Apache Doris 2.1 版本
资源负载隔离机制升级

王博 – SELECTDB 资深研发，Apache Doris PMC

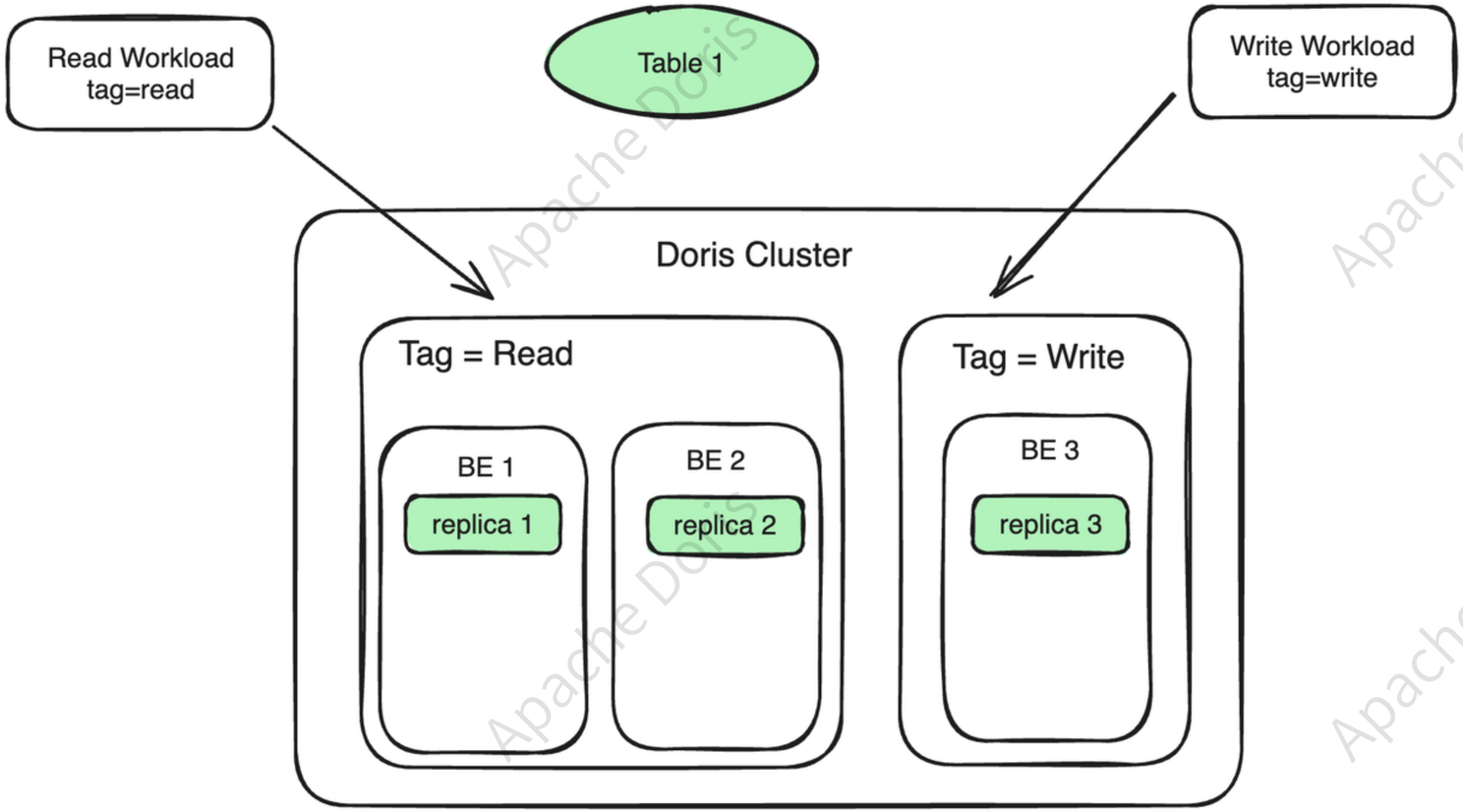
直播将于 19:30 准时开始，请耐心等待 ~

需求&问题收集

关注 SelectDB

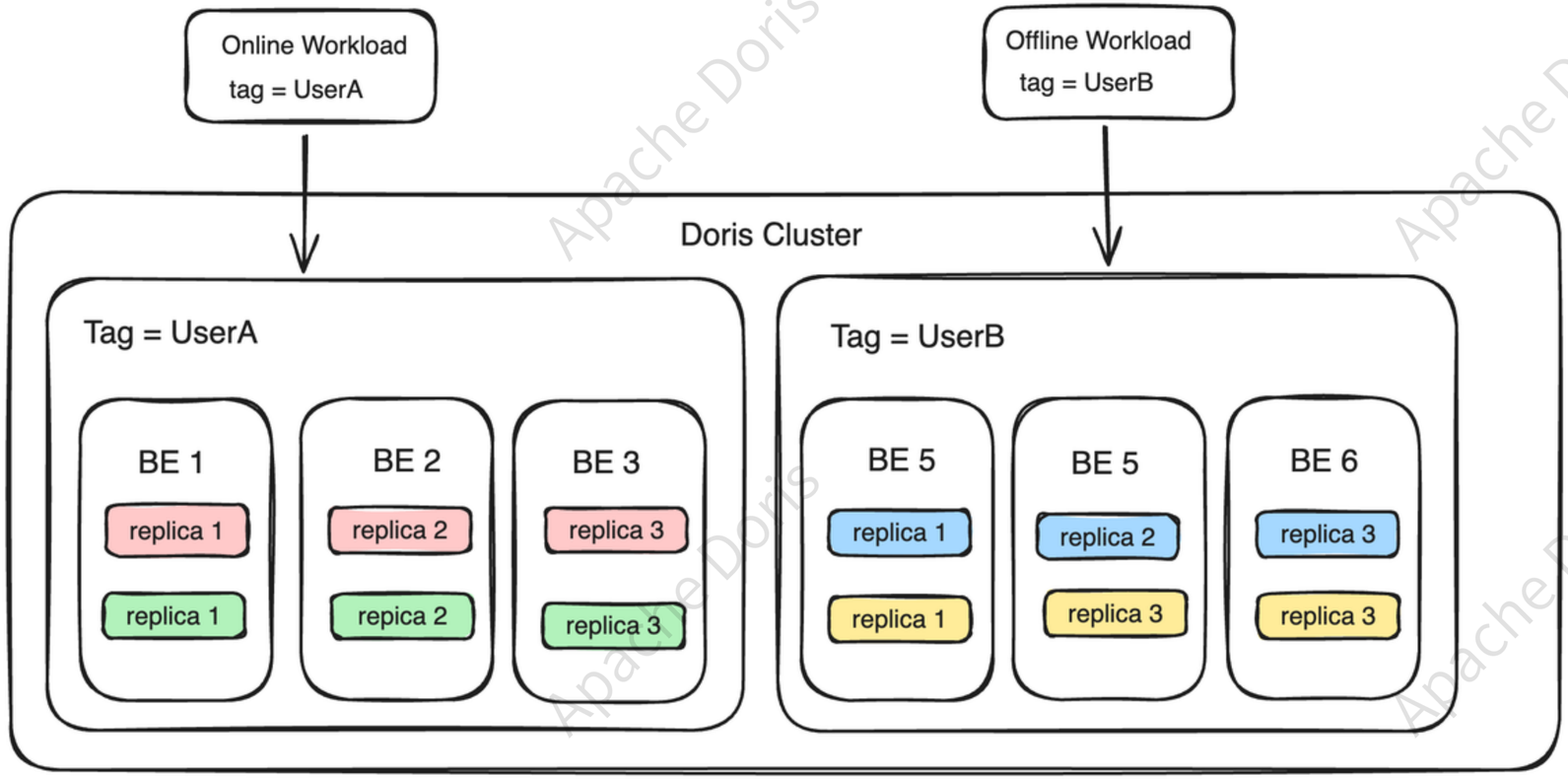
目录

Resource Tag 实现读写分离



- 把 BE 节点分成两组：tag=read 和 tag=write
- 把一张表的两个副本绑定到 read 组，把一个副本绑定到 write 分组
- 对工作负载进行分组，读负载路由到 read 组机器，写负载路由到 write 分组机器，从而进程级别的读写隔离

