专题：智算网络

智算中心网络的亚毫秒级网络监控系统

张冬月[1]，辛奇[2]，韩博文[1]，徐博华[1]，曹畅[1]，顾茹雅[3]
（1. 中国联合网络通信有限公司研究院，北京 100048；
2. 北京邮电大学，北京 100876；
3. 北京基流科技有限公司，北京 100083）

摘　要：随着智算中心成为支撑数字经济高质量发展的核心基础设施，千亿参数大模型训练对网络性能提出严苛要求，传统监控手段因采样精度不足、缺乏细粒度观测难以应对万卡集群的通信瓶颈。提出了亚毫秒级网络监控系统（sub-millisecond-level network monitoring system，sMon），通过在工作请求处理流水线中集成智能计数器与动态带宽分析模块，实时追踪网卡端口队列深度及流量波动。基于滑动窗口算法实现带宽吞吐的动态计算，在保持亚毫秒级时序精度的同时，采用异步日志采集机制将系统开销中央处理器（central processing unit，CPU）占用率降低至0.8%。在128节点A100集群的测试中，系统成功捕获了网络端口的亚毫秒级流量细节。实验证明该系统较传统监控方案在数据粒度上提升2个数量级。该方案通过高精度网络状态感知能力，为构建“超大规模、超高带宽、超强可靠”的智算中心网络提供了实时监控与性能保障，有效支撑大规模人工智能（artificial intelligence，AI）训练任务的需求。

关键词：智算中心网络；网络监控；亚毫秒级
中图分类号：TP393
文献标志码：A
doi: 10.11959/j.issn.1000−0801.2025172

Sub-millisecond-level network monitoring system for intelligent computing centers network

ZHANG Dongyue[1], XIN Qi[2], HAN Bowen[1], XU Bohua[1], CAO Chang[1], GU Ruya[3]
1. China Unicom Research Institute, Beijing 100048, China
2. Beijing University of Posts and Telecommunications, Beijing 100876, China
3. Infrawaves, Co., Beijing 100083, China

Abstract: As intelligent computing centers become core infrastructure supporting the high-quality development of the digital economy, the training of hundred-billion-parameter large models imposes stringent requirements on network performance. Traditional monitoring methods struggle to address communication bottlenecks in ten-thousand-card clusters due to insufficient sampling precision and lack of fine-grained observation. A sub-millisecond-level network

收稿日期：2025−03−25；**修回日期：**2025−04−12
通信作者：徐博华，xubh15@chinaunicom.cn
基金项目：国家重点研发计划项目（No.2024YFB2908704）
Foundation Item: The National Key Research and Development Program of China (No.2024YFB2908704)

monitoring system (sMon) was proposed, which integrated intelligent counters and dynamic bandwidth analysis modules into the workflow processing pipeline to enable real-time tracking of NIC port queue depth and traffic fluctuations. By implementing dynamic bandwidth calculation through sliding window algorithms, sub-millisecond temporal accuracy was maintained while reducing system overhead to 0.8% CPU utilization via asynchronous log collection mechanisms. Testing on a 128-node A100 cluster, the system's capability was demonstrated to capture sub-millisecond traffic details at network ports. Experimental results show a two-order-of-magnitude improvement in data granularity compared with conventional monitoring solutions. The proposed system provides real-time monitoring and performance assurance for constructing "ultra-large-scale, ultra-high-bandwidth, ultra-reliable" intelligent computing center networks through high-precision network state perception, effectively supporting the requirements of large-scale AI training tasks.

Key words: intelligent computing centers network, network monitoring, sub-millisecond-level

0 引言

近年来，生成式人工智能（artificial intelligence，AI）与大语言模型的爆发式增长（如ChatGPT、Sora等）推动全球算力需求呈指数级攀升。以千亿乃至万亿参数规模的模型训练为代表，人工智能技术对算力基础设施提出了“超低时延、超高带宽、零丢包”的严苛要求。在此背景下，智算中心作为支撑人工智能发展的核心基础设施，在中国进入规模化建设阶段。据《中国综合算力指数报告（2024）》统计[1]，截至2024年6月，全国已建及在建智算中心超250个，单集群规模从千卡迈向万卡甚至十万卡级别。

然而，智算中心网络的复杂性与性能瓶颈日益凸显。一方面，传统网络监控手段（如SNMP轮询、分钟级采样）受限于精度不足与业务关联性缺失，难以捕捉微秒级时延抖动、瞬时流量微突发（microburst）等关键指标[2]；另一方面，万卡集群中高并发通信对网络稳定性要求极高，传统监控系统缺乏细粒度观测能力[3]，导致故障定位效率低下、资源利用率不足等问题。部分智算中心网络拥塞导致图形处理单元（graphics processing unit，GPU）闲置率超过30%，严重制约训练效率。此外，现有监控技术在高带宽（400~800 Gbit/s）、高密度（单机柜20~30 kW）场景下的扩展性与实时性亦面临挑战[4-5]。

