



Performance of Apache Ozone on NVMe

Wei-Chiu Chuang (jojochuang) Ritesh Shukla (kerneltime)

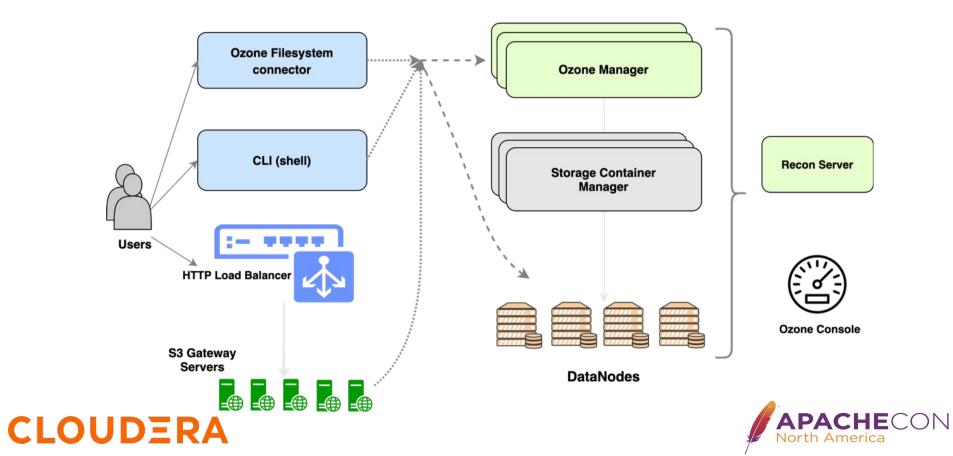
Agenda

- 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

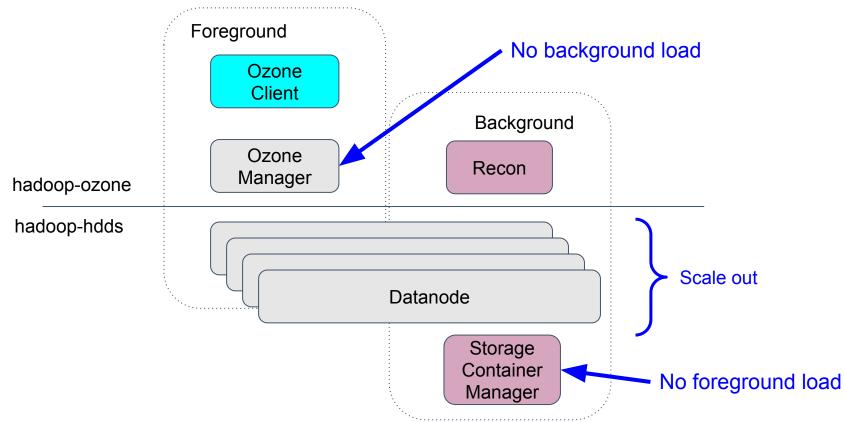




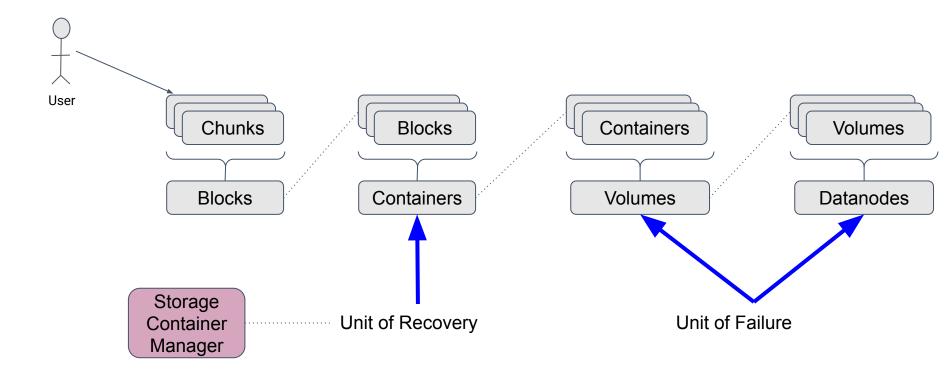
Ozone Architecture



Why does Ozone Scale? Separation of concerns



Why does Ozone Scale? Aggregation via containers



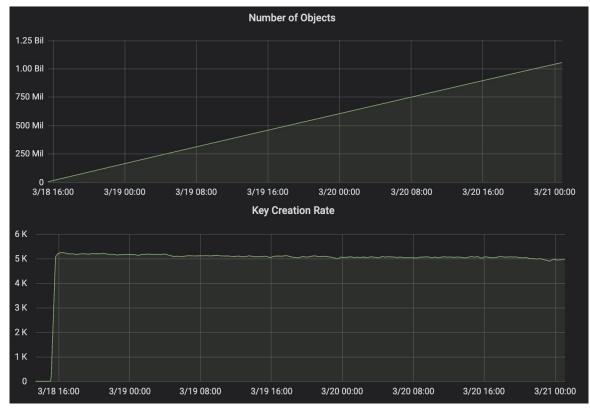
Why does foreground Scale?

- 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



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

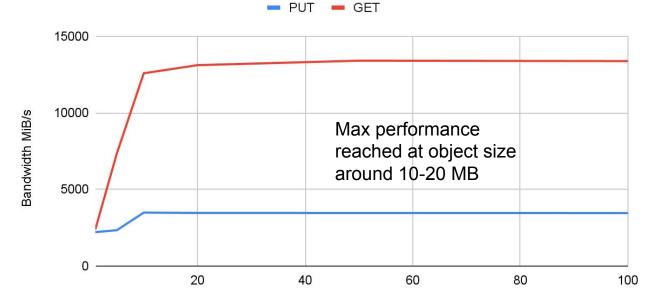
Capability	Ozone	HDFS
Storage Density	1000's of nodes at 600TB per node	1000's of nodes at 100TB per node
Scalability	10B Objects	400M Objects
Recovery	Fast recovery (< 5 min restart)	Slow startup based on size
High Availability	Active - Active	Active - Standby
Protocol Support	Hadoop / S3 API	Hadoop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

CLOUDERA



Small objects are welcome

