PR #33109 完整报告

sgl-project/sglang

[Docs] Add verified H200 and B200 DeepSeek-V4 Flash Official results

合并时间：2026-08-01 16:05

原文链接：http://prhub.com.cn/sgl-project/sglang/pull/33109

执行摘要

本 PR 是 #33083 的 follow up：在 DeepSeek-V4 Flash Official (0731) 部署配方基础上，补齐 4 台 H200 与 8 卡 B200 的 FP4 端到端验证，将 6 个 flash-official 部署 cell 标记为 verified，并发布 12 条 SGLang v0.5.16 实测压测记录。部署参数有两处直接影响用户复制的修正：B200 三个策略从 TP4/DP4 切换为实测的 TP8/DP8，并为 H200 balanced 增加 --mem-fraction-static 0.88、为 B200 low-latency 增加 0.90，以规避 FlashInfer MoE workspace 与 draft CUDA-graph 捕获阶段的 OOM。纯文档与基准数据变更，不涉及运行时源码，已获 zijiexia 批准。

功能与动机

- 承接未完成的验证：PR body 明确说明 "follow up on #33083 with completed Flash Official FP4 validation on 4xH200 and 8xB200"。上一版配方中 H200/B200 的 flash-official cell 全部是 verified: false，本 PR 将状态补齐。
- 默认内存参数在真实硬件上不可用：H200 上 FlashInfer MoE workspace 需要 1.34 GiB 而仅剩 1.06 GiB；B200 默认 mem-fraction-static 0.944 在 draft CUDA-graph 捕获阶段需 632 MiB、仅剩 371 MiB。必须把验证过的内存预留值固化进配方，否则用户按默认值启动必然 OOM。
- 社区数据诉求：Swipe4057 询问是否加入 FP8 checkpoint 对比，作者澄清本 PR 针对官方 0731 版本（deepseek-ai/DeepSeek-V4-Flash-0731），不覆盖 FP8 变体。

实现拆解

1. 压测数据入档（deepseek-v4-benchmarks.jsx，+60 行）：为 B200 与 H200 的 flash-official + FP4 三个策略（low-latency / balanced / high-throughput）各新增一个 cell，共 6 个 cell、12 条记录；每条记录含 ttft_ms、tpot_ms、tokens_per_sec_per_gpu，版本标注为 v0.5.16。数据按 match 元组（hw / variant / quant / strategy / nodes）与部署 cell 一一对应，旧的 flash 非 official 记录完整保留，便于对照新旧配方。
2. 部署配方更新（deepseek-v4.jsx，+13/-11）：6 个 flash-official cell 的 verified 改为 true；B200 三个策略从 --tp 4 / --dp 4 切换为实测的 --tp 8 / --dp 8；为 B200 low-latency 增加 --mem-fraction-static 0.90，为 H200 balanced 增加 --mem-fraction-static 0.88。这两处内存参数是本次实测中必须的避坑项。
3. 手册一致性（DeepSeek-V4.mdx，+1/-1）：模型表格中 Flash Official (0731) 的验证描述由 "verified for low-latency serving on 4×GB300" 扩展为 "verified on 8×B200, 4×GB300, and 4×H200"。

4. 质量校验配套：作者执行了 pre-commit、esbuild 语法检查、npx mint validate、npx mint broken-links --check-anchors、矩阵断言（三个 H200/B200 flash-official cell 均为 verified、B200 的 TP8/DP8 与三条压测记录匹配）及 git diff --check。CI 中 PR Test (Base) 通过、PR Test (Extra) 失败，失败原因未在 PR 内说明。

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

部署配方核心：6 个 flash-official cell 从 verified: false 改为 true，B200 三个策略从 TP4/DP4 切换为实测的 TP8/DP8，并为 B200 low-latency 与 H200 balanced 新增 --mem-fraction-static 0.90/0.88。这些 flags 会被 cookbook 页面直接渲染为用户可复制的启动命令，是本 PR 对用户行为影响最大的文件。

```
// DeepSeek-V4 Flash Official 部署配方 cell（docs_new/src/snippets/configs/deepseek-ai/
deepseek-v4.jsx）
// 每个 cell 通过 match 元组（硬件 / 变体 / 量化 / 策略 / 节点数）与 benchmarks.jsx
的压测记录一一对应。
// verified: true 表示该组合已在目标硬件上跑通并留有正式基准数据，配方可直接照抄到部署命令。

{
  // B200 低延迟：实测 TP8 + FlashInfer MXFP4 + DSpark
  match: { hw: "b200", variant: "flash-official", quant: "fp4", strategy: "low-latency", nodes:
  "single" },
  verified: true,
  env: [],
  flags: [
    "--trust-remote-code",
    "--model-path {{MODEL_NAME}}",
    "--tp 8", // 8 卡 B200 实测配置，替代原 TP4
    "--moe-runner-backend flashinfer_mxfp4",
    "--speculative-algorithm DSPARK",
    "--chunked-prefill-size 4096",
    "--disable-flashinfer-autotune",
    "--swa-full-tokens-ratio 0.1",
    "--mem-fraction-static 0.90", // 默认 0.944 在 draft CUDA-graph 捕获阶段 OOM（需 632
    MiB、仅剩 371 MiB）
    "--host {{HOST_IP}}",
    "--port {{PORT}}",
  ],
},

{
  // H200 均衡：实测 TP4 + FlashInfer MXFP4 + DSpark
  match: { hw: "h200", variant: "flash-official", quant: "fp4", strategy: "balanced", nodes: "single"
  },
  verified: true,
  env: [],
  flags: [
    "--trust-remote-code",
    "--model-path {{MODEL_NAME}}",
```

```
    "--tp 4",
    "--moe-runner-backend flashinfer_mxfp4",
    "--speculative-algorithm DSPARK",
    "--mem-fraction-static 0.88", // 默认静态内存下 FlashInfer MoE workspace OOM（需 1.34
    GiB、仅剩 1.06 GiB）
    "--host {{HOST_IP}}",
    "--port {{PORT}}",
  ],
},
```

