September 2023

Solution Proposal:Leveraging Ultra-Low Latency Lossless Ethernet to Power Distributed Storage Clusters

V1.5 | ASTERFUSION DATA TECHNOLOGIES

Table of Content

1. High-Performance Distributed Storage in the All-Flash Era Poses Challenges to Traditional Networks

Technologies such as cloud computing, big data, IoT, and artificial intelligence have gradually matured and scaled into production across industries. In this technical landscape, data volumes continue to experience exponential growth, driving a surge in the demand for large-capacity storage. Simultaneously, the expansion of compute capabilities and upper-layer business scales requires higher performance from underlying storage infrastructures.
Consequently, enterprises face a dilemma between centralized SAN/NAS storage (highly reliable and performant but expensive and proprietary) and distributed storage (large-capacity and low-cost but lower in performance). There is an urgent need for storage infrastructure that unifies these strengths: high performance, high reliability, massive scalability, and low overall cost.

1.1 High-Performance, Flexible, and Scalable Distributed Storage is the Future Trend

1.1.1 Centralized SAN/NAS Storage

Prior to the mass adoption of distributed storage, centralized storage architectures were predominantly deployed to address enterprise storage requirements. Centralized SAN/NAS storage physically utilizes a "Controller + Disk Enclosure" architecture. Mid-to-high-end storage arrays are typically configured with multiple controllers to guarantee high availability and scale performance. These controllers are tightly coupled, interconnected via PCIe buses or InfiniBand networks, and share a unified cache pool, delivering short I/O paths and ultra-low access latency.

Storage Controller-A	Storage Controller-B
Storage Interface (FC / SAS / iSCSI)	Storage Interface (FC / SAS / iSCSI)
BBU | Shared Cache	BBU | Shared Cache
Backend Bus	Backend Bus
PCIe / IB Interconnect	
Disk Enclosure (JBOD / SSD Storage Pool)	

Figure 1: Hardware Architecture of Centralized Storage

To ensure stability and reliability, traditional centralized SAN/NAS storage utilizes internal Battery Backup Units (BBU) or external UPS systems for power-fail protection. This guarantees that in-flight data residing in the volatile cache pool is cached and written to non-volatile disks during an abrupt power outage. This is combined with active-active clustering, disaster recovery, and continuous data protection (CDP) technologies to secure business continuity.
Traditional centralized storage hit the market early, features high technological maturity, and offers robust stability alongside high IOPS, low latency, and strong data consistency. These advantages match

core database storage requirements in mission-critical environments like financial services and healthcare. However, its proprietary architecture dictates that horizontal scaling is constrained by the physical limits of the storage controllers, rendering it poorly suited for modern, massive scale-out cloud environments and high-concurrency access scenarios.

1.1.2 Distributed Storage

To manage rapidly expanding pools of unstructured and mass data, a more elastic and horizontally scalable storage architecture emerged: distributed storage. As a modern software-defined technology, distributed storage utilizes standard commodity server hardware combined with intelligent distributed storage software. Standard x86 or ARM servers are interconnected via high-speed Ethernet, using software layers to abstract and pool local server drives into a single, unified, logical storage resource pool.

Distributed storage successfully achieves complete decoupling between storage hardware and software. This enables enterprises to build highly agile, standardized storage platforms within data centers, drastically lowering Total Cost of Ownership (TCO) and aligning with software-defined data center (SDDC) mandates.

Distributed storage fundamentally overcomes the scaling limitations of centralized storage. A cluster can scale horizontally up to thousands of physical nodes, capacity can extend into Petabyte (PB) or Exabyte (EB) dimensions, and overall performance increases linearly alongside capacity growth. Because distributed architecture allows multiple storage nodes to handle read/write operations simultaneously without relying on a centralized controller head, they provide exceptional aggregate throughput and concurrent I/O performance.

