PR #31363 完整报告

sgl-project/sglang

docs(cookbook): re-benchmark DeepSeek-V4 on sglang 0.5.15

合并时间：2026-07-22 06:55

原文链接：http://prhub.com.cn/sgl-project/sglang/pull/31363

执行摘要

本 PR 将 DeepSeek-V4 cookbook 中全部 36 个单节点基准测试单元格的数据从 sglang v0.5.12.post1 更新到 v0.5.15/v0.5.15.post1，所有数值均在 cache-cold 模式下重新测量并使用 P50 百分位。同时增强了渲染引擎 _deployment.jsx，支持 per-cell 级别的 latencyPercentile 覆盖，以便准确展示旧版 Mean 数据。新增了 B200/B300 NVFP4 低延迟单元格，并修复了 h200 pro 单元格的 detokenizer 挂死问题（调整 mem-fraction-static 至 0.90）。Review 过程中进行了详尽的数据一致性审计，所有异常点均被重新测量并验证。

功能与动机

根据 PR body，本次变更的核心目标是：

- "Re-benchmarks the DeepSeek-V4 cookbook on sglang v0.5.15 / v0.5.15.post1(single-node cells), replacing the prior 0.5.12.post1 numbers."
- 新增 B200/B300 NVFP4 低延迟单元格（此前在 benchmarks.jsx 中缺失）。
- 解决 h200 pro fp4 low-latency 单元格在 --mem-fraction-static 0.83 下的 detokenizer 挂死问题，通过提高至 0.90 得以正常运行。
- 统一所有数据为 P50 百分位，并通过 per-cell 覆盖机制保留旧版 Mean 单元格（如 GB200）。

实现拆解

1. 数据更新（deepseek-v4-benchmarks.jsx）：逐个更新 36 个单元格的 sglang_version 和性能指标。测量方法：sglang.launch_server 配合精确 cookbook 参数，然后 sglang.bench_serving 对于每个 concurrency 点。所有运行均使用 --flush-cache 确保 cache-cold。每次测量进行 3 次运行，取第一次运行（run1）的值，并记录了 run2/run3 以证明可复现性。新增 B200/B300 NVFP4 低延迟单元格。GB200 单元格恢复但带上 latencyPercentile: "Mean" 以正确标注。
2. 渲染引擎增强（_deployment.jsx）：修改 renderBenchmarkCard 中 pct 变量的解析逻辑，从 config.latencyPercentile || "P50" 改为 (entry && entry.latencyPercentile) || config.latencyPercentile || "P50"。这个极小的改动使得每个基准条目可以携带自己的百分位覆盖，例如 GB200 的 Mean 数据。添加了相应的文档注释说明解析优先级。
3. 模型配置调整（deepseek-v4.jsx）：三处关键修改——全局 latencyPercentile 从 "Mean" 改为 "P50"；speed benchmark 命令添加 --flush-cache 参数以匹配 cache-cold 测量方法；h200 pro fp4 low-latency 单元格的 --mem-fraction-static 从 0.83 改为 0.90 以解决

detokenizer 挂死。

4. 模板与技能文档同步：benchmarks.jsx.tmpl 和 config.jsx.tmpl 更新以反映新的数据契约（per-cell latencyPercentile 注释、--flush-cache 参数）。cookbook-review-pr/SKILL.md 新增审核检查点，要求 benchmark 命令必须包含 --flush-cache，并检查 per-cell 百分位标注是否正确。authoring-reference.md 同步更新。

