中国移动
China Mobile

新一代智算中心网络
技术白皮书

（2022 年）

中国移动

2022 年 11 月发布

目 录

前 言

2022 年 2 月 18 日，国家正式启动“东数西算”工程，突显了数字经济在国家发展中的战略地位。IDC 预测，数字经济的占比将持续增加，到 2022 年，全球 65%的 GDP 将由数字化推动；在中国，到 2025 年，在新基建等战略驱动下，数字经济占 GDP 的比例将超过 70%。数据在未来企业的成长过程中扮演越来越重要的角色，对数据价值利用的深度将决定企业数字化转型高度。而算力是数字经济发展的基础设施和核心生产力，是国家经济发展的重要基础设施。据《2021-2022 全球计算力指数评估报告》显示，计算力指数平均每提高 1 个百分点，数字经济和 GDP 将分别增长 3.5‰和 1.8‰。

算力网络是联接算力供给端和需求端的重要桥梁，也是未来经济发展的重要衡量指标之一。“算力为中心，网络为根基”，网络贯穿算力的生产，传输和消费全流程，一张具有超大带宽、超低时延、海量联接、多业务承载的高品质网络是关键。

本白皮书主要研究智算中心发展情况、智算中心网络发展趋势以及满足智算中心发展需求的智算中心网络关键技术，希望通过在超大规模网络关键技术、超高性能网络关键技术、超高可靠网络关键技术以及网络智能化关键技术等方面的探索，为未来面向智算中心的新型网络架构提供参考。

本白皮书由中国移动通信研究院牵头编制，联合编制单位：华为技术有限公司、上海云脉芯联科技有限公司、中科驭数(北京)科技有限公司、中兴通讯股份有限公司等。

1. 智算中心发展情况

1.1. 政策形势

当下，新一轮科技革命方兴未艾，各行各业开启全面数字化。大数据、云计算、人工智能、区块链等数字化技术落地应用，对计算能力提出更高要求。算力，与经济社会发展的联系愈发密切，成为驱动产业变革的新兴动力。

信息技术浪潮推动人类社会由“电力时代”迈向“算力时代”，以算力为根基的智能化数字经济世界即将来临。为打造经济发展新高地、应对国际激烈竞争、抢抓战略制高点，近年来，党中央、国务院高度重视数字经济发展，推动算力相关技术研发，加快部署各类算力中心。

2020 年 4 月 20 日，国家发展改革委首次明确新型基础设施范围，将智能计算中心作为算力基础设施的重要代表纳入信息基础设施范畴。随着 AI 产业化和产业 AI 化的深入发展，智算中心受到越来越多地方政府的高度关注并开展前瞻布局，已成为支撑和引领数字经济、智能产业、智慧城市、智慧社会发展的关键性信息基础设施。中国智能算力占全国总算力的比重也由 2016 年的 3%提升至 2020 年 41%，预计到 2023 年智能算力的占比将提升至 70%。

2021 年 5 月 24 日，国家发改委等四部门联合发布了《全国一体化大数据中心协同创新体系算力枢纽实施方案》，明确提出布局全国算力网络枢纽节点，启动实施“东数西算”工程。今年 2 月，“东数西算”上升为国家战略，国家发改委等部门确定了 8 个国家算力枢纽节点，并规划了 10 个国家数据中心集群。

政策方面的扶持和激励，特别是东数西算工程的全面启动，给智算中心的快速发展注入了强大的助推剂。智算中心承载以模型训练为代表的非实时性算力需求尤为适合实施"东数西算"，以智算中心为算力底座，在我国东西部地区开展人工智能领域的算力协同合作，“东数西训”将是我国推动"东数西算"工程落地的重要场景之一。

1.2. 产业趋势

近年来，自动驾驶、生命医学、智能制造等领域发展迅速，随之而来的是超大规模人工智能模型和海量数据对算力需求的不断提高，智算中心建设正当其时。

工信部数据表示，截至 2021 年底，我国在用数据中心机架总规模达 520 万标准机架，

在用数据中心服务器规模 1900 万台，算力总规模超过 140 EFLOPS。全国在用超大型和大型数据中心超过 450 个，智算中心超过 20 个。

据不完全统计，从 2021 年 1 月 1 日到 2022 年 2 月 15 日，全国共有至少 26 个城市在推动或刚刚完成当地智算中心的建设，其中投入使用的有 8 个，包括南京、合肥等地的智算中心。除了这些投入使用的，全国至少还有 18 个城市签约、开工、招标、计划建设智算中心项目，包括深圳、长沙的项目都已经开工建设。

其中几个典型的智算中心规模如下：

8 月 30 日，阿里云宣布正式启动张北超级智算中心。该智算中心总建设规模为 12 EFLOPS（每秒 1200 亿亿次浮点运算）AI 算力，将超过谷歌的 9 EFLOPS 和特斯拉的 1.8 EFLOPS，成为全球最大的智算中心，可为 AI 大模型训练、自动驾驶、空间地理等人工智能探索应用提供强大的智能算力服务。

在 WAIC2020 大会期间，商汤科技宣布，上海“新一代人工智能计算与赋能平台”临港超算中心启动动工。该算力中心占地面积近 80 亩，总投资金额超过 50 亿元人民币，一期将安置 5000 个等效 8000W 的机柜。算力中心建成并投入使用后，总算力规模将超过 3700PFLOPS, 可同时接入 850 万路视频，1 天即可完成 23600 年时长的视频处理工作。

南京智算中心采用浪潮 AI 服务器算力机组，搭载寒武纪思元 270 和思元 290 智能芯片及加速卡。目前已运营系统的 AI 计算能力达每秒 80 亿亿次（AI 算力远超传统数据中心提供的基础算力供给），1 小时可完成 100 亿张图像识别、300 万小时语音翻译或 1 万公里的自动驾驶 AI 数据处理任务。

1.3. 技术趋势

随着算力经济的发展，以及人工智能产业越来越成熟，各种专用算力芯片在市场上也是呈爆发式发展趋势，对应的智能算力在总算力中的占比也在逐渐提高，传统的通用算力占比在下降。在新一代智能算力集群中，由各种算力协同一起完成一个大规模复杂的计算任务，各种类型的资源首先需要池化，如存储资源池、GPU 资源池等。

服务器作为算力的主要载体，开始踏入了更高速的车道。以 AI 为核心的算力需求激增，多元异构算力增速超过通用算力成为主流。越来越多的行业使用人工智能技术分析、挖掘日常海量数据，以图像、语音、视频为主的非结构化数据导致深度学习模型的规模和复杂性不断增加。到 2030 年，以 GPU、NPU 等为代表的智能算力增长近 500 倍，远超 10 倍增速的

通用算力，成为全球算力主流。随着摩尔定律逼近极限，以 CPU 为主的通用计算性能提升放缓，为保证数据处理效率， GPU、DPU、FPGA 等异构加速芯片将有望取代 CPU 成为智算中心的主算力。

存储系统实现应用数据的持久化，向应用提供数据访问服务。随着社会智慧程度的提高，海量数据收集、分析、处理带来的挑战越来越大，智算中心必须解决好数据“存得下、读得出、用得好”的问题。需要多方面的提升存储能力，首先，存储介质由单一的 HDD 向 SSD、SCM、HDD 等异构存储介质演进，采用高速存储协议 NVMe，满足上层多样化的数据存储需求；其次，需要提升数据存取效率，重点解决处理器内部、处理器和内存、内存和外存以及服务器之间等不同层级数据存取的效率问题，包括提升 L1、L2、3 的 Cache 缓存能力、构建大规模持久内存池、引入 RDMA/DMA 协议等，实现端到端数据存取加速，最终实现降低访问时延、大幅提升传输效率的目的；最后，传统集中式存储在弹性扩展能力等方面存在力不从心，基于通用硬件构建的分布式存储快速发展。

