DRIVENETS

采用DriveNets Network Cloud-AI解决人工智能后端网络瓶颈

白皮书

网络瓶颈会减慢应用的速度，妨碍新用例和业务的可行性，这似乎是早已让人司空见惯的认知。

在拨号调制解调器时代、2G时代和早期3G时代都是这种情况。任何多媒体应用，无论是视频会议、流媒体还是具有高级图形的在线游戏，都只能局限于本地或园区网络。这是因为，远程连接（特别是WAN和访问）要么使得这些应用无法使用，要么使得体验质量非常差。

但这都是老黄历了。从那之后，xDSL、FTTX、4G、5G等技术消除了主要的网络瓶颈，不仅在应用领域带来了巨大的进步，还推动了OTT（Over-The-Top）和基于云的应用的兴起，从根本上消除了LAN和WAN之间的界限。

当今的网络瓶颈

如今，大多数应用的使用都不受位置或内容的限制影响。因此，当用户组处理文档、观看电影或玩游戏时，无论文件存储在本地还是共享驱动器上，无论观看的是在线4K电影还是本地设备存储的电影，用户都能获得一致的体验。

那么，除了旧话重提，我们此刻为什么还要来探讨网络的瓶颈呢？

原来，网络的进步带来的一些用例是这种进步根本不足以应对的，于是网络再次成为瓶颈。

这听起来似乎有些反常识，但再想想看并行计算用例吧。

高性能计算（HPC）

在并行计算中，多个计算设备（例如CPU、GPU）之间的连接通过服务器内总线或集群内网络运行。做到这一点只是因为现在的网络基础设施可以达到与内部I/O总线相同的速率和性能。

对于所有需要运行大规模计算任务的人来说，这都是个好消息。

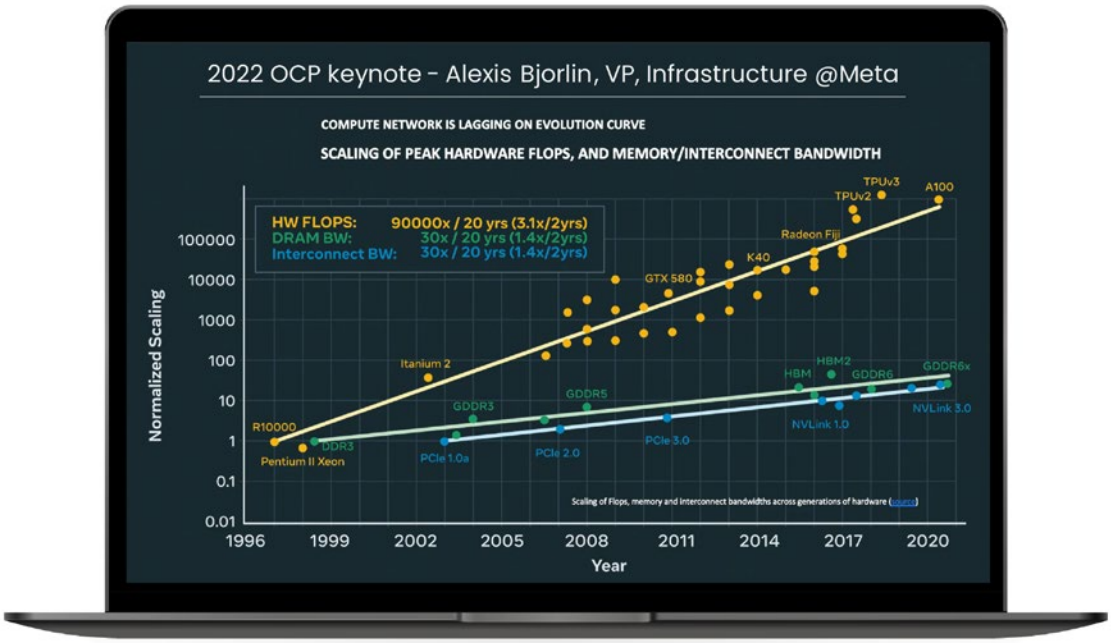
计算网络差距：下一个网络瓶颈

外部和内部I/O机制能够提供（大致）相同的带宽，因而它们之间开始具备可互换性。以PCIe 5.0为例，它能够支持128GB/s（1024Gbps），而以太网能够支持800Gbps，两者大致相同。

