PR #23715 完整报告

sgl-project/sglang

docs(DeepSeek-V4): mark h200lbiglpd-disagg verified + recipe fixes

合并时间：2026-04-25 22:49

原文链接：http://prhub.com.cn/sgl-project/sglang/pull/23715

执行摘要

本 PR 将 DeepSeek-V4 在 H200 big 硬件上的 PD 分解部署方案从“待验证”标记为“已验证”，并引入三个关键配方修复：DeepEP 调度缓冲区上限、强制 deepep MoE 后端、以及内存预算调整。同时修复了多节点 PD 场景下 --dist-init-addr 与 multiNodeFlags 冲突的问题。

功能与动机

h200lbiglpd-disagg 配方此前因缺乏 4 节点 H200 集群（带共享 IB 网络）而标记为“待验证”。现在验证工作已在可用集群上完成，需要更新已验证集合，并加入实际部署中发现的关键配置修正，以确保文档生成的命令开箱即用。

实现拆解

1. 标记已验证：在 VERIFIED_RECIPES 集合中添加 "h200lbiglpd-disagg"，删除之前注释说明。
2. DeepEP 缓冲区上限：在 buildRole 函数中，当 hardware === "h200" && modelSize === "big" 时，设置环境变量 SGLANG_DEEPEP_NUM_MAX_DISPATCH_TOKENS_PER_RANK=128，防止 DeepEP 在 tp=16 多节点下因 token 过多而断言失败。
3. MoE 后端强制 deepep：将 --moe-a2a-backend deepep 的条件从 isBlackwell 扩展为包括 H200 big，因为 tp=16 时 FP8 block_n=128 无法整除 MoE intermediate_size_per_partition (3072/16=192)，必须将 expert 保留在单 rank 内而非 TP 分片。
4. dist-init-addr 条件修复：修改为 !isGB300 && !multinode，避免多节点 PD 时覆盖 multiNodeFlags 中已设置的正确跨节点地址。
5. 内存预算调整：添加 --cuda-graph-max-bs 128 和 --mem-fraction-static 0.9，解决显存不足（available_gpu_memory ~17.93 GB vs reserve target 87 GB）并提升解码吞吐。

核心代码片段如下：

关键源码片段

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

唯一变更文件，包含已验证配方标记更新和三个 H200 big 部署修复逻辑。

```
// 已验证配方集合（新增 h200lbiglpd-disagg）
```

```
const VERIFIED_RECIPES = new Set([
  /* ... 其他已验证配方 ... */
  "h200|big|pd-disagg", // 此前为注释状态，现已验证通过
  /* ... */
]);

// 在 buildRole 函数内部，为 H200 big 添加特定环境变量和参数
if (hardware === "h200" && modelSize === "big") {
  // DeepEP dispatch buffer 上限，防止 per-rank token 超出限制
  roleEnv.push("SGLANG_DEEPEP_NUM_MAX_DISPATCH_TOKENS_PER_RANK=128");
}
// ... 省略中间代码 ...
// MoE 后端：H200 big 的 tp=16 下 FP8 block_n=128 无法整除中间层大小，
// 必须使用 deepep 保持 expert 在单 rank 内
if (isBlackwell || (hardware === "h200" && modelSize === "big")) {
  flags.push(" --moe-a2a-backend deepep");
}
// 修复 dist-init-addr：multinode 时不覆盖已在 multiNodeFlags 中设置的地址
if (!isGB300 && !multinode) {
  flags.push(` --dist-init-addr 127.0.0.1:${distPort}`);
}
// 内存预算：提高 CG 批量大小至 128，mem-fraction-static 调至 0.9
if (hardware === "h200" && modelSize === "big") {
  flags.push(" --cuda-graph-max-bs 128");
  flags.push(" --mem-fraction-static 0.9");
}
```

评论区精华

review 仅有一条来自 gemini-code-assist[bot] 的评论，提出两点改进建议：

作者未回复，PR已合并，建议未被采纳。日期笔误虽不影响功能，但可能对后续读者造成困惑。

风险与影响

- 风险：极低。变更仅限于文档代码片段，不涉及运行时逻辑。唯一的轻微风险是 dist-init-addr 条件修改，如果在非 H200 big 的 multinode PD 场景下有未覆盖的边界情况，可能导致引导地址错误。但由于条件增加了 !multinode，反而使逻辑更安全。
- 影响：对 H200 集群用户，现在可以直接使用 h200|big|pd-disagg 配方，无需手动验证。文档生成的部署命令与已验证配置一致，减少用户试错成本。

关联脉络

本 PR 是 DeepSeek-V4 部署文档系列更新的一部分，与以下 PR 属于同一演进线：

- PR#23691：标记 GB300 多个配方验证状态并添加专属修复
- PR#23689：标记 B200/h200 small 配方验证状态
- PR#23698：调整 GB300 Pro PD 的 --mem-fraction-static

这些 PR 共同构成了 DeepSeek-V4 部署方案的持续验证和参数调优过程，反映了团队在多硬件平台、多部署模式下的系统性测试验证工作流。