确定性光传输支撑广域长距算力互联

Deterministic Optical Transmission for Wide Area and Long-distance Computing Power Interconnection

王光全[1]，满祥锟[1]，徐博华[1]，吕福华[2]，孟万红[2]（1. 中国联通研究院，北京 100048；2. 华为技术有限公司，广东 深圳 518129）
Wang Guangquan[1]，Man Xiangkun[1]，Xu Bohua[1]，Lü Fuhua[2]，Meng Wanhong[2]（1. China Unicom Research Institute，Beijing 100048，China；2. Huawei Technologies Co.，Ltd.，Shenzhen 518129，China）

摘　要：
高性能算力产业的需求随着人工智能应用的普及和不断发展而持续增加，出现了越来越多的算力协同场景。介绍了算力互联、数据传输中，影响RDMA长距吞吐量的因素，提出了超大带宽及确定性体验的网络解决方案，以实现高性能算力互联。

关键词：
长距 RDMA；全光网；OTN；OXC
doi：10.12045/j.issn.1007-3043.2024.02.002
文章编号：1007-3043（2024）02-0007-07
中图分类号：TN913
文献标识码：A
开放科学（资源服务）标识码（OSID）：

Abstract：
The industrial demand for high-performance computing has been increasing continuously with the development and popularization of artificial intelligence applications，and more and more computing collaborative scenarios have emerged. It introduces the key factors that affect RDMA long-distance throughput in computing power interconnection and data transmission，and a network solution with ultra-high bandwidth and deterministic experience is proposed to acheive high-performance computing interconnection.

Keywords：
Long-distance RDMA；All-optical network；OTN；OXC

引用格式：王光全，满祥锟，徐博华，等. 确定性光传输支撑广域长距算力互联[J]. 邮电设计技术，2024（2）：7-13.

0 引言

2022年1月，国务院印发《“十四五”数字经济发展规划》，提出加快建设信息网络基础设施，有序推进基础设施智能升级，加快实施“东数西算”工程的要求。随着国家东数西算战略的推进，越来越多的算力协同场景以及跨地域大数据搬移场景开始涌现。数据和算力已经不再局限于单一的数据中心，更多的新型计算任务和大量数据需要在多个算力中心间流转并进行算力协同，算力中心间的长距高性能传输能力已成为影响业务性能的关键因素。

收稿日期：2024-01-16

算力互联意味着将算力中心内部的DCN网络进行延伸，典型的DCN网络覆盖范围在10 km以内，且高性能计算DCN网络当前主流的协议为远程内存直接访问（Remote Direct Memory Access，RDMA），由于RDMA协议要求无损传输，当将DCN网络扩展到广域百公里至千公里的范围时，会导致超长的链路传输时延，进而导致网络状态反馈滞后。然而，现有的传输层协议的拥塞控制算法存在不足之处（例如，在长距离传输中，Cubic算法的带宽利用率低，丢包现象较为严重），无法有效地利用带宽。为了应对超长距传输的挑战，满足高性能算力互连的需求，承载网必须具备长距无损确定性传输能力，并且需要与终端侧进行协同，以确保高性能协议的传输效率。因此，如何构

建大带宽的确定性网络以实现千公里级RDMA的无损传输是当前广域算力互联领域的研究热点。

1 RDMA现状及应用于广域算力互联的挑战

1.1 RDMA技术介绍

传统的TCP/IP存在着网络传输和数据处理延迟过大、多次数据拷贝和中断处理、复杂的TCP/IP协议处理等问题。RDMA[1-2]支持本端节点“直接”访问远端节点内存的操作，本端节点可以像访问本地内存一样，绕过传统以太网中复杂的TCP/IP网络协议栈读写远端内存。由网卡直接进行内存读写操作，能够释放CPU算力并降低数据的传输时延，这是一种为了解决网络传输中服务器端数据处理延迟问题而产生的技术。

RDMA有3种传输模式：RDMA Send、RDMA Read和RDMA Write。如图1所示，其协议传输的主要特征是：以数据块为单元，一次把所要传输数据根据PMTU大小进行切片，直到所有数据块传输完毕；采用PSN系列号机制确认数据的完整性，如果有丢包，则进行重传；可以配置多队列、多数据块请求、调整PMTU大小、设置网卡队列缓存大小等参数，提升RDMA的传输效率。针对丢包，采用Go Back N重传机制，检测到PSN序列号丢失时，则请求从该PSN序列号之后的报文全部重传。目前RDMA协议不支持选择性重传，因此，一旦网络有丢包，则无法保证RDMA协议的传输效率。

