INTEL® SOLUTIONS REFERENCE ARCHITECTURE

High-Availability and Scalable Cluster-in-a-Box HPC Storage Solution

Using RAIDIX® Storage Software Integrated with Intel® Enterprise Edition for Lustre*

Audience and Purpose

Intensive growth of hardware computing power and storage capabilities has transformed corporate business models in different industries, such as Finance, Media, Life Sciences, Climate, Manufacturing and others. Fast and scalable Intel® Enterprise Edition for Lustre* provides the benefits of High-Performance Computing for data-intensive corporate applications. The next challenge is to build High-Availability storage solutions for business critical IT environments.

This solution reference architecture provides information for companies and organizations that require scalable, high-performance and high-availability data storage solutions for mission critical workloads. Presenting an integrated RAIDIX and Intel Enterprise Edition for Lustre solution helps IT managers to build appropriate storage systems for business critical workloads with minimal downtime and data loss costs.

This document describes practical experience of integrated RAIDIX software for high performance storage solutions with Intel EE for Lustre* installed on Intel Xeon® processor-based servers. The solution combines the advantages of RAIDIX and Intel software and provides high-availability and scalable storage systems for high-performance parallel I/O operations. Integrated solutions installed in a single box provide companies great cost benefits.

"Lustre provides wide scalability, both in performance and storage capacity in a POSIX compliant global namespace and it can store billions of objects. Lustre can be used for a wide range of tasks including the increasingly important one of Big Data analytics."

– Sergey Platonov,
Product Owner,
RAIDIX

Table of Contents

Executive Summary

High Performance Computing (HPC) as a technology is no longer just a researchers' tool; now more and more companies are discovering the competitive advantages of HPC for their own business models. They generate huge volumes of data and use high-performance data processing applications to analyze and derive value from their data flows. This is a characteristic for several industries, and a common challenge most of them face is that they use HPC for business critical tasks; therefore, reliability and availability have the same importance as performance. These companies require a storage infrastructure that can scale-out endlessly and deliver high I/O capability for high-throughput data processing, together with high-availability and no tolerance for data loss.

Intel Enterprise Edition for Lustre* inherits the best values of well-known Lustre software and provides business customers with a commercial-grade version of Lustre optimized to address key storage and data throughput challenges of HPC-class computing in business. The main benefits of Intel Enterprise Edition for Lustre are high-performance, scalable capacity, proprietary software management tools, and corporate-class 24/7 support.

RAIDIX® provides a storage solution with fast and reliable failover, high-performance data processing, broad functionality for data integrity and monitoring. RAIDIX software, integrated with Intel Enterprise Edition for Lustre, comprises a high-availability bundle installed on Intel Xeon® processor based systems. RAIDIX erasure coding algorithms specially adapted to Intel Xeon processors provide great operations performance. This approach has significant cost reduction benefits, which are very important for commercial companies.

RAIDIX has implemented a joint solution powered by RAIDIX HPC and Intel Enterprise Edition for Lustre. The solution includes RAIDIX storage software run on commercial off-the-shelf (COTS) hardware with Lustre OSS/OST or MDS/MDT, representing an essential building block for Lustre HPC storage infrastructure. Such building blocks can contain 8 to 128 hard drives in a high density chassis with performance of up to 12 GB/s. Separate storage nodes are combined into a horizontally scalable system using Intel Enterprise Edition for Lustre.

RAIDIX storage as a building block for HPC storage infrastructure meets demands for high performance, fault tolerance and continuity of work by delivering high-bandwidth, low-latency and reliability, due to the use of parallel computing and unique patented and patent-pending methods and proprietary mathematical algorithms for RAID 6 and RAID 7.3 arrays. These algorithms provide data coding at the rate of up to 37 GB/s (RAID 6) and 25GB/s (RAID 7.3) per processor core. During the data reconstruction processes, the algorithms provide up to 25GB/s (RAID 6) and up to 15GB/s (RAID 7.3) per processor core.

Compared to traditional methods where Lustre OSS and MDS servers require additional hardware and each server needs to be configured individually, the RAIDIX approach to build HPC storage infrastructure from integrated building blocks reduces TCO of such infrastructure due to:

- Reduced hardware cost
- Reduced interconnection costs
- Simplified configuration and ease of setup and maintenance
- Reduced failover time and increased overall availability of the storage infrastructure

Introduction

RAIDIX consists of software services designed to create high-performance storage systems using widely used Intel processor based hardware platforms. To increase storage availability, RAIDIX has implemented a dual controller active-active storage cluster software module. Storage Bridge Bay (SBB) compliant platforms are more suitable for dual controller configurations, as this type of platform already contains some components required for HA storage.

General configuration requirements for RAIDIX dual controller platform:

CPU	Intel® Xeon® processors E5-2640. V3 and above
Motherboard	Compatible with the model of the processor and support PCI Express 3.0 x8/x16.
Internal Cache Memory	Compatible with the motherboard above, 64GB + per node
Enclosure	Dual power supply recommended, Dual motherboard recommended
SAS Controller (internal drives connection, additional could be used for external JBOD connection)	From RAIDIX® Hardware Compatibility List. LSI* 93xx is recommended
HBA (Controller Cache Synchronization)	From RAIDIX Hardware Compatibility List. Mellanox* ConnectX-3 VPI is recommended
HBA (Lustre Network Connection)	Mellanox ConnectX-3 VPI is recommended
HDD	For dual controller, SAS HDD is required
L2 Cache Devices	Intel® SSD DC P3700
Lustre Network	Infiniband* QDR/FDR/EDR, Ethernet 10GbE/40GbE/100GbE
Management Network	Ethernet 1GbE

