DAOS Async API Support & Performance Tunning in Spark

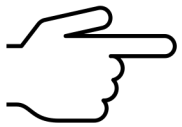
Jiafu Zhang

Nov. 2021

Agenda

- Spark DAOS Overview
- DAOS Async API Support
- Performance Tunning in Spark

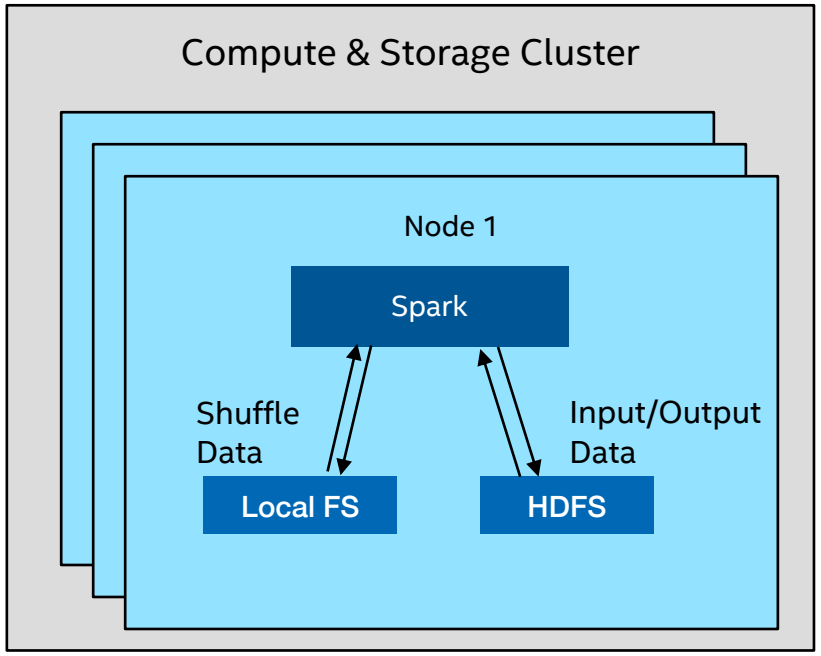
Agenda



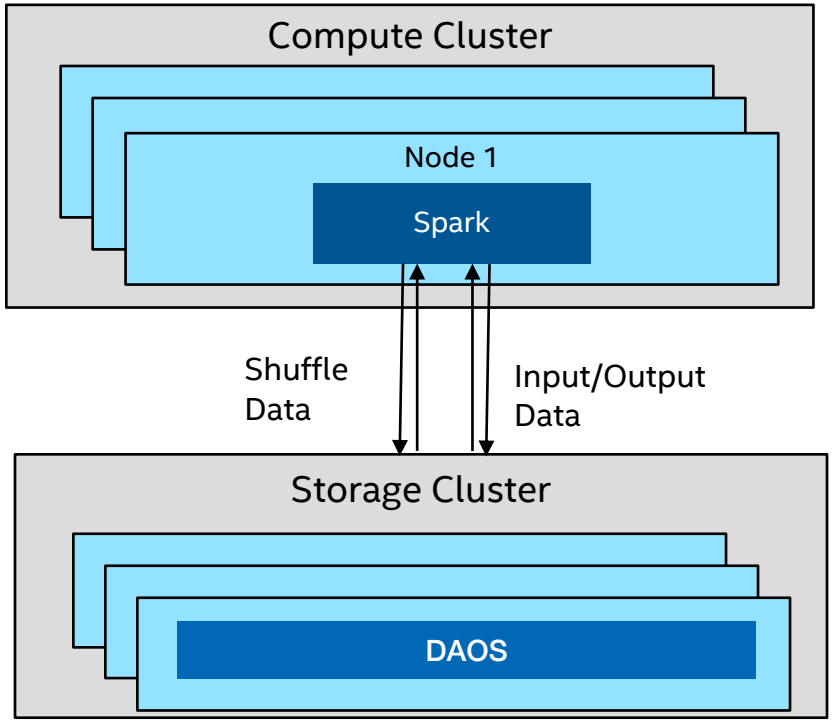
- Spark DAOS Overview
- DAOS Async API Support
- Performance Tunning in Spark

Enable and Accelerate Spark with DAOS

- Spark Input/Output Storage: From HDFS to DAOS
- Spark Shuffle Data Storage: From Local FS to DAOS

