AWS re:Invent

DECEMBER 2 – 6, 2024 | LAS VEGAS, NV

ARC311

Scaling Prime Video for peak NFL streaming on AWS

Elliott Nash
(He/Him)
Head of Infrastructure Scaling
Prime Video

Ralph Chaker
(He/Him)
Principal Product Manager
Prime Video

Tulip Gupta
(She/Her)
Sr. Solutions Architect, AWS Strategic Accounts
AWS

© 2024, Amazon Web Services, Inc. or its affiliates. All rights reserved.

Agenda

- Introduction to NFL Thursday Night Football on Prime Video
- Prime Video multi-Region scaling strategies
- Prime Video elastic scaling strategies
- AWS benefits – AWS Well-Architected Framework

© 2024, Amazon Web Services, Inc. or its affiliates. All rights reserved.

120+

© 2024, Amazon Web Services, Inc. or its affiliates. All rights reserved.

prime video

Business opportunity

Cost-effectively scale and delight millions of subscribers around the world during live sports

Challenge

Scale out infrastructure to meet the demands of NFL games viewership

Solution

Adopting autoscaling and multi-Region architecture

© 2024, Amazon Web Services, Inc. or its affiliates. All rights reserved.

FAQs: Thursday Night Football on Prime Video

1. September 7, 2017, the first NFL game was co-exclusively streamed on Prime Video

2. In 2022, exclusively streamed on Prime Video

3. In 2023, 17 million fans watched TNF on Prime Video

4. 2024 Season 3 average was 18% higher than 2023

5. 2024 biggest NFL event in Prime Video history

© 2024, Amazon Web Services, Inc. or its affiliates. All rights reserved.

A multi-Region scaling story

We are global

240+

Countries and territories

30+

Languages

Chinese
Danish
Dutch
English
French
German
Hindi

Indonesian
Norwegian
Polish
Portuguese
Italian
Japanese
Korean

Russian
Spanish
Swedish
Tamil
Telugu
Thai
Turkish

200M+

Prime members worldwide

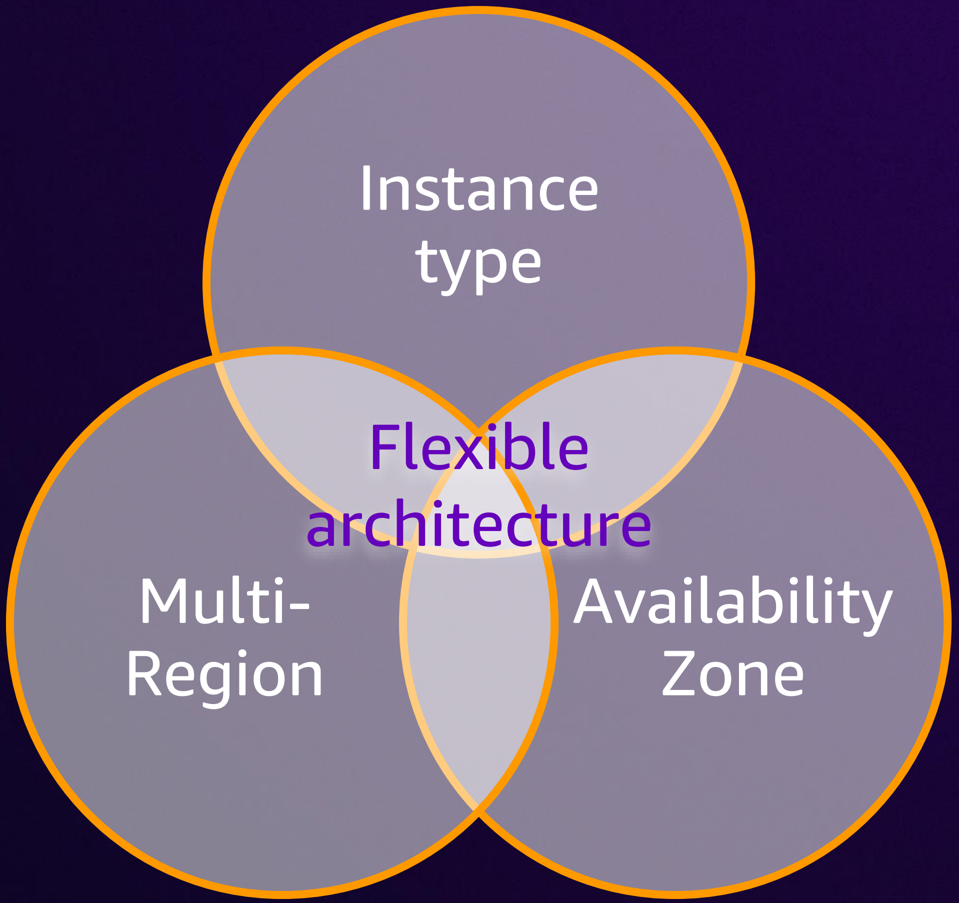
prime video

© 2024, Amazon Web Services, Inc. or its affiliates. All rights reserved.

Designing for high availability

- Availability requirements:
  - Better than broadcast
  - Live content is perishable

- Flexibility across three layers

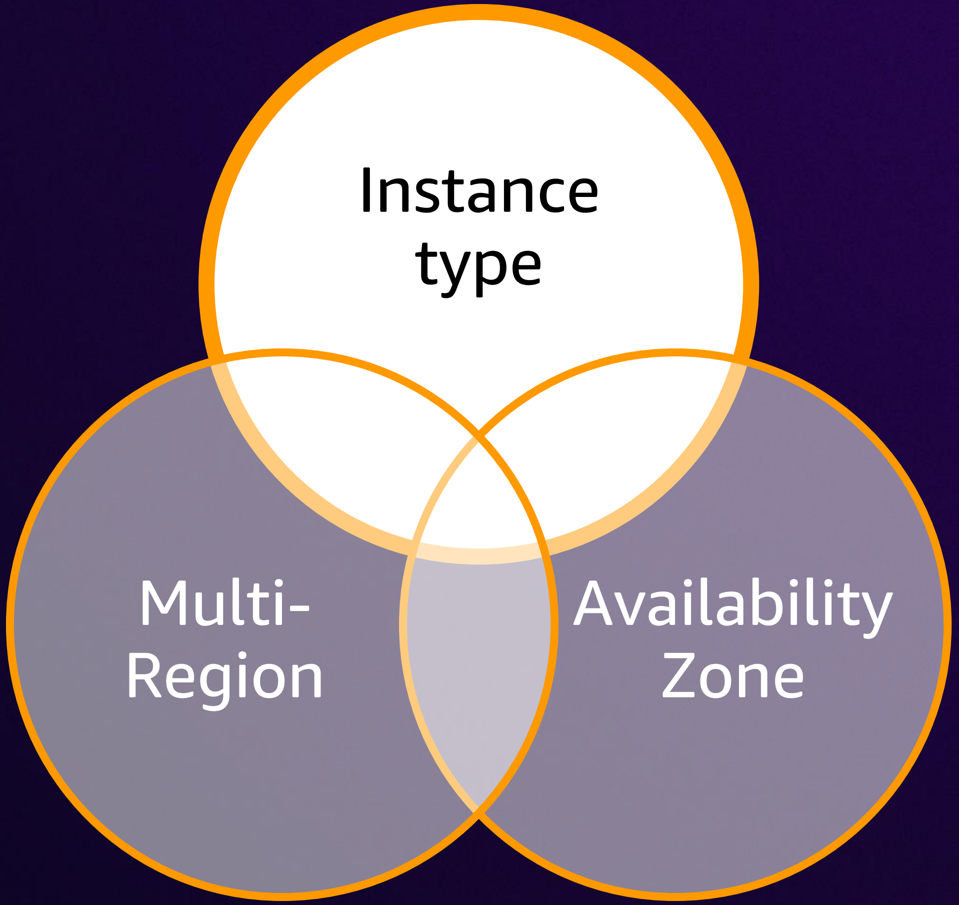


© 2024, Amazon Web Services, Inc. or its affiliates. All rights reserved.

Designing for high availability

- Availability requirements:
  - Better than broadcast
  - Live content is perishable

- Flexibility across three layers

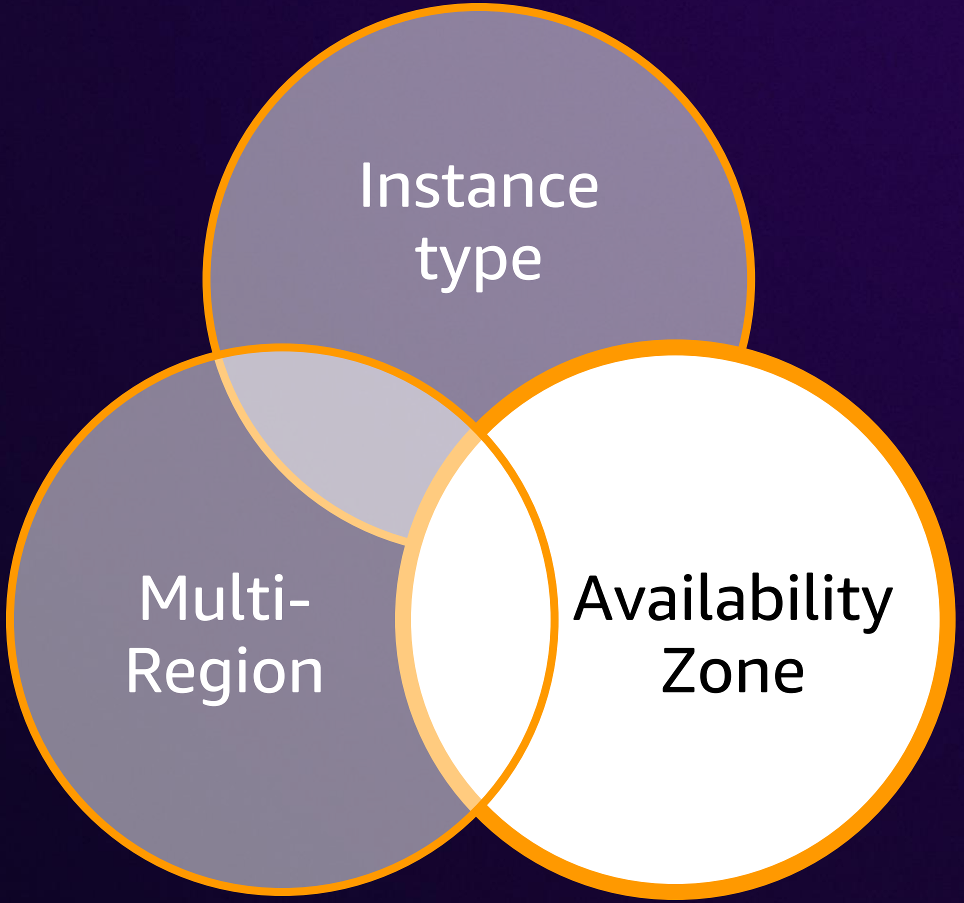


© 2024, Amazon Web Services, Inc. or its affiliates. All rights reserved.

Designing for high availability

- Availability requirements:
  - Better than broadcast
  - Live content is perishable

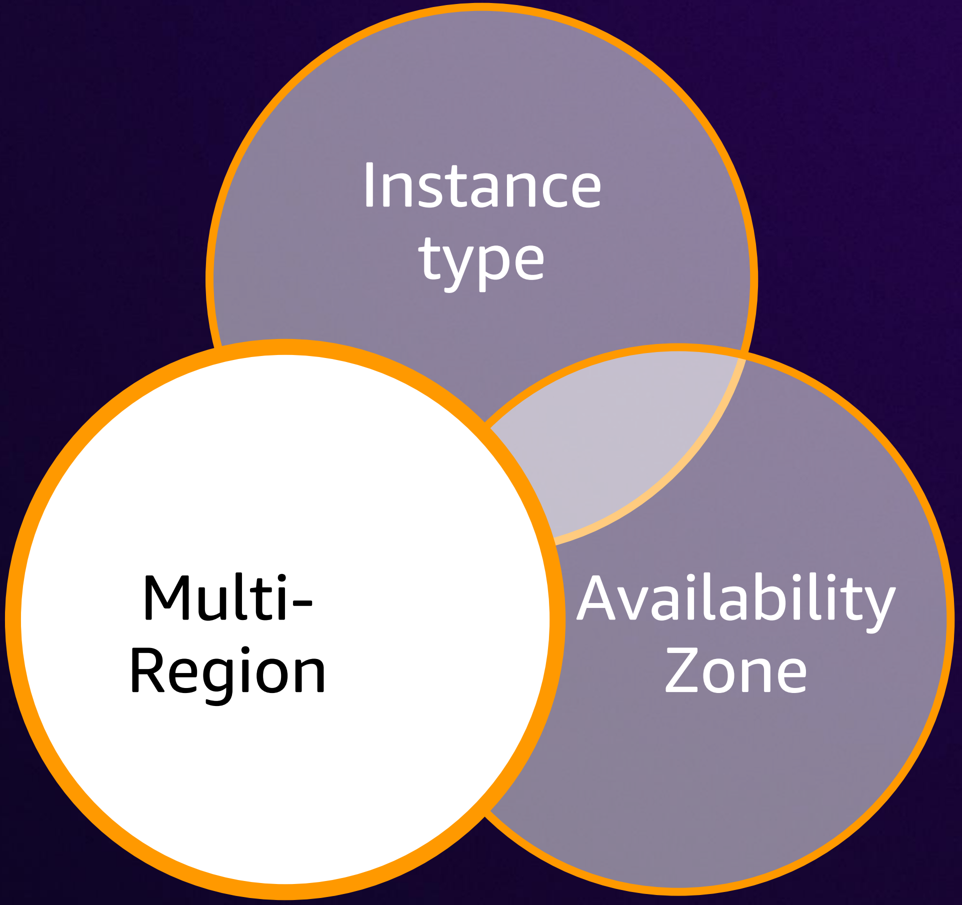
- Flexibility across three layers



© 2024, Amazon Web Services, Inc. or its affiliates. All rights reserved.

Designing for high availability

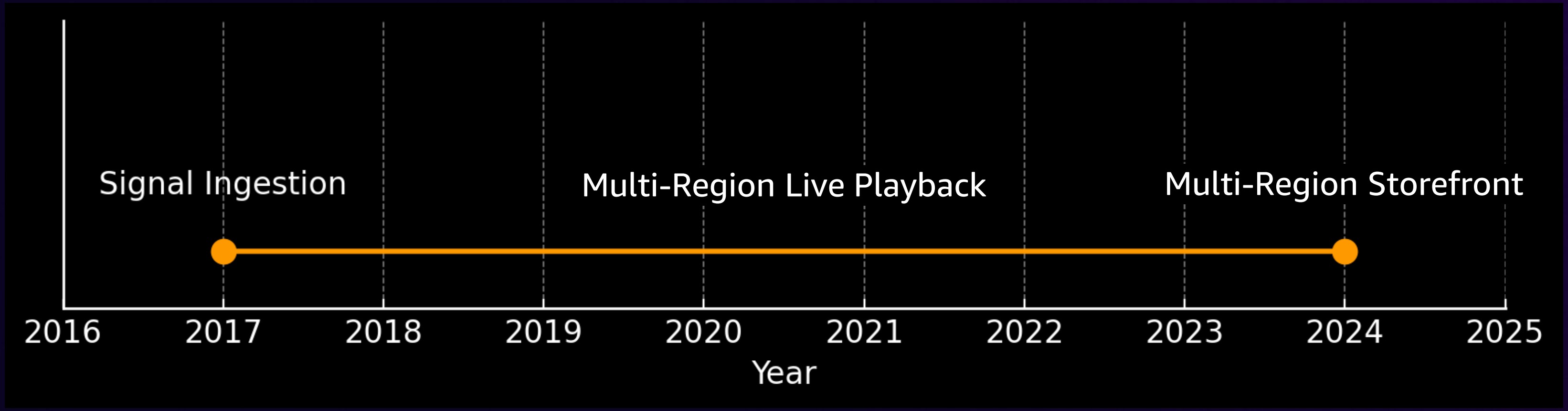
- Availability requirements:
  - Better than broadcast
  - Live content is perishable
- Flexibility across three layers



© 2024, Amazon Web Services, Inc. or its affiliates. All rights reserved.

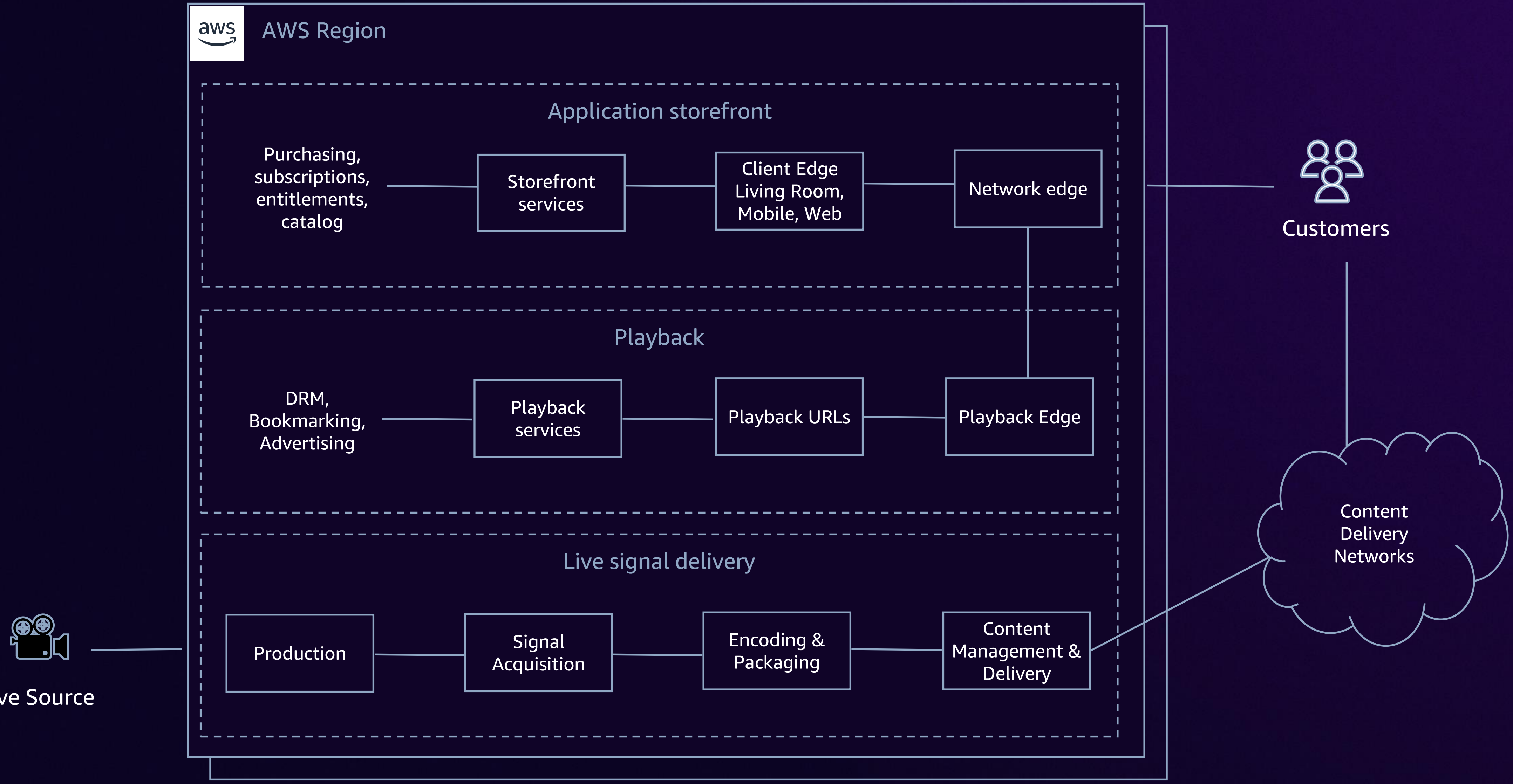
Architecture evolution

- 2017: Launch TNF on Prime Video (non-exclusive)
- 2022: First TNF exclusive season
- 2024: All critical customer journeys support multi-Region



