# OceanStor 100D Object Competitive Analysis - vs Ceph

Boston Storage Solution Group

FUTUREWEI Technologies

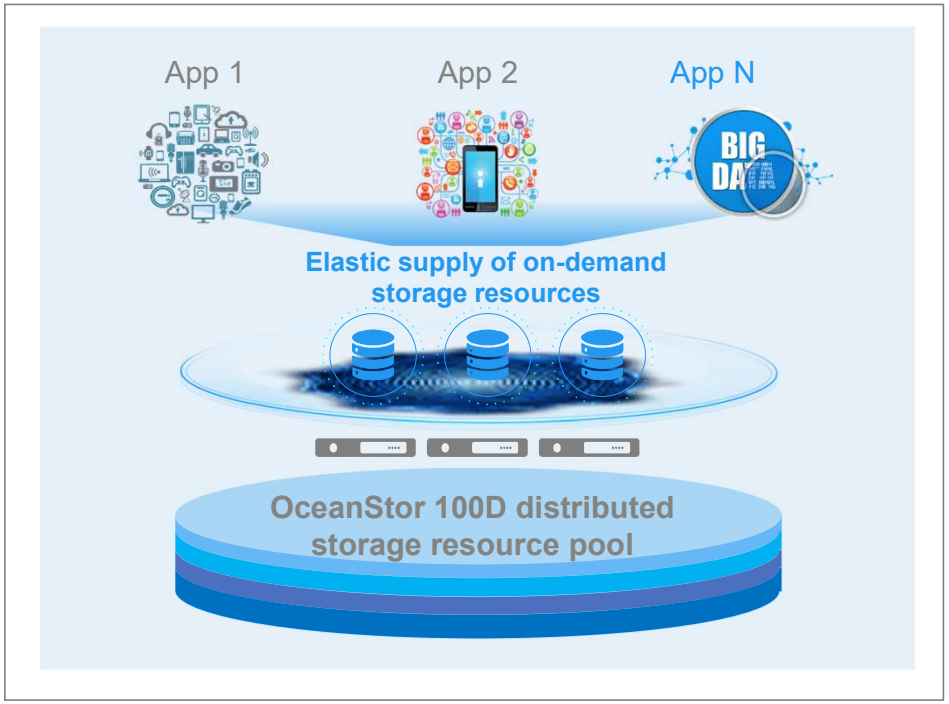
# OceanStor 100D Enterprise Scale Out Object Storage

- Challenges of object storage
- OceanStor 100D quick introduction
- Product spec comparison

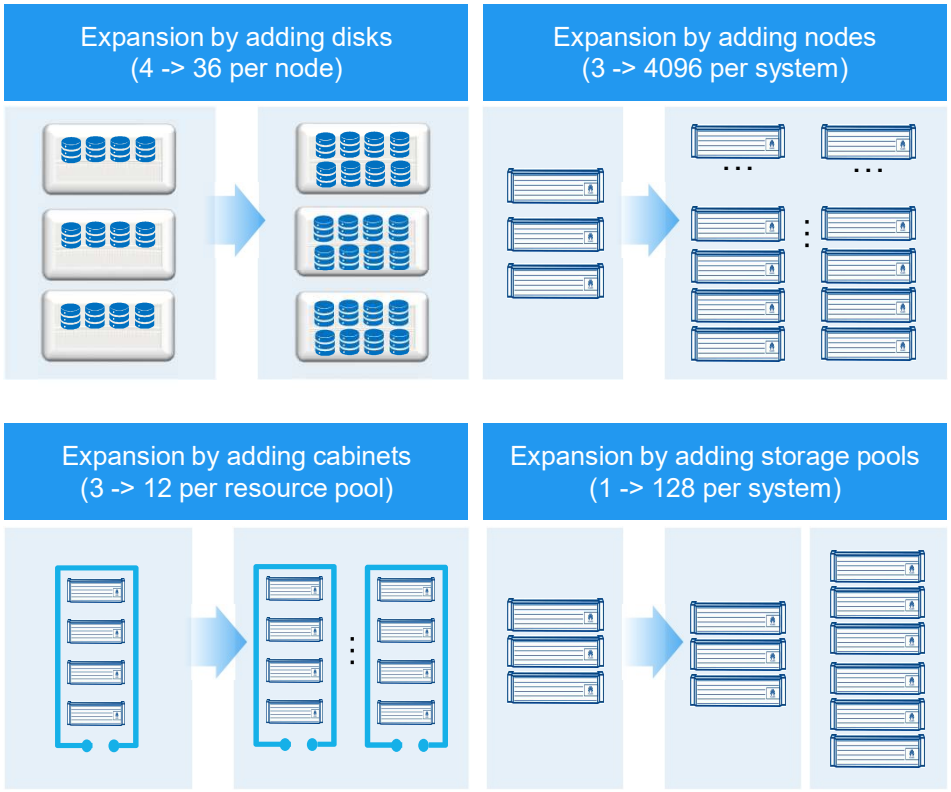


FUTUREWEI Technologies

# OceanStor 100D Quick Introduction



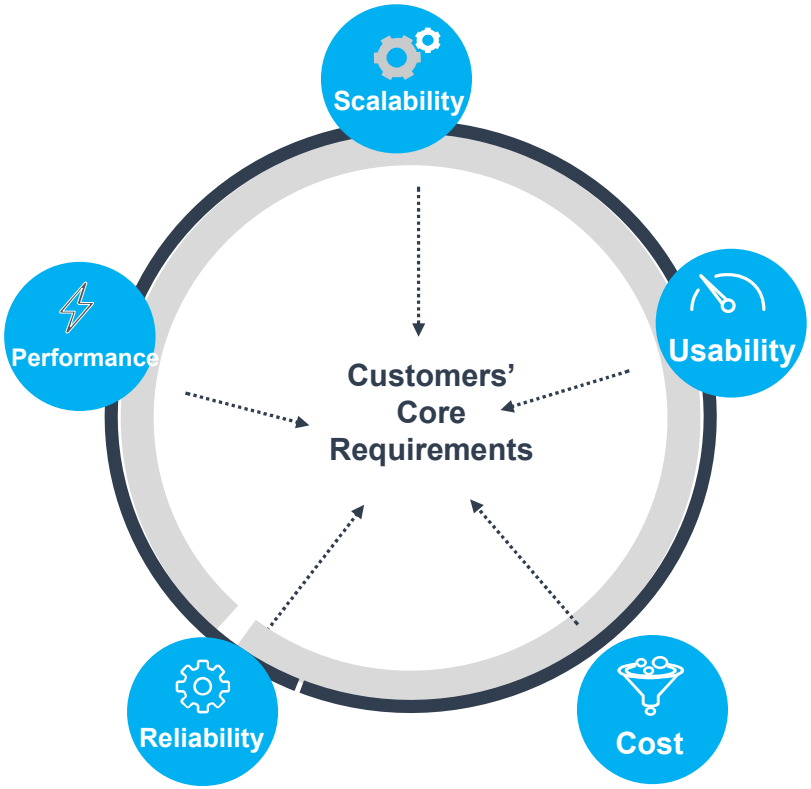

OceanStor 100D provides single name space resource pooling for multiple applications.




Its scale-out architecture allows flexible expansion up to 4096 nodes.

FUTUREWEI Technologies

# OceanStor 100D vs Core Storage Requirements




Huawei OceanStor 100D (originally called FusionStorage) is an intelligent, distributed storage product with scale-out architecture to support the business needs of both today and tomorrow.

- High scalability: Scale-out architecture supports up to 4096 nodes.
- High performance: High performance based on specially designed hardware and many software features.
- High Reliability: Designed for core enterprise systems. 100D supports fast failover and failback.
- Low cost: Off-the-shelf servers or specialized nodes. Customers can pick best fit nodes and mix them. Tiering is supporting to provide best TCO.
- High usability: One cluster provides industry standard interfaces, such as block, HDFS, object, and file storage, for upper-layer applications. This greatly simplifies operations complicated by siloed storage systems.

FUTUREWEI Technologies

# Object Storage Spec Comparison

| | OceanStor 100D | Ceph on Comparable Hardware |
|---|---|---|
| Max #nodes per cluster | 4096 with near linear performance growth | Performance degrades quickly when cluster size grow, 1000+ nodes cluster is very rare |
| Max front end speed | 100GE / Infiniband w/ RDMA | 10/25GE |
| Max backend speed | 100GE / Infiniband w/ RDMA | 10/25GE |
| Capacity density | Up to 1.6PB/5U (pacific node) | Around 0.5PB/4U |
| Rack capacity | Up to 6 Pacific chassis with capacity up to 9.6PB in 42U rack | Up to 9x 4U nodes with capacity up to 6PB |
| Power density per 4U | 790w-1200w (depends on configuration) | Around 1500w |
| Max capacity in 5kw rack | 3024TB (6 node cluster) | 1440-2400TB (depends on model) |
| Max capacity in 8kw rack | 5040TB (6 node cluster) | 2400-4000TB (depends on model) |
| Rebuild speed | 2TB/hour (HDD) | < 0.25TB/hour (HDD) |
| Max #bucket | 100 million | 0.1 million |
| Max #object per bucket | 100 billion | 1 billion |

FUTUREWEI Technologies

# OceanStor 100D Enterprise Scale Out Object Storage

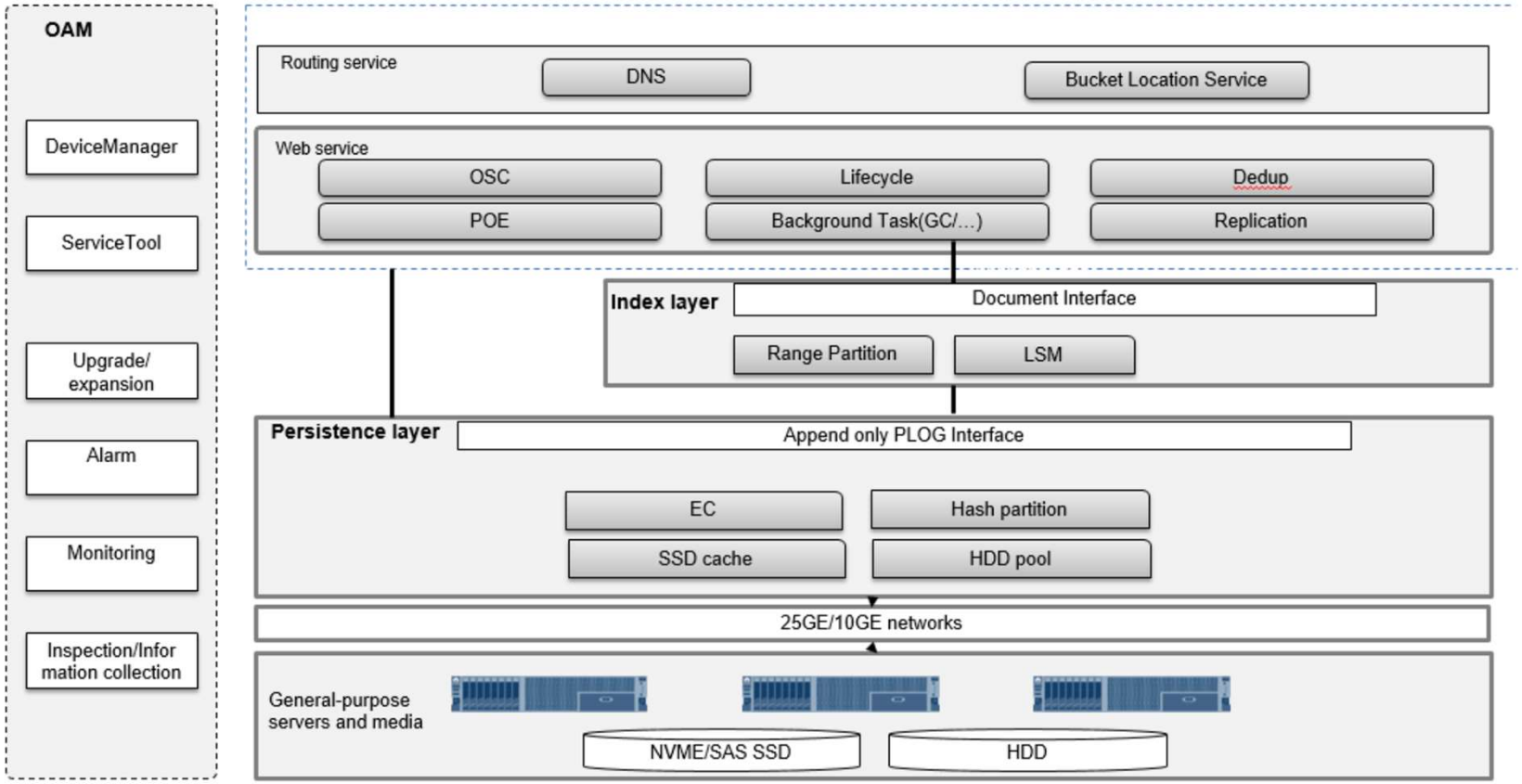
Overview | Software | Performance

**System Architecture** | Enterprise Features | Management

- Architecture overview
- Scalability overview
- Reliability overview
- Product comparison

