驱动人工智能
支撑数据中心的半导体生态系统

Powering AI: The Semiconductor Ecosystem
at the Foundation of Data Centers

美国半导体行业协会（SIA）与德勤（Deloitte）联合报告

2026年6月

目录

执行摘要

一、AI核心处的众多芯片

推动AI进步的芯片创新

半导体AI芯片分类

二、拆解AI硬件栈：AI数据中心服务器中的芯片

计算托盘

加速器互连托盘

电源托盘

网络与智能平台管理接口（IPMI）托盘

冷却液分配单元（CDU）托盘

三、芯片如何协同执行AI训练工作流

四、AI增长曲线：市场视角

五、巨额支出：半导体内容价值分析

六、从设计到数据中心：了解AI数据中心的全球供应链

七、AI基础设施的新兴前沿

八、结论

作者简介

术语表

参考文献

执行摘要

半导体是人工智能（AI）的基础，这项技术正在改变我们的经济和社会，使整个产业更具生产力和创新性，并推动重大科学突破。当今的AI系统建立在半导体生态系统数十年创新的基础之上。随着芯片技术的不断进步，AI将变得更有能力、更节能、更具成本效益。更强大的AI反过来将有助于改进芯片设计、优化半导体制造，并推动对支撑AI的广泛芯片类别的更多需求。

关键要点：

1. 半导体是AI的基础使能技术。芯片为现代AI系统提供了基础硬件层，在现代AI服务器总价值中占据重要份额：

- 单个AI服务器机架包含超过4,500颗封装芯片，由约20,000个独立晶粒（即独特的集成电路）组成。
- 半导体占领先AI服务器机架内容价值的95%以上，以及建设和运营AI数据中心所需总资本支出的50%以上。

2. AI需要全方位的半导体技术。要运行复杂的AI训练和推理工作负载，当今的AI数据中心需要大量的计算、存储和内存带宽、电力分配和网络能力——所有这些都由全套芯片技术提供。这些芯片技术中的每一项对于推动美国的AI建设都至关重要，而这些领域中任何关键依赖都可能阻碍这一建设。AI数据中心中的芯片包括：

- 先进逻辑芯片，如AI加速器、专用集成电路（ASIC）、现场可编程门阵列（FPGA）、中央处理单元（CPU）、数据处理单元（DPU）和网络芯片。
- 存储芯片，如高带宽存储（HBM）、动态和静态随机存取存储器（DRAM和SRAM）以及非易失性闪存（NAND）。
- 模拟和基础芯片，如电源芯片、收发器、控制器和传感器。

3. AI是整个半导体行业芯片的主要需求驱动力。在一个正反馈循环中，AI的进步推动了对改进的半导体性能和效率的需求，而半导体技术的进步则使更强大、更先进的AI系统成为可能：

- 为满足全球对新AI应用的需求，政府和行业将在2028年前向新数据中心基础设施投资超过4万亿美元，其中高达2.8万亿美元将用于半导体。
- 部署在AI数据中心的半导体年收入到2028年可能超过1.2万亿美元，四年内增长近十倍。
- AI数据中心市场正经历前所未有的增长，预计2022年至2028年的复合年增长率（CAGR）为88.8%。虽然最初的势头由生成式AI的快速采用推动，但持续的需求保持强劲，预计2025年至2028年的CAGR为56.3%。

整个半导体供应链使AI基础设施的建设成为可能。没有半导体，就没有AI。要在这项变革性技术中保持领先，政府和行业必须共同努力推进政策，加速全谱芯片技术的增长和创新，并与全球合作伙伴密切合作，建立强大和有弹性的供应链。

一、AI核心处的众多芯片

人工智能（AI）近年来经历了爆炸性增长，吸引了大量关注于那些训练和部署AI模型的人和组织。各种各样的半导体作为AI硬件栈的支柱和使能技术，芯片技术的进步推动了AI应用在处理能力、计算效率和整体性能方面的提升。半导体支撑着嵌入日常数字体验中的AI系统。

本报告通过拆解最先进的AI数据中心服务器——现代AI基础设施的基础单元——提供了关于构成AI基础设施核心的多种芯片的独特由内而外的视角。与传统停留在系统级性能或市场规模分析的报告不同，本报告深入探讨服务器每个子系统内部的半导体内容本身——映射出驱动当今数据中心的芯片、晶粒和支持组件。

我们进一步通过聚焦这些服务器系统中价值集中的位置以及哪些技术至关重要来补充这一分析，从基于领先工艺节点的尖端逻辑到成熟节点组件——如电源管理集成电路（PMIC）、电可擦可编程只读存储器（EEPROM）、化合物半导体和微控制器——所有这些都是AI系统和基础设施功能不可或缺的。

推动AI进步的芯片创新

尽管AI看似是现代发展，但其基础贯穿数十年的先进计算能力发展。早期的AI系统在大型机和通用CPU上运行，计算成本高昂、速度慢且耗电量大，将AI限制在学术和国防实验室。随后由于训练方面的限制，AI经历了多次寒冬。

随着微处理器性能的提升和内存成本的降低，AI研究进入了更实用的应用领域，如语音识别和决策树。高性能计算（HPC）集群开始支持大规模并行计算，GPU迅速取代CPU成为训练工作负载的主力处理器。ResNet等突破性模型推动了对定制AI加速器的需求。2012年AlexNet在GPU上的训练标志着现代AI的黎明， ushering in an era of massive training，推动了对定制AI加速器的需求。

2022年生成式AI中大语言模型（LLM）的出现推动了对更专门的AI加速硬件的需求，通常与高带宽存储（HBM）配对。如今的AI基础设施要求前所未有的复杂性，AI训练运行需要数以万计的AI加速器，通常与HBM共同封装并通过光互连连接。2.5D/3D封装、芯粒和液冷方面的创新正在重塑数据中心架构。瓶颈已从纯计算转向热能管理、内存带宽和互连速度。

在过去几年中，逻辑、存储、网络、电源和冷却技术的规模和复杂性的持续改进为高性能AI系统的广泛部署铺平了道路。这些进步催生了AI数据中心的兴起。

虽然传统数据中心已经存在数十年，用于管理企业IT运营、网站托管和存储，但现代AI数据中心代表的不仅仅是增量演变，而是一种根本性的能力专业化。每个AI数据中心服务器机架都集成了一组由先进半导体器件组成的复杂组件，旨在支持并行化、数据邻近性和可扩展性。单个AI服务

器机架由约20,000个独立半导体晶粒组成，整合为超过4,500颗封装芯片。这些包括提供高吞吐量计算的逻辑处理器、超低延迟存储子系统、电源管理单元和网络组件。

AI数据中心中的芯片数量

AI数据中心中的每个服务器机架包含超过4,500颗芯片，而这些芯片又由约20,000个独立半导体晶粒组成。一个领先的数据中心可以容纳超过45,000,000颗芯片。

> 注：假设领先的AI数据中心拥有10,000个AI计算机架

随着各行业组织竞相部署AI驱动的解决方案，对AI数据中心容量以及先进半导体的需求skyrocketed。这种需求激增的影响贯穿整个半导体价值链。芯片设计师面临着缩短创新周期的压力，更频繁地发布新一代尖端器件。与此同时，晶圆代工厂必须交付重大的制造技术升级，以实现AI工作负载所要求的性能飞跃。这些工作负载正在推动前几代硬件的极限，暴露了架构优化、散热和数据在庞大系统中的移动等关键问题。作为回应，芯片制造商正在共同设计硬件和软件，并追求更紧密的内存和计算集成，从而推动了实现高密度、高带宽配置的创新封装技术的发展。

半导体需求增长循环

这是半导体与AI之间自我强化的创新循环：半导体技术和AI系统的进步使开发者生态系统日益成熟，而这反过来又要求日益强大的AI系统和半导体技术。随着开发者生态系统成熟度的提高，AI模型规模扩大，需要更多数据、更快的处理、系统间更紧密的协调以及更多计算，将现有芯片所能支持的能力推向极限。这导致了向高度专业化半导体设计的转变，包括更高效、更高性能的处理器、更专业化的存储堆栈以及能够支持整个数据中心分布式AI工作负载的高速互连。事实上，半导体设计师和制造商正越来越多地利用AI方法来推进下一代产品。

> **AI工作负载：训练与推理**
>
> 在整个报告中，我们经常提到两个主要的AI工作负载——即训练和推理，它们代表AI计算的不同阶段，并决定了芯片的设计方式：
>
> - **训练**是通过将模型暴露于非常大的数据集来教导模型的过程。例如，要构建一个猫识别模型，神经网络可能会被展示数千张猫和非猫的图像。通过反复 exposure，模型学习识别模式，如耳朵、胡须或体型，这些特征将猫与其他物体区分开来。
> - **推理**是将训练好的模型应用于新的、未见过的数据。继续猫的例子，一旦模型训练完成，它就可以生成一张新猫的图片。
>
> 简而言之，训练是模型学习的方式，而推理是它将所学应用于现实世界情境（如生成对查询的响应、做出预测或识别模式）的方式。

半导体AI芯片分类

AI服务器机架依赖多种协同工作的半导体技术，每种技术都量身定制以满足现代AI工作负载的苛刻要求。

图3 半导体类型——通过计算、存储、连接为AI数据中心提供动力

集成电路（IC）——在单个封装中集成多种功能的单晶粒或多晶粒芯片：

- **数字芯片**：执行计算、控制和内存处理以管理二进制工作负载
  - **逻辑芯片**：处理数据，为AI和云工作负载提供计算
    - **处理器**：按顺序执行任务（CPU）；控制AI数据中心中的系统编排和工作负载处理
    - **加速器**：并行运行许多操作，针对AI工作负载（如训练和推理）进行优化
    - **网络芯片**：跨服务器移动数据，实现低延迟AI模型访问
  - **存储/存储器**：存储、检索和处理用于训练和推理的数据
    - **DRAM**：为AI训练提供快速、易失性存储
    - **HBM**：通过提供快速、高带宽内存访问来加速AI工作负载
    - **NAND**：提供非易失性存储，实现快速AI数据集和数据集读取
    - **SRAM**：用作CPU和AI加速器中的缓存和缓冲区，实现低延迟数据访问
  - **信号与接口**：管理用于转换和接口的信号
    - **串行/并行转换器**：将模拟信号转换为数字数据（反之 versa），实现传感器输入控制

