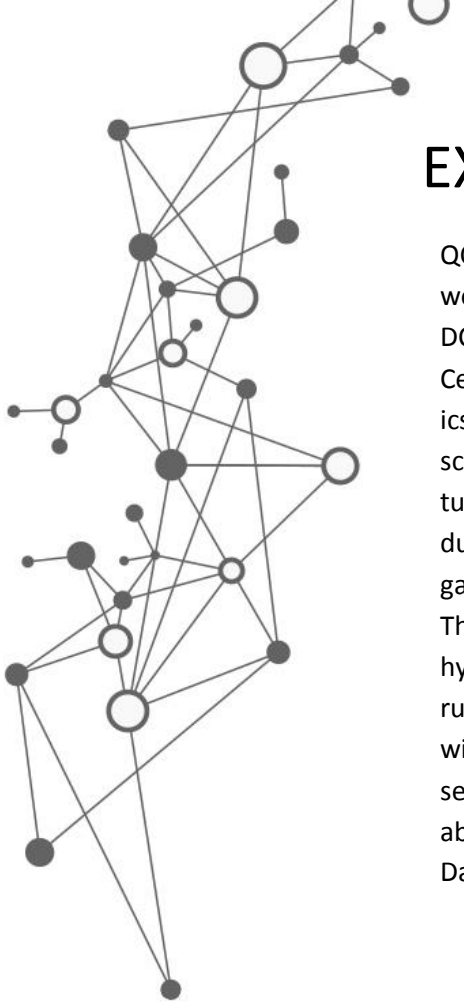
REFERENCE ARCHITECTURE

Real-time Analytics with All-Flash Ceph Data Lake Architecture

Version 1.0



EXECUTIVE SUMMARY

QCT storage team worked on a Data Lake reference architecture for Big Data analytics workloads based on All-Flash Ceph built on Intel Xeon Scalable processors and Intel SSD DC series. Combined with high performance hardware along with massively scalable Ceph storage, this configuration is optimized to meet the need of real-time data analytics processings. The four major benefits for disaggregated architecture are: on demand scaling on standard high-volume servers with performance guarantees - this architecture allows users to easily scale compute and storage independently to efficiently reduce costs and management complexity. Second, erasure coding with Ceph RADOS gateway provides 50% of storage savings compared to HDFS or RAID based systems. Third, comparable performances between disaggregated Ceph architecture and HDFS hyper-converged architecture are proved, ensuring that users do not have to sacrifice runtime performance using disaggregated data lake architecture. Lastly, compatible with Amazon S3, Ceph storage allows users to deploy on-premise Data Lake solution to service wide variety of analytics processing workloads on public cloud service. With the above benefits, the Data Lake with All-Flash Ceph architecture is ideal for large scale Big Data workloads.

Data Lake with All-Flash Ceph for Data Analytics

Paradigm Shift from Data-Generating to Data-Powered Enterprise

With the rise of social media, portable device, and cloud technologies, the amount of data has grown beyond expectation according to the numerous research predicting the exponential data growth by the year of 2020. Moreover, people have come to realize that data are no longer numbers or graphs that simply lie on reports, but important indicators of human behavior. Bearing this notable information in mind, enterprises are now eager to transform themselves from data-generating centers to data-powered companies, which means not just using existing data but finding the insights behind data and putting it to actual business uses.

Disaggregated Architecture Designed by QCT to Meet Big Data Demands

To meet enterprises' demand for data analytics, the open source Apache Hadoop (Hadoop) data analytics framework is widely adopted due to its reliability, low cost, and all-fit approach for both structured and unstructured data. However, its distributed computing cluster is implemented as disk-based direct-attached storage (DAS), which sacrifices overall scalability and makes it difficult for enterprise to scale their business.

To solve this problem, Quanta Cloud Technology (QCT) collaborates with leading processor, server systems, and storage technology provider Intel to design a Data Lake with All-Flash Ceph Storage architecture to meet the demands of data analytics workloads, as shown in Figure 1. Composed of the Hadoop ecosystem as well as a disaggregated Ceph storage, the reference architecture separates compute and storage resources, allowing users to scale the resources independently according to their actual workloads so as to achieve better optimization and efficiency of purchased appliances.

