Western Digital.

Technical Brief

Next-Generation Storage for VMware Environments: Virtual RAID Appliance Powered by Western Digital OpenFlex™ Data24 and xiRAID Opus

Problem Statement

With the increasing virtualization of applications in different spheres, traditional data storage systems are proving no longer suitable. These applications demand storage solutions that can offer high performance, scalability, and efficient resource utilization.

This paper introduces an innovative solution that integrates high-performance storage with virtualized environments, with a particular focus on maximizing the potential of Ethernet-attached Bunch of Flash (EBOF) storage in hypervisor ecosystems.

The proposed solution, called Virtual RAID Appliance (VRA), is a lightweight virtual machine that leverages network storage, bypassing the VMware® ESXi™ stack, to create RAID-protected virtual block devices. These volumes are then re-exported to the virtualization cluster for creating Datastores on top of them. For exceptionally fast volumes, direct connection from the virtual machine is possible.

By leveraging Western Digital's OpenFlex Data24 NVMe-oF™ Storage Platform and the high-performance xiRAID Opus software RAID engine, this solution offers:

- Performance and security. It achieves maximum storage performance with minimal resource consumption using a single virtual machine. It creates protected volumes using advanced erasure coding technology, surpassing simple data replication methods.
- Cost-effectiveness. This solution leverages existing virtual infrastructure, providing advanced features without expensive hardware upgrades.
- Compatibility. It ensures smooth operation between ESX nodes and EBOF storage across diverse environments.
- Enhancement of OpenFlex Data24. While OpenFlex Data24 provides disaggregated storage with resource utilization levels comparable to large web-scale enterprises, this solution enables secure volume creation and flexible capacity allocation.

Western Digital's OpenFlex Data24 offers Shared Storage with performance levels similar to or exceeding high-end SAN arrays, but at a significantly lower cost and with much more efficient resource allocation. The VRA builds upon this foundation, implementing classic storage services such as volume creation and striping. This paper will explore how the VRA empowers organizations to use the full potential of their storage resources.

This document is structured to provide a comprehensive overview of the proposed storage solution. It begins with a detailed technology overview, describing the specifics of Western Digital's OpenFlex Data24 and Xinnor's xiRAID Opus software RAID engine, forming the foundation of the solution. The paper then describes the deployment topology, explaining how the solution integrates within existing infrastructures. The test environment overview details the hardware components, network configuration, and software setup used for testing. This is followed by extensive performance tests, providing empirical evidence of the solution's capabilities in both random and sequential operations across various scenarios. The paper concludes with a summary of the findings and the implications for enterprise storage strategies.

OpenFlex Data24 Technology Overview

OpenFlex is Western Digital's architecture that supports OCI through storage disaggregation. The OpenFlex Data24 is a 2U rack-mounted data storage enclosure, built on the OpenFlex platform. This Just-a-Bunch-Of-Flash (JBOF) platform leverages the 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, in turn makes that storage widely available to multiple applications and hosts. For more details, refer to OpenFlex Data24 NVMe-oF Storage Platform.

Enabling applications to share a common pool of storage capacity allows data to be easily allocated and / or shared between applications whilst independent 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.

Western Digital RapidFlex™ NVMe-oF fabric adapters are used to share storage to servers using the NVMe-oF protocol, either in direct connectivity or switched topology. OpenFlex Data24 is a vertically integrated Western Digital design with NVMe SSDs, fabric adapters, and a JBOF platform architecture. Each OpenFlex Data24 chassis is scalable up to 368TB[1] of low-latency dual-port SSDs in 2U 24-bay platform.

