YaoBase 多租户资源隔离设计

系统架构分析报告

一、系统概述

YaoBase 是一个分布式数据库系统，采用极致解耦设计。该项目旨在实现多租户资源隔离功能，在共享集群中为多个租户提供逻辑隔离的数据库服务。系统由四个核心服务器组件构成：AdminServer（集群管理）、SqlServer（SQL 处理）、TransServer（增量数据）和 DataServer（基线数据）。

二、架构特点

服务组件	职责	隔离策略
AdminServer	集群管理、副本负载均衡	共享
SqlServer	解析 SQL、生成执行计划	隔离
TransServer	L0 增量数据管理	共享
DataServer	L1 基线数据管理	隔离

三、资源隔离策略

资源类型	隔离层级	隔离难点	解决方案
CPU	SqlServer	已有项目改造	支持 cgroup 精确控制
内存	SqlServer	多租户共享 MemTable	仅在 SS 实现隔离
磁盘	DataServer	多租户共享基线数据	区分控制使用空间
网络	SqlServer	网络线程共享	按需配置

事务	TransServer	RAFT 架构独立	共享增量能力

四、核心设计原理

本项目采用最小侵入、渐进式的设计原则。CPU 和内存隔离在 SqlServer 实现，磁盘隔离在 DataServer 实现，而 TransServer 作为共享的增量服务不加入隔离对象。

4.1 CPU 资源隔离设计

- **隔离模式**：支持基础模式（线程池比例分配）和 cgroup 模式（精确限制）
- **线程池管理**：全局 120 个工作线程，按租户配额动态分配
- **租户线程组**：每个租户拥有独立的线程组和无锁任务队列
- **cgroup 集成**：可选的 Linux cgroup 支持，实现精确的 CPU 限制
- **监控机制**：实时监控 CPU 使用率、限制命中率、告警告知

4.2 内存资源隔离设计

- **隔离范围**：在 SqlServer 层面实现内存配额控制
- **共享设计**：TransServer 的 MemTable 为所有租户共享，自动合并释放内存
- **配额检查**：请求处理前检查租户内存配额，防止超限
- **监控统计**：提供租户内存使用量查询接口

4.3 磁盘资源隔离设计

- **隔离层级**：在 DataServer 层面控制租户磁盘使用总量
- **控制策略**：超限时禁止写入操作
- **检测延迟**：在数据合并后检测，不会立即报错
- **资源管理**：定期清理过期数据，提升空间利用率

五、技术实现栈

模块	关键类	功能
租户管理	TenantContext, TenantManager	租户生命周期管理
资源统计	ResourceStats, BasicResourceStats	统一的资源统计
CPU 隔离	CpuResourceManager, CpuQuotaChecker	CPU 管理和监控

内存隔离	MemoryResourceManager, MemoryQuotaChecker	内存配额控制
磁盘隔离	DiskResourceManager, DiskQuotaChecker	磁盘配额管理
线程管理	ThreadPoolManager, TenantThreadGroup	线程池和任务队列
认证授权	TenantAuthenticator	租户身份认证
cgroup 集成	CgroupController	Linux cgroup 支持

六、请求处理流程

1. **连接建立**：客户端建立连接时携带租户标识（user@tenant 格式）
2. **租户认证**：TenantAuthenticator 验证租户身份和连接权限
3. **CPU 配额检查**：CpuQuotaChecker 检查租户 CPU 资源是否充足
4. **内存配额检查**：MemoryQuotaChecker 验证内存限制
5. **线程分配**：ThreadPoolManager 为请求分配租户线程组中的工作线程
6. **SQL 执行**：在分配的线程上执行 SQL 处理
7. **监控统计**：CpuMonitor 实时记录 CPU/内存使用情况
8. **资源释放**：处理完成后释放分配的资源
9. **结果返回**：将结果返回给客户端

七、关键特性

特性	说明
无锁数据结构	使用 LockFreeQueue 实现任务队列，提升并发性能
动态配置	支持运行时调整租户配额，无需重启
性能开销低	设计目标：资源隔离性能开销 < 5%
可扩展性	支持 1000+租户并发管理

监控告警	提供租户资源使用视图和告警机制
容错能力	cgroup 失效时自动降级到基础模式
GoogleTest 集成	90 个单元测试 + 集成测试完整覆盖

八、设计约束与限制

- **磁盘控制延迟性**：磁盘超限在数据合并后才检测
- **内存隔离不作用于 TransServer**：TS 为共享资源，自动合并释放内存
- **数据隔离程度有限**：多租户共享 RAFT 架构、LSM-tree、raft 日志
- **事务处理能力受限**：受共享的 TS RAFT 组结构限制
- **使用建议**：对于严格数据隔离需求，建议使用独立集群

九、总结

本项目采用最小侵入的设计原则，通过在 SqlServer 和 DataServer 分别实现 CPU、内存、磁盘资源隔离，实现了多租户在共享集群中的逻辑隔离。系统设计具有高度的可配置性、可监控性和容错能力。整个实现过程遵循了工程最佳实践，包括接口抽象、依赖注入、无锁数据结构、全面的单元测试和集成测试。这个设计为 YaoBase 提供了灵活的多租户支持能力，既提升了资源利用率，又保障了租户间的公平性。