PUT/GET Throughput 8 Datanodes 8 Clients 20 Threads



Size of Object MiB





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

CLOUDERA

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 PCle 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?

j j j j j j j j	P jāvā/ jāvā/ jāva/ jāpach	Bite Jack Bite Jack Bite Jack Bite Jack Bite Jack Bite Jack Bite Jack Bite Jack	org/apa s org/apa s Th org/apache/ th	nread_native_entry(Thread*)	ParM ParM BitM BitM CMSC CMSC CMSC Yiel
java/lang/Thread		oop/ipc/Server\$Handler.run		tart_thread	

Function: org/rocksdb/RocksDB.get (2,170 samples, inln=8, 8.70%

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 +

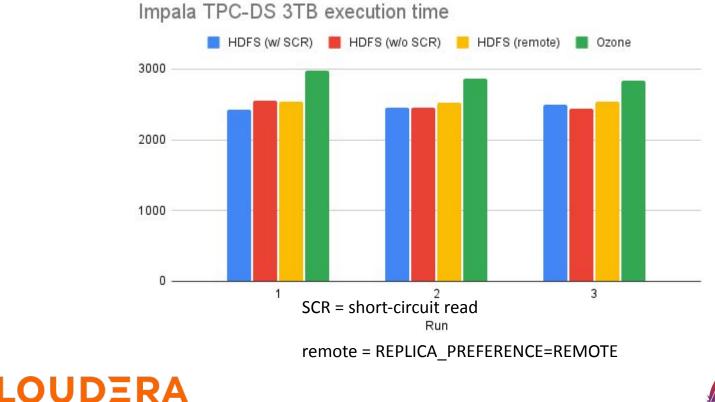
- IMPALA-11457 Fix regression with unknown disk id
- HDDS-4970 Significant overhead when DataNode is over-subscribed
- HDDS-7135 ofs file input stream should support StreamCapabilities interface
- HDDS-7136 Memory leak due to ChunkInputStream.close() not releasing buffer
- HDDS-7161 Make Checksum.int2ByteString()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