针对上述问题，本文聚焦智算中心网络的亚毫秒级监控需求，提出一种新型实时监测系统。该系统突破传统监控的时空粒度限制，通过动态感知网络状态、精准定位异常事件，为构建“超可靠、超弹性”的智算中心网络提供底层技术支撑。相较于既有方案，本文系统在确保低开销（CPU占用率<1%）的前提下，首次实现全栈网络指标的亚毫秒级可视化，为智算中心自动化运维与资源优化开辟新路径。

1 研究背景和动机

1.1 智算中心网络的发展

随着人工智能、大数据和云计算的迅速发展，智算中心网络经历了从传统以太网架构向高性能、低时延网络的演变[6]。早期的数据中心主要依赖树状或胖树（Fat-Tree）结构的以太网，随着计算需求激增，这种架构逐渐暴露出带宽不足和时延较高的问题。为满足图形处理单元（graphics processing unit，GPU）/张量处理单元（tensor processing unit，TPU）集群间高速数据传输的需求，网络技术逐步引入了RDMA、InfiniBand等先进技术，使得数据传输效率得到了显著提升[7]。

当前，智算中心网络不仅需要支持超大规模集群（如千卡、万卡甚至十万卡级别）的计算任

务，还面临高并发通信带来的网络拥塞、负载均衡以及异常监控等挑战。传统监控手段如SNMP轮询和分钟级采样已经无法满足实时捕捉微秒级流量变化和故障定位的需求，这促使网络监控系统向亚毫秒级别演进，以实现对网络状态的精准监测和快速响应[8]。

未来，智算中心网络的发展将更多依赖于智能调度、动态带宽管理和超低时延监控系统，这些技术不仅能进一步提升网络可靠性和扩展性，还能为大规模模型训练等高负载应用提供坚实的支撑，为数字经济和人工智能等前沿领域的发展注入持续动力。

1.2　传统网络监控系统

传统网络监控系统最初主要依靠NetFlow和sFlow等技术，网络监控系统概览见表1，这类系统通常以秒到分钟级别的时间粒度进行监测，适用于常规流量统计和网络健康检测。随着网络设备性能的不断提升和技术进步，近来的研究工作如Trumpet和Sketch已经将监测时间粒度缩小到毫秒级，使得对网络流量的动态变化能够进行更细粒度的捕捉[9]。总体来看，传统网络监控系统根据任务主要分为两大类：一类是流量管理系统，另一类是网络事件检测系统。现有毫秒级网络监控系统如图1所示。

流量管理系统通常采用接口级别的流量统计方法，对网络各接口的总体流量进行监控。这种方法在大规模部署中具有较低的存储和计算成本，但由于其统计粒度较粗，往往难以捕捉到亚

表1　网络监控系统概览

工具	主要思路	时间粒度	优点	缺点
NetFlow	基于流（flow）进行采样，记录五元组与流量统计	秒级或分钟级	采样率可调，能在不增加过多开销的情况下获得基本流量信息	难以捕捉突发流量，无法精确定位短时异常
sFlow	采用报文采样（packet sampling），周期性采集	秒级或分钟级	硬件支持广泛，对高速网络适用，开销相对可控	报文抽样精度受限，无法实现亚毫秒级监控
SNMP	通过轮询网络设备（如路由器、交换机）的MIB信息	秒级或分钟级	部署简单，设备支持度高，适合基础监控	数据粒度较粗，仅提供接口级计数，无法获取精细化流量或报文级信息
Trumpet	触发式（trigger-based）监控，利用硬件事件检测网络异常	毫秒级	监测精度高，可快速捕捉网络事件	对设备的定制化需求高，部署门槛较高
FlowRadar	将流信息编码至Sketch结构，实现对流量的解码与检测	毫秒级	在较低的存储与计算开销下可识别大量流信息	编码和解码过程复杂，对网络设备的处理能力有一定要求
Sketches	采用哈希与计数结构（如count-min sketch）进行近似计量	毫秒级	存储与计算开销相对可控，可在高速网络中实现较高监测精度	存在一定误差率，需权衡精度与资源开销；实现方式对网络设备可编程性有要求
MiliSampler	利用高频率的采样策略（如基于时间片的快速轮询）实现细粒度流量统计	毫秒级	能够捕捉比秒级更细的流量波动，采样方法灵活	采样频率的提升会带来额外开销，难以进一步扩展至亚毫秒级
Lumina	基于可编程交换机或智能网卡的实时遥测（telemetry）技术	毫秒级	能实现更高精度和实时性，对网络拓扑和流量变化进行动态监测	需要对交换机或网卡进行定制开发，商业交换机兼容性相对不足
Valinor	在分布式环境中采用分段式数据收集与分析的方式，对端到端流量进行精细监控	毫秒级	有较好的可扩展性，能够在大规模集群中对流量进行细粒度观测	实现复杂度较高，需对数据中心网络进行较大程度的改造
ConQuest	专注于拥塞检测和队列深度分析，通过可编程数据面（data plane）采集信息	毫秒级	能精准定位拥塞热点与突发流量，帮助进行实时的流量调度	对设备可编程能力依赖较高，通用性和易用性在传统商业交换机环境中仍有局限
BurstRadar	专门用于探测微突发（micro-burst）的系统，通过快速捕捉队列长度和流量变化	毫秒级	能够有效发现短时高峰（micro-burst），在精细化拥塞管理和性能优化中具有重要价值	需要额外的高速计数与存储资源，对网络设备的性能提出较高要求

