Token原生的Scale Across网络

Token Native Scale Across Network

苏远超/Su Yuanchao
（阿里云计算有限公司，中国 杭州 310030）
(Alibaba Cloud, Hangzhou 310030, China)

DOI:10.12142/ZTETJ.202603007
网络出版地址：https://link.cnki.net/urlid/34.1228.TN.20260626.1349.010
网络出版日期：2026-06-26
收稿日期：2026-04-15

摘要：传统广域网架构的设计初衷是承载普通流量，并未考虑人工智能（AI）工作负载的特殊性。基于第一性原理，提出一种新的Token原生Scale Across网络，支持各类AI负载在广域网中的传送。该网络采用模型与基础设施协同设计、每任务流量工程以及统一的互联网协议第6版（IPv6）寻址方案。相比传统网络，其容量扩展能力提升100倍，收敛时间缩短至原来的1/10，成本降低50%。

关键词：模型与基础设施协同设计；每任务流量工程；统一的IPv6寻址方案

Abstract: Traditional wide-area network (WAN) architectures are originally designed to carry regular traffic, and the specificities of artificial intelligence (AI) workloads are not taken into account. Based on first principles, a new token native Scale Across network is proposed to support the transmission of various AI workloads over the WAN. The network is built upon model-infrastructure co-design, per-task traffic engineering, and a unified Internet Protocol version 6 (IPv6) addressing scheme. Compared with traditional networks, its capacity scalability is improved by 100 times, convergence time is reduced to one-tenth, and cost is reduced by 50%.

Keywords: model-infrastructure co-design; per-job traffic engineering; unified IPv6 addressing scheme

引用格式：苏远超. Token原生的Scale Across网络 [J]. 中兴通讯技术, 2026, 32(3): 37-39. DOI: 10.12142/ZTETJ.202603007
Citation: Su Y C. Token native scale across network [J]. ZTE technology journal, 2026, 32(3): 37-39. DOI: 10.12142/ZTETJ.202603007

1 Scale Across全新架构的必要性

随着人工智能（AI）技术的飞速发展，特别是大语言模型和多模态模型的兴起，AI训练/推理对网络基础设施提出了前所未有的挑战。Agent技术的爆炸式发展，不仅催生了Token经济体系，也正推动计算底层架构向全新的范式演进。然而，传统广域网架构以承载普通流量为设计初衷，既未考虑AI工作负载的特殊性，如海量参数同步、键值缓存（KV Cache）交换，以及对超低延迟和超高带宽的严格要求，也未能针对Token的传输特征进行优化[1-3]。

从第一性原理出发，我们需要回答一个核心问题："如果Token是第一等的用户实体，那么怎样的AI基础设施才是最合适的？"若聚焦广域网范畴，问题则转化为：如何构建最适合Token传输的Scale Across网络。

本文提出Token原生的Scale Across架构设计，旨在构建面向未来3~5年的下一代广域网络。该架构并非对现有网络的局部修补，而是从底层设计到上层应用的系统性重构。

该架构的底层逻辑是面向生产Token、传输Token、消费Token的全过程，通过模型与基础设施的协同设计，构建一张具备Token粒度监控与调度能力的超大规模广域专用网络。

Token原生的Scale Across架构设定了3项关键技术指标[4]，并通过以下技术路径实现：

1）容量可扩展能力：提升100倍——通过路由芯片迭代与Fabric-router技术的协同演进，支持系统容量实现百倍级扩展。

2）故障收敛时间：缩短至原来的1/10——综合运用波道保护、链路检测（LD）与快速检测（RD）等技术，实现百毫秒级（100 ms）故障恢复。

3）单位成本：降低50%——通过全面自研芯片与智能流量调度技术的深度融合，单位带宽成本降低50%。

2 Token原生的Scale Across网络核心应用场景

Token原生的Scale Across架构主要服务于四大关键场景[5]，这些场景背后的业务需求与技术挑战，共同构成了架构设计的直接驱动力。

1）计算到计算互联。该场景涵盖预填充-解码（PD）分离架构、强化学习等应用，覆盖同城与跨城互联需求，是实现分布式算力协同的基础。

2）计算到存储互联。该场景主要面向KV Cache的存储访问优化，旨在实现同城高速互联。随着模型规模持续扩大，训练数据集的读取与模型检查点的保存均对存储访问带

宽提出了极高要求。

3）前端与后端跨可用区互联。该场景旨在打通不同可用区（AZ）之间的前端与后端网络，支撑业务的地理分布式部署，提升系统容灾与弹性调度能力。

4）跨可用区训练。该架构以前瞻性设计支撑未来大规模分布式AI训练，可确保未来3～5年的技术领先性。

上述4个场景从当前迫切需求到未来演进方向层层递进，形成了完整的应用场景图谱，同时也为后续技术选型提供了清晰的方向指引。

3 Token原生的Scale Across架构设计

Token原生的Scale Across在架构上继承了阿里巴巴eCore网络的多空间、多平面、多域设计及单栈单片技术体系，并在此基础上进行了针对性优化[6]：

1）采用专用物理网络，彻底隔离AI业务流量与其他业务流量，从根本上消除复杂的服务质量（QoS）策略配置与相互干扰。

2）全部采用大带宽端口互联（端口速率从400 Gbit/s起步）及大容量单机设备（交换容量从51.2 Tbit/s起步），最大程度降低由AI“大象流”引发的哈希冲突，从而有效减少拥塞丢包。