Table 1 - General Configuration Requirements for RAIDIX Dual Controller Platform

For more information on the Storage Bridge Bay specification, see:

http://www.snia.org/sites/default/orig/sdc_archives/2008_presentations/monday/MarkHall_SBB_2_0_Spec_Overview.pdf

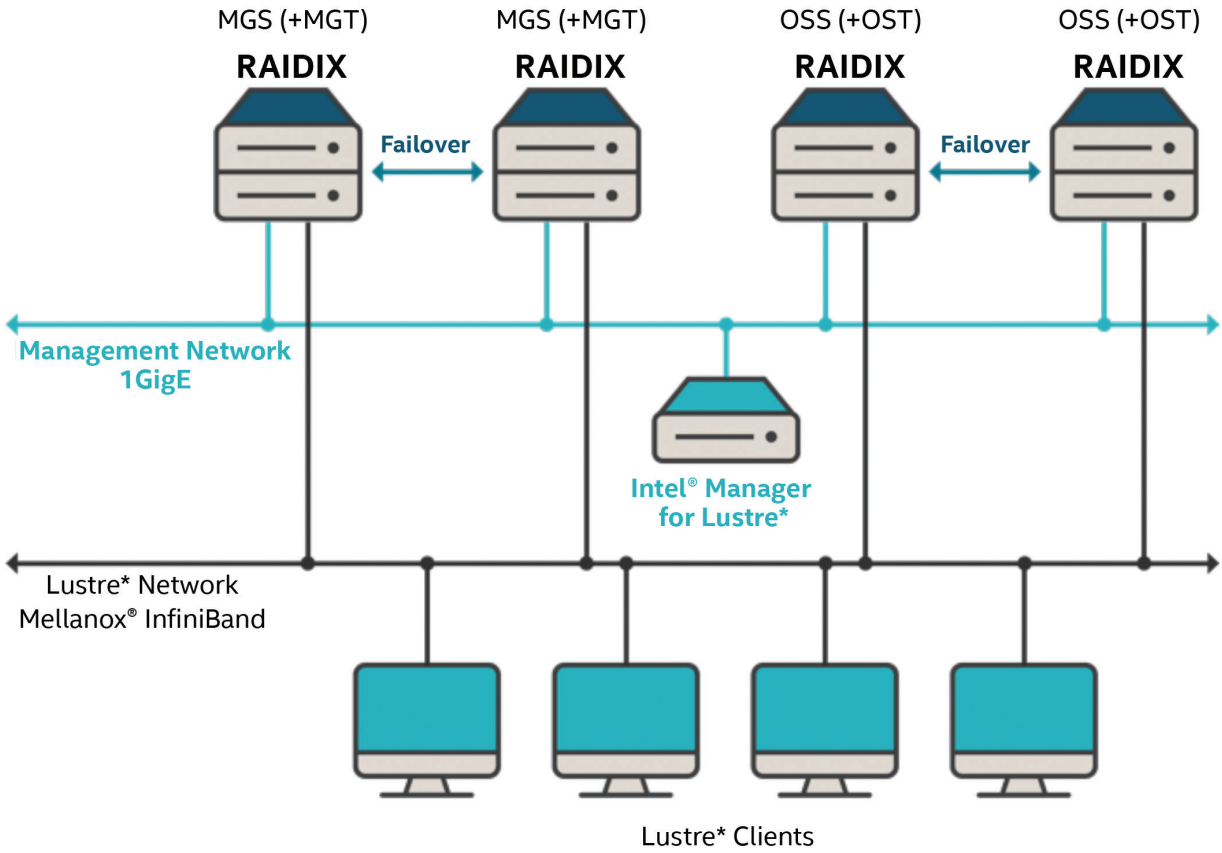


Figure 1 - Recommended Deployment Scheme for Typical HPC Application

The following deployment scheme is recommended for a typical HPC infrastructure:

1. RAIDIX recommends building an HPC Lustre infrastructure on dual controller RAIDIX building blocks to improve availability of each OST.
2. Each RAIDIX controller in RAIDIX DC used for OST has Lustre OSS installed in an active-active configuration.
3. Each OST of RAIDIX cluster is registered on both OSS's that are installed on the nodes of the cluster. Native RAIDIX failover is configured, and in a case of a failure, the RAIDIX failover mechanism transfers control of the OST to the other, non-failing OSS.
4. Lustre MGS and MDS also have to be configured in failover mode on RAIDIX DC to improve availability of MGT and MDT.
5. Intel Manager for Lustre is installed into the infrastructure to provide advanced management and monitoring functionality
6. 1GbE Ethernet is used for management networking
7. 56Gb Infiniband interconnect is used for Lustre networking
8. Each client has Lustre client installed

The above recommendations enables the creation of a highly available, manageable HPC storage infrastructure.

Product Overview

HPC storage solutions must provide the following crucial and important features: data security, data availability, scalability and guaranteed high performance. RAIDIX storage software with Intel Enterprise Edition for Lustre addresses the above issues; a modular structure and use of COTS hardware components make the solution cost very attractive.

Management

Lustre*Cluster Management

The Lustre cluster is managed through Intel Manager for Lustre, a web application built on a REST API and fully featured CLI. It has the following functionality:

- to provision and monitor Lustre file systems;
- to configure servers, volumes and power controls;
- monitoring tools to control performance level and resource utilization. Administrators can easily view file systems, check resource consumption for jobs, and monitor performance on the dashboard.

