本地部署开源大模型

Ch.1 如何选择合适的硬件配置

为了在本地有效部署和使用开源大模型，**深入理解硬件与软件的需求至关重要。**在硬件需求方面，关键是**配置一台或多台高性能的个人计算机系统或租用配备了先进GPU的在线服务器**，确保有足够的内存和存储空间来处理大数据和复杂模型。至于软件需求，**推荐使用Ubuntu操作系统**，因其在机器学习领域的支持和兼容性优于Windows。编程语言建议以Python为主，结合TensorFlow或PyTorch等流行机器学习框架，并利用DeepSpeed等优化工具来提升大模型的运行效率和性能。

所以在本系列课程中，我们将从硬件选择入手，逐步引导大家理解并掌握如何为大模型部署选择合适的硬件，以及如何高效地配置和运行这些模型，从零到一实现大模型的本地部署和应用。首先来看硬件方面，提前规划计算资源是必要的。目前，我们主要考虑以下两种途径：

1. **配置个人计算机或服务器**，组建一个适合大模型使用需求的计算机系统。
2. **租用在线GPU服务**，通过云计算平台获取大模型所需的计算能力。

一、大模型应用需求分析

大模型的本地部署主要应用于三个方面：**训练（train）、高效微调（fine-tune）和推理（inference）**。这些过程在算力消耗上有显著差异：

- **训练**：算力最密集，通常消耗的算力是推理过程的至少三个数量级以上。
- **微调**：微调是在预训练模型的基础上对其进行进一步调整以适应特定任务的过程，其算力需求低于训练，但高于推理。
- **推理**：推理指的是使用训练好的模型来进行预测或分析，是算力消耗最低的阶段。

总的来说，在算力消耗上，**训练 > 微调 > 推理。**

从头训练一个大模型并非易事，这不仅对个人用户，对于许多企业而言也同样困难。因此，如果个人使用，关注点应该放在**推理和微调**的性能上。在这两种应用需求下，对**硬件的核心要求体现在GPU的选择上，对CPU和内存的要求并不高。**无论是选择租用在线算力还是配置本地计算机，如果想在本地运行大模型，我们可以拆分成两个关注点：

- 模型：选择什么基座模型或微调模型，这可以直接下载至本地。
- 硬件：希望在什么硬件平台上来执行，可以分为 CPU 和 GPU 两大类。

大部分开源大模型支持在 CPU 和 Mac M系列芯片上运行，但较为繁琐且占用内存至少 32G 以上，因此更推荐在 GPU 上运行。针对本地部署大模型，**在选择GPU时，可以遵循的简单策略是：在满足具体的大模型的官方配置要求下，选择性价比最高的GPU。**

GPU的性能主要由以下三个核心参数决定：

1. **计算能力**：这是最关注的指标，尤其是32位浮点计算能力。随着技术发展，16位浮点训练也日渐普及。对于仅进行预测的任务，INT 8 量化版本也足够；
2. **显存大小**：大模型的规模和训练批量大小直接影响对显存的需求。更大的模型或更大的批量处理需要更多的显存；

3. **显存带宽**：决定了GPU处理器能够多快地从显存中读取数据和向显存中写入数据。显存带宽越高，GPU处理大量数据时的性能通常也越好；

注：显存带宽相对固定，选择空间较小。

二、硬件配置的选择标准

无论是个人使用、科研团队进行项目研究，还是企业寻求商业应用的落地，不同的应用场景和目标任务（如微调或推理）都需要相应的硬件配置方案来支持。所以**在选择硬件配置时应根据具体的模型需求和预期用途来确定。**

因此，我们的建议是：**根据部署的大模型配置需求，先选择出最合适的 GPU，然后再根据所选 GPU 的特性，进一步搭配计算机的其他组件，如CPU、内存和存储等，以确保整体系统的协调性和高效性能。最简单的匹配GPU的标准是显存大小和性价比。**因为训练不纯粹看一个显存容量大小，而是和芯片的算力高度相关的。因为实际训练的过程当中，将海量的数据切块成不同的batch size，然后送入显卡进行训练。显存大，意味着一次可以送进更大的数据块。但是芯片算力如果不足，单个数据块就需要更长的等待时间显存和算力，必须要相辅相成。

简单来说，在深度学习的训练和推理中，GPU的显存主要用于以下几个方面：

1. **权重存储**：模型的参数，包括权重和偏置，都需要在显存中存储。这些参数是模型进行预测或分类所必需的。
2. **中间过程数据存储**：在模型计算的前向传播和反向传播过程中，会产生并且需要暂时存储大量的中间计算结果。这些数据同样存储在显存中。
3. **计算过程**：GPU专门为并行处理大量的矩阵和向量运算而设计，这正是深度学习中常见的计算类型。这些计算直接在显存中进行，以利用GPU的高速运算能力。

显存的大小和速度直接影响到模型的处理速度和能处理的模型大小。显存越大，意味着可以处理更大的模型和更复杂的计算任务，但同时也需要更多的能源和可能导致更高的成本。

> "芯片"通常指的是集成电路，它们被集成到各种电脑硬件组件中，如CPU、GPU和主板等。CPU本身就是一种芯片。它是计算机的大脑，负责执行程序和处理数据。显卡上的核心组件是图形处理器（GPU），它也是一种芯片。GPU负责处理图形和视频渲染。

所谓的"算力"大小，通常指的是整个计算系统的处理能力，尽管在特定上下文中，它有时特指GPU的处理能力。

我们以ChatGLM-6B模型为例，官方给出的硬件配置说明如下：

量化等级	推理时 GPU显存占用	微调时 GPU显存占用
单精度	20G	22G
半精度	13G	14G
INT 8	8G	9G
INT 4	6G	7G

> 模型量化是一种用于优化模型的技术，特别是在推理时。它通过减少模型中使用的数值精度来减小模型的大小，加快推理速度，并降低内存和能源消耗。模型量化常用于将模型部署到资源受限的设备上，如手机或嵌入式系统，量化程度越高，对硬件的要求就会越低。 单精度通常指的是32位浮点数（FP32），使用32位表示，包括1位符号位、8位指数位和23位尾数位。FP32是标准的训练和推理格式，但由于半精度（FP16）提供了相似的结果且计算速度更快，更节省内存，因此在资源受限或需要

> 高速计算的场景中越来越受欢迎。而 int 4，就是所有的参数只保留了四位数，当然，保留的参数越少，它的计算量就会越小，对应的输出结果的精度也就会越差。

2.1 选择满足显存需求的 GPU

关于如何选择GPU，当前市场 NVIDIA 和 AMD 是两大主要显卡生产商。但在人工智能、大数据、深度学习领域，NVIDIA（通常被称为N卡）几乎独占鳌头。主要原因还是NVIDIA在很早期就开始专注于AI和深度学习市场，开发了强大的软件工具和库，例如cuDNN、TensorRT，这些都是专门为深度学习优化的，与流行的深度学习框架（如TensorFlow、PyTorch等）紧密集成，同时NVIDIA的CUDA（Compute Unified Device Architecture）作为独特的平行计算平台和编程模型，它允许开发者利用NVIDIA的GPU进行高效的通用计算。这一点对于深度学习和大数据分析等需要大量并行处理的应用来说至关重要。

- **英伟达是一家什么公司?**

这时候可能有小伙伴说了，英伟达是一家卖游戏显卡的，这个说法呢，对，但也不对，从财报来看，英伟达目前主要有四块业务，分别是游戏GPU，数据中心产品，自动驾驶芯片和其他业务。占比分别为33.6%，55.%，3.3%和7.4%。游戏GPU，数据中心产品，自动驾驶芯片其实都可以归类为计算芯片这个门类下面，换句话说，如果从财报公布的业务情况来分析的话，英伟达确实就是一家卖计算芯片的公司。但如果真的把英伟达当做一家卖芯片的公司，那就大错特错了。英伟达的确是靠着游戏显卡起家，并且在人工智能爆发的现在靠着一手AI计算芯片市值突破了万亿美元，但其实它并不是一家卖芯片的公司，我对英伟达的定位是，它是一家卖**人工智能系统**的公司。这就有两个核心的概念，一个是英伟达的计算芯片，一个是英伟达针对自家芯片做的计算架构CUDA，二者缺一不可。这种定位就像智能手机时代的苹果公司，苹果依靠着A系列芯片和ios操作系统收割了智能手机行业超过80%的利润。人工智能大发展的时代，英伟达就依靠着GPU和计算芯片与CUDA计算架构，共同组成的AI生态系统赢得了市场青睐，根据相关机构的统计数据，在独立显卡领域，英伟达的市占率高达85%，在AI算力芯片领域，在未来可能达到90%，现在做深度学习，英伟达的卡就是刚需，没有其他的选择。