2. 智算中心网络发展趋势

在过去十年，数据中心网络技术经历了两个发展阶段：

（1）虚拟化时代（~2020），以应用为中心，提供远程服务：各类敏捷智能的微服务应用的发展，推进了企业的数字化转型。在这一阶段，分布式和虚拟化技术替代了大型机、小型机，满足了当时企业业务扩展带来的弹性需求，通过 ESXI/OPS/Docker 等虚拟化技术，实现生产系统上云，推动数据中心高速发展。

（2）云化时代（~NOW），以多云为中心，提供云化服务：多云之间算力无损调度需求，推进了云化计算和算力网络发展。在这一阶段，出现了资源池化技术，把计算和存储资源分离，再规模化编排和调度，提供了超大规模的计算和存储资源池。GPU 高速发展、算力普惠，带来算力中心集约化建设，数据中心正从“云化时代”转向“算力时代”。

传统数据中心，面向传统的计算处理任务，或离线大数据计算，以服务器/VM 为池化对象，网络提供 VM/服务器之间连接，聚焦业务部署效率及网络自动化能力。智算中心是服务于人工智能的数据计算中心，包括人工智能、机器学习、深度学习等需求，以 GPU 等 AI 训练芯片为主，为 AI 计算提供更大的计算规模和更快的计算速度，以提升单位时间单位能耗下的运算能力及质量为核心诉求。

智算中心将算力资源全面解耦，以追求计算、存储资源极致的弹性供给和利用，以算力资源为池化对象，网络提供 CPU、GPU、存储之间总线级的高速连接，如图 2-1 所示。智算中心网络作为连接 CPU、xPU、内存、存储等资源重要基础设施，贯穿数据计算、存储全流程，算力水平作为三者综合衡量指标，网络性能成为提升智算中心算力的关键要素，智算中心网络向超大规模、超高带宽，超低时延、超高可靠等方向发展。

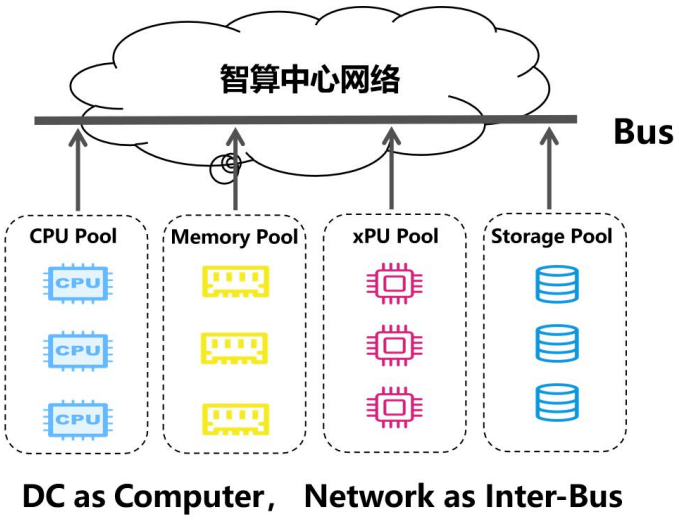

图 2-1 池化总线级智算中心网络

系统级端网协同体系创新是智算中心高性能网络性能提升关键，端侧通过智能网卡硬件卸载网络协议栈，提升网络规模及处理性能，网侧构建低时延、高吞吐的高速通道。如图 2-2 所示，智能网卡与网络设备协同工作，优化拥塞控制算法、网络态势感知、动态路径切换、端到端带内遥测等能力，打造极致的网络性能与运营能力。

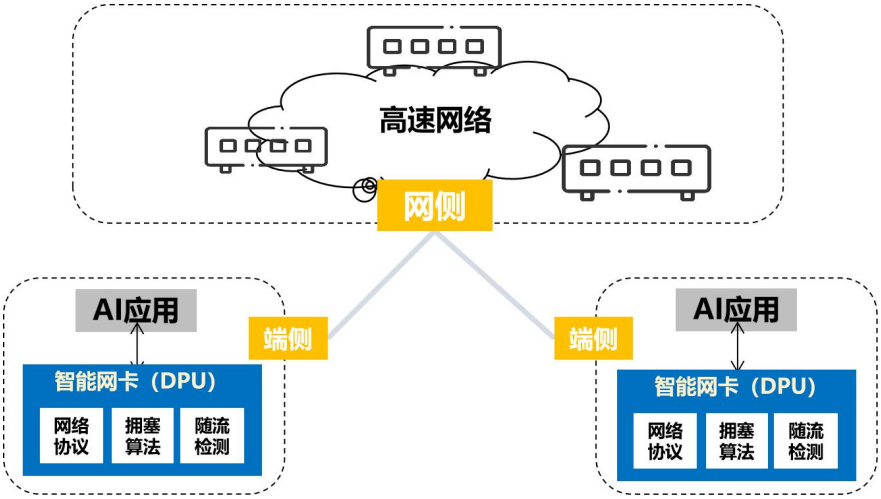

图 2-2 端网协同的下一代高性能网络体系

新一代智算中心将从数据中心的内部做体系化创新，从以往的以云为中心，进入以 AI 为中心的体系架构。元宇宙、生命科学等超大算力需求呈现爆发式增长，超大算力中心、异构算力协同应运而生。但新一代智算中心网络当前还面临四大关键挑战：

1、超大规模网络

随着 AI、5G、IoT 等技术的爆炸式发展，海量数据流的产生和多元化的应用场景为智能计算产业带来发展机遇。在这一过程中，基于 CPU 架构和工艺提升的创新日益趋缓，已无法满足新场景下多样化算力快速增长的需求，算力提升的核心动力正从 CPU 扩展到以 NPU（Neural-Network Processing Unit，嵌入式神经网络处理器）、VPU（Vector Processing Unit，矢量处理器）、GPU（Graphics processing unit，图形处理器）等为代表的计算单元。XPU 直出以太等技术持续发展使得计算/存储资源实现解构。未来会出现融合以太、总线、信元技术的超融合网络，满足计算/存储/内存池化需求。智算中心内节点的数量将增长 10 倍，从现在的十万台服务器增长到百万台 XPU 互联。

2、超高性能网络

当前 AI 应用已采用 GPU 甚至专用 AI 芯片，计算速度相比传统 CPU 提升 100~1000 倍之多。同时 AI 应用计算量也呈几何级数增长，算法模型向巨量化发展，人工智能模型参数在过去十年增长了十万倍，2025 年向百万亿参数模型演进，训练数据集规模百倍增长。同时，存储介质 SSD 访问性能较传统 HDD 已提升 100 倍，而采用 NVMe 接口协议的 SSD（简称 NVM 介质），访问性能相比 HDD 甚至可以提升 10000 倍，在存储介质大幅降低的情况下，网络时延占比从原来小于 5%上升到 65%左右，这意味着存储介质有一半以上的时间是空闲通信等待。如何降低计算通信时延、提升网络吞吐是新一代智算中心能够充分释放算力的核心问题。

3、超高可靠网络

算力资源边缘部署逐渐成为产业趋势，自动驾驶、智能工厂、机器协作、远程医疗等 2B 行业蓬勃发展，对业务高速切换数据不中断等提出新的可靠性要求。百毫秒乃至秒级网络故障对集中式存储、分布式数据库等业务会造成影响，如 OLTP 在线交易类业务，网络故障时交易都失败，甚至会影响节点状态，降低系统可靠性，出现分钟级的业务中断。业务中断会给企业及社会带来重大损失，新一代智算中心超高可靠能力不可或缺，故障收敛性能需提升至亚毫秒级。