3）网络拓扑完全扁平化，减少转发跳数，显著提升数据平面转发效率。

4）路由架构完全扁平化，端到端统一采用互联网协议第6版（IPv6）地址空间，减少协议封装与地址转换，提升流量调度效率。

5）在运维层面，支持完善的监控体系，包括框架层监控、任务级诊断与异常检测、多主机关联分析、网络为中心的定位能力，以及覆盖计算（图形处理器（GPU）、内存）、通信（远程直接内存访问（RDMA）、NVLink、外围组件互连高速（PCIe））和网络交换的全栈监控。这一体系确保了Per-job的可观测性和可运维性。Token原生的Scale Across网络拓扑如图1所示。

3.1 模型与基础设施协同设计

模型与基础设施的协同设计，标志着网络架构正从“被动适配业务”向“主动协同业务”转变。上层模型的结构、并行策略、调度方式及内容传送机制等，均会对Scale Across网络的流量模型产生深远影响。

因此，当Scale Across网络出现性能瓶颈并需要调优时，若仅在网络层面进行优化，往往既耗时又收效甚微。阿里巴巴的工程实践表明，需要从两个维度开展协同设计：一是模型与基础设施的协同设计，二是端侧（网卡）与网络侧的协同设计。

一个典型场景是采用Scale Across网络承载预填充-解码（PD）分离架构下的远程直接内存访问（RDMA）流量。若不经任何修改，直接将原本部署在数据中心内的PD分离RDMA流量搬移至广域网传输，则很可能遭遇首字延迟（TTFT）显著恶化的问题，从而导致业务不可用。该问题的

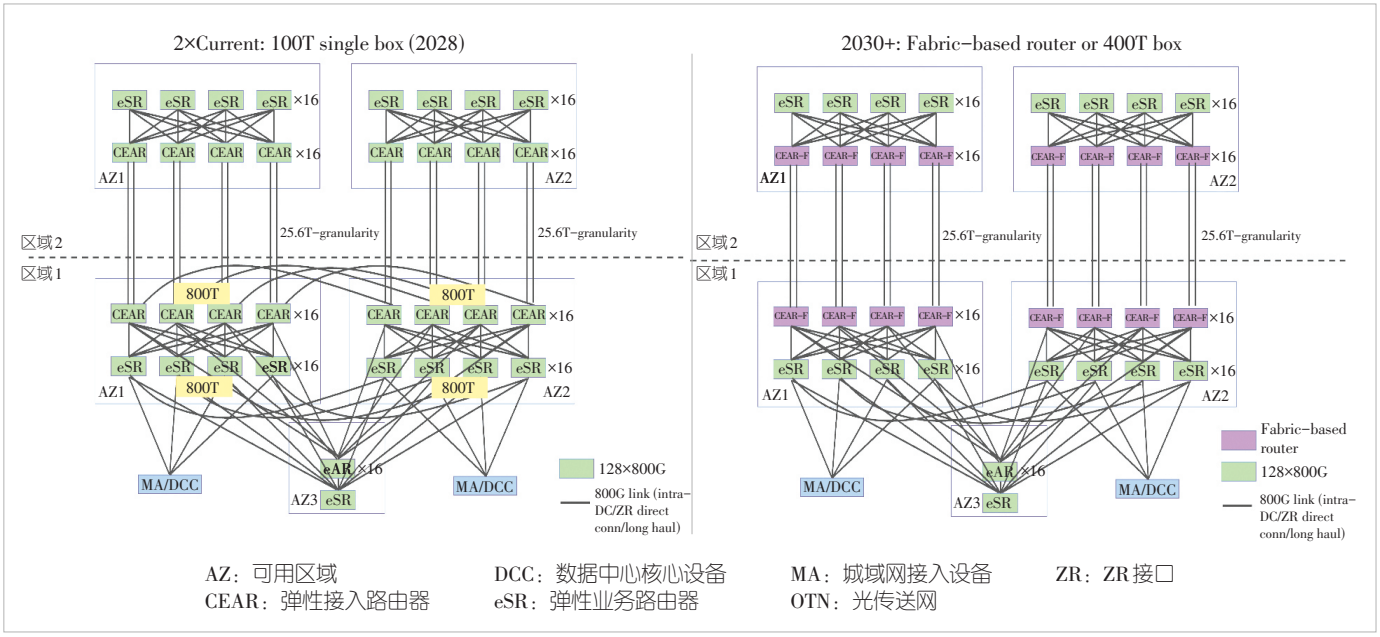

图1 Token原生的Scale Across网络拓扑

根源在于，GPU直接内存访问（GDR）要求发送端与接收端的显存逻辑地址连续，而PD分离场景中张量并行（TP）规模往往不一致，这会触发大量GDR小包。在数据中心内，此类小包仅带来微秒级时延，影响有限；但在广域网中，时延跃升至毫秒级，将导致网卡侧未完成工作队列元素（WQE）快速堆积，进而造成吞吐率下降，具体体现为TTFT显著增长。通过调整端侧网卡的WQE参数，并结合后文阐述的每任务流量工程机制，可在基础设施层面优化小包传送性能。然而，该方案仅能解决约50%的此类问题[7]。