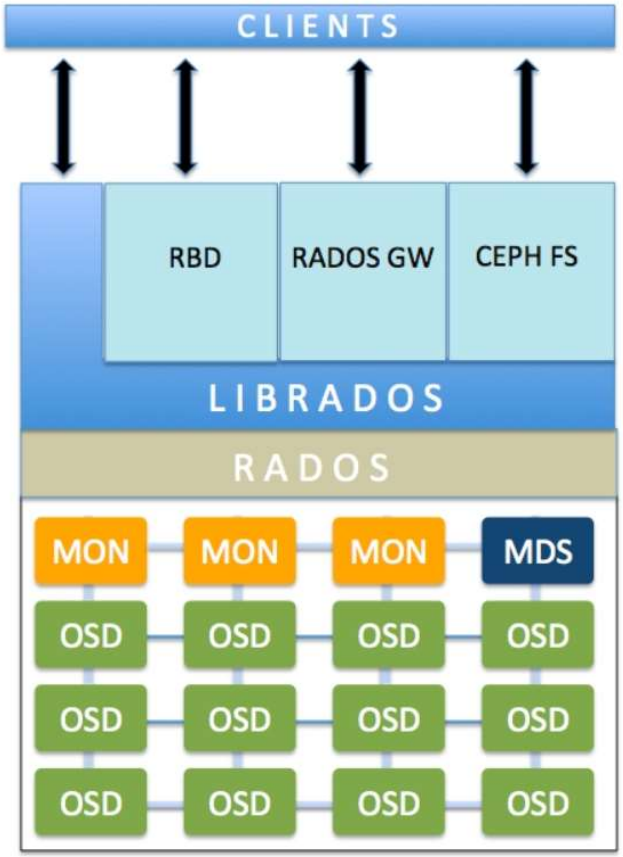
FUTUREWEI Technologies

# 100D Object Software Architecture




FUTUREWEI Technologies

# Ceph Software Architecture

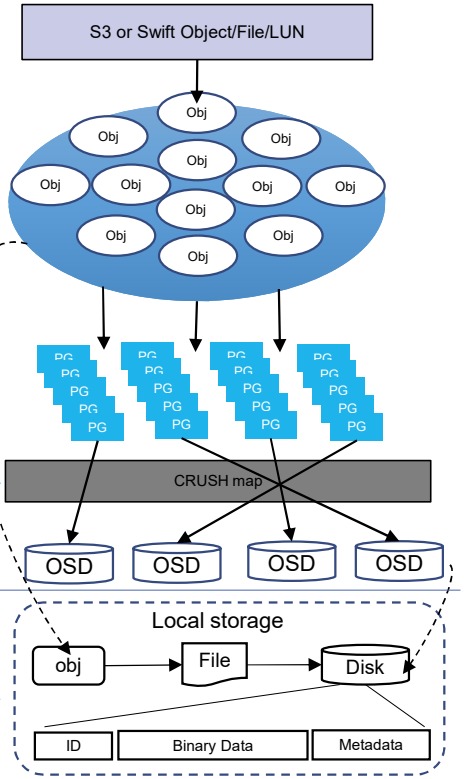



Mainstream open-source software in the industry

Three steps to query the mapping: LUNs,files and objects are mapped to multiple continuous objects on the local OSD.

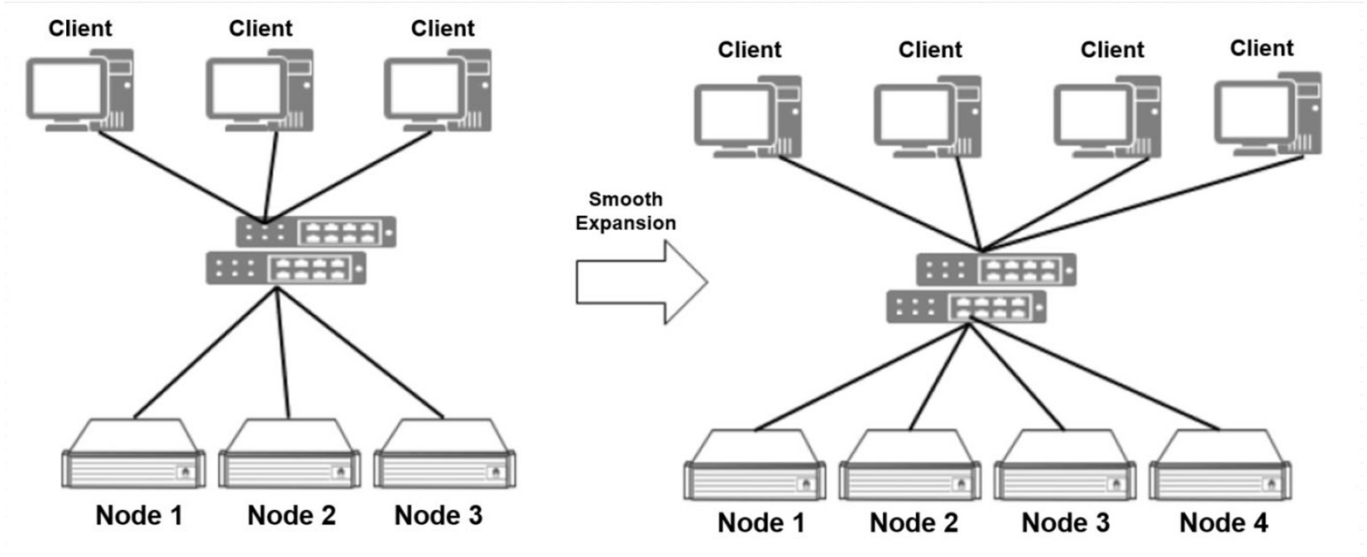
Placement groups (PGs) have a great impact on system performance and layout balancing. They need to be dynamically split and adjusted based on the storage scale. This adjustment affects system stability.

The minimum management granularity is by object. The size of an object is adjustable, which affects metadata management, space consumption, performance, and even read/write stability.




FUTUREWEI Technologies

# OceanStor Scalability: Elastic Expansion to 4096 nodes

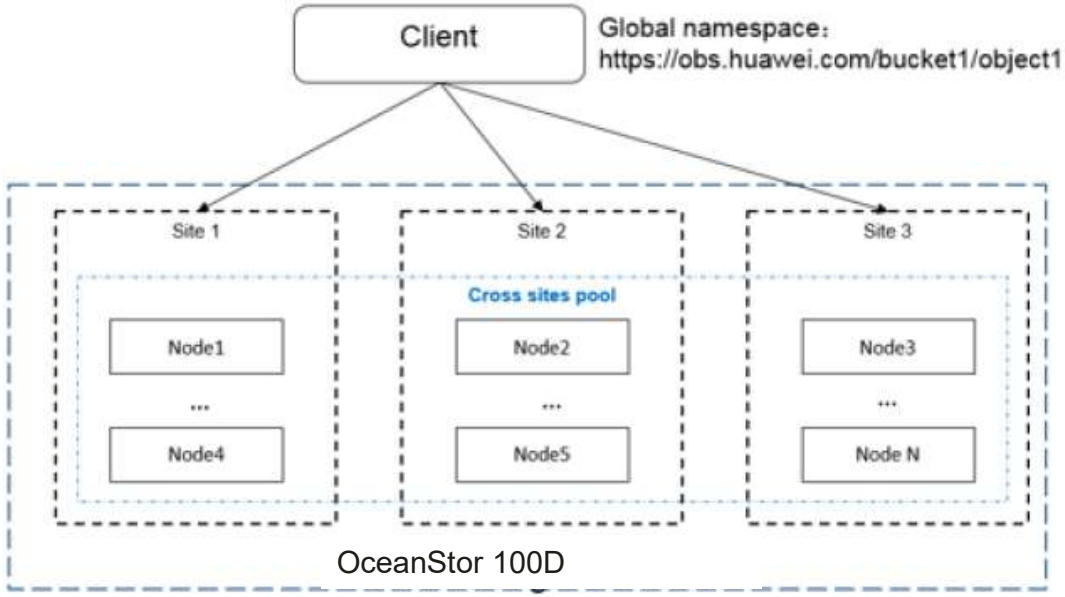



OceanStor 100D

- **Flexible expansion:** You can add disks and nodes to expand capacity.
- **Fast load balancing:** After new nodes are added, 100D object storage implements fast load balancing and avoids migration of a large amount of data.
- **Linear performance growth:** Computing, storage, and cache resources are evenly distributed on each node. The system TPS, throughput, and cache linearly increase as more nodes are added.

FUTUREWEI Technologies

# Reliability Overview




1. Data Redundancy protection: N+M EC, Node-level and Cabinet-level Security, Cross-site EC.
2. Data Consistency: online data consistency check and background data consistency check.
3. Fast Data Reconstruction: restoration speed is up to 2TB/hour
4. Cluster Reliability: fully symmetric architecture (any node can process user requests), smooth service switch-over in case of node failure.
5. Hardware Reliability
6. Link Reliability: no single point of failure (port, switch).



HUAWEI

# Product Level Comparison

| | OceanStor 100D Object | Ceph on Comparable Hardware |
|---|---|---|
| Performance | • Support RDMA.<br>• Support end-to-end NVMe.<br>• Good small IO perf with EC aggregation + plog append-only write.<br>• Self-developed EC algorithm with higher efficiency and low CPU usage. | • Do not support RDMA.<br>• Do not support NVMe.<br>• The performance for small I/Os and EC is poor. |
| Reliability | • Active-Active Cross AZ redundancy: Up to 12 nines.<br>• Quick recovery.<br>• Data Evenly distributed between nodes and disks.<br>• Subhealth management<br>• Support EC and EC folding, (No multi-copy)<br>• Support 3AZ EC | • Design-based CRUSH algorithm constraints: uneven data distribution, uneven disk space usage.<br>• Insufficient subhealth processing (only subhealth disk check, no subhealth network check)<br>• Support multi-copy (1 ~ 6) and EC.<br>• No 3AZ EC |
| Scalability | • One cluster can scale up to 4096 storage nodes.<br>• Cluster does fast node balancing when add a node, avoid large data moving but make new resource immediately effective.<br>• Adding node will not impact primary IO traffic.<br>• Disks, computing nodes, and storage nodes could be added separately or together. | • Restricted by the CRUSH algorithm, adding nodes costs high, and large-scale capacity expansion is difficult. |
| Usability | • Three-Layer Intelligent O&M.<br>• Work well with other ecological management system, including VMWare vCenter.<br>• One-key non-disruptive upgrade (SW, FW). | • Lack of cluster management.<br>• Poor usability at maintenance interfaces. |
| Cost | • High EC ratio (22+2,  ratio 91.67%). Cross-site EC ratio (20+16) is 55.5%<br>• Deduplication and compression are supported. | • EC is CPU intensive, Very low data ratio (33%) for three-way mirroring.<br>• Deduplication is not supported.<br>• Inline compression is supported. |

# OceanStor 100D Enterprise Scale Out Object Storage

- Scalability
- Performance Optimization
- Efficiency
- Reliability

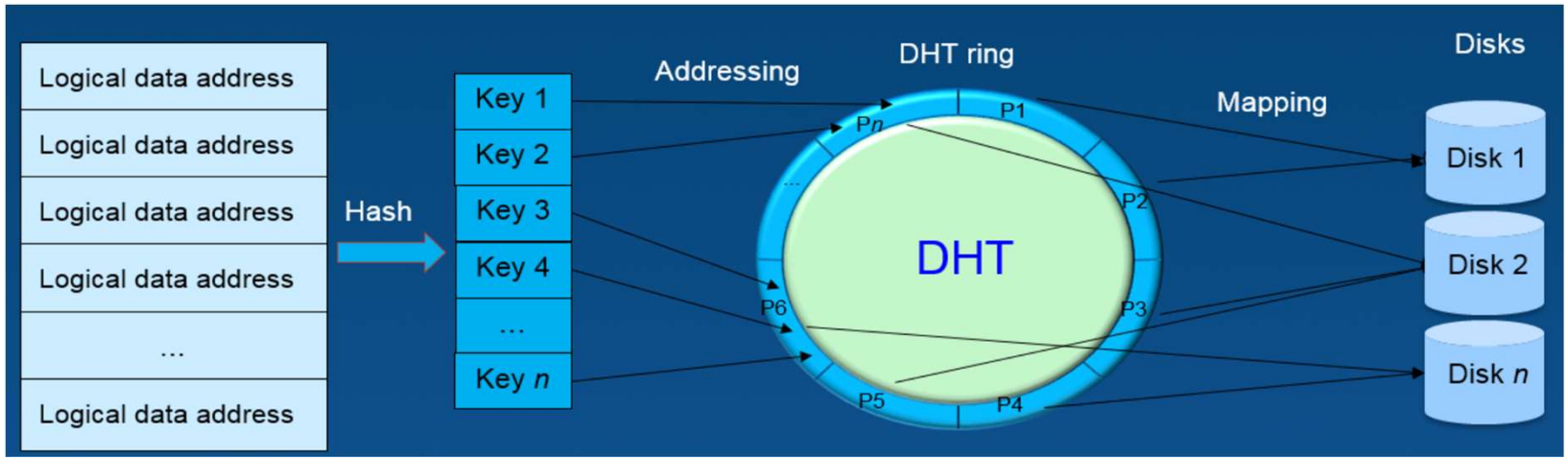
Overview | **Software** | Performance

System Architecture | Enterprise Features | Management

FUTUREWEI Technologies

# Scalability: Data Routing

OceanStor 100D object data routing




OceanStor 100D Object

- **Outstanding performance:** The DHT ring enables data to be evenly stored and processed on all disks, eliminating read and write performance bottlenecks incurred by frequent data access on certain disks.
- **High reliability:** The partition allocation algorithms are flexible. Identical data copies are not stored onto the same disk, server, or cabinet.
- **Rapid scale-out:** When new physical nodes are added, only part of the data needs to be migrated for load balancing.

