人工智能数据中心：
扩大规模与拓展规模

执行摘要

本文件深入分析了现代人工智能数据中心实现规模扩大和扩展的技术基础。通过突出关键行业发展和高级扩展技术的持续演变，作者旨在强调在扩展的关键方面，包括人工智能硬件创新、模块化基础设施规划和复杂冷却方法等方面，需要行业范围内的协作方法。

您将了解到：

01

人工智能与机器学习简介

获得人工智能（AI）、机器学习（ML）和大型语言模型（LLMs）的全面概述。本节解释了基本概念，包括模型、训练和推理。

02

人工智能自2017年以来的演变

探索重要的AI里程碑，例如变革性的Transformer模型的出现。本节还强调了朝着更大规模的AI模型和增强计算能力的发展趋势。

03

扩大人工智能基础设施规模

探讨现代AI数据中心扩张策略，关注扩张规模和扩展性。

04

人工智能硬件的进步

了解半导体技术、芯片模块和封装技术的创新。深入了解高速网络和先进冷却系统。

05

人工智能的未来趋势

探讨新兴趋势，例如分段模型、频率降低的同步和扩展分布式系统。本节还讨论了数据中心互联（DCI）对中长距离连接日益增长的需求。

作者：

艾伦·凯泽，高级技术顾问，AFL

本·阿瑟顿 技术作者，澳大利亚足球联赛（AFL）

生成式人工智能已出现并迅速发展，其规模和速度出乎我们意料。它建立在多种技术平台之上，所有这些技术也在规模和速度上取得了进步。这是一个非凡的技术故事。

人工智能（AI）、机器学习（ML）和大型语言模型（LLMs）简介

人工智能（AI）是指那些旨在执行通常需要人类智能的任务的机器或软件（例如，理解自然语言、视觉感知、语音识别、语言翻译、学习和问题解决）。

机器学习（ML）训练算法以推断意义并提供对独特提示的准确类似人类的响应。深度学习（DL）是无需人类干预的机器学习。深度学习使用称为人工神经网络（ANN）的算法，这些算法在多个阶段处理输入刺激，并能识别复杂数据集中的关系。大型语言模型（LLM）是专门处理语言的深度学习模型。

DL算法可以处理任何其元素之间存在关系的数字化信息。例如，LLMs可以生成针对查询或提示的人类语言响应（例如，GPT-4），并且也可以在图像和编码等某些非语言领域工作。

基本概念：模型、训练和推理

机器学习过程旨在开发能够进行准确推理、做出逻辑决策并展现类人智能的模型。训练阶段涉及准备选定数据并优化响应以实现最佳性能。在推理阶段，训练好的模型分析新数据，应用优化后的模式识别，并自动生成逻辑响应。本节提供了对模型、训练和推理的基础理解。

模型

在机器学习（ML）的语境下，一个模型代表了一组被训练以识别模式和预测新、未见数据中一致关系的算法。例如，模型用途可能包括预测天气、识别图像以及根据用户行为提供高度个性化的电子商务体验。

以下模型代表早期模型类型（欲了解更近期的模型，请参阅下一节，标题为“自2017年以来的人工智能演变”）：

监督学习模型

监督学习模型从经过批准的示例中学习。例如，一个监督学习模型可以根据包含相同对象相似表示的图像，视觉上识别出该对象。

非监督学习模型

无监督学习模型在未标记数据中寻找隐藏的模式或分组。技术包括聚类（将相似数据点分组）和降低数据复杂性（简化数据以实现更有效的分析）。

强化学习模型

强化学习模型通过与环境的互动来学习——反馈被登记为奖励或惩罚。这类模型在游戏和机器人等领域得到应用。

培训

训练阶段教会模型进行准确的预测。此阶段包括在要求模型做出预测之前，将模型暴露于一个预定的数据集。随后进行参数调整，有助于最小化错误。

训练模型涉及多个阶段：

数据收集

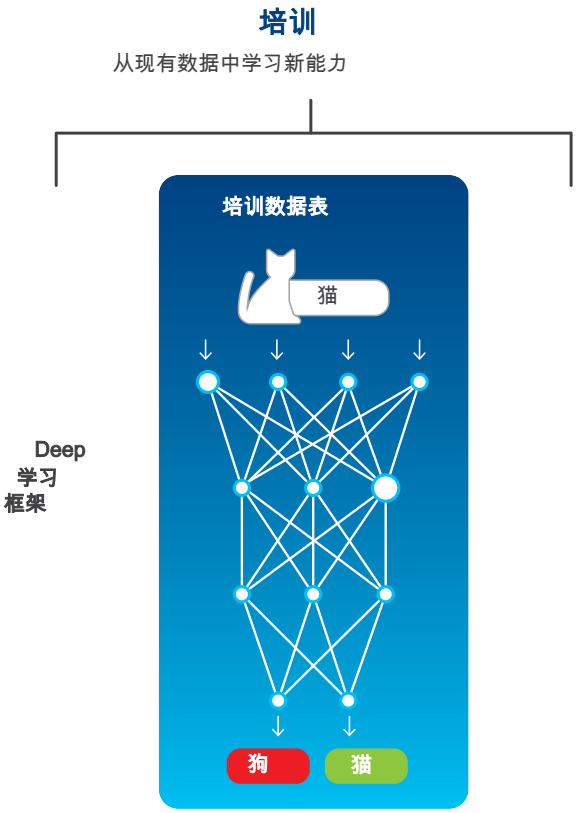
利益相关者，如机器学习工程师和数据科学家，必须收集和整理与模型最终目的相关的大量且多样化的数据集。

数据预处理

数据预处理涉及对数据进行清理和转换，以便为训练做准备。这可能包括识别数据间隙和填充缺失值，对相似变量进行归一化以提高学习效率，以及将数据分为训练集（用于训练模型）和验证集（用于评估模型性能）。

模型选择

模型选择包括多个考量因素。例如，大型、标注数据集可能适合神经网络。在多个模型之间进行迭代测试可以帮助评估和确定最佳匹配模型以达到最佳性能。



优化

首先，机器学习工程师和数据科学家必须定义一个损失函数——一个衡量预测准确性的性能指标。接下来，算法会反复调整和更新模型的参数以实现收敛——即使经过多次参数调整，损失函数也不再显著改善。

评估

最终阶段可能包括多个阶段。例如，工程师可以使用验证数据来获取无偏的性能评估。这意味着要检查过度拟合（训练数据表现良好，验证数据表现不佳）和欠拟合（整体表现不佳）。常见的解决方案包括正则化（即，对过度镜像训练数据添加惩罚）、dropout（即，随机移除数据子集以防止过度依赖）和交叉验证（即，分割并交叉引用训练和验证数据集，再次以防止过度依赖）。