"Intel EE for Lustre software is backed by Intel, the recognized technical support providers for Lustre, including 24/7 service level agreement (SLA) coverage."

Figure 2 - Sample of Intel Manager for Lustre Dashboard

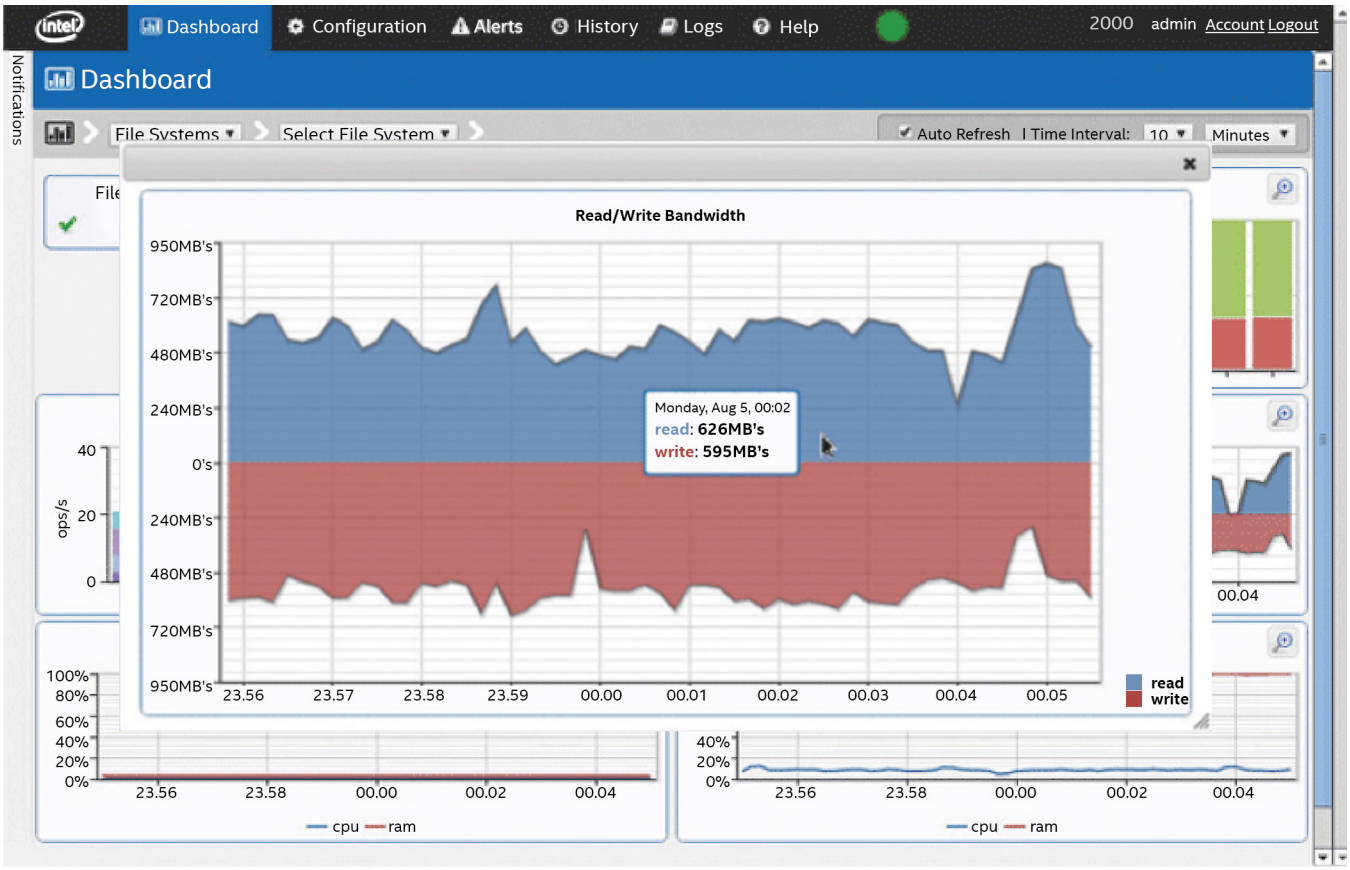


Figure 3 - Sample Detail from Intel Manager for Lustre Dashboard

Storage Management

RAIDIX powered storage has a web based interface which allows configuring and monitoring storage volumes, performance metrics etc. It gives clear information on the state of the storage.

