Performance of Apache Ozone on NVMe

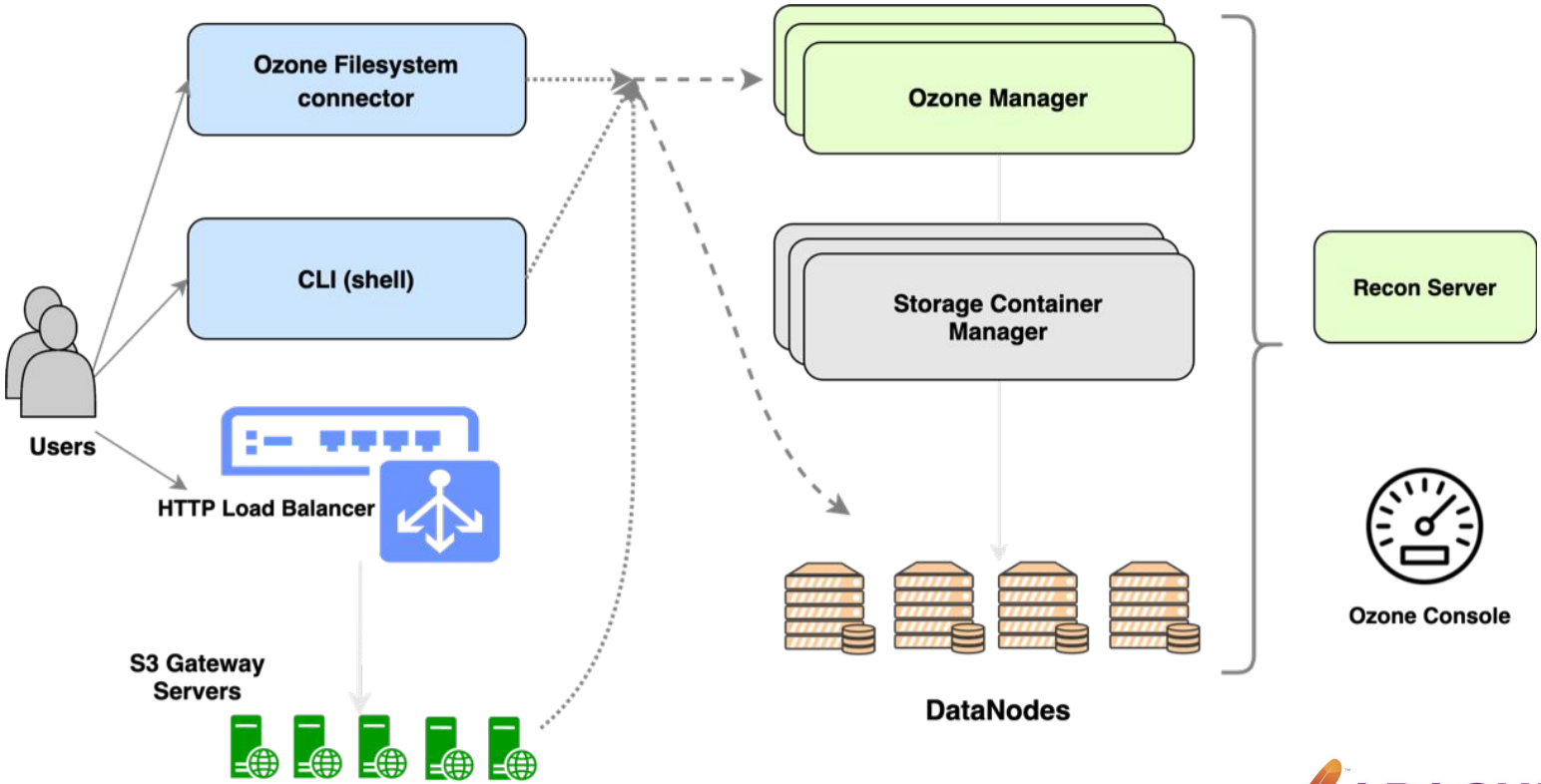
Wei-Chiu Chuang (jojochuang)

Ritesh Shukla (kerneltime)

