

PR #39713 完整报告

sgl-project/sglang

[Router] Bound the e2e worker memory budget so prefill graph capture stops OOMing

合并时间: 2026-09-16 15:13

原文链接: <http://prhub.com.cn/sgl-project/sglang/pull/39713>

执行摘要

- 一句话: 为 Router e2e worker 限定显存预算并禁用 prefill 图, 消除启动 OOM
- 推荐动作: 建议精读, 但重点不在 diff (只有 22 行、单一测试基建文件), 而在 PR body 的诊断链路。值得关注三点: 一是「KV 池按卡容量分配、slack 却被激活预留独占」这条内存计费缺陷的完整账本, 它解释了许多人遇到过的「小模型反而更容易 CUDA graph OOM」现象; 二是把测试夹具参数收敛成单个 CI_WORKER_ARGS 常量、复用既有 worker_args 契约而不改读侧的设计手法; 三是作者主动标注的后续动作——应针对 memory profiler 开 issue, 让 KV sizing 为 graph capture 预留额度。评审 / 维护者后续应确认两件事: 新增 e2e spec 是否强制携带 CI_WORKER_ARGS, 以及在更大显存或非 H100 runner 上 0.6 这个比例是否仍然合适。

功能与动机

PR body 直接给出失败现场: tier-3 — e2e job 间歇性报 `RuntimeError: sglang worker exited during startup with code -9`, 作者强调这是「a router test that never reaches the router」。根因是 worker 走裸引擎默认值: `mem_fraction_static=0.83` 留给 KV 池的显存几乎等于卡容量 (0.6B 模型分到 556K tokens, 而测试只生成 16 个 token), 只留 `pre_model_load_memory * (1 - mem_fraction_static)` (约 13 GB, 计量点在 `python/sglang/srt/mem_cache/kv_cache_configurator.py:2193`、`python/sglang/srt/distributed/bootstrap.py:145`) 同时覆盖激活与所有 CUDA graph, 而默认 `chunked_prefill_size=8192` 的激活工作集按引擎自算的 `pre_capture_activation_reserve_mb` (`python/sglang/srt/runtime_context.py:1842`) 已接近 12.6 GB, 导致 58-bucket breakable prefill graph 在第 26 个 bucket 就 CUDA OOM。PR body 用三条两天内、三个不同分支的 CI run (35038908194 / 35009011984 / 34929381917) 证明这不是某个 PR 引入的问题, 而是套件级基建缺陷。

实现拆解

1. 新增模块级内存预算常量: 在 `experimental/sgl-router/tests/e2e/infra/model_specs.py` 中定义 `CI_WORKER_ARGS: list[str]`, 内含 `--mem-fraction-static=0.6` 与 `--cuda-graph-backend-prefill=disabled` 两项。之所以用 0.6 而不是其他值, 是为了把「KV 池吃满剩余显存」改成显式比例, 从而为激活工作集和 (被跳过的) 图捕获留出预算; 禁用 prefill 图则直接把最容易触顶的 58-bucket capture 环节移出启动路径, 顺带省掉每次 spawn 的捕获时间——这对「每个测试都重启 worker」的 e2e 套件是额外收益。文件

里保留了十余行注释，记录 (1 - mem_fraction_static) slack 与 chunked_prefill_size=8192 激活预留之间的账本推导，便于后续调参者理解数值来源。