因此，我们建议还是选择 NVIDIA 的显卡。如果对应的ChatGLM-6B模型的硬件配置说明，我们就可以这样选择GPU。理论上，**在进行少量对话时:**

量化等级	推理时 GPU显存占用	微调时 GPU显存占用	最低显卡配置	显卡显存
单精度	20G	22G	3090	24G
半精度	13G	14G	3090	24G
INT 8	8G	9G	2080Ti	11G
INT 4	6G	7G	2060s	8G

在选择显卡时，必须遵循的首要准则是：显卡的显存容量一定要高于大模型官方要求的最低显存配置。这是确保模型能够有效运行的基本要求。显存容量越大，其推理或微调的能力就会越强。当然，随着显存容量的增加，显卡的价格也相应提高。以下是目前最主流的几款大模型的显卡型号及其显存容量：

显卡型号	显存容量
H100	80 GB
A100	80/40 GB
H800	80 GB
A800	80 GB

显卡型号	显存容量
4090	24 GB
3090	24 GB

其组合形式可以分为以下四类：

1. 纯CPU：基于不同架构的CPU配置，适用于不需要或不能使用GPU加速的场景。 **（不推荐）**
    - x86 (如Intel或AMD)
    - ARM (如Apple、Qualcomm、MTK)

2. 单机单卡：使用一块GPU进行计算，适用于大多数个人使用和一些中等计算负载的场景。 **（典型配置）**
    - Nvidia系列GPU
    - AMD系列GPU
    - Apple系列GPU
    - Apple Neural Engine（较少见，支持有限）

3. 单机多卡：在一台机器上使用多张GPU卡，适用于高计算负载的场景，如模型分割处理。 **（典型配置）**
4. 多机配置：使用多台计算机进行集群计算，通常超出个人使用范围，主要用于从头预训练基座模型等高负载任务。

所以，在单个显卡的显存容量不足以满足需求时，也可以采用多显卡配置来增加整体的显存容量。只要总显存超过官方推荐的配置要求就可以。此外，在选择显卡时，除了考虑整体显存容量，还要根据不同显卡的性能和成本进行权衡。根据具体需求和预算，决定是选择单张高性能显卡，还是部署多张成本效益更高的低版本显卡。实现最优的性价比。比如在理论上，在进行多轮对话时或需要微调时，采用单机多卡：

量化等级	推理时 GPU显存占用	微调时 GPU显存占用	最低显卡配置	显卡显存
单精度	30G	22G	3090双卡	48G
半精度	20G	14G	3090	24G
INT 8	12G	9G	3080Ti	12G
INT 4	10G	7G	2080Ti	11G

2.2 主流显卡性能分析

对于 NVIDIA的显卡（N卡）卡来说，我们可以按照以下几个维度来划分：

按照产品线划分：

系列	特点	主要应用领域
GeForce系列（G系列）	消费级GPU产品线，注重提供高性能的图形处理能力和游戏特性，性价比高，适合游戏和深度学习推理、训练。	主要面向游戏玩家和普通用户。

系列	特点	主要应用领域
Quadro系列（P系列）	专业级GPU产品线，针对商业和专业应用领域进行了优化，适用于设计、建筑等专业图像处理。	设计、建筑、专业图像处理等领域。
Tesla系列（T系列）	主要用于高性能计算和机器学习任务，集成深度学习加速器，提供快速的矩阵运算和神经网络推理。	高性能计算、机器学习任务等领域。
Tegra系列	移动处理器产品线，用于嵌入式系统、智能手机、平板电脑、汽车电子等领域，具备高性能的图形和计算能力，低功耗。	嵌入式系统、智能手机、平板电脑、汽车电子等领域。
Jetson系列	面向边缘计算和人工智能应用的嵌入式开发平台，具备强大的计算和推理能力，适用于智能摄像头、机器人、自动驾驶系统等。	边缘计算、人工智能、机器人等领域。
DGX系列	面向深度学习和人工智能研究的高性能计算服务器，集成多个GPU和专用硬件，支持大规模深度学习模型的训练和推理。	深度学习、人工智能研究和开发等领域。

按照架构划分：

架构	年份	芯片代号	特点	代表产品
Tesla	2006	GT	第一个通用并行计算架构，主要用于科学计算和高性能计算。	Tesla C870/GeForce 8800 GTX
Fermi	2010	GF	引入CUDA架构、ECC内存等，用于科学计算、图形处理和高性能计算。	Tesla C2050/GeForce GTX 480
Kepler	2012	GK	功耗效率和性能改进，引入GPU Boost技术，适用于科学计算、深度学习和游戏。	Tesla K40/GeForce GTX 680
Maxwell	2014	GM	提高功耗效率，引入新技术如多层次内存系统，应用于游戏、深度学习和移动设备。	Tesla M40/GeForce GTX 980
Pascal	2016	GP	16nm FinFET制程技术，加强深度学习和AI计算支持，引入Tensor Cores，应用于深度学习和高性能计算。	Tesla P100/GeForce GTX 1080

架构	年份	芯片代号	特点	代表产品
Volta	2017	GV	深度学习优化特性，如Tensor Cores，主要用于深度学习、科学计算和高性能计算。	Tesla V100
Turing	2018	TU	实时光线追踪技术、深度学习技术，适用于游戏、深度学习和专业可视化。	Tesla T4/GeForce RTX 2080 Ti
Ampere	2020	GA	第二代深度学习架构，更多Tensor Cores、改进的Ray Tracing技术，应用于深度学习、科学计算和高性能计算。	Telsa A100/GeForce RTX 3090
Ada Lovelace	2022	AD	专为光线追踪和基于AI的神经图形设计，第四代Tensor Core，第三代RT Core，提高GPU性能。	GeForce RTX 4090
Hopper	2022	GH	下一代加速计算平台，支持PCIe 5.0，专用的Transformer引擎，适用于大型语言模型和对话AI，提供企业级AI支持。	Telsa H100

按照应用领域划分：

类型	系列	描述	应用领域	代表产品
游戏娱乐	GeForce RTX™系列	面向大众消费级游戏和创作者用户的图形加速卡。在性能、功耗和成本之间达到最佳平衡点,提供极致的游戏和创作体验。	游戏、娱乐、内容创作	如RTX 3090、RTX 4090等
专业设计和虚拟化	NVIDIA RTX™系列	面向专业可视化和创意工作负载的高性能GPU,提供强大的计算性能、大容量视频内存等。服务于工业设计、建筑设计、影视特效渲染等专业用户。	工业设计、建筑设计、影视特效渲染	高端专业可视化工作站级显卡
深度学习、人工智能和高性能计算	A系列、H系列、L系列、V系列、T系列	不同系列针对不同AI和计算需求。A系列是AI计算加速器; H系列是AI超算; L系列是边缘AI推理; V系列是虚拟工作站; T系列是AI推理优化解决方案。	数据中心AI训练和推理、边缘AI、虚拟桌面、AI推理加速	A100、A30、A40、H100、L40、DeepStream加速器等

像大模型领域这种生成式人工智能，需要强大的算力来生成文本、图像、视频等内容。在这个背景下，NVIDIA先后推出V100、A100和H100等多款用于AI训练的芯片，其中 A100 是 H100 的上一代产品，于2020年发布，使用7纳米工艺，支持AI推理和训练。而H100，该显卡是2022年3月发布，可谓是核弹级性能显卡，采用了台机电4纳米工艺，具备800亿个晶体管，采用最新 Neda Hopper架构，同时显存还支持hbm3，最高带宽可达 3TB每秒。第四代MNLINK的带宽，900G每秒。是PCIE5.0的7倍，比上一代的A100显

卡高一倍，显卡对外总带宽达到超高的 4.9TB每秒。性能上H100显卡相对于上一代A100来说，可谓是质的飞跃，各项基础性能是A100的三倍之多，H100的单片显卡售价24万元左右。

