# ORACLE

# Architecture Principles and Implementation Practices for Remote Replication Using Oracle ZFS Storage Appliance

An Oracle ZFS Storage Appliance White Paper

## DISCLAIMER

This document in any form, software or printed matter, contains proprietary information that is the exclusive property of Oracle. Your access to and use of this confidential material is subject to the terms and conditions of your Oracle software license and service agreement, which has been executed and with which you agree to comply. This document and information contained herein may not be disclosed, copied, reproduced or distributed to anyone outside Oracle without prior written consent of Oracle. This document is not part of your license agreement nor can it be incorporated into any contractual agreement with Oracle or its subsidiaries or affiliates.

This document is for informational purposes only and is intended solely to assist you in planning for the implementation and upgrade of the product features described. It is not a commitment to deliver any material, code, or functionality, and should not be relied upon in making purchasing decisions. The development, release, and timing of any features or functionality described in this document remains at the sole discretion of Oracle.

Due to the nature of the product architecture, it may not be possible to safely include all features described in this document without risking significant destabilization of the code.

Table of Contents

# INTRODUCTION

An increase in storage demand increases the complexity of protecting data. The storage-based remote replication capability of the Oracle ZFS Storage Appliance products offers a simple and effective automated solution for businesses that require offsite copies of production data in addition to local backups. By maintaining a replica of the primary data at a remote site, disaster recovery time is dramatically reduced compared to traditional offline backup architectures.

To help system planners seeking to protect local copies of data by replicating them to a remote site, this paper addresses considerations in planning and implementing remote replication with Oracle ZFS Storage Appliance. The information particularly addresses system planners seeking to simplify offsite data protection.

This paper presents the following topics:

- Overview of the remote replication capability of the Replication feature of the Oracle ZFS Storage Appliance family of storage systems
- Implementation guidelines for deploying remote replication
- Application-specific implementation guidelines
- Example architecture
- Remote replication benefits
- Remote replication considerations

Considerations:

- Note that the functions and features of Oracle ZFS Storage Appliance described in this white paper are based on the latest available Oracle ZFS Storage Appliance firmware release.
- This white paper describes fundamental remote replication features and there is some overlap with the more recent Implementing Remote Replication Using Oracle ZFS Storage Appliance, which describes enhanced replication features.

# OVERVIEW OF ORACLE ZFS STORAGE APPLIANCE REPLICATION

The Oracle ZFS Storage Appliance products support snapshot-based replication of projects and shares from a source appliance to any number of target appliances or to a different pool in the same appliance. Replication can be executed manually or on a schedule for the following use cases:

- **Disaster recovery**: Replication can be used to maintain a replica of an Oracle ZFS Storage Appliance system for disaster recovery. In the event of a disaster that impacts service of the primary appliance (or even an entire data center), administrators activate service at the disaster recovery site, which takes over using the most recently replicated data. When the primary site has been restored, data changed while the disaster recovery site was in service can be migrated back to the primary site and normal service restored. Such scenarios are fully testable before such a disaster occurs.
- **Data distribution**: Replication can be used to distribute data (such as virtual machine images or media) to remote systems across the world in situations where clients of the target appliance would not ordinarily be able to reach the source appliance directly, or such a setup would have prohibitively high latency. One example uses this scheme for local caching to improve latency of read-only data (such as documents).
- **Disk-to-disk backup**: Replication can be used as a backup solution for environments in which tape backups are not feasible. Tape backup might not be feasible, for example, because the available bandwidth is insufficient or because the latency for recovery is too high.
- **Data migration**: Replication can be used to migrate data and configuration between Oracle ZFS Storage Appliance systems, or to move data to a different pool within the same Oracle ZFS Storage Appliance when upgrading hardware or rebalancing storage. The Shadow Migration feature of Oracle ZFS Storage Appliance also can be used for data migration.

Oracle ZFS Storage Appliance replication has the following properties:

- **Snapshot based**: The replication subsystem takes a snapshot as part of each update operation and sends the entire project contents up to the snapshot, in the case of a full update. In the case of an incremental update, only the changes since the last replication snapshot for the same action are sent.

- **Block level**: Each update operation traverses the filesystem at the block level and sends the appropriate filesystem data and metadata to the target.

- **Asynchronous**: Because replication takes snapshots and then sends them, data is necessarily committed to stable storage before replication even begins sending.

- **Resumable replication**: This refers to resumption of an initial replication process or resumption from last snapshot replication process. A replication process that was stopped, due to, for instance, a network failure or system outage, can be restarted so that the replication process can continue from the point it stopped instead of having to retransmit all previously replicated data from the last snapshot or an initial replication process.

- **Includes metadata**: The underlying replication stream serializes both user data and Oracle Solaris ZFS metadata, including most properties configured on the Shares screen. These properties can be modified on the target after the first replication update completes, though not all take effect until the replication connection is severed; for example, to allow sharing over NFS to a different set of hosts than on the source. The replication documentation in the "Remote Replication" section in the Oracle ZFS Storage Appliance Administration Guide provides more information.

- **Secure**: The replication control protocol used among Oracle ZFS Storage Appliance products is secured with secure socket layer (SSL). Data can optionally be protected with SSL as well. An Oracle ZFS Storage Appliance system can replicate only to/from another Oracle ZFS Storage Appliance system after an initial manual authentication process. The Oracle ZFS Storage Appliance Administration Guide provides more details.

- **Protocol independent**: Oracle ZFS Storage Appliance supports both file-based (CIFS and NFS) and block-based (FC and iSCSI LUNs) storage volumes. The replication mechanism is protocol independent.

Oracle ZFS Storage Appliance replication also includes the following important features:

- **Adaptive multithreading and dynamic compression level of replication data streams**: The data in replication streams at the source is compressed to make better use of the available network bandwidth between source and target node. This is especially beneficial in situations where distance and network bandwidth are data throughput limiting factors. The compression rate and number of compression threads is dynamically adjusted based upon CPU utilization at the source and available network I/O throughput between the source and target Oracle ZFS Storage Appliance system.

- **Monitoring recovery point objective (RPO) target(s)**: An RPO target can be specified for each replication action. An alert is generated when the replication update exceeds the desired RPO target thresholds, which can inform the system administrator to address potential issues or perform additional system tuning.

The site that hosts the project to be replicated is called the source, and the receiving side is called the target. Administrators must define replication actions on the source appliance. Actions are the primary administrative control point for replication. The replicated data at the target is referred to as a package.

Replication can be configured either 1:1, 1:many, or many:1, between the Oracle ZFS Storage Appliance systems. Likewise, more than one action can be added to a project. In this context, many projects can be replicated to many targets simultaneously. Replication is supported between any Oracle ZFS Storage Appliance platforms. It is also possible to replicate between different storage profiles.

One Oracle ZFS Storage Appliance node can act as both source and target for projects to be replicated.

For example, node 1 contains three projects: project1, project2, and project3. Node 2 contains two projects: project1 and project2. Node 1 replicates project1 to node 2 and node 2 replicates project2 to node 1. In this scenario, each node keeps a copy of the project of the other node. Additionally, node 1 can replicate project3 to itself. This setup can be used to migrate a project to a different pool or to copy data from a deduplicated pool to a non-deduplicated pool.

The Oracle ZFS Storage Appliance Administration Guide provides more details on these replication procedures. This guide is also available via the online help function.

In this paper, the term replication update refers to a cycle of operations starting with an implicit snapshot followed by streaming of the data to the target to completion.

## Replication Modes

The Oracle ZFS Storage Appliance remote replication capability supports three different replication modes to give administrators flexibility when supporting new deployments and complex legacy applications:

- **On demand**: Replication is manually triggered by the user at any time, either from the web BUI, command line interface, or via REST API script.
- **Scheduled**: Replication is automatically executed according to a predefined schedule. Periodic replication updates can be scheduled to start every 5, 10, 15, 20, or 30 minutes; every 1, 2, 4, 8, or 12 hours; and every day, every week, or every month. More complex update patterns can be created by defining multiple schedules for a single replication action. If a scheduled replication update cannot be started because the previous update has not completed, the scheduled update will automatically be attempted again once that update does complete. Additionally, if a replication update fails for any reason, it will be retried.

The replication process is the same for each mode except that the time interval between the replications is different. The replication modes can be changed from one to another at any time to support different and changing business requirements.

For every mode, an RPO target can be specified for each replication action, and a warning alert and error alert thresholds can be specified to monitor the RPO target for each replication action.

