PR #36233 完整报告

sgl-project/sglang

[NVIDIA] Add CUDA 13.4 container for initial Rubin support

合并时间：2026-08-27 06:02

原文链接：http://prhub.com.cn/sgl-project/sglang/pull/36233

执行摘要

- 一句话：新增 CUDA 13.4 容器与 nightly 流水线，初步支持 Rubin sm_107
- 推荐动作：值得容器与 CI 基建团队精读。该 PR 展示了在一项硬件工具链尚未正式发布时如何搭建「可用优先」的构建栈：预览源引导、公共 NCCL 回退、C++ 标准参数化、三大 wheel 内联重编、cutedsl 架构降级，这些 trade-off 与『同层清理预览源而非 apt pin』等细节都值得借鉴。对应用层工程师价值有限，可只关注 nightly-cu134 镜像的可用性与 flashinfer 冷启动的注意事项。

功能与动机

CUDA 13.4 Developer Preview 为 Rubin 硬件（sm_107）提供了初步支持，而 sglang 需要尽早具备在该硬件上跑通 DeepSeek-V4-Pro 等模型的能力。PR body 明确说明：'This PR adds a new docker image with CUDA 13.4 for functional Rubin support, which would be built alongside the current cu12 and cu13 containers.' 即在现有 cu12/cu13 容器之外并行提供 cu134 镜像，作为 Rubin 功能验证的入口，并在 body 中给出 DeepSeek-V4-Pro GSM8K 精度（0.978 / 0.977）作为正确性证据。

实现拆解

实现分为 5 个步骤：

1. 新增 CUDA 13.4 容器基础栈（docker/Dockerfile.cu134，+1167 行）：顶部把 CUDA 版本、预览仓库地址、torch nightly 版本全部参数化（如 CUDA_PKG_VERSION=13-4、TORCH_NIGHTLY_VERSION=2.15.0.dev20260818+cu134、MANYLINUX_IMAGE=pytorch/manylinuxaarch64-builder:cuda13.4）。cuda_base 阶段从 ubuntu:24.04 起步，通过 nvidia-packages-preview.sources 指向 prerelease/cuda/13.4.0 源安装 cuda-toolkit-13-4，再软链 /usr/local/cuda。原因：13.4 官方 nvidia/cuda base 镜像尚未发布，只能用预览源引导。随后在公共 CUDA 仓库安装 libnccl2 / libnccl-dev（预览源不发 NCCL，而 pynccl_allocator 在 symmetric memory 下 JIT 编译需要 -lnccl），并必须在同一层内删除预览源——不能用 apt pin 抑制，因为 devtools 源会发布元数据完全相同的包，pin 会误伤 nsight-systems-cli 的安装。
2. sgl-kernel 构建配置适配 C++20（python/sglang/kernels/aot/CMakeLists.txt，+5/-4）：torch 2.15 将编译标准从 c++17 升到 c++20，因此把写死的 "-std=c++17" 替换为新增的缓存变量 SGL_KERNEL_CXX_STANDARD（默认 '17'，CACHE STRING），host 端 CMAKE_CXX_STANDARD 与 CUDA 侧 SGL_KERNEL_CUDA_FLAGS、FA3、InFLLM 三

处均改为 "-std=c++${SGL_KERNEL_CXX_STANDARD}"。默认值保持 17 保证 cu12/cu13 存量构建不受影响，cu134 镜像构建时以 -DSGL_KERNEL_CXX_STANDARD=20 覆盖。

