PR #20234 完整报告

sgl-project/sglang

Use device-agnostic helpers for Mamba tests and core ops

合并时间：2026-05-01 07:14

原文链接：http://prhub.com.cn/sgl-project/sglang/pull/20234

执行摘要

- 一句话：Mamba 测试与内核设备无关化以支持 XPU
- 推荐动作：值得关注 get_device、get_device_count 等辅助函数的设计模式，可作为后续设备无关化重构的参考。

功能与动机

Replace hardcoded CUDA device checks with device-agnostic helpers (get_device, get_device_count, torch.get_device_module) to support both CUDA and XPU devices.

实现拆解

1. 在三个测试文件中，将 torch.cuda.is_available() 替换为 get_device() + 允许设备列表 ["cuda", "xpu"] 检查；将 torch.cuda.device_count() 替换为 get_device_count()。
2. 将测试函数和设备生成函数的默认参数从 device="cuda" 改为 device=None，并在函数内调用 get_device() 解析实际设备。
3. 在多 GPU 测试中，将 torch.cuda.set_device() 替换为 torch.get_device_module(device).set_device()。
4. 在五个内核操作文件（layernorm_gated.py、mamba_ssm.py、ssd_bmm.py、ssd_chunk_state.py、ssd_state_passing.py）中，将 torch.cuda.device() 上下文管理器替换为 torch.get_device_module(device).device()。
5. 在测试文件中新增导入 from sglang.srt.utils import get_device, get_device_count。

关键文件：

- test/registered/layers/mamba/test_mamba_ssm_ssd.py（模块 SSD 测试；类别 test；类型 test-coverage；符号 generate_random_inputs）：测试文件改动最大，替换了所有设备检查和默认参数，并修改了 generate_random_inputs 函数签名
- test/registered/layers/mamba/test_mamba_ssm.py（模块 SSM 测试；类别 test；类型 test-coverage）：测试文件，替换了三个测试函数中的 CUDA 可用性检查和默认设备
- test/registered/layers/mamba/test_mamba2_mixer.py（模块 Mixer 测试；类别 test；类型 test-coverage）：测试文件，替换了多 GPU 测试中的设备可用性、设备计数和 set_device

- python/sglang/srt/layers/attention/mamba/ops/ssd_chunk_state.py（模块 SSD 块状态；类别 infra；类型 infrastructure）：内核操作文件，替换了三处 torch.cuda.device 上下文管理器
- python/sglang/srt/layers/attention/mamba/ops/layernorm_gated.py（模块 层归一化；类别 infra；类型 infrastructure）：内核操作文件，替换了一处 torch.cuda.device 上下文管理器
- python/sglang/srt/layers/attention/mamba/ops/mamba_ssm.py（模块 Mamba SSM；类别 infra；类型 infrastructure）：内核操作文件，替换了一处 torch.cuda.device 上下文管理器
- python/sglang/srt/layers/attention/mamba/ops/ssd_bmm.py（模块 SSD 批乘；类别 infra；类型 infrastructure）：内核操作文件，替换了一处 torch.cuda.device 上下文管理器
- python/sglang/srt/layers/attention/mamba/ops/ssd_state_passing.py（模块 SSD 状态传递；类别 infra；类型 infrastructure）：内核操作文件，替换了一处 torch.cuda.device 上下文管理器

关键符号：generate_random_inputs, test_selective_state_update, test_selective_state_update_with_batch_indices, test_mixer2_gated_norm_multi_gpu

关键源码片段

test/registered/layers/mamba/test_mamba_ssm_ssd.py

测试文件改动最大，替换了所有设备检查和默认参数，并修改了 generate_random_inputs 函数签名

```
from sglang.srt.utils import get_device

def generate_random_inputs(batch_size, seqlen, n_heads, d_head, itype, device=None):
    """生成随机输入张量，device 为 None 时由 get_device() 自动选择。"""
    if device is None:
        device = get_device() # 自动适配当前设备（CUDA 或 XPU）
    # 仅允许在 CUDA 或 XPU 上运行，否则跳过测试
    if device not in ["cuda", "xpu"]:
        pytest.skip("Test only supports CUDA and XPU devices")
    torch.manual_seed(0)
    A = -torch.exp(torch.rand(n_heads, dtype=itype, device=device))
    dt = F.softplus(torch.randn(batch_size, seqlen, n_heads, dtype=itype, device=device) - 4)
    X = torch.randn((batch_size, seqlen, n_heads, d_head), dtype=itype, device=device)
    B = torch.randn((batch_size, seqlen, n_heads, d_head), dtype=itype, device=device)
    C = torch.randn((batch_size, seqlen, n_heads, d_head), dtype=itype, device=device)
    return A, dt, X, B, C

def generate_continuous_batched_examples(..., device=None, ...):
    A, dt, X, B, C = generate_random_inputs(
        num_examples, full_length, n_heads, d_head, itype, device
    )
```

```
device = X.device # 从生成的张量捕获实际设备，确保后续一致
```

评论区精华

PR 经历了两次 CI 检查，由 mingfeima 批准后合并。没有收到实质性 review 评论或设计争论。

- 暂无高价值评论线程

风险与影响

- 风险：风险极低：所有替换都是机械的 API 映射，不改变原有计算逻辑。潜在风险包括：如果 get_device() 在特定环境返回意外值，可能导致测试在原本可用的设备上被跳过；XPU 支持仅通过允许列表实现，未覆盖所有可能后端，但已有 CI 覆盖 CUDA 和 AMD。
- 影响：对现有 CUDA 用户无影响；XPU 用户可以运行 Mamba 测试和依赖这些内核的模型；整体提升了代码可移植性和多后端支持能力。影响范围限于 Mamba 相关测试和内核模块。
- 风险标记：缺少 XPU 实际测试覆盖，get_device 行为依赖环境变量

关联脉络

- PR #24297 Rename SGLANG_USE_JIT_ALL_REDUCE to SGLANG_OPT_USE_CUSTOM_ALL_REDUCE_V2: 同样涉及环境变量和设备抽象的重构，属于多后端适配的连续工作