The flexible and open storage architecture provides distributed storage with numerous advantages over traditional centralized storage, including "infinite" capacity expansion, linear performance scaling, lower construction costs, and high manageability. However, this decentralized layout also introduces inherent performance trade-offs, such as longer physical I/O routing paths, high read/write tail-latency, and challenges with weak data consistency. Consequently, in legacy infrastructure designs, mission-critical application database backends remained on centralized storage, while distributed storage was primarily utilized for massive, cold, or secondary unstructured data pools.

1.2 The Network Has Become the Performance Bottleneck of Distributed Storage

By dissecting the precise internal mechanics of a distributed storage write/read lifecycle, we can better understand the root causes behind long I/O paths, latency spikes, and data synchronization bottlenecks. When a client writes a file to distribute storage, the file is first segmented into distinct data shards. Taking a single data shard write as an example: the shard is first transmitted across the network and committed to the disk of the primary replication node (Master Node). The primary node then handles replication management, distributing identical data copies across the network to secondary nodes (Slave Nodes). Only after the primary node receives verification that all secondary nodes have successfully committed the data to disk will it issue a final "Write Complete" acknowledgment back to the client application. Due to disaster recovery domain design constraints, replica data is typically written across different server racks. Therefore, a single file writes triggers multiple backend east-west network hops and cross-rack disk commits. Assuming a file is split into 100 shards under a standard 3-replica configuration without write-cache buffering, committing this single file requires 300 total write operations across various nodes over the network—100 writes to primary nodes, and 200 secondary

network distribution replication writes. The client must wait for all secondary confirmations before the operation finishes.
When a client reads data, it directly fetches shards from the primary node. However, because shards composing a single file are scattered across different physical storage nodes throughout the cluster, the client must pull chunks from multiple servers simultaneously and reassemble them locally. This read operation similarly triggers multiple cross-rack network transactions.
This read/write mechanism highlights the core source of distributed storage performance drops: the long internal network I/O path and typical one-to-many/many-to-one traffic patterns heavily congest network switches. Total storage I/O latency can be calculated as network latency + storage media latency + software processing latency.
When the underlying storage media consists of legacy mechanical Hard Disk Drives (HDD), storage media latency dominates the equation, rendering network latency negligible. However, the introduction of solid-state drives (SSDs) natively leveraging the Non-Volatile Memory Express (NVMe) protocol has delivered orders-of-magnitude improvements in both throughput and latency over legacy HDDs. This has led to the rise of **All-Flash Distributed Storage**.

Switch Type	Typical Latency	Storage Media	Typical Latency
Traditional Legacy Switch	8 µs	HDD Random Read/Write	1300 µs
Asterfusion ULL Switch	**0.4 µs**	**NVMe SSD Random Read/Write**	**10 µs**

Table 1: Latency Comparison Across Switching Architecture & Storage Media Generations

In all-flash distributed storage deployments, the network throughput easily reaches 20 to 30 GB/s per node. As storage media latency drops significantly due to hardware upgrades, the network transit delays become the absolute performance bottleneck of the entire storage environment. If network transit delays can be reduced, the aggregate performance of all-flash distributed storage scales up naturally. Beyond hardware component upgrades, distributed storage system vendors continue to optimize their backend replication engines to mitigate latency and data consistency issues. For instance, modern storage software platforms can write to three replica nodes concurrently rather than following a strict primary-then-secondary serialized sequence. Crucially, the integration of **RDMA (Remote Direct Memory Access)** technology completely bypasses host CPU kernels and OS protocol stacks, offloading data movement straight to network interface cards (NICs) to slash inter-node communication latency.
In conclusion, rapid iterations in storage media and storage software engines have elevated distributed platforms far beyond their historical performance limitations. However, these advancements place much higher demands on the network fabric. Building a low-latency, zero-packet-loss, high-performance network is now the absolute key to breaking flash storage performance bottlenecks.

2. Asterfusion Ultra-Low Latency Lossless Ethernet Breaks Storage Performance Bottlenecks

Based on a deep understanding of distributed storage workloads and RDMA protocol behaviors, Asterfusion leverages its CX-N series ultra-low latency cloud switches to deliver a low-latency, zero-

packet-loss storage fabric tailored for all-flash distributed storage clusters.

