Virtuozzo

Whitepaper
virtuozzo.com

Virtuozzo Storage

An Overview of the Features and Benefits of Virtuozzo Software-Defined Storage

OCTOBER, 2017

© 2018 Virtuozzo. All rights reserved.

Table of Contents

© 2018 Virtuozzo. All rights reserved.

Introduction

To support growing demands for both high performance and high data availability, modern data centers need a fast, flexible storage solution. But this requirement often presents challenges. In addition to the difficulties involved in managing and maintaining storage, there are flexibility and cost issues. Redundant storage is typically either not flexible (as is the case with local RAID arrays) or too expensive for many organizations (as is the case with SAN storage).

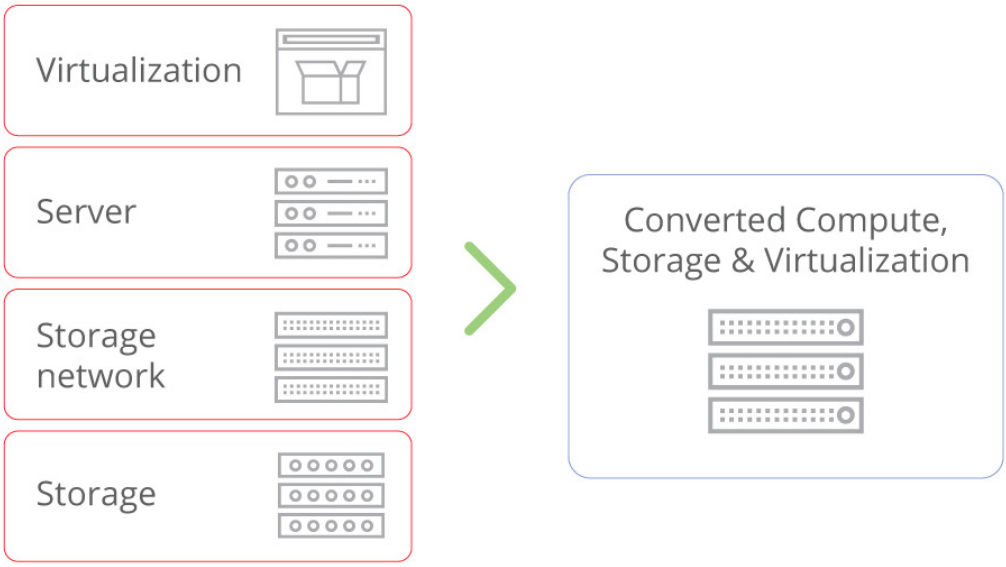
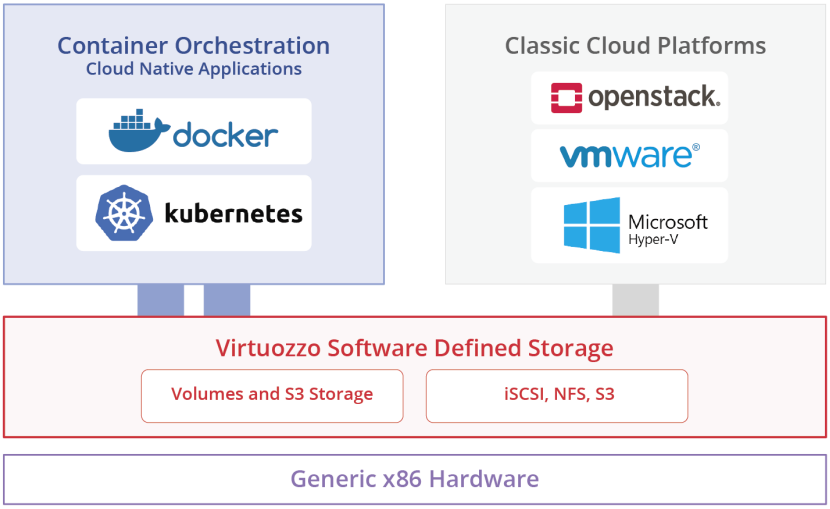


Figure 1. Hyperconverged infrastructure transforms datacenter in a set of standard, simple and scalable servers.

Virtuozzo Storage is designed to solve both problems. It can run on commodity hardware, and is designed to use existing compute servers. As a result, it is easy to set up and scales on demand.

Software-defined storage also provides flexible redundancy to handle any failures, and can completely eliminate the need for SAN or NAS solutions.

