新型网络芯片技术

New Network Chip Technology

成伟/CHENG Wei，王俊杰/WANG Junjie，
杨勇涛/YANG Yongtao

（苏州盛科通信股份有限公司，中国 苏州 215125）
(Suzhou Centec Communications Co., Ltd. Suzhou 215125, China)

DOI：10.12142/ZTETJ.202406011
网络出版地址：http://kns.cnki.net/kcms/detail/34.1228.TN.20250123.1445.001.html
网络出版日期：2025-01-23
收稿日期：2024-10-17

摘要：针对大规模AI模型训练的高强度通信需求，从高性能交换架构、高性能端口、低时延、无损流控和多维负载均衡等关键技术维度，提出新型网络芯片的整体解决方案。通过整合先进网络架构与多种优化手段，该方案能够有效降低通信开销，提升训练效率，推动AI和高性能计算的规模化应用落地。

关键词：网络芯片；高性能交换架构；高性能端口；低时延；无损流控；负载均衡

Abstract: Aiming at the high-intensity communication requirements for large-scale AI model training, an overall solution of new network chip is proposed from the key technical dimensions of high-performance switching architecture, high-performance ports, low-latency, lossless flow control, and multi-dimensional load balancing. By integrating advanced network architecture and various optimization means, this solution can effectively reduce communication overhead, improve training efficiency, and promote the large-scale application of AI and high-performance computing.

Keywords: network chip; high-performance switching architecture; high-performance port; low latency; lossless flow control; load balancing

引用格式：成伟, 王俊杰, 杨勇涛. 新型网络芯片技术 [J]. 中兴通讯技术, 2024, 30(6): 68-73. DOI: 10.12142/ZTETJ.202406011
Citation：CHENG W, WANG J J, YANG Y T. New network chip technology [J]. ZTE technology journal, 2024, 30(6): 68-73. DOI: 10.12142/ZTETJ.202406011

1 新型网络芯片产业现状

随着ChatGPT等生成式人工智能（AI）的爆发式发展，AI大模型的参数规模从百亿、千亿到超万亿量级增长，这对算力资源提出了空前的需求。在Scaling law原则下，模型训练使用的算力卡数量也从万卡级别向十万卡、百万卡发展。与之对应，智算网络规模也需要同步扩大，以支持更大规模的高速无损互联。

当前，大规模智算网络互联面临两个主要挑战：一是网络设备单点带宽容量需要大幅提升，从400G、800G、1.6T向更高性能演进；二是组网更大规模演进，支持万卡、十万卡集群互联，确保端到端通信的可靠性。

高性能数据中心网络（DCN）已经通过采用Leaf-Spine（叶脊全互联架构），实现了机间网络的扩展，提高了网络的扩展性和可靠性。然而，DCN在性能优化方面仍存在以下不足：

1）带宽利用率不高。由于负载均衡和流量调度的局限，网络资源未得到充分利用。

2）时延不可控。动态负载下的拥塞和排队时延增加，影响网络的可预测性。

3）无损传输难以实现。传统传输控制协议/互联网协议（TCP/IP）难以避免丢包和重传，在高性能计算中成为瓶颈。

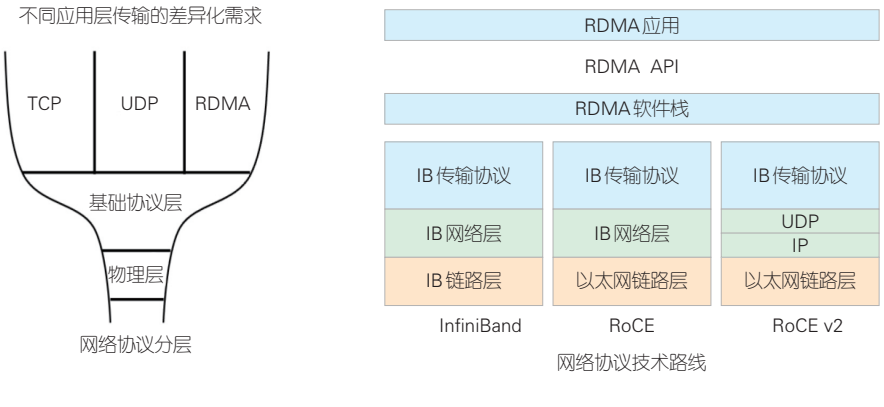
如图1所示，InfiniBand具有高带宽、低时延的特点[1-2]，但在实践中仍面临以下挑战：

