HPDA Solution Technical White Paper

Issue 01

Date 2022-04-21

Huawei Technologies Co., Ltd.

Trademarks and Permissions


and other Huawei trademarks are trademarks of Huawei Technologies Co., Ltd.

All other trademarks or registered trademarks mentioned in this document are the property of their respective holders.

Notice

The purchased products, services, and features are stipulated by the contract made between Huawei and the customer. All or part of the products, services, and features described in this document may not be within the purchase scope or the usage scope. Unless otherwise specified in the contract, all statements, information, and recommendations in this document are provided "AS IS" without warranties, guarantees or representations of any kind, either express or implied.

The information in this document is subject to change without notice. Every effort has been made in the preparation of this document to ensure accuracy of the contents, but all statements, information, and recommendations in this document do not constitute a warranty of any kind, express or implied.

Huawei Technologies Co., Ltd.

Address: Huawei Industrial Base
Bantian, Longgang
Shenzhen 518129
People's Republic of China

Website: https://e.huawei.com/

About This Document

Overview

This document describes the technology evolution trends, solution architecture, and main functions of the HPDA storage solution.

Intended Audience

This document is intended for:

- Pre-sales engineers
- Delivery engineers

Symbol Conventions

The symbols that may be found in this document are defined as follows.

Symbol	Description
DANGER	Indicates an imminently hazardous situation which, if not avoided, could result in death or serious injury.
WARNING	Indicates a potentially hazardous situation which, if not avoided, could result in death or serious injury.
CAUTION	Indicates a potentially hazardous situation which, if not avoided, could result in minor or moderate injury.
NOTICE	Indicates a potentially hazardous situation which, if not avoided, could result in equipment damage, data loss, performance deterioration, or unanticipated results. NOTICE is used to address practices not related to personal injury.
NOTE	Supplements the important information in the main text. NOTE is used to address information not related to personal injury, equipment damage, and environment deterioration.

Release History

Issue	Date	Description
01	2012-11-15	This issue is the first official release.
02	2021-04-30	This issue is the second official release.
03	2022-04-30	This issue is the third official release.

Contents

1 Solution Overview

With the application and popularization of new technologies such as 5G, AI, big data, and blockchain, data is growing explosively, and users' requirements for storage systems become higher and higher. Conventional high-performance computing (HPC) construction faces many challenges, such as insufficient industrialization capabilities, high O&M costs, and insufficient computing power and application support. The design goal of supercomputing is also changed from purely pursuing the computing capability of floating-point operations per second to simultaneously pursuing computing and data processing capabilities. Traditional scientific computing is converged in view of new computing architectures such as High Performance Data Analytics (HPDA) and HPC-based AI.

Designed for new data-intensive supercomputing scenarios, the HPDA storage solution integrates cutting-edge technologies such as HPC, AI, and big data. Based on storage, the HPDA solution is built from bottom to top to provide diversified computing storage platforms, improving service efficiency, reducing O&M costs, and ensuring data security. It is applicable to scenarios such as the national supercomputing center, new AI computing center, scientific apparatus, national laboratory, and enterprise HPC platform.

- National supercomputing center: With world-leading supercomputing resources, the national supercomputing center enhances the convergence of HPC, AI, and big data to form a hybrid computing and storage centers with diversified computing capabilities, bringing social benefits and enabling industry clusters based on application requirements. It provides powerful supercomputing resources for enterprises and scientific research institutions.
- New AI computing center: Based on AI computing clusters, it provides full-stack AI capabilities from bottom-layer chip computing power to top-layer application enablement. AI computing centers lay a foundation to enable applications in various industries.
- Scientific apparatus: Large scientific devices are completed through large-scale investment and engineering construction. After the construction is completed, these devices can achieve important scientific and technological objectives through long-term stable operation and continuous scientific and technological activities. The involved technologies are comprehensive and complex. A large number of non-standard devices need to be developed, and diversified HPC platforms are required.
- ● National laboratory: A national scientific research institution that carries out basic research, cutting-edge high-tech research, and social welfare research for

national modernization and social development. In general, it actively undertakes major national scientific research tasks.

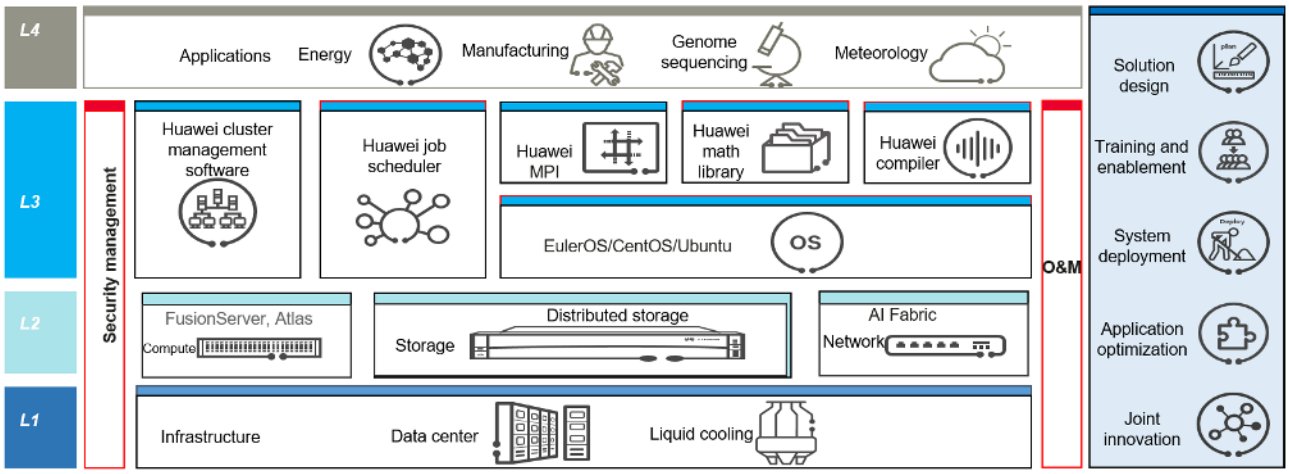
- Enterprise HPC platform: Enterprises build their own HPC platforms and use HPC technologies to efficiently process service data, improving service efficiency and accelerating service innovation.

2 Solution Architecture

2.1 HPDA Solution Architecture

Figure 2-1 shows the overall system architecture of the Huawei HPDA solution.

Figure 2-1 HPDA solution architecture



The architecture of the HPDA solution comprises four layers. From the lowest to the highest, they are the equipment room facility, hardware platform, cluster software platform, and application platform layers. The following table describes major components in the solution and their functions.

Table 2-1 HPDA solution components and functions:

Function Layer	Function
L4 application platform	Service application system used by a user
L3 cluster software platform	• Manages and monitors cluster resources, and

Function Layer	Function
	provides middleware such as compilers, parallel libraries, and math libraries. • Provides management and scheduling software to allow visualized job submission and quick application integration, and supports various scheduling policies and cluster management capabilities, effectively improving scheduling policies.
L2 hardware platform	Storage: provides high-performance parallel file storage, object storage, and HDFS storage. Computing: provides CPU computing power and GPU/NPU computing power. Network: provides high-performance networks and management networks.
L1 equipment room facilities	Provides cabinets, power supply, and heat dissipation capabilities.

This document describes the HPDA storage solution that features ultra-high energy efficiency, heterogeneous acceleration, and autonomous controllability.

Huawei HPDA Solution uses OceanStor Pacific series parallel file storage as its core and integrates peripheral resources such as compute, network, and scheduling software to provide more efficient, cost-effective, and reliable storage services. The HPDA solution fits perfectly in any HPC environment, and features the following highlights:

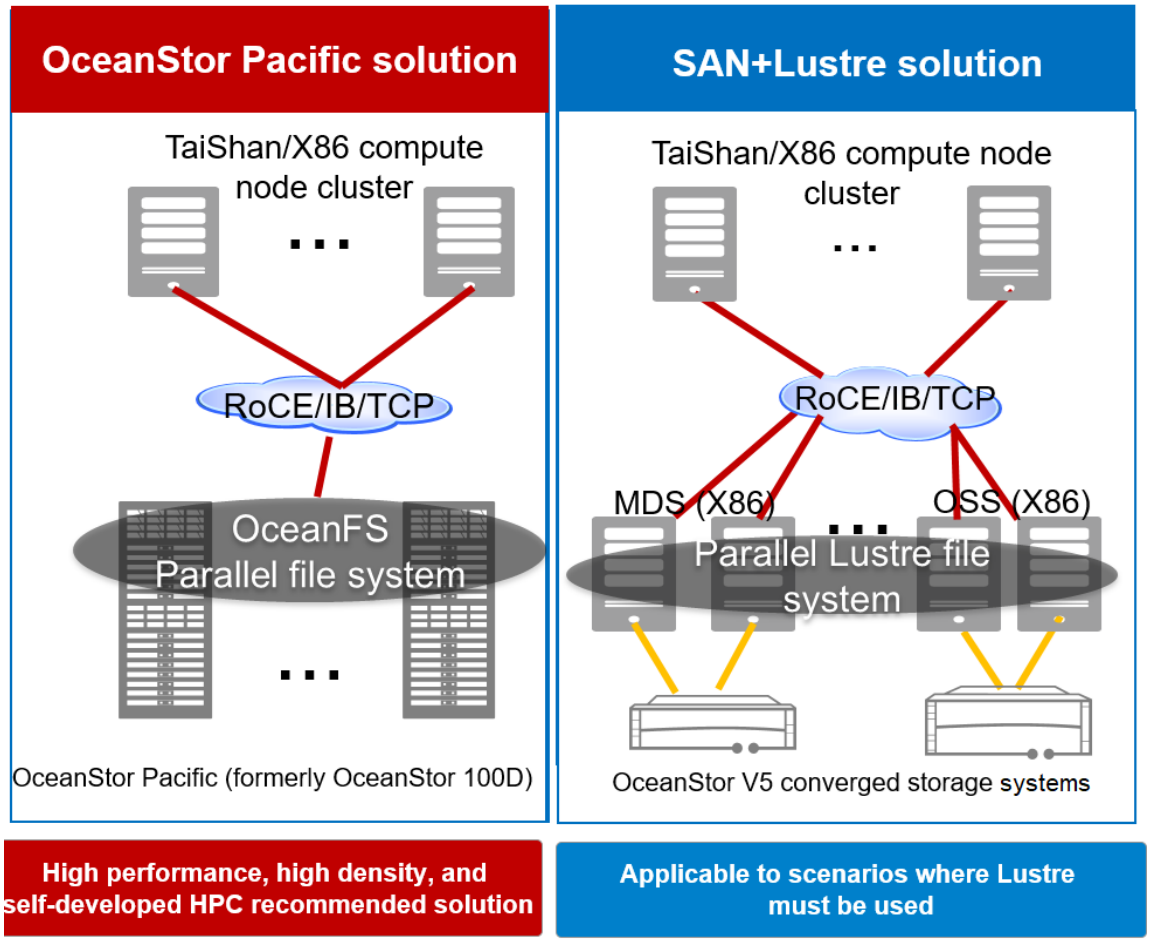
- Large capacity and flexible expansion, addressing the challenges of rapid data growth, reducing the capital expenditure (CAPEX), and protecting long-term investment
- Lossless interworking among file, object, and HDFS protocols and zero data migration in HPDA scenarios, reducing TCO and power consumption
- Innovative architecture and high-density all-flash hardware for high bandwidth and OPS in one storage system
- Innovative algorithm and large-ratio elastic erasure coding (EC), improving utilization without compromising performance
- Innovative solutions and intelligent tiering of cold and hot data, ensuring both performance and cost-effectiveness
- End to end (E2E) DIF verification and comprehensive sub-health check, protecting mass data security

NOTE

This document focuses on the HPDA storage solution. Components such as computing, network, scheduling software, and equipment room facilities are mainly included in Huawei-partner cooperation solutions, but not offerings of this solution.

2.2 HPDA Storage Solution Architecture

Figure 2-2 HPDA storage solution architecture



The HPDA storage solutions include the OceanStor Pacific solution and SAN+Lustre solution.

Architecture	Description
OceanStor Pacific solution	A fully symmetric parallel storage system uses single-layer storage hardware devices, provides MPI-IO parallel access and POSIX interfaces, and supports access protocols such as NFS, SMB, and HDFS. The storage nodes are deployed in the full symmetric architecture. System data and management data (metadata) are distributed on each node, eliminating system resource contending and system bottlenecks.
SAN+Lustre solution	An asymmetric parallel storage system uses two-layer storage hardware devices (I/O servers and block storage devices) and provides MPI-IO parallel access and POSIX interfaces. In this system, metadata and file data are separated and stored on different servers.

