Optimizing Quality of Service with SAP HANA on Power Rapid Cold Start

How SAP HANA on Power with Rapid Cold Start helps clients quickly restore business-critical operations

Contents

In a world where success can be measured in milliseconds, it's imperative that IT departments deliver uninterrupted service to their business users. But they also need to ensure the systems they manage are up-to-date, rapid and secure. It's a dilemma many IT professionals grapple with: How do you install the upgrades required to optimize performance without disrupting the service you provide to your constituents?

Whether you're performing scheduled maintenance, or are dealing with the very rare occurrence of an unplanned interruption, your objective is always to get business users back online and working productively in the fastest possible time.

1. About this document

This paper presents an evaluation of how the SAP HANA in-memory database improves cold start performance and Quality of Service (QoS) on IBM® Power Systems™.

We examine how the IBM Non-Volatile Memory Express (NVMe) adapter improves the performance ramp-up time after a SAP HANA DB (re-) activation. We also explore the advantages that IBM NVMe provides in terms of QoS and IT costs.

By studying use cases and performance results, we analyze the performance of SAP HANA High Availability (HA) in accelerating cold starts to ensure business-critical systems are available to users as quickly as possible after any planned – or unplanned – interruption.

1.1 Introduction

SAP HANA is an innovative, in-memory database and data management platform that stores all relevant data in main memory. This method of storage enables it to provide significantly accelerated data processing operations, ensuring timely, accurate analyses of enormous volumes of data. The result is a single, reliable version of the truth that business users can depend upon to make intelligent decisions.

SAP HANA is fully supported on IBM Power Systems, utilizing the SUSE Linux operating system, to provide flexibility, resiliency and performance.

1.2 SAP HANA High Availability (HA)

A valuable feature of SAP HANA is its performance in the event of a service interruption.

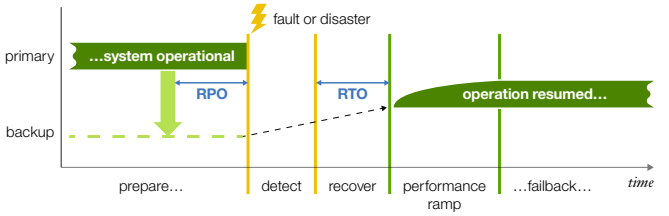


Figure 1: Phases of database failure and recovery. **Source:** "Introduction of High Availability for SAP HANA," scn.sap.com/docs/DOC-65585, January 2016. SAP.

SAP HANA uses business data persisted on non-volatile storage devices to restore data into volatile memory in the event of planned or unplanned server offline times. Data can also be persisted to NVMe cache as warm or hot data in SAP HANA. Thus, SAP HANA provides a comprehensive fault and disaster recovery solution and high availability for business continuity. IT professionals will find this capability of particular interest because mission critical systems require high availability, and downtime is no longer considered an option.

Recovery Point Objective (RPO) and Recovery Time Objective (RTO) are the key measure of recovery parameters. The Quality of Service Time (QoST) metric is also of critical importance for very large in-memory databases. Figure 1 shows the phases of high availability.

Fast recovery and QoS is a function of performance ramp-up time, which increases linearly with database size. This is because performance ramp-up time depends on the amount of data that has to be relocated from persistent storage into the memory of the now-activated standby server. For large databases, this calls for the use of a high-end back-end Storage Area Network (SAN) to facilitate fast data loads.

Consider a scale-up system with a nine terabyte (TB) database. If we need a minimum 80 percent of the database in memory to achieve desired QoS, a system connected to backend flash through two 2-port 16 Gigabit/second (Gb/s) adapters would theoretically require this approximate time:

$$T = \frac{0.8 \times 9 \times 1024\ (GB)}{6\ GB/s} = \sim 20\ minutes$$

The time required increases linearly with growing database volume. IBM Power Systems with IBM NVMe technology improve database performance ramp-up time and your ability to respond to business QoS requirements.

We have seen customer environments with a mid-range storage environment providing ~3GB/s[1] read bandwidth which would need much higher time to load the database.

2. IBM and Non-Volatile Memory

Non-Volatile Memory Express specifications enable Solid State Drives (SSD) to utilize the Peripheral Component Interconnect Express (PCIe) bus for internal system communication. This results in storage that has low latency and high throughput, calculated as Input/Output operations Per Second (IOPS), and bandwidth.

