DRIVENETS

大型人工智能计算集群高效互连解决方案——Distributed Disaggregated Chassis (DDC)

白皮书

目录

引言

近年来，随着人工智能（AI）应用覆盖范围的拓宽，人工智能不断发展，用例和应用范围越来越大。虽然用户体验得到简化，更为直观易用，但生成这种逻辑所需的计算模型却变得更为庞大而复杂。微软、谷歌、Meta和字节跳动等提供大型云基础设施的公司，正在开发新的人工智能模型，将人工智能功能引入各自的基础设施产品，从而为用户提供更为广泛的服务。这些新部署服务的盈利能力尚不清楚，但更不清楚的是，未来会有哪些应用加入这一阵容，以及需要什么样的基础设施来实现这些应用。

处理这些新开发模型以及即将推出的模型中的计算需求所需的基础设施，正在推动市场开发强大的计算能力。市场规模的急剧增长引发了一场与主要"武器"供应商英伟达的"军备竞赛"，而其他供应商也逐渐进入这一领域。

本白皮书将介绍计算端的市场格局，重点关注大型服务器阵列互联的组网解决方案。在探讨这些解决方案的属性、优势和缺点限制的同时，本文还将深入探讨Distributed Disaggregated Chassis (DDC) 架构以及它如何适应人工智能组网。

人工智能计算服务器架构

自世纪之交图形处理单元（GPU）问世以来，计算领域的主角逐渐从以x86架构为主的中央处理单元（CPU）变为以GPU为主的系统。在此类系统中，CPU的作用是管理GPU资源，而不再限于用作计算资源。

GPU的主导地位

高性能计算（HPC）行业始终关注着世界500强超级计算机。2018年，高性能计算领域诞生了世界上最快的计算机Summit，它基于英伟达的GPU和IBM的CPU，采用Power架构（后已弃用）。2014年3月，美国能源部宣布拨款4.25亿美元用于研发构建超级计算机，标志着CPU在高性能计算领域主导地位的终结。Summit的原定算力达到150petaflops，即每秒150千万亿次浮点运算，几乎是中国天河二号超级计算机的三倍。

当今的顶级计算服务器由多个GPU组成，这些GPU与其他此类服务器之间以网格形式互联。英伟达最常见的架构是DGX/HGX，可支持8个GPU以及它们之间的专有整体结构（称为NVLink）。其他架构也采用AMD和英特尔的GPU，同时还有开发出来的其他技术，具有针对特定计算任务的能力，但不一定具有"通用"能力。部署最为广泛的是谷歌的TPU产品组合，目前已是第五"代"——每一代都专注于不同的工作负载类型。

虽然谷歌的TPU是最知名（或传播最广泛）的加速硬件，但其他超大规模公司也在开发自建加速器，其中一些已经投入部署。

GPU服务器架构挑战

在GPU主导高性能计算领域的同时，人工智能领域也在不断发展，并在容量和计算能力方面超越了高性能计算领域。这引发了另一场初创公司之间的竞赛，意图在人工智能计算领域占据可观市场份额。Cerebras、Groq、SambaNova、Graphcore等公司都在致力于构建更大型的计算基础设施，预计很快就会实现。

功耗

GPU服务器阵列通常内置于标准19英寸机架中。随着每一代高容量GPU功耗的提高，单个机架中托管的服务器数量取决于该机架的供电。

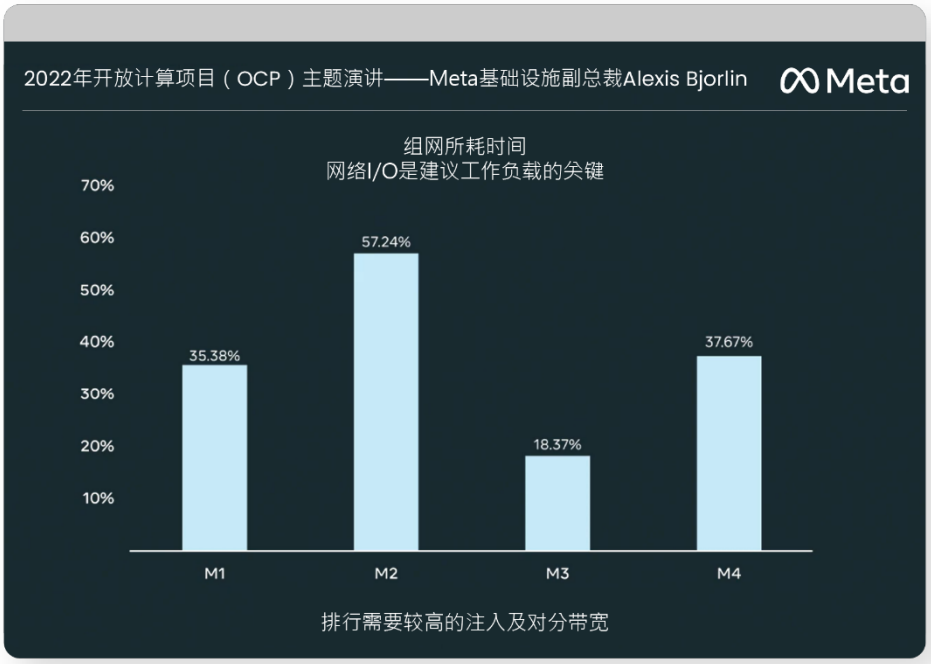
架顶式（ToR）拓扑

标准的数据中心设有架顶式（ToR）交换机/路由器，直接连接该机架中托管的所有服务器，并通过多层网络将它们连接到其他ToR设备。

随着各种不同的服务器架构与众多计算引擎的出现，标准ToR拓扑使得灵活采用统一架构连接不同类型设备成为可能。

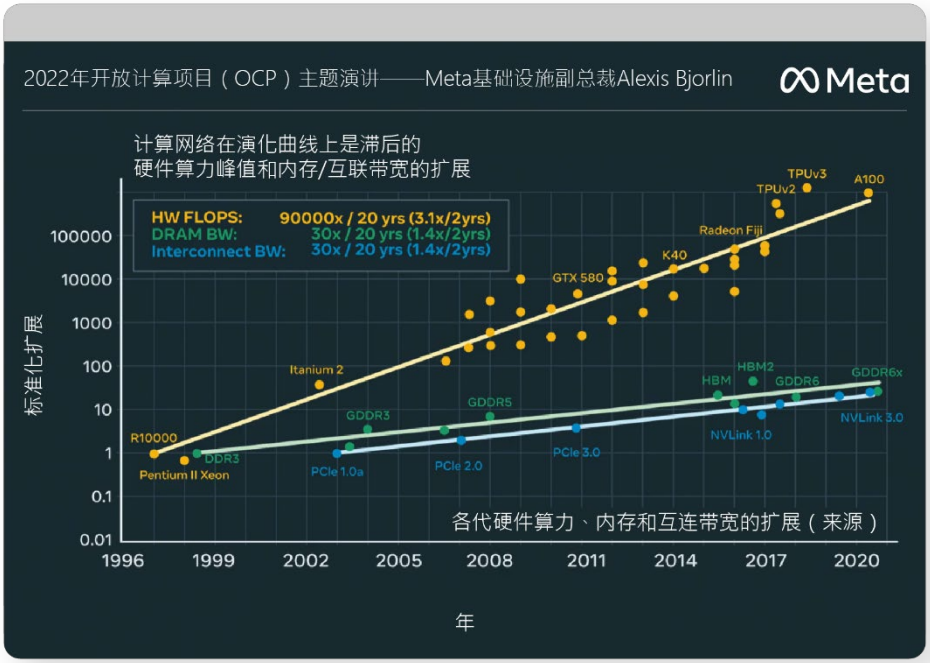
用于处理任务的GPU分配

不同类型计算引擎面临的另一个明显挑战是分配GPU从而处理特定任务。对于跨越数十甚至数百个处理器的并行计算任务，计算引擎的分配也是一项任务，如果处理得当，则能提高这些计算引擎的利用率。这是构建高效计算阵列的关键因素，因为构建此类基础设施需要大量费用，而这些阵列中最昂贵的元件是GPU服务器本身（占费用的80-90%）。



条形图显示等待网络提供所需负载（payload）时浪费的时间。浪费的时间即GPU空闲周期，也即对于已投入资金的浪费。来源：2022年"开放计算项目"（Open Compute Project）峰会Meta公司主题演讲

虽然合理分配计算资源可以提高GPU利用率，但前提是网络直接连接阵列中的所有GPU（可扩展到数千个），且保证无损高带宽和无拥塞网络连接。在为各个任务合理分配GPU的情形下，这种假设对于作业调度程序来说是相当合理的，但要建立如此高性能的网络却也是巨大的挑战。