4、智能化网络

LinkedIn 最新数据显示，网络故障持续增加：人机接口变为机器与机器间的接口，网络

不可视；网络、计算和存储边界模糊，定界困难；数据海量，网络故障难以快速定位和隔离。同时，由于应用策略及互访关系日益复杂，传统的网络运营和运维手段已无法适应智算中心网络的发展，需要引入新的智能引擎，依托大数据算法，对应用流量与网络状态进行关联分析，及时准确地预测、发现、隔离网络故障，形成网络采集、分析、控制三位一体的闭环系统。同时，依托 Telemetry 以及边缘智能等技术，网络设备数据可实现信息的高速采集和预处理，主动上报智能引擎，为业务网络提供自愈能力，实现新一代智算中心网络智能化。

3. 智算中心网络关键技术

3.1. 超大规模网络关键技术

3.1.1. 新型拓扑

5G、万物互联的智能时代产生海量数据，算力要求快速增长，算力扩容成本高昂，需要支持超大规模组网实现集群高速互联。当前智算中心网络通常采用 CLOS 网络架构，主要关注通用性，无法满足超大规模超算场景下低时延和低成本诉求，业界针对该问题开展了多样的架构研究和新拓扑的设计。

如图 3-1 所示，直连拓扑在超大规模组网场景下，因为网络直径短，具备低成本、端到端通信跳数少的特点。64 口盒式交换机 Dragonfly 最大组网规模 27w 节点，4 倍于 3 级 CLOS 全盒组网。以构建 10 万个节点超大规模集群为例，传统的 CLOS 架构需要部署 4 级 CLOS 组网，端到端通信最大需要跨 7 跳交换机。使用 Dragonfly 直连拓扑组网，端到端交换机转发跳数最少减少至 3 跳，交换机台数下降 40%。同时，通过自适应路由技术实时感知网络流量负载，动态进行路由决策，充分利用网络链路带宽，提升网络整体吞吐和性能。

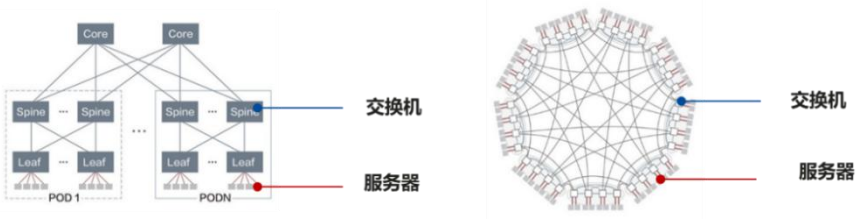

图 3-1 CLOS 和直连拓扑组网架构图

3.1.2. 高效能 IPv6 演进

随着机器学习、人工智能大模型的快速发展，AI 训练集群内的节点数量及所需的 IP 地址越来越多。同时业务应用逐步采用容器、Serverless 等部署方式大大提升了智算中心内计算资源的虚拟化比例，导致智算中心内需要的 IP 地址数量呈指数级上升。但是全球可供分配的 IPv4 协议地址已经枯竭，所有的运营商不能再申请到公网的 IPv4 地址池。这将促使为移动终端和固定终端申请 IPv6 地址，以支撑各种业务的开展，实现万物互联和智能连接。

传统数据中心通常采用 VxLAN 技术提供多租户及跨 TOR 的子网内 IP 地址互通能力，若智算中心网络采用 IPv6 Over IPv6 的 VxLAN 隧道将会在原始 IPv6 报文基础上增加 70~74 字节的封装。双层 IPv6 报文头导致报文封装成本上升、转发能效下降，假设原始 IPv6 报文（仅包含 IPv6 基本头）转发能效为 1，如图 3-2 所示，对于 Payload 长度小于等于 256 字节的报文，IPv6 VxLAN 封装的转发能效出现明显下降。

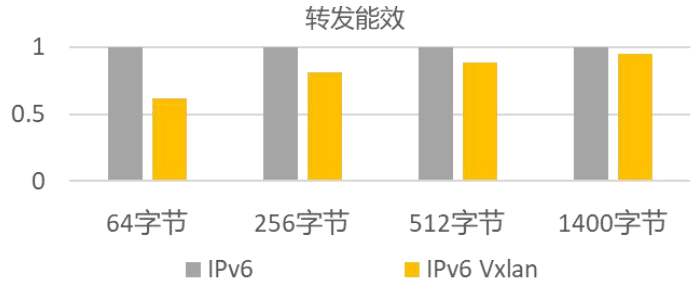

图 3-2 IPv6 和 IPv6 VxLAN 转发能效对比图

智算中心 IPv6 网络中，报文无需添加 Underlay IPv6 头部封装，仅需增加一个 IPv6 扩展头（12 字节）的封装成本，网络转发能效远超 IPv6 VxLAN 封装、接近原始 IPv6 报文，如图 3-3 所示：

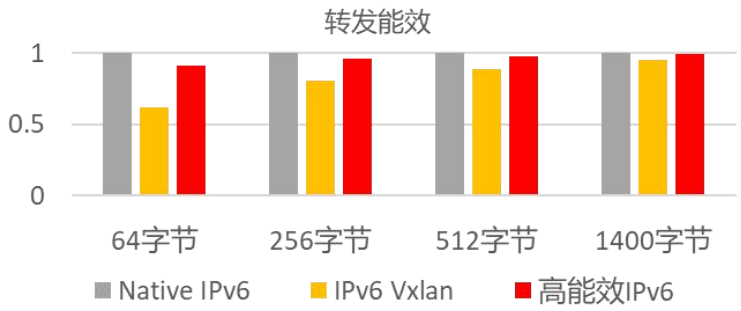

图 3-3 高能效 IPv6 转发能效对比图

智算中心网络存在业务多租户及安全等要求，不同业务、不同安全级别、不同租户间的业务根据需要进行隔离/互通控制。智算中心 IPv6 网络中，通过 IPv6 扩展头携带租户标识、安全组标识及业务信息，可以支持智算中心内及跨智算中心的租户隔离/互通、微分段及业务链能力。

3.1.3. 智算中心间网络连接

随着国家东数西算战略的推进以及越来越多的分布式算力协同场景的出现，AI 算力已经不再局限于单一的智算中心内部，更多的新型计算任务需要依赖“横向互联”和“纵向延伸”的多智算中心协同完成，通过跨智算中心网络连接在逻辑上形成算力层面的超级虚拟智算中心。

智算中心之间的长距连接成为影响业务性能的关键。为了支撑高效的数据搬移，相较于普通广域网，互联网络提出了更高的要求：

1、超高的带宽利用率。大管道是算力时代的标配。核心算力中心间几百 G 甚至上 T 的链路将带来超高的成本。充分利用带宽，减缓扩容节奏，将成为超长距连接的首要目标。

2、超低的丢包率。极少丢包甚至零丢包将极大减少丢包重传带来的带宽资源消耗，在高带宽利用率的同时，保证有效吞吐，提升数据搬移效率。

然而，现有网络技术面临多方面的挑战，无法满足算力网络需求：

（1）上千公里的长距，带来超长的链路传输时延，网络状态反馈滞后，现有的传输层协议拥塞控制算法存在不足：基于丢包的 Cubic 算法在长距传输表现出低的带宽利用率、同时丢包较多；TCP BBR（Bottleneck Bandwidth and Round-trip propagation time）算法虽然能获得较高的带宽利用率，但丢包率较高。

