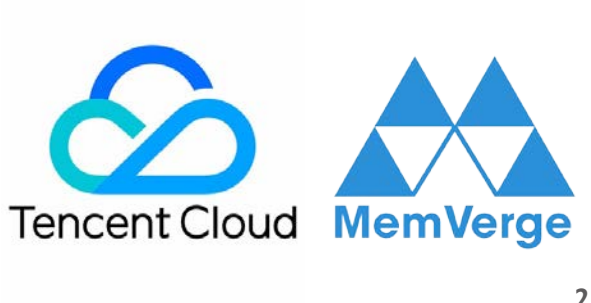
Elastify Cloud-Native Spark Application with PMEM

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

Table of Contents

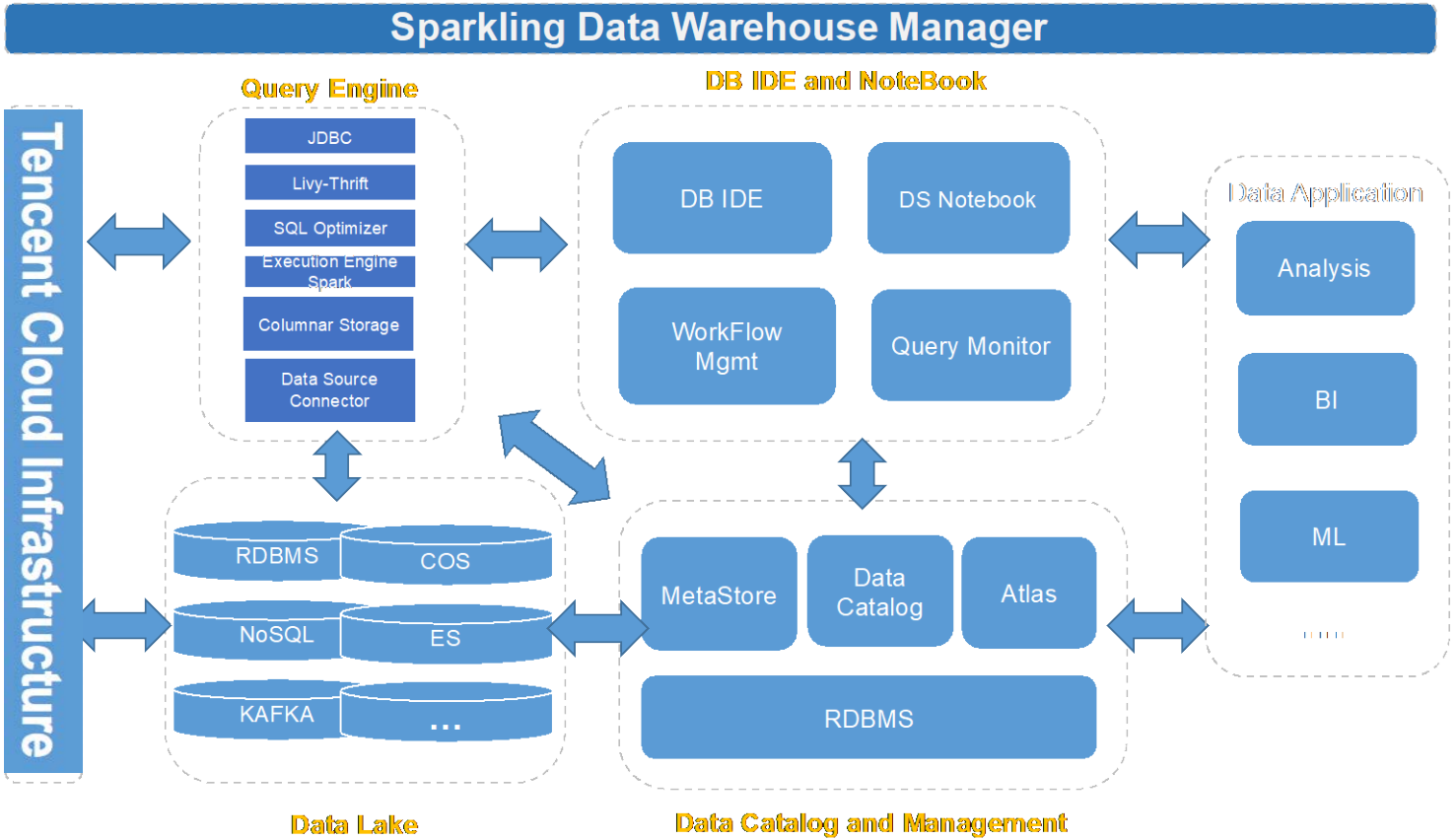
- Sparkling: The Tencent Cloud Data Warehouse
- MemVerge PMEM Centric Elastic Spark Solution
- Performance
- Ongoing Work
- Summary



