PR #33945 完整报告

sgl-project/sglang

feat: support deterministic FA4 for GLM-4.7-Flash

合并时间：2026-08-12 16:57

原文链接：http://prhub.com.cn/sgl-project/sglang/pull/33945

执行摘要

- 一句话：支持 GLM-4.7-Flash 在 Blackwell 上确定性 FA4 推理
- 推荐动作：值得精读。核心看点：确定性推理与 batch-size 相关 kernel（cutedsl）冲突时的处理范式——既在 auto 路径自动规避，又对显式选择直接报错；测试基类用类身份取代参数探测来判定抽象基类，是易被忽略的整洁改动。

功能与动机

PR body 明确目标是 "Enable deterministic GLM-4.7-Flash inference with the FA4 attention backend on Blackwell"。实现上存在两个阻塞点：其一，确定性推理的架构校验名单未包含 Glm4MoeLiteForCausalLM，该模型无法通过 absorbed-MLA 确定性校验；其二，SM10x 上 BF16 GEMM 的 auto 后端默认选择 cutedsl，其 kernel 选择依赖 batch-size，会破坏推理确定性，需要将确定性模式下的默认改为 torch 并禁止 cutedsl 组合。

实现拆解

1. 扩展 absorbed-MLA 模型名单：python/sglang/srt/server_args.py 的 _handle_deterministic_inference 与 python/sglang/srt/arg_groups/overrides.py 的 _deterministic_is_deepseek_model 同步加入 Glm4MoeLiteForCausalLM。前者用于确定性推理的注意力后端校验，后者驱动未显式指定 backend 时自动解析为 triton；同时把错误信息从 "DeepSeek models" 泛化为 "absorbed-MLA models"，覆盖 GLM-4.7-Flash。
2. 约束 BF16 GEMM 后端选择：python/sglang/srt/layers/quantization/unquant.py 的 initialize_bf16_gemm_config 中，SM10x 上 auto 后端在 enable_deterministic_inference 为真时选择 torch（cuBLAS）而非 cutedsl；并在 cutedsl 分支显式抛出 ValueError，拒绝与确定性推理组合，避免静默产生 batch-size 相关的不确定结果。
3. 修正确定性测试基类判定：python/sglang/test/test_deterministic_utils.py 的 TestDeterministicBase.setUpClass 原通过探测 server args 是否含 --attention-backend 判断是否跳过基类，改为按类身份（cls is TestDeterministicBase）判断，使故意不指定 backend 的子类成为有效测试用例。
4. 新增回归测试：新增 test/registered/attention/test_glm4_moe_lite_deterministic.py，注册为 4-gpu-gb300 的 nightly 测试。TestGlm4MoeLiteFa4Deterministic 显式指定 --attention-backend fa4 跑确定性套件；TestGlm4MoeLiteAutoBackendDeterministic 不指定 backend，并通过 /server_info 断言自动解析结果为 triton。

5. 文档配套：docs/docs/advanced_features/server_arguments.mdx 中 --bf16-gemm-backend 说明由 "SM100/SM103" 统一为 "SM10x"，补充确定性模式下选 torch 的行为及 torch 选项的完整描述。

关键文件：

- python/sglang/srt/layers/quantization/unquant.py（模块 量化层；类别 source；类型 core-logic；符号 initialize_bf16_gemm_config, Bf16GemmBackend）：BF16 GEMM 后端选择的实际控制点：确定性模式下 auto 回落为 torch，并拒绝 cutedsl 组合，是本 PR 保证确定性的关键逻辑。
- python/sglang/srt/server_args.py（模块 启动参数；类别 source；类型 core-logic；符号 _handle_deterministic_inference）：确定性推理校验入口：把 Glm4MoeLiteForCausalLM 纳入 absorbed-MLA 名单，并泛化校验报错文本，决定 GLM-4.7-Flash 能否通过确定性启动校验。
- python/sglang/srt/arg_groups/overrides.py（模块 配置覆盖；类别 source；类型 core-logic；符号 _deterministic_is_deepseek_model）：自动注意力后端解析的镜像名单，与 server_args.py 同步加 Glm4MoeLiteForCausalLM，决定未显式指定 backend 时解析为 triton 而非 flashinfer。
- test/registered/attention/test_glm4_moe_lite_deterministic.py（模块 回归测试；类别 test；类型 test-coverage；符号 TestGlm4MoeLiteFa4Deterministic, TestGlm4MoeLiteAutoBackendDeterministic, test_auto_backend_resolves_to_triton）：新增回归测试，同时覆盖显式 fa4 与不指定 backend 的 auto 解析路径，是 BBuf review 指出的测试缺口的直接回应。
- python/sglang/test/test_deterministic_utils.py（模块 测试基类；类别 test；类型 test-coverage；符号 TestDeterministicBase.setUpClass）：测试基类判定逻辑修正，从探测 --attention-backend 参数改为类身份判断，是 auto 后端测试类能真正运行的先决条件。
- docs/docs/advanced_features/server_arguments.mdx（模块 文档；类别 other；类型 documentation）：参数文档同步：将 SM100/SM103 统一为 SM10x，并补充确定性模式下 auto 选择 torch 的行为，回应 BBuf 的文档不一致评论。

