Vol.2017-IOT-38 No.15
2017/6/24

比較的安価なNASの定性的・定量的な性能評価

柏崎 礼生[1,a)]

概要：ストレージの性能評価には様々な手法がある．本稿ではテラバイト単価 100USD 程度の比較的安価なストレージ製品を定性的，定量的に評価する．本稿は費用対効果を示すとともに，持続可能でベンダー横断的な評価へと展開する展望について述べる．

Qualitative and quantitative performance evaluations of less expensive network attached storage products.

Hiroki Kashiwazaki[1,a)]

Abstract: There are various approach to evaluate performance of storage systems. This paper shows qualitative and quantitative evaluations of storage products that cost around 100 USD per 1 TiB. The paper also proposes cost effectiveness of them and future prospect to develop these efforts to sustainable, cross-vendor evaluations of products.

Keywords: storage products, evaluation

1. Background

ICT systems consist of three kinds of elements. First of all, without computing nodes, no information can be processed. In the second place, without network communication interfaces and lines, no information can be communicated. At the end, third, without storage systems, any data can not be stored. The storage systems are connected to computing nodes with network communication interfaces and lines. Typically, the storage systems are connected with TCP/IP on Ethernet, FCP on Fibre Channel and so on. The storage systems can provide storage interfaces such as iSCSI, NFS, CIFS and so on. Especially in the case of very small ICT systems, storage is only one of components of computing nodes and there are no needs to implement external storage. Furthermore, recent rises of software defined storage can obsolete external storage systems. However, the systems are still main component for many ICT systems.

There are a lot of and various products of storage systems from ones for personal use to ones for enterprise use. Prices of the products also range widely. The budget is always limited. Designers of ICT systems always must also find an equilibrium point between a requirement specification and a budget limitation. There are two types of point of to evaluate products, qualitative views and quantitative views. An example of the qualitative views is a diagram of specifications. The designer can understand an overview of each product and can not understand a detail performance of each specification. According to some reports, company of the products show specifications but several functions of the specifications do not work properly [1]. Some failure of specification implementations can be fixed. Unfortunately, designers can only guess duration of the fixing.

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

[1] 大阪大学
Osaka University
[a)] reo@cmc.osaka-u.ac.jp

publishes a document about best practices for performance[*1]. 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.

It is useful for designers to share information about results of real performance evaluations of various storage products. If evaluations are done under unfair conditions, a result of the evaluations can become one of the FUD for the product. In some cases, unfair evaluations are prohibited by an end user license agreement (EULA) of the product. The evaluations should be done with cooperation from a maker, vendors of each product under fair conditions. This paper intends to introduce several methods to evaluate storage systems and shows results of quantitative evaluations of less expensive network attached storage products. The definition of "less expensive" is that the unit price per 1 TiB is around 100 USD. One of objects of the paper is to share results of quantitative evaluations under fair conditions and to call for collaboration to designers, makers, and vendors.

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[*2], 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 [*3] 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[*4] 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[*5].

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

[*1] https://www.emc.com/collateral/software/white-papers/h10938-vnx-best-practices-wp.pdf
[*2] http://www.oracle.com/technetwork/server-storage/vdbench-downloads-1901681.html
[*3] https://www.snia.org/emerald
[*4] https://www.snia.org/emerald/download/Spec_v2.1
[*5] http://www.storageperformance.org/results/benchmark_results_spc1_active/

storage component configurations. Additional details are available in an SPC-1/E presentation available for viewing or download.

2.3 IOzone

IOzone[*6] 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[*7]. 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 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[*8]. 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 last year, the author accidentally met an opportunity to get two storage products. One is NETGEAR ReadyNAS 516 (RN51600)[*9] and the other is Synology DS2015xs[*10]. The unit price per TiB of these two products are shown in Table 1. RN51600 is no longer on sale at the end of May 2017. The unit of price is USD. The unit price per TiB of these two products is around 100 USD.

表 1 A comparison of unit price per TiB of each storage product.

	RN51600	DS2015xs
purchase season	Mar. 2016	Mar. 2017
chassis price	1,650	1,560
HDD	WD60EFRX	WD80EFZS
HDD price	272	341
number of HDDs	6	8
total price	3,282	4,288
RAID level	5	6
total capacity	27 TiB	43 TiB
unit price per TiB	121.5	99.7

表 2 A comparison of specifications of each storage product.