毫秒级的突发流量行为，可能忽略一些关键的网络异常情况。而网络事件检测系统则依赖于可编程交换机（如使用P4语言编程的交换机）来实现高精度的监测，能够实时捕捉细微的网络事件和瞬时流量变化，但这类系统在实际应用中往往与现有的商业交换机兼容性较差，制约了其在大规模环境下的推广[10-11]。

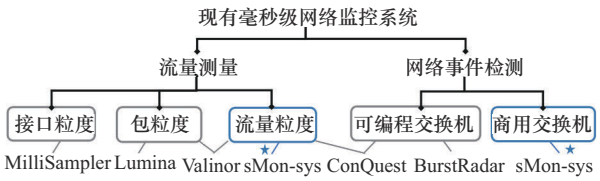


图1　现有毫秒级网络监控系统

目前，虽然现有监控技术在流量管理或事件检测上各有侧重，但仍缺乏一种能够同时支持大规模流量测量与亚毫秒级拥塞分析的通用网络监控系统。

1.3　亚毫秒级网络监控系统的重要性

在传统网络监控系统的局限性已明确体现出对更高精度和实时性需求的背景下，亚毫秒级网络监控系统（sub-millisecond-level network monitoring system，sMon）的重要性主要体现在以下几方面。

首先，亚毫秒级监控能够捕捉到传统秒级或分钟级监控手段无法检测的瞬时流量波动和短时突发现象。这种细粒度的数据采集对于发现网络中微小异常或局部拥塞极为关键，有助于实现更早的故障预警和问题定位[12]。其次，借助亚毫秒级监控数据，网络调度与流量管理系统可以实现实时动态调控，快速响应网络拥塞状况，及时调整资源分配，从而降低时延检测引发的连锁故障风险，提高整体网络的运行效率和稳定性。此外，亚毫秒级监控提供的高精度数据为后续的自动化故障处理和智能化网络管理奠定了坚实基础。基于细粒度数据，运维系统能够更准确地分析网络异常模式，支持基于数据驱动的自适应调度和系统自愈，进一步提升网络的可靠性和安全性[13]。

总的来说，亚毫秒级网络监控系统通过提升监测精度和实时性，为网络流量测量、异常检测及自动调控提供了有效手段，有效弥补了传统监控系统在高负载和高并发场景下的不足，是实现高效、智能网络管理的必然发展方向。

2　系统设计

2.1　概览

本文系统基于远程直接内存访问技术构建，针对智算中心网络中微秒级突发流量与瞬时拥塞事件的检测需求，设计了一套跨节点的精细化监控架构。亚毫秒级网络监控系统概览如图2所示。系统以模块化方式划分为数据发送端（模块A）与数据接收端（模块B），通过队列化流水线、计数器协同及无锁内存访问机制，实现从数据采集、跨节点传输到状态反馈的全链路闭环管理。

2.1.1　核心组件与功能定义

模块A（数据发送端）作为监控数据的采集与调度中枢，由先进先出队列、工作请求封装单元、发送队列及计数器构成。网络流量通过智能网卡捕获后，首先进入先进先出队列进行时序化缓冲，确保突发流量事件的原始时序特征不被破坏。工作请求封装单元将原始流量转换为结构化指令，包含目标内存地址、数据块长度及全局唯一的标识索引。目标内存地址指向模块B预分配的接收缓冲区，数据块长度定义单次写入操作的负载规模，而标识索引则用于跨节点数据包溯源与状态跟踪。封装后的工作请求进入发送队列，通过远程直接内存访问技术的写入操作，绕过操作系统内核直接推送至模块B的内存空间。

