Preprints.org

Communication

Not peer-reviewed version

Design and Performance Evaluation of HEPS Data Center Network

Shan Zeng *, Tao Cui, Yanming Wang, Mengyao Qi, Fazhi Qi

Posted Date: 17 October 2025

doi: 10.20944/preprints202510.1395.v1

Keywords: HEPS; data center network; RoCE; performance

Preprints.org is a free multidisciplinary platform providing preprint service that is dedicated to making early versions of research outputs permanently available and citable. Preprints posted at Preprints.org appear in Web of Science, Crossref, Google Scholar, Scilit, Europe PMC.

Copyright: This open access article is published under a Creative Commons CC BY 4.0 license, which permit the free download, distribution, and reuse, provided that the author and preprint are cited in any reuse.

Preprints.org (www.preprints.org) | NOT PEER-REVIEWED | Posted: 17 October 2025 doi:10.20944/preprints202510.1395.v1

Disclaimer/Publisher's Note: The statements, opinions, and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions, or products referred to in the content.

Communication

Design and Performance Evaluation of HEPS Data Center Network

Shan Zeng *, Tao Cui, Yanming Wang, Mengyao Qi and Fazhi Qi

Computing Center, Institute of High Energy Physics, Chinese Academy of Sciences, Beijing 100049, China

* Correspondence: zengshan@ihep.ac.cn; Tel.: +86-10-88236018

Abstract

Among the 15 beamlines in the first phase of the High-Energy Photon Source (HEPS) in China, the maximum peak data generation volume can reach 1 PB per day, with the maximum peak data generation rate reaching 3.2 Tb/s. This poses significant challenges to the underlying network system. This paper designed a high-performance and scalable network architecture based on RoCE (RDMA over Converged Ethernet) to meet the needs of storage, computing, and analysis of HEPS scientific data. Test results show that the RoCE-based HEPS data center network system achieves high bandwidth and low latency characteristics, stably maintains reliable transmission performance during the interaction of scientific data storage, computing, and analysis, while also exhibiting good scalability and being able to adapt to the future expansion needs of HEPS beamlines.

Keywords: HEPS; data center network; RoCE; performance

1. Introduction

As the first 4th-generation synchrotron light source in Asia and the first high energy source in China, HEPS is a greenfield light source. Its storage ring energy is 6 GeV and its ring circumference is 1,360 meters [1]. HEPS can provide essential support for the breakthroughs in technological and industrial innovation as well as a state-of-the-art and multi-disciplinary experimental platform for basic science researchers. The 15 beamlines in Phase I are under construction and commissioning, and are expected to undergo acceptance inspection by the end of 2025, after which they will be made available for public services [2].

According to the estimated data rates, the maximum peak data generation rate reaching 3.2 Tb/s while 800 TB raw experimental data will be produced in HEPS per day in average from the 15 beamlines, and the data volume will be even greater with the completion of over 90 beamlines in Phase II. All the data will be transferred from beamline to HEPS data center, which requires a high-performance network environment with high bandwidth, ultra-low latency, and zero packet loss to ensure both the high-throughput computing (HTC) and high-performance computing (HPC) needs from beamline scientists.

With the evolution of Ethernet through features such as RoCE and improved congestion control, the dominated era of Infiniband or proprietary networks have been ended in the TOP500 supercomputers and open science platforms since 2020. Meanwhile, Mellanox market shares as well as the storage, hyperscale and hyperconverged markets shares have shifted heavily towards Ethernet and the trend is projected to continue [3]. According to the statistics of the TOP500 in November 2022, Ethernet occupies the first place in the interconnect family system share, with the proportion of 46.6%. Nearly all supercomputer architectures as well as leading data center providers utilize RDMA in production today [4]. Based on the former research and evaluation of RoCE in IHEP data center [5], we designed a high-performance and scalable network architecture based on RoCE to meet the needs of storage, computing, and analysis of HEPS scientific data.

