AIX and Flash
By Jaqui Lynch

Introduction

Flash technology is becoming more pervasive in data centers and it is expensive, so it is important to understand where and when it can benefit your workloads. The implementation varies amongst competitors with some mixing SSDs with flash and with others using pure Flash. Additionally, various vendors target their flash systems differently – some aim at simple cache acceleration while others aim at using it as a permanent data tier.

Flash should not be compared to spinning disks on a cost per capacity basis – you can install many terabytes of spinning disks for much less than flash. However, if performance matters, then cost per transaction or, more likely, cost per IOPS (I/O per second) is a significant factor. From a performance perspective you have Flash at the top, then SSDs and then spinning disks. Flash can easily have 10 to 50 times less latency than spinning disks as long as the array is correctly configured and the data is laid out appropriately. Details on this are provided in the reference on FlashSystems and SAP below.

IBM Offerings

IBM currently has 2 offerings in their all-flash arrays; the FlashSystem 900 and the FlashSystem v9000. These are all flash storage systems that offer superb performance, low latency, very high IOPS and advanced flash management. As a non-storage person I was quickly able to zone my FlashSystem 900 to make it usable to my POWER servers.

One significant feature of IBM FlashSystems is the ability to speed up access using preferred reads. Basically, the data in stored on a slower medium (spinning disk) and mirrored to the FlashSystem. The system is then told to read data in that filesystem from the fastest storage while still mirroring any writes between the devices. This provides for full redundancy without the need to purchase two FlashSystem arrays. You also have the option of mirroring between two FlashSystem arrays rather than spinning disk (or SSD) and flash.

FlashSystem 900

The FlashSystem 900 can scale from 2TB to 57TB (usable) in a single system using up to 12 hot swappable MicroLatency flash modules. These modules are either 1.2TB, 2.9TB or 5.7TB in size. All modules are the same size so the module chosen impacts the total usable potential for the FlashSystem. The FlashSystem uses a form of variable stripe RAID-5 to maintain performance and capacity if a flash chip was to fail. The system constantly monitors the chips and reports any issues. To ensure reliability the system provides for redundant hot swappable controllers, interface cards, power supplies, batteries and fans.

Minimum write latency can be as low as 90us and minimum read latency as low as 155us. The maximum IOPS for 4KB reads (100% random) is listed as 1,100,000. The array supports up to 16 x 8Gb fibre channel connections although there are options for 8Gb, 10Gb infiniband, 10Gb FCoE and 10Gb iSCSI. All of this in just 2U in the rack with a very minimal draw on power.

FlashSystem V9000

The FlashSystem V9000 is designed to provide a virtualized Flash environment. Storage behind the V9000 can be virtualized, making it easy to move data between external and internal storage non-disruptively. This provides for easy data migration between spinning disks and the FlashSystem. The FlashSystem V9000 is where the advanced software functions such as Real-Time Compression, external storage virtualization, dynamic tiering, thin provisioning, snapshots, cloning, replication, data copy services and high-availability configurations are available depending on which solution you choose.

The V9000 supports a maximum of four building blocks and four additional storage enclosures, each of which can support 2.2TB to 57TB of RAID-5 usable storage. So total usable storage can range from 2.2TB to 456TB in the internal enclosures. There is also provision for up to 32PB usable in external storage. Maximum IOPS for 4KB reads (100% random) can get to as high as 2,520,000 depending on design and the number of building blocks. The V9000 supports 16 x 16/8/4Gb fibre channel connections with options for 10Gb FCoE and 10Gb iSCSI.

The V9000 is very scalable, supporting up to 2048 LUNs (logical unit numbers) per building block with the system being able to support up to 2048 host connections with up to 256 host connections for each interface port.

When and Where to use FlashSystems

Performance is the reason most people look at flash. The best way to get that performance is to pay attention to recommendations for getting the most out of the flash. This may involve using EasyTier and allowing it to auto balance data across the various tiers (spinning disk, SSDs, flash, etc) or it may involve setting up specific volume groups that are flash, disk or SSD only and allocating the data there yourself.

