Western Digital.

RAIDIX

REFERENCE ARCHITECTURE JULY 2021

High-Performing RAID-Enabled Data Protection Solution on OpenFlex™ F3200 Powered by RAIDIX

Authors: Puspanjali Panda, Pavan Gururaj, Saravana Kumar Pandian

Western Digital.

RAIDIX

REFERENCE ARCHITECTURE JULY 2021

Contents

1. Executive Summary

NVM Express or NVMe™ is the next evolution in storage connectivity and protocol adopted by the enterprise. NVMe offers greater performance for local drives, storage systems and devices connected across a storage network. Compared to legacy protocols like SAS and SATA, NVMe is changing the way storage products are developed and implemented.

The NVMe protocol has significantly expanded data center opportunities. NVMe drives not only offer all the features of traditional SSDs, but also deliver new levels of performance and latency, making them the ideal choice for a range of top-notch use cases, such as OLAP, OLTP, ML & AI, advanced ERP applications and autonomous production lines. These applications still demand a great deal of long-term investment, but with NVMe that becomes easier and more affordable. At the same time, there are some tangible factors that limit the wide distribution of NVMe drives. End users need to understand how NVMe will bring about fundamental changes in the storage architectures we use today. NVMe devices differ from traditional storage, and NVMe over Fabrics (NVMe-oF™) will become the dominant storage protocol for the enterprise and hyper-scale public cloud soon.

Two things are crucial when using NVMe in enterprise-grade data centers: the ability to work within fault-tolerant RAID for data protection and the opportunity to use NVMe-oF as a transport to network-attached drives. Existing software or hardware RAID technologies either use excessive redundancy in array mirroring or lose up to 65% of the performance in RAID 5 or RAID 6. In addition, not all of these approaches have the ability to support NVMe-oF for network access.

The OpenFlex F3200 is a fabric device that leverages the Open Composable Infrastructure (OCI) approach in the form of disaggregated data storage using NVMe-oF. RAIDIX ERA is an easily deployed RAID block storage solution, developed with a focus on high performance, simple installation and minimal management effort.

This paper provides an overview of the system architecture of the combined OpenFlex F3200 with RAIDIX ERA configuration and a performance analysis of the solution. You will see that OpenFlex with RAIDIX ERA provides significant performance, data protection and energy efficiency comparable to an all-flash NVMe system, but at a much lower cost, which may enable a significant TCO improvement for performance-critical applications.

2. Problem Statement

With the growth in both size and amount of content, total storage capacity is rapidly increasing. Workloads and data changes impact the choice of storage medium. Hard disk drives (HDDs) are historically the most popular storage devices, sometimes with magnetic tape or storage in the cloud used to back up or archive content. The cloud (tiered storage in data centers) is often used for storing and sharing content for collaborative editing. Although HDDs will remain important storage media for parking content and storing content not frequently accessed, they have inherent performance limitations in the context of editing today's high-capacity and high-bandwidth 4K+ content.

Thankfully, as manufacturing efficiencies have improved and competition has become fiercer, the economics of NAND flash memory has improved enough such that many organizations are now able to consider this storage media for a larger percentage of post-production workflows. Like most developments in the technology industry, things rarely stay stagnant for long, and such is the case with flash memory devices. Although Serial AT Attached (SATA) and Serial Attached SCSI (SAS) SSDs are widely used, these older interfaces (originally created to support the needs of HDDs) can restrict the data rate and latency that SSDs are capable of achieving. This has led to the wide use of an interface that brings more of the internal performance of the SSD to the computers it is connected to. This new interface is called Non-Volatile Memory Express (NVMe). Many applications now must accommodate faster data access while avoiding any delay with data protection. The OpenFlex F3200 is a fabric device that leverages the OCI approach in the form of disaggregated data storage using NVMe-oF. RAIDIX ERA is a software RAID developed with close attention paid to NVMe features. Innovative RAIDIX technologies and a focus on the new protocol's strengths make RAID engines capable of achieving up to 97% of the total drive's performance in RAID 6 and lower latency even for mixed workloads. So with both OpenFlex and RAIDIX, organizations can enable greater performance, including collecting and processing more data with higher velocity; capturing, storing and processing of millions of events; and delivering faster responses to analytics and query requests for insights and machine learning applications.

3. Solution Highlights

Western Digital, a pioneer in reliable, high-density, industry-standard hardware for software-defined storage projects, is partnering with RAIDIX for providing a high-performance RAID storage solution. The following sections of this paper provide an overview of the RAIDIX ERA RAID solution built on Western Digital.