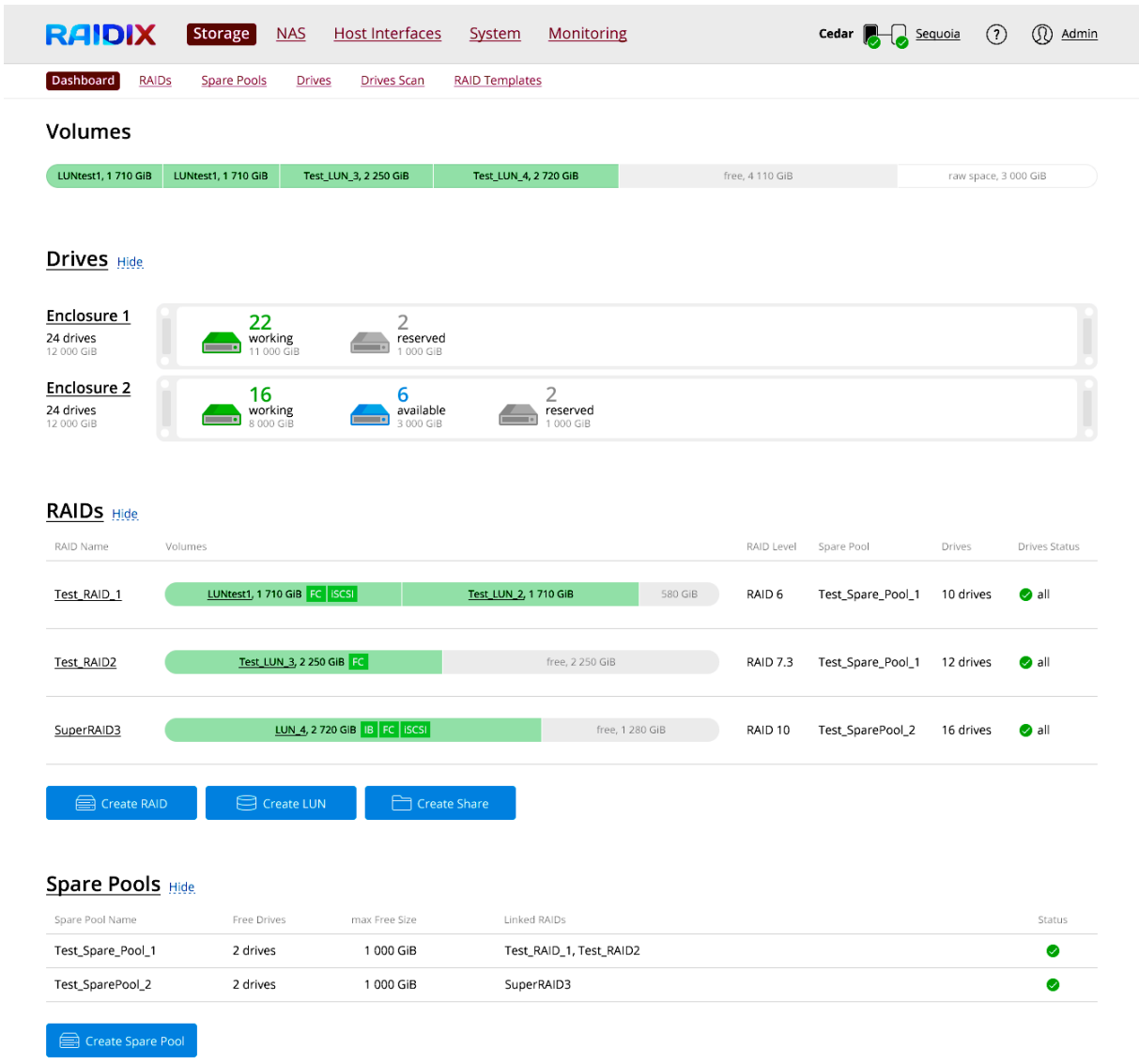


Figure 4 - RAIDIX Storage Web Monitoring Interface

Intel® Enterprise Edition for Lustre™ Value

A key component of the solution is Intel Enterprise Edition for Lustre. The choice to create a scale-out cluster solution based on RAIDIX storage software and Intel Enterprise Edition for Lustre was dictated by the great value of open source Lustre and the Intel proprietary enhancements:

- Essential manageability with Intel Manager for Lustre
- Improved I/O performance of enterprise applications like MapReduce
- Intel® Xeon® Phi™ client support
- Hadoop* connector allows use of the Lustre cluster for Hadoop applications
- Complete Hierarchical Storage Management
- Single thread improvement patch

Volume Data Safety and Protection

RAIDIX storage software uses erasure coding based on proprietary patented RAID algorithms optimized for high performance tasks. RAIDIX supports different RAID levels to provide flexibility for IT administrators to set up an appropriate level of volume data protection, including:

- Classical **RAID 6** can sustain a complete failure of two drives in the same group.
- **RAID 7.3** can sustain a complete failure of three drives in the same group: interleaving blocks with triple distributed parity, allowing restoration of data when up to 3 drives fail and providing increased reliability. RAID 7.3 is based on proprietary RAIDIX algorithms ensuring high sustainable performance without additional load on CPU. RAID 7.3 is recommended for arrays of more than 32 TB.
- **RAID N+M** is the level of interleaving blocks with M checksums, based on proprietary RAIDIX mathematical algorithms. RAID N+M allows the user to choose the number of disks for checksums allocation. RAID N+M requires at least 8 drives and can sustain a complete failure of up to 32 drives in the same group (depending on the number of parity drives). By selecting different parity numbers, the user can set a proper balance between the performance and data protection level, depending on how critical the data is and what performance requirements are.
- **RAID 10** sustains a complete failure of half of the array drives.

Besides supported RAID levels, two additional features improve volume data protection of RAIDIX-powered data storage systems.

- **Silent Data Corruption Detection and Correction**
  Business critical data should be unconditionally protected and physical failure of disks is not the only threat to data safety. Silent data corruption on healthy drives is a serious problem for storage systems under high workload. RAIDIX implemented a forward silent error correction algorithm that analyzes RAID metadata in the background mode to detect and fix corruptions and soft errors, while regular drive operations are performed without performance degradation.
- **Partial Reconstruction**
  Data reconstruction for failed high capacity drives can take a significant amount of time, increasing the probability of other drives failing and increased data loss risk. To eliminate the risk, RAIDIX software has a mechanism of RAID Partial Reconstruction that allows restoring only a particular area containing corrupted data on a hard drive, reducing the array recovery time. Partial reconstruction is extremely effective for large arrays.

