VTIER Storage

Defined Storage for the 21st Century

Vikas Rana

Founder

Table of Contents

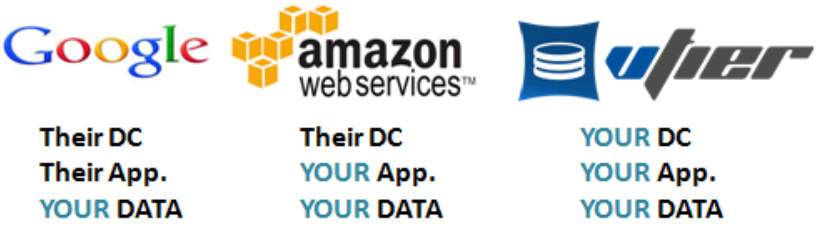
1. Executive Summary - The Need for Massively Scalable Storage

The ubiquity of the Internet has radically transformed the IT landscape. Every Internet user queries at least one search engine and checks email daily, often using multiple accounts. Users upload photos to online albums; connect with colleagues and friends over social networks, author product, pore over travel and restaurant reviews, post videos, and share personal experiences using multimedia rich content. Who is driving the worldwide explosion of data? Everyone – from businesses, to consumers as well as devices and machines. We are all contributing to the massive explosion of user-generated information.

Leading social and e-commerce web sites like Facebook, eBay, Yahoo or Netflix were designed to perform at Cloud scale. In contrast, many leading IT vendors did not design their solutions to handle this volume of data. Most traditional IT vendors, including leading providers of storage, have found it difficult to match the pace of growth and innovation demanded by the Internet and Cloud. Few have had the luxury to start over with a "clean sheet of paper" design. As has happened many times in the history of technology, disruptive innovation has come from newer, more agile players – emerging companies not encumbered by a portfolio of legacy products and technologies. These companies have been successful in inventing new classes of products, providing solutions that are much larger in scope and that offer improved functionality based on a fundamental rethinking of core technology principles. Such breakthroughs invariably improve IT performance as well as economics, enabling a large, mainstream market to deploy a class of solutions that had previously only been affordable for a few, highend customers.

To solve data and storage challenges associated with new online services, leading Internet and e-commerce companies had to invent new platforms rather than rely on solutions from traditional storage and infrastructure vendors. Internet innovators viewed the last generation of IT solutions as being constrained in many ways. They offered limited storage capacity and scalability, could not provide multi-site data services and could not achieve Cloud scale capacity or performance in a cost-effective way. Today, for example, the most scalable commercial NAS solutions, offers, at best, a maximum of 20 petabytes of raw storage, a capacity that is inadequate to handle large data services such as those required by an online photo sharing site needing to store at least four times this size. Traditional solutions are simply too complex, cumbersome and costly to support applications at Cloud scale.

In the absence of readily available, cost-effective, massively scalable commercial storage products able to support hundreds of millions of users and hundreds of petabytes of data, innovative companies in need of such capacity were forced to design their own storage systems. Using internal engineering and R&D teams and the insights of leading university computer scientists, these companies developed their own solutions based on open source software. Companies like Google, Facebook and Amazon succeeded in providing game-changing platforms, embodying radically new approaches to their internal IT systems and operations. These fundamentally innovative platforms enabled these companies to lead the emerging Internet commerce and Cloud revolutions. In 2011, VTIER, saw an opportunity to develop a commercial datacenter-grade product based on many of the insights, concepts and new computing models developed by the top Internet and Cloud companies. Figure 1 compares VTIER's approach to that of two leading Internet companies in regard to data center, application and data ownership.



The VTIER unified scalable storage solution brings the technological innovations developed to support leading Internet Commerce and Cloud Service Providers to the enterprise.

A few years ago, a wave of innovation and invention prompted the development of a number of similar solutions. These similarities are manifestations of tremendous market developments commonly referred to as the consumerization of IT. A decade ago, CIOs looked for inspiration, to the technologies and infrastructure deployed by large telco and bank data centers; now it is the practices and technologies developed and used by the big Internet players that drive IT innovation.

Beginning in 2011, a team of engineers at VTIER designed, from the ground-up, a unique storage software approach: A technology that is completely hardware agnostic, able to store exabytes of data with a very high durability and with multiple flexible data access methods, and can be deployed and managed at a very efficient cost.

