Apache Ozone
State of the Union

ApacheCon 2022

Siyao Meng, Ethan Rose

Speakers

Siyao Meng

❑ Engineer at Cloudera Storage Team

❑ Apache Ozone PMC and Committer, Apache Hadoop Committer

❑ GitHub @smengcl

Ethan Rose

❑ Engineer at Cloudera Storage Team

❑ Apache Ozone PMC and Committer

❑ GitHub @errose28

Agenda

❑ History of Apache Ozone

❑ Current state of Ozone

❑ New features and improvements in 1.3.0

❑ Roadmap

Brief History of Apache Ozone

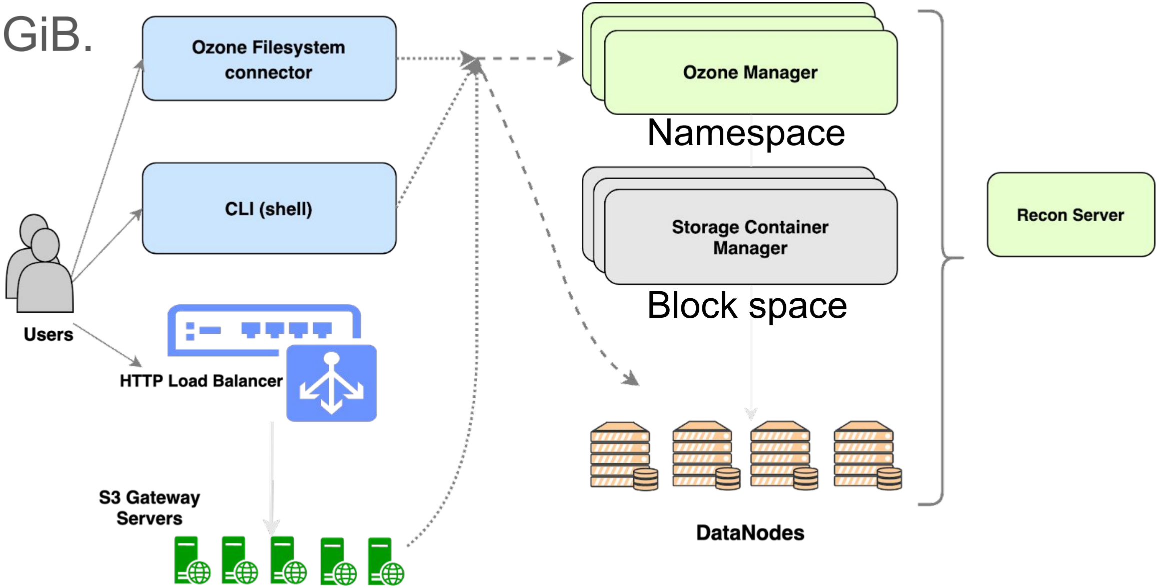
- ❏ To address scalability issue of HDFS.
- ❏ Designed to store **billions of** objects in a single cluster.
- ❏ Ozone started as a sub-project under Hadoop (HDFS-7240).
- ❏ Ozone is built by the Apache Hadoop community.
- ❏ Ozone was established as a Top Level Project (TLP) after 4 alpha releases and 1 beta release in Oct 2020.

What is Apache Ozone

- Distributed key-value store
- **Object Store** for Apache Hadoop
- Stores metadata in high-performance embedded **RocksDB**, relying on off-heap memory
- Provides strong consistency
- Uses Raft in high availability and 3x replication
- Built-in security: Kerberos authentication, pluggable authorizer, encryption
- Seamlessly works with YARN, MapReduce, Hive, Spark with the Hadoop Compatible FileSystem (HCFS) interface.

Building Blocks of Ozone

- Ozone separates **namespace** management and **block space** management
    - Ozone namespace layout: `/`**`volume`**`/bucket/key`
- Scales by not tracking individual data blocks. Instead, SCM tracks containers*, which aggregates blocks. By default, each container* can be as large as 5 GiB.