Resource Tag 实现多租户隔离



适用于将一个 Doris 集群进一步划分成多个小集群的场景

Resource Tag 功能总结

优势

- 基于进程级别的资源隔离，隔离性比较好

不足

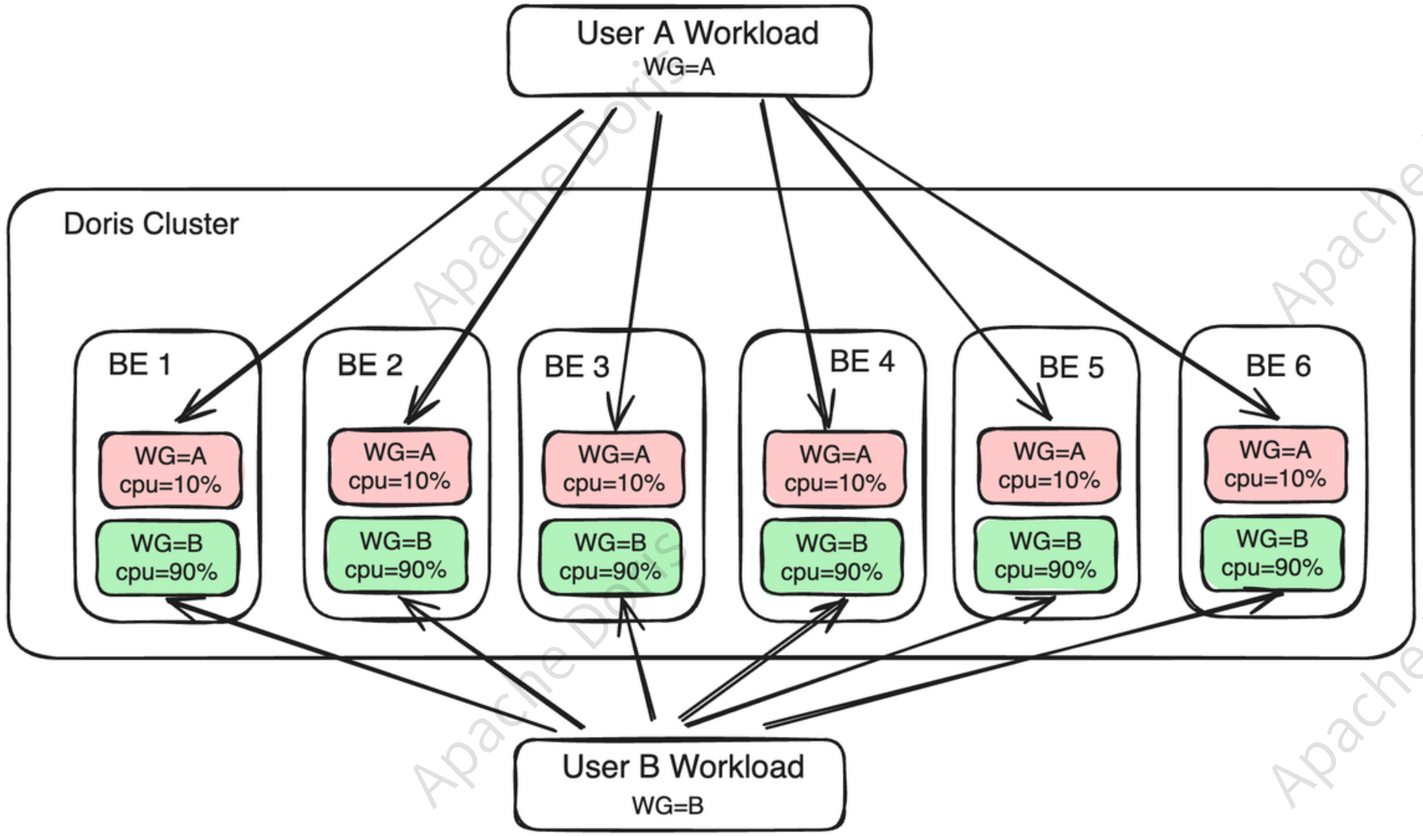
- 资源利用率较低，缺乏弹性
- 同一分组的多个业务方也会存在资源的竞争

版本说明

- 从 1.2 版本起开始支持

目录

基于 Workload Group 的资源隔离方案



- 主要关注进程内的 CPU 和内存的资源划分
- 多个 Workload Group 在同一个进程内部竞争资源
- Workload Group 只关注单机计算资源的划分，不关注存储资源

使用 Workload Group 管理 CPU 资源

CPU 软限

- 当 BE 内的 CPU 资源存在空闲时，单个 Workload Group 的最大可用 CPU 为整个 BE 的 CPU 资源
- 当 BE 内的各个 Workload Group 都达到最大负载时，Workload Group 的 CPU 分配根据事先配置的权重决定，权重越高，获得的 CPU 时间也就越长