© 2025 by the author(s). Distributed under a Creative Commons CC BY license.

Preprints.org (www.preprints.org) | NOT PEER-REVIEWED | Posted: 17 October 2025 doi:10.20944/preprints202510.1395.v1

2. Network Architecture Design

We designed a spine-leaf architecture comprising two spine switches and eight leaf switches, with 25G and 100G Ethernet connectivity to meet diverse bandwidth demands [6]. Seen in Figure1. Access layer gateway functions are hosted at leaf nodes, reducing latency and optimizing traffic routing between compute/storage resources and the core network. Between the spine and leaf layers, dynamic routing protocols enable adaptive path establishment, while Equal-Cost Multipath (ECMP) distributes traffic across redundant links—ensuring load balancing, maximizing throughput, and enhancing fault tolerance. This design delivers scalable, low-latency communication with efficient bandwidth utilization, making it well-suited for data-intensive applications such as AI training, HPC, and real-time analytics.

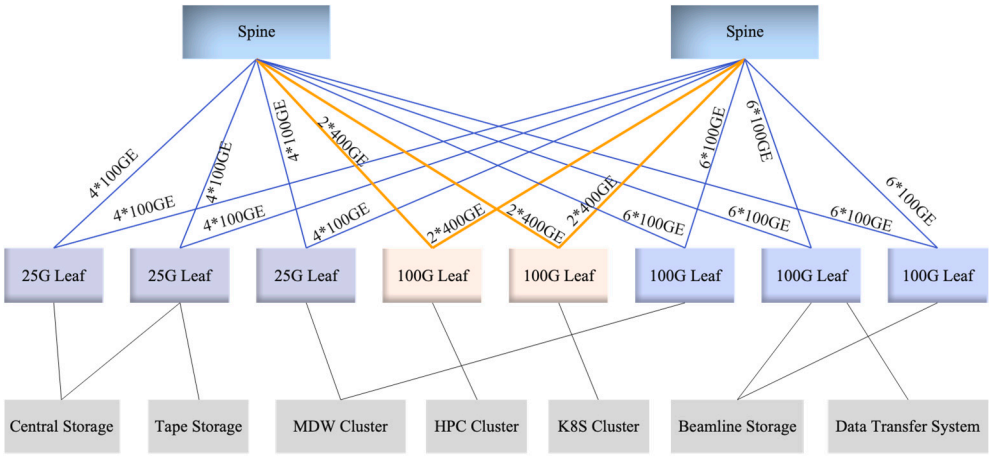

Figure 1. This is a schematic diagram of the network architecture in HEPS data center.

Data Center Quantized Congestion Notification (DCQCN), a critical congestion control framework standardized under IEEE 802.1Qau, is specifically designed for lossless, low-latency RDMA traffic—distinguishing it from traditional TCP-based approaches. This framework addresses critical challenges in modern data centers, including buffer overflow and in-cast congestion. A defining innovation of DCQCN lies in its compatibility with Priority-based Flow Control (PFC), which mitigates buffer starvation across distinct traffic classes [7]. Through integration with PFC, DCQCN balances congestion control and traffic prioritization, facilitating efficient coexistence between latency-sensitive RDMA flows and best-effort TCP traffic. Furthermore, its adaptive rate adjustment mechanism accommodates the bursty workloads characteristic of AI training, HPC, and cloud-scale analytics, where instantaneous bandwidth demands exhibit significant fluctuations. Consequently, DCQCN has been implemented in the HEPS data center to minimize end-to-end latency and optimize bandwidth utilization efficiency.

3. Results

We choose 7 servers that are actually in operation at the HEPS Data Center to conduct network throughput tests. Among the 7 servers, which includes data transfer server (transfer01), Kubernetes nodes (k8sgn08/k8sgn09/k8sgn12) and software framework nodes (daisycn01/daisygn02/daisygn01).