推断

推断阶段需要训练好的模型基于新的、未见过的数据进行预测。这是在模型可以自信地应用于现实世界之前，训练和验证的最后阶段。

推理过程包括：

输入处理

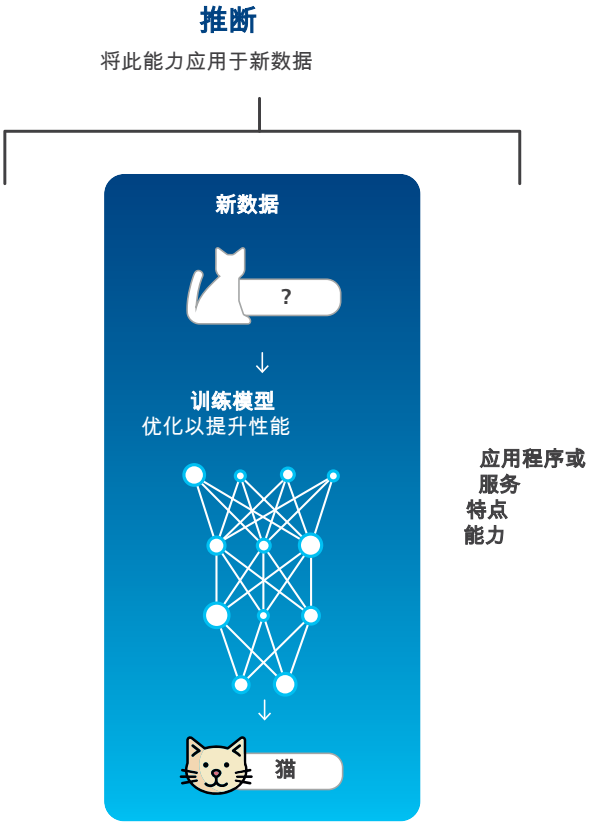
这一步骤的目的是通过向模型中输入与训练数据准备阶段所使用的方法相匹配的数据，以提高预测准确度。例如，采用相同的方法处理缺失值（例如，用众数或中位数填充缺失值）可以确保有效的模型输入处理，从而提高一致性。

预测

在此阶段，该模型必须处理新的、未见过的数据（无论是单个数据点还是数据批次）并在实时内做出预测或提供有价值的见解。

后期处理

后处理阶段包括几个不同的阶段，用于精炼输出。例如，结合多个预测以创建一个过滤的、更准确的最终预测。



人工智能自2017年以来的演变

2017年标志着深度学习（DL）领域多个重要里程碑的实现——特别是在语音识别和图像识别方面。大规模标注数据集的日益可用性与行业在机器学习（ML）训练技术方面的进步相结合，使得模型学习效率得到了提升。

例如，卷积神经网络（CNNs – 专注于图像识别的图案检测和分类的特殊深度学习算法）和循环神经网络（RNNs – 利用先前输入的记忆来为顺序数据，如自然语言，提供当前预测的深度学习模型）的改进，使得从2012年的15.3%的错误率有所降低。[1] 2017年仅2.3%[2] - 更低的错误率转化为更少的预测错误，从而有益于结果。

同样，深度学习模型在语音识别领域的性能得到提升，微软研究人员在2017年实现了5.1%的对话语音里程碑。[3] 人类 parity 词错误率。这些前所未有的新水平强调了人工智能在各个应用中的变革性和进步潜力。

引言：谷歌Transformer模型简介

2017年出版，“注意力即所需”[4] (Vaswani et al) 介绍了Transformer模型，这可能是当时新兴人工智能技术领域的突出突破性进展。由谷歌的八位科学家呈现的这篇论文概述了一种新的深度学习架构方法——基于2014年的注意力机制（Bahdanau et al）。自从那时起，Transformer模型已经成为LLM架构设计的基石，为现代人工智能应用如谷歌的Bard——于2024年2月8日更名为Gemini——的发展铺平了道路。

变压器模型在革命性发展后将焦点完全转移到注意力机制上，消除了行业对循环神经网络（RNNs）和卷积神经网络（CNNs）的依赖。其即时影响包括自然语言处理（NLP）方面的重大进步。

变压器架构

然而，与RNNs和LSTMs按顺序处理数据流不同，Transformers同时处理有序标记（即同时管理输入数据的所有部分，而不是一个接一个），从而提高了并行化和效率，以加速训练和推理过程中的开发周期。

Transformer模型采用自注意力机制，使序列标记相对于彼此进行分析，从而提高训练速度（此方法还使模型能够捕捉数据中的长距离依赖关系）。

自2017年以来，Transformer架构已成为支持自然语言编程（NLP）领域现代最先进进展的主要、基础模型类型——例如BERT和GPT。这一AI领域的关键时刻为AI技术的广泛快速采用及其随后的大规模投资奠定了基础。

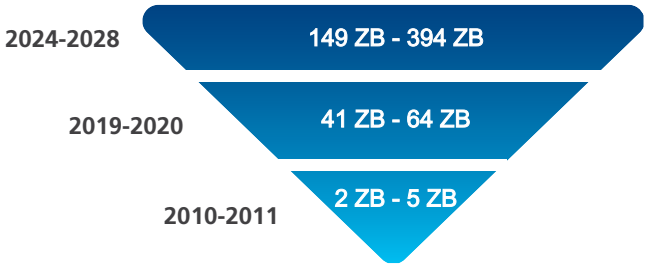
1. PyTorch
2. Statista
3. Microsoft
4. arXiv

趋势：转向更大的AI模型

随着时间的推移，模型参数数量的显著增加（从早期拥有数百万参数的模型转变为现代拥有数十亿甚至数万亿参数的模型）标志着向更大规模AI模型的趋势。例如，2019年发布的OpenAI的GPT-2拥有15亿参数。 [5] 参数。然而，一年后，2020年的GPT-3模型远远超过了这个数字，达到了1750亿参数。到2023年，GPT-4的发布标志着模型参数的又一次里程碑式增长，估计有1800亿参数。

日益增长的训练数据量和复杂人工智能过程所需的计算能力与人工智能模型规模的行业增长相平行——OpenAI报告称，自2012年以来，管理人工智能模型所需的计算能力每3.4个月翻一番 [6] 同样，最近全球数据量的激增促使大量数据集用于训练得到广泛应用。

在2010年至2011年期间，全球互联网上创建的总新数据从2泽字节增长到5泽字节。 [7] 2019年至2020年期间，新增数据的总量从41泽字节增长至64泽字节。 [7] 预测显示，2024年至2028年间，该数字将从149泽字节增长到394泽字节。 [7] .



这些数据突出了在人工智能发展前景中存在的巨大潜力和显著的挑战——例如存储和可扩展网络解决方案等方面——的前景。

市场规模与投资

在2023年，全球人工智能市场价值达到约1966.3亿美元。 [8] ，预计在2024年至2030年间的复合年增长率（CAGR）为36.6%。推动这一增长的因素包括持续的研究和创新，尤其是在深度学习（DL）、自然语言处理（NLP）和生成式人工智能（generative AI）领域，所谓的“大科技公司”包括微软、Meta和亚马逊预计将投资1万亿美元。 [9] 在未来几年内，人工智能领域。

学术贡献与产业贡献对比

学术界和人工智能产业对AI模型开发都做出了显著贡献，每个领域都发挥着独特、独立而又相互补充的作用。

历史上，由学术界驱动的AI基础研究专注于理论进步和长期创新。然而，对AI技术的需求巨大，导致产业而非学术界做出了回应，模型规模扩大了29倍。 [10] 相较于学术界的模型数量。在2023年，行业产生的模型数量达到了51 [11] 与学术界支持的模型数量相比，后者达到了15（学术界与工业界的合作也导致了21个模型的出现）。 [11] 显著的模型。大量的金融资源和计算硬件的可用性意味着私营公司可以加速人工智能的进步。

5. OpenAI 6. MIT Technology Review 7. Statista 8. Grand View Research 9. Investopedia 10. MIT Sloan 11. 人工智能指数报告2024，斯坦福 12. Techopedia 13. 国际贸易管理局

