International Journal of Applied Mathematics
Volume 38 No. 3s, 2025
ISSN: 1311-1728 (printed version); ISSN: 1314-8060 (on-line version)

MONITORING AND ANALYZING LATENCY AND PERFORMANCE IN ULTRA-LOW-LATENCY ENVIRONMENTS POWERED BY RDMA

[1]Ajay Prasad,[2] Hardik Mahant

[1]Fremont, CA, USA
Independent Researcher

Email: rushtoajay@gmail.com

[2]San Jose, CA, USA

Independent Researcher

Email: hardik.s.mahant@gmail.com

Abstract

This paper explores methods for monitoring and analyzing end-to-end (E2E) or round-trip time (RTT) latency in ultra-low latency (ULL) environments, focusing on Remote Direct Memory Access (RDMA) for real-time workloads such as high-performance computing (HPC), AI/ML training, and distributed systems. It highlights the adoption of RDMA-enabled technologies like InfiniBand and RoCE in modern data centers to achieve microsecond-level performance, while addressing challenges including limited observability due to kernel bypass, measurement overhead, scalability constraints, and security concerns such as packet injection attacks. The analysis categorizes ULL systems across layers such as hardware/kernel, storage, network, application, system-wide, and security, linking each to relevant metrics, thresholds, and tools. Unlike existing solutions that often rely on high-overhead traditional monitoring or cloud-dependent zero-overhead approaches such as proprietary systems like Zero+, the proposed framework advocates for lightweight, custom monitoring solutions that integrate time-series databases, background daemons, and domain-specific prototypes to bridge observability gaps, reduce perturbations in hybrid environments, and enable scalable, vendor-agnostic diagnostics tailored for RDMA's unique architectural constraints.

Keywords: RDMA, ultra-low-latency (ULL), round-trip time, high-performance computing, observability, security vulnerabilities, time-series databases

Introduction

Many organizations are investing heavily to modernize their data centers to support ULL environments and use RDMA for fast data access. InfiniBand and RDMA-enabled solutions have become their default choice as business demands are driving these investments. With frequent technology refresh cycles for both hardware and software, in many cases software patches and upgrades are not only performance-driven but also mandated by compliance policies. The underlying assumption is that these updates will enhance performance and reliability. The growth of HPC systems increases the complexity with respect to understanding resource utilization, system management, and performance issues [1]. To ensure seamless operations, organizations rely on monitoring tools to track the health of systems, devices,

Received: July 27, 2025

International Journal of Applied Mathematics
Volume 38 No. 3s, 2025
ISSN: 1311-1728 (printed version); ISSN: 1314-8060 (on-line version)

applications, and databases. These tools are often automated to generate tickets in the event of application outages or service disruptions, enabling faster resolution and improved customer service. In ULL and HPC environments, monitoring E2E or RTT latency is especially critical. Real-time applications depend on minimal delay, and even small inefficiencies can degrade performance. Achieving and maintaining these minimal latencies is key to delivering optimal outcomes in modern digital enterprises. Monitoring these critical systems and devices is challenging.

There exists a gap in how to monitor and what to monitor on these systems, including a lack of holistic monitoring frameworks [2,27]. Current monitoring approaches suffer from fundamental architectural mismatches between traditional network monitoring paradigms and RDMA's unique operational characteristics. Research reveals that existing tools operate in isolated silos—monitoring network flows without understanding RDMA verb semantics, tracking host performance without correlating NIC internal states, and measuring application metrics without visibility into underlying transport anomalies [2]. This fragmented approach creates blind spots where performance degradation occurs at the intersection of layers, precisely where ULL applications are most vulnerable.

This study offers comprehensive guidance on monitoring strategies and key metrics in ULL environments, systematically breaking down the system into layered components such as hardware/kernel, storage, network, application, system-wide, and security, to enable precise diagnostics, performance optimization, and security assurance in RDMA-enabled deployments.

Table 1 shows each layer and showcases what to monitor and some parameters for monitoring thresholds.

