Experiences in testing Oracle Database with NVMe Storage

Bhargavaram Akula
IBM Systems Group
August 2020

Table of contents

Abstract

This paper describes the findings of NVMe (Non-Volatile Memory express) storage testing with an Oracle Database 19c standalone database instance on an IBM Power® Systems server. We also show a comparison of performance test results gathered when using NVMe storage and IBM FlashSystem™ storage.

Introduction

IBM® Non-Volatile Memory express was developed as an industry specification for accessing non-volatile storage via the PCIe interface. The NVMe specification capitalizes on the internal characteristics of flash storage and enables storage accesses with low latency, high efficiency, and high scalability.

NVMe, like SATA or USB, allows for multiple vendors to develop products compliant with the specification which are all supported by the same host device driver, therefore removing software compatibility as an adoption inhibitor.

The key areas of improvement in the NVMe specification are:

- Increased queue depth
- Reduced register access per command
- Lightweight protocol requiring minimal path length
- Multiple MSI-X supported

In this paper we will show the results of testing Oracle Database 19c with NVMe storage and we'll compare the results with those gathered using IBM FlashSystem 840 storage. We performed testing using the Orion tool and other Oracle Database workload generators.

System topology

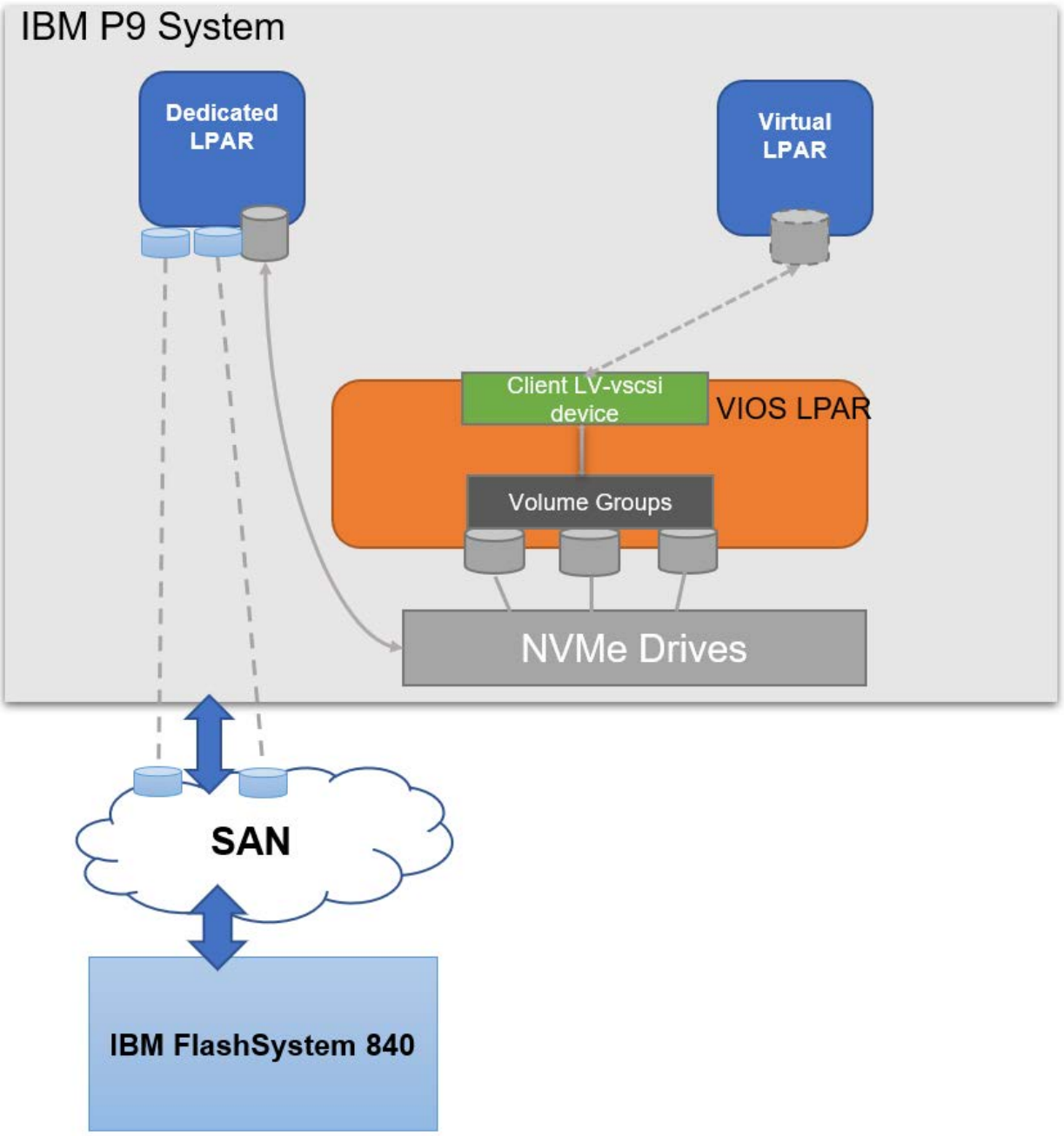

Figure 1. System Configuration

