HPN：阿里云大模型训练网络架构

HPN: Alibaba Cloud's Data Center Network Architecture for Large Language Model Training

钱坤/QIAN Kun，翟恩南/ZHAI Ennan，操佳敏/CAO Jiamin
（杭州阿里云飞天信息技术有限公司，中国 杭州 310030）
(Hangzhou AliCloud Apsara Information Technology, Hangzhou 310030, China)

DOI：10.12142/ZTETJ.202406010
网络出版地址：http://kns.cnki.net/kcms/detail/34.1228.TN.20250109.0925.004.html
网络出版日期：2025-01-09
收稿日期：2024-10-16

摘要：介绍了阿里云用于大型语言模型（LLM）训练的数据中心网络架构高性能网络（HPN）。HPN通过双上联、多轨、双平面的网络架构设计，避免了单链路故障带来的严重连通性影响，并且避免了哈希极化的产生。实验表明，HPN将LLM训练的端到端性能提升超过14.9%。HPN已在阿里的生产环境中部署了超过1年。

关键词：大模型训练；网络架构；数据中心网络

Abstract: The Alibaba cloud's data center network architecture for high-performance network (HPN) used in the training of large language models (LLMs) is introduced. HPN is designed with a dual-top of rank (ToR), rail-optimized, and dual-plane architecture, which avoids severe connectivity impacts caused by single-link failures and prevents hash polarization. Experiments have shown that HPN improves the end-to-end performance of LLM training by over 14.9%. HPN has been deployed in Alibaba's production enviroment for over a year.

Keywords: large-scale model training; network architecture; data center network

引用格式：钱坤, 翟恩南, 操佳敏. HPN：阿里云大模型训练网络架构 [J]. 中兴通讯技术, 2024, 30(6): 63-67. DOI: 10.12142/ZTETJ.202406010
Citation：QIAN K, ZHAI E N, CAO J M. HPN: Alibaba cloud's data center network for large language model training [J]. ZTE technology journal, 2024, 30(6): 63-67. DOI: 10.12142/ZTETJ.202406010

大语言模型（LLM）包含超过100亿个参数，并且由多个模型层构成。这些模型的高效训练需要数千个图形处理器（GPU）协同工作。主流的训练框架（例如Megatron-LM[1]和Deepspeed[2]）通常通过多种并行策略的混合来实现大规模训练。

1）数据并行（DP）。训练数据集均匀分布在所有GPU之间，每个GPU都拥有整个模型的一个副本。在每次迭代中，所有GPU都使用AllReduce来同步计算出的梯度。

2）流水线并行（PP）。模型被划分为多个阶段，每个阶段包含一系列连续的模型层，并由不同的GPU提供服务。流水线中的每个GPU都接收来自前一阶段的输入，并将输出发送到流水线中的下一阶段。

3）张量并行（TP）。在PP中，整个模型或每个层都可以进一步水平分割。因此，每个层都分布在一组GPU之间。在每次迭代中，同一TP组中的GPU使用AllReduce/AllGather来同步计算输出和相应的梯度。

考虑到大模型训练过程中的各种并行策略，训练过程中观察到的流量模式与弹性云计算或传统深度神经网络（DNN）训练中的流量模式非常不同，这样的流量模式的差异给智算集群网络带来了新的挑战。

