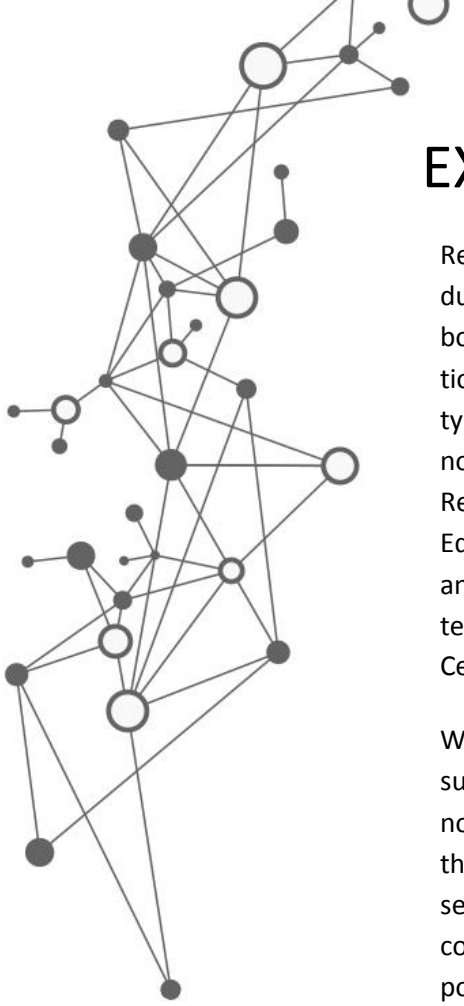
SOLUTION BRIEF

Red Hat Ceph Storage 3.0 on QCT Servers

Version 1.0



EXECUTIVE SUMMARY

Red Hat Ceph Storage has been a leading software-defined storage (SDS) in the market due to its high flexibility and scalability. The object-based storage solution is suitable for both structured and unstructured data. The solution is considered an ideal storage solution for business operating at large-scale data. Users can easily scale out storage capacity without impacting cluster performance. Together with Red Hat, Quanta Cloud Technology (QCT) has designed the QxStor Red Hat Ceph Storage Edition, which integrates Red Hat Ceph Storage with QCT in-house server systems. QxStor Red Hat Ceph Storage Edition comes in three configurations: throughput-optimized, cost/capacity-optimized, and IOPS-optimized. Each configuration adopts different hardware based on the characteristics of defined workloads, providing customers a selection of products to begin with Ceph.

With the release of Red Hat Ceph Storage 3.0, Ceph File System (CephFS) is officially supported by Red Hat. This is an exciting news for users since Red Hat Ceph Storage is now able to support three types of storage: object, block, and file. QCT understands that the pre-defined configurations need to be re-examined with newly-supported feature. A series of scenarios are designed to scrutinize the optimal configuration for CephFS. According to the test results, the underlying storage media in terms of CephFS metadata pool can influence CephFS performance.

In this brief, QCT will introduce the proposed hardware configuration for Red Hat Ceph Storage 3.0. With the pre-defined QxStor Red Hat Ceph Storage, enterprises will no longer need to spend time on studying the new Ceph features and matching these software features to suitable hardware devices. That is to say, enterprises can focus on applying the storage solution to their business, accelerating time-to-value.

CONTENT

Brief on Red Hat Ceph Storage 3.0

Red Hat Ceph Storage is an object-based, distributed software-defined storage with high scalability and reliability. It is designed to meet the dynamic storage demand faced by enterprise business. Major updates were made to Red Hat Ceph Storage, including the fully support of Ceph File System (CephFS), iSCSI gateway and installation of NFS Ganesha by using Ansible. This paper would focus mainly on validating CephFS performance.

CephFS is certified as a POSIX-compatible file system which requires at least one Metadata Server (MDS) node to run. The MDS daemon manages metadata of the files stored on CephFS and provides the access for these files to the shared Ceph storage cluster. There are two major differences between traditional network file system such as NFS or CIFS and CephFS. First, CephFS uses the POSIX semantics, which can maintain a strong cache coherency across the clients. With this feature, processes communicating on different hosts via the file system behave as if they are running on the same host. Second, CephFS provides high availability (HA) and scalability. The data on CephFS is distributed to all object storage daemon (OSD) nodes to prevent single point of failure (SPOF). When clients read data from CephFS, the data is streamed from multiple nodes instead of one single gateway.

Primarily, Red Hat Ceph Storage provides only block and object storages with high performance and flexible configurations. With the release of Red Hat Ceph Storage 3.0, it can now provide file storage as well, which allows users to aggregate multiple application scenarios in one single cluster.

''With the release of Red Hat Ceph Storage 3.0, it can now provide file storage as well, which allows users to aggregate multiple application scenarios in one single cluster.''