An IBM Power E980 system with 40 CPUs and 2TB memory was used for the testing. Two Logical Partitions (LPAR) were created, one is a dedicated I/O LPAR with 16 CPUs and other is a virtual I/O LPAR with 16 CPUs where NVMe devices are assigned from a VIOS partition. For the dedicated I/O LPAR IBM FlashSystem™ 840 volumes were attached via a SAN network and a ~2.9TB dedicated NVMe drive was added. On the VIOS server, logical volumes were created from the NVMe disks volume group

and these logical volumes backed by virtual SCSI devices were assigned to the virtual I/O LPAR. Various Oracle Database workload tests were performed using these logical partitions.

Orion test results

Orion is a I/O calibration tool provided by Oracle along with the Oracle Database software. Orion simulates Oracle Database I/O workloads using the same I/O software stack as the Oracle Database. For more information regarding the Orion tool refer to:

https://docs.oracle.com/en/database/oracle/oracle-database/19/tgdba/IO-configuration-and-design.html

The Orion tool testing was done using the dedicated I/O LPAR to which both NVMe and Flash LUNs were attached. For this testing we used 2.9TB NVMe disk and 256GB IBM FlashSystem™ 840 SAN volumes. By default, the NVMe disk sector size is 4K and IBM FlashSystem™ 840 disk sector size is 512B. We can also create 4K sector size flash LUNs on the IBM FlashSystem™ 840. The Orion testing was performed on NVMe with 4K sector size and on Flash LUNs with 512B sector size and 4K sector size.

The following Orion test tool commands were used for testing:

1) <ORACLE_HOME>/bin/orion -run dss
2) <ORACLE_HOME>/bin/orion -run oltp
3) <ORACLE_HOME>/bin/orion -run advanced -testname R70W30 -size_small 8 -num_large 0 -write 30 -matrix row -verbose
4) <ORACLE_HOME>/bin/orion -run advanced -testname R80W20 -size_small 8 -num_large 0 -write 20 -matrix row -verbose

Note: Flash 512Byte sector size data is taken as the baseline.