The following diagram provides a high-level overview of the replication concept. Oracle ZFS Storage Appliance systems are used as both "source" (original data system) and "target" (any system receiving a replica). A project with two filesystems (NFS and CIFS protocols) and a block-based volume [LUN]) are replicated. At time t1, a full replication happens and subsequently at time t2, only the changes between the time t1 and t2 are replicated.
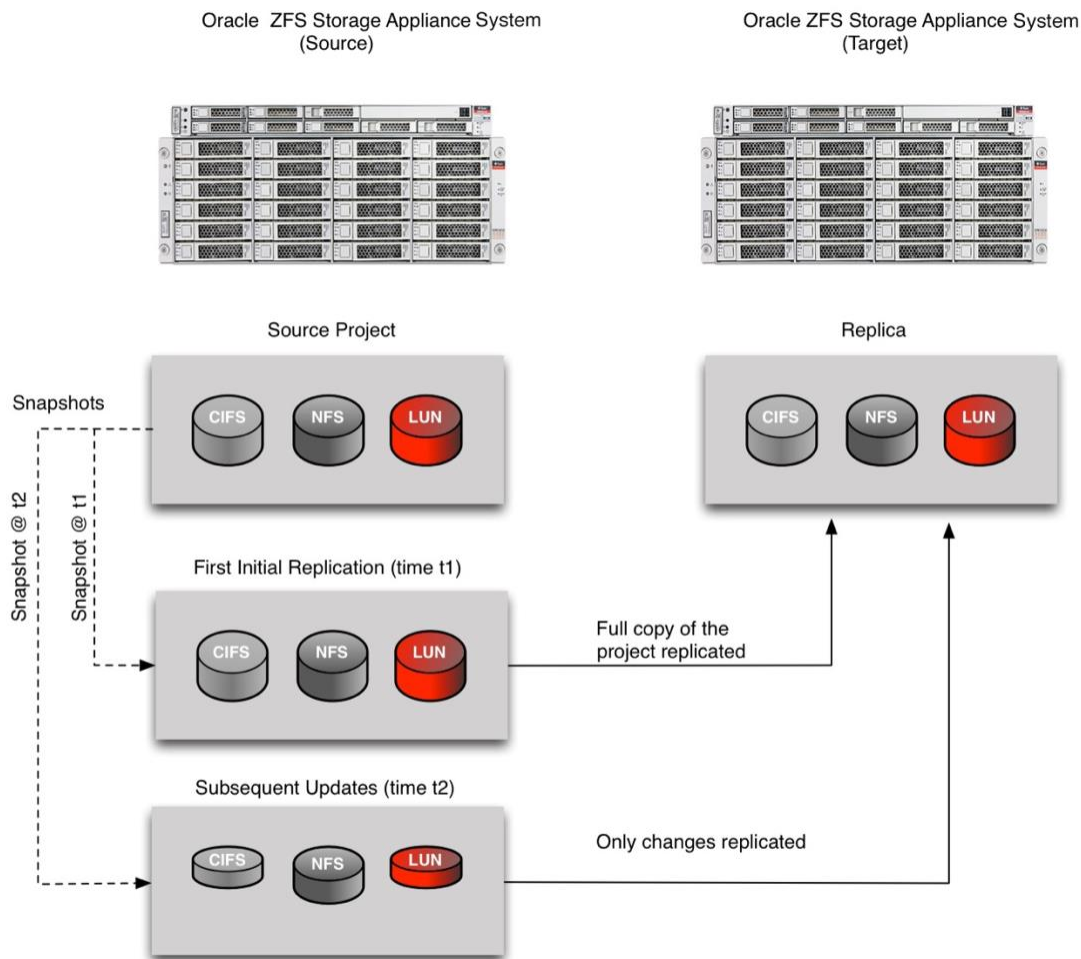
Figure 1. Replication process overview

## Using the Replica at a Target Site

The replicated package can be used for failover, test, backup, and reporting purposes after at least one successful transmission to the target site.

The Oracle ZFS Storage Appliance products provide several ways to use the replica target:

- **Read-only access**: Enable shares to be mounted from the client for read-only purposes using the Export option.
- **Clone/export replica**: Create a new writable project from the replica for alternate business processing at the target:
- **Sever link**: Convert a replication package into a writable local project by severing the replication connection.
- **Role reversal (reverse replication)**: Sever the link (see preceding) and establish replication back to the original source; this operation reverses the roles of the source and target sites.

The replica at the target site can be promoted into production operation in the event of a disaster in which the source site is completely lost. Furthermore, snapshots can be taken of the exported clone of the replica to allow for save points to protect system administrators from mistakes that can be made when acting quickly during unfamiliar circumstances.

When a node at the original primary site suffered from a partial loss of data (user error or loss of some disk enclosure), only a partial full replication has to be performed from the secondary site. The related projects/volumes that have a replicated copy in the secondary site need to have the current target severed from the primary and a new replication action configured to replicate the data back to the original primary node.

# Intelligent Replication Compression

A critical factor in a remote replication solution is the time lag between the original data and its produced copy. It is a key element in the recovery point objective (RPO) requirement. The RPO metric defines the point back in time where the last consistent copy of a set of data is available for recovery of data services after a disaster. The lower the RPO value, the less loss of data after a disaster. There are a number of factors that influence the RPO:

- Size of data set to be replicated and its rate of change (ROC)
- Frequency of sending changes to the target site
- Available bandwidth of the replication data link and its latency

The available bandwidth of a replication link is influencing the first two factors. An increase in available bandwidth will result in the ability to keep up with the ROC of a data set and increase the frequency of sending changes captured in snapshots to the target site—thus lowering the RPO. Quite often the bandwidth of a replication link between the source and target site is determined by physical characteristics of links that can be obtained from network providers.

By compressing the replication data stream at the source, the effective replication throughput can be increased beyond the physical link bandwidth capabilities. Including the compress function within a storage subsystem has a risk of slowing down other storage services in the subsystem. This adds an extra burden for system administrators, who must manage the various storage subsystem resources to minimize the impact on the performance of the storage subsystem as a whole.

The Intelligent Replication Compression feature of Oracle ZFS Storage Appliance uses an algorithm to dynamically adjust the level of compression and the number of concurrent replication threads, depending on system load and replication network performance of the Oracle ZFS Storage Appliance system.

The dynamic compression selection algorithm chooses between the use of GZIP and LZJB. GZIP compression offers a higher compression rate but needs more CPU resources than the LZJB type of compression. The selection algorithm uses available CPU resources and replication link speed as selection criteria to dynamically switch between use of GZIP or LZJB compression. When using GZIP, data buffers can be compressed fast enough to keep up with a 1 Gbs network. For high-speed network links, LZJB proves to be more efficient.

An adaptive multithreading algorithm is used to provide a replication stream with enough parallel threads to fully utilize the replication network bandwidth. If the available replication bandwidth decreases, the number of threads in a replication stream is decreased, too. The algorithm is clever enough to handle multiple replication streams over different networks. Replication threads are shared from a common pool, so when a replication stream on one network experiences a bandwidth drop, threads are moved to replication streams that run over a network connection with a higher bandwidth.

The following graphs show the benefit of using compression on the replication throughput compared to throughput over the network. The graphs also show the compression option turned off. Measurements were performed on a data set of 30 GB with a compression rate of 3x.
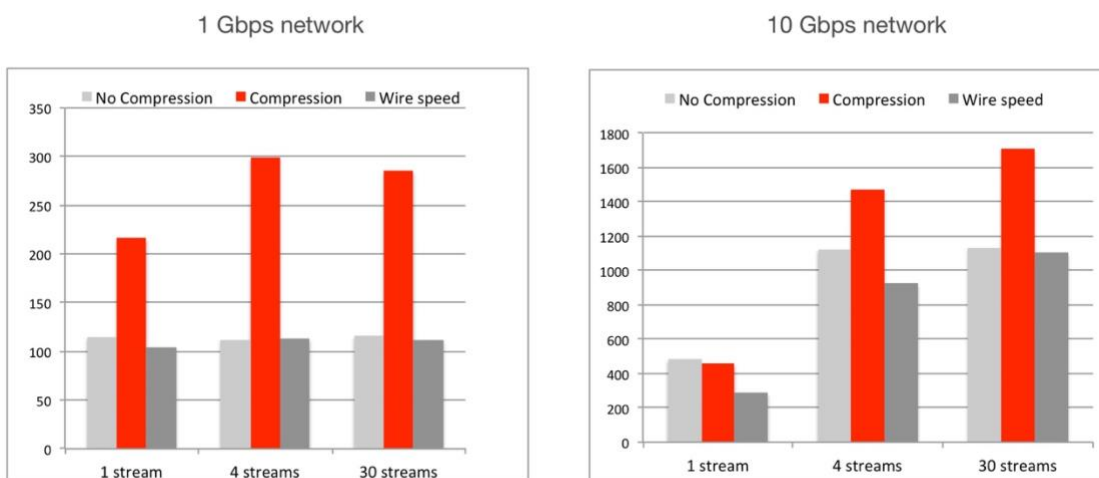


Figure 2. Comparing replication throughput using no compression and compression option and network throughput in MB/sec

## Monitoring Replication

## Monitoring Recovery Point Objective Settings

For each data set to be replicated, an RPO objective can be set via its related replication action. The RPO objective, expressed in time, should be derived from a wider customer disaster recovery or business continuity plan in which RPOs are specified for business processes using the data repositories by the related business applications.