Introduction to QCT In-House Server for SDS

QuantaGrid D52T-1ULH server is an ultra-dense 1U server, as shown in Fig. 1. In this design, the sever is configured as a Ceph OSD node. Each OSD node is equipped with 12 HDDs and 4 SSDs. 12 HDDs are configured as a Ceph data pool while 3 SSDs are configured as a Ceph journal pool. The remaining one SSD is used for storing CephFS metadata. The exquisite hot-swappable SSD slots can save HDD space, increasing space utilization rate by 25%, compared to 1U 3.5" storage servers in market.

Figure 1. QuantaGrid D52T-1ULH storage server

Table 1 summarizes the hardware configuration of OSD node. QuantaGrid D52T-1ULH server with high-density storage features the Intel® Xeon® Scalable Processors and Intel 10GbE NICs, providing the computing power to fulfill Ceph storage tasks and networking bandwidth to accelerate throughput. To ensure the HA of OSD nodes, two dual port NICs are adopted for public and cluster networks. Each NIC is connected to the designated CPU with cross-NIC LACP teaming configured on each node.

Type	Commodity Name	Quantity Per Node
CPU	Intel® Xeon® Silver 4114 Processor	2
RAM	DDR4 2666MHz 32GB	4
NIC	Intel X710 Dual port 10GbE	1
NIC	Intel X722 Dual port 10GbE	1
SSD	Intel S4500 960GB	3+1
HDD	SAS 6TB	12
HBA	Avago 3008A	1
OS Drive	PCIe M.2 or Intel SSD	1

Table 1. Hardware configuration of each OSD node

Overview of Testing Tools and Testing Architecture

FIO and Smallfile are the tools adopted to benchmark CephFS performance. FIO can generate a number of threads to execute read/ write operations. FIO is used to test 100% sequential read and 100% sequential write performance over CephFS. On the other hand, Smallfile is a distributed metadata-intensive workload generator for POSIX-like file system. Smallfile is used to emulate the commonly-known operations in the file system to benchmark system performance. The total of 128 files, in which each file is 64GB, are pre-populated to CephFS before the testing.

The Ceph cluster contains 3 OSD nodes. The required Ceph services including Monitor (MON) node, Manager (MGR) node, and Metadata Server (MDS) are hosted on one server. The individual server with Zabbix and Ceph-metrics is set up to monitor system behavior. On the other hand, a total of 8 client nodes are deployed in which high-end hardware is used to ensure there's no hardware resource constraints on the client side. Figure 2 illustrates the topology of lab environment.

The entire system is connected to one single switch. Ceph public and cluster networks are separated by VLAN on the network switch. Each client node is equipped with one dual port 40GbE NIC. The dual port NIC LACP teaming is applied to each client node to form a total of 80GbE network bandwidth. On the other hand, each Ceph OSD node configured with two NICs with cross-NIC LACP teaming provides 20GbE network bandwidth for both public and cluster networks.

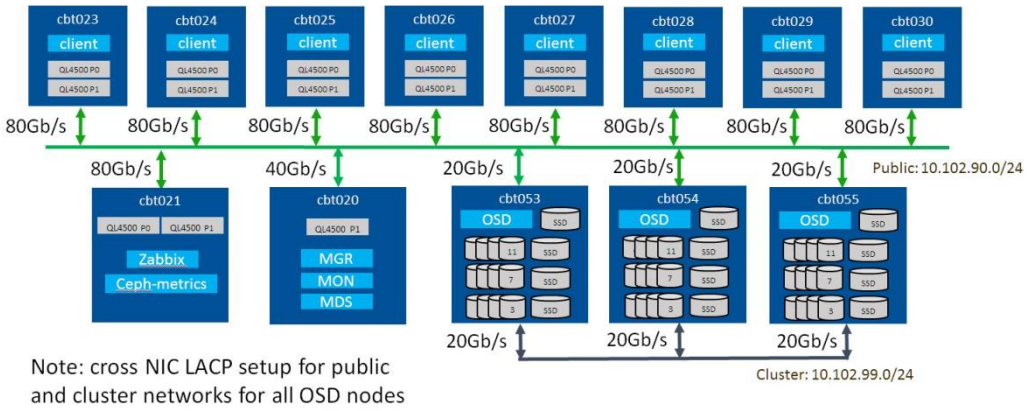


Figure 2. Testing topology of QCT lab environment.

Prior to reviewing CephFS testing results, the performance data of object and block storages are referred to as the baseline. Ideally, Ceph is expected to display stable performance across object and block storages. According to the test results in Table 2, when Ceph is used as an object storage, the sequential read throughput for the entire cluster is up to 3314 MB/s. Moreover, when Ceph is used as a block storage, the sequential read throughput for the entire cluster is up to 3360 MB/s. It is inferred that when Ceph is used as a file storage, an analogous throughput performance is expected.