这种可互换性促进了并行计算的发展，让其中大量计算元件（主要是GPU）能够在同一计算作业上并行工作。这种服务器集群基本上是一台非常大的计算机，或者说超级计算机。服务器内I/O协议（例如PCIe和NVLink）和服务器间I/O协议（例如以太网和InfiniBand）在这方面起着类似且同等重要的作用。

然而，事实证明，正如Meta的工程、基础设施副总裁Alexis Bjorlin在2022年开放计算项目（OCP）全球峰会的主题演讲中所描述的那样，这种架构中出现了一种新的差距。

这种不断增长的新差距存在于计算能力/容量（以FLOPS为单位）与内存访问和互连的带宽之间，具体如下面的时间线图所示。



来源：2022年OCP主题演讲——Meta基础设施副总裁Alexis Bjorlin

这种差距再次使网络成为瓶颈。

在计算过程严重依赖于服务器间或GPU间连接的系统中，这一瓶颈则变得更加严重。在人工智能集群中就是这种情况，尤其是在大规模集群中，这种组网性能滞后会导致GPU空闲周期。

在这种情况下，显著增长的硬件算力会因组网而降级。如果用户拥有一款非常强大的GPU，却因为要等待同一集群中另一个GPU的信息，而不得不有一半以上的时间都处于空闲状态，那就真是太令人遗憾了。这种延时是由于互连网络中的延迟、抖动或数据包丢失造成的。

如前所述，这在人工智能集群网络中极其重要。幸运的是，有几种解决方案可以以某种方式应对这一痛点，而其中一些解决方案会比其他的更具优势。

利用DriveNets Network Cloud-AI解决瓶颈

人工智能后端网络具有规模性——单个集群中有1000个400/800Gbps端口，以及在线性——受益于前端和后端的相同技术，目标是优化其计算（GPU）资源以实现最快的作业完成时间（即最佳JCT性能）。因为这些，后端网络产生了一些独特的需求。

