PR #27049 完整报告

sgl-project/sglang

docs: add DeepSeek-V4 EPLB Waterfill tips

合并时间：2026-06-03 15:37

原文链接：http://prhub.com.cn/sgl-project/sglang/pull/27049

PR #27049 分析报告

执行摘要

本 PR 在 DeepSeek-V4 cookbook 文档中新增了 EPLB + DeepEP Waterfill的配置指南，覆盖录制与回放流程、非 PD 与 PD 分离部署参数示例，以及 CUDA graph 和 MegaMoE 的兼容性约束。纯文档变更，无代码逻辑修改，对使用 DeepSeek-V4 进行 Expert Parallel 部署的团队有直接帮助。

功能与动机

为使用 DeepEP 的 EP 部署提供明确的 EPLB 和 Waterfill 启用指导。用户在高并发下专家路由不均衡时可参考文档中的命令行示例进行调优，并理解 --enable-deepep-waterfill 与 --moe-a2a-backend deepep 的依赖关系。

实现拆解

1. 新增独立小节：在 docs_new/cookbook/autoregressive/DeepSeek/DeepSeek-V4.mdx 中插入 EPLB + DeepEP Waterfill (Experimental) 章节。
2. 录制指南：引用已有文档，指导用户生成 expert_distribution_recorder_*.pt。
3. 命令行示例：分别给出非 PD 和 PD 分离场景的启动命令，标注 --moe-a2a-backend deepep、--deepep-mode 等关键参数。
4. 兼容性注释：明确 --deepep-mode normal 与 CUDA graph 不兼容，MegaMoE 不支持 Waterfill，并解释了原因。

docs_new/cookbook/autoregressive/DeepSeek/DeepSeek-V4.mdx

唯一变更文件，新增 EPLB + DeepEP Waterfill 配置小节，包含录制方法、命令行示例和兼容性约束。

```
<!-- docs_new/cookbook/autoregressive/DeepSeek/DeepSeek-V4.mdx -->

**EPLB + DeepEP Waterfill (Experimental)**

For EP deployments that use DeepEP, enable EPLB when routed experts become imbalanced under high concurrency. Use `--enable-deepep-waterfill` to route shared expert routing through DeepEP for load balancing.

For recorded/static EPLB reproduction, first record an expert-distribution file by following
```

[Capture expert selection distribution in MoE models](../../../docs/basic_usage/native_api.mdx#capture-expert-selection-distribution-in-moe-models).
Please checkout to latest main branch for this feature.

For non-PD reproduction, use:
```bash Command
--moe-a2a-backend deepep \
--deepep-mode auto \
--init-expert-location /path/to/expert_distribution_recorder_*.pt \
--enable-deepep-waterfill
```

For PD-Disagg reproduction, use normal mode on the prefill server and low_latency mode on the decode server. Add the same --init-expert-location flag to both commands:
```bash Command

## prefill

--moe-a2a-backend deepep \ --deepep-mode normal \ --init-expert-location /path/to/expert_distribution_recorder_*.pt \ --enable-deepep-waterfill

## decode

--moe-a2a-backend deepep \ --deepep-mode low_latency \ --init-expert-location /path/to/expert_distribution_recorder_*.pt \ --enable-deepep-waterfill You can also add `--ep-num-redundant-experts` and `--eplb-algorithm` to customize EPLB placement. MegaMoE is not supported with this DeepEP Waterfill recipe yet. Waterfill routes the shared expert through DeepEP for load balancing, so `--enable-deepep-waterfill` requires `--moe-a2a-backend deepep`.

## prefill: --moe-a2a-backend deepep --deepep-mode normal --init-expert-location ... --enable-deepep-waterfill

## decode: --moe-a2a-backend deepep --deepep-mode low_latency --init-expert-location ... --enable-deepep-waterfill

MegaMoE is not supported with this recipe yet. Waterfill routes the shared expert through DeepEP for load balancing, so `--enable-deepep-waterfill` requires `--moe-a2a-backend deepep`.

### 评论区精华

- gemini-code-assist[bot]：建议将 "add EPLB" 改为 "enable EPLB"，并说明 --deepep-mode normal 与 CUDA graph 不兼容。
- Fridge003：要求小节标题标注 Experimental、链接指向 /docs_new、询问 MegaMoE 兼容性。
- xutizhou：通过实测确认 MegaMoE 不兼容，并补充了原理说明。
```

风险与影响

风险：无代码变更，技术风险极低。但用户若忽略文档中的实验性标注和约束（如 CUDA graph 不兼容），可能在生产环境中遇到问题。

影响：仅影响查阅 DeepSeek-V4 文档的用户，提供更完整的配置参考，无运行时行为变更。

关联脉络

此 PR 与近期多个 MoE/DeepSeek 相关的 PR（如 #25556、#25655、#26349）均涉及 MoE 推理优化，但本 PR 仅聚焦文档补充，无代码关联。