OpenFlex Data24 3200 NVMe-oF Enclosure

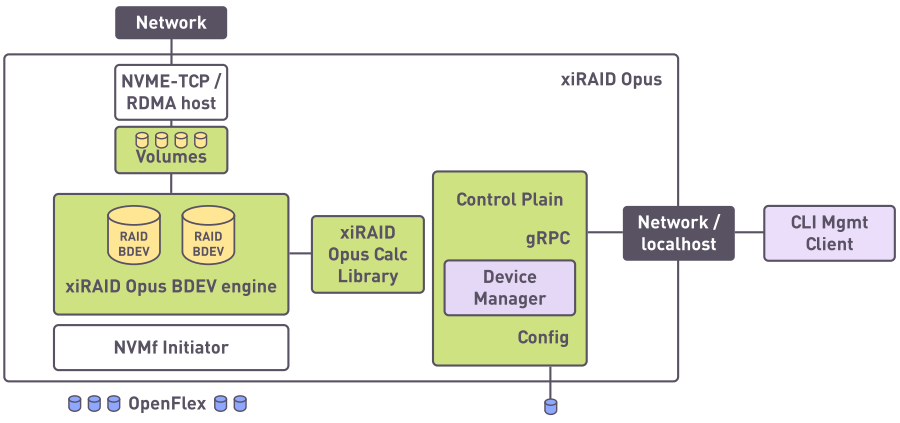
Composable Infrastructure seeks to disaggregate compute, storage, and networking fabric resources into shared resource pools that can be available for on-demand allocation (i.e., "composable"). Composability occurs at the software level, while disaggregation occurs at the hardware level using NVMe-oF.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

xiRAID Opus Overview

xiRAID Opus (Optimized Performance in User Space) is a high-performance software RAID engine operating in Linux® user space. The user-space engine enables the system to achieve high speeds and eliminates dependencies on the operating system version, drivers, and libraries.

By bypassing traditional OS-level drivers, xiRAID Opus facilitates direct interaction with hardware storage devices, resulting in more efficient I/O operations for network-connected devices. The inclusion of internal user-space drivers, specifically optimized for high performance, further enhances the efficiency of data operations.

xiRAID Opus is providing the VRA solution with a comprehensive approach to storage management. It seamlessly connects to disaggregated drives, creates secure volumes, and re-exports these volumes within the virtual infrastructure using NVMe-oF and NVMe/TCP protocols. A key advantage of this system is its unified management interface, which provides a single console and a consistent set of commands for all operations, eliminating the need for various third-party utilities and significantly streamlining storage administration.



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

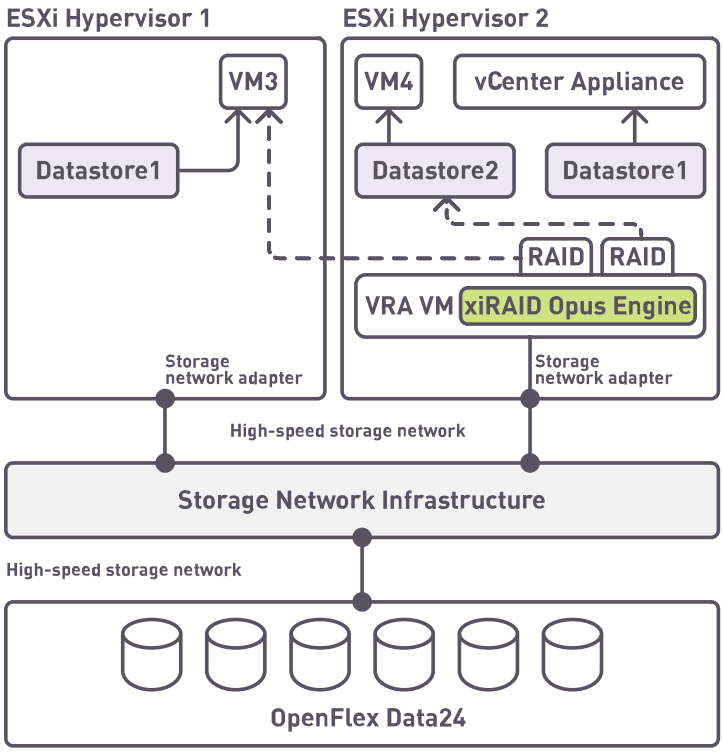
The storage system consists of the following components:

- Network-connected drives accessed in user space using an integrated NVMe-oF initiator.
- Device Manager component responsible for connecting, disconnecting, and monitoring physical storage devices and their drivers within the data path engine and OS.
- Block device (BDEV) subsystem that represents unified logical disk API for different disk types such as a backend disk, RAID, or other block device within a layered block device topology.
- RAID engine based on the fast xiRAID technology. The RAID engine implements various RAID features such as RAID structure initialization, general IO operations, degraded mode IO, disk replacement, and disk data reconstruction (including Xinnor's partial reconstruction algorithms).
- Frontend target subsystem that connects xiRAID block devices or other BDEVs configured on top of a RAID to a client IO engine.
- Configuration database that persists and reports the storage topology and its component parameters. The database enables the automatic recovery and restart of the storage system in the event of an engine, operating system, or physical node (server) restart.
- Management logic for configuring and monitoring all storage components.

