FLINK FORWARD # ASIA 2022 ONLINE | 实时即未来

# Flink OLAP 在字节跳动的查询优化和落地实践

何润康 | 字节跳动基础架构工程师

# 01 字节 Flink OLAP 介绍

1. 业务落地情况
2. 总体架构 & 业务架构
3. 业务落地挑战

# 业务落地情况

业务规模

12+ 核心业务方

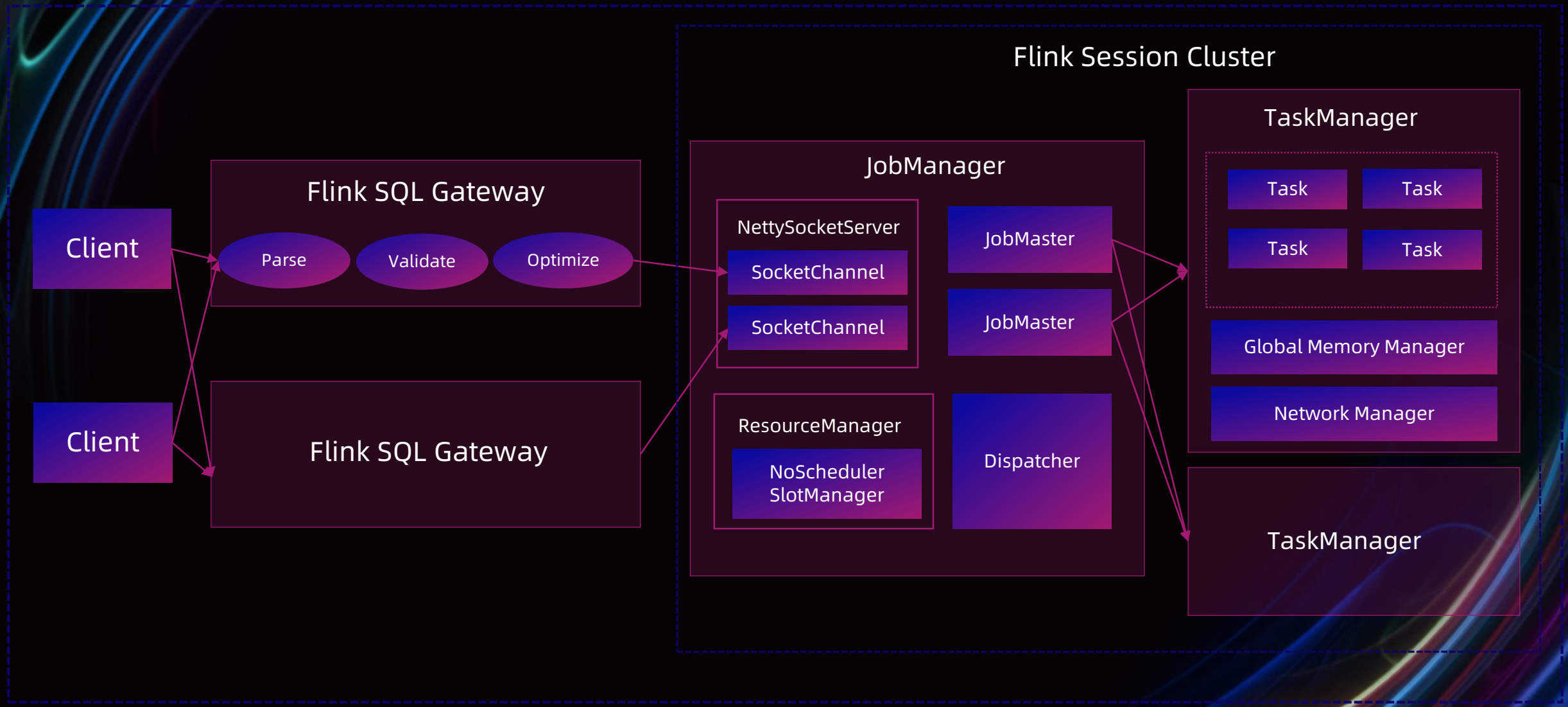
集群规模

1.6w Core 资源

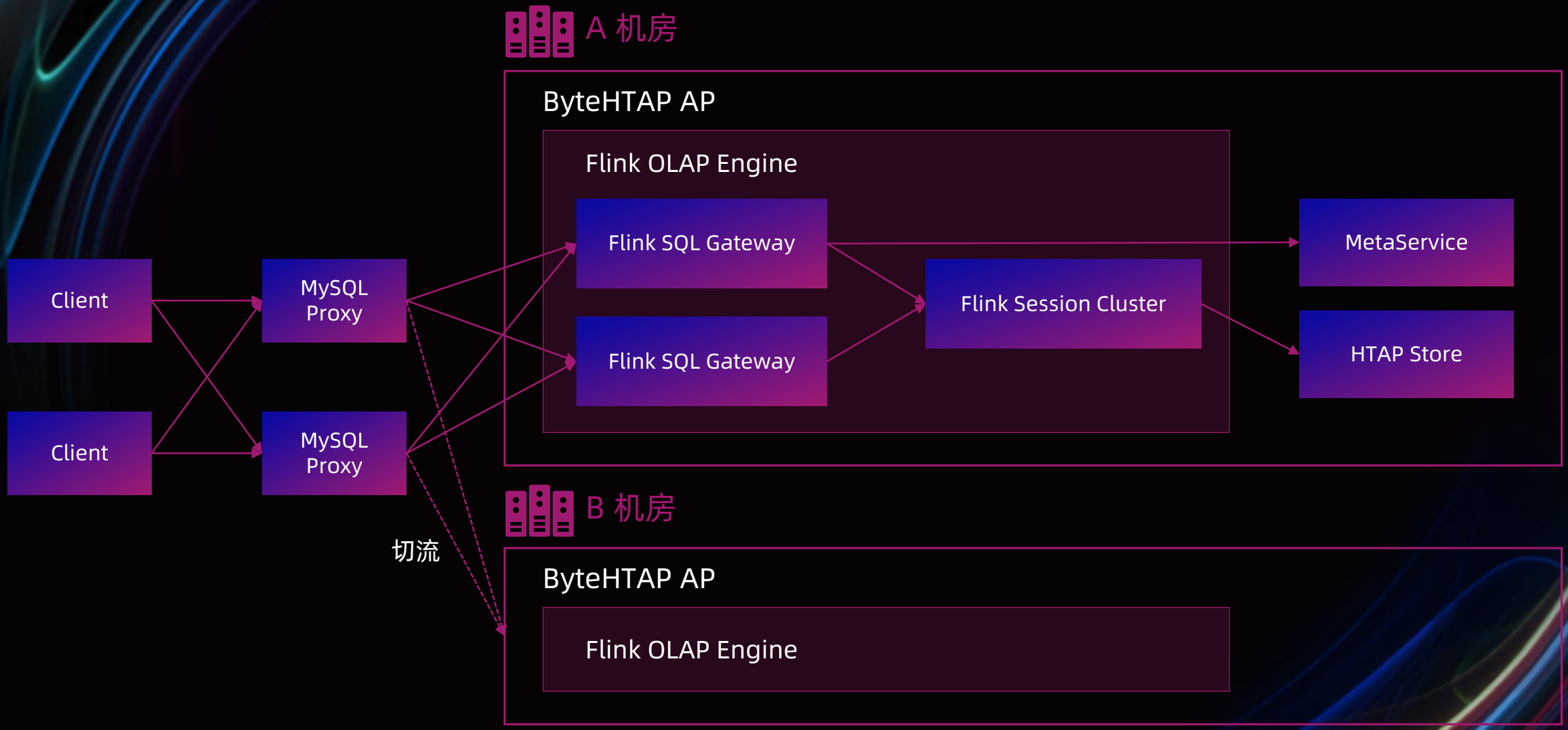
查询规模

每天 Query 50w+

# 总体架构


Flink Session Cluster
Client
Client
Flink SQL Gateway
Parse
Validate
Optimize
Flink SQL Gateway
JobManager
NettySocketServer
SocketChannel
SocketChannel
JobMaster
JobMaster
ResourceManager
NoScheduler
SlotManager
Dispatcher
TaskManager
Task
Task
Task
Task
Global Memory Manager
Network Manager
TaskManager