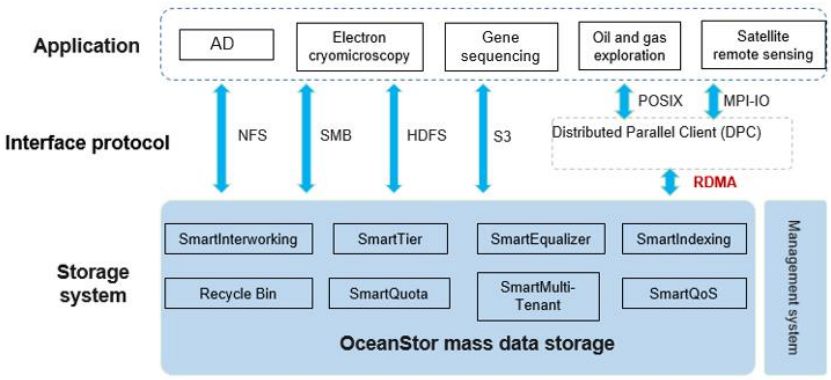
3 Solution Introduction

3.1 Storage System

3.1.1 OceanStor Pacific Solution

Huawei OceanStor Pacific storage system works in all-active share-nothing mode, which features massive scale-out capability and hundreds of PB global unified namespace. The storage system integrates distributed file, object, and HDFS services. It allows file, object, and HDFS services to be deployed in the same cluster and managed in a unified manner, and provides abundant service functions and value-added features. Concurrent access to different areas of the same file from multiple nodes is supported, implementing high-concurrency read/write and achieving high-performance access.

Figure 3-1 OceanStor mass data storage



Highlights of Huawei OceanStor Pacific storage system:

- **Fully symmetric distributed architecture, enabling ultimate elastic scalability**

  OceanStor Pacific series adopts a fully symmetric distributed architecture. It enables a linear growth in system capacity and performance by increasing storage nodes, requiring no complex resource requirement plans. It can be easily expanded to contain thousands of nodes and provide EB-level storage capacity. This helps meet your future storage demands. OceanStor Pacific series evenly distributes data and metadata on nodes and automatically balances loads, eliminating metadata access bottlenecks and ensuring system performance after capacity expansion. The system leverages FlashLink technologies such as intelligent stripe aggregation and I/O priority scheduling, as well as multi-level caches, big data passthrough, and other key technologies to deliver high bandwidth and low latency. It is adaptable to any customer need for I/O, latency, bandwidth, or capacity and supercharges the technology of today for the business of tomorrow.

- **Convergence and interworking of unstructured data storage services**

  OceanStor Pacific series supports the interworking of file, object, and HDFS unstructured data storage services. One piece of data can be shared and accessed by all unstructured services. The storage system provides file services based on the NFS protocol, SMB protocol, standard POSIX interface, and MPI-IO library, meeting customers' requirements on high bandwidth and low latency in HPDA scenarios. Working with OceanFS, a Huawei-developed parallel file system that supports MPI-IO, the unstructured services also provide the object service compatible with the Amazon S3 protocol and the HDFS service. Multiple unstructured services support protocol interworking and mutual data access, eliminating data migration and saving space.

- **Superb capacity and performance density design**

  OceanStor Pacific series leverages high-density hardware to help customers save equipment room space. Specifically, the ultra-large capacity design helps store more data with fewer devices; the superb performance density design helps quickly adapt to the all-flash trend. OceanStor Pacific 9550 adopts the superb capacity density design where a high-cohesion structure is used to maximize the disk density. In addition, dual air channels and counter-rotating pressurized fans are used to resolve the heat dissipation problem. It is an ultra-high-density storage device that accommodates to 1.1 m cabinets. OceanStor Pacific 9950 adopts the superb all-flash performance density design. It is an all-PCIe 4.0 flash storage device that adopts the small-node design and uses Huawei-developed half-palm NVMe SSDs to provide customers with superb performance experience.

3.1.1.1 Parallel File System

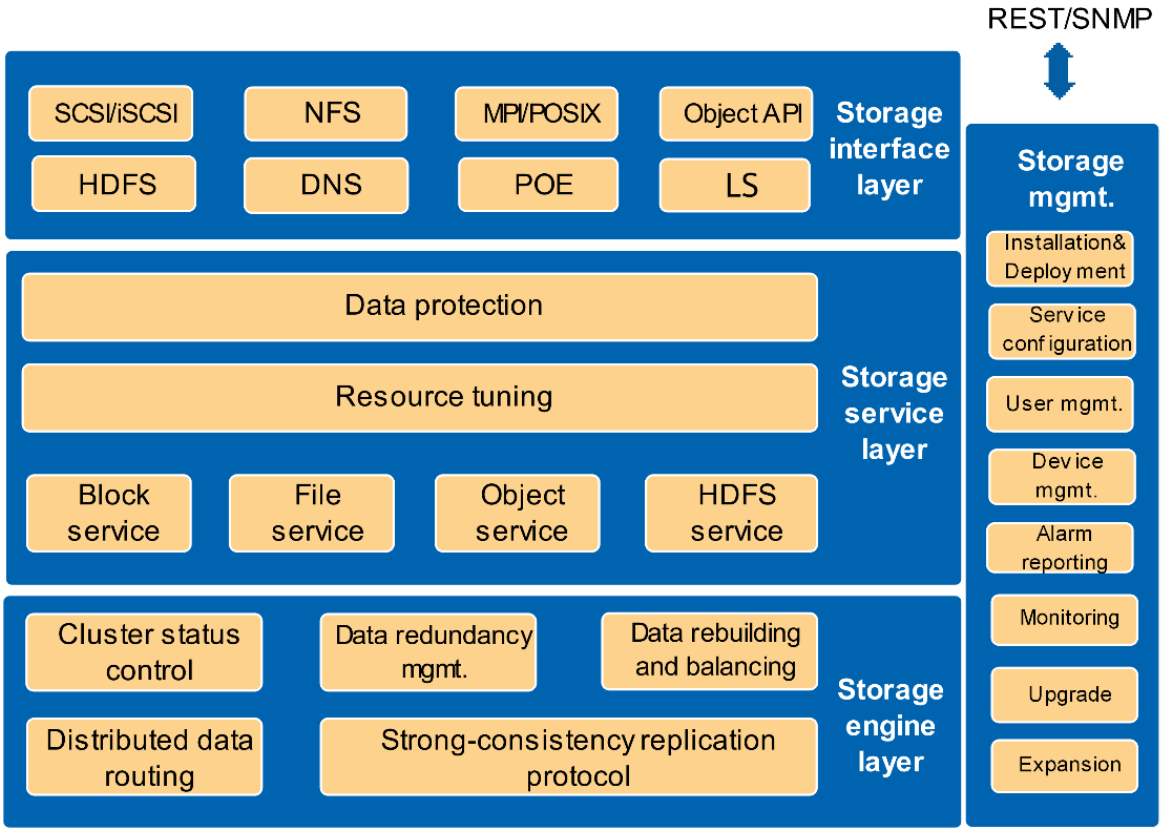
The HPDA storage solution uses Huawei OceanStor Pacific series parallel file storage to provide file storage services.

As shown in Figure 3-2, the functional architecture consists of the storage interface layer, storage service layer, storage engine layer, and storage management.

- Storage interface layer: Provides standard interfaces for applications to access the storage system and supports MPI, POSIX, NFS, CIFS, S3, and HDFS protocols.
- Storage service layer: Provides block, HDFS, object, and file services and enriched enterprise-grade features.

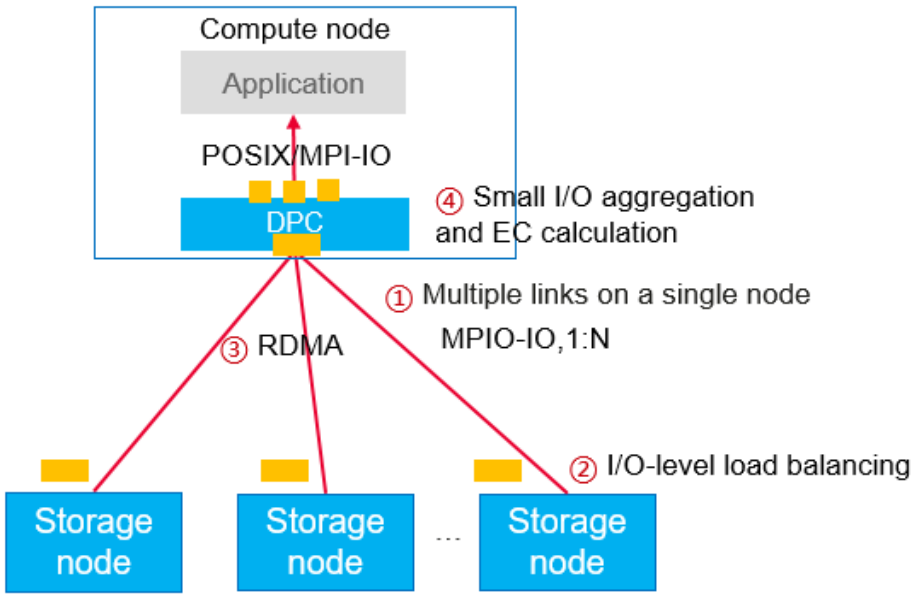
- Storage engine layer: Implements persistent storage. This layer leverages the Plog interface, an Append Only redirect-on-write (ROW) mechanism, to provide EC, data rebuilding and balancing, disk management, and data read/write capabilities.
- Storage management: Operates, manages, and maintains the system, and provides functions such as installation, deployment, service configuration, device management, monitoring, alarm reporting, upgrade, and capacity expansion.

Figure 3-2 Functional architecture of OceanStor Pacific series parallel file storage



MPI parallel computing communication application interfaces are often used in HPDA scenarios. To cope with the challenges posed by parallel I/O access to storage systems, OceanStor Pacific series file service launches the Distributed Parallel Client (DPC) to carry standard POSIX semantics and MPI-IO semantics, improving the performance of a single stream and a single client and reducing the access latency. DPC runs on a compute node and functions as a storage client. It exchanges data with back-end storage nodes through network protocols. DPC is compatible with standard POSIX and MPI-IO semantics and provides parallel interfaces and an intelligent data cache algorithm, so that upper-layer applications can access storage space more intelligently.

Figure 3-3 DPC



1. A single client is connected to multiple storage nodes. MPI-IO is supported. A single file can be concurrently processed among multiple storage nodes, improving the single-stream and single-client bandwidth.
2. I/O-level load balancing is implemented. A single client can access multiple nodes at the same time, preventing service load forwarding between storage nodes.
3. RDMA is used for communication between DPCs and storage nodes to achieve lower latency and lower CPU overhead.
4. Random small I/Os are aggregated into large I/Os. After calculating the EC redundancy on the DPC side, the large I/Os are written to the persistent node to improve data write performance.

3.1.1.2 Unstructured Data Convergence and Interworking

OceanStor Pacific series launches a technical solution that converges and interconnects three unstructured services: file, object, and HDFS. The solution has the following key capabilities:

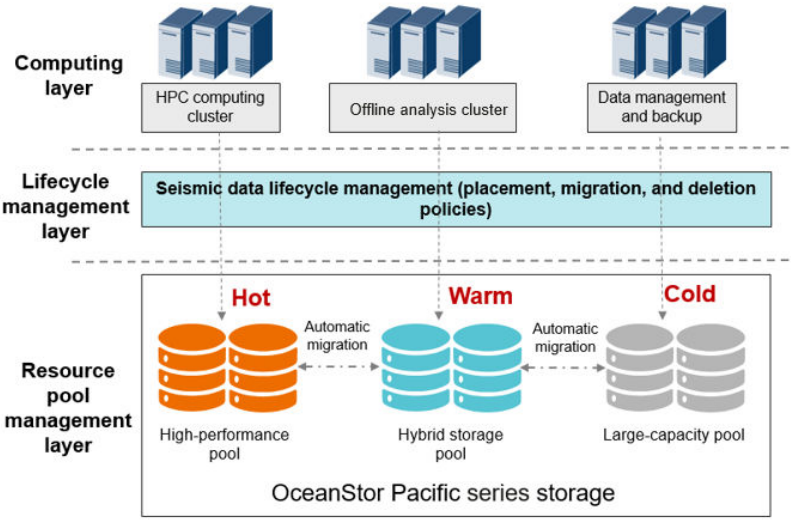
- Unified management plane: unified resource models, such as tenants/users, networks, and storage pools
- Shared storage pools: Unstructured services are provided by the same storage pool. Multiple services share the same data.
- I/O semantic interworking: full native semantics, lossless protocols, and no external plug-ins
- Value-added feature interworking: Unstructured services share and interconnect advanced features, such as QoS, tiering, and quota, without semantic loss. The scalability, performance, and serviceability are unified.