        - **ADC/DAC转换器**：调节电压和电源以支持系统运行
- **模拟芯片**：调节电源、管理热量、转换信号以支持数字基础设施
    - **电源芯片**：执行多种电源相关功能，包括控制开关、整流和多级电压调节、排序、监控和保护
    - **时钟/缓冲器/振荡器**：为AI数据中心中的芯片提供精确时序和同步
    - **光电芯片**：将电信号转换为光信号（反之 versa），用于通信、传感和显示
- **分立器件**：执行单一功能的独立组件
    - 处理电源的单功能设备，如晶体管
    - **传感器和换能器**：检测和转换信号以进行系统监控，包括温度/湿度、电气（电流分流）等环境传感器

> 注：
>
> 1. 信号中的某些芯片将包含混合信号，即模拟和数字属性
> 2. 示例并非详尽无遗，而是指示性的，由于AI半导体的复杂性质，多晶粒芯片封装包含来自不同分类的元素
> 3. 在本报告中，网络芯片被归类为逻辑芯片
> 4. 分立器件包括模拟和数字半导体，但该类别以模拟功能为主。纯数字分立器件已过时且罕见，因为数字逻辑通常集成到IC中以提高效率

每个服务器内部都有许多专用芯片，包括执行运行复杂AI训练和推理模型所需的并行处理的AI加速器。这些芯片同时执行数十亿次操作以高效处理大量信息。

支撑这一计算层，存储半导体对性能至关重要，实现快速可靠的数据访问。随着AI模型规模扩大，系统数据量急剧上升。高性能存储有助于确保处理器不断加载数据，避免瓶颈并保持系统响应性。在AI任务日益成为下一代AI系统设计的核心之际，存储和逻辑的专门化也在推进。

众多电源和网络半导体使AI服务器内和跨系统的高效能量传输和无缝互连成为可能。AI服务器同时使用许多芯片，每个芯片与其余芯片紧密协调，使电源、数据和计算资源的快速可靠管理成为必要。与此同时，分布式AI工作负载依赖于节点之间的快速、低延迟通信，使网络半导体对于协调互连服务器组中的计算至关重要。

图4 标准AI硬件栈：从硅晶粒到机架

层级	定义
晶粒（Die）	半导体层次结构中最小的功能单元——包含执行逻辑、存储或I/O功能的晶体管的硅片（如AI加速器晶粒、DRAM晶粒、PHY）。
芯片（Chip）	集成一个或多个晶粒的封装半导体器件。封装提供电气连接、热管理和机械保护。芯片设计用于安装在电路板上。
电路板（Board）	承载芯片、无源器件、存储器和连接器的印刷电路板（PCB）。电路板还可以承载辅助模块（如计算板上的SSD），在单个电路板 footprint 内形成分层组件。

托盘（Tray）	为一个或多个针对特定功能配置的电路板的模块化组件（如计算托盘、互连托盘、电源托盘）。托盘设计用于在机架内热插拔安装，并实现子系统模块化。
机架（Rack）	容纳多个托盘以及共享的电源、网络和热基础设施的完整服务器机柜。机架是AI数据中心的顶层构建块，在全系统范围内扩展计算能力。

二、拆解AI硬件栈：AI数据中心服务器中的芯片

如以下章节所述，拆解AI服务器机架硬件揭示了一个高度模块化、垂直集成的系统。它由服务器内每个子系统中的数万个相互依赖的半导体组件组成，从CPU和加速器到信号调理芯片、电源稳压器、存储晶粒和控制逻辑。

图5 AI数据中心服务器拆解示意

AI数据中心服务器机架由多种服务器类型组成——通常称为"托盘"——每种都设计用于执行不同且关键的功能。机架架构和单个托盘配置都经过定制，以满足客户需求和工作负载要求。

托盘类型	数量/机架	功能	芯片数量	价值/机架
网络与IPMI托盘	1-2个	提供控制平面基础设施，结合带外管理、高速网络和硬件根安全，以管理和保护机架级操作	约100颗	约$17,000-$25,000
电源托盘	1-2个	机架的电气基础，容纳电源单元（PSU）和电源管理模块（PSMM），以高效率和实时监控转换、调节和分配千瓦级负载	约600颗	约$50,000-$290,000
CDU托盘	1-2个	通过连接到CPU、AI加速器和存储器上冷板的闭环系统管理机架级液冷，确保超过空气冷却极限的稳定热性能	约10颗	约$15,000-$30,000
计算托盘	2-4个	AI服务器机架的大脑，托管主要处理逻辑。每个托盘通常包含一个或多个计算板、本地固态硬盘（SSD）用于存储、网络接口卡（NIC）用于连接，通常还有数据处理单元（DPU）用于管理数据	约70颗	约$1.5M-$3.5M
加速器互连托盘	8-18个	建立统一的高性能计算架构，使用交换机ASIC、电源分配块和有源铜缆（ACC）芯片来同步AI加速器之间的存储器和计算	约4,000颗	约$10,000-$50,000

> 注：本图仅供参考；实际服务器组件配置可能因规格、价格、供应商和超大规模云服务商要求而有显著差异。此外，托盘尺寸也可根据托盘/功能类型、供应商来源、数据中心要求等在1U-8U之间变化。

服务器机架包含一组协调的托盘，每个托盘设计用于提供特定功能。每个托盘或子系统包含半导体和支持电子元件的组合——加载了半导体内容——协同工作以提供AI工作负载所需的吞吐量、

能源效率和可靠性。理解这种分层架构是认识支撑它的半导体供应链战略重要性的关键。随着AI继续扩展，保护、设计和集成这些组件的能力将变得与计算本身同样关键。

一般来说，现代数据中心的AI数据中心服务器机架由五种类型的托盘组成，即1）计算托盘，2）电源托盘，3）网络与智能平台管理接口（IPMI）托盘，4）AI加速器互连托盘，和5）冷却液分配单元（CDU）托盘。每个托盘容纳具有关键作用的不同半导体组件。服务器机架中托盘的确切组成因数据中心而异，变化由OEM选择、规模、布局、电源配置以及数据中心的建造者和/或运营商驱动。

本报告逐托盘拆解通用AI数据中心服务器中的半导体内容。系统中的每个组件，从AI加速器到单个电压稳压器，在确保大规模性能、效率和可靠性方面都发挥着至关重要的作用。

计算托盘

计算托盘是AI服务器机架的大脑，为AI工作负载提供所需的主要处理逻辑。每个计算托盘通常包含一个或多个计算板、非易失性存储器 express（NVMe）固态硬盘（SSD）用于存储、网络接口卡（NIC）用于连接、数据处理单元（DPU）用于将数据组织成离散处理单元，以及电源分配单元（PDU）。该托盘承载了服务器机架中最昂贵、最先进和最集中的半导体组件，因为它直接负责每瓦性能和AI工作负载吞吐量。

图6 计算托盘组件映射

A. AI加速器

AI加速器执行核心AI模型计算，并行处理海量数据并与HBM紧密接口。**单价：约$10,000-$40,000**（内置HBM价格）

B. 高带宽存储（HBM）

HBM直接向AI加速器提供超快速、片上内存带宽，以实现不间断计算。**单价：约$30-50/GB**；假设机架中有14TB，成本约为$400-600K

C. 中央处理单元（CPU）

CPU作为服务器的控制中心，管理内存访问、控制系统启动和关闭、处理轻量级AI推理，并在CPU和DPU之间分配工作负载以处理数据缓冲和任务切换。**单价：约$7,500-$15,000**

D. 系统内存（RAM）

为服务器操作期间的CPU和DPU提供快速、临时存储，用于数据缓冲和任务切换。**单价：约$7-8/GB**；假设机架中有17-24TB，成本约为$135-190K

E. 非易失性存储（SSD）

通过PCIe快速存储和服务大型AI数据集和模型，为AI加速器和CPU提供计算任务的数据。用于加载AI训练数据集、中间检查点和推理就绪模型。**单价：约$0.20/GB**；假设机架中有200TB，成本约为$40,000

F. 数据处理单元（DPU）

专用的数据轨道控制器，从CPU卸载网络、存储访问、加密和遥测工作负载。**单价：$1,000-$3,500**

G. 网络接口卡（NIC）

执行硬件和软件系统之间的指令、计算和数据流管理。**单价：$7,500-$15,000**

H. 电源分配单元（PDU）

将电源从机架级母线降压、调节和管理到单个芯片（CPU、AI加速器、DPU、NIC）。**单价：$1,000-$2,000**

计算板

每个计算托盘容纳一到两个计算板，AI计算在其中进行协调和执行。每个计算板作为计算托盘内的高密度处理模块。它通常将一个或多个高性能AI加速器与通用CPU相结合，通过高带宽互连紧密耦合，并由本地随机存取存储器（RAM）支持。计算板设计用于并行性、热效率和信号完整性。

表1 典型AI计算板的子组件

子组件	描述	数量	芯片类型
AI加速器	带有片上HBM和互连晶粒的高性能加速器	1-2	逻辑和存储——DRAM

中央处理单元（CPU）	用于任务编排、输入/输出（I/O）和通用工作负载的多核微处理器	1-2	逻辑
系统内存（RAM）	用于CPU内存访问的高速DDR5或LPDDR5模块	8-32个DIMM	存储——DRAM
电压稳压模块（VRM）	电源稳压器、电源管理IC和分立金属氧化物半导体场效应晶体管（MOSFET）或绝缘栅双极晶体管功率级	50-100+	模拟和分立——功率晶体管
控制和支持IC	板管理控制器、温度传感器、振荡器、缓冲器、时钟发生器和/或其他锁相环	10+	逻辑、传感器和模拟
网络和互连	通信逻辑IC（如InfiniBand、以太网）	1+	逻辑

