
AWS
re:Invent

DAT209

What's new in Amazon ElastiCache

Barry Morris
General Manager
AWS

Kevin McGehee
Principal Engineer
AWS

AWS re:Invent

© 2020, Amazon Web Services, Inc. or its affiliates. All rights reserved.

Agenda

Real-time applications

Amazon ElastiCache

What’s new in 2020

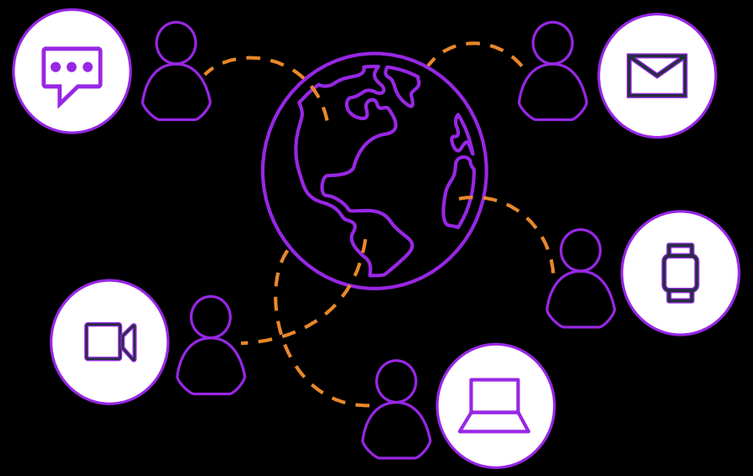


AWS re:Invent

© 2020, Amazon Web Services, Inc. or its affiliates. All rights reserved.