This is what the industry today calls Software-Defined Storage.

This white paper describes the philosophy, architecture and design choices underlying VTIER Storage.

2. Requirements for a New Generation of Exabyte-Scale Storage Solutions

Today's enterprise computing environment suffers from many of the same challenges that confronted Internet and Cloud leaders. Applications that produce petabyte and exabyte scale data, once thought exotic, are increasingly common. In areas as diverse as media and entertainment, oil and gas, biotechnology, financial services and high performance computing, the amount of data that must be managed is outstripping the capability of existing technologies. When attempts are made to use existing technology to solve petabyte scale problems, customers usually find the cost to be prohibitive.

In developing new storage solutions for the petabyte and exabyte scale era of computing, several requirements must be met to ensure that tomorrow's massive data stores will enjoy the same or better levels of access, protection and usability as today's enterprise repositories. These requirements include:

- **Storage at Exascale**: A next generation storage system must be capable of supporting users and data at Cloud scale. This requires unlimited storage, scaling to billions or trillions of objects, files and other entities without degrading performance.
- **High Availability**: Data must be available to users continuously, without interruption, even when the storage system is performing rebuild or data recovery operations, or when configurations are undergoing maintenance or upgrades.
- **Data Durability:** Data must be stored exactly as it was intended and must, upon access, prove to be identical to the data that was stored. The system must protect against the possibility of data corruption for any and all data stored in the system. Automated self-healing must guarantee protection against both software and hardware failures, servers or disks, data center disasters, loss of power and any other failure model.
- **High Performance Access:** Data access operations must exhibit high throughput, high IOPS and low latency to ensure that the system can support mission-critical applications accessed by millions of users.
- **Universal Data Access:** Storage systems must be transparent to the applications that access them. Applications must be able to store and retrieve data using their existing file system protocols without requiring changes to underlying applications. This means that storage systems must be able to interoperate with traditional file protocols such as NFS and CIFS as well as with newer environments, such as HTTP/REST and Hadoop HDFS.
- **Geo-aware File Storage:** The system must be able to synchronize and replicate data efficiently across dispersed data centers and offer multiple access points via data propagation.
- **Simple Storage Management:** Storage systems should simplify and automate storage management tasks such as provisioning, replication, and backup and recovery.
- **Auto-Tiering:** Policy--based methods should ensure that data is written or moved to the storage tier that provides the right balance of storage cost and performance based on the lifecycle of each type of data stored on the system.
- **Cost-Effective Scalability**: Next generation storage should enable cost-effective scalability so that petabyte storage is affordable for mainstream IT buyers. TCO is also affected by automation capabilities and a hardware agnostic approach driven by a software-defined philosophy. At scale, intelligent power management also plays a key role in the global financial equation.
- **Storage Powered by Software:** Decoupling storage from client applications is essential for true scalability, availability and cost efficiency. The goal of this new approach is to allow full programmatic control in and by the software, and the building of a large multitenant storage pool from a farm of heterogeneous physical servers.

3. Limitations of Last-Generation Storage Technologies

Last-generation storage solutions, such as NAS and SAN, fail to meet the requirements identified in the last section of this document when operating at anywhere near petabyte scale. Operational constraints at web scale make traditional approaches to data management, resiliency, durability and data protection fundamentally inadequate.

As noted previously, leading Internet and Cloud companies such as Google and Amazon recognized early on that they needed to innovate in order to achieve the unprecedented scalability and high performance required for the support of a global user base. The table below outlines limitations inherent in last-generation approaches to storage that led Internet innovators to seek new solutions. Among these limitations, three are the most critical: 1) Scalability limits, 2) Inadequacy of RAID data protection, and 3) WAN and geographic distribution limitations.