Ozone Releases

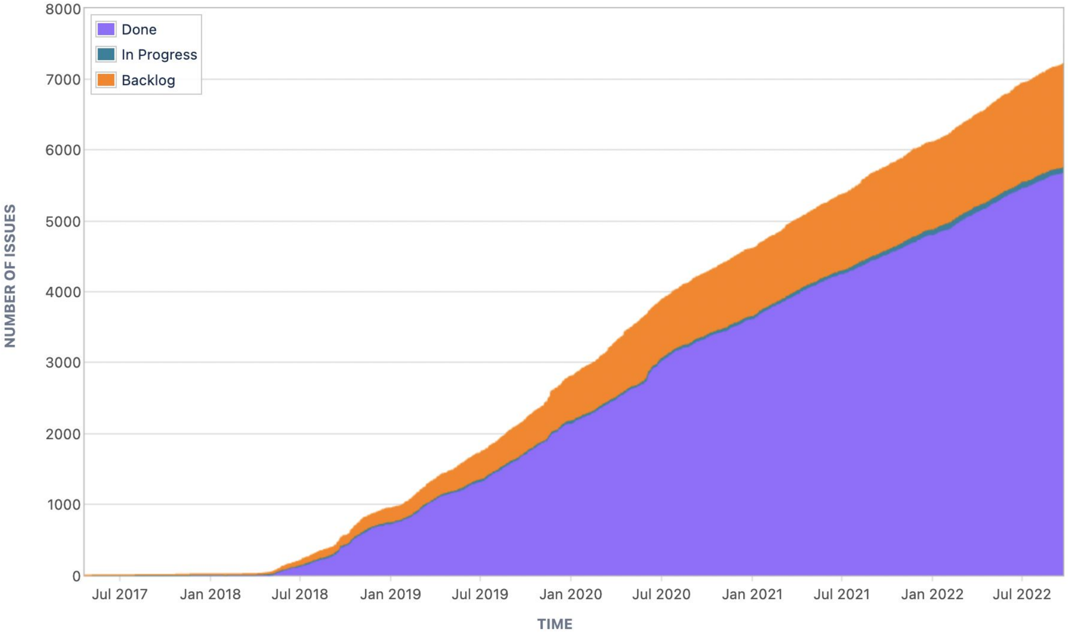
- ❑ Generally Available since 1.0.0 in Sep 2020
- ❑ Latest stable 1.2.1, released in Dec 2021
- ❑ Version 1.3.0 is in-progress
    - ❑ **Tons** of new features and improvements
        - ❑ Erasure Coding
        - ❑ Container Balancer
        - ❑ S3 Multi-Tenancy
        - ❑ S3 gRPC improvements
        - ❑ …
    - ❑ 983 new commits since 1.2.1 release and counting
        - ❑ 2,265 changed files with 150,474 additions and 36,212 deletions

Apache Ozone Committee and Community

- Ozone PMC Chair: Sammi Chen
- 28 PMC members (+1 this year), 61 Committers (+10 since last SotU)
  - Committers / PMC members located in US, Hungary, India, China, Germany, …
  - from Cloudera, Target, Tencent, Infinstor, Oracle, Microsoft, Intel, G-Research, …
- 199 contributors (who has at least one PR merged), 127 active contributors in the past two years.
- 4975 commits in total on the main branch, 2067 merged in the past two years.

Apache Ozone JIRA

- 7,200+ JIRAs opened under Apache Ozone (HDDS) project and counting
  - The original HDFS-7240 uber jira also has another 594 task JIRAs opened under HDFS tag
- 2,968 JIRAs opened, 2,134 of them resolved in the past 2 years



New Feature: Erasure Coding (HDDS-3816)

