



MEASURING STORAGE PERFORMANCE

TABLE OF CONTENTS

Introduction..... 3

Storage Performance Metrics..... 4

Factors Affecting Storage Performance 5

Provisioning IOPS in Hardware-Defined Solutions 6

Provisioning IOPS in Software-Defined Solutions 7

Best Practices for RAID and Cache Sizing..... 8

INTRODUCTION

In today's era of big data and ever-increasing demands for real-time analysis of that data, it is imperative that IT organizations understand how to measure storage performance. This document guides IT personnel through the process of measuring the performance of both the newer software-based storage and traditional hardware-based storage. By understanding how to measure storage performance, IT personnel will be able to better predict storage needs as they apply to the needs of the business and develop benchmarks for RFPs and product evaluations.

This document focuses on measuring the impact of the following factors on storage performance and the application of best practices in modern software-defined storage systems:

- Storage performance metrics (IOPS, throughput)
- Factors affecting storage performance (RAID penalty, READ/WRITE ratios)
- Provisioning IOPS in legacy (hardware-defined) storage solutions
- Measuring Storage Performance in ZFS-based (software-defined) storage solutions
- Best Practices for RAID and cache sizing in ZFS-based storage

This document is intended for solution architects, storage network engineers, and system administrators involved with storage evaluation, configuration, deployment, and management. A working knowledge of basic storage concepts is assumed.

STORAGE PERFORMANCE METRICS

This section introduces the key storage performance metrics – IOPS and throughput – and how to measure them. The relationship between throughput and IOPS is measured as:

$$\text{Throughput (MB/sec)} = \text{IOPS} * \text{Block size (MB)}$$

IOPS CALCULATIONS

IOPS is measured by the number of I/O operations, i.e. READs and WRITEs, per second, and can be classified as follows:

- Per Disk IOPS is the rated IOPS of a single SATA/SAS/FC disk of varying RPMs.
- Frontend IOPS is the IOPS of the application, installed on storage LUN, which consumes storage. This is the IOPS classification used when talking about a requirement for 100, 200, 1,000, or 1 million IOPS.
- Backend IOPS is the IOPS required by the storage subsystem to deliver the required frontend IOPS and is dependent on RAID penalties.

CALCULATING PER-DISK IOPS

METRIC	HOW IT IS CALCULATED
Average READ seek time	Rated and published by disk vendors in data sheets and other product specifications.
Average WRITE seek time	Rated and published by disk vendors in data sheets and other product specifications
Average rotational latency	Half the time required for a rotation in milliseconds (ms). For example, 7200 RPM (120 rotations per second) translates to one rotation every 8.33 ms. Half the rotation takes 4.16 ms. Thus, the average rotational latency for a 7200 RPM drive is 4.16 ms.
IOPS per disk	$1 / ((\text{average read seek time} + \text{average write seek time}) / 2) / 1000 + (\text{average rotational latency} / 1000)$ Below are three example calculations: For a 7200 RPM disk, per disk IOPS = $1 / (((8.5 + 9.5) / 2) / 1000) + (4.16 / 1000) = 1 / ((9 / 1000) + (4.16 / 1000)) = 1000 / 13.16 = 75.98$. For a 10K RPM SAS/FC disk, per disk IOPS = $1 / (((3.8 + 4.4) / 2) / 1000) + (2.98 / 1000) = 1 / ((4.10 / 1000) + (2.98 / 1000)) = 1000 / 7.08 = 141.24$ For a 15K RPM SAS/FC disk, per disk IOPS = $1 / (((3.48 + 3.9) / 2) / 1000) + (2.00 / 1000) = 1 / ((3.65 / 1000) + (2 / 1000)) = 1000 / 5.65 = 176.99$ These examples illustrate the reason for minor variations in the rated disk IOPS from different models/vendors for the same RPM disks.