Modern application requirements

REQUIRES MORE PERFORMANCE, SCALE, AND AVAILABILITY





E-commerce



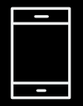
Media streaming



Social media



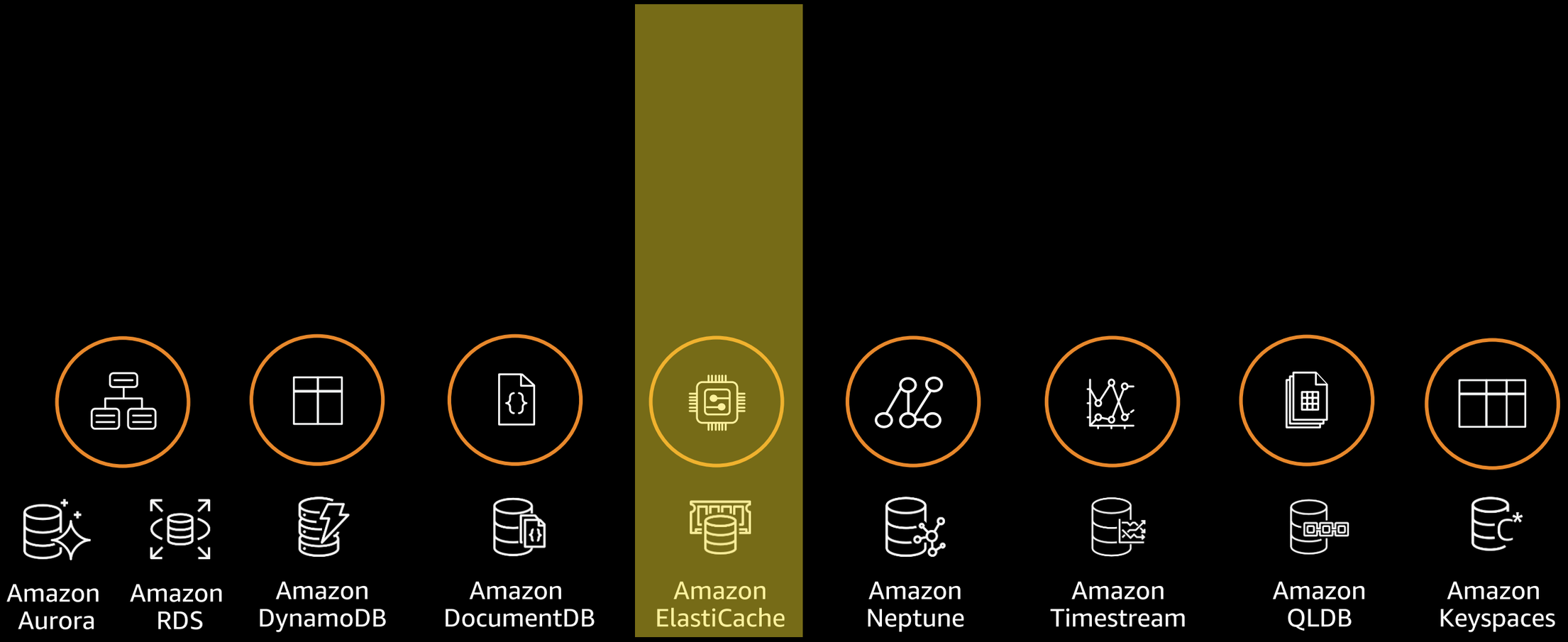
Online gaming



Shared economy

Users	1M+
Data volume	Terabytes – petabytes
Locality	Global
Performance	Microsecond latency
Request rate	Millions per second
Access	Mobile, IoT, devices
Scale	Up-down-out-in
Economics	Pay-as-you-go
Developer access	Open API

Purpose-built databases



In-memory engines

Memcached

Fast, simple, and robust distributed cache


Caching


Session store


Mobile applications


Web applications


API counters

Redis

Powerful, rich, highly available data structure store


Microservices


Mobile applications


Web applications


Gaming leaderboards


Message queues


Geospatial


Chat apps


Real-time analytics


