ESCALA

Application Roll-over Facility V7

Administrator's Guide

REFERENCE
86 A2 95EF 14

ESCALA

Application Roll-over Facility V7

Administrator's Guide

Software

November 2009

Bull Cedoc
357 avenue Patton
BP 20845
49008 Angers Cedex 01
FRANCE

REFERENCE
86 A2 95EF 14

Trademarks and Acknowledgements

We acknowledge the right of proprietors of trademarks mentioned in this manual.

All brand names and software and hardware product names are subject to trademark and/or patent protection.

Quoting of brand and product names is for information purposes only and does not represent trademark misuse.

Table of Contents

List of Figures

Preface

Intended Readers

This guide is intended for administrators of AIX systems who need to install, configure and maintain *Application Roll-over Facility* software.

Note *Application Roll-over Facility Administrator's Guide* is delivered on the ARF media in PDF format. It is also available on the Bull Support Web site at: http://support.bull.com/ols/product/system/aix/infodoc.

Read the Software Release Bulletin

Read carefully the SRB *(Software Release Bulletin*) for *Application Roll-over Facility* that comes with the software.

The SRB includes environment requirements and restriction as well as late-breaking news. It also includes the procedure to install the software.

Highlighting

The following highlighting conventions are used in this guide:

Bold	Identifies the following: • Interface objects such as menu names, labels, buttons and icons. • File, directory and path names. • Keywords to which particular attention must be paid.
Italics	Identifies references such as manuals or URLs.
`monospace`	Identifies portions of program codes, command lines, or messages displayed in command windows.
< >	Identifies parameters to be supplied by the user.

Related Publications

For more information about the administration tasks of AIX systems, refer to the AIX documentation, which is delivered with the AIX system.

Chapter 1. Concepts Overview, Requirements, Licenses

Application Roll-over Facility is designed to move easily an application and its environment from one system to another system for maintenance purpose, or in answer to hardware failure, system overload or disaster recovery.

Application Roll-over Facility is based on classical high availability configuration using at least two nodes and shared disk storage. *Application Roll-over Facility* is designed to detect system failure, to notify it to the system administrator who can check the situation and manage failover to a recovery node.

1.1. Functionalities

Application Roll-over Facility enables administrators to cleanly stop one or more applications running on a system, and to relocate them on another system. With *Application Roll-over Facility* it is possible to balance the workload across the nodes, to perform maintenance operations on a system keeping the applications available, or, in case of node failure, to move automatically applications. *Application Roll-over Facility* also detects and notifies node errors. The Application Monitoring periodically retrieves the status of running applications and execute appropriate actions when failures are detected.

Application Roll-over Facility is based on two main functionalities:

- **heartbeat mechanism** to monitor the nodes availability,
- **application relocation** to move applications from one node to another, either automatically (in case of node failure) or under administrator control.

Application Roll-over Facility can be configured in **automatic** or **manual take over** mode :

- When the **automatic take over mode** is set for one or several applications, the heartbeat mechanism must be used (the monitoring daemon must be started). In case of node failure, the applications are taken over by the node defined as "take over node".
- When the **manual take over mode** is used, the heartbeat mechanism and the application relocation work independently:
  - the heartbeat mechanism warns the administrator in case of node unavailability,
  - the administrator uses the application relocation if he decides to move the application from one node to another. It is the administrator's responsibility to relocate the application (it is not an automatic action).

Application Roll-over Facility provides sample application scripts and templates.

Terminology: Local Node / Remote Node

By convention:

- "**local node**" refers to the node on which you will run the tool to configure *Application Roll-over Facility*,
- "**remote node**" refers to the other node(s).

1.1.1. Heartbeat Mechanism

On each node, a daemon regularly sends heartbeats to the other daemons, using each configured network, in order to detect node failure and to warn the administrator of the node status.

To detect if a node is failed, when the heartbeats are not received anymore by none of the networks, the local heartbeat daemon tries to ping several addresses on the local area network. If other addresses are reachable, this means that the node is probably failed or that it has a network problem (for example an adapter failure). If no address is reachable, a local area network failure is detected.

When a remote monitored node is detected as failed, the local monitoring daemon checks if it has to recover an application or if it has just to warn the administrator about the node status.

In case of system outage, a notification allows the system administrator to determine if the recovery action must be performed immediately (for example, a failure could result in an overloaded node). This manual failover mechanism allows the administrator to avoid such a situation. Once the take-over decision is made, the administrator must perform the application failover using the appropriate SMIT menus.

Application Roll-over Facility offers a set of tools to define the heartbeat environment and to activate and de-activate the monitoring.

There are two types of heartbeat mechanisms:

- IP heartbeating
- Disk heartbeating

The mechanism for both types is illustrated below:

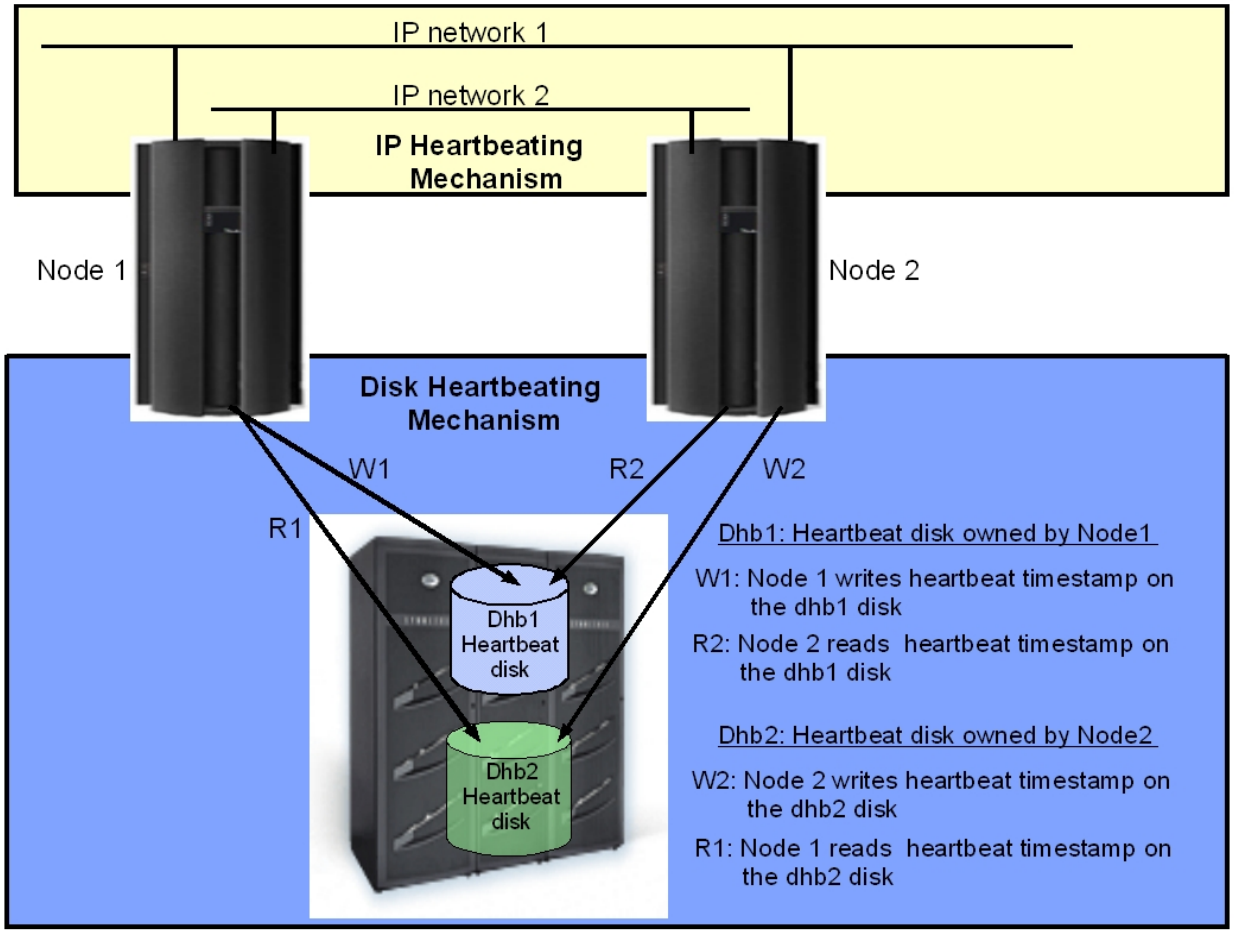


Figure 1. IP and Disk Heartbeating mechanism

IP heartbeating:

IP heartbeating consists in sending heartbeat messages on the IP configured networks.

Disk heartbeating:

The disk heartbeating mechanism is different from IP heartbeating in the sense that it does not use the TCP/IP stack to monitor the ARF nodes.

Disk heartbeating uses disks shared by all the cluster nodes and configured for this purpose. Each cluster node writes a timestamp on a disk it owns. Conversely, each cluster node reads the timestamp on all the disks it does not own.

To allow each cluster node to read the heartbeat timestamp on all disks, it is necessary to set the disk reservation attributes as follows:

- **reserve_lock** (for PowerPath) to 'no'.
- **reserve_policy** to 'no_reserve'.

Notes
- The recommended size (lun size) of the disks used for heartbeat is 1 MB.
- The disk managed by the heartbeat mechanism is used in raw mode, so do NOT create a volume group, logical group or file system on the disk.

1.1.2. Application Definition and Relocation

An application is associated with the environment needed to be executed correctly on a group of nodes.

The nodes can communicate each other through one or more networks and share storage subsystems. An address can be defined for each node, on each network, allowing to communicate with other nodes.

An **application environment** is composed of:

- Resources such as:
  - a list of specific addresses used by the clients of the application,
  - a list of volume groups and/or file systems on shared disks containing application data.
- Methods to start and stop the application.

To activate an application means to make its resources available and to launch the application on a node previously defined. Before and after each step of application activation, it is possible to execute specific methods. If resources cannot be made available, the activation is aborted.

To de-activate an application means to stop application processes and to release its resources on a node, as wished.

Before and after each step of application de-activation, it is possible to execute specific methods. If resources cannot be released, the de-activation is aborted.

For management purpose, the execution of activation and de-activation is logged and main events are recorded.

Application Roll-over Facility offers a set of tools to define nodes and application environment, to activate and de-activate application, and to display information about the configuration.

1.1.3. Application Monitoring

Application Monitoring allows you to check periodically the status of *Application Roll-over Facility* applications. It is performed by a daemon started automatically at the boot time, or by the administrator.

An *Application Monitoring* is associated to an application as defined in *Application Roll-over Facility*. It is possible to define several *Application Monitorings* and associate them to the same application, allowing the same application to be monitored partly in one way and partly in another way.

1.1.4. Running an application in an Application Workload Partition (WPAR)

Usually *Application Roll-over Facility* starts an application in the system environment. It is also possible to run an application in a workload partition (WPAR).

1.1.5. Disaster Recovery

Application Roll-over Facility supports disaster recovery solutions using AIX mirroring or Data Replication mechanisms.

1.2. Standard Edition / Enterprise Edition

Application Roll-over Facility is available in two different editions, in order to provide high availability and disaster recovery for any enterprise requirement:

- *Application Roll-over Facility Standard Edition* for building a high availability solution at a local single site. It includes the NFS option.
- *Application Roll-over Facility Enterprise Edition*, for off-site recovery with the support for thirdparty storage replication technologies. The following features are supported:
  - GLVM (Geographic Logical Mirroring)
  - MirrorView for EMC CLARiiON subsystems
  - MetroCluster for NetApp subsystems
  - RDR/RDR-CO (Remote Data Replication/Consistency Option) for StoreWay FDA subsystems
  - SRDF (Symmetrix Remote Data Facility) for Symmetrix subsystems
  - ERM (Enhanced Remote Mirroring) for IBM DS5020 / DS4000 subsystems

1.3. License Key

All servers that use *Application Roll-over Facility Standard Edition* require a product license key.

1.3.1. Temporary Key

A temporary key is contained in a file included on the enclosed CD-ROM. This key is valid for a limited period (usually 3 months), for any server.

License Key for Enterprise Edition

An additional product license key is required to use *Application Roll-over Facility Enterprise Edition*. As for basic license key, a temporary key, valid for a limited period, for any server, is provided on the enclosed CD-ROM.

1.3.2. Definitive Key

A definitive key can be obtained for each server using *Application Roll-over Facility Standard Edition*. This key is valid for an unlimited period and only for the server whose system id is the one you have provided to obtain the key.

As for basic license key, a definitive key can be obtained for each server using *Application Roll-over Facility Enterprise Edition*.

To obtain the definitive key, you can either:

- contact your Bull representative
- or, if you have a support contract, order the key directly on the Bull Support Web site:
  http://support.bull.com/ols/online/keys
  Your WAC (Web Access Code) and your order number are required.

To obtain the definitive keys you have to provide the system ID of all the servers. (The system ID is the result of the **uname -Mu** command) .

1.4. Hardware Requirements

1.4.1. Servers Requirements

Application Roll-over Facility works with SMP (Symmetrical Multiprocessor System) servers in a "no-single-point-of-failure" server configuration. *Application Roll-over Facility* supports the Bull Escala models designed for server applications and meets the minimum requirements for internal memory, internal disk, and I/O slots.

The minimum configuration and sizing of each machine is highly dependent on the user's database package and other applications.

Actual configuration requirements are highly localized according to the required function and performance needs of individual sites.

1.4.2. Network Requirements

It is highly recommended to have two physically different IP networks, in order to make the difference between network failure and node failure:

- A **main network**, used by the Disk heartbeating, on which the application addresses will be aliased.
- A **second network**, used by the IP heartbeating or Disk heartbeating mechanisms to avoid cluster partitioning.

A standard hardware configuration requires:

- For the main network, at least two network adapters to configure an Etherchannel interface (network availability is ensured by Etherchannel interface),
- Two I/O adapters to support I/O multipathing mechanism.

1.4.3. External SCSI Disk Devices

If you have SCSI Disk devices, you must verify that each SCSI device connected to the shared SCSI bus has a unique ID. For example:

- For SCSI-2 Differential adapters enter the command:

```
lsattr -E -l scsi1 | grep id
```

- For SCSI-2 Differential Fast/Wide adapters, enter the command:

```
lsattr -E -l ascsi1 | grep external_id

      SCSI ID
Id    7     Adapter Card SCSI ID
```

To change the SCSI ID, use the **chdev** command, or use the **smit chgscsi** fast path. Example:

```
chdev -l scsi1 -a id=6
```

1.5. Software Requirements

- AIX5L V5.2, 5.3 or AIX Version 6.1.

  AIX Version 6.1 and later is mandatory if you plan to run an application in a WPAR or if you want to use Live Partition Mobility feature.

- Software for I/O multipathing

Chapter 2. Installation and Configuration Overview

This chapter describes the installation and configuration procedure.

2.1. Installing and Configuring Application Roll-over Facility

Install and Configure SSH

With ssh, the security of the system is enhanced. Install and configure ssh for use in the *Application Roll-over Facility* configuration.

See *Installing and Configuring SSH*, on page 3-1.

Install *Application Roll-over Facility* Software

Install *Application Roll-over Facility* software on each *Application Roll-over Facility* node and install license keys.

See *Installing Software and License Keys*, on page 4-1.

Tailor AIX for *Application Roll-over Facility*

Review or edit various AIX files to ensure a proper configuration for network options and for various host files.

See *Tailoring AIX for Application Roll-over Facility*, on page 5-1.

Define Shared LVM Components

Create the shared volume groups, logical volumes, and file systems for your *Application Roll-over Facility* configuration.

See Appendix A *Defining Shared LVM Components*, on page A-1.

Configure the *Application Roll-over Facility* Software

Define the components of your *Application Roll-over Facility* configuration.

See *Configuring an Application Roll-over Facility Environment*, on page 6-1.

Configure a Fibre Channel Network for heartbeat

A Fibre Channel Network can be used for heartbeat monitoring.

See Appendix C *Configuring Fibre Channel Network for ARF*, on page C-1.

Configure AIX mirroring for Disaster Recovery

AIX mirroring function can be used for disaster recovery solution.

See *Configuring AIX Mirroring for Disaster Recovery*, on page 7-1.

Set Up ARF for Live Partition Mobility

Live Partition Mobility allows you to migrate the running AIX partitions and their hosted applications from one physical server to another without disrupting the infrastructure services.

See *Support of Live Partition Mobility*, on page 8-1.

Configure DLPAR/PoD Resources

On partitioned systems, to respond to workload peaks, Dynamic Logical Partitioning (DLPAR) and Power on Demand (POD) can be used.

See *Configuring DPLAR/PoD Resources for ARF*, on page 9-1.

2.2. Configuring Disaster Recovery Optional Features

The following options are available with *Application Roll-over Facility Enterprise Edition*.

Configure GLVM

GLVM (Geographic Logical Mirroring) is an AIX feature for real time geographic data mirroring over standard TCP/IP networks.

See *Configuring GLVM for ARF*, on page 10-1.

Configure MirrorView for ARF

MirrorView is an EMC feature providing disaster recovery to protect your most critical data in the event of an outage.

See *Configuring MirrorView for ARF*, on page 11-1.

Configure NetApp MetroCluster

NetApp MetroCluster is a unique synchronous replication solution protecting your critical data against site disasters. It provides the capability to force a failover when an entire storage system (including the controllers and storage) is destroyed or unavailable.

See *Configuring NetApp MetroCluster for ARF*, on page 12-1

Configure RDR/RDR-CO

RDR/RDR-CO (Remote Data Replication/Consistency Option) is a StoreWay FDA feature providing disaster recovery protection.

See *Configuring RDR/RDR-CO for ARF*, on page 13-1.

Configure SRDF

Symmetrix Remote Data Facility (SRDF®) is a business continuance solution that maintains a mirror image of data at the device level in Symmetrix® arrays located in physically separate sites.

See *Configuring SRDF for ARF*, on page 14-1 .

Configure ERM

Enhanced Remote Mirroring (ERM) is an option of the IBM DS4000 Storage Manager software and is used for replication data between DS4000 Storage Subsystem over a remote distance.

In the event of disaster or unrecoverable error at one storage system, *Application Roll-over Facility* promotes automatically the second storage system to take over responsibility for normal I/O operations.

See *Configuring Enhanced Remote Mirror (ERM) for ARF*, on page 15-1

2.3. Managing ARF

Set up and use ARF Watch

Set up ARF Watch, a monitoring facility which helps the administrator to detect critical situations. Then use the Graphical User Interface to monitor your *Application Roll-over Facility* configuration.

See *Setting Up and Using ARF Watch*, on page 16-1 for details.

View the ARF configuration

To display the configuration, to show the application and monitoring status, see *Viewing the Configuration*, on page 17-1.

Maintain the ARF Environment

To activate/deactivate application environment and monitoring, to modify an ARF configuration, to customize pre and post events, see *Maintaining the ARF Environment*, on page 18-1.

Check the status of running applications

The Application Monitoring allows the administrator to periodically check the status of running applications.

See *Running Application Monitoring, on page 19-1* for details.

Save and restore ARF configuration

The snapshot utility allows saving and restoring an ARF configuration.

See *Saving and Restoring Configurations*, on page 20-1 for details.

Diagnosing the ARF resources

The diagnostic tool allows you to generate diagnostic information.

See *Diagnosing the ARF Resources*, on page 21-1 for details.

Troubleshooting

See *Troubleshooting*, on page 22-1 for a description of the log files and ODM files that can help you to diagnose problems.

Chapter 3. Installing and Configuring SSH

This chapter describes how to install and configure SSH.

Note SSH is used by default. To use RSH instead of SSH, set the value of the **"Use ssh?"** field to 'no', as described in *Configuring the Node Monitoring*, on page 6-13.

3.1. Installing SSH

The **openssh** fileset and its pre-requisite **openssl** rpm package must be installed on each *Application Roll-over Facility* node.

If virtual disks (from Virtual I/O Servers) are used as "Shared Volume Groups" in an *Application Roll-over Facility* configuration, **openssh** and **openssl** must also be installed on the Virtual I/O servers.