RDMA技术主要包括IB、RoCE和iWARP。

IB（InfiniBand）：基于InfiniBand架构的RDMA技术，需要专用的IB网卡和IB交换机。

RoCE（RDMA over Converged Ethernet）：基于以太网的RDMA技术，需要交换机支持无损以太网传输，此时要求服务器使用RoCE网卡。

iWARP（Internet Wide Area RDMA Protocal）：基于TCP/IP协议的RDMA技术，由IETF标准定义，目前使用较少。

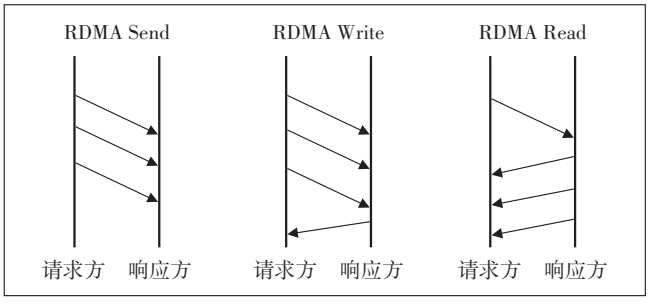

图1 RDMA 3种传输模式对比

目前，IB主要在DC内应用，因为其链路层采用Credit机制，所以无法实现满速率的长距传输。因此，本文主要讨论RoCE对网络的要求及相应的解决方案。在RoCE网络中，为了确保网络传输过程中不丢包，需要构建无损以太网。目前，主要采用2种机制：PFC机制和ECN机制[3]。PFC机制是交换节点逐级向上游设备反压，上游设备缓存报文，若Buffer达到阈值，则会继续向上游反压；ECN机制是报文在网络节点中发生拥塞并触发ECN时，使用IP报文头的ECN字段标记数据包，这表明该报文遇到网络拥塞，并将它发送给源端服务器，源服务器收到后，通过降低相应流发送速率，缓解网络设备拥塞，从而避免丢包。

1.2 RDMA应用于广域算力互联的主要挑战

RDMA技术最典型的落地业务场景是高性能计算（HPC/AI）。为了满足超长距高性能算力互连要求，传统广域基于TCP/IP传输协议的互联网络，面临3个方面的挑战：首先，高性能计算互联单次突发数据量为MB/GB级别的大流，而TCP/IP机制需要把数据切分为小分片（MTU默认1 500），导致有效载荷低；其次，互联网网络采用逐层收敛结构，业务传输跳数多，网络上的数据突发和拥塞都会造成不可预知的时延、抖动和丢包。为保证业务端到端可靠传输，RDMA的丢包重传机制额外耗费了网络带宽，降低了业务吞吐率，进一步导致性能下降。第三，原生RDMA技术对丢包敏感，难以直接用于有损的广域网络传输，因此，需要设计高品质无收敛的网络互联架构与技术，让RDMA数据流可以直接承载在具有确定性品质的无损网络上，中间不再经过多级交换汇聚设备，以减少拥塞，提升吞吐率；考虑到算力互联间带宽以100G~400G的大颗粒为主，适合在源点和宿点之间构筑波长级的一跳直达连接，以避免网络拥塞和丢包导致的效率降低；而广域拉远带来的传输时延是客观存在的，通过确定性的传输时延，与端侧RDMA协议协同调整RDMA传输参数，也是提升RDMA广域传输效率的有效手段。因此通过架构、技术、协议等多方面的优化和改进，可以有效提高RDMA跨广域传输吞吐率。

2 RDMA在广域算力互联的影响因素研究

RDMA的吞吐率受到诸如距离、丢包、QP数量和传输块大小等多种因素的影响。本文基于全光网络

的长距环境，对RDMA的吞吐量进行了研究。验证环境的组网如图2所示，通过OTN全光无损网络提供低于10^{-15}误码率的高质量长距传输链路，包括2条不同长度（200 m和600 km）的光纤链路，这2条链路的带宽均为100 Gbit/s。基于这条OTN链路使用性能测试工具（IB write）进行吞吐量测试。