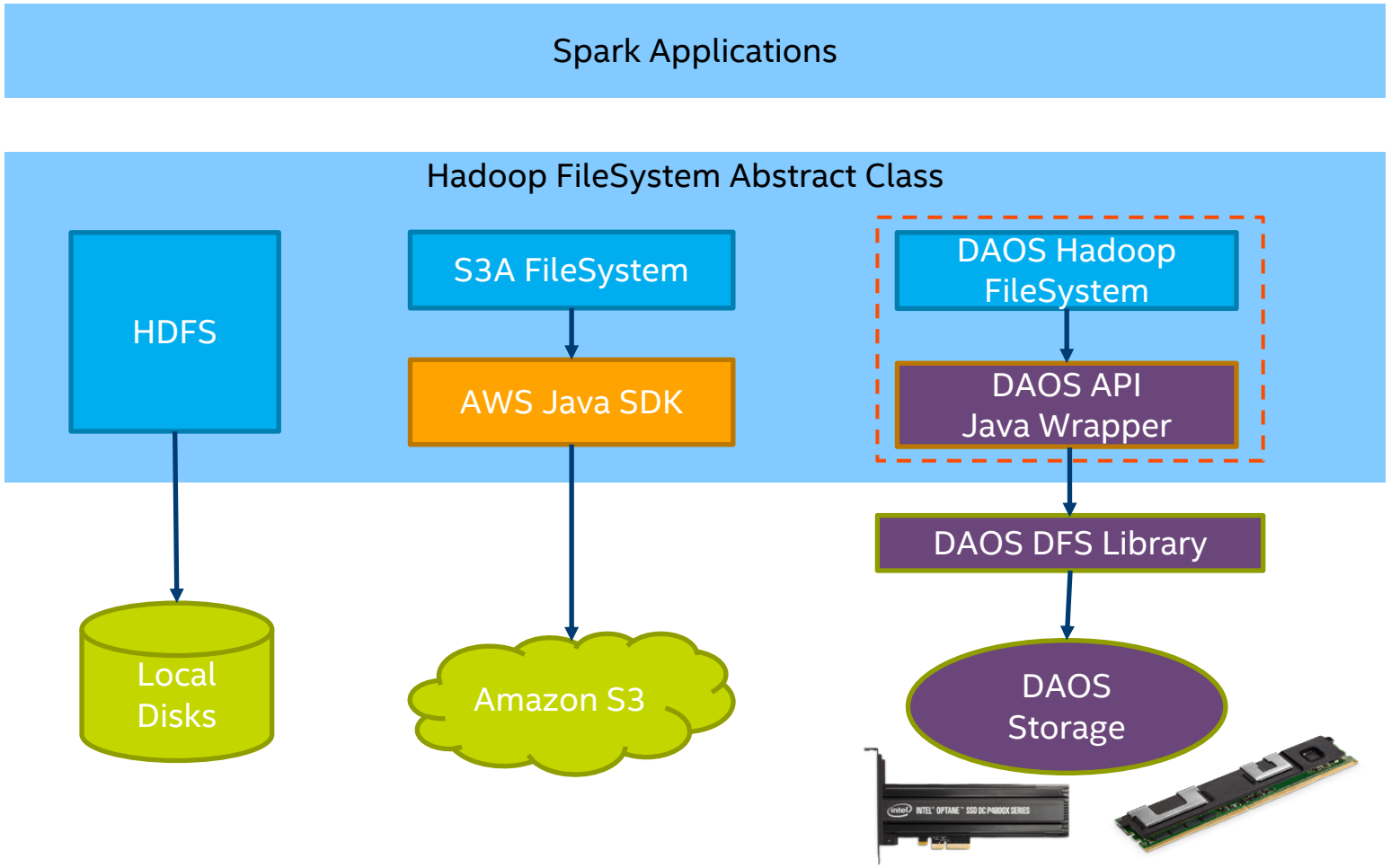


Colocated Cluster

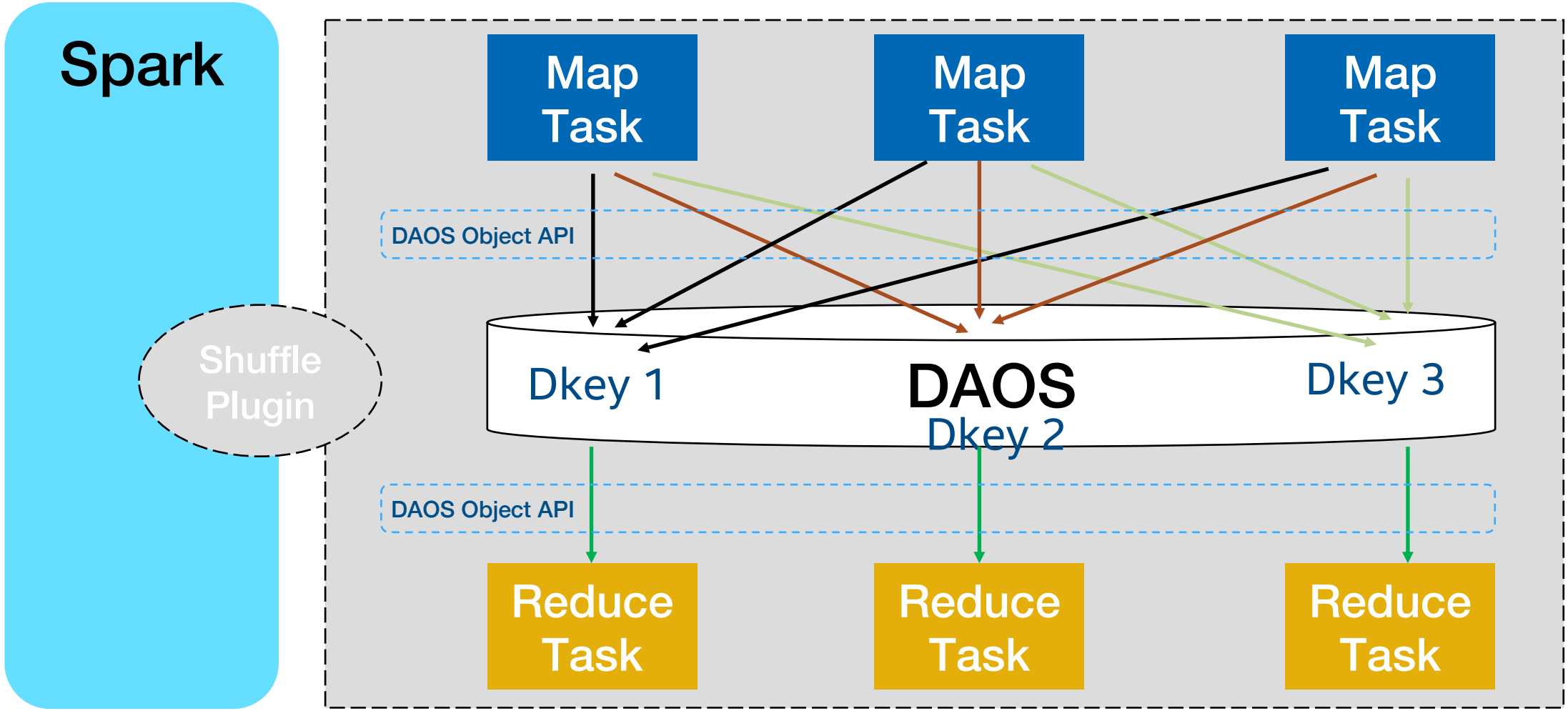


Disaggregated Cluster

DAOS Hadoop Filesystem Interface

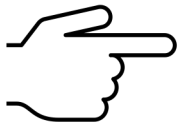


Spark Shuffle to DAOS



Agenda

- Spark DAOS Overview



- DAOS Async API Support

- Performance Tunning in Spark

DAOS Event Queue

❑ Event-driven

A task or operation is bound to specific reusable event belonging to an Event Queue from which we can later poll completeness of the task or operation.

❑ One Event Queue per Process

For some platforms, like JVM, tasks run in threads. If there are large number of threads in one process, it could be a bottle-neck for some NIC, like Intel OPA.

DAOS Async API Design in Java

❑ Java Wrapper

To use DAOS async API, we need to access native Event Queue and Event. Thus, some Java wrapper classes are provided first for both Hadoop DAOS and Remote Shuffle Plugin.

