引用本文：杨泽崇，刘逢清. 基于DSCM的数据中心光互连架构设计[J]. 光通信技术，2026，50(1)：94-98.

基于DSCM的数据中心光互连架构设计

杨泽崇，刘逢清*

（南京邮电大学 电子与光学工程学院、柔性电子(未来技术)学院，南京 210023）

摘要：为突破传统电交换在数据中心机架内面临的带宽与高时延瓶颈，设计一种基于数字子载波复用(DSCM)的数据中心光互连架构。该架构融合DSCM技术与光无源器件，构建以超级节点为核心的转发系统，通过子载波动态分配与光域无队列直通传输机制，实现服务器间高效通信。仿真结果表明：在均匀流量满载(ρ=1.0)时，所提架构端到端时延较传统电交换架构降低近3个数量级，始终低于0.03 ms，并实现99.62%的带宽利用率，吞吐量接近800 Gb/s的理论上限；在流量集中场景下，其丢包率仍趋近于零，展现出优异的稳定性。

关键词：数字子载波复用；架顶交换机；点对多点传输；光互连

中图分类号：TN929.1　**文献标志码：**A　**文章编号：**1002-5561(2026)01-0094-05

DOI：10.13921/j.cnki.issn1002-5561.2026.01.016

Design of data center optical interconnection architecture based on DSCM

YANG Zechong, LIU Fengqing*

(College of Electronic and Optical Engineering & College of Flexible Electronics(Future Technology), Nanjing University of Posts and Telecommunications, Nanjing 210023, China)

Abstract: To overcome the bandwidth and high latency bottlenecks faced by traditional electrical switching within data center racks, an optical interconnection architecture for data centers based on digital subcarrier multiplexing (DSCM) is designed. This architecture integrates DSCM technology with optical passive devices to construct a forwarding system centered on a super node, enabling efficient communication between servers through dynamic subcarrier allocation and an optical-domain queue-free cut-through transmission mechanism. Simulation results show that under uniform traffic at full load (ρ=1.0), the end-to-end latency of the proposed architecture is reduced by nearly three orders of magnitude compared to traditional electrical switching architectures, consistently remaining below 0.03 ms, while achieving a bandwidth utilization of 99.62% and a throughput approaching the theoretical upper limit of 800 Gb/s. In traffic-concentrated scenarios, the packet loss rate remains close to zero, demonstrating excellent stability.

Key words: digital subcarrier multiplexing, top-of-rack switch, point-to-multipoint transmission, optical interconnection

0 引言

以DeepSeek、ChatGPT为代表的人工智能大模型的快速发展，对智算数据中心与超算数据中心的算力与互联性能提出了前所未有的高要求[1-2]。然而，支撑其运行的现有数据中心网络普遍采用传统电交换技术，面临交换容量受限、传输时延高、可扩展性差等关键瓶颈。光交换技术凭借其高带宽、低时延与低能耗的固有优势，被视为突破上述瓶颈、推动数据中心向高效化与智能化升级的下一代网络核心技术，近年来受到学术界与工业界的共同关注[3-6]。

针对数据中心的光互连与网络架构创新，业界提出了多种技术路线。例如：华为团队验证了基于零差相干检测的600 Gb/s双偏振64进制正交幅度调制(DP-64QAM)信号短距传输方案，为未来800G/1.6T互连提供了潜在路径[7]；微软构建了基于阵列波导光栅路由器的全光Sirius数据中心网络[8]；谷歌则实践了基

收稿日期：2025-03-19。

基金项目：国家自然科学基金项目(62004105)资助。

作者简介：杨泽崇(1998—)，男，山西长治人，硕士研究生，现就读于南京邮电大学电子与光学工程学院、柔性电子(未来技术)学院电子信息专业，主要研究方向为数据中心光互连网络。

