https://axonops.com

axonops

Storage considerations when running Apache Cassandra on Kubernetes

Hayato Shimizu
Co-founder

Sergio Rua
Principal DevOps

kubernetes

cassandra

Introduction

Microsoft Azure

aws

Google Cloud

Introduction

Why run Apache Cassandra on Kubernetes?

Kubernetes is now ubiquitous

Record number of organizations are using or evaluating Kubernetes as the technology goes mainstream and users start to move up the stack - CNCF 2022

Auto-healing and fault tolerance

Some of the advantages of K8s are that applications will recover from most failures such as a node failure.

Homogenous lifecycle management

Developers can deploy immutable Cassandra images using the same tools as the applications, for example, GitOps provisioning models.

Very quick provisioning and decommissioning

- Quick provisioning
- Containerised deployments are fast
- Immutable configuration

Because it's cool!

Cassandra on Kubernetes

Increased interest in running databases on K8s

The *Data On Kubernetes* community was strongly featured at the latest **KubeCon** 2022

Deployment Model

- **Simple Cluster Helm** - no operational management
- **Operator pattern** - essential for managing Cassandra on Kubernetes

Several Cassandra Operators available

- K8ssandra (https://k8ssandra.io/)
- CassKop (https://github.com/cscetbon/casskop)
- Cassandra Operator by Sky UK
- Instaclustr (sunset)

Challenges of running Apache Cassandra on Kubernetes

Kubernetes was not intended for stateful distributed systems

Kubernetes was designed for running *microservices*. *StatefulSets* were added later on. Dynamic IPs, heavily DNS based, designed to be auto-scaled.

Fluidity of pod execution hosts

- Worker node failure
- Kubernetes upgrades
- What storage should I use?

Ingress solutions are difficult to set up

If the clients are outside of Kubernetes, connecting to the cluster can be challenging.

- BGP
- HostPort
- NodePort
- Load Balancer
- Ingress (SNI)

Stargate is a great solution!

Cassandra Storage
in
Kubernetes

Cassandra Storage Requirements

High Throughput

- CommitLog
- Memtable flushes
- Compaction
- Anti-entropy

Low Latency

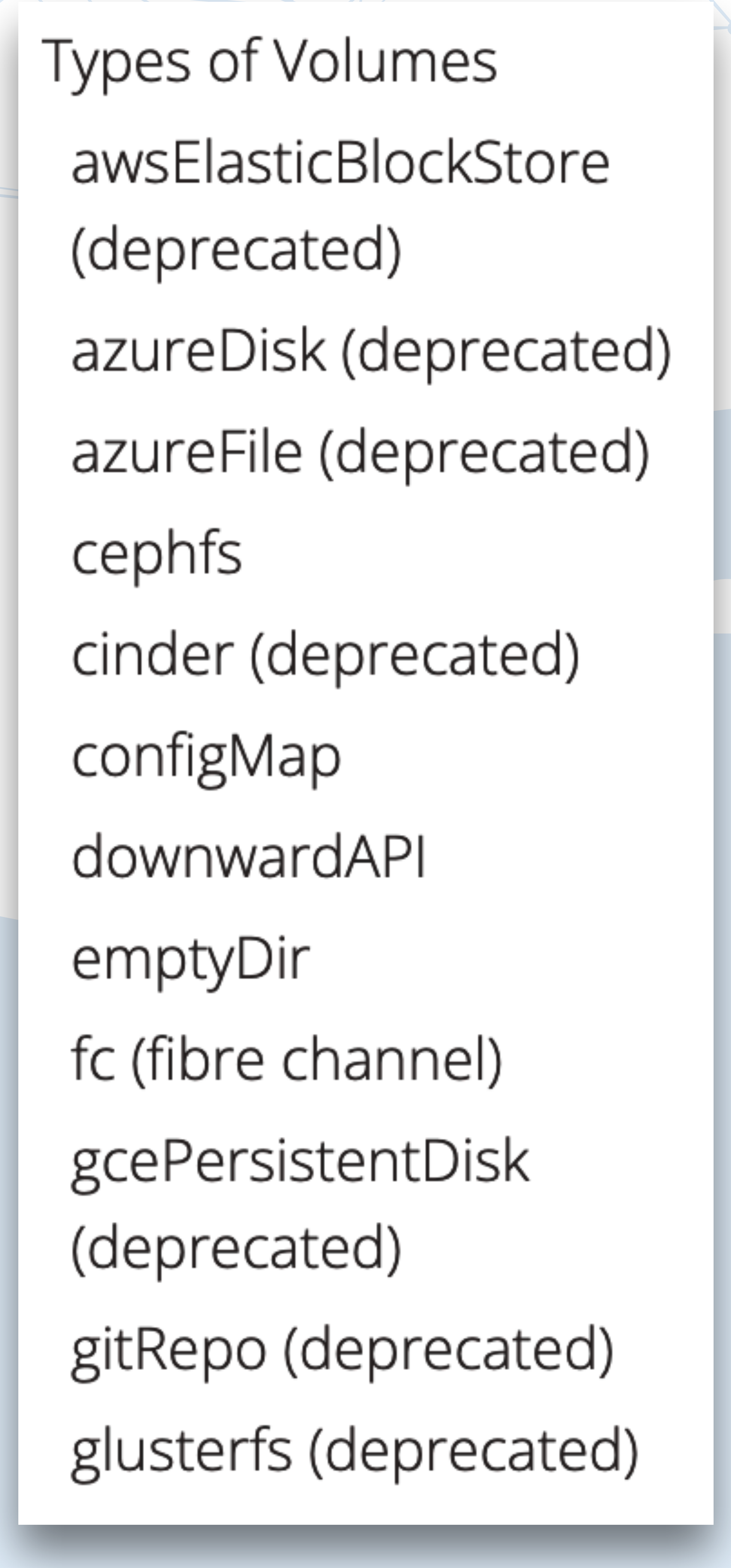
- Queries

High IOPS

- Queries
- Compaction

Kubernetes Storage: types

- **Local disk**
  - Local ephemeral filesystem
  - Distributed local block storage
- **Remote storage**
  - Public cloud remote storage - EBS etc
  - iSCSI
  - NFS
  - Longhorn
  - OpenEBS



Types of Volumes

- awsElasticBlockStore (deprecated)
- azureDisk (deprecated)
- azureFile (deprecated)
- cephfs
- cinder (deprecated)
- configMap
- downwardAPI
- emptyDir
- fc (fibre channel)
- gcePersistentDisk (deprecated)
- gitRepo (deprecated)
- glusterfs (deprecated)

Kubernetes Storage

ROOK

Using local disks on managed Kubernetes is challenging

- Patching and Upgrading
- Pod management

Many other storage providers also supported

There are many more providers supported, some *in-tree*

and others by installing additional CSI drivers (*out-of-tree*).

Avoid DIY distributed storage

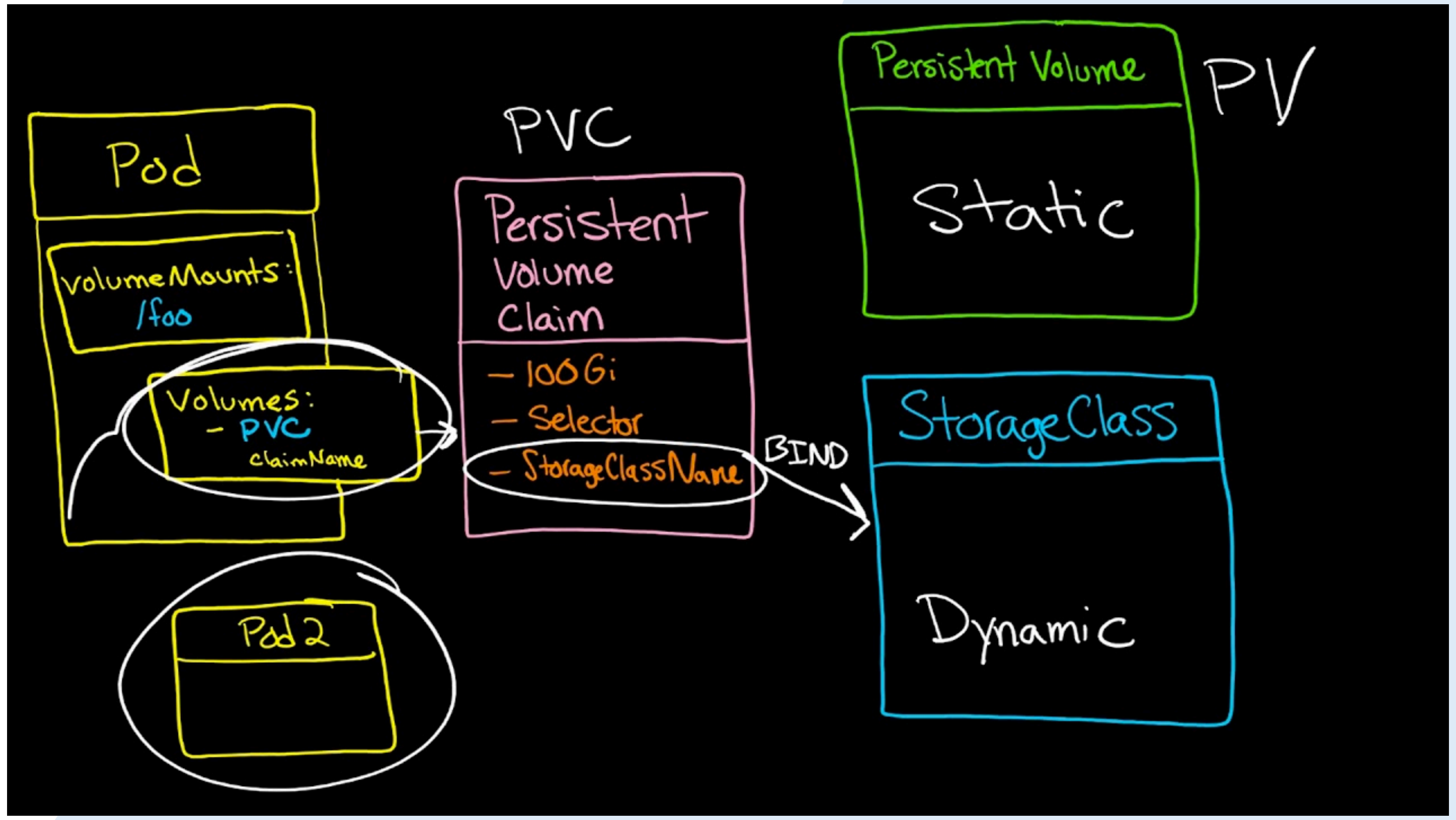
- YADSTM - increases complexity
- High resource requirements
- Replicating replicated data