3.1. Experimental Environment

The network topology is shown in Figure2. Each server is uplink to a 100G leaf switch which uplinks to the two spine switches.

© 2025 by the author(s). Distributed under a Creative Commons CC BY license.

Preprints.org (www.preprints.org) | NOT PEER-REVIEWED | Posted: 17 October 2025 doi:10.20944/preprints202510.1395.v1

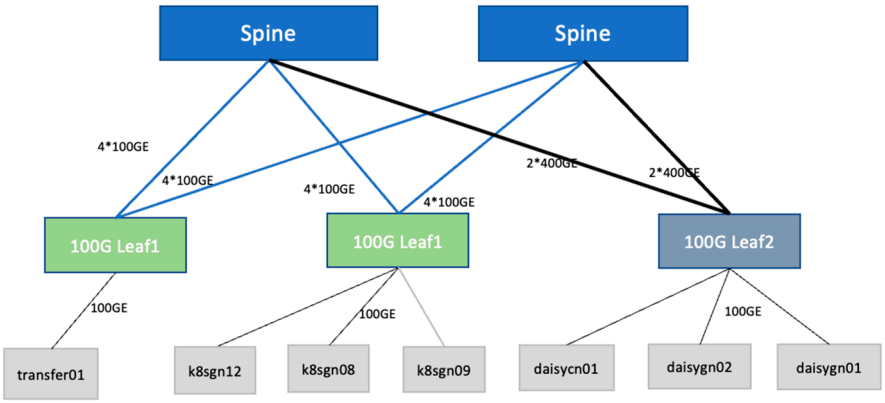

Figure 2. Network topology for the test.

The specifications of the test environment, including network interface card models, operating systems, network driver versions of the testing servers, as well as the manufacturers and models of network switches are presented in Table 1.

Table 1. Details of experimental setup.

Name	Type and version	Vendor
Testing server	UniServer R4900 G5	H3C
OS/Kernel	AlmaLinux release 9.3	-
NIC	Mellanox MT27800	NVIDIA
NIC Drivers/Firmware	mlx5_core 24.10-1.1.4	NVIDIA
100G Leaf1	CE8850	HUAWEI
100G Leaf2	CE8851	HUAWEI
Spine	CE9860	HUAWEI

3.2. PerfTest

PerfTest is an open-source performance testing toolkit specifically designed to evaluate the performance of InfiniBand and RoCE networks. It focuses on low-level hardware and protocol metrics, making it indispensable for validating, benchmarking, and optimizing high-speed interconnects in data centers [8].

PerfTest includes a suite of specialized tools tailored to different InfiniBand/RDMA operations, with common examples like:

ib_write_bw / ib_write_lat: Test bandwidth and latency for RDMA Write operations.

ib_read_bw / ib_read_lat: Focus on RDMA Read operations.

In the test, we selected two test servers connected to different leaf nodes and conducted ib_write_bw and ib_write_lat tests respectively, with the test results shown in Figure 3 and Figure 4. Figure 3 shows that the average network bandwidth of the ib_write_bw test between two 100G servers across spines can reach 92 Gb/s. Figure 4 shows that the average network latency of the ib_write_lat test between two 100G servers across spines is approximately 3.72 us. Test results indicate that the network throughput between the two servers communicating across spines can approach the theoretical peak, and the network latency is also as expected.

© 2025 by the author(s). Distributed under a Creative Commons CC BY license.

