

PR #2700 完整报告

radixark/miles

feat(ci): add stage-c-8-gpu-b200 for the Blackwell runner

合并时间: 2026-08-30 14:37

原文链接: <http://prhub.com.cn/radixark/miles/pull/2700>

执行摘要

- 一句话: 新增 stage-c-8-gpu-b200, 接入 8xB200 CI runner
- 推荐动作: 值得快速精读。虽然单看只是注册表加一行 + 一个 workflow job, 但其中蕴含两个可复用的 CI 设计原则: 一是「测试自声明 GPU 预算 (ray start --num-gpus / torchrun --nproc-per-node) 而不是读取可见设备」, 这在不分区主机的资源分配语义清晰; 二是「合入时零测试注册是安全的, 因为空 stage 收集零测试并退出 0」, 支撑了增量迁移的分步落地。对参与 CI 基础设施维护的工程师, 建议重点关注 pr-test.yml 中 job 注释与 tests/ci/run_suite.py 的 suite 注释, 这两处把「为什么这么设计」讲得最清楚。

功能与动机

PR body 明确这是 Blackwell CI 的 Chain 1/5, 目标是把 8xB200 主机纳入现有 CI stage 体系, 为后续 b200-ci-hardware-axis、b200-ci-baseline-identity、b200-ci-dispatch、b200-ci-enable-tests 逐步在 Blackwell 硬件上跑测试铺路。作者强调两个约束: 一是主机刻意不分区——「an 8-GPU runner cannot share a node with 2/4-GPU runners without both claiming the same devices, and GitHub has no cross-runner resource lock」; 二是合入前置条件——runner 必须已在线, 否则「the nightly cron would queue a ["b200", "8gpu"] job that no runner can claim」。prerequisite 已确认完成: b200-oma-8gpu-0 online, radixark/miles:dev 在 sm100 上验证通过 (torch 2.11.0+cu130, capability (10,0), 8 devices, bf16 matmul)。

实现拆解

1. Suite 注册: 在 tests/ci/run_suite.py 的 CI_SUITES[HWBackend.CUDA] 列表末尾追加 stage-c-8-gpu-b200, 并重写模块顶部注释, 说明三个 8-GPU 整节点 suite (H100/H200/B200) 与 H200 拆分车队 (2+2+4, 按 CUDA_VISIBLE_DEVICES 分区) 的差异, 以及 Blackwell 主机不分区、子 8-GPU 测试也注册到本 suite 的原因。
2. Stage 选择与 runner 映射: 在 tests/ci/stage_selection.py 的 PR_GPU_STAGES 中加入 stage-c-8-gpu-b200, 使 PR 变更路径扇出时能选中该 stage; 在 tests/ci/file_run.py 的 CUDA_SUITE_RUNS_ON 中加入 "stage-c-8-gpu-b200": ["b200", "8gpu"] 映射, 保证 /rerun-test 文件运行链路 (run-ci-file.yml 的 resolve-file-run job) 能解析到新 runner。
3. Job 接线: 在 .github/workflows/pr-test.yml 末尾追加 stage-c-8-gpu-b200 job, 与其它 GPU stage 保持同一 gate 模式 (needs: [resolve-ci-policy, resolve-ci-image, stage-a-cpu], 包含 skipped_stages fail-open 判断和 bypass_fastfail 门), 复用

_run-ci.yml 可复用 workflow，以 --suite stage-c-8-gpu-b200 调用 run_suite.py。

4. 锁测试同步：更新 tests/ci/test/test_run_suite.py 中所有 stage 数量断言——test_cuda_suites_exact 追加新 suite 名、command 块计数 7→8、_run-ci.yml 引用 5→6、manual_scope 计数 5→6、bypass/fail-open gate 计数 5→6。这些锁测试确保新 stage 必须遵守现有策略 seam 和 gate 约定，防止配置漂移。
5. 文档配套：tests/ci/README.md 记录 b200-oma-8gpu-0 主机约定（整节点、CVD 未设置、刻意不分区）；docs/ci/00-stage.md 将 stage 命名规则的 hw 集合扩展为 {cpu, h100, h200, b200, mi350}，并在 roster 表新增一行；docs/ci/contributor-guide.md 补充 Blackwell suite 选择指引（nvfp4、mxfp8 等需要 Blackwell 硬件的测试）。

