cloudera®

MapReduce Performance on SSDs

Yanpei Chen (Software Engineer – Performance)

Karthik Kambatla (Software Engineer – Platform)

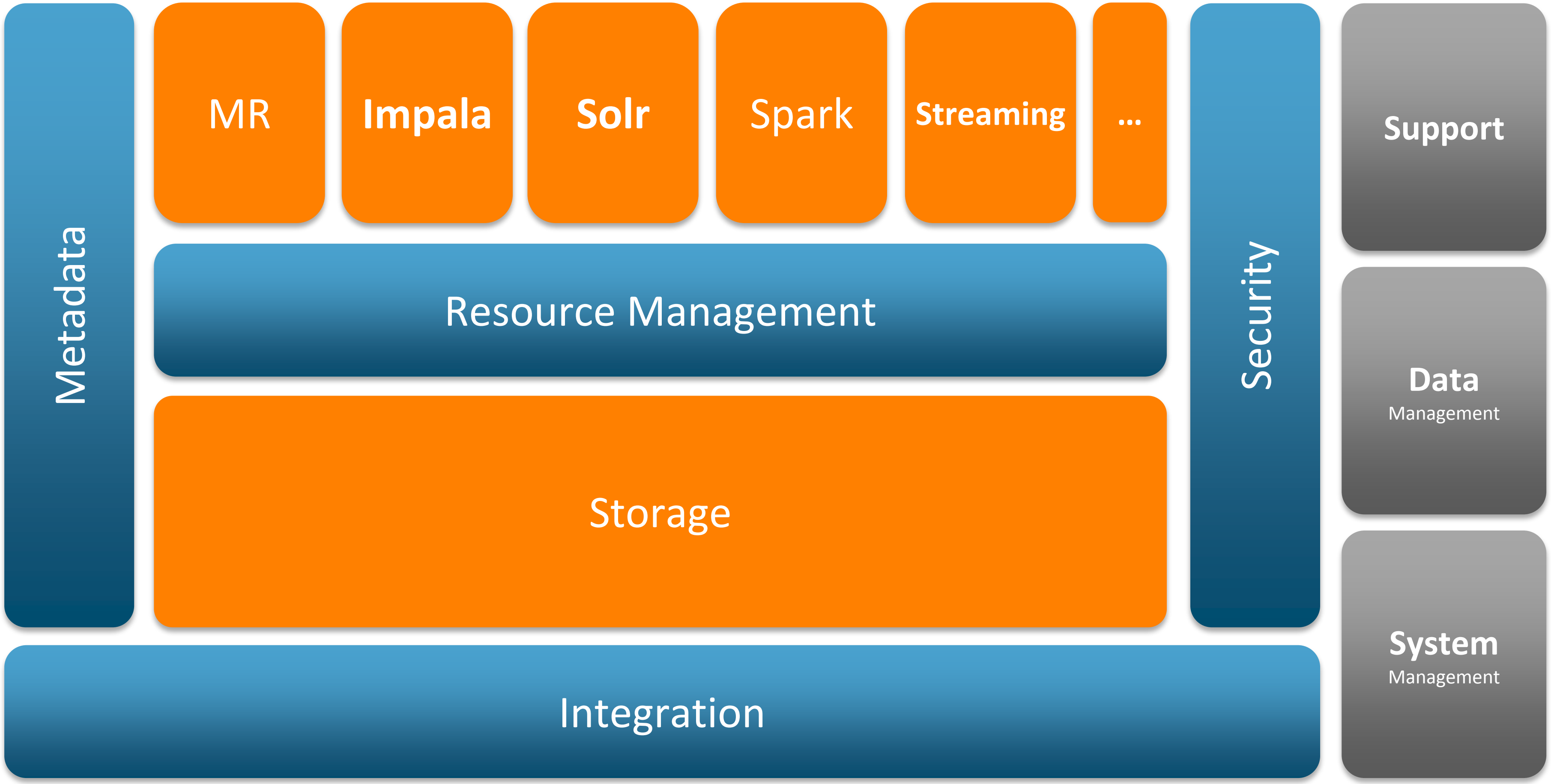
In a nutshell

- MapReduce + SSDs = ?
- Findings
  - Achieve up to 70% higher performance
  - Have 2.5x higher cost-per-performance
  - Should be split into multiple local directories in hybrid clusters
- Meta-finding on SSD trends
  - Compare **cost-per-performance**, not just **cost-per-capacity**

Motivation

- Identify EDH components that would benefit from the use of SSDs
- Provisioning resources for a given workload
  - New clusters: should one prefer HDDs, SSDs or a combination
  - Expansion time: add SSDs or HDDs?

Enterprise Data Hub



cloudera

Background - SSDs

- Typically smaller in capacity
- More expensive than HDDs
- Superior performance
  - Higher sequential read/write throughput
  - Even higher random read/write throughput
  - No seek overhead as in spinning disks

Background – prior work

- Simulate SSD using OS buffer cache
  - Found HDFS code paths that bottleneck HBase
- Real SSD, virtualized cluster
  - Found Hadoop 3x better on SSDs
- Simulate SSD using mathematical models
  - Found small SSD cache gives 3x perf. at 5% more cost
- Actual SSD vs HDD, albeit non-uniform BW and cost
  - Found SSDs can accelerate shuffle phase in Terasort