图示单个GPU的计算增长和网络容量之间产生的差距。这更加凸显了网络端在合理扩展GPU阵列方面的作用。来源：2022年"开放计算项目"（Open Compute Project）峰会Meta公司主题演讲

人工智能计算服务器阵列中的组网挑战

组网领域存在各种独特群体。电信、企业、园区、数据中心，各不相同——在此仅举例几个不同的网络类型。在这些类别中还存在子类别，而在数据中心组网中通常有两种类型。

数据中心组网的两种类型

一种网络类型是本地以太网（Naive Ethernet）。称其"朴素"，是因为这种网络在扩展彼此互联且连接外部域（例如互联网）的数以万计的服务器时，它是能够实现有损技术的"最大努力"。这项技术存在很多供应商，以及健康的、庞大的竞争性市场。

数据中心内的另一种网络类型通常称为后端网络。在数据中心里，这种网络不可互操作，负责连接存储节点或计算节点，与数据中心外部断开连接。"后端"表示位置，但因其属性而区别于本地以太网。后端通常对其网络有更严格的要求：

- 无损行为
- 吞吐量和延迟可预测
- 高带宽
- 高密度部署环境

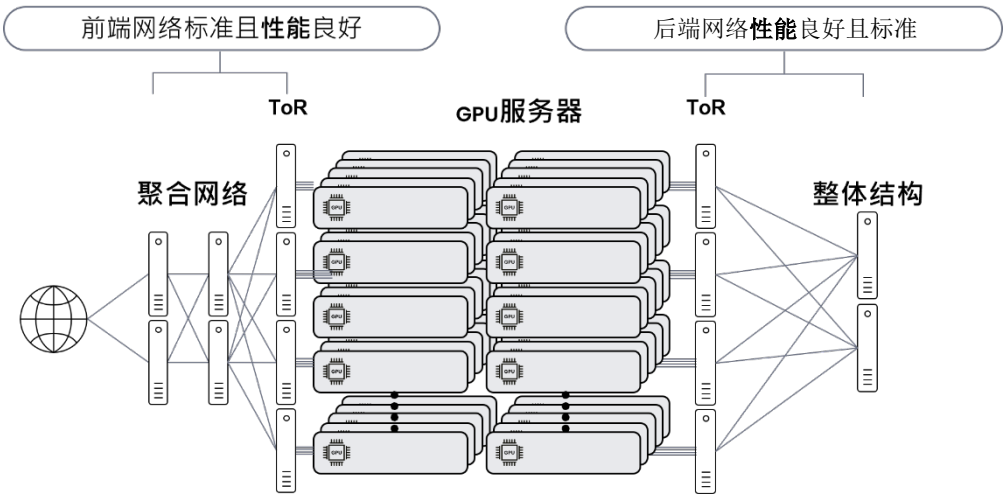
正是这些属性让本地以太网被弃用，从而让利基标准和专有技术有机会应用于数据中心。高性能计算领域就是后端网络技术（例如Slingshot、InfiniBand）蓬勃发展的用例之一。

无损网络的影响具有三重效果。为了提高GPU利用率，需要了解GPU周期被浪费在哪里。

- 作业GPU分配不当，意味着可用的GPU未分配并保持空闲状态，或者分配了错误GPU——这里说的错误GPU意思是这种分配会导致作业完成时间（JCT）的缩短；这是作业调度程序的任务，它"假设"GPU之间的流量始终到达其目的地的，但这不在再本文讨论的范围。
- 本文将详细探讨在网络本身内明显增加的交叉对分带宽（CBB）。
- 准备GPU要处理的数据集，可以通过GPU服务器本身更好的互连来改进。具有更好卸载（offload ）功能和远程直接内存访问（RDMA）的网络接口卡（NIC）能够将网络承载的有效负载准备成数据集，以便GPU更快地处理。这使GPU免于处理此类任务（浪费计算周期）和/或等待网络接口卡准备要处理的新数据集。从网络借口卡的角度来看，为了使RDMA发挥作用，网络必须是无损的。无论是通过InfiniBand（无限带宽）还是基于融合以太网的远程直接内存访问（RoCE）实施，远程直接内存访问都是通过绕过在服务器之间移动数据的软件程序来实现其效率，其中一些软件程序还包括检查流量是否已到达且完好无损。故障网络（有损或噪声网络）将会终结使用远程直接内存访问的优势。

第三种数据中心组网类型：人工智能基础设施

随着人工智能设施的兴起，尽管与高性能计算有很大不同，但计算阵列仍然需要来自网络的与后端相同的属性。然而，这种人工智能基础设施应该是云的一部分，支持与数据中心外部域的本地连接。这种组合创建了第三种类型的数据中心组网，应具备良好的性能，还应基于广泛使用的标准。



典型的数据中心阵列结构显示两个独立的网络，而人工智能的引入将把两个网络推向第三种类型，或者可能是两种类型的融合。来源：DriveNets

对于第三种类型的数据中心组网是否会消耗现有两个域中的一个或甚至将两个域纳入其自身，存在一些争议。本文不涉及这一争议，而是重点关注这一新形成的类型及其属性。

GPU分配挑战

GPU服务器正在快速发展，GPU容量可以达到400G的网络馈送，预计很快将增长到800G。当运行繁重的工作负载时，一组GPU都可以充分利用其网络链路的最大容量，在网络内创建最大吞吐量。在数以千计的GPU规模下，显然无法将所有GPU连接到单个设备，因此网络中存在第二层。网络第二层承载的带宽称为交叉对分带宽（CBB）。当多个GPU同时传输/接收大量工作负载时，交叉对分带宽会出现峰值。这是相当常见的场景，由并行计算算法中GPU分配方式导致。

工作负载行为挑战

网络面临的另一个挑战是消息在端点之间通过网络运行所需的时间——不仅是平坦延迟（flat latency），还包括消息之间的延迟变化。这是因为GPU一直处于挂起状态，直到完全确认计算周期已结束为止。长尾延迟将消除任意最小头部延迟值。

规模挑战

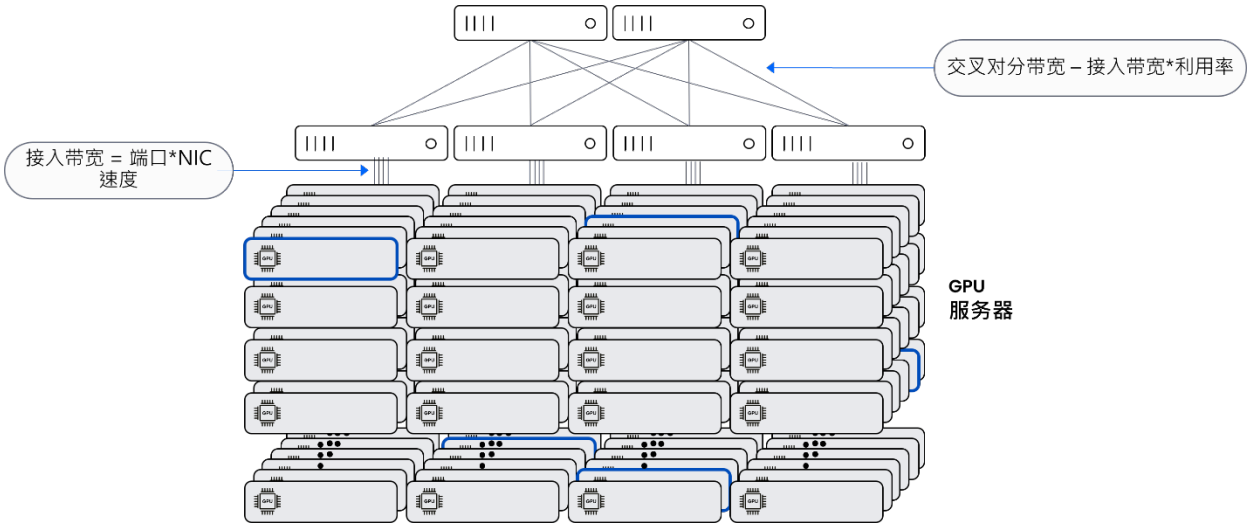
人工智能工作负载很大，且在不断增大。模型的准确性和超越竞争对手的竞赛需要更大规模的基础设施来运行。网络的挑战或属性在中等规模时都是明显的，当规模增加时都会成倍增加。