Ceph Object

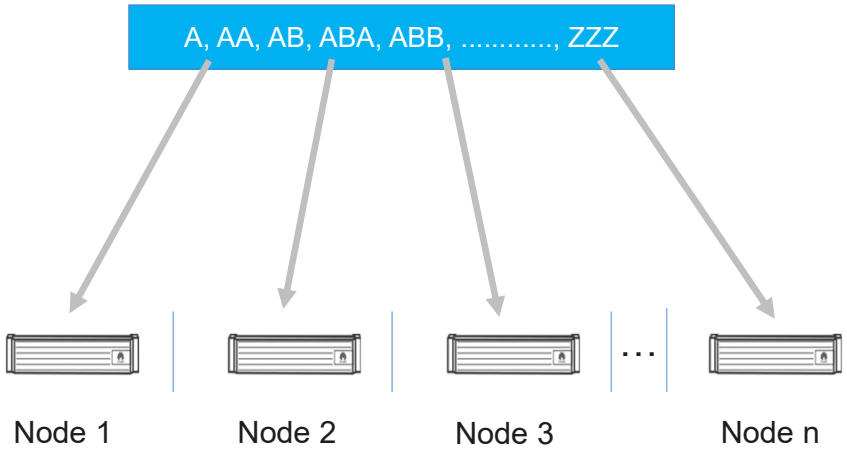
CRUSH mapping for obj->PG and PG->OSD:

- Restricted by the CRUSH algorithm, adding nodes costs high, and large-scale capacity expansion is difficult.
- uneven data distribution, uneven disk space usage.

FUTUREWEI Technologies

# Performance: Range Partitioning and WAL Submission Improve Performance and Process Hundreds of Billions of Objects

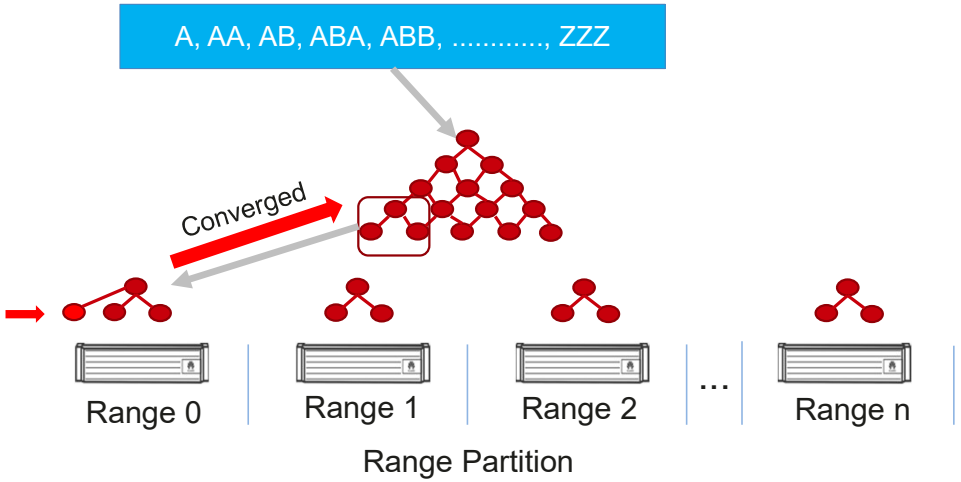
Range




1. Metadata is evenly distributed in lexicographic order + range partitioning mode. Metadata is cached to SSDs.

- You can locate the node where the metadata index resides based on the bucket name and prefix to quickly search for and traverse metadata.

WAL



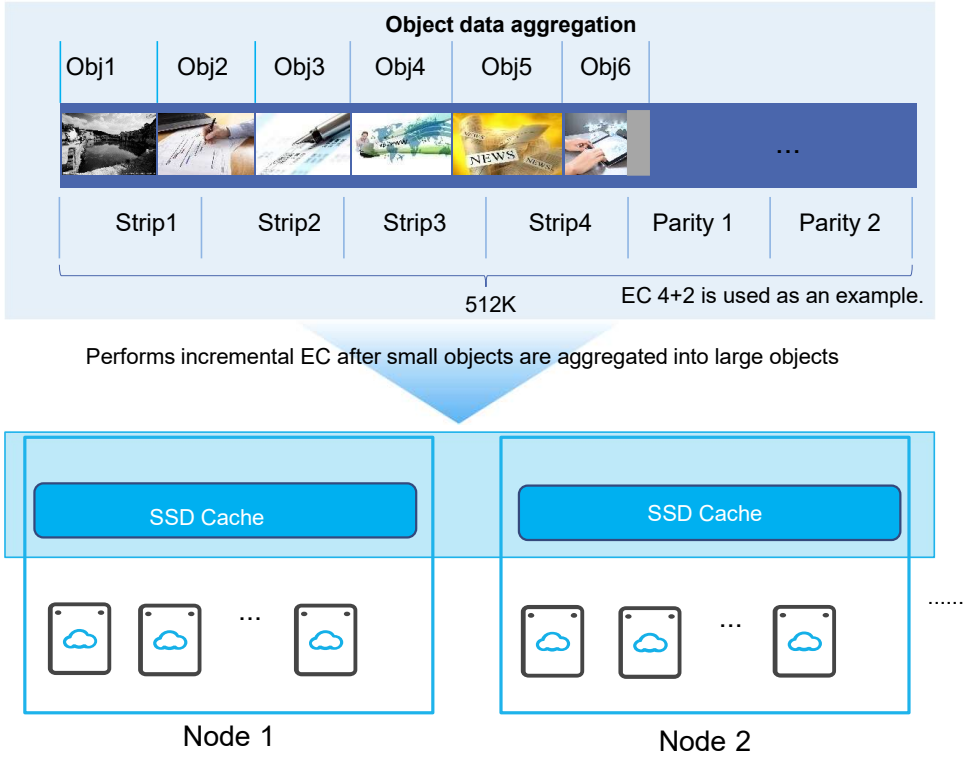

1. The WAL mode is used. The foreground records only one write log operation for one PUT operation.
2. Data is compacted in the background.

- The WAL mode reduces foreground write I/Os.
- SSDs improve the access speed.

FUTUREWEI Technologies

# Performance: Online Aggregation of Massive Small Objects Improves Performance and Capacity

When the storage file object is smaller than the system strip, a large number of space fragments are generated, which greatly affects the space usage and access efficiency.

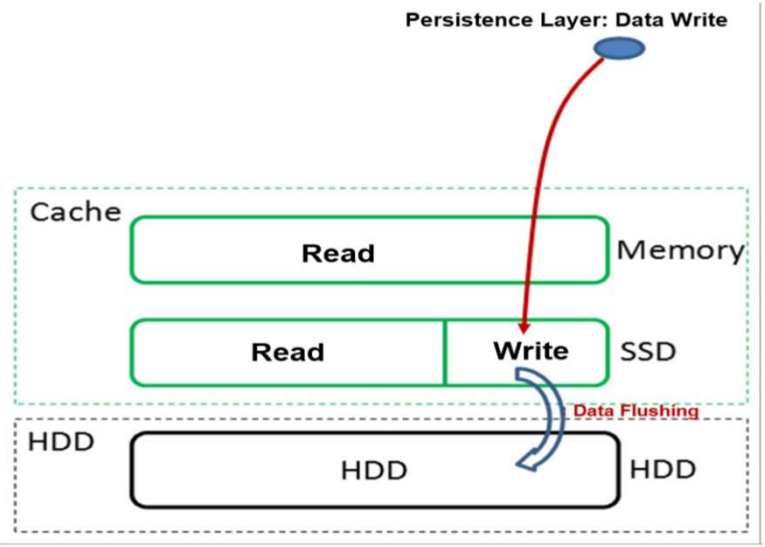



- Incremental EC aggregates small objects into large objects without performance loss.
- Reduce storage space fragments in massive small file storage scenarios, such as government big data storage, carrier log retention, and bill/medical image archiving.
- Improve the space utilization of small objects from 33% (three copies) to over 80% (12+3).
- SSD cache is used for object aggregation, improving the performance of a single node by six times (PUT 3000 TPS per node).

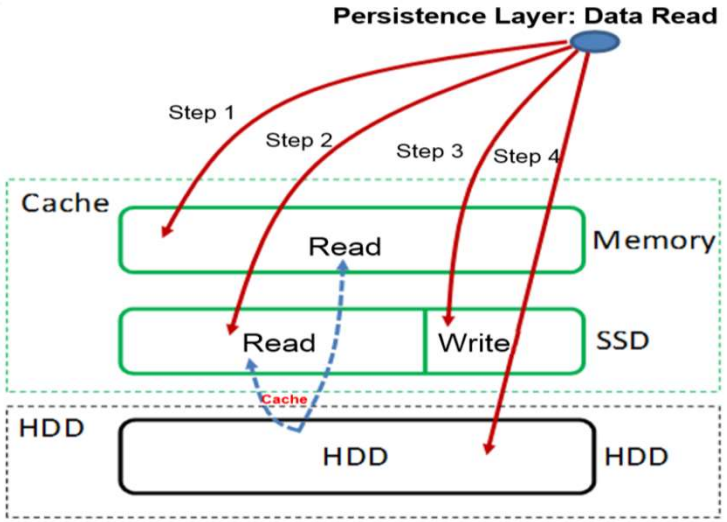
FUTUREWEI Technologies

# Performance: Cache

100D Object Write Cache




100D Object Read Cache

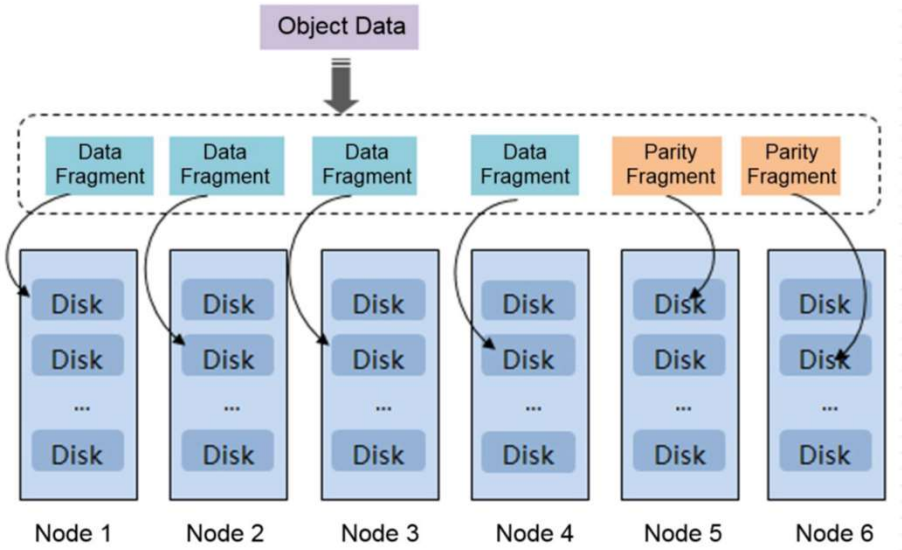



OceanStor 100D Object:
- Multi-level cache mechanism improves I/O performance.
- SSD cache is used for object aggregation.
- Read memory cache uses LRU mechanism and SSD cache is used for hotspot.
- Supports prefetching.
- Large IO (>256KB) passthrough.

Ceph Object:
- Cache tiering is deprecated: it does not provide any performance improvement for most workloads and introduces stability issues.
- Future version will use dm-cache:
- Default stochastic multiqueue(SMQ) cache policy;
- Use SSD for random reads/writes;
- Sequential reads/writes are not cached.

FUTUREWEI Technologies

# Efficiency: EC




EC 4+2

Data written into 100D is divided into *N* data fragments, and *M* parity fragments are generated for the *N* data fragments using EC. If *M* fragments are damaged in an EC group, the system implements data recovery from the *N* fragments.

The EC data protection mode maintains high disk utilization of $N/(N + M)$.

When nodes are faulty and the number of normal nodes does not meet the minimum quantity required by an EC scheme, users are allowed to reduce data fragments to ensure that the reliability does not deteriorate. The reduction rule is $(N/2 + M)$. For example, EC scheme 4+2 is shrunk to 2+2, EC scheme 8+2 to 4+2.

