Blink Runtime解析

马国维(黎钢) · 阿里巴巴 /高级技术专家

Apache Flink Meetup 杭州 – 2019年03月02日

Apache Flink

CONTENT

目录 >>

01 /

Stream Architecture

02 /

Blink In Alibaba

03 /

Blink Runtime Improvements

04 /

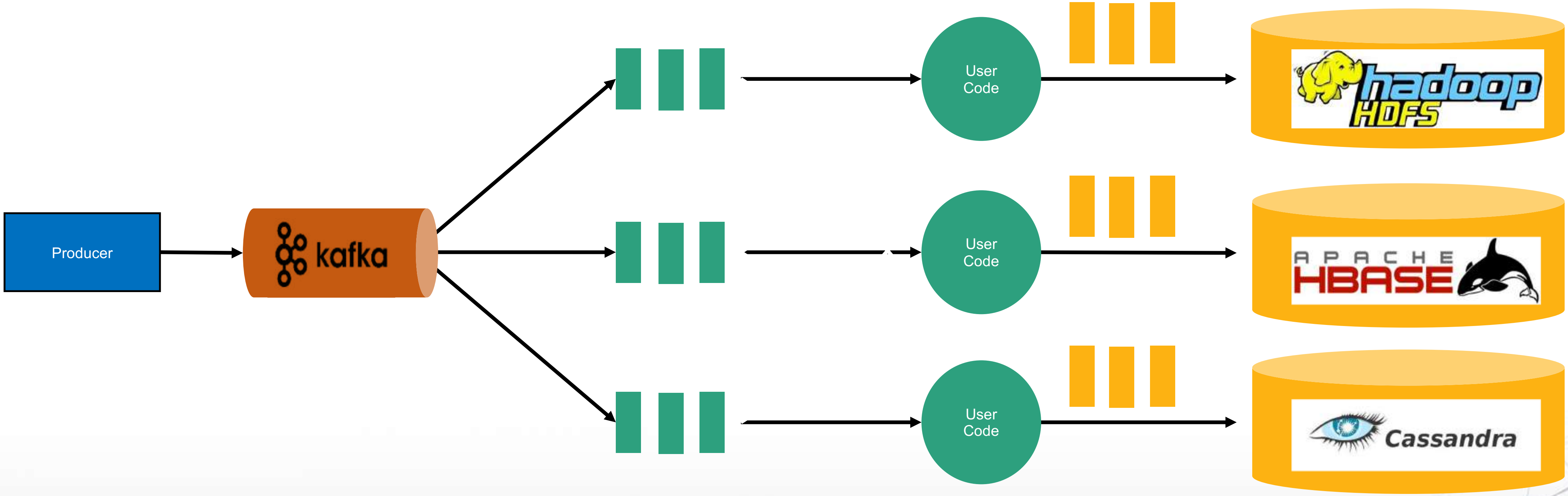
Future Plans

01

Stream Architecture

Apache Flink

Unbounded Stream



Stateful Stream

为什么要有状态

需要处理跨多条信息的计算

状态的一致性

Exactly Once/At Least Once

状态的管理

Checkpoint/Recovery

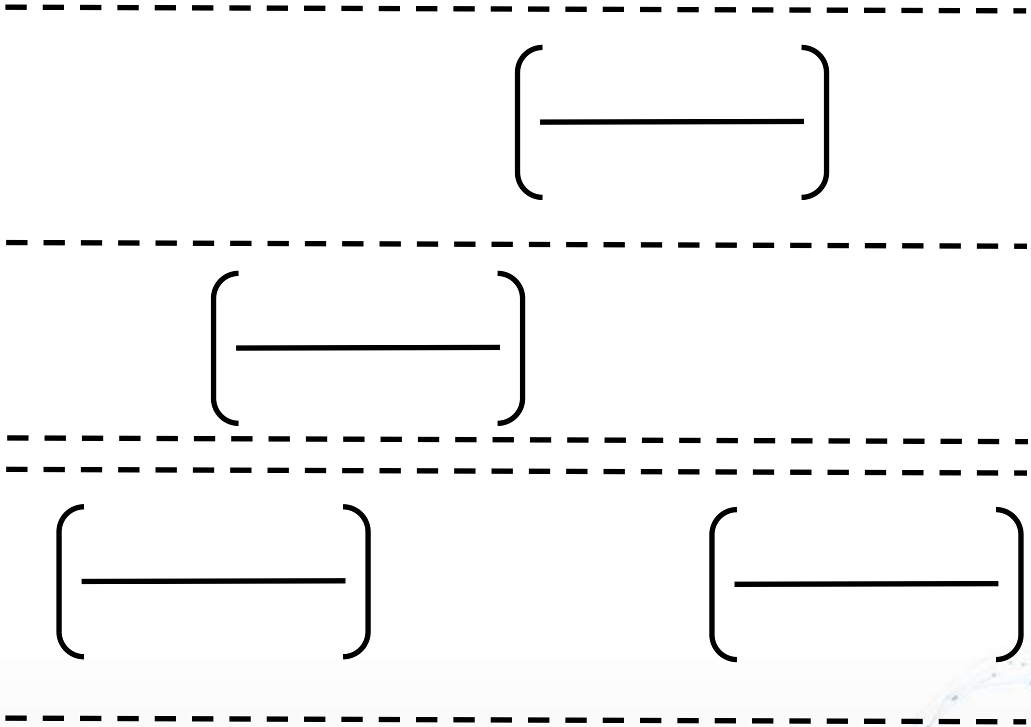
场景

Max/Min/TopN…

Window

Time-driven (examples: every 30 seconds)

Data-driven (examples: every 1000 records)



- Sliding window (no overlap)
- Tumbling window(with overlap)
- Session window (punctuated by a gap of inactivity)

