



Elastify Cloud-Native Spark Application with PMEM

Junping Du --- Chief Architect, Tencent Cloud Big Data Department
Yue Li --- Cofounder, MemVerge

Table of Contents

- Sparkling: The Tencent Cloud Data Warehouse
- MemVerge PMEM Centric Elastic Spark Solution
- Performance
- Ongoing Work
- Summary

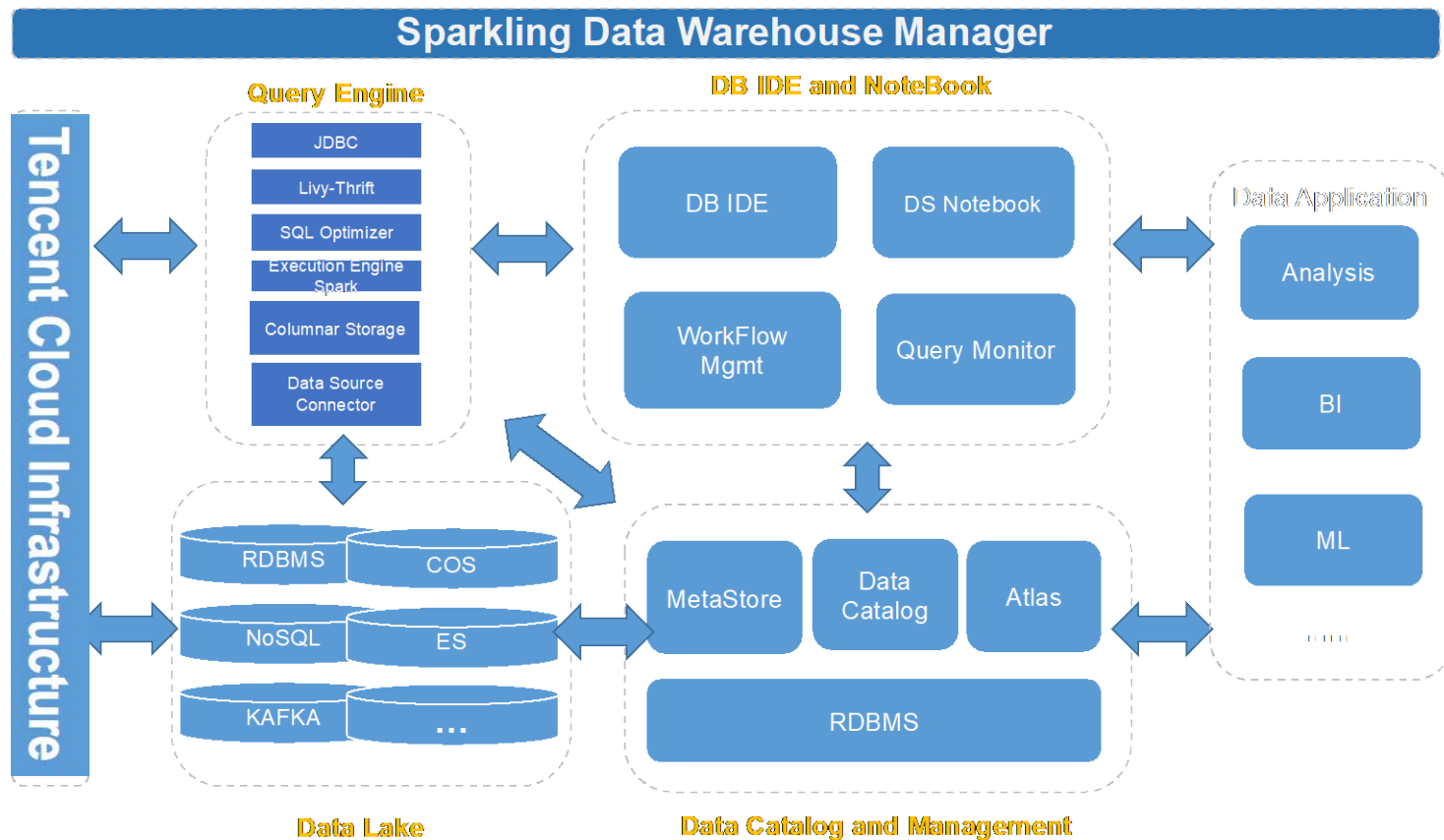


Sparkling Cloud Data Warehouse

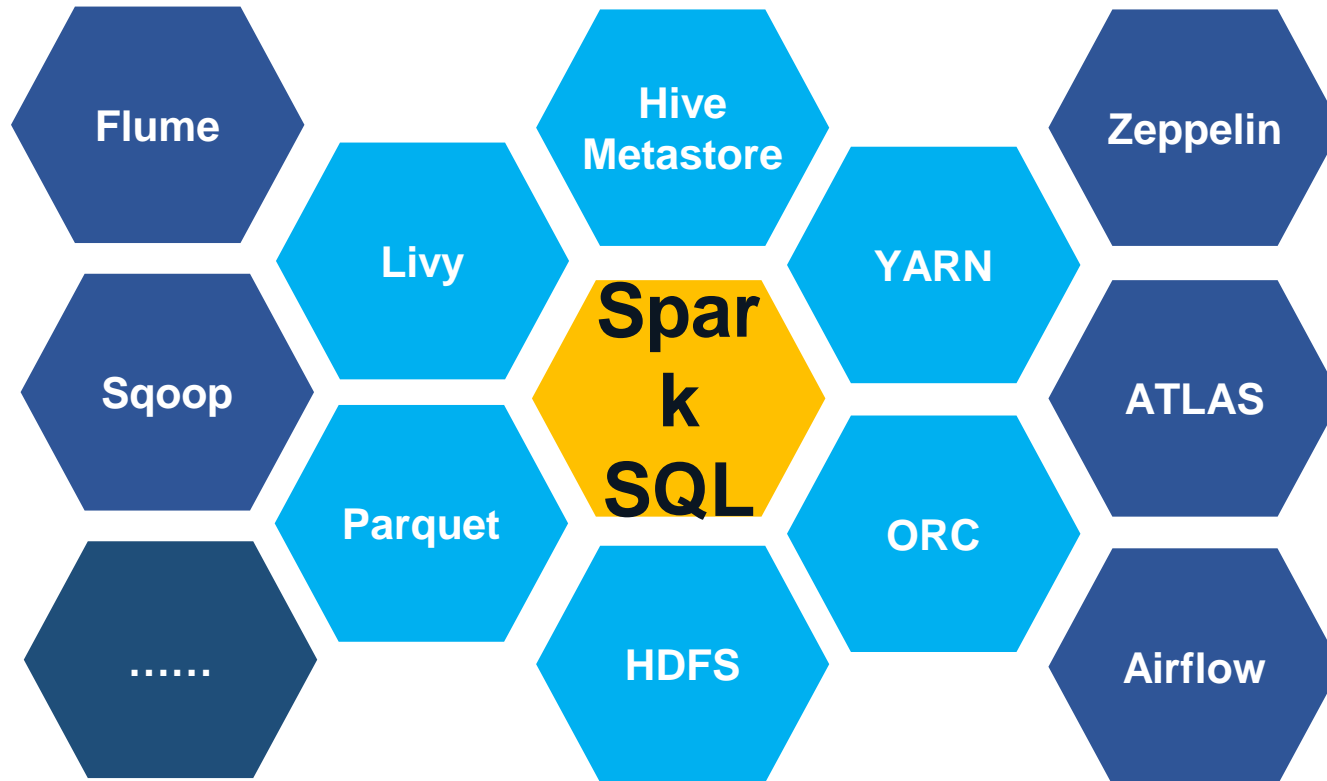
- A PB scale elastic data warehouse