© 2018 Virtuozzo. All rights reserved.



In addition, Virtuozzo's software-defined solution supports multiple use cases, enabling block storage, storage for Kubernetes and Docker application containers, NFS, SMB, and S3 Object storage.

This white paper provides a description of the architecture, use cases and major features of Virtuozzo Storage.

Architecture Overview

Virtuozzo Storage is a highly available distributed storage system with built-in replication. It runs on top of commodity hardware, using locally attached hard drives to create storage clusters that span multiple machines. This approach ensures that data is always highly available, while eliminating disk fragmentation and optimizing overall cluster I/O performance.

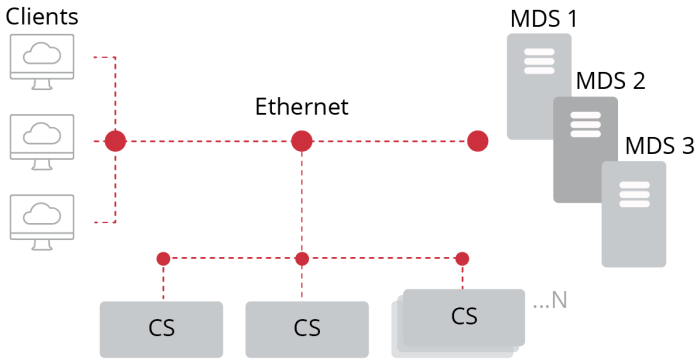


Figure 2. Key components of Virtuozzo Storage

© 2018 Virtuozzo. All rights reserved.

Figure 2 illustrates the logical structure and basic components of Virtuozzo Storage.

The key components are:

- **Chunk Servers (CS)**. All data in a Virtuozzo Storage cluster, including disk images of both containers (CTs) and virtual machines (VMs), is stored in the form of fixed-size chunks on chunk servers, which provide access to the data as needed. The cluster automatically replicates the data chunks and distributes them across the available chunk servers to provide high availability of the data at all times. As a result, to ensure this level of high availability, a Virtuozzo Storage cluster must have multiple chunk servers.
- **Metadata Servers (MDS)**. To keep track of data chunks and their replicas, the cluster stores metadata about them (such as their file names) on metadata servers. In addition to managing metadata, the MDSs control how files are split into chunks and where the chunks are stored. They also track versions of chunks, ensure that the cluster has enough replicas, and keep a global log of important events that happen in the cluster. As is the case with CSs, multiple MDSs are needed to provide high availability.
- **Clients.** Clients manipulate data stored in the cluster by sending different types of file requests, for example, to modify an existing file or create a new one. Clients access a storage cluster by communicating with the MDSs and CSs. Virtuozzo CT and VM clients can be run natively, i.e. directly from the storage cluster. You can also mount storage as a conventional file system (although Virtuozzo Storage is not POSIX-compliant). In addition, you can create and mount image files residing on storage as loop devices, and export them using iSCSI.

A recommended cluster setup typically consists of three to five MDS instances (allowing you to survive the loss of one or two of the MDSs, respectively) and three or more CSs to provide storage capacity.

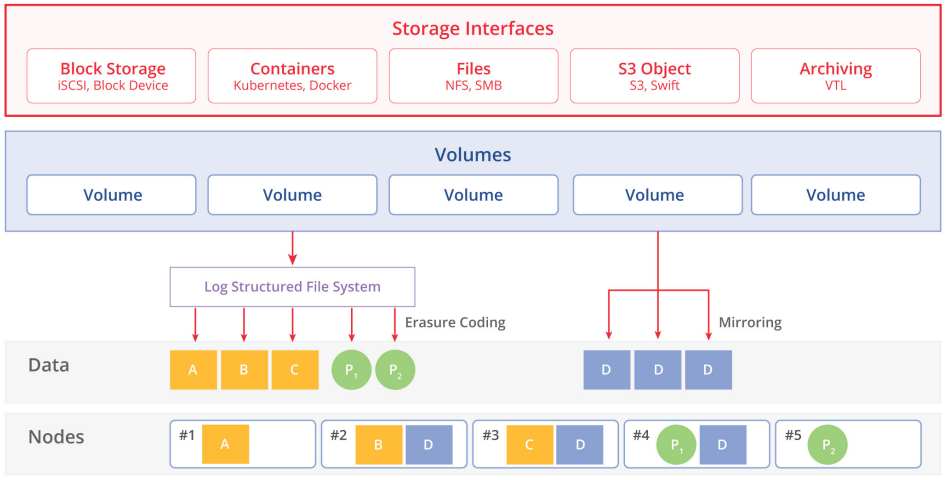


Figure 3. Key concepts of Virtuozzo Storage

The entire capacity is exposed as single pool to all storage clients. On top, aggregated capacity storage creates volumes. Each volume can be stored using mirroring (RF2 and RF3) or erasure coding.

