Cloudian Sizing and Architecture Guidelines

The purpose of this document is to detail the key design parameters that should be considered when designing a Cloudian HyperStore architecture. The primary factors that should be taken into account include the following;

- Capacity
- Performance
- Availability
- Data Protection
- Disaster Recovery
- Cost

This document can be used in conjunction with the sizing excel spreadsheet that will calculate the specification and number of cluster nodes required to deliver the target architecture based upon the defined technical and business requirements.

Capacity

Typically we start with a usable storage capacity target – how much data needs to be stored by the HyperStore cluster.

Translating Usable storage capacity to raw storage required.

Data Protection Overhead

As we are architecting a scale out object storage platform, typical data protection techniques such as RAID and sync/async replication used in traditional file and block storage platforms are not appropriate, as not only do we have to deal with disk failures, but also node and DC failures.

Cloudian HyperStore offers two types of data protection schemes that provides data protection and data availability against hardware (node and disk) failures.

1. Replication Factor (RF) - similar to data mirrors in RAID but data distributed across nodes rather than just disks. It is possible to configure as many copies of data objects as required and is expressed as RF=2 (2copies), RF=3 (3 copies), RF=4 (4 copies), etc.

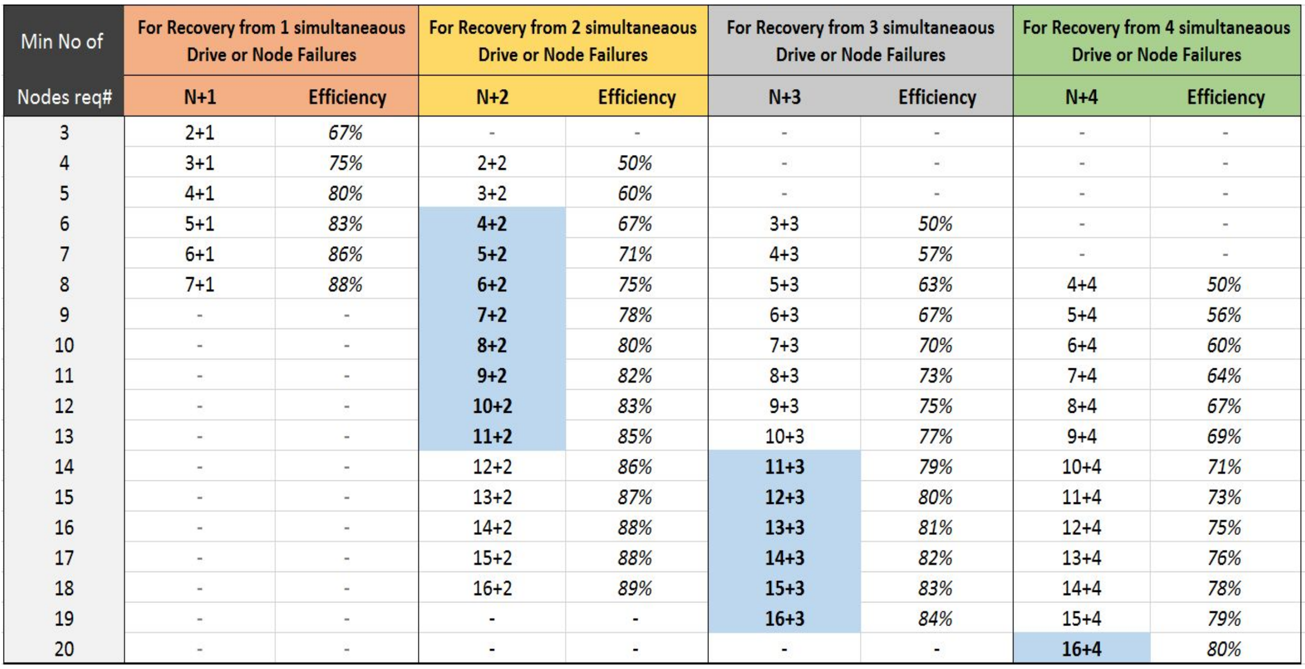
   The chart below illustrates the storage capacity overhead of various RF configurations to assist in calculating the raw storage required to meet the usable storage capacity target once data protection has been factored in.