但在2022年10月，漂亮国政府为了限制我国的人工智能发展，发布禁令：禁止NVIDIA向中国出售A100和H100显卡。数据显示，2022年中国市场的人工智能芯片规模高达70亿美元，而这70亿的市场，被NVIDIA垄断了90%，虽然NVIDIA的A100，H100这样的顶级芯片不能卖给中国，但NVIDIA作为商业公司，也是要做生意的，于是为了合规，NVIDIA针对传输速率进行了限制，提出了中国大陆特供版的A800和H800，即：H100、A100的阉割版。

性能参数	V100 PCIe	A100 80GB PCIe	A800 80GB PCIe	H100 80GB PCIe
微架构	Volta	Ampere		Hopper
FP64	7TFLOPS	9.7TFLOPS		26 TFLOPS
FP32	14TFLOPS	19.5TFLOPS		51 TFLOPS
FP16 Tensor Core		312TFLOPS		756.5 TFLOPS
INT8 Tensor Core	62 TOPS	624 TOPS		1513 TOPS
GPU 显存	32/16GB HBM2	80GB HBM2e		80GB
GPU 显存带宽	900 GB/s	1935GB/s		2TB/s
最大热设计功耗 (TDP)	250 瓦	300 瓦		300-350W
多实例 GPU		最多 7 个 MIG 每个 10GB		
外形规格		PCIe 双插槽风冷式 或单插槽液冷式		PCIe 双插槽风冷式
互连技术	NVLink：300 GB/s PCIE：32 GB/s	搭载 2 个 GPU 的 NVIDIA" NVLink" 桥接器：600GB/s PCIe 4.0：64GB/s	搭载 2 个 GPU 的 NVIDIA" NVLink" 桥 接器：400GB/s PCIe 4.0：64GB/s	NVLink：600GB/s PCIe 5.0：128GB/s
服务器选项		搭载 1 至 8 个 GPU 的合作伙伴认证系统和 NVIDIA 认证系统		

也就是说，由于漂亮国的禁令，我们现在使用的GPU都是中国特供版的，说白了就是阉割版的，像A100，到国内就成了A800，H100到国内就成了H800，那么 A ~ H的差距在哪里呢?

规格	A100 SXM	A800 SXM	H100 SXM	H800 SXM
双精FP64	9.7 TFLOPS		34 TFLOPS	1TFLOPS
双精FP64 Tensor Flow	19.5 TFLOPS		67 TFLOPS	1TFLOPS
单精FP32	19.5 TFLOPS		67 TFLOPS	
单精TF32	156 TFLOPS | 312 TFLOPS*		495 TFLOPS | 989 TFLOPS*	
半精FP16	312 TFLOPS | 624 TFLOPS*		990 TFLOPS | 1,979 TFLOPS*	
FP8	NA		3,958 TFLOPS	
显存大小	80G		80G	
显存带宽	2039 GB/s		3.35TB/s	
互联带宽	600 GB/s	400GB/s	900GB/s	400GB/s
NVLINK链路	12条	8条	18条	8条
功耗	400W	400W	700W（最高）	

直接用 SXM 版本的H800进行对比，只能说这个参数对比，对于不了解的人来说，还是比较出人意料的，除了 FP64 和 NVLink传输速率上的明显削弱，其他参数和H100都是一模一样的。FP64上的削弱主要影响是H800在科学计算、流体计算、有限元分析等超算领域的应用，受到影响最大的还是NVLINK上的削减，但因为架构上的升级，虽然比不上同为 Hoper架构的H100，但是比AMPERRE架构的A100还是要强上不少，说白了，老黄想要抓住国内市场，就算是阉割，也不会阉割的特别过分，漂亮国政府想限制国内的超算，那就算把超算性能砍掉，传输速率减小，换个名字，GPU照卖。只要保证H800在大部分场景下的性能不受影响，能满足大部分人的使用需求就足够了。毕竟也不会有人跟钱过不去，所以 其实H800和 H 100的性能差距并没有想象的那么夸张，就算是砍掉了FP64和 NVLINK的传输速率，性能依旧够用。最关键的是，它合法呀。所以如果不是追求极致性能的话，也没必要冒着风险去选择H100。

而就在今年的10月份，漂亮国又玩起了变卦，10月份刚升级了芯片禁令，开启了新一轮的出口管制，先预留了30天的窗口期，随后又要求立即生效，连30天都没了，也就是说，从10月份开始，中国将无法再获得NVIDIA5类的GPU显卡 （A800、A800、H100、A100，L40S），其实早在8月份的时候，BAT的一些大厂不知道是收到风声还是控制风险，就向NVIDIA提前订购了10万个A800芯片，结果这次也是彻底泡汤。其实从19年开始，漂亮国对芯片的打压手段就步步升级，现在已经有点丧心病狂了，反正不管合不合理，只要跟芯

片大搭点边，全都想禁。主打的就是一个全面限制，围追堵截。而现在，中国现在是无法在获得任何NVIDIA的尖端AI芯片了，漂亮国就是亮牌，高端AI芯片，必禁无疑。所以对于目前的 A100系列和H100系列，因为是漂亮国断供之前出的芯片，所以现在国内还有货，只不过市场渠道比较乱，需要甄别。

同时需要说明的是，GeForce 系列显卡虽被官方定位为面向消费级市场，适合游戏爱好者。但这类显卡在深度学习领域同样展现出了出色的性能，很多人用来做推理、训练，单张卡的性能跟深度学习专业卡Tesla系列比起来其实差不太多，但是性价比却高很多。对于大模型来说，同样可以使用GeForce 系列显卡。

那么个人使用或者实验室针对大模型的推理和微调需求配置服务器，高端显卡目前我们可选的就是A100、A800、H100和4090等，应该如何选呢？

2.3 单卡4090 vs A100系列

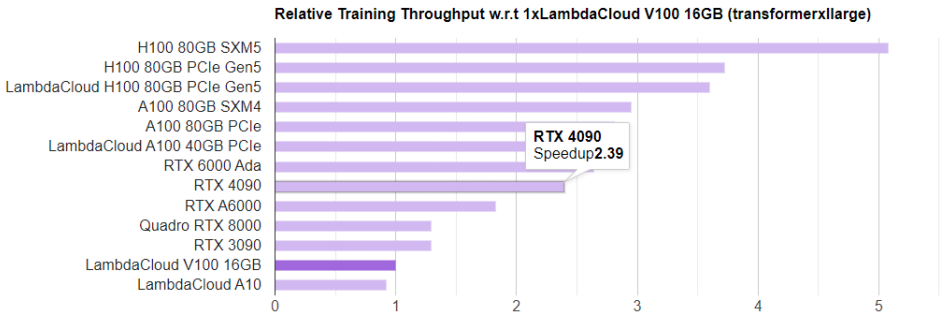
先说结论：**没有双精度需求，追求性价比，选4090。有双精度需求，选A100，没有A100选A800。如果是做大模型的训练，GeForce RTX 4090 是不行的。但在执行推理（inference/serving）任务时，使用 RTX 4090 不仅可行，而且在性价比方面甚至略优于 A100。同时如果做微调，也勉强是可以的，但建议多卡。**

GPU 型号	Tensor FP16 算力	Tensor FP32 算力	内存容量	内存带宽	通信带宽	通信时延	售价（美元）
H100	989 Tflops	495 Tflops	80 GB	3.35 TB/s	900 GB/s	~1 us	30000~40000
A100	312 Tflops	156 Tflops	80 GB	2 TB/s	900 GB/s	~1 us	15000
4090	330 Tflops	83 Tflops	24 GB	1 TB/s	64 GB/s	~10 us	1600

- **推理**

从数据对比来看，A100 和 GeForce RTX 4090 显卡在通信能力和内存容量方面存在显著差异，但在算力上差距并不大。在 FP16 算力方面，两者几乎相当，4090 甚至略有优势。相较于 A100，其较高的性价比主要源于推理过程通常涉及单一模型，在这种场景下，显卡的算力才是关键因素，而 4090 在这方面表现出色。虽然内存带宽同样重要，但在推理任务中，4090 的内存带宽通常足以应对需求，不会成为显著的制约因素。