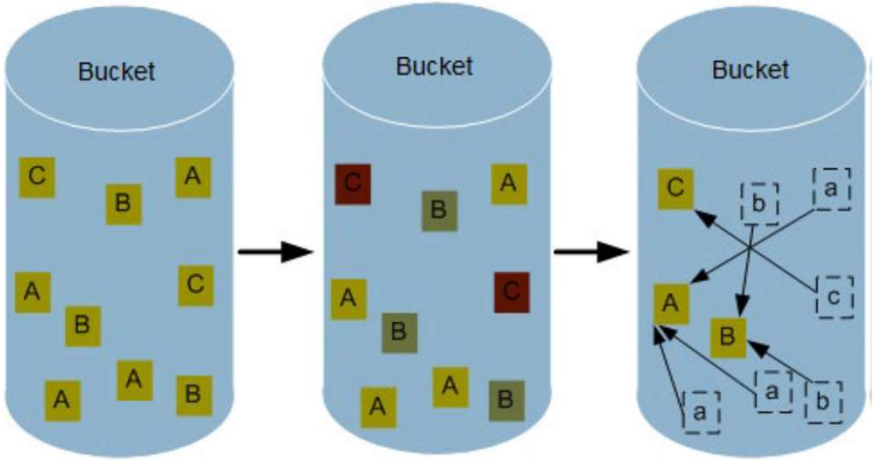
# Efficiency: EC Comparison

| OceanStor 100D Object | Ceph |
| --- | --- |
| Home-developed algorithm with higher efficiency and less CPU usage. | Ported open-source libraries (Jerasure by default).<br>EC is cpu and memory intensive, multi-copy is recommended to use in production.<br>EC is not supported on Ceph Block Device. |
| EC-folding :1 | No EC-folding. Allow degraded write. |
| Plog write-append: reduces read/write amplification and compute overhead, improves peak performance | Write in place. |
| EC aggregation improves storage efficiency and reduces write amplification. | Small objects are not aggregated during write. |
| Support maximum 22+2 or 20+4 EC scheme. (22 +2 is 92% storage efficiency). | Default EC profile is 2+1. |
| Node-level and cabinet-level fault domain | Failure domains such as device, rack, room, etc |
| Cross-site EC | No 3AZ EC |

FUTUREWEI Technologies

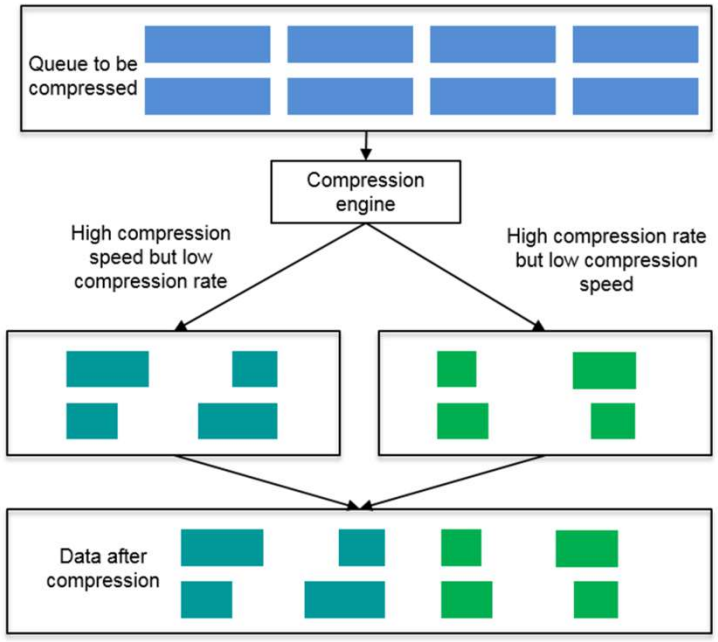
# Efficiency: Deduplication and Compression

Object-level Deduplication




OceanStor 100D Object:
- Object-level dedup can be enabled for entire system.
- Compression engine runs in a combination of two compression algorithms. User can config one compression algorithm for each storage pool

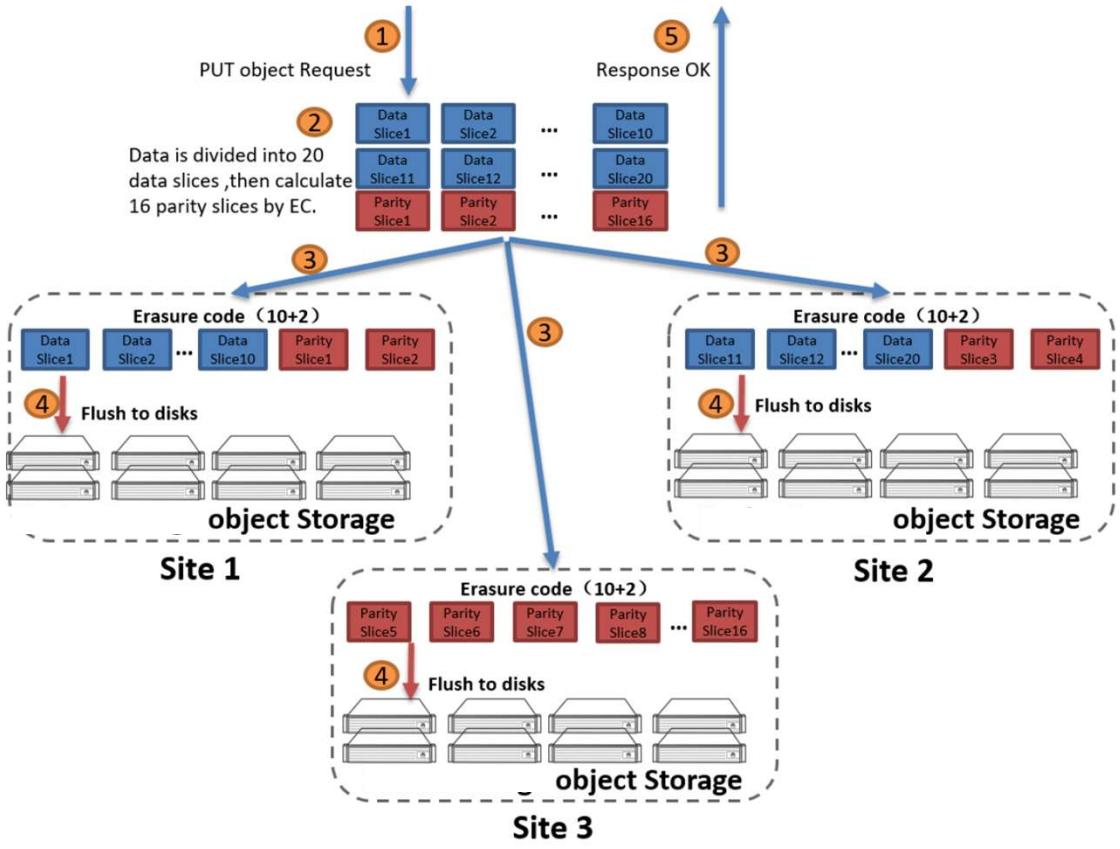
Compression




Ceph Object:
- Dedup is not supported.
- Inline compression is supported (uses open-source algorithms), QAT acceleration.

FUTUREWEI Technologies

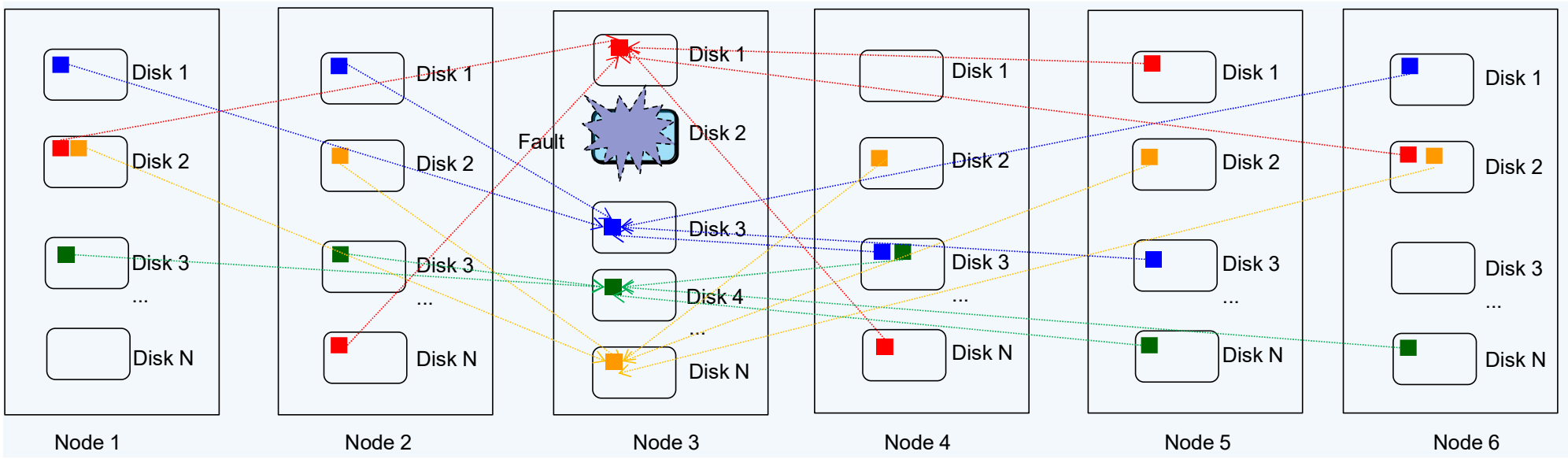
# Reliability: Cross-site EC adds reliability up to 99.9999999999% (twelve nines).

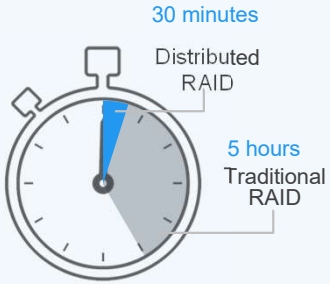



- OceanStor 100D object storage uses cross-site Erasure Coding (EC) to provide data redundancy protection among sites. Compared with the traditional two-copy redundancy mode, cross-site EC improves space utilization by more than 12% without changing data durability.
- The data utilization of the 20+16 cross-site EC scheme is 55.5% and that of the two-site copy mode (EC configuration of a single site is 10+2) is 10/(12 + 12) = 41.6%. Cross-site EC enables more efficient use of storage space.
- Data is synchronous among the three sites and the RPO is 0.

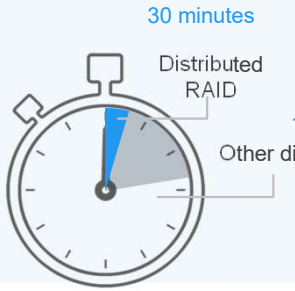
FUTUREWEI Technologies

# Reliability: Fast Rebuild




**10 times** faster than traditional storage




**Twice** faster than other distributed storage

FUTUREWEI Technologies

# OceanStor 100D Enterprise Scale Out Object Storage

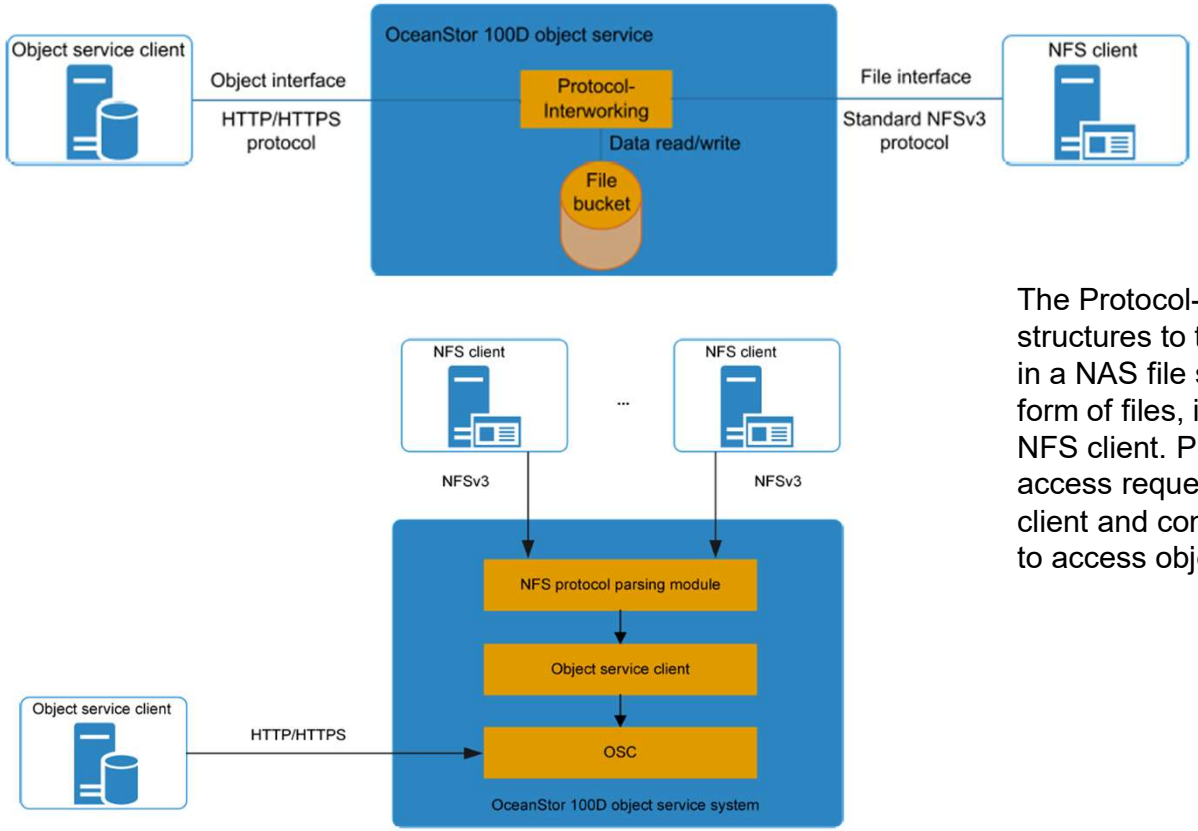
Overview | Software | Performance

System Architecture | **Enterprise Features** | Management

- Protocol Interworking
- Async replication
- Object versioning
- Object lifecycle management
- QoS
- Multi-tenant management
- Quota
- WORM
- Access permission control
- Bucket access logging

FUTUREWEI Technologies

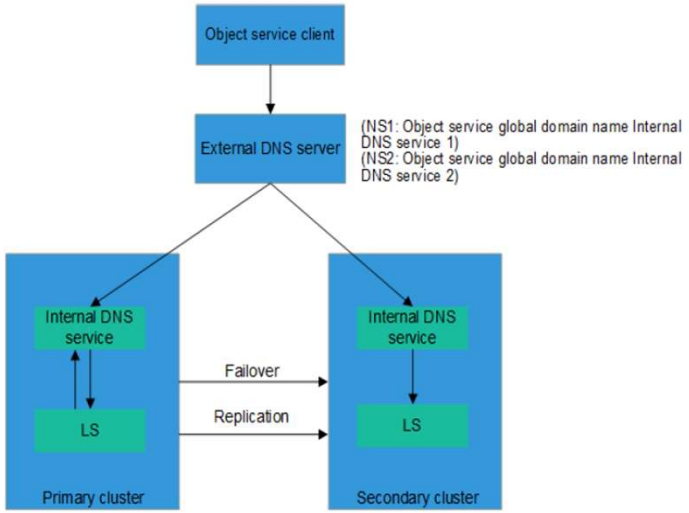
# Protocol Interworking




The Protocol-Interworking feature maps object structures to the directories or files in a NAS file system, and maps object data, in the form of files, in a bucket to an NFS client. Protocol-Interworking also receives file access requests sent by the NFS client and converts them into object service requests to access object service data.