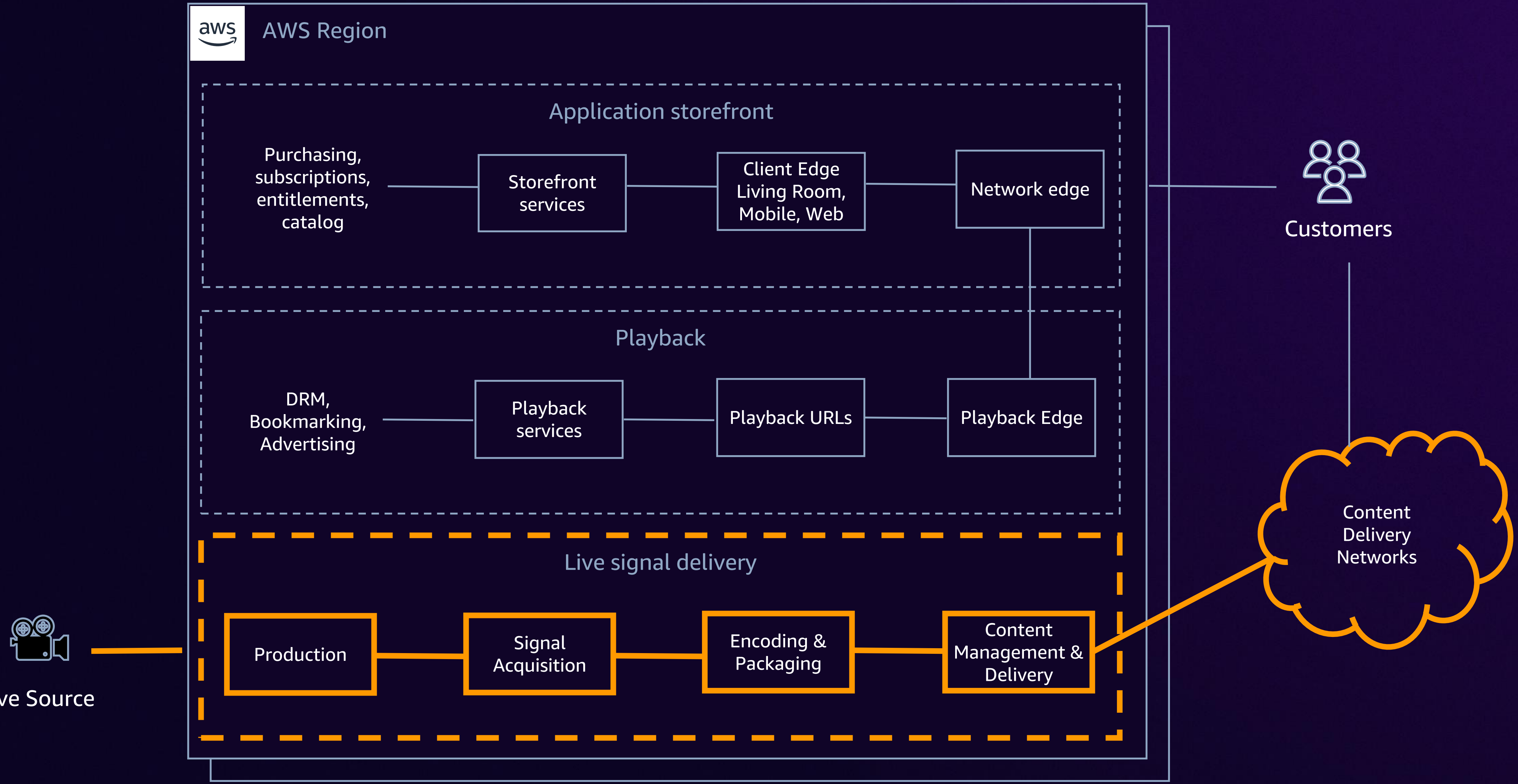
© 2024, Amazon Web Services, Inc. or its affiliates. All rights reserved.

10,000 ft architecture



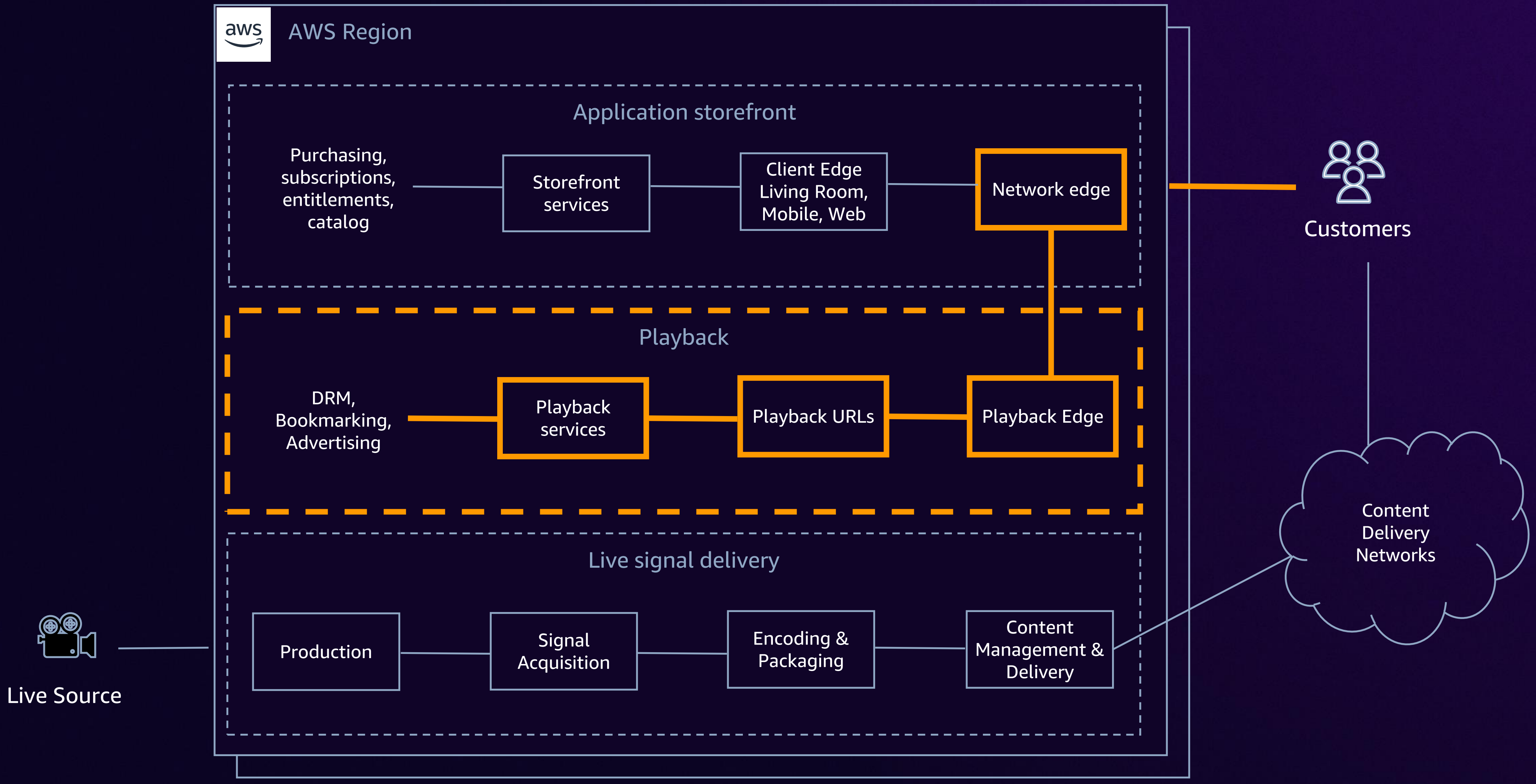
© 2024, Amazon Web Services, Inc. or its affiliates. All rights reserved.

10,000 ft architecture



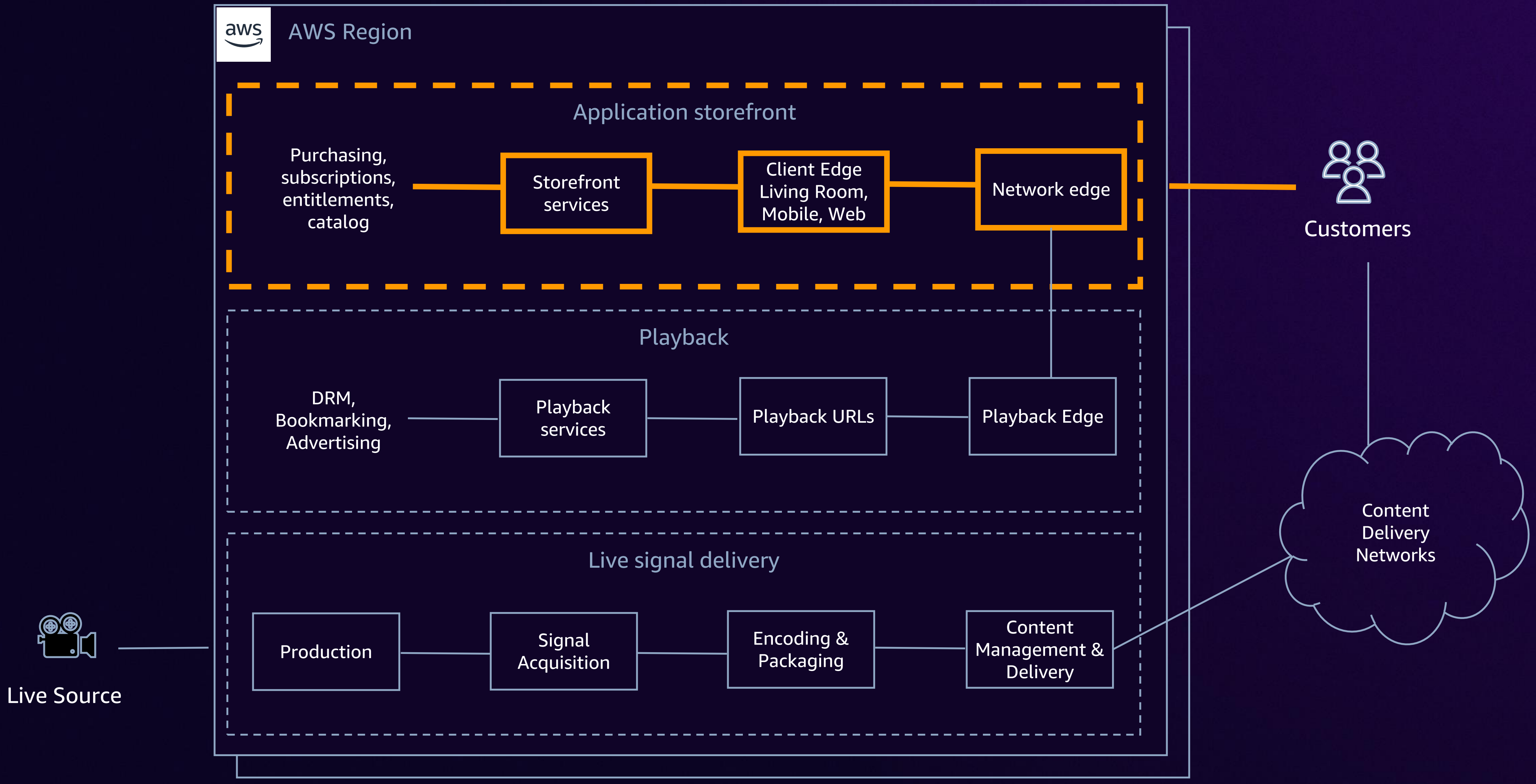
© 2024, Amazon Web Services, Inc. or its affiliates. All rights reserved.

10,000 ft architecture



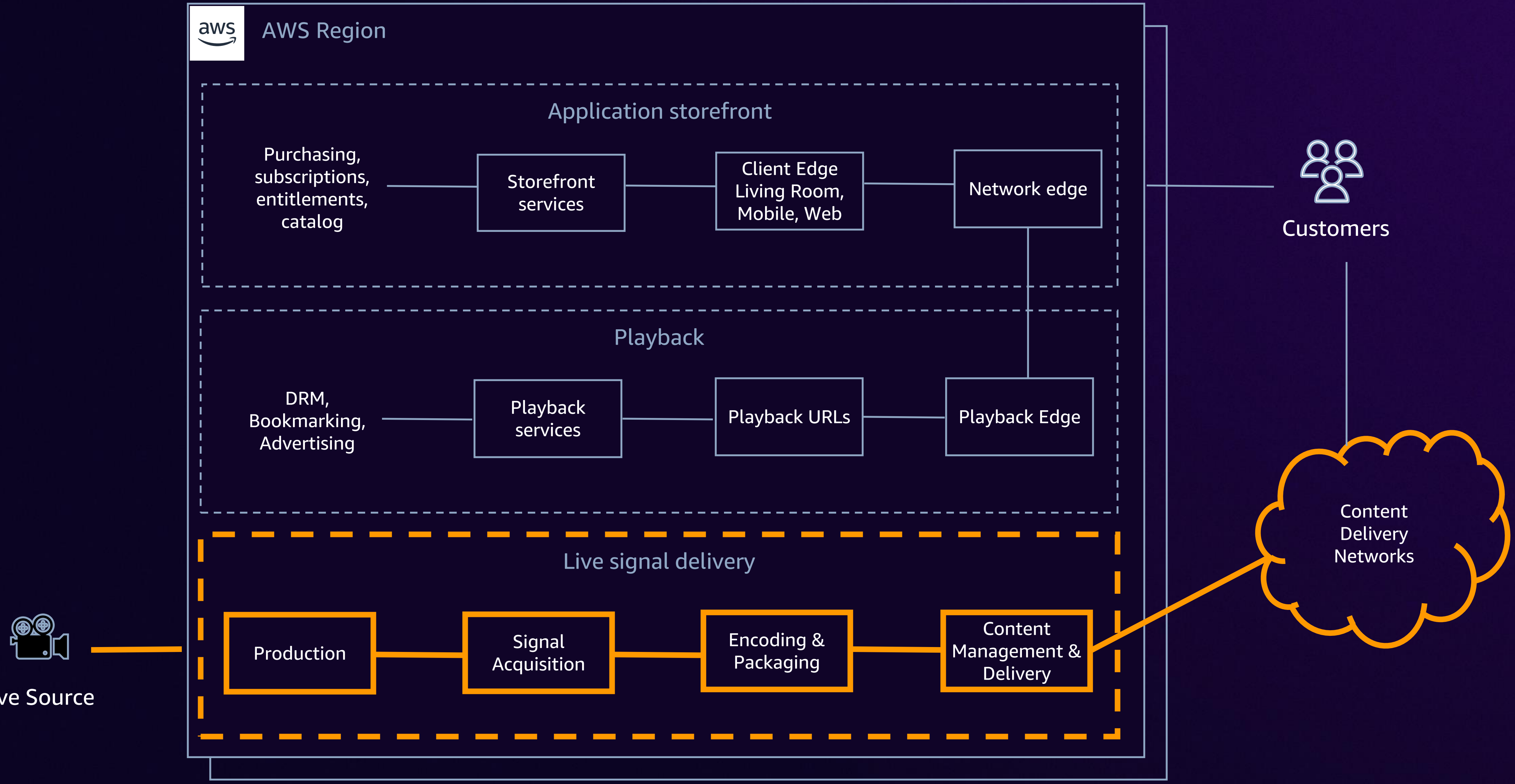
© 2024, Amazon Web Services, Inc. or its affiliates. All rights reserved.

10,000 ft architecture



© 2024, Amazon Web Services, Inc. or its affiliates. All rights reserved.

10,000 ft architecture



© 2024, Amazon Web Services, Inc. or its affiliates. All rights reserved.

Signal ingestion: Stadium source

© 2024, Amazon Web Services, Inc. or its affiliates. All rights reserved.

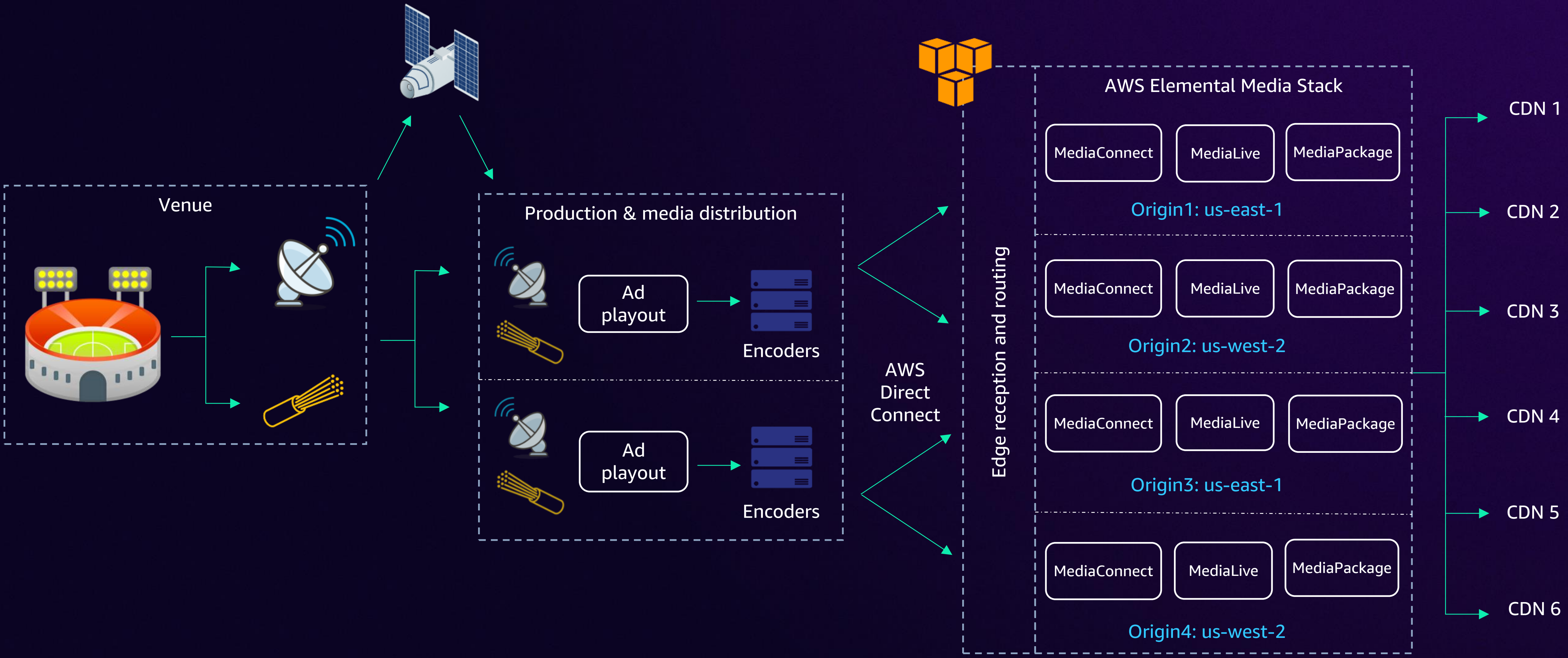
Signal ingestion: Mobile production studio

© 2024, Amazon Web Services, Inc. or its affiliates. All rights reserved.

Signal ingestion: Eyes on glass

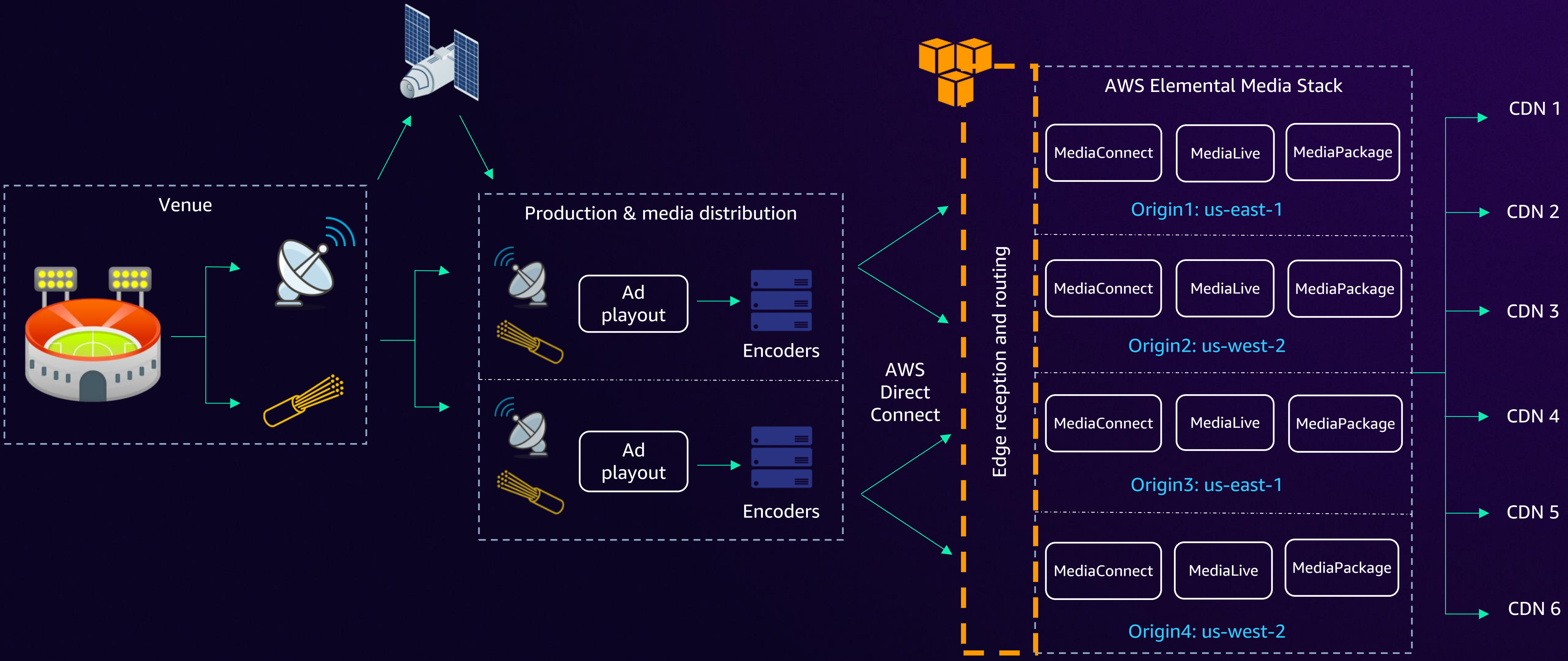
© 2024, Amazon Web Services, Inc. or its affiliates. All rights reserved.