DAOS Async API Design in Java

❑ One Event Queue per Thread instead of JVM Process

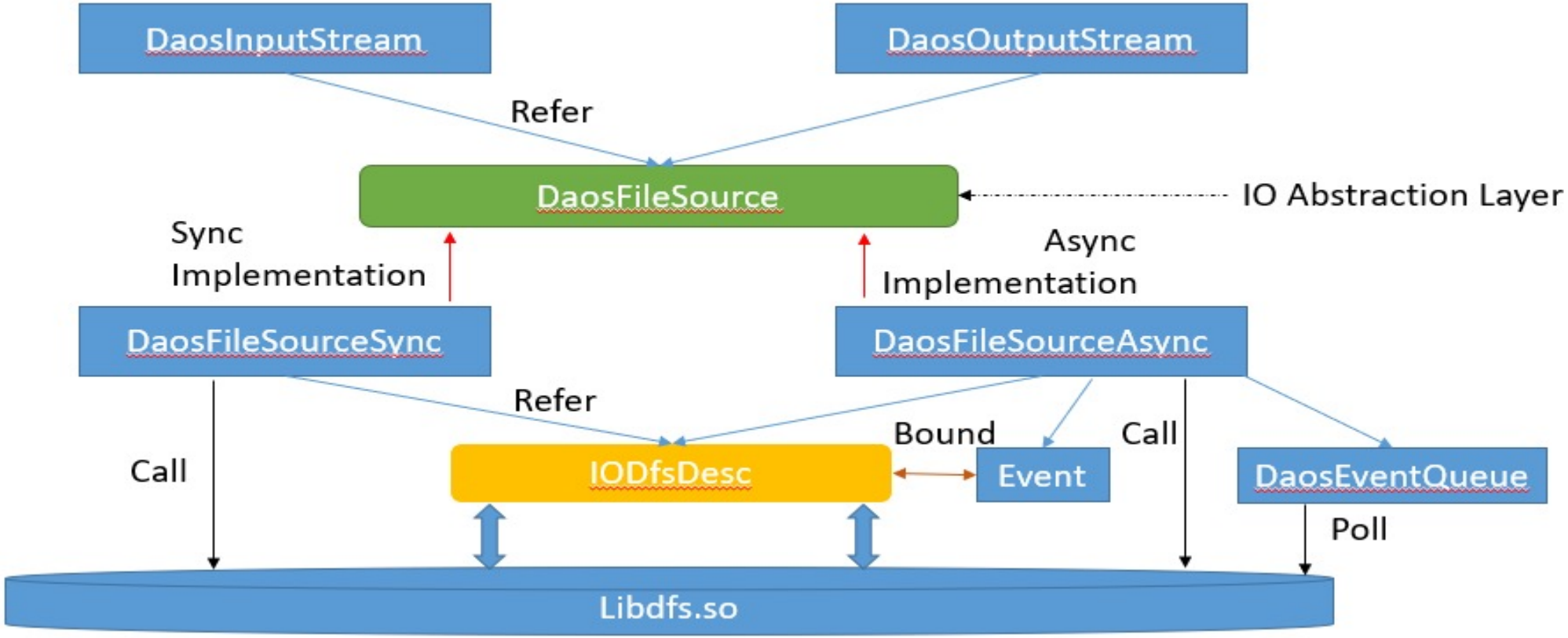
1. Avoid potential bottle-neck
2. Avoid memory sync for async objects among threads
3. Configurable number of events per EQ
4. Intel OPA test

Case: Read 256 files with 1 GB each in workload dfsioe using one executor (process) with 30 threads.

Async Read	Sync Read
1144 MB/s	24.4 MB/s

Hadoop DAOS Impl.

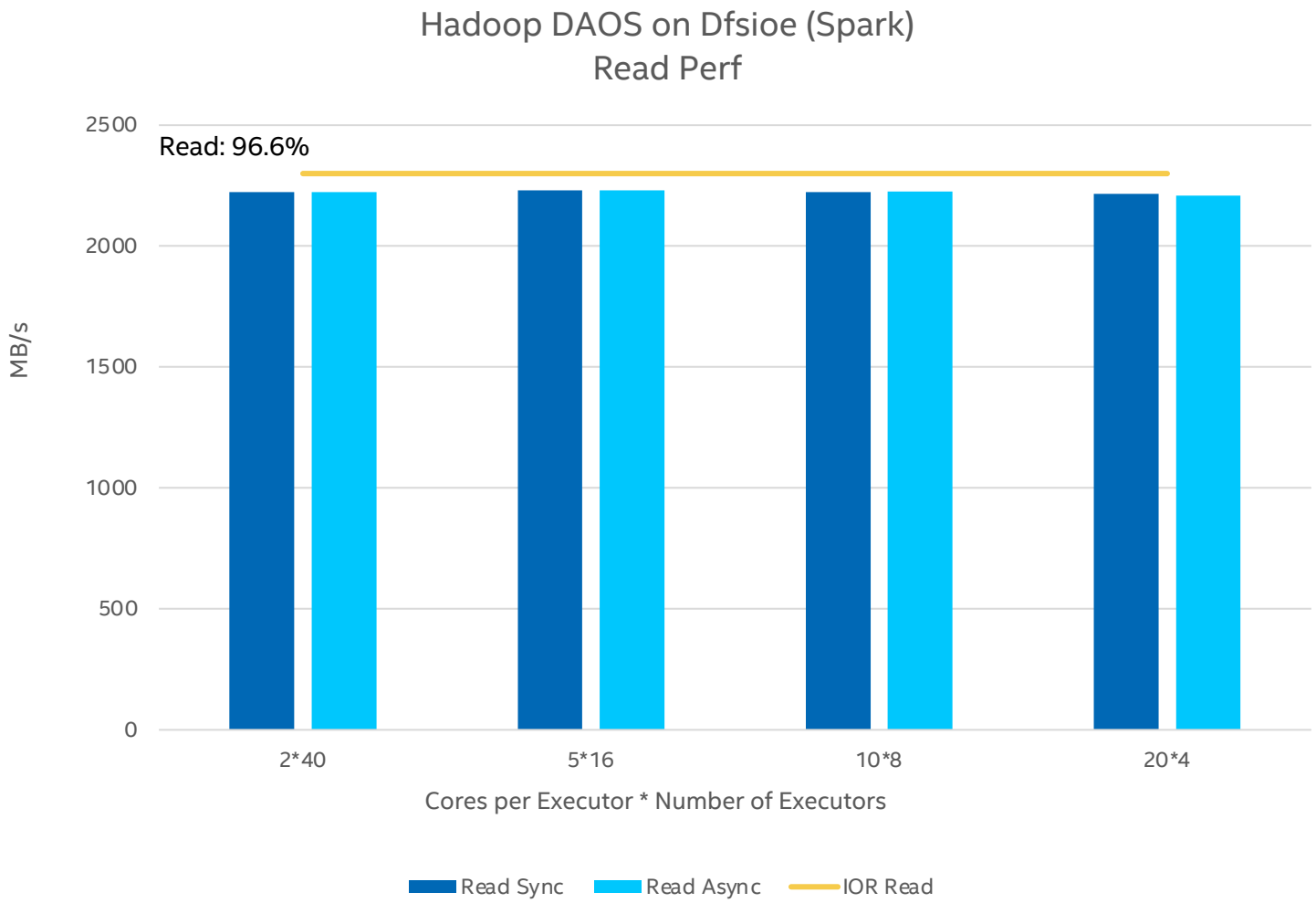
- Refactor Code to Abstract IO Layer and Sync/Async Impl