RAIDIX volume in the integrated solution is used for OST and MDT to guarantee OST's and MDT's data integrity.

Silent Data Corruption

Silent Data Corruption happens due to a variety of causes - bugs in drivers, disk firmware, memory errors, electromagnetic interference, etc. Occurring at a higher rate than one in every 1016 bits, it happens without warning and typically goes unnoticed. An unintentionally changed block appears in the storage and there is no indication that the stored data is incorrect. With the growing capacity of hard drives the problem becomes more acute. Silent Data Corruption is especially treacherous during data reconstruction, when the corrupted block considered normal is replicated across the entire storage system. A risk of real data loss becomes viable.

Embedded Silent Data Corruption Detection and Correction Module

RAIDIX's forward error-correction algorithm uses mathematical properties of RAID checksums to analyze RAID metadata, detect and correct silent data corruption while regular disk operations are performed. RAIDIX implements error correction at the block level, rendering compatibility with any OS and the file system chosen by the platform's administrator.

Guaranteed High Performance

- The integrated solution has several features that help to increase overall performance:
- All RAID algorithms are calculated on standard Intel Xeon processors with a high level of parallelism, so a high level of data protection is combined with high, sustainable read and write performance
- Intel Xeon processors' high processing throughput provides advantages for RAIDIX software to implement an Advanced Reconstruction mechanism. This feature accelerates overall disk array performance by eliminating the drives with the lowest read rates from reading operations. The data expected from these drives is recalculated on the fly instead of physical reading of the drive. Calculation at a rate of up to 25 GB/s per core provides data at a much higher rate than the physical process of reading. Please see the last section of the paper On RAIDIX Erasure Coding Library (p.13) for more details.
- Lustre as a parallel file system allows the client to write and read to multiple OST's volumes simultaneously, increasing the overall performance.

Data High Availability

The Clustering system of RAIDIX 4.x generation software creates a fault-tolerant high-performance cluster (by configuring dual-controller mode) and places RAIDs asymmetrically on the nodes. Each RAID can be accessed through one of the nodes.

- Increases fault tolerance due to automatic and manual failover modes (switching between the nodes)
- Helps to increase system performance due to ability to migrate RAIDs from any node of the cluster to balance the nodes' workload.

Lustre Integration inside the RAIDIX dual controller allows the user to:

- Place several Lustre OST's asymmetrically on each RAIDIX cluster node and balance the load of each node to get maximum performance from the single cluster
- Provide high availability for data stored on OST or MDT; in case of one node failure, data will be available on the other node. This provides business continuity for an HPC cluster, as OST is always operational without interrupting HPC check pointing or data reading
- Integrate Lustre OST and MDT failover as a part of the storage node failover process. This eliminates the need for additional services like Corosync* and Pacemaker*, as the RAIDIX cluster runs Lustre failover itself.

Reference Architecture & Performance Analysis

Hardware Configuration

This document has been prepared based on the Proof of Concept small-scale cluster installation, built on AIC* HA201-WP. As a platform for the solution, RAIDIX recommends to have cluster nodes in one chassis and therefore to use similar SBB type hardware available from different vendors. It also can be extended with additional JBODs to increase capacity and performance. AIC HA201-WP is a 2U high-availability, Cluster-in-a-Box Storage Server Solution that leverages the SBB (Storage Bridge Bay) specification with the use of COTS components. The dual node configuration comes with two Intel® Server Boards (S2600WP). Each node supports dual Intel® Xeon® Processor E5-2600/E5-2600 v2/E5-2600 v3 series processors. HA201-WP is a high availability active-active solution, which features fully redundant, hot-swappable compute nodes, 24x 3.5" hot-swappable HDD bays and 5x PCIe Gen3 slots per node.

Detailed configuration of the platform used for POC:

Platform	AIC HA201-WP SBB
CPU	Dual Intel® Xeon® processor E5-2640 v2 for each motherboard
Motherboard	Intel® Server Board S2600WP
Internal Cache Memory	64 GB per node
Enclosure	AIC HA201-WP, dual motherboard, dual power supply, 24 hot swap drive bays
SAS Controller (internal backplane connection)	LSI 9300 8-i
HBA (Controller cache synchronization)	Mellanox ConnectX®-3 Dual-Port Adapter
HBA (Lustre network connection)	Mellanox® ConnectX-3
HDD	24x NL-SAS 7.2K
RAIDIX software	v.4.3.1
Intel® Enterprise Edition for Lustre	v2.2

Table 2 – Hardware Used in This Solution Reference Architecture

Figure 5 - AIC HA201-WP SBB Module – Front and Back

Interconnect

Each node has 2x Infiniband 56G HBAs installed: one to synchronize controller cache and the other to connect to Lustre infrastructure. Each node has a SAS HBA installed for JBOD expansion. Each node in the dual controller configuration is connected to its pair with 1 GbE heartbeat. Each node has to be connected to 1 GbE management network.