# 业务架构


A 机房
ByteHTAP AP
Flink OLAP Engine
Flink SQL Gateway
Flink SQL Gateway
Flink Session Cluster
MetaService
HTAP Store
Client
Client
MySQL Proxy
MySQL Proxy
切流
B 机房
ByteHTAP AP
Flink OLAP Engine

# 性能挑战



流式

端到端 latency 和稳定性

批式

处理速度和吞吐

OLAP

亚秒级的 latency 和高查询 QPS

# 运维和稳定性挑战



运维

测试流程

无感升级

监控

监控体系

稳定性

容灾能力

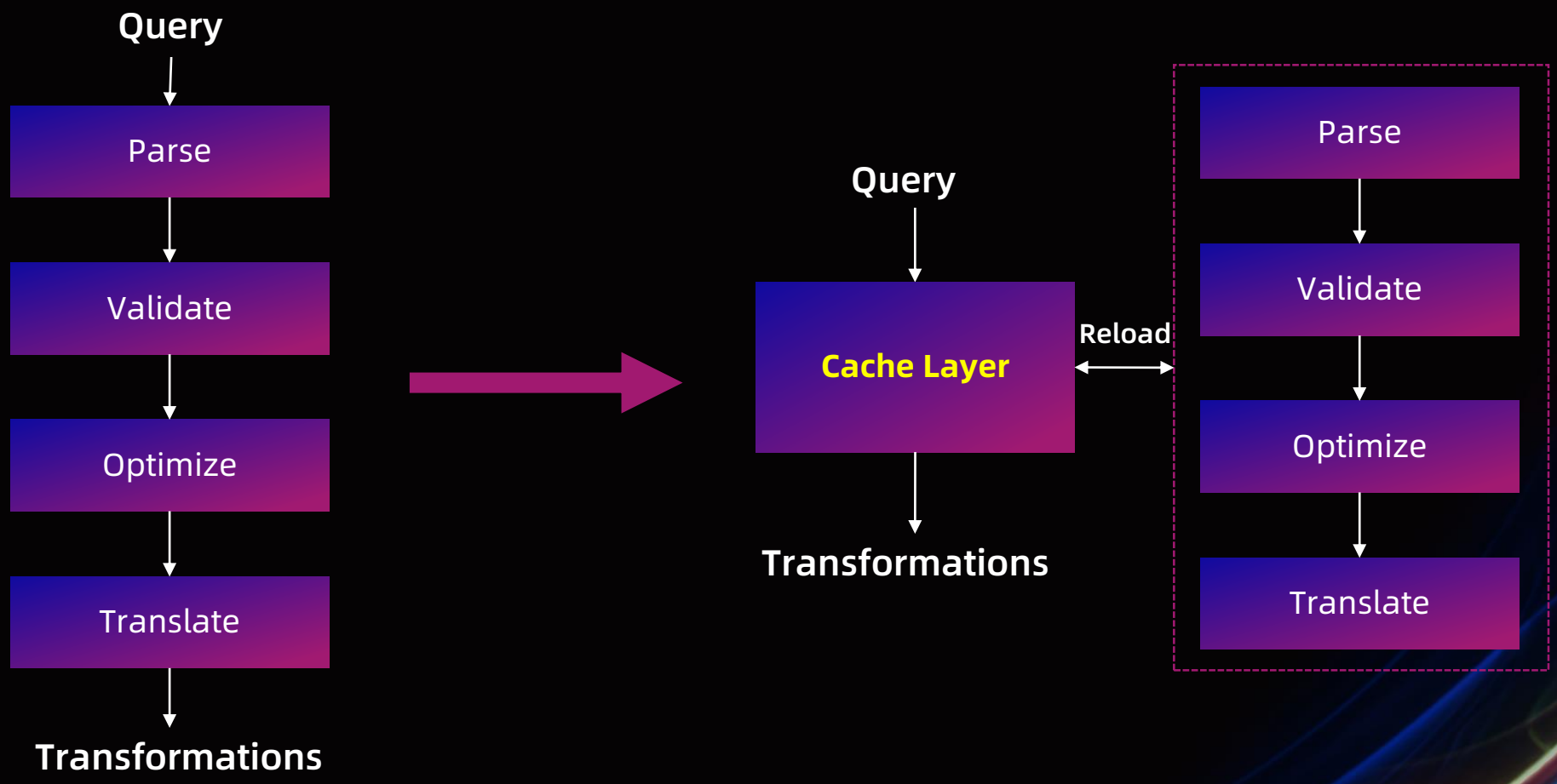
Full GC 治理

# 02 查询优化

1. Query Optimizer 优化
2. Query Executor 优化

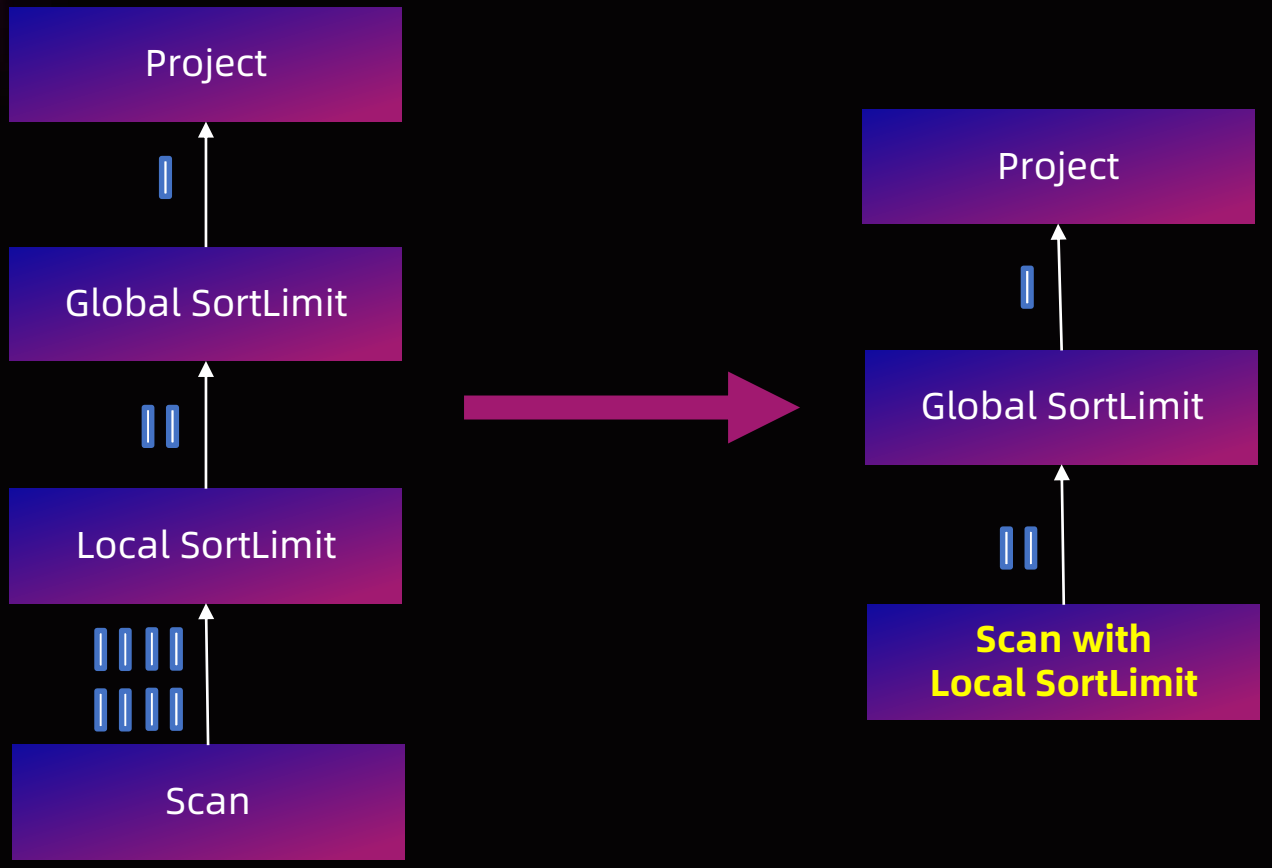
# Query Optimizer 优化

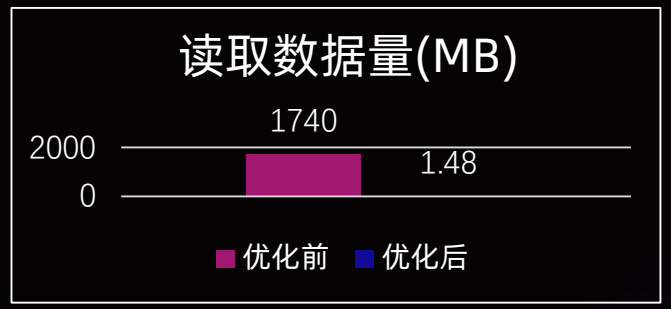
## Plan 缓存

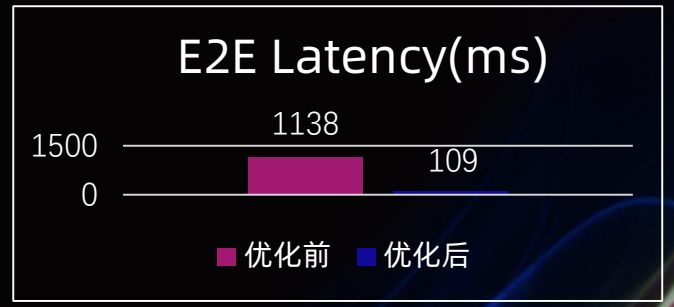


