PR #44251 完整报告

vllm-project/vllm

[Perf] Add tuned selective_state_update configs for H200 and RTX PRO ⋯

合并时间：2026-06-03 14:59

原文链接：http://prhub.com.cn/vllm-project/vllm/pull/44251

执行摘要

本 PR 为 NVIDIA H200 和 RTX PRO 6000 Blackwell Server Edition GPU 添加了调优的 selective_state_update 内核配置文件，是 PR #43083 的延续。H200 在 float16 下获得最高 2.9x 内核加速，端到端吞吐量提升 2.6%。所有配置均通过精度验证，风险极低，推荐合并。

功能与动机

PR #43083 已为 B200、GB200 和 H100_80GB_HBM3 提供了调优配置。但在 H200 和 RTX PRO 6000 Blackwell 上，Triton 加载器会回退到内置启发式，导致明显性能损失。本 PR 通过自动化基准测试脚本生成并验证针对这两款 GPU 的配置，确保它们也能获得最佳性能。

实现拆解

1. 调优脚本：使用 benchmarks.kernels.benchmark_selective_state_update 脚本，固定 headdim=64, dstate=128，分别以 float16 和 float32 运行，nheads 从 8 到 256 不等。
2. 生成配置：对每个有效批量大小（8-262144），脚本搜索最优的 BLOCK_SIZE_M 和 num_warps 组合，并输出 JSON 文件。
3. 验证：每个配置都通过数值验证（atol=0.01），保证与内核默认实现一致。
4. 存放：4 个 JSON 文件按设备名和数据类型命名，直接部署到 vllm/model_executor/layers/mamba/ops/configs/selective_state_update/ 目录，加载器自动识别。

关键源码片段

vllm/model_executor/layers/mamba/ops/configs/selective_state_update/headdim=64,dstate=128,device_name=NVIDIA_H200,cache_dtype=float16.json

H200 float16 配置，加速效果最为显著（最高 2.9x），覆盖主要使用场景

```
{
  // Triton 版本标识，用于向后兼容
  "triton_version": "3.6.0",
  // 键为有效批量大小 (effective batch)，值为 {BLOCK_SIZE_M, num_warps}
  "8": { "BLOCK_SIZE_M": 4, "num_warps": 2 },
  "16": { "BLOCK_SIZE_M": 4, "num_warps": 1 },
  "32": { "BLOCK_SIZE_M": 4, "num_warps": 1 },
  "64": { "BLOCK_SIZE_M": 16, "num_warps": 4 },
```

```
    "128": { "BLOCK_SIZE_M": 8, "num_warps": 2 },
    "256": { "BLOCK_SIZE_M": 8, "num_warps": 2 },
    "512": { "BLOCK_SIZE_M": 16, "num_warps": 1 },
    "1024": { "BLOCK_SIZE_M": 16, "num_warps": 1 },
    "2048": { "BLOCK_SIZE_M": 8, "num_warps": 2 },
    "4096": { "BLOCK_SIZE_M": 16, "num_warps": 2 },
    "8192": { "BLOCK_SIZE_M": 32, "num_warps": 2 },
    // 更大的批量通常需要更大 BLOCK_SIZE 和更多 warp
    "196608": { "BLOCK_SIZE_M": 16, "num_warps": 2 },
    "262144": { "BLOCK_SIZE_M": 16, "num_warps": 2 }
  }
```

评论区精华

审核人 tomeras91直接批准，仅留下“LGTM!”，作者随后请求合并。无技术讨论。

风险与影响

- 风险：极低。配置变更不影响原有设备，且加载失败时自动回退；所有数值通过验证。
- 影响：正面。H200 和 RTX PRO 6000 Blackwell 用户可获得显著的 SSM 内核加速（特别是 float16 场景），而其他用户无影响。

关联脉络

本 PR 是 PR #43083 的后续，共同构成完整的 selective_state_update 调优覆盖链，目前覆盖 B200、GB200、H100、H200 和 RTX PRO 6000 Blackwell。未来可能还需要为其他 GPU（如 AMD MI300）添加类似配置。