Universal Storage

Innovation to Break Decades of Tradeoffs

FEBRUARY 2020

VAST

AN END TO DECADES OF STORAGE COMPLEXITY AND COMPROMISE

SUMMARY

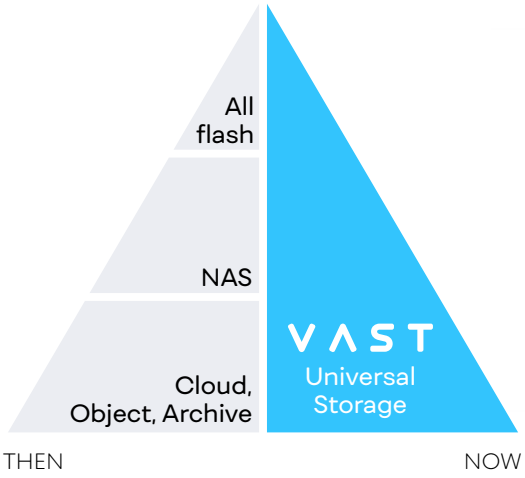
When it's possible to store all of your data in a single tier of storage that's fast enough for all of your demanding applications, large enough to manage all of your data, and affordable enough such that the economic arguments for HDDs no longer apply, everything is simple. When exabytes of data are available in real time, new insights become possible.

THE VAST DATA ORIGIN STORY

Building on their experience of ushering in the age of flash for primary, high-performance applications –VAST Data's founders sought to address the challenges found across the capacity tier of storage, to unleash access to vast data sets and bring an end to the hard disk era.

With new concepts that make it possible to break decades of long-standing storage compromises, the VAST team has reinvented the storage experience and made it possible to now build data centers entirely from flash. These new concepts are only possible beginning in 2018, as a new collection of technologies enable an entirely new storage architecture.

Our goal is simple:



Universal Storage Defined:

Tier-1 Performance.

Tier-5 Economics.

Exabyte-Scale Namespace.

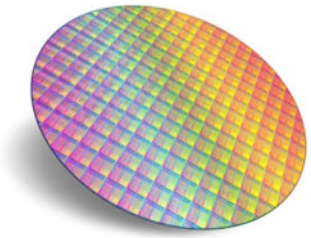
No More Tiers. No More Compromises.

STARTING FROM A NEW FOUNDATION

The VAST Data engineering team had the opportunity to rethink how storage could be built by inventing on a collection of technologies that weren't commercially available until 2018. More than a retrofit to classic storage architectures, these technologies are used in different and counter-intuitive ways to result in something altogether different.

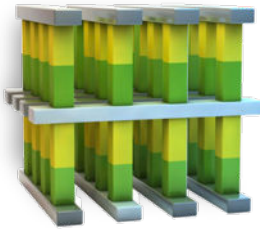
NVMe-OVER-FABRICS

Data-center-scale storage protocol that enables remote NVMe devices to be accessed with direct attached performance.



QLC FLASH

A new flash architecture that costs dramatically less than enterprise flash while delivering enterprise levels of performance.



3D XPOINT

Persistent, NVMe memory that can be used to reliably buffer perfect writes to QLC and create large, global metadata structures to enable added efficiency.

