

PR #31500 完整报告

sgl-project/sglang

[AMD][DI][CI] 5/N Add DSV4 wide-EP16 4-node 2P1D nightly recipes

合并时间: 2026-08-04 13:08

原文链接: <http://prhub.com.cn/sgl-project/sglang/pull/31500>

执行摘要

- 一句话: DSV4 宽 EP16 4 节点 2P1D 夜间配方新增
- 推荐动作: 值得精读, 尤其适合关注 AMD 分离式部署与 RDMA/RoCE 调优的读者。重点学习: ① 2P1D 拓扑中窄 prefill + 宽 decode 的设计取舍; ② 硬件约束 (ionic 4 GiB MR 上限、GID 作用域、RDMA ABI) 如何转化为可复现的配方参数; ③ 字节级非回归门控思路——所有新逻辑仅在 `nodes-per-engine > 1` 时生效, 保证既有 EP8 配方逐字节不变。注意最终合并 diff 仅为配方与配置, launcher 能力需结合同系列其他 PR 阅读。

功能与动机

PR body 将本 PR 定位为 "Phase-4 of the AMD disaggregation nightly": 为 mi355x 集群新增 DSV4 Wide-EP16 腿 (no-MTP + MTP), 目标是在 4 节点上运行 2P1D EP16 (2 个 prefill EP8 + 1 个 decode EP16)。拓扑采纳了 functionstackx 的建议——wide EP 只对 decode 有益 ("wideEP is mainly helpful on decode as long as the weights fill in a single prefill node"); 同时说明 prefill TP8 → decode TP16 的非对称 KV 传输对 DeepSeek MLA 天然布局兼容 (KV 是按 token 复制的 latent, 不做 head 分片)。此外, 最初面向 spur 集群的方案因 "spur's ionic fabric could not cross-rail the MORI MoE all-to-all" 而放弃, 整体迁移到 bnxt RoCE 的 mi355x 集群, 并以 GSM8K $0.927 \geq 0.91$ 作为正确性门禁验证。

实现拆解

1. 配方矩阵 (8 个新 YAML): 新增 `scripts/ci/slurm/recipes/mi355x-{fp4,fp8}/dsv4{pro,flash}/1k1k/2p1d-ep16{-mtp}.yaml`, 声明 `resources` (`prefill_workers: 2 / decode_workers: 1`)、`backend.sglang_config` (`prefill TP/EP/DP=8, decode TP/EP/DP=16`)、`runtime` (`attention_backend: dsv4, moe_a2a_backend / kv_transfer_backend: mori, ib_devices: rdma0..7, dist_socket_ifname: eno0, rocm700a: 0`)。
2. `wide_ep` 参数定标 (硬件约束驱动): ① decode KV 池硬上限——ionic NIC 拒绝超过 4 GiB 的单一 RDMA 内存区, mori 把每个 KV buffer 注册为单一 region (无 chunking), 最大 `buffer = max_total_num_tokens × 256 B`, 上限 16,777,216 tokens, MTP 配方统一标为 7,000,000; ② MTP decode 的 MORI dispatch 缓冲 (512 / 2048 / 1024)——MTP 的 decode CUDA 图捕获会把 MoE dispatch 推到 `cuda_graph_bs × draft_tokens tokens/rank`, 缓冲不足会在捕获阶段中止; ③ 非对称 `mem-fraction` (`prefill 0.8 / decode 0.85, pro-fp8 MTP prefill` 因权重最重提到 0.85)。

3. nightly leg 登记: scripts/ci/slurm/nightly-configs.yaml 新增 8 条 runner: mi355x 的 leg (multinode: true、disagg: true) , 与既有 dp8ep8 leg 并列; Pro MTP 两条腿在 search-space 排除 conc256, 原因是分离式 decode 的 SWA hybrid pool retract→get_cpu_copy 为上游 NotImplementedError, 会在并发 256 时崩溃。
4. 正确性门禁: 每个配方在 bench 段配置并发扫描 (1..256) 之前先跑完整 GSM8K (1319 题、8-shot、准确率阈值 0.91) , 防止“快但错”的结果混入夜间测试。
5. 跨 PR 的配套加固: launcher (经同系列其他 PR 进入 main) 按 nodes-per-engine = ceil(TP/8) 推导引擎节点数, 跨节点 --nnodes / --node-rank / --dist-init-addr 仅对多节点引擎发射; MORI_IB_GID_INDEX=1 规避 link-local 与 global GID 不匹配; g20 节点先以条件例外 + SLURM_DIST_TAIL 作为 decode 非根 rank 使用, 随后在共享 host ionic 挂载就绪后重新放开全局排除; MORI_BOOTSTRAP_TIMEOUT=300 缓解 bootstrap 竞态。所有门控保证 TP=8 既有配方生成参数逐字节不变。