# Query Optimizer 优化

## TopN 下推

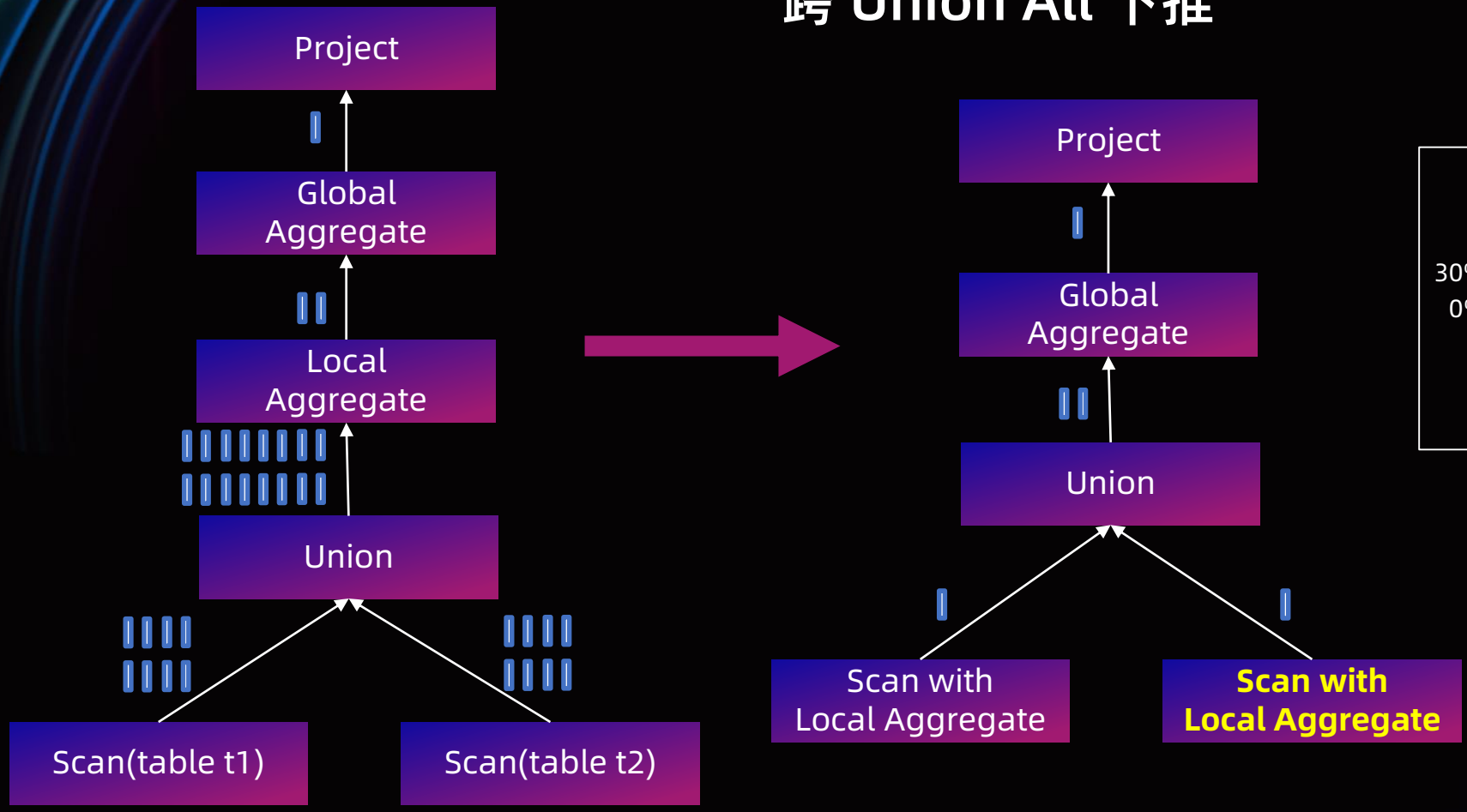


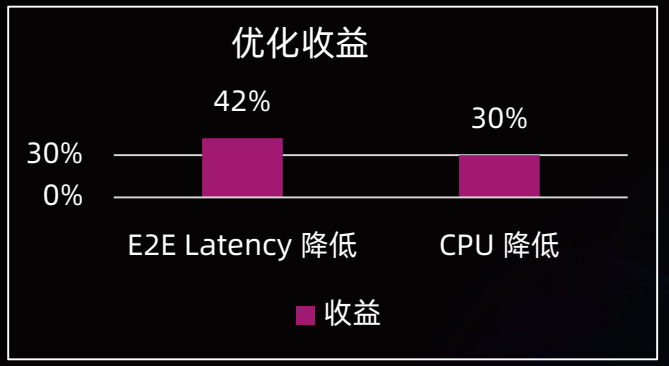



# Query Optimizer 优化

## 跨 Union All 下推

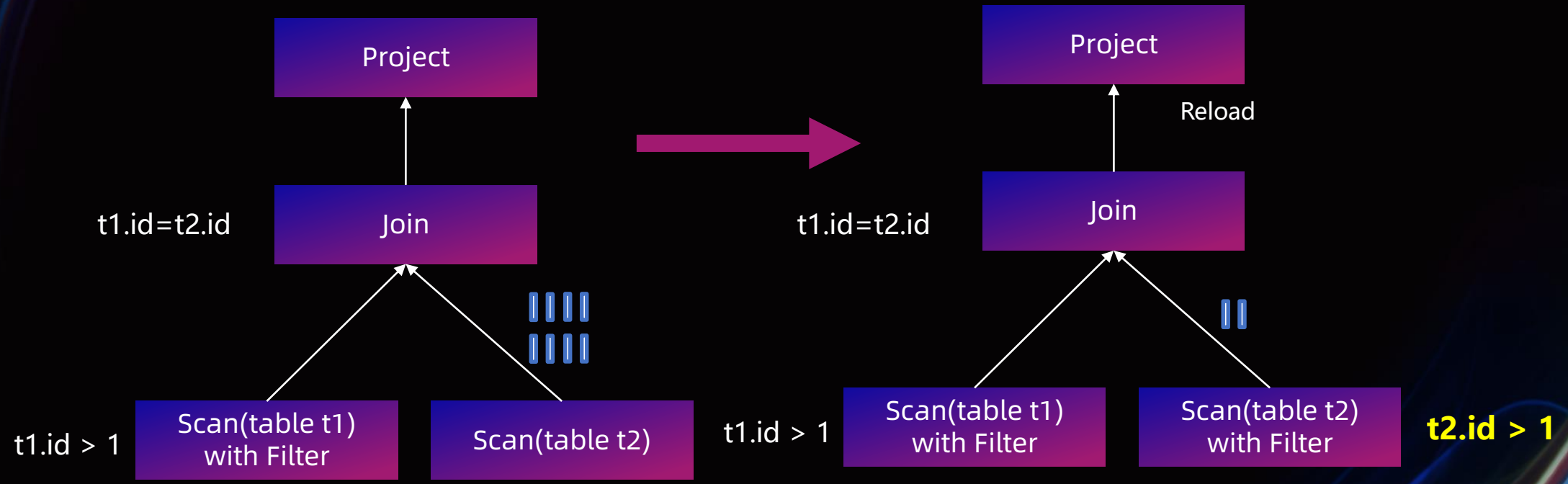


# Query Optimizer 优化

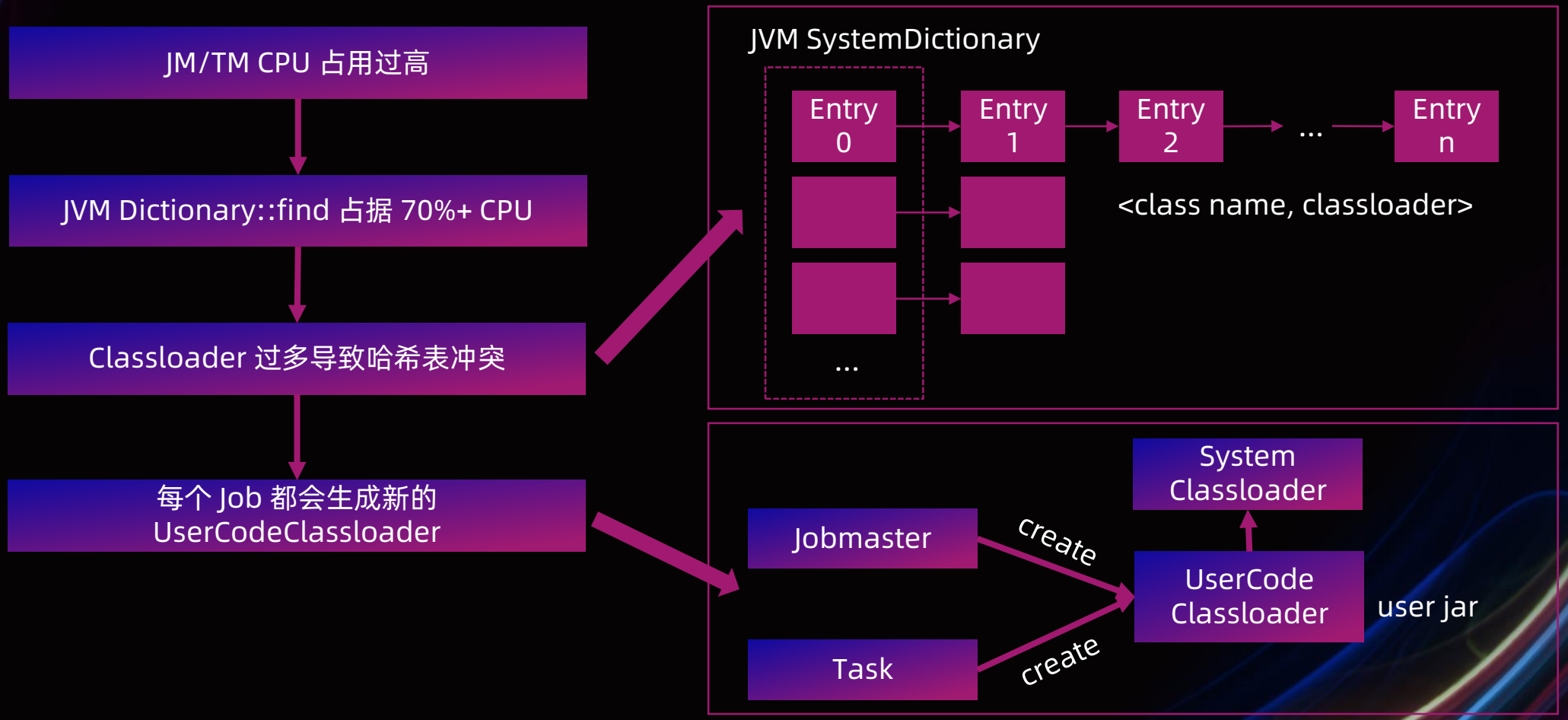
```
SELECT *
FROM t1 JOIN t2
ON t1.id = t2.id AND t1.id > 1
```