Time

Event Time

和处理Event的时间无关

Watermark

完整性和延迟的一种权衡机制

Late Event Handling

提供更高完整性

Processing Time

低延迟，近似解

02

Blink In Alibaba

Blink In Alibaba

2016
Blink 接受双十一的检验

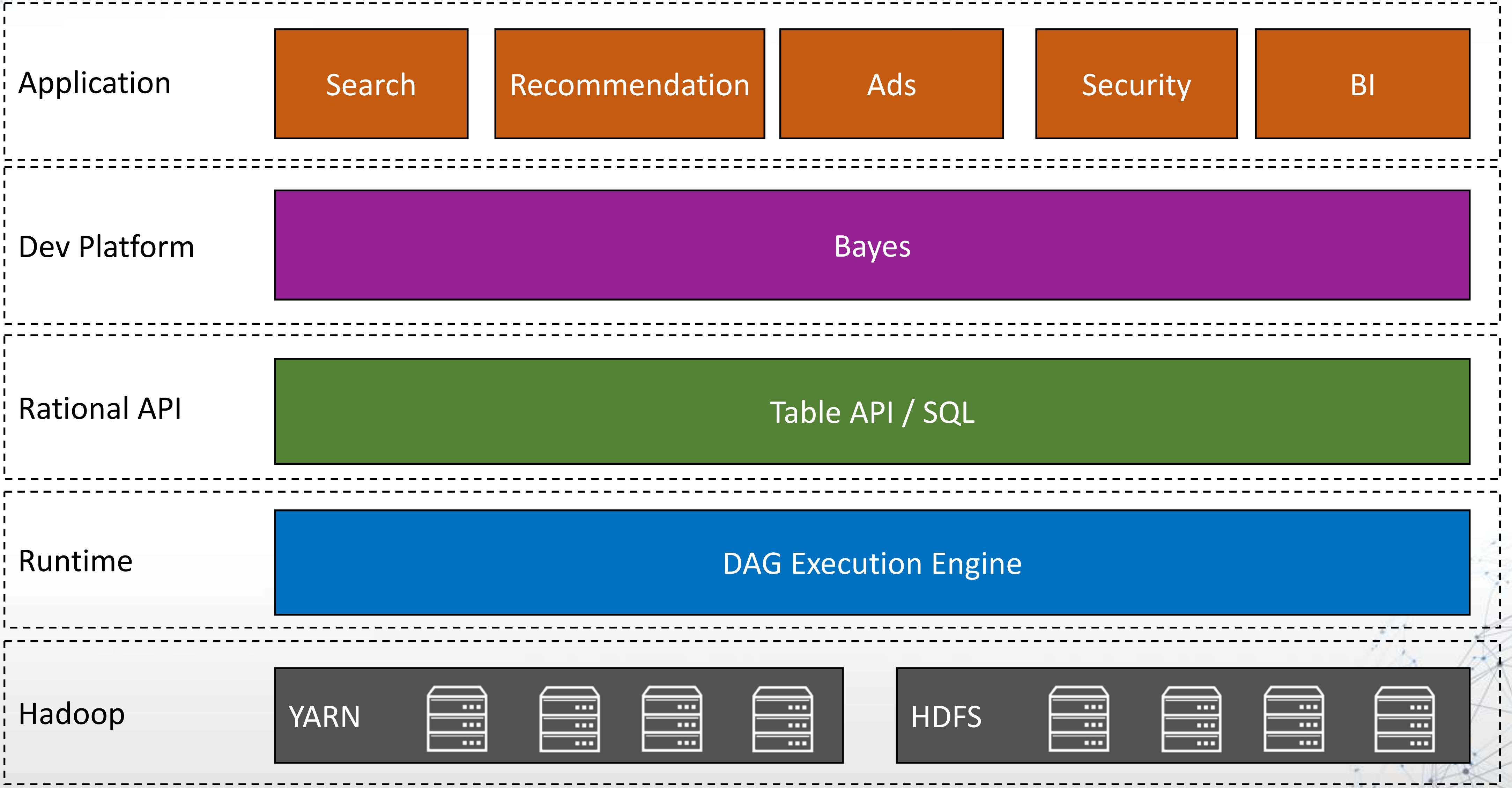
2017
Blink统一阿里实时流计算

2018
Blink 流批统一引擎

2019
Blink 在Github上正式开源

Apache Flink

Blink In Alibaba



Application: Search, Recommendation, Ads, Security, BI

Dev Platform: Bayes

Rational API: Table API / SQL