***通信作者：**刘逢清(1975—)，男，博士，副教授，硕士生导师，主要研究方向为光纤通信及其接入技术，侧重于光网络的优化设计。

于微机电系统光开关的Apollo项目[9]。这些方案的研究重点多集中于数据中心的核心或汇聚层。实际上，随着数据中心网络流量模式从“南北向”为主转变为以“东西向”为主（约80%的流量发生在机架内部[5]），承担机架内互联任务的架顶（ToR）交换机面临巨大压力，其功耗可占网络交换机总功耗的90%[4]，已成为整体网络性能的主要瓶颈之一。为此，面向机架内场景的特定光互连方案被相继提出，如基于耦合器矩阵的节点/网络内的无源光交叉（POXN）架构[10]，基于集中式媒质访问控制的面向机架规模互连的无源光网络（POToRI）架构[11]，采用纳秒级光子开关的单级网络[12]，机架内与机架间协同设计的时间波长混合复用（TWDM）融合架构[13]，以及基于可重排Clos网络的多波长路由方案[14]等。这些方案虽各具特色，在一定程度上可以改善带宽、时延或能耗性能，但普遍采用固定的波长或时隙分配策略。面对机架内计算与存储服务器协同处理激增带来的需要子波长粒度动态连接重构的新需求，现有方案在灵活性、响应速度和成本效益方面仍显不足。

数字子载波复用（DSCM）技术通过单一光载波上的多子载波调制，可将物理链路灵活划分为多个动态可配的逻辑信道。该技术特别适合应对数据中心内“东西向”突发流量，能在不显著增加系统复杂度的前提下，有效解决传统光互连方案在灵活性与控制复杂度之间的矛盾，为实现子波长粒度的动态带宽调整与连接重构提供了一种高效且经济的技术途径。为此，本文设计一种基于DSCM的数据中心光互连架构，提升网络的灵活性与对动态业务的快速响应能力。

1 DSCM架构设计

本文提出的基于DSCM的机架内全光互连架构（简称DSCM架构）原理示意如图1所示。该架构在单个机架内部署了17台服务器，其中包括15台计算服务器和2台存储服务器。计算服务器采用高密度异构计算单元设计，每台配备双中央处理器（CPU）与6个图形处理器（GPU）。服务器之间通过CPU专用通道直连网络接口卡，并经由ToR交换机接入基于外部分层胖树（Fat-tree）拓扑的网络，以实现跨机架通信。为优化GPU间通信效率，本架构引入了基于光接口的星形拓扑：采用数字相干可插拔收发器、光分路器、光带通滤波器与光耦合器等器件构建全光传输网络，实现在光域内完成所有GPU之间的数据交换。

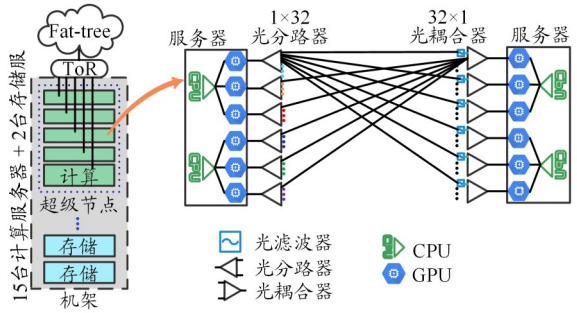

图1 DSCM架构原理示意图

本架构的核心设计理念与技术特征如下：其设计借鉴了Summit超级计算机节点与机架设计的核心理念[15]，通过创新性地融合DSCM技术与光无源器件，构建了一种能够支持大规模计算服务器与存储服务器间全光数据传输的新型互连架构，其技术特征可概括为以下3个方面：

1）硬件兼容性与平滑升级：点对多点（P2MP）模块在物理接口层面与传统点对点（P2P）模块保持插槽兼容性，同时支持以25 Gb/s为增量单位对光通道进行动态重构；

2）带宽资源智能调度：通过动态调整子载波组合，可在无需远端设备进行硬件改造的情况下，实现多级传输速率的灵活适配；

3）端到端可编程性：基于片上系统的光电协同控制机制，使系统能够自适应优化传输距离与功率预算，显著提升数据中心网络拓扑的灵活性。

该架构通过灵活调节子载波数量与收发器容量，可动态构建不同传输速率的可扩展网络。以800G可插拔收发器为例，其可划分为32个25 Gb/s的子载波通道，支持由5台服务器构成的超级节点（包含10个CPU与30个GPU）。超级节点内首台服务器的发射端GPU通过1×32光耦合器将数据广播至所有30个GPU，接收端则通过光带通滤波器筛选特定子载波频率，并经由32×1光耦合器实现信号合路，其它GPU采用相同机制进行通信。系统预留2个空闲子载波通道，既可用于连接机架内的2台存储服务器以提供高带宽存储服务，也可用于超级节点间的通信，从而构建P2MP甚至多点到多点的全光互连网络。