Install first the **openssl** fileset from the *AIX Toolbox for Linux Applications* CD, then install the **openssh.base** fileset from the *Expansion Pack* CD.

3.2. Configuring SSH Access

Configuring SSH access consists of generating the ssh keys (with "no password access") on each server and then propagating the ssh keys on all the servers (nodes and VIO servers) of the configuration.

3.2.1. Generating the SSH keys

To generate an RSA key pair of the ssh protocol:

1. Log in as **root** user (or as **arfw** user for configuring SSH access for ARF Watch).
2. Enter the command **ssh-keygen -t rsa**.
3. Accept default location file and do not enter passphrase.

A text similar to the following is displayed:

```
Generating public/private rsa key pair.
Enter file in which to save the key (//.ssh/id_rsa):
Enter passphrase (empty for no passphrase):
Enter same passphrase again:
Your identification has been  saved in //.ssh/id_rsa.
Your public key has been saved in //.ssh/id_rsa-pub.
The key finger print is:
d6:3f:11:da:44:63:ee:17:0a:e0:98:ca:3b:16:4d:fe root@nodeA
```

3.2.2. Propagating the SSH keys

1. Append the contents of **~/.ssh/id_rsa.pub** to **~/.ssh/authorized_keys** on the local server:

   for the root user:

   ```
   cat ~/.ssh/id_rsa.pub >> ~/.ssh/authorized_keys
   ```

   for the arfw user:

   ```
   cat ~/.ssh/id_rsa.pub >> ~/.ssh/authorized_keys
   ```

2. Propagate the ssh keys to the remote servers:

   ```
   cat ~/.ssh/id_rsa.pub|ssh <remote server> 'cat - >>
   ~/.ssh/authorized_keys'
   ```

3. Verify the communication between the local node (for example: *nodeA*) and all the nodes (for example: *nodeB*) and all the virtual I/O servers. Run the following steps:

   - Enter the following command (no password prompt):

     ```
     ssh nodeB date
     ```

   - The first time you run this command, the following message is displayed:

     ```
     The authenticity of host 'nodeB (172.18.1.220)' can't be established.
     RSA key fingerprint is
     ed:6e:a5:32:55:84:d6:25:a6:11:d0:69:1d:0c:11:23.
     Are you sure you want to continue connecting (yes/no) ?
     ```

   - Answer : **yes**

   - The following message is displayed including the output of the date command:

     ```
     Warning: Permanently added 'nodeB, 172.18.1.220' (RSA) to the list of
     known hosts.
     Mon Oct 8 16:54:12 CDT 2007
     ```

   - The **~/.ssh/known_hosts** file on nodeA is updated with authorization to the nodeB address.

4. Verify by entering again the command:

   ```
   ssh nodeB date
   ```

   The output of the date comamnd is immediately displayed:

   ```
   Mon Oct 8 17:05:02 CDT 2007
   ```

5. Repeat the steps 1. to 4. on each server (node and VIO servers) of the *Application Roll-over Facility* configuration.

3.2.3. Authorizing additional node addresses for SSH

If the node has several addresses defined in the *Application Roll-over Facility* configuration, (**Addresses List** field in **Managed Node** menu), you have to authorize each additional address on all nodes for SSH.

For example, if NodeB has two addresses: **nodeB** and **nodeB1**, to authorize nodeB1 address for ssh you have to run steps 3. and 4. from NodeA, and to perform this operation on all nodes for all additional addresses.

3.3. Additional Tasks

If you plan to use **DLPAR and On/Off PoD resources** with the *Application Roll-over Facility* application, you have to authorize HMC(s) ssh access for all nodes. Refer to chapter *Configuring DLPAR and On/Off Pod Resources with ARF*, section *Enable no password HMC access*, on page 9-3.

If you plan to use **ARF Watch**, you have to authorize ssh access for all nodes for the **arfw** user. To do this, refer to chapter *Using ARF Watch*, section *Setting_up ARF Watch*, on page 16-2.

If you plan to use **virtual disks from Virtual IO Server(s)** with the *Application Roll-over Facility* application, you have to authorize ssh access for all nodes to the **root** user. To do this, refer to chapter *Configuring an Application Roll-over Facility Environment*, section *Configuring Virtual Disks for an Application Roll-over Facility Cluster*, on page 6-15.

Chapter 4. Installing Software and License Keys

This chapter gives recommendations related to installation and configuration.

4.1. Read the Software Release Bulletin

Read carefully the SRB *(Software Release Bulletin*) for *Application Roll-over Facility* that comes with the software.

The SRB includes environment requirements and restriction as well as late-breaking news. It also includes the procedure to install the software.

4.2. Installing Application Roll-over Facility

Application Roll-over Facility product is NOT factory pre-loaded.

You must install it as described in the *SRB for Application Roll-over Facility.*

Then, you can check the installed LPPs by entering the following command **on each node** of your configuration:

```
lslpp –L Bull.approllf.*
```

4.3. Installing License Key(s)

To obtain the key file, refer to *License key*, on page 1-5. Then copy this file on each server, as described below.

4.3.1. Standard Edition

If you use *Application Roll-over Facility Standard Edition*, install the corresponding license key as follows.

- If it is a temporary key (for example a temporary key in the **barf.key.tmp** file):
  - Copy the **barf.key.tmp** file into the **/usr/sbin/barf/data/barf.key** file on each server.
- If it is a definitive key (for example a definitive key in the **barf.key.foo** file for the **foo** server and in the **barf.key.bar** file for the **bar** server):
  - Copy the **barf.key.foo** file into the **/usr/sbin/barf/data/barf.key** file on the **foo** server .
  - Copy the **barf.key.bar** file into the **/usr/sbin/barf/data/barf.key** file on the **bar** server.
- If you plan to use **Live Partition Mobility** feature, refer to *Support of Live Partition Mobility*, on page 8-1 for key management.

4.3.2. Enterprise Edition

If you use *Application Roll-over Facility Enterprise Edition*, install the corresponding license key as follows.

- If it is a temporary key (for example a temporary key in **barf.ee.key.tmp** file):
  - Copy the **barf.ee.key.tmp** file into the **/usr/sbin/barf/data/barf.ee.key** file on each server.

- If it is a definitive key (for example a definitive key in the **barf.ee.key.foo** file for the **foo** server and in the **barf.ee.key.bar** file for the **bar** server):
  - Copy the **barf.ee.key.foo** file into the **/usr/sbin/barf/data/barf.ee.key** file on the **foo** server.
  - Copy the **barf.ee.key.bar** file into the **/usr/sbin/barf/data/barf.ee.key** file on the **bar** server.

Chapter 5. Tailoring AIX for Application Roll-over Facility

This chapter discusses several general tasks necessary to make sure that your *Application Roll-over Facility* environment works as planned.

5.1. Overview

The following AIX items must be configured as expected in an *Application Roll-over Facility* configuration:

- Users and groups
- **/etc/hosts file**
- **/.rhosts** file (if rsh is used) or **authorized _keys** and **known_hosts** files (if ssh is used)
- Error notification facility
- Shared LVM components.

5.2. Checking Users, Groups and Passwords

If a node fails, users should be able to log on to the surviving nodes without experiencing problems caused by mismatches in the user or group IDs. To avoid mismatches, make sure that user and group information is propagated to nodes as necessary. User and group IDs should be the same on all nodes.

5.3. Updating /etc/hosts and Name Server Configuration

When applying configuration settings, the configuration tool must be able to access the remote nodes in order to run appropriate configuration commands on this node.

Consequently, for the configuration to work properly, you have first to update the **/etc/hosts** file on both local and remote nodes.

- Add entries for all interfaces of the *Application Roll-over Facility* remote nodes.

  Edit the **/etc/hosts** file (and the **/etc/resolv.conf** file, if the name server configuration is used) on each node in *Application Roll-over Facility* configuration to make sure that the IP addresses of all *Application Roll-over Facility* interfaces are listed.

- Also, make sure that the **/etc/hosts** file has the following entry:

```
127.0.0.1   loopback    localhost
```

5.4. /etc/inittab File

The **/etc/inittab** file is automatically modified by *Application Roll-over Facility* scripts to remove temporary files used by *Application Roll-over Facility* and to automatically start *Application Roll-over Facility* monitoring mechanism.

5.5. Remote Commands (ssh, rsh)

Remote commands are used in many *Application Roll-over Facility* operations (activation, propagation...). You can use **ssh** or **rsh**.

5.5.1. Updating .rhosts (if rsh is used)

The **/.rhosts** file must be updated on both local and remote nodes, if you plan to use rsh for remote commands.

For security reasons, you can add entries to the **/.rhosts** file only if necessary, and delete them when they are no longer required.

The *Application Roll-over Facility* synchronization and verification functions, however use **rcmd** and **rsh** and thus require these **/.rhosts** entries.

To ensure that it will work as expected, edit the **/.rhosts** file on each node of the configuration. Add entries for all the nodes, with access right granted to root.

5.5.2. Updating ssh files (if ssh is used)

The **/.ssh/authorized_keys** and **/.ssh/known_hosts** files must be updated if you plan to use ssh for secure remote commands.

In the **/.ssh/authorized_keys** file, add all the ssh nodes keys.

In the **/.ssh/known_hosts** file, add authenticity information for all the nodes addresses defined in the *Application Roll-over Facility* configuration.

5.6. AIX Error Notification Facility

The AIX Error Notification facility allows you to detect an event not monitored by *Application Roll-over Facility* (typically the status of disk resources) and to program a response to the event.

Permanent hardware errors on disk drives, controllers, or adapters may impact the fault resiliency of data. By monitoring these errors through error notification methods, you can assess the impact of a failure on the *Application Roll-over Facility* ability to provide high availability. A simple implementation of error notification would be to send a mail message to the system administrator to investigate the problem further. A more complex implementation could include logic to analyze the failure and decide whether to continue processing, stop processing, or escalate the failure to a node failure and have a takeover node make the volume group resources available to clients.

It is strongly recommended that you implement an error notification method for all errors that affect the disk subsystem. Doing so ensures that degraded fault resiliency does not remain undetected.

- See *Automatic Error Notification*, on page 5-2 for information about the errors that *Application Roll-over Facility* automatically takes into account.
- *Application Roll-over Facility* provides a SMIT interface to configure the AIX Error Notification facility. To implement your own notification methods see *Configuring Error Notification*, on page 6-10.

5.6.1. Automatic Error Notification

At *Application Roll-over Facility* installation time, error notification is automatically configured for the errors listed below.

By default, the associated notification method performs the following:

- sends a message to the system console,
- sends a mail to the mailing list defined in the *Application Roll-over Facility* Run Time Parameters,
- sends an SNMP trap.

Automatic error notification applies to:

- all HARDWARE PERMANENT errors,
- DISK_ERR3 and SC_DISK_ERR3 errors: the message indicates that a resource is failed and suggests to activate the application on another node, if the failing resource is a disk defined as a resource of an application environment and if there is no more good copies of that disk; if the automatic take over mode is set, the halt -q command is issued to force the application to move from one node to another,
- SC_DISK_ERR7, RPVC_IO_TIMEOUT, BMPIO_ALL_PATHS_DEAD errors: the message indicates that all paths are failed for a disk and suggests to activate the application on another node, if the failing resource is a disk defined as a resource of an application environment and if there is no more good copies of that disk; if the automatic take over mode is set, the halt -q command is issued to force the application to move from one node to another,
- PPATH_PATH_DEL or EMCP_PATH_DEAD errors: the message indicates that a PowerPath trespass occurs for a disk,
- PPATH_DEVICE_GONE or EMCP_VOL_DEAD and EMCP_ALL_PATHS_DEAD errors: the message indicates that a PowerPath trespass failure occurs for a disk and suggests to activate the application on another node, if the failing disk is defined as a resource of an application environment and if there is no more good copies of that disk. If the automatic take over mode is set, the **halt -q** command is issued to force the application to move from one node to another.

5.7. Defining Shared LVM Components

Refer to Appendix A *Defining Shared LVM Components*.

Chapter 6. Configuring an Application Roll-over Facility Environment

This chapter describes how to configure an *Application Roll-over Facility* environment.

6.1. Overview

Perform the following steps to define the *Application Roll-over Facility* configuration:

6.2. Defining a Topology

Complete the following procedure to define the *Application Roll-over Facility* topology. You need to perform these steps only on one node. When you will propagate the configuration its definition will be copied to the other nodes.

6.2.1. Defining Nodes

To define the *Application Roll-over Facility* nodes:

1. Type **smit barf** and select the following options: **Configuration Definition > Managed Nodes > Add a Managed Node** or use the **smit barf_add_node** fast path. When you press Enter, SMIT displays the **Add a Managed Node** following screen :

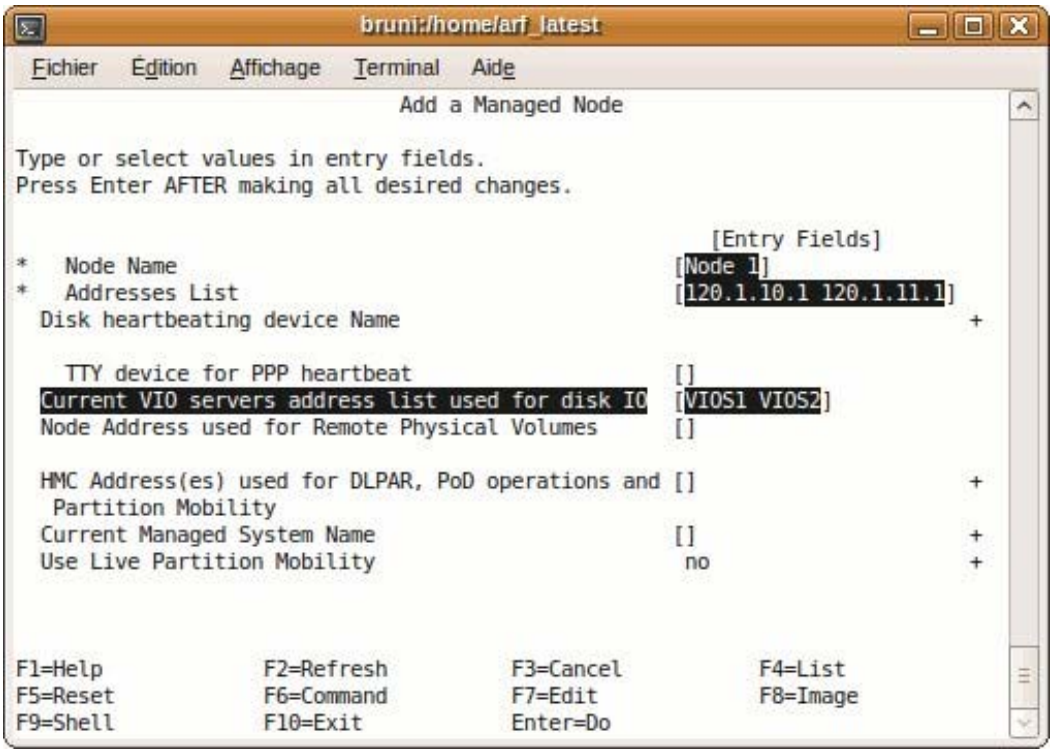


Figure 2. Add a managed node screen

Node Name
: Name of the node. A name cannot exceed 31 characters. It can include alpha and numeric characters and underscores. It is not mandatory to indicate the same name as the host name.

Addresses List
: List of the addresses to reach the node (they must be known in the **/etc/hosts** file). Each address must correspond to a dedicated network physically different.

Disk heartbeating device name
: Device name of the disk heartbeating. Select this name in the list.

TTY device for PPP hearbeat
: TTY device for PPP protocol if you use this feature (old configurations).

Current VIO servers address list used for disk IO
: Name or address of the VIO server.

Node Address used for Remote Physical Volumes
: Node adress for the Remote Physical Volumes client-server communication.
See *Configuring GLVM for ARF*, on page 10-1 for more information.

HMC Address(es) used for DLPAR and PoD operations and Live Partition Mobility
: IP address (dot format) of the HMC that manages the partition node.
See *Configuring DLPAR and On/Off PoD Resources with ARF*, on page 9-1 for more information.

Current Managed System Name
: Managed System Name to which the partition node belongs.
See *Configuring DLPAR and On/Off PoD Resources with ARF*, on page 9-1 for more information.

Use Live Partition Mobility
: Type 'yes' if you plan to use **Live Partition Mobility** feature.
See *Support of Live Partition Mobility*, on page 8-1 for more information.

2. Repeat this operation for all the nodes participating to the configuration.
3. Press F3 until you return to the Configuration Definition screen, or F10 to exit SMIT.

6.2.2. Synchronizing the Node Definition Across Nodes (optional)

You can already synchronize the *Application Roll-over Facility* Definition across the nodes using the following command :

1. Type **smit barf** and select the following options: **Configuration Definition > Propagate Configuration** or use the **smit barf_sync_conf** fast path. When you press Enter, SMIT displays the following screen.

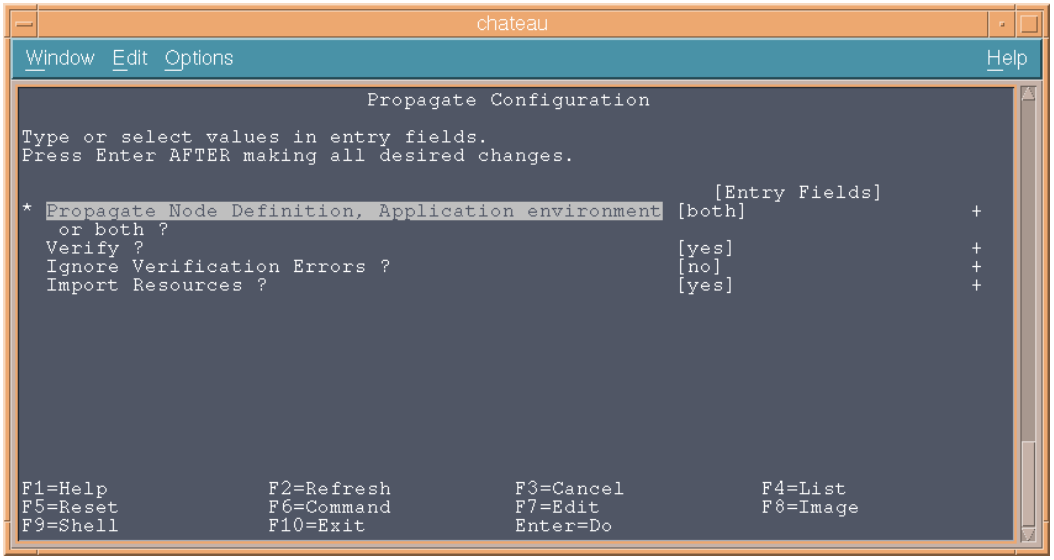


Figure 3. Propagate Configuration screen (Node Definition propagation)

Propagate Node Definition, Application environment or both ?
: Choose 'node' to make only the synchronization of the node definition.

Verify ?
: By default, this field is set to 'yes' and the topology verification program is run. To save time in the synchronization process, you can toggle this entry field to 'no'. Doing so, the verification will be skipped.

Ignore Verification Errors
: By choosing 'yes', the result of the verification is ignored and the configuration is synchronized even if verification fails. By choosing 'no', the synchronization process terminates if errors are found; view the error messages in the system error log to determine the configuration problem.

Import Resources
: If you set this field to 'yes', the synchronization will try to vary on all the volume groups defined on each node with the same major number than the local node, if this number is free, except if the volume group is already imported.

2. Press Enter. The *Application Roll-over Facility* definition (including all nodes information) is copied to the other nodes.
3. Press F10 to exit SMIT.

6.3. Defining Applications

Configuring applications means pointing the *Application Roll-over Facility* event scripts to the scripts that they call to start and stop the application, and taking into account the resources defined for each application.
Note that this section does not discuss writing the start and stop scripts. See the vendor documentation for specific product information on starting and stopping a particular application.

6.3.1. Adding Applications

Complete the following steps to create applications on any *Application Roll-over Facility* node.

