WHITE PAPER

Proving the Scalability of Software-Defined Storage With 25/50Gb Ethernet Networking

Contents

©2016 Mellanox Technologies. All rights reserved.

Executive Summary

Software-Defined Storage (SDS) is growing in popularity because it offers lower costs and more deployment flexibility than traditional storage arrays, but it comes in many different architectures and designs which support varying degrees of scale and performance. NexentaEdge is an ideal scale-out SDS choice for OpenStack clouds requiring high performance block and object storage. Benchmark testing of a 12-node cluster proved its ability to deliver high throughput with flash storage and also demonstrated the value of a high-speed Mellanox network in supporting that performance.

Advantages and Challenges of Software-Defined Storage

Software-Defined Storage is growing in popularity due to its ability to lower hardware costs significantly compared to traditional storage arrays. It also provides greater hardware flexibility as customers can choose the best and most efficient servers, storage media, and networking options for their needs, and change those hardware choices as needed. Some SDS choices, including NexentaEdge, offer the option to deploy as either traditional scale-out storage or as hyper-converged infrastructure, depending on the workload and application. The vast majority of cloud and Web 2.0 customers, as well as a growing number of enterprises, are choosing SDS as a way to achieve rack-scale deployments efficiently with high performance at an affordable cost.

However, there are often concerns that SDS does not support the same level of performance as traditional storage arrays sold on dedicated, and sometimes customized, hardware. Customers seek guidance and best practices on how to architect their SDS solutions for the best scalability, highest efficiency, and easiest manageability.

Test Goals

Mellanox, Nexenta and Micron desired to demonstrate the scalability and performance of NexentaEdge using Micron high-capacity SSDs, as well as the ability of higher-speed networking to enable higher levels of throughput from the storage cluster. Secondary goals included testing the performance at different network speeds and measuring the performance benefits of inline deduplication and compression.

NexentaEdge Design and Scalability

NexentaEdge is designed from the ground up to deliver high-performance block and object storage services and limitless scalability to next-generation OpenStack clouds, petabyte scale active archives, and Cloud Native Application infrastructures. NexentaEdge runs on shared-nothing clusters of industry-standard Linux servers, and builds on Nexenta patent pending Cloud Copy On Write (CCOW) technology to break new ground in terms of reliability, functionality, and cost efficiencies

NexentaEdge is a truly distributed, scale-out architecture, consisting of three or more physical servers interconnected using a dedicated Ethernet (10/25/40/50 GbE) network for cluster communication. The connected servers form a cluster that maintains redundancy and resilience of data throughout the system using strong cryptographic checksums for data integrity, and replication technology to ensure hardware-level redundancy. A single cluster can provide a global namespace across petabytes of storage, providing storage through block (iSCSI) or object (Swift and S3) interfaces. It supports cluster-wide in-line deduplication and compression to increase both storage efficiency and performance. Nexenta Replicast, a patented multicast technology, allows NexentaEdge to write multiple copies of incoming data with very low latencies.

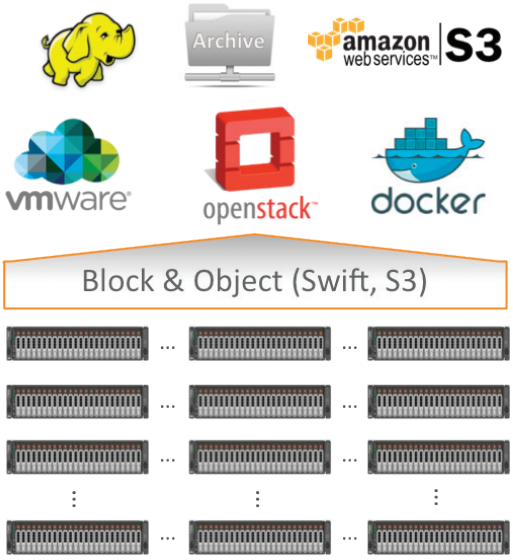


Figure 1. *NexentaEdge supports scale-out block and object storage.*

©2016 Mellanox Technologies. All rights reserved.