与传统数据中心依赖电交换实现机架内通信的模式（服务器通过双端口网络接口卡连接至ToR交换机）不同，本设计在物理层和网络层均进行了重要改进。在物理层，通过引入基于DSCM的智能可插拔收发器及光无源器件，在服务器部署与超级节点划分中以全光互连取代了传统电交换，从而突破了电交换的带宽瓶颈与高功耗限制，实现了频谱资源的高效复用

与低损耗传输。在网络层，本架构摒弃了基于架顶（ToR）交换机的集中式电交换模式，通过动态调节子载波数量并配合接收端光带通滤波器的灵活配置，实现了带宽的动态分配，构建出支持超级节点内与节点间点对多点通信的灵活可扩展全光互连架构。

2 仿真实验与结果分析

本文基于OMNeT++ 6.0平台构建了DSCM架构网络的系统级性能评估模型，仿真具体设置如下：

1）对比架构：DSCM架构中，32个GPU节点通过1×32光分路器与可调带通滤波器实现点对多点直接互连；基准架构（ToR交换机）中，32个服务器节点通过以太网链路连接至ToR交换机。

2）流量模型：采用依据泊松到达过程生成业务流的流量生成器。

3）路由策略：DSCM架构采用广播-选择路由策略；ToR架构部署开放式最短路径优先（OSPF）路由策略。

4）性能评价指标：主要考察平均端到端时延、系统吞吐量及丢包率。

关键仿真参数如表1所示。

表1 仿真参数

参数类别	参数	分布、设定值
流量生成	数据包到达间隔	服从负指数分布（泊松过程）[16]
	流数目/（流/s）	服从泊松分布、500[17]
	数据流的大小	服从均匀分布[16] 95%：1~100 kB； 5%：100 kB~10 MB
	流到达间隔/ms	服从负指数分布、2 [18]
网络拓扑	节点数/个	32
	目的地址分布	服从均匀分布（全随机）[18]
	路由策略	广播-选择/OSPF
队列管理	队列长度L_q	64、128、256、512、1 024
	队列策略	先入先出（FIFO）
性能配置	节点端口速率/（Gb/s）	25
	仿真时间T/s	10（忽略前2 s瞬态数据）
	负载强度范围	0.1~1.0

2.1 吞吐量-时延性能对比

为建立公平的性能基准，本文将所提DSCM架构与经队列优化后的ToR交换机架构（下文简称ToR架构）进行对比。通过前期仿真分析[19-20]发现，ToR架构的性能高度依赖于队列长度L_q配置：随着队列长度L_q增大，其吞吐量效率与丢包率性能提升，但平均时延也随之增长。综合权衡下，本文确定L_q=512为ToR架构的最优配置，其在负载强度ρ=1.0（满载）条件下可实现97.38%的吞吐量效率、1.29%的丢包率与79.67 ms的平均时延，代表ToR架构在机架内场景下的较优性能水平。本文基于此优化配置，客观评估DSCM架构的性能。

2.1.1 时延性能对比

2种架构的网络时延随负载强度ρ的变化趋势如图2所示。可以看出，DSCM架构的时延在全负载范围内始终保持在毫秒级，即使在ρ=1.0时也仅为0.03 ms。相比之下，ToR架构的时延随负载增加呈非线性指数增长：当ρ=0.1时，其值已达0.319 ms，远超DSCM架构满载时的时延；当ρ=1.0时，ToR架构时延飙升至79.674 ms，较DSCM架构高出近3个数量级。这一悬殊差距源于二者底层传输机制的差异：ToR架构依赖缓存队列处理突发流量，数据包排队导致时延累积；而DSCM架构基于子载波路由实现零缓存直通传输，从根本上消除了排队延迟。

