PR #26695 完整报告

sgl-project/sglang

[docs] Qwen3.5 cookbook: multi-node, MTP TP overrides, dense mamba flag

合并时间：2026-05-30 03:10

原文链接：http://prhub.com.cn/sgl-project/sglang/pull/26695

PR #26695 分析报告

执行摘要

此 PR 更新了 Qwen3.5 部署 cookbook，新增多节点部署支持（397B H100 BF16）、MTP 条件性 TP 覆盖、密集模型 mamba 调度策略标志，并调整了多个模型 / 硬件组合的 TP 和 mem 参数。同时修复了 React 状态残留导致错误发射 mamba 标志的 bug。变更主要集中在 JSX 配置生成器，并更新了 Docker 标签。

功能与动机

PR body 明确说明需要多节点部署支持以运行 397B 模型（TP=16 横跨 2 节点），并为 MTP 模式提供正确的 TP 覆盖（如 35B H100 BF16 MTP 开启时 TP=2）。此外，密集模型在 NVIDIA+MTP 时需 --mamba-scheduler-strategy extra_buffer，但原有条件门控在状态切换时会失效。commit 消息详细记录了所有调整。

实现拆解

1. 多节点支持：添加 multiNodeFlags 和 prependMultiNodeNote 辅助函数，在 397B H100 BF16 配置中设定 multinode: true, nnodes: 2，生成对应命令行参数。
2. MTP 条件性 TP 覆盖：对 35B/27B/122B 等模型，在 MTP 开启时合并基础配置与覆盖参数，覆盖中设置 mem: undefined 以跳过 --mem-fraction-static。
3. TP/mem 基础值调优：更新 122B、35B 等多个配置的 tp 和 mem 值，提高内存利用率。
4. 密集模型 mamba 标志：当硬件为 NVIDIA 且 MTP 开启时，强制设置 mambaCache 为 'v2'，并在规则中跳过条件检查，始终发射 --mamba-scheduler-strategy extra_buffer。
5. B300 CUDA 网格溢出规避：为 0.8B/2B BF16 添加 --max-running-requests 4064。
6. 状态残留修复：commit 2 修正了 mambaCache 状态残留，确保仅在 MTP 开启时设置。
7. Docker 标签更新：将 nightly-dev-* 改为 latest。

docs_new/src/snippets/autoregressive/qwen35-deployment.jsx

核心变更文件，实现多节点、MTP 覆盖、内存分数调整等所有逻辑配置。

```
// 辅助函数：生成多节点部署标志
const multiNodeFlags = (spec) => {
  // spec 包含 multinode, nnodes 等字段
  const flags = [];
  if (spec.multinode) {
```

```
    flags.push(`--nnodes ${spec.nnodes}`);
    flags.push('--node-rank $RANK'); // 实际使用中会被适当替换
    flags.push('--dist-init-addr $MASTER_ADDR:$MASTER_PORT');
  }
  return flags;
};

// 辅助函数：在命令前添加多节点说明
const prependMultiNodeNote = (spec) => {
  if (spec.multinode) {
    return `# Multi-node deployment: ${spec.nnodes} nodes required\n`;
  }
  return '';
};

// 在 emitFlags 中使用条件化 mem 发射（相关部分）
const emitFlags = (values) => {
  // ...
  // 当 spec 中无 mem 字段时，不发射 --mem-fraction-static
  if (spec.mem !== undefined) {
    flags.push(`--mem-fraction-static ${spec.mem}`);
  }
  // MTP 条件性 TP 覆盖：合并基础配置与 MTP 覆盖
  if (mtpEnabled && mtpOverrides[modelKey]) {
    spec = { ...spec, ...mtpOverrides[modelKey] };
    // 覆盖中可能设置 mem 为 undefined 以跳过发射
  }
  // ...
};
```

评论区精华

gemini-code-assist[bot] 指出的状态残留 bug 是核心讨论：当用户从 MoE+MTP 切换到密集 + 无 MTP 时，mambaCache 保持 'v2'，导致误加 --mamba-scheduler-strategy extra_buffer。解决方案（已实施）是只在 MTP 开启时强制设置。

风险与影响

- 风险：前端状态管理复杂，可能仍有边缘 case 导致错误参数；TP/mem 值基于特定 GPU 配置，用户环境差异可能引起 OOM 或性能回退。
- 影响：为用户提供准确的最新部署命令，减少配置错误；团队维护成本降低（与 DeepSeek-V4 对齐）。影响范围仅限文档，不涉及运行时。

关联脉络

此 PR 与 DeepSeek-V4 cookbook 的实现模式对齐（如 multiNodeFlags 等辅助函数），体现了文档代码共享的趋势。无直接关联的历史 PR。