- 2. 挂载到模型规格表：MODEL_SPECS 的 qwen3-0.6b、llama-1b、llama-8b 三个条目各新增 "worker_args": CI_WORKER_ARGS。选择复用同一常量而非逐条写死，是为了避免三条 spec 的参数随时间漂移；同时 worker_args 本身就是模块 docstring 第 13 行早已声明的可选字段 (optional extra sglang.launch_server flags)，因此这是填充既有契约，而不是扩展契约。
- 3. 读侧零改动：get_model_spec() 依旧原样返回 spec dict，get_models_with_feature() 依旧只按 features 过滤，e2e 用例侧无需感知新参数；影响范围收敛在 worker 启动命令行拼接这一处。
- 4. 参数合法性核对（无代码改动，但属于配套验证）：作者在 PR body 中说明已把两个 flag 对照 python/sglang/srt/arg_groups/fields/schedule.py:34、python/sglang/srt/arg_groups/fields/exec_.py:476 以及 python/sglang/srt/model_executor/cuda_graph_config.py:47 的 Backend.ALL 校验，确认 disabled 是允许的 prefill backend，也正是引擎自身 OOM 提示推荐的取值 (prefill_cuda_graph_runner.py:254)。
- 5. 测试与配置配套：本 PR 没有新增测试文件，验证路径是 pre-commit、get_model_spec 能解析出新参数、以及 PR 自身的 tier-3 — e2e 真实 GPU job (PR body 的 Tests 段落明确写了「Real verification is the tier-3 — e2e job on this PR」)。横向证据是：套件里唯一从未踩到该 OOM 的真实 GPU 测试正是已固定 --max-total-tokens 的 test_hicache_storage_tiers.py:96。

模型条目	变更前	变更后
qwen3-0.6b	无 worker_args，走引擎默认	worker_args = CI_WORKER_ARGS
llama-1b	无 worker_args，走引擎默认	worker_args = CI_WORKER_ARGS
llama-8b	无 worker_args，走引擎默认	worker_args = CI_WORKER_ARGS

关键文件：

- experimental/sgl-router/tests/e2e/infra/model_specs.py (模块 测试夹具；类别 test；类型 configuration；符号 CI_WORKER_ARGS, MODEL_SPECS, get_model_spec)：本 PR 唯一改动文件。新增模块级 CI_WORKER_ARGS (--mem-fraction-static=0.6、--cuda-graph-backend-prefill=disabled) 并挂到三个模型 spec 的 worker_args，是全套 e2e worker 显存口径的唯一来源；文件内的长注释保留了 OOM 归因推导，是后续调参的主要依据。

关键符号：CI_WORKER_ARGS, MODEL_SPECS, get_model_spec

关键源码片段

experimental/sgl-router/tests/e2e/infra/model_specs.py

本 PR 唯一改动文件。新增模块级 `CI_WORKER_ARGS` (`--mem-fraction-static=0.6`、`--cuda-graph-backend-prefill=disabled`) 并挂到三个模型 spec 的 `worker_args`，是全套 e2e worker 显存口径的唯一来源；文件内的长注释保留了 OOM 归因推导，是后续调参的主要依据。

```
# experimental/sgl-router/tests/e2e/infra/model_specs.py

# 本套件里每个 worker 只为几条 8-1024 token 的 router 断言而存在，
# 引擎的默认服务配置在这里是反向优化：--mem-fraction-static=0.83 会把卡上
# 剩余显存几乎全部交给 KV 池，只留 free_memory_at_dist_init * (1 - mem_fraction_static)
# 的 slack（空闲 80 GB H100 上约 13 GB）同时覆盖激活与所有 CUDA graph；
# 而默认 chunked_prefill_size=8192 的激活工作集（pre_capture_activation_reserve_mb）
# 已吃掉其中大部分，58 个 num-token bucket 的 breakable prefill graph 无处捕获，
# 在 graph.capture_end() 里 CUDA OOM，worker 以 -9 退出，最终表现为与 router
# 转发逻辑无关的 e2e 失败。
CI_WORKER_ARGS: list[str] = [
    # 收紧 KV 池：把「吃满剩余显存」改为显式 0.6 比例，为激活与图捕获留出预算
    "--mem-fraction-static=0.6",
    # 本套件不需要 prefill 图；跳过它还顺带省掉每次 worker spawn 的捕获时间
    "--cuda-graph-backend-prefill=disabled",
]

MODEL_SPECS: dict[str, dict] = {
    "qwen3-0.6b": {
        "model": _resolve_model_path("Qwen/Qwen3-0.6B"),
        "memory_gb": 4,
        "tp": 1,
        "features": ["chat", "streaming"],
        # 三个条目共用同一份预算参数，避免逐条漂移
        "worker_args": CI_WORKER_ARGS,
    },
    "llama-1b": {
        "model": _resolve_model_path("meta-llama/Llama-3.2-1B-Instruct"),
        "memory_gb": 4,
        "tp": 1,
        "features": ["chat", "streaming"],
        "worker_args": CI_WORKER_ARGS,
    },
    "llama-8b": {
        "model": _resolve_model_path("meta-llama/Llama-3.1-8B-Instruct"),
        "memory_gb": 16,
        "tp": 1,
        "features": ["chat", "streaming"],
        "worker_args": CI_WORKER_ARGS,
    },
}

def get_model_spec(model_id: str) -> dict:
```

```
"""读侧契约不变: worker_args 是模块 docstring 早已声明的可选字段,
e2e 拉起 worker 时把它拼到 sglang.launch_server 命令行即可。
"""

if model_id not in MODEL_SPECS:
    raise KeyError(
        f"Unknown model: {model_id}. Available: {list(MODEL_SPECS.keys())}"
    )
return MODEL_SPECS[model_id]
```

