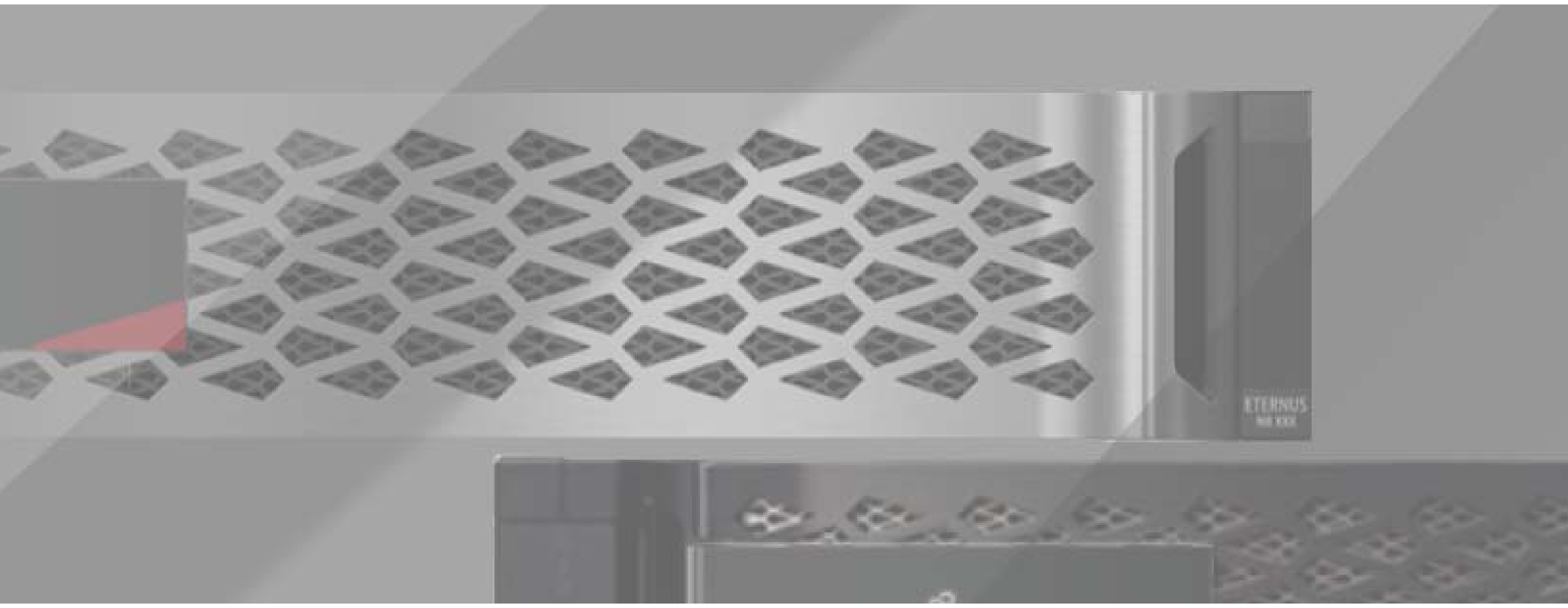
C140-0068-01ENZ3

Fujitsu Storage
ETERNUS AX series All-Flash Arrays

Data Availability and Integrity with All SAN Array (ASA)



FUJITSU

Table of Contents

List of Figures

List of Tables

Preface

This document describes the various data protection and data integrity features of ETERNUS AX series All SAN array (ASA) systems, plus the best practices for designing, implementing, and managing a SAN network to achieve maximum reliability.

First Edition
November 2023

Trademarks

Third-party trademark information related to this product is available at:
https://www.fujitsu.com/global/products/computing/storage/eternus/trademarks.html

Trademark symbols such as ™ and ® are omitted in this document.

About This Manual

Intended Audience

This manual is intended for system administrators who configure and manage operations of the ETERNUS AX, or field engineers who perform maintenance. Refer to this manual as required.

Related Information and Documents

The latest information for the ETERNUS AX is available at:
https://www.fujitsu.com/global/support/products/computing/storage/manuals-list.html

Document Conventions

- Notice Symbols

The following notice symbols are used in this manual:

Caution Indicates information that you need to observe when using the ETERNUS AX. Make sure to read the information.

Note Indicates information and suggestions that supplement the descriptions included in this manual.

1. Introduction

ONTAP is a powerful data-management platform with native capabilities that include inline compression, nondisruptive hardware upgrades, and the ability to import a LUN from a foreign storage system. Up to 12 nodes can be clustered together, simultaneously serving data to your SAN through iSCSI, Fibre Channel (FC), and Nonvolatile Memory Express (NVMe) protocols. In addition, Snapshot technology is an integral part of ONTAP that enables the creation of tens of thousands of backups of critical datasets and the near-instantaneous cloning of datasets. It also offers comprehensive disaster recovery capabilities.

2. All SAN Array

All SAN Array (ASA) systems are built on all-flash systems running ONTAP and provide an enterprise-class SAN solution for customers who want to consolidate and share storage resources for multiple workloads.

ASA systems deliver the following:

- Industry-leading >99.9999% availability
- Massive clusters that scale both up and out
- The best enterprise performance in the industry
- Industry-leading storage efficiency
- Among the most complete cloud-enabled connectivity available
- Cost-effective seamless data protection

ASA builds on the all-flash system platform to deliver continuous SAN availability. ASA systems provide uninterrupted access to data during a planned or unplanned storage failover and deliver streamlined implementation, configuration, and management through a solution that's dedicated only to running SAN workloads. Fujitsu recommends ASA configurations when your requirements include the following:

- Mission-critical workloads such as databases that must have symmetric active-active paths from hosts to storage
- Preference for a dedicated system to isolate SAN workloads

All-flash systems running ONTAP are also suitable for the following customers who:

- Need to scale out SAN clusters to up to 12 nodes.
- Do not have a requirement for active-active SAN path management.
- Prefer a cluster that supports unified protocols to support mixed NAS and SAN workloads.

3. ASA Architecture: Data Availability and Data Integrity

There are two fundamental requirements for any storage system: make sure that data is protected and make sure that data is available.