Hadoop DAOS Test

❑ Performance (Read)

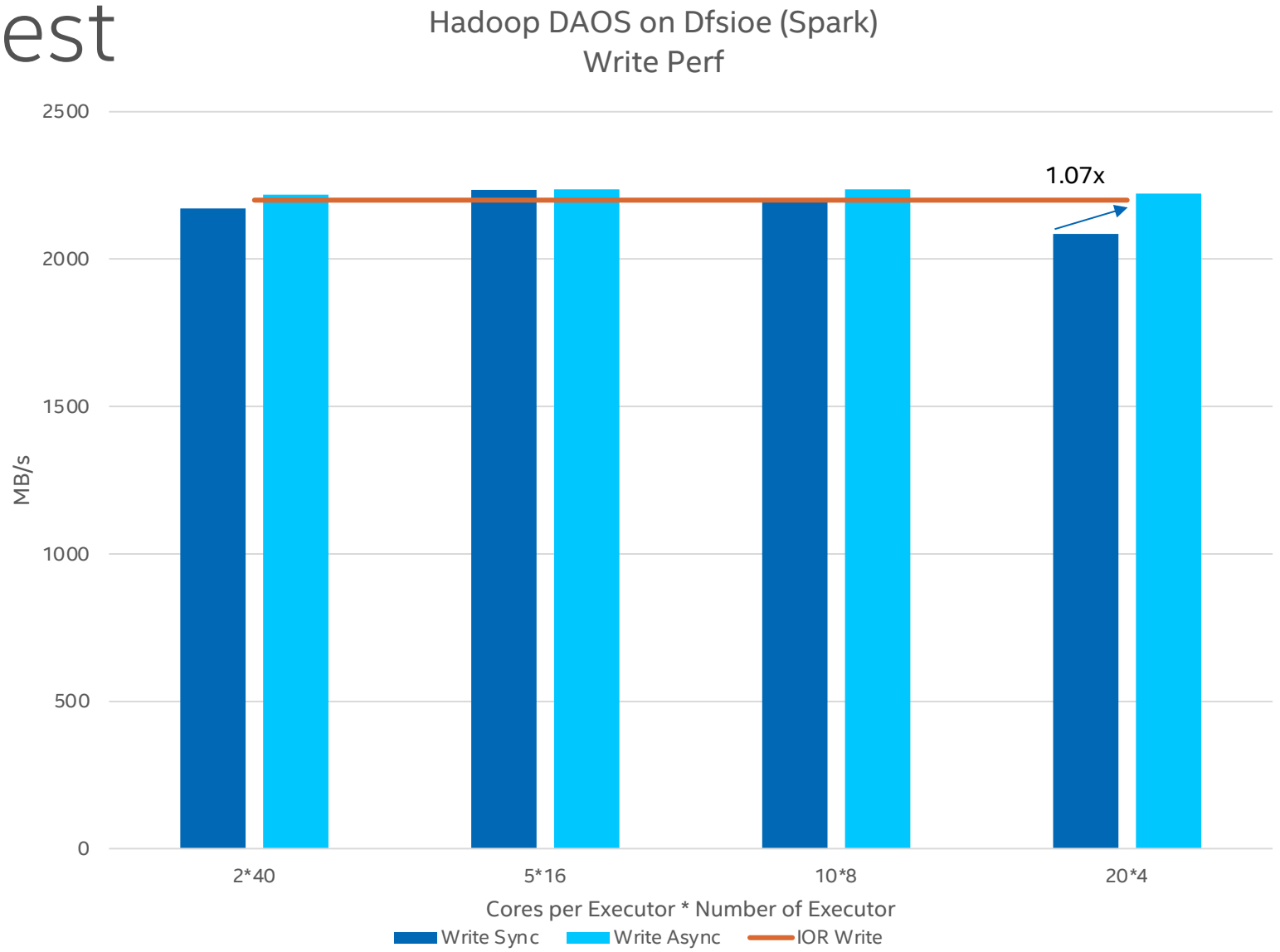
Sync and async performs equally, about 96.6% of roofline in average.



Hadoop DAOS Test

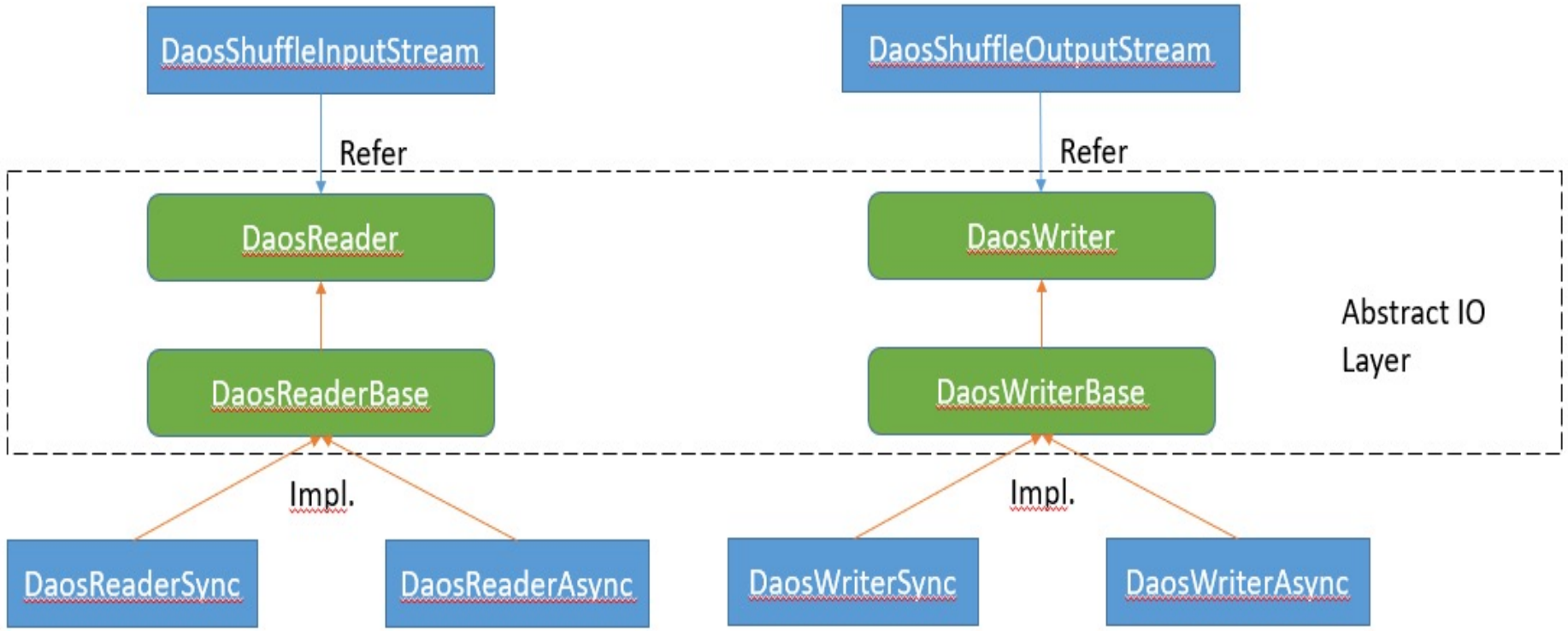
❑ Performance (Write)

Async is 1.07x of sync write when 20 cores per executor.