AI加速器

AI加速器是一个总称，包括图形处理单元（GPU）、现场可编程门阵列（FPGA）和专用集成电路（ASIC），是专用半导体硬件，旨在通过优化计算密集型操作来加速人工智能工作负载，否则这些操作将压垮通用处理器。这些加速器主导着AI数据中心中的大规模训练和推理，其中最大化吞吐量和性能是优先事项。在边缘端，一类新的加速器已经出现：神经处理单元（NPU）。NPU专门用于以卓越的能效执行神经网络工作载，优化对推理至关重要的矩阵和张量操作。

虽然存在多种AI加速器架构，但GPU仍然是当今AI训练和高性能推理中最具市场相关性和最广泛部署的平台。它是现代AI基础设施的计算动力源。GPU架构最初设计用于渲染图形，允许极端的并行处理能力来处理数十亿次计算。与其他现代AI加速器一样，它们集成逻辑晶粒（采用7nm等尖端工艺技术制造）和HBM堆栈，实现实时AI应用和大语言模型处理所需的快速数据吞吐。

这些AI加速器依赖于先进的封装技术，如带有高带宽存储（HBM）的2.5D集成和新兴的3D方法，这远远超出了将单个晶粒安装在封装中的传统封装。与传统封装方法不同，先进封装将多个晶粒和存储器紧密集成，以最大化带宽、减少延迟并提高能效，使其成为驱动AI数据中心工作负载的关键。值得注意的是，这些AI加速器本身由有源和无源子组件组成。这些子组件的通用物料清单（BOM）分解见图7。

图7 现代AI加速器和HBM的示意图分解

高带宽存储（HBM）

HBM是一种封装在AI加速器内的动态随机存取存储器（DRAM），为现代AI加速器提供极快、高效的内存性能。与系统内存（下文进一步描述）不同，HBM使用垂直电连接（即硅通孔或"TSV"）垂直堆叠存储层，以提高与逻辑处理器之间数据传输的速度、能效和可扩展性。AI加速器中HBM与逻辑晶粒的物理邻近性以及HBM的大规模并行访问能力缓解了否则会限制AI模型性能的存储瓶颈，特别是对于需要快速访问数十亿参数的大语言模型。HBM通常用于大规模AI模型的训练和提

供高吞吐量推理工作负载，其中存储带宽直接影响用户体验和运营效率。然而，并非所有AI系统都需要HBM。许多AI芯片，特别是针对成本敏感、延迟容忍或较窄推理工作负载优化的芯片，依赖替代存储架构，如片上DDR、LPDDR或更大的片上SRAM，以平衡性能、功耗和系统成本。

表2 集成AI GPU的半导体组件

子组件	描述	数量	芯片类型
逻辑晶粒（ASIC）	AI加速器的"计算引擎"，一次执行数百万次数学运算，驱动训练和推理。通常包含张量核心、矩阵单元和缓存	1-2	逻辑
互连PHY（物理链路）	这是芯片间的高速通信端口，连接AI加速器与其他加速器或CPU	1+	逻辑
HBM（高带宽存储）	堆叠DRAM，每个堆栈多达16个晶粒，与逻辑晶粒共同封装，由硅通孔（TSV）连接。实现AI工作负载所需的极快数据访问	6-12个堆栈（96-192个晶粒）	存储——DRAM
共同封装的中介层或桥接晶粒	一种特殊的互连层，连接AI加速器晶粒、互连Phy、HBM和其他芯粒，使它们能够以统一晶粒的效率和带宽运行。用于先进AI加速器和CPU异构片上系统架构	1	逻辑
VRM和无源器件	电压稳压模块（VRM）和无源元件（电容、电感、扼流圈、电阻）是AI服务器硬件的幕后功臣。VRM本身不是单个芯片，而是一个小型电源子系统，通常由多相控制器构建，驱动由MOSFET组成的功率级，MOSFET又驱动驱动电路、电感和电容。它们一起将机架级电压降至GPU、CPU和存储器所需的精确1V以下轨线。每个加速器板可以容纳数十个VRM和数百个支持无源器件，在整个AI服务器机架中，这些组件数以千计。虽然单个成本低廉且基本商品化，但它们在AI工作负载极端且快速变化的需求下维持稳定、高效的电源传输方面是不可或缺的。	数十个VRM+数百个无源器件	模拟和分立

中央处理单元（CPU）

CPU是AI服务器内的编排器和通用计算引擎，负责监督数据准备（数据预处理和后处理）、协调工作负载执行以及管理整体系统控制。CPU逻辑晶粒采用先进到当前代节点构建，取决于性能和成本目标。在AI基础设施中，CPU对数据摄取、模型服务协调以及运行支持AI工作负载的操作系统和框架至关重要。现代服务器CPU使用多芯片模块封装来组合多个具有大缓存层次结构的半导体晶粒，支持分布式AI训练和推理操作的任务调度和存储器管理。

系统内存/随机存取存储器

系统内存/随机存取存储器为AI服务器提供主要的数据暂存和缓冲能力，作为永久存储和处理单元之间的高容量、高速存储层。系统内存/RAM由动态随机存取存储器晶粒构成，排列在双列直插式存储器模块（DIMM）上或直接焊接到主板上。系统内存的容量和带宽直接影响AI数据服务器可以同时处理的数据集大小以及服务器可以支持的并发AI工作负载数量，使其成为确定整体系统吞吐量和成本效益的关键因素。

图8 DRAM DIMM的示意图分解

表3 DRAM DIMM的半导体组件

子组件	描述	数量	芯片类型
DRAM晶粒	使用易失性存储器单元高速存储和检索数据	每个DIMM 8-16个晶粒	存储——DRAM
寄存时钟驱动器（RCD）	缓冲和同步来自存储器控制器的时钟和命令信号	1	模拟——信号转换
电源管理IC（PMIC）	本地管理并向DRAM芯片分配电源，减轻主板负担	1	模拟——电源管理
串行存在检测（SPD）	存储系统BIOS正确识别存储器所需的配置数据	1	存储——闪存EEPROM
温度传感器	监控模块温度以进行热节流或关断	1	传感器

架构权衡：串行 vs. 并行 vs. 可重构计算

CPU和AI加速器主要在其处理架构上有所不同。CPU针对串行处理进行优化，以低延迟顺序执行少量复杂任务。CPU可以在不需要专用加速器的情况下处理较小、基础、低参数AI模型的推理，使其足以满足延迟和规模要求较低的AI任务。

AI加速器专为大规模并行而构建，同时处理数千个更简单的操作。这使AI加速器每秒能够处理比CPU多10-100倍的计算，并实现更大的数据处理吞吐量，使其特别适用于涉及大规模矩阵操作的AI训练和推理工作负载。因此，AI加速器已成为AI数据中心的首选计算引擎。

截至2015年，高端GPU还存在于游戏机中，而非数据中心。大多数数据中心任务由CPU处理——一次执行一个任务的串行处理器。2016年，研究人员发现用于游戏的GPU由于其并行性可以高效运行机器学习模型。数据中心随后开始采用为AI优化的GPU，尽管市场规模仍然不大。2022年生成式AI中大语言模型（LLM）的出现推动了对更专业的AI加速硬件的需求，通常与高带宽存储（HBM）配对。这些单元依赖于数据中心级或服务器级CPU，在架构上类似但不同于标准CPU，以及其他芯片。

FPGA作为AI基础设施中的适应性加速器，占据了一个低延迟、确定性执行和接口灵活性至关重要的细分市场。与其他擅长大规模并行矩阵操作的AI加速器或管理通用控制任务的CPU不同，FPGA因其针对工作负载的特定加速而受到重视，专门用于推理。它们优化数据路径，减少输入-输出瓶颈，并与各种传感器和网络协议无缝集成，使其非常适合边缘AI、电信和配置及每瓦性能最重要的专业数据中心用例。

FPGA是AI数据中心中其他AI加速器的补充，在专业推理细分市场中开辟价值。

非易失性存储器 express 存储

非易失性存储器 express 固态硬盘（NVMe SSD）存储结构化和非结构化数据以支持现代AI工作负载。利用PCIe接口和为并行性优化的精简命令集，NVMe驱动器具有超低延迟和极高吞吐量，能够快速摄取大型数据集、快速模型检查点以及无缝推理模型加载。这些能力消除了可能阻碍AI加速器利用的I/O瓶颈，使NVMe存储成为端到端AI管道效率的核心推动因素。在推理中心系统中，这些存储器驱动器通常充当模型缓存，按需向存储器提供大型模型以减少冷启动延迟。除了AI加速器和CPU等逻辑组件外，存储组件也由多个子组件组成，如典型NVMe SSD的BOM分解（表4）所示。

表4 典型数据中心非易失性存储器 express 固态硬盘的子组件

子组件	描述	数量	半导体类型

NAND闪存晶粒	使用闪存技术在存储器单元中存储持久数据	256个（封装中32层 x 8堆栈）	存储——闪存
SSD控制器	管理数据流、温度控制、磨损均衡和闪存转换层	1	逻辑
DRAM缓存	临时存储元数据和用户数据以实现快速访问	1	存储——DRAM
I/O IC	处理接口协议（PCIe/NVMe）和信号转换	1	模拟——接口
PMIC	调节所有有源SSD组件的电源传输	1	模拟——电源管理

数据处理单元（DPU）

数据处理单元是一种专用的、可编程的基础设施处理器，可从AI数据中心的CPU和AI加速器上卸载非计算任务。它运行在网络、存储和安全的交叉点，作为高效资源编排的控制器和数据路径推动者。在AI中心工作负载中，海量数据集在AI加速器之间流式传输和shuffle，DPU优化网络数据流（东西向流量），处理安全任务（如隔离工作负载），并收集实时系统数据。现代DPU在使AI基础设施更灵活方面发挥着关键作用，允许通过软件动态分配和重新配置计算、存储和网络资源。与AI加速器一样，DPU也由其他半导体组件组成（表5）。

表5 典型DPU的半导体组件

子组件	描述	数量	半导体类型
嵌入式逻辑晶粒	DPU内部的小型处理器，运行自己的迷你操作系统；处理软件和管理任务，不涉及主服务器CPU	1	逻辑
包处理引擎	在网络、存储器和设备之间快速移动数据的定制引擎；处理加密、流量引导和遥测等复杂任务，速度比普通CPU更快	1	逻辑
片上DRAM	用于缓冲、控制平面操作和软件执行的外部存储器	1-2	存储——DRAM
物理链路（PHY）接口	通过光纤或铜缆发送和接收以太网流量；转换为物理链路	1	模拟——接口
闪存（SPI/EEPROM）	存储启动固件、遥测日志和操作系统镜像	1	存储——闪存

