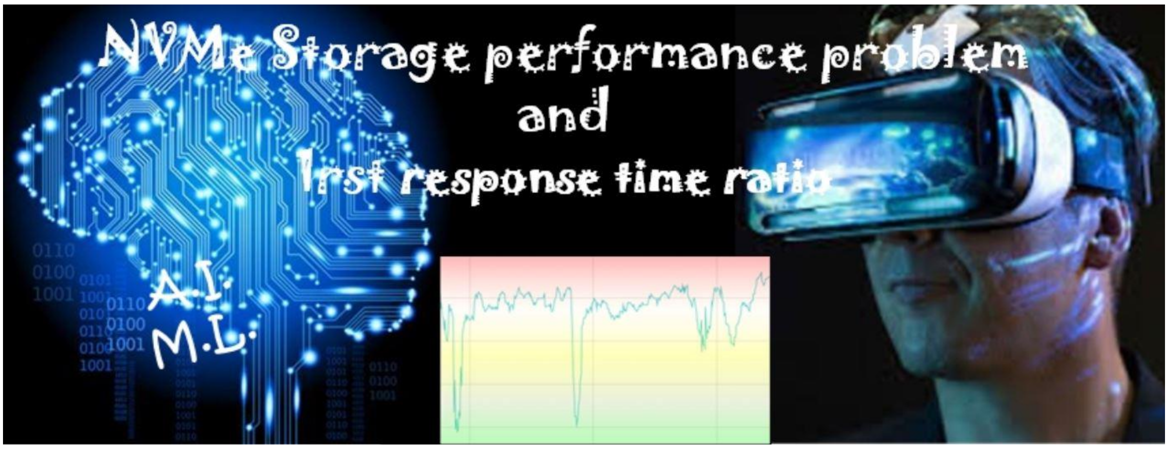


NVMe Data-Storage performance problem and 1st response time ratio.
NVMe 数据存储性能问题和第一响应时间比

业界已经创建了一种名为 Non-Volatile Memory Express（NVMe）的新块协议，它利用了非易失性存储器的性能特征 - 例如，闪存可以比 HDD 更容易地支持并行数据访问。 但是现实呢？ 如果您购买新的顶级 NVMe（内部和外部）企业级闪存数据存储，您是否认为可以实现宣布的性能？ 您是拥有还是拥有验证它的工具或方法？ 我将尝试解释，为什么检查它以及如何执行它很重要。。

The industry has created a new block protocol called Non-Volatile Memory Express (NVMe), which takes advantage of the performance characteristics of non-volatile memory - for example, flash memory can support parallel data access more easily than HDD. But what about reality? If you buy a new top-level NVMe (internal and external) enterprise-level flash data storage, do you think you can achieve declared performance? Do you own or have tools or methods to verify it? I'll try to explain why it's important to check it and how to implement it.

本文代表我个人的观点，想法发展和意见，不表达我的雇主或文章中提到的公司的观点或意见。

This article represents my personal views, idea development and opinions and do not express the views or opinions of my employer or companies named in the article.

Introduction

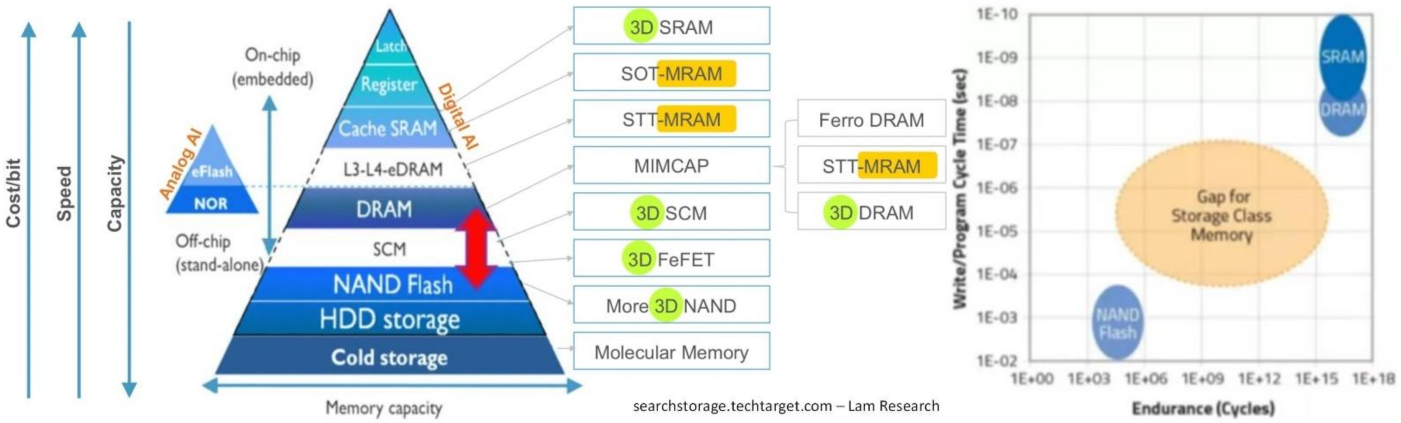
As NAND flash evolved, the SCSI protocol itself started limiting flash storage performance. So, the industry created a new block protocol called non-volatile memory express (NVMe) that capitalizes on the performance characteristics of non-volatile memory -- such as flash's ability to easily support data being accessed in parallel at a much greater degree than was ever imagined for HDDs.

The initial target for NVMe is PCI Express bus interfaces to unlock the SCSI performance bottleneck and so to run CPUs more efficiently than before. Connecting NVMe devices to PCIe slots reduces I/O overhead. That allows the devices and systems, they're on, to fully benefit from the parallelism of modern SSDs, bringing data closer to the processor.

Today's AI and machine learning applications are all about speed, processing data much faster than in the past. They also rely on much larger data sets, particularly to train smart system algorithms. AI and machine learning tools can require the scanning of millions, and even billions, of small files. New super computer and processor are hungry on data, this means you have to be able to deliver the data faster than they need, because waiting is too expensive. NVMe & NVMeoF provides the bandwidth and low latency that these demanding workloads require, making it a mainstream option for AI storage.

Recently, new classes of non-volatile memory (such as 3D XPoint) have emerged that work much faster than traditional 3D NAND flash and approaches dynamic RAM speeds. Called storage class memory (SCM), it also can be addressed at the byte level versus the page level of NAND memory.