3.1.1.3 SmartTier

For HPDA services and applications, not all the data is important. Among the data, some is frequently accessed, some is seldom accessed, and some even has not been accessed for several years. A large amount of data with low value occupies high-performance and reliable system resources and large storage space, causing a waste of resources. However, the data requires long-term data storage. Automatic tiered storage technology migrates data and stores the data in the proper storage space, solving the preceding problems.

Unstructured services of OceanStor Pacific series provide the SmartTier feature and allow the system to divide different types of physical nodes in the same storage pool into different disk pools. A disk pool is a collection of nodes that have the same characteristics such as the physical type and access performance. SmartTier allows users to define the value of data in workflows based on file pool policies. High-value files are stored on storage devices of high availability and high performance, and low-value files are stored on those of low cost, low performance, and low availability.

Figure 3-4 Tiering diagram



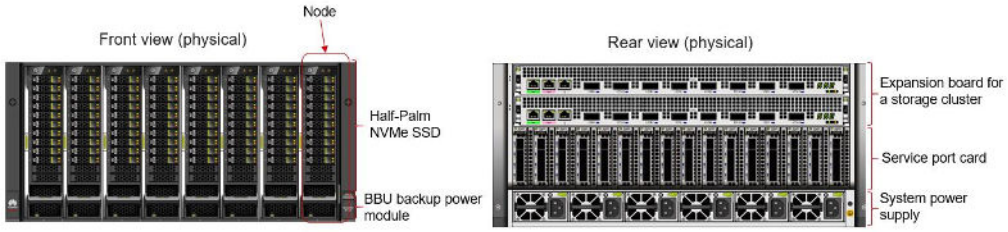
- Supports three service levels: file, object, and HDFS.
- Supports three tiers of storage pools. Users can configure proper storage media as required.
- Cross-pool data migration has no impact on upper-layer applications. The migration speed can be dynamically adjusted based on service loads. Data can be directly accessed after the migration without being migrated back.
- Multiple tiering policies are supported, including tiering based on the file name, file size, creation time, modification time, access time, status change time, and UID/GID. Users can manually plan their own services and store important directories and files in high-performance tiers based on their service characteristics.

3.1.1.4 High-Density Hardware

OceanStor Pacific 9950 is a 5 U high-density all-flash storage device. It adopts the multi-node design to deliver excellent storage performance.

- Each OceanStor Pacific 9950 device houses eight independent storage nodes and 80 half-palm NVMe SSD slots, and delivers 160 GB/s bandwidth and 32 GB/s density per U, ranking No.1 in the industry.
- The entire system adopts the PCIe 4.0 design. The industry's first all-PCIe 4.0 flash storage device doubles the bandwidth compared with the existing PCIe 3.0 system, providing higher bandwidth and lower latency.

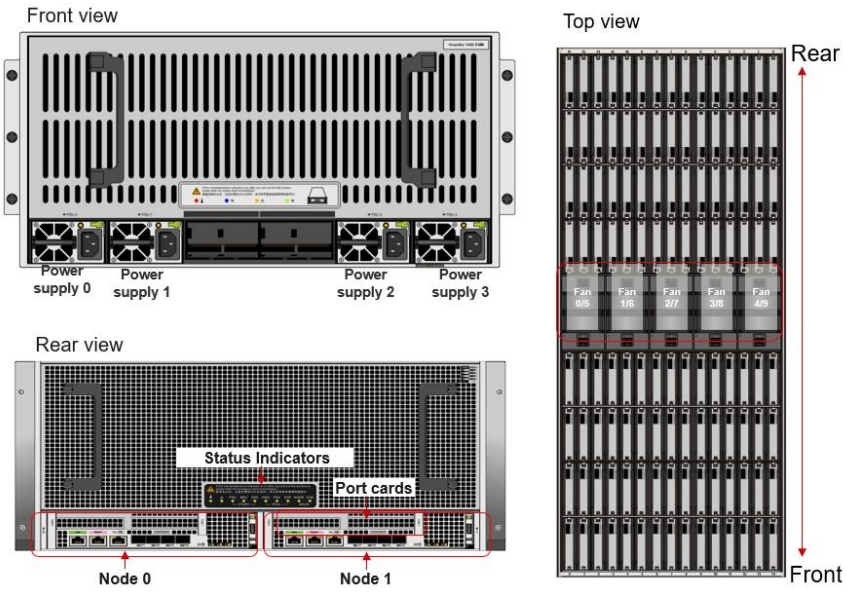
Figure 3-5 OceanStor Pacific 9950



OceanStor Pacific 9550 is a 5 U high-density and large-capacity storage device. It adopts dedicated two-node distributed hardware design to deliver superb reliability and disk density.

- Each OceanStor Pacific 9550 chassis houses two nodes and 120 disk slots, providing the highest density (24 disks/U) in the industry.
- The bi-directional drawer slide and tank chain maximize the disk density, resolves the conflict between the ultimate density and O&M space, and resolves the disk heat dissipation and online independent maintenance problems.

Figure 3-6 OceanStor Pacific 9550



3.1.2 SAN+Lustre Solution

Parallel storage is a cluster storage file system designed to address mass data storage. Among a large number of supercomputer systems across the globe, Lustre is the most widely used file system. Lustre is an object-based file system that separates file metadata from file data and stores them on different servers. File metadata is stored in a metadata server, and file data is stored in an object storage server.

A Lustre file system consists of the metadata server (MDS), metadata target (MDT), object storage server (OSS), object storage target (OST), and client. The components are described as follows:

MDS: An MDS manages metadata of a Lustre file system, including file names, directories, permissions, and file structures. It generates metadata and stores the metadata on one or more MDTs and provides services for clients. A Lustre file system can contain multiple MDSs, but only one of them functions as the active one, and the rest work in standby mode.

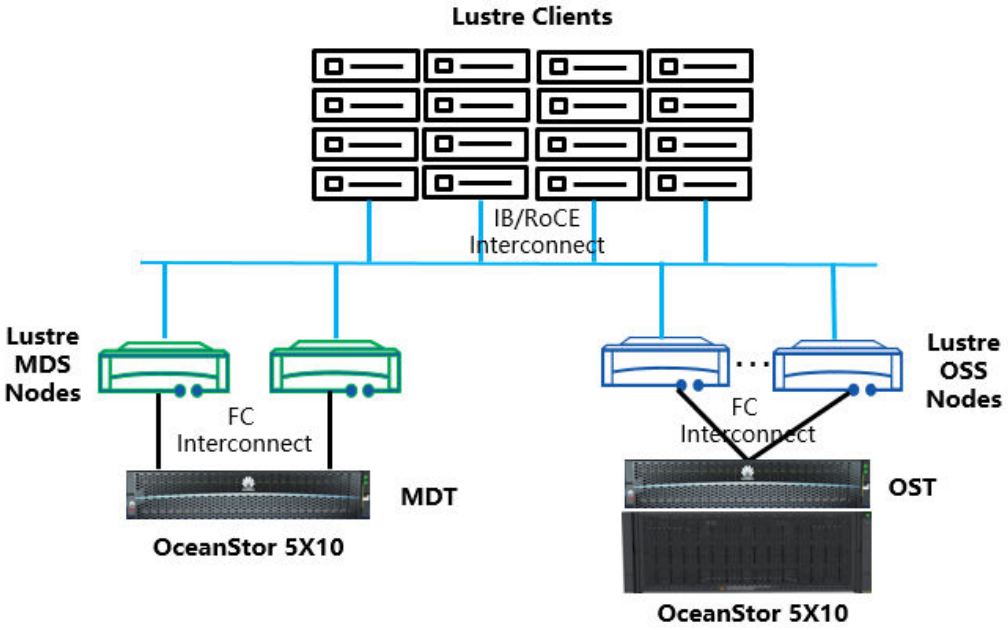
MDT: Each file system has an MDT. The MDT can be a local disk of the MDS (when there is only one MDS) or a LUN of a remote storage device. An MDT can be simultaneously mapped to two hosts to support access from multiple MDSs with only one MDS accessing at the same time, improving MDS availability.

OSS: An OSS provides the file I/O service for Lustre clients. A Lustre client obtains metadata from the MDS, accesses file data from the OSS based on the metadata, and stores the file data on the OST connected to the OSS.

OST: In a Lustre file system, user files are stored in one or more objects. Each object corresponds to one independent OST. Each file can be stored on one OST or across multiple OSTs. An OST can be mapped to two hosts simultaneously for high OSS availability.

Lustre client: A Lustre client refers to a compute, virtualization, or desktop node that is installed with Lustre client software and can be mounted with a Lustre file system. In a high-performance cluster, it is usually a compute node or a management node.

Figure 3-7 OceanStor intelligent hybrid flash storage + Lustre storage system



Highlights of OceanStor intelligent hybrid flash storage + Lustre solution:

- **Superb performance**: OceanStor intelligent hybrid flash storage provides TB/s-level bandwidth, and each cabinet provides 120 GB/s bandwidth.
- **Elastic scalability**: Supports more than 10,000 compute nodes and 512 PB scalability for a single namespace.
- **Robust reliability**: Features solution-level reliability, full redundancy of all components, and high disk fault tolerance.

3.2 Heterogeneous Computing

Compute nodes provide computing power for HPDA clusters. They largely determine the performance of an HPDA cluster. Based on their different functions, compute nodes are classified into general-purpose compute nodes, heterogeneous acceleration nodes, and fat nodes. This technical white paper focuses on storage systems and does not describe specific computing products.

3.3 Network

HPDA networks comprise the computing network, storage network, service management network, and out-of-band management network. The service management network and out-of-band management network use the universal network switches of the data center. This document focuses on storage systems. For details about the network switches used in the computing network and storage network, see the *CloudEngine 8800 and 6800 Series Switches V300R020C10 Product Documentation*.

3.4 Cluster Suite and Basic Software

The cluster suite and basic software mainly adopt Huawei-developed service software and support open-source software such as Slurm and OpenMPI. This document focuses on storage systems. For details about cluster suites and basic software, see the *Kunpeng Computing HPC Solution 20.0.2 Product Documentation*.

3.5 Scenario-Specific Solutions

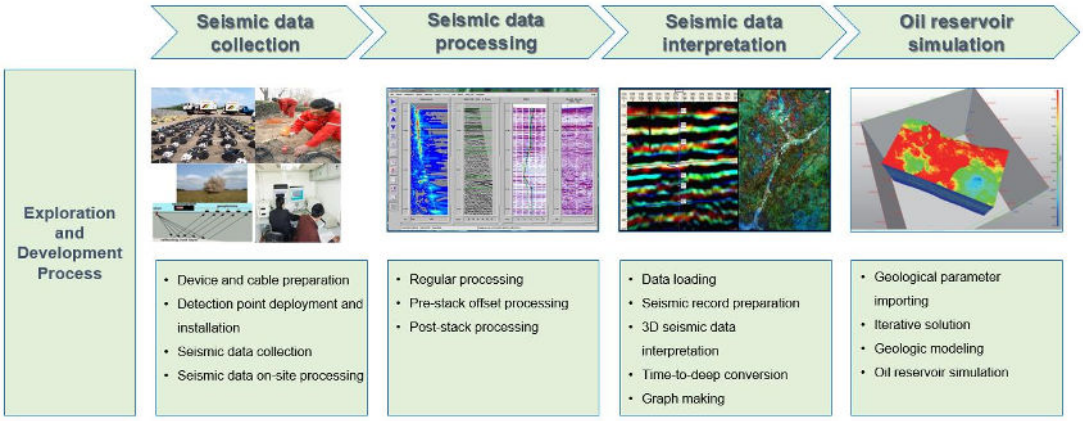
- Large-scale commercial use of HPDA: gene sequencing, oil and gas exploration, autonomous driving, education and scientific research (electron cryomicroscopy), and supercomputing center (Lustre solution)
- Restricted commercial use of HPDA: meteorological prediction, enterprise computer aided engineering (CAE), satellite remote sensing, and supercomputing center (OceanStor Pacific solution)

3.5.1 Oil & Gas Exploration

3.5.1.1 Scenario Characteristics

The HPC technology is used to process and calculate raw data acquired in field seismic surveys to transfer it into high-quality and reliable seismic information. The information in turn can provide an intuitive and reliable basis and related geological data for subsequent oil and gas exploration. Broadly, oil and gas exploration comprises seismic data acquisition, processing, and interpretation, as well as oil reservoir simulation.