Storage Type	Test Tool	Cluster Maximum Throughput (100% seq-read)
Object Storage	CBT RADOS bench	3314.88 MB/s
Block Storage	CBT librbdfio	3360.24 MB/s

Table 2. Maximum sequential read throughput when Ceph used as object or block storage

Performance Results of QxStor Red Hat Ceph Storage 3.0 Edition

FIO

"Ceph performs identically across the object, block and file storage."

The throughput is tested when Ceph is respectively used as object, block, and file storages. According to the test results in Table 3, Ceph performs identically across the three different storage types.

Storage Type	Test Tool	Cluster Maximum Throughput (100% seq-read)
Object Storage	CBT RADOS bench	3314.88 MB/s
Block Storage	CBT librbdfio	3360.24 MB/s
File Storage	CephFS FIO	3101.72 MB/s

Table 3. 100% sequential read throughput when Ceph used as object, block, and file storage

Users who retrieve the data via CephFS are directed to CephFS metadata pool first before accessing data in Ceph OSD; thus, the throughput performance of the underlying storage media in terms of CephFS metadata pool is likely to impact the cluster's perfor-

mance. To validate this inference, the throughput performances are compared when HDD and SSD are used as the underlying storage media in terms of CephFS metadata pool in 100% sequential read and 100% sequential write, respectively.

For 100% sequential read, the use of SSD shows better throughput performance under the concurrent jobs set between 62 and 112, as shown in Fig. 3. For 100% sequential write, either the use of SSD or HDD shows no apparent throughput difference, as shown in Fig. 4.

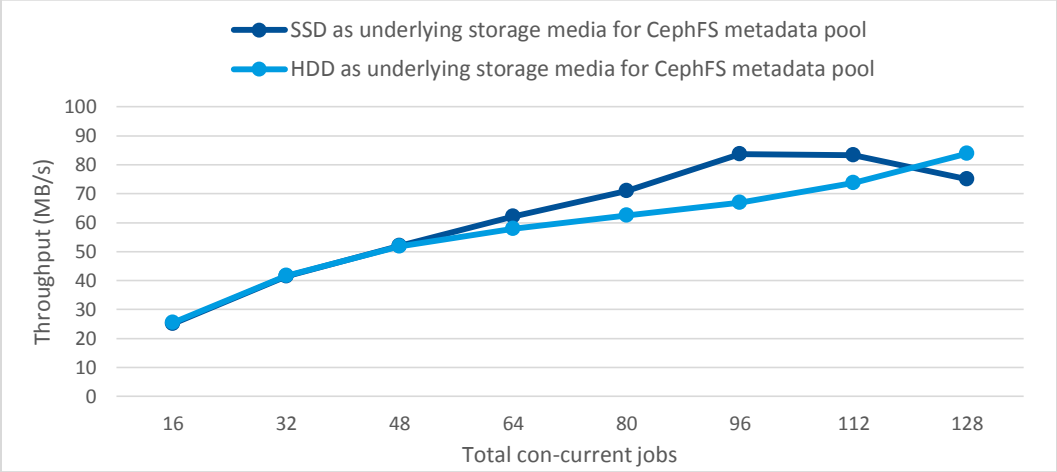


Figure 3. 100% sequential read throughput per OSD (64 GB file size, 4MB block size, kernel driver)

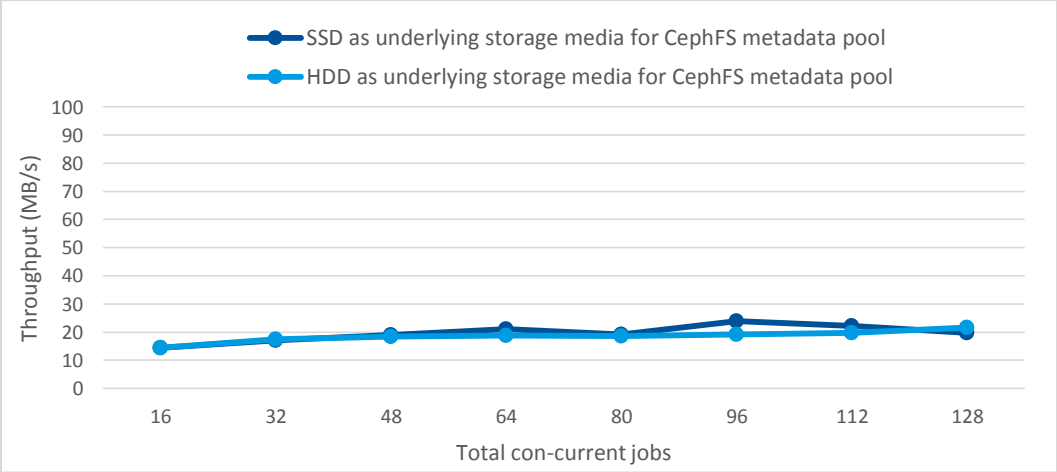


Figure 4. 100% sequential write throughput per OSD (64 GB file size, 4MB block size, kernel driver)

CephFS mounted with either kernel driver or Filesystems in Userspace (FUSE) are compared to evaluate the throughput performance in 100% sequential read and 100% sequential write, as respectively shown in Fig. 5 and Fig. 6.

