PR #23698 完整报告

sgl-project/sglang

docs(DeepSeek-V4): bump GB300 Pro PD decode --mem-fraction-static 0.83 → 0.9

合并时间：2026-04-25 16:35

原文链接：http://prhub.com.cn/sgl-project/sglang/pull/23698

执行摘要

将 DeepSeek-V4 GB300 Pro PD decode 角色的 --mem-fraction-static 默认值从 0.83 提升至 0.9，在避免 OOM 的前提下扩大 KV cache 容量，提升内存利用率。变更仅涉及部署指南中的参数值及注释，风险极低。

功能与动机

原默认值 0.83 在 GB300 Pro 上过于保守，导致显存利用不充分。通过 mem-fraction 扫描（0.83/0.87/0.89/0.91 均通过静态烟雾测试），发现 0.9 仍能安全运行，同时提供约 1M token 的 KV 缓存空间，每个 GPU 保留约 14 GB 用于 mooncake 传输和激活峰值，从而在不触发 OOM 的前提下提升解码吞吐。

实现拆解

步骤 1：修改参数值

在 docs_new/src/snippets/autoregressive/deepseek-v4-deployment.jsx 文件中，将 --mem-fraction-static 从 0.83 改为 0.9。

步骤 2：更新注释

补充 mem-fraction 扫描结果和选择 0.9 的技术理由，帮助用户理解为什么提升至此值。

变更代码片段（原注释简化为关键点）：

```
  if (isGB300 && modelSize === "big") {
    flags.push("  --max-running-requests 128");
    // mem-frac sweep 0.83/0.87/0.89/0.91 均通过静态烟雾测试;
    // 0.9 保留 ~14 GB/GPU 后 CG 余量，同时提供 ~1M token KV pool。
    flags.push("  --mem-fraction-static 0.9");
    flags.push("  --cuda-graph-max-bs 128");
  } else {
    flags.push("  --max-running-requests 256");
  }
```

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

包含所有变更：mem-fraction 参数值从 0.83 改为 0.9，以及相关注释更新。

```
  // ... 前置上下文
  if (isGB300 && modelSize === "big") {
    flags.push("  --max-running-requests 128");
```

```
    // 原为 0.83，经静态烟雾测试验证 0.83/0.87/0.89/0.91 均通过，
    // 0.9 可保留约 14 GB/GPU 后 CG 余量，同时提供约 1M token KV pool。
    flags.push("  --mem-fraction-static 0.9");
    flags.push("  --cuda-graph-max-bs 128");
  } else {
    flags.push("  --max-running-requests 256");
  }
  // ... 后续上下文
```

评论区精华

该 PR 无 review 评论。提交信息说明了静态烟雾测试覆盖范围。

风险与影响

- 风险：几乎不存在技术风险，因为这只是文档中的默认值调整，不修改任何代码逻辑。但极端 workload 下（如长序列、高并发）可能因 mem-fraction 提高而触发 OOM，不过注释已说明保留约 14 GB 余量，风险可控。
- 影响：影响范围为 GB300 Pro PD *decode* 角色的部署配置，提升 KV *cache* 容量，改善长序列解码性能，并减少用户手动调参成本。

关联脉络

本 PR 与近期多个 DeepSeek-V4 文档更新 PR（#23690、#23691、#23697）位于同一文件，持续优化 GB300 平台部署参数。该系列 PR 的演进方向是：在实验验证的基础上，逐步放宽保守的默认值，提升资源利用率和性能。