Figure 3-8 Oil and gas exploration service flow



- Seismic data collection is based on geology and physics, and uses new technologies in the fields of electronics and information theory to cause the earth's crust to vibrate artificially. For example, explosives are used to generate artificial earthquakes, and precision instruments are used to record the vibration of each point on the ground after the explosion.
- Seismic data processing uses a computer to process and transform raw data acquired in field seismic surveys to transfer it into high-quality and reliable

seismic information which in turn can provide an intuitive and reliable basis and related geological information for subsequent oil and gas exploration. Field seismic data contains information about the terrain structure and lithology. However, the information is superimposed on the interference background and distorted by external factors. The information is often intertwined and cannot be directly used for seismic data interpretation. Therefore, it is necessary to process the acquired field seismic data.

- Seismic data interpretation is the process of turning processed information into geological results. The wave theory and geology knowledge, together with geology, well drilling, and well logging data, are used to make structural interpretation, stratigraphic interpretation, lithology and hydrocarbon detection, and comprehensive interpretation. Maps are drawn to evaluate oil and gas reservoirs in the survey area and provide suggestions on drilling sites.
- Oil reservoir simulation is used to study the proper development of various types of oil reservoirs. The effect of important technical decisions on oilfield development can be studied in a model. For example, sensitivity analysis can be performed to analyze the impacts of the water injection time, oil reservoir pressure, and yield on the final oil recovery rate.

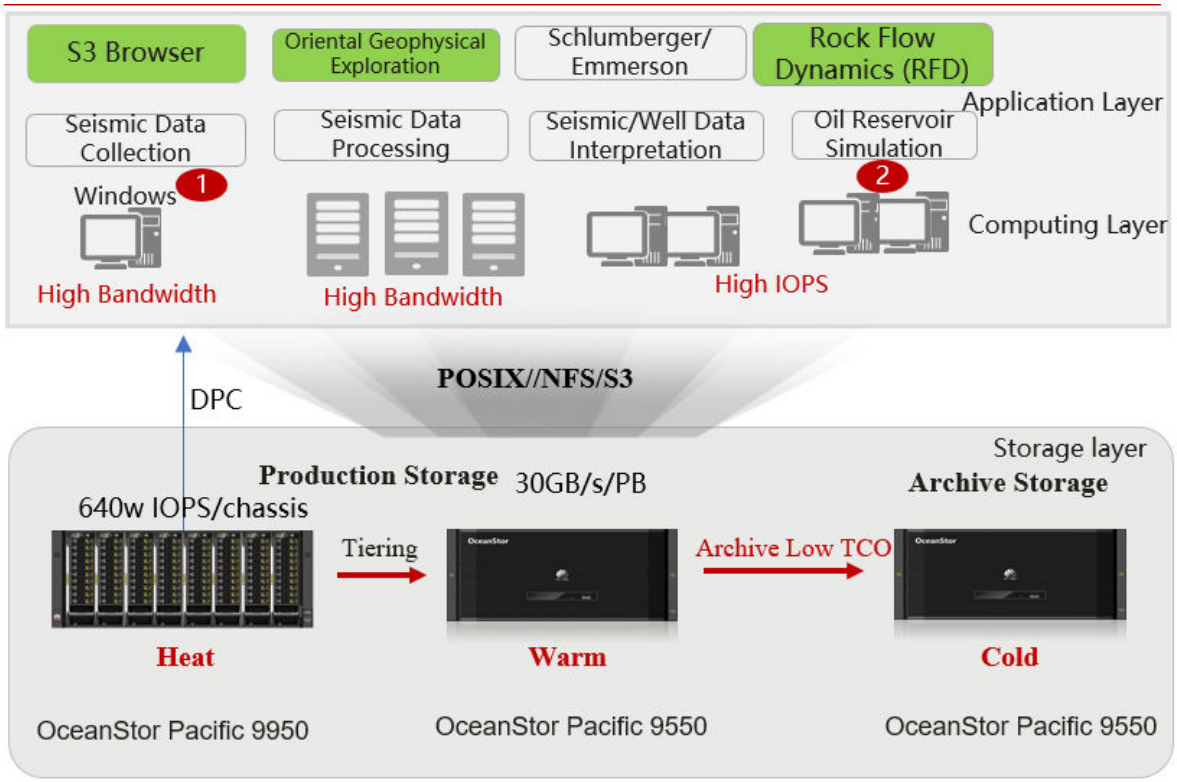
Oil and gas exploration applications have the following characteristics:

- Seismic data collection: Data is imported and exported at a high speed and needs to be portable in the field. Data must be processed on the day when it is collected. However, data replication and migration are time-consuming and may take several days.
- Seismic data processing: The data volume increases by 10 to 20 times in the entire process. The data volume can reach the 10 PB level. GB-level sequential read and write for large I/Os of large files are required. The aggregated bandwidth is generally up to 20 GB/s. These pose high requirements on storage capacity and stability and consume a long time.
- Seismic data interpretation and oil reservoir simulation: GB-level random and sequential small I/Os of large files are required, with the latency of a single I/O within 100 microseconds.

3.5.1.2 Scenario-Specific Solution

The following figure shows the solution designed to address major service challenges and pain points in oil and gas exploration scenario.

Figure 3-9 Solution architecture for the oil and gas exploration scenario



This section describes the architecture of the oil and gas exploration solution in terms of:

- Seismic data collection: S3 Browser is used to import seismic data to object storage, greatly improving the import efficiency. High-performance file interfaces are provided for seismic data processing and interpretation through protocol interworking.
- Integration of processing and interpretation: GeoEast of BGP of CNPC uses one storage system to meet the requirements of seismic data processing, interpretation, and reservoir simulation, and data migration is not needed. Full lifecycle management of hot and cold seismic data is implemented based on data tiering. In addition, ultimate bandwidth performance is available, which is several times higher than traditional solutions such as NFS.
- Reservoir simulation: The tNavigator of Rock Flow Dynamics (RFD) is used to build a reservoir stimulation and a complete oil and gas exploration workflow.

NOTE

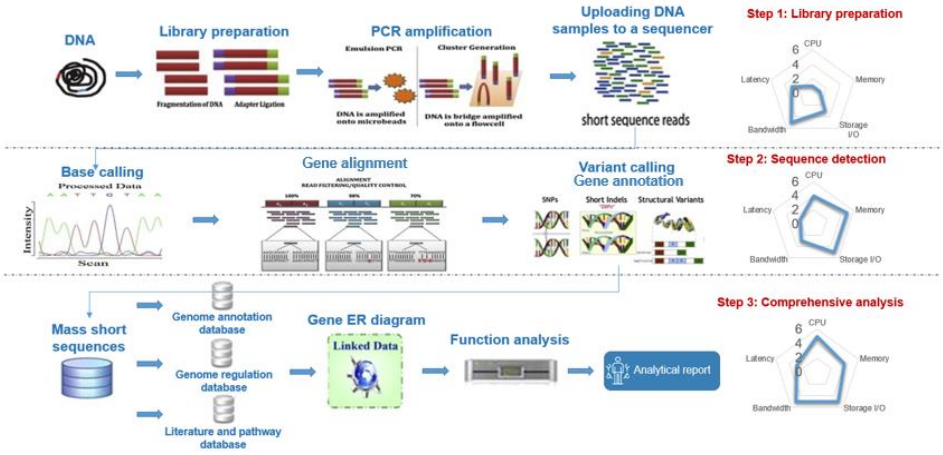
Huawei provides storage solution for oil and gas exploration solution industry. In this solution, switches are involved in GE networks on the management plane and high-performance networks (IB/RoCE/Ethernet, including 100 Gbit/s IB, 25 Gbit/s/100 Gbit/s Ethernet, and 25 Gbit/s/100 Gbit/s RoCE) connecting to the computing cluster (client).

3.5.2 Gene Sequencing

3.5.2.1 Scenario Characteristics

The human genome contains 3 billion base pairs. A genome-wide profiling generates more than 100 GB data. Only protein-coding genes and a small number of non-protein-coding genes can be used in scientific research and clinical applications. These genes account for 1% to 2% of the total data. However, only 1% of the such data is fully researched and applied after being associated with diseases. In addition, the cost of obtaining gene data is high. In the future, with the development of technologies, more and more encoding genes will help humans. Therefore, it is important to further explore the future. All genome test data cannot be easily discarded. A large amount of data is generated during gene sequencing. How to use new technologies to store massive data and quickly extract key information from massive amounts of data has become a great challenge. Currently, the breakthrough and innovation of the second-generation or next generation sequencing (NGS) technology makes gene sequencing fast, low-cost, and high-throughput, which helps popularize the application and makes consumers affordable. In the term of commercialization, sequencing is a tool whose market space of the clinical application domain is huge. In the future, the rapid growth and continuous prosperity of the gene sequencing industry will be an irresistible trend. The following figure shows the process of the second-generation sequencing technology.

Figure 3-10 Gene sequencing process



Gene sequencing poses the following challenges to storage systems:

- Mass data storage: A sequencer generates approximately 6 TB raw data every day. One person's gene data is approximately 1.5 TB and needs to be stored permanently, thus requiring hundreds of PBs in space scalability and low-cost storage capability.
- Converged data processing: Traditional statistical methods, emerging big data analytics, and object interface requirements for data distribution require storage systems to provide the converged data processing capability.

In addition, the requirements on CPU, memory, storage I/O, bandwidth, and latency vary in different phases, especially in the sequence analysis phase.

- There are more than 1000 types of sequence analysis software. Different types of software are selected for different service directions. The resource consumption of software can be classified into CPU consumption, memory consumption, storage I/O consumption, and comprehensive consumption.
- Local resources have limited specifications and cannot meet all service requirements. As a result, resources often remain idle or are inadequate, for example, inadequate storage space, unsatisfactory storage performance, consistent full-load operation of CPUs, and periodic idleness of memory.

Table 3-1 Key requirements of gene sequencing

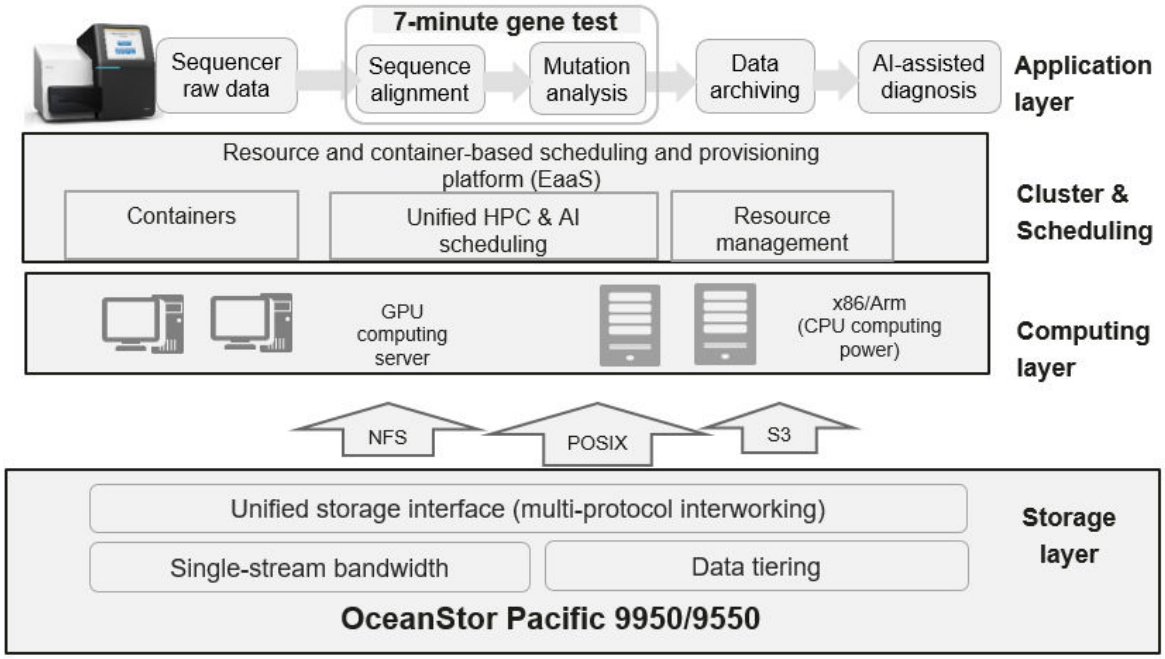
Example Sequencing of a Marine Life	Data Volume	Operation	I/O Model	Performance Requirement
1. Data collection	(MB to GB) x Total number of samples	The sequencer transmits data to the storage device over CIFS, FTP, or HTTP.	Write of GB to TB level large files	Low
2. Format conversion	GB to TB level	Convert the *XXX* (such as BCL and Bam) format to the FastQ format.	Read and write of GB to TB level large files	Low
3. Genome assembly	GB to TB level	Read and write a large number of temporary files.	Single job: maximum read bandwidth 340 MB/s and maximum write bandwidth 220 MB/s	500 MB/s
4. Gene alignment	GB to TB level	Append data to process files, read raw files in sequence, read reference files randomly, and write temporary small files in sequence.	For a single job, the maximum read bandwidth is 1.12 GB/s, and the maximum write bandwidth is 60 MB/s.	1.12 GB/s Note: According to the actual test, the maximum read bandwidth reaches 1.12 GB/s. Therefore, the recommended values are for reference only. The specific performance requirements need to be designed based on actual situations.