如果要尝试对这些需求进行分类，以此为依据来评估不同的解决方案，那么以下三类则至关重要：

架构灵活性

- 多个、多样化的应用
- 支持扩展
- Web连接（与孤立的HPC不同）

大规模高性能

- 支持扩展
- 超过机箱限制的大规模GPU部署
- 通过灵活性、高可用性、最小爆炸半径等实现最快JCT。

值得信赖的生态系统

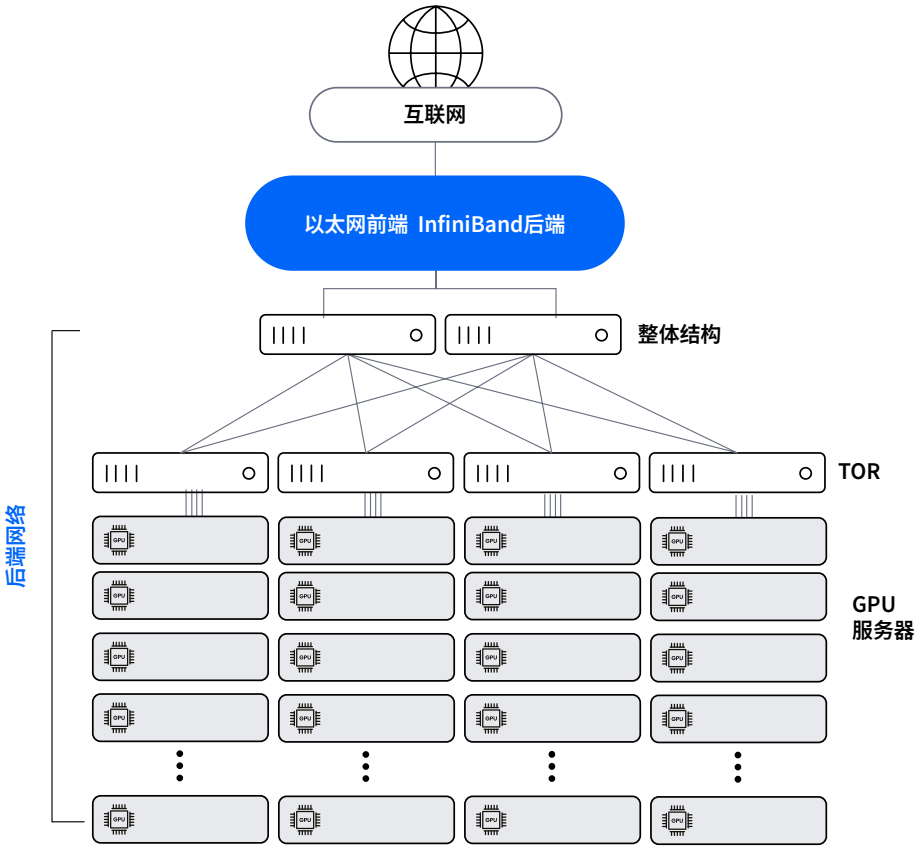
- 标准接口，可实现多供应商混合搭配——避免硬件/ASIC供应商锁定
- 经过现场验证的互连解决方案——降低风险

人工智能的行业解决方案及其缺点

人工智能后端网络有几个著名的行业解决方案 。

非以太网（例如英伟达的InfiniBand）

这种半专有的非以太网解决方案作为一种具有无损性和可预测性的架构，能够提供卓越的性能，实现满足需求的JCT性能。而另一方面，它在组网方面和GPU方面又会导致实际的供应商锁定。同时，它还缺乏及时调整不同应用的灵活性，需要特定的操作技能，还创建了无法在相邻前端网络中使用的隔离性设计 。

