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

DAT214

HBO Max achieves scale and performance with Amazon ElastiCache

Steven Hancz (he/him)
Specialist Solution Architect
Amazon Web Services

Itay Maoz
General Manager, In-Memory Database Services
Amazon Web Services

Leelavinod Bandla
Principal Software Engineer
Warner Bros. Discovery

Shabbir Yusuf
Senior Manager Data Engineering
Warner Bros. Discovery

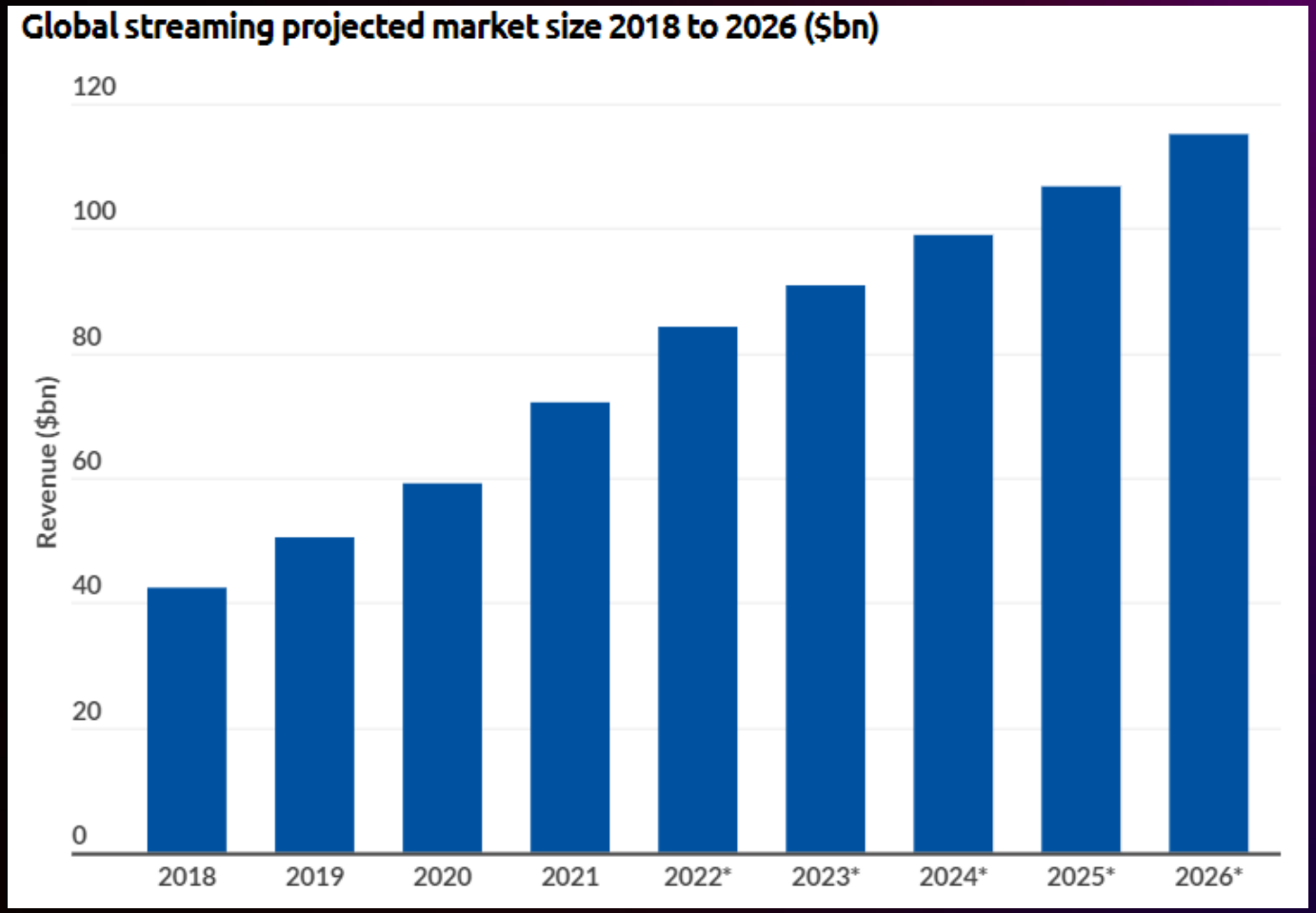
Agenda

- Customers expect fast response
- How HBO Max scaled to millions of subscribers with ElastiCache
- Overview of Amazon ElastiCache
- What's New with ElastiCache
- Q&A

Video streaming industry is growing fast

Video streaming app revenue

The video streaming industry reached $72.2 billion in 2021, with most of the revenue coming from the United States. It is projected to reach $115 billion by 2026.[1]



[1] Source: BusinessOfApps

Users demand real-time performance

- Most businesses operate online or mobile
- Users expect real-time performance

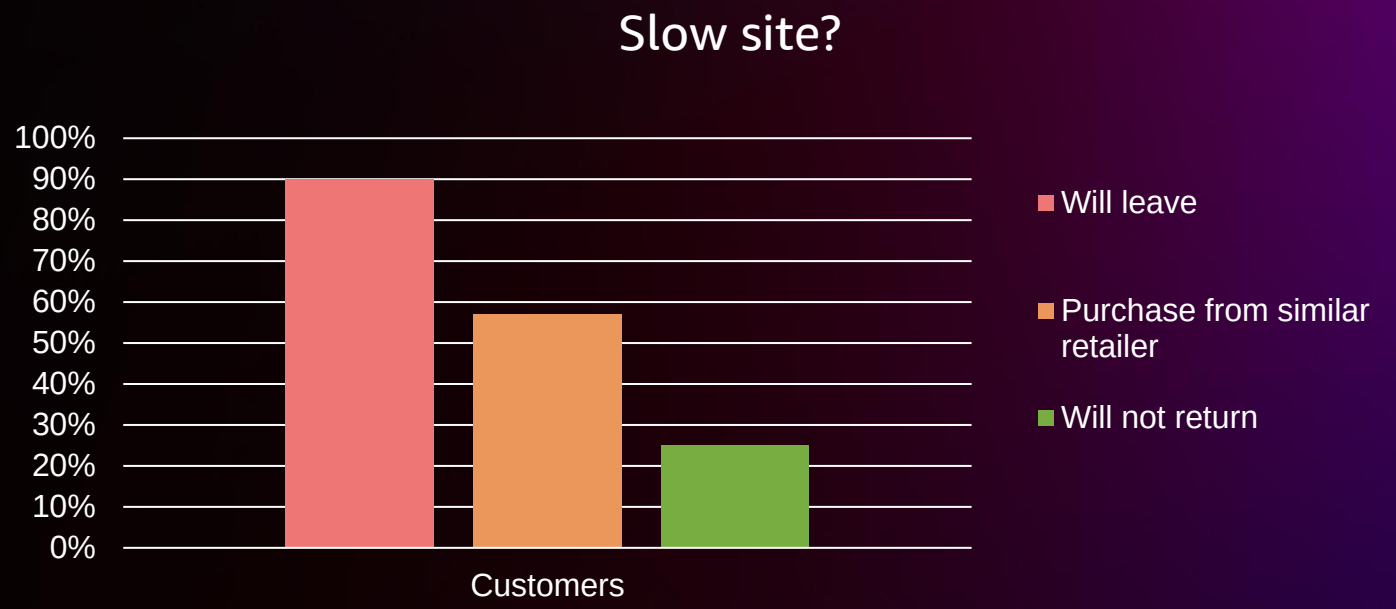
Latency is not an option

The less interactive an app becomes, the more likely users will move on to a competitor

Why does performance matter?