 - ♦ fast deployment, resource elasticity, performant, cost-effective



Sparkling: Architecture Overview



Sparkling: Hiring and Evolving Open Source Technologies



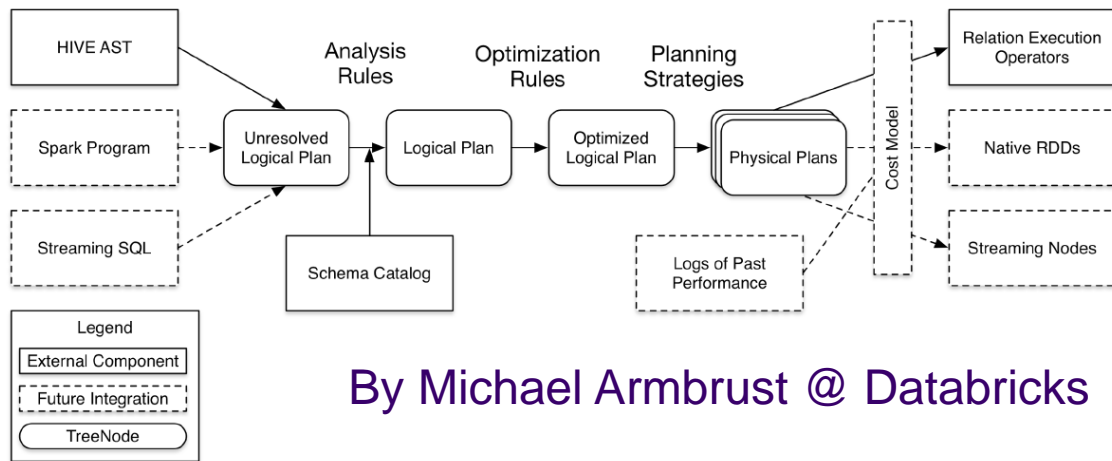
SparkSQL

➤ Outstanding query engine

- ◆ Open source with large and active community
- ◆ Fully compatibility with ANSI SQL 2003
- ◆ High performance

➤ Integration with Spark ecosystem

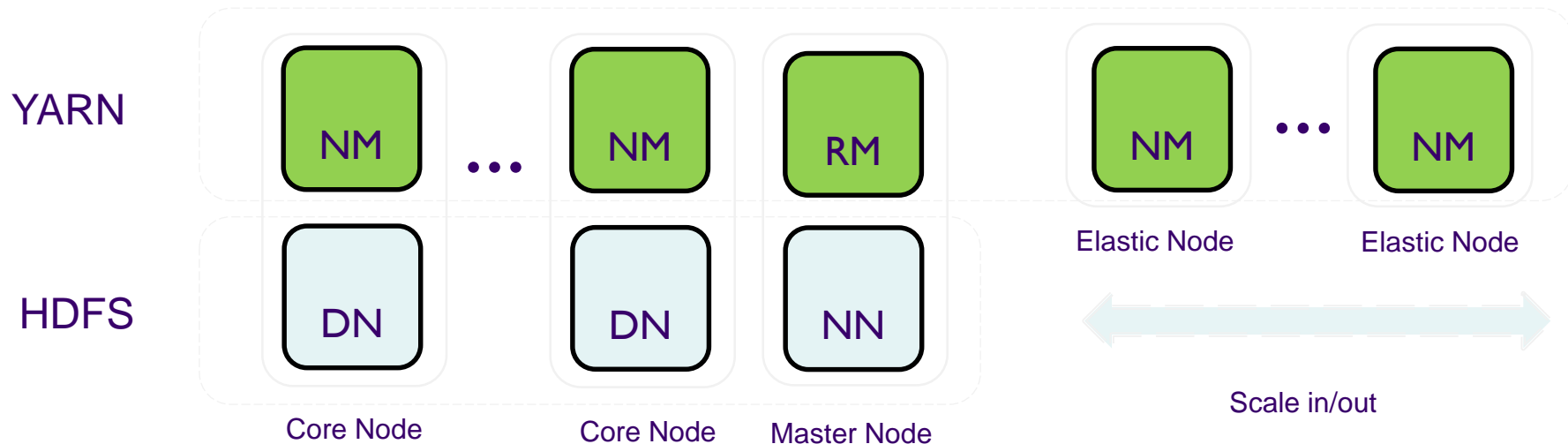
- ◆ Spark Streaming
- ◆ GraphX
- ◆ MLlib/ML



