PR #30312 完整报告

sgl-project/sglang

[NPU]Add support --pre-warm-nccl

合并时间：2026-07-07 17:17

原文链接：http://prhub.com.cn/sgl-project/sglang/pull/30312

执行摘要

- 一句话：为 NPU 平台添加 HCCL 预热支持
- 推荐动作：建议精读：该 PR 展示了如何以最小改动扩展硬件支持，值得关注 _handle_nccl_pre_warm 的通用适配模式。

功能与动机

当使用多 GPU 张量并行时，首次集体通信操作会触发 HCCL 通信器初始化，导致前 2-3 个请求的 P99 TTFT 严重退化。本 PR 在服务启动期间预暖 NCCL/RCCL/HCCL，以消除冷启动延迟。

实现拆解

1. 更新 _handle_nccl_pre_warm 方法（server_args.py）：将平台检查条件从 not (is_cuda() or is_hip()) 扩展为 not (is_cuda() or is_hip() or is_npu())，当检测到 NPU 时不再禁用预暖功能。
2. 更新日志与注释（model_runner.py）：将日志和注释中的 "NCCL/RCCL" 更新为 "NCCL/RCCL/HCCL"，以准确反映支持范围。预暖逻辑本身（一个 all_reduce + synchronize）保持不变。

关键文件：

- python/sglang/srt/server_args.py（模块 配置管理；类别 source；类型 core-logic；符号 _handle_nccl_pre_warm）：核心逻辑变更：修改 _handle_nccl_pre_warm 白名单以包含 NPU。
- python/sglang/srt/model_executor/model_runner.py（模块 模型执行器；类别 source；类型 data-contract）：预暖代码位置声明性变更：注释和日志中增加 "HCCL" 字样。

关键符号：_handle_nccl_pre_warm, init_torch_distributed

关键源码片段

python/sglang/srt/server_args.py

核心逻辑变更：修改 _handle_nccl_pre_warm 白名单以包含 NPU。

```
# python/sglang/srt/server_args.py
```

```
def _handle_nccl_pre_warm(self):
    # 原本只允许 CUDA 或 HIP；现在也允许 NPU（Ascend）
    if self.pre_warm_nccl and not (is_cuda() or is_hip() or is_npu()):
        logger.warning(
            "pre_warm_nccl is only applicable for CUDA or HIP hardware or NPU hardware. "
            "Ignoring pre_warm_nccl setting on current hardware."
        )
        self.pre_warm_nccl = False
```

python/sglang/srt/model_executor/model_runner.py

预暖代码位置声明性变更：注释和日志中增加 "HCCL" 字样。

```
# python/sglang/srt/model_executor/model_runner.py

# Pre-warm NCCL/RCCL/HCCL to eliminate cold-start latency in first request
# Controlled by --pre-warm-nccl flag (default: enabled on AMD GPUs)
if self.server_args.pre_warm_nccl and (
    self.tp_size > 1 or self.pp_size > 1 or self.moe_ep_size > 1
):
    warmup_start = time.perf_counter()
    tp_group_handle = get_tp_group().device_group

    # Single warmup all_reduce to initialize NCCL/RCCL/HCCL communicator
    warmup_tensor = torch.zeros(1, device=torch.cuda.current_device())
    dist.all_reduce(warmup_tensor, group=tp_group_handle)
    current_platform.synchronize()

    warmup_elapsed = time.perf_counter() - warmup_start
    logger.info(
        f"NCCL/RCCL/HCCL warmup completed in {warmup_elapsed:.3f}s "
        f"(tp_size={self.tp_size}, pp_size={self.pp_size}, ep_size={self.moe_ep_size})"
    )
```

评论区精华

Reviewer iforgetmyname 在 server_args.py 上评论 "NPU hardware"，作者 loading66 回复 "modified" 并修正了警告消息从 "Ascend hardware" 改为 "NPU hardware"。

- NPU 白名单条件 (correctness): 将警告消息中的 'Ascend hardware' 修改为 'NPU hardware'，使文案与 is_npu() 一致。

风险与影响

- 风险：风险极低：变更仅涉及平台白名单和日志文本，不修改预暖核心逻辑。需确认 NPU 上 is_npu() 函数正确返回 True，否则预暖将被跳过。
- 影响：对用户：Ascend NPU 用户启用 --pre-warm-nccl 后，可显著降低 P99 TTFT。对系统：无功能影响，仅优化延时。
- 风险标记：暂无

关联脉络

- 暂无明显关联 PR