PR #23689 完整报告

sgl-project/sglang

docs(DeepSeek-V4): mark b200|small|pd-disagg + h200|small|{cp,pd-disagg} verified

合并时间：2026-04-25 11:57

原文链接：http://prhub.com.cn/sgl-project/sglang/pull/23689

执行摘要

该 PR 为 DeepSeek-V4 部署文档标记了 B200 small PD-Disagg、H200 small CP 和 H200 small PD-Disagg 三种配方为已验证，同时引入了 TBD（To Be Done）机制，为尚未提供配方的组合输出友好占位符，并更新了 H200 镜像可用状态及 PD-Disagg 部署的权限注意事项。

功能与动机

根据 PR 提交信息，多个硬件 / 模型组合的端到端验证已在生产环境完成，需要在文档中标记为已验证以确保用户直接获得可运行的部署命令。作者还处理了部分未能提供配方的场景（如 h200|big|cp），通过添加 TBD 机制给出清晰提示，避免输出被注释掉的无效命令。

实现拆解

1. 标记已验证配方

在 docs_new/src/snippets/autoregressive/deepseek-v4-deployment.jsx 中，向 VERIFIED_RECIPES 集合新增了三个已验证配方：

- "b200|small|pd-disagg"
- "h200|small|cp"
- "h200|small|pd-disagg"

同时添加注释说明 h200|big|pd-disagg 仍在等待 4 节点 H200 集群验证。

2. 添加 TBD 机制

该文件还新建了 TBD_RECIPES 集合和 TBD_PLACEHOLDER 常量，用于标记 暂时无法提供命令的配方。在命令生成函数 generateCommand() 中，优先检查 TBD 集合，若命中则直接返回 "# to be provided" 占位符，而不是输出无效的注释后命令。

```
const TBD_RECIPES = new Set([
  "h200|big|cp",
]);
const TBD_PLACEHOLDER = "# to be provided";

// 在命令生成函数中：
if (TBD_RECIPES.has(verifyKey)) return TBD_PLACEHOLDER;
```

这一改动将“未提供”与“未验证”两种状态区分开，用户体验更好。

3. 更新 H200 文档注释

在 docs_new/cookbook/autoregressive/DeepSeek/DeepSeek-V4.mdx 中将 H200 镜像和检查点的描述从“即将推出”改为“已公开可用”，并补充了 PD-Disagg 部署的权限注意事项：

> PD-Disagg recipes on H200 may require docker run --privileged --ulimit memlock=-1 (or --device /dev/infiniband:/dev/infiniband --cap-add IPC_LOCK) so mooncake can discover the IB HCAs; without IB exposure mooncake silently falls back to TCP, which can lead to garbled KV transfer on large checkpoints.

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

主变更文件，更新已验证配方集合、新增 TBD 机制、调整命令生成逻辑

```
// docs_new/src/snippets/autoregressive/deepseek-v4-deployment.jsx
// 已验证配方集合，标记了端到端验证通过的硬件 | 模型大小 | 部署方案组合
const VERIFIED_RECIPES = new Set([
  "b200|small|low-latency",
  "b200|small|balanced",
  "b200|small|max-throughput",
  "b200|small|cp",
  "b200|small|pd-disagg", // 新增：B200 small PD-Disagg 已验证
  "b200|big|low-latency",
  "b200|big|balanced",
  "b200|big|max-throughput",
  "b200|big|cp",
  "h200|small|low-latency",
  "h200|small|balanced",
  "h200|small|max-throughput",
  "h200|small|cp", // 新增：H200 small CP 已验证
  "h200|small|pd-disagg", // 新增：H200 small PD-Disagg 已验证
  // h200|big|pd-disagg: pending verification (needs 4-node H200 cluster with
  // shared IB fabric: 2-node prefill + 2-node decode).
]);

// 配方命令暂未提供的集合（例如因上游限制），会显示友好占位符
const TBD_RECIPES = new Set([
  "h200|big|cp", // H200 big CP 暂时无法提供配方
]);
const TBD_PLACEHOLDER = "# to be provided";

// 生成命令主函数（片段）
const generateCommand = () => {
  // ... 各种 flags 组装 ...
  const verifyKey = `${hardware}|${modelSize}|${recipe}`;
  // 先检查是否是 TBD 配方，如果是则直接返回占位符
  if (TBD_RECIPES.has(verifyKey)) return TBD_PLACEHOLDER;
  // 正常检查已验证集合
  return VERIFIED_RECIPES.has(verifyKey)
    ? withMultinode
```

```
    : `${BEING_VERIFIED_NOTE}\n${commentOutCommand(withMultinode)}`;
};
```

评论区精华

无外部审核评论。PR 作者的多次 commit 展现了逐步完善的过程：

- 初始提交标记验证结果
- 随后补充 H200 CP/PD-Disagg 验证和笔记刷新
- 将特权说明措辞从“需要”改为“可能需要”
- 最后为 h200lbiglcp 添加 TBD 占位符

风险与影响

- 风险：极低。纯文档变更，不涉及任何运行时代码。主要风险在于验证结果是否准确，以及后续 checkpoint 变更是否需要更新文档状态，但文档已有回退机制和注释说明。
- 影响：用户可直接使用这些已验证的部署命令，减少验证成本；内部维护者可通过 TBD 机制清晰管理配方状态。

关联脉络

该 PR 与 #23690、#23691 同属 DeepSeek-V4 部署文档验证标记系列，覆盖了 B200、H200 和 GB300 三种硬件平台的验证状态标记。整体趋势是随着多节点、多方案验证逐步完成，文档从“待验证”状态逐步演进为“已验证”，并引入了 TBD 机制应对暂时无法提供配方的场景。

与近期的 #22998、#23682、#23671 等性能优化或 bugfix PR 对比，该 PR 属于轻量级文档维护，但作为基础设施文档的最终呈现，其准确性直接影响用户首次体验。