OpenEBS

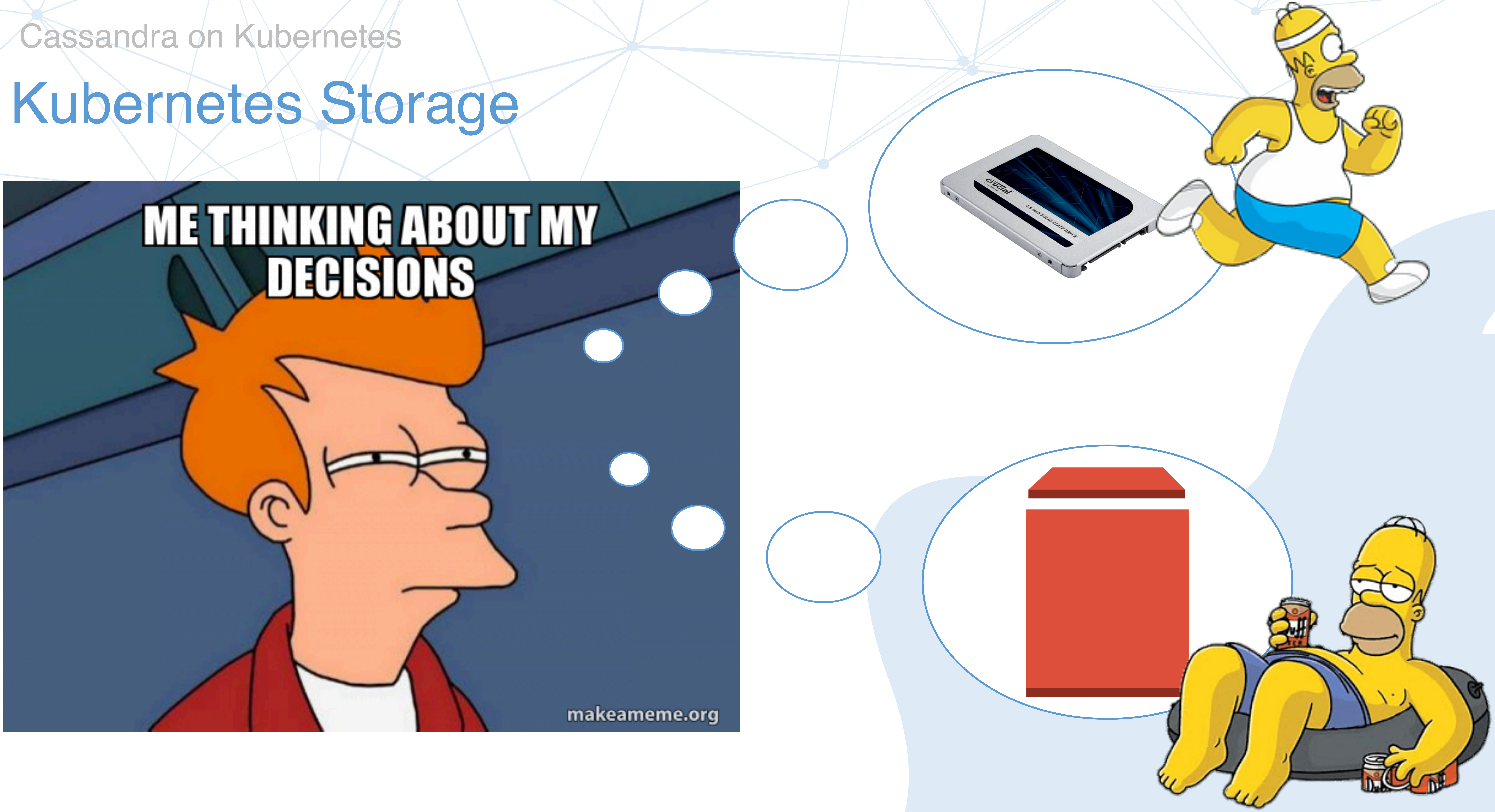
LONGHORN

AWS/Google/Azure are supported in-tree

Cloud Remote Storage provisioners for the main providers are supported as part of the main Kubernetes distribution. This is the old model and it is no longer recommended.



Kubernetes Storage



Storage Classes: configuration

Defines storage driver to use for provisioning

Both local and remote, defines what storage to provision and assign to pods.

Use the Container Storage Interface (CSI) provisioner

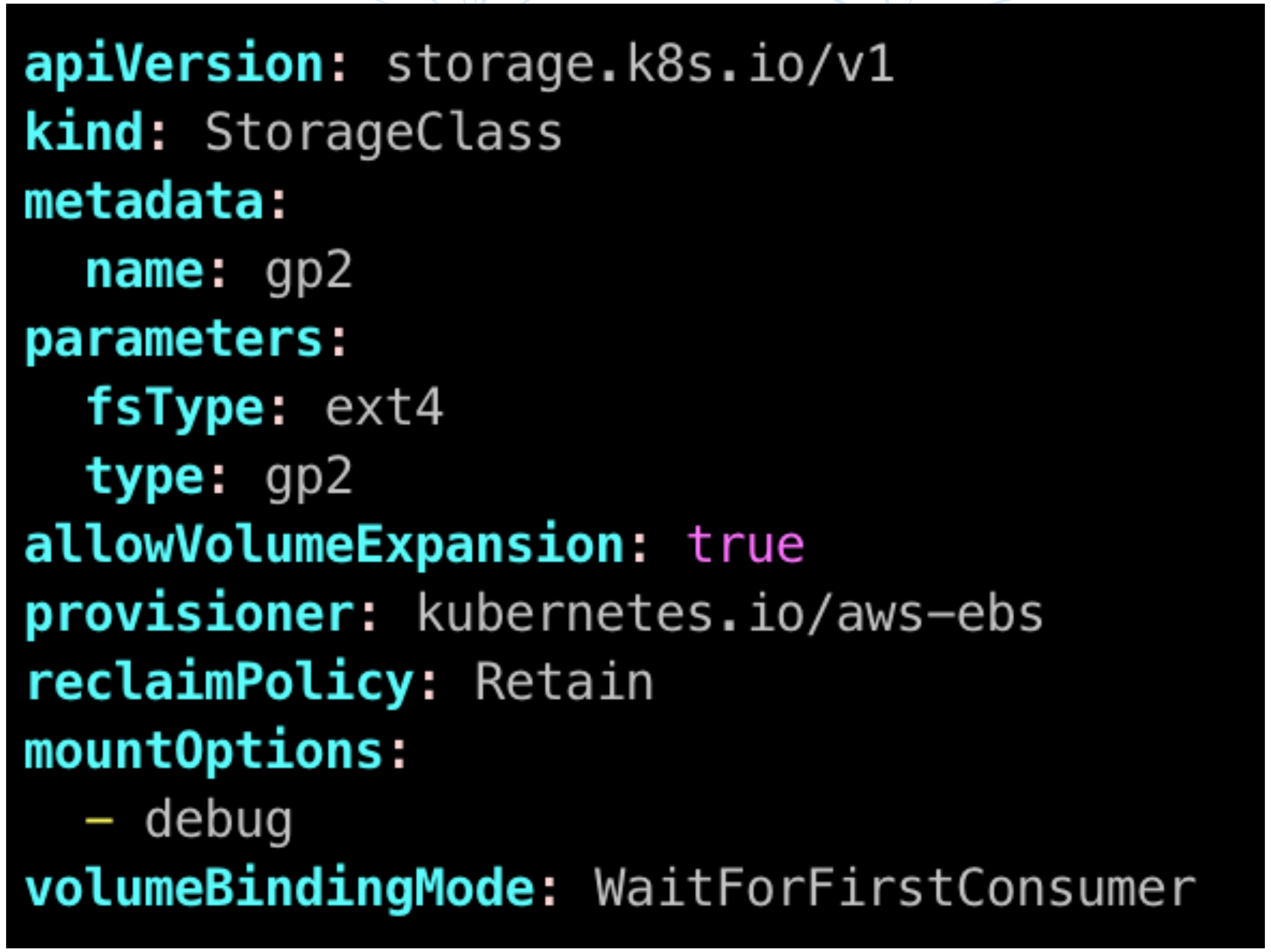
More up-to-date than *in-tree* drivers, supported by the cloud platforms and with more fine grained options available.