PMIC/VRM	SoC和高速I/O域的电源调节和分配	6-8	模拟——电源管理

网络接口卡（NIC）

现代AI工作负载需要巨大的带宽。网络接口卡实现了分布式AI训练和推理服务所需的高速互连，支持服务器机架、计算托盘和存储系统之间的太比特级数据传输。具有远程直接存储器访问（RDMA）和硬件加速网络协议等先进功能的NIC最小化通信开销和延迟，直接提高大规模AI部署中的训练效率和推理响应时间。核心半导体内容包括网络控制器ASIC以及用于信号完整性和时钟/数据恢复的模拟物理链路（PHY）晶粒。网络ASIC有两种类型：纵向扩展（scale-up），促进系统内的低延迟数据移动；以及横向扩展（scale-out），实现更大、分布式系统之间的连接和协调。这些芯片通常与板上存储器缓冲器封装，以支持拥塞控制和数据包排队（表6）。

表6 高性能NIC的子组件

子组件	描述	数量	芯片类型
网络控制器ASIC	带有嵌入式DRAM的片上系统，处理数据包处理、RDMA和PCIe/CXL连接	1	逻辑
光收发器	将高速电信号转换为光信号的模块	1	光电子
SerDes/物理链路（PHY）接口	将数字数据转换为模拟信号以进行高速串行传输的电路	4-8	模拟——接口
DDR存储器模块	用于缓冲、排队和控制平面操作的板上DRAM	1	存储——DRAM
PCIe接口	通过PCIe Gen4/Gen5实现主机通信的连接器和逻辑	1	逻辑
PMIC/VRM	为核心ASIC、I/O模块和存储器的电源传输和调节	2-4	模拟——电源管理
数字信号处理器	增强和均衡长光纤传输的信号，特别是在DR4/DR8光模块中	1	逻辑——数字信号处理器
振荡器	为NIC时序提供稳定的参考时钟信号	1	分立——时序
时钟缓冲器	将振荡器参考时钟分配到多个NIC域，同时保持信号完整性	1-2	模拟——时序/缓冲
时钟发生器	振荡器和缓冲器的替代品；从单个参考源派生多个频率	1	模拟——时序/时钟发生器

电源分配单元（PDU）

在先进的AI计算托盘中，电源分配单元将高压输入的电能转换并分配给CPU、AI加速器、存储器和其他子系统所需的精确低压轨线。在先进的AI计算托盘中，PDU接收机架级48V或更高电压，并向具有高峰值负载的高密度组件提供严格调节的电压，通常低于1V。内部，它们容纳多个数字和模拟半导体组件，包括电源管理IC、电压稳压模块（VRM）、集成场效应晶体管（FET）和微控制器。这些芯片通常采用当前代到成熟工艺节点制造，以确保耐用性和热稳定性。

随着AI数据中心需要更多电力，负责管理和转换电力的芯片正在发展为提供更高的效率。化合物半导体，如氮化镓（GaN）和碳化硅（SiC），越来越多地用于电源管理系统，因为它们可以处理更高的电压，能量损失和热量更少。此外，AI数据中心有可能从当前高压交流电与48V直流电混合供电给机架的方式，转向提议的400V直流电全系统供电，这将带来各种效率提升。甚至有一些计划转向800V高压直流（HVDC）架构，以支持在不久的将来单个需要1 MW（即1,000 kW）电力的机架，相比之下，当今最强大的机架需求为100-120 kW。

表7 AI服务器计算托盘组件典型电源分配单元的半导体组件

子组件	描述	数量	芯片类型
中间总线转换器（IBC）	大功率DC-DC转换器，将机架级48-51V总线降至中间轨线（通常为12V）	2-6	分立
负载点（PoL）转换器	本地电压稳压器，将中间轨线（12V/6-8V）转换为低压轨线（5V、3.3V、1.8V、1.0V）	6-12	分立
微控制器或PMBus SoC	用于闭环电源管理、遥测和I2C/PMBus控制	1-2	逻辑——MCU
电源管理IC（PMIC）	用于排序、遥测和动态电压调节的控制逻辑	2-6	模拟——电源管理
电流和电压传感器	用于实时监控和故障检测的模拟/数字IC	5-10	模拟——放大器
驱动器和FET	在IBC级执行实际电流开关和调节（如48V->12V）	2-4	分立

> **超越硅：化合物半导体**
>
> 化合物半导体代表了材料科学的战略演进，在硅达到物理极限的地方扩展性能范围。氮化镓（GaN）、碳化硅（SiC）、砷化镓（GaAs）和磷化铟（InP）等材料实现了宽带隙、更高电子迁移率和卓越热导率等特性，这些对下一代数据中心扩展和光技术至关重要。这些进步使器件能够在更高电压、频率和温度下运行，减小支持基础设施的尺寸和冷却要求。
>
> 电源和存储在芯片数量上占主导地位；电源芯片还因使用GaN/SiC的材料多样性而脱颖而出。

加速器互连托盘

加速器互连托盘创建一个统一的、高性能的计算架构，使AI服务器内的AI加速器能够作为一个单元运行。该托盘包含高带宽交换机专用集成电路（ASIC）、电源分配块和信号调理有源铜缆（ACC）芯片，这些芯片协调服务器机架内数十个AI加速器之间的无缝通信（图10）。这使它们能够共享存储器、同步计算并共同处理超出单个加速器能力的大规模AI工作负载（表8）。

图10 加速器互连托盘

交换机ASIC

交换机ASIC实现服务器机架内或跨机架的多个AI加速器之间的高速、低延迟通信。它们的性能直接影响AI工作负载的可扩展性和效率，特别是在需要跨多个加速器同步处理的大语言模型训练和其他计算密集型应用中。

电源分配块

加速器互连托盘内的电源分配块管理本地化的、高电流的输送，为交换机ASIC和互连组件供电，在整个服务器机架中汇集和重新分配电力。内部，加速器互连托盘的这一元素集成了电源芯片，为托盘分配和调节电力。

有源铜缆（ACC）芯片

ACC芯片是嵌入在ACC两端的小型模拟信号调理IC，实现托盘或机架之间的高速、短距离通信（最长5米）。

背板连接器

背板连接器作为连接计算托盘上多个AI加速器的主要枢纽。背板连接器是机电组件，不是半导体器件，但为了系统互连的完整性而在此保留。

表8 GPU互连托盘的半导体组件

子组件	描述	数量	芯片类型
高速PHY	将数字数据转换为极快的电信号，用于加速器之间的数据发送	2	逻辑——数字信号处理器
交换机ASIC（结构交换机）	在多个加速器之间路由流量，使它们都能高效地相互通信	2个/互连托盘	逻辑
SerDes收发器	将宽数据流转换为窄的高速串行通道，再转换回来	每个交换机ASIC 16-72+通道	光电子
ACC芯片	管理多个AI加速器模块之间低级控制、遥测和协调的专用控制器IC	1-2	模拟——信号转换
PMIC/VRM	集成到电源分配块中，以高效调节和分配电力	4-6	模拟——电源管理
时钟发生器/锁相环（PLL）IC	为同步SerDes、交换机ASIC和PHY提供精确时序信号	1-2	逻辑——数字信号处理器
振荡器	为AI加速器互连时序提供稳定的参考时钟信号	1	分立——时序
时钟缓冲器	将振荡器参考时钟分配到多个GPU/SerDes域，同时保持信号完整性	1-2	模拟——时序/缓冲
时钟发生器	振荡器和缓冲器的替代品；从单个源派生多个频率	0-1	模拟——时序/时钟发生器

电源托盘

电源托盘作为AI服务器机架的电气基础，转换、调节和分配电力给机架内的所有系统组件。该托盘容纳必须可靠地输送千瓦级电力负载的主电源单元（PSU），同时保持高效率并提供实时监控能力。电源托盘内的关键半导体组件包括将电流转换的桥式整流器、优化输入电源质量和效率的功率因数校正（PFC）控制器IC，以及快速开关和调节电压转换以实现高效率的功率晶体管（通常为GaN基）。

与位于计算托盘上或附近并负责本地化电压调节和向计算托盘上单个组件进行精细电流输送的PDU不同，电源托盘在系统级运行，处理整个服务器机架的 bulk power conversion 和智能电源编排。这些电源部件共同组成了一个为高密度、高可靠性计算环境设计的分层电力输送网络。

图11 电源托盘

电源管理模块（PSMM）

电源管理模块作为AI服务器中电力分配和监控的智能编排器。PSMM系统通过在需要时限制电力来防止过热，平衡服务器不同部件之间的电力，并通过持续监控电源质量来支持主动维护。为此，PSMM集成了多种芯片，包括智能微控制器、本地电源电压稳压器、远程访问功能（以太网PHY）和传感器，协同工作以保持服务器高效运行并最小化停机时间。

电源单元（PSU）

在电源托盘内，PSU作为AI服务器的主要电源转换和调节系统，将数据中心基础设施的高压交流电转换为服务器组件所需的多个直流电压轨（表9）。

表9 服务器级电源单元的半导体组件

子组件	描述	数量	芯片类型
桥式整流器	将来自市电的输入交流电（AC）转换为未调节的直流电（DC）；构成PSU输入级的一部分	1	分立——整流器
功率因数校正（PFC）控制器IC	确保输入电流与电压波形同相，提高功率效率并减少谐波失真	1	模拟——电源管理
功率MOSFET（如GaN）	用于开关级的高速晶体管，用于调节电压转换；通过快速开关实现高效的AC-DC转换	2-4	分立——功率晶体管
脉宽调制（PWM）/谐振控制器IC	通过控制占空比或谐振时序来调节变压器的开关行为，优化电力传输和效率	1	模拟——电源管理
栅极驱动器IC	在控制器IC和功率MOSFET之间接口，提供切换晶体管所需的正确时序和电压	2-4	模拟——电源管理
同步整流器FET	使用低电阻晶体管替代传统二极管进行整流，减少功率损耗并提高转换效率	2-4	分立——功率晶体管
微控制器/电源管理总线片上系统（PMBus SoC）	用于监控、遥测、保护和通信（如通过PMBus或I2C协议）的嵌入式控制单元；管理整体PSU逻辑	1	逻辑——MCU