模块B（数据接收端）的核心功能在于实时处理监控数据并反馈系统状态，其核心组件包括

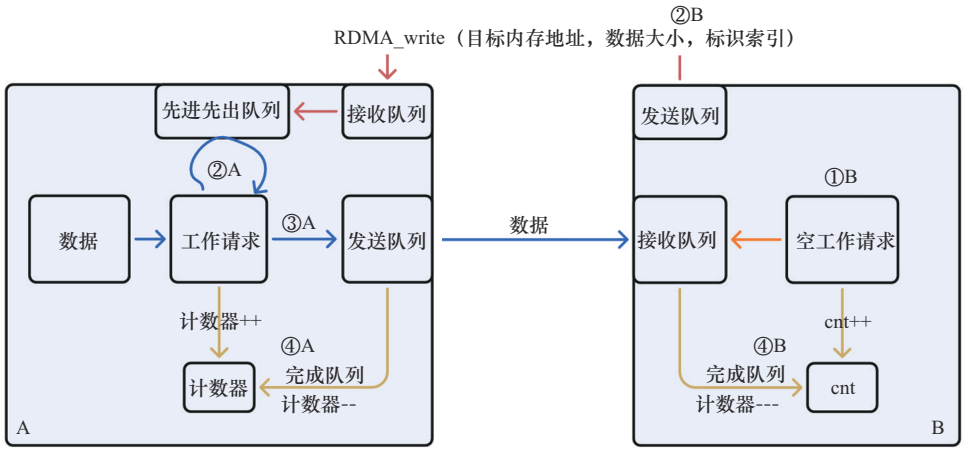


图2　亚毫秒级网络监控系统概览

接收队列、空工作请求预分配池、计数器及完成队列。接收队列采用动态内存映射机制，根据标识索引将写入的数据包分发至对应的处理线程。为了避免接收端缓冲区耗尽导致的丢包问题，模块B在初始化阶段即预生成空工作请求池，这些空请求不携带实际数据，仅作为内存占位符等待发送端的写入操作填充。当模块A的写入指令抵达时，标识索引与预分配的空请求自动匹配，实现缓冲区资源的无缝衔接。计数器作为跨节点状态同步的核心，在发送端与接收端分别记录活跃任务数：模块A每推送一个工作请求，计数器递增；模块B每接收并绑定一个数据包，计数器再次递增；任务处理完成后，两端计数器同步递减。这种双向递增机制确保系统始终掌握全网未完成操作的总量，为拥塞预警提供量化依据。

2.1.2　数据流转与状态同步机制

系统的运行流程可分解为数据注入、跨节点传输、资源绑定及完成反馈4个阶段，各阶段通过严格的时序逻辑与状态联动实现微秒级精度控制。

在数据注入阶段，智能网卡以硬件加速方式捕获网络端口流量，生成包含时间戳、队列深度、流控标识等元数据的监控信息。原始数据进入先进先出队列后，封装单元根据预设规则将其转换为结构化工作请求。以检测交换芯片瞬时拥塞为例，当某个端口在50 μs内队列深度从10%飙升至95%时，系统会生成高优先级请求，目标内存地址指向模块B中预留给拥塞分析线程的专用区域，数据块长度设置为足以容纳完整的队列状态快照，标识索引则关联至该端口的物理拓扑编号。封装后的请求按优先级排序进入发送队列，触发远程直接内存访问写入操作。此时，发送端计数器递增，记录当前处于传输中的任务数量。

跨节点传输阶段依托远程直接内存访问技术的无CPU旁路特性，直接将数据从模块A的发送队列写入模块B的接收队列。此过程完全绕过操作系统内核与协议栈，端到端传输时延可稳定控制在100 μs以内。写入操作的核心参数——目标内存地址、数据块长度及标识索引——共同决定了数据在接收端的存储位置与处理逻辑。

资源绑定阶段的核心在于空工作请求的动态分配机制。模块B在初始化时预先生成大量空请求并注册至全局资源池，每个空请求包含预留的内存地址与空闲标识索引。当模块A的写入操作携带的标识索引与某个空请求匹配时，该请求被激活并绑定至实际数据包，同时释放其占用的内存资源供后续使用。这种机制有效避免了传统通信中反复协商缓冲区的开销，尤其适应万卡集群场景下的高并发需求。例如，在1 024条并发写

入请求场景下，模块B通过弹性扩缩容策略动态调整空请求池的规模：当空请求数量低于阈值时，批量预分配新请求；当资源闲置时，逐步释放冗余内存以降低开销。

完成反馈阶段通过完成队列实现闭环状态同步。当模块B的处理线程解析完数据包并执行相应操作后，会向完成队列提交处理结果。完成队列携带原始标识索引，向模块A发送异步回调信号，触发两端计数器的递减操作。在此过程中，系统通过时间窗口聚合算法，将离散的微秒级事件关联为完整的拥塞传播路径。若检测到连续3个时间窗（每个窗口为100 μs）内计数器值异常波动，则自动触发流控协议参数调整，例如，动态提高RoCEv2协议的CNP报文发送频率，以加速拥塞缓解。

2.2 流量级监控

流量级监控的核心目标是精准捕捉智算中心网络中的流量动态变化，及时识别突发流量和潜在拥塞风险，从而优化整体网络性能。本文系统依托智能网卡的硬件加速能力，结合远程直接内存访问技术，实现了高精度、低开销的流量监测方案。智能网卡可实时捕获网络端口的流量数据，包括流量速率、队列深度、流控信号、协议类型等，并通过硬件时间戳确保数据的精准同步。这种方式不仅降低了传统软件轮询的开销，同时避免了系统中断和调度带来的数据时延，提高了监控时效性。

在流量级监控过程中，系统对采集到的数据进行分类，以便于更有针对性地进行分析和处理。首先，系统区分端到端业务流，针对计算流量（如AI训练、存储访问）和控制流量（如流控信号、CNP报文）采用不同的监测策略[14]。其次，为了精准识别突发流量，系统通过短时间窗口内流量速率变化来判定异常情况。例如，当某个端口在100 μs内负载从20%突增至90%时，系统会标记该端口可能存在突发流量或潜在拥塞风险。此外，基于流控信号，系统能够进一步识别当前网络中的流控热点，并结合历史数据分析其发展趋势，以提前预警可能的拥塞演化过程。