The solution combines:

- **Western Digital OpenFlex**, which leverages an open composable infrastructure approach in the form of disaggregated data storage using NVMe-oF, making it a perfect fit for scaling big data AI environments.
- **RAIDIX ERA**, a software RAID for flash drives that uses up to 97% of maximum total hardware potential. With parallelized computation and lockless architecture, ERA increases the full speed of all-flash devices to skyrocket storage performance. This is absolutely essential in large systems for enterprise infrastructures, in HPC-computing and for companies in the M&E industry.

By combining RAIDIX ERA with Western Digital OpenFlex F3200, organizations will benefit from:

- Robust performance
- Data protection
- Scalability
- Data durability
- Data integrity
- Data throughput
- Easy deployment and integration with existing infrastructure
- Easy data management
- Optimization for highly concurrent access
- Easy cloud integration

4. Technology Overview

4.1 OpenFlex™ E3000 Enclosure with F3200 Fabric Device Overview

Designed for High Density and Flexibility

The OpenFlex E3000 is a 3U rack-mounted data storage enclosure built on the OpenFlex platform. OpenFlex is Western Digital's architecture that supports OCI through storage disaggregation. The OpenFlex F3200 is a fabric device that leverages this OCI approach in the form of disaggregated data storage using NVMe-oF. NVMe-oF is a networked storage protocol that allows storage to be disaggregated from compute to make that storage widely available to multiple applications and servers. For more details refer to OpenFlex-Composable-Infrastructure.

NVMe-oF enables applications to share a common pool of storage capacity, making data easily sharable between applications and needed capacity allocatable to an application regardless of location. Exploiting NVMe device-level performance, NVMe-oF promises to deliver the lowest end-to-end latency from application to shared storage. NVMe-oF enables composable infrastructures to deliver the data-locality benefits of NVMe DAS (low latency, high performance) while providing the agility and flexibility of sharing storage and compute.

Figure 1: OpenFlex E3000 Fabric Enclosure

Figure 2: OpenFlex E3000 Fabric Device

The maximum data storage capacity of E3000 is 614TB[1] when leveraging a full set of 10 F3200 fabric devices. F3200 is capable of scaling up to 2 million IOPS and cumulatively each E3000 can be scaled up to 20 million IOPS in a 3U solution.

Composable infrastructure seeks to disaggregate compute, storage and networking fabric resources into shared resource pools that can be available for on-demand allocation (i.e., "composable"). Composability occurs at the software level and disaggregation occurs at the hardware level using NVMe-oF. This vastly improves compute and storage utilization, performance and agility in the data center.

Western Digital's vision for OCI is based on four key pillars:

Open

- Open in both API and form factor.
- Designed for robust interoperability of multi-vendor solutions.

Scalable

- Ability to compose solutions across the entire network.
- Enable self-organizing systems of composable elements that communicate horizontally.

Disaggregated

- Pools of resources available for any use case that is defined at run time.
- Independent scaling of compute and storage elements to maximize efficiency and agility.

Extensible

- Inclusive of both disk and flash.
- Entire ecosystem of composable elements managed and orchestrated using a common API framework.
- Prepared for yet-to-come composable elements – e.g., memory, accelerators.

Open Composable API. Western Digital's new Open Composable API is designed for data center composability. It builds upon existing industry standards utilizing the best features of those standards as well as practices from proprietary management protocols.

OpenFlex is Western Digital's architecture that supports OCI through storage disaggregation – both disk and flash natively attached to a scalable fabric. OpenFlex does not rule out multiple fabrics, but whenever possible, ethernet will be used as a unifying connect for both flash and disk because of its broad applicability and availability.

OpenFlex F3200 Specification Summary

Specification	Value
Max raw data storage capacity per device	61.4TB
Data ingest capability	2x 50G Ethernet
Data transfer rates	12 GBps*
Number per enclosure	Up to 10
Hot-swappable	Yes

4.2 RAIDIX ERA Overview

RAIDIX ERA allows for the creation of a highly performative RAID from NVMe and SAS/SATA SSD for the most demanding enterprise-grade tasks, ensuring fast and effective access to data. It is easy to maintain and more suitable for operating in large server infrastructures. For more details refer to RAIDIX.

RAIDIX ERA is a software RAID presented by Linux® kernel module and management utility (CLI).

