PR #29622 完整报告

sgl-project/sglang

Budget EAGLE/STANDALONE draft KV pool in SWA pool configurators

合并时间：2026-06-30 15:58

原文链接：http://prhub.com.cn/sgl-project/sglang/pull/29622

执行摘要

- 一句话：修复 SWA 池因未预算 draft KV 导致 OOM
- 推荐动作：这是一次必要且安全的 bugfix，逻辑清晰。特别推荐给关注 SWA 与 speculative decode 内存管理的工程师精读，尤其是 pool_configurator.py 中 _cell_size 的计算方式。

功能与动机

DefaultPoolConfigurator 已正确为 EAGLE/STANDALONE draft KV 预算内存（pool_configurator.py lines 132-149），但 SWA 配置器缺失此调整，导致混合 SWA 模型（如 Gemma, Command-R, MiMo）配合 EAGLE3 时出现 PD decode OOM。

实现拆解

1. 在 HybridSWAPoolConfigurator.__init__ 中新增 draft 层检测：通过 mr.spec_algorithm.is_eagle() 或 is_standalone() 判断并获取 mr.eagle_draft_num_layers，仅对非 draft 工作节点设置 self._draft_full_layers_num。
2. 修正 _cell_size 计算：在 full_layers_num == 0 分支增加 self._full_per_token * self._draft_full_layers_num；在 else 分支将 full 项由 self._full_layers_num 改为 self._full_layers_num + self._draft_full_layers_num。
3. 同步修正 SWAChunkCapPoolConfigurator.calculate_pool_sizes：full_cell_size 同样改为 self._full_per_token * (self._full_layers_num + self._draft_full_layers_num)，确保预算分配正确。
4. 无测试文件变更：仅源代码逻辑调整，CI 已通过基本验证。

关键文件：

- python/sglang/srt/model_executor/pool_configurator.py（模块 内存分配；类别 source；类型 data-contract；符号 HybridSWAPoolConfigurator.init, SWAChunkCapPoolConfigurator.calculate_pool_sizes）：唯一的变更文件。在两个 SWA 配置器中新增 draft 层预算逻辑，修正了 _cell_size 和 full_cell_size 的计算公式。

关键符号：HybridSWAPoolConfigurator.init, SWAChunkCapPoolConfigurator.calculate_pool_sizes

评论区精华

review 中无实质性讨论，所有评论均为 bot 配额提示和 CI rerun 命令。

- 暂无高价值评论线程

风险与影响

- 风险：【回归风险】当 draft_full_layers_num = 0（即无 speculative decode 场景）时，行为与修改前完全一致，计算式中 draft 项为 0，无回归风险。【性能风险】涉及 CUDA graph replay metadata 生成的变更，但此 PR 仅修正内存预算逻辑，不改变算子或图生成流程，性能影响可忽略。【兼容性】eagle_draft_num_layers 属性在修改前已存在，仅在非 draft worker 上使用，无兼容问题。
- 影响：影响范围：仅修改 pool_configurator.py，限制于 SWA 配置器（HybridSWAPoolConfigurator 和 SWAChunkCapPoolConfigurator）。直接影响使用 SWA 模型 + EAGLE/STANDALONE speculative decoding 的用户，解决 PD decode OOM。对于无 speculative decoding 的场景，行为无变化。
- 风险标记：无测试覆盖

关联脉络

- PR #29499 [DSA] Optimize DSA CUDA graph replay metadata generation: 同为 speculative decoding 相关内存优化，但本 PR 专注于 SWA 配置器的内存预算修正。