"Composed of the Hadoop ecosystem as well as a **disaggregated Ceph storage**, the reference architecture separates compute and storage resources, allowing users to scale resources independently according to their actual workloads."

By leveraging the compatibility of Ceph RADOS gateway (RGW) with S3 RESTful API, Ceph storage can be easily connected to data analytics clusters. The main reason for using RGW rather than RADOS block device (RBD) as the bridge to Ceph storage is better user authentication support. Using RBD as the interface, all clients access to an uniform resource pool, which means it is possible for clients to gain access to other clients' data easily. This scenario is likely to impose limitations for certain use cases. However, RGW can successfully solve this problem. Because RGW connects to data analytics clusters via Amazon S3, all users need to be assigned with an account to gain access to the storage pool. Under such mechanism, administrators can have better power management over the clusters, ensuring data confidentiality to allow multiple data analytics clusters to run consequently.

A load balancer in the architecture efficiently distributes network traffic across multiple RGWs to avoid performance bottleneck which can significantly improve overall performance.

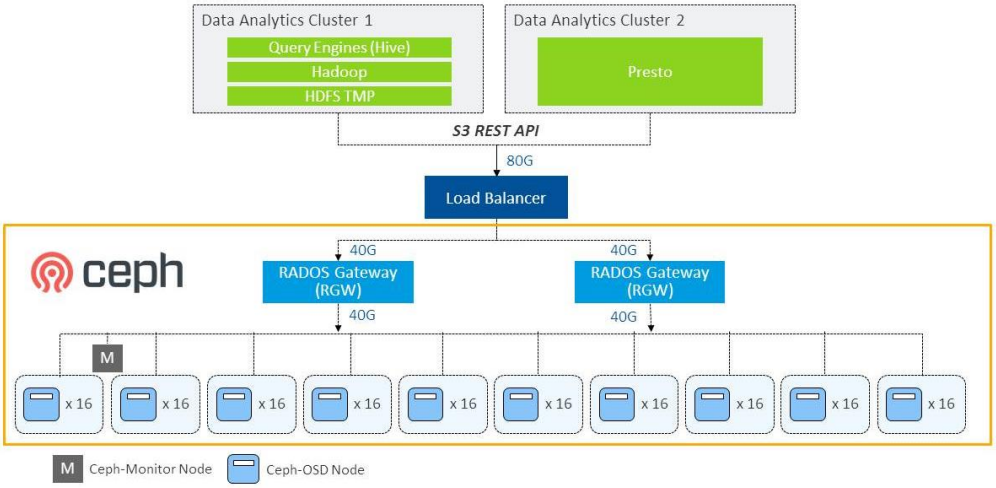


Figure 1. Data Lake with All-Flash Ceph Architecture

Software-Defined Storage Ceph for Petabyte Scale Storage

Ceph Storage is selected as the backend storage for the disaggregated architecture. As data grows faster than ever, the data growth cannot be managed by traditional storage systems; thus, a self-healing and self-managing software-defined storage system like Ceph would be highly preferred to meet customers' demand. With high scalability feature and compatibility for object, block, and file storage, Ceph has been particularlly designed to handle today and future's data storage needs.

Best-In-Class Hardware Applied to Stimulate Performance Capacity

The configuration is designed based on QCT in-house storage server models developed for tier 1 cloud service providers, ensuring optimized price capacity for different user scenarios. In order to comply with the high hardware requirement for data analytics workload, QCT QuantaPlex and QuantaGrid system hardware with Intel® SSD DC P4500 NVMe SSD is applied to augment the computing power and IOPS capacity to its maximum. The hardware specification is detailed in Fig 2.