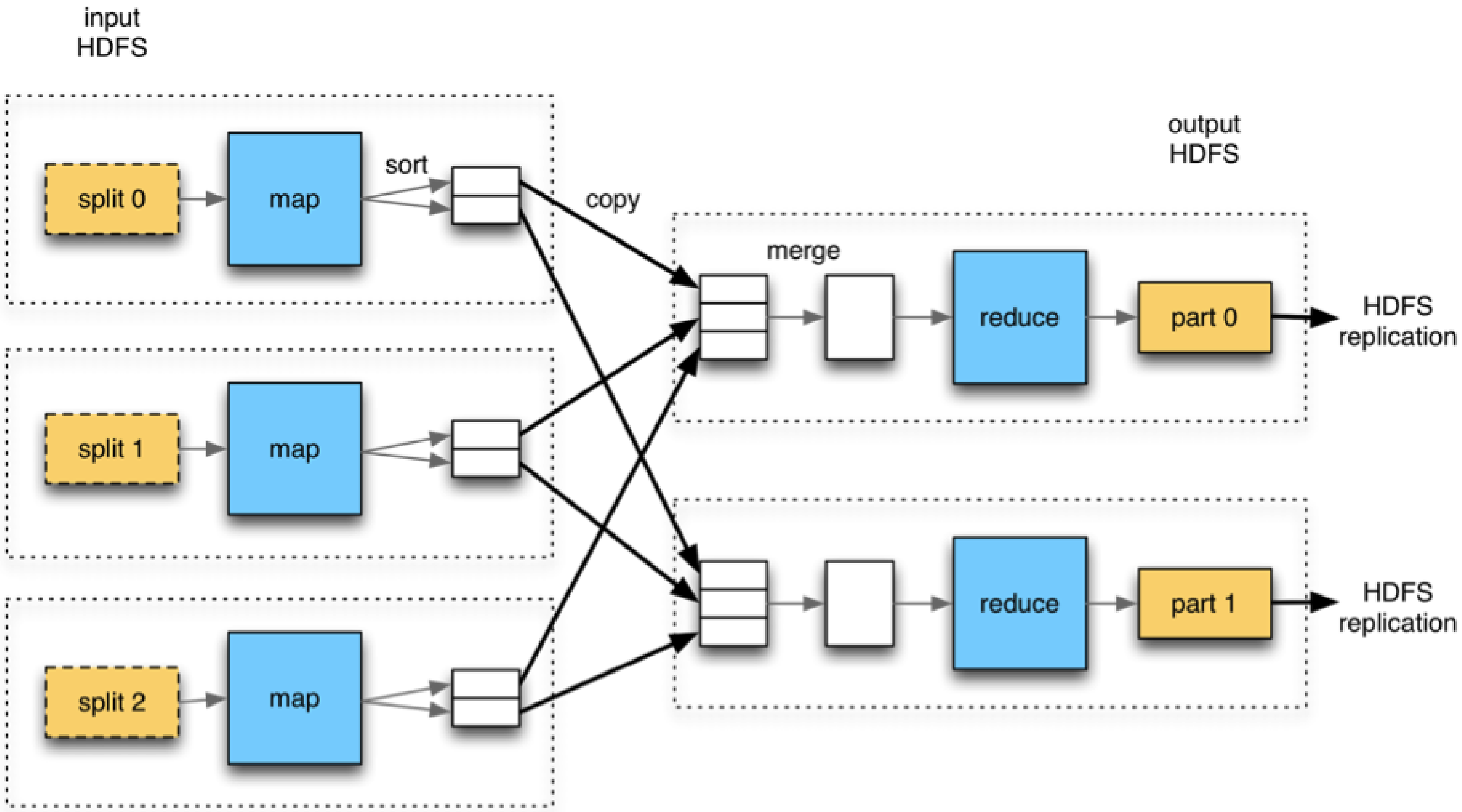
Methodology – build on prior work

- Actual SSDs vs HDDs under equal-bandwidth constraints
- Consider both new (single-medium) and hybrid clusters
- Run stand-alone jobs with a variety of IO/compute mixes
- Run multi-job workloads (did not get to this …)

Hardware used

Setup	Storage	Capacity	Sequential R/W bandwidth	Price
HDD-6	6 HDDs	12 TB	720 MBps	$2,400
HDD-11	11 HDDs	22 TB	1300 MBps	$4,400
SSD	1 SSD	1.3 TB	1300 MBps	$14,000
Hybrid	6 HDDs + 1 SSD	13.3 TB	2020 MBps	$16,400