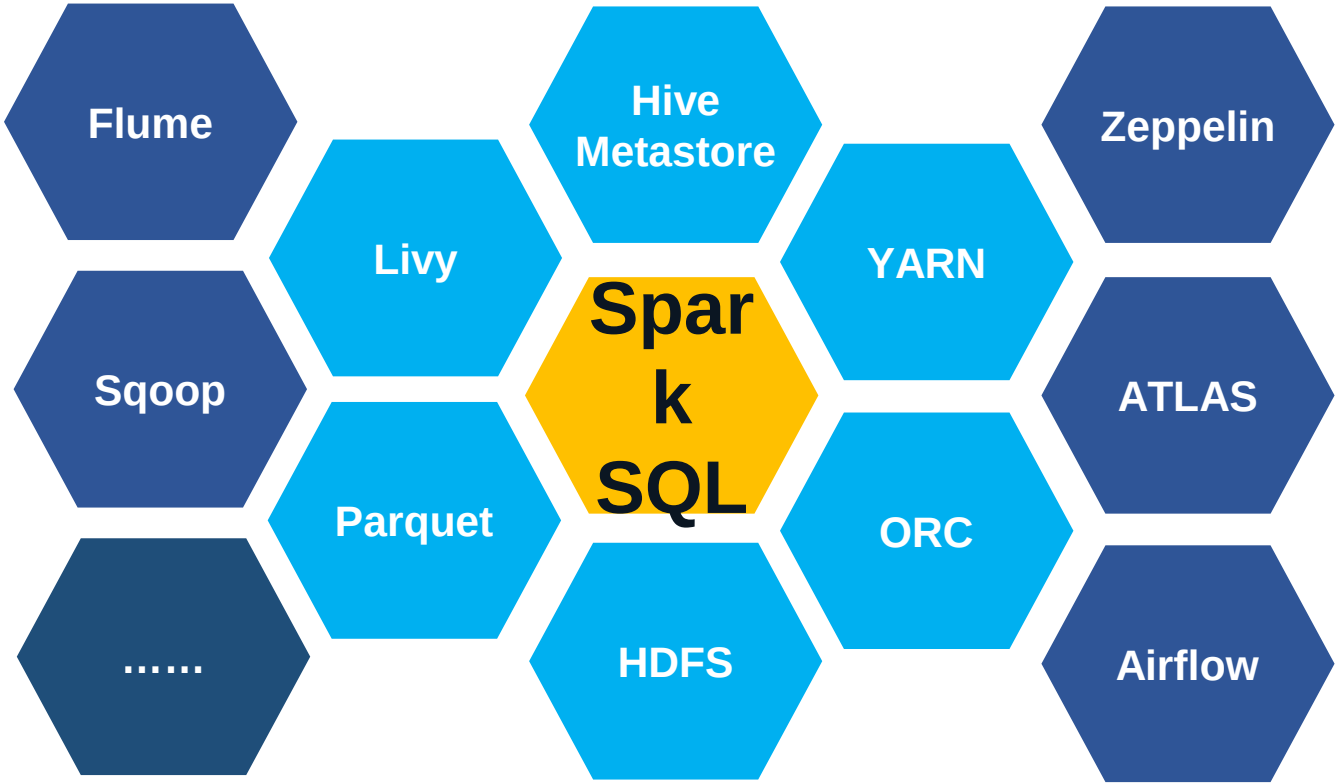
Sparkling Cloud Data Warehouse

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

Sparkling: Architecture Overview

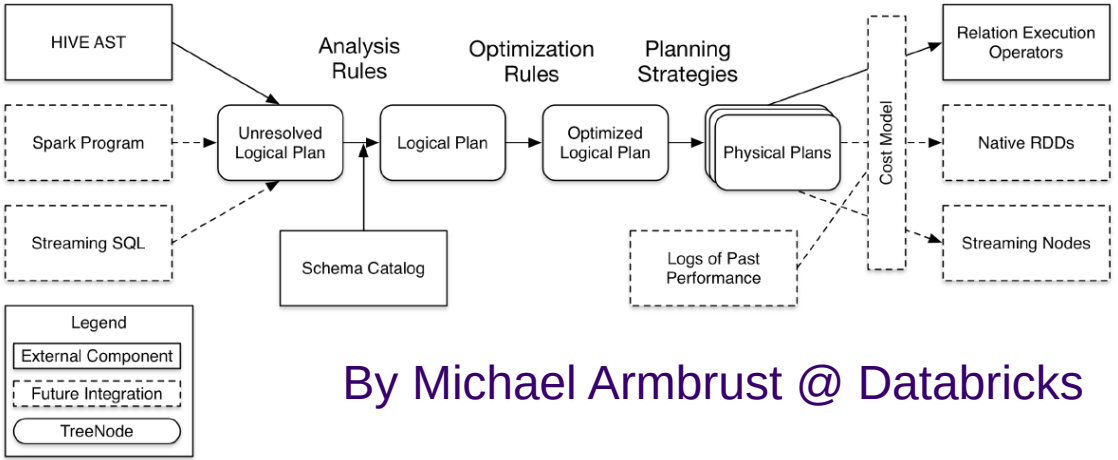


Sparkling:
Hiring and Evolving Open Source Technologies



SparkSQL

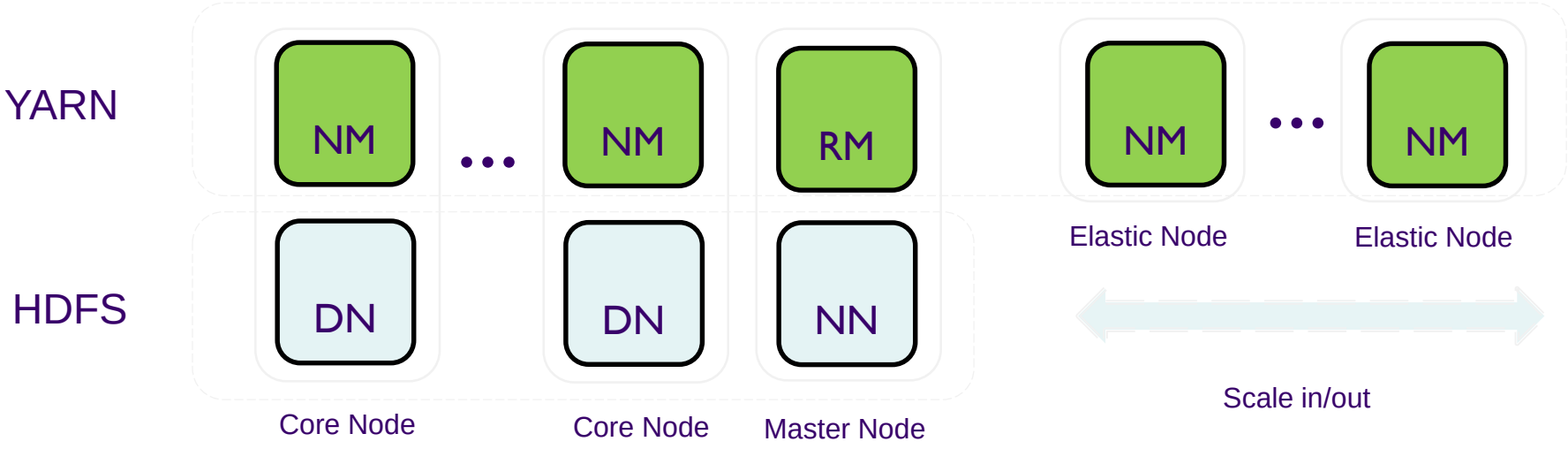
- Outstanding query engine
  - Open source with large and active community
  - Fully compatibility with ANSI SQL 2003
  - High performance
- Integration with Spark ecosystem
  - Spark Streaming
  - GraphX
  - MLlib/ML