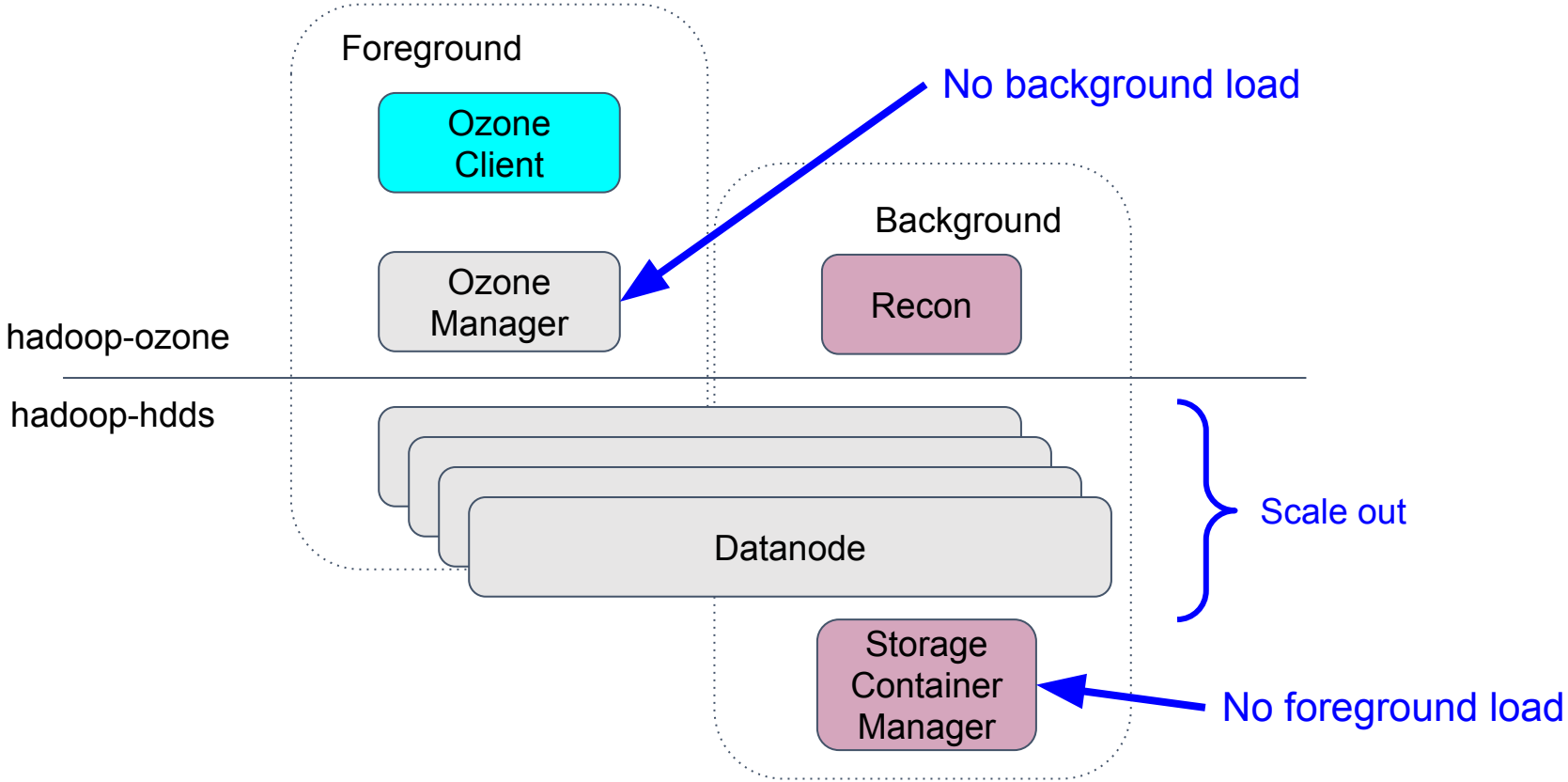
Agenda

- Overview of how Ozone and how it scales
- Why NVME is important for Ozone for scaling
- Benefits of using NVME
- Impala performance results from NVME clusters
- Write path improvements results from NVME clusters
- Summary
- Questions