Min No. of Nodes Req.	RF Config.	Efficiency	Node failure Tolerance
1	RF= 1*	100%	0
2	RF=2*	50%	1
3	RF=3	33%	2
4	RF=4	25%	3

* Not a recommended data protection schema

Erasure Coding (EC) - similar to data parity protection such as RAID5/6 in RAID but data distributed across nodes rather than just disks. It is possible to configure whatever EC scheme as required and is expressed as EC 4+2 (4 data copies & 2 parity), 9+3 (9 data copies & 3 parity), 12+4 (12 data copies & 4 parity), etc.

The chart below illustrates the storage capacity overhead of various EC configurations to assist in calculating the raw storage required to meet the usable storage capacity target once data protection has been factored in.



Min No of	For Recovery from 1 simultaneaous Drive or Node Failures		For Recovery from 2 simultaneaous Drive or Node Failures		For Recovery from 3 simultaneaous Drive or Node Failures		For Recovery from 4 simultaneaous Drive or Node Failures	
Nodes req#	N+1	Efficiency	N+2	Efficiency	N+3	Efficiency	N+4	Efficiency
3	2+1	*67%*	-	-	-	-	-	-
4	3+1	*75%*	2+2	*50%*	-	-	-	-
5	4+1	*80%*	3+2	*60%*	-	-	-	-
6	5+1	*83%*	**4+2**	*67%*	3+3	*50%*	-	-
7	6+1	*86%*	**5+2**	*71%*	4+3	*57%*	-	-
8	7+1	*88%*	**6+2**	*75%*	5+3	*63%*	4+4	*50%*
9	-	-	**7+2**	*78%*	6+3	*67%*	5+4	*56%*
10	-	-	**8+2**	*80%*	7+3	*70%*	6+4	*60%*
11	-	-	**9+2**	*82%*	8+3	*73%*	7+4	*64%*
12	-	-	**10+2**	*83%*	9+3	*75%*	8+4	*67%*
13	-	-	**11+2**	*85%*	10+3	*77%*	9+4	*69%*
14	-	-	12+2	*86%*	**11+3**	*79%*	10+4	*71%*
15	-	-	13+2	*87%*	**12+3**	*80%*	11+4	*73%*
16	-	-	14+2	*88%*	**13+3**	*81%*	12+4	*75%*
17	-	-	15+2	*88%*	**14+3**	*82%*	13+4	*76%*
18	-	-	16+2	*89%*	**15+3**	*83%*	14+4	*78%*
19	-	-	-	-	**16+3**	*84%*	15+4	*79%*
20	-	-	-	-	-	-	**16+4**	*80%*

Note: The use of data protection schemas higher than EC 3+2 are recommended.

Hard drive capacity

The capacity of a hard disk drive, as reported by an operating system to the end user, is smaller than the amount stated by the drive manufacturer; this is due to the operating system using some space and different units used while calculating capacity. There are two different number systems which are used to express units of storage capacity;

Binary 1 KB (kibibyte) is equal to 1024 bytes

Decimal 1 KB (kilobyte) is equal to 1000 bytes.

The storage industry standard is to market storage capacity in decimal representation of a GB which shows greater capacity than you actually can use. A good rule of thumb to calculate actual usable capacity of a hard drive is to apply a 7% overhead on top of the advertised hard drive capacity number.

Similarly there is an overhead of an operating systems file system, which needs to utilize space on the drive to hold the file system meta-data/reference tables. Again, a good rule of thumb is to apply an overhead of 2% for the filesystem overhead on top of the advertised hard drive capacity.

Hard Drive Capacity	Overhead (decimal/binary conversion & file system)		Usable storage available to application
	%	Capacity	
1TB	0.09	90 GB	910 GB
2TB	0.09	180 GB	1,820 GB
4TB	0.09	360 GB	3,640 GB
6TB	0.09	540 GB	5,460 GB
8TB	0.09	720 GB	7,280 GB
10TB	0.09	900 GB	9,100 GB
12TB	0.09	1080 GB	10,920 GB
14TB	0.09	1260 GB	12,740 GB

Now that we understand how much raw storage is required to fulfil our criteria for usable capacity and data protection schemes and what the actual hard disk drives will deliver from a system use point of view we can start to calculate how many drives and nodes we need to architect a cluster.