By Michael Armbrust @ Databricks

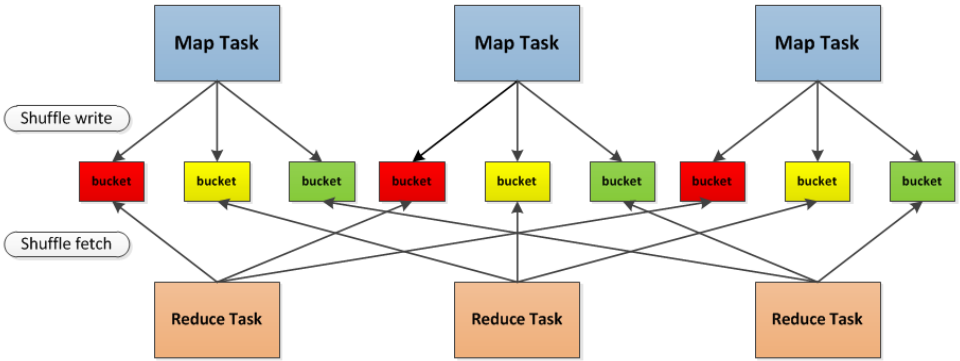
Elastic Deployment Architecture

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



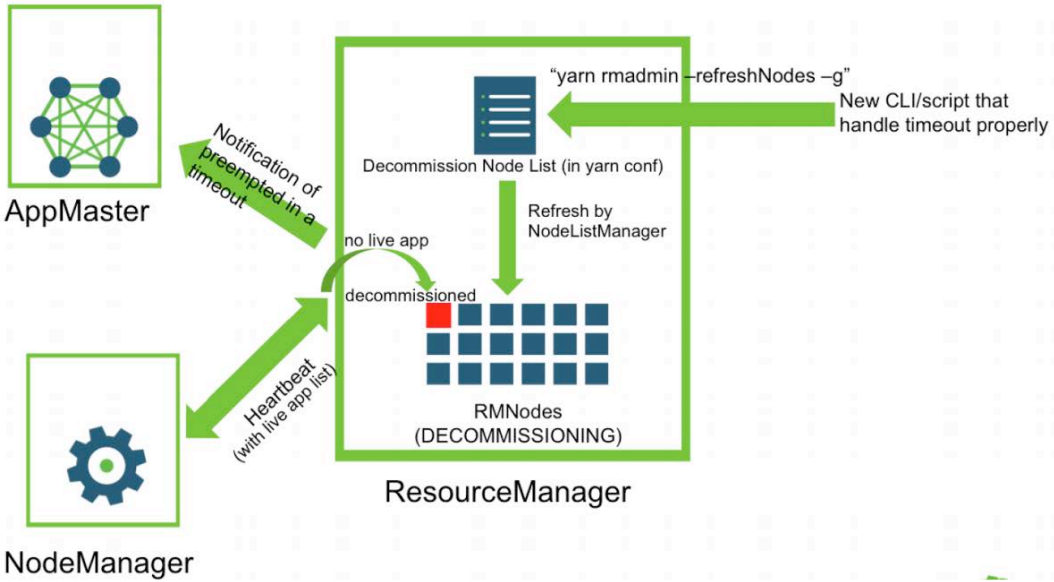
However, Elastic story is not so easy…

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