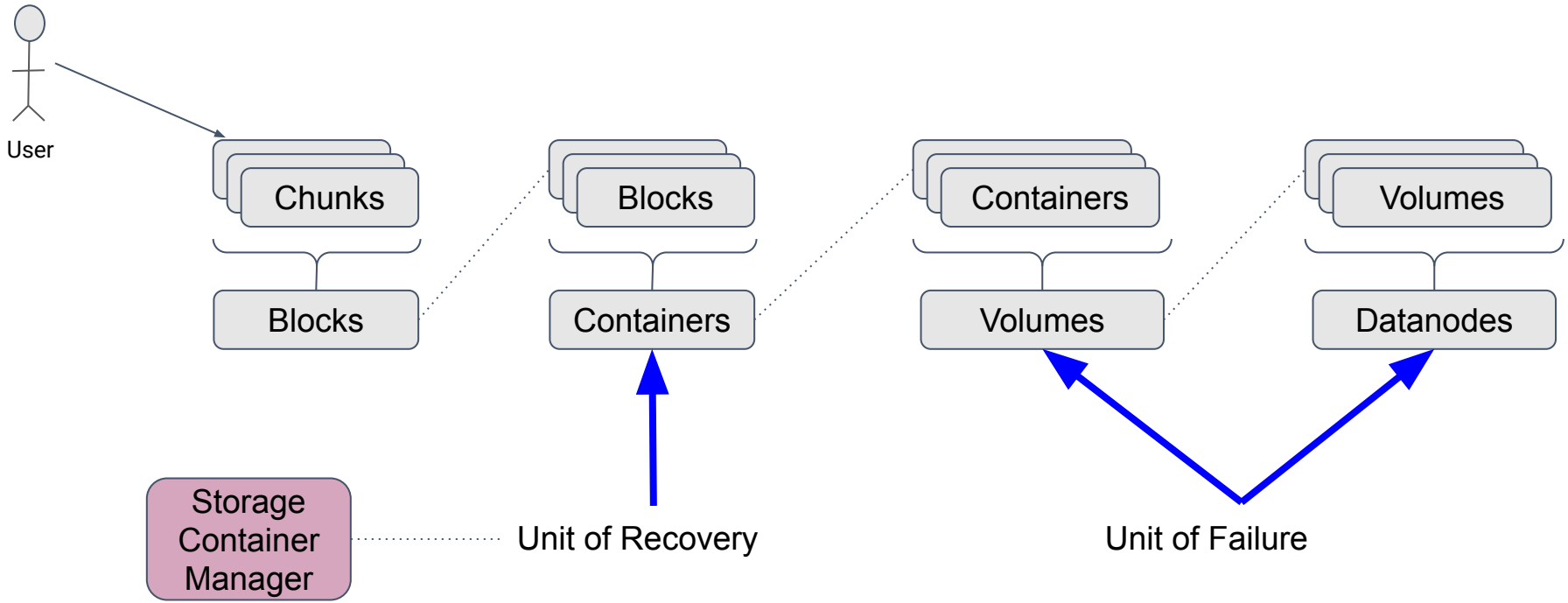
Ozone Architecture



Why does Ozone Scale? Separation of concerns



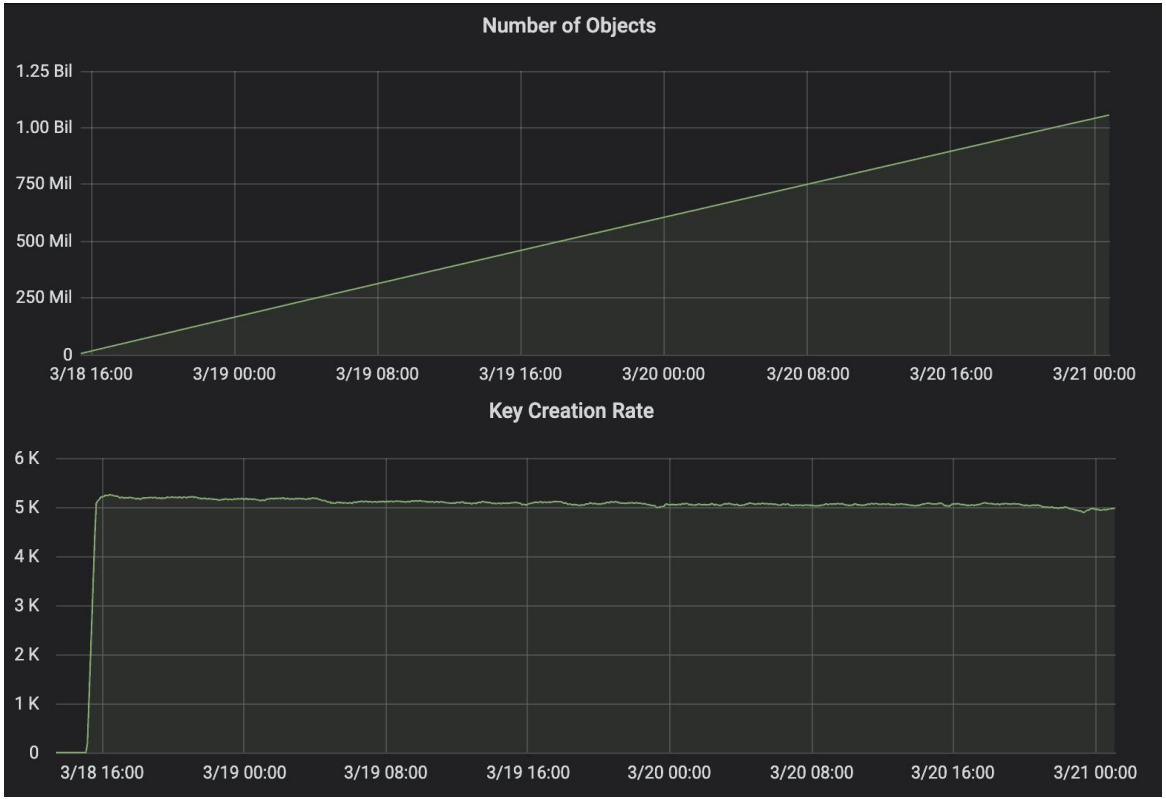
Why does Ozone Scale? Aggregation via containers



Why does foreground Scale?

- No heap limitations, working set can be cached in memory and unused data can be destaged to disk
- OM uses NVME to store RocksDBs
- Future projects such as Snapshots leverage RocksDB to preserve simplicity of IO path.

Ozone scales!



CLOUDERA

APACHECON North America

Does background scale up and scale out?

- Datanode count can scale beyond HDFS
  - No memory pressure on OM due to block reports/object counts/heap limitations
- Container abstraction allows scaling of Datanodes and any background processing.
- Much higher density per Datanode than HDFS

