# 总结表

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| **型号** | **显存** | **算力FP16（TFLOPS）** | **支持虚拟化** | **支持并发** | **GLM推理** | **GLM轻量微调** | **GLM 全量微调** |
| 4060ti | 16G | 22 | 否 | 1～2 | 1 x 实例 | 不支持 | 需要 x 20卡 |
| 4090 | 24G | 164 | 否 | 8 | 1 x 实例 | 支持 | 需要 x 14卡 |
| 3090 | 24G | 164 | 否 | 8 | 1 x 实例 | 支持 | 需要 x 14卡 |
| A6000 | 48G | 77 | 是 | 2～3 | 3 x 实例 | 支持 | 需要 x 7卡 |
| A100 | 40G | 165 | 是 | 8 | 2 x 实例 | 支持 | 需要 x 8卡 |
| A100 | 80G | 312 | 是 | 15 | 5 x 实例 | 支持 | 需要 x 4卡 |
| H100 | 80G | 1513,1979,3979（不同接口） | 是 | 75～200 | 5 x 实例 | 支持 | 需要 x 4卡 |
| 昇腾910 | 32G | 313 | 是 | 15 | 2 x 实例 | 支持 | 需要 x 10卡 |
| 昇腾910B | 64G | 370 | 是 | 18 | 4 x 实例 | 支持 | 需要 x 5卡 |

## 并发性能计算过程 （纯理论计算）

根据智谱提供的信息，8卡昇腾910 能提供126路并发。  
得出 => 单卡并发为15个

由于昇腾910 的算力为313 FP16，能提供15个并发。  
得出 => GLM3-6B的单并发需求为【20.8 FP16】

综上结论：  
由于4060ti 的FP16 为22 FLOPS，刚好满足单并发任务的需求。

## 实例计算过程

GLM3-6B以 FP16 精度加载，运行上述代码需要大概 13GB 显存，如果是GLM3-6B-32K版本则需要14G显存。 以上仅为加载到GPU最低显存需求，实际推理过程中，显存会出现上涨浮动。在使用过程中，发现GLM3-6B对显存的使用达到过17.3G显存  
可以理解为15G～16G显存为正常运行推理的最低配置需求。

## 能否进行微调计算过程

【轻量微调】  
粗略按照推理内存消耗的两倍计算

【全量微调】  
根据GLM2-6B官方培训资料，微调需要 A100 x 4 卡，约 320G 显存为参考标准。  
以下为官方资料截图：

# AI算力数据

入门级AI卡，RTX 4060 ti 的【单精度浮点算力】大约为22.06 TFLOPS  
平民顶配AI卡，RTX 4090 ti 的【单精度浮点算力】大约为82.06 TFLOPS  
4060ti 的算力约为 4090的 1/4 左右。  
A6000显卡的【单精度浮点算力】大约为38.7 TFLOPS

# 参考资料：

真实性能如何？RTX 4060 Ti 测试报告  
https://zhuanlan.zhihu.com/p/631651468

2023年最新最全的显卡深度学习AI算法算力排行（包括单精度FP32和半精度FP16的对比）：  
https://zhuanlan.zhihu.com/p/665120615?utm\_id=0

如何评价华为 8.23 正式推出 AI 处理器昇腾 910 和全场景 AI 计算框架？  
https://www.zhihu.com/question/342327559/answer/3261672301

GPU A100 性能测试报告：  
https://zhuanlan.zhihu.com/p/645052868?utm\_id=0

GLM3官方github仓库  
https://github.com/THUDM/ChatGLM3

4060Ti-16G、4070Ti、4090显卡的深度学习性能测试和结论  
https://www.bilibili.com/read/cv22000735/

6\*RTX4090+静音---当下最强深度学习工作站/集群硬件配置  
https://www.bilibili.com/read/cv22718070/