© 2018 Virtuozzo. All rights reserved.

How Virtuozzo Storage Works

Mirroring

To ensure high availability, Virtuozzo Storage stores a number of replicas of each chunk across multiple physical servers. If it detects the unavailability of one or more chunk servers for a predefined amount of time, it automatically starts the replication process to ensure the availability of a sufficient number of copies of the data. Figures 4-6 illustrate the replication process in a cluster that consists of three CSs. With a redundancy factor of 2, each chunk should have two replicas. If CS #1 becomes unavailable for some reason, chunks green and yellow end up with only one replica. In this case, Virtuozzo Storage immediately detects that these chunks have fewer replicas than they should, and creates however many additional replicas are needed to bring the number up to the specified replication level.

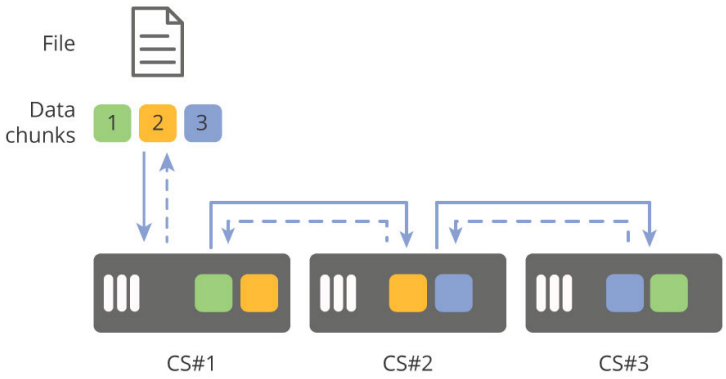


Figure 4. This healthy cluster has three CSs and two replicas of each of three data chunks.

Figure 5. This cluster, with one unavailable CS, has become degraded: Chunks green and yellow have only one replica.

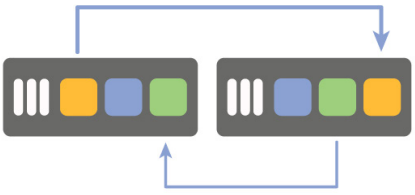


Figure 6. The cluster is healthy again after replication, with two replicas of every chunk.

© 2018 Virtuozzo. All rights reserved.

Virtuozzo Storage performs replication in a manner similar to the RAID process, but it has two key differences:

- Replication is much faster with Virtuozzo Storage than with a typical local RAID 1/5/10 rebuild because Virtuozzo Storage is able to replicate in parallel, across all servers in the cluster. In contrast, typical local RAID is only able to copy data from a single hard drive to another single hard drive.
- The more CSs you have in a cluster, the less time it takes to replicate each chunk.

Replication performance is important because it minimizes the time that the system operates with a lower than normal redundancy level. To ensure the overall reliability of the storage solution, the replication time must be much shorter than the expected time between CS faults, because the briefer the period of reduced redundancy, the less likely you are to lose any data.

Factors affecting replication performance include:

- Number of available CSs, because replication is a process that runs in parallel, i.e. the more available replication sources and destinations you have, the faster it goes.
- Performance of the local disk system used for storing chunks.
- Network performance, as each chunk is read, it's transferred by the network to write an additional copy.
- Chunk distribution, because some CSs may have much more data to be replicated than others, and therefore may become overloaded during replication.
- I/O activity in the cluster under replication.

Performance measurements show that with Virtuozzo Storage, replication performance scales almost linearly as the number of servers in the cluster grows. Tests demonstrate that even with Virtuozzo Storage running on commodity hardware, its replication time is a fraction of what RAID requires, resulting in higher reliability.

Although Virtuozzo Storage was designed to operate efficiently when running even on modest hardware, it's able to make full use of all the resources of more powerful systems.

© 2018 Virtuozzo. All rights reserved.

Erasure Coding

Virtuozzo Storage features support for erasure coding. The primary benefit of erasure coding data protection is that, compared to mirroring, it reduces the overall storage capacity needed to protect the same amount of data, which helps you reduce costs.

