PR #23697 完整报告

sgl-project/sglang

update: b300 container for dsv4

合并时间：2026-04-25 13:59

原文链接：http://prhub.com.cn/sgl-project/sglang/pull/23697

执行摘要

PR#23697 为 DeepSeek-V4 的部署方案新增 NVIDIA B300 硬件平台支持。通过将 B300 映射为 B200 的配置，在部署代码片段和深度指南文档中增加了 B300 选项和对应的容器镜像说明，改动量小且风险极低。

功能与动机

NVIDIA B300 是 Blackwell 架构的新 GPU，需要纳入 DeepSeek-V4 已验证的硬件部署矩阵。该 PR 让使用 B300 的用户能够从部署交互组件中获得正确的启动命令和容器镜像指引，与近期多项 DeepSeek-V4 文档 PR（如 #23690、#23691）共同完善硬件覆盖。

实现拆解

1. 硬件选择器增加 B300 选项：在 deepseek-v4-deployment.jsx 的 options.hardware.items 数组中新增 id 为 "b300"、标签为 "B300 (FP4)" 的条目。
2. B300→B200 映射逻辑：在 generateCommand() 和 buildPDDisaggCommand() 中，将 rawHardware 参数在值为 "b300" 时映射为 "b200"。这样无需为 B300 定义独立的 HW_SIZE_SPEC、环境变量列表等，所有配置直接复用 B200 已验证的条目。

```
const generateCommand = () => {
  const { hardware: rawHardware, modelSize, recipe, reasoningParser, toolcall } = values;
  // B300 的使用方式与 B200 完全相同 — 做别名以不重复每个 spec 条目
  const hardware = rawHardware === "b300" ? "b200" : rawHardware;
  const specKey = `${hardware}|${modelSize}`;
  const spec = HW_SIZE_SPEC[specKey];
  const { slug, tp, multinode, nnodes } = spec;
  const isBig = modelSize === "big";
  // ...
};
```

1. 文档新增 B300 容器镜像行：在 DeepSeek-V4.mdx 的硬件表格中插入 B300 行，显示 lmsysorg/sglang:deepseek-v4-b300 镜像，与已存在的 B200、H200、GB300 行并列。

```
<tr>
  <td>NVIDIA B300</td>
  <td><code>lmsysorg/sglang:deepseek-v4-b300</code></td>
</tr>
```

该 PR 不含测试改动，也不包含运行时代码变更，纯属文档和部署代码片段的更新。

docs_new/src/snippets/autoregressive/deepseek-v4-deployment.jsx

实现了 B300 的硬件选择器和 B300→B200 映射逻辑，是部署命令生成的核心文件。

```
const options = {
  hardware: {
    name: "hardware",
    title: "Hardware Platform",
    items: [
      { id: "b200", label: "B200 (FP4)", default: true },
      { id: "b300", label: "B300 (FP4)", default: false }, // 新加的 B300 选项
      { id: "gb300", label: "GB300 (FP4)", default: false },
      { id: "h200", label: "H200 (FP8)", default: false },
    ],
  },
  // ...
};

const generateCommand = () => {
  const { hardware: rawHardware, modelSize, recipe, reasoningParser, toolcall } = values;
  // B300 的使用方式与 B200 完全相同 — 做别名以不重复每个 spec 条目
  const hardware = rawHardware === "b300" ? "b200" : rawHardware;
  // ... 后续逻辑全部使用 hardware 变量
};

const buildPDDisaggCommand = (rawHardware, modelSize) => {
  // B300 的使用方式与 B200 完全相同 — 做别名以不重复每个 spec 条目
  const hardware = rawHardware === "b300" ? "b200" : rawHardware;
  // ...
};
```

评论区精华

本 PR 无人工 review 评论。两个自动化 bot 留言（gemini-code-assist 的配额警告和 mintlify 的预览部署通知）不涉及技术讨论。

风险与影响

风险：低。B300 完全复用 B200 配置，没有引入新逻辑分支，不会影响现有 B200/GB300/H200 的部署命令生成。文档表格增加一行，不影响其他行。唯一潜在差异在于 B300 的 GPU 特性（如 SMS 数量、内存带宽）可能与 B200 不同，但当前作为初步支持是合理的。

影响：对用户而言，B300 用户现在能获得正确的部署命令和容器镜像；对系统和团队无负面影响。影响范围限于文档和交互式代码片段。

关联脉络

该 PR 是近期 DeepSeek-V4 部署文档完善系列的一部分。此前多个 PR 已标记了 B200（#23689）、GB300（#23690、#23691）等硬件上的各方案验证状态。本 PR 将 B300 纳入

同一部署矩阵，确保覆盖率完整。随着更多硬件验证完成，未来可能需将 B300 与 B200 的配置分离（如环境变量差异），但在当前阶段统一的映射策略是高效且低风险的选择。