

PR #30651 完整报告

sgl-project/sglang

cookbook(deepseek-v4): add MORI disagg backend for AMD + bump MI355X image

合并时间: 2026-07-15 15:33

原文链接: <http://prhub.com.cn/sgl-project/sglang/pull/30651>

执行摘要

此 PR 为 DeepSeek-V4 Playground 添加 AMD 专属的 MORI PD 分离传输后端，并同步更新 MI355X 的 Docker 镜像和多项性能调优参数。变更仅涉及 cookbook 配置和 UI 逻辑，无引擎代码修改。

功能与动机

- 动机: AMD 用户 (MI300X/MI355X) 需要除 Mooncake/NiXL 之外的 PD 分离传输后端选项。MORI 来自 InferenceX 的自研传输方案。
- 参考: PR body 关联 SemiAnalysisAI/InferenceX#1818 (MORI disagg)、#2093 (DSV4 non-MTP 更新)、#2108 (DSV4 MTP 更新)。

实现拆解

1. 添加 MORI 后端: 在 docs_new/src/snippets/configs/deepseek-ai/deepseek-v4.jsx 的 pdDisagg.transferBackends 数组中插入 { id: "mori", label: "MORI", hide: { hw: ["h100", "h200", "b200", "b300", "gb200", "gb300", "rtx6000"] } }, 通过 hide 约束使 MORI 仅对 ROCm 平台可见。
2. 更新 MI355X 镜像和参数: Docker 镜像从 v0.5.13.post1 升级到 v0.5.14; --swa-full-tokens-ratio 从 0.1 改为 0.15; 将 --enable-prefill-delayer 和 --prefill-delayer-max-delay-ms 5000 替换为 --enable-two-batch-overlap; 为 DP 注意力单元添加 SGLANG_SHARED_EXPERT_TP1=1、SGLANG_DP_SHARED_EXPERT_LOCAL=1、SGLANG_DP_USE_REDUCE_SCATTER=1 环境变量。
3. Playground 回退逻辑: 在 docs_new/src/snippets/_playground.jsx 的 pdDisagg.revertHidden 中增加对 transferBackend 的检测, 避免硬件切换后选中被隐藏的后端导致 UI 状态卡住。

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

主要变更文件: 添加 MORI 传输后端、更新 MI355X 镜像和性能参数。

```
// 展示 MORI 后端的添加和镜像更新 (来自 deepseek-v4.jsx)
```

```
// AMD daily-updated lmsysorg/sglang-rocm images. Bump the dated tag when you  
// re-verify on a newer build.
```

```
mi300x: "lmsysorg/sglang-rocm:v0.5.13.post1-rocm720-mi30x-20260623",
```

```
mi355x: "lmsysorg/sglang-rocm:v0.5.14-rocm720-mi35x-20260708", // 镜像标签从 v0.5.13.post1
```

升级

```
// ----- Card 5: "PD Disaggregation" -----
pdDisagg: {
  modes: [
    { id: "off", label: "Off" },
    { id: "prefill", label: "Prefill role" },
    { id: "decode", label: "Decode role" },
  ],
  transferBackends: [
    { id: "mooncake", label: "Mooncake", ... },
    { id: "nixl", label: "NiXL" },
    // MORI-IO transport is AMD-only — hidden on every non-ROCm platform.
    { id: "mori", label: "MORI",
      hide: { hw: ["h100", "h200", "b200", "b300", "gb200", "gb300", "rtx6000"] } },
  ],
  // ...
},
```

```
// 示例：某个 DP-Attention 单元的参数替换（两处类似的修改）
// 之前：--enable-prefill-delayer --prefill-delayer-max-delay-ms 5000
// 之后：--enable-two-batch-overlap
"--swa-full-tokens-ratio 0.15", // 从 0.1 提升
"--enable-two-batch-overlap", // 替换老化的 prefill delayer
// 环境变量新增
"SGLANG_SHARED_EXPERT_TP1=1",
"SGLANG_DP_SHARED_EXPERT_LOCAL=1",
"SGLANG_DP_USE_REDUCE_SCATTER=1",
```

[docs_new/src/snippets/_playground.jsx](#)

添加 transferBackend 隐藏时的回退逻辑，确保硬件切换后后端选择不会被卡住。

```
// 在 pdDisagg.revertHidden 函数中添加 transferBackend 检测（位于 _playground.jsx）
pdDisagg: {
  initState: () => ({ mode: "off", transferBackend: "mooncake", ibDevice: "auto" }),

  revertHidden: (value, fc, base, h) => {
    let changed = false;
    const next = { ...value };
    // 原有的 mode 和 ibDevice 检测...
    // 新增：若当前 transferBackend 在 fc.transferBackends 中被隐藏，则回退到 mooncake
    if (next.transferBackend !== "mooncake" && fc.transferBackends
      && h.isHidden(fc.transferBackends, next.transferBackend, base)) {
      next.transferBackend = "mooncake"; changed = true;
    }
    return changed ? next : value;
  },
  // ...
}
```

评论区精华

无实质性讨论；仅有 bot 配额提示。

风险与影响

- 风险：极低。仅影响 cookbook 配置，无引擎变更。需确认 MI355X 新参数在正式环境中的稳定性。
- 影响：AMD MI300X/MI355X 用户可在 Playground 中选择 MORI 后端；所有 MI355X 预设命令和镜像更新为最新调优基准。非 AMD 用户无感知。

关联脉络

- 关联上游仓库 SemiAnalysisAI/InferenceX 的三个 PR (#1818、#2093、#2108)，属于 AMD 基准测试矩阵的持续更新。
- 当前仓库无其他直接关联 PR。