LambdaLabs 有个很好的 GPU 单机训练性能和成本对比：https://lambdalabs.com/gpu-benchmarks ， 我们来看：



可以看到，4090的速度是2.39，A100是2.29，差别不大，但价格相差10倍，所以4090的性价比是非常高的。

- **微调**

反观训练需求下，4090在训练的时候表现不佳的原因主要是其有限的通信能力和内存容量。比如训练 LLaMA-2 70B 时需要2400块 A100 ，同时据说训练ChatGPT用了上万块 A100，主要还是因为训练过程除了存储模型参数外，还需要处理大量数据以及各层之间的中间数据和参数。因此，大容量内存和高通信带宽会比较关键，以便高效地处理和协调这些信息。首先就是把n个T的数据，分发到不同的GPU上去，然后训练，这叫数据并行。第二个并行就是会把这个模型的数据在一块GPU里可能放不下，所以要按照每一层，把某几层放在不同的GPU上面，进行串联。这就叫流水线并行。第三个就是Tensor张量并行。主要是我们目前训练的Transform模型都是多头的，每一个头都是可以按照张量来进行并行训练。所以整个 LLaMA-2 70B他会通过张量，流水线、数据三种并行方式，从模型内层，到模型层之间，和训练数据三个维度进行计算空间的划分。

2400块GPU之间要进行大量的协调和通讯计算，这种复杂的并行结构需要 GPU 之间进行大量的协调和通信。4090 的通信带宽仅为 64 GB/s，与 A100 的 900 GB/s 相比差距过大，导致在这类大规模训练任务中通信成为瓶颈，进而影响整体性价比。因此，尽管 4090 在某些方面表现优秀，但在大模型训练中的局限性仍然明显。尽管微调过程对硬件的要求相较于训练相对较低，但这个过程仍然需要足够的内存以存储模型参数，以及有效的通信带宽来处理数据和模型层之间的交互。所以对于需要高通信带宽和大内存容量的大模型微调任务，A100等高端GPU可能是更合适的选择。

我们拿 GPT 3 来说，GPT 3的参数将近700亿，假设每个参数使用4字节（通常使用float 32）进行存储，训练运算储备需求是 4200 GB，完成一次GPT 3训练的总算力是：3.15 * 10 ^23 Flops,仅考虑算力的情况下，单块 A100 需要45741天，几乎是128年（假设有效算力是78Tflpos），单块4090 需要91146天，几乎是250年，（假设有效算力是40 Tflpos）。任何一张单卡训练一次都需要超过100年，对于参数量达到10亿级别的大模型，参数本身就大，而且大模型通常需要更高的显存来来存储参数、中间计算结果和梯度等，既然要多卡运行，数据的同步效率就会显得非常重要，那么内存带宽、通信带宽、通信延时等性能将非常非常重要。，4090 24g的显存小，而且内存带块、通信贷款、和通信延时都相对较弱，这就有点像短板理论。最弱的那一项就决定了显卡的能力。综上，4090在较大的大模型没有什么发挥的余地，但随着现在的大模型越来越小，对显存和算力需求相对较小的大模型，再加上推理的算力需求更低，如LLama 7B 13B模型，单卡的4090 都可以运行，至少对于学习和研究大模型的个人或实验室来说，4090还真是不错的选择。

大模型的工业级实践要求，大模型全量微调需要至少4张A100 80G显卡；（ChatGLM6B模型 全量微调差不多也是需要这个配置）

2.4 单卡4090 vs 双卡3090

如果预算差不多的情况下，对于两张3090与一张4090的选择，推荐使用两张3090显卡。虽然从算力角度看，两张3090与一张4090大致持平，但两张3090显卡提供的总显存会更多，这对于处理大型模型尤为重要。目前，大多数深度学习计算框架都支持各种并行计算技术，如流水线并行、张量并行和CPU卸载。这些技术使得即使是显存较小的显卡也能处理大型模型。在这种情况下，双3090配置可以更有效地利用流水线并行，同时，与单4090配置相比，CPU卸载的需求会降低。此外，企业环境一般都是多卡并行，使用双3090配置还可以练习如何写多卡代码。现有技术如串行反向传播，进一步增强了多卡系统的效率，使其成为一个经济高效的选择，尤其是在需要处理大量数据和复杂模型的情况下。因此，从目前的技术和应用需求来看，选择两张3090显卡无疑是更优的选择。

2.5 风扇卡与涡轮卡如何选择

一定要买涡轮卡。

- 供电接口位置与散热方向
    - 风扇卡与涡轮卡的供电接口位置不同，涡轮卡的供电接口位置在接口尾部，供电线比风扇卡的线更短，这样是方便安装和理线，而风扇卡供电接口一般在显卡顶部，接线后线缆会高于机箱最高面，在服务器中使用风扇卡，服务器盖板盖不上。
- 在散热方向上面，涡轮卡散热方向是朝尾部散热，并于服务器风向是一致的，而风扇卡的散热是朝四面八方来散热的，平常的PC机箱放一张是可以适应的，但用作服务器上（很多时候是多卡）就不适合了，很容易因为温度过热出现宕机。
- 风扇卡与涡轮卡的尺寸大小不同
    - 涡轮卡与风扇卡的尺寸大小也是不一样的，风扇卡的尺寸一般是2.5-3倍宽设计，而涡轮卡的尺寸大小是双宽设计，因为涡轮卡为了方便放入服务器里，所以涡轮卡的尺寸和高度都远远低于风扇卡，从而服务器可以支持4卡或者8卡，如果用风扇卡代替涡轮卡装在服务器里，那位置够不够还是一回事儿呢。
- 面对市场不同
    - 风扇卡无论是公版显卡还是非公版显卡，风扇卡都是面向个人的，是应用在个人游戏行业的，4090风扇卡的特点就是外观炫酷，而个人游戏行业就是为了风扇卡的外观和玩游戏的性能。而4090涡轮卡是定制版，是面向AI科技产业，因为做工精巧、支持多卡安装、性价比高等一系列优点，4090涡轮卡深受广大AI深度学习用户的喜爱。

2.6 整机参考配置

确定GPU后，根据GPU搭配合适的计算机组件，具体来说，计算机八大件：CPU、散热器、主板、内存、硬盘、显卡、电源、风扇、机箱。个人使用的计算机，典型的配置是单GPU或双GPU，一般不超过四个GPU，否则常规的机箱放不下，且运行时噪声很大，而且容易跳闸。

目前国内实验室主流的还是4090和3090,10万+的预算配置4张4090是没问题的，20~30万的预算则可以考虑8张4090，或者两张A100 80G，如果预算不限，A100 8卡服务器一定是最佳选择。

这里给出一个本地部署ChatGLM-6B，同时也适用于大多数消费级实验环境的配置：

- GPU：3090双卡，涡轮版；总共48G显存，能够适用于大多数试验和复现性质深度学习任务；同时双卡也便于模拟多卡运行的工业级环境；
- CPU：AMD 5900X；12核24线程，模拟普通服务器多线程设置；
- 存储：64G内存+2T SSD数据盘；内存主要考虑机器学习任务需求；
- 电源：1600W单电源；双卡GPU的电源在1200W-1600W均可；
- 主板：华硕ROG X570-E；服务器级PCE，支持双卡PCIE；
- 机箱：ROG太阳神601；atx全塔式大机箱，便于高功耗下散热；

A800 工作站的典型配置信

配置项	规格
CPU	Intel 8358P 2.6G 11.2UFI 48M 32C 240W *2
内存	DDR4 3200 64G *32
数据盘	960G 2.5 SATA 6Gb R SSD *2

配置项	规格
硬盘	3.84T 2.5-E4x4R SSD *2
网络	双口10G光纤网卡（含模块）*1
	双口25G SFP28无模块光纤网卡（MCX512A-ADAT ）*1
GPU	HV HGX A800 8-GPU 8OGB *1
电源	3500W电源模块*4
其他	25G SFP28多模光模块 *2
	单端口200G HDR HCA卡(型号:MCX653105A-HDAT) *4
	2GB SAS 12Gb 8口 RAID卡 *1
	16A电源线缆国标1.8m *4
	托轨 *1
	主板预留PCIE4.0x16接口 *4
	支持2个M.2 *1
原厂质保	3年 *1

总的来说：

- 3090比4090综合性价比更高，不过4090计算速度几乎是3090的两倍，有需求亦可考虑升级，不过4090需要的机箱空间更大、电源配置也要求更高；
- 双卡GPU升级路线：3090—>4090—>A100 40G （2.5w左右）—>A100 80G（6~7w左右）；
- 大模型的工业级实践要求，大模型全量微调需要至少4张A100 80G显卡；（ChatGLM6B模型 全量微调差不多也是需要这个配置）

