

PR #33888 完整报告

sgl-project/sglang

config: delete the dead `get_server_args()` bindings across the repo

合并时间: 2026-08-08 13:41

原文链接: <http://prhub.com.cn/sgl-project/sglang/pull/33888>

执行摘要

- 一句话: 清理 bag 迁移遗留的 `get_server_args()` 死绑定
- 推荐动作: 值得快速浏览, 重点看两点: 一是 `topk.py` “调用有副作用、绑定是死代码”的区分, 二是 `closure-aware AST` 扫描的死代码发现方法 (先断言行内容再删除)。该 PR 可作为大规模机械清理 + 语义保留的参考范式; 若团队后续还要继续清理 `get_server_args()` 残留, 建议补一个针对 `_eplb_remap_enabled` 发布探测语义的回归测试。

功能与动机

仓库此前把全局 `server_args` 读取迁移到 `get_exec()` / `get_spec()` / `get_parallel()` 等 namespace bag, 但若干文件留下了“赋值后从未读取”的 `server_args = get_server_args()` 绑定。这些死绑定既误导读者 (看起来像在使用全局配置), 又妨碍后续彻底移除旧的 `get_server_args()` 接口。PR body 明确定位为 “dead `get_server_args()` bindings that earlierbag-migrationflipsleftbehind”, 并通过 `closure-awareAST` 扫描证明删除后零残留。

实现拆解

1. 发现与核验: 对 `python/` 与 `test/` 做 `closure-aware AST` 扫描, 判定条件是“某名字在其作用域 (含嵌套函数) 内从未被加载”; 普通扫描会对闭包误报 (如 `triton_symm_mem_ag`), 因此必须感知嵌套作用域。每个删除点按精确行号执行, 且先断言行内容再删除。
2. 逐文件删除绑定与孤立 import: 涉及 `sampling_batch_info.from_schedule_batch`、`deepseek_v2` 的 `forward_normal` / `forward_normal_dual_stream` / `determine_num_fused_shared_experts` (3 处)、`forward_mla._skip_rope_for_dsa_tilelang_fused`、`minimax_m3_vl._determine_num_fused_shared_experts`、`deepseek_v4_memory_pool.__init__` / `get_ring_size` (2 处)、`flexkv_radix_cache` / `lmc_radix_cache` 的 `cache_finished_req`、`indexer_topk.__init__`、`loader._get_weights_iterator`、`routed_experts`、`inkling_sconv_backend`、`schedule_batch`。同步删除被孤立的 `get_server_args` import。
3. 保留 load-bearing 调用: `layers/moe/topk.py` 的 `_eplb_remap_enabled` 中, `get_server_args()` 裸调用承担“配置是否已发布”的探测职责——except `ValueError` 表示未发布、不做 EPLB remap, 因此只删赋值、调用保留。
4. 测试配套: `test_sampling_batch_info.py` 的 7 个 `@patch(...get_server_args)` mock 恰好打在已删除的绑定上; 生产代码读的是 `get_exec()` bag (测试 `setUp` 已注入并按测试覆盖

deterministic / custom-logit 开关) , 于是移除装饰器与 mock 参数。

5. 验证: 涉及子系统 (sampling、mem_cache、models、layers、managers) 单测全绿; 对 base commit 跑完整 16 分区 CPU 测试矩阵, 零新增失败; AST 复扫确认零残留。

关键文件:

- python/sglang/srt/models/deepseek_v2.py (模块 模型层; 类别 source; 类型 data-contract; 符号 forward_normal, forward_normal_dual_stream, determine_num_fused_shared_experts) : 本次清理覆盖面最大的模型核心文件, 删除 3 处热路径死绑定 (forward_normal、forward_normal_dual_stream、determine_num_fused_shared_experts) , 最能体现 bag 迁移后的读取方式。
- python/sglang/srt/models/deepseek_common/attention_forward_methods/forward_mla.py (模块 注意力前向; 类别 source; 类型 data-contract; 符号 _skip_rope_for_dsa_tilelang_fused) : MLA 注意力前向的公共文件, 删除 import 与 _skip_rope_for_dsa_tilelang_fused 内的死绑定, 是 attention 侧 bag 迁移的典型代表。
- test/registered/unit/sampling/test_sampling_batch_info.py (模块 采样模块; 类别 test; 类型 test-coverage; 符号 test_basic_construction, test_greedy_detection, test_logit_bias_construction, test_deterministic_seed) : 删除 7 个 @patch(...get_server_args) mock 与对应参数, 是本次“测试联动清理”的核心证据文件, 验证生产路径已完全转向 get_exec() bag。
- python/sglang/srt/layers/moe/topk.py (模块 MoE 路由; 类别 source; 类型 core-logic; 符号 _eplb_remap_enabled) : 唯一保留裸调用点的文件: _eplb_remap_enabled 中 get_server_args() 承担“配置是否已发布”的探测职责, 只删绑定不删调用, 语义最微妙。
- python/sglang/srt/mem_cache/deepseek_v4_memory_pool.py (模块 KV 缓存; 类别 source; 类型 dependency-wiring; 符号 DeepSeekV4SingleKVPool.init, get_ring_size) : 删除 2 处死绑定 (__init__与 get_ring_size) , 是 DSV4 统一 KV 池构建路径上的清理。
- python/sglang/srt/models/minimax_m3_vl.py (模块 模型层; 类别 source; 类型 data-contract; 符号 _determine_num_fused_shared_experts) : 多模态模型入口, 删除 import 与 _determine_num_fused_shared_experts 内的死绑定, 与 deepseek_v2 的融合判定清理同步。
- python/sglang/srt/sampling/sampling_batch_info.py (模块 采样模块; 类别 source; 类型 dependency-wiring; 符号 from_schedule_batch) : 采样批构造的核心入口, from_schedule_batch 的死绑定被删除, 且与测试文件 mock 清理形成联动。
- python/sglang/srt/mem_cache/storage/flexkv/flexkv_radix_cache.py (模块 KV 缓存; 类别 source; 类型 dependency-wiring; 符号 cache_finished_req) : FlexKV 存储路径的 cache_finished_req 删除函数内 import 与死绑定, 属于“commit-prefix 路径”清理。
- python/sglang/srt/mem_cache/storage/lmcache/lmc_radix_cache.py (模块 KV 缓存; 类别 source; 类型 dependency-wiring; 符号 cache_finished_req) : LMCache 存储路径的 cache_finished_req 同步清理, 与 FlexKV 形成对称修改。
- python/sglang/srt/state_capturer/indexer_topk.py (模块 状态捕获; 类别 source; 类型 dependency-wiring; 符号 init) : 状态捕获器初始化路径上的死绑定清理, 删除 __init__内的局部 import 与绑定。

- python/sglang/srt/model_loader/loader.py (模块 权重加载; 类别 source; 类型 data-contract; 符号 _get_weights_iterator) : 权重加载路径 (safetensors 分支) 的死绑定清理。
- python/sglang/srt/state_capturer/routed_experts.py (模块 状态捕获; 类别 source; 类型 entrypoint) : 路由专家状态捕获文件的死绑定清理。
- python/sglang/srt/layers/attention/linear/inkling_sconv_backend.py (模块 注意力后端; 类别 source; 类型 core-logic) : Inkling 卷积注意力后端文件, 单处死绑定清理。
- python/sglang/srt/managers/schedule_batch.py (模块 调度模块; 类别 source; 类型 core-logic) : 调度批处理核心文件, 单处死绑定清理, 属于调度子系统收尾。

关键符号: SamplingBatchInfo.from_schedule_batch, DeepseekV2MoE.forward_normal, DeepseekV2MoE.forward_normal_dual_stream, deepseek_v2.determine_num_fused_shared_experts, forward_mla._skip_rope_for_dsa_tilelang_fused, minimax_m3_vl._determine_num_fused_shared_experts, deepseek_v4_memory_pool.get_ring_size, flexkv_radix_cache.cache_finished_req, lmc_radix_cache.cache_finished_req, topk._eplb_remap_enabled, loader._get_weights_iterator

