

PR #34654 完整报告

sgl-project/sglang

[Docs] Add decode context parallelism to advanced features

合并时间: 2026-08-14 04:51

原文链接: <http://prhub.com.cn/sgl-project/sglang/pull/34654>

执行摘要

- 一句话: 新增 DCP 高级功能文档页并登记导航与参数
- 推荐动作: 值得对 DCP 感兴趣的架构与推理团队精读, 尤其是拓扑约束、LSE 合并与通信后端对比部分; 对运维而言可直接复制命令示例。需要留意页面警告的启动检查缺口, 文档描述的约束强于当前代码校验, 建议后续在实现中补齐。

功能与动机

PR body 说明目的是在 Advanced Features 下新增 DCP 页面, 覆盖 MLA KV 条带化、LSE 合并、通信后端以及与 DPA、DSpark、PD、HiCache L2 的组合, 并登记侧边栏、概述页和三个新 server 参数。文档化填补了用户启用 DCP 时的操作指南空白, 属于功能实现后的配套文档补齐。

实现拆解

1. 新增 DCP 主文档页

在 `docs/docs/advanced_features/dcp.mdx` 新增约 180 行内容: 先介绍 DCP 的条带化模型 ($\text{owner}(p) = p \bmod c$) 和 LSE 精确合并原理, 再给出启用参数表、拓扑约束、命令行示例, 随后展开虚拟 KV 布局、MLA decode 数据流、Q 投影复制和四种通信后端对比, 最后说明与投机解码、PD 分解、DPA、HiCache 的组合规则。

2. 注册导航与重定向

`docs/docs.json` 添加 `/advanced_features/dcp.html` 到 `/docs/advanced_features/dcp` 的重定向, 并在 Advanced Features 侧边栏 `quantized_kv_cache` 与 `expert_parallelism` 之间插入 `docs/advanced_features/dcp` 条目, 保证旧 URL 能跳转且新页面可导航。

3. 补充 server 参数文档

`docs/docs/advanced_features/server_arguments.mdx` 新增三行表格对应 `--dcp-size` (含别名 `--decode-context-parallel-size`)、`--dcp-comm-backend` (`ag_rs`、`a2a`、`fi_a2a`) 和 `--dcp-replicate-q-proj`, 并给出默认值和取值说明; 同时在概览章节加入入口提示。

4. 更新 overview 页面

`docs/docs/advanced_features/overview.mdx` 在 Expert Parallelism 与 LoRA 之间加入 Decode Context Parallelism 链接一行业。

5. 提交演进与校验

4 次提交依次完成初稿、裁剪为操作员向内容、澄清投机解码段（删除 DSpark 术语）、删除 `tokenspeed_mla` 并链接 DCP/Helix 追踪 issue；PR 使用 `mint validate` 校验，并由 Mintlify 生成预览。

关键文件：

- `docs/docs/advanced_features/dcp.mdx`（模块 文档站；类别 docs；类型 core-logic）：新增约 180 行的 DCP 核心文档，覆盖 KV 条带化、LSE 合并、通信后端、拓扑约束与组合场景，是本次 PR 的主体。
- `docs/docs.json`（模块 站点配置；类别 config；类型 configuration）：为 DCP 页面添加重定向并在 Advanced Features 侧边栏注册导航条目，决定页面是否可被发现。
- `docs/docs/advanced_features/server_arguments.mdx`（模块 文档站；类别 docs；类型 core-logic）：新增 `--dcp-size`、`--dcp-comm-backend`、`--dcp-replicate-q-proj` 的官方参数说明，是操作者查询命令行开关的唯一入口。
- `docs/docs/advanced_features/overview.mdx`（模块 文档站；类别 docs；类型 core-logic）：在 Advanced Features 总览列表中加入 DCP 入口，保证从 overview 页面也可进入新文档。

关键符号：未识别

评论区精华

该 PR 没有人工 review 评论；唯一外部交互是 mintlify[bot] 生成的预览部署通知（lmsysorg docs preview）。设计取舍主要通过 4 次提交历史表达：初稿包含成本公式与测试清单，随后裁剪为 operator 向内容；投机解码段落改写为“draft KV 全量复制 + DCP 感知 MLA kernel”的描述；最终移除 `tokenspeed_mla` 并指向 DCP/Helix 追踪 issue。

- 暂无高价值评论线程

风险与影响

- 风险：主要风险不是运行时回归，而是文档与实现的同步：页面临时警告启动检查只校验 `tp_size % dcp_size == 0`，比文档要求的“DCP 组嵌套在 attention-TP/DP 副本内”弱，用户按 DPA+DCP 组合部署时可能遇到未被启动期拦截的坏拓扑；`--dcp-comm-backend` 选择、Kimi K3 默认行为、`cutedsl_mla` 后端等描述依赖实现演进，容易过时；页面中 mermaid 图和内部链接（DPA、PD、HiCache、Kimi K3 cookbook）缺少自动化校验，改路径后可能 404。纯文档变更，不涉及代码回归、性能或安全面。
- 影响：对用户：DCP 从此有官方操作文档，包含示例配置、拓扑约束和组合矩阵，可显著降低 MLA 长上下文部署的试错成本。对系统与代码：无任何运行时文件改动，服务行为不变。对团队：新增一个需要随 DCP 实现一起维护的文档页面，后续功能演进（如 Helix 追踪项落地、启动期强校验）需同步刷新。
- 风险标记：文档与实现同步风险，启动检查弱于文档约束，链接缺少自动化校验

关联脉络

- PR #34766 [Fix] Carry the backend on Kimi-K3 deferred preprocessing configs: 同一功能线: DCP 文档页针对 Kimi K3 专门说明默认后端选择与 replicated Q 行为, 二者共同支撑 Kimi K3 的 MLA 长上下文部署。
- PR #34692 [PD] Add the missing Prefill bootstrap timeout for NIXL: 同一功能方向: DCP 文档描述 DCP×PD 组合中 Mooncake 与 NIXL 的传输计划; #34692 修复 NIXL bootstrap 超时, 是 DCP×PD 场景稳定性的前置修复。