PR #24222 完整报告

sgl-project/sglang

[CI] Restore SMG *e2e* on 2-gpu-h100 / 4-gpu-h100 runners

合并时间：2026-05-02 14:55

原文链接：http://prhub.com.cn/sgl-project/sglang/pull/24222

执行摘要

- 一句话：恢复 SMG *e2e* 测试到 H100 GPU 集群并修复 gRPC 兼容性
- 推荐动作：此 PR 值得精读，尤其是 setup_backend.py 中关于 pytest-parallel 兼容性问题的分析和 grpc_server.py 中向后兼容的优雅处理方式。展示了如何处理上游库的不兼容性并保持功能降级路径，对 CI 和测试基础设施维护者很有参考价值。

功能与动机

原有的 4-gpu-a10 运行器镜像存在 transformers/kernels 兼容性问题，导致所有 SGLang 工作进程 import 崩溃，于 PR #24166 中通过临时跳过所有工作进程相关测试来禁用 SMG *e2e* 测试。此 PR 旨在将 SMG *e2e* 套件迁移到健康的 2-gpu-h100 和 4-gpu-h100 运行器上，并修复迁移过程中发现的 gRPC 启动兼容性问题（#22500 引入的新参数不被 smg-grpc-servicer ≤ 0.5.2 接受）。

实现拆解

1. CI 工作流调整：在 .github/workflows/pr-test-rust.yml 中将 build-wheel 从 4-gpu-a10 移到 ubuntu-latest（仅 CPU 编译），将 gateway-*e2e*、benchmarks、*e2e*、chat-completions 等矩阵条目指向 2-gpu-h100 或 4-gpu-h100。添加基于运行器类型的作业级并发锁，避免 GPU 资源冲突。保留 responses 条目为禁用占位符，便于追踪覆盖缺口。
2. gRPC 兼容性修复：在 python/sglang/srt/entrypoints/grpc_server.py 中使用 inspect.signature 检测 _serve_grpc 是否接受 on_request_manager_ready 参数。若不接受（如 smg-grpc-servicer ≤ 0.5.2），则不传入该钩子，导致 HTTP sidecar（Prometheus metrics 和 profiling 端点）不启动。当用户显式启用 --enable-metrics 时，则抛出 RuntimeError 以避免静默丢失 metrics；否则仅记录警告。核心 gRPC 服务不受影响。
3. 测试夹具作用域变更：在 sgl-model-gateway/*e2e*_test/fixtures/setup_backend.py 中将 setup_backend 夹具从 scope="class" 改为 function 作用域。原因是 pytest-parallel 对非 function 作用域的 finalize 处理有 bug，导致类级夹具的 teardown 从未执行，从而模型池引用泄漏并死锁。改为 function 作用域后，每次测试结束后都能正确释放资源。
4. GPU 资源管理增强：在 sgl-model-gateway/*e2e*_test/fixtures/hooks.py 中添加了 _count_gpus_without_cuda 函数，通过 nvidia-smi 获取可用 GPU 数量，并正确解析 CUDA_VISIBLE_DEVICES 环境变量（包括 -1 表示零 GPU）。在 pytest_collection_modifyitems 中，为 GPU 需求超过可用量的测试添加 pytest.mark.skip

；同时修剪模型池中 TP 超过可用 GPU 的键。在 pytest_collection_finish 中，若所有测试均因 GPU 原因被跳过，则抛 UsageError，避免运行器配置错误时静默通过。

5. 辅助工具改进：在 sgl-model-gateway/e2e_test/infra/process_utils.py 中重构了 detect_ib_device 函数，从 /sys/class/infiniband 枚举设备，优先选择 mlx5_ib*（原生 IB）而非 mlx5_eth*（RoCE），避免 PD 模式下 KV 传输失败。在 gpu_allocator.py 中，get_open_port 添加了端口上限 55536 的检查，防止 --grpc-mode 下端口溢出。