From a hardware perspective, IBM NVMe adapters are Half Height-Half Length (HH-HL) and are PCIe Gen3 compatible. The current version of cards provides a read bandwidth of ~3GB/s. For example, 8 NVMe adapters provide ~24GB/s compared to ~6 GB/s of a traditional SAN attached storage subsystem with two 16GB/s Fiber Channel (FC) adapters.

NVMe adapters can be directly attached through the PCIe slots or through an external IO box in situations where it is shared across multiple systems.

More information can be found at NVM Express (http://www.nvmexpress.org/).

2.1 Features and benefits

The following features of IBM NVMe adapters make them particularly advantageous when used with SAP HANA:

1. Very low latency (<100 us)
2. High read IOPs and read bandwidth (750K IOPs, 3.0GB/s read bandwidth (BW))
3. High Endurance (> 3 Drive Writes Per Day (DWPD))

Table 1 provides information on the hardware features and the input-output (IO) performance capabilities of the IBM NVMe adapter. The NVMe adapters have a Mean Time Between Failure (MTBF) of 2 million hours, excellent write endurance, and high resiliency.

The adapters can handle high IOPS and the bandwidth measured per card is close to 3.0 GB/s. The read bandwidth scales linearly as the data is striped across multiple NVMe devices.

Hardware Features	Workload	Target
• PCIe Gen 3×4 • Half Height Half Length (HH-HL) • Power ≤ 25 W • Block Size 4096 • Non Volatile Write Buffer • Endurance ≥ 3 DWPD • Warranty ≥ 5 years • Hot Plug Capable • eMLC NAND Flash Technology • RAIF: Tolerant to single flash die failures • ECC ≥ 100 bits per 4KB • MTBF ≥ 2 million hours • Boot: Option ROM BAR ≥ 128KB	Read	750K IOPs
	Write	180K IOPs
	Mixed R/W (70/30)	310K IOPs
	Read Data Tp	3.0 GB/s
	Write Data Tp	1.8 GB/s
	Read Latency	90 µs
	Write Latency	25 µs

Table 1: Features and performance measurements (internal measurements) based on ideal laboratory conditions.

2.1.1 Rapid cold start

SAP HANA provides capabilities to maintain the reliability of its data in the event of failures and also to resume operations with the memory re-loaded as quickly as possible. Planned and unplanned cold start can occur after a software update of SAP HANA, or a software or machine failure. To significantly speed up the data load phase of this process, NVMe is used in parallel with traditional storage or SAN as SAP HANA persistency. As discussed earlier, database size and IO bandwidth determine the amount of time taken to load the database into memory and achieve QoS in terms of query performance before and after failover.

For any software upgrades to take effect, or when any configuration or parameter changes occur, it is necessary to restart SAP HANA core services. The services must also be restarted in the very unlikely event of a database crash. These occurrences necessitate a cold start in which the data has to be loaded into the memory from persistent storage. The time it takes to load the data into the memory, and achieve the expected QoS, is accelerated by using the NVMe based solution.

2.1.2 Configuration

A RAID100 configuration adds robustness and resiliency to traditional configurations that are in use today. Figure 2 shows how RAID0 configurations using traditional SAN and NVMe are used in parallel. RAID1 created on top of the RAID0s makes sure the data is in sync. The data filesystem is created over the RAID100. This configuration provides the following advantages:

- The data is always in sync between NVMe and storage
- Read bandwidth is limited by the NVMe devices (by setting them as the preferred read path)
- Write bandwidth is limited by the external storage (SAN in this case)
- Failure of one of the RAID0 does not affect /hana/data filesystem

The RAID configurations are managed by mdadm component of the Linux kernel.

