Qualitative and quantitative performance evaluations of relatively inexpensive storage products (2)

Hiroki Kashiwazaki[1,,a)]

Abstract: In 2017, the author introduced a qualitative and quantitative performance evaluation of the relatively inexpensive Network Attached Storages (NAS) that cost approximately 100 USD per terabytes and are connected with Gigabit Ethernet. At that time, it was difficult to configure with 10Gigabit Ethernet with keeping 100 USD per terabytes. After three years since then, cheap network modules have been offered and switch prices have fallen. Not only 10Gigabit Ethernet but also 40Gigabit or 100Gigabit Ethernet can be a candidate of a storage area network (SAN). Due to the falling price of SSDs, an all-flash storage system can be configured easily. In this paper, the author evaluates disk array systems that cost 100 USD per terabytes and all-flash storage systems that cost 500 USD per terabytes with quantitative and qualitative approaches.

1. Introduction

The distributed system consists of three parts: a computer, a network line, and a power supply source. The computer is composed of a central processing unit, a main storage device, an auxiliary storage device, and the like. With the improvement of the I/O performance of the auxiliary storage device, the transfer speed with the main storage device has been increased. Further, as the capacity of the auxiliary storage device has increased, it has become necessary to install the auxiliary storage device outside the computer. Early external storage was connected via interfaces such as SCSI and Fiber Channel. Eventually, computers and storages connected by L2 via Ethernet, etc. used the iSCSI protocol to treat the storages as block devices, and NFS to mount the file systems provided by the storages. When connecting as a block device, a network connecting a computer and a storage is called a storage area network (SAN). Storage products that provide NFS services are called Network Attached Storage (NAS).

Fourteen years have passed since the concept of cloud computing was first proposed. When it was first proposed, it was thought that there would be a shift from ownership to the use of computing resources. But now, users in various organizations are still using a certain number of on-premises Systems because they want to avoid storing sensitive information on external devices and they hate degradation of performance due to delays. With the commoditization of computing resources, there is no longer any difference between manufacturers in computing. If users ask server vendors the number of CPU cores, their operating frequency, memory capacity, and the capacity of secondary storage, they can get the computer they want. There are only a few differences that can be considered as vendor-specific, such as management console features and so on. Meanwhile, external storage products still have a diversity of specifications.

Some storage vendors benchmarked their products and publish them as performance sheets. However, not all vendors publish them, and the benchmarking conditions are not uniform. The vendors develop their products for use in their focused target markets. Therefore they prefer the results of benchmarks that can better perform their products to the customers in the markets. It is not a rare case that only the result that shows superior to other vendors can be published. While such arbitrary publication is effective in terms of vendors and their stock prices, users can neither see the results of multiple vendors nor make quantitative and cross-sectional comparisons.

Therefore, in 2017, the author evaluated a relatively inexpensive NAS qualitatively and quantitatively. A target of the evaluation is NAS products that cost 100 USD per terabytes and consist of Gigabit Ethernet [1], [2] Network Interface Card (NIC). This constraint was set because 10Gigabit Ethernet [3] enabled switches were still expensive at that time. Three years later, the price of 10Gigabit Ethernet-enabled switches has also fallen. Not only 10Gigabit Ethernet but 40Gigabit Ethernet and 100Gigabit Ethernet [4] enabled switches have also fallen. In 2020, the unit price per line rate of 40Gigabit Ethernet switch is more expen-

[1] National Institute of Informatics, Chiyoda, Tokyo 101–8430, Japan
[a)] reo_kashiwazaki@nii.ac.jp

sive than one of some 100Gigabit Ethernet switch. In the past, the modules to connect between NICs and the port of switches were also expensive and this has been a critical commercial problem to build a cheaper distributed system. Over the past several years, some vendors have started to provide inexpensive and adequate quality modules. Because of price competition, the price of the modules has been declining. The issue of the cost has already resolved.

