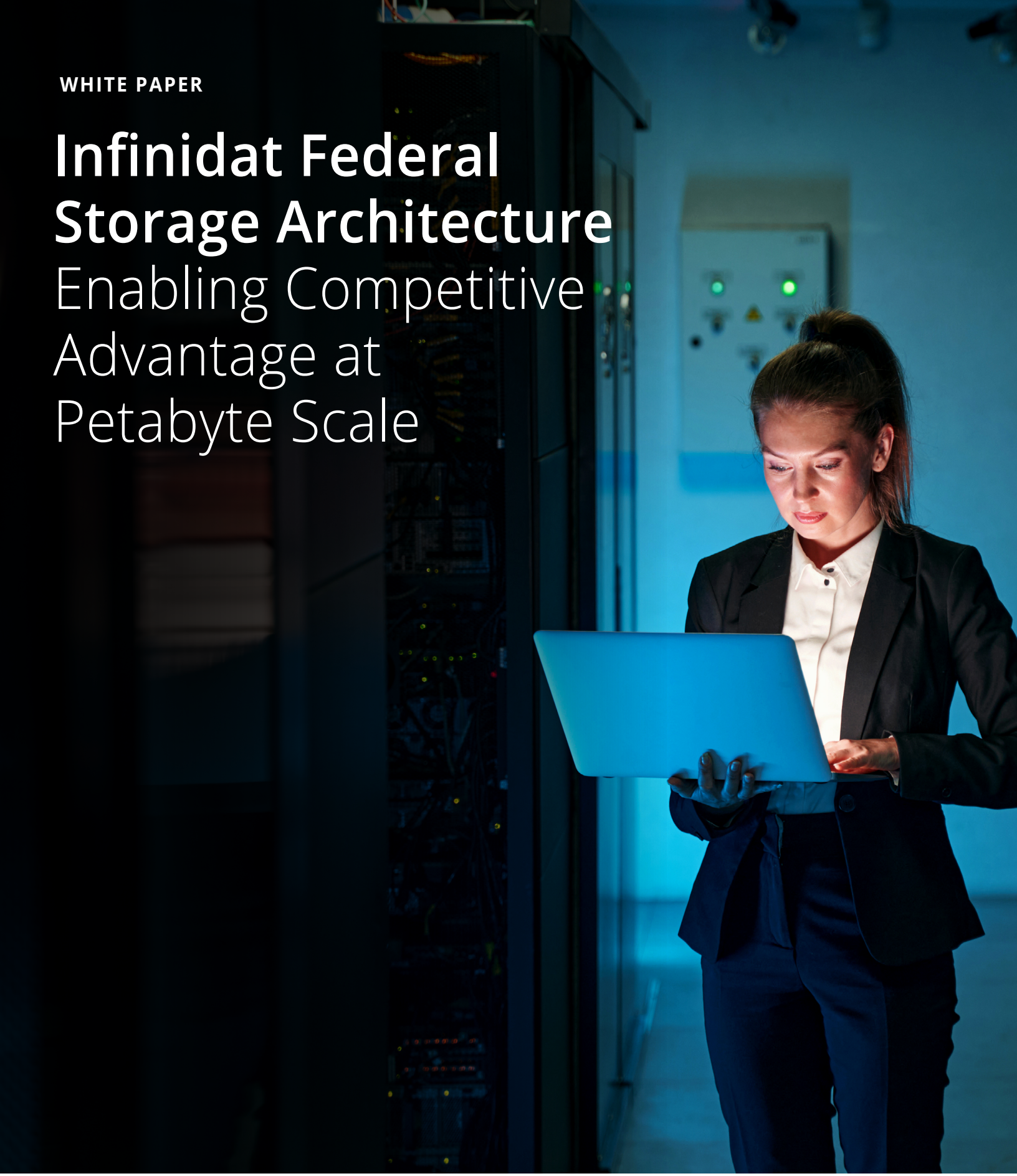
WHITE PAPER

Infinidat Federal Storage Architecture

Enabling Competitive Advantage at Petabyte Scale



INFINIDAT FEDERAL

Abstract

Infinidat Federal enterprise storage solutions are based upon the unique and patented InfiniBox storage architecture—a fully abstracted set of Software-Defined Storage (SDS) functions integrated with the best-of-breed commodity off-the-shelf hardware. InfiniBox's software-focused architecture, an evolution and revolution in data management design over 30 years in the making, solves the conflicting requirements to make storage bigger, faster, and less expensive. This paper discusses the technology that enables Infinidat Federal to be the only enterprise storage provider that achieves multi-petabyte capacity with groundbreaking performance (35 microsecond latency[1]) and an unprecedented 100% availability, all at the lowest Total Cost of Ownership (TCO).