Background – MapReduce internals

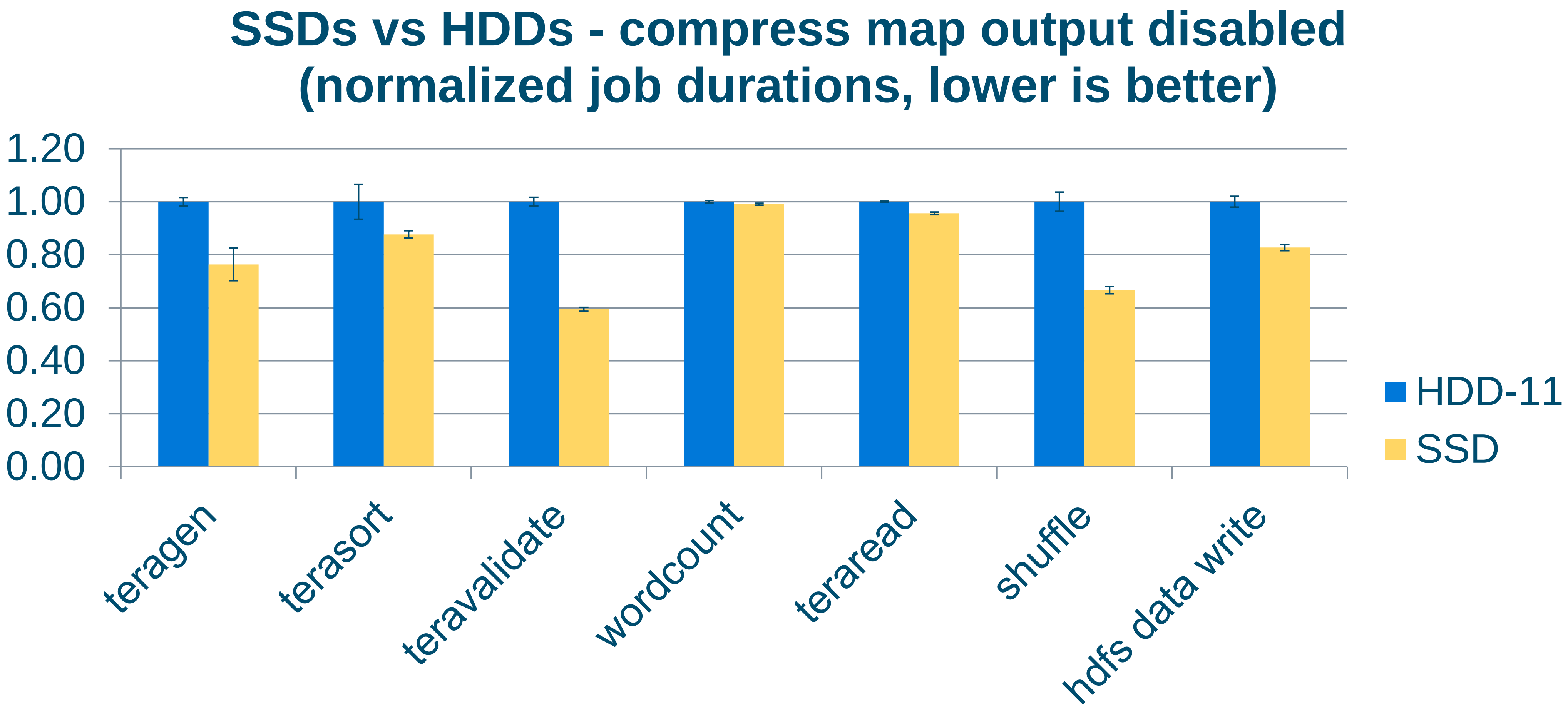


MapReduce jobs used

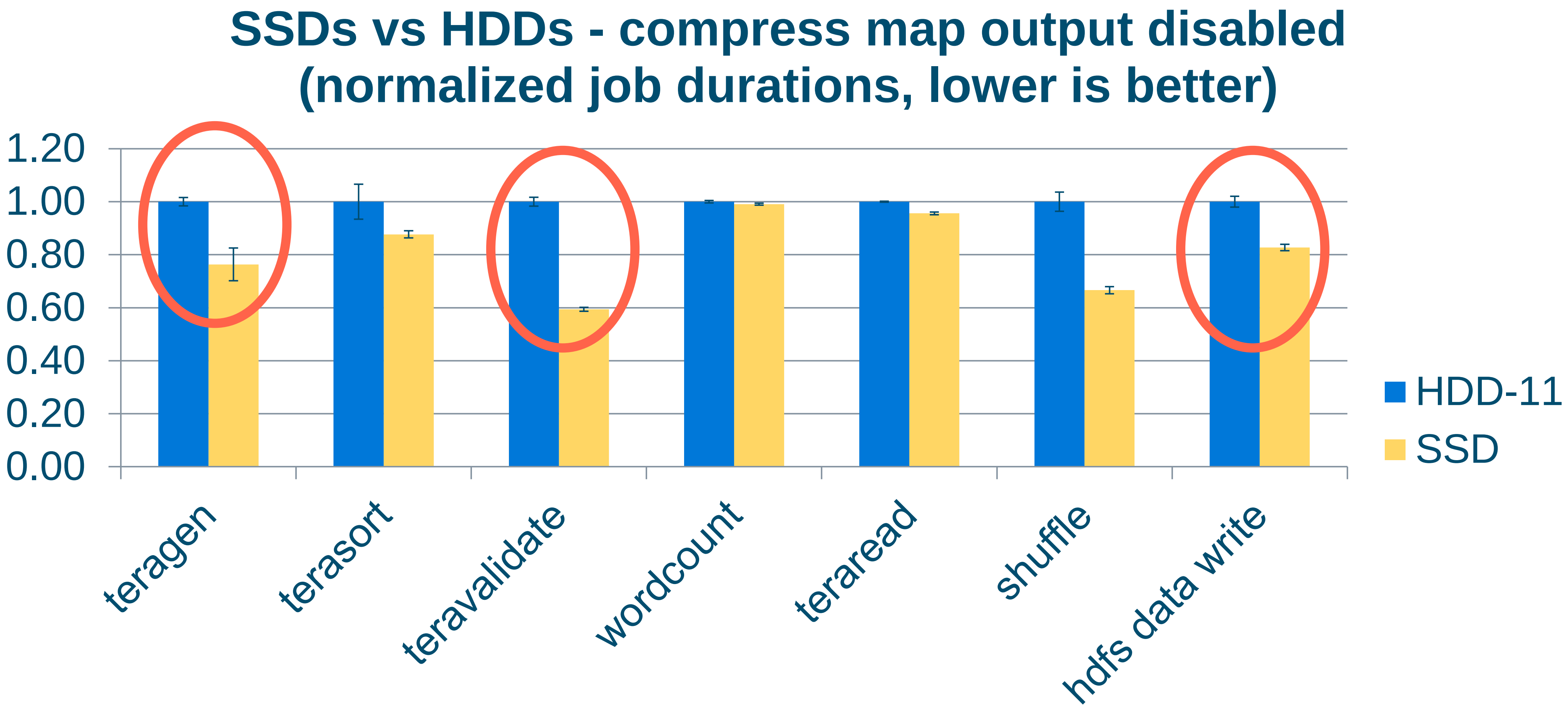
Job	Input size	Shuffle size	Output size	CPU utilization
Teragen	0	0	3	60%
Terasort	1	1	1	61%
Teravalidate	1	0	0	36%
Wordcount	1	0.09	0.09	90%
Teraread	1	0	0	75%
Shuffle	0	1	0	61%
HDFS Data Write	0	0	1	57%

New clusters:
Pure SSD/HDD

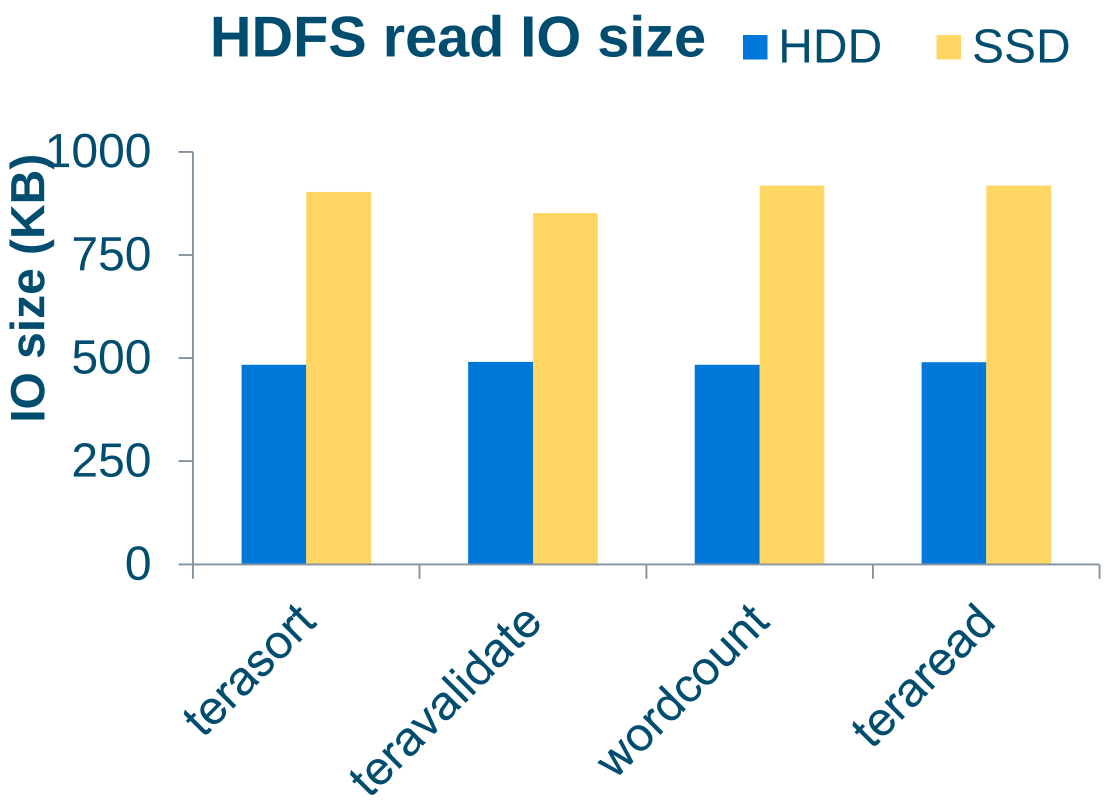
SSDs > HDDs for equal hardware bandwidth

