Innovative Performance Monitoring for Data Storage Area Network

高端数据存储区域网络的创新性能监控

面部或车牌识别系统，在线支付系统，自动驾驶和其他需要实时回答。这种高负载系统对延迟没有太大的容忍度，它们需要始终如一的高性能和最短的响应时间，即使在最糟糕的情况下，例如双十一日或全国庆的安全检查（ID /人脸检测)。但是，AI 系统要求具有高 IOPS 性能的内存非常重要，尽管它取决于处理随机 I / O 操作的能力。“当你分析深度学习时，它是随机密集的，而输出可以忽略不计 - 它可以是千字节，"Gartner 的决定说。“它并不一定需要高 IOPS，而是一种可随意优化读取的架构。”根据 Gartner 的研究，到 2021 年，深度学习和 AI 等工作负载将成为数据中心设计和架构的关键因素。随着数据中心网络架构师为 AI 准备基础架构，他们必须优先考虑可扩展性，这需要高带宽，低延迟的网络和创新的架构。闪存现在很常见，而 NVMe Flash 是需要最快访问存储在 GPU 附近的数据的应用程序的首选媒体。但真正的性能数据呢？应用程序对响应时间的敏感度是多少？客户对应用程序的性能要求是什么？在本文中，我们将看到大多数供应商没有给您机会详细查看性能，或者由于缺乏概念而无法提供。几个月前，我发表了两篇文章：“大型企业环境中的延迟和性能监控 - 挑战和愿景”。 IO 延迟监控和 ROI"。在这篇，我想介绍一种用于性能监控和监控规划的创新方法，包括一些非常有用的用户案例。

Facial or license plate recognition system, online payment system, automatic driving and other needs to be answered in real time. Such high-load systems do not have much tolerance for latency, and also the users. They need consistent high performance and shortest response time, even in the worst case, such as the Double Eleventh Day or National Day security check (ID/Face Detection). However, AI systems require memory with high IOPS performance, although it depends on the ability to handle random I/O operations. "When you analyse in-depth learning, it's random and intensive, and the output is negligible - it can be kilobytes," Gartner decided. "It does not necessarily require a high IOPS, but rather a readable architecture that can be optimized at will." According to Gartner's research, workloads such as in-depth learning and AI will be key factors in data center design and architecture by 2021. As data center network architects prepare the infrastructure for AI, they must give priority to scalability, which requires high bandwidth, low latency networks and innovative architectures. Flash is now common, and NVMe Flash is the preferred medium for applications that need the fastest access to data stored near the GPU. But what about real performance data? How sensitive is the application to response time? What are the customer's performance requirements for the application? In this article, we'll see that most vendors don't give you the opportunity to look at performance in detail or are unable to provide it because of a lack of concepts. A few months ago, I published two articles: "Delay and Performance Monitoring in Large Enterprise Environments - Challenges and Vision". "IO delay monitoring and ROI". In this article, I want to introduce an innovative approach to performance monitoring and monitoring planning, including some very useful user stories.

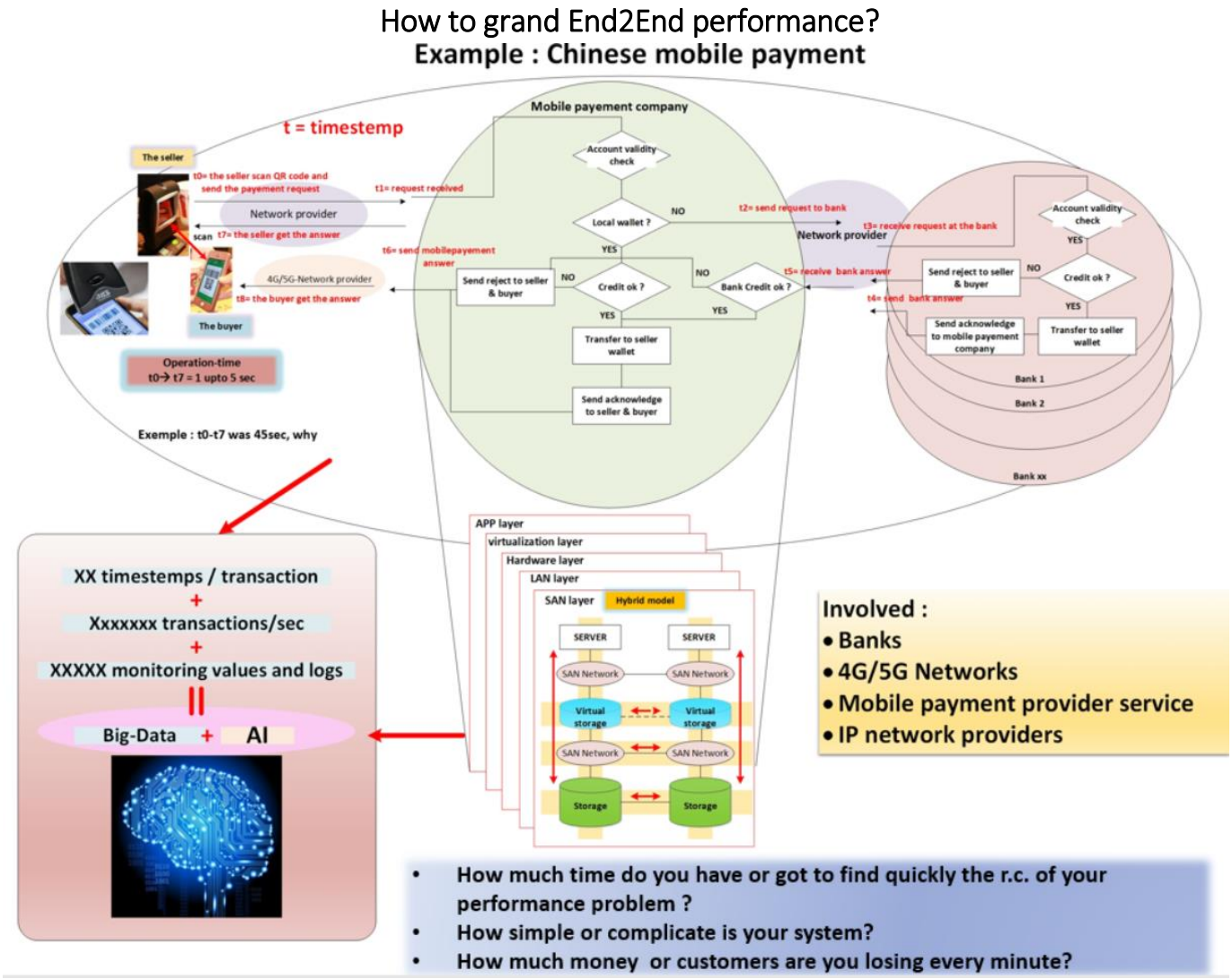
本文代表我个人的观点，想法发展和意见，不表达我的雇主或文章中提到的公司的观点或意见。

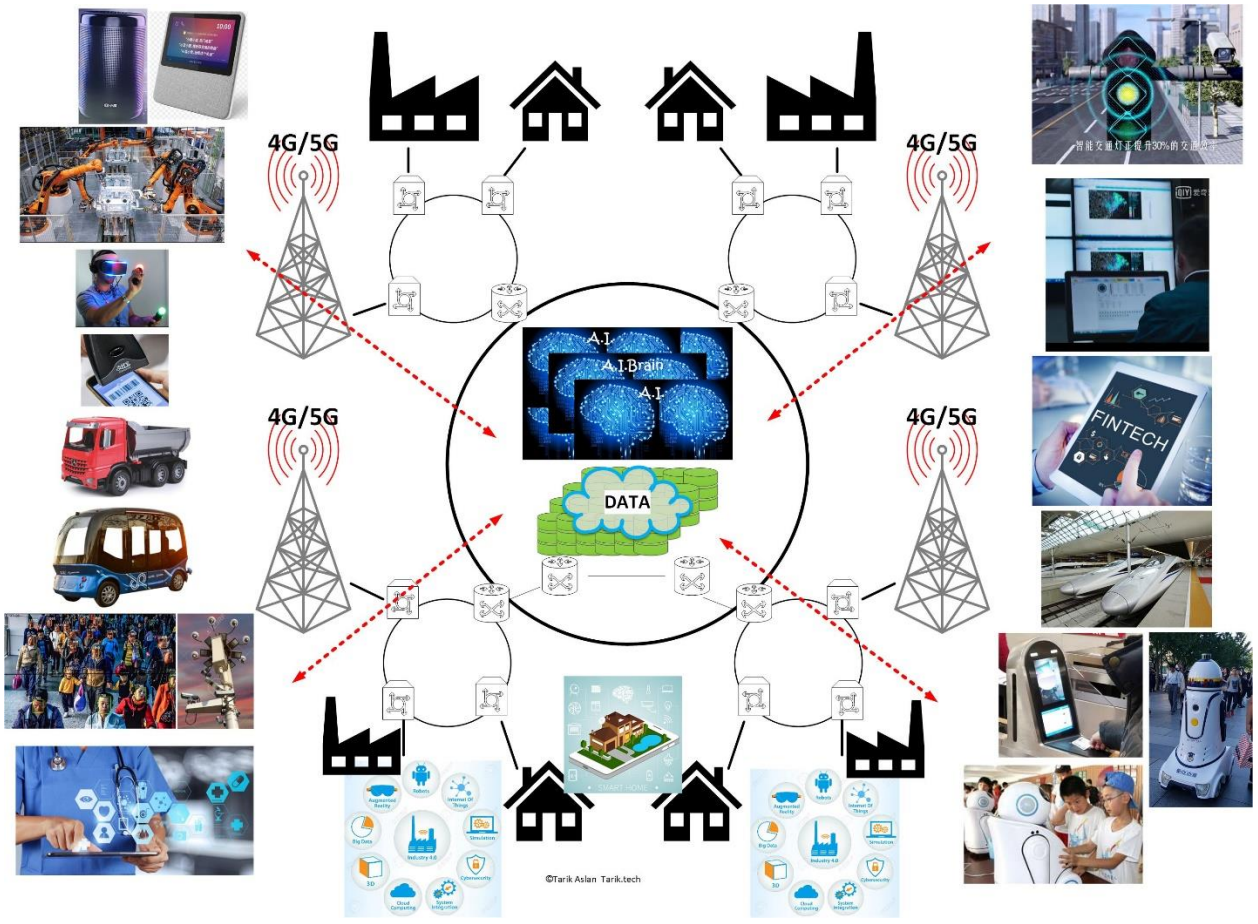
This article represents my personal views, idea development and opinions and do not express the views or opinions of my employer or companies named in the article.

Introduction

Nowadays the demand for performance by computing is getting higher and higher, also the huge amount of data that you need to analyze in short time or real time (automatic driving, AI, speech recognitions / assistant, global mobile payment, face / object / car recognition on the street subway / airport etc.). Meanwhile, they are also everywhere, at home soon (i.e. no home key anymore but face recognition, speech assistant to command any objects (小米 Just ask how you can guarantee that you have the performance you need 100% of the time? Latencies in automatic driving,

mobile payments, security checks and medical remote operations are unacceptable. Buying all-flash memory and supercomputers is not enough to guarantee this. The question is:





AI is a broad term that covers a wide range of use cases and applications, as well as different ways of processing data. Machine learning, deep learning, and neural networks all have their own hardware and software requirements and use data in different ways.

Machine learning is a subset of AI, and deep learning is a subset of machine learning.