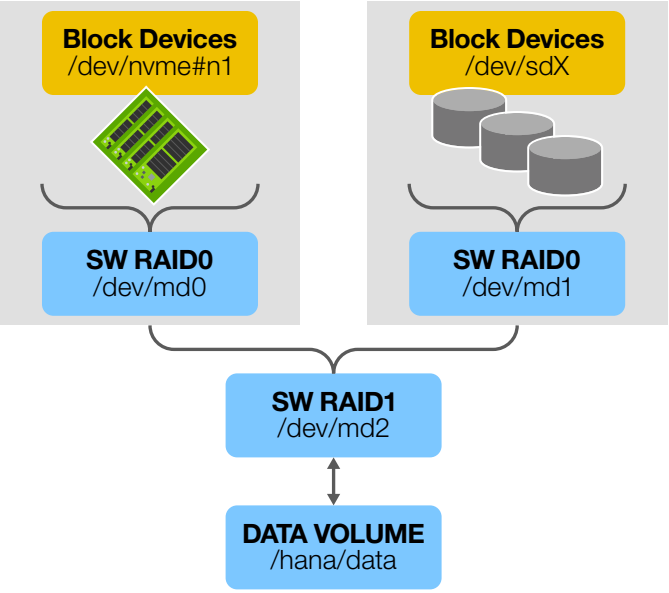
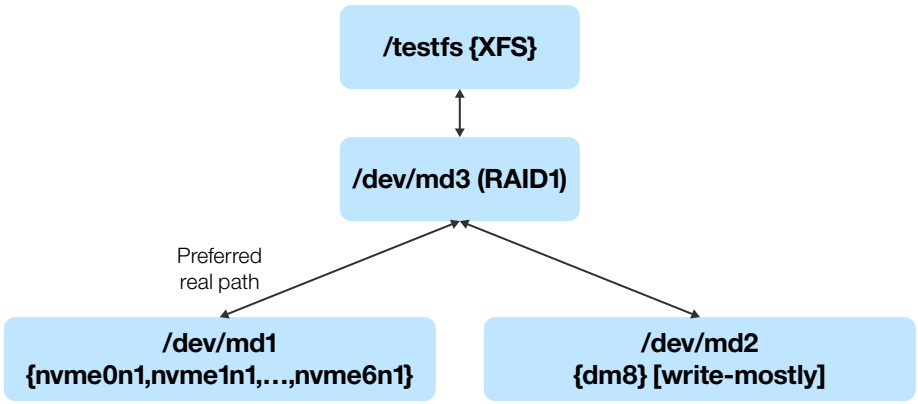


Figure 2: A typical RAID configuration for added resiliency

2.1.3 Performance results

fsperf is provided by SAP as part of the HWCCT (Hardware Check Tool) package along the SAP Tailored Data Center Integration process. Fsperf is the component that measures the IO Key Performance Indicators (KPI) of a customer's SAP HANA infrastructure. We used this tool to test the configuration below.

The first table in Figure 3 shows the fsperf results collected when a database persists directly to external storage (/dev/dm-8), the second table, the RAID100 configuration (/dev/md3) shows the performance achieved with NVMe. The last table in Figure 3 shows the overhead introduced with NVMe cache has no negative performance impacts on database reads. Additionally, these results indicate that the overhead due to multiple Linux software raid (md raid) layers are minimal.



/dev/dm-8 FC LUN performance

Block size	Read BW MB/s	Ratio Trig time/IO time	Read Latency (µs)
4K	96.2	0.0064	228
16K	333.8	0.0066	309
64K	766.9	0.0033	428
1M	785.5	0.0002	1960
16M	786.7	0.0002	21885
64M	787.0	0.0001	83471

/dev/md3 RAD100 performance (NVMe cache path)

Block size	Read BW MB/s	Ratio Trig time/IO time	Read Latency (µs)
4K	824.86	0.0782	93
16K	3219.45	0.0730	273
64K	11969.58	0.0618	700
1M	18812.92	0.0050	1197
16M	20213.76	0.0005	4463
64M	20023.29	0.0003	12905

/dev/md3 RAID100 performance (FC LUN Path)

Block size	Read BW MB/s	Ratio Trig time/IO time	Read Latency (µs)
4K	96.694	0.0065	230
16K	342.381	0.0058	265
64K	774.45	0.0032	360
1M	786.96	0.0002	1983
16M	784.775	0.0002	22127
64M	785.686	0.0001	83285

Figure 3: IO KPIs using 7 NVMe cards

3. HANA tuning

The higher bandwidth requirements need additional tuning of global.ini and the SAP HANA persistence layer.

The following change must be made to the global.ini file:

Global.ini -> [parallel] -> tables_preloaded_in_parallel = 100

The hdbparam also needs to be changed to improve the performance of the SAP HANA persistence layer.

size_kernel_io_queue = 2048

max_parallel_io_requests = 2048

num_submit_queues = 8

4. Sample configuration

A two terabyte database was used to compare the speed up in terms of time. The SAP HANA DB is started using the "HDB start" command and the time it takes to load ~95 percent of data is measured.