"The No. 1 reason people abandon a website after viewing just one page is that it's slow to load."[1]

"A 100-millisecond delay in website load time can hurt conversion rates by 7 percent."[2]



[1] Source Why Slow Website Performance Hurts Retail Websites 2022

[2] Source Akamai Online Retail Performance Report: Milliseconds Are Critical 2017

Speed alone is not sufficient

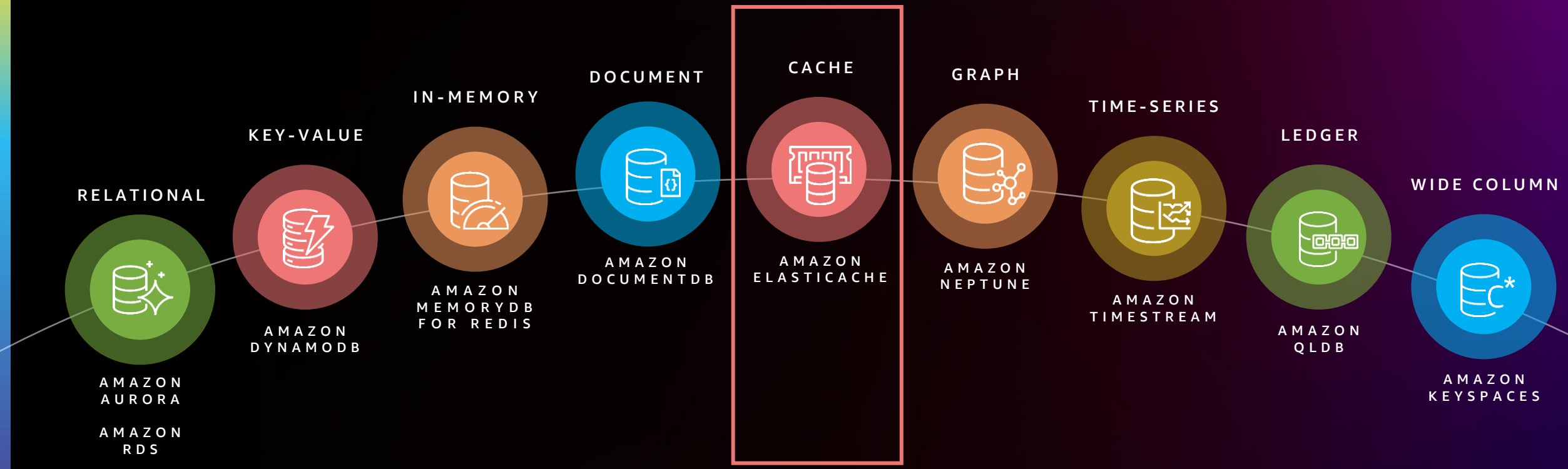
Session management

Relevant content recommendation

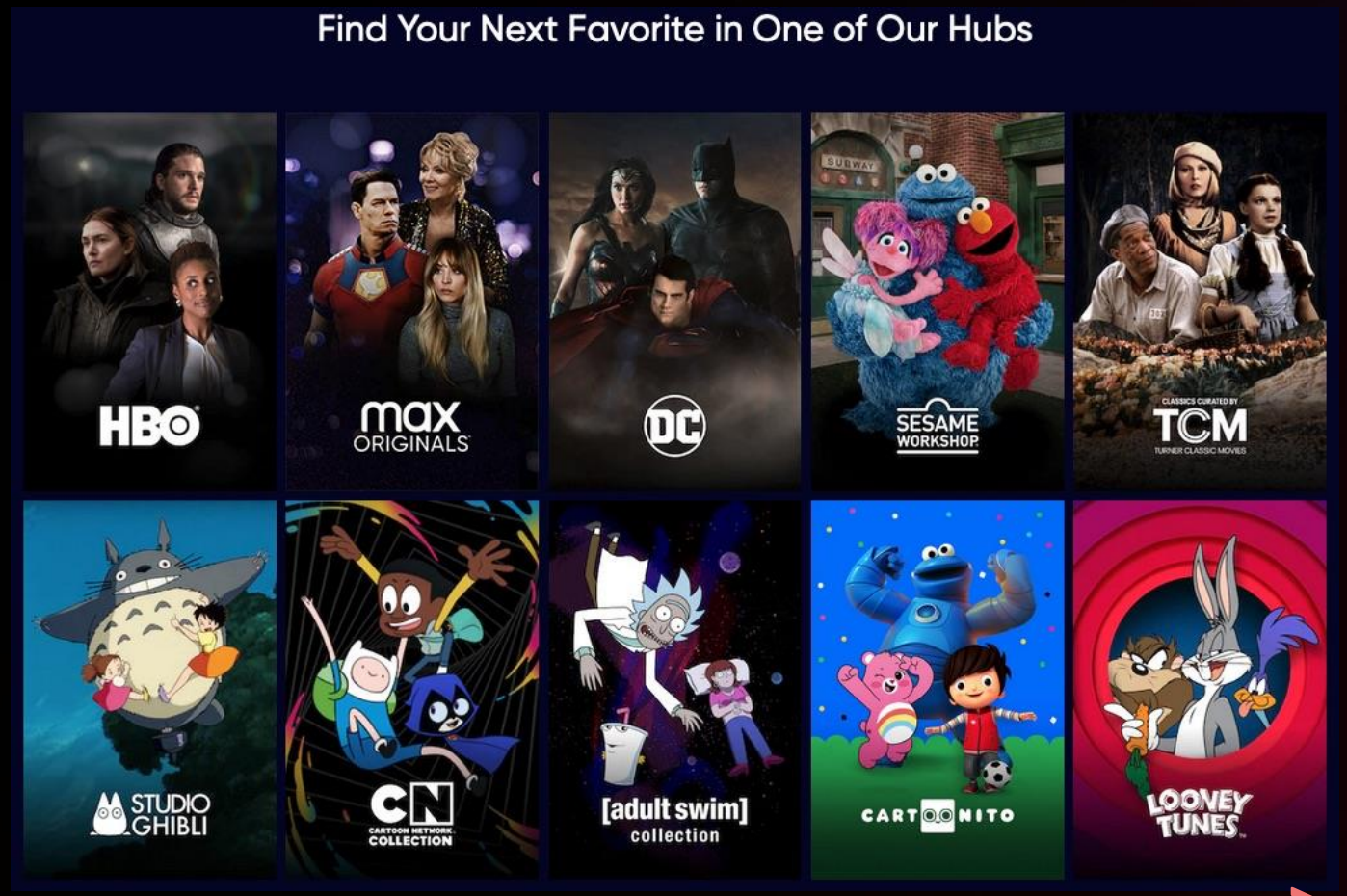
Scale to meet bursts in demand

Global reach

Broadest portfolio of purpose-built databases



How HBO Max achieves scale and performance with Amazon ElastiCache



HBO max

Watch films & shows you love

Across 61 countries

>94.9M subscribers in <2 years

Who we are and what we do

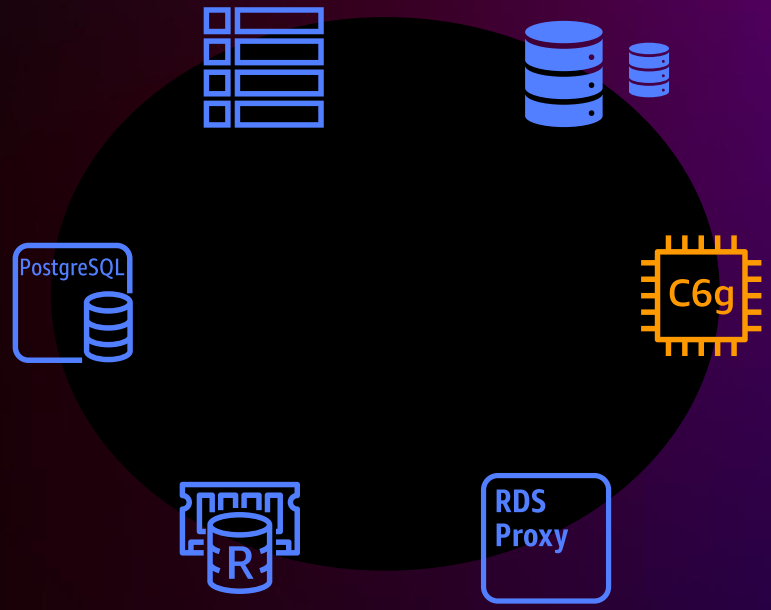
Choose from the Purpose-built datastores