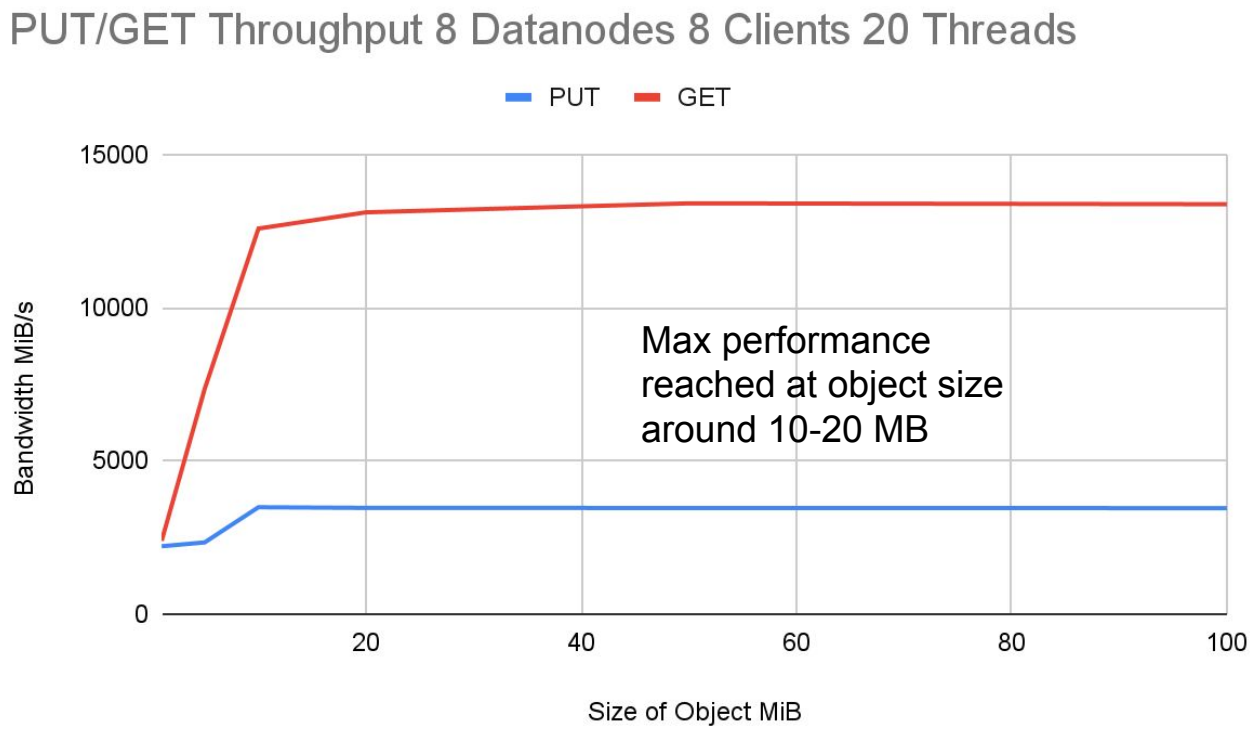
Datanode scales out and scale up

- Testbed used:
  - ~400 TB/Datanode
  - Tested with 200k containers per datanode => 1 PB per datanode.
- Cisco UCS M6
  - Capacity node: 256 TB per datanode
- Cisco UCS S3260
  - Extreme Capacity: 384 TB per datanode

Ozone vs. HDFS

Capability	Ozone	HDFS
Storage Density	1000's of nodes at 600TB per node	1000's of nodes at 100TB per node
Scalability	10B Objects	400M Objects
Recovery	Fast recovery (< 5 min restart)	Slow startup based on size
High Availability	Active - Active	Active - Standby
Protocol Support	Hadoop / S3 API	Hadoop API

Small objects are welcome



Hardware trends

- Cloudera recommends Ozone's metadata reside on NVME
- Not just metadata increasing number of customers using all NVME clusters for Ozone
- Ozone certified against Cisco all NVME data intelligence platform
  - https://www.cisco.com/c/en/us/td/docs/unified_computing/ucs/UCS_CVDs/cisco_ucs_cdip_allnvme_intersight.html
- Customers see TCO benefits with all NVME clusters

Why NVME

- Enables destaging of data with minimal impact to performance.
  - Long tail latency is a small percentage of the overall latency
- Vendors increasing shipping configurations with NVME
- Bet in the right direction of hardware trends.
- Low latency metadata can stay on NVME
  - Data at scale can be on spindles.

Disk characteristics (rule of thumb)

	HDD	(SATA) SSD	NVMe SSD
Transfer rate	Typically 100 MB/s - 200MB/s Up to ~500MB/s	Typically 400 MB/s - 550MB/s Up to 600MB/s	Typically 3,000 MB/s-5,000 MB/s Up to 7,000 MB/s
Latency (4kb)	~10 ms	~200 us	~60 us
Size	1TB - 16TB each Up to 20TB	500GB - 4TB each Up to 15TB	500GB - 4TB each Up to 15TB
Cost	Low	High	Somewhat same as SATA SSD

Testbed

3 x master nodes, 16 x DataNodes

Master nodes

CPU	2 x Intel(R) Xeon(R) Gold 6230 CPU @ 2.10GHz/20 cores
memory	384GB (12 x 32GB DDR4 @ 2933MHz)
OS Boot	Cisco Boot optimized M.2 Raid controller with 2 x 240GB SATA SSD
SSD	3.8TB SATA SSD Enterprise Value
Storage Controller	Cisco 12G Modular Raid Controller with 2GB cache
Network Adapter	Cisco UCS VIC 1387 2 x 40Gbps ports x8 PCIe Gen3