注: 本 PR 的 commit 历史包含 launcher 修改, 但最终合并 diff 为 9 个 YAML 文件 (951 行纯新增) , launch_mi355x.sh 的变更在合并前已随系列其他 PR 落入 main。

关键文件:

- scripts/ci/slurm/recipes/mi355x-fp4/dsv4flash/1k1k/2p1d-ep16-mtp.yaml (模块 CI 配方; 类别 infra; 类型 configuration) : MTP 分支的宽 EP16 配方代表, 集中体现硬件约束定标: decode KV 池上限 (ionic 4 GiB MR 限制 → 7M token) 、MTP 专用 MORI dispatch 缓冲 (512/2048/1024) 、非对称 mem-fraction, 以及 GSM8K 正确性门禁。
- scripts/ci/slurm/recipes/mi355x-fp4/dsv4pro/1k1k/2p1d-ep16-mtp.yaml (模块 CI 配方; 类别 infra; 类型 configuration) : Pro FP4 MTP 配方, 与 Flash 版本共享同一拓扑与参数骨架, 构成 Pro/Flash × fp4/fp8 矩阵的 MTP 分支。
- scripts/ci/slurm/recipes/mi355x-fp8/dsv4flash/1k1k/2p1d-ep16-mtp.yaml (模块 CI 配方; 类别 infra; 类型 configuration) : Flash FP8 MTP 配方, FP8 精度分支的代表, 与 fp4 分支形成精度矩阵。
- scripts/ci/slurm/recipes/mi355x-fp8/dsv4pro/1k1k/2p1d-ep16-mtp.yaml (模块 CI 配方; 类别 infra; 类型 configuration) : Pro FP8 MTP 配方, 权重最重, 其 prefill mem_fraction 需单独提到 0.85 (低于此值会中止) , 是同类配方中参数差异最大的一条。
- scripts/ci/slurm/recipes/mi355x-fp4/dsv4flash/1k1k/2p1d-ep16.yaml (模块 CI 配方; 类别 infra; 类型 configuration) : 非 MTP 分支配方, decode KV 上限为 16M, decode MORI 缓冲为 64/332/128, 与 MTP 分支形成对比, 体现投机解码对 dispatch 缓冲的放大效应。
- scripts/ci/slurm/recipes/mi355x-fp4/dsv4pro/1k1k/2p1d-ep16.yaml (模块 CI 配方; 类别 infra; 类型 configuration) : Pro FP4 非 MTP 配方, 构成矩阵的一角。
- scripts/ci/slurm/recipes/mi355x-fp8/dsv4flash/1k1k/2p1d-ep16.yaml (模块 CI 配方; 类别 infra; 类型 configuration) : Flash FP8 非 MTP 配方, 最终验证中唯一失败的腿 (MORI bootstrap 竞态) 即此分支。
- scripts/ci/slurm/recipes/mi355x-fp8/dsv4pro/1k1k/2p1d-ep16.yaml (模块 CI 配方; 类别 infra; 类型 configuration) : Pro FP8 非 MTP 配方, 构成矩阵的一角。

- scripts/ci/slurm/nightly-configs.yaml (模块 CI 配置; 类别 infra; 类型 configuration) : 夜间矩阵入口, 登记 8 条 mi355x EP16 leg (multinode + disagg), 并记录 Pro MTP 需排除 conc256 的上游缺陷, 是配方与调度器之间的契约层。