Preprints.org (www.preprints.org) | NOT PEER-REVIEWED | Posted: 17 October 2025 doi:10.20944/preprints202510.1395.v1


```
[root@k8sgn11 ~]# ib_write_bw -d mlx5_1 -R --run_infinitely --report_gbits 10.5.34.112
 WARNING: BW peak won't be measured in this run.
---------------------------------------------------------------------------------------
                    RDMA_Write BW Test
 Dual-port       : OFF          Device         : mlx5_1
 Number of qps   : 1            Transport type : IB
 Connection type : RC           Using SRQ      : OFF
 PCIe relax order: ON           Lock-free      : OFF
 ibv_wr* API     : ON           Using DDP      : OFF
 TX depth        : 128
 CQ Moderation   : 1
 Mtu             : 1024[B]
 Link type       : Ethernet
 GID index       : 3
 Max inline data : 0[B]
 rdma_cm QPs     : ON
 Data ex. method : rdma_cm
---------------------------------------------------------------------------------------
 local address: LID 0000 QPN 0x01db PSN 0xf8daef
 GID: 00:00:00:00:00:00:00:00:00:00:255:255:10:05:124:51
 remote address: LID 0000 QPN 0x004b PSN 0xbebce0
 GID: 00:00:00:00:00:00:00:00:00:00:255:255:10:05:34:112
---------------------------------------------------------------------------------------
 #bytes     #iterations    BW peak[Gb/sec]    BW average[Gb/sec]   MsgRate[Mpps]
Conflicting CPU frequency values detected: 4000.000000 != 3300.002000. CPU Frequency is not max.
 65536      878788           0.00               92.15              0.175755
Conflicting CPU frequency values detected: 4000.000000 != 3300.002000. CPU Frequency is not max.
 65536      878629           0.00               92.13              0.175724
Conflicting CPU frequency values detected: 4000.000000 != 3299.997000. CPU Frequency is not max.
 65536      878621           0.00               92.13              0.175723
```

Figure 3. Bandwidth testing result for RDMA write.

```
[root@k8sgn12 ~]# ib_write_lat -d mlx5_0 -R --report_gbits -n 100 10.5.32.251
---------------------------------------------------------------------------------------
                    RDMA_Write Latency Test
 Dual-port       : OFF          Device         : mlx5_0
 Number of qps   : 1            Transport type : IB
 Connection type : RC           Using SRQ      : OFF
 PCIe relax order: OFF          Lock-free      : OFF
 ibv_wr* API     : ON           Using DDP      : OFF
 TX depth        : 1
 Mtu             : 1024[B]
 Link type       : Ethernet
 GID index       : 3
 Max inline data : 220[B]
 rdma_cm QPs     : ON
 Data ex. method : rdma_cm
---------------------------------------------------------------------------------------
 local address: LID 0000 QPN 0x0090 PSN 0xfc2134
 GID: 00:00:00:00:00:00:00:00:00:00:255:255:10:05:125:52
 remote address: LID 0000 QPN 0x0159 PSN 0x78ab3a
 GID: 00:00:00:00:00:00:00:00:00:00:255:255:10:05:32:251
---------------------------------------------------------------------------------------
 #bytes #iterations    t_min[usec]    t_max[usec]  t_typical[usec]    t_avg[usec]    t_stdev[usec]   99% percentile[usec]   99.9% percentile[usec]
Conflicting CPU frequency values detected: 2100.000000 != 2599.059000. CPU Frequency is not max.
Conflicting CPU frequency values detected: 2100.000000 != 2557.536000. CPU Frequency is not max.
 2       100          3.70          3.80          3.72               3.72           0.00          3.80                   3.80
---------------------------------------------------------------------------------------
```

Figure 4. Latency testing result for RDMA write.

4. Discussion

The core working hypothesis of this study is that a RoCE-based spine-leaf network architecture, combined with Data Center Quantized Congestion Notification (DCQCN) and Priority-based Flow Control (PFC), can meet the high-bandwidth, ultra-low-latency, and scalable requirements of HEPS scientific data transmission (covering data offloading from beamlines, computing-node interaction, and storage access). The experimental results from the 7 in-service HEPS servers (including transfer, Kubernetes, and software framework nodes) strongly validate this hypothesis.

