PR #33403 完整报告

sgl-project/sglang

[Scheduler] Honor explicit min-free-slots thresholds

合并时间：2026-08-05 16:44

原文链接：http://prhub.com.cn/sgl-project/sglang/pull/33403

执行摘要

- 一句话：修复 min-free-slots 显式阈值被封顶，非 DFlash 可调
- 推荐动作：值得精读：它演示了一个小型调度器准入参数如何在“用户显式配置 vs 自动启发式”之间做优先级设计，并伴随 PP 约束收紧。重点看 resolve_min_free_slots 的新分支顺序和 server_args 的启动校验，以及测试如何覆盖边界（8 以下集群、max_running 上限、1 禁用）。

功能与动机

PR body 明确：显式阈值“currently capped to the DFlash auto-default formula, whose maximum is four”，导致非 DFlash 用户“deliberately accumulating a larger admission batch”时配置被忽略。该 PR 旨在让用户在非 DFlash 工作负载上按需放行更大的准入批处理。

实现拆解

1. 准入阈值解析重构（python/sglang/srt/managers/min_free_slots_delayer.py） - resolve_min_free_slots 改为两段式判断：显式 user_value 存在时直接返回 min(user_value, max_running_requests)（<=1 返回 None 禁用）；未设置时仅 DFlash 家族且 max_running_requests >= 8 才使用 min(4, max(2, (max_running_requests + 5) // 6))。原“用户值一律被公式封顶、小集群一律禁用”的约束被移除。 - 原因：让非 DFlash 工作负载能按业务需要累积更大批，同时不破坏 DFlash 的默认启发式行为。 - 影响：<8 的集群守卫只作用于自动默认，显式值在小型集群下仍生效。
2. 启动参数与兼容性校验（python/sglang/srt/server_args.py） - 更新 min_free_slots_delay 的 help 文本：显式值优先、受 max_running_requests 上限、1 禁用。 - 在 check_server_args 中新增断言：pp_size > 1 时不允许设置该参数，因为流水线并行下可分配槽位数受 pp-max-micro-batch-size 限制，阈值可能永远无法满足。 - 影响：PP 部署若显式设置该参数会在启动阶段报错，属一次性破坏性校验。
3. 单元测试同步（test/registered/scheduler/test_min_free_slots_delayer.py） - 删除旧用例 test_small_cluster_disables 和 test_caps_to_formula，新增 test_explicit_value_survives_small_cluster、test_non_dflash_uses_explicit_value、test_explicit_value_is_capped_to_max_running_requests、test_explicit_one_disables_dflash_default 等，覆盖显式值优先级、max_running_requests 封顶、1 禁用、小集群守卫仅作用于自动默认等边界。 - CI 回归：通过 /rerun-test 重跑 test_min_free_slots_delayer.py（ubuntu-latest）与

test_dflash.py（1-gpu-5090），均通过。

4. 无附加配置 /schema/ 部署改动：serving 启动参数不再受公式隐藏限制。

关键文件：

- python/sglang/srt/managers/min_free_slots_delayer.py（模块 准入控制；类别 source；类型 core-logic；符号 resolve_min_free_slots, MinFreeSlotsDelayer.should_delay）：核心准入阈值解析逻辑重写：显式值优先并受 max_running_requests 上限，DFlash 自动默认保留公式，<8 守卫仅作用于自动默认。
- test/registered/scheduler/test_min_free_slots_delayer.py（模块 单元测试；类别 test；类型 test-coverage；符号 test_small_cluster_disables, test_explicit_value_survives_small_cluster, test_caps_to_formula, test_respects_smaller_user_value）：单元测试同步新优先级，覆盖显式值在 <8 集群、max_running 上限和 1 禁用等边界，删除旧的 formula 封顶用例。
- python/sglang/srt/server_args.py（模块 服务参数；类别 source；类型 configuration；符号 ServerArgs.check_server_args）：CLI 帮助文本反映新行为，并在 check_server_args 中禁止 PP 下使用该参数。

关键符号：resolve_min_free_slots, MinFreeSlotsDelayer.should_delay, ServerArgs.check_server_args, test_explicit_value_survives_small_cluster, test_non_dflash_uses_explicit_value, test_explicit_value_is_capped_to_max_running_requests, test_explicit_one_disables_dflash_default

