PR #27977 完整报告

sgl-project/sglang

[Spec] Remove the dead spec V1 scheduler paths

合并时间：2026-06-12 09:31

原文链接：http://prhub.com.cn/sgl-project/sglang/pull/27977

执行摘要

- 一句话：移除 Spec V1 调度器死代码，统一 V2 路径
- 推荐动作：值得精读，特别是通过条件逆反消除 is_spec_v2 的方法，以及 deprecation 策略。整体设计干净，适合作为清理技术债务的范例。

功能与动机

依据 PR 描述，所有 speculative 算法（EAGLE、NGram、DFLASH、STANDALONE）均通过 supports_spec_v2() 返回 True，使得 batch.is_spec_v2 退化为 'is speculative'，V1 调度路径不可达。为清理技术债务，移除这些死代码。

实现拆解

1. 清理 Eagle 验证输入和输出类：在 python/sglang/srt/speculative/eagle_info.py 中，删除了约 220 行的 EagleVerifyInput.verify 方法、EagleVerifyOutput 类以及工厂函数 create_idle，同时清理了不再使用的导入（如 torch.nn.functional、is_cuda 等）。
2. 移除 V1 Logprob 计算路径：在 python/sglang/srt/layers/utils/logprob.py 中，删除了函数 add_output_logprobs_for_spec_v1（约 120 行）及其类型依赖（EagleVerifyOutput, NgramVerifyInput），该功能已被 compute_spec_v2_logprobs 取代。
3. 消除 is_spec_v2 区分：在 python/sglang/srt/managers/schedule_batch.py 中，删除了 ScheduleBatch.is_spec_v2 属性和 new_tokens_required_next_decode 中的 V1 token 估算分支，所有 speculative batch 统一走 _new_tokens_required_next_decode_spec_v2。同时删除了 filter_batch 方法的 v1_spec_info_filtered 参数。
4. 简化 scheduler 和 batch_result_processor 的分支逻辑：在 scheduler.py 中，多处 is_spec_v2 判断替换为 not spec_algorithm.is_none()，移除 V1 专用的 relay 和 token-estimate 代码。在 batch_result_processor.py 中，移除了 is_spec_v1 检测和对应的 continue 分支，统一处理 V2 和非 spec 场景。
5. 调整插件注册的行为：在 python/sglang/srt/speculative/spec_registry.py 中，移除了 CustomSpecAlgo.supports_spec_v2 方法（原来委托给 supports_overlap），改为所有非 None 算法都返回 True。create_worker 中为 supports_overlap=False 的插件添加 deprecation 警告（日志）并继续以同步 V2 模式运行。
6. 测试配套：更新了 test/registered/unit/spec/test_spec_registry.py，调整 test_supports_spec_v2_follows_supports_overlap 为测试新行为，新增

test_supports_overlap_false_warns_deprecation 检查警告。

关键文件：

- python/sglang/srt/speculative/eagle_info.py（模块 推测解码；类别 source；类型 dependency-wiring；符号 verify, EagleVerifyOutput, create_idle）：删除 V1 核心 verify 方法和 EagleVerifyOutput 类，削除 525 行，是本次清理的最大块。
- python/sglang/srt/layers/utils/logprob.py（模块 LogProb 处理；类别 source；类型 core-logic；符号 add_output_logprobs_for_spec_v1）：删除 add_output_logprobs_for_spec_v1 函数（118 行），V1 logprob 处理路径彻底清除。
- python/sglang/srt/managers/schedule_batch.py（模块 调度批处理；类别 source；类型 core-logic；符号 is_spec_v2, new_tokens_required_next_decode, filter_batch）：删除 is_spec_v2 属性和 V1 token 估算分支，所有 speculative batch 统一走 V2 路径。
- python/sglang/srt/managers/scheduler_components/batch_result_processor.py（模块 结果处理器；类别 source；类型 core-logic）：移除 is_spec_v1 分支，统一 post-processing 逻辑。
- python/sglang/srt/managers/scheduler.py（模块 调度器；类别 source；类型 core-logic）：多处用 not spec_algorithm.is_none() 替换 is_spec_v2，移除 V1 relay/token-estimate 代码。
- python/sglang/srt/speculative/spec_registry.py（模块 推测注册；类别 source；类型 dependency-wiring；符号 supports_spec_v2, create_worker）：修改 CustomSpecAlgo.supports_spec_v2 逻辑，添加 deprecation warning。
- test/registered/unit/spec/test_spec_registry.py（模块 推测注册测试；类别 test；类型 test-coverage；符号 test_supports_spec_v2_follows_supports_overlap, test_supports_overlap_false_warns_deprecation）：测试覆盖新增 deprecation 警告行为，调整原有测试。

