
AWS
re:Invent

CMP308

Enable high-performance block storage for enterprise applications

Venky Nagapudi

Principal PM, Elastic Block Store
AWS

Prudhvi Karumanchi,

Senior Software Engineer
Netflix

AWS re:Invent

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

Agenda

Running enterprise workloads on Amazon EBS: Introduction

Scaling to high single-instance performance with EBS

Achieving reliability for high-performance workloads

Clustering multiple instances for high performance

Cache warming using EBS Multi-Attach at Netflix



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

Running enterprise workloads on EBS: Introduction



AWS re:Invent

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

Enterprise workloads on Amazon EBS



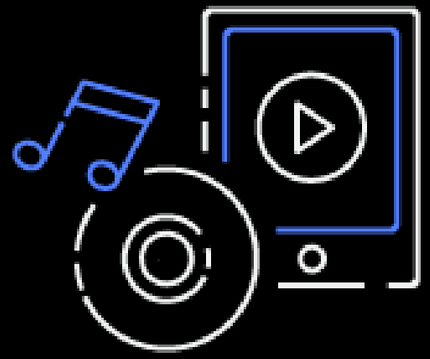
Databases

Oracle DB, SQL Server,
SAP HANA, MongoDB



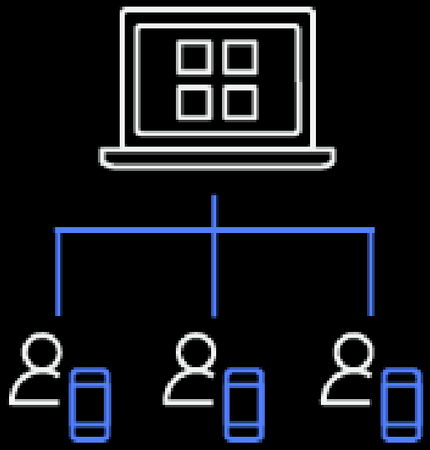
Data & analytics

Kafka, Splunk, Hadoop,
data warehousing



Media

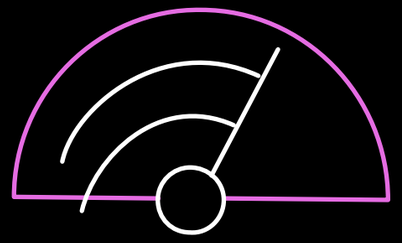
Transcoding, encoding,
render farms



File

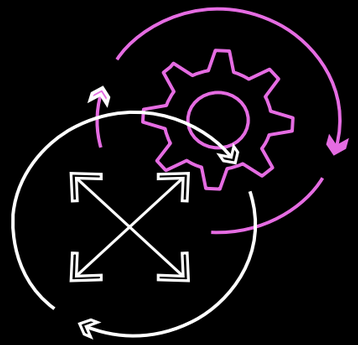
CIFS, NFS
archive

Storage requirements for enterprise workloads



High performance

Several 100K IOPS
Multiple GBs of bandwidth
Consistent low latencies



Scalability

Scale compute
Scale IOPS and bandwidth
Scale capacity

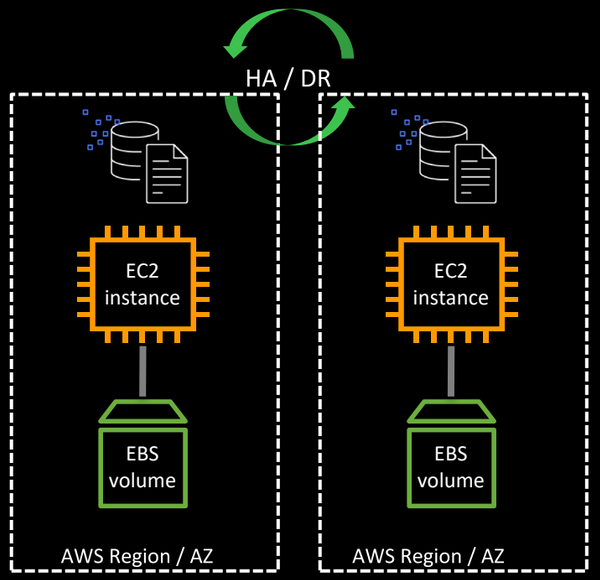


Reliability

Backup & restore
High availability (HA)
Disaster recovery (DR)