全球格局

多数关键国家在人工智能研究和开发领域处于世界领先地位。美国位居前列，拥有2490亿美元的投资。 [12] 在私人融资方面，并且几乎60%。 [12] 美国顶尖人工智能研究人员在大学和公司工作。

中国紧随其后，筹集了950亿美元。 [12] 在私人投资中并贡献了11% [12] 为顶级人工智能研究人员。

英国排名第三，占市场份额210亿美元 [13] 人工智能市场估值预计到2035年将达到1万亿美元。其他显著贡献者包括以色列、加拿大和法国，每个国家都培养着充满活力的初创企业生态系统。这些进步不仅加速了人工智能技术的发展，而且塑造了全球人工智能政策和伦理标准。

大模型与更多训练的影响

人工智能的快速发展被一个持续的趋势所标记：模型规模和复杂度的指数级增长。现代人工智能模型，如大型语言模型（LLMs）和推荐系统，现在包含数百亿到数万亿的参数。这种增长对训练和部署这些模型所需的计算基础设施产生了深远的影响。在过去十年中，训练前沿模型所需的计算每三到四个月就会翻一番。

训练大型人工智能模型所需的计算

年份 模型 近似训练计算（FLOPs*）	
2012 AlexNet 4.7 x 10	17
2014 VGG16 1.2 x 10	19
2015 ResNet-152 1 x 10	19
2016 AlphaGo Zero 3.4 x 10	23
2018 BERT Large 2.8 x 10	20
2019 GPT-2 1.9 x 10	21
2020 GPT-3 3.1 x 10	23
2021 Megatron-Turing NLG 1.17 x 10	24
2022 PaLM 2.5 x 10	24
2023 GPT-4 (1+T est.) 2.1 x 10	25

前缀缩写指数计算机性能存储容量			
吉伽- G 10	9	吉弗洛普斯（GLOPs）吉字节（GB）	
太拉- T 10	12	teraflops (TFLOPs) 兆字节 (TB)	
百万亿（petabyte）P 10	15	petaFLOPS (PFLOPs) petabyte (PB)	
exa- E 10	18	exaFLOPS (EFLOPs) exabyte (EB)	
zetta- Z 10	21	zettaFLOPS (ZFLOPs) zettabyte (ZB)	
yotta- Y 10	24	yottaFLOPS (YFLOPs) yottabyte (YB)	

浮点运算（FLOP）- 计算工作量的一种度量，用于需要大量浮点计算的领域（例如，机器学习）。

计算大型模型的计算需求

> 我们估计，在一个月内训练一个万亿参数模型需要70,000个与NVIDIA H100等效的加速器。
>
> 艾伦·凯泽 - 高级技术顾问，AFL

高级处理器

模型参数的增加对并行计算系统提出了挑战，需要更多更强大的处理器来处理密集计算。由于传统CPU的通用设计和有限的并行性，它们无法完成这项任务。因此，已经出现了向专用硬件的显著转变：

图形处理单元（GPUs）

人工智能优化的GPU提供巨大的并行性（即同时管理多个计算或处理过程），这使得它们非常适合训练大型神经网络。像NVIDIA和AMD这样的公司已经开发了专门针对人工智能工作负载优化的GPU，例如NVIDIA H100和AMD MI325X设备。

应用专用集成电路（ASICs）

应用特定集成电路（ASICs）针对特定的人工智能计算任务进行定制。

张量处理单元（TPUs）

由谷歌研发，TPU（张量处理器）是针对机器学习任务设计的定制型ASIC（专用集成电路），提供高吞吐量和能效。

晶圆级引擎

多个处理器及其相关内存和网络的集成可以制造为单晶圆系统。例如，包括Cerebras WSE-2和Tesla Dojo。

GPU、TPU和ASIC模块，辅以支持AI优化的组件，被称为加速器。一个或多个加速器与一个或多个CPU组合形成节点。CPU用于在外部网络上进行通信、准备训练数据、管理训练过程以及执行一般维护任务，而加速器则执行核心AI模型计算。

更大、更快的网络

训练大型模型需要在众多处理器和节点间分配大量工作负载。这种分配需要高速网络以确保高效的通信和同步。

高带宽互连

技术如NVIDIA的NVLink、AWS的NeuraLink和PCIe在节点内GPU之间以及局部节点组（Pod）之间提供高速连接。InfiniBand和RDMA增强型以太网将大量加速器连接起来，形成大型集群。

低延迟网络协议

针对最小延迟优化的协议对于同步训练方法至关重要，在这些方法中，及时的数据交换是关键。

无损耗协议保持数据包顺序

通过在以太网协议中采用RDMA，消除了以太网的固有数据包丢失问题。

电力消耗和热量产生所带来的挑战

功耗

训练大型人工智能模型所需的能量是巨大的。

能源成本

训练单个大型模型可能消耗数百万兆瓦时的电能，从而产生显著的运营成本。对于我们的百亿参数模型，训练集群和支持设备将需要大约84兆瓦，每进行一次训练运行将消耗约66兆瓦时。

环境影响

高能耗导致碳排放增加，引发了关于人工智能发展可持续性的担忧。不出所料，几乎所有大规模人工智能计算组织都实施了重要的可再生能源和其他缓解方案。核能的进步 [15] 可能提供多个优势，包括不间断的稳定产出、可靠性、可扩展性和低碳排放。

发热

电力消耗几乎全部转化为计算系统中的热量。

热管理

高容量冷却解决方案是必要的，以从加速器、节点和开关中移除热量。为此，以当前的人工智能硬件为例，每个比比萨盒大小的1RU机架可能产生高达4,000瓦的功率，而每个机架的功率可能高达120,000瓦。

冷却技术

为了有效散热，高端AI赋能数据中心必须采用先进的冷却方法，例如直接液体冷却和浸没冷却。

扩大基础设施规模

数据中心基础设施必须随着人工智能计算的增长而扩大。

数据中心扩展

需要更大规模的基础设施以容纳更多服务器，实现紧密、低延迟的连接，并提供所需的电力和冷却能力。

边缘计算

将计算分配得更靠近终端用户可以降低延迟，并为对数据主权敏感的应用提供更高的安全性。

云解决方案

利用基于云的AI平台，无需大量前期投资即可获取可扩展资源，将成为大多数用户的实际解决方案。

15. 分析洞察

如何竞争格局驱动创新

全球对人工智能技术的兴趣推动了主要玩家和初创企业之间的激烈竞争。持续的竞争环境见证了不断增加的投资为突破性的、变革性的跨行业创新铺平道路。这一追求创新的过程加速了硬件设计、能源效率和冷却方法的技术进步，同时将简化的可扩展性解决方案融入现代人工智能系统的架构之中。

主要玩家和风险投资家投资

新兴人工智能领域的投资激增，资金来源包括传统科技巨头和风险投资家。例如，谷歌、微软和亚马逊等公司已向人工智能初创企业投入数十亿美元——从2021年的110亿美元增长到2023年的近三倍。 [16] 在2023年，仅微软一家就投资了100亿美元。 [17] 在OpenAI中，尽管人工智能初创公司Anthropic据报道已获得谷歌2亿英镑的投资。 [18] 40亿英镑来自亚马逊 [19] .