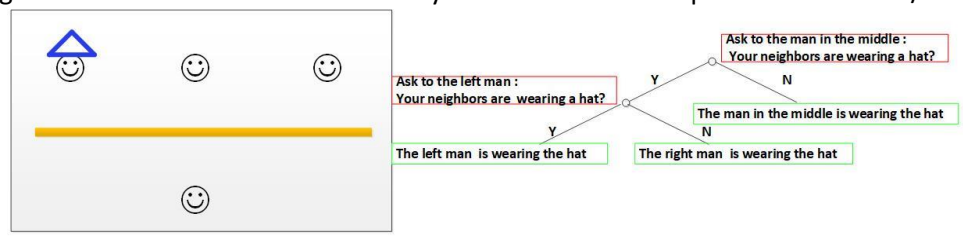
- Deep learning, for example, will carry out several passes of a data set to decide and learn from its predictions based on the data it reads. Based on the accuracy or inaccuracy of predictions, it can automatically re-learn or self-adjust how it learns from data. A deep learning application data set will be an order of magnitude larger, easily running to millions of data points.
- Machine learning is simpler and relies on human-written algorithms and training with known data to develop the ability to make predictions. If the results are incorrect, the data scientists will change the algorithms and retrain the model. A machine learning application could draw on thousands of data points.

Also the storage and I/O requirements of AI are not the same throughout its lifecycle. AI systems need training, and during that phase they will be more I/O-intensive, which is where they can make use of flash and NVMe

For example, suppose you manage a large data storage area network that may be a hybrid network (multiprotocol transport). What methods and data do you need to perform true performance monitoring for statistical analysis and automated machine learning analysis? The typical information "My 5-minute AVR response time is 2.5 milliseconds" is not interesting information, but we need information that describes network and data storage unit usage, as well as a criterion that show performance improvement or degradation. In such a large network, there is a large correlation between various parameters, namely transmission event and response time parameters (crc> scsi-timeout> high latency> bottleneck the new method should also allow the detection of configuration and load distribution problems or the recording of improvements. To be useful, it has to work almost in real time. In general, the goal is to provide a tool and method that lets you say everything in less than a minute, without having to log in to a storage or switch, and without much technical knowledge, regardless of the size of the network.

Weight of information

What is the weight or value of information? You may know the best example from Statistics / Probability Lecture:



You have 3 men in one room, one wears a hat, the other does not, you cannot see the men, but you can ask questions. The game says: What is the minimum of questions you should ask to find out who is wearing the hat. It means that each question brings information, some bring more information "heavy", some less useful information "light", that means you win by asking questions that bring the "heaviest" information (useful).

Similarly, if we want to monitor or qualify the performance of a system that can be very extensive with thousands of individual elements and different technologies, we need to look for the source of "heavy" information (which will be used as the primary source), and perhaps (ignore or use as a secondary source) "light" information.

In practice, lets show some "light" and "heavy" information in a data storage network environment

- 1rst example

I/O Latency

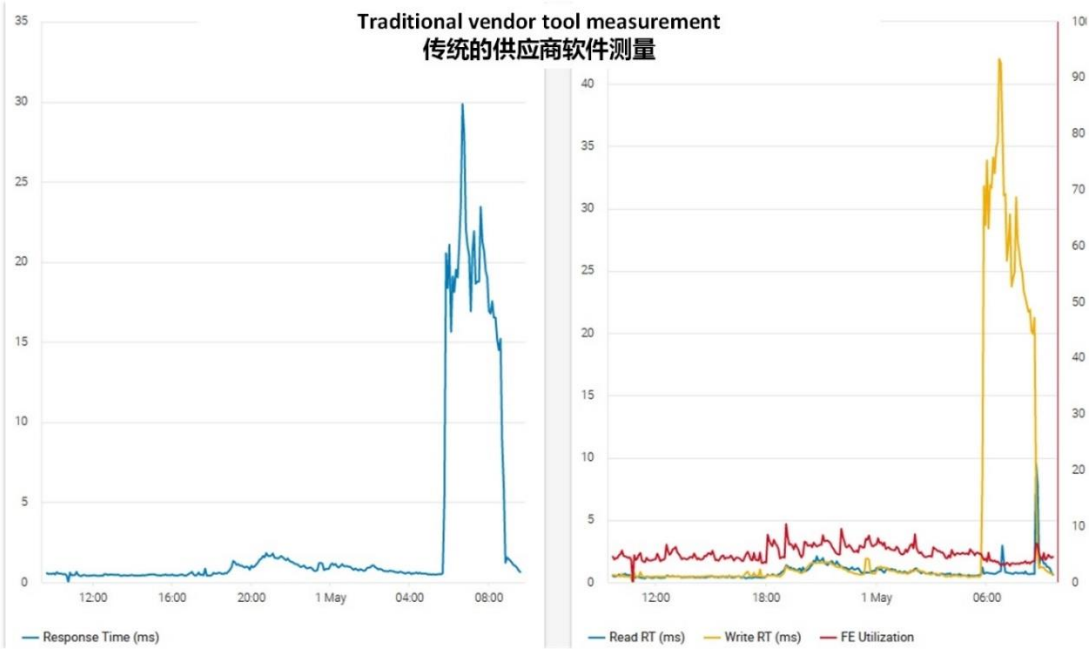
Metric	IO Size	Max			Avg		
		10sec	5min	All	10sec	5min	All
RD CMD -> Status Time	0K-<8K	139.9m	240.1m	2.896s	5.495m	6.269m	8.543m
	8K-<64K	41.78m	79.16m	2.546s	4.682m	7.467m	8.073m
	64K-<512K	29.03m	36.99m	331.8m	8.531m	5.895m	4.411m
	512K->512K						
	ALL	139.9m	240.1m	2.896s	5.481m	6.287m	8.537m

Here we can read the I / O latency information provided by a CLI command. You can see the average and maximum values in the last 10 seconds, 5 minutes and since the statistic was last reset, and all information for metric and block size. At first glance, the information may help a few, but nothing more. It lacks a lot like how many spikes are in the time interval, was it just high peaks or even medium? etc. The average value shows no critical value. So, the

question is, should we do something? If so, how and what? If not, let's wait and see if someone sends a complaint. This information here is a piece of weak "light" information, you cannot decide if you have a problem and you have no way to track the RC.

- 2d example

Imagine that you have a faulty interface adapter. This means that optical power is slowly becoming too low for clear data transport, consequence: this leads to frame transport errors, and automatically latency.

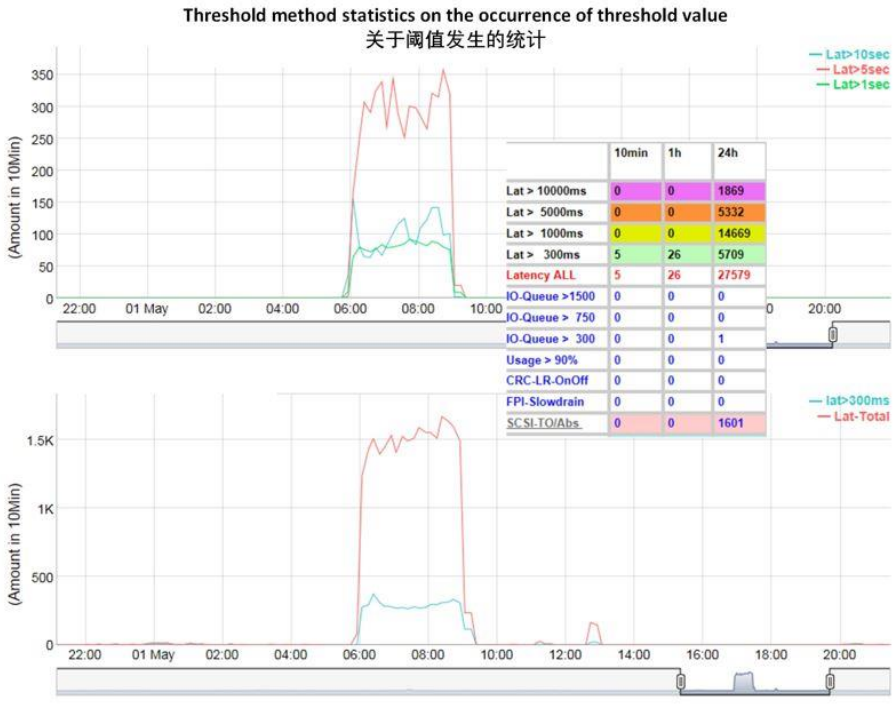


Here we see an increase in latency, but I would say it's always not so "critical" that it's only 30ms. This average is calculated in a time window, some are 5 minutes / 10 minutes / 15 minutes, depending on the supplier tool. You can also read the maximum value within this time window, it shows no value over 100 ms, even without specification, when, how often, etc. The information here is also weak "light", but you have the information that something happened. Remember 2 tables from the 2d article:

1000 IOs/sec 2ms/IO	"Realtime-near" Avr 30sec	SLA/tool Avr 5min	SLA/tool Avr 10min	SLA/tool Avr 15min
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
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

For this example, I recorded the following latency spikes with the new method:



- 3rd example:

The hat game from the introduction was about two aspects, how much useful "heavy" information your question can contain and how many questions you have to ask to solve the problem. It means that you need to have a clear workflow in mind to know the quickest way to solve the problem, not to store the details or weak information, just the relevant data. Your brain does not have unlimited storage capacity and more data you have, the more time you will need to analyse it.

That's why: If you have a tool that gives you 1 GB of statistics / day, and you need to retrieve it to find the very small useful information with a script or an Excel spreadsheet, this tool is (for me) too weak, how much time and man day, do you need to analyse it? Maybe you cannot recognize the "heavy" information due to the large amount.

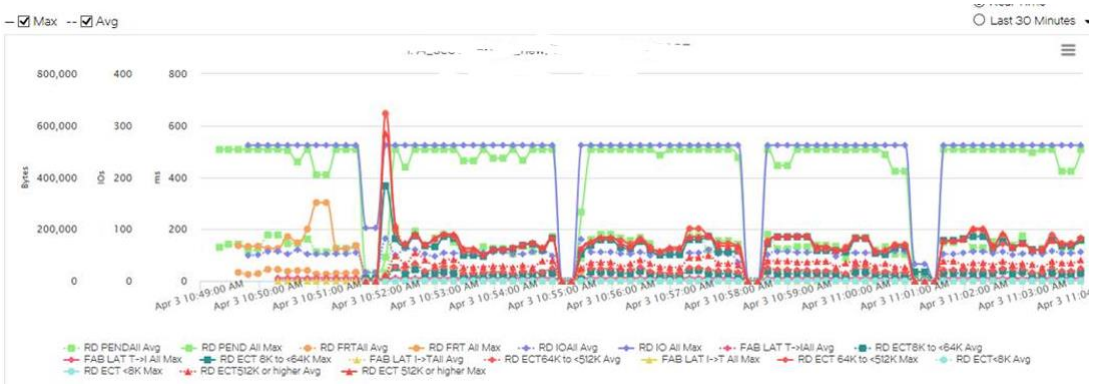
Also, the workflow of your tool is important. A tool that has to set up 100 steps before starting to register some data is equivalent to asking for 100 hat questions before starting the analysis of the answer, ie. :

1. create a filter (to locate the elements to be monitored)
2. Create a collection of items that you want to monitor based on your filter. Can you (Capacity Limit ☹) record all elements in a collection? If not, go back to starting position (it's like in the monopoly game ☹ - "go back to starting position") and create so much filter and collection that you need it
3. Start recording (oh My Gott, the capacity of the recording hardware is exceeded ☹, the event is not recurrent ☹, or I cannot record any longer because other people need the analyser ☹, etc ...)
4. Let's go to flow monitoring menu
5. Select the flow of your collection that you want to monitor
6. Select the parameters. You can either show a parameter for all elements of the collection ☹, or you can create as much window multi parameter as the element ☹, please go back to position 5. ☹
7. "A little later" ☹ ☹ you can start the analysis. "What was the problem? I forgot it ☹" (The 3 guys with the hat are already dead ☹.)