Develop Infrastructure as Code modules

Collaborate with Service engineering

Provision the chosen databases

Size, Scale, Deploy & Tune



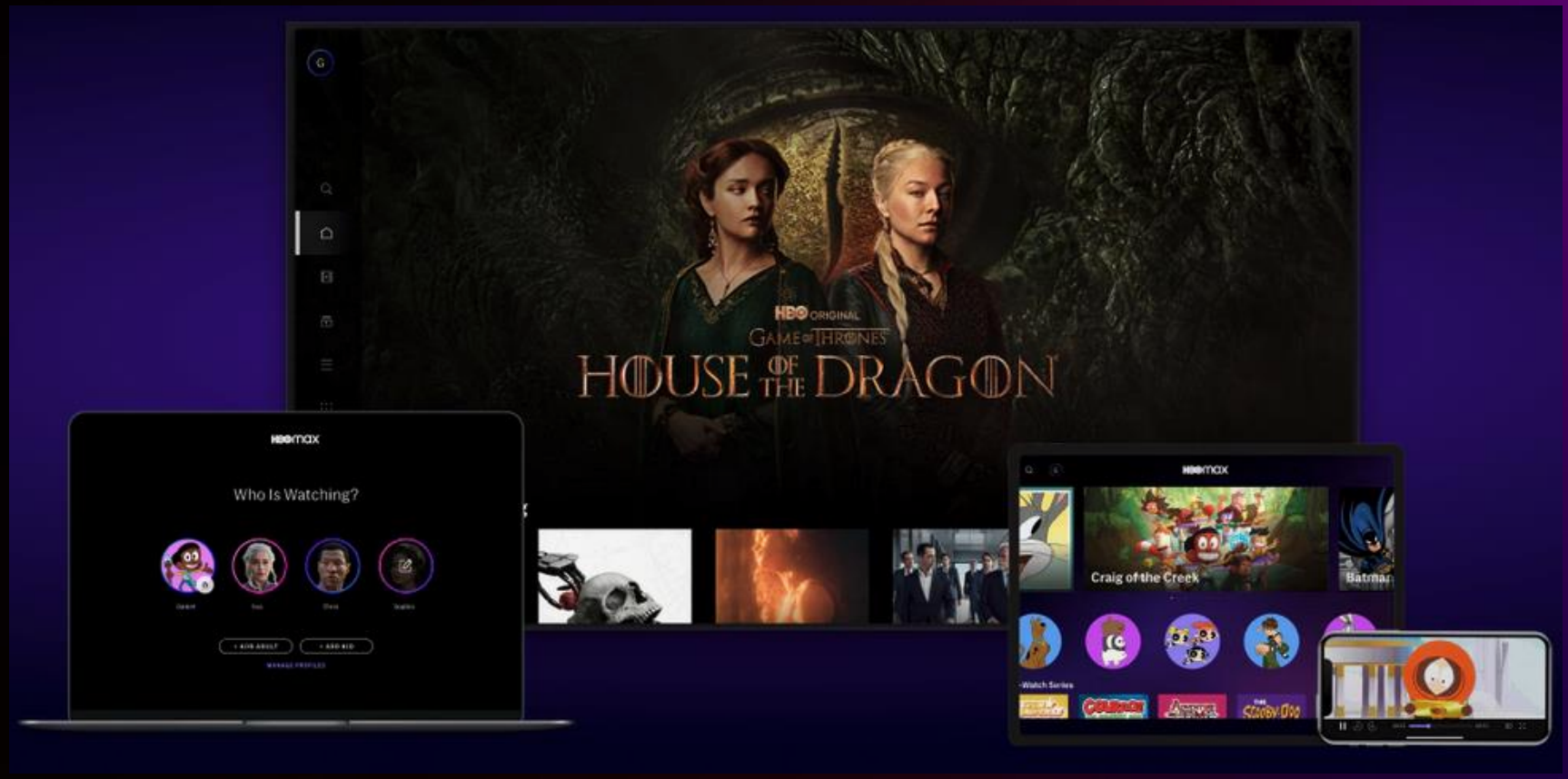
Handling Bursts in Demand

Prime time releases

Retry Storm

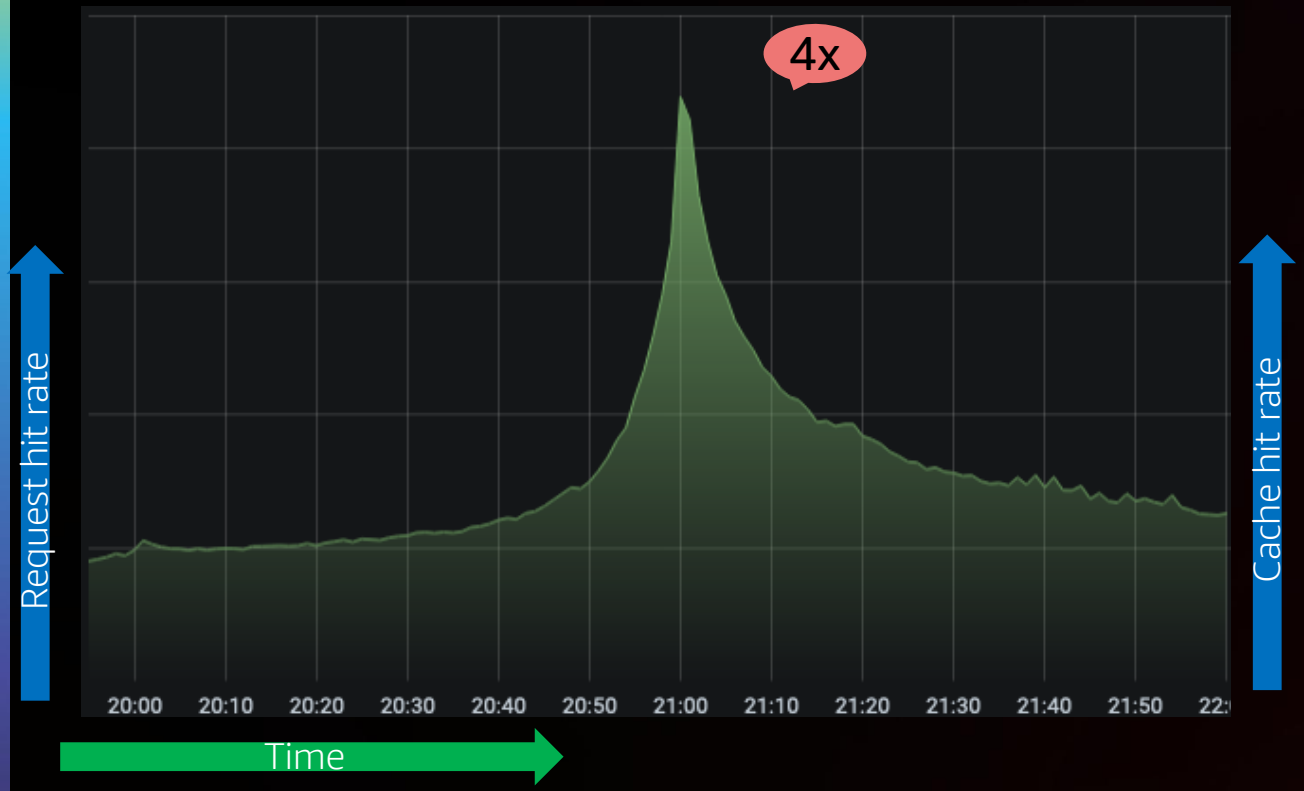