关键文件：

- .github/workflows/pr-test.yml（模块 CI workflow；类别 infra；类型 infrastructure）：新增 stage-c-8-gpu-b200 job 是本次核心接线：runs_on ["b200", "8gpu"]、复用 _run-ci.yml、与其它 GPU stage 完全一致的四重 gate（policy/image/cpu-gate/bypass），并附带整段设计注释说明为什么不分区。
- tests/ci/run_suite.py（模块 套件注册；类别 test；类型 configuration；符号 CI_SUITES）：CI_SUITES 是 suite 注册的事实来源，新增 stage-c-8-gpu-b200 后，所有 policy 解析、测试筛选、skip 报告都会自动感知新 suite；同步改写的注释阐明了不分区设计。
- tests/ci/test/test_run_suite.py（模块 套件测试；类别 test；类型 test-coverage；符号 test_cuda_suites_exact, test_every_stage_consumes_resolved_policy, test_cpu_and_gpu_stages_use_dedicated_reusable_workflows, test_dispatch_has_no_scope_input_but_runs_all_cuda_domains）：锁测试是本 PR 的验收网：所有 stage 数量断言（7→8、5→6 等）确保新 stage 必须遵守既有策略 seam、gate 约定和 fail-open skip 列表，防止配置漂移静默通过。
- tests/ci/file_run.py（模块 文件运行；类别 test；类型 configuration；符号 CUDA_SUITE_RUNS_ON）：CUDA_SUITE_RUNS_ON 是 /rerun-test 文件运行链路解析 runner 标签的映射表，缺了它，显式请求单个 Blackwell 测试文件时就无法路由到 b200 runner。
- tests/ci/stage_selection.py（模块 阶段选择；类别 test；类型 configuration；符号 PR_GPU_STAGES）：PR_GPU_STAGES 决定 PR 变更路径扇出到哪些 GPU stage，新增条目后，任何改动 Blackwell 测试或相关源码的 PR 都会自动把 stage-c-8-gpu-b200 纳入可选中集合。
- tests/ci/README.md（模块 测试文档；类别 docs；类型 documentation）：记录 b200-oma-8gpu-0 主机约定（整节点、CVD 未设置、刻意不分区），是后续维护者理解 Blackwell fleet 资源语义的第一手文档。
- docs/ci/00-stage.md（模块 CI 文档；类别 docs；类型 documentation）：stage 命名规则与 roster 表是 CI stage 体系的规范文档，hw 集合加入 b200 并新增一行，保证新硬件类有据可查。
- docs/ci/contributor-guide.md（模块 CI 文档；类别 docs；类型 documentation）：贡献者选择 suite 的对照表新增 Blackwell 指引，帮助测试作者为 nvfp4、mxfp8 等硬件相关测试选对归宿。

关键符号: CI_SUITES, PR_GPU_STAGES, CUDA_SUITE_RUNS_ON,
test_cuda_suites_exact, test_every_stage_consumes_resolved_policy,
test_cpu_and_gpu_stages_use_dedicated_reusable_workflows,
test_dispatch_has_no_scope_input_but_runs_all_cuda_domains,
test_gpu_gates_consume_shared_bypass_output

关键源码片段

[.github/workflows/pr-test.yml](#)

新增 stage-c-8-gpu-b200 job 是本次核心接线: runs_on ["b200", "8gpu"]、复用 _run-ci.yml、与其它 GPU stage 完全一致的四重 gate (policy/image/cpu-gate/bypass), 并附带整段设计注释说明为什么不分区。

```
# Blackwell fleet 是单台未分区的 8-GPU 主机: 8-GPU runner 无法与
# 2/4-GPU runner 共存于同一节点, 否则 GitHub 会把 job 派发到
# 重叠的设备上 (GitHub 没有跨 runner 的资源锁)。
# 少于 8 卡的 Blackwell 测试也注册到本 suite; 不设分区矩阵,
# 单个 runner 串行处理 shard, 当前预算远在 job 超时之内。
stage-c-8-gpu-b200:
  needs: [resolve-ci-policy, resolve-ci-image, stage-a-cpu]
  if: |
    always() && !cancelled() &&
    needs.resolve-ci-policy.result == 'success' &&
    needs.resolve-ci-image.result == 'success' &&
    !contains(fromJSON(needs.resolve-ci-policy.outputs.skipped_stages || '[]'), 'stage-c-8-gpu-b200') &&
    (needs.stage-a-cpu.result == 'success' ||
     needs.resolve-ci-policy.outputs.bypass_fastfail == 'true')
  uses: ../github/workflows/_run-ci.yml
  with:
    runs_on: ["b200", "8gpu"]
    container_image: ${needs.resolve-ci-image.outputs.container_image}
    ref: ${inputs.ref || ''}
    execute_command: >-
      python tests/ci/run_suite.py --hw cuda --suite stage-c-8-gpu-b200
      --cadence ${needs.resolve-ci-policy.outputs.cadence}
      --labels ${needs.resolve-ci-policy.outputs.raw_labels}
      ${github.event_name == 'workflow_dispatch' && '--match-all-labels' || ''}
  secrets: inherit
```

[tests/ci/run_suite.py](#)

