PR #33083 完整报告

sgl-project/sglang

[Docs] Add DeepSeek-V4 Flash Official (0731) recipe

合并时间：2026-08-01 03:14

原文链接：http://prhub.com.cn/sgl-project/sglang/pull/33083

执行摘要

本 PR 为 DeepSeek-V4-Flash-0731 checkpoint 新增了独立的 Flash Official文档配方：在 Docs 站点上扩展 flash-official variant、22 个 FP4 硬件 / 策略矩阵单元，并发布 GB300 / H200 上实测的 serving 与 accuracy 数据。与此同时，它改造了所有模型页面共用的 Playground 交互引擎，使 --speculative-dspark-block-size 可被识别和清理，并在 DSpark 与 DP Attention、PD 拆分、ROCm 组合不成立时正确地隐藏或禁用选项。该 PR 不涉及服务端代码，但 Playground 引擎改动影响面覆盖整个文档站点。

功能与动机

DeepSeek 发布的 0731 Flash refresh 自带 DSpark draft head，部署方式与 EAGLE 配方的 Flash/Pro 完全不同：不需要单独的 draft model，也不需要 --speculative-num-steps / --speculative-eagle-topk / --speculative-num-draft-tokens。PR body 明确写道：要“add a separate Flash Officialoption for deepseek-ai/DeepSeek-V4-Flash-0731”，并“prevent unsupported DSpark combinations with DP Attention, PD disaggregation, and ROCm”。同时作者做了大量实测，包括 4xGB300 与 4xH200 上的六个正式 serving workload，目的是让用户拿到经过端到端验证的启动命令与调优指引，而不是未经证实的猜测性配置。

实现拆解

1. 模型配置扩展（docs_new/src/snippets/configs/deepseek-ai/deepseek-v4.jsx） - 新增 flash-official variant（label “Flash Official”，subtitle “284B • 0731”），并将 flash-officialIfp4 映射到 deepseek-ai/DeepSeek-V4-Flash-0731。 - speculative.options 增加 dspark 预设（--speculative-algorithm DSPARK），对 flash / pro 隐藏，并在 dpAttnOn: true 或 AMD mi300x / mi355x 时禁用；原有的 mtp-314 / mtp-112 对 flash-official 隐藏。 - pdDisagg 新增 showWhen: (base) => base.specAlgorithm !== "DSPARK" 与 incompatibleSpeculativeAlgorithms: ["DSPARK"]，从配置层面声明 DSpark 与 PD 拆分的互斥。 - 新增 flagSelects 调优控件 dsparkDraftTokens：一个 slider，覆盖 --speculative-dspark-block-size 1–5，且仅在 flash-official + DSPARK 时显示。
2. Playground 引擎适配（docs_new/src/snippets/_playground.jsx） - speculative axis 的 deriveFromBase 与 apply 把 --speculative-dspark-block-size 纳入识别集合和 strip 集合，避免切换算法时残留 DSpark 参数。 - pdDisagg.apply 与 getRenderHints 增加算法兼容性守卫：当 --speculative-algorithm 为 DSPARK 时直接保持 flags 不变，不输出任何

PD 拆分参数。 - applyAllDeltas 从当前 flags 中实时提取 specAlgorithm 并注入 liveSel，后续 axis 的禁用条件可以直接读取；渲染阶段还新增一次 applyAllDeltas 计算 constraintEffective，用于推导 pdMode 与 specAlgorithm 的最终值。 - MTP 警告提示从固定文案改为按实际算法显示名称（EAGLE → MTP、DSPARK → DSpark、NGRAM → N-gram 等）。

3. 基准数据落地（docs_new/src/snippets/configs/deepseek-ai/deepseek-v4-benchmarks.jsx） - 为 GB300 flash-official 的 low-latency（concurrency 1/16）、balanced（64/256）、high-throughput（1024/4096）三个单元补充实测 speed 数据，low-latency 单元还带 GPQA / AIME25 / GSM8K accuracy；版本标记为 0.5.16。
4. Cookbook 文档（docs_new/cookbook/autoregressive/DeepSeek/DeepSeek-V4.mdx） - 模型介绍表格新增 Flash Official (0731) 行（304B / 13B 激活），新增 “3.4 DSpark (Speculative Decoding)” 章节，给出 4xGB300 已验证命令（--tp 4、--moe-runner-backend flashinfer_mxfp4、--speculative-algorithm DSPARK、--mem-fraction-static 0.90、--chunked-prefill-size 4096、--swa-full-tokens-ratio 0.1），并说明冷启动耗时 10–15 分钟、draft block-size 调优方法与验收指标。
5. 测试与验证配套 - 没有新增单元测试。作者在 PR body 中声明了 pre-commit、npx mint validate、npx mint broken-links --check-anchors、矩阵不变量检查，以及 live Mintlify preview 验证（确认 DSpark 在 DP Attention 与 ROCm 下被禁用、切换算法时 block-size 参数被正确清除）。