Flash can be used for many things. It comes down to what matters for you for performance. As an example, I have seen it used for /saswork on SAS systems. Not many people would think of using expensive flash for work space, but due to the nature of SAS this makes a significant difference to performance. I have also seen it used for the database for IBM Spectrum Protect (was Tivoli Storage Manager), for Oracle redo logs and for actual data. I have seen applications such as oracle, DB2, SAS and SAP take advantage of flash and I am sure there are many more. Determining what to put on the flash normally involves running a POC (proof of concept) or some kind of test. The tests could be as simple as

using the ndisk component of nstress or you may want to reach out to your business partner or IBM team and ask about using the FlashSystem POC Toolkit.

Setting up Flash on AIX for Performance

There are some specific recommendations on how to set up flash for AIX. In general, the best performance is obtained when there are at least 32 LUNs in the volume group. You should also ensure that the FC SCSI I/O controller (fscsi0 …) is set to fast fail and dynamic tracking. This can be done as follows (using fscsi0 as an example):

chdev -l fscsi0 -a fc_err_recov=fast_fail
chdev -l fscsi0 -a dyntrk=yes

Any time the fibre adapters are deleted and re-added you will need to reset these. You will also want to increase the adapter queue size (num_cmd_elems) on the fibre adapters and, most likely, the DMA size (max_xfer_size). By default they are set to 200 and 16MB respectively. The command below set them to 2048 and 128MB. A reboot is required after the change to activate it.

chdev -l fcs0 -a num_cmd_elems=2048 -a max_xfer_size=0x200000 -P

If you are using NPIV to attach the flash then num_cmd_elems cannot be set to more than 256 per fibre adapter (FCS) as of AIX v6.1 tl8 and AIX v7.1 tl2). However, the VIO servers can still be set to the higher values and need to be the same as or larger than the largest client LPAR FCS setting.

Summary

There are many uses for flash and IBM offers two excellent options to choose from, the FlashSystem 900 and the FlashSystem V9000. The decision to purchase a FlashSystem 900 versus a FlashSystem V9000 is going to be made based on whether or not you want to take advantage of the V9000's ability to virtualize storage and whether you need any of the advanced options such as cloning, thin provisioning, etc. Additionally, scalability is a consideration, given the ability of the V9000 to scale much higher than the 900.

Flash will definitely help improve performance for many workloads provided it is properly configured and implemented. Use of flash may well allow you to elongate the life of your current storage systems as the use of a small amount of flash combined with Easy Tier can provide a significant performance boost. This is a good time to look at how your overall I/O is performing and to review how flash may play a role in improving I/O performance.

References

Implementing IBM FlashSystem 900
http://www.redbooks.ibm.com/redbooks/pdfs/sg248271.pdf

Introducing and Implementing IBM FlashSystem v9000
http://www.redbooks.ibm.com/redbooks/pdfs/sg248273.pdf

IBM FlashSystem V9000 Product Guide Redpiece 5317
http://www.redbooks.ibm.com/redpapers/pdfs/redp5317.pdf

IBM FlashSystem home page
http://www-03.ibm.com/systems/storage/flash/index.html

IBM System Storage Reference Architecture featuring IBM FlashSystem for SAP Landscapes
http://www-03.ibm.com/support/techdocs/atsmastr.nsf/WebIndex/WP102587

IBM FlashSystem all-flash storage and Oracle Database 12c
http://www-01.ibm.com/support/docview.wss?uid=tss1prs5327&aid=1

IBM FlashSystem 900 Specifications
http://www-03.ibm.com/systems/storage/flash/900/specifications.html

nstress
https://www.ibm.com/developerworks/community/wikis/home?lang=en#!/wiki/Power%20Systems/page/nstress