CPU硬限

- 不管当前 BE 是否存在空闲的 CPU 资源，Workload Group 的 CPU 用量都不能超过配置的百分比

技术实现

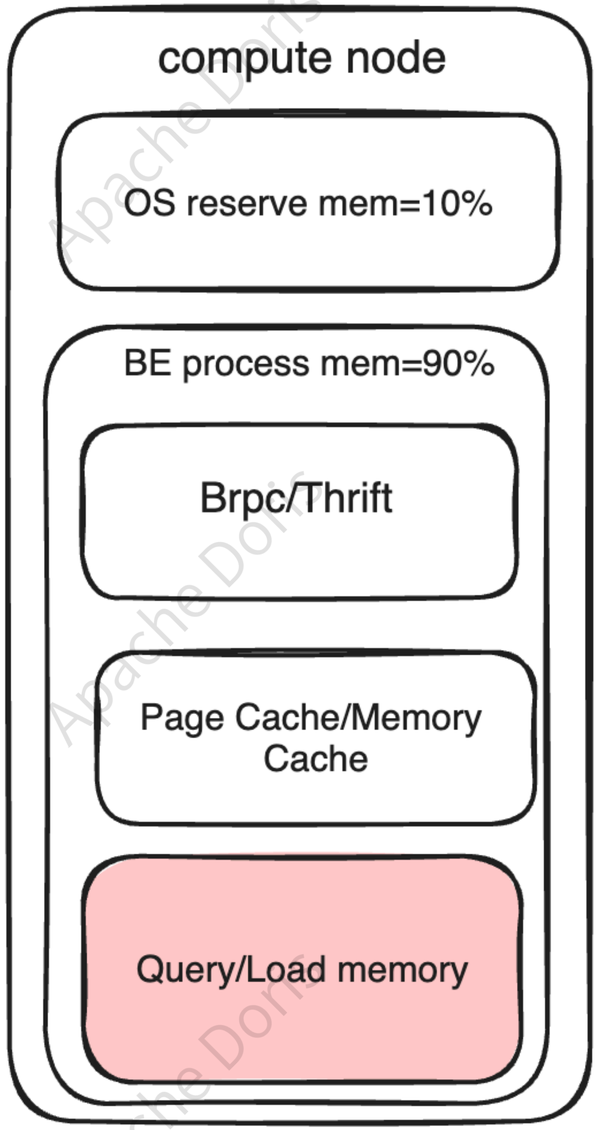
- 目前主要通过 cgroup 实现 CPU 资源的管理

CPU 软限和 CPU 硬限的对比分析

CPU 软限	CPU 硬限
资源利用率较高，支持进程内的弹性计算。	不支持弹性，无法利用 BE 上的空闲 CPU 资源。
适用于**用户偏好吞吐比较高**的场景。 性能波动可能会比较大。	适用于对于性能的稳定性有要求的场景： 用户的客户端负载不发生变化，Doris 侧的最大 CPU 资源一直不变，那么性能预期不会有太大波动。

- 两种CPU限制方式各适用于不同的需求场景。

使用 Workload Group 管理内存资源



在一个 BE 节点上，内存分为以下几部分：

- 操作系统保留内存
- BE 进程内存，而BE 进程内存又可以分为以下几个部分：
  - Doris 导入/查询使用的内存
  - Doris 公共组件内存，比如 page cache/内存块缓存
  - 第三方组件内存，比如 brpc/thrift 网络框架内存