2.1 Leveraging RoCEv2 to Reduce Transport Protocol Latency

In distributed storage networks, standard TCP/IP processing consumes substantial host CPU cycles due to multiple kernel space memory copies and context switching. In the NVMe all-flash era, this legacy protocol stack creates severe software processing overhead. RDMA addresses this by enabling a network interface card to read/write memory addresses directly across physical hosts without remote kernel involvement, maximizing bandwidth while driving CPU utilization down toward zero.
Therefore, RDMA can be understood as leveraging hardware and network features to let Server A's NIC read/write Server B's memory directly. The application layer bypasses kernel intervention, enabling high throughput, minimal jitter, and low host utilization resource metrics.
Currently, RDMA is implemented over three types of networks: InfiniBand, iWARP, and RoCE. While InfiniBand delivers excellent native performance, it requires proprietary, high-cost host channel adapters (HCAs) and dedicated IB switches, leading to vendor lock-in. To lower TCO, the industry developed Ethernet-compatible alternatives: iWARP and RoCE (which includes RoCEv1 and RoCEv2). Among these, **RoCEv2 (RDMA over Converged Ethernet v2)** offers the best balance of performance, routing capabilities, and cost efficiency, making it the preferred standard for enterprise all-flash distributed clusters. The Asterfusion ultra-low latency lossless Ethernet fabric is fully optimized to carry RoCEv2 traffic, delivering a high-performance network that rivals InfiniBand at standard Ethernet cost points.

2.2 Ultra-Low Latency Switching Silicon to Reduce Network Forwarding Latency

Asterfusion CX-N cloud switches leverage advanced cut-through forwarding architectures to minimize port-to-port transit delays down to sub-microsecond levels. This ultra-low latency profile easily handles the tight timing budgets of scale-out distributed environments where inter-node data replication paths span multiple hops, significantly improving overall storage cluster transaction response speeds.

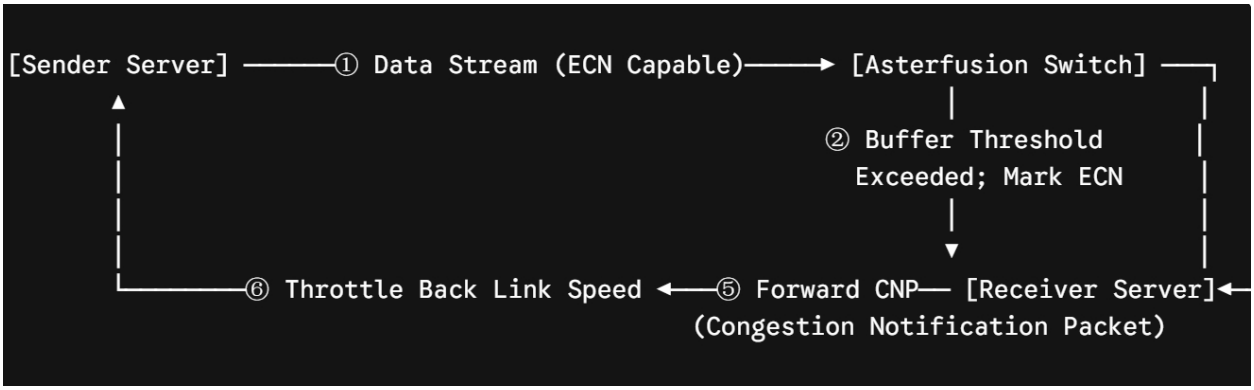
2.3 Deploying PFC High-Priority Queues to Ensure Zero Packet Loss for Storage Traffic

Distributed architectures natively generate massive many-to-one or one-to-many synchronous storage traffic patterns. This is known as the Incast traffic model, and it represents the primary cause of buffer overflow and packet drops within enterprise data centers. To combat this, Asterfusion utilizes **PFC (Priority-based Flow Control)** to logically isolate storage traffic, ensuring zero packet loss for RoCEv2 frames.
PFC operates by slicing a single physical Ethernet link into 8 independent virtual channels, each mapped to a distinct Class of Service (CoS) priority queue with dedicated buffer spaces. By assigning storage replication traffic to a designated high-priority queue, the switch can selectively pause non-storage workloads during periods of congestion without dropping or delaying critical storage traffic.