Layer	**Key Metrics (General + RDMA-Specific)**	**Why Monitor?**	**Thresholds for Ultra-Low Latency**
Hardware/Kernel	Interrupt latency, CPU scheduling delays, RDMA NIC counters such as QP retransmissions, completion errors	Detect OS/RDMA hardware bottlenecks; RDMA bypasses kernel but still faces interrupts.	<10µs interrupts, <1% retransmissions.
Storage	I/O latency (random/sequential), NVMe command completion times, NVMe-over-Fabrics latency, persistent memory access times, storage queue	Critical for data-intensive ultra-low-latency applications, storage often becomes bottleneck in optimized systems; RDMA storage disaggregation requires monitoring	Local NVMe <10µs (4K random), NVMe-oF <50µs E2E, persistent memory <500ns, queue depth <32

Received: July 27, 2025

International Journal of Applied Mathematics
Volume 38 No. 3s, 2025
ISSN: 1311-1728 (printed version); ISSN: 1314-8060 (on-line version)

	depth, cache hit ratios, RDMA storage verbs latency		
Network	RTT, jitter, packet loss, microbursts, hop-by-hop delays, RDMA-specific ECN markings, queue depth (via INT), out-of-order packets, PFC pauses	Identify propagation/congestion in RDMA fabrics; monitor for lossless assurance.	RTT <1ms (RoCE: <4µs), jitter <100µs, zero PFC-induced pauses.
Application	E2E processing time, job launch latency, API response times, RDMA verbs latency such as send/receive completions	Ensure real-time performance in RDMA-accelerated apps (e.g., AI training).	E2E <500µs, verb completions <2µs.
Security	Packet injection attempts, access token validations, performance isolation metrics, encryption overhead	Protect against vulnerabilities in direct memory access; ensure isolation in multi-tenant setups.	Zero unauthorized accesses; isolation overhead <5%
System-Wide	Throughput vs. latency trade-offs, error rates, RDMA link utilization, power consumption such as DRAM in RDMA op	Balance HPC scales; monitor for efficiency in RDMA clusters.	100Gbps+ with <1ms latency; power variance <5%.

Table 1: Monitoring layers

3. Service Components

Systems designed for a specific service or purpose, such as HPC, distributed AI training, or real-time data analytics, can be systematically categorized into five core components, compute, storage, network, application, and security. Each component plays a pivotal role in the system's functionality, and the overall performance of an application is inherently interdependent on their seamless integration and optimization. In RDMA-enabled ULL environments, where data transfers must occur with minimal overhead and microsecond-level responsiveness, this

Received: July 27, 2025

International Journal of Applied Mathematics
Volume 38 No. 3s, 2025
ISSN: 1311-1728 (printed version); ISSN: 1314-8060 (on-line version)

categorization becomes even more critical. RDMA allows direct memory-to-memory communication bypassing the CPU and OS, enabling high throughput and low latency, but it also introduces unique challenges across these layers, such as visibility gaps and security vulnerabilities. Below, I'll elaborate on each component, their roles in RDMA systems, and how they collectively influence application performance.

3.1 Compute

The emergence of hardware such as NVMe, SSD, and low-latency networks has driven a shift toward performing more processing at the edge and within caching tiers. The architecture typically consists of three primary tiers which aer data caching, content page caching, and database query result set caching. These systems provide very fast read IO. In memory Databases such as Couchbase , Oracle Times Ten, AWS CDN and Cloudflare's Caching falls into this area [6]. Monitoring I/O latency and network latency, RTT, in compute is essential for optimizing application performance. I/O latency refers to the time taken for a storage device to respond to data requests, which can significantly impact how quickly applications retrieve and process data. On the other hand, network latency, including RTT, measures the delay in data transmission between the source and destination, affecting the responsiveness of applications and services.

Ultra-low latency RDMA systems represent the pinnacle of high-performance networking, where microsecond-level improvements can translate to significant competitive advantages. These systems demand careful orchestration of compute components including specialized CPUs, memory subsystems, cache hierarchies, and interconnects. The challenge lies not only in optimizing individual components but also in monitoring their performance in real-time applications where traditional observability tools fall short. Current monitoring tools face fundamental limitations in microsecond-level environments. Traditional network monitoring operates at millisecond granularity, while RDMA systems achieve latencies down to 10 microseconds [3] or even sub-microsecond levels [4]. This creates a significant observability gap where critical performance events occur below the measurement threshold of conventional tools. RDMA operations bypass the traditional network stack, making them invisible to standard monitoring tools. The kernel bypass and zero memory copy techniques [3] that provide RDMA's performance benefits also eliminate many of the instrumentation points that monitoring systems rely upon. This creates blind spots in, Direct memory access operations, Hardware-level error handling, NIC-to-NIC communication patterns. Hence, existing monitoring solutions can not capture events occurring at microsecond timescales. The deployment of effective monitoring in ultra-low latency RDMA systems requires addressing fundamental architectural limitations. Organizations implementing high-frequency trading systems [5] or AI/ML distributed training workloads face the challenge of maintaining performance visibility while minimizing monitoring overhead.