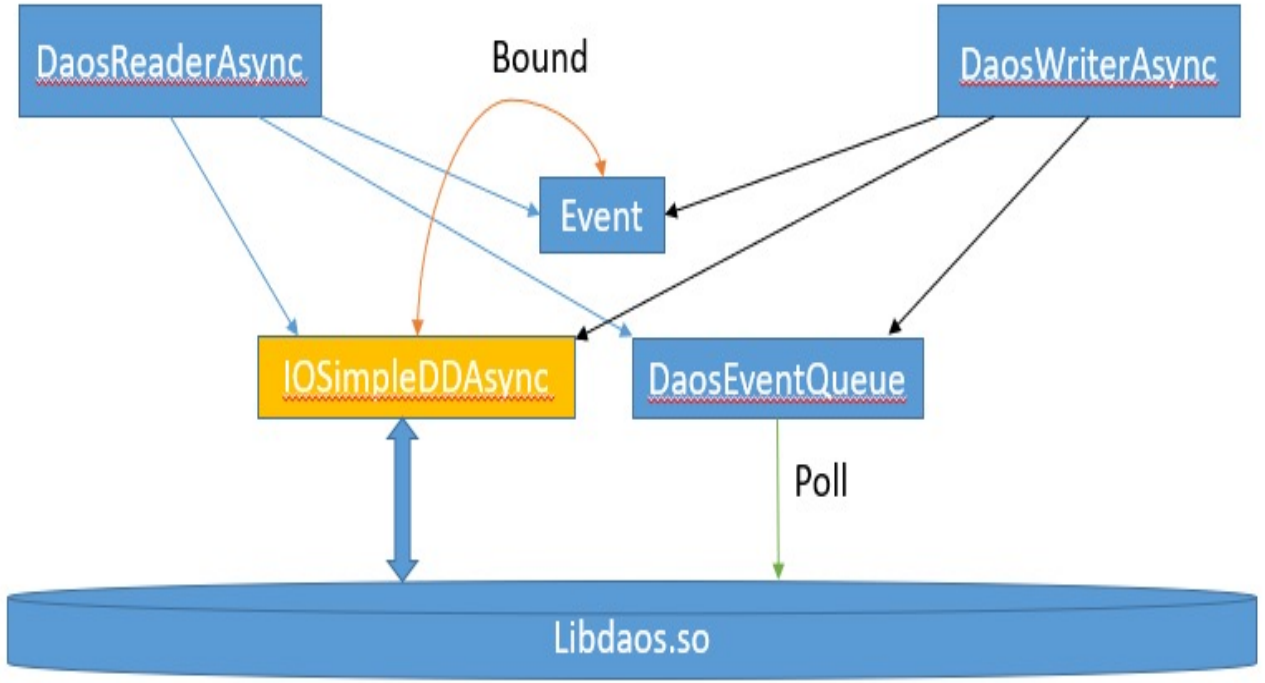
Shuffle Read/Write Impl.

- Abstract IO Layer



Shuffle Read/Write Impl.

- Async Impl.

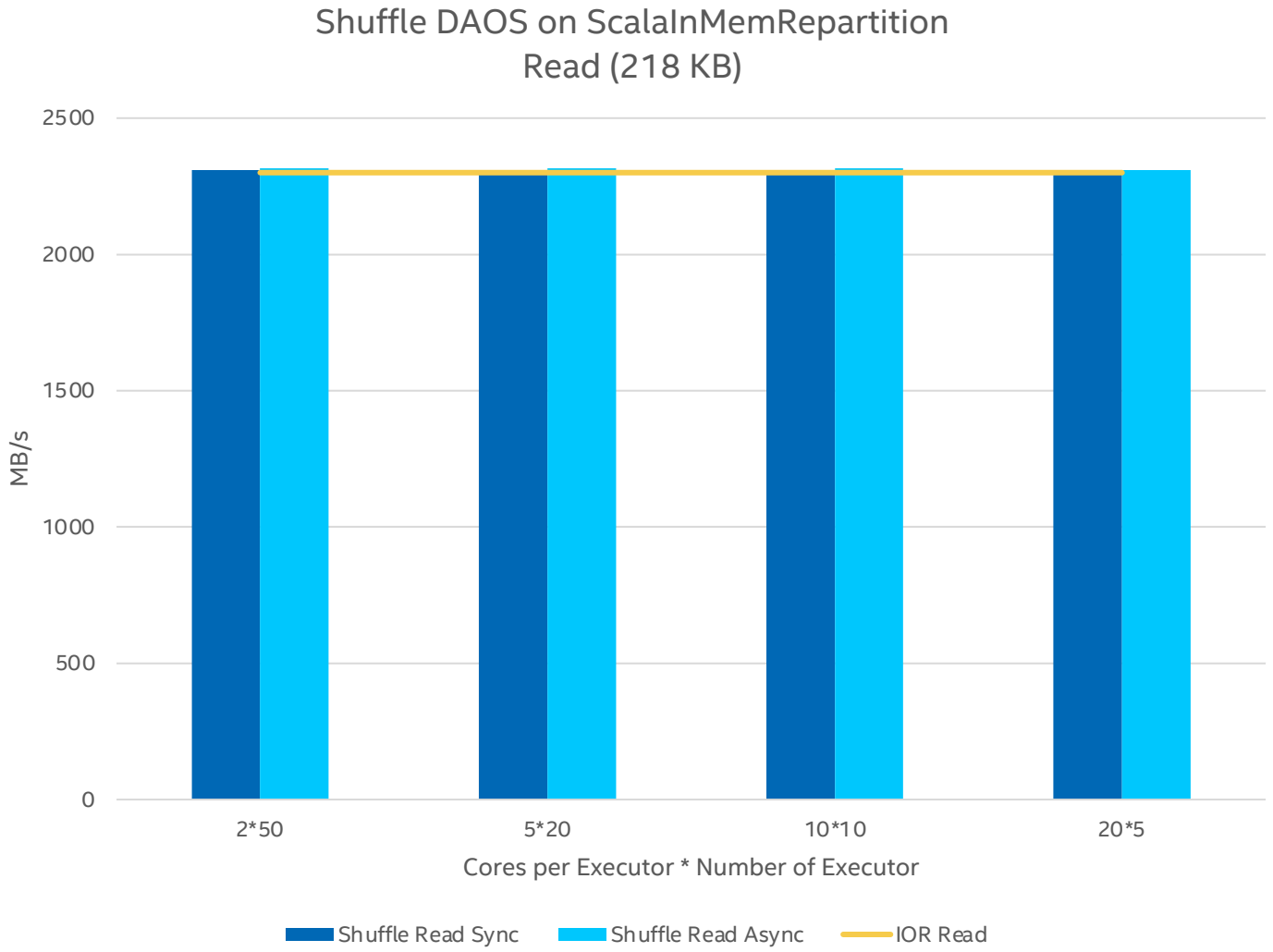


Shuffle Plugin Test

❑ Performance 1 (Read)

Shuffle Block Size	Total Shuffle Size
218 KB	106,496 MB

Both sync and async reach rootfline.