In the PerfTest evaluations, the average bandwidth of RDMA Write operations (via ib_write_bw) between 100G servers across spine switches reached 92 Gb/s—this is close to the theoretical peak bandwidth of 100G Ethernet (≈94 Gb/s, after accounting for protocol overheads such as frame headers). This performance confirms that the dual-spine (HUAWEI CE9860) and multi-leaf (HUAWEI CE8850/CE8851) topology, paired with Equal-Cost Multipath (ECMP) for traffic distribution, effectively eliminates inter-layer bandwidth bottlenecks. ECMP's ability to spread traffic across redundant links not only maximizes throughput but also ensures load balancing—critical for handling HEPS' bursty data generation (e.g., sudden spikes near the 3.2 Tb/s peak rate).

Meanwhile, the average latency of RDMA Write operations (via ib_write_lat) was approximately 3.72 us, which is far lower than the latency of traditional TCP/IP networks (typically 50–100 us for 100G links) and comparable to dedicated InfiniBand networks (≈4–5 us for similar-scale deployments). This ultra-low latency stems from two key design decisions: (1) Locating access-layer gateway functions at leaf nodes, which reduces routing hops between compute/storage resources and the core network (avoiding additional latency from intermediate gateways); (2) Integrating DCQCN with PFC, which addresses buffer overflow and in-cast congestion—two major causes of latency spikes in RDMA networks. Notably, the consistency of results across multiple test iterations (e.g., stable 92 Gb/s bandwidth and 3.72 us latency) further proves the architecture's ability to maintain

© 2025 by the author(s). Distributed under a Creative Commons CC BY license.

Preprints.org (www.preprints.org) | NOT PEER-REVIEWED | Posted: 17 October 2025 doi:10.20944/preprints202510.1395.v1

reliability under the variable workloads of HEPS (e.g., alternating between real-time data analysis and offline tape storage).

References

1. He, P., Cao, J., Lin, G., Li, M., Dong, Y., Pan, W. and Tao, Y., 2023. Update on HEPS Progress. Synchrotron Radiation News, 36(1), pp.16-24.
2. Li, X., Zhang, Y., Liu, Y., Li, P., Hu, H., Wang, L., He, P., Dong, Y. and Zhang, C., 2023. A high-throughput big-data orchestration and processing system for the High Energy Photon Source. Synchrotron Radiation, 30(6), pp.1086-1091
3. Kalia, A., Kaminsky, M. and Andersen, D.G., 2016. Design guidelines for high performance {RDMA} systems. In 2016 USENIX annual technical conference (USENIX ATC 16) (pp. 437-450).
4. Li, W., Zhang, J., Liu, Y., Zeng, G., Wang, Z., Zeng, C., Zhou, P., Wang, Q. and Chen, K., 2024, March. Cepheus: accelerating datacenter applications with high-performance roce-capable multicast. In 2024 IEEE International Symposium on High-Performance Computer Architecture (HPCA) (pp. 908-921). IEEE.
5. S Zeng, S., Qi, F., Han, L., Gong, X. and Wu, T., 2021. Research and Evaluation of RoCE in IHEP Data Center. In EPJ Web of Conferences (Vol. 251, p. 02018). EDP Sciences.
6. Liu, Y., Geng, Y.D., Bi, X.X., Li, X., Tao, Y., Cao, J.S., Dong, Y.H. and Zhang, Y., 2022. Mamba: a systematic software solution for beamline experiments at HEPS. Synchrotron Radiation, 29(3), pp.664-669.
7. Hu, Y., Shi, Z., Nie, Y. and Qian, L., 2021, October. Dcqcn advanced (dcqcn-a): Combining ecn and rtt for rdma congestion control. In 2021 IEEE 5th Information Technology, Networking, Electronic and Automation Control Conference (ITNEC) (Vol. 5, pp. 1192-1198). IEEE.
8. OFED perftest. 2024. https://github.com/linux-rdma/.

Disclaimer/Publisher's Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.

© 2025 by the author(s). Distributed under a Creative Commons CC BY license.