	RN51600	DS2015xs
CPU	Intel Core i3-3220	Annapurna Labs Alpine AL-514
number of cores	2	4
CPU Frequency (GHz)	3.3	1.7
memory (GB)	4	4
number of NIC	GbE x 2	GbE x 2, 10GbE x 2
size (mm)	287.5 x 192 x 259	157 x 340 x 233
internal file system	BTRFS	EXT4

Specification of two products are described in Table 2. As the most significant aspect of DS2015xs, the product provides 2 port 10GbE SFP+ at the price. Unfortunately, evaluation with the 10GbE is not done in the paper because the author do not have any 10 GbE Switches. The author strongly hope some merciful network vendor give or rent me some 10 GbE Switches for an evaluative use. If the vendor would do so, the author intends to evaluate the products under fair conditions, and then also publishes the results and advertises the vendor on the workshop or symposium of IPSJ SIG-IOT and so on. If you are interested, please send the author E-mail.

And DS2015xs can also provide read-write and read-only SSD cache technology.

The paper only introduces Synology DS2015xs benchmark results because of space limitation. In the future paper, comparative evaluations between two or among more products will be published. Do not miss it.

[*6] http://www.iozone.org
[*7] http://freecode.com/projects/fio
[*8] http://www.oracle.com/technetwork/jp/topics/index-096484-ja.html
[*9] https://www.netgear.com/business/products/storage/readynas/RN51600.aspx
[*10] https://www.synology.com/en-global/products/DS2015xs

図 1 A diagram of benchmark setup for IOzone

3.2 setting up

If a laboratory of the author would have enough budgets, the paper can show evaluation results of network attached storage with 10GbE environment. A diagram of an environment of evaluations is shown in Figure 1. DS2015xs is connected to CentreCOM AT-x210-24GT GbE L2 network switch with 1000Base-T. And AT-x210-24GT is also connected to Apple Mac mini (Late 2012) with 1000Base-T. macOS Sierra (10.12.5) is running on Mac mini. Benchmark programs are executed on macOS. DS2015XS provides Apple Filing Protocol (AFP) over TCP/IP and macOS mounts a share point.

3.3 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[*11], 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 information 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 ex-

[*11] http://www.iozone.org/docs/IOzone_msword_98.doc

ample 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)

A file size of the benchmark is constantly 4 GB, a record size is dynamically changed doubling from 4 KB to 16 MB. Each benchmark evaluation is executed 40 times. Results of the benchmarks are shown in Figure 2. To compare with an more expensive network attached storage product, throughput performance of NFS storage attached to virtual machines on Hokkaido University Data Science Cloud System is also measured (Figure 3). Plotted points are averaged values of each combination of record size and its throughput. These points are connected with lines. Standard deviations of each combination are also described with error bars. The results of **Record Rewrite** are supposed to be effected by cache. The graph of **Record Rewrite** is omitted.

For the results of DS2015xs AFS throughput, the results of **write**, **rewrite**, **read** and **reread** show constant throughput in each record size. Total average **write**, **rewrite**, **read**, **reread** throughput through all record sizes are 70.1 MB/sec, 46.1 MB/sec, 95.9 MB/sec and 96.45 MB/sec. A detailed explanation of **rewrite** throughput degradation is waiting a response from product company.

Meanwhile the results of **random read**, **random write**, **bkwd read** and **stride read** show a characteristic pattern of throughput changes in each record size. Every results show the worst throughput in 16 KB record size. In **random read**, **random write** and **bkwd read**, throughput with 4KB and 8 KB record size show approximately same throughput with 512 KB record size. Each throughput with more than 1024 KB record size increase linearly. Standard deviations in **stride read** are larger than results of the other operations. The result shows the product may increase throughput jitters under stride read operations.

Although detail information for NFS storage of Hokkaido University Data Science Cloud System is not clear, the author makes sure that the storage must be more expensive. The cloud system consists of Citrix (Accelerite)) CloudPlatform[*12] and XenServer[*13][*14]. Users of the system can make virtual machines (VMs) and can attach NFS storage to VMs. In the comparative evaluation, number of virtual CPU of a VM is 10, main memory is 40GB, system storage is 222 GB. Capacity of attached NFS storage is 1 TB. An operation system of the VM is Ubuntu 17.04. Evaluation benchmarks are executed with the same operations and same options of benchmarks on Synology DS2015xs.

As contrasted to the benchmark results of Synology DS2015xs, the results of **write** and **rewrite** show linear increments between record size and throughput from record size 4 KB to 4096 KB. Although the results of **read** and **reread** also show linear increments, the peak throughput is on record size 1024 KB or 2048 KB. The peak throughput of **write**, **rewrite**, **read** and **reread** are 254.1 MB/sec, 280.3 MB/sec, 258.1 MB/sec and 254.2 MB/sec respectively. According to the results, the bandwidth between VMs and NFS storages can be supposed to be more than 1 Gbps. The results of **random read**, **random write** and **stride read** also show the similar characteristics between the record size and the throughput on **write**, **read** and **reread** operations. The characteristics of **rewrite** is similar to the one of **random write**, the peak of throughput is located between record size 4096 KB and 16384 KB.