An RPO target for a data set specifies the maximum time a copy data set is allowed to lag behind the source data set at the time of a disaster, that is, loss (of access) to the primary dataset.

The RPO target is specified in the replication action for a specific data set. The RPO target can be monitored by specifying a warning and error threshold level. A related warning/error alert is issued by the Oracle ZFS Storage Appliance system when the actual replication target time lag crosses the set alert threshold. The actual replica time lag can be monitored in real time via the command-line interface (CLI) and RESTful API scripting interface.

## Monitoring and Analyzing Replication Actions

After a replication action is created and the associated replication is scheduled, the replication progress can be monitored. Its progress is shown in the BUI under the replication rule, and it shows percentage of the data replicated, the replication throughput, and the estimated remaining time to completion. The same information can also be determined using the CLI.



Figure 3. Replication progress monitoring

Replication performance can be investigated in further detail using specific replication analytics statistics available under "Add statistic…" in the analytics worksheet window.



Figure 4. Replication analytics worksheet statistics

Each statistic can be broken down by direction, by type of operation, by peer, by pool name, by project, or by data set.

Figure 5. Replication analytics worksheet example

## Replication Event Logging and Action Auditing

Replication events are logged in the event log. These events can be monitored via email messages, sending SNMP information, or using the remote syslog mechanism. Replication events also can be configured to execute a specific workflow in Oracle ZFS Storage Appliance.

The Oracle ZFS Storage Appliance system posts alerts for the category remote replication when any of the following replication-related events occur:

- Manual or scheduled replication update starts or finishes successfully (both source and target).
- When any replication update fails, including as a result of explicit cancellation by an administrator (both source and target). The event includes the reason of failure.
- A scheduled replication update is skipped because another update for the same action is still in progress.
- A replica time lag exceeds its specified RPO threshold.

The following replication configuration actions are logged in the auditing log so that creation and deletion of all actions and changes to actions can be tracked. For replication actions this includes:

- Creating, modifying, or destroying replication actions
- Adding or removing shares from a replication group
- Creating, modifying, cloning, reversing, severing, or destroying replication packages on the target
- Creating, modifying, or destroying replication targets



Figure 6. Replication actions audit log messages

## GENERAL IMPLEMENTATION GUIDELINES

The following general guidelines and considerations are for administrators who are applying remote replication capabilities in the Oracle ZFS Storage Appliance environment.

Oracle ZFS Storage Appliance products use asynchronous communication between the source and the target to ensure that network latency does not slow production operations. This technology cannot guarantee that updates to the source will be present at the target site after a loss of the source site; however, the image of the project at the target site is guaranteed to be write-order consistent as of the time of the most recently completed data transfer.

## Configuring a Replication Mode

An administrator of Oracle ZFS Storage Appliance systems can configure or change the replication mode between scheduled and on demand at any time. The administrator also can configure different replication modes for the same target when using multiple replication actions.

Choosing the optimal configuration depends on the technical and operational requirements of the specific implementation. With regard to the replication mode, the administrator should consider the following details:

- Recovery point objective (RPO)
- Recovery time objective (RTO)
- Available bandwidth between the source and target sites and related to this, the rate of change (ROC) of data on the data set to be replicated
- Balancing available CPU and memory resources for the various data functions within the Oracle ZFS Storage Appliance system

Scheduled replication mode makes the best use of available resources. As long as the chosen scheduled replication interval is big enough to be able to send the changed data to the secondary site, a stable RPO level can be expected in this mode. With scheduled replication, the Oracle ZFS Storage Appliance source periodically replicates a point-in-time image (snapshot) of the source project to the remote site. This reduces network traffic while preserving consistent and timely copies of the primary data set. With the option to set the replication interval as low as five minutes, low, predictable RPO values can be maintained.

On-demand replication mode is designed for applications that need to put data into a specific state before the replication can occur. On-demand replication updates can be triggered from arbitrary locations in the application-processing stack through the RESTful API automated scripting language of the Oracle ZFS Storage Appliance CLI.

## Replication Network Setup

When there are enough network ports available on the Oracle ZFS Storage Appliance controller, it makes sense to dedicate a specific port for replication traffic. This is to ensure that replication data to the target IP uses that specific port on the Oracle ZFS Storage Appliance system. An entry in the routing table can be added using a specific /32 static route to the target's IP address over the replication source interface. After the entry in the routing table is added, the replication rule can be set up for the target IP address.

A clustered configuration ensures the replication for the specific network interface is migrated to the other node in the event of a node failure or scheduled maintenance activities. If no static route is defined to the target's IP address, replication traffic may be delivered over a private interface, such as an administrative interface, that becomes unavailable in the event of a takeover within the cluster.

The Oracle ZFS Storage Appliance DTrace Analytics feature can be used to verify that the replication data traffic is using the intended interface on the source Oracle ZFS Storage Appliance system.

Of course, the same setup needs to be repeated on the target Oracle ZFS Storage Appliance system, so when a replication role reversal is executed, the target node also uses its dedicated network port for the reversed replication traffic.

## Performing the Initial Seeding of a Target Node

When setting up a remote replication configuration, the initial replication of the data set can take a long time when the available bandwidth between two locations is limited. To overcome this problem, the initial replication synchronization can be done by locating source and target Oracle ZFS Storage Appliance systems next to each other. Once data is synchronized, the system intended for the remote site can be shipped to its location. Once this Oracle ZFS Storage Appliance system is connected to the network, its new target IP address can be changed in the source Oracle ZFS Storage Appliance system.

A second option is to use the offline replication capability when creating a replication action. With this option, the initial replication data stream is stored on an NFS share as specified under Export Data Path from a local available system. All other replication action properties can be specified as required.

Figure 7. Initial replication seeding by exporting to NFS share

When the export replication action is finished, the system with the NFS share can be moved to the remote site or the replication data can be copied from the NFS share to a transportable media like tape or DVD. When setting up a replication action, the initial replication data stream will be stored on a local system's NFS share (visible in the following figure in Import Data Path). As part of the initial replication action, the source node already informs the target node of the replication action setup of the share/project. On the node on the secondary site, the external replication data set then can be imported via NFS from a system that contains the shipped copy of the external replication data set using the "Import update from external media" option.
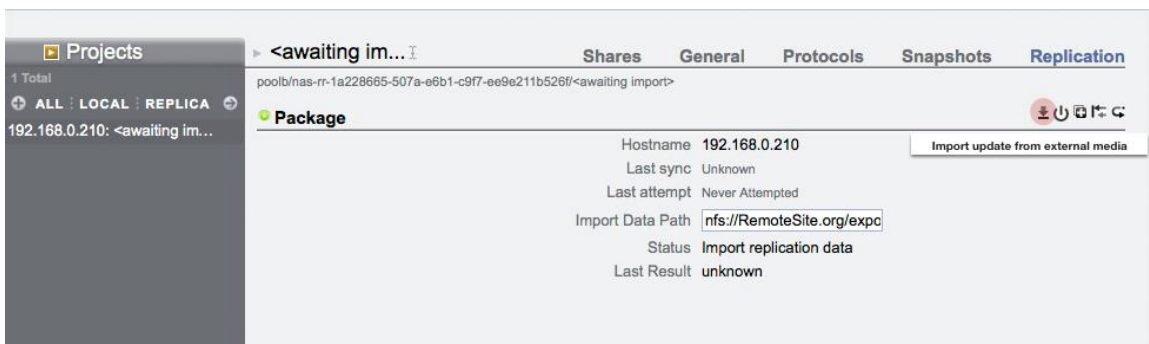


Figure 8. Import replication set from NFS share

## Using the Intelligent Replication Compression Feature

Intelligent Replication Compression is enabled by default. When upgrading Oracle ZFS Storage Appliance systems to a firmware level that includes Intelligent Replication Compression, and replication actions are already present, these replication actions will have Intelligent Replication Compression enabled as part of the firmware upgrade process.

In general, there is no need to disable Intelligent Replication Compression for any replication action. It is advised not to use Intelligent Replication Compression for the following exceptional cases, where:

- Source data sets to be used for replication are not compressible at all.