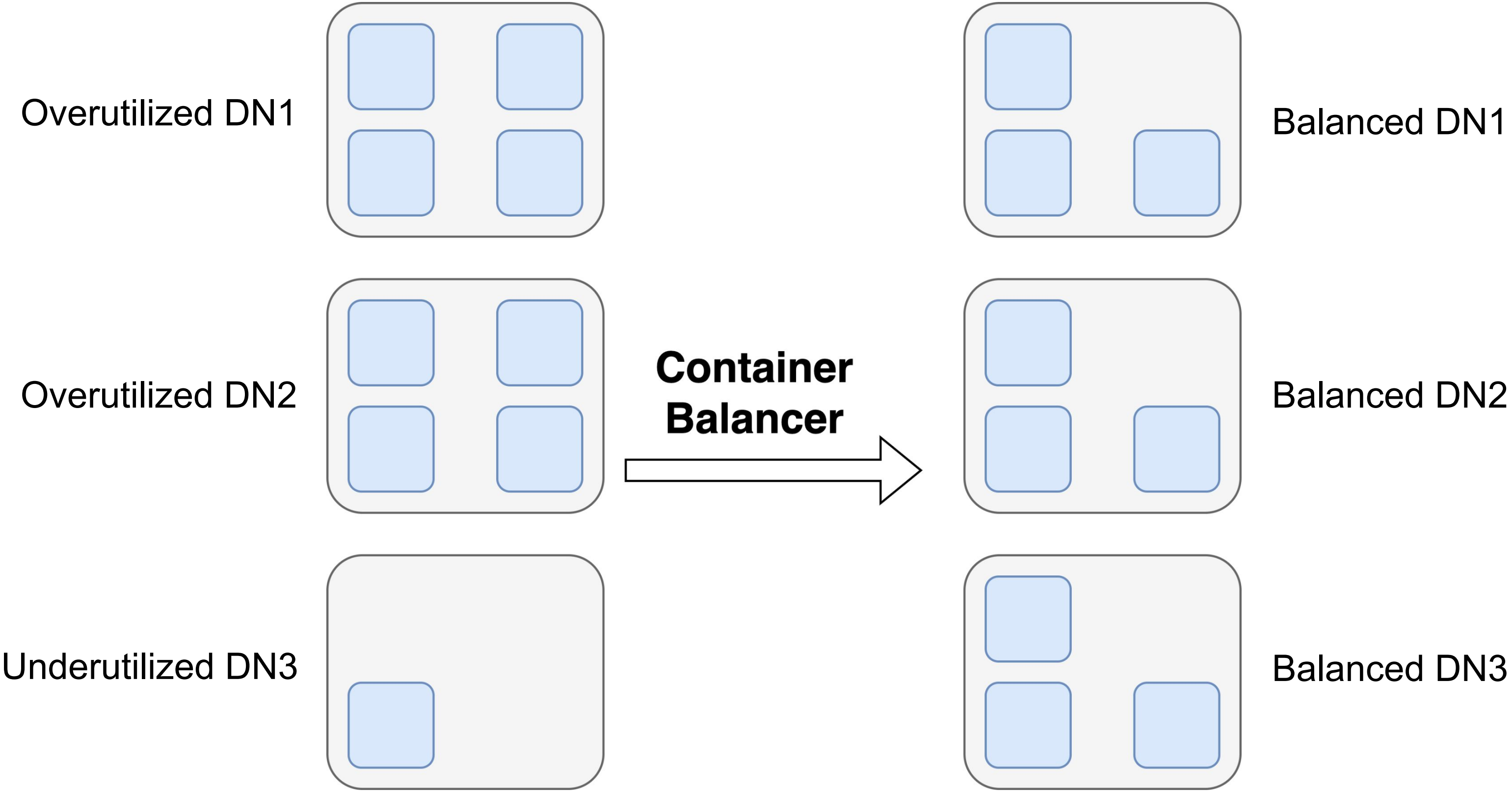
- Much better **storage efficiency** than traditional 3x replication
- Potentially helps reduce tail latency when fetching data
- Check out this dedicated session by Uma (yesterday) for more details
    - *Reduce Your Storage Footprint with Apache Ozone Erasure Coding*

	Data blocks	Parity blocks	Data durability	Storage efficiency
Single replica	1	0	0	100 %
Three replicas	3	0	2	33 %
RS(6,3)	6	3	3	66 %
RS(10,4)	10	4	4	71 %

New Feature: Container Balancer (HDDS-4656)

