PR #27184 完整报告

sgl-project/sglang

docs: fix Nemotron Super MTP deployment command (spec-v2 + B200)

合并时间：2026-06-04 05:52

原文链接：http://prhub.com.cn/sgl-project/sglang/pull/27184

执行摘要

修复了 Nemotron3 Super 部署文档代码片段中 MTP 启用时的命令生成逻辑，修正了 radix cache 被错误禁用及 B200 上 attention backend 不兼容的问题。这是一个纯文档变更，但包含了有价值的工程设计决策。

功能与动机

PR Body 指出：启用 MTP 后生成的命令会丢弃 radix cache（因为使用了 --disable-radix-cache），且在 B200 硬件上默认的 attention backend（flashinfer）因 per-step plan() 的 host-sync 会阻塞 spec-v2 重叠调度器。修复目标：生成适用于不同硬件（H200/B200）的正确部署命令。

实现拆解

1. 修改 commandRule 函数签名：将 mtp 选项的 commandRule 由接收仅一个参数修改为接收 (value, state)，使函数能够访问完整的表单数据，特别是 hardware 选择。
2. 替换 MTP 命令片段：
   - 删除 --disable-radix-cache 参数。
   - 添加 --mamba-scheduler-strategy extra_buffer，从而在 spec decode 下保持 radix cache 开启。
   - 根据 state.hardware === 'b200' 条件，追加 --attention-backend trtllm_mha（仅在 B200 上）；H200 使用默认 fa3 无此需求。
3. 修改 generateCommand 函数：
   - 当 MTP 启用时，在命令前添加 SGLANG_ENABLE_SPEC_V2=1 环境变量前缀。
   - 将启动命令由 python3 -m sglang.launch_server 改为 sglang serve（第二个 commit）。
4. 更新 commandRule 调用：在 generateCommand 循环中，将 option.commandRule(values[key]) 改为 option.commandRule(values[key], values)，将完整状态传入。

docs_new/src/snippets/autoregressive/nemotron3-super-deployment.jsx

主变更文件，修复了 MTP 命令生成的核心逻辑，涉及 commandRule 签名与实现、generateCommand 函数及环境变量设置。

```
// 以下代码片段展示了关键变更：
// - commandRule 现在接收 (value, state) 以访问硬件选项
```

```
// - 返回值不再包含 --disable-radix-cache，而是使用 --mamba-scheduler-strategy extra_buffer
// - 只在 B200 上追加 --attention-backend trtllm_mha
commandRule: (value, state) => value === 'enabled'
  ? '--speculative-algorithm EAGLE \\
  --speculative-num-steps 3 \\
  --speculative-eagle-topk 1 \\
  --speculative-num-draft-tokens 4 \\
  --mamba-scheduler-strategy extra_buffer'
    + (state.hardware === 'b200'
      ? ' \\
  --attention-backend trtllm_mha'
      : '')
  : null
```

评论区精华

无 review 讨论。

风险与影响

风险：低。仅为文档代码片段变更，不影响运行时。但未来硬件或最佳实践变化时，此处的硬编码字符串 'b200' 需要同步更新。

影响：仅影响 Nemotron3 Super 部署文档页面。用户复制后得到的命令将正确保留 radix cache 并兼容 B200 硬件。

关联脉络

该 PR 与之前的 PR #25198 同属 Nemotron 系列文档更新。此外，PR #26997（spec v2 tree drafting reland）与 --mamba-scheduler-strategy extra_buffer 等参数相关，可参考了解 spec v2 的整体设计。