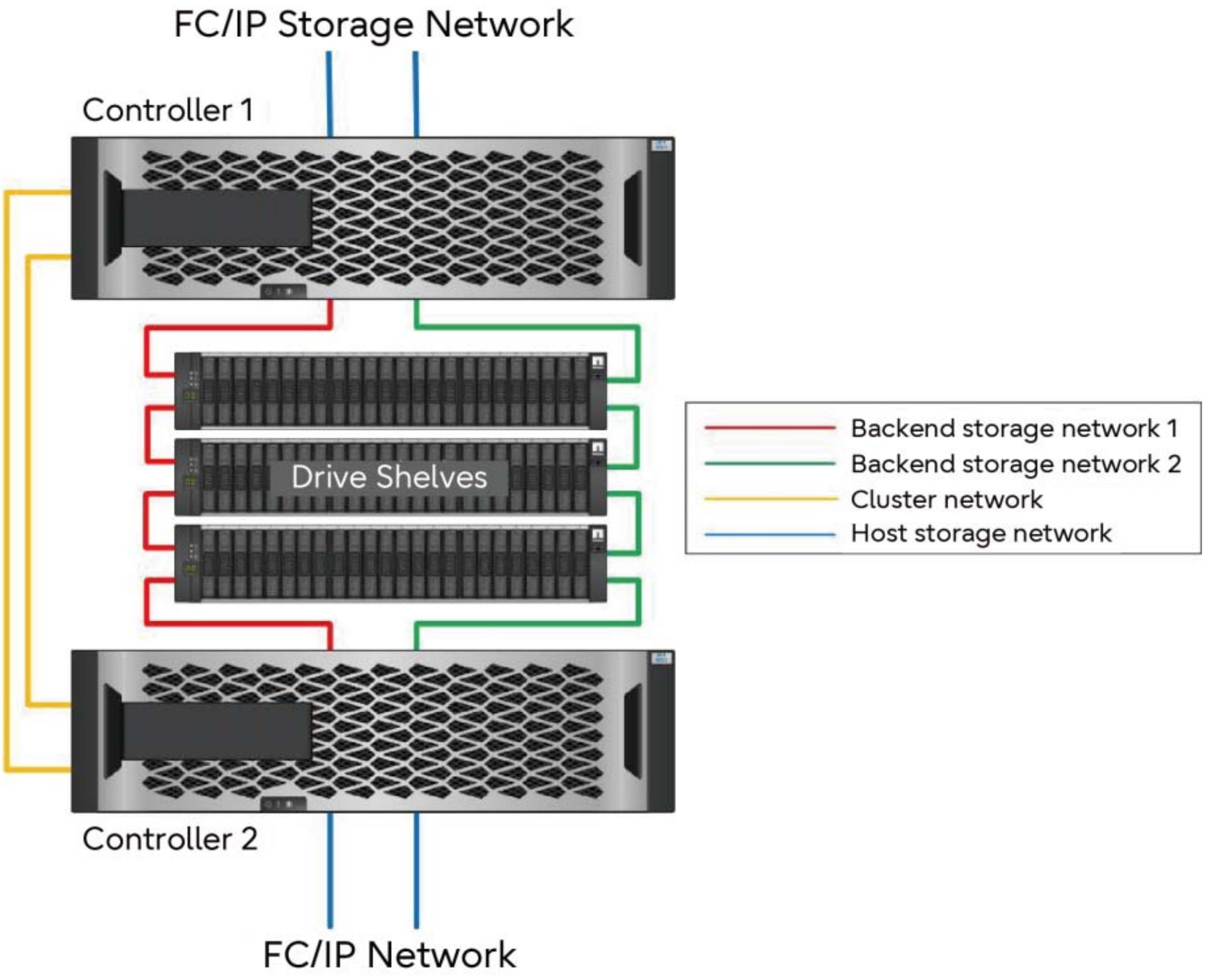
High Availability

A complete description of ONTAP high availability features is beyond the scope of this document. However, as with data protection, a basic understanding of this functionality is important when designing a database infrastructure.

HA Pairs

The basic unit of high availability is the HA pair.

Figure 1 HA pair



NVRAM

Each pair contains redundant links to support replication of NVRAM data. NVRAM is not write cache. The RAM inside a controller serves as the write cache. The purpose of NVRAM is to temporarily journal data as a safeguard against unexpected system failure. In this respect, it is similar to a database transaction log. Both NVRAM and a database transaction log are used to store data quickly, allowing changes to data to be committed as fast as possible. The update to the persistent data on drives does not take place until later during a process called a checkpoint. Neither NVRAM data nor database redo logs are read during normal operations.

If a controller fails abruptly, there are likely to be pending changes stored in NVRAM that have not yet been written to the drives. The partner controller detects the failure, takes control of the drives, and applies the required changes that have been stored in NVRAM.

Cabling Redundancy

The diagram above is an example of HA-pair cabling, but the precise layout varies based on controller and drive type. In all cases, redundant data paths exist, whether they are physical cables or electrical traces on a backplane within a single chassis. A controller has at least two paths to other controllers in the cluster to replicate changes via NVRAM and to facilitate intra-cluster communications for servicing I/O operations or to nondisruptively relocate data within the cluster.

Power Redundancy

All controllers, drive shelves, and other components have redundant power supplies. Systems are typically deployed in server racks with dual PDUs (Power Distribution Units), each attached to a different UPS-protected circuit in the data center.

Takeover and Giveback

Takeover and giveback refer to the process of transferring responsibility for storage resources between nodes in an HA pair. There are two aspects to takeover and giveback:

- Management of the network connectivity used by hosts to access the storage system
- Management of the drives within the storage system

Network interfaces supporting SAN block protocols such as iSCSI and FC are not immediately relocated during takeover and giveback on an ASA system. If a controller suddenly fails, the partner controller continues to serve data using its interfaces. Later in the process, the failed interfaces are brought online on the partner controller either by moving the IP address (iSCSI LIF failover) or relocating the HBA WWN (FCP and NVMe/FC Persistent Ports) so the host does no longer sees any failed paths to storage.

Note

Additional paths to additional controllers can also be configured to support relocating data between nodes in a larger cluster, but this is not part of the HA process.

The second aspect of takeover and giveback is the transfer of drive ownership. The exact process depends on multiple factors including the reason for the takeover or giveback and the command line options issued. The goal is to perform the operation as efficiently as possible. Although the overall process might appear to require a few minutes, the actual moment in which ownership of the drive is transitioned from node to node can generally be measured in seconds.

