

PR #29784 完整报告

sgl-project/sglang

[AMD][DI][CI] 2/N Add DSV4 DP8/EP8 and MTP MI355X 1P1D nightly recipes

合并时间: 2026-07-01 23:46

原文链接: <http://prhub.com.cn/sgl-project/sglang/pull/29784>

执行摘要

- 一句话: 为 AMD MI355X 添加 DSV4 DP8/EP8 和 MTP 夜间测试配置
- 推荐动作: 值得精读, 以理解如何为 AMD 平台扩展复杂的分离式测试拓扑。重点关注 `launch_mi355x.sh` 的解析逻辑和 `recipe` 结构, 可作为类似 CI 配置的模板。同时也需留意上游 `get_cpu_copy` 的修复进展, 以便及时解除并发限制。

功能与动机

PR 标题为 '[AMD][DI][CI] 2/N Add DSV4 DP8/EP8 and MTP MI355X 1P1D nightly recipes'。根据 body, 目的是扩展 AMD MI355X 的夜间测试, 覆盖 DeepSeek-V4 的多种并行拓扑和推测解码, 以尽早发现回归问题。

实现拆解

1. 新增 `recipe` 文件: 在 `scripts/ci/slurm/recipes/` 下按模型精度和拓扑创建 12 个 YAML 文件。每个文件定义 `backend.sglang_config` 中的 TP/EP/DP 大小、runtime 参数 (镜像、IB 设备、端口等)、`ben th` 配置 (并发序列、正确性门控), 以及可选的 `mtp` 段 (启用 MTP 时设置 `num_steps/eagle_topk/num_draft_tokens`)。基础变体 (无 MTP、EP1/DP1) 沿用现有配置。
2. 更新 `nightly-configs.yaml`: 在 `scripts/ci/slurm/nightly-configs.yaml` 中为每个模型精度组合新增三个配置条目 (`-mtp`、`-dp8ep8`、`-dp8ep8-mtp`), 指向对应的 `recipe` 文件。
3. 增强 `launch_mi355x.sh`: 修改启动脚本以解析 `recipe` 中的 `expert-parallel-size`、`data-parallel-size` 和 `mtp` 节, 并转换为 `--ep-size`、`--enable-dp-attention` `--dp-size`、`--speculative-*` 等命令行参数。基础 `recipe` 产生空额外标志, 保持向后兼容。
4. 处理已知限制: 在 `dsv4pro-fp8-mtp` 的 `recipe` 中通过注释说明并发 256 会因上游 `NotImplementedError` (`get_cpu_copy` 未实现) 崩溃, 因此将 `concurrencyes` 上限设为 128。同时尝试调高 `swa_full_tokens_ratio` 无效。
5. 排除坏节点: 在 `launch_mi355x.sh` 的 `salloc` 阶段添加节点排除逻辑, 屏蔽有 RDMA 驱动 ABI 不匹配的节点 (如 `mia1-p01-g20`)。
6. 日志收集: 末次 `commit` 修改日志脚本, 将 MI355X 分离式测试的日志打包到结果目录。

关键文件:

- `scripts/ci/slurm/nightly-configs.yaml` (模块 CI 配置; 类别 `infra`; 类型 `infrastructure`): 核心聚合配置文件, 新增 12 个测试条目定义每个变体的模型、精度、搜索空间和 `recipe`

引用。

- `scripts/ci/slurm/recipes/mi355x-fp8/dsv4pro/1k1k/1p1d-mtp.yaml` (模块 CI 配置; 类别 infra; 类型 infrastructure) : 代表 MTP 变体的 recipe, 包含详细的并发限制注释和上游未实现问题的 workaround, 是测试配置决策的典型示例。
- `scripts/ci/slurm/launch_mi355x.sh` (模块 启动脚本; 类别 infra; 类型 infrastructure) : 启动脚本修改, 新增对 `expert-parallel-size`、`data-parallel-size` 和 `mtp` 段的解析, 转换为启动参数, 是功能实现的入口。