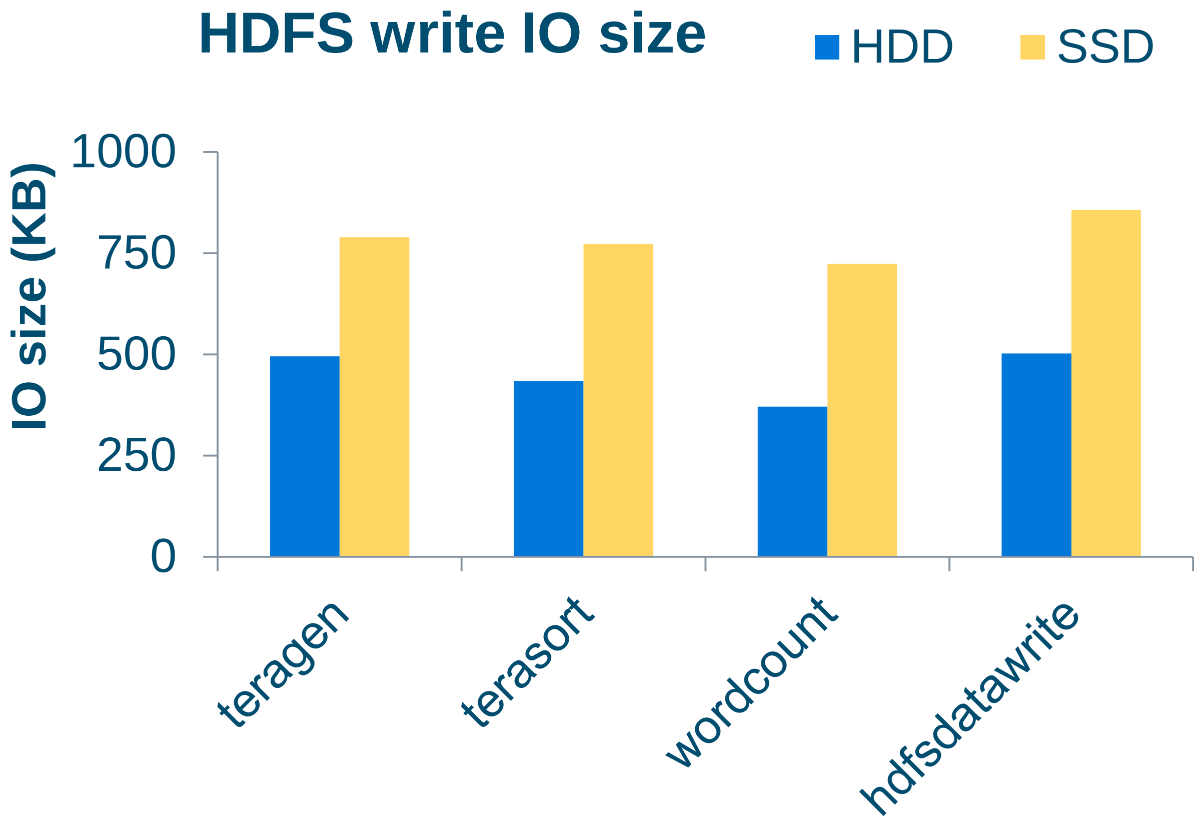


SSDs > HDDs for equal hardware bandwidth

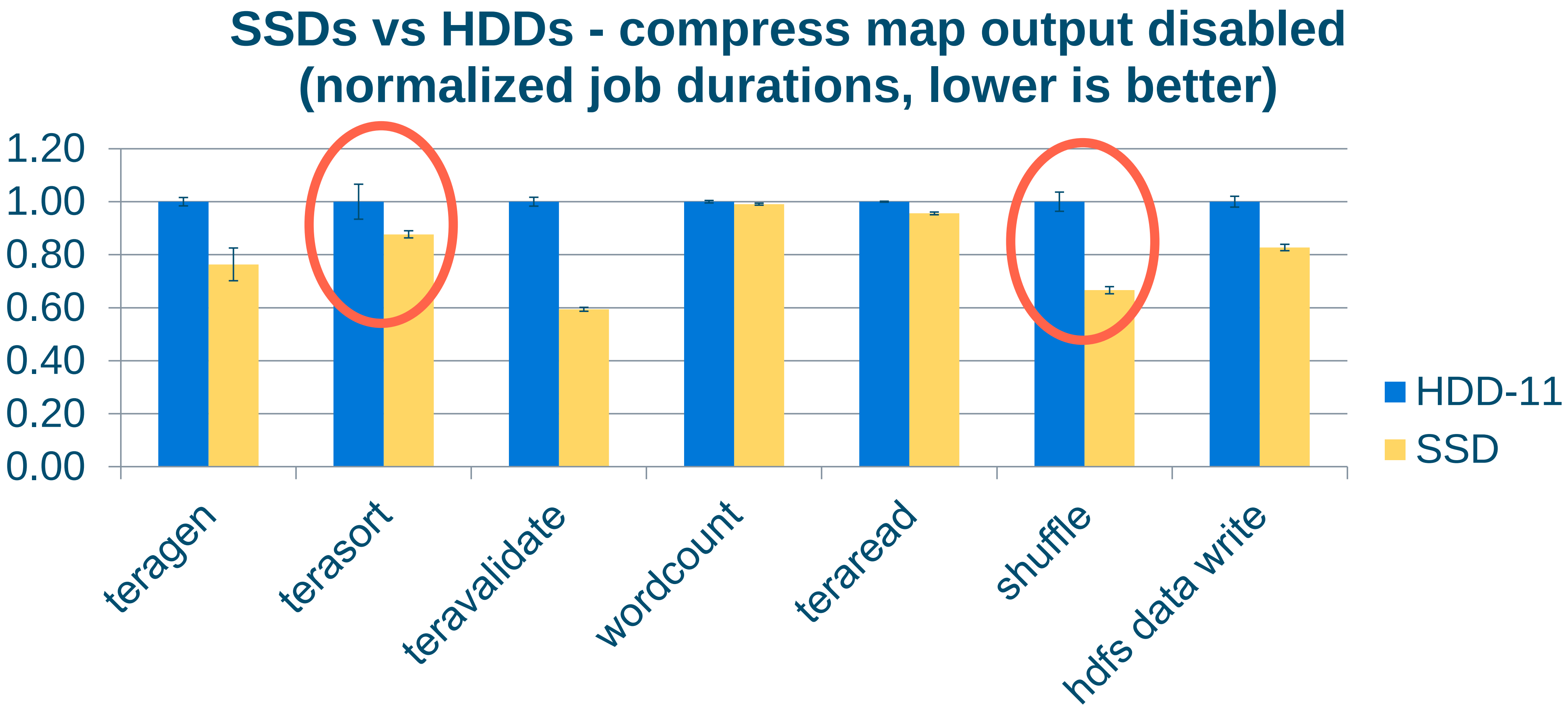