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



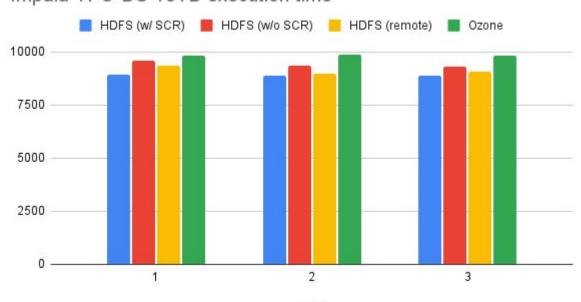


Ozone has a small overhead compared to HDFS (13% more than HDFS, and 12% more than remote HDFS).





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



Run

SCR = short-circuit read

C

LOUDERA

remote = REPLICA_PREFERENCE=REMOTE



Flamegraph	(Impalad)		
tcp_r. inet sock new_sy new_sy vfs_read vfs_vrad vfs_write do_sysca entry_SY. entry_SY.	d. d. d. d. d. d. d. d. d. d.	 	 upd jdk jdk jdk jav org org/a org/apache/hadoop/hd org/apache/hadoop/doz org/apache/hadoop/foz org/apache/hadoop/foz org/apache/hadoop/foz org/

p.. ep.. 64.. sys.. y... bin.. wn]