3.2 Storage

As distributed storage systems like Ceph or custom solutions increasingly rely on high-speed fabrics such as InfiniBand or RoCE for low-latency data access, monitoring becomes critical to ensure performance, detect congestion, and maintain reliability. Traditional storage metrics

Received: July 27, 2025

International Journal of Applied Mathematics
Volume 38 No. 3s, 2025
ISSN: 1311-1728 (printed version); ISSN: 1314-8060 (on-line version)

such as IOPS and throughput are still relevant, but InfiniBand/RDMA introduces unique challenges like kernel bypass, which can make traffic harder to inspect, and the need for lossless networking to avoid packet drops. Diagnosing end-to-end host bottlenecks in storage systems is crucial for achieving fast data access, particularly in environments utilizing RDMA. Real-time performance monitoring in HPC clusters has emerged as a critical requirement, with tools evolving to integrate with open-source stacks like Prometheus and Grafana for visualization and alerting. The implementation of Prometheus and Node Exporter for real-time metrics collection has become standard, offering comprehensive monitoring for distributed systems. Additionally, the ELK (Elasticsearch, Logstash, Kibana) stack provides robust log-based monitoring and analysis capabilities [8]. 400G line rate requirements demand PCIe 5.0 x16 bandwidth (~64 GB/s) to avoid bottlenecks [9]. Ceph distributed storage systems benefit significantly from RDMA monitoring integration. CERN operates several Ceph storage clusters with over 100 PB capacity, providing block storage for OpenStack, CephFS for containers and HPC clusters, and S3 object storage for cloud-native applications [10]. With the emergence of cloud-native RDMA solutions like Zero+, there is a shift toward more efficient, low-overhead monitoring approaches. This trend has important implications for monitoring strategy selection and implementation. The Zero+ monitoring system, developed for Alibaba Cloud, achieves zero overhead in collecting raw metrics using one-sided RDMA. This system supports 1–10k hosts with 0.1–1s sampling intervals using a single thread for network I/O, while maintaining 80–95% bandwidth utilization when monitoring 1k hosts [11]. While every organization would love Zero+-like monitoring, there is a big upfront investment in such systems. Also, many organizations do not want complete dependency on the cloud and would like to stay in a hybrid environment. So, it becomes challenging to decide how to monitor and what to monitor.

3.3 Network

ULL environments have become critical for real-time applications, including high-frequency trading, autonomous systems, and distributed machine learning. RDMA technology has emerged as a cornerstone for achieving microsecond-level latencies by bypassing traditional kernel networking stacks [11]. However, monitoring and diagnosing performance in these environments presents unique challenges due to the extreme precision required and the risk of measurement overhead affecting the very performance being measured. Unlike traditional TCP/IP networking, RDMA bypasses the kernel networking stack, eliminating context switches and memory copies that introduce latency. This technology provides applications with bare-metal latencies in the order of single-digit microseconds [12].

ULL monitoring requires tracking multiple complementary metrics, including End-to-end RTT, One-way latency, Jitter measurements, Tail latency analysis. Different applications require tailored monitoring approaches. For example, industrial control applications may target reliable end-to-end latency down to hundreds of microseconds [11]. Modern systems enable continuous latency metrics without injecting any traffic into the network [13]. This always-on monitoring capability is essential for maintaining performance visibility without impacting the applications being monitored. Despite significant advances, several challenges continue to

Received: July 27, 2025

International Journal of Applied Mathematics
Volume 38 No. 3s, 2025
ISSN: 1311-1728 (printed version); ISSN: 1314-8060 (on-line version)

hinder ULL monitoring. Even passive measurement techniques can introduce subtle performance overhead, while ensuring scalability and maintaining precision across large, distributed systems remain difficult. Further complications arise from gaps in standardization and the absence of unified monitoring frameworks that work consistently across different RDMA implementations.