1）成本高昂。专用的硬件和协议增加了部署和维护成本。

2）生态封闭。与以太网等主流网络技术兼容性差，这限制了其应用范围。

3）扩展性受限。在大规模系统中，InfiniBand网络的架构复杂度和成本成倍增加。

从机内到机间的互联扩展面临性能损失与复杂度的挑战。机内互联协议是针对芯片或单机内部的高带宽、低时延需求进行优化的，当其扩展到多节点甚至跨集群网络时，不同层级的网络通信需要不同协议转换与适配。此过程不仅引入额外的时延和系统复杂度，还造成资源利用率的下降，难



▲图1 网络协议分层与技术路线

以满足高性能计算对低时延和高效互联的要求。

1）规模扩展性。片上网络（NOC）扩展性差，难以适应大规模计算集群。网络部署方案对路由算法和拓扑设计提出了更高的要求。高速串行计算机扩展总线标准（PCIe）的拓扑结构主要为树状或点对点连接，当节点数量增加时，会出现带宽瓶颈和路由复杂度增加的问题。

2）协议兼容与转换。机内互联技术通常使用专有协议（如NVLink），而机间通信则依赖于融合以太网协议（如RoCE）。从机内互联扩展到机间互联时，必须进行协议转换和适配。这不仅增加了系统复杂度，还会引入显著的时延开销，影响高性能计算的整体性能。

2 新型网络芯片关键技术

新型网络芯片的关键技术包括高性能交换架构、高效能物理层、无损级低时延、双向联合流控、多维负载均衡、开放生态底座。网络芯片技术的发展，只有通过整合先进架构与多种网络优化技术，才能有效应对未来高性能计算和AI训练中的通信瓶颈与传输挑战。

1）高性能芯片架构。高性能芯片架构是网络交换芯片的核心。通过采用高性能的交换架构设计，网络芯片可以实现高吞吐量和低时延的数据包处理能力，满足大规模并行计算对高速数据交换的需求。

2）高性能端口。高性能端口是实现高速数据传输的关键。高速SerDes（串行器/解串器）和四电平脉冲幅度调制（PAM4）高效调制技术，使得网络芯片具备更高的物理传输速率、更低的传输损耗。

3）低时延。低时延是高性能网络的核心指标。通过简化数据处理流程，降低通信开销，尤其是在网络丢包的情况下，网络芯片能够配合流控保障无损不丢包。这样可以显著降低传输时延和业务整体时延，满足AI和高性能计算对低时延和无损的要求。

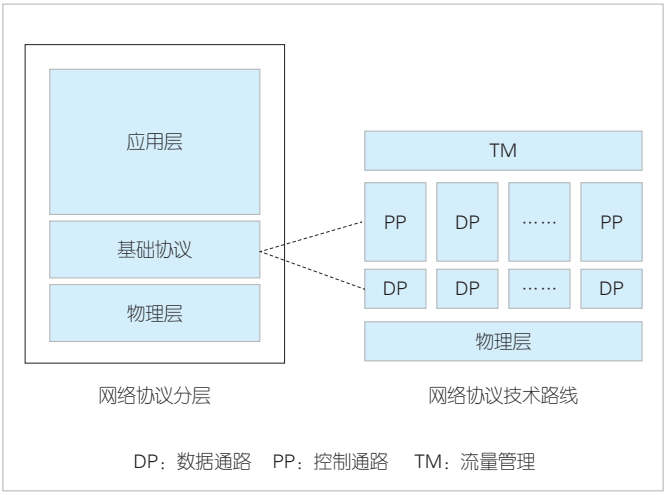
4）无损流控。无损流控是通过基于优先级的流量控制（PFC）在发送端和接收端之间建立协同机制，并根据不同的业务对数据流实施精细化的水线配置。这样可以有效防止网络拥塞和丢包，确保网络的可靠性和稳定性。

5）多维负载均衡。通过感知网络的多维状态，结合动态负载均衡算法，系统可以将本地路径决策提升到全局维度的智能调度和路径选择。这样可以有效避免网络热点和单点故障，提升整体网络吞吐量和资源利用率。