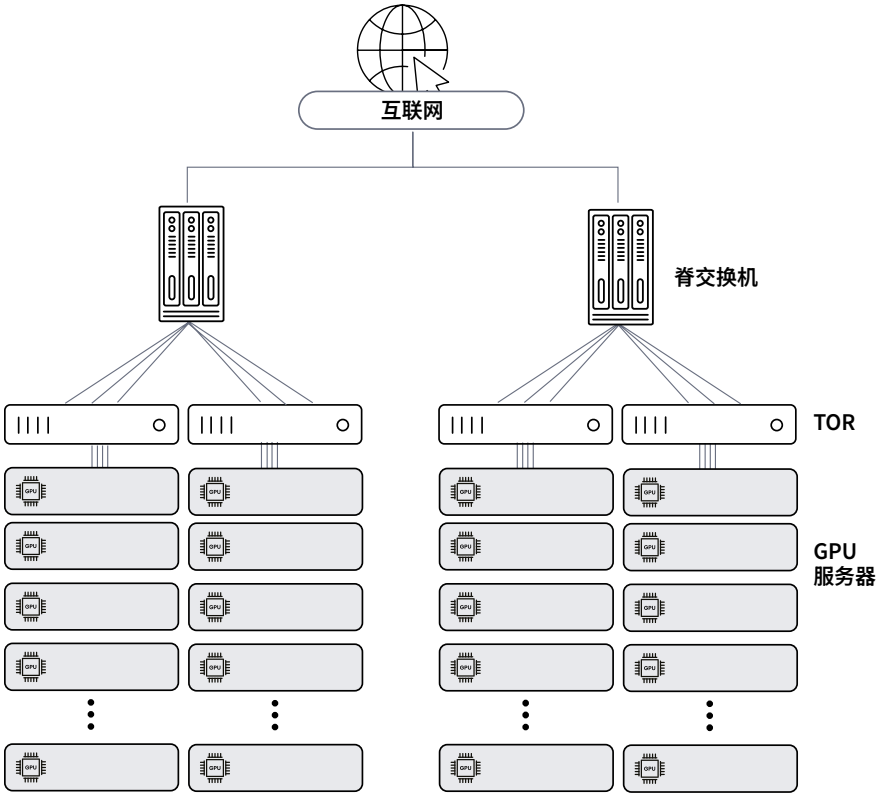


以太网——Clos架构

以太网是组网领域事实上的标准，因而它的规划和部署非常容易。

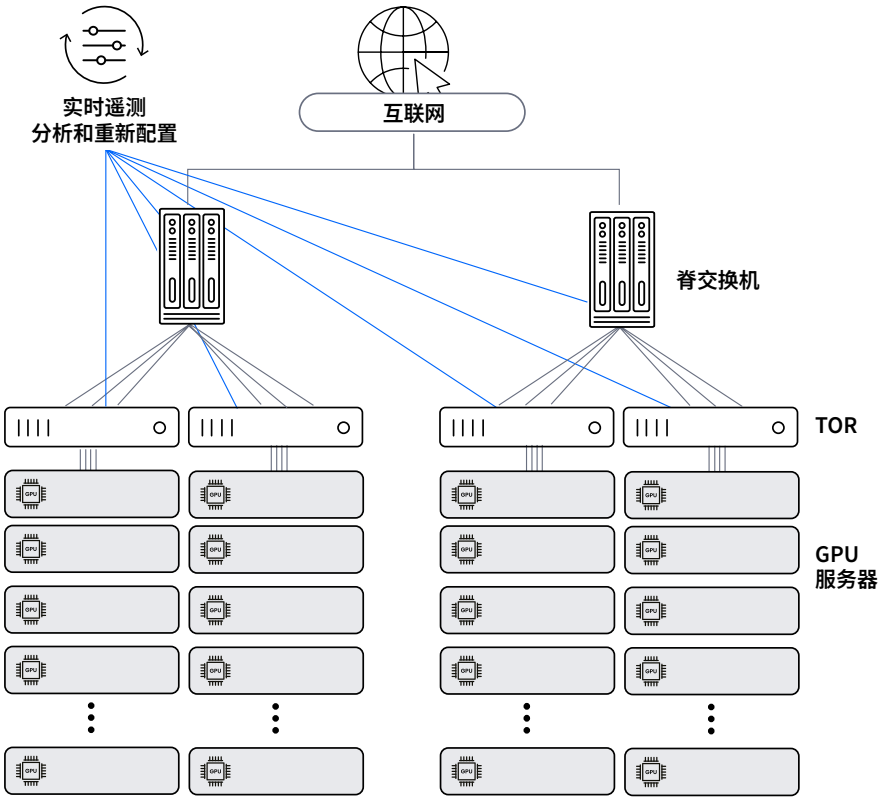
当构建在Clos架构（具有ToR叶层和基于机箱的脊层）中，它实际上会有无限的规模。但另一方面，其性能会随着规模的增长而下降，其固有的延迟、抖动和丢包会导致GPU空闲周期，从而降低JCT性能。