在模型层面，需解决因TP规模不一致而产生大量GDR小包的问题。具体做法是在发送端与接收端均执行内存重排操作，使RDMA操作基于连续大块内存进行读写，以提升GDR传输效率。

模型与基础设施的协同设计，需以任务完成时间（JCT）、TTFT及每个输出Token的时间（TPOT）为共同优化目标，并在组织层面建立高效的协同机制。

3.2 每任务流量工程(Per-job TE)

上述端网协同设计是架构最具创新性的核心体现。传统网络通常采用基于五元组的流量识别方法，但对于远程RDMA流量而言，这种识别远不够精细。本架构通过解析RDMA的队列对（QP）字段，并结合目的IP地址，能够精确识别每一条RDMA流。新一代网卡支持单QP多路径能力，使每任务流量工程（Per-job TE）与网卡多路径机制可正交协同工作，实现更为灵活的流量调度。

基于上述设计，即可建立从Token到Job，再经框架层、通信库、QP，直至Scale Across网络的完整端到端映射链路。这一映射机制构成了Token Native架构最为核心的技术基石。

3.3 统一的IPv6寻址方案

为支持上述精细化调度，架构采用统一的IPv6地址规划，覆盖从容器、网卡到高性能网络及Scale-Across的完整数据路径。该设计的核心优势在于免去IPv6-in-IPv6封装，从而减少报文处理开销。同时，基于重新定义的SRv6地址格式，网络可根据业务需求灵活选择默认路径、低时延路径或低成本路径，以适配不同业务的差异化需求。

3.4 关键技术特性

Per-job流量工程：基于（目的QP号、目的IP）二元组精确匹配RDMA流，实现更为精细的流量调度粒度。

链路检测（LD）/接收检测（RD）：依据G.709 Annex C标准，在网络降级条件下实现无损且快速的ECMP故障切换，确保业务连续性。其中，LD与RD作为光传输系统中的关键监测机制，为快速感知链路状态变化提供支撑。

快速拥塞通知包（CNP）：加速拥塞反馈信号的传递，使发送端能够更快感知网络状态变化，并及时调整发送速率，从而有效缓解拥塞扩散。

Packet Trimming：当报文因拥塞被丢弃时，仅保留L3与L4头部并发送至源端，支持单包重传。该设计避免了传统模式下需由CPU介入处理完整报文的问题，提升了重传效率。

4 结束语

Token原生的Scale Across架构标志着Token时代广域网络面向AI场景的一次重大演进。这一设计既是对第一性原理的遵循，也是Token内在传输需求的最自然体现。该架构将为未来大规模AI训练与推理提供坚实的网络底座，有力支撑业务的持续创新与增长。

参考文献

[1] Filsfils C, Camarillo P, Leddy J, et al. Segment routing over IPv6 (SRv6) network programming RFC 8986 [EB/OL]. (2026-05-20) [2026-05-25]. https://datatracker.ietf.org/doc/rfc8986/
[2] Filsfils C, Dukes D, Previdi S, et al. RFC 8754: IPv6 segment routing header (SRH) [EB/OL]. (2020-03) [2026-04-25]. https://www.rfceditor.org/info/rfc8754/.https://doi.org/10.17487/RFC8754
[3] Filsfils C, Ketan T, Daniel V, et al. Segment routing policy architecture RFC 9256 [EB/OL]. [(2022-08) [2026-04-20]. https://datatracker.ietf.org/doc/rfc9256/. DOI: 10.17487/RFC9256
[4] Singh R, Agarwal S, Calder M, et al. Cost-effective cloud edge traffic engineering with cascara [EB/OL]. [2026-04-15]. https://www.usenix.org/conference/nsdi21/presentation/singh
[5] Xu Z Y, Yan F Y, Singh R, et al. Teal: learning-accelerated optimization of WAN traffic engineering [C]//Proceedings of the ACM SIGCOMM 2023 Conference. ACM, 2023: 378-393. DOI: 10.1145/3603269.3604857
[6] Singh A, Ong J, Agarwal A, et al. Jupiter rising: a decade of clos topologies and centralized control in Google's datacenter network [EB/OL]. (2015-08-17) [2026-04-20]. https://dl. acm. org/doi/10.1145/2829988.2787508
[7] Zhong Z Z, Ghobadi M, Khaddaj A, et al. ARROW: restoration-aware traffic engineering [EB/OL]. (2021-08-09) [2026-04-18]. https://dl.acm.org/doi/abs/10.1145/3452296.3472921

作 者 简 介

苏远超，阿里云计算有限公司广域网架构与研发总监；负责阿里巴巴全球网络规划、建设和开发，主要研究领域为IP和光网络架构、协议及流量工程系统；发表多篇论文，拥有多项专利。