通用性挑战

网络面临的另一大挑战是处理不同的流量模式和工作负载特征。各种人工智能模型的运行方式不同，通常使用不同的通信算法，例如：

- 多对多
- 多对一
- 一对多
- 全局归约（All-reduce）

数据包大小和流量大小体现了网络需要传播的流量的一些不同特征。人工智能网络无法针对一种特定的流量模式进行微调，但必须满足所有这些模式。



有标记的GPU是指分配给同一作业的GPU，意味着这些端点之间的高流量。
网络需要无差别地满足任意流量模式，以提高其利用率。
来源：DriveNets

这些挑战为这两种类型数据中心组网的网络解决方案带来了障碍。然而，这些解决方案都有缺点，让它们无法最大限度地提高网络利用率。

人工智能集群组网解决方案

数据中心组网有多种解决方案，都可以被视为人工智能集群网络解决方案的候选方案。本文将研究部分主要候选方案的优缺点。

本地以太网

本地以太网是数据中心内部署最多的网络类型，其强势的领先地位得益于自身的几个属性。这一解决方案包括：

- 连接到服务器的ToR交换机/路由器
- 连接到ToR交换机的聚合层（通常为1或2层）交换机
- 在Clos-3或Clos-5拓扑中的聚合层路由上行链路它具备以下属性：
- GPU网络端口的带宽规模非常大且很高
- 标准行为以及与数据中心外部设备的通用连接

许多芯片供应商都实现了此类网络，主要如下：

- 博通（Tomahawk、Trident ASIC）
- 英伟达（Spectrum ASIC）
- Marvell（包括Xpliant和Innovium产品）
- 思科（多个自建ASIC）

- 英特尔（Barefoot、QLogic、以及自建）

系统和软件包括系统供应商的垂直集成解决方案以及开源解决方案，例如用于硬件的开放计算项目（OCP）和用于软件的SONiC。

虽然本地以太网解决方案在开放性和标准化方面得分很高，但在性能方面却不尽如意。

- 尽管GPU的带宽很高，但当网络的整体吞吐量预计上升时，交叉对分带宽会迅速下降。
- 虽然通过单个交换机的单个数据包的平坦延迟可能非常低（采用直通交换延迟），但数据包在端点之间采取的不同路径会导致较高的尾部延迟，并影响GPU性能。
- 由于流量是通过基于哈希的分配算法（如ECMP）从网络的第一层转移到更高层，那它理所当然地对流量特征敏感，因此不平衡。

这些缺点限制意味着，在大型环境中，由于网络的"朴素"行为，许多GPU将长时间处于空闲状态。

对本地以太网Clos进行改进的尝试

如前所述，完美网络的"朴素"（又称为"最大努力"）假设实际上意味着"最差努力"，鉴于是最差努力，那么显然可以改进。目前已经提出且实施有几种改进本地以太网Clos（脊叶Spine-Leaf）拓扑的尝试。

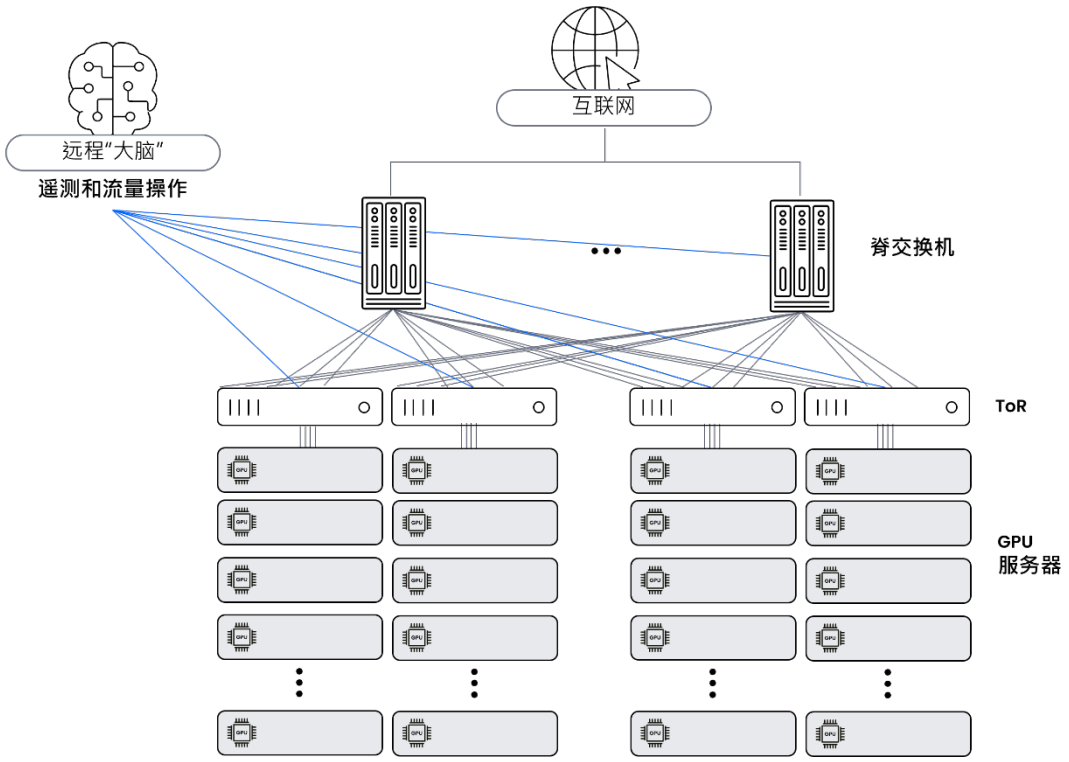
部分示例包括：

- 具有前向和后向通知的细粒度流量控制机制
- 噪声引起的哈希函数，可强制在多个第2层（脊交换机）设备上分配大流量
- 自适应路由，可为标准路由协议之外的部分流量强制实施替代路径
- 以及更多其他示例

这些附加功能（其中一些是专有的）使得朴素Clos解决方案不再那么标准，而它们带来的性能改进却微乎其微。

GPU作业分配

处理Clos拓扑的另一种方法是在Clos的同一分支内进行GPU作业分配。如此，GPU之间的流量不会绕过网络的更高层，并且根本不会需要任何交叉对分带宽。虽然没有真正解决而是避免了Clos的部分固有问题，但这样使得Clos能够应用在某些特定的人工智能用例。这种方法将这项工作的负担转移给了作业调度程序，需要考虑到GPU的物理位置。在某些情况下，这意味着即使集群有足够的计算能力来运行作业，作业也不会运行，因为这些是分配"错误"的GPU。



典型的带有"增强功能"的本地以太网Clos——一种在有损技术上强制执行无损行为的"最大努力"尝试。来源：DriveNets

InfiniBand

InfiniBand于1999年推出（同年英伟达推出首款GPU），作为在高容量性能网络中连接服务器的解决方案，它具备成功后端解决方案的所有特征。

作为一个新形成的标准，它包含了以太网原本不具备的功能，包括：

- 防损流量控制，可实现InfiniBand无损行为
- 自适应路由，可避免使用哈希算法
- 周密的流量处理，使网络能够应对独特的流量特征

高性能计算领域的主导标准

InfiniBand逐渐成为高性能计算领域的主导技术，其中计算集群很少（或者几乎不会）连接到计算集群外部的设备。虽然InfiniBand逐渐在高性能计算领域占据主导地位，但它未能对其他数据中心域进行任何有意义的渗透（以太网仍然是主要选择）。这导致除了一家之外的大多数InfiniBand供应商跳出了这一技术领域，放弃了各自的产品组合。开发InfiniBand解决方案的唯一供应商是Mellanox Technologies，于2020年被英伟达收购。

英伟达是唯一的InfiniBand供应商

虽然InfiniBand是正式标准，但其单一供应商模式在很大程度上否定了部署协议的想法。将InfiniBand网络连接到其计算网络之外的设备非常麻烦，而且会造成性能障碍。单一供应商技术是一种风险，但可以肯定的是，英伟达不会很快终止这项业务（或者瓦解，这将产生比网络影响更大的冲击）；尽管如此，从人工智能集群构建方的角度来看，从同一供应商采购太多组件仍然是一个风险。

有实力的供应商是好的，因为他们能让顾客安心。占据主导地位的供应商则不然，因为他们支配着客户的基础设施，让客户失去控制权。

人工智能集群可引入更多端点

与InfiniBand有关的部分其他缺点限制与高性能计算和人工智能计算集群之间的差异有关，具有相似之处，但并不完全相同。

人工智能集群的规模可以远大于高性能计算集群，并且还在不断成长之中。InfiniBand是一种软件定义网络（SDN）技术，专为数百到数千个端点而设计。人工智能集群已经可以针对单一集群的数万个端点。虽然理论上InfiniBand可以达到这些规模，但在实践中，当SDN尝试大幅扩展时，会反复出现失败的结果。