All components described above (except for physical storage devices) are packed with an OS into the VRA.

Deployment Topology

Scheme below represents proposed deployment topology:



All elements required to implement solution (with all software components described in the previous section) are packed into VRA. It is important to emphasize that:

- VMware vCenter is used as a deployment environment for the VRA containing most of the software components.
- Minimal storage network infrastructure required: at least one switch to connect OpenFlex Data24 system with ESXi hypervisors.
- At least one high-speed storage network adapter is required per ESXi host, with SR-IOV support and corresponding option configured – this feature is used to allow storage network access for ESXi host, VRA and client virtual machines.
- VRA acts as a consumer of raw storage resources provided by OpenFlex Data24 system: drives are attached using NVMe-oF, configure RAIDs and provide storage resources on top of these RAIDs to the client systems.
- These entities are considered as client systems for VRA:
  - ESXi hosts (even the same host on which VRA is running): VRA-provided storage resources are attached and used as a base for creating specialized VMFS Datastore. This datastore is used as a storage for common virtual disks attachable to the virtual machines. Corresponding vSphere storage configuration should be done.
  - Specific virtual machine, which can consume VRA resources provided directly over NVMe-oF. Corresponding virtual machine network hardware configuration should be done.

Typical deployment steps are:

1. Installation of all required storage network infrastructure components, their system and network configuration.
2. Installation, connection, and configuration of the storage network adapters on the ESXi's side (and enabling SR-IOV feature, if needed).
3. VRA deployment, using common VMware tools + VRA hardware and network configuration, to provide access to the storage network.
4. VRA configuration to attach NVMe drives provided by OpenFlex Data24, to configure RAID(s) on top of them, and to publish these new storage resources for clients.
5. Client system configuration (vSphere or specific virtual machine) to consume storage resources published by VRA.

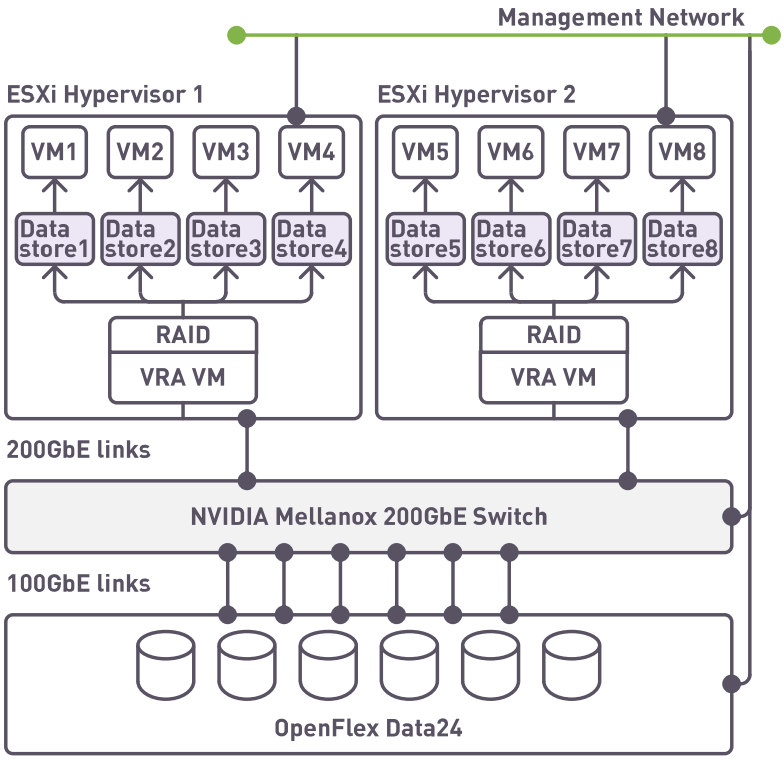
In the next section you can see a more specific environment description and a set of tests done for it.

Test Environment Overview

The test environment contains the following hardware components:

- NVIDIA® Mellanox® Spectrum®-3 SN4600 200GbE Ethernet Switch.
- Western Digital OpenFlex Data24-3200 NVMe-oF Storage Platform with a set of Western Digital Ultrastar® DC SN840 drives, 24 devices in total, 7.68 TB each.
  - Platform includes 6 network interfaces, all of them connected to the NVIDIA Mellanox switch (each link is 100 Gb/s) and 2 management network interfaces.
- Two Dell® PowerEdge® R650 Servers, which serve as ESXi hosts. Each server has:
  - Intel® Xeon® Gold 6454 CPU, 18 cores, 36 threads
  - 512 GiB RAM.
  - 1.5 TB Samsung NVMe PM1733 NVMe drive.
  - Mellanox Connect®X-6 200GbE network adapter, connected to the ethernet switch.
  - Network interface used for management purposes.