关键符号：initialize_bf16_gemm_config, _handle_deterministic_inference, _deterministic_is_deepseek_model, TestDeterministicBase.setUpClass, test_auto_backend_resolves_to_triton

关键源码片段

python/sglang/srt/layers/quantization/unquant.py

BF16 GEMM 后端选择的实际控制点：确定性模式下 auto 回落为 torch，并拒绝 cutedsl 组合，是本 PR 保证确定性的关键逻辑。

```
# python/sglang/srt/layers/quantization/unquant.py
def initialize_bf16_gemm_config(server_args: ServerArgs) -> None:
    global _BF16_GEMM_BACKEND, _cutedsl_bf16_gemm, _use_cutedsl_bf16_gemm

    from sglang.srt.utils import is_sm100_supported
```

```
    backend_str = server_args.bf16_gemm_backend
    # SM10x 上 auto 默认走 cutedsl（CuTe DSL），但确定性推理要求
    # kernel 选择与 batch-size 无关，因此自动回落为 torch 的 cuBLAS
    # 路径（F.linear），保证 run-to-run 与 batch 组合不变性。
    if backend_str == "auto" and is_sm100_supported():
        backend_str = (
            "torch" if server_args.enable_deterministic_inference else "cutedsl"
        )

    backend = Bf16GemmBackend(backend_str)

    if backend.is_cutedsl():
        # cutedsl 的 kernel 选择随 batch-size 变化，与确定性目标冲突，
        # 直接拒绝组合而非静默产出不确定结果。
        if server_args.enable_deterministic_inference:
            raise ValueError(
                "--bf16-gemm-backend cutedsl is batch-size dependent and cannot "
                "be combined with --enable-deterministic-inference"
            )
        if not is_sm100_supported():
            raise ValueError("--bf16-gemm-backend cutedsl requires an SM10x GPU")

        from sglang.kernels.ops.gemm.cutedsl_bf16_gemm import (
            cutedsl_bf16_gemm,
            use_cutedsl_bf16_gemm,
        )

        _cutedsl_bf16_gemm = cutedsl_bf16_gemm
        _use_cutedsl_bf16_gemm = use_cutedsl_bf16_gemm

    _BF16_GEMM_BACKEND = backend
```

python/sglang/srt/server_args.py

确定性推理校验入口：把 Glm4MoeLiteForCausalLM 纳入 absorbed-MLA 名单，并泛化校验报错文本，决定 GLM-4.7-Flash 能否通过确定性启动校验。

```
  # python/sglang/srt/server_args.py：确定性推理校验的架构探针
  is_deepseek_model = False
  if parse_connector_type(self.model_path) != ConnectorType.INSTANCE:
      try:
          hf_config = self.get_model_config().hf_config
          model_arch = hf_config.architectures[0]
          # absorbed-MLA 模型名单：fa4 后端通过 flash_attn.cute 的 qv 参数
          # 实现确定性 absorbed MLA，仅 SM100/SM110 支持。GLM-4.7-Flash 的
          # Glm4MoeLiteForCausalLM 与 DeepSeek 系列共用这条校验与自动
          # backend 选择路径。
          is_deepseek_model = model_arch in [
              "DeepseekV2ForCausalLM",
              "DeepseekV3ForCausalLM",
```

```
            "DeepseekV32ForCausalLM",
            "MistralLarge3ForCausalLM",
            "PixtralForConditionalGeneration",
            "GlmMoeDsaForCausalLM",
            "Glm4MoeLiteForCausalLM", # 本次新增
        ]
    except Exception:
        pass
```