[1] Internal Latency as low as 35 microseconds based on customer production telemetry data, read I/O hits (DRAM cache) actual results may vary. March 2022.

Design Principles

When designing a storage architecture to meet modern data center needs, multiple requirements must be satisfied:

CATEGORY	REQUIREMENT
Reliability	Businesses operate 24x7; downtime is not tolerated
Capacity	Exponentially increasing volumes of data, accelerated by digital transformation, disaggregated big data architectures, artificial intelligence (AI), and machine learning (ML)
Performance	As data scales, performance must keep pace to provide the same (or better) results in the same (or shorter) timeframes
Simplicity	Administrators expect simple operations, wide ecosystem integration and built-in tools to transition to DevOps models to spend less time managing storage, and more time on applications and business processes
Consolidation	Point technologies are a thing of the past; modern storage should accommodate all use cases for maximum efficiency, simplicity and cost savings
Cost	Budgets aren't scaling in line with capacity and performance growth requirements; a disruptive change in architecture is required
Resiliency	Organizations require resilience both within their infrastructure and from external threats such as cyber attacks

At the same time, public cloud infrastructure providers such as Amazon, Google, and Azure claim to reduce cost for the entire IT stack, and they often do for small customers who can't afford a large IT staff and often rely on one or two "jacks of all trades" to maintain their entire IT operation. However, for large organizations, as well as regional cloud and managed services providers, adopting a more efficient IT stack— one that meets their business, technology, and financial requirements and provides them with all the benefits of the cloud within their own infrastructure, while reducing costs and maintaining sovereignty over their data.

InfiniBox® Architecture

InfiniBox® was designed with the key principles in mind, to meet all these challenges:

PRINCIPLE	REASONING	CHALLENGE
Innovative software design	Software, unlike hardware, is optimized over time, improving performance instead of degrading it. InfiniBox is based on over 140 software patents—the true meaning Software Defined Storage (SDS). InfiniBox's innovative software includes it's patented Neural Cache algorithms, metadata management structures, and next-generation storage features, among many others.	**Performance** **Simplicity** **Reliability** **Cost**
Design for resiliency	When designing for scale, resiliency is critical. InfiniBox is designed for seven nines (99.99999%) reliability, uses a triple-redundant architecture where all critical components (software and hardware) have at least two (2) redundancies (N+2) protecting from downtime and data loss. InfiniBox offers a 100% system availability guarantee for InfiniBox and InfiniBox SSA based on tens of thousands of operating hours of experience backed by independent and industry-leading Gartner Peer Insights Reviews.	**Resiliency** **Cost** **Simplicity** **Consolidation**
Architecting for scale	Achieving capacity and performance at a disruptive cost requires scale. InfiniBox was designed for large customers scaling to 10PB or more effective capacity in a single 42U rack.	**Consolidation** **Cost** **Simplicity** **Capacity**
Resilient integration of hardware and software	Infinidat Federal evaluates all leading hardware component vendors, selecting only the most reliable ones to be used as part of the InfiniBox solution. This best-of-breed approach means that customers get a fully integrated and tested solution, as opposed to the typical SDS requiring complicated hardware integration and administration work.	**Reliability** **Simplicity** **Consolidation**
Commodity off-the-shelf (COTS) hardware	Using commodity hardware and avoiding long development cycles allows simpler adoption of new technologies. These include CPUs, memory types, and storage media. Using commodity hardware and its associated software, also brings more stability, as the same hardware is used in thousands of systems worldwide.	**Cost** **Reliability** **Capacity** **Simplicity** **Performance**

Building upon Strong Architecture

The InfiniBox architecture continues to deliver unprecedented performance, reliability, scalability, simplicity and economics. This architecture, introduced with InfiniBox, has also been the foundation for continued innovation. Today, Infinidat Federal offers even more choice, performance and scale extolling the virtue of this common architecture.