3. 升级 FlashMLA 依赖（python/sglang/kernels/aot/cmake/flashmla.cmake，+2/-2）：FetchContent 的 FlashMLA 归档从 05e2664 升级到 c1dee56（对应 PR body 中声明的 commit 15f13e5），并同步更新 SHA256。这是 Rubin 上运行 FlashMLA 的前置条件（review 中有确认）。
4. 内联源码重编三大 wheel（Dockerfile 内联构建阶段）：sgl-kernel、sgl-deep-gemm、sgl-deep-ep 都依赖 torch Python ABI，torch 2.15 升级后必须从源码重编；部分 wheel 需要 c++17→c++20 补丁。cutedsl 4.6.2 尚不支持 sm_107，全部 cutedsl kernel 通过 CUTE_DSL_ARCH / CUTE_EXPERIMENTAL_DSL_ARCH 回退到 sm_100f 目标。flashinfer-jit-cache 无 13.4 发布版本，跳过安装，flashinfer kernel 将走 JIT 编译路径，PR body 明确提示启动时间会很长。
5. 新增 nightly 发布流水线（.github/workflows/release-docker-cu134-nightly.yml，+104 行）：schedule 每天 10:00 触发（与现有 nightly docker 构建错峰，因为本构建耗时长），支持 workflow_dispatch 手动指定 image_repo / docker_target / build_only；仅跑 linux/arm64，超时 360 分钟（冷构建要重编三个 CUDA wheel），产出 nightly-cu134-{date}-{sha} 与滚动 tag nightly-cu134，并带 5 次重试的 push 逻辑。

测试与验证配套：无单元测试，但 PR body 提供了 DeepSeek-V4-Pro GSM8K 两组精度数据（TP4 MTP 场景 Accuracy 0.978、DEP4 MegaMOE 场景 Accuracy 0.977，Invalid 均为 0.000）；CI 侧 Extra 测试通过。

关键文件：

- docker/Dockerfile.cu134（模块 容器镜像；类别 infra；类型 infrastructure）：本 PR 的核心交付物：CUDA 13.4 开发者预览版容器，包含预览源引导、NCCL 回退、torch 2.15 nightly 固定、三大 wheel 源码重编与 cutedsl sm_100f 降级等全部关键逻辑。
- .github/workflows/release-docker-cu134-nightly.yml（模块 发布流水线；类别 infra；类型 infrastructure）：新增的 nightly 发布流水线，负责自动构建并推送 aarch64 的 cu134 镜像，是镜像可被用户获取的通道；含错峰调度、手动触发、build_only 验证与 360 分钟超时等设计。
- python/sglang/kernels/aot/CMakeLists.txt（模块 内核构建；类别 infra；类型 configuration；符号 SGL_KERNEL_CXX_STANDARD, CMAKE_CXX_STANDARD）：sgl-kernel 构建入口的共享改动：把写死的 c++17 编译标准参数化为 SGL_KERNEL_CXX_STANDARD，是适配 torch 2.15（c++20）的关键，同时默认值 17 保护了既有 cu12/cu13 构建。
- python/sglang/kernels/aot/cmake/flashmla.cmake（模块 依赖管理；类别 other；类型 dependency-bump；符号 repo-flashmla）：FlashMLA 源码归档升级，是 Rubin 上运行 FlashMLA 的前置依赖；虽然改动仅 4 行，但被 reviewer 专门质疑过必要性，属于需要明确归属的依赖变更。