Machine learning


Caching


Session store


API counters

Redis – “most loved” data store

OPEN SOURCE IN-MEMORY DATA STRUCTURE STORE

Intuitive, powerful programming model

~200 commands, Lua scripting, and pub/sub

Data structures first

Strings, lists, hashes, sets, sorted sets, streams, geospatial, bitmaps, HyperLogLogs

Atomic operations

Supports transactions

Fast

<1 ms latency for most commands

Highly available

Multi-AZ replication with autofailover

Backup/restore

Daily and on-demand snapshotting

Leaderboard example – Sorted Set

Populating a leaderboard

```
> zadd leaderboard 556 Andy
> zadd leaderboard 819 Bob
> zadd leaderboard 105 Carl
> zadd leaderboard 1312 Derek
```

Querying a leaderboard

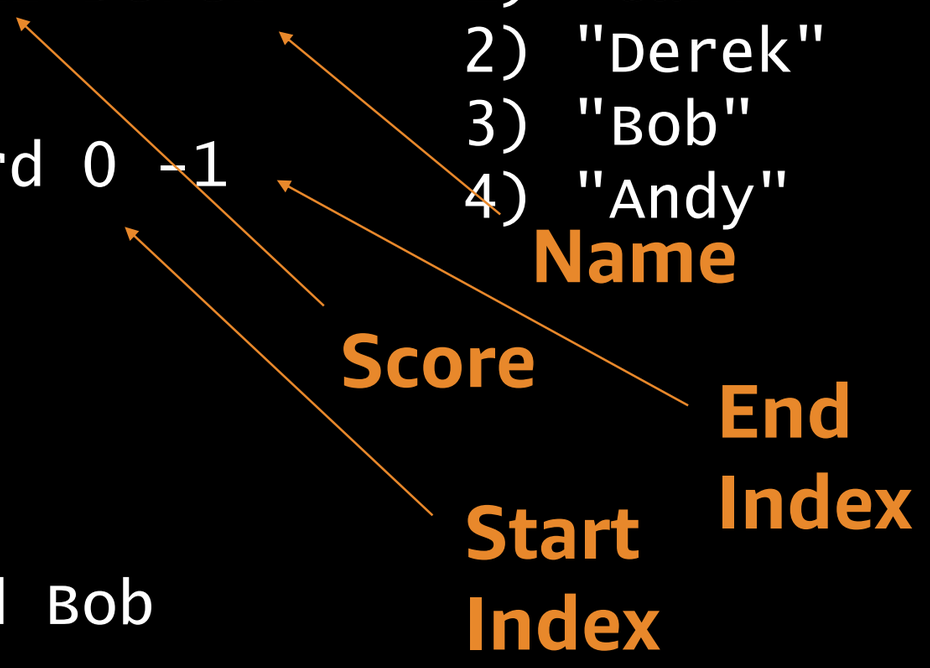
```
> zrevrange leaderboard 0 -1
1) "Derek"
2) "Bob"
3) "Andy"
4) "Carl"

> zrevrank leaderboard Bob
1
```