Running enterprise workloads on EBS

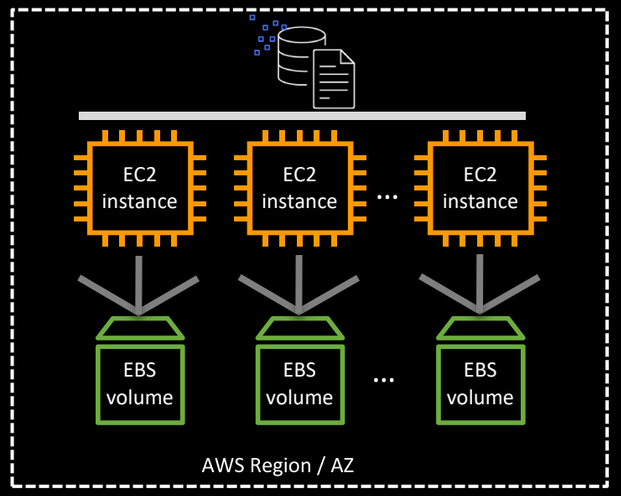
Single-instance applications



1. Optimize single-instance performance

2. Build in reliability with backup, HA, and DR

Clustered applications



3. Optimize multi-instance performance with built-in reliability

EBS enables high performance single-instance and clustered applications

Scaling to high single-instance performance with EBS

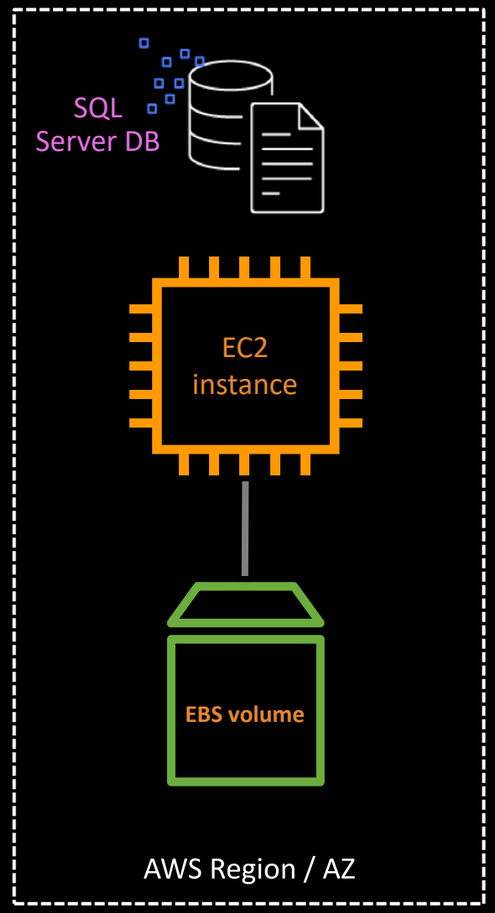


AWS re:Invent

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

Customer scenario: SQL Server DB example

Goal: I want to achieve 250K IOPS & 7.5GB/s

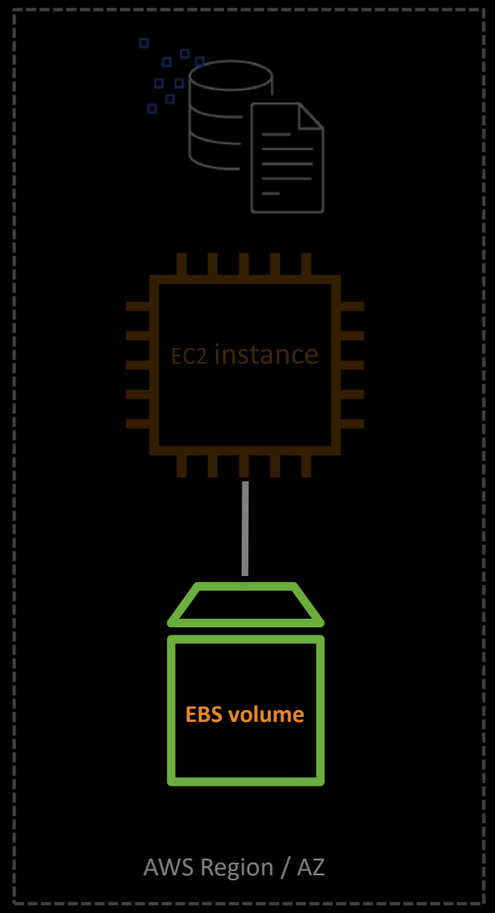


Considerations:

- What and how many volumes should I use?
- What instance should I use?
- How do I scale capacity and performance?
- What is the recommended solution?
- How do I save on license costs?

Use SSD volumes and EBS optimized EC2 instances for high performance

What volumes to use : General purpose SSD

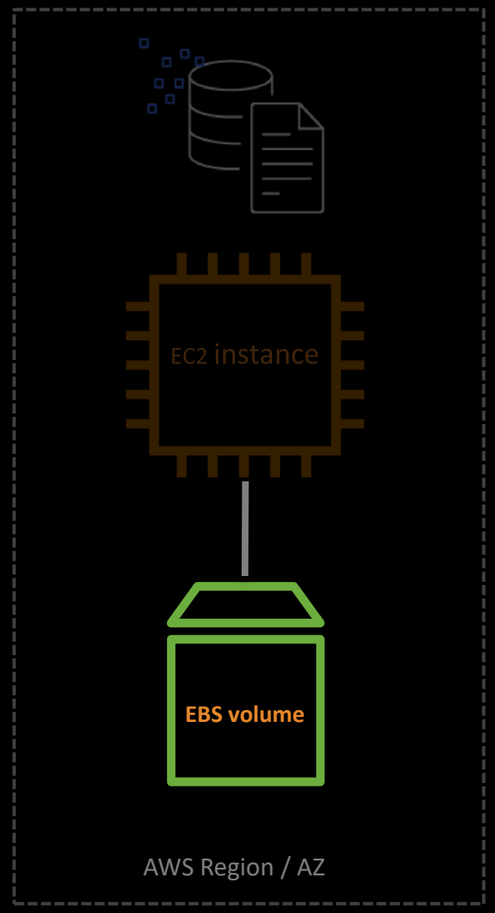


gp2 General purpose SSD

- **Baseline: 100 to 16,000 IOPS; 3 IOPS per GiB**
- **Burst: 3,000 IOPS (for volumes up to 1,000 GiB)**
- **Throughput: Up to 250 MiB/s**
- **Latency: Single-digit ms**
- **Capacity: 1 GiB to 16 TiB**