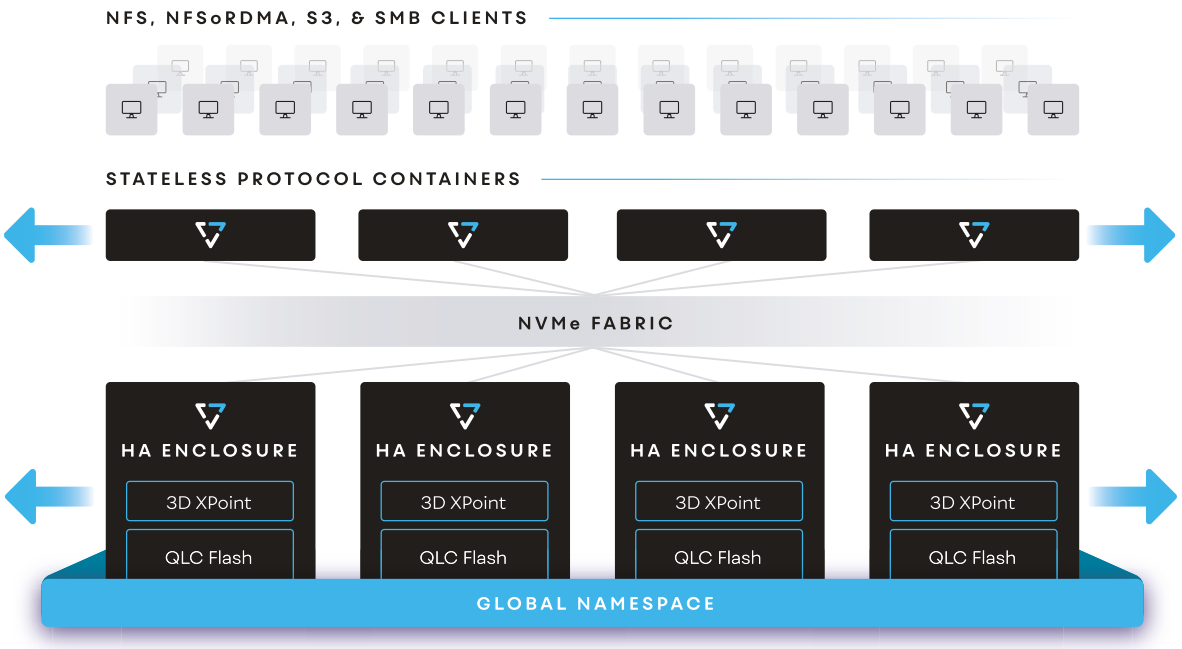
VAST DATA HAS REINVENTED SCALABLE STORAGE AND BROKEN ALL OF THE TRADEOFFS WITH ITS DASE ARCHITECURE

DISAGGREGATED, SHARED-EVERYTHING

VAST's DASE storage architecture breaks from the idea that scalable storage needs to be built from shared-nothing clusters. When servers are disaggregated from storage, everything is better:

- Servers are stateless, and failures of any server never require data reconstruction across a network
- Servers are loosely coupled in a scalable cluster – each operating independently while all accessing a shared, global namespace, thereby eliminating cluster cross-talk and enabling virtually limitless scale
- New, global algorithms are implemented to achieve game-changing levels of efficiency and resilience

THE VAST CLUSTER ARCHITECTURE



VAST SERVERS

A cluster can be built with 2-10,000 stateless servers. Servers can be collocated with applications as containers and made to auto-scale with application demand.

QLC FLASH

A scalable, shared-everything cluster can be built by connecting every server and device in the cluster over commodity data center networks (Ethernet or IB).

NVMe ENCLOSURES

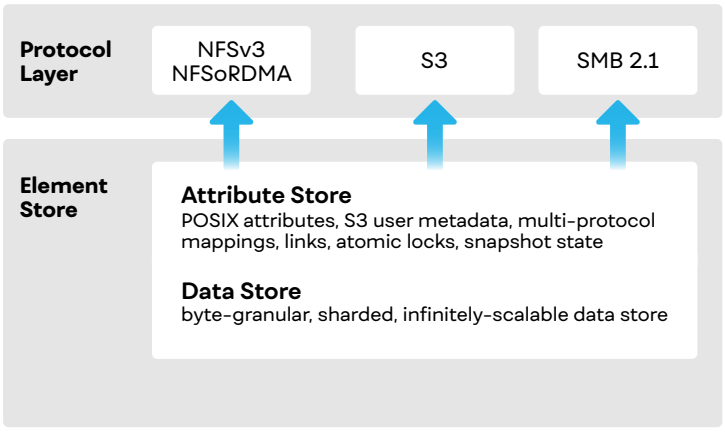
Highly-Available NVMe Enclosures manage over one usable PB per RU. Enclosures can be scaled independent of Servers and clusters can be built to manage exabytes.

