大模型训练需要大量算力怎么办？H100提供超强算力，北京H100算力租赁价格，猿界算力GPU服务器租赁

产品名称	大模型训练需要大量算力怎么办？H100提供超强算力，北京H100算力租赁价格，猿界算力GPU服务器租赁
公司名称	北京猿界云计算科技有限公司
价格	150000.00/件
规格参数	型号:英伟达 型号:H100 北京:猿界算力租赁服务提供商
公司地址	北京市大兴区宏业东路2号院1号楼3层0111（注册地址）
联系电话	18621019618 18621019618

产品详情

如果大模型训练需要大量算力，算力租赁是一种常用的解决方案。以下是选择合适的算力租赁解决方案的途径：

1. **选择高性能的算力租赁公司**：寻找那些具备高性能硬件设施和强大计算能力的算力租赁公司。这样可以确保能够满足大模型训练的需求。

2. **选择适合的GPU型号**：了解租赁公司提供的各种GPU型号，选择适合大模型训练的GPU型号。通常，高性能的GPU可以加速大模型训练的速度。

3. **定制化算力租赁服务**：某些算力租赁公司提供定制化服务，可以根据大模型训练的具体需求，提供更高的级别算力和专业支持，以满足特定要求。

4. **考虑租赁期限和费用**：与算力租赁公司协商租赁期限和费用，确保根据大模型训练的预计时间和预算进行合理安排。

5. **评估服务质量和技术支持**：了解租赁公司的服务质量和技术支持，确保能够及时解决问题和提供必要的支持，以避免训练过程中的延误和困扰。

综合以上考虑因素，可以选择最适合大模型训练的算力租赁解决方案。确保算力租赁的性能和服务能够满足大模型训练的需求，并在租赁过程中与租赁公司保持沟通和合作，以确保训练的顺利进行。

H100核心采用的其实是台积电目前最先进的4nm工艺，而且是定制版，CoWoS 2.5D晶圆级封装，单芯片设计，集成多达800亿个晶体管，号称世界上最先进的芯片。

完整版有8组GPC(图形处理器集群)、72组TPC(纹理处理器集群)、144组SM(流式多处理器单元)，而每组SM有128个FP32 CUDA核心，总计18432个。

显存支持六颗HBM3或者HBM2e，控制器是12组512-bit，总计位宽6144-bit。

Tensor张量核心来到第四代，共有576个，另有60MB二级缓存。

扩展互连支持PCIe 5.0、NVLink第四代，后者带宽提升至900GB/s，七倍于PCIe 5.0，相比A100也多了一半。整卡对外总带宽4.9TB/s。

性能方面，FP64/FP32 60TFlops(每秒60万亿次)，FP16 2000TFlops(每秒2000万亿次)，TF32 1000TFlops(每秒1000万亿次)，都三倍于A100，FP8 4000TFlops(每秒4000万亿次)，六倍于A100。

H100计算卡采用SXM、PCIe 5.0两种形态，其中后者功耗高达700W，相比A100多了整整300W。

按惯例也不是满血，GPC虽然还是8组，但是SXM5版本只开启了62组TPC(魅族GPC屏蔽一组TPC)、128组SM，总计有15872个CUDA核心、528个Tensor核心、50MB二级缓存。

PCIe 5.0版本更是只有57组TPC，SM虽然还是128组，但是CUDA核心只有14952个，Tensor核心只有456个。

H100系统集成八颗H100芯片、搭配两颗PCIe 5.0 CPU处理器(Intel Sapphire Rapids四代可扩展至器?)，拥有总计6400亿个晶体管、640GB HBM3显存、24TB/s显存带宽。

性能方面，AI算力32PFlops(每秒3.2亿亿次)，浮点算力FP64 480TFlops(每秒480万亿次)，FP16 1.6PFlops(每秒1.6千万亿次)，FP8 3.2PFlops(每秒3.2千亿次)，分别是上代DGX A100的3倍、3倍、6倍，而且新增支持网络内计算，性能3.6TFlops。

PCIe 5.0版本的性能基本都再下降20%。

同时配备Connect TX-7网络互连芯片，台积电7nm工艺，800亿个晶体管，400G

GPUDirect吞吐量，400G加密加速，4.05亿/秒信息率。

GPU租赁市场价格波动较大，近期在北京租H100
GPU租赁价格在15万/月左右，具体看节点、配置、台数以及租期等因素都会影响价格。