为了高效存储和处理这些监测数据，系统采用远程直接数据存取（remote direct memory access，RDMA）直写方式将流量信息传输至监控服务器，绕过操作系统内核，提高数据传输效率。所有监测数据被存储在环形缓冲区中，近期的流量状态可以通过时间窗口聚合算法进行归纳。

基于监测结果，系统能够提供针对性的优化策略，以提升网络的整体稳定性和吞吐能力。当网络流量处于正常状态时，系统维持低采样率，以减少监测开销；而在检测到异常流量时，会动态提高采样精度，以获取更详细的数据。此外，针对持续触发显式拥塞通知（explicit congestion notification，ECN）信号的端口，系统可以主动调整流控参数，例如，增加拥塞通知包（congestion notification packet，CNP）报文发送频率，加速拥塞的缓解过程。同时，系统还具备实时告警能力，一旦检测到严重的突发流量或异常的队列积压情况，会立即触发预警，提醒管理系统采取相应的流量调度策略。本文系统通过流量级监控，实现了对微秒级流量动态的精细化分析，为后续优化网络性能提供了基础数据支撑。

2.3 动态带宽计算

动态带宽计算是本文系统流量监控的重要组成部分，旨在精准衡量网络中的数据传输速率，并及时反映带宽占用情况。在智算中心的高性能计算环境中，流量状态会随计算任务的变化而剧烈波动，短时间内的带宽突增或骤降会影响整体通信效率。因此，单次数据传输的统计方式难以准确描述带宽变化趋势，而基于滑动窗口的动态带宽计算方法能够提供更具时效性和准确性的测量结果。

系统在网络传输过程中，每当某个数据请求

的传输完成，即事件状态被清零时，便记录当前时间以及本次传输的数据量，并将其存入日志队列。然而，单次数据记录的时间间隔仅能表明该数据请求在该时间范围内完成传输，并不能准确反映整个时间段的平均带宽。因此，本文系统引入滑动窗口机制，通过综合多个数据记录的传输量与时间间隔，计算更平滑的带宽数值。滑动窗口大小通常设定为W，即计算带宽时始终基于最新的W条传输记录，使得带宽测量不仅能捕捉短时突发流量，还能平滑数据波动，避免异常情况对整体测量的干扰。

当新的数据记录到来时，系统会将滑动窗口向前推进，移除最早数据记录，并加入最新数据，保持窗口大小不变。设窗口内第 i 条数据的传输完成时间为T_i，对应的传输数据量为D_i，则滑动窗口内的带宽计算式如下：

$$B=\frac{\sum_{i=1}^{W}D_i}{T_W-T_1} \tag{1}$$

其中，$\sum_{i=1}^{W}D_i$为窗口内所有数据传输量之和，T_W-T_1为窗口内最早和最晚数据记录之间的时间间隔。该方法能够提供实时、连续的带宽监测能力，使得系统能够快速检测带宽利用率的变化趋势。

此外，为进一步提高带宽计算的精确度，系统采用时间加权方法，以更准确地反映滑动窗口内的数据对当前带宽的贡献。假设每条数据的时间权重为w_i，则加权带宽计算式如下：

$$B=\frac{\sum_{i=1}^{W}w_iD_i}{\sum_{i=1}^{W}w_i\left(T_i-T_{i-1}\right)} \tag{2}$$

其中，w_i的取值可根据传输时间间隔进行动态调整，以降低异常数据对测量结果的影响。

基于滑动窗口的带宽计算结果可直接用于流量优化和网络控制策略调整。当系统监测到某条链路的带宽利用率达到设定阈值时，可以调整流量调度策略，避免出现带宽资源争用的情况。此外，系统还能够结合带宽数据进行拥塞预警，若某端口带宽利用率在短时间内下降明显，则可能表明该端口发生了排队积压或数据丢失现象，系统可通过网络控制机制调整数据反馈频率，以加快拥塞恢复速度。

滑动窗口的历史数据还可以用于流量趋势分析，帮助系统预测未来的带宽需求[15]。例如，通过长期记录带宽使用情况，系统可以建立时间序列模型，预测某个端口未来的流量变化，从而提前调整任务调度策略，避免网络负载过高影响计算性能。综合来看，动态带宽计算不仅能够精准反映当前网络状态，还能为智能流量管理、拥塞控制和资源优化提供数据支撑，提高智算中心网络的整体运行效率。