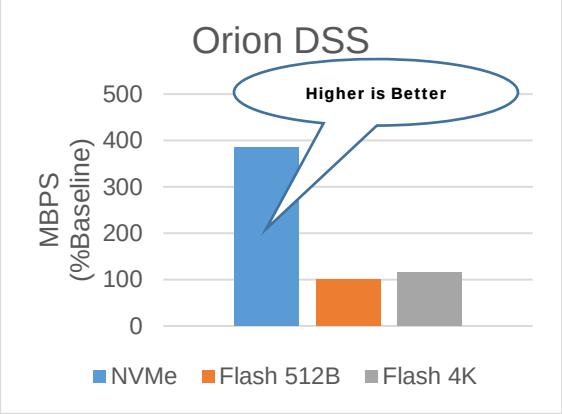

Figure 2

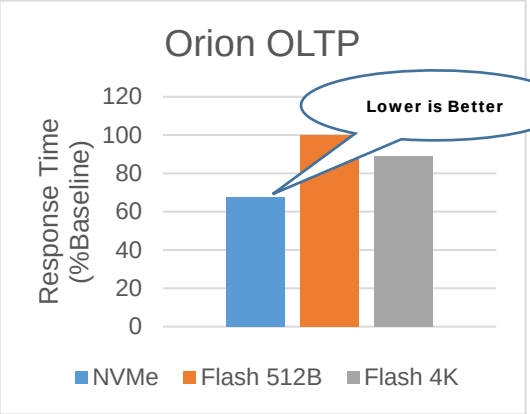

Figure 3

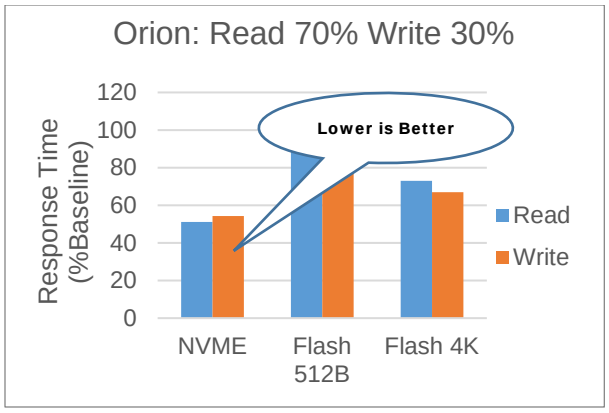

Figure 4

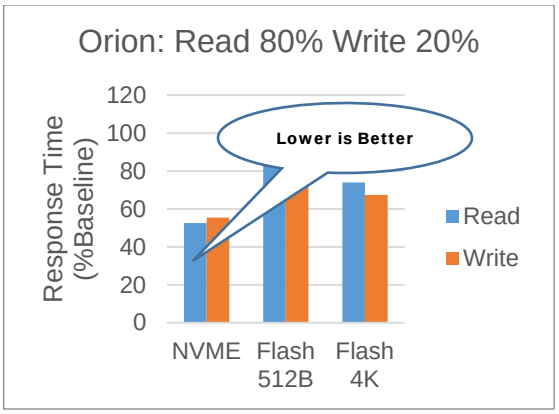

Figure 5

The results shown in the figures above clearly show that NVMe storage provides low latencies and high bandwidth when compared with flash volumes.

NVMe dedicated and via VIOS

An OLTP workload was used for testing the NVMe storage attached with a dedicated I/O LPAR and virtual I/O LPAR. This workload simulates a brokerage firm with customers who generate transactions related to trades, account inquiries, and market research. The brokerage firm in turn interacts with financial markets to execute orders on behalf of their customers and updates relevant information.

In our testing we have created an OLTP database with 50,000 customers and a database size of ~1TB. An Oracle Database 19c single instance was created on both LPARs, each having an SGA size of 350GB. We have used a transaction mix that would generate a reasonable I/O load.

The chart below shows the amount of read/write activity that happened during the workload run for one user. The maximum bandwidth value is taken as baseline and percentages are calculated for remaining time period. During the warmup time we can see more reads were happening and writes were increasing slowly.