电压和温度传感器	监控内部电压轨、电流消耗和热条件以实现系统保护和动态控制	3-5	传感器

网络与智能平台管理接口（IPMI）托盘

网络与智能平台管理接口（IPMI）托盘建立在半导体组件的基础上，这些组件为AI服务器操作实现安全、可扩展的控制和连接。该托盘内的关键半导体包括基板管理控制器、硬件安全芯片、本地存储器和高速网络交换机。

图12 网络与IPMI托盘

基板管理卡（BMC）

基板管理卡实现远程电源循环、硬件健康监控和系统配置，而不会中断正在运行的AI工作负载（表10）。

表10 AI服务器机架BMC的子组件

子组件	描述	数量	芯片类型
SoC（微控制器）	运行固件和IPMI堆栈以进行带外管理的中央处理单元	1	逻辑——MCU
固件闪存	用于BMC固件、日志和配置数据的非易失性存储	1	存储——闪存
以太网PHY + 磁性元件	通过LAN提供远程管理的物理和电气连接	1	模拟——接口
电源管理IC（PMIC）	为BMC SoC、存储器和I/O接口提供稳压电压	1-2	模拟——电源管理
看门狗定时器IC	如果检测到故障或冻结，自动重置系统	1	逻辑——MCU
时钟发生器	为BMC和外围接口提供参考时钟信号，在某些设计中替代振荡器和缓冲器	1	模拟——时序/时钟发生器

现场可更换单元（FRU）

现场可更换单元电可擦可编程只读存储器（FRU EEPROM）充当服务器组件的"数字名称标签"，存储身份、配置和服务历史等基本详细信息。这使数据中心能够轻松自动跟踪、管理和维护硬件。

带外（OOB）交换机

带外（OOB）交换机使用专用半导体芯片为服务器管理任务创建一个单独的网络。这些交换机（由网络处理器和微控制器驱动）帮助维护对服务器的可靠监督和控制，即使系统处于重负载下。这种分离由OOB交换机内部的先进网络半导体实现，确保管理员的安全和不间断访问。

表11 带外交换机的子组件

子组件	描述	数量	芯片类型
交换机ASIC	使用MAC表、VLAN和路由逻辑实时在端口之间路由数据包的核心逻辑芯片	1	逻辑
以太网PHY接口	将数字信号转换为通过物理以太网介质传输的电信号，并处理链路协议	24	模拟——接口
管理SoC/CPU	运行交换机操作系统并处理CLI、API、遥测和系统管理功能的嵌入式处理器	1	逻辑——MCU
闪存	用于保存固件、配置文件、操作系统和事件日志以实现恢复的非易失性存储	1	存储——闪存
DRAM	为管理处理器使用的有源数据提供临时存储的易失性系统存储器	1	存储——DRAM
电源管理IC	调节所有有源SSD组件的电源传输	1	模拟——电源管理

带内（IB）交换机

带内（IB）交换机通过支持400G/800G以太网或InfiniBand等先进网络协议，管理AI数据中心内计算节点和存储系统之间的高速数据流量。IB交换机确保可扩展的同步通信，这对于多机架AI集群中的训练和推理至关重要。

表12 带内交换机的子组件

子组件	描述	数量	芯片类型
交换机ASIC	以超高速接收数据包并将其路由到正确端口的中央芯片	1	逻辑
SerDes收发器	为每个端口将并行数据转换为高速串行数据（每通道112G-224G）	100+	光电子

以太网/InfiniBand PHY接口	根据端口类型（如OSFP）将数据转换为电信号/光信号	1	模拟——接口
光收发器	用于高速光纤链路的可插拔光学器件（800G、1.6T）	72	光电子
数字信号处理器	增强和整形高速信号以实现清洁的光纤传输	1	逻辑——数字信号处理器
电源管理IC	为交换机ASIC和收发器的不同部分提供和调节电力	10	模拟——电源管理
风扇/热管理控制器	用风扇控制温度，通过串行和以太网端口启用本地和远程设置，并通过指示灯显示系统状态	1	逻辑——MCU
闪存	用于保存固件、配置文件、操作系统和事件日志以实现恢复的非易失性存储	1	存储——闪存
DRAM	为管理处理器使用的有源数据提供临时存储的易失性系统存储器	1	存储——DRAM

可信平台模块（TPM）

可信平台模块为AI服务器提供基于硬件的安全基础，包括安全启动功能、加密密钥存储和硬件认证服务。TPM也被称为安全模块。在AI环境中处理敏感数据或专有模型时，TPM模块确保系统完整性，并通过提供防篡改的安全锚点来实现安全的多租户操作。这些能力支撑新兴的机密计算架构，其中数据和模型不仅在静止和传输中受到保护，而且在通过CPU和加速器中的硬件强制隔离和加密执行期间也受到保护。虽然机密计算还不是AI训练或推理的默认选择，但随着标准和平台支持的成熟，采用率正在增加，特别是对于受监管或多租户环境中的推理工作负载。

共封装光学器件

随着AI模型变得越来越复杂，大型加速器集群必须协同工作以开发具有数十亿参数的前沿模型。同样，基于Agentic AI的推理解决方案需要 progressively 更大的集群规模来部署可同时操作的多个代理。为满足这种对计算和存储器永不满足的需求，数据中心正在实施高速纵向扩展网络，在集群或pod内提供高带宽连接。

到目前为止，这些计算集群一直依赖电信号来传输数据。然而，随着带宽需求增加和集群规模扩大，电连接越来越难以满足这些网络的传输距离和带宽需求。高速电信号在网络规模和距离增加时显示出显著的信号质量下降。光学器件，特别是共封装光学器件，提供了一条实现更高带宽和更节能的纵向扩展网络的途径。共封装光学器件技术利用过去十年在芯粒技术方面的进步，将硅光子解决方案集成到加速器和交换机封装上。通过将光链路定位在靠近加速器的位置，它们提供了一种节能的解决方案。

尽管取得了这些进步，共封装光学器件解决方案的实施仍面临若干挑战。虽然CMOS工艺技术和供应链已经建立，但光学技术仍处于采用的早期阶段。为光通信建立标准协议对于构建供应链生态系统合作伙伴能够轻松访问的供应链至关重要。激光模块（光学引擎的光源）和可插拔光纤模块等标准组件对于将光学器件从实验室规模扩展到大批量制造至关重要。

AI基础设施的下一阶段将受到热通量、机械翘曲、界面稳定性以及组装/可制造性的制约，其程度不亚于计算。在AI/HPC预测中，封装面积扩展到约9,000-10,000 mm2，晶粒面积趋势约为4,000-5,000 mm2，平均加速器晶粒功率预计在未来十年内将超过约5,000 W，这意味着有效功率密度至少翻倍。这些条件增加了对TIM行为在实际平整度/曲率、夹紧负载限制和循环可靠性（硅开裂、分层、互连疲劳）下的敏感性，无论是在封装级还是系统级组装期间。

展望未来，热机械设计将随着新兴的系统级晶圆/系统级面板概念达到拐点，其中模块/托盘级功率可能推向>25 kW级别，随着更深入的光子集成以减少延迟，对热稳定性和机械对齐提出了更严格的要求。实际的赢家将是那些将热学视为机械使能系统的架构：在OSAT/EMS规模上为可制造性而设计，通过与测试载具相关联的验证，针对任务概况循环进行强化，并为高效的现场服务而设计。

冷却液分配单元（CDU）托盘

冷却液分配单元托盘为高性能AI服务器管理机架级液冷。随着AI工作负载中现代AI加速器和CPU产生比传统服务器高得多的热负载，仅靠空气冷却已不再足够。CDU通过连接到CPU、AI加速器和存储器模块上冷板的闭环系统循环液冷剂来确保可靠的热管理。这不仅防止了热节流，还使密集配置中更高的持续计算性能成为可能。半导体使这些托盘中的各种泵、传感器、电路控制阀和

电源单元成为可能。例如，配备嵌入式逻辑半导体的CDU控制器板管理流速、监控压力和温度，并调节泵速和阀位以实现精确的热控制。包含半导体芯片（PHY和MAC控制器）的以太网接口转换信号并管理数据包，实现远程监控和控制，而本地LCD显示屏提供实时诊断和操作状态。

图13 CDU托盘

表13 冷却液分配单元的半导体组件

子组件	描述	数量	芯片类型
流量传感器	确保MCU与设备之间的电压兼容性	3	传感器
压力传感器	包括泵、流量计、阀门、冷板、散热器、冷却风扇和机械连接，用于在服务器机架中循环冷却剂	4	传感器
温度传感器	实时测量液体流速（L/min）	4	传感器
CDU控制器板（MCU）	监控冷却剂温度差（ΔT）	1	逻辑——MCU
以太网/PM总线接口	处理系统监控、风扇/泵控制、以太网接口	1	模拟——接口

三、芯片如何协同执行AI训练工作流

大规模训练AI模型需要一个紧密协调的专用芯片系统，每个芯片都针对工作流中的特定角色进行优化。图14说明了CPU、AI加速器、高带宽存储器（HBM）、非易失性存储器和光互连如何协同工作以管理并行工作负载、跨节点同步计算，并最终提供一个完全训练好的模型，准备部署。

图14 数据中心中的数据：芯片如何在AI训练工作流中协同工作

AI模型训练从数据摄取开始，由作为web规模数据进入系统协调器的数据处理单元处理。一旦摄取，数据被写入NVMe SSD，这些SSD针对检查点和同时供给多个加速器进行了优化。这种架构实现了快速吞吐量和训练期间的低延迟，特别是当与GPUDirect等技术配对时，允许从存储器直接访问AI加速器存储器，完全绕过CPU。