Burst requests at top of the hour

Services auto scale

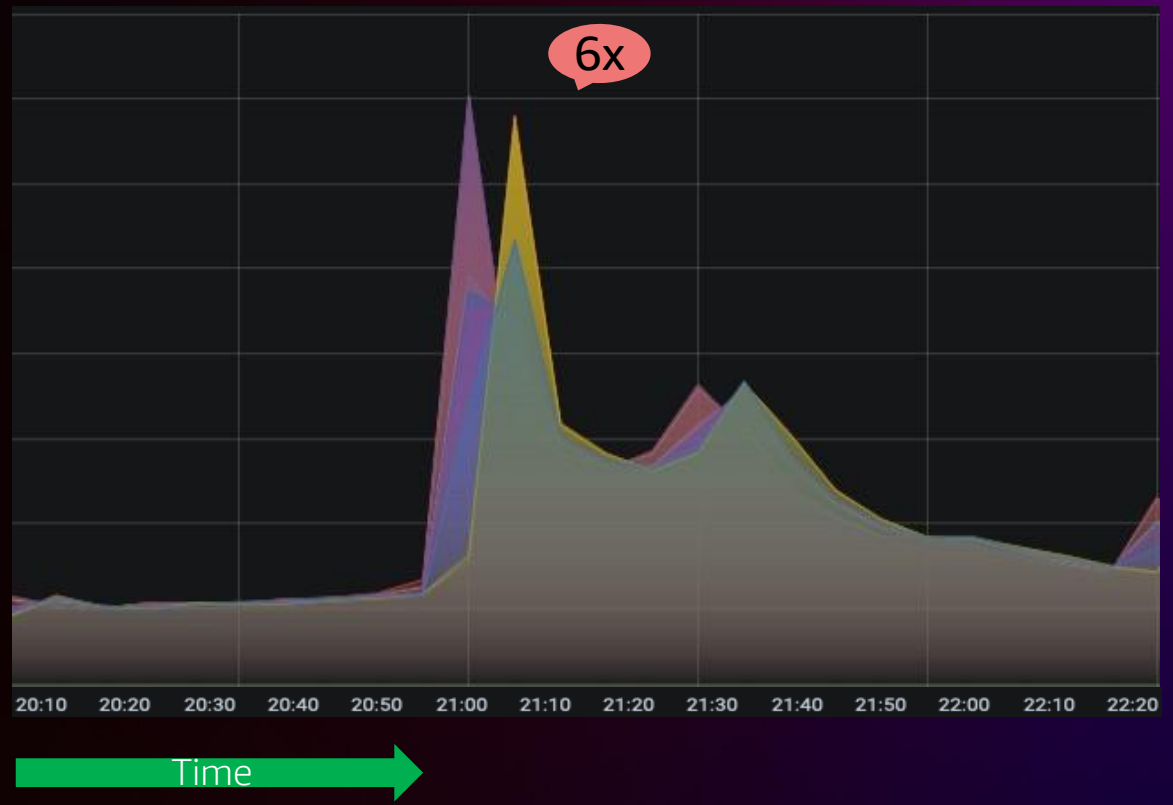


Massive Spikes caused by simultaneous user demand

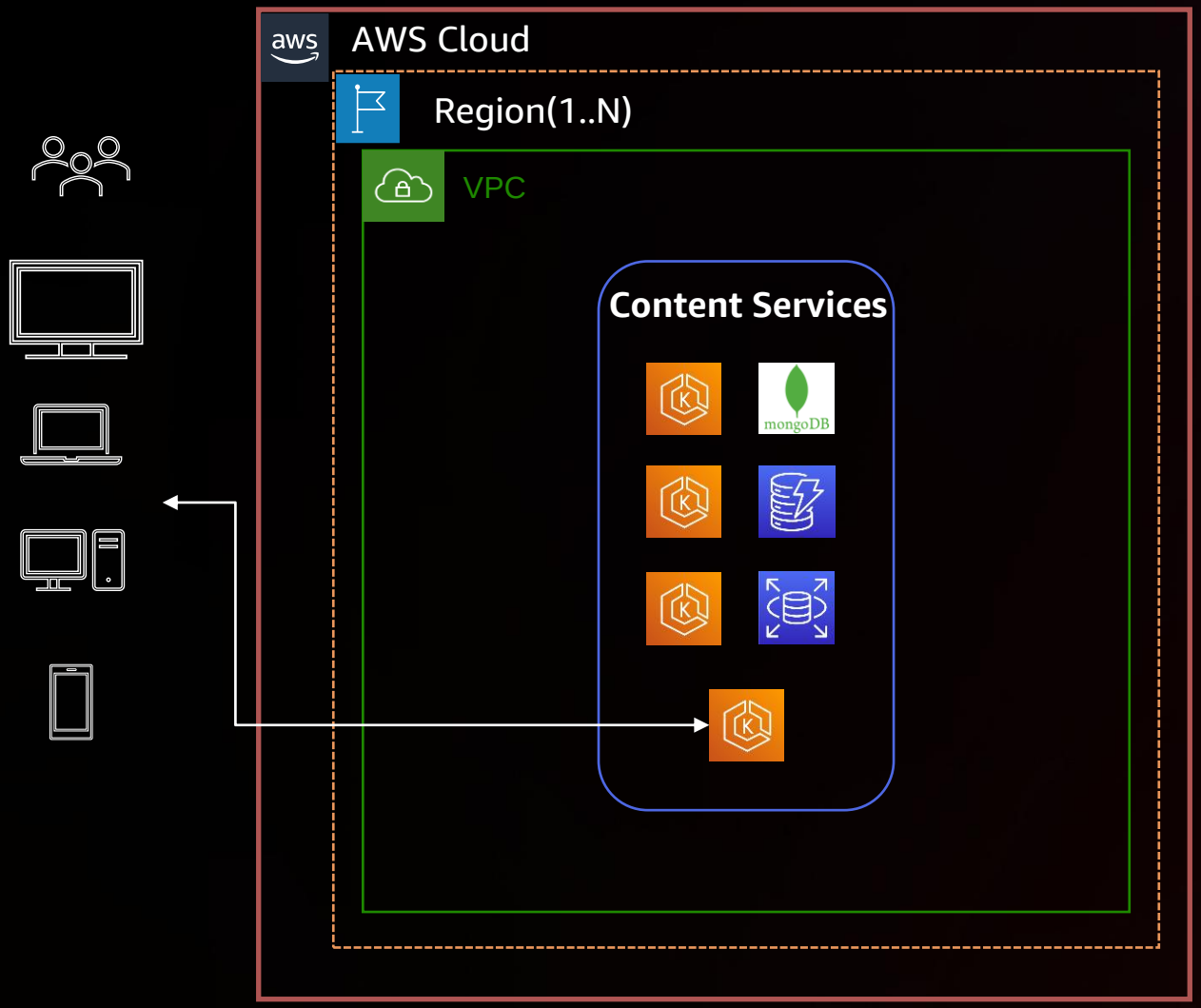
Spike in request pattern



Spike in cache hits



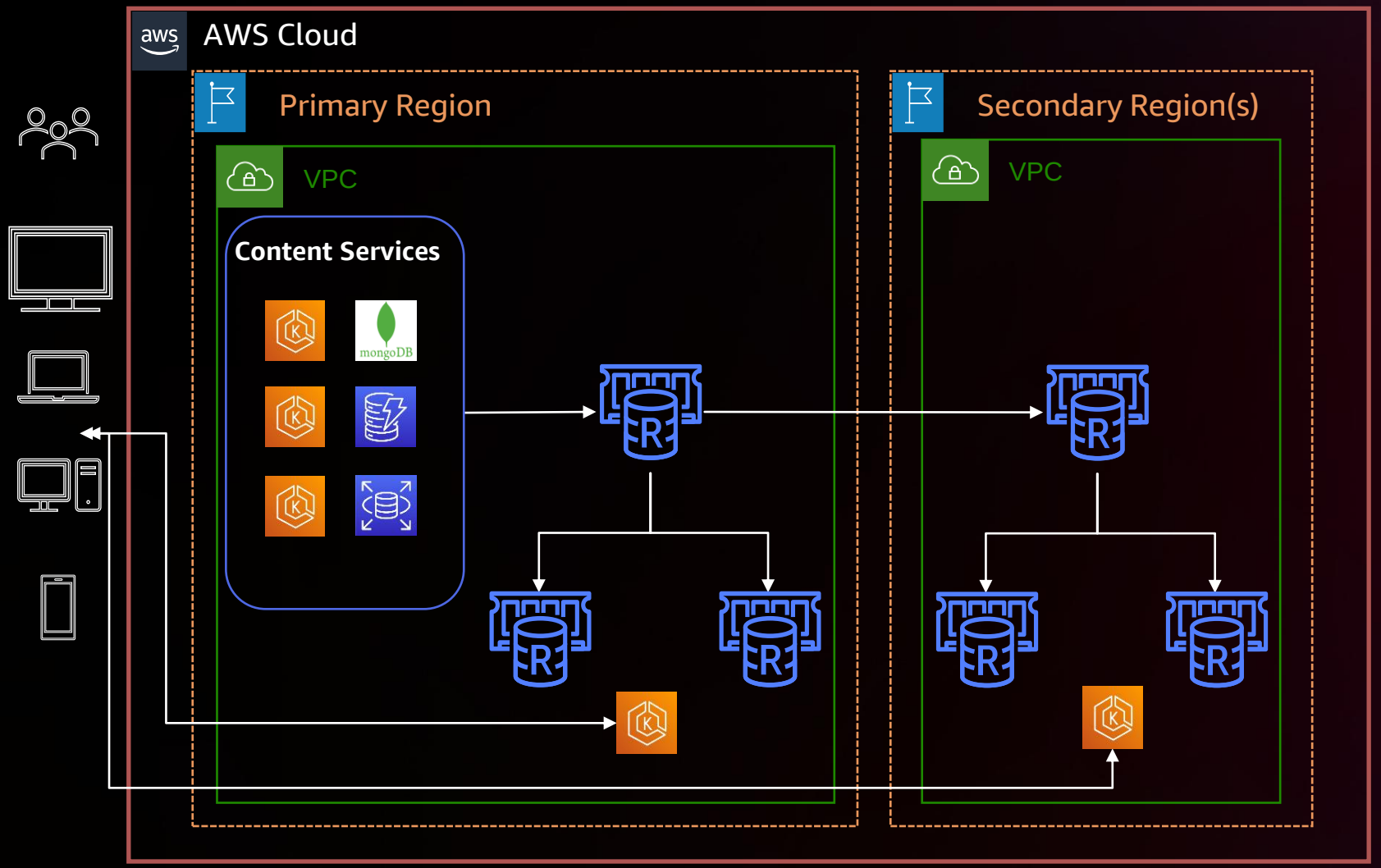
Prequel world: without Cache layer



Disadvantages:

- Multiple levels of penetration
- Many services need to scale
- Dependent on multiple service availabilities
- Increase/unpredictable response time
- Multiple possible points of failure
- Financially costly

New world: ElastiCache for Redis to the rescue!