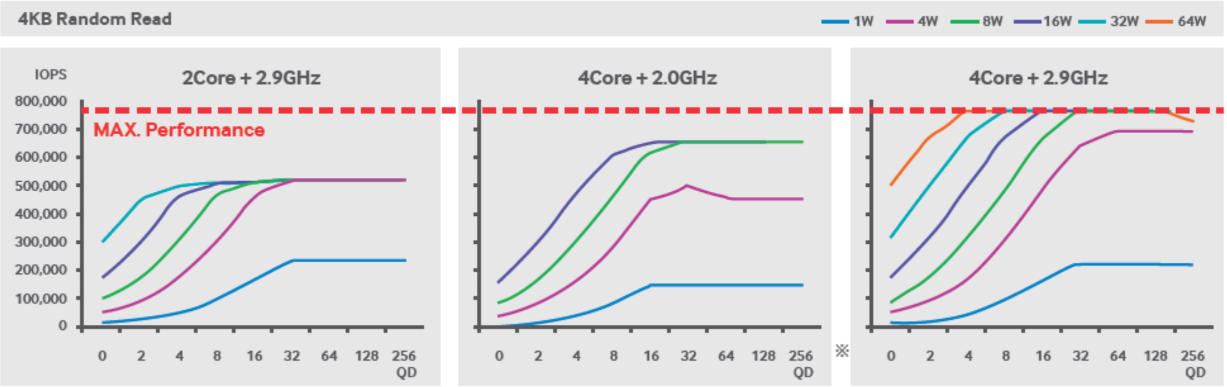
Traditional NAND-based devices are slower than DRAM and need careful monitoring to manage their life-cycle. Intel Optane memory (3D XPoint memory) is faster than NAND, denser than DRAM and non-volatile. While DRAM holds a charge and data in a single cell, Optane’s 3D XPoint technology has removed the charge so it is possible to keep writing to the cell and deliver greater capacity in the same space. Combining Optane memory media, controllers and interconnectedness between controllers reduces latency for hot-data workloads and is ideal for devices, applications and services requiring fast access to large amounts of data.



searchstorage.techtarget.com – Lam Research

Most all-flash arrays boast latencies of less than a millisecond, and many leading AFAs have latencies of less than 500 microseconds. SCM-based arrays using NVMeoF could improve latencies by another order of magnitude, approaching 50 microseconds. NVMe can do this because its command set requires less than half the number of CPU instructions to process an I/O request than SCSI and ATA command sets. NVMe supports 64,000 commands in one message queue and as many as 64,000 queues. A SAS device, on the other hand, supports only up to 256 commands per queue and SATA supports up to 32 commands.

We can see that the latency problem in the future or today already is no longer located at the storage disk self's but move to the other components (server/network/storage-Unit), i.e. described in a white paper *SM953_whitepaper-0.pdf*, Samsung write:" The best advantage of the NVMe is that it provides the highest performance by removing the HBA bottleneck by connecting directly to the CPU. The NVMe performance is affected by the CPU's core and frequency. For the highest performance, a certain number of cores and clock speeds are required. Figure 2-3 shows the evaluation results for the NVMe SSD, which provides the best performance with a four-core CPU and at least a 2.5GHz clock speed."



[Figure 2-3] NVMe Best Performance by Number of Cores and Frequency

The multi-core system typically used for a server system significantly degrades the NVMe performance. In the Non-Uniform Memory Access (NUMA) structure, the soft interrupt between the CPUs may cause low performance. In systems, data moves between the memory and a processor. But at times this exchange causes latency and power consumption, which is sometimes referred to as the memory wall.

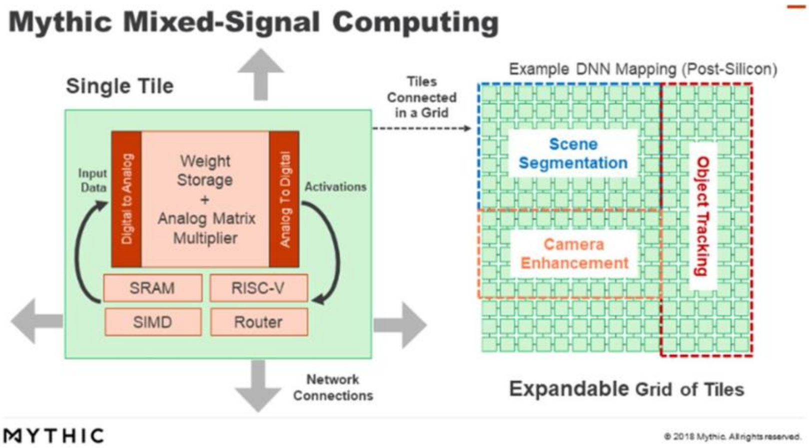
The industry is also working on other solutions to get performance for the AI. “Everybody is striving for a chip that has 100 Tera-OPS of performance,” said Steve Pawlowski, vice president of advanced computing solutions at Micron Technology. “But to get the efficiency of that chip, you must have several things going on simultaneously. This means having to bring data into the chip and get it out of the chip as fast as possible.” Additionally, CPUs are constructed of a few cores, which performs a single calculation at a time. In contrast, modern GPUs can contain thousands of cores that process separate threads simultaneously.

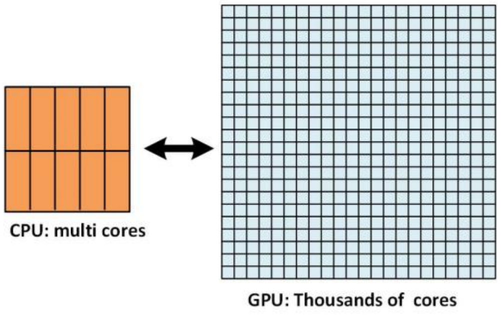
The nature of deep learning algorithms means they use an enormous amount of matrix math, making them well suited to execution on GPUs initially designed to make thousands of simultaneous floating-point calculations on pixel data. Unlike computer graphics, neural networks and other deep learning models don't require high-precision floating point results and are commonly accelerated further by a new generation of AI-optimized GPUs and CPUs that support low-precision 8- and 16-bit matrix calculations, an optimization that can turn storage systems into even bigger performance bottlenecks.

The diversity of deep learning models and data sources, along with the distributed computing designs commonly used for deep learning servers, means systems designed to provide storage for AI must address the following factors:

- A wide variety of data formats, including binary large object (BLOB) data, images, video, audio, text and structured data, which have different formats and I/O characteristics.
- Scale-out system architecture in which workloads are distributed across many systems, usually four to 16, for training and potentially hundreds or thousands for inference.
- Bandwidth and throughput that can rapidly deliver massive quantities of data to compute hardware.
- IOPS that can sustain high throughput regardless of the data characteristics; that is, for both many small transactions and fewer large transfers.
- Latency to deliver data with minimal lag since, as with virtual memory paging, the performance of training algorithms can significantly degrade when GPUs are kept waiting for new data. In addition, GPU are too expensive to let them wait.

That's where in- or near-memory computing fits, bringing memory closer or integrating it into processing tasks to boost the system. Both technologies are attractive for other reasons, it may give the industry another option besides traditional chip scaling.





Also, by AI the trend to in-memory computing grow. This is particularly true for applications such as computer vision in cars, where LiDAR and camera sensors will generate streaming video, and for artificial intelligence/machine learning/deep learning, where large volumes of data need to be processed quickly. "If you can process data where it resides, it's much more efficient," said Dan Bouvier, chief architect of client products at AMD.

You will also have to answer a question if you delete the source data (cache / fast memory size is limited) after it has been parsed or keep a copy, so you can occasionally re-analyse (offline) it, so hold open a possibility for later code improvement. This will affect your network and data storage design.