Reason 1: SSDs > HDDs for seq IO size



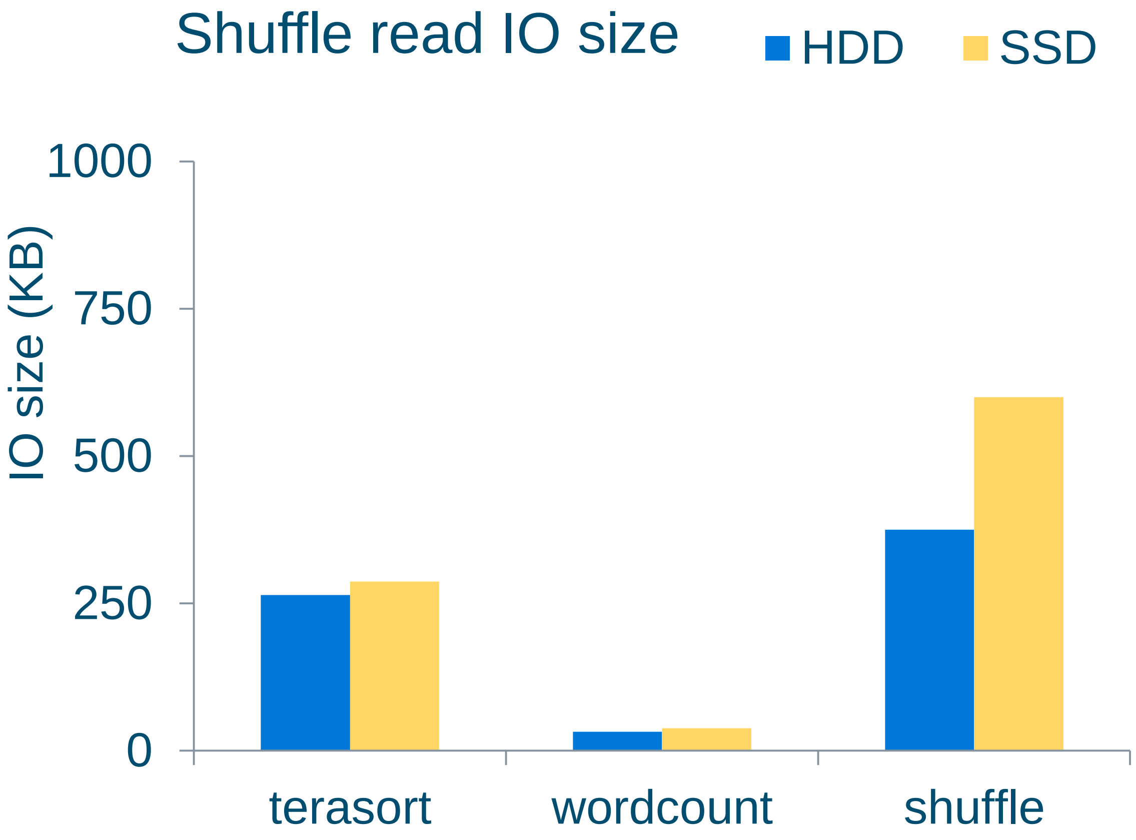


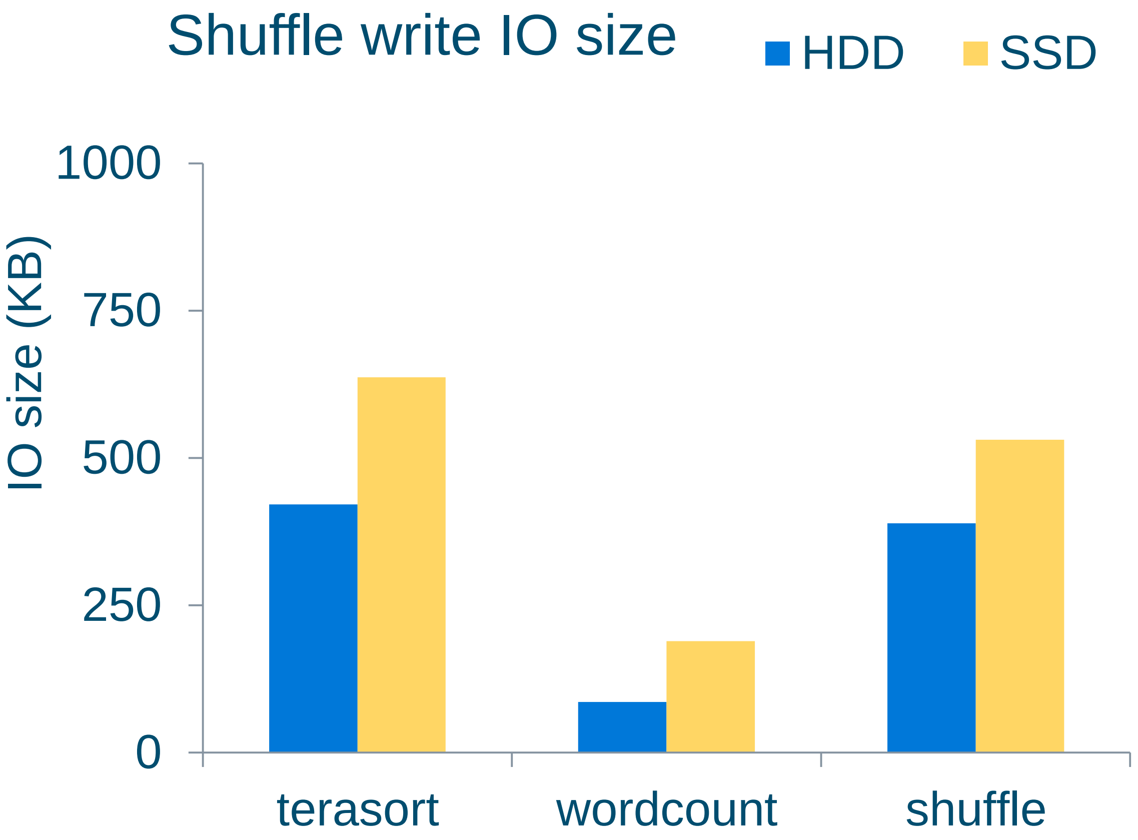
SSDs > HDDs for equal hardware bandwidth



