WHITE PAPER

All-Flash Storage Solution for SAP HANA:

Storage Considerations using SanDisk® Solid State Devices

SanDisk
a Western Digital brand

Western Digital Technologies, Inc.
951 SanDisk Drive, Milpitas, CA 95035

www.SanDisk.com

Table of Contents

Preface

This publication describes in-memory computing configurations with SanDisk all-flash storage and servers that are based on SAP® HANA® and utilize the Lenovo® families of System x® M5™ and X6™ flagship systems with Optimus MAX™ SSD 3.84TB SSDs from SanDisk for the example. This white paper also describes the architecture and components of the solution and how this all-flash storage can result in advantages for customers.

This white paper is intended for SAP administrators and technical solution architects. It is also for SanDisk business partners who would like to understand all-flash storage components of an SAP HANA offering and how it compares to existing persistent storage for SAP HANA solutions. The solution described here as well as a number of other specific in-memory data base capacity point configurations utilizing Optimus MAX 3.84TB SSD can be found for purchase on the configuration templates for the SAP HANA all-flash storage solution provided by Lenovo.

Why SanDisk?

Data is at the heart of everything you do. Innovative flash storage solutions from SanDisk gives everyone, from small businesses to the most advanced data centers, the power to think big. Flash accelerates the flow of data so businesses can get more out of their technology. Flash-transformed data centers are built with fewer servers, fewer storage systems, less software and less energy while delivering better performance, scalability and reliability. With high capacity flash, less is more. The solution described here has been tested and known to comply and exceed the SAP Hana performance requirements for an SAP HANA supporting storage infrastructure.

Why Lenovo?

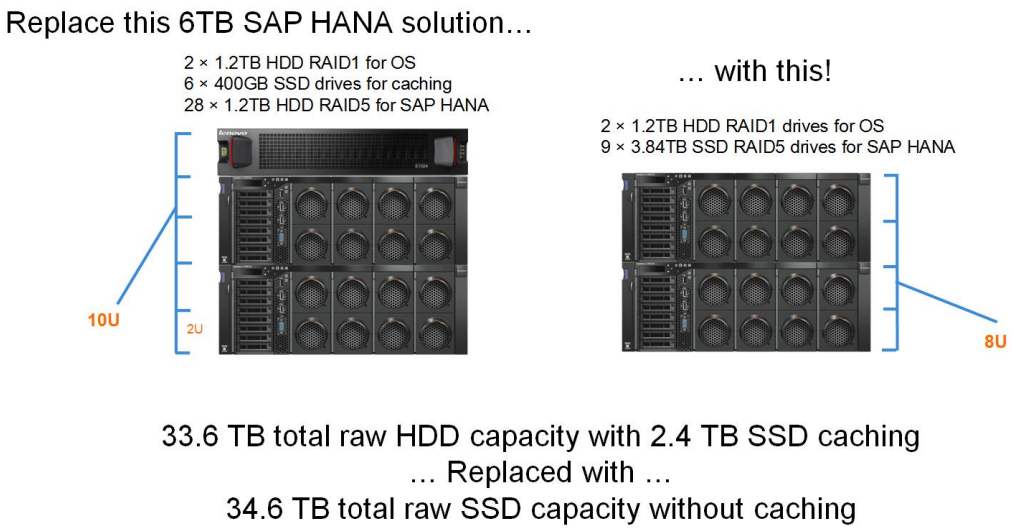
The Lenovo server portfolio includes data warehouse appliances that integrate database, server and storage into a single, easy-to-manage appliance that requires minimal setup and ongoing administration, and provides fast and consistent analytic performance. Lenovo also offers pre-built and pre-integrated workload-optimized data warehousing and analytics platforms and data warehouse software for operational intelligence. These offerings are enhanced with additional support for big data and new types of analytics workloads, including continuous and fast analysis of massive volumes of data-in-motion. The solution described in this white paper as well as a number of other SanDisk based SAP HANA all-flash storage solutions can be purchased in full from Lenovo including a Lenovo-branded version of the Optimus MAX 3.84TB SSD referred to as the Lenovo 3.84TB enterprise Capacity SSD.