在2023年，500亿美元 [20] 在风险投资（VC）公司关键投资推动下的人工智能创新，其中超过70轮融资每轮估值均超过1亿美元。值得注意的贡献型VC公司包括红杉资本、Khosla Ventures和安德森·霍洛维茨。 [21] 投资者面临着强烈竞争，以识别并支持下一个重要的AI突破，这推动了对AI领域的持续且不断增长的资本流入。

投资如何驱动人工智能创新

来自大型公司和风险投资公司的投资促使了具有改进计算能力的先进人工智能模型的高速开发。大型语言模型（LLMs）的发展代表此类创新的关键领域之一。

OpenAI的GPT-4和Google的PaLM 2都显著推进了自然语言处理（NLP）领域，使得针对提示的文本生成更加准确和具有情境意识。例如，GPT-4的高级功能展示了40%的... [22] 相较于OpenAI先前的GPT-3.5模型，PaLM 2返回事实性响应的可能性有所增加。同样地，PaLM 2已进化为使用五倍多的文本数据。 [23] 比PaLM（尽管参数数量减少，使得PaLM 2比PaLM更轻、更高效）。

这些模型的可获取性使得先进的人工智能能力变得普及，使各种规模的企业能够利用人工智能进行各种相关应用（例如，客户服务、内容创作）。

竞争格局培育了一种快速迭代和系统改进的文化，其中公司持续优化AI模型以保持竞争力。这一创新的持续循环与投资创造了先进AI技术与新投资之间的良性协同效应。

16. Livy AI 17. Forbes 18. CNBC 19. Amazon 20. CCN 21. CGAA 22. OpenAI 23. CNBC

数十个大型语言模型公开可用。

人工智能开发者竞赛见证了大量大型公开可用的大型语言模型（LLM）的涌现。著名的LLM包括：

模型描述 注释		
深度学习推荐 Meta AI 的模型（DLRM）	利用深度学习进行大规模 缩放推荐任务。	集成了用于优化的功能 高效处理稀疏数据。
基于变换器的 推荐模型	适配 Transformer 架构至 推荐系统。	具备建模长远范围的能力。 个性化所依赖的。
YouTube的深度神经网络 YouTube用的网络 建议	个性化内容 YouTube上的推荐	可扩展的架构结合 候选人生成及排序。
阿里巴巴的AIRec	AI驱动的推荐引擎 用于电子商务应用。	实时处理，多模态 数据集成、个性化。
亚马逊的个性化 推荐系统	高级算法定制产品 建议针对个人用户。	使用行为数据和基于内容的 滤波和协作滤波。
深度兴趣网络（DIN） 深度兴趣演变 网络（DIEN）	DIN 和 DIEN 专注于建模 用户兴趣。	捕捉用户行为序列 并且模型动态利益。
Pinterest的Pixie	实时推荐引擎 针对低延迟推荐。	利用基于图的算法方法。 可扩展至大型用户群。
Netflix的推荐 算法	采用复杂算法以 推荐电影和电视剧。	协作过滤、行为 建模和内容分析。
-wide & deep learning (Google) 关于推荐	记忆化（大模型）与 泛化（深度模型）。	平衡已知关联 随着发现新的模式。
深度检索（Meta AI）	端到端的检索和排序 推荐系统。	整合检索到深度 学习相关性模型。
Meta的LLaMA系列	开源的 LLaMA 3.1 版本拥有 405 数十亿参数。	免费使用 – 民主化接入 向高级人工智能能力发展。
迷雾AI模型	创新架构，分组 查询注意力，和开放权重。	开放式权重替代方案 专有模型。
Databricks 公司的 DBRX	开源语言模型，拥有1320亿参数 参数。	开源多专家混合 变压器模型。
EleutherAI 的 GPT 模型	开源替代方案 专有模型。	在民主化进程中起到关键作用 访问先进的AI功能。
阿布扎比制造的“猎鹰” 技术创新研究院	成为领先的开源 模型。	集成到各种应用中 例如，阿联酋卫生系统。

来自风险投资公司和大型公司的投资为初创公司和成熟企业提供了扩大运营和推动新兴人工智能技术边界的必要资源。随着竞争和投资的持续加剧，该行业预计将看到更多突破性的AI创新。下一节将探讨现代人工智能模型的发展。

现代人工智能模型进展

机器学习模型演变及训练要求

过去十年间，人工智能经历了显著转型，得益于模型架构的进步和模型尺寸的增加，导致了不断攀升的培训需求。如今的生成式人工智能展示出的能力，将令十年初的研究者感到惊讶。

时间段 模型类型 模型大小 训练需求			
2010年代初 传统机器学习 并且，黎明的曙光 深度学习	- **传统算法：** 支持向量机（SVMs） 决策树，随机森林， K-近邻算法 - **早期神经网络：** 浅 在有限的能力范围内	- 小型模型，具有少量 千参数 受硬件限制和 数据可用性	- 在标准CPU上的培训 - 小型数据集如MNIST （60,000 张图片）
2012–2014 深度崛起 学习	- **卷积神经网络** **（卷积神经网络）**：AlexNet在 2012 - **循环神经网络** **(循环神经网络)**：引入，但 存在培训问题	大约 6000 万 参数 - 模型变得更加深入且 更复杂	- 采用GPU加速 - 使用更大的数据集如 ImageNet（百万张图片）
2014–2016 建筑 创新	- **高级架构：** VGGNet, GoogLeNet (Inception), ResNet - **改进的RNNs：** LSTMs和GRUs 解决的训练问题。 - **生成模型：** 生成对抗网络（GANs） 于2014年推出	- **VGGNet:**高达1.38亿 参数 - **ResNet:** 超过100层 - 优化	- 数据增强 技术 - 分布式训练跨 多个GPU和机器
2017 《变形金刚》 革命	- **变形金刚：** 由...引入 注意即你所需。 - **关注机制：** 模型 关注特定的输入部件， 消除重复需要	- 数亿个 参数	- 序列并行化 较长的上下文 显著GPU资源 必需的
2018–2020 扩大规模 与预训练的 语言 模型	- **BERT (2018):**启用上下文- 意识与理解 - **GPT系列：** 推动了边界 语言生成	- **BERT Large:**~3.4亿 参数 - **GPT-2 (2019):**十五亿 参数 - **GPT-3 (2020):**一千七百五十亿 参数	- 在大量数据集上进行训练 类似于Common Crawl 高性能计算 聚类；数周培训/ 月份
2021–至今 出现 极具规模 模型	- **多模态模型：** 合并 文本和图像（CLIP, DALL·E） - **大型语言模型：** GPT-4 并且其他 - **高效模型：** 稀疏注意力， 边缘优化架构	- **GPT-4 (2023):**估计 万亿参数 - **巨兽-图灵自然语言生成** **(2021):**5300亿 参数	- 超级计算机资源 （TPU pods，大型GPU集群） - 优化训练技术 - 能源消耗问题

24. NVIDIA开发者 25. GPT-3 演示

24 25
开发者与GPT-3演示。
信息来源于公开可获得的在线资源，包括NVIDIA。

关键驱动因素推动演变

01 硬件进步

GPU 和 TPU： 专用处理器在并行计算领域变得可用，性能不断提高，从而实现了更大规模模型的使用以及在大数据集上的训练。

内存改进： 更大的内存容量，以适应大型模型。在加速器中嵌入的HBM内存。

02 算法创新

优化技术： Adam优化器，学习率调度程序和正则化方法。

高效架构： 稀疏模型、知识蒸馏以及模型压缩技术。

03 数据可用性

开放数据集： 大规模数据集在训练和基准测试中的广泛应用。

网络抓取： 利用互联网规模数据对语言模型进行训练。

04 社区和开源贡献

框架： TensorFlow、PyTorch以及其他简化模型开发的库的发展。

协作研究： 共享预训练模型和研究成果加速了进步。

过去十年见证了人工智能模型从相对简单的算法发展到具有前所未有的复杂架构。模型类型已转向深度学习，其中变换器已成为自然语言处理中的事实标准，并在其他领域取得进展。模型规模呈指数增长，这得益于对性能提升的追求，导致模型拥有数百亿个参数甚至更多。这种增长得益于计算硬件的进步，但同时也带来了成本、可访问性和可持续性方面的挑战。