3.4 Application

Similar to storage, compute, and network layers, application layer monitoring presents unique challenges for application performance monitoring (APM) due to stringent timing requirements and specialized network protocols. Traditional monitoring approaches often introduce unacceptable overhead or fail to capture the nuanced performance characteristics of RDMA-based systems [14]. This analysis examines key insights and approaches for effective application layer monitoring in these demanding environments. The application component is central to delivering functionality and user value. Application-level monitoring focuses on tracking the performance, availability, and behavior of software applications in real-time, going beyond infrastructure metrics to emphasize end-user experience, code efficiency, and business impact. Unlike lower-level monitoring, it provides visibility into how applications perform under load, detects bottlenecks in code execution, and ensures seamless user interactions. It involves collecting telemetry data such as response times, error rates, resource usage, and transaction traces to identify issues like slow database queries or API failures. Important aspects include, digital experience monitoring, Application performance monitoring (APM), Code-level insights, Database query efficiency, Infrastructure monitoring tied to application health.

In environments using RDMA for ultra-low latency, it reduces overhead, provides higher throughput, and achieves latencies as low as single-digit microseconds. RDMA's efficiency comes at the cost of visibility, as it operates outside traditional network stacks. This leads to several monitoring hurdles, such as, kernel bypass and limited visibility, Intra-host and inter-host bottlenecks, Performance isolation in multi-tenant environments, Scalability and overhead in large-scale deployments, Application-specific requirements, as not all applications are RDMA-aware. RDMA middleware monitoring, as exemplified by X-RDMA, provides comprehensive visibility with minimal overhead [15]. For high-throughput environments, eBPF-based monitoring tools offer the necessary performance to monitor multi-gigabit RDMA traffic without impacting application performance [16]. RDMA application monitoring requires domain-specific approaches that account for the unique performance characteristics and deployment requirements of ultra-low latency systems.

3.5 Security

Ultra-low latency environments that use RDMA face significant security challenges related to direct memory access capabilities. The technology allows applications to access remote memory directly without CPU involvement, which creates potential attack vectors [17]. The NeVerMore research demonstrates critical vulnerabilities in RDMA protocols, showing how unprivileged users can inject packets into RDMA connections and bypass operating system security mechanisms [18]. These attacks can significantly increase short flow completion times

Received: July 27, 2025

International Journal of Applied Mathematics
Volume 38 No. 3s, 2025
ISSN: 1311-1728 (printed version); ISSN: 1314-8060 (on-line version)

(by factors of 10x or more), creating a fundamental trade-off between performance optimization and security isolation [19]. RDMA protocols face unique network security challenges. The RoCE protocol relies on Priority-based Flow Control (PFC) to enable a drop-free network, but PFC can lead to poor application performance due to problems like head-of-line blocking and unfairness. These network-level vulnerabilities can be exploited to degrade system performance or create denial-of-service conditions [20].

InfiniBand implements multilayered security through hardware-enforced key mechanisms that act as secure access tokens, along with Globally Unique Identifiers (GUIDs) that are hard-coded into every node and port [21]. For application-level security, the Pilaf system demonstrates how to detect read-write races without client-server coordination by implementing data structures that can verify their own integrity [22]. Current RDMA security mechanisms have significant limitations. The NeVerMore research reveals fundamental vulnerabilities in RDMA protocols that current security mechanisms do not adequately address, such as the ability for unprivileged users to inject packets into RDMA connections [18]. Performance isolation remains a significant challenge, with existing congestion control protocols like DCQCN unable to resolve the fundamental trade-off between performance isolation and work conservation [23]. This limitation makes RDMA systems vulnerable to performance-based attacks in multi-tenant environments.

4. Discussion

Research on monitoring RDMA-enabled ultra-low latency environments has emerged as a critical area of inquiry due to the increasing adoption of RDMA technology in data centers and cloud-native infrastructures, which demand high throughput and minimal latency for applications such as distributed machine learning, storage, and real-time analytics [24].

