PR #21463 完整报告

sgl-project/sglang

Migrate all callers from /get_server_info to /server_info

合并时间：2026-04-02 12:17

原文链接：http://prhub.com.cn/sgl-project/sglang/pull/21463

执行摘要

- 一句话：将所有调用者从弃用的 /get_server_info 迁移到新的 /server_info 端点，清理 API 表面。
- 推荐动作：建议开发团队快速浏览此 PR，关注 API 清理的最佳实践和弃用管理。对于新贡献者，理解如何管理端点弃用是有价值的学习案例。变更直白，无需深度技术分析，但值得参考以确保未来类似重构的顺利进行。

功能与动机

根据 Issue #21054，/get_server_info 是弃用的别名，服务器已记录弃用警告，但约 49 个文件仍引用旧端点。PR 旨在迁移所有内部调用者，减少 API 表面混淆并确保新代码使用正确路径。

实现拆解

变更按模块拆解：1) 文档更新（如 server_arguments.md、sgl_model_gateway.md），修正端点引用；2) 客户端调用迁移，包括 Python SDK 的 runtime_endpoint.py 中 get_server_info 方法、基准测试脚本（bench_serving.py）和 profiler；3) 测试文件全面更新，覆盖 speculative decoding、MLA、EP、DP、HiCache、quantization、AMD、Ascend 等模块；4) sgl-model-gateway 中添加 /server_info 路由并保留 /get_server_info 别名，设置 TODO 注释用于一个发布周期的弃用窗口。

关键文件：

- python/sglang/lang/backend/runtime_endpoint.py（模块 Python SDK）：更新 Python SDK 的 get_server_info 方法，直接影响所有客户端调用
- sgl-model-gateway/src/server.rs（模块 sgl-model-gateway）：添加 /server_info 路由并设置弃用别名，关键网关变更
- test/registered/core/test_srt_endpoint.py（模块 Testing）：更新核心测试中的端点调用，确保测试覆盖和正确性
- docs/advanced_features/server_arguments.md（模块 Documentation）：更新文档中的端点引用，影响用户指南和 API 说明

关键符号：get_server_info (in runtime_endpoint.py), get_server_info (in server.rs route)

评论区精华

review 评论由 gemini-code-assist[bot] 提供，主要关注 TODO 注释，强调弃用窗口的重要性。例如，在 mini_lb.py 中添加的 TODO 注释被标记为中等优先级，提醒未来移除别名；在 discover_metadata.rs 中讨论了函数重命名和回退机制的未来清理。没有实质性争议，讨论集中于确保平滑过渡和代码管理。

- 弃用窗口管理 (documentation): 已添加 TODO 注释，计划在一个发布周期后移除 /get_server_info 别名

风险与影响

- 风险：风险较低，主要包括：1) 遗漏调用者可能导致测试失败或功能异常，尤其在广泛的文件修改中（48 个文件）；2) 弃用窗口设置不当可能过早移除别名，破坏向后兼容性；3) 文档或代码中残留旧引用，引发混淆。具体风险点在于测试覆盖率是否全面，需验证所有变更文件。
- 影响：影响范围广但程度低：1) 对用户（开发者），API 更加一致，无直接功能变化，但需注意弃用警告；2) 对系统，服务器端已有弃用包装器，变更透明，性能无影响；3) 对团队，减少维护负担，简化 API 表面，但需确保所有 CI 测试通过。
- 风险标记：遗漏调用者，弃用窗口管理

关联脉络

- 暂无明显关联 PR