尚未准备好纵向扩展到足以支持更快的GPU服务器

另一个与扩展相关的问题是GPU服务器的纵向扩展因素。虽然目前还是400G（即InfiniBand中的NDR，最近刚刚发布），但800G即将到来，InfiniBand速度的变化周期通常为4-5年。英伟达可能会延迟推出支持800G的GPU，并冒着失去其作为GPU供应商优势的风险，从而为NDR InfiniBand提供更多的机会。

缺乏调整

InfiniBand的最后一个缺点限制源自它对特定应用程序和工作流程属性提供的调校。在高性能计算领域，这是一个优势，可以提高性能，使InfiniBand优于其他方案。然而，在人工智能领域，由于工作流程和流量模式不断变化，这种微调的能力并不实用，而且缺乏调整会导致性能不佳。

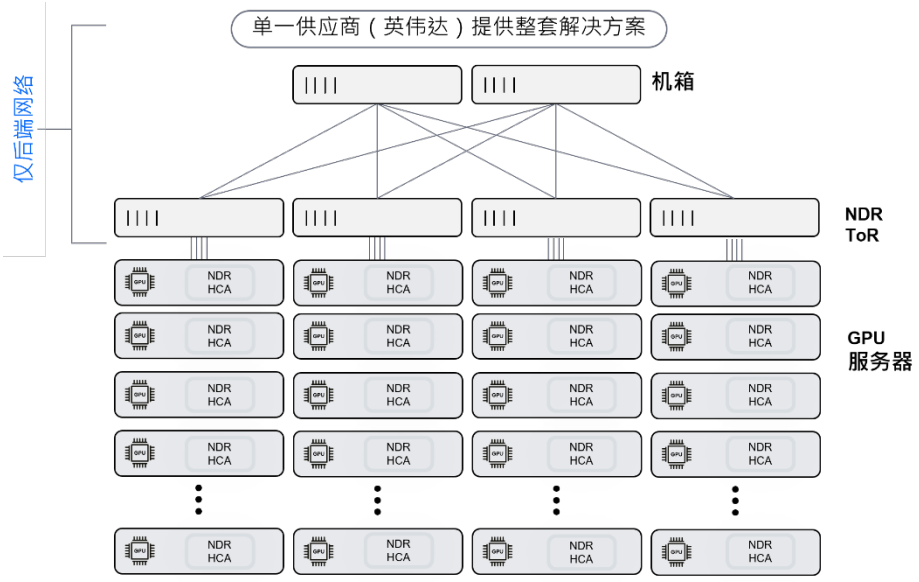
针对大型人工智能集群的实施尚无测试结果

考虑到针对大型人工智能集群中InfiniBand的测试结果非常有限，甚至没有已知的测试结果，可以肯定得是没有理由隐瞒此类结果。然而，众所周知，ChatGPT等大型语言模型（LLM）是通过InfiniBand网络进行训练的。很明显，这一解决方案能够达到一定的性能水平，尽管明显低于英伟达期待的市场反应。

英伟达供应商锁定导致性能风险

英伟达提出的另一种说法也值得关注：从网络到GPU的速度与SerDes完全一致，从而可以实现性能提升。那么是否可以假设更换互连会降低GPU本身的性能？是否还可以假设，当GPU不是来自英伟达时，InfiniBand（或任何其他英伟达网络）的性能就会不佳？

因此，InfiniBand在规模方面和工作负载变化方面的表现并不突出，并且还导致GPU提供商锁定，这一点很麻烦。



典型HPC Clos拓扑InfiniBand（唯一）后端网络，由单一供应商提供，最大速度为NDR（400G）。来源：DriveNets

定制的以太网

就标准而言，以太网有多个级别。虽然这类实施可以与以太网帧格式保持一致，但它们也可能引入非标准流量处理行为。我们将这些选项称为"扭曲的以太网"。

这类解决方案：

- 在高性能计算领域表现良好（如HPE的Slingshot），其人工智能行为有待观察
- 并入更大的公司（如Qlogic和Barefoot并入英特尔），并且可能会改名换姓
- 是颇具营销声势却有待观察的技术（例如思科的Silicon One集群和英伟达的Spectrum-X解决方案）

不同的流量

这些解决方案都尚未发布合理的描述说明，也没有在该领域内传达部署，因而本文最多只能进行部分分析。一些解决方案声称具有最佳规模（Spectrum-X和Silicon One），一些声称具有出色的SDN控制和数据包流操作（Barefoot），还有一些声称具有出色的高性能计算性能（Slingshot）。但值得注意的是，所有这些解决方案都清楚其承载的流量，由于流量模式不断变化，并且流量之间可能存在重复特征，可能会导致交叉对分带宽不佳和性能严重下降。

人工智能工作负载尚未经过测试

另外值得注意的是，新发布的解决方案（Spectrum-X和Silicon One）声称是为人工智能工作负载而设计的。因此，虽然没有关于它们如何处理流量的详细信息，但可能存在一些隐藏的机制。思科声称数据包被"喷射"在脊交换机层，从而产生待定的性能。英伟达声称，其解决方案决定了来自端点的流量路径，并应用自适应路由（很像InfiniBand），使得解决方案具有专有性，并要求采用完整的端到端英伟达解决方案（同样，很像InfiniBand）。

单机箱

具有一定效果的潜在解决方案应该能够避免网络的所有潜在事故，并将所有GPU连接到单个网络元素。如果所有GPU都连接到同一个基于机箱的设备，那么与干扰流、流量突发、可预测性等相关的所有问题都将被消除。

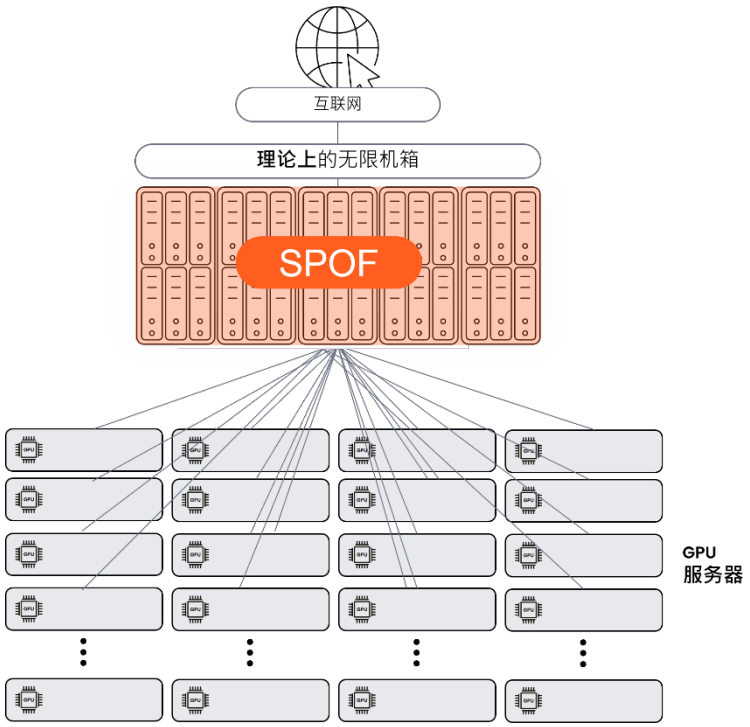
不切实际的人工智能集群解决方案

虽然这种拓扑的解决方案确实存在，但主要是在网络技术匮乏和资金充裕的领域，但不是出于性能的原因，而且肯定不是大规模的。以具有2000个端口的人工智能集群（按如今标准相当于小型计算阵列）为例，几乎不可能构建相同尺寸的机箱。有人声称有这样的InfiniBand机箱，但也仅存在于理论层面。使用这样的机器，需要供电、冷却以及托管所用的标准机架，都会产生物流和运营开销。

运营噩梦

一旦就位，要从大约100个相邻机架拉入2000根光缆连接到机箱，就安装一项而言就是一场运营噩梦，对这些光纤链路的维护更是如此。使用铜缆（无源或有源）更加可靠且不易出现故障，但无法将如此众多的不同机架连接在一起。

虽然单个机箱在理论上可能是一个很好的解决方案，但实际上它的操作难题无法克服，并且会构成单点故障（SPOF）。



潜在的单一"无限规模"机箱结构会造成布线混乱，给整个计算阵列造成单点故障。来源：DriveNets

Distributed Disaggregated Chassis (DDC)

Distributed Disaggregated Chassis (DDC) 是开放计算项目（OCP）电信项目下定义的一项新兴技术。根据AT&T定义使用博通的转发ASIC，UfiSpace、Delta和Edgecore已实现其白盒硬件的安装启用，并应用DriveNets和思科的软件。AT&T也是第一家在其网络中实施DDC模型以运行核心MPLS和对等路由器的公司。