Despite RDMA’s advantages, monitoring these ultra-low latency environments presents distinct challenges. Existing network monitoring tools and diagnostic mechanisms often fail to capture intra-host bottlenecks and subtle performance anomalies inherent to RDMA subsystems [25]. The AMD Elba evaluation exposed measurement and instrumentation challenges, feature-induced performance collapse, many interdependent metrics requiring careful interpretation, testbed and workload-generation constraints, and complexity when silicon offloads fall back to CPU processing [26]. Traditional network monitoring focuses on inter-host traffic flows, leaving the internal host datapath comprising RNIC-to-PCIe-to-CPU/GPU connections, largely invisible to diagnostic tools [27]. The Hostping system, deployed across thousands of servers in distributed ML clusters, revealed that loopback latency and bandwidth tests within individual hosts could expose six distinct classes of previously undetected bottlenecks that external monitoring completely missed. They are, CPU root port failures, memory channel flapping, disabling ATS resulting in high PCIe latency, enabling "slow start" on the RNIC, setting "Tx window" too small on the RNIC, and overloaded inter-socket buses[27]. These internal path degradations can silently deteriorate application throughput even when inter-host links and switch fabrics appear healthy, demonstrating the inadequacy of conventional monitoring approaches. The complexity of RDMA hardware interactions creates additional diagnostic blind spots. Furthermore, RDMA's architectural

Received: July 27, 2025

International Journal of Applied Mathematics
Volume 38 No. 3s, 2025
ISSN: 1311-1728 (printed version); ISSN: 1314-8060 (on-line version)

characteristics actively undermine standard telemetry collection mechanisms. The stateful nature of RDMA connections, limited concurrent writer capabilities, and one-sided operation semantics make traditional monitoring approaches both technically impractical and performance-disruptive [28,29]. Zero+ demonstrated that conventional CPU-intensive monitoring can itself perturb latency-sensitive services during high-load events, while Direct Telemetry Access (DTA) analysis revealed that RDMA's constrained primitives require specialized translator and aggregation layers to enable scalable telemetry collection [28,29]. These fundamental incompatibilities between RDMA operation and existing monitoring paradigms necessitate entirely new approaches to performance visibility in ultra-low latency environments.

This study provides an alternative solution to monitor the RDMA enabled Ultra-Low Latency environments. Shown below in **Table 2,** are some tools available to diagnose and identify issues with systems running on RDMA.

Layer	Tool
Hardware/Kernel	perf, cyclictest, hwlatdetect, irqbalance, intel-pcm , ibstat, ibdiagnet , perfquery , rdma-core tools , mlx5_core counters
Network	iperf3, netperf, tcpdump/Wireshark, ss, ethtool, ibping, rping, rcopy, ucmatose, rdma_bw/rdma_lat, INT monitoring,PFC, monitoring, ECN monitoring, DCQCN monitoring,
Storage	fio, ioping, iotop, nvme-cli, blktrace, iostat, nvme-of-cli, spdk_tgt, nvmf_tgt, rdma-rxe, fio RDMA engine, pmempool, ndctl, daxctl, pcm-memory, perf mem, ceph-mon, gluster-cli, lustre-monitoring, beegfs-ctl, Custom coherency checker, perf stat cache, Intel PCM cache, NVMe device cache, Application cache profiler
Application	verbs-based profilers, MPI profilers, NCCL profilers, UCX profilers, strace, ltrace, gprof, Intel VTune, perf record/report
System-Wide	Prometheus plus Grafana, InfluxDB plus Telegraf, Zabbix, Nagios, Custom RDMA dashboard, Fabric Manager, UFM, OpenSM, ibnetdiscover
Security	IPsec, sRDMA, libreswan, Wireshark, tcpdump, scapy, mlxconfig, OVS, DPDK, BlueField Security SDK

Table2: Tools available to monitor

These tools can be incorporated into a monitoring program as shown in **Fig. 1,** and can be deployed onto each building block of the ultra-low latency environment to run as a service or a “daemon” on Unix-based systems. Utilizing a time series database and integrating it with tools like SolarWinds, organizations can create their very own custom-built monitoring systems. This will allow effective tracking of performance trends over time as well. Time series

Received: July 27, 2025

International Journal of Applied Mathematics
Volume 38 No. 3s, 2025
ISSN: 1311-1728 (printed version); ISSN: 1314-8060 (on-line version)

databases are specifically designed to handle time-stamped data, making them ideal for capturing and analyzing latency metrics.