The test result reveals that CephFS mounted with kernel driver generally shows better throughput performance than CephFS mounted with FUSE for each FIO session. Specifically, the throughput performance in 100% sequential read can increase up to 60% when the total concurrent job is 96.

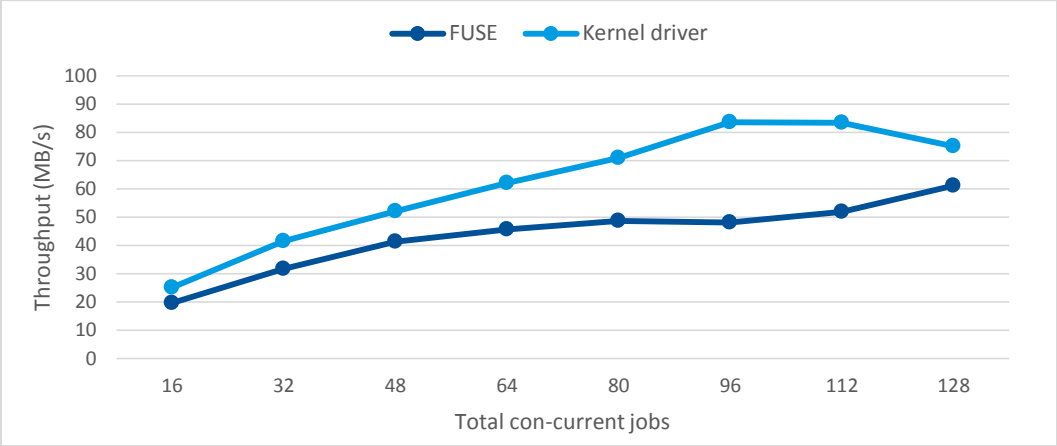
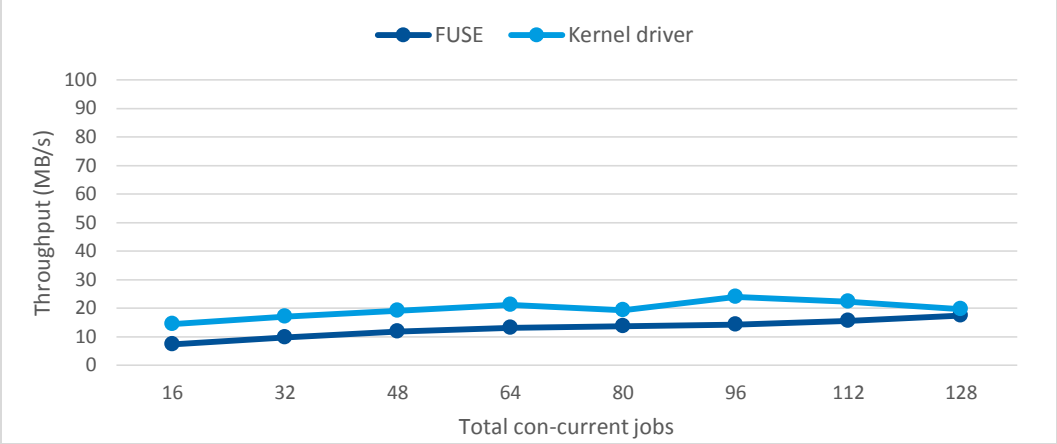


Figure 5. 100% sequential read throughput per OSD with CephFS metadata pool located on SSD



(64GB file size and 4MB block size)

Table 6. 100% sequential write throughput per OSD with CephFS metadata pool located on SSD (64GB file size and 4MB block size)

Smallfile

To provide a comprehensive configuration guide, five user scenarios, including (1) number of con-current threads, (2) file size, (3) CephFS mounted with either kernel driver or FUSE, (4) MDS cache memory size, and (5) the underlying storage media in terms of CephFS metadata pool are designed to evaluate CephFS performance using Smallfile.

As to the number of con-current threads, the performances among 32, 64, and 128 concurrent threads are compared. Generally, the result shows that the more the concurrent jobs is implemented, the better the file per second is achieved, as shown in Fig. 7.