Executive Summary

For in-memory database implementations, moving to the SanDisk powered SAP HANA all-flash solution with high capacity 3.84TB enterprise SSDs enables significant efficiency gains and reduction of complexity.

Summary of benefits in this example[1]:

- Overall storage subsystem power requirements are reduced by 82%
- Cost of powering and cooling the storage subsystem is reduced by 82%
- Hardware footprint drops by 20% by removing the DAS enclosure
- Storage device count is reduced from 36 to 11
- 30 or more points of failure are removed from the configuration (storage devices, DAS enclosures cables, RAID controllers, etc.)
- Complexity of multiple storage tiers is removed



While the original Lenovo Solutions for SAP HANA appliance leveraged a balanced storage tier concept consisting of high-performance flash devices combined with high capacity hard disk drives, upgrading to an all-flash solution for standalone systems can provide advantages. These advantages include reduced real estate, power consumption, and increased reliability.

As prices for flash memory devices continue to decline, an all-flash solution will continue to become more appealing in the foreseeable future.

Basic Concepts of In-memory Computing

In-memory computing is a method that allows the analysis and processing of a dataset in a server's main memory to provide results in the fastest possible timeframe. This accelerated dataset processing occurs as a result of using the fastest media with capacity points large enough to house and access the dataset. Utilizing fast media large enough to house the entire dataset is a requirement to prevent the need for complex swapping of portions of the dataset. This fast media is

[1] All performance testing and data was completed by Lenovo.

currently main system memory. HDDs and SSDs are both slower and don't provide the fastest response times and while CPU cache and register sets are even faster, they don't provide the capacity points to be useful for this task. SAP HANA is an example of in-memory computing in which the current working dataset is held in main system memory. To reach the fastest possible performance, SAP HANA follows these guidelines:

- Keep the entire dataset in main system memory
- Minimize data movement
- Compress data for most efficient use of main memory space
- Perform calculations on the dataset at a database level
- Scales the dataset across multiple SAP HANA servers to provide additional system memory, more than a single SAP HANA server could provide

Of course, keeping data in main memory means the data is not persistent. One of the core requirements for an enterprise database is durability, so a level of protection must be put in place for this configuration. In database technology, atomicity, consistency, isolation, and durability (ACID) must be met to ensure that database transactions are processed reliably:

- A transaction must be atomic. This means if part of a transaction fails, the entire transaction must fail and leave the database state unchanged.
- The consistency of a database must be preserved by the transactions that it performs.
- Isolation ensures that no transaction interferes with another transaction.
- Durability means that after a transaction is committed, it remains committed.
- When a dataset primarily lives in main system memory, additional functionality that is not required with a standard storage-based database must be put in place to ensure that durability is achieved.

Providing Durability

A popular method to provide this needed durability is to utilize a persistent log that is stored on persistent storage media. This log contains saved transaction pages. Each committed transaction generates a log entry that is written to non-volatile storage, which ensures that all transactions are permanent. Today these log entries are generally written to enterprise solid-state storage devices, while the transaction pages are written to protected (RAID and/or GPFS) 10K RPM HDDs or, in some cases, an attached SAN. The persistence is supplied by a two-part solution of logs and saved transaction pages. The overall persistent storage capacity will be about 4-5x the size of the in-memory database. In the following discussion we will focus on direct attached storage (DAS) based storage solutions for SAP HANA. We will specifically use the Lenovo DAS-based SAP Hana solution as an example to illustrate a real world solution that can be built today.

To obtain the needed performance from the persistent storage, transaction logs can be placed on enterprise SSDs, while saved transaction pages are placed on 10K RPM HDDs. Both tiers would each be protected by RAID or some other method of data protection. The enterprise SSDs used for logging require a dedicated RAID controller, and the HDD based persistent storage will require another RAID controller.