评论区精华

本 PR 的 `review_comments_count` 与 `comments_count` 均为 0，作者 Kangyan-Zhou 同时是 merge 人，没有任何 review 交锋或线程可提炼。唯一有价值的「讨论」来自 PR body 的自我审查段落，可以视为作者预先写下的设计权衡：

“Note for a follow-up (not in this PR) — The engine-side issue is that KV sizing charges nothing for CUDA-graph capture: the $(1 - \text{mem_fraction_static})$ slack is consumed by the activation reserve alone whenever `chunked_prefill_size` is large relative to the card, so any small model at default settings is one capture away from this. Worth an issue against the memory profiler rather than papering over it in every test suite.”

“This gives the e2e worker specs a bounded memory budget and no prefill CUDA graph, which removes the failure class.”

作者还把二次崩溃 (`num_active_captures_ > 0 INTERNAL ASSERT FAILED ... markCaptureEnd called with no captures in progress`) 显式定性为清理逻辑在首次 capture 失败后重复调用 `capture_end()` 的次生现象，而不是根因——这是一个容易被误判的细节，值得读 PR body 时留意。

- 引擎侧 KV sizing 不为 CUDA graph capture 计费（根因不在本 PR 修）(design): 本 PR 只在测试侧绕开，明确倾向“针对 memory profiler 开 issue”，不采用在每个测试套件里逐一手工规避的做法；后续需要引擎侧为图捕获预留额度。
- 修复的验证方式：仅靠 tier-3 e2e 与参数声明比对，无独立单测 (testing): 作者接受“以 CI 作为回归证据”的验证标准，未新增测试文件；这使参数值 (0.6) 的边界正确性缺乏可回归的单测锚点。

风险与影响

- 风险：
 - e2e 覆盖削弱：--cuda-graph-backend-prefill=disabled 会让 router e2e 套件整体不再经过 breakable prefill CUDA graph 路径。若未来出现只在 prefill 图里暴露的回归（例如 padding 契约、num_token bucket 校验、input_embeds 刷新类问题，参见 PR#39574 修过的图填充计数与 prefill 图刷新缺陷），tier-3 将无法拦截。decode 图与 sampling 路径不受影响。
 - KV 池口径收紧的长期副作用：--mem-fraction-static 从 0.83 降到 0.6，池容量下降约三成。当前测试 prompt/ 输出都很短 (8-1024 token)，但一旦新增长上下文、大并发或

prefix-cache 压力型的 router e2e, 可能撞到池容量上限, 届时需要按模型覆盖或再调比例; 而参数被集中成单个常量, 覆盖时需要显式拆分, 容易遗漏。