[*12] `https://accelerite.com/products/cloudplatform/`
[*13] `https://xenserver.org/`
[*14] `https://www.citrix.com/content/dam/citrix/en_us/documents/customers/hokkaido-university-en.pdf`

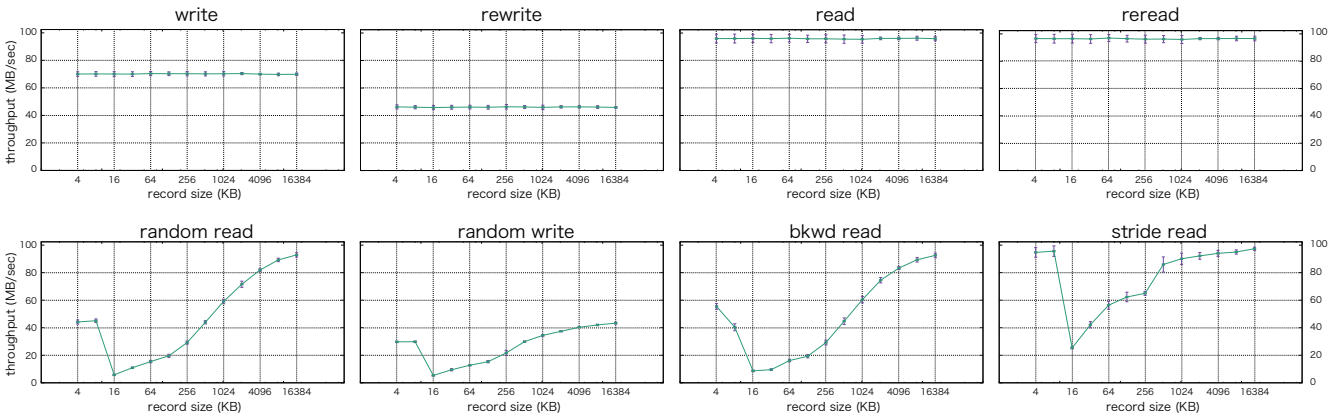

図 2 DS2015xs AFS I/O throughput results with 8 access variations (IOzone)

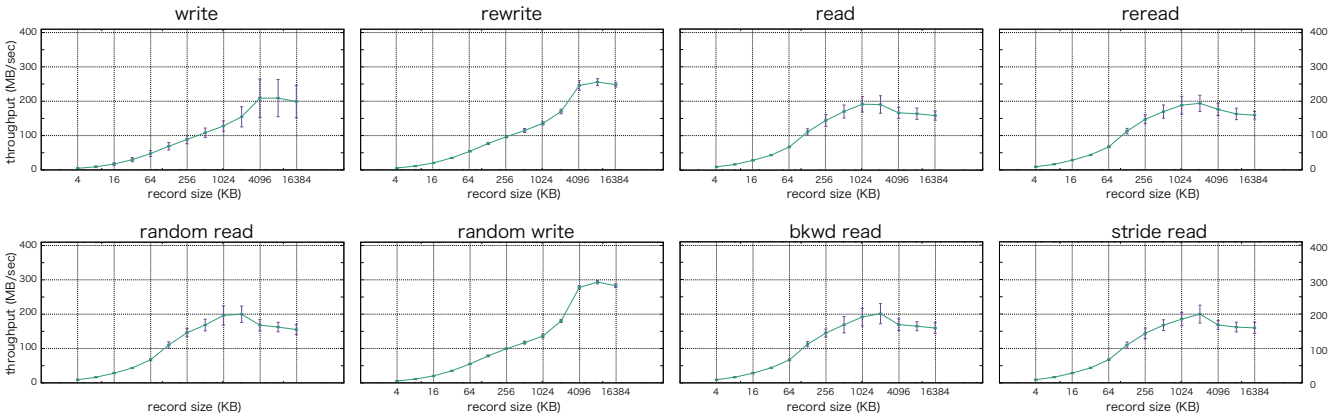

図 3 NFS storage of Hokkaido University Data Science Cloud System I/O throughput results with 8 access variations (IOzone)

4. Pay forward cycle

5. 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.

参考文献

[1] 中村　豊，佐藤彰洋：次世代ファイアーウォール機器の評価検証について，インターネットと運用技術シンポジウム 2016 論文集，Vol. 2016, pp. 106–106 (2016).