	NAS	SAN	Next Gen Storage
Nature	File system	Block device	Unified
Fault tolerance	Low: must have all components available (LUN, partitions…)		High: Only needs an ID to locate data (independent of physical topology)
Logical entity	Network file system	Volume, LUN	Object, Bucket
Access Methods	Byte level via file path name	Block level (512Bytes, 4kB) via /dev	Multiple with Block, File and Object
Access Protocols	NFS, CIFS, pNFS, FTP	SCSI, iSCSI, FC, FCoE, IB	Object based, such as HTTP/REST, Amazon S3 and CDMI; Compatibility with traditional Block and File is a plus
Data Protection	RAID double parity and spares are not aligned with large--scale requirements; Limited and costly geo--redundancy solution at Block or File level		Replication, Erasure Coding, Fault tolerance across nodes
Distance tolerance	Medium (can be extended with WAN Acc./Opt.) but traditional file sharing protocols were not designed for the Internet	Limited (local: DC, building, a few kilometers with channel extenders)	High (designed for the Internet)
Advantages	Flexible (NAS clients embedded in OS); Well adopted: IOPS for Scale-Out NAS, Bandwidth	Well deployed and adopted; IOPS, Bandwidth, Low latency	Very flexible with programmatic API and legacy compatibility; IOPS, Bandwidth; Close to the application; Geo redundancy: Hardware agnostic
Limitations	Division between two file sharing protocols (NFS and CIFS); Maximum number of files; Maximum file size; File services can't be used over the Internet; Georedundancy	Rigid; Disaster Recovery; Distance; Number and size of volumes; RAID doesn't scale; Georedundancy	New Data models
Use cases	Vertical IT/Industry; Generic file share; Office documents	Database, VM and applications with low latency requirements	Staas, Vertical IT/Industry, Unstructured content at scale
Cost	$$$	$$$$	$

NAS, SAN and Next Gen Characteristics and Limitations

4. Next Generation Technology

As the previous section has indicated, it is becoming increasingly difficult and cost prohibitive to support larger storage deployments using traditional NAS or SAN technology. Storage systems have already reached their physical limits using existing scale-up methods.

To deal with these limitations, leading universities and major Internet firms have introduced a number of concepts essential to building a very scalable and agile IT infrastructure. Underlying this work in large-scale distributed computing has been the industry's actual experience of frequent failure of CPU, disk or network components as clusters increase in size. This section describes a number of theoretical principles informing the design of the next generation of scalable storage systems. These principles are of fundamental importance to VTIER's core architecture, which is discussed in later sections of this paper.

GENESIS

To address these limitations and deliver a more scalable approach to enterprise storage, the IT industry has begun to consider new models employing *scale-out* and *shared-nothing* paradigms. In these models, data is distributed and managed together with its associated metadata as a single object – constituting a new logical entity.

SCALE-OUT AND SHARED-NOTHING

The most demanding applications require both significant computing power and highly scalable storage capacity. Such applications must harness the resources of hundreds or thousands of servers, where computational power is scaled horizontally, across many separate compute nodes, rather than vertically, where the computational power of each node is increased with additional internal resources.

A *scale-out model* is one in which many loosely coupled, independent components cooperate to deal with large amounts of data. Scale-out is a radically different IT concept, based on distributed computing. Instead of using clusters of large, proprietary systems, organizations use commodity (Commodity Off The Shelf or "COTS") servers using intelligent software to manage their integration. This approach delivers functionality that is superior to older proprietary systems and, ultimately, improves performance, availability and scalability, as well as reducing hardware costs far beyond what proprietary systems can achieve.

The scale-out approach is related to a recently introduced computing architecture called *shared-nothing*, where each server brings its own resources to the cluster, and where the only shared resource is the network connecting the servers themselves. Clusters are built using self-contained peers, connected over a relatively high-speed network. Each of these nodes uses standards components: x86 CPUs, internal disk or SSD drives, an Ethernet network, IP protocol and a Linux OS.

NEW IT SOLUTIONS CHARACTERISTICS

The following table compares new and legacy IT models.

	Legacy solutions	Next Gen. solutions
Infrastructure	Virtualized	IaaS and software defined
	Dedicated (1 tenant)	Elastic, multi-tenant and shared
	Enterprise-grade	Carrier-grade
	Proprietary hardware	Commodity hardware
Application architecture	Centralized	Distributed
	Stateful	Stateless
	Synchronous	Asynchronous
	Scale-up	Scale-out
Configuration management	Manual	Automated
	Layer specific	Converged
Operational owner	IT	DevOps
Preferred management tools	On--premise	SaaS
Access methods	Local only (block or file based)	Ubiquitous, shared and global with object (HTTP) and file
Data protection	RAID and limited geo--copy	Data replication, erasure coding (geo distributed)

OBJECT STORAGE FOR MASSIVE SCALE

The need for massively scalable storage requires a new approach. Object storage was introduced to enable unlimited storage capacity, high performance, linear levels of service, and the capability to share content remotely and transparently across geographies. The leading Internet and Cloud providers developed their own object model based on their new requirements for global scalability. While the enterprise is beginning to adopt object storage, the first generation of object storage deployments were for public Cloud services.