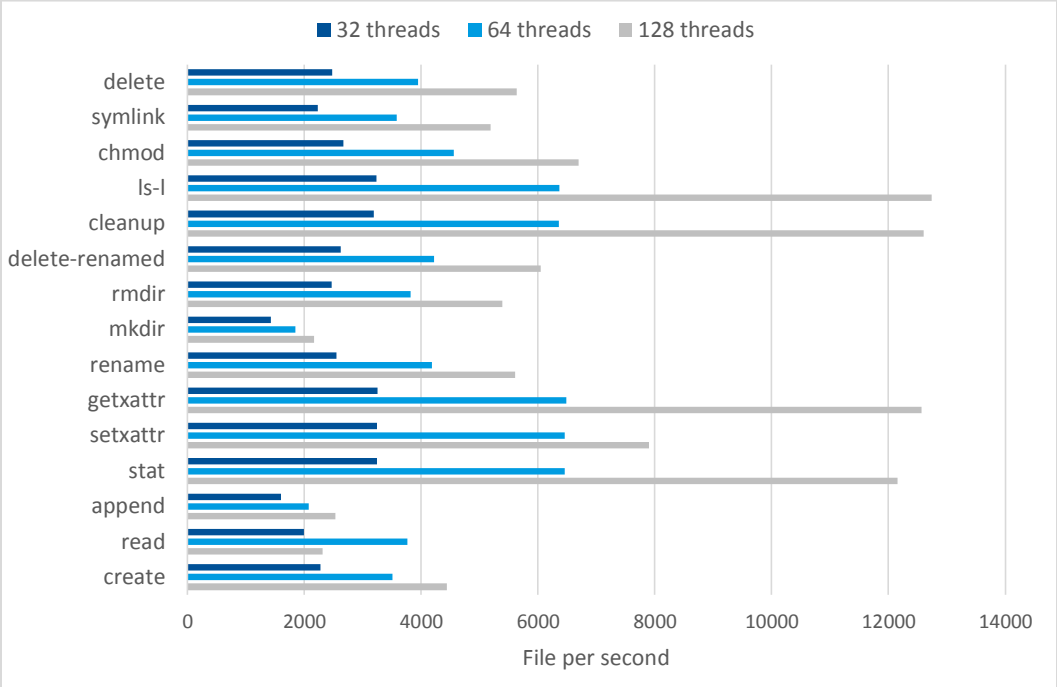


Figure 7. File per second performance for different file system operations based on 32, 64, and 128 threads (64KB file size, kernel driver)

As to the file size, the performances among 4KB, 16KB, and 64KB are compared. The file per second performances among the three file sizes are considerably similar, as shown in Fig. 8. It is supposed that the performance result for 4KB read is caused by cache behavior.

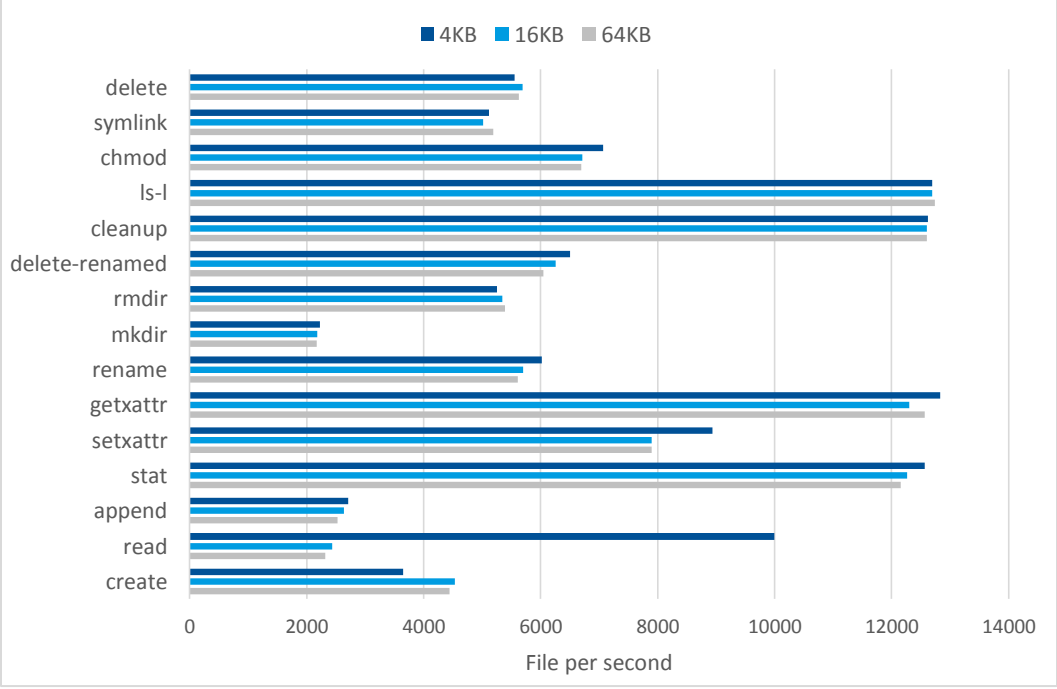


Figure 8. File per second performance for different file system operations based on file size of 4KB, 16KB and 64KB (128 con-current threads, kernel driver)

As to CephFS mounted with either kernel driver or FUSE, the result indicates that CephFS mounted with kernel driver provides better performance than CephFS mounted with FUSE, as shown in Fig. 9.