庞大的规模还会导致管理的复杂性，因为每个节点（叶或脊）都需要单独管理。



以太网——具有增强型遥测的Clos架构

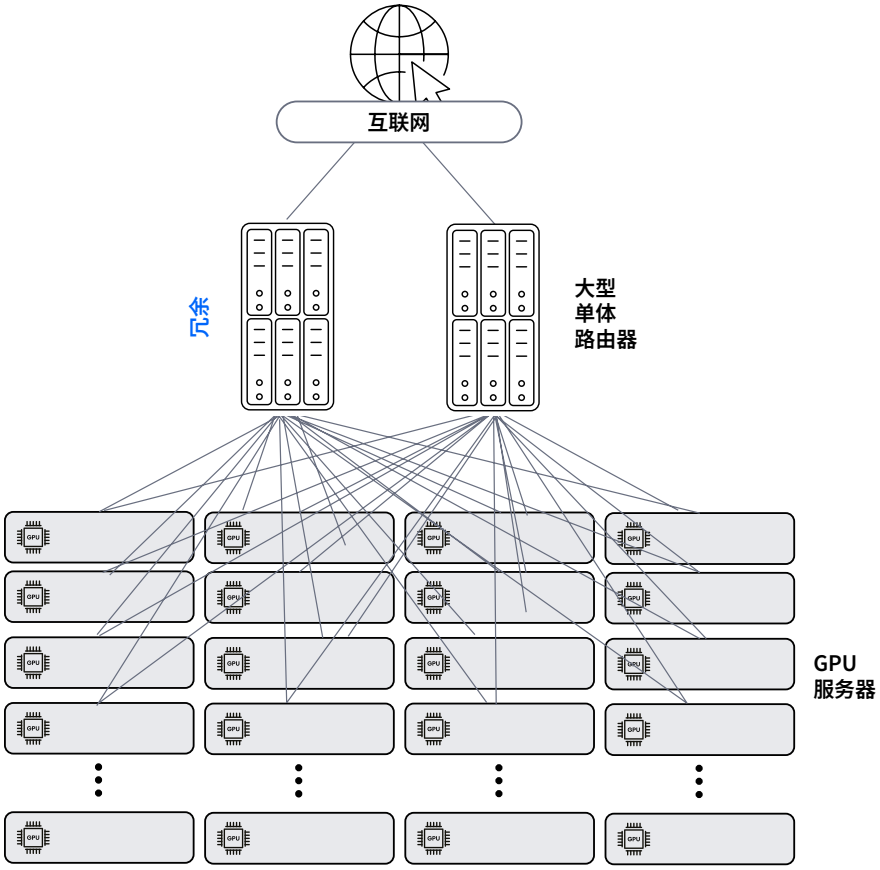
增强型遥测可以通过监控整个网络的缓冲区/性能状态并主动监管流量，在一定程度上提高Clos架构以太网解决方案的性能。尽管如此，这种解决方案仍然缺乏大规模人工智能网络所需的性能。