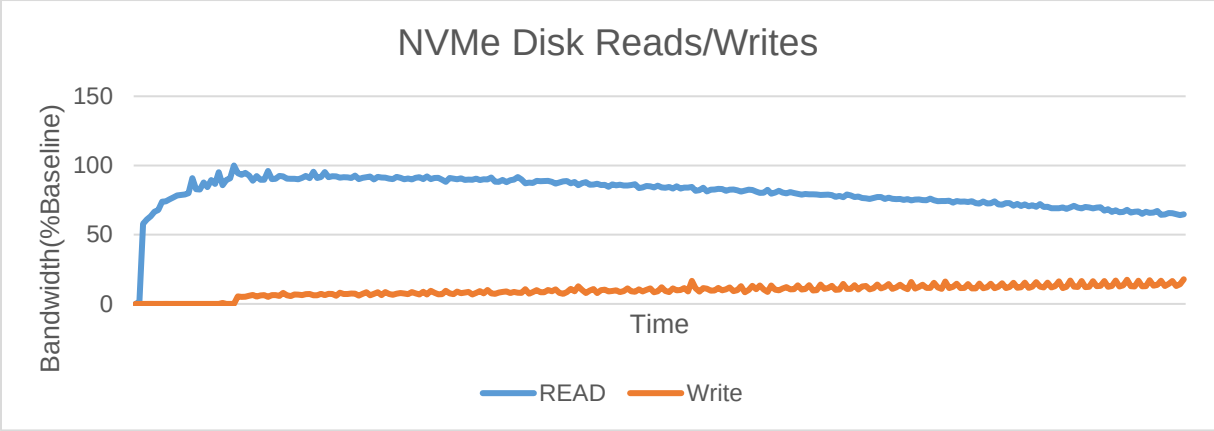

Figure 6. OLTP - Read/Write Bandwidth

The following chart shows the throughput (TPS) percentage for 1 to 5 users on the dedicated I/O LPAR and virtual IO LPAR. Here, the one user TPS output on virtual I/O LPAR is taken as the baseline. The results show there is a slight improvement in throughput when we use dedicated NVMe disk compared to virtual. Since we have used an SGA size as 350GB for the 5 user run, more data got populated in the database cache and this in turn reduced the number of physical I/Os. Therefore, we have observed a small difference in TPS between the dedicate I/O LPAR and virtual I/O LPAR as we increase the number of users.

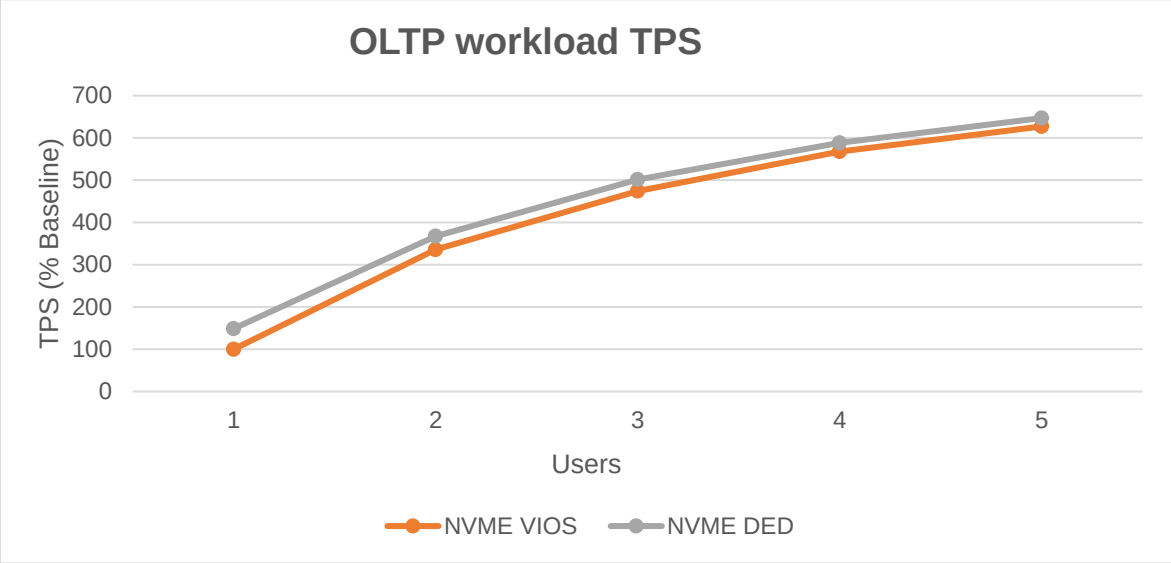

Figure 7. OLTP – Throughput (TPS)

Note: On AIX® for NVMe the disk iostat tool currently does not provide data for read service time, write service time and queue service time sections.

```
Adapter:                    xfers             time
-------------------- ------------------------- ---------
                       bps    tps  bread  bwrtn
nvme1                41.0M 3734.5  407.1K  40.6M  04:08:26

Disks:                         xfers                           read                          write                          queue
-------------------- ------------------------------ ------------------------------ ------------------------------ ------------------------------
                     %tm    bps   tps  bread  bwrtn   rps   avg   min   max time fail   wps   avg   min   max time fail   avg   min   max   avg   a
                     act                               serv  serv  serv outs            serv  serv  serv outs       time  time  time  wqsz  sq
hdisk0               8.2  41.0M 3734.5 407.1K  40.6M
```