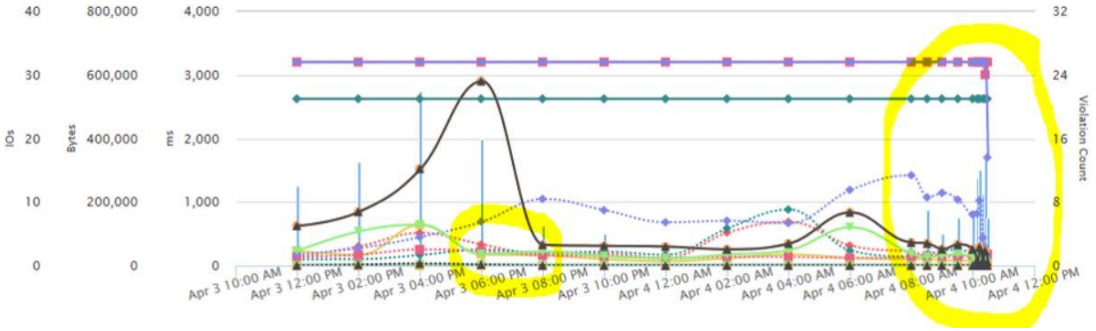
, or the need to interact with 20 different tools to get the true result (that's the number of questions) also does not make sense.

- 4th example

Your supplier only provides you with a granularity of information. This means that details can only be retrieved for a limited amount of time (ie 24 hours, hurry, the information is deleted, or granularity is reduced), and no information about the global infrastructure. That means you may be able to solve your problem but need to search more (ask more questions) and much more time (days, weeks) and more involved people (supplier support, support etc), or that you may not be able to solve it your problem at all (not enough information in the answer, no more interest, it only happened once).



Not-usable automatic-scaling

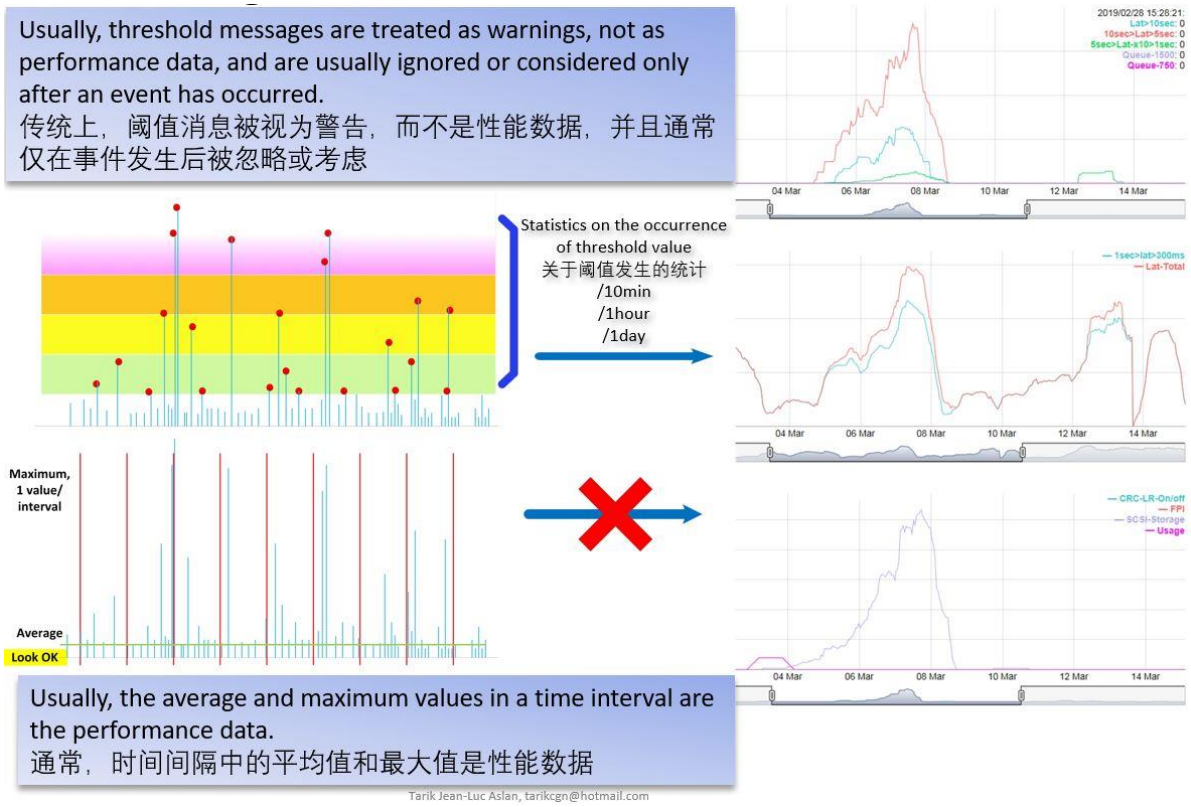


So, let's begin:

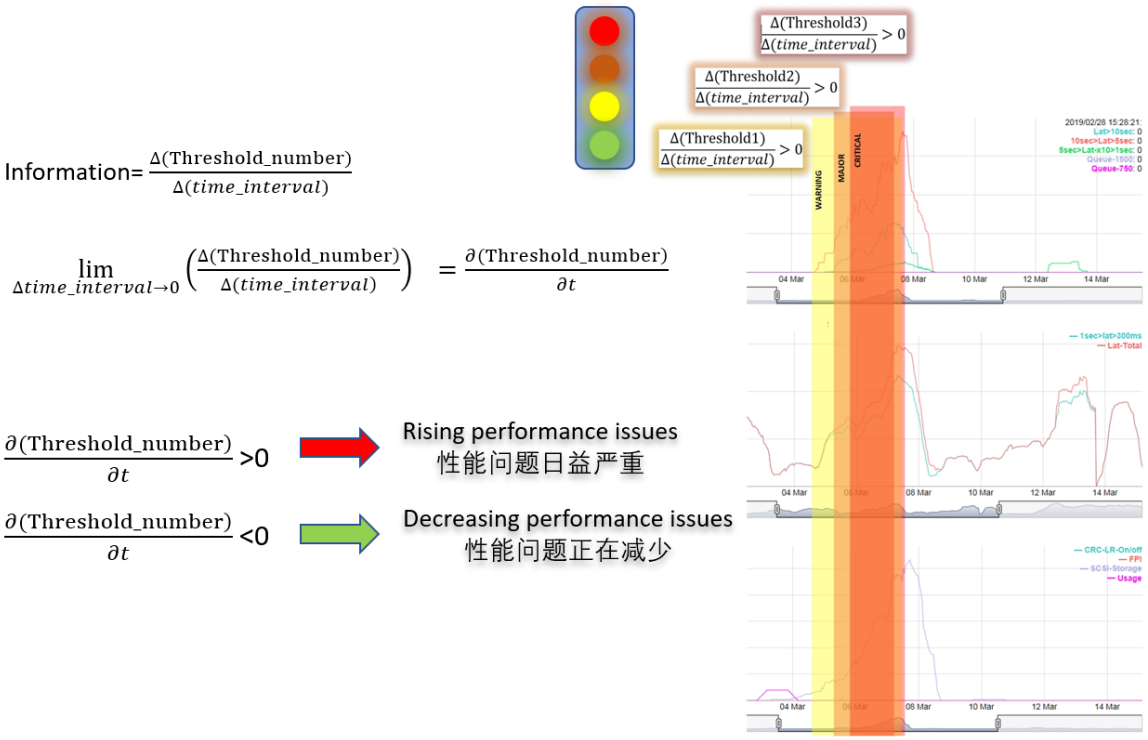
Concept:

In IT, most actions are discrete actions and not continued actions, ie read / write, etc. Even if you get a nice curve from your tool, this means that they fit a curve to represent information about discrete events

The idea is : if I want to display the performance of my system, I will not care if my IOs response time(IO-level/vr/max) is 2,5ms or 2,7ms for every IOs but more to create the statistics of all values that exceed a certain threshold (the number and value of the thresholds can be freely defined or can also adapt to the situation). So , we see it on the picture, on the bottom left you can see the classic supplier method (time window 30s/5min/10min/..), top left is the basis for the new method : define x threshold limit value and so you defined x level or area, from this point you can begin to record all events that correspond to the level, with timelines and detailed information on the type of events, who, from…to…, occurred y-time within last second.



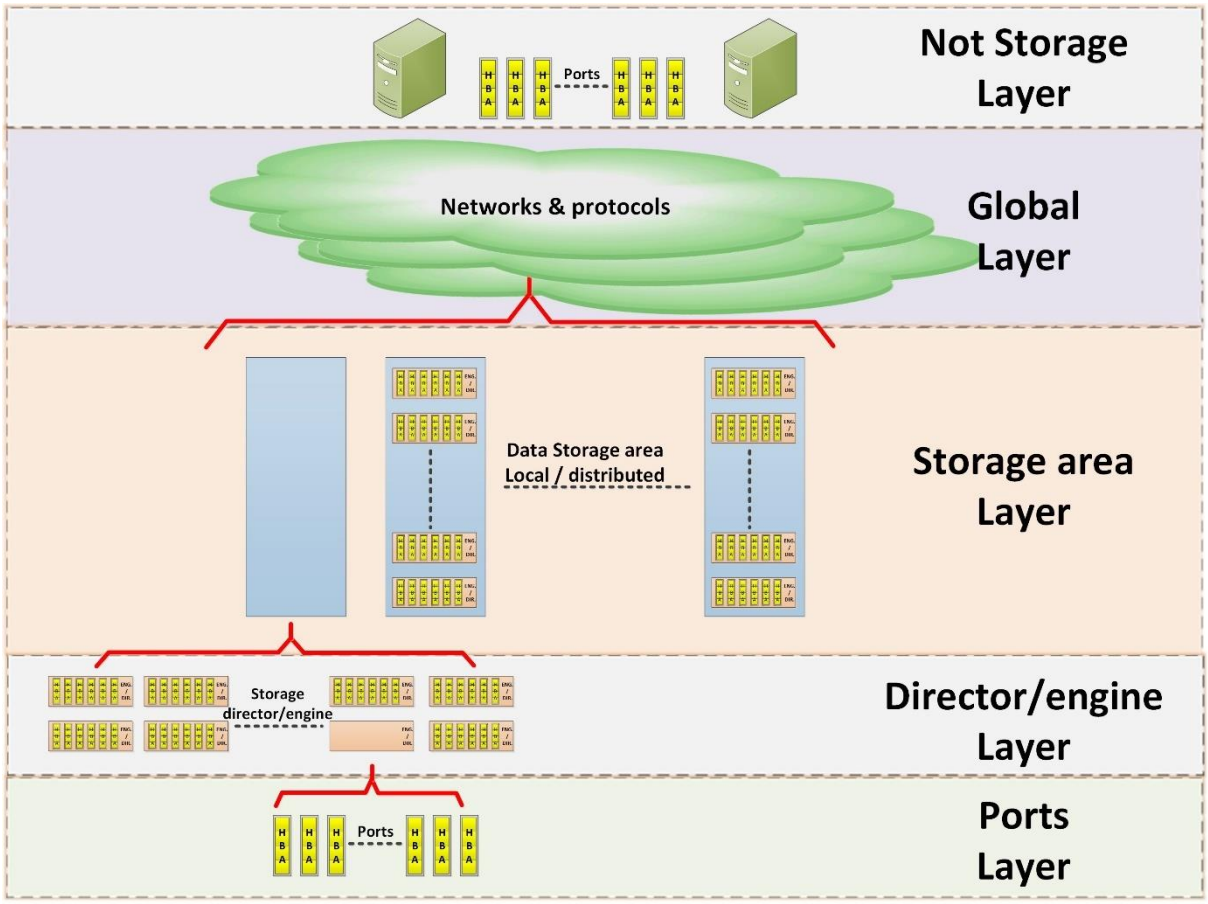
Data analysis of threshold information - 阈值信息的数据分析



A system with poor performance has a high number (occurrences) in the upper threshold. It is also possible to track the evolution of your problem over time. This idea applies not only to the latency, but also to a multitude of other parameters. This information is "heavy" information. Blindly compare the amount of information your supplier tool record anywhere for avr / max. After 24 hours, the data must be resampled to reduce the amount. The dataset of the new method is low. You can continue to use them for weeks or months without having to down sample (sometimes you need to check when the occurrence of an event has started in the past).