Runtime: DAG Execution Engine

Hadoop: YARN, HDFS

03

Blink Runtime Improvements

Outline

Architecture

架构

Performance

性能

Availability

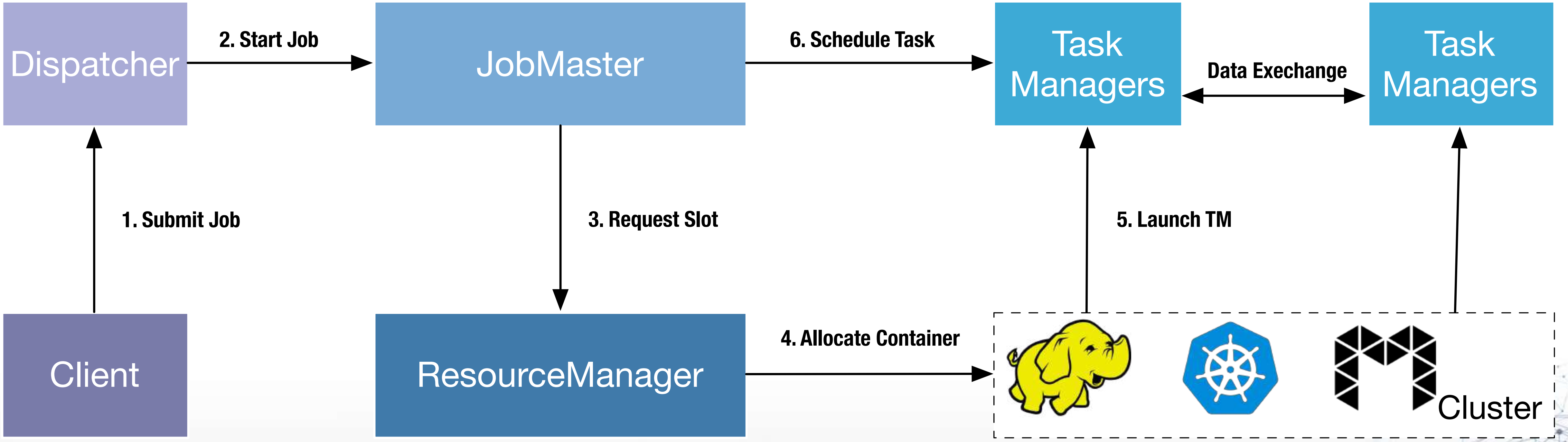
稳定性

Ecosystem

生态

Apache Flink

Blink Architecture

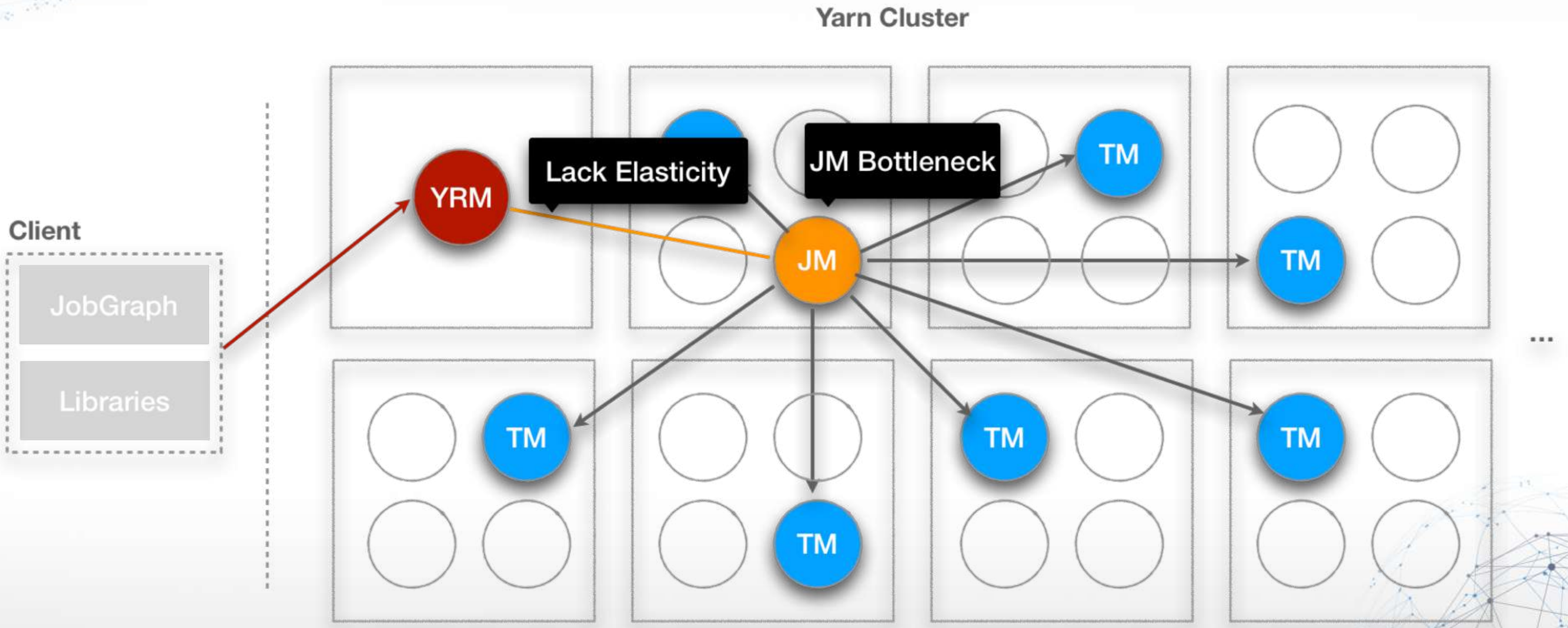


Apache Flink

Blink Runtime Improvements

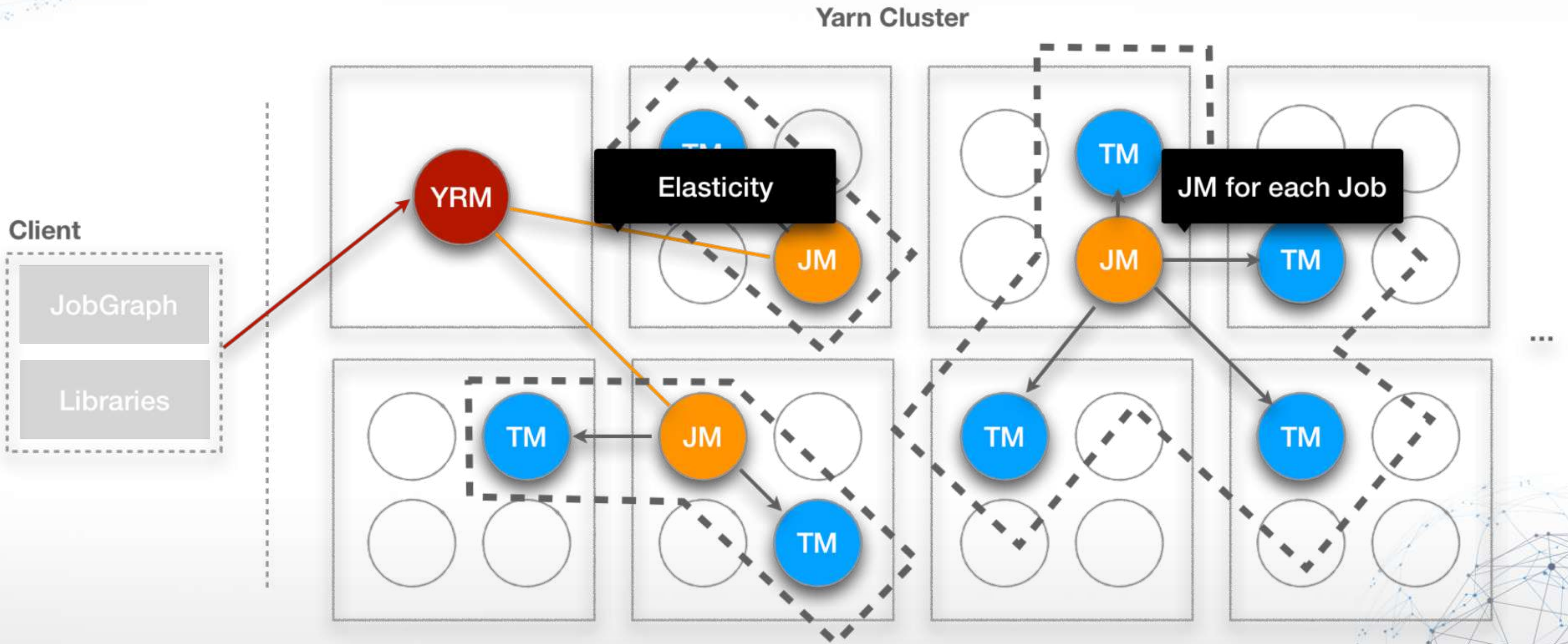
Architecture

Before FLIP-6



Apache Flink

FLIP–6



Apache Flink