## Join Filter 传递



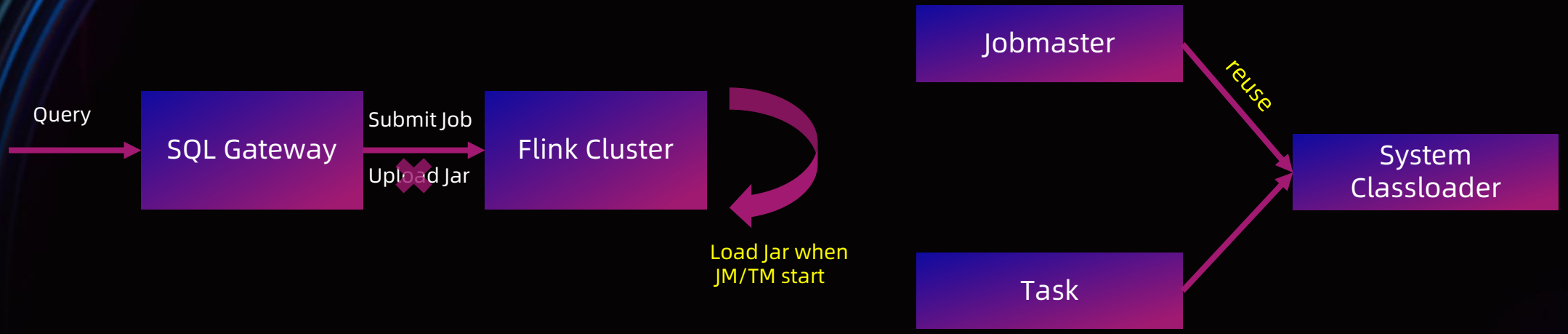
# Query Executor 优化

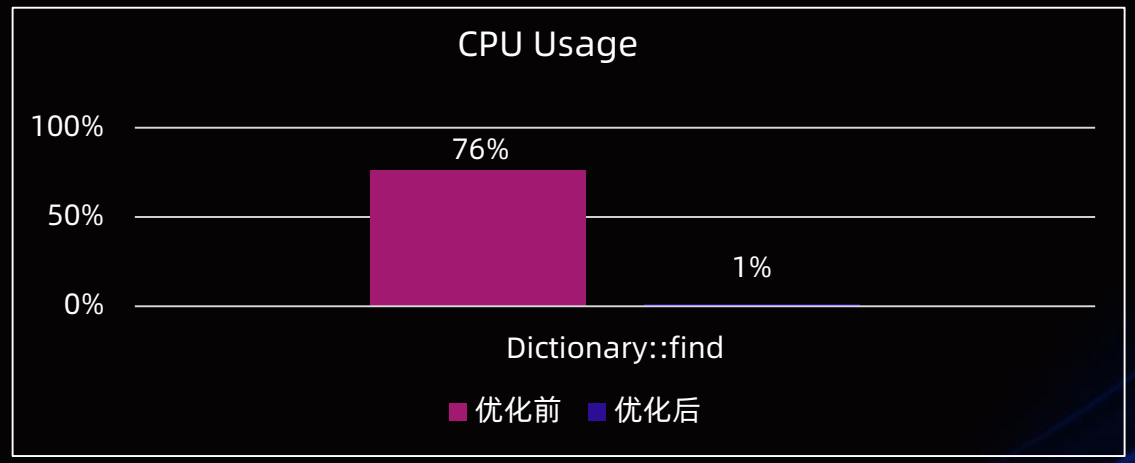
## Classloader 问题分析



# Query Executor 优化

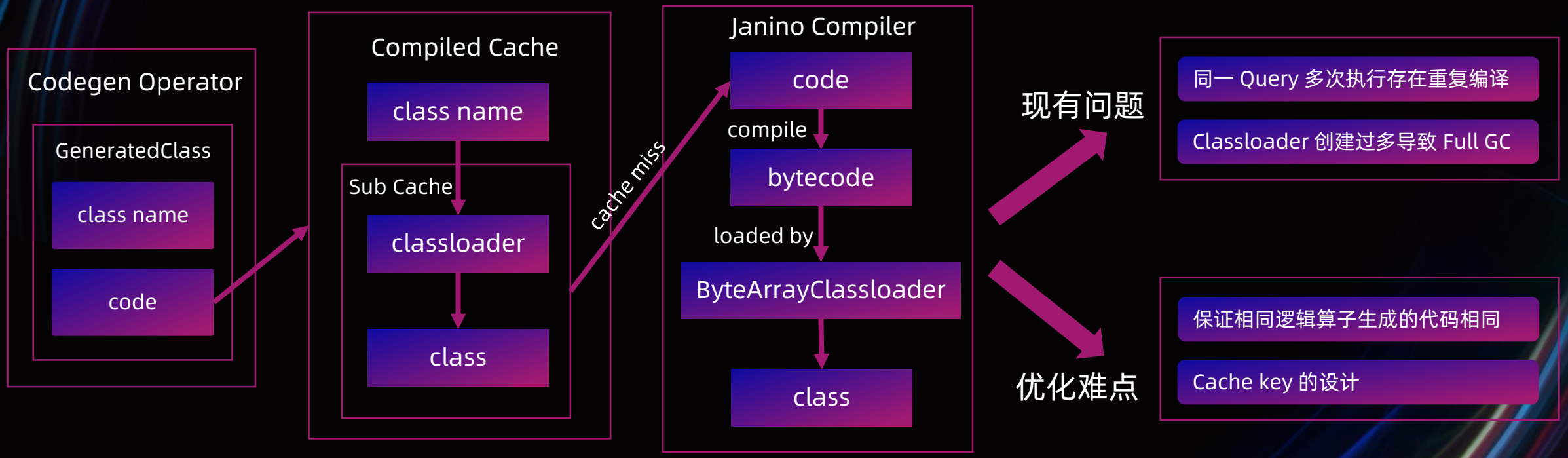
## Classloader 复用



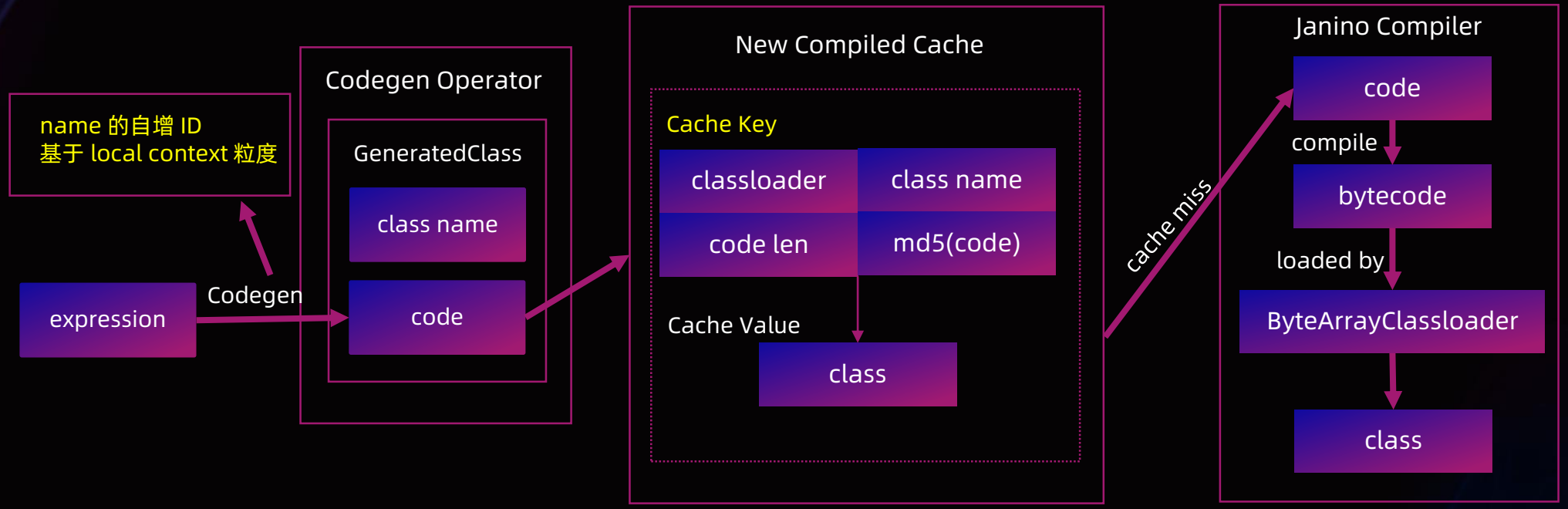
# Query Executor 优化

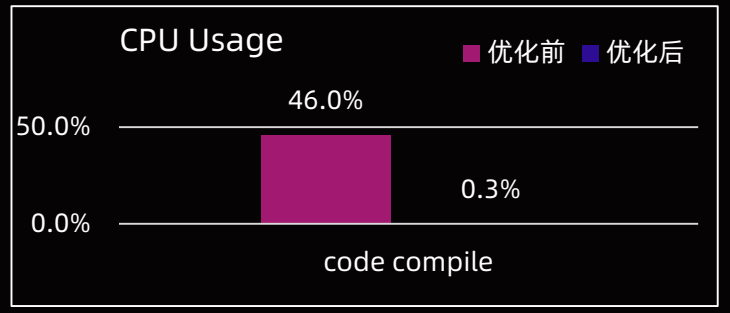
## Codegen 问题分析

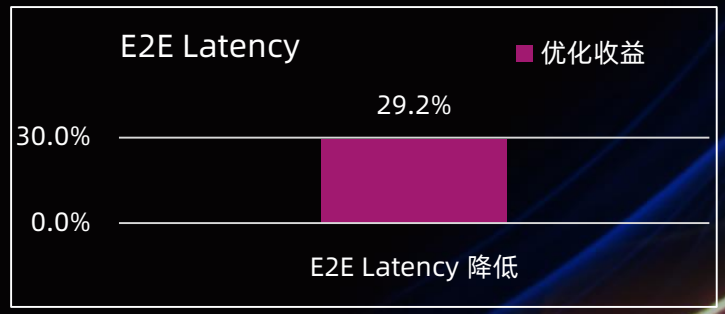