Great for boot volumes, low-latency applications, and bursty databases

What volumes to use : Provisioned IOPS



io1/io2 NEW Provisioned IOPS

Baseline: 100 – 64,000 IOPS

Throughput: Up to 1,000 MiB/s

Latency: Single-digit ms

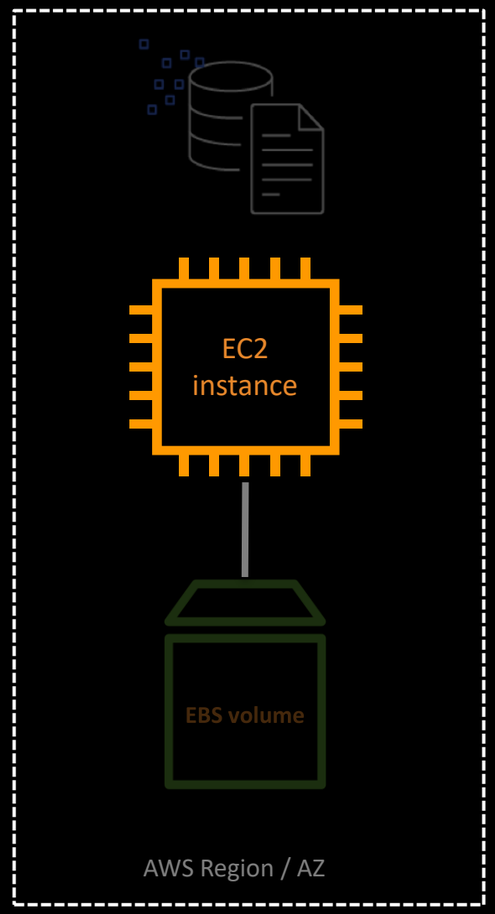
Capacity: 4 GiB to 16 TiB

99.999 Durability and availability,

500:1 IOPS: GB ratio with io2 at the same price at io1

Ideal for critical applications and databases with sustained IOPS

What instances to use: R5/R5b for databases

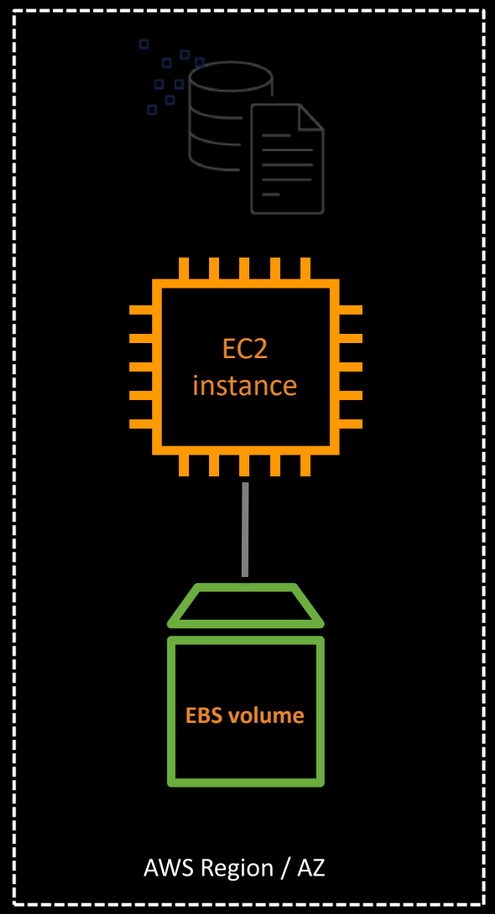


R5/R5b EBS optimized EC2 instances

- R5/R5n: EBS bandwidth (2,375 MB/sec) and 80K IOPS
- New R5b: 3x higher EBS bandwidth (7,500MB/sec) and 260K IOPS
  - Supports io1 now; io2 support coming soon
- Dedicated network bandwidth for Amazon EBS I/O
- Updated network stack, Scalable Reliable Datagram (SRD)

Use R5b for I/O heavy database use cases

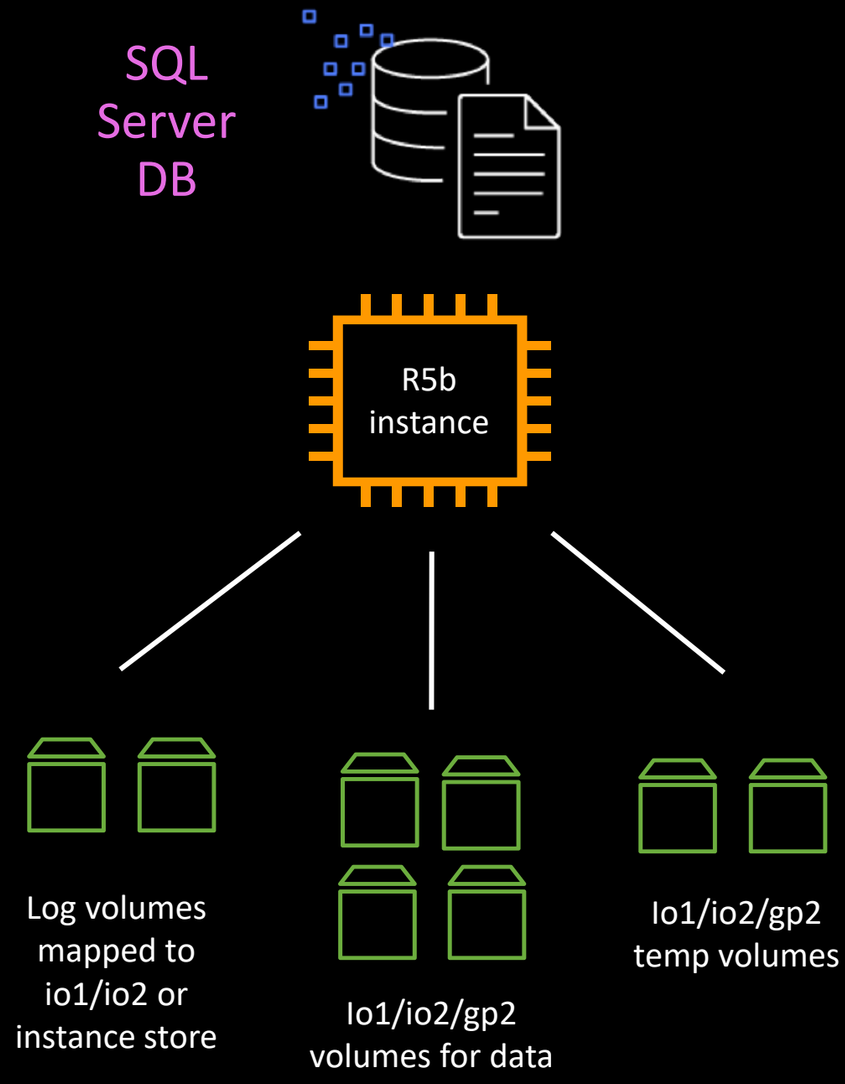
Scaling capacity and performance



	Compute scaling				Storage scaling					
Example config	**R5**		**R5B**		**gp2**			**Io1/io2**		
	vCPUs	Max BW (GB/s)	vCPUs	Max BW (GB/s)	# Vol	Total max IOPS	Total BW (GB/s)	# Vol	Total max IOPS	Total BW (GB/s)
40K IOPS (mini)	12	1.15	4	1.25	5	80K	1.25	2	128K	2
80K IOPS (small)	24	2.3	8	2.5	10	160K	2.5	3	192K	3
160K IOPS (medium)	-	-	16	5	20	320K	5	5	320K	5
240K IOPS (large)	-	-	24	7.5	30	480K	7.5	8	512K	8

- With R5b, you can use **3x fewer cores** for same IOPS
- Use R5b when you need more than 80K IOPS/instance
- Use Elastic Volumes (EV) to dial in the IOPS for io1/io2
- Creating striped io2 volumes results in the best durability

High-performance single instance SQL Server



Use high-performance instance

- R5b supports up to 96 vCPUs
- 260K IOPS & 7.5GB/s of EBS performance
- Use io1/gp2 (io2 support coming soon)

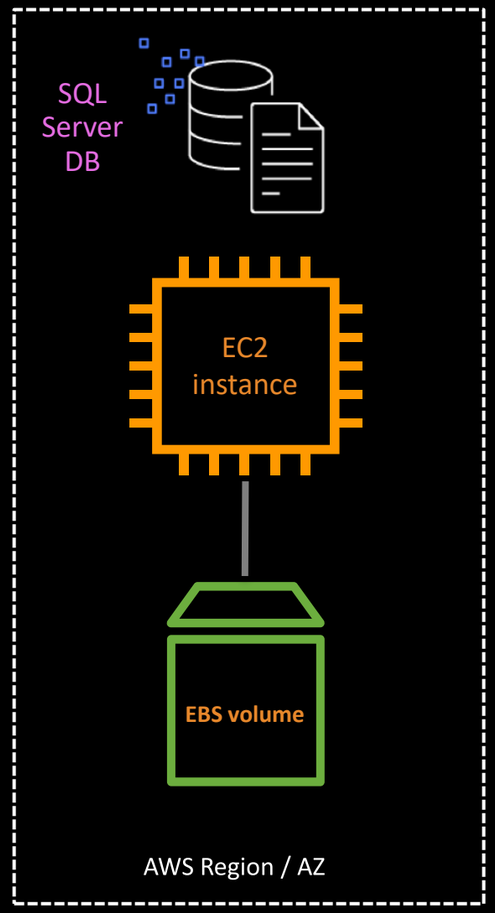
Use multiple io1/io2 or gp2 data volumes