Signal ingestion: Architecture deep dive



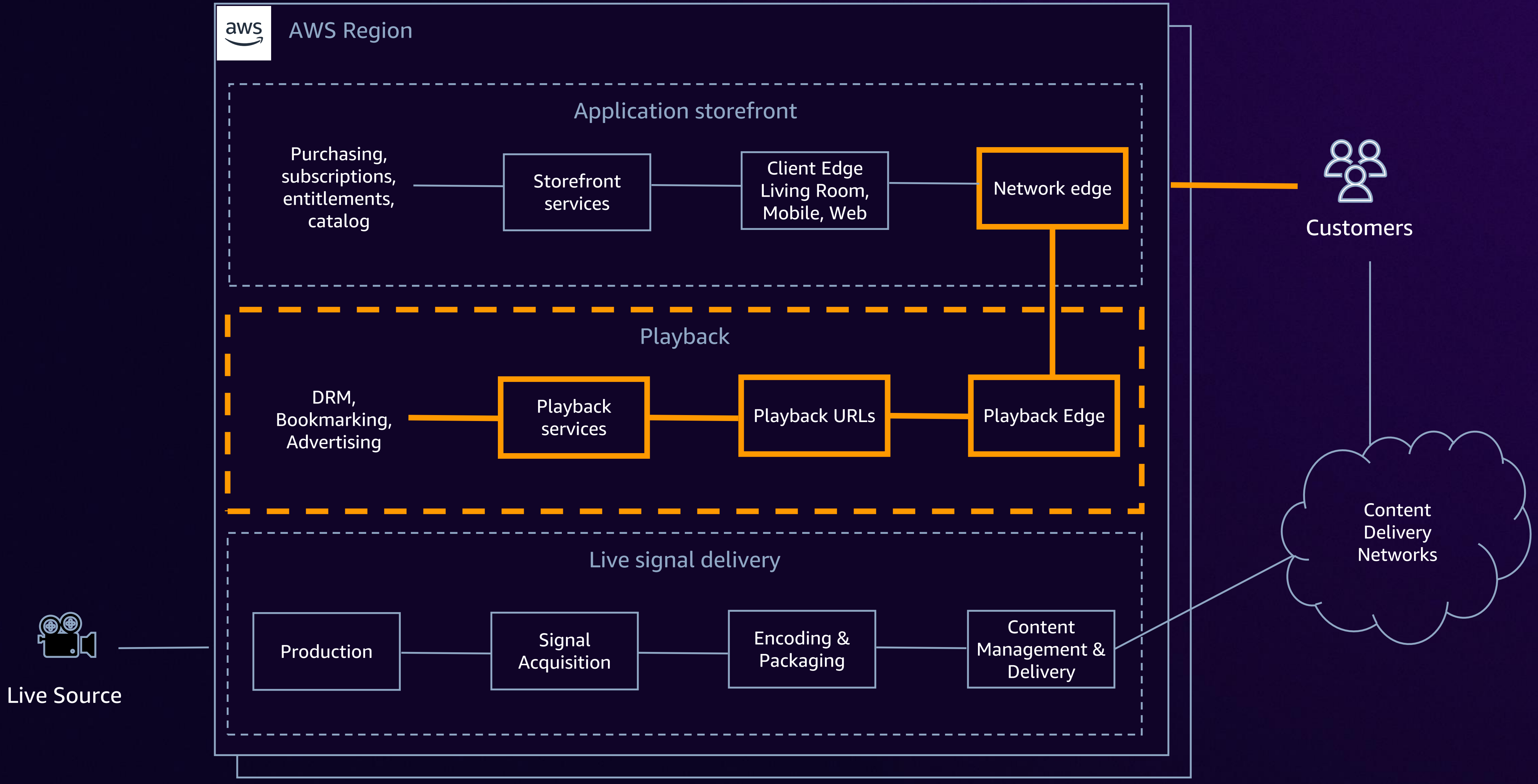
© 2024, Amazon Web Services, Inc. or its affiliates. All rights reserved.

Signal ingestion: Architecture deep dive



© 2024, Amazon Web Services, Inc. or its affiliates. All rights reserved.

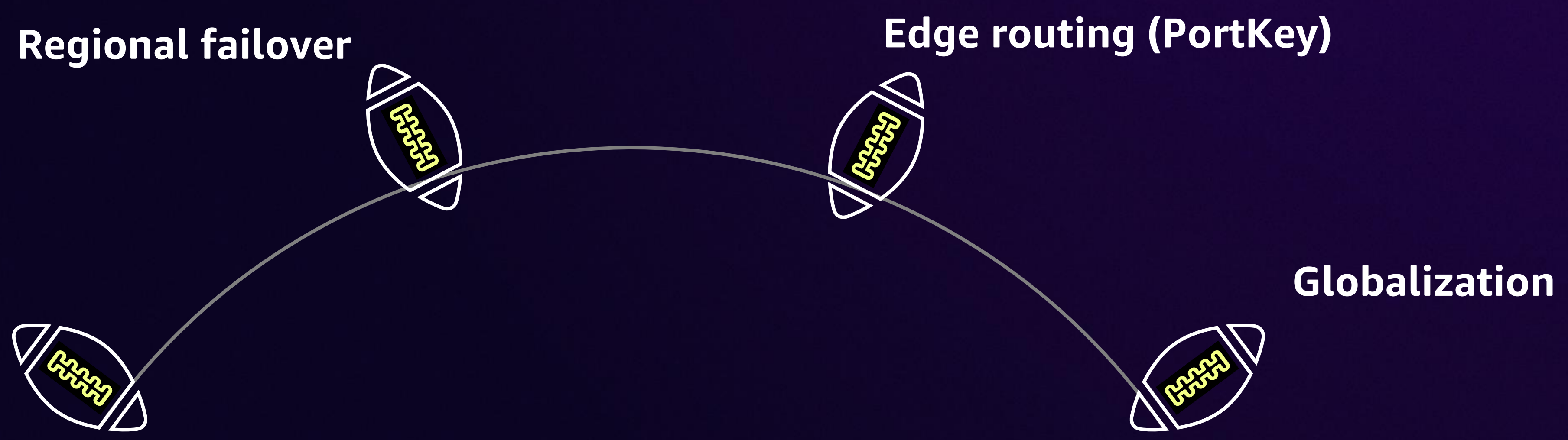
10,000 ft architecture



© 2024, Amazon Web Services, Inc. or its affiliates. All rights reserved.

Multi-Region playback

Enhanced availability and resilience for NFL Thursday Night Football



Active/active

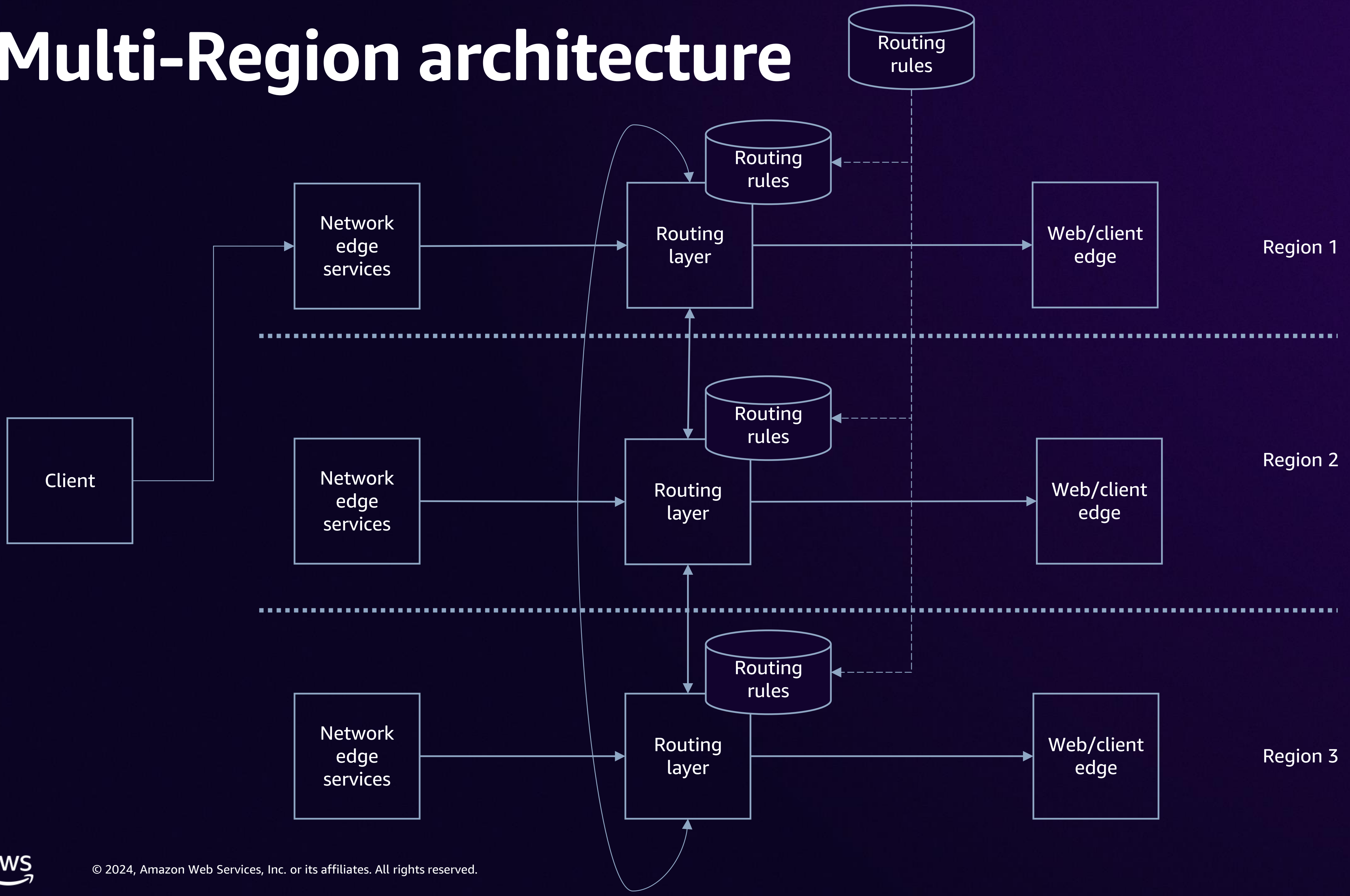
Regional failover

Edge routing (PortKey)

Globalization

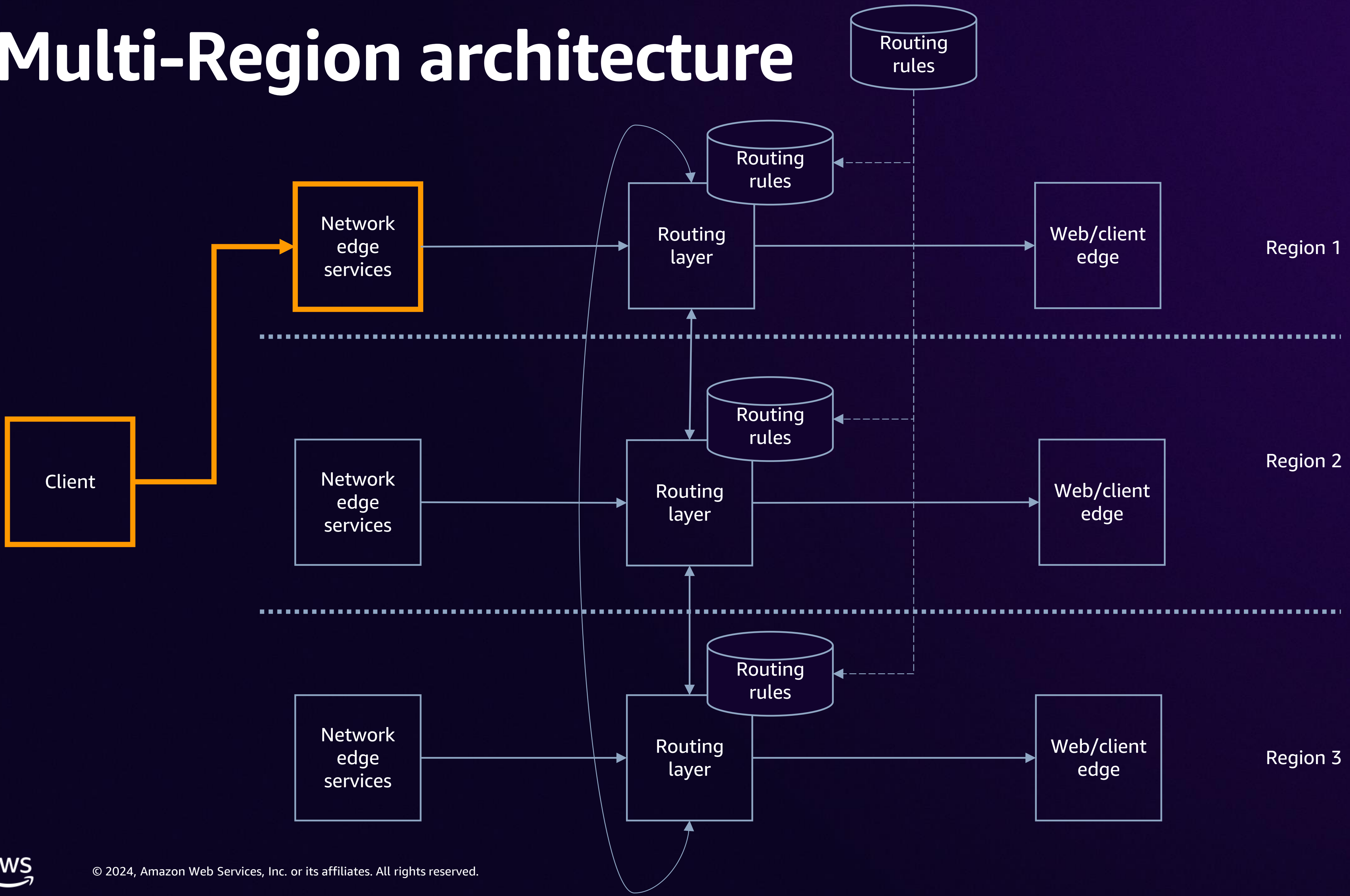
© 2024, Amazon Web Services, Inc. or its affiliates. All rights reserved.

Multi-Region architecture



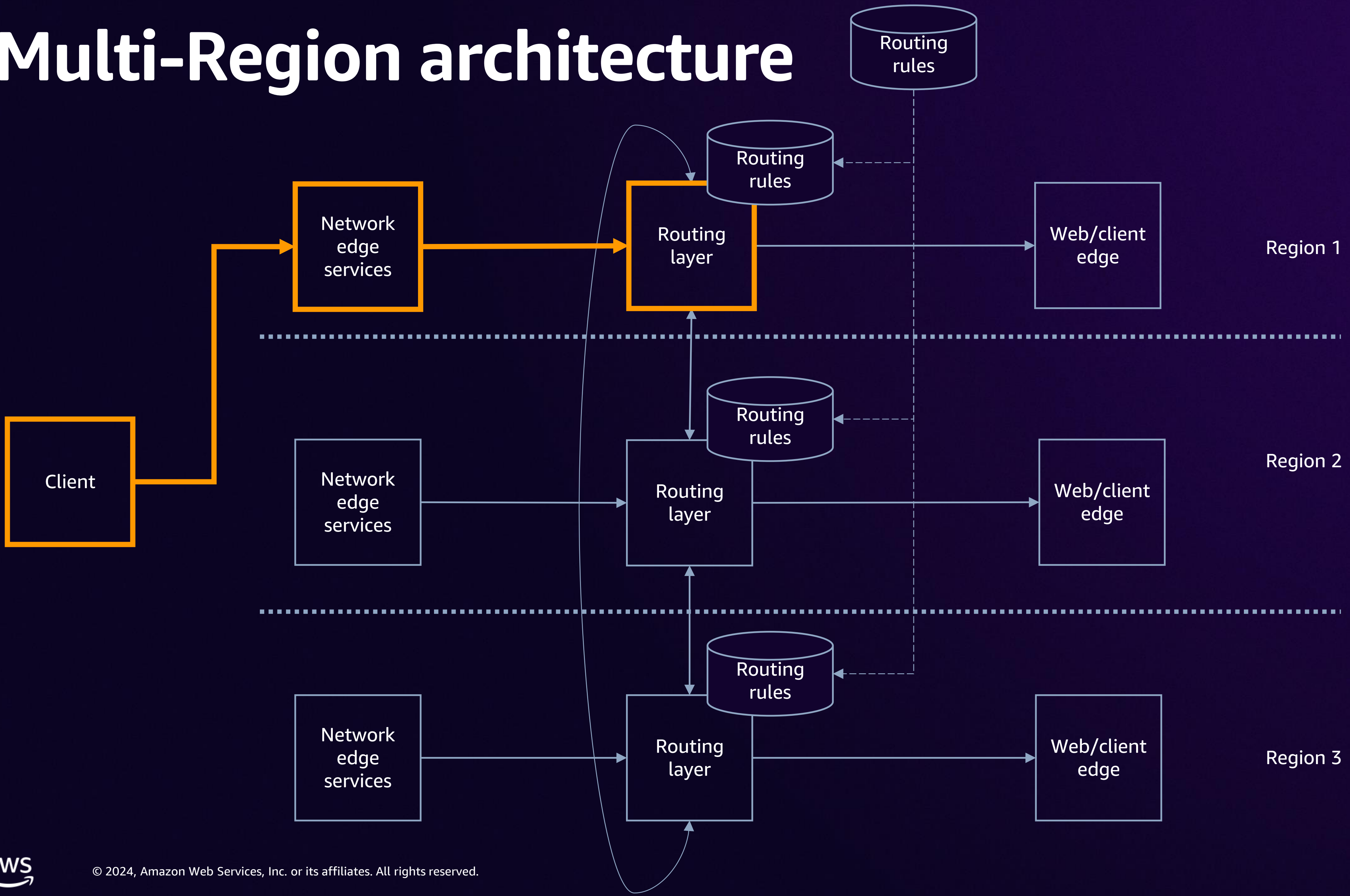
© 2024, Amazon Web Services, Inc. or its affiliates. All rights reserved.