2.4 Implementing ECN Congestion Control Algorithm to Eliminate Network Congestion

While PFC guarantees zero packet loss, severe backpressure can cause pause frame propagation across upstream switches. To proactively mitigate congestion, Asterfusion implements **ECN (Explicit**

Congestion Notification) to avoid packet retransmissions, decrease network jitter, and enhance storage throughput.
When egress buffer queues cross preset ECN thresholds, the switch marks the ECN bits in the IP header of outbound packets to signal downstream congestion. Upon receiving these marked packets, the destination server generates a CNP (Congestion Notification Packet) and sends it back to the source host. The source server then throttles its transmission rate for that specific flow, proactively easing network congestion before buffers overflow and trigger PFC pause actions.



2.5 Simplifying Lossless Network Deployment and Operations with "One-Click RoCE"

Storage engineers focus heavily on tuning host-side operating systems and storage software, often finding the complex configurations of underlying lossless networks cumbersome. To simplify operations, Asterfusion developed the "One-Click RoCE" framework within its enterprise operating system, AsterNOS.
This feature abstracts atomic network configuration parameters into unified, business-level commands. With a single command execution, engineers can instantly apply recommended PFC and ECN thresholds, set up queue mappings, and access a centralized dashboard showing the real-time health and operational status of the lossless fabric, saving valuable time for upper-layer business optimization.

2.6 High-Density 100G/400G Interfaces and Two-Tier Network Architectures to Reduce Hops

The massive bandwidth needs of all-flash storage necessitate regular network upgrades. Asterfusion CX-N cloud switches offer up to 25.6 Tbps of bidirectional switching capacity per device, providing flexible configurations with up to 32 x 400G or 128 x 100G interfaces to support scaling storage environments. This high-density configuration allows enterprise architectures to build non-blocking, two-tier Spine-Leaf networks. Compared to legacy three-tier network topologies, this streamlined architecture guarantees that communication between any two storage servers never exceeds three switch hops, driving network forwarding latency down to a minimum.

3. Best Practices for Distributed Storage Scenarios

3.1 Performance Testing in Real-World Scenarios

The following tables illustrate the testing environment and performance benchmark results. The empirical results clearly demonstrate that Asterfusion's CX532P-N open Ethernet switch outperforms the Mellanox SB7700 InfiniBand switch across critical storage vectors.

Device Type	Model	Parameters / Specs	Qty
Compute Node	DELL R840	Intel Xeon Gold 6230 CPU @ 2.10GHz, 512G RAM, 100G ConnectX-5 NIC	2
Storage Node	ThinkSystem SR650	Intel Xeon Silver 4214 CPU @ 2.20GHz, 64G RAM, 1.6TB NVMe SSD*3, 100G ConnectX-5 NIC	3
100G IB Switch	Mellanox SB7700	32-Port 100G InfiniBand Switch	1
Ethernet Switch	Asterfusion CX532P-N	32-Port 100G ULL Open Ethernet Switch	1

Table 2: Hardware Environment for Performance Testing

Test Benchmark Vector	Mellanox SB7700 (100G IB)	Asterfusion CX532P-N (100G ULL Ethernet)
Latency: 4k Random Read (latr)	141.79 µs	**132.84 µs (Faster)**
Latency: 4k Random Write (latw)	79.67 µs	**71.6 µs (Faster)**
Latency: 8k Random Read (latr-8k)	150.64 µs	**145.83 µs (Faster)**
Latency: 8k Random Write (latw-8k)	80.89 µs	**73.89 µs (Faster)**
IOPS: 4k Random Read (2 Stress Hosts)	2548k	**2633k (Higher)**
IOPS: 4k Random Write (2 Stress Hosts)	850k	**916k (Higher)**
IOPS: 1024k Random Read (2 Stress Hosts)	17474	**21.2k (Higher)**

