PR #25041 完整报告

sgl-project/sglang

Optimize uvicorn startup command

合并时间：2026-05-12 15:28

原文链接：http://prhub.com.cn/sgl-project/sglang/pull/25041

执行摘要

- 一句话：优化 Uvicorn 启动命令，修复 worker 健康检查超时问题
- 推荐动作：本 PR 是一次干净、标准的修复，值得合并。其设计思路（移除无效的 monkey patch，改用原生参数传递）可视为同类问题的正确处理模式。

功能与动机

根据 Issue #25040 的描述，在多 worker 并行启动时，tokenizer 初始化期间 CPU 和 GIL 负载高，导致 worker 的健康检查线程无法及时响应，Uvicorn 父进程误认为 worker 已死并将其杀死，造成 worker 周期性重启。原有的 monkey patch 方案因 supervisor 显式传参 timeout= 而失效。

实现拆解

1. 移除无效的 monkey patch：从 python/sglang/srt/managers/multi_tokenizer_mixin.py 中删除函数 monkey_patch_uvicorn_multiprocessing 及其调用，该函数通过 partialmethod 修改 Process.is_alive 默认超时，但实际上 Uvicorn multiprocess supervisor 会显式传递 timeout 参数，覆盖了默认值，因此该 patch 从未生效。
2. 新增环境变量控制超时：在 python/sglang/srt/environ.py 中添加 SGLANG_UVICORN_WORKER_HEALTHCHECK_TIMEOUT = EnvInt(10)，默认值设为 10 秒，比 Uvicorn 原默认值 5 秒更宽松。
3. 将超时参数直接传递给 Uvicorn：在 python/sglang/srt/entrypoints/http_server.py 的多 worker 分支中，调用 uvicorn.run() 时新增参数 timeout_worker_healthcheck=envs.SGLANG_UVICORN_WORKER_HEALTHCHECK_TIMEOUT.get()，该参数是 Uvicorn 原生支持的，能正确控制 supervisor 的健康检查间隔。

关键文件：

- python/sglang/srt/managers/multi_tokenizer_mixin.py（模块 Token 管理器；类别 source；类型 dependency-wiring；符号 monkey_patch_uvicorn_multiprocessing）：移除了无效的 monkey patch 函数及相关导入，是主要变更之一。
- python/sglang/srt/entrypoints/http_server.py（模块 HTTP 入口；类别 source；类型 core-logic）：调用了 Uvicorn run 时传递新的超时参数，并移除了 monkey patch 调用。
- python/sglang/srt/environ.py（模块 环境变量；类别 source；类型 core-logic）：新增环境变量 SGLANG_UVICORN_WORKER_HEALTHCHECK_TIMEOUT，是配置中心。

关键符号：monkey_patch_uvicorn_multiprocessing

关键源码片段

python/sglang/srt/environ.py

新增环境变量 SGLANG_UVICORN_WORKER_HEALTHCHECK_TIMEOUT，是配置中心。

```
# python/sglang/srt/environ.py
# 在 Envs 类的 HTTP Server 区域新增：

# HTTP Server
SGLANG_TIMEOUT_KEEP_ALIVE = EnvInt(5)

# Uvicorn multiprocess supervisor pings each worker on this interval; default 5s is
# too short when many workers cold-start and load tokenizers in parallel.
# 新增：允许用户通过环境变量调整健康检查超时，默认 10 秒
# 使用方式：export SGLANG_UVICORN_WORKER_HEALTHCHECK_TIMEOUT=15
SGLANG_UVICORN_WORKER_HEALTHCHECK_TIMEOUT = EnvInt(10)

# HTTP/2 Server
SGLANG_GRANIAN_PARENT_PID = EnvInt(None)
```

评论区精华

无 review 评论，但 PR 由一名 reviewer 批准，未发现争议。

- 暂无高价值评论线程

风险与影响

- 风险：低风险。仅涉及启动配置，不改变运行时逻辑。新增环境变量默认与旧 monkey patch 的超时值一致（10 秒），对已有部署无影响。若用户自定义了该环境变量，需确保数值合理，过短会导致同样问题，过长可能延迟检测真正故障的 worker。
- 影响：直接解决多 tokenizer worker 模式下 worker 被误杀重启的 bug，影响使用 --tokenizer-worker-num > 1 的用户，尤其是大规模部署场景。系统稳定性提升，无需用户手动调整。
- 风险标记：暂无

关联脉络

- 暂无明显关联 PR