Updating a leaderboard

```
> zadd leaderboard 1502 Carl

> zrevrange leaderboard 0 -1
1) "Carl"
2) "Derek"
3) "Bob"
4) "Andy"
```



Amazon ElastiCache

MANAGED, REDIS OR MEMCACHED-COMPATIBLE IN-MEMORY DATA STORE



Scalable

Write and memory scaling with sharding; non-disruptive scaling; read scaling with replicas (Redis)



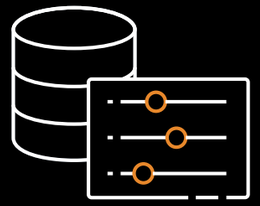
Extreme performance

In-memory data store and cache for sub-millisecond response times



Fully managed

AWS manages all hardware and software setup, configuration, monitoring



OSS Compatible

Fully compatible with open source Redis and Memcached engines

Amazon ElastiCache for Redis



Extreme performance

In-memory data store for sub-millisecond response times



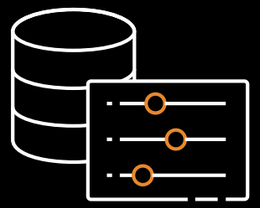
Secure

Role-based access control (RBAC), HIPAA eligible, PCI compliant, FedRAMP, encryption at rest, in-transit, Amazon VPC



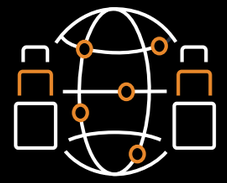
Fully managed

Hardware provisioning and software patching, configuration, monitoring, failure recovery, and backups



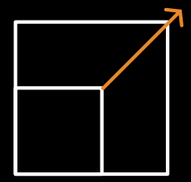
Redis-compatible

Supports Redis 6, maintains compatibility with open source Redis, Redis APIs, works with Redis clients



Highly available, reliable

Supports Redis cluster and non-cluster modes, automatic failover, cross-region replication, multi-AZ



Easily scalable

Non-disruptive scaling up to 500 nodes and 500 shards, read scaling with replicas, write scaling with sharding

Amazon ElastiCache for Redis customers



Using for **caching layer** for real-time analytics

A/B testing on Expedia.com website with **~200 million messages daily**



Using for **caching layer** for real-time data computation for **daily bookings**



Session state for applications

Help developers build **container-based stateless applications** on AWS

Amazon ElastiCache: What's new in 2020



AWS re:Invent

© 2020, Amazon Web Services, Inc. or its affiliates. All rights reserved.

What’s new in 2020

Performance



Graviton2 instance support

AWS Outposts and Local Zones

Security



Resource-level API permissions

Role-based access control for Redis

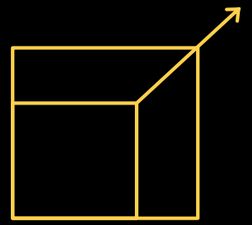
Auto-update of clusters

Microservice integration



AWS Controllers for Kubernetes (ACK)

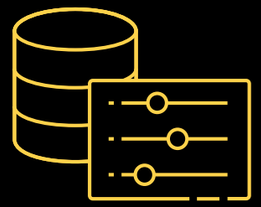
Scalability and availability



Global Datastore

500 node clusters

OSS compatibility



Redis 6 with RBAC, client-side caching

Memcached 1.6

Performance



© 2020, Amazon Web Services, Inc. or its affiliates. All rights reserved.

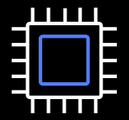
AWS Graviton2 processor

Enabling the best price / performance for your cloud workloads

Graviton Processor



First Arm-based processor available in major cloud



Built on 64-bit Arm Neoverse cores with AWS-designed silicon using 16 nm manufacturing technology



Up to 16 vCPUs,10 Gbps enhanced networking, 3.5 Gbps Amazon EBS bandwidth

Graviton2 Processor



4x compute cores, and 5x faster memory



Built with 64-bit Arm Neoverse cores with AWS-designed silicon using 7 nm manufacturing technology



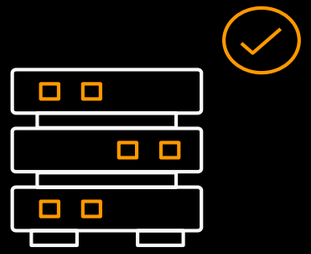
Up to 64 vCPUs, 25 Gbps enhanced networking, 18 Gbps Amazon EBS bandwidth

