IOSG ｜人工智能训练与推理技术栈：从硅到智能



人工智能的迅猛发展基于复杂的基础设施。AI技术栈是一个由硬件和软件构成的分层架构，它是当前AI革命的支柱。在这里，我们将深入分析技术栈的主要层次，并阐述每个层次对AI开发和实施的贡献。最后，我们将反思掌握这些基础知识的重要性，特别是在评估加密货币与AI交叉领域的机会时，比如DePIN（去中心化物理基础设施）项目，例如GPU网络。



## 1.硬件层:硅基础

在最底层是硬件，它为人工智能提供物理计算能力。

CPU（中央处理器）：是计算的基础处理器。它们擅长处理顺序任务，对于通用计算非常重要，包括数据预处理、小规模人工智能任务以及协调其他组件。

GPU（图形处理器）：最初设计用于图形渲染，但因其能够同时执行大量简单计算而成为人工智能的重要组成部分。这种并行处理能力使GPU非常适合训练深度学习模型，没有GPU的发展，现代的GPT模型就无法实现。

AI加速器：专门为人工智能工作负载设计的芯片，它们针对常见的人工智能操作进行了优化，为训练和推理任务提供了高性能和高能效。

FPGA(可编程阵列逻辑)：以其可重编程的特性提供灵活性。它们可以针对特定的人工智能任务进行优化，特别是在需要低延迟的推理场景中。





## 2. 底层软件:中间件

AI技术栈中的这一层至关重要，因为它构建了高级AI框架与底层硬件之间的桥梁。CUDA、ROCm、OneAPI和SNPE等技术加强了高级框架与特定硬件架构之间的联系，实现了性能的优化。

作为NVIDIA的专有软件层，CUDA是该公司在AI硬件市场崛起的基石。NVIDIA的领导地位不仅源于其硬件优势，更体现了其软件和生态系统集成的强大网络效应。

CUDA之所以具有如此大的影响力，是因为它深度融入了AI技术栈，并提供了一整套已成为该领域事实上标准的优化库。这个软件生态构建了一个强大的网络效应：精通CUDA的AI研究人员和开发者在训练过程中将其使用传播到学术界和产业界。

由此产生的良性循环强化了NVIDIA的市场领导地位,因为基于CUDA的工具和库生态系统对AI从业者来说变得越来越不可或缺。

这种软硬件的共生不仅巩固了NVIDIA在AI计算前沿的地位，还赋予了公司显著的定价能力，这在通常商品化的硬件市场中是罕见的。

CUDA的主导地位和其竞争对手的相对默默无闻可以归因于一系列因素,这些因素创造了显著的进入壁垒。NVIDIA在GPU加速计算领域的先发优势使CUDA能够在竞争对手站稳脚跟之前建立起强大的生态系统。尽管AMD和Intel等竞争对手拥有出色的硬件，但他们的软件层缺乏必要的库和工具，并且无法与现有技术栈无缝集成，这就是NVIDIA/CUDA与其他竞争对手之间存在巨大差距的原因。



## 3. 编译器:翻译者

TVM(张量虚拟机)、MLIR(多层中间表示)和PlaidML为跨多种硬件架构优化AI工作负载的挑战提供了不同的解决方案。

TVM源于华盛顿大学的研究，因其能够为各种设备(从高性能GPU到资源受限的边缘设备)优化深度学习模型而迅速获得关注。其优势在于端到端的优化流程，在推理场景中尤为有效。它完全抽象化了底层供应商和硬件的差异，使得推理工作负载能够在不同硬件上无缝运行，无论是NVIDIA设备还是到AMD、Intel等。

然而，在推理之外，情况变得更加复杂。AI训练的硬件可替代计算这一终极目标仍未解决。不过,在这方面有几个值得一提的倡议。

MLIR，Google的项目，采用了更基础的方法。通过为多个抽象级别提供统一的中间表示，它旨在简化整个编译器基础设施，以针对推理和训练用例。

PlaidML，现在由Intel领导，将自己定位为这场竞赛中的黑马。它专注于跨多种硬件架构(包括传统AI加速器之外的架构)的可移植性，展望了AI工作负载在各类计算平台上无缝运行的未来。