6）开放生态底座。通过支持标准化接口规范和开放标准协议，达到异构系统互联互通3个目标：通过标准协议确保系统互操作性、基于统一开放接口降低开发门槛、基于可扩展架构支持功能演进。

2.1 高性能交换架构

高性能交换架构是满足未来AI算力中心发展的基础。其中，网络芯片带宽的持续提升是必不可少的一环。如图2所示，高性能大带宽网络芯片采用多核架构，通过控制通路（PP）和数据通路（DP）按需分配来提升架构灵活性。相比于单核架构，多核架构不仅增加了前端设计的复杂度，还因



▲图2 网络协议分层与网络芯片架构

PP和DP的交互带来额外的设计开销，同时也给芯片后端布局带来多种挑战。

高性能网络芯片主要组成部分为：物理层、数据通路（DP）+控制通路（PP）+流量管理（TM）、芯片内存Memory。

1）物理层。物理层主要包含端口Serdes串行器/解串器、物理编码子层（PCS）、多媒体访问控制（MAC）等模块，这3个模块决定了芯片对外体现的端口形态和数量。

2）DP。DP是报文接收发送的物理通道，在一定程度上决定了芯片内部带宽性能。

3）PP。PP主要包含入方向PP和出方向PP。报文进到DP并在接收到一定长度后，生成一个Message，再被送往入方向PP进行处理；报文载荷是会继续接收并存储在内部缓存。

4）TM。TM主要是管理报文在Buffer中的存储和读取、队列管理。

5）芯片内存。芯片内存是指业务化的、查表的内存。根据介质不同，一般分为静态随机存取存储器（SRAM）和三态内容寻址存储器（TCAM）。TCAM常见的是访问控制列表（ACL）、掩码路由表。静态随机存取存储器主要用于二层桥接转发表、主机路由表、下一跳编辑表。

在高性能网络芯片架构设计中，性能、功耗和面积是必须考虑的3个核心指标。为实现性能、功耗、面积（PPA）的最佳平衡，需要针对具体应用场景进行权衡，以满足实际应用需求。

1）性能。通过优化芯片架构和数据通路，提升数据传输效率。

2）功耗。采用低功耗设计策略，如电源管理、时钟门控等来降低能耗。

3）面积。利用先进工艺制程和高密度集成技术，在控制芯片面积的同时增加功能模块。

4）工艺。采用更先进的半导体工艺，以降低功耗和芯片面积。

5）模块化。优化功能模块化设计布局，以提高芯片面积利用率。

2.2 高性能端口

高性能端口主要负责传输和接收数据，它将数据链路层的相关报文进行封装/解封装，在数据包之间添加/删除间隔（IPG）和起始定界符，并对传输的数据帧进行编/解码。根据对应的端口速率、传输介质类型，我们将数据转换为电信号或光信号，并通过介质发送/接收对端。

随着单芯片交换容量的提高，单端口转发能力也在不断提高。网络芯片具备全面支持400 Gbit/s端口的同时，还支持800 Gbit/s、1.6 Tbit/s的高性能端能力，这对芯片端口物理层设计提出了新的挑战。

高性能以太网端口的标准经历了多次演进，2010年的IEEE 802.3ba定义了100 Gbit/s端口；2017年的IEEE 802.3bs定义了200 Gbit/s和400 Gbit/s端口；2024年的IEEE 802.3df定义了800 Gbit/s和1.6 Tbit/s端口[3–4]。根据以太网“摩尔定律”，端口速率平均每2～3年翻一番。

如图3所示，芯片端口物理层包括物理编码子层（PCS）、物理介质连接层（PMA）、物理介质相关层（PMD）。

1）PCS。PCS主要对数据进行编/解码，并对结果进行校验。PCS位于MAC层和PMA层之间，将MAC送来的数据进行物理层编码后再送给PMA，再将PMA送到PCS的数据进行解码后再发送给MAC。