The layers

A storage area network can be very complicated, not just a host / storage / switch, but also a host connected to storage or storage connected to network and then to host, but also storage, storage, and more local building or a distant building with a distance of a few 100m to several 1000km and also include different protocols with different characteristics. You can have classic storage or cloud-based storage, even distributed storage, from one supplier or from multiple suppliers (also means different tools). Well, in the end everyone comes up with the same problem or the same question, what is the actual performance, how can a problem be highlighted, how can the future problem be prevented, how can the solution of an incident be improved without x tools to use and log in xx elements and try to find a correlation between. The next image shows a different abstracted layer in your Storage Area Network. At each layer you will find a different answer to the different questions. For some questions, you must choose the right layer.



Global Layer can:

- describe quickly whether your new design / architecture has resulted in improved utilization / performance, etc.
- describe quickly the impact of some events on your infrastructure. Eg 11.11. (Single day 双十一), 20.12. - 01.06. (Christmas - New Year), 28.01. - 14.01. (Spring Festival 春节), Football World Cup, etc..
- describe quickly the performance and problem of the global site-to-site.
- describe Global, regardless of your supplier, your current / past situation on and part of the future (performance planning)
- describe the impact of introducing new technologies on your environment? (ie increasing AI, etc.)
- Show very quickly if there is an increased performance incident or not

Storage area layer can:

- Quickly describe the potential improvement or degradation in performance after a new firmware / code upgrade
- Identify quickly which storage unit could be overloaded and which not, also in connection with the special events.
- Quick recognition of transport or hardware defects (part errors) without having to go to warehouse, parts errors are often not recognized in supplier tools

Director/Engine Layer can:

- Describe a hardware director defect quickly. All director ports have errors
- Describe the effects of overusing the core on the director's performance.
- Quickly describe an indirect performance effect of one port (customer) to another (High IO)
- Describe the global one quickly SCSI Timeout on a Director (local for one port or for all ports)

Port Layer:

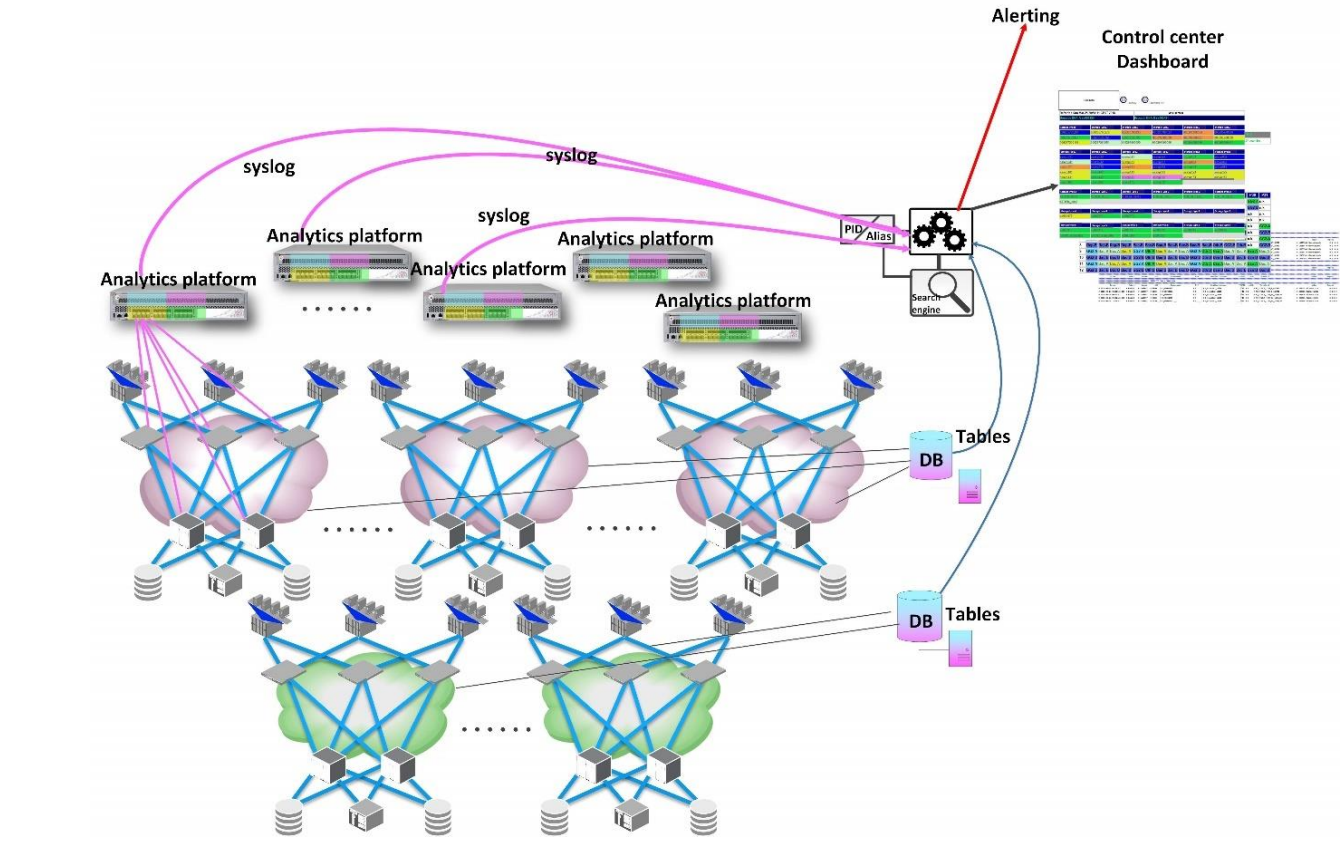
- Describe the use of the queues quickly
- Describe other parameters quickly with time / source / destination of the traffic
- Describe if a problem occurs, occurred or could occur
- Describe the I / O, bandwidth / queue / Latency / SCSI-TO and many other parameters
- Describe quickly which user can influence or influence the performance of all users on the port

Not-storage Layer:

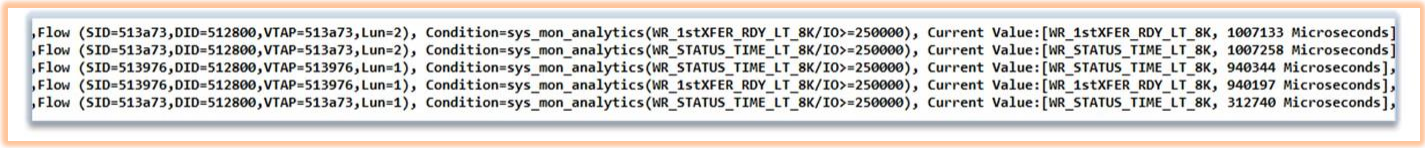
- Describe the customer's behaviour quickly
- Describe the customer's traffic performance quickly
- Describe if a problem occurs, what happened or might happen
- Describe the I / O, bandwidth / queue / latency / SCSI-TO utilization and many more parameters

The realisation / design / solution:

The only setup action is the threshold, provide basic information about the data storage unit (type, name, location), and provide minimal information about the switches (type, name, location); the rest is done automatically, creating the full structure for the data storage unit. The analytics platform delivers the syslog messages. The Analytics platform can be an AMP (Brocade) or a Switch Brocade/Broadcom (IO Insight Gen7). I think Cisco and VI should also be able to provide such information, it has to be checked.



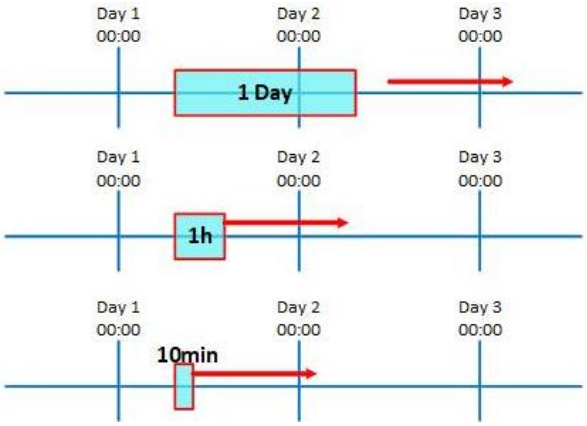
The translator core translates PID (0xAABBCC) into human-readable information, PID need not be unique.



