The IOPS cheat sheet !

Over 35 recommendations, explanations and general know how's

Bas van kaam

Note : Let’s start at the beginning, IOPS stands for: input / output operations per second, which in general is either a read or write operation. Simply put, if you have a disk that is capable of doing a 100 IOPS, it means that it is theoretically capable of issuing a 100 read and or write operations per second. However, being able to issue a 100 read and or write operations isn’t the same as actually processing them, reading and writing data takes time. This is where latency comes in. If our disk subsystem can handle, or issue, a 100 IOPS but they are processed at around 20 milliseconds per operation (which is slow by the way), then it will only be able to actually handle 50 operation per second as apposed to the issued 100.

Note : In the above example the 20 milliseconds is what we would call the latency with which our operations are performed. It tells us how long it will take for a single IO request to take place, or be processed.

Note : Remember that a random IOPS number, on its own, doesn’t say anything. For example, we can do a million IOPS ! Well, ok, that’s nice, but how did you test? Were they read or write operations? If mixed what was the percentage reads vs writes? Writes are more resource intensive. Did you read from cache? What was the data block size? How many host and disk controllers were involved? What type of storage did you use? Was there RAID involved? Although using RAID will probably negatively impact the IOPS number, but still. The same applies to data tiering. Physical disks? Probably. If so, are we dealing with sequential or random reads and writes?

Note : In addition to the above, and this is probably the most important one, how much, in milliseconds, latency is involved? This will range from around 2 milliseconds, which is comparable to a locally installed physical disk, to 20+ milliseconds at which performance, if any, will be highly impacted, an overview:

1.	0 - 12 milliseconds	Is fine, the lower the number the better off you are.
2.	10 - 15 milliseconds	Still acceptable in most cases, user might notice a small delay.
3.	15 - 20 milliseconds	Step up and take action, most of your users won’t be happy.
4.	20 - 25 milliseconds	You might as well shut it all down.

Note : Latency tells us how long it takes to process a single read or write I/O request.

“ If you should remember anything from this guide, it should be that a high number of IOPS is useless unless latency is low ! Even with SSD’s which are capable of providing a huge number of IOPS compared to traditional HDD’s, latency matters “

Note : With ‘legacy’ physical disks, overall speed and latency greatly depends on the rotations, or revolutions, per minute (RPM) a certain disk is capable of, the laws of physics apply. Today we can classify hard disk speeds (HDD) as follows: 5400 rpm, 7200 rpm, 10.000 rpm and 15.000 rpm. A higher rpm equals higher read and write speeds. Another factor impacting performance is disk density. The higher the density the more data a disk is able to store on its ‘platter’ data will be written closer together on the disk and as a result the disks read/write head will have to travel shorter distances to access the data, resulting in higher read and write speeds. This may sound like a small note to some, but imagine having a SAN or Filer holding hundreds of disks. Having 15.000 rpm and high density disks makes a real difference!

Note : So when a random vendor tells you that their storage appliance is capable of doing an crazy high number of IOPS you probably have a few questions to ask them, right?! I think it’s also clear that the more IOPS we can actually process, as apposed to issue, per second the better our overall performance will be!

“ Latency is king, the less you have the faster you’ll infrastructure will be! “

Note : There is no standard when it comes to measuring IOPS! There are to many factors influencing overall performance and thus the number of IOPS.

Note : Not all IOPS are the same, sure, you could boil it down to it being either a read or a write, but that’s not the whole truth now is it? First of, read and writes can be random or sequential, reads can be re-read and writes can be re-written, single and multiple treads, reads and writes taking place at the same time, random writes directly followed by sequential reads of the same data, different block sizes of data that get read or written, ranging from bytes to Megabytes and all that’s in between, or a combination of the above.

Note : It is important to understand your application workloads and their characteristics with regards to the IOPS they need. This can be a very tricky process. Take block size for example (just one of many examples) having a huge amount of smaller data blocks as apposed to having a relatively small number of larger data blocks can make a huge difference, have a look at the article below, it uses some clear analogies. It’s also on my ‘reference materials’ list.

http://recoverymonkey.org/2012/07/26/an-explanation-of-iops-and-latency/

Note : Ask your storage provider for detailed test procedures, how did they test and what did they use. In addition, at a minimum you will want to know these three 'golden' parameter:

1. the latency, in MS, involved,
2. the read vs write ratio
3. data block sizes used.

