
AWS
re:Invent
NOV. 28 – DEC. 2, 2022 | LAS VEGAS, NV

BOA311

Being good neighbors: Rate limiting in a modern application world

Liel Ran

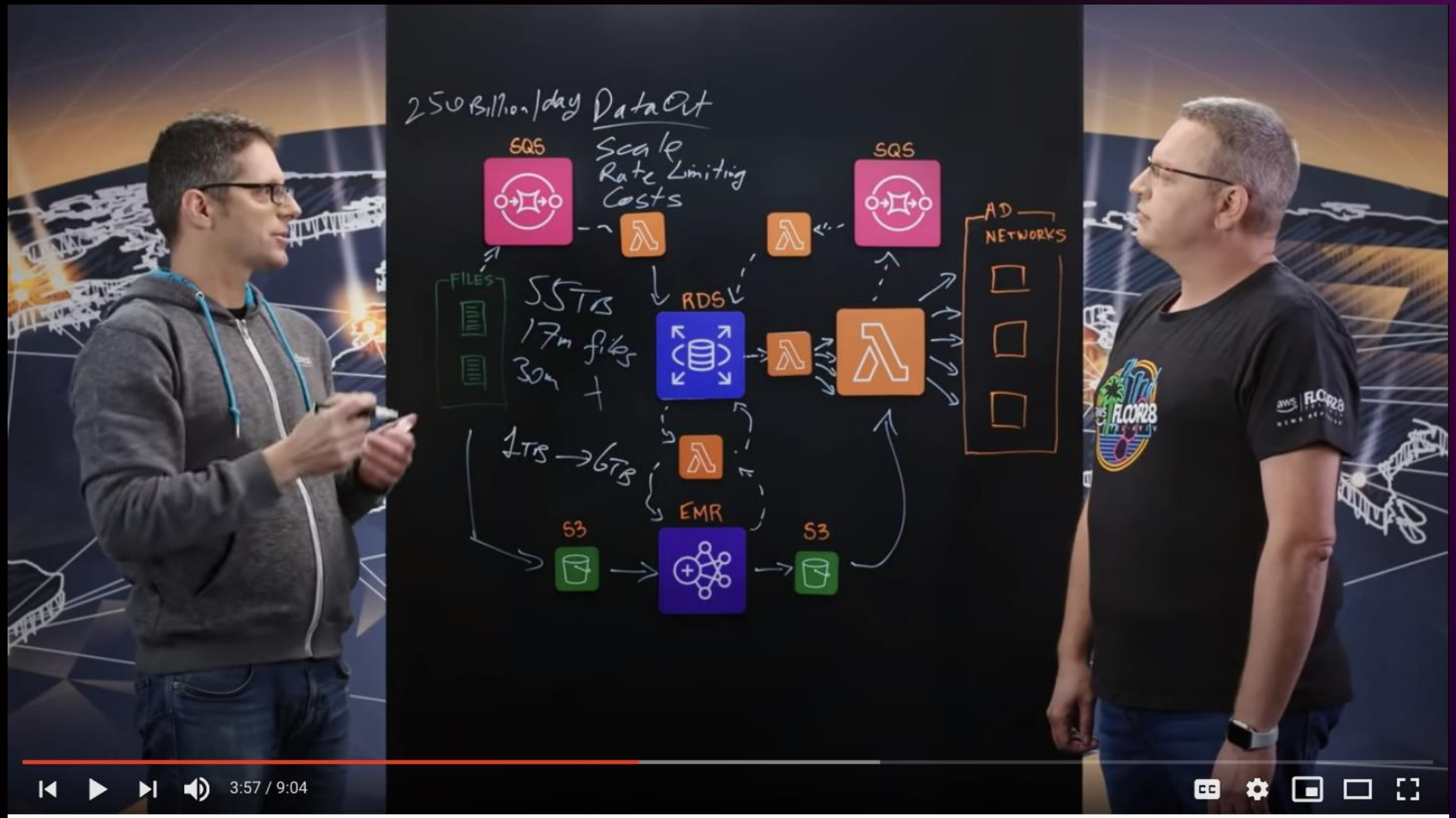
CTO & Co-Founder
DoControl

Boaz Ziniman

Principal Developer Advocate
AWS

aws

© 2022, Amazon Web Services, Inc. or its affiliates. All rights reserved.

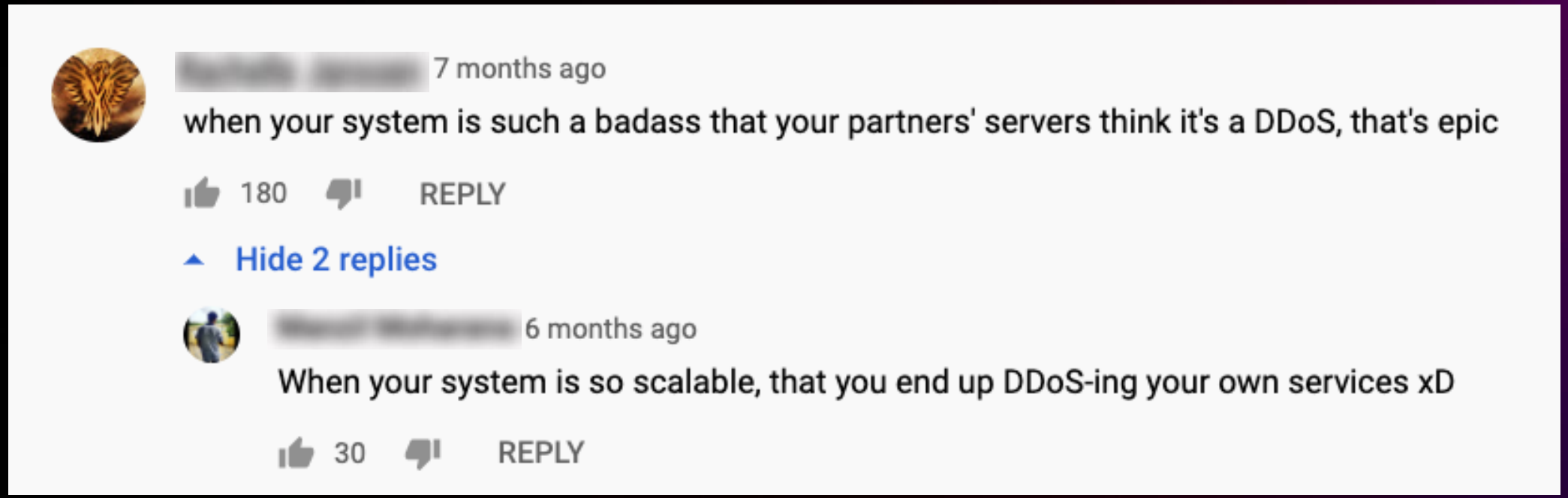


#AWS

Nielsen: Processing 55TB of Data Per Day with AWS Lambda

© 2022, Amazon Web Services, Inc. or its affiliates. All rights reserved.

Nielsen: Processing 55 TB of data per day with AWS Lambda



© 2022, Amazon Web Services, Inc. or its affiliates. All rights reserved.

What do our customers want in a modern app?