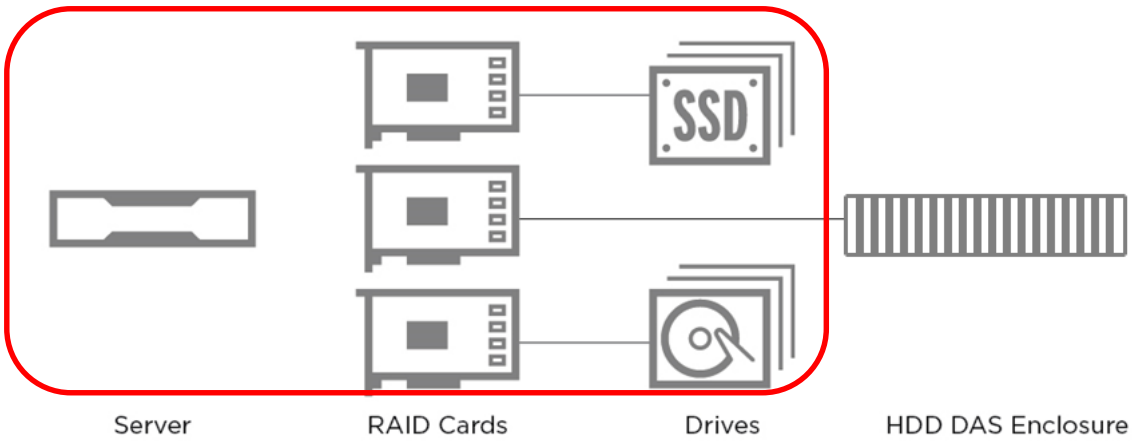
If the in-memory database size is smaller, then the server will have sufficient drive bays to hold all of the HDDs needed to provide persistent storage for the transactions pages. However, if the in-memory database is larger, the server that houses the database may not have enough drive bays to hold all of the HDDs needed to provide the persistent storage for these saved transaction pages. In this case an external drive enclosure, also known as direct attached storage or DAS, will have to

be attached to the server running the in memory database. Attaching an external DAS enclosure requires another RAID controller. This occurs because the RAID controllers currently being used inside of the server have internal connection points only and the DAS enclosure requires an external connection. Large configurations that require an external DAS enclosure attached need a minimum of three (3) RAID controllers, including one for each DAS enclosure(s). Additionally both internal and external SAS cables are required to attach these controllers to the SSD/HDD drives and external enclosures.

Another deployment configuration stores both the logging and saved transaction pages on the same 10K RPM HDD media, and a persistent enterprise SSD-based cache is used to accelerate the persistent tier. This behaves the same way as the first example, since all log entries are written to both the cache and the back end storage. This allows the system to move forward as soon as the persistent cache has committed the log entry. The hardware required to build this configuration is the same as the example above. This solution requires a RAID caching feature to be added to the installation. The same hardware is needed, but the configuration is slightly different in that the SSD storage capacity doesn't increase overall storage capacity.

There are important limitations in both examples described above which can be virtually eliminated by moving to an all-flash solution. The caching solution doesn't allow for the SSD capacity to be utilized as actual usable capacity. As a result, more capacity HDDs must be purchased, and the added complexity of a caching layer must be added. The use of SSDs in the solution increases the number of attach points, which could increase the number of RAID controllers needed. With larger size in-memory databases, this can even require an external enclosure to be attached to the server to hold the overflow of HDDs needed for persistent storage of saved transaction pages.

Similarly, utilizing SSDs as a logging tier in front of an HDD based saved-transaction-pages-tier creates unneeded complexity. The solution no longer requires caching software, but the logging tier now has to be large enough to encompass the entire logging set, since the results won't be held in a much larger monolithic dataset but rather a standalone storage tier. The hardware limitations described above, including the requirement for multiple RAID controllers and the need for external DAS, add complexity.



Providing Durability with an All-Flash Solution

There is another way to provide persistent storage for the SAP HANA in-memory database. Instead of using a separate array for logging on SSDs and HDDs for the saved transaction pages, or using a caching layer accelerating an HDD persistent storage element, we can simplify.

The persistent storage solution can be simplified to a single all-flash-based tier, holding both the logging and the saved transaction pages with no cache element. The ratio of capacity points per SSD versus cost-per-GB have improved to a point that we can now produce solutions with moderate performing flash devices, such as the Optimus MAX 3.84TB SSD, that can replace storage systems that previously were only feasible with HDDs.