- CPU resources are already heavily utilized by other services in Oracle ZFS Storage Appliance.
- The replication network is extremely fast and dedicated to replication only, so there is no need to save network bandwidth used.
- A WAN accelerator that performs compression is already in use on the replication link.

Be careful not to enable SSL on the Oracle ZFS Storage Appliance system when using a WAN accelerator, as enabling SSL substantially degrades replication throughput in the WAN accelerator.

The benefit to using the compression function of the Oracle ZFS Storage Appliance system is that it will increase the replication throughput as the SSL encryption is performed against compressed data.

A single controller supports up to 120 parallel running replication actions (the sum of all incoming and outgoing replication updates). When a large number of replication actions are active, systems (both at source and at target) should be carefully monitored for CPU resource usage. When CPU utilization exceeds 90 percent, adding more replication actions that could run in parallel with already existing replication actions—due to overlapping schedules—is not recommended. There should be no more than 120 replication actions (threads) running at the same time.

For situations where data can be further compressed by using deduplication, such as VDU-type data images or Oracle Recovery Manager (Oracle RMAN) backup images, the replication deduplication over-the-wire option can be used. However, this option should be used carefully as it requires extra CPU and memory resources to deal with the deduplication process in the Oracle ZFS Storage Appliance system. The efficiency of the deduplication on a specific replication stream can be monitored by looking at the project messages in the BUI Logs > Alerts page that begin with "Finished replicating…"
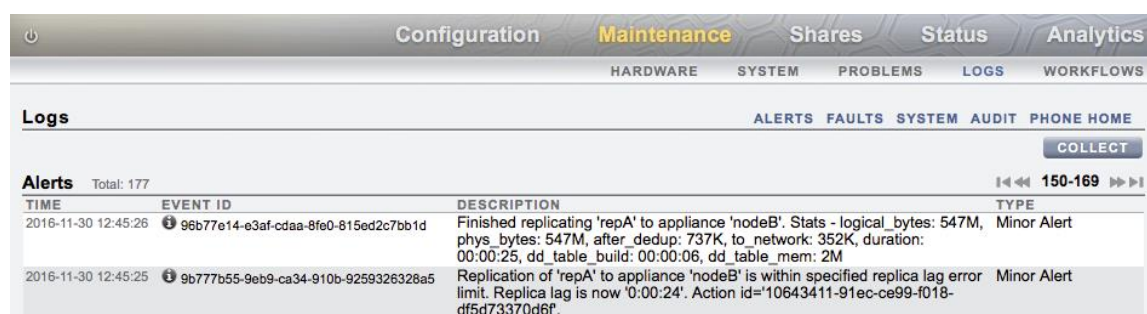


Figure 9. Deduplication statistics in replication finished alert logs

## Project-Level Compared to Share-Level Replication

Remote replication can be performed on both the project level and individual share level. Project-level share replication is the preferred option for a number of reasons:

- **Data consistency**: A project acts as a data consistency group for all shares in the project when they are replicated with a project replication action. This is very important when data transactions are using multiple shares and integrity, between the shares for these transactions, needs to be guaranteed.
- **Limiting the number of snapshots and disk space overhead**: Replication snapshots are always taken on a project level. When replication actions are created on individual shares, more snapshots are likely to be created, resulting in higher capacity consumption on the pool.
- **Reversing replication**: When replication is reversed and shares are replicated individually, the share is placed in its own project, resulting in fragmentation of the original project and loss of the share's property inheritance relationship with its project.

## Using the Deduplicated Replication Feature

The Deduplicated Replication feature in the Oracle ZFS Storage Appliance product provides the ability to reduce the amount of data sent over the replication connection to the target node, increasing network efficiency. This feature is different and independent of the appliance on-disk data deduplication feature. On-disk data deduplication is chosen on the project or share level and aims to eliminate duplicated blocks of data stored on disk. When a project or share that has the data deduplication feature enabled, is replicated, its data is deduplicated before being send to the secondary node. This process is independent from the use of the Deduplicated Replication feature.

As with the on-disk data deduplication feature, the Deduplication Replication feature has to be used with caution. This feature requires CPU resources for data preprocessing and memory resources for lookup tables. The amount of these resources needed is also dependent on certain share property settings. Furthermore, the effectiveness of the deduplication is very much dependent on the type of data sets used. The benefits of using Deduplicated Replication has a specific narrow sweet spot use case. The following basic requirements define the sweet spot environment:

- Data should have sufficient duplicates, such as weekly full RMAN images or VDI images.

- The replication connection between the nodes uses a low-bandwidth, high-delay network.
- The checksum property used for the share/project is set to either SHA-256 or SHA-256-MAC and if not, on-disk data deduplication should be set.
  The Deduplication Replication feature uses crypto-type check summing. When they already have been generated at the time of storing data, they don't need to be recreated. When using on-disk data deduplication, strong type checksums are already provided.
- When encryption is used, a deduplication 'friendly' encryption mode must be set: GCM type encryption never creates repeatable data patterns so deduplication can never be effective here. CCM type encryption should be used and on-disk data deduplication must be switched on.
- Do not use small on-disk block sizes: When small block sizes are used, a larger number of blocks need to be processed, it takes longer to build deduplication tables and tables are much larger with the risk that there is not enough memory available to build the tables needed. Also when there are relatively small updates to the data set, the time it takes to build the tables outweigh the benefits of time gained in the transfer process. It is recommended to use disk block sizes of 1MB for data sets to be used for the Deduplication Replication feature.

If these conditions are not met, the replication transfers are slower and bulkier but will still work.

The effectiveness of the Deduplication Replication feature on data sets can be monitored. The appliance shows a summary of the deduplicated statistics in the event log under alerts.



| 2017-3-3 14:52:03 | ⓘ ae9becc8-0ddb-624a-fde1-beeb290a335c | Finished replicating 'repa2' to appliance 'nodeB'. Action - e807948c-bd22-c229-b67f-d210318af6f5. Stats - logical_bytes: 1.12K, phys_bytes: 9.07K, after_dedup: 5.72K, to_network: 9.64K, duration: 00:00:03, dd_table_build: 00:00:00, dd_table_mem: 0. | Minor Alert |

Figure 10. Deduplication replication event reporting example

More detailed information can be obtained via the CLI interface using the 'stats' node of the replication action for a share:

```
nodeA:shares repa2/vola2-1 action-000> stats

nodeA:shares repa2/vola2-1 action-000 stats> ls

Properties:

           replica_data_timestamp = Fri Mar 03 2017 15:02:00 GMT+0000 (UTC)

                        last_sync = Fri Mar 03 2017 15:02:03 GMT+0000 (UTC)

                         last_try = Fri Mar 03 2017 15:02:03 GMT+0000 (UTC)

                      last_result = success

                last_logical_bytes = 1.125K

                  last_phys_bytes = 9.06640625K

                 last_after_dedup = 5.72265625K

                  last_to_network = 9.68261719K

                    last_duration = 00:00:03

               last_dd_table_build = 00:00:00

                 last_dd_table_mem = 0

                    total_updates = 91

              total_logical_bytes = 606.117188K

                 total_phys_bytes = 865.257813K

                total_after_dedup = 511.277344K

                 total_to_network = 886.063477K

                   total_duration = 00:05:37

                  dd_total_updates = 27

            dd_total_logical_bytes = 30.375K

               dd_total_phys_bytes = 244.816406K

              dd_total_after_dedup = 154.226563K
```

```
dd_total_to_network = 260.385742K

dd_total_duration = 00:01:39

dd_total_table_build = 00:00:00

dd_total_table_mem = 0
```

## Determine the Replication Snapshot Frequency of Scheduled Replication Actions

To determine the appropriate replication snapshot frequency, both the recovery point objective (RPO) and the rate of change (ROC) of the data repository to be replicated need to be defined. An RPO is one of the key elements in a business continuity plan, specifying the amount of transactions or data loss that can be tolerated by the business in case of a disaster impacting the primary storage repository. The RPO specifies how far back in time a snapshot of the data repository is available at an alternative location. Equally important is knowledge of the ROC of data in the primary data repository. The ROC determines the minimum required transmission speed of the replication link between the primary and secondary site and, to some degree, the replication interval. Both the RPO and ROC determine the frequency of a scheduled replication action for a specific data repository or project in Oracle ZFS Storage Appliance context. For the ROC, it defines the amount of data changed between two replication intervals and the requirement that that data must be transmitted to the secondary location within the subsequent snapshot intervals.