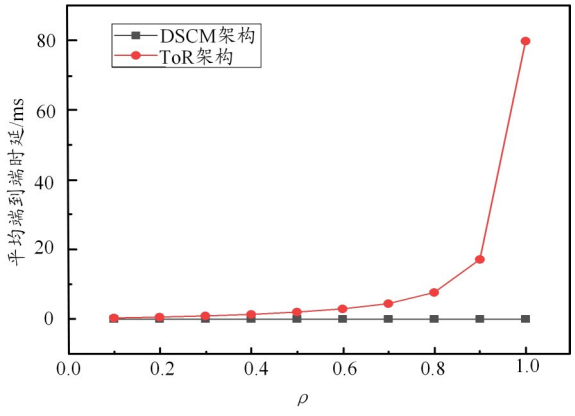

图2 2种架构网络时延对比

2.1.2 吞吐量性能对比

2种架构的吞吐量对比情况如图3所示。可以看出，随着ρ从0.1增至1.0，2种架构的吞吐量均相应提升，而DSCM架构呈现近似线性增长，在ρ=1.0时达到796.923 Gb/s，对应理论带宽利用率为99.62%，接近32端口×25 Gb/s的理论上限800 Gb/s。相比之下，ToR架构在同等负载下吞吐量仅为758.422 Gb/s，利用率为

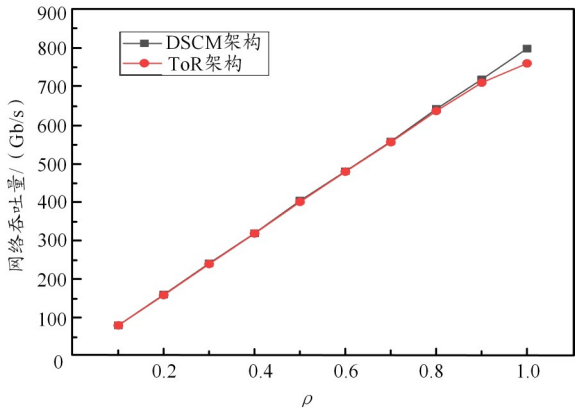

图3 2种架构网络吞吐量对比

94.80%，较DSCM架构低38.501 Gb/s。尤其在$\rho \geq 0.8$的高负载区间，DSCM架构的吞吐量增长斜率显著高于ToR架构，表明其能够更高效地利用带宽资源，避免电交换中因队列溢出导致的吞吐量瓶颈。

综合吞吐量与时延的对比结果可知，随着负载强度增加，DSCM架构的性能优势愈发显著：其端到端时延始终远低于ToR架构，吞吐量在低负载下略优，并在高负载下差距进一步扩大。尤其在ρ=1.0条件下，DSCM架构的吞吐量接近理论极限，时延较ToR架构降低近千倍，充分验证了光互连在高吞吐、低时延应用场景中的不可替代性。相比之下，ToR架构虽经队列优化可在一定程度上控制丢包，但其本质上仍受限于电交换“存储-转发”模式，难以突破时延与吞吐量的理论极值。与基于纳秒级光子开关的节点内高性能计算网络架构[12]相比，本文提出的架构在32节点机架上，实现了全负载范围内平均时延均低于30 μs的性能，且吞吐量呈现严格的线性增长特性，直至系统饱和；无需依赖昂贵的纳秒级光开关器件与频繁的交换重构，并能够在同一时刻支持P2MP或多点到多点通信，突破了传统P2P光互连的拓扑限制。

2.2 不同目的服务器数量对性能的影响规律

实验中的变量包括负载强度（ρ=0.6, 0.8, 1.0）与目的服务器数量（1、8、16、24、323 2台），，覆盖了从集中到均匀的多种流量模式。通过分析时延与吞吐量数据，揭示了流量分布与负载变化对网络性能的影响机制。