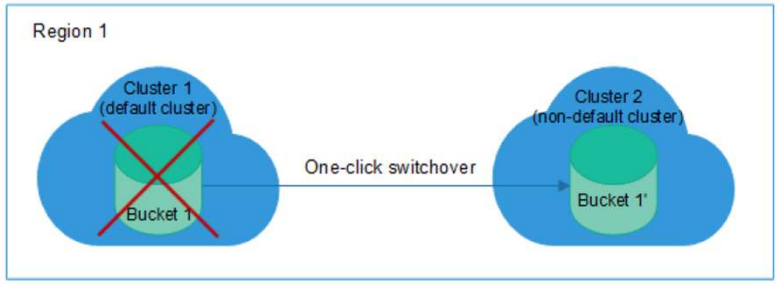
FUTUREWEI Technologies

# HyperReplication Asynchronous Replication



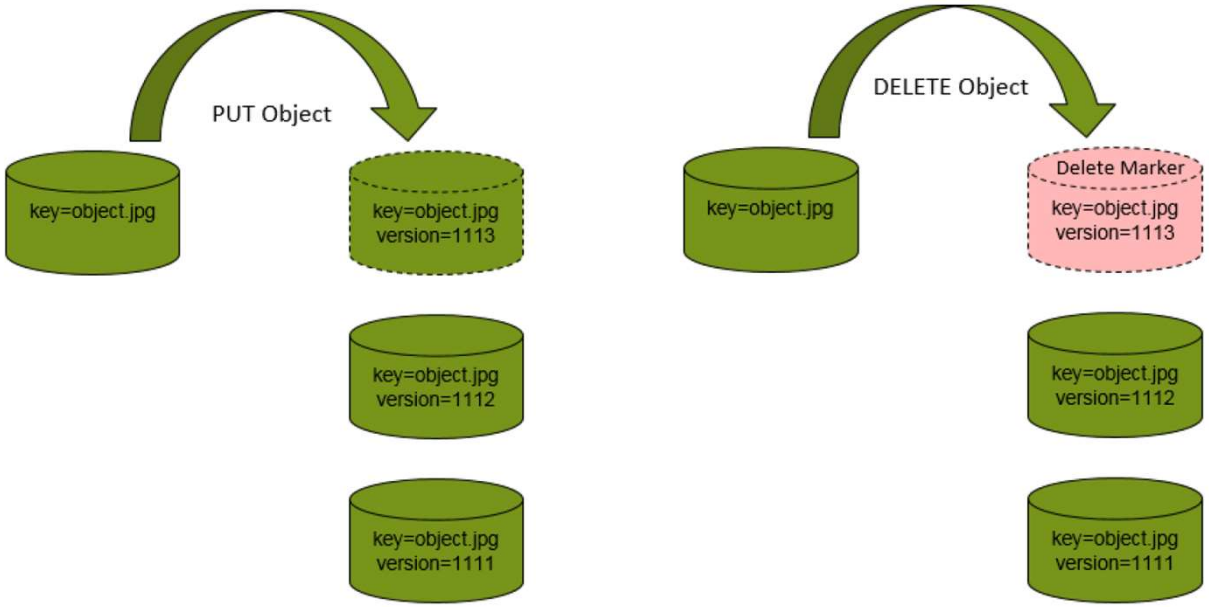

An asynchronous remote replication relationship is established between the production center and a secondary DR center.

- After the initial synchronization is complete, the asynchronous remote replication enters the periodic incremental synchronization phase. Based on the customized interval, the system periodically starts the synchronization task and synchronizes the incremental data written by the production host from the end of the previous period to the current time.
- Support for one-to-one and one-to-many
- Simple management and one-click DR drill and recovery




FUTUREWEI Technologies

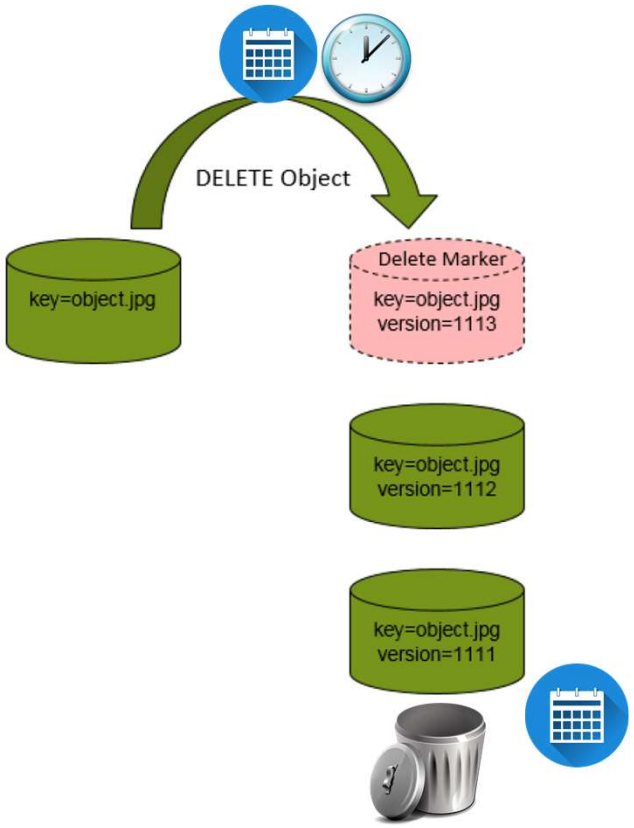
# Object versioning




When object versioning feature is enabled, object changes will be kept via different versions.

- Older versions can be accessed by specifying a version number.
- Deleted objects are marked by delete markers and can still be accessed by issuing version numbers.

# Object lifecycle management



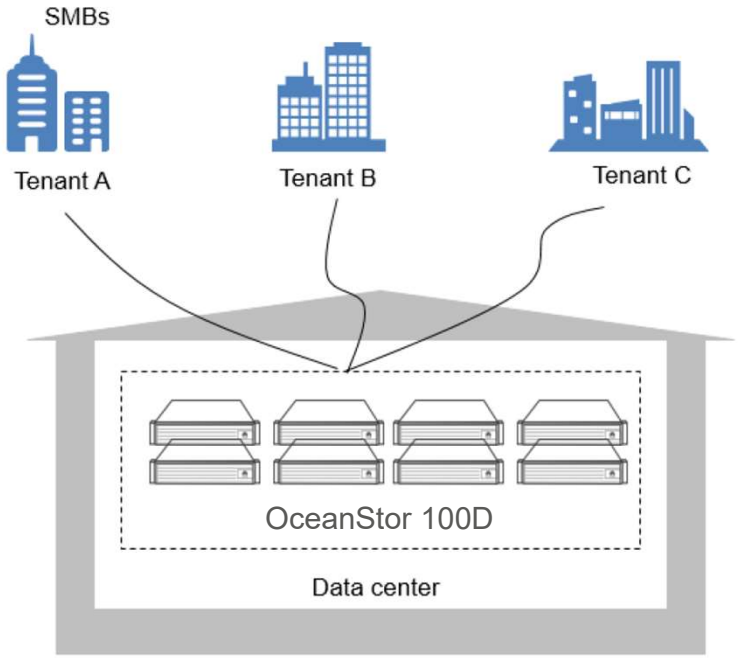

OceanStor 100D object storage provides object lifecycle management to store objects cost-effectively throughout their lifecycle.

By using this function, you can define whether objects expire after being stored for a specific period of time or at a specific point in time. This function is also available for noncurrent versions of objects when object versioning is enabled.

For example:

- You may need periodic log files for a week or month. After that, you might want to delete them.
- Some documents are frequently accessed for a period of time but then infrequently accessed. You can archive these documents and delete them later.

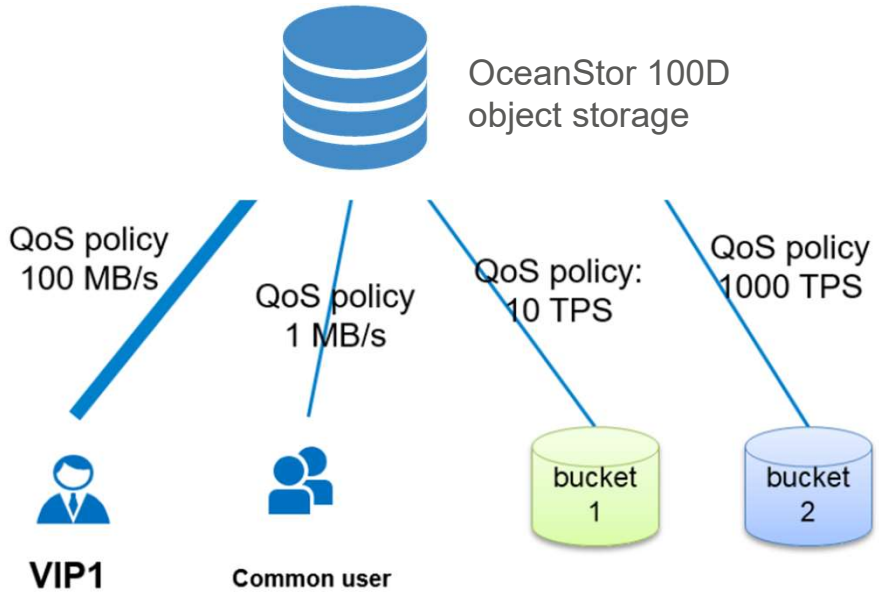
# Multi-tenant management




OceanStor 100D object storage provides multi-tenant management. Data of different tenants is logically isolated to facilitate resource allocation. Multi-tenant management has the following benefits:

- A single system provides a variety of client services, reducing initial investments.
- The system is centrally managed, data is logically isolated, and online storage is supported.
- Encrypted HTTPS transmission and user authentication are supported to ensure data transmission security.

# QoS




In multi-tenant scenarios such as private cloud, customers require that transactions per second (TPS) and bandwidth resources in storage pools be properly allocated to tenants or buckets with different priorities and that the TPS and bandwidth resources of mission-critical services be sufficient. To meet customer requirements, FusionStorage object storage provides the following refined QoS capabilities:

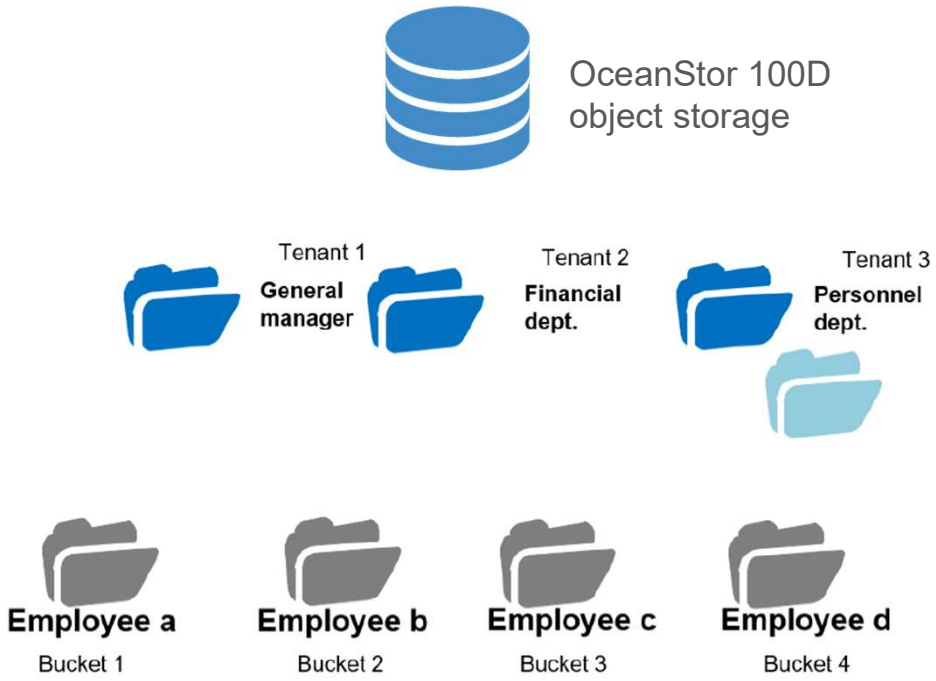
- **Refined I/O control**

Enables the system to provide differentiated services for tenants and buckets with different priorities.