FACTORS AFFECTING STORAGE PERFORMANCE

This section reviews the concept and impact of RAID penalties and how different operations impact READ/WRITE ratios

RAID PENALTY

Because WRITES to a disk are complete only when the data and the parity information have been fully written to the disk, extra time is required for writing the parity information. This extra time is called the “RAID penalty”. It applies only to WRITE I/OS, not to READ I/Os. Measurement begins at RAID Penalty 1, which means that there is no RAID penalty. Other common examples are given in the table below:

RAID TYPE	SCENARIO	PENALTY
RAID0	Striping	There is no parity to be calculated, so there is no associated WRITE penalty. The READ penalty is 1 and the WRITE penalty is 1.
RAID1	Mirroring	The WRITE must be to the mirrored pair, so while the READ penalty is still 1, the WRITE penalty increases to 2.
RAID5	Distributed parity	This entails reading old data block, reading old parity block, writing new data block, and writing new parity block for each change to the disk, so while the READ penalty is still 1, the WRITE penalty increases to 4.
RAID6	Dual distributed parity	Now the operations involve reading data, reading parity1, reading parity2, writing data, writing parity1, and writing parity2 for each change to the disk. The READ penalty is still 1 but the WRITE penalty is now 6

HOW DO DIFFERENT OPERATIONS IMPACT READ/WRITE RATIOS?

The following table explains the average READ and WRITE percentages (approximated) for a range of operations.

APPLICATION	RANDOM/SEQUENTIAL	READ PERCENT	WRITE PERCENT	BLOCKS (IN KB)
File Copy (SMB)	Random	50	50	64
Mail Server	Random	67	33	8
Database (transaction processing)	Random	67	33	8
Web Server	Random		100	64
Database (log file)	Sequential		100	64
Backup	Sequential		100	64
Restore	Sequential	100		64
Mail Server	Sequential	100		≥ 64

PROVISIONING IOPS IN HARDWARE – DEFINED SOLUTIONS

To calculate the number of disks needed to meet a frontend IOPS requirement on a legacy (hardware-based) storage system, use the following equation:

$$N = (\% \text{READS} * \text{READ penalty} * \text{frontend IOPS} + \% \text{WRITE} * \text{WRITE penalty} * \text{front-end IOPS}) / (\text{per disk IOPS})$$

Note that legacy storage systems can significantly increase IOPS when a large amount of caching or all-flash (solid state drive) arrays are involved.

PROVISIONING IOPS IN HARDWARE – DEFINED SOLUTIONS

This section discusses the factors to consider when planning a software-defined storage solution using ZFS-based systems as the example

CALCULATING FRONT-END IOPS

The theoretical front end IOPS is limited by the number of virtual devices (raid groups) provided all VDEVs have similar disks. The practical front end IOPS can be viewed in performance analytics tools like the I/O meter, which may vary depending on available network bandwidth.

ASSESSING READ PERFORMANCE

- RAID1 or mirroring RAID group of n disks (VDEV) gives n times a single disk's IOPS. For a pool with multiple VDEVs (raid groups), read IOPS for the pool = $n * \text{number of VDEVs} * \text{single disk IOPS}$
- RAID-Z RAID group (VDEV) gives a single disk's IOPS. For a pool with multiple VDEVs (raid groups), read IOPS for the pool = $\text{number of VDEVs} * \text{single disk IOPS}$

Performance can be improved with a RAID 1+0 configuration by adding multiple RAID1 groups in a pool.

THE IMPACT OF DYNAMIC STRIPING

ZFS dynamically stripes data across all virtual devices (RAID groups) in a pool. Multiple RAID 1 groups in a pool lead to RAID 1+0. Multiple RAID-Z1/RAID-Z2 groups in a pool lead to RAID 50 and RAID 60. Dynamic striping delivers the best of both worlds - striped performance and underlying redundancy. Striped mirrors (1+0) always outperform RAID-Z in both sequential and random READs and WRITES.

BEST PRACTICES FOR RAID AND CACHE SIZING

This last section applies industry best practices and recommendations for RAID and cache sizing, again using ZFS based systems as the example.

