Gridstore Proves Storage Resiliency

Storage Grid Survives Node Failures

A DeepStorage Technology Validation Brief

Prepared for
Gridstore

About DeepStorage

DeepStorage, LLC. is dedicated to revealing the deeper truth about storage, networking and related data center technologies to help information technology professionals deliver superior services to their users and still get home at a reasonable hour.

DeepStorage Reports are based on our hands-on testing and over 30 years of experience making technology work in the real world.

Our philosophy of real world testing means we configure systems as we expect most customers will use them thereby avoiding "Lab Queen" configurations designed to maximize benchmark performance.

This report was sponsored by Gridstore. However, DeepStorage always retains final editorial control over our publications.

Introduction

Server virtualization has been a major driver of the adoption of SAN and NAS shared storage in the SME market. However, traditional storage arrays were designed long before virtualization, and they simply are not optimized for virtual workloads.

The virtual environment is highly dynamic, making it particularly difficult to address with traditional scale-up storage systems. Buyers of scale-up systems get a fixed amount of controller performance, and controller performance frequently becomes the limiting factor, especially when SSDs are added. With only two controllers, these systems also suffer a significant performance penalty when a controller fails or is taken offline.

Scale-out storage, which adds controller horsepower as well as capacity, can address both the expansion problem and make the system more resilient, but most scale-out systems don't scale down to fit the needs of most SMEs.

Gridstore's Storage Grid uses a unique host-centric virtual controller coupled with external storage nodes to deliver a scale-out storage system specifically for virtualization environments. Gridstore hired DeepStorage Labs to test the Storage Grid's resiliency by examining how several workloads were affected when a node was removed from a Storage Grid of three to four nodes

The Bottom Line

Gridstore's Storage Grid takes a unique approach to delivering storage, specifically to support virtual workloads. By using a host-based-plus-storage-node scale-out architecture, the Storage Grid offers a compelling combination of performance, management—including storage QoS—and resiliency.

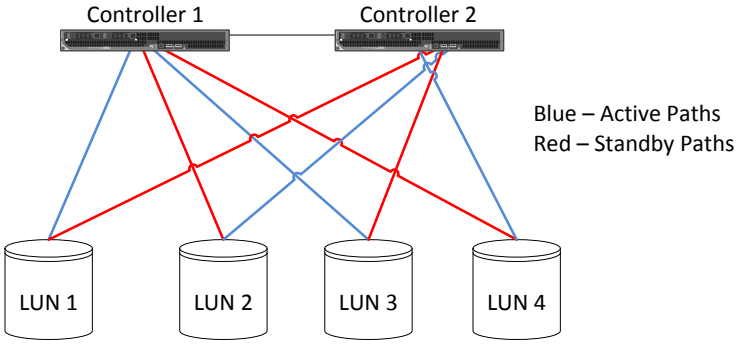
In order to explore just how resilient Gridstore's system really is, we tested the Gridstore grid using several different workloads and configurations. These tests, totaling more than 20 iterations, each simulated one or more workloads accessing the storage grid during a node failure. In addition to simulating node failures, we also directly caused a node failure using thermite to create the video now on YouTube

We learned:

- *None of the test sets resulted in a reported data error.*
- *When failing from a three-node to a four-node Storage Grid:*
- *Our 8KB OLTP workload retained 94% of its performance.*
- *Our 64KB sequential read workload saw no performance loss.*
- *The Storage Grid automatically detects when a node rejoins the grid and rebuilds the effected volumes.*

The Problem with Traditional Failover

Most dual-controller storage systems use a dual-active architecture, where each controller is responsible for some set of volumes while the other controller "owns" the other volumes. When a controller fails, the other controller takes over responsibility for the half of the volumes the failed controller was managing. This process of failover takes time, sometimes long enough to cause a disk error and crash an application.

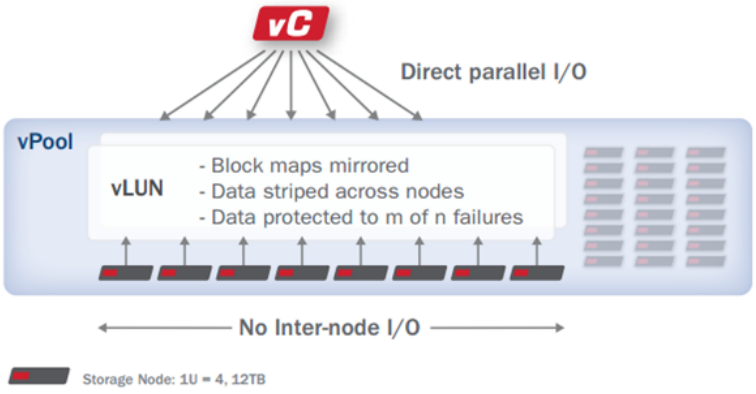


A Typical Dual-Active Storage System

If, as is all too frequently the case, each of the controllers is running at 60 or 70% of its capacity, when a controller fails, there will be not only a stutter as the controllers fail over but a significant loss of performance as the surviving controller struggles to do the work of two.