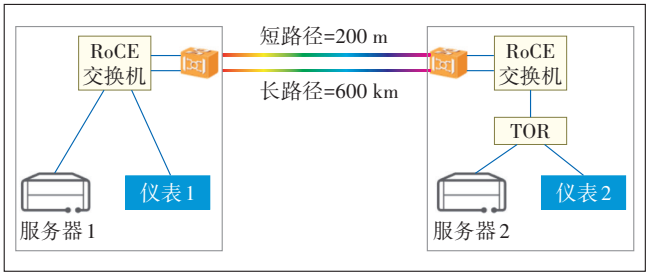

图2　RDMA吞吐量影响测试连接

2.1 距离/时延对长距RDMA吞吐量的影响

为了测试长距带来的时延对RDMA吞吐量的影响，采用OMSP保护方式，构建2条不同长度的光路，一条为200 m，另一条为600 km，默认将OMSP保护组工作在短路由上，即服务器1和服务器2之间的业务流工作在短距离连接上，链路误码率为10^{-15}，服务器1和服务器2通过IB write打流，链路最高吞吐量为100 Gbit/s。具体如图3所示。

然后，通过触发LOS将工作路由切换到长路径上，再次用IB write打流测试。结果显示，吞吐量只有原来的1/10，即约10 Gbit/s。这表明，随着传输距离的增加，ACK回复变慢，导致网卡出口缓存被占满，业务吞吐率下降。在调整IB write参数的情况下，增大RDMA块大小或者QP数量，RDMA在600 km长距离下达到满速100 Gbit/s（百分百吞吐量）。因此，在链路无损的情况下，RDMA协议需要根据传输距离设置合适的QP数量或块大小，以保证长距离吞吐量不下降

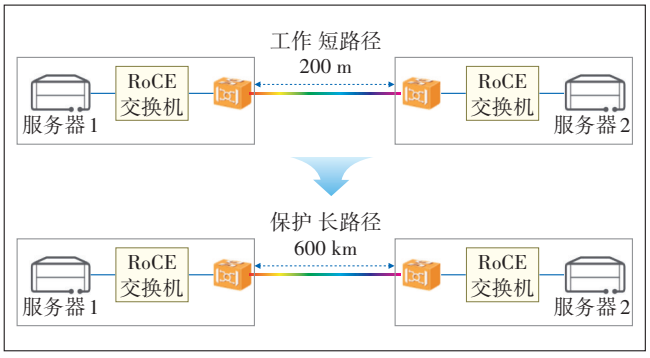

图3　不同距离下对长距离RDMA影响测试

（见图4）。

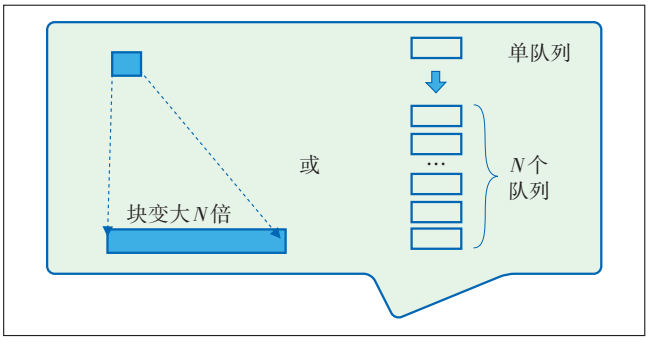

图4　RDMA协议中对应QP数量或块大小调整示意

2.2 丢包对长距离RDMA吞吐量的影响

长距离丢包的主要原因有2类，一类是以太客户侧的丢包，例如尾纤和连接器出问题时导致的丢包，另一类是光线路侧的丢包。长距离传输虽然会出现误码，但是相干系统自带FEC纠错功能，所以光线路侧丢包主要是由瞬断导致的。

以太客户侧的丢包主要原因有：尾纤问题和连接器问题。

a）尾纤问题：尾纤是用于传输电信号的光纤，如果尾纤质量不佳或者使用不当，就会在信号传输过程中出现丢失、反射、噪声等问题，从而导致丢包。例如尾纤损坏或者有污点，都可能在信号传输过程中出现丢失现象，从而导致丢包。

b）连接器问题：连接器是用于连接尾纤和设备的接口，如果连接器质量不佳或者使用不当，就会在信号传输过程中出现丢失、反射、噪声等问题，从而导致丢包。例如连接器损坏或者有污点，都可能在信号传输过程中出现丢失现象，从而导致丢包。