1. Type **smit barf** and select the following options: **Configuration Definition> Application Environment > Add an Application Environment** or use the **smit barf_add_appli** fast path. When you press Enter, SMIT displays the **Add Application Environment** following screen:

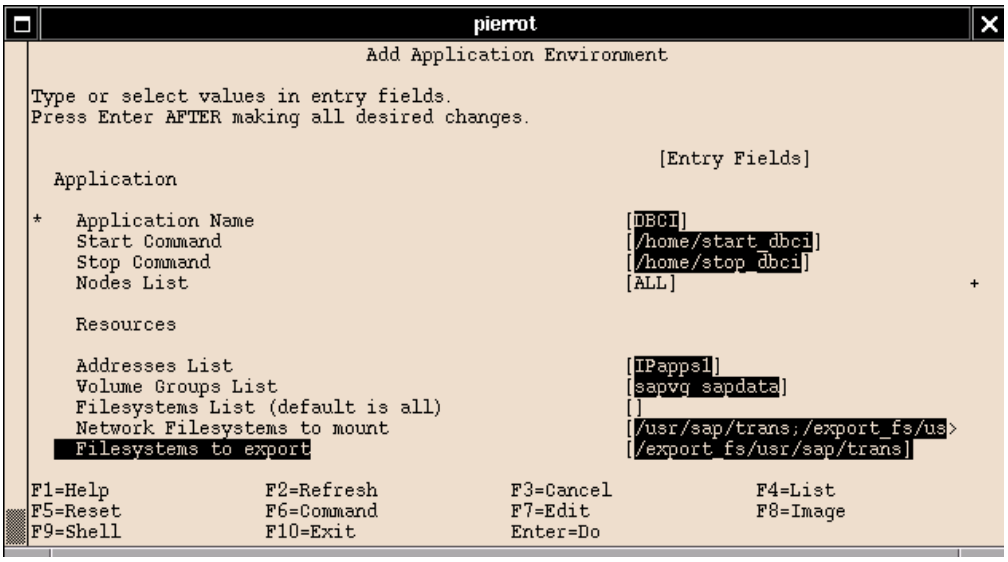


Figure 4. Add Application Environment screen

Application Name	Enter an ASCII text string that refers the application. This name can include alphabetic and numeric characters and underscores. Use no more than 31 characters.
Start Command	Enter the pathname of the script (followed by arguments) called by the event scripts to start the application. This script must be in the same location on each node that might start the application. The contents of the script, however, may differ.
Stop Command	Enter the pathname of the script called by the event scripts to stop the application. This script must be in the same location on each node that may start the server. The contents of the script, however, may differ.
Nodes List	List of the nodes on which the application can be running.
Addresses List	List of the addresses associated to this application.
Volume Groups List	List of all the volume groups that will be used by the application, and will be varied on.
File systems List	List of all the file systems used by the application, and mounted.
Network Filesystems to mount	Identify the filesystems or directories to be NFS mounted.

An NFS mount point is required to mount a filesystem or a directory via NFS. Specify the NFS mount point, then the local mount point, separated by a semicolon. For example:
`/nfs_mount_point;/local_mount_point`
If there are more entries, separate them with a space:
`/nfs_mnt_pt1;/local_mnt_pt1`
`/nfs_mnt_pt2;/local_mnt_pt2`
Note : The NFS mount point must be outside the directory tree of the local mount point.

Filesystems to export — Identify the filesystems or directories to be NFS exported by an application of type *NFS server* to *Application Roll-over Facility* application(s) of type *NFS client*. If there are several entries, separate them by a space.

2. Press Enter to add this information to the ODM on the local node.
3. Press F3 to return to previous SMIT screen to add other application.

6.3.2. Synchronizing the Applications Definition Across Nodes (optional)

You can synchronize the *Application Roll-over Facility* Application Environment across the nodes using the following command:

1. Type **smit barf** and select the following options: **Configuration Definition> Propagate Configuration** or use the **smit barf_sync_conf** fast path. When you press Enter, SMIT displays the following screen.

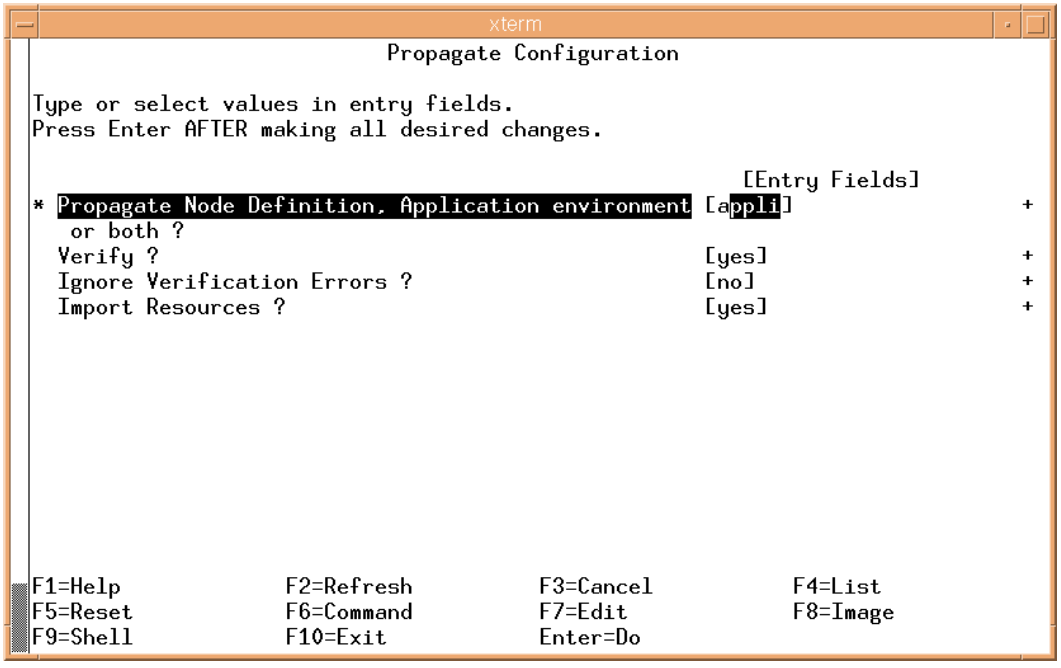


Figure 5. Propagate Configuration screen (Application environment propagation)

Propagate Node Definition, Application environment or both ? — Choose '**appli**' to make only the synchronization of the application environment.

Verify ? — By default, this field is set to '**yes**' and the topology verification program is run. To save time in the synchronization process, you can toggle this entry field to '**no**'. By doing so verification will be skipped.

Ignore Verification Errors — By choosing '**yes**', the result of the verification is ignored and the configuration is synchronized even if verification fails. By

	choosing 'no', the synchronization process terminates if errors are found; view the error messages in the system error log to determine the configuration problem.
Import Resources	If you set this field to 'yes', the synchronization will try to vary on all the volume groups defined on each node with the same major number than the local node, if this number is free, except if the volume group is already imported.

2. Press Enter. The *Application Roll-over Facility* definition (including all nodes information) is copied to the other nodes.
3. Press F10 to exit SMIT.

6.4. Configuring Custom Events

Application Roll-over Facility custom events are scripts defined by the user, which can be executed before or after events, during activation or de-activation of an application.
The following events will be executed for each volum group, or file system or alias: **mountvg**, **umountvg**, **mountfs**, **umountfs**, **configalias**, **unconfigalias**.
To configure events, you indicate the script that handles the event, as described below. You can define multiple customized pre-event and post-event scripts.
To define your customized event scripts, take the following steps.

1. Type **smit barf** and select the following options: **Configuration Definition> Application Environment** or use the **smit barf_conf_app_menu** fast path. SMIT displays the menu choices for adding, changing, or removing a custom Post/Pre event.

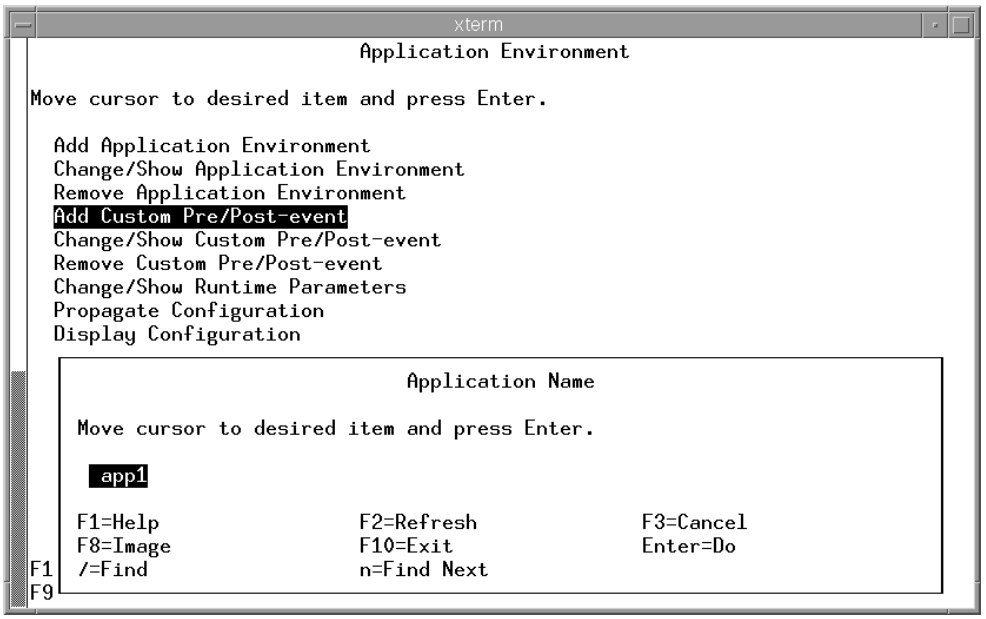


Figure 6. Application Environment screen

2. Select **Add a Custom Post/Pre Event** from the menu. When you press Enter, SMIT displays the following screen:

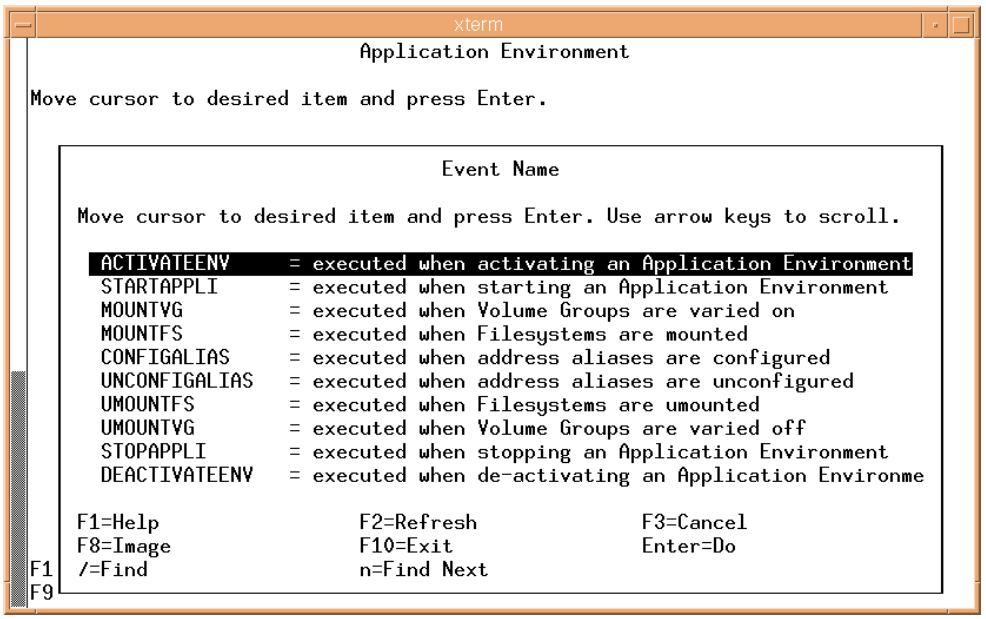


Figure 7. Event Name selection screen

3. Select an event in the list and press Enter to validate. SMIT displays the following screen:

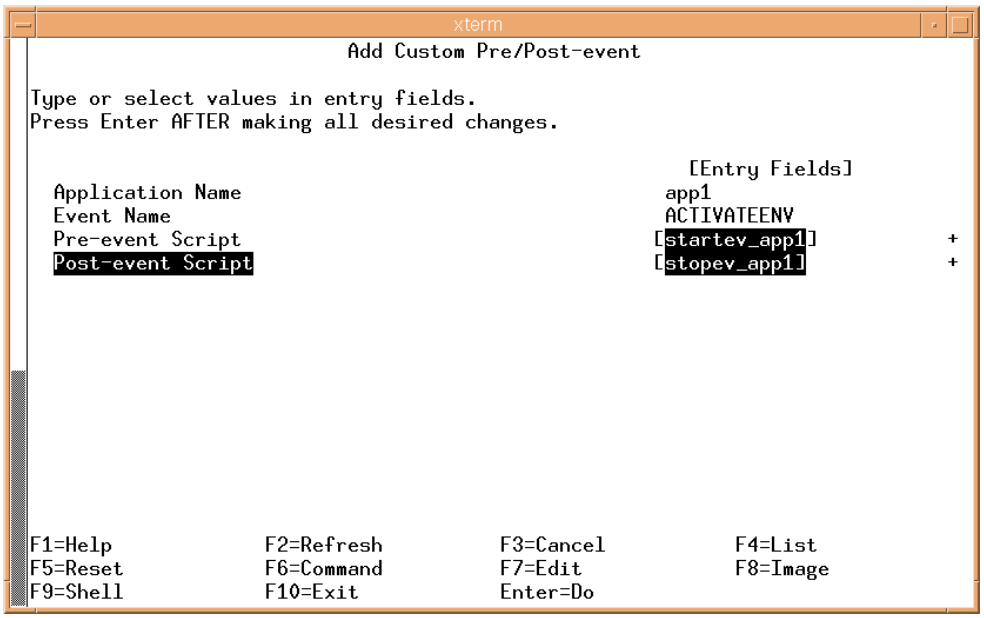


Figure 8. Add Custom Pre/Post-event screen

Application Name	Name of application on which you want to configure events.
Event Name	Type of event you want to custom (chosen from Event Name list).
Pre-Event Script Filename	(optional). If you have defined custom events, press F4 for the list. Or, enter the name of a custom-defined event to run before the event command executes. This command provides pre-processing before an event occurs. Remember that the *Application Roll-over Facility* Manager will not process the event until this pre-event script or command has completed.
Post-Event Script Filename	(optional). If you have defined custom events, press F4 for the list. Or, enter the name of the custom event to run after the event command executes successfully. This script provides post-processing after an event.

4. Press Enter to add the information to *Application Roll-over Facility* in the local ODM.
5. Repeat this procedure for all the post and pre-events you want to add.
6. Synchronize your changes across all nodes by selecting the Propagate Configuration in Configuration Definition > Propagate Configuration or use the **smit barf_sync_conf** fast path and select “appli" for Propagate option.

Note Synchronizing does not propagate the actual new or changed scripts; you must add these to each node manually.

6.5. Customizing Log and Trace Files

You can redirect log files and trace files from their default directory to a directory of your choice. If you do this, keep in mind that the requisite (upper limit) disk space for most log and trace files is 2MB. 14MB is recommended for *Application Roll-over Facility*.

Note Logs and traces should be redirected to local file systems and not to shared or NFS file systems. Having logs and traces on those file systems may cause problems if the file system needs to unmount during a failover event. Redirecting logs to NFS file systems may also prevent *Application Roll-over Facility* services from starting during node reintegration.

The mount point or mount-file system operation must have read-write access.

Be sure to synchronize the configuration directly before redirecting a log file in order to avoid failure of the redirection process.

Redirecting a Log or a Trace File

To redirect a log or a trace file from its default directory to another destination, take the following steps:

1. Type **smit barf** and select **Configuration Definition > Change/Show Runtime Parameters** or use the **smit barf_ch_param** fast path. The following screen is displayed :

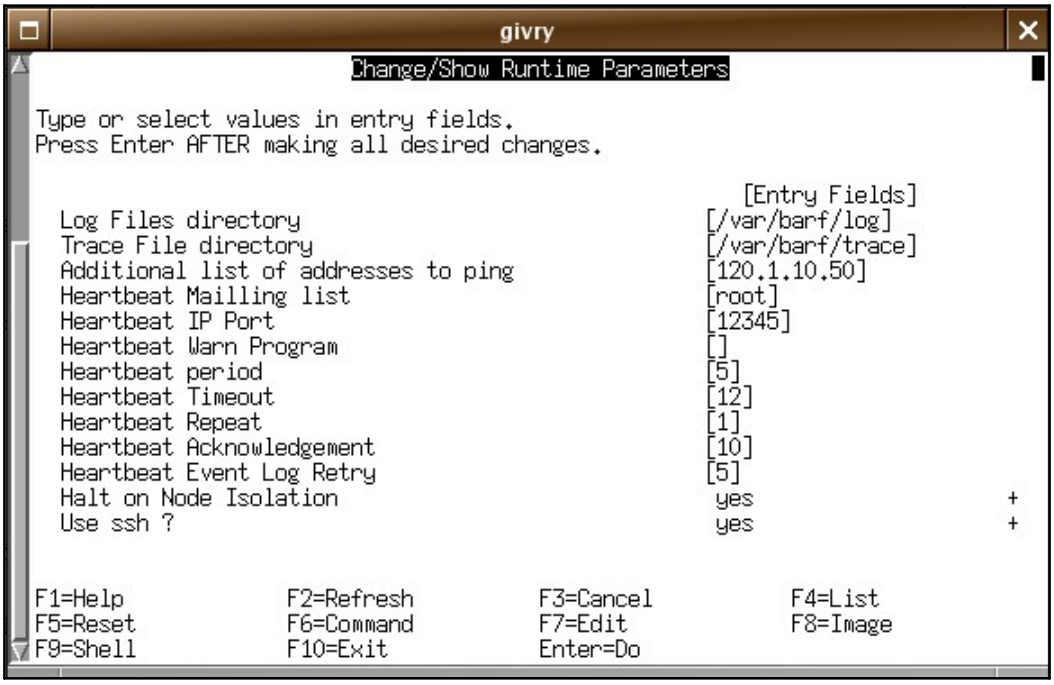


Figure 9. Change/Show Runtime Parameters screen

2. Modify the **Log Files directory** or the **Trace File directory** parameter to indicate where you want to redirect the log files or the trace files.
3. Press Enter to return to the **Configuration Definition** screen and to propagate the new configuration.

Note Existing log files will not be moved to the new location.

6.6. Configuring Error Notification

It is stronlgly recommended to implement your own notification methods for all errors that affect the disk subsystems.