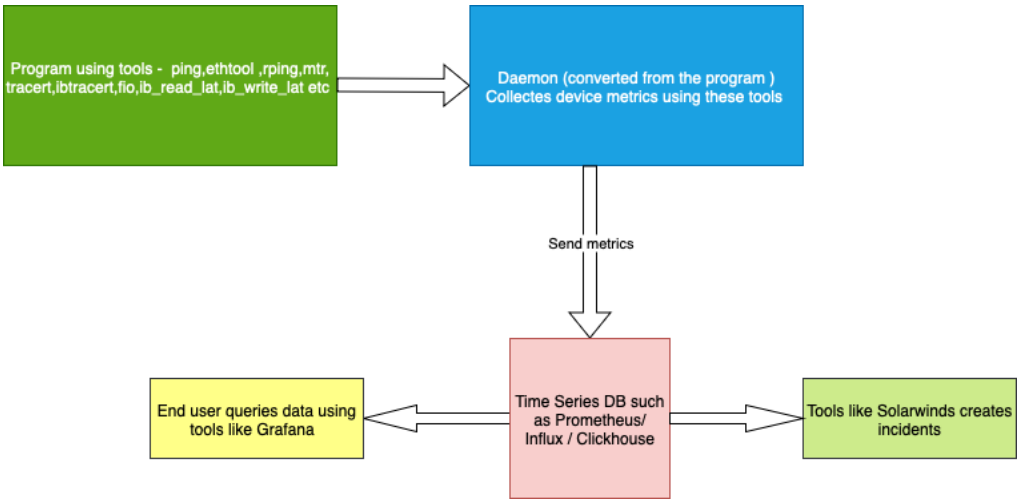

Fig1: Metric Collection and Alerting Pipeline in RDMA Systems

In light of these persistent challenges, future directions in RDMA monitoring must evolve toward AI-driven telemetry for predictive optimization in ultra Ethernet and InfiniBand interconnects, enabling proactive anomaly detection and performance tuning in heterogeneous environments.[30] Emerging approaches could incorporate advanced encryption and authentication protocols tailored for RDMA traffic, alongside software-defined scheduling to expose RNIC resources without compromising low-latency guarantees. Ultimately, bridging these gaps requires innovative frameworks that prioritize zero-CPU overhead, interoperability across RoCE and iWARP, and integration with emerging standards for planetary-scale RDMA, ensuring resilient performance for next-generation AI and HPC workloads.

5. Conclusion

This study examines methods for monitoring and analyzing E2E round-trip time (RTT) latency in ULL environments, with a focus on leveraging RDMA for real-time workloads such as HPC, AI/ML training, and distributed systems. It emphasizes the growing adoption of RDMA-enabled technologies such as InfiniBand and RoCE in modern data centers to achieve microsecond-level performance, while also addressing key challenges including limited observability due to kernel bypass, measurement overhead, scalability constraints, and security concerns such as packet injection attacks.

To structure the analysis, the study categorizes ULL systems across multiple layers—hardware/kernel, storage, network, application, system-wide, and security, linking each to relevant metrics and thresholds, including RTT below 4 µs for RoCE and interrupt latency below 10 µs. It also surveys monitoring tools such as Prometheus, Grafana, eBPF, as well as RDMA-specific utilities including rdma_bw and perf for diagnostics. Component-level considerations are highlighted, such as caching strategies in compute, Ceph with NVMe-oF in

Received: July 27, 2025

ISSN: 1311-1728 (printed version); ISSN: 1314-8060 (on-line version)

storage, jitter and tail latency in networks, low-overhead APM, and InfiniBand key management with performance isolation trade-offs for security.

Ultimately, the study advocates for lightweight, custom monitoring solutions integrating time-series databases and background daemons to bridge observability gaps, reduce overhead, and deliver consistent performance across hybrid environments designed for real-time workloads.

References

[1] Izadpanah, R., Naksinehaboon, N., Brandt, J., Gentile, A., and Dechev, D., 2018, August. Integrating low-latency analysis into HPC system monitoring. In Proceedings of the 47th International Conference on Parallel Processing (pp. 1-10).

[2] Li, Y., Lu, Y., Duan, J., Liu, H., Zhao, Y., Liu, Y., Duan, L., and Cui, L., 2022. Collie: Finding Performance Anomalies in RDMA Subsystems. In Proceedings of the 19th USENIX Symposium on Networked Systems Design and Implementation (NSDI '22) (pp. 287-305).

[3] RDMA Transports in Datacenter Networks: Survey. (2024). IEEE Network.

[4] Stuedi, P., Metzler, B., and Trivedi, A. (2013). jVerbs: ultra-low latency for data center applications. In Proceedings of the 4th annual Symposium on Cloud Computing (pp. 1-14).