ElastiCache Graviton2-based instances

- Significant price-performance improvement over prior generation
- Always-on DRAM encryption
- Supports m6g and r6g node families
  - Single-node dataset sizes from 6GiB (m6g.large) to over 400GiB (r6g.16xlarge)
- Available for seamless upgrades or new clusters

AWS Local Zones

- New type of AWS infrastructure deployment
- Places compute, storage, database, and other services closer to customers
- For demanding applications that require single-digit latencies



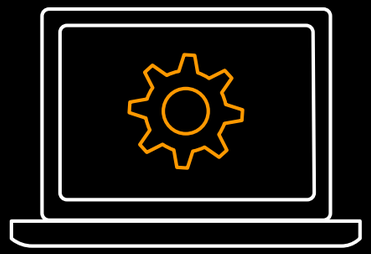
AWS infrastructure at the edge



Local compute, storage, database, and other services



Connect to services in AWS Regions



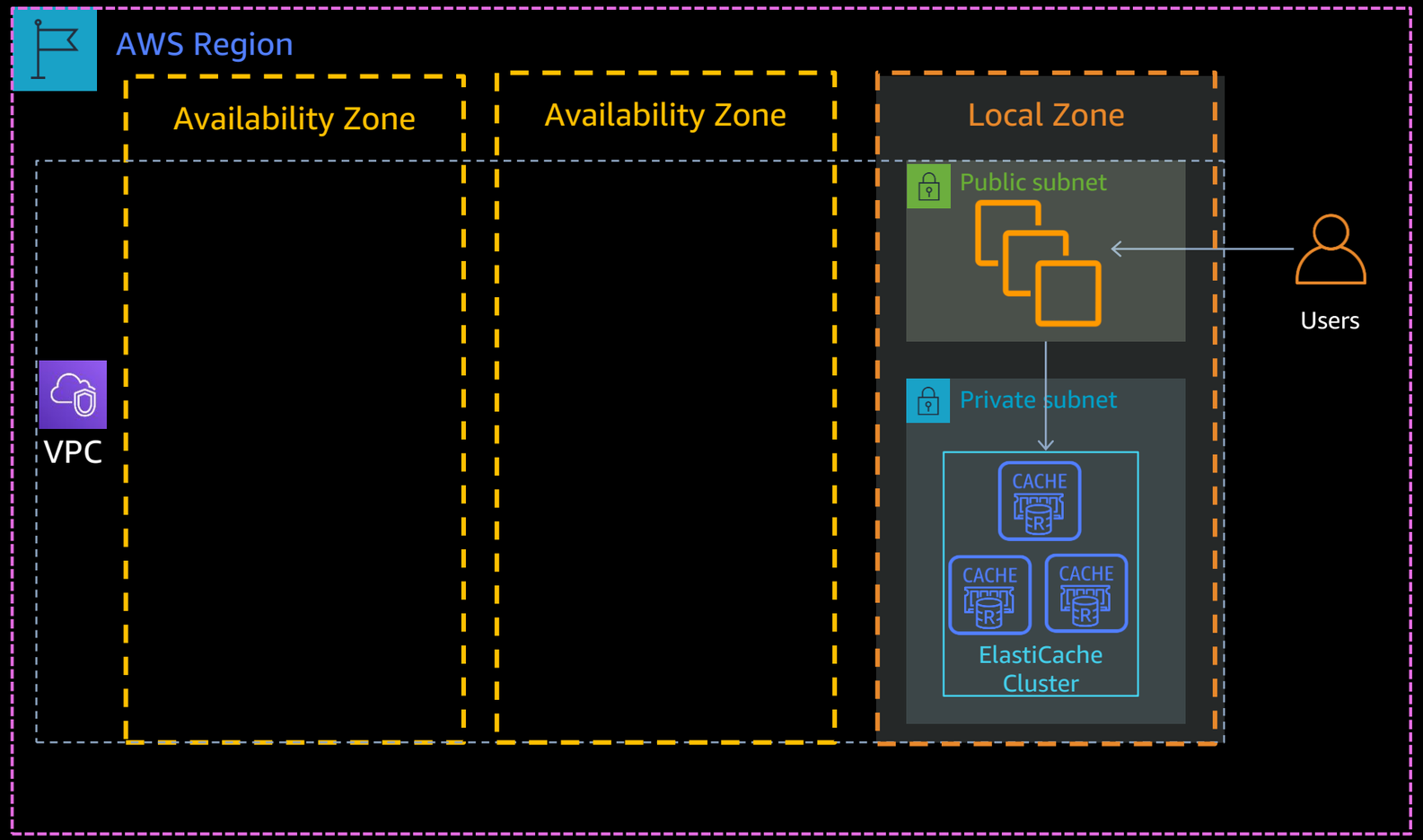
Deliver new low-latency apps

AWS Local Zones use cases

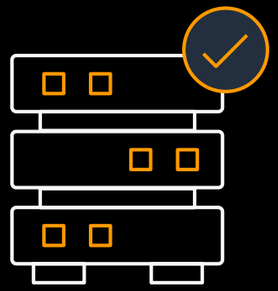
Highly demanding applications that require single-digit millisecond latencies to end users, including:

- Media and entertainment content creation
- Real-time multiplayer gaming
- Electronic design automation
- Machine learning inference

AWS Local Zones



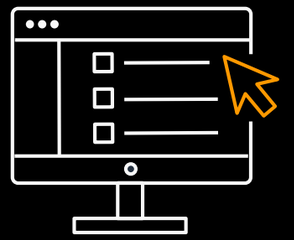
AWS Outposts: Bringing AWS on premises



Same AWS-designed infrastructure as in AWS data centers (built on AWS Nitro System)

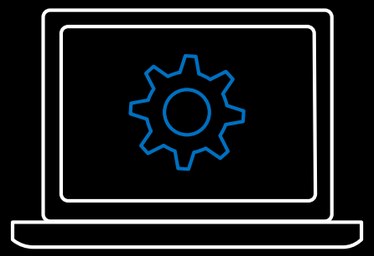


Fully managed, monitored, and operated by AWS as if in AWS Regions



Single pane of management in the cloud providing the same APIs and tools as in AWS Regions

Applications that need to remain on premises

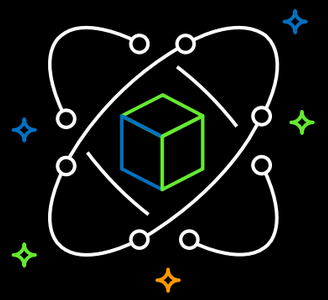


Latency sensitive

Equipment and processes sensitive to compute latency

Interactive workloads such as AR/VR, design and visualization

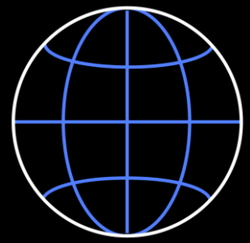
Complex workloads that span a variety of host and storage systems



Local data processing

In hybrid workflows, transcoding, filtering, caching, and alerting applied at the edge

