基于双卡 RTX 4090 搭建家用深度学习主机

Yue Shui

2024-12-21T12:00:00+08:00

Table of Contents

## 租用 GPU 还是购买 GPU？

在构建深度学习工作环境之前，首先需要综合考虑 **使用周期**、**预算**、**数据隐私** 以及 **维护成本**。如果长期（例如超过一年以上）且对数据安全要求较高，自建 GPU 服务器通常能带来更低的综合成本和更可控的环境；如果只是短期项目，或对数据隐私不敏感，那么租用云上 GPU（如 Azure、AWS、GCP 等）或使用免费平台（Colab、Kaggle）则更加灵活。

* **租用 GPU 的优点**：
  + 无需一次性投入高额硬件成本
  + 可根据项目需求弹性扩容
  + 云厂商通常提供数据合规与安全保障，省去硬件运维烦恼
* **购买 GPU 的优点**：
  + 长期大规模使用时，整体成本更低
  + 对内部数据和模型有更高的隐私与可控性
  + 硬件可随时调整、升级，部署更灵活

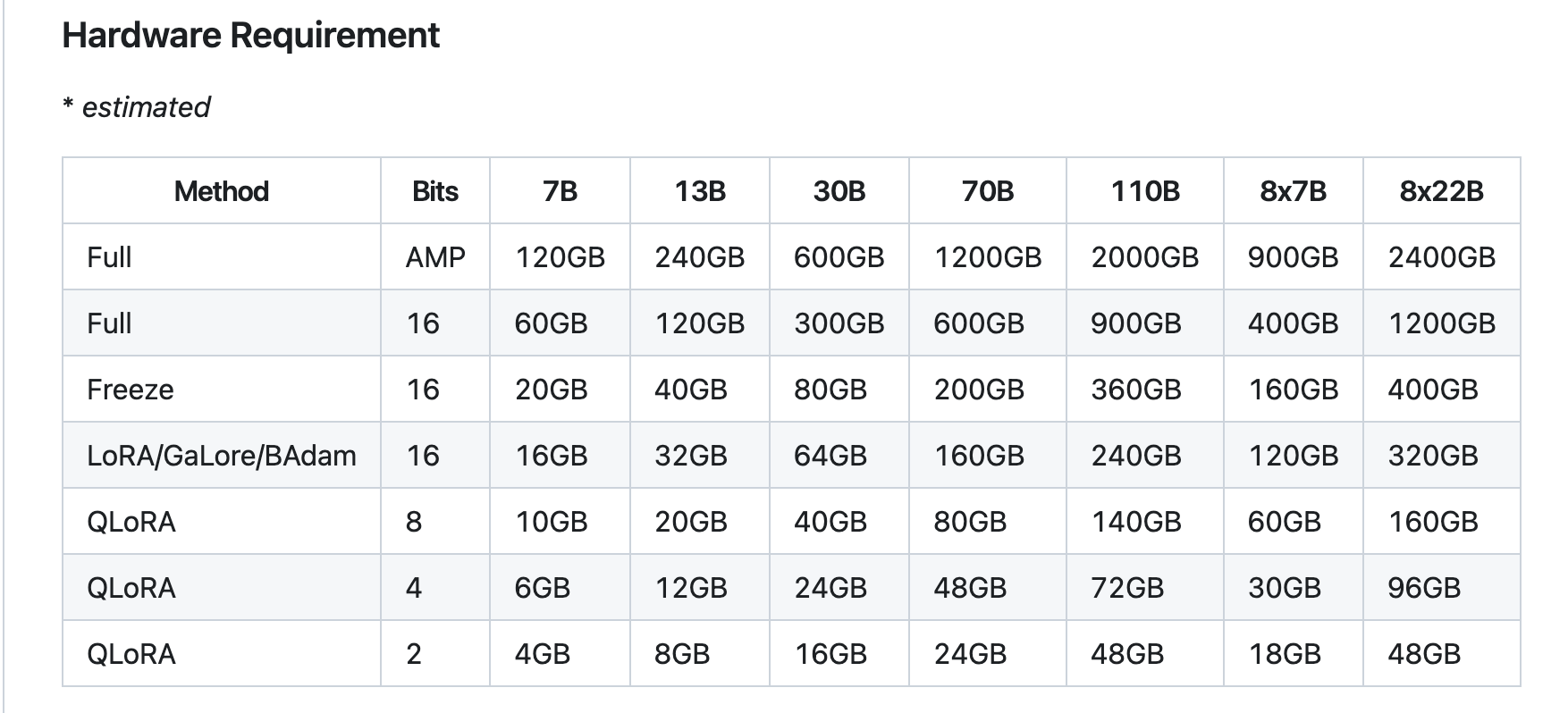
**个人建议**  
1. 如果预算有限或只是初学阶段，可先使用 Colab、Kaggle 或云 GPU；  
2. 当算力需求和隐私需求上升时，再考虑自建多卡服务器或租用多机多卡集群。