Pluggable Shuffle Architecture

Low Resource Utilization

Task 执行完无法释放TM，其他JOB无法使用

Pluggable Shuffle Architecture

提供标准的Interface让用户扩展

External Shuffle

需要从API、Compile、JM、TM都需要改动

External Shuffle Service

实现若干接口即可

Ecosystem

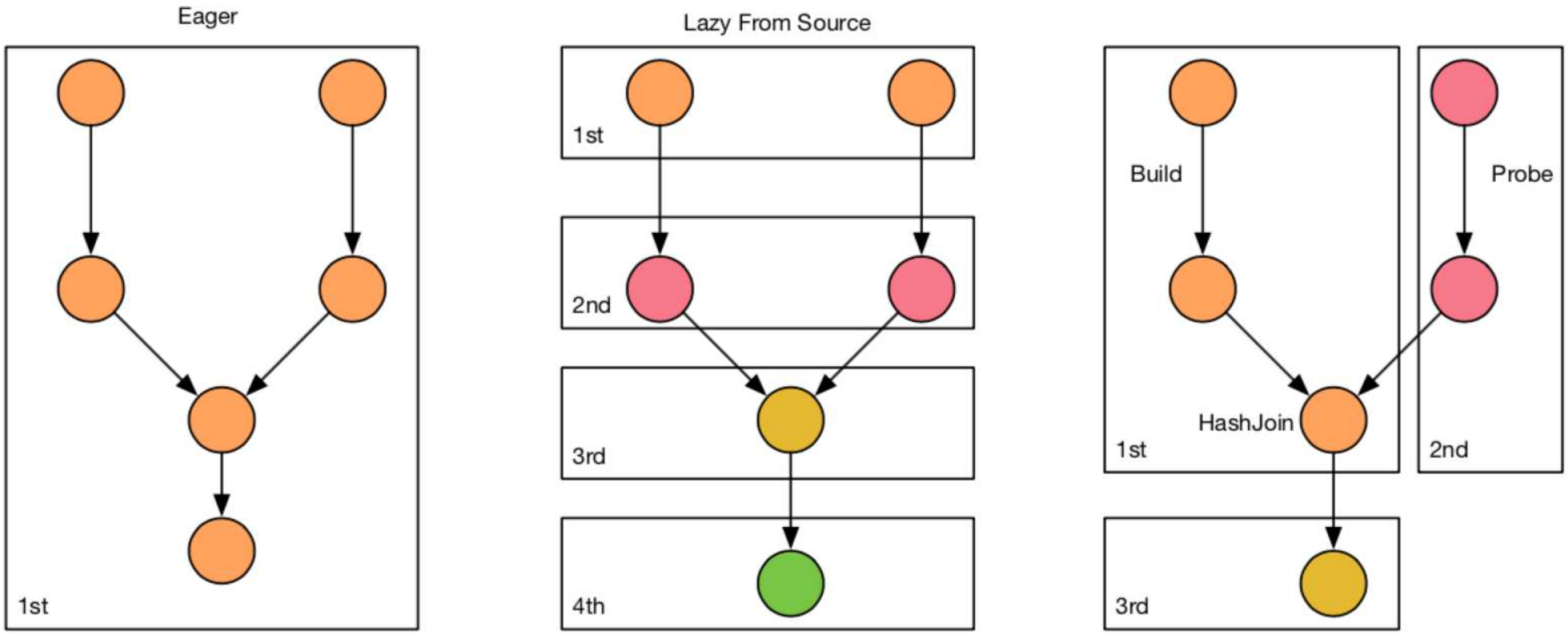
Shuffle 生态不够丰富没有用户进行扩展

Customized

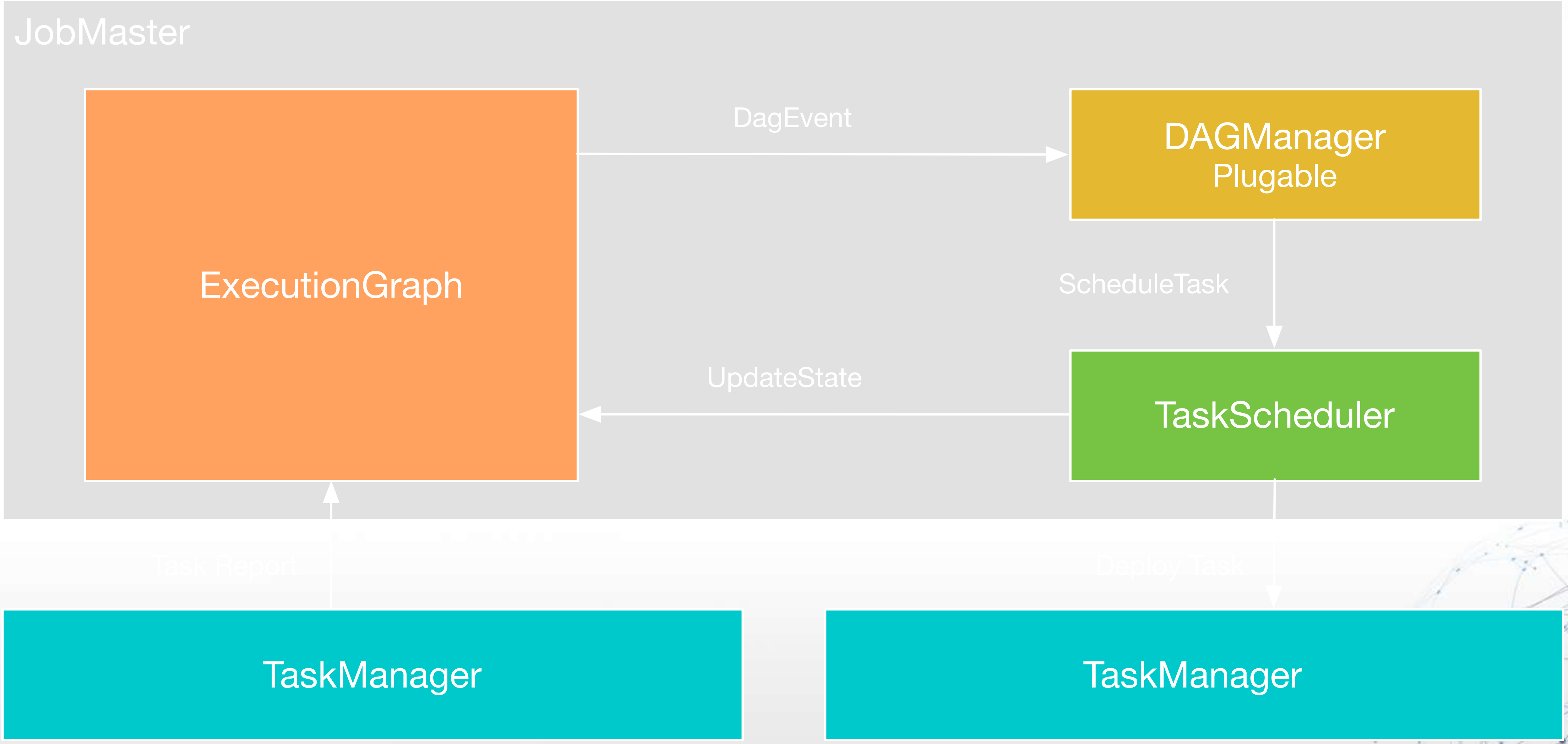
用户可以根据新硬件，新架构定制Shuffle，丰富Shuffle生态

Apache Flink

Pluggable Scheduler Architecture