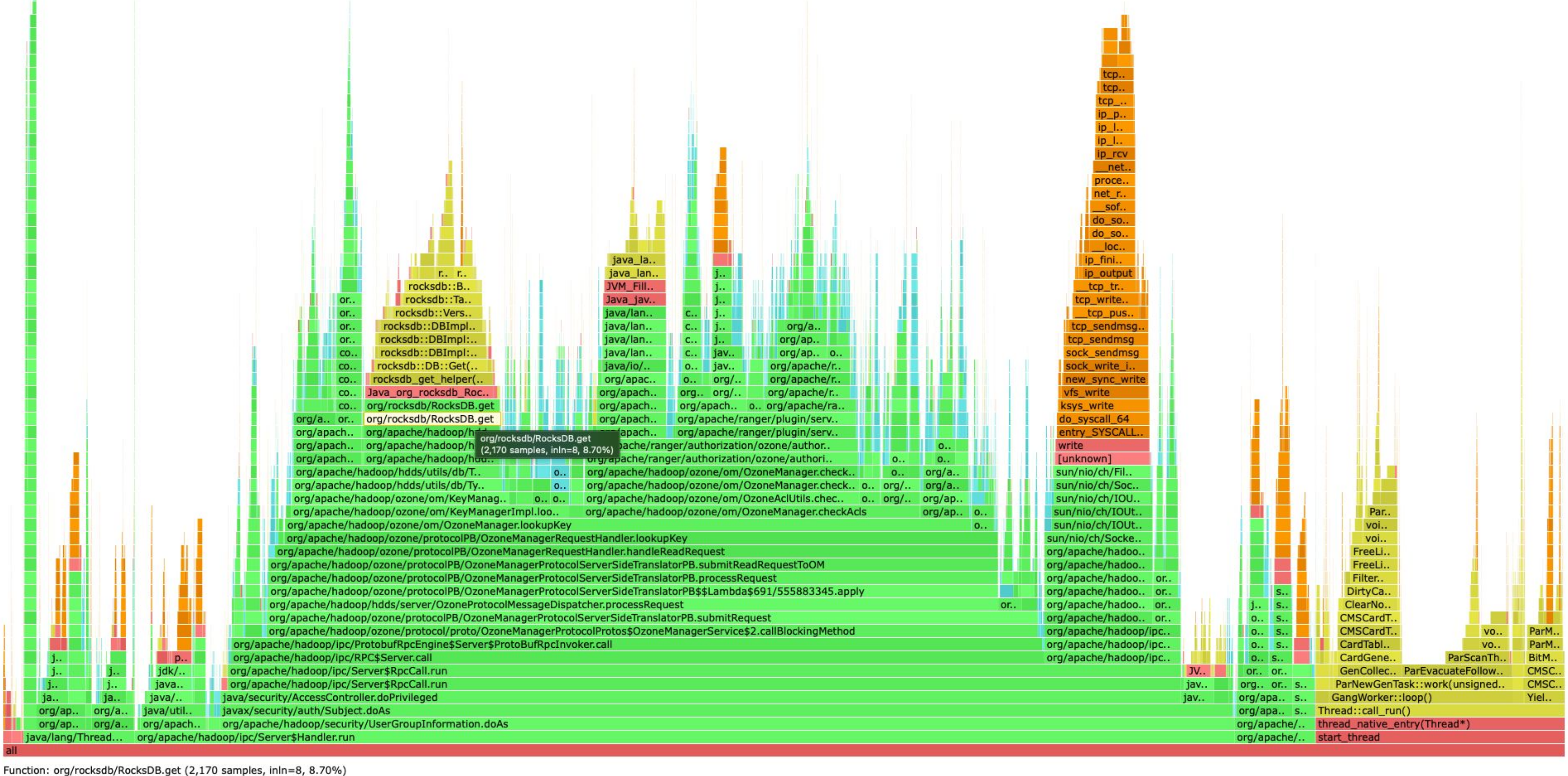
Data Nodes

CPU	2 x Intel(R) Xeon(R) Gold 6262V CPU @ 1.90GHz/24 cores
memory	384GB (12 x 32GB DDR4 @ 2933MHz)
OS Boot	Cisco Boot optimized M.2 Raid controller with 2 x 240GB SATA SSD
NVMe	10 x 8TB Intel P4510 U.2 High Performance Value
Storage Controller	NA
Network Adapter	Cisco UCS VIC 1387 2 x 40Gbps ports x8 PCIe Gen3

Tests conducted

- Freon read load post hard restart (minimal caching)
- Warp test to measure network saturation when using S3
- Impala TPCDS benchmark
- Ratis streaming performance tests

How much does disk read cost with NVME?



Impala TPCDS

Why Impala and Ozone?

- Data Warehouse is the most common use case. ($$$)
- Impala historically optimized on HDFS -> what will it do on Ozone

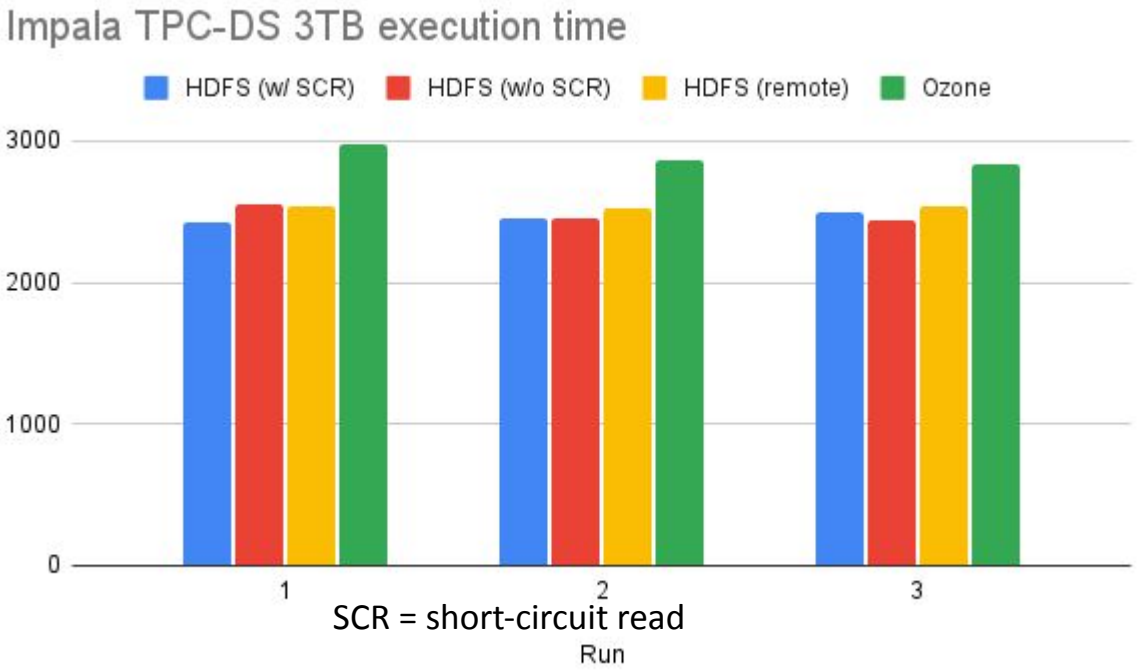
Software under test

CDP Private Cloud Base 7.1.8 +

- IMPALA-11457 Fix regression with unknown disk id
- HDDS-4970 Significant overhead when DataNode is over-subscribed
- HDDS-7135 ofs file input stream should support StreamCapabilities interface
- HDDS-7136 Memory leak due to ChunkInputStream.close() not releasing buffer
- HDDS-7161 Make Checksum.int2ByteString() zero-copy