光纤瞬断是光线路侧丢包的典型问题，常见的瞬断原因包括如下3种。

a）光纤质量问题。光纤质量问题是导致光纤瞬断的主要原因之一，光纤的质量直接影响其传输能力和可靠性。如果光纤存在质量问题，如损坏、污染、弯曲度过度过大等，就会导致光纤传输过程中出现短期中断。

b）环境因素。环境因素包括温度、湿度、光照等。在某些环境下，如高温、低温、高湿度、低光照等，光纤的传输性能会受到影响，从而导致光纤瞬断。

c）人为因素。人为因素包括光纤的意外弯曲、拉断、碰撞等，这些因素可能会导致光纤出现短期中断。

由于RDMA对丢包敏感，一旦光纤瞬断导致丢包

频繁出现，会引起RDMA协议层Go Back N机制重传丢包后的所有报文，导致RDMA吞吐率急速下降（见图5）。

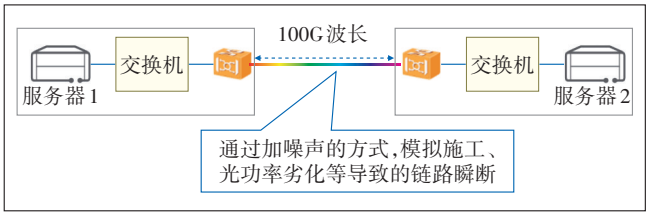

图5　光纤瞬断对RDMA传输的影响

实验室通过在光路上增加噪声的方式，模拟线路出现大量误码造成光纤瞬断的情况。通过抓包观察，发现此时RDMA业务会出现时断时续的现象，测试结果显示，当业务丢包率达到万分之六以上，RDMA的吞吐量会直线下降（见表1）。

表1　丢包率与带宽关系表

丢包率	带宽/(Gbit/s)
万分之六	6.4
千分之一	4.5
百分之一	0.2

2.3 流控机制对长距离RDMA吞吐量的影响

基于优先级的流量控制（Priority-based Flow Control，PFC）是一种能够有效避免丢包的流量控制技术。PFC基于优先级的流量控制，将流量按照优先级进行分类，从而实现对不同优先级流量的控制。当下游设备的无损队列发生拥塞时，下游设备会通知上游设备停止发送该队列的流量，从而实现零丢包传输。

在长距离传输场景中，当宿端发生拥塞时，当前典型的DCN内交换机是小缓存配置，无法支持2倍RTT的流量缓存能力。因此，流控信号需要长距离传输的网络设备进行响应，这就要求OTN传输设备具备PFC流控响应能力，能缓存网络上流量并保证不丢包，同时具备逐级向上反压流量的能力，从而与DCN交换机协同实现长距离无损。

在实验室中构造如图6所示的测试场景。首先，通过2台仪表构建1条25 Gbit/s的背景业务流，然后从服务器1发起流量为80 Gbit/s的RDMA业务到600 km外的服务器2。因为原宿节点交换机端口和OTN均为100G端口，所以在宿端Spine交换机将业务流转发到TOR交换机时，总带宽超过100 Gbit/s，会出现流量拥塞。实验结果如下。

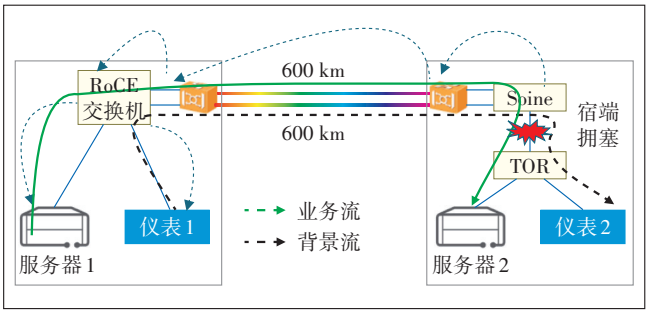

图6　实验室RDMA测试连接示意

a）OTN不开启PFC流控：由于宿端DCN网络发生拥塞并导致丢包，RDMA的重传进一步加剧了拥塞，导致更多的丢包，服务器业务的有效带宽从80 Gbit/s降到9 Gbit/s。

b）OTN开启PFC流控：服务器丢包现象消失，由于OTN收到宿端Spine交换机发送的拥塞信号，首先响应PFC流控，并通过Buffer缓存正在发送的流量。同时，它生成新的PFC信息，向上游发送流控帧通知，逐级反压到达服务器端侧，端侧降速以达到端网协同，防止无效重传。实验室测试结果显示，开启OTN流控功能后，服务器业务的有效带宽从9 Gbit/s提升到76 Gbit/s。

2.4 小结

实验结果显示，物理网络的参数和服务器端侧参数都会对长距离RDMA的吞吐量产生影响。其中，距离是影响RDMA吞吐量的最重要因素之一，随着距离的增加，RDMA的吞吐量会逐渐降低；网络侧丢包/误码也会对RDMA的吞吐量产生影响，当发生丢包时，RDMA的Go Back N机制需要从丢包的位置重新发送后续的全部数据，这导致业务的有效吞吐量显著降低；服务器端侧队列对（QP）数量也会对RDMA的吞吐量产生影响，QP数量越多，RDMA吞吐量就越高；服务器端侧的块大小也是影响RDMA吞吐量的因素之一，较大的块大小可以提高RDMA的吞吐量。