1 AI大模型训练的网络挑战

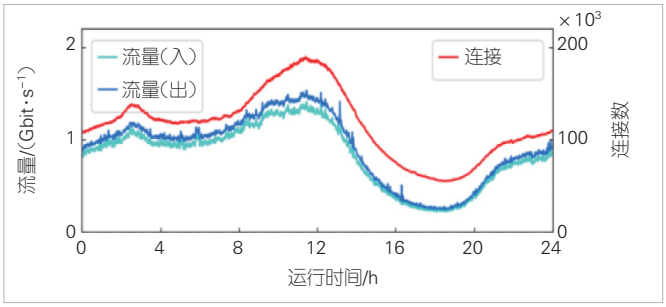
1.1 挑战1：AI训练流量特性导致负载均衡困难

传统的数据中心网络架构（例如fat tree[3]）主要用于一般的弹性云计算。我们观察到，LLM训练的流量模式与一般云计算有所不同。

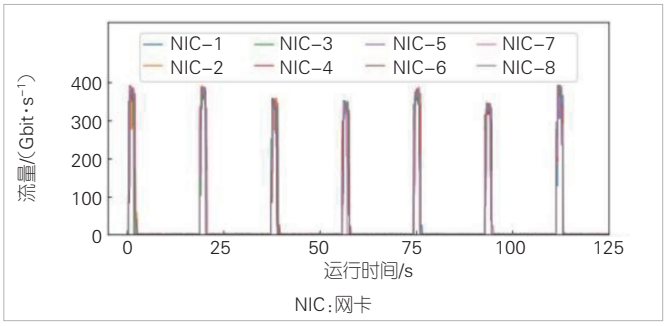
1）网络利用率周期性突发。在实际生产中，传统云计算会产生数百万个流量，流量利用率通常保持在20%以下。整体流量模式相对连续稳定，每小时缓慢变化（如图1所示）。

相反，LLM训练产生的连接极少，但流量存在周期性突发（如图2所示）。图2显示了我们实际生产中LLM训练中一块网卡（2×200 Gbit/s）的吞吐量。网卡周期性地传输大量数据，瞬间就会达到端口最大容量（400 Gbit/s），并会持续几秒到几十秒。在每个迭代中，需要在不同的并行组（每个具有许多GPU）之间进行数据同步。

网络利用率的突发意味着LLM训练需要极高的网络带



▲图1 传统云计算流量模式



▲图2 模型训练期间的网卡出方向流量

宽。因此，我们需要确保LLM训练的网络能够为突发情况提供足够带宽，以避免丢包。此外，流量的同步性表明，LLM训练对长尾延迟特别敏感。任何长尾流量都将阻碍整个集合通信操作的完成，这会使所有参与这次集合通信的所有节点陷入等待。

2）流量数量较少。如图1所示，一般的云计算实例通常会产生数十万个连接；相反，LLM训练中的每个节点产生的连接非常少，一个GPU仅使用几十到几百个连接。结合之前训练过程中提到的突发高网络利用率，每个流需要发送的实际数据量是相当大的。

3）负载不均。传统数据中心网络采用等价多路径路由（ECMP）作为负载均衡方案。ECMP假设哈希算法能够有效地将流量均匀地分布在网络中所有等效路径上，当网络中有大量流数目时，此假设是成立的。然而，在LLM训练中，这种假设不再成立，因为LLM训练仅涉及少量大流。在我们使用传统数据中心进行LLM训练的实践中，已经遇到了很多由此引起的性能问题。

更严重的是，由于传统数据中心网络都采用了3层网络架构，大流的转发需要经过3次哈希计算（即柜顶交换机、聚合和核心层）。由于每次哈希的输入（即流的五元组）保持不变，这种“级联”哈希的影响会导致更严重的负载不平衡（即哈希极化[4]）。我们在生产中也确实观察到了很多由于哈希极化而引起的负载不均的现象。这样的问题在跨Pod通信场景中尤为常见。

1.2 挑战2：AI大模型训练对网络故障更敏感

1）LLM训练对故障更为敏感。在LLM训练中，多个GPU合作完成每个迭代，并且我们需要许多次迭代（持续几十天）来完成整个训练过程。因此，任何一个GPU或主机的故障都可能直接导致整个LLM训练过程崩溃。

2）预防单点故障很重要。传统网络架构中尽管Tier2和Tier3层具有丰富的冗余链路，但每个网卡（NIC）只通过一条链路连接到柜顶交换机（ToR），存在单点故障风险。当接入链路（即连接NIC和ToR的链接）中断时，对应的主机会出现连接断开的情况。更糟糕的是，ToR的故障可以使数十甚至数百台主机不可用，这会导致严重的服务质量下降。LLM训练需要数千个GPU进行协作训练，涉及数十个ToR和数千个光模块和链路。在如此大规模的情况下，几乎不可能保证没有网络设备发生故障。监控和故障排除系统等工具可以在事后定位故障的根本原因，但无法防止训练崩溃。在我们的运行集群中，每个月有0.057%的NIC–ToR链接失败，并且大约有0.051%的ToR交换机遇到严重错误和崩溃。在如此高的故障率下，单个LLM训练作业每个月会遇到1~2次崩溃。此外，每天会发生5 000 ~ 60 000次链路抖动，导致模型性能的暂时下降。