2. 平台泛化未验证: 显存账本是在空闲 80 GB H100 上算的, PR body 也显示该 PR 没有 AMD ROCm run (No AMD PR run found for this commit)。在显存更小或非 NVIDIA 的 runner 上, 0.6 比例加禁用 prefill 图是否仍足以避免启动 OOM, 本 PR 没有证据。
3. 引擎侧根因仍存在: `python/sglang/srt/mem_cache/kv_cache_configurator.py:2193` 的 KV sizing 不为 CUDA graph capture 预留任何额度, 配合 `python/sglang/srt/runtime_context.py:1842` 的 `pre_capture_activation_reserve_mb`, 使「小模型 + 默认 chunked_prefill_size」在真实部署里也可能在 capture 阶段 OOM。本 PR 只在测试侧绕开, 属于作者自述的 papering over, 风险会转移到未来任何忘记设置 mem-fraction-static 的测试套件或用户部署上。
4. 一致性风险: 新增 spec 若忘记复用 `CI_WORKER_ARGS`, 会重新引入默认参数导致的口径不一致, 而目前没有任何断言或 lint 强制 `MODEL_SPECS` 条目必须携带 `worker_args`。- 影响: 对 CI/ 团队: 直接消除 `pr-test-ssl-router.yml` 中 tier-3 — e2e 的一类间歇性失败 (worker 启动期 -9 退出), 减少无关 PR 被误判为失败、以及反复重跑带来的机器时间消耗; 同时跳过 prefill 图捕获后每次 worker spawn 更快, 对「每个测试重启一次 worker」的套件是实打实的墙钟时间收益。

对开发流程: 所有读取 `MODEL_SPECS` 的 router e2e 用例的 worker 启动参数统一变化, 任何依赖 prefill CUDA graph 或在默认 KV 池容量下运行的用例行为都会改变; 这是隐性但全局的口径变更, 值得 e2e 用例作者知晓。

对生产 / 用户: 不涉及 `experimental/sgl-router` 的 Rust 转发逻辑, 也不涉及 SRT 引擎代码, 对线上推理路径零影响。间接价值是 PR body 给出的显存账本为「引擎应为 graph capture 单独预留内存」这一问题提供了可复现的分析素材。

对外部集群: `ROUTER_LOCAL_MODEL_PATH`、`memory_gb`、`tp` 等既有字段未动, 本地缓存模型与单卡拓扑假设保持不变。

- 风险标记: 禁用 prefill CUDA graph, KV 池容量下调, 引擎侧根因未修, 仅 CI 验证无单测, 测试夹具全局口径变更

关联脉络

- PR #39110 [Router] Real-GPU e2e coverage for storage-tier-aware cache routing (3/4): 与本 PR 同属 `experimental/sgl-router/tests/e2e` 基建目录 (`model_pool.py` / `test_hicache_storage_tiers.py`)。本 PR body 明确点名 `test_hicache_storage_tiers.py:96` 固定了 `--max-total-tokens`, 是套件里唯一从未触发该 OOM 的真实 GPU 测试, 等于用它的既有做法反向佐证了修法。
- PR #39004 [router] Resolve a wire protocol per worker at registration: 本 PR 列举的三次失败 run 中有两次发生在 `router-h2c-in-and-out` 与 `router-h2c-2-worker-protocol` 分支, 正是 h2c wire-protocol 这条工作线的分支; 说明该 OOM 是横跨多个 router 特性线的共性基建缺陷, 而非某个 PR 引入。
- PR #39015 [Router] Add the shutdown-drain configuration surface: 同期 router 系列的配置面 / 测试基建改动 (`config.rs` 与 `tests/proxy.rs`), 与本 PR 同处“先铺配置与测试

环境、再验证功能”的节奏，属于同一条 CI 稳定性脉络。