By Michael Armbrust @ Databricks

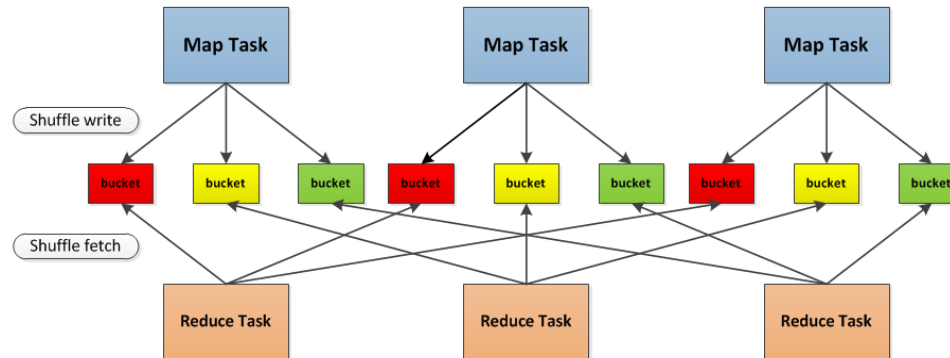
Elastic Deployment Architecture

- ▶ Three types of nodes: Master, Core and Elastic
- ▶ Scale in/out for elastic nodes



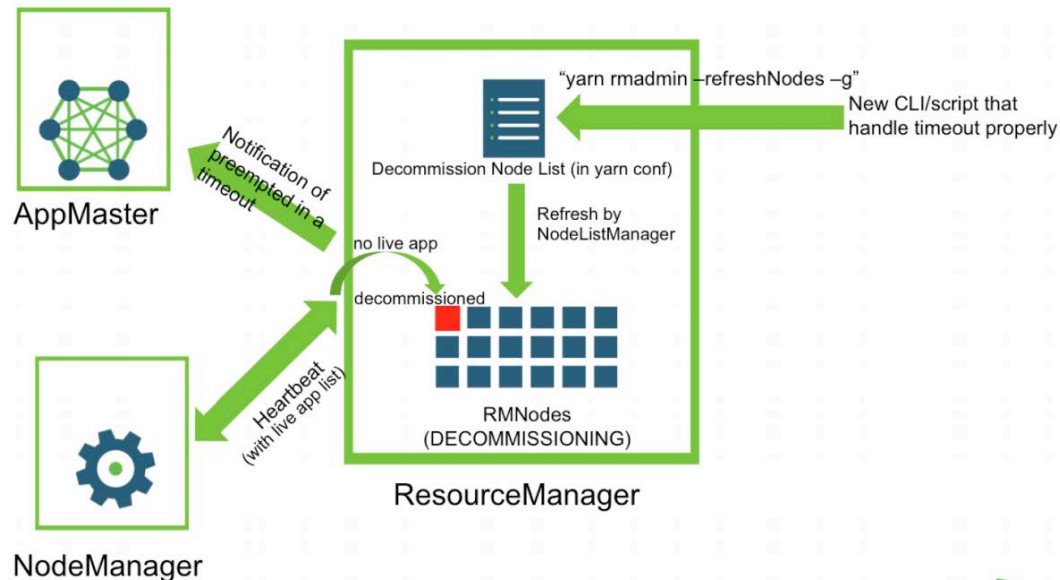
However, Elastic story is not so easy...

- Ideally, compute node should be stateless
- Actually, it is not...
 - ◆ Node decommission will bring down running tasks
 - ◆ Map output get lost during node decommission

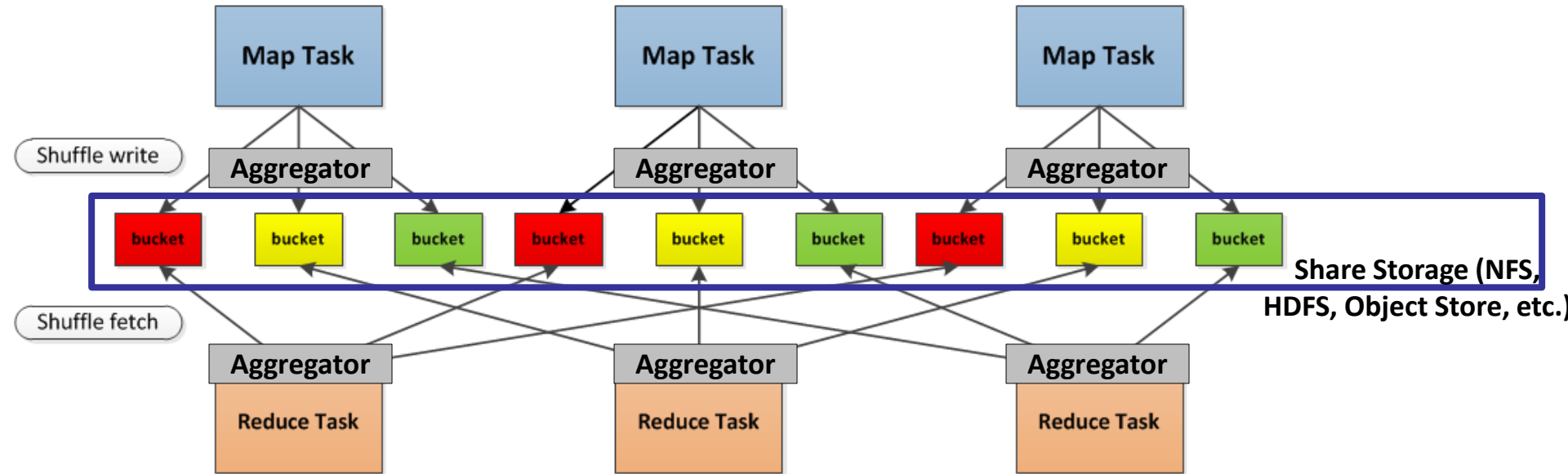


Key solution: Gracefully Decommission

- Add new state of NM – “decommissioning”
- No new containers get launched in decommissioning NMs
- Decommission nodes if all running containers are finished or timeout
- Umbrella JIRA: YARN-914