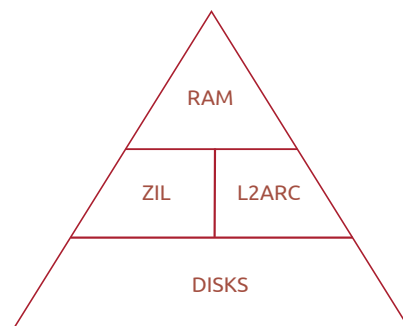
RAID GROUP SIZING

For RAID-Zp, $2^n + p$ is the recommended number of disks in a RAID group, where n can linearly increase (1, 2, 3...) to provision the required storage performance and capacity

RAID CONFIGURATION	NUMBER OF DISKS IN RAID GROUP
RAID-Z1	3, 5, 9, 17...
RAID-Z2	4, 6, 10, 18, ...

CACHE SIZING AND PERFORMANCE

A middle cache tier can significantly improve performance, an approach which is not possible with legacy systems, which follow a direct RAM-to-DISK operation.



As shown in Figure 1 above, the Adaptive Replacement Cache (ARC) resides in RAM and is the first destination for all data written to a ZFS Pool. It is the fastest source for data READs from a ZFS pool. When data is requested from ZFS, it first looks in the ARC; if data is present in the ARC, it can quickly be retrieved by the application. The contents of the ARC are balanced between the most recently used (MRU) and the most frequently used (MFU) data. The second level (L2) cache resides in SSD and is populated by data first placed in the ARC.

The amount of RAM needed for L2ARC will vary according to individual requirements, but as an example, about 15 GB of RAM is required to reference 600 GB of L2ARC at an 8 KB ZFS record size. For a 16 KB record size, the RAM required is halved to 7.5 GB. If insufficient RAM is configured, L2ARC will not completely populate with the MRU and MFU data.

RAID GROUP SIZING

The optimal ZFS record size for L2ARC is 8 KB. Higher record sizes reduce the IOPS, whereas smaller record sizes hog the RAM. As the SSD has to be populated with the MRU MFU data, L2ARC takes a while to warm up.

RAID GROUP SIZING

WSS is the subset of total data that is actively worked upon - for example, 0.2X out of the total X GB; it is a great deal easier to size ARC, L2ARC, and disk space requirements with historical data from production systems. To get the maximum cache hits and fewer cache misses, it is helpful to have as much active data in one of the two levels of cache as possible.

ZIL, L2ARC, AND SSDS

The ZIL device is used for WRITE caching and need not be more than

$10 \text{ sec} * \text{Speed of SSD in S GB/Sec} = 10 * S \text{ GB.}$

.In terms of recommended disk type, if SSDs of same type are used, ZIL/L2ARC benefits are not available for all SSD arrays. Some vendors offer optimized SSDs that are only meant only for handling READ and WRITE caching. If these SSDs are used for caching, large capacity SSDs can be used in place of slow-spinning RPM-based drives to give an allSSD array.

DOES SSD FAILURE MATTER WHEN USED FOR ZIL/L2ARC?

SSD failure when used for ZIL/L2ARC will affect performance but not data. Here's what will happen:

For L2ARC, losing one SSD in L2ARC means MRU/MFU data access requests must be served from slow spinning drives. However, L2ARC best practice uses multiple striped drives.

For ZIL/SLOG, any data in the ZIL/SLOG is also in the ARC until it is flushed to the spinning HDDs. Data loss will occur only if the ZIL device fails and the controller loses power within the ensuing 10 seconds, so mirrored ZIL drives are used.

ABOUT CLOUDBYTE

CloudByte is an award-winning software-defined storage company that offers guaranteed quality of service (QoS) to every application from a shared storage platform, allowing cloud service providers and enterprises building private clouds to easily host performance applications in the cloud. With CloudByte's on-demand performance management, cloud environments can easily scale to thousands of applications while economically guaranteeing performance for each application. Established in 2011 by technology executives from companies such as HP, IBM, NetApp, and Novell, CloudByte is backed by Fidelity Worldwide Investment, Nexus Venture Partners and Kae Capital.

For more information, go to www.cloudbyte.com, www.facebook.com/cloudbyte, or www.twitter.com/cloudbyteinc.

CLOUDBYTE ECOSYSTEM