DATABASE CONSISTENCY

With VAST's transactional file system, data is managed using append-on-write semantics. Data is never overwritten in place – new writes are committed to new write stripes via a layer of indirection created through pointers stored in non-volatile, storage class memory. No volatile caches. No batteries. Since every write is atomic, the file system can never be corrupted through loss of power or a system crash. Without journals, there is no need for legacy fsck tools.

MULTI-PROTOCOL SYSTEM

Each VAST protocol server has direct access to an exabyte-scale global namespace. Data access is fast from any protocol, data is unbounded.



INNOVATION TO DRIVE RADICAL STORAGE ECONOMICS

VAST GLOBAL QLC FLASH TRANSLATION

Legacy storage systems were not designed to work with large, multi-GB flash erase blocks and low-endurance drives. While QLC devices cost significantly less than traditional enterprise flash, it takes a new architecture concept to be able to use these devices and ensure over a decade of device longevity.

INDIRECT-ON-WRITE FILE SYSTEM

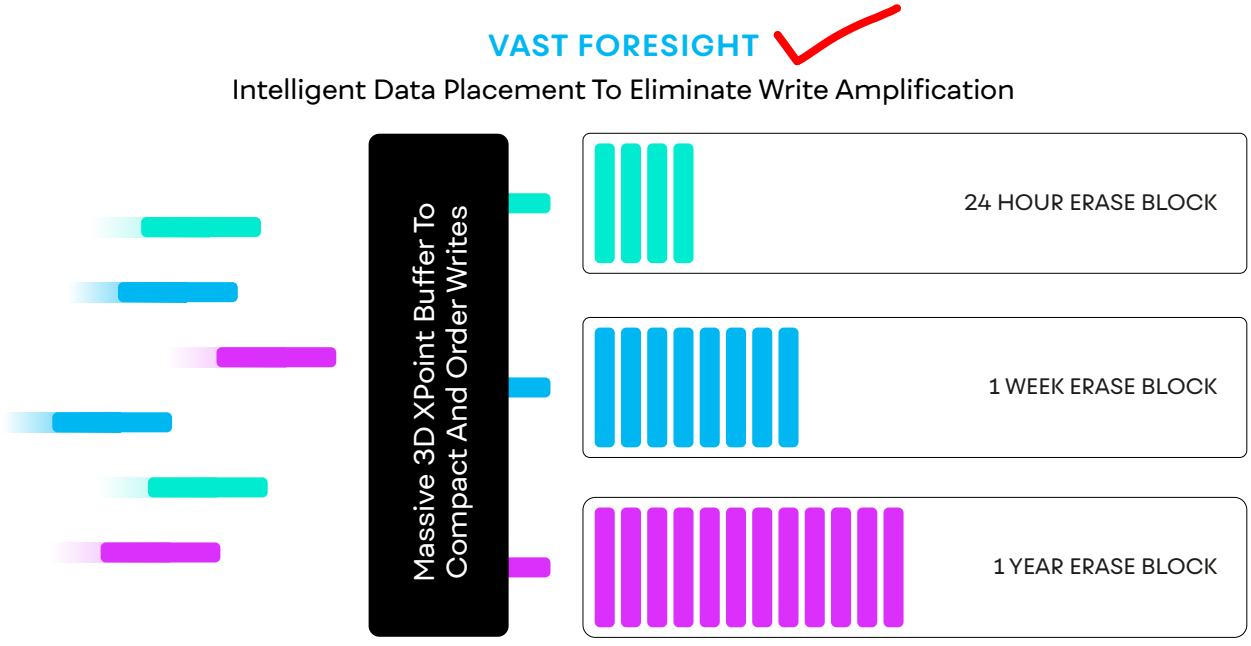
The DASE cluster architecture leverages TBs to PBs of Storage Class Memory (3D XPoint) to buffer writes into global and persistent fabric-attached memory. VAST's log-based file system creates a layer of indirection for new writes and appends. This indirection enables the system to only write in full, multi-GB QLC flash erase blocks - vastly reducing garbage collection, and enabling an unnaturally higher level of device endurance.

GLOBAL WEAR LEVELING AND WRITE AMORTIZATION

QLC flash wear leveling is done at the global level, and the system balances the needs of transactional workloads with the needs of archive data to deliver an averaged endurance that is greater than what QLC devices provide themselves.