Test Benchmark Vector	Mellanox SB7700 (100G IB)	Asterfusion CX532P-N (100G ULL Ethernet)
IOPS: 1024k Random Write (2 Stress Hosts)	3673	**4820 (Higher)**

Table 3 Empirical Benchmark Performance Testing Results

3.2 Recommended Network Topology for Compute-Storage Disaggregated Scenarios

With the widespread adoption of all-flash distributed storage, performance-intensive business applications are migrating onto software-defined storage blocks. However, to unleash the maximum hardware capacity of NVMe SSDs, network optimization is required alongside software tuning to build a low-latency, zero-loss, high-throughput network fabric.

In a disaggregated compute and storage scenario, it is highly recommended to deploy two physically isolated Spine-Leaf networks. The storage backend network will exclusively occupy a dedicated physical fabric to guarantee that critical cluster data tasks (multi-replica synchronization, cluster node rebalancing, and drive rebuilds) run non-blocking. Meanwhile, the storage frontend network and general business network share a separate physical fabric. Based on varying reliability requirements, compute cluster nodes can connect via single-homed links, whereas storage cluster nodes should always utilize dual-homed multi-chassis link aggregation (MC-LAG) downlinks to secure active-active high availability at the network edge.

Following physical wiring, full-network traffic prioritization must be planned. Switches should be configured with PFC and PFC Deadlock Prevention to assign varied applications into distinct queues, ensuring RoCEv2 storage flows take absolute precedence. Concurrently, ECN congestion parameters must be enabled globally. Switch configuration policies must align precisely with host NIC properties regarding queue maps, buffer watermarks, PFC limits, and ECN triggers to maximize cluster throughput.