DPU还以线速率对数据进行预处理，在将其直接传递给计算单元之前进行解析和格式化。这最小化了CPU的工作负载，并确保AI加速器接收到干净、结构化的数据进行训练。SmartNIC集成的DPU可以同时支持存储卸载、网络加速和协议处理，实现更大的可扩展性和更低的功耗。

一旦数据到达AI加速器集群，通过在数千个核心上进行高度并行化计算来加速训练，由HBM支持以管理跨分布式节点的大量模型权重和激活。同时，CPU编排训练工作负载，管理同步、存储器协调和任务分配。它们处理控制流逻辑并将激活卸载以平衡存储器利用率，从而支持更大的模型和批量大小。利用互连，AI加速器以高速交换模型状态和中间激活，使大规模模型能够在许多设备上迭代高效地训练。训练完成后，CPU聚合输出并完成模型组装，为下游部署做好准备。

四、AI增长曲线：市场视角

AI应用正在快速增长，推动了对为这些工作负载提供支持的数据中心所使用的半导体的更大需求。部署在AI数据中心的半导体收入预计到2028年将达到超过1.2万亿美元，五年内增长近十倍（图15）。数据中心逻辑芯片（主要是AI加速器）的市场在2022年约为300亿美元，2024年扩大到700亿美元，WSTS预计到2026年将达到1900亿美元。

图15 AI数据中心服务器部署的半导体总收入

> 来源：德勤研究
> 注：本收入预测不包括边缘AI
> 乐观情景（35%增长）/保守情景（15%增长）

不同的AI应用需要不同的计算基础设施，对半导体硬件设计和部署有影响。虽然数据中心的大部分半导体需求一直由AI模型开发的训练驱动，但推理市场（如将训练好的模型应用于现实世界场景）正在增长。

训练和推理代表不同的计算任务。AI模型训练是数据中心处理海量数据集以学习模式并理解复杂现实世界变量的过程，需要具有显著计算能力的半导体。为AI训练设计的芯片提供极端并行性并支持巨大的存储器带宽，使模型能够数百万次甚至数十亿次调整其内部参数，直到能够做出准确的预测或建议。这需要具有先进互连、密集晶体管布局和高速度存储器层次的硬件。

训练工作负载是间歇性的且效率越来越高，这将导致对训练专用芯片的需求趋于平稳。模型架构和训练技术的进步减少了训练大型模型所需的时间和计算。此外，预训练基础模型的可用性进一步降低了对大规模训练运行的需求。

训练完成后，模型被部署用于推理——执行现实世界任务、生成对查询的响应、做出预测或识别模式。推理发生在数据中心（通常称为"云端"）或设备上（称为"边缘"，如AI PC、智能手机、车辆或工厂系统内）。在数据中心，推理芯片针对吞吐量和延迟进行优化，利用高效率计算在毫秒内执行操作。在边缘端，设备上AI加速器芯片优先考虑低延迟、小尺寸和能效，实现本地决策而不需要持续的服务器通信。鉴于AI训练与推理工作负载的独特特征，数据中心运营商越来越多地选择垂直定制的专用集成电路（ASIC）来优化性能和效率增益。

预测推理工作负载将成为AI相关芯片需求的主要驱动力。推理计算需求随着消费者对生成式AI工具的使用、企业采用以及嵌入应用中的AI代理的兴起而扩展。与超大规模训练集群不同，边缘AI部署的资源密集度低得多，依赖更小、更节能的推理优化芯片。不断增长的推理工作负载指向对定制、节能硅的需求：针对吞吐量优化的AI加速器、专用推理ASIC/FPGA、针对延迟敏感任务调整的CPU以及相关的存储器和网络芯片。

目前，训练和推理工作负载在相同的硬件组件上运行，尽管它们的工作负载特征不同。也就是说，芯片制造商正在专门设计用于推理工作负载的产品，受对更高效率或更高吞吐量需求的驱

动。随着产品差异化的推进，该行业可能会看到训练专用和推理专用硬件之间更明显的分化。训练芯片可能集中在最大化并行计算上，而推理芯片则向超低延迟、节能设计演进。

行业预测表明，推理工作负载的收入将在未来几年增长数倍，而训练收入预计将趋于平稳。例如，波士顿咨询集团估计，从2023年到2028年，训练将以30%的复合年增长率增长。相比之下，推理在同一时期将以122%的速度增长。为了具体说明这种转变，行业专家预测推理在总需求中的份额可以从2024年的20%增长到2032年的80%。

AI有望在本十年剩余时间内推动半导体行业增长的很大一部分，麦肯锡预计半导体行业年收入到2030年将达到1.6万亿美元，主要受AI和数据中心的推动。2025年半导体市场为7917亿美元，生成式AI可能因此代表该行业未来几年增长的40%以上。

五、巨额支出：半导体内容价值分析

为满足全球AI需求，预计在2023年至2030年的八年内，将累计投资4.0万亿美元用于建设AI数据中心，其中预计高达2.8万亿美元将用于AI服务器设置的半导体和其他硬件。在典型的AI数据中心中，大部分资本支出位于计算基础设施，占AI数据中心总资本支出的50%以上。这项投资集中在AI服务器机架中——这是一种模块化、高价值的系统，每个成本为150万至400万美元，专为处理AI工作负载的独特需求而设计。一个领先的数据中心可容纳多达10,000个机架。

为进一步了解这项投资将落在半导体价值链的哪个环节，我们分析了标准、行业级AI服务器内的半导体内容，以及这些内容如何转化为全球半导体供应链的经济价值。（注：图16基于单个AI数据中心机架，价值在150万至400万美元之间。所示百分比代表该机架内组件内容的大致份额。）

图16 AI数据中心机架中半导体内容价值的份额

逻辑和存储是数据中心组件中部署最广泛的芯片领域，合计占半导体价值的85%以上。对于现代、领先的AI服务器机架，半导体约占内容价值的95%。在组件层面，AI加速器占AI数据服务器机架中半导体内容的70%以上。每个GPU模块将一个或多个大型、领先节点逻辑晶粒、HBM存储器堆栈、中介层和有助于整体价值的专用互连进行共同封装。CPU在架构上也至关重要，占8%，DPU进一步增加2%。总体而言，逻辑芯片占服务器机架中半导体内容的65%，还包括AI加速器、CPU、DPU、NIC控制器、BMC和ASIC。

存储和存储半导体占机架总半导体内容价值的20%以上，由集成在各种服务器托盘中的HBM、DRAM和NAND闪存器件组成。这些数字反映了现代AI工作负载的架构性质：庞大的矩阵、分布式模型和大型训练集，不仅需要原始计算能力，还需要极端的存储器带宽和低延迟数据访问。

正如前面提到的，整个AI数据服务器的稳定性还依赖于一长串低成本、当前代节点组件和新兴材料创新，它们占机架半导体内容价值的剩余约10%。虽然单个成本较低，但这些芯片对电压稳定性、信号完整性和热控制至关重要，特别是在具有兆瓦级功率密度的紧密堆积机架中。此类别包括模拟半导体，如电源管理IC、电压稳压器和时钟发生器，确保可靠的电源传输和信号时序。它还包括化合物半导体（如GaN和SiC功率器件），实现液冷泵和高电压分配的高效率功率转换。这些支持芯片共同提供了使高价值逻辑和存储组件能够大规模执行所需的弹性、效率和余量。

如前所述，给定服务器内的半导体价值组成将在一定程度上因某些因素而异（方法论和更多细节见附录）。然而，这种变化主要由特定的逻辑芯片选择驱动。集成尖端AI加速器的服务器机架表现出65%至75%的逻辑芯片组成价值，而使用中端AI加速器的机架仅显示40%至50%的逻辑价值，余量转向存储和存储组件。价格较高的系统通常具有更大的计算密度和更高的峰值性能。

> **性能-实用主义权衡**
>
> 系统集成商和超大规模云服务商通过首先建立性能、容量、功耗和总拥有成本（TCO）的要求来设计AI平台。这些要求由预期的工作负载组合、部署规模和运营约束驱动，并指导结构化的组件选择过程，而非单一的"最佳芯片"结果。从这个角度来看，加速器选择反映了一系列审慎的权衡：
>
> - **性能目标（TFLOPs）：**并非所有AI工作负载都需要尖端GPU提供的最大计算吞吐量。对于许多训练和推理用例，中端加速器在提高利用效率和系统平衡的同时提供足够的性能。
> - **总拥有成本（TCO）：**加速器选择是整体评估的，综合考虑采购成本、功耗、冷却要求、机架密度和预期运营寿命。对于不受性能限制的工作负载，低功耗或中端加速器可以提供更优的TCO。
> - **容量和可用性考虑：**供货周期、平台兼容性和部署时间表影响加速器选择，特别是在大规模情况下，可预测性和可重复性至关重要。

到目前为止，计算托盘占数据服务器中大部分半导体足迹——约95%的内容价值，近90%的半导体晶粒总数，以及超过80%的单个芯片数量（图18）。

图17 每种托盘类型的工艺节点技术价值分布

先进节点晶粒几乎完全集中在计算托盘中以处理AI工作负载，而其他处理非核心功能的托盘主要依赖成熟和遗留节点。

图18 每种托盘类型的内容价值和晶粒/芯片数量份额

托盘的价值分布与半导体技术的先进程度或工艺节点高度相关。本报告将先进节点定义为sub-10纳米代，通常使用极紫外（EUV）光刻制造。除了计算托盘占半导体晶粒的大部分外，这些晶粒中的大多数都采用先进工艺技术制造（图18）。

加速器互连托盘和IPMI托盘也利用了一些先进的晶粒，但这些托盘的整体价值要低得多，因为存在的晶粒数量较少，而且用于制造这些晶粒的技术节点相对成熟。

价值集中与成本动态

虽然少量逻辑芯片占AI数据服务器价值的大部分，但按数量计算，绝大多数组件位于低成本端，但对系统功能同样至关重要（图19）。

图19 AI数据中心中跨价格区间的芯片分布

> 仅3%的高价值芯片就驱动了服务器大部分内容价值。服务器中超过一半的芯片价格低于$10。

这种在领先边缘的价值集中与更广泛的半导体 spectrum 中的关键依赖之间的二元性，突显了塑造AI基础设施未来的机遇和脆弱性。

六、从设计到数据中心：了解AI数据中心的全球供应链

到本报告这一刻，我们已经将AI数据中心服务器中的芯片拆解到最详细的程度。然而，要制造这些芯片中的一个，需要整合整个生态系统。数据中心芯片制造需要在全球分布的供应链中经历数千个步骤。该过程涉及两个主要供应链环节：一个用于从初始设计创建封装芯片，另一个用于将这些芯片集成到准备好进行数据中心部署的系统中。