Advantages:

- High Availability
- One layer serves multiple user requests
- Pre-generate timelines
- Write in primary region read in secondary regions
- Increased performance
- Financial savings

Amazon ElastiCache Global Datastore

Multi-Region, Multi-AZ

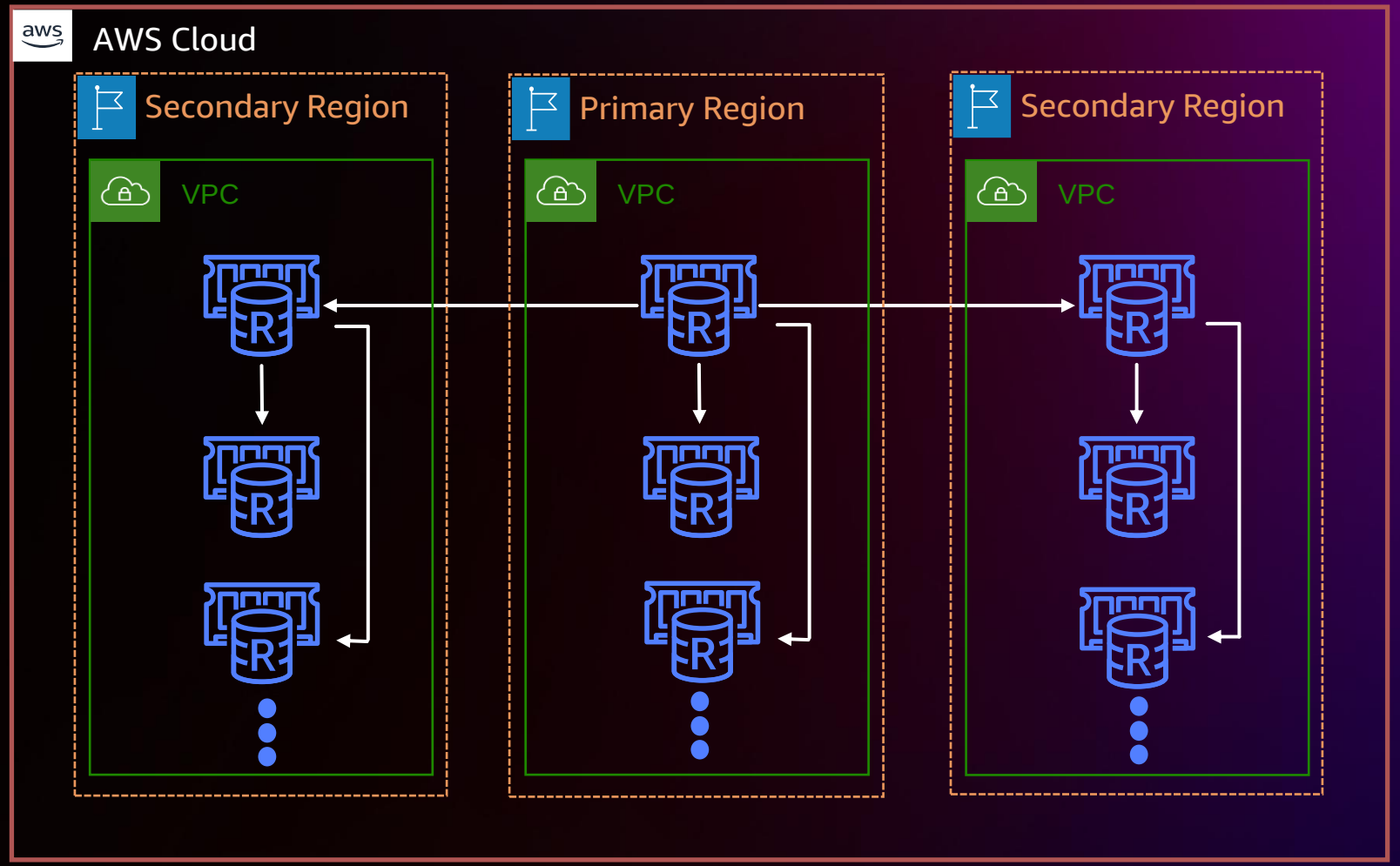
Cluster Mode Off

Periodic Writes for future timelines

Read current active timelines

Uses Graviton2 Instances

Highly performant



Redis Customization for spiky traffic

Parameter group
- repl-backlog-size
- maxmemory-policy

Vertical scale
- Network Throughput
- Memory
- Engine CPU

Horizontal scale
- 15 replicas in each Region

Batch size
- Lua script batches of 200 keys per iteration

No eviction policy (not using TTL or LRU eviction)
- Client side garbage collection for data purging
- To preserve graph timelines for extended period
- Pause graph builds for popular event

Reducing cost with Global Datastore

Architecture:

- Instead of replicating entire back end databases to meet regional demands
- Front them with scalable ElastiCache and distribute “hot” data regionally with Global Datastore
- Use Cluster Mode Enabled with minimum number of shards and replicas to reduce cost

Used by content services:

- Video-metadata service: uses PostgreSQL to store video related asset details
- Catalog service: uses MongoDB to store asset information

Financial advantages of pairing databases with ElastiCache:

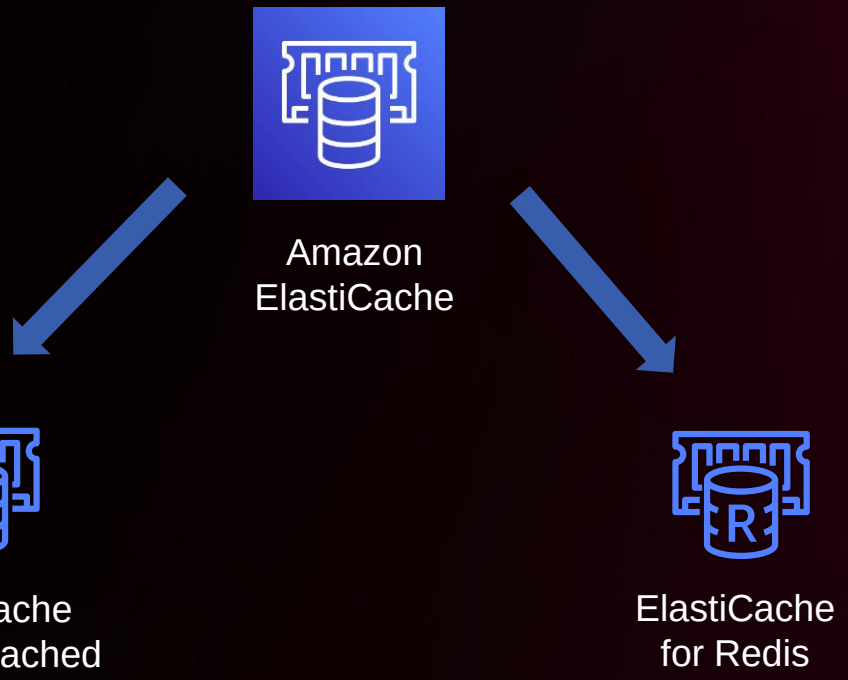
- Cost savings: hundreds of thousands of dollars on the above two services