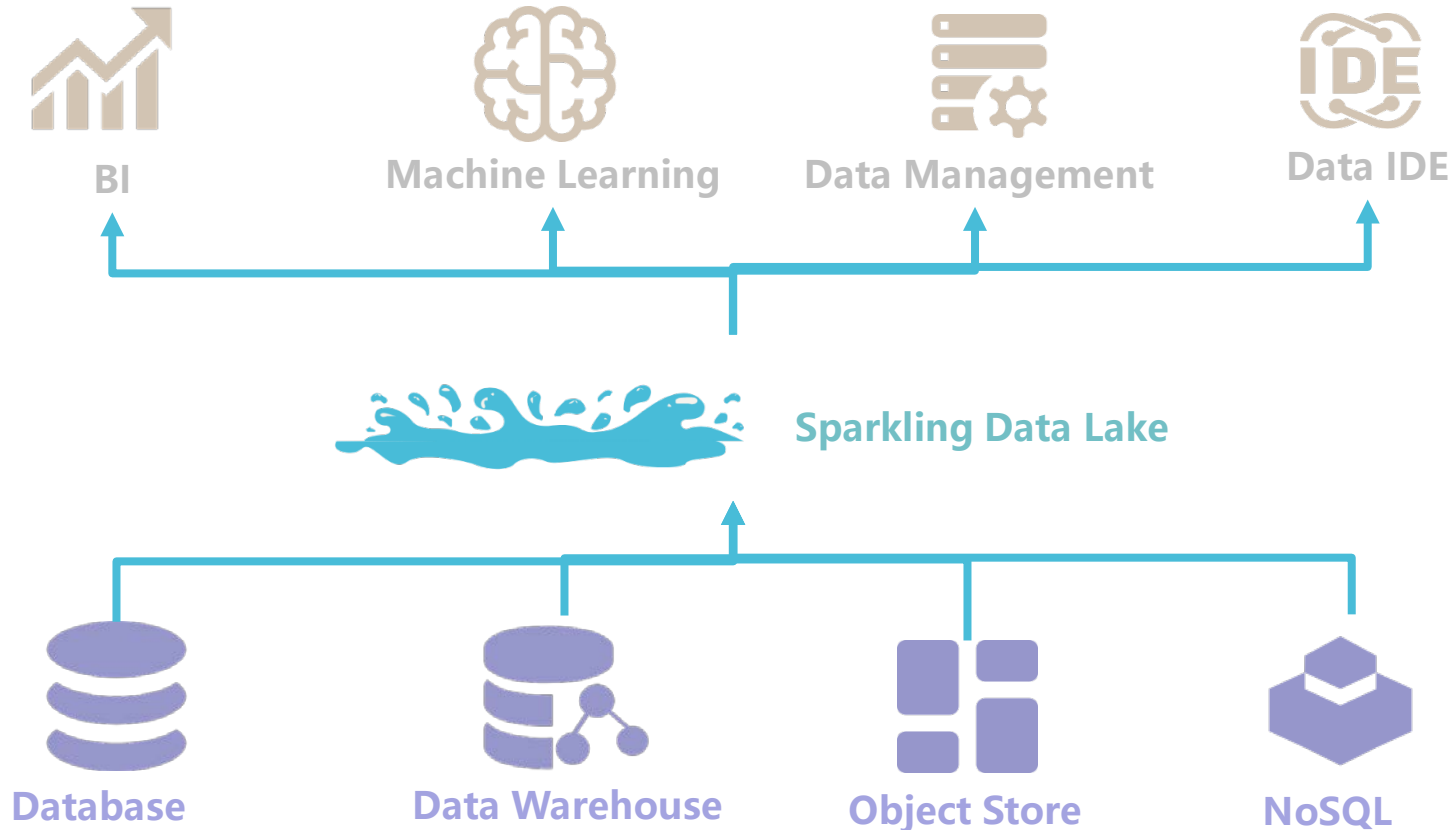


Key Solution: Independent Shuffle Service

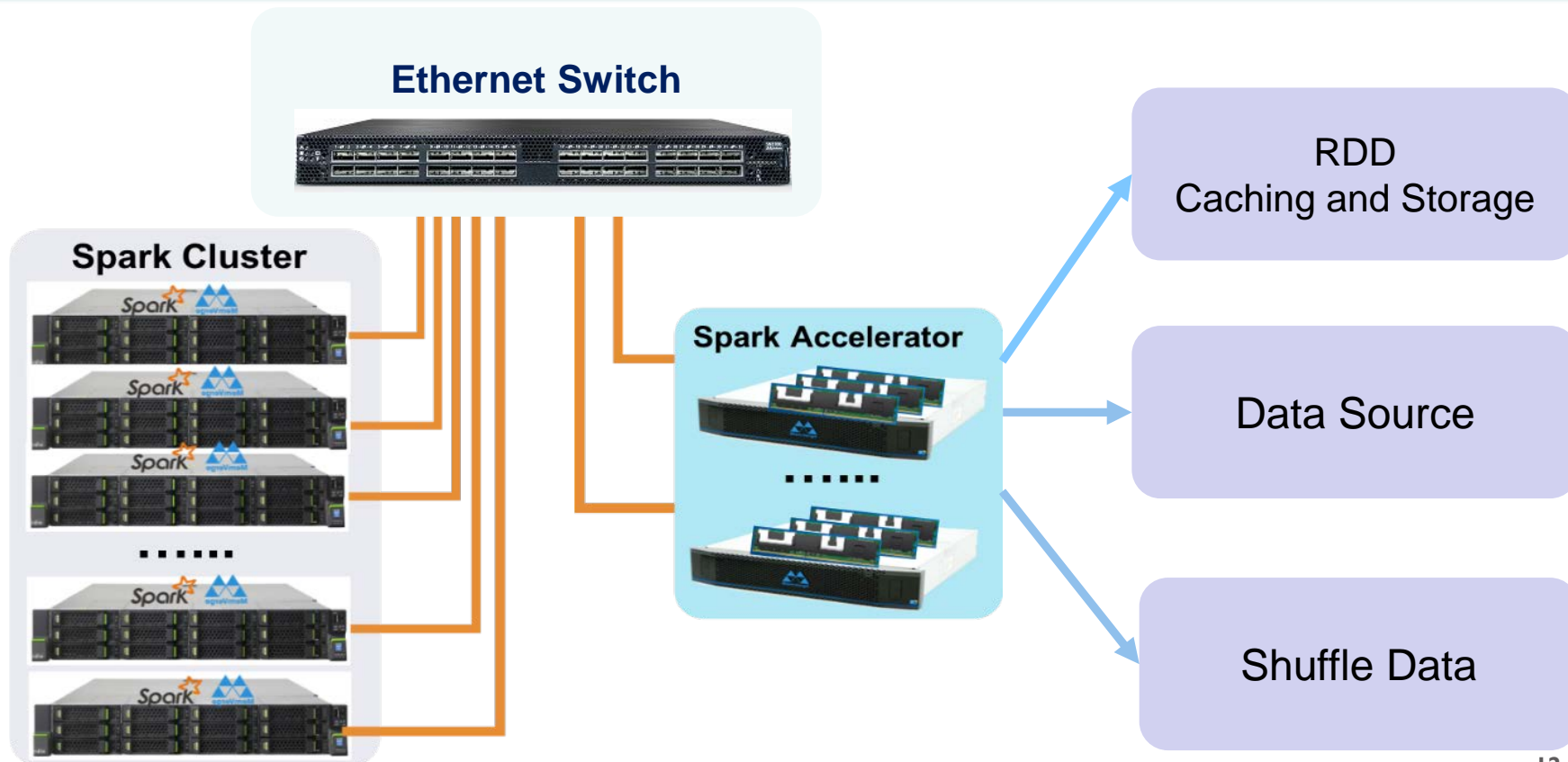


- Shuffle I/O are **decoupled** from a specific network/storage.
- Shuffle read and write can be implemented