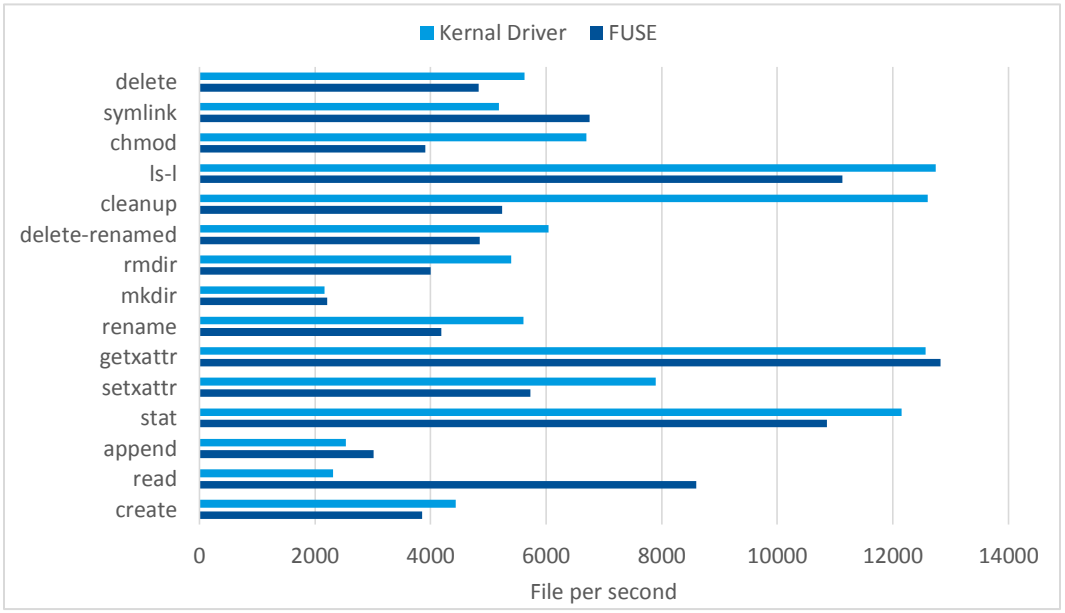


Figure 9. File per second performance for different file system operations with kernel driver or FUSE (64KB file size and 128 threads)

As to MDS cache memory size, the performances for 1GB and 64GB MDS cache memory are compared. It is observed that the MDS cache memory size in 64GB can generally provide better performance than the MDS cache memory size in 1GB, as shown in Fig. 10. For users' reference, Red Hat Ceph Storage sets MDS cache memory size at 1GB by default.

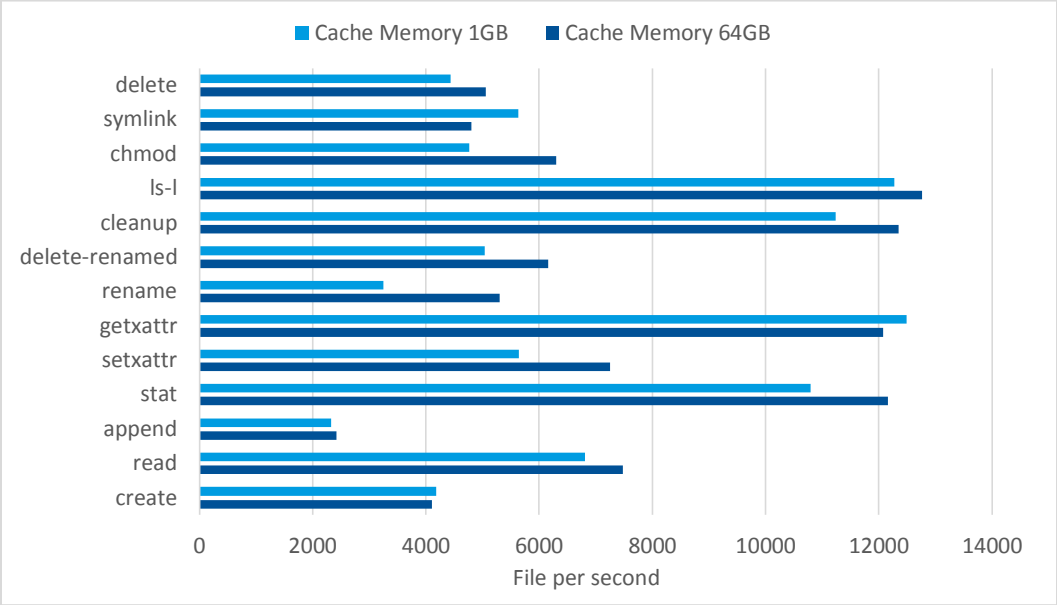


Figure 10. File per second performance for difference file system operation with cache memory size set at 1GB and 64GB (64KB file size and 128 con-current threads, kernel driver)

As to the underlying storage media in terms of CephFS metadata pool, the performances of using SSD and HDD are compared. It is observed that the use of SSD generates better file per second performance, compared to the use of HDD, as shown in Fig. 11. Ceph metadata pool is recommended to be configured on low-latency and high-IOPS storage media such as SSD or NVMe to accelerate file system operations.