关键源码片段

python/sglang/srt/models/deepseek_v2.py

本次清理覆盖面最大的模型核心文件, 删除 3 处热路径死绑定 (forward_normal、forward_normal_dual_stream、determine_num_fused_shared_experts), 最能体现 bag 迁移后的读取方式。

```
# deepseek_v2.py — determine_num_fused_shared_experts 变更后头部
# 变更点: 删除 `server_args = get_server_args()` 死绑定;
# 后续 enable / disable 判定全部改读 get_exec().moe.* 与 get_parallel()
# 等 namespace bag, 该局部变量在旧版本中从未被使用。
def determine_num_fused_shared_experts(
    self, architecture: str = "DeepseekV3ForCausalLM"
):
    self.num_fused_shared_experts = 0

    # 共享专家融合的总开关来自 get_exec().moe 的 namespace bag
    if get_exec().moe.disable_shared_experts_fusion:
        return

    disable_reason = None
    if get_exec().moe.enforce_shared_experts_fusion:
        pass
    elif is_sbo_enabled() or is_tbo_enabled():
        disable_reason = "SBO/TBO enabled: incompatible with fusing shared expert into MoE kernel."
    elif is_deepep_class_backend():
```

```

disable_reason = "DeepEP: fusion off by default (use --enforce-shared-experts-fusion to
enable)."
elif (
    self.config.architectures[0] != architecture
    # n_routed_experts 白名单: 256 -> DeepSeek-V3 / R1;
    # 384 -> Kimi-K2.5, 仅限 Quark MXFP4 checkpoint (预融合),
    # 标准 compressed-tensors checkpoint 的共享专家是松散存储的,
    # 走融合路径会静默加载错误, 因此被排除。
    or self.config.n_routed_experts not in (256, 384)
    or self.config.n_shared_experts != 1
    or (
        self.config.n_routed_experts == 384
        and (
            self.quant_config is None
            or self.quant_config.get_name() != "quark"
        )
    )
):
    disable_reason = "Config does not support fused shared expert(s)."
# ... 后续其余 disable_reason 分支与最终决策逻辑保持不变

```

python/sglang/srt/models/deepseek_common/attention_forward_methods/forward_mla.py

MLA 注意力前向的公共文件，删除 import 与 `_skip_rope_for_dsa_tilelang_fused` 内的死绑定，是 attention 侧 bag 迁移的典型代表。

```

# forward_mla.py — _skip_rope_for_dsa_tilelang_fused 变更后
# 变更前函数体内有一行从未被读取的 `server_args = get_server_args()` 绑定，
# 本次连同 import 列表中的 get_server_args 一起删除。
# 本函数真正依赖的运行时配置来自 get_exec() 返回的 namespace bag
# （由 runtime_context 在 runner 启动时注入），与全局 server_args 无关。
def _skip_rope_for_dsa_tilelang_fused(self: DeepseekV2AttentionMLA) -> bool:
    """
    Check if we should skip rope and use fused rope+cache path for
    TileLang DSA on gfx95.
    """
    # 跳过 rope 的条件只看 gfx95 后端开关、当前 attention backend,
    # 以及 get_exec().kernel 中 dsa 的 decode / prefill 后端是否为 tilelang。
    return (
        _use_aiter_gfx95
        and self.current_attention_backend in ("dsa", "nsa")
        and (
            get_exec().kernel.dsa_decode_backend == "tilelang"
            or get_exec().kernel.dsa_prefill_backend == "tilelang"
        )
    )