CI_SUITES 是 suite 注册的事实来源, 新增 stage-c-8-gpu-b200 后, 所有 policy 解析、测试筛选、skip 报告都会自动感知新 suite; 同步改写的注释阐明了不分区设计。

```
# CUDA suites: 每个 suite 由 .github/workflows/pr-test.yml 中对应的
# workflow job 提供。stage-c-8-gpu-h100 / stage-c-8-gpu-h200 /
# stage-c-8-gpu-b200 运行在整节点 8-GPU 主机上; H200 拆分车队是一台
# 8-GPU 节点通过每个 runner 的 CUDA_VISIBLE_DEVICES 分为 2+2+4 个
```

```

# worker (参见 pr-test.yml 中 stage-c-4-gpu-h200 / stage-b-2-gpu-h200 /
# stage-c-2-gpu-h200 job 注释)。Blackwell 单台主机不分区，因此
# 2/4-GPU 的 Blackwell 测试也注册到 8-GPU suite —— 测试通过
# ray start --num-gpus 或 torchrun --nproc-per-node 声明自己的预算，
# 而不是读取可见设备。
CI_SUITES = {
    HWBackend.CPU: [
        "stage-a-cpu",
        "stage-b-cpu",
    ],
    HWBackend.CUDA: [
        "stage-b-2-gpu-h200",
        "stage-c-8-gpu-h100",
        "stage-c-8-gpu-h200",
        "stage-c-4-gpu-h200",
        "stage-c-2-gpu-h200",
        "stage-c-8-gpu-b200", # 新增：Blackwell 8-GPU 整节点 suite
    ],
    HWBackend.ROCM: [
        # 由 pr-test-rocm.yml 消费。
        "stage-c-4-gpu-mi350",
        # 由外部 sgl-project/sglang MI350 nightly 消费。
        "nightly-stage-c-8-gpu-mi350",
        "nightly-stage-c-4-gpu-mi350",
        "nightly-stage-c-2-gpu-mi350",
    ],
}

```

评论区精华

本 PR 没有产生实质 review 讨论：review_comments_count 为 0，唯一的 review 交互是 claude[bot] 的自动提醒（提示仓库配置了手动 review，可 @claude review）和 guapisolo 的空 body APPROVED。设计上的两个关键权衡——「不分区」和「零测试注册也合入」——都在 PR body 与代码注释中预先论证，未经过 review 交锋：不分区是因为 GitHub 没有跨 runner 资源锁，任何 2/4-GPU runner 与 8-GPU runner 共存都会争抢同一批物理设备；零测试注册是因为 stage 在没有启用注册时收集零测试并退出 0，属于增量迁移期的预期状态，验证责任在链路第 5 个 PR (b200-ci-enable-tests)。

- 暂无高价值评论线程

风险与影响

- 风险：

1. 外部 runner 依赖：PR body 明确指出合入前提是 b200-oma-8gpu-0 已在线（已确认）。但若该 runner 后续下线或标签缺失，schedule 触发的 nightly cron 对 skipped_stages 为空，会持续排队一个无人认领的 ["b200", "8gpu"] job，造成 CI 挂起。

2. 零测试覆盖空洞：当前没有任何启用的测试注册到 stage-c-8-gpu-b200，stage 收集零测试并以 0 退出，属于有意设计。真正的执行验证（sm100 上的 torch/cuda 兼容、BF16 matmul、容器内 ray/torchrun 启动）要等链路第 5 个 PR（b200-ci-enable-tests）才发生，在此之前 b200 路径从未被真实 job 跑过。
3. 资源利用率：不分区意味着 2/4-GPU 的 Blackwell 测试会空闲 4-6 张 GPU；单分片 job 串行化 shard，若后续 Blackwell 测试量增长超出 job 超时预算，会需要重新引入分区矩阵或并发 runner。
4. 镜像与硬件兼容性：验证了 radixark/miles:dev 在 sm100 上可运行（capability (10,0)），但后续若镜像构建目标不含 Blackwell 相关的 CUDA 能力（nvfp4/mx8 kernel），启用测试后会暴露新的失败模式。
 - 影响：对系统：CI stage 分类体系新增第四种 NVIDIA 硬件类 b200，PR 与 schedule 两条触发路径都会把新 stage 纳入策略解析和 gate 计算；/rerun-test 文件运行链路同步具备解析能力。对团队：Blackwell 相关测试（nvfp4、mx8 等）从此有了明确的归属 suite，贡献者指南给出选择指引，降低后续接入成本。对用户：无任何面向终端用户的行为变化。影响程度为低到中——纯 CI 基础设施配置变更，但它是 5-PR 链的入口，后续 4 个 PR 将逐步把 Blackwell 测试纳入常规 CI，最终影响面会扩大。
 - 风险标记：依赖外部 runner 在线，新 stage 零测试覆盖，GitHub 无跨 runner 资源锁，Blackwell 新硬件平台

关联脉络

- PR #2768 [AMD] fix(docker): drop the obsolete ROCm Megatron fused-kernels patch: 两个 PR 共享 tests/ci/test/test_run_suite.py，属于同一套 CI stage 治理 / 锁测试机制的连续演进；2768 从 AMD 侧调整 stage 相关验证，2700 从 Blackwell 侧扩展 stage 体系。
- PR #2813 feat(ci): add user allowlists for command tiers: 同为 CI 基础设施治理方向，扩展 .github/workflows 与 tests/ci 工具链，反映仓库对 CI 权限与 runner 编排的持续投入。
- PR #2773 ci: default a file run to the PR's own image: 与 file_run.py 的 suite→runner 映射同属 CI 文件运行链路（run-ci-file.yml / resolve-file-run），本 PR 让该链路对 b200 suite 可见。