- 目前 Workload Group 可以管理的内存主要是导入/查询使用的内存
- 支持内存的软限和硬限

使用 Workload Group 管理内存资源

当前的实现方式

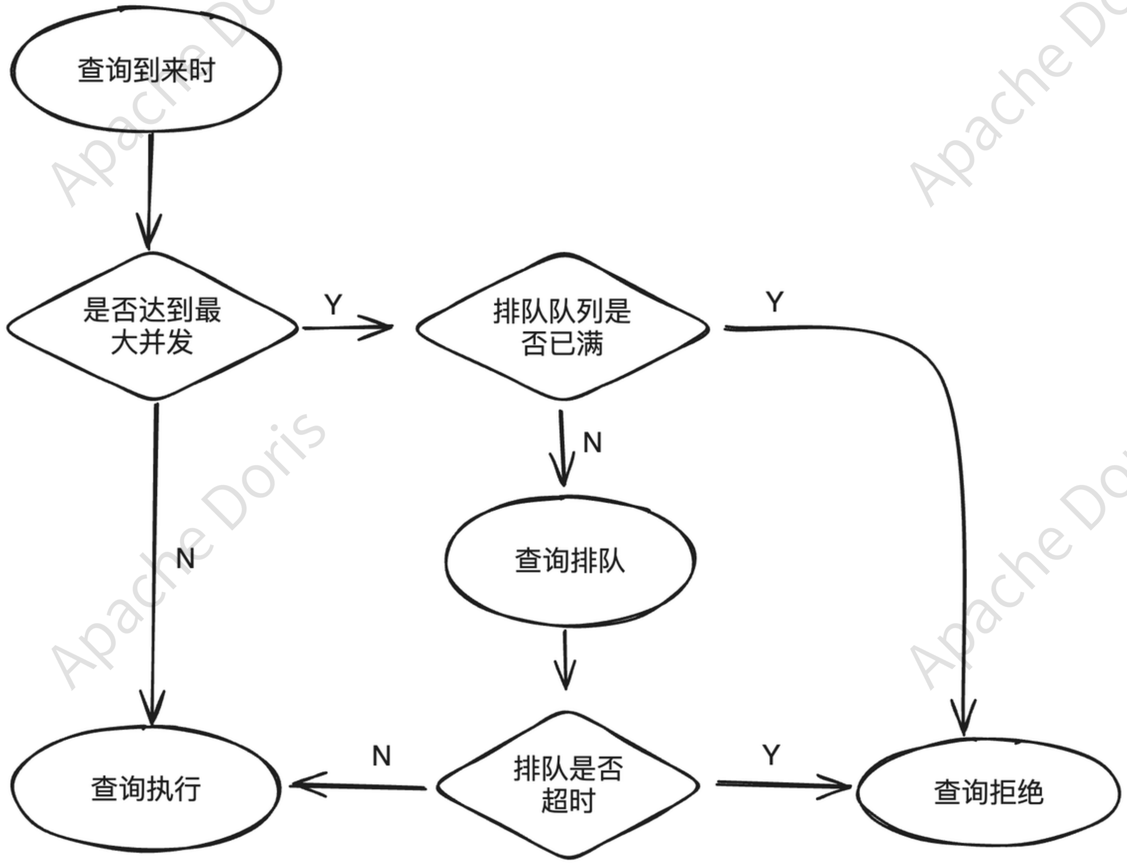
- BE 进程内有一个定时的内存 GC 线程计算导入/查询的内存用量，通过 cancel query 的方式释放内存
- 当 BE 进程的内存使用到达低水位时，会尝试释放 Workload Group 借用的内存
- 当 BE 进程的内存使用到达高水位时:
  - 会尝试释放 Workload Group 的借用的内存
  - 如果系统内存还是不足，那么会尝试从全局粒度释放查询/导入的内存