Multi-Region architecture



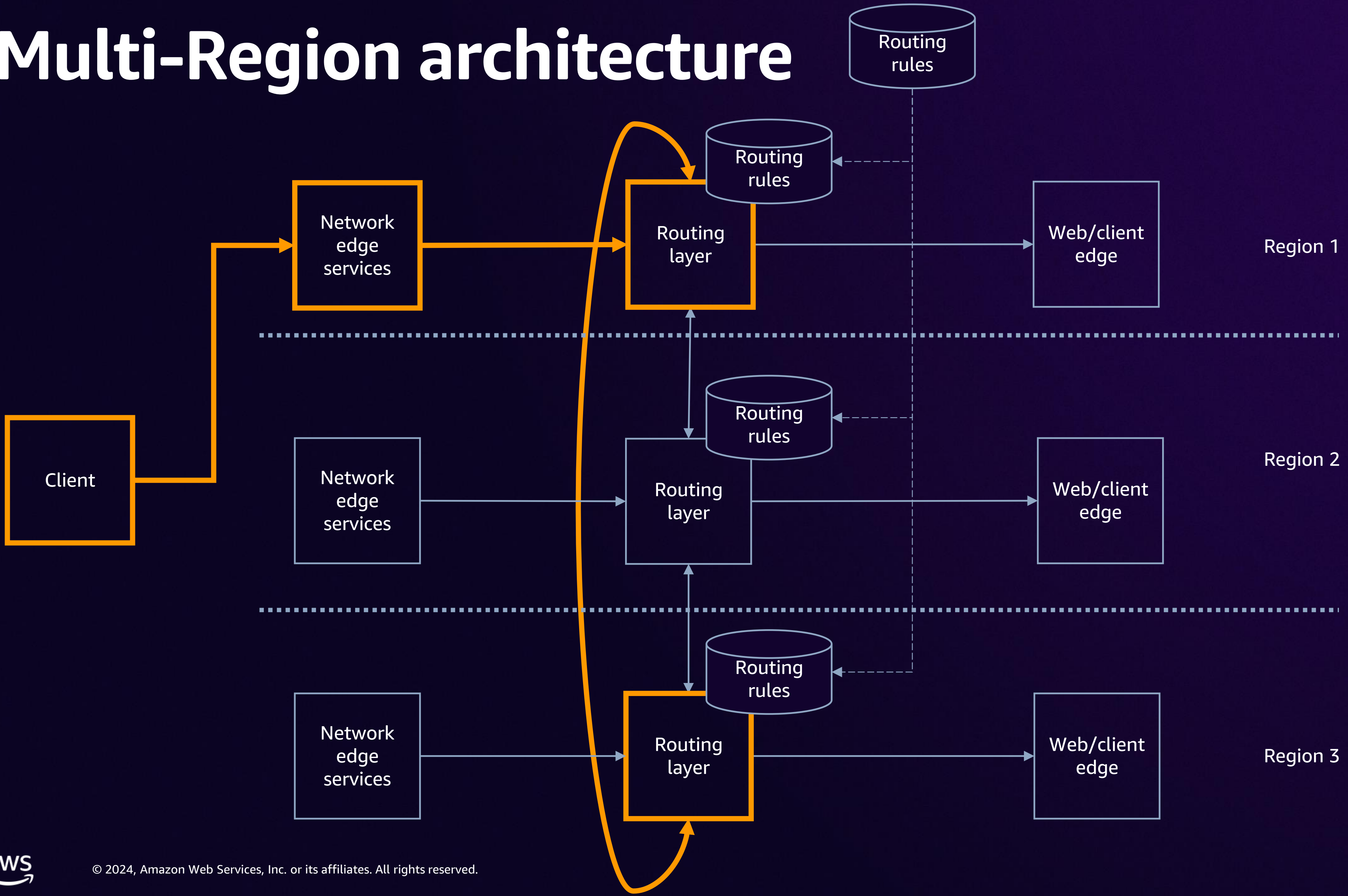
© 2024, Amazon Web Services, Inc. or its affiliates. All rights reserved.

Multi-Region architecture



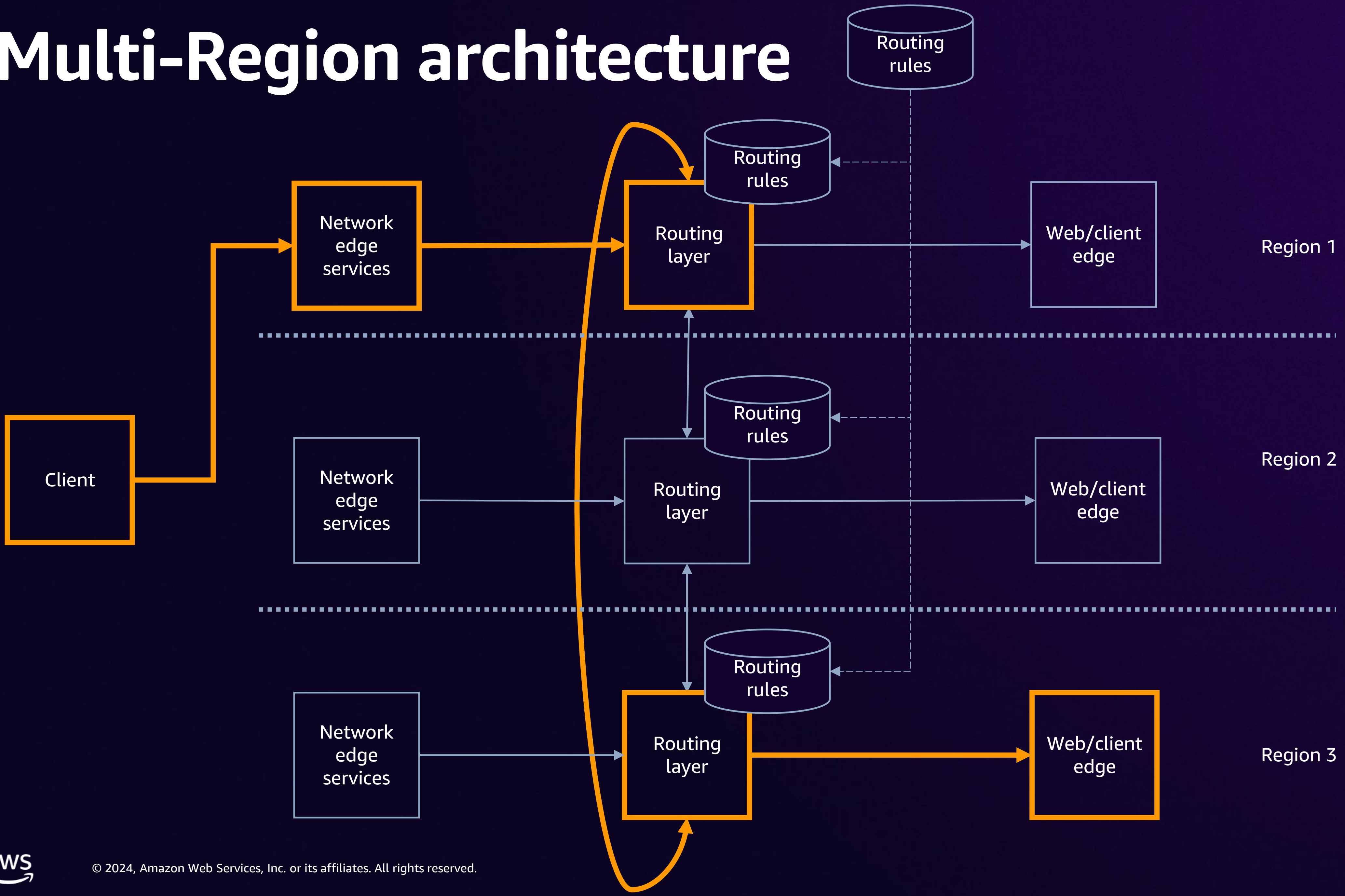
© 2024, Amazon Web Services, Inc. or its affiliates. All rights reserved.

Multi-Region architecture



© 2024, Amazon Web Services, Inc. or its affiliates. All rights reserved.

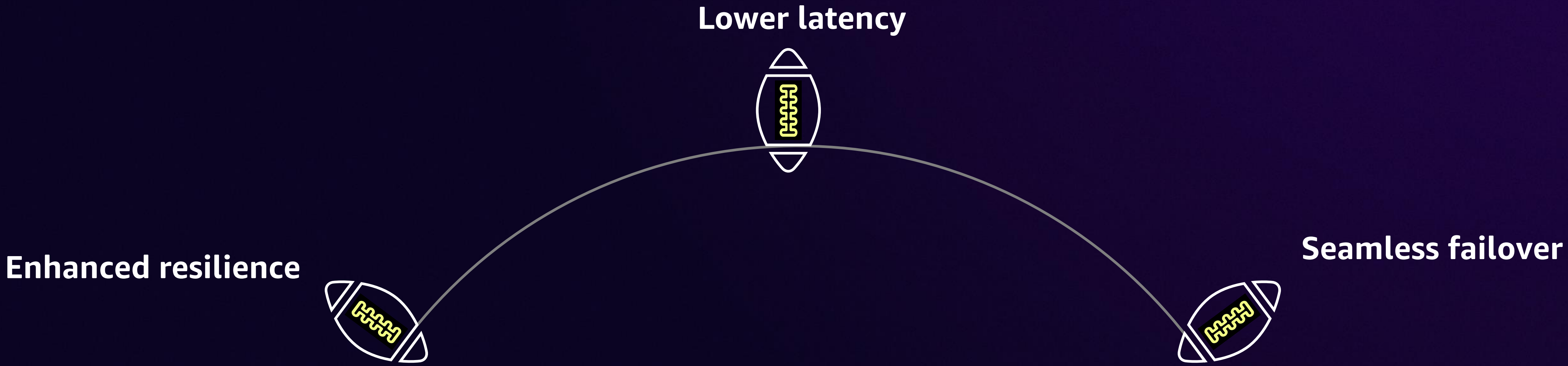
Multi-Region architecture



© 2024, Amazon Web Services, Inc. or its affiliates. All rights reserved.

Key benefits of multi-Region playback

Enhanced availability and resilience for NFL Thursday Night Football



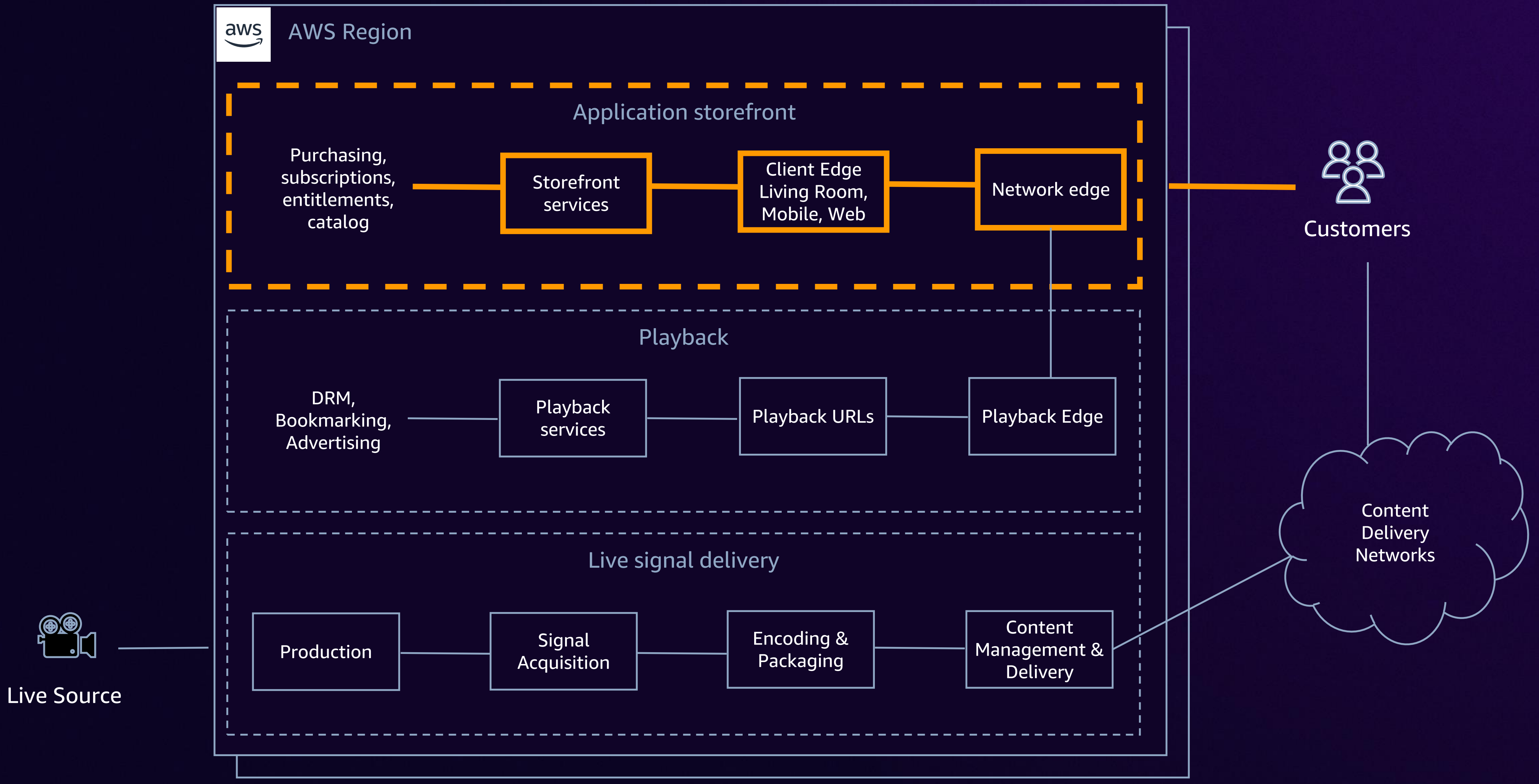
Lower latency

Enhanced resilience

Seamless failover

© 2024, Amazon Web Services, Inc. or its affiliates. All rights reserved.

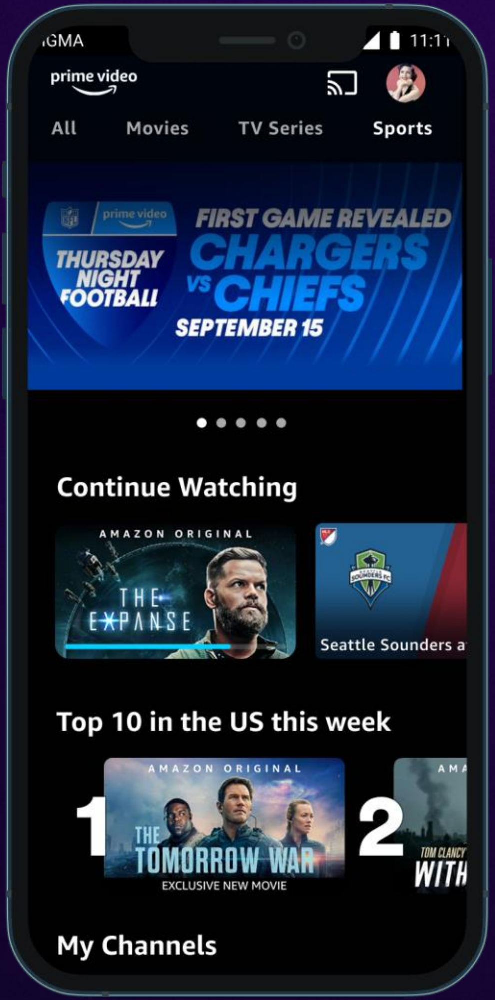
10,000 ft architecture



© 2024, Amazon Web Services, Inc. or its affiliates. All rights reserved.

Multi-Region storefront and detail pages

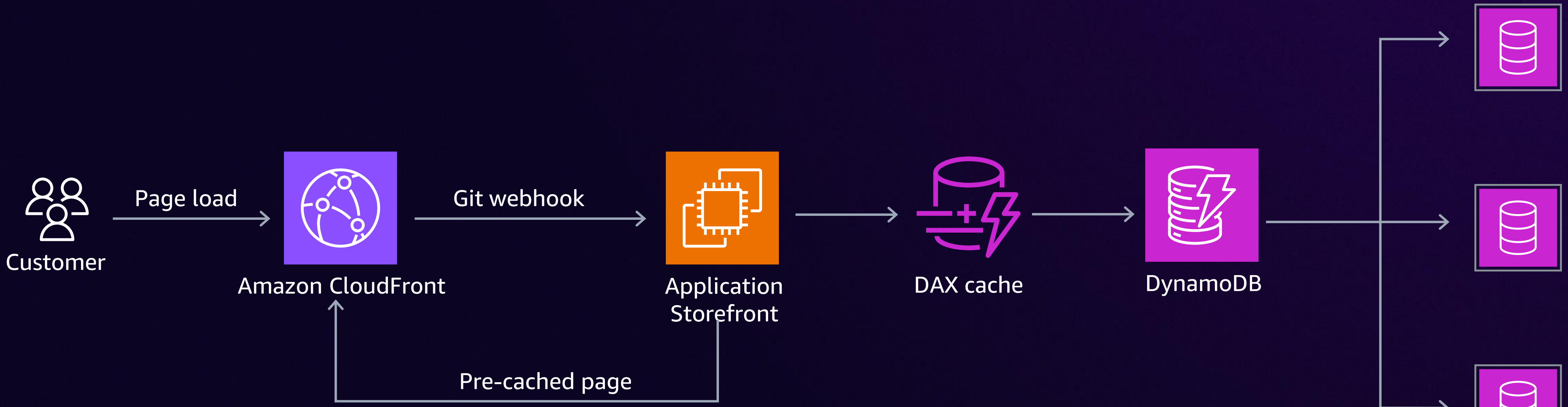
- App start, home page, and detail page
- Globalization: Feature parity between Regions
- Data replication and consistency
- Shadow testing
- PhasedlLaunch



© 2024, Amazon Web Services, Inc. or its affiliates. All rights reserved.

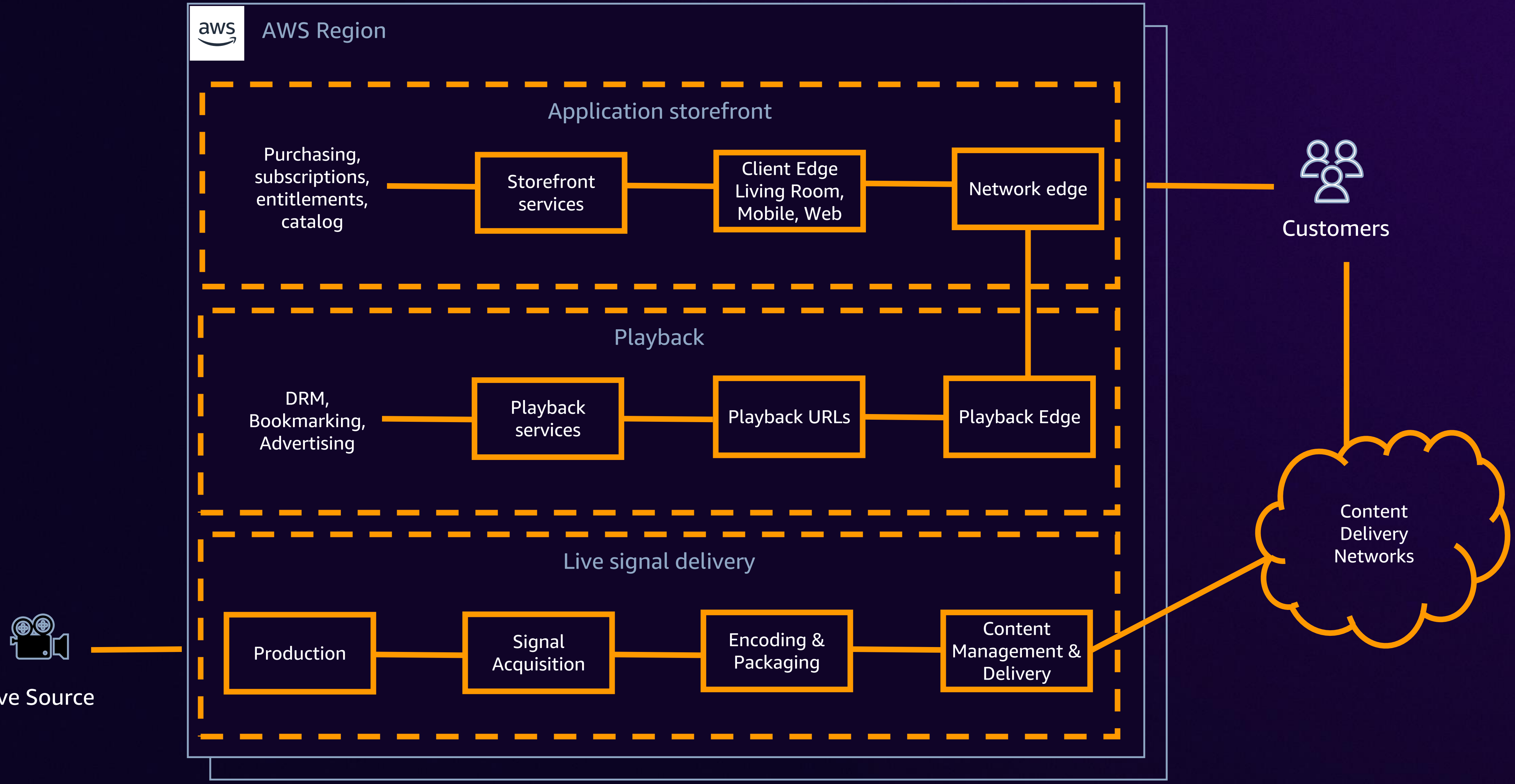
Data replication and consistency

- Leveraged Amazon DynamoDB global tables for low-latency access to critical metadata
- Bookmarking multi-master cross-region replication strategy



© 2024, Amazon Web Services, Inc. or its affiliates. All rights reserved.

10,000 ft architecture



© 2024, Amazon Web Services, Inc. or its affiliates. All rights reserved.