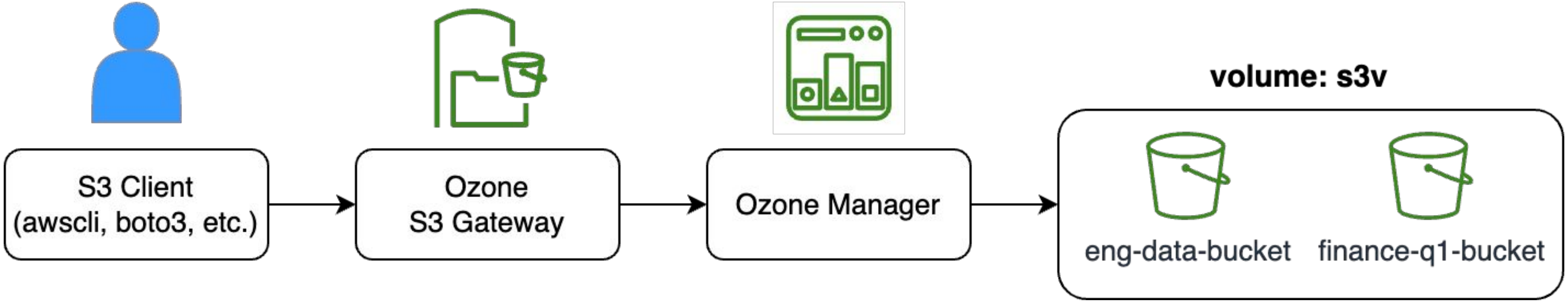
- Stateless service, built into Storage Container Manager (SCM)
- Use Cases
  - New DataNodes are added to a cluster, need to move some existing containers to those empty nodes.
  - DataNodes' utilization become skewed overtime. e.g. due to data deletion.
- We can start the container balancer with admin command:
  - **`ozone admin containerbalancer start`**
- Configurable: util threshold, max iterations, max size to move in each iter, percentage% of datanodes to be involved in each iter, ...
- Check out the talk by *Lokesh* and *Siddhant* for more depth into the feature
  - *Balancing data in Apache Ozone* https://youtu.be/l6L3E6q0dpk

New Feature: Container Balancer (HDDS-4656)



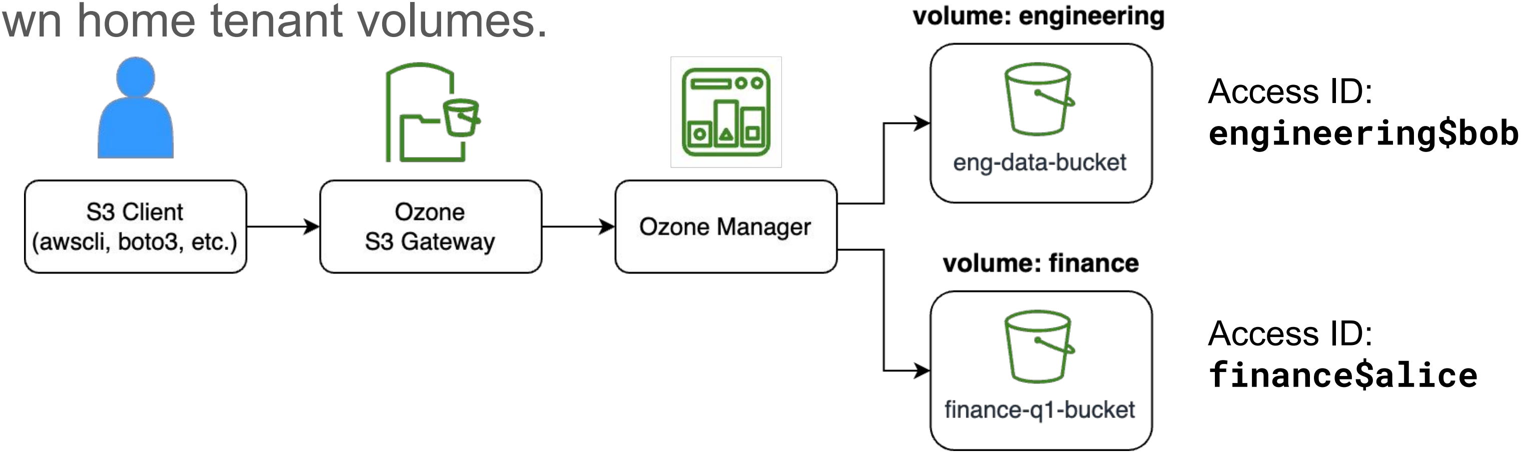
New Feature: S3 Multi-Tenancy (HDDS-4944)

