VMware® vSAN Tuning Guide for AMD EPYC™ 7002 Processors

Publication #	**56781** Revision: **1.0**
Issue Date:	**December 2019**
Author:	**Travis Hindley**

Advanced Micro Devices

Contents

Revision History

Date	Revision	Description
December, 2019	1.0	Initial public release.

Chapter 1 Introduction

AMD launched the 2nd Generation EPYC 7002 Series Processors in August 2019. The AMD EPYC™ 7002 Series Processors are built with leading-edge 7nm technology, Zen2 core and microarchitecture. The AMD EPYC™ SoC offers a consistent set of features across 8 to 64 cores, including 128 lanes of PCIe® Gen 42, 8 memory channels and access to up to 4 TB of high-speed memory.

This VMware vSAN tuning guide is intended for partners who wish to deploy this HCI software solution on OEM hardware within the VMware ecosystem. Due to the number of permutations available, this guide is intended to be used as a best practice reference.

This tuning guide assumes some level of familiarity with x86 server hardware, infrastructure found in a typical datacenter, and VMware virtualization, and VMware vSAN. For a more complete overview of 2nd Generation EPYC Series Processors, recommended BIOS settings, and other best practices as it relates to enterprise virtualization, see *VMware vSphere Tuning Guide for AMD EPYC™ 7002 Series Processors*.

Chapter 2 Tuning Guide Testbed

2.1 Goals and Assumptions

VMware vSAN can have a number of variables that make it difficult to provide a common set of best practices. A change from SATA SSD to SAS SSD or SAS SSD to all NVMe can have a significant impact on the underlying performance despite still falling into the category of vSAN All Flash solution. The goal of this tuning guide is to focus on the 2nd Generation EPYC processors.

Therefore, the Systems Under Test (referred to as SUTs in this document) had as many bottlenecks removed as possible in order to expose the limitations of the overall solution.

2.2 4 Node vSAN Cluster

VMware vSAN can be deployed with up to 64 ESXi hosts in a vSAN cluster. Alternatively, clusters can be quite small as well – 2 nodes + vSAN Witness Host, and 3 Nodes without a vSAN Witness. However, a 4-node vSAN Cluster is ideal for availability purposes. Further, it is not possible to compare RAID1 and RAID5 with only 3-node in a vSAN cluster because RAID5 has a 4-node minimum.

See the following for more details on the vSAN Witness:
https://blogs.vmware.com/virtualblocks/2018/09/17/vsan-2n-witness-consolidation/

For the above reasons, all testing was conducted on 4 identically configured servers.

2.3 vSAN Cluster Configuration

All four SUTs were configured identically with the below specifications

- 1 x Single Socket AMD EPYC 2nd Generation Server
- 1 x AMD EPYC™ 7742 2nd Generation Processor
- 8 x 64GB 3200 MHz DIMMs (1DPC, Dual Rank)
- 10 x 1.6TB Samsung PM1725b NVMe
- 1 x Dual Port Mellanox ConnectX-5 Ex 100 GbE QSFP NIC

Unless otherwise specified, all BIOS, ESXi, and vSAN options were configured to defaults. The one notable exception is that the entire vSAN network was configured for 100 Gbps Ethernet at MTU 9000. While 100 Gbps might not be as prevalent, there is no expectation that 25 Gbps should be substantially different. Further, standard frame sizes are more typical in the datacenter, but enabling jumbo frames is one of the most commonly recommended best practices for vSAN clusters.

Most testing was completed with two Disk Groups on the cluster. This means that of the ten NVMe devices, two were configured as cache drives and 8 were configured as capacity drives. In order to limit the scope of this tuning guide, all vSAN testing is categorized as an All-Flash Architecture.

The default BIOS settings on the SUTs and most Second Generation EPYC servers is listed below. It should be assumed these settings are configured unless otherwise specified.

- NUMA Per Socket: NPS1
- L3 Cache as NUMA Domain: disabled
- CPU Power Management: OS Controlled
- Memory Frequency: 3200 MHz
- IOMMU Support: enabled
- PCIe Preferred IO Device: disabled
- Turbo Boost: enabled
- C States: enabled
- Memory Refresh Rate: 1x

Chapter 3 BIOS Settings Best Practices

This chapter covers some of the common BIOS settings that may impact performance in an AMD vSAN cluster. Consider the following:

- HCIBench is a synthetic IO benchmark, and a number of these settings are best practices in an environment where typical enterprise virtualization is taking place. However, these settings could negatively impact vSAN performance when no other work is ongoing *except* driving IO load.
- Every option comes with a tradeoff. There is no single set of tuning options that will be best for vSAN in all environments. Some options are throughput oriented, while others are latency oriented. This is true outside vSAN environments as well; therefore, carefully consider what types of workloads will be likely when determining what is best.

3.1 BIOS Option: NUMA Per Socket (NPS)

2nd Generation EPYC features user configurable Non-Uniform Memory Access settings. NPS stands for NUMA Nodes Per Socket, which has been further simplified and optimized since first generation EPYC. Briefly, they are as follows:

- NPS1
  - One NUMA node per socket
  - Larger memory domain, potentially higher latency
  - Higher throughput achieved with minimal tuning
- NPS2
  - Two NUMA nodes per socket
- NPS4
  - Four NUMA nodes per socket
  - Smaller memory domains, assists the ESXi scheduler to achieve lower latency
  - Throughput suffers without proper NUMA node distribution

The NPS setting that works best will be highly dependent on the emphasis of the cluster. For large block sequential reads, NPS1 achieves almost double the throughput of NPS2 and NPS4 at almost half the latency. However, few vSAN clusters will truly need that much throughput.

NPS2 achieves exceptional small block random read performance, and NPS4 is close as well.