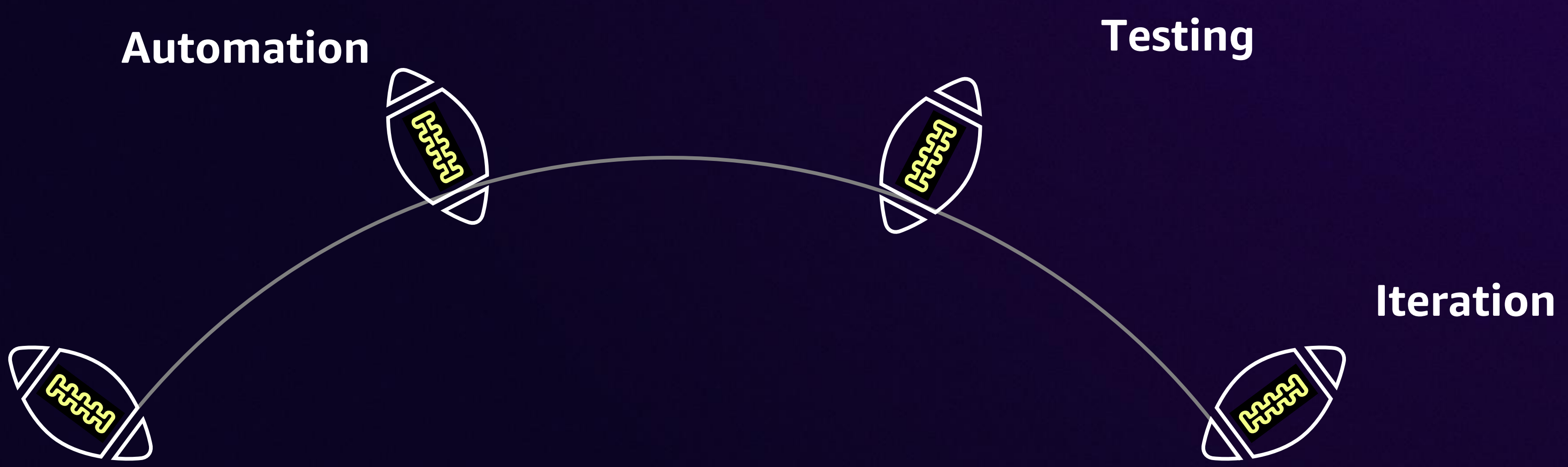
Real-world success stories

- In 2024 we served 50% of customers from another Region
- Millions of concurrent users
- Smooth user experiences despite high demand surges

© 2024, Amazon Web Services, Inc. or its affiliates. All rights reserved.

Lessons learned and future enhancements

Enhanced availability and resilience for NFL Thursday Night Football



© 2024, Amazon Web Services, Inc. or its affiliates. All rights reserved.

An elastic scaling story

TRIVIA

In 2024, what was the peak audience for the Cowboys-Giants Thursday Night Football game on Prime Video?

- A. 18 million
- B. 10 million
- C. 15 million
- D. 20 million

© 2024, Amazon Web Services, Inc. or its affiliates. All rights reserved.

TRIVIA

In 2024, what was the peak audience for the Cowboys-Giants Thursday Night Football game on Prime Video?

A. 18 million

B. 10 million

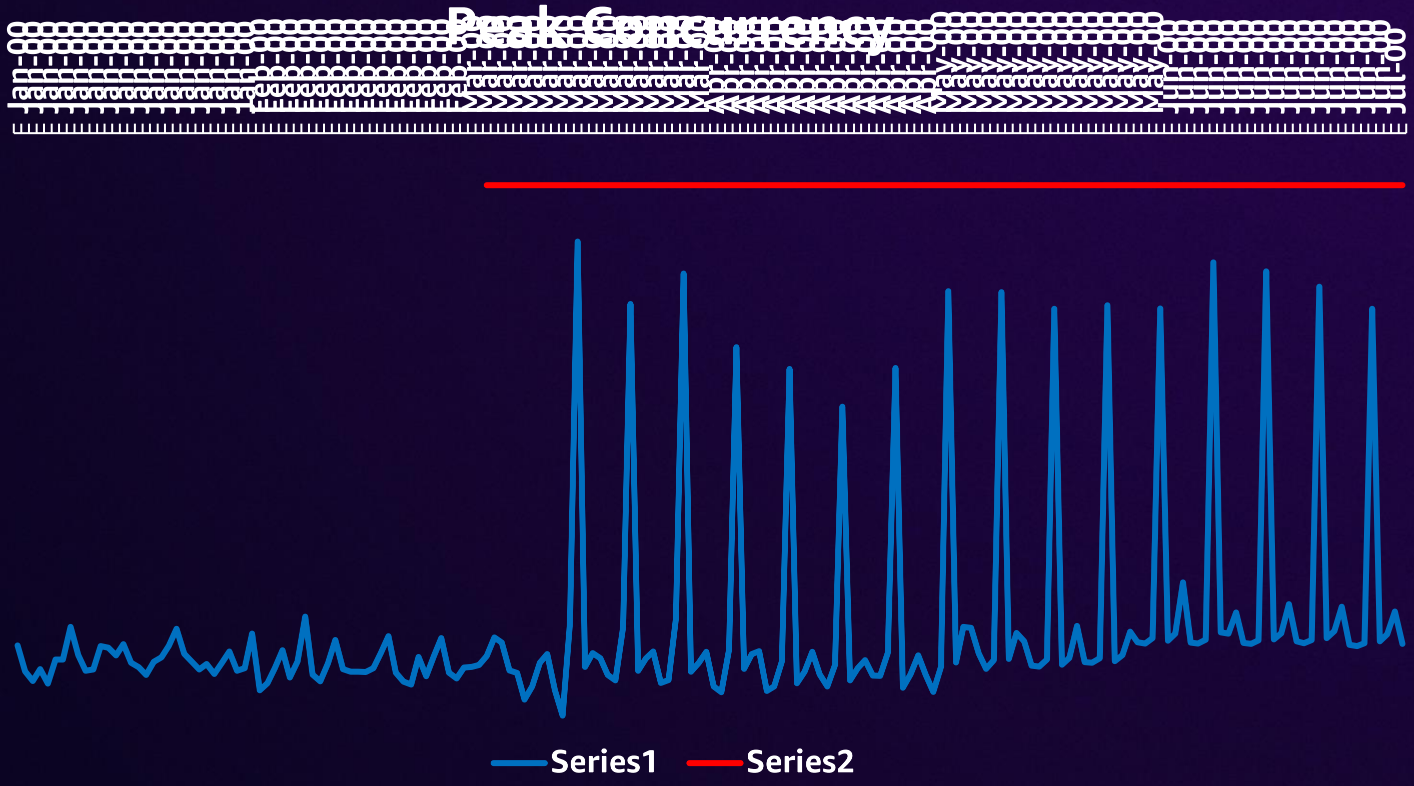
C. 15 million

D. 20 million

© 2024, Amazon Web Services, Inc. or its affiliates. All rights reserved.

Scaling for Live Sports at Prime Video

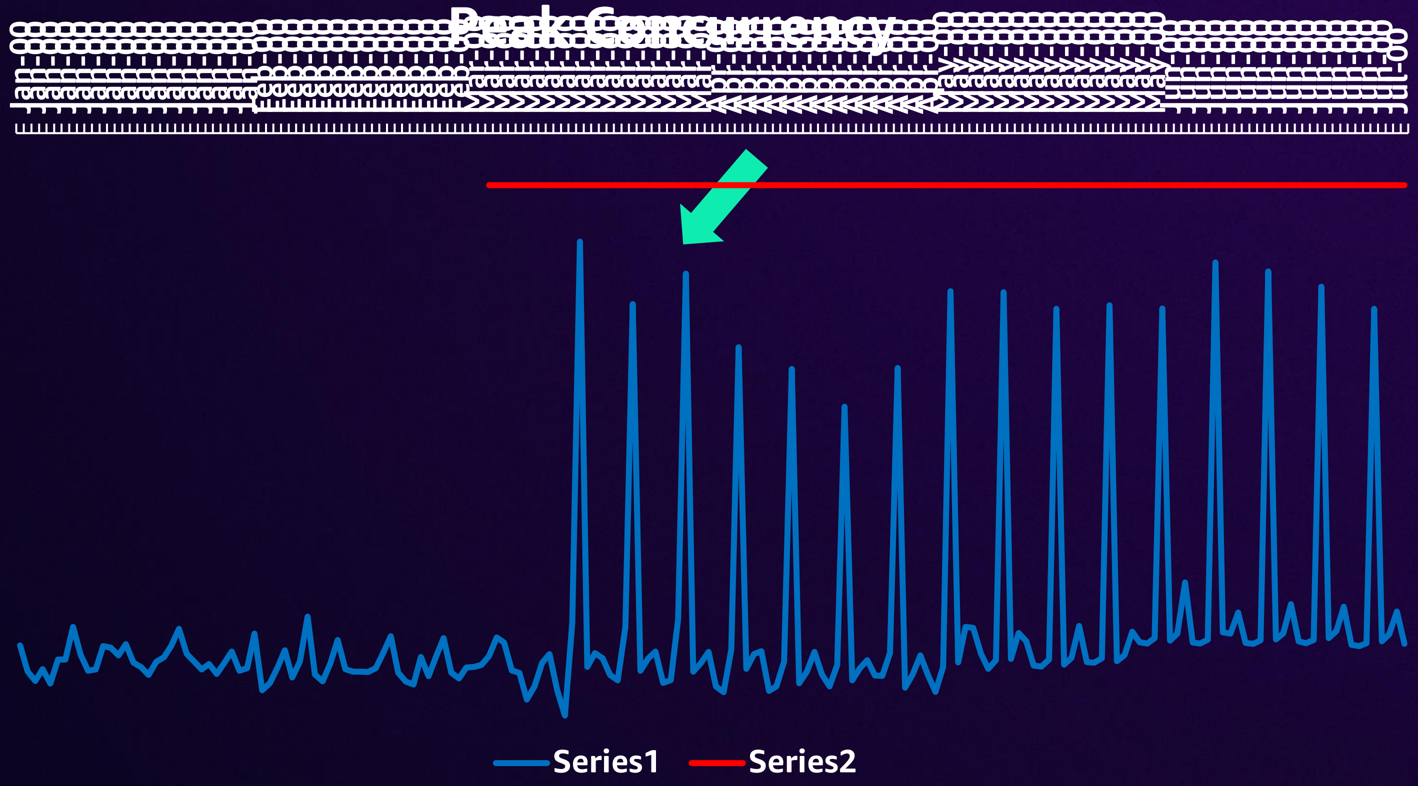
- Impact of live sports
- Prime Video’s role in streaming Thursday Night Football (TNF)



© 2024, Amazon Web Services, Inc. or its affiliates. All rights reserved.

Scaling for Live Sports at Prime Video

- Impact of live sports
- Prime Video’s role in streaming Thursday Night Football (TNF)



© 2024, Amazon Web Services, Inc. or its affiliates. All rights reserved.

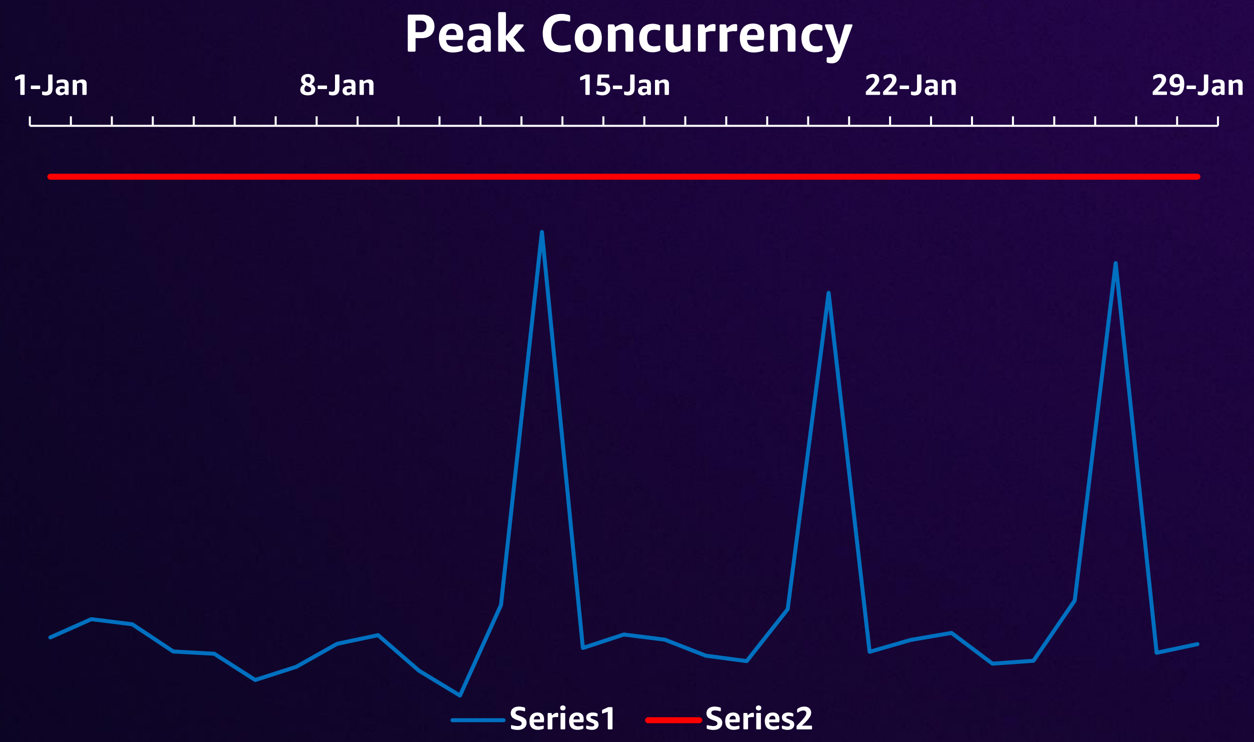
Scaling for Live Sports at Prime Video

- Impact of live sports
- Prime Video's role in streaming Thursday Night Football (TNF)

© 2024, Amazon Web Services, Inc. or its affiliates. All rights reserved.

Key challenges to scaling for peak NFL events

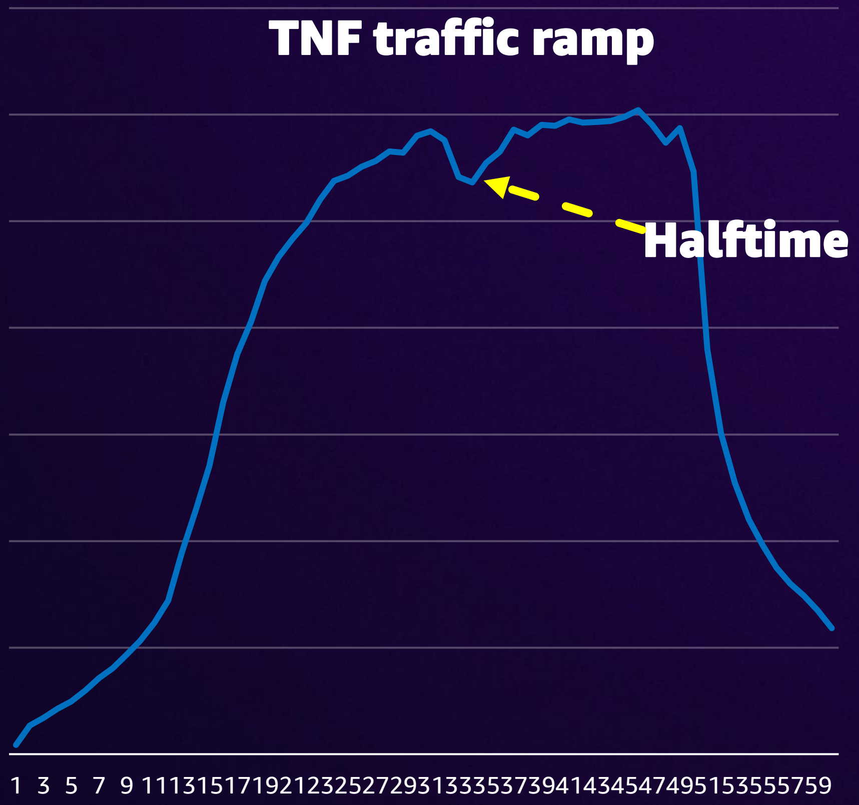
- Overall scale
- Spiky traffic patterns
- Coordination Effort & Lead time
- Lack of Automation



© 2024, Amazon Web Services, Inc. or its affiliates. All rights reserved.

Key challenges to scaling for peak NFL events

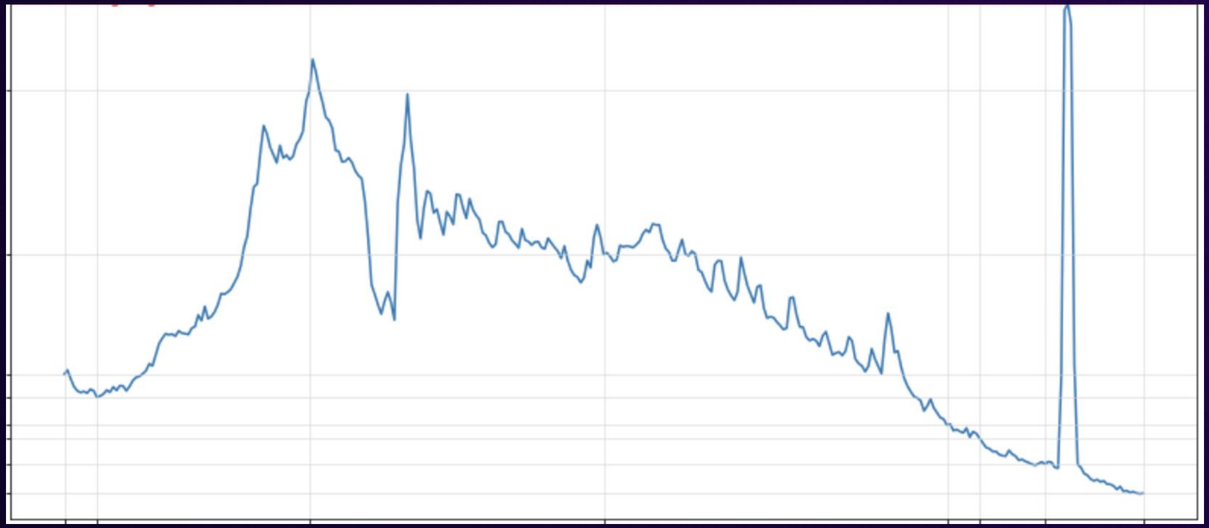
- Overall scale
- Spiky traffic patterns
- Coordination Effort & Lead time
- Lack of Automation



© 2024, Amazon Web Services, Inc. or its affiliates. All rights reserved.

Key challenges to scaling for peak NFL events

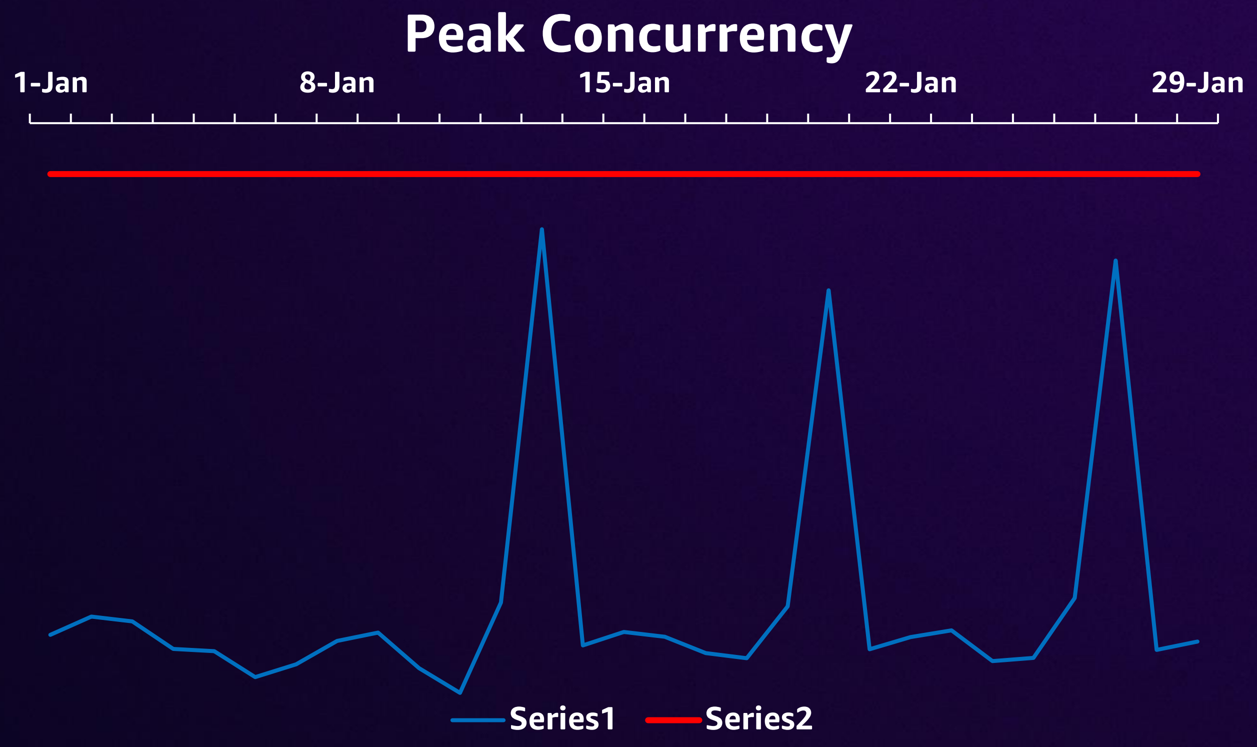
- Overall scale
- Spiky traffic patterns
- Coordination Effort & Lead time
- Lack of Automation



© 2024, Amazon Web Services, Inc. or its affiliates. All rights reserved.