The drives that compose the NAS have also changed dramatically over the past several years. In addition to the price reduction of disk drives, the capacity increasing and popularization of SSD has had a major impact. The users can easily create a storage system with a lot of 2.5-inch SSDs in a 2RU chassis. If the SSD is connected with SATA 6.0Gbps, the bandwidth of the interface can be a bottleneck of performance. When a motherboard that multiple NVMe M.2 SSDs with PCIe3.0 x4 are installed, users can create high-performance storage that does not degrade with the bandwidth of the interface. Thus, the network bandwidth can become the bottleneck in the next stage. 10Gigabit Ethernet is already not enough, and now 40Gigabit Ethernet installed storage systems are already sold in the market.

In this paper, we quantitatively and qualitatively evaluate a disk array with a unit price of about 100 USD per terabyte and an all-flash storage system with a unit price of 500 USD per terabyte that can be equipped with 40Gigabit Ethernet NICs. The author hopes to share the performance and tuning knowledge widely.

2. Benchmark applications

There are a lot of applications to evaluate storage systems. This section introduce several useful applications to evaluate the system.

2.1 Vdbench

Vdbench is a command line utility specifically created to help engineers and customers generate disk I/O workloads to be used for validating storage performance and storage data integrity[*1], developed by Oracle. It is written in Java with the objective of supporting Oracle heterogeneous attachment. At this time I/O has been tested on Solaris Sparc and x86, All flavors of Windows, HP/UX, AIX, Linux, Mac OS X, zLinux and RaspBerry Pi. The objective of the utility is to generate a wide variety of controlled storage I/O workloads, allowing control over workload parameters such as I/O rate, LUN or file sizes, transfer sizes, thread count, volume count, volume skew, read/write ratios, read and write cache hit percentages, and random or sequential workloads. This applies to both raw disks and file system files and is integrated with a detailed performance reporting mechanism eliminating the need for the Solaris command iostat or equivalent performance reporting tools.

SNIA Emerald [*2] is one of a program to provide public access to storage system power usage and efficiency through use of a well-defined testing procedure, and additional information related to system power. Measurement specification of SNIA Emerald[*3] adopts Vdbench.

2.2 SPC-1, SPC-1/E

SPC-1 Results provide a source of comparative storage performance information that is objective, relevant, and verifiable. That information will provide value throughout the storage product life-cycle, which includes development of product requirements, product implementation, performance tuning, capacity planning, market positioning, and purchase evaluations. The SPC-1 Benchmark is designed to be vendor/platform independent and are applicable across a broad range of storage configuration and topologies. Any vendor should be able to sponsor and publish an SPC-1 Result, provided their tested configuration satisfies the requirements of the SPC-1 benchmark specification[*4].

SPC-1 consists of a single workload designed to demonstrate the performance of a storage subsystem while performing the typical functions of business critical applications. Those applications are characterized by predominately random I/O operations and require both queries as well as update operations. Examples of those types of applications include OLTP, database operations, and mail server implementations. Otherwise, SPC-1/E is the second SPC benchmark extension, which consists of the complete set of SPC-1 performance measurement and reporting plus the measurement and reporting of energy use. This benchmark extension expands energy use measurement and reporting to larger, more complex storage configurations, complementing SPC-1C/E, which focuses on storage component configurations. Additional details are available in an SPC-1/E presentation available for viewing or download.

2.3 IOzone

IOzone[*5] is a filesystem benchmark tool. The benchmark program generates and measures a variety of file operations including sequential read/write, sequential reread/rewrite, backwards read, random read/write, record rewrite, strided read, fread/fwrite, freread/frewrite and pread/pwrite. Iozone has been ported to many machines and runs under many operating systems.

2.4 fio

fio is an I/O tool meant to be used both for benchmark and stress/hardware verification[*6]. It has support for 19 different types of I/O engines (sync, mmap, libaio, posixaio, SG v3, splice, null, network, syslet, guasi, solarisaio, and more), I/O priorities (for newer Linux kernels), rate I/O, forked or threaded jobs, and much more. It can work on

[*1] http://www.oracle.com/technetwork/server-storage/vdbench-downloads-1901681.html
[*2] https://www.snia.org/emerald
[*3] https://www.snia.org/emerald/download/Spec_v2.1
[*4] http://www.storageperformance.org/results/benchmark_results_spc1_active/
[*5] http://www.iozone.org
[*6] http://freecode.com/projects/fio

Table 1 A comparison of the specifications of each storage product.