- **TPS- and bandwidth-based QoS for tenants and buckets**

Accurately controls operations, such as PUT, GET, DELETE, and LIST.
QoS allocates buckets with different TPS and bandwidth capabilities for applications of different priorities. This maximizes storage pool resource utilization and prevents mission-critical services from being affected by other services. Different QoS policies can be configured for VIP and common tenants in the same system to ensure service quality for high-priority tenants.

# Quota




OceanStor 100D object storage supports bucket and tenant capacity quotas as well as object resource statistics. The following figure shows a capacity quota example where company departments represent tenants and employees in the departments represent buckets. You can set a 40 TB quota for the financial department (tenant 2) and a 10 TB quota for employee b (bucket 2) in the department.

The capacity quota function of 100D object storage has the following characteristics:

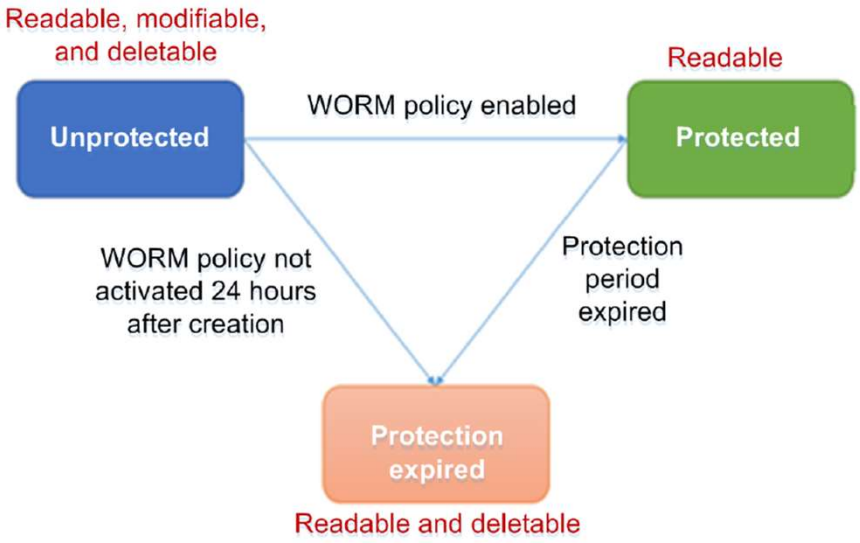
- **Bucket capacity quota**

Specifies the maximum size of a bucket. When the bucket size reaches the specified upper limit, new data cannot be written into the bucket.

- **Tenant capacity quota**

Specifies the maximum capacity assigned to a tenant. When the total size of buckets in a tenant reaches the specified upper limit, the tenant and all its users cannot write new data.

# WORM



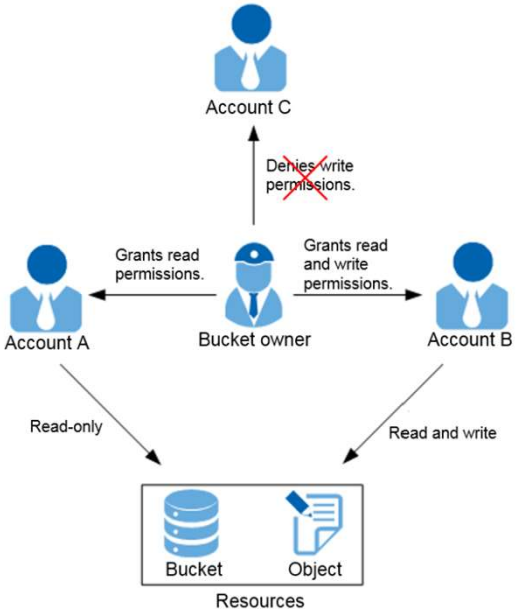

Write Once Read Many (WORM) is a technology that allows data to be read-only once being written. Users can set protection periods for objects. During protection periods, objects can be read but cannot be modified or deleted. After protection periods expire, objects can be read or deleted but cannot be modified.

WORM policies can be configured for buckets. Different buckets can be configured with different WORM policies. In addition, you can specify different object name prefixes and protection periods in WORM policies. For example, you can set a 100-day protection period for objects whose names start with **prefix1** and a 365-day protection period for objects whose names start with **prefix2**.

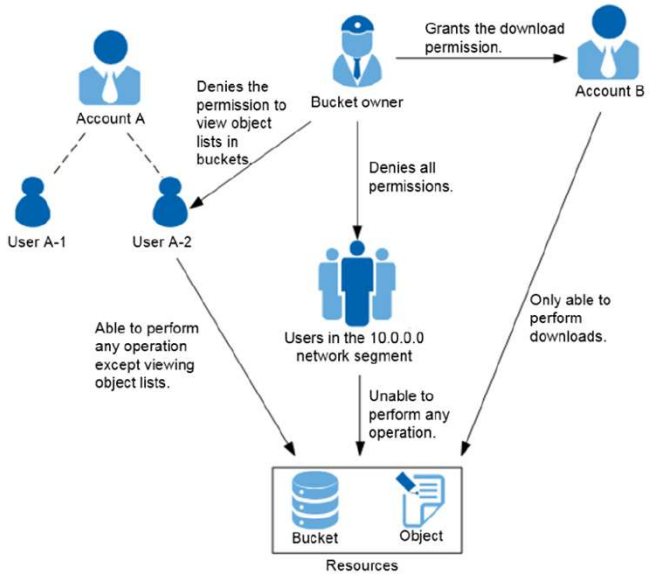
Objects enabled with WORM have three states: unprotected, protected, and protection expired.

- Unprotected: Objects in the unprotected state can be read, modified, and deleted, same as common objects.
- Protected: After a WORM policy is enabled for a bucket, the objects that meet the WORM policy enter the protected state and can only be read.
- Protection expired: When the WORM protection period of objects expires, the objects enter the protection expired state. In this state, the objects can only be read or deleted.

# Access permission control




ACLs grant accounts (also called tenants in 100D object storage) the permission to access resources. Each entry in an ACL specifies permissions (read-only, write, or read and write) of specific accounts. ACLs can grant but cannot deny permissions.




Bucket policies control access from accounts and users to buckets and objects. Bucket policies can both grant and deny permissions. Bucket policies provide more refined permission control than ACLs. For example, bucket policies can control specific operations (such as PUT, GET, and DELETE), forcibly enable HTTPS access, control access from specific IP address segments, allow access to objects with specific prefixes, and grant access permissions to specific clients.

# Enterprise Feature Comparison

| Features | OceanStor 100D Object (Enterprise-class quality) | Ceph |
|---|---|---|
| Protocol Interworking | S3, Swift, NFS | Support but not recommended by Ceph |
| 3AZ* | Allow 3AZs to form a high availability solution using 20+16 EC. | Multi-site solution is based on async log replication. |
| Asynchronous Replication | Policy based asynchronous replication with bucket level. | Supported as multi-site solution. Allow “active-active” solution to write into non-master nodes. Provides eventual consistency and the last write wins. |
| Object versioning | Object updates and deletions create versions for the old data. | Support ROW style snapshots, performance degradation as snapshot are taken |
| Object lifecycle management | Objects can be automatically deleted after a designated time. Same as versions. | Supports expiration dates for objects. |
| QoS | Bandwidth and TPS limitations | QoS is only for block. RADOS GW with dmclock is experimental only. |
| Multi-tenant management | Data is logically isolated | Support simple multi-tenant features(separate user and bucket namespaces for each tenant) |
| Quota | Bucket capacity quota and tenant capacity quota | User Quota and bucket quota |
| WORM | Policies configured for buckets | Object Lock API support merged into master on Jun 2019** |
| Authentication & ACL | Define ACLs and bucket policies | Support LDAP, Multi-factor and OpenConnect ID for authentication. ACLs and bucket policies. |
| Bucket access logging | Operations are logged in a bucket for logs | Ceph allows usage logging for each user. |
| Encryption | Solution level / external (server side) encryption | Support server side encryption, QAT acceleration supported. |
| Policy | Bucket policies | Bucket policies. Do not yet support setting policies on users, groups, or roles. |
| S3 Select | Roadmap | Partially Support(Limited operators) |
| Tiering | Roadmap | SSD Cache tier and HDD tier |

*: available on certain release versions

**: recently updated and not sure whether GUI support is added.

# OceanStor 100D Enterprise Scale Out Object Storage

- OceanStor performance
- Ceph performance
- Quick comparison

Overview | Software | **Performance**

System Architecture | Enterprise Features | Management

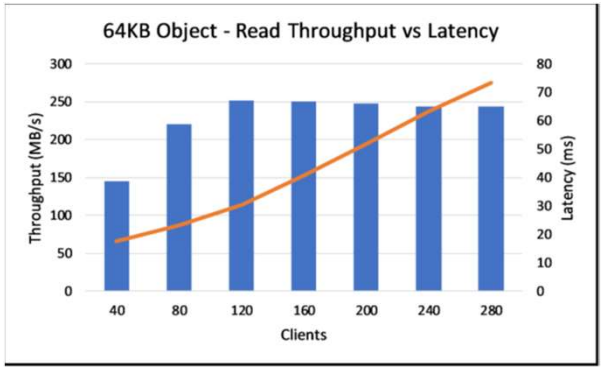
FUTUREWEI Technologies

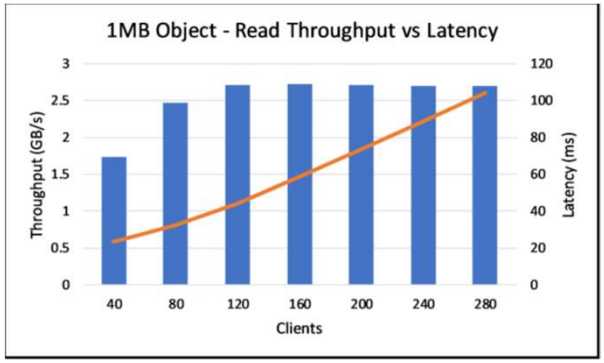
# OceanStor 100D S3 Benchmark

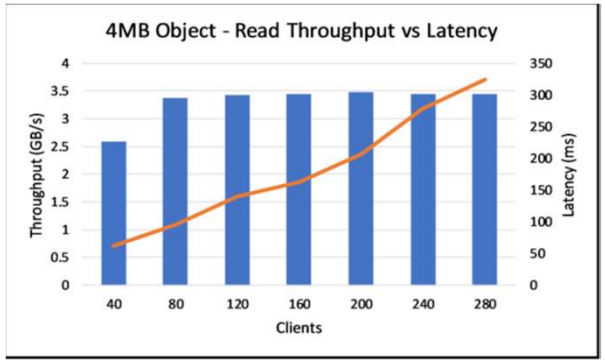
| Operation | Size | Concurrency | Performance per Node (3 node cluster - TaiShan 5280 2*Kunpeng 920@48core EC : 4+2) | | Performance per Node (3 node cluster - TaiShan 2280 2*Kunpeng 920@48core EC : 4+2) | |
|---|---|---|---|---|---|---|
| | | | Sustained performance (80% capacity used) | Initial performance (<50% capacity used) | Sustained performance (80% capacity used) | Initial performance (<50% capacity used) |
| PUT | 32k | 100 | 4000 tps | 5500 tps | 3500 tps | 5000 tps |
| GET | 32k | 100 | 5500 tps | 7500 tps | 5000 tps | 7500 tps |
| PUT | 128k | 100 | 1500 tps | 2000 tps | 1450 tps | 2000 tps |
| GET | 128k | 100 | 5000 tps | 7000 tps | 4000 tps | 4500 tps |
| PUT | 256k | 100 | 1200 tps | 2000 tps | 1100 tps | 1800 tps |
| GET | 256k | 100 | 2200 tps | 2500 tps | 1500 tps | 1800 tps |
| PUT | 512k | 100 | 500 MB/s | 750 MB/s | 400 MB/s | 500 MB/s |
| GET | 512k | 100 | 900 MB/s | 1000 MB/s | 400 MB/s | 500 MB/s |
| PUT | 1M | 100 | 700 MB/s | 900 MB/s | 450 MB/s | 650 MB/s |
| GET | 1M | 100 | 1100 MB/s | 1500 MB/s | 600 MB/s | 750 MB/s |
| PUT | 4M | 100 | 1000 MB/s | 1600 MB/s | 700 MB/s | 900 MB/s |
| GET | 4M | 100 | 1500 MB/s | 2000 MB/s | 900 MB/s | 1100 MB/s |
| PUT | 128M | 100 | 1300 MB/s | 1600 MB/s | 700 MB/s | 900 MB/s |
| GET | 128M | 100 | 1500 MB/s | 1900 MB/s | 1000 MB/s | 1200 MB/s |
| PUT | 1G | 1 | 300 MB/s | 350 MB/s | 250 MB/s | 300 MB/s |
| GET | 1G | 1 | 600 MB/s | 750 MB/s | 500 MB/s | 600 MB/s |