根据上述特征，为了保证RDMA数据传输的效率，防止拥塞、无效重传、时延不稳定导致的性能下降，传输链路应具备确定性的长距离无损能力。

a）传输链路应具备稳定的低时延能力，从源端到宿端光层一跳直达，尽量减少电交换设备，以实现极致低时延，从而最大化传输效率。

b）传输链路应保持低误码率，误码率越低越好，以避免因误码导致的丢包、闪断等重传问题，从而确保性能的稳定。

c）传输链路要避免拥塞，应尽量使用确定性的无拥塞管道传输，防止网络设备拥塞导致丢包影响业务，产生无效重传。

d）传输链路应具备与端侧协同的能力，传输设备应能与服务器端侧之间互通状态信息，当端侧能够感知到物理层状态参数信息时，就能灵活调整RDMA发送参数，从而实现长距离高吞吐量传输。

e）传输链路提供超大带宽能力，缩短搬移周期，促进数字经济的高速发展。

3 确定性光传输广域RDMA解决方案

如图7所示，当前算力中心之间有2种互联方式：一是通过互联网出口互联，这种方式容易受到互联网拥塞、丢包的影响，从而导致RDMA广域传输性能严重劣化；二是通过专线方式互联，这种方式可以解决互联网拥塞等问题，但数据中心内部经过大量的交换机及服务器处理转发，也会导致RDMA广域传输性能受限。以某超算为例，DCN内需要经过约15跳节点处理才能到DC专线出口路由器节点，导致转发处理时延长。

为了实现RDMA广域高性能传输，算力互联网络架构需要优化：构建算间全光高速平面，将DCN网络的Spine/leaf节点直连OTN光传输设备，OTN设备基于物理层参数数据与端侧业务参数协同，实现高吞吐长距离传输。

全光网[4]具备高品质、确定性、高安全、低时延、低抖动等优势，是实现RDMA无损拉远的理想技术，可视为新型算力协同互联的最佳解决方案。通过全光网络承载提供高品质、高可靠的算网保障，可有效保证长距传输时RDMA的高吞吐量，以实现高效算力协同。因此，构建全光算力网方案需要从以下几个方面展开。

3.1 Mesh化组网架构

以算力为中心，规划“1 ms-5 ms-20 ms”覆盖从城域至枢纽的多级时延圈，通过确定性链路带宽、时延和可用率，以及网络端到端硬隔离、安全可靠品质实现分布式算力节点间Mesh化连接。这种连接方式具备灵活高效调度能力，使算力能效最大化。具体如图8所示。

算力节点间互联采取Mesh化、立体化拓扑进行组网，全面部署OXC，通过联动OTN实现光电协同高效调度。链路路由去行政化，减少路由迂回，实现最低的网络时延。枢纽内算力互联以400G/800G系统为主，枢纽间算力互联以单波400 Gbit/s的系统为主，同时具备向单波800 Gbit/s及更高速率演进的能力，频谱从C波段扩展到L波段，单纤容量得到显著提升（相比当前提升4~8倍以上），单位比特的能耗大幅降低，最大化机房、光缆等基础设施的利用率。