以太网——单机箱

机箱能够将从任意GPU到任意GPU的以太网跳数减少到一，从而解决了多跳Clos架构的固有性能问题和复杂性。但是，

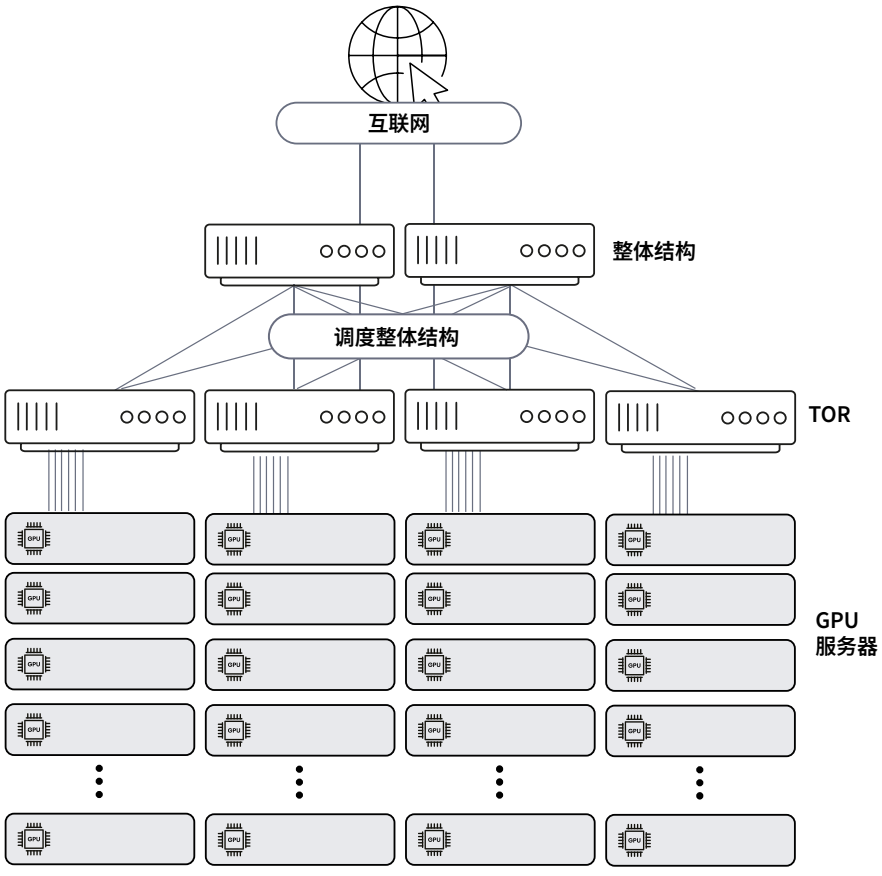
它无法根据需要进行扩展，并且还带来了复杂的布线管理难题。



以太网——Distributed Disaggregated Chassis (DDC)