PREDICTIVE DATA PLACEMENT

Moreover, VAST exploits additional context provided by the applications to predictively place data in QLC erase blocks that have a common life expectancy. By eliminating write amplification, VAST systems can be deployed for over a decade... all backed by by VAST Data's 10-Year Endurance Warrantee.



VAST GLOBAL ERASURE CODES

VAST has broken the cost vs. reliability tradeoff in data protection by developing a new, global approach that provides unprecedented efficiency and resilience.

SHARED-EVERYTHING ERASURE-ENCODED WRITES

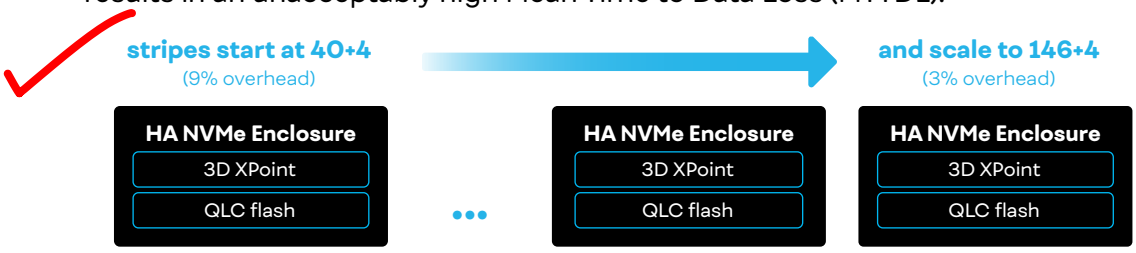
Shared-nothing storage nodes have CPUs responsible for a limited number of storage devices in a shared-nothing cluster. The VAST DASE architecture, on the other hand, enables any CPU to directly write to 10s to 1,000s of drives simultaneously, making it possible for the system to create very wide write stripes without imposing any cluster cross–talk or erasure code coordination.

A MASSIVE, PERSISTENT WRITE BUFFER TO ENSURE WRITE PERFORMANCE

Before VAST, the only way to achieve the right balance of performance and write stripe efficiency would be to use large amounts of DRAM, which is both costly and complicates system architectures because of cache power management and coherency issues. The VAST architecture employs 3D XPoint in a fabric-attached, persistent write buffer to enable fast cluster write speeds while also giving the cluster the time to craft large, QLC-optimized write stripes without the need for expensive DRAM and without the need for cache-coherence or batteries.

A NEW ERASURE CODE TO BREAK THE COST/RESILIENCE TRADEOFF

At the core of nearly every scale-out storage system in the market today is an adaptation of a Reed-Solomon error correction. The basic data reconstruction principle of Reed-Solomon is simple: when a drive within a write stripe is lost, all of the other drives in the system must be read from in order to perform a recovery. When stripes grow very large, Reed Solomon makes large striping impossible because the added time to perform a device rebuild by reading very wide stripes results in an unacceptably high Mean Time to Data Loss (MTTDL).

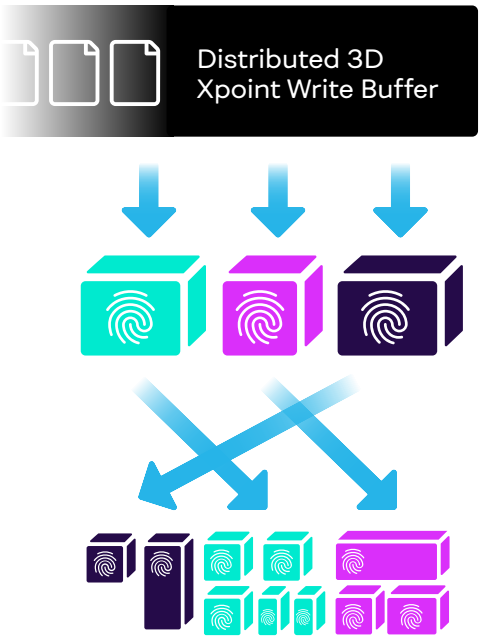


VAST's new Global Erasure Codes enable each data protection slice in a write stripe to act as a force-multiplier for cluster rebuild speed. VAST's declustered error correction can recover a failed device in a fraction of the time that it takes Reed-Solomon, As clusters grow, stripes are distributed across enclosures - the overhead of the system goes down to 3% while the system becomes more resilient. VAST's erasure codes are also fail-in-place, to enable instant recovery.