Takeover Triggers

Takeovers can occur for a variety of reasons, including the following:

- Manual initiation of a takeover with the `storage failover takeover` command.
- A software or system failure occurs that leads to a controller panic. After the panic completes and the controller reboots, the storage resources are given back, returning the system to normal.
- A controller has a complete system failure, such as loss of power, and cannot reboot.
- A partner controller fails to receive a heartbeat message. This situation could happen if the partner experiences a hardware or software failure that does not result in a panic but still prevents it from functioning correctly.

NDO

The term "nondisruptive operations" includes the ability to not only handle a sudden controller failure but also to allow online upgrades and maintenance of a controller. Once a controller has ceded responsibility for data services to its partner, administrators can then upgrade the ONTAP OS, replace failed hardware, add new adapters, or even update the controller itself.

Takeover Time

With the Fujitsu ASA, active-active paths through both controllers means that the host OS does not need to wait for active paths to come down before activating alternative paths. The host was already using all paths on all controllers and the host always has active paths whether the system is in a steady state or performing a controller failover operation.

In addition, ASA includes a unique ability that significantly accelerates the SAN failover process. Each controller constantly replicates key LUN metadata to its partner. Each controller is thus prepared to immediately begin serving data upon sudden failure of its partner, even before the failover process is complete, because it already has the core information required to begin using the drives previously managed by the failed controller.

Table 1 Takeover times

Takeover Type	IO Resumption Time
Planned takeover	2 - 3 seconds
Unplanned takeover	2 - 3 seconds

This time reflects the complete I/O resumption time at the operating system. The takeover time is quicker if you only measure the storage system's ability to respond to IO, but a more important number is the complete IO resumption time as viewed from the host.

Data Integrity

Logical data protection within ONTAP consists of three key requirements:

- Data must be protected against data corruption.
- Data must be protected against drive failure.
- Changes to data must be protected against loss.

These three needs are discussed in the following sections.

Network Corruption: Checksums

The most basic level of data protection is the checksum, which is a special error-detecting code stored alongside data. Corruption of data during network transmission is detected with the use of a checksum and, in some instances, multiple checksums.

For example, an FC frame includes a form of checksum called a cyclic redundancy check (CRC) to make sure that the payload is not corrupted in transit. The transmitter sends both the data and the CRC of the data. The receiver of an FC frame recalculates the CRC of the received data to make sure that it matches the transmitted CRC. If the newly computed CRC does not match the CRC attached to the frame, the data is corrupt, and the FC frame is discarded or rejected. An iSCSI I/O operation includes checksums at the TCP/IP and Ethernet layers, and, for extra protection, it can also include optional CRC protection at the SCSI layer. Any bit corruption on the wire is detected by the TCP layer or IP layer, which results in retransmission of the packet. As with FC, errors in the SCSI CRC result in a discard or rejection of the operation.

Drive Corruption: Checksums

Checksums are also used to verify the integrity of data stored on drives. Data blocks written to drives are stored with a checksum function that yields an unpredictable number that is tied to the original data.
When data is read from the drive, the checksum is recomputed and compared to the stored checksum. If it does not match, then the data has become corrupt and must be recovered by the RAID layer.

Data Corruption: Lost Writes

One of the most difficult types of corruption to detect is a lost or a misplaced write. When a write is acknowledged, it must be written to the media in the correct location. In-place data corruption is relatively easy to detect by using a simple checksum stored with the data. However, if the write is simply lost, then the prior version of data might still exist on the media, and the checksum of the underlying blocks would be correct. If the write is placed at the wrong physical location, the associated checksum would once again be valid for the stored data, even though the write has destroyed other data.

The solution to this challenge is as follows:

- A write operation must include metadata that indicates the location where the write is expected to be found.
- A write operation must include some sort of version identifier.

When ONTAP writes a block, it includes data on where the block belongs. For example, if a subsequent read identifies a block, but the metadata indicates that it belongs at location 123 when it was found at location 456, then the write has been misplaced.

Detecting a wholly lost write is more difficult. The explanation is very complicated, but essentially ONTAP stores metadata in a way that a write operation results in updates to two different locations on the drives. If a write is lost, a subsequent read of the data and associated metadata shows two different version identities. This indicates that the write was not completed by the drive.

Lost and misplaced write corruption is exceedingly rare, but, as drives continue to grow and data sets push into exabyte scale, the risk increases. Lost write detection should be included in any storage system supporting critical datasets.

Drive Failures: RAID4, RAID DP, and RAID-TEC

If a block of data on a drive is discovered to be corrupt or the entire drive fails and is wholly unavailable, the data must be reconstituted. This is performed in ONTAP with parity drives. Data is striped across multiple data drives, and then parity data is generated. This is stored separately from the original data.

ONTAP originally used RAID 4, which uses a single parity drive for each group of data drives. The result was that any one drive in the group could fail without resulting in data loss. If the parity drive failed, no data was damaged and a new parity drive could be constructed. If a single data drive failed, the remaining drives could be used with the parity drive to regenerate the missing data.

When drives were small, the statistical chance of two drives failing simultaneous was negligible. As drive capacities have grown, so has the time required to reconstruct data after a drive failure. This has increased the window in which a second drive failure would result in data loss. In addition, the rebuild process creates a lot of additional I/O on the surviving drives. As drives age, the risk of the additional load leading to a second drive failure also increases. Finally, even if the risk of data loss did not increase with the continued use of RAID 4, the consequences of data loss would become more severe. The more data that would be lost in the event of a RAID-group failure, the longer it would take to recover the data, extending business disruption.

These issues led Fujitsu to develop RAID DP technology, a variant of RAID 6. This solution includes two parity drives, meaning that any two drives in a RAID group can fail without creating data loss. Drives have continued to grow, which eventually led Fujitsu to develop the RAID-TEC technology, which introduces a third parity drive.

Some historical SAN best practices recommend the use of RAID 1+0, also known as striped mirroring. This offers less data protection than even RAID DP because there are multiple two-drive failure scenarios, whereas in RAID DP there are none.

There is also some legacy SAN best practice documentation that indicates RAID 1+0 is preferred to RAID 4/5/6 options due to performance concerns. These recommendations sometimes refer to a RAID penalty. Although these recommendations are generally correct, they are inapplicable to the implementation of RAID within ONTAP. The performance concern is related to parity regeneration. With traditional RAID implementations, processing a write requires multiple drive reads to regenerate the parity data and complete the write. The penalty is defined as the additional read IOPS required to perform write operations.