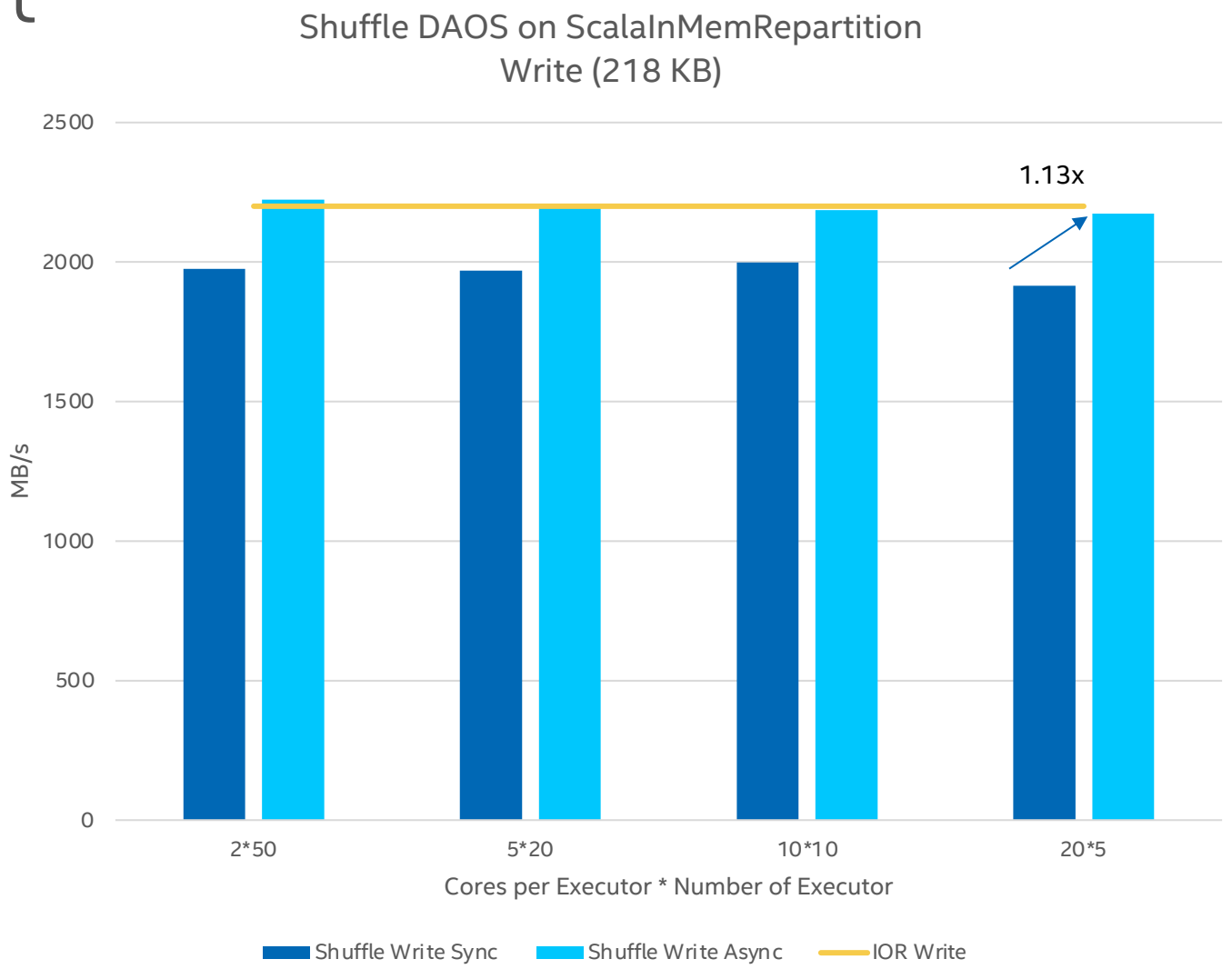


Shuffle Plugin Test

❑ Performance 1 (Write)

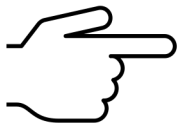
Shuffle Block Size	Total Shuffle Size
218 KB	106,496 MB

Async is about 1.13x of sync write when 20 cores per executor, 1.11x in average.



Agenda

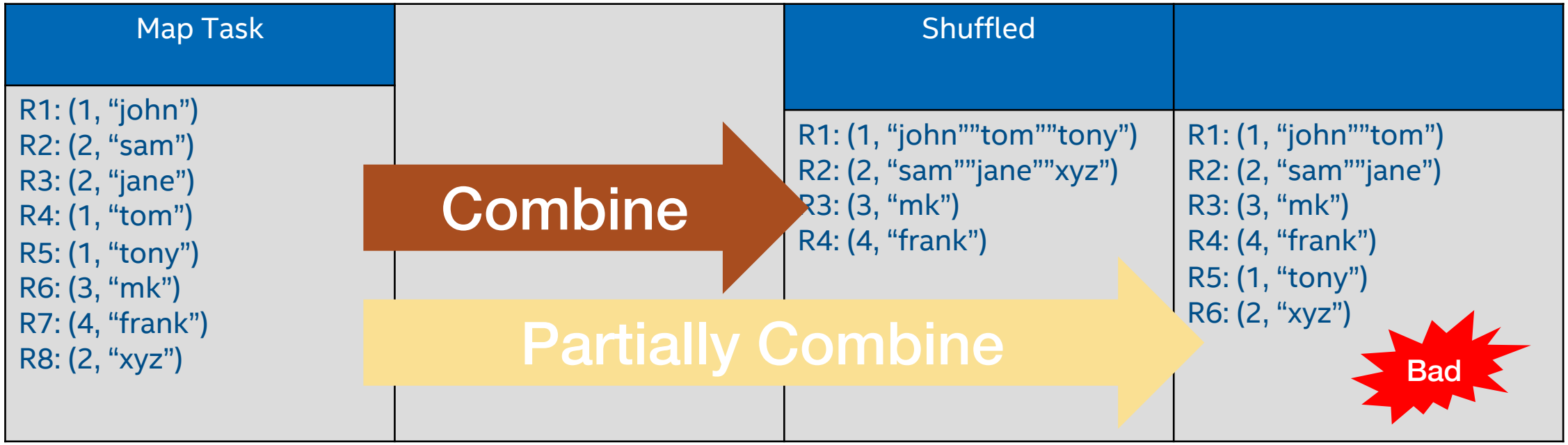
- Spark DAOS Overview
- DAOS Async API Support
- Performance Tunning in Spark