Key ElastiCache Takeaways

- Configure clusters for high availability
- Enable cluster mode for more flexible scalability
- Use Graviton-based instances for best price per performance
- Pair databases with ElastiCache for better application performance at a lower cost
- Monitor slow-logs and engine logs for anomalies
- Perform periodic load tests

Overview of Amazon ElastiCache

Amazon ElastiCache



Amazon ElastiCache for Redis service

Redis compatible

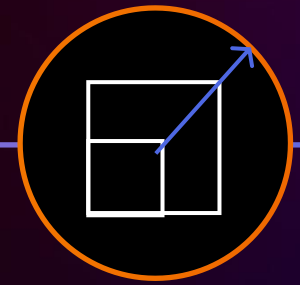
Fully managed

Highly available

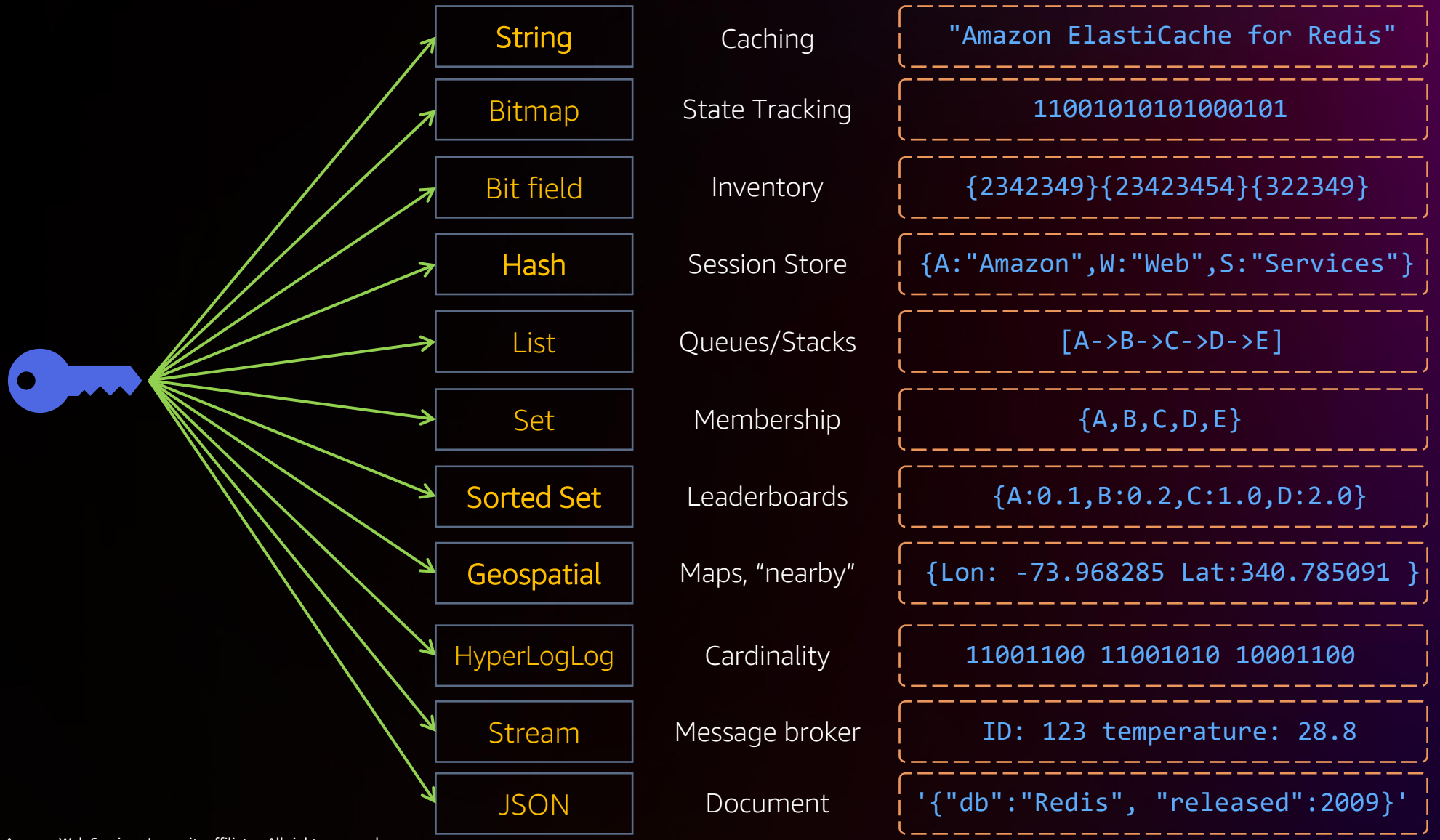
Extreme performance

Secure and compliant

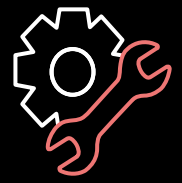
Easily scalable



Amazon ElastiCache is Redis Compatible

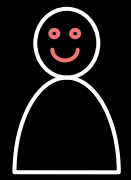


ElastiCache for Redis Advantages vs Self-Managed



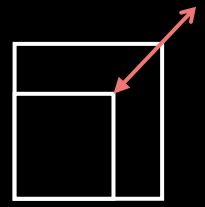
Managed by AWS

Server provisioning, software patching, setup, configuration, and backups



Highly available

Multi-AZ
with
Automatic failover
Multi Region



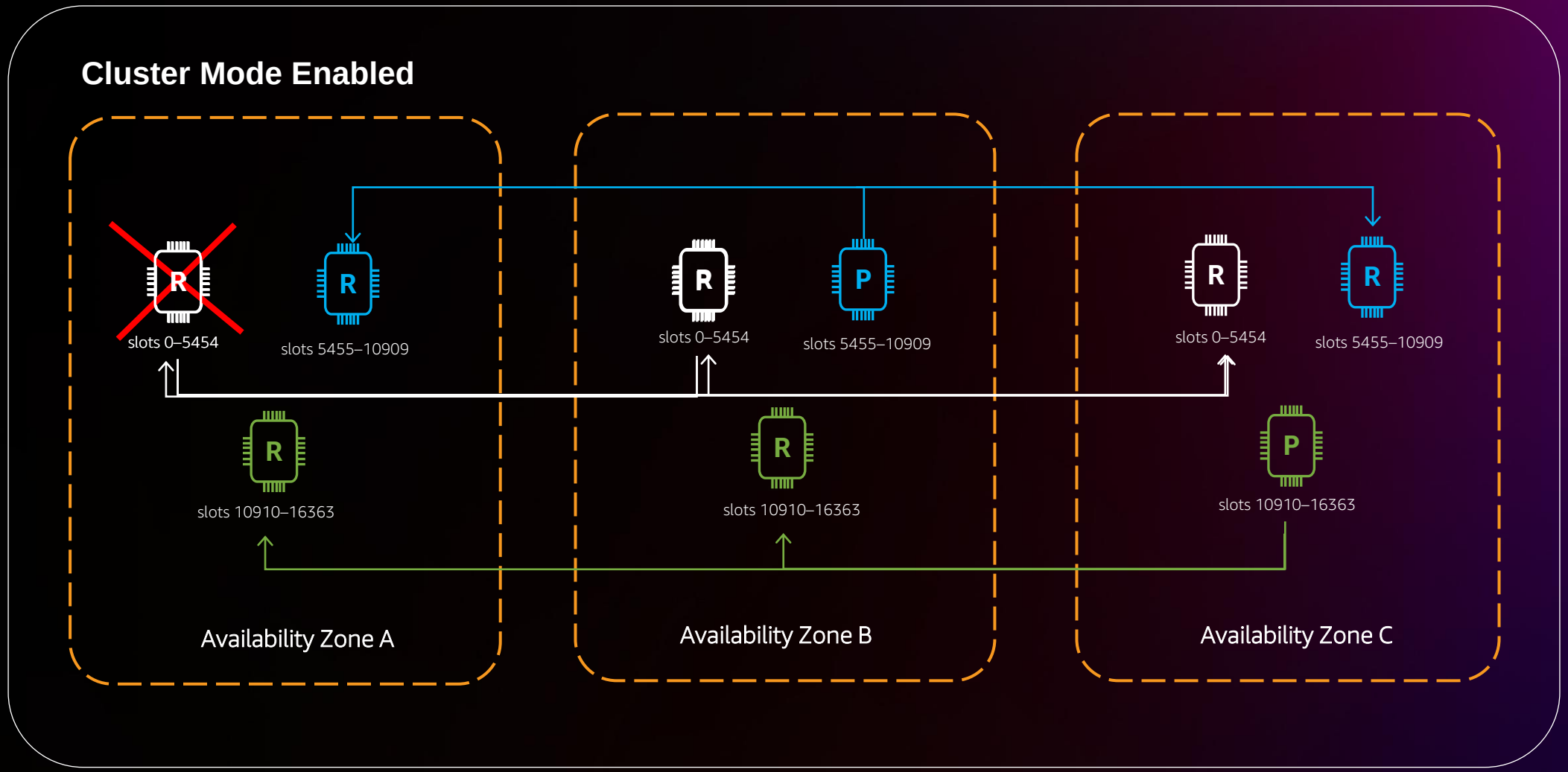
Auto scaling

Scale capacity to accommodate variable workloads

Cost Effective

Per instance billing
Without additional network charges
Reserved and Data Tiering instances

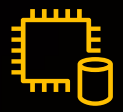
Highly Available with Automatic Failover



Extreme Performance

Microsecond response time

Millions of transactions per second



Scale up to **340 TB** of **in-memory capacity** and up to **1PB with data tiering**

Enhanced **I/O** handling boosts network processing

Optimized for Graviton chip and Nitro systems

Secure and Compliant

In-Transit Encryption

TLS encryption

All network flows encrypted

At-Rest Encryption