关键符号: launch_mi355x.sh, nightly-configs.yaml, 2p1d-ep16*.yaml 配方

关键源码片段

scripts/ci/slurm/recipes/mi355x-fp4/dsv4flash/1k1k/2p1d-ep16-mtp.yaml

MTP 分支的宽 EP16 配方代表, 集中体现硬件约束定标: decode KV 池上限 (ionic 4 GiB MR 限制 → 7M token)、MTP 专用 MORI dispatch 缓冲 (512/2048/1024)、非对称 mem-fraction, 以及 GSM8K 正确性门禁。

```
# MI355X DeepSeek-V4-Flash FP4 4 节点 2P1D 分离式配方 (窄 prefill EP8 + 宽 decode EP16 + MTP)
```

```
# Oren 拓扑: wide EP 只对 decode 有收益; 2 个单节点 prefill 引擎 (EP8) 承载路由分发,
```

```
# 1 个 decode 引擎跨 2 节点 (EP16), 共 4 节点, mori 同时承载 MoE all-to-all 与 P->D KV 传输。
```

```
resources:
```

```
  prefill_workers: 2
```

```
  decode_workers: 1
```

```
backend:
```

```
  sglang_config:
```

```
    # prefill 保持窄 EP8 (节点内 XGMI all-to-all, 已验证可靠); decode 放宽到 EP16。
```

```
    prefill:
```

```
      tensor-parallel-size: 8
```

```
      expert-parallel-size: 8
```

```
      data-parallel-size: 8
```

```
    decode:
```

```
      tensor-parallel-size: 16
```

```
      expert-parallel-size: 16
```

```
      data-parallel-size: 16
```

```
runtime:
```

```
  image: lmsysorg/sglang-rocm:v0.5.15.post1-rocm720-mi35x-20260715
```

```
  attention_backend: dsv4
```

```
  # RoCE HCA (每节点 8 个) 同时供 mori MoE all-to-all 与 P->D KV 传输使用。
```

```
  ib_devices: rdma0,rdma1,rdma2,rdma3,rdma4,rdma5,rdma6,rdma7
```

```
  moe_a2a_backend: mori
```

```
  kv_transfer_backend: mori
```

```
  dist_socket_ifname: eno0 # 宽 decode 引擎跨节点 torch-distributed 初始化的 NIC
```

```
  rocm700a: 0 # rocm720 0715 镜像需关闭 ROCm-7.0.0-alpha 路径 (已验证)
```

```
  prefill_port: 30025
```

```
  decode_port: 30026
```

```
  prefill_bootstrap_port: 8998
```

```
  decode_bootstrap_port: 9001
```

```
  lb_port: 8000
```

```
mem_fraction_static: 0.90
page_size: 256
max_running_requests: 256
chunked_prefill_size: 131072
swa_full_tokens_ratio: 0.1
```

wide_ep 参数块：按角色拆分定标，覆盖单节点 EP<=8 配方默认的对称配置。

wide_ep:

```
kv_cache_dtype: fp8_e4m3
prefill_mem_fraction_static: 0.8
decode_mem_fraction_static: 0.85
prefill_chunked_prefill_size: 131072
prefill_max_running_requests: 1024
decode_max_running_requests: 1024
common_extra_flags: "--moe-dense-tp-size 1 --enable-dp-lm-head --decode-log-interval 100 --
watchdog-timeout 3600 --load-balance-method round_robin"
prefill_extra_flags: "--context-length 9217 --max-total-tokens 262144"
# decode KV 池硬上限：ionic NIC 拒绝超过 4 GiB 的单一 RDMA 内存区，而 mori 把每个
# KV buffer 注册为单一 region（无 chunking）；最大 buffer = max_total_num_tokens * 256
B,
# 故上限为 16,777,216 tokens。MTP 定标 7,000,000（已验证值，低于所有 MTP leg
的自然池）。
# 下方 cuda-graph-bs 处为 1 到 128 的完整枚举（此处省略号仅为展示缩写）。
decode_extra_flags: "--max-total-tokens 7000000 --cuda-graph-bs 1 2 3 ... 128 --prefill-
round-robin-balance"
prefill_extra_env:
  MORI_MAX_DISPATCH_TOKENS_PREFILL: 8192
  MORI_MAX_DISPATCH_TOKENS_DECODE: 256
  SGLANG_MORI_NUM_MAX_DISPATCH_TOKENS_PER_RANK: 16384
# MTP 的 decode CUDA 图捕获会把 MoE dispatch 推到 cuda_graph_bs * draft_tokens
# tokens/rank，缓冲不足会在捕获阶段中止；512/2048/1024 为 mi355x 实测值。
decode_extra_env:
  MORI_MAX_DISPATCH_TOKENS_DECODE: 512
  MORI_MOE_MAX_INPUT_TOKENS_DECODE: 2048
  SGLANG_MORI_NUM_MAX_DISPATCH_TOKENS_PER_RANK: 1024
```

MTP / EAGLE 投机解码（NextN 头来自基础模型），prefill 与 decode 均启用。

mtp:

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

bench:

```
concurrencyes: [1, 8, 16, 32, 64, 128, 256]
num_prompts_factor: 4
random_range_ratio: 1.0
```

正确性门禁：性能扫描之前先走 PD 路径跑完整 GSM8K（1319 题 8-shot），

准确率低于 0.91 即判失败，避免“快但错”的结果通过夜间测试。

```
accuracy:
  enabled: true
  num_shots: 8
  num_questions: 1319
  threshold: 0.91
```

