

PR #37898 完整报告

sgl-project/sglang

sm120 32GB mem-tier: raise decode cuda-graph max_bs 24->48 + chunked_prefill 2k->4k

合并时间: 2026-09-04 18:47

原文链接: <http://prhub.com.cn/sgl-project/sglang/pull/37898>

执行摘要

- 一句话: 提升 32GB GPU 内存层的 CUDA Graph batch size 和 chunked prefill 上限, 修复高并发解码吞吐量崩溃问题。
- 推荐动作: 这是一个直接且效果显著的性能调优 PR。建议技术管理者关注其在 32GB GPU 产品线上的性能提升, 并考虑将其作为基准性能的一部分。对于工程师, 值得精读 `memory_hook.py` 中关于 GPU 显存层级划分和参数设置的设计, 理解其如何根据硬件特性动态优化配置。PR 提供的详细验证方法和数据也值得在后续性能优化工作中借鉴。

功能与动机

在 RTX 5090 (32GB, sm_120) 上, 之前解码 CUDA Graph 的 `max_bs=24` 限制导致在 batch size ≥ 32 时无法使用 CUDA Graph, 被迫采用 eager 解码, 这严重崩溃了高并发吞吐量, 使其远低于 vLLM。PR body 明确指出“32GB Blackwell (RTX 5090) can hold decode cuda graphs well past bs=24; the previous cap forced eager decode at bs $\geq$ 32 and collapsed high-concurrency throughput vs vLLM.”

实现拆解

变更入口: 修改 `python/sglang/srt/arg_groups/memory_hook.py` 文件中的 `handle_gpu_memory_settings` 函数。该函数根据 GPU 显存大小自动设置关键运行时参数。

1. 调整 32GB 内存层默认参数: 针对 `gpu_mem < 35 * 1024` (对应 A10、4090、5090 等 32GB GPU) 的内存层, 将 `chunked_prefill_size` 的默认值从 2048 调整为 4096。同时, 将 `decode_cuda_graph_config.max_bs` 的默认值从 24 (tp<4 时) 提升至 48, 从 80 (tp $\geq$ 4 时) 提升至 160。
2. 分析参数修改的影响: 根据 PR body 描述, 虽然增大了 CUDA Graph 配置, 但对最终的 KV 池大小 (`mem_fraction_static`) 影响是正面的。因为 `chunked_prefill` 对应的激活 - token 项取值逻辑被 `max(chunked_prefill, 2048)` 限制, 其大小不变; 而增大的 CUDA Graph 工作区在 KV 尺寸计算前被捕获, 因此仅需为实际捕获成本预留内存, 净效果是 KV 池可能增加约 2GB。
3. 补充说明: 此变更为纯配置调整, 不涉及算法逻辑修改, 因此无需配套测试。PR 在提交前已在 RTX 5090 x8 集群上对多个模型 (Qwen3.5-9B、Qwen3.5-4B、Qwen2.5-VL-7B、gpt-oss-20b) 进行了充分验证, 结果显示解码性能大幅提升, 解决了吞吐量崩溃问题。

关键文件:

- python/sglang/srt/arg_groups/memory_hook.py (模块 内存配置; 类别 source; 类型 configuration; 符号 handle_gpu_memory_settings) : 此文件包含根据 GPU 显存自动设置关键运行时参数 (如 CUDA Graph batch size、chunked prefill 大小) 的核心逻辑 handle_gpu_memory_settings。本次 PR 仅修改了此文件, 调整了 32GB 内存层的默认配置, 直接影响系统在该硬件上的性能表现。