## 背景

在 2023 年 9 月，为了在工作之余继续对大模型（LLM）进行探索和研究，我组装了一台 **双 RTX 4090** 的个人 AI 实验服务器。该服务器已运行近一年，整体体验如下：

* **噪音**：服务器放在脚边，满负荷训练时风扇噪音较大，但在日常推理或中等负载下可接受
* **推理性能**：双卡共计 48GB 显存，采用 4bit 量化方案时可运行到 70B 级别的模型（如 Llama 70B、Qwen 72B）
* **训练性能**：在使用 [DeepSpeed](https://github.com/microsoft/DeepSpeed) 的分布式和 offload 技术（ZeRO-3 + CPU offload）后，可对 34B 左右的模型（如 CodeLlama 34B）进行微调
* **性价比**：对于个人或小团队的日常实验和中小规模模型训练而言，该配置较为实用；但若进行超大规模模型的全参数训练，仍需更多专业卡（如多卡 A100 或 H100 集群）

下图展示了不同大小模型、不同训练方法对显存的需求（参考 [LLaMA-Factory](https://github.com/hiyouga/LLaMA-Factory#hardware-requirement)）：



硬件需求示意图

## 搭建思路与配置详情

整机预算在 **4 万元人民币（约 6000 美元）** 左右，以下是我最终选用的配置清单，仅供参考：

|  |  |  |
| --- | --- | --- |
| 配件 | 型号 | 价格 (元) |
| **显卡** | RTX 4090 \* 2 | 25098 |
| **主板 + CPU** | AMD R9 7900X + 微星 MPG X670E CARBON | 5157.55 |
| **内存** | 美商海盗船(USCORSAIR) 48GB\*2 (DDR5 5600) | 2399 |
| **SSD** | SOLIDIGM 944 PRO 2TB \*2 + 三星 990PRO 4TB | 4587 |
| **电源** | 美商海盗船 AX1600i | 2699 |
| **风扇** | 追风者 T30 12cm P \* 6 | 1066.76 |
| **散热** | 利民 Thermalright FC140 BLACK | 419 |
| **机箱** | PHANTEKS 追风者 620PC 全塔 | 897.99 |
| **显卡延长线** | 追风者 FL60 PCI-E4.0 \*16 | 399 |

**总计**：约 42723.3 元

### GPU 选择

对于大模型研究，**浮点运算性能（TFLOPS）** 和 **显存容量** 是最核心的指标。专业卡（A100、H100 等）虽有更高显存以及 NVLink，但价格动辄数十万人民币，对个人用户并不友好。根据 [Tim Dettmers](https://timdettmers.com/2023/01/30/which-gpu-for-deep-learning/) 的调研，RTX 4090 在单位美元算力方面表现非常亮眼，且支持 BF16、Flash Attention 等新特性，因此成为高性价比的选择。

#### 散热方式：涡轮 vs 风冷 vs 水冷

|  |  |  |  |
| --- | --- | --- | --- |
| **散热方式** | **优点** | **缺点** | **适用场景** |
| **涡轮风扇** | 体积紧凑；适合并行多卡部署 | 噪音大、整体散热效率一般 | 企业服务器机柜、多卡密集部署 |
| **风冷散热** | 性能与噪音平衡佳、维护简单 | 显卡体型通常较大 | 家用或个人研究（主机摆放空间足够） |
| **水冷散热** | 散热能力突出、满载噪音更低 | 可能会出现漏液、价格更高 | 对静音要求极高或极限超频场景 |