关键符号: 未识别

关键源码片段

`scripts/ci/slurm/recipes/mi355x-fp8/dsv4pro/1k1k/1p1d-mtp.yaml`

代表 MTP 变体的 recipe, 包含详细的并发限制注释和上游未实现问题的 workaround, 是测试配置决策的典型示例。

```
# MI355X DeepSeek-V4-Pro-FP8 2 节点 1P1D 分离式 recipe — TP8 + MTP
# 注意: 并发 256 因上游 get_cpu_copy 未实现而排除
```

```
resources:
```

```
  prefill_workers: 1
  decode_workers: 1
```

```
backend:
```

```
  sglang_config:
    prefill:
      tensor-parallel-size: 8
      expert-parallel-size: 1
      data-parallel-size: 1
    decode:
      tensor-parallel-size: 8
      expert-parallel-size: 1
      data-parallel-size: 1
```

```
runtime:
```

```
  image: lmsysorg/sglang-rocm:v0.5.13.post1-rocm720-mi35x-20260623
  attention_backend: dsv4
  ib_devices: rdma0,rdma1,rdma2,rdma3
  prefill_port: 30025
  decode_port: 30026
  # ... ( 其他运行时参数 )
  swa_full_tokens_ratio: 0.1 # 曾尝试调至 0.2/0.3 但无法避免 conc256 crash
```

```
mtp: # MTP / EAGLE 推测解码配置
```

```
  enabled: true
  num_steps: 3
  eagle_topk: 1
  num_draft_tokens: 4
```

bench:

```
# 限制并发上限: conc256 会触发 disagg-decode SWA hybrid pool 的 get_cpu_copy 未实现错误
concurrencyes: [1, 8, 16, 32, 64, 128] # 注意不含 256
num_prompts_factor: 4
random_range_ratio: 1.0
accuracy:
  enabled: true
  num_shots: 8
  num_questions: 1319
  threshold: 0.91
```

评论区精华

从 PR body 和评论区提取的关键讨论:

- 并发限制决策: bingxche 在评论中报告了测试失败 (设置 swa_full_tokens_ratio 为 0.2 和 0.3 均失败), 最终作者决定将 dsv4pro-fp8-mtp 的并发上限限制在 128。PR body 详细解释了根本原因: disagg-decode SWA hybrid KV pool 填满后, scheduler 回收请求时调用 offload_kv_cache 的 get_cpu_copy, 而 SWA hybrid pool 的 get_cpu_copy 是未实现的 NotImplementedError。
- 上游依赖: recipe 注释中明确标注该限制, 并注明需要上游实现 get_cpu_copy 才能解除。
- 无设计争议: 讨论中没有架构层面的争议, 主要围绕实际运行中的问题定位与 workaround。
 - swa_full_tokens_ratio 调优与并发限制 (correctness): 接受 conc128 限制, 等待上游修复 get_cpu_copy。
 - 节点排除与 RDMA 驱动问题 (infra): 在 salloc 阶段添加节点排除逻辑, 避免调度到坏节点。

风险与影响

- 风险: 风险较低, 因为修改的是 CI 测试配置, 不影响产品代码。但需注意以下潜在风险:
 - 固化限制: dsv4pro-fp8-mtp 限制在 conc128 可能成为硬性约束, 若上游修复 get_cpu_copy, 需手动更新 recipe 以恢复全并发覆盖。
 - 节点排除误判: 节点排除逻辑基于已知问题节点, 若未来节点恢复或新节点出现类似问题, 可能需要更新排除列表。
 - 配置扩散: 新增 12 个配置条目增加了 CI 配置的复杂度, 但结构清晰风险可控。
 - 影响: 对用户无直接影响。对 AMD MI355X 平台的影响是扩展了夜间测试覆盖, 有助于捕获 DeepSeek-V4 在不同并行策略下的回归。对团队的影响是增加了 CI 配置文件的维护量, 但结构清晰且模板化, 降低了长期维护成本。影响程度中等, 仅限于 CI 环境。
- 风险标记: 依赖上游修复, 固化并发限制, 节点排除可能过时

关联脉络

- PR #29290 [AMD] Cover DeepSeek-R1 MXFP4 TP4 MTP nightly CI: 同为 AMD 平台添加 DeepSeek 模型夜间测试的 PR, 为本次扩展提供了 CI 基础设施的先例。

- PR #29726 [AMD] Rebalance stage-c-large-8-gpu-mi35x partitions to fix 60-min timeout: 同是 AMD MI355X 相关的 CI 调整，涉及分区和超时问题，与本 PR 的 CI 稳定性改进方向一致。