Flamegraph (Oz		taNc	ode)		i 0		
	jdk/internal/misc/ jav			- 📕 📔		jav jav		
copy_user_enhanced_fast.	jdk/internal/misc/ org jdk/internal/misc/ org			0		org		
copyout	jdk/internal/misc/ org			0		org		
copy_page_to_iter	sun/misc/Unsafe.co org			0		org		
generic_file_buffered_read	org/apache/ratis/t org			o copy		org		
xfs_file_buffered_aio_read?[xfs] xfs_file_read_iter?[xfs]	org/apache/ratis/t org org/apache/ratis/t org			o copyin		org		
do_iter_ready_writev	org/apache/ratis/t org			o tcp_sen		org		
do_iter_read	org/apache/ratis/t org		t	o tcp_send		org		
vfs_readv	org/apache/ratis/t org		t.,	o sock_sen		org		
do_readv	org/apache/ratis/thirdpa		S	o sock_wri		org		
do_syscall_64 entry_SYSCALL_64_after_hwframe	org/apache/ratis/thirdpa org/apache/ratis/thirdpa		d	o do_iter		org		
readv	org/apache/ratis/thirdpa		d	o vfs writev	0	org		
[unknown]	org/apache/ratis/thirdpa		V	o do_writev	0	org		
[unknown]	org/apache/ratis/thirdpa		d	o do_sysca	0	org		
sun/nio/ch/FileDispatcherImpl.re jlong_disjoint_ar	org/apache/ratis/thirdpa		d	o entry_SY	0	org		
sun/nio/ch/FileDispatcherImpl.re sun/nio/ch/IOVecWr sun/nio/ch/IOUtil.read	org/apache/ratis/thirdpa		е	oGIwr o [unknown]		org		
sun/nio/ch/FileChannelImpl.read	org/apache/hadoop/hdds/p org/apache/ratis/thirdpa		 [o [unknown]	o copyin	org		
memset_avx2 java/nio/channels/FileChannel.read	org/apache/ratis/thirdpa		[or org/apach	o tcp_sendmsg.	org		
MemAllocator: org/apache/hadoop/ozone/container/keyvalue/helpers/Chu	org/apache/hadoop/hdds/p		0	or org/apach	or tcp_sendmsg	org		
CollectedHeap org/apache/hadoop/ozone/container/keyvalue/helpers/Chu	org/apache/ratis/thirdpa		0	org/apache/rat	or sock_sendmsg	org		
OptoRuntime:: org/apache/hadoop/ozone/container/keyvalue/helpers/Chu	org/apache/ratis/thirdpa		org	org/apache/rat	or sock_write_i	org		
org/apache/had org/apache/hadoop/ozone/container/keyvalue/helpers/Chu	org/apache/hadoop/hdds/p		org	org/apache/rat	or do_iter_read	org		
org org/apache/hadoop/ozone/container/keyvalue/impl/FilePerBlockStrategy.re org/ org/apache/hadoop/ozone/container/keyvalue/impl/ChunkManagerDispatcher	org/apache/ratis/thirdpar org/apache/ratis/thirdpar		org	org/apache/rat org/apache/rat	or do_iter_write or vfs writev	org		
org/ org/apache/hadoop/ozone/container/keyvalue/KeyValueHandler.handleReadCh	org/apache/ratis/thirdpar.		org	org/apache/rat	or do_writev	org		
org/org/apache/hadoop/ozone/container/keyvalue/KeyValueHandler.dispatchRequ	org/apache/ratis/thirdpar.		org	org/apache/rat	or do_syscall_64	org		
org/a org/apache/hadoop/ozone/container/keyvalue/KeyValueHandler.handle	org/apache/ratis/thirdpar.		org	org/apache/rat	or entry_SYSCALL.			
org/apache/hadoop/ozone/container/common/impl/HddsDispatcher.dispatchRequest	org/apache/ratis/thirdpar.		org	org/apache/rat	orGIwritev	org		
org/apache/hadoop/ozone/container/common/impl/HddsDispatcher.lambda\$dispatch\$0	org/apache/ratis/thirdpar.	jav	org	org/apache/rat	or [unknown]	org		
org/apache/hadoop/ozone/container/common/impl/HddsDispatcher\$\$Lambda\$1076/653072 org/apache/hadoop/hdds/server/OzoneProtocolMessageDispatcher.processRequest	org/apache/ratis/thirdpart org/apache/ratis/thirdpart	java JVM	org org	org/apache/rati org/apache/rati	or [unknown] org/ org/apache/ra	org		
org/apache/hadoop/rozone/container/common/impl/HddsDispatcher.dispatch	org/apache/ratis/thirdpart	Java	org	org/apache/rati	org/ org/apache/ra	org		
org/apache/hadoop/ozone/container/common/transport/server/GrpcXceiverService\$1.onNex		java/	org	org/apache/rati	org/apache/ratis/thi	org		
org/apache/hadoop/ozone/container/common/transport/server/GrpcXceiverService\$1.onNex	dt 👘	jdk/ java/	org	org/apache/rati	org/apache/ratis/thir.	org		
org/apache/ratis/thirdparty/io/grpc/stub/ServerCalls\$StreamingServerCallHandler\$Streamin	gServerCallListener	java java/	org	org/apache/rati	org/apache/ratis/thir.	org	BlockO., Cl.,	
org/apache/ratis/thirdparty/io/grpc/ForwardingServerCallListener.onMessage		java java/	org/	org/apache/ratis	org/apache/ratis/thir.	org	CMSCardTable:	
org/apache/hadoop/hdds/tracing/GrpcServerInterceptor\$1.onMessage org/apache/ratis/thirdparty/io/grpc/internal/ServerCallImpl\$ServerStreamListenerImpl.mess	sages Available Internal	java java/ java/util/co	org/ org/	org/apache/ratis org/apache/ratis	org/apache/ratis/thir org/apache/ratis/thir	org	CMSCardTable:	
org/apache/ratis/thirdparty/io/grpc/internal/ServerCallImpl\$ServerStreamListenerImpl.mess		java/util/con	org/	org/apache/ratis/thir		org	CardGeneratio	SpinPause
org/apache/ratis/thirdparty/io/grpc/internal/ServerImpl\$JumpToApplicationThreadServerStre		org/apache/ra		org/apache/ratis/thin		org		ParEvacuateFollow
org/apache/ratis/thirdparty/io/grpc/internal/ContextRunnable.run		org/apache/rat		org/apache/ratis/thin		org	ParNewGenTask::wo	ork(unsigned int)
org/apache/ratis/thirdparty/io/grpc/internal/SerializingExecutor.run		org/apache/ratis		/ratis/thirdparty/io/ne		org	GangWorker::loop())
java/util/concurrent/ThreadPoolExecutor.runWorker		org/apache/ratis		/ratis/thirdparty/io/ne		org	Thread::call_run()	(Thread *)
java/util/concurrent/ThreadPoolExecutor\$Worker.run java/lang/Thread.run		org/apache/ratis	org/apache	/ratis/thirdparty/io/ne	etty/util/internal	org	thread_native_entry start_thread	y(mead*)
all							ocon_cmedu	