评论区精华

核心讨论围绕拓扑设计与集群环境验证展开。① functionstackx 建议把 wide EP 只放在 decode (2 个单节点 prefill EP8 + 1 个跨 2 节点 decode EP16)，HaiShaw 认可、Lzy17 据此把 8 个配方从 1P1D 重写为 2P1D。② bingxche 的纯合并验证发现全局 SLURM_EXCLUDE=mia1-p01-g20 导致 4 节点请求只剩 3 节点可用而无法启动；进一步验证表明 g20 的排除不能简单删除 (EP8 路径在 g20 上遭遇 ionic ABI 不匹配)，最终采用条件例外 + SLURM_DIST_TAIL 固定 g20 为 decode 非根 rank。③ 完整 8 腿验证 7/8 通过，唯一失败为 Flash FP8 base 的 MORI bootstrap 节点内竞态。

- 拓扑设计：从 1P1D 调整为 2P1D (宽 EP 只放 decode) (design): 采纳 2P1D 拓扑: 2 个单节点 prefill EP8 + 1 个跨 2 节点 decode EP16; 配方文件更名 1p1d-ep16→ 2p1d-ep16。
- 全局 SLURM_EXCLUDE=mia1-p01-g20 导致 4 节点请求失败 (correctness): EP16 腿需要条件例外并加 SLURM_DIST_TAIL=mia1-p01-g20, 让 g20 作为 decode 非根 rank 参与四节点请求。
- g20 的 ionic ABI 不匹配导致排除不能简单删除 (correctness): 最终方案: 共享 host ionic 挂载修复 ABI 后重新放开 g20 (保留 wide-EP 根节点顺序), 相关 commit 由 bingxche 完成 (e9085598 / d7a79633 / 0d0dcbd5)。
- 完整 8 腿验证 7/8 通过, Flash FP8 base 遭遇 MORI bootstrap 竞态 (performance): PR 设置 MORI_BOOTSTRAP_TIMEOUT=300 放大握手窗口, 竞态本质未根除, 属已知不稳定点。

风险与影响

- 风险: ① MORI bootstrap 竞态未根除: 最终验证 7/8 中失败的正是节点内握手竞态 (Rank 2/4/6 连接失败、AcceptSocket 超时), MORI_BOOTSTRAP_TIMEOUT=300 仅放大握手窗口, 竞态本质仍在, Flash FP8 base 腿存在偶发失败风险。② KV 池上限的长期影响: ionic 4 GiB MR 限制把 decode max-total-tokens 压到 16.78M (MTP 7M), 长上下文或高并发下可能成为吞吐瓶颈, 且依赖 mori 是否支持 buffer chunking。③ 多节点路径覆盖有限: 跨节点 EP16 仅在 mi355x 集群 8 条夜间腿中回归, g20 节点排除 / 固定逻辑与环境强耦合, 集群拓扑变化时易回归。④ MTP 依赖上游缺陷: Pro MTP 排除 conc256 (SWA get_cpu_copy NotImplementedError)、overlap-plan-stream 需对 MTP 关闭 (DSV4 MLA 后端未实现 update_verify_buffers_to_fill_after_draft), 上游修复前覆盖不完整。⑤ 对生产代码零影响: 合并 diff 全为 CI 配置, 风险集中在夜间测试有效性层面。
- 影响: 为 AMD MI355X 集群新增 8 条 4 节点 2P1D 分布式夜间覆盖, 首次把跨节点宽 EP16 (MoE all-to-all + KV 传输双通道走 mori) 纳入常规回归; 对 AMD DI 团队而言是系

列化推进的关键里程碑，沉淀了硬件约束（RoCE GID、ionic MR 上限、RDMA ABI）驱动的配方定标方法论。对用户与生产服务无任何运行时影响（纯 CI 基础设施变更）。

- 风险标记：MORI bootstrap 竞态未根除，4 GiB RDMA MR 上限约束 KV 池，节点排除 / 固定逻辑与环境强耦合，MTP 依赖上游未实现 API，多节点路径仅 CI 验证

关联脉络

- PR #33333 [AMD][DI][CI] 6/N Add Kimi-K2.6 MXFP4 wide-EP16 2P1D nightly recipes: 同一 AMD DI CI 系列的下一个 PR，复用本 PR 确立的多节点引擎与 mi355x 宽 EP16 启动路径。
- PR #33374 [AMD][DI][CI] Auto-mount latest host ionic/ibverbs userspace in the MI355X nightly launcher (fix ABI mismatch): 与本 PR 的 g20/ionic ABI 处理同源：g20 节点最终放开的方案依赖共享 host ionic 挂载，两边由同一批作者协同完成。
- PR #29084 Merged EP8 recipes (PR body 提及): 既有 EP8 配方，本 PR 的非回归底线：所有新逻辑门控保证其生成参数逐字节不变。
- PR #29784 Merged EP8 recipes (PR body 提及): 既有 EP8 配方，同上，属非回归保护对象。
- PR #29855 Merged EP8 recipes (PR body 提及): 既有 EP8 配方，同上，属非回归保护对象。