与Clos拓扑相同的优点

DDC 和机箱结构一样采用Clos拓扑，既具有机箱的可靠性和性能等优势，也具有使用小白盒的Clos拓扑的灵活性。机箱中的大多数组件均已作为独立设备安装启用，可以通过软件"重构"为统一的网络元件。

基于软件

机箱中没有专用设备在DDC中予以体现的一个组件是机箱设备本身。将所有部件组合在一起并确保机箱的诸多控制、管理和数据平面等属性的金属机箱已被软件取代。这类软件负责为分布式机箱协调所有各种组件。

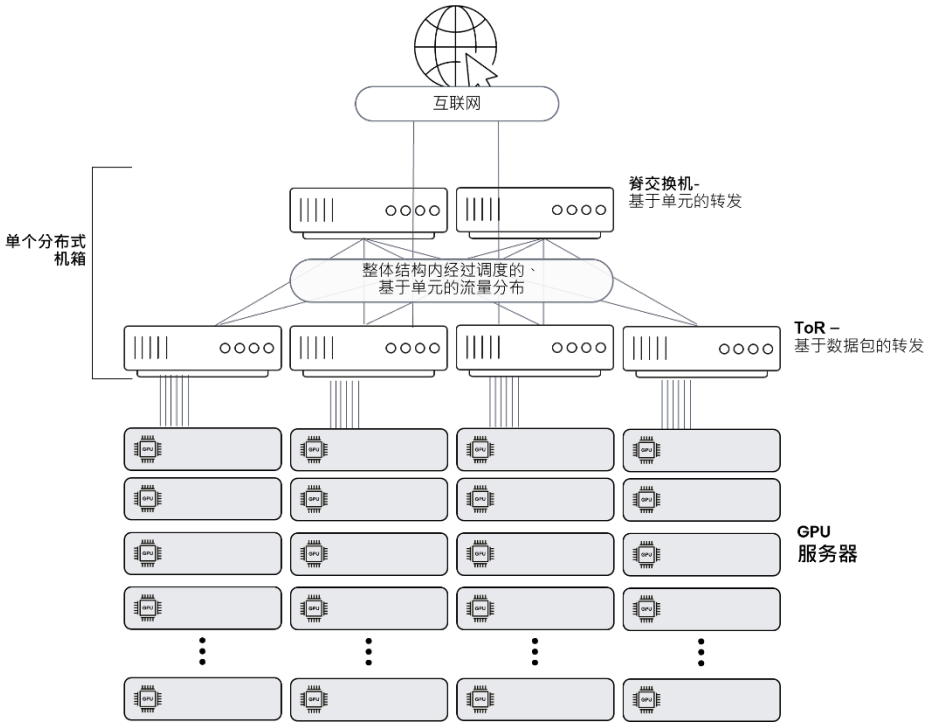
可扩展

除了优点，传统的金属机箱还是一种限制。在部署的规模、托管端口的数量、消耗的功率以及散发的热量方面存在物理限制。DDC的实施意味着能够摆脱这些限制，增长到（甚至超出）ASIC的最大基数。电力和热量分布在各种独立设备之间，每个独立设备都可以放置在距离直连端点更近的不同机架中。

DDC是数据中心互连（DCI）的解决方案

虽然最初的构建意图是用作大型多功能多服务路由器，但DDC模型可通过DriveNets Network Cloud Packet Forwarder（网络云数据包转发器，NCP）用作数据中心互连（DCI）解决方案，其中NCP充当ToR，DriveNets Network Cloud Fabric（网络云整体结构，NCF）充当脊交换机。这种解决方案具有运营商级路由器的所有属性以及几乎无限的ASIC密度规模。

DDC满足人工智能集群需求的方式以及DDC解决方案的差异化将在下一章讨论。



DDC拓扑的运行方式和物理外观类似于Clos，并通过标准以太网连接到外部设备，而其内部运行方式类似于调度整体结构。来源：DriveNets

内部整体结构

内部整体结构是一种非常常见的GPU连接方法。由于英伟达的GPU在人工智能市场中占据主导地位，并且该公司通过其DGX/HGX服务器设计来推广NVLink的实施，因此这种方法非常普遍。AMD的Infinity Fabric以及Cerebras的固有数据平面整体结构机制（可互连其晶圆解决方案内的区域）均提供了等同等解决方案。

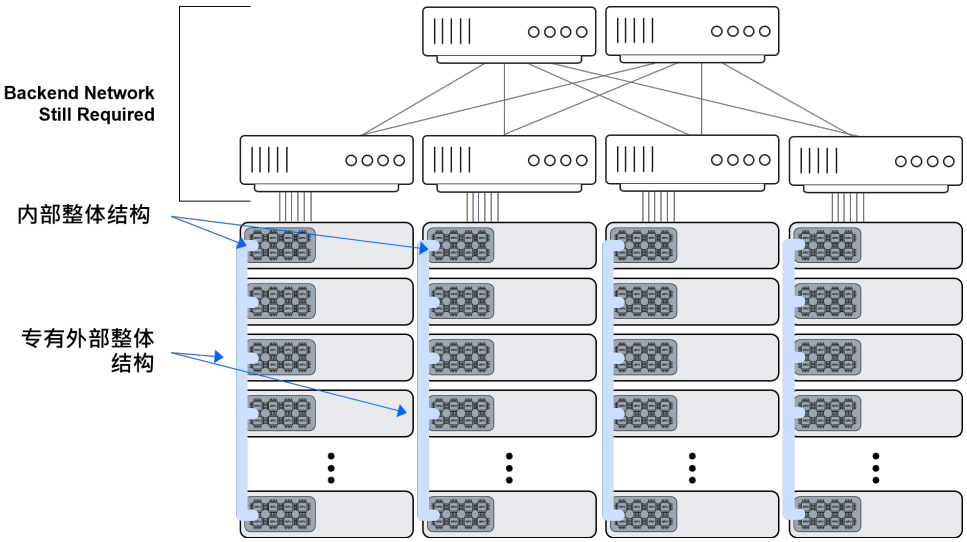
黑盒功能

内部整体结构解决方案都有点神秘性。虽然确实有用，但是它们的工作原理目前还不清楚，因为都是各大供应商专有且受保护的知识产权。它们确实能够实现计算引擎之间的无缝连接以及整体结构连接的计算引擎之间的高效吞吐量。在某种程度上，内部整体结构还提供了服务器内计算引擎之间的某种平衡，从而使调度程序的任务变得更容易，并使编写应用程序代码更容易实现。

供应商锁定问题

只要解决方案的内部整体结构类型保持其内部性，从用户的角度来看，这是可以容忍的。无论是单个计算引擎，还是一对、一个八位组还是整个晶圆，都仍然被视为"即买即用的现成"单个构建块，其专有性质也被认为是可以接受的。将这种"未知"整体结构扩展到大规模服务器网络上更像是一种供应商锁定劣势（除了它安置工作更容易的优点之外）。

它给调度程序行为和应用程序代码架构带来影响，还妨碍采购团队混合匹配各种计算解决方案。



服务器中包含8个GPU的Pod，通过专有内部整体结构连接。外部后端网络不会因内部整体结构的存在而失效。将内部整体结构扩展到跨服务器专有结构要求所有GPU来自同一供应商，同时要求相应地编写软件。来源：DriveNets

Distributed Disaggregated Chassis (DDC) 的优势

适合人工智能集群的网络属性包括：

- 高带宽
- 高交叉对分带宽
- 在具有高度差异性的多个工作流中实现延迟的可预测性和一致性
- 无损行为（故障中断会导致非常高的代价）
- 采用通用标准接口的互联网连接

DDC 最初是作为标准电信级路由器构建的，其硬件和软件旨在维持99.999%的可用性标准（通常称为“运营商级”）。在将其全国核心网络委托给这项新技术之前，AT&T执行了完整的测试周期，确认DDC的所有设施都是标准的，因为它的实施环境是实时网络，连接的是该网络中的任意现有设备。