With erasure coding, Virtuozzo Storage breaks the incoming data stream into fragments of a certain size, then splits each fragment into a certain number (M) of 1-megabyte pieces and creates a certain number (N) of parity pieces for redundancy. All pieces are distributed among M+N storage nodes, that is, one piece per node. On storage nodes, the pieces are stored in regular chunks, but these chunks are not replicated as redundancy has already been achieved. The cluster can survive the failure of any N storage nodes without data loss. The erasure coding approach is much more efficient in terms of physical storage utilization than replication, however some types of workloads can demonstrate performance loss.

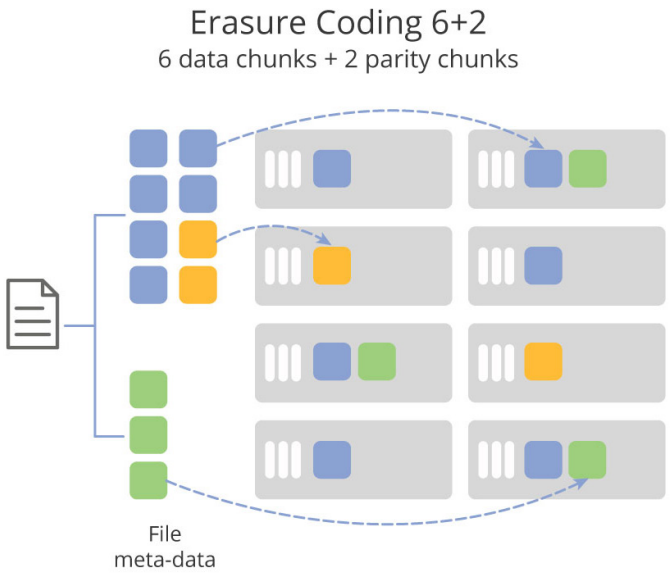


Figure 7. Erasure coding data distribution.

© 2018 Virtuozzo. All rights reserved.

The following table offers a comparison of using different redundancy schemes:

Redundancy	Minimum number of servers	How many nodes can fail simultaneously	Storage overhead	Raw space required to store 100Gb of data
Redundancy Factor (RF2)	2	1	100%	200Gb
RF3	3	2	200%	300Gb
EC 3+2	5	2	67	167GB
EC 5+2	7	2	40	140GB
EC 7+2	9	2	29	129GB
EC 17+3	20	3	18	118GB

Why Virtuozzo Storage

1. Complete, software-defined hyperconvergence; nodes can run all compute and storage services.
2. Flexible storage for all needs—multi-API support:
   - iSCSI—block storage for VMs
   - NFS, SMB/cifs—file storage
   - S3—object storage
   - Kubernetes and Docker volume storage for application containers
3. Optimized TCO across OPEX and CAPEX
   - No dedicated hardware required, low resource requirements
   - Use generic x86 servers with hardware failure handled on software level
   - Planned maintenance, no rush to replace a failed drive or server
   - Web UI and APIs that simplify and automate management procedures
   - Erasure coding/compression to reduce physical space needed
4. Designed for hot and cold data scenarios
   - 'Best-in-class' performance with SSD solutions
   - Optimal cost efficiency for cold data application

© 2018 Virtuozzo. All rights reserved.

Key Features & Benefits

INTERFACES

Block Storage	iSCSI interface for classic virtualization platforms like OpenStack, ESXi or Hyper-V.
File Storage	Storage supports NFS v3 and v4. SMB/CIFS interfaces are coming soon.
Container Storage	Persistent Volumes for Kubernetes and Docker containers.
S3 Object Storage	S3 object storage for modern applications.
Archiving Storage	VTL Virtual Tape Library interface to support existing backup and archiving systems.

DEPLOYMENT

Hyperconverged Solution	Virtuozzo Storage is designed to run side by side with the compute cluster. It needs only one CPU core and several Gb of RAM to operate. As a result, almost all CPU and RAM resources are available for compute technology running on top.
Generic x86 Hardware	There are no special hardware requirements: the solution can be built using commodity hardware (such as SATA or SAS drives and 1Gbit+ Ethernet) and will run on your current hardware stack. Although additional hardware resources, such as SSD drives or 10Gbit Ethernet, can significantly boost performance, they are not required to get the major benefits of Virtuozzo Storage, which provides a distributed storage solution with built-in replication and high availability.
Scale Out and Scale In	You can easily scale your storage capacity and performance by simply adding more nodes to the cluster or adding more disks to existing nodes. Nodes in a cluster can have different configuration depending on particular demand for compute, networking and storage.