2.7 显卡博弈的形式分析

除此之外，在2023年11月13日老黄又放出来两个核弹。第一个核弹就是它推出了全新的超算GPU H200，直接说是当世最强，听起来很嚣张，但其实一点没有吹牛。在AI超算领域，对手只有看NVIDIA车尾灯的份。从数据层面看，H200强在大模型推理上，以700亿参数的Llama2 二代大模型为例，h200推理速度几乎比前代的h100快了一倍。而且能耗还降低了一半。显存从h100的80GB，直接拉到了141gb，带宽也从3.35TB/s，提升到了4.8TB/s，最新的GPU H200，跟前一代H100相比，最大的提升就是它的内存，达到了惊人的1.15TB/s，相当于在1s内传输了 230步FHD的高清电影。如果每一部的容量按5G来算的话。这个跟我们以前的计算机里的内存条就不一样了，它采用的最新技术是HBM3e，HBM就是高带宽内存，这个实现是把DRAM内存用3D封装的技术叠了起来，然后把它和GPU芯片放在同一个GPU的底板上，它们之间的通信就通过这块晶元直接做了，这样内存的容量就大大提高了，同时他们和GPU的通信速度也有显著的增长，。达到了每秒钟4.8个TB。然后又把所有的软件做了优化，这样就使得ChatGPT这样 大模型的推理速度大大的提升，跟A100相比提高了 18倍。第二个核弹就是CPU和GPU的合体，GH200， 就是把ARM的CPU和它的GPU封装在了同一块GPU晶圆板上，这样CPU和GPU之间的传输速度就非常快，而且可以共享内存。内存也达到

了惊人的624GB。而且上面有72个ARM的计算核。这个算力跟X86的芯片相比，提高了100倍。而功耗却只有1/2。

炸一听好像是王炸升级，刚装满h100的企业要哭晕在厕所了。但实际上，它可能只是h100的一个中期改款，单论峰值算力，H100和H200其实是一模一样的。，真正提升的是显存和带宽，然而对于AI芯片的性能，讨论最多的是训练能力。，在GPT 3 175B大模型的训练中，H200相较于H100，只强了10%，提升并不明显，这操作，大概率是老黄有意为之，以前为了打造大模型，对GPU的首要要求是训练，但是到了现在，随着各种AI大预言模型的落地，大家开始卷的是推理速度。于是H200的升级，就忽略了算力升级，转向推理方面的发力，老黄的刀法依旧精准。哪怕只是小提升，依然当得起最强的称号。谁让NVIDIA的显卡，在AI芯片这块，这就是遥遥领先。

但这因为是断供后的新卡，国内现在基本买不到。

在H200没出现以前，H100是地表最强GPU，NVIDIA每一个层级的性能基本都是翻倍的，H100，其中微软采购了 15w片，mate 采购了 15w片，谷歌、亚马逊、甲骨文、腾讯都是5w片，那么谷歌的gemini发布晚，原来是因为缺少GPU哈。一共是 48w片，和外界传的 一年H100的产量50w基本吻合。在2024年预计出货量在200万张。中国采购的用户 H800要比H100量大，而且H800的售价比H100还要高，为什么性能不行价格还是高呢，主要原因还是有一份建议国企采购产品中的文件，里面只有A800和H800，没有A100和H100，这就导致国企采购更愿意采购A800的原因，

同时需要说的是，在今年的10月份，漂亮国再次禁用 H800、A800芯片后，NVIDIA计算再次推出中国特供AI芯片，初步计划是3款，分别是h20，L20和 L2，这三款基于H100进行阉割。使以性能符合禁令的要求。其中最强的是H20，但与H100相比，性能被封印了80%，只有H100的20%左右的性能，对于NVIDIA而言，中国这笔70亿美元的大市场肯定不能丢，必须推出AI芯片来抢占，不过，近日有消息传出，这三款特供版芯片要跳票了，只有L20可能会按期推出，H20和L2都可能延期。特别是 H20这个最强的，什么时候推出，会不会推出都是未知数，最重要的原因，还是目前中国的市场已经发生了变化。没有NVIDIA想像的那么美好了。

有一些朋友可能还不知道这回事，而知道的朋友有些也不清楚，为什么偏偏是显卡会成为国际博弈的工具，这些卡一张卖 十几 甚至几十万元，这么赚钱的生意，怎么就不让做了。在1999年之前的人类文明早期，世界上是没有显卡的，但已经有了电子游戏了，那时候的游戏画面是由CPU生成的，游戏玩家说，需要有高画质，于是就有了显卡。1999年，NVIDIA声称自己发明了GPU，也就是 GFFORCE 256，所谓的GPU，就是图形计算单元，他是显卡最最核心的部件，再给他配上其他一系列零件，就成为了一张显卡。GPU跟显卡，严格来说不是一个概念，只是大家平时很少特意区分。这玩意为啥能提升游戏画质？因为渲染游戏画面这件事，难就难在计算量太大了，比如游戏中任何一个3D物体，它的位置、方向、大小、光源、物体表面等变化，都需要电脑来计算。

渲染画面这件事，就像再做10000道加减乘除，CPU的核心很强，但数量少。每个核心就像出于一个智力巅峰的高三学生，他能熟练的解出模拟卷上的最后一道大题，但让他算 1000道，他得累死。而显卡上面密密麻麻的分布着几千个小核心，每个核心都像是一个小学生，高考题肯定是不会，但他们能同时并排启动，在10秒只能，把10000道题做完。所以渲染特别快。显卡能提升画质，就是因为他有这种强大的并行计算能力。而显卡从此脱离游戏领域，被别的领域盯上，乃至成为国际博弈的筹码。也是因为这种强大的并行计算能力，。发布了没几年，斯坦福大学实验室就盯上了显卡，它们想要显卡解决其他领域的问题，但是并不简单，显卡算力虽强，但是它们都是小学生，需要合适的软件能驾驭才行。沿用刚才的比方，GPU就是一万个小学生在同时工作，计算能力很强，但前提是你得能把一道难解的大题，分解成无数个小学生能解决的简单问题才行。否则显卡再强又有什么用呢？转换到现实中，就是得让开发者能方便的写出代码。利用上显卡的并行计算能力才行。所以在2006年，带领团队出现了至今仍然在不断更新的 CUDA，CUDA就是更方便的让开发人员能够面向GPU编程。如果绝大多数 AI模型的训练，背后都离不开CUDA的支持。除了CUDA，

OpenCL、ROCm平台，作用也类似，每一个拿显卡干活的人，都绕不开它们，在硬件层面，显卡有强大的并行计算能力，在软件层面，配套的编程平台也成熟了，这就意味着，GPU可以完全离开游戏领域，走向更大的世界了。

第一次感受到GPU，就是挖矿，也就是挖比特币，挖矿其实就是用计算机 来解决数学问题，比如任何数据，都可以通过哈希算法生成一串哈希值，原始数据不管发生多小的变化，最后生成的哈希值都完全不同。我们有时候下载大文件时，也会利用哈希算法的这种特性，来做一次校验。。看看下载的文件是否完整。很多挖矿，就是要生成一个符合要求的哈希值，这就需要计算机去反复的尝试，所以挖矿就跟游戏画面一样，属于那种不难，但是计算量非常大的事情。恰好能够利用显卡的算力。于是在加密价格持续攀升的日子里，显卡涨价，缺货。一路推动NVIDIA的市值从140亿美元暴涨到了 1750亿美元。但显卡跟加密货币之间只是一段露水情缘，随着专用矿机的出现，虚拟比特币的价格跳水，以太坊等调整了挖矿规则等原因，显卡跟虚拟货币的关系已经大不如前了。但显卡就跟上了更大、更革命的科技浪潮，就是AI。现在所有人都知道，AI是可能引起新一轮科技革命的巨大产业，而几乎所有的AI模型训练，都需要显卡。

就拿现在正火的ChatGPT来说，它的模型训练中涉及到大量的矩阵运算，这些矩阵运算本身不难，但是量很大很大，所以就需要GPU来并行处理。AI是可能改变世界的，而AI的基础是 算法、算力和数据。而提到的A100和H100，售价高到十几万、甚至几十万的专业显卡，还供不应求。有报道说，训练ChatGPT需要相当于300块A100显卡的算力，光这一个项目就需要花几十个亿来购买显卡，这也是为什么 从2022年10月开始，NVIDIA的市值在半年时间内就飙升了34倍。

