# PR #21700 完整报告

sgl-project/sglang

Support HTTP2 server

合并时间：2026-04-08 00:42

原文链接：http://prhub.com.cn/sgl-project/sglang/pull/21700

## 执行摘要

本 PR 引入了 HTTP/2 服务器支持，通过集成 Granian ASGI 服务器作为 Uvicorn 的可选替代，旨在减少连接开销并提升高并发客户端的吞吐量。实现包括添加命令行标志、依赖管理、服务器逻辑调整和新增测试，是一个有意义的服务器层改进。

## 功能与动机

PR 的动机是解决 HTTP/1.1 在高并发流式工作负载下的连接开销问题。如 PR body 所述："HTTP/2 multiplexing allows many in-flight requests over a single TCP connection, reducing connection overhead and improving throughput for high-concurrency clients. This is particularly useful for streaming workloads where multiple SSE streams can share one connection instead of each occupying a separate HTTP/1.1 connection." 基准测试显示，使用 Granian HTTP/2 后请求速率从 126.34 req/s 提升至 2645.94 req/s。

## 实现拆解

实现主要涉及以下模块：

1. 依赖管理(python/pyproject.toml)：添加 granian>=2.6.0 作为可选依赖项 http2。
2. 服务器逻辑(python/sglang/srt/entrypoints/http_server.py)：
   - 新增 _init_granian_worker 函数初始化 Granian 工作进程。
   - 新增 _close_main_process_sockets 函数关闭主进程的 ZMQ 套接字以避免冲突。
   - 新增 _run_granian_server 函数启动 Granian 服务器，支持 HTTP/1.1 和 HTTP/2 自动协商。
3. 命令行参数(python/sglang/srt/server_args.py)：
   - 添加 --enable-http2 标志，并集成验证逻辑，限制与 --enable-ssl-refresh 和 --tokenizer-worker-num > 1 的兼容性。
4. 环境变量(python/sglang/srt/environ.py)：添加 SGLANG_GRANIAN_PARENT_PID 以支持进程 ID 覆盖。
5. 进程协调(python/sglang/srt/managers/multi_tokenizer_mixin.py)：调整 get_main_process_id 函数，支持环境变量覆盖。
6. 测试(test/registered/openai_server/basic/test_http2_server.py)：新增测试文件，验证服务器启动、健康检查、完成请求和 HTTP/2 协议支持。

## 评论区精华

review 讨论较为简单，仅有一个来自 gemini-code-assist[bot] 的风格建议：

> "For better code clarity and maintainability, please add a type hint for the server_args parameter. Based on its usage, it should be ServerArgs."

该建议被采纳，代码在后续提交中更新，没有其他争议。

## 风险与影响

风险分析：

- 依赖风险：引入第三方包 granian 可能带来版本兼容性或安装失败问题。
- 功能限制：当前不支持多 tokenizer 工作线程 (--tokenizer-worker-num > 1) 和 SSL 证书热重载 (--enable-ssl-refresh)，限制了使用场景。
- 并发风险：_close_main_process_sockets 函数在关闭套接字时可能因时机不当导致资源泄漏或进程间通信冲突。
- 测试覆盖：新增测试验证了基本功能，但高并发、边缘情况或长期运行下的稳定性可能未充分覆盖。

影响分析：

- 对用户：提供可选 HTTP/2 支持，需要安装额外依赖 sglang[http2]，但对现有 API 无破坏性变更。
- 对系统：显著提升服务器吞吐量，减少 TCP 连接数，优化流式工作负载性能。
- 对团队：新增维护点，需关注 Granian 集成的稳定性，并可能影响未来服务器架构演进。

## 关联脉络

从近期历史 PR 看，本 PR 与基础设施和依赖管理相关变更有相似之处。例如：

- PR 22162 添加了 mlx 和 mlx-lm 依赖，同样涉及 pyproject.toml 的修改。
- PR 22267 调整测试套件，与本 PR 的新增测试文件类似，都关注测试基础设施。

这些关联表明仓库在持续优化服务器层和测试框架，以支持更多硬件后端和协议特性。