- Smaller stripe with io2 volumes
- Higher durability with io2 (99.999%)

Use low-latency volumes for logs

- Either io1/io2 or instance store

R5b vs. R5 performance benchmarks



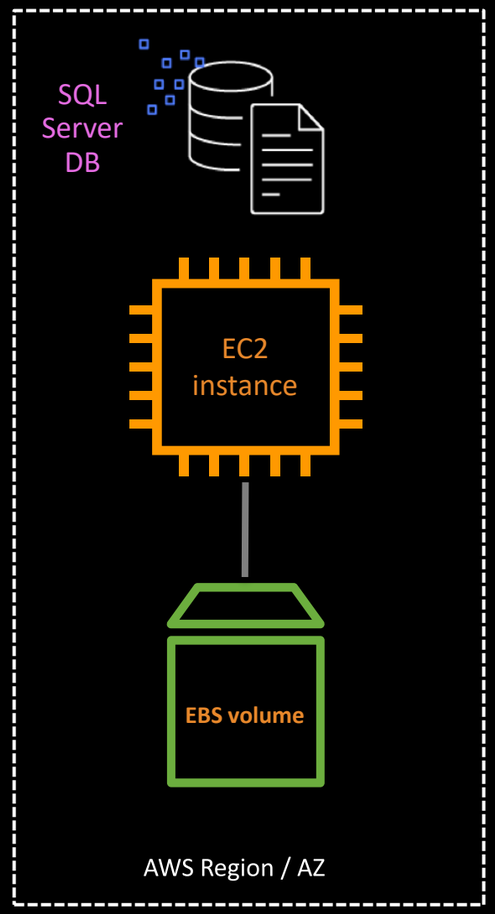
- TPC-C benchmarks run using R5 and R5b
  - Comparing R5.24xl and R5b.24xl
  - Io1 volumes with 40K IOPS used
  - NOPM test with 30,000 warehouses

IOPS	Instance	IO1 volumes	Avg TPC-C NOPM
80K	R5.24xl	4 @ 20K IOPS	271K
120K	R5b.24xl	4 @ 30K IOPS	531K
200K	R5b.24xl	4 @ 50K IOPS	773K
256K	R5b.24xl	4 @ 64K IOPS	871K

- Scale IOPS with Elastic Volumes

R5b achieves 3.2x higher TPC-C performance than R5

Key takeaways



Use R5b and SSD volumes for high-performance SQL Server

Scale to high performance (>200K IOPS) with 99.999% durability using io2 stripes

Achieve **>3x higher** SQL Server performance with R5b vs. R5

Save **up to 66%** on SQL Server license costs IO- heavy use cases with R5b vs R5

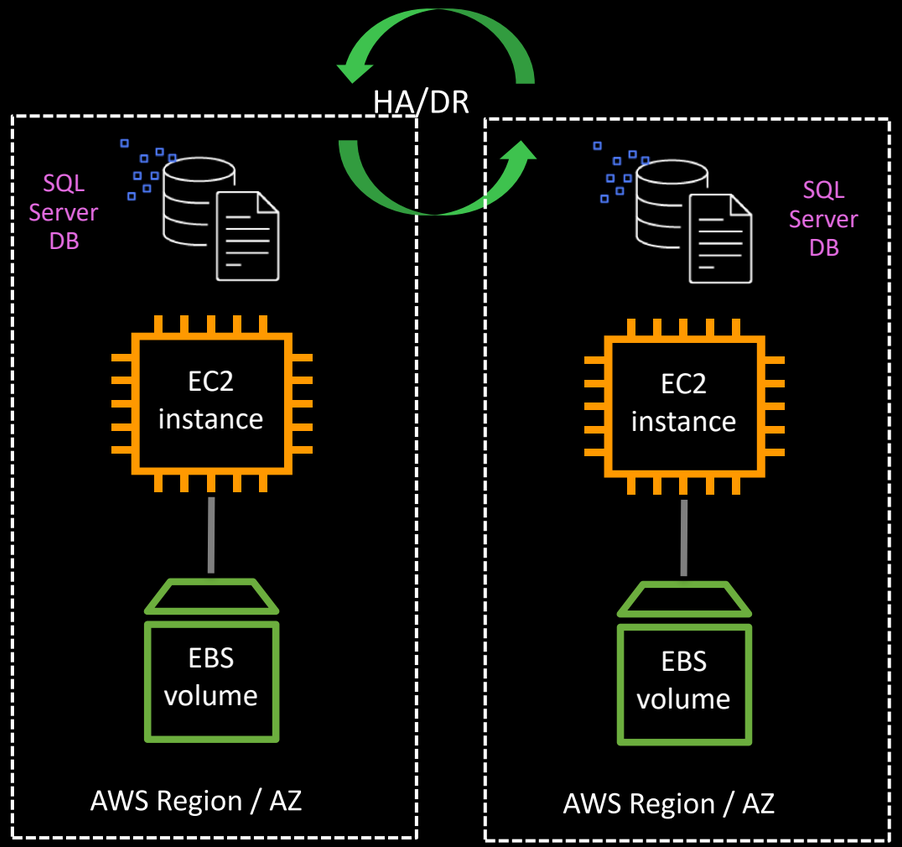
Achieving reliability for high-performance workloads