2.2.1 时延性能对比

2种架构在不同目的服务器数量S下的时延表现如图4所示。可以看出，DSCM与ToR架构在时延行为上存在显著差异。在相同负载下，2种架构的时延均随

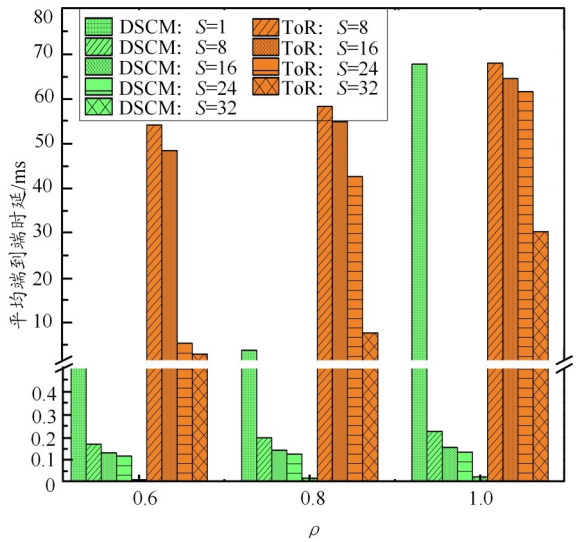

图4 2种架构在不同目的服务器数量下的时延对比

目的服务器数量的减少而上升，但上升幅度与机理截然不同。

在流量集中场景（目的服务器数量$S \leq 8$）中，DSCM架构在S=1时因子载波资源局部拥塞，时延从ρ=0.6时的1.476 ms激增至ρ=1.0时的67.311 ms；而当S增至8时，其时延仅微增0.056 ms，说明动态子载波分配能有效缓解拥塞。相比之下，ToR架构在S=8时时延显著增加25.5%，暴露出电交换队列在高负载集中流量下的严重拥塞问题。此外，ToR架构在S=1时因队列深度不足甚至无法完成通信，进一步凸显其对极端集中流量的容忍度缺陷。

在均匀流量场景（$S \geq 16$）中，DSCM架构时延始终低于0.03 ms，负载增加带来的时延增幅不足0.012 ms，体现了其无队列直通与并行处理机制的优势；而ToR架构时延则从2.93 ms大幅上升至30.14 ms，反映出其串行处理模式在高负载均匀流量下的效率瓶颈。这一显著差异的根本原因在于光互连避免了电交换中的排队时延，而ToR架构则受限于缓存拥塞与调度延迟之间的恶性循环。

2.2.2 吞吐量性能对比

2种架构在不同目的服务器数量下的吞吐量特性如图5所示。可以看出，2种架构在S从1~32的所有场景中，吞吐量随着ρ的增大均保持增长，但在S=8和16时，DSCM架构的吞吐量明显高于ToR架构；当ρ=1.0时，无论S取何值，DSCM架构的吞吐量均高于ToR架构。这验证了DSCM架构基于子载波并行的全局无损传输能力。

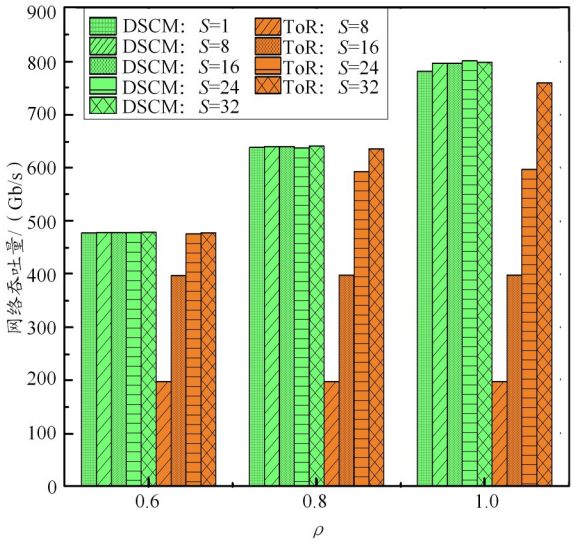

图5 2种架构在不同目的服务器数量下的网络吞吐量对比

3 结束语

为满足高性能数据中心机架内服务器间高带宽、低时延的通信需求，本文设计了一种基于DSCM的数据中心光互连架构。通过优化架构布局，并借助光载波上的多子载波调制与光域无队列直通传输机制，实现了网络整体性能的显著提升。仿真结果表明：在均匀流量满载（ρ=1.0）时，所提架构端到端时延较ToR构架降低近3个数量级，始终低于0.03 ms，并实现了99.62%的带宽利用率，吞吐量接近800 Gb/s的理论上限；在流量集中场景下，其丢包率仍趋近于零，展现出优异的稳定性，为下一代数据中心光互连网络的架构演进提供了新的技术思路。未来研究工作将聚焦于软件定义网络控制下的子载波动态分配机制，以及多业务共载场景下的多跳路由优化策略，进一步推动数据中心网络在能效比与可扩展性方面的突破。

参考文献：

[1] 马思聪，孙吉斌，孙一豪. 东数西算场景下的算力网关研发及应用[J]. 中兴通讯技术，2023，29（4）：2-7.