Key challenges to scaling for peak NFL events

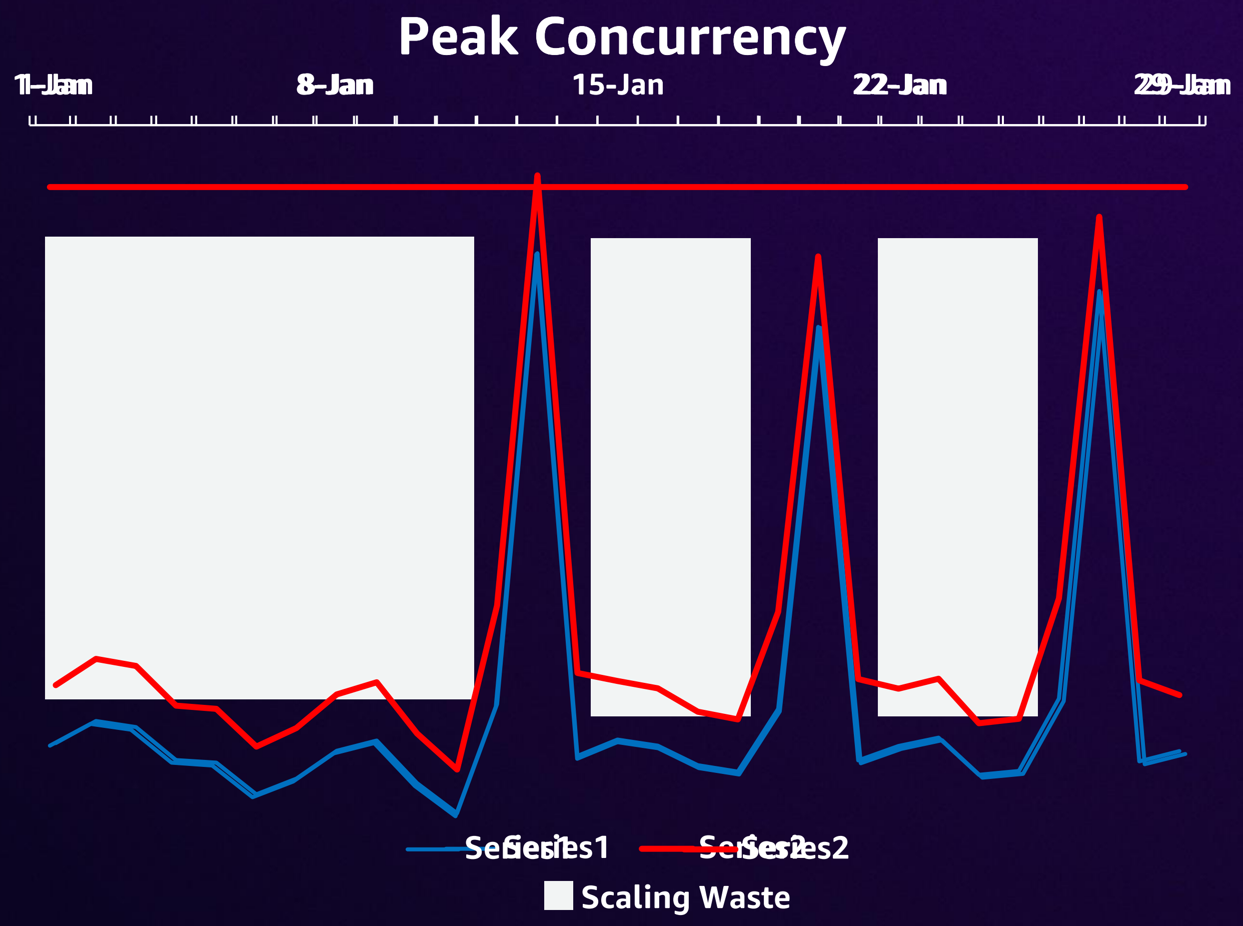
- Overall scale
- Spiky traffic patterns
- Coordination Effort & Lead time
- Lack of Automation



© 2024, Amazon Web Services, Inc. or its affiliates. All rights reserved.

Opportunity with auto scaling (elasticity)

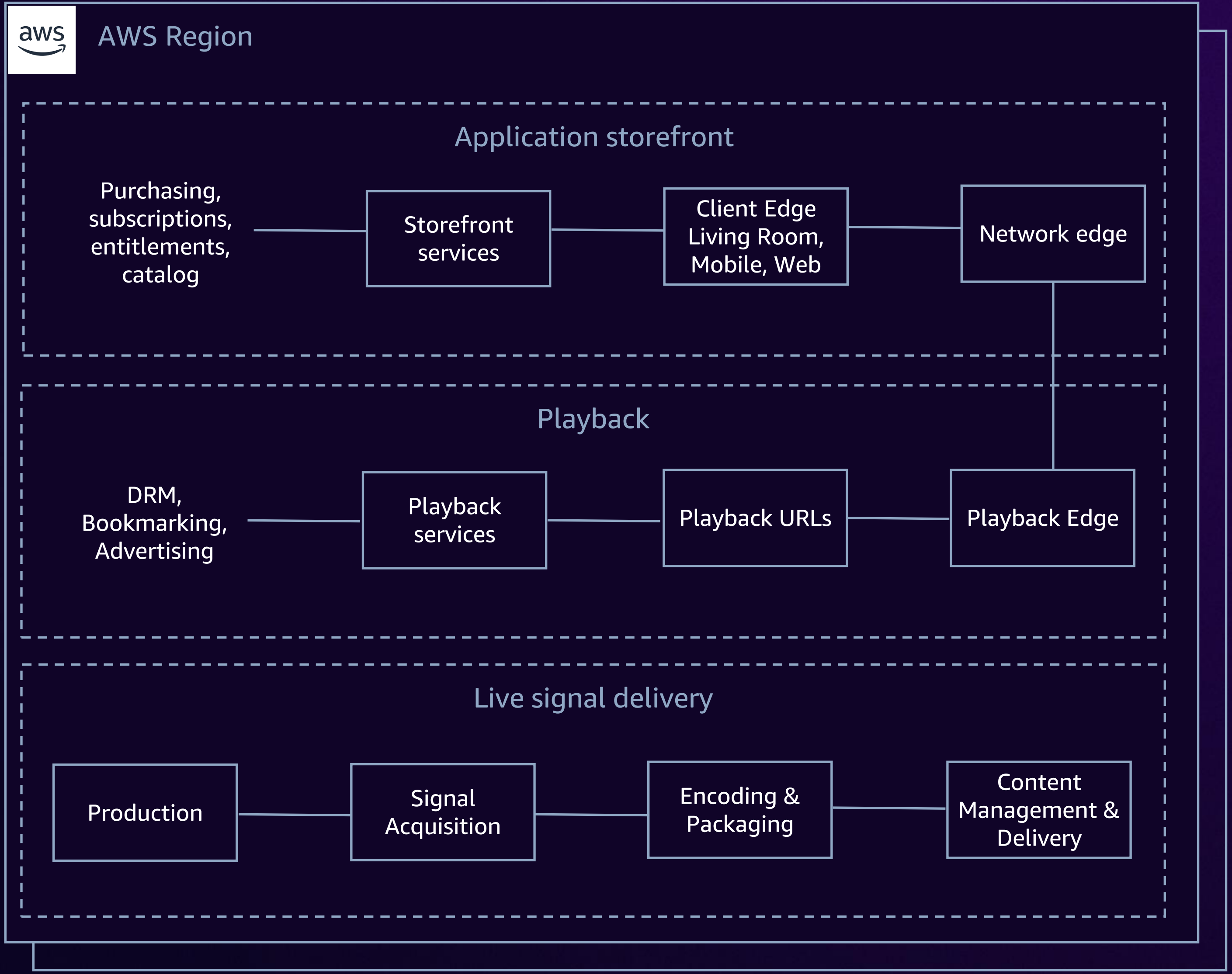
- Address the area under the curve
- Reduce number of days at peak scale
- Reduce infra $$
- Reduce engineering effort



© 2024, Amazon Web Services, Inc. or its affiliates. All rights reserved.

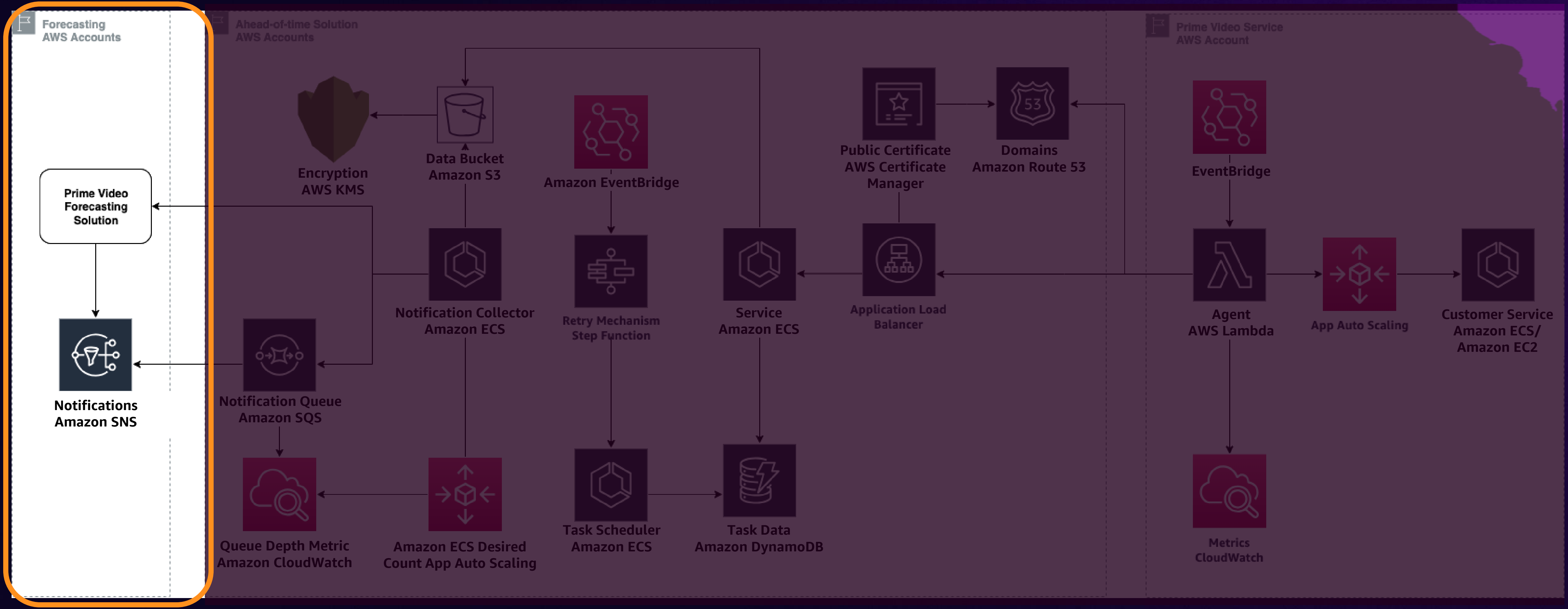
Signaling solution to enable auto scaling

- Proactive fully-automated signaling solution
- Building a solution that is AWS product-agnostic, and can be leveraged across hundreds of teams



© 2024, Amazon Web Services, Inc. or its affiliates. All rights reserved.

Prime Video automated signaling solution

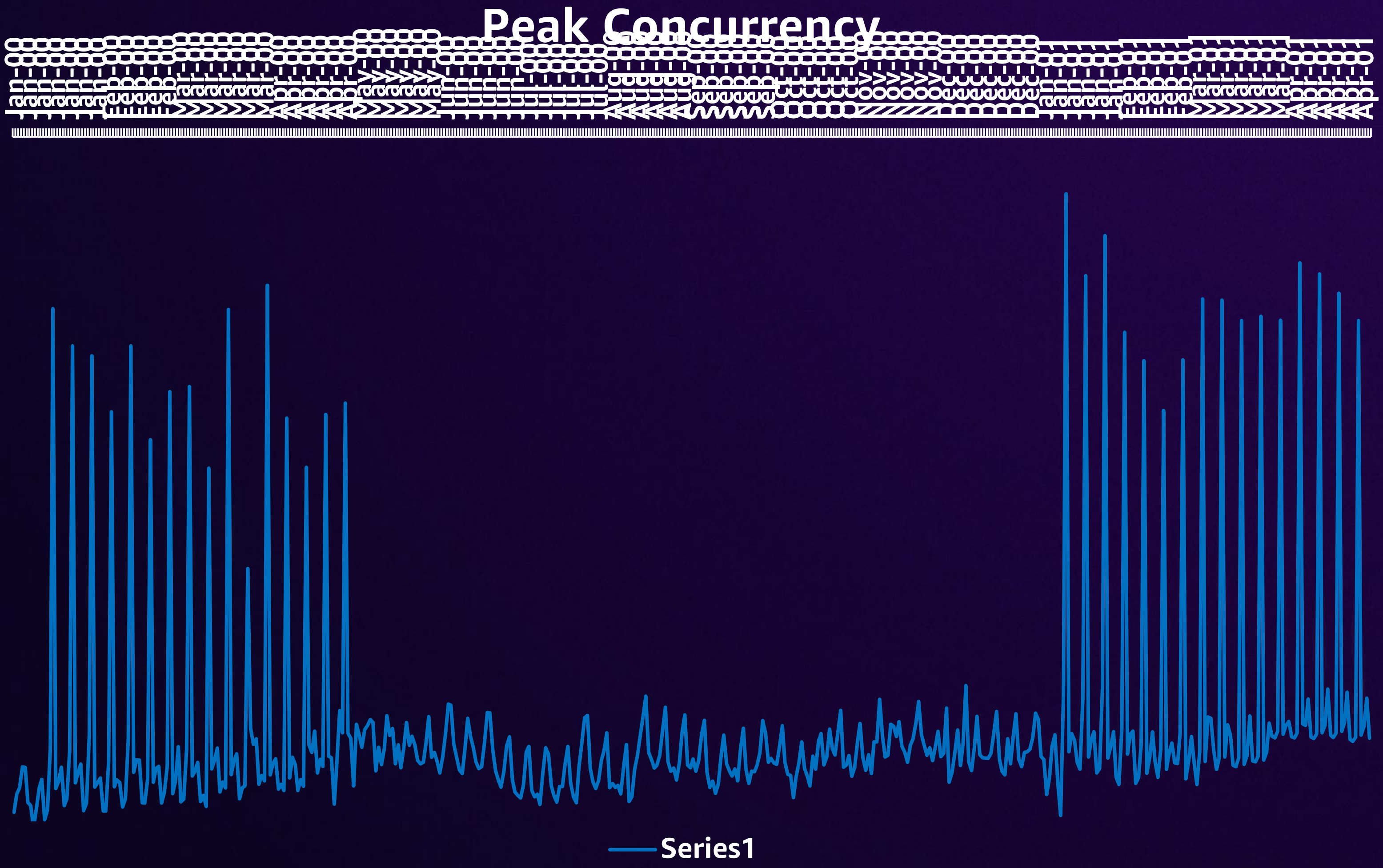


Forecasting AWS Accounts
Prime Video Forecasting Solution
Notifications Amazon SNS
Ahead-of-time Solution AWS Accounts
Encryption AWS KMS
Data Bucket Amazon S3
Amazon EventBridge
Public Certificate AWS Certificate Manager
Domains Amazon Route 53
Notification Collector Amazon ECS
Retry Mechanism Step Function
Service Amazon ECS
Application Load Balancer
Notification Queue Amazon SQS
Queue Depth Metric Amazon CloudWatch
Amazon ECS Desired Count App Auto Scaling
Task Scheduler Amazon ECS
Task Data Amazon DynamoDB
Prime Video Service AWS Account
EventBridge
Agent AWS Lambda
App Auto Scaling
Customer Service Amazon ECS/ Amazon EC2
Metrics CloudWatch

© 2024, Amazon Web Services, Inc. or its affiliates. All rights reserved.

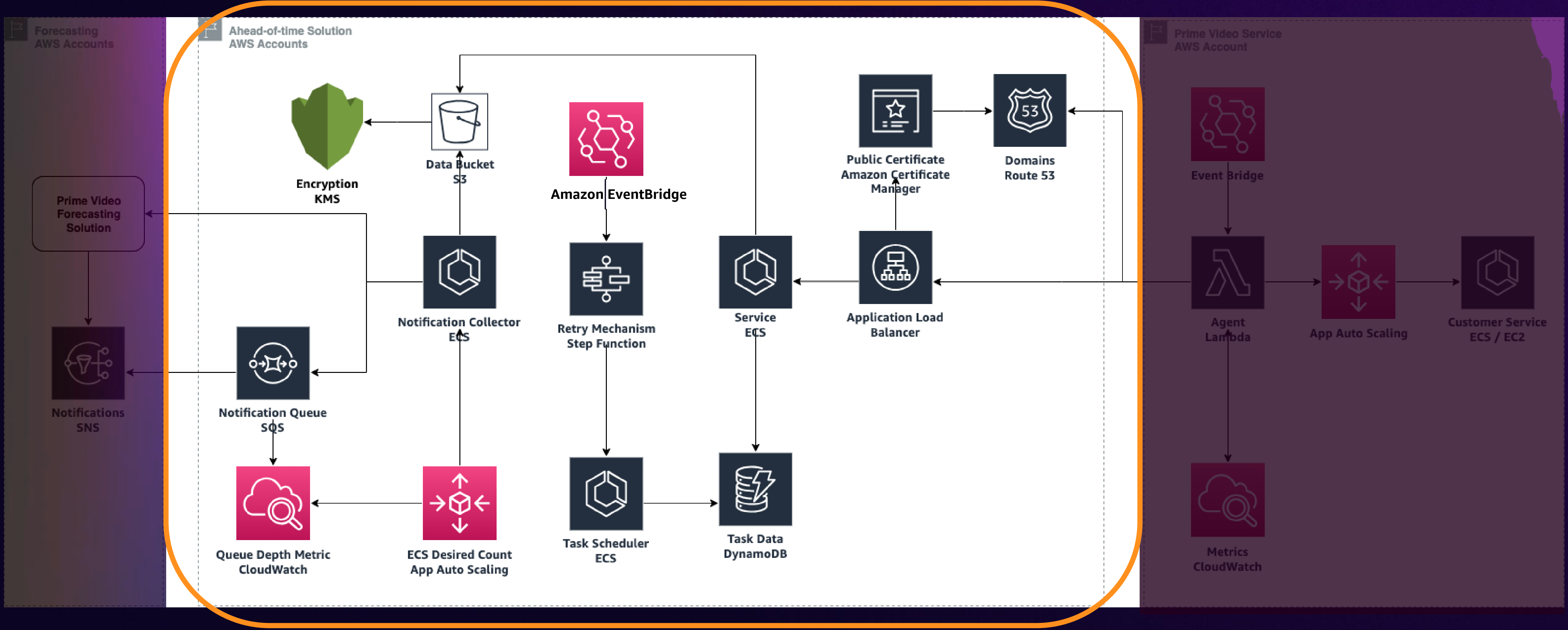
Proactive demand forecasting

- Utilizing machine learning and business inputs to predict NFL demand
- Importance of accuracy in capacity planning



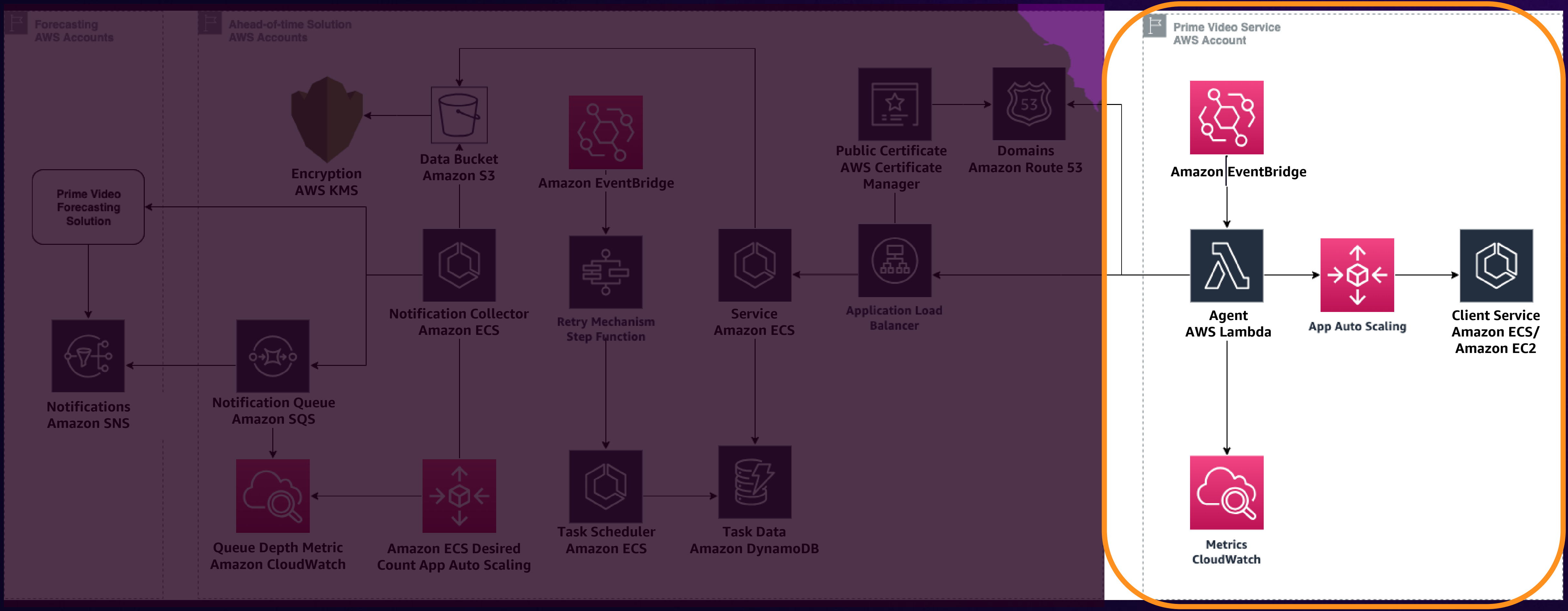
© 2024, Amazon Web Services, Inc. or its affiliates. All rights reserved.