using **configurable** network transports and backend storage
- Listeners are inserted into different stages of the shuffle to apply hooks.

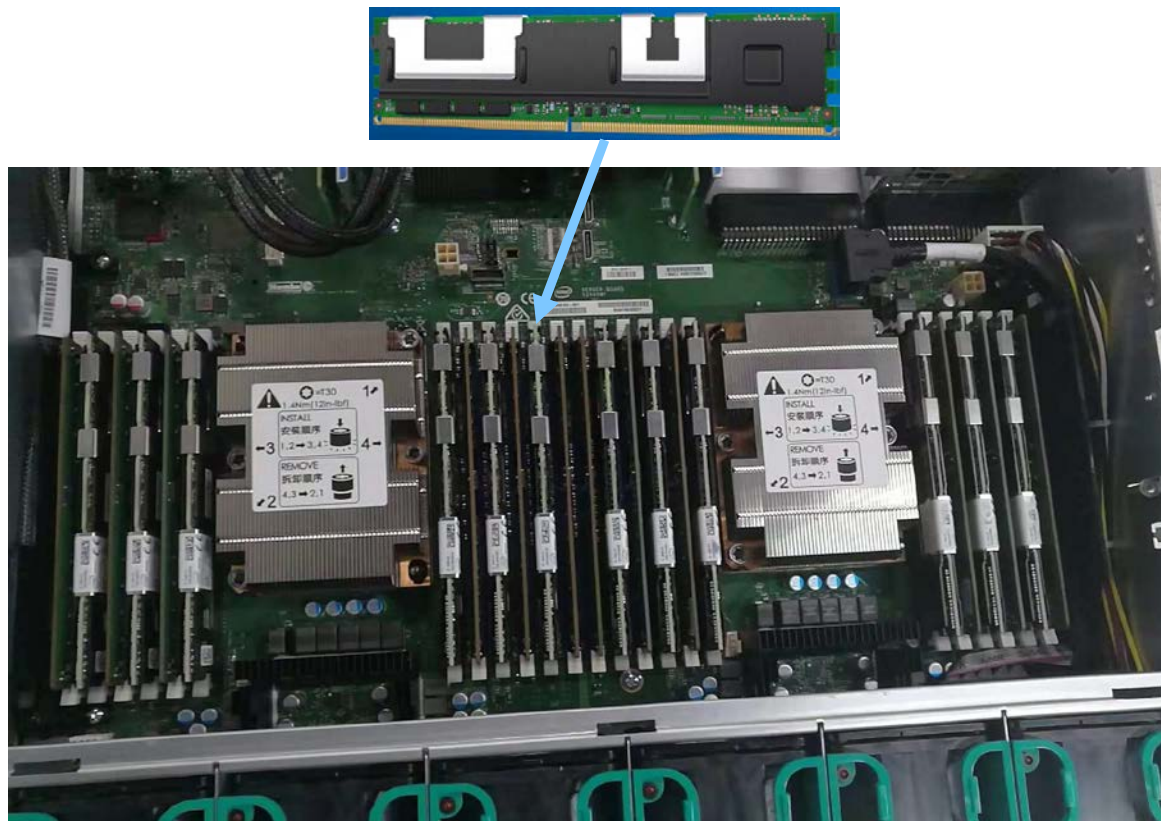
Sparkling: Evolving from Data Warehouse to Data Lake