关键源码片段

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

DeepSeek-V4 配置核心：新增 flash-official variant、DSpark 预设与禁用条件、PD 拆分兼容性声明及 22 个矩阵 cell。

```
// ----- Card 4: "Speculative Decoding" -----
// Flash Official (0731) 使用 checkpoint 内置的 DSpark draft head,
// 因此 EAGLE / MTP 预设在该 variant 下隐藏；DSpark 仅对 flash-official 开放,
// 并声明与 DP Attention、AMD ROCm 的互斥条件。
speculative: {
  options: [
    { id: "current", label: "Inherited from base" },
    { id: "off", label: "Off (greedy)" },
    { id: "mtp-314", label: "EAGLE / MTP 3-1-4",
      flags: ["--speculative-algorithm EAGLE", "--speculative-num-steps 3",
              "--speculative-eagle-topk 1", "--speculative-num-draft-tokens 4"],
      hide: { variant: ["flash-official"] } },
    { id: "mtp-112", label: "EAGLE / MTP 1-1-2",
      flags: ["--speculative-algorithm EAGLE", "--speculative-num-steps 1",
              "--speculative-eagle-topk 1", "--speculative-num-draft-tokens 2"],
      hide: { variant: ["flash-official"] } },
    // DSpark：单 flag 预设，不附带 draft model 路径——权重内置于 checkpoint。
    { id: "dspark", label: "DSpark",
      flags: ["--speculative-algorithm DSPARK"],
      hide: { variant: ["flash", "pro"] },
```

```
      disable: [
        { when: { dpAttnOn: [true] },
          reason: "DSpark is not compatible with DP Attention on the current release." },
        { when: { hw: ["mi300x", "mi355x"] },
          reason: "DSpark currently requires CUDA." },
      ] },
    { id: "ngram", label: "NGRAM",
      flags: ["--speculative-algorithm NGRAM",
              "--speculative-num-draft-tokens 16",
              "--speculative-ngram-max-bfs-breadth 10"],
      disable: { dpAttnOn: [true] },
      disableReason: "NGRAM is incompatible with DP-Attention. Turn DP-Attention off in the
      Attention card above to use NGRAM." },
    { id: "dflash", label: "DFlash", disabled: true,
      disableReason: "Coming soon — pending DFlash kernel integration." },
  ],
},

// ----- Card 5: "PD Disaggregation" -----
// DSpark 与 PD 拆分互斥：配置层直接声明不兼容算法，引擎据此隐藏整张卡。
pdDisagg: {
  showWhen: (base) => base.specAlgorithm !== "DSPARK",
  incompatibleSpeculativeAlgorithms: ["DSPARK"],
  modes: [
    { id: "off", label: "Off" },
    { id: "prefill", label: "Prefill role" },
    { id: "decode", label: "Decode role" },
  ],
  // ...transferBackends / ibDevices / router 保持不变，省略。
},

// 新增的 Playground 调优控件：DSpark 每步提议的 draft token 数。
// showWhen 依赖引擎注入的 specAlgorithm——只有 Flash Official + DSPARK 才显示。
flagSelects: [
  {
    id: "dsparkDraftTokens",
    title: "DSpark Proposed Draft Tokens",
    showWhen: (base) => base.variant === "flash-official" && base.specAlgorithm === "DSPARK",
    control: "slider",
    stripPrefixes: ["--speculative-dspark-block-size"],
    options: [
      { id: "auto", label: "Checkpoint default" },
      { id: "1", label: "1", flags: ["--speculative-dspark-block-size 1"] },
      { id: "2", label: "2", flags: ["--speculative-dspark-block-size 2"] },
      { id: "3", label: "3", flags: ["--speculative-dspark-block-size 3"] },
      { id: "4", label: "4", flags: ["--speculative-dspark-block-size 4"] },
      { id: "5", label: "5", flags: ["--speculative-dspark-block-size 5"] },
    ],
  },
```

],

评论区精华

本 PR 没有产生人类 review 讨论。唯一一条 reviewer 批准来自 zijiexia，且没有附带说明；评论区只有三条机器人消息：两条 Gemini Code Assist 停用通知和一条 Mintlify 预览链接。因此，最有价值的信息源是作者在 PR body 中留下的实测记录与决策解释，例如：

> "The balanced cell therefore remains DP4 DeepEP target-only: the TP4 DSpark candidate is a useful lower-concurrency operating point, but it does not replace the balanced topology across both published workloads."
>
> "The first balanced concurrency-64 warmup exposed a FlashInfer MoE workspace OOM: a 1.34 GiB request had only 1.06 GiB free at the default static-memory setting. That failed run is excluded."

这些记录展示了"测量驱动决策"的文档维护方式：哪个矩阵单元被保留、哪个被降级为参考点，都有明确的数据依据。

风险与影响

- 全局组件回归：_playground.jsx 是所有模型 Playground 页面的公共引擎，本次修改了 pdDisagg.apply / getRenderHints 签名并新增 constraintEffective 的第二遍 apply 管线，逻辑错误会影响非 DeepSeek 模型页面；该文件没有单元测试。
- 约束管线重复执行：渲染时 applyAllDeltas 会被执行两次（一次用于最终命令，一次用于 constraintEffective），依赖 handler 的纯函数性质；若有状态副作用，pdMode 或 flags 可能在两次执行间不一致。
- 未验证矩阵单元可见：B200 等 flash-official cell 标注为 verified: false，但仍会在页面中展示，用户可能误以为这些命令已经过端到端验证。
- DSpark 互斥约束分散：DSpark 与 DP Attention / PD 拆分 / ROCm 的互斥同时由配置数据（disable、incompatibleSpeculativeAlgorithms）和引擎逻辑（liveSel 注入）保证，未来 checkpoint 行为变化时需要两处联动维护。

关联脉络

本 PR 是 DeepSeek-V4 文档矩阵的一次重要扩展，与近期 speculative decoding 与 GB300 测试相关的 PR 处于同一演进脉络：#32920 优化了 speculative target-verify mask 的压缩逻辑，与 DSpark 的 verify 图行为相互影响；#33044 固定 GB300 PR 测试的 NCCL 端口，说明 GB300 测试环境正在逐步稳定，也支撑了本 PR 中 4xGB300 实测数据的可复现性。整体来看，SGLang 正在把"模型 + 硬件 + 算法"的部署知识沉淀为文档站点的结构化配置，并通过 Playground 引擎把约束内嵌到交互命令生成中。