PR #20410 完整报告

sgl-project/sglang

[AMD] Add SGLANG_DISAGGREGATION_NUM_PRE_ALLOCATE_REQS env var for configurable KV transfer overlap

合并时间：2026-03-31 05:37

原文链接：http://prhub.com.cn/sgl-project/sglang/pull/20410

执行摘要

PR #20410 通过添加环境变量 SGLANG_DISAGGREGATION_NUM_PRE_ALLOCATE_REQS，允许在 PD 解聚模式中配置 KV 缓存传输重叠的额外槽位，从而提升解码吞吐量，特别适用于 AMD 大内存硬件如 MI355X，测试显示吞吐量改善约 1.5%。

功能与动机

此变更旨在解决 PD 解聚服务中的性能瓶颈：当 max_num_reqs > 32 时，之前的硬编码逻辑限制了 KV 传输与解码执行的重叠，影响吞吐量。PR body 强调：“在 PD 解聚服务中，多个 KV 传输应与正在进行的解码执行重叠”，以最大化硬件资源利用率，尤其是在 AMD 平台上。

实现拆解

- 配置层：在 environ.py 中新增环境变量定义，默认值为 0。
- 核心逻辑：修改 model_runner_kv_cache_mixin.py 的 _init_pools 函数： 这替代了原硬编码 0，允许用户在不增加 max_running_requests 的前提下配置额外槽位。

评论区精华

review 讨论极少，仅有一条来自 hnyls2002 的代码风格建议：

> “Import from top level.”

这促使导入语句优化，确保模块结构清晰，可能已在后续 commit 中调整。

风险与影响

- 风险：新增环境变量需文档支持以避免配置错误；增加槽位可能占用更多内存，需用户根据硬件资源谨慎设置。
- 影响：正面提升 AMD 平台解码性能，测试显示吞吐量从 7.40 req/s 增至 7.53 req/s；对现有系统默认无影响，但需注意内存使用增长。

关联脉络

该 PR 是近期 AMD 优化系列的一部分，与 PR #21234（AMD MXFP4 支持）和 PR #21315（AMD 融合 rope kv store）相关联，表明仓库正聚焦于 AMD 硬件的性能增强。这反映了团队在扩展硬件兼容性和优化资源利用方面的持续投入。