Performance Issue Identified with Map-side Combine Workload

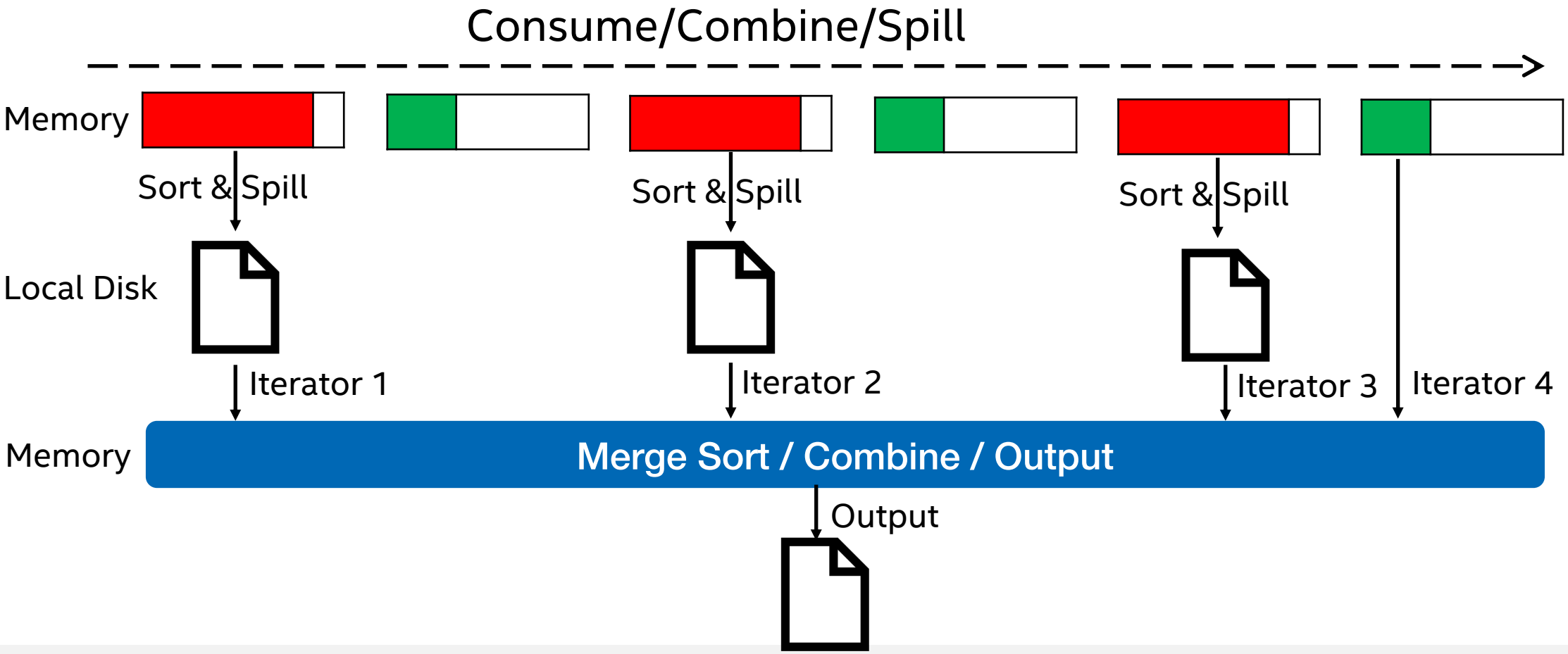
- Performance Issue - Partially Combined due to No Spill

Record Format: (key, value)



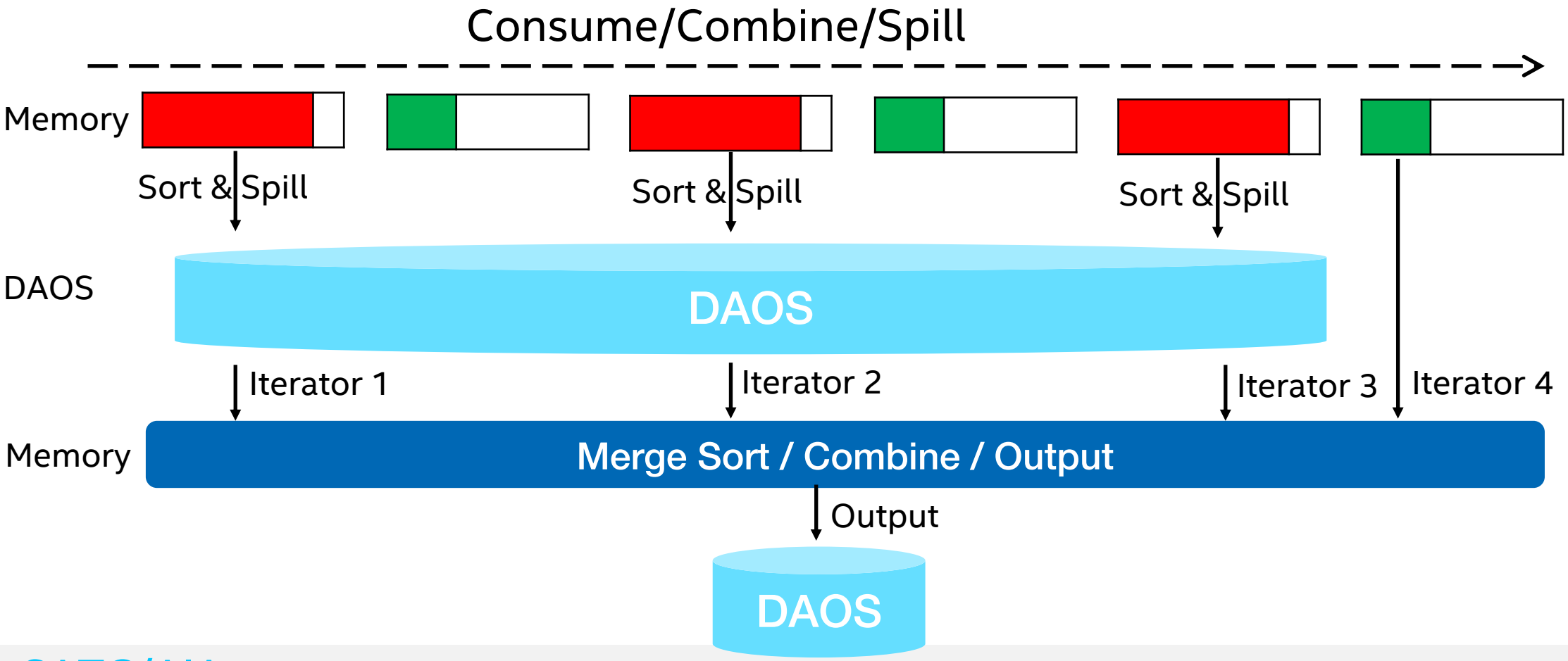
Spill Temporary Shuffle Records to Local Disk