关键符号：EagleVerifyInput.verify, add_output_logprobs_for_spec_v1, ScheduleBatch.is_spec_v2, CustomSpecAlgo.supports_spec_v2, CustomSpecAlgo.create_worker, ScheduleBatch.new_tokens_required_next_decode, BatchResultProcessor.process_batch_result_decode, BatchResultProcessor._normalize_decode_outputs, ScheduleBatch.filter_batch

关键源码片段

python/sglang/srt/layers/utils/logprob.py

删除 add_output_logprobs_for_spec_v1 函数（118 行），V1 logprob 处理路径彻底清除。

```
# 此函数替代了被删除的 add_output_logprobs_for_spec_v1 (spec V1)。
# 在所有推测解码场景（V2）下计算接受 token 的 logprobs。
def compute_spec_v2_logprobs(
    batch,
    logits_output,
    predict: torch.Tensor,
    accept_index: torch.Tensor,
    speculative_num_steps: int,
```

```
):
    """Compute logprobs for accepted tokens after spec v2 verify sampling.

    Gathers logits at accepted positions, applies log_softmax (temperature-scaled
    if not greedy), and populates logits_output.next_token_logprobs (plus optional
    top-k / token-ids logprobs) so they flow through copy_to_cpu().
    """
    bs = len(batch.seq_lens)
    max_accept = speculative_num_steps + 1
    device = predict.device

    flat_accept_idx = accept_index.long().reshape(-1)
    gathered_logits = logits_output.next_token_logits[flat_accept_idx]

    # 如果全部 greedy 或设置了 SGLANG_RETURN_ORIGINAL_LOGPROB，则不做温度缩放
    if batch.sampling_info.is_all_greedy or envs.SGLANG_RETURN_ORIGINAL_LOGPROB.get():
        gathered_logprobs = torch.nn.functional.log_softmax(gathered_logits, dim=-1)
    else:
        temperatures = torch.repeat_interleave(
            batch.sampling_info.temperatures,
            max_accept,
            dim=0,
        )
        gathered_logprobs = torch.nn.functional.log_softmax(
            gathered_logits / temperatures, dim=-1
        )
    # ... 后续写入 logits_output.next_token_logprobs 等属性
```

python/sglang/srt/managers/schedule_batch.py

删除 is_spec_v2 属性和 V1 token 估算分支，所有 speculative batch 统一走 V2 路径。

```
# 删除 is_spec_v2 属性和 V1 分支后的方法
# 所有 speculative 算法统一走 V2 估算
    def new_tokens_required_next_decode(
        self, selected_indices: Optional[List[int]] = None
    ):
        page_size = self.token_to_kv_pool_allocator.page_size
        requests = (
            self.reqs
            if selected_indices is None
            else [self.reqs[i] for i in selected_indices]
        )

        if self.spec_algorithm.is_none():
            new_pages = sum(1 for r in requests if r.kv_committed_len % page_size == 0)
            return new_pages * page_size

        # 不再检查 is_spec_v2，直接调用 V2 精确估算
        return self._new_tokens_required_next_decode_spec_v2(requests, page_size)
```

is_spec_v2 属性已完全移除

评论区精华

该 PR 没有收到 review 评论，唯一评论来自 gemini-code-assist 的配额提示，与技术内容无关。

- 暂无高价值评论线程

风险与影响

- 风险：主要风险是插件兼容性：外部注册了 supports_overlap=False 的算法之前运行在 V1 schema，现在运行在同步 V2 schema。由于行为相近（同步），风险较低，且 deprecation 警告会通知开发者。另一个风险是回归，但本 PR 已通过单元测试（bookkeeping-ownership、spec-registry）和 E2E 测试（DFLASH sync/overlap、EAGLE constrained decoding、非 spec return_logprob smoke），验证充分。无性能风险，因为减少了分支。
- 影响：对用户透明（无接口变更），对内部开发者：代码量减少 697 行，维护性提升；插件开发者需迁移至 V2 以支持 overlap 调度。
- 风险标记：插件兼容性警告，低风险移除死代码

关联脉络

- PR #27964 [Spec] Retire Spec V1: 先行移除了 SGLANG_ENABLE_SPEC_V2 环境变量，本 PR 在此基础上进一步清除代码路径。
- PR #27959 [Spec] Remove the DFLASH V1 worker path: 移除了 DFLASH V1 worker，本 PR 清理了调度器中残留的 V1 通用路径。
- PR #27950 [Spec] Fold the DFLASH worker base into DFlashWorkerV2 on BaseSpecWorker: 将 DFLASH worker 基类合并到 V2，为本次清理奠定了基础。