INNOVATION TO DRIVE RADICAL STORAGE ECONOMICS

VAST DATA SIMILARITY-BASED DATA REDUCTION

VAST has broken the tradeoff between data reduction efficiency and performance with its Similarity-Based Data Reduction technology. This breakthrough approach combines the global nature of deduplication with the fine-grained byte-granularity of pattern matching to achieve unprecedented levels of storage efficiency without compromising performance or endurance.



LOCAL DECODABILITY = SUB-MS READS

Data is first persisted to 3D XPoint so that write speeds are fast, while the system also has time to do aggressive data reduction in the background, all without wearing down flash.

Data is then chunked and fingerprinted with a variable-length hashing algorithm. Unlike conventional deduplication systems, VAST's hash is not intended to find exact block matches, rather it's engineered to create a signature of the data used determine the relative distance between other data that is already in the system.

The system then measures the 'distance' of new data with other data in the system to find relative similarity, the data is compressed against a cluster of similar blocks, providing the opportunity to find and compress patterns across files with byte-range granularity (1000s of times less sensitive to noise in data than classic deduplication methods).

Data in a similarity cluster is locally decodable, such that the system can simply retrieve the reference block and the deltas to perform a decompressed read within less than a millisecond. You don't need to decompress petabytes to perform a 4K read, as would be needed with legacy approaches to compressing multiple files.

When run against real-world datasets including unstructured data, backup data and even compressed data, VAST Data's Similarity-Based Data Reduction delivers on average anywhere between 2:1 to 20:1 storage-level data reduction.

UNLEASH APPLICATION INSIGHT AND AGILITY

VAST clusters are in production in some of the most data-intensive environments today. The goal of these deployments is to eliminate reliance on mechanical media and enable consolidation across workflows. The result is faster pipelines and workflows for business-critical applications.

VERTICAL APPLICATIONS

LIFE SCIENCES

FINANCIAL SERVICES

CONTENT DELIVERY

HPC

FILM ANIMATION

From AI to backup, from HPC to Cloud... never before has it been practical to consolidate a such a diverse set of applications on a single storage platform in a way where nothing is compromised on – performance, capacity, resilience, cost – get the best of all worlds for data intensive applications.