Software Configuration

1. Each node has to be configured as a dual controller storage according to RAIDIX administrative guide (available upon request from RAIDIX Technical Support support@raidix.com)
2. Create two arrays with RAID level 6 or 7.3 containing 12 drives each per system. Created RAID sets have to have affinity configured to different nodes so that each RAID set has affinity to one node and each node would serve one RAID set. Each RAID set needs at least 8GB of cache configured.
3. Create a full size LUN for each RAID set
4. Configure LNet, check that all nodes see each other
5. Mount file systems on MDT, MGT, and OST
6. Configure Lustre failover scripts on RAIDIX
7. Mount Lustre file system on a client side
8. Install Intel Manager for Lustre software
9. Add RAIDIX nodes, wait for the agent to be installed, choose Monitored only mode
10. Scan servers for Lustre file system

Test Stand

3 SBB DC nodes work as OSS/OST

1 SBB DC works as MDS/MDT

16 clients with Intel® Xeon® E5 v2 CPUs, 128GB RAM, FDR IB HCA, RHEL 6.6

Test Methodology

We used IOR HPC Performance benchmark to obtain test results

Writing data to File System

IOR -w -k -t 1m -b 32g -vv -e -g -F –o /mnt/Lustre

Reading data from File System

IOR -r -k -t 1m -b 32g -vv -e -g -F –o /Lustre/file

Test Results

WRITE TEST		READ TEST	
Number of Processes	Throughput(MB/s)	Number of Processes	Throughput(MB/s)
2	3715	2	3322
4	11804	4	9679
8	10312	8	8124
16	11084	16	8321
32	11092	32	7915
64	11101	64	7980
128	12603	128	7956
256	11905	256	8024

Table 3 – Write and Read Test Results

The write performance result was higher than 12GB/s. Each OST drive performs at 170 MBps in RAID 6. RAIDIX algorithms enable to achieve the maximum performance level from particular hardware configuration

Things to Consider

Future development of the solution reference architecture presented can follow two directions:

- Improving the current solution
  - » Configure SSD drive storage for MDT based on NVMe based Intel SSD P3700 for increased performance
  - » Integration with OpenZFS as a file system instead of ext4 based file system
  - » Hardware with advanced Intel Xeon processor E5-2600 v3 CPUs will improve results, especially with expansion JBODs and number of volumes per cluster > 4. With multiple JBOD configuration and RAID N+M, it becomes possible to configure the system to survive a failure of an entire JBOD. A sample configuration is presented on the scheme below:

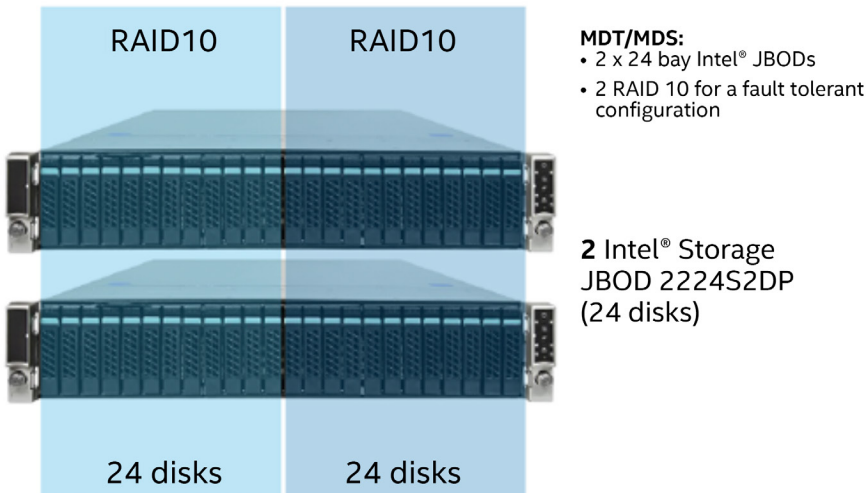


Figure 6 - Distributed File System Storage Solution: RAIDs Configuration for MDT/MDS