关键文件：

- python/sglang/srt/entrypoints/grpc_server.py（模块 gRPC 服务；类别 source；类型 dependency-wiring；符号 serve_app, _on_request_manager_ready）：修复了 gRPC 启动时因 #22500 引入的 TypeError：旧版 smg-grpc-servicer 不接受 on_request_manager_ready 参数。通过 inspect.signature 检测兼容性并优雅降级。
- sgl-model-gateway/e2e_test/fixtures/setup_backend.py（模块 测试夹具；类别 test；类型 test-coverage；符号 setup_backend, _setup_pd_backend）：将 setup_backend 夹具从 class 作用域改为 function 作用域，并添加 try/finally 确保资源释放。修复了 pytest-parallel 下的模型池引用泄漏死锁。
- sgl-model-gateway/e2e_test/fixtures/hooks.py（模块 测试钩子；类别 test；类型 test-coverage；符号 _count_gpus_without_cuda, pytest_collection_modifyitems, pytest_collection_finish）：添加 GPU 资源校验逻辑：通过 nvidia-smi 计数 GPU，跳过超规格测试，修剪不适配模型，并在全跳过时 loud-fail。引入 CUDA_VISIBLE_DEVICES 正确解析。
- sgl-model-gateway/e2e_test/infra/process_utils.py（模块 进程工具；类别 test；类型 test-coverage；符号 detect_ib_device）：重构 InfiniBand 设备检测函数，优先选择原生 IB 设备（mlx5_ib）而非 RoCE 设备（mlx5_eth），避免 PD 模式下 KV 传输失败。
- .github/workflows/pr-test-rust.yml（模块 CI 配置；类别 infra；类型 infrastructure）：CI 工作流核心调整：移动矩阵条目到新运行器，添加作业级并发锁，移除 pytest-parallel，调整环境配置。
- sgl-model-gateway/e2e_test/infra/gpu_allocator.py（模块 GPU 分配；类别 test；类型 test-coverage；符号 get_open_port）：改进 get_open_port 函数，添加端口上限 55536 检查，防止 grpc_port 溢出 16 位范围。
- sgl-model-gateway/e2e_test/chat_completions/test_function_calling.py（模块 函数调用测试；类别 test；类型 test-coverage；符号 test_multi_tool_scenario_required）：跳过 Mistral 的 test_multi_tool_scenario_required 测试，因 SMG 路由器解析 tool_choice='required' 输出时有已知 bug。

关键符号：serve_app, _on_request_manager_ready, setup_backend, _setup_pd_backend, pytest_collection_modifyitems, _count_gpus_without_cuda, detect_ib_device, get_open_port, test_multi_tool_scenario_required

关键源码片段

python/sglang/srt/entrypoints/grpc_server.py

修复了 gRPC 启动时因 #22500 引入的 TypeError：旧版 smg-grpc-servicer 不接受 on_request_manager_ready 参数。通过 inspect.signature 检测兼容性并优雅降级。

```
import inspect

    async def _on_request_manager_ready(request_manager, srv_args, sched_info):
        # 内部实现：启动 sidecar 和管理端点
        ...

    # 检查已安装的 smg-grpc-servicer 是否接受 on_request_manager_ready 参数
    # 旧版本 (≤ 0.5.2) 只接受 (server_args, model_info),
    # 会拒绝该关键字参数导致 TypeError
    serve_kwargs: dict = {}
    sidecar_supported = (
        "on_request_manager_ready" in inspect.signature(_serve_grpc).parameters
    )
    if sidecar_supported:
        # 新版本 servicer：正常传入钩子，启动 HTTP sidecar
        serve_kwargs["on_request_manager_ready"] = _on_request_manager_ready
    elif server_args.enable_metrics:
        # 用户明确要求 metrics，但 servicer 不支持 sidecar
        # 失败 loudly，而不是静默产生一个没有 /metrics 的服务器
        raise RuntimeError(
            "--enable-metrics requires smg-grpc-servicer ≥ 0.5.3 "
            "(the version that accepts 'on_request_manager_ready'); "
            "installed version lacks the hook so the HTTP sidecar "
            "would never start. Upgrade smg-grpc-servicer or remove "
            "--enable-metrics."
        )
    else:
        # 无 metrics 要求，静默降级：无 sidecar，仅记录警告
        logger.warning(
            "Installed smg-grpc-servicer does not accept "
            "'on_request_manager_ready'; HTTP sidecar disabled "
            "(no /metrics, /start_profile, /stop_profile). "
            "Upgrade smg-grpc-servicer to ≥ 0.5.3 to enable it."
        )

    try:
        # 最终调用 _serve_grpc，根据兼容性传入或不传入钩子
        await _serve_grpc(server_args, model_info, **serve_kwargs)
    finally:
        if sidecar_runner is not None:
            # 清理 sidecar（如果启动过）
            try:
                await sidecar_runner.cleanup()
            except Exception as e:
                logger.exception(
                    "Failed to cleanly shut down HTTP sidecar server: %s", e
```

```
            )
```