Here is a connection scheme for the hardware components described above and some required software-based components. This environment corresponds to the deployment scheme described in the previous section.



Software configuration of the test environment:

- For NVIDIA Mellanox Ethernet Switch:
  - Switch uses Cumulus Linux 5.6.0 as an operation system.
  - Using vendor's recommended procedure, all network interfaces used in the environment are added to the single untagged VLAN, additional configuration parameters are applied to the VLAN:
    - Roce enabled, mode: lossless
- For Western Digital OpenFlex Data24 Storage Platform:
  - 100 Gb/s IPv4 network is configured for each network interface. MTU used is 5000.
- For each Dell server:
  - VMware ESXi 8 Update 1 hypervisor is installed on the server.
  - Storage and management networks are configured, vendor's procedure used to install hardware drivers and to enable network interface's SR-IOV feature.
  - VMware vCenter® 8.0.2 is installed on one of the ESXi hosts.
  - ESXi hosts are added to the vCenter inventory.
  - Two VRA instances are deployed to the vCenter/ESXi environment (one instance for each host).
  - A set of test client virtual machines (Ubuntu Server 22.04.2) is deployed to the vCenter/ESXi environment.
  - Using vCenter tools, ESXi storage subsystem is configured to allow 200GbE network interface usage as SR-IOV device and/or host network interface.

Performance Tests

The main objectives of performance testing were to evaluate:

- the scalability of performance when increasing the number of virtual machines under loads with random pattern
- the full saturation of the network channel under loads with sequential pattern

Two virtual machines with installed and configured VRA were created on each ESXi host. For the random operations tests, configuration of 8 drives was used. For the sequential operations tests, 9 drives were used to ensure the I/O size matched the stripe width.

Each RAID was presented targets to the corresponding ESXi host as 4 NVMe-oF RDMA. VMFS6 datastore was created for each NVMe-oF RDMA target.

For load tests with a random pattern, 8 virtual machines were created, 4 on each ESXi host. For load tests with a sequential pattern, 2 virtual machines were created, 1 on each host. All virtual machines have the same configuration. Each virtual machine was connected to a virtual disk created on the dedicated VMFS Datastore on the NVMe-oF RDMA target.

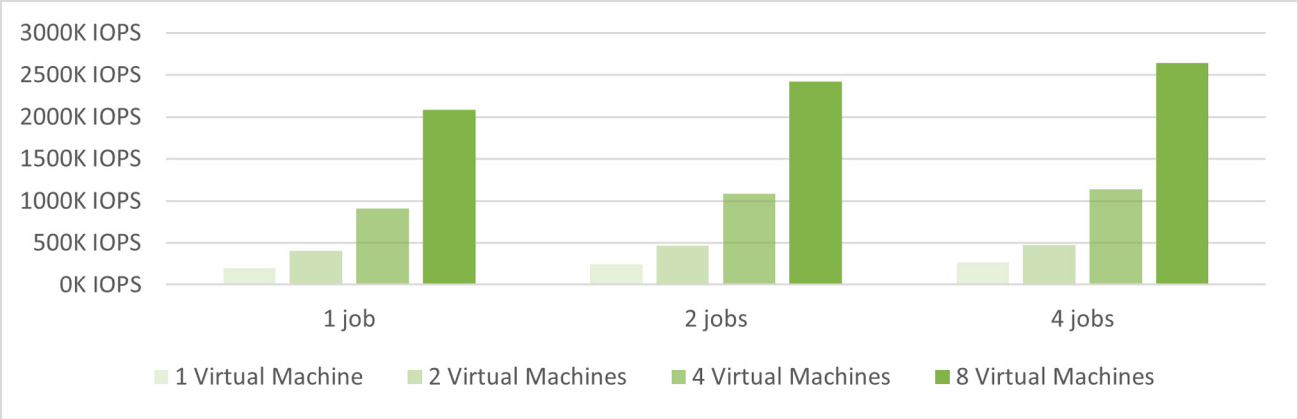
A typical configuration of virtual machines is given in Appendix 1.