培训需求的升级突显了更加高效算法和硬件的需求。随着人工智能的持续进步，平衡性能与伦理和环境考量将至关重要。下一个十年可能会专注于优化这些大型模型，使其更加高效、公平且易于获取，确保人工智能的益处能广泛地惠及社会。

硬件创新：芯片、系统和封装

半导体技术的进步是提升人工智能能力的关键。芯片设计、系统架构和封装技术的创新使得开发出支持现代人工智能工作负载的高性能硬件成为可能。处理和网络方面的进步高度依赖于半导体技术的演进。人工智能的需求及其产生的财务流强烈推动了半导体技术的发展。

半导体发展可以描述为节点进化的过程。在此背景下，节点是指晶圆制造的一整套完整工艺，这是一个涉及数百个单独步骤的复杂过程。节点以基于其微电子元件最小特征尺寸的简称来表征。

工艺节点转换

工艺节点

节点名称指的是芯片上最小特征的尺寸，以纳米（nm）或埃（Å）为单位。埃是纳米的十分之一。

N5（5纳米）

广泛应用于当前一代处理器。

N3（3纳米）

下一代技术提供改进的性能和效率。预计于2024年从几家先进的晶圆厂发货。

18Å（1.8纳米）

代表着微型化技术的下一个目标，首批产品预计将在2025年发货。

14Å（1.4纳米）

节点位于18Å之外。开发工作已经开始。

作者评论

摩尔定律依然有效，尽管有些传言相反，但每一步都变得越来越困难且成本更高。

人工智能的启示

性能提升

较小的晶体管使得处理器更快且更高效。每片芯片上的晶体管数量增加，扩展了处理器和开关芯片的功能。

能源效率

每操作次的功耗降低。注意，尽管每操作次的功耗下降，但功能密度的增加可能会导致每芯片的总功耗需求增加。

制造业挑战

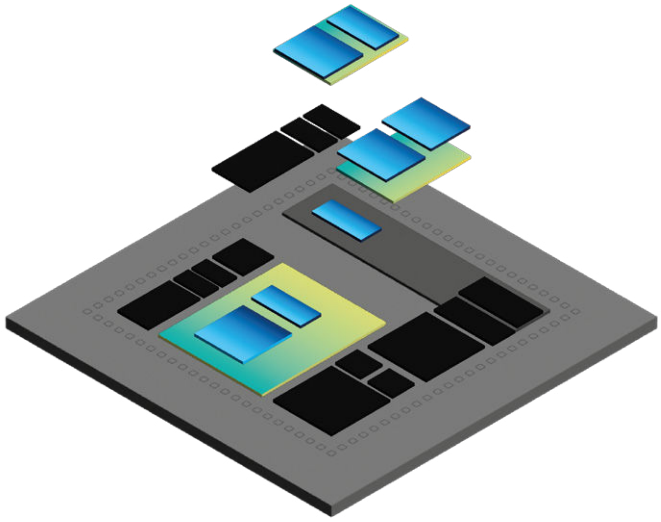
先进的光刻技术是必需的，这增加了复杂性和成本。据称，ASML最新的极紫外光工具每套成本高达3.5亿美元。额外的新的工艺需求包括深蚀刻、晶圆和芯片级封装，以及更加关键的洁净室设备。到2024年，一个先进且经济规模化的晶圆厂可能成本超过250亿美元。

芯片片

芯片拼图是未封装的、模块化集成电路，可以组合成更大的单一系统单元，例如处理器或交换机。芯片拼图可以专门化，例如用于存储和电光功能。

人工智能领域的优势与应用	
可扩展性 允许混搭不同的芯片粒 并且芯片技术，以创造更广泛的和 功能性解决方案。 **收益率提升** 个体芯片的制造收益率更高。 预先测试，导致系统产出率更高 降低成本。	**领先的加速器制造 利用芯片组件** 将具有不同功能的芯片片相结合 （例如，CPU、GPU、内存）增强性能。 **互连标准** 通用芯片片互连表达式：UCIe是一种 对开模互连的开放式规范 芯片间的串行总线。

包装技术



先进封装技术正在开发中，这些技术能够实现将越来越复杂的系统组装成单个微电子封装。使用极紫外光刻（EUV - 在半导体制造中，EUV技术使用光在硅晶圆上创建精确图案，比传统光刻方法具有更高的分辨率）制造的最大的单个芯片大约为800平方毫米。 [26] 并且可能拥有1500亿到2000亿个晶体管。为了实现更高的复杂性，需要集成两个或更多个单个晶圆。

由台湾半导体制造公司（TSMC）制造的NVIDIA Blackwell芯片包含2080亿个晶体管 [27] 。巴克威尔（Blackwell）的尺寸超过了EUV光刻掩模的限制，并且是通过在共享的中间基板上制造两个芯片片（chiplets）来制备的。

台积电（SoIC、CoWoS、InFO）、英特尔（EMIB、Foveros）以及其他公司，已开发了在单一高密度封装中组装和互连多个复杂芯片的技术。领先的加速器在很大程度上依赖于这种先进封装。

