PR #34908 完整报告

sgl-project/sglang

Support Intern-S2-Mobius FP8

合并时间：2026-08-20 01:58

原文链接：http://prhub.com.cn/sgl-project/sglang/pull/34908

执行摘要

- 一句话：新增 Intern-S2-Mobius FP8 checkpoint 加载支持
- 推荐动作：建议精读 python/sglang/srt/models/interns2_mobius.py 中 _load_fused_mobius_expert_weight 与 _expected_mobius_load_slots 的改动，它展示了一种用查表替代手写分支来兼容量化 scale 张量的方法，对维护多精度 checkpoint 加载有借鉴意义。同时值得关注其“不写单元测试”的权衡：对于纯序列化映射，端到端验证确实更具说服力，但建议至少在注释中记录回归验证步骤。

功能与动机

PR body 明确说明：internlm/Intern-S2-Mobius-FP8 is now public, but it does not load on current SGLang. 原因是序列化 FP8 配置中 modules_to_not_convert 将 .linear_attn 父路径列为精确匹配项，而 SGLang 把忽略层名称当作前缀匹配，导致量化的子投影被跳过；同时严格加载需要映射 fused expert scale 张量，并允许缺失 KV-cache scale（attn.k_scale / attn.v_scale）。

实现拆解

1. 模型加载器核心调整（python/sglang/srt/models/interns2_mobius.py）：新增 _is_optional_mobius_parameter(name)，将 .attn.k_scale 与 .attn.v_scale 判定为可选参数；在 _load_fused_mobius_expert_weight 中将 gate_up/down 的 scale 后缀通过查表映射到 experts.w13_weight_scale_inv / experts.w2_weight_scale_inv，替代原先手写的单分支替换逻辑。
2. 预期槽位计算同步：_expected_mobius_load_slots 将 *_scale_inv 纳入 .meta_mlp 专家权重槽位，并在剩余分支末尾通过 _is_optional_mobius_parameter 跳过可选参数，保证严格加载时既不会漏报也不会误报。
3. 加载路由扩展：_load_mobius_weights_strict（及其内部路由）将 gate_up_proj_scale_inv / down_proj_scale_inv 纳入 fused expert 张量分发条件，匹配 FP8 checkpoint 的序列化命名。
4. 文档配置更新（docs/src/snippets/configs/internlm/intern-s2-mobius.jsx）：新增 FP8 量化轴、defaultlfp8 模型名映射，加入 H200 已验证的 low-latency / high-throughput FP8 命令 cells，并将 B200 FP8 标记为未验证。

5. Cookbook文案调整（docs/cookbook/autoregressive/InternLM/Intern-S2-Mobius.mdx）：明确现有速度与精度数据仅适用于 BF16 checkpoint，FP8 仅完成启动、文本、视觉与 stop-reason 验证；同步更新模型表格与描述。
6. 测试策略：未新增单元测试。作者在 PR body 中说明这是序列化 checkpoint 集成路径，mock 加载器测试只会复述参数映射；真实路径由端到端验证覆盖（GSM8K 1319 例 96.82% acc、NEXTN 文本 / 视觉冒烟、292 token 有界生成）。

关键文件：

- python/sglang/srt/models/interns2_mobius.py（模块 模型加载；类别 source；类型 data-contract；符号 _is_optional_mobius_parameter, _load_fused_mobius_expert_weight, _expected_mobius_load_slots）：核心加载逻辑调整：新增可选张量判定、扩展 fused expert scale 映射与槽位计算，是 FP8 checkpoint 能通过严格加载的关键。
- docs/src/snippets/configs/internlm/intern-s2-mobius.jsx（模块 文档配置；类别 docs；类型 configuration）：Cookbook 配置新增 FP8 变体、模型名映射及 H200 已验证 / B200 未验证的命令 cells，是用户实际生成部署命令的数据源。
- docs/cookbook/autoregressive/InternLM/Intern-S2-Mobius.mdx（模块 模型文档；类别 docs；类型 documentation）：更新模型描述、表格与速度声明，明确 FP8 无性能基准，防止用户误读数据适用范围。

关键符号：_is_optional_mobius_parameter, _load_fused_mobius_expert_weight, _expected_mobius_load_slots

评论区精华

本 PR 没有实质性的 review 评论或讨论线程，唯一审核人 zijiexia 直接给予 APPROVED。PR body 对测试取舍做了明确说明：No unit test is added: this is a serialized-checkpoint integration path, and a mock loader test would only restate the configured parameter mappings. 从提交历史看，作者通过多次 rerun-failed-ci 修复 CI 波动，最终合并。

- 暂无高价值评论线程

风险与影响

- 风险：
  1. 加载路径回归风险：_load_fused_mobius_expert_weight 与 _expected_mobius_load_slots 是 BF16/FP8 共用的严格加载逻辑，改动可能影响现有 BF16 checkpoint 加载，虽然代码结构上支持原有后缀，但缺少自动化防护。
  2. 缺少单元测试：PR 明确不添加单测，将来若重构加载流程或改动 modules_to_not_convert 语义，FP8 加载可能静默失效，回归只能靠端到端人工门禁发现。
  3. B200 配方未验证：文档中 B200 FP8 recipes 标记为 unverified，若用户在 B200 上直接套用，可能遇到命令无法启动或精度 / 性能未达预期的问题。

4. 性能承诺缺失：FP8 未做速度基准，用户无法预期相对于 BF16 的收益，文档已明确，但若社区误读仍可能产生预期落差。 - 影响：对使用 Intern-S2-Mobius 的用户：可以部署公开的 FP8 checkpoint，显著降低显存占用（H200 上模型权重仅 34.22 GB），并可使用已验证的 NEXTN 推理配方；对系统：模型加载器新增了可选张量与 fused scale 的通用处理模式，为后续同类混合精度 checkpoint 提供参考；对团队：文档侧建立了“已验证 / 未验证”的清晰边界，降低误导风险。整体影响集中在特定模型，范围可控。 - 风险标记：核心加载逻辑变更，缺少单元测试覆盖，B200 配方未验证

关联脉络

- PR #33691 (PR body 中引用的基准 PR): 本 PR 的 BF16 速度与精度声明基于该 PR 的合并版本（main @ e0828ee3 + PR #33691），属于同一 cookbook 功能演进线。