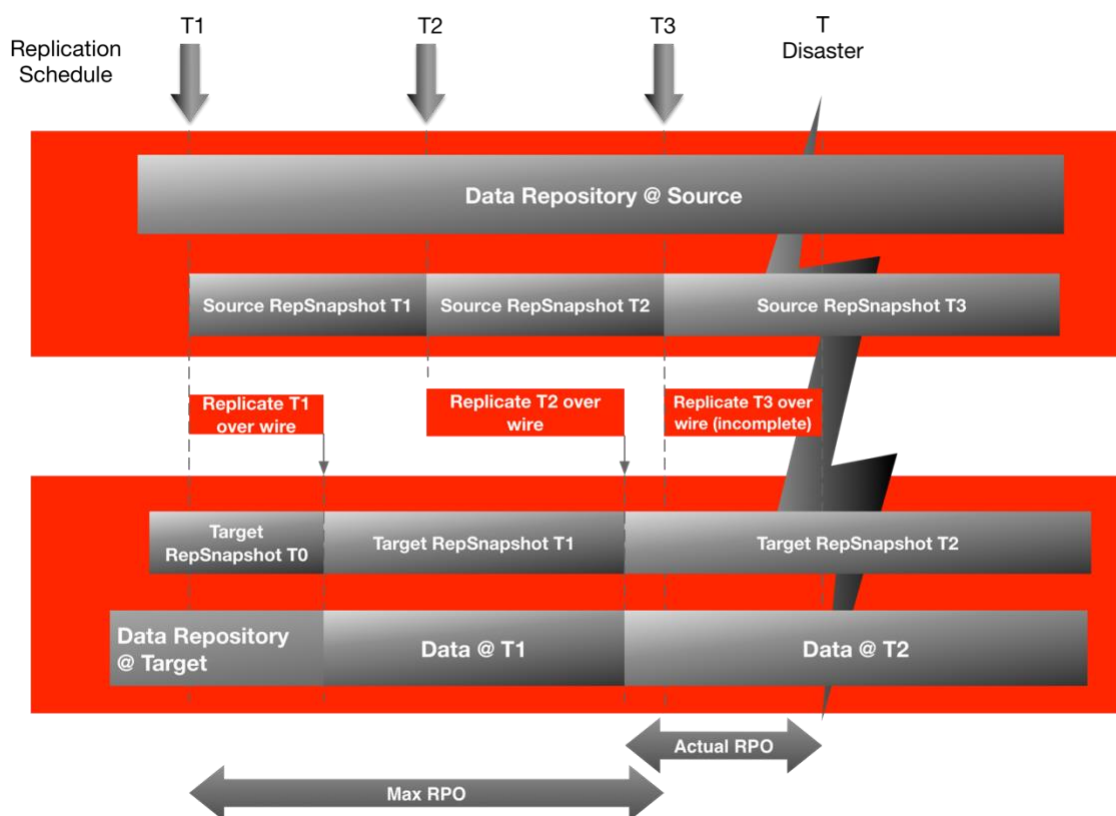For the RPO, the recommendation is to use a replication frequency of one-half the required RPO.



Figure 11. RPO and snapshot frequency relationship

The worst case RPO situation occurs when a replication snapshot contains almost no changed data followed by a snapshot that almost takes the whole time slot between two replication actions. The resulting RPO is twice the snapshot time interval, based on the assumption that the replication of a snapshot will always finish before the next snapshot is taken. When it takes longer for a replication snapshot to be transmitted, the next replication snapshot action is deferred until the in-progress update completes.

When replicating multiple data repositories to a secondary site, the bandwidth requirements of the data connection between the sites are determined by the total ROC of all data repositories.

When replicating compressed data repositories, the actual replicated data will be less than the actual "logical" ROC of the data in the repository. The Oracle ZFS Storage Appliance analytics function can be used to determine the physical ROC on the actual disks as a measure to determine the required bandwidth for replicating data between the primary and secondary site.

## Setting Up Required RPO in Replication Action

For each replication action an RPO objective can be set along with the related warning and error level threshold.



Figure 12. Specifying RPO and RPO monitoring thresholds

A benefit of using scheduled replication is the more predictable RPO level. With the option of setting replication frequency intervals as low as five minutes, low RPO values can be achieved.

## How to Configure Replication Alert Events Monitoring

An Oracle ZFS Storage Appliance system contains an alerting system that collects and manages all event and threshold alerts occurring within the system. Oracle ZFS Storage Appliance has a very flexible mechanism for managing the way the system reacts to these alert messages. Threshold alerts can be defined based on analytics metrics like, for instance; capacity percentage threshold used for a specific pool.

For each category of alerts, alert actions can be defined. Alert actions are used to define the type of action to be taken by the system. Alert actions are added via the Configuration > Alerts menu.
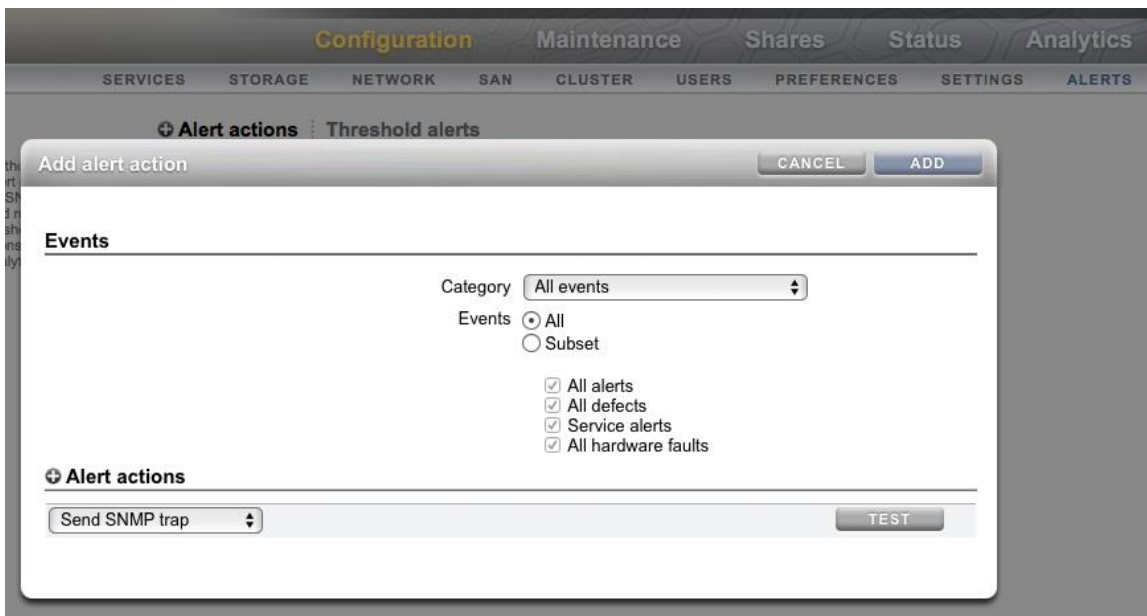
Figure 13. Add alert action dialog

The following alert actions are available:
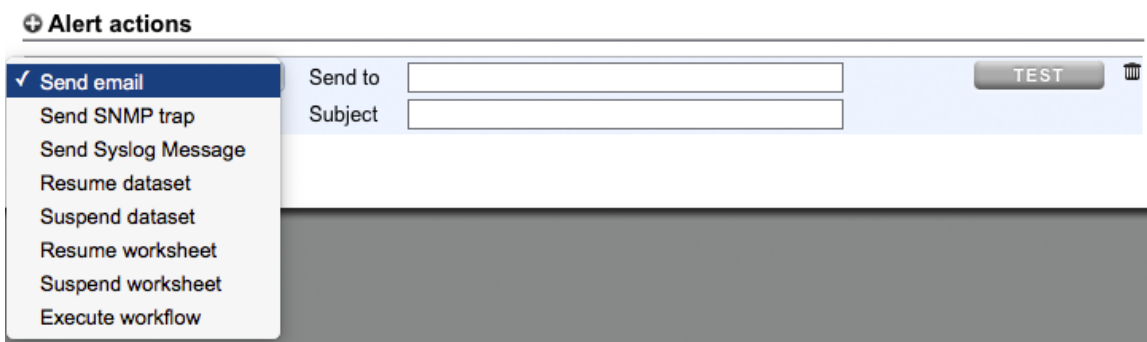


Figure 14. Type of event alert actions

Note that for the SNMP and syslog alert action options, their respective service has to be defined and started via the Configuration > Services section. In a windows environment, the Microsoft Management Console (MMC) can be used to access the event logs. The Oracle ZFS Storage Appliance Administration Guide (also available via the BUI online help) under SMB MMC Integration provides further details.