InfiniBox - Enterprise storage delivering AI-exploiting Neural Cache.

InfiniBox SSA - All solid-state components for the most demanding workloads.

InfiniGuard - Purpose-built data protection and recovery appliance with built-in cyber resilience.

The strength of this architecture has been demonstrated through the consistent delivery of innovation, competitive differentiation, and customer value over the past decade and into the future.

Performance Acceleration

InfiniBox uses a combination of DRAM, flash media (SSD), and high capacity NL-SAS disks to write, read, and store data. Below is the explanation of how reads and writes are accelerated to achieve maximum performance at minimum latency. The internal latency customers are experiencing with InfiniBox SSA, for example, is 35 microseconds. The algorithm used for data placement optimization is called Neural Cache. This section explains how Neural Cache provides customers with the industry's lowest latency by leveraging smart software algorithms. It is important to remember that most transactional applications require at least two separate I/Os (one to write a transaction to the logs and one to write data into the database), making latency the key component in determining both the user experience and the application's maximum performance.

METADATA LAYER

Metadata layer response times immediately affect I/O latency. InfiniBox accelerates metadata operations by:

- **All metadata is in DRAM**—metadata is kept in DRAM, accelerating both reads and writes.
- **Metadata structure**—A complete history of all data written to InfiniBox is managed in a metadata structure called a "Trie". This patented implementation captures the data placement, relevant organizational and virtual addressing information, and multiple layers of data protection.
- **Trie efficiency**—all inserts, modifications, and deletions from the Trie operate at the same latency, providing consistent performance from the first bytes of data to multiple petabytes.

WRITE ACCELERATION

InfiniBox accepts all writes without any pre-processing (such as pattern removal, compression, encryption, etc.) into its DRAM, and makes a second copy of the write in another node's DRAM over low-latency InfiniBand before sending the acknowledgment to the host. Accepting the write from DRAM (directly attached to the CPU), instead of an external flash device, allows InfiniBox to complete writes in the lowest possible latency.

Unlike many architectures, where write cache is broken down into small buckets (as in matrix-architectures and dual controller architectures), InfiniBox uses a single, large memory pool to accept writes. This allows larger write bursts to be sustained, allows data that changes frequently to get overwritten at DRAM latency, and allows Neural Cache time to make smart decisions, prioritizing which data blocks will benefit from DRAM speeds and which should be destaged to SSDs and HDDs. By keeping data longer in the write cache, Neural Cache avoids unnecessary workload on the CPU and back-end persistency layers.

Prior to destaging, each cache cycle collects randomly written data and reassembles it into larger sequential writes based on a number of factors including relevancy of the data as it was written, which aids the Neural Cache in predictive analytics later to determine which data may subsequently be needed in concert with each read operation.

READ ACCELERATION

Unlike traditional storage arrays, which aim to place the most active data (a.k.a. "hot data") in flash cache to achieve performance parity with traditional all-flash arrays, InfiniBox uses its innovative Neural Cache that aims to place all of the hot data in DRAM. The InfiniBox Neural Cache allows most reads to complete at DRAM speed, which is 1000 times faster than flash.

The company's global data fabric spans many exabytes of data and Neural Cache has been proven to provide almost all reads from DRAM, allowing customers to experience an "All-DRAM-Array"-like experience at a TCO lower than competing arrays.

Since Neural Cache is a learning algorithm, it optimizes performance over time. InfiniBox leverages a thick SSD flash layer, which serves as a "cushion" for DRAM-misses. As Neural Cache learns the I/O patterns and optimizes DRAM data placement, the flash layer changes its function from handling DRAM-misses to handling changes in I/O patterns, which the algorithm may not be able to predict (e.g. periodic audit that requires data not in DRAM).

Software Architecture

When designing InfiniBox to sustain 100% availability, software is used to overcome the unpredictability of hardware failures. InfiniBox leverages an active-active-active software architecture and N+2 design providing constant monitoring, self-healing, and graceful recoveries from hardware failures on all levels.