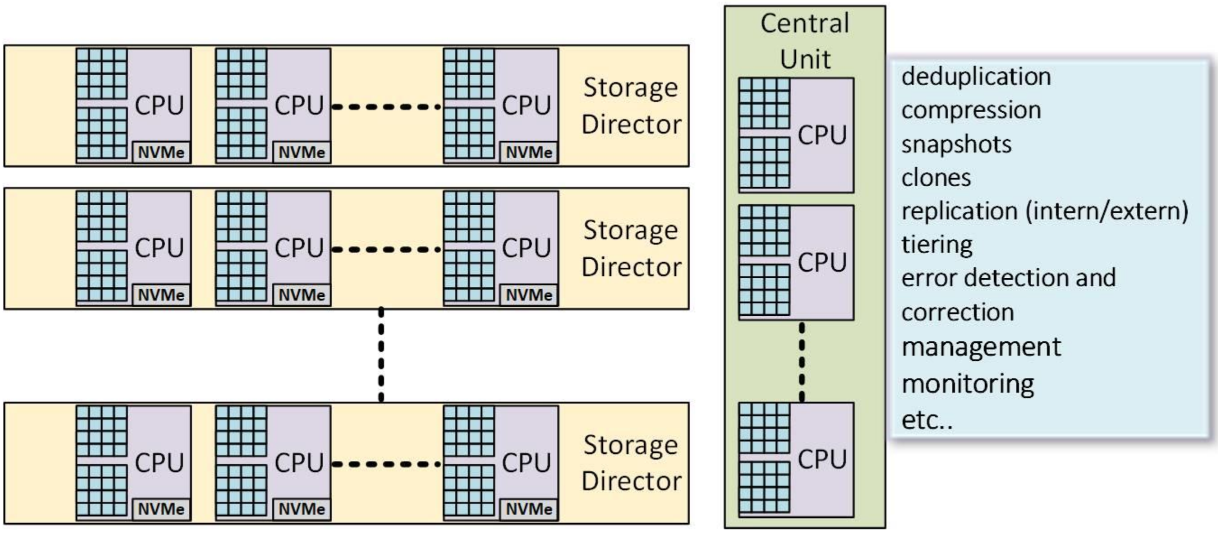
As you can see, there is no scope for latencies, not powerful or not well tested hardware.

Exposing other weaknesses

Enterprise storage systems are made up of media, the network and storage software, each of which can contribute to latency issues. However, as NVMe flash storage eliminates media latency, issues associated with other parts of the storage architecture become much more visible. This is made clear by the fact that most storage systems today are able to deliver only a fraction of the total raw performance of the NVMe drives they use. NVMe-oF will eventually eliminate latency issues associated with the storage network, leaving the software as the main latency culprit.

Marc Staimer (Dragon Slayer Consulting) wrote:" "The root cause of this NVMe performance challenge isn't hardware. It's storage software that wasn't designed for CPU efficiency. Why bother with efficiency when CPU performance was doubling every 18 to 24 months? Features, such as deduplication, compression, snapshots, clones, replication, tiering and error detection and correction, were continually added to storage software. And many of these features were CPU intensive. When storage software is consuming CPU resources, they aren't available for storage I/O to the high-performance drives.

Some believe storage class memory (SCM), the next-generation of non-volatile memory, will fix this NVMe performance challenge. It won't. SCM technologies will only exacerbate it, because their increased performance puts even more load pressure on the CPU."



The NVMe protocol exposes components, specifically the software, used to hiding behind media latency in the environment. Storage vendors have taken four approaches to try to improve storage software performance:

- Keep software basically the same but combine it with more powerful processors. The problem is that the standard Intel processors driving most of these software offerings have improved performance by increasing the number of cores, not the performance of each core.
- Throwing more CPUs -- servers or storage controllers -- and interconnect at it. This is the most common approach, but it comes with a high cost and diminishing marginal returns.
- Turn software into hardware by using field-programmable gate arrays (FPGAs) or even custom silicon. Turning software into hardware enables storage services to run on dedicated hardware and processing. The FPGA (near-memory computing) or silicon approach adds cost versus using off-the-shelf Intel CPUs. It also makes software upgrades more difficult, and an organization will need to periodically reprogram the FPGAs in the storage system.
- Rewrite software from the ground up to take full advantage of various changes in hardware. Rewriting starts with creating truly parallel threads that can stripe across cores instead of being dedicated to one core. A rewrite should go further by also rewriting the algorithms for core functions like RAID, metadata tracking for snapshots, deduplication, replication and thin provisioning to ensure they are optimized for the high core count of today's processors and the very low latency of current storage media.

George Crump (Storage Switzerland) write "An easy way to verify this latency gap is to examine the raw performance of an NVMe-based flash drive. Many NVMe drives claim more than 500,000 IOPS. Yet most storage systems, even though they have 24 of these drives, deliver only 10% of the potential raw performance of a single drive. A typical name brand NVMe array with 24 NVMe drives may have the raw potential to deliver almost 12 Mio IOPS, but once the overhead of the storage ecosystem is factored in, it often only delivers less than 1 Mio IOPS"

In this context, the emerging new companies will be the winners, starting with the right innovative code and architecture, with no constrains on compatibility with legacy hardware and software, i.e. faster on the market, taking the full advantage of the new powerful technology.

The 1st response ratio

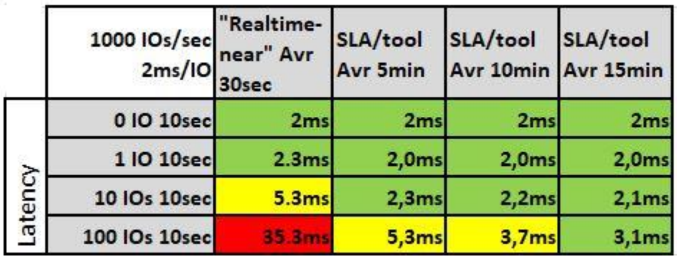
Now back to the harsh reality. What do you do if you want to buy a new high-end storage unit?

1. Invite the sales representative of a storage supplier
2. With ppt, he will present you the new features of the last new powerful hardware and lets you dream on unlimited new possibilities and performances.
3. Perhaps you may also do some "simple load / failover and availability test" and check the result with the supplier's tool, keeping the dream response time of 1 to 2 msec constant.
4. Buy the new storage and use it

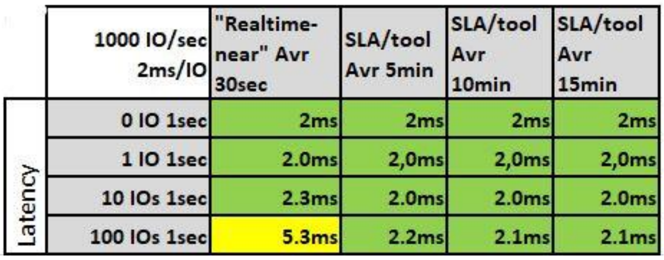
One difficulty of testing is to create a similar global IO load as will exist later. However, some performance issues are only visible from an IO/s load level. If you cannot check it out later on performance issues, you'll have to hear that this is your new AI application (code), which is causing a performance issue, or the transport network, but not data-storage unit, since it's up to that moment, no visible detectable problem was.