docs_new/src/snippets/configs/deepseek-ai/deepseek-v4-benchmarks.jsx

核心数据文件，包含所有 36 个单节点单元格的基准测试数据，从 0.5.12.post1 更新到 0.5.15/0.5.15.post1，新增 NVFP4 单元格，恢复 GB200 单元格（标注为 Mean）。

```
// DeepSeek-V4 per-cell benchmark numbers, keyed by the same `match` tuple as
// deepseek-v4.jsx cells. See _deployment.jsx for the speed/accuracy schema.
// Measured on sglang v0.5.15 / v0.5.15.post1 (per-cell sglang_version).
// tokens_per_sec_per_gpu is total (input+output) tok/s/GPU = output/GPU × (isl+osl)/osl.
export const benchmarks = [
  // ============================================================================
  // B200 + FP4
  // ============================================================================
  {
    match: { hw: "b200", variant: "flash", quant: "fp4", strategy: "low-latency", nodes: "single" },
    sglang_version: "0.5.15",
    speed: [
      { workload: { dataset: "random", isl: 8192, osl: 1024, max_concurrency: 1 },
        ttft_ms: 302, tpot_ms: 2.91, tokens_per_sec_per_gpu: 677 },
      { workload: { dataset: "random", isl: 8192, osl: 1024, max_concurrency: 16 },
        ttft_ms: 454, tpot_ms: 8.76, tokens_per_sec_per_gpu: 3059 },
    ],
  },
  // B200 NVFP4 (newly added)
  {
    match: { hw: "b200", variant: "flash", quant: "nvfp4", strategy: "low-latency", nodes: "single" },
    sglang_version: "0.5.15",
    speed: [
      { workload: { dataset: "random", isl: 8192, osl: 1024, max_concurrency: 1 },
        ttft_ms: 308, tpot_ms: 2.88, tokens_per_sec_per_gpu: 682 },
      { workload: { dataset: "random", isl: 8192, osl: 1024, max_concurrency: 16 },
        ttft_ms: 466, tpot_ms: 8.67, tokens_per_sec_per_gpu: 3059 },
    ],
  },
  // GB200 cell with per-cell latencyPercentile set to "Mean"
  {
    match: { hw: "gb200", variant: "flash", quant: "nvfp4", strategy: "low-latency", nodes: "single"
    },
    sglang_version: "0.5.12.post1 (PR #25820)",
    latencyPercentile: "Mean", // overrides global P50; legacy data not re-measured
    speed: [ /* ... */ ],
  },
]
```

docs_new/src/snippets/_deployment.jsx

渲染引擎，负责将基准数据渲染到 cookbook 页面。本次修改为支持 per-cell latencyPercentile，通过回退链（entry → config → "P50"）实现，保持向后兼容。

```
const renderBenchmarkCard = (entry) => {
  // [key, label, unit, compute?]. Optional compute(measurement) supplies
  // derived metrics (preferred over measurement[key] when present).
  // pct 的解析策略：优先使用 entry 自带的 latencyPercentile（用于遗留 Mean 数据），
  // 其次回退到 config 级别设置（页面级 P50 或 Mean），最后默认 "P50".
  const pct = (entry && entry.latencyPercentile) || config.latencyPercentile || "P50";
  const SPEED_LABELS = [
    ["ttft_ms", `TTFT (${pct})`, "ms"],
    ["tpot_ms", `TPOT (${pct})`, "ms"],
    // ...
  ];
  // ...
}
```