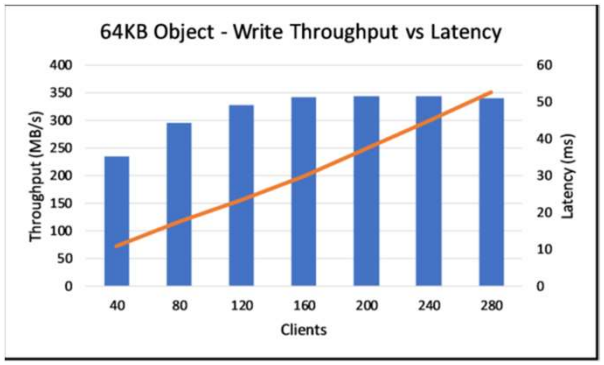
# Ceph 3.2 S3 Benchmark on Commodity Server

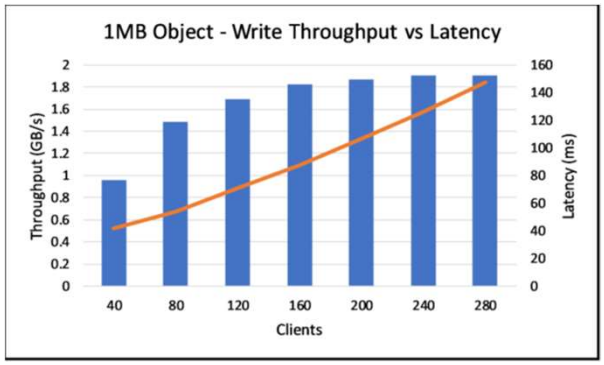
Cluster of 4X Dell R740xd 4U servers, each with 2x 25GbE

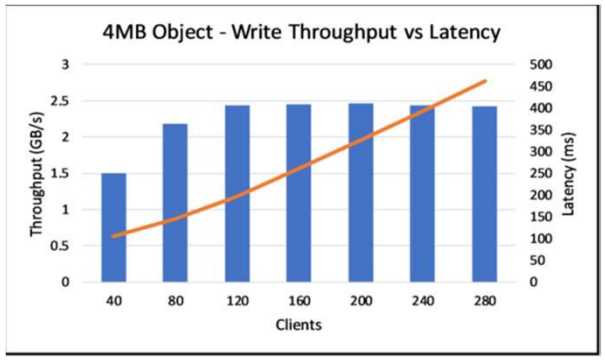



Reference: [Ceph 3.2 S3 benchmark report]



FUTUREWEI Technologies

# S3 Performance Comparison – OceanStor 100D vs Ceph

All data below are ‘normalized’ to reflect single node throughput with 100 threads. Taishan 5280 servers and Dell R740xd are comparable hardware, both are 4U servers with dual sockets and similar amount of disks.

This comparison shows OceanStor 100D can achieve **over 2X performance** on comparable hardware; or achieve better performance on half sized hardware (Taishan 2280 2U server).

| Throughput MB/sec | 100D (5280,4U) | 100D (2280,2U) | Ceph (R740, 4U) | 100D advantage |
|---|---|---|---|---|
| 1MB GET | 1500 | 750 | 630 | **2.38X** |
| 1MB PUT | 900 | 650 | 400 | **2.25X** |
| 4MB GET | 2000 | 1100 | 850 | **2.35X** |
| 4MB PUT | 1600 | 900 | 570 | **2.80X** |

# OceanStor 100D Enterprise Scale Out Object Storage

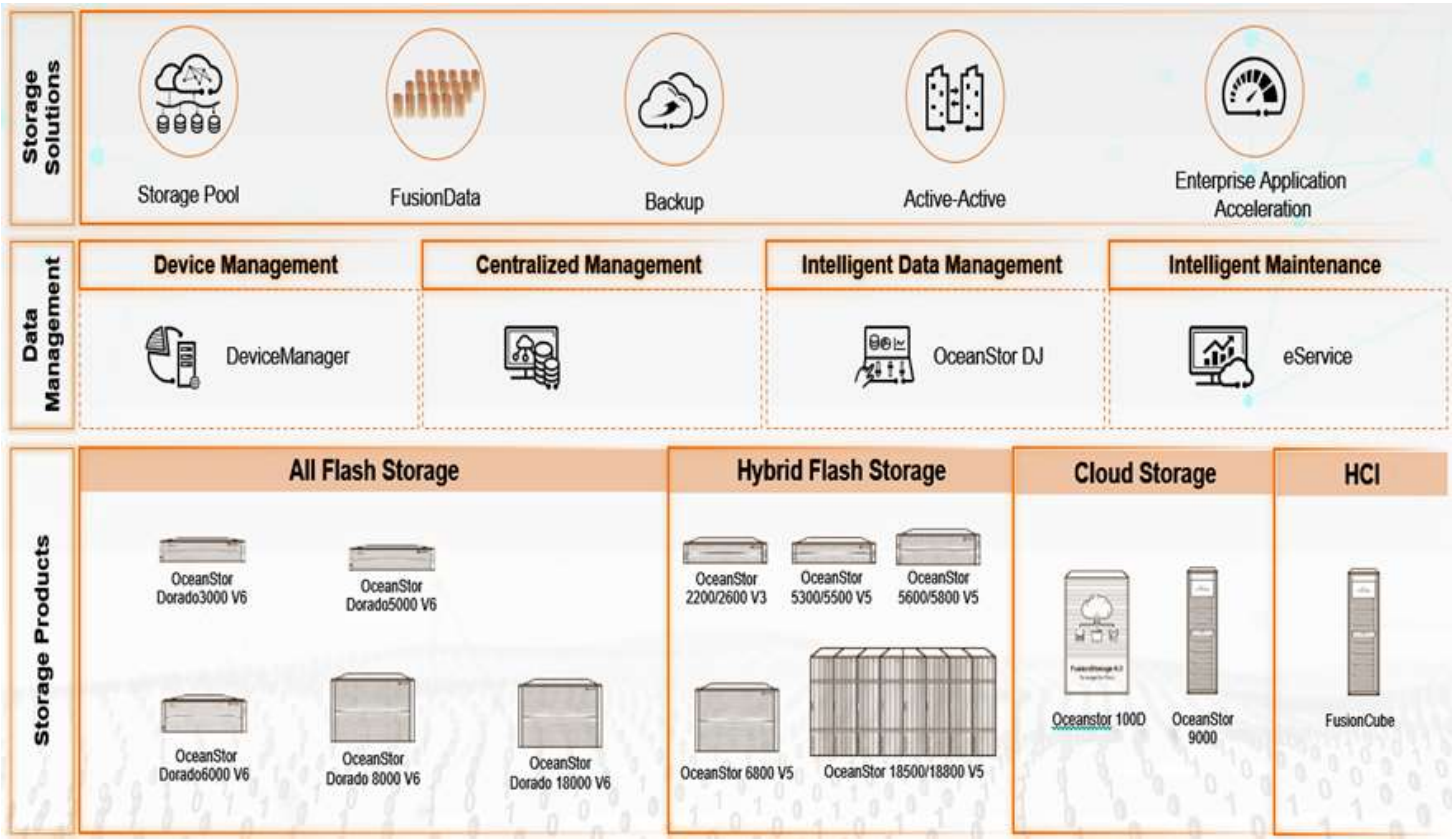
Overview | Software | Performance

System Architecture | Enterprise Features | **Management**

- OceanStor 3-tier management
- OceanStor DeviceManager
- OceanStor cloud AI

FUTUREWEI Technologies

# OceanStor 3 Tiers of Management Software




1. DeviceManager: Embedded/unified storage system management software to provide full storage system operation capacities, and many ease-of-use features
2. OceanStor DJ (DME): Powerful storage provisioning, abstraction, orchestration and automation engine, including multi-vendor support
3. eService: Cloud- and AIOps-based one-stop IT operation platform. Next-generation intelligent O&M platform designed around Huawei's storage system management for self-driving autonomous systems

FUTUREWEI Technologies

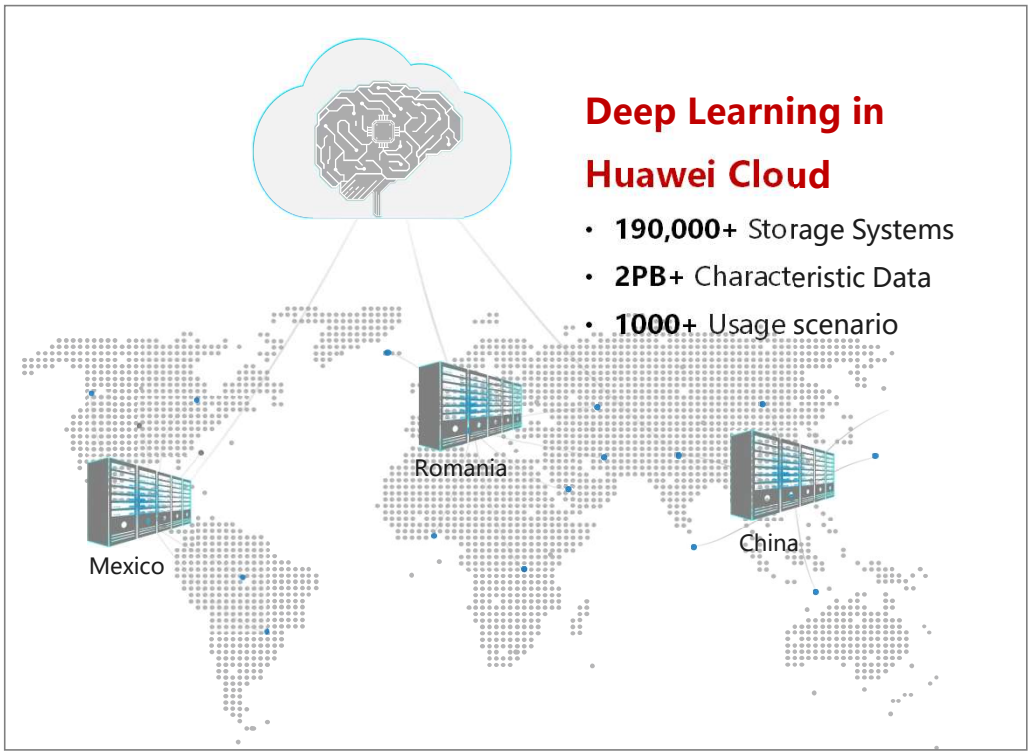
# OceanStor 100D DeviceManager GUI

◆Convenient 2-step disk / node expansion through GUI
◆Deploy EBS module within 2-3 minutes
◆GUI monitor with logging up to 90 days
◆Support GUI/CLI , Restful API, SNMP

◆Visualized disk management
◆Visualized pool and node management
◆Disk info, status and usage details through GUI
◆SMART data and disk locator LED through GUI

FUTUREWEI Technologies

# OceanStor eService Cloud based Management with AI Assistance




**14d**

## Disk Failure Prediction

Industry leading, prevent DU/DL, up to 14 days ahead of time

**60d**

## Usage Prediction

Analyze usage pattern and potential performance bottlenecks, up to 60 days ahead.

**365d**

## Capacity Prediction

- Precisely analyze workload, improve system utilization ratio by 20%
- Suggest optimal expansion plan, lower TCO
- Predict capacity trend up to 1 year ahead

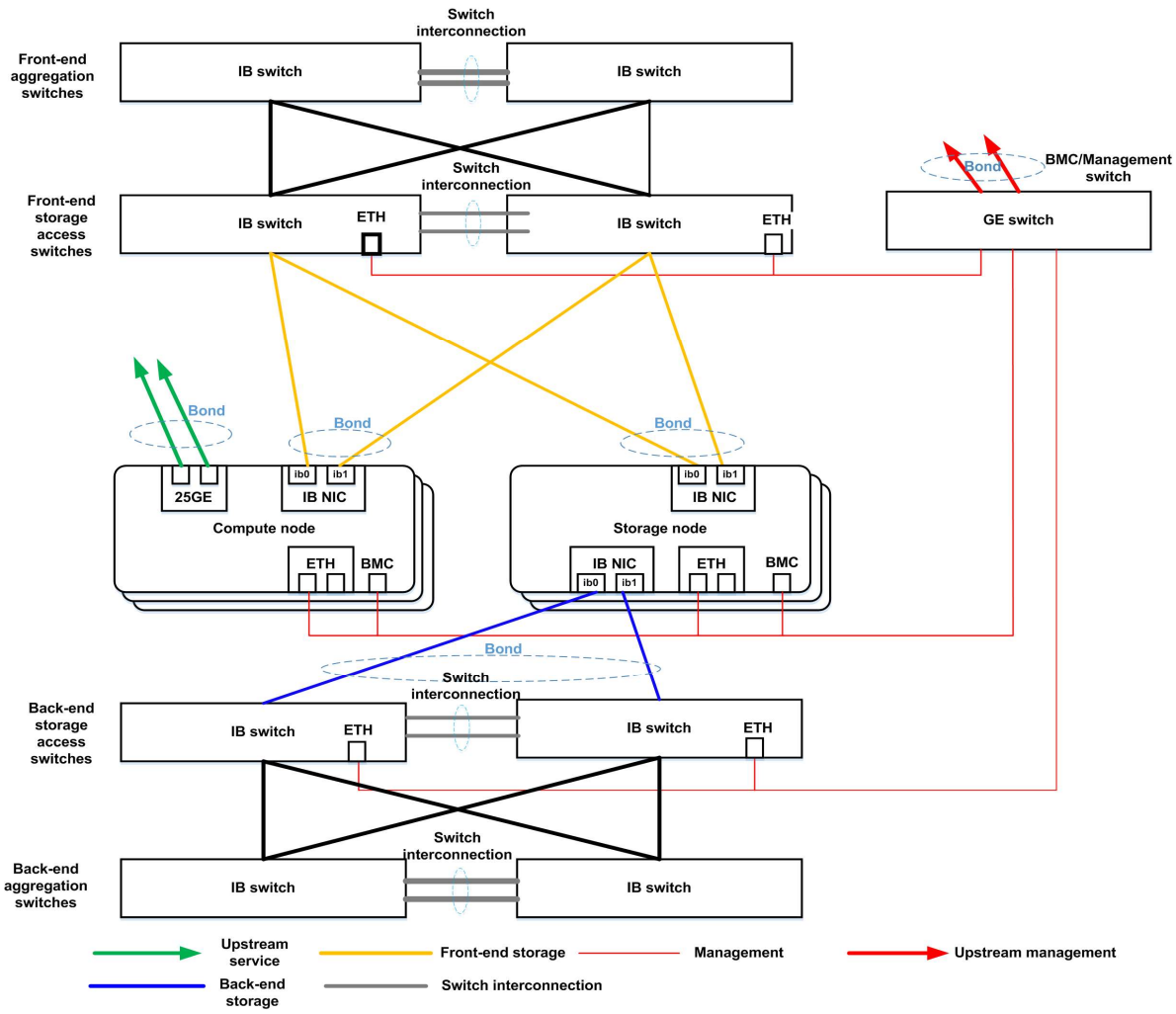
# Thank You.



