PR #37092 完整报告

sgl-project/sglang

[AMD] Update v4 amd cookbook 0830

合并时间：2026-08-30 14:55

原文链接：http://prhub.com.cn/sgl-project/sglang/pull/37092

执行摘要

PR #37092 是一次面向 DeepSeek-V4 AMD cookbook 的例行维护更新：为所有 AMD ROCm 配置单元追加 TORCH_BLAS_PREFER_HIPBLASLT=1 环境变量，将 tp8 档位的 --chunked-prefill-size 从 8192 提升到 16384，并把 AMD 镜像日期标签从 20260828 统一 bump 到 20260829。变更仅涉及网站渲染数据源 deepseek-v4.jsx 与用户文档 DeepSeek-V4.mdx，不含任何运行时代码，属于低风险文档型PR，由HaiShaw直接批准合并。

功能与动机

PR body 列出了三点改动动机：

- 新增 TORCH_BLAS_PREFER_HIPBLASLT=1，让 PyTorch 在 ROCm 上的 GEMM 优先走 hipBLASLt 实现；
- tp8 档位的 chunked prefill 从 8192 提升到 16384，减少 prefill 分块与调度开销；
- AMD 镜像版本从 0828 bump 到 0829，跟随每日更新的 lmsysorg/sglang-rocm 镜像。

PR 未关联 Issue，Speed Tests and Profiling 一节标注为 None，属于 AMD 侧重新验证后对 cookbook 推荐参数的例行刷新。

实现拆解

1. 变更入口：docs/src/snippets/configs/deepseek-ai/deepseek-v4.jsx 是 DeepSeek-V4 cookbook 网站的数据源，站点上所有硬件配置单元（match / env / flags 三段式对象）都由它渲染。
2. 环境变量统一补充：所有 AMD ROCm 单元（MI300X、MI355X）的 env 数组统一追加 TORCH_BLAS_PREFER_HIPBLASLT=1，与已有的 SGLANG_USE_ROCM700A=0、SGLANG_HACK_FLASHMLA_BACKEND=unified_kv_triton、AITER_BF16_FP8_MOE_BOUND=0 等开关共同构成 ROCm 7.2 上的推荐组合。
3. chunked-prefill 调参：将原为 8192 的 tp8 单元（如 MI300X Flash FP8 low-latency、MI355X Pro FP4 low-latency 等）统一调整为 16384；dp8 的 balanced / high-throughput 单元维持 65536 不变，说明本次调参只针对 tp8 场景。
4. 镜像标签同步：JSX 的 dockerImages.mi300x / mi355x 与 MDX 文档中的 docker pull / run 示例同步更新为 20260829，保证网站渲染与正文一致。
5. 配套改动：无测试、无 schema、无部署配套；各单元的 verified 字段未随本次变更调整，即新调参结论尚未被标记为“已完整验证”。

docs/src/snippets/configs/deepseek-ai/deepseek-v4.jsx

DeepSeek-V4 cookbook 网站的数据源文件，本次所有核心配置变更（env 追加、chunk 调整、镜像标签）都发生在这里，是 PR 的主体。

```
// DeepSeek-V4 AMD cookbook 数据源（deepseek-v4.jsx）
// 本片段展示 MI355X Pro FP4 低延迟单元，覆盖本次三项核心变更：
// 1) env 新增 TORCH_BLAS_PREFER_HIPBLASLT=1，让 torch 侧 GEMM 优先走 hipBLASLt；
// 2) --chunked-prefill-size 由 8192 提升至 16384，减少 prefill 分块与调度开销；
// 3) AMD 每日镜像日期标签统一 bump 至 20260829。

{
  match: { hw: "mi355x", variant: "pro-official", quant: "fp4", strategy: "low-latency", nodes:
  "single" },
  verified: false,
  // 与 SGLANG_HACK_FLASHMLA_BACKEND=unified_kv_triton、AITER 相关开关配合，
  // 该 env 组合是 ROCm 7.2 上本次重新验证后的推荐配置。
  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",
    "--attention-backend dsv4",
    "--page-size 256",
    "--mem-fraction-static 0.90",
    "--swa-full-tokens-ratio 0.15",
    "--enforce-shared-experts-fusion",
    "--kv-cache-dtype fp8_e4m3",
    // tp8 档位统一从 8192 提到 16384；dp8 的 balanced / high-throughput 单元保持 65536。
    "--chunked-prefill-size 16384",
    "--host {{HOST_IP}}",
    "--port {{PORT}}",
  ],
}

// dockerImages 中 AMD 每日镜像标签同步 0828 -> 0829
// 该日期标签与 MDX 文档中的 docker pull / run 示例保持一致。
dockerImages: {
  mi300x: "lmsysorg/sglang-rocm:v0.5.18-rocm720-mi30x-20260829",
  mi355x: "lmsysorg/sglang-rocm:v0.5.18-rocm720-mi35x-20260829",
}
```

评论区精华

该 PR 没有任何评论或 review 评论（comments_count 与 review_comments_count 均为 0），HaiShaw 以空 body 直接 APPROVED，因此没有可提炼的技术交锋。唯一值得留意的是 CI 状态：Base PR Test 通过，Extra 测试失败但未留下原因说明，且 AMD ROCm 7.2 的 PR job 并未为该 commit 运行——对一个 AMD 文档 PR 而言属于轻度覆盖缺口。

风险与影响

- chunked-prefill-size 从 8192 增至 16384 会提高单次 prefill 的中间激活显存峰值，实际影响取决于序列长度与 batch；PR 未附 speed / memory benchmark（Speed Tests 一节标 None），用户在大 batch 场景下需自行验证。
- 20260829 标签依赖 Docker Hub 上镜像真实发布且可用；文档直接引导用户 pull，若镜像尚未同步，用户会立即拉取失败。
- JSX 与 MDX 两处镜像标签需要同步维护，存在漏改一处导致文档不一致的长期维护成本。
- 影响范围仅限在 MI300X / MI355X 上部署 DeepSeek-V4 的用户与复现流程；不涉及任何运行时路径，无代码回归风险。

关联脉络

- 与 PR #36954（FlashInfer 0.6.18 bump，同时更新 Kimi-K3.mdx 的镜像指引）同属“cookbook 文档镜像日期标签维护”模式：日期化、每日更新的镜像标签已成为仓库 cookbook 的惯例。
- 该 PR 是 DeepSeek-V4 AMD cookbook 持续迭代的一环（配置注释中仍保留 0813 的字样），后续预计会随 ROCm 镜像每日更新继续 bump 标签。
- 与仓库近期高强度的核心开发（Unified Memory、HiCache 链路、PD 协议对齐等）完全解耦，属于独立的文档维护线。