PR #23742 完整报告

sgl-project/sglang

docs(DeepSeek-V4): add h200|big verified recipes + tune H200 Pro parameters

合并时间：2026-04-26 12:44

原文链接：http://prhub.com.cn/sgl-project/sglang/pull/23742

执行摘要

本 PR 在 DeepSeek-V4 部署命令生成器中，将 H200 Pro (big) 的 low-latency、balanced、max-throughput 三种配方标记为“已验证”，并基于测试数据对其关键启动参数进行了调优。主要变更包括降低 DeepEP dispatch-token 上限、减少 CUDA graph 批量和最大运行请求数，以及提高内存占比参数。变更已合并，但 review 指出的注释和参数范围问题尚未修复，需在后续迭代中关注。

功能与动机

PR body 明确说明动机：“Mark h200|big|low-latency, h200|big|balanced, h200|big|max-throughput as verified”以及“Tune H200 Pro (big) parameters based on testing”。目的是让 H200 Pro 用户能够直接使用经过验证的部署命令，同时通过参数调优提升部署性能。

实现拆解

1. 标记已验证配方：在 VERIFIED_RECIPES Set 中添加 "h200|big|low-latency"、"h200|big|balanced"、"h200|big|max-throughput"，使这些配方在命令生成器中呈现为可运行的命令（而非注释掉的新版）。
2. 低延迟（low-latency）配方调优：
   - 将 --cuda-graph-max-bs 从 32 减少至 8，--max-running-requests 从 64 减少至 32。
   - 将 --mem-fraction-static 从 0.82 提升至 0.88（对所有 big 模型生效）。
3. balanced 配方调优：
   - 为 H200 Pro (big) 添加 --cuda-graph-max-bs 8 和 --max-running-requests 32。
   - 设置 --mem-fraction-static 0.88（仅 H200 Pro big）。
   - 将 DeepEP dispatch-token 上限从 256 降至 128。
4. max-throughput 配方调优：同样将 H200 Pro big 的 dispatch-token 上限从 256 降至 128。

以下是 low-latency 配方中参数调整的关键代码片段：

```
// low-latency 配方：H200 Pro (big) 使用更保守的并发参数
if (hardware === "h200" && isBig) {
  flags.push(" --cuda-graph-max-bs 8"); // 原为 32
  flags.push(" --max-running-requests 32"); // 原为 64
}
```

```
// 所有 big 模型的 mem-fraction-static 提升至 0.88
if (isBig) flags.push(" --mem-fraction-static 0.88"); // 原为 0.82
```

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

唯一的变更文件，包含了所有配方标记和参数调优逻辑，是整个 PR 的核心。

```
// 已验证配方集合：新增 H200 Pro (big) 的三种配方
const VERIFIED_RECIPES = new Set([
  // ... 原有条目省略
  "h200|big|low-latency", // 新增：H200 Pro low-latency 已验证
  "h200|big|balanced", // 新增：H200 Pro balanced 已验证
  "h200|big|max-throughput", // 新增：H200 Pro max-throughput 已验证
  "h200|big|pd-disagg", // 原有
  // ... 其余条目省略
]);

// balanced 配方：为 H200 Pro (big) 设置专属参数
if (recipe === "balanced") {
  if (hardware === "h200") {
    recipeEnv.push(isBig
      ? "SGLANG_DEEPEP_NUM_MAX_DISPATCH_TOKENS_PER_RANK=128" // H200 Pro big 使用
      128
      : "SGLANG_DEEPEP_NUM_MAX_DISPATCH_TOKENS_PER_RANK=256" // H200 small 使用
      256
    );
  } else {
    // 其他硬件保持原有逻辑
  }
}

// max-throughput 配方：类似 balanced 的调整
if (recipe === "max-throughput") {
  if (hardware === "h200") {
    recipeEnv.push(isBig
      ? "SGLANG_DEEPEP_NUM_MAX_DISPATCH_TOKENS_PER_RANK=128"
      : "SGLANG_DEEPEP_NUM_MAX_DISPATCH_TOKENS_PER_RANK=256"
    );
  } else {
    // 其他硬件保持原有逻辑
  }
}
```

评论区精华

- 注释与代码不一致：Copilot 指出 low-latency 配方的注释仍提及旧参数 cg=32 max-run=64，但代码已改为 cg=8 max-run=32；balanced 配方注释同样过时（写 cg=128 max-run=128，实际已改为 cg=8 max-run=32）。这些注释未更新，可能导致困惑。

- mem-fraction-static 改动范围过宽：Copilot 建议将 0.88 仅应用于 H200 Pro，而非所有 big 模型，以避免影响已验证的 B200/GB300 配方。该建议未被采纳。

风险与影响

- mem-fraction-static 影响其他平台：将 --mem-fraction-static 从 0.82 改为 0.88 对所有 big 模型的 low-latency 配方生效，可能影响 B200/GB300 等已验证平台的显存分配，存在 OOM 或性能回归风险。
- 注释过时：low-latency 和 balanced 配方的注释未同步更新，降低了代码可读性，可能误导阅读者。
- 无 CI 验证：PR 未触发运行测试，参数调优的稳定性依赖离线测试，缺少自动化回归保障。

关联脉络

本 PR 是 DeepSeek-V4 文档系列的第 3 次更新，与 #23715（标记 H200 big pd-disagg 已验证）和 #23725（添加 GB200 平台）同属一个演进线。此外，与 #23698（调整 GB300 Pro 的 mem-fraction-static）相呼应，显示团队正在系统性地为各硬件平台优化部署参数。