Bringing AWS compute to large datasets that can’t be easily moved



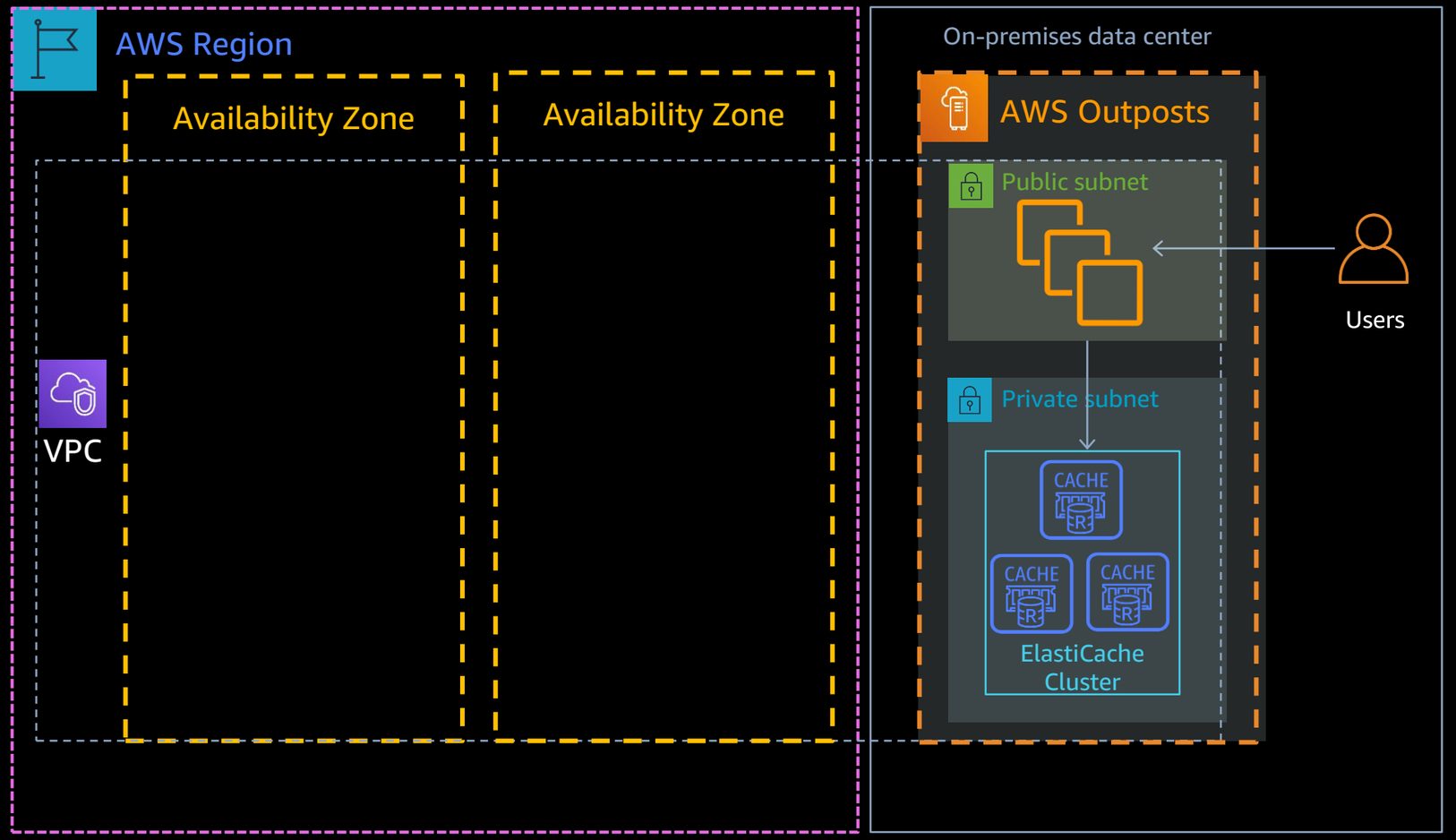
Residency

Where regulations dictate that data and infrastructure reside in specific countries

Where contracts specify where applications are deployed

Where enterprises are not ready to move to AWS Regions for Infosec or other reasons

Run ElastiCache on premises



OSS compatibility



AWS re:Invent

© 2020, Amazon Web Services, Inc. or its affiliates. All rights reserved.

ElastiCache for Redis 6.x support

- Role-based access control (RBAC)
  - Define users with separate authentication and authorization
- Client-side caching
  - Tracks user requests and sends invalidation messages
- Significant operational improvements
  - Improved replication under low-memory conditions
  - Improved expiry algorithm
- Versionless engine support
  - ElastiCache manages your minor version automatically

aws.amazon.com/blogs/aws/new-redis-6-compatibility-for-amazon-elasticache/

Role-based access control

	Redis 5	Redis 6
Users	A "default" user	Supports multiple users
Authentication	Optional auth token	Optional per-user passwords
Authorization	Access to all data by default; can rename commands to hide them	Per-user access to specific keys and commands
Updates	Rotation of auth token and online update of renamed commands	Online update to users (add/remove/rotate password/update access)
Visibility & auditing	No metrics	AuthenticationFailures KeyAuthorizationFailures CommandAuthorizationFailures

Access control – example

```
ec2> redis-cli -h <primary_endpoint> --user readonly --askpass
Please input password: ***********
> get a
"123"
> set a 345
(error) NOPERM this user has no permissions to run the 'set' command or its subcommand
```

```
ec2> redis-cli -h <primary_endpoint> --user marketing --askpass
Please input password: ***********
> set mktg:a 1
OK
> get mktg:a
"1"
> get a
(error) NOPERM this user has no permissions to access one of the keys used as arguments
```

Microservice integration



© 2020, Amazon Web Services, Inc. or its affiliates. All rights reserved.