ONTAP does not incur a RAID penalty because writes are staged in memory where parity is generated and then written to drive as a single RAID stripe. No reads are required to complete the write operation.

In summary, when compared to RAID 1+0, RAID DP and RAID-TEC deliver much more usable capacity, better protection against drive failure, and no performance sacrifice.

Hardware Failure Protection: NVRAM

Any storage system servicing latency-sensitive workloads must acknowledge write operations as quickly as possible. Furthermore, a write operation must be protected from loss from an unexpected event such as a power failure. This means any write operation must be safely stored in at least two locations.

ASA systems rely on NVRAM to meet these requirements. The write process works as follows:

1. The inbound write data is stored in RAM.
2. The changes that must be made to data on drive are journaled into NVRAM on both the local and partner node. NVRAM is not a write cache; rather it is a journal similar to a database redo log. Under normal conditions, it is not read. It is only used for recovery, such as after a power failure during I/O processing.
3. The write is then acknowledged to the host.

The write process at this stage is complete from the application point of view, and the data is protected against loss because it is stored in two different locations. Eventually, the changes are written to drive, but this process is out-of-band from the application point of view because it occurs after the write is acknowledged and therefore does not affect latency. This process is once again similar to database logging. A change to the database is recorded in the redo logs as quickly as possible, and the change is then acknowledged as committed. The updates to the datafiles occur much later and do not directly affect the speed of processing.

In the event of a controller failure, the partner controller takes ownership of the required drives and replays the logged data in NVRAM to recover any I/O operations that were in-flight when the failure occurred.

Redundancy Failure: NVFAIL

As discussed earlier, a write is not acknowledged until it has been logged into local NVRAM and NVRAM on at least one other controller. This approach makes sure that a hardware failure or power outage does not result in the loss of in-flight I/O. If the local NVRAM fails or the connectivity to other nodes fails, then data would no longer be mirrored.

If the local NVRAM reports an error, the node shuts down. This shutdown results in failover to a partner controller when HA pairs are used. With MetroCluster, the behavior depends on the overall configuration chosen, but it can result in automatic failover to the remote controller.

In any case, no data is lost because the controller experiencing the failure has not acknowledged the write operation. Data loss means loss of an acknowledged write. A key principle of block storage management with filesystems and applications is that an unacknowledged write may or may not exist on persistent storage prior to acknowledgement because there is no way for an OS to know whether a write was lost before it was received by the storage system, or whether it was only the acknowledgement that was lost. Until acknowledgment is received, the state of a write is indeterminate.

4. Data Protection

The prior chapter addresses data availability and data integrity with respect to the storage hardware. An equally important aspect of data availability and integrity is the ability to recover from inevitable user and application errors. Any enterprise that demands 99.9999% uptime from storage systems should also plan a backup and recovery strategy that ensures increasingly larger datasets can be rapidly and reliably recovered.

Data Protection with Snapshot Copies

The foundation of ONTAP data protection software is Snapshot technology. The key values are as follows:

- Simplicity
  A Snapshot copy is a read-only copy of the contents of a container of data at a specific point in time.
- Efficiency
  Snapshot copies require no space at the moment of creation. Space is only consumed when data is changed.
- Manageability
  A backup strategy based on Snapshot copies is easy to configure and manage because Snapshot copies are a native part of the storage OS. If the storage system is powered on, it is ready to create backups.
- Scalability
  Up to 1024 snapshots of a single LUN can be retained locally. For complex datasets, multiple containers of data can be protected by a single, consistent set of Snapshot copies.
  Performance is unaffected, whether a volume contains 1024 Snapshot copies or none.

As a result, protecting a dataset running on ONTAP is simple and highly scalable. Backups do not require movement of data. Therefore, a backup strategy can be tailored to the needs of the business rather than the limitations of network transfer rates, large numbers of tape drives, or expensive drive staging areas.

Data Restoration with ONTAP SnapRestore

Rapid data restoration in ONTAP from a Snapshot copy is delivered by SnapRestore technology. The key values are as follows:

- Individual files or LUNs can be restored in seconds, whether it is a 16TB LUN or a 4KB file.
- An entire container (a FlexVol volume) of LUNs and/or files can be restored in seconds, whether it is 10GB or 100TB of data.

When a critical application is down, critical business operations are down. Tapes can break, and even restores from drive-based backups can be slow to transfer across the network. SnapRestore avoids these problems by delivering near instantaneous restoration of critical datasets. Even petabyte-scale databases can be completely restored with just a few minutes of effort.

Remote Data Protection with SnapMirror

SnapMirror is an easy-to-manage, highly scalable, highly efficient replication technology. It can also replicate not only the data but the snapshots as well. You can selectively store some or all of your backups at remote locations or even in the cloud. The ensures backups are available no matter what, and the efficiency of snapshot technologies minimizes the demands on network infrastructure as well as storage capacity.

Data Restoration with ONTAP FlexClone

Not all datasets can be simply restored in-place. Sometimes you need to repair a dataset rather than restoring it. The same technology that allows you to restore data also allows you to clone it without affecting the current data.

- Individual files or LUNs can be cloned in seconds, whether it is a 16TB LUN or a 4KB file.
- An entire container (a FlexVol volume) of LUNs and/or files can be cloned in seconds, whether it is 10GB or 100TB of data.
- Clones can be made from any copy of the data: local, remote, or in the Cloud.

Administrators can then examine the clone, extract data as required, and repair the dataset.

5. Disaster Recovery

A single ASA system provides maximum availability at the hardware level and rapid restore capability to address user and application errors, but what about disaster? How do you provide constant data availability if the power is out completely or a site is destroyed?

ASA systems support two options: MetroCluster and SnapMirror Business Continuity.

Note

Snapshots on an ASA system can also be replicated asynchronously for disaster recovery needs if data loss during a disaster is tolerated.

MetroCluster Technology

MetroCluster delivers a highly available, zero data-loss solution for mission-critical workloads. In addition, integrated solutions such as MetroCluster simplify today's complicated, scale-out enterprise applications and virtualization infrastructures. MetroCluster replaces multiple external data protection products and strategies with one simple, central storage system that provides integrated backup, recovery, disaster recovery, and high availability (HA) within a single clustered storage system.

HA with MetroCluster

MetroCluster replication is based on SyncMirror technology, which is designed to efficiently switch into and out of synchronous mode. This capability meets the requirements of customers who demand synchronous replication, but who also need high availability for their data services. For example, if connectivity to a remote site is severed, it is generally preferable to have the storage system continue operating in a nonreplicated state.

