RAIDIX ERA Software RAID with Ultrastar® DC SN640 NVMe™ SSDs

Introduction

NVMe™ SSDs are seeing an ever-increasing adoption rate in public cloud and on-premise infrastructure. NVMe SSDs are optimized for flash memory providing lower latencies and higher data bandwidth versus legacy SAS and SATA SSDs which are based on rotating media protocols and are bottlenecked with system level hubs. Western Digital's Ultrastar DC SN640 NVMe SSDs have 6x the read performance of SATA SSDs, providing for better TCO and higher performance.

The transition to all NVMe storage requires newer hardware RAID controllers designed to provide easy drive management, reliable drive redundancy and fast rebuild times all while minimizing the impact to the NVMe SSDs native high performance and low latencies. Both hardware (HW) and software (SW) RAID options are available for NVMe systems. This technical brief will illustrate the advantages and benefits of deploying software NVMe RAID with RAIDIX ERA.

RAIDIX ERA for NVMe

Compared to the established ecosystem of SATA and SAS RAID options, working with NVMe is not as straightforward. SW RAID options have been considered inferior to hardware in terms of performance and additional CPU cycle requirements and a lack of features. However, SW RAID makes up for that with flexibility, zero associated hardware costs and being vendor-agnostic in terms of compatibility.

RAIDIX ERA is a software RAID presented by Linux® kernel module and management utility. The core features include innovative RAID technologies which provide stable performance for mixed data workloads. With I/O handling parallelization and lockless data path, ERA maximizes hardware capabilities to help achieve the highest possible performance. ERA supports a wide range of RAID levels and is lightweight with low requirements for CPU and RAM usage.

Benchmarking RAIDIX ERA

Western Digital benchmarked RAIDIX ERA performance, with Ultrastar DC SN640 NVMe SSDs, in a classic RAID-5 8 drive scenario against a baseline of 8 drives with no RAID. These drives present mainstream NVMe performance with low latency and are positioned as a replacement to SATA SSDs. System HW included AMD's EPYC™ 7702P with 128 PCIe lanes per socket and 64 high performance cores, and 128GiB DDR4 RAM. The single AMD EPYC CPU can process heavy I/O applications without oversubscription or the need for a PCIe switch. Oracle® Linux 8.3 with Unbreakable Enterprise Kernel UEKR6 was used as the OS.

Western Digital followed the SNIA PTS methodology for single-drive performance evaluation. Steps were taken to measure unbuffered steady-state performance of the arrays. Sequential, random and mixed workloads were measured. All tests were performed on a single RAID-5 volume consisting of 8 drives to represent an average usage scenario. Minimal system tuning was performed as advised by various NVMe performance guides and ERA and MD RAID user guides:

BIOS:

```
System C-states disabled
PCIe Maximum Payload set to 256b
PCIe Maximum Read Request set to 4096b
```

OS:

```
tuned-adm profile througput-performance
modprobe -r nvme && modprobe nvme poll _
queues=36
echo 0 | tee
/sys/block/nvme*n1/queue/io _ poll _ delay
echo 0 | tee
/sys/block/nvme*n1/queue/iostats
```

ERA RAID tuning for sequential writes (set me=0 to lower random latency):

```
eraraid modify -me=1 --merge-wait=10000
```

RAID creation commands:

```
mdadm --create --chunk 64K /dev/md1 --level=5
--raid-devices=8 /dev/nvme{0..7}n1
eraraid create -n wdcera -l 5 -ss 64 -d /
devnvme{0..7}n1
```