最后，DDC提供了一个两全其美的解决方案。它创建了一种单跳以太网架构，具有非专有性、灵活性和可扩展性（最多可扩展至32,000个800Gbps端口）。它能够为工作负载实现JCT效率，因为它不仅能够提供无损的网络性能，还能保持易于构建的Clos物理架构。

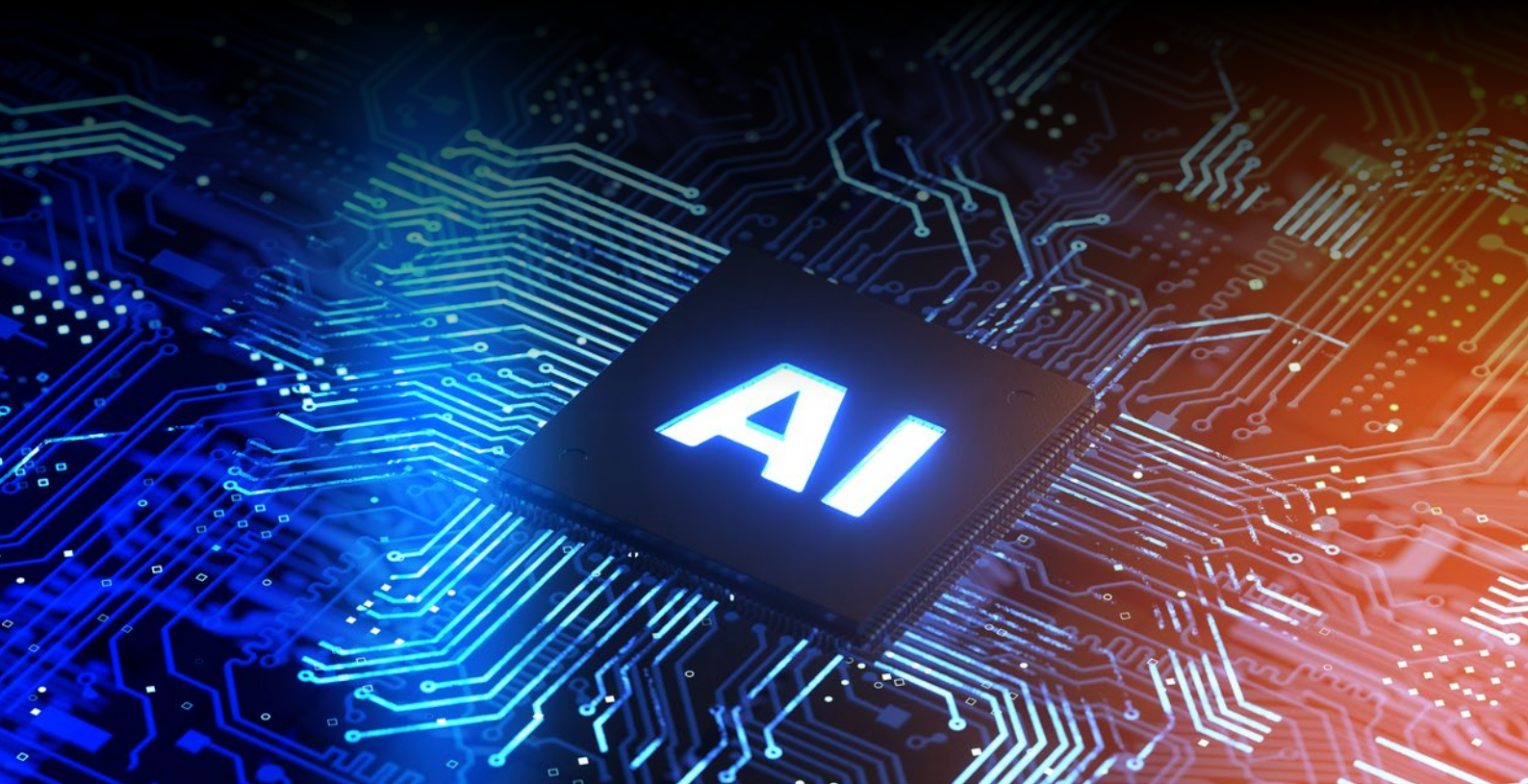
在这种架构中，叶脊都是同一个以太网实体，它们之间的整体结构连接是基于单元的、可调度的且有保证的。