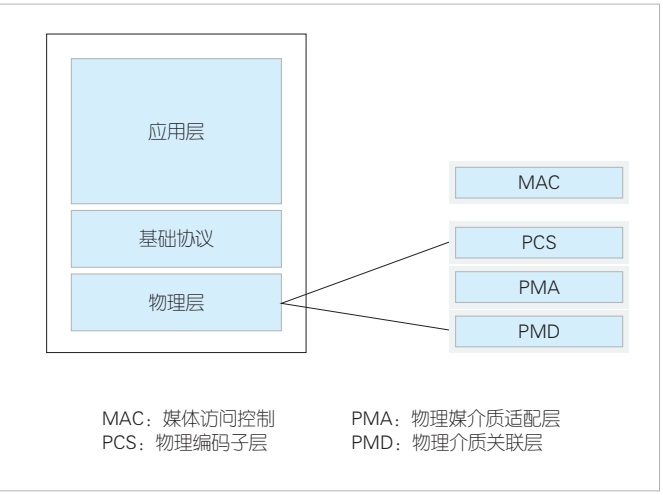
2）PMA主要用于电路的串/并转换。PMA子层集成了SerDes串行器/解串器、发送和接收缓冲、时钟发生器及时钟恢复电路。

3）PMD主要负责将串行信号转换到相应的物理介质上。物理层的PCS子层通过介质无关接口连接数据链路层，并通过PMD对外连接物理介质。

2.3 低时延

低时延是高性能网络的核心指标。实现低时延，不仅需要从芯片设计、简化业务模型处理流程等方面入手，更关键的是在网络丢包的情况下，通过网络芯片硬件级的检测和重传错误数据包。这样可以避免重新发送整个数据包，减少通信时间。

面向AI新型网络的低时延，不仅是指在网络轻负载情况下的单包测试时延，还指动态负载的实际时延，即数据流



▲图3 网络芯片与高性能端口

的完成时间[5]。如图4所示，从网络芯片时延优化的视角来看，可以将网络设备转发整体时延分解为静态时延和动态时延。

静态时延包括数据串行时延、设备转发时延和光电传输时延。这类时延由网络芯片的转发能力和链路传输的距离决定，具有确定的量级。芯片队列缓存和丢包重传对网络时延的影响是动态不可控的。

如图5所示，网络芯片时延的影响因素为芯片主频、数据包长、直通转发、端口形态、业务配置模型、实际流量模型、模块串行设计、模块并行设计。

芯片主频是直接影响芯片带宽的因素，芯片主频越高，转发带宽越大，时延就越低。芯片主频受限于工艺，随着先进工艺的提升，芯片主频也在刷新，但并不是线性的提升。同时，昂贵的流片费用也不断攀高。

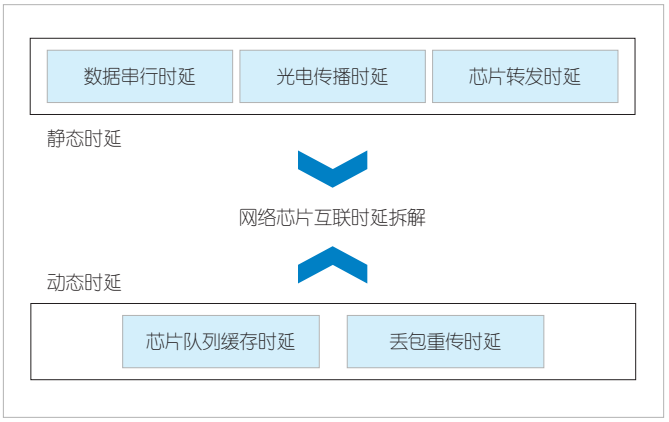
在同等条件下，相比于100G到100G端口的转发时延，全端口400G端口的转发时延会有所降低。不同业务的流量模式也会导致时延差异，64字节小包长的转发时延相比4 096字节长包的转发时延更低。

直通转发是指在数据包完整接收前即可进行后续处理。如图6所示，数据包从数据接收处理引擎直通至数据存储转发控制模块。当接收到预设数据长度（如128字节）时，系统产生信号并将数据送入下一模块处理，同时启动芯片查表转发。此时，当前数据包的后续部分持续存入数据存储转发控制模块的存储器中，并按预设长度分段送入下一级模块处理。