（2）超长距传输连接数少时，容易损失吞吐。

（3）超长距光纤传输无法避免错包。

（4）超大的带宽时延积 BDP（Bandwidth Delay Product）容易发生拥塞丢包。要想实现无损流控，设备接收端缓存需要大于 BDP，这也对网络设备提出了更高要求。此外，接收端的缓存也会由于丢包导致接收数据块不连续，无法提交给应用，而快速消耗，进而影响吞吐。

为了应对超长距传输的挑战，满足高性能算力互连要求，新一代智算中心内部网络应具备如下的典型能力：（1）传输层协议可硬件卸载，支持超长距的 RDMA。（2）吞吐能力

接近瓶颈链路带宽。（3）支持加密传输。

同时，考虑到智算中心间网络引入了大延时和大抖动，智算中心内的无损网络方案无法做到及时的拥塞控制和流量调整，需要新的技术方案解决。智算中心之间的互联网络可以看作是智算中心内部无损网络的延伸，DCI 网络引入了大延时和大抖动，仅靠智算中心内部的网络方案无法做到及时的拥塞控制和流量调整，需要承载网提供长距无损或者一定的确定性能力，目前业界的研究方向有全光网络直连、广域确定性承载网络、智算中心与承载网智能协同，空心光纤等。引入空芯光纤，不仅可以实现容量距离积的大幅提升，更可大幅降低约 1/3 的传输物理时延、并提高时间确定性，为构建低时延时间保证光互联网络提供基础支撑能力。

3.2. 超高性能网络关键技术

3.2.1. 自适应路由

传统数据中心网络通常采用最短路径算法指导流量转发。对于均匀随机流量，吞吐率和延迟均可达到最优，如遇到持续大象流，最短路径会非常重载，而非最短路径处于空闲状态。

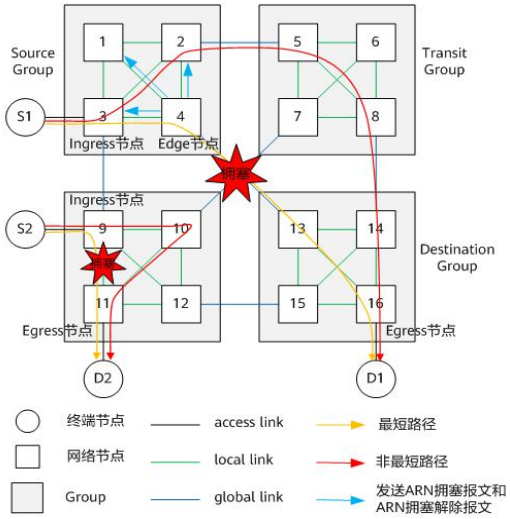

图 3-4 自适应路由转发示意图

如图 3-4 所示，自适应路由的目标是提升整网的有效吞吐以及网络韧性，能够快速感知网络链路负载状态变化，识别出关键拥塞路径，快速调整网络转发路径，做到毫秒/亚毫秒级别的链路快速切换，动态选择轻载链路进行转发，满足超高性能网络的可靠性要求。

3.2.2. 静态转发时延优化

应用时延=计算操作的步数*每步时延，过大的网络延时则直接影响系统性能，严重浪费系统算力。从引起时延的性质来看，网络设备转发时延主要有两部分构成：静态时延、动态时延。

静态时延是指网络设备硬件转发固有的时延，目前随着转发设备的硬件能力提升，静态时延已下降到微秒级，一般都小于 1us。

动态时延是指多打一流量造成网络设备的端口队列拥塞，队列深度增大带来的队列时延，也包括因队列缓存溢出丢包，导致业务报文重传带来的延迟。

如图 3-5 所示，转发芯片主要有如下模块构成，Serdes、PHY/MAC、上行包处理（PP）、缓存管理(BM)、下行包处理（PP）等，报文转发必须经过这些模块。各模块时延分布大致为：Serdes ~30ns, PHY/MAG ~300ns (含 FEC), PP ~400ns, BM ~100ns(直通转发)，各转发芯片模块划分和实现存在差异，该时延分布仅供参考。为进一步降低报文静态转发时延，可以针对各模块进行低时延设计优化。

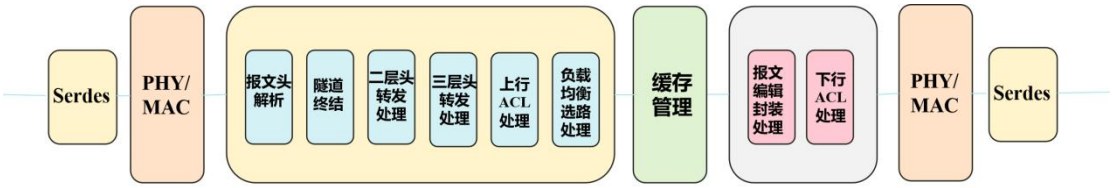

图 3-5 数据中心交换机转发芯片模块构成

- PHY/MAC 模块

高速接口物理链路误码率高，需要通过 FEC（前向纠错）技术实现纠错。FEC 纠错技术需要收齐一定长度的 bit 流（码字）后才能开始纠错处理，这个会带来时延的增加，RS(544,514) FEC 应用在 50G 单 lane 接口时的解码延时为 148ns，时延相当大。为了降低 FEC 纠错时延，业界引入了更短的码字，如 RS272-FEC 相对 RS544-FEC 只需要收齐一半的 bit 流就可以开始纠错处理，解码时延可以减低一半，RS272-FEC 相对 RS544-FEC 纠错能力下降，只能在链路误码率较低的场景使用。为了支持更广泛的场景应用，在保证接口可靠性的同时追求更低的时延，新的接口形态和编码算法有待进一步探索。

- 包处理（PP）模块

不同业务（L2/L3/VxLAN）包处理模块内处理流程差异较大，VxLAN 出入隧道转发相对基本 L2/L3 转发会多查一些转发表，如隧道终结表、隧道封装表，这些额外的处理会带来报文处理模块时延的增加。要降低包处理模块的时延需要简化业务部署，关闭报文转发路径上不需要的子模块，避免部署 VxLAN 业务，设备上未部署下行 ACL 时，可以考虑关闭下行 ACL 功能。

包处理模块内存在较多的查表（MAC 表/FIB 表）过程，主要表项因为容量较大普遍采用算法查找，查表深度也会影响转发时延。为了追求更低的时延，需要探索更好的并行查表设计，高效的查表算法。

3.2.3. 端网协同

3.2.3.1. 端网协同流控

由于网络中流量的随机性以及路径的多样性，拥塞的出现不可避免。网络出现拥塞后，会造成排队时延增大（排队长/丢包高/触发 PFC 等）、网络利用率低（欠吞吐）等影响，导致应用性能出现恶化。现在有很多拥塞控制手段，通过不断调整端侧发送的速率，最终达到进入的网络的容量尽量逼近网络的承载量，来解决网络中的拥塞问题。当前，主要从带宽、时延、收敛速度、公平性等角度评价不同算法。