Scales to millions of users



Has global availability



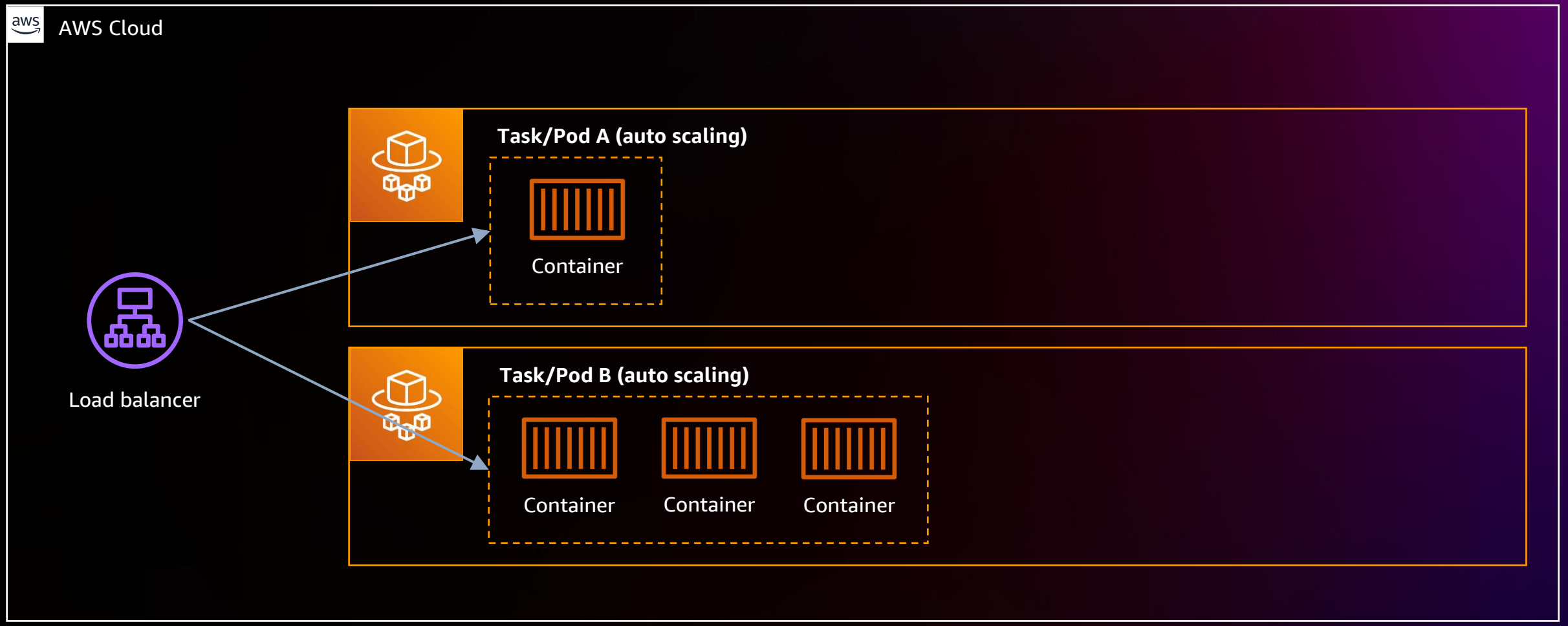
Responds in milliseconds



Handles petabytes of data

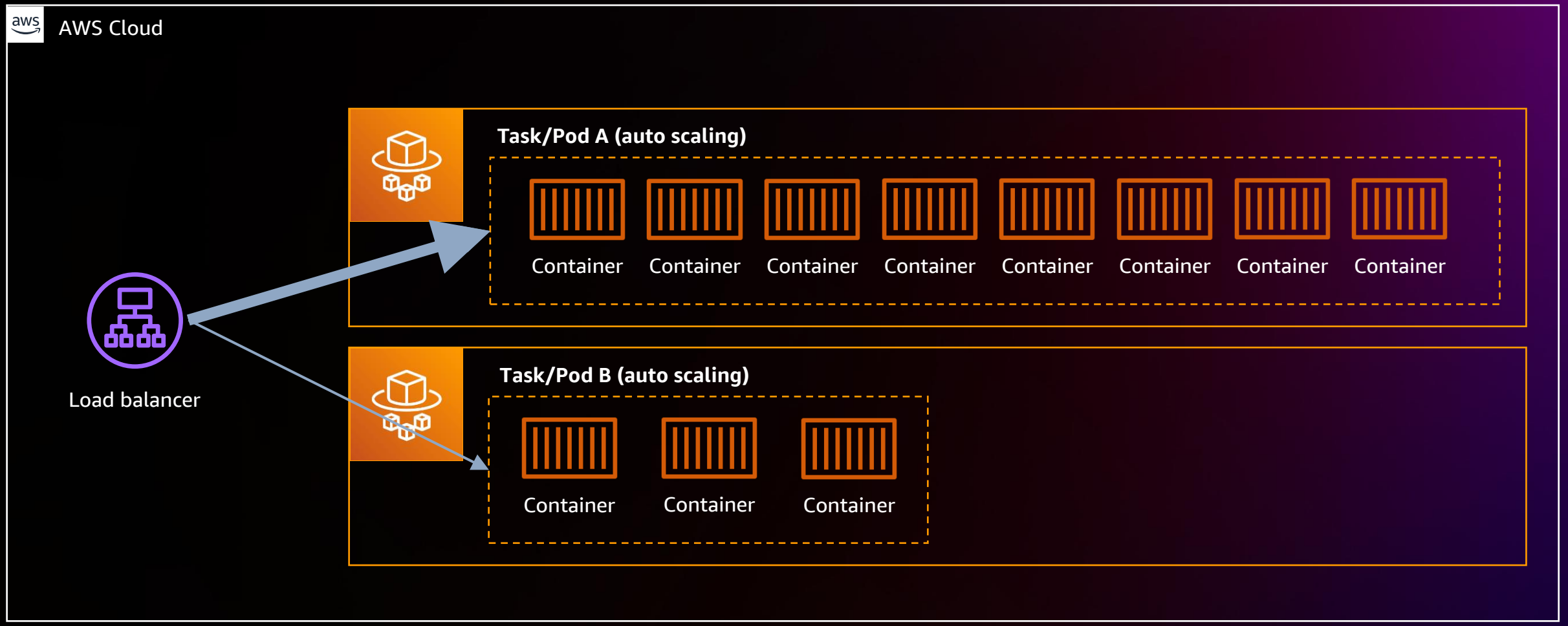
© 2022, Amazon Web Services, Inc. or its affiliates. All rights reserved.

Automatic scaling



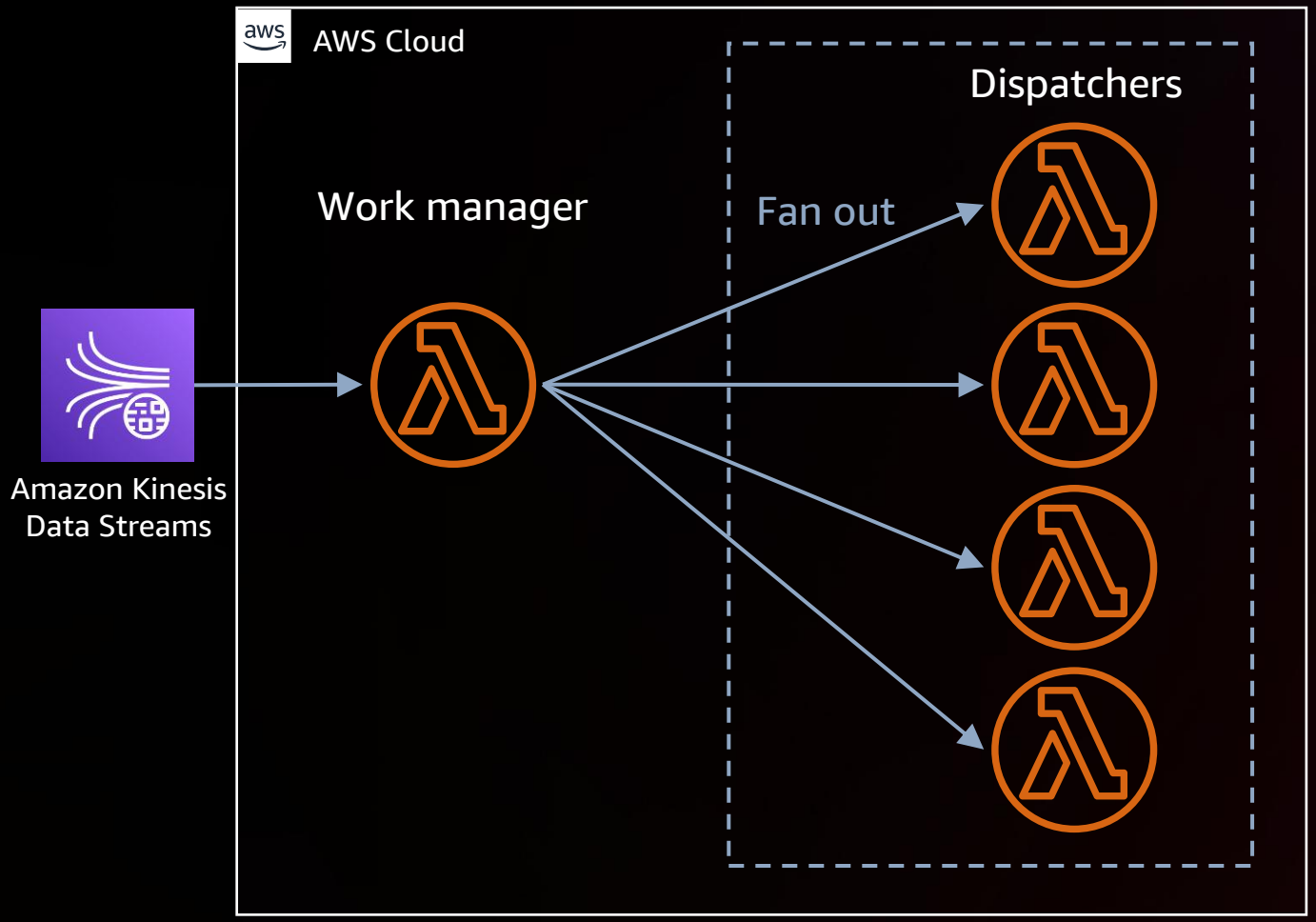
© 2022, Amazon Web Services, Inc. or its affiliates. All rights reserved.