Watch out for default options

The default storage type is most likely unsuitable for Cassandra. Also, most cloud providers default to **Delete** the storage when the pod is terminated. Hint: ***ReclaimPolicy=Retain***

Storage Classes

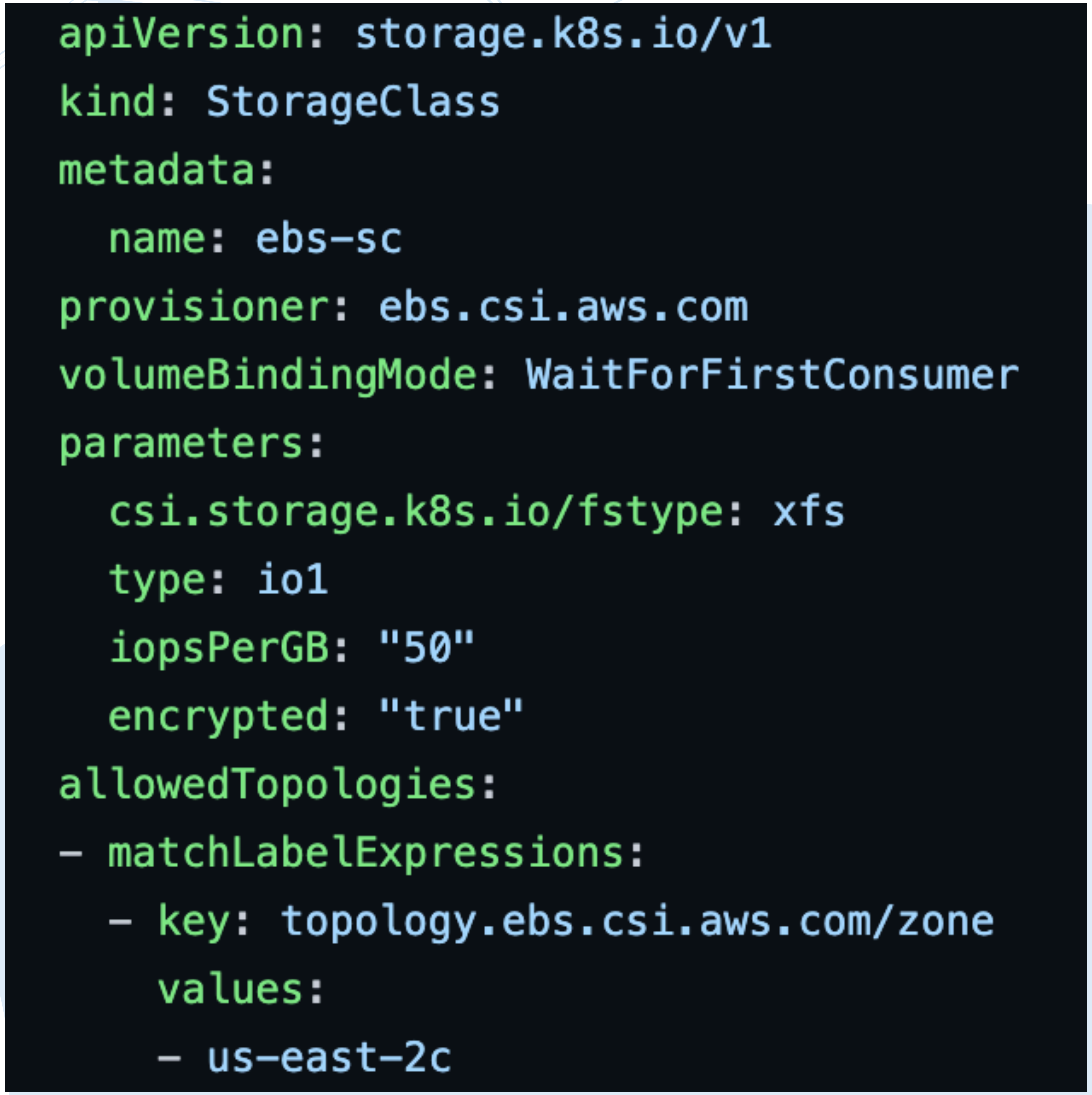
Limited Options



```yaml
apiVersion: storage.k8s.io/v1
kind: StorageClass
metadata:
  name: gp2
parameters:
  fsType: ext4
  type: gp2
allowVolumeExpansion: true
provisioner: kubernetes.io/aws-ebs
reclaimPolicy: Retain
mountOptions:
  - debug
volumeBindingMode: WaitForFirstConsumer
```

AWS in-tree

Good Practice



```yaml
apiVersion: storage.k8s.io/v1
kind: StorageClass
metadata:
  name: ebs-sc
provisioner: ebs.csi.aws.com
volumeBindingMode: WaitForFirstConsumer
parameters:
  csi.storage.k8s.io/fstype: xfs
  type: io1
  iopsPerGB: "50"
  encrypted: "true"
allowedTopologies:
- matchLabelExpressions:
  - key: topology.ebs.csi.aws.com/zone
    values:
    - us-east-2c
```

AWS CSI Driver

Choosing a remote disk type

IOPS

Each of the storage types have a different threshold. Some allow you to configure the disks to meet performance requirements (provisioned IOPS).

Encryption

Enable encryption for your volumes. It's easy in the public cloud.

Size

Public cloud vendors generally have very high maximum size.

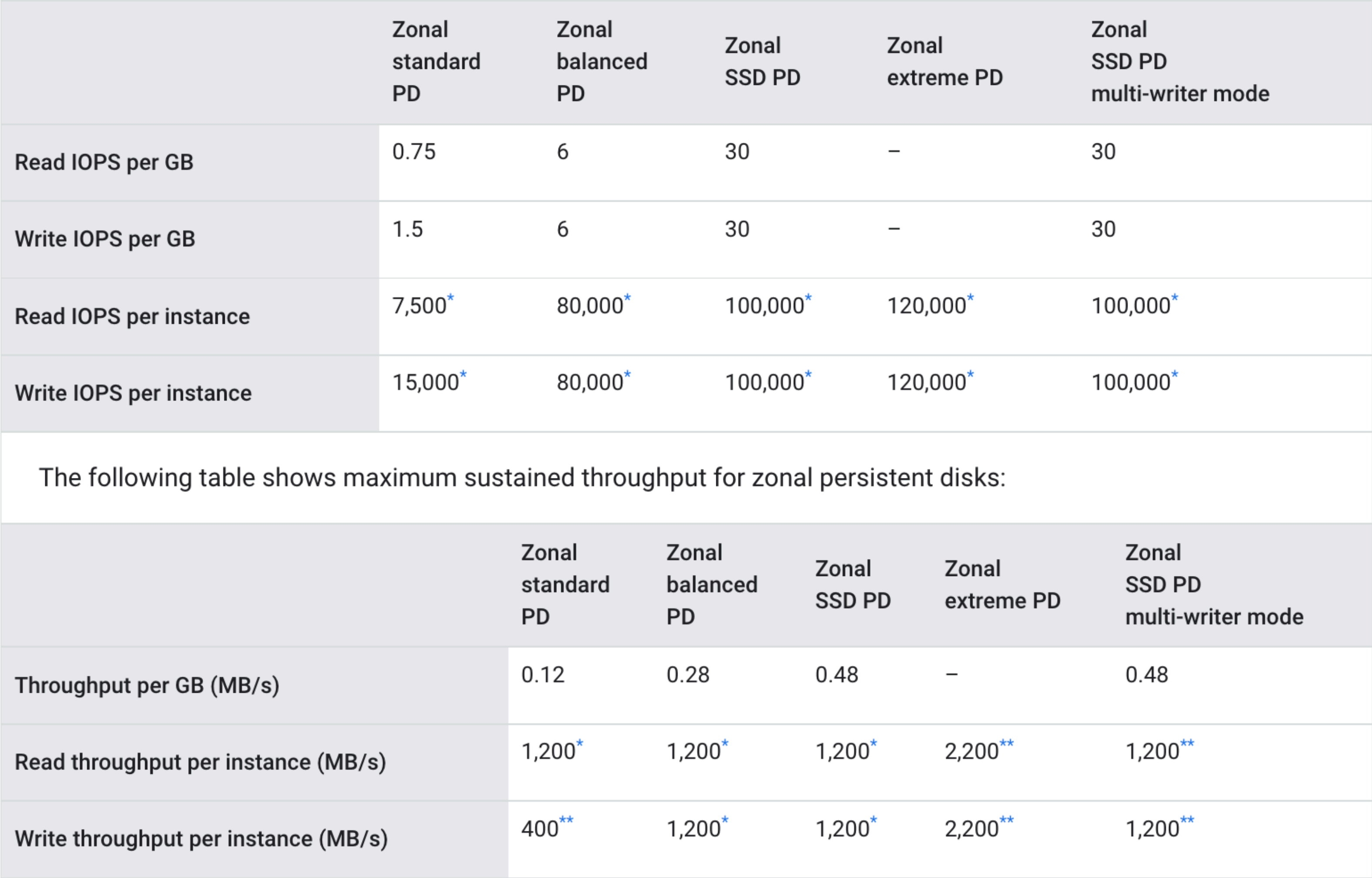
Throughput / IOPS

Depending on the cloud providers IOPS and throughput are determined by the provisioned volume size.