AWS re:Invent

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

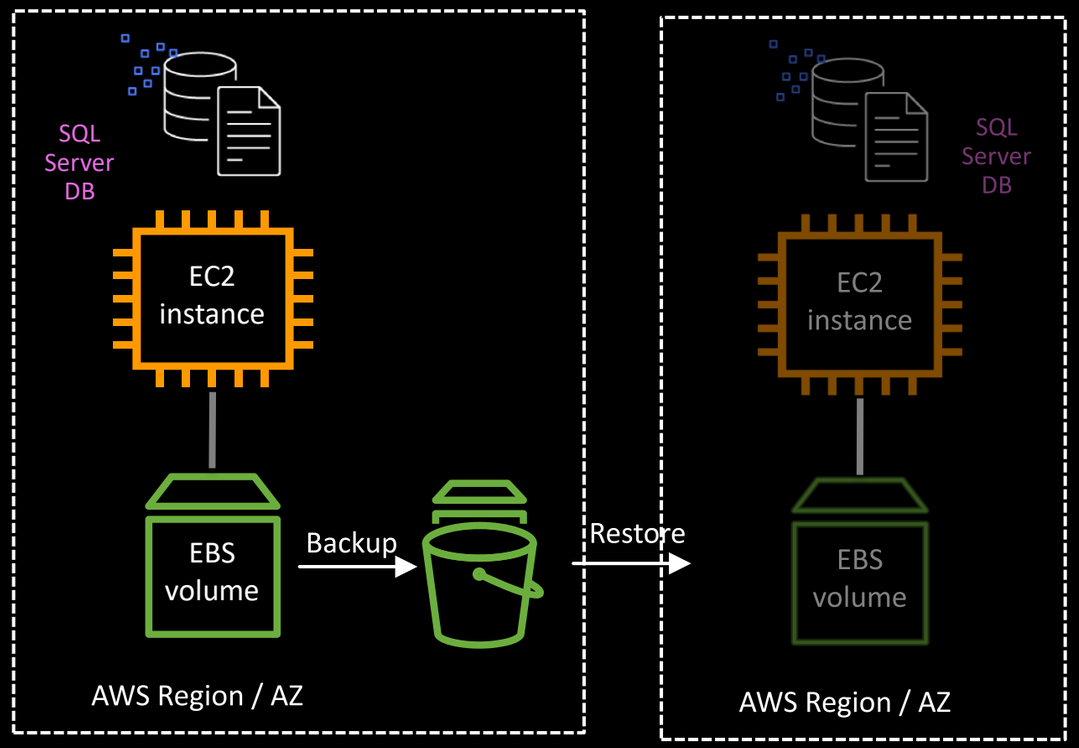
Customer scenario: SQL Server DB example



Considerations:

- How do I perform app-consistent backup & restores?
- How do I achieve HA/DR?

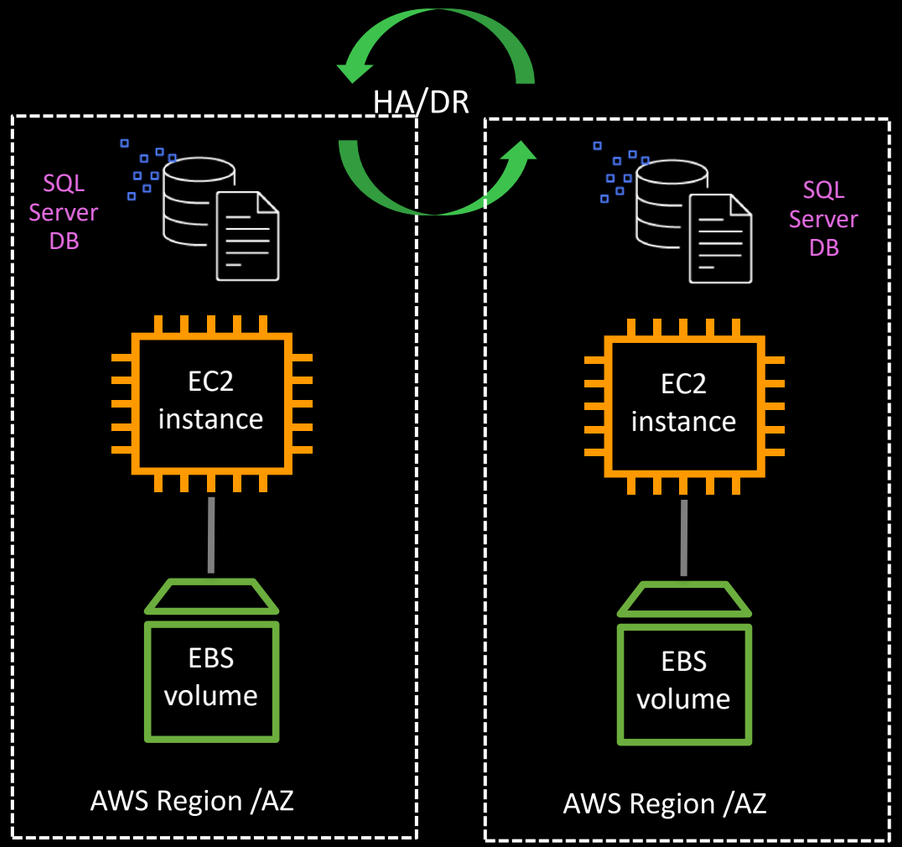
Backup & restore with EBS snapshots



- App-consistent backup with VSS integration
- Restore only when needed to save costs
- Restore to different AZ or Region
- RPO and RTO in hours
- Fast snapshot restore (1TB/hr)

Achieve application-consistent backups with EBS and VSS

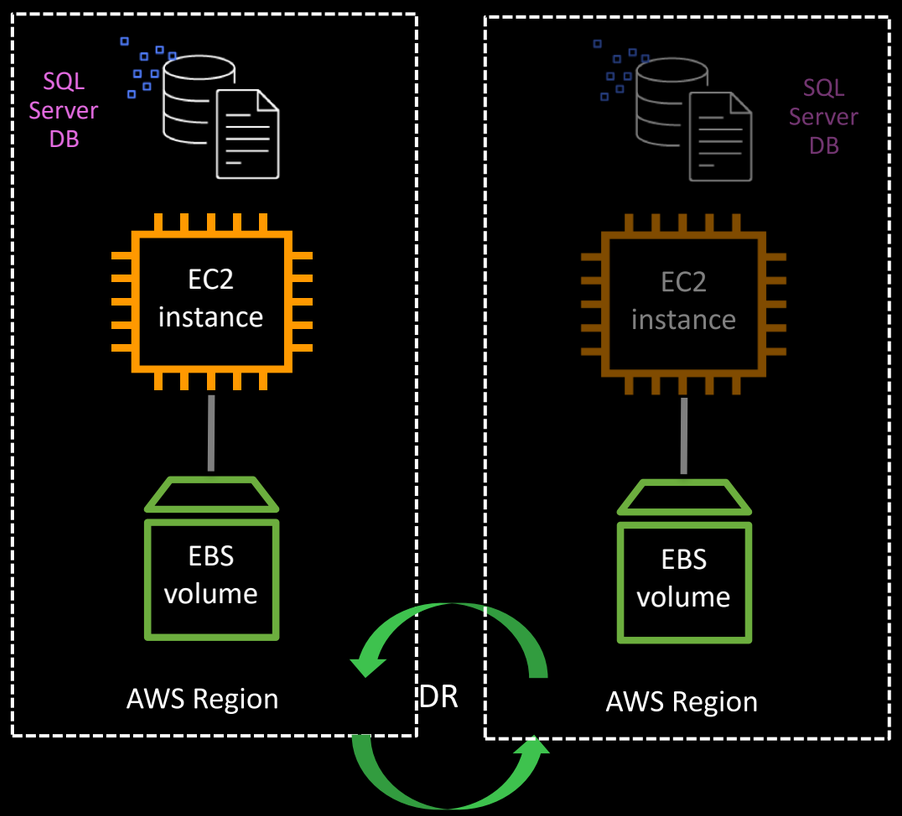
HA/DR with built-in SQL Server options



- Use SQL Server Always On availability groups (AG)
- 2 nodes with standard license
- More nodes need enterprise license
- Needs EC2 instances and EBS volumes to be always available
- Close to zero RPO & RTO