Three types of event categories are available for the replication function:

- Remote replication, source only
- Remote replication, target only
- Remote replication (which contains both source and target events)

Figure 15. Available event categories via add or edit alert action option

The replication events category contains detailed replication events that can be selected to be included in the alert action messages.

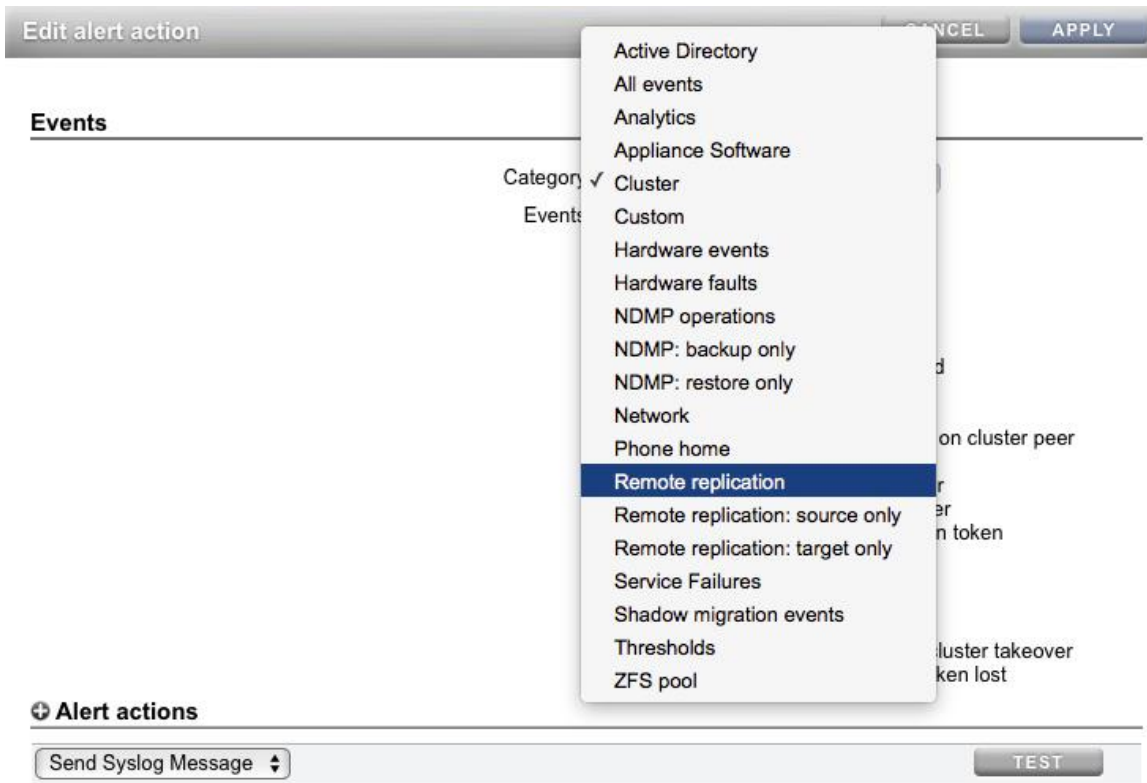## Events

Category    Remote replication ▾

Events   ⦿ All
       ○ Subset

☑ Receive failed (unsupported version)
☑ Receive failed (cancelled)
☑ Receive failed (encryption key unavailable)
☑ Receive failed (all others)
☑ Receive failed (NDMP-held snapshot)
☑ Receive failed (Retained snapshot)
☑ Receive failed (out of space)
☑ Receive failed (package not upgraded)
☑ Receive failed (clone origin missing)
☑ Receive finished
☑ Receive not making progress
☑ Receive started
☑ Receive resumed making progress
☑ Send failed (unsupported version)
☑ Send failed (unsupported feature)
☑ Send failed (cancelled)
☑ Send failed (encryption key unavailable)
☑ Send failed (all others)
☑ Send failed (connectivity)
☑ Send failed (out of space)
☑ Partial send failure (cancelled)
☑ Partial send failure (connectivity)
☑ Send failed (remote verification)
☑ Send failed (out of receive streams)
☑ Send failed (package severed/reversed)
☑ Send failed (replication service/package disabled)
☑ Send failed (package offline)
☑ Send failed (target pool in discovery)
☑ Send finished
☑ Send finished (target downrev)
☑ Export finished
☑ Send skipped (already running)
☑ Replica lag exceeded warning limit
☑ Replica lag exceeded error limit
☑ Replica lag within warning limit
☑ Replica lag within error limit
☑ Send not making progress
☑ Send started
☑ Export started
☑ Send resumed making progress
☑ No replica dedup

## ⊕ Alert actions

Send email ▾    Send to [               ]     TEST
                   Subject [               ]

Figure 16. Replication events for alert actions

## Replicating Between Nodes with Different Software Release Versions

Replication is compatible between most Oracle ZFS Storage Appliance software releases. Compatibility issues occur when different releases support different ZFS pool features that are incompatible, and those features that are applied to the node by executing the deferred updates function after a software upgrade procedure. Also, issues occur when newer releases contain new replication options and features that may not be supported in combination with earlier releases. More information about compatibility and deferred update features can be found in the Oracle ZFS Storage Appliance Remote Replication Compatibility document (Doc ID 1958039.1) on My Oracle Support https://my.oracle.support.

## Replicating Encrypted Shares

The software on both source and target nodes must support the Oracle ZFS Storage Appliance encryption function. Before starting the replication action for an encrypted project or share, the encryption keystore must be set up at both nodes. When using Oracle Key Manager, both nodes must point to the same Oracle Key Manager server and the same wrapping keys must be set up. When using a local keystore, both nodes must use the same wrapping key names. Failure to do so will result in replication failure alerts in the Oracle ZFS Storage Appliance log system available under Maintenance > LOGS in the BUI.

## Use at the Target (Remote) Site

The replica at the target site contains all the information present at the source site as of the most recently completed replication transaction. The replica can be used as a direct substitute for the original source in the event that the source is lost, or clones of the replica can be used for additional business processing, such as backup, test, or analysis.

## Accessing Read-Only Copy

The replicated package can be mounted and can be used as a read-only source from the client. Setting the Export option for the replicated project makes the project and its shares mountable from the client. The data can be used for backup and other read-only purposes.

## Role Reversal (for Production Failover)

The client application can be failed over to the target site in the event of a primary site disaster, or during a maintenance operation at the primary site. The target becomes the "production" site in this mode. Only the data that is changed after the failover is replicated back when the primary site is back up again. If the original primary site is set to become the production site again, the role-reversal operation is performed at the new target. It is strongly recommended to disable the replication before initiating this operation.

If a replication action is not disabled before the reversal process is started, the replication action at the old source will be automatically disabled when it attempts to send the next update to the new source.

## Test, Backup, and Reporting

To access the replicated package for read/write purposes, a package is cloned. This operation essentially clones the project that is received. Clones of the replication target can be used for test, reporting, and any other type of business processing.

Cloning the replication target efficiently uses storage resources because space allocation for clones is performed only when updates to the primary or cloned data set are performed.

Clones can be recloned and deployed to these different applications in cases where multiple applications need access to copies of the same data set.

Alternatively, many clones from the same replica can be created. Space utilization is further enhanced in this case because the read-optimized flash devices in the L2ARC store a copy of the on-disk image, which is only the original data set plus the updates to the data set.

Multiple test applications can take advantage of the performance of read-optimized flash devices without requiring large amounts of read-optimized flash storage in cases where few updates are performed on the data set. This provides a high-performance, cost-effective solution for using replication resources to perform value-added business functions.

## Multiclone Setup

The target site allows multiple clones from the package at different points in time. For example, a clone project, clone-t1, is created at time t1, which has the point-in-time data as of the last received replica. The replica is refreshed by a replication update after t1. At time t2, another clone project, clone-t2, is created from the same replica.

Likewise, any number of clones can be created at the target. If the clone projects are not used for updates, the additional space occupied in this setup at the target is only a minimal amount used for metadata. This multiclone setup is useful when the requirements are to maintain few point-in-time images at the target site for various purposes.

## Managing Backups

The replication target offers an alternative solution when the resources of the primary site are insufficient or unavailable to perform backups. Network Data Management Protocol (NDMP) backups can be used to back up NFS and CIFS shares using the read-only mount, and arbitrary backup applications can be used to back up data over any protocol supported by Oracle ZFS Storage Appliance systems.

## Backup Using NDMP

Local, remote, and three-way NDMP backups can be performed by Oracle ZFS Storage Appliance systems on CIFS and NFS shares. In either case, the backup is performed by exporting the project as read-only.