6.6.1. Add an errnotify Method

To add a new **errnotify** method, take the following steps:

Type **smit barf** and select the following options: **Error Notification> Add a Notify Method** or use the **smit barf_add_notifymeth.dialog** fast path . When you press Enter, SMIT displays the following screen:

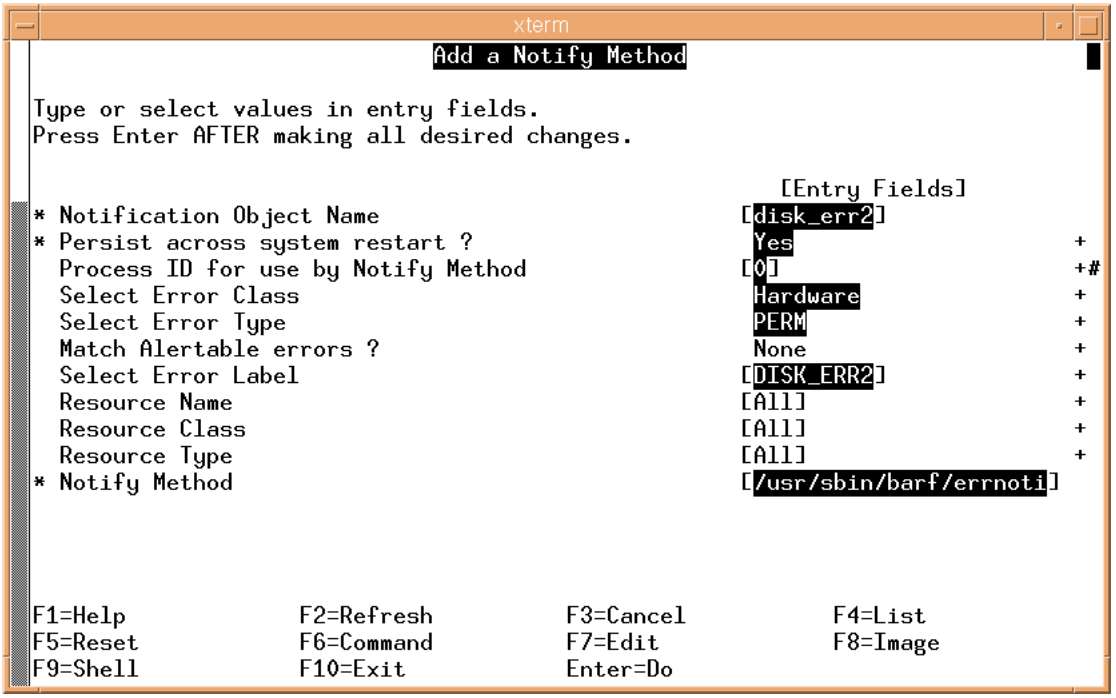


Figure 10. Add a Notify method screen

Notification Object Name
: A name assigned by the creator of the Error Notification object to uniquely identify the object. The creator will use this unique name when the creator wants to modify or remove the object.

Persist across system restart ?
: Choose 'No' to ensure obsolete error notification objects are deleted when the system is restarted. Choose 'Yes' to allow the object to persist after a system is restarted.

Process ID for use by Notify Method
: Select a Process ID which can be used by the Notify Method, or leave it blank. Objects which have a PID specified should choose 'No' for 'Persist across system restart?'.

Select Error Class
: Choose the appropriate error class. If an error occurs which consists of a matching class, the selected notify method will be run. Valid classes are 'All', 'Hardware', 'Software', and 'Errlogger'. Choose 'None' to ignore this entry.

Select Error Type
: Choose the appropriate error type. If an error occurs which consists of a matching type, the selected notify method will be run. Valid types are 'All', 'PENDing', 'Permanent ', PERFormance, 'TEMPorary', and 'UNKNown'. Choose 'None' to ignore this entry.

Match Alertable errors ?
: Choose 'Yes' to match alertable errors. Choose 'No' to match non-alertable errors. Choose 'All' to match both. Choose 'None' to ignore this entry.

Select Error Label
: Select an error label from the **/usr/include/sys/errids.h** file. If this particular error occurs, the selected Notify Method will be run.

Resource Name
: The name of the failing resource. For the hardware error class, a resource name is the device name. For the software error class, a resource name is the name of a failing executable.

Resource Class
: The class of the failing resource. For the hardware error class, the resource class is the device class. The resource error class is not applicable for the software error class.

Resource Type
: The type of the failing resource. For the hardware error class, a resource type is the device type a resource is known by in the devices object.

Notify Method
: Enter the full-path name of an executable file to be run whenever an error is logged which matches any of the defined criteria.

6.6.2. Change / Show an errnotify Method

To change or show a notify method, perform the following steps:

Type **smit barf** and select the following options: Error Notification> Change/Show a Notify Method or use the **smit barf_change_notifymeth_select** fast path. Choose the method that you want to modify and press Enter. SMIT displays the following screen:

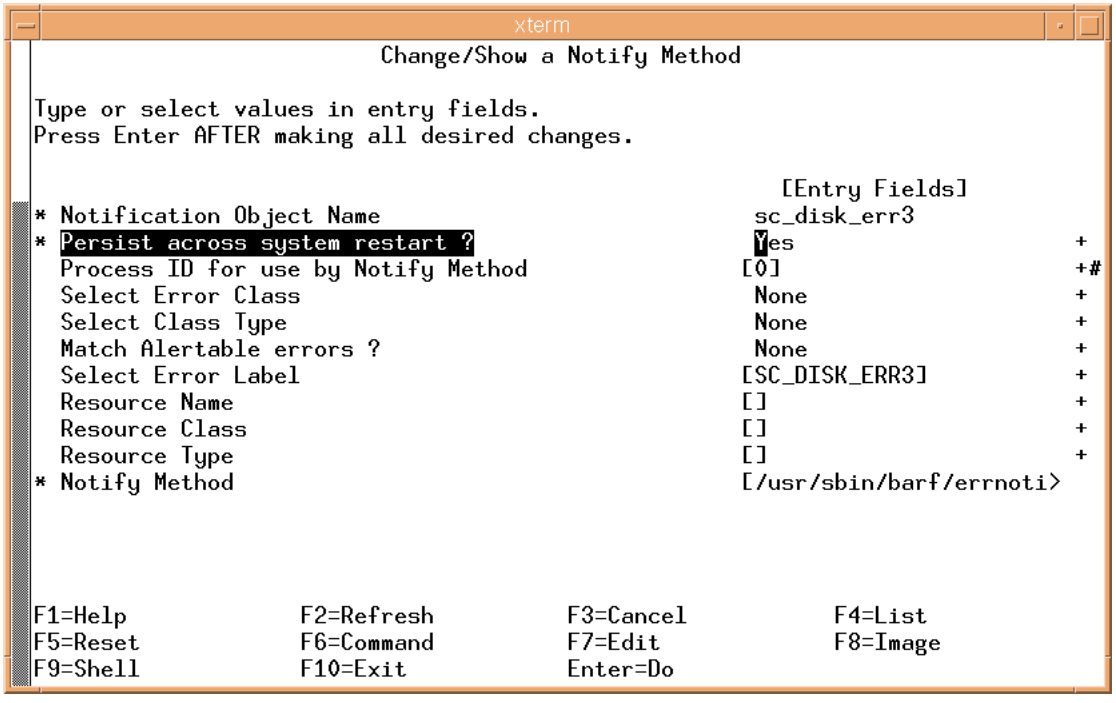


Figure 11. Change/Show a Notify method screen

Note This menu accesses all the errnotify methods that exist on the system, and not only those specific to *Application Roll-over Facility*.

Notification Object Name
: A name assigned by the creator of the Error Notification object to uniquely identify the object. The creator will use this unique name when the creator wants to modify or remove the object.

Persist across system restart ?
: Choose 'No' to ensure obsolete error notification objects are deleted when the system is restarted. Choose 'Yes' to allow the object to persist after a system is restarted.

Process ID for use by Notify Method
: Select a Process ID which can be used by the Notify Method, or leave it blank. Objects which have a PID specified should choose 'No' for 'Persist across system restart?'.

Select Error Class
: Choose the appropriate error class. If an error occurs which consists of a matching class, the selected notify method will be run. Valid classes are 'All', 'Hardware', 'Software', and 'Errlogger'. Choose 'None' to ignore this entry.

Select Error Type
: Choose the appropriate error type. If an error occurs which consists of a matching type, the selected notify method will be run. Valid types are 'All', 'PENDing', 'Permanent ', PERFormance, 'TEMPorary', and 'UNKNown'. Choose 'None' to ignore this entry.

Match Alertable errors ?
: Choose 'Yes' to match alertable errors. Choose 'No' to match non-alertable errors. Choose 'All' to match both. Choose 'None' to ignore this entry.

Select Error Label
: Select an error label from the **/usr/include/sys/errids.h** file. If this particular error occurs, the selected Notify Method will be run.

Resource Name
: The name of the failing resource. For the hardware error class, a resource name is the device name. For the software error class, a resource name is the name of a failing executable.

Resource Class
: The class of the failing resource. For the hardware error class, the resource class is the device class. The resource error class is not applicable for the software error class.

Resource Type
: The type of the failing resource. For the hardware error class, a resource type is the device type a resource is known by in the devices object.

Notify Method
: Enter the full-path name of an executable file to be run whenever an error is logged which matches any of the defined criteria.

6.6.3. Remove an errnotify Method

You can remove an errnotify method from an active *Application Roll-over Facility* system dynamically.

Take the following steps to remove an errnotify method:

Type **smit barf** and select the following options: Error Notification> Delete a Notify Method or use the **smit barf_EN_menu** fast path. When you press Enter, SMIT displays a list of errnotify methods. Choose one, and press enter.

6.7. Configuring the Node Monitoring

Configuring the node monitoring means to accept or modify the default values of the runtime parameters.

To change the runtime parameters for a node, do the following steps:

1. Type **smit barf** and select the following options: **Configuration Definition> Change/Show Run Time Parameters** or use the **smit barf_ch_param** fast path. SMIT displays the following screen :

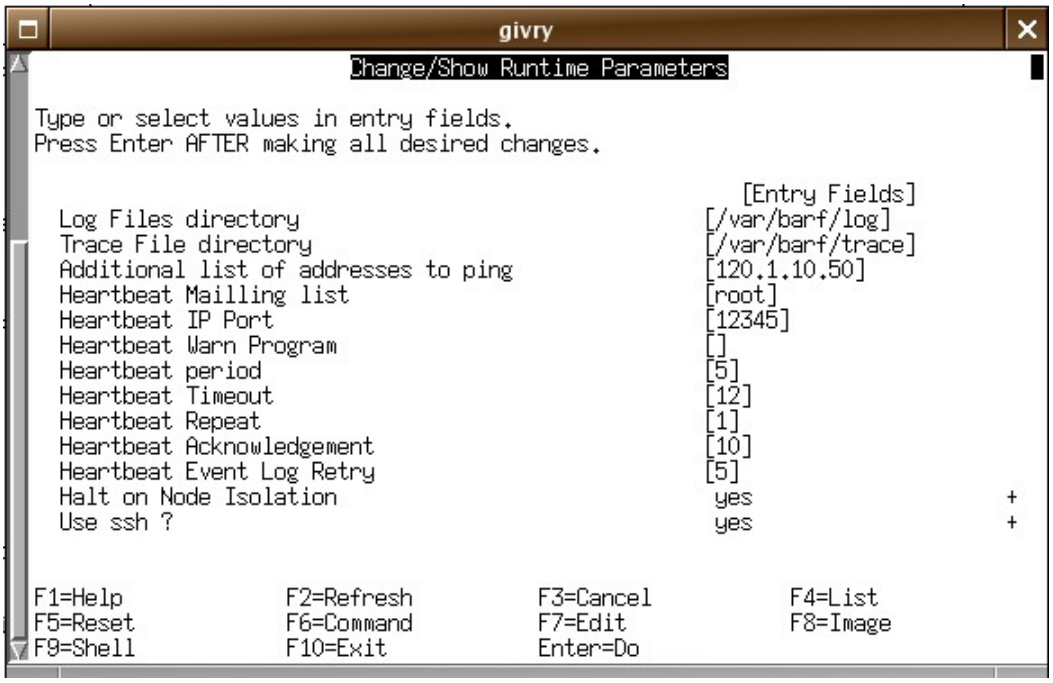


Figure 12. Change/Show Runtime Parameters screen

Log Files Directory Name of directory where the log files are registered. Default Value: **/var/barf/log**.

Trace File directory Name of the directory where the trace files are registered. Default Value: **/var/barf/trace**.

Additional list of addresses to ping

This address (or list of addresses) allow the system to determine if a node failure or a lan failure occurs. If only one physical network (a lan) exists in the configuration, it is not possible to determine if a node failure or a lan failure occurs.

If this address (or list of addresses) located on the local area network is reachable, a node failure is detected, otherwise if no address is reachable a lan failure is detected.

So, in a configuration with only one physical network (a lan), it is recommended to configure an additional address to ping and to set the **Halt on Node Isolation** parameter to **yes**.

Heartbeat Mailing List

List of Email Addresses where a message is sent when a node is unreachable. Each Email address is separated by a space. Default Value: **root**.

Heartbeat IP Port IP port of the Node daemon (numbers composed of digits). Default Value: 12345.

Heartbeat Warn Program
: Full Path Name of the user-defined program to be run by the daemon when a node is unreachable (full pathname composed of letters, digits and slash).
Refer to Appendix D *Customizing the Warning Program File* for an example.

Heartbeat Period
: Time (in seconds) between the sending of two Heartbeats of the Node Daemon (time composed of digits). The period of sending a Heartbeat will be the maximum of the Heartbeat Timeout divided by two (Heartbeat Period < Heartbeat Timeout/2). Default Value: 5.

Heartbeat TimeOut
: Time (in seconds) to wait for a Heartbeat from the Monitoring Node Daemon (time composed of digits). As during the wait of the acknowledgement, no Heartbeat is sent, the Heartbeat Acknowledgement must be less than the Heartbeat Timeout (Heartbeat Acknowledgement < Heartbeat Timeout). Default Value: 12.

Heartbeat Repeat
: Acceptable number of times that a Remote Node is being unreachable (digit). The elapsed time is equal to the Node Heartbeat Period + the Node Heartbeat Timeout + the Node Heartbeat Acknowledgement (described hereafter). Minimum and default Value: 1.

Heartbeat Acknowledgement
: Time (in seconds) to wait for an acknowledgement from the Monitoring Node Daemon (time composed of digits). Default Value: 10.

Heartbeat Event Log Retry
: Maximal number of time to retry to:
- record an event in the **clsmd.log** file.
- send a mail to the administrator in case of a node failure.
- send a message to the system console in case of a node failure.
- execute the customized Heartbeat Warn Program script in case of a node failure.
If the value is 0, the event is always retried. Default Value: 5.

Halt on Node Isolation
: Node halt setting. Type: '**no**' for no halt or '**yes**' for halt. Default value: yes.
When the **Halt on Node Isolation** parameter is set to yes, the takeover node is prevented from running an application which is still active on the isolated node.

Use ssh?
: Indicates which protocol is to be used for remote commands. Type '**yes**' to mean ssh is used, type '**no**' to mean rsh is used. Default value: yes.

2. Press Enter to validate the new values into the *Application Roll-over Facility* ODM database.
3. To propagate these new values to the other nodes, you must synchronize the applications. See *Synchronizing the Application Roll-over Facility Configuration*, on page 6-21 for more information.

6.8. Configuring Virtual Disks for an Application Roll-over Facility Cluster

This function is available on Escala servers that support the virtualization feature.

What follows explains the way to use virtual disks (from Virtual IO Server) as shared Volume Groups in an ARF cluster.

It is assumed that one Virtual SCSI Server adapter has been created in the Virtual IO Server (VIOS) for each Virtual IO Client partition (ARF node) and that it is associated to the Virtual SCSI Client adapter created in each Virtual IO Client partition.

6.8.1. Prerequisites if rsh is used for remote commands

1. Prerequisites for each VIOS:
   - Modify the VIOS file **/etc/inetd.conf** to allow client partitions to execute rsh commands on the VIOS:
     . login to VIOS as **padmin** user.
     . Get the root privileges with the **oem_setup_env** command.
     . Edit **/etc/inetd.conf** and un-comment the line for **rshd** daemon:

     ```
     shell  stream tcp6   nowait root   /usr/sbin/rshd       rshd
     ```
   - For each VIOS, create **/.rhosts** file to authorize the root user of all ARF nodes (client partitions) to execute rsh commands.
   - If not yet installed, install **Bulltools** fileset from *Bull Enhancements* CD-ROM in each VIOS. This fileset provides the **bsan** command, used to reset disks reservation.
2. Prerequisite for the client partitions:
   - Modify the **/.rhosts** file to authorize the root user of the VIOS they use to perform remote commands.

6.8.2. Prerequisites if ssh is used for secure remote commands

1. Install **openssh** fileset and **openssl** RPM package on each Virtual I/O Server
2. Authorize ssh access to the **root** user, for all the nodes on each Virtual I/O Server. Refer to *Generating the SSH keys*, on page 3-1.

6.8.3. Configuring Virtual Disks

1. On the first VIOS of the first server, create Virtual Target Devices to map the physical disks to the client partitions.
   - Login as **padmin** user.
   - Assign a PVID to each physical disk:

     ```
     chdev -dev hdisk(power)<x>  -attr pv=yes  -perm
     ```

   In case of redundant VIOS (two VIOS per machine) and MPIO used in client partitions:
   - List disk reservation attribute:

     ```
     lsdev -dev hdisk(power)<x> -attr
     ```

- Look for attribute such as **reserve_lock** (for PowerPath) or **reserve_policy**.
- Change reservation policy on the disk, to be able to map it from both VIOSs:

  ```
  chdev -dev hdiskpower<x> -attr reserve_lock=no
  or
  chdev -dev hdisk<x> -attr reserve_policy=no_reserve
  ```

- Map physical disks to client partitions (ARF nodes):

  ```
  mkvdev -vdev hdisk(power)<x> -vadapter vhost<x> -dev <vhdiskx>
  ```

  where:
  `vhost<x>` is the virtual SCSI Server adapter associated to the client partition.

  `<vhdiskx>` is an example of name given to the Virtual Target Device associated to the physcal disk `hdisk(power)<x>`.

  To know about hdisk(power)<x> and vhost<x>, enter:

  ```
  lsmap -all
  ```

Or get the root privileges with the **oem_setup_env** command and enter:

```
lsdev -Cc disk
lscfg -vp | grep vhost
```

To view current disk mapping, login as **padmin** user and enter:

```
lsmap -all
```

Repeat this step (**chdev** and **mkvdev**) for the second (redundant) VIOS, if it exists.

2. Once all the Virtual Target Devices have been created on the VIOS of the first Escala, run **cfgmgr** or reboot the client partitions of this VIOS to configure the virtual disks to be used as ARF shared Volume Groups.

   To list the virtual disks and adapters, enter:

   ```
   lsdev -Cc disk
   lsdev -Cc adapter
   ```

3. On the client partition(s) of this VIOS:
   - Create ARF shared Volume Groups:

     ```
     smit mkvg
     ```

     `smit chvg` (to change VG characteristics to not activate volume group automatically at system restart).
   - Create Logical Volumes for JFS2 (or JFS) log and for JFS2 (or JFS) File Systems.
   - Format log Logical Volume:

     ```
     logform /dev/logxxx
     ```

   At this time only one VIOS and its client partition(s) have been configured, in one server.

4. To be able to configure VIOS and its client partition(s) on the second server, ARF shared Volume Groups must be de-activated by client partitions:

   ```
   varyoffvg <VG name>
   ```

   In the VIOS, Virtual Target Devices must be in **defined** state to remove disk reservation. As **padmin** user, for each Virtual Target Device (`lsmap -all`), run:

   ```
   rmdev -dev <vhdiskx>  -ucfg
   ```

5. On the VIOS of the second server repeat step 1. to create Virtual Target Devices to map the physical disks containing the shared Volume Groups.

6. In client partitions (ARF nodes), run `cfgmgr` (or `shutdown -Fr`) to access the disks.

   Then import ARF shared Volume Groups:

   `lspv` (to identify the correct hdisk)

   `smit importvg`

   `smit chvg` (to change VG characteristics not to be automatically activated at boot time).

7. ARF shared Volume Groups must be de-activated by client partition:

   `varyoffvg <VG name>`

   In all VIOS, Virtual Target Devices must be in **defined** state to remove disk reservation. As **padmin** user, for each Virtual Target Device (`lsmap -all`), run:

   `rmdev -dev <vhdiskx> -ucfg`

8. Now that the shared volume groups are known by client partitions, ARF can be configured. When you configure an ARF node that is a VIO Client partition, specify the IP address of the VIOS it uses.

6.9. Displaying the Disk Mapping in an Application Roll-over Facility Configuration

What follows explains how to display the disk mapping between virtual disks on ARF nodes and physical disks on Virtual IO Servers (VIOS).

Prerequisites

- The nodes and VIOS must be already defined in the *Application Roll-over Facility* configuration.
- The nodes definition must be already synchronized.