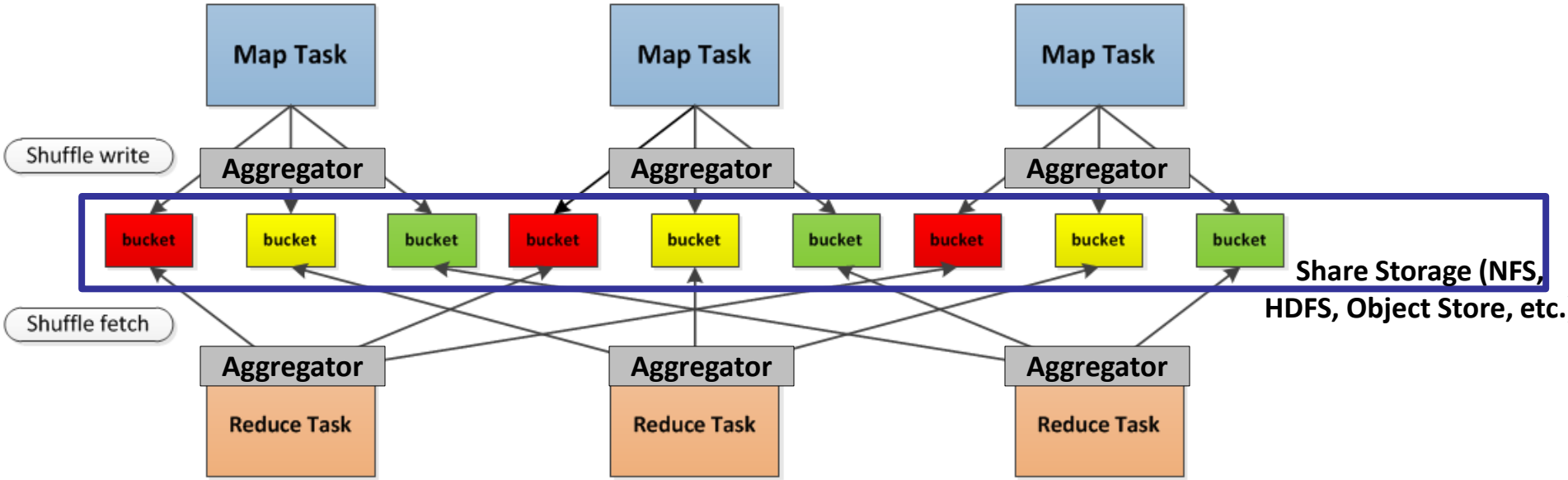


Key solution: Gracefully Decommission

- 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

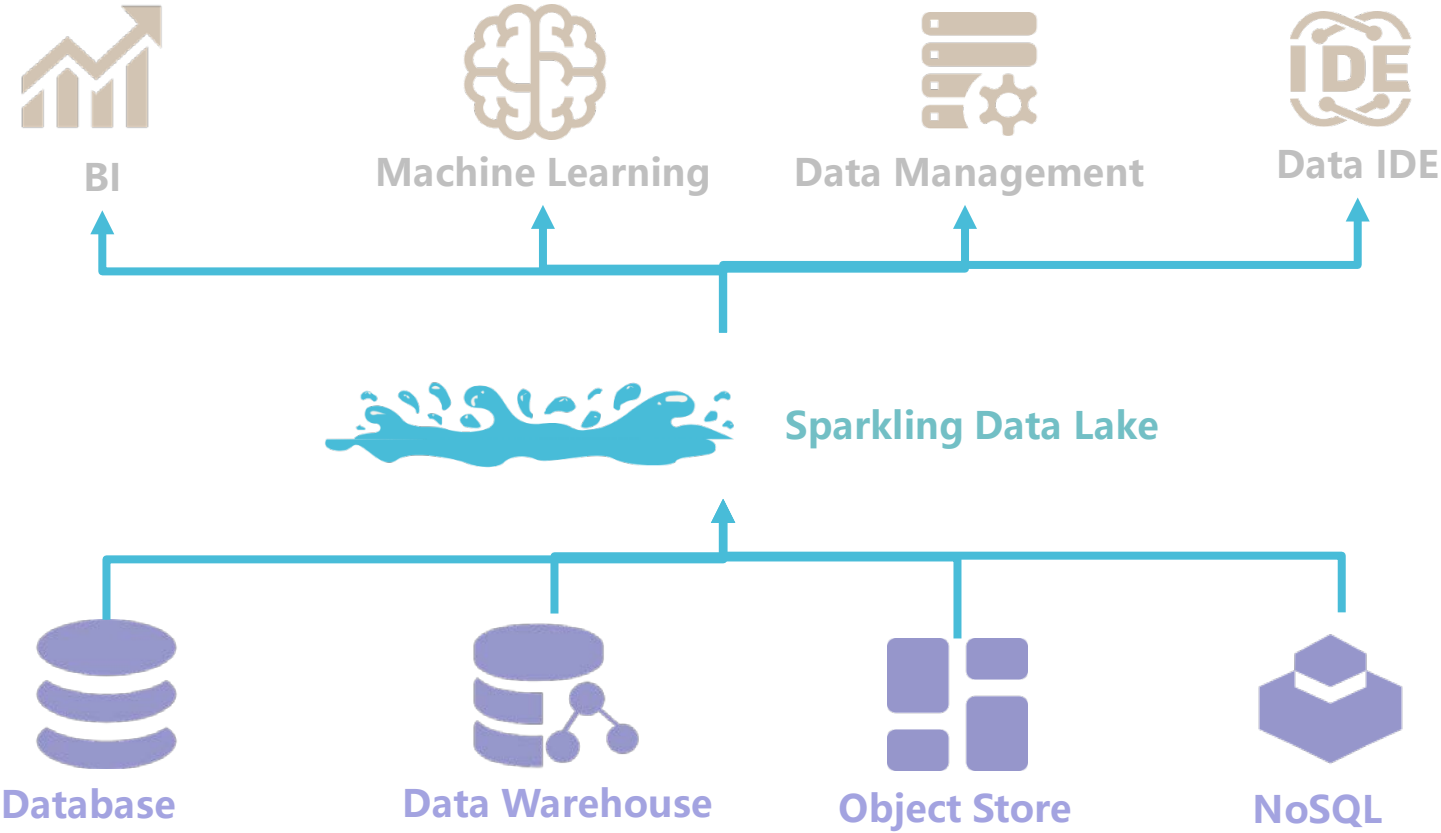


Key Solution: Independent Shuffle Service

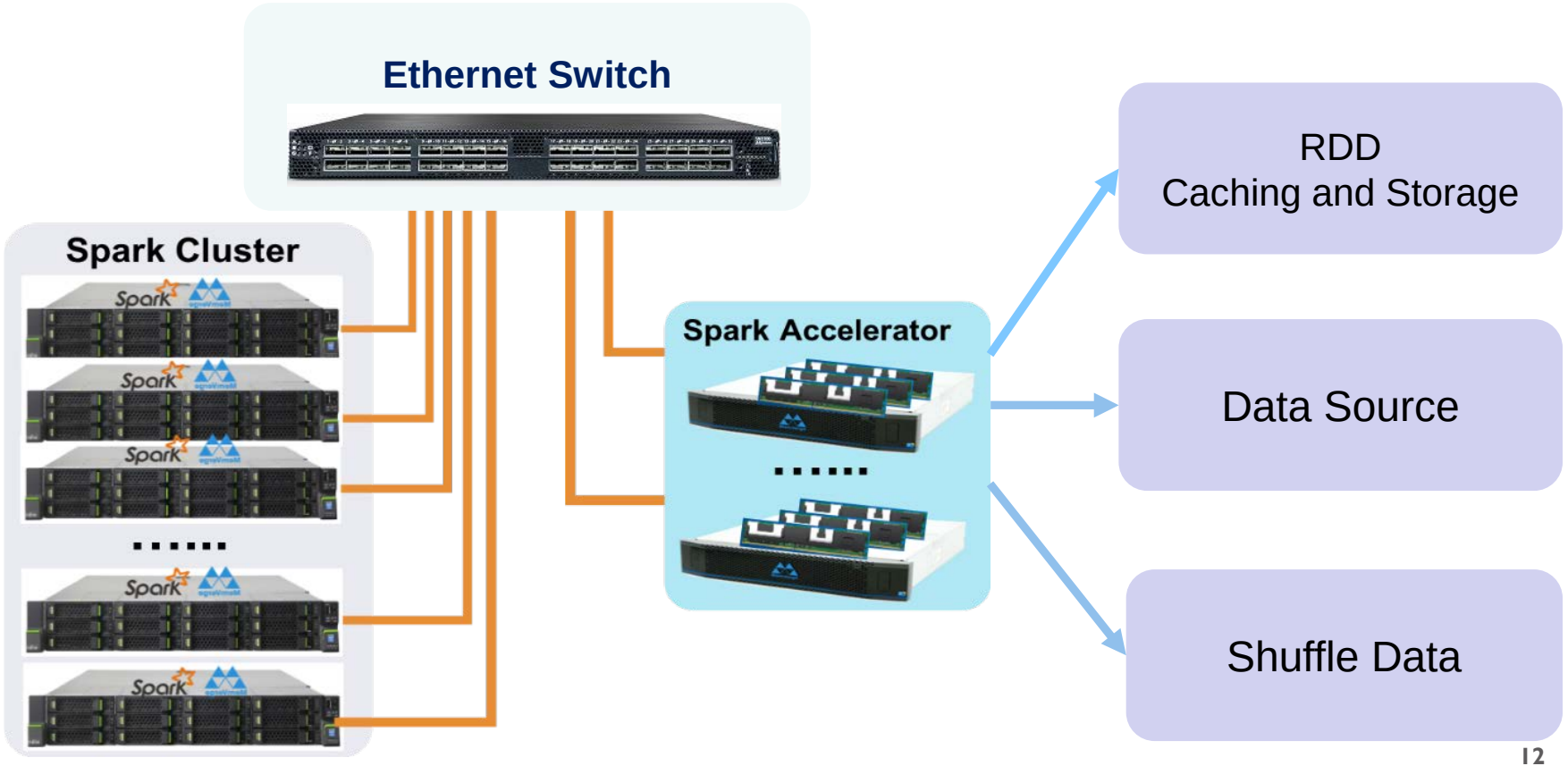


- Shuffle I/O are **decoupled** from a specific network/storage.
- Shuffle read and write can be implemented using **configurable** network transports and backend storage
- Listeners are inserted into different stages of the shuffle to apply hooks.

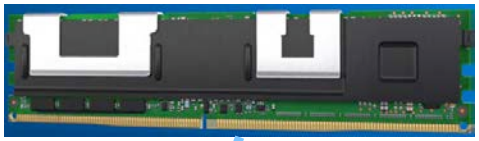
Sparkling:
Evolving from Data Warehouse to Data Lake