Many synchronous replication solutions are only capable of operating in synchronous mode. This type of all-or-nothing replication is sometimes called domino mode. Such storage systems stop serving data rather than allowing the local and remote copies of data to become unsynchronized. If replication is forcibly broken, resynchronization can be extremely time consuming and can leave a customer exposed to complete data loss during the time that mirroring is reestablished.

Not only can SyncMirror seamlessly switch out of synchronous mode if the remote site is unreachable, it can also rapidly resync to an RPO = 0 state when connectivity is restored. The stale copy of data at the remote site can also be preserved in a usable state during resynchronization, which ensures that local and remote copies of data exist at all times.

MetroCluster and SyncMirror

Synchronous replication in ONTAP is delivered by SyncMirror. At the simplest layer, SyncMirror creates two complete sets of RAID-protected data in two different locations. They could be in adjoining rooms within a data center, or they could be located many kilometers apart.

SyncMirror is fully integrated with ONTAP and operates just above the RAID level. Therefore, all the usual ONTAP features, such as Snapshot copies, SnapRestore, and FlexClone, work seamlessly. It is still ONTAP. It just includes an additional layer of synchronous data mirroring.

A collection of ONTAP controllers managing SyncMirror data is called MetroCluster. Many configurations are available, and the primary purpose of MetroCluster is to provide high-availability access to synchronously mirrored data in a variety of typical and disaster recovery failure scenarios.

The key values of data protection with MetroCluster and SyncMirror are as follows:

- In normal operations, SyncMirror delivers guaranteed synchronous mirroring across locations. A write operation is not acknowledged until it is present on nonvolatile media on both sites.
- If connectivity between sites fails, SyncMirror automatically switches into asynchronous mode to keep the primary site serving data until connectivity is restored. When restored, it delivers rapid resynchronization by efficiently updating the changes that have accumulated on the primary site. Full reinitialization is not required.

SnapMirror is also fully compatible with systems based on SyncMirror. For example, a primary database might be running on a MetroCluster cluster spread across two geographic sites. This database can also replicate backups to a third site as long-term archives or for the creation of clones in a DevOps environment.

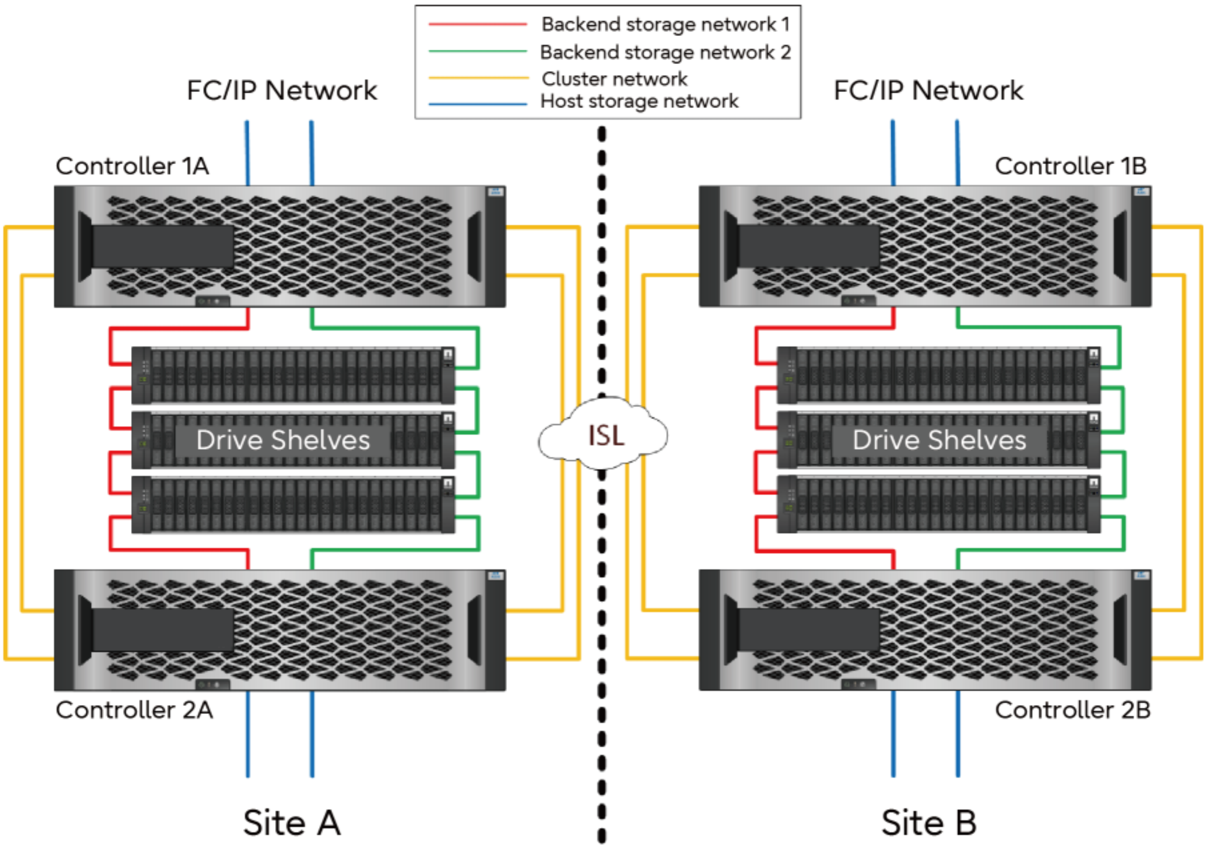
MetroCluster Architecture

A complete explanation of MetroCluster is beyond the scope of this document, but you should understand its core availability functions. The following sections use an IP-based MetroCluster for illustration. Because high-speed and low-latency IP circuits are more readily available these days, and the infrastructure requirements are simpler, most customers choose IP connectivity.

For additional information, see the official ONTAP documentation.

MetroCluster systems using IP connectivity are configured with HA pairs on each site.

Figure 2 MetroCluster IP basic architecture



MetroCluster Resiliency Features

As shown in the figures above, there are no single points of failure in a MetroCluster solution:

- Each controller has two independent paths to the drive shelves on the local site.
- Each controller has two independent paths to the drive shelves on the remote site.
- Each controller has two independent paths to the controllers on the opposite site.
- In the HA-pair configuration, each controller has two paths to its local partner.