- Adjusted for the most popular Linux distribution (Ubuntu, CentOS and Oracle®).
- Works with local and remote drives.
- Provides RAID as a standard Linux block device.
- POSIX API support.

I/O handling parallelization and lockless data path in RAIDIX ERA allow for the removal of array internal barriers and deliver unprecedented performance. RAIDIX ERA is also able to sustain high performance levels and low latency (< 0.5ms) even in mixed workloads. To protect data, RAIDIX ERA delivers a wide range of RAID level support: RAID 1/0/5/6/7.3/50/60/70. Moreover, drive failure causes a low performance loss, which helps business applications run smoothly. That comes from an innovative approach to erasure coding calculations.

High Speed of the Checksums Calculations

The core innovation of the RAIDIX product is the unique software RAID that calculates array parity faster than any other alternatives in the storage industry. The RAID engine reads and writes parity blocks with the record speed (about 25GBps for 1 CPU core) and therefore keeps high array performance even when the drive goes down.

During the sequential workloads, a drive failure reduces total storage performance by less than 10%. This result is better than that delivered by any other existing storage solution.

Reduced RAID Rebuild Time

Rebuild (or reconstruction) of the RAID after a drive failure is a potentially fraught and dangerous time frame for storage administrators.

First of all, reconstructing data to a new drive usually consumes a significant part of the total array performance. Second, it increases the risk of data loss because the number of drives acceptable for the failure is down by one.

With fast checksums calculation, RAIDIX software arrays require significantly less time to perform rebuild operations compared to existing solutions at the global storage market.

Advantages of the RAIDIX Software RAID

Due to its fast coding and decoding ability, RAID provides a stable performance level needed for smooth and uninterrupted business operations. Fast RAID rebuild protects storage from extensive system downtime and mitigates the impact on workflows even if a few drives fail.

That is crucial for data-intensive systems and high-density storage infrastructures where even a single drive failure can cause the checksum recalculation for a vast amount of data.

RAIDIX has developed a range of technologies that apply a fast RAID engine to enhance software-defined storage functionality.