Automatic scaling



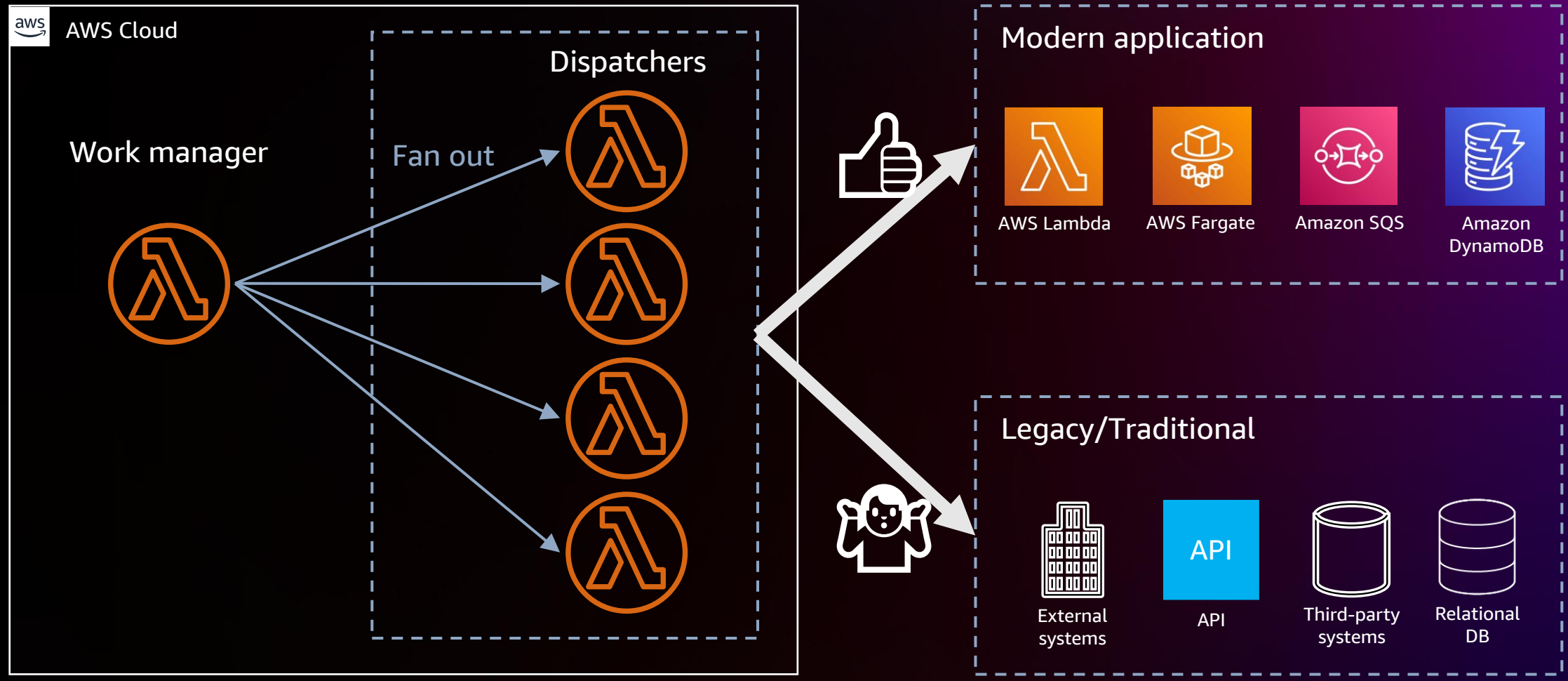
© 2022, Amazon Web Services, Inc. or its affiliates. All rights reserved.

Automatic scaling



© 2022, Amazon Web Services, Inc. or its affiliates. All rights reserved.

Automatic scaling



© 2022, Amazon Web Services, Inc. or its affiliates. All rights reserved.

Non-modern scaling might be hard

- No scaling
- Limited scaling



© 2022, Amazon Web Services, Inc. or its affiliates. All rights reserved.

Serverless functions and containers can be isolated and egoistic



aws

© 2022, Amazon Web Services, Inc. or its affiliates. All rights reserved.

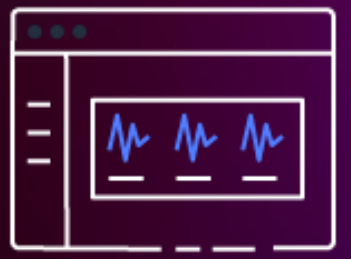
What can possibly go wrong?



API throttling/ blocking



Timeouts



Downtime



Cost



Customer experience



DDoS

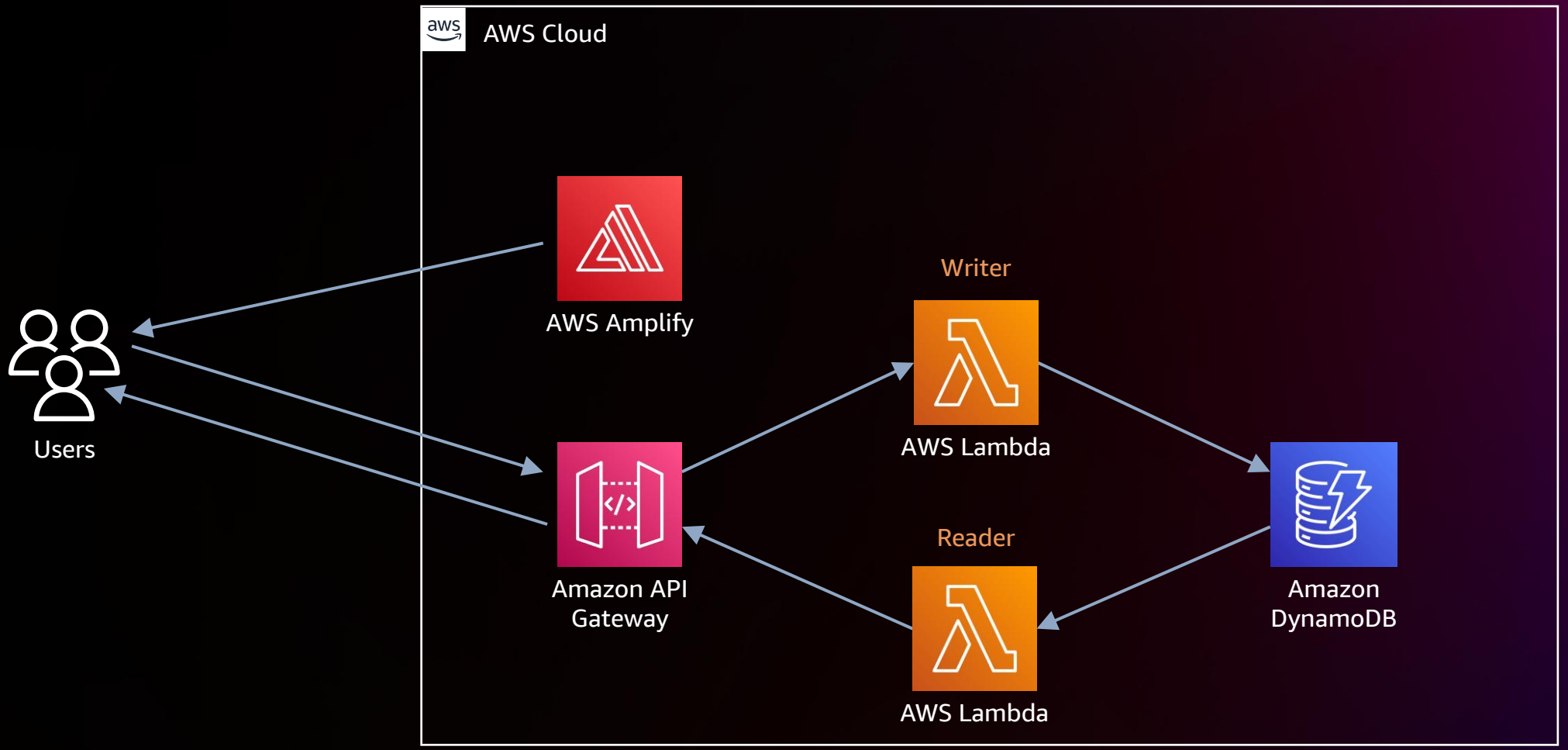
aws

© 2022, Amazon Web Services, Inc. or its affiliates. All rights reserved.

Limits

© 2022, Amazon Web Services, Inc. or its affiliates. All rights reserved.

Architecture slide – Amazon DynamoDB

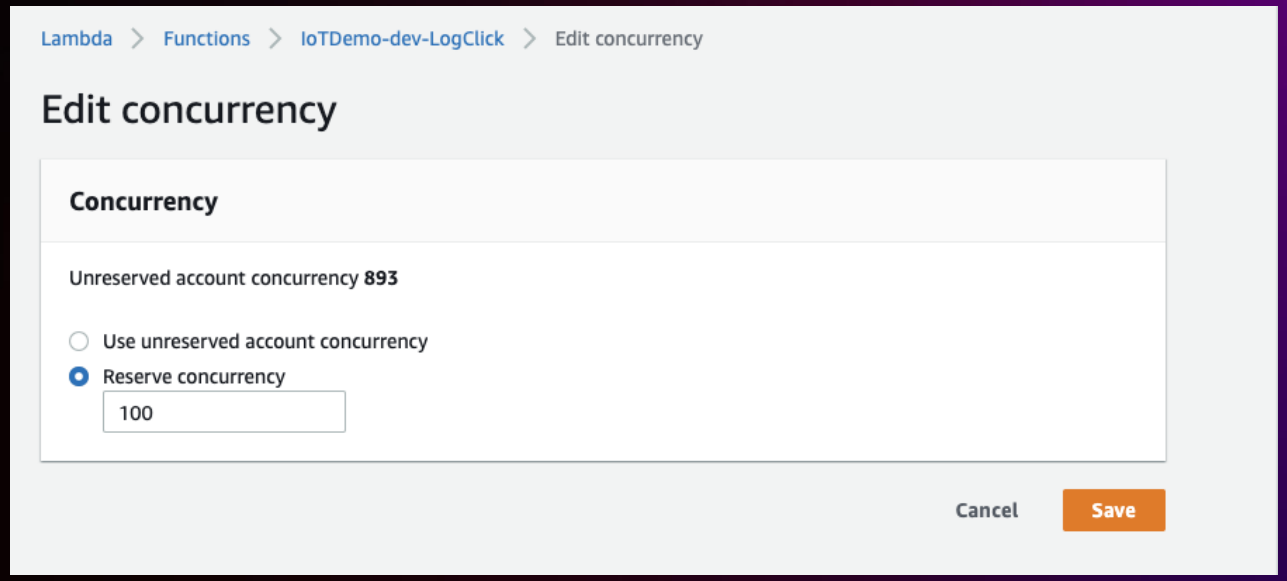


© 2022, Amazon Web Services, Inc. or its affiliates. All rights reserved.