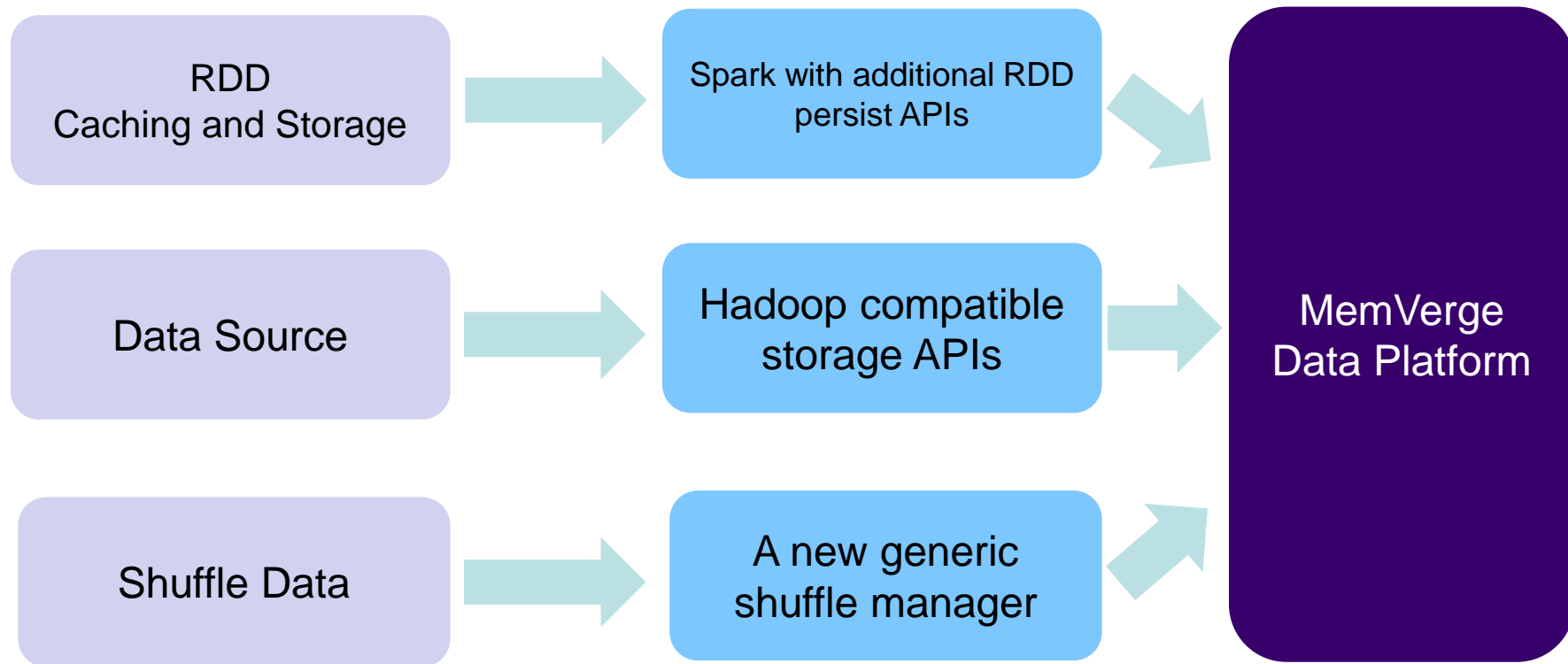
MemVerge Elastic Spark Solution



Intel® Optane™ DC Persistent Memory



Spark Integration



MemVerge Splash Shuffle Manager

- A flexible shuffle manager
 - ◆ supports user-defined storage backend and network transport for shuffle data

- Open source
 - ◆ <https://github.com/MemVerge/splash>

- Spark JIRA: SPARK-25299

The screenshot shows the GitHub repository for MemVerge/splash. At the top, it says "MemVerge / splash" with options to Unwatch (13), Unstar (14), and Fork (1). Below this are tabs for Code, Issues (0), Pull requests (0), Projects (0), Wiki, Insights, and Settings. A description states: "Splash, a flexible Spark shuffle manager that supports user-defined storage backends for shuffle data storage and exchange". There are tags for spark, shuffle, apache-spark, bigdata, scala, java, storage, elasticity, and disaggregation. Statistics show 14 commits, 3 branches, 0 releases, 4 contributors, and Apache-2.0 license. A "New pull request" button is visible. A list of files is shown with their commit messages and dates. The README section is partially visible, titled "Splash", with a build status of "passing" and coverage of "76%".

MemVerge / splash

Unwatch 13 Unstar 14 Fork 1

Code Issues 0 Pull requests 0 Projects 0 Wiki Insights Settings

Splash, a flexible Spark shuffle manager that supports user-defined storage backends for shuffle data storage and exchange Edit

spark shuffle apache-spark bigdata scala java storage elasticity disaggregation Manage topics

14 commits 3 branches 0 releases 4 contributors Apache-2.0

Branch: master New pull request Create new file Upload files Find file Clone or download

File	Commit Message	Time Ago
doc	Typo/language/grammar (#11)	12 days ago
src	Integrate continuous integrations tools on Cloud (#12)	11 days ago
.gitignore	Setup external storage interface.	2 months ago
.travis.yml	Integrate continuous integrations tools on Cloud (#12)	11 days ago
CONTRIBUTING.md	Update CONTRIBUTING.md	12 days ago
LICENSE	Setup external storage interface.	2 months ago
README.md	Integrate continuous integrations tools on Cloud (#12)	11 days ago
clean	Setup external storage interface.	2 months ago
integration-test	Setup external storage interface.	2 months ago
make	Get ready for open source.	12 days ago
make.notest	Get ready for open source.	12 days ago
pom.xml	Integrate continuous integrations tools on Cloud (#12)	11 days ago

README.md

Splash

build passing coverage 76%

A shuffle manager for Spark that supports different storage plugins.

The motivation of this project is to supply a fast, flexible and reliable shuffle manager that allows the user to plug in his/her favorite backend storage and network frameworks for holding and exchanging shuffle data.