The FIO utility was used for performance testing. Typical FIO configurations are given in Appendix 2.

Random Operations

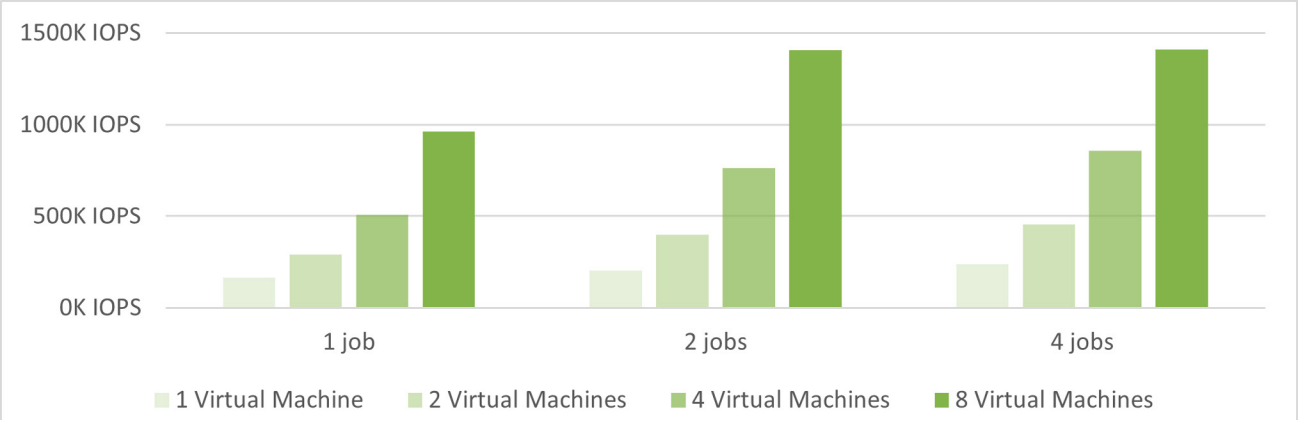
Performance tests using 8 drives were performed sequentially in series with a step-by-step increase of the job number with values of 1, 2, and 4. Each series of tests was performed sequentially on one, two, four, and eight virtual machines simultaneously. For the test series with one and two virtual machines, one ESXi host was used, for the test series with four and eight virtual machines, two ESXi hosts were used with an even distribution of virtual machines.

Random Read Tests



	1 job, K IOPS	2 jobs, K IOPS	4 jobs, K IOPS
1 Virtual Machine	205	250	271
2 Virtual Machine	405	471	473
4 Virtual Machine	917	1091	1140
8 Virtual Machine	2089	2422	2650

Random Write Tests



	1 job, K IOPS	2 jobs, K IOPS	4 jobs, K IOPS
1 Virtual Machine	168	205	240
2 Virtual Machine	293	400	455
4 Virtual Machine	505	762	857
8 Virtual Machine	965	1407	1410

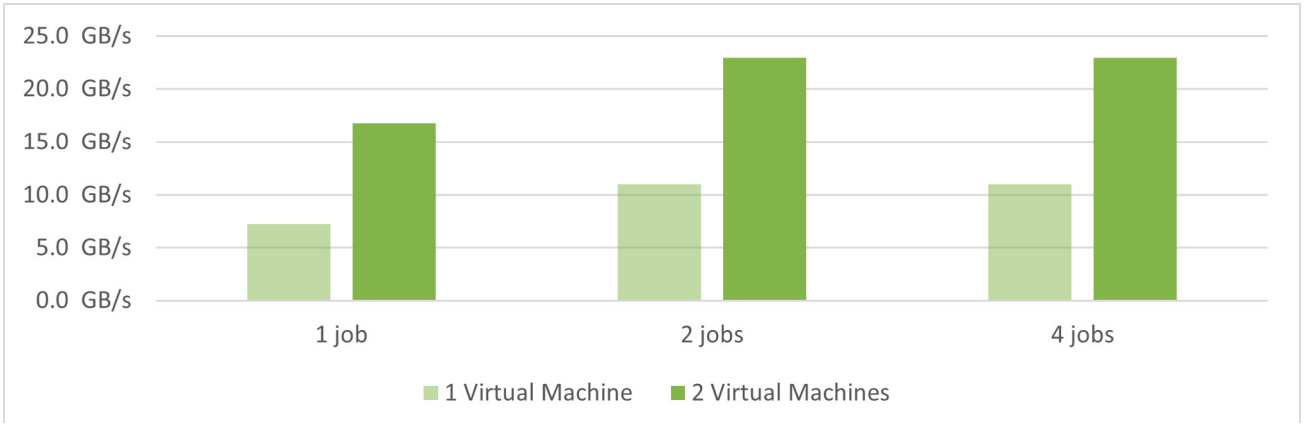
The performance test results demonstrate impressive scalability. For random read operations, the system exhibits near-linear scaling as the number of virtual machines increases, reaching up to 2,650K IOPS with 8 VMs and 4 jobs. The results demonstrate reaching the performance scaling limits of the virtual machines.