Node Specification

The typical recommended hardware configuration of a cluster node is detailed within the Cloudian HyperStore software installation manual as detailed below.

- Intel compatible hardware
- Processor: 1 CPU, 8 cores, 2.5GHz
- Memory: 128GB RAM
- Data disks: 12 x 4TB HDD
- OS/Meta-data disks: 2 x 960GB SSD
- RAID: RAID-1 on SSD recommended for the OS/Metadata, JBOD for the Data Drives
- Network: 2x10GbE Ports
- Power Supply: Redundant Platinum level PSUs

They key variables that should be considered for capacity sizing are;

- CPUs – CPU Clock speed is preferred over number of cores, but of course the more cores the better.
- Memory – More memory is required as a node has more drives/capacity
- OS/Meta-data disks - More capacity is required for meta-data storage as a node has more drives/capacity
- Network Interface – For better performance it is highly recommended to deploy nodes with 2 x 10GbE interfaces as a minimum, more ports can be utilized by bonding ports together for resilience and greater bandwidth capability.

A guideline for the per node specification based on Cloudian appliance models is as follows;

Total capacity (per node)	RAM Memory	SSD Storage (per node)	Ratio
48 - 168TB	128GB	960GB	~0.6% - 2%
280 - 490 TB	128GB	1.92TB	~0.4% - 0.7%

Different use cases require considerations around design to ensure performance is optimized as appropriate.

Networking Considerations for node design

Cloudian HyperStore operates with two networks in mind;

- Public Network – How clients will attach to the storage service using the S3 API to read and write data. Typically this network will be exposed to the internet.
- Private Network – Internal cluster network that allows the nodes to communicate with each other for distribution of meta-data, data copies and cluster chatter. This network should be hidden behind the DC firewall and is not for public access.

It is possible to configure both private and public network access over the same NICs and then use VLANs to segregate traffic. This setup may be appropriate when there are two NIC ports, but you would like to configure for resilience. NIC Port bonding can be used to create a single virtual port but connected to 2 VLANs (public and private). Bonding is usually used with Round Robin traffic management and in this setup, a port failure will not impact the node operation.

Typically for performance best practices, you would look to separate the public and private network connections onto separate NIC ports. And to take it one step further it would be possible to utilize 4 ports (maybe 2 dual port NICs) with 2 ports bonded for public and 2 ports bonded for private) which would provide resilience at the port and NIC level for both network connections and enhance the overall bandwidth capability of the cluster.

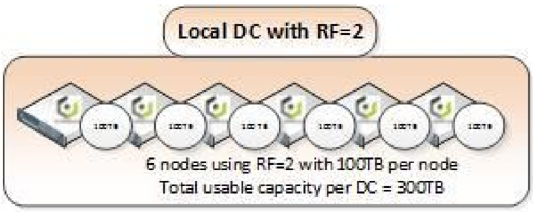
Good design

Best practices for building a HyperStore cluster should take into account that when adding additional nodes to increase capacity and performance, the system will move data between nodes to keep a balanced cluster. Available capacity should be maintained within the cluster to facilitate this process therefore it is recommended to try and maintain 15-20% of free space in the cluster before adding additional nodes.

Example Sizing for single site/DC implementation

For the following examples for both single and multi-site implementation, it is assumed that a node capacity has the drive and FS overheads taken into account and that the 100TB number per node is the usable storage presented by that node to the cluster before data protection scheme is applied.

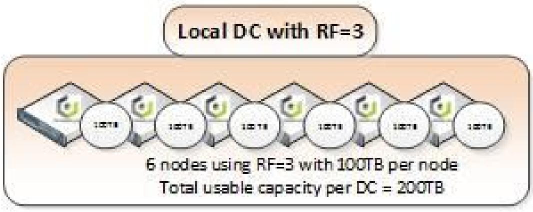
The following diagram illustrates options for data protection strategies for a single site/DC deployment with corresponding storage overheads.