The basics – Set function concurency

Reserved concurrency

- Sets ceiling on maximum number of execution environments – maximum concurrency for a given function
- Also reserves that concurrency from the account's quota




```
aws lambda put-function-concurrency \
--function-name rate-limiting-demo-prod-WriteFeedback \
--reserved-concurrent-executions 0
```

Calling the invoke API action failed with this message: Rate Exceeded.

© 2022, Amazon Web Services, Inc. or its affiliates. All rights reserved.

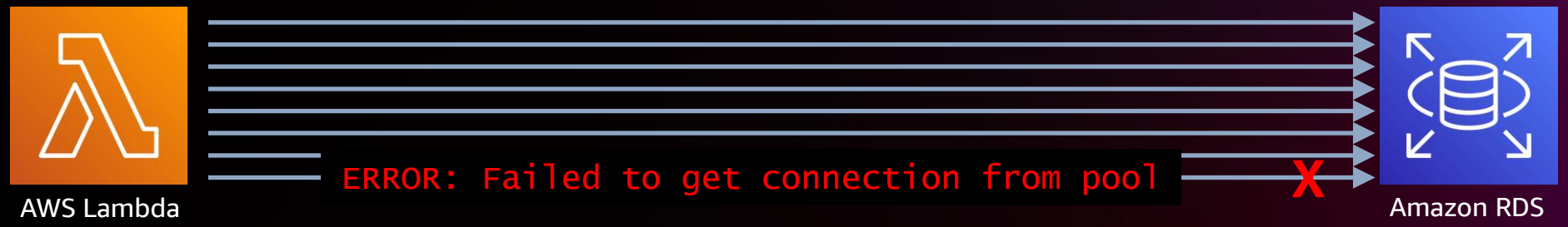
AWS Lambda: Event filtering

- Filter incoming messages before function invocation
- Reduce traffic to function, reduce cost
- Currently supports: Amazon Kinesis Data Streams, Amazon DynamoDB Streams, Amazon SQS, Amazon EventBridge, AWS SMS
  NEW - MSK, Self-Managed Kafka, MQ for Apache ActiveMQ, and MQ for RabbitMQ

```
FleetTirePressureMapping:
  Type: AWS::Lambda::EventSourceMapping
  Properties:
    FunctionName: fleet-tire-pressure-evaluator
    BatchSize: 100
    StartingPosition: LATEST
    EventSourceArn:   arn:aws:kinesis:...:stream/fleet-telemetry
    FilterCriteria:
      Filters:
        - Pattern: |
          {
            "data": {
              "tire_pressure”:
                [{"numeric": ["<", 32]}]
            }
          }
```

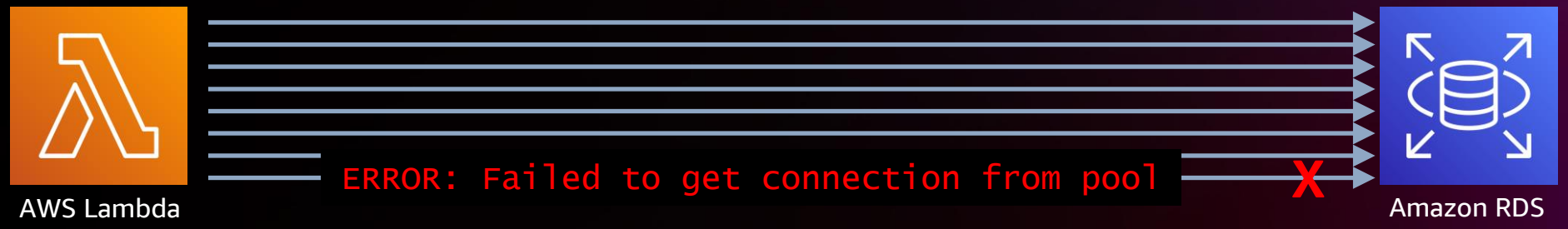
© 2022, Amazon Web Services, Inc. or its affiliates. All rights reserved.

Handling Amazon RDS connections



© 2022, Amazon Web Services, Inc. or its affiliates. All rights reserved.

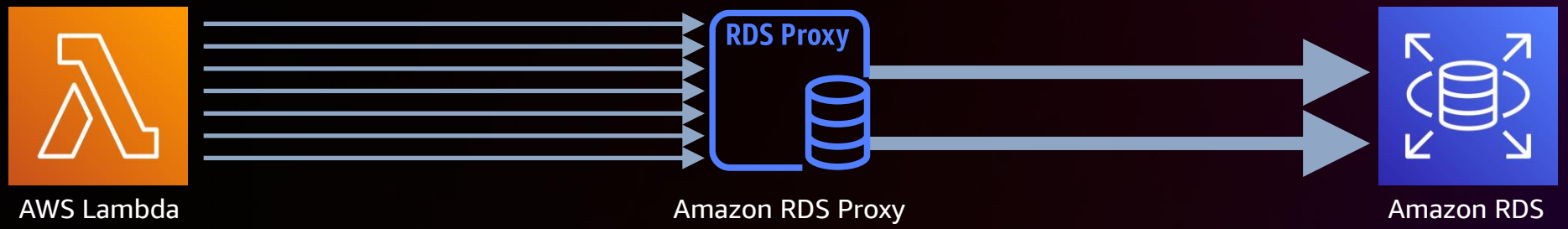
Handling Amazon RDS connections



Add buffers (queues)

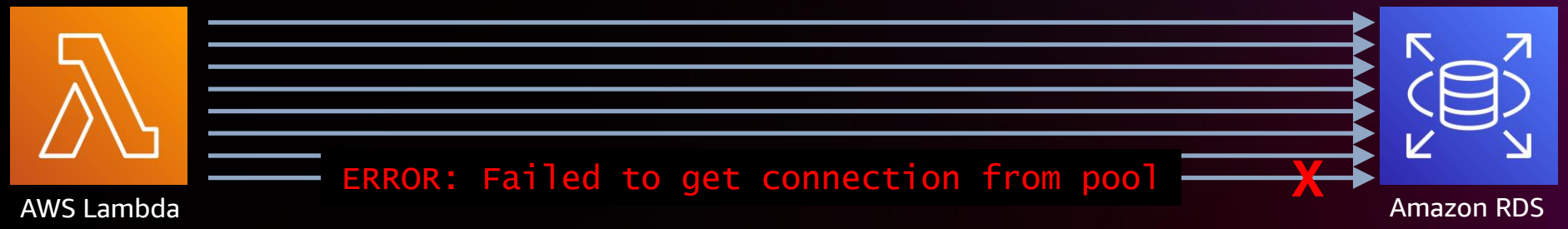


Use aggregators (RDS Proxy)



aws
© 2022, Amazon Web Services, Inc. or its affiliates. All rights reserved.

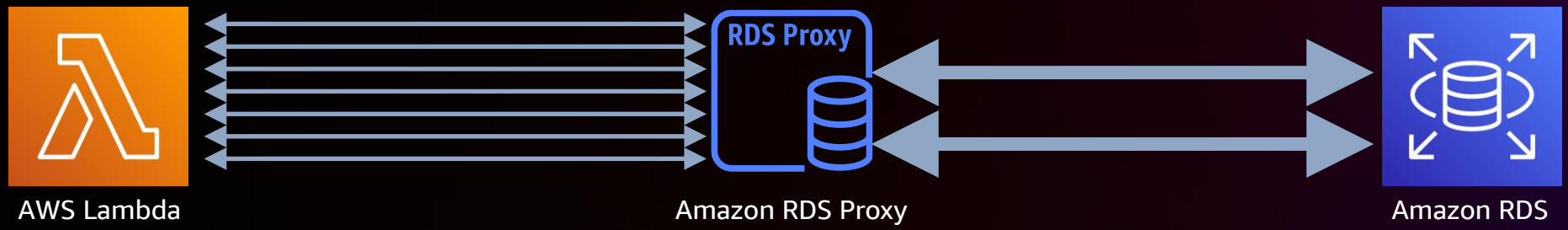
Handling Amazon RDS connections



Add buffers (queues)



Use aggregators (RDS Proxy)



aws
© 2022, Amazon Web Services, Inc. or its affiliates. All rights reserved.



Liel Ran
CTO & Co-Founder
DoControl

© 2022, Amazon Web Services, Inc. or its affiliates. All rights reserved.

About me



HELLO!

I'm Liel Ran.
CTO & Co-Founder

DoControl.

© 2022, Amazon Web Services, Inc. or its affiliates. All rights reserved.