3.2 光电协同提供波长级超大带宽，并支持端网协同实现最大吞吐量

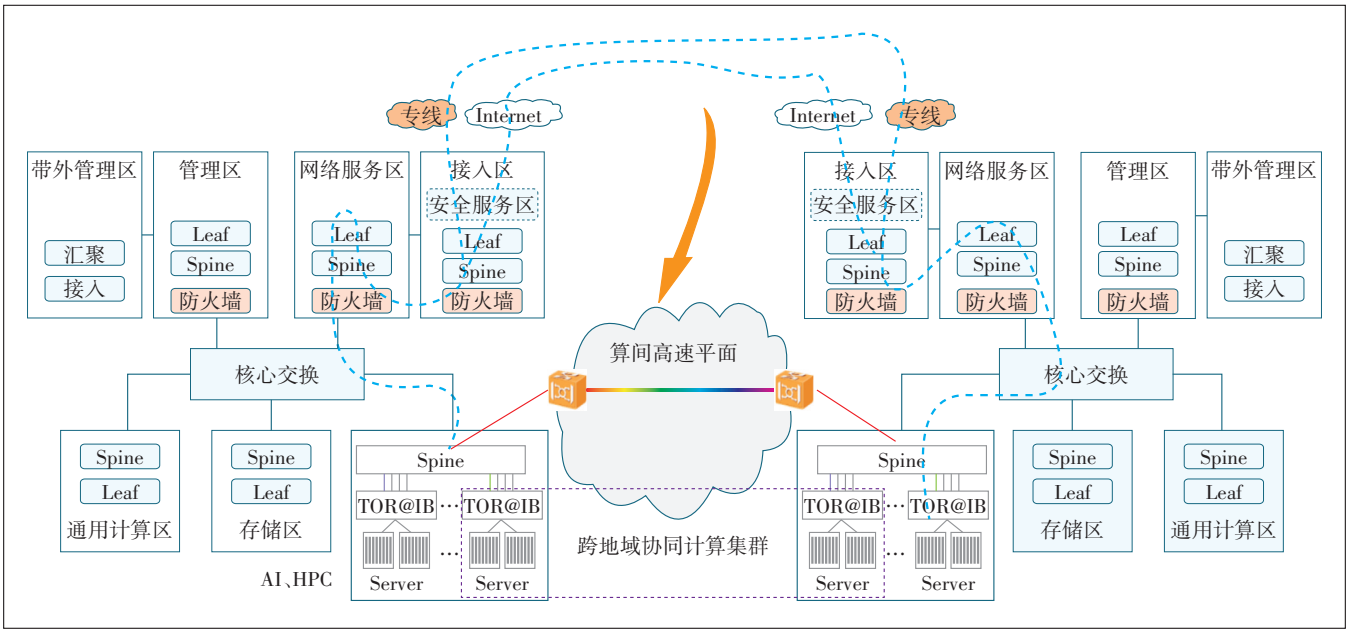

图7　算力中心互联方式

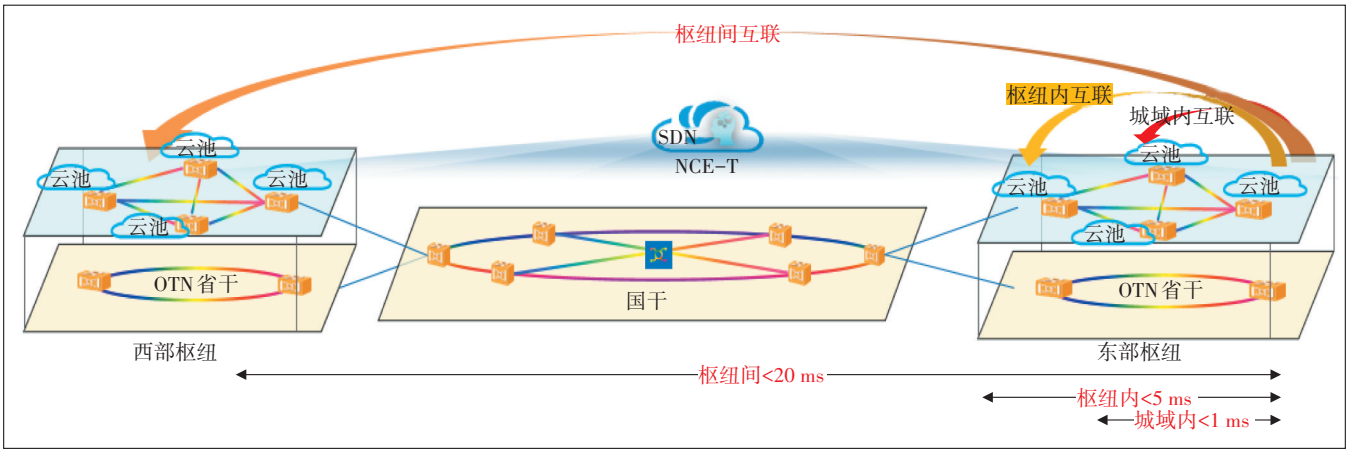

图8　算力节点间组网架构示意

网络需要端到端的波长级调度能力，通过在部分站点预留端口资源，并与超级备波资源一起构建站点资源池化能力，以支持波长在任意方向的无阻塞调度。在网络进行波长级调度或者工作保护路径倒换后，网络的时延等变化需要通知端侧，端侧RDMA根据变化后的时延等调整RDMA的参数（如QP数、块大小、RDMA MTU），以确保RDMA的最大吞吐率。光电交叉协同示意如图9所示。