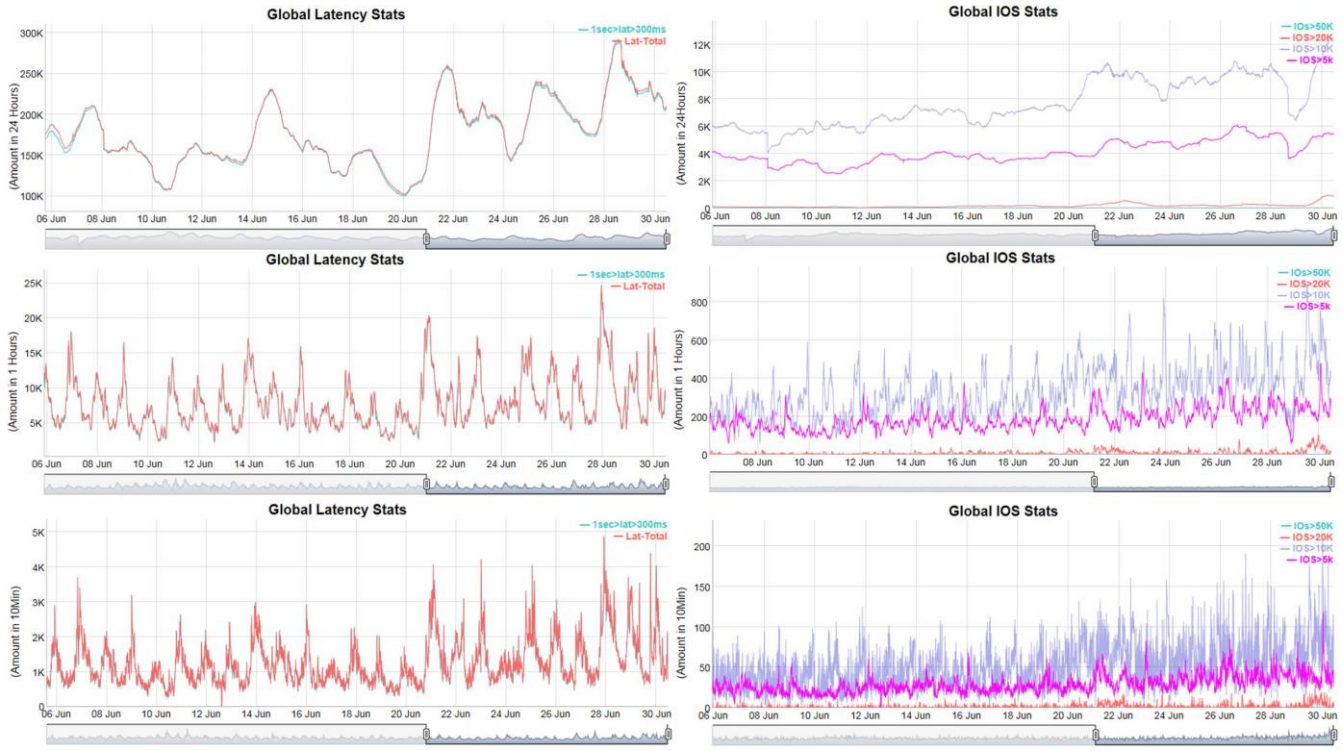
The goal is not only to monitor performance, but also to analyse and correlate every "non-normal" event directly in the dashboard. Here are the base events listed, all other "non-normal" events are not listed in the following table

SCSI - Block <8k, 8-64K, 64-512K, >512K	FC transport	SCSI-Protocol	Fabric performance
Read 1rst response time	Link-Reset	SCSI-Abords	Slowdrain-Quarantine
Read completion time	Off-On	SCSI-TimeOut	FPI
Write 1rst response time	C3TxTo		Fabric latency
Write completion time	CRC		
Other completion time	Utilisation > 80%		
Rd-IO Queuing			
Wr-IO Queuing			
Others Queuing			
IO/s			

For the use of the AMP is the interesting part of the idea is more the use of the platform: The limit of the platform is 20000 IT / IT lun data flow, but a large server with 8 HBA ports, which is connected to 16 storage HBA Ports and 1000Luns-Mapped already generate 128000 ITL flows. In this situation the AMP is only used by performance incident. The idea is to forget the IT & ITL flow limit and to consider only the IO/s limit (5 million IO/S). Then you can monitor many storage ports with 1 AMP. Remember that I / O operations are discrete events. All storage ports will never read/write the maximum IO/s at the same time. This means, the true IO/S flow is lower.
Also, the only information I use is the syslog messages (MAPS), which say that the xx Threshold has been exceeded ("Heavy" information). I do not save any other data (weak information) from the AMP. I have also set the AMPs data retention to minimum, it give more flexibility ie, if a server no longer has I / O, its data flow is directly deleted and capacity is freed up. The statistics result uses 3 different time windows, 10 minutes / 1 hour / 24 hours, 10 minutes window to see the dynamics of your network 1 hour window allows you to filter the dynamics of the system and provide useful information for the trend analysis 24 hours window is useful for very long term analysis from up to 1 or 2 years.

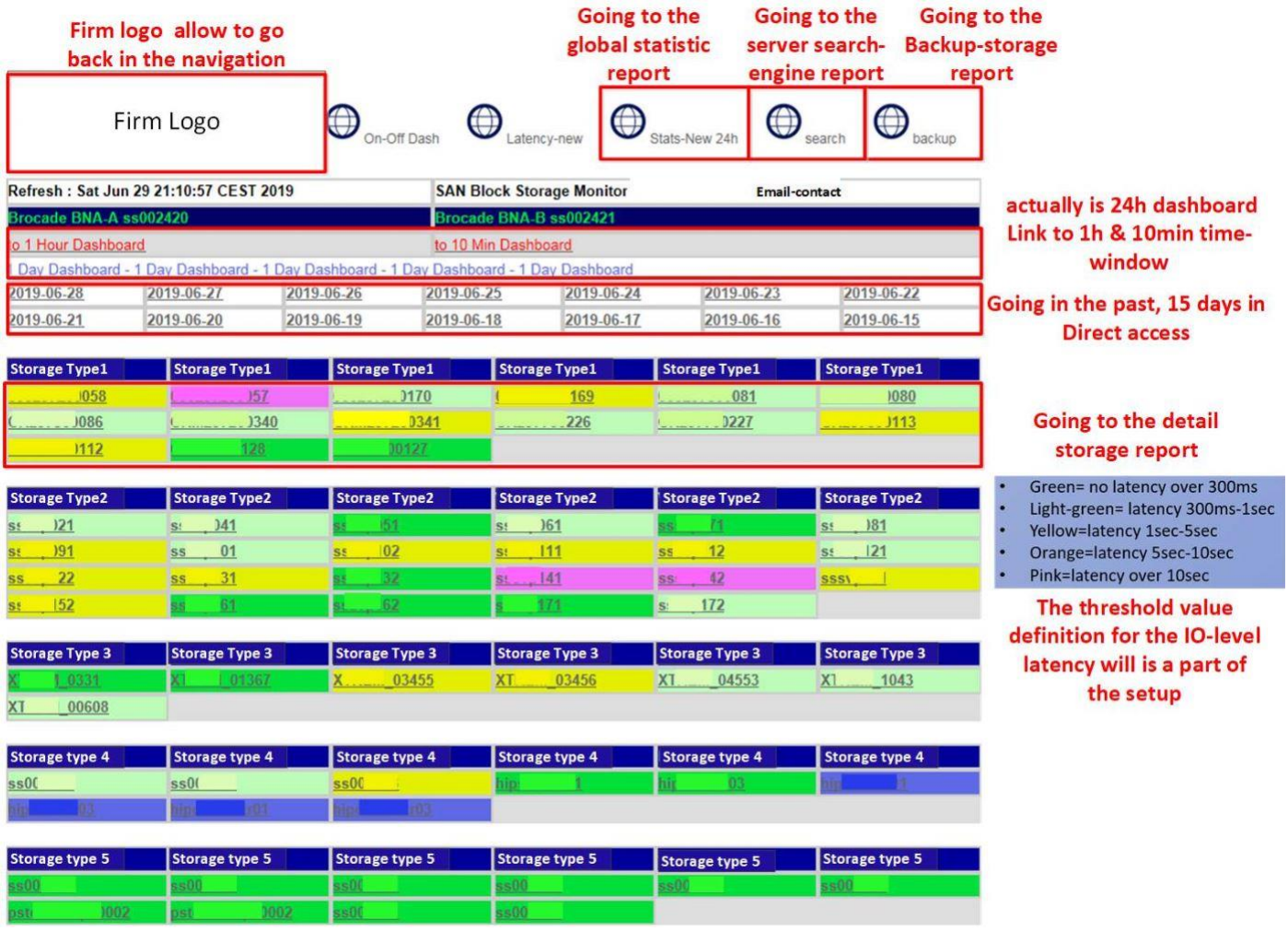


The following picture shows the difference between these 3 time windows:



This version of the dashboard is limited to the FC network. An extension to the NVMe and the IP network should not be difficult.

The global Layer



The data storage area layer:

Depending on the storage type, the design of the physical port is different. If a director is dead, you will see it directly. The Director line is red.

Going to the Performance graphic 1D/1H/10Min time-window

NAME	On	Off	n.k.	SerialNR	Location	FINA-A	FINA-B
_)057	151	0	41	)057	.F07	2420-works	2421-works

Graphic Day-Long	Graphic Day-window	Graphic Hour-window	Graphic 10Min-window
o 1 Hour Dashboard	to 10 Min Dashboard		

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

Switch from 1Day to 10min or 1h time-window

2019-06-28	2019-06-27	2019-06-26	2019-06-25	2019-06-24	2019-06-23	2019-06-22
2019-06-20	2019-06-19	2019-06-18	2019-06-17	2019-06-16	2019-06-15	2019-06-14

Going in the past, 15 days in Direct access

	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	12	0	0	0	36	0	0	0
Lat > 5000ms	0	0	0	0	0	0	0	0	0	5
Lat > 1000ms	0	0	61	20	25	28	60	58	20	38
Lat > 300ms	339	1912	49085	52604	57803	42255	54215	56860	49208	41608
Latency ALL	339	1912	49158	52624	57828	42283	54311	56918	49228	41651
IO-Queue >1500	0	0	0	0	0	0	0	0	0	0
IO-Queue > 750	0	0	12	0	13	9	0	24	0	11
IO-Queue > 300	0	0	0	0	0	1	0	16	0	0
Usage > 90%	0	0	0	0	0	0	2	0	0	0
CRC-LR-OnOff	0	0	0	0	0	0	0	0	0	0
FPI-Slowdrain	0	0	0	0	0	0	0	0	0	0
SCSI-TO/Abs	0	0	46	209	0	0	100	0	0	0
IOs > 50000	0	0	0	0	0	0	0	0	0	0
IOs > 20000	0	0	0	0	0	0	0	0	0	2
IOs > 10000	0	4	225	150	333	413	514	310	445	313
IOs > 6000	55	222	6763	5461	6836	9676	5906	6930	8016	7151
Fab-Lat	0	0	13	3	3	0	16	0	18	0
SCSI-Res	0	0	0	0	0	0	0	0	0	0

Global statistic for this storage unit

Going to the the parameter report, incl source-destination info

DIR-M	DIR-H	P04	P05	P06	P07	P08	P09	P10	P11	P24	P25	P26	P27	P28	P29	P30	P31
01-mi	01-ho	Rep-A	Ban-A	Ban-A	GG2-A	Bac-A	Bac-A	n.k.	n.k.	Bac-A	Bac-A	Ban-A	n.k.	n.k.	n.k.	GG2-A	n.k.
02-mi	02-ho	Rep-B	Ban-B	Ban-B	GG2-B	Bac-B	Bac-B	n.k.	n.k.	Bac-B	Bac-B	Ban-B	n.k.	n.k.	n.k.	GG2-B	n.k.
03-mi	03-ho	n.k.	Ban-A	Ban-A	GG2-A	Bac-A	Bac-A	n.k.	n.k.	Bac-A	Bac-A	Ban-A	n.k.	GG2-A	n.k.	n.k.	n.k.
04-mi	04-ho	n.k.	Ban-B	Ban-B	GG2-B	Bac-B	Bac-B	n.k.	n.k.	Bac-B	Bac-B	Ban-B	Oth-B	GG2-B	n.k.	n.k.	n.k.
05-mi	05-ho	n.k.	Ban-A	Ban-A	n.k.	Bac-A	Bac-A	n.k.	n.k.	Bac-A	Bac-A	Ban-A	Oth-A	GG2-A	Oth-A	n.k.	GG2-A
06-mi	06-ho	n.k.	Ban-B	Ban-B	n.k.	Bac-B	Bac-B	n.k.	n.k.	Bac-B	Bac-B	Ban-B	Oth-B	GG2-B	Oth-B	n.k.	GG2-B
07-mi	07-ho	Rep-A	Rep-A	Rep-A	Rep-A	Bac-A	Bac-A	Ban-A	Ban-A	Bac-A	Bac-A	Ban-A	Oth-A	GG2-A	Oth-A	n.k.	GG2-A
08-mi	08-ho	Rep-B	Rep-B	Rep-B	Rep-B	Bac-B	Bac-B	Ban-B	Ban-B	Bac-B	Bac-B	Ban-B	Oth-B	GG2-B	Oth-B	n.k.	GG2-B
09-mi	09-ho	GG2-A	Bac-A	Bac-A	Bac-A	GG2-A	Oth-A	Ban-A	Bac-A	Bac-A	GG2-A	n.k.	Ban-A	Bac-A	Bac-A	Ban-A	Bac-A
10-mi	10-ho	GG2-B	Bac-B	Bac-B	Bac-B	GG2-B	Oth-B	Ban-B	Bac-B	Bac-B	GG2-B	n.k.	Ban-B	Bac-B	Bac-B	Ban-B	Bac-B
11-mi	11-ho	GG2-A	Bac-A	Bac-A	Bac-A	GG2-A	Oth-A	Ban-A	Bac-A	Bac-A	GG2-A	n.k.	Ban-A	Bac-A	Bac-A	Ban-A	Bac-A
12-mi	12-ho	GG2-B	Bac-B	Bac-B	Bac-B	GG2-B	Oth-B	Ban-B	Bac-B	Bac-B	GG2-B	n.k.	Ban-B	Bac-B	Bac-B	Ban-B	Bac-B