Local full backups can be performed by attaching tape storage devices to an Oracle ZFS Storage Appliance system, and remote backups are performed over the network when the tape devices are not attached to an Oracle ZFS Storage Appliance system but instead are connected to a media server. The data to be backed up is transferred over the network to the media server.

A three-way backup is performed by executing a remote backup to an alternate Oracle ZFS Storage Appliance system with access to tape devices.

The local NDMP backup is the most efficient mode for the backup because it does not incur any data transfer over the network. The other modes require the data to be transferred over the network, so the architecture design should take into consideration the bandwidth requirements. Using the read-only mount, both full and incremental backups can be performed for replicated shares.

With the introduction of the NDMP replica backup enhancement a few releases ago, NDMP is now capable of creating a ZFS format backup (full or incremental) directly from a replication package without requiring the creation of a clone and without having to export a read-only mount point.

## Snapshot Management for Replication

Snapshots can be generated for each project, filesystem, or LUN. These snapshots are instant and offer a frozen read-only view of a project, filesystem, or LUN at the time of the snapshot. Snapshots can be taken by a user (manually) or by a scheduled action. Scheduled snapshots are taken at regular intervals and can include how many snapshots to retain for a specific project, filesystem, or LUN.

Whenever in a replication action the "Include snapshots" property is set, snapshots created via a schedule or manually, will be replicated from the source site to the target site.

Scheduled snapshots of projects, filesystems, or LUNs have an associated retention policy property. This retention policy can be applied to the target site, too, or can be made independent from the source by specifying a different retention policy for the target site in the replication action for the project, filesystem, or LUN. When a replication role reversal takes place, only the snapshots that were originally present on the source site are replicated back and not any that had been deleted at the source site before the replication role reversal took place.

User-created snapshots are always replicated from the source site to the target site. When deleting a snapshot from the source site, the default behavior is that it will also be deleted from the target site. However, there is an option in the replication action to prevent this, so a snapshot at the target will remain there until the user deletes the snapshot. This makes using replication for remote back-up solutions more flexible and gives the user control over the number of snapshots kept at the target independent of the number of snapshots at the source and deletion of snapshots on the source.

Replicating clones from the source site to a target site takes a bit more consideration. Clones are created from a snapshot of a filesystem or a LUN at the source site (this type of cloning is different from cloning a replica at the target). They are thin clones, meaning the data that is not changed is shared between the clone and its originating share or LUN. Any data written to them is stored in the clone instance only. Clones are often used as a quick way to create copies from data repositories for testing or to serve as a seed for a new instance of a fresh installation.

In some cases, clones are put in a different project as the origin and, as such, a different replication action for the clone needs to be created. In such cases, replicating the data in the clone instance itself is not enough for the clone to be useable at the target site. So, the originated filesystem also must be present at the target site.

There are two ways to accomplish this. Either replicate the clone origin first before replicating the project containing the clone, or use the "Include clone origin as data" property in the replication action for the project containing the clone.

When putting clones in a different project as the origin, there is a convenience/space trade-off to be made:

» If the clone origin is replicated before the clone, on the target the relationship between the origin and clone will remain intact and thus the clone will remain a thin clone.

» When using the automatic inclusion option by specifying the "Include clone origin as data" property in the replication action, the origin-clone relationship is lost at the target when the clone is located in a different project as the origin and the clone contains a full copy of the origin of the clone at the source.

Therefore, from a replication perspective, the best practice is to keep clones within the same project as the origin.

# APPLICATION-SPECIFIC IMPLEMENTATION GUIDELINES

This section augments the general guidelines presented in the previous section. The principles described in this section can be applied to other applications when the behavior of the application is similar to that of the applications described.

## Business Applications and Middleware

Metadata and configuration details, as well as unstructured business application data, can be protected by storing the information on an Oracle ZFS Storage Appliance system and replicating the content to a target system. Different applications' data can be stored on different shares in the same project in cases where the data must be kept consistent with the data of a disparate application. A consistent image of a federated system can be maintained at the remote site in this model. scheduled or on-demand replication can be used, depending on the specific requirements of the implementation.

## Consolidating Virtualized Environments

Protecting the virtualized environment infrastructure is critical. The virtual disk images (VMDK) files form the crux of the virtualized environment. The disk images are stored on either iSCSI LUNs or over NFS files. It is recommended to place the related disk images in a single project for replication. Then, a number of projects can be created and replicated.

A VMware VMFS3 filesystem can be created on the iSCSI LUN to host VMDKs, or the iSCSI LUN can be attached directly to a virtual machine using raw device mapping (RDM).

Differing replication schedules can be configured for different types of virtual machines or for subsets of virtual disk types; for example, OS/boot versus production data. For the replication, the recommendation is to configure multiple projects, each to replicate on its own schedule.

If the images need to be in a quiesced state, a manual mode using scripting at the virtual machine server that quiesces the image and then initiates the replication is preferred.

## Protecting Mail Servers

Mail servers are a critical piece of the business component that need to be protected either by backup or by replicating to a remote site. In the event of a primary site problem, the mail server can then fail over to the target site to avoid downtime. Microsoft Exchange Server uses database and log components. Also, only iSCSI LUNs are supported to host these files.

In order to perform full replication, the logs and the database files are stored in the same project and replicated using any of the modes. Note that Microsoft recommends that, for better protection, logs and database files be separated on different physical drives. However, having a mirrored storage pool in an Oracle ZFS Storage Appliance system alleviates the problem and provides better performance.

# PROVIDING A DISASTER RECOVERY SOLUTION — AN EXAMPLE ARCHITECTURE

The following figure shows an example scenario for an enterprise's disaster recovery setup.
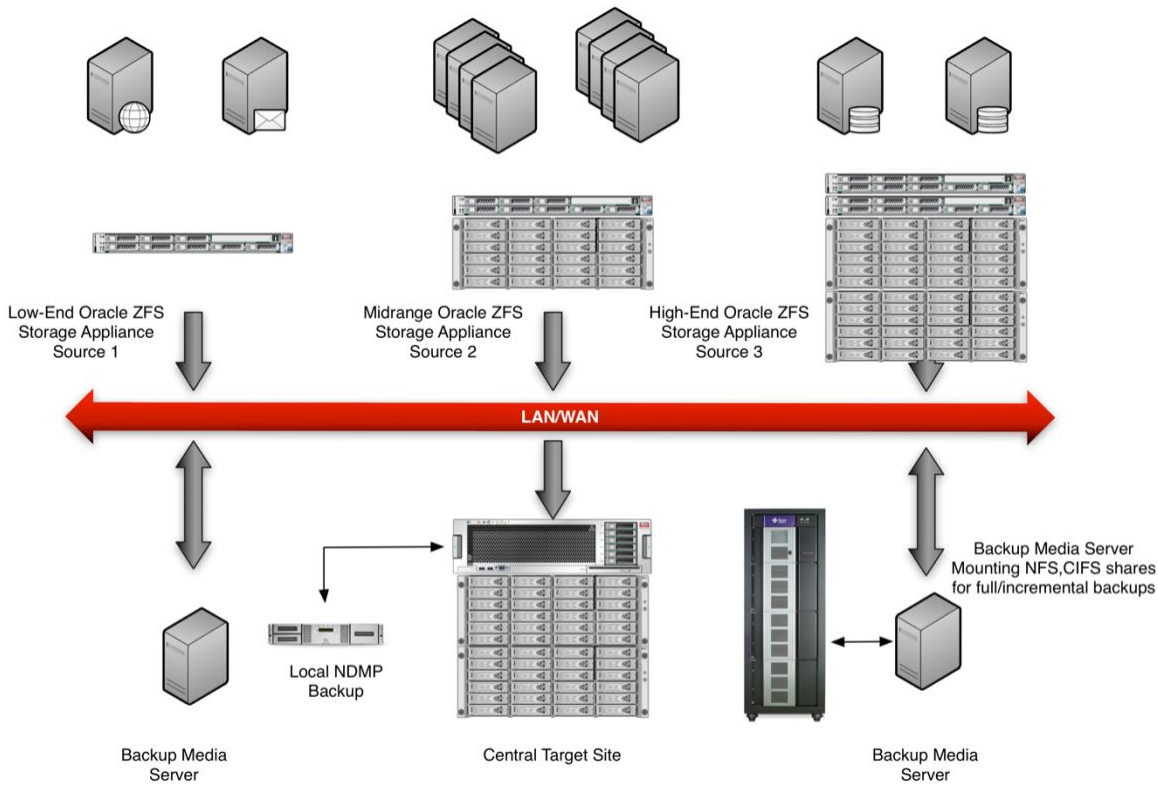


Figure 17. Replication setup for an enterprise

## Source Site Configuration

In the example architecture, distributed Oracle ZFS Storage Appliance systems that are serving various application needs are deployed across the enterprise. These individual storage systems host a number of projects with varied types of application usage. All the critical data from all these sites is replicated to the high-end Oracle ZFS Storage Appliance system. Each site's Oracle ZFS Storage Appliance system may differ in the mode of replication, size, and I/O access pattern.