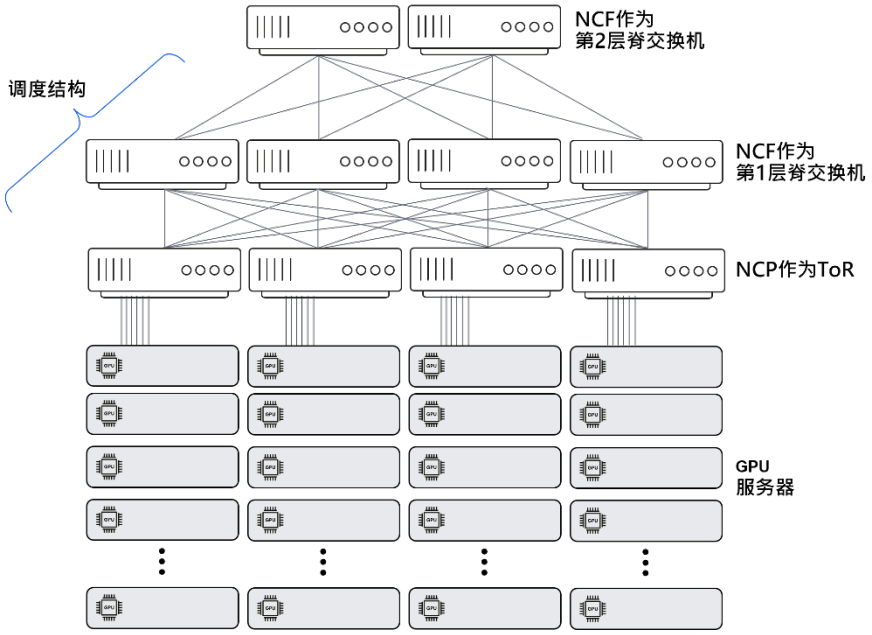
虽然人工智能集群会因网络故障而遭受经济损失（代价昂贵的GPU闲置），但电信领域的故障中断带来的损失也同样严重。如果中断服务或者中断政府规定的紧急服务（如911救助服务电话），运营商会被罚款，故障会让电信公司付出高昂代价。这就是推动DDC架构发展的动力。

在DDC诞生的背景下，DDC解决方案具备这些人工智能网络的属性就不足为奇了。让我们看看DDC架构如何应对关键挑战。

最大规模

DDC通过DriveNets的网络云数据包转发器（NCP）和网络云整体结构（NCF）实施，两者均作为独立元件构建。NCP以Clos拓扑连接NCF，从而实现ASIC基数和规模的最大展开（fanout），而这在金属外壳机箱中是不可能的。

在Clos-5拓扑中进一步连接NCF作为NCP和第二层NCF之间的中间层，正是这种能力实现了第二级规模。这样能够将DDC的展开（fanout）从单个ASIC (N)的基数增加到该基数的幂[N^2/2]。Clos拓扑能够确保每个网络流的平坦延迟。



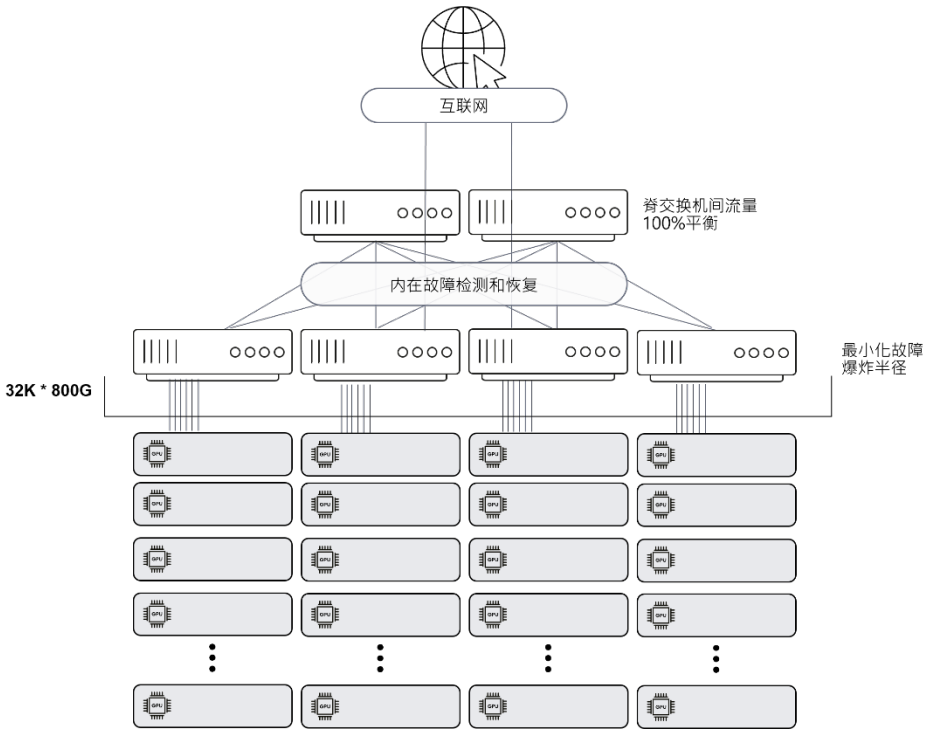
位于Clos拓扑中的NCP和NCF构建两层整体结构的DDC基础设施

高带宽和降低的交叉对分带宽

DDC的最初设计采用了博通 Jericho2 ASIC，设有100G和400G接口，通过Ramon ASIC运行的400G整体结构进行连接。下一代Jericho引入了100G SerDes和800G接口，并具有向后兼容性，因此可以将其添加到现有DDC集群或用于构建新集群。这能够为当今最新GPU服务器实现连接，并为预计在2024年推出的下一代服务器做好准备。由于网络接口卡和交换机速度的变化周期一致，我们可以假设下一代1.6T速度也将从时间线和互操作性的角度保持一致。

查看DDC上接收到的数据包，从而分析该数据包应该从DDC内流出的目标端口。然后，数据包被分成均匀的单元，将这些单元喷射到所有结构设备，在这些整体结构设备中被转发到出口端口（egress port）进行重组。每个数据包的这种单元分布能够确保所有整体结构元件的均衡利用。这意味着，平均而言，当整体结构利用率为90%时，每个结构的实际利用率也为90%。

任何其他在不同脊交换机设备之间分配流量的方法都会受到脊交换机设备之间利用率差异的影响，这样会有效降低整个网络的交叉对分带宽。



DDC拓扑显示连接到某一整体结构的高达32,000个GPU的规模，具有99.999%的可靠性和接近100%的网络利用率。来源：DriveNets

快速故障恢复

任何具有连接设备的网络都能够检测物理故障并触发结果。DDC在这方面没有什么不同，除了一个细微的区别，即DDC内的*任何*故障都被视为内部故障。DDC是一种众所周知的拓扑结构，其运行方式也是如此。任何故障都会由硬件指示器捕获，并根据每个元件在本地预先计算此类故障的含义。对数据平面的影响和所需的反应是在纳秒级恢复中实现的。此外，故障指示的传播是主动进行的，因此整个DDC系统都会意识到故障以及如何最小化由此导致的数据包丢失。

这与传统网络的采样性质相比是有利的，在传统网络中，故障恢复缓慢对GPU的测量利用率及有效利用率有很大影响。由于人工智能作业在多个GPU上运行，当发生网络故障时，作业无法运行完成。结果会导致作业需要全部重新运行或从它到达的最近检查点重新运行。这意味着即使运行作业的GPU处于活动状态，它们运行的工作也需要重新运行。

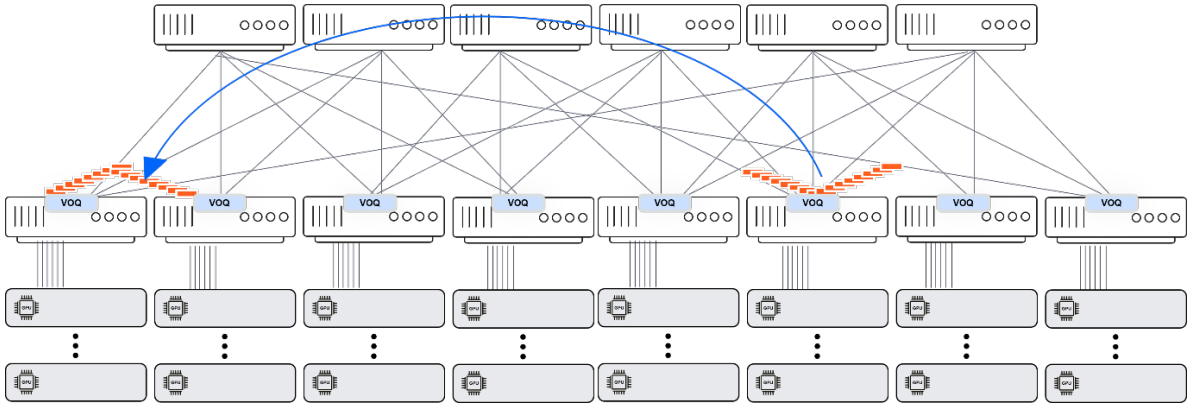
也就是说，尽管有效利用率以实际利用率衡量，但还是受到了影响。鉴于现代人工智能集群能够到达数以万计GPU的规模，用于互连所有这些GPU的元件数量（>1000）也就意味着故障场景会很常见。