1. Type **smit barf** and select **Configuration Definition > Display Disk Mapping (VIOS Partitions)**, or use the **smit barfviosmapping** fast path. The following screen appears:

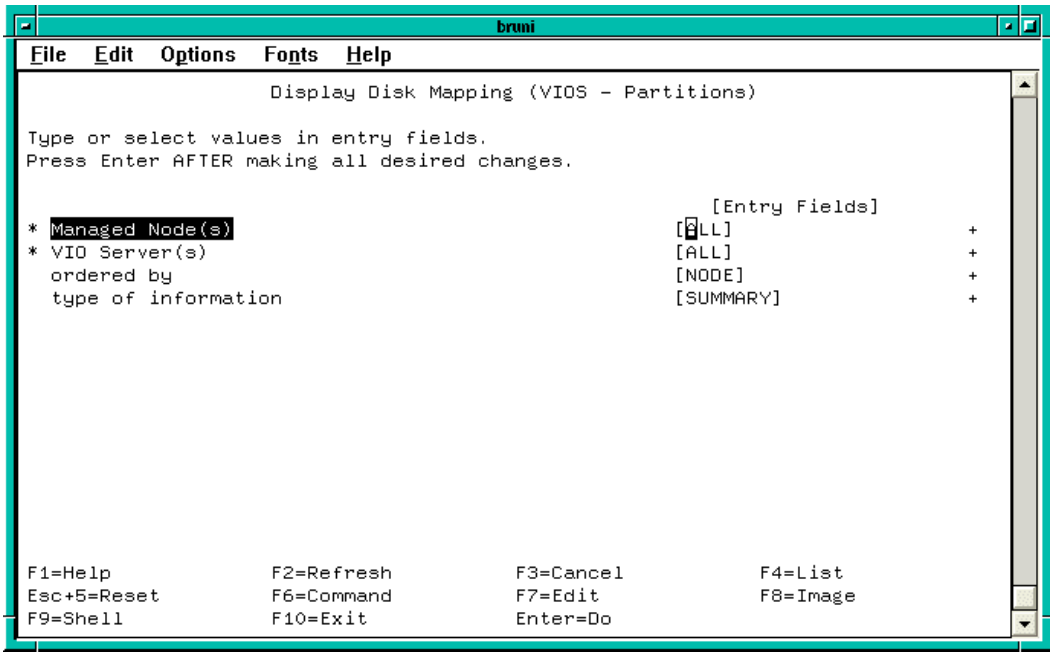


Figure 13. Display Disk Mapping (VIOS - Partitions) screen

Managed Nodes(s)	By default, the disk mapping is displayed for all ARF nodes. You can select specific nodes from the list; in this case, the disk mapping will be displayed only for these nodes.
VIO Server(s)	By default, the disk mapping is displayed for all VIO Servers defined in the configuration. You can select specific VIOS from the list; in this case, the disk mapping will be displayed only for these VIOS.
ordered by	By default, the output is ordered by NODE. You can select another order: by VIO.
type of information	By default, a SUMMARY output is displayed for disk mapping. You can select another type of information (DETAILED or ALL) from the list. The SUMMARY option outputs the following fields: node name, virtual disk name on node, PVID on node, VIOS name, physical disk name on VIOS, PVID on VIOS. The DETAILED option outputs the following fields : node name, virtual disk name on node, virtual parent on node, VIOS name, physical disk on VIOS, physical adapter on VIOS, virtual disk on VIOS, virtual adapter on VIOS. The ALL option outputs the same fields as the DETAILED option for all the physical disks on VIOS.

Note The logical volume - virtual disk mapping is not displayed. Only the mapping between virtual disks and physical disks is taken into account.

2. Press Enter. The disk mapping between virtual disks on nodes and physical disks on VIOS is displayed.
3. Press F10 to exit SMIT.

Examples of output:

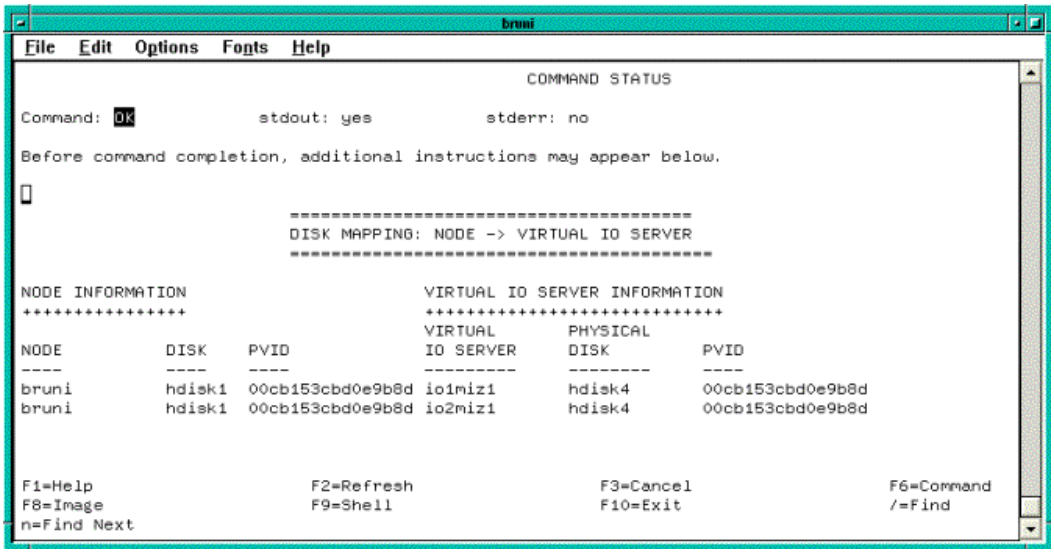


Figure 14. Disk Mapping: summary

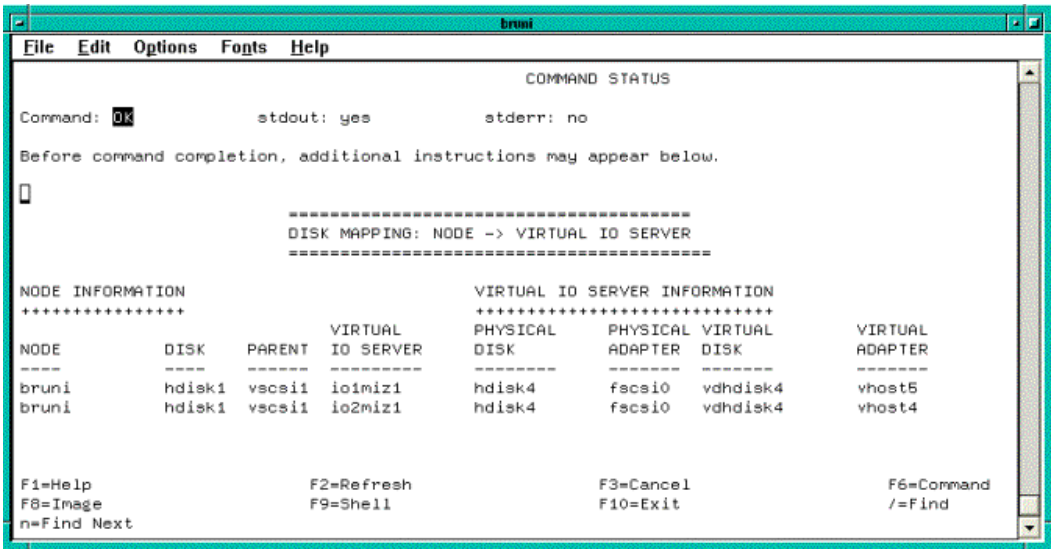


Figure 15. Disk Mapping: details

6.10. Verifying the Application Roll-over Facility Environment

This section describes how to verify the *Application Roll-over Facility* environment, including the node configurations. This process ensures that all nodes agree on topology and assignment of resources.

After defining the node environment, run the verification procedure on one node to check that all nodes agree on the assignment of *Application Roll-over Facility* resources.

To verify the node configuration:

1. Enter **smit barf** command and select **Verify Configuration**.
2. Press Enter.
   SMIT runs the **barf_chk_conf** utility. The output from the verification is displayed in the SMIT Command Status window. If you receive error messages, make the necessary changes and run the verification procedure again.

To synchronize all nodes, use the **Propagate Configuration** option on the Configuration definition SMIT screen. See *Synchronizing the Application Roll-over Facility Configuration*, on page 6-21 for more information.

6.11. Synchronizing the Application Roll-over Facility Configuration

This section describes how to propagate the configuration on the different nodes. This process ensures that all nodes agree on topology and assignment of resources.

You can ask *Application Roll-over Facility* to try to import each volume group of each resource, on each node of the configuration, even if the volume group is already varyon on one node.

1. Type **smit barf** and select the following options: **Configuration Definition > Propagate Configuration** or use the **smit barf_sync_conf** fast path. When you press Enter, SMIT displays the following screen.

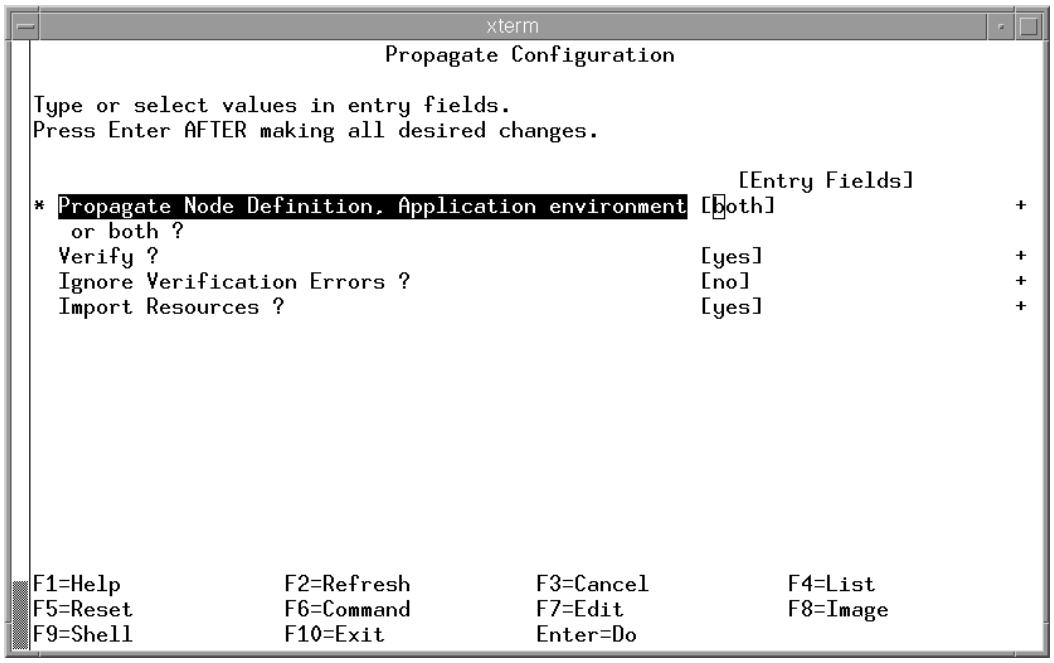


Figure 16. Propagate Configuration screen (Node and Application)

Propagate Node Definition, Application environment or both ?
: Choose '**both**' to do the synchronization of the node definition and Application environment.

Verify ?
: By default, this field is set to '**yes**' and the topology verification program is run. To save time in the synchronization process, you can toggle this entry field to '**no**'. By doing so verification will be skipped.

Ignore Verification Errors
: By choosing '**yes**', the result of the verification is ignored and the configuration is synchronized even if verification fails. By choosing '**no**', the synchronization process terminates if errors are found; view the error messages in the system error log to determine the configuration problem.

Import Resources
: If you set this field to '**yes**', the synchronization will try to import all the volume groups defined on each node with the same major number than the local node, if this number is free, except if the volume group is already imported.

2. Press Enter. The *Application Roll-over Facility* definition is copied to the other nodes.
3. Press F10 to exit SMIT.

6.12. Verifying the Application Roll-over Facility License Key(s)

This section explains how to verify the Application Roll-over Facility license key(s).

This verification must be done before Starting Application Roll-over Facility Services or Activating Node Monitoring Services to avoid failed execution of these services if the license key is not valid.

1. Type **smit barf** and select **License Key** or use the **smit barfkey** fast path. The following screen appears:

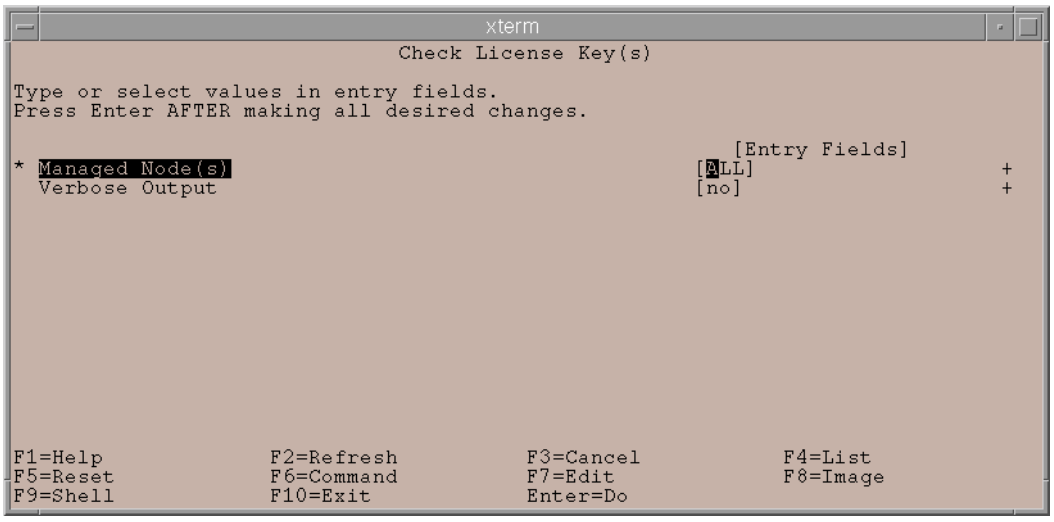


Figure 17. Check License Key(s)

Managed Nodes	By default, the license key is verified on all the managed nodes. You can select specific nodes from the list; in this case, the license key will be verified only on these nodes.
Verbose Output	By choosing '**yes**', information about license is displayed (expiration date, host, version, issuer, comment, ...).

2. Press Enter. The license key(s) are verified on the specified managed nodes.
3. Press F10 to exit SMIT.

6.13. Configuring Application Monitoring

6.13.1. Application Monitoring Daemon

The *Application Monitoring* daemon runs the following tasks:

- It updates its own configuration and knowledge by reading configurations and Operating System Data in ODM data base, to detect for example on which node the daemon is running or to know the status of active applications (the application monitorings are then added or removed from the scheduler),
- It schedules the configuration updates, as soon as possible, according to the filling of the “FIFO",
- It “monitors" an application (this means it runs the **status** command and proceeds suitable tasks according to the returned status).

The first application status retrieving is scheduled to be executed at a given time after which the application is detected to be started. The length of this period of time is defined in the **Gracious Time** field. The next status is scheduled after a period of time defined in the **Interval** field. When a restart command is performed, the next status is scheduled as if the application has just started.

6.13.2. The daemon commands

The commands delivered with the *Application Monitoring* facility are "common" commands. It is also possible to write specific commands for showing daemon status or restart daemon for example. In this case, it may be helpful to use the following environment variables:

BARF_MON_NAME	monitoring name.
BARF_MON_APPNAME	name of the monitored application.
BARF_MON_STEP	running step in the monitoring process. The possible values are: STATUS, RESTART, ABANDON, FAILURE, ROLLOVER...
BARF_MON_CUR_NODE	current node.
BARF_MON_FAILURE	number of detected failures.
BARF_MON_TOLERANCE	number of consecutive errors that are ignored.

Scripts delivered with the monitored application may also be used here.

The common command names are indicated in the following. The usual PATH for commands is: **/usr/sbin/barf/appmon/bin**.

barf_am_abandon	stops the application. This is the default value for the **Abandon Command** parameter.
barf_am_restart	restarts the application. This is the default value for the **Restart Command** parameter.
barf_am_rollover	stops the application on the local node and tries to restart this application on the nodes dfined in the take-over list of the application. This is the default value for the **Rollover Command** parameter.
barf_am_status	retrieves the application status. This is the default value for the **Status Command** parameter. It may be used with the following parameters: - -h : prints help - -p <pid> : checks if the <pid> process is present.

- -n <string> : checks if there is at least one process running with the given string <string> as program name.
- -c <string> : checks if there is at least one running process whose the associated command is <string>.

6.13.3. Configuring Daemon Parameters

To configure *Application Monitoring* daemon parameters:

1. Type **smit barf** and select the following options: **Configuration Definition> Change/Show Application Monitoring Daemon Parameters** or use the **smit barf_ch_appmondmn** fast path. The following menu appears:

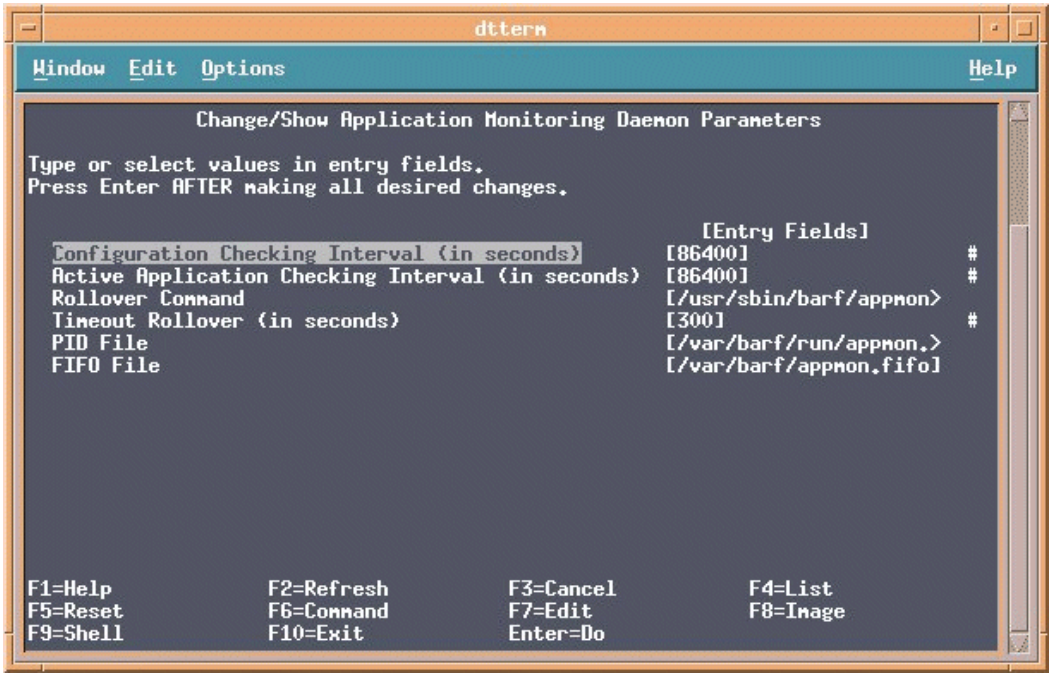


Figure 18. Change/Show Application Monitoring Daemon Parameters

Configuration Checking Interval (in seconds)
: Maximum time spent between two checks of the configuration. The daemon accesses the ODM data base to read configurations related to the application monitoring, the daemon and the topology. It also collects the IP addresses set on the network interfaces. The default value is 86400.

Active Application Checking Interval (in seconds)
: Maximum time spent between two checks of started and stopped applications. The daemon accesses to the ODM data base to read the active applications, thus deducing the newly stopped and started applications. The default value is: 86400

Rollover Command
: Command to be executed when a rollover is expected. The default value is: barf_am_rollover.

Timeout Rollover (in seconds)
: Maximum time for running the rollover command. If the running time exceeds this value, the process is killed and an erreor generated. The default value is: 300.

PID File
: File path containing the daemon's process ID. This ID is used by the scripts to check if the daemon is running or not. The default value is: /var/barf/run/appmon.pid.

FIFO File — Path to the FIFO used to wake up the daemon. The default value is: /var/barf/run/appmon.fifo.