- Spill When No More Memory Available



Spill Temporary Shuffle Records to DAOS

❑ Spill When No More Memory Available



Shuffle Write without Spill

Aggregated Metrics by Executor

Show 20 entries Search:

Executor ID	Logs	Address	Task Time	Total Tasks	Failed Tasks	Killed Tasks	Succeeded Tasks	Excluded	Input Size / Records	Shuffle Write Size / Records
0	stdout stderr	10.100.0.35:41395	7.5 min	16	0	0	16	false	15.3 GiB / 24693816	13.5 GiB / 15987097
1	stdout stderr	10.100.0.35:35365	7.5 min	15	0	0	15	false	15 GiB / 24135191	13.2 GiB / 15573058

Showing 1 to 2 of 2 entries Previous 1 Next

Tasks (31)

Show 20 entries Search:

Index	Task ID	Attempt	Status	Locality level	Executor ID	Host	Logs	Launch Time	Duration	GC Time	Input Size / Records	Shuffle Write Size / Records	Errors
0	0	0	SUCCESS	PROCESS_LOCAL	0	10.100.0.35	stdout stderr	2021-06-30 10:38:12	29 s	0.7 s	1 GiB / 1608841	901.6 MiB / 1038593	
1	1	0	SUCCESS	PROCESS_LOCAL	0	10.100.0.35	stdout stderr	2021-06-30 10:38:12	31 s	0.7 s	1 GiB / 1609061	901.5 MiB / 1034995	
2	2	0	SUCCESS	PROCESS_LOCAL	0	10.100.0.35	stdout stderr	2021-06-30 10:38:12	31 s	0.7 s	1 GiB / 1608762	901.4 MiB / 1032535	
3	3	0	SUCCESS	PROCESS_LOCAL	0	10.100.0.35	stdout stderr	2021-06-30 10:38:12	31 s	0.7 s	1 GiB / 1610215	901.3 MiB / 1019886	
4	4	0	SUCCESS	PROCESS_LOCAL	0	10.100.0.35	stdout stderr	2021-06-30 10:38:12	31 s	0.7 s	1 GiB / 1609083	901.5 MiB / 1041793	
5	5	0	SUCCESS	PROCESS_LOCAL	1	10.100.0.35	stdout	2021-06-30 10:38:12	31 s	0.9 s	1 GiB / 1610001	901.5 MiB / 1036877	

Shuffle Write with Spill

Aggregated Metrics by Executor

Show 20 entries Search:

Executor ID	Logs	Address	Task Time	Total Tasks	Failed Tasks	Killed Tasks	Succeeded Tasks	Excluded	Input Size / Records	Shuffle Write Size / Records
0	stdout stderr	10.100.0.35:40663	21 min	16	0	0	16	false	15.3 GiB / 24693628	6.7 GiB / 319984
1	stdout stderr	10.100.0.35:39433	21 min	15	0	0	15	false	15 GiB / 24138156	6.5 GiB / 299985

Showing 1 to 2 of 2 entries Previous 1 Next

Tasks (31)

Show 20 entries Search:

Index	Task ID	Attempt	Status	Locality level	Executor ID	Host	Logs	Launch Time	Duration	GC Time	Input Size / Records	Shuffle Write Size / Records	Errors
0	0	0	SUCCESS	PROCESS_LOCAL	0	10.100.0.35	stdout stderr	2021-08-23 17:31:20	1.5 min	29 s	1 GiB / 1609696	444.5 MiB / 19999	
1	1	0	SUCCESS	PROCESS_LOCAL	0	10.100.0.35	stdout stderr	2021-08-23 17:31:20	1.5 min	29 s	1 GiB / 1609688	444.4 MiB / 19999	
2	2	0	SUCCESS	PROCESS_LOCAL	0	10.100.0.35	stdout stderr	2021-08-23 17:31:20	1.5 min	29 s	1 GiB / 1608526	444.5 MiB / 19999	
3	3	0	SUCCESS	PROCESS_LOCAL	0	10.100.0.35	stdout stderr	2021-08-23 17:31:20	1.5 min	29 s	1 GiB / 1608742	444.4 MiB / 19999	
4	4	0	SUCCESS	PROCESS_LOCAL	0	10.100.0.35	stdout stderr	2021-08-23 17:31:20	1.5 min	29 s	1 GiB / 1609725	444.5 MiB / 19999	

Shuffle Read without Spill

Page: 1 — 1 Pages. Jump to 1 . Show 100 items in a page. Go

Stage Id ▾	Description	Submitted	Duration	Tasks: Succeeded/Total	Input	Output	Shuffle Read	Shuffle Write
0	map at ScalaWordCount.scala:39 +details	2021/06/30 10:38:10	1.6 min	31/31	30.3 GiB			26.7 GiB

Page: 1 — 1 Pages. Jump to 1 . Show 100 items in a page. Go

▾ Failed Stages (1)

Page: 1 — 1 Pages. Jump to 1 . Show 100 items in a page. Go

Stage Id ▾	Description	Submitted	Duration	Tasks: Succeeded/Total	Input	Output	Shuffle Read	Shuffle Write	Failure Reason
1	runJob at SparkHadoopWriter.scala:83 +details	2021/06/30 10:39:46	20 min	0/10 (10 running)					Job 0 cancelled

Shuffle Read with Spill

Page: 1

1 Pages. Jump to 1 . Show 100 items in a page. Go

Stage Id ▾	Description	Submitted	Duration	Tasks: Succeeded/Total	Input	Output	Shuffle Read	Shuffle Write
1	runJob at SparkHadoopWriter.scala:83 +details	2021/08/23 17:35:35	1.4 min	90/90		26.4 GiB	13.2 GiB	
0	map at ScalaWordCount.scala:41 +details	2021/08/23 17:31:18	4.3 min	31/31	30.3 GiB			13.2 GiB

Future Plan

Test and benchmark to cover more use cases.

Any question?