2.8 国产AI超算芯片期待

这么着急赶尽杀绝，不惜拉上自己企业垫背。答案就是我国在半导体自主研发上，已经触碰到了他们的痛处。很多人总以为，我们依赖国外的AI芯片，是自身技术上不过关，其实真相是我们的芯片都没有真正的上过牌桌，为什么？ AI芯片只有在实际应用中才能够发现问题，加快迭代，而我们的国产芯片，起步晚、性能差，所以国内的厂商大多不愿意使用，这也就造成了国产芯片无法获得正向反馈，发展速度只会越来越慢，这是一个矛盾的循环。在还有选择的时候，考虑到性能也好，成本也好，中国企业往往愿意选择像NVIDIA的芯片，所以在很长一段时间，美国对中国的高端芯片的制裁，也不是彻底封死，因为国产的AI芯片不是它的对手，都还不够好用，但现在，局面彻底改变了。出口禁令全面升级，中国企业没有了其他的任何选择，下一步只能规模化的采购国产芯片，并且培育出本土的产业链，逐步实现真正意义上的国产替代。那可能有人说，这件事这么简单，能实现岂不是早就实现了。不太可能。其实还真不一定，放眼全球算力芯片市场，不可否认，NVIDIA占据了九成多的份额，出于高度垄断的地位。但是，目前国产AI芯片的可替代方案，也不少。

如果单看并行计算这个领域，有两家国产GPU公司值得关注：分别是摩尔线程和壁任科技。

摩尔线程2020年10月成立，在2023年10月17日，第一款产品摩尔芯用了7纳米工艺，支持CUDA平台和算法模型，性能超过每秒20万亿次浮点计算，仅成立三年就上了白宫严选名单。成为老美的制裁对象之一。是现在唯一可能买到的实际产品，并且一直在更新驱动的公司。壁任科技一款产品 壁任一号，7纳米工艺，支持CUDA平台和算法模型，性能超过每秒30万亿次浮点计算。去年发布了一个GPU 叫BR100，性能就直逼英伟达的H100，但是这个芯片并没有使用，原因还是台积电不给生产，准确的说是美国政府不让台积电给我们生产，这就是一个不公平的竞争，华为的遭遇大家就更熟悉了，19年以后 芯片的生产、制造全都被摁的死死的，但是麒麟芯片出来了，说明我们大概率已经突破了先进芯片制造的生产流程了，顶多就是成本高一点。

这些当中除了华为之外都面临相同的问题，那就是没有针对芯片专门优化的计算架构，换句话说，这些公司不具备与CUDA抗衡的能力，如果能成为芯片供应商，去识别其他的计算架构，理论上也是可以的。但是对于使用者来说这样做效率就太低了。一旦训练十几天，一旦出现BUG，不就前功尽弃了吗。所以这事还是得看华为。华为就厉害了，在人工智能领域的布局包括但不限于 昇腾系列计算芯片，CANN异构计算架构、Mindspore深度学习框架，ModelArts一站式开发平台，盘古大模型还有各种应用。海思还能做鲲鹏

920服务器CPU，这么看英伟达能做的，华为都能做，英伟达做不到的，华为也能做。华为可能是唯一可能可能布局出全栈式人工智能的公司，也是唯一一个全栈式国产化公司，困扰华为的最大问题还是美国的制裁，芯片没法生产，再怎么布局也是白搭，一旦能解决芯片问题，

有句话说的好，天下苦英伟达久矣。有时候看似把我们逼入绝境，但其实是给我了我们绝地求生的机会，就看我们怎么把握了。尽管中国企业会迎来一段痛苦的过渡期，但从长远来看，这是一条不得不走的路。而面对当下的巨大差距，最关键的是终于有了攻坚克难的凝聚力。相信度过这段阵痛期，有的人，有的国家，只会后悔的拍大腿，眼睁睁的看着被逆袭、被超越，毕竟每一次我们都是这样过来的。

三、组装计算机硬件选型策略

计算机八大件：CPU、散热器、主板、内存、硬盘、显卡、电源、风扇、机箱，我们依次来看对于一套需要部署大模型的个人计算机，如何搭配。

3.1 GPU选型策略

1. **选择厂商**

目前独立显卡主要有AMD和NVIDIA两家厂商。其中NVIDIA在深度学习布局较早，对深度学习框架支持更好。**建议选择NVIDIA的GPU。**

> 桌面显卡性能天梯图：https://www.mydrivers.com/zhuanti/tianti/gpu/index.html

2. **选择系列及品牌**

对个人用户来说，就是从NVIDIA的RTX系列中，选择出合适的GPU。就部署大模型的需求来说，只需考虑计算能力和显存大小就可以了。显存带宽通常相对固定，选择空间较小。目前各级别显卡的均价如下：

显卡型号	显存大小	单卡价格
4090	24G	1.9w+
4080	16G	9k+
3090Ti	24G	1w+
3090	24G	8.5k+
2080Ti	11G	2.2k+

但需要注意：不同的显卡品牌，及市场行情，显卡价格会有上下浮动。详细品牌介绍如下：

- **一线（三大厂）**

品牌	华硕	微星	技嘉
顶级旗舰			
旗舰	ROG猛禽	超龙X	大雕
次旗舰	TUF	魔龙	超级雕/小雕
中端	巨齿鲨	/	雪鹰/魔鹰
丐版	DUAL	万图师	猎鹰

华硕显卡品控优做工好，高品牌信仰足，但溢价严重，这个牌子最会宣传，卡不错但性价比不是很高。

高端优先推荐华硕ROG猛禽，当然缺点就是：贵，另外主流用户个人更推荐TUF，更低端的巨齿鲨和DUAL不太推荐

微星显卡30系列之前更推荐魔龙，30系列更推荐超龙

- **准一线**

品牌	七彩虹
顶级旗舰	九段
旗舰	火神/水神
次旗舰	adoc
中端	ultra
丐版	战斧

推荐七彩虹，用料良心，保修出名的好，深受矿老板们的喜爱，支持个人送保，ultra，三风扇，散热性能极好，噪音小，白色颜值高，不带rgb灯效，喜欢rgb灯效的可以选择adoc。

- **二线**

品牌	影驰	索泰	映众	耕升	铭瑄
顶级旗舰					
旗舰	名人堂	PGF/AMP	冰龙寒霜		MGG 大玩家
此旗舰	GAMER/星耀	天启	超级冰龙		电竞之心
中端	金属大师	/	冰龙	炫光/星极	
丐版	黑将/大将	X-gaming	黑金至尊	追风	turbo/终结者

二线品牌中，影驰和索泰属于二线中实力比较强的两个，可以说是强二线，索泰重点推荐PGF（排骨饭）和天启，其中天启就是之前经典的至尊PLUS系列的升级

影驰：原为是我国香港的品牌，在08年被台湾同德收购，目前也算台湾省的品牌，旗下名人堂属于卡皇级别的产品，颜值高，性能强，次旗舰GAMER和星耀一个主打DIY一个主打RGB，都是非常有特点的产品

- 企业级显卡

  参考第二部分的GPU推荐。

- 服务器推断卡

  除了用于训练，还有一类卡是用于推断的（只预测，不训练），如：

型号	架构	价格(元)	显存(GB)	CUDA核	Tensor核	FP32(TFLOP
Tesla P4	Pascal	14500	8	2560	NA	5.5
Tesla P40	Pascal	43999	24	3584	NA	12
Tesla P100	Pascal	48900	16	3584	NA	10.6
Tesla V100	Volta	79600	32	5120	640	15.7
Tesla T4	Turing	19500	16	2560	320	8.1

这些卡全部都是不带风扇的，但它们也需要散热，需要借助服务器强大的风扇被动散热，所以只能在专门设计的服务器上运行，性价比首选 Tesla T4，但是发挥全部性能需要使用 TensorRT 深度优化，目前仍然存在许多坑，比如当你的网络使用了不支持的运算符时，需要自己实现。

- **避免踩坑**

如果选择配置单机多卡，采购显卡的时候，**一定要注意买涡轮版的**，不要买两个或者三个风扇的版本，除非只打算买一张卡。因为涡轮风扇的热是往外机箱外部吹的，可以很好地带走热量，如果买三个风扇的版本，插多卡的时候，上面的卡会把热量吹向第二张卡，导致第二张卡温度过高，影响性能。