Object storage has evolved from OSD (Object-based Storage Devices) and CAS (Content Addressable Storage) to wider use cases and interfaces. It is now defined by these key characteristics:

- An object is a self-describing opaque entity that contains data and associated metadata.
- An object belongs to a single flat namespace. This simple namespace guarantees transparent scalability.
- An object is location-independent and does not utilize nested directories, file paths or other complex addressing schemes.
- Policies and user-defined metadata exist at the object-level or bucket-level.
- Object storage is, by nature, multi-tenant.
- Object storage provides vertical consistency, in that the model is simple and end-to-end, from the application to the object itself, without regard for volume size, number of objects, directory structure or file system layout.
- Object storage performance has predictable, linear response, and is not degraded by central authority control mechanisms or lookups.
- The object requires a specific HTTP API, often REST-based, to connect to the application and to deliver content.
- Object storage provides a self-service mode and offers provisioning and metering capabilities.

5. VTIER Unified Storage

As the example of top tier Cloud and Internet innovators indicates, IT models and approaches need to change fundamentally to satisfy emerging requirements for orders of magnitude of additional computational power and storage capacity. Several years ago, VTIER anticipated these requirements and understood that meeting them would require a paradigm shift, a fundamentally new design and a new class of IT storage solutions. Designed originally to deliver massively scalable consumer email platforms, VTIER developed a software-defined, large scale data storage platform able to store petabytes of data, and billions of files while providing high levels of data durability without compromising performance.

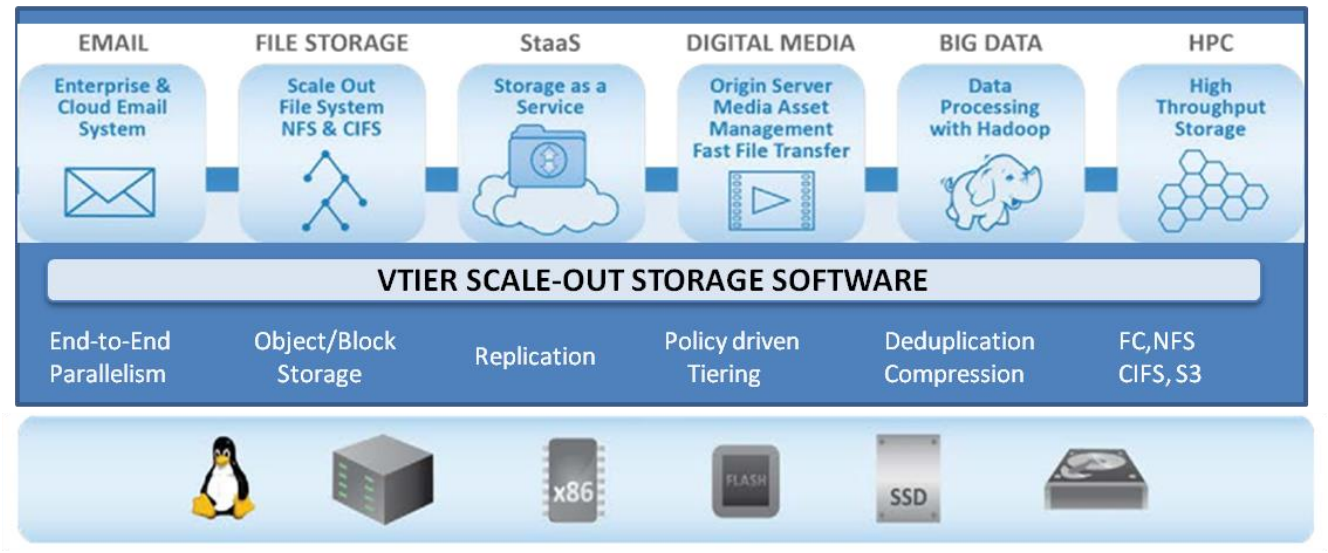
- **Store** an unlimited amount data;
- **Protect** data locally and globally to maximize its durability;
- **Serve** data to applications using multiple flexible access methods, each capable of delivering satisfying performance.
- **Compute** data, if needed, using the host's processing power within the storage cluster itself.