SaaS Security Platform

Asset management
Users, collaborators, assets, groups, domains

Continuous monitoring
Anomaly detection, consolidated SaaS events, alerts

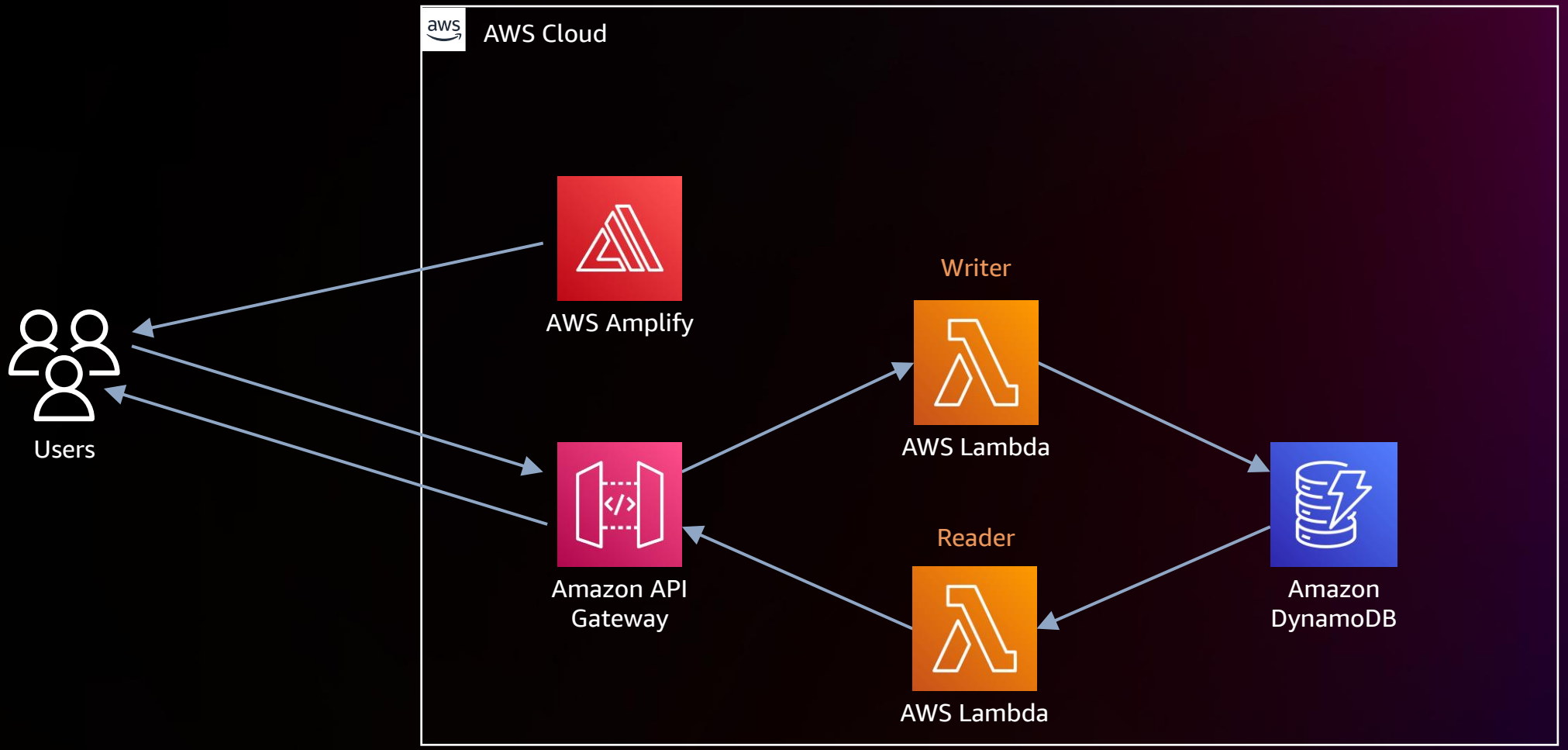
Security workflows
Trigger security workflows based on any SaaS event

Self-service remediation tools
Remove collab, revoke links, Slack DM approval, delete asset



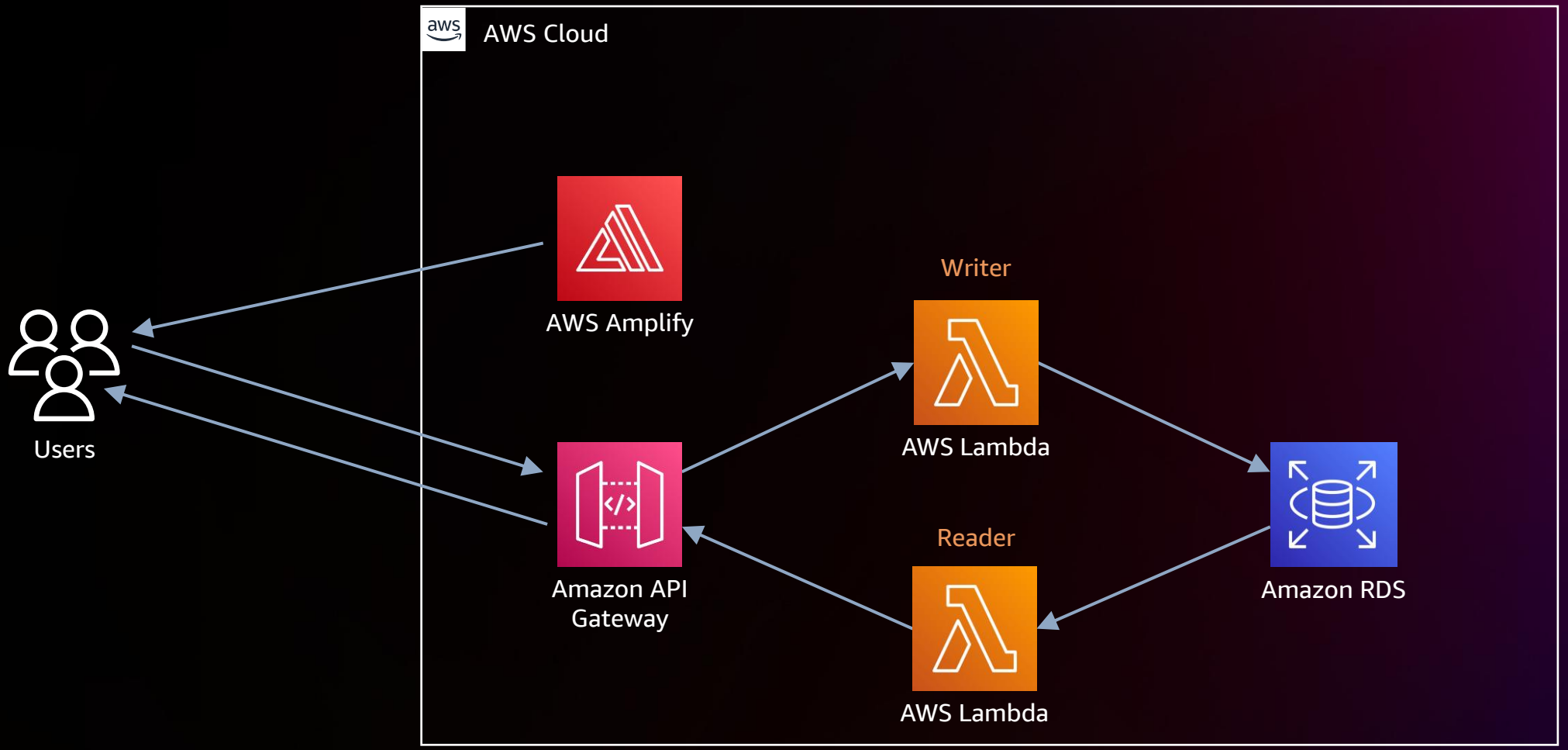
© 2022, Amazon Web Services, Inc. or its affiliates. All rights reserved.