3.2 CPU选型策略

CPU在大模型使用中起到什么作用？当在GPU上运行大模型时，CPU几乎不会进行任何计算。最有用的应用是数据预处理。CPU负责将数据从系统内存传输到GPU的显存中，同时也处理GPU完成计算后的数据。有两种不同的通用数据处理策略，具有不同的CPU需求。

- 训练时处理数据：高性能的多核CPU能显著提高效率。建议每个GPU至少有4个线程，即为每个GPU分配两个CPU核心。每为GPU增加一个核心，应该获得大约0-5%的额外性能提升。
- 训练前处理数据：不需要非常好的CPU。建议每个GPU至少有2个线程，即为每个GPU分配一个CPU核心。用这种策略，更多内核也不会让性能显著提升。

在这种情况下，GPU通常承担大部分计算负担，CPU的作用更多是管理和协调，因此需要高核心数，同时也需要快速的数据预处理，同样需要高频率，所以**高核心 + 高频率，虽然不是必须，但推我们推荐还是能高即高，标准是：要与选择的GPU和CPU的性能水平相匹配**，避免将一款高端显卡与低端CPU或一款高性能CPU与低端显卡匹配，因为这可能导致性能瓶颈。比如：

- NVIDIA GeForce RTX 3090 * 2 搭配 Intel Core i7-13700K/KF 或 AMD 5900X CPU；
- NVIDIA GeForce RTX 3080 搭配 Intel Core i9-11900K CPU。

但相对来说，瓶颈没有那么大，一般以一个GPU对应 2~4 个CPU核数就满足基本需求，比如单卡机器买四核CPU，四卡机器买十核CPU。在训练的时候，只要数据生成器（DataLoader）的产出速度比 GPU 的消耗速度快，那么 CPU 就不会成为瓶颈，也就不会拖慢训练速度。

1. **选择品牌**

目前消费市场CPU品牌就两家，Intel和AMD，由于历史原因，英特尔目前市场占有率比较高，而且在过去很长的一段时间了里，英特尔一直是绝对霸主的存在，代表最先进的生产力，时至今日，AMD在产品性能层面已经完全可以和Intel正面硬刚了。

> CPU性能天梯图：https://www.mydrivers.com/zhuanti/tianti/cpu/index.html

- **Intel 系列命名规范**

可以通过CPU名称得到一些信息，如i7-10700K，代表产品型号是i7，后面的10代表是第10代，然后700代表性能等级高低，K代表这个CPU可以超频，当然后缀字母还有T、X、F等，X后缀代表高性能处理器，而T代表超低电压，/F代表无CPU无内置显卡版本。

1. 系列：由低到高 Celeron（赛扬） / Pentium（奔腾） /酷睿系列的i3 / i5 / i7 / i9
2. 世代：第1组数字代表是第几代
   例如这三个CPU：I7-8700、I7-9700，i7-10700第1个是第八代，第2个是第九代、第3个是第十代，还是比较容易理解的。
3. 性能：第2组(3个数)是表示性能等级
   例如：I5-12400、I5-12500，数字越大表示越好。
4. 后缀：K→可超频，F→没有核显
   可超频K版CPU要搭配可超频的Z系列主板才行，才可以超频，搭载不能超频的主板也可以用，只是不能超频了而已。
   没有核显的F版CPU要搭配独立显卡才能开机点亮屏幕

> 超频简单理解就是可以提升处理器性能，核显就是把一张入门级的显卡塞进处理器里。

i3是家用级别，i5是游戏级别，i7是生产力和游戏发烧友级别，i9是最顶级的。后缀带K可以超频，带F表示没有核显。

- **AMD系列命名规范**

和Intel类似：

1. 系列：由低到高 APU / Althlon（速龙） / Ryzen（锐龙）系列 R3 / R5 / R7 / R9
2. 世代：第1个数字代表第几代
3. 例如这两个CPU：R7-2700X、R7-3700X，第1个是第二代，第2个是第三代。
4. 性能：第2组数字（3个数字）表示性能等级
   数字越大性能越好，例如 R7 3800X的性能大于R7 3700X。
5. 后缀：字母G表示有核显，字母X没有明确意思，一般性能强一点。如R5 3600X比R5 3600性能高一点，主频高一点。

其中 r3家用，r5游戏级别，r7生产力级别，r9顶级生产力，x表示高频版本。带g 代表有核显。

要选Intel还是AMD，其实都可以。如果追求性价比，AMD性价比高一些，如果主要玩游戏，且对价格不敏感，建议选择英特尔Intel，英特尔Intel一般主频较高，一些游戏主要依赖主频，所以高主频的Intel玩游戏更推荐一些。除了品牌维度的分析，目前**主流的大模型训练硬件通常采用 Intel + NVIDIA GPU**。但具体情况具体分析，只能简单说：一分钱一分货，一般来说贵的好。

- **选购CPU误区**

电子产品有一个说法是，“买新不买旧”，一般新产品，会用更新的工艺架构，性能更强，功耗更低，比较值得购买。当然有些时候，老一代的性价比很高，也可以考虑，但如果是好几代前的老产品，就不要考虑了，有些商家会卖几年前的i7电脑主机，它的性能可能还不如最新的i3，主要是忽悠小白的，要注意辨别。

目前消费级市场，我们最常听到的i3 i5 i7 等，是英特尔的酷睿系列产品，主要面向一般消费市场，数字大的性能更强，（注意这里只在同代产品中成立）。AMD与之对应的是R3 R5 R7。这里值得注意是，同代产品i7比i5强，如果拿老一代的i7和新一代的i5比，就未必成立，部分商家经常会营销i5免费升级i7，其实是把最新一代的i5换成立了老一代的i7，性能方面可能还不如没升级呢？比如i5-8400的性能就高于i7-7700.

3.3 散热选型策略

CPU 不断地更新换代和性能提升，其功耗和发热量也越来越大，如果温度过高，就会出现自动关机或者是蓝屏死机等情况，所以需要单独的散热器来压制，目前**CPU散热器分两种：水冷和风冷。**

风冷和水冷系统都是用于GPU的散热解决方案。它们各有优势和不足：通常，**水冷系统在散热效率方面优于风冷系统。**以Intel的i9-13900KF为例，这款CPU性能目前位于CPU性能天梯榜第二位，很多用户认为使用水冷系统是必要的。但如果这款CPU没有超频需求，使用高质量的风冷系统其实也能够有效地散热。只有在超频的情况下，水冷系统由于其更出色的冷却效果，才成为更佳的选择。

但需要注意，**风冷和水冷与GPU无关。**在计算机硬件中，CPU和GPU（显卡）的散热策略和要求各有不同。CPU通常需要配单独的散热器，我们可以根据需要选择并购买不同类型的散热器，例如水冷或风冷系统，并且可以根据性能要求进行升级。由于CPU的**高主频和较少的核心数（通常是几个到二十几个核心）**，高性能的CPU在运行时会产生较多热量，因此需要更有效的散热解决方案来维持合适的运行温度。与此相对，显卡通常采用一体化的散热设计，其散热器是显卡的重要组成部分。显卡散热器的设计已经由各厂商经过测试和优化，因此用户一般不需要担心显卡的散热问题。显卡采用的是**多核心、低频率**的策略，即使是高端显卡如Nvidia的4090，其频率也相对较低，通常在3000MHz左右，而同代的高端CPU（如Intel i9）的频率可能是其两倍。显卡的散热器可以直接接触GPU核心和显存，从而高效散热。因此，在正常满载情况下，显卡的温度达到70多或80多度是正常现象，通常不会成为性能瓶颈。

对于大模型部署来说，首要原则还是**CPU的等级要和GPU相匹配。**对于中低端处理器，如Intel的i5系列，以及AMD的R5和R7系列，一般推荐使用风冷系统。这些处理器的热设计功耗（TDP）通常较低，风冷系统足以提供有效的散热。而对于更高性能的处理器，如Intel的i7 13700KF及更高级别的i7和i9系列，建议至少使用240mm规格的水冷系统。考虑到这些处理器较高的性能和热输出，水冷系统能提供更为高效和稳定的冷却效果。因此，尽管风冷在某些情况下依然可行，但为了确保最佳性能和稳定性，对于高性能处理器，更推荐使用水冷系统。