Cost

The cost between remote storage types is widely different in each cloud and across vendors. You want to strike a balance between expenditure and performance.

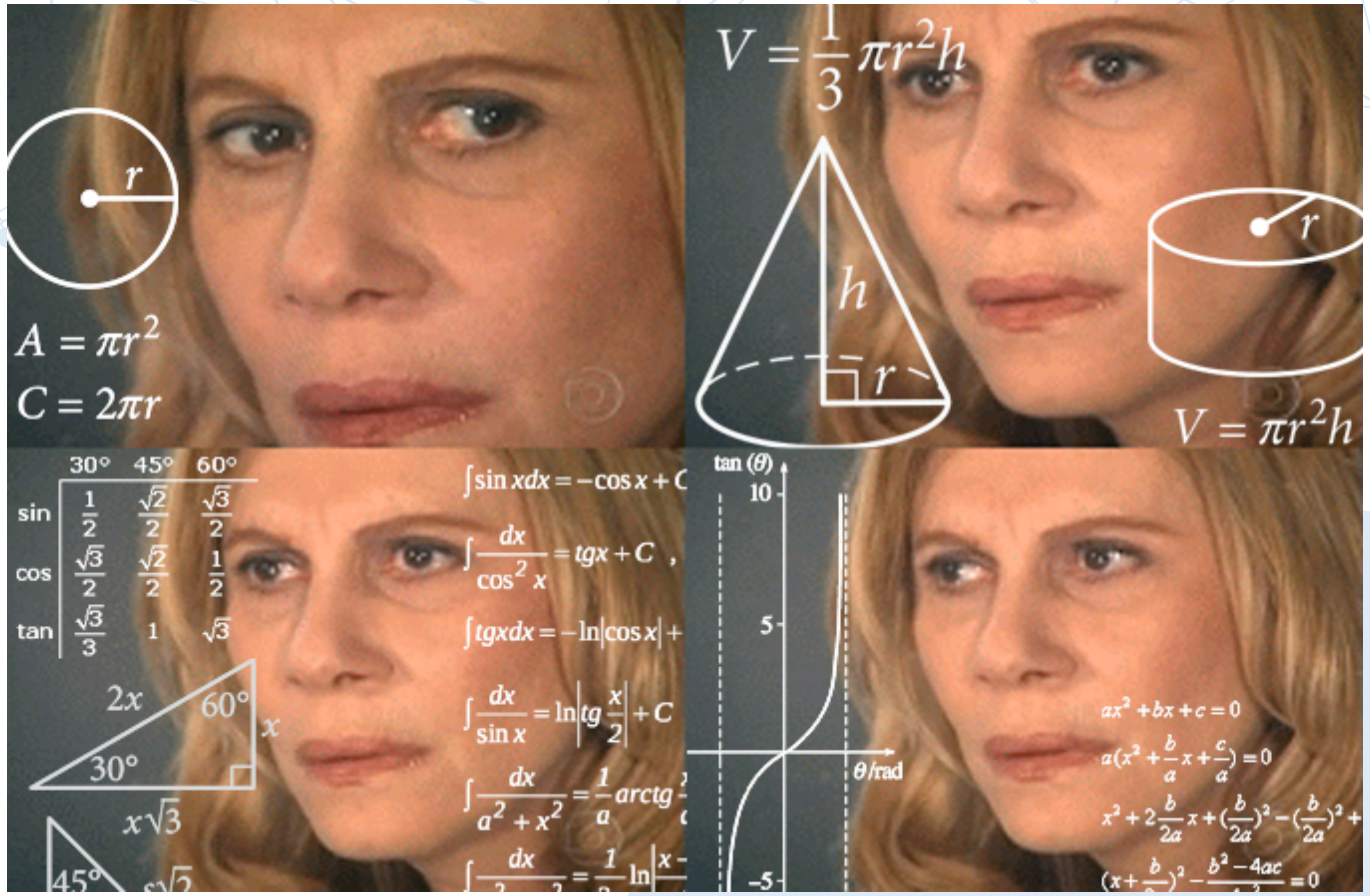
Storage Cost planning

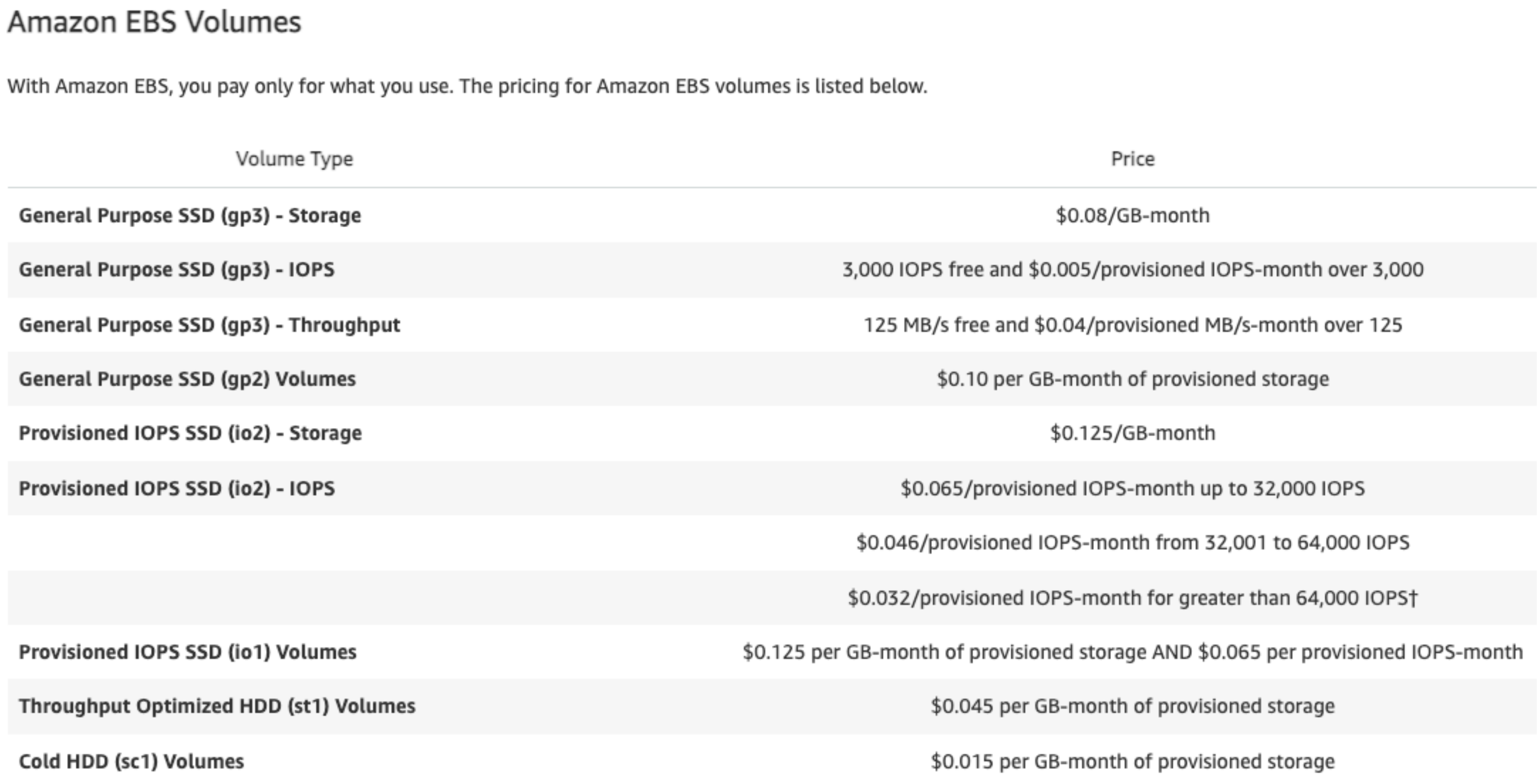


	Zonal standard PD	Zonal balanced PD	Zonal SSD PD	Zonal extreme PD	Zonal SSD PD multi-writer mode
Read IOPS per GB	0.75	6	30	–	30
Write IOPS per GB	1.5	6	30	–	30
Read IOPS per instance	7,500*	80,000*	100,000*	120,000*	100,000*
Write IOPS per instance	15,000*	80,000*	100,000*	120,000*	100,000*