Pluggable Scheduler Architecture



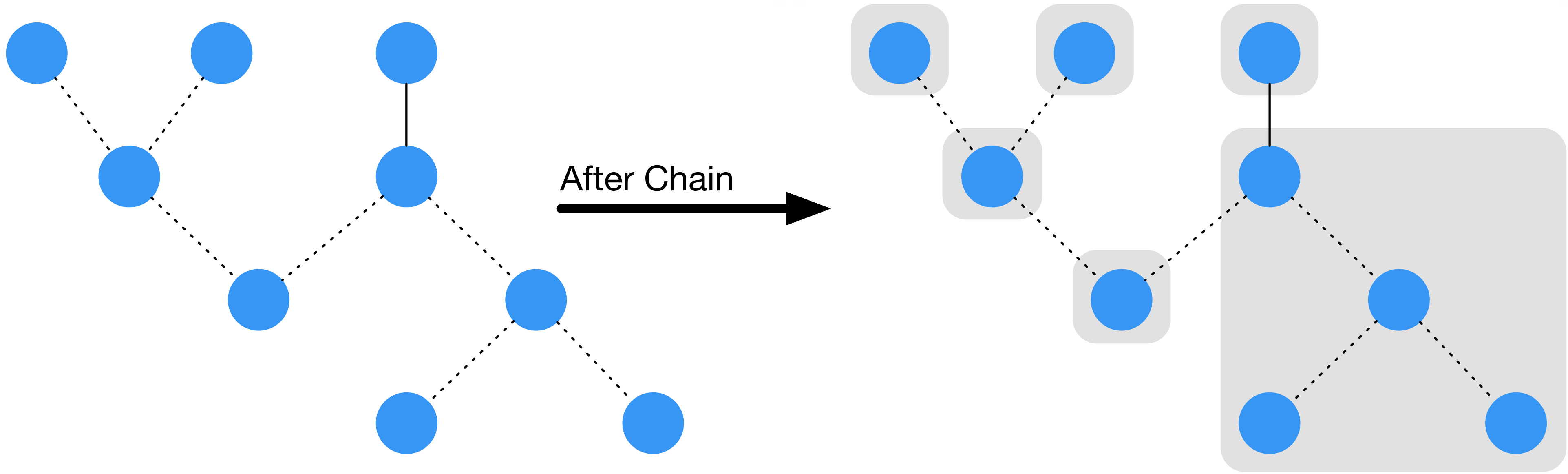
Apache Flink

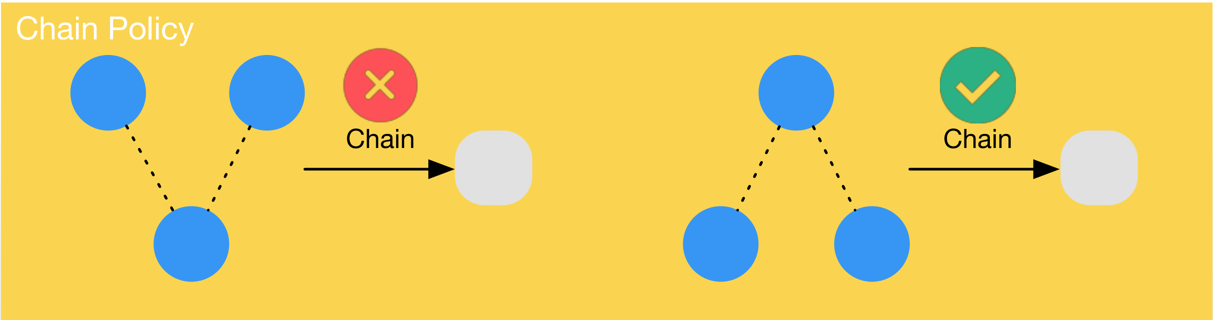
Blink Runtime Improvements

Performance

Apache Flink

Operator Dag

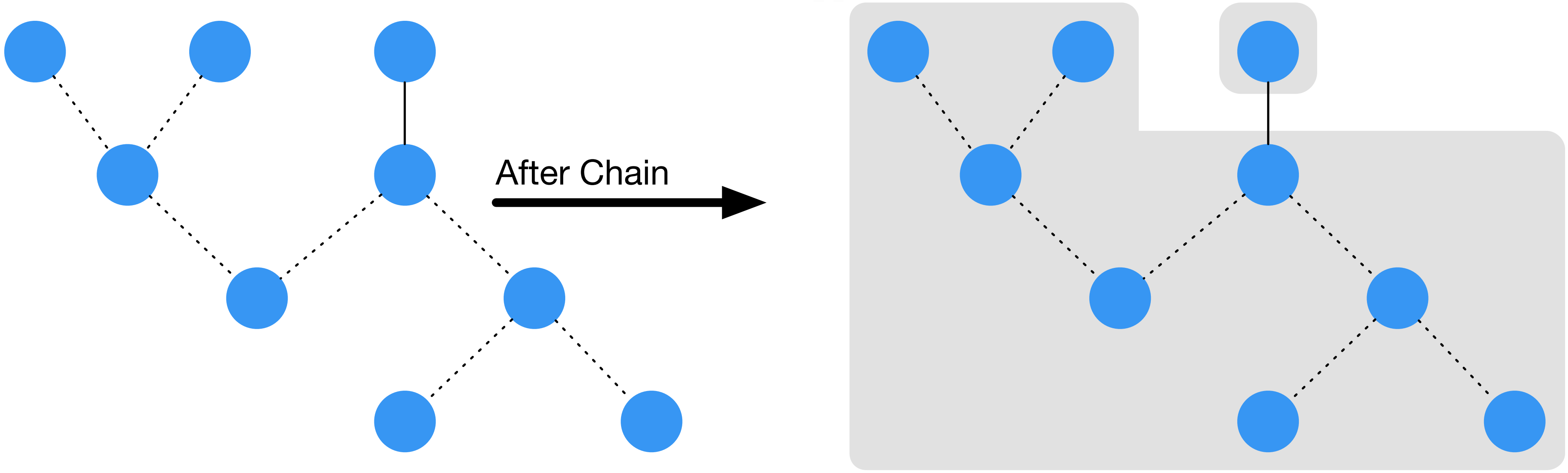


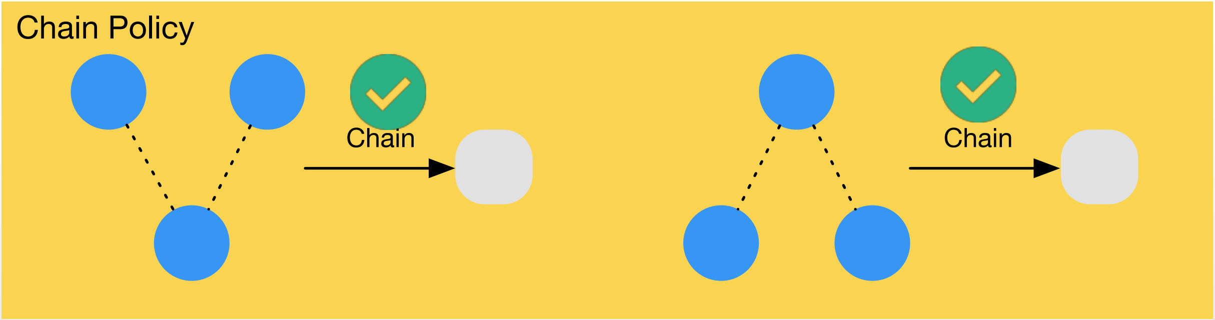


▢ JobVertex ● Operator

...... Forward ―― KeyBy

Operator Dag





▢ JobVertex ● Operator

...... Forward ——— KeyBy

Apache Flink

Batch Shuffle