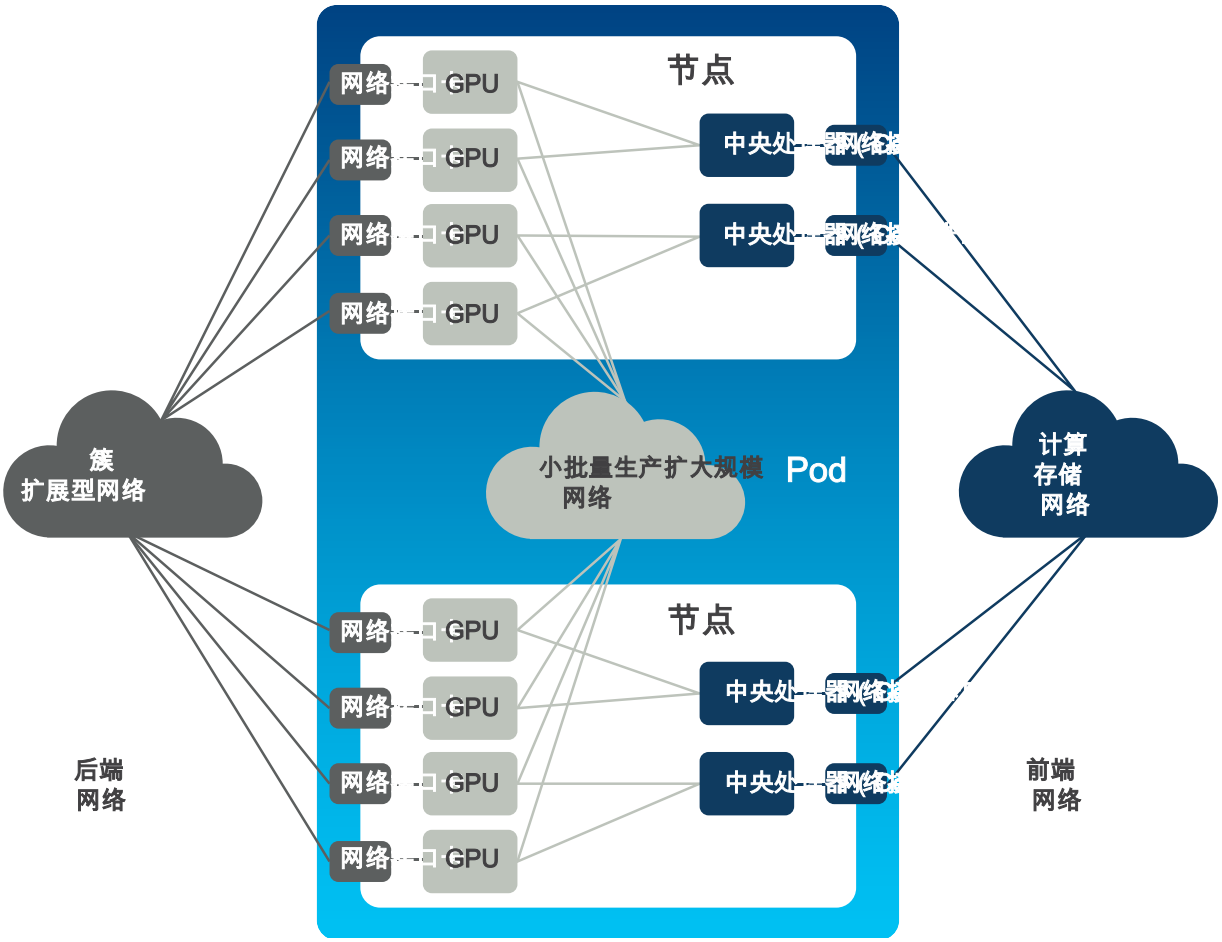
在提供电力、冷却以及连接这些设备方面仍存在挑战。今天的先进加速器消耗1200瓦或更多。按照目前的轨迹，我们预计四年内将出现4000瓦的设备。该领域正在快速发展。

2026年 AWS re:Invent 2024 (YouTube)
27. NVIDIA

扩展人工智能：规模化提升和扩展策略

人工智能的进步需要制定强大的策略来有效扩展AI基础设施。扩展AI涉及两种主要方法：垂直扩展和水平扩展。每种方法都旨在优化性能，高效管理资源，并满足AI应用不断增长的计算需求。最近的研究表明，对AI基础设施的投资正在激增，预计到2025年全球投资额将达到2000亿美元。[28]

28. 高盛（Goldman Sachs）



在人工智能数据中心中进行规模扩大和规模扩展

节点代表人工智能计算集群的基本构建块。一个节点由一个或多个专用处理器组成，这些处理器统称为加速器（GPU、TPU和定制ASIC），以及一个或多个通用处理器（CPU）。这些加速器针对矩阵乘法和加法等任务进行了优化，这些任务是CNN模型处理的核心，但在通用计算功能上并不有效。在节点中将GPU和CPU与共享内存结合使用，允许GPU在模型计算上以最高效率运行，同时CPU负责通信、数据处理、监控和日常管理。

通过将附近节点中的GPU通过非常高速、低延迟的连接连接起来，可以形成更强大的AI计算单元。这个GPU“舱”可以同步共享训练和推理模型参数及中间结果，同时降低网络设备和能源成本，提高集群性能。连接GPU内部的网络被称为“升级”网络（见图示）。GPU的直接连接是最优的，但这限制了网络的跨度至几米，通常是一到两个机架。NVIDIA的GB200 NVL 72就是一个升级系统的例子，它在一到两个机架中利用了72个GB200加速器。

架构超过10万加速器的超大型系统

非常庞大的AI系统架构（如具有10,000个以上加速器的系统）管理着巨大的计算工作负载，需要结合CPU、GPU和专业的AI加速器。在大规模AI机器学习集群中连接所有加速器是“扩展”网络的功能。

xAI 巨型超级计算机 [29] 在孟菲斯，美国建造的该系统是一个价值数十亿美元的超大型系统实例，由10万个NVIDIA H100 GPU组成。Supermicro液冷机架构成了Colossus架构的基础。每个机架包含八个4U服务器，每个服务器容纳八个NVIDIA H100 GPU，每个机架总计64个GPU。一个完整的GPU计算机架由八个GPU服务器、一个Supermicro冷却剂分配单元（CDU）和必要的相关硬件组成。请注意，每个节点配备八个GPU且没有公开披露的Pod级网络，这个系统代表了一个有限的扩展系统——从工程角度来看，这可能是权宜之计而非优雅设计的例子。

29. 服侍之家

同步人工智能训练中的网络连接

人工智能机器学习训练使用并行计算方法将大型模型和大型数据集分割成更小的作业，这些作业可以通过具有节点和Pod支持的个别加速器来管理。每个工作加速器处理训练任务的微小部分，更新模型参数，并将更新与所有其他工作器共享。所有工作器必须完成更新，才能由任何工作器启动下一个训练任务。人工智能机器学习训练的同步性质对后端（扩展）网络提出了关键要求。

后端网络要求

高带宽

高带宽网络在同步人工智能模型训练中扮演着至关重要的角色。后端网络流量可能间歇且快速变化。每个加速器在每轮训练阶段之间必须交换数以吉字节计的训练数据和模型参数。后端网络设计者会选择当前新建项目中可用的最高带宽交换机和收发器，每个端口运行400G和800G。

低延迟

飞行时间（即数据包在各个点之间传输所需的时间）在后台网络中对总训练时间有所贡献。尽可能缩短电缆长度——通常少于200米——可以增加训练吞吐量。交换机和FEC组件必须进行延迟优化。

无损

由于网络冲突和其他故障导致的丢包在大多数以太网应用中是可以接受的，因为重传只会减缓单个过程或系统。对于同步训练集群，整个集群必须等待所有工作者完成，重传的代价过高，无法接受。成功的大规模人工智能机器学习集群需要网络监控和优化。

保持数据包顺序

基于GPU的加速器不是高效的通用处理器，作为端点设备，它们无法有效地识别传输故障和重新排序数据包。

处理器负载低 – RDMA

网络操作必须能够直接在内存之间移动数据，无需CPU或加速器的干预。这需要远程直接内存访问（RDMA）。RDMA是InfiniBand的本地特性，但对于以太网网络，则需要RDMA over converged Ethernet（ROCE）扩展。

高可靠性

长训练运行在非常大的集群中可能会遇到定期设备故障。由于集群是一个同步运行的大型分布式系统，单个单元的故障无法通过冗余实时解决。缓解措施是通过整个集群监控和周期性检查点实现的。

当检测到加速器、节点或Pod故障时，整个集群中的当前训练阶段将暂停。随后，选择备用硬件，加载最后一个有效的检查点数据，并重新启动训练阶段。

硬件必须尽可能可靠，将所有形式的故障降低到最低实际水平。AI机器学习培训操作员在上线前可以选择对新安装的设备进行老化测试。强烈推荐最佳光纤布线实践，例如在连接前进行检查和清洁，以及仔细规划电缆和跳线管理。