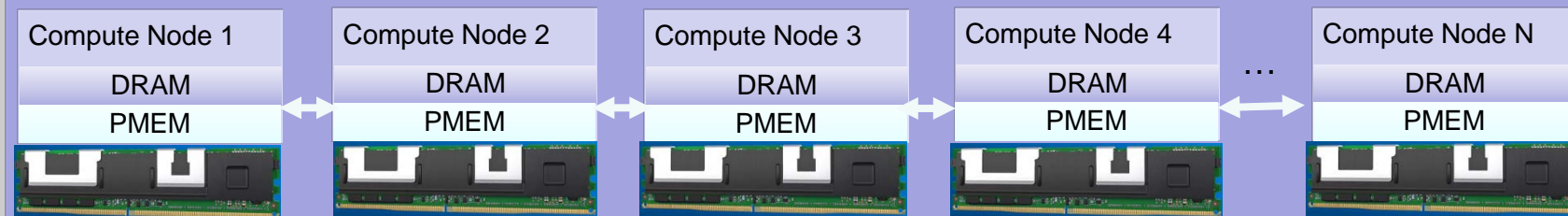
PMEM Centric Data Platform



MemVerge Spark Adaptors

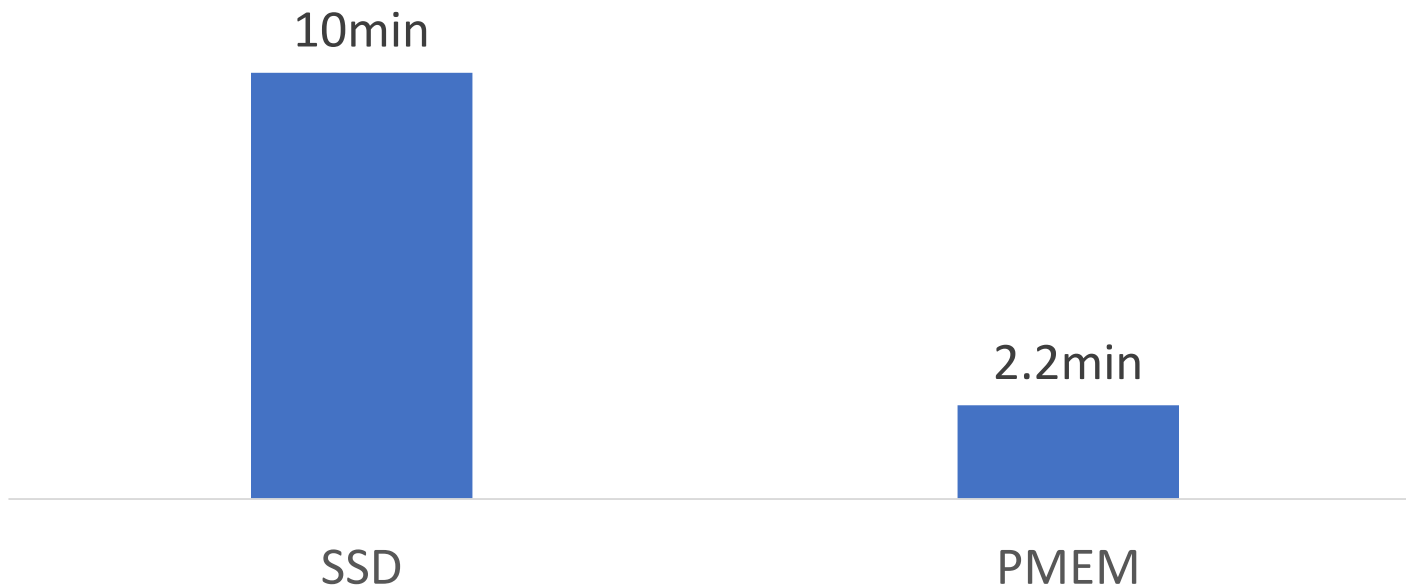
MemVerge SDK

Cluster Shared Persistent Memory



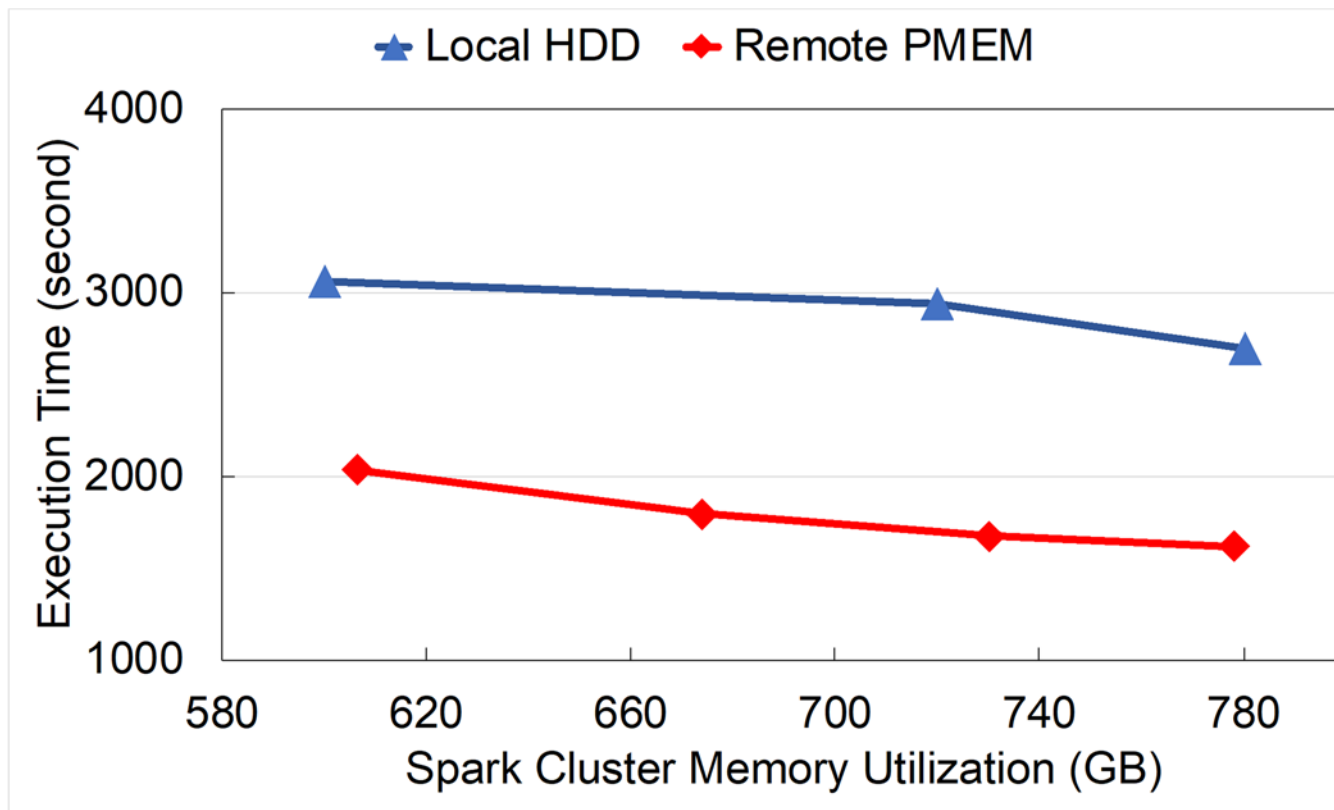
Source Data in Remote PMEM

HiBench Wordcount Time



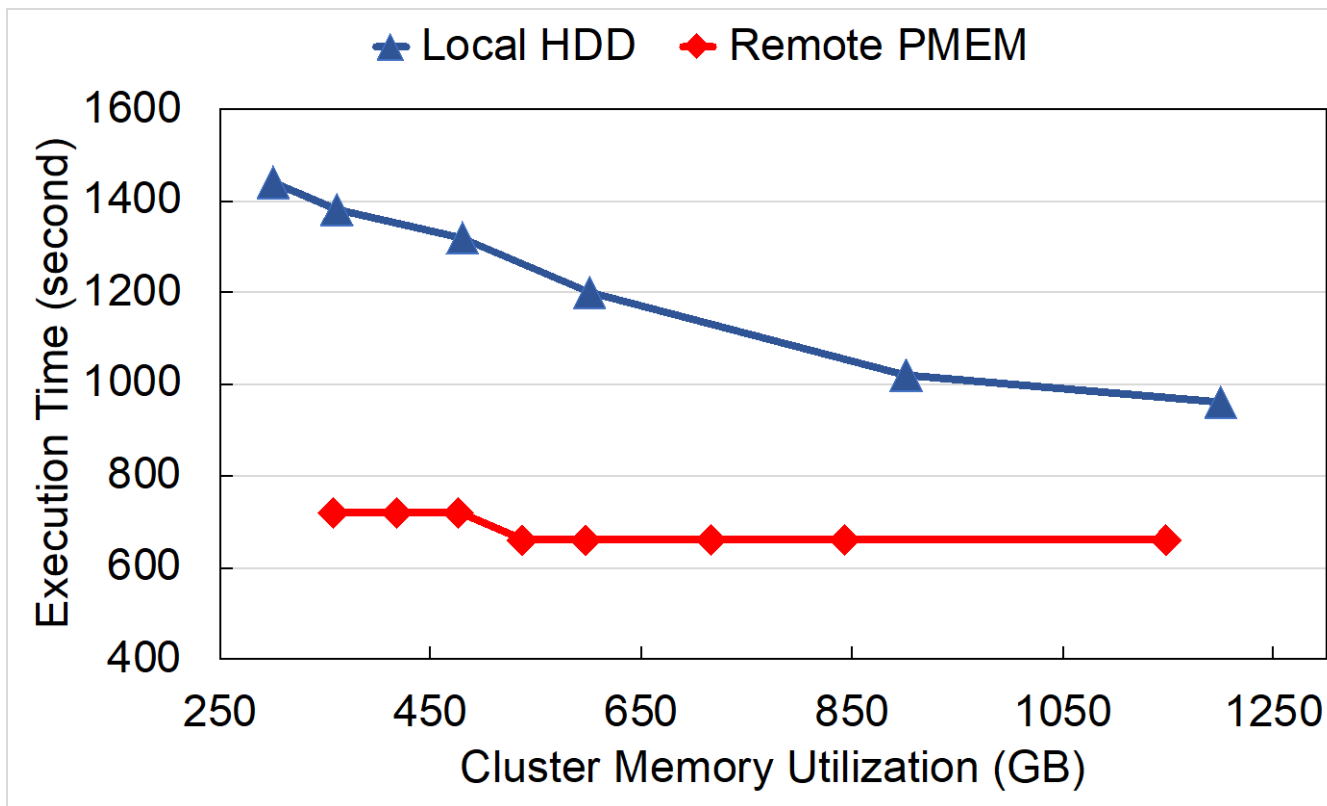
*5 Spark compute node, 1 remote PMEM node, Data size: 610GB

Persisting RDD to Remote PMEM



*10 compute nodes, 1 remote PMEM node, production analytics workload

Shuffling with Remote PMEM



*10 compute nodes, 1 remote PMEM node, production data warehouse workload

➤ TPC-DS performance study

➤ Better cloud readiness

- ◆ Virtual machine support
- ◆ Container support

- PMEM will bring fundamental changes to ALL data centers and enable a data-driven future
- MemVerge and Tencent Cloud deliver better scalability and performance at a lower cost not just for Spark
 - ◆ AI, Big Data, Banking, Animation Studios, Gaming Industry, IoT, etc.
 - ◆ Machine learning, analytics, and online systems
- Thank you Intel for supporting our work!

Junping Du: junpingdu@tencent.com

Yue Li: yue.li@memverge.com