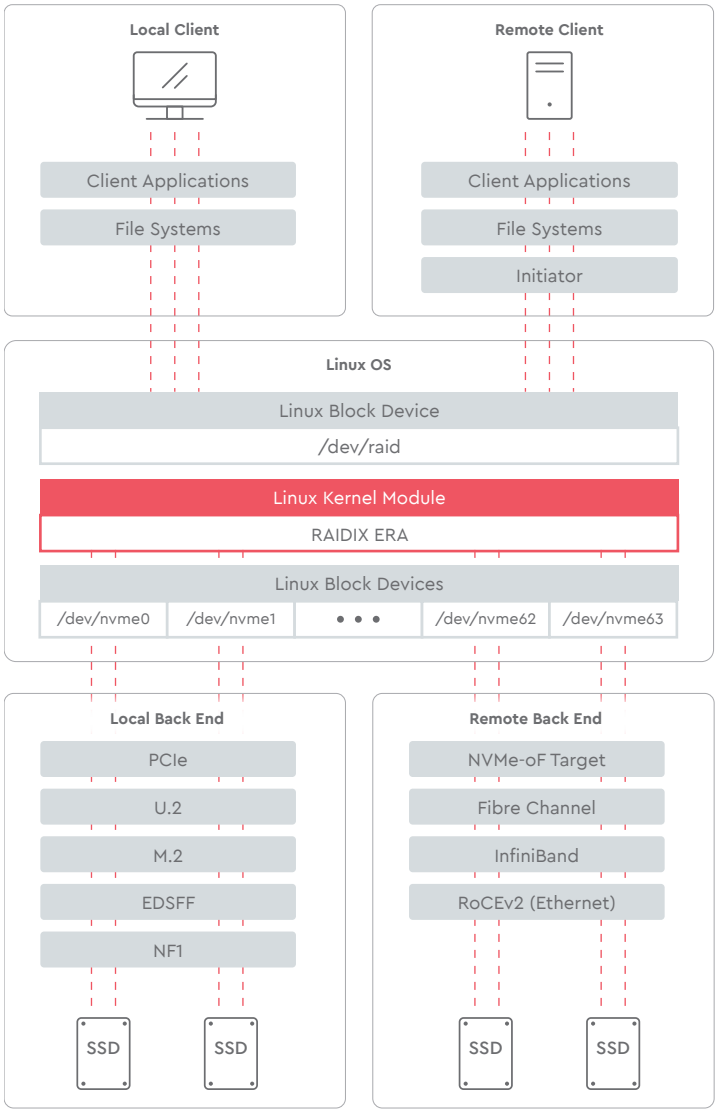


Figure 3: RAIDIX ERA Architecture

5. RAIDIX ERA Installation & Configuration Details on OpenFlex™

We have validated an alternate RAID solution on an OpenFlex F3200 device powered by RAIDIX after running multiple experiments with different types of RAIDIX ERA RAID configuration. As part of the test, we have used an FIO tool to run a performance test on the OpenFlex F3200 device after configuring various RAIDIX raid volumes (e.g., RAID0, RAID1, RAID5, RAID50, RAID6 and RAID N+M).

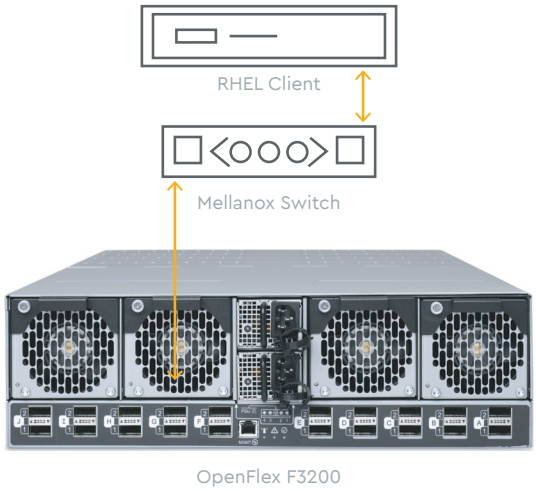


Figure 4: Topology Diagram

To set up OpenFlex system refer to OpenFlex-User-Guide.

The below configuration has been set up for test on OpenFlex F3200.

Product	OpenFlex 3200 Fabric Device
Interface	Dual QSFP28 (2×50Gb)
Host Server CPU Core	Intel® Xeon® Gold 5118 CPU @ 2.30GHz
Host Server CPU Core Details	Dual socket server, 24-core CPU each; 96 logical cores with HT enabled
Host OS	Red Hat® Enterprise Linux® release 8.2 (Ootpa)
Kernel	4.18.0–193.28.1.el8_2.x86_64
Host NIC	CX5 – MCX516A-CCAT
NIC Firmware Version	16.28.2006 (MT_0000000012)
CXS OFED package version	5.0–2.1.8
ERA RAID Version	3.2
Memory	128GiB
No. of Volumes	8

5.1 Setting up RAIDIX ERA

You need a valid license and Era RAID utility on your system.

The RAIDIX ERA 3.2 distribution consists of four software packages for Linux OS.

- *eraraid* – the main functionality which is the driver for RAID arrays;
- *eraraid-util* – a user management utility for RAIDs and licenses;

The package type depends on your Linux distribution. To get the license and rpm packages up refer to RAIDIX-ERA.

After getting the rpm, to install RAIDIX ERA 3.2, unpack the archive with the distribution to a folder and go to the folder by running the following commands.

 # tar xzf raidix_era.tar.gz

 # cd raidix_era 1.2

To install RAIDIX ERA 3.2, run the installation script installer.sh, which is located in the archive with the program packages.

 # ./installer.sh

During its running the script will perform the below.

- ✓ Ask you to accept the terms of the license agreement. You will see the text of the agreement on your screen.
- ✓ Ask you to confirm the installation of DKMS.
- ✓ Check the presence of the required package with the kernel header files.
- ✓ Install the required dependencies from the repositories (For more details refer to RAIDIX_ERA_Installation_Guide). If the installation fails, check and install all the required dependencies to proceed with the installation. After installing all required dependencies, rerun *./installer.sh* script.
- ✓ Once the installation is done, it will ask to start the ERA services.