BIG DATA

LOG STORAGE

MACHINE & DEEP LEARNING

ENTERPRISE BACKUP

THREE OPTIONS TO DEPLOY VAST

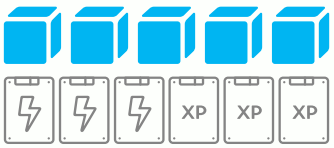
FULL APPLIANCE

VAST Provides a Turnkey Appliance

SOFTWARE DEFINED

Enclosures & Container SW

SOFTWARE-ONLY



Run Vast SW on Certified QLC HW

SYSTEM SPECIFICATIONS

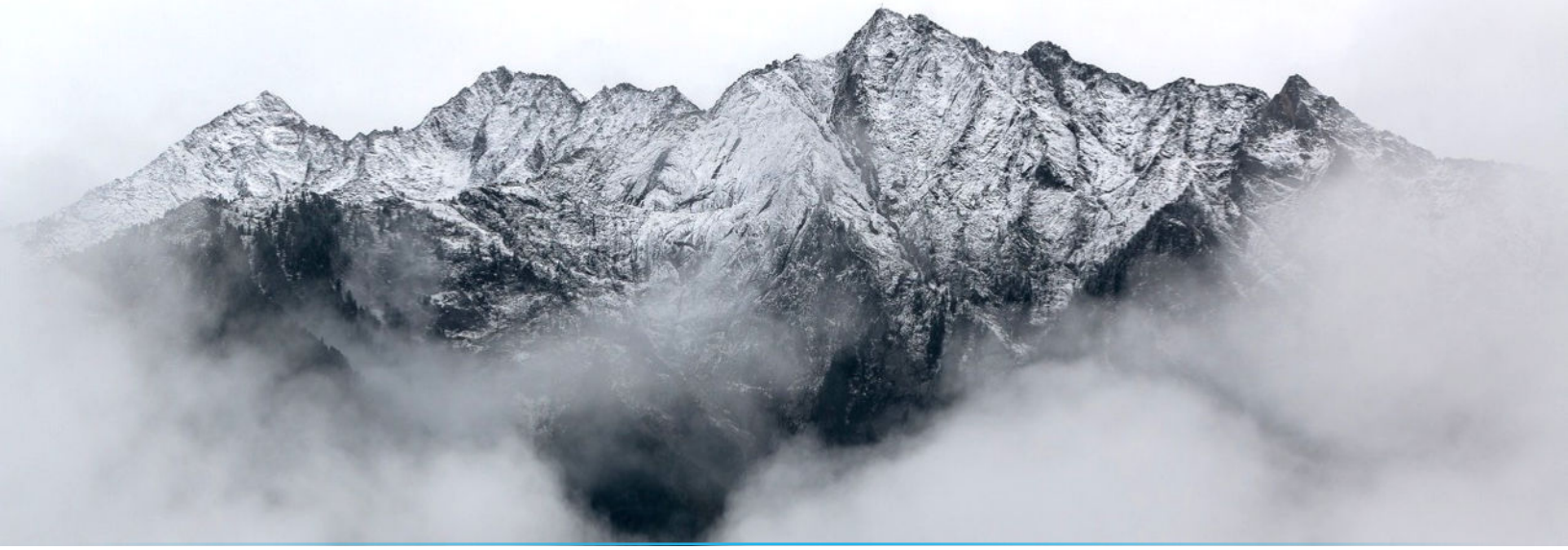
VAST DF-5615
Active/Active NVMe Enclosure

I/O Modules	2 x Active/Active IO Modules
I/O Connectivity	4 x 100Gb Ethernet or 4 x 100Gb InfiniBand
Management (optional)	4 x 1GbE
NVMe flash Storage	44 x 15.36TB QLC flash
NVMe Persistent Memory	12 x 1.5TB U.2 Devices
Dimensions (without cable mgmt.)	2U Rackmount H: 3.2", W: 17.6", D: 37.4"
Weight	85 lbs.
Power Supplies	4 x 1500W
Power Consumption	1200W Avg / 1450W Max
Maximum Scale	Up to 1,000 Enclosures

VAST Quad Server Chassis

Servers	4 x Stateless VAST Servers
I/O Connectivity	8 x 50 Gb Ethernet 4 x 100 Gb InfiniBand
Management (optional)	4 x 1GbE
Physical CPU Cores	80 x 2.4 GHz
Memory	32 x 32GB 2400 MHz RDIMM
Dimensions	2U Rackmount H: 3.42", W: 17.24", D: 28.86"
Weight	78 lbs.
Power Supplies	2 x 1600W
Power Consumption	750W Avg / 900W Max
Maximum Scale	Up to 10,000 VAST Servers

BREAK ALL THE TRADEOFFS



CLUSTER MANAGEMENT

GUI
CLI
REST (SWAGGER)

VAST CALL HOME

VAST's Cloud-Based Remote Management Service

VAST MULTI-PROTOCOL ACCESS

NFS v3, NFSv3 over RDMA, S3, SMB 2.1, Docker CSI, Containers

VAST DATA SERVICES

Exabyte-Scale Namespace	Similarity-Based Data Reduction	Continuous Snapshots
Distributed File Locking	Multi-Tenancy	Quotas
LDAP	POSIX ACLs	S3 Fan-Out

VAST INFRASTRUCTURE SERVICES

Transactional Storage System	Rapid Rebuild Error Correction
Distributed, Persistent Metadata	Tenant-Based Container Pooling
Global QLC Flash Translation	Fail-in-Place Architecture
At-Rest Encryption	Replication to Object Storage