© 2018 Virtuozzo. All rights reserved.

DATA PROTECTION

Flexible Redundancy	Virtuozzo Storage supports: • Mirroring with RF2 and RF3 • Erasure coding Encoding schemes are defined per volume.
Integrity Checks	Features scrubbing and check summing. Although Virtuozzo Storage itself is highly available and consistent, there is always a chance that local storage could become corrupted. To guard against such silent data corruptions, which could occur, for example, due to hardware faults, Virtuozzo Storage periodically performs data checks. When data is first written or modified, the storage solution calculates and remembers checksums for each piece of data. Then, in background mode, Virtuozzo Storage periodically inspects data from local storage and compares saved checksums with newly calculated ones to ensure that there is no corruption. If it detects any corrupted data chunks, it replaces them with replicas of known good copies.
Availability Domains	Availability Domains offer greater protection from hardware failures by allowing Virtuozzo Storage clusters to survive the failure of a node or rack. Availability domains are created based on the granularity at which failures are likely to occur.
S3 Geo Replication	Easily configure active-active Geo-Replication for S3 data between multiple datacenter locations with full data consistency and collision resolution.

EFFICIENCY

Erasure Coding	Erasure coding technology provides resilience and can save up to 2x more physical space. It optimizes space usage with the same level of software redundancy. Erasure coding in general requires a bit more CPU and has comparable performance with a mirroring approach.
Data Compression	Data can be compressed inline as it is written to the system, for a particular date (log files for example) significant improvements (20x compression rate) can be achieved without any noticeable performance impact.

© 2018 Virtuozzo. All rights reserved.

PERFORMANCE

Auto Tiering	Since fast storage costs more than slower storage, Virtuozzo Storage lets you define allocation policies for different types of local drives to deliver the best price/ performance ratio. For instance, you may choose to store frequently accessed data on fast, but relatively expensive SSD drives, while placing rarely needed backup data on inexpensive SATA drives.
Data Locality	Storage ensures that as much of a VM's data as possible is stored on the node where the VM is running. This negates the need for read I/O to go through the network. Keeping data local optimizes performance and minimizes network congestion.
Cluster Load Balancing	Virtuozzo Storage automatically balances workloads among all available CSs. In storing new chunks, Virtuozzo Storage takes into account both the amount of free disk space and the distribution of the I/O load, in order to ensure high performance. Additionally, when you add new CSs to a cluster, Virtuozzo Storage automatically balances existing data across all CSs.
SSD Caching	**SSD cache for a chunk server journal**—You can attach an SSD drive to a CS in the cluster and configure the drive to store a write CS journal—a step that can boost the performance of random write operations in the cluster by a factor of two or more. The journal can also maintain chunk checksums to improve storage reliability. **SSD read cache on the client**—You can attach an SSD drive to a client and configure the drive to store a local cache of frequently accessed data— a step that can increase overall cluster read operations performance by a factor of 10 or more.
Thin Replication	Virtuozzo Storage supports "thin" replication, in which only the changed parts of chunks are updated on recovery, rather than the whole chunks. This approach significantly boosts overall performance because it involves fewer reads, writes, and data transfers.

© 2018 Virtuozzo. All rights reserved.

MANAGEMENT

Web Management Control Panel	Simple and powerful web management interface covers all day-to-day administration routines.
API	Simple API for storage task automation.
Non-Disruptive Updates	Update storage components without any disruption for running services.
Monitoring & Alerting	Virtuozzo Storage has a built-in monitoring and alerting system. Additionally, you can integrate cluster monitoring to a 3rd party solution via the SNMP.

SECURITY

Data Encryption at Rest	Virtuozzo Storage can encrypt data stored on disks with the AES-256 standard to provide protection against lost or stolen data. Virtuozzo Storage stores disk encryption keys in the cluster's metadata (MDS). Encryption can be enabled or disabled only for the newly created chunk services (CS). Once tier encryption is enabled, you can decrypt disks (CSs) by manually releasing them from encrypted tiers.
Role Based Security	Support role based security with a flexible role definition and integration with LDAP or Active Directory services.

For More Information

www.virtuozzo.com
info@virtuozzo.com

© 2018 Virtuozzo. All rights reserved.