关键符号：未识别

关键源码片段

docker/Dockerfile.cu134

本 PR 的核心交付物：CUDA 13.4 开发者预览版容器，包含预览源引导、NCCL 回退、torch 2.15 nightly 固定、三大 wheel 源码重编与 cutedsl sm_100f 降级等全部关键逻辑。

```
# ============ CUDA base 阶段 ============
# 官方 nvidia/cuda 13.4 基础镜像尚未发布，因此从 ubuntu 24.04 起步，
# 通过 NVIDIA 开发者预览仓库安装 cuda-toolkit-13-4，再软链 /usr/local/cuda
FROM ${UBUNTU_BASE_IMAGE} AS cuda_base

ARG CUDA_PKG_VERSION
ARG CUDA_PREVIEW_REPO
ARG CUDA_PREVIEW_SUITE

RUN export DEBIAN_FRONTEND=noninteractive \
    && apt-get update \
    && apt-get install -y --no-install-recommends ca-certificates wget gnupg \
    # 安装预览仓库 keyring，并把 prerelease/cuda/13.4.0 写入 apt 源
    && wget -q -O /tmp/nvidia-preview-keyring.deb \
        "${CUDA_PREVIEW_REPO}/nvidia-preview-keyring.deb" \
    && dpkg -i /tmp/nvidia-preview-keyring.deb \
    && rm -f /tmp/nvidia-preview-keyring.deb \
    && printf '%s\n' \
        'X-Repolib-Name: NVIDIA Packages (frozen)' \
        'Types: deb' \
        "URIs: ${CUDA_PREVIEW_REPO}" \
        "Suites: ${CUDA_PREVIEW_SUITE}" \
        'Components: main' \
        'Signed-By: /usr/share/keyrings/nvidia-packages-preview.gpg' \
        'Enabled: yes' \
        > /etc/apt/sources.list.d/nvidia-packages-preview.sources \
    && apt-get update \
    # 预览仓库只发布 cuda-toolkit 13.4 包，安装后软链并验证 nvcc
    && apt-get install -y --no-install-recommends "cuda-toolkit-${CUDA_PKG_VERSION}" \
    && cuda_dir="$(ls -d /usr/local/cuda-1* 2>/dev/null | head -1)" \
    && test -n "${cuda_dir}" \
    && ln -sfn "${cuda_dir}" /usr/local/cuda \
    && /usr/local/cuda/bin/nvcc --version \
    && rm -rf /var/lib/apt/lists/*

# 预览源不发 NCCL 包，而后续 pynccl_allocator 在 symmetric memory 场景下
# JIT 编译需要 -lnccl，因此要回退公共 CUDA 仓库安装 libnccl2 / libnccl-dev。
# 注意：必须在同一层内删除预览源，不能用 apt pin——devtools 源会发布
# 元数据完全相同的包，任何 pin 都会误伤 nsight-systems-cli 的安装。
```

python/sglang/kernels/aot/CMakeLists.txt

sgl-kernel 构建入口的共享改动：把写死的 c++17 编译标准参数化为 SGL_KERNEL_CXX_STANDARD，是适配 torch 2.15（c++20）的关键，同时默认值 17 保护了既有 cu12/cu13 构建。

```
# torch 2.15 起编译标准从 c++17 升到 c++20，而既有 cu12/cu13 容器
# 仍按 c++17 构建。把标准做成 CMake 缓存变量，构建时由 Dockerfile 传入，
# 默认值保持 17，避免影响存量构建链。
set(SGL_KERNEL_CXX_STANDARD "17" CACHE STRING "C++ standard used for host and CUDA compilation")
set(CMAKE_CXX_STANDARD ${SGL_KERNEL_CXX_STANDARD})
set(CMAKE_CXX_FLAGS "${CMAKE_CXX_FLAGS} -O3")

# CUDA flags 中写死的 -std=c++17 统一改为引用该变量，保证 host 与 device
# 两端编译标准一致（FlashInfer、FA3、InFLLM 三处均按此处理）。
set(SGL_KERNEL_CUDA_FLAGS
    "-Xcompiler" "-fPIC"
    "-gencode=arch=compute_90,code=sm_90"
    "-std=c++${SGL_KERNEL_CXX_STANDARD}"
    "-DFLASHINFER_ENABLE_F16"
    "-DCUTE_USE_PACKED_TUPLE=1"
    "-DCUTLASS_ENABLE_TENSOR_CORE_MMA=1"
)
```

评论区精华

PR 的 review 评论不多（3 条），核心交锋集中在两点：

1. FlashMLA 升级必要性：Fridge003 在 flashmla.cmake 第 5 行质疑 'Is this hash change necessary for rubin image?'，trevor-m 回复确认必需，指出需要 commit https://github.com/sgl-project/FlashMLA/commit/15f13e5030374295491c5ce31b02d7e63a7772c6 才能在 Rubin 上运行 FlashMLA。这是「依赖升级是否有明确归属」的一次标准审问，结论清晰。
2. workflow 实测责任：nvpohanh 在新增的 release-docker-cu134-nightly.yml 上直接指派 '@Fridge003 to test if this workflow works'，要求合入前实测流水线。该评论没有书面回复，但 PR 最终合并且 CI Extra 通过，说明验证未阻塞合入。