Large Scale

Task会产生大量的文件

IO

IOPS高，数据量大

Cache

操作系统自动淘汰

Merge

减少文件个数

Reduce IO

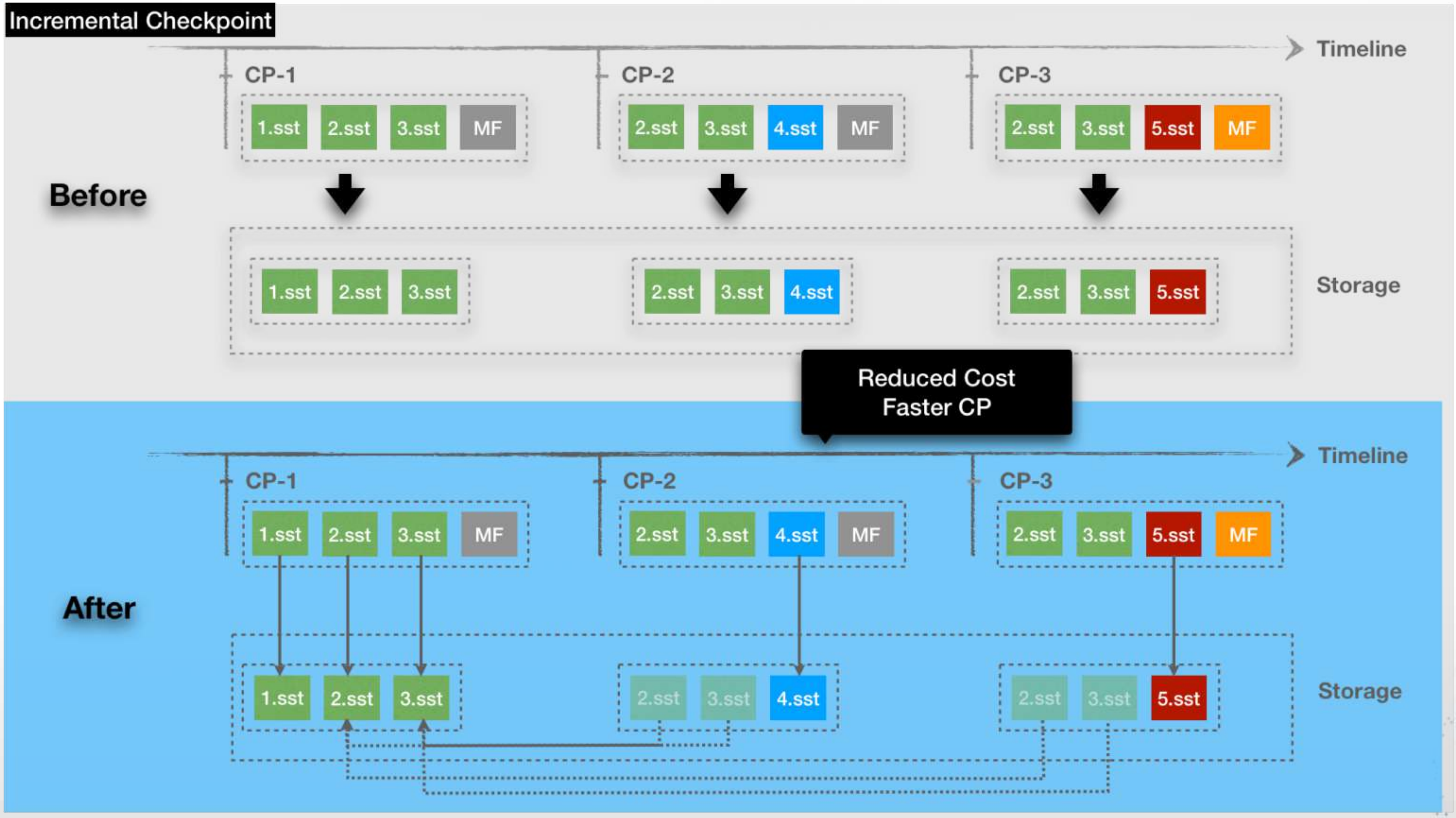
减少不必要IO，通过压缩减少数据量

Managed Cache

主动Prefetch和Drop

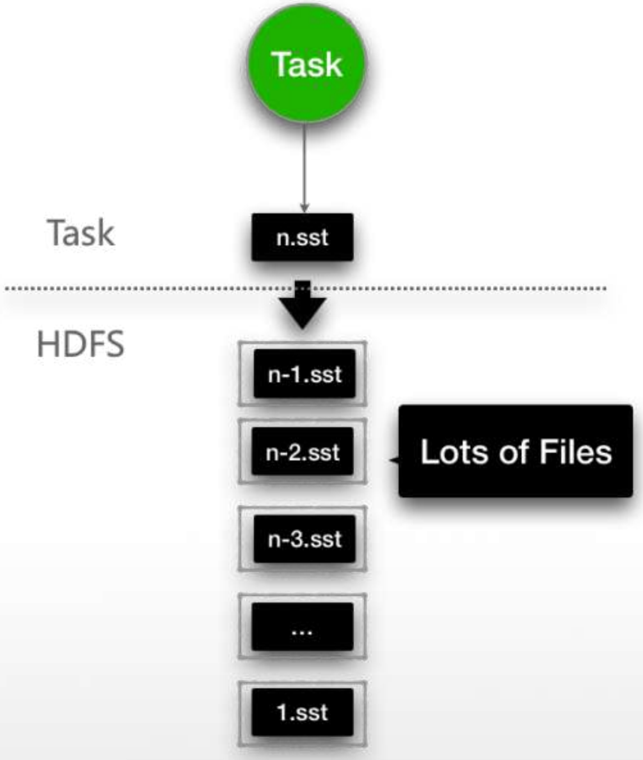
Apache Flink

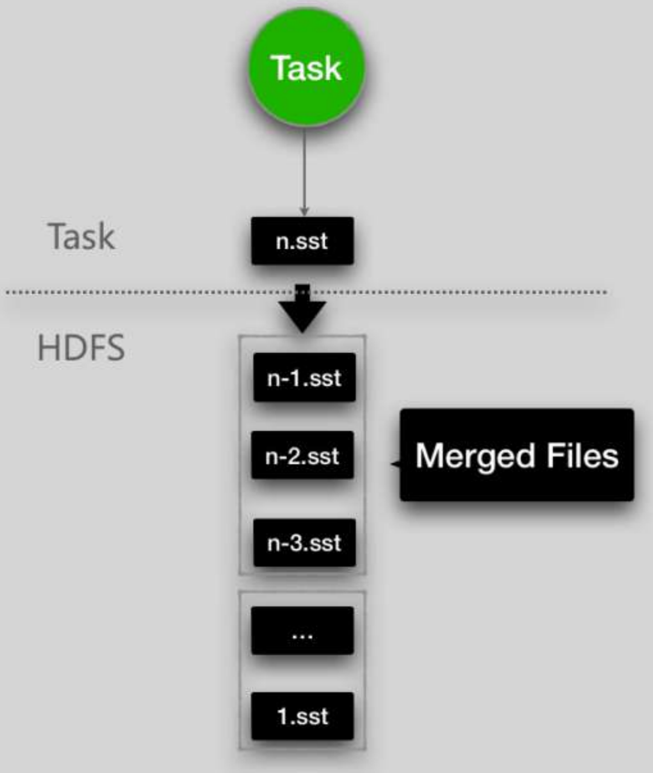
Incremental Checkpoint



Merge Checkpoint File

Checkpoint Performance



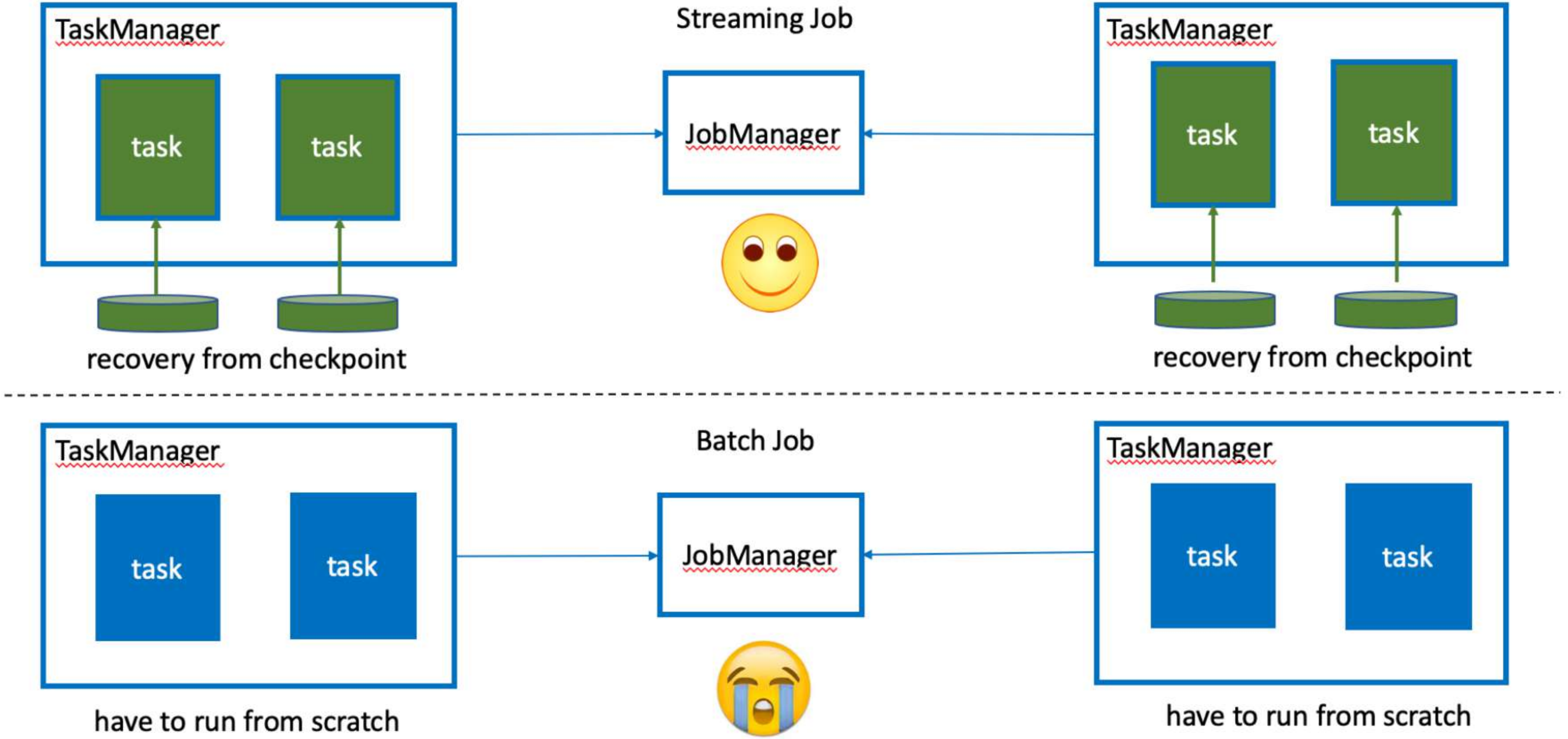


Apache Flink

Blink Runtime Improvements

