Fabric

性能白皮书

文档版本 01

发布日期 2024-12-31

华为云计算技术有限公司

商标声明


和其他华为商标均为华为技术有限公司的商标。
本文档提及的其他所有商标或注册商标，由各自的所有人拥有。

注意

您购买的产品、服务或特性等应受华为云计算技术有限公司商业合同和条款的约束，本文档中描述的全部或部分产品、服务或特性可能不在您的购买或使用范围之内。除非合同另有约定，华为云计算技术有限公司对本文档内容不做任何明示或暗示的声明或保证。

由于产品版本升级或其他原因，本文档内容会不定期进行更新。除非另有约定，本文档仅作为使用指导，本文档中的所有陈述、信息和建议不构成任何明示或暗示的担保。

华为云计算技术有限公司

地址： 贵州省贵安新区黔中大道交兴功路华为云数据中心 邮编：550029

网址： https://www.huaweicloud.com/

目 录

1 推理性能白皮书

本文提供Fabric推理服务使用性能测试平台进行性能测试的方法和测试数据报告。

测试环境

- 局点：华为云-Fabric测试环境。
- 测试时间：2024年11月30日。
- 推理服务及其相关资源列表：

推理服务名称	使用模型类型	资源规格	算力数量（MU）	实例数量
LLama-3-8B	LLAMA_3_8B	mu.llama3.8b	2	1
LLama-3-70B	LLAMA_3_70B	mu.llama3.70b	8	1
LLama-3.1-8B	LLAMA_3.1_8B	mu.llama3.1.8b	2	1
LLama-3.1-70B	LLAMA_3.1_70B	mu.llama3.1.70b	8	1
QWEN-2-72B	QWEN_2_72B	mu.qwen2.72b	8	1
GLM-4-9B	GLM_4_9B	mu.glm4.9b	2	1

测试工具

本文使用JMeter进行测试。JMeter是一款用于测试性能的开源软件，它可以模拟多种协议的服务器和客户端，例如HTTP、FTP 、SMTP等。它可以用来测试Web应用程序，也可以用于测试数据库连接、 FTP服务器等。 JMeter还支持自定义和预定义的脚本，可以模拟不同的负载，并支持分布式测试。

JMeter依赖于JDK，所以必须确保当前计算机上已经安装了JDK，并且配置了环境变量。您可以从**Apache JMeter官网**下载。

测试方法

1. 登录Fabric控制台，在目标工作空间区域，单击“进入工作空间”。
   您也可以单击“创建工作空间”进行创建。
2. 在左侧菜单栏选择“资源与资产 > 模型”，然后在页面右上角单击“创建模型”，填写模型的基本信息，包括名称、描述等，并选择模型文件的OBS路径，然后单击“立即创建”。
3. 在左侧菜单栏选择“资源与资产 > 推理端点”，然后在页面右上角单击“创建推理端点”，填写端点的名称、使用的资源规格和数量，单击“立即创建”。
4. 在左侧菜单栏选择“开发与生产 > 推理服务”，然后在页面右上角单击“创建推理服务”，填写推理服务的名称、描述等基本信息，并选择推理端点和模型，配置资源最小值和最大值，单击“立即创建”。
   “模型”支持选择“我的模型”或者“公共模型”。
5. 在左侧菜单栏选择“开发与生产 > 试验场”，选择目标推理服务进行推理调试。
6. 使用测试工具并发推理。

测试指标

RPM（Request Per Minute）是指每分钟请求数，是衡量系统性能的一个重要指标。它表示在一分钟内，系统能够处理的请求数量。RPM是衡量模型处理能力的一个关键指标，反映了模型在给定时间内能够处理的请求数量。

测试数据

- **输入数据1：**
  该数据是比较简短的一个问题，且回答的max_tokens是256。

```
{
 "type":"ChatCompletionRequest",
 "messages": [
  {
   "role": "user",
   "content": "What is LLM? What is different between different LLM?"
  }
 ],
 "max_tokens": 256,
 "stream":true
}
```