	FS6400	SA3400
number of drives	72	36
drive	WD Red SA500	WD120EFAX-RT
total price (USD)	52,000	27,000
total capacity (TiB)	144 TiB	432 TiB
unit price per TiB	360	63

block devices as well as files. fio accepts job descriptions in a simple-to-understand text format. Several example job files are included. fio displays all sorts of I/O performance information, including complete IO latencies and percentiles. Fio is in wide use in many places, for both benchmarking, QA, and verification purposes. It supports Linux, FreeBSD, NetBSD, OpenBSD, OS X, OpenSolaris, AIX, HP-UX, Android, and Windows.

2.5 Oracle ORION

ORION (Oracle I/O Calibration Tool) is a standalone tool for calibrating the I/O performance for storage systems that are intended to be used for Oracle databases[*7]. The calibration results are useful for understanding the performance capabilities of a storage system, either to uncover issues that would impact the performance of an Oracle database or to size a new database installation. Since ORION is a standalone tool, the user is not required to create and run an Oracle database.

3. Evaluations

3.1 target products

In the end of last year, the author accidentally met an opportunity to get two storage products. One is Synology FlashStation FS6400[*8] and the other is Synology SA3400[*9]. The unit price per TiB of these two products are shown in Table 1. The unit of price is USD[*10]. The unit price per TiB of FS6400 is less than 500 USD and the one of SA3400 is less than 100 USD. I adopted Cisco Nexus 9332C[*11] as SAN 100Gigabit Ethernet Switch. Including the switch, the unit price of FS6400 is around 500 USD and the one of SA3400 is around 100 USD each. Each total capacity is a number with 2 external units. Each total price includes a cost of the units.

Specification of two products are described in Table 2. Each product certifies 40Gigabit Ethernet NICs such as Mellanox ConnectX series[*12].

[*7] http://www.oracle.com/technetwork/jp/topics/index-096484-ja.html
[*8] Synology FlashStation FS6400 https://www.synology.com/en-global/products/FS6400
[*9] Synology SA3400https://www.synology.com/en-global/products/SA3400
[*10] USD-JPY currencies rate is 106.9 JPY/USD (18 Juune 2020)
[*11] Cisco Nexus 9332C and 9364C Fixed Spine Switches Data Sheet https://www.cisco.com/c/en/us/products/collateral/switches/nexus-9000-series-switches/datasheet-c78-739886.html
[*12] Mellanox ConnectX Ethernet Adapters https://www.mellanox.com/products/ethernet/connectx-smartnic

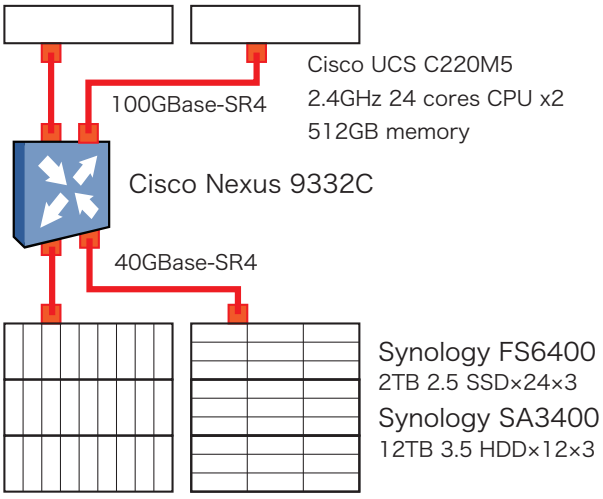

Fig. 1 A diagram of benchmark setup for IOzone

3.2 setting up

This paper shows evaluation results of network attached storage with 100GbE/40GbE environment. A diagram of an environment of evaluations is shown in Figure 1. FS6400 and SA3400 are connected to Cisco Nexus 9332C Fixed Spine Switches with 40GBase-SR4. And Nexus 9332C is also connected to Cisco UCS C220M5 with 100GBase-SR4. Ubuntu Linux 20.04 LTS (Focal Fossa)[*13] is running on the C220M5. Benchmark programs are executed on Ubuntu Linux. FS6400 and SA3400 provides block device with iSCSI and NFS service.