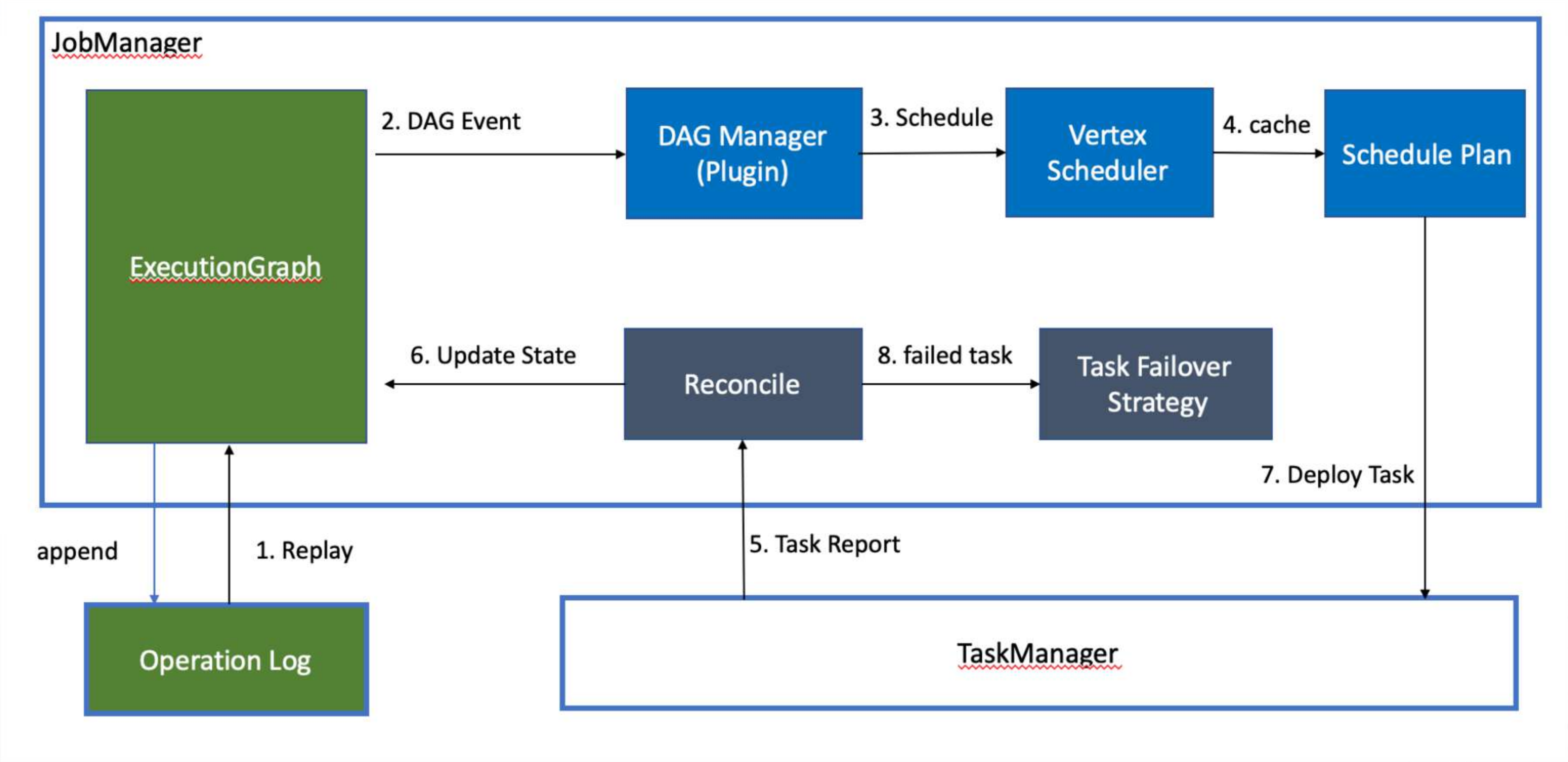
Availability

JM FailOver


TaskManager
task
task
recovery from checkpoint
Streaming Job
JobManager
TaskManager
task
task
recovery from checkpoint
TaskManager
task
task
have to run from scratch
Batch Job
JobManager
TaskManager
task
task
have to run from scratch

Apache Flink

JM FailOver



Apache Flink

Blink Runtime Improvements

Ecosystem

Ecosystem

K8S

Native On K8s

Connectors

Gemni–StateBackend

Java Based

Yarn Shuffle

Yarn Aux Service

04

Future Plans

Apache Flink

Future Plans

- Deployment
  - Unified Elastic Session For FLIP-6
- Job Schedule
  - Dynamic Update of JobGraph for Batch Job
  - Hotupdate of JobGraph for Streaming Job
- Network Stack
  - More kinds of external ShuffleService : RDMA
- Checkpoint
  - CheckPoint for Batch processing
- New API Stack

THANKS

Flink China社区大群

扫一扫群二维码，立刻加入该群。