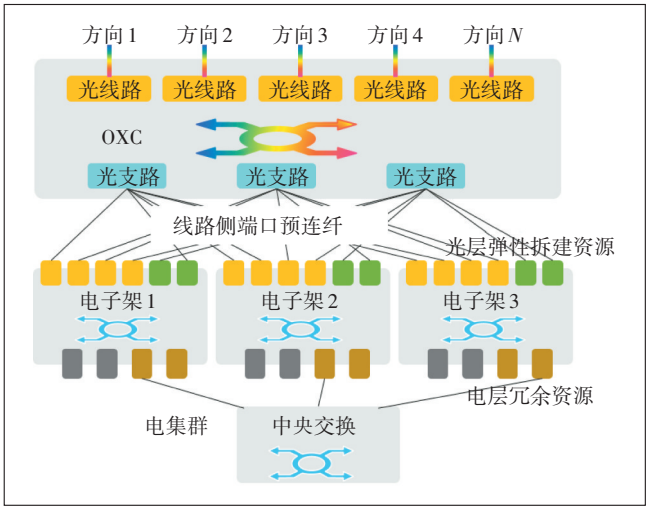

图9　光电交叉协同示意

a）光电跨层协同算路：在光+电协同算路中，首先引入光层数字孪生技术，构建实时、高精度反映BER/OSNR/PDL/SOP/非线性/色散等光学物理量变化的数字光底座，在线评估预开通波长链路的可达性。基于数字光底座，引入光电联动智能规划算路、光电交叉同步配置、光系统自动调测、光性能自动均衡等管控自动化技术，实现光传输L0/L1层协同算路，即根据业务SLA自适应选择线路速率、码型、谱宽等参数，自动计算出满足时延、业务可用率要求的工作、保护光链路（见图10）。

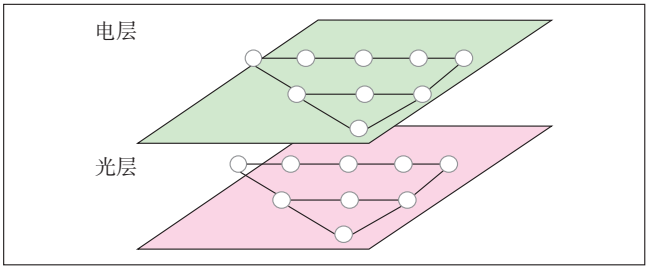

图10　光电协同算路示意

b）光电交叉同步创建：光交叉、电交叉同步打通，业务一次性创建，无需分步骤等待（见图11）。

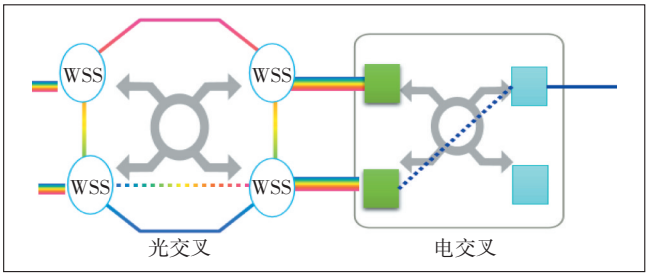

图11　光电交叉同步创建示意

c）自动调测一键式开通：OCH创建后光层自动调测，在线自动化插损预置，无需人工干预，业务自动打通（见图12）。

在光电链路路径切换后，网络链路的带宽和时延都可能出现变化，为了达到最大的吞吐量，RDMA的并发QP对数量和块大小都需做相应的调整。工作保护路径切换同样也存在类似的诉求。在网络路径因为

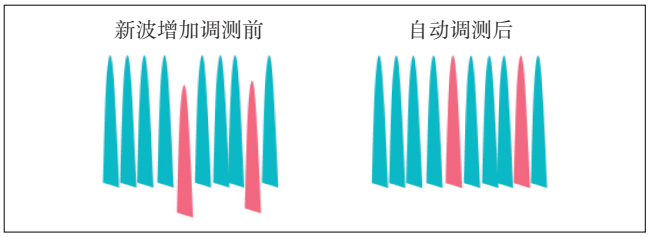

图12　波长调整示意

链路故障导致保护倒换后，网络时延RTT会发生变化，从而导致RDMA传输性能下降。实测表明，工作路径为200 m，RDMA的QP数为1，块大小为20 KB时，RDMA吞吐量即可达到80 Gbit/s。倒换到保护路径（600 km）后，QP数需增加到25个，块大小为1 KB，才能达到80 Gbit/s。所以，在波长调度或者路径保护倒换后，网络将最新的带宽和时延信息通知端侧的RDMA网卡，端侧收到信息后调整QP数和块大小，从而实现最优传输性能。