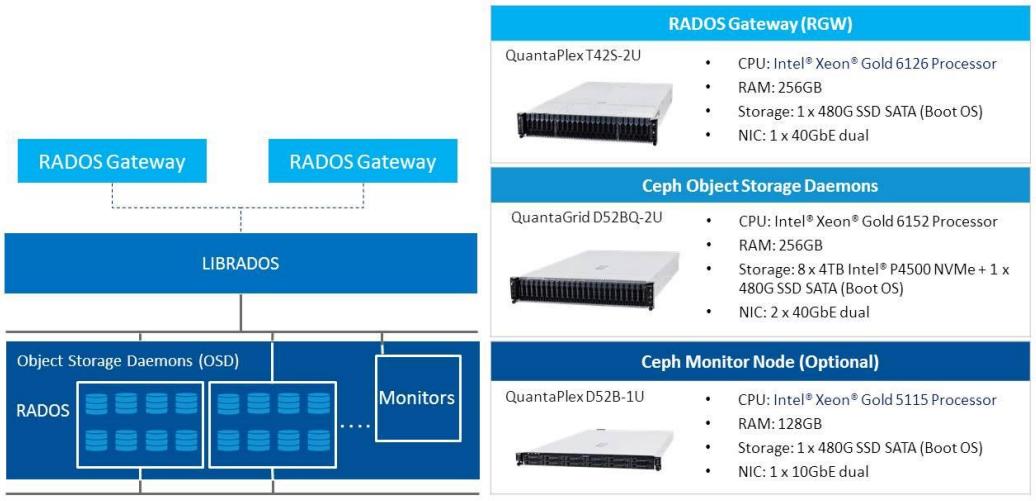


Figure 2. Hardware configuration for Data Lake with All-Flash Ceph Architecture

QCT QuantaPlex and QuantaGrid systems are versatile 2-socket rackmount servers with Intel® Xeon® Scalable processor scalable technology, flexible enough to support various software-defined workloads with 1.5x memory bandwidth and 3.9x higher throughput compared to previous version processors. With the increased computing power, this server plays a critical role as Hadoop data node serves read and write requests from the file system clients, and also performs block creation, deletion, and replication upon instruction from the name node.

All-Flash/NVMe configuration powered by Intel SSD DC series is adopted in current configuration based on couple reasons. NVMe is a high-performance PCIe interface specifically designed for SSD devices which offers significantly high IOPS. Users no longer have to add extra disks to cluster to meet the high IOPS requirement. Simultensouly, compared to conventional spindle disks, NVMe displays higher network traffic on incoming and outgoing traffic, higher disk read throughput, and up to 9 times increase in CPU utilization and disk latency while running Hive on the disaggregated Ceph architecture. The results indicate that NVMe reveals extraordinary performance matrix as shown in Table 1.

CPU Utilization	Network Traffic	Disk Read Throughput	Disk Read Latency
x9.24	Incoming: **x3.9** Outgoing: **x16.1**	**x2.81**	**x9.77**

Table 1. Performance improvement with NVMe compared to spindle disk

The price of NVMe SSD in recent years has made the All-Flash/NVMe configuration more attractive. Moreover, the storage capacity of SSD has now reached to the scale offered by spindle disk (or beyond thereof), enabling high storage density nodes in small

"NVMe can successfully run multiple analytical clusters consequently without sacrificing overall performance, which makes it suitable for mission-critical applications."

form factors. Last but not least, NVMe can meet the high bandwidth demand for mission-critical applications. With the high parallelism provided by NVMe, the cluster can run multiple applications concurrently without sacrificing overall performance.

In the testing scenario, the Intel® SSD DC P4500 4TB drive is used for the Hadoop/HDFS workload. Intel® SSD DC P4500 provides up to 3270 MB/s on the sequential read, 1860 MB/s on the sequential write, and 20 μs in write latency.

Hadoop Framework SQL Engines for Big Data Workloads

Figure 3 displays a simplified diagram of the testing architecture. Two architectures are built for comparison, one is based on Hadoop hyper-converged structure and the other is the proposed disaggregated architecture with Ceph as the backend storage. As for the data analytics workload, Hadoop is chosen as the framework for processing, storing, and analyzing big data due to its popularity. Two popular SQL engines, Apache Hive (Hive) and Presto, are selected as the workload for this testing. These SQL query engines used for querying database play important roles in analytics platform. Two column store database formats, orc and parquet, are selected for the testing to better resemble to real-life data analytics use cases.