AI服务器不是即插即用的。它们必须插入预先配置的电源域，匹配最小化延迟的网络拓扑，并连接到确保吞吐量的分布式存储集群。设施运营商、云超大规模云服务商和基础设施提供商都在这一部署编排中发挥作用，并经常与半导体提供商密切合作以确保满足计算性能要求。随着AI基础设施继续扩展，供应链预计将交付从第一天起就准备好运行复杂工作负载的芯片和完整集成系统。

如图20所示，旅程始于半导体设计，工程师使用电子设计自动化（EDA）工具来架构和模拟集成电路和印刷电路板（PCB），确保制造前的功能正确性和可制造性。接下来，在半导体制造中，设计在洁净室环境中使用各种复杂的原材料和先进设备（如光刻和蚀刻工艺）印在硅晶圆上。在封装阶段，通过严格测试的晶粒使用先进的2.5D/3D技术进行高性能芯片互连，而标准芯片则使用传统方法（如引线键合）。结果是完全封装和电气验证的芯片，准备好集成到系统中。芯片通过电气验证后，进入模块和电路板组装，封装晶粒安装在PCB上，如DIMM（用于DRAM）、SSD电路板（用于NAND）或计算电路板（用于AI加速器）。系统集成和最终组装涉及将计算、存储、网络和热管理系统组合成完全功能的服务器单元。单元的精确配置根据客户对性能、冷却和电力传输的要求进行调整。

图20 从设计到数据中心

阶段	关键位置	说明
1. 产品设计	美国、韩国	EDA解决方案，用于架构和模拟IC和PCB
2.1 晶圆制造	台湾、韩国、美国	将设计印在硅晶圆上；高性能AI芯片（如加速器、HBM、CPU）的关键阶段
2.2 晶圆测试与晶粒切割	台湾、韩国、马来西亚、越南、菲律宾和美国	正常芯片（如NAND、DRAM、旧代CPU）的测试和切割
3.1 先进封装	台湾、韩国、马来西亚和美国	高性能AI芯片的2.5D/3D封装技术

3.2 传统封装	台湾、中国大陆、韩国、马来西亚、越南和菲律宾	标准芯片的引线键合等传统方法
4. 模块/电路板组装	中国大陆、台湾、越南	将封装晶粒安装在DIMM、SSD板、计算板上
5. 系统集成与测试	马来西亚、中国大陆、越南、台湾	组合计算、存储、网络和散热系统
6. 安装与部署	全球	将服务器安装到数据中心设施中

七、AI基础设施的新兴前沿

随着AI数据中心的成熟，一系列技术转变和市场力量正在重塑基础设施的设计、部署和维持方式。本节重点介绍开始推动下一阶段增长的关键趋势：

边缘需求的影响

推理需求正在从集中式数据中心转向结合云、边缘和设备上计算的混合模型。边缘AI正在通过将工作负载从集中式数据中心重新分配到分布式环境来重塑计算需求。到2025年，预计超过50%的数据将由边缘设备生成，其中许多设备被设计为独立运行AI模型，而不依赖云基础设施。这一趋势通过采用可在本地部署的较小、任务特定模型而得到加强。混合计算正成为企业战略的核心要素，由对更快响应时间、减少网络依赖和更强隐私保护的需求驱动。不断扩展的边缘推理可能将市场 segment 拉向低功耗、成本优化的半导体，如嵌入式AI SoC和NPU、微加速器以及专用PMIC。这一趋势也可能维持对当前代和成熟节点芯片、传感器和连接硅的需求，因为边缘部署优先考虑体积、集成和效率而非原始性能。

工艺技术演进

半导体在密度和效率方面继续进步，但原始晶体管扩展的速度正在放缓。该行业正从FinFET转向纳米片FET设计，这改善了功率控制并减少能量损失。与此同时，研究人员正在探索二维半导体和化合物半导体等新材料的潜力，以将性能扩展到硅的极限之外。随着仅靠扩展已不再足够，进步越来越多地来自系统级创新。虽然3nm技术目前正在推动晶体管扩展的边界，但未来节点可能包含新材料，如纳米片FET、全环绕栅极（GAA）晶体管以及高数值孔径极紫外（High-NA EUV）等先进光刻技术。这些方法使芯片能够在摩尔定律放缓的情况下提供更大的性能、能效和成本效益。

架构与封装转变

另一个重大转变是向领域特定架构的迁移，该架构紧密集成了高性能逻辑、存储器和互连。未来的AI工作负载将越来越多地利用模块化设计，允许异构组件（即"芯粒"）以最小延迟惩罚进行共同封装。这一架构转变得到了2.5D和3D封装进步的支持，这些进步实现了计算和存储元件的更紧密集成，同时改善了功率效率和热管理。光子学与共封装光学（CPO）也是一项新兴突破，将光互连直接嵌入交换机ASIC或加速器，以提供更高带宽和每位更低能耗。光互连更广泛地被定位为下一代延迟降低和节能解决方案，特别是对于具有多节点AI加速器架构的大规模训练集群。

存储与数据移动

HBM和下一代DRAM的持续创新至关重要。仅HBM市场预计到2025年将达到210亿美元，凸显了其在实现高带宽计算方面的关键作用。未来几代HBM将具有更高的堆叠高度、改进的能效以及与计算核心更紧密的集成。同样，新的存储器层次结构预计将出现，结合SRAM、DRAM和持久存储器，以实现优化的数据局部性和减少的数据移动。

硬件-软件协同设计

另一个关键趋势是跨硬件和软件栈的日益集成和优化，为AI工作负载驱动更高的效率。通过端到端控制硅架构、封装和部署策略，这些组织和超大规模云服务商可以针对特定性能、延迟或能耗目标调整系统。这一趋势预计将影响市场动态，减少对现成解决方案的依赖，并增加内部半导体能力的战略重要性。

能源与可持续发展要务

随着AI基础设施的能源需求增加，能效和可持续发展将成为更紧迫的问题，推动对节能芯片设计、动态工作负载分配和智能冷却系统的更大投资。基础设施提供商需要优先考虑每瓦能源的计算效率，主要是作为节省成本的措施，也是应对全球可持续发展目标和能源市场波动的需要。主动投资适应性、节能系统的运营商将更好地 positioned to sustain long-term value as AI deployments diversify across sectors and geographies。

随着每一代新数据中心的推出，AI计算正变得显著更节能，这一趋势正在重塑现代数据中心的经济学和架构。半导体硬件能效的持续改进（从计算相关芯片到管理冷却、稳定电力需求、分配和转换电力以及组织整个设施工作负载的芯片）是释放AI潜力的关键。每瓦性能的提升将使数据中心运营商能够更密集地打包服务器并在不成比例增加能源的情况下提升AI性能。随着芯片公司开发降低每计算单元能源需求的解决方案，一个10 MW的数据中心配合更高效的芯片最终可能提供与使用早期代半导体技术的50 MW数据中心相同的处理能力。随着电力成本和电网需求的上升，更高的能效不再是一个"加分项"——它是大规模部署AI的唯一途径。

最后，整体AI能效不仅由硅本身决定，还由端到端数据中心系统决定。这包括电力如何引入设施（这是一个以可与芯片开发媲美的速度发展的领域），以及功率转换损耗、热管理效率、利用率和工作负载放置。AI数据中心的电力传输架构现在几乎每年都在变化，在某些情况下甚至更快，共同决定了每瓦传输电力的有效计算量。

八、结论

本报告说明了支撑AI基础设施的半导体供应链中输入的多样性——超越了高端处理器和加速器，延伸到信号调理芯片、电源稳压器、存储晶粒和控制逻辑。随着AI继续扩展，维持有竞争力的半导体创新系统的能力将是AI持续进步的基础。整个半导体供应链，涉及美国半导体设计、制造和制造设备公司，使AI基础设施的建设成为可能。简而言之，没有半导体，就没有AI。要在这项变革性技术中保持领先，政府和行业必须共同努力推进政策以加速半导体创新——加强全谱芯片技术的能力——并与全球合作伙伴密切合作，建立强大和有弹性的供应链。

作者简介

David Isaacs

政府事务副总裁
disaacs@semiconductors.org

David Isaacs是SIA政府事务副总裁，负责协会与政府政策相关的所有工作。

Jeroen Kusters

主要负责人
jekusters@deloitte.com

Jeroen Kusters是Deloitte Consulting LLP的美国半导体负责人和合伙人，负责监督半导体行业业务，为跨多个领域的半导体客户提供建议。

Mary Thornton

全球政策副总裁
mthornton@semiconductors.org

Mary Thornton是SIA副总裁，负责SIA的全球政策和经济安全议程。

Duncan Stewart

研究总监
dunstewart@deloitte.ca

Duncan Stewart是Deloitte Canada TMT研究总监，是Deloitte US TMT中心和Deloitte Global半导体主题的首席研究员。