All fixes are upstreamed in Apache Ozone 1.3.0 + Apache Impala 4.1.1

Ozone has a small overhead compared to HDFS (13% more than HDFS, and 12% more than remote HDFS).



SCR = short-circuit read

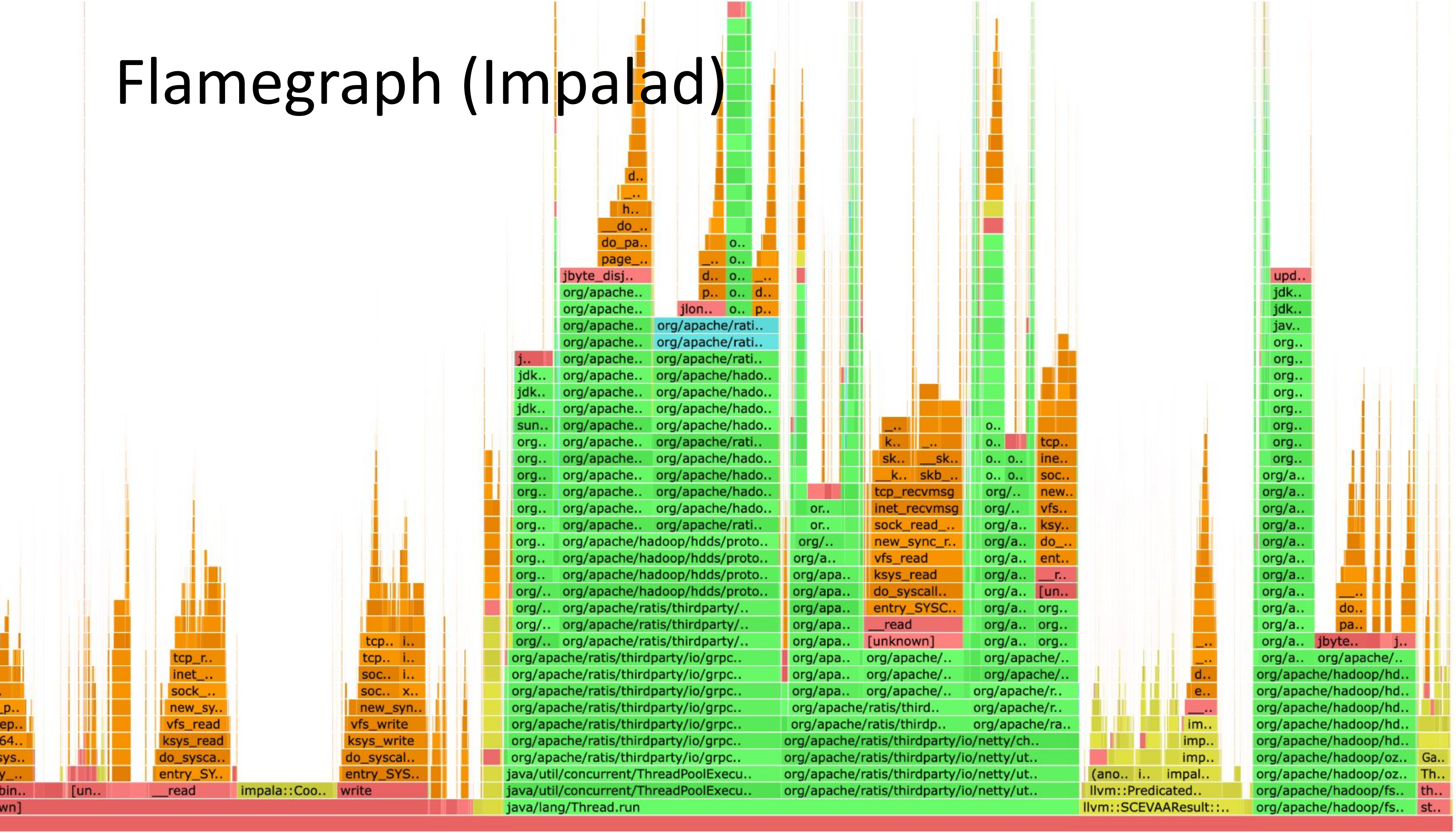
remote = REPLICA_PREFERENCE=REMOTE

Ozone has a small overhead compared to HDFS (5% more than HDFS).