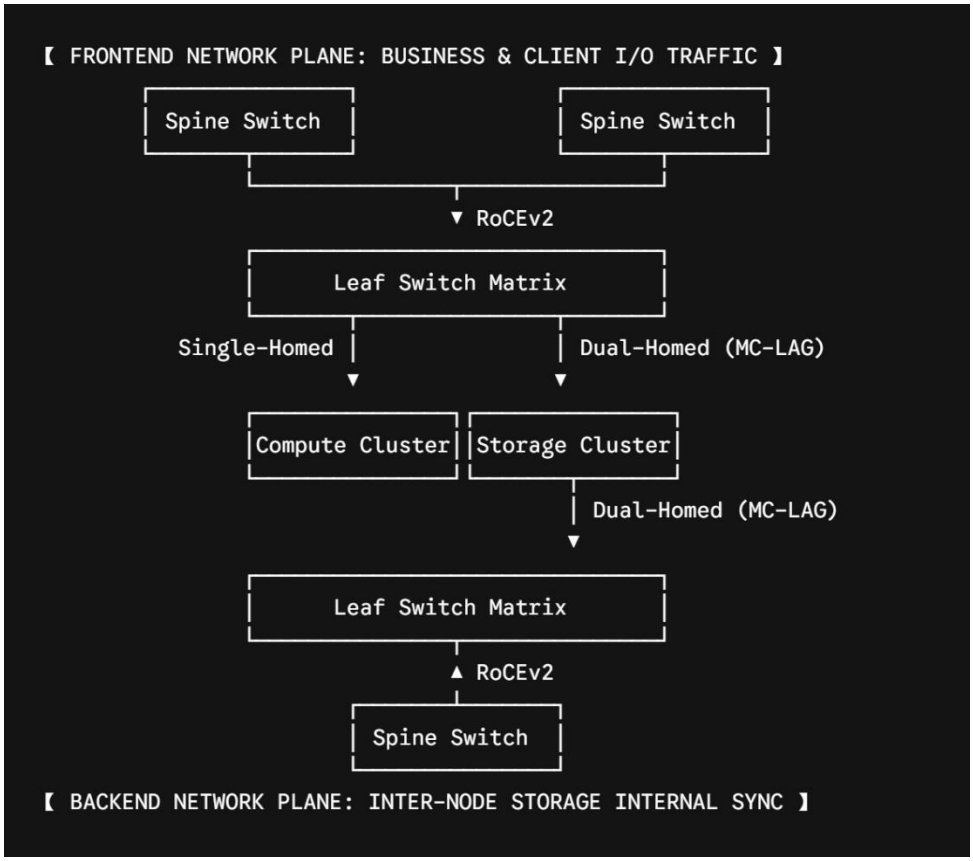


Figure 4: Network Blueprint for Compute-Storage Disaggregated Environments

3.3 Planning and Calculation for Storage Network Designs

This section outlines the design calculations for a storage backend network supporting 512 physical servers in a distributed storage cluster. The fabric uses 100GE links to interconnect the Spine and Leaf tiers, with storage nodes connecting to Leaf switches via 25GE RoCEv2 NICs. Switch selections based on cluster scaling parameters are defined in Table 3-4.

Server Node Scale	Leaf Switch Model	Spine Switch Model
N <= 512	CX308P-48Y-N/1U (48x25G + 8x100G)	CX532P-N/1U (32x100G)
512 < N	CX308P-48Y-N/1U (48x25G + 8x100G)	CX11008-N Modular Switch–

Table 3-4: Device Selection Guide for Scale-Out Storage

The design deploys the CX532P-N (32 x 100GE wire-speed ports) at the Spine tier, and the CX308P-48Y-N (48 x 25GE access ports and 8 x 100GE uplink ports) at the Leaf tier, using a target network oversubscription ratio of 4:3. Servers connect via MC-LAG active-active dual-homing. Based on these

parameters, the precise switch infrastructure quantities are calculated as follows:
1. Calculate Effective Uplink Bandwidth per Leaf Switch: When deploying Leaf switches in pairs using MC-LAG, 2 * 100GE interfaces on each switch are dedicated to the inter-chassis peer link. This leaves 6 * 100GE interfaces available as active uplinks to the Spine layer, providing an aggregate uplink bandwidth of **600 Gbps** per Leaf Switch (**6 * 100GE = 600 Gbps**).
2. Calculate Maximum Node Density per Leaf Switch: To maintain the target 4:3 oversubscription ratio with 600 Gbps of uplink capacity, each Leaf switch requires exactly **800 Gbps** of access-facing bandwidth (**600 Gbps * (4 / 3) = 800 Gbps**). At 25GE per client link, this allows each Leaf switch to support up to 32 downstream servers (**800 Gbps / 25GE = 32 Servers**).
3. Calculate Total Required Leaf Switches:
To dual-home 512 servers via MC-LAG with 32 active server connections per Leaf, the design requires a total of 32 Leaf switches:
Total Leaf Switches = (512 Servers / 32 Links) * 2 (for MC-LAG redundancy) = 32 Leaf Switches
4. Calculate Total Required Spine Switches: With 32 Leaf switches deploying 6 uplinks each, the fabric requires 192 total 100GE ports at the Spine layer (**32 * 6 = 192 Ports**). Since each CX532P-N Spine switch provides 32 x 100GE ports, the network requires exactly 6 Spine switches: **Total Spine Switches = 192 Ports / 32 Ports per Switch = 6 Spine Switches**

4. Conclusion

The ultra-low latency lossless Ethernet solution built on Asterfusion CX-N series cloud switches completely eliminates standard network forwarding bottlenecks via optimized, validation-backed system designs. By deploying high-throughput, sub-microsecond, zero-packet-loss storage fabrics, it unlocks the maximum linear performance scaling potential of modern all-flash distributed clusters.

About Asterfusion

Asterfusion Data Technologies is a leading provider of next-generation open cloud networking architectures. We deliver full-stack open networking solutions designed to help enterprises build open, secure, and cost-effective data center infrastructure, re-engineering the cloud networks infrastructure with our global ecosystem partners.