2 HPN网络接入：非叠双ToR上联

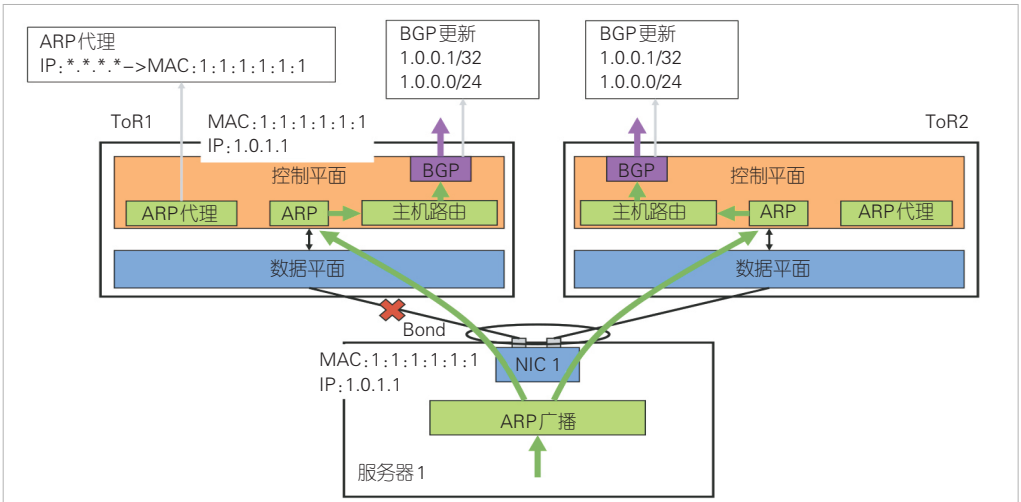
在传统数据中心网络中，每个网卡的两个端口通过一根连接到ToR交换机的电缆/光纤进行汇聚，称为单ToR设计（目前大多数云提供商广泛使用[5]）。然而，单ToR设计非常容易受到交换机/链路故障的影响，严重影响LLM训练。

非堆叠双ToR设计将每个网卡的两个端口以主–备方式连接到不同的ToR。这两个端口配置相同的IP和媒体接入控制（MAC）地址。如果一个ToR（或一个端口）宕机，另一个仍可继续工作。此外，由于同一网卡中的两个端口共享相同的队列对（QP）上下文，流量切换不会导致活动流的中断，并对上层应用透明。

然而，这样的设计引入了一个新的挑战：如何在没有直接连接的情况下同步两个不同的ToR的状态？应对这个挑战并不容易。在已有的堆叠双ToR方案中，由于两个ToR通过一个链接直接连接，它们可以通过直接链接协商一个共享的sysID。这使得主机可以通过链路聚合控制协议（LACP）与叠加式双ToR交换机进行通信。然而，因为我们想要消除ToR之间的直接链接，使它们相互独立，这意味着它们不能再使用LACP进行协商。因此，我们需要设计一种新的技术，通过一种隐式方法来“伪装”两个ToR，使主机可以通过LACP与双ToR进行通信。

如图3所示，构建非堆叠双ToR并不容易，因为我们必须确保在LACP协商过程中，双ToR交换机使用相同的MAC地址和不同的portID。我们与交换机供应商深度合作，实现了定制的LACP模块，以实现这一目标。