3.3 network performance benchmark

When configuring a SAN with 100Gigabit Ethernet, it is necessary to measure in advance whether the network will become a bottleneck. The network-tuning method of 100Gigabit Ethernet will be announced around 2016 by the Internet2 Technology Exchange by Energy Science Network (ESnet)[*14] and Stanford Research Computing Center[*15]. According to these documents, in the environment where two hosts with CentOS 7.2 installed are connected to a switch of 100 Gigabit Ethernet, the network performance measurement tool nuttcp[*16], 79Gbps throughput is measured. There are several parameters to change in performance tuning: TCP buffer, CPU governor, and MTU. In a LAN environment in which all measuring hosts are connected to one switch, jumbo frames with an MTU of 9000 are effective. However, this time, when measured in an environment of 100 Gigabit Ethernet, it was found that MTU=9000 does not always record the maximum throughput. Throughput

[*13] Ubuntu 20.04 LTS (Focal Fossa) https://releases.ubuntu.com/20.04/
[*14] Recent Linux TCP Updates, and how to tune your 100G host https://www.es.net/assets/Uploads/100G-Tuning-TechEx2016.tierney.pdf
[*15] 100g Network Adapter Tuning — Stanford Research Computing Center https://srcc.stanford.edu/100g-network-adapter-tuning
[*16] nuttcp - network performance measurement tool http://nuttcp.net/nuttcp/

Table 2 A comparison of specifications of each storage product.

	FS6400	SA3400
CPU	Intel Xeon Silver 4110	Intel Xeon D-1541
number of cores	8	8
CPU Frequency (GHz)	2.1 ~3.0	2.1 ~2.7
memory (GB)	512	128
number of NIC	40GbE x2	40GbE x2
size (mm)	264 x 482 x 724	264 x 482 x 724
internal file system	Btrfs/EXT4	Btrfs/EXT4

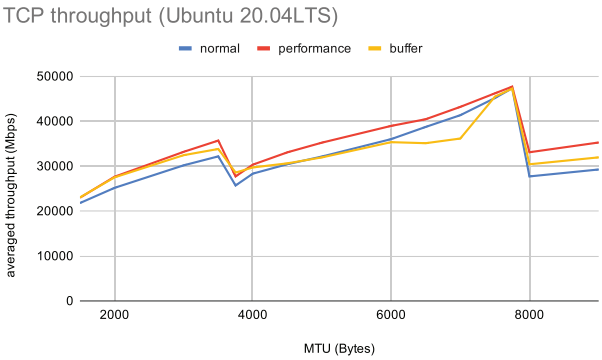

Fig. 2 TCP throughput changes and MTU change on Ubuntu 20.04 LTS

does not change linearly with increasing MTU.

Therefore, we constructed a program for searching the MTU value that records the maximum throughput. The user specifies three or more MTU values in advance. The program sets the specified MTU for the network interfaces of two opposing hosts and repeatedly executes the nuttcp benchmark a specified number of times. The MTU of the switch port is fixed at the maximum value of 9000. The average value of the benchmark results is calculated, and the change rate is calculated in the section between all the measurement points that have already been measured. When the rate of change becomes positive and negative in the adjacent sections, the midpoint of each of the two sections is set as a new search point. When the rate of change is positive or negative in all sections, the average of the lengths of all sections is calculated, and the midpoint of the section longer than the average section length is set as a new search point. By repeating this, the highest throughput value can be heuristically searched.

For this measurement, the Cisco UCS C220M5 connected to the Nexus 9332C has three OSs: CentOS 8[*17], Ubuntu 18.04 LTS and 20.04 LTS. Was installed respectively and the search results of the highest throughput value were compared. The figure 2 and 3 shows the change in the average throughput value with respect to the change in MTU.