**家用推荐**：**风冷卡** 兼顾散热效率、噪音和维护成本；相对于涡轮卡和水冷卡更友好。

### CPU & 主板

在深度学习场景中，CPU 主要负责数据预处理、管道调度以及多进程/多线程并行管理，确保数据能够以高吞吐量、低延迟的方式传递到 GPU。因此，CPU 的核心需求主要包括 **充足的 PCIe 通道** 和 **卓越的多线程性能**。

* **Intel**：13/14 代 i9（如 13900K）拥有 20 条 PCIe 主通道，能够满足双卡 x8 + x8 的需求
* **AMD**：Ryzen 7000/9000 系列（如 7950X）提供 28 条（可用 24 条）PCIe 通道，支持双卡 x8 + x8，并为 M.2 SSD 提供足够带宽

#### 微星 MPG X670E CARBON 主板

* **扩展性**：支持 PCIe 5.0 和 DDR5 内存，具备充足的未来升级空间
* **稳定性**：高规格供电设计，保障 CPU 与多显卡的稳定运行
* **接口丰富**：支持多块 M.2 SSD 和 USB4，满足多样化使用需求

#### AMD Ryzen 9 7900X 特点

* **核心与线程**：12 核心、24 线程，在深度学习场景中的数据预处理和多任务处理方面表现强劲
* **PCIe 带宽**：提供 28 条（可用 24 条）PCIe 5.0 通道，可轻松支持双卡 x8 + x8，并为 M.2 SSD 提供高速带宽
* **能效比**：基于 Zen 4 架构，性能与能耗平衡优秀，适合高性能计算需求

#### 主板选购要点

1. **空间布局**
   * RTX 4090 尺寸庞大且卡槽较厚，需确认主板是否能同时容纳两张显卡；若存在空间或散热冲突，可使用显卡延长线竖放第二张卡。
2. **PCIe 通道拆分**
   * 主板需至少支持双 PCIe 4.0 x8 + x8 配置，以避免出现 x16 + x2 的情况。x16 + x2 的带宽分配会显著限制第二块 GPU 的数据传输能力，进而影响 GPU 与 CPU 之间的数据交换效率。在大模型训练中，这种带宽瓶颈可能导致性能显著下降，严重影响整体训练效率。
3. **扩展性**
   * 在双显卡插满的情况下，仍需确保主板具有足够的 M.2 SSD 插槽和外设接口

综合扩展性、性能与性价比等因素，我最终选择了 **AMD Ryzen 9 7900X 搭配 微星 MPG X670E CARBON 主板** 的组合。通过显卡延长线解决了 4090 双卡过厚导致的插槽冲突问题。

#### BIOS 设置建议

1. **内存优化**
   * 开启 **XMP/EXPO**（对应 Intel/AMD）以提升内存频率，增强带宽性能。
2. **超频调整**
   * 如果需要进一步提升性能，可在 BIOS 中启用 **PBO（Precision Boost Overdrive）** 或 Intel Performance Tuning，并结合系统监控工具观察稳定性。
3. **温度与稳定性**
   * 避免过度超频，注意控制温度，避免因崩溃或过热导致系统不稳定。

### 内存

深度学习训练中，内存会被大量占用用于数据加载、模型优化状态储存（尤其在多 GPU Zero-stage 并行场景下）。**内存容量最好 ≥ 显存总容量的两倍**。本配置中，使用了 48GB \* 2（共 96GB），满足日常多任务和分布式训练的需求，减少内存不足导致的频繁 swap。

### 硬盘

* **优先选用 M.2 NVMe SSD**：其读写性能更优，对加载超大模型权重、缓存中间文件、训练日志等都有显著速度提升
* **容量建议 ≥ 2TB**：随着大模型文件越来越庞大，2TB 往往很快就会被占满，建议根据自身需求选 4TB 或更多
* **SSD 品牌**：三星、海力士或西部数据等主流大厂都拥有稳定的高端产品线

### 电源

双 4090 满载时整机功耗可达 **900W~1000W 左右**，CPU、主板和硬盘等还需额外功率余量。通常建议选择 **1500W 以上** 的铂金或钛金电源，以确保在高负载下电流供给稳定、降低电压波动带来的系统不稳定。  
我在此使用美商海盗船 AX1600i（数字电源），可以通过软件监控实时功耗，并提供充足冗余。