NVMe and Flash storage - performance

In this section we have performed Oracle Database workload testing on both NVMe disk and flash LUNs. NVMe disks by default have a sector size of 4K while the IBM FlashSystem 840 volumes have a default sector size of 512Byte. The FlashSystem 840 also provides an option for creating 4K sector volumes. We did a series of tests with NVMe and flash LUNs using an Oracle Database generated I/O workload. We have created four 256GB flash 512B sector size LUNs and four 256GB flash 4k sector size LUNs, and assigned those LUNs to the dedicated I/O LPAR. An I/O based database named as IOP was created with 25 warehouses, the sga_target was tuned to have more physical reads. Tests were performed by varying

workload users from 5 to 40. The oracle datafiles were moved between the NVMe diskgroup and the flash LUNs ASM diskgroup.

Figure 8 below shows the Read Bandwidth that was recorded when using NVMe disk, flash 512B LUNs and flash 4k LUNs for the 40 concurrent users workload run. Here the Read Bandwidth data on flash 512B LUN is taken as the baseline and the percentage is calculated for the other drives tested. NVMe storage on average delivered ~130% more throughput compared to the baseline. Clearly, we can see that the NVMe disk performs better at reads when compared to flash volumes.

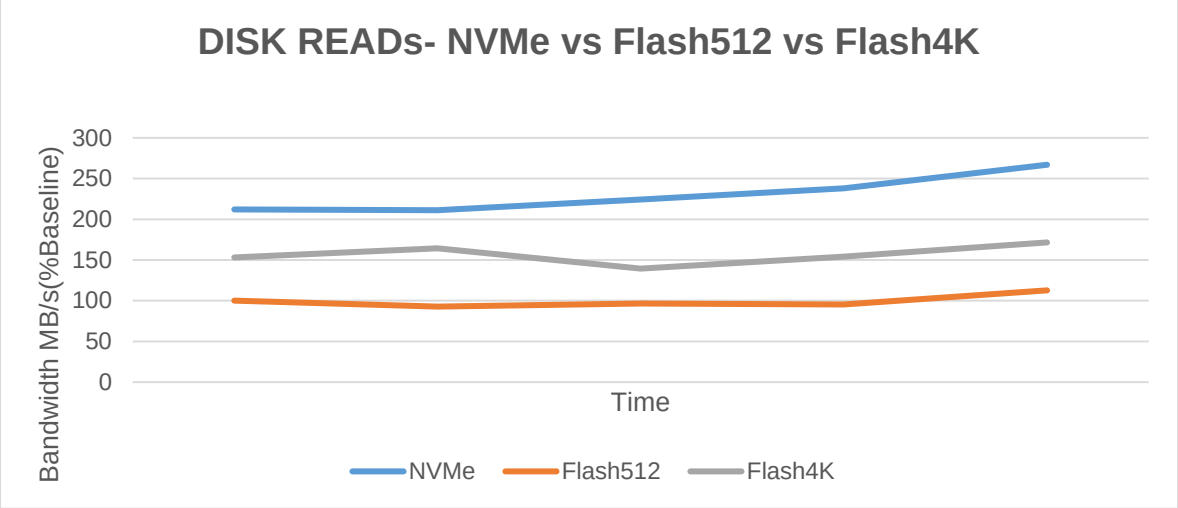

Figure 8. Read Bandwidth

In Figure 9 below, we show the Write Bandwidth that was recorded on NVMe disk, flash 512B LUNs and flash 4k LUNs for the 40 concurrent users workload run. NVMe on an average delivered ~50% more throughput when compared to the baseline. Clearly, we can see that the NVMe disk performs better writes when compared to flash volumes.