3.4 Software Defined Storage Extension

In some documents about best practices, designers can be confused because of a lack of logical and quantitative proofs. For instance, a famous storage maker EMC publishes a document about best practices for performance[*18]. The document says that "Storage pools have multiple RAID options per tier for preferred type and drive count...(snip) Use RAID 6 for NL-SAS tier: Preferred drive counts of 6+2, 8+2, or 10+2 provide the best performance versus capacity balance. Using 14+2 provides the highest capacity utilization option for a pool, at the expense of slightly lower availability and performance". Why drive counts of 4+2 provides less performance than 6+2, 8+2 or 10+2? Why drive counts of 16+2 provides less capacity utilization option for a pool? Designers can not find any additional logical or quantitative explanations of the description.

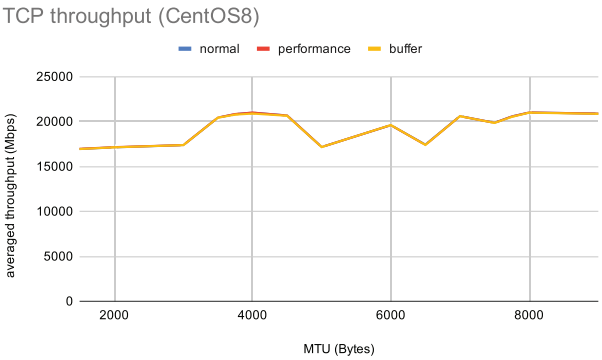

Fig. 3 TCP throughput changes and MTU change on CentOS 8

In order to eliminate this deception, we decided to introduce a mechanism for expanding the FS6400 and SA3400 and changing the RAID configuration in a programmable manner. Both FS6400 and SA3400 are provided with a management operation system called Synology DSM, and APIs for operating this DSM in a programmable manner are also provided. However, this API does not provide the function to change the RAID configuration. A Command Line Interface (CLI) tool is also provided, but neither is it provided with the ability to change the RAID configuration. Therefore, I created a program to add a disk to RAID F1 by imitating HTTPS communication so that an administrator could manually operate DSM using a Web browser. This makes it possible to measure storage performance according to the number of drives that make up RAID F1. This is called Software Defined Storage Extension in this paper.

3.5 benchmark with IOzone

In the paper, IOzone benchmark is used to measure throughput of storage systems. The benchmark can evaluate storage systems with various type of operations. According to the documentation of iozone[*19], the evaluation use 9 types of operations. The operations and their descriptions are as follows:

Write This test measures the performance of writing a new file. When a new file is written not only does the data need to be stored but also the overhead informa-

[*17] CentOS Project `https://www.centos.org/`

[*18] `https://www.emc.com/collateral/software/white-papers/h10938-vnx-best-practices-wp.pdf`

[*19] `http://www.iozone.org/docs/IOzone_msword_98.pdf`

tion for keeping track of where the data is located on the storage media. This overhead is called the "metadata" It consists of the directory information, the space allocation and any other data associated with a file that is not part of the data contained in the file. It is normal for the initial write performance to be lower than the performance of re-writing a file due to this overhead information.

Re-write This test measures the performance of writing a file that already exists. When a file is written that already exists the work required is less as the metadata already exists. It is normal for the rewrite performance to be higher than the performance of writing a new file.

Read This test measures the performance of reading an existing file.

Re-Read This test measures the performance of reading a file that was recently read. It is normal for the performance to be higher as the operating system generally maintains a cache of the data for files that were recently read. This cache can be used to satisfy reads and improves the performance.

Random Read This test measures the performance of reading a file with accesses being made to random locations within the file. The performance of a system under this type of activity can be impacted by several factors such as: Size of operating system's cache, number of disks, seek latencies, and others.

Random Write This test measures the performance of writing a file with accesses being made to random locations within the file. Again the performance of a system under this type of activity can be impacted by several factors such as: Size of operating system's cache, number of disks, seek latencies, and others.