Greg LaRocca

市场研究与经济政策总监
glarocca@semiconductors.org

Greg LaRocca是SIA市场研究与经济政策总监，负责监督行业数据分析。

Amy Scimeca

经理
aseiler@deloitte.com

Amy Scimeca是Deloitte Consulting LLP经理，为半导体客户提供市场战略、价值捕获和供应链设计业务支持。

Erik Hadland

技术政策总监
ehadland@semiconductors.org

Erik Hadland是SIA技术政策总监，负责协会的研究、开发和技术活动以及教育和劳动力发展工作。

Karan Aggarwal

经理
karaaggarwal@deloitte.com

Karan Aggarwal是Deloitte Consulting LLP经理，领导跨美国和亚洲半导体客户的大型战略和企业转型项目。

Pranav Salhan

顾问
psalhan@deloitte.com

Pranav Salhan是Deloitte Consulting LLP顾问，为半导体客户在运营优化、供应规划和企业转型计划方面提供支持。

致谢

我们感谢SIA成员公司和主题专家为本报告提供反馈所做出的宝贵贡献。

本报告的完成离不开SIA同事Aaron Woolf、Alex Gordon、Emma Rafaelof和Molly O'Leary的贡献。

关于SIA

美国半导体行业协会（SIA）是半导体行业的代言人，是美国最大的出口行业之一，也是美国经济实力、国家安全和全球竞争力的关键驱动力。SIA代表了美国半导体行业收入的99%和近三分之二的非美国芯片公司。通过这一联盟，SIA致力于通过与国会、政府和全球关键行业利益相关者合作，鼓励推动创新、促进商业和驱动国际竞争的政策，来加强半导体制造、设计和研究的领导地位。了解更多请访问 www.semiconductors.org。

关于德勤

Deloitte指Deloitte Touche Tohmatsu Limited（一家英国私人担保有限公司）及其成员公司网络中的一个或多个。请访问 www.deloitte.com/about 了解Deloitte Touche Tohmatsu Limited及其成员公司的法律结构详细描述。请访问 www.deloitte.com/us/about 了解Deloitte LLP及其子公司的法律结构详细描述。根据公共会计的规则和法规，某些服务可能无法向鉴证客户提供。

术语表

缩写	定义（AI数据中心语境）	缩写	定义（AI数据中心语境）
2.5D	2.5维——先进的芯片封装技术，在硅中介层上组合多个晶粒以增强AI计算性能	I/O	输入/输出——对AI数据处理至关重要，影响数据中心的存储器带宽和吞吐量
3D	三维——指用于增加AI硬件密度和效率的堆叠芯片架构或3D封装	推理	将训练好的AI模型应用于新数据以进行预测或决策，通常实时进行
先进封装	2.5D和3D集成等技术，用于组合多个芯片以提高性能	中介层	在先进半导体封装中连接多个芯片的层
AI数据中心	专为支持计算密集型AI任务（如大规模训练和推理）而设计的设施	MCU	微控制器单元——帮助管理数据中心内环境控制和监控的嵌入式系统
ASIC	专用集成电路——为加速特定AI任务而设计的定制芯片，广泛用于超大规模AI数据中心	NAND	与非门——一种广泛用于数据中心存储系统的非易失性闪存
BMC	基板管理控制器——用于监控AI数据中心服务器健康和性能的嵌入式控制器	NIC	网络接口卡——使AI数据中心中的服务器能够通过高速网络高效通信的硬件
CDU	冷却液分配单元——AI数据中心液冷系统中的关键组件，用于热效率	nm	纳米——半导体技术节点的测量单位；更小的尺寸提高AI芯片的性能和能效
芯粒	与其他模块集成以形成完整芯片系统的模块化半导体单元	NVMe	非易失性存储器Express——提高SSD访问速度的接口协议，对AI工作负载中的大数据处理至关重要
CPU	中央处理单元——在AI数据中心中用于编排和管理AI工作负载及支持传统任务	OEM	原始设备制造商——为其他公司品牌生产AI数据中心设备的公司
DRAM	动态随机存取存储器——在AI数据中心中常用于模型执行期间的临时数据存储	PCB	印刷电路板——在数据中心基础设施中组装AI服务器组件的基础
EDA	电子设计自动化——用于设计AI数据中心AI芯片的工具	PCIe	外围组件互连Express——在AI数据中心服务器中连接GPU和加速器的高速接口标准
边缘AI	在本地设备上运行AI模型以减少延迟和对云基础设施的依赖	光子计算	使用光信号实现更快数据传输和更低功耗的计算

EUV	极紫外——用于AI计算节点高性能芯片的光刻技术	PHY	物理接口——实现数据中心AI芯片与存储器或互连之间高速连接的层
晶圆厂	制造——指半导体制造，是生产AI数据中心用AI处理器的基础	PMIC	电源管理集成电路——管理AI数据中心AI芯片的电压和功率效率
Fabless	设计芯片但外包制造的公司	RAM	随机存取存储器——在AI数据中心训练和推理期间保存工作数据和模型的关键组件
FET	场效应晶体管——现代数据中心AI加速器供电芯片中的基本组件	SoC	片上系统——将多种功能整合到单个晶粒上的集成电路，越来越多地用于AI加速器和边缘AI以提高性能、功率效率和系统集成
代工厂	根据第三方设计制造芯片的半导体制造设施	SRAM	静态随机存取存储器——AI芯片中用于存储中间模型权重或数据的高速缓存
FPGA	现场可编程门阵列——可重构芯片，在AI数据中心中实现高效、任务特定的加速	训练（AI）	AI模型使用计算资源从大型数据集中学习模式的阶段
GAA	全环绕栅极——增强AI硬件性能和能效的先进晶体管架构	TSV	硅通孔——实现3D芯片堆叠的垂直互连，提高AI计算模块的性能和密度
GPU	图形处理单元——AI加速器的一个子类	VRM	电压稳压模块——调节AI服务器中处理器的电源传输以实现稳定运行
HBM	高带宽存储器——支持现代数据中心芯片高速AI训练工作负载的存储器类型		

参考文献

1. Krystal Hu, "ChatGPT sets record for fastest-growing user base—analyst note," Reuters, February 2, 2023.
2. Fatima Hameed Khan, Muhammad Adeel Pasha, and Shahid Masud, "Advancements in microprocessor architecture for ubiquitous AI—an overview on history, evolution, and upcoming challenges in AI implementation," Micromachines 12, no. 6 (2021): p. 665.
3. Jesse Noffsinger et al., "The cost of compute: A $7 trillion race to scale data centers," McKinsey Quarterly, April 28, 2025.
4. TrendForce News, "Foundry giants' advanced node expansion efforts beyond 2025: TSMC, Intel, Rapidus and more," January 13, 2025.
5. World Semiconductor Trade Statistics, 2025 Product Classification Guide, December 11, 2024.
6. Ondrej Burkacky et al., "Generative AI: The next S-curve for the semiconductor industry?," McKinsey & Company, March 29, 2024.
7. SK Hynix, "HBM technology: The silent catalyst of the AI revolution," January 30, 2024.
8. Intel, "Field programmable gate arrays (FPGAs) for artificial intelligence (AI)," accessed August 23, 2025.
9. Future Market Insights, FPGA market report, August 23, 2025.
10. Manuel Mota, "Die-to-die connectivity with high-speed SerDes PHY IP," IBM Semiconductor Engineering, March 12, 2020.
11. Matteo Buffolo et al., "Review and outlook on GaN and SiC power devices: Industrial state-of-the-art, applications, and perspectives," IEEE Transactions on Electron Devices 71, no. 3 (March 2024): pp. 1344-55.
12. Nidhi Govil, "Nvidia unveils 800V high-voltage DC power system for next-gen AI data centers," The Outpost, May 28, 2025.
13. Amr Elmeleegy, "NVIDIA contributes NVIDIA GB200 NVL72 designs to Open Compute Project," NVIDIA Developer Blog, October 15, 2024.
14. Duncan Stewart et al., "Silicon photonics: Gen AI communicates at lightspeed," Deloitte, November 19, 2024.
15. World Semiconductor Trade Statistics (WSTS), "H2-2025 Forecast," December 2, 2025.
16. Jim Rowan et al., "State of Generative AI in the Enterprise: Quarter four report," Deloitte, January 2025.
17. CloudSyntrix, "The impact of Generative AI on hyperscalers, chipmakers, and inference workload," February 10, 2025.
18. SambaNova Systems, "SN40L RDU AI Chip," accessed October 2025.

19. Qualcomm, "Cloud AI 100 Ultra," Qualcomm Technologies, accessed October 2025.

20. David Crawford, Jue Wang, Roy Singh, "AI's Trillion-Dollar Opportunity," Tech Report 2024, Bain & Company, 2024.

21. Intel Corporation, "Gaudi AI Accelerator Products," Intel, accessed October 2025.

22. Amazon Web Services (AWS), "AWS Trainium," accessed August 23, 2025.

23. Vivian Lee et al., "Breaking barriers to data center growth," Boston Consulting Group, January 20, 2025.

24. Walters et al., "Rethinking AI demand part 1: AI data centers are experiencing a surge of training demand—what happens when the surge is over?," Alvarez & Marsal," February 25, 2025.

25. McKinsey & Company, "Hiding in plain sight: The underestimated size of the semiconductor industry," January 16, 2026.

26. SIA/WSTS, "Global Annual Semiconductor Sales Increase 25.6% to $791.7 Billion in 2025," Feb. 7, 2025.

27. Jesse Noffsinger et al., "The cost of compute: A $7 trillion race to scale data centers," McKinsey Quarterly, April 28, 2025.

28. Uvation, "Why AI server cost per user is the new metric that matters," July 4, 2025.

29. Josh Schneider and Ian Smalley, "What is a neural processing unit (NPU)?," IBM, accessed August 23, 2025.

30. Adita Agrawal, "The convergence of edge computing and 5G," Control Engineering, August 7, 2023; Baris Sarer et al., "AI and the evolving consumer device ecosystem," CIO Journal by Deloitte for The Wall Street Journal, April 24, 2024.

31. Chris Thomas et al. "Is your organization's infrastructure ready for the new hybrid cloud?," Deloitte Insights, June 30, 2025.

32. Murray Slovick, "From FinFETs to Nanosheets: ICs evolve to keep pace with 'Moore's Law'," The IP&E Specialist, July 6, 2021.

33. Slovick, "From FinFETs to Nanosheets: ICs evolve to keep pace with 'Moore's Law'." July 6, 2021.

34. Vishnu Kumar, "Global semiconductor market to grow 14% in 2025, Gartner report indicates," Circuit Digest, October 29, 2024.

35. Walters et al., "Rethinking AI demand part 1: AI data centers are experiencing a surge of training demand—what happens when the surge is over?" February 25, 2025.

36. Martin Stansbury et al., "Can US infrastructure keep up with the AI economy?," Deloitte Insights, June 24, 2025.