The Gridstore Architecture

Gridstore's Storage Grid uses a host-based virtual controller process to form a scale-out storage cluster across multiple storage nodes without requiring the high performance backend network scale-out clusters usually require. A virtual controller (vController) process runs as a Windows service in each Hyper-V host connecting the host to the storage nodes that make up the Storage Grid.



GridStore's Storage Grid

When an application writes data to one of the Storage Grid's virtual LUNs (vLUNs), the virtual controller writes that data directly to multiple storage nodes. By default, data for a vLUN is written across three storage nodes in a 2+1 RAID-5 layout. With larger grids, administrators can choose additional protection levels, and multiple vLUNs are spread across storage nodes to balance the load.

What We Did

To see just how resilient the Storage Grid, is we ran several workloads in Hyper-V virtual machines hosted on a Storage Grid and then removed one node.

At the Ranch

Our testing started at Rocking Horse Ranch, where we had a grid of four nodes. We connected three Hyper-V hosts to the storage grid, created a vLUN for each host, and then a Windows Server 2008 R2 virtual machine in each vLUN. The data was laid out across the storage nodes and disk drives as shown in the table below.

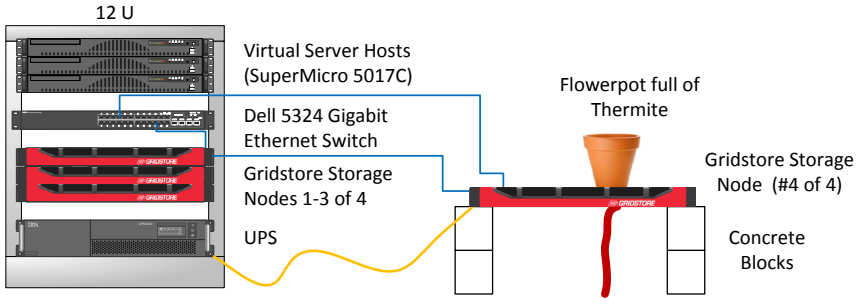
vLUN / Virtual Machine	Node A	Node B	Node C	Node D
vLUN1/VM-1	Disk1	Disk1	Disk1	
vLUN2/VM-2		Disk2	Disk2	Disk1
vLUN3/VM-3	Disk2	Disk3		Disk2

Virtual Machine Layout on the Storage Grid

On each virtual machine, we installed the VLC media player application, and we downloaded an MPEG-4 video file of a public-domain feature film to each VM. We started playing the movie files:

VM-1	The Charlie Chaplin Festival
VM-2	The Little Shop of Horrors (Roger Corman version)
VM-3	His Girl Friday (Cary Grant)

We used Hyper-V manager to open a console window into all three VMs from our management station and used Camtasia Studio on the management station to record its screen. While all three videos were playing, we destroyed storage node D with thermite.



Configuration for the Video Demo

We then reviewed the Camtasia recording and were not able to detect any frame loss, pixelization, or other artifact. The GridBusters video documents this process better than we could here.

Back in the Lab

We returned to the lab with a Storage Grid reduced in number to three. We then upgraded those three nodes from Gridstore's C2100 capacity nodes to the H2100 high-performance hybrid node with the addition of PCIe flash cards and 10Gbps NICs. We connected the virtual server hosts to the grid through our Brocade 8000 10Gbps Ethernet switch.

To get a more detailed view of how the Storage Grid behaved when a node failed, we used the latest 1.1 version of the venerable IOmeter benchmark. In addition to a 64KB sequential read workload, which we thought would be a pretty good approximation of the video player, we also used an 8KB OLTP-like workload to see how a database engine might be affected when a storage node failed.

Since we were interested in the failure behavior of the Storage Grid, not its ultimate performance, we ran both workloads with a queue depth of 16, a level that generated significant traffic but left the grid with a bit of performance headroom.

We ran each workload for a total of 30 minutes, disconnecting one node of the storage grid approximately fifteen minutes into the test. While disabling a storage node with thermite was fun, and dramatic, that technique does have a couple of disadvantages for more formal testing, not the least of which is a lack of repeatability. In the lab, we removed nodes from the grid by disabling the ports they were connected to through the Brocade 8000's web interface.