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

承载本 PR 的 12 条 v0.5.16 实测压测记录：B200 与 H200 的 flash-official FP4 三策略各 2 个并发点（TTFT/TPOT/tok per s per GPU）。数据按 match 元组与 deepseek-v4.jsx 的 cell 强关联，是 verified 状态的数据支撑。

```
// 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.
// 新增记录统一标注 sglang v0.5.16，tokens_per_sec_per_gpu 为总吞吐（输入 + 输出）。

export const benchmarks = [
  // ============================================================================
  // B200 + FP4 Flash Official（v0.5.16 实测，8 卡 TP8）
  // ============================================================================
  {
    match: { hw: "b200", variant: "flash-official", quant: "fp4", strategy: "low-latency", nodes:
    "single" },
    sglang_version: "0.5.16",
    speed: [
      { workload: { dataset: "random", isl: 8192, osl: 1024, max_concurrency: 1 },
        ttft_ms: 218.90, tpot_ms: 1.28, tokens_per_sec_per_gpu: 481 },
      { workload: { dataset: "random", isl: 8192, osl: 1024, max_concurrency: 16 },
        ttft_ms: 189.89, tpot_ms: 3.38, tokens_per_sec_per_gpu: 3383 },
    ],
  },
  {
    // 高吞吐策略：MegaMoE target-only，无 DSpark
    match: { hw: "b200", variant: "flash-official", quant: "fp4", strategy: "high-throughput", nodes:
    "single" },
    sglang_version: "0.5.16",
    speed: [
      { workload: { dataset: "random", isl: 8192, osl: 1024, max_concurrency: 1024 },
        ttft_ms: 74853.51, tpot_ms: 51.72, tokens_per_sec_per_gpu: 5464 },
      { workload: { dataset: "random", isl: 8192, osl: 1024, max_concurrency: 4096 },
        ttft_ms: 216241.01, tpot_ms: 51.80, tokens_per_sec_per_gpu: 5301 },
    ],
  },
  // ... H200 三条 flash-official 记录结构相同，tp 4、mem-fraction-static 0.88 见 deepseek-v4.jsx
];
```

评论区精华

PR 无行级 review 评论，zijiexia 以空 body 批准合并；实质讨论集中在 Issue 评论区：

> Swipe4057：Is it possible to add a comparison with the checkpoint ... DeepSeek-V4-Flash-FP8 on H200?
>
> JustinTong0323：This PR is to add the bench result for Flash official checkpoint released yesterday: deepseek-ai/DeepSeek-V4-Flash-0731

即本 PR 数据范围限定为官方 0731 版本，FP8 对比不在范围内。其余为 gemini-code-assist 服务下线提示与 mintlify 预览部署的机器人消息，无技术内容。

风险与影响

- 参数与实测环境强绑定：mem-fraction-static 0.88/0.90 是在 4×H200（driver 580.159.03）与 8×B200（driver 580.126.09、NV18 互联）上标定的，其他驱动、GPU 组合（如 8×H200、4×B200）或 CUDA 版本下所需预留空间可能不同。
- 内存预留降低容量的副作用：下调 mem-fraction-static 会压缩 KV cache 可用空间，PR 验证的 workload 固定为 isl 8192 / osl 1024，长上下文或更高并发场景下可能减少可服务请求数。
- 跨版本对比仅具方向性：PR body 自述与旧 H200 Flash FP4 数据、与 GB300 的对比是 " directional across checkpoint and SGLang-version changes"，不建议当作严格性能结论。
- 新旧配方并存易混淆：B200 的 flash-official 已切到 TP8/DP8，而相邻的 flash（EAGLE）cell 仍是 TP4/DP4，读者需注意 variant 字段差异。
- CI 状态不一致：PR Test (Extra) 失败但未在 PR 内澄清原因，虽为文档改动，仍存在文档构建链路隐患待确认。

关联脉络

- #33083（Add DeepSeek-V4 Flash Official (0731) recipe）：本 PR 的直接上游，前者只提供配方与 GB300 验证，后者补齐 H200/B200 实测并修正参数，二者构成 " 配方发布 + 实测回填 " 的完整闭环。
- #33157（Add RTX 5090 DeepSeek-V4 recipe）：同样修改 deepseek-v4.jsx，与本次共同完善 DeepSeek-V4 在多硬件（GB300、H200、B200、RTX 5090）上的部署覆盖。
- 从更长的演进线看，sglang 仓库正通过 cookbook/recipe + benchmarks 数据矩阵的方式，把 " 经过验证的部署配方 " 作为一等交付物，本 PR 是这一流程在 DeepSeek-V4 上的落地样板。