传统的拥塞控制以被动拥塞控制为主，即收到拥塞信号后被动探测式地调整速率，典型的如 DCQCN 算法，发送端根据接收到的 ECN 标记报文，利用 AI/MD 机制（additive-increase/multiplicative-decrease，线性增速乘性降速）调整发送速率。由于 1 个比特的 ECN 信号无法定量地表示拥塞程度，发送端设备只能探测式地调整发送速率，导致收敛速度慢，性能较差。目前，业界典型的优化思路分为两类：一类是更加精细化的被动控制，如 HPCC（High Precision Congestion Control，高精度拥塞控制），利用相比 ECN 更精细的信息，提高调速的准确率，避免长时试探；第二类是提前预留/主动分配式的主动控制，如 HOMA（一种接收端拥塞控制算法）等，主动为后面的包做资源预留以及分配，避免拥塞的发生。

但是当前主流的优化思路仍然在端侧实现，仍然需要至少 1 个 RTT 的响应时长，同时针对网络中存在的多拥塞点问题，仍然需要多个周期才能收敛。因此需要一种新型的端网协同的拥塞控制算法，网络提供的更精细信息以及更主动的控制，端侧更精准的调控速率，实现满带宽、低时延、快速收敛、公平性优等目标，有效提升网络的传输效率，保障大规模分

布式 AI 任务的高效完成。

在 200 打 1 场景下，不同网络拥塞控制算法对应的缓存排队时延如表 3-1 所示。可见端网协同时的拥塞控制效果最好。

时延（us）	端网协同 CC	HPCC	DCQCN
50%-ile	0.155	3.023	116.612
90%-ile	0.238	6.662	121.82
99%-ile	0.321	8.204	125.48
99.9%-ile	0.401	9.094	127.131

表 3-1 端网协同拥塞控制算法与业界拥塞控制算法仿真实验数据对比

目前业界为满足不同业务场景需求，会开发一些定制化的拥塞控制算法，通过与数据中心交换机协同工作，满足精细化的流量拥塞控制需求，这就对网卡的可编程能力提出新的要求。DPU 具备灵活的网络业务配置能力和可编程的拥塞控制算法开发能力，是实现端网协同，网络流量细粒度调度管理的首选。

3.2.3.2. RoCE 协议改进

RoCEv2 协议作为业界主流远程直接内存访问（RDMA）协议，存在三大限制，对网络传输性能有比较明显的影响：

（1）每连接单路径的限制。RoCEv2 协议每个 RC 都映射到唯一的一对五元组。故障情况下，会导致流量跌落多、流量中断时间长；整网负载均衡性差，导致网络带宽利用率降低；更容易产生拥塞，不能调路，造成时延性能劣化。

（2）硬件 RC 连接数的限制。RoCEv2 将协议栈卸载到网卡中，其中也包括应用通讯的连接关系数据库，但受限于网卡芯片内的表项空间限制，芯片内的连接数有限，当连接数超过某个数量的情况下，就会发生网卡芯片与主机内存的连接表交换，从而导致网络传输性能下降。

（3）Go Back N 重传能力的限制。RoCEv2 协议为保障可靠传输，协议栈实现了重传机制，目前典型的重传机制是 Go Back N 重传，即发生丢包后，从上一次确认接收的位置之后进行全量重传，而不是仅针对丢弃的报文进行有限重传。这也是当前 RoCEv2 依赖开启 PFC

反压的主要原因，由于丢包后重传的代价巨大，需要依赖 PFC 反压尽量杜绝网络上的丢包。

（4）大 QP 规格下流控机制限制。在 QP 数量较多的场景下，基于公平轮询原则，单个 QP 调度时间周期比较长，造成 QP 的 CPN 反馈、QP 升速和降速不及时，从而造成流量控制不精准。

RoCEv2 的这些限制已经越来越广泛的被业界所认知，同时业界也在针对以上限制进行不断的改进，与上述限制相对应，RoCE 协议需在以下方面进行优化改进：

改进 1，支持每连接多路径的能力优化。所谓的每连接多路径是指，可以基于多个五元组的会话进行数据包的传输，连接上的数据可以分担到多个不同的五元组。这样的好处，首先是可靠性的提升，在智算中心 fat-tree 组网存在充分的等价路径的前提下，任意一个单点故障只会影响部分路径的转发，但不会导致整个连接都中断，从而可靠性得到提升。同时网络均衡性会提高，可以使得网络的利用率得到改善和提高，从而提高 RoCE 传输的性能。AWS 已经将多路径技术应用到其自研的协议 SRD 中，并在流量收敛性能上得到了显著的优化。

改进 2，从 RC 模式往连接数依赖更小的模式演进。目前基于 RC 的通讯是为每一对需要通讯的 QP 建立、维护一组连接，因此导致了连接数的规模巨大，限制了组网规模，影响了性能。针对这块有两种思路：思路 1，不提供更粗粒度的传输服务，这方面 AWS 的 SRD 就是基于此思路的尝试，协议栈不提供面向连接的保序传输可靠传输能力，硬件协议栈仅负责可靠报文传递，保序这类复杂的服务由驱动软件完成；思路 2，进行连接的层次拆分优化，构建连接池，实现连接的动态共享，Mellanox 的 DC 技术就是此思路的代表。

改进 3，从 Go Back N 往选择性重传优化。Go Back N 重传是一种简单的重传方式，所以在早期芯片资源受限的情况下硬件卸载的协议栈选择此方式来实现重传，加上有 PFC 加持，一般来说丢包概率非常低（在 PFC 参数配置合理的情况下，一般只会在出现链路错包，链路故障的情况下才会发生丢包），芯片实现 Go Back N 重传不失为一种合理的选择。但随着 RoCE 组网规模不断增加，引发对 PFC 风暴整网流量骤停的担忧，同时半导体工艺的提升帮助网卡硬件芯片能够实现更为复杂的协议，RoCE 的重传方式将会逐渐从 Go Back N 的全量重传演进到选择性重传。

改进 4，基于大 QP 组的拥塞控制。将两个节点间共享同一转发路径 QP 资源归为一个 QP 组，如图 3-6 所示，可以通过五元组或引入带内遥测机制进行识别。一个大 QP 组内所有的信息可以实现共享，如 CNP 反馈信息、速率信息、令牌信息等，在大 QP 组内，实现各个 QP 的速率快速精准控制。当网络出现拥塞或恢复时，QP 组根据自身策略进行速率调

整，策略包括：（1）每个小 QP 单独计算自己的速率，汇总到大 QP 组。QP 组计算一个调整比例系数，告知各个小 QP。（2）QP 组计算出来组速率，分解到各个小 QP，然后告知各个小 QP 具体的速率值。

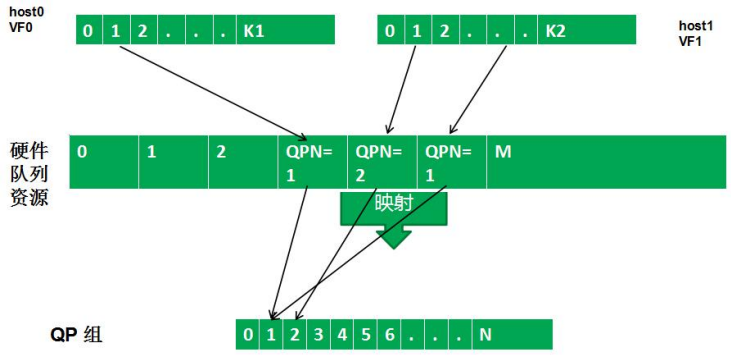

图 3-6 QP 与 QP 组映射关系

当 RoCE v2 协议延伸到更复杂的超长距互联网络时，问题将变得更为复杂。当单一的技术手段无法满足需求的时候，尝试将 AI、PFC、ECN、TDM 等多种技术手段进行融合将是一种必然的选择，采用智能化多维度分析调度的手段才能达到最佳的效果。