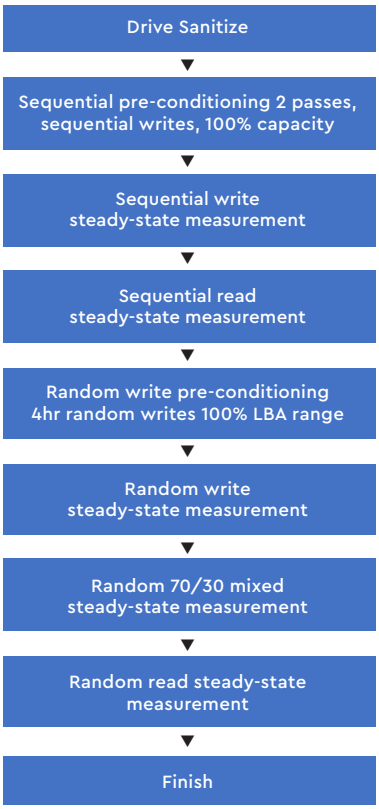
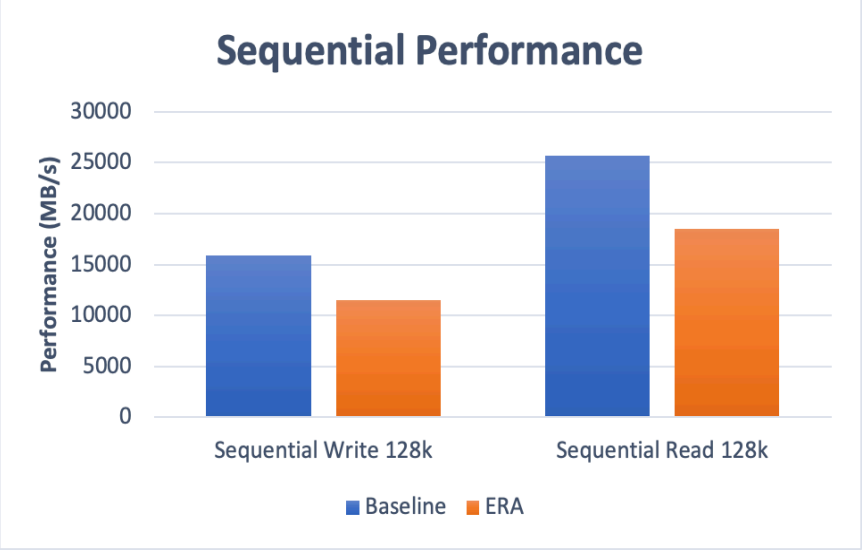


Figure 1 Benchmark Process Flow

Performance

RAID-5 volumes for RAIDIX ERA effectively use 7 drives with 1 strip used for parity. These are benchmarked against a baseline of 8 drives without RAID.

The ERA RAID solution showed solid performance for sequential, random and mixed workloads. Sequential performance for the ERA solution was approximately 82% of the baseline. Random read performance of ERA was 90% of the baseline, while random write performance was 50% of the baseline. RAID-5 layout penalizes random writes due to an increased number of I/O operations required, and usually for such workloads RAID-5 performance is approximated as 25% of raw device random write performance. The ERA solution combined with the asymmetric nature of SSD performance has demonstrated a significant increase in random write IO/s, making RAID-5 a viable choice even for medium write-intensive workloads.



Compared to the raw device baseline there is a 2x penalty on writes which is respectable for RAID-5. Usually for writes, a 4x penalty is used for RAID-5 performance approximation, but with the asymmetric nature of SSD performance and ERA optimizations, it is only 2x, making RAID-5 a viable choice even for medium write-intensive workloads. Read performance of ERA is very high, reaching more than 90% of raw device performance.

20 drive RAID-5 performance

To test the limits of hardware and software combination, it was decided to run an IOPS test on a larger drive population. An ERA RAID-5 volume was created consisting of 20 Ultrastar DC SN640 drives.
After proper pre-conditioning, the resulting performance was as follows:

```
Random writes at 128 jobs / 32 iodepth: 1.4M IOPS
Random 70/30 mix at same settings: 3.75 M IOPS
Random reads: 10.3M IOPS
```

Conclusion

Western Digital conducted benchmarking that demonstrated RAIDIX ERA software RAID with the Western Digital DC SN640 NVMe SSDs is a viable solution for data protection. The ERA solution can support the performance boost that NVMe brings to the enterprise workloads and should be viewed as a step-up from SATA and SAS RAID configurations. A basic RAID-5 is demonstrated with great performance and low rebuild times. Other options like RAID-6, 50, 60 and RAIDIX unique RAID 7.3 could offer an even better experience at the cost of extra parity drives.

Western Digital.

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