PR #23691 完整报告

sgl-project/sglang

docs(DeepSeek-V4): mark gb300|{small,big}|{cp,pd-disagg} verified + GB300-specific fixes

合并时间：2026-04-25 12:21

原文链接：http://prhub.com.cn/sgl-project/sglang/pull/23691

执行摘要

该 PR 标记了 DeepSeek-V4 在 GB300 硬件上的 cp 和 pd-disagg 共 4 个部署配方为已验证，并针对 GB300 的特性（更大显存压力、DeepEP 断言）添加了必要的环境变量和参数调整。同时更新了文档，说明跨 pod MNNVL 问题的解决办法。变更范围小，仅涉及文档片段和配置常量，风险极低。

功能与动机

本 PR 的目的是将 GB300 上的 DeepSeek-V4 部署配方（cp 和 pd-disagg）从“待验证”状态提升为“已验证”，确保用户能够一键复制可运行的命令。动机来源于作者团队在 GCP a4x 集群上的实际测试（journal 2026-04-25-001-gb300-cp-pd-cookbook-verify），验证过程中发现了内存不足和 DeepEP 断言失败等问题，需要在配方中注入特定修复。

实现拆解

1. 标记 GB300 已验证配方

文件：docs_new/src/snippets/autoregressive/deepseek-v4-deployment.jsx 在 VERIFIED_RECIPES 集合中新增 4 项：gb300|small|cp、gb300|big|cp、gb300|small|pd-disagg、gb300|big|pd-disagg。这些字符串对应文档中生成的部署表格，标记后该单元格的命令将显示为可执行状态。

2. 调整 GB300 big CP 的内存比例

在 cp 配方生成逻辑中，原来统一设置 --mem-fraction-static 0.78。但对于 GB300 big（Pro 模型 1.6T 权重，tp=4 时约 224 GB/ 卡），0.78 的目标值（273 × 0.78 = 212.94 GB）小于权重占用，导致 KV pool 初始化失败。修改为：

```
// 显存：GB300 总 273 GB，Pro 权重 224 GB，0.88 后约 240 GB，剩余 16 GB KV + 33 GB
runtime
if (hardware === "gb300" && isBig) {
  flags.push("  --mem-fraction-static 0.88");
} else {
  flags.push("  --mem-fraction-static 0.78");
}
```

此修改仅影响 GB300 big 的 cp 配方，其他硬件和其他配方不变。

3. 添加 DeepEP 分派 buffer 上限环境变量

在 PD 模式的环境变量生成中，新增对 GB300 的判断：

```
// 防止 deep_ep.cpp:1233 断言：x.size(0) <= num_max_dispatch_tokens_per_rank
// big 模型 max-running-requests 为 128，per-rank=32 ≤ 256
// small 模型 max-running-requests 为 256，per-rank=64 ≤ 1024
if (isGB300) {
  roleEnv.push(modelSize === "big"
    ? "SGLANG_DEEPEP_NUM_MAX_DISPATCH_TOKENS_PER_RANK=256"
    : "SGLANG_DEEPEP_NUM_MAX_DISPATCH_TOKENS_PER_RANK=1024");
}
```

此环境变量在 B200 上仅用于 *decode* 角色，而 GB300 上 prefill 和 *decode* 都需要设置。

4. 更新文档说明跨 pod MNNVL 问题

文件：DeepSeek-V4.mdx 的 §3.2 节新增一段说明，指导用户在跨 pod KV 传输失败时添加 MC_FORCE_MNNVL=1 NCCL_MNNVL_ENABLE=1 NCCL_CUMEM_ENABLE=1 到 sglang serve 命令前。

docs_new/src/snippets/autoregressive/deepseek-v4-deployment.jsx

核心变更文件，包含所有配方标记和环境变量修复逻辑

```
// 在 VERIFIED_RECIPES 中添加 GB300 的 cp 和 pd-disagg 配方
const VERIFIED_RECIPES = new Set([
  "b200|small|low-latency",
  // ... 其他已验证配方
  "gb300|small|cp",
  "gb300|big|cp",
  "gb300|small|pd-disagg",
  "gb300|big|pd-disagg",
]);

// cp 配方中，GB300 big 需要更高的 mem-fraction-static 以避免内存不足
// GB300 big CP: Pro 1.6T 权重在 tp=4 时约 224 GB / 卡，
// 0.78 会导致 KV pool 初始化失败，0.88 为经验值
if (hardware === "gb300" && isBig) {
  flags.push("  --mem-fraction-static 0.88");
} else {
  flags.push("  --mem-fraction-static 0.78");
}

// GB300 PD 模式下设置 DeepEP 分派 buffer 上限，防止 deep_ep.cpp 断言失败
if (isGB300) {
  // big 模型 max-running-requests 128，per-rank=32 ≤ 256
  // small 模型 max-running-requests 256，per-rank=64 ≤ 1024
  roleEnv.push(modelSize === "big"
    ? "SGLANG_DEEPEP_NUM_MAX_DISPATCH_TOKENS_PER_RANK=256"
    : "SGLANG_DEEPEP_NUM_MAX_DISPATCH_TOKENS_PER_RANK=1024");
}
```

评论区精华

本 PR 无 review 评论。

风险与影响

- 风险：仅修改文档片段和配置常量，不涉及运行时代码，风险极低。但参数（--mem-fraction-static 0.88、SGLANG_DEEPEP_NUM_MAX_DISPATCH_TOKENS_PER_RANK）基于特定集群验证，不同硬件拓扑或未来模型权重变化后可能失效。
- 影响：用户可直接复制已验证的 GB300 部署命令，减少试错成本；团队内部降低 GB300 部署支持负担。

关联脉络

该 PR 是 #23690（"Small udpate gb300 recipe for deepseek v4"）的延续，后者标记了低延迟和平衡配方，本 PR 补全了 cp 和 pd-disagg 配方。整体构成了 GB300 上 DeepSeek-V4 部署配方的完整验证。