Never forget that the data storage companies are under time pressure to bring ever faster new products on the market, due to their competitors and rapidly developing technology (yesterday flash, today NVMe, tomorrow SCM etc..), very fast increasing performance requirements and less time for testing the newly developed systems

But how about the real serious test on the IO level? Not a lot of companies are able to do it, there is often a shortage of measuring equipment, and also most of storage suppliers are not granting IO level performance but SLA level 5/10/15min average response time.

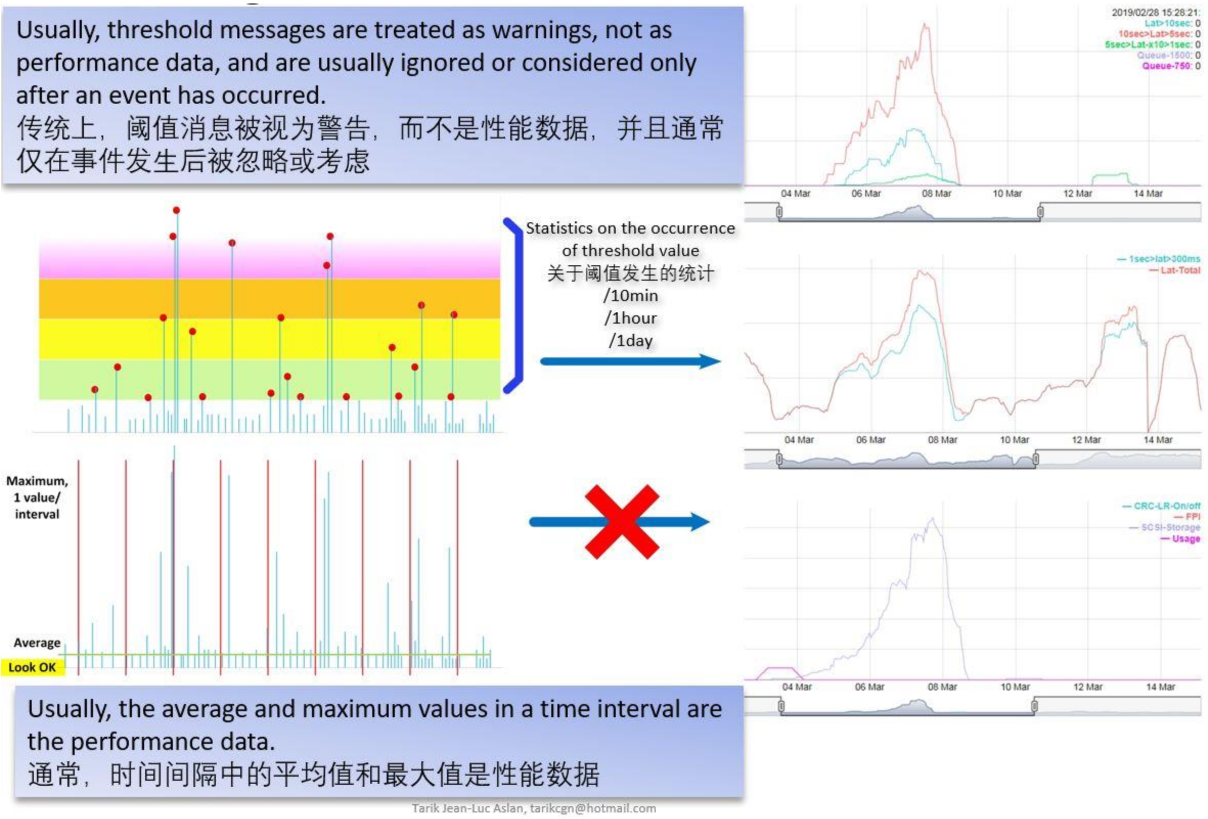


	1000 IOs/sec 2ms/IO	"Realtime-near" Avr 30sec	SLA/tool Avr 5min	SLA/tool Avr 10min	SLA/tool Avr 15min
Latency	0 IO 10sec	2ms	2ms	2ms	2ms
	1 IO 10sec	2.3ms	2,0ms	2,0ms	2,0ms
	10 IOs 10sec	5.3ms	2,3ms	2,2ms	2,1ms
	100 IOs 10sec	35.3ms	5,3ms	3,7ms	3,1ms



	1000 IO/sec 2ms/IO	"Realtime-near" Avr 30sec	SLA/tool Avr 5min	SLA/tool Avr 10min	SLA/tool Avr 15min
Latency	0 IO 1sec	2ms	2ms	2ms	2ms
	1 IO 1sec	2.0ms	2,0ms	2,0ms	2,0ms
	10 IOs 1sec	2.3ms	2.0ms	2.0ms	2.0ms
	100 IOs 1sec	5.3ms	2.2ms	2.1ms	2.1ms

In my last article I introduced a new method, performance monitoring based on the statistics of events like IO-based latency is above i.e. the theshold-level1, the theshold-level2, the theshold-level3 inside a 10min/1h/24h time window.



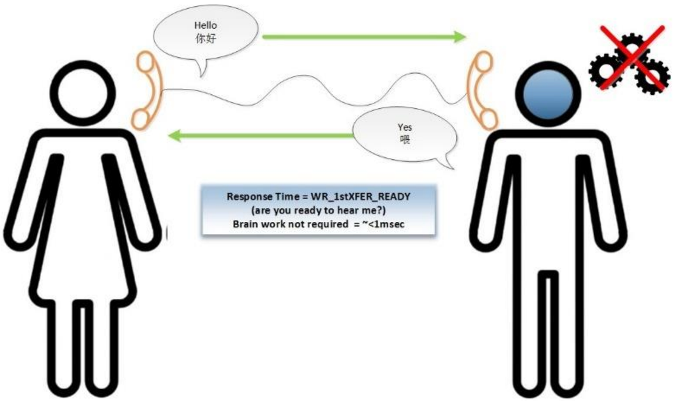
Usually, threshold messages are treated as warnings, not as performance data, and are usually ignored or considered only after an event has occurred.
传统上，阈值消息被视为警告，而不是性能数据，并且通常仅在事件发生后被忽略或考虑

Statistics on the occurrence of threshold value
关于阈值发生的统计
/10min
/1hour
/1day

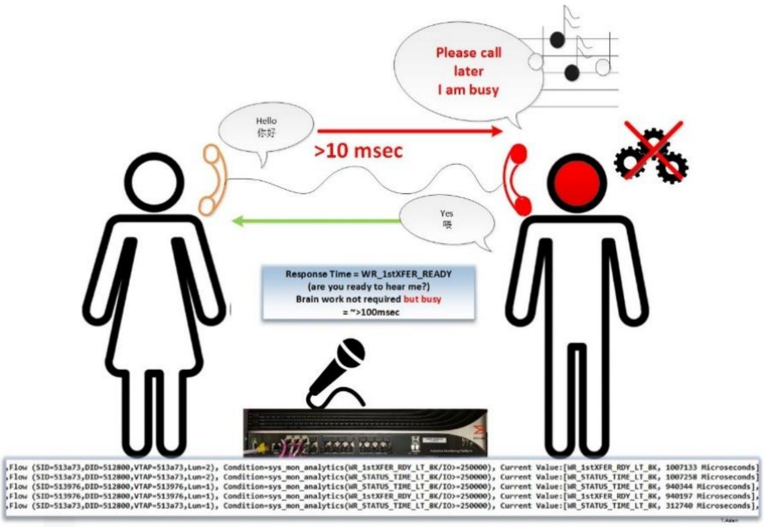
Usually, the average and maximum values in a time interval are the performance data.
通常，时间间隔中的平均值和最大值是性能数据