A midrange Oracle ZFS Storage Appliance system is used as storage for Microsoft Exchange Mail Server over iSCSI, and VMware ESX disks over NFS and iSCSI. Scheduled replication is set up for Microsoft Exchange Mail Server to replicate every night for backup purposes. A virtualized storage project is scheduled for replication every four hours.

A low-end Oracle ZFS Storage Appliance system is used as a repository for the application layer of the infrastructure. It does not demand too much of the storage. A manual replication is set up and will be triggered whenever any of the components are changed.

All these systems have a dedicated gigabit Ethernet infrastructure to the target site.

## Target Site Configuration

In this example, the target site is used for several purposes:

- As a disaster recovery site for all the applications
- To serve as test/development/report
- For backups

A high-end Oracle ZFS Storage Appliance system is configured to handle the workload and to provide high-availability capability. One controller of the cluster node is configured to receive the application components from Oracle ZFS Storage Appliance midrange and low-end systems. The second head is the target for the mail server and VMware components from the Oracle ZFS Storage Appliance midrange system. The cluster controllers also are used as the disaster recovery destination for all the applications.

## Target Backup

Two backup servers are used at the target. One is to perform the NDMP backup of back-office databases using "local" mode.

The other backup server is used for incremental backups and doing some non-NDMP backup. The tape library is attached to the backup media server. The replicas of Microsoft Exchange and VMware disks are cloned. The iSCSI replica from the mail server is backed up as a full backup. Then a combination of full and incremental backup is initiated from all these clones.

In all the backup setups, the replica is never destroyed.

## Testing and Reporting at the Target Site

The site is also used for testing and reporting purposes. Any application development or QA testing can be done on the replica. A snapshot is taken from the cloned replica to revert quickly to the previous state while doing the testing.

## Failover to the Target

All the applications can be failed over to the target site. Production failover or a role reversal can be performed, depending on the type of application and the type of failure that occurs at the source site.

## Target Considerations

When considering a target site for testing, reporting, QA, backup, and failover, some of the items that need to be identified and planned include the following:

- Capacity requirement to hold all the data
- Storage tiering and archiving requirement
- Storage profile (mirror or RAIDZ-DP preferred)
- SLAs for failover applications
- Network bandwidth to handle all the concurrent replication updates

# FEATURES AND BENEFITS OF REMOTE REPLICATION

The remote replication capability of the Oracle ZFS Storage Appliance provides the following features:

- Replication is supported across Oracle ZFS Storage platforms and across storage profiles.
- Replication can be performed either at the project level or at the share level.
- No dedicated link is necessary; any network can be used for replication. However, in a more complex environment requiring predictable and guaranteed bandwidth for a replication service, a separated subnet/network to be used for replication is highly recommended.
- By default, the auto-tune compression and adaptive multithreading function is enabled for replication actions, providing more efficient network bandwidth utilization, which is especially important in slower WAN-type replication networks. This function can always be disabled at will if it turns out that data that is replicated cannot be sufficiently compressed to provide any replication performance gain.
- Data deduplication over the wire is an extra option that can be used to further increase network bandwidth utilization efficiency. This option should be used carefully as it requires extra CPU and memory resources from the Oracle ZFS Storage Appliance system.
- SSL encryption of the replication data stream sent over the network is an optional feature for data protection, and it can be disabled when performance is more important.
- A project from the source can be replicated to one or more targets with different schedules.
- For faster and efficient target site catch-up, only changes are replicated (except during the initial replication).
- When a replication operation is interrupted by a network failure, system outage or operator action, the replication will be automatically resumed from the point of disruption.
- Administrators have the flexibility to change the mode of replication from one mode to another.
- The target site replica can be used for many purposes, alleviating the performance pressure from the source site.
- The clone from the replica is treated as any other project, including taking snapshots of the shares, which provides more flexibility for QA and test purposes.
- Different retention policies can be specified at source and target for scheduled snapshots of projects/shares being replicated.
- Deleting a manually created snapshot at the source side does not automatically lead to the same snapshot being deleted at the target side. This enables more flexibility for the use of snapshots for creating backups at the target site.
- Efficient single-click reverse replication is provided to reverse the role of the source and target for a project, enabling faster DR.

# CONSIDERATIONS FOR REMOTE REPLICATION USE

- Synchronous mode is not supported, so a zero data-loss requirement cannot be met. Consider using scheduled replication mode at 5-minute intervals to provide an alternative with minimal data loss in the event of a disaster.
- The write ordering and write consistency is maintained at the granularity of the replicated component. The write ordering is preserved within the share if the replication is set at the share level. However, the write ordering is not preserved across the shares if more than one share is replicated. The write ordering at the target for all the shares in the project is preserved if the replication happens at the project level. The write ordering is not preserved across the projects. Refer to the administration guide of a particular model of Oracle ZFS Storage Appliance for details.
- Business availability and recovery objectives, such as RTO, RPO, and SLA, should be considered in deciding the mode of replication. The rate of change, latency, bandwidth, and number of projects to replicate all influence the decision-making process.
- The target site is not verified for the space requirement when the replication is established. Before initiating the replication, it is necessary to verify that the target site has enough storage space to receive the replica.

## CONCLUSION

This paper explains considerations for planning and implementing remote replication with Oracle ZFS Storage Appliance systems. An example architecture and various options are described to enable system planners to simplify offsite data protection. Administrators can use these guidelines in tailoring their application of remote replication to their particular storage environments.

## REFERENCES

- Oracle ZFS Storage Appliance product information:
  - https://www.oracle.com/storage/nas/

- Recommended related Oracle ZFS Storage Appliance white papers in http://www.oracle.com/technetwork/server-storage/sun-unified-storage/documentation/index.html white paper repository.
  - Implementing Remote Replication Using Oracle ZFS Storage Appliance
  - Networking Best Practices with Oracle ZFS Storage Appliance
- Oracle ZFS Storage Appliance documentation library, including installation, analytics, customer service, and administration guides: https://docs.oracle.com/cd/F13758_01/index.html

## REPLICATION TERMINOLOGY, ABBREVIATIONS, AND TERMS INDEX

**Clone**: A clone action on a filesystem or LUN creates an independent copy of a filesystem or a LUN.

**Clone a replica**: This action creates a read/write clone of the replica at the target site. The clone is used to promote the most recent received snapshot to a full r/w copy of that replication package.

**Replication source**: This is an Oracle ZFS Storage Appliance system node that is configured to send replication data to another system's node (target).

**Replication target**: This is an Oracle ZFS Storage Appliance system node that is configured to receive replication data from another system node (source).

**Replication action**: This action specifies a replicating action from the current node (source) to a target for a specific project/share. The action defines various replication attributes such as target address and pool, security, deduplication, thresholds for recovery point objective (RPO) target monitoring, and so forth. Replication actions are run by user-defined schedules.

**Replication schedule**: Each replication action can contain one or more defined schedules. A schedule defines the starting point of each replication action. Schedule definitions are defined by user RPO requirements. RPOs can be monitored by defining RPO level thresholds.

**RPO**: Recovery point objective (RPO) is the point in time that the restarted business process has to go back to "find" the last consistent copy of data to work with. Basically, RPO is the rollback that will occur as a result of the recovery. To reduce a RPO it is necessary to increase the synchronicity of data replication.

**RTO**: Recovery time objective (RTO) is the time that is needed to have a business (process) up and running again after a disaster.

ROC: The rate of change (ROC) is the amount of data that is changed over time on the data volumes that are to be replicated to a second location. ROC can be expressed as a peak value or as an average over a period of time, such as a day. The ROC is used in combination with RPO requirements when determining the speed requirements of a replication link.

**WOC**: When replicating data on multiple volumes the order of updates (in time) to each volume must be kept consistent between the primary and target repository. This is referred to as write order consistency (WOC). When replicating a data repository spread over multiple volumes, all volumes must be kept in the same project in the Oracle ZFS Storage Appliance system, and the data replication must be set up at project level.

CONNECT WITH US

Call +1.800.ORACLE1 or visit oracle.com.
Outside North America, find your local office at oracle.com/contact.

blogs.oracle.com          facebook.com/oracle          twitter.com/oracle

White Paper Architecture Principles and Implementation Practices for Remote Replication Using Oracle ZFS Storage Appliance

Author: Peter Brouwer (original author), Cindy Swearingen (update)

October 2020