主机能够通过将每个ARP消息复制到NIC上的两个端口的方法来同时更新两个ToR上的ARP信息。到目前为止，所有主流的主机和交换机都能支持该非堆叠双ToR方案。



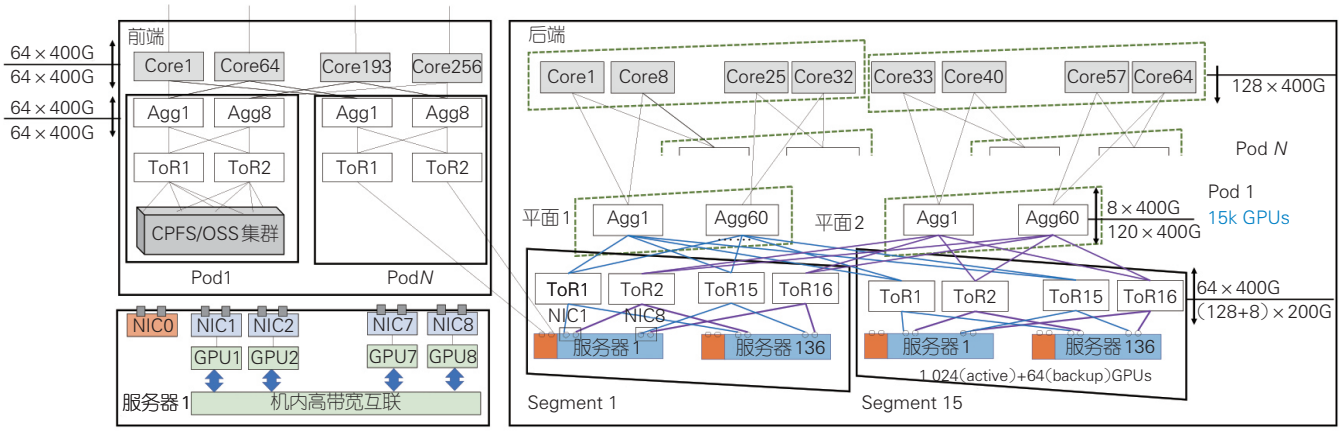
▲图3 非堆叠双上联

3 构建容纳千卡规模的一层网络

如图4所示，我们在高性能网络（HPN）中使用了最新的51.2 Tbit/s以太网单芯片交换机。在Tier1（一个Segment）中，每个交换机具有128个可用加8个备用的200 Gbit/s下行端口和60个上行的400 Gbit/s端口。这种设计确保了接近1:1的超额预订（实际上是1.067:1）。每个ToR交换机保留8个备用下行端口。我们使用这些端口连接备用主机，可以在主机端故障时快速更换主机。

1）单芯片交换机。ToR交换机的带宽容量直接决定了同一Tier1网络中GPU的数量。业界已经有支持更高带宽容量的多芯片框式交换机[6]。然而，阿里云在操作数据中心网络方面的长期经验表明，和单芯片交换机相比，多芯片框式交换机引入了更多的稳定性风险。具体来说，我们线上实际运营的单芯片交换机数量是多芯片交换机的32.6倍。相反，多芯片交换机遇到的关键硬件故障总数比单芯片交换机高3.77倍。根本原因在于多芯片交换机是一个分布式的交换系统。内部结构、芯片间相互作用、芯片与CPU的通信故障都会导致整体关键故障。因此，我们决定对所有新设计的网络架构都采用单芯片交换机。

2）多轨组网。主机内的8个GPU通过高带宽的主机内网络进行连接。虽然不同类型的GPU的主机内网络带宽不同，但是它比NIC提供的2×200 Gbit/s带宽高出4~9倍。NVIDIA是第一个提出多轨组网设计的[7]，此种网络设计已经广泛应用于训练集群中。在多轨组网中，同一铁路中的