2. Press Enter to Change/Show the *Application Monitoring* Daemon Parameters.
3. Press F10 to exit SMIT.

6.13.4. Adding, Changing, Removing an Application Monitoring

An *Application Monitoring* may be added, changed or removed.

6.13.4.1. Creating an Application Monitoring

To create an *Application Monitoring:*

1. Type **smit barf** and select the following options: **Configuration Definition> Application environment> Configure Application Monitoring** or use the **smit barf_cfg_appmon** fast path. The following menu appears:

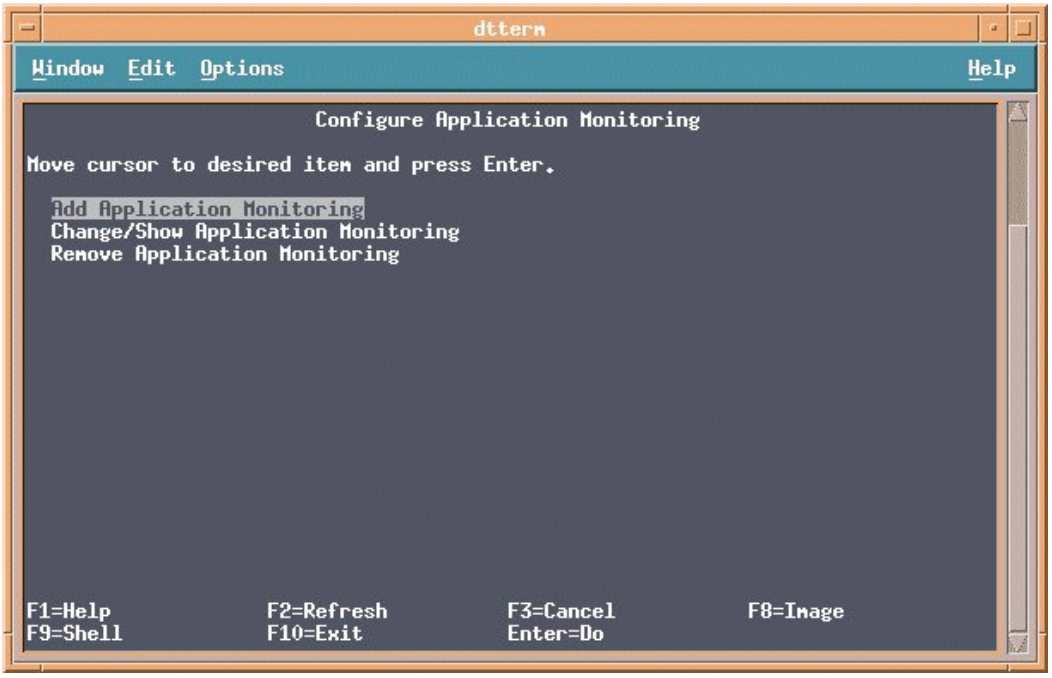


Figure 19. Configure Application Monitoring (create)

2. In the **Configure Application Monitoring** Menu, select **Add Application Monitoring**.
3. Select the name of the application to monitor. The monitoring does not start while the specified application is not detected as '**started**' on the current node.

The following menu appears:

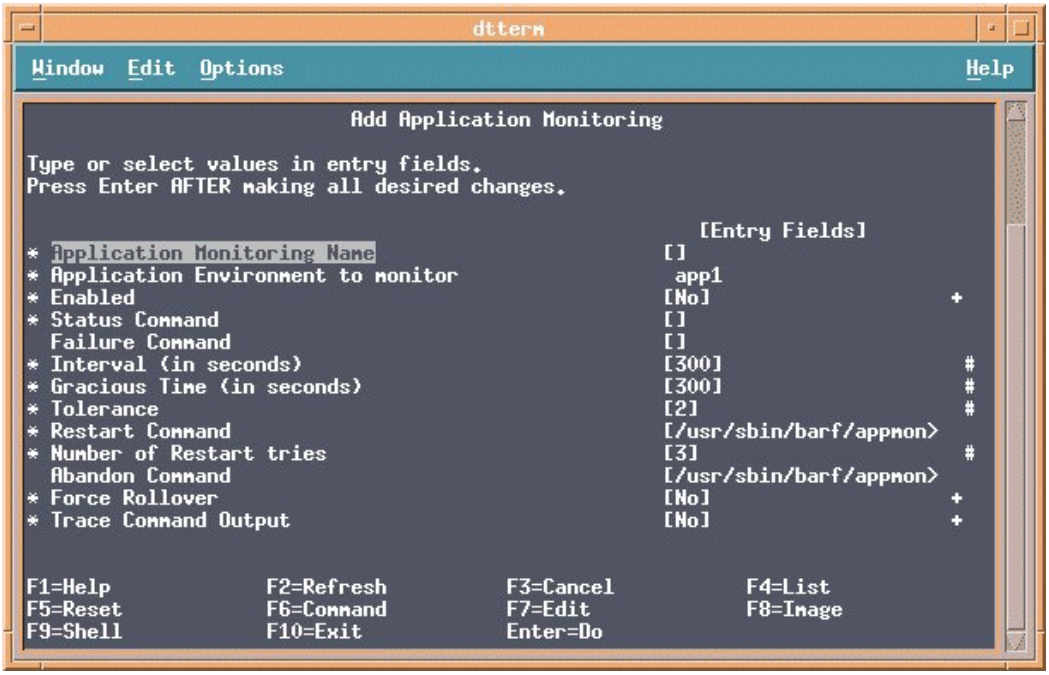


Figure 20. Add Application Monitoring

Application Monitoring Name
: Name of the Application Monitoring.

Application Environment to monitor
: Name of the selected application to monitor.

Enabled
: 'Yes' to start the monitoring, 'No' to stop it. The default value is: No.

Status Command
: Command to be executed to retrieve the Application status. The error counter is updated according to the return value. If the return value is 'ok' (return code=0), the error counter is set to '0'. If the return value is 'false' (return code non equal to 0), or if the command was not terminated, or if the command could not be run, the error counter is incremented.

Failure Command
: Command to be executed when the error counter is incremented.

Interval (in seconds)
: Elapsed time between two starts of status retrieving. This also defines the maximum time of execution for Status, Failure, Restart and Abandon Commands. The default value is: 300.

Gracious Time (in seconds)
: Elapsed time between the application start and the first execution of the status command. The default value is: 300.

Tolerance
: Number of consecutive errors to be ignored. The default value is: 2.

Restart Command
: Command to be executed when the number of errors is greater than the tolerance number.

Number of Restart tries
: Maximum number of restart tries allowed when the restart command fails. The default value is: 3.

Abandon Command
: Command to be executed when the number of restarts reaches the maximum allowed.

Force Rollover
: 'Yes' to force a rollover, 'No' otherwise. The default value is: No.

Trace Command Output 'Yes' to write the output of command execution in the Trace Files, 'No' otherwise. The default value is: No.

4. Propagate the new configuration: return to the **Configuration Definition** menu and select the following option: **Propagate Configuration**.
5. Verify the new configuration: return to the main menu and select the following option: **Verify Configuration**.
6. Press F10 to exit SMIT.

WARNING

If rollover is selected, it means that the application will be run on a backup node if a restart fails. During this switch, some ARF commands may be called. If an "abandon" command is defined on this monitoring AND if this command is run in the background mode AND if it calls ARF commands, execution problems may appear due to possible conflicts between the abandon command and the rollover command. So it is strongly recommended to avoid the setting of the "Force Rollover" parameter to 'yes' and the "Abandon Command" parameter to 'bar_am_abandon &', for example.

6.13.4.2. Modifying an Application Monitoring

To modify an *Application Monitoring:*

1. Type **smit barf** and select the following options: **Configuration Definition> Application environment> Configure Application Monitoring** or use the **smit barf_cfg_appmon** fast path. The following menu appears:

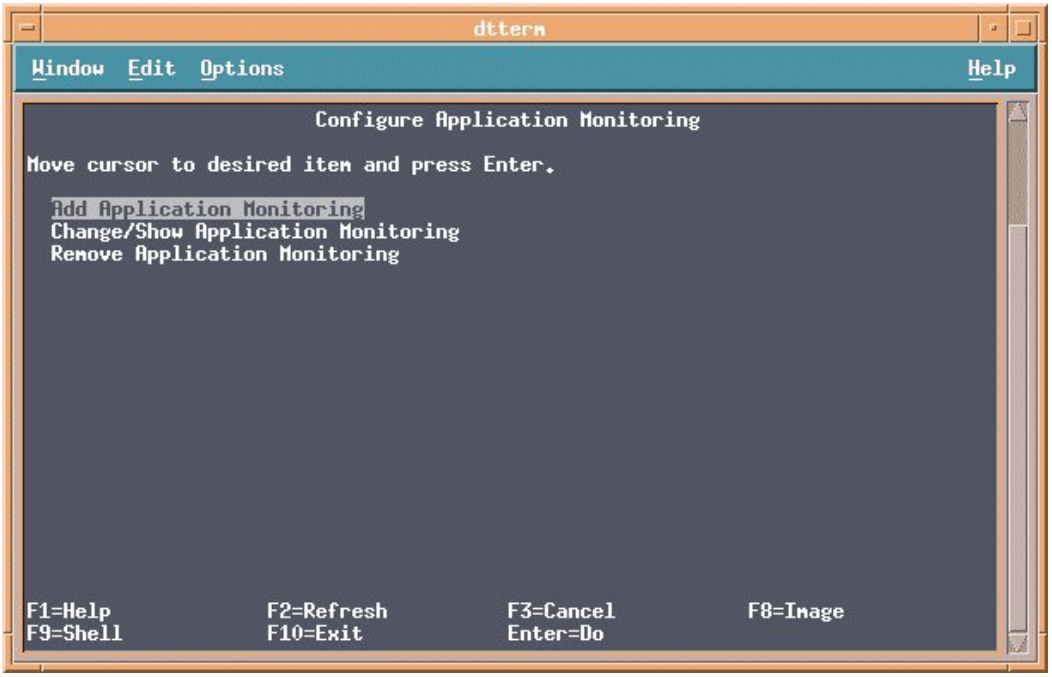


Figure 21. Configure Application Monitoring (modify)

2. In the **Configure Application Monitoring** Menu, select **Change/Show Application Monitoring**.
3. Select the **Application Monitoring** to modify. The following menu appears:

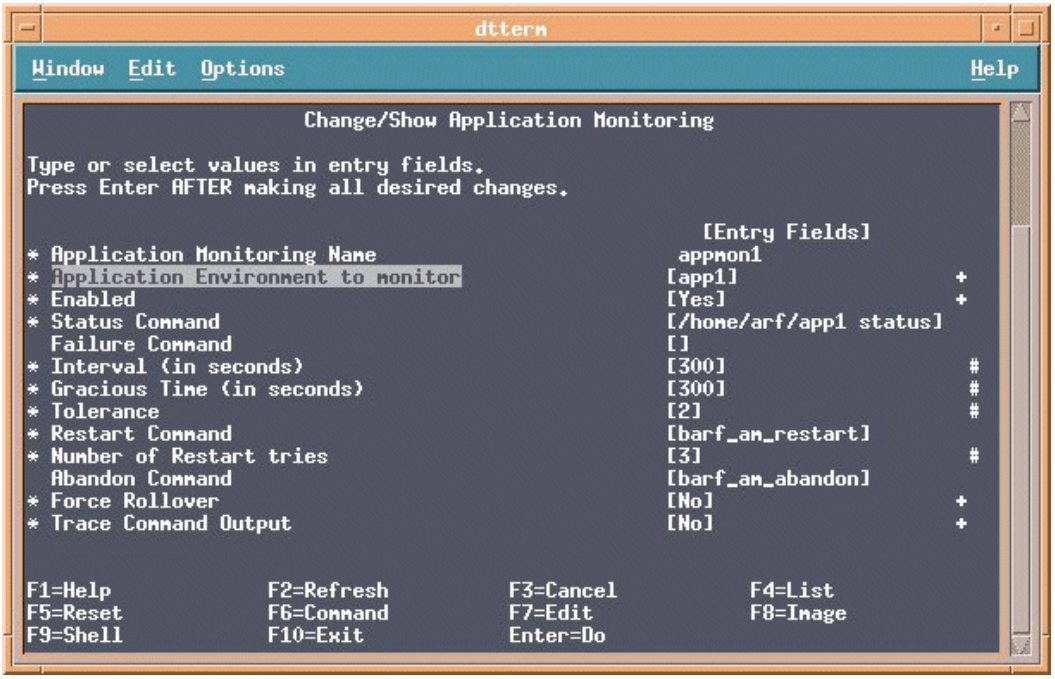


Figure 22. Change/Show Application Monitoring

All the parameters values are those defined at the monitoring creation and may now be modified.

Application Monitoring Name	Name of the Application Monitoring.
Application Environment to monitor	Name of the selected application to monitor.
Enabled	'Yes' to start the monitoring, 'No' to stop it.
Status Command	Command to be executed to retrieve the Application status. The error counter is updated according to the return value. If the return value is 'ok' (return code=0), the error counter is set to 0. If the return value is 'false' (return code non equal to 0), or if the command was not terminated, or if the command could not be run, the error counter is incremented.
Failure Command	Command to be executed when the error counter is incremented.
Interval (in seconds)	Elapsed time between two starts of status retrieving. This also defines the maximum time of execution for Status, Failure, Restart and Abandon Commands.
Gracious Time (in seconds)	Elapsed time between the application start and the first execution of the status command.
Tolerance	Number of consecutive errors to be ignored.
Restart Command	Command to be executed when the number of errors is greater than the tolerance number.
Number of Restart tries	Maximum number of restart tries allowed when the restart command fails.
Abandon Command	Command to be executed when the number of restarts reaches the maximum allowed.
Force Rollover	'Yes' to force a rollover, 'No' otherwise.
Trace Command Output	'Yes' to write the output of command execution in the Trace Files, 'No' otherwise.

4. Propagate the new configuration: return to the **Configuration Definition** menu and select the following option: **Propagate Configuration**.
5. Verify the new configuration: return to the main menu and select the following option: **Verify Configuration**.
6. Press F10 to exit SMIT.

6.13.4.3. Removing an Application Monitoring

To remove an *Application Monitoring:*

1. Type **smit barf** and select the following options: **Configuration Definition> Application environment> Configure Application Monitoring** or use the **smit barf_cfg_appmon** fast path. The following menu appears:

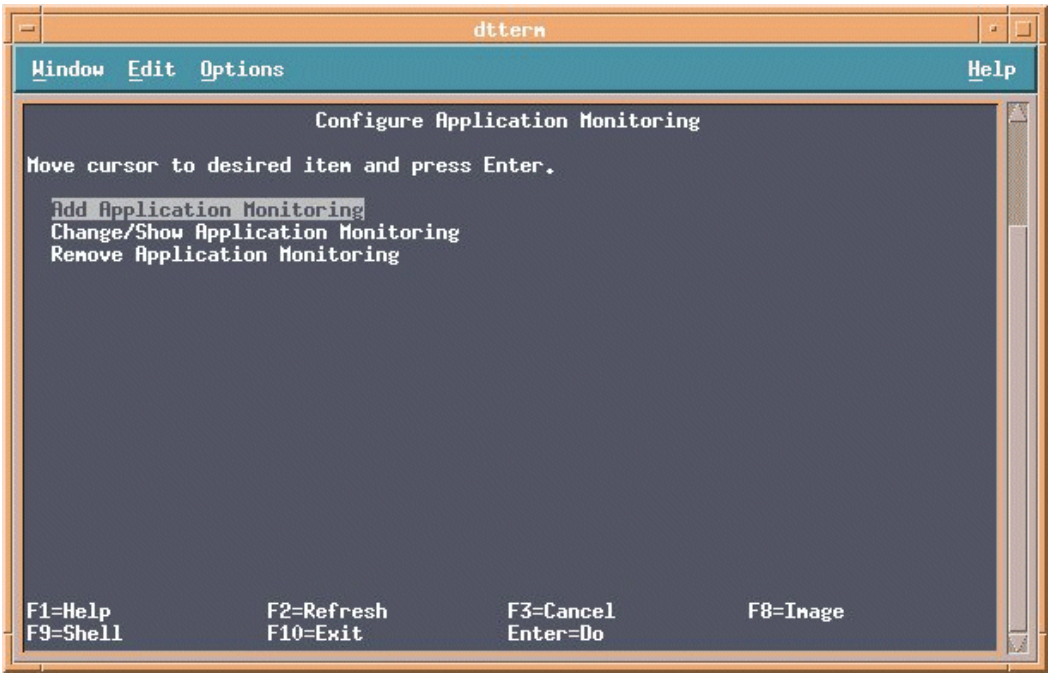


Figure 23. Configure Application Monitoring (remove)

2. In the **Configure Application Monitoring** Menu, select **Remove Application Monitoring**.

3. Select the **Application Monitoring** to remove. The following menu appears:

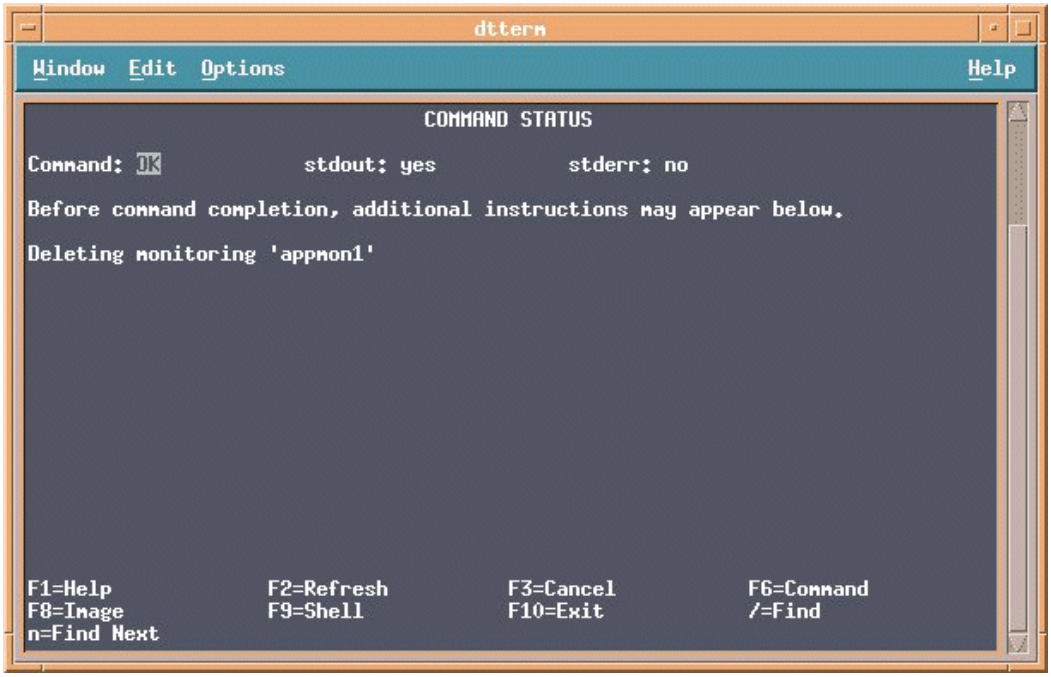


Figure 24. Remove Application Monitoring

4. Propagate the new configuration: return to the **Configuration Definition** menu and select the following option: **Propagate Configuration**.
5. Verify the new configuration: return to the main menu and select the following option: **Verify Configuration**.
6. Press F10 to exit SMIT.

6.13.5. Example of configuring an Application Monitoring

This example shows how to write and configure in the Application Monitoring a program that checks the application status.

Prerequisites:

- An application is already created in ARF (named **appweb** in this example).
- When this application is running, a daemon runs on the system. Its command is **/usr/HTTPServer/bin/httpd**.