Example Sequencing of a Marine Life	Data Volume	Operation	I/O Model	Performance Requirement
5. Gene annotation	GB to TB level	Read input files in sequence, read reference files randomly, and write result files in sequence.	For a single job, the maximum read bandwidth is 1.12 GB/s, and the maximum write bandwidth is 60 MB/s.	1.12 GB/s
6. Data archiving or distribution	GB to TB level	Distribute result data over CIFS or FTP.	Read of GB to TB level large files	Low

3.5.2.2 Scenario-Specific Solution

Figure 3-11 shows the architecture of the solution that uses OceanStor Pacific series parallel file storage (distributed storage) as its core.

Figure 3-11 Solution architecture for the gene sequencing scenario



This section describes the architecture of the gene sequencing solution in terms of:

- Cluster & scheduling: provides functions such as cluster user management, resource management, job scheduling, and application image management by cooperating with partners (such as ClusterTech EaaS).
- Computing layer: based on Huawei TaiShan servers or provided by a third party.

- Network layer: based on Huawei Ethernet/RoCE switches or provided by a third party.
- Storage layer: stores and processes gene data. OceanStor Pacific is used for storage, and Atempo is used for archiving.

NOTE

- When delivering the HPDA Storage Solution, ensure that the computing performance is not a bottleneck. However, services are directly running at the computing layer. Under the premise of promoting Chinese-made servers in the genome sequencing scenario, it is recommended that Arm be used to evaluate the specifications of compute nodes. Other servers are subject to the specific project.
- In this solution, switches are involved in GE networks on the management plane and high-performance networks (IB/RoCE/Ethernet, including 100 Gbit/s IB, 25 Gbit/s/100 Gbit/s RoCE, and 25 Gbit/s/100 Gbit/s Ethernet) connecting to the computing cluster (client).
- The IB/RoCE network is required when services access the storage system through the DPC (POSIX). The Ethernet is required when services access the storage system through the NFS and S3 protocols.

3.5.3 Education and Scientific Research (Electron Cryomicroscopy)

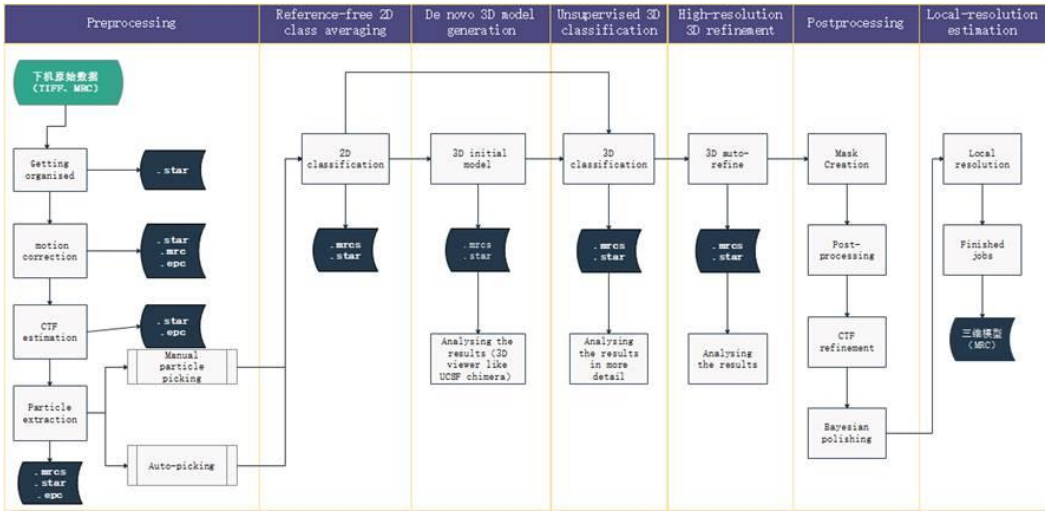
3.5.3.1 Scenario Characteristics

Cryogenic electron microscopy, or cryo-EM, is a technique used to study the forms of transmissive electron microscopy (TEM) samples at ultra-low temperature (usually liquid nitrogen temperature -196°C).

TEMs are often used to record high-resolution images up to thousands of identical but randomly distributed particles (molecules) from each sample. The images are then grouped, aligned, and balanced using an image classification algorithm to distinguish multiple directions of the 3D molecules.

Using high quality samples, electron cryomicroscopy can resolve molecular structures with a resolution close to 2 Å. Cryo-EM is an important tool for high-level research. This technique has been widely adopted by life scientists and has become a benchmark for high-performance computing.

Figure 3-12 Data processing steps of cryo-EM



- Pre-processing

  Cryo-EM irradiates a specific position of a sample for multiple times to obtain an image with sufficient illuminance. Generally, the image is in the TIFF or MRC format. The preprocessing performs drift correction, CTF estimation, and particle selection on the raw data.
- Drift correction: During the irradiation process, any drift of a sample affects the stacking effect of images. Therefore, the impact of motion must be eliminated before stacking.
- Contrast conversion function (CTF) processing: determines the defocus parameter of a contrast conversion function for explaining the luminance of a sample in an image.
- Particle selection: extracts particles that meet subsequent process quality requirements from images from different angles to form a star file (location description file, usually several MB) and a stack file (particle image data file, usually hundreds of GB). Particles can be selected manually or automatically.
- 2D classification

  Particle clustering and alignment: classifies, rotates, and aligns similar particles.
- 3D reconstruction and refinement

  Data transformation and 3D reconstruction: 2D Fourier transform, Fourier space construction, and 3D inverse Fourier transform are used to generate a 3D model. Based on the model, classification, alignment, transformation, and reconstruction are repeated to continuously optimize the 3D model.
- Post-processing
- Local analysis result

At different phases in cryo-EM, data characteristics are different and require varying levels of HPC storage resources.

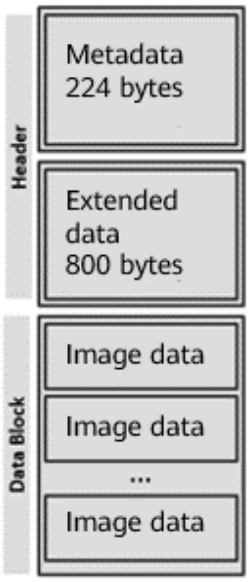
Cyro-EM Phase	Data Volume	Operation	I/O Model	Performance Requirement
1. Data collection	10 TB to 30 TB in each experiment	Upload the images taken by cryo-EM to the storage.	Sequential write for large files	10 TB to 30 TB/18 hours 300 MB/s to 500 MB/s
2. Image processing	Hundreds of GB	Compress a 6 GB to 7 GB image to 50 MB to 60 MB.	Sequential read and write for small I/Os of large files	Aggregate bandwidth: 120 GB/s (depending on project requirements) Single-client I/O: 2.6 GB/s
3. 3D reconstruction	Single-project images Median: 2000 images (50 MB per image)	Convert a 2D image to a 3D image.	Random read is used by default. Small I/Os account for 75%, and medium I/Os account for 25%.	Aggregate bandwidth: 120 GB/s (depending on project requirements) Single-client I/O: 2.6 GB/s

Cyro-EM Phase	Data Volume	Operation	I/O Model	Performance Requirement
4. Download and analysis	50 TB to 100 TB	Download and analyze the result file.	Few storage I/O interactions	Few storage I/O interactions

- Each job generates 2000 to 3000 MRCS or TIFF files that contain raw sequence data, occupying several TB to dozens of TB space and requiring long-term storage.

  NOTE

  - MRCS or MRC is the main file format of the cryo-EM system. It is a container for storing image data (including raw, intermediate, and result data) and metadata.

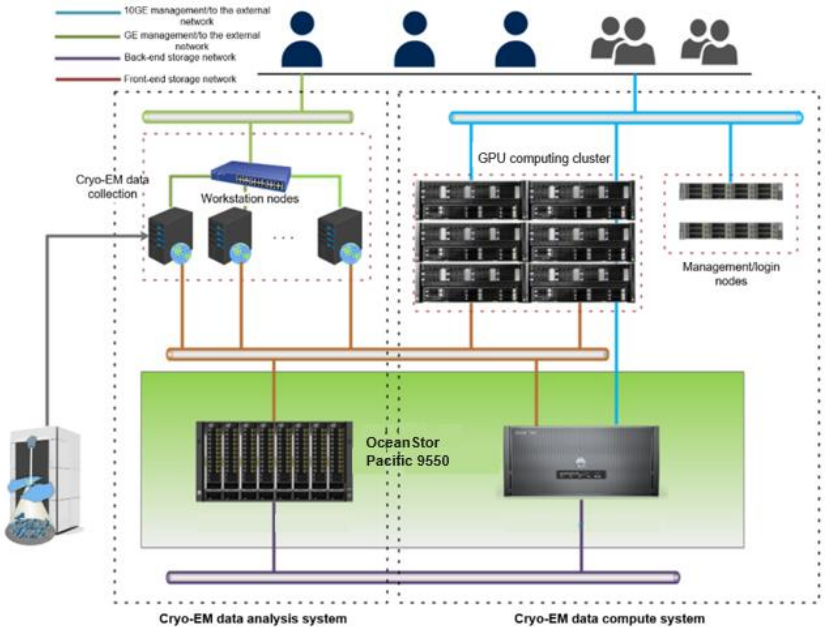


  - The metadata describes the dimensions and data format of images, as well as the description information related to the job.
  - The extended field is used to carry the private extension of a system provider or software provider.
  - The image data includes 2D or 3D data.
- To save storage resources, cryo-EM vendors compress image data stored in the files using the LZW algorithm or using GZIP. The amount of information in the image data is very valuable, and the compression processing must be binary lossless.
- Massive image data generated by cryo-EM is used only in the 3D modeling process. However, because sample preparation takes a long time, raw data and secondary images generated by using the gain correction technique are stored for a long time. Currently, the cryo-EM data processing platform does not have the archiving capability. Typically, historical data is manually migrated.
- Data processing software needs to load large image files (hundreds of GB) containing mass particles. The throughput of the existing system is only GB-level, which causes a bottleneck.

3.5.3.2 Scenario-Specific Solution

The HPDA storage solution designed for the education and scientific research (electron cryomicroscopy) scenario uses OceanStor Pacific series parallel file storage (distributed storage) as its core and works with compute, network, and Relion software. Figure 3-13 shows the solution architecture.

Figure 3-13 Solution architecture for the education and scientific research (electron cryomicroscopy) scenario



In the education and scientific research (electron cryomicroscopy) project, two storage levels are usually constructed: primary storage and secondary storage.

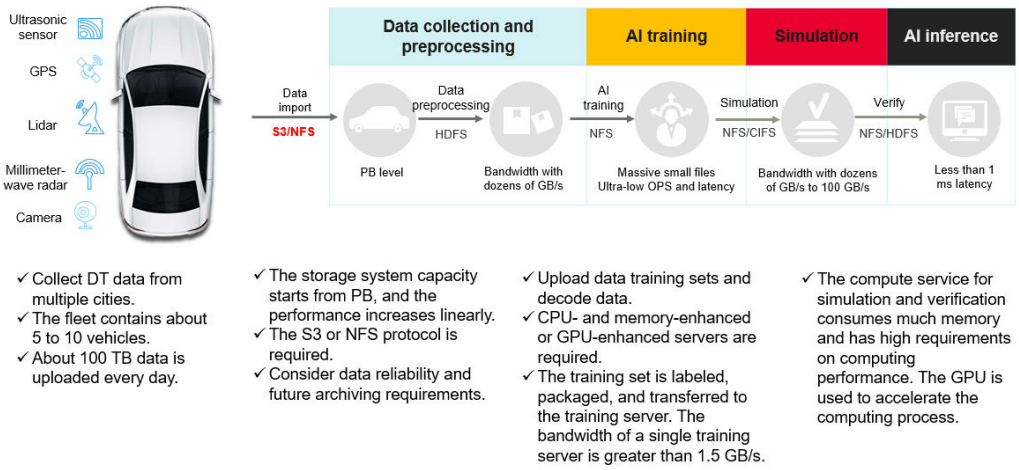
- The primary storage uses SSDs to improve I/O bandwidth and receive data from cryo-EM. Its data volume is large and continuous reads and writes have stringent requirements on storage performance.
- The secondary storage is used for cluster computing and archiving of data from electron cryomicroscopy. The NFS protocol is used for processing and analyzing the data.