与毫秒级相比，更快的纳秒级故障恢复使实际和有效的GPU利用率保持一致，从而缩短作业完成时间（JCT）。

延迟保持一致

DDC的"分段和重组"（SAR）层能够确保所有数据包从源到目的地都经过相同的路径，从而防止数据包延迟发生变化。然而，由于系统具有多个输入和多个输出，且采用任意到任意的流量模式，因此队头阻塞（HOL）场景仍然可能对吞吐量和延迟造成严重损害。出现队头阻塞场景意味着存在无法转发的流（因出口端口拥塞或Incast现象等实际网络原因而导致）。这条位于转发线路头部的流在等待"它的"出口端口可用性的时候，会阻塞其后的其他流，不然的话，这些流是可以转发的，因为"它们的"网络部分是空闲的。实际网络会因队头阻塞场景而发生急剧恶化，随着网络利用率的增长而变得非常普遍，还会影响数据包延迟。

DDC实施与出口端口关联的虚拟输出队列（VOQ）。队头数据包在等待出口端口的可用信用，然后会被发送到分段和重组层，而后进入整体结构中。此时，流向其他目标出口端口的流不会在同一队列中等待，因此不会被队头数据包阻挡。数据包在DDC中延迟的唯一情形是当某个直接连接GPU的出口端口的利用率高于其所连接的GPU的容量时。这种情形并不正常，实际上表明人工智能工作负载调度程序存在缺陷行为。



DDC拓扑的内部流量依赖于基于信用的转发机制，这种转发会触发数据包从VOQ传播到单元中，然后在接收端进行重组。任何流量都将以类似方式处理。来源：DriveNets

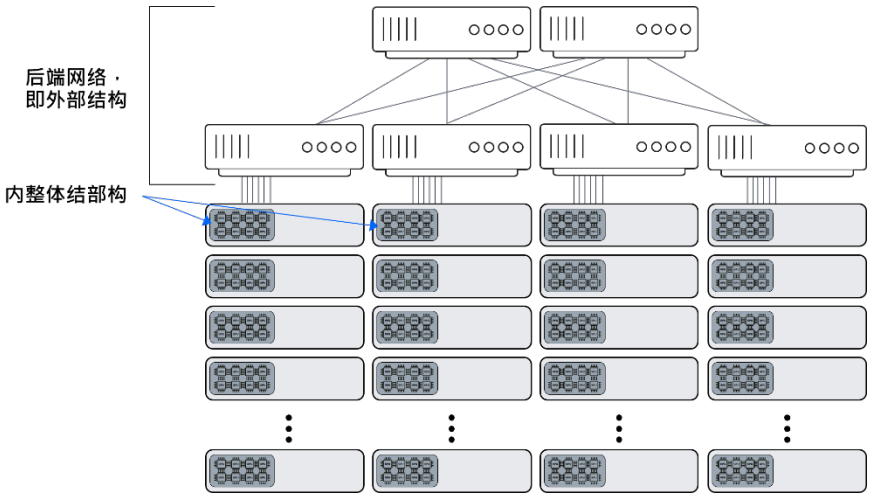
与流无关（Flow Agnostic）

通过查看每个特定的工作负载，可以对网络进行训练，从而确定后续工作流的需求。但是进行这样的调整非常困难，需要特定的专业知识。确实，调整后的网络需要能够对流量进行此类外部调整，SDN就是这种网络的一个常见例子。非恒定、不断变化的GPU阵列会导致流量模式和工作负载特征发生变化，当工作流在这种阵列上运行时就会出现问题。

分配一组分散的GPU来运行特定作业是由作业调度程序处理的任务，作业调度程序会考虑每个GPU的可用性和容量。GPU的位置以及连接这些GPU的网络的现有工作负载对于GPU和调度程序来说都是未知的。如此便会导致这样一种情景：针对特定流进行网络调整没有用处，而针对每个流行调整则无法实现。

DDC架构避免了针对特定流选择最优路径。由于DDC的单元分布，每个流以均衡的方式假定入口点和出口点之间的所有可能路径。这也解决了每个作业分配网络感知GPU的巨大挑战。

GPU作业分配的另一个角度涉及前面提到的内部整体结构以及将该结构扩展到服务器外部。整体结构的优点对于业界来说显而易见。能够对无缝GPU分配的整体结构优势进行拓展，同时也能忽略整个集群上单个服务器之外的集群中GPU的物理位置，这正是DDC结构的优势。在这种情况下，扩展整体结构和后端网络正是提供外部高带宽以及与集群甚至外部世界的可靠连接的元件。这种元件还能够提供内部任意对任意GPU连接，无需考虑GPU的物理位置。



人工智能集群，显示服务器内内部整体结构（NVLink/Infinity/其他）的使用，而服务器间连接由标准DDC结构实现。来源：DriveNets

无损行为

如上所述，DDC架构处理数据包的排队和分段，最大限度提高网络利用率和提高交叉对分带宽（CBB）。

根据定义，以太网具有有损网络行为，将数据包转发到目的地，同时假定数据包会经过每一个网络结点和最终目的地的出口。如果数据包遇到任何网络问题（意味着发生了问题），则会发出警报以避免数据包丢失。如果仍然出现数据包丢失，以太网会假定更高层会负责重新传输丢失的数据包。这种行为被称为“最大努力”。

不过，DDC实施三种机制来避免数据包丢失，以此确保无损行为。

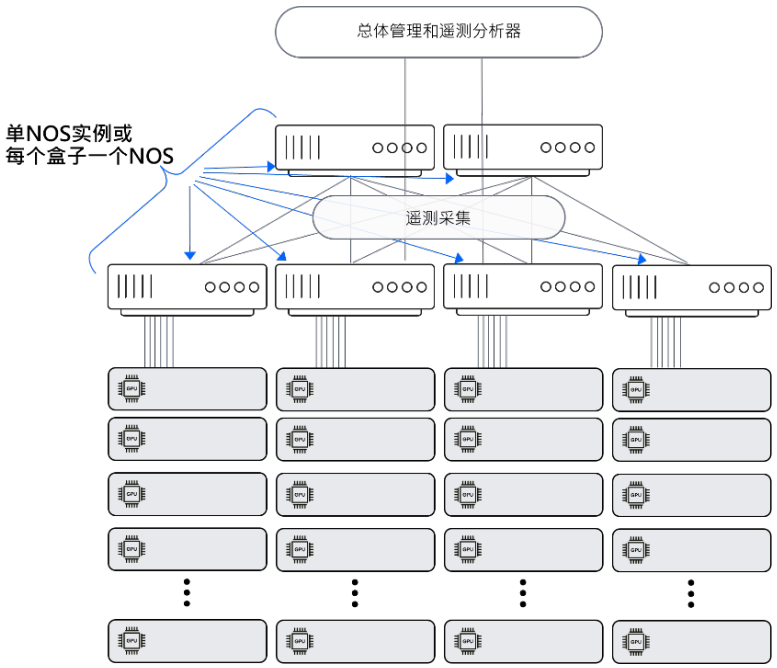
1. 入口端口（ingress port）中的虚拟输出队列（VOQ）能够防止如上所述的队头阻塞
2. 分段和重组层能够最大限度提高网络的交叉对分带宽，实际上可以防止这些所谓的“网络问题”
3. 基于信用的转发机制。出口端口向入口端口发出可以转发数据包的信用信号。也就是说，通过整体结构发送的数据包不会出现容量不足的情况。这意味着DDC架构具有一大优势，能够在新作业启动时或在作业之间更好地应对小规模突发流量，从而避免出现流量消耗整体结构资源和数据包最终被丢弃的情况。

遥测采集

在构建用于承载数据中心流量的大型网络时，网络内有许多称为遥测的指示点。这些指示被采集并用于触发有关网络行为的结论和行动。最为人所熟知的例子是延迟的测量和/或丢包可能性的计算。虽然这两个属性在DDC中得到了根本解决，但遥测采集的能力仍然存在，并且可用于对网络运行状况进行更深入的分析和故障预防。

管理不同的网络操作系统

大型人工智能阵列通过大型网络连接。这是显而易见的。没有使用DDC的情况下，大型路由网络会带来令人头痛的管理难题。从管理协议和路由表，到管理不同的网络操作系统（NOS）类型和版本以及它们之间支持或不支持的互操作性，这些都是使用DDC即可轻松避免的任务。DDC可以将单个集中式NOS实例作为分布式功能运行在所有DDC设备上，也可以运行在独立设备上运行的多个NOS实例。虽然推荐的最佳实践是用于集中式NOS，但用户也仍然可以选择用于多个NOS的实践。