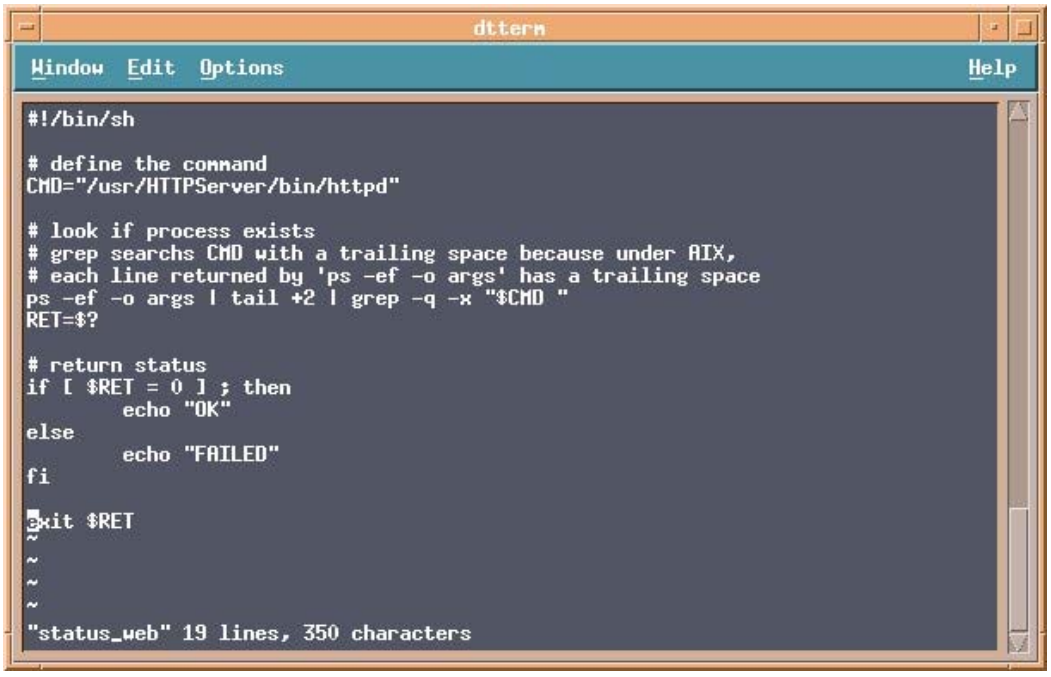
6.13.5.1. Write the application status supervisor program

The program verifies that the daemon is still running: it checks that there is at least one process having the **/usr/HTTPServer/bin/httpd** command.

The program returns 0 is the application status is good, otherwise it returns 1.

Assuming the name of this program is **status_web**:

1. Create the program (under **/tmp** for example), with the following contents:



Note In the above example, the command given in the CMD parameter is a binary file. If the application is a shell script (/tmp/start_app for example), the CMD parameter should be set to /bin/sh /tmp/start_app.

2. Give to the /tmp/status_web program the executable status:
   chmod +x /tmp/status_web
3. Copy this program and check it is still executable on each node.

6.13.5.2. Add an Application Monitoring

Modify the parameters to fit the needs, as shown in the following menu:

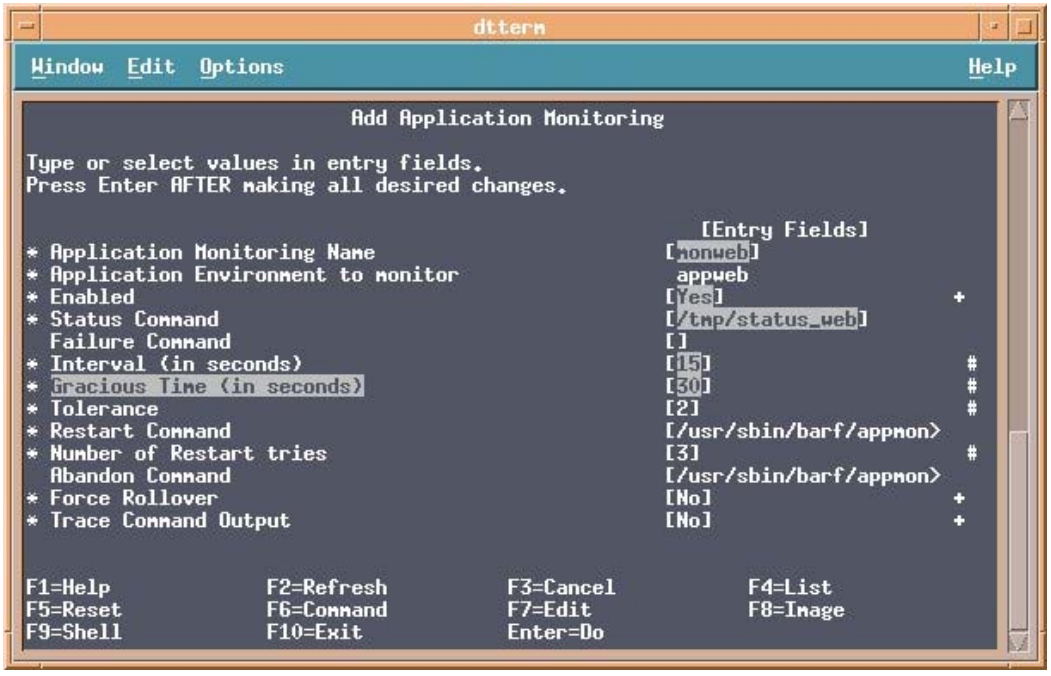


Figure 25. Add Application Monitoring

Application Monitoring Name	Arbitrary Name: **monweb**.
Application Environment to monitor	Name of the application to monitor (previouly selected).
Enabled	'**Yes**' to activate this application monitoring.
Status Command	Name of the command previouly created: **/tmp/status_web**.
Interval	Check every 15 seconds.
Gracious Time	The monitoring starts 30 seconds after that the application is detected as '**started**'.
Tolerance	2: Default Value. Ignores the two first consecutive errors and tries to restart at the third consecutive error.
Restart Command	Default Value: the generic command **/usr/sbin/barf/appmon/bin/barf_am_restart**.
Number of Restart tries	Default Value: three consecutive restarts before abandon.
Abandon Command	Default Value: the generic command **/usr/sbin/barf/appmon/bin/barf_am_abandon**.
Force Rollover	Default Value: No.
Trace Command Output	Default Value: No.

6.14. Configuring an Application to start in an Application WPAR

CAUTION

- The application which runs in a WPAR must have been defined as an ARF Application, then enabled to be started as an Application WPAR.
- *Application Roll-over Facility* will be able to start an application inside a WPAR if the software requirements are met. See *Software Requirements*, on page 1-6 for more information.

When *Application Roll-over Facility* starts an application in a WPAR, the Application WPAR exists only while the application is running. (Application start script should never end).

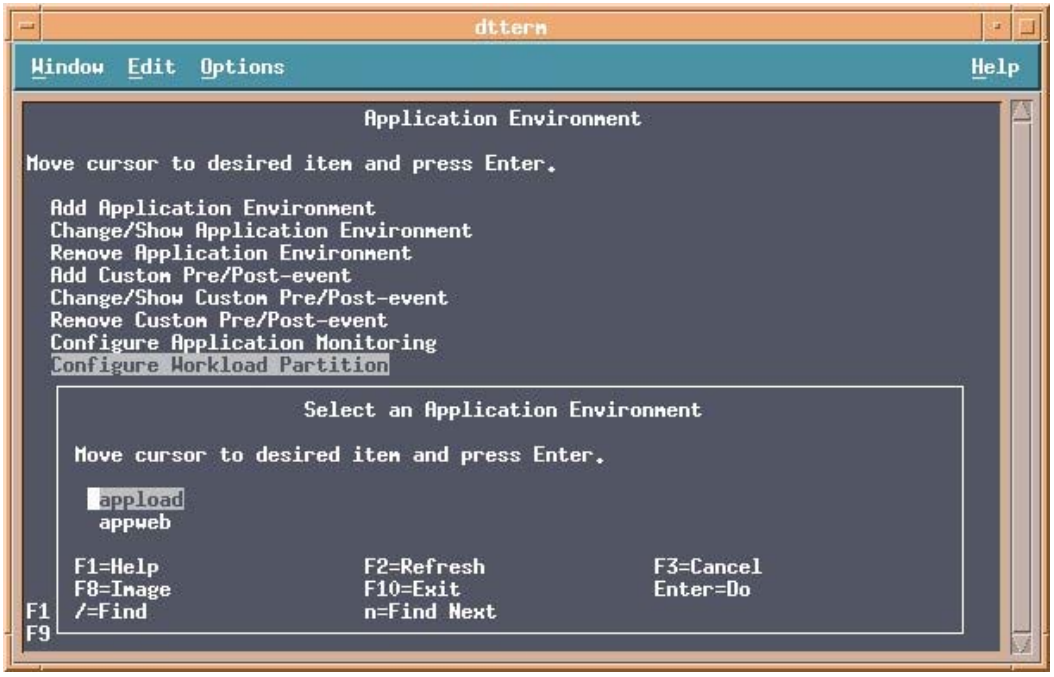
To configure the execution of an application inside a WPAR:

1. Type **smit barf** and select the following option: **Configuration Definition > Application Environment** or use the **smit barf_conf_app_menu** fast path. The following menu appears:

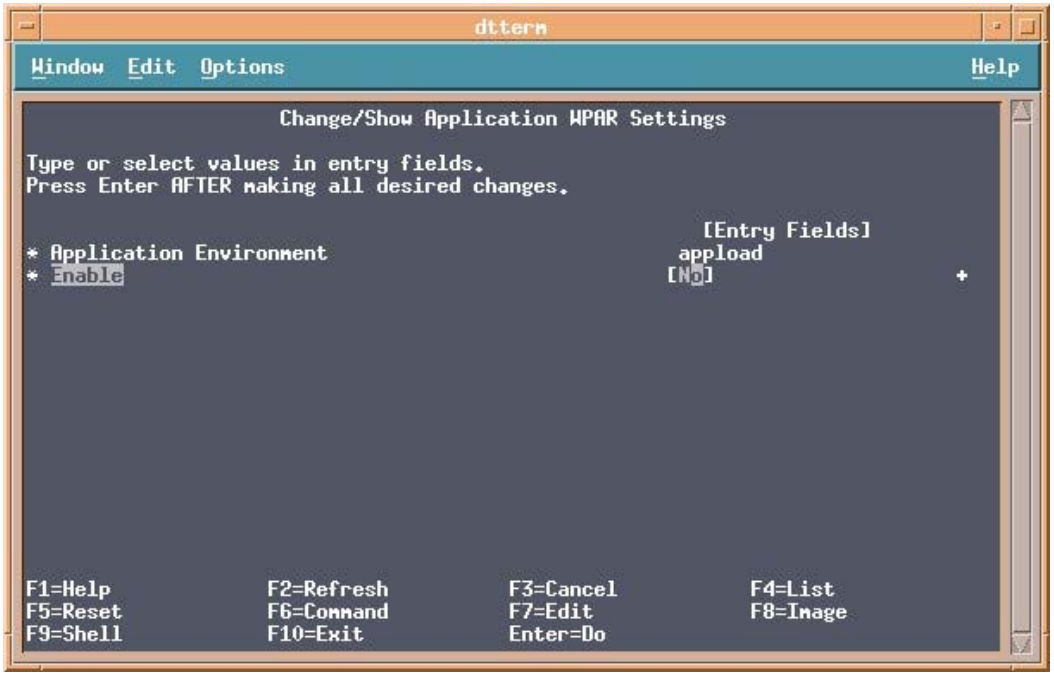
Figure 26. Configure Workload Partition

2. Select **Configure Workload Partition**.

3. Select the Application to modify (for example: **appload**).



4. The following menu appears:



Application Environment	Name of the Application
Enable	'Yes' to start the Application in a WPAR. 'No' to start the Application in the ARF node (general environment).

5. Propagate the new configuration: return to the **Configuration Definition** menu and select the following option: **Propagate Configuration**.
6. Verify the new configuration: return to the main menu and select the following option: **Verify Configuration**.
7. Press F10 to exit SMIT.

6.15. Configuring an Application to Start in a System WPAR

6.15.1. Configuring reservation on disks

Check that there is no reservation on the disks of VIOS:

```
lsattr -El hdiskpowerx
```

The **reserve_lock** parameter must be set to "**no**". If it is not the case, change it:

```
chdev -l <hdisk> -a reserve_lock=no
```

6.15.2. Creating a System WPAR

Before configuring an Application to start in a System WPAR, the System Workload Partition must be created.

To create a System WPAR, perform the following steps:

The name of the WPAR is the same on the two nodes.

1. On the first node:
   a. Create a volume group for the WPAR.
   b. Vary on the volume group for the WPAR.
   c. Create a System Workload Partition. You can either use SMIT:

      ```
      smit wpar
          Administer SYSTEM Workload Partition
              Create a System Workload Partition or specification File
                  Create a System Workload Partition
      ```

      Or run the following command:

      ```
      mkwpar -n <wparname> -d <base directory> -g <volumegroup>
      ```

   d. Verify that the System Workload Partition can be running, using the **startwpar** command.
   e. Stop the System Workload Partition.
   f. Vary off the volume group for the WPAR.
2. On the second node,
   a. Import the volume group for the WPAR.
   b. Create the System Workload Partition with the same name and with preserve mode, by running the following commands:

      ```
      mkwpar -p -n <wparname> -d <base directory> -g <volume group>
      chmod 700 <basename directory>
      mount <base directory>
      mkdir -p <base directory>/usr
      mkdir -p <base directory>/opt
      mkdir -p <base directory>/proc
      ```

Modify the /etc/filesystems file:
Add stanzas for <base directory>/usr, <base directory>/opt, <base directory>/proc as in the /etc/filesystems file of the first node.
Run the command:

```
umount <base directory>
```

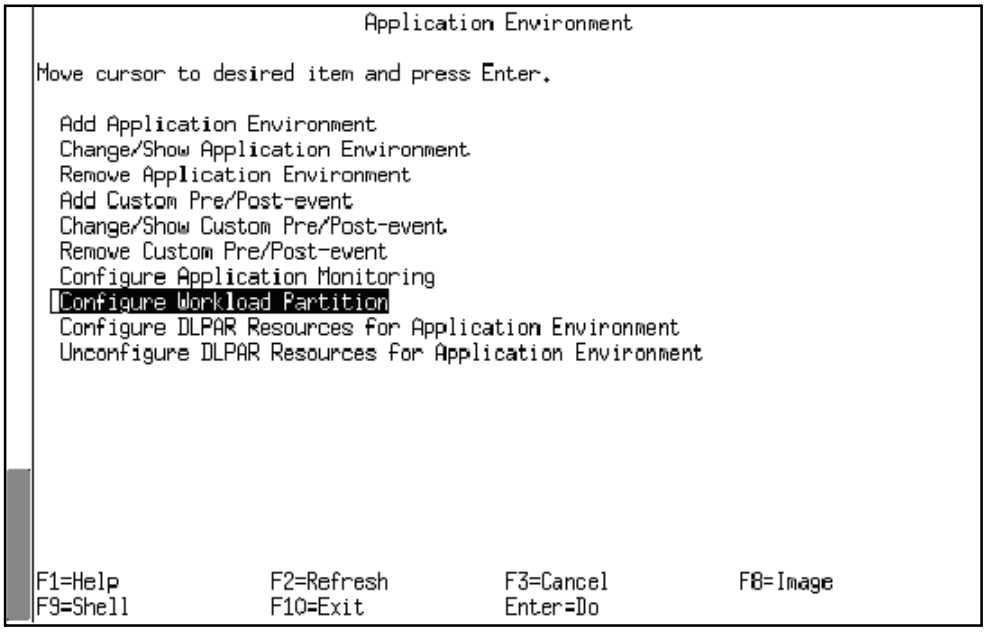
c. Run the **startwpar** command to verify that the System Workload Partition can be running.

d. Stop the System Workload Partition.

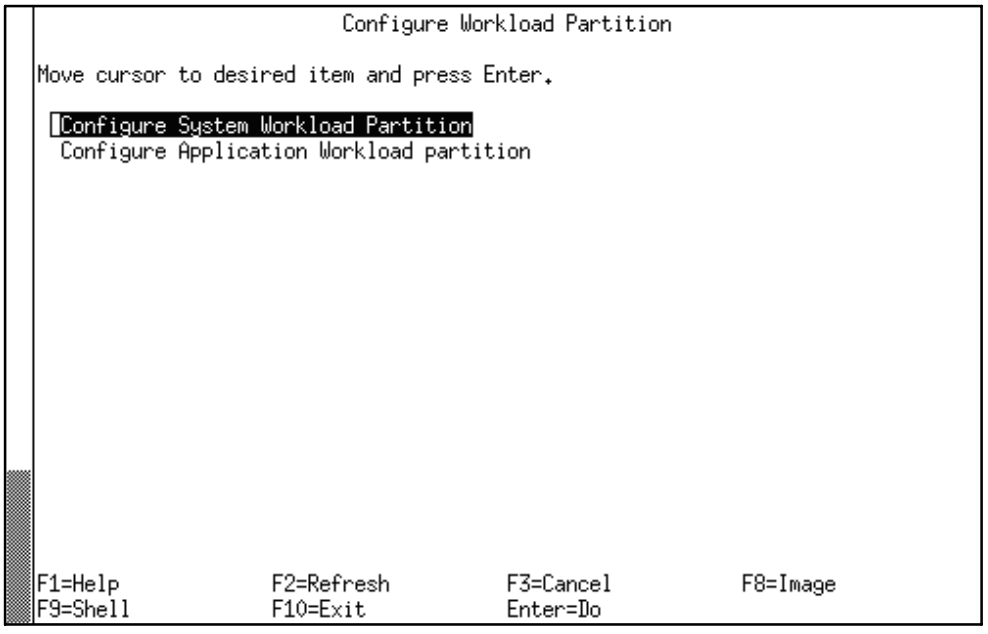
e. Vary off the volume group for the WPAR.