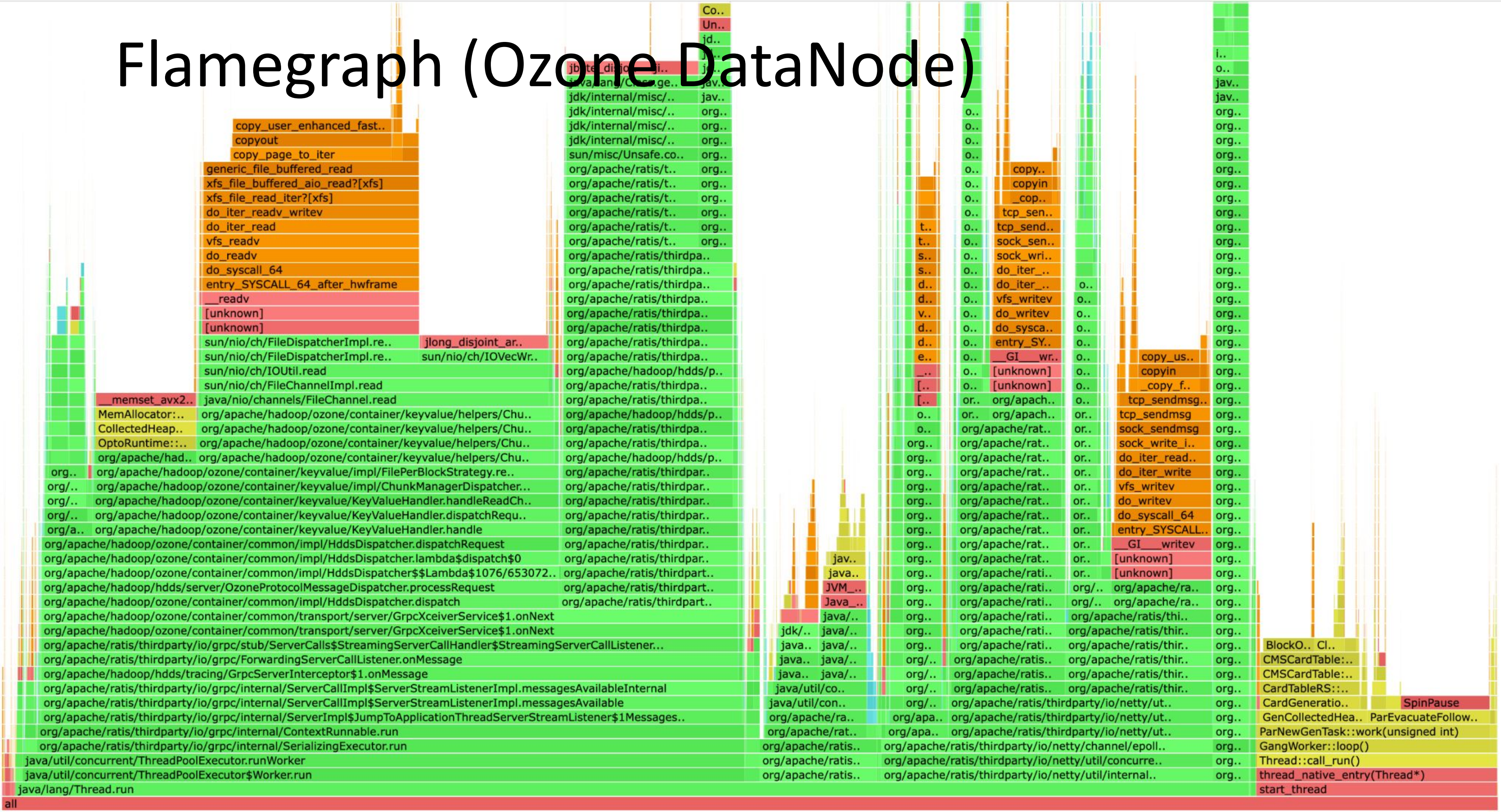
SCR = short-circuit read

remote = REPLICA_PREFERENCE=REMOTE

Flamegraph (Impalad)



Flamegraph (Ozone DataNode)



Lesson Learned

Too many rocksdb instances is bad

One RocksDB to manage the metadata of a 5GB container

But a DataNode can be up to a few hundred TB → 100k rocksdb instances.

Very slow to load (HDDS-3892, HDDS-4427, HDDS-4488, HDDS-5785)

Error prone (HDDS-5756/rocksdb issue:8617)

→ HDDS-3630 (Merge rocksdb in datanode)

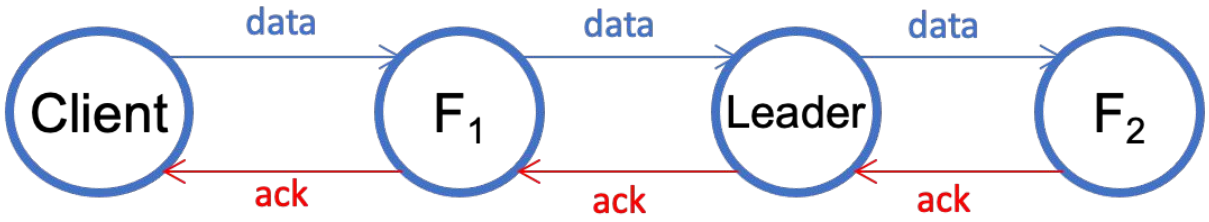
- One rocksdb instance to manage the containers of a disk

