PR #22329 完整报告

sgl-project/sglang

[AMD] Add prealloc token env for mori-ep

合并时间：2026-04-10 00:34

原文链接：http://prhub.com.cn/sgl-project/sglang/pull/22329

执行摘要

本 PR 为 AMD MORI-EP 新增环境变量 SGLANG_MORI_PREALLOC_MAX_RECV_TOKENS，允许用户配置接收令牌的预分配数量，以在内存占用和缓冲区溢出风险间取得平衡。同时引入向后兼容性检查函数，并更新相关文档。变更主要影响 AMD 平台的 MoE 调度模块，为内存敏感场景提供调优能力。

功能与动机

根据 PR body，主要动机是：

1. 添加环境变量：SGLANG_MORI_PREALLOC_MAX_RECV_TOKENS，允许用户自定义 MORI-EP 的令牌预分配数量。
2. 提供向后兼容性：通过 check_mori_compatibility 函数处理不同 MORI 库版本的参数兼容性。
3. 文档修复：更新环境变量文档和服务器参数帮助文本。

这解决了用户需要灵活控制内存占用与性能平衡的需求，特别是在内存受限的 AMD 硬件环境中。

实现拆解

实现涉及三个文件：

文件	关键变更	说明
python/sglang/srt/layers/moe/token_dispatcher/moriep.py	新增 check_mori_compatibility 函数；在 init_mori_op 中读取环境变量并传递给 MORI 配置。	核心逻辑，通过动态反射检查 MORI 库兼容性，并集成新环境变量。
docs/references/environment_variables.md	新增环境变量条目：SGLANG_MORI_PREALLOC_MAX_RECV_TOKENS。	提供使用指南：有效范围 1 到 world_size*SGLANG_MORI_NUM_MAX_DISPATCH_TOKENS_PER_RANK，默认 0 表示最大值，设置过小可能导致缓冲区溢出。

文件	关键变更	说明
python/sglang/srt/server_args.py	修正帮助文本，将“DeepEP MoE”扩展为“DeepEP or MoriEP MoE”。	次要更新，反映更广泛的 MoE 支持。

关键代码片段：

```
def check_mori_compatibility(kwargs: dict) -> None:
    """Remove kwargs not accepted by the installed mori's EpDispatchCombineConfig."""
    import dataclasses
    config_cls = mori.ops.EpDispatchCombineConfig
    valid_kwargs = {f.name for f in dataclasses.fields(config_cls)}
    invalid_kwargs = set(kwargs.keys()) - valid_kwargs
    for arg in invalid_kwargs:
        logger.warning(f"[MORI compat] Removing incompatible argument {arg} ")
        del kwargs[arg]
```

评论区精华

review 讨论较少，但包含关键点：

> @billishyahao Can you leave comments in code/doc on how SGLANG_MORI_PREALLOC_MAX_RECV_TOKENS should be used and default?

作者通过提交补充了文档描述，明确了环境变量的作用、默认值和风险警告，解决了 reviewer 的疑虑。

风险与影响

风险：

1. 兼容性风险：check_mori_compatibility 依赖运行时反射，可能在不同 MORI 版本中行为不一致。
2. 配置风险：环境变量设置过小可能导致缓冲区溢出，文档已警告但缺乏运行时验证。
3. 回归风险：修改 init_mori_op 参数传递逻辑，可能影响现有功能，但 UT 测试通过提供了基本保障。

影响：

- 用户：AMD 平台用户获得内存调优能力，需自行权衡内存占用与溢出风险。
- 系统：扩展了 MoE 调度配置灵活性，增强平台适应性。
- 团队：需关注 MORI 库版本演进，确保兼容性逻辑稳定。

关联脉络

与近期 PR 的关联：

- PR 22424：同属 AMD 平台优化，涉及性能调优和内核调度。
- PR 20089：同属 MoE 分发路径（EP）相关功能，本 PR 扩展了 MORI-EP 支持。
- PR 22335：同属 AMD 平台问题修复，涉及兼容性处理。

这反映了 sglang 仓库在 AMD 平台上持续优化 MoE 和调度能力的趋势，特别是内存管理和性能调优方向。