▲图4 高性能网络整体概览

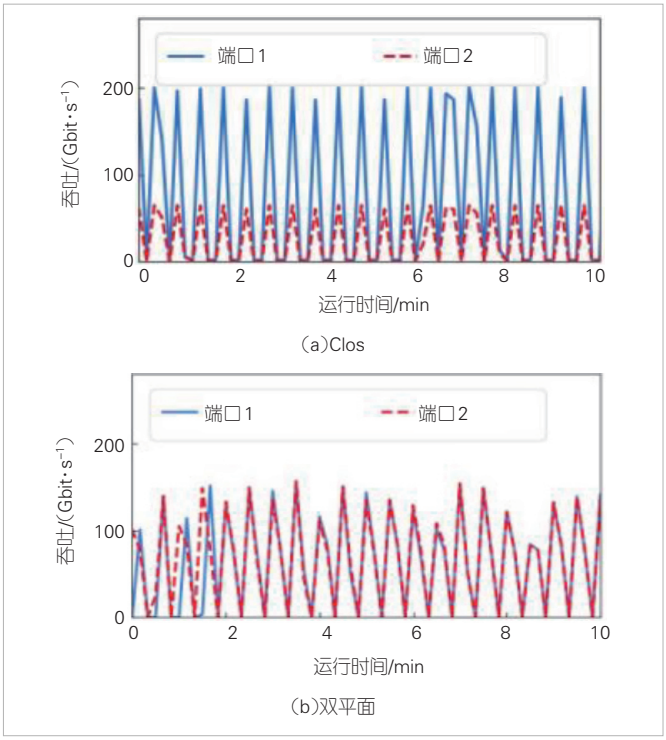
NIC通过同一套非堆叠交换机进行连接。不同轨道的NIC可以通过主机内+主机间转发的组合进行通信。例如，在图5中，如果服务器1中的GPU1想要与服务器3中的GPU2进行通信，转发路径为：服务器1的GPU1→服务器1的GPU2→ToR3→服务器3的GPU2。

4 构建容纳万卡规模的二层网络

在Tier1网络中使用双ToR，在ToR和聚合交换机之间简单部署典型的Clos拓扑结构，仍会存在哈希偏极化。在下行方向，双ToR设计导致存在2个可达下一跳，这引起了从60个聚合交换机到2个ToR交换机的高度收敛的流量。图6（a）展示了双ToR设置中两个下行端口的出口流量，流向同一网卡。我们对在生产环境中运行的GPT-3 175B的实际训练作业期间进行了测量。这两个端口的负载显著不同（吞吐量的差别高达3倍）会降低训练性能。

为了避免负载极端不均问题，我们需要在一个Pod中消除哈希偏极化。如图6（b）所示，在双平面设计中，每个双ToR设置中的ToR交换机被分为两个独立的组。有了这个设计，一旦一个流进入ToR中的任何一个上行链路，其在Pod内的转发路径就完全确定了。因此，在Pod中，哈希偏极化被完全消除了。部署双平面设计后，如图6（b）所示，不同端口的输入流量变得更加均匀，而在ToR下行端口的队列长度减少了91.8%。实际测试表明，双平面设计为跨段流量贡献了高达71.6%的性能优化。通过对512个GPU同时运行4个AllReduce作业的测试，这种优化的路径选择可以将集体通信性能提升34.7%。

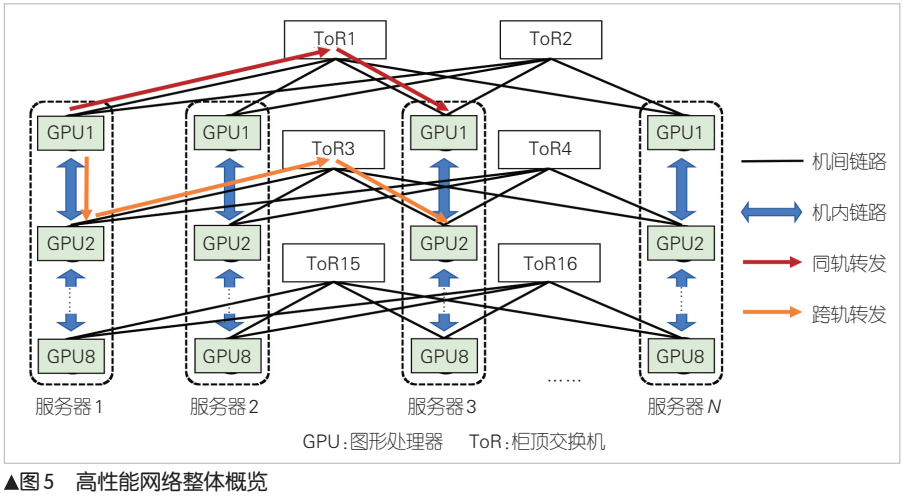
由于双平面设计，当搜索不相交路径时，我们只需搜索每个ToR交换机中的链路（即最多搜索60条链路），这样大大减少了时间消耗。HPN能够减少1或2个数量级的计算复杂性。