[5] Gupta, D. B., Purohit, A., and Naresh, R. (2024). FPGA for High-Frequency Trading: Reducing Latency in Financial Systems. In Proceedings of the 2024 International Conference on Advanced Computing and Robotics Systems (pp. 1-6).

[6] Cloudflare. (n.d.). What is caching? Retrieved from https://www.cloudflare.com/learning/cdn/what-is-caching/#:~:text=What%20is%20CDN%20caching%3F,them%20along%20to%20other%20servers.

[7] Cui, H., Keeton, K., Roy, I., Viswanathan, K., and Ganger, G.R., 2015, August. Using data transformations for low-latency time series analysis. In Proceedings of the Sixth ACM Symposium on Cloud Computing (pp. 395-407).

[8] Kamath, M. S. S. (2024). Real-time Performance Monitoring of HPC Clusters: Techniques and Challenges. Indian Scientific Journal Of Research In Engineering And Management, 10.55041/ijsrem35707.

[9] AscentOptics. (2025). 400G Ethernet RDMA Network Cards: Revolutionizing Data Center Performance. Retrieved from https://ascentoptics.com/blog/400g-ethernet-rdma-network-cards/.

[10] Song, Z., Wu, J., Ma, T., Wang, Z. F., Kong, L., Kong, L., Wen, Z., Li, J., Kong, L., Yang, Y., Ma, T., Zheng, L., and Chen, G. (2024). Monitoring Large-Scale Cloud-Native Infrastructure Using One-Sided RDMA. IEEE ACM Transactions on Networking, 10.1109/tnet.2024.3394514.

Received: July 27, 2025

ISSN: 1311-1728 (printed version); ISSN: 1314-8060 (on-line version)

[11] Cecil, R., Setka, V., Tolar, D., and Sikora, A. (2020). RETIS – Real-Time Sensitive Wireless Communication Solution for Industrial Control Applications. IEEE.

[12] Stuedi, P., Metzler, B., and Trivedi, A. (2013). jVerbs: ultra-low latency for data center applications. Symposium on Cloud Computing.

[13] Sundberg, S., Brunström, A., Ferlin-Oliveira, S., Høiland-Jørgensen, T., and Brouer, J. D. (2023). Efficient Continuous Latency Monitoring with eBPF. Lecture Notes in Computer Science.

[14] Chang, H., Hanafy, W. A., Mukherjee, S., and Wang, L. (2024). INSERT: In-Network Stateful End-to-End RDMA Telemetry. Journal Article. https://doi.org/10.1109/infocom52122.2024.10621203.

[15] Ma, T., Ma, T., Zhuo, S., Jingxuan, L., Huaixin, C., Chen, K., Jiang, H., and Wu, Y. (2019). X-RDMA: Effective RDMA Middleware in Large-scale Production Environments. International Conference on Cluster Computing, 1-12. https://doi.org/10.1109/CLUSTER.2019.8891004.

[16] Sundberg, S., Brunström, A., Ferlin-Oliveira, S., Høiland-Jørgensen, T., and Brouer, J. D. (2023). Efficient Continuous Latency Monitoring with eBPF. Lecture Notes in Computer Science. https://doi.org/10.1007/978-3-031-28486-1_9.

[17] Mitchell, C. (2015). Building Fast, CPU-Efficient Distributed Systems on Ultra-Low Latency, RDMA-Capable Networks.

[18] Taranov, K., Rothenberger, B., De Sensi, D., Perrig, A., and Hoefler, T. (2022). NeVerMore: Exploiting RDMA Mistakes in NVMe-oF Storage Applications. arXiv preprint arXiv:2202.08080.

[19] Snyder, J., Lebeck, A. R., and Zhuo, D. (2023). RDMA Congestion Control: It Is Only for the Compliant. IEEE Micro.

[20] Zhu, Y., Eran, H., Firestone, D., Guo, C., Lipshteyn, M., Liron, Y., Padhye, J., Raindel, S., Yahia, M. H., and Zhang, M. (2015). Congestion Control for Large-Scale RDMA Deployments. Proceedings of ACM SIGCOMM.

[21] Schultz, S. (2025). InfiniBand Multilayered Security Protects Data Centers and AI Workloads. NVIDIA Technical Blog. Retrieved from https://developer.nvidia.com/blog/infiniband-multilayered-security-protects-data-centers-and-ai-workloads/.