DDC将整个集群的遥测数据传输到分析仪中。容器化NOS功能或每个盒子的完整NOS能够提供操作灵活性。来源：DriveNets

遥测和集中管理

DDC的另外两个相对不为人知的功能是遥测和集中管理。

实例：DDC的应用

接下来一章将探讨实际测试场景得出的一些结果。

当单个流量流的故障损伤或其他突出行为影响同一网络内的其他流量流（即使流量之间没有关系），就会出现明显的相邻干扰（noisy neighbor）场景。

多个作业的平均JCT时间

下列表格显示了2,000个GPU规模的集群中随机安排的多个作业的平均JCT时间。

在第一个表格中，导入任意故障损伤后的性能与建模的理论性完美网络仅存在4%的偏差，这一结果极为出色。

第二个表格列出的是将链路容量降低至70%的某种故障损伤，作业#1的性能下降了24%，而所有其他作业根本不受影响。

最后一个表格列出的是将容量降低至50%的更严重的故障损伤，结果是直接受影响的作业出现了73%的性能下降，而在这种情况下，仍然没有影响同一网络中运行的其他作业。对Clos拓扑进行的其他测试表明，相比于4%的偏差值，与最佳值的偏差在70%场景中已提高至5.5%，在50%场景中则提高至6.5%。

2048个节点的DNX架构，随机安排，无故障损伤

配置	作业	作业	第100个
2,000个节点的DNX架构，多个作业（8x256），随机安排，无故障损伤	6	1.040	1.041
	2	1.040	1.041
	8	1.040	1.042
	4	1.040	1.041
	7	1.040	1.042
	5	1.040	1.041
	3	1.040	1.042
	1	1.040	1.042

2048个节点的DNX架构，随机安排，一个NIC损伤70%

配置	作业	作业	第100个
2,000个节点的DNX架构，256Nx8，随机安排，导致PCIe带宽降至峰值70%的RX损伤，单一作业损伤	1	1.238	1.240
	7	1.040	1.041
	3	1.040	1.042
	5	1.040	1.041
	8	1.040	1.042
	6	1.040	1.041
	2	1.040	1.042
	4	1.040	1.042

2048个节点的DNX架构，随机安排，一个NIC损伤50%

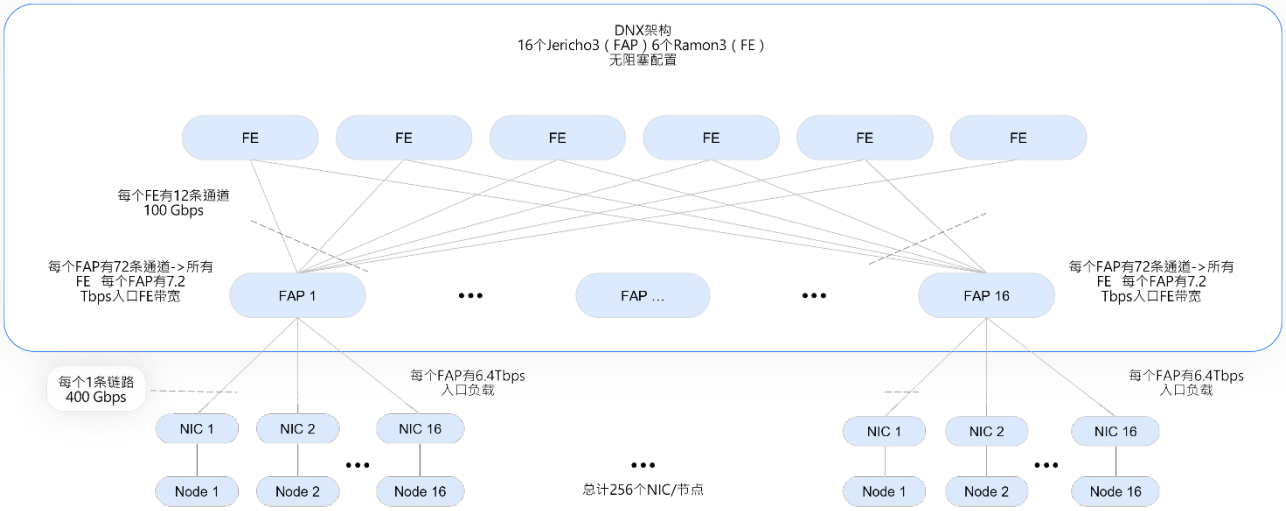
配置	作业	作业	第100个
2,000个节点的DNX架构，256Nx8，随机安排，导致PCIe带宽降至峰值50%的RX损伤，单一作业损伤	1	1.730	1.733
	7	1.040	1.041
	3	1.040	1.042
	5	1.040	1.042
	8	1.040	1.042
	6	1.040	1.042
	2	1.040	1.042
	4	1.040	1.042

这些结果突出强调了需要监控稳态下的GPU分配决策以及故障场景下的附带损害影响。

这些结果得出的另一个结论是，DDC解决方案能够"线性"扩展运行作业的数量。这与流量感知网络拓扑相比更具优势，因为在流量感知网络拓扑中，作业数量的扩展会对通过网络运行的现有作业产生影响。

检查故障场景的影响

下面的实例是256个GPU规模的场景，用于检查故障场景对DDC拓扑的影响。请注意，总体上，根据定义而言，DNX架构在结构访问处理器（FAP）和整体结构元件（FE）之间的链路上具有固有的过载能力（oversubscription capacity）。



下表显示了单脊交换机和双脊交换机故障场景中测得的JCT。结果显示，有效带宽下降与结构元件故障导致的带宽损失百分比一致。

这得出两个结论：

- 故障发生后，由于基于单元的流量分配，流量在剩余上行链路上达到均衡。
- 平均结果与第100个结果之间的最小差异值，表明不同测量值之间的差异可以忽略不计。

256个节点的DNX架构，单个作业，交换机受损

配置	作业	作业	第100个
256个节点的DNX架构，一个FE交换机故障受损	1.249	1.095	1.097
256个节点的DNX架构，两个FE交换机故障受损	1.562	1.362	1.365

在ECMP等基于流的技术会受到交叉干扰且预测结果差异较大的场景中，DDC解决方案则显示了一致且独立的测试结果。

人工智能集群解决方案对比表

	专有	以太网			
	InfiniBand及其他	Clos拓扑[本地以太网]	定制的以太网	单机箱	DDC
规模	理论上可行，但尚未大规模使用	高，但性能不够	尚待大规模应用	仅限于机箱	高
带宽和交叉对分带宽	带宽变化周期较慢，交叉对分带宽根据应用调整	接口速度为最新一代，但交叉对分带宽很差	接口速度处于领先水平，但大规模的交叉对分带宽仍是未知数	接口速度落后一步，而交叉对分带宽接近100%	接口速度处于领先水平，而交叉对分带宽接近100%
延迟	正确调整和预配置后可实现超低延迟和抖动	低延迟，对延迟变化没有任何控制	低延迟，对延迟变化没有任何控制	中等延迟，变化最小	中等延迟，变化最小
流量处理的灵活性	差	无论单个还是多个流均低于标准	未知	与流无关，规模有限但很好	与流无关
故障恢复	集中式SDN控制器的故障反应本质上较慢	通过协议进行本地化操作和缓慢传播	通过协议进行本地化操作和缓慢传播	基于硬件，立即恢复	基于硬件，立即恢复
无损行为	基于信用的无损协议	有损，尝试在此基础上模拟无损	有损，尝试通过端点管理模拟无损	无损设备	无损整体结构
遥测和集中管理	作为"岛屿"进行管理。全部为专有且集中式	最初应用遥测技术改善流量行为	更密集地使用非标准遥测	无法遥测机箱内部统计数据	充分利用遥测作为DDC的固有功能

DDC应对人工智能数据中心基础设施面临的挑战

将人工智能引入数据中心给网络层带来了诸多挑战。用于互连人工智能集群的许多网络解决方案都各有利弊，迫使终端用户做出权衡。

DDC似乎能够应对人工智能数据中心基础设施的所有挑战，即使（或许也正因为）DDC主要是为完全不同的网络用例而设计的。DDC与所有其他可选择的网络解决方案之间的一个主要架构差异是DDC不是网络，而是网络元件。

被应用于数据中心领域时，DDC能够扩展为一个极大规模的网络元件。由于具有管理和控制软件，它也仍会是一个逻辑元件。DDC外部是网络接口，内部是整体结构。整体结构行为正是满足计算人工智能服务器阵列需求所需要的。

DDC能够确保运行高带宽工作负载的大规模服务器阵列的无损连接，无差别处理流量，最小化网络故障造成的影响，这些均出于传统、易于管理的类似Clos的拓扑。

DRIVENETS

DriveNets是云原生网络软件和网络解耦化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