The following table shows maximum sustained throughput for zonal persistent disks:

	Zonal standard PD	Zonal balanced PD	Zonal SSD PD	Zonal extreme PD	Zonal SSD PD multi-writer mode
Throughput per GB (MB/s)	0.12	0.28	0.48	–	0.48
Read throughput per instance (MB/s)	1,200*	1,200*	1,200*	2,200**	1,200**
Write throughput per instance (MB/s)	400**	1,200*	1,200*	2,200**	1,200**

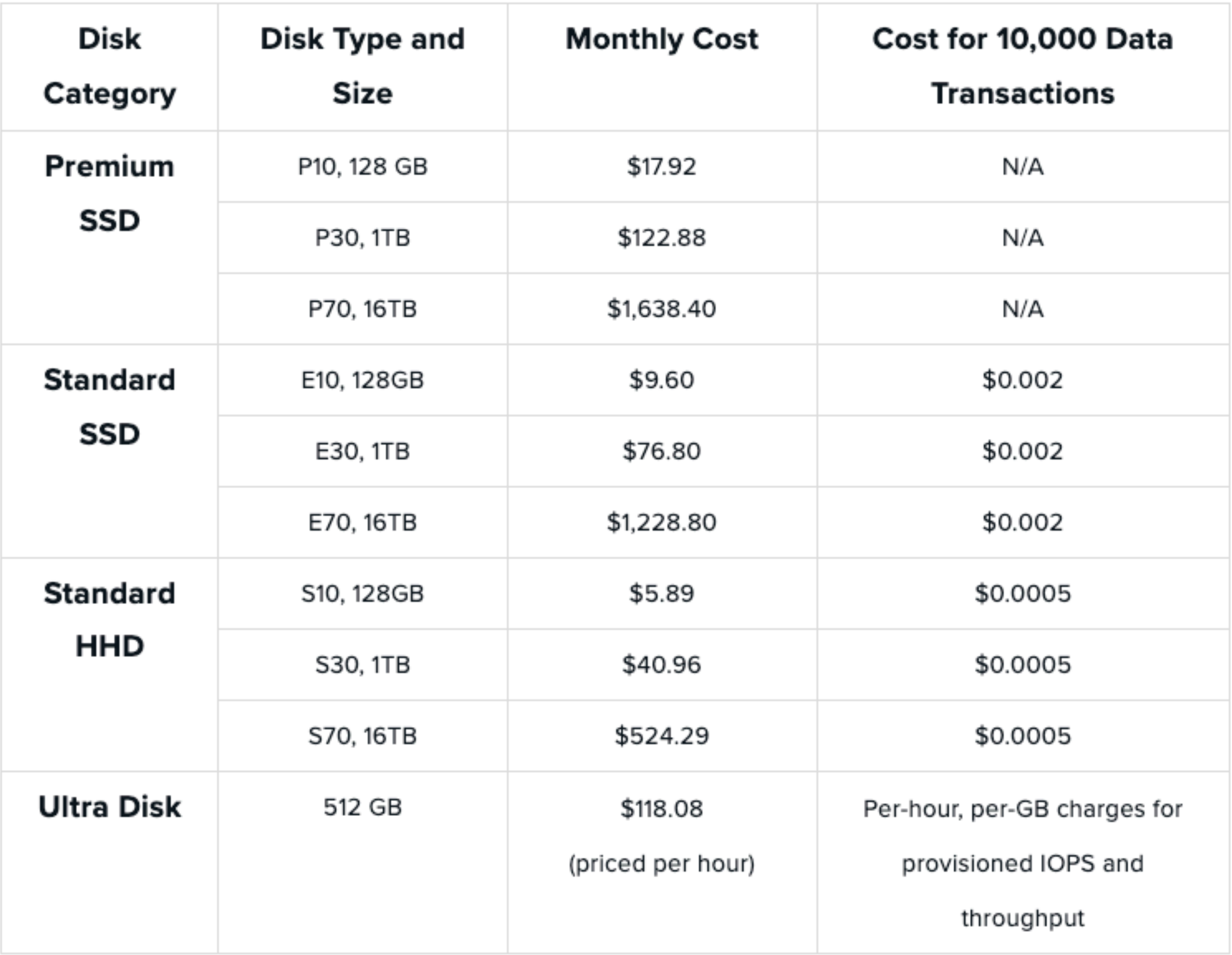

r
$A = \pi r^2$
$C = 2\pi r$
$V = \frac{1}{3}\pi r^2 h$
h
r
$V = \pi r^2 h$
30° 45° 60°
sin
cos
tan
2x
60°
30°
x
$x\sqrt{3}$
45°
$\int \sin x dx = -\cos x + C$
$\int \frac{dx}{\cos^2 x} = tgx + C$
$\int tgx dx = -\ln|\cos x| +$
$\int \frac{dx}{\sin x} = \ln|tg\frac{x}{2}| + C$
$\int \frac{dx}{a^2 + x^2} = \frac{1}{a} arctg$
tan (θ)
10
5
-5
θ/rad
$ax^2 + bx + c = 0$



Amazon EBS Volumes

With Amazon EBS, you pay only for what you use. The pricing for Amazon EBS volumes is listed below.

Volume Type	Price
General Purpose SSD (gp3) - Storage	$0.08/GB-month
General Purpose SSD (gp3) - IOPS	3,000 IOPS free and $0.005/provisioned IOPS-month over 3,000
General Purpose SSD (gp3) - Throughput	125 MB/s free and $0.04/provisioned MB/s-month over 125
General Purpose SSD (gp2) Volumes	$0.10 per GB-month of provisioned storage
Provisioned IOPS SSD (io2) - Storage	$0.125/GB-month
Provisioned IOPS SSD (io2) - IOPS	$0.065/provisioned IOPS-month up to 32,000 IOPS
	$0.046/provisioned IOPS-month from 32,001 to 64,000 IOPS
	$0.032/provisioned IOPS-month for greater than 64,000 IOPS†
Provisioned IOPS SSD (io1) Volumes	$0.125 per GB-month of provisioned storage AND $0.065 per provisioned IOPS-month
Throughput Optimized HDD (st1) Volumes	$0.045 per GB-month of provisioned storage
Cold HDD (sc1) Volumes	$0.015 per GB-month of provisioned storage



Disk Category	Disk Type and Size	Monthly Cost	Cost for 10,000 Data Transactions
Premium SSD	P10, 128 GB	$17.92	N/A
	P30, 1TB	$122.88	N/A
	P70, 16TB	$1,638.40	N/A
Standard SSD	E10, 128GB	$9.60	$0.002
	E30, 1TB	$76.80	$0.002
	E70, 16TB	$1,228.80	$0.002
Standard HHD	S10, 128GB	$5.89	$0.0005
	S30, 1TB	$40.96	$0.0005
	S70, 16TB	$524.29	$0.0005
Ultra Disk	512 GB	$118.08 (priced per hour)	Per-hour, per-GB charges for provisioned IOPS and throughput

Performance Analysis

Testing Cloud Remote Block Storage for K8ssandra

AWS

- **gp2**: General Purpose
- **gp3**: Lower cost than gp2 and higher IOPS
- **io1**: Provisioned IOPS SSD volumes

Azure

- **Standard**: default, general purpose
- **Premium**: low latency and high IOPS and throughput
- **Ultra**: Provisioned IOPS SSD volumes

Google

- **pd-balanced**: Cost-effective and reliable block storage
- **pd-ssd**: Fast and reliable block storage
- **pd-extreme**: Provisioned IOPS SSD volumes

Methodology

Cluster

- 2TB disk volume
- 4 nodes with **4 DCs**, **RF={DC1:1, DC2:1, DC3:1, DC4:0}**
- **Write Consistency=ANY** and **Read Consistency=ALL**
- Each node uses a different storage type
- **XFS** Filesystem
- Adaptive Repairs running

Measurements

- IOPS
- Disk R/W latency
- IOWait
- Average Queue Size

Tool

- NoSQLbench running for 1 day

Three Storage Types per Cloud Provider

- DC1: default storage
- DC2: medium performance storage
- DC3: high performance with provisioned IOPS

apachecon
cassandra
apachecon-eks
Cluster Overview
5m
Cluster Overview
Performance
Overview
System
Cache
Compactions
Coordinator
Application
CQL
Data
Dropped Messages
Entropy
Keyspace
Table
Thread Pools
Security
Reporting
Logs & Events
Alerts & Notifications
Service Checks
Operations
Repairs
Rolling Restart
Backups
Restore
PDF Reports
Settings
Cluster Overview
GRAPH VIEW
LIST VIEW
Context
Alerts And Status

Testing Cassandra on Kubernetes

Methodology

DC2 using medium performance storage

DC1 using standard performance storage

DC3 using provisioned IOPS

DC4 coordinator/injector

apachecon
cassandra
apachecon-gke
Cluster Overview
5m

Configuration

Node ID: 192.168.5.18
Agent ID: 6c7f4594-06d5-4768-9ff3-18d3da772bb9

OS
CASSANDRA
JVM
TASKS
NODESTATS

Search: concurrent

CASSANDRA

concurrent_compactors	
concurrent_counter_writes	32
concurrent_materialized_view_builders	1
concurrent_materialized_view_writes	32
concurrent_reads	256
concurrent_replicates	
concurrent_validations	0
concurrent_writes	256
max_concurrent_automatic_sstable_upgrades	1
native_transport_max_concurrent_connections	-1
native_transport_max_concurrent_connections_per_ip	-1
native_transport_max_concurrent_requests_in_bytes	-1