All components are implemented in software, from the RAID to the clustered services, to allow constant optimization with each new release. Over the first five years since its first Generally Availability (GA) release, InfiniBox's maximum performance has improved by over 4x, just by non-disruptively upgrading the software. This is the power of a true software-defined solution.

CLUSTERED SERVICES

All data services run on all nodes, in accordance with the N+2 architectural design, and are active on all nodes (no passive nodes in the cluster). The data services are designed to run in user-space, including low-level components such as Fiber Channel (FC) drivers. Since no data services run in the kernel, no single service failure can affect other services in the system, or the node's availability. Further, each service can independently restart in mere seconds. These design principles apply to front-end services such as data protocols (NFS, iSCSI, FC, NVMe-oF) as well as to back-end data services such as Neural Cache, InfiniRaid®, and InfiniSnap®.

Data services are launched and monitored by the Cluster Manager (CLM), which identifies any service issues and can restart services when necessary. A service experiencing any failure will restart and self-test before re-joining the cluster. Any service failing to start correctly will not join the cluster to avoid failing when in the cluster (Byzantine failure). If the Cluster Manager identifies a service that tried to restart several times unsuccessfully on a specific node, it stops restarting it and notifies support. Any service failure—whether automatically recovered or not—is reported back to Infinidat Federal's data analytics platform to detect software issues and continuously improve code quality.

DISK LAYOUT

The InfiniBox's disk layout is managed by patented software innovation called InfiniRaid. InfiniRaid is software-defined Redundant Array of Independent Disks (RAID) controlling all data placement, data protection and the recovery from failure scenarios. InfiniRaid is a declustered RAID, which is a type of RAID that separates the data layout from the physical layer and uses thousands of virtual RAID groups, spreading data across all the drives and preventing any hot spots. InfiniRaid creates the RAID groups so that every two drives in the system only share up to 2.5% of their RAID groups. This low percentage of overlapping RAID groups has multiple benefits:

- **Self-healing**—any potential hot spots are automatically solved by the data layout optimization.

- **Virtual spares**—Space capacity is evenly spread across all disks in the system. There are no physical hot spares, allowing the rebuild process to redistribute data optimally and minimize unnecessary cost. The system holds enough spare capacity for up to 12 drives to fail in an F6000.
- **Performance protection**—a single drive failure (data is still protected) will only generate a low priority RAID rebuild ("Rebuild-1"), one that prioritizes application performance.
- **Fast recovery**—When a second drive fails, the system will prioritize the rebuild for the common 2.5% of RAID groups that are shared between the two failed drives ("Rebuild-2"), before reverting to the lower priority Rebuild-1 as there will be no more unprotected RAID groups.
- **InfiniSpares**—Beyond the capacity guaranteed for the equivalent of 12 spares, InfiniBox can also leverage free capacity as spare capacity, if needed. This innovation allows up to 100 disks to fail without losing protection.

Data Protection Services

InfiniBox offers many data protection services, to help customers protect their assets:

- **Snapshots**—The InfiniBox snapshot mechanism is called InfiniSnap, and is based on a non-locking, redirect-on-write mechanism that yields consistent performance with or without snapshots. Each dataset can have up to 1000 snapshots, each can be either read only (for data protection) or writeable (for testing and development environments). InfiniSnap performs snapshots in DRAM without requiring any writes to the persistent layer.
- **Immutable Snapshots**—InfiniSnap snapshots can also be marked "immutable", meaning that they cannot be modified or deleted once written pursuant to the requirements and timers established when the immutable snapshot is created. This capability provides excellent recovery data protection for Ransomware and similar threats.
- **Low RPO Asynchronous replication**—The asynchronous replication engine can achieve and maintain the industry's lowest Recovery Point Objective (RPO) with a 4 second replication interval, while using IP infrastructure to reduce cost and complexity.
- **Synchronous replication**—The synchronous replication engine provides synchronous data protection with zero RPO while maintaining latency below 400µs (microseconds) of storage latency. In the case of problems with the WAN (high latency, loss of connectivity), the InfiniBox synchronous replication engine automatically fails back to asynchronous mode. When the WAN is restored, the replication will automatically replicate all the missing data and resume sync replication without disrupting I/O.
- **Active-Active replication**—Active-active replication with InfiniBox systems allows simultaneous read and write to consistency groups over metropolitan distances. They maintain an external image of the volumes appearing as if they are multi-paths to the same volume, leveraging synchronous replication to keep the volumes consistent at all times. Without any master-slave relationship, there are no extraneous round-trips needed to perform write updates to any given volume. An external, lightweight "witness" can exist on a stand-alone node, or even in a virtual machine in a cloud.
- **Concurrent 3rd-Site replication**—Any consistency group in an Active-Active replication relationship can also be simultaneously replicated asynchronously to a third location without any additional performance penalty. Since each InfiniBox can sustain a second remote replica of any consistency group, each can be replicated to the same third InfiniBox system, or even further replicated to a fourth InfiniBox in a separate location.
- **InfiniSafe**—InfiniSafe extends cyber resilience capabilities to the InfiniBox family of products. This technology leverages immutable snapshots, creates a local logical air-gap for separation, establishes a fenced forensic environment, and provides near-instantaneous recovery from cyber attacks.

Data Reduction

InfiniBox employs multiple methods of data reduction, to further reduce the cost of storage, including:

- **Thin provisioning by default**—All volumes are thin provisioned by default. Since InfiniBox also offers smart capacity pools, the risk of over-allocation / over provisioning can easily be mitigated by setting alert thresholds and emergency buffers on the pool, protecting application availability.
- **Zero-reclamation**—As hosts (physical or virtual) clear space in a disk (LUN), they write zeros into that space either through the write-same operation (more efficient) or simply by writing individual zeros into that space. InfiniBox identifies both cases and removes this space, as if it were never written to, further improving thin provisioning.
- **Compression**—InfiniBox compresses data only once it is destaging from the write cache (DRAM) to disk. This accelerates writes (no added latency due to data reduction) while avoiding compressing any transient data that gets overwritten after a few seconds (saving CPU resources). InfiniBox compression leverages LZ4 with a 64KiB chunk size, generating a higher compression ratio than traditional small-block compression (commonly used in all flash arrays).
- **Snapshots**—InfiniBox snapshots are space-efficient by design, and help customers avoid the capacity and performance penalties of a full copy.

Network Architecture

For all network-based services, network accessibility is critical for availability. Specifically, for Internet Protocol (IP)-based services (iSCSI, NFS, SMB, asynchronous replication, synchronous replication), IT administrators typically expect the storage system to handle failover and quickly overcome configuration issues. InfiniBox has innovated in this domain by using instant IP failover in the event of a connectivity problem, moving IP addresses to network interfaces that can provide the relevant services.

Instant IP failover applies to all failure scenarios, including both hardware (node failure, Ethernet port/network card failure) or software (service failure on a specific node). To minimize the impact on other services, InfiniBox moves the minimal number of IP addresses, so that IPs of a different service on that node, or IPs on other nodes, will not be moved.

InfiniBox also leverages Virtual MAC addresses (VMAC) and assigns each IP address to a VMAC. When IP addresses move, VMAC addresses also move with them. This eliminates the failover time, allowing the configuration change to happen on the switch without propagating the change to each host. It also helps avoid the gratuitous ARP problems and increases availability.

InfiniBox employs smart network monitoring (using ICMP ping over IPv6) to identify potential misconfigurations, such as accidentally blocking a storage network interface from accessing a VLAN that is used for data services. Each network configured in InfiniBox is constantly monitored, often giving storage administrators the answer to the question, “Why has this application lost access to the storage?” long before they ask the question themselves.

Hardware Architecture

InfiniBox is a software-defined storage system, leveraging COTS hardware. As part of its design, Infinidat Federal has invested in software to make the COTS hardware more reliable, more cost-effective, and simpler to administer and support. The most critical design principle is N+2: all components have at least triple redundancy to design for seven nines reliability and when coupled with InfiniBox software, 100% system availability.