sgl-model-gateway/e2e_test/fixtures/hooks.py

添加 GPU 资源校验逻辑：通过 nvidia-smi 计数 GPU，跳过超规格测试，修剪不适配模型，并在全跳过时 loud-fail。引入 CUDA_VISIBLE_DEVICES 正确解析。

```
available_gpus = _count_gpus_without_cuda()

for item in items:
    # ... 提取 model_id, test_gpus ...

    # Mark over-capacity tests as skipped 当可用 GPU 不足时
    # 包括 available_gpus == 0 的情况，以便 collection_finish 检测全跳过
    if test_gpus > available_gpus:
        item.add_marker(
            pytest.mark.skip(
                reason=(
                    f"requires {test_gpus} GPUs (model={model_id}, "
                    f"tp={MODEL_SPECS.get(model_id, {}).get('tp', 1)}); "
                    f"only {available_gpus} available on this runner"
                )
            )
        )

# Prune workers that can never launch on this runner
for key in list(_worker_counts.keys()):
    spec = MODEL_SPECS.get(key[0], {})
    if spec.get("tp", 1) > available_gpus:
        del _worker_counts[key]
_first_seen_order[:] = [k for k in _first_seen_order if k in _worker_counts]

# ... 日志输出 ...

# pytest_collection_finish 中检测全跳过情况
if _max_test_gpu_requirement > 0 and all(
    "GPU" in item.get_closest_marker("skip").args[0] # 伪代码，实际检查 skip 原因
    for item in items
):
    pytest.fail("All tests skipped due to insufficient GPUs")
```

评论区精华

- gemini-code-assist[bot]指出 if available_gpus > 0 and test_gpus > available_gpus: 中的 > 0 守卫在 available_gpus == 0 时阻止了测试跳过，使得 pytest_collection_finish 无法检测全跳过情况。建议移除守卫。已采纳。
- gemini-code-assist[bot]指出工人池修剪时的 available_gpus > 0 守卫在零 GPU 时阻止修剪，导致 model_pool 尝试启动不适配的模型。建议移除守卫。已采纳。

- gemini-code-assist[bot]指出 CUDA_VISIBLE_DEVICES=-1 是 CUDA 禁用所有设备的标准方式，当前实现会将其计为 1 个 GPU。建议过滤掉 -1 值。已采纳。
- 修复 GPU 跳过逻辑中 available_gpus > 0 守卫导致的静默通过 (correctness): 已采纳建议，将条件改为 if test_gpus > available_gpus:，使其在零 GPU 时也正确跳过测试并触发 loud-fail。
- 修复工人池修剪中 available_gpus > 0 守卫导致的静默通过 (correctness): 已采纳，无条件进行修剪，确保零 GPU 运行器也正确裁剪模型需求。
- 处理 CUDA_VISIBLE_DEVICES=-1 使 GPU 计数为零 (correctness): 已采纳，在 _count_gpus_without_cuda 中添加了对 -1 的过滤，函数现在正确返回可用 GPU 数量。

风险与影响

- 风险:
  - gRPC 服务降级：当 smg-grpc-servicer 版本低于 0.5.3 时，HTTP sidecar 不启动，--enable-metrics 会导致服务器启动失败，没有 metrics 和 profiling 能力。
  - 测试时间增加：setup_backend 从 class 作用域改为 function 作用域后，每个测试都重启路由器（约 1-2 秒），可能显著延长 e2e 测试执行时间。
  - GPU 计数依赖 nvidia-smi：_count_gpus_without_cuda 回退时若 nvidia-smi 失败或超时，日志记录可能不准确，但影响限于 CI 环境。
  - IB 设备检测硬编码：优先 mlx5_ib* 模式可能不适用于非 Mellanox 硬件，但当前 CI 运行器均为 Mellanox。
  - pytest-parallel 移除：e2e 测试改为串行执行，可能增加 CI 耗时，但避免了死锁风险。
  - 影响：用户影响：无直接用户可见变更，gRPC 服务在旧版 smg-grpc-servicer 下缺失 metrics/profile 功能已有降级警告。系统影响：SMG e2e 测试恢复在 H100 集群上运行，CI 稳定性提升；模型池引用泄漏问题得到修复。团队影响：需要维护 2-gpu-h100 和 4-gpu-h100 运行器镜像，并注意 fixture 作用域变更后的测试执行时间。
  - 风险标记：gRPC sidecar 静默降级，测试执行时间增加（function scoped），GPU 计数依赖 nvidia-smi，IB 设备检测硬件假设，pytest-parallel 移除后串行执行

关联脉络

- PR #22500 [Observability] Add HTTP sidecar endpoints and FlushCache gRPC RPC for gRPC mode: 引入了 on_request_manager_ready 参数但未更新 smg-grpc-servicer 最低版本，导致旧版 servicer 启动时 TypeError。本 PR 修复了该兼容性问题。
- PR #24166 未在历史列表中提供，但根据 PR 描述，该 PR 是临时禁用 SMG e2e 的跳过的 PR。: 临时跳过了 SMG e2e 测试，本 PR 移除了该跳过。
- PR #23314 未在历史列表中提供，PR 描述提到借鉴了其并发锁模式。: 本 PR 的作业级并发锁模式参考了 #23314 的 per-hardware-group 模式。