The following two configurations of storage were used:

1. IBM NVMe configuration with 10NVMe cards
2. Midrange flash storage

The IO rate versus time was plotted to study the results. The result shows significant improvement in database load time by a factor of 4.6x. The IBM NVMe configuration took 150 seconds to load most of the data while the mid-range flash storage took 690 seconds.

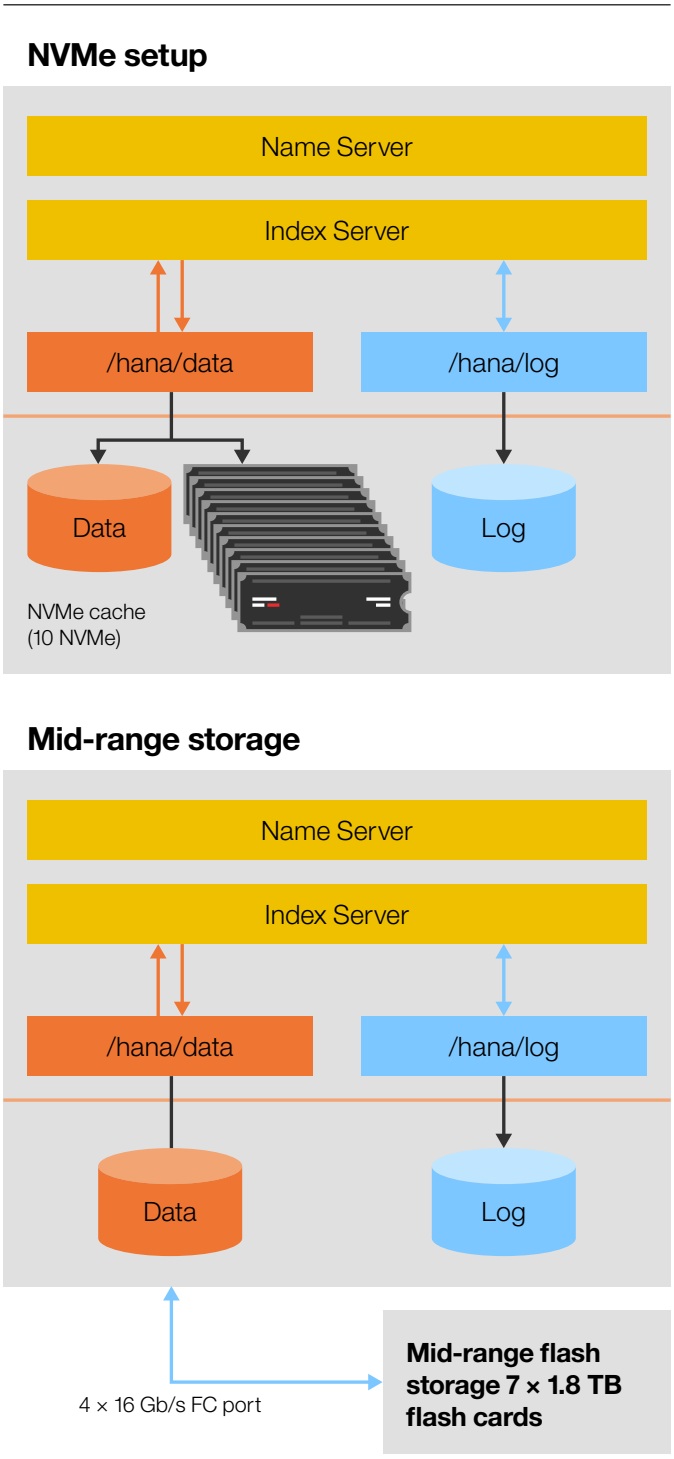


Figure 4: Configuration used to compare IBM NVMe setup versus mid-range storage

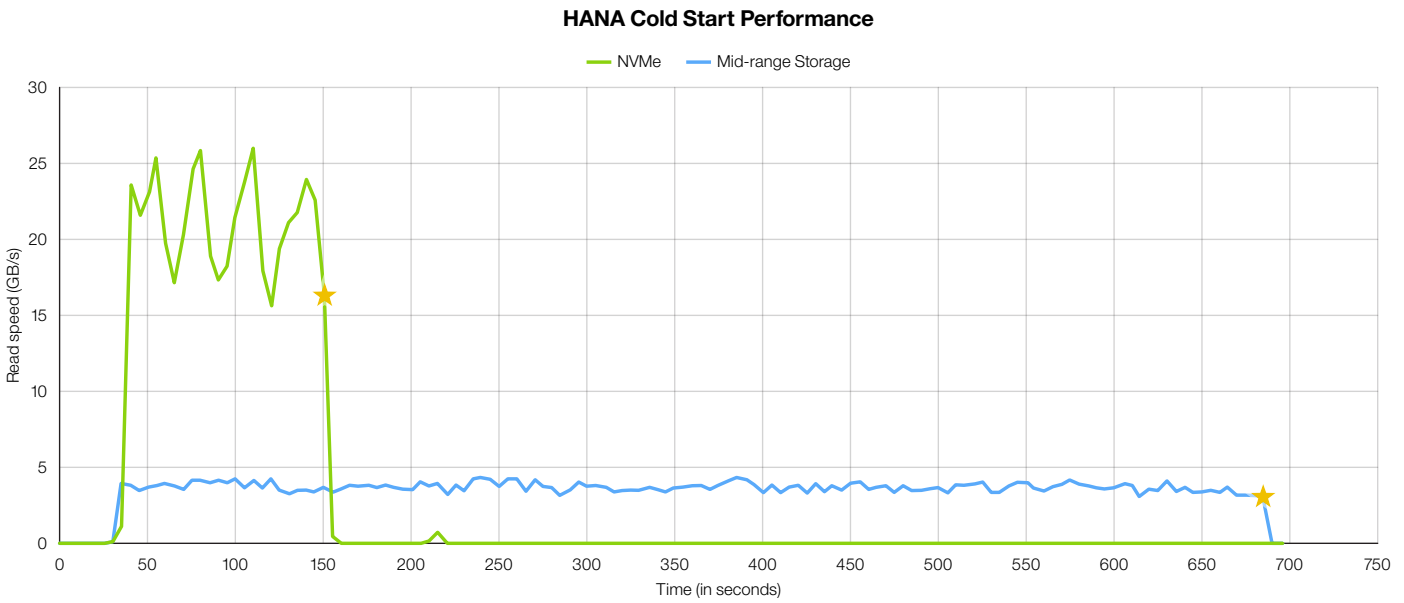


Figure 5: IO rate and time for IBM NVMe setup versus mid-range storage

5. Key takeaways

IBM NVMe adapters help clients:

1. Go live faster by a factor of 4.6 from test to production
2. Reduce planned downtime resulting in a higher level of QoS
3. Enhance existing mid-range storage infrastructure with the IBM NVMe configuration that provides the high bandwidth benefits typically seen in high-end storage

IBM NVMe adapters help your organization move test and QA environments to production faster. NVMe for rapid cold start also enables you to minimize planned maintenance by shortening the time it takes to bring business users back online so they can be productive.

For more information

To learn more about this offering, contact your IBM sales representative or visit the following web site:
ibm.com/power/hana

IBM Corporation
IBM Systems
Route 100
Somers, NY 10589

Produced in the United States of America
September 2016

This document is current as of the initial date of publication and may be changed by IBM at any time. Not all offerings are available in every country in which IBM operates.

The performance data discussed herein is presented as derived under specific operating conditions. Actual results may vary

It is the user’s responsibility to evaluate and verify the operation of any other products or programs with IBM products and programs

Actual available storage capacity may be reported for both uncompressed and compressed data and will vary and may be less than stated.

THE INFORMATION IN THIS DOCUMENT IS PROVIDED “AS IS” WITHOUT ANY WARRANTY, EXPRESS OR IMPLIED, INCLUDING WITHOUT ANY WARRANTIES OF MERCHANT-ABILITY, FITNESS FOR A PARTICULAR PURPOSE AND ANY WARRANTY OR CONDITION OF NON-INFRINGEMENT. IBM products are warranted according to the terms and conditions of the agreements under which they are provided.

[1] Based on HWCCT data collected from multiple customer environment

Please Recycle

POW03190-USEN-02