▲图6　同一网卡的两个端口流量

更重要的是，当发生故障时，主机只需要从ToR交换机那里获得新的等价多路径（ECMP）组，并重新计算不相交路径（而不是在全局控制器中维护来自不同层的ECMP组）。

双平面设计带来了另一个重要的好处：在ToR和聚合之间减少了一半的链路连接。这使得聚合交换机可以支持同一Pod中更多的Segment。因此，Tier2网络的规模翻了一番。另外，我们设置了聚合–核心的收敛比为15:1，并额外增加了聚合交换机上87.5%的端口，用于容纳更多的Segment。最终，我们实现了将15 000卡放置在同一Pod中，并为每个GPU提供了400 Gbit/s的网络接入能力。



▲图5　高性能网络整体概览

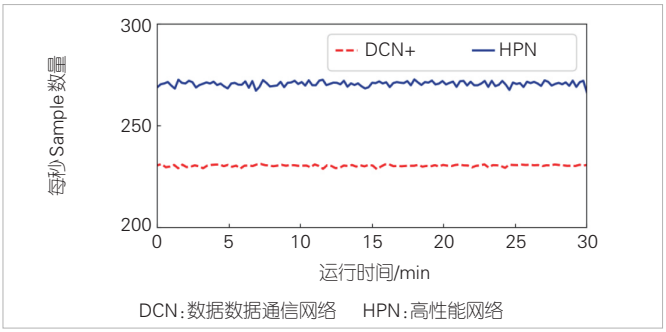
5 HPN性能评价

我们通过阿里云自主研发的大模型在集群上的训练效果来充分展示HPN所带来的性能提升。这个模型的训练采用了2 300多个GPU（超过288台服务器）。该大模型最初是在数据通信网络（DCN+）上进行训练，然后迁移到HPN上。在DCN+中，训练任务使用了19个Segment，而在HPN中，训练任务只需要3个Segment。我们观察到，迁移后性能会显著提升。图7显示，端到端训练性能提高了14.9%以上。这种端到端的

性能提升在实际生产环境中具有很大的价值。考虑到整个训练集群的构建可能会花费数十亿美元，14.9%的性能提升则可带来显著的成本节省。聚合交换机承载跨Segment流量，其统计数据直接反映网络状态。根据图8显示，跨Segment流量平均减少了37%。较少的跨Segment流量使得网络中的拥塞大幅下降。图9展示了聚合交换机下行链路队列长度分布。在DCN+中，大流量和哈希冲突不断积累队列长度；而在HPN中，该问题在很大程度上得到了解决。

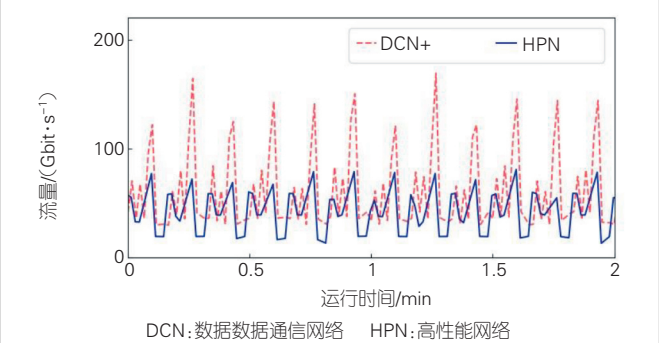
6 结束语

这篇文章介绍了HPN——一种用于大模型训练GPU集群的全新网络架构。该架构已在阿里云中大规模部署超过1年。HPN避免了传统数据中心拓扑中由单ToR设计引起的单点故障，通过双层网络连接了15 000个GPU，消除了哈希极化，并简化了最佳路径的选择。HPN使LLM训练的端到端性能提升超过14.9%。