We had IOmeter record the system's performance once a second, giving us 1,800 data-points. To compare the system's performance in the healthy and degraded states, we averaged the IOPS, throughput, and latency for the period from the start of the test up to 15 seconds before the storage node was disconnected and the period from 15 seconds after the node was disconnected.

What We Found

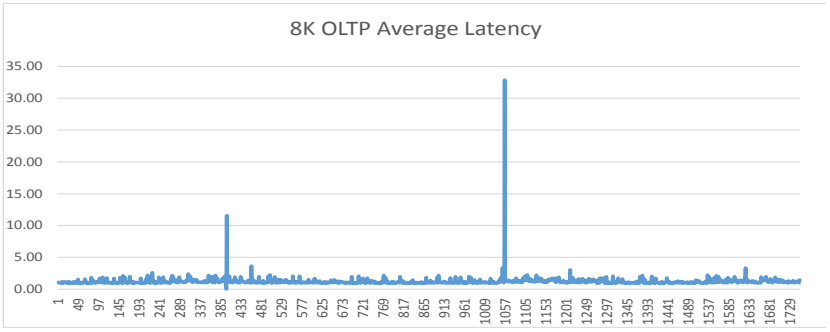
The first thing we found was that, in more than 20 test runs with various workloads removing a node from the grid, we never saw a disk error or timeout. While applications could see a latency spike, the vControllers always reconfigured the grid fast enough to prevent the kind of error that would crash an application or corrupt a database.

Workload	Grid State	Errors	IOPS	Throughput (Mbps)	Average Latency (ms)	Maximum Latency (ms)
64KB Sequential Read	Healthy	0	6926	432	2.63	237
64KB Sequential Read	Degraded	0	6921	432	2.65	243
8KB OLTP	Healthy	0	14432	112	1.15	45.0
8KB OLTP	Degraded	0	13652	101	1.21	38.4

IOmeter results

Almost as importantly, the performance of our VMs was almost unaffected. The 8KB OLTP workload ran 92% as fast on the two-node grid as it had on the three-node, and the 64KB sequential read workload's performance was essentially unaffected.

The graph below shows the average latency over the duration of our 8KB OLTP workload testing. Note that there is a brief peak when the node is removed from the Storage Grid but that performance has returned to normal within seconds.



Average latency of 8KB OLTP workload. Peak to 33ms as node disconnected at 1056s.

Conclusions

Our testing showed that Gridstore has managed to deliver on the resiliency long promised by scale-out architectures. By using the host's virtual controller process to manage access to the storage nodes, Gridstore has simplified the failover process when a node fails.

In our testing over more than 20 attempts, we were unable to cause an application error by removing a node from the grid. While our workloads were modest compared to the maximum capacity of the system, as an average workload should be, when degraded to a two-node grid, the system delivered almost the same performance as when running with three nodes.

The Storage Grid also includes features and functionality specifically designed to support virtual workloads, including demultiplexing requests from multiple VMs, reversing the dreaded I/O blender effect that fools traditional arrays into treating sequential I/O requests from multiple VMs as random I/O. It also supports storage quality of service (QoS), which limits how much one virtual machine, the so-called noisy neighbor, can consume storage performance to the detriment of the other VMs.

All in all, Gridstore's grid is an attractive alternative to more traditional storage systems for Hyper-V hosting.

The Test Environment

We used three SuperMicro 5017C servers to host our test workloads under Hyper-V. Each server ran a full installation, with GUI, of Windows Server 2012R2. Each has one Intel Xeon E3-1230 processor, 16GB RAM, dual Intel Gigabit Ethernet ports, and an Emulex OCe14002 10Gbps CNA.

The Hyper-V hosts were connected to the four nodes of the storage grid for the video testing through a Dell PowerConnect 5324 switch and to the three surviving nodes through a Brocade 8000 10Gbps switch.

IOmeter Access Specifications

Workload	Transfer Request Size	Sequential/ Random	Burst Frequency	Burst Length	Read/ Write	Alignment
64KB Sequntial Read	64K	100/0	0	1	100/0	64KB
8KB OLTP	8KB	10/90	0	1	60/40	8KB

Hyper-V Virtual Machine:

1 vCPU
4GB Memory
26GB System drive
20GB IOmeter test drive
20GB Movie File Drive

All VM resources stored on one vLUN.

Paperless Productivity
1402 Third Ave.
Suite 812
Seattle, WA 98122
Tel 1.888.838.0042
info@PaperlessProductivity.com

www.PaperlessProductivity.com