PR #23891 完整报告

sgl-project/sglang

[NPU] Support radix-cache with mamba-extra-buffer for Qwen3.5

合并时间：2026-05-11 09:40

原文链接：http://prhub.com.cn/sgl-project/sglang/pull/23891

执行摘要

- 一句话：NPU 启用 Qwen3.5 前缀缓存支持
- 推荐动作：PR 改动简单明确，建议直接合并。有兴趣了解 NPU 端 mamba 调度策略演进的人员可以关注后续针对 NPU 的测试文档更新。

功能与动机

为 Qwen3.5 模型适配前缀缓存（radix-cache），并更新 NPU 设备上 mamba extra_buffer 策略的条件检查，以支持通过 --mamba-scheduler-strategy extra_buffer 启用该优化。

实现拆解

1. 修改断言条件：在 python/sglang/srt/server_args.py 的 _handle_mamba_radix_cache 函数中，将原本只允许 CUDA 和 MUSA 设备使用 mamba extra_buffer 的断言条件扩展，增加对 NPU 设备的支持。具体将 is_cuda() or is_musa() 改为 is_cuda() or is_musa() or is_npu()。
2. 更新错误提示信息：同步更新断言失败时的错误提示字符串，将 "CUDA and MUSA" 更新为 "CUDA and MUSA and NPU"，更准确地反映支持的硬件范围。

关键文件：

- python/sglang/srt/server_args.py（模块 服务器参数；类别 source；类型 core-logic；符号 _handle_mamba_radix_cache）：这是唯一的变更文件，修改了 mamba extra_buffer 策略的设备支持断言，将 NPU 设备纳入支持范围。

关键符号：_handle_mamba_radix_cache

关键源码片段

python/sglang/srt/server_args.py

这是唯一的变更文件，修改了 mamba extra_buffer 策略的设备支持断言，将 NPU 设备纳入支持范围。

```
# 修改位于 _handle_mamba_radix_cache 方法中
# 变更前：is_cuda() or is_musa()
# 变更后：is_cuda() or is_musa() or is_npu()
assert (
```

```
    is_cuda() or is_musa() or is_npu()
), "Mamba extra_buffer is only supported on CUDA and MUSA and NPU devices with FLA backend"
```

评论区精华

Reviewer Hexq0210 询问一处条件逻辑是否应为 "and" 改为 "or"（原 diff 显示有误），作者 silencejade 回复已修复。最终合并的代码正确使用了 `or` 逻辑。

- 断言条件逻辑修正 (design): 作者 silencejade 回复 'already fix'，最终合并版本正确使用了 'or' 逻辑。

风险与影响

- 风险：该变更仅涉及一个断言条件，添加了对 NPU 的支持，没有改变其他设备的逻辑。风险极低，因为只是放宽了检查，不会引入回归。潜在的兼容性风险是如果 NPU 后端实际上不支持 extra_buffer 策略，但断言被允许通过，可能导致运行时错误；但从 PR 上下文看，NPU 后端已适配该功能。
- 影响：对用户：NPU 用户现可为 Qwen3.5 模型启用 mamba extra_buffer 策略和 radix-cache，可能带来性能提升。对系统：无负面影响。对团队：只需更新一个文件，无维护负担。
- 风险标记：暂无

关联脉络

- 暂无明显关联 PR