In summary, any one component in the configuration can be removed without compromising the ability of MetroCluster to serve data. The only difference in terms of resiliency between the two options is that the HA-pair version is still an overall HA storage system after a site failure.

Site Failure Protection: NVRAM and MetroCluster

MetroCluster extends NVRAM data protection by replicating NVRAM data to both the local partner and a remote partner. A write is not acknowledged until it is replicated to all partners.

This architecture protects in-flight I/O from site failure. This process is not involved with drive-level data replication. The controller that owns the aggregates is responsible for data replication by writing to both plexes in the aggregate, but there still must be protection against in-flight I/O loss in the event of site loss. Replicated NVRAM data is only used if a partner controller must take over for a failed controller.

Site and Shelf Failure Protection: SyncMirror and Plexes

SyncMirror is a mirroring technology that enhances, but does not replace, RAID DP or RAID-TEC. It mirrors the contents of two independent RAID groups. The logical configuration is as follows:

- Drives are configured into two pools based on location. One pool is composed of all drives on site A, and the second pool is composed of all drives on site B.
- A common pool of storage, known as an aggregate, is then created based on mirrored sets of RAID groups. An equal number of drives is drawn from each site. For example, a 20-drive SyncMirror aggregate would be composed of 10 drives from site A and 10 drives from site B.
- Each set of drives on a given site is automatically configured as one or more fully redundant RAID DP or RAID-TEC groups, independent of the use of mirroring. This use of RAID underneath mirroring provides data protection even after the loss of a site.

Figure 3 SyncMirror

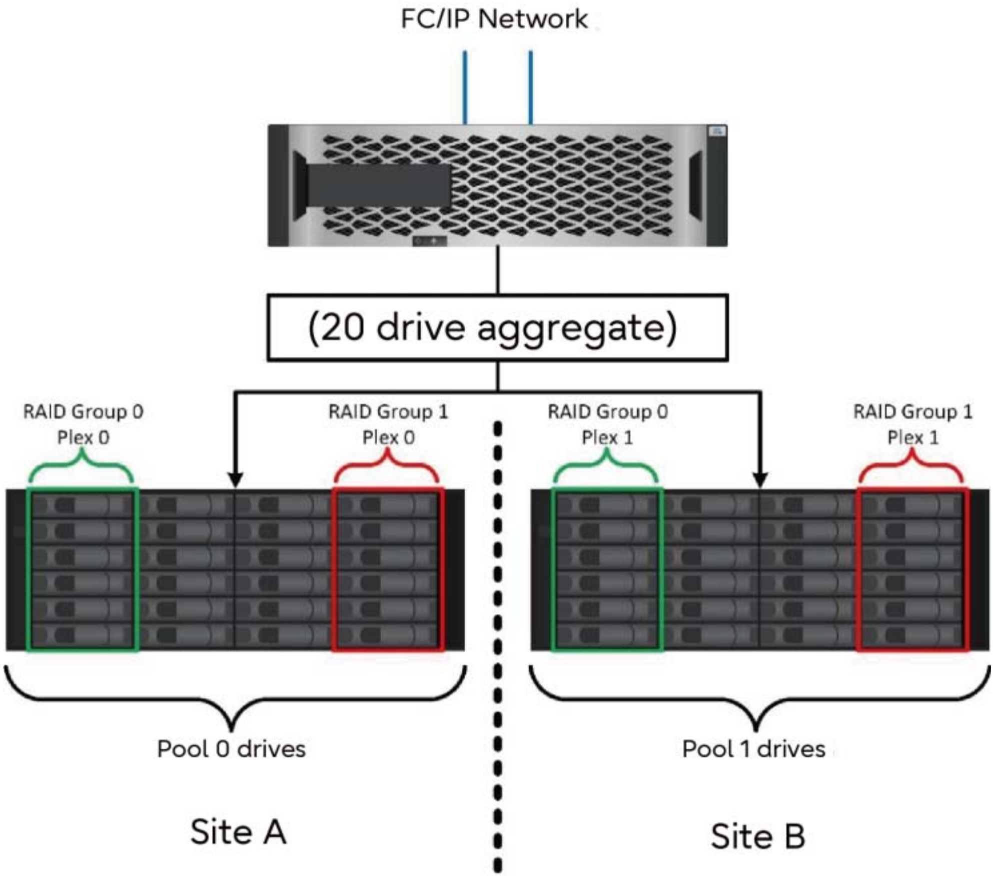


Figure 3 illustrates a sample SyncMirror configuration. A 24-drive aggregate was created on the controller with 12 drives from a shelf allocated on site A and 12 drives from a shelf allocated on site B. The drives are grouped into two mirrored RAID groups. RAID group 0 includes a 6-drive plex on site A mirrored to a 6-drive plex on site B. Likewise, RAID group 1 includes a 6-drive plex on site A mirrored to a 6-drive plex on site B.

SyncMirror is normally used to provide remote mirroring with MetroCluster systems, with one copy of the data at each site. On occasion, it has been used to provide an extra level of redundancy in a single system. In particular, it provides shelf-level redundancy. A drive shelf already contains dual power supplies and controllers and is overall little more than sheet metal, but in some cases the extra protection might be warranted. For example, one Fujitsu customer has deployed SyncMirror for a mobile real-time analytics platform used during automotive testing. The system was separated into two physical racks supplied with independent power feeds and independent UPS systems.

Hardware-Assisted Takeover

The service processor is an out-of-band management device embedded in ONTAP systems. It is accessed by its own IP address and is used for direct console access and other management functions irrespective of whether the controller is operational.

ONTAP by itself can trigger a takeover of a failed node after it no longer detects the heartbeat from the partner node, but there are timeouts involved. Hardware-assisted takeover uses the service process to speed up the takeover process by more quickly detecting failures and immediately initiating the takeover. It does not wait for ONTAP to recognize that the partner's heartbeat has stopped.

Switchover and Switchback

The terms switchover and switchback refer to the process of transitioning volumes between remote controllers in a MetroCluster configuration. This process only applies to the remote nodes. When MetroCluster is used in a four-volume configuration, local node failover is the same takeover and giveback process described previously.

Planned Switchover and Switchback

A planned switchover or switchback is similar to a takeover or giveback between nodes. The process has multiple steps and might appear to require several minutes, but what is actually happening is a multiphase graceful transition of storage and network resources. The moment when control transfers occurs much more quickly than the time required for the complete command to execute.