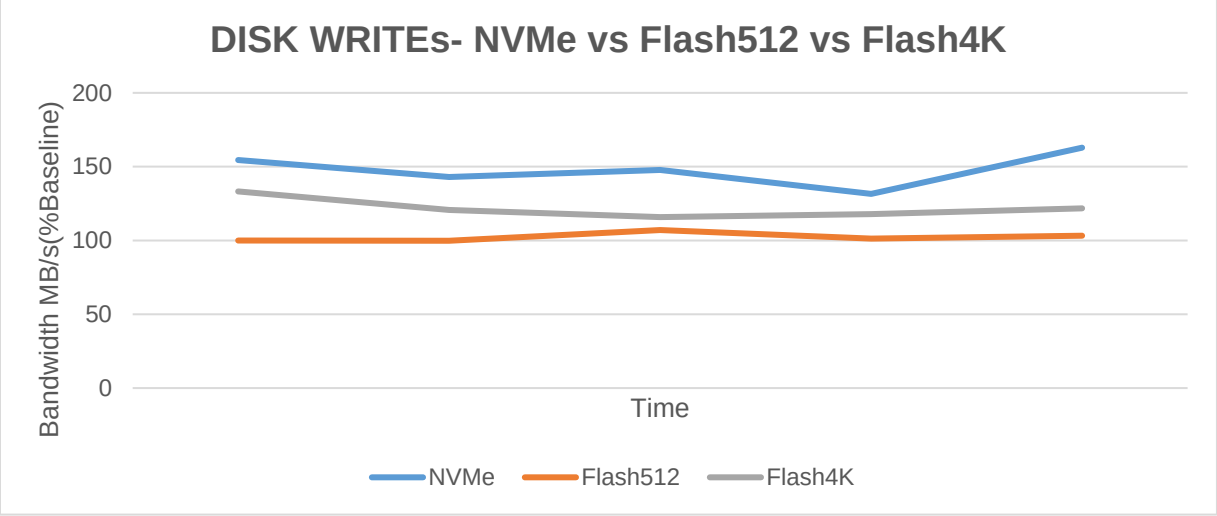

Figure 9. Write Bandwidth

The Oracle Database I/O workload test was an on NVMe and flash volumes by varying workload users from 5 to 40. Figure 10 below shows the throughput (TPS)% distribution when the IOP database is on NVMe and flash storage. Here the FlashSystem 5 user run TPS is taken as the baseline and the percentage is calculated for the other drives accordingly.

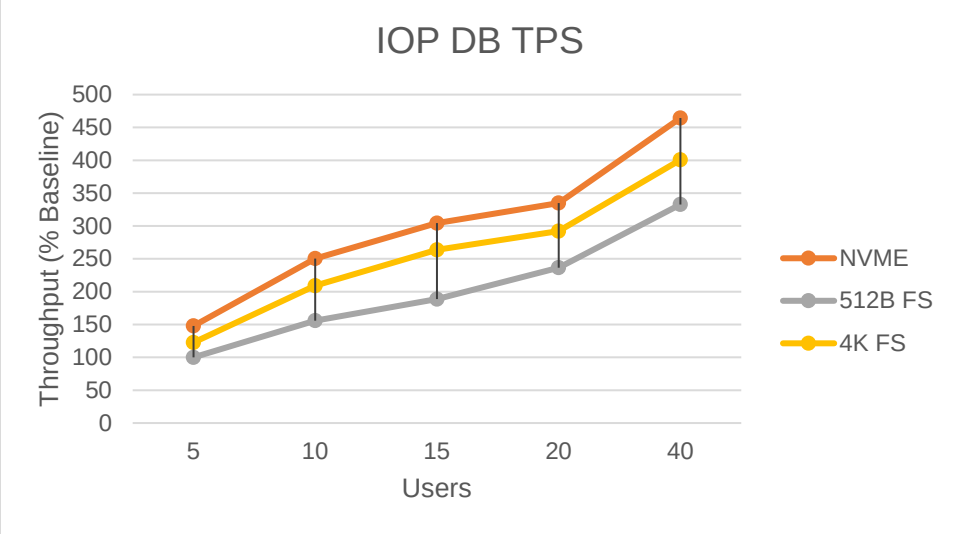

Figure 10 IOPDB- Throughput (TPS)

For five workload users run when using NVMe storage we see ~50% improvement in throughput (TPS) when compared to flash 512 byte sector storage run. From Figure 10 it clearly shows that when the database is stored on NVMe storage the measured TPS (transactions per second) are better when compared to when the database is stored on flash volumes.

To avoid single point of failure when using NVMe storage for the Oracle Database you can use application level high availability techniques such as Oracle Data Guard.