# Query Executor 优化

## Codegen 缓存优化



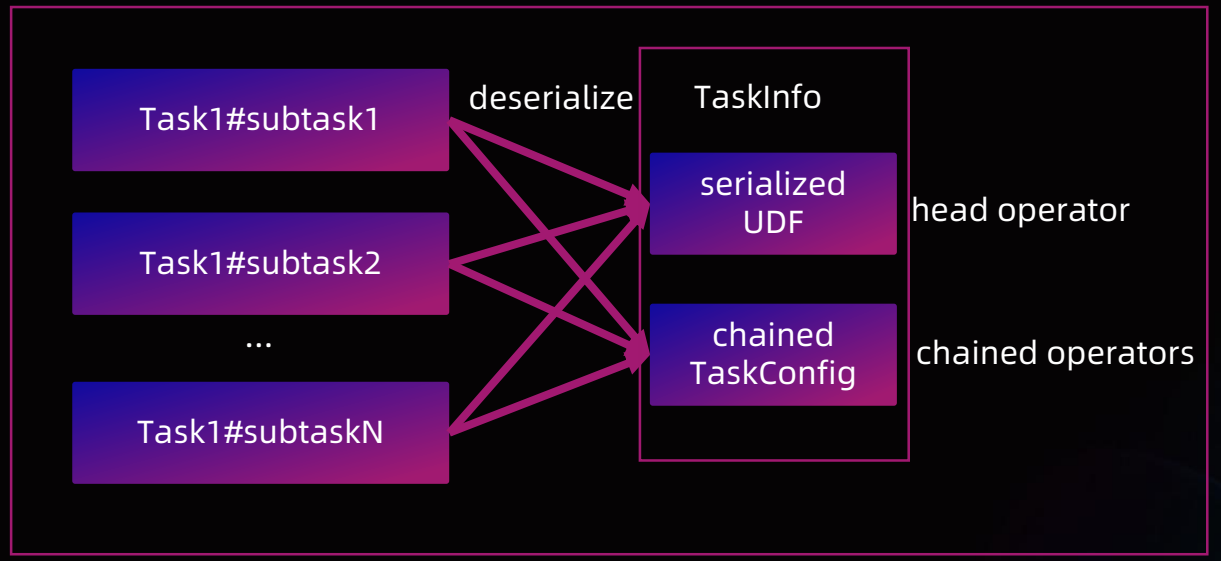
# Query Executor 优化

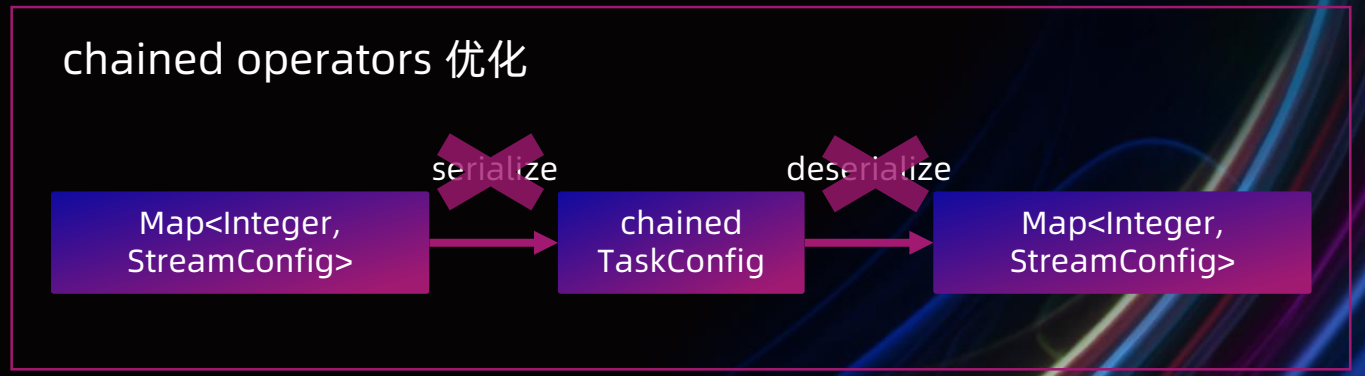
## 反序列化优化

TM Task 初始化阶段 CPU 占用较高

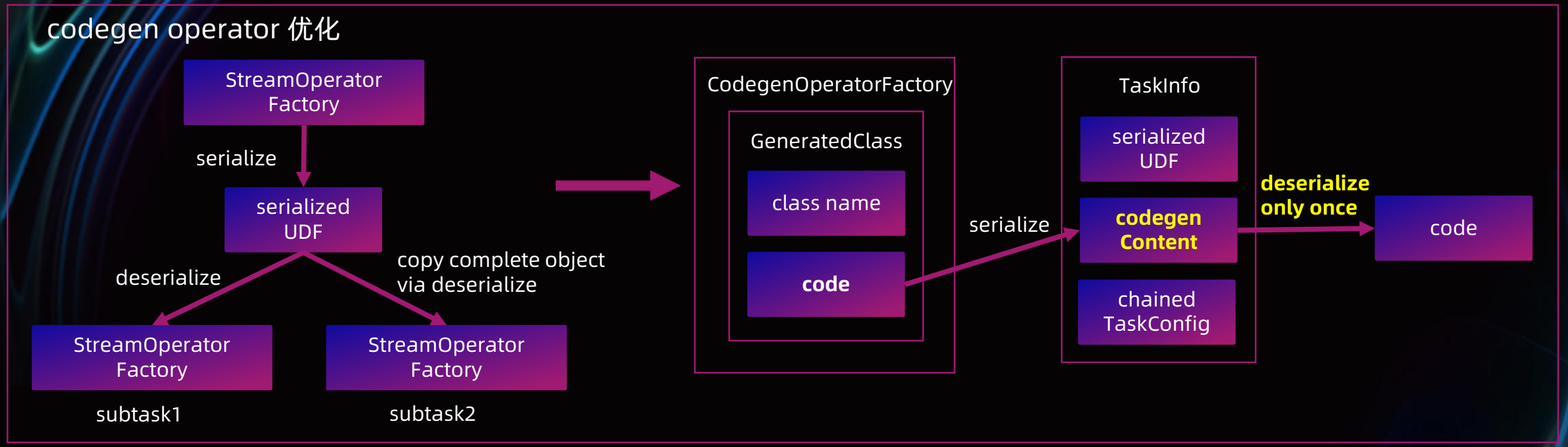
部署信息反序列化占比较高

多个 Subtask 存在冗余反序列化

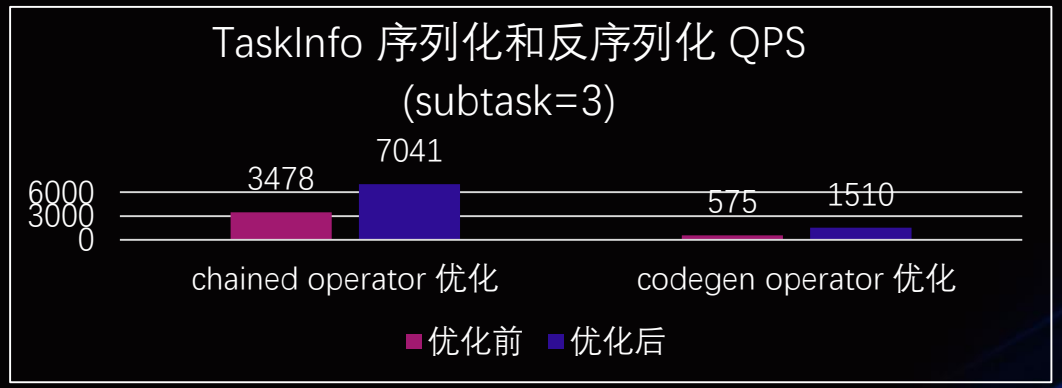


# Query Executor 优化

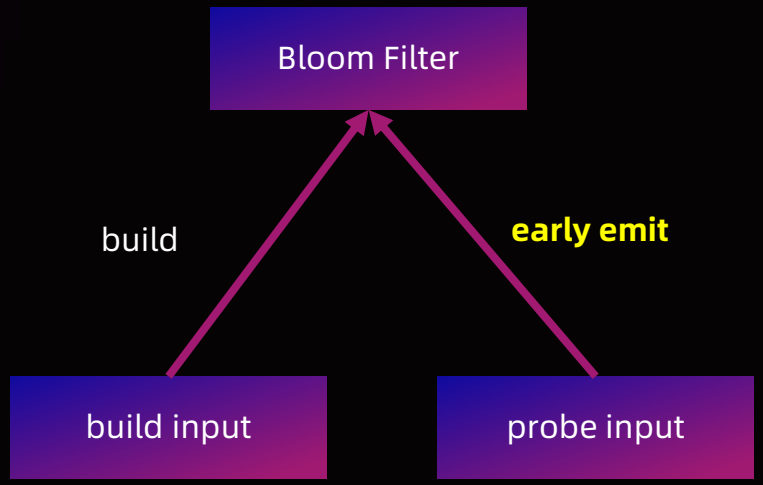