Via AWS Key Management Service (AWS KMS)

All data at rest is encrypted

Automatic memory encryption with AWS Graviton2

Authentication/Authorization

IAM authentication Redis v7

Access Control Lists Redis v6

Compliance

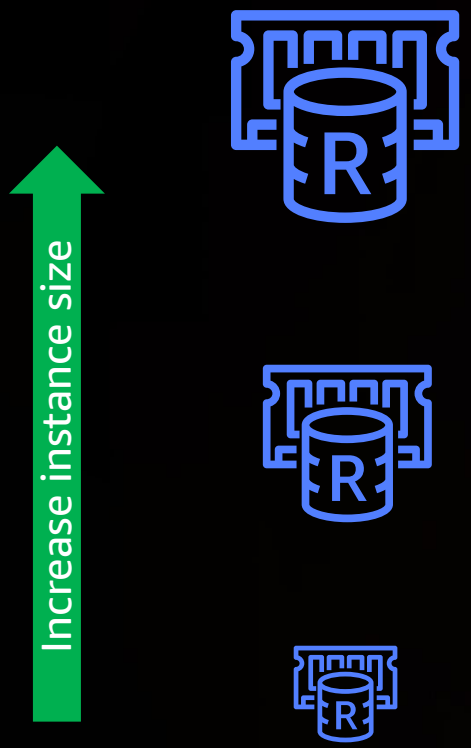
HIPAA eligible, ISO

PCI DSS compliant, SOC

FedRAMP authorized

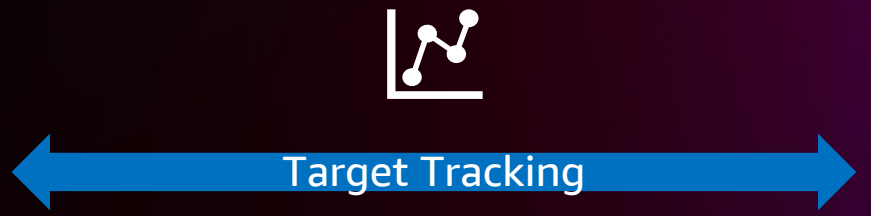
Scalable to Meet your Workload Demands

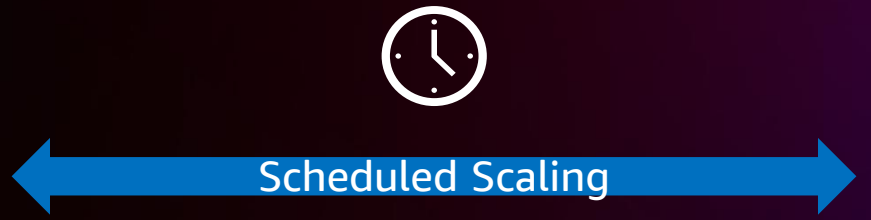
Cluster Mode Disabled



Cluster Mode Enabled

Autoscaling





Cross-Region Replication with Global Datastore

ElastiCache Use Cases

Web and mobile

Session management

Geospatial indexing

Retail

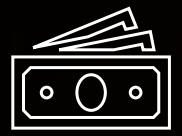
Customer profiles

Inventory tracking

Gaming

Leaderboards

Session history



Banking and finance

User transactions

Fraud detection

Machine Learning

Fast processing of data

Automate decision making

IoT

Streaming device data

Real time analytics

What’s new with ElastiCache

ElastiCache new features added in 2022

New and improved management console

Memcached v 1.6.12

Memcached TLS

Memcached FedRAMP and HIPAA eligible

IPv6 for both Redis and Memcached

ElastiCache new features added in 2022

Redis log delivery through Kinesis Data Firehose and CloudWatch Logs

Native JSON support

Private Link

Redis 7 (ACL v2, functions, sharded pub/sub)

Redis IAM authentication

ElastiCache 2022 availability and changes

How to get started, access technical resources

Amazon ElastiCache

Access webinars and videos, technical blogs, whitepapers, and presentations.

Deepen your skills with digital learning on demand.

Access ElastiCache **Resources**

Thank you!

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