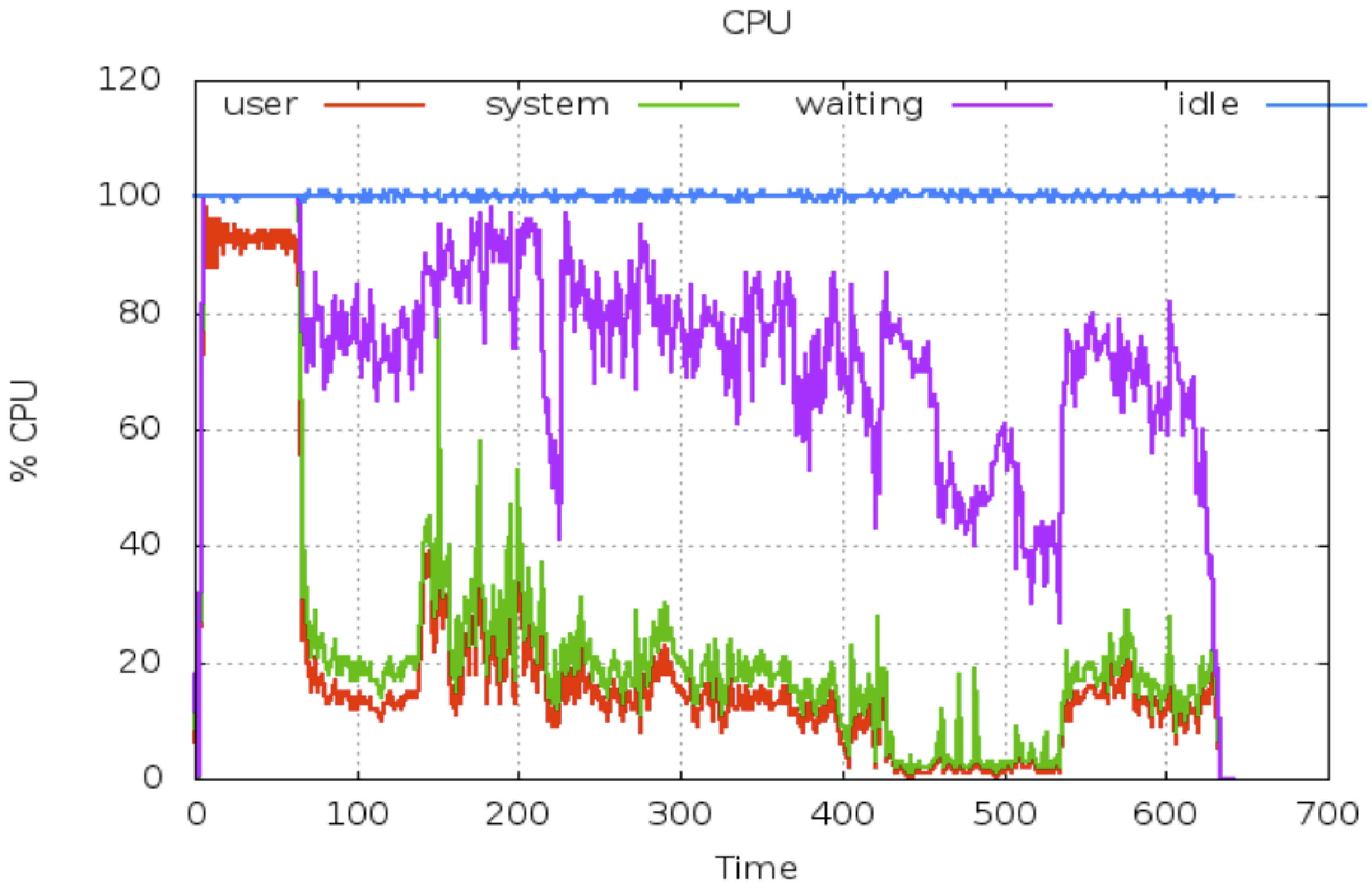
cloudera

Reason 2: SSDs > HDDs for small IO in shuffle

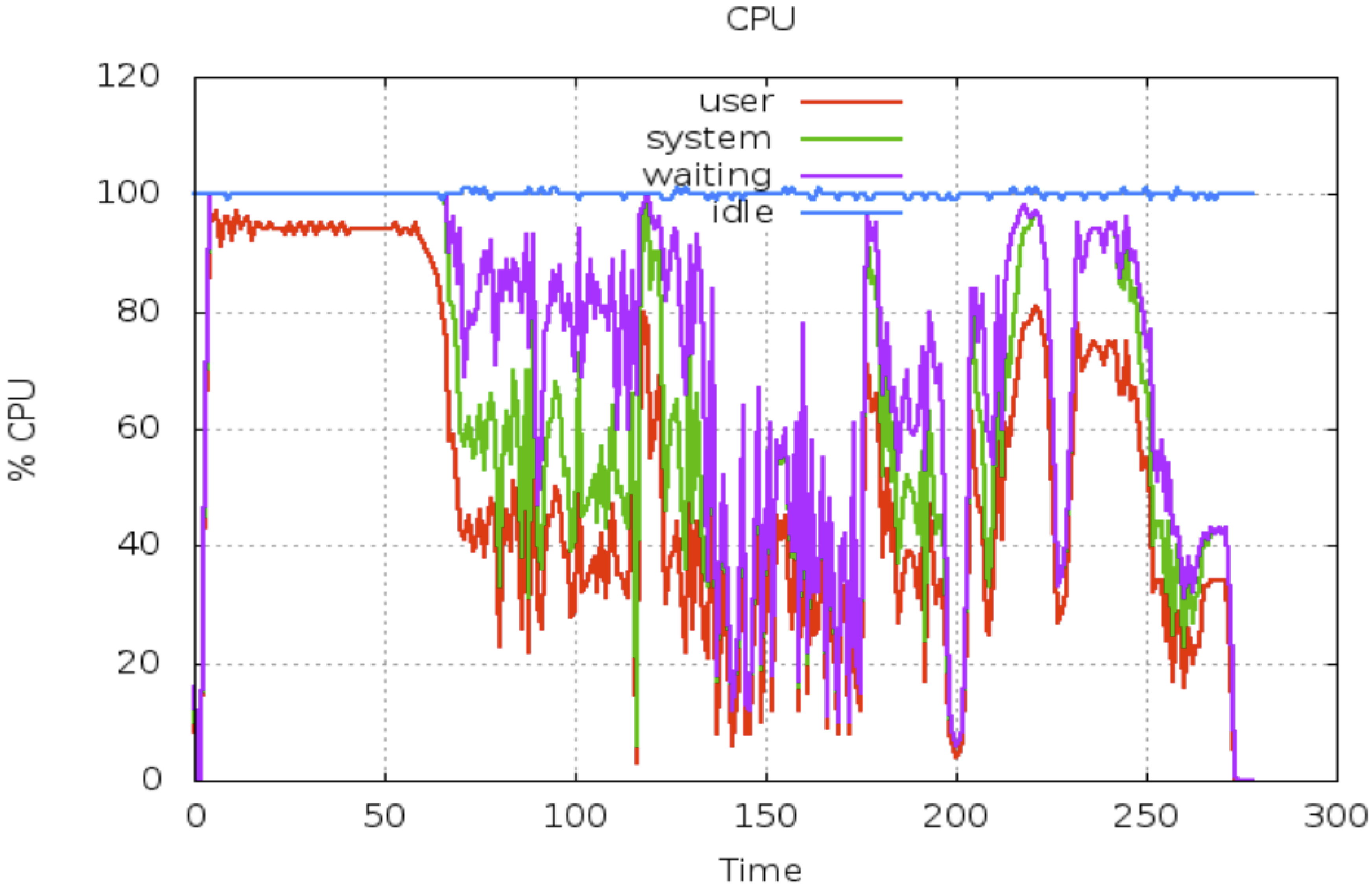




CPU utilization on HDD-6 for Shuffle

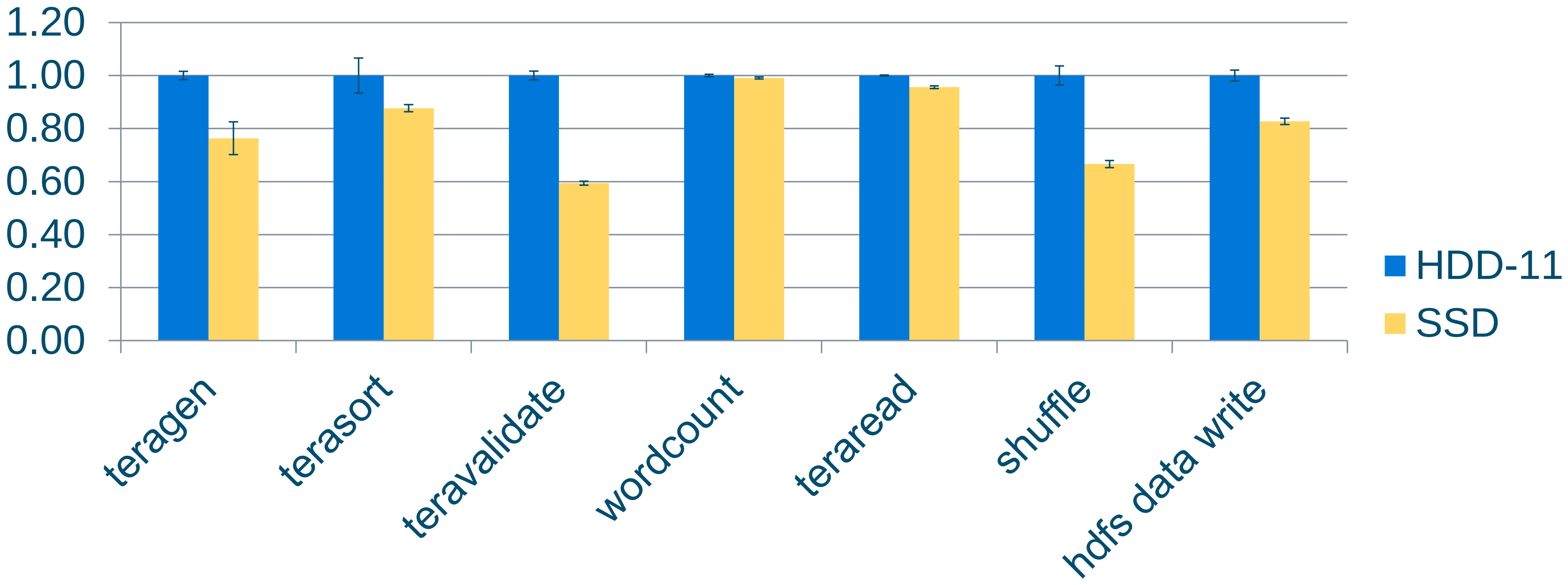