Use popular SQL Server AG with EBS volumes

DR with CloudEndure AWS solution



- CloudEndure uses EBS snapshots and volumes
- Sub-second RPOs, RTO in minutes
- Use instances and licenses only on disaster

Use CloudEndure to optimize DR costs for SQL Server DB

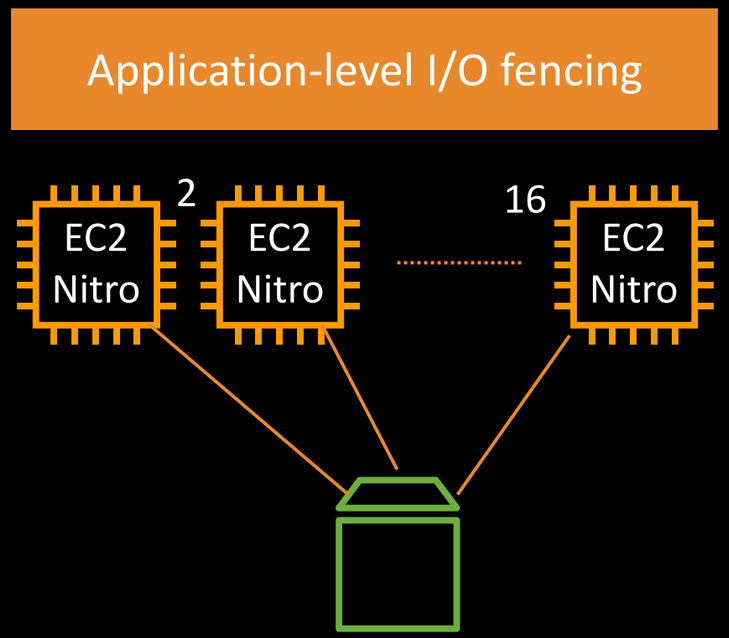
Clustering multiple instances for high performance



AWS re:Invent

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

Sharing EBS volumes with Multi-Attach



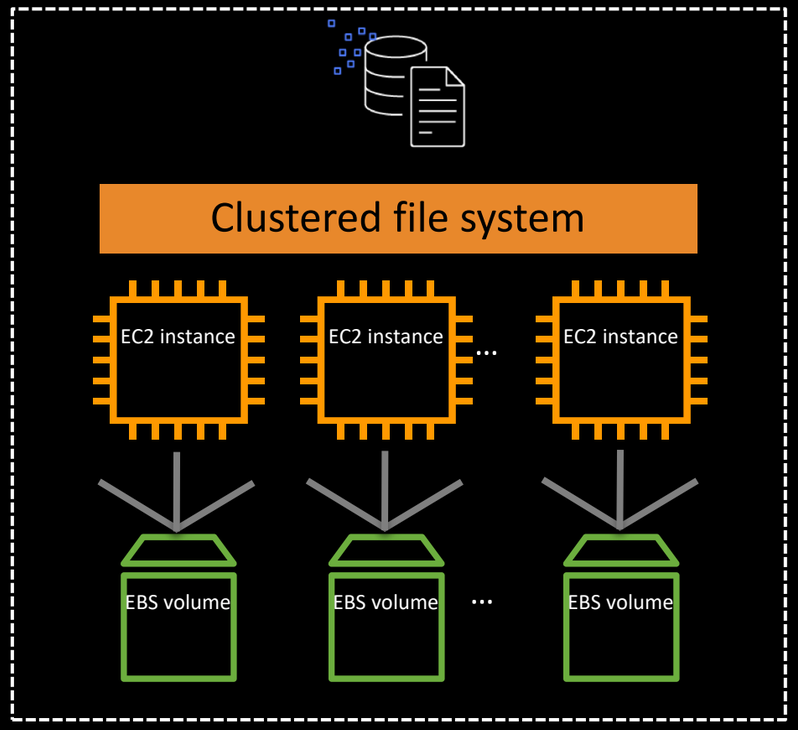
Enables io1 volume shared by up to 16 EC2 instances in the same AZ