# Backup Material

# System Architecture Overview

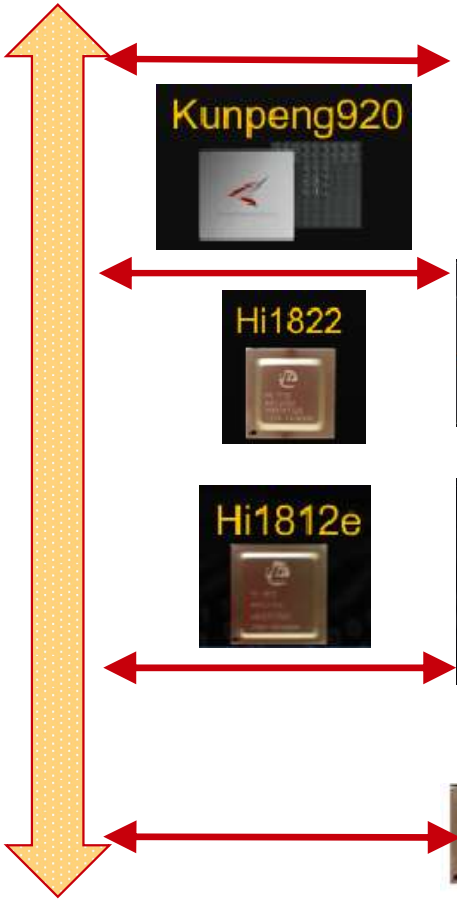
- Huawei 100D storage system features a fully symmetric distributed architecture with component level failover.
- Extensive scale-out capabilities support up to **4096 nodes** in one cluster, and single FS up to **256 nodes** and up to **200PB**.
- Bleeding edge technology integration – **RDMA, NVME, 40/100GE**
- Superior performance in a super-large single file system up to 200PB for shared storage of unstructured data, accessible through **multiple front end protocol** including File (NFS, SMB) and Object (HDFS, S3) **on one copy of data**.



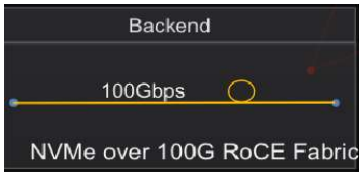

Independent deployment, where front- and back-end storage networks are separated

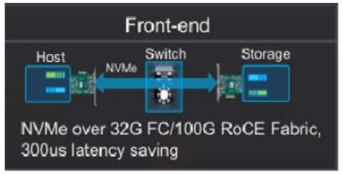
FUTUREWEI Technologies

# Flashlink™ Self-developed Intelligent Chipset

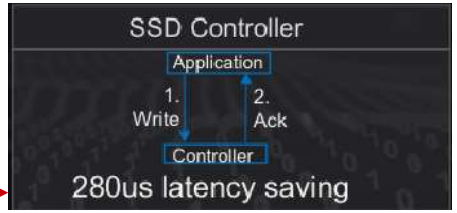



- OceanStor 100D can choose Kunpeng server with up to 48 * 2 cores per node.




- **High speed 100Gps**
- **Low latency**
- RDMA/FC/Eth/InfiniB/TCP




- **0.5ms consistent latency**
- **3M hr MTBF, highest in industry**
- Local wear leveling
- Bad block mgmt.
- Data Redundancy
- Background inspection
- Full SAS/NVMe support

**BMC Chip Hi1710**

- **93% troubleshooting accuracy**

- Sensor Monitoring
- Error diagnostics

FUTUREWEI Technologies

# OceanStor 100D Supports Commodity Servers with both x86 and ARM

## Kunpeng ARM Server

**TaiShan 200(Model 2280) 2U Node**:
2*Kunpeng 920 CPU
12/25 disk slots or 24 NVMe slots
GE / 10GE / 25GE / 100G IB

**TaiShan 200(Model 5280) 4U Node**:
2*Kunpeng 920 CPU
36 disk slots
GE / 10GE / 25GE / 100G IB

## X86 Server

**2288H V5 2U Node**:
2*X86 CPU
12/25 disk slots or 24 NVMe slots
GE / 10GE / 25GE / 100G IB

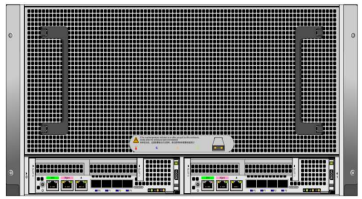
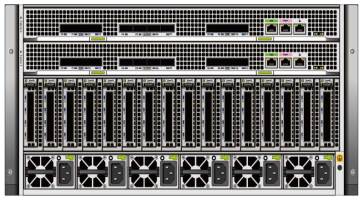
**5288H V5 4U Node**:
2*X86 CPU
30 disk slots
GE / 10GE / 25GE / 100G IB

| Node Model | Optimal Scenario | Hardware Model |
|---|---|---|
| C110 | Capacity | 5288 V5 |
| C100 | | TaiShan 200 (Model 5280) |
| P110 | Performance | 2288H V5 |
| P100 | | TaiShan 200 (Model 2280) |
| F110 | All Flash | 2288H V5 |
| F100 | All Flash | TaiShan 200 (Model 2280) |

FUTUREWEI Technologies

# OceanStor 100D Specialty Hardware

## Optimized for high capacity or high performance

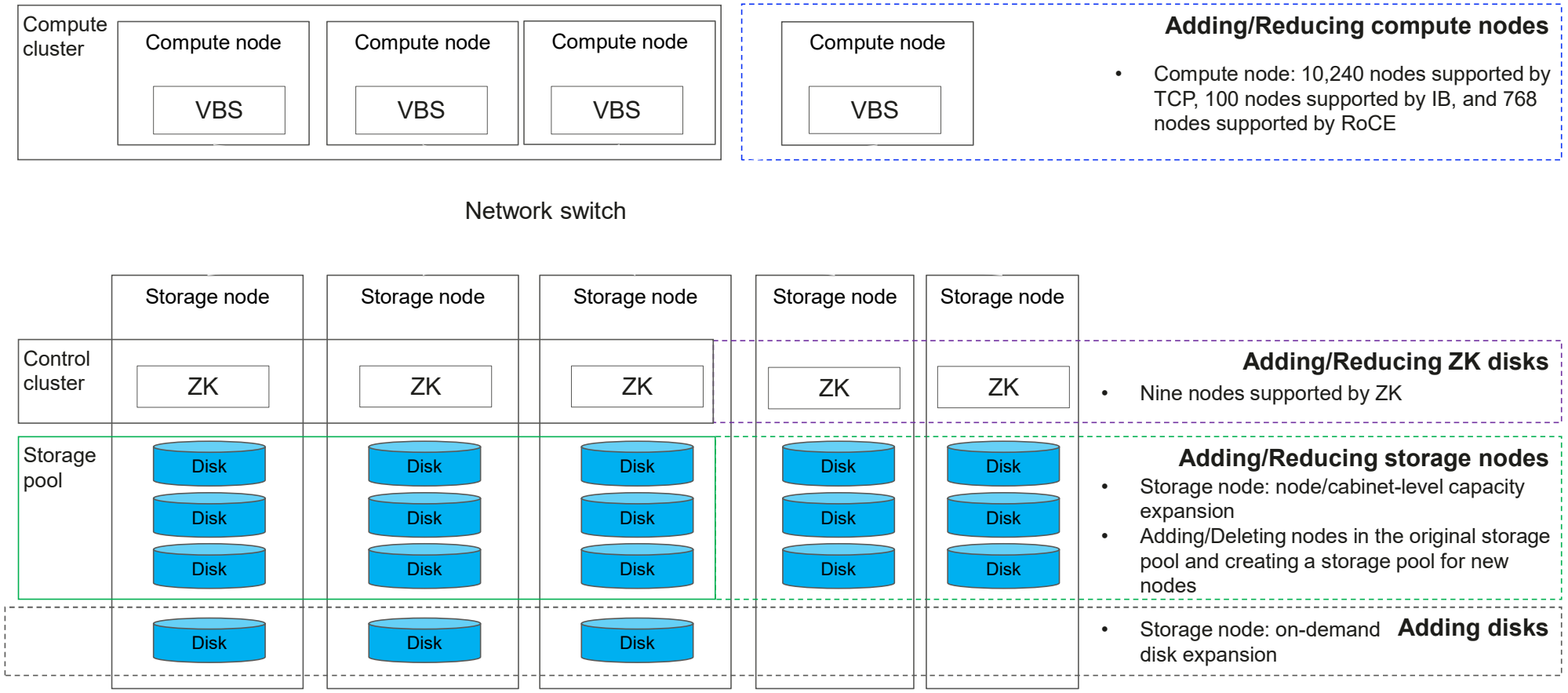
| | Front | Rear |
|---|---|---|
| High-density HDD (Pacific) |  5 U, 2 nodes, 60 HDDs per node, 120 HDDs in total |  Two I/O cards per node, four onboard 25GE ports |
| High-performance all-flash (Atlantic) |  5 U, 8 nodes, 10 SSDs per node, 80 SSDs in total |  Two I/O cards per node, a total of 16 I/O cards. Two switch boards, supporting eight 100GE ports and eight 25GE ports |

In addition to generic commodity server hardware, OceanStor 100D also offers rich selection of specialty hardware that is optimized for different business scenarios.

Pacific Hardware: High Disk Density, Ultimate TCO, Dual-Controller Switchover, High Reliability, Smooth Data Upgrade, and Ever-New Data.

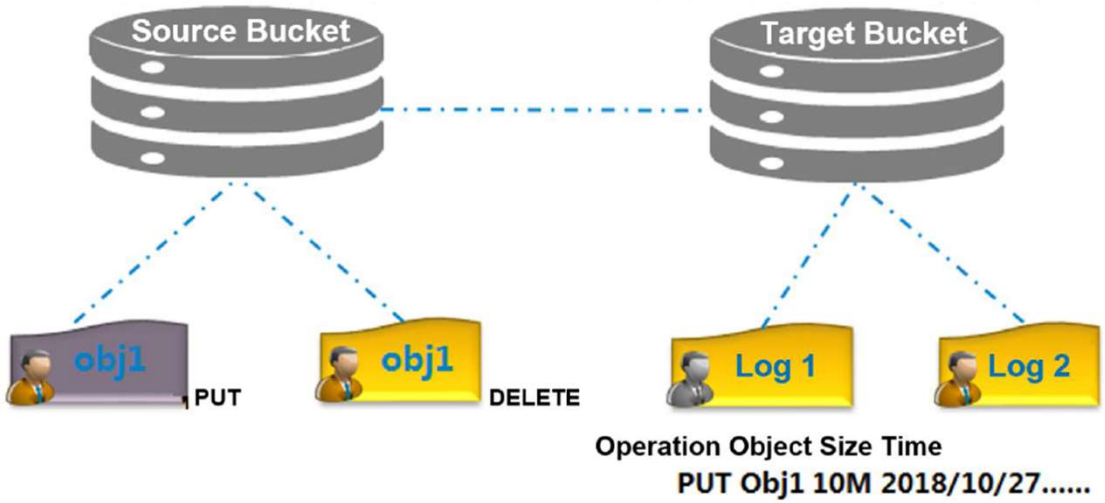
Atlantic Hardware: Flash Storage Native Design with Ultimate Performance, Built-in Switch Networking, Smooth Upgrade, and Ever-New Data.

FUTUREWEI Technologies




FUTUREWEI Technologies

# Bucket access logging




OceanStor 100D object storage implements bucket access logging for security audit and system operation tracing. After access logging is enabled for a bucket, the system records all operations (including PUT, GET, and DELETE) on the bucket in logs, consolidates the logs into log files, and saves the log files in a specified bucket. Besides the operations, requesters, bucket names, request time, response status, and error codes (if any) will also be recorded in the logs.

In the figure, the source bucket is enabled with access logging and the target bucket is where access logs of the source bucket will be stored. Operations, such as PUT and DELETE, will be recorded in logs, and the logs will eventually be stored in the target bucket. You can list the logs in the target bucket and download desired logs.