Tarik Jean-Luc Aslan, tarikcgn@hotmail.com

Only with the help of these statistics, it is not possible to understand the cause of these latencies. It lacks a feature, the **1st response ratio**. Let's refresh some protocol elements (from the 1st article) and explain the value of this feature:

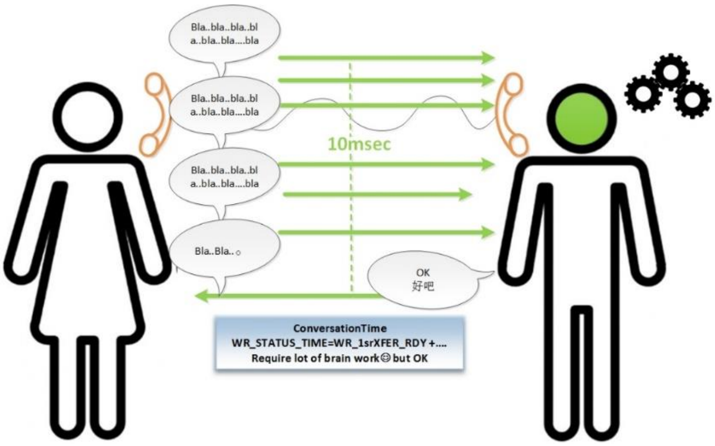

Hello
你好
Yes
喂
Response Time = WR_1stXFER_READY
(are you ready to hear me?)
Brain work not required = ~<1msec

The first response time (" are you ready to receive data")

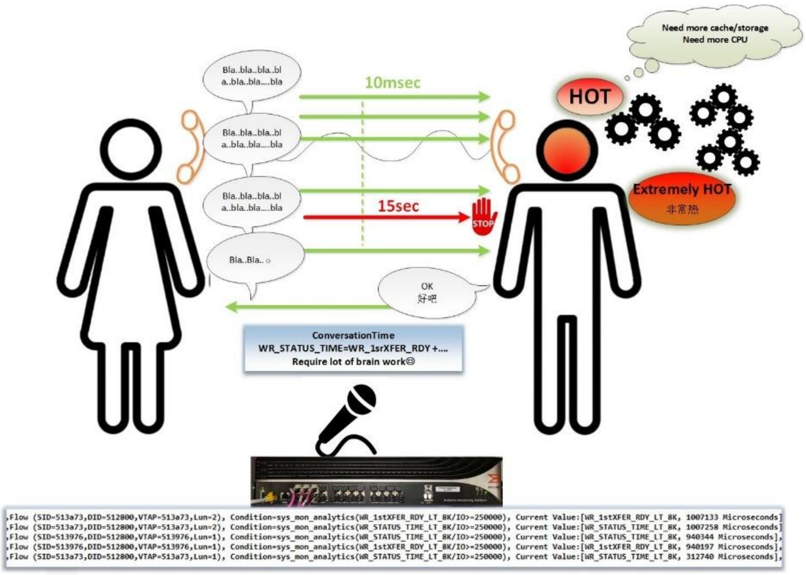

Please call
later
I am busy
Hello
你好
>10 msec
Yes
喂
Response Time = WR_1stXFER_READY
(are you ready to hear me?)
Brain work not required but busy
= ~>100msec

```
,Flow (SID=513a73,DID=512800,VTAP=513a73,Lun=2), Condition=sys_mon_analytics(WR_1stXFER_RDY_LT_8K/IO>=250000), Current Value:[WR_1stXFER_RDY_LT_8K, 1007133 Microseconds]
,Flow (SID=513a73,DID=512800,VTAP=513a73,Lun=2), Condition=sys_mon_analytics(WR_STATUS_TIME_LT_8K/IO>=250000), Current Value:[WR_STATUS_TIME_LT_8K, 1007258 Microseconds]
,Flow (SID=513976,DID=512800,VTAP=513976,Lun=1), Condition=sys_mon_analytics(WR_STATUS_TIME_LT_8K/IO>=250000), Current Value:[WR_STATUS_TIME_LT_8K, 940344 Microseconds],
,Flow (SID=513976,DID=512800,VTAP=513976,Lun=1), Condition=sys_mon_analytics(WR_1stXFER_RDY_LT_8K/IO>=250000), Current Value:[WR_1stXFER_RDY_LT_8K, 940197 Microseconds],
,Flow (SID=513a73,DID=512800,VTAP=513a73,Lun=1), Condition=sys_mon_analytics(WR_STATUS_TIME_LT_8K/IO>=250000), Current Value:[WR_STATUS_TIME_LT_8K, 312740 Microseconds],
```

The First response with latency:



The completion time with data transfer:



```
,Flow (SID=513a73,DID=512800,VTAP=513a73,Lun=2), Condition=sys_mon_analytics(WR_1stXFER_RDY_LT_8K/IO>=250000), Current Value:[WR_1stXFER_RDY_LT_8K, 1007133 Microseconds]
,Flow (SID=513a73,DID=512800,VTAP=513a73,Lun=2), Condition=sys_mon_analytics(WR_STATUS_TIME_LT_8K/IO>=250000), Current Value:[WR_STATUS_TIME_LT_8K, 1007258 Microseconds]
,Flow (SID=513976,DID=512800,VTAP=513976,Lun=1), Condition=sys_mon_analytics(WR_STATUS_TIME_LT_8K/IO>=250000), Current Value:[WR_STATUS_TIME_LT_8K, 940344 Microseconds],
,Flow (SID=513976,DID=512800,VTAP=513976,Lun=1), Condition=sys_mon_analytics(WR_1stXFER_RDY_LT_8K/IO>=250000), Current Value:[WR_1stXFER_RDY_LT_8K, 940197 Microseconds],
,Flow (SID=513a73,DID=512800,VTAP=513a73,Lun=1), Condition=sys_mon_analytics(WR_STATUS_TIME_LT_8K/IO>=250000), Current Value:[WR_STATUS_TIME_LT_8K, 312740 Microseconds],
```

The Completion time with delay in data transfer

Here is the definition:

$$1st\ Response\ Ratio = \frac{(Nr\ of\ 1st\ response\ Latency\ \times 100)}{Nr\ of\ completion\ Latency}\ \%$$