3.4 主板选型策略

在构建大模型的系统时，低端主板通常不适用。根据所选的CPU和GPU规格，应从中端或高端主板中选择出合适的。

制造商	系列	定位	支持CPU超频	支持内存超频	适用用户
Intel	Z系列	高端	是	是	追求高性能和定制化设置的用户
Intel	B系列	中端	否	是	需要一定性能但预算有限的用户
Intel	H系列	低端	否	否	预算有限或对性能要求不高的基本用途
AMD	X系列	高端	是	是	追求最高性能和高度定制化的用户
AMD	B系列	中端	否	是	寻求性价比的用户
AMD	A系列	低端	否	否	有限预算或基本计算需求的用户

选择主板时，核心因素是**确保它与CPU的性能和超频能力相匹配。**以Intel处理器为例：对于中高端CPU（如i5系列及以上），更适合选择B660到Z690系列的主板。对于如13600KF这样的高性能CPU，至少应选择B660系列的主板作为起点。需要考虑的是CPU是否支持超频（如带有“K”后缀）。可超频的CPU更适合搭配支持超频的高端主板，如Z系列

其次，需要**检查CPU和主板型号是否匹配及合理。**

通常情况下，每一种型号的CPU都需要搭配对应型号的主板，每代CPU和主板都有自己的针角及接口类型，Intel cpu不能用于AMD系列主板，某些主板可能会通用几代cpu，但有的主板只能兼容某一代，例如intel 十代 的i510400f，不能用于早期的四代 b85系列主板，而是否匹配，指的是高性能CPU搭配低性能主板，h610是入门主板，虽然可以点亮，但是低端主板因为供电有限，无法发挥出cpu的全部性能，以及无法超频，这样就失去了cpu本身的性能和意义。

最后，**考虑PCIe通道。**

PCIe通道是一种高速接口，用于将GPU连接到计算机的主板。通过这些通道，GPU可以与CPU以及系统内存快速交换数据。每个PCIe通道（或称为“通道”）都提供一定的数据传输带宽。更多的通道意味着更高的总体带宽。例如，PCIe 3.0 x16接口意味着有16个通道，每个通道的速度是PCIe 3.0标准的速度。

GPU的性能部分取决于它与主板之间的通信速度，这是由PCIe通道的数量和版本（如PCIe 3.0、4.0或5.0）决定的。更高版本的PCIe提供更高的传输速率，从而可能提高GPU的性能。以下是需要考虑的几个关键点：

1. **PCIe版本**：

- PCIe有多个版本，如PCIe 3.0、4.0、5.0等，每个版本的带宽和速度都有所不同。新版本（如PCIe 4.0和5.0）提供更高的数据传输速率，这对于高性能GPU和其他高速设备非常重要。

2. **PCIe槽数量和布局**：
   - 主板上的PCIe槽数量决定了可以安装多少个扩展卡。如果计划安装多个GPU或其他PCIe设备，需要确保主板有足够的槽位。
   - 槽位布局也很重要，尤其是在安装大型GPU时，需要确保它们之间有足够的空间，避免过热或物理干扰。

3. **PCIe通道分配**：
   - 主板上的PCIe通道是从CPU和芯片组分配的。不同的主板可能有不同的通道分配方式，这可能会影响到扩展卡的性能，特别是在多GPU配置中。
   - 确认主板是否支持您所需的PCIe配置，例如双向或四向GPU设置。

4. **与GPU的兼容性**：
   - 虽然大多数现代GPU兼容大多数主板的PCIe槽，但是为了最佳性能，最好确认GPU与主板的PCIe版本相匹配。

综上所述，因为需要通过PCIe通道连接和使用GPU，因此在选择主板时考虑PCIe通道的版本、数量、布局和通道分配非常重要。

3.5 硬盘选型策略

首先考虑接口类型。主流固态硬盘主要有两种接口：SATA和M.2。

- **SATA接口**的固态硬盘体积较大，形状类似于传统的机械硬盘，主要用于升级老式电脑，因为这些电脑通常不具备M.2接口。SATA接口硬盘的最高速度为600MB/s。
- **M.2接口**的硬盘则较小，可以直接安装在主板上的专用接口。它们采用新的硬盘协议，速度上可以达到4GB/s。

推荐选择 M.2接口的硬盘。

然后考虑协议。M.2接口的固态硬盘分为SATA协议和NVMe协议两种。

- M.2接口的**SATA协议硬盘**速度较慢，实际上就是标准SATA硬盘的形状变化，速度仍然是最高600MB/s，这类硬盘多用于旧电脑。
- **NVMe协议硬盘**则速度更快，适合对速度有较高要求的应用。

在面对大量小文件的时候，使用 NVMe 硬盘可以一分钟扫完 1000万文件，如果使用普通硬盘，那么就需要一天时间。**推荐选择 NVME协议的M.2接口的硬盘。**

最后考虑 PCIe等级。当前市面上最新的是PCIe 5.0，但更常见的是PCIe 3.0和PCIe 4.0。PCIe等级越高，硬盘的速度潜力越大。但重要的是检查主板是否支持相应的PCIe等级。例如，一些主板可能最高只支持到PCIe 4.0。一般来说，选择PCIe 4.0的即可。

硬盘不会限制深度学习任务的运行，但如果小看了硬盘的作用，可能会让你追、悔、莫、及。想象一下，如果你从硬盘中读取的数据的速度只有100MB/s，那么加载一个32张ImageNet图片构成的mini-batch，将耗时185毫秒。

3.6 内存选型策略

建议内存容量应大于GPU的显存。例如，对于搭载单卡GPU的系统，建议配置至少16GB内存。如果是四卡GPU系统，则建议至少配置64GB内存。由于数据生成器（DataLoader）的存在，数据不需要全部加载到内存中，因此内存通常不会成为性能瓶颈。

内存不用太纠结，是GPU显存的一到两倍。目前，128G 就可以，64G 也凑合。而且内存没那么贵，可以配满。

内存大小不会影响深度学习性能，但是它可能会影响你执行GPU代码的效率。内存容量大一点，CPU就可以不通过磁盘，直接和GPU交换数据。所以应该配备与GPU显存匹配的内存容量。

在选择的时候，**注意检查主板是否支持内存的数量及型号。**目前常见的 ddr3 ~ 5，每一代内存都需要对应主板的插槽类，ddr 5代内存 是无法混插在 ddr 4代内存上的。另外需要确定主板的内存插槽数量，如果只有两个插槽，买了四个，那么根本插不进去。

其次检查cpu主板是否支持内存频率。内存条的频率上限 受到 cpu 和主板的控制， 例如 i5的 10400f + b460主板 = 2666，如果你买的内存是 3600频率的，无疑发挥不出内存本身的优势。

3.7 电源选型策略

在选择电脑电源时，需要**检查电源的瓦数是否足以支持整机的功耗。**并非越高瓦数越好，但瓦数过低可能会导致关机或黑屏等问题。在确定合适的电源瓦数之前，应综合评估整机硬件的功耗，尤其要考虑CPU和显卡这两个功耗大户。通常，将CPU和显卡的TDP功耗相加后乘以2可以得到一个合适的电源瓦数估计。例如，对于一个65W的CPU和125W的显卡，合适的电源瓦数应该在400W或450W左右。

双卡最好1000W以上，四卡最好买1600W的电源

3.8 机箱选型策略

最后，选择完所有上述所有配件之后，选择机箱就相对简单，只需要确保这个机箱足够宽敞，能够容纳所选的所有配件。我们需要检查以下几项内容：

1. **核对主板与机箱尺寸匹配性**：
    - 确保所选主板的大小与机箱兼容。例如，ITX主板应与ITX机箱相匹配。这就像选择合适大小的鞋子一样重要。

2. **确认机箱支持显卡尺寸**：
    - 对比显卡的长度与机箱对显卡长度的限制。建议选择机箱的显卡限长至少比显卡长度长出30毫米以上，以确保有足够空间进行安装和通风。

3. **检查散热器与机箱的兼容性**：
    - 非常重要的一点是比较散热器的尺寸与机箱的散热限高。如果散热器太高，可能无法正常安装侧盖。
    - 考虑到许多内存条配备较高的散热马甲，需要确认散热风扇是否会受到内存条的干扰。
    - 如果选择水冷散热系统，确保机箱的水冷位能够容纳水冷冷排的尺寸。例如，360mm的水冷冷排无法安装在仅适用于240mm的位置上。

4. **检查机箱对电源尺寸的支持**：
    - 常见的电源类型包括SFX、ATX和TFX。由于不同规格的电源在形状和大小上有所不同，必须确认机箱的电源仓是否适合所选电源的尺寸。