使用 Workload Group 管理内存资源

当前内存管理的问题与未来规划

- 目前释放内存的方式主要是通过 cancel 查询的方式，这对用户来说不够友好
- 未来会通过算子落盘的方式实现内存回收

Workload Group排队功能



现状

- 支持最大并发/排队/排队超时的功能

未来规划

- 支持基于内存的反压，当内存使用较高时，使得新来的查询排队

效果测试：CPU 软限

测试环境：16 核 64G 内存单机环境，1FE1BE，测试时需要关闭 Doris 的 page cache

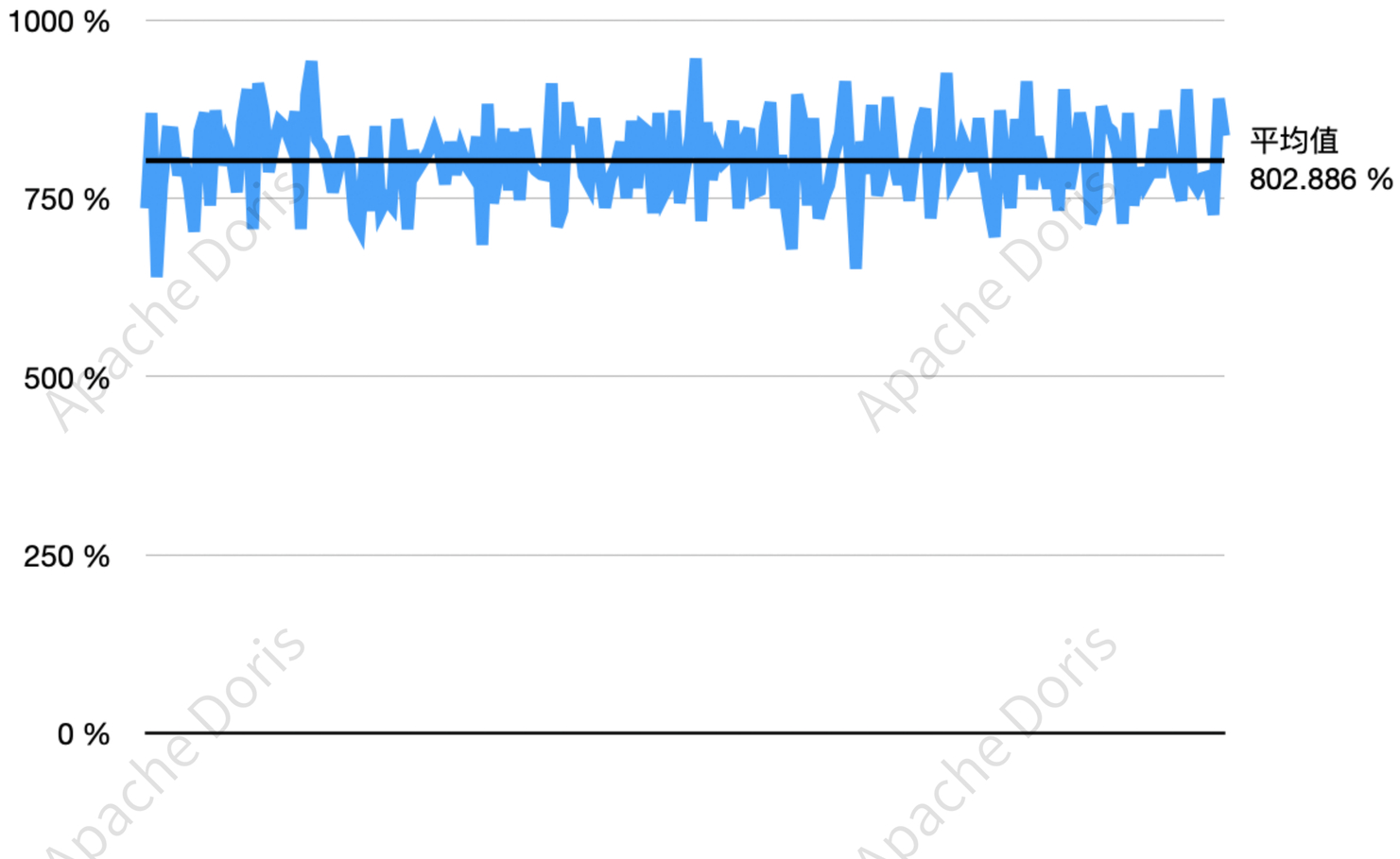
测试 1：启动两个客户端，持续提交 clickbench 的 q23，不使用 Workload Group 进行限制，运行 5 分钟

测试 2：启动两个客户端，持续提交 clickbench 的 q23，使用 Workload Group 进行限制，运行 5 分钟

测试编号	测试名称	客户端并发	Workload group	总执行sql个数	吞吐量	吞吐量比例（客户端1：客户端2）
1	客户端1	4	无	153	30.3/min	1:1
	客户端2	4	无	153	30.3/min	
2	客户端1	4	cpu_share=2048	188	37.4/min	2:1
	客户端2	4	cpu_share=1024	96	19.1/min	

效果测试：CPU 软限

在一个 16 核的机器上，配置 Workload Group 的 CPU 硬限为 50%；
使用 top 命令观察 CPU 用量应该一直是 800% 左右，也就是 8 个核。



目录

总结

- Resource Tag 适用于多租户/读写分离的场景，用户需要关注为每个分组分配几台机器。
- Workload Group 适用于 BE 进程内的资源隔离，用户需要关注进程内为每个分组分配多少 CPU 和内存。
- 对用户来说有一定的理解和使用成本。

未来规划

- 支持通过落盘的方式释放内存。
- 支持根据 BE 负载自动在 FE 进行排队。
- 统一 Workload Group 和 Resource Tag 功能。

目录

欢迎加入 Apache Doris

加入社区用户微信群

扫码添加 Doris 小助手，备注“加群”

提问赢好礼

提问被选中回复的小伙伴，请添加小助手微信领取 Doris 精美周边，数量有限，先到先得～

Doris 问答论坛

地址：ask.selectdb.com

Thanks !