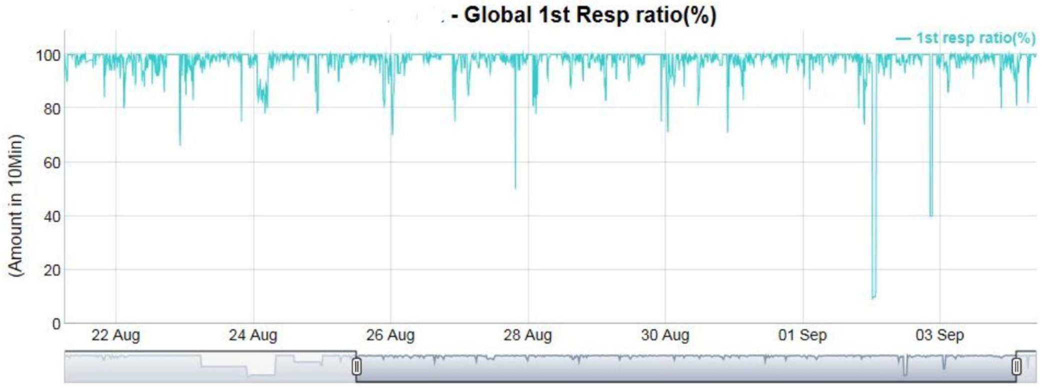
$$if\ Nr\ of\ completion\ Latency = 0\ (means\ no\ latency), 1st\ Response\ Ratio = -1$$

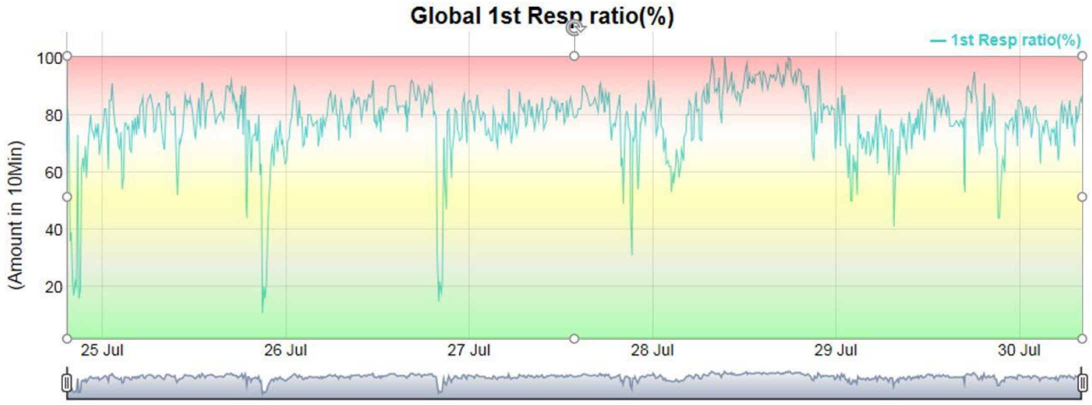
$$Completion\ time = 1st\ response\ time + data\ transfer\ \ time$$

$$Completion\ time \geq 1st\ response\ time$$

Here is the interpretation:

1st Response Ratio	1st Response time	Data Transfer time	Completion Time
0%	0%	100%	100%
50%	50%	50%	100%
100%	100%	0%	100%



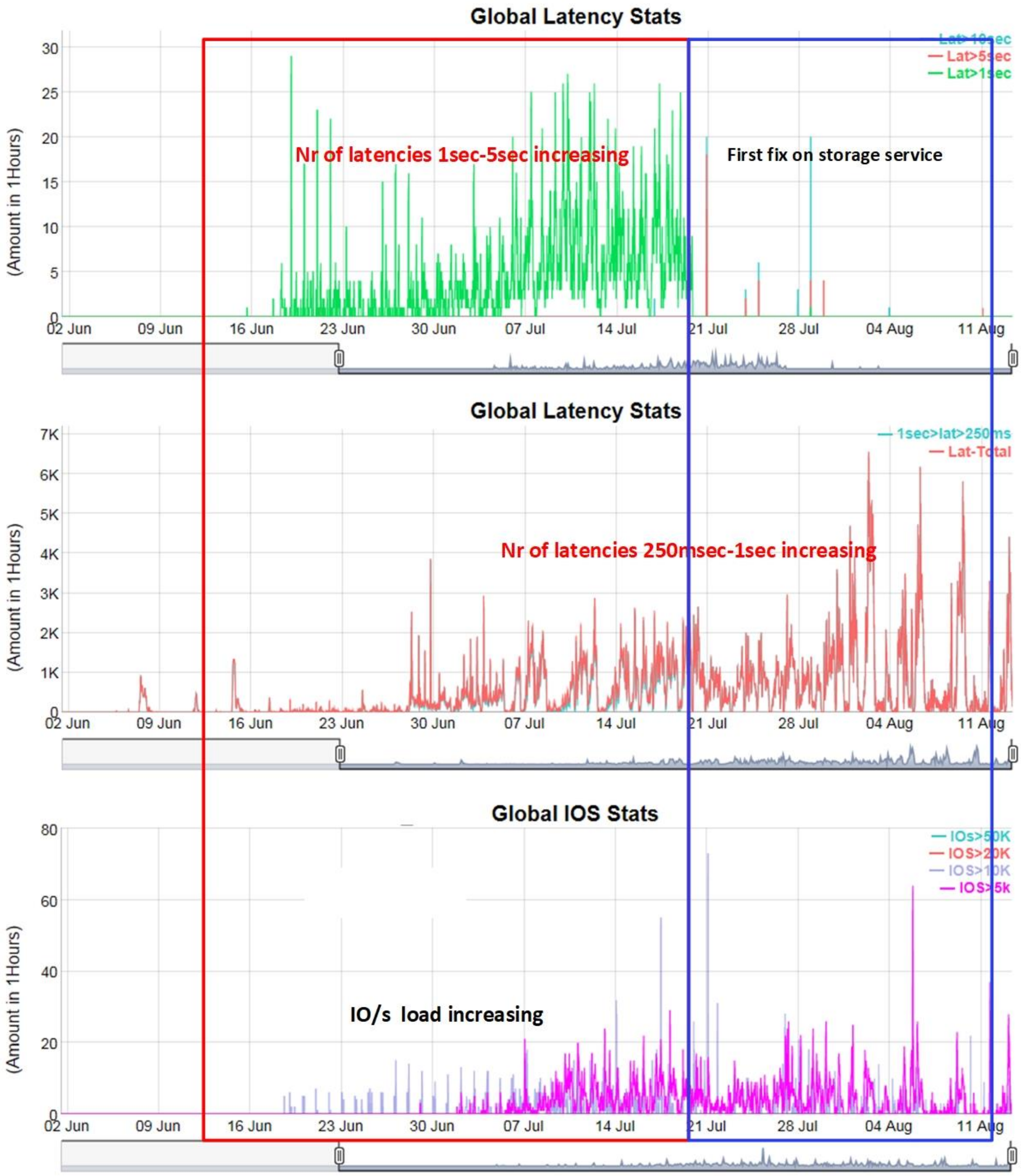


- At the storage port level:
  - 1st Response ratio >50%-100% too many IO/s at this port, you should use more path to other storage ports
  - 1st Response ratio = 0% possible storage Back-End performance problem or slow-drain, or data transfer blocks size >512KB
- At the storage director level:
  - 1st Response ratio >50-100% too many IO/s for this director, too few cores available (cores may be reserved for other purposes, such as synchronous replication between 2 storages and not normal traffic available)
- At the complete storage level:
  - 1st Response ratio 20-50% = too much IOs for this storage, you are waiting for the "I am ready"
  - 1st Response ratio >50% global storage performance issue, means the central unit responsible for many services can slow down your global traffic (i.e. by deduplication check the block footprint) if one of these services is not performant (code) enough. All your data transfers are waiting for the "I am ready"