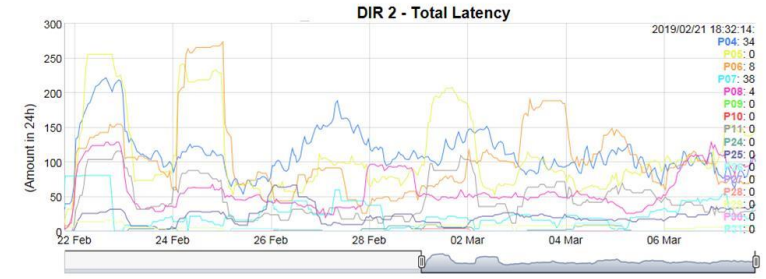
Pink>10000ms, Orange>5000ms, Yellow>1000ms, Light-Green>300)ms, Red = Offline, Dark-Blue= SWITCH has NO AMP connectivity, Light-Blue= NO-AMP-port

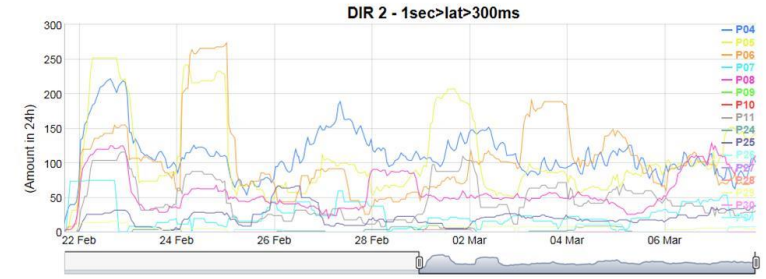
Going to the Director/ Engine graphic 1h/10min time-window

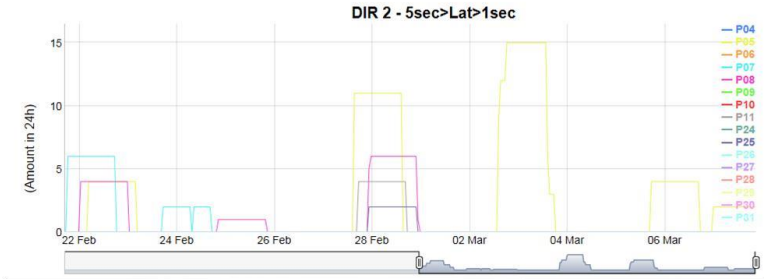
Going to the storage-port detail report

Director graphic example:

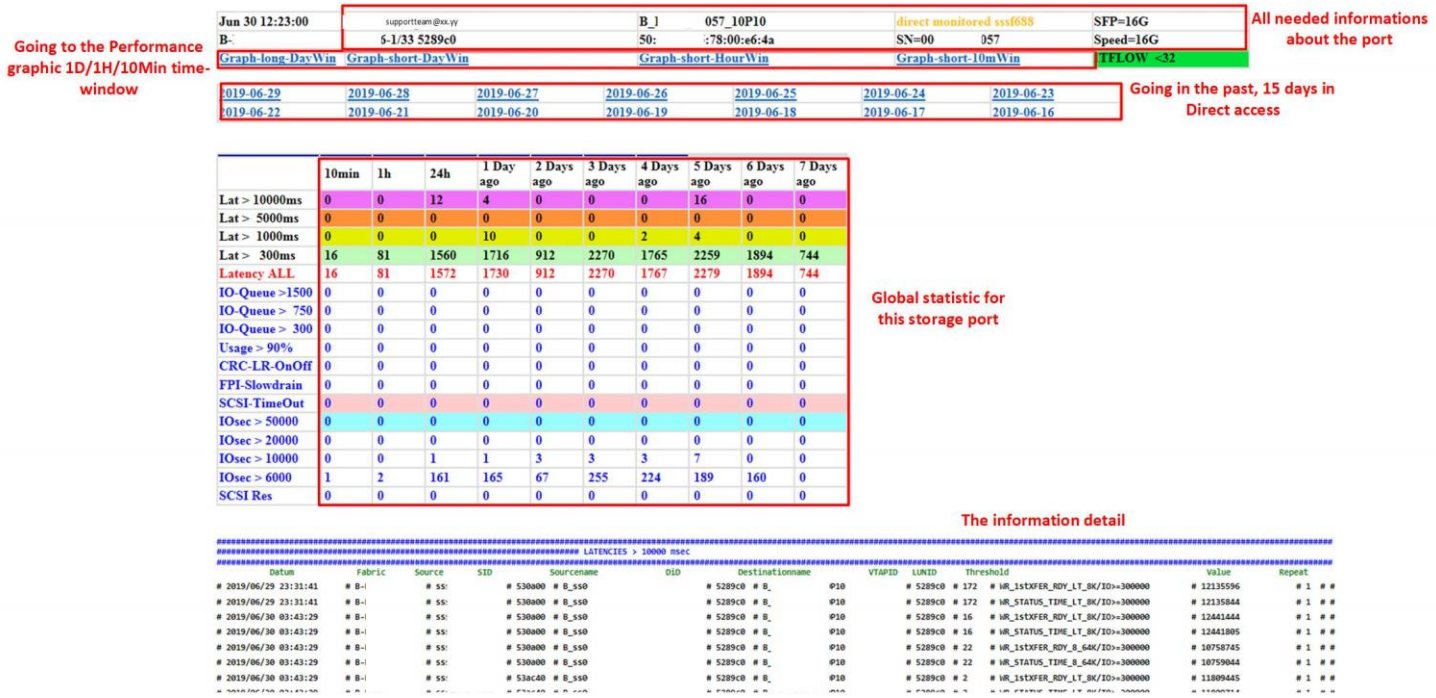
- If latency is increasing the same for all ports, it means that all cores are IO-over-used. You can also find directly the ports with the highest IO load.
- If the SCSI-TO is going up for all ports of the director, this means the director is going defect.







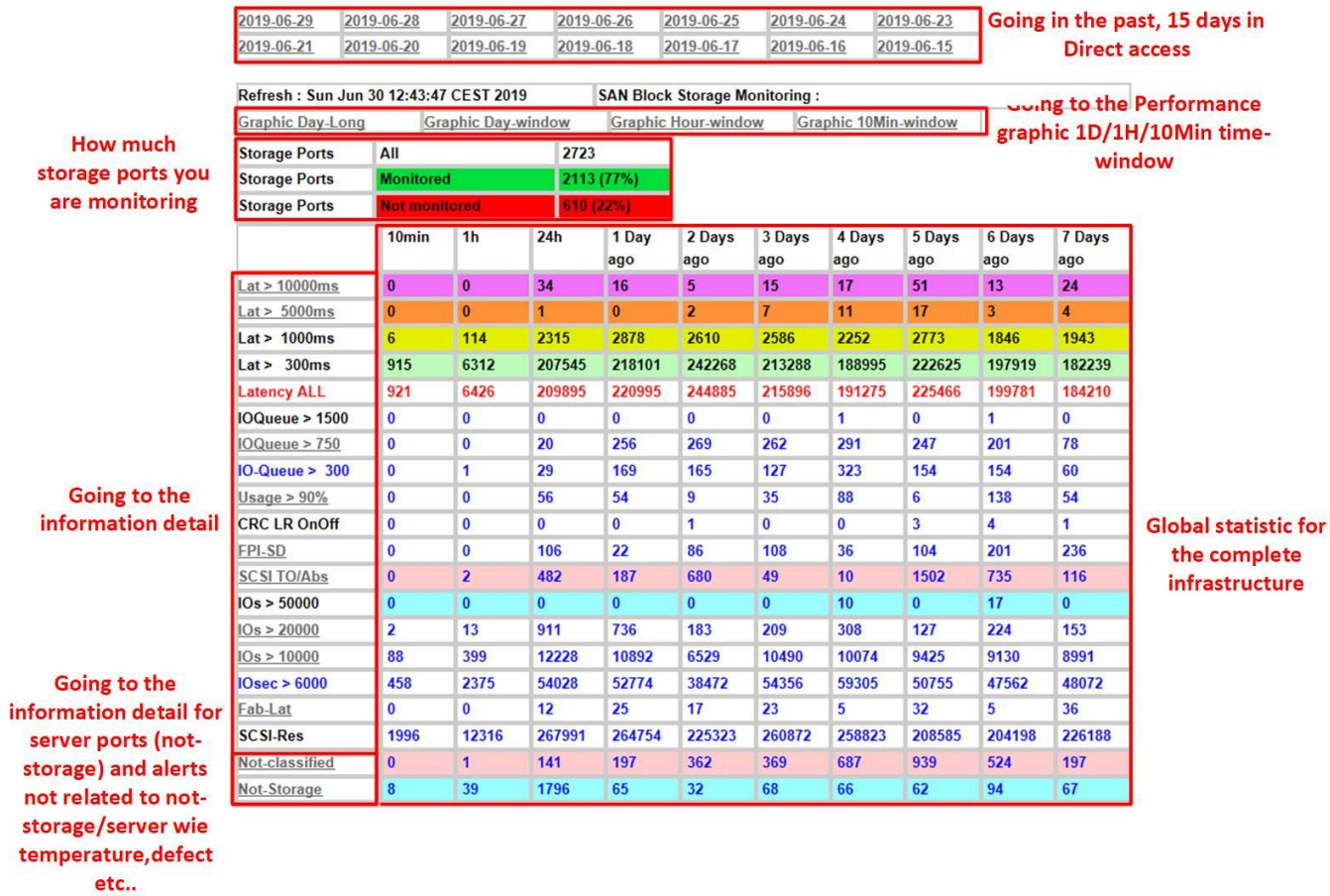
Port Layer:



	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	12	4	0	0	0	16	0	0
Lat > 5000ms	0	0	0	0	0	0	0	0	0	0
Lat > 1000ms	0	0	0	10	0	0	2	4	0	0
Lat > 300ms	16	81	1560	1716	912	2270	1765	2259	1894	744
Latency ALL	16	81	1572	1730	912	2270	1767	2279	1894	744
IO-Queue >1500	0	0	0	0	0	0	0	0	0	0
IO-Queue > 750	0	0	0	0	0	0	0	0	0	0
IO-Queue > 300	0	0	0	0	0	0	0	0	0	0
Usage > 90%	0	0	0	0	0	0	0	0	0	0
CRC-LR-OnOff	0	0	0	0	0	0	0	0	0	0
FPI-Slowdrain	0	0	0	0	0	0	0	0	0	0
SCSI-TimeOut	0	0	0	0	0	0	0	0	0	0
IOsec > 50000	0	0	0	0	0	0	0	0	0	0
IOsec > 20000	0	0	0	0	0	0	0	0	0	0
IOsec > 10000	0	0	1	1	3	3	3	7	0	0
IOsec > 6000	1	2	161	165	67	255	224	189	160	0
SCSI Res	0	0	0	0	0	0	0	0	0	0

A high initial response time WR_1stXFER_RDY mainly indicates a problem with the storage port (too many I/O? Too many hosts?), While the completion time is more likely to indicate a cache or backend problem, but a high initial answer is always on high problem brings completion time (completion time = 1 response time + data transfer time)

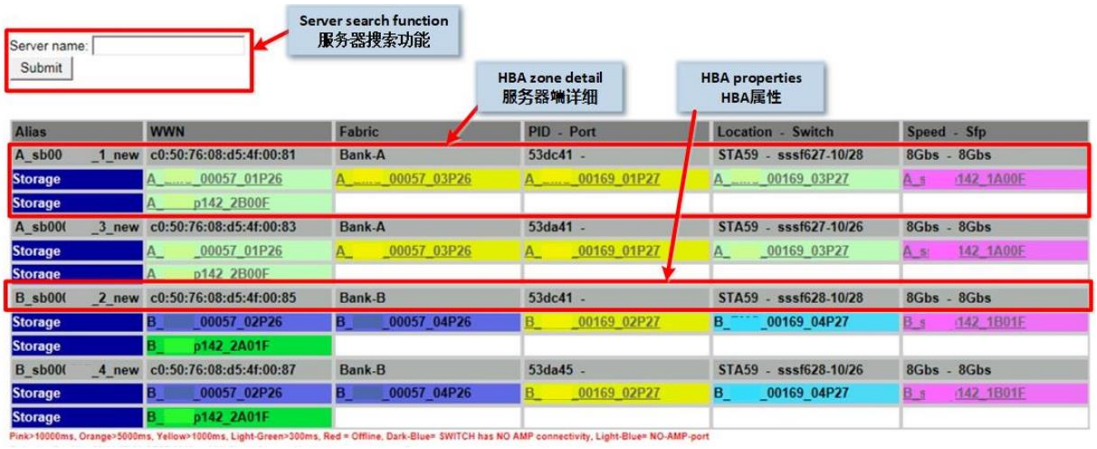
Global Layer Stats report



	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	34	16	5	15	17	51	13	24
Lat > 5000ms	0	0	1	0	2	7	11	17	3	4
Lat > 1000ms	6	114	2315	2878	2610	2586	2252	2773	1846	1943
Lat > 300ms	915	6312	207545	218101	242268	213288	188995	222625	197919	182239
Latency ALL	921	6426	209895	220995	244885	215896	191275	225466	199781	184210
IOQueue > 1500	0	0	0	0	0	0	1	0	1	0
IOQueue > 750	0	0	20	256	269	262	291	247	201	78
IO-Queue > 300	0	1	29	169	165	127	323	154	154	60
Usage > 90%	0	0	56	54	9	35	88	6	138	54
CRC LR OnOff	0	0	0	0	1	0	0	3	4	1
FPI-SD	0	0	106	22	86	108	36	104	201	236
SCSI TO/Abs	0	2	482	187	680	49	10	1502	735	116
IOs > 50000	0	0	0	0	0	0	10	0	17	0
IOs > 20000	2	13	911	736	183	209	308	127	224	153
IOs > 10000	88	399	12228	10892	6529	10490	10074	9425	9130	8991
IOsec > 6000	458	2375	54028	52774	38472	54356	59305	50755	47562	48072
Fab-Lat	0	0	12	25	17	23	5	32	5	36
SCSI-Res	1996	12316	267991	264754	225323	260872	258823	208585	204198	226188
Not-classified	0	1	141	197	362	369	687	939	524	197
Not-Storage	8	39	1796	65	32	68	66	62	94	67

Example of performance graphic:

The server search page



Some user case:

1. performance improvement or deterioration after storage code upgrade

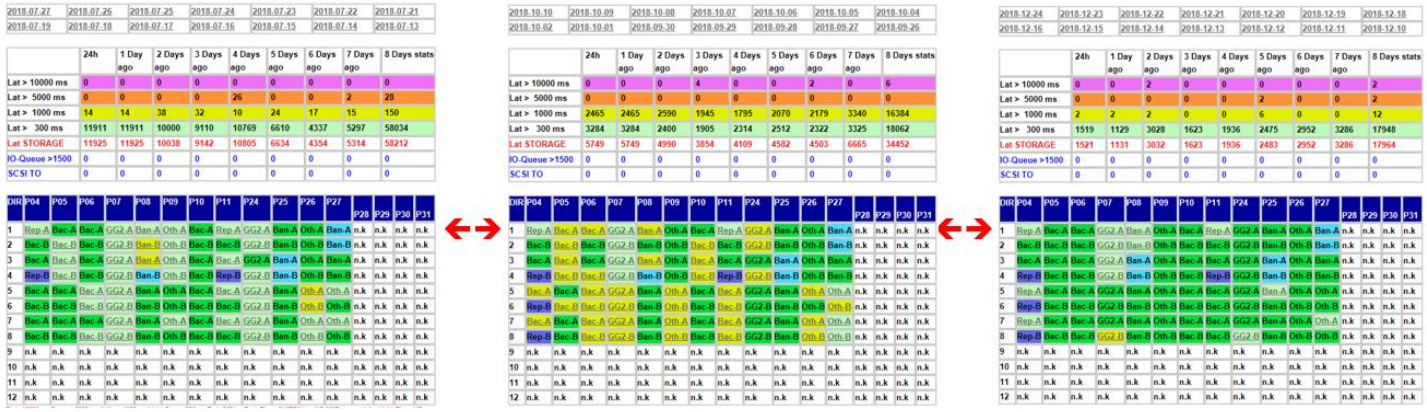
Before firmware upgrade
在固件升级之前

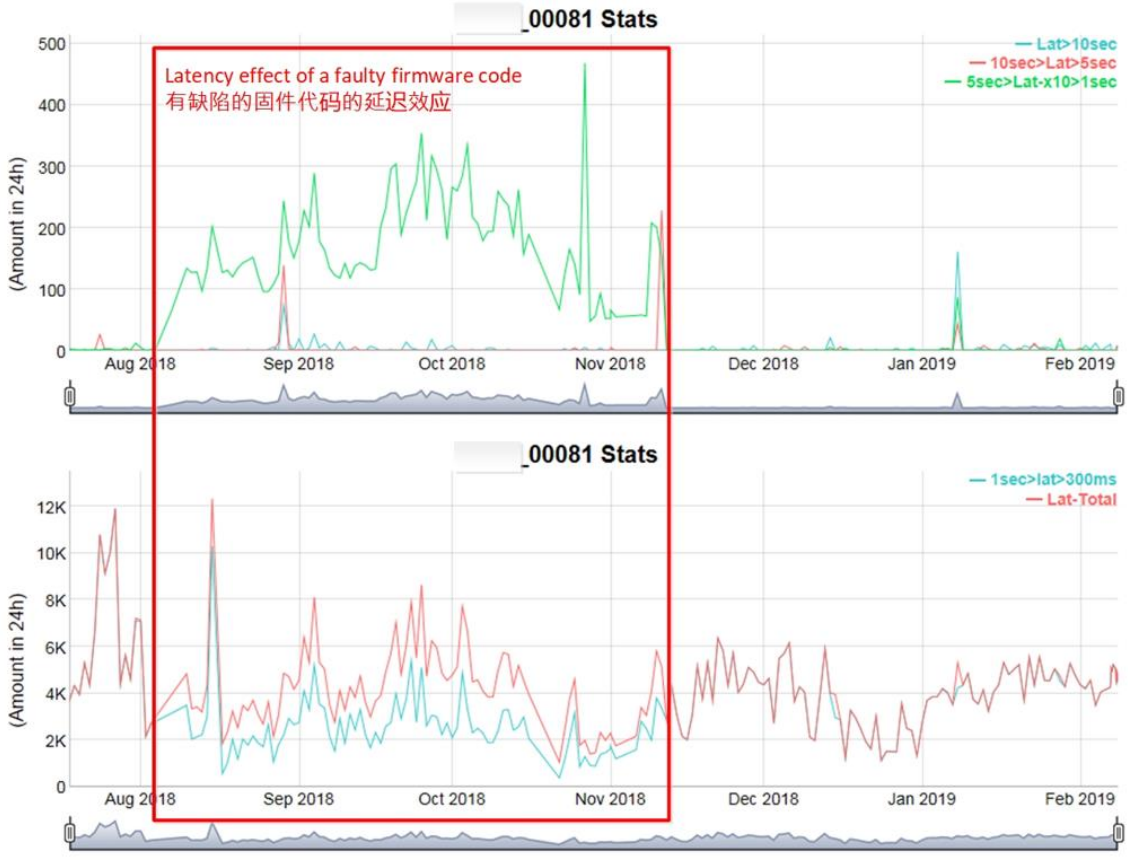
↔

Performance deterioration
after firmware upgrade
固件升级后性能下降

↔

Performance improvement
after patch upgrade
补丁升级后的性能提升





2. fabric switch problem. How to find it, easy look the storage unit detail, you can directly see that all the errors are related to one fabric or 2 fabrics

	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	142	564	1536	0	0	0	0	0	0	0
Lat > 5000ms	321	1367	4247	0	0	0	0	0	0	0
Lat > 1000ms	899	4212	12164	0	0	0	0	0	0	0
Lat > 300ms	310	1446	4611	58	533	428	268	261	0	0
Latency ALL	1672	7589	22558	58	533	428	268	261	0	0
IO-Queue >1500	0	0	0	0	0	0	0	0	0	0
IO-Queue > 750	0	0	0	0	0	0	0	0	0	0
IO-Queue > 300	0	0	0	0	0	0	0	0	0	0
Usage > 90%	0	0	0	0	0	0	0	0	0	0
CRC-LR-OnOff	0	0	0	0	0	0	0	0	0	0
FPI-Slowdrain	0	0	0	0	0	0	0	0	0	0
SCSI-TO/Abs	72	406	1136	0	1	0	0	0	0	0
IOs > 50000	0	0	0	0	0	0	0	0	0	0
IOs > 20000	0	0	0	0	0	0	0	0	0	0
IOs > 10000	0	0	11	0	8	12	7	12	0	0
IOs > 6000	0	0	12	0	29	25	32	48	0	0
Fab-Lat	0	0	0	0	0	0	0	0	0	0
SCSI-Res	0	0	0	0	0	0	0	0	0	0