### 散热与风扇

我采用 **风冷** 方案，包括：

* **CPU 散热器**：利民 FC140（双塔式气冷方案，兼顾了较高的散热效率和相对低噪音）
* **机箱风扇**：追风者 T30 12cm \* 6，保持机箱内部正压或者稍微正压的风道布局，保证显卡和供电模块的进风顺畅

在 GPU 长时间高负载训练（如分布式训练大型模型）时，机箱内的风道管理和风扇配置非常重要。建议使用监控软件及时查看 CPU、GPU、主板供电模块温度，适度调节风扇转速。

**散热进阶**  
- 若对静音有更高要求，可考虑 *Hybrid* 散热（半水冷方案）或更精细的风扇调速曲线。  
- 适度清理机箱灰尘、使用防尘网并定期更换导热硅脂也能提升散热和稳定性。

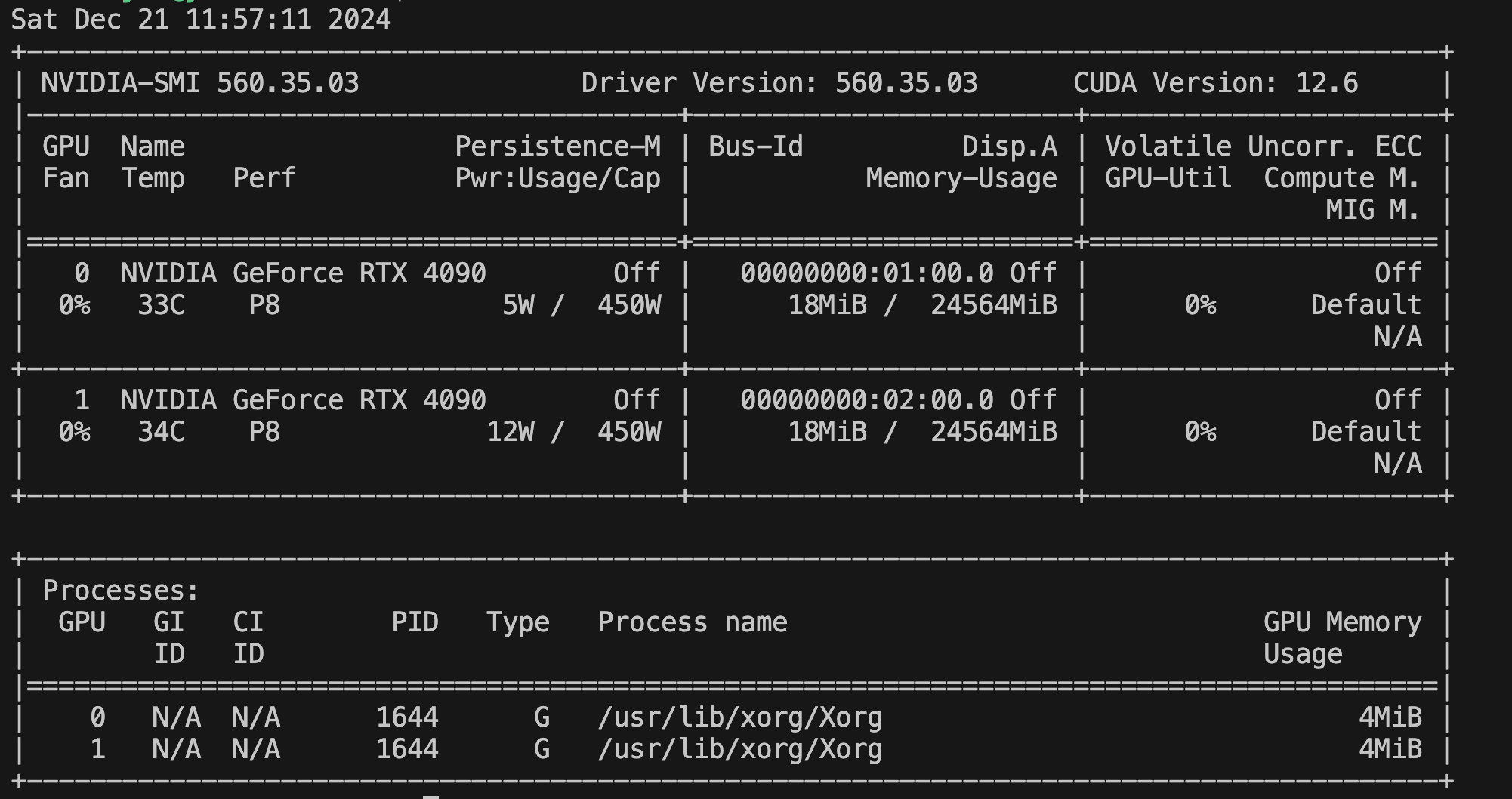
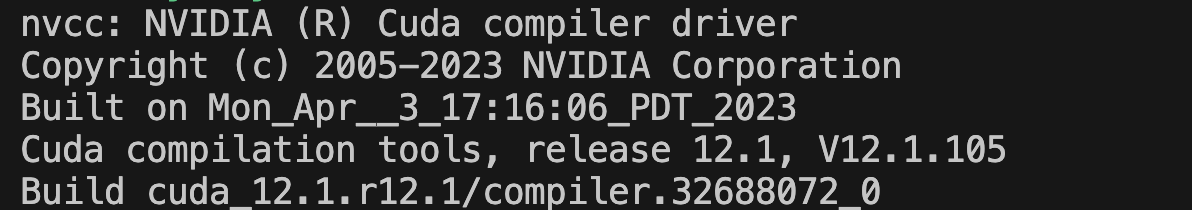
### 机箱

