PR #34357 完整报告

sgl-project/sglang

Update Qwen3.5 B200 NVFP4 MTP config

合并时间：2026-08-12 07:12

原文链接：http://prhub.com.cn/sgl-project/sglang/pull/34357

Qwen3.5 B200 NVFP4 MTP 配置更新

执行摘要

本次 PR 将 Qwen3.5-397B-A17B 在 B200 上使用 NVFP4 且开启 MTP 时的推荐部署配置从 tp=4 调整为 tp=2 + expert parallelism 2（TEP2），并同步更新交互式配置生成器与 cookboo文档。改动虽小，但直接决定用户生成的部署命令，属于一次有数据支撑的配置调优。

功能与动机

PR body 指出："tp=2 with expert parallelism 2 when MTP is enabled ... outperforms the current tp=4 setting across the concurrency sweep"，最佳配置参考了外部仓库 SemiAnalysisAI/InferenceX#2550 的基准结果。即对 MoE 模型在 MTP 场景下，TEP2 的并行切分优于纯 TP4。

实现拆解

1. 更新 docs/src/snippets/autoregressive/qwen35-deployment.jsx 顶部的 GPU 需求注释，标注 B200 NVFP4 在 MTP 下为 tp=2 ep=2。
2. 在 generateCommand() 中新增条件分支：当 model=397b、hardware=b200、quantization=fp4 且 speculative=enabled 时，将 hwConfig 覆盖为 { tp: 2, ep: 2, mem: 0.8 }，后续命令拼接逻辑会自动输出 --tp 2 与 --expert-parallel-size 2。
3. 同步修改 docs/cookbook/autoregressive/Qwen/Qwen3.5.mdx：规格表中 B200 行标注为 4 / 2 + EP2 (MTP)，并在部署说明中补充对应描述。
4. 无新增测试；纯文档与配置生成器改动，CI 通过。

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

配置生成器核心逻辑所在，新增了 B200 NVFP4 MTP 场景的 tp/ep 覆盖规则，直接决定用户生成的部署命令。

```
// 部署命令生成器核心逻辑（整理自 Qwen3.5 配置生成组件）：
// 根据用户选择的模型 / 硬件 / 量化 / 推测解码选项，输出 sglang serve 命令。
const generateCommand = () => {
  const { model, hardware, quantization, speculative } = values;

  // 1. 先取默认并行配置，若未定义则返回提示语
  let hwConfig = modelConfigs[model]?.[hardware]?.[quantization];
```

```
if (!hwConfig) {
  if (quantization === 'fp4') {
    return '# FP4 requires B200/B300 (Blackwell) and is only available for Qwen3.5-397B-
    A17B';
  }
  return '# Please select a valid hardware and quantization combination';
}

// 2. 针对 MTP 场景覆盖并行策略:
// 35B / 27B 在 H100 BF16 + MTP 时提升至 tp=2
if ((model === '35b' || model === '27b') && hardware === 'h100' && quantization === 'bf16'
&& speculative === 'enabled') {
  hwConfig = { ...hwConfig, tp: 2, mem: undefined };
}
// 122B 在 H100 FP8 + MTP 时提升至 tp=4
if (model === '122b' && hardware === 'h100' && quantization === 'fp8' && speculative ===
'enabled') {
  hwConfig = { ...hwConfig, tp: 4, mem: undefined };
}
// 397B 在 B200 NVFP4 + MTP 时改为 tp=2 + EP=2（TEP2），
// 基准显示该配置在并发扫描下优于 tp=4（参考 InferenceX PR#2550）
if (model === '397b' && hardware === 'b200' && quantization === 'fp4' && speculative ===
'enabled') {
  hwConfig = { ...hwConfig, tp: 2, ep: 2, mem: 0.8 };
}

// 3. 根据量化格式选择 checkpoint 名称
let modelName;
if (quantization === 'fp4') {
  // AMD MI355X 使用 MXFP4，Blackwell 使用 NVFP4-V2
  modelName = hardware === 'mi355x'
    ? 'amd/Qwen3.5-397B-A17B-MXFP4'
    : 'nvidia/Qwen3.5-397B-A17B-NVFP4-V2';
} else {
  const suffix = MODEL_SUFFIX[model];
  const quantSuffix = quantization === 'fp8' ? '-FP8' : '';
  modelName = 'Qwen/Qwen3.5-' + suffix + quantSuffix;
}

// 4. 组装命令：tp > 1 时追加 --tp，ep 存在时追加 --expert-parallel-size
let cmd = 'sglang serve --model-path ' + modelName;
if (hwConfig.tp > 1) {
  cmd += ' --tp ' + hwConfig.tp;
}
if (hwConfig.ep) {
  cmd += ' --expert-parallel-size ' + hwConfig.ep;
}
// …… 其余参数（如 --mem-fraction-static、多节点地址等）继续拼接
```

```
    return cmd;
  };
```

评论区精华

PR 没有 review 评论。作者在 Issue 评论区 @ 相关人员请求合并，说明配置已经实测验证（"It has been tested as in the description"）。

风险与影响

- 影响面窄：仅改变 397b + B200 + NVFP4 + MTP 组合的输出命令，其他组合不受影响。
- 一致性风险：jsx 生成器与 mdx 文档需同时更新，本 PR 已同步。
- 实测依据来自外部基准，若 SGLang 的 MoE 并行实现或显存策略变化，该配置可能不再最优。
- mem: 0.8 将在 tp=2 下保留 80% 显存给 KV cache，与之前 tp=4 的默认行为不同，使用者需注意显存余量。

关联脉络

仓库近期有大量 Qwen3.5 部署文档与配置相关工作（如 FP8、Ling 模型配方）。本 PR 的 TEP2 调优思路与 SemiAnalysisAI/InferenceX#2550 的外部基准关联，反映了 MoE 部署经验从纯 TP 向 TP+EP 组合演进的趋势；仓库内目前无直接关联 PR。