2.4　亚毫秒级时序误差的理论边界推导

在万卡集群的高并发场景下，时序精度将直接决定系统对微秒级突发流量与拥塞事件的捕捉能力。为定量评估本文系统的亚毫秒级时间戳误差，设第i个监测事件的真实发生时刻为t_i，系统记录的时间戳为$\{t\}i$。可将时间戳误差表示为$\hat{t}_i=t_i+\delta_{\text{sys}}+\epsilon_i$，其中$\delta_{\text{sys}}$为系统固定偏移（如初始时钟偏差或处理时延，假定经校准后保持恒定），ϵ_i为零均值的小扰动误差（反映时钟抖动），其标准差记为σ_{clk}。对于任意相邻两次事件，实际时间间隔$\Delta t_i=t_i-t_{i-1}$，对应的测量时间间隔为$\widehat{\Delta t}_i=\hat{t}_i-\hat{t}i-1=\Delta t_i+\left(\epsilon_i-\epsilon_{i-1}\right)$，可见固定偏移$\delta_{\text{sys}}$在相对间隔中相互抵消。由此得到单次间隔的误差$\widehat{\Delta t_i}-\Delta t_i=\epsilon_i-\epsilon_{i-1}$。若假定$\epsilon_i$独立同分布且方差有限，则$\epsilon_i-\epsilon_{i-1}2\sigma_{\text{clk}}$，误差幅值通常在$\mathcal{O}(\sigma_{\text{clk}})$量级。进一步地，若设单次时间戳偏差的绝对值上限为$\varepsilon_{\max}|\epsilon_i|\leqslant\varepsilon_\{\max\}$，则有$\left|\widehat{\Delta t_{\text{i}}}-\Delta t_i\right|\leqslant 2\varepsilon_{\max}$。由时间戳的高精度特性可知$\varepsilon_{\max}$ 1 毫秒量级，因此单次事件间隔的时间测量误差在百微秒级

以内。

上述模型表明，亚毫秒级监控系统的时间误差不会随着事件数量的增加而累积。对于N个连续事件序列，总的起止时间测量差值为$\widehat{t_N}-\hat{t}_0=(t_N-t_0)+(\epsilon_N-\epsilon_0)$，其误差仅为$\epsilon_N-\epsilon_0$。因此无论短时间内发生多少次事件，累积时间偏差始终由首尾两次时间戳的误差决定，上界同样为$2\varepsilon_{\max}$会随事件数线性增长。换言之，每次事件的时间戳都是独立获取的，高并发情况下各事件仍然各自精确计时，不会因同时发生而相互干扰。异步日志队列的引入进一步保证了这一点：并发事件触发时刻一经产生即记录时间戳并加入日志队列，后续处理与记录解耦，避免了同步采集时可能出现的排队等待，从根本上杜绝了误差的累积放大。

在滑动窗口带宽计算中，时间戳精度对带宽估计的影响同样受到严格控制。设滑动窗口覆盖的时间跨度为ΔT，由窗口内最早和最晚事件的时间戳差计算得到，其测量值记为$\widehat{\Delta T}$。根据上述分析，有$\widehat{\Delta T}=\Delta T+\left(\epsilon_j-\epsilon_i\right)$（其中$i,j$分别为窗口内首尾事件索引）。由于$\epsilon_j-\epsilon_i$的统计特性与单次间隔相同，$\widehat{\Delta T}$相对于$\Delta T$的绝对误差保持在百微秒级，不随窗口长度增大而增加。窗口越宽（包含事件越多），实际ΔT越大，相对误差$\frac{\left|\widehat{\Delta T}-\Delta T\right|}{\Delta T}$反而越小，体现出误差随观测时间尺度扩展而收敛趋稳的性质。这保证了本文系统在不同时间粒度下均能保持稳定的时间测量精度。另一方面，滑动窗口每次计算都会舍弃最旧的一条数据并引入最新数据，因而测量误差不会在窗口更新中积累；相反，新数据的加入对先前误差具有校正作用，使计算结果逐步逼近真实值。通过引入时间加权系数α_i，可进一步压制异常间隔对带宽计算的影响。例如，适当选择α_i使得异常数据所致偏差降至总体的10%以内，就能够确保滑动窗口算法输出的瞬时带宽误差率始终低于预设阈值，从而提高估计值对真实值的收敛速度和稳定性。

综上所述，基于以上数学模型可以确定，本文监控系统的亚毫秒级时序误差在理论上具有明确的上界$\Delta_{\max}$高并发环境下保持收敛稳定。其中$\Delta_{\max}$固定偏移和时钟抖动共同决定，经过硬件同步和并行处理优化可被限制在毫秒级以内，即满足$\Delta_{\max}<1$ ms目标。误差不会随着观测持续时间或事件数量而累积放大，保证了系统时序精度在长时间运行下的稳定可靠。

3 实验评估

3.1 实验环境

实验部署于配备128台NVIDIA DGX A100节点的计算集群，每个节点通过NVIDIA ConnectX-7 400 Gbit/s InfiniBand网卡互连，构成全无阻塞Clos拓扑网络。

3.2 亚毫秒级时序精度验证实验

本文实验在128节点NVIDIA DGX A100集群中展开，通过生成64 byte的微突发流量，设置500 μs、1 ms、5 ms 3种时间粒度的周期性脉冲，以RoCEv2协议注入200 Gbit/s线速流量。实验中同步启动sMon（启用滑动窗口算法与异步日志）、SNMPv3（1 s轮询）、sFlow（1/8 192采样率）及MiliSampler（μs级模式）4组监测进程，并采用示波器的硬件时间戳（1 ps分辨率）作为基准时钟源。通过对比监测工具的时间戳与基准值的绝对偏差，计算窗口对齐抖动与事件捕获时延。