Architecture slide – Amazon DynamoDB



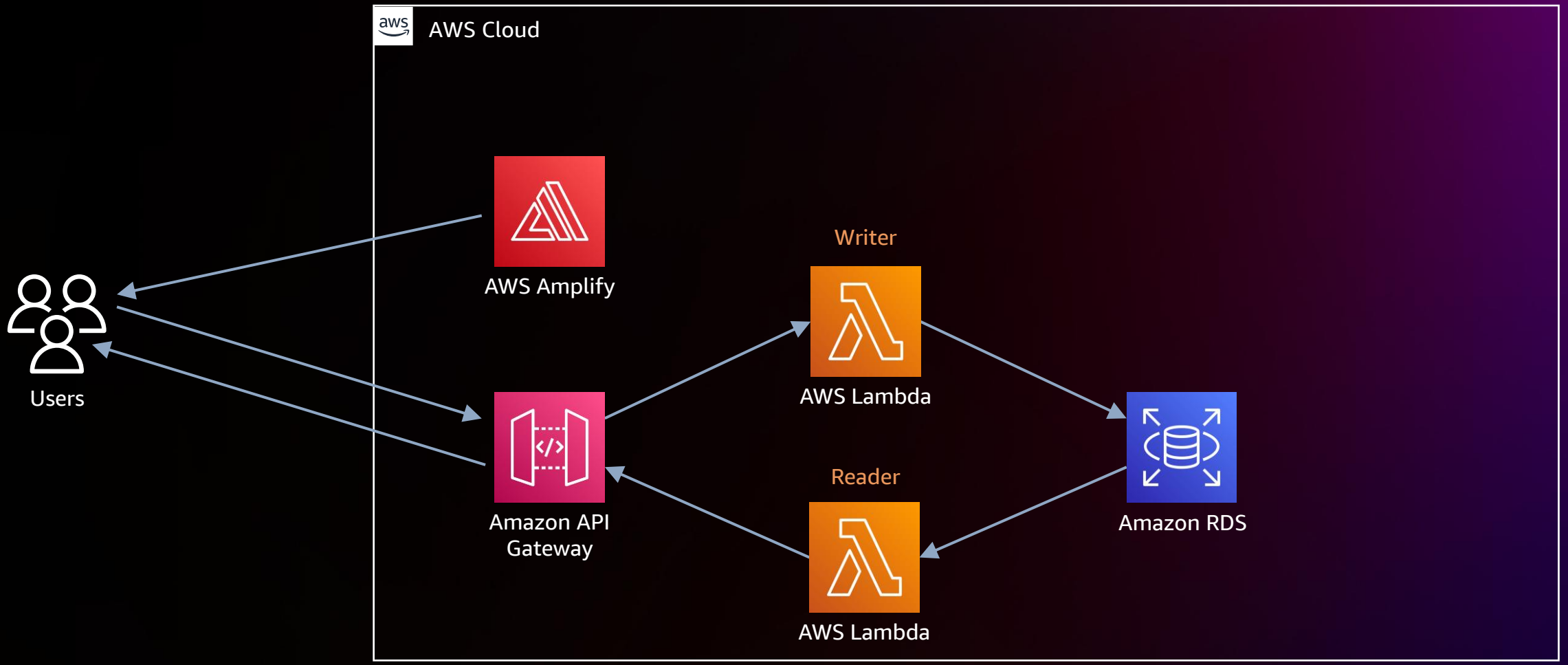
© 2022, Amazon Web Services, Inc. or its affiliates. All rights reserved.

Architecture slide – Amazon RDS



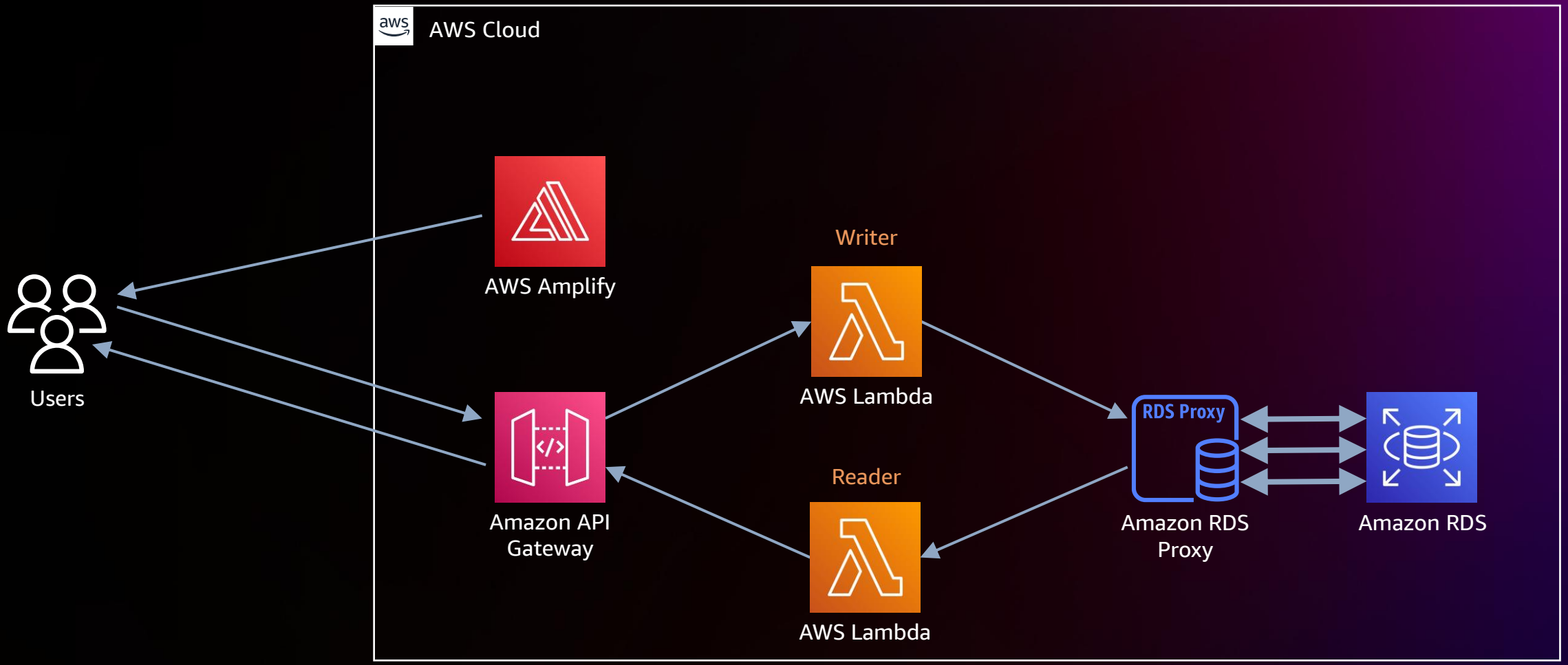
© 2022, Amazon Web Services, Inc. or its affiliates. All rights reserved.

Architecture slide – Amazon RDS



© 2022, Amazon Web Services, Inc. or its affiliates. All rights reserved.

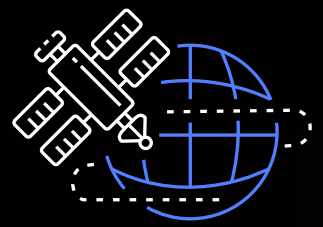
Architecture slide – Amazon RDS



© 2022, Amazon Web Services, Inc. or its affiliates. All rights reserved.

Amazon RDS Proxy

Fully managed, highly available database proxy feature for Amazon RDS; pools and shares DB connections to make applications more scalable, more resilient to database failures, and more secure



Pool and share DB connections for improved app scaling



Increase app availability and reduce DB failover times



Manage app data security with DB access controls



Fully managed DB proxy, compatible with your database

aws

© 2022, Amazon Web Services, Inc. or its affiliates. All rights reserved.

A few more options . . .

© 2022, Amazon Web Services, Inc. or its affiliates. All rights reserved.

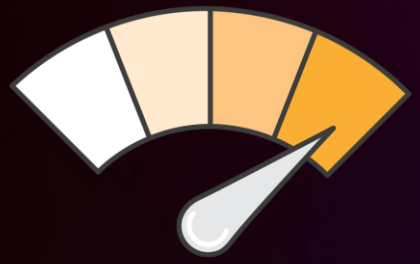
API Gateway throttling

- Protects backend systems
- Prevents one customer from consuming all your backend system capacity (with usage plans)
- Lets you decide how to allocate capacity among your API consumers with quotas and request rates

Legacy system plan:
10 RPS, up to 100 calls/day



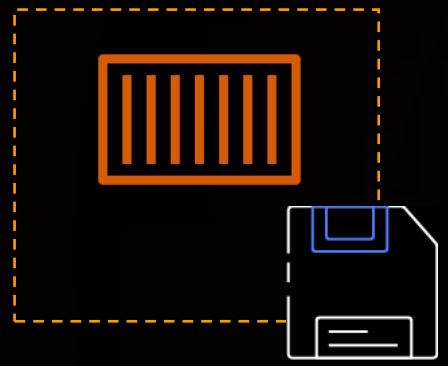
Modern system plan:
500 RPS, no limit on calls/day



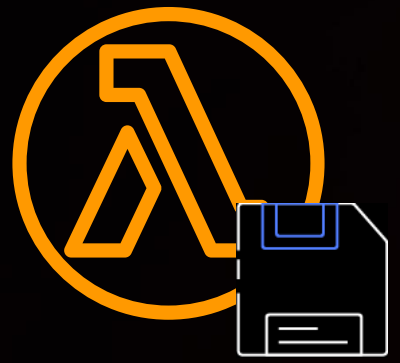
aws
© 2022, Amazon Web Services, Inc. or its affiliates. All rights reserved.