5.2 RAIDIX ERA Installation Check & RAID Creation

1. To ensure that the packages were installed successfully, run:
   # eraraid show
2. To check the license details run:

 # [root@node1 ~]# eraraid license --show
Kernel version: 4.18.0–193.28.1.el8_2.x86_64
hwkey: 480722A1A9248A59
license_key:
42BAA83DDB24B367262251BF1D428466702E8FBF804C94E0DCA
BE9807BA0F093C712D862A1CDB7FC9E8D27A6F2ZZF84F53889A
787DDACTTSABC457ACBC83254C09D2B12BA403F498939AF3913
5522C3942F6D6428D3B0A6AADBNN
version: 1
crypto_version: 1
created: 2021–3–15
expired: 2021–5–15
disks: 30
levels: 70
type: nvme
disks_in_use: 8
status: valid

3. To create a RAID of levels 0, 1, 5, 6, 7.3, or 10, run:

   # eraraid create -n <raid_name> -l {0|1|5|6|7|10} –d (drives) [–ss {16|32|64|128|256}] [-bs {512|4096}]

   # [root@node1 ~]# eraraid create -n era1 -l 5 -d /dev/mapper/mpathd /dev/mapper/mpathe /dev/mapper/mpathf /dev/mapper/mpathg /dev/mapper/mpathh /dev/mapper/mpathi /dev/mapper/mpathj /dev/mapper/mpathk -ss 16 -bs 4096

To create a RAID of levels 50, 60, or 70, you must also point to the -gs parameter:

 # eraraid create -n <raid_name> -l {50|60|70} –d (drives) -gs <group_size> [–ss {16|32|64|128|256}] [-bs {512|4096}]

 # [root@node1 ~]# eraraid create -n era1 -l 50 -d /dev/mapper/mpathe /dev/mapper/mpathf /dev/mapper/mpathg /dev/mapper/mpathh /dev/mapper/mpathi /dev/mapper/mpathj /dev/mapper/mpathk /dev/mapper/mpathl -gs 4 -ss 16 -bs 4096

To create a RAID N+M, add the parameter -sc:

 # eraraid create -n <имя_raid> -l nm –d (block_devices) -sc <number of syndroms> [-ss {16|32|64|128|256}] [-bs {512|4096}]

 # eraraid create -n era1 -l nm -d /dev/mapper/mpathe /dev/mapper/mpathf /dev/mapper/mpathg /dev/mapper/mpathh /dev/mapper/mpathi /dev/mapper/mpathj /dev/mapper/mpathk /dev/mapper/mpathl -sc 4 -ss 16 -bs 4096

Required Parameters		
-n	--name	RAID name.
-l -d	--level --drives	RAID level: 0, 1, 5, 6, 7, 10, 50, 60, 70, or N+M. Space-separated list of block devices.
-g	--group_size	Only for RAIDs 50, 60, or 70. Set the drives number for one RAID group of level 5, 6, or 7.3 of an appropriate RAID 50, 60, or 70. Possible values are integers from 4 to 32.
-bs	--block-size	RAID block size: 512 or 4096 bytes.
-sc	--synd-cnt	Only for RAIDs N+M. Number of syndromes M. Possible values are integers from 4 to 32. Additional conditions: N+M ≤ 64 and M ≤ N.

For more details on eraraid parameters, refer to

 # eraraid -s

4. To show RAID information:

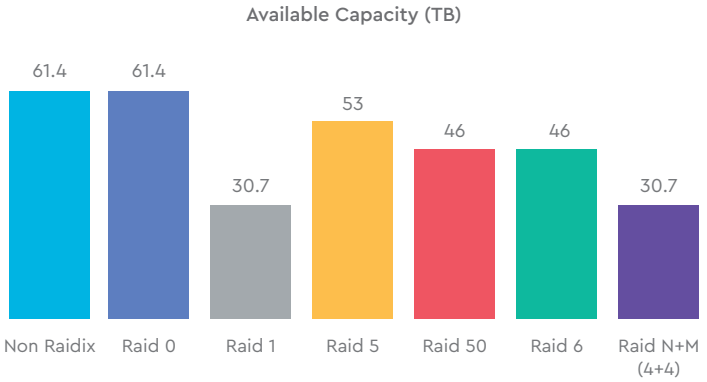
```
root_node1 ~]# eraraid show
RAIDs
```

name	static	state	devices	info
era1	size: 50069 GiB level: 5 strip_size: 16 block_size: 4096 active: True config: True	online initialized read_only	0 /dev/mapper/mpathd online 1 /dev/mapper/mpathe online 2 /dev/mapper/mpathf online 3 /dev/mapper/mpathg online 4 /dev/mapper/mpathh online 5 /dev/mapper/mpathi online 6 /dev/mapper/mpathj online 7 /dev/mapper/mpathk online	memory_usage_mb : –

6. RAIDIX ERA Performance Details on OpenFlex

Note: RAIDIX ERA RAID experiments were conducted on a single OpenFlex F3200 device of 61.4TB capacity. An FIO tool was used to measure the performance on the ERA RAID volumes.