"It is observed that the use of SSD as the underlying storage media in terms of CephFS metadata pool generates better file per second performance, compared to the use of HDD. "

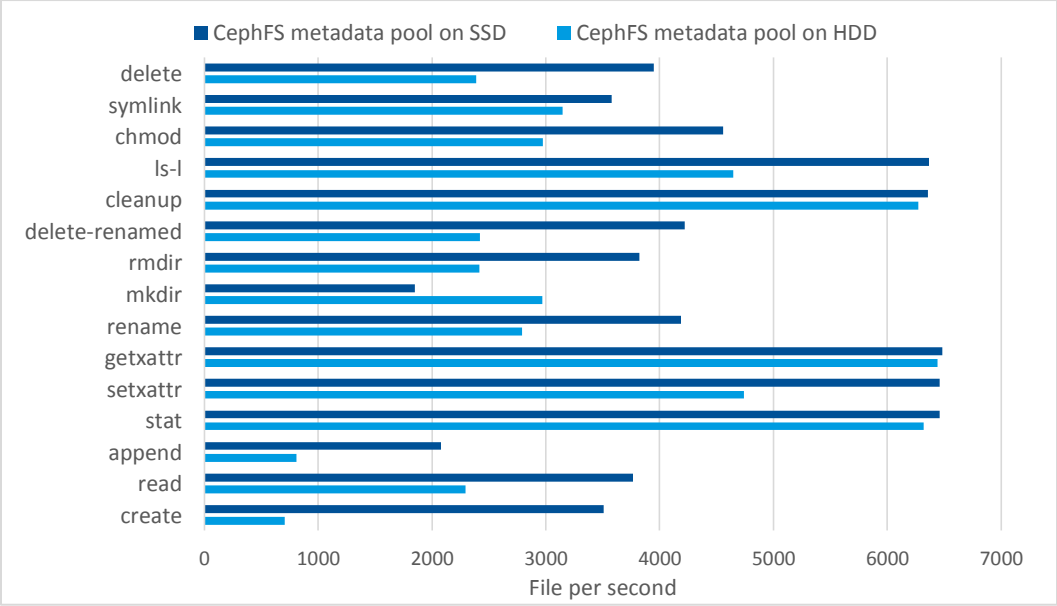
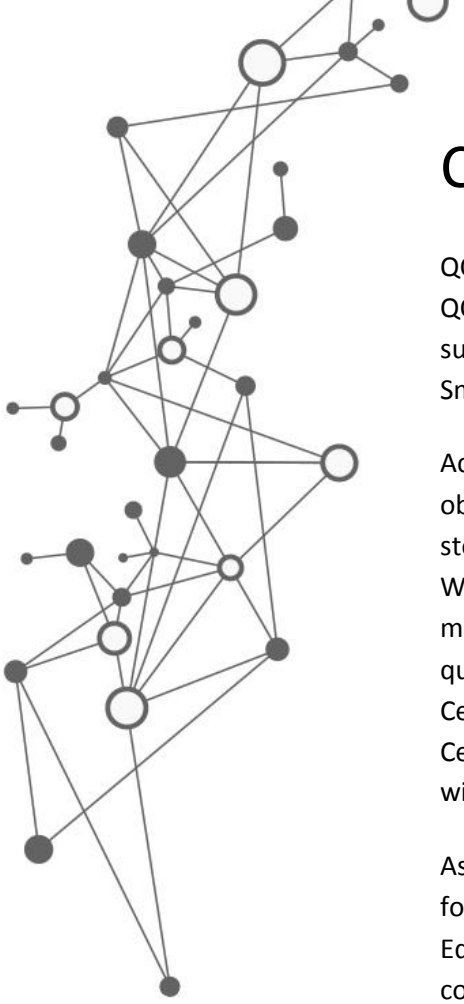


Figure 11. File per second performance for different file system operations with CephFS metadata pool located on SSD or HDD (64KB file size and 64 con-current threads, kernel driver)



CONCLUSION

QCT designed a series of testing scenarios to benchmark Red Hat Ceph Storage 3.0 on QCT in-house server, QuantaGrid D52T-1ULH. The test mainly focused on the newly-supported feature of Red Hat Ceph Storage - CephFS. The benchmark tools, FIO and Smallfile, are adopted to evaluate CephFS performance.

According to the testing results, it's proven that Red Hat Ceph Storage 3.0 used as object, block, or file storages shows similar performances. Moreover, the underlying storage media in terms of CephFS metadata pool is a key factor of cluster performance. When more file system operations are sent to the cluster, the objects stored in CephFS metadata pool are more frequently accessed. That is, when the underlying storage can quickly respond to a client's request, the client can relatively retrieve metadata from CephFS metadata pool faster to improve the performance of file system. To sum, CephFS metadata pool is highly recommended to be configured on the storage media with low latency and high IOPS.