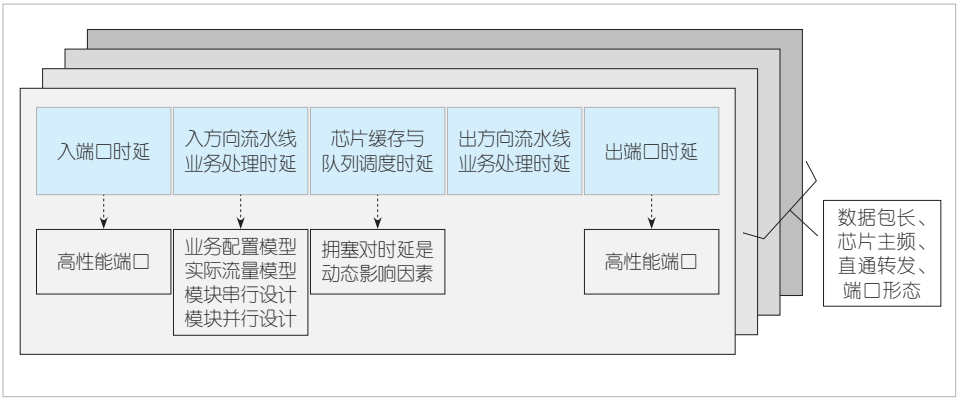
2.4 无损流控

无损流控是指PFC。通过在发送端和接收端之间建立协同机制，双向联合流控能够根据不同的业务需求对数据流进行精细化水线配置。这样可以有效防止网络拥塞和丢包，确保网络的稳定性与可靠性。

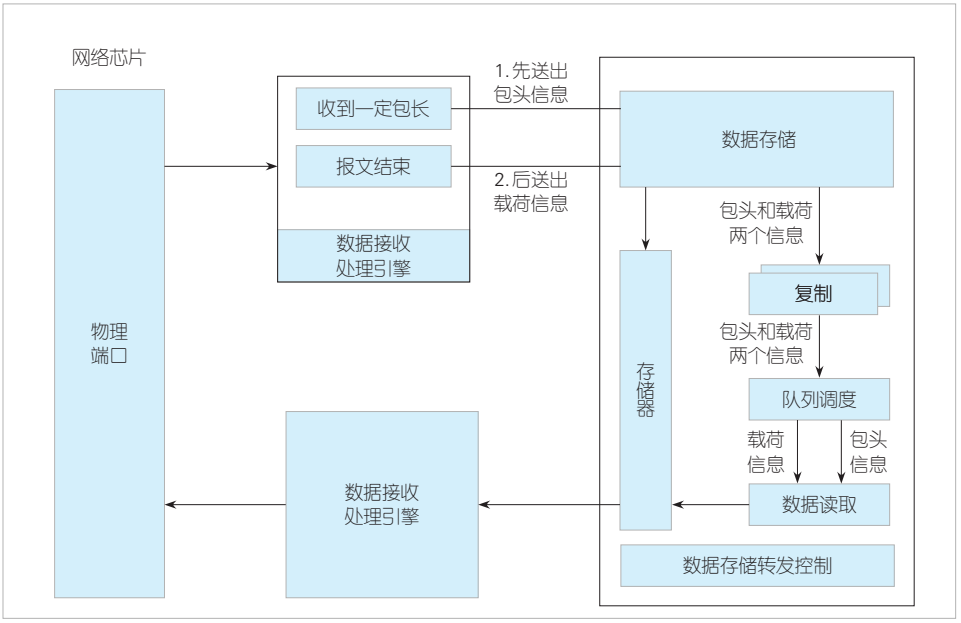
如图7所示，传统远程直接内存访问（RDMA）无损以太网采用PFC来处理拥塞场景下的丢包[5]，并使用PFC+显式拥塞通知（ECN）的方式。这在一定程度可以提前感知芯片队列拥塞，并及时通过数据中心量化拥塞通知