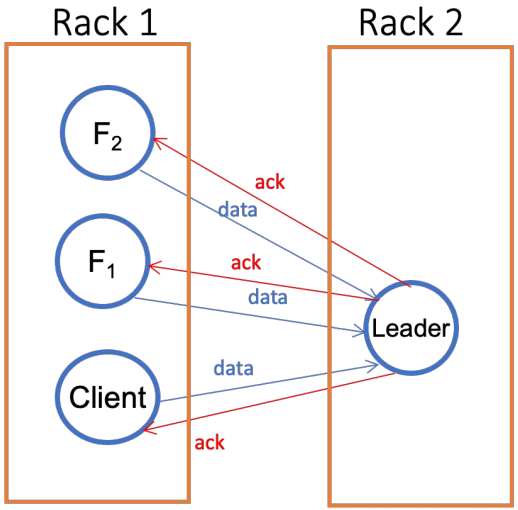
Write path improvements in the pipeline

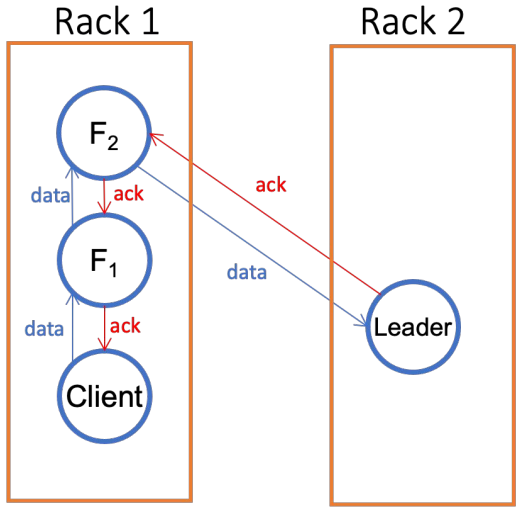
with Ratis Streaming API (RATIS-979)

- The Leader does not get more traffic
  - It is no longer the performance bottleneck.
- Better network topology awareness
  - Client writes to the **closest datanode** instead of the Leader
- Netty zero buffer copy
  - No gRPC buffer problem

Ratis streaming



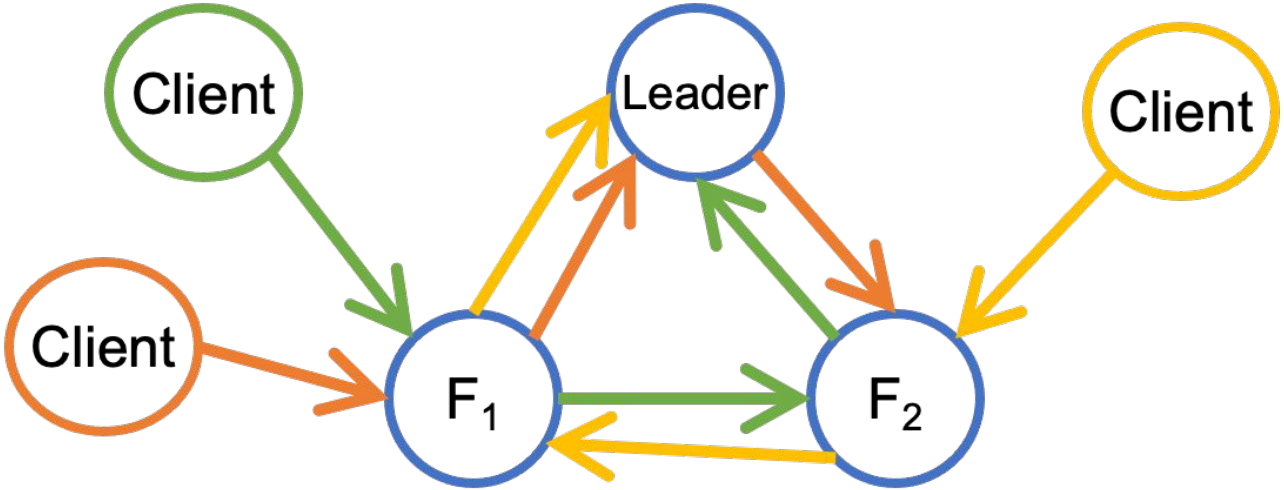




Benchmark – Observation

V1 (Async API) vs V2 (Streaming API)

- **V2 Streaming** multiple-client cases can be **~3x** of **V1 Async**
  - Streaming can use the full power of all **three** datanodes.



Performance roadmap ahead

1. Ratis streaming merge
2. OM Performance improvements
3. DN saturation of network
4. Better leveraging benefits of NVME
    a. Squeezing every bit of latency from each request processing
    b. Better caching architectures from computation down to disk to leverage HW.

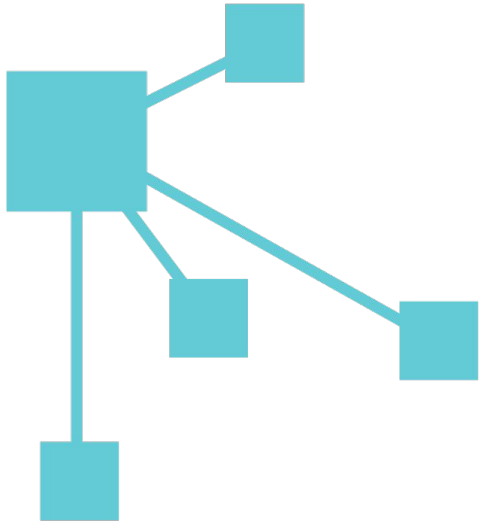
Conclusion

- Ozone architecturally addresses scale issues
- Hardware trends in the right direction for Ozone architecture.
  - NVME for Ozone Manager
  - High density datanodes with higher node counts
- Tests validate the architecture and direction for Ozone.

Acknowledgement

Cisco

Apache Ozone and Ratis communities



The unexpected: JDK performance problems

JDK lock contention JDK-7092821 (resolved in JDK 8u241 and 11u07)

Token verification (SHA256withRSA) HDDS-7256

Contributions welcome!

github.com/apache/ozone/

Questions?