PR #38435 完整报告

vllm-project/vllm

[Core][Metrics] expose waiting request breakdown via labeled metric (capacity/deferred)

合并时间：2026-04-14 03:30

原文链接：http://prhub.com.cn/vllm-project/vllm/pull/38435

执行摘要

本 PR 新增了 vllm:num_requests_waiting_by_reason 标签化指标，将等待请求队列细分为“容量约束”和“延迟约束”，帮助操作员精确诊断调度瓶颈，同时保持向后兼容性，体现了对 Prometheus 最佳实践的遵循。

功能与动机

现有 vllm:num_requests_waiting 指标合并了两种不同请求队列：一是因调度容量不足而等待的请求（waiting 队列），二是因 LoRA 预算、异步 KV 传输等约束被延迟的请求（skipped_waiting 队列）。这导致在生产负载下无法区分负载驱动积压与约束驱动阻塞，如 PR body 所述：“makes it impossible to distinguish load-driven backlog from constraint-driven blocking”。本 PR 旨在通过标签化指标提供细粒度洞察，提升监控和调试能力。

实现拆解

- vllm/v1/metrics/stats.py：在 SchedulerStats 类中添加 num_waiting_reqs 和 num_skipped_waiting_reqs 字段，分别记录两个队列的长度。
- vllm/v1/core/sched/scheduler.py：更新 make_stats() 函数，设置 num_waiting_reqs=len(self.waiting) 和 num_skipped_waiting_reqs=len(self.skipped_waiting)。
- vllm/v1/metrics/loggers.py：定义常量 WAITING_REASON_CAPACITY 和 WAITING_REASON_DEFERRED，创建标签化指标 vllm:num_requests_waiting_by_reason，并更新日志输出，当延迟请求数大于 0 时显示“Deferred: N reqs”。
- 测试文件：在 tests/v1/core/test_scheduler.py 中添加 test_scheduler_stats_waiting_queues() 验证统计正确性；在 tests/entrypoints/serve/instrumentator/test_metrics.py 中将新指标加入期望列表。

评论区精华

- 标签化设计采纳：markmc 建议“使用标签化指标模式”，类似 vllm:prompt_tokens_by_source，以遵循 Prometheus 最佳实践并支持未来扩展。该建议被采纳，初始实现从单独指标重构为标签化指标。

- 术语统一决策：讨论中涉及“skipped” vs “deferred”等术语，最终选择“deferred”作为用户友好标签，明确表示约束延迟。
- 代码一致性优化：*gemini-code-assist*[bot] 指出初始化应使用 *create_metric_per_engine* 工具，后续提交中已修复，确保代码风格统一。

风险与影响

- 风险：向后兼容性风险极低，原始 vllm:num_requests_waiting 指标通过求和两个队列保持语义不变；性能影响可忽略，仅增加少量指标记录开销；安全无关；兼容性无破坏。
- 影响：用户（操作员）能更精细监控调度瓶颈，快速识别容量 vs. 约束问题；系统支持未来指标细分（如按 LoRA、KV 传输类型）；团队维护简单，测试覆盖充分，无重大学习负担。

关联脉络

本 PR 是 PR #35781 的后续扩展，后者引入了 skipped_waiting 队列。通过标签化指标，进一步提升了监控能力，体现了 vLLM 在指标系统上向 Prometheus 最佳实践的演进。结合近期历史 PR 分析，该变更属于核心调度和监控的持续改进，与类似指标增强 PR（如 #39572 的性能数据导出）形成协同，共同增强系统可观测性。