Top Considerations for Disk Subsystem Performance

Written by André Rossouw, Advisory Technical Solutions Education Consultant, EMC Corporation
EMC Proven Professional, EMCIEe, EMCTAe

The reality is that data storage environments are experiencing exceedingly rapid growth rates. Consequently, so is the need for experienced storage professionals, but since that’s a longer-term issue to solve, this article will focus on helping the existing pool of IT professionals to work a little smarter. Specifically, this is directed at those who either manage storage as a primary responsibility, or confront storage-related tasks during database, systems, or network administration.

Why is disk performance a topic for concern? Consider your own experience when you start a program running on your personal computer—much of the delay from the moment you type the program name, or click the link, to the time the program actually launches, is delay caused by the disk. Now imagine if you were sharing that disk with several other people—the delay would be a lot more significant, and would start having an impact on your productivity.

That impact on productivity particularly applies in a business environment where disk subsystems, shared by many users, are vital to business operation. Making those systems perform optimally is an important task—one that we’ll take an introductory look at in this article.

Consider the following with regard to disk subsystem performance:

1) **Know your workload**
Performance issues are often related to factors other than the disks or disk subsystem. In order to get an idea of what is a reasonable level of performance in any environment, it is vital to have a good idea of what type of data access patterns the hosts are using. Some applications, an example being backup applications, perform largely sequential operations, where the data read from, or written to, disk is contiguous, or mostly contiguous. Other applications, examples being messaging or OLTP (online transaction processing) applications, perform largely random reads and writes, where data is scattered all over the data area of the disks. These access patterns produce very different performance levels.

---

**Common Terms**

**RAID**: Acronym for Redundant Array of Independent Disks. The “array” part of the name means that a group of disks are used together, usually to improved performance, improved data protection, or both. While RAID principles are fairly well-defined, the precise implementation is often vendor-specific.

**Host**: For our purposes, a computer system which is attached to, and uses for data storage, the disk subsystem we are discussing.

**Striping**: Dividing data into equal-sized pieces, and distributing those pieces evenly among the disks in a set.

**Stripe depth**: The maximum amount of data that will be written to a single disk when a single stripe is written.

**Bandwidth**: The amount of data read and/or written in a given time interval, typically the number of megabytes (MB) per second.

**Throughput**: The number of read and/or write operations in a given time interval, typically expressed as I/Os per second, or IOPs.
Another very important part of the workload to be aware of is the ratio of disk reads to disk writes. Writes take longer than reads in environments that do not use caching, and may impose an even greater performance penalty in RAID environments, where multiple disk operations may be associated with a single host write.

2) **Understand what disks are capable of**

Disks are mechanical devices which perform only one operation at a time. Those operations will be data reads, data writes, or seeks—moving the read/write heads from one location to another. Reads and writes, viewed in isolation, are relatively fast processes for the most frequently encountered I/O (input/output) sizes. Seeks, being movement of mechanical parts, take a long time, and have a dramatic effect on performance.

Let’s take a look at an oversimplified example. We’ll assume that a read or write operation can be completed in 1.0 ms (millisecond—1/1000 of a second), and that a seek takes 4.0 ms to perform. If data access is purely sequential, with no read/write head movement, then our theoretical disk can perform 1,000 operations per second. If each read or write is preceded by a seek, the time taken per read or write operation increases to 5.0 ms. The disk can then perform only 200 operations per second. Every “real world” disk will have a data sheet which lists the times taken for seeks, and the expected data transfer rates for the disk. The numbers are generated in carefully controlled environments, so they should not be used as the expected level of performance in any specific environment.

3) **Separate sequential and random workloads**

The example above shows how differently sequential and random workloads perform. When random data access patterns are mixed with sequential access patterns on a disk, some interference occurs, and disk performance will be reduced.

If possible, then, determine which applications produce random workloads, and which applications produce sequential workloads, and keep data from those application types on different disks.

4) **Use disk caching where possible**

Modern disks all have some memory that is used for buffering or caching of data, usually in the 2 MB to 8 MB range. This memory can speed up disk operations by absorbing bursts of writes, and holding data which the disk reads ahead of time. Some operating systems allow the use of this memory by, for example, giving the user a choice of write-back or write-through operations. Write-through caching means that data is written to the cache and to the physical disk at the same time—the operation takes as long as a write directly to disk, but leaves the data in cache, where it can be accessed more quickly for any subsequent reads. Write-back caching means that data is written to the cache, and only written to the physical disk at some later time; the disk will try to arrange the writes to achieve the best performance. The cached data will be lost if power is removed from the drive before the data has been written onto the physical disk. As a result, write-back caching is usually only used when the disk enclosure (the host, or an external enclosure) is protected against power failures. Most intelligent disk storage systems have their own cache, and will disable the cache on the physical disks.
The following graphic depicts the steps that take place when data is read from/written to the drive:

**Disk Drive Performance: Data Transfer Rate**

External transfer rate measured here

Internal transfer rate measured here

HBA

Interface

Buffer

Disk Drive

Source: Storage Technologist Foundations Course, Section: Storage Systems Architecture

The **data transfer rate** describes the MB/second that the drive can deliver data to the HBA. Internal and external factors can impact performance transfer rates.