- Ozone namespace layout: /**volume**/bucket/key
- Before S3 Multi-Tenancy feature, all S3 requests to Ozone (via S3 Gateway) are limited to a dedicated **s3v** volume only.
- What if users want the power of Ozone volumes with the compatibility of S3 interface?
- The following is a diagram shows a typical S3 request path: From S3 Client → S3 Gateway → Ozone Manager → s3v volume → bucket



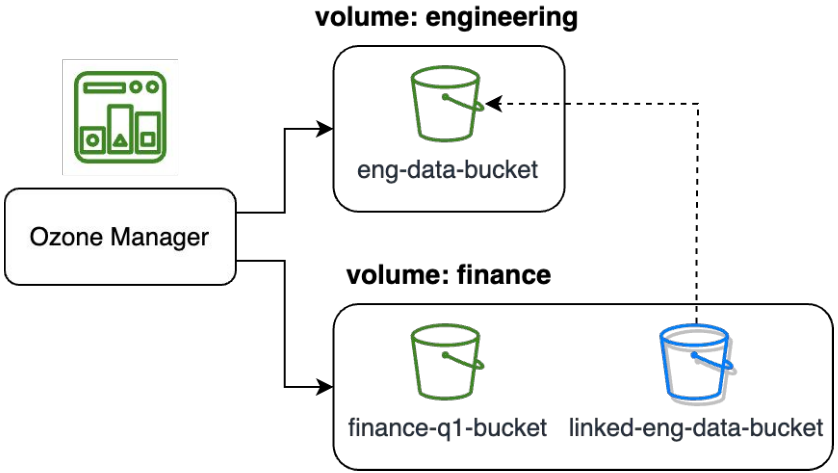
New Feature: S3 Multi-Tenancy (HDDS-4944)

- Now with S3 Multi-Tenancy, Ozone admins can use CLI to create tenants with their own volumes, assign tenant users.
  - `ozone tenant create finance`
  - `ozone tenant user assign alice --tenant=finance`
- Optionally, Ozone admins can assign tenant admins that can manage their own tenants (e.g. assign new tenant users).
- Most importantly, Requests from tenant users are now transparently routed to their own home tenant volumes.



New Feature: S3 Multi-Tenancy (HDDS-4944)

- Because access to different volumes from S3 are naturally isolated, if users need to access buckets from other tenant volumes, such cross-volume sharing is achieved by creating bucket symlinks.
- Access control policy must be configured (with Apache Ranger) to allow user access to the source bucket. See this document section for more.