Local DC with RF=2

6 nodes using RF=2 with 100TB per node
Total usable capacity per DC = 300TB

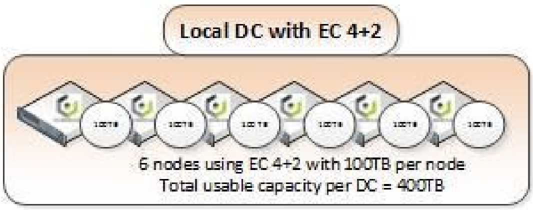
- Raw Capacity = 600TB
- Usable Capacity = 300TB
- No. of Object Copies = 2
- Protection overhead = 0.5
- Utilisation = 50%
- Can lose 1 node



Local DC with RF=3

6 nodes using RF=3 with 100TB per node
Total usable capacity per DC = 200TB

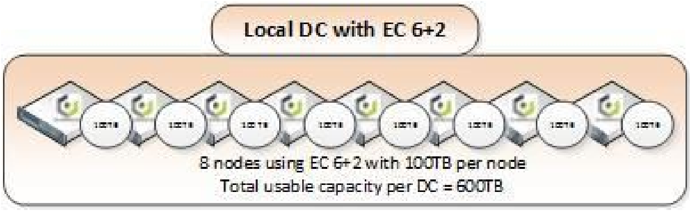
- Raw Capacity = 600TB
- Usable Capacity = 200TB
- No. of Object Copies = 3
- Protection overhead = 0.666
- Utilisation = 33%
- Can lose 2 nodes



Local DC with EC 4+2

6 nodes using EC 4+2 with 100TB per node
Total usable capacity per DC = 400TB

- Raw Capacity = 600TB
- Usable Capacity = 400TB
- No. of Protection Nodes = 2
- Protection overhead = 0.333
- Utilisation = 66%
- Can lose 2 nodes



Local DC with EC 6+2

8 nodes using EC 6+2 with 100TB per node
Total usable capacity per DC = 600TB

- Raw Capacity = 800TB
- Usable Capacity = 600TB
- No. of Protection Nodes = 2
- Protection overhead = 0.25
- Utilisation = 75%
- Can lose 2 nodes

Multi-site/DC implementation

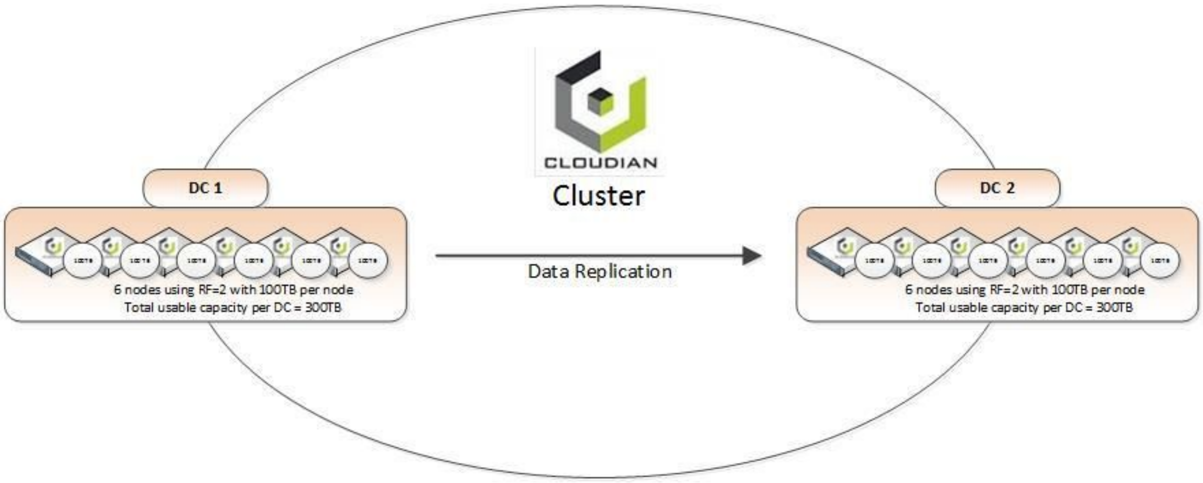
The above sizing guidelines are based on a cluster implemented within a single site/DC/location. Typically Cloudian HyperStore is deployed in multiple locations to provide Disaster Recovery capabilities and scales across geographies. In these cases there are additional sizing considerations required to factor into the design.