For vSAN clusters that are write heavy, none of the NPS settings seem to have much of an impact, likely because all writes are being consumed by the cache drive before de-staging to the capacity drives. An imbalance of capacity drives in NPS2 or NPS4 will adversely impact read throughput. However, a two-disk group configuration will run into other bottlenecks before NPS2 or NPS4 distribution becomes a problem.

Best practice recommendation for vSAN is:

- Large block, sequential read heavy workloads should utilize NPS1
- Small block, random read heavy workloads should utilize either NPS2 or NPS4

3.2 BIOS Option: L3 Cache as NUMA Domain

In addition to the NPS settings detailed above, there is a new BIOS option in the2nd Generation EPYC that can improve performance in some virtualization scenarios – L3 Cache as NUMA Domain. Understanding this setting requires knowledge about the AMD EPYC™ 7002 processors. For more information, see *https://developer.amd.com/wp-content/resources/56338_1.00_pub.pdf*

By default, ESXi schedules vCPUs for a single VM onto the same NUMA node. This does not guarantee that those same vCPUs will be able to leverage a common 16MB L3 cache. Some workloads, such as VMmark3, where defining each CCX and associated 16MB L3 cache as its own NUMA domain can be highly beneficial. However, there are other workloads, such as HCIBench on vSAN, where there is no benefit from an IO perspective.

A performance improvement could be realized in a vSAN cluster with this option enabled, but that is because most vSAN clusters are running more than a synthetic IO benchmark. Purely from a storage performance standpoint, there is no performance benefit from enabled L3 Cache as NUMA Domain – in fact, quite the opposite.

When L3 Cache as NUMA Domain is enabled, throughput is equivalent or lower than when it is disabled. Similarly, latency is significantly higher when it is enabled. While the comparison with L3 Cache as NUMA Domain enabled is especially poor for large block sequential reads, NPS1 performs exceptionally well in this metric.

There are many reasons to enable L3 Cache as NUMA Domain in VMware environments. In fact, it could even be considered a best practice. However, it should never be enabled *solely* for improving HCIBench performance.

Best practice recommendation for vSAN is to leave L3 Cache as NUMA Domain disabled.

3.3 BIOS Option: Memory Frequency

2nd Generation EPYC Processors support memory frequencies up to 3200MHz. It is also possible to run at a lower memory frequency when necessary. There are multiple reasons for this, but one reason to consider is that any processor operates within a thermal and power envelope. It should be expected that the higher the operating memory frequency, power consumption and heat generated will increase. Workloads that are not memory bandwidth dependent may see improvement operating at a lower memory frequency.

A clear tradeoff can be observed as memory frequency is scaled down that peaks at 2667 MHz in all test cases. Only at 2400 MHz does the reduction in memory frequency and data fabric negatively impact performance but is still a clear improvement over 2933 MHz.

VMware vSAN observes no advantage running at 3200 MHz. Memory frequency running at 2933 MHz might be the best compromise between memory throughput and latency in vSAN environments, but both 2667 MHz and 2400 MHz are higher performing. The best performance was achieved at 2667 MHz.

As with L3 Cache as NUMA Domain, vSAN IO performance does not exist in isolation. Memory throughput intensive VMs such as SAP HANA, even if residing on a vSAN cluster, could be negatively impacted by scaling down frequency too far. While 2667 MHz might give the very best vSAN performance, 2933 MHz might be a better general-purpose frequency.

Best practice recommendation for vSAN is to run memory frequency at 2667 MHz for best IO Performance and 2933 MHz for best general-purpose performance where vSAN is configured.

3.4 Power Management Settings

The impact Power Management Settings can have on a system is a frequent topic of debate. Historically, recommendations to disable all power management is a first step at troubleshooting or a best practice. However, in recent years, this has adverse effects on frequency, particularly on systems that are not operating at 100% utilization. AMD EPYC mitigates this to a certain degree, but it can still have an impact.

The following two options affect power management that can be altered individually or together:

3.4.1 BIOS Power Management

All OEM server vendors implement BIOS Power Management differently and with different names. For the most part, these all bundle in a series of BIOS options to force the system to operate at a higher performing state. However, particularly in lightly utilized systems, this is not always the case.

The configured options are:

- C States – disabled
- Memory Frequency – Maximum Performance
- OS Control – disabled

However, as an example, memory frequency does not always translate into higher performance, particularly where vSAN is concerned.

3.4.2 ESXi Power Management

Within ESXi, VMware provides a variety of power management governors that can impact performance. This setting is found on each ESXi host, in the “Configure” tab, below the

“Hardware” tree in the “Power Management” Section. The default Active Policy is “Balanced”, but “High Performance” is a common recommendation.

3.4.3 Conclusion

No matter which method of Power Management is applied, Performance is within acceptable noise for run-to-run variation. The notable exception is sequential write throughput. Latency and throughput are negatively impacted whenever C States are disabled in the BIOS. However, if there are VMs within the cluster that require power management turned off, there will not be a significant impact other than in a few, write heavy cases.

Best practice recommendation for vSAN is to leave all C States enabled, keep all OS control enabled, and leave VMware Power Management at the default of “Balanced”.

3.5 BIOS Option: Preferred I/O

In some scenarios, it may be possible to improve the performance of certain PCIe devices by enabling Preferred IO Mode. Depending on the OEM, these steps can vary. However, this feature requires both for Preferred I/O to be enabled as well as the address of the Bus / Device / Function for the PCIe device to be prioritized.

Large block sequential writes are negatively impacted, and large block sequential reads are positively impacted by setting the NIC to Preferred IO device. However, both workloads are very close to the run to run variation window.

Best practice recommendation for vSAN is to leave Preferred IO unconfigured as it does not have an overly significant impact on vSAN performance.