DEFINITION

VTIER is a web-scale, software storage solution. VTIER is based on a patented storage tiering technology with full scale-out file system support. It is built using a distributed, shared-nothing architecture with no single point-of-failure. Built-in tiering provides maximum flexibility for storage configuration and data movement, and ensures low latency and high performance.

VTIER is designed to support very large volumes of unstructured data and to sustain heavy traffic and heavy data workloads. VTIER is cost efficient to operate and delivers comprehensive data protection. The VTIER OS operates seamlessly on any commodity server hardware, turning generic x86 servers into a reliable, high performance storage platform. These commodity servers provide the storage media, and VTIER's software provides the storage provisioning and management, data protection, self-healing operations, high availability and automated-tiering.

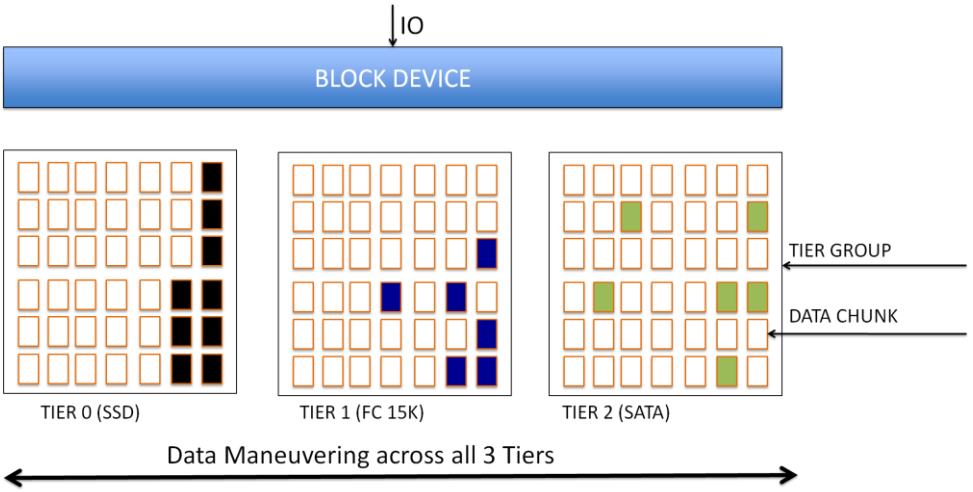
VTIER's architecture enables it to overcome traditional scalability limits, easily storing and managing petabytes and exabytes of data. VTIER's architecture supports virtually any application, including high performance computing, infrastructure for top-tier Cloud Service providers, massive compliance archives, storage of digital media, enterprise-class email and high volume storage of business files.



ARCHITECTURE AND COMPONENTS

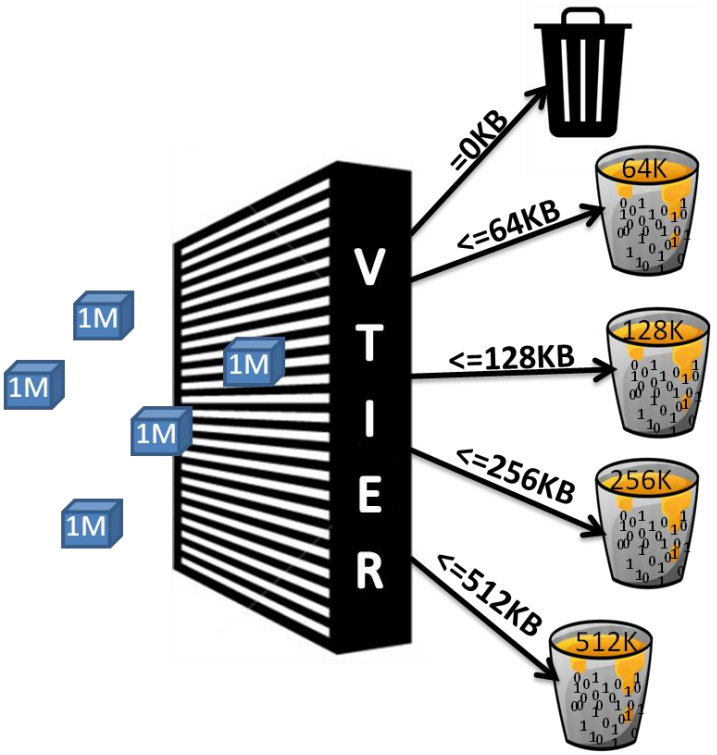
VTIER's distributed architecture achieves limitless scalability and high levels of availability and durability. Three principles drive the VTIER's distributed design:

- **Age/Access based Tiering**: VTIER OS automatically moves active data to high-performance storage tiers and inactive data to low-cost, high-capacity storage tiers. The result is higher performance, lower costs, and a denser footprint than conventional systems.



- **Analyze workload**: VTIER OS performs predictive analysis of IO streams and places data as per the defined IO profile. Random works automatically placed on higher tiers and serial IO streams are routed directly to lower tiers. This ensures efficient data placement.

• **Zero reclaims, Transparent compression and Deduplication:** VTIER OS watches for zero pages and reclaim them during storage tiering. Central to VTIER's unique product capabilities is the ability to compress and de-duplicate data that typically achieve ratios of 8:1, rising to 16:1 in some cases. Compressed data is placed into buckets. Four buckets are configured per engine. A block size of 1MB is scanned by VTIER engine and placed into its appropriate bucket after running through deuplication and compression engine. Zero blocks are discarded and added to freelist. Each bucket is btree based and designed to achieve concurrent updates and high throughput. Btree's resides in memory and SSD space to achieve maximum performance.



IO OPERATIONS

At the heart of the VTIER system, IO daemons, known as *engines*, are responsible for the persistence of data on physical media. Their role is to write the data passed to the node on the same machine, monitor physical storage and ensure durability. Each *engine* is local to one machine, managing local storage space, and communicating only with the storage node instances present on that same machine. There is no exchange between a node of a machine and the *engine* of another machine. Multiple *engine* run on the same physical machine

Each *engine* controls its own file system and its data containers built on storage nodes. These containers are, in fact, elementary storage units of the VTIER cluster that receive written objects directed to the *engines* from node requests initiated by any connector.

The use of a local file system on each local disk provides VTIER and the administrator the capability to use standard Linux commands to copy, migrate, repair and scrub disks if required. Containers used by the VTIER are large files grouping thousands of objects. As such, this design does not incur any performance impact and adds no overhead in terms of disk utilization.