For each Data Centre location within the cluster, a separate data protection policy is required locally to each DC. This can be set using either Replication Factor (RF) or Erasure Coding (EC).

Replication Factor over DCs

In this first example, we will consider a setup where we have two DCs configured with 2 object replicas in each DC.

- Local protection in DC1 of losing 1 node
- Local protection in DC2 of losing 1 node
- Cluster wide protection of losing an entire DC


CLOUDIAN
Cluster
DC 1
100TB
100TB
100TB
100TB
100TB
100TB
6 nodes using RF=2 with 100TB per node
Total usable capacity per DC = 300TB
Data Replication
DC 2
100TB
100TB
100TB
100TB
100TB
100TB
6 nodes using RF=2 with 100TB per node
Total usable capacity per DC = 300TB

This is at a cost of local RF=2 protection at both sites of 0.5% and an overhead of the replication of 0.5 of the total storage provisioned.

The Calculation for this is as follows

Total usable storage = Local Data Protection Policy overhead x (No. of Nodes x Node capacity usable) x No. of DCs / No. of DCs

For the diagram example

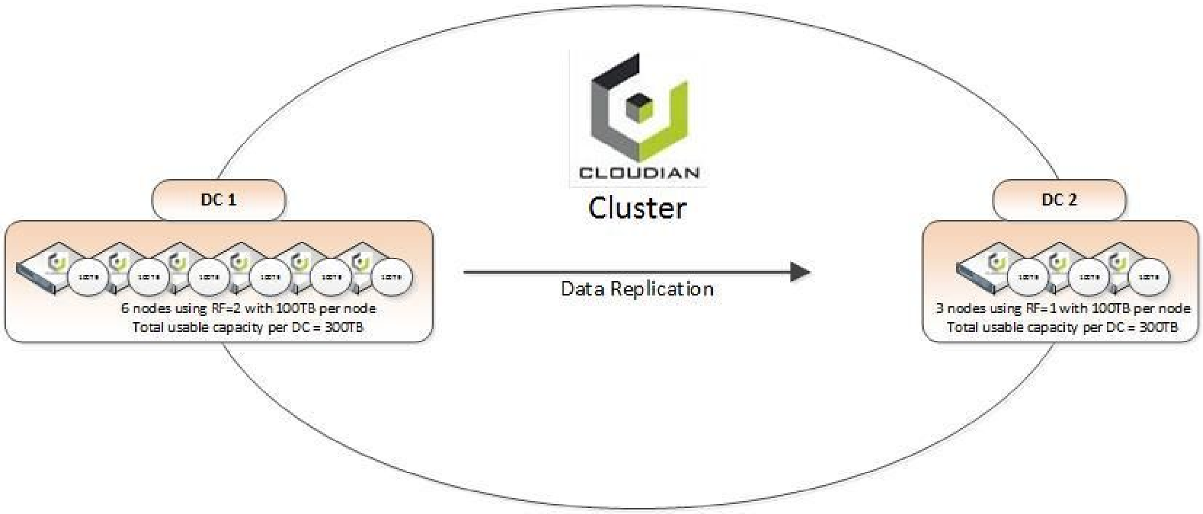
Total usable storage = 0.5 x (6 x 100TB) x 2 /2

Total usable storage = 300TB

Raw Storage = 1200TB

In this second example of using RF over multiple DCs, we will configure RF=2 on the primary DC1 and RF=1 on the DR DC2 site. This reduces the amount of storage required on the second site, but introduces less data protection overall. This scenario may be considered when cost is a significant factor but some form of DR is required.

- Local protection in DC1 of losing 1 node
- No protection in DC2 of losing any nodes
- Cluster wide protection of losing an entire DC


CLOUDIAN
Cluster
DC 1
6 nodes using RF=2 with 100TB per node
Total usable capacity per DC = 300TB
Data Replication
DC 2
3 nodes using RF=1 with 100TB per node
Total usable capacity per DC = 300TB

This is at a cost of local RF=3 protection overall across all sites of 0.666% overhead.