3.5.4 Autonomous Driving

3.5.4.1 Scenario Characteristics

An autonomous driving car is an unmanned ground vehicle for transmitting power. It can sense the environment and navigate without human intervention. In the next three years, ADAS-based autonomous driving will evolve from L2 to L4, and the number of road data collection sensors will increase from 10 to 20, improving the detecting precision. The road data collection dataset will increase by 10 times (from 50 PB to 500 PB). The data volume increases by an order of magnitude, but the version iterations remain unchanged or even shorter.

Figure 3-14 Autonomous driving service flow



Data collection: The road test automobile collects data in real time on the real road using multiple sensors, such as the camera, radar, ultrasonic wave, lidar, and GPS. The data is stored on the local hard disk. The removable disk storing the collected data is delivered to the data processing site of the automobile manufacturer on the same day.

Data import: indicates the process of loading collected data into an autonomous driving development data platform. Every night, an operator inserts the removable hard disk into a dedicated data import server and uploads the data to the storage pool in batches.

Data pre-processing: After data is uploaded to a large-capacity storage pool, it needs to be pre-processed in multiple phases before being used in machine learning, deep learning, software simulation, and hardware simulation.

AI model and algorithm training: main data consumption scenarios, including machine learning in AI scenarios and neural network deep learning. The autonomous driving algorithm model development and training (MiL), such as **perception algorithm, convergence algorithm, and decision-making algorithm**. 30% of the data volume is used as the training set.

Software and hardware simulation: Before large-scale commercial use of autonomous vehicles, functional security and performance security tests must be conducted to ensure the safety of consumers and the public. The tests include software-in-the-Loop (SIL) simulation and hardware-in-the-loop (HIL) simulation.

Dataset and result data analysis: This step is implemented based on the Hadoop big data platform and requires comprehensive analysis of massive data, including simulation results and DT data. The data to be analyzed is imported to the big data platform. If the data platform used for simulation supports HDFS big data interface access, massive data synchronization is not required.

Customer pain points and challenges in autonomous driving include:

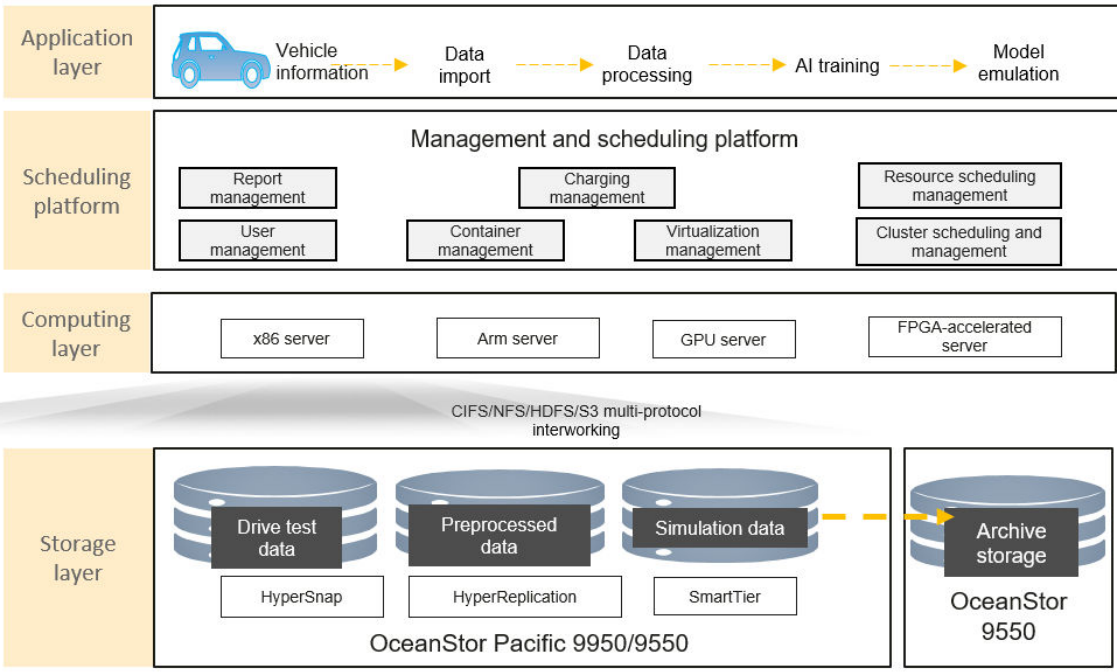
- **Massive amounts of road data generated due to the AD upgrade**: Compared with L3, the L4 road collection data volume increases by three to five times. Hundreds of TB data needs to be imported and stored every day. The amount of data upon pre-processing is about 15% of that of collected data, and PB-level data needs to be archived for more than 10 years.

- **One service involving multiple protocols:** Test data transmission, import, preprocessing, training, simulation, and result analysis require different protocols (object/NAS/HDFS). Data is siloed to a high degree and data copy lasts for a longer time than processing and analysis, reducing efficiency.
- **Complex service models and ultimate performance requirement:** There are various types of road test data sensors and complex service I/O models. Therefore, ultimate performance is required in vehicle model algorithm training and function simulation phases.

3.5.4.2 Scenario-Specific Solution

The HPDA storage solution designed for autonomous driving uses OceanStor Pacific series parallel file storage (distributed storage) as its core and works with compute, network, as well as training and simulation software. Figure 3-15 shows the solution architecture.

Figure 3-15 Solution architecture for the autonomous driving scenario



Highlights:

- Unified storage base and multi-protocol interworking: The data preprocessing platform directly reads data. OceanStor distributed storage provides native HDFS, object, and file interfaces. Interworking from the S3/NFS protocol to the HDFS/NFS protocol is supported. A unified hardware platform is used, and data does not need to be copied, reducing data migration.
- Hot and cold data tiering: Processed scenario databases and unstructured data are stored to the OceanStor Pacific series storage system by tier based on the cold and hot data tiering policies, reducing the costs of data storage and maintenance.
- Ultimate performance: The training server uses the NFS protocol to read training dataset from the OceanStor Pacific series storage system. The training sets are

packed into GB-level large files in advance. For large-file reads and writes, a bandwidth of hundreds of GB/s is provided.

3.5.5 Supercomputing Center

3.5.5.1 Scenario Characteristics

The HPDA Storage Solution serves the following national supercomputing centers and university-level supercomputing platforms:

- China: 8 national supercomputing centers (in Wuxi, Tianjin, Jinan, Shenzhen, Changsha, Guangzhou, Zhengzhou, and Kunshan), 11 regional supercomputing centers (Harbin, Xi'an, Chengdu, Chongqing, Wuhan, Wenchang, Shanghai, Zhejiang, Inner Mongolia, Fuzhou, and Qingdao), and the supercomputing platforms of key universities (such as Tsinghua University, Nanjing Agricultural University, and Huazhong University of Science and Technology).
- Outside China: National supercomputing centers and university-level supercomputing platforms, such as the national supercomputing center of the Czech Republic and the supercomputing platform of Tokyo University.

A supercomputing center or platform:

- Provides resources and tools to support diversified services and customers at the application layer.
- Mainly carries services with large loads, such as those from scientific research, government, and manufacturing, and also supports small-load services. The supercomputing center or platform must support flexible resource scheduling according to the workload to reduce power consumption and improve resource utilization.

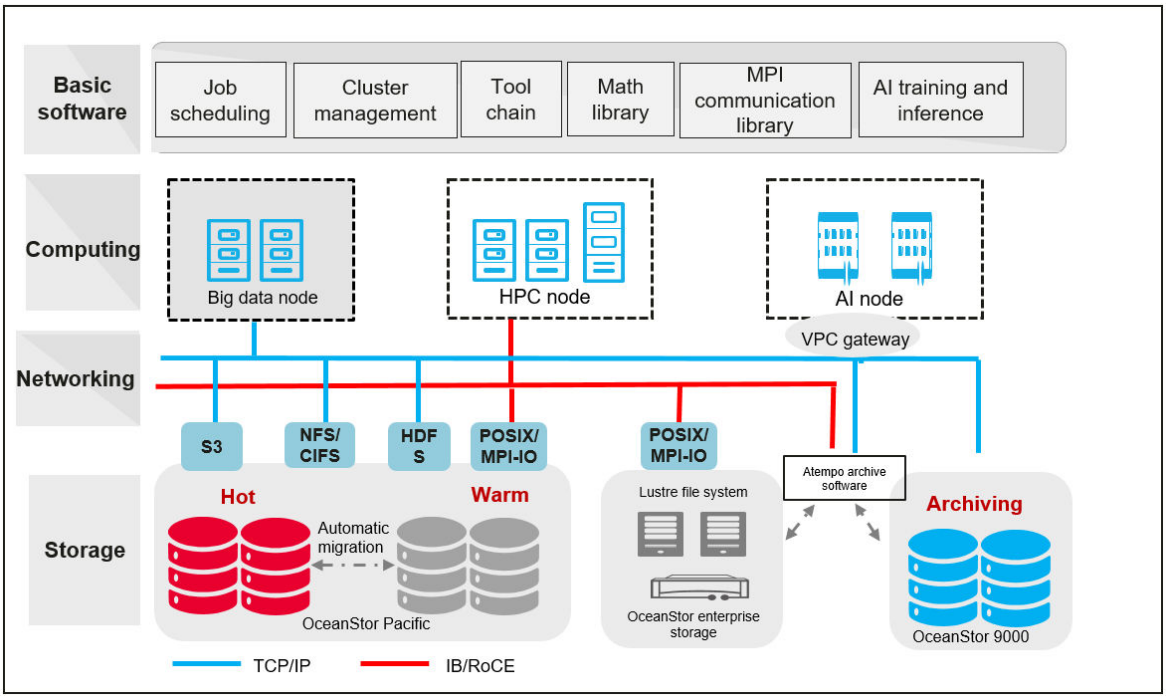
The challenges facing the supercomputing center scenario:

- Complex service model: A supercomputing center receives access requests from different service systems whose I/O models vary greatly. Therefore, a storage system that supports multiple service models, such as high bandwidth, high IOPS, and high OPS is needed. The storage system must provide more efficient data access for HPC, big data analytics, and AI training services, and currently this is impossible with conventional systems.
- Resource isolation: One storage system carries different service systems. Resources of different service systems must be logically isolated to prevent them from affecting each other's performance.

3.5.5.2 Scenario-Specific Solution

A supercomputing center is a platform solution that applies to oil and gas exploration, meteorology and oceanography, and industrial manufacturing scenarios. Figure 3-16 shows the architecture of the solution for the supercomputing center.

Figure 3-16 Solution architecture for the supercomputing center scenario



The supercomputing center solution has the following functions and features:

- The self-developed OceanStor Pacific parallel file system is promoted to provide high-performance storage under hybrid loads. It meets the performance requirements of HPC and AI platforms, ranking top 1 among commercial parallel file systems on the IO500 list.
- Service data in HPC and AI supports multi-protocol interworking. A unified data foundation is provided, achieving zero data migration and improving service efficiency.
- The HPC platform supports the Lustre solution. Based on the high-performance and high-reliability SAN service of Huawei OceanStor enterprise storage, a scalable Lustre file system is constructed.
- The Lustre file system can be archived to the Huawei OceanStor 9000 distributed file storage system.

NOTE

In the supercomputing center scenario, this solution mainly provides the storage service. The computing and network resources are provided by Intelligent Computing and Data Communication or third parties. The Lustre software is provided by partners.

The Atempo Miria software (resold) is used for archiving. Archive servers are provided by a third party and configured as follows:

3.5.6 Weather Forecast

3.5.6.1 Scenario Characteristics

The weather prediction scenario involves four types of services: weather forecast, environment monitoring, marine prediction, and climate prediction.

Category	Service	Common Model
Weather forecast	Forecasts the weather in an area in the next few days.	MM5, WRF, and GRAPES
Environment monitoring	Predicts air quality in an area in the next few days.	CMAQ, WRF-CHEM, and CAUSE
Marine prediction	Predicts changes in currents and water temperature in a sea or river area.	FVCOM, ROMS, and HYCOM
Climatic prediction	Predicts the statistical features of weather elements after 10 days and within 2 years.	CESM and CCSM

Numerical weather prediction describes the law of atmospheric motions by solving hydrodynamics and thermodynamics differential equations. Specifically, numerical weather prediction predicts the atmospheric motions and weather phenomena in a future period of time based on the spatio-temporal analysis of seven variables. These are the highest temperature, lowest temperature, precipitation, humidity, atmospheric pressure, wind direction, and wind speed.