Testing Cassandra on Kubernetes

apachecon
cassandra
apachecon-eks
Repairs
Cluster Overview
Performance
Overview
System
Cache
Compactions
Coordinator
Application
CQL
Data
Dropped Messages
Entropy
Keyspace
Table
Thread Pools
Security
Reporting
Logs & Events
Alerts & Notifications
Service Checks
Operations
Repairs
Rolling Restart
Backups
Restore
PDF Reports
Settings
Repairs
ADAPTIVE REPAIR
SCHEDULED REPAIR
SHOW ADVANCED SETTINGS

Adaptive Repairs

✓ Active

8 Running Adaptive Repairs

Keyspace	Tables	State	Segments	Failures	Status	Estimated Remaining Duration (Hours)
system_distributed	view_build_status		3 / 32	-	✓	-
baseline1	tabular		0 / 32	-	✓	-
baseline2	tabular		0 / 32	-	✓	-
system_auth	roles		0 / 32	-	✓	-
system_auth	role_permissions		0 / 32	-	✓	-
system_auth	role_members		0 / 32	-	✓	-
system_auth	resource_role_permissons_index		0 / 32	-	✓	-
system_auth	network_permissions		0 / 32	-	✓	-

Number Of Items Per Page 10 20 50 100

0 Pending Adaptive Repairs

Keyspace	Tables	Last Run	Last Status

apachecon
cassandra
apachecon-aks
Performance
System
5m
Cluster Overview
Performance
Overview
System
Cache
Compactions
Coordinator
Application
CQL
Data
Dropped Messages
Entropy
Keyspace
Table
Thread Pools
Security
Reporting
Logs & Events
Alerts & Notifications
Service Checks
Operations
Repairs
Rolling Restart
Backups
Restore
PDF Reports
Settings
Data Center
Rack
Node
Interface
CPU And Load
Average CPU Usage per DC
CPU usage per host
Load Average (15m)
Avg IO wait CPU per Host
100.00%
80.00%
60.00%
40.00%
20.00%
0.00%
10
8
6
4
2
0
15:49:40
15:50:00
15:50:20
15:50:40
15:50:50
15:51:00
15:51:20
15:51:40
15:52:00
15:52:20
15:52:30
15:52:40
15:53:00
15:53:20
15:53:40
15:54:00
15:54:10
15:54:20
Disk Statistics
Disk % Usage /var/lib/cassandra
Used Disk Space Per Node
Bytes Read Per Second
1.61 TB
1.07 TB
537 GB
0 B
107.4 MB
85.92 MB
64.44 MB
42.96 MB
21.48 MB
Disk WeightedIO time
IOPS
Disk avgqsz
6mins
5mins
3mins
1min
0ms
1.5k
1k
500
2k
Bytes Write Per Second
% File Descriptors Allocated
Time Spent on Disk IO
429.6 MB
322.2 MB
214.8 MB
107.4 MB
1s
800ms
600ms
400ms
200ms
0ms
Memory Statistics
Used Memory Percentage
Cached memory
Used memory
107.4 GB
53.7 GB
16.11 GB
10.74 GB
5.37 GB
all

Mount Point: /var/lib/cassandra

Partition: sdb, sdc

Azure: Summary

Bytes Write Per Second
all
451.08 MB
429.6 MB
408.12 MB
386.64 MB
365.16 MB
343.68 MB
322.2 MB
300.72 MB
279.24 MB
257.76 MB
236.28 MB
214.8 MB
193.32 MB
171.84 MB
150.36 MB
128.88 MB
107.4 MB
85.92 MB
64.44 MB
42.96 MB
21.48 MB
0 B
15:50:00
15:50:20
15:50:40
15:51:00
15:51:20
15:51:40
15:52:00
15:53:00
15:53:20
15:53:40
15:54:00
15:54:20
15:54:40

15/09/2022, 15:51:57
● ultra - 10.244.3.12 - sdb 348.61 MB
● premium - 10.244.2.11 - sdc 253.53 MB
● standard - 10.244.1.10 - sdb 105.45 MB
● coordinator - 10.244.0.18 - sdb 4.4 kB

386MB/s

257MB/s

107MB/s

Azure: Bytes Write Per Second

Bytes Read Per Second
all
225.54 MB
214.8 MB
204.06 MB
193.32 MB
182.58 MB
171.84 MB
161.1 MB
150.36 MB
139.62 MB
128.88 MB
118.14 MB
107.4 MB
96.66 MB
85.92 MB
75.18 MB
64.44 MB
53.7 MB
42.96 MB
32.22 MB
21.48 MB
10.74 MB
0 B
14:26:40
14:35:00
14:43:20
14:51:40
15:00:00
15:08:20
15:16:40
15:25:00
15:33:20
15:41:40
15:50:00

15/09/2022, 15:39:06
● ultra - 10.244.3.12 - sdb 100.33 MB
● premium - 10.244.2.11 - sdc 67.56 MB
● standard - 10.244.1.10 - sdb 1.83 MB
● coordinator - 10.244.0.18 - sdb 429.35 kB

Azure: Bytes Read Per Second

128MB/s

50MB/s

Test Results

Azure: IOPS

IOPS
all
1.6k
1.5k
1.4k
1.3k
1.2k
1.1k
1k
900
800
700
600
500
400
300
200
100
0
1.2K
800
400
15:55:20
15:55:40
15:56:00
15:56:20
15:56:40
15:57:00
15:57:20
15:57:40
15:58:00
15:58:20
15:58:40
15:59:00
15:59:20
15:59:40
16:00:00

15/09/2022, 15:57:32

● ultra - 10.244.3.12 - sdb	1.17k
● premium - 10.244.2.11 - sdc	648.4
● standard - 10.244.1.10 - sdb	511
● coordinator - 10.244.0.18 - sdb	0.4

Test Results

Azure: CPU

Average CPU Usage per DC
all
100.00%
95.00%
90.00%
85.00%
80.00%
75.00%
70.00%
65.00%
60.00%
55.00%
50.00%
45.00%
40.00%
35.00%
30.00%
25.00%
20.00%
15.00%
10.00%
5.00%
0.00%
15:56:00
15:56:20
15:56:40
15:57:00
15:57:20
15:57:40
15:58:00
15:58:20
15:58:40
15:59:00
15:59:20
15:59:40
16:00:00
16:00:20
16:00:40

15/09/2022, 15:58:27

● ultra	29.00%
● standard	28.00%
● premium	22.00%
● coordinator	6.00%

50%

30%

10%

Google: Summary

apachecon
cassandra
apachecon-gke
Performance
Overview
5m
Cluster Overview
Performance
Overview
System
Cache
Compactions
Coordinator
Application
CQL
Data
Dropped Messages
Entropy
Keyspace
Table
Thread Pools
Security
Reporting
Logs & Events
Alerts & Notifications
Active
Availability & connections statistics
UP vs Down endpoints
🟩 up (100%)
Native connections
all
4
3
2
1
0
17:38:20
17:40:00
17:41:40
Number of Endpoints Down Per ...
all
Table Count
Latency Statistics Per Node
Coordinator Read Latency - 75t...
all
600µs
400µs
200µs
0µs
17:38:20
17:40:00
17:41:40
Coordinator Scan Latency - 75t...
all
15ms
10ms
5ms
0µs
17:38:20
17:40:00
17:41:40
Coordinator Write Latency - 75t...
all
15ms
10ms
5ms
0µs
17:38:20
17:40:00
17:41:40
Read Latency Per Keyspace Per ...
all
400µs
300µs
200µs
100µs
0µs
17:38:20
17:40:00
17:41:40
Range Read Latency Per Keyspa...
all
500µs
400µs
300µs
200µs
100µs
0µs
17:38:20
17:40:00
17:41:40
Write Latency Per Keyspace Per ...
all
600µs
400µs
200µs
0µs
17:38:20
17:40:00
17:41:40

Data Center

Rack

Node

GroupBy: dc

Percentile: 75thPercentile

Keyspace

Table

No data found in the time range for this chart.