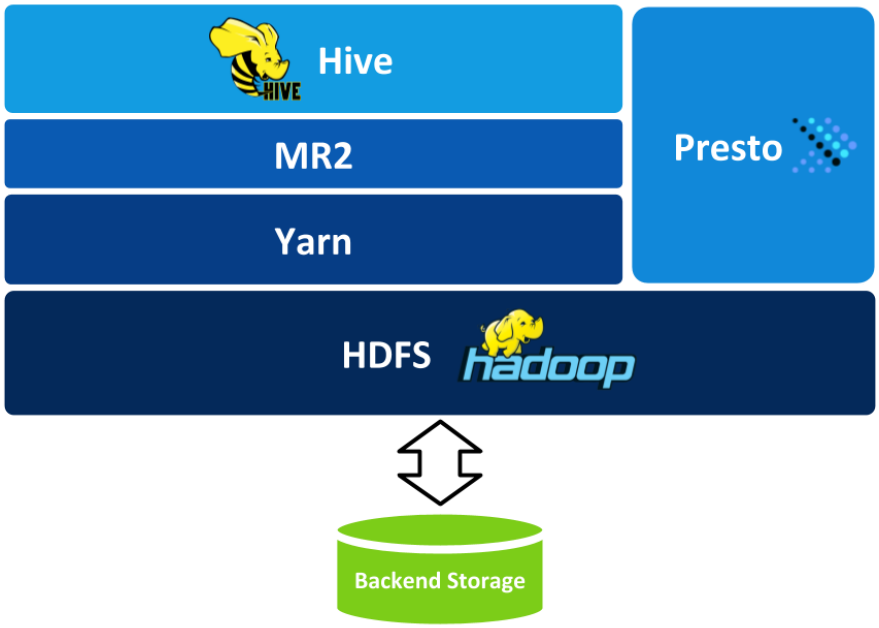


Figure 3. Testing architecture

TPC-DS chosen as the benchmark tool is the most common industry-standard decision support benchmark released by the Transaction Processing Performance Council (TPC). The benchmark test is defined as the execution of a database load test followed by a performance test. The database load is the process of building the test database while the performance test runs intensive SQL queries to measure the execution power of the system. Compute or IO intensive queries from TPC-DS are selected for better testing efficiency.

To provide broader reference, different scenarios are tested to better resemble actual workloads. Both 1TB and 10 TB data sizes are respectively tested with the two commonly used data types: ORC and Parquet. Thus, the total of 8 scenarios are tested.

"Performance penalty decreases as data size increases, which means that the **benefit of disaggregated architecture is more significant in larger data size**."

Comparable Runtime Results for Hive

The comparison between the two architectures show that the total runtime in the disaggregated architecture for Hive workload increases 5% to 20%, as shown in Fig 4. It is observed that the performance penalty decreases as data size increases from 1TB to 10TB, which means the benefit of disaggregated architecture is more significant in the larger data size. One potential reason for the results is that the compute nodes are now free from storage-related tasks in the disaggregated architecture, creating more available compute resource for analytical workloads.