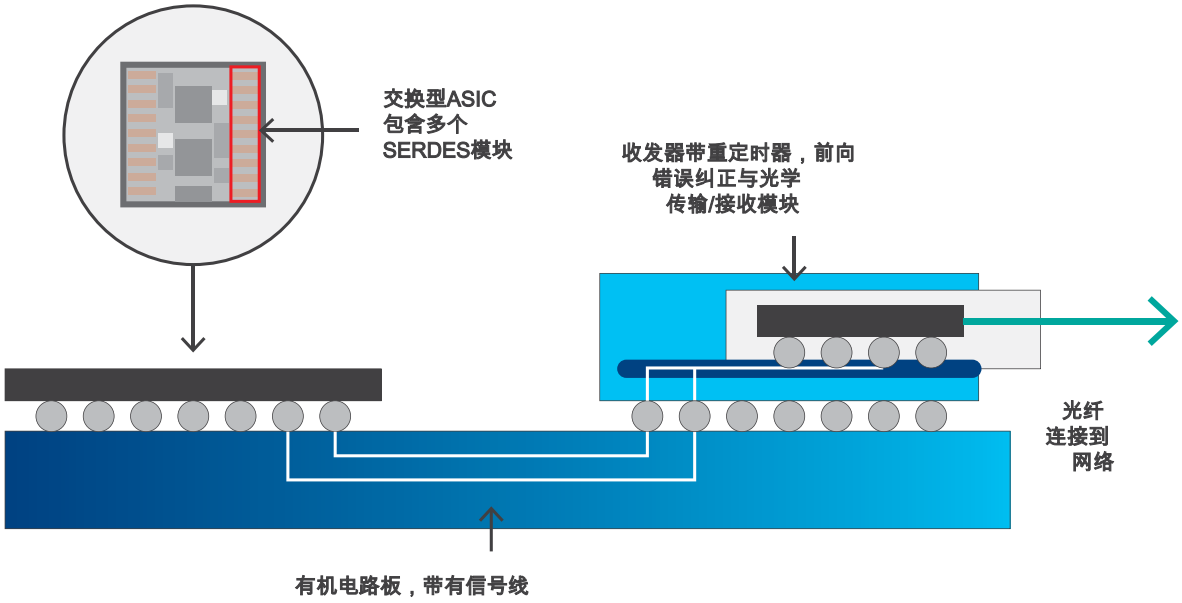
后端网络的性能指标是训练效率——所有加速器用于训练的总时间比例。关于人工智能机器学习大型集群效率的公开信息表明，高效性能水平超过五十个百分点。

序列化/反序列化器（SERDES）

计算使用由多个二进制位组成的数字、字符和其他符号进行。为了提高效率，这些符号以并行方式存储和处理。网络也使用符号来表示地址、数据包序列号以及其他更多内容。

处理器和交换设备在并行信息上操作。然而，超过电路板或机箱的通信几乎总是沿着单一的通道介质，如光纤。这意味着在每个芯片到媒体接口处，芯片并行信息必须转换为串行比特串，反之亦然。

此转换是通过片上组件SERDES（串行器/解串器）完成的。



SERDES设备在目标媒体带宽之上运行数据速率（例如，对于100 Gbps链路，数据速率为112 Gbps）。IEEE根据802.3标准组建立了以太网的SERDES速率，并以两倍步进增加（例如，28 Gbps、56 Gbps、112 Gbps和224 Gbps）。收发器可以通过组合多个信号通道来实现总端口带宽。例如，一个400G收发器可能从八个SERDES芯片端口接收八个56G的电气通道，并使用PAM4信号将这些通道组合成四个光纤对。

通过结合提高SERDES的基本速率和更多通道使用更多光纤、更多波长或更高信号密度，实现了更高的数据媒体和端口带宽。SERDES的数据速率受限于芯片性能、有机电路板中的信号退化以及链路中所有半导体组件的功耗。尽管如此，为了提升到更高的数据速率，全链路仍需要改进。

网络系统，其SERDES速率为224 Gbps，将于2025年开始进入市场。我们预计基于448 Gbps SERDES的系统将在2028年之后可用。

高频宽收发器

随着人工智能模型复杂度的增加，人工智能集群内部以及数据中心之间的快速数据传输需求日益加剧。高带宽收发器在满足这些需求中扮演着关键角色，为人工智能应用提供更多、更快的通信通道。

人工智能领域的高带宽收发器

需求

人工智能训练集群

人工智能模型训练是一项涉及成千上万个单独加速器的同步活动。在每一个训练阶段，每个加速器都在处理模型和训练数据的一部分。为了找到最佳模型权重和嵌入相关性的最佳匹配，这些结果将与集群中所有其他加速器共享，从而导致大量数据传输的突发。所有传输都必须在开始下一个训练阶段之前完成。

训练网络使用高性能拓扑和每个链路最高的实际带宽。带宽受限于终端设备能力（例如，交换芯片SERDES速率）、收发器带宽和媒体能力。对于超过几米的链路，使用光纤和光收发器。

分布式训练

同步跨多个节点进行训练需要低延迟、无损耗、维护数据包顺序的高带宽远程直接内存访问（RDMA）连接，以实现最大的训练效率。

带宽演进

从400G迁移到800G

2024年，采用800GB交换机和网卡端口的全新高端AI训练网络已出现。我们预计2025年将达到1.6TB的联网能力，随着SERDES设备的使用（从2026年开始），将过渡到3.2TB。

太比特收发器

工作正在进行中，以开发超过1 Tbps的收发器，以适应未来的应用。例如，请参阅正在开发的IEEE 802.3dj标准图表，该标准涵盖了200 Gb/s、400 Gb/s、800 Gb/s和1.6 Tbit/s，使用200 Gbit/s通道，预计将于2026年初发布。

以太网比率	信号比率	AUI背板铜缆			SMF 500m	SMF 2公里	固体模型化纤维 10km	SMF 20kg	SMF 40公里
200 Gb/s	200 Gb/s	200GAUI-1 C2C C2M	200GBASE-KR1	200GBASE-CR1	200GBASE-DR1	200GBASE-DR1-2			
400 Gb/s	200 Gb/s	400GAUI-2 C2C C2M	400GBASE-KR2	400GBASE-CR2	400GBASE-DR2	400GBASE-DR2-2			
800 Gb/s	200 Gb/s	800GAUI-4 C2C C2M	800GBASE-KR4	800GBASE-CR4	800GBASE-DR4 800GBASE-FR4-500	800GBASE-DR4-2 800GBASE-FR4	800GBASE-LR4		
	800 Gb/s	800GBASE-LR1		800GBASE-ER1-20	800GBASE-ER1				
1.6 Tb/s	100 Gb/s	1.6TAUI-16 C2C C2M							
	200 Gb/s	1.6TAUI-8 C2C C2M	1.6TBASE-KRS	1.6TBASE-CRS	1.6TBASE-DRS	1.6TBASE-DRS-2			

表格：IEEE P802.3dj接口和物理层规范摘要。图片来源于以太网联盟

最新创新

线性光学（LRO/LPO）

高频宽接收发射器采用数字信号处理（DSP）来补偿信号退化（重定时）和比特错误（前向纠错，FEC）。对于使用15瓦以上的800G接收发射器，一半或更多的功耗来自于DSP。考虑到网络接口卡和交换芯片中的高性能SERDES、管理的电路板信号长度和有限的纤维链路范围，即使没有DSP功能，也能实现高完整性通信。这可能导致总网络功耗显著降低，并提高延迟。

全线性收发器且不带任何数字信号处理功能的被称为线性可插拔光学（LPO），而仅在发射端配备数字信号处理的收发器则被称为线性接收光学（LRO）。选择线性光学需要关注系统兼容性。此类设备将用于短距离应用，如机架、行和区域连接。LRO/LPO设备现已进入生产阶段，并将于2025年部署。