The primary difference between takeover-giveback and switchover-switchback is with the effect on SAN connectivity. With local takeover-giveback, a host experiences the loss of all SAN paths to the local node and relies on its native MPIO to change over to available alternate paths. Ports are not relocated. With switchover and switchback, the virtual FC target ports on the controllers transition to the other site. They effectively cease to exist on the SAN for a moment and then reappear on an alternate controller.

ONTAP Mediator with MetroCluster IP

The ONTAP Mediator is used with MetroCluster IP and certain other ONTAP solutions. It functions as a traditional tiebreaker service.

Its primary function is determining whether NVRAM and SyncMirror are synchronized. With MetroCluster, NVRAM and the underlying aggregate plexes are in sync, it is safe to proceed with switchover without risk of data loss because it makes sure that the data at each site is identical.

ONTAP does not permit a failover or switchover when the data is out of sync unless the failover or switchover is forced. Forcing a change in conditions in this manner acknowledges that data might be left behind in the original controller and that data loss is acceptable.

SnapMirror Business Continuity

MetroCluster is an ideal solution when an entire environment needs zero data loss in the event of a disaster with minimal disruption to services. However, not all customers want RPO=0 data protection for an entire array. Sometimes only select datasets require RPO=0 synchronous data protection.

SM-BC (SnapMirror Business Continuity) was introduced in ONTAP 9.9.1 to address this need. SM-BC and SM-S (SnapMirror Synchronous) share a replication engine; however, SM-BC includes additional features such as transparent application failover and failback.

Modes

SM-BC operates in one of two modes. Synchronous mode is similar to MetroCluster. In normal operation, RPO=0 is preserved, and all writes are committed to the local and remote system. However, if a write cannot be replicated, synchronous mode times out and allows processing to continue. Site failure would result in data loss because the remote site would no longer be in sync with the original data.

This is the preferred mode for most customers, but for those workloads where a change must be committed to both replicas or not at all, SM-BC includes StrictSync mode. In this case, a prolonged inability to replicate a change results in an error reported back to the operating system performing the I/O. This typically results in application shutdown.

Path Access

SM-BC makes storage devices visible to host operating systems from both the primary and remote storage systems.

The paths to the local controllers are designated Active/Optimized paths and the paths to the remote controllers are Active/Nonoptimized paths. In normal operation, all I/O is serviced by the local controllers hosting Active/Optimized paths. In the event of a site failure or storage failover to a remote site, the Active/Nonoptimized paths are transitioned to optimized paths.

Failover

SM-BC supports two types of storage failover operations: planned and unplanned, which work in slightly different ways.

A planned failover is initiated manually by the administrator for quick switchover to a remote site whereas unplanned failover is initiated automatically by the Mediator on the third site. The primary purpose of a planned failover is to perform incremental patching and upgrades, to perform disaster recovery testing, or to adopt a formal policy of switching operations between sites throughout the year to prove full business continuity capability.

Storage Hardware

Unlike other disaster-recovery storage solutions, SM-BC offers asymmetric platform flexibility—the hardware at each site does not need to be identical. This capability allows you to right-size the hardware used to support SM-BC. The remote storage system can be identical to the primary site if it needs to support a full production workload, but if a disaster results in reduced I/O, then a smaller system at the remote site might be more cost-effective.

ONTAP Mediator

ONTAP Mediator is a software application that automates failover operations for both the primary and remote site storage cluster. It can be deployed on a small virtual machine (VM) hosted either on-premises or in the cloud. After it is configured, it acts as a third site to monitor failover scenarios for both the sites.

The Mediator recognizes this split-brain scenario and resumes I/O on the node that holds the master copy. When connectivity between sites is back online, the alternate site performs automatic resync.

Contact Fujitsu support personnel for obtaining the Mediator.

6. SAN Configuration Best Practices

The following best practices are critical to ensuring maximum SAN availability. The majority apply to the host and FC network configuration and are a result of various aspects and limitations of SAN implementations, operating systems, and multipathing software. While some deviations from these best practices might sometimes be worthwhile, administrators should carefully consider possible consequences and risks.

Independent FC Fabrics

FC SAN hosts obviously require redundant network connections to make sure that a single port failure on a host or network switch does not cause an outage. These two network connections should also use independent FC fabrics. Full-mesh fabrics lead to an excessive number of paths being exposed to a host and increase the risk of a user error affecting the entire SAN.

Independent IP Subnets

iSCSI and NVMe/TCP hosts that require maximum availability should use at least two network adapters (NICs) on independent subnets. Using a common subnet for all TCP/IP communication increases the risk of disruption to that one entire subnet, leading to an outage. Furthermore, many OSs have internal routing tables that result in one and only one of the available NICs being used for network communication. Extra NICs might exist, but, if they share a common subnet, they cannot be used by the OS.

Host bonding, such as LACP trunking, can also be used for each subnet.

As an example, HA iSCSI or NVMe/TCP could be configured as follow:

- NIC #1 on the host with an address of 192.168.1.10/24
- NIC #2 on the host with an address of 192.168.2.10/24
- Two-port LACP trunk on ONTAP controller #1 with an address of 192.168.1.1/24
- Two-port LACP trunk on ONTAP controller #2 with an address of 192.168.2.1/24

The result would be load-balanced availability of SAN resources from ONTAP controllers on trunked interfaces, plus load balanced redundancy on the hosts. The use of two subnets helps to make sure that network disruption does not completely interrupt SAN connectivity.

LUN Path Limits

SAN hosts in modern networks should not normally require more than four paths to a LUN or namespace and should never be configured for more than eight paths.

An excessive number of paths leads to delays in OS booting and path failovers. In some cases, an excessive number of paths exposes host OS bugs in path discovery and management. Finally, the risk of user errors administering SAN devices on the host is increased as the number of exposed paths increases.

LUN/Namespace (NS) Sizing

Even the largest ONTAP controller can be driven to 100% performance capacity with as few as eight LUNs or namespaces. More might be required if a single application is expected to consume the maximum theoretical performance capacity, but the incremental performance improvement beyond 8 LUNs or namespaces is rarely significant.

A larger number of LUNs or namespaces means an increased number of paths, which leads to the same problems discussed above. To avoid problems, use fewer but larger LUNs. For example, an ordinary database that is 8TB in size should be placed on four 2TB LUNs or namespaces. If I/O is especially high, eight 1TB LUN/namespaces might be beneficial.