关键符号: handle_gpu_memory_settings

关键源码片段

python/sglang/srt/arg_groups/memory_hook.py

此文件包含根据 GPU 显存自动设置关键运行时参数 (如 CUDA Graph batch size、chunked prefill 大小) 的核心逻辑 `handle_gpu_memory_settings`。本次 PR 仅修改了此文件, 调整了 32GB 内存层的默认配置, 直接影响系统在该硬件上的性能表现。

```
# --- 变更后的代码片段 (handle_gpu_memory_settings 函数内) ---
elif gpu_mem < 35 * 1024:
    # A10, 4090, 5090
    # (chunked_prefill_size 4k, max_bs 48 if tp < 4 else 160)
    # 32GB Blackwell (RTX 5090) can hold decode cuda graphs well past
    # bs=24; the previous cap forced eager decode at bs>=32 and
    # collapsed high-concurrency throughput vs vLLM.
    if cfg.chunked_prefill_size is None:
        declare_resolution(
            server_args,
            "_handle_gpu_memory_settings",
            # 将默认的 chunked prefill 大小从 2048 提升到 4096, 以利用更大的显存。
            chunked_prefill_size=4096,
        )
    if decode_cuda_graph_config.max_bs is None:
        if cfg.tp_size < 4:
            # 单卡或 tp<4 场景: 将最大解码 batch size 从 24 提升到 48。
            # 这允许在更高并发下继续使用 CUDA Graph, 避免退化为 eager 解码。
            decode_cuda_graph_config.max_bs = 48
        else:
            # tp>=4 场景: 将最大解码 batch size 从 80 提升到 160。
            decode_cuda_graph_config.max_bs = 160
```

评论区精华

由于 PR 评论区为空, 讨论点主要源自 PR body 和代码注释:

- 问题确认: PR body 明确指出了原有参数 (max_bs=24) 在 32GB GPU 上导致性能崩溃的根因是“forced eager decode at bs>=32”。
- 参数选择依据: 代码注释更新为“32GB Blackwell (RTX 5090) can hold decode cuda graphs well past bs=24”, 表明新参数 (48/160) 是基于对 32GB 显存容量的重新评估。
- 性能验证: PR 提供了详尽的对比数据, 显示调整后在不同模型和 batch size 上均超越 vLLM, 为变更的合理性提供了强有力的实证支持。

- 32GB GPU 默认参数选择与性能验证 (performance): 通过实证数据确立了新的、更优的默认参数配置, 解决了已知的性能瓶颈。
- 参数调整对显存分配的净影响 (correctness): 参数调整在提升性能的同时, 可能优化了资源利用, 其影响比表面看起来更积极。

风险与影响

- 风险:

1. 显存溢出风险: 虽然 PR 声称新参数是安全的, 但提升 `max_bs` 和 `chunked_prefill_size` 意味着需要捕获和运行更大的 CUDA Graph 及处理更大的 prefill 批次, 这会消耗更多显存。如果用户同时运行多个实例或使用极端配置, 仍有触发 OOM 的可能。
2. 兼容性影响: 此变更仅适用于符合 `gpu_mem < 35 * 1024` 条件的 GPU。对于其他内存层 (如 40GB、80GB) 的 GPU 行为无影响。但若用户手动覆盖了 `chunked_prefill_size` 或 `cuda_graph_config.max_bs` 参数, 则此默认调整不会生效, 可能造成混淆。
3. 缺乏自动化测试: 此配置变更没有配套的单元或集成测试, 完全依赖于人工验证。未来若内存管理逻辑或 CUDA Graph 捕获成本发生变化, 此默认值可能再次变得不适用, 但缺乏测试来自动捕获此类回归。

- 影响:

1. 用户影响: 直接惠及所有使用 RTX 5090、RTX 4090、A10 等 32GB 显存 GPU 的用户。在高并发 (`batch size >= 32`) 解码场景下, 吞吐量将获得大幅提升, 性能可与或超越 vLLM。根据验证数据, 提升幅度在 4% 到 71% 之间。
2. 系统影响: 改变了系统在特定硬件上的默认资源分配策略。在解码性能提升的同时, 理论上 KV 缓存容量也可能略微增加, 这有助于改善整体服务容量。
3. 团队影响: 为 32GB GPU 提供了更优的开箱即用性能, 减少了用户调参的必要性。但同时也增加了内存配置参数的复杂性, 需要文档跟进说明不同内存层的默认值。 - 风险标记: 显存管理风险, 缺少自动化测试

关联脉络

- PR #37930 [CI] Fix handle_platform_cp_compatibility reading legacy CP flags off the record: 两个 PR 都涉及对 `sglang/srt/arg_groups/` 下模块中配置处理逻辑的修改, 可能影响系统初始化和运行时行为。
- PR #37546 [sp] Make attention-TP sequence sharding a per-forward batch property: 此 PR 也涉及对调度和运行时状态管理的优化, 旨在减少全局状态带来的问题, 与本 PR 优化调度性能 (避免 eager decode) 的目标有间接关联。