关键源码片段

python/sglang/srt/managers/min_free_slots_delayer.py

核心准入阈值解析逻辑重写：显式值优先并受 max_running_requests 上限，DFlash 自动默认保留公式，<8 守卫仅作用于自动默认。

```
from typing import Optional


def resolve_min_free_slots(
    user_value: Optional[int],
    max_running_requests: int,
    is_dflash_family: bool = False,
) -> Optional[int]:
    """解析 min-free-slots 阈值（None 表示禁用）。

    优先级约定：
    1. 显式 user_value 始终优先，按 max_running_requests 封顶（<=1 视为禁用）；
    2. 未设置且是 DFlash 家族时，fallback 到旧公式自动启用（<8 集群禁用）；
    3. 其余情况保持禁用。
    注意：<8 守卫只属于自动默认，不再压制显式值。
    """
    max_running_requests = max(0, int(max_running_requests))

    if user_value is not None:
```

```
        # 显式值只受 max_running_requests 限制，不再被 DFlash 公式压缩
        threshold = min(user_value, max_running_requests)
        return threshold if threshold > 1 else None

    if is_dflash_family and max_running_requests >= 8:
        # 保留 DFlash 无配置时的历史行为
        return min(4, max(2, (max_running_requests + 5) // 6))

    return None


class MinFreeSlotsDelayer:
    """当运行槽位不足时推迟新 prefill 准入，凑成一批再放行。

    典型场景：DFlash 的 draft prefill 每次准入成本高；阈值按 rank 本地计算，
    每个 DP rank 独立判断，避免一个拥挤 rank 拖住空闲 rank。
    """

    def __init__(self, min_free_slots: int):
        self._min_free_slots = min_free_slots

    def should_delay(self, *, running_bs: int, num_allocatable_reqs: int) -> bool:
        # 无 running 请求时不延迟；空闲 slot 不足阈值时才延迟
        return running_bs > 0 and num_allocatable_reqs < self._min_free_slots
```

python/sglang/srt/server_args.py

CLI 帮助文本反映新行为，并在 check_server_args 中禁止 PP 下使用该参数。

```
if self.pp_size > 1:
    assert (
        self.disable_overlap_schedule and self.speculative_algorithm is None
    ), "Pipeline parallelism is not compatible with overlap schedule, speculative decoding"
    # 流水线并行下 per-microbatch 槽位受 pp-max-micro-batch-size 限制，
    # 阈值可能永远达不到，因此启动阶段直接拒绝该组合
    assert self.min_free_slots_delay is None, (
        "--min-free-slots-delay is not supported with pipeline "
        "parallelism: allocatable slots per microbatch are bounded by "
        "pp-max-micro-batch-size, so the threshold may never be reached"
    )
```

评论区精华

该 PR 没有正式的 review 评论；关键讨论集中在 issue 评论与提交演进中：

- hnyls2002 通过 /rerun-test test_min_free_slots_delayer.py 和 /rerun-test test_dflash.py 触发回归，github-actions 反馈两个重跑均通过。
- 首次重跑被 github-actions 拦截："Your PR is diverged relative to required base commit cdff33d"，要求先 rebase 到最新 main，最终通过合并 main 的提交完成。

- 设计收敛体现于提交历史：从“Make min-free-slots override explicit”到“Keep DFlash min-free-slots automatic”，再到“scope dflash guards to auto-default; trim test comments”，最终确定“显式值优先、公式只做 DFlash 默认、小集群守卫不压制显式值”。
- CI 重跑与 rebase 需求 (testing): 重跑均通过（ubuntu-latest 与 1-gpu-5090），最终通过合并 main 的提交完成 rebase。
- 显式阈值与 DFlash 自动默认的优先级设计 (design): 合并策略：显式值始终优先；公式只做 DFlash 默认；<8 guard 不压制显式值；PP 下禁止使用该参数。

风险与影响

- 风险：
  - 调度准入行为变化：非 DFlash + 显式阈值时，较大阈值会让 prefill 等待更久以攒批，可能抬高 TTFT；PR body 明确未做基准测试，建议后续补充吞吐与首 token 延迟对比。
  - 小集群行为变化：<8 集群下显式阈值从“禁用”变为“生效”，可能改变既有服务的准入节奏。
  - PP 兼容性破坏：check_server_args 新增断言会让 pipeline parallelism 下设置该参数的服务启动失败，升级前需检查配置。
  - DFlash 回归面：未设置时行为不变，但显式值现在可大于 4；test_dflash.py 回归通过，风险可控。
- 影响：
  - 用户：非 DFlash 部署可通过 --min-free-slots-delay 控制准入批处理规模；显式设置 1 可关闭 DFlash 自动默认；PP 下必须移除该参数，否则启动失败。
  - 系统：准入调度逻辑变更，影响 prefill 延迟与批处理聚合；改动集中在 resolve_min_free_slots，MinFreeSlotsDelayer 本体决策逻辑未变。
  - 团队：CLI 文档与单元测试同步更新，行为契约发生变化，建议补充性能基准测试。
  - 风险标记：调度准入门控变更，缺少性能基准，PP 配置兼容性破坏，显式阈值改变小集群行为

关联脉络

- PR #33545 Allow optimistic prefill with L2 hierarchical cache and write-back policy: 同为 scheduler 准入策略调整（prefill 准入与缓存协同），与本 PR 的准入延迟逻辑相关。
- PR #33537 Multiple flexibility fixes for DP attention: 修复调度与解码图相关灵活性，涉及 forward_batch_info 与 decode runner，与本 PR 同处调度 / 准入链路。