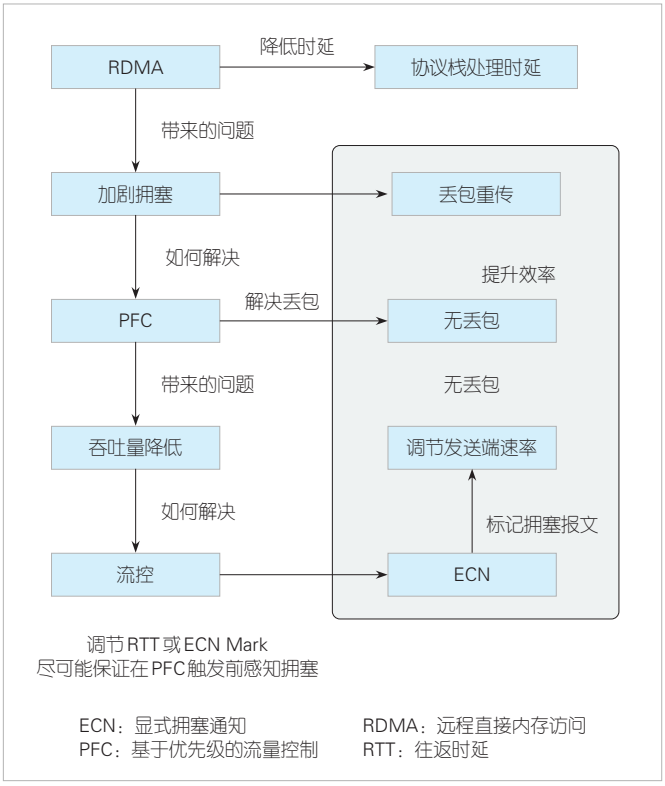
▲图4 网络互联与芯片时延



▲图5 网络芯片时延影响因子



▲图6 网络芯片直通转发模式



▲图7 远程直接内存访问无损流控机制

（DCQCN）调节发送端速率[6]，以缓解拥塞并减少丢包的出现。PFC后向流控本质上是无法解决拥塞和反馈不及时的问题。在大规模部署时，运维团队还面临PFC死锁的风险，过多的PFC Pause会降低吞吐量，同时水线的配置和调整也会给运维带来挑战[7]。

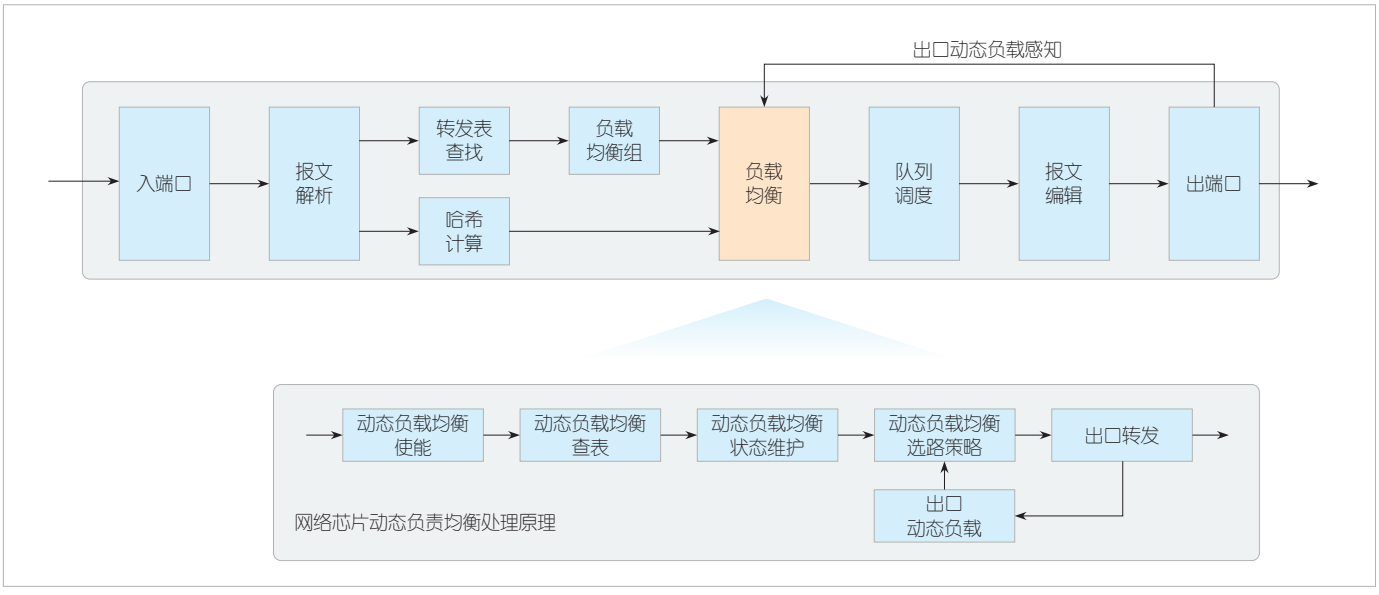
2.5 多维负载均衡

多维负载均衡是指通过感知网络的多维度状态，结合动态负载均衡算法，将本地路径决策升维到智能调度和路径选择。这样可以有效避免网络热点和单点故障，提升整体网络吞吐和利用率。