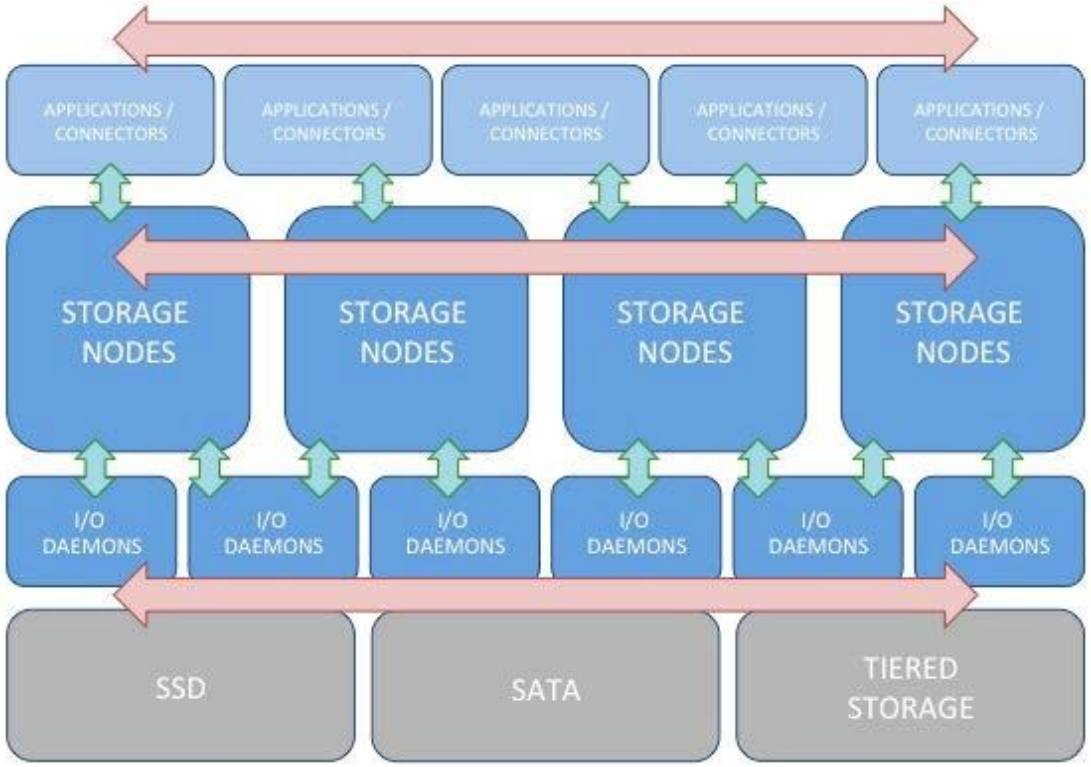


Figure 8: Parallelism from Connectors to Storage Nodes to Engines

6. VTIER feature list

Features	Benefits
Multi-Protocol Support (FC, iSCSI, CIFS, NFS, Object)	Support SAN, NAS and VTL in a single architecture to reduce hardware cost and share SSD Tier.
SSD Acceleration	Random IO Workload benefit from SSD Tier.
Read/Write Optimized layout	I/O profile detection ensures sequential data streamed directly to lower tier.
Transparent compression	Compresses SATA tier blocks when passes user-defined threshold. Compresses data 8x-16x
Disk pools	Multiple disk pools with separate IO threads for IO Parallelism.
Thin Provisioned storage	Every volume is Thin. All unused space is available to any application with no stranded space in any tier.
Performance	SSD tier is virtualized across multiple volumes. Enhance performance by adding additional SSDs.
Snapshot/Clones	Space efficient snapshots for backup, test, development and disaster recovery(DR)
Replication	Replicate volumes to a remote VTIER system. Bandwidth savings with Dedup and compression.
Encryption	Disk encryption. Encrypted remote replication.
VTL	VTL emulation for backup applications.
Performance knobs	User configurable settings for each volume to satisfy application performance requirements.
Management	AD integration.
Connectivity	LACP, 10gbe

7. Conclusion: A Storage Solution Operating at Exascale

VTIER provides an advanced, robust and massively scalable way of deploying and managing storage based on distributed models, insights and operation guidelines similar to those developed to support the world's largest and most successful Cloud and e-commerce companies. VTIER OS delivers a proven storage solution without any inherent limits. It enables companies to address their requirements for petascale or exascale storage with an easy to manage, cost-effective, high performing, and fully scalable software storage solution.

Now in its fourth generation, VTIER offers data center class functionality while overcoming the high cost, capacity and performance limitations of traditional storage solutions. Whatever the application or usage model, VTIER can deliver an interface customized to an organization's specific storage requirements. VTIER can store and exchange data in any combination of block, file or native object modes, and can provide users with seamless access to information, whether user data is stored in HTTP/web-based systems or in industry--standard file systems, such as NFS.

VTIER storage offers five unique benefits:

1. **Exascale:** Unlimited capacity and high performance.
2. **Multi-Geo:** Flexible topologies with complete disaster recovery, business continuity and multiple points of presence.
3. **Data Protection:** Data protection based on replication, erasure coding and intelligent tiering across nodes.
4. **Universal Data Access:** Comprehensive support for the broadest range of object APIs and file system interfaces.
5. **Ecosystem:** A software--defined storage solution that is hardware agnostic while also providing tested and proven OEM hardware reference platforms. VTIER's ecosystem incorporates a growing range of partners to deliver innovative storage solutions for a variety of markets and applications. VTIER's vision is guided by the following architectural goals:
    - **Comprehensive Unified Storage** with Block, File and Object interfaces, and new application storage APIs such as HDFS.
    - **Ubiquitous Access** means local and remote data points without any need for the user to know the location of the data.
    - **High Scalability** for both Performance and Capacity without inherent limits.
    - **Advanced Data Services** with real-time policy and quality of service, content indexing and other service features.
    - **Convergent IT Platform** with capabilities to run multiple applications within the cluster.