3.2.4. 在网计算

近年来，随着深度学习、高性能计算等一批新型应用负载的需求量大增长，导致分布式系统规模越来越大，例如我国的超级计算机太湖之光已达到千万核级别。在计算机科学领域，有一条著名的经验法则，叫做阿姆达尔定律，代表了并行计算之后效率提升的能力。根据阿姆达尔定律，并行系统的加速比受限于串行部分（即无法通过并行加速的部分）的性能。系统规模增大，系统内各节点之间的协同开销也随之增大，加剧了无法通过并行计算加速的串行计算部分的占比。

算力需求的爆炸式增长促进了计算产业的繁荣，例如，过去 8 年，英伟达 GPU 算力增长了 317 倍并持续提升。与算力指数级增长不匹配的是，决定并行计算中串行部分的网络带宽增长却是线性的。数据中心网络带宽从过去的 10Gbps/25Gbps 发展到现如今主流的 40Gbps/100Gbps，增长速度远远落后于算力增长。因此，两者之间的差距鸿沟，需要系统级的网络-应用协同设计才能跨越。

典型的网络-应用协同设计涵盖了高性能计算与深度学习领域广泛使用的集合通信操作，

包括 AllReduce 全规约和 Broadcast 广播。

高性能计算（High Performance Computing, HPC）是指利用聚合的算力来解决复杂的、大规模的科学计算问题，如天气预测、数学建模、物理分析等，其中涉及到多个算力节点之间的小规模数据集合通信操作（mpi incast 现象）。对于小规模数据来说，网络的转发时延是其集合通信时间的主要组成部分，因此网络通信效率将会影响 HPC 应用的完成时间。但是随着聚合算力的规模不断增长、计算复杂度的增加，集合通信中数据交互的次数也会有明显的增长，网络通信效率对 HPC 应用完成时间的制约作用也越来越明显。如图 3-7 所示，以目前较流行的集合通信操作 mpi ring all-reduce 为例，需要 2(N-1)次的数据交互才能完成，其中 N 为参与的节点数量。深度学习同样需要调用 AllReduce 操作进行梯度聚合，且每个节点的传输数据量是深度学习模型尺寸的 2(N-1)/N 倍，当 N 比较大时，传输量接近原始模型尺寸的 2 倍，相当于额外增添了网络带宽的负担。

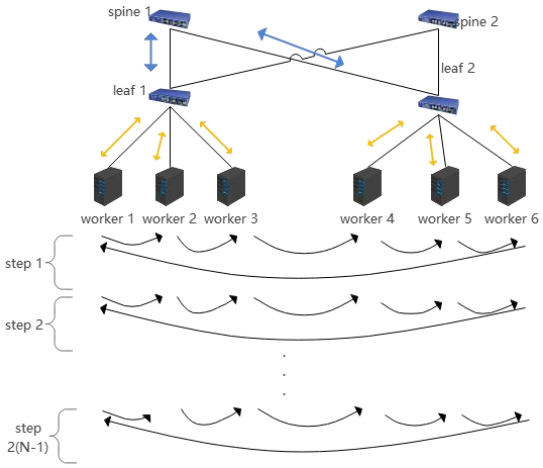

图 3-7 集合通信操作 AllReduce 示意图

近年来，随着可编程交换机的兴起和部署，利用在网计算压缩数据流量，提升计算传输效率成为一个有效的提升分布式系统的方法。在集合通信原语中，Reduce 和 AllReduce 含有计算的语义，因此可以使用在网计算进行加速，减少了数据交互次数和入网数据量。

组播是分布式计算系统中最常使用的通信模式之一。例如，超算系统 Mira 中，MPI_Bcast 原语的执行时间占 MPI 通信总时间的 14%，时间占比在 MPI 集合通信原语中仅次于 MPI_AllReduce。当前 MPI_Bcast 普遍采用应用层组播的方式实现组播通信，即在应用层多次调用下层单播，将数据重复发送多次，从而使得多个目的节点都能获得源节点的数据。由于数据被重复发送，应用层组播任务完成时间大于数据量与通信带宽之比。可靠组播技术利用交换机完成组播报文的复制分发，以网络层组播替代应用层组播，避免了相同数据的重复发送，使得组播任务完成时间逼近理论最优值（即数据量与带宽之比），相比于应用层组播

任务完成时间有约 50%的减少。

3.2.5. DPU 卸载

DPU 作为一种新型可编程异构计算处理器，为高带宽、低延迟和数据密集型新型智算场景提供计算引擎，与 CPU 和 GPU 一起成为智算中心的三大支柱。DPU 作为智算中心内部资源互联的网络端点，是连接异构算力资源，加速数据在 CPU 和存储及网络之间的移动，实现异构算力间数据高速互联互通的关键设备。为了更好的支持智算中心网络，聚合智能算力，提供高性能弹性可伸缩的智能计算能力，DPU 在可以从以下几个方面提升端网协同的网络加速能力。

- NVMe-oF 卸载

基于 NVMe 原生提出的 NVMe-oF（NVMe over Fabric）可以使 NVMe 从支持本地存储（DAS）发展为支持网络存储（NAS）且无需转换其他存储协议，在网络存储中延续保持 NVMe 存储访问低时延、高吞吐的特点。随着存储介质从机械硬盘逐渐向固态硬盘转变，存储介质的访问延时从毫秒量级缩短到几十甚至几微秒，使得存储性能瓶颈从存储介质、网络传输逐渐转移到主机侧对存储网络协议栈的处理。

传统方式下，主机侧 CPU 至少需要运行三层协议栈才能将报文从网卡转发出去。通过 DPU 对 NVMe-oF Initiator 和 NVMe-oF Target 端进行卸载加速，能够有效解决存储性能遇到的瓶颈，在基于 DPU 的存储架构中主机侧只负责发出存储命令，即只需要运行一层存储协议栈。其他协议栈将卸载到 DPU 中执行，降低主机端 CPU 的占用率，是在分布式高性能存储高速发展的趋势下的必然。

根据实现方式不同，NVMe-oF 的加速方案可分为分为半卸载与全卸载两种。半卸载指将原运行在主机端的存储协议栈卸载到 DPU 中的 CPU 核心中处理，结合 DPU 的专用加速单元如加解密、压缩解压缩实现存储的加速。DPU 存储全卸载仍然将运行在主机端的存储协议栈转移到 DPU 中执行，但 DPU 中的 CPU 核心负责配置存储控制器的参数，例如，IO 队列数、队列深度、可并发命令数等。在 DPU 存储全卸载的模式下，主机发起的存储命令将直接通过 DPU，经由网络卸载引擎直接转发出去。类似的，接收网络传来的数据直接经过后端 DPU 的存储加速单元写入主机内存，进一步降低存储访问延时同时提高存储访问的并行度。

NVMe-oF 在 DPU 上实现卸载加速的基础是实现 NVMe 设备虚拟化和 RoCEv2 的大规

模连接能力，考虑 NVMe-oF 的性能最大化，需要在 NVMe-oF Initiator 和 Target 同时实现卸载加速。同时，NVMe-oF 的存储服务能力也是必不可少的，如存储数据压缩/解压缩、加密/解密、RAID 和纠删码（Erasure Code，EC）等。

- GPU Direct RDMA 能力

在当前 GPU 的算力能力下，100Gbps 或更大的数据量才能够充分发挥单个 GPU 的算力。在这样的发展趋势下，基于 RDMA 协议的 GPU Direct RDMA 技术，在 DPU 与 GPU 通信的过程中，可绕过主机内存，直接实现对 GPU 内存的读写能力，并且 DPU 上全硬件实现的 RDMA 能够支持单流百 G 以上的数据收发能力，进而实现了 GPU 算力聚合并且最大化提升了 GPU 集群算力。GPU Direct RDMA 技术已经是当前算力资源总线级互联高性能网络的主流技术。