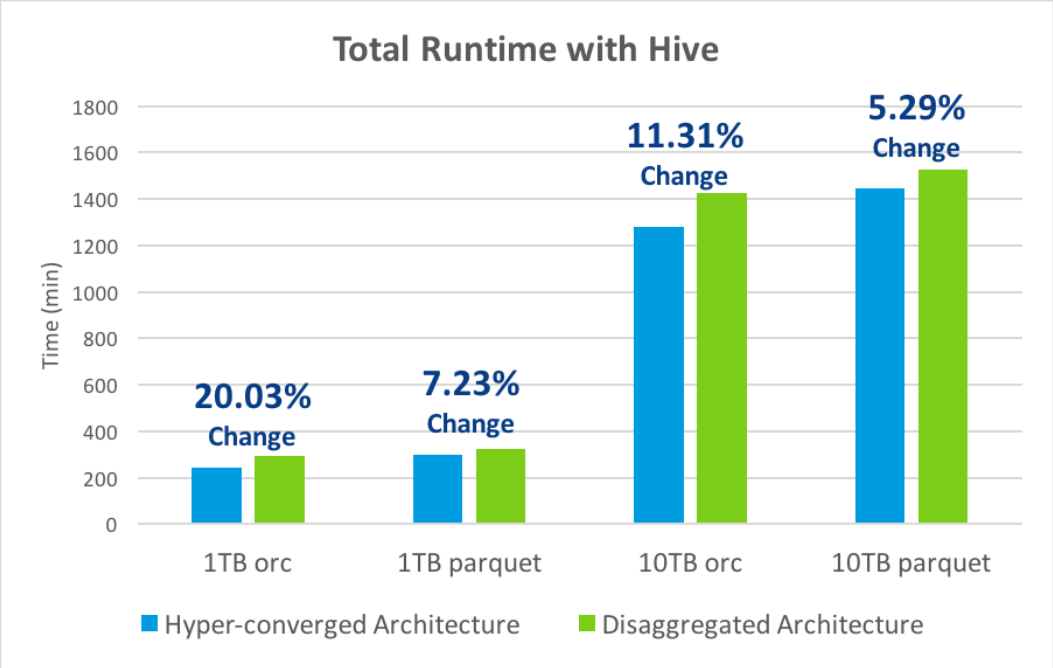


Figure 4. Testing results with Hive

Improved Runtime Results for Presto

"The results proved that disaggregating storage from the compute nodes unlocks the computing power of Presto."

The comparison between the two architectures shows that the disaggregated architecture performs better runtime than the hyper-converged architecture when running Presto. Especially in Parquet data storage format, up to 22% decrease in runtime is observed. The results proved that disaggregating storage from the compute nodes unlocks the computing power of Presto, as shown in Fig 5.

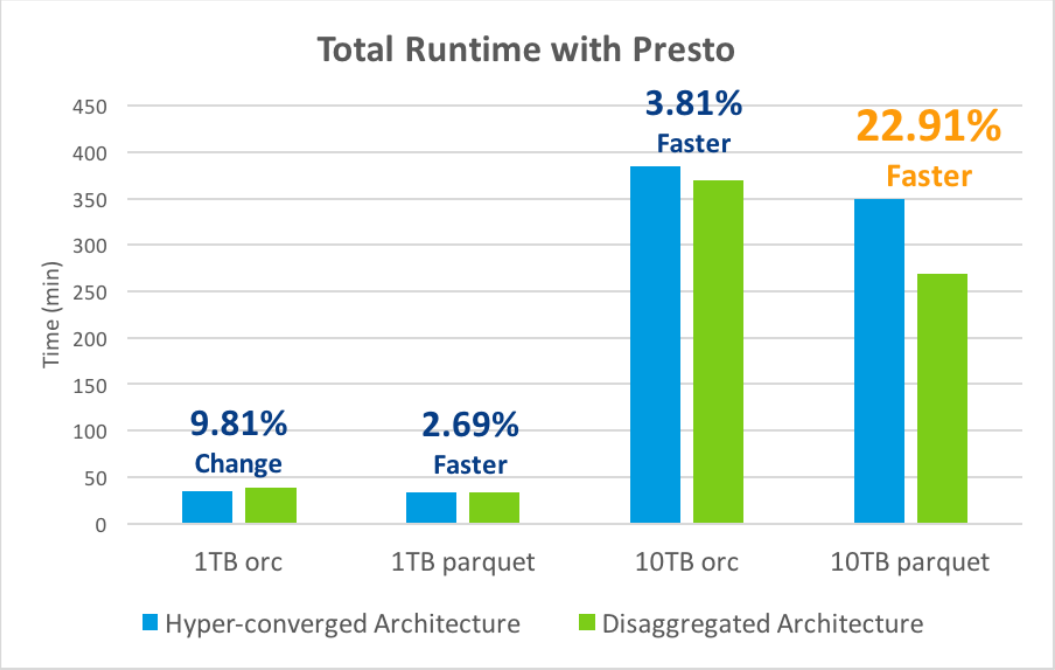
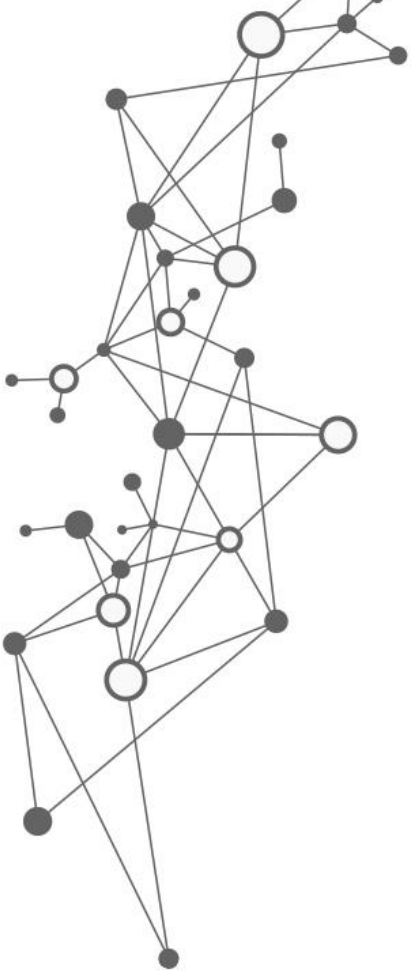