Backwards Read This test measures the performance of reading a file backwards. This may seem like a strange way to read a file but in fact there are applications that do this. MSC Nastran is an example of an application that reads its files backwards. With MSC Nastran, these files are very large (Gbytes to Tbytes in size). Although many operating systems have special features that enable them to read a file forward more rapidly, there are very few operating systems that detect and enhance the performance of reading a file backwards.

Record Rewrite This test measures the performance of writing and re-writing a particular spot within a file. This hot spot can have very interesting behaviors. If the size of the spot is small enough to fit in the CPU data cache then the performance is very high. If the size of the spot is bigger than the CPU data cache but still fits in the TLB then one gets a different level of performance. If the size of the spot is larger than the CPU data cache and larger than the TLB but still fits in the operating system cache then one gets another level of performance, and if the size of the spot is bigger than the operating system cache then one gets yet another level of performance.

Strided Read This test measures the performance of reading a file with a strided access behavior. An example would be: Read at offset zero for a length of 4 Kbytes, then seek 200 Kbytes, and then read for a length of 4 Kbytes, then seek 200 Kbytes and so on. Here the pattern is to read 4 Kbytes and then Seek 200 Kbytes and repeat the pattern. This again is a typical application behavior for applications that have data structures contained within a file and is accessing a particular region of the data structure. Most operating systems do not detect this behavior or implement any techniques to enhance the performance under this type of access behavior. This access behavior can also sometimes produce interesting performance anomalies. An example would be if the application's stride causes a particular disk, in a striped file system, to become the bottleneck.

IOzone also has a lot of command line options. In the evaluation three options are enabled to exclude effects of cache. The options and their descriptions are as follows:

-c Include close() in the timing calculations. This is useful only if you suspect that close() is broken in the operating system currently under test. It can be useful for NFS Version 3 testing as well to help identify if the `nfs3_commit` is working well.

-e Include flush (fsync,fflush) in the timing calculations

-I Use `DIRECT IO` if possible for all file operations. Tells the filesystem that all operations to the file are to bypass the buffer cache and go directly to disk. (not available on all platforms)

The author wants to show the result but because of the end user license agreement, he has not obtain the agreement yet. In this proceedings, only the setting and results of network performance comparison are shown.

4. Conclusion

To share performance information under fair condition, the paper shows one result of benchmark with one storage product. The cost of evaluation is not negligible. To reduce the cost, the author will propose a system to collect various results. The author also hope to build a pay forward environment among evaluators, system designers, and product vendors.

References

[1] Frazier, H.: The 802.3z Gigabit Ethernet Standard, *IEEE Network*, Vol. 12, No. 3, pp. 6–7 (1998).
[2] : IEEE Standard for Information Technology - Telecommunications and information exchange between systems - Local and

Metropolitan Area Networks - Part 3: Carrier Sense Multiple Access with Collision Detection (CSMA/CD) Access Method and Physical Layer Specifications - Physical Layer Parameters and Specifications for 1000 Mb/s Operation over 4 pair of Category 5 Balanced Copper Cabling, Type 1000BASE-T, *IEEE Std 802.3ab-1999*, pp. 1–144 (1999).

[3] : IEEE Standard for Information technology - Local and metropolitan area networks - Part 3: CSMA/CD Access Method and Physical Layer Specifications - Media Access Control (MAC) Parameters, Physical Layer, and Management Parameters for 10 Gb/s Operation, *IEEE Std 802.3ae-2002 (Amendment to IEEE Std 802.3-2002)*, pp. 1–544 (2002).

[4] : IEEE Standard for Information technology– Local and metropolitan area networks– Specific requirements– Part 3: CSMA/CD Access Method and Physical Layer Specifications Amendment 4: Media Access Control Parameters, Physical Layers, and Management Parameters for 40 Gb/s and 100 Gb/s Operation, *IEEE Std 802.3ba-2010 (Amendment to IEEE Standard 802.3-2008)*, pp. 1–457 (2010).