Bytes Write Per Second
all
590.7 MB
537 MB
483.3 MB
429.6 MB
375.9 MB
322.2 MB
268.5 MB
214.8 MB
161.1 MB
107.4 MB
53.7 MB
0 B
17:33:40
17:34:00
17:34:20
17:34:40
17:35:00
17:35:20
17:35:40
17:36:00
17:36:20
17:36:40
17:37:00
17:37:20
17:37:40
17:38:00
17:38:20

530 MB

06/10/2022, 17:36:50

● ssd - 192.168.2.10 - sdb 540.67 MB
● extreme - 192.168.4.9 - sdb 531.58 MB
● standard - 192.168.1.8 - sdb 452.89 MB
● coordinator - 192.168.0.10 - sdb 6.19 MB

Google: Bytes Write Per Second

Bytes Read Per Second
all
236.28 MB
214.8 MB
193.32 MB
171.84 MB
150.36 MB
128.88 MB
107.4 MB
85.92 MB
64.44 MB
42.96 MB
21.48 MB

Google: Bytes Read Per Second

150 MB

06/10/2022, 17:44:16

● ssd - 192.168.2.10 - sdb 83.39 MB
● extreme - 192.168.4.9 - sdb 81.3 MB
● standard - 192.168.1.8 - sdb 80.42 MB
● coordinator - 192.168.0.10 - sdb 453.99 kB

Test Results

Google: IOPS

IOPS
all
3000 IOPS
2000 IOPS
3.5k
3k
2.5k
2k
1.5k
1k
500
0
17:46:00
17:46:20
17:46:40
17:47:00
17:47:20
17:47:40
17:48:00
17:48:20
17:48:40
17:49:00
17:49:20
17:49:40
17:50:00
17:50:20
17:50:40

06/10/2022, 17:49:00

● standard - 192.168.1.8 - sdb	2.53k
● extreme - 192.168.4.9 - sdb	2.44k
● ssd - 192.168.2.10 - sdb	2.31k
● coordinator - 192.168.0.10 - sdb	26.6

Test Results

AWS: Summary

apachecon
cassandra
apachecon-eks
Performance
System
30m
Cluster Overview
Performance
Overview
System
Cache
Compactions
Coordinator
Application
CQL
Data
Dropped Messages
Entropy
Keyspace
Table
Thread Pools
Security
Reporting
Logs & Events
Alerts & Notifications
Service Checks
Operations
PDF Reports
Settings
Data Center
Rack
Node
Mount Point
Partition
Interface
CPU And Load
Average CPU Usage per DC
CPU usage per host
Load Average (15m)
100.00%
80.00%
60.00%
40.00%
20.00%
0.00%
8
6
4
2
0
14:15:00
14:23:20
14:31:40
Avg IO wait CPU per Host
14:10:00
14:11:40
14:13:20
14:15:00
14:16:40
14:18:20
14:20:00
14:21:40
14:23:20
14:25:00
14:26:40
14:28:20
14:30:00
14:31:40
14:33:20
14:35:00
14:36:40
14:38:20
Disk Statistics
Disk % Usage All
Used Disk Space Per Node
1.07 TB
859.2 GB
644.4 GB
429.6 GB
214.8 GB
0 B
Bytes Read Per Second
214.8 MB
161.1 MB
107.4 MB
53.7 MB
0 B
Bytes Write Per Second
644.4 MB
429.6 MB
214.8 MB
0 B
IOPS
4k
3k
2k
1k
0
Disk avgqsz
80
60
40
20
0
Disk WeightedIO time
15s
10s
5s
0ms
% File Descriptors Allocated
Time Spent on Disk IO
1s
800ms
600ms
400ms
200ms
0ms
Memory Statistics
Network Statistics
Network Received (bytes)
372.44 MiB
186.22 MiB
0 B
Network Transmitted (bytes)
558.66 MiB
372.44 MiB
186.22 MiB
0 B
NTP offset (milliseconds)
80ms
60ms
40ms
20ms
0ms
all

Test Results

AWS: Bytes Write Per Second

Bytes Write Per Second
all
698.1 MB
644.4 MB
590.7 MB
537 MB
483.3 MB
429.6 MB
375.9 MB
322.2 MB
268.5 MB
214.8 MB
161.1 MB
107.4 MB
53.7 MB
0 B
14:13:20
14:15:00
14:16:40
14:18:20
14:20:00
14:21:40
14:23:20
14:25:00
14:26:40
14:28:20
14:30:00
14:31:40
14:33:20
14:35:00
14:36:40
14:38:20
14:40:00
14:41:40

23/09/2022, 14:36:46

● io1 - 10.123.3.4 - nvme2n1	624.49 MB
● gp2 - 10.123.2.142 - nvme2n1	213.07 MB
● gp3 - 10.123.2.176 - nvme2n1	111.15 MB

590 MB

268 MB

107 MB

Test Results

AWS: IOPS

IOPS
all
4k
3.8k
3.6k
3.4k
3.2k
3k
2.8k
2.6k
2.4k
2.2k
2k
1.8k
1.6k
1.4k
1.2k
1k
800
600
400
200
0
14:13:20
14:15:00
14:16:40
14:18:20
14:20:00
14:21:40
14:23:20
14:25:00
14:26:40
14:28:20
14:30:00
14:31:40
14:33:20
14:35:00
14:36:40
14:38:20
14:40:00
14:41:40

23/09/2022, 14:39:50

● io1 - 10.123.3.4 - nvme2n1	2.54k
● gp2 - 10.123.2.142 - nvme2n1	1.54k
● gp3 - 10.123.2.176 - nvme2n1	669.4

3000 IOPS

1500 IOPS

800 IOPS

Test Results

AWS: Bytes Disk Read Per Second

Bytes Read Per Second
all
236.28 MB
214.8 MB
193.32 MB
171.84 MB
150.36 MB
128.88 MB
107.4 MB
85.92 MB
64.44 MB
42.96 MB
21.48 MB
0 B
14:13:20
14:15:00
14:16:40
14:18:20
14:20:00
14:21:40
14:23:20
14:25:00
14:26:40
14:28:20
14:30:00
14:31:40
14:33:20
14:35:00
14:36:40
14:38:20
14:40:00
14:41:40

23/09/2022, 14:39:10

● gp2 - 10.123.2.142 - nvme2n1	47.17 MB
● gp3 - 10.123.2.176 - nvme2n1	41.11 MB
● io1 - 10.123.3.4 - nvme2n1	31.68 MB

171 MB

107 MB

85 MB

42 MB

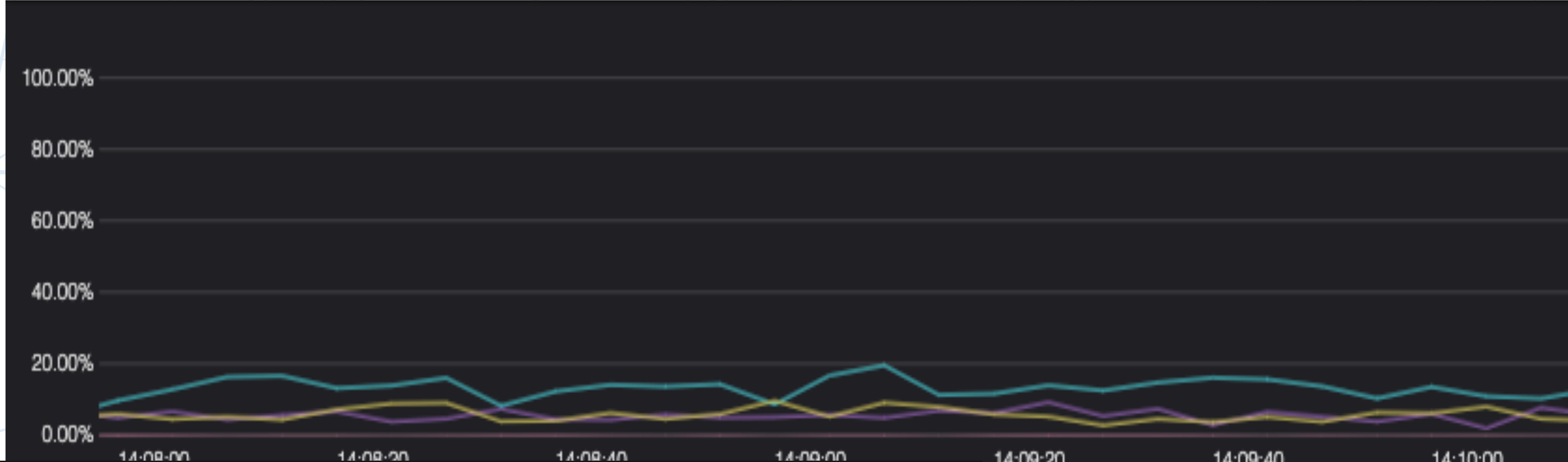
Wait, there is more...