## 反序列化优化

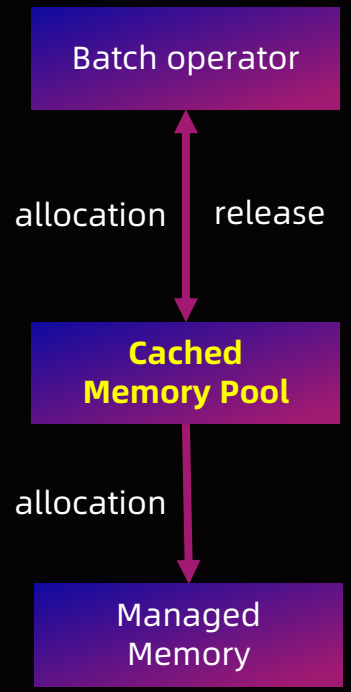


# Query Executor 优化

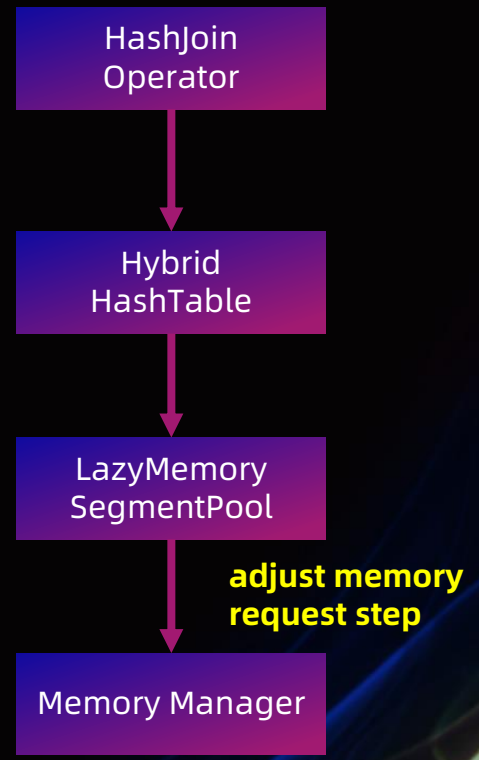
## Join Probe 提前输出




## 内存池化



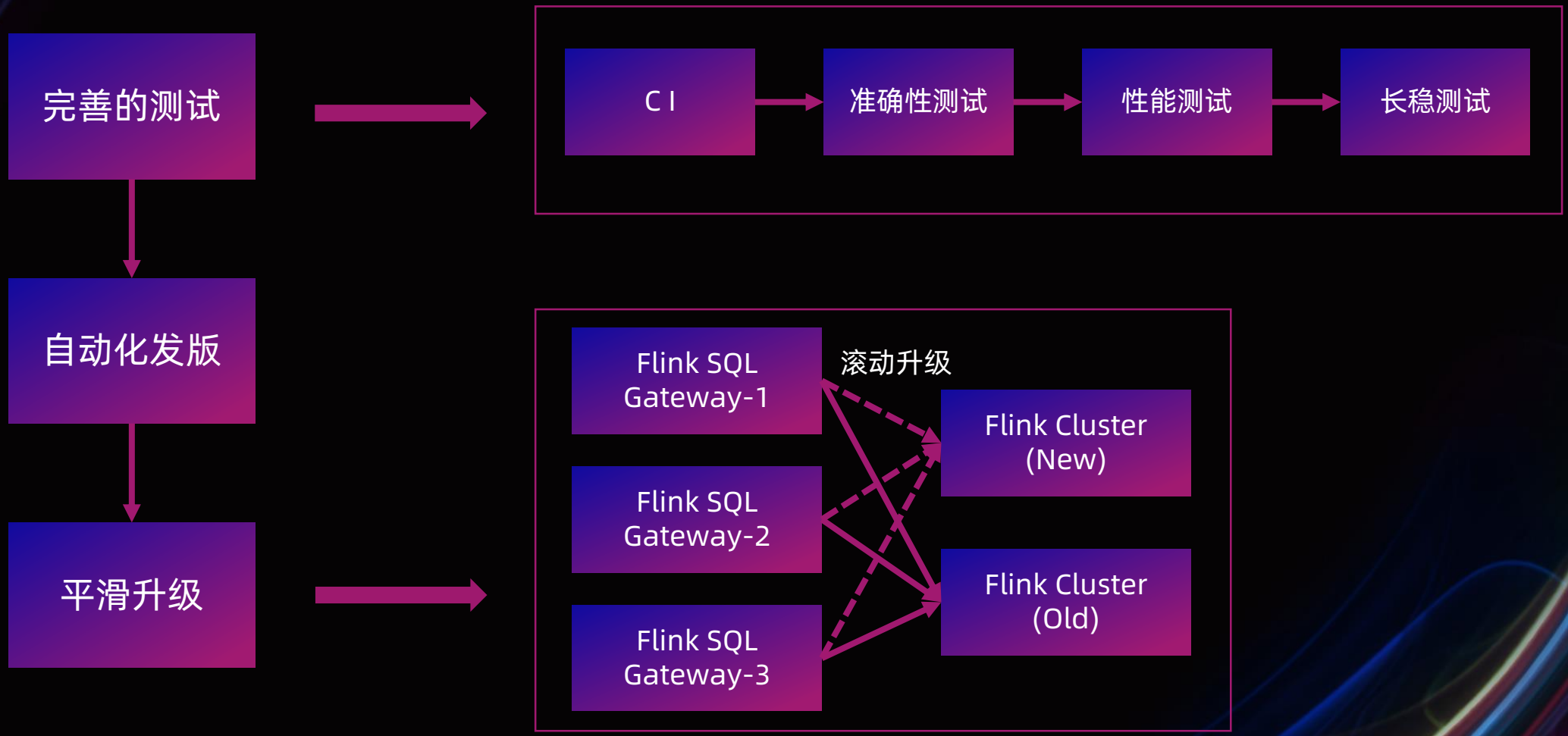

## 内存使用优化



# 03 集群运维和稳定性建设

1. 运维体系完善
2. 监控体系完善
3. 稳定性治理

# 运维体系完善


完善的测试
自动化发版
平滑升级
C I
准确性测试
性能测试
长稳测试
Flink SQL Gateway-1
Flink SQL Gateway-2
Flink SQL Gateway-3
滚动升级
Flink Cluster (New)
Flink Cluster (Old)