The maximum recommended number of LUNs or namespaces supporting a single dataset is 64 LUNs per controller or 16 namespaces per controller. This is not the maximum number of storage resources that can be advertised from a single controller to a single host; it is however the maximum number of resources that should be used for a single dataset workload. For example, ten databases might represent 10 different workloads, each with eight LUNs for a total of 80.

Single-Initiator Zoning

Always use single-initiator zoning. While multi-initiator zoning is often harmless, some operating systems or HBA/firmware combinations will have intermittent problems caused by initiator crosstalk. It can be both severe and unexpected when it occurs. Single-initiator zoning avoids this problem by insulating one initiator from another. Multi-target zoning is acceptable.

SAN Configuration against the SAN Host Utilities Documentation

Although most operating systems work correctly as installed, certain configurations might require additional settings to function properly.

Use of sanlun Utilities to Verify Path Health

Host Utilities should be installed on all operating systems where it is supported. The critical utility is the `sanlun` command. Users can run `sanlun lun show -p` to verify the path health. This is especially important before performing ONTAP or SAN infrastructure upgrades. Many support cases reporting an outage ultimately turn out to be a result of missing paths, either because only a single controller was zoned to the SAN when the host was initially installed, or due to interim changes in the SAN since configuration.

Verifying that the correct number of paths exist and including both controllers of an HA pair prevent this type of oversight. Verification also detects possible OS misconfiguration or malfunction before changes are made to the SAN that could result in an outage.

In addition to the `sanlun` command, you can use the relevant OS multipath management tools.

Note on Linux LVM

There is a design flaw in the linux LVM that can lead to I/O errors and application crashes during path changes. During boot time, the multipath driver and the LVM driver start at about the same time, which creates a race condition. In most cases, multipathd finishes creating devices before LVM starts, but this is not guaranteed.

It is possible that LVM will create PV devices using the single path devices because the multipath device was not present when LVM scanned devices. If the LV using that PV is mounted and there is a failover, the PV disappears because its only path is no longer available. Although only a few configurations have enough LUNs or namespaces to trigger this bug, it has been observed with Fujitsu customers.

One sign that you might have an unsafe condition is in the `pvs` output. It can warn you that certain PVs are not using a multipath device.

```
  WARNING: Not using device /dev/mapper/3600a0980383038616e3f4a53716a4c7a for PV 4ZZweF-tjt9-
wLxC-CdPU-oQmT-78Wy-My6st2.
  WARNING: Not using device /dev/mapper/3600a0980383038616e3f4a53716a4d32 for PV O3IihV-zEaH-
J82B-fF8B-NGvz-dlPe-uUgb1r.
  WARNING: Not using device /dev/mapper/3600a0980383038616e3f4a53716a4d31 for PV XvjZty-T1qx-
7aHc-nrtI-yh3N-CWAv-U5gwrX.
  WARNING: Not using device /dev/mapper/3600a0980383038616e3f4a53716a4d30 for PV tl9BmZ-3dCY-
Lfvs-s7xR-3jfN-NLLT-dFGLc0.
```

This issue can be addressed by changes to `/etc/lvm/lvm.conf`. The default setting is as follows, and results in lvmd scanning all devices for physical volumes.

```
filter = [ "a|.*/|" ]
```

In general, the following setting works, but it must be tested carefully.

```
filter = [ "a|^/dev/sda[1-9]$|", "a|^/dev/mapper/*|", "r|^/dev/*|" ]
global_filter = [ "a|^/dev/sda[1-9]$|", "a|^/dev/mapper/*|", "r|^/dev/*|" ]
```

This filter results in lvmd scanning for physical volumes on `/dev/sda*` and `/dev/mapper/*` only. If your boot device is not a `/dev/sda` partion, this setting can interfere with rebooting. If, for example, your server's local boot device might appear as a `/dev/xda` device. Consult the official LVM documentation for more details.

Caution

If you change this file, reboot the server to make sure that a reboot is successful. Also, be prepared to log on at a console to fix errors.

Note on /etc/sysconfig/oracleasm Errors

When using an Oracle database with ASMlibm, make sure that `/etc/sysconfig/oracleasm` is not detecting single path devices. Single-path devices on Linux still work alongside multipath devices. ASMlib should be configured to only discover multipath devices.

Example:

```
# ORACLEASM_SCANORDER: Matching patterns to order disk scanning
ORACLEASM_SCANORDER="mpath dm"  (OR ORACLEASM_SCANORDER="dm")

# ORACLEASM_SCANEXCLUDE: Matching patterns to exclude disks from scan
ORACLEASM_SCANEXCLUDE="sd"
```

Note on host_config Script with Solaris

In particular, Solaris requires specific configuration steps to ensure that ONTAP multipath devices are properly recognized. Failure to follow the instructions on host configuration can affect resilience and result in severe ZFS performance problems.

NVFAIL

Any SAN volume on ONTAP storage with critical data should have the `nvfail` parameter set to `on`.

SAN workloads are especially vulnerable to corruption if a failover or switchover is forced because applications such as databases maintain large internal caches of data on drive. If a forced failover or switchover occurs, previously acknowledged changes are effectively discarded. The contents of the storage system effectively jump backward in time, and the state of the database cache no longer reflects the state of the data on drive.

The `nvfail` setting also protects the volume from a catastrophic failure of NVRAM journaling that puts data integrity in question. The `nvfail` parameter takes effect during startup. If NVRAM errors are detected, then there might be uncommitted changes that have been lost, and the drive state might not match the database cache. ONTAP then sets volumes with an `nvfail` parameter of `on` to `in- nvfailed-state`. As a result, any process attempting to access the data receives an I/O error, which leads to a protective crash or shutdown of the database.

Fujitsu Storage
ETERNUS AX series All-Flash Arrays
Data Availability and Integrity with All SAN Array (ASA)

C140-0068-01ENZ3

Date of issuance: November 2023
Issuance responsibility: Fujitsu Limited

- The content of this manual is subject to change without notice.
- This manual was prepared with the utmost attention to detail.
  However, Fujitsu shall assume no responsibility for any operational problems as the result of errors, omissions, or the use of information in this manual.
- Fujitsu assumes no liability for damages to third party copyrights or other rights arising from the use of any information in this manual.
- The content of this manual may not be reproduced or distributed in part or in its entirety without prior permission from Fujitsu.

FUJITSU