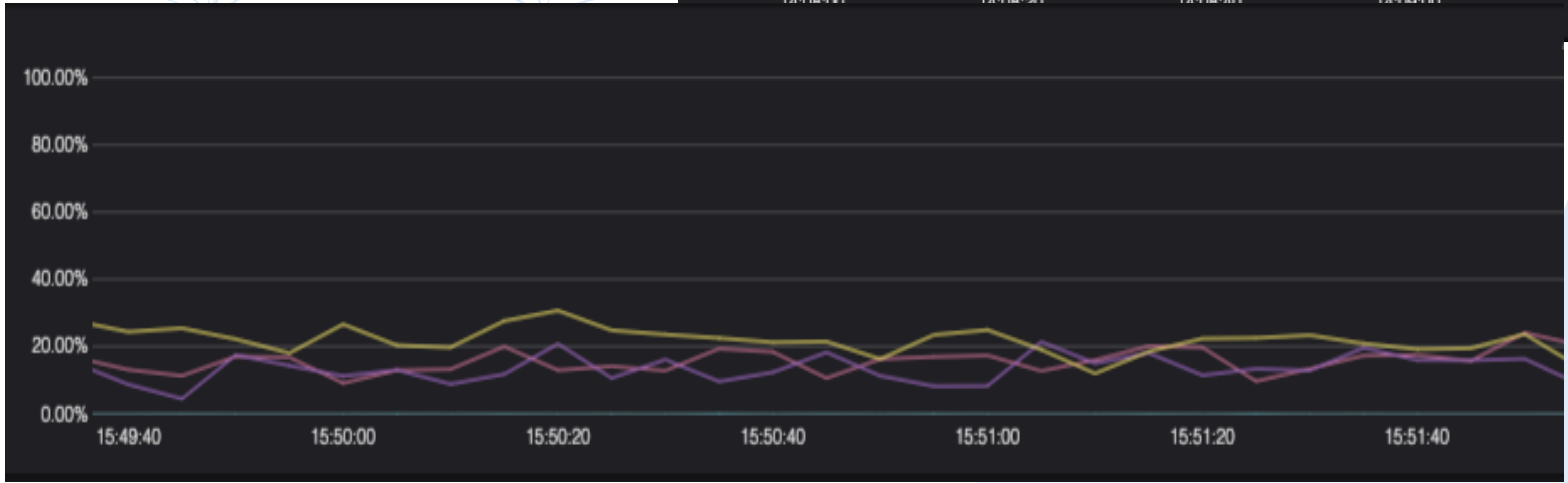
Wait for it...

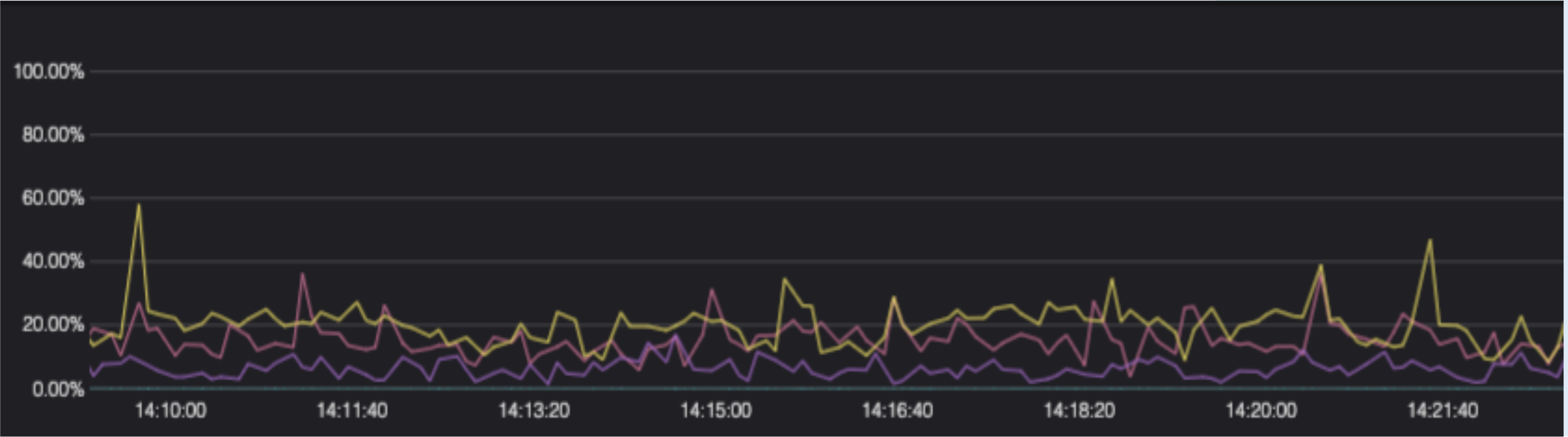
Wait for it some more...

Test Results

IOWait - Measure of Impatience

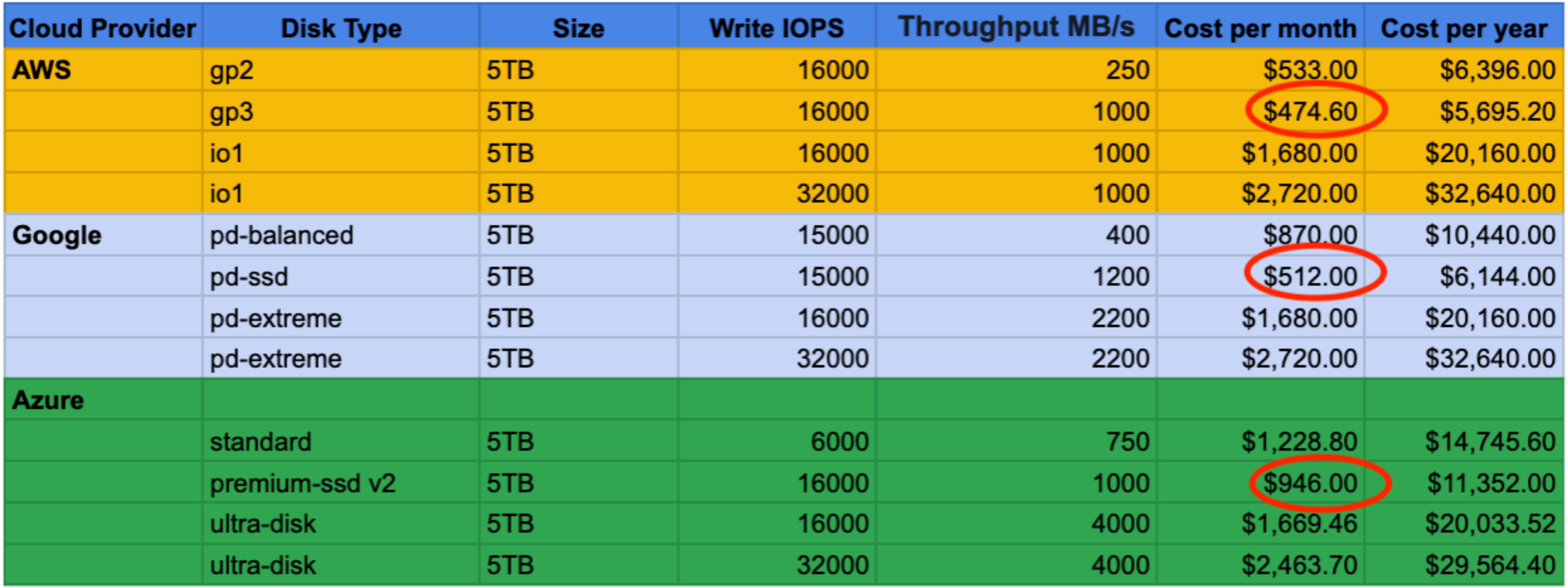






Final Thoughts

Costs



Cloud Provider	Disk Type	Size	Write IOPS	Throughput MB/s	Cost per month	Cost per year
AWS	gp2	5TB	16000	250	$533.00	$6,396.00
	gp3	5TB	16000	1000	$474.60	$5,695.20
	io1	5TB	16000	1000	$1,680.00	$20,160.00
	io1	5TB	32000	1000	$2,720.00	$32,640.00
Google	pd-balanced	5TB	15000	400	$870.00	$10,440.00
	pd-ssd	5TB	15000	1200	$512.00	$6,144.00
	pd-extreme	5TB	16000	2200	$1,680.00	$20,160.00
	pd-extreme	5TB	32000	2200	$2,720.00	$32,640.00
Azure						
	standard	5TB	6000	750	$1,228.80	$14,745.60
	premium-ssd v2	5TB	16000	1000	$946.00	$11,352.00
	ultra-disk	5TB	16000	4000	$1,669.46	$20,033.52
	ultra-disk	5TB	32000	4000	$2,463.70	$29,564.40

Final thoughts

Conclusions

- **Storage selection for Cassandra on Kubernetes requires some R&D**
- **Remote storage is slow and expensive**
- **Local SSDs will give you much better performance but tasks like upgrading the K8s version could become a very lengthy exercise for a large cluster**
- **Remote storage is convenient but the performance suffers**
- **Remote storage is pricey - pays for the beers ApacheCon!**

Final thoughts

Recommendations if you're going to use K8ssandra

Disk Spec

Watch out for the minimum requirements for disk size and your required IOPS. If unsure, a good starting size is 32GB but fewer than 2TB may not be enough.

Throughput

Each of the storage types has a different throughput. The virtual machine types selected must accommodate the network bandwidth for both Cassandra and remote disks.

Do performance testing

You will not know if you have the right set up until you tested. *cassandra-stress* and *nosqlbench* are good tools for this purpose.

Keep up with K8s releases

Public cloud managed Kubernetes versions have EOL dates well defined.

Get comfortable with the operator

Test your node restoration process for your chosen storage types, especially if you go with the local ephemeral volumes.

Trial and error

You may not get it right the first time for your ever changing workload. Prepare to change storage type if needed.

Thank You

https://axonops.com
info@axonops.com

Hayato Shimizu
https://www.linkedin.com/in/hayatoshimizu/

Sergio Rua
https://www.linkedin.com/in/sergiorua/