如果这些编译器中的任何一个能够很好地集成到技术栈中，不影响模型性能，也不需要开发人员进行任何额外修改，这极可能威胁到CUDA的护城河。然而，目前MLIR和PlaidML还不够成熟，也没有很好地集成到人工智能技术栈中，因此它们目前并不会对CUDA的领导地位有明显威胁。





## 4. 分布式计算:协调者

Ray和Horovod代表了AI领域分布式计算的两种不同方法，每种方法都解决了大规模AI应用中可扩展处理的关键需求。

由UC Berkeley的RISELab开发的Ray是一个通用分布式计算框架。它在灵活性方面表现出色，允许分配机器学习之外的各种类型的工作负载。Ray中基于actor的模型极大简化了Python代码的并行化过程，使其特别适用于强化学习和其他其他需要复杂及多样化工作流程的人工智能任务。

Horovod，最初由Uber设计，专注于深度学习的分布式实现。它为在多个GPU和服务器节点上扩展深度学习训练过程提供了一种简洁而高效的解决方案。Horovod的亮点在于它的用户友好性和对神经网络数据并行训练的优化，这使得它能够与TensorFlow、PyTorch等主流深度学习框架完美融合，让开发人员能够轻松地扩展他们的现有训练代码，而无需进行大量的代码修改。



## 5.结束语:从加密货币角度

与现有AI栈的集成对于旨在构建分布式计算系统的DePin项目至关重要。这种集成确保了与当前AI工作流程和工具的兼容性，降低了采用的门槛。

在加密货币领域，目前的GPU网络，本质上是一个去中心化的GPU租赁平台，这标志着向更复杂的分布式AI基础设施迈出的初步步伐。这些平台更像是Airbnb式的市场，而不是作为分布式云来运作。尽管它们对某些应用有用，但这些平台还不足以支持真正的分布式训练，而这是推进大规模AI开发的关键需求。

像Ray和Horovod这样的当前分布式计算标准，并非为全球分布式网络设计，对于真正工作的去中心化网络，我们需要在这一层上开发另一个框架。一些怀疑论者甚至认为，由于Transformer模型在学习过程中需要密集的通信和全局函数的优化，它们与分布式训练方法不兼容。另一方面，乐观主义者正在尝试提出新的分布式计算框架，这些框架可以很好地与全球分布的硬件配合。Yotta就是试图解决这个问题的初创公司之一。

NeuroMesh更进一步。它以一种特别创新的方式重新设计了机器学习过程。通过使用预测编码网络(PCN)去寻找局部误差最小化的收敛，而不是直接去寻找全局损失函数的最优解，NeuroMesh解决了分布式AI训练的一个根本瓶颈。

这种方法不仅实现了前所未有的并行化，还使在消费级GPU硬件(如RTX 4090)上进行模型训练成为可能，从而使AI训练民主化。具体来说，4090 GPU的计算能力与H100相似，但由于带宽不足，在模型训练过程中它们未被充分利用。由于PCN降低了带宽的重要性，使得利用这些低端GPU成为可能，这可能会带来显著的成本节省和效率提升。

GenSyn，另一家雄心勃勃的加密AI初创公司，以构建一套编译器为目标。Gensyn的编译器允许任何类型的计算硬件无缝用于AI工作负载。打个比方，就像TVM对推理的作用一样，GenSyn正试图为模型训练构建类似的工具。

如果成功,它可以显著扩展去中心化AI计算网络的能力，通过高效利用各种硬件来处理更复杂和多样的AI任务。这个雄心勃勃的愿景，虽然由于跨多样化硬件架构优化的复杂性和高技术风险而具有挑战性，但如果他们能够执行这一愿景，克服诸如保持异构系统性能等障碍，这项技术可能会削弱CUDA和NVIDIA的护城河。

关于推理：Hyperbolic的方法，将可验证推理与异构计算资源的去中心化网络相结合，体现了相对务实的策略。通过利用TVM等编译器标准，Hyperbolic可以利用广泛的硬件配置，同时保持性能和可靠性。它可以聚合来自多个供应商的芯片(从NVIDIA到AMD、Intel等)，包括消费级硬件和高性能硬件。

这些在加密AI交叉领域的发展预示着一个未来，AI计算可能变得更加分布式、高效和可访问。这些项目的成功不仅取决于它们的技术优势，还取决于它们与现有AI工作流程无缝集成的能力，以及解决AI从业者和企业实际关切的能力。