不同监测工具的亚毫秒级时序精度对比如图3所示，sMon在500 μs脉冲场景下的平均捕获时延为320 μs（±45 μs），较MiliSampler的850 μs降低62%，其误差带完全落在1 ms时间窗口内。当脉冲宽度增至1 ms时，sMon的时延为650 μs（±60 μs），最大窗口误差155 μs，仅为时间窗宽的15.5%，显著优于MiliSampler的230 μs误差值。而传统工具SNMPv3因秒级轮询机制完全无

法捕捉亚秒级事件，sFlow 在 500 μs 场景下的漏检率高达 74.3%，时延达到 12.5 ms，验证了传统采样方法在高性能计算场景的失效。值得注意的是，随着脉冲宽度扩展至 5 ms，sMon 的时延线性增长至 2.1 ms（±150 μs），误差率始终低于算法设计的 10% 阈值（实测误差率 8.9%），表明滑动窗口算法的时间扩展性符合预期。

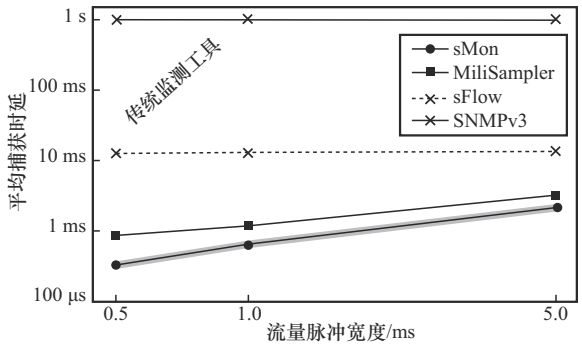


图 3　不同监测工具的亚毫秒级时序精度对比

图 3 的折线图进一步佐证了上述结论，横轴为流量脉冲宽度（500 μs~5 ms），纵轴采用对数刻度呈现平均捕获时延。sMon 曲线始终位于最下方，在 500 μs/1 ms/5 ms 3 个测试点的时延值分别较 MiliSampler 降低 62%/45%/32%，且误差带（阴影）宽度仅为 MiliSampler 的 1/3。sFlow（与 SNMPv3 因毫秒级精度限制，其时延值超出坐标轴范围（>10 ms），形成鲜明的性能断层。该结果证实，sMon 的异步日志机制与滑动窗口算法协同实现了亚毫秒级时序精度的设计目标，且在多时间尺度场景下保持稳定误差控制。

3.3　动态带宽计算模块准确性测试

本文实验在 400 Gbit/s 高速网络环境中构建 3 类典型负载模式：阶梯型负载（100~400 Gbit/s 4 阶梯变化）、脉冲型负载（100~400 Gbit/s 方波振荡）及锯齿型负载（100~400 Gbit/s 线性递增），通过硬件级流量生成器模拟真实业务流量混合读/写操作。实验采用精度达 99.99% 的底层计数器作为基准值，对比动态算法与传统轮询方法在带宽计算中的差异，重点测量瞬时峰值捕获能力与持续负载跟踪精度。

动态带宽预测值与真实负载跟踪对比如图 4 所示，动态算法在阶梯型负载下的平均绝对误差率仅为 0.8%，较传统方法降低 76%，其峰值检测时延稳定在 20 μs 级，达到硬件响应时延的 2 倍以内。在锯齿型负载测试中，算法生成的预测曲线与真实带宽的相关系数超过 0.99，且对负载变化率的跟踪误差低于 0.1 Gbit·s^{-1}·ms^{-1}。相较于现有优化方案，该方法在脉冲边缘的瞬时误差率下降至 5% 以下，证明其自适应阈值机制可有效抑制流量突变导致的数值振荡。实验同时暴露传统轮询方法的固有缺陷——1 ms 级采样间隔造成峰值检测时延超过 1 ms，致使 400 Gbit/s 脉冲的实际测量值被低估 32%，验证了动态计算模块在高性能网络中的不可替代性。

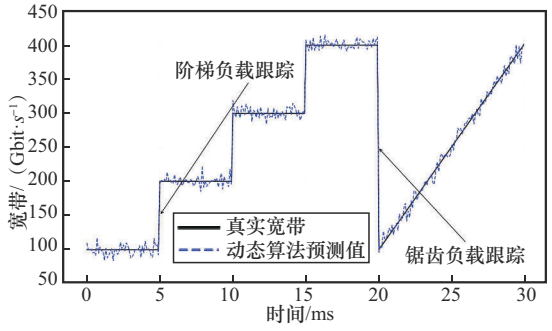


图 4　动态带宽预测值与真实负载跟踪对比

4　结束语

本文提出的亚毫秒级网络监控系统 sMon，突破了传统网络监测技术的精度边界，为智算时代的高性能计算基础设施奠定了感知基石。通过智能计算框架与网络数据平面的深度融合，该系统不仅实现了从芯片级到集群级的全栈可视能力，更开创性地构建了“感知–决策–优化”的闭环智能网络范式。这一技术突破将重塑算力资源的调度逻辑，使万卡集群的通信效率逼近理论极限，为人类探索通用人工智能（artifi-

