PR #36768 完整报告

sgl-project/sglang

:memo: [NPU] Use vendor-neutral wording in quantization comments

合并时间：2026-08-30 04:05

原文链接：http://prhub.com.cn/sgl-project/sglang/pull/36768

执行摘要

- 一句话：NPU 量化注释去除厂商措辞，统一为 NPU
- 推荐动作：无需精读技术细节，但值得快速过一眼关键文件的注释，尤其是 linear_method_npu.py 和 mxfp4_npu.py 中保留的量化布局规范（FRACTAL_NZ、双级 scale 形状、DualLevelQuantBatchMatMul 的 A5-only 限制），这些是 NPU 量化调试的重要线索。该 PR 适合作为注释术语规范的参考模板。

功能与动机

PR body 指出：#34829 清理了 MXFP8/MXFP4 线性方法的厂商措辞，但另外五个文件中仍残留两类文案——一是引用外部项目的注释，二是将 Ascend 作为 docstring 中的 prose 品牌词用于命名空间已是 npu 的模块。作者认为应移除第三方归属、保留真实信息（如在线 MXFP4 路径的实验性警告），并将硬件表述统一为 NPU，避免品牌绑定。

实现拆解

1. 删除外部项目引用：在 init_routing.py 的 scale 归一化 docstring、moe_methods.py 的 fused-MoE 权重布局注释、qwen3_moe.py 的 router-gate 注释中移除 vllm-ascend 引用；在 linear_method_npu.py 的 NPUDualLevelMXFP4LinearMethod docstring 中移除 MindIE-SD 引用；在 modelslim_mxfp4_scheme.py 的 scheme 头部 docstring、mul_scale 关键注释及 mxfp4_npu.py 中移除 MindIE-SD 引用。这些被删行均为纯归属说明，周边注释已包含 FRACTAL_NZ 格式、L0 scale 转置、smooth-quant 等真实约束。
2. 保留并改写关键事实：在线 MXFP4 diffusion 路径的“实验性”说明保留，但将“仅 MindIE-SD 有离线路径”改为中性的“这些模型已有的 MXFP4 路径仅支持离线”；modelslim_mxfp4_scheme.py 中 mul_scale 缺失会导致 mosaic 输出的 CRITICAL 警告保留其技术内容。
3. 将 Ascend 改写为 NPU：linear_method_npu.py 内六个类 docstring 与三处硬件要求行（如 Ascend 950 (A5) → A5 NPU）、moe_methods.py 中两个 dispatcher dtype 注释与 NPUMXFP8MoEMethod 头、online_moe_methods.py、modelslim_mxfp4_scheme.py、modelslim_mxfp8_scheme.py、mxfp4_npu.py 的模块与类 docstring 均统一改写。
4. 刻意保持标识符与用户可见字符串不变：AscendQuantInfo、AscendRunnerCore、MoeRunnerBackend.ASCEND、is_ascend_fuseep()、ascend_tp / ascend_fuseep、ascend attention backend 以及向用户显示的错误消息均不动，避免破坏 API 与运行时行为。

5. 质量验证：未加测试，仅通过 python -m black --check 与 AST 解析验证所有文件语法；无行为变更。

关键文件：

- python/sglang/multimodal_gen/runtime/layers/quantization/mxfp4_npu.py（模块 量化层；类别 source；类型 documentation）：同时体现了删除 MindIE-SD 引用和 Ascend -> NPU 措辞改写两种清理，且保留在线 MXFP4 路径的实验性警告，是本次清理的核心样例文件。
- python/sglang/srt/hardware_backend/npu/quantization/linear_method_npu.py（模块 量化层；类别 source；类型 documentation）：SRT 侧 NPU 密集量化方法的集中文件，6 个类 docstring 与硬件要求行统一从 Ascend 改写为 NPU，并移除 MindIE-SD 引用。
- python/sglang/multimodal_gen/runtime/layers/quantization/modelslim_mxfp4_scheme.py（模块 量化层；类别 source；类型 documentation）：offline ModelSlim MXFP4 scheme 的头部 docstring 与 mul_scale 关键注释中移除 MindIE-SD 引用，但保留 smooth-quant 校准对正确性的影响说明。
- python/sglang/srt/models/qwen3_moe.py（模块 MoE；类别 source；类型 documentation）：Qwen3 MoE router-gate 注释中删除 vllm-ascend 引用，说明 ModelSlim 离线路径才可量化 gate 的逻辑保持原样。
- python/sglang/srt/hardware_backend/npu/quantization/moe_methods.py（模块 MoE 量化；类别 source；类型 documentation）：MoE 量化方法文件，dispatcher dtype 注释与 NPUMXFP8MoEMethod docstring 中的 Ascend 措辞改写为 NPU，并删除 vllm-ascend 对照引用。
- python/sglang/multimodal_gen/runtime/layers/quantization/modelslim_mxfp8_scheme.py（模块 量化层；类别 source；类型 documentation）：模块 docstring 简单描述从 Ascend NPU 改为 NPU，属于全文统一的最小改动。
- python/sglang/srt/hardware_backend/npu/quantization/online_moe_methods.py（模块 MoE 量化；类别 source；类型 documentation）：online MoE 方法模块头与类 docstring 中的 Ascend 指称改为 NPU，保持与其他文件的一致性。
- python/sglang/srt/hardware_backend/npu/moe/init_routing.py（模块 MoE 路由；类别 source；类型 documentation）：init-routing scale 归一化 docstring 删除 vllm-ascend 对照说明，纯归属清理。