Random write operations also show significant performance improvements with additional VMs, peaking at 1,410K IOPS. Here the maximum possible performance for eight drives with RAID 5 penalty is achieved.

Sequential Operations

Performance tests using 9 drives were performed sequentially in series with a step-by-step increase of the job number with values of 1, 2, and 4. Each series of tests was performed sequentially on one and two virtual machines simultaneously. For the test series with one virtual machine, one ESXi host was used, for the test series with two virtual machines, two ESXi hosts were used.

Sequential Read Tests



	1 job	2 jobs	4 jobs
1 Virtual Machine	7.1 GB/s	10.7 GB/s	11.0 GB/s
2 Virtual Machine	16.3 GB/s	21.8 GB/s	22.0 GB/s

For both read and write operations, performance scales significantly when moving from 1 to 2 virtual machines. This near-doubling of performance when using two VMs indicates excellent scalability across multiple ESXi hosts.

The results also show that increasing the number of jobs from 1 to 2 provides substantial performance gains for both read and write operations. However, further increasing to 4 jobs offers minimal to no additional improvement in most cases, which proves achieving maximum performance levels potentially available with two 100 Gigabit interfaces.

Conclusions

The comprehensive performance tests conducted on the Virtual RAID Appliance solution, integrated with Western Digital's OpenFlex Data24 NVMe-oF Storage Platform, validate the solution's ability to efficiently manage and scale storage resources in VMware virtualized environments while maintaining high performance.

Both random and sequential I/O tests exhibited excellent scalability as the number of virtual machines increased across multiple ESXi hosts. This scalability directly addresses the need for efficient resource management as storage demands grow. The performance metrics underscore the solution's success in maximizing storage performance with minimal resource consumption, seamlessly integrating with existing virtualized infrastructures.

The test results validate the system's capacity to overcome the compatibility issues between hypervisor nodes and Ethernet-attached Bunch of Flash (EBOF) storage, while also providing advanced storage services often lacking in EBOF systems. By offering high-performance storage capabilities through a virtual appliance, the solution delivers enterprise-grade features without requiring costly hardware upgrades, instead utilizing existing virtual infrastructure.

Overall, the Virtual RAID Appliance for VMware, based on xiRAID Opus and coupled with OpenFlex Data24, reaches performance levels of several million IOPS for both read and write operations, fully utilizing 100 Gigabit interfaces when using only eight drives. The solution's scalability is a key advantage, allowing to enhance performance by deploying additional VRAs with minimal resource requirements. This approach enables to achieve performance levels comparable to high-end flash arrays, leveraging external EBOF and utilizing server resources efficiently. This solution optimizes IT operations and infrastructure investments, providing a cost-effective approach to storage management in modern data centers.

Typical Configuration of Virtual Machines

- CPU: 8 CPU(s)
- Memory: 8 GiB
- Network adapter: 1 VM Network
- Compatibility: ESXi 8.0 and later (VM version 20)
- System Hard Disk: 40 GB Thick Provision Lazy Zeroed
- OS: Ubuntu 22.04.4 LTS
- Kernel: 5.15.0-113-generic

FIO Utility Configurations

Random Loads

```
[global]
ioengine=libaio
direct=1
runtime=600
group _ reporting
numjobs=1
iodepth=32
bs=4k
norandommap
gtod _ reduce=1
rw=randread/randwrite
[vda]
filename=/dev/sdb
```

Sequential Loads

```
[global]
ioengine=libaio
direct=1
runtime=600
group _ reporting
numjobs=1
iodepth=32
bs=1M
rw=read/write
ioffset _ increment=2%
[vda]
filename=/dev/sdb
```

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

February 2025 / Revision 02