PR #42607 完整报告

vllm-project/vllm

Update Intel Xeon model list and vLLM Benchmark Suite BKMs

合并时间：2026-05-15 13:14

原文链接：http://prhub.com.cn/vllm-project/vllm/pull/42607

执行摘要

本 PR 更新了 Intel Xeon CPU 的模型支持列表和性能基准测试配置，为 vLLM 0.20.2 版本发布做准备。主要变更包括：将 CPU 基准测试配置中的测试用例从多 TP 和 ShareGPT 组合切换为以 TP=1 为主的随机输入测试，并新增大量量化模型和多尺寸模型；同步更新 CPU 模型文档，记录新支持模型和量化变体的状态。

功能与动机

根据 PR body 描述，目的是 "update Intel Xeon model list and vLLM Benchmark Suite BKMs for 0.20.2 release"。随着 CPU 后端支持越来越多的模型和量化格式，需要及时更新文档和基准测试配置，以便用户了解支持情况，并确保在发布前进行覆盖这些新增模型的性能验证。

实现拆解

1. 更新 CPU 基准测试 JSON 配置（.buildkite/performance-benchmarks/tests/serving-tests-cpu-text.json）：删除了原来的 12 个基于 ShareGPT 和不同输入输出长度的测试用例（包括 TP=2 和 TP=4 的配置），替换为 10 个新的测试用例，全部采用随机数据集、输入长度 128、输出长度 128，且 tensor_parallel 固定为 1（70B 模型也使用 TP=1 单机测试）。新增模型包括：Llama-3.2-1B/3B、Llama-3.3-70B、Granite-3.2-2B、Qwen3 系列（1.7B/4B/8B/14B），以及这些模型的 INT4 AWQ 和 INT8 w8a8 量化变体。通过使用固定参数模板和环境变量，大幅简化了配置，同时显著扩大了模型覆盖范围。
2. 更新文档（docs/models/hardware_supported_models/cpu.md）：在文本模型表格中新增约 20 个模型条目，包括 unsloth/gpt-oss-20b、meta-llama/Llama-3.2-1B、meta-llama/Llama-3.3-70B-Instruct、多个 RedHatAI w8a8 量化模型、hugging-quants AWQ INT4 模型、TheBloke AWQ/GPTQ 模型、Qwen3 全系列（包括 MoE 版本）、Phi-4、Mistral AWQ 等。在多模态表格中新增了 Llama-4、Gemma-3、Whisper 等模型。
3. 测试验证：作者在 AWS m8i.24xlarge Intel 实例上运行所有测试并通过，保证了配置的正确性。

以下展示基准测试配置中新增的代表性测试用例（部分）：

```
// 基准测试配置头部保持不变
"server_parameters": {
  "model": "meta-llama/Llama-3.1-8B-Instruct",
  "tensor_parallel_size": 1,
  "dtype": "bfloat16",
```

```
    ...
  },
  "tests": [
    // 新增测试用例：INT4 量化模型
    {
      "test_name": "serving_llama8B_int4_tp1_random_128_128",
      "server_parameters": {
        "model": "hugging-quants/Meta-Llama-3.1-8B-Instruct-AWQ-INT4"
      },
      "client_parameters": {
        "model": "hugging-quants/Meta-Llama-3.1-8B-Instruct-AWQ-INT4",
        "dataset_name": "random",
        "random-input-len": 128,
        "random-output-len": 128
      }
    },
    // 新增测试用例：INT8 量化模型
    {
      "test_name": "serving_llama8B_int8_tp1_random_128_128",
      "server_parameters": {
        "model": "RedHatAI/Meta-Llama-3.1-8B-Instruct-quantized.w8a8"
      },
      "client_parameters": {
        "model": "RedHatAI/Meta-Llama-3.1-8B-Instruct-quantized.w8a8",
        "dataset_name": "random",
        "random-input-len": 128,
        "random-output-len": 128
      }
    },
    // 新增测试用例：70B 大模型
    {
      "test_name": "serving_llama70B_tp1_random_128_128",
      "server_parameters": {
        "model": "meta-llama/Llama-3.3-70B-Instruct"
      },
      "client_parameters": {
        "model": "meta-llama/Llama-3.3-70B-Instruct",
        "dataset_name": "random",
        "random-input-len": 128,
        "random-output-len": 128
      }
    }
    // ... 其他新测试用例类似
  ]
```

文档更新部分示例（新增条目以 + 标记，实际为静态表格）：

```
  <!-- 文本模型表格（部分） -->
  | Model | Architecture | Supported |
  | ---------------------------------- | -------------------------------------- | --------- |
```

unsloth/gpt-oss-20b	GptOssForCausalLM	✅
meta-llama/Llama-3.2-1B	LlamaForCausalLM	✅
meta-llama/Llama-3.3-70B-Instruct	LlamaForCausalLM	✅
RedHatAI/Meta-Llama-3.1-8B-quantized.w8a8	LlamaForCausalLM	✅
hugging-quants/Meta-Llama-3.1-8B-Instruct-AWQ-INT4	LlamaForCausalLM	✅
Qwen/Qwen3-14B	Qwen3ForCausalLM	✅
Qwen/Qwen3-30B-A3B	Qwen3MoeForCausalLM	✅
microsoft/Phi-4-reasoning	Phi3ForCausalLM	✅
<!-- 多模态表格也新增了 Llama-4、Gemma-3 等模型 -->

评论区精华

本 PR 没有实质性的 review 讨论。仅有的机器人评论（Claude Code Review 和 Gemini Code Assist）都来自自动化工具，未产生技术讨论；审批人 bigPYJ1151 直接批准，说明变更清晰无争议。

风险与影响

- 风险：低风险。仅修改测试配置和文档，不涉及核心代码。主要风险是基准测试配置中模型名称或路径错误可能导致测试失败，但作者已实际运行并通过。文档与配置的同步需要确保一致性，但本次更新同时修改了二者，避免了不一致。
- 影响：
  - 对用户：文档更新提供了更准确的 CPU 支持信息，特别是量化模型的可用性。
  - 对测试：基准测试配置简化（删除多 TP 和长文本测试），新增多个模型覆盖，有助于提升回归测试的效率和覆盖面。
  - 对团队：减少维护成本，统一测试策略为 TP=1 随机输入，便于对比不同模型的性能。

关联脉络

本次 PR 是 vLLM CPU 后端持续演进的一部分。近期 PR #40119 添加了 RVV 优化的注意力 kernel，提升了 CPU 推理性能；本次 PR 则从模型支持文档和基准测试角度，保障了这些优化的可验证性。两者共同体现了 Intel 团队对 CPU 后端的系统性投入。