NexentaEdge Target Use Cases & Differentiators

The image below represents the typical NexentaEdge use cases, although this application now extends our OpenStack and Cloud Service Provider support and infrastructure up to many petabytes of block and object storage by allowing to run high performance and I/O intensive workloads. Some additional examples include:

- High performance storage pool for private and public clouds on OpenStack
- Analytics
- Multiple distributed databases
- Network services
- Dev/ops

OpenStack Cloud

openstack

iSCSI Cinder and Swift Object API
Low latency block services
Inline cluster-wide deduplication
Inline compression
Instant snapshots and clones

Active Archive

Swift and S3 Object API
Simple multi-PB scaling
Cloudscale availability management
Automated capacity balancing
Inline data reduction

Scale-Out Storage for VMware

vmware

Low latency iSCSI services
Simple multi-PB scaling
Cloudscale availability management
Inline cluster-wide deduplication
Instant snapshots and clones

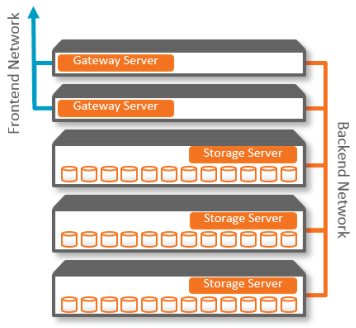
Container-Converged

docker

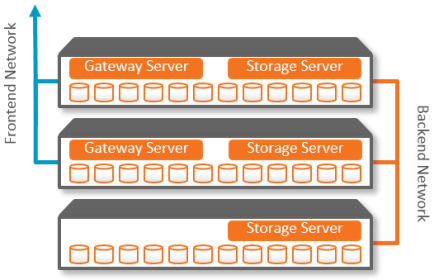
Deployed as storage microservice
Support stateful container mobility
Flocker volume plug-in
High performance container block driver
Inline data reduction
Instant snapshots and clones

NexentaEdge Deployment Options

NexentaEdge: Deployment Models



Dedicated Nodes Configuration



Mixed Nodes Configuration

A NexentaEdge cluster consists of these main components:

- Gateway nodes – Connect the storage system to the outside world through any of the supported access protocols: block/iSCSI or object (S3 and Swift)
- Data nodes – Store the actual data
- Networking infrastructure –Ethernet networking that is IPv6 capable, such as the Mellanox ConnectX® family of adapters and Spectrum Ethernet switches.

The gateway nodes can be separate from the data (storage) nodes or each node can be both a data and gateway node, which is a hyper-converged configuration. NexentaEdge uses two logical networks,

©2016 Mellanox Technologies. All rights reserved.

a cluster network for management and replication and a public network for connecting client traffic. These logical networks can be separated for performance or security reasons or combined onto one physical network if it has sufficient bandwidth. Write traffic, or incoming data, is typically written to disk two or three times for redundancy and availability, and this replication means that the cluster network typically sustains 2x or 3x the load of the front-end network. However, NexentaEdge is able to reduce this overhead by utilizing in-line compression, deduplication and intelligent multicast based networking.

Hyper-converged with combined gateway and data nodes and top-of-rack Ethernet switch

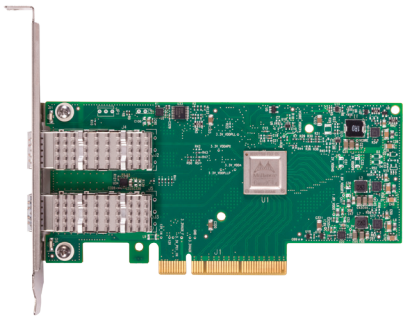


Figure 2. *Mellanox ConnectX-4 Lx Ethernet Adapter*

Mellanox End-to-End Ethernet Solution

Mellanox offers the ideal end-to-end Ethernet networking solution for deploying NexentaEdge, including adapters, switches, and cables. Mellanox adapters support the most popular Ethernet speeds for scale-out storage including 10, 25, 40, 50, and 100Gb Ethernet, and feature low latency, high throughput, efficient offloads, and low power usage. They offer advanced offload and hardware acceleration capabilities for general networking, virtualization and cloud deployments, including stateless TCP offloads, RDMA (RoCE and InfiniBand), SR-IOV, and overlay network acceleration for NVGRE, VXLAN, and GENEVE.

Mellanox Ethernet switches offer high performance which allows storage performance to scale linearly and predictably to dozens or hundreds of nodes. They deliver low latency with no jitter, at any packet sizes and network speeds, eliminate avoidable packet loss, and have fair bandwidth allocation using any port configuration. Innovative designs allow support for many different speeds and very dense port configurations, including a 1RU high, half-width switch ideal for building highly-available storage clusters.