关键符号：未识别

关键源码片段

python/sglang/multimodal_gen/runtime/layers/quantization/mxfp4_npu.py

同时体现了删除 MindIE-SD 引用和 Ascend -> NPU 措辞改写两种清理，且保留在线 MXFP4 路径的实验性警告，是本次清理的核心样例文件。

```
"""Online MXFP4 quantization for Diffusion models on NPU.

Provides ``NPUMXFP4Config`` (registered as ``"mxfp4_npu"``) and
``NPUMXFP4DiffusionLinearMethod`` which quantises FP16/BF16 weights to MXFP4
```

```
    at load time using dual-level MX quantization, and uses
    ``npu_dynamic_dual_level_mx_quant`` + ``npu_dual_level_quant_matmul`` for
    inference.

    The ``"mxfp4_npu"`` key is distinct from upstream's ROCm ``"mxfp4"``
    (``Mxfp4Config`` in ``mxfp4.py``) which targets AMD MI350+ via aiter kernels.

    NOTE: Online weight quantization via ``npu_dynamic_dual_level_mx_quant`` is
    experimental; the established MXFP4 path for these models is offline
    (pre-quantized) only. The online path quantizes FP16/BF16 weights at load
    time, which may produce different numerical results than the offline
    calibrated path.
    """

    # 本次 PR 的注释清理要点：
    # 1. docstring 中 “Ascend NPU” 改为 “NPU”，去除厂商品牌词；
    # 2. “MindIE-SD only uses an offline path” 改写为通用的
    # “the established MXFP4 path for these models is offline only”；
    # 3. 实验性警告与数值差异说明保留，因为它们承载了真实技术约束。
```

python/sglang/srt/hardware_backend/npu/quantization/linear_method_npu.py

SRT 侧 NPU 密集量化方法的集中文件，6 个类 docstring 与硬件要求行统一从 Ascend 改写为 NPU，并移除 MindIE-SD 引用。

```
class NPUDualLevelMXFP4LinearMethod(NPUSingleLevelMXFP4LinearMethod):
    """NPU W4A4 online quantization: dual-level MXFP4 (higher accuracy).

    This is the sole online ``--quantization mxfp4`` linear path. Instead of
    a single UE8M0 (power-of-2) block scale, dual-level MX quant produces a
    finer L0 (FP8 E4M3) scale immediately followed by the L1 scale.

    # 本次 PR 的清理：
    # 原 docstring 中 “Reference: MindIE-SD ... ” 一行已删除，
    # 对应的 “Hardware: Ascend 950 (A5) only” 改写为
    # “Hardware: A5 NPU only”，并且保留“A2/A3 上
    # DualLevelQuantBatchMatMul 不可用”这一关键硬件限制。
    """
```

评论区精华

PR 合并时由维护者 ping1jing2 操作，其评论：> "i merged it without running CI as there is no code change in this PR" 表明虽然 PR 状态显示多个 CI run 为失败 / 未完成，但因纯注释变更，维护者选择跳过 CI 直接合并。除此之外没有代码 review 评论。

- 跳过 CI 直接合并 (other): 维护者确认无行为变更，直接合并，未运行 CI。

风险与影响

- 风险：无行为变化，因此不存在运行时回归、性能或安全风险。主要风险是维护性：移除 MindIE-SD、vllm-ascend 等外部引用后，后续维护者定位 NPU 量化实现的上游参考来源会变困难；但 PR 保留了技术约束描述（如 FRACTAL_NZ format 29、L0 scale 转置、mul_scale 必须应用等），风险有限。另外 CI 未运行，虽然影响很小，但严格来说合并流程未覆盖。
- 影响：对用户无影响，所有用户可见错误信息和量化配置键均未改变。对开发者影响主要是注释可读性：去品牌化后 NPU 量化模块的措辞更中性，不再暗示与华为 Ascend 或 MindIE-SD 绑定，便于多硬件后端语境下阅读。对团队而言，这是继 #34829 之后统一注释风格的收尾，可减少后续 PR 中混用 Ascend / NPU 的情况。
- 风险标记：CI 未运行，上游参考链接移除

关联脉络

- 暂无明显关联 PR