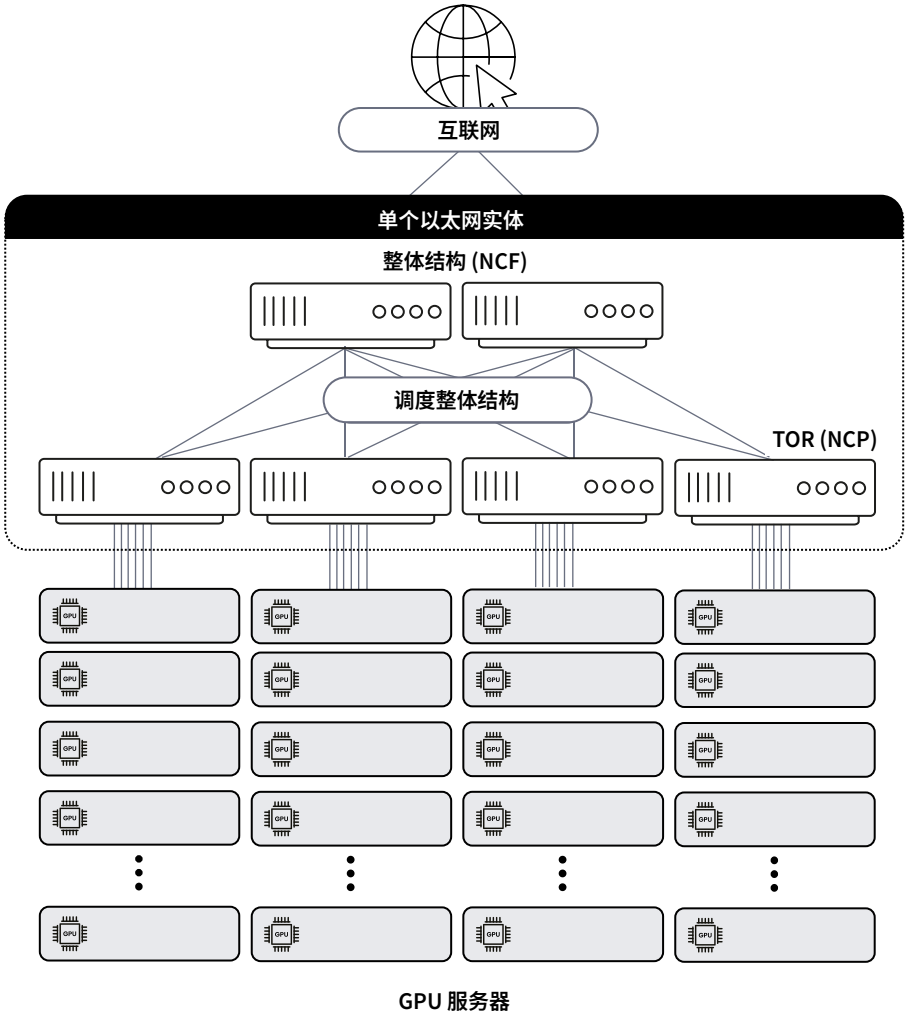
DDC ——人工智能组网的最佳解决方案

下表根据之前定义的类别总结了上述各种解决方案：

解决方案	专有	以太网				
	InfiniBand及其他	Clos拓扑	具有增强遥测功能的Clos架构	单机箱	DDC（非DriveNets）	DriveNets Network Cloud-AI
架构灵活性	**低** 前端和后端采用不同技术	**高**	**高**	**高**	**高**	**高** • 无缝互联网连接——后端和前端采用同一技术 • 知名协议——以太网端口年出货量6亿 • 支持多个应用 • 支持扩展 • 不受限于ASIC和ODM
大规模高性能	**高**	**低** Clos性能不佳	**低-中** Clos性能中等 机箱扩展性差	**低** 机箱扩展性差	**高**	**高** • 高达320,000x800Gbps • 基于单元的整体结构 • JCT性能提升10-30%：可能会带来100%的系统投资回报率，因为组网成本占系统成本的10%
值得信赖的生态系统	**中** 封闭式解决方案，ASIC及硬件供应商锁定	**中-高** 通常而言并非开放式解决方案（供应商锁定）	**中** 非开放式解决方案（供应商锁定）	**低** 供应商锁定	**低** 未经过现场验证非开放式	**高** • 基于认证的开放计算项目（OCP）概念 • 赋能全球最大DDC网络（负责传输AT&T核心网络52%以上的流量） • 性能得到中美两国超大规模数据中心验证



不难看出，DDC解决方案最适合人工智能组网需求。虽然有多家供应商声称拥有基于DDC的解决方案，但DriveNets Network Cloud-AI是唯一上市且经过现场验证的解决方案 。



最高性能的以太网人工智能整体结构

随着人工智能工作负载和基础设施建设的快速增长，人工智能集群整体结构中使用的网络解决方案需要不断发展，从而最大限度利用昂贵的人工智能资源（人工智能加速器、GPU等），提供标准连接以支持供应商互操作性。人工智能训练集群需要能够提供无损、可预测连接的组网整体结构。

DriveNets Network Cloud-AI基于全球最大规模的Distributed Disaggregated Chassis (DDC) 架构，能够最大限度提高了人工智能基础设施的利用率并大幅降低其成本。根据超大规模数据中心的新近试验，这款解决方案不仅实现了无损连接，还将大规模、高性能人工智能工作负载的JCT性能提高了10%-30%。这意味着DriveNets Network Cloud-AI是人工智能基础设施最具成本效益的以太网解决方案。它在不放弃供应商互操作性的前提下提供基于标准的实施，最大限度地利用人工智能资源，从而有效地“收回成本”。

DriveNets Network Cloud-AI将为高性能人工智能工作负载设立组网标准，为人工智能组网提供高性能的以太网解决方案。

DRIVENETS

DriveNets是云原生网络软件和网络解耦解决方案领域的领导者。DriveNets成立于2015年，总部位于以色列，为服务提供商和云提供商提供全新的网络构建方式，能够通过改变技术和经济模式来大幅提高盈利能力。DriveNets推出解决方案Network Cloud（网络云），能够将云的架构模型提升为电信级网络。网络云是一款云原生软件，可在标准白盒的共享物理基础设施上运行，从根本上简化网络运营，以更低的成本实现电信规模的性能和灵活性。

欲了解更多信息，请访问www.drivenets.com