Figure 5. Testing results with Presto



CONCLUSION

As data grows faster than ever, traditional storage systems can no longer fulfill enterprise demand. Enterprises are also eager to transform from data-generating centre to data-powered company. To meet the need, QCT and Intel design a Data Lake with All-Flash Ceph for Data Analytics architecture. By disaggregating compute and storage appliances, users can easily scale out and conveniently run multiple analytic clusters in parallel.

The configuration adopts QCT QuantaGrid and QuantaPlex systems to provide high computing power via Intel Xeon Scalable processors and Intel SSD DC P4500 series to drive maximum IOPS. With Ceph storage, this configuration is suitable for mission-critical analytics workloads. According to the test result, both the performances of disaggregated and hyper-converged architectures are highly dependent on the workload characteristic and its flexibility for parameters tuning.

Major advantages gained from this architecture includes flexible scaling, less storage capacity purchased, assured performance level, and easy approach to connect to public cloud. By separating compute and storage, different components can be scaled out or replaced independently, which greatly reduces the cost and complexity for resource management. For instance, if users use SQL engine Presto for CPU and memory intensive workload, users can simply increase the number of compute nodes in current disaggregated cluster rather than purchase both storage and compute altogether.

Furthermore, with Ceph Storage, users can adopt erasure-coding technique to reduce the storage space required for data durability. The minimum storage capacity of Ceph using erasure-coding technique is significantly smaller than that of HDFS or RAID based storage system. The test results also proved that there's a 5% to 20% runtime difference between the disaggregated Ceph architecture and HDFS hyper-converged architecture, ensuring performance guarantee in All-Flash Ceph Data Lake architecture. Finally, compatible with Amazon S3, Ceph allows users to conveniently connect their cloud analytic clusters to on-premise Data Lake storage. Based on the aforementioned benefits and comparable performance results, the Data Lake with All-Flash Ceph for Data Analytics solution can be considered a suitable architecture for enterprise to conduct analytical tasks.

For more about QCT's QxStor Ceph solution, 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.

Intel technologies' features and benefits depend on system configuration and may require enabled hardware, software or service activation. Performance varies depending on system configuration. No computer system can be absolutely secure. Check with your system manufacturer or retailer or learn more at intel.com.
Intel disclaims all express and implied warranties, including without limitation, the implied warranties of merchantability, fitness for a particular purpose, and non-infringement, as well as any warranty arising from course of performance, course of dealing, or usage in trade.
Intel technologies' features and benefits depend on system configuration and may require enabled hardware, software or service activation. Performance varies depending on system configuration. No computer system can be absolutely secure. Check with your system manufacturer or retailer or learn more at intel.com.
Intel, the Intel logo, , and Xeon are trademarks of Intel Corporation in the U.S. and/or other countries.

ABOUT QCT

QCT (Quanta Cloud Technology) 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