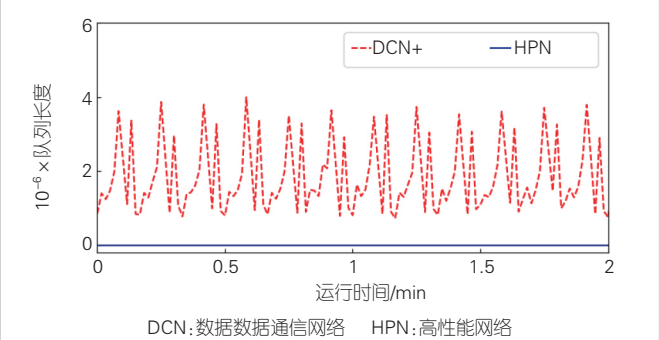
DCN：数据数据通信网络　HPN：高性能网络

▲图7　端到端训练性能



DCN：数据数据通信网络　HPN：高性能网络

▲图8　聚合层交换机入方向流量



DCN：数据数据通信网络　HPN：高性能网络

▲图9　聚合层交换机队列长度

参考文献

[1] Megatron-LM. Megatron-LM & megatron-core [EB/OL]. [2024-10-04]. https://github.com/NVIDIA/Megatron-LM

[2] DeepSpeed. DeepSpeed-extreme speed and scale for DL training and inference [EB/OL]. [2024-10-04]. https://www.microsoft.com/en-us/research/project/deepspeed/

[3] AL-FARES M, LOUKISSAS A, VAHDAT A. A scalable, commodity data center network architecture [C]//Proceedings of the ACM SIGCOMM 2008 Conference on Data Communication. ACM, 2008: 63-74. DOI: 10.1145/1402958.1402967

[4] ZHANG Z, ZHENG H, HU J, et al. Hashing linearity enables relative path control in data centers [C]// 2021 USENIX Annual Technical Conference (USENIX ATC 21). USENIX, 2021: 855-862

[5] POUTIEVSKI L, MASHAYEKHI O, ONG J, et al. Jupiter evolving: Transforming google's datacenter network via optical circuit switches and software-defined networking [C]//ACM SIGCOMM 2022 Conference (SIGCOMM '22). ACM, 2022: 66 - 85. DOI: 10.1145/ 3544216.3544265

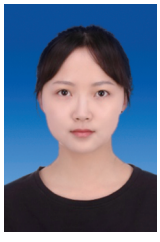
[6] Cisco. Cisco nexus 9800 series switches data sheet data sheet [EB/OL]. [2024-10-04]. https://www.cisco.com/c/en/us/products/collateral/switches/nexus-9000-series-switches/nexus9800-series-switches-ds.html

[7] FS. Troubleshoot the switchport packet loss [EB/OL]. [2024-10-06]. https://img-en.fs.com/file/user_manual/switch-port-packet-loss-troubleshooting.pdf

作　者　简　介

钱坤，阿里云高级技术专家；研究领域主要包括高性能智算网络、高性能存储网络和跨集群传输网络中的性能和稳定性优化，负责阿里云智算集群网络的监控和稳定性系统建设；发表论文10余篇。

翟恩南，阿里云资深技术专家、网络研究团队负责人，并担任SIGCOMM、NSDI、ACM SoCC等国际顶级会议程序委员会委员；研究领域包括计算机网络、分布式系统安全、程序验证等；发表论文30余篇。



操佳敏，阿里云技术专家；主要研究方向为高性能网络系统，包括面向大模型的网络性能优化、可编程芯片和可编程网络、软件定义网络等；曾参与多项国家自然科学基金、国家重点研发计划等项目；获得SIGCOMM2024最佳论文提名奖和ICCCN 2019最佳论文奖；发表论文20余篇，申请发明专利7项。