

PR #39702 完整报告

sgl-project/sglang

[AMD] Update deepseek-v4 PDI and cache policy setting for agentic workload

合并时间: 2026-09-16 12:16

原文链接: <http://prhub.com.cn/sgl-project/sglang/pull/39702>

执行摘要

本 PR (#39702) 是一次面向 DeepSeek-V4 在 AMD MI300X/MI355X 上 agentic 工作负载的文档与配置对齐更新: ROCm 镜像标签从 20260911 升到 20260914, MI355X FP4 各档位 `--prefill-decode-interval` 从 10 调为 20, 并在 cookbook 中新增 `sglang_router --policy cache_aware` 路由建议、删除旧的 prefill-delayer 稳定性说明。变更与 SemiAnalysis InferenceX#3120 对齐, 属于低风险文档维护, 无任何运行时代码改动。

功能与动机

PR body 明确说明目的是「Align with SA submission」, 即与 [SemiAnalysisAI/InferenceX#3120](#) 保持配置一致, 具体三项:

- Bump 镜像到 0914 版本。
- 将 PDI (`--prefill-decode-interval`) 设为 20。
- 将缓存路由策略切换为 `cache-aware`。

核心动机是: 针对 MI355X 上长上下文 agentic 负载, 原先的调度稳定性方案 (PDI 10 + prefill-delayer) 被新的组合 (PDI 20 + `cache-aware` 路由) 取代, SGLang 侧文档需要同步, 避免用户按旧指引部署得到次优行为。仓库内无关联 Issue。

实现拆解

1. 部署面板配置片段 (`docs/src/snippets/configs/deepseek-ai/deepseek-v4.jsx`): 将 `dockerImages` 中 `mi300x`、`mi355x` 两行的镜像标签从 `v0.5.19-rocm720-mi30x/mi35x-20260911` 升到 `20260914`; 随后把 MI355X + FP4 Pro Official 的 `low-latency`、`balanced`、`high-throughput` 三个策略单元的 `--prefill-decode-interval` 从 10 统一改为 20。该文件是 Playground 部署面板生成启动命令的数据源, 改动直接影响面板产出的部署参数。
2. cookbook 文档 (`docs/cookbook/autoregressive/DeepSeek/DeepSeek-V4.mdx`): 同步更新两处镜像标签与 `docker pull` 示例; 在 §3.7 agentic recipe 中把两处 `--prefill-decode-interval` 改为 20; 删除旧段落中「PDI 10 用于 scheduler stability + `--enable-prefill-delayer`」的说明; 在 §3.7 结尾新增 `cache-aware` 路由建议 (`sglang_router --policy cache_aware`, 并发超 160 时追加 `--balance-abs-threshold 32`); §3.8 PD 场景镜像标签同步更新。
3. 测试与代码配套: 无。本次不涉及 Python/Rust 源码, 也没有新增测试, 配置有效性依赖文档构建与人工复核。

关键源码片段

[docs/src/snippets/configs/deepseek-ai/deepseek-v4.jsx](#)

部署面板的配置数据源：AMD 镜像标签升到 0914，MI355X FP4 三档策略的 --prefill-decode-interval 统一从 10 调到 20，是本 PR 的核心配置载体。

```
// AMD daily-updated lmsysorg/sglang-rocm images.
// Bump the dated tag when you re-verify on a newer build.
// 本次从 20260911 升到 20260914，与 SA 侧的镜像验证结果对齐。
dockerImages: {
  mi300x: "lmsysorg/sglang-rocm:v0.5.19-rocm720-mi30x-20260914",
  mi355x: "lmsysorg/sglang-rocm:v0.5.19-rocm720-mi35x-20260914",
},

// MI355X + FP4 — Pro Official (0813), low-latency 单元。
// PDI (--prefill-decode-interval) 从 10 调至 20：降低 prefill 与 decode
// 批次交替频率，适配长上下文 agentic 负载的调度稳定性诉求。
// balanced / high-throughput 两个单元也做了同样调整。
{
  match: { hw: "mi355x", variant: "pro-official", quant: "fp4", strategy: "low-latency", nodes:
    "single" },
  verified: false,
  env: ["SGLANG_USE_ROCM700A=0", "TORCH_BLAS_PREFER_HIPBLASLT=1",
    "SGLANG_HACK_FLASHMLA_BACKEND=unified_kv_triton",
    "AITER_BF16_FP8_MOE_BOUND=0", "SGLANG_OPT_USE_AITER_BATCHED_GEMM=true"],
  ,
  flags: [
    "--trust-remote-code",
    "--model-path {{MODEL_NAME}}",
    "--tp 8",
    "--prefill-decode-interval 20",
    "--attention-backend dsv4",
    "--enable-deepseek-v4-fp4-indexer",
    "--page-size 256",
    "--mem-fraction-static 0.90",
    "--swa-full-tokens-ratio 0.15",
    "--enforce-shared-experts-fusion",
    "--kv-cache-dtype fp8_e4m3",
    "--chunked-prefill-size 16384",
    "--speculative-algorithm DSPARK",
    "--host {{HOST_IP}}",
    "--port {{PORT}}",
  ],
},
```

评论区精华

本 PR 没有任何 review 评论或讨论线程，yichiche 与 HaiShaw 两位审核人均直接 APPROVE（yichiche 回复「LGTM.」）。不存在技术交锋或未解决疑虑。

风险与影响

- 参数变更缺基准验证：--prefill-decode-interval 从 10 调到 20 会改变 scheduler 的 prefill/decode 批次交替频率，直接影响调度稳定性与首 token 延迟，但 PR 的 Speed Tests 与 Accuracy Tests 均标注 None，缺少实测数据背书。该参数原先承载「scheduler stability」的职责，调整后文档删除了 prefill-delayer 的旧防护说明，长 prefill 独占引擎的问题在新方案下如何兜底未明确。
- 镜像标签可用性：20260914 是 daily 构建标签，若该镜像被下架或内容变动，文档固定标签会失效；PR 未说明 0914 相对 0911 的修复内容。
- router 策略依赖：cache-aware 路由建议依赖 sglang_router 对应策略可用，用户 router 版本不匹配时按文档部署会失败。
- 影响面：仅影响阅读 cookbook 的 AMD 用户与 Playground 部署面板生成的参数，对运行时代码零影响，整体风险低。

关联脉络

本 PR 属于外部驱动文档对齐：上游是 SemiAnalysisAI/InferenceX#3120，说明 SGLang 与 SA 在 DeepSeek-V4 部署配方上存在协同迭代机制。仓库内部，该 PR 与 DeepSeek-V4 部署链路的近期演进同属一条线——包括 [PR#38984]（INTRA_NODE_NVLINK 传输批量优化）与 [PR#36534]（DSV4 BCG 显存优化），两者都涉及 V4 在 AMD/ 调度 /KV 上的推荐配置变化；本 PR 的价值在于把这些已落地或新提出的参数选择沉淀为面向用户的文档契约，供下游部署者直接消费。