PR #35628 完整报告

sgl-project/sglang

[AMD] Increase gfx950 DSA model indexer topk_transform kernel occupancy

合并时间：2026-08-28 13:20

原文链接：http://prhub.com.cn/sgl-project/sglang/pull/35628

执行摘要

- 一句话：提升 gfx950 topk_transform 内核占用率
- 推荐动作：建议精读。该 PR 虽然改动简单，但体现了通过精细控制共享内存预算来平衡性能与精度的工程实践。值得关注其决策过程：如何通过分析内核的 LDS 需求，在占用率和正确性之间找到平衡点。对于维护 AMD 内核的工程师，此调整提供了有用的参考。

功能与动机

原实现在 gfx95x 上为 topk 内核分配了 128KB 动态 LDS，导致每个 CU 只能运行 1 个 block，约 2/3 的线程能力闲置。PR 正文指出：'The large buffer is unnecessary: it only backs the radix tie-candidate list, which needs ~length/256 entries.' 通过缩减缓冲区到 40KB，在保持正确性边界的前提下，将占用率提升至 3 blocks/CU，捕获约 2 倍于可用 ~2.5x 的理论加速。背景是 DeepSeek-V4 的 topk_transform 内核性能优化。

实现拆解

本 PR 的改动非常集中，仅涉及一个文件的配置调整。

1. 修改文件：python/sglang/kernels/aot/setup_rocm.py。
2. 核心变更：将 topk_dynamic_smem_bytes 的计算逻辑从 48 * 1024（gfx942）或 32 * 1024 * 4（gfx95x/gfx1250）改为 48 * 1024（gfx942）或 40 * 1024（gfx95x/gfx1250）。
3. 关键符号：topk_dynamic_smem_bytes，该变量通过 -DSGL_TOPK_DYNAMIC_SMEM_BYTES 宏传递给 hipcc 编译的 topk 内核。
4. 影响分析：该宏直接影响内核中动态共享内存的分配，从而影响占用率。40KB 的动态预算结合静态约 7KB，使得单个 block 的 LDS 约 47KB，在 gfx950 的 160KB LDS/CU 限制下可容纳 3 个 block。同时，40KB 预算对应 5120 个 tie 候选条目，可覆盖约 1.3M token 的序列，超出 DeepSeek-V4 的 1M 默认上下文，保证了正确性。
5. 测试与部署：PR 未新增或修改测试文件，但提供了详细的基准数据（见 PR 正文表格），展示了不同 batch size 和序列长度下的性能提升。该改动属于编译时配置，无需额外部署步骤。

关键文件：

- python/sglang/kernels/aot/setup_rocm.py（模块 内核构建；类别 source；类型 core-logic；符号 topk_dynamic_smem_bytes）：本 PR 唯一改动文件，调整了编译期动态共享内存预算，直接决定 topk 内核占用率。

关键符号：未识别

关键源码片段

python/sglang/kernels/aot/setup_rocm.py

本 PR 唯一改动文件，调整了编译期动态共享内存预算，直接决定 topk 内核占用率。

```
# python/sglang/kernels/aot/setup_rocm.py
# ...

# Dynamic shared-memory budget for the TopK kernels.
# - gfx942 (MI300/MI325): LDS is typically 64KB per workgroup -> keep dynamic smem <=
~48KB
# (leaves room for static shared allocations in the kernel).
# - gfx95x (MI350) and gfx1250: LDS is larger. Large dynamic budget wastes LDS
# and pins occupancy to 1 block/CU. Keep it small (40KB) for better occupancy.
topk_dynamic_smem_bytes = 48 * 1024 if amdgpu_target == "gfx942" else 40 * 1024

# 该值通过编译宏传递给 topk 内核，直接影响 LDS 分配。
# 40KB 动态预算 + 约 7KB 静态 LDS = 约 47KB/block，
# 在 gfx950 的 160KB LDS/CU 限制下可容纳 3 个 block，
# 从而将 512 线程块的占用率从 1 提升至 3。
# 对应 tie 候选列表 5120 条目，可覆盖约 1.3M token 序列，
# 高于 DeepSeek-V4 默认 1M 上下文，保证正确性。

hipcc_flags = [
    "-DNDEBUG",
    f"-DOPERATOR_NAMESPACE={operator_namespace}",
    "-O3",
    "-Xcompiler",
    "-fPIC",
    "-std=c++17",
    f"--amdgpu-target={amdgpu_target}",
    "-DENABLE_BF16",
    "-DENABLE_FP8",
    fp8_macro,
    f"-DSGL_TOPK_DYNAMIC_SMEM_BYTES={topk_dynamic_smem_bytes}",
]
```

评论区精华

PR 获得审核者 1am9trash 的批准，评论为 'LGTM, only effect AMD code path.'。PR 正文中详细讨论了 128KB、48KB、40KB、32KB 等不同预算的权衡，并最终选择 40KB 作为性能与精度的平衡点。无其他未解决疑虑。

- 暂无高价值评论线程

风险与影响

- 风险：技术风险较低，但存在以下潜在问题：
  1. 正确性边界：40KB 动态 LDS 对应的 tie 缓冲可覆盖约 1.3M token，而 DeepSeek-V4 的默认上下文为 1M token。若未来模型上下文长度扩展超过此界限，可能导致 tie 缓冲不足，影响 topk 结果的精度。该风险在 PR 中已明确说明。
  2. 性能回归：对于小 batch（如 B=256）或短序列（S=8192），40KB 预算的性能提升有限（约 0.99x），无明显回归。但对于不同 batch 和序列组合，可能存在未知的性能波动。
  3. 兼容性：本次改动仅影响 AMD 架构上的编译宏，对非 AMD 平台无影响。gfx942 保持不变，兼容性风险低。 - 影响：影响范围：仅影响 AMD 架构（gfx95x 和 gfx1250）上的 DeepSeek-V4 topk_transform 内核性能。对于运行 DeepSeek-V4 的 MI350/MI355 用户，长上下文场景可获约 2.3-2.4 倍性能提升。代码改动量极小（+3/-2），无 API 或配置兼容性影响。对团队而言，此类调整展示了通过编译期参数调优内核的实践，但需注意未来模型上下文长度扩展时的回归风险。 - 风险标记：正确性边界，无测试配套

关联脉络

- PR #36684 [AMD] Enable deepseek-v4 topk_transform v2 kernel: 同为 DeepSeek-V4 topk_transform 内核的 AMD 优化，改动了相同内核的宏配置，本 PR 可能基于该 PR 调整占用率。