To get a reasonable performance access density, referred to here as input/output operations per second (IOPS) per GB, using HDDs the capacity points of the HDDs used are relatively small. This makes multiple RAID controllers necessary to allow direct attachment of all the storage devices. Some SAP HANA configurations may require external storage enclosures to house the HDDs that are needed for the SAP HANA database size. This increases the RAID controller count even farther as the data would now be split among even more RAID controllers.

Large capacity SSDs, like the Optimus MAX 3.84TB SSD, that go beyond the capacity that can be provided with 10K RPM HDDs today can ensure both the logging and saved transaction pages can both fit completely inside of the server, while still meeting the needed performance requirements of the persistent storage subsystem. The Optimus MAX 3.84TB SSD reduces the storage device count significantly. As an example, a 6TB in-memory database based on the Lenovo System x3950 X6 and utilizing HDDs for transaction pages and SSDs for caching, requires 34 storage devices. This can be reduced to as a little as nine Optimus MAX 3.84TB devices. The reduced device count means that a RAID controller is eliminated, thus removing a persistent cache license and reducing complexity as well. Of course, the external DAS enclosure, as well as the cables needed to attach it, is also removed leaving the entire solution encompassed inside of the server running the in-memory database instance.

When building a persistent storage layer with all-flash devices, we can produce a number of benefits in addition to simplifying and consolidating the solution.

Server, RAID Cards, and SSDs…

Better Protection for the Saved Transaction Pages

Enterprise-based flash devices typically have 10x better data protection compared to HDDs with respect to the uncorrectable bit error rate (UBER). Advertised UBER on the typical HDD is 10^{-16} while the Optimus MAX 3.84TB enterprise SSD referred to in this white paper has an UBER of 10^{-18} which is 100 times (100x) better. One could assume that RAID protection makes this irrelevant, but that would be incorrect. An uncorrectable bit error could be a contributing factor in a RAID array marking a drive defunct, which puts the persistent storage array in an unprotected state in a RAID 5 array. Until the array is rebuilt, which may take days with cached HDDs, the data is in jeopardy. Another scenario is when a RAID 5 array is already in a critical state (such as one drive in a defunct state) for any number of reasons and encounters an uncorrectable bit error. While the array is in a critical mode, there is no data protection or ability to generate lost data from parity, and the data packet affected by the uncorrectable bit error would be lost. Since the enterprise SSD has 100x better UBER, there is a 100x less chance of this happening during the rebuild.

Efficient Use of the Existing Capacity

In a solution in which the persistent storage is accelerated with a cache, the capacity of that cache adds no additional capacity to the overall solution. In other words, this means the persistent HDD tier and the cache tier add up to a capacity point that is larger than the needed persistent storage layer. When consolidating the cache accelerator and the persistent HDD storage into a single tier, there is no portion of the data that is held as a copy. The entire persistent data set is protected by the RAID array parity, so full RAID protection is granted, but the capacity point that was once flash-based cache is now a usable portion of the base persistent array. In a 6TB in-memory database, there is about 2.4TB of cache capacity that can now be counted as part of the persistent storage, where it would not be otherwise. This means that less storage capacity needs to be purchased to build the solution as nothing is wasted on a cache copy. Also, by removing the cache, and instead having all of the data on a SanDisk enterprise SSD tier, this means the scenario of a “cache miss” has been removed completely.

Accelerator Cache or Logging Element without High Performance SSDs

When designing a solution with an HDD-based persistent storage element and a flash based accelerating layer in front of it, the idea is to move as much of the workload to the front-end accelerating layer as possible. To do this the flash-based accelerating layer needs to be fast enough to perform this function while also keeping up with the write workload of the entire persistent storage capacity (as you would typically write all new incoming data into cache.) In doing so, this provides a solution that is faster than moving the entire persistent storage element into the high speed flash devices that comprise the cache element. A problem develops in that the necessary high-performance enterprise SSDs carry a cost premium for the performance, as well as for the endurance characteristics. While these faster devices can handle a heavier write profile, the price premium is large. However, as the enterprise solid-state prices continue to fall, large capacity point devices with a moderate performance level at an attractive cost per GB are now possible. The cache can now be removed completely and we can simply run the entire persistent element on flash with no cache present at all.