Provides application HA for applications that manage writes

Note: Applications have to perform I/O fencing

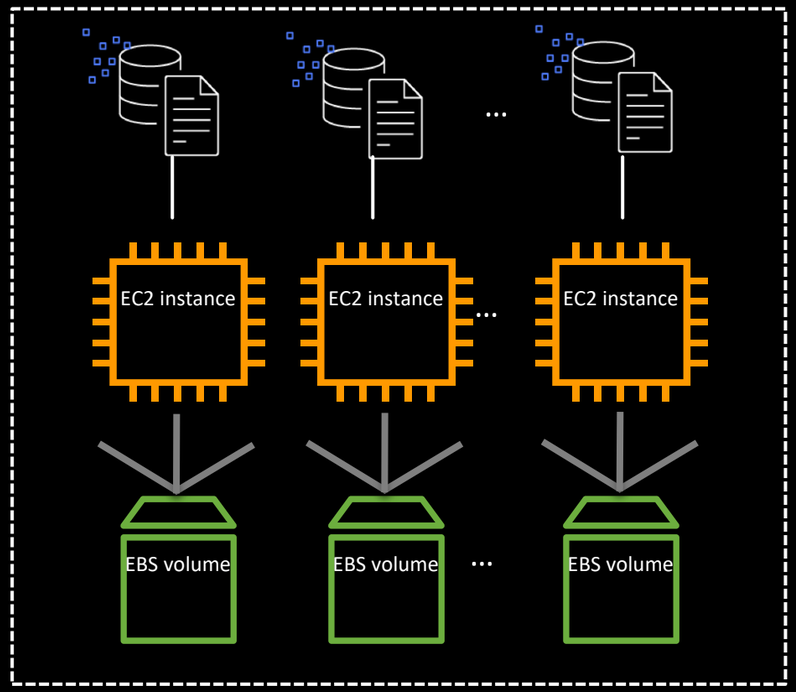
Clustered applications on EBS

Application running on clustered file system



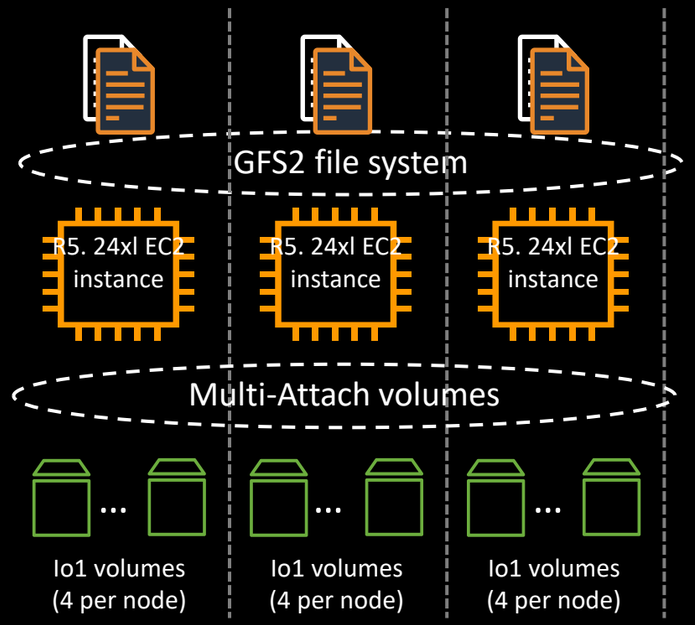
File system implements I/O fencing

Application running without clustered file system



Application implements I/O fencing

Clustered GFS2 Filesystem with Multi-Attach



GFS2 3-node cluster with Multi-Attach

Enables highly available shared file systems

Resilient to node failures

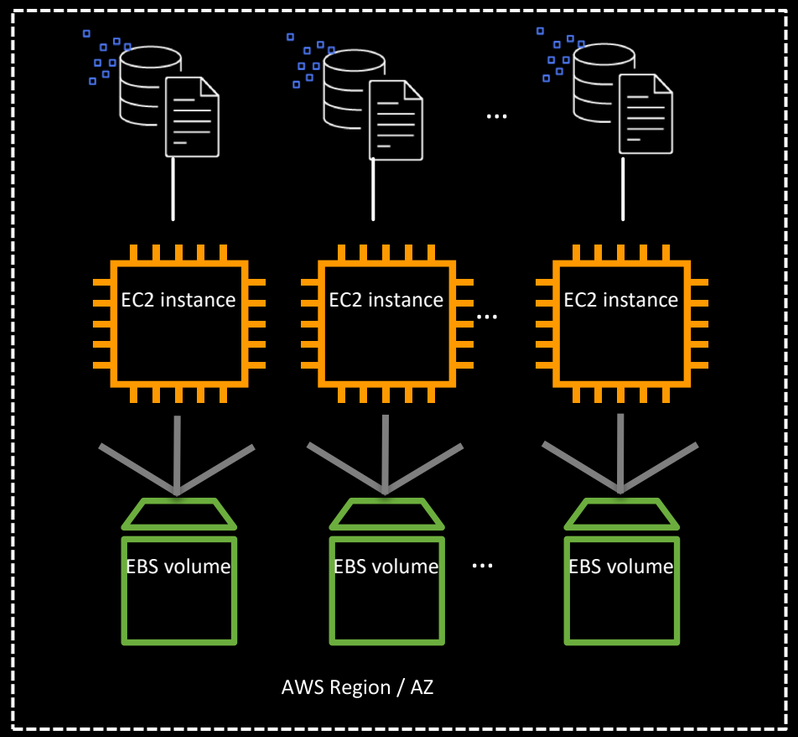
Achieves high cluster performance*

- 80K IOPS/node for random 16KB I/O for 100R and 60R-40W mix
- 2.3GB/s per node for 1MB I/O for sequential reads and writes

*Measured performance with 3 R5.24xl and 4 io1 volumes / instance

Multi-Attach enables high-performance clustered file systems

Custom use cases with Multi-Attach



Memcached cache warming with Multi-Attach on io1 volumes

Use Multi-Attach to accelerate performance of custom applications

Cache warming using EBS Multi-Attach at Netflix



AWS re:Invent

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



Clustered Memcached optimized for AWS and tuned for Netflix use cases

Distributed Memcached	Tunable replication	Highly resilient
Topology aware	Highly scalable	Dynamic

EVCache

Two variants of memcached

- In-memory
- External storage*

* https://netflixtechblog.com/application-data-caching-using-ssds-5bf25df851ef

What services touch EVCache?