The chart below shows the available capacity of different RAID configuration created using RAIDIX ERA on an OpenFlex F3200 device.

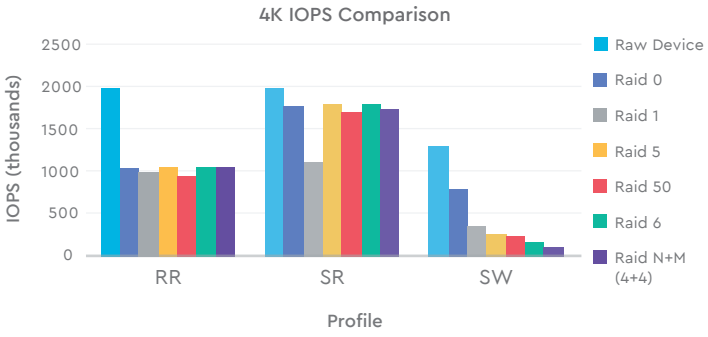


The table below shows the 4K IOPS performance comparison of OpenFlex F3200 device (raw) block performance and block performance results on different ERA RAID volumes.

4K IOPS (thousands)

Profile	Block Performance	Raid 0	Raid 1	Raid 5	Raid 50	Raid 6	Raid N+M (4+4)
SR	2000	1787	1121	1810	1715	1814	1744
SW	1300	790	349	270	240	170	100
RR	2000	1047	1003	1065	943	1068	1059

The chart below shows the 4K IOPS performance comparison of OpenFlex F3200 device (raw) block performance and block performance results on different ERA RAID volumes.

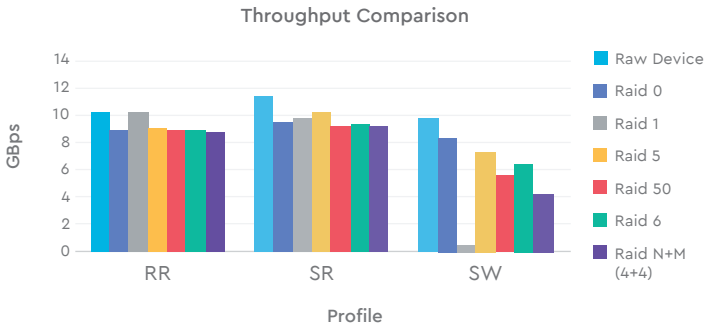


The table below shows the 1M large block throughput performance comparison of OpenFlex F3200 device (raw) and the throughput results on different ERA RAID volumes.

1M (GBps)

Profile	Block Performance	Raid 0	Raid 1	Raid 5	Raid 50	Raid 6	Raid N+M (4+4)
SR	11.5	9.7	10.1	10.3	9.4	9.5	9.4
SW	10	8.5	1.5	7.6	6	6.7	4.7
RR	10.5	9.1	10.3	9.2	9.1	9.1	9

The chart below shows the 1M large block throughput performance comparison of OpenFlex F3200 device (raw) and the throughput results on different ERA RAID volumes.



7. Conclusion

For enterprises, Western Digital OpenFlex with RAIDIX ERA unlocks the opportunities of leading technologies. For storage vendors and system integrators, it is a simple and affordable way to get a technological advantage for building NVMe solutions. This powerful software RAID solution with OCI achieves high speeds and low latency with different workloads in order to optimize the performance potential of NVMe drives in fault-tolerant array.

The results prove that high performance with data protection is achievable with little or no loss in speed and latency.

- Customers can also save up to 70% in costs by eliminating overprovisioned storage and compute resources with this solution.
- With low latency, consistent high performance and easy scalability, OpenFlex will provide a new range of operational capabilities.
- OpenFlex & RAIDIX can be used in e-commerce, AL-ML and HPC deployments which can realize TCO savings of up to $10 million per application when compared to other alternatives.

8. References

OpenFlex: https://www.westerndigital.com/products/data-center-platforms/openflex-composable-infrastructure

RAIDIX: https://www.raidix.com/products/

OpenFlex User Guide: https://documents.westerndigital.com/content/dam/doc-library/en_us/assets/public/western-digital/product/platforms/openflex/user-guide-openflexf3100-western-digital.pdf

Western Digital.

5601 Great Oaks Parkway
San Jose, CA 95119, USA
www.westerndigital.com

[1] One gigabyte (GB) is equal to one billion bytes and one terabyte (TB) is equal to one trillion bytes. Actual user capacity may be less due to operating environment.