Collapsing a relatively small but fast tier and a slower larger capacity tier into a single tier can allow for the use of balanced performance, mid-range enterprise SSDs. Spreading the high performance requirement out over a larger capacity point can allow for the use of less expensive SSDs, resulting in a global cost per GB for storage, that may be equal to or less than the global cost per GB of the combined expensive fast tier and a slower back-end tier.

Let's assume, as an example, a large capacity 6TB in-memory database, again using a Lenovo System x3950 X6 server. The storage consists of six 400GB enterprise flash devices to provide a high-performance cache layer, and the base persistent storage is comprised of (28) 1.2TB 10K HDDs, a portion of which require an external DAS enclosure to hold them. For the relatively small cache size needed to meet the performance requirement, they will need to be fast-performing devices (at a premium cost) to ensure the best quality of service (QoS) and performance by the relatively small SSD controller count (six SAS/SATA ports). If the entire persistent element is on moderate-performing flash devices, there is no need for a high performing front end of flash for caching. We simply have a single capacity point built of moderate-performing SSDs. The overall capacity in the solution is larger. In this case the overall capacity provided increased by 1.0TB as the cache layer isn't needed, so we can either have a larger capacity point for persistent storage or we can reduce the amount of capacity purchased.

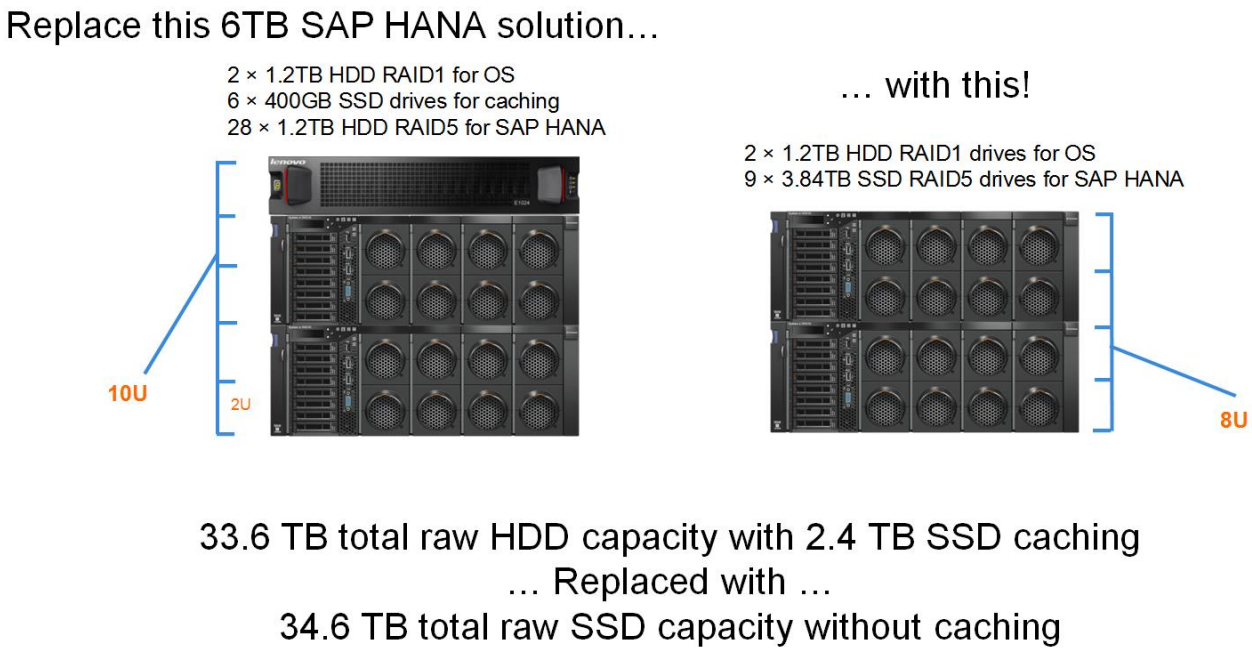
The fact that the total capacity point is built on flash, the SSD controller count went up – in this case to nine controllers from the six SSD controllers that previously accelerated the solution. Nine controllers allow for the performance to be spread out, thus utilizing a more moderate performance needed per device, allowing for a lower cost, lower-performance flash device to be utilized, while actually increasing the performance of the persistent storage subsystem. A combination of removing the need to buy high-performing flash devices, and moving to a higher count of lower-performing devices, allows for meeting the same or better performance characteristics at a similar cost point, in a smaller footprint.