3.2.6. 智能 ECN

智算中心网络同时承载计算、存储和管理等多种业务流量。不同业务追求目标不同，对网络的诉求不同。传统方式的 ECN 门限值是通过手工配置的，存在一定的缺陷。首先，静态的 ECN 取值无法兼顾网络中同时存在的时延敏感老鼠流和吞吐敏感大象流。ECN 门限设置偏低时，可以尽快触发 ECN 拥塞标记，通知源端服务器降速，从而维持较低的缓存深度（即较低的队列时延），对时延敏感的老鼠流有益。但是，过低的 ECN 门限会影响吞吐敏感的大象流，限制了大象流的流量带宽，无法满足大象流的高吞吐。

结合了 AI 算法的无损队列智能 ECN 功能可以根据现网流量模型进行 AI 训练，对网络流量的变化进行预测，并且可以根据队列长度等流量特征调整 ECN 门限，进行队列的精确调度，保障整网的最优性能。

如图 3-8 所示，支持智能 ECN 的设备会对现网的流量特征进行采集并上送至 AI 业务组件，AI 业务组件将根据预加载的流量模型文件智能的为无损队列设置最佳的 ECN 门限，保障无损队列的低时延和高吞吐，从而让不同流量场景下的无损业务性能都能达到最佳。

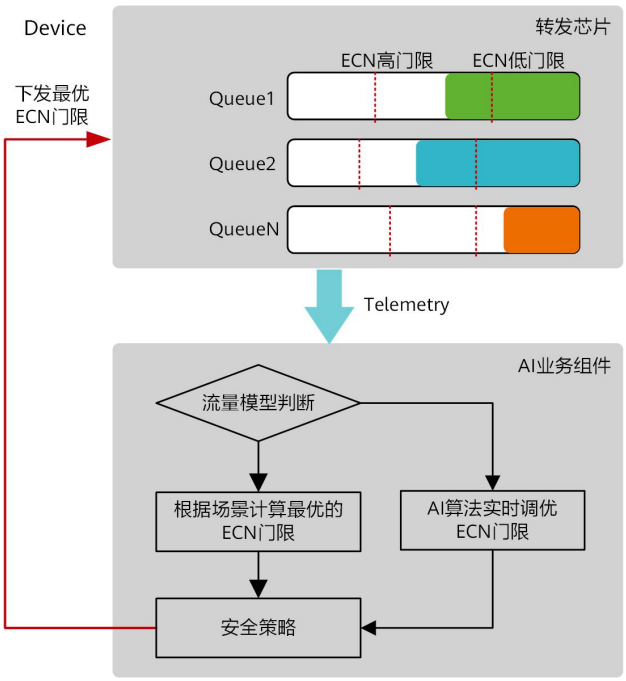

图 3-8 数据中心交换机转发芯片模块构成

1. Device 设备内的转发芯片会对当前流量的特征进行采集，比如队列缓存占用率、带宽吞吐、当前的 ECN 门限配置等，然后通过 Telemetry 技术将网络流量实时状态信息推送给 AI 业务组件。
2. AI 业务组件收到推送的流量状态信息后，将根据预加载的流量模型文件对当前的流量进行场景识别，判断当前的网络流量状态是否是已知场景。如果是已知场景，AI 业务组件将从积累了大量的 ECN 门限配置记忆样本的流量模型文件中，推理出与当前网络状态匹配的 ECN 门限配置。如果是未知的流量场景，AI 业务组件将结合 AI 算法，在保障高带宽、低时延的前提下，对当前的 ECN 门限不断进行实时修正，最终计算出最优的 ECN 门限配置。
3. 最后，AI 业务组件将符合安全策略的最优 ECN 门限下发到设备中，调整无损队列的 ECN 门限。
4. 对于获得的新的流量状态，设备将重复进行上述操作，从而保障无损业务的最佳性能。

无损队列的智能 ECN 功能可以根据现网流量模型进行 AI 训练，对网络流量的变化进行预测，并且可以根据队列长度等流量特征调整 ECN 门限，进行队列的精确调度，保障无损业务的最优性能。

3.2.7. 基于信元交换的网络级负载均衡

基于流的转发负载分担衍生出很多扩展的负载分担方法，比如 ECMP（equal cost multipath）、UCMP（unequal cost multipath），前者不同的路径之间在进行负载均衡选择时完全等价，后者不同的路径在进行负载均衡时会有差异化的权重，至于权重的设定则是可以由控制面逻辑计算而设定。但是不论是何种衍生扩展，他们都存在共同的限制。网络设备在接收到一条流进行转发时，此流经过 hash 计算确定一个转发路径，若不发生网络路径的变化，此流所有的报文都将持续在确定的路径上转发。由于 Hash 计算本身就是一个范围收敛的计算，会导致不同的流选择的路径会有重叠，一般来说网络中流的数量要远远大于路径的数量，通过大量流的叠加，一般来说可以保障网络上各个路径使用相对均衡；但若在网络中流大小极其不均衡、流的数量有限的情况下（一般流的数量规模低于路径数*10^3 就认为流的数量少），不同路径叠加后的流量压力就容易产生较大偏差，这就是大家经常说的负载分担不均衡。针对小规模、大小不均衡流的负载均衡问题，一直是困扰网络数据面转发的难题。

在 AI/ML 的应用中，GPU 或其他类型的 AI/ML 计算单元之间他们有着非常简单的通讯关系（流的数量非常少）；并且由于他们有着极高的计算能力，导致一对通讯单元间的数据吞吐极高（单个流很大，所需的网络带宽极大），这就导致在这样的应用中存在极端的负载分担不均衡，而且这种不均衡一旦引发网络丢包，就会对整体 AI/ML 的任务完成时间带来显著的负面影响。

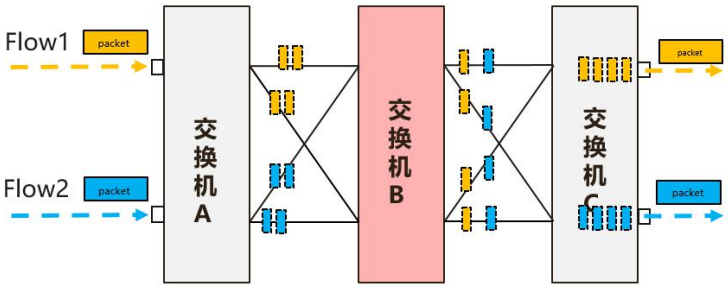

图 3-9 信元转发负载分担示意图

如图 3-9 所示，在基于信元交换的网络级负载均衡机制下，接收端设备接收到报文后，会将报文拆分成的若干信元，信元会基于目的端发送的调度信令选择空闲的链路进行转发，到的目的后，信元被重新拼装成报文发出设备。在这样的机制下，不同于流转发，一个固定的流仅能利用单条路径，交换机 A 和交换机 C 之间的所有链路都可以利用，而且是动态的基于微观负载实时调整的均衡利用。信元交换本身并不是一项崭新的技术，在目前广泛应用

的框式设备中，线卡芯片与网板芯片之间的流量交换普遍都采用了信元交换的技术，以实现机框内无阻塞交换。不过信元交换以前主要应用在框式设备系统内部，往往都是各个交换机设备厂商自定义的信元格式和调度机制，不具备跨厂商互通的能力。此项技术可以进一步扩展，应用到整个网络上，是解决智算中心网络负载均衡问题的方向之一。