RTX 4090 体型巨大，且双卡堆叠时需要充足的内部空间和散热风道。全塔机箱能提供更好的走线空间和气流组织。我选用了 PHANTEKS 追风者 620PC，除了体型大、空间充裕外，也自带良好的线缆管理通道。

装机完成后的示意图如下：  


## 系统与软件环境

**操作系统**方面强烈推荐 **Linux**，例如 **Ubuntu 22.04 LTS**，因其对 CUDA、NVIDIA 驱动以及常见深度学习框架有更好的支持和兼容性。大致流程如下：

1. **安装 OS**：使用 Ubuntu 或其他 Linux 系统即可。
2. **安装 NVIDIA 驱动**：确保 nvidia-smi 能正确识别两张 4090:  
   
3. **安装 CUDA 工具链**：通过 nvcc -V 确认版本信息:  
   
4. **安装 cuDNN**：确保深度学习框架可以调用 GPU 加速卷积和 RNN 等操作
5. **测试框架**：使用 [PyTorch](https://pytorch.org/)、[TensorFlow](https://www.tensorflow.org/) 或 [JAX](https://github.com/google/jax) 简单测试模型推理/训练是否正常
6. **Docker 容器化**：
   * 利用 [nvidia-container-toolkit](https://github.com/NVIDIA/nvidia-container-toolkit) 让容器直接访问 GPU 资源，避免主机环境污染。
   * 在多机多卡环境下，还能结合 **Kubernetes**、**Ray** 或 **Slurm** 等进行集群调度与资源管理。

## 常用工具与框架推荐

1. **训练框架**
   * [**LLaMA-Factory**](https://github.com/hiyouga/LLaMA-Factory)：对大语言模型训练/推理流程有较好封装，新手友好
   * [**DeepSpeed**](https://github.com/microsoft/DeepSpeed)：支持大模型分布式训练、多种并行策略和优化功能
   * [**Megatron-LM**](https://github.com/NVIDIA/Megatron-LM)：NVIDIA 官方的大规模语言模型训练框架，适合多机多卡场景
2. **监控 & 可视化**
   * [**Weights & Biases**](https://wandb.ai/) 或 [**TensorBoard**](https://www.tensorflow.org/tensorboard)：实时监控训练过程中的损失函数、学习率等指标，支持远程可视化
3. **推理工具**
   * [**ollama**](https://github.com/jmorganca/ollama)：基于 [llama.cpp](https://github.com/ggerganov/llama.cpp) 的本地推理部署，可快速启动
   * [**vLLM**](https://github.com/vllm-project/vllm)：主打高并发、多用户场景下的推理吞吐量优化

|  |  |  |
| --- | --- | --- |
| * **Framework** | * **Ollama** | * **vLLM** |
| * **作用** | * 简易本地部署 LLM | * 高并发 / 高吞吐的 LLM 推理 |
| * **多请求处理** | * 并发数增加时，推理速度下降明显 | * 并发数增大也能保持较高吞吐 |
| * **16 路并发** | * ~17 秒/请求 | * ~9 秒/请求 |
| * **吞吐对比** | * Token 生成速度较慢 | * Token 生成速度可提升约 2 倍 |
| * **极限并发** | * 32 路以上并发时，性能衰减较严重 | * 仍能平稳处理高并发 |
| * **适用场景** | * 个人项目或低并发应用 | * 企业级或多用户并发访问 |