docs_new/src/snippets/configs/deepseek-ai/deepseek-v4.jsx

DeepSeek-V4 模型的页面配置，包括硬件支持、变体、量化、启动参数等。本次修改涉及三大配置变化：latencyPercentile 从 Mean 改为 P50，speed 命令添加 --flush-cache，h200 pro 的 mem-fraction-static 从 0.83 改为 0.90。

```
export const config = {
  modelName: "DeepSeek-V4",

  latencyPercentile: "P50", // changed from "Mean"; all re-measured cells use P50

  // ... supportedHardware, hardware, variants, quantizations, strategies, nodesOptions ...

  benchmarkCommands: {
    speed:
`python3 -m sglang.bench_serving \
  --backend sglang \
  --host {{CURL_HOST}} --port {{CURL_PORT}} \
  --model {{MODEL_NAME}} \
  --dataset-name {{DATASET}} \
  --random-input-len {{ISL}} --random-output-len {{OSL}} \
  --num-prompts {{NUM_PROMPTS}} --max-concurrency {{MAX_CONCURRENCY}} \
  --warmup-requests 64 --flush-cache`, // added --flush-cache to match cache-cold
  measurement

    // ... accuracy commands
  },

  cells: [
    // ... h200 pro fp4 low-latency cell
    {
```

```
      match: { hw: "h200", variant: "pro", quant: "fp4", strategy: "low-latency", nodes: "single" },
      verified: true,
      env: [],
      flags: [
        "--trust-remote-code",
        "--model-path {{MODEL_NAME}}",
        "--tp 8",
        "--mem-fraction-static 0.90", // changed from 0.83; avoids detokenizer hang
        "--host {{HOST_IP}}",
        "--port {{PORT}}",
        // ...
      ],
    },
  ],
}
```

评论区精华

zijiexia: "Ran a full consistency audit of the re-benchmarked numbers: joined each cell with its serve command's --tp, checked all 78 measurement rows against the closed-loop identity... Overall the data holds up well..." 审核者对所有 78 个测量行进行了 closed-loop 上界检验，发现并发 4096 时 TPOT 不应低于并发 1024 的 TPOT（b300 flash ht），以及 b300 pro ht 的 TTFT 缩放异常大等问题。这些均被作者重新测量后更正。

zijiexia（关于 GB200 单元格）: "The removal rationale here (Mean-era numbers would be mislabeled by the global latencyPercentile: "P50") applies equally to the two gb200 nvfp4 cells below... Suggest treating them the same way (blank them) or annotating them explicitly." 最终作者选择更优方案：添加 per-cell latencyPercentile 支持，保留 GB200 数据并正确标注。

zijiexia: "maybe just keep as latest, we would recommend users to use the latest version of sglang" 关于 dockerImages 固定版本还是使用 latest 的讨论，作者采纳 latest。

dougyster: 对每个被标记的异常点都提供了 3 次 cache-cold 重复测量数据（如 b200 flash ht conc-4096 run1/2/3 的 tps/GPU 分别为 8345/8368/8371），证明了数据的可复现性。

风险与影响

- 数据准确性：所有测量均经过 closed-loop 一致性检验（通过 throughput <= conc x (isl+osl) / (TTFT + (osl-1) x TPOT) / n_gpu），并通过 3 次重复测量验证，风险较低。但部分 B300 数据由于硬件可用性问题，测量时间窗口较短，可能未捕捉到全部变异性。
- GB200 旧数据：保留的 Mean 数据与页面 P50 标签不同，虽然通过 per-cell 覆盖机制正确处理，但读者可能忽略版本差异（0.5.12.post1 vs 0.5.15）。
- 引擎改动影响：_deployment.jsx 的改动极小且 fallback 链完整，不会破坏现有其他 cookbook 页面。但无独立测试覆盖。

- 对用户的影响：cookbook 页面展示的性能数据更准确、更完整；新增的 NVFP4 单元格填补空白；h200 pro 单元格重新可用。
- 对团队的影响：确立了 cache-cold + 3 次运行取第一次 + closed-loop 审计的基准测试标准流程，为后续 cookbook 维护提供了可复用的方法论。

关联脉络

- 历史 PR #25820：GB200 原始基准测试 PR，本次保留其 Mean 数据。
- 历史 PR #31373：MegaMoE a2a 后端支持，本次 b200 pro balanced 单元格切换使用该后端。
  - 本 PR 与近期其他 PR（如 #31941 移除废弃 auto-benchmark）共同体现了 sglang 团队对基准测试数据质量和可复现性的持续投入。
  - per-cell 百分位覆盖机制是部署引擎的一个微小但重要的外扩，为未来其他模型（如 Laguna、Inkling）的 cookbook 维护提供了灵活性。