- **输入数据2：**
  该数据是中等长度的问题，且回答的max_tokens是2048。

```
{
 "type":"ChatCompletionRequest",
 "messages": [
  {
   "role": "user",
    "content": "Please write a novel and the word size should more than 2000,
requirements:1.Setting: Village, ancient forest, bustling city, forgotten island, futuristic metropolis, enchanted castle. 2.Protagonist: Orphaned child, disgraced knight, brilliant scientist, secret agent, reclusive artist, adventurous explorer.3.Antagonist: Shadowy figure, corrupt politician, malevolent sorcerer, rival adventurer, robotic overlord, vengeful ghost.4.Conflict: Quest for revenge, search for a lost artifact, battle for power, love triangle, struggle against fate, resistance against tyranny.2000-Word Requirement Guideline: Writing a 2000-word novel can be challenging, but it's also a great way to hone your writing skills and tell a concise, compelling story. Here are some tips to help you meet the word count while maintaining quality:1.Outline Your Story: Before you start writing, take some time to outline your story. Decide on your main plot points, character arcs, and the overall theme you want to explore. This will help you stay focused and ensure that your story has a clear structure.2.Focus on Key Scenes: With a limited word count, you need to prioritize the most important scenes. Focus on the scenes that drive the plot forward, reveal character development, and create
```



```
tension. Avoid unnecessary descriptions and subplots that don't contribute to the overall story.3.Show, Don't Tell: Use vivid, sensory details to bring your story to life. Instead of telling readers what's happening, show them through dialogue, actions, and internal monologue. This will make your writing more engaging and help you use your words more effectively.4.Edit Ruthlessly: As you write, be prepared to cut out anything that doesn't add value to your story. This might include redundant descriptions, unnecessary characters, or scenes that don't move the plot forward. Remember, every word should count."
    }
  ],
  "max_tokens": 2048,
  "stream":true
}
```

测试结果

- 基于**输入数据1**进行测试，并发度为64，测试结果如下：

  表 1-1 输入数据 1 的测试结果

模型名称	测试类型	数量	max token	测试时间（s）	成功率	状态码	总请求数	平均时延（ms）	TP99时延（ms）	每秒事务数	RPM
LLama-3-8B	并发	64	256	300	100%	200	2090	9231	32615	7.01	420.6
LLama-3-70B	并发	64	256	300	100%	200	420	43072	68082	1.79	107.4
LLama-3.1-8B	并发	64	256	300	100%	200	960	20453	51011	3.27	196.2
LLama-3.1-70B	并发	64	256	300	100%	200	679	29975	44826	2.29	137.4
QWEN-2-72B	并发	64	256	300	100%	200	8706	2212	4915	29.02	1741.2
GLM-4-9B	并发	64	256	300	100%	200	578	35655	66167	1.93	115.8

- 基于**输入数据2**进行测试，并发度为16，测试结果如下：

  说明

  推理请求的响应时长根据请求的输入Token、输出Token和参数的变化而变化，下表的数值仅供参考，实际可能差异比较大。

表 1-2 输入数据 2 的测试结果

模型名称	测试类型	数量	max token	测试时间（s）	成功率	状态码	总请求数	平均时延（ms）	TP99时延（ms）	每秒事务数	RPM
LLama-3-8B	并发	16	2048	300	100%	200	96	51636	96797	0.32	19.2
LLama-3-70B	并发	16	2048	300	100%	200	82	64296	74727	0.27	16.2
LLama-3.1-8B	并发	16	2048	300	100%	200	192	26072	38645	0.68	40.8
LLama-3.1-70B	并发	16	2048	300	100%	200	64	85552	103198	0.22	13.2
QWEN-2-72B	并发	16	2048	600	100%	200	197	51260	75031	0.33	19.8
GLM-4-9B	并发	16	2048	300	100%	200	137	37630	52302	0.46	27.6