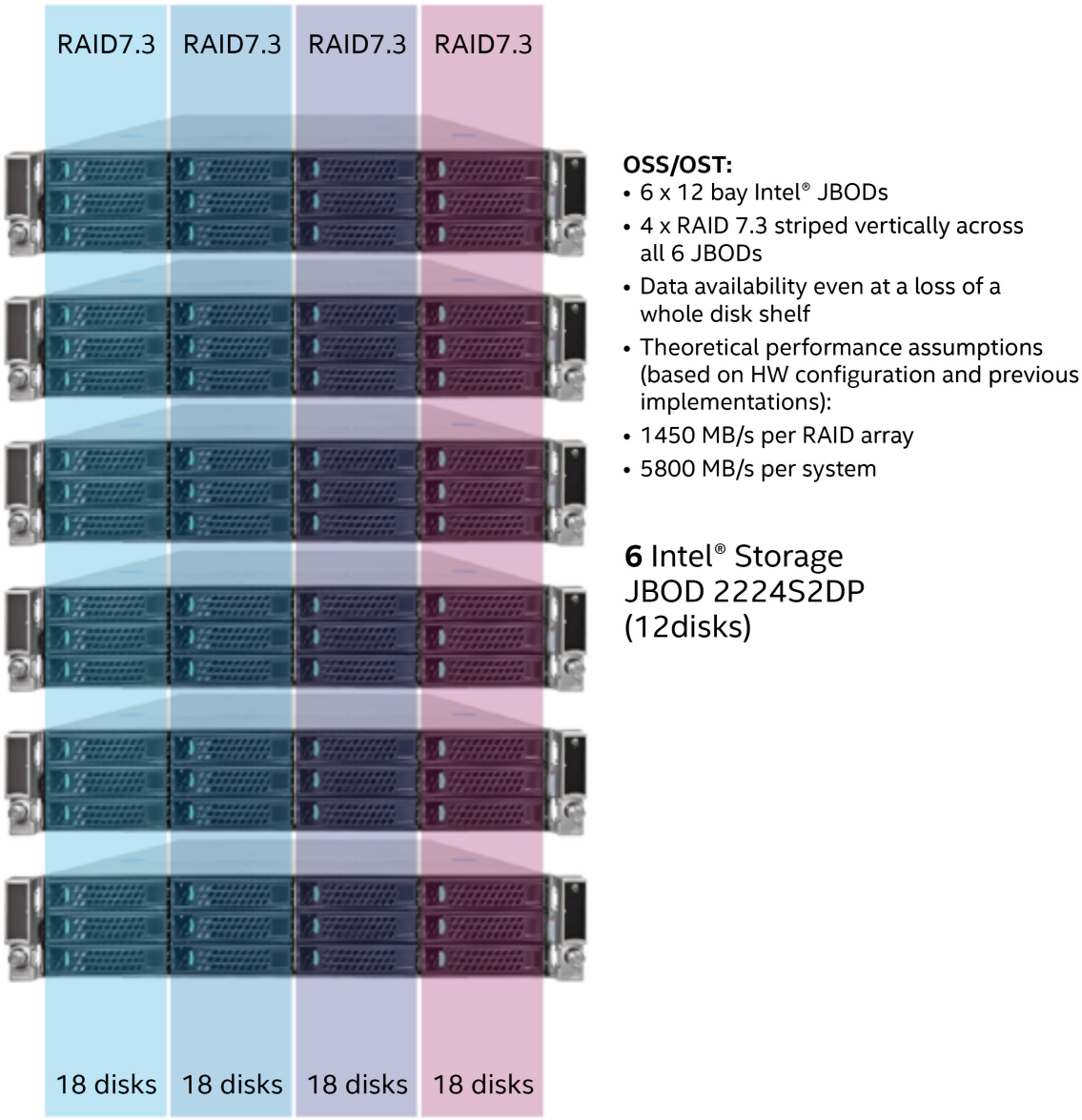


Figure 7 - Distributed File System Storage Solution: RAIDs Configuration for OSS/OST

» In future RAIDIX versions, Intel Non-Transparent Bridging (NTB) can be used for cache synchronization between the controllers

"We are pleased to welcome RAIDIX to the Lustre community. They bring a deep understanding of high performance storage software and its tailored integration into industry-specific solutions."

– Brent Gorda, General Manager of Intel's High Performance Data Division

Conclusion

Joint solution powered by RAIDIX HPC and Intel Enterprise Edition for Lustre offers a solid building block for HPC storage infrastructure. The solution meets the demand for high performance, fault tolerance and continuity of work by delivering high-bandwidth, low-latency and reliability. Compared to traditional methods where Lustre OSS and MDS and MGS servers require additional hardware and each server needs to be configured individually, this integrated approach to building HPC storage infrastructure from building blocks reduces TCO of such infrastructure due to:

- Reduced hardware cost
- Reduced interconnection costs
- Simplified configuration and ease of setup, management and maintenance
- Reduced failover time and increased overall availability of the storage infrastructure

Links

http://www.intel.com/content/www/us/en/software/intel-enterprise-edition-for-lustre-software.html

http://www.raidix.com/products/raidix-hpc/

http://www.snia.org/sites/default/orig/sdc_archives/2008_presentations/monday/MarkHall_SBB_2_0_Spec_Overview.pdf

Glossary

OSS – Lustre object storage server

MDS – Lustre Metadata server

MDT – Lustre Metadata Storage Target

OST – Lustre Object Storage Target

HBA – Host Bus Adapter

HA – High Availability

MGT – Lustre Management Target

MGS – Lustre Management Server

DC – dual controller

SBB – Storage Bridge Bay

COTS – Commodity off-the-shelf hardware

On RAIDIX Erasure Coding Library

RAIDIX proprietary erasure coding library is developed and patented by RAIDIX. The library performance was tested through the following method:

Description of the testing algorithm:

1. Allocate memory for data blocks and checksums blocks. Total (N+M) blocks with the size of 4096 bytes should be allocated. Memory is allocated with posix_memalign() function.
2. Fill data blocks with random values
3. Choose positions of failed blocks. Amount of failed blocks is equal to the number of checksums. Positions of failed blocks are random values
4. Allocate memory for coding matrix encode_matrix and decoding matrix decode_matrix;
5. Generate coding matrix using gf_gen__matrix() function
6. Generate tables for encoding;
7. Perform checksum calculation;
8. Fill failed blocks with zero.
9. Start timer using clock_gettime()
10. Generate decoding matrix;
11. Generate tables for decoding matrix;
12. Perform data recovery using table calculated for decoding matrix.
13. Stop timer.
14. Check recovered data
15. Free all allocated memory

16. Calculate the time between stopping and starting the timer

This algorithm is repeated 10000 times.

Library performance is calculated as P(MB/s) = V (MB) / T (nanoseconds), where V – is amount of generated data and T – is average time of data encoding or decoding provided by the tool.