The InfiniBox system comes pre-assembled in a rack, as shown here:

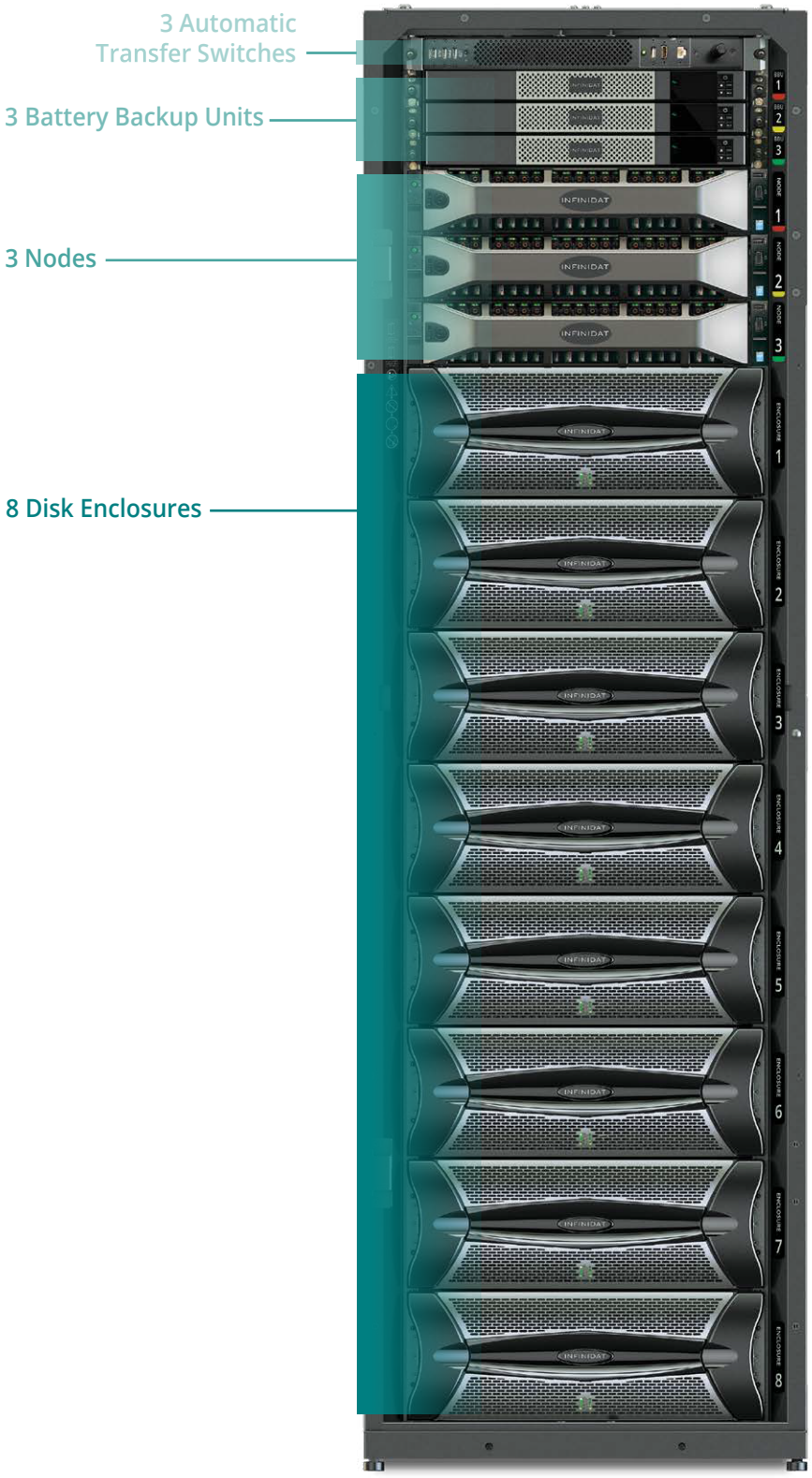


IMAGE 1 *InfiniBox Rack - front view*

NODES

The nodes are the storage controllers in the InfiniBox. The three fully redundant nodes work in an Active-Active-Active cluster, allowing I/Os to flow seamlessly through all three nodes. The nodes are directly interconnected with fast InfiniBand for direct access to memory using RDMA, which allows new writes to quickly replicate between nodes at the lowest possible latency.