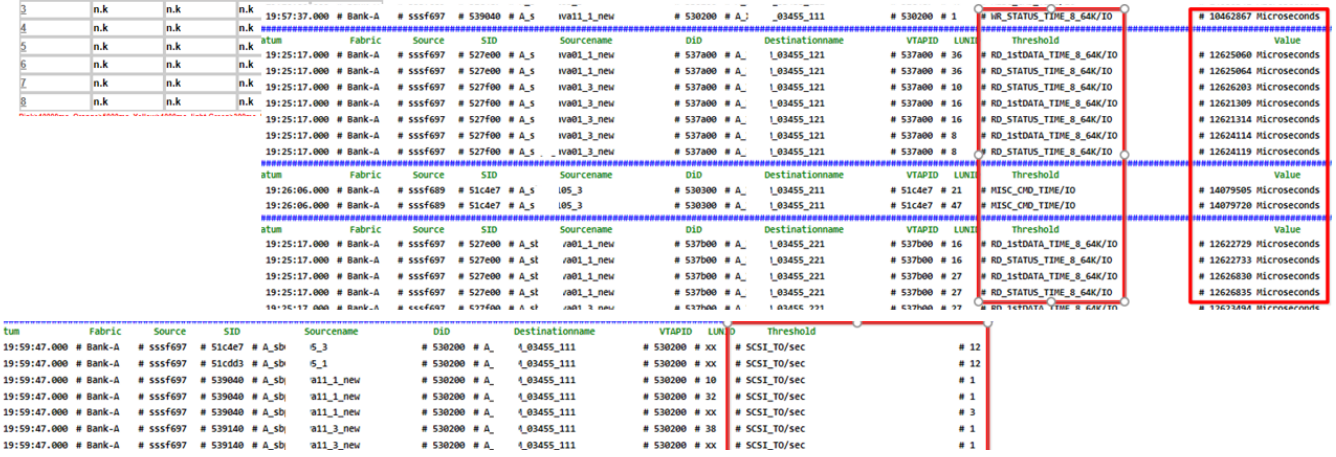
DIR-D	DIR-H	P04	P05	P06	P07	P08	P09	P10	P11	P24	P25	P26	P27	P28	P29
01-D	01-H	Rep-A	Bac-A	Bac-A	GG2-A	Ban-A	Oth-A	Bac-A	Rep-A	GG2-A	Ban-A	Oth-A	Ban-A	n.k.	n.k.
02-D	02-H	Bac-B	Bac-B	Bac-B	GG2-B	Ban-B	Oth-B	Bac-B	Bac-B	GG2-B	Ban-B	Oth-B	Ban-B	n.k.	n.k.
03-D	03-H	Bac-A	Bac-A	Bac-A	GG2-A	Ban-A	Oth-A	Bac-A	Bac-A	GG2-A	Ban-A	Oth-A	Ban-A	n.k.	n.k.
04-D	04-H	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.	n.k.
05-D	05-H	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.	n.k.
06-D	06-H	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.	n.k.
07-D	07-H	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.	n.k.
08-D	08-H	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.	n.k.

3. code upgrade impact (latency up to 19sec)

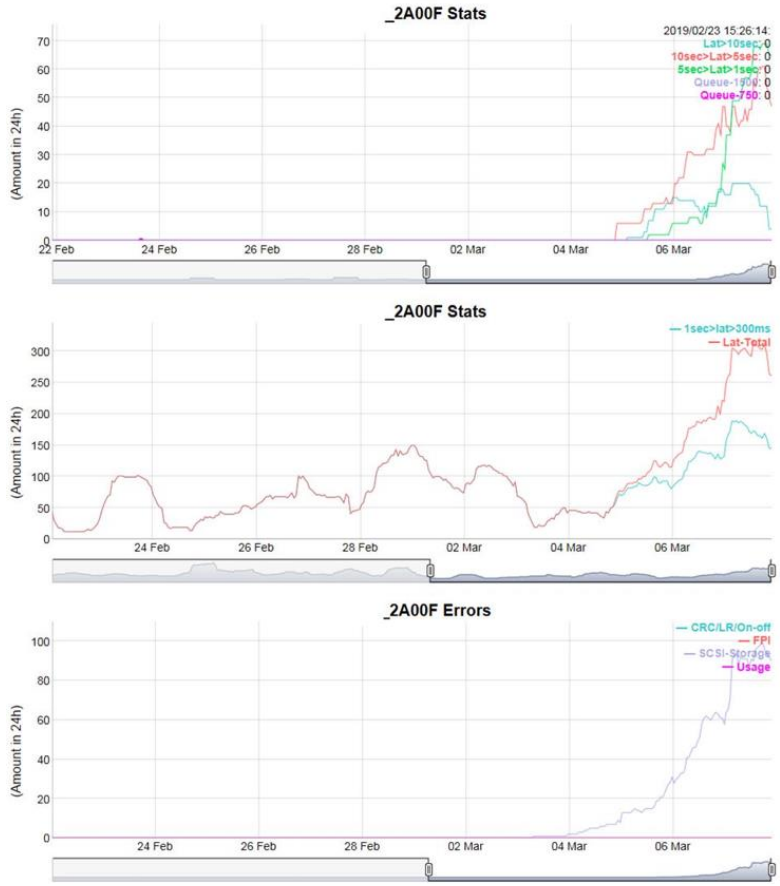
	24h	1 Day ago	2 Days ago	3 Days ago	4 Days ago	5 Days ago	6 Day ago
Lat > 10000 ms	26	0	0	0	0	0	0
Lat > 5000 ms	4	0	0	0	0	0	0
Lat > 1000 ms	105	0	0	0	0	0	0
Lat > 300 ms	4	0	0	0	0	0	0
Lat STORAGE	139	0	0	0	0	0	0
IO-Queue >1500	0	0	0	0	0	0	0
SCSI TO	29	0	0	0	0	0	0
Usage	0	0	0	0	0	0	0

DIR	11	12	21	22
1	Ban-A	Ban-B	Ban-A	Ban-B
2	Ban-A	Ban-B	Ban-A	Ban-B
3	n.k	n.k	n.k	
4	n.k	n.k	n.k	
5	n.k	n.k	n.k	
6	n.k	n.k	n.k	
7	n.k	n.k	n.k	
8	n.k	n.k	n.k	

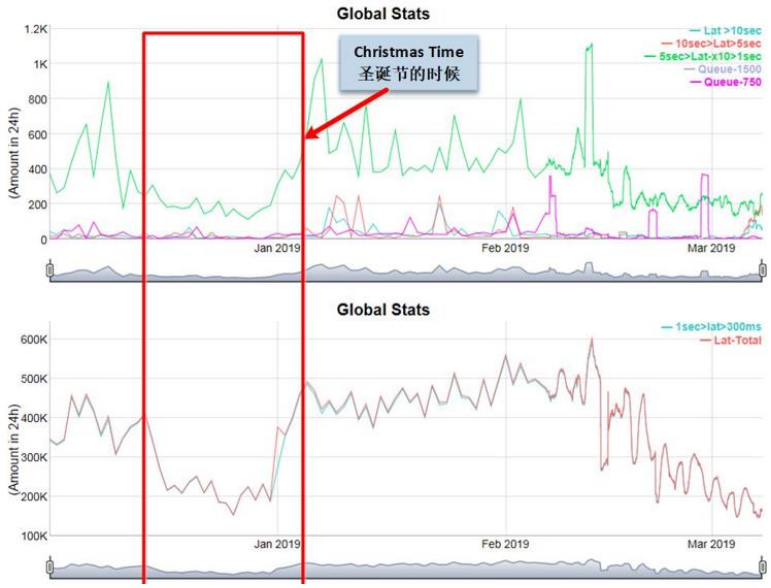
Upgrade is not disruptive but during the upgrade the Latency is going up to 20-30sec, this means applications performance impact.
Consequence : if your customer at the same time run task with load, the task will not be able to finish in the contract time and you will need to pay penalties



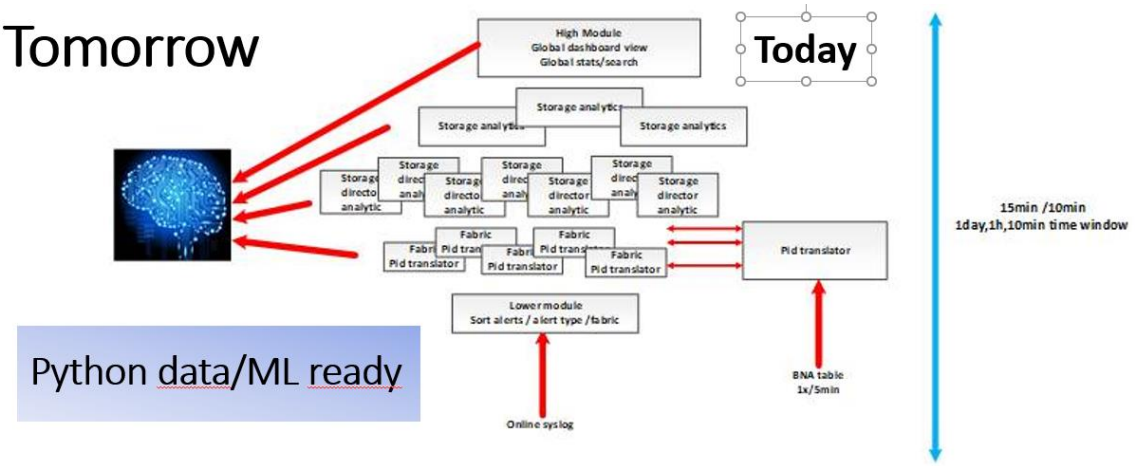
4. Correlation between different parameter, SCSI-timeout & high latency



5. Capacity planning or special events effect



The next future



The last point of this article:

I do not want to talk directly about the ROI, but rather try to outline the actual cost of an incident. Just look at the following table. The costs per hour are assessed. The idea is to show where they come from. Only the value of business revenue from the impacted application is missing.

Player	Team	number of poeple	cost/h	cost/h
	OS/Storage/hardware/network	6	100-200€/poeple/h	1200€/h
	escalation management	2	100-200€/poeple/h	400€/h
Company	management	2 depending of the impact	200-300€/poeple/h	600€/h
Customer	Customer company	1-5 depending of impact	200-300€/poeple/h	1500€/h
Supplier	support team	2-5	100-200€/poeple/h	1000€/h
Service outage	customer-customer		xx.000€/h-xxx.000€/h	xxx.000€/h
Penalities			xx.000€/inc-xxx.000€/Inc	xxx.000€/Inc
		3h		x14.100€
		6h		X28.200€
		12h		x.x56.200€
		24h		x.x02.200€
		48h		...
Incident		72h		
				add penalities

As conclusion

My claim:

Enterprise storage quality does not come from the average response time, but from the number of unacceptable IO response times per period.

- The method preserves the heavily information and also reduce the data volume allowing long-term graphics
- The workflow is easy and optimized, also no complicate setup
- You can directly in a few minutes say everything is ok, or last weekend /night was nothing, you can also be proactive, before the problem become a critical one. The system also sends alerts by critical events in the 10min time window.
- I also think, that it is a good way to test the performance of new storage unit (buy or self-developed)
- With this new method you can see what you normally cannot see. Sometimes you may be surprised that it can exist.
- The method preserves the "heavily" information and reduces the data volume, so that long-term graphics are possible.
- The workflow is simple and optimized, no complicated settings.
- You can say in a few minutes, everything is fine. or last weekend / night was nothing, you can also be proactive before the problem gets critical. The system also sends alerts for critical events in the 10-minute window.

The idea, concept and code were developed in my private time and are not based on a company requirement, the realization time amounts to less than 3 months after work.

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， deutsch, English, francais
Contact:

email: tarikcgn@hotmail.com
linkedin : http://www.linkedin.com/in/tarik-jean-luc-aslan-45a0a93b
wechat: tarikzh