一致光学

光学相干技术采用了光载数字信号的替代编码和解码方式，从而在消耗更少能耗的情况下实现更高的频带宽度和更长的传输距离。光学相干技术在长期远程电信应用中已有应用历史。在过去的几年里，适用于数据中心应用的小型化收发器中的光学相干技术已进入市场。例如，400G ZR收发器支持达40km的无需放大大连接，借助EDFA放大器，这一距离可扩展至120km。光学相干技术在数据中心间流量（DCI）中的应用正在日益增加。

随着数据速率的增加，相干光学的相对功率优势将降低其传输距离的折衷点。预计相干光学将从3.2T网络一代开始，在数据中心内部链路中具有可行性。

在板光学、近封装光学和共封装光学（OBO、NPO、CPO）

过去十年中，许多观察家预测光学连接将很快在内部和外部连接的设备中实现。一个主要驱动因素是电路板上铜迹信号传输相对较差。由于SERDES、DSP、电路板材料和一般学习的持续发展找到了在箱内连接中实现高带宽的实用解决方案，而没有求助于光学，因此OBO的到来在整个十年中被推迟。

然而，潜在的局限并未消失，直接连接到模块和芯片的光学连接在许多技术路线图中仍然存在。

芯片片技术及标准的开发降低了将光电组件直接集成到微电子设备中的障碍。

进一步推动因素是每片或模块外部连接数量的增长，使得周边或“海滩前”区域更加密集。先进的封装技术与光学芯片片组相结合是可能的解决方案。OBO/NPO/CPO的益处包括降低功耗和更低的延迟，这两者均受DSP功能需求减少的影响。

一个关键挑战是实现高可靠性和灵活性。标准可插拔收发器易于现场更换，并且可以选择适用于应用，无需对基础系统进行修改。我们的观点是，可插拔光模块与集成光模块之间将持续竞争，其中CPO（共封装光模块）将成为首选的集成光模块形式。

人工智能超级集群：接下来是什么？

挑战：超级扩展

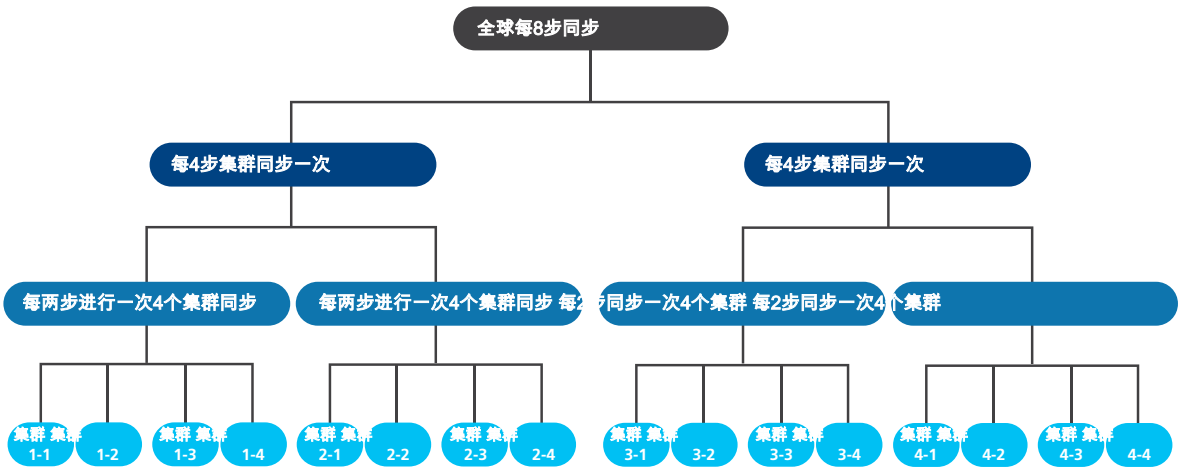
目前正在建设的最大单一地点人工智能设施拥有10万至20万台加速器。[30] 并且消耗300兆瓦[31]电力。同步训练将此类系统的物理跨度限制在不到一公里。过去十年的轨迹表明，通过在更多数据上训练更大的模型可以获得更好的模型，这意味着该行业将需要比目前正在进行的设施还要大的设施。没有在模型架构和高性能计算方面的突破，将需要更大的AI ML设施。

分段模型

一条攻击策略是模型的分割和训练，使得超大规模的训练工作量可以分布在多个段之间。模型研究基于几个有希望的攻击策略。在一种方法中，分割模型由单独的集群处理，参数更新定期进行（但不是在每个训练阶段，限制网络需求，放宽延迟要求）。

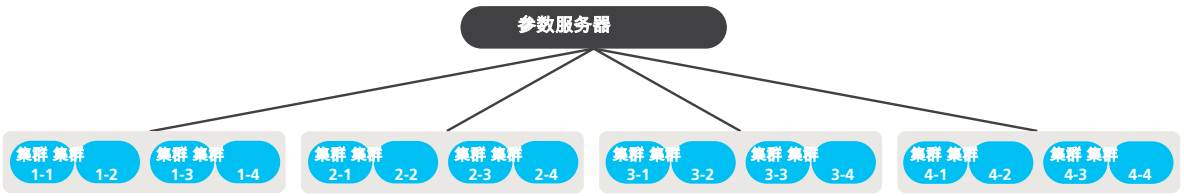
30. TECHSPOT 31. 数据中心知识

分段模型与周期同步



另一种方法是将模型参数存储在一个中央库中，并通过与分布式工作集群交换周期性更新来进行维护。非同步训练算法也正在开发中。

分段模型、全局参数服务器以及非同步更新



地理分布式的训练

随着进一步的发展，分段模型可能在广泛分离的设施中得以处理。这些设施可以位于电力供应良好且自然冷却可行的地点。站点之间的通信将需要专用的高带宽链路。

数据中心互联（DCI）中的中长距离链路

中距离（40公里以内）和长途链路（80公里及以上）目前在高效数据中心互联（DCI）解决方案中扮演着主导角色，为地理上分散的设施提供无缝通信和高速、可靠的 数据传输。这些链路将需要升级以支持地理分布式的AI ML训练。

也值得关注

- 纤维和光电子技术的发展将支持更高的带宽和密度。
- 持续的ANN模型开发将导致更小、更容易训练的模型。
- 将探索基于模拟技术和晶圆级集成的创意处理器。
- 热管理和技术分布将得到改善，但持续成为约束。
- 主要半导体晶圆制造和封装技术的开发将产生性能更高的加速器和网络设备。

结论

人工智能的快速发展对AI数据中心基础设施和高性能计算所需的先进硬件提出了巨大需求。从传统机器学习模型向复杂AI优化架构的转变导致了对更强大和高效的处理器以及更强大的网络的需求不断增长。半导体技术、芯片模块和封装技术的创新进一步加速了处理器和网络的发展。

我们观察到，极其大型且功能强大的AI基础设施的快速涌现是基于一个整合了多种个别技术的平台，每一种技术都极为出色。随着AI持续融入商业和个人生活的各个方面，推动这一多方面技术竞赛的动力在未来几年内将持续存在。

有效的扩展策略必须结合向上扩展和向外扩展，同时仔细关注实际数据中心设计和先进的冷却方法。随着人工智能模型持续增大和复杂化，对强大、高速网络和高效热管理的需求也将增加——如分段模型和扩展分布式系统等新兴因素将进一步塑造人工智能数据中心未来的格局。