[2] 浪潮信息. 2023-2024年中国人工智能计算力发展评估报告[R]. 济南：浪潮信息，2023.

[3] Safarnejadian A, Mohammadi A, Rusch L A, et al. A power-efficient SDM structure for next-generation data center interconnect[J]. Lightwave Technology, IEEE/OSA Journal of(J-LT), 2024, 42(7): 8-8.

[4] Sato K I, Matsuura H, Konoike R, et al. Prospects and challenges of optical switching technologies for intra data center networks[J]. Journal of Optical Communications and Networking, 2022, 14(11): 903-915.

[5] Ben Y S J. Prospects and challenges of photonic switching in data centers and computing systems[J]. Journal of Lightwave Technology, 2022, 40(8): 2214-2243.

[6] Yang H, Xiang M, Cheng W, et al. 800G low-latency photonic data-center inter connections over 5 km hollow-core fiber[J]. IEEE Communications Magazine, 2025, 63(3):122-128.

[7] Gui T, Wang X, Tang M, et al. Real-time demonstration of homodyne coherent bidirectional transmission for next-generation data center interconnects[J]. Journal of Lightwave Technology, 2021, 39(4): 1231-1238.

[8] Ballani H, Costa P, Behrendt R, et al. Sirius: a flat datacenter network with nanosecond optical switching[C]//ACM. Proceedings of Annual Conference of the ACM Special Interest Group on Data Communication (SIGCOMM '20). New York: ACM, 2020: 782-797.

[9] Urata R, Liu H, Yasumura K, et al. Apollo: large-scale deployment of optical circuit switching for datacenter networking [C]//Optical Fiber Communication Conference and Exhibition, March 5-9, 2023, San Diego, USA. Washington: Optica Publishing Group, 2023: M2G -1-1-M2G -1-13.

[10] Ni W, Huang C, Liu Y L, et al. POXN: a new passive optical cross-connection network for low-cost power-efficient datacenters[J]. Journal of Lightwave Technology, 2014, 32(8): 1482-1500.

[11] Cheng Y, Fiorani M, Lin R, et al. POToRI: a passive optical top-of-rack interconnect architecture for data centers[J]. IEEE/OSA Journal of Optical Communications & Networking, 2017, 9(5): 401-411.

[12] Maniotis P, Dupuis N, Schares L, et al. Intra-node high-performance computing network architecture with nanosecond-scale photonic switches [J]. Journal of Optical Communications and Networking, 2020, 12(12): 367-377.

[13] Drainakis G, Baziana P, Bogris A. Optical intra- and inter-rack switching architecture for scalable, low-latency data center networks[C]//IEEE. Proceedings of 2023 IEEE Symposium on Computers and Communications(ISCC). Gammarth: IEEE, 2023: 1348-1351.

[14] 王金涛，刘逢清. 光互连数据中心网络架构与业务路由技术研究[J]. 光通信研究，2024(5)：96-101.

[15] Stunkel C B, Graham R L, Shainer G, et al. High - speed networks for the summit and sierra supercomputers[J]. IBM Journal of Research and Development, 2020, 64(3/4): 1-10.

[16] Benson T, Anand A, Akella A, et al. Understanding data center traffic characteristics[J]. ACM SIGCOMM Computer Communication Review, 2010, 40(1): 92-99.

[17] Benson T, Akella A, Maltz D A. Network traffic characteristics of data centers in the wild [C]//2010 ACM Internet Measurement Conference (IMC '10), October 25-27, 2010, San Jose, USA. New York: ACM, 2010: 267-280.

[18] Roy A, Zeng H, Bagga J, et al. Inside the social network's(datacenter) network[C]//Annual Conference of the ACM Special Interest Group on Data Communication (SIGCOMM '15), August 17-21, 2015, London, UK. New York: ACM, 2015: 123-137.

[19] Xiao_mage. 云游戏云电脑等云化应用场景[EB/OL].[2025-03-19]. https://www.cnblogs.com/XiaoMaGeYe/articles/17846285.html.

[20] 国际电信联盟. 沉浸式媒体体验指南[R]. 日内瓦：国际电信联盟，2023：15-18.