3.3 高通量RDMA广域无损传输

为解决当前广域网数据传输存在的问题（即采用TCP传输协议导致物理链路传输通量无法得到有效提升问题）和业务节点因网络转发大量消耗CPU算力的问题，建议采用RDMA传输方式替换TCP传输方式，以实现高性能算力互联（见图13）。

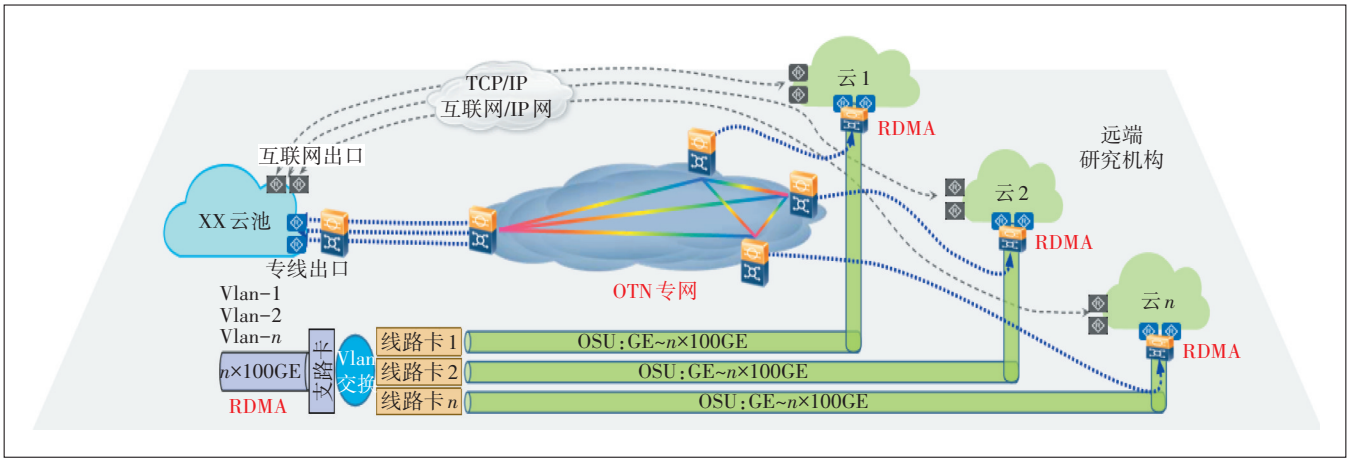

图13　采用OTN承载高通量RDMA广域无损传输示意

利用OTN网络零丢包、稳定低时延、大带宽的承载品质，通过RDMA长距离无损流控技术、端网协同感知，配置最优业务参数等核心技术方案，使RDMA传输协议可应用于广域网下的海量数据搬运，从而使链路通量无限逼近物理链路带宽。

4 总结

长距离RDMA作为新一代广域高性能算力互联的一种技术，是业界的研究热点，然而，目前该技术的技术标准和产业生态都不够完善，需要进一步结合新型全光网络架构提供的超大带宽及确定性体验特性，持续优化RDMA协议。同时，考虑RDMA协议层与全光网络物理层的上下感知联动，以实现超长距离下的高吞吐量无损传输。

参考文献：

[1] MACARTHUR P, RUSSELL R D. A Performance Study to Guide RDMA Programming Decisions [C]//IEEE International Conference on High Performance Computing, Data, and Analytics. IEEE, 2012. DOI:10.1109/HPCC.2012.110.

[2] I. T. Association. Infiniband architecture specification volume 1 release 1.2.1annex A17: RoCEv2, 2014 [EB/OL]. [2023-12-02]. https://cw.infinibandta.org/document/dl/7781.

[3] ZHU Y, ZHANG M, ERAN H, et al. Congestion Control for Large-Scale RDMA Deployments[J]. ACM SIGCOMM Computer Communication Review, 2015, 45(5): 523-536.

[4] 中国信息通信研究院技术与标准研究所. 全光运力研究报告（2022年）[EB/OL]. [2023-12-02]. http://www.caict.ac.cn/kxyj/qwfb/ztbg/202302/P020230217517703022811.pdf.

作者简介：
王光全，教授级高级工程师，长期从事通信网络的规划、设计和研究工作；满祥锟，高级工程师，主要从事通信网络的规划、设计和研究工作；徐博华，高级工程师，主要从事通信网络的规划、设计和研究工作；吕福华，光传输解决方案专家，主要从事光传输网络规划和设计工作；孟万红，2012蓝军实验室高级专家，主要从事长距RDMA、算力网络相关研究工作。