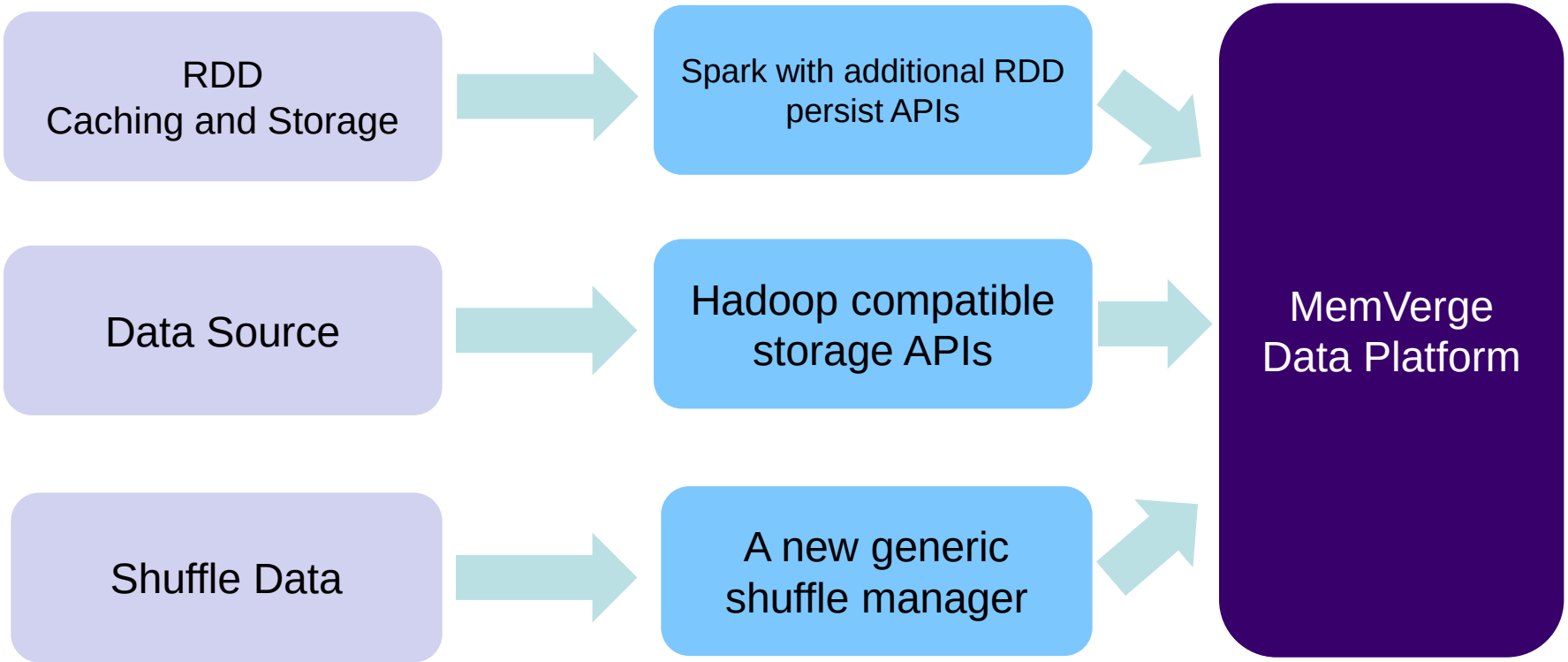
MemVerge Elastic Spark Solution



Intel® Optane™ DC Persistent Memory

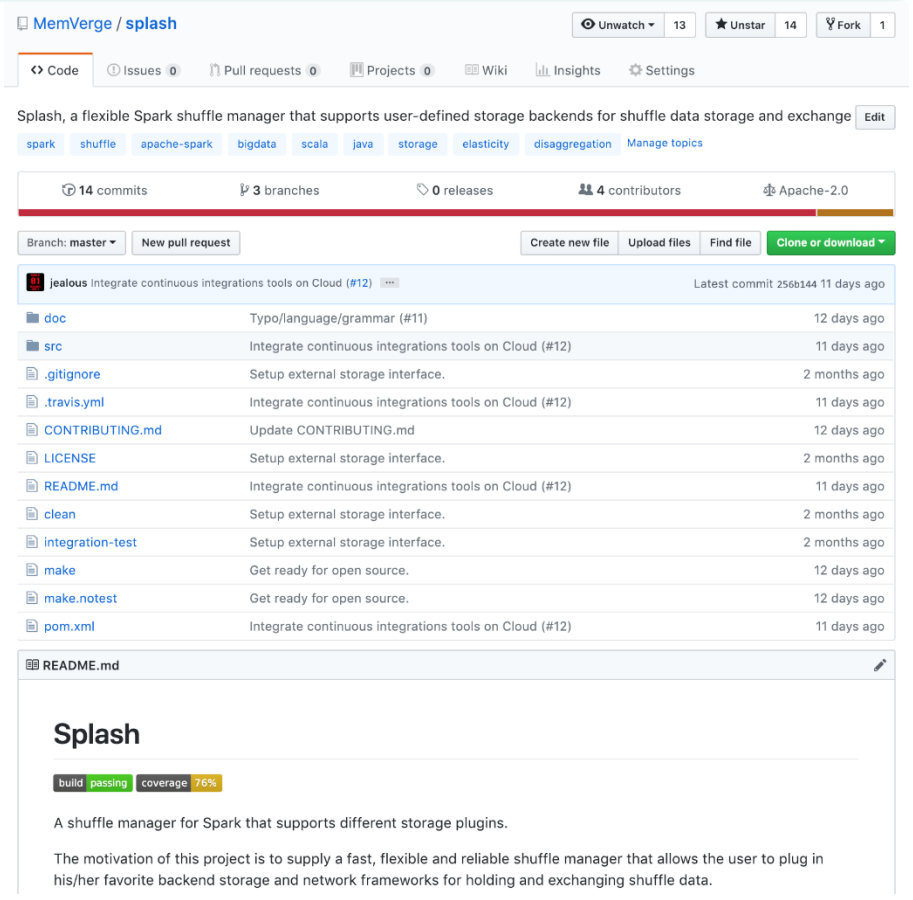


Spark Integration



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

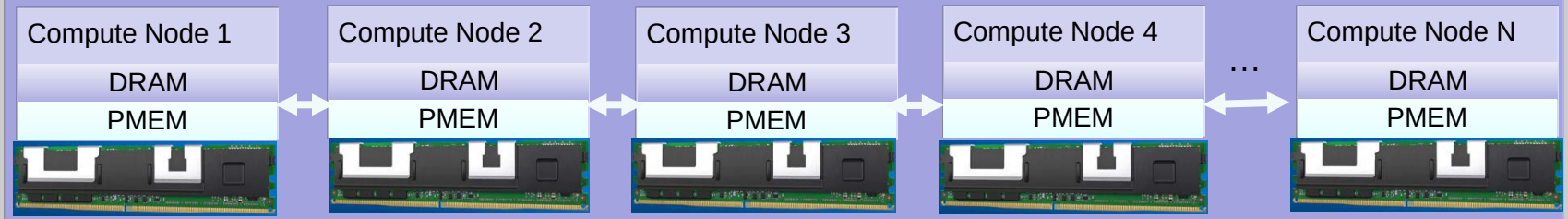


