PR #15829 完整报告

sgl-project/sglang

[feat] Support `extra_buffer` in Mamba2-based models

合并时间：2026-05-26 16:03

原文链接：http://prhub.com.cn/sgl-project/sglang/pull/15829

执行摘要

- 一句话：为 Mamba2 模型支持 extra_buffer 调度策略
- 推荐动作：该 PR 展示了如何将 extra_buffer 策略从 FLA 扩展到 Mamba2，其 chunk size 统一思路值得借鉴。但由于合并后出现回归，建议暂停部署，待作者修复后重新 review。阅读此 PR 可重点关注 _init_track_ssm_indices 中的索引映射逻辑和 mamba_cache_chunk_size 的动态计算过程。

功能与动机

近期对 Qwen3-Next 模型的更新使其能够同时使用 radix cache 和 overlap scheduler（extra_buffer 策略）。本 PR 旨在为 Mamba2 模型提供相同的支持，使其也能享受这一调度优化。

实现拆解

1. Kernel 修改：在 mamba_chunk_scan_combined 中增加 return_intermediate_states=True，使 prefill 过程中间状态可被提取。对应文件 python/sglang/srt/layers/attention/mamba/mamba.py。
2. 状态跟踪索引统一：在 hybrid_linear_attn_backend.py 中，将 _init_track_conv_indices 和 _init_track_ssm_indices 中原先硬编码的 FLA_CHUNK_SIZE 替换为动态的 mamba_cache_chunk_size（来自 server_args），并区分 FLA 和 Mamba2 分支计算 num_h_states。
3. ServerArgs 重构：mamba_cache_chunk_size 属性现在优先读取模型配置中的 mamba_chunk_size（Mamba2 模型），否则回退到 FLA_CHUNK_SIZE；移除了原有的整除性断言。对应文件 python/sglang/srt/server_args.py。
4. 调度器和缓存适配：修改 schedule_batch.py 和 mamba_radix_cache.py 以使用统一的 chunk size，确保 Mamba2 模型能正确进行 radix cache 跟踪。
5. 模型适配：为 NemotronH 和 FalconH1 的注意力后端增加 forward_batch 参数传递。调整 GraniteMoEHybrid 的默认 attention backend 为 flashinfer。
6. 测试：
   - test_disaggregation_hybrid_attention.py：新增基于 Nemotron Nano v2 的 Mamba2 测试类（含 extra_buffer 和 DP decode），原 Qwen3-Next 测试类重命名为 GDN。

- test_nvidia_nemotron_nano_v2.py：新增 TestNvidiaNemotronNanoV2BF16ExtraBuffer（含 KL 散度和前缀缓存分支）和 TestNvidiaNemotronNanoV2SpeculativeDecodingExtraBuffer。
- test_granite_moe_hybrid.py：新增 TestGraniteMoEHybrid 和 TestGraniteMoEHybridExtraBuffer。

关键文件：

- test/registered/disaggregation/test_disaggregation_hybrid_attention.py（模块 分离测试；类别 test；类型 test-coverage；符号 TestDisaggregationHybridAttentionMamba, TestDisaggregationHybridAttentionGDN, TestDisaggregationHybridAttentionMambaExtraBuffer, TestDisaggregationHybridAttentionGDNExtraBuffer）：将 P/D 分离测试从 Qwen3-Next 切换为 Nemotron Nano v2（Mamba2 模型），并新增 extra_buffer 和 DP decode 测试用例。
- test/manual/models/test_nvidia_nemotron_nano_v2.py（模块 模型测试；类别 test；类型 test-coverage；符号 TestNvidiaNemotronNanoV2FP8, TestNvidiaNemotronNanoV2BF16ExtraBuffer, TestNvidiaNemotronNanoV2SpeculativeDecodingExtraBuffer）：新增 extra_buffer 和 speculative decoding 测试，使用 KLDivergenceMixin 和 PrefixCacheBranchingMixin 验证正确性。
- python/sglang/srt/layers/attention/hybrid_linear_attn_backend.py（模块 注意力后端；类别 source；类型 core-logic）：核心变更：统一了 conv state 和 SSM state 的索引跟踪逻辑，使其同时支持 FLA（chunk size=64）和 Mamba2（chunk size=256）的不同 chunk size。
- python/sglang/srt/server_args.py（模块 服务配置；类别 source；类型 core-logic）：重构 mamba_cache_chunk_size 属性以支持 Mamba2 模型，新增 fallback_attention_backend 参数。
- python/sglang/srt/layers/attention/mamba/mamba.py（模块 Mamba2 模型；类别 source；类型 dependency-wiring）：让 MambaMixer2.forward 返回 intermediate_states，供 extra_buffer 路径提取缓存。
- test/manual/models/test_granite_moe_hybrid.py（模块 模型测试；类别 test；类型 test-coverage；符号 TestGraniteMoeHybrid, TestGraniteMoeHybridExtraBuffer）：新增文件，测试 Granite MoE Hybrid 模型的 extra_buffer 支持。

关键符号：_init_track_ssm_indices, _init_track_conv_indices, mamba_cache_chunk_size, MambaMixer2.forward, Mamba2AttnBackend.forward, _handle_mamba_radix_cache

评论区精华

Review 中 Hanming Lu 指出不应直接用 mamba_cache_chunk_size 替换 FLA_CHUNK_SIZE，应区分 backend；经讨论后作者改为使用统一属性并保留分支。他还质疑 schedule_batch.py 中 mamba_track_interval 的改动，认为可能因值恰好为 256 才未触发问题。最终他 approve 了 PR。合并后 ispobock 报告了回归（disaggregation 测试失败），要求回退并提供复现步骤。fzyzcjy 关联了 PR #25656 的类似失败。ispobock 进一步分析指出

_init_track_ssm_indices 中 Mamba2 分支的 track_ssm_h_src 计算可能存在 global vs per-sequence 索引错位。

- Chunk size 统一逻辑设计 (design): 最终达成一致，使用 server_args.mamba_cache_chunk_size 动态获取。
- 合并后回归问题 (correctness): PR 被回退，回归未修复。
- mamba_track_interval 改动可能错误 (correctness): 作者未直接回应但后续 approve。

风险与影响

- 风险：核心风险在于索引计算逻辑的变更可能影响缓存提取的正确性，尤其是 SSM state 的 src/dst 索引映射。统一 chunk size 后，Mamba2（chunk size=256）与 FLA（chunk size=64）的路径复用同一份代码，若未充分测试所有边界条件（如对齐 / 非对齐序列），可能导致缓存内容错误或越界。此外，修改了 schedule_batch.py 中的 track 计算，可能影响调度器的正确性。该 PR 合并后已导致 disaggregation 测试失败（回归），证明了上述风险。
- 影响：对用户：成功时可让 Mamba2 模型同时受益于 radix cache 和 overlap scheduler，提升推理吞吐。对系统：修改了多个核心模块（server_args、schedule_batch、attention backend），影响范围广，任何 Mamba2 模型（Nemotron、Granite、FalconH1）都会受影响。对团队：需要修复回归后重新合并，且需加强 Mamba2 场景的测试覆盖（尤其是边界序列、PD 分离模式）。
- 风险标记：核心路径变更，回归风险，索引计算复杂，缺少边界测试

关联脉络

- PR #26838 Skip flaky mamba extra_buffer disagg test: 该 PR 跳过了因本 PR 导致回归的 flaky 测试。