Mellanox LinkX® interconnect solutions includes cables, transceivers and module with the highest reliability, lower power consumption, and lowest bit error rate (BER) in the industry. Available in copper, fiber Active Optical Cables (AOCs), and transceivers, Mellanox LinkX supports innovative technology to support high speeds over long distances and enable very dense 10, 25, and 50GbE port configurations on Mellanox switches.

Mellanox Ethernet adapters and switches are certified by Nexenta for use with NexentaEdge.

Figure 3. *Mellanox Spectrum SN2100 half-width 1U Ethernet switch*

©2016 Mellanox Technologies. All rights reserved.

Hardware Setup and Test Configuration

The NexentaEdge cluster was set up as follows:

Cluster Setup:

- 12 Nodes, running in mixed mode (hyper-converged), each running as a gateway + data node

NexentaEdge Server Hardware (each node):

- 8x SSD, Micron 510dc 960GB
- 256GB RAM (some nodes with 512GB)
- CPU: 2x E5-2697 v3 @2.6GHz

Network Setup

- Mellanox ConnectX-4 Lx 50Gb Ethernet adapter, one per server using a single port each
- Mellanox Spectrum SN2700 switch with ports configured for 50GbE
- Mellanox splitter cables, 100GbE QSFP28 to 2x50GbE QSFP28
- Two nodes at 50GbE connected to each switch port, via splitter cables

Software:

- Ubuntu 14.04.3 LTS with 4.2 Linux Kernel
- NexentaEdge 1.1.0-fp2
- NUMA is used to lock Client to CPU0/MEM0 and storage services to CPU1/MEM1 respectively

Benchmark Tool and Settings

- FIO - 2.2.8
- 20x LUN x12 nodes = 240 LUNs 100GB LUN = 24TB working set
- 256KB 100% rand write tests
- Compression : 0%
- Deduplication : 0%
- Replication : 2x
- fio settings - "allrandrepeat=0 direct=1 rw=randrw rwmixread=0/100 buffer_compress_percentage=0 dedupe_percentage=0 refill_buffers norandommap randrepeate=0 ioengine=psync bs=256k iodepth=2 numjobs=80 time_based group_reporting"

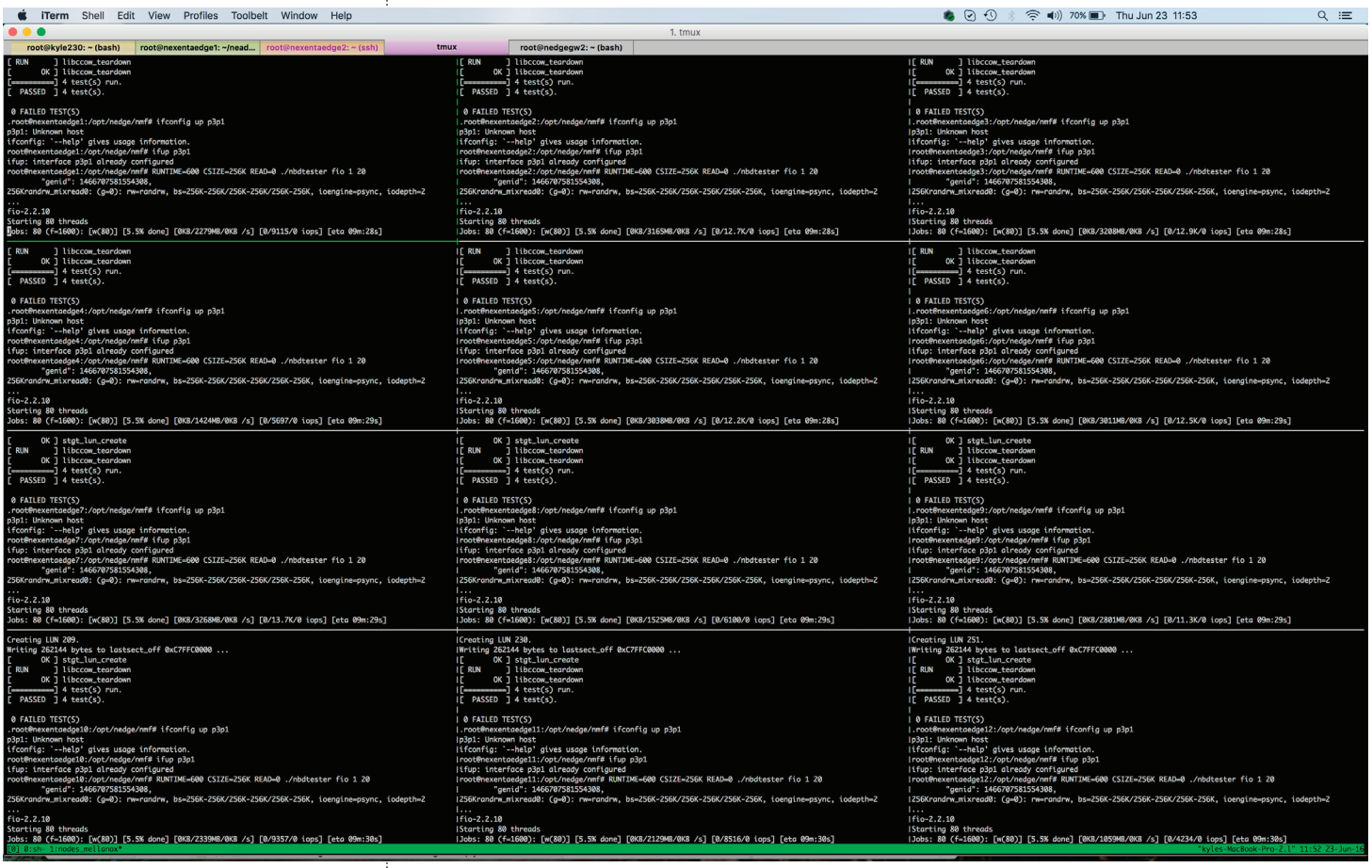
High Throughput Random Large Write using 50GbE Ethernet