Test stand specifications:

- **Motherboard**: Intel® S2600WTT
- **CPU**: Intel® Xeon® E5-2695 v3
- **RAM**: Micron* 36ASF2G72PZ-2G1A2 x 12

Test Results

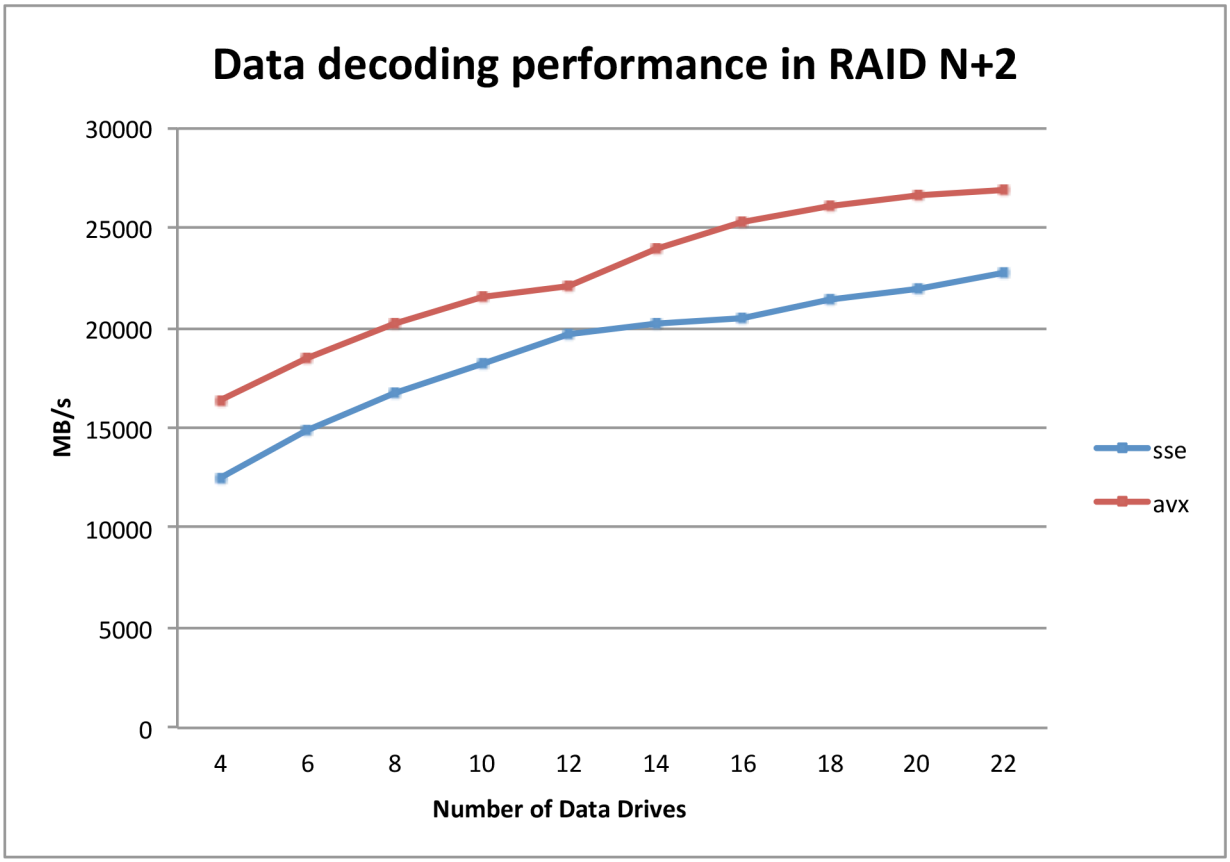
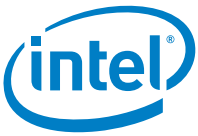


Figure 8 – Test Results - Throughput Achieved Using RAIDIX N+2 Algorithms



Disclaimers

Intel technologies' features and benefits depend on system configuration and may require enabled hardware, software or service activation. Performance varies depending on system configuration. No computer system can be absolutely secure. Check with your system manufacturer or retailer or learn more at intel.com.

Software and workloads used in performance tests may have been optimized for performance only on Intel® products and microprocessors.

Performance tests, such as SYSmark and MobileMark, are measured using specific computer systems, components, software, operations and functions. Any change to any of those factors may cause the results to vary. You should consult other information and performance tests to assist you in fully evaluating your contemplated purchases, including the performance of that product when combined with other products. For more complete information visit http://www.intel.com/performance.

INFORMATION IN THIS DOCUMENT IS PROVIDED IN CONNECTION WITH INTEL® PRODUCTS. NO LICENSE, EXPRESS OR IMPLIED, BY ESTOPPEL OR OTHERWISE, TO ANY INTELLECTUAL PROPERTY RIGHTS IS GRANTED BY THIS DOCUMENT. EXCEPT AS PROVIDED IN INTEL'S TERMS AND CONDITIONS OF SALE FOR SUCH PRODUCTS, INTEL ASSUMES NO LIABILITY WHATSOEVER, AND INTEL DISCLAIMS ANY EXPRESS OR IMPLIED WARRANTY, RELATING TO SALE AND/OR USE OF INTEL PRODUCTS INCLUDING LIABILITY OR WARRANTIES RELATING TO FITNESS FOR A PARTICULAR PURPOSE, MERCHANTABILITY, OR INFRINGEMENT OF ANY PATENT, COPYRIGHT OR OTHER INTELLECTUAL PROPERTY RIGHT. UNLESS OTHERWISE AGREED IN WRITING BY INTEL, THE INTEL PRODUCTS ARE NOT DESIGNED NOR INTENDED FOR ANY APPLICATION IN WHICH THE FAILURE OF THE INTEL PRODUCT COULD CREATE A SITUATION WHERE PERSONAL INJURY OR DEATH MAY OCCUR.

*Other names and brands may be claimed as the property of others.

Printed in USA Please Recycle