[22] Mitchell, C., Li, Y., Li, J., and Ganger, G. R. (2013). Using One-Sided RDMA Reads to Build a Fast, CPU-Efficient Key-Value Store. In Proceedings of the 2013 USENIX Annual Technical Conference (ATC '13) (pp. 103-114).

[23] Zhang, Y., Tan, Y., Stephens, B., and Chowdhury, M. (2019). RDMA Performance Isolation With Justitia. arXiv preprint arXiv:1905.04437.

Received: July 27, 2025

ISSN: 1311-1728 (printed version); ISSN: 1314-8060 (on-line version)

[24] Liu, K., Zhang, J., Jiang, Z., Wang, W., Zhong, X., Tan, L., Pan, T., and Huang, T. (2024). Diagnosing End-Host Network Bottlenecks in RDMA Servers. IEEE ACM Transactions on Networking, 1–15. https://doi.org/10.1109/tnet.2024.3416419.

[25] Li, Y., Lu, Y., Duan, J., Liu, H., Zhao, Y., Liu, Y., Duan, L., and Cui, L. (2022). Collie: Finding Performance Anomalies in RDMA Subsystems. In Proceedings of the 19th USENIX Symposium on Networked Systems Design and Implementation (NSDI '22) (pp. 287-305).

[26] Advanced Micro Devices, Inc., "ARCHITECTURE MATTERS — Performance of a Multi-Stage SDN Pipeline on Arm® vs AMD Pensando™ Programmable Silicon," AMD white paper, 2024. https://www.amd.com/content/dam/amd/en/documents/pensando-technical-docs/white-papers/pensando-comparison-of-dpu-hardware-strategies.pdf

[27] Liu, K., Jiang, Z., Zhang, J., Wei, H., Zhong, X., Tan, L., Pan, T. and Huang, T., 2023. Hostping: Diagnosing intra-host network bottlenecks in {RDMA} servers. In *20th USENIX Symposium on Networked Systems Design and Implementation (NSDI 23)* (pp. 15-29).

[28] Song, Z., Wu, J., Ma, T., Wang, Z., Kong, L., Wen, Z., Li, J., Lu, Y., Yang, Y., Ma, T. and Liu, Z., 2024. Zero+: Monitoring Large-Scale Cloud-Native Infrastructure Using One-Sided RDMA. *IEEE/ACM Transactions on Networking*, *32*(4), pp.3499-3514.

[29] Langlet, J., Ben Basat, R., Oliaro, G., Mitzenmacher, M., Yu, M. and Antichi, G., 2023, September. Direct telemetry access. In *Proceedings of the ACM SIGCOMM 2023 Conference* (pp. 832-849).

[30] Aramide, O.O., 2025. Advanced Network Telemetry for AI-Driven Network Optimization in Ultra Ethernet and InfiniBand Interconnects. *SAMRIDDHI: A Journal of Physical Sciences, Engineering and Technology*, *17*(01).

Abbreviations:

RDMA: Remote Direct Memory Access

ULL: Ultra-low latency

RTT: Round-trip time

E2E: End to end

HPC: High-Performance Computing

QP: Queue Pair

NVMe: Non-Volatile Memory Express

NVMe-oF: NVMe over Fabrics

INT: In-band Network Telemetry

PFC: Priority-based Flow Control

ECN: Explicit Congestion Notification

Received: July 27, 2025

Volume 38 No. 3s, 2025
ISSN: 1311-1728 (printed version); ISSN: 1314-8060 (on-line version)

SSDs: Solid-State Drives

I/O: Input/Output

CDN: Content Delivery Network

IOPS: Input/Output Operations Per Second

ELK: Elasticsearch, Logstash, Kibana

PCIe: Peripheral Component Interconnect Express

PB: Petabyte

S3: Simple Storage Service

TCP/IP: Transmission Control Protocol/Internet Protocol

APM: Application Performance Monitoring

eBPF: extended Berkeley Packet Filter

CPU: Central Processing Unit

OS: Operating System

DoS: Denial-of-service

GUIDs: Globally Unique Identifiers

DCQCN: Data Center Quantized Congestion Notification

TSDB: Time Series Database

RNIC: RDMA Network Interface Card

DTA: DTA": Direct Telemetry Access

iWARP: Internet Wide Area RDMA Protocol

ATS: ATS": Address Translation Service

Tx: Transmit

UPI: Ultra Path Interconnect

RoCE: RDMA over Converged Ethernet

NOC: Network on Chip

Received: July 27, 2025