The first test was designed to highlight scalability of the NexentaEdge software and network links, which required synchronous replication so as to force the most traffic over the network links while maintaining high throughput numbers.

Large blocks writes were used as small block does not adequately stress the wide network link scaling from 10gbit/s-50gbit/s without first saturating the CPU of the gateways. A write workload was chosen as the write path sends "replica count" additional copies of every block we are able to amplify network traffic with 2x replication by sending traffic synchronously. In this case, a replication factor of 2x meant that each 1MB write request actually sent 2MB of data over the network. Initial testing with writes all in-cache allowed write rates significantly higher than the sustained disk bandwidth is capable of. Using 50Gb Ethernet ensured the network was not the bottleneck and allowed one physical network to support both the public and cluster networks.

This test yielded peak network bandwidth figures across the 12-node cluster of 28.9GB/s, or approximately 231 Gb/s. Including the 2x replication traffic, actual cluster throughput was 57.8 GB/s (462 Gb/s) of back-end data traffic across the network. The per-node throughput of just over 19Gb/s of data and 38Gb/s of network traffic demonstrated the need for a network faster than 10GbE.

©2016 Mellanox Technologies. All rights reserved.

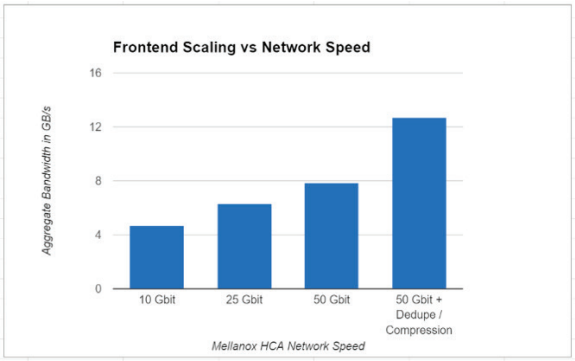


©2016 Mellanox Technologies. All rights reserved.

Advantage of Faster Network Speeds

The second test wrote data to actual flash media instead of to RAM cache, showing that increasing the speed of the network allowed the NexentaEdge cluster to support increasing amounts of write throughput. By varying only the network speed and holding all other aspects constant, the following aggregate front-end writes performance was obtained on the NexentaEdge cluster. Performance was measured at the front-end (client-side) with writes to flash storage, 100% random write workload with 256KB I/O size.