# 监控体系完善

**集群监控**

资源使用

- CPU
- 内存
- 网络
- 磁盘
- 细粒度 CPU

进程状态

- GC Time/Count
- Thread 数

退出码

- JM 退出码
- TM 退出码

查询负载

- 作业 QPS
- 同时运行作业数

**作业监控**

全链路 Latency

- Parse Latency
- Optimize Latency
- Submit Latency
- Schedule Latency
- Job Latency
- Result Push Latency
- E2E Latency

慢查询

- 慢查询 JobID
- 慢查询 QPS

失败查询

- 失败查询 QPS
- 失败查询 Latency

外部 IO

- HTAP MetaClient Latency
- HTAP Store Scan Latency

流 & 批　　OLAP

# 稳定性治理

## High Available

1. 双机房热备，支持故障切流
2. 支持 JobManager HA

## 限流 & 熔断

1. 支持 SQL Gateway QPS 限流
2. 限制 Flink 集群最大运行作业数
3. 作业 Failfast，避免集群雪崩

## GC 优化

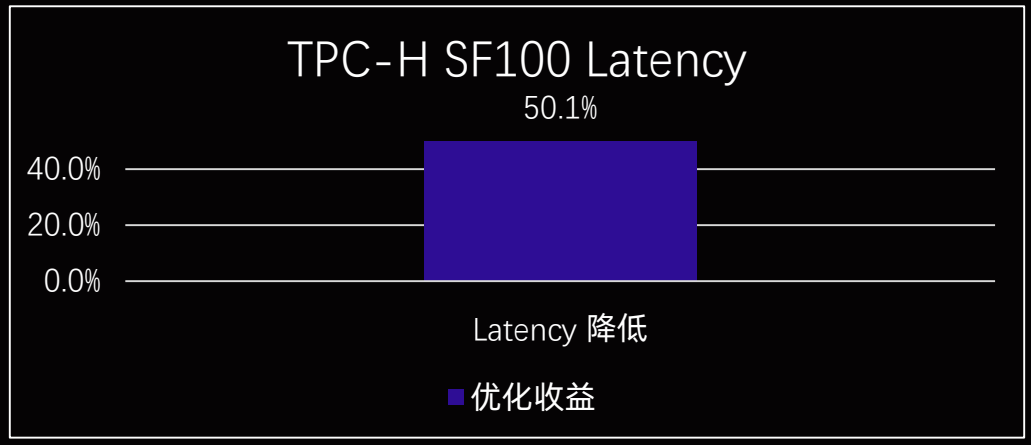
1. 移除 Task 级别的 metric，JM Full GC 频率降低 88%
2. Codegen 缓存优化，TM Metaspace Full GC 次数降低为接近 0

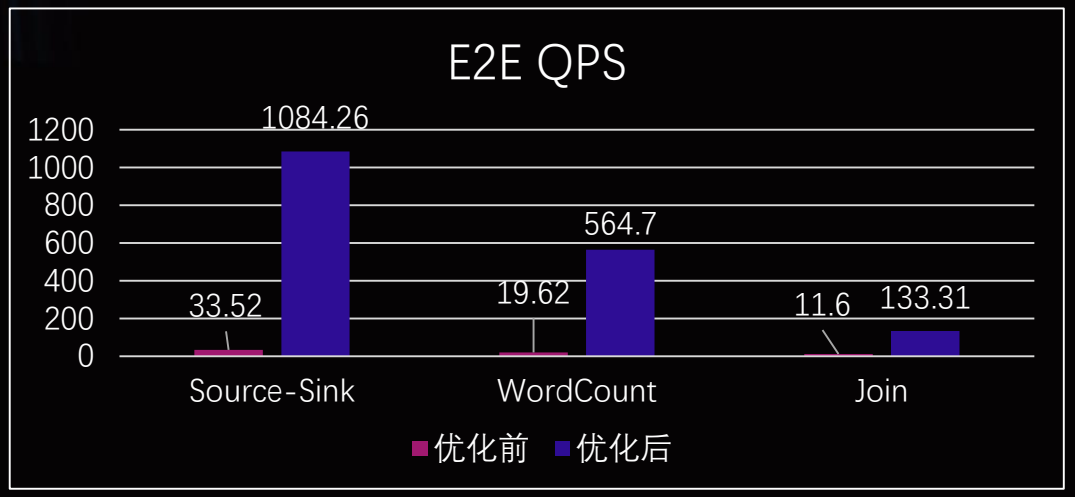
## JM 稳定性提升

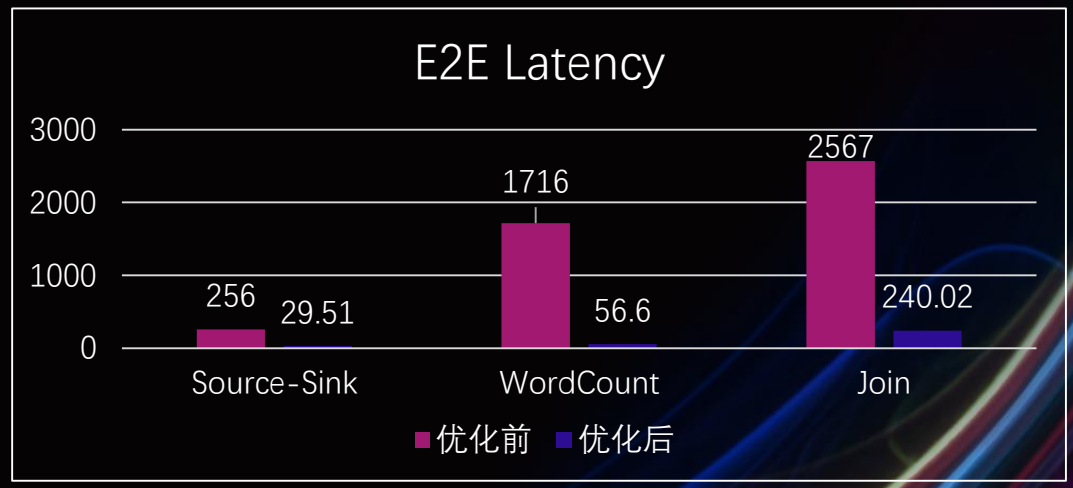
1. Jobmaster 去除 zk 依赖
2. 限制 Flink UI 展示的作业数
3. 关闭 Flink UI 自动刷新

# 04 收益

# Benchmark 收益



# 业务性能和稳定性收益

Job Latency 降低 48.3%

TM avg CPU 降低 27.3%

JM Full GC 频率降低 88.0%

TM Full GC 时间降低 71.5%

# 05 未来规划

# 未来规划

- 产品化完善
- 向量化引擎
- 物化视图
- Optimizer 演进

FLINK FORWARD # ASIA 2022 ONLINE | 实时即未来

# THANK YOU

谢 谢 观 看