CPU utilization on SSD for Shuffle



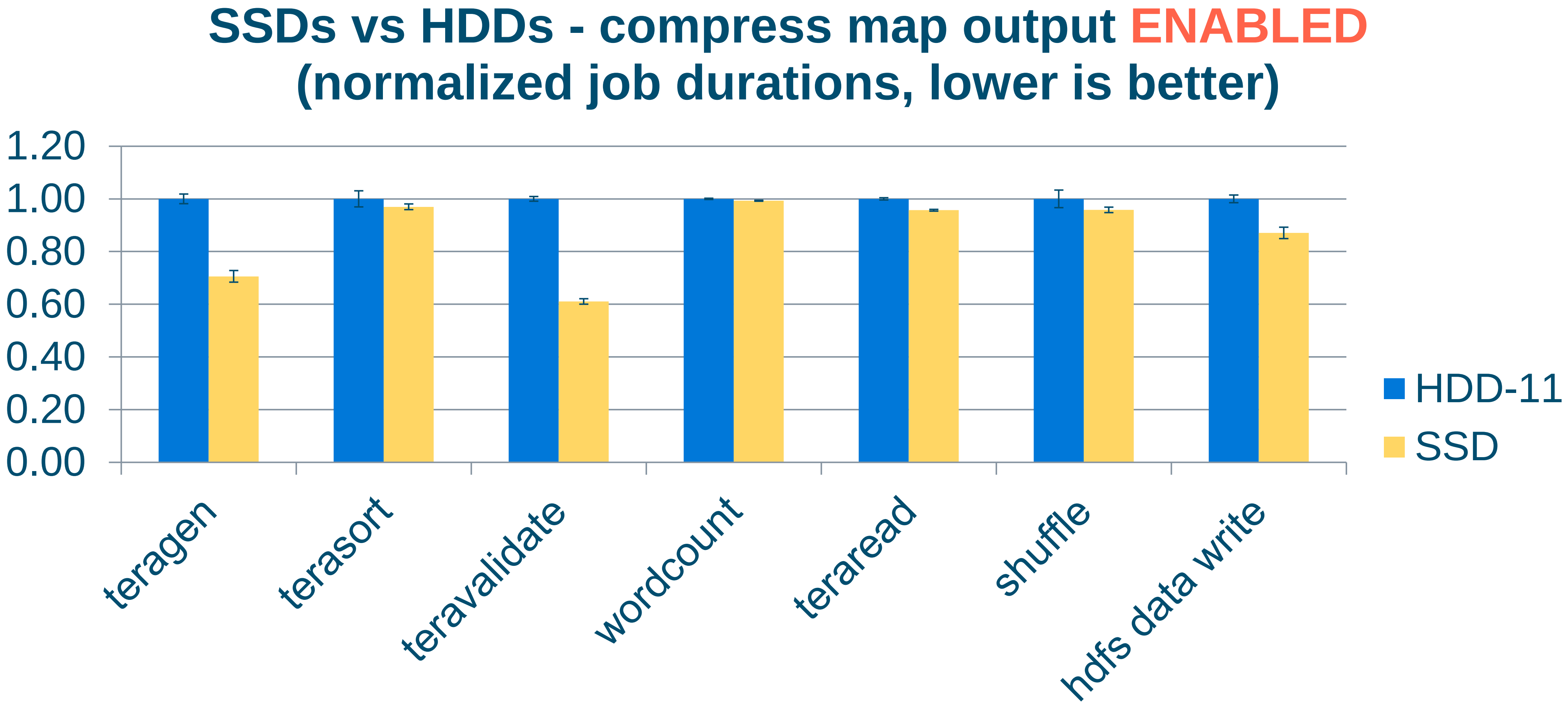
Compression shifts IO vs CPU tradeoff

SSDs vs HDDs - compress map output DISABLED (normalized job durations, lower is better)