The Calculation for this is as follows

Total usable storage = Total Cluster RF Policy overhead x (No. of Nodes x Node capacity usable)

For the diagram example

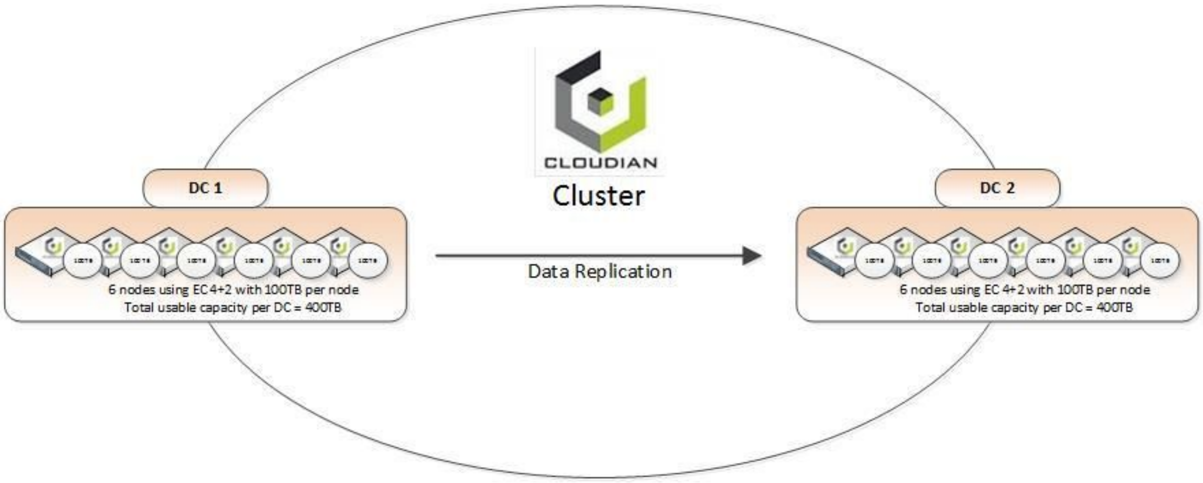
Total usable storage = 0.333 x (6 x 100TB)) + (3 x 100TB))

Total usable storage = 300TB

Raw Storage = 900TB

Replication over Erasure Coding

For example in the following diagram, we are showing 2 DCs, each have 6 nodes and we have configured EC 4+2 protection scheme in each DC. This means that we can lose 2 nodes and still have access to data within the DC at a cost of 2 nodes worth of raw storage.


CLOUDIAN
Cluster
DC 1
6 nodes using EC 4+2 with 100TB per node
Total usable capacity per DC = 400TB
Data Replication
DC 2
6 nodes using EC 4+2 with 100TB per node
Total usable capacity per DC = 400TB

To bring in disaster recovery for protection against losing an entire DC, then we need to deploy a 2nd DC with a similar setup. Now we provide replication over the top of the two EC implementations to deliver an infrastructure that provides;

- Local protection in DC1 of losing 2 nodes
- Local protection in DC2 of losing 2 nodes
- Cluster wide protection of losing an entire DC

This is at a cost of local EC protection at both sites 0.333% and an overhead of the replication of 0.5 of the total storage provisioned.

The Calculation for this is as follows

Total usable storage = Local Data Protection Policy overhead x (No. of Nodes x Node capacity usable) x No. of DCs / No. of DCs

For the diagram example

Total usable storage = 0.333% x (6 x 100TB) x 2 /2

Total usable storage = 400TB

Raw Storage = 1200TB

Multi Region implementation

When operating with multiple Regions, each region is considered to be an independent cluster with its own data protection schemes that do not interact. RF or RF over EC cannot be spread across multiple regions.

Cross Region Replication (CRR) can be used to replicate data from one region to another at the bucket level and this required like for like buckets established in two different regions. A data bucket will have its own local protection schema applied.

Scale-out Performance

As a Cloudian HyperStore cluster can scale performance linearly as more nodes are added, it can be assumed that the performance of a node will increase the overall performance of the cluster as the node is added.

Performance varies greatly depending on a number of factors;

- Read/write ratio of IO
- Sequential/Random nature of IO workload
- Object size
- Bandwidth of network used for both public and private network interfaces of the cluster
- Latency tolerance of application

For this reason, it is best to contact your Cloudian Sales Engineer to assist in determining the correct cluster size to meet the specific workflow and workload requirements.