- 4.76 GB @ 10GbE
- 6.35 GB @ 25GbE
- 7.87 GB @ 50GbE



Performance Benefits of Inline Deduplication and Compression

NexentaEdge supports inline deduplication and compression, which is very rare in enterprise and cloud storage solutions as most other storage systems perform dedupe and compression as a "post" process, after writing the data. With this enabled and still writing to actual flash media, performance increased over time, climbing from 7.87GB/s to 12.7 GB/s of actual front-end data throughput. The performance increases over time because as more data is written there is an increasing chance that some of that data is not unique, and that it can be automatically deduplicated in-line by NexentaEdge, increasing the performance and reducing the amount of data that needs to be written to flash media.

Conclusion

NexentaEdge provides a highly-scalable SDS solution capable of achieving 29GB/s (232 Gb/s) of write throughput with twelve nodes, which actually required 57.8GB/s (462Gb/s) of network throughput with 2x replication. Connecting the servers in the NexentaEdge cluster with Mellanox Ethernet adapters and switches provided a non-blocking high-speed network with ease of deployment/management and industry-leading cost/performance. Deploying NexentaEdge with Mellanox high-speed Ethernet networking and Micron SATA data center SSDs allowed a high level of write performance which scaled up in a near-linear fashion as the number of nodes was increased. In addition, cluster write throughput increased as the network speed was increased from 10 to 25 to 50Gb/s. Turning on NexentaEdge's inline dedupe and compression increased performance further, showing the advantages of inline data efficiency vs. post-process data efficiency features.

Testing suggested that using more SSDs per node or faster (SAS or NVMe) SSDs would allow the same 12-node cluster—using the same server hardware, software, and networking equipment—to support even faster performance. Together, Nexenta and Mellanox break new ground for software-defined storage in terms of reliability, functionality and cost efficiency, and the joint solution enables next generation OpenStack clouds, petabyte scale active archives and Big Data applications.

©2016 Mellanox Technologies. All rights reserved.

About Micron

Micron Technology, Inc. is a global leader in advanced semiconductor memory systems. Micron's broad portfolio of high-performance technologies—including DRAM, NAND and NOR Flash—is the basis for solid state drives, modules, multichip packages and other system solutions. Backed by more than 35 years of technology leadership, Micron's memory solutions portfolio enables the world's most innovative computing, consumer, enterprise storage, networking, mobile, embedded and automotive applications. Micron's common stock is traded on the NASDAQ under the MU symbol. To learn more about Micron Technology, Inc., visit www.micron.com.

About Mellanox

Mellanox Technologies is a leading supplier of end-to-end InfiniBand and Ethernet interconnect solutions and services for servers and storage. Mellanox interconnect solutions increase data center efficiency by providing the highest throughput and lowest latency, delivering data faster to applications and unlocking system performance capability. Mellanox offers a choice of fast interconnect products: adapters, switches, software, cables and silicon that accelerate application runtime and maximize business results for a wide range of markets including high-performance computing, enterprise data centers, Web 2.0, cloud, storage and financial services. More information is available at www.mellanox.com.

About Nexenta

Nexenta is the global leader in Open Source-driven Software-Defined Storage (OpenSDS) with 6,000+ customers, 400+ partners, 42 patents, and more than 1,500 petabytes of storage under management. Nexenta is 100% Software-based; and 100% hardware-, protocol-, cloud platform-, and app-agnostic providing organizations with Total Freedom protecting them against punitive "vendor-lock-in", "vendor-bait-n-switch", and "vendor-rip-n-replace" gimmicks. Nexenta enables everyday apps from rich media-driven Social Living to Mobility; from the Internet of Things to Big Data; from OpenStack and CloudStack to Do-It-Yourself Cloud deployments – for all types of Clouds – Private, Public, and Hybrid. Founded around an "Open Source" platform and industry-disrupting vision, Nexenta delivers its award- and patent-winning software-only unified storage management solutions along with enterprise-scale 24x7 - around the globe - All Love - service and support with a global partner network.

Mellanox TECHNOLOGIES

350 Oakmead Parkway, Suite 100, Sunnyvale, CA 94085
Tel: 408-970-3400 • Fax: 408-970-3403
www.mellanox.com

© Copyright 2016. Mellanox Technologies. All rights reserved.
Mellanox and ConnectX are registered trademarks of Mellanox Technologies, Ltd. LinkX and Spectrum are trademarks of Mellanox Technologies, Ltd. All other trademarks are property of their respective owners.