Cache layer

REDUCE NUMBER OF CALLS TO LEGACY SYSTEMS



Container



AWS Lambda

Storage
(Lambda /tmp 512 MB – 10 GB , local disk, EFS)
or
memory



Amazon ElastiCache



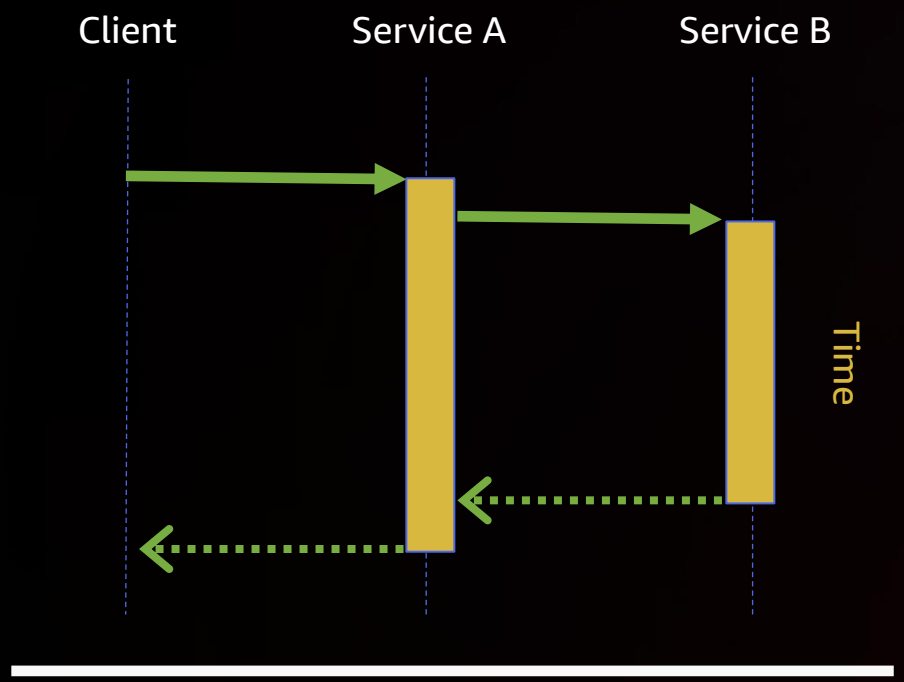
aws

© 2022, Amazon Web Services, Inc. or its affiliates. All rights reserved.

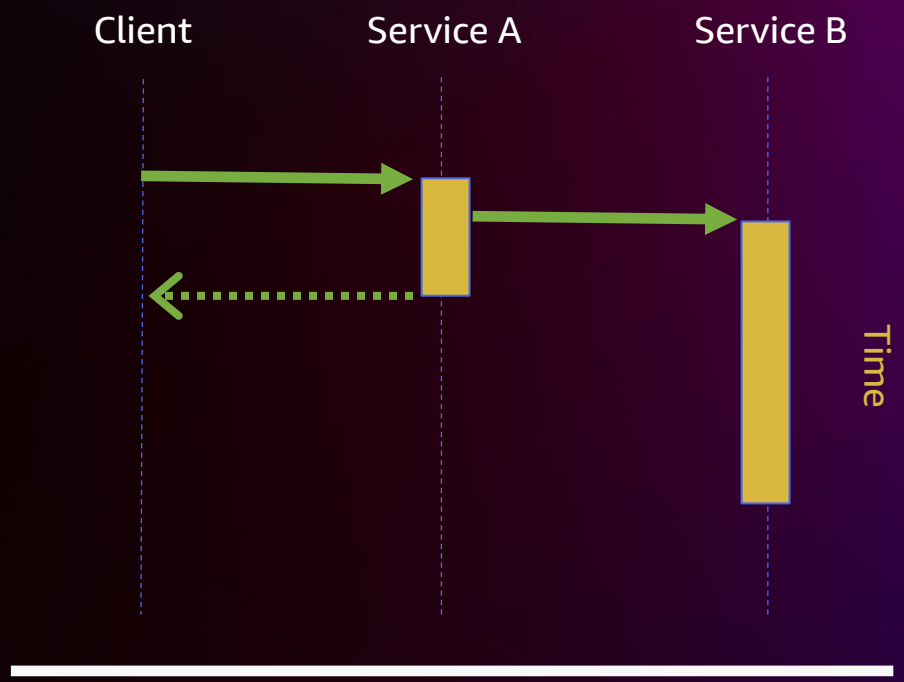
Asynchronous workloads

© 2022, Amazon Web Services, Inc. or its affiliates. All rights reserved.

Think asynchronously



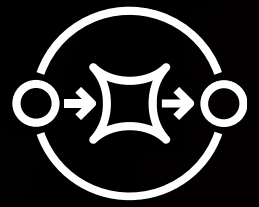
Synchronous commands



Asynchronous events

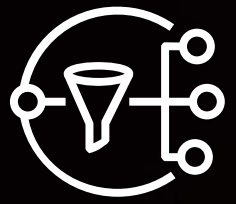
© 2022, Amazon Web Services, Inc. or its affiliates. All rights reserved.

Use asynchronous invocations and buffers



Amazon SQS

Queues
Simple
Fully managed
Any volume



Amazon SNS

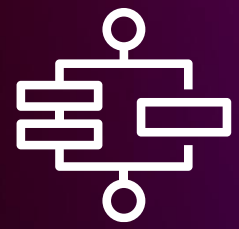
Pub/sub
Simple
Fully managed
Flexible



Amazon EventBridge

Synchronization
Rapid
Fully managed
Real time

API destinations
rate limiting



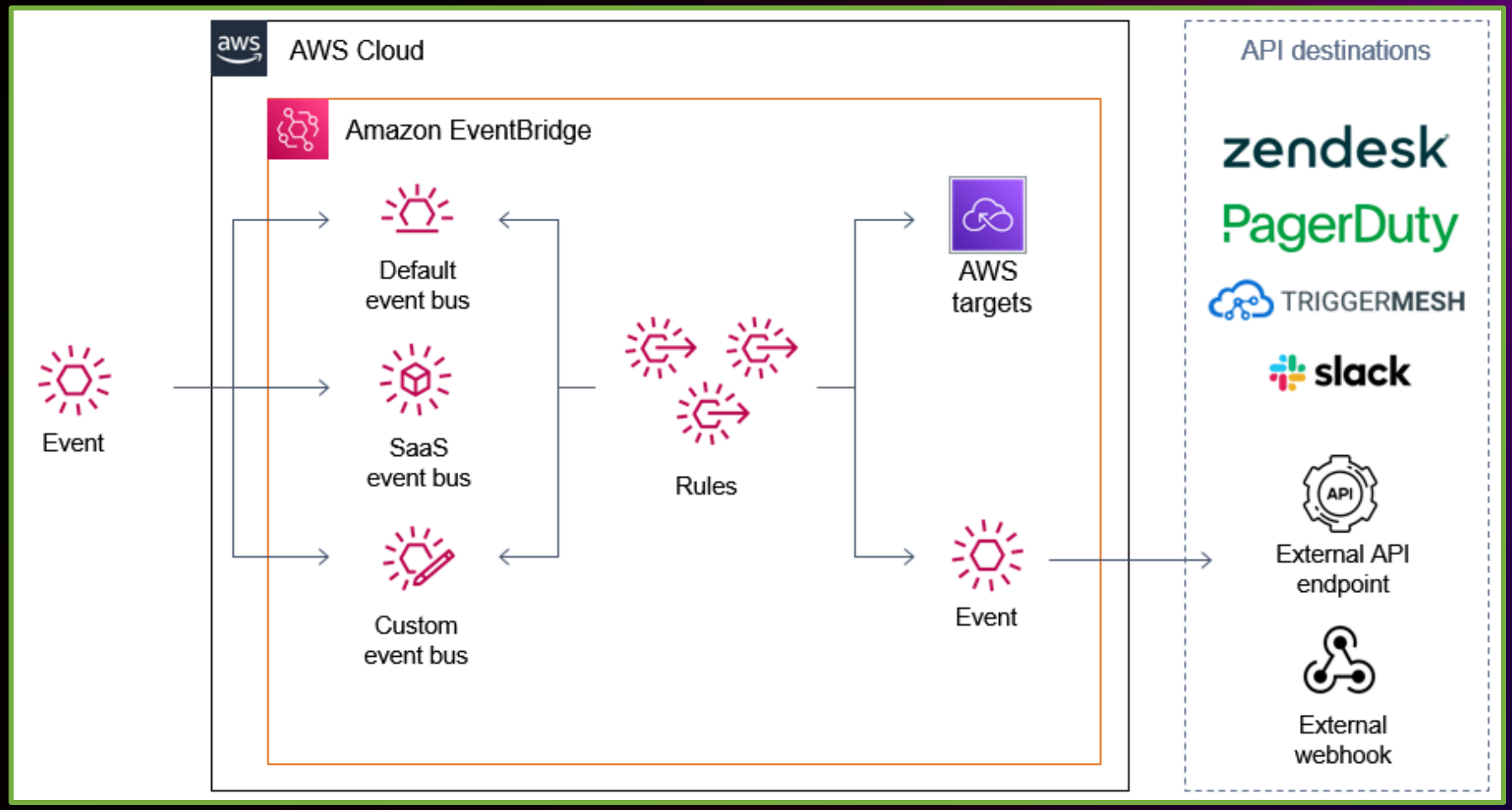
AWS Step Functions

Orchestration
Powerful
Fully managed
Low code

© 2022, Amazon Web Services, Inc. or its affiliates. All rights reserved.