1. **WebUI**
   * [**Open-WebUI**](https://github.com/open-webui/open-webui)：基于 Web 界面的多合一 AI 界面，支持多种后端推理（ollama、OpenAI API 等），便于快速原型和可视化

## 进阶建议

1. **开发与调试效率**  
   使用 SSH 工具提升远程开发效率，制作自定义容器镜像减少环境配置时间。
2. **量化与剪枝**  
   通过 4bit、8bit 量化和剪枝技术，减少模型参数和显存需求，优化推理性能。
3. **混合精度训练**  
   使用 BF16 或 FP16 提升训练速度，结合 GradScaler 提高数值稳定性。
4. **CPU 协同优化**  
   使用多线程、多进程或 RAM Disk 缓存优化数据加载，支持流式加载大规模预训练数据集。
5. **多机集群部署**  
   通过 InfiniBand 或高速以太网搭建集群，使用 Kubernetes 实现高效资源调度。

## 总结

通过以上配置与思路，我成功搭建了一台 **双卡 RTX 4090** 深度学习主机。它在 **推理** 和 **中小规模微调** 场景中表现良好，对于想要在个人或小团队环境下进行大模型（LLM）科研或应用探索的人来说，这种方案兼具 **性价比** 与 **灵活性**。当然，如果要大规模全参数训练上百亿乃至上千亿参数的大模型，依然需要更多 GPU（如多卡 A100/H100 集群）。

就个人体验而言，双 4090 在预算范围内提供了较好的训练与推理性能，可以满足绝大部分中小规模研究与实验需求，值得有条件的个人或小团队参考。

## 参考资料

1. [Tim Dettmers: Which GPU for Deep Learning? (2023)](https://timdettmers.com/2023/01/30/which-gpu-for-deep-learning/)
2. [Intel 14900K PCIe 通道规格](https://www.intel.com/content/www/us/en/products/sku/236773/intel-core-i9-processor-14900k-36m-cache-up-to-6-00-ghz/specifications.html)
3. [AMD R5 7600X PCIe 通道规格](https://www.amd.com/en/products/processors/desktops/ryzen/7000-series/amd-ryzen-5-7600.html)
4. [MSI MPG X670E CARBON 规格](https://www.msi.com/Motherboard/MPG-X670E-CARBON-WIFI/Specification)
5. [nvidia-container-toolkit](https://github.com/NVIDIA/nvidia-container-toolkit)
6. [LLaMA-Factory](https://github.com/hiyouga/LLaMA-Factory)
7. [DeepSpeed](https://github.com/microsoft/DeepSpeed)
8. [Megatron-LM](https://github.com/NVIDIA/Megatron-LM)
9. [ollama](https://github.com/jmorganca/ollama)
10. [vLLM](https://github.com/vllm-project/vllm)
11. [Ollama vs VLLM: Which Tool Handles AI Models Better?](https://medium.com/@naman1011/ollama-vs-vllm-which-tool-handles-ai-models-better-a93345b911e6)
12. [Open-WebUI](https://github.com/open-webui/open-webui)

## 版权声明与引用

**声明**：本文所涉及的配置清单、价格与建议仅供技术交流与研究参考。实际购买与部署请结合个人预算和需求进行综合评估。若因参考或采纳文中信息导致任何直接或间接后果，本文作者恕不承担责任。  
**引用**：转载或引用本文内容，请注明原作者与出处。

**Cited as:**

Yue Shui. (Dec 2024). 基于双卡 RTX 4090 搭建家用深度学习主机. https://syhya.github.io/posts/2024-12-21-build-gpu-server

Or

```bibtex @article{syhya2024build, title = “基于双卡 RTX 4090 搭建家用深度学习主机”, author = “Yue Shui”, journal = “syhya.github.io”, year = “2024”, month = “Dec”, url = “https://syhya.github.io/posts/2024-12-21-build-gpu-server/”