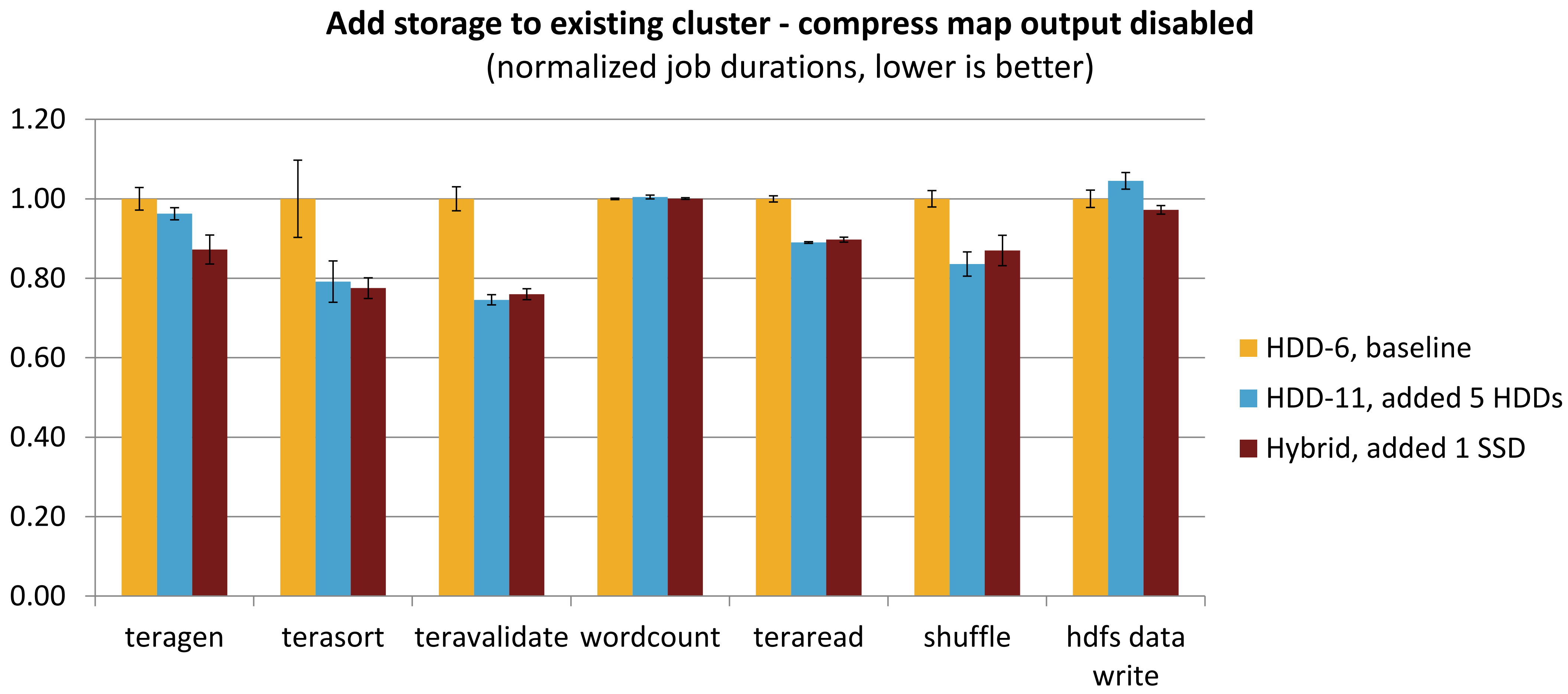
cloudera

Compression shifts IO vs CPU tradeoff



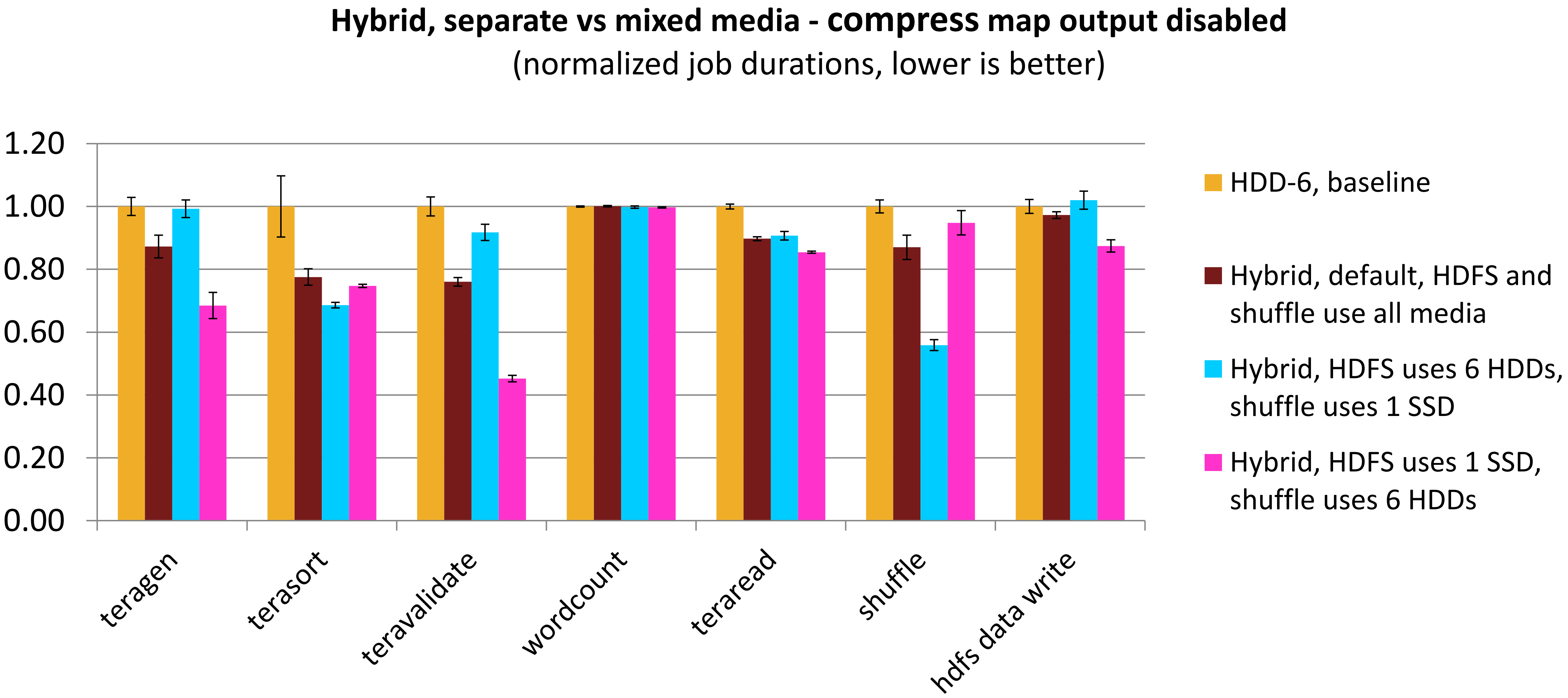
Hybrid clusters

Hybrid clusters – default settings



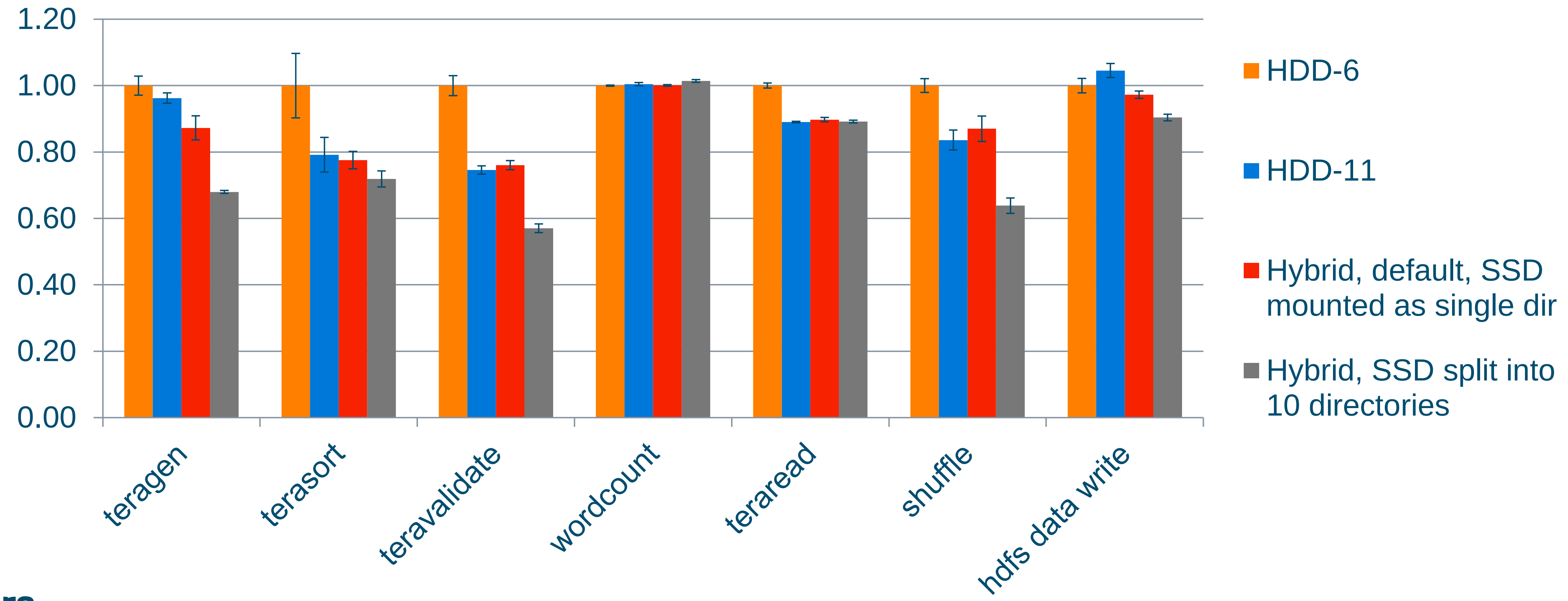
cloudera

Hybrid clusters – SSDs for HDFS/Shuffle



Hybrid clusters – SSD split further

Hybrid, SSD split into 10 directories - compress map output disabled
(normalized job durations, lower is better)



cloudera

Need to consider cost-per-performance

- So SSDs or HDDs?

Setup	Unit cost	Capacity	Unit BW	US$ per TB	Cost per performance
Disk	$400	2 TB	120 MBps	200 (1x baseline)	HDD-11 (1x baseline)
SSD	$14,000	1.3 TB	1300 MBps	10,769 (54x baseline)	SSD (2.5x baseline)

- Willing to pay 2.5x premium for higher performance?
- Willing to work with lower SSD capacity?
- Energy efficiency?

Future work – revisit for new SSDs/HDDs

- Different cost/performance

Setup	Unit cost	Capacity	Unit BW	US$ per TB	Cost per MBps
Disk	$250	4 TB	120 MBps	62.5 (1x baseline)	2.1 (1x baseline)
SSD	$6,400	2 TB	2000 MBps	3,200 (51x baseline)	3.2 (1.5x baseline)

- Use hardware setup under constant cost constraints
- Explore TCO, especially OpEx (energy cost)

Future work – you can help ☺

- Run multi-job MapReduce workloads (SWIM)
- Investigate other enterprise data hub components
  - HBase, Impala, Search, Spark
  - All four aggressively cache data

cloudera

Thank you