Massive and real-time computing of numerical weather prediction requires HPC.

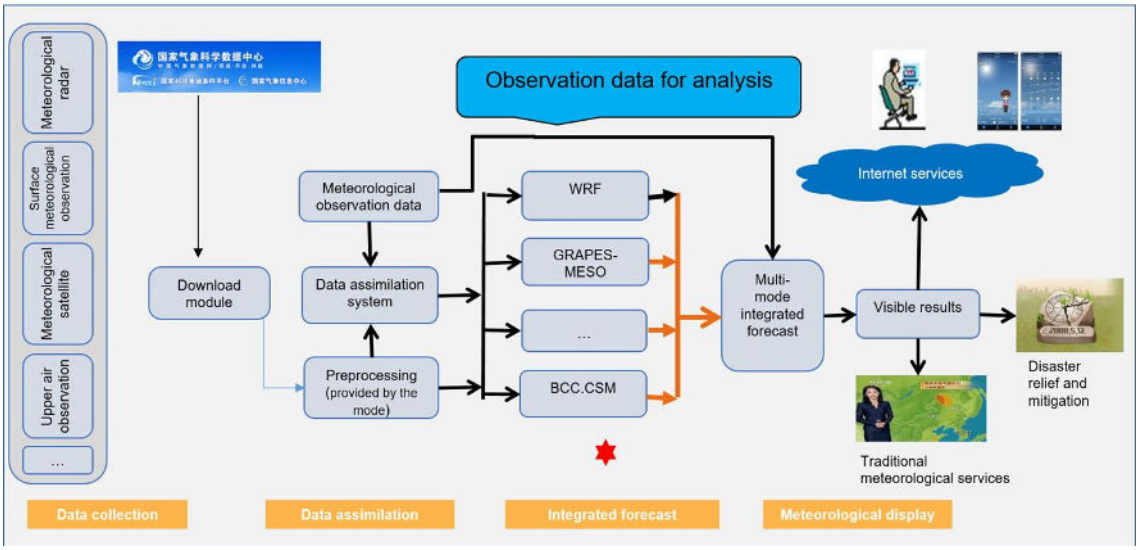
A numerical weather prediction system comprises a series of subsystems for observation data acquisition, data quality control, objective analysis, prediction mode, post-processing, and interpretation application, covering the E2E service flow.

NOTE

This document applies to WRF scenarios. If other software is used, design and deliver solutions based on project requirements.

Figure 3-17 shows the data flow of numerical weather forecast using WRF.

Figure 3-17 Numerical weather forecast process



- In data collection, raw data is collected through the satellite remote sensing system, hot air balloon detection system, radar monitoring system, and observation station system.
- In data assimilation, the weather evolution process is described as hydrodynamics and thermodynamics equations. New observation data is fitted in the process of solving the scheme group dynamically, and the assimilated data of velocity, temperature, humidity, atmospheric pressure, and air density is obtained, improving the prediction accuracy of the model.
- In predictive analytics, application systems based on software such as GRAPES, CUACE, and WRF use parallel computing and high-performance storage to calculate and analyze the numerical equations and provide the model prediction result.
- In meteorological display, weather forecast products are provided. Required meteorological elements (such as rain and snow forecast) are extracted from the calculation results, and the results are displayed on the Internet and mobile terminals in the form of graphics, animations, and web pages. Meteorological data extension services are provided for industries such as aviation, environmental protection, oceanography, electric power, and agriculture.

Data characteristics vary at different phases in numerical weather prediction, and therefore require different levels of HPDA storage resources.

Service Scenario	Data Volume	I/O Model	Bandwidth/Latency	Data Retention Period
Data collection	100 TB/day (closely related to the size of involved areas and	Sequential write for large I/Os of large files	Aggregated bandwidth: 20 GB/s	Raw data retained for three months

Service Scenario	Data Volume	I/O Model	Bandwidth/Latency	Data Retention Period
	resolution)			
Data assimilation	700 GB structured data/four times a day	Hybrid file model: sequential read for large I/Os of large files and massive small files	-	Process data retained for three months
Integrated forecast	About 3 TB data/day, mainly structured data	Sequential read/write for large I/Os of large files	Total aggregated bandwidth: 200 GB/s Microsecond-level latency High IOPS	Process data retained for three months
Meteorological display	9 TB to 15 TB data/day	Random read/write for small I/Os of large files	Aggregated bandwidth: 5 GB/s 100,000 IOPS	Process data retained for three months

NOTE

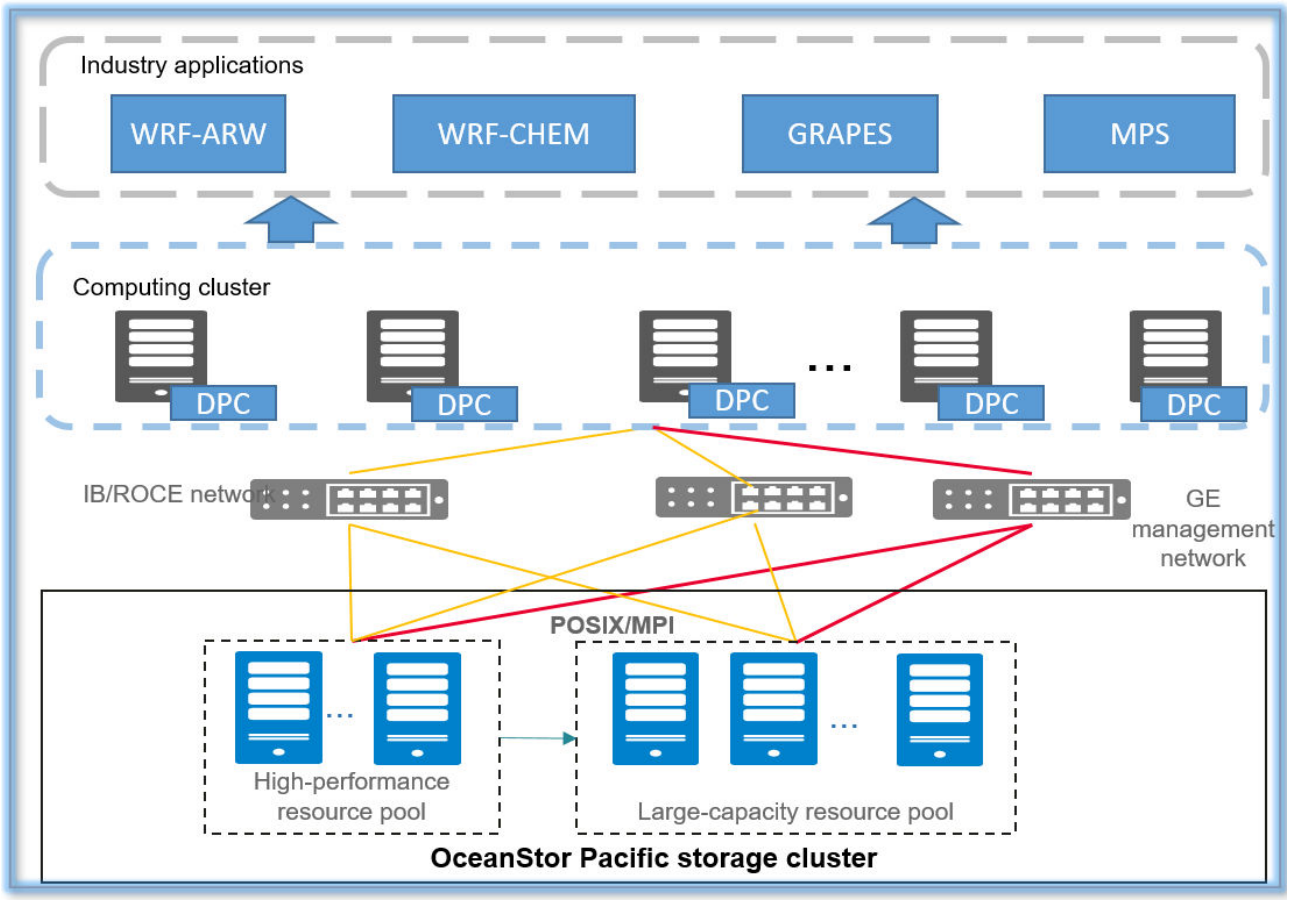
The data in the preceding table is for reference only.

3.5.6.2 Scenario-Specific Solution

For the weather forecast scenario, Huawei cooperates with ISVs in the industry to provide the following distributed storage solution: ISV application + Huawei-developed Arm server + distributed storage device. In addition, the solution supports smooth transition to HPDA.

Figure 3-18 shows the architecture of the solution that uses OceanStor Pacific series parallel file storage (distributed storage) as its core.

Figure 3-18 Solution architecture for the weather forecast scenario



Highlights:

- One storage system supports high bandwidth, high IOPS, and MPI parallel I/Os, meeting complex performance requirements.
- High-density large-capacity hardware and automatic tiering reduce the overall data storage costs.
- E2E DIF verification enables services to be quickly taken over within 10 seconds upon a node fault, ensuring always-on services.

NOTE

This solution is compatible with application software such as WRF-ARW, WRF-CHEM, WRFDA, COAWST, GRAPES, WPS and CMAQ. For details, see the *HPDA Solution ISV Compatibility List*.

Recommended storage solution configuration:

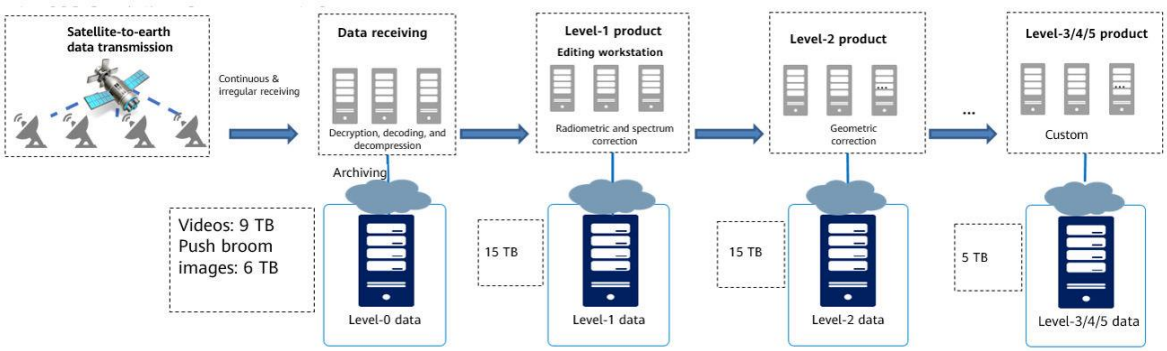
- If 6 PB capacity is required, use 20% of the SSDs to achieve high performance of hot data. The bandwidth must be higher than 100 GB/s. DPC clients support MPI-IO.
- Use 80% of the high-density HDDs for cold data storage, reducing the overall cost.
- The front-end and back-end networks are separated, and the front-end network is an IB high-speed network.

3.5.7 Satellite Remote Sensing

3.5.7.1 Scenario Characteristics

Figure 3-19 shows the basic service flow of satellite data processing:

Figure 3-19 Basic service flow of satellite data processing



- Satellite data generation: Different satellites have different functions. The collected data consists of two parts: video and push broom images. The acquired data is compressed first, and then stored on satellites.
- Satellite data collection: Continuous satellite data is generated irregularly. The amount of data increased a day is about 15 TB (9 TB video and 6 TB images). After transcoding and quick-look processing, and being added with some supplementary information, the received data becomes the level-0 data.
- Level-1 processing: The level-1 data is generated after radiometric and spectrum correction of the level-0 data. The planned data volume is 15 TB per day.
- Level-2 processing: The level-2 product data is generated after the control point and geometric rough correction of the level-1 data. The planned data volume is 15 TB per day.
- Level-3, level-4, and level-5 product data: The data is customized based on industry applications and no regular processing is required. 5 TB data is reserved every day.

NOTE

Data volume per day: 15 TB x 3 + 5 TB = 50 TB. (Level-0: 15 TB; level-1/2: 15 TB each; level-3/4/5: 5 TB reserved in total) Product data at each level must be archived in online mode for 3 months and in offline mode for 1 to 10 years.