3.3. 网络可靠性及智能运维关键技术

3.3.1. 数据面故障感知与恢复

故障收敛是网络保障连通性的重要手段，整个流程依次为：故障感知，即网络设备检测故障是否发生；故障传递，即网络设备间互相通告故障信息；故障恢复，即网络设备重新计算流量路径并引流至新路径。

早期网络故障收敛过程全部依赖控制面，即通过轮询或中断感知物理故障，通过协议保活机制感知链路层以上故障，再由控制面路由协议完成故障传递与处理，所有流程均需要软件参与，典型场景收敛性能为秒级。后来为提升故障收敛性能，业界引入 BFD（双向转发检测）等检测技术来提升故障感知性能，采用 FRR（快速重路由）来提升故障处理性能，其共同特征是将部分故障收敛过程由数据面硬件卸载，降低网络故障场景控制面参与并获得显著的收益，典型场景的故障收敛性能提升至百毫秒量级。

然而随着网络基础带宽的持续提升，以及 AI 计算、高性能存储业务对可靠性的更高要求，百毫秒量级的收敛性能已无法满足业务发展的需求，需进一步降低故障收敛控制面参与度，将故障收敛流程硬件卸载，完全由数据面感知、传递、处理故障，提升故障收敛性能至亚毫秒级。

3.3.2. 基于意图的网络仿真校验

基于意图的网络，本质是围绕用户的意图，借助 AI 和大数据技术，将用户意图转换为网络系统可理解、可配置、可度量、可优化的对象及属性，实现网络设计和运维操作。由意图生成的网络，在下发到物理网络前，理论上是一个逻辑网络，叠加实际的网络数据后，就具备了网络仿真和演算能力。

网络仿真演算技术的本质，首先是通过对于网络配置层面、资源层面和转发层面的建模，形成一张和现网行为无限接近的虚拟网络。然后，在这张虚拟网络通过形式化的数学方法，

快速的验证网络是否能够提供可承诺的 SLA，包括连通性、隔离性、必经路径、转发黑洞、策略一致性、时延丢包。

网络仿真的关键价值在于验证，包括在线配置仿真验证、离线配置仿真验证和事后验收。实际的验证过程是以现网配置、拓扑和资源信息作为输入，通过网络建模和形式化验证算法，基于现网仿真剩余网络资源是否足够、呈现详细的连通性互访关系、数字化模拟用户重大意图的执行、验证意图的预期效果、分析和评估变更对原有业务影响，并持续验证原始业务意图是否已经被满足，进而保障客户网络可靠性。

3.3.3. 智能运维闭环网络

当前数据中心网络运维处于工具辅助、人工决策阶段。传统运维流程存在多个人工断点，首先，人工汇总各个维度的信息判断当前状态是否正常，该信息的来源非常单一，一般主要来自设备自身上报的日志/告警信息以及管理系统采集的性能 KPI 数据阈值告警；其次，NOC 监控人员检测到网络异常，需要通过工单通知网络维护人员，网络维护人员在进行人工的问题定位；最后，网络维护人员基于定位的结果，基于对整体网络的理解给出应急恢复措施以及影响评估。整体的故障处理流程较长，并且依赖运维人员的能力，业务影响较大。

作为人工智能在运维领域的创新应用，智能运维已成为智算中心应对复杂技术架构、严苛运行要求等一系列挑战的必然选择。

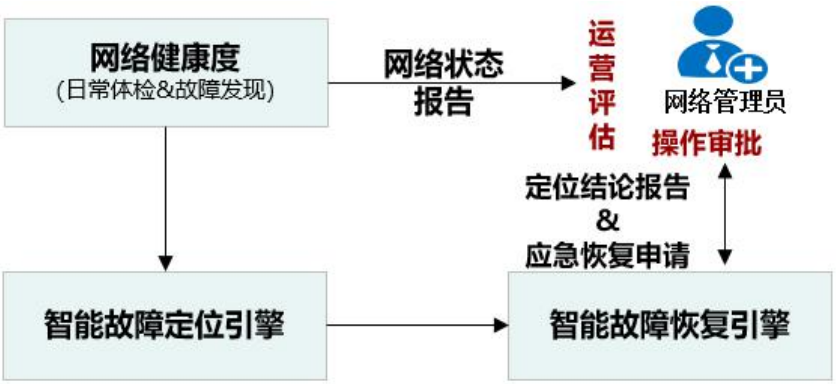

图 3-10 智算中心智能运维流程图

如图 3-10 所示，智能运维主要通过网络健康度、智能故障定位引擎及智能故障恢复引擎等三个模块提供故障预防到故障定位，再到故障闭环的智能保障能力。其中，

（1）网络健康度模块主要对智算中心的物理硬件、网络链路、运行协议、网络业务及业务

流量进行整体的监控，通过 Telemetry 机制，整合网络配置数据、表项数据、日志数据、KPI 性能数据及业务流数据，实时发现网络中各个层面的问题，同时，利用机器学习算法，对网络设备的行为从空间、时间维度挖掘网络状态、流行为预测进行预测，提前发现风险。

（2）智能故障定位引擎通过对多维数据的关联分析，从时间相关性、空间相关性进行关联分析，给出故障的根因。

（3）智能故障恢复引擎综合故障节点及故障类型给出不同的隔离/恢复预案，并且会根据故障节点当前承载的业务情况，给出隔离/恢复的影响范围。

未来新一代智算中心应从网络可靠性、网络一致性、网络容量、网络性能、网络稳定性等多个维度全面分析和评估智算中心网络中可能存在的风险，在网络故障发生之前，根据不同的风险类型提供影响分析及优化建议。

4. 总结和展望

新一代智算中心网络是一个很大的新课题，新协议、新架构、新设备形态、新技术的出现，必然给智算中心网络带来全新的升级。本白皮书从智算中心发展情况、智算中心网络演进趋势和智算中心网络关键技术三个方面，开展了相关研究，以期抛砖引玉，更盼得到更多同行的参与和讨论。

中国移动也希望按照高价值优先、先易后难的原则，逐步推动智算中心网络关键技术成熟与落地，我们期盼与众多合作伙伴一起，汇聚行业力量，共同打造高效、开放、解耦的新一代智算中心网络。

术语与缩略词表

英文缩写	英文全称	中文全称
BDP	Bandwidth Delay Product	超大的带宽时延积
OTN	Optical Transport Network	光传送网
RDMA	Remote Direct Memory Access	远程直接数据存取
AI	Artificial Intelligence	人工智能
VxLAN	Virtual eXtensible Local Area Network	虚拟扩展局域网
PP	Packet Process	包处理
BM	Buffer Management	缓存管理
PS	Parameter Server	参数服务器
HPC	High Performance Computing	高性能计算
DPU	Data Processing Unit	数据处理单元
NVMe	NVM Express	非易失性内存主机控制器接口规范
NVMe-oF	NVMe over Fabric	基于网络的非易失性内存主机控制器接口规范
DAS	Direct-Attached Storage	直连式存储
NAS	Network Attached Storage	网络附属存储
ECN	Explicit Congestion Notification	明确的拥塞通知
PFC	Priority-Based Flow Control	基于优先级流量控制
ML	Machine Learning	机器学习
ECMP	Equal Cost Multipath	等价多路径
UCMP	Unequal Cost Multipath	非等价多路径
ICT	Information and Communications Technology	信息与通信技术
KPI	Key Performance Indicator	关键绩效指标
NOC	Network Operations Center	网络操作中心
CC	Congestion Controlled	拥塞控制