最终 Fridge003 以 APPROVED 状态合入本 PR。

- FlashMLA 归档升级是否为 Rubin 镜像必需 (question): trevor-m 回复确认必需：需要 FlashMLA commit 15f13e5030374295491c5ce31b02d7e63a7772c6 才能在 Rubin 上运行 FlashMLA，与 PR body 声明的 bump 目标一致。
- 新增 nightly workflow 需要实测验证 (testing): 该评论没有书面回复；从 PR 的 CI Extra 状态通过及最终合入来看，验证未阻塞合入，但该 workflow 首次真实 nightly 调度的结果仍是后续观察项。

风险与影响

- 风险：
  1. 预览版工具链依赖（docker/Dockerfile.cu134）：cuda-toolkit-13-4 来自 NVIDIA prerelease 源，包内容可能随预览迭代变化甚至撤回，镜像可复现性依赖对归档 / 源的固定引用，存在隐性漂移风险。

2. NCCL 版本不匹配：13.4 无专属 NCCL，回退公共版 libnccl2/libnccl-dev。Rubin 的集合通信特性是否被该版 NCCL 完整支持未知；且『同一层内删除预览源』的实现非常脆弱，任何后续层尝试从预览源解析包都会失败。
3. 构建链脆弱：sgl-kernel / sgl-deep-gemm / sgl-deep-ep 全部从源码编译，workflow 超时 360 分钟，冷构建失败概率高；torch 2.15 nightly 版本硬编码（dev20260818），nightly 轮换后可能失效。
4. 运行时冷启动：无 flashinfer-jit-cache，flashinfer kernel 首次运行 JIT 编译，启动与首批请求延迟显著变长。
5. 性能留白：cutedsl 4.6.2 不支持 sm_107，全部 cutedsl kernel 回退到 sm_100f，Rubin 上的理论性能未充分释放。
6. 平台覆盖：镜像与流水线仅 aarch64，x86 用户无法使用 cu134 容器。
7. 存量构建回归：CMakeLists.txt 默认保持 SGL_KERNEL_CXX_STANDARD=17，对 cu12/cu13 构建风险较低，但仍属于共享构建入口改动，需要重点关注。 - 影响：对用户：Rubin（sm_107）早期用户可拉取 lmsysorg/sglang:nightly-cu134 镜像做功能验证与精度测试；镜像仅 aarch64，x86 用户暂不可用。对系统：新增一条与 cu12/cu13 并行的夜间构建流水线，sgl-kernel 构建入口增加 C++ 标准参数（SGL_KERNEL_CXX_STANDARD），为后续其他 torch 版本适配提供可复用机制。对团队：需要持续跟踪 CUDA 13.4 正式发布节奏，待官方 base 镜像、NCCL、flashinfer-jit-cache 就绪后收敛 Dockerfile 中的各类 workaround；nightly 构建运行时间长（约 6 小时超时预算），占用 arm 构建节点资源。总体影响范围集中在基础设施层，不触及运行时核心调度 / 缓存逻辑。 - 风险标记：依赖预览版 CUDA 工具链，无 13.4 专属 NCCL，torch 2.15 nightly 硬编码，flashinfer 冷启动时间长，cutedsl 回退 sm_100f，镜像仅 aarch64，长构建时间（360 分钟超时）

关联脉络

- PR #36397 [NVIDIA] Tune custom all reduce v2 for sm_107: 同属 Rubin sm_107 硬件支持线：本 PR 铺好 CUDA 13.4 容器后，社区紧接着对 sm_107 做 custom allreduce 性能调优，显示 Rubin 支持从『能跑』走向『调优』。
- PR #36465 [Kernel] Declare the PyTorch ABI dependency in sgl-kernel wheels: 与 sgl-kernel 的 torch ABI 兼容直接相关：本 PR 因 torch 2.15 升级不得不从源码重编 sgl-kernel，印证了在 wheel 中声明 PyTorch ABI 依赖的必要性。