PMEM Centric Data Platform

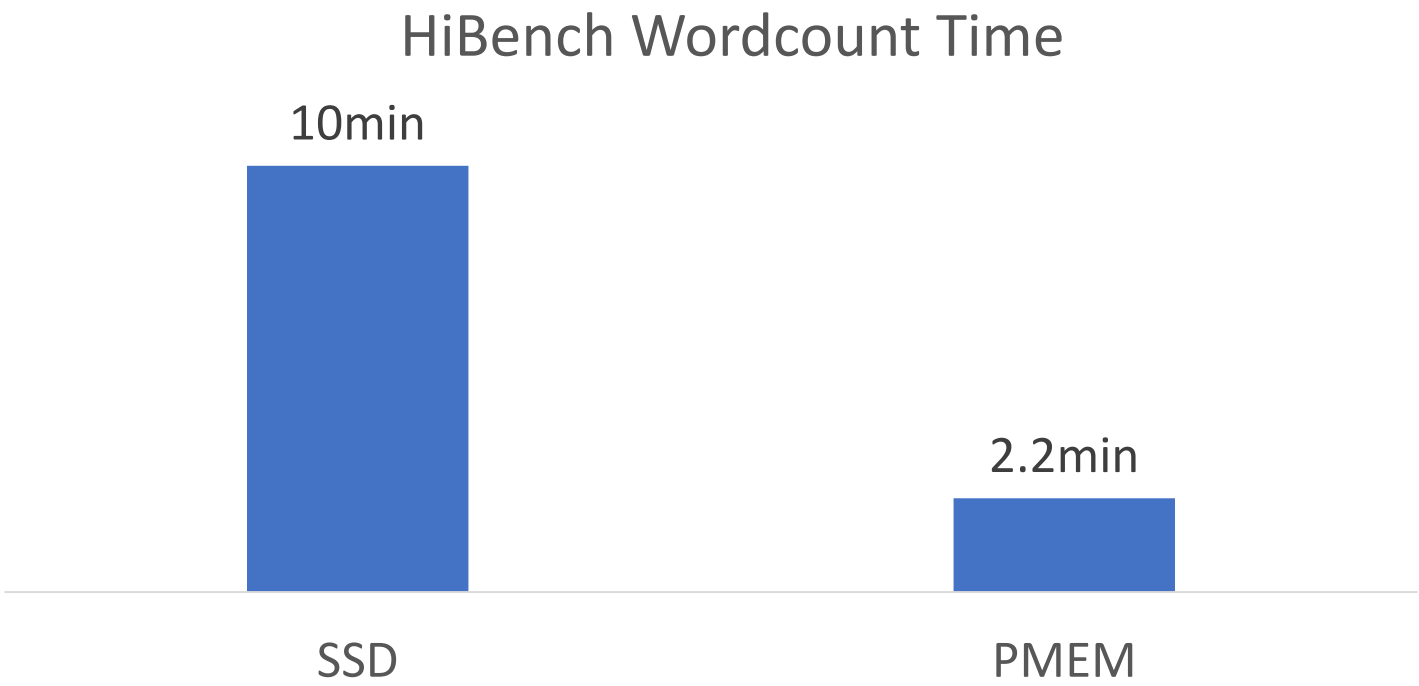
MemVerge Spark Adaptors

MemVerge SDK

Cluster Shared Persistent Memory

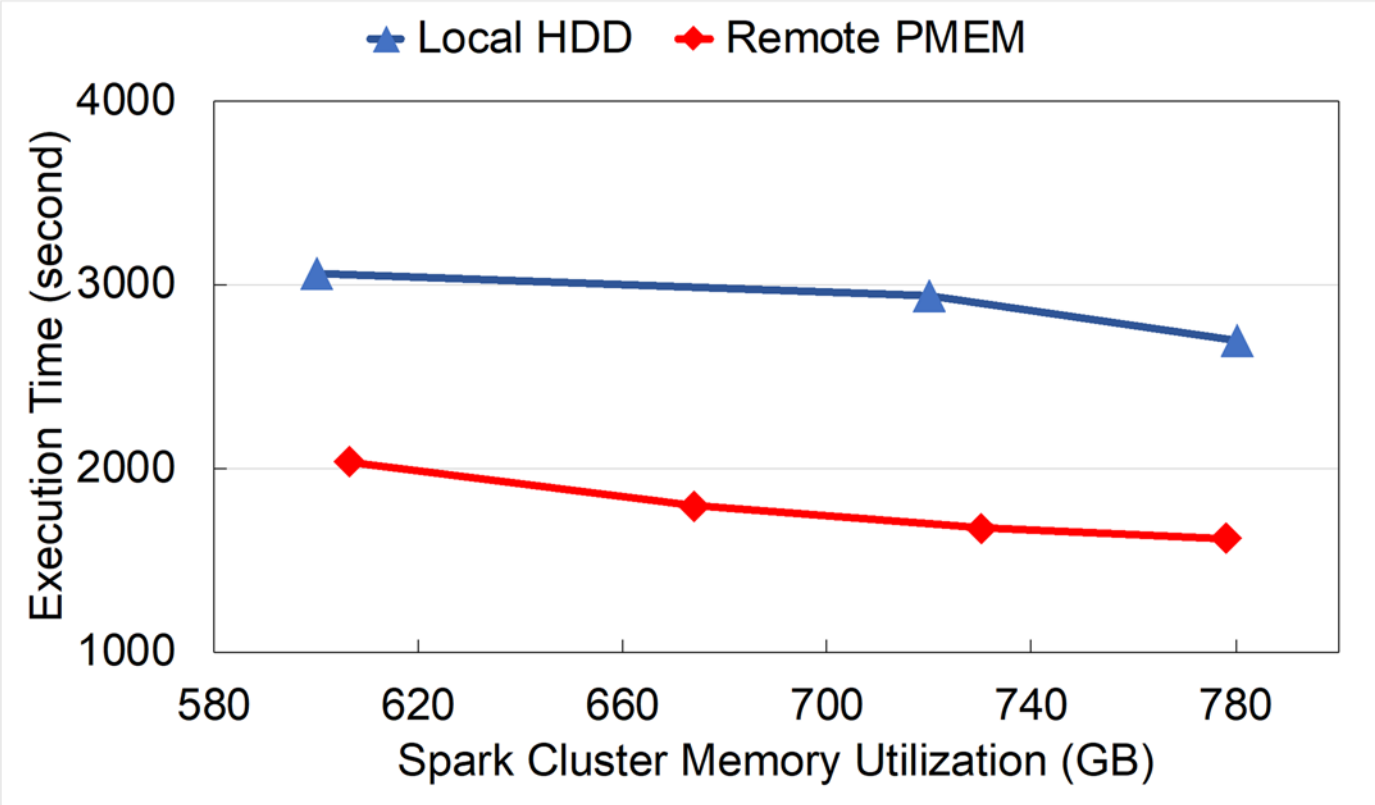


Source Data in Remote PMEM



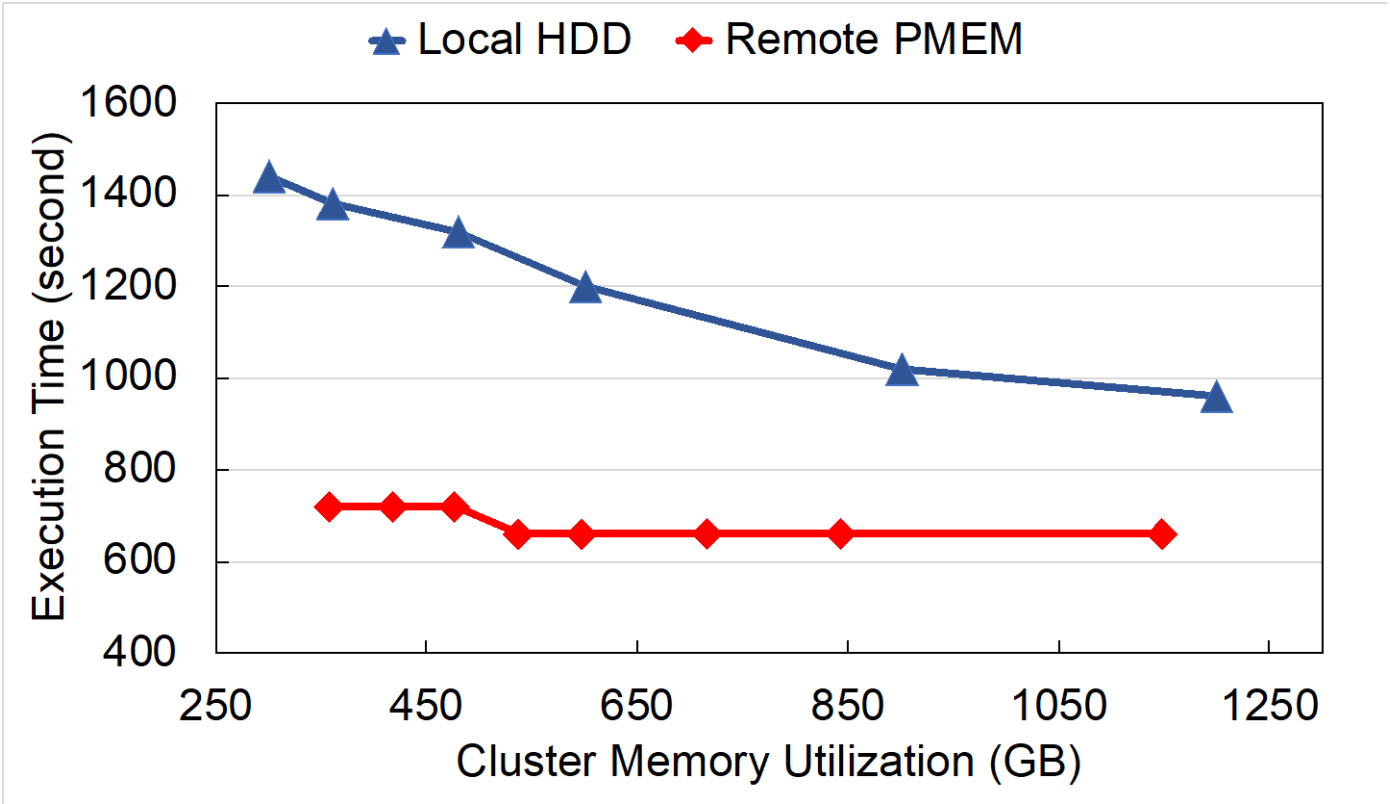
*5 Spark compute node, 1 remote PMEM node, Data size: 610GB

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

Ongoing Work

- TPC-DS performance study
- Better cloud readiness
  - Virtual machine support
  - Container support

Summary

- 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