```

test/registered/unit/sampling/test_sampling_batch_info.py

删除 7 个 `@patch(...get_server_args)` mock 与对应参数，是本次“测试联动清理”的核心证据文件，验证生产路径已完全转向 `get_exec()` bag。

```
# test_sampling_batch_info.py — 测试去掉 @patch 装饰器
# 变更前这 7 个测试都带 `@patch("...sampling_batch_info.get_server_args")`
# 装饰器与 mock_server_args 参数；而生产路径 from_schedule_batch 早已改读
# get_exec() 返回的 bag，这些 mock 恰好打在本次删除的死绑定上，
# 属于无用 mock，一并移除。测试所需开关由 setUp 注入的 _exec_ns 按测试覆盖。
def test_deterministic_seed(self):
    """Test that explicit seed=123 is kept and missing seed defaults to 42."""
    self._exec_ns.deterministic.enable_deterministic_inference = True

    reqs = [self._make_req(seed=123), self._make_req(seed=None)]
    batch = MagicMock()
    batch.reqs = reqs
    batch.device = DEVICE
    info = SamplingBatchInfo.from_schedule_batch(batch, VOCAB_SIZE)
    self.assertIsNotNone(info.sampling_seed)
    self.assertEqual(info.sampling_seed[0].item(), 123)
    self.assertEqual(info.sampling_seed[1].item(), 42) # 缺省 seed 默认值
```

评论区精华

diff 上无行内评论，两轮自审 (reviewer: ch-wan) 均给出 LGTM 结论。核心确认点：(1) 每个被删绑定都是“赋值后从未加载”，剩余控制流已读 `get_exec()` / `get_spec()` / `get_model()` / `get_schedule()` bag；(2) `_eplb_remap_enabled` 是唯一有副作用的调用点，保留裸调用作为 `except ValueError` 下的发布探测；(3) `test_sampling_batch_info` 的 `setUp` 已注入 `get_exec()` bag，去掉 7 个 mock 安全。两轮审核均为 0 bug / 0 suggestion / 0 nit，作为 commit stack 的中间 commit 合入。

- 删除范围与正确性核查 (design): 0 bug / 0 suggestion / 0 nit, LGTM, 可作为 commit stack 中间 commit 合入。
- `topk.py` 发布探测调用保留 (correctness): 保留裸调用作为有副作用的生效路径，两轮自审均确认无遗留消费者被误删。
- 测试 mock 移除安全性 (testing): 确认移除后相关子系统单测全绿，16 分区 CPU 矩阵零新增失败。

风险与影响

- 风险：行为风险低：所有删除均被 AST 证明为死代码，纯删除、无逻辑改写；但改动横跨 14 个文件、触及 `deepseek_v2.py` 的 `forward_normal` / `forward_normal_dual_stream` 等热路径函数，与并行迁移 PR 合并时冲突面较大。`topk.py` 的 `_eplb_remap_enabled` 依赖 `get_server_args()` 对未发布配置抛 `ValueError` 这一契约；若未来该函数不再抛错，EPLB remap 判定会静默改变。当前调用保留，风险可控，但值得在后续退役 `get_server_args()` 时专项测试。测试清理依赖 `setUp` 注入的 `_exec_ns` 与生产路径 `get_exec()` 的等价性，单测全绿与 16 分区 CPU 矩阵零新增失败已覆盖主要风险。

- 影响：对用户无行为影响（纯死代码删除）。对系统而言，消除了“看起来在读配置、实际是死代码”的误导，降低后续维护者误用 `server_args` 绑定或拷贝该模式的可能；测试文件去掉 7 个 mock 后更简洁，也避免了 mock 对象与被删绑定的隐性耦合。对团队而言，这是 `get_server_args()` 旧式全局配置读取的系统性退役的一部分，为后续收紧 `runtime_context` 接口、统一 namespace bag 读取铺路。
- 风险标记：纯删除、行为零变更，`topk.py` 探测依赖 `except ValueError` 契约，跨 14 文件合并冲突面大，热路径函数改动

关联脉络

- PR #33889 moe: the shared-experts-fusion decision is a per-runner value the loader installs: 与本次 PR 同处 commit stack (PR body 称“middle commit of the stack”)，共享 `deepseek_v2.py` / `minimax_m3_vl.py` 等文件；本次清理的正是这类共享专家融合重构遗留的死绑定。
- PR #33925 config: route DCP topology reads through `get_parallel()`: 同属“配置读取收敛到 namespace bag”的迁移家族，方向一致：逐步撤销遗留的 `server_args` 全局读取，说明 `get_server_args()` 正在系统性退役。