This particular example required an external DAS enclosure attached to the in-memory database server, along with a RAID

controller for the SSDs, another for the internal HDDs, and a third with connection to the DAS enclosure (which holds the remaining HDDs needed to reach the total purchased capacity of 33.6TB.) An Optimus MAX 3.84TB SSD-based all-flash persistent storage solution moves all the storage into the main system. This reduces the hardware footprint of the solution by 20%. The RAID controller count drops from 3 controllers to 2 and the added cost of acquiring a DAS enclosure and cable set also goes away. We have a solution that loses 25 points of device failure and the device count goes from 36 storage devices to 11. Numerous potential points of failure and complexity also go away by removing external components such as cables, enclosures, enclosure fans, power supplies, etc. The all-flash solution fully removes the chance of ever having a cache miss. There is no wasted capacity due to data copies in the cache. The chances of encountering an uncorrectable bit error are reduced by a factor of 100x per device with the added benefit of 70% less devices in the solution that could experience one. This of course also means the overall solution MTBF is improved because of the removal of so many points of failure.

In this example the power usage of the storage subsystem is also reduced by 82%. Let's examine the various variables used in calculating the subsystem power reduction. Assume the need for one DAS enclosure – assuming 100 watts power draw for the enclosure. Each of the 28 HDDs requires 10 watts and each of the six SSDs require 6 watts. Assume the 3 RAID controllers referenced above require 10 watts each. This means a total of 446 watts is required for the standard hybrid solution.

The all-flash solution requires two RAID controllers at 10 watts and nine 3.84TB enterprise-capacity flash drives at 7 watts each for a total of 83 watts for the subsystem. There is no need for data to ever leave the server housing the in-memory database. 83 watts versus 446 watts is an 82% power reduction in the subsystem.



Summary

In summary, moving to an all-flash SAP HANA configuration option brings several advantages.

Summary of benefits for the Optimus MAX 3.84TB SSD all-flash, SSD-based solution[2]:

- Hardware footprint drops by 20% by removing the DAS enclosure
- Need for cables to attach and power the DAS enclosure are removed
    - Removing the external enclosure cable removes a single point of external vulnerability that if unplugged or damaged, access to saved transaction pages will result
- RAID controller count goes from 3 down to 2
- Storage device count is reduced from 36 down to 11
- Fewer controllers and storage devices means a better solution MTBF
- Complexity of multiple storage tier is removed
- Need for cache tier and cache software license is removed
- Risk of a "cache miss" is removed by removing the cache element
- Wasted capacity that can't be claimed as part of the persistent storage is removed
- Overall storage subsystem power requirements are reduced by 82%
- Costs of powering and cooling the storage subsystem is reduced by 82%
- Odds of experiencing an uncorrectable bit error resulting in data loss are reduced significantly

Specifications are subject to change. ©2016 Western Digital Corporation or its affiliates. All rights reserved. SanDisk and the SanDisk logo are trademarks of Western Digital Corporation or its affiliates, registered in the U.S. and other countries. Optimus MAX is a trademark of Western Digital Corporation or its affiliates. Other brand names mentioned herein are for identification purposes only and may be the trademarks of their respective holder(s). 5101EN 20160616

Western Digital Technologies, Inc. is the seller of record and licensee in the Americas of SanDisk® products.

[2] All performance testing and data was completed and provided by Lenovo.