Prime Video automated signaling solution


Forecasting AWS Accounts
Prime Video Forecasting Solution
Notifications SNS
Ahead-of-time Solution AWS Accounts
Encryption KMS
Data Bucket S3
Amazon EventBridge
Public Certificate Amazon Certificate Manager
Domains Route 53
Notification Collector ECS
Retry Mechanism Step Function
Service ECS
Application Load Balancer
Notification Queue SQS
Queue Depth Metric CloudWatch
ECS Desired Count App Auto Scaling
Task Scheduler ECS
Task Data DynamoDB
Prime Video Service AWS Account
Event Bridge
Agent Lambda
App Auto Scaling
Customer Service ECS / EC2
Metrics CloudWatch

© 2024, Amazon Web Services, Inc. or its affiliates. All rights reserved.

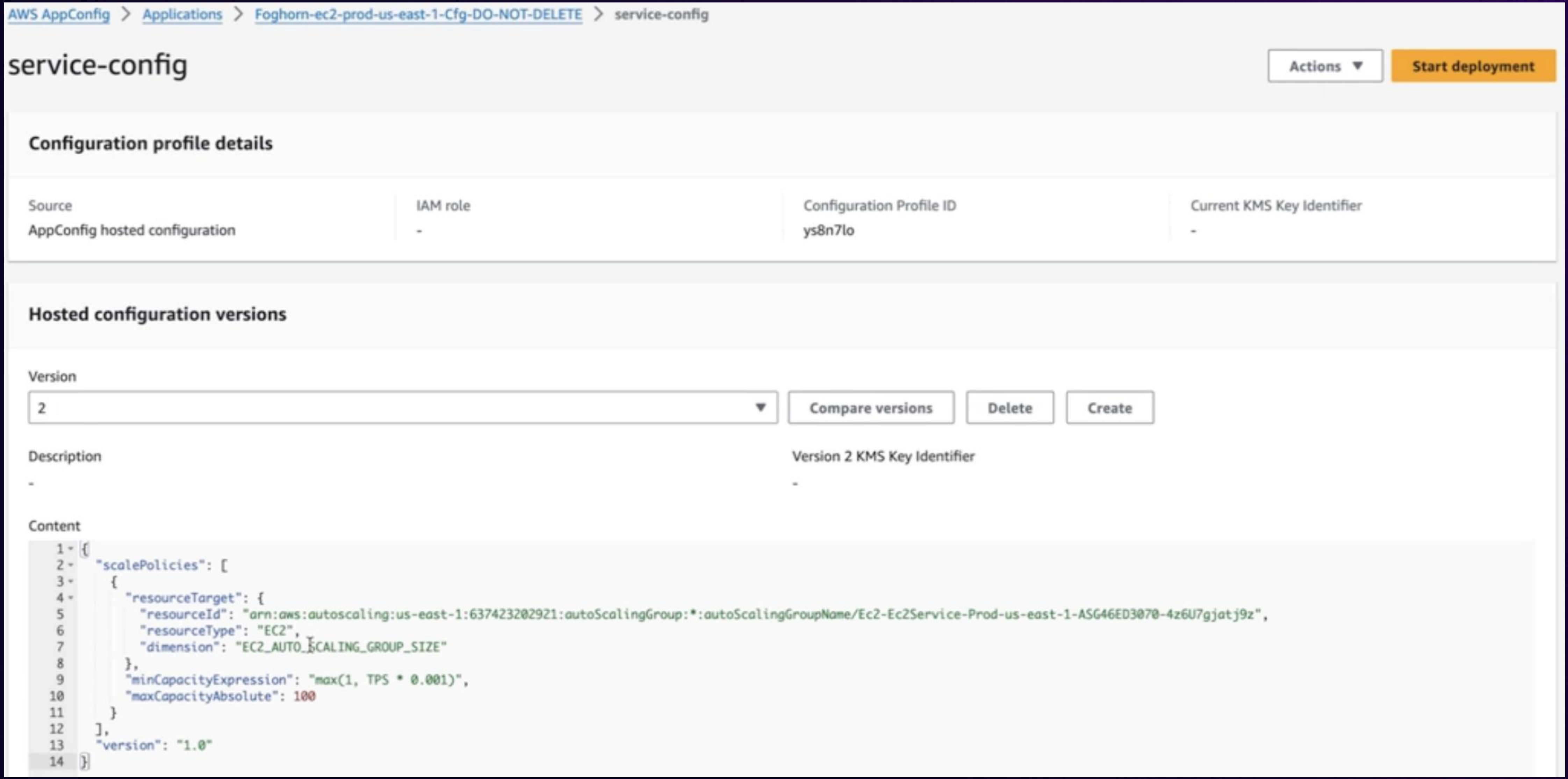
Prime Video automated signaling solution



Forecasting AWS Accounts
Ahead-of-time Solution AWS Accounts
Encryption AWS KMS
Data Bucket Amazon S3
Amazon EventBridge
Public Certificate AWS Certificate Manager
Domains Amazon Route 53
Prime Video Forecasting Solution
Notification Collector Amazon ECS
Retry Mechanism Step Function
Service Amazon ECS
Application Load Balancer
Notifications Amazon SNS
Notification Queue Amazon SQS
Queue Depth Metric Amazon CloudWatch
Amazon ECS Desired Count App Auto Scaling
Task Scheduler Amazon ECS
Task Data Amazon DynamoDB
Prime Video Service AWS Account
Amazon EventBridge
Agent AWS Lambda
App Auto Scaling
Client Service Amazon ECS/ Amazon EC2
Metrics CloudWatch

© 2024, Amazon Web Services, Inc. or its affiliates. All rights reserved.

Scheduled (proactive) Auto Scaling



AWS AppConfig > Applications > Foghorn-ec2-prod-us-east-1-Cfg-DO-NOT-DELETE > service-config

service-config

Actions ▼
Start deployment

Configuration profile details

Source	IAM role	Configuration Profile ID	Current KMS Key Identifier
AppConfig hosted configuration	-	ys8n7lo	-

Hosted configuration versions

Version: 2 ▼

Compare versions
Delete
Create

Description: -

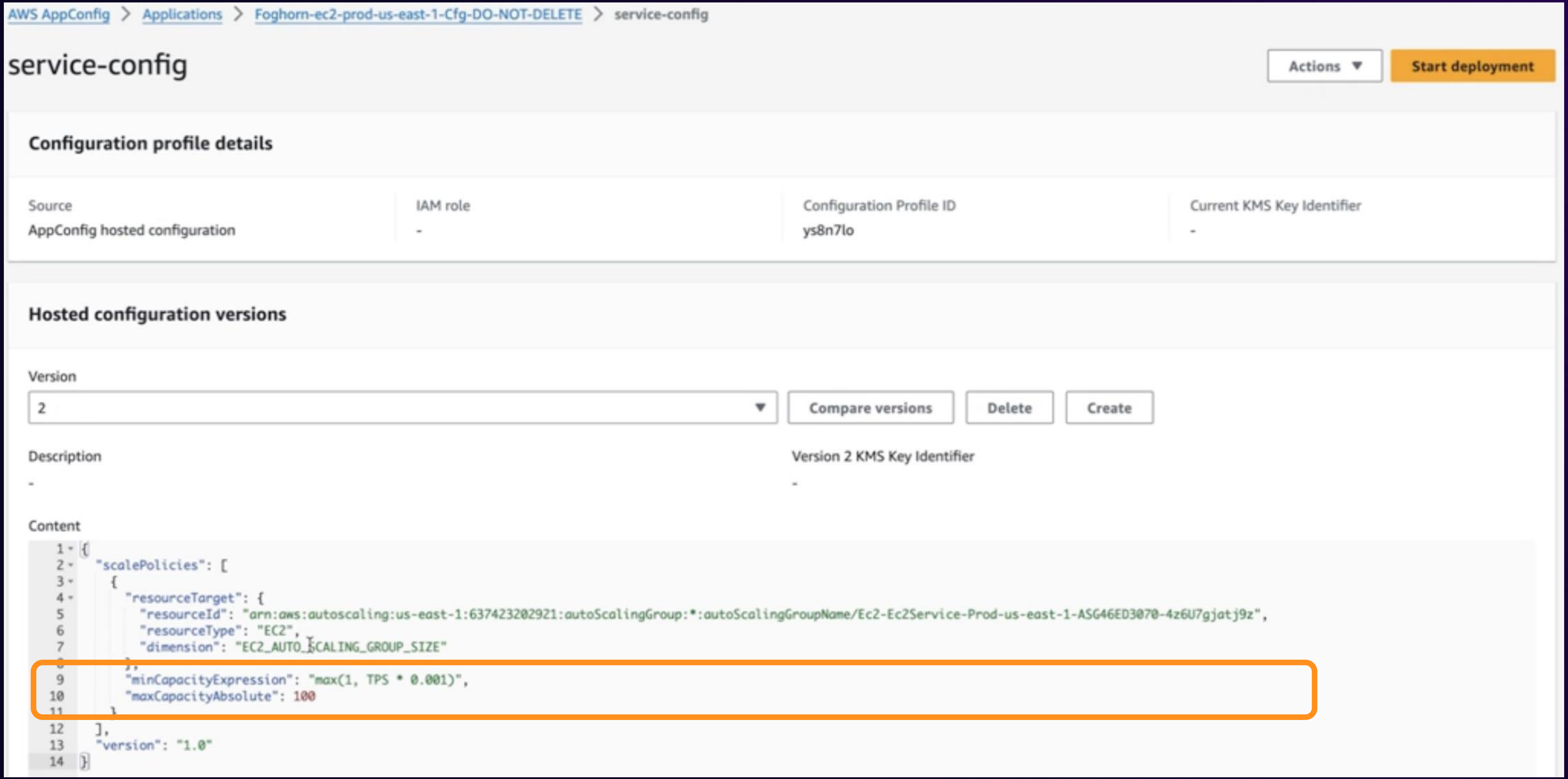
Version 2 KMS Key Identifier: -

Content

```
{
  "scalePolicies": [
    {
      "resourceTarget": {
        "resourceId": "arn:aws:autoscaling:us-east-1:637423202921:autoScalingGroup:*:autoScalingGroupName/Ec2-Ec2Service-Prod-us-east-1-ASG46ED3070-4z6U7gjatj9z",
        "resourceType": "EC2",
        "dimension": "EC2_AUTO_SCALING_GROUP_SIZE"
      },
      "minCapacityExpression": "max(1, TPS * 0.001)",
      "maxCapacityAbsolute": 100
    }
  ],
  "version": "1.0"
}
```

© 2024, Amazon Web Services, Inc. or its affiliates. All rights reserved.

Scheduled (proactive) Auto Scaling


AWS AppConfig > Applications > Foghorn-ec2-prod-us-east-1-Cfg-DO-NOT-DELETE > service-config
Actions ▼
Start deployment
Compare versions
Delete
Create

service-config

Configuration profile details

Source
AppConfig hosted configuration

IAM role
-

Configuration Profile ID
ys8n7lo

Current KMS Key Identifier
-

Hosted configuration versions

Version
2 ▼

Description
-

Version 2 KMS Key Identifier
-

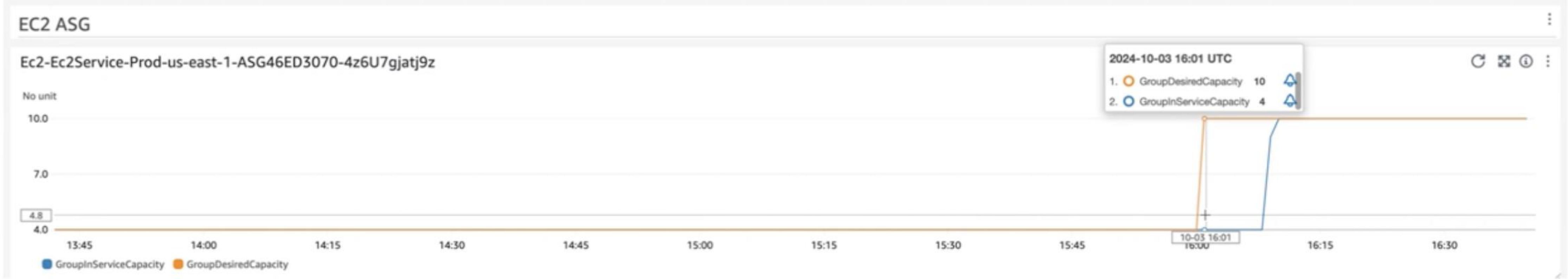
Content

```
{
  "scalePolicies": [
    {
      "resourceTarget": {
        "resourceId": "arn:aws:autoscaling:us-east-1:637423202921:autoScalingGroup:*:autoScalingGroupName/Ec2-Ec2Service-Prod-us-east-1-ASG46ED3070-4z6U7gjatj9z",
        "resourceType": "EC2",
        "dimension": "EC2_AUTO_SCALING_GROUP_SIZE"
      },
      "minCapacityExpression": "max(1, TPS * 0.001)",
      "maxCapacityAbsolute": 100
    }
  ],
  "version": "1.0"
}
```

© 2024, Amazon Web Services, Inc. or its affiliates. All rights reserved.

Scheduled (proactive) Auto Scaling

CloudWatch Metric Dashboard



EC2 ASG

Ec2-Ec2Service-Prod-us-east-1-ASG46ED3070-4z6U7gjatj9z

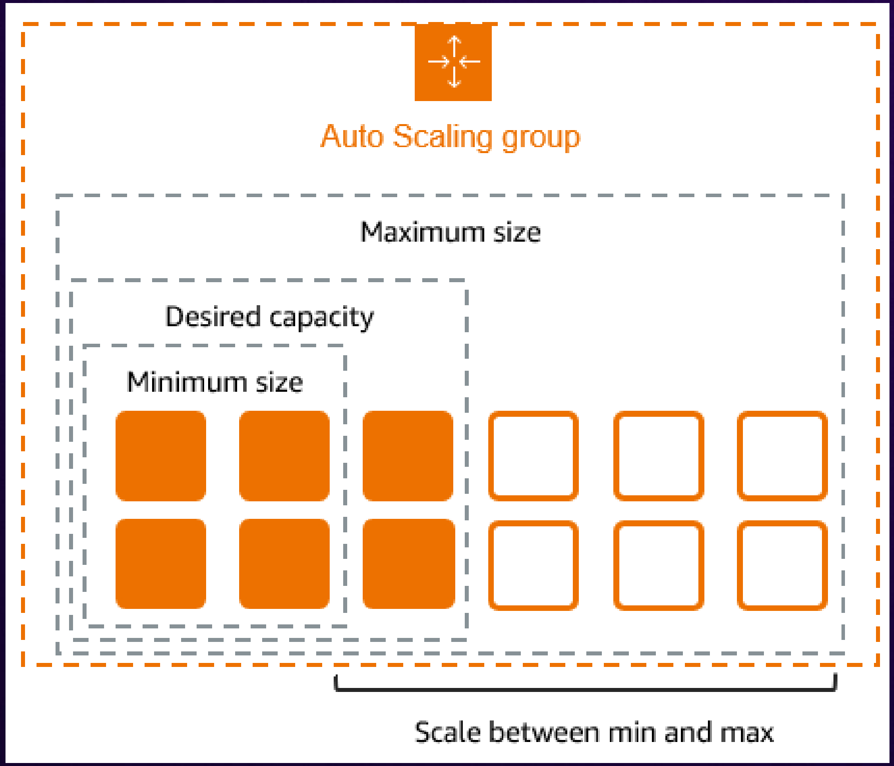
2024-10-03 16:01 UTC

1. ○ GroupDesiredCapacity 10
2. ○ GroupInServiceCapacity 4

© 2024, Amazon Web Services, Inc. or its affiliates. All rights reserved.

Dynamic (reactive) auto scaling

- Real-time scaling to handle unanticipated spikes in traffic:
  - Popular Cowboys matchup exceeding forecasts
  - Taylor Swift surprise appearance!
- Benefits of immediate responsiveness
  - Monitors key metrics and scales out resources within minutes

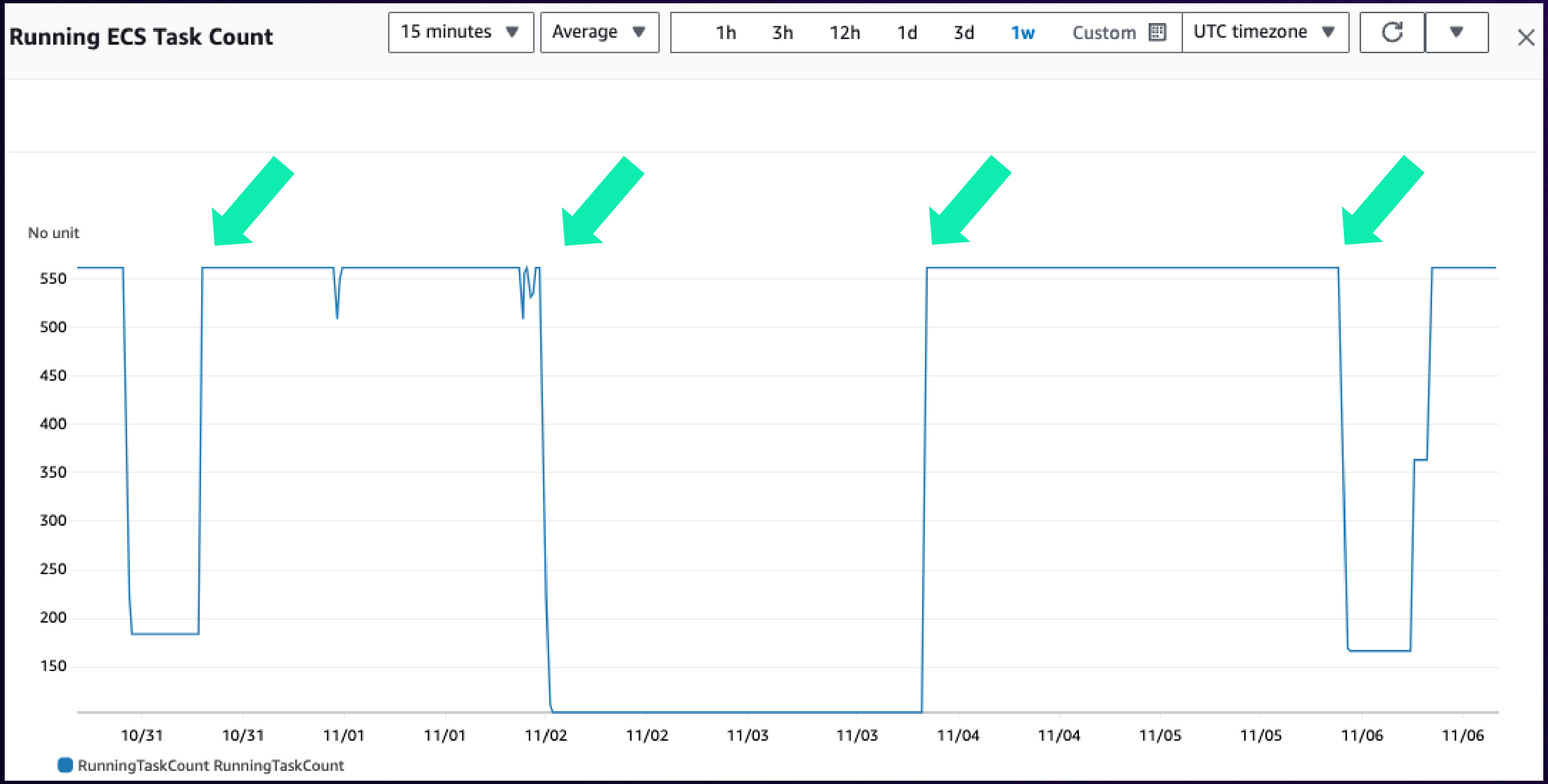


© 2024, Amazon Web Services, Inc. or its affiliates. All rights reserved.