A node failure is handled by the remaining two nodes taking over its responsibilities, resynching any part of the write cache that is no longer replicated to resume full data protection and maintaining operations non-disruptively. The N+2 node architecture also simplifies maintenance operations on a specific node (e.g., replacing a component) as the system still has two Active-Active nodes running and protecting the data.

Physical Connectivity

Front-end connectivity from the nodes to the customer's fabric:

- **Fiber Channel (FC)**—Eight ports per node, 24 ports in total. All ports are active, so each host sees multiple paths (at least one per node; two per node are recommended). Multi-pathing allows a port or HBA failure to only impact the individual path, and not impact the applications.
- **Ethernet (Eth) ports**—Up to six ports per node, 18 ports in total, offering either copper or optical connections, and supporting the iSCSI, NVMe/TCP, NFS, SMB, synchronous replication and asynchronous replication protocols. These ports support smart IP failover to prevent any physical failure from impacting system accessibility.

Internally, the nodes also provide the redundant back-end connectivity:

- **InfiniBand (IB) ports**—Used for the cluster interconnect. Any InfiniBand failure causing a node disconnection from another node will cause these two nodes to communicate through the third node. If a node becomes disconnected from both remaining nodes, it will be gracefully removed from the cluster until the disconnection is resolved.
- **SAS ports—**Connecting the nodes to all the disk enclosures. Any SAS failure leading to loss of access of a specific node to some of the disks will use the InfiniBand to access these disks remotely through another node.

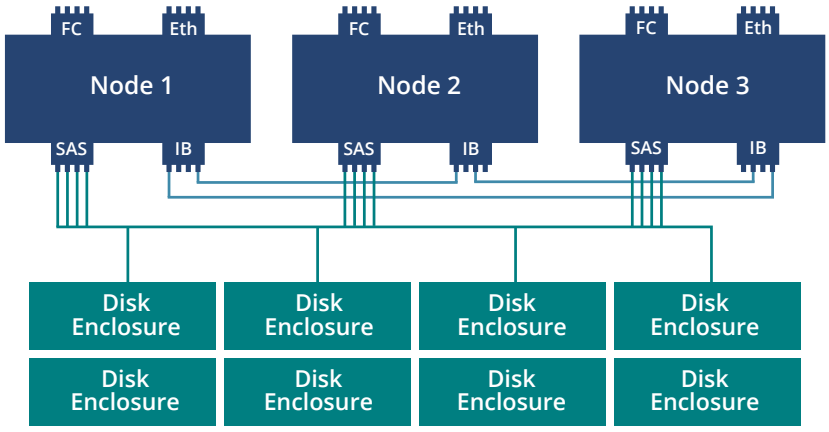


IMAGE 2 *InfiniBox front-end and back-end connectivity*

The nodes have redundant power supplies, and are fed from different Battery Backup Units (BBUs), which in turn are fed through multiple power inlets to allow non-disruptive operation though power issues.

Automatic Transfer Switches

The Automatic Transfer Switches (ATS) control the power feeds into the BBUs, and make sure the battery will always get input current, even in the event of a power outage in one of the power sources. The ATS can instantly switch between two power sources when one of them fails, keeping the power to the BBU uninterrupted.

Battery Backup Units

The BBUs maintain power to the InfiniBox nodes through short power outages (e.g. until generators are fully active), avoiding the need to shut down the system. They also provide power to properly remove (destage) data from the DRAM cache in the event of a longer power outage, allowing InfiniBox to always achieve proper shutdown procedures. The BBUs are monitored and each of them is automatically tested once a week, making sure their batteries are in order and ready to protect the system in case of a real power outage.

Conclusion

The unique InfiniBox architecture breaks the traditional compromises between reliability, performance, capacity, and total cost of ownership. The focus on software as the driving force for storage innovation now makes it possible to implement solutions that improve over time. For the first time, IT organizations can work within their limited budget while still enabling their business to execute new initiatives. With InfiniBox, organizations can more easily and affordably acquire, store, analyze, and protect their most critical corporate data to achieve competitive advantage.