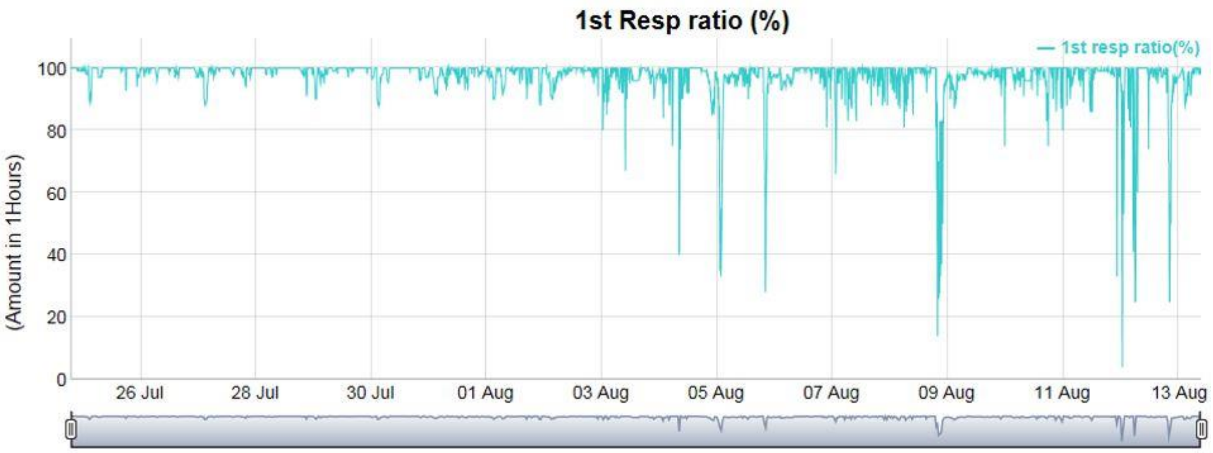
One user case

Let's first describe the start situation. We have a completely new storage model unit, new technology that is promising for high performance.

Using the method and performance dashboard from the last article, I was able to see the following performance issue at the Global storage level:

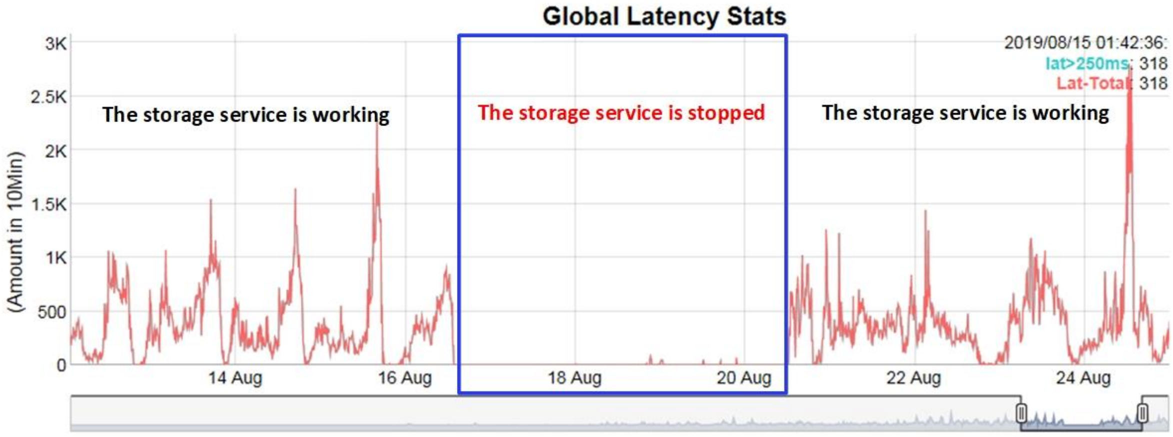


By introducing the 1st response ratio, I got the following, a nearly constant 100%



This means: 1st Response ratio >50% = global storage performance issue, means (Marc Staimer Dragon Slayer Consulting) the central unit responsible for a lot of service can slow down your global traffic (i.e. by deduplication check the block footprint) if one of these services is not performant enough.

How can you prove that?



Global Latency Stats

The storage service is working

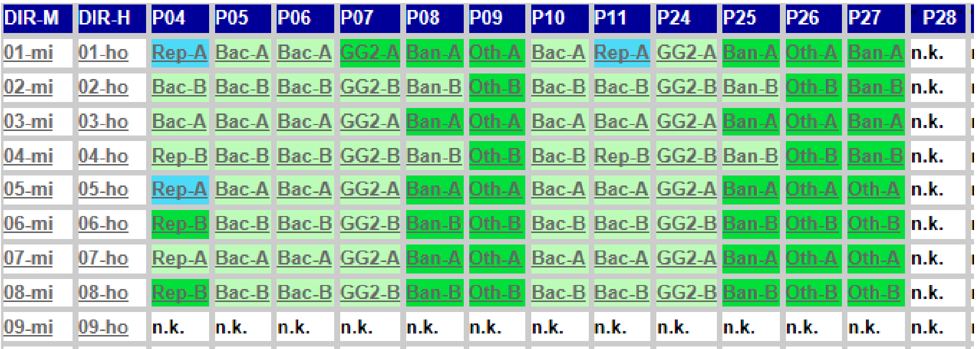
The storage service is stopped

The storage service is working

2019/08/15 01:42:36:
lat>250ms: 318
Lat-Total: 318

To the questions:

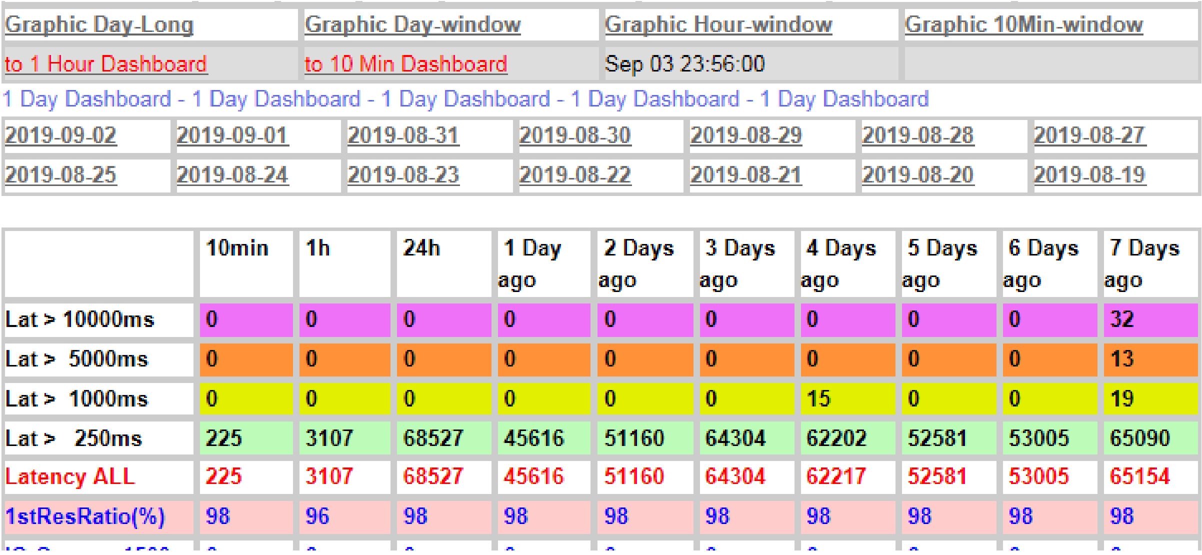
- Why did I think it was a global storage problem? I found this behaviour on different storage (same type), the latencies were on different directors, and different fabrics (light green).



DIR-M	DIR-H	P04	P05	P06	P07	P08	P09	P10	P11	P24	P25	P26	P27	P28
01-mi	01-ho	Rep-A	Bac-A	Bac-A	GG2-A	Ban-A	Oth-A	Bac-A	Rep-A	GG2-A	Ban-A	Oth-A	Ban-A	n.k.
02-mi	02-ho	Bac-B	Bac-B	Bac-B	GG2-B	Ban-B	Oth-B	Bac-B	Bac-B	GG2-B	Ban-B	Oth-B	Ban-B	n.k.
03-mi	03-ho	Bac-A	Bac-A	Bac-A	GG2-A	Ban-A	Oth-A	Bac-A	Bac-A	GG2-A	Ban-A	Oth-A	Ban-A	n.k.
04-mi	04-ho	Rep-B	Bac-B	Bac-B	GG2-B	Ban-B	Oth-B	Bac-B	Rep-B	GG2-B	Ban-B	Oth-B	Ban-B	n.k.
05-mi	05-ho	Rep-A	Bac-A	Bac-A	GG2-A	Ban-A	Oth-A	Bac-A	Bac-A	GG2-A	Ban-A	Oth-A	Oth-A	n.k.
06-mi	06-ho	Rep-B	Bac-B	Bac-B	GG2-B	Ban-B	Oth-B	Bac-B	Bac-B	GG2-B	Ban-B	Oth-B	Oth-B	n.k.
07-mi	07-ho	Rep-A	Bac-A	Bac-A	GG2-A	Ban-A	Oth-A	Bac-A	Bac-A	GG2-A	Ban-A	Oth-A	Oth-A	n.k.
08-mi	08-ho	Rep-B	Bac-B	Bac-B	GG2-B	Ban-B	Oth-B	Bac-B	Bac-B	GG2-B	Ban-B	Oth-B	Oth-B	n.k.
09-mi	09-ho	n.k.	n.k.	n.k.	n.k.	n.k.	n.k.	n.k.	n.k.	n.k.	n.k.	n.k.	n.k.	n.k.

- Did it have any customers impact? The answer is no, because the problem was found proactive very early on. The difficulty was convincing the support team and proving that the problem existed. But with this dashboard (and the Maps message of Amp`s Analytic monitoring platform Broadcom) it's a breeze. After that, this is the job of the storage developer/support team to find what is definitively blocking the box.

The ratio is now graphed from port level up to global level, and also in the statistical summary i.e. for this storage



Graphic Day-Long	Graphic Day-window	Graphic Hour-window	Graphic 10Min-window
to 1 Hour Dashboard	to 10 Min Dashboard	Sep 03 23:56:00	

1 Day Dashboard - 1 Day Dashboard - 1 Day Dashboard - 1 Day Dashboard - 1 Day Dashboard

2019-09-02	2019-09-01	2019-08-31	2019-08-30	2019-08-29	2019-08-28	2019-08-27
2019-08-25	2019-08-24	2019-08-23	2019-08-22	2019-08-21	2019-08-20	2019-08-19

	10min	1h	24h	1 Day ago	2 Days ago	3 Days ago	4 Days ago	5 Days ago	6 Days ago	7 Days ago
Lat > 10000ms	0	0	0	0	0	0	0	0	0	32
Lat > 5000ms	0	0	0	0	0	0	0	0	0	13
Lat > 1000ms	0	0	0	0	0	0	15	0	0	19
Lat > 250ms	225	3107	68527	45616	51160	64304	62202	52581	53005	65090
Latency ALL	225	3107	68527	45616	51160	64304	62217	52581	53005	65154
1stResRatio(%)	98	96	98	98	98	98	98	98	98	98

Conclusion:

We read about the rapid technology evolution, even the next, which is at the door (SCM). We also read about difficulties to mastering the new technology requirements inside storage units (compatibility with legacy hardware/software, not optimized software code, high pressure to bring new technology business ready etc..).

The 1st response ration quickly provide the correct information that it is at a local-port/director/global problem, even if it is a transport problem (completion time) or a core/software performance problem 1st response time, without xxx Gb data logs to be reviewed also the need to be visible in the supplier monitoring software.

I can only recommend making similar test regardless of the supplier software. This will save you later stress, time and impact on production performance. In the User case the supplier was needing first to trace the storage to obtain similar information and find the RC responsible for this behaviour.

Questioning the supplier why and how they could not see the problem during the development and testing of the new storage type is management business, but it shows the need to carry out supplier-independent thorough testing by yourself.

Here are the links to the first three articles:
Latency and performance monitoring in big Enterprise environments - challenges and vision. (2018.02.01)
Latency and performance monitoring in big Enterprise environments - Part 2: IO-latency monitoring and ROI. (2018.05.09)

Innovative Performance Monitoring for High-end Data Storage Area Network 高端数据存储区域网络的创新性能监控 (2019.07.02)

About the author:

Education: Electrical Engineer (master's degree), State University, Liege, Belgium
Complementary education: Applied Data science: machine learning, EPFL (Ecole Polytechnique Fédérale de Lausanne)
Certified : BCFA Gen5, BCFD Gen5, BCEFP 2015, SCSE, SCSA, SCSN-E, EMCISA-v2
Work experience:

- E & IT-Engineer : German Aerospace Center : parallel computing medical reasearch
- SAN Architect-engineer: German Telekom: SAN design, implementation, operation, support
- SAN Solution Architect / Senior Data Storage Engineer: Swisscom IT-services: SAN design, implementation, operation, support, automation, monitoring, new technology integration and development

Award: GTB Innovation Award 2017: Swisscom & Brocade - Project: Analytics monitoring platform
Languages: 中文，German, English, French
Contact:

Email: tarikcgn@hotmail.com
linkedin: http://www.linkedin.com/in/tarik-jean-luc-aslan-45a0a93b
WeChat: tarikzh (WeChat-QR)