Note : We already highlighted read and write IOPS, both will be part of your workload profile. However, a lot of application vendors will refer to an average amount of IOPS that is needed by their workload to guarantee acceptable performance. This is also referred to as Steady State IOPS. A term also used by Citrix when they refer to their VDI workloads. After a virtual Windows machine boots up, users login and applications are launched, your users will start their daily routines. Seen from an IOPS perspective, this is the Steady State. It is the average amount of read and write IOPS processed during a longer period of time, usually a few hours at least.

" Although the average amount of IOPS, or the Steady State, does tell us something, it isn't sufficient. We also need to focus on the peak activity measured between the boot and the Steady State phases, and size accordingly "

Note : When we mention the 20 / 80 read / write ratio we are usually referring to the Steady State. Something you may have heard about during one of the many MCS vs PVS discussions. As you can see, the Steady State consists mainly out of write I/O, however, the peaks that occur as part of the boot and or logon process will be much higher. Again, these rules primarily apply to VDI like workloads.

Note : There are several tools available helping us to measure the IOPS needed by Windows and the applications installed on top. By using these tools we can get a idea of the IOPS needed during the boot, logon and Steady State phases as mentioned earlier. We can use Performance Monitor for example, using certain perfmon counters it will tell us something about the reads and writes taking place as well as the total amount of IOPS and the Disk queue length, telling us how many IOPS are getting queued by Windows. Have a look at these counters:

1. Disk reads/sec - read IOPS
2. Disk writes/sec - write IOPS
3. Disk transfers/sec - total amount of IOPS
4. Current Disk Queue length - IOPS being queued by Windows.

Note : Here are some more interesting tools for you to have a look at, they will either calculate your current IOPS load or help you predict the configuration and IOPS needed based on your needs and wishes.

1. IOMeter - measures IOPS for a certain workload
    a. http://www.iometer.org/doc/downloads.html
2. ESXTOP - specific to ESX, provides certain disk states, totals, reads and writes.
    a. http://www.yellow-bricks.com/esxtop/
3. WMAROW - web interface, used to calculate performance, capacity, random IOPS.
    a. http://www.wmarow.com/strcalc/
4. The Cloud Calculator - web interface, disk RAID and IOPS calculator.
    a. http://www.thecloudcalculator.com/calculators/disk-raid-and-iops.html
5. Process Monitor - general analyses of IOPS
    a. http://technet.microsoft.com/en-us/sysinternals/bb896645.aspx
6. Login VSI - VDI workload generator, simulate user activity on your infrastructure.
    a. http://www.loginvsi.com/

Note : As we will see shortly, there is a distinct difference between the boot and logon phase. Both (can) create so called 'storms' also referred to as a boot storm and or a logon storm, potentially impacting overall performance. This is where the read IOPS peaks mentioned earlier come in.

Note : Take your time! It's important to get the numbers right, or as close as possible, and to check that what you've build holds up in practice. Check out some of the tooling mentioned above and use it wisely.

Note : make sure to check out the whitepaper below, it's written by Jim Moyle, if anybody knows anything about IOPS and all that's related, it's him! Great info, which I used for this document as well.

http://jimmoyle.com/wordpress/wp-content/uploads/downloads/2011/05/Windows_7_IOPS_for_VDI_a_Deep_Dive_1_0.pdf

Note : Know that storage throughput isn't the same as IOPS. When we need to be able to process large amounts of data then bandwidth becomes important, the number of GB/sec that can be processed. Although they do have an overlap, there is a distinct difference between the two. Check out the following article, it's short but spot on!

http://timradney.com/2012/07/06/iops-verses-throughput/

Note : Be aware that RAID configurations bring a write penalty, this is because of the parity bit that needs to be written as well. A write can't be fully completed until the both the data and the parity information are written to disk. The time it takes for the parity bit to be written to disk is what we refer to as the write penalty. Of course this does not apply to reads.

Note : When looking at VDI workloads, we can break it down into five separate phases: boot, user logon, application launch, the Steady State and logoff / shutdown.

Note : During the boot process, especially in large environments, dozens of virtual machines might be booted simultaneously creating the earlier highlighted boot storm. Booting a machine creates a huge spike in read I/O, as such, and depending on the IOPS available, booting multiple machines at once might negatively impact overall performance.