As a total solution provider, QCT has been aiming at providing suitable sizing guidance for Red Hat Ceph Storage solution users. By adopting QxStor Red Hat Ceph Storage Edition, customers can greatly save time and effort in selecting the optimal configurations and software settings for Red Hat Ceph Storage 3.0. For more information about QCT QxStor Red Hat Ceph Storage Edition, please visit: http://go.qct.io/solutions/software-defined-storage/qxstor-red-hat-ceph-storage-edition/

LEGAL DISCLAIMER

INFORMATION IN THIS DOCUMENT IS PROVIDED IN CONNECTION WITH QUANTA CLOUD TECHNOLOGY (QCT) PRODUCTS. NO LICENSE, EXPRESS OR IMPLIED, BY ESTOPPEL OR OTHERWISE, TO ANY INTELLECTUAL PROPERTY RIGHTS IS GRANTED BY THIS DOCUMENT. EXCEPT AS PROVIDED IN QCT'S TERMS AND CONDITIONS OF SALE FOR SUCH PRODUCTS, QCT ASSUMES NO LIABILITY WHATSOEVER AND QCT DISCLAIMS ANY EXPRESS OR IMPLIED WARRANTY, RELATING TO SALE AND/OR USE OF QCT PRODUCTS INCLUDING LIABILITY OR WARRANTIES RELATING TO FITNESS FOR A PARTICULAR PURPOSE, MERCHANTABILITY, OR INFRINGEMENT OF ANY PATENT, COPYRIGHT OR OTHER INTELLECTUAL PROPERTY RIGHT.

UNLESS OTHERWISE AGREED IN WRITING BY QCT, THE QCT PRODUCTS ARE NOT DESIGNED NOR INTENDED FOR ANY APPLICATION IN WHICH THE FAILURE OF THE QCT PRODUCT COULD CREATE A SITUATION WHERE PERSONAL INJURY OR DEATH MAY OCCUR.

Quanta Cloud Technology (QCT) may make changes to specifications and product descriptions at any time, without notice. Designers must not rely on the absence or characteristics of any features or instructions marked "reserved" or "undefined." QCT reserves these for future definition and shall have no responsibility whatsoever for conflicts or incompatibilities arising from future changes to them. The information here is subject to change without notice. Do not finalize a design with this information.

The products described in this document may contain design defects or errors known as errata which may cause the product to deviate from published specifications. Current characterized errata are available on request.

All products, computer systems, dates, and figures specified are preliminary based on current expectations, and are subject to change without notice. Contact your local QCT sales office or your distributor to obtain the latest specifications and before placing your product order.

ABOUT QCT

Quanta Cloud Technology (QCT) is a global datacenter solution provider extending the power of hyperscale datacenter design in standard and open SKUs to all datacenter customers.

Product lines include servers, storage, network switches, integrated rack systems and cloud solutions, all delivering hyperscale efficiency, scalability, reliability, manageability, serviceability and optimized performance for each workload.

QCT offers a full spectrum of datacenter products and services from engineering, integration and optimization to global supply chain support, all under one roof.

The parent of QCT is Quanta Computer Inc., a Fortune Global 500 technology engineering and manufacturing company.

http://www.QCT.io

UNITED STATES

QCT LLC., Silicon Valley office
1010 Rincon Circle, San Jose, CA 95131
TOLL-FREE: 1-855-QCT-MUST
TEL: +1-510-270-6111
FAX: +1-510-270-6161
Support: +1-510-270-6216

QCT LLC., Seattle office
13810 SE Eastgate Way, Suite 190, Building 1,
Bellevue, WA 98005
TEL: +1-425-633-1620
FAX: +1-425-633-1621

CHINA

云达科技, 北京办公室（Quanta Cloud Technology）
北京市朝阳区东大桥路 12 号润诚中心 2 号楼
TEL +86-10-5920-7600
FAX +86-10-5981-7958

云达科技, 杭州办公室（Quanta Cloud Technology）
浙江省杭州市西湖区古墩路浙商财富中心 4 号楼 303 室
TEL +86-571-2819-8650

JAPAN

Quanta Cloud Technology Japan 株式会社
東京都港区芝大門 2-5-8 芝大門牧田ビル 3F, 105-0012
TEL +81-3-5777-0818
FAX +81-3-5777-0819

GERMANY

Quanta Cloud Technology Germany GmbH
Hamborner Str. 55, 40472 Düsseldorf
TEL +492405-4083-1

TAIWAN

雲達科技（Quanta Cloud Technology）
桃園市龜山區文化二路 211 號 1 樓
1F, No. 211 Wenhua 2nd Rd., Guishan Dist., Taoyuan City 33377, Taiwan