Amazon EventBridge – API destinations

- API keys
- Basic Auth
- OAuth

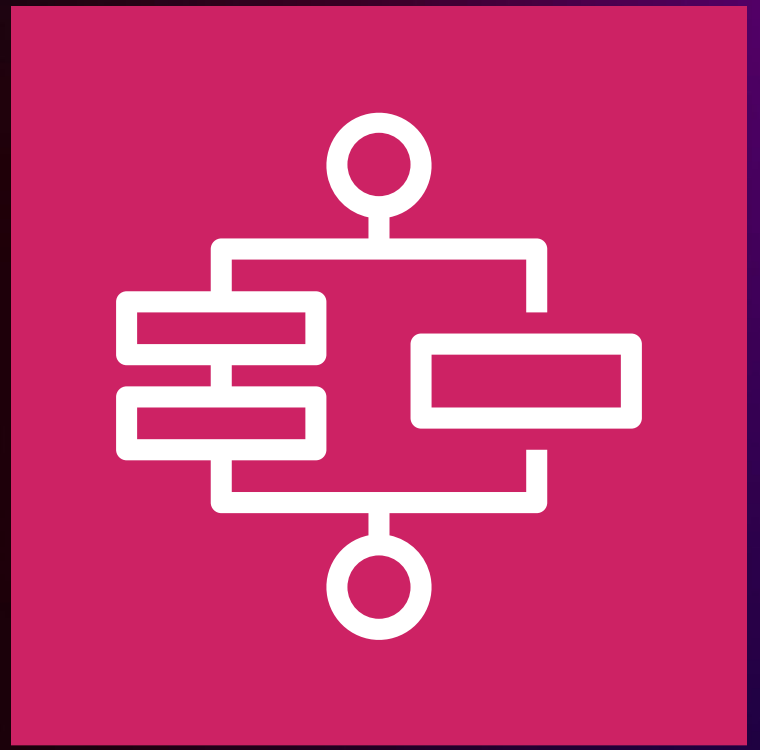


© 2022, Amazon Web Services, Inc. or its affiliates. All rights reserved.

AWS Step Functions

FULLY MANAGED STATE MACHINES ON AWS

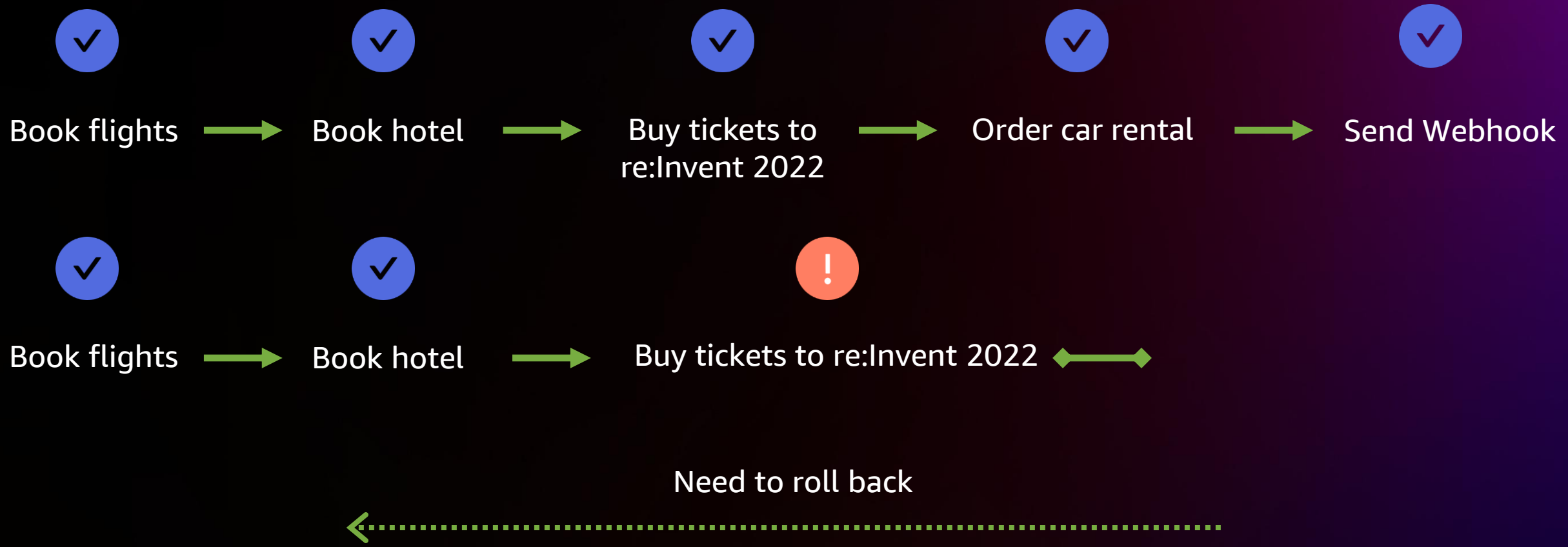
- Resilient workflow automation
- Built-in error handling
- Powerful AWS service integration
- First-class support for integrating with your own services
- Auditable execution history and visual monitoring



aws

© 2022, Amazon Web Services, Inc. or its affiliates. All rights reserved.

Working with state machines



© 2022, Amazon Web Services, Inc. or its affiliates. All rights reserved.

What can possibly go wrong?



Server-side error and general backend errors

5xx , network error, maintenance, downtime



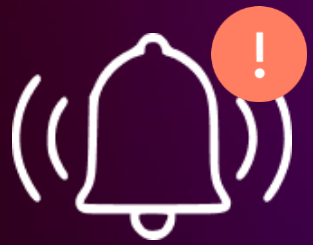
Runtime errors

“States.Runtime”



Timeout

“States.Timeout”



Rate limits

429 Too Many Requests

Need to roll back in case of any error

© 2022, Amazon Web Services, Inc. or its affiliates. All rights reserved.

External API calls – 429 Too Many Requests



© 2022, Amazon Web Services, Inc. or its affiliates. All rights reserved.

HTTP Too Many Requests



429 Too Many Requests

The HTTP `429 Too Many Requests` response status code indicates the user has sent too many requests in a given amount of time ("rate limiting").

A `Retry-After` header might be included to this response indicating how long to wait before making a new request.

Status

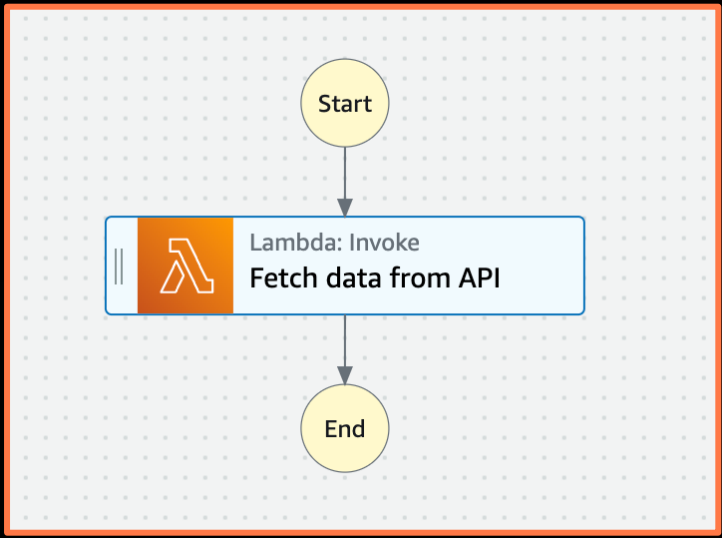
```
429 Too Many Requests
```

Example

```
HTTP/1.1 429 Too Many Requests
Content-Type: text/html
Retry-After: 3600
```

© 2022, Amazon Web Services, Inc. or its affiliates. All rights reserved.

Retry on error



Exponential backoff:
double the sleep latency
between any/selected errors

Interval - *optional*

The number of seconds before the first retry attempt.

2 seconds

Must be greater than zero.

Max attempts - *optional*

The maximum number of retry attempts.

6

Must be zero or greater.

Backoff rate - *optional*

The multiplier by which the retry interval increases with each attempt.

2

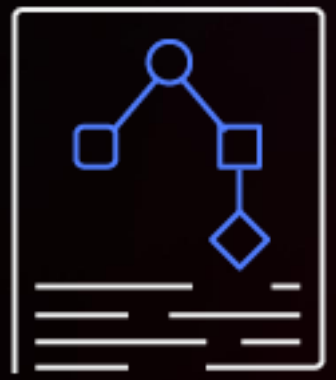
Must be greater or equal to 1.0.

aws
© 2022, Amazon Web Services, Inc. or its affiliates. All rights reserved.

AWS Step Functions demo



Basic implementation



Step Function
With default error handling



Step Function
Respect retry-after



github.com/lielran/reinvent2022-cdk-step-function

Disclaimer:
1. All external API calls are mocked with a 70% probability of getting an error
2. In case of error, 70% probability of getting a Rate Limit 429 error

aws
© 2022, Amazon Web Services, Inc. or its affiliates. All rights reserved.

Error-handling best practices

- Manage state machine outside of the code
- Use retries for catching general errors (runtime, timeouts, network)
- Respect rate limits and retry-after
- Use managed services(API Destination) with built-in support in backoff and retry-after



aws
© 2022, Amazon Web Services, Inc. or its affiliates. All rights reserved.



Thank you!

Liel Ran
CTO & Co-Founder
DoControl



© 2022, Amazon Web Services, Inc. or its affiliates. All rights reserved.

Be a good neighbor!
Implement rate limiting

© 2022, Amazon Web Services, Inc. or its affiliates. All rights reserved.

Thank you!

Liel Ran
@lielran

Boaz Ziniman
@ziniman



Please complete the session survey in the **mobile app**

aws

© 2022, Amazon Web Services, Inc. or its affiliates. All rights reserved.