Success Metrics and Impact - Pilot

Success metrics for evaluating auto scaling performance for pilot services

- Reduction in days at peak scale of ~70%
- Improvements in CPU utilization by ~20-30%
- Cost reduction of ~30%



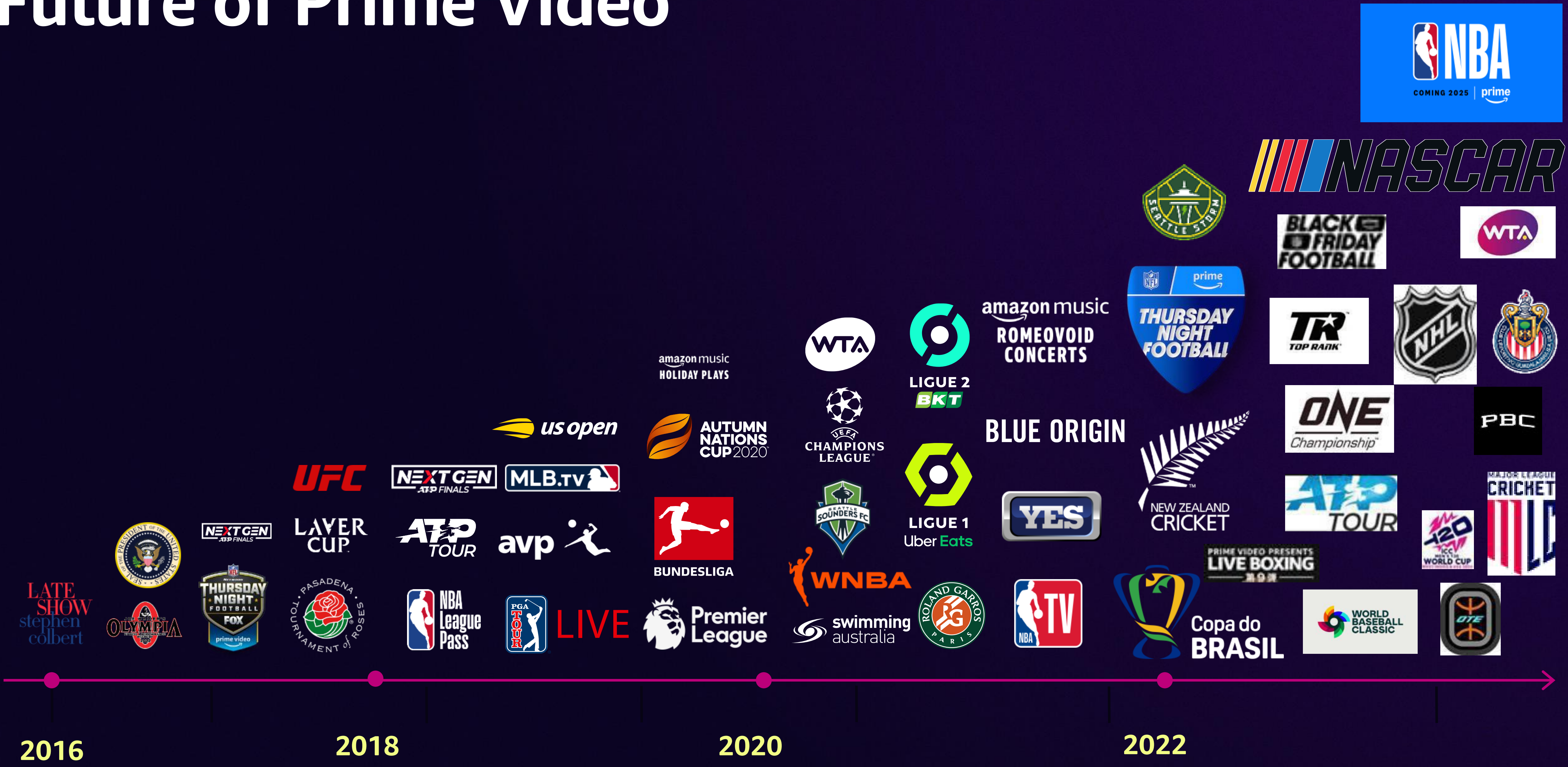
© 2024, Amazon Web Services, Inc. or its affiliates. All rights reserved.

Lessons learned and best practices

- Understand your workloads (scheduled vs. dynamic)
- Complement with dynamic auto scaling
- Start simple, gradually increase
- Monitoring, Alarms, and optimize KPIs

© 2024, Amazon Web Services, Inc. or its affiliates. All rights reserved.

Future of Prime Video



© 2024, Amazon Web Services, Inc. or its affiliates. All rights reserved.

Recap

© 2024, Amazon Web Services, Inc. or its affiliates. All rights reserved.

Recap

1

Delivery – Premium video at scale to millions of customers in across thousands of different devices

2

Customer experience – Millions of requests and terabits of data over highly variable networks

3

Availability and resiliency – Achieve 100% up-time with multi-region and autoscaling strategies

© 2024, Amazon Web Services, Inc. or its affiliates. All rights reserved.

Number of AWS services?

© 2024, Amazon Web Services, Inc. or its affiliates. All rights reserved.

120+

© 2024, Amazon Web Services, Inc. or its affiliates. All rights reserved.

AWS Well-Architected Framework

- Operational excellence
- Security
- Reliability
- Cost optimization
- Sustainability
- Performance efficiency

© 2024, Amazon Web Services, Inc. or its affiliates. All rights reserved.

Reliability

Reliability (Goal)

The ability of a system to perform its required function correctly and consistently

Resilience (Approach)

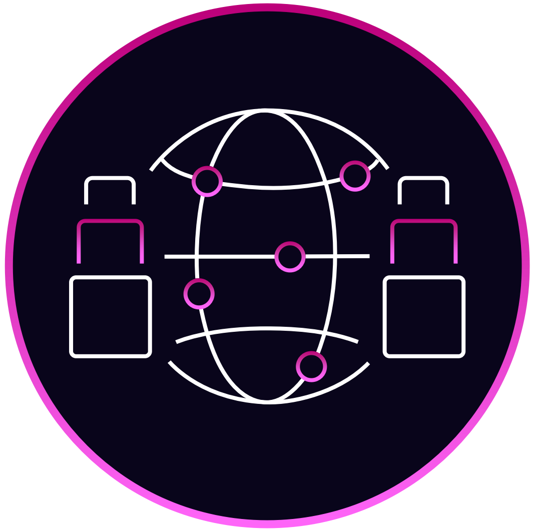
The ability of a system to recover quickly and restore to a fully functional state after a disruption or failure

https://bit.ly/reliability-pillar

© 2024, Amazon Web Services, Inc. or its affiliates. All rights reserved.

Reliability design principles

- Automatically recover from failure
- Test recovery procedures
- Scale horizontally to increase aggregate workload availability
- Stop guessing capacity
- Manage change through automation



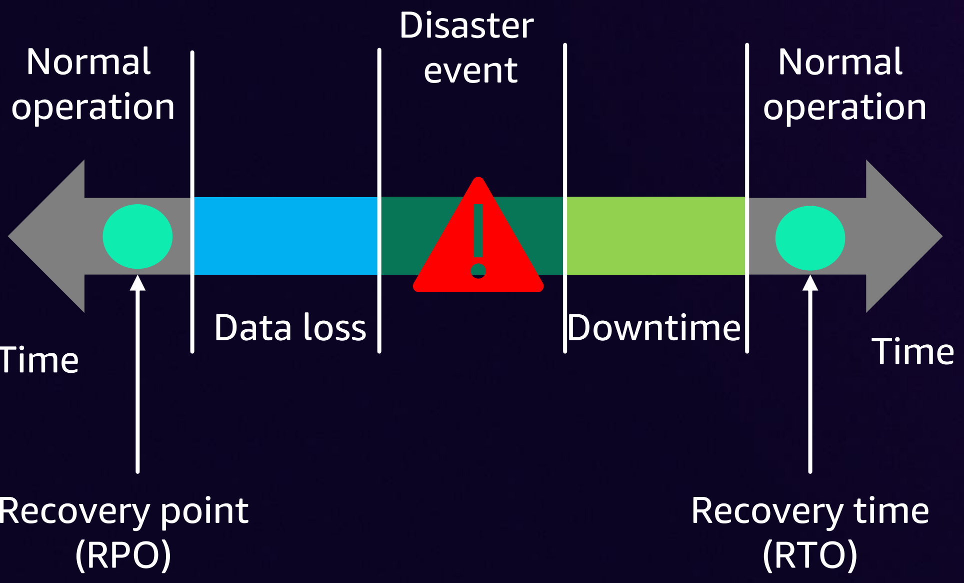
© 2024, Amazon Web Services, Inc. or its affiliates. All rights reserved.

Multi-Region best practices

- Understand consistency and replication options
- Utilize asynchronous writes across Regions
- Multi-Region writes should be idempotent
- Observability from multiple viewpoints

© 2024, Amazon Web Services, Inc. or its affiliates. All rights reserved.

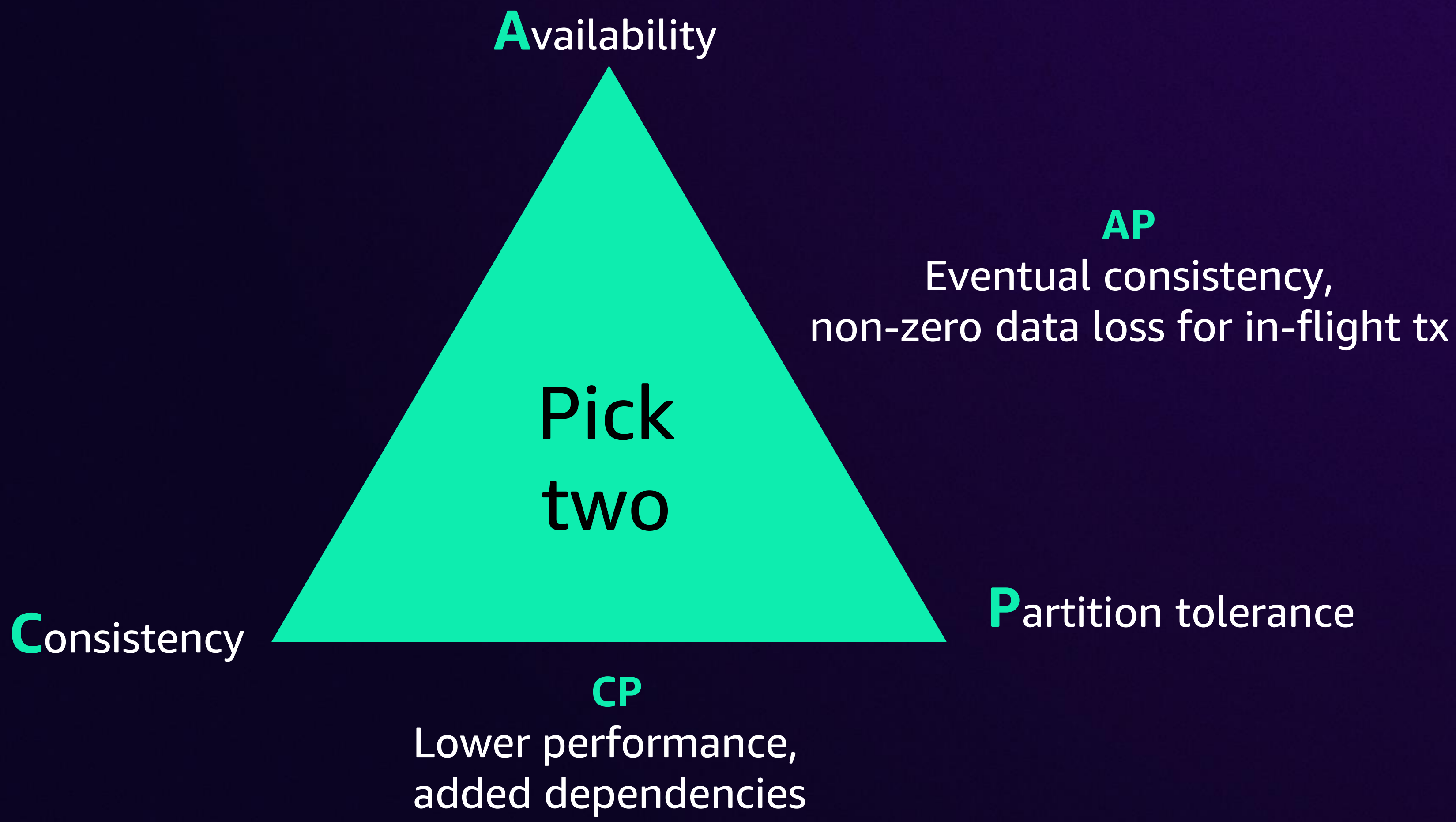
Fundamental #1: Understand the requirements



- Why multi-Region?
- Cost to run your application multi-Region vs. business impact
- What's the RTO/RPO requirements?
- Development cycles to improve application resilience?

© 2024, Amazon Web Services, Inc. or its affiliates. All rights reserved.

Fundamental #2: Understand data consistency



Availability

Pick two

Consistency

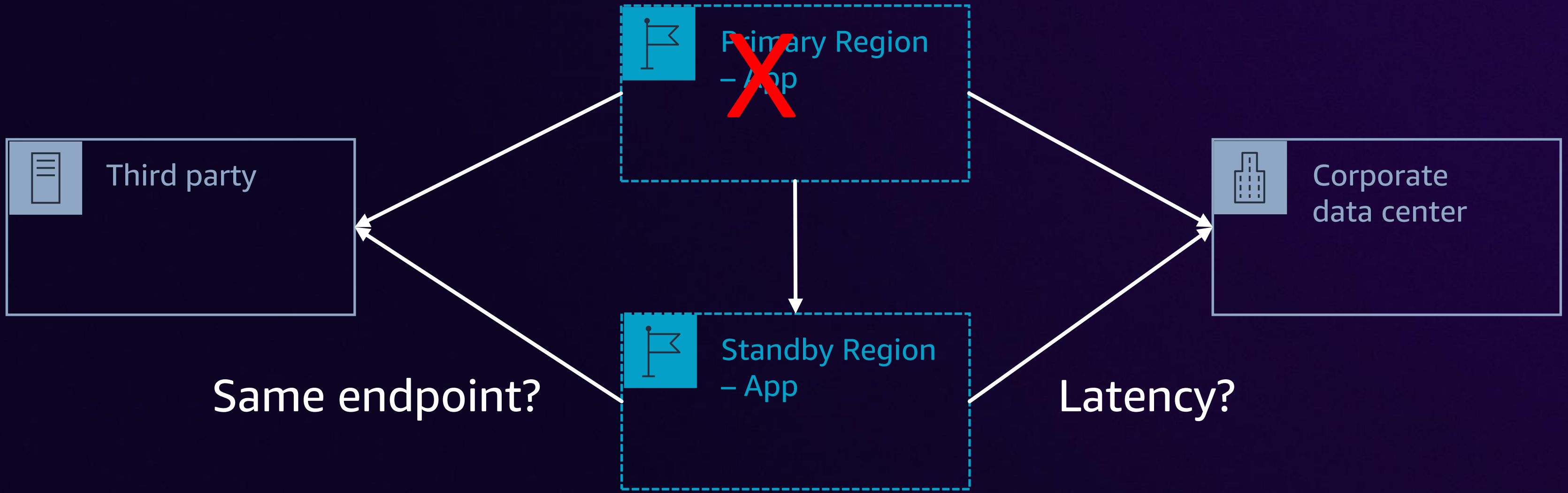
Partition tolerance

AP
Eventual consistency,
non-zero data loss for in-flight tx

CP
Lower performance,
added dependencies

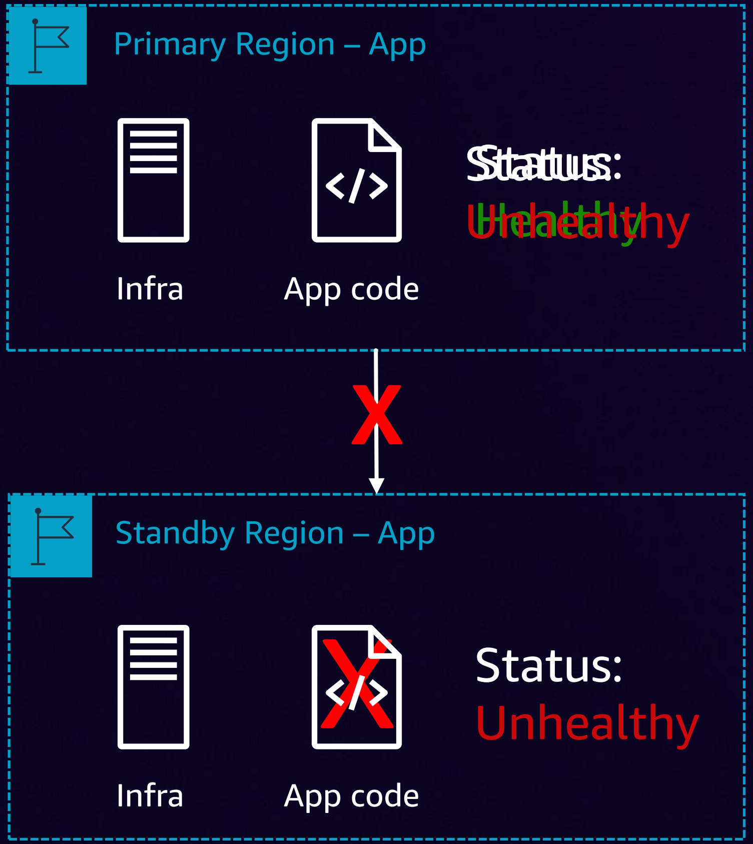
© 2024, Amazon Web Services, Inc. or its affiliates. All rights reserved.

Fundamental #3: Understanding dependencies



© 2024, Amazon Web Services, Inc. or its affiliates. All rights reserved.

Fundamental #4: Operational readiness



- AWS account limit awareness
- Use cross-Region health checks
- Make informed decisions
- Automate & test operation readiness using playbooks

© 2024, Amazon Web Services, Inc. or its affiliates. All rights reserved.

Key takeaways

© 2024, Amazon Web Services, Inc. or its affiliates. All rights reserved.

Key takeaways

Operational resilience

Support growth for millions of users on TNF days

Dynamic scaling

Peak-to-mean ratio is **highly variable** (baseline days to NFL days)

Optimized cost

~20% reduction in cost and operations

Carbon savings

43% reduction in carbon footprint by reducing number of EC2 instances

Resiliency

Optimization

© 2024, Amazon Web Services, Inc. or its affiliates. All rights reserved.

> Everything fails all the time. We needed to build systems that embrace failure as a natural occurrence.

Dr. Werner Vogels

CTO, Amazon

© 2024, Amazon Web Services, Inc. or its affiliates. All rights reserved.

Thank you!

Please complete the session survey in the mobile app

Tulip Gupta

✉ tulipgpt@amazon.com

in linkedin.com/in/tulip-gupta

Elliott Nash

✉ ellinash@amazon.com

in linkedin.com/in/elliottnash

Ralph Chaker

✉ chakerr@amazon.com

in linkedin.com/in/ralphchaker

© 2024, Amazon Web Services, Inc. or its affiliates. All rights reserved.