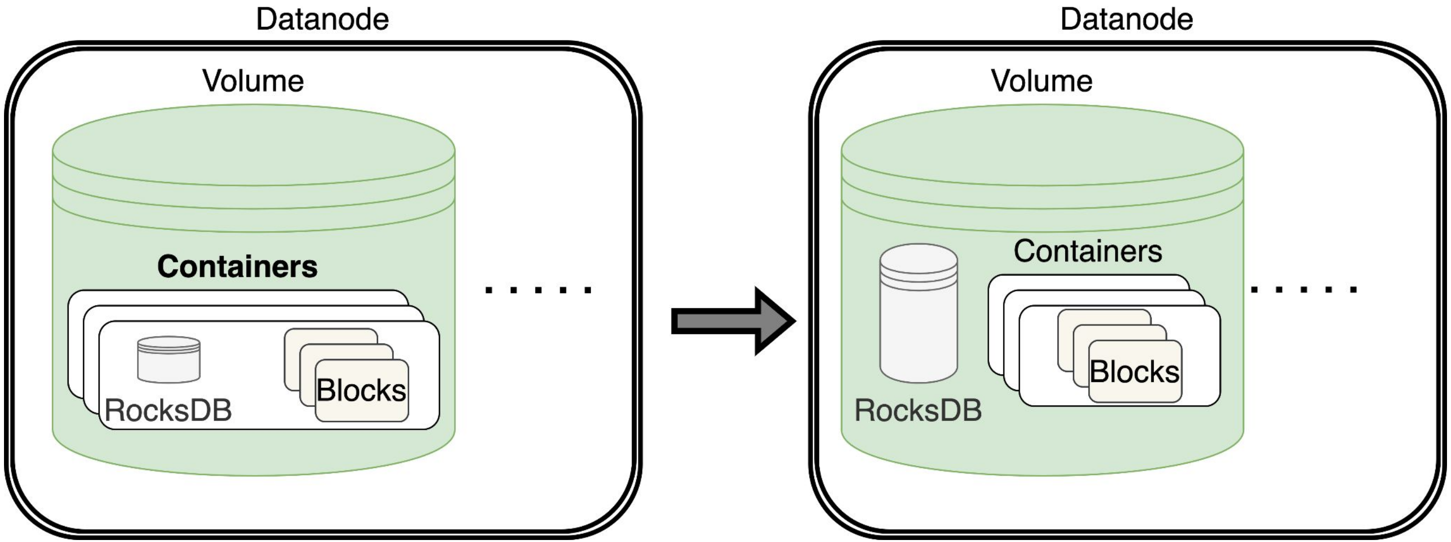


Ozone Manager Performance Improvements

- S3 Gateway
  - Client to OM now supports gRPC for S3 Gateway
    - Per client performance with on the wire encryption in gRPC is significantly faster.
  - S3 Gateway now supports **persistent** client connection to OM.
- Ozone Manager
  - Improving OM ops per second with OM container cache (HDDS-7223)
  - OM locking improvements in the works (HDDS-6402 and more in the pipeline)

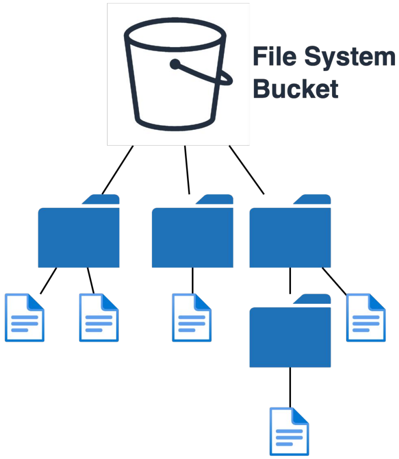
1 RocksDB per Datanode Volume

- Original container design: 1 RocksDB per container
  - Resulted in many small RocksDB instances **affecting performance** and **stability**
- New container design: 1 RocksDB per volume
  - All containers share 1 RocksDB on the volume (disk)

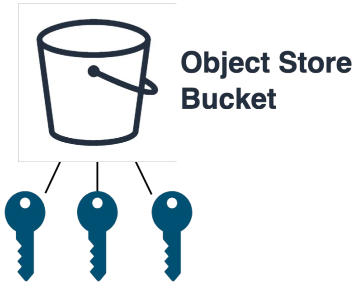


Bucket Layout Types

- File System Optimized (FSO)
  - Hadoop compatible
  - Directories and files
  - Atomic directory rename and delete



- Object Store (OBS)
  - S3 compatible
  - Flat namespace



Roadmap

- ❏ Snapshot support (HDDS-6517)
- ❏ Certificate rotation
- ❏ Recon UI/UX improvements and new features
- ❏ Storage tiering
- ❏ Rolling upgrades

Q&A

Thank you!

❑ More Ozone talks in ApacheCon 2022
  ❑ *Reduce Your Storage Footprint with Apache Ozone Erasure Coding*
    ❑ Monday, Oct 3 02:20 PM CDT
  ❑ *Inside an Apache Ozone Upgrade*
    ❑ Monday, Oct 3 03:10 PM CDT
  ❑ ***Performance of Apache Ozone on NVMe***
    ❑ **Thursday, Oct 6 12:10 PM CDT**

❑ *Ozone Birds of a Feather* sessions
  ❑ Monday, Oct 3 05:50 PM CDT
  ❑ **Wednesday, Oct 5 05:50 PM CDT**

For more

- ❏ Ozone homepage: https://ozone.apache.org
- ❏ Ozone repo: https://github.com/apache/ozone
- ❏ Ozone dev wiki: https://cwiki.apache.org/confluence/display/OZONE
- ❏ Developer mailing list: dev@ozone.apache.org