传统静态负载均衡（SLB）可以根据报文哈希值分配流量，但无法动态感知网络负载状态[8]。等价多路径路由（ECMP）无法区分不同流量的包长和带宽差异，无法感知出口动态负载。例如，当发往同一机柜顶部交换机的不同网口的转发流量通过负载均衡选路到同一下一跳设备时，该设备的下行端口容易出现拥塞。单节点的本地负载均衡算法难以解决全局冲突问题，而负载均衡哈希冲突会导致流量重叠，引发负载不均衡。为解决这一问题，动态负载均衡应运而生。

考虑到设备端口利用率，动态负载均衡将不同的子流分配到不同的端口。如图8所示，芯片动态负载均衡的处理流程为：在入方向流水线转发查表以获取下一跳负载均衡组，根据数据流特征计算哈希作为流量标识，组合生成动态负载均衡流表索引，最后基于动态负载均衡流表的维护状态更新与选路信息。动态负载均衡选路机制是关键，可以支持多种选路策略并最终选择负载均衡出口。芯片监测负载均衡出口带宽/出口队列深度的动态负载，并根据阈值配置来划分质量等级。

全局负载均衡通过引入全局网络路径信息来支持更智能的流量分配。设备接收远端设备通告报文，解析之后配置芯片表项设置出口权重。本地芯片根据收到的远端质量表，结



▲图8 芯片动态负载均衡的原理

合本地的端口质量表，计算出最终的转发出口权重。全局负载均衡能够识别人工智能网络业务的远程直接内存访问流量，并进行调度优化。在网络拥塞时，它可以通过全网智能调度来保障多路径负载均衡，实现全局业务流量的高吞吐，避免网络拥塞对远程直接内存访问性能的损失。

3 结束语

总体来看，新型网络芯片的发展需要集成高性能交换架构、高性能端口、低时延、无损流控和多维负载均衡等关键技术。只有这样，才能有效应对未来高性能计算和人工智能训练中的通信瓶颈与传输挑战。

展望未来，在开放生态系统的构建上，网络芯片技术的进一步发展将依赖于产业界多方协同推进创新。网络芯片通过支持标准化的开放协议和接口，实现与主流网络技术的深度兼容，从而推动产业链的协同合作与发展。通过构建开放、协作的生态系统，网络芯片技术能够更快速地响应市场需求，推动创新，并最终为各行各业的数字化转型提供强大的网络基础设施支持。未来，随着开放生态底座的不断完善和扩展，网络芯片技术将迎来更加光明的发展前景，为构建下一代智能互联网络奠定坚实的基础。

参考文献

[1] Infiniband. Infiniband architecture specification [EB/OL]. [2024-10-13]. https://www.infinibandta.org/ibta-specification/
[2] GUO C X, WU H T, DENG Z, et al. RDMA over commodity ethernet at scale [EB/OL]. [2024-10-12]. https://dl.acm.org/doi/10.1145/2934872.2934908
[3] Ethernet Alliance. 2024 Ethernet roadmap [EB/OL]. [2024-10-10]. https://ethernetalliance.org/technology/ethernet-roadmap/
[4] IEEE. IEEE 802.3 ethernet working group [EB/OL]. [2024-10-10]. https://www.ieee802.org/3/
[5] 开放数据中心委员会. ODCC无损网络技术白皮书 [R]. 2017
[6] IEEE. The lossless network for data centers [R]. 2018
[7] 刘军, 韩骥, 魏航, 等. 数据中心RoCE和无损网络技术 [J]. 中国电信业, 2020(7): 76-80
[8] 沈耿彪, 李清, 江勇, 等. 数据中心网络负载均衡问题研究 [J]. 软件学报, 2020, 31(7): 2221-2244
[9] 毛鹏轩. 下一代网络拥塞控制关键算法的研究 [D]. 北京: 北京交通大学, 2013

作 者 简 介

成伟，苏州盛科通信股份有限公司副总裁；负责产品、战略和产业生态，主要研究方向为高性能网络、确定性网络、边缘计算网络、可编程网络、光电融合网络领域。

王俊杰，苏州盛科通信股份有限公司标准总工；负责技术标准工作，主要研究方向为高性能互联、网络协议、开源系统等。

杨勇涛，苏州盛科通信股份有限公司资深总监；负责网络交换芯片的技术市场与推广工作，对软件定义网络与白盒交换机有深入研究。