" If IOPS are limited, try (pre) booting your machines at night. Also, make sure your users can't reboot the machines them selves "

Note : Using the above methods will only get you so far, there might be several reasons why you may need to reboot multiple, if not all, machines during day time. Something to think about as your VDI environment might not be available for a certain period of time.

Note : Although we are primarily focussing on IOPS here, du note that the underlying disk subsystem isn't the only bottleneck per se. Don't rule out the storage controllers for example, they can only handle so much, CPU, memory and network might be a potential bottleneck as well. Also account for RAID penalties, huge amount of writes for a particular workload, data compression and or deduplication taking place, and so on.

Note : Logon storms are a bit different in that they will always take place during the morning / day. It isn't something we can schedule during the night, users will always first need to logon before they can start working. Although this may sound obvious, it's still something to be aware of.

Note : Logons generate high reads (not as high as during the boot process) and less writes, although it's near to equal. This is primarily due to software that starts during the logon phase and the way how user profiles are loaded. Using application streaming, folder redirection, flex profiles etc. will greatly enhance overall performance.

Note : Make sure you understand the difference between state full, aka persistent, and stateless VDI's. Both will, or might, have different storage requirements as well as (re) boot and logon schedules.

" Launching applications will generate high read I/O peaks and initial low writes. Chances are that after users logon they will start, either automatically or manually, their main applications. Again something to take into account as this will probably cause a application launch storm, although it's usually not recognized as such "

Note : Steady State, we already discussed this one earlier, this is where write I/O will take over from read I/O, on average this will be around the earlier mentioned 20 / 80 read / write ratio. So if we scale for the peaks, read as well as write, we can't go wrong, at least that's the theory.

Note : Today the market is full with solutions and products helping us to overcome the everlasting battle against IOPS. Some are 'patch' like solutions which help speed up our 'legacy' SAN and NAS environments using SSD's and flash orientated storage in combination with smart and flexible caching technologies. While other focus on converged like infrastructures and in-memory caching. These in-memory solutions are somewhat special, they don't necessarily increase the number of IOPS available, instead, they decrease the number of IOPS needed by the workload because writes go to RAM instead of disk, ultimately achieving the same, or even better, result(s). Check out this article :

http://virtualfeller.com/2014/06/30/the-latest-xenapp-7-5-readwrite-ratios/

And this one as well:

http://blogs.citrix.com/2014/07/22/diving-deeper-into-the-latest-xendesktop-7-5-iops-results/

Using any of these products will greatly increase the number of IOPS available (or decrease the number of IOPS needed) and decrease the latency that comes with it. Just know that no matter which solution you pick, you will still have to determine the number of IOPS needed and scale accordingly. Even when using the most enhanced IOPS accelerator today won't guarantee that you will be able to boot your entire VDI infrastructure during day time and won't run into any issues.

" By leveraging RAM for writes, aka RAM Cache with Overflow to Disk in terms of Citrix PVS write cache, we can significantly reduce the number of IOPS needed. In fact, Citrix claims to only need 1 to 2 IOPS per users on a XenApp environment without any complex configurations or hardware replacement "

Note : Although this guide isn't about which product or vendor does it best, I'd like to make an exception (since I already highlighted Citrix PVS as well) and point out ThinIO, find out why for your self. More information can be found here:

http://thinscaletechnology.com/thinio/

Note : Give your Anti Virus solution some extra attention, especially with regards to VDI. You will be glad that you did. I suggest you start here:

http://www.brianmadden.com/blogs/rubenspruijt/archive/2013/01/15/project-vrc-antivirus-impact-and-best-practices-on-vdi.aspx

Note : During logoff and shutdown we will see a large write peak and very little read activity. Just as we need to plan for peak reads at the time of boot and logon, the same applies for write peaks during logoff. But since we also scaled to handle the peak writes during the Steady State phase we should be fine. Again I'd like to emphasize that although there might be a ton of IOPS available, it's the speed with which they are handled that counts! Less latency equals higher speeds.

Hope this helped!

Resources and reference material used:

1. http://www.en.wikipedia.org
2. http://www.jimmoyle.com
3. http://www.recoverymonkey.org
4. http://www.brainmadden.com
5. http://www.timradney.com