At host level you can use:

- VG Mirroring
- RAID 1/10 configuration

Conclusion

In this document we have tested an Oracle Database 19c instance on NVMe storage with a dedicated I/O LPAR and a virtual I/O LPAR. The dedicated LPAR with NVMe drives attached performed slightly better when compared to the VIOS-based drive attachment. We also tested an Oracle Database workload with NVMe SSD and FlashSystem SSDs. The NVMe disks delivered lower read and write latencies and higher efficiency when compared SAN based flash LUNs. Thus we have demonstrated that the NVMe interface enables storage accesses with low latency and high efficiency with improved database performance.

Resources

These Web sites provide useful references to supplement the information contained in this document:

- IBM Knowledgecenter – NVMe subsystem
  https://www.ibm.com/support/knowledgecenter/ssw_aix_72/kerneltechref/nvme.html

- Oracle Database 19c on AIX
  https://docs.oracle.com/en/database/oracle/oracle-database/19/axdbi/index.html

- IBM eServer pSeries [System p] Information Center
  http://publib.boulder.ibm.com/infocenter/pseries/index.jsp

- IBM Publications Center
  http://www.elink.ibmlink.ibm.com/public/applications/publications/cgibin/pbi.cgi?CTY=US

- IBM Redbook – IBM Storage NVMe
  https://www.redbooks.ibm.com/redpapers/pdfs/redp5437.pdf

About the authors

Bhargavaram Akula is a Technical Consultant with IBM India, Hyderabad. He collaborates with the specialists at the IBM Oracle International Competency Center based in Foster City, and Redwood Shores, California, US, working on Oracle product certifications on the IBM Power AIX platform. He has extensive experience on Oracle Products. He can be reached at bhargaku@in.ibm.com.

Wayne Martin is the IBM Systems and Technology Group Technology Solutions Manager responsible for the technology relationship between IBM and the developers of Oracle Corporation Database and Fusion Middleware for all IBM server brands. His responsibilities include driving mutual understanding between IBM and Oracle on technology innovations that can generate benefits for mutual customers, managing the process of getting that technology implemented in products, and ensuring that availability of the products to customers is timely. Wayne has held a variety of technical and management roles at IBM that have focused on driving enhancements of ISV software that uses IBM mainframe, workstation, and scalable parallel products. He can be reached at wmartin@us.ibm.com.

Trademarks and special notices

References in this document to IBM products or services do not imply that IBM intends to make them available in every country.

IBM, the IBM logo and AIX, FlashSystem, and Power are trademarks or registered trademarks of International Business Machines Corporation in the United States, other countries, or both.

Other company, product, or service names may be trademarks or service marks of others.

The information provided in this document is distributed “AS IS” without any warranty, either express or implied.

The information in this document may include technical inaccuracies or typographical errors.

All customer examples described are presented as illustrations of how those customers have used IBM products and the results they may have achieved. Actual environmental costs and performance characteristics may vary by customer.

Information concerning non-IBM products was obtained from a supplier of these products, published announcement material, or other publicly available sources and does not constitute an endorsement of such products by IBM. Sources for non-IBM list prices and performance numbers are taken from publicly available information, including vendor announcements and vendor worldwide homepages. IBM has not tested these products and cannot confirm the accuracy of performance, capability, or any other claims related to non-IBM products. Questions on the capability of non-IBM products should be addressed to the supplier of those products.

All statements regarding IBM future direction and intent are subject to change or withdrawal without notice, and represent goals and objectives only. Contact your local IBM office or IBM authorized reseller for the full text of the specific Statement of Direction.

Some information addresses anticipated future capabilities. Such information is not intended as a definitive statement of a commitment to specific levels of performance, function or delivery schedules with respect to any future products. Such commitments are only made in IBM product announcements. The information is presented here to communicate IBM's current investment and development activities as a good faith effort to help with our customers' future planning.

Performance is based on measurements and projections using standard IBM benchmarks in a controlled environment. The actual throughput or performance that any user will experience will vary depending upon considerations such as the amount of multiprogramming in the user's job stream, the I/O configuration, the storage configuration, and the workload processed. Therefore, no assurance can be given that an individual user will achieve throughput or performance improvements equivalent to the ratios stated here.

Photographs shown are of engineering prototypes. Changes may be incorporated in production models.

Any references in this information to non-IBM Web sites are provided for convenience only and do not in any manner serve as an endorsement of those Web sites. The materials at those Web sites are not part of the materials for this IBM product and use of those Web sites is at your own risk.