Support for AWS Controllers for Kubernetes

- Define and use AWS resources directly from Kubernetes
- Unified, seamless way to manage your application and dependencies
- Recently launched ElastiCache in Developer Preview
  - Replication Groups
  - Cache Subnet Groups
  - Snapshots

aws.github.io/aws-controllers-k8s/

AWS Service	Release Status	Controller
Amazon API Gateway V2	DEVELOPER PREVIEW	apigatewayv2
Amazon DynamoDB	DEVELOPER PREVIEW	dynamodb
Amazon ECR	DEVELOPER PREVIEW	ecr
Amazon S3	DEVELOPER PREVIEW	s3
Amazon SQS	BUILD	sqs
Amazon SNS	DEVELOPER PREVIEW	sns
Amazon ElastiCache	DEVELOPER PREVIEW	elasticache

Support for AWS Controllers for Kubernetes

Example policy YAML document

```
apiVersion: elasticache.services.k8s.aws/v1alpha1
kind: ReplicationGroup
metadata:
       name: reinvent
spec:
       engine: redis
       replicationGroupID: reinvent
       replicationGroupDescription: Reinvent replication group
       cacheNodeType: cache.t3.micro
       numNodeGroups: 3
       atRestEncryptionEnabled: true
       transitEncryptionEnabled: true
```

Support for AWS Controllers for Kubernetes

```
ec2> kubectl apply -f ~/Documents/repl.yaml
replicationgroup.elasticache.services.k8s.aws/reinvent created
> kubectl describe replicationgroup
Name: reinvent
Namespace: default
Labels: <none>
Annotations: ...
API Version: elasticache.services.k8s.aws/v1alpha1
Kind: ReplicationGroup
...
Spec:
    At Rest Encryption Enabled: true
    Cache Node Type: cache.t3.micro
    Engine: redis
    Num Node Groups: 3
    Replication Group Description: Reinvent replication group
    Replication Group ID: reinvent
    Transit Encryption Enabled: true
...
```

Scalability and availability



© 2020, Amazon Web Services, Inc. or its affiliates. All rights reserved.

Extreme scale

- Launch clusters with up to 310 TiB of in-memory data and 225M TPS
- 500 nodes within a single cluster
  - Configurations range from 83 shards (1 primary, 5 replicas) to 500 shards (no replicas)
- Online horizontal and vertical scaling for reads and writes
  - Scale out your existing 5.0.6+ cluster
- Same high-availability guarantees and SLA

Global Datastore

- AWS offers 77 Availability Zones in 24 Regions, with more on the way
- Global Datastore provides a fully managed, fast, reliable, and secure cross-region ElastiCache cluster



Global Datastore use cases

Low-latency reads for a global application:

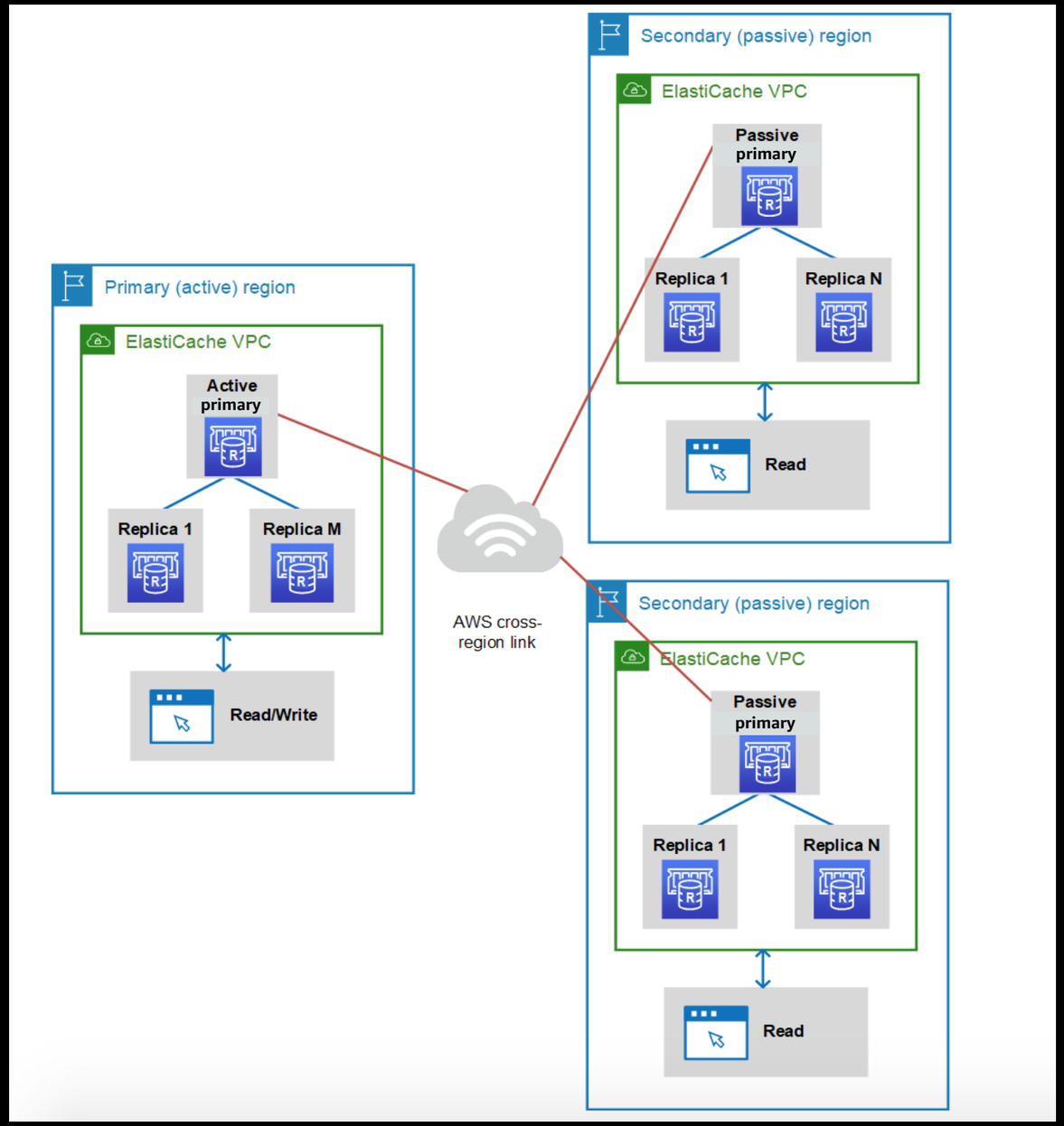
- A globally distributed application can read consistent data within Region
- E.g., a globally distributed leaderboard with fast retrieval

Allows disaster recovery across Regions:

- Recovery in the event of regional degradation to an application
- Secondary clusters can be promoted to primary with full read / write capabilities
- Once initiated, the promotion typically completes in less than 1 minute

Global Datastore design

- Active-passive replication between regions
- Asynchronous replication from primary Region to up to two secondary Regions
  - Replication lag typically under 1 second



Global Datastore scalability

Scaling write capacity:

- Supports both Redis cluster mode (sharded) and non-cluster modes
- Each regional cluster can be up to 500 nodes
- Online horizontal and vertical scaling
- All regional clusters stay identical in node type and number of primary nodes

Scaling read capacity:

- Independent regional scaling of read capacity – add / remove replicas

aws.amazon.com/elasticache/redis/global-datastore/

What’s new in 2020

Performance



Graviton2 instance support

AWS Outposts and Local Zones

Security



Resource-level API permissions

Role-based access control for Redis

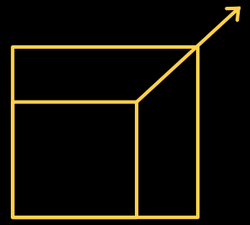
Auto-update of clusters

Microservice integration



AWS Controllers for Kubernetes (ACK)

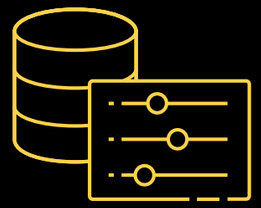
Scalability and availability



Global Datastore

500 node clusters

OSS compatibility



Redis 6 with RBAC, client-side caching

Memcached 1.6

Amazon ElastiCache resources

ElastiCache page

aws.amazon.com/elasticache

ElastiCache documentation

docs.aws.amazon.com/elasticache

Blog

aws.amazon.com/blogs/database/tag/elasticache

Contact us

pages.awscloud.com/ElastiCache-Contact-Us.html

Related session

DAT306

Beyond caching: Advanced design patterns in Redis

AWS re:Invent

© 2020, Amazon Web Services, Inc. or its affiliates. All rights reserved.

Thank you!



AWS re:Invent

© 2020, Amazon Web Services, Inc. or its affiliates. All rights reserved.



Please complete the session survey

AWS re:Invent

© 2020, Amazon Web Services, Inc. or its affiliates. All rights reserved.