Data in a satellite remote sensing scenario has the following features:

- Large amount of satellite remote sensing data

  From launch and throughout its lifecycle, a satellite generates hundreds of GB to TB data per day. Therefore, several PB to even dozens of PB of storage capacity is required for long-term or permanent storage of satellite image data files.
- HPC requirements

  Back-end satellite data analysis and processing consist of data transmission management, data aggregation and processing, service scheduling management, product distribution, and load balancing systems. Many service systems, almost all of which are core real-time systems, have a large amount of

data to be processed and have stringent requirements on processing performance. Large files (100 GB-level) and small files (KB-level) coexist, requiring multiple GB/s to even dozens of GB/s of throughput and high OPS. At the same time, raw data must not be lost.

- High-bandwidth and high-speed transmission network

  After the satellite passes by, it scans and collects data in real time and sends the data to the ground station through multiple parallel channels. The data rate at each channel reaches hundreds of MB/s-level and increases continuously. Different types of files with sizes ranging from KB-level to 100 MB-level coexist. Therefore, the storage system must provide stable response bandwidth. Otherwise, data loss may occur.

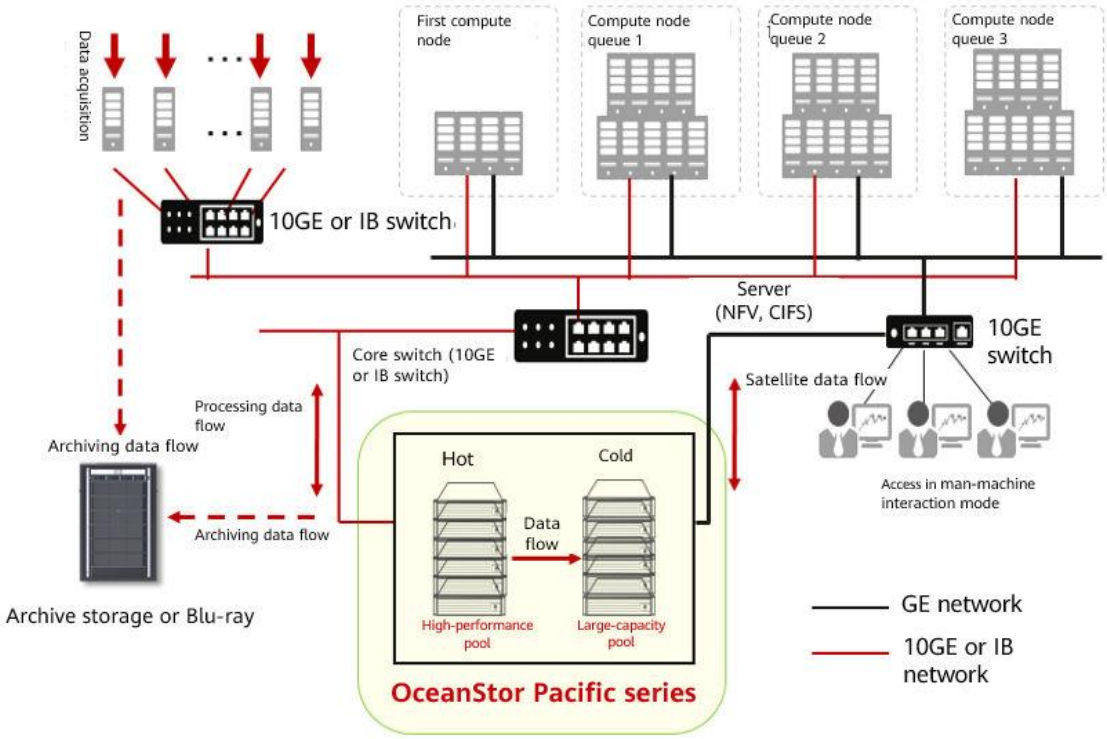
- Heterogeneous sharing and nearline and offline archiving

  Satellite receivers, including Windows industrial computers and Linux/UNIX hosts, are typical NAS storage systems. Intelligent tiering of storage, analysis, and archiving requires data analysis interfaces and intelligent archiving policies to eliminate data migration costs.

3.5.7.2 Scenario-Specific Solution

The HPDA storage solution designed for the satellite remote sensing scenario uses OceanStor Pacific series parallel file storage (distributed storage) as its core and works with compute, network, and archive software. Figure 3-20 shows the solution architecture.

Figure 3-20 Solution architecture for the satellite remote sensing scenario



Highlights:

- High cost-effectiveness
  - Mass data storage: Meets the requirements of 100 PB-level satellite data growth in the future.
  - Large EC ratio: The storage utilization reaches 90% or above.
  - High-density storage: up to 1.6 PB ultimate capacity.
- Fast storage
  - Fully symmetric distributed architecture and TB-level aggregate bandwidth capability.
  - DPC provides a single-stream bandwidth of more than 3 GB/s.
  - A client can be connected to multiple storage nodes, delivering high single-client and single-stream performance.
  - MPI-IO is supported, delivering good performance of MPI applications.
  - I/O-level load balancing is supported, minimizing load differences among storage nodes.
  - Network transmission is mainly based on RDMA, reducing the latency.
- Ease of use
  - Protocol interworking: Interworking among file, HDFS, and object protocols makes cross-department access more convenient.
  - Archiving solution: Atempo's integrated archiving solution offers optimal TCO and price-performance ratio.

3.5.8 Enterprise CAE

3.5.8.1 Scenario Characteristics

Computer aided engineering (CAE) uses the CAE simulation technology to optimize product design and accelerate development. It is applied to structural mechanics analysis, fluid mechanics design, electromagnetic design, collision design, noise design, and motor design.

CAE Segmentation	**Description**	***Main Application (Open Source)***	***Main Application (Commercial)***	***Application Characteristics***
Fluid simulation	Appearance, air channel, and pipe design; external air flow field, air channel, and air flow resistance etc.	*OpenFoam* *Code Saturne* *SU2, PALABOS*	*Ansys Fluent* *StarCCM+*	*The memory access is intensive. A single job supports a maximum of tens of thousands of cores. The memory capacity requirement is not high. Typical applications such as OpenFOAM and Ansys Fluent mainly use double-precision scalar computing.*
Structure (implicit)	Structure and appearance design; structural strength, fatigue,	*Code Aster*	*Abaqus* *Ansys Mechanical*	*In I/O- or memory-intensive access scenarios, the scalability is poor (hundred-core level), the I/O bandwidth of a single*

CAE Segmentation	Description	*Main Application (Open Source)*	*Main Application (Commercial)*	*Application Characteristics*
	heat, and NVH analysis			*process is about 200 MB/s, the total I/O bandwidth increases linearly with the number of processes, and the instruction set supports AVX/AVX2.*
Collision (explicit)	Collision safety analysis, high-speed impact analysis, and explosion analysis	-	LS-Dyna PamCrash Altair RADIOSS	The scalability is good, and the memory bandwidth and capacity requirements are not high. According to the data of an autonomous enterprise collected using Ls-Dyna, the memory access bandwidth of a single client is about 5 GB/s.
Electromagnetic emulation	Antenna, chip, and PCB design; electromagnetic interference simulation and optimization	-	Altair Feko Ansys HFSS	The memory capacity requirement is extremely high and the scalability is poor.

From the perspective of the service process, CAE includes CAE pre-processing, simulation calculation and solution, post-processing, and data archiving. The following table lists the data characteristics.

Scenario	Operation	Data Volume	I/O Model	Bandwidth/ Latency	Data Retention
CAE pre-processing	Engineering or product modeling, geometric modeling, and grid division	About 200 TB	Concurrent read and write (> 100 channels) of small files	Single cluster CIFS > 30,000 OPS	10 to 15 years
Simulation calculation and solution	Calculates grids through concurrent and iterative computing, which is usually an MPI application.	About the PB level	Mixed large and small files, sequential I/Os, 7:3 read/write ratio	Single-client read: 1 GB/s to 7 GB/s; single-thread: > 300 MB/s	/
Post-processing	Determine whether the design solution is reasonable based on the model design requirements.	About 10 TB to 50 TB	Small files and analysis reports	/	/

Scenario	Operation	Data Volume	I/O Model	Bandwidth/ Latency	Data Retention
Data archiving	CAE engineering files and simulation files	About 200 TB to 500 TB	/	/	10 to 15 years

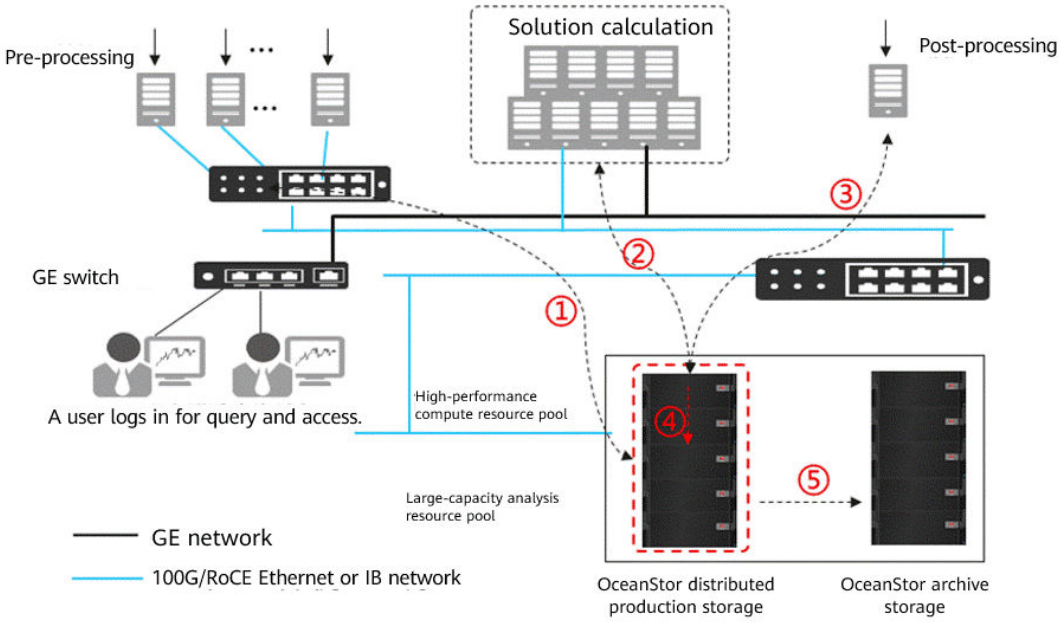
Data in enterprise CAE scenarios has the following characteristics and challenges:

- **Data silos in pre-processing and post-processing**: Modeling and CAE pre-processing and post-processing create geometric models on Linux workstations and divide computing grids. It is difficult for CAE to collaborate with pre-processing and post-processing. Processing tools in different phases have different requirements on storage protocols, and data needs to be migrated for multiple times, resulting in data silos.
- **MPI-IO requirements:** The CAE simulation cluster requires MPI-based efficient communication, and the storage supports MPI-IO for improved service performance.
- **Complex workloads**: A type of service generates multiple types of workloads which must be carried by different storage systems. These workloads include bandwidth-intensive, OPS-intensive, and concurrent (N:1 and N:N) workloads, resulting in low efficiency and complex management.

3.5.8.2 Scenario-Specific Solution

The HPDA storage solution designed for the CAE scenario uses OceanStor Pacific series parallel file storage (distributed storage) as its core and works with compute, network, and simulation software. Figure 3-21 shows the solution architecture.

Figure 3-21 Solution architecture for the enterprise CAE scenario



- In the CAD pre-processing (commercial software on the Windows/Linux platform) phase, product prototype modeling and computing grid division are performed for projects or products. Generated files are stored in the OceanStor Pacific storage system through the NFS protocol. High concurrency and high OPS operations are supported.
- In the CAE high-performance solution calculation phase, large and small files are mixed. The NFS protocol can be used to process these files, ensuring the I/O performance for both large and small files. For applications that require MPI, DPC is used to access storage, and then MPI is used to access storage after multiple iterative calculations, improving the simulation performance.
- In the post-processing phase, the simulation result is evaluated, processed, checked, and optimized (the calculation result is displayed and the product performance is evaluated). The CAD software, private clients, and post-processing software can communicate and share data with each other through NFS or the interworking between DPC and NFS.

4 Terms, Acronyms and Abbreviations

Table 4-1 HPDA storage solution terms, acronyms, and abbreviations

Name	Description
AI	Artificial Intelligence
CCportal	Cluster Computing Portal
CCScheduler	Cluster Computing Scheduler
DPC	Distributed Parallel Client
GID	Group identification
HPC	High-performance computing
HPDA	High Performance Data Analytics
Hyper MPI	Hyper message passing interface
MPI	Message passing interface
PCIe	Peripheral Component Interconnect Express
QoS	Quality of service
RDMA	Remote direct memory access
RoCE	RDMA over Converged Ethernet (RoCE)
TEM UID	Transmission electron microscopy User identification