test/registered/attention/test_glm4_moe_lite_deterministic.py

新增回归测试，同时覆盖显式 fa4 与不指定 backend 的 auto 解析路径，是 BBuf review 指出的测试缺口的直接回应。

```
# test/registered/attention/test_glm4_moe_lite_deterministic.py
GLM_MODEL = "zai-org/GLM-4.7-Flash"

# COMMON_SERVER_ARGS 是共享模块状态，必须拷贝后再扩展：
# 原地 append 会把 fa4 标志泄漏给下面的 auto 测试类，使其
# 悄悄变成第二个 fa4 测试。
SERVER_ARGS = COMMON_SERVER_ARGS + [
    "--chunked-prefill-size",
    "2048",
    "--max-prefill-tokens",
    "2048",
    "--mem-fraction-static",
    "0.8",
]


class TestGlm4MoeLiteFa4Deterministic(TestDeterministicBase):
    @classmethod
    def get_model(cls):
        return GLM_MODEL

    @classmethod
    def get_server_args(cls):
        return SERVER_ARGS + ["--attention-backend", "fa4"]


class TestGlm4MoeLiteAutoBackendDeterministic(TestDeterministicBase):
    @classmethod
    def get_model(cls):
        return GLM_MODEL

    @classmethod
    def get_server_args(cls):
        return SERVER_ARGS

    def test_auto_backend_resolves_to_triton(self):
```

```
# 守护架构探针本身：如果 Glm4MoeLiteForCausalLM 不再被
# 视为 absorbed-MLA 模型，自动填充会返回 flashinfer，
# 而确定性推理会在启动时拒绝该后端。
info = requests.get(DEFAULT_URL_FOR_TEST + "/server_info").json()
self.assertEqual(info["attention_backend"], "triton")
```

评论区精华

BBuf 在整体 review 中评价 "Overall the implementation looks sound, and the manual determinism validation is useful"，但指出测试缺口：验证命令显式设置 `--attention-backend fa4`，未覆盖新改的 auto 解析路径，也未覆盖确定性 BF16 GEMM `auto -> torch` 在 SM10x 上的选择。作者回复 "Added" 并追加提交，新增了覆盖 auto 路径的 nightly 测试。BBuf 的另一条行内评论指出：`initialize_bf16_gemm_config()` 用 `is_sm100_supported()` 接受全部 SM10x（含 SM107），文档却写 SM100/SM103，建议统一为 SM10x；作者在后续文档提交中修正。

- auto 解析路径与 auto -> torch 分支缺少自动化回归 (testing): 作者回应 "Added" 并新增 test_glm4_moe_lite_deterministic.py，在 nightly CI 覆盖显式 fa4 与 auto 解析两个路径；auto -> torch 的 BF16 GEMM 选择仍未独立单测，BBuf 标注 non-blocking。
- 文档 SM100/SM103 与运行时 is_sm100_supported() 的 SM10x 不一致 (documentation): 作者在提交 "docs: describe bf16 GEMM backends as SM10x and document torch" 中统一为 SM10x，并补全 torch 选项描述。

风险与影响

- 风险：
  1. 错误信息措辞变更：server_args.py 中确定性校验错误从 "DeepSeek models" 改为 "absorbed-MLA models"，若外部脚本或监控依赖旧文本匹配，会受影响，概率低。
  2. 显式拒绝组合的行为变更：unquant.py 对 cutedsl + --enable-deterministic-inference 直接抛 ValueError。之前该组合在 SM10x 上可能被接受（虽不确定），升级后启动即失败，属于非渐进式行为变更。
  3. 性能回退：确定性模式下 auto -> torch 意味着 SM10x 上 BF16 GEMM 放弃 CuTe DSL 加速，确定性路径吞吐下降，属预期取舍，但需要让用户明确感知（已文档化）。
  4. 名单语义膨胀：_deterministic_is_deepseek_model 名称仍为 deepseek，但已包含 GLM 系列，未来维护时容易误改；两处名单（server_args 与 overrides）需保持同步。
  5. 测试覆盖缺口：新增 nightly 测试覆盖 auto 解析，但 auto -> torch 的 BF16 GEMM 分支仍无独立单测，BBuf 提出的该点仅部分解决。 - 影响：对用户：GLM-4.7-Flash 用户可在 Blackwell（SM10x）上启用 --enable-deterministic-inference --attention-backend fa4 获得确定性推理；不指定 backend 时自动解析为 triton。对系统：确定性模式的 BF16 GEMM 默认行为在 SM10x 上从 cutedsl 变为 torch，并新增参数组合合法性校验。对团队：新增 nightly 测试条目（估计耗时 900 秒），依赖 4-gpu-gb300 硬件资源。 - 风险标记：确定性路径行为变更，显式拒绝 cutedsl 组合，auto→torch 分支缺少独立单测，架构名单双处维护需同步

关联脉络

- PR #33997 Bump FlashInfer to 0.6.17 and remove Kimi K3 workarounds: 同样涉及 SM10x 上 GEMM/ 量化 kernel 后端选择（mxfp4、trtllm_gen_moe），与本 PR 的 BF16 GEMM backend 演进属于同一内核调度生态。
- PR #34195 [CI] Align rerun-test environment with the test stages: 本 PR 新增 4-gpu-gb300 的 nightly 测试注册，与 CI 环境对齐和 rerun-test 基础设施相关。