Sign up*

Personalization*

Playback*

Searching*

Viewing history

Algorithm & ML training and many more

* Multiple caches involved

EVCache footprint at Netflix

Petabytes of data

Trillions of items stored

Hundreds of millions of ops/sec

Hundreds of distinct clusters

Tens of thousands of instances

3 regions

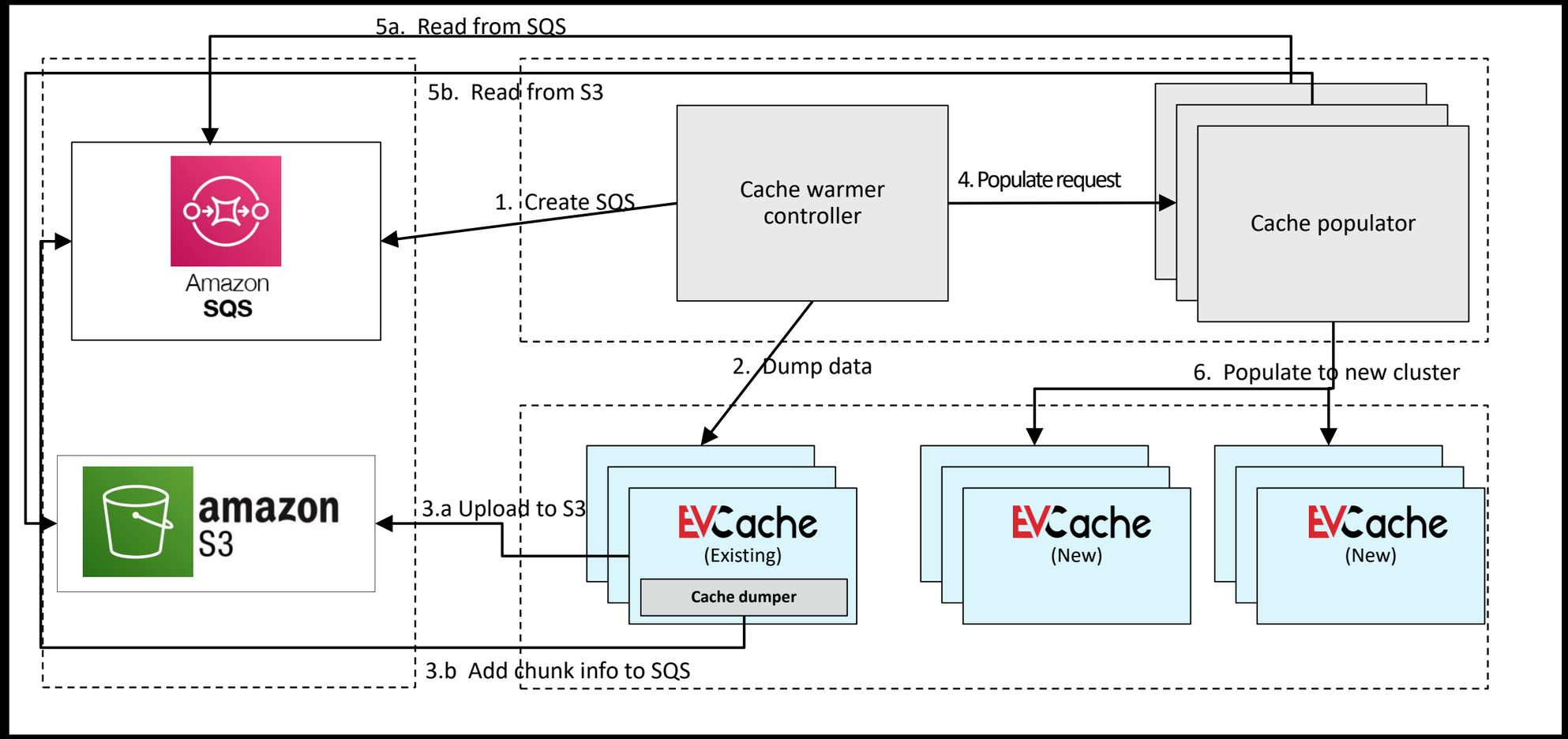
Why is cache warming needed

- New deployments are quite common
  - Increase / decrease capacity
  - Memcached upgrade (fixes / new features)
  - Base AMI upgrades
  - Instance type changes
- New deployments cannot take read traffic right away
  - Wait until item TTL duration
  - TTL can range from few hours to few weeks

Design goals of cache warmer

- Limit network impact on current EVCache clients
- Minimize the memory and disk usage on EVCache nodes
- Shorten the warm-up time
- Have no restrictions on when (peak / non-peak periods) to warm up

Old cache warmer architecture



More context: https://netflixtechblog.com/cache-warming-agility-for-a-stateful-service-2d3b1da82642

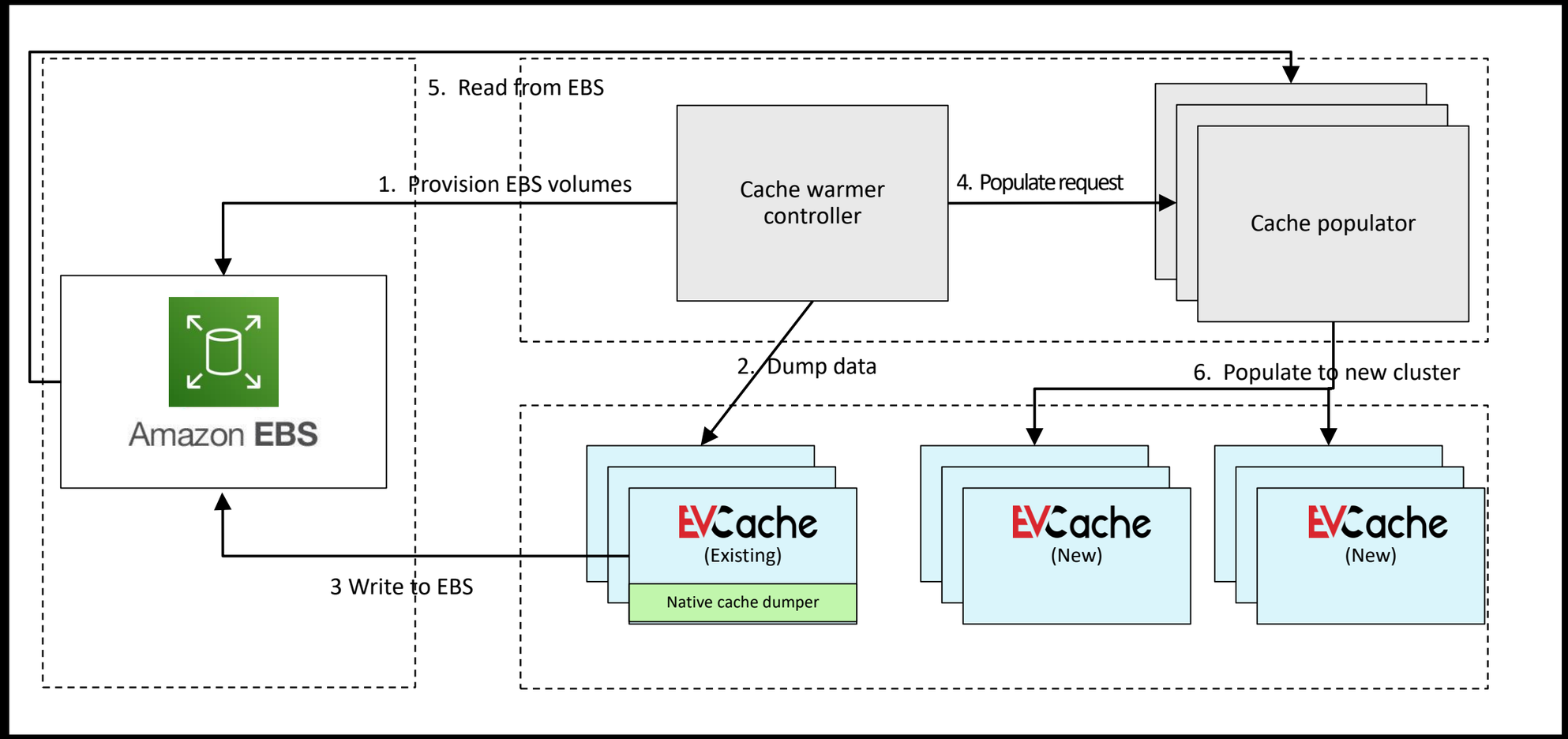
Limitations of current warmer architecture

- Impact on existing production traffic
- S3 network bandwidth
- JVM crash on OutofMemory errors

How EBS Multi-Attach fixed these limitations

- Dedicated network
- No disruption to applications
- EBS Multi-Attach enables pipelining reads and writes

New cache warmer architecture



Hmm, how does this work?

What file system is used

- xfs

Source side

- mount with wsync option
- fsync() on files, directories

Destination side (in a loop)

- mount with ro,norecovery

If not, how do you check for the integrity of the data?

Checksum at source and validate at the destination

Performance and quality of warmup data

Time to warm up

	Data size: 1000 TB, 69 billion keys
Existing architecture	> 40 hours
New architecture	6–9 hours

Are all the limitations addressed?

Note: The data size is cumulative for a cluster and not per node data. Cumulative IOPs provisioned – 1million IOPs

Acknowledgements

- Thanks to all members of EVCache team at Netflix, especially Sailesh Mukil, Shashi Madappa, and Tharanga Gamaethige, for their contributions
- Thanks to Nimisha Ranganath and Marc Olson from AWS

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.