**Read**
1. Data moves from the disk platters to the heads.
2. Data moves from the heads to the drive’s internal buffer.
3. Data moves from the buffer through the interface to the rest of the system, shown here as an HBA.

**Write**
1. Data moves from the interface through the drive’s internal buffer.
2. Data moves from the buffer to the read/write heads.
3. Data moves from the disk heads to the platters.

---

5) **Choose the correct RAID level**

Disk subsystems that use RAID can achieve higher levels of performance if the correct RAID level is used for the data access pattern produced by the application hosted on those disks. Commonly used RAID levels are RAID 0, RAID 1, RAID 5, and a combination of RAID 0 and RAID 1, often referred to as RAID 0+1 or RAID 1+0.

RAID 0 stripes data across multiple disks, but adds no redundant data for protection. Performance is good because of the striping, but the entire disk set will be unavailable if one or more disks in the set fail.

RAID 1 mirrors data onto two disks, a primary and a secondary. Each disk holds a full copy of the data, so if one disk fails, the data is still accessible from the other. Read performance is good; write performance is somewhat slower because two copies of the data are written for each host write.

RAID 5 stripes data across multiple disks, and adds parity protection to that data. Read performance is good; write performance can be poor in environments that perform a high proportion of random writes with small I/O sizes.
Choosing the right RAID Level for an e-mail application

Let’s use an e-mail system as the example of an application and see what happens when RAID 5, and then RAID 1 + 0, is used to store that data.

E-mail systems typically use small data blocks—4 KB is common. The data access pattern tends to be very random, largely because of the large number of users accessing the storage at any time. A write to an e-mail system using RAID 5 will incur a fairly high performance penalty—the old data and parity must be read and used in calculations with the new data to obtain the new parity. The new data and the new parity can then be written to disk. A single write from the host has resulted in four disk operations: two reads and two writes.

A write to a RAID 1 + 0 system will result in two disk writes—one to the primary disk in the mirror, and one to the secondary disk. Only half the disk operations are performed than was the case with a RAID 5 write, leading to better performance in this environment from RAID 1 + 0 than from RAID 5.

6 Align data to the disk structure

Some operating systems reserve a portion of the data area at the beginning of a disk for special operating system data. This reserved area may consist of an odd number of disk sectors and may cause user data to be misaligned with the formatted data structure on the disk. In disk subsystems that use striping, and especially striping with parity, this data misalignment may cause performance degradation. Data should be aligned by using the utility offered by the specific operating system. As an example, Windows allows data alignment with the diskpar and diskpart utilities.

Let’s take a look at another example. We’ll use RAID 5 storage, with a stripe depth of 64 KB, and a host data access pattern of 64 KB I/Os. If data is aligned at the 64 KB level, then each host I/O will involve only one disk in the RAID 5 disk group. If the data is not aligned, then each host I/O will involve accesses to two disks. In the event of a write, this will mean that eight disk operations will be performed—four for each of the physical disks. This doubling of disk operations without a corresponding increase in host operations is costly in terms of performance.

In closing, note that the topics we’ve looked at could be covered in more detail, and additional topics could be added. Among those are SAN design and its effect on performance, optimizing host storage systems and host applications, and analyzing performance data. But let’s bring it up to a lighter, higher level conclusion. Remember why we bother with performance considerations in the first place...the text book response would be “to increase the utilization of expensive data center components.”

Another way to think of it is as follows: disk performance and storage professionals have parallel interests. The art of performance monitoring requires both (data center components and you) to work efficiently and push performance limits to avoid being underutilized and a bottleneck. Consider the above suggestions as storage words of wisdom to increase your personal bandwidth throughout the day.
About the Author

André Rossouw has worked in the IT industry since CP/M was a state-of-the-art operating system for small computers. His roles have included Technical Support Engineer for a repair center environment, Customer Service Engineer, Course Developer, and Instructor. Certifications include A+, Network+, and he was previously an MCNE, MCNI, and SCO Authorized Instructor. He lives in North Carolina with his wife, daughter, and a garden full of squirrels.

About EMC Education Services

EMC® Education Services offers extensive education on EMC’s storage management software and solutions. Leveraging the EMC Proven™ Professional training and certification framework, EMC curriculum provides flexible learning options within a proven learning structure.

For more information about storage design and management, EMC Education Services has recently developed Storage Technology Foundations (STF), the first course in the Storage Design and Management curriculum. The course provides a comprehensive introduction to storage technology which will enable you to make more-informed decisions in an increasingly complex IT environment. You will learn about the latest technologies in storage networking (including SAN, NAS, CAS, IP-SAN), DAS, business continuity, backup and recovery, information security, storage architecture, and storage subsystems.

The unique, open course, which focuses on concepts, principles, and design considerations of storage technology rather than on specific products, is available from EMC Education Services as well as through Global Knowledge and Learning Tree International.

For course schedules and locations, please visit: http://education.EMC.com.

EMC², EMC, and where information lives are registered trademarks and EMC Proven is a trademark of EMC Corporation. All other trademarks used herein are the property of their respective owners.

© Copyright 2006 EMC Corporation. All rights reserved.