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 \rightarrow 100k rocksdb instances.

Very slow to load (<u>HDDS-3892</u>, <u>HDDS-4427</u>, <u>HDDS-4488</u>, <u>HDDS-5785</u>)

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

 \rightarrow <u>HDDS-3630</u> (Merge rocksdb in datanode)

One rocksdb instance to manage the containers of a disk
 CLOUDERA

Write path improvements in the pipeline

with Ratis Streaming API (RATIS-979)

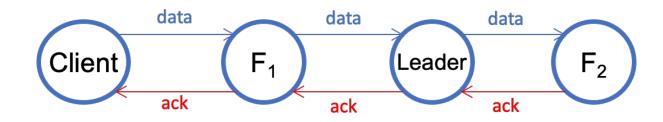
- The Leader does not get more traffic
 - It is no longer the performance bottleneck.
- Better network topology awareness
 - Client writes to the **closest datanode** instead of the Leader
- Netty zero buffer copy

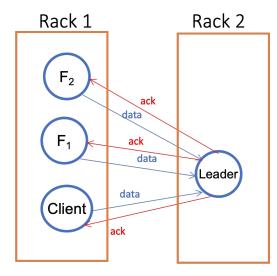
 \odot No gRPC buffer problem

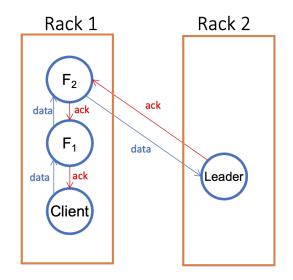




Ratis streaming





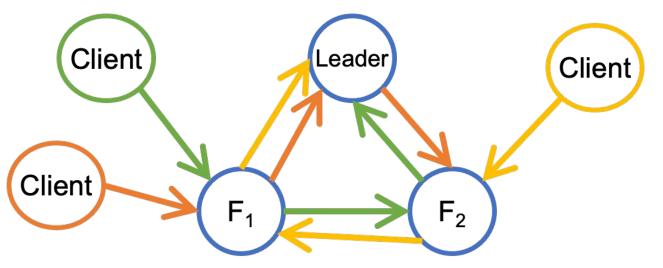


Benchmark – Observation

V1 (Async API) vs V2 (Streaming API)

• V2 Streaming multiple-client cases can be ~3x of V1 Async

 \odot 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.

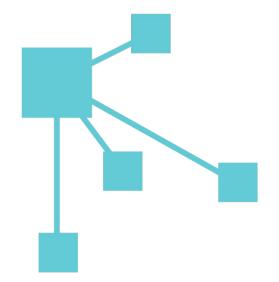




Acknowledgement

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?