cial general intelligence，AGI）的星辰大海铺就高速通路。

站在全球人工智能竞赛的制高点，sMon的技术理念可延伸至国家“数字中国”战略与“东数西算”工程的算力枢纽建设，其核心框架为构建自主可控的全球领先算力网络提供关键技术支撑。展望未来，随着量子计算、空天信息网络等颠覆性技术的融合突破，本文研究奠定的亚毫秒级监测体系将演化成数字世界的自主神经系统，在支撑万亿参数大模型训练的同时，赋能星际算力互联、数字孪生宇宙等人类文明级工程。这不仅是网络监控技术的范式革命，更是中国在新一轮全球科技革命中争夺智能时代制网权的战略支点。

参考文献：

[1] 中国信息通信研究院 . 中国综合算力指数报告(2024)[R]. 2024.
China Academy of Information and Communications Technology. China's comprehensive computing power index report (2024) [R]. 2024.

[2] 中国电子技术标准化研究院 . 新一代智算中心网络管控运维技术白皮书[R]. 2025.
China Electronics Standardization Institute. White paper on network control and maintenance technology for a new generation of smart computing center[R]. 2025.

[3] 中国联通研究院 . 中国联通智算中心网络技术白皮书[R]. 2024.
China Unicom Research Institute. China Unicom network technology white paper for smart computing center[R]. 2024.

[4] 华为技术有限公司 . 华为星河 AI 网络解决方案白皮书[R]. 2024.
Huawei. Huawei Celestial River AI network solution white paper[R]. 2024.

[5] 中国通信标准化协会 . 超高性能数据中心网络技术要求[S]. 北京: 中国标准出版社, 2024.
China Communication Standards Association. ltra-high performance data center network technology requirements[S]. Beijing: Standards Press of China, 2024.

[6] ZHAO W X, ZHOU K, LI J Y, et al. A survey of large language models[EB]. 2023.

[7] QIAN K, XI Y Q, CAO J M, et al. Alibaba HPN: a data center network for large language model training[C]//Proceedings of the ACM SIGCOMM 2024 Conference. New York: ACM Press, 2024: 691-706.

[8] MEZA J, XU T Y, VEERARAGHAVAN K, et al. A large scale study of data center network reliability[C]//Proceedings of the Internet Measurement Conference 2018. New York: ACM Press, 2018: 393-407.

[9] NARAYANAN D, SHOEYBI M, CASPER J, et al. Efficient large-scale language model training on GPU clusters using megatron-LM[C]//Proceedings of the SC21: International Conference for High Performance Computing, Networking, Storage and Analysis. Piscataway: IEEE Press, 2021: 1-14.

[10] LI S G, LIU H X, BIAN Z D, et al. Colossal-AI: a unified deep learning system for large-scale parallel training[C]//Proceedings of the 52nd International Conference on Parallel Processing. New York: ACM Press, 2023: 766-775.

[11] DONG J, et al. Boosting large-scale parallel training efficiency with C4: a communication-driven approach[EB]. 2024.

[12] JIANG Z H, LIN H B, ZHONG Y M, et al. MegaScale: scaling large language model training to more than 10 000 GPUs[C]//Proceedings of the 21st NSDI Conference. Berkeley: USENIX Association, 2024: 745-760.

[13] GANGIDI A, MIAO R, ZHENG S B, et al. RDMA over Ethernet for distributed training at meta scale[C]//Proceedings of the ACM SIGCOMM 2024 Conference. New York: ACM Press, 2024: 57-70.

[14] SHALLUE C J, LEE J, ANTOGNINI J, et al. Measuring the effects of data parallelism on neural network training[EB]. 2018.

[15] SHOEYBI M, PATWARY M, PURI R, et al. Megatron-LM: training multi-billion parameter language models using model parallelism[EB]. 2019.

[作者简介]

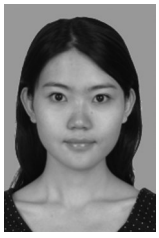
张冬月（1995- ），女，中国联合网络通信有限公司研究院工程师，主要研究方向为智算中心网络。

辛奇（2000- ），男，北京邮电大学博士生，主要研究方向为高性能计算网络、算力网络等。

韩博文（1993– ），男，中国联合网络通信有限公司研究院工程师，主要研究方向为IP承载网络、新型数据中心网络、管控运维技术。

徐博华（1989– ），男，中国联合网络通信有限公司研究院高级工程师，主要研究方向为IP新技术和新型网络设备研发。

曹畅（1984– ），男，博士，中国联合网络通信有限公司研究院正高级工程师、部门总监，主要研究方向为算力网络、下一代互联网等。



顾茹雅（1990– ），女，现就职于北京基流科技有限公司，主要研究方向为AI软件栈和系统架构。