PR #27144 完整报告

sgl-project/sglang

Dllm radix cache npu

合并时间：2026-06-25 11:44

原文链接：http://prhub.com.cn/sgl-project/sglang/pull/27144

执行摘要

- 一句话：修复 NPU 上 DLLM 启用 radix cache 时页面大小不同步问题
- 推荐动作：该 PR 改动极小但修复了关键的一致性问题，建议合并。可快速阅读理解，无需精读全部文件。

功能与动机

在 NPU 上使用 DLLM 推理并启用 radix cache 时，由于 page_size 未正确同步到 server_args，导致缓存操作异常。PR body 中指出需 "Force page_size the same with block size" 以支持 Radix cache。基准测试显示修复后，启用 radix cache 使 GSM8K 100 样本延迟从 100s 降至 60s。

实现拆解

1. 在 python/sglang/srt/managers/scheduler.py 的 init_model_config 方法中，NPU 环境下如果 dllm_config.block_size < self.page_size，现有逻辑已将 self.page_size 回退为 dllm_config.block_size，但未更新 self.server_args.page_size。
2. 增加一行 self.server_args.page_size = self.dllm_config.block_size，使两者保持一致。
3. 该修改确保后续 radix cache 初始化及查询均使用与 block_size 一致的 page_size，从而正确命中缓存。

关键文件：

- python/sglang/srt/managers/scheduler.py（模块 调度器；类别 source；类型 core-logic）：核心调度器文件，该文件中的 init_model_config 方法是变更点，修复 page_size 与 server_args.page_size 不同步的问题。

关键符号：init_model_config

关键源码片段

python/sglang/srt/managers/scheduler.py

核心调度器文件，该文件中的 init_model_config 方法是变更点，修复 page_size 与 server_args.page_size 不同步的问题。

```
# python/sglang/srt/managers/scheduler.py (partial)
```

```
    def init_model_config(self):
        self.model_config = ModelConfig.from_server_args(self.server_args)
        if _is_npu:
            # 确保 page_size 不超过 block_size 和 chunked_prefill_size
            from sglang.srt.dllm.config import DllmConfig

            self.dllm_config = (
                DllmConfig.from_server_args(self.server_args)
                if self.server_args.dllm_algorithm is not None
                else None
            )
            if self.dllm_config:
                if self.dllm_config.block_size < self.page_size:
                    logger.warning(
                        "WARNING: "
                        f"The page size {self.page_size} should not be larger "
                        f"than dllm block size {self.dllm_config.block_size}."
                        f"Page size now falls back to {self.dllm_config.block_size}"
                    )
                    self.page_size = self.dllm_config.block_size
                    # 关键修复：同步 server_args 中的 page_size，
                    # 确保后续 radix cache 使用正确的页大小
                    self.server_args.page_size = self.dllm_config.block_size
```

评论区精华

PR 仅有一条来自 gemini-code-assist[bot] 的自动评审，确认修改正确且直接。无人工讨论。

- 暂无高价值评论线程

风险与影响

- 风险：风险极低。修改仅增加一行赋值，且仅在 NPU 后端 + DLLM 启用 + page_size 回退的特定条件下执行。不会影响其他后端或非 DLLM 场景。但缺少单元测试覆盖该条件分支，建议后续补充测试用例。
- 影响：影响范围限于 NPU 后端上使用 DLLM 推理并启用 radix cache 的场景。修复后 radix cache 可正确工作，显著提升推理性能（GSM8K 延迟降低 40%）。不影响其他平台或推理模式。
- 风险标记：缺少测试覆盖

关联脉络

- PR #27939 Support online MXFP8 quantization for ungated MoE: 同属 NPU 后端改进系列，但无直接代码依赖。
- PR #26980 [fix] Skip routed expert capture for draft model under spec v2: 同样涉及 NPU 后端条件和 scheduler 逻辑，但功能无关。