To configure the execution of an application inside a System WPAR:

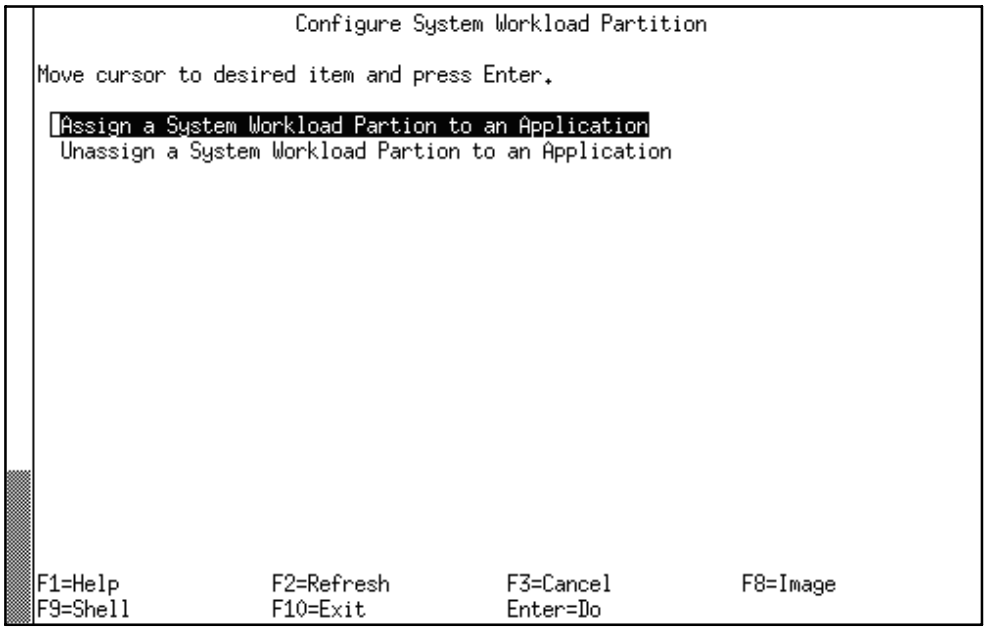
1. Type **smit barf** and select the following options: **Configuration Definition > Application environment**, or use the **smit barf_conf_app_menu** fast path. The following menu appears:



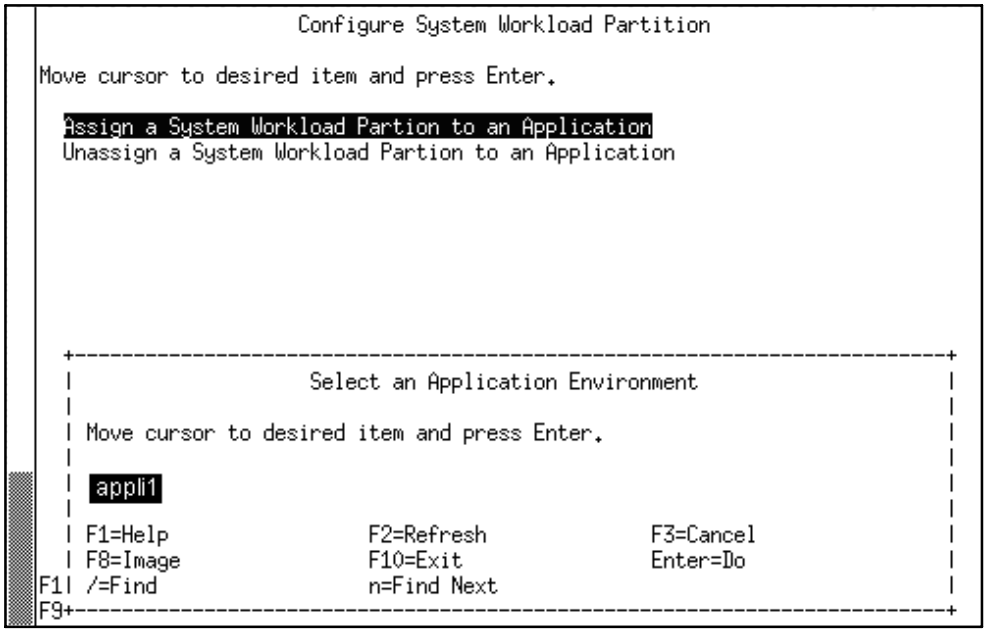
2. Select **Configure Workload Partition**. The following menu appears:



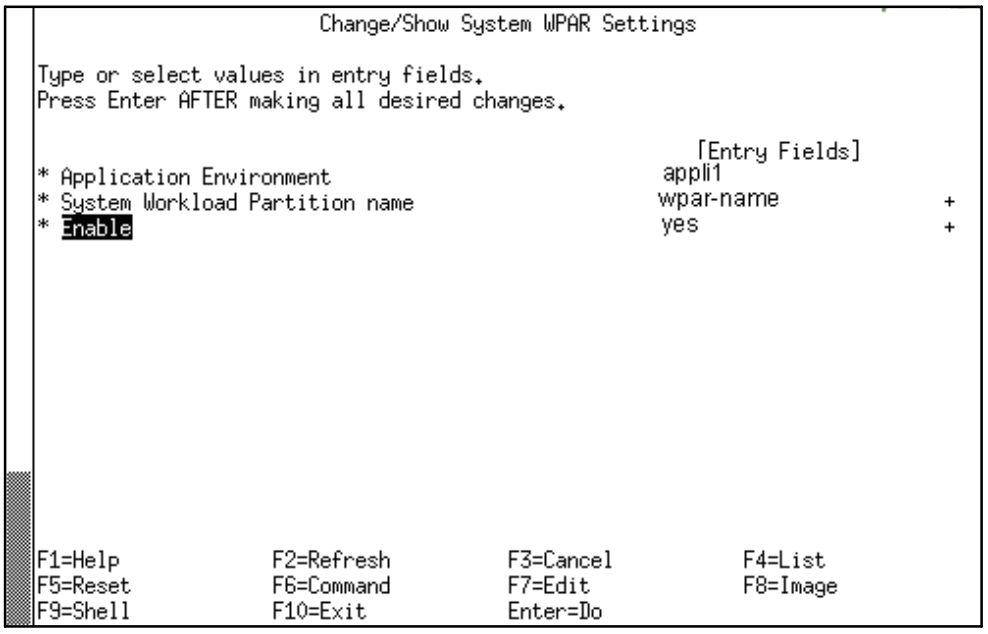
3. Select Configure System Workload Partition. The following menu appears:



4. Select Assign a System Workload Partition to an Application. The following menu appears:



5. Select the Application to modify (for example appli1). The following menu appears:



Application Environment	Name of the Application
System Workload Partition Name	Name of the System WPAR where the application is started.
Enable	Select 'Yes' to start the Application in a System WPAR.

6. Propagate the new configuration: Return to the **Configuration Definition** menu and select the **Propagate Configuration** option. Check that the **Verify** parameter is set to '**yes**'.
7. Press F10 to exit SMIT.

6.15.3. Installing the ARF Application Scripts

Important The scripts to start and stop ARF application must be installed in the /usr/sbin/barf/scripts directory.

Chapter 7. Configuring AIX Mirroring for Disaster Recovery

In a disaster recovery solution (two sites) using AIX mirroring function, *Application Roll-over Facility* uses the AIX function allowing to copy a logical volume defined on a subsystem disk in one site to another logical volume defined on another subsystem disk in the other site.

7.1. Activating Disaster Recovery

ALL the Volume Groups for ALL the applications managed by ARF, MUST be in AIX mirroring mode.

To make effective the AIX mirroring environment (previously configured using AIX SMIT menus), it is necessary to run the following SMIT menus before starting *Application Roll-over Facility* on the cluster nodes:

```
smit barf
     Configuration Definition
         Disaster Recovery Configuration
             ACTIVATE Disaster Recovery     Yes
```

7.2. Configuring fc_err_recov to fast_fail for Disaster Recovery

For Disaster Recovery configuration using two AIX mirrored EMC disk subsystems with PowerPath, it is mandatory to change the **fc_err_recov** attribute of all **fscsi<x>** adapters to **fast_fail**, on ALL nodes:

```
# lsdev -C | grep fscs
# chdev -l fscsi<x> -a fc_err_recov=fast_fail -a dyntrk=yes -P
```

To check that the **fc_err_recov** attribute value is set to **fast_fail** (NOT **delayed_fail**), run:

```
# lsattr -El fscsi<x>
```

Then reboot AIX:

```
# shutdown -Fr
```

Chapter 8. Support of Live Partition Mobility

8.1. Live Partition Mobility Overview

The Live Partition Mobility feature is available on Power6-based servers running AIX5L 5.3 TL7 and later or AIX Version 6.1 and later.

Live Partition Mobility allows you to migrate running AIX partitions and their hosted applications from one physical server to another without disrupting the infrastructure services. The migration maintains complete system transactional integrity. This operation, which takes only a few seconds, transfers the entire system environment, including processor state, memory, attached virtual devices, and connected users.

Live Partition Mobility helps you meet continuously increasingly stringent service-level agreements (SLAs) because it allows you to proactively move running partitions and applications from one server to another.

An application managed by *Application Roll-over Facility* is migrated from one server to another without disrupting the application services and without modifying the behaviour of ARF.

For more information about Live Partition Mobility, refer to:

- *AIX Documentation*
- *IBM System p Live Partition Mobility Redbook* (ref. SG24-7460-00).
  Refer to http://www.redbooks.ibm.com/

8.2. License Key

To use Live Partition Mobility with ARF, it is mandatory to install the ARF license key for the system you want to migrate to (Destination system). To do this, you need to ask for a definitive key, providing the **system id** of one partition of the Destination system, for example the Destination VIO Server.

Note The **system id** is the result of the **uname -Mu** command

When you have obtained a definitive key, append the content of the key file into the **/usr/sbin/barf/data/barf.key** file on the partition you want to migrate to.
So, the **/usr/sbin/barf/data/barf.key** file will have at least two or more key numbers, each key number corresponding to each system you want to migrate the partition.

For more information refer to *License Key*, on page 1-5.

8.3. Configuring ARF for Live Partition Mobility

- When you configure a node for Live Partition Mobility using the **Add a managed node** menu, it is mandatory to:
  - Complete the **HMC Address(es) used for DLPAR and PoD operations and Live Partition Mobility** field
  - Set the **Use Live Partition Mobility** parameter to 'yes'.

  See *Defining Nodes*, on page 6-1 for more details.
- In order to use remote command operations on the HMC, **ssh** must be installed on all the nodes. To do this, refer to *Installing and Configuring ssh*, on page 9-2.

8.4. Inactive ARF Partition Mobility

After the migration of an inactive ARF partition, the *Application Roll-over Facility* configuration must be updated as follows:

1. Update the **Current VIO Server** and the **Current System Name**:
   - Type **smit barf** and select the following options:
     Configuration Definition > Managed Nodes > Change/Show a Managed Node
     (or use the **smit barf_ch_node** fast path).
   - Update the following fields with the value corresponding to your current configuration:
     . **Current VIO servers address list used for disk IO**
     . **Current Managed System Name**

     Press Enter.
2. Synchronize the *Application Roll-over Facility* configuration:
   - Type **smit barf** and select the following options:
     Configuration Definition > Application Environment > Propagate Configuration
     (or use the **smit barf_sync_conf** fast path).
   - Press Enter.
     The *Application Roll-over Facility* definition is copied to the other nodes.
   - Press F10 to exit SMIT.

Chapter 9. Configuring DLPAR and On/Off PoD Resources with ARF

9.1. DLPAR and PoD Overview

To respond to workload peaks on partitioned systems, the **Dynamic Logical Partitioning (DLPAR)** and the **Power on Demand (POD)** can be used.

DLPAR allows you to dynamically allocate additional resources (such as memory and CPU) from the free spool to each logical partition (LPAR), if needed without stopping the partition.

Power on Demand (PoD) is one of the features of the DLPAR function that lets you activate pre-installed but yet inactive (and unpaid for) CPU and memory resources.

There are several types of **PoD** licenses. Only the **On/Off PoD** is integrated in Application Roll-over Facility.

Note The term **Capacity Upgrade on Demand** is also used to designate **Power on Demand**.

9.1.1. On/Off PoD

On/Off PoD allows you to **temporarily** activate and deactivate processors and memory units. The On/Off PoD feature provides you with a key that you must enter on your server using the **Hardware Management Console** (HMC). This enables your server for the use of temporary capacity.
When needed, you can request temporary activation of resources, specifying quantities of capacity and number of days. Granularity is one processor and one GB memory per day.

The ordering of the On/Off enablement (by SFR) includes a specific contract you sign with Bull. This contract requires you to report billing data at least once per month, regardless of whether you have used temporary capacity during the period.

For more information about the **On/off PoD** feature, contact your Bull representative.

9.1.2. On/Off PoD in ARF

ARF can activate **On/Off PoD** resources and dynamically allocate them to the LPAR node before the application is started and release them after the application is stopped. In case of application failover, this allows the application to run without loss of performance.

When you configure an application environment, you can define the minimum and desired amount of resources required for that application. ARF determines if additional resources need to be allocated for the node hosting the application and tries to allocate as many resources as possible to meet the desired amount for that application.

If the amount of resources in the free pool is insufficient to satisfy the total amount requested for allocation, ARF requests resources from **On/Off PoD** if allowed.

ARF starts counting the extra resources required for the application from the minimum amount of the partition. The minimum partition resources are retained for the node's overhead operations and are not used to host an application.

If the amount of resources that can be allocated from the free spool and from On/Off PoD is less than the minimum amount specified for the application, the application does not start.

When the application moves to another node, ARF releases the resources that are no longer necessary to support this application on the node.

9.2. Prerequisites and Preliminary Tasks

To use the DLPAR and On/Off PoD functions in ARF, the following requirements must be satisfied:

- AIX 5L V5.3 or AIX Version 6.1 installed on ARF nodes
- Openssh 3.4p1 or greater installed on ARF nodes
- Version 5 or greater on HMC
- Activation code for On/Off PoD entered on HMC
- The LPAR partition name, the AIX hostname and the ARF node name must all match
- HMC SSH access configured.

9.3. Installing and Configuring SSH

In order to use remote command operations on the HMC, **ssh** must be installed on all the nodes.

9.3.1. Installing SSH

openssh fileset and its prerequisite **openssl** rpm package must be installed on each ARF node.

Install first **openssl** from the *AIX Toolbox for Linux Applications* CD then install **openssh.base** fileset from *Expansion Pack* CD.

9.3.2. Configuring HMC SSH Access

To configure HMC ssh access, do the following steps:

- Enable HMC ssh access on HMC
- Generate SSH keys on ARF nodes
- Enable no password HMC access

9.3.3. Enable HMC SSH access on HMC

On HMC GUI, select

HMC Maintenance
    System Configuration
        Enable/Disable Remote Command Execution

Select the box to enable ssh.

Repeat these steps on each HMC.

9.3.4. Generate SSH keys

On the node, use the following steps to generate an RSA key pair of the SSH protocol. This is the default starting with OpenSSH.

1. Log in as **root** user and go to the directory **/.ssh**
2. Enter the command **ssh-keygen –t rsa**
3. Accept default location key file and do not enter passphrase.

4. The output is the following:

```
Generating public/private rsa key pair.
Enter file in which to save the key (//.ssh/id_rsa):
Enter passphrase (empty for no passphrase):
Enter same passphrase again:
Your identification has been saved in //.ssh/id_rsa.
Your public key has been saved in //.ssh/id_rsa.pub.
The key fingerprint is:
d6:3f:11:da:44:63:ee:17:0a:e0:98:ca:3b:16:4d:fe root@lpar2
Repeat these steps on each node.
```

9.3.5. Enable no password HMC access

1. Transfer the file **authorized_keys2** of user **hscroot** from HMC (IP address *129.183.12.32*) to the **/tmp** directory of the node using the **scp** command:

   ```
   scp hscroot@129.183.12.32:~/.ssh/authorized_keys2 /tmp
   ```

   Answer **yes** at **Are you sure you want to continue connecting** question.

   Enter **hscroot** password when asked.

   ```
   The authenticity of host '129.183.12.32 (129.183.12.32)' can't be
   established.
   RSA key fingerprint is
   7b:dd:d5:4a:53:02:d8:a0:46:e2:82:30:e4:b5:40:99.
   Are you sure you want to continue connecting (yes/no)? yes
   Warning: Permanently added '129.183.12.32' (RSA) to the list of
   known hosts.
   Password:
   authorized_keys2                    100%    0     0.0KB/s   00:00
   ```

2. Append the file **/.ssh/id_rsa.pub** to the **/tmp/authorized_keys2** file:

   ```
   cat /.ssh/id_rsa.pub >> /tmp/authorized_keys2
   ```

3. Transfer back the **authorized_keys2** file from node to the HMC:

   ```
   scp /tmp/authorized_keys2
   hscroot@129.183.12.32:~/.ssh/authorized_keys2
   ```

4. Verify the communication between node and HMC by entering the command (no password prompt):

   ```
   ssh hscroot@129.183.12.32 date

   Tue Aug 29 14:55:22 CEST 2006
   ```

Repeat these steps for each node and for each HMC.

9.4. Configuring DLPAR/PoD for ARF

9.4.1. Defining HMC and Managed System Names

When defining a node for ARF, enter the HMC IP address (dot format) and the managed system name to which the partition node belongs. The managed system name is the name that appears in HMC GUI.

```
                              Add a Managed Node

Type or select values in entry fields.
Press Enter AFTER making all desired changes.

                                                      [Entry Fields]
*   Node Name                                        [lpar2]
*   Addresses List                                   [172.16.108.52]

  TTY device for PPP hearbeat                        []
  VIO servers address list used for disk IO          []
  Address(es) used for Remote Physical Volumes       []

  HMC Address(es) used for DLPAR and PoD operations [172.16.108.112]       +
  Managed System Name                                [plmiz1]              +
```

9.4.2. Defining Application Environment DLPAR/PoD Resources

To define DLPAR/PoD resources for an application environment, enter:

```
smit barf
   Configuration Definition
      Application Environment
         Configure DLPAR Resources for Application Environment
```

Select the application environment in the list.

The following screen appears:

```
          Configure DLPAR Resources for Application Environment

Type or select values in entry fields.
Press Enter AFTER making all desired changes.

                                                      [Entry Fields]
* Application Environment                             app1
* Minimum number of CPUs (process units)             [1.5]
* Desired number of CPUs (process units)             [2.5]
* Minimum amount of memory (MB)                       [512]                #
* Desired amount of memory (MB)                      [1024]                #

* Use PoD if resources are insufficient?            [Yes]                  +
  (this might result in extra costs by
  using the PoD license)
* Number of days for PoD resources                   [1]                    #
```

Minimum number of CPUs

Enter the minimum number of process units to acquire when the application environment is activated. Default value is 0.
ARF checks how many process units the node currently has above its LPAR

minimum value, compares with the value entered in this field and allocate more process units to the partition node, if needed, to satisfy the request (in the limit of maximum LPAR value allowed).

When evaluating the process units to allocate to the node, the LPAR minimum is a starting point for calculations.

Desired number of CPUs
: Enter the desired number of process units to acquire when the application environment is activated. Default value is 0.

  ARF may allocate fewer if there is not enough available in the free spool.

Minimum amount of memory
: Enter the minimum amount of memory (in 256 MB increments) to acquire when the application environment is activated. Default value is 0.
  ARF checks how much memory the node currently has above its LPAR minimum value, compares with the value entered in this field and allocate more memory to the partition node, if needed, to satisfy the request (in the limit of maximum LPAR value allowed).

  When evaluating the process units to allocate to the node, the LPAR minimum is a starting point for calculations.

Desired amount of memory
: Enter the desired amount of memory (in 256 MB increments) to acquire when the application environment is activated. Default value is 0.

  ARF may allocate fewer if there is not enough available in the free spool.

Use PoD if resources are insufficient?
: Default is no. Enter yes if you want that ARF use Power on Demand (PoD) to obtain enough resources to fulfill the minimum request (in addition to dynamic allocation of resources available in the free spool).

  Using PoD requires an activation code to be entered on the HMC (Hardware Management Console) and may result in extra costs.

Number of days for PoD resources
: The number of days for which the On/Off PoD resources are requested.

9.4.3. Removing Application Environment DLPAR/PoD Resources

To remove DLPAR/PoD resources for an application environment, enter:

```
smit barf
   Configuration Definition
      Application Environment
         Unconfigure DLPAR Resources for Application
Environment
```

Select the application environment in the list for removing its DLPAR resources configuration.

9.4.4. Displaying Application Environment DLPAR/PoD Resources status

To display DLPAR and PoD resources status, enter:

```
smit barf
   Manage Application Environment
      Show DLPAR and PoD Resources Status
```

The following screen appears:

```
                    Show DLAR and PoD Resources Status

Type or select values in entry fields.
Press Enter AFTER making all desired changes.

                                                        [Entry Fields]
* Managed Node(s)                                    [ALL]                  +
* DLPAR and PoD Resources Type                       [ALL]                  +
```

Managed Node(s)
: Select ALL for all nodes or a specific node from the list.

DLPAR and PoD Resource Type
: Select ALL for resource of type memory and processor or a specific type from the list

Example of output:

```
                                COMMAND STATUS

Command: OK            stdout: yes           stderr: no

Before command completion, additional instructions may appear below.

---- SYSTEM ON/OFF PoD RESOURCES ----
SYSTEM      PARTITION       TYPE STATE   ACTIV AVAIL UNRET DAYS HOURS DAYS
                                                        LEFT LEFT AVAIL
plmiz1        lpar2           proc This system is not On/Off PoD capable.
plmiz1        lpar2           mem  This system is not On/Off PoD capable.
PL1650R       navajo          proc Avail   0     6     0     0    0     354
PL1650R       navajo          mem  Running 1024  3072  0     1    21    995

---- PARTITION RESOURCES ----
SYSTEM      PARTITION       TYPE FREE  ALLOC MIN   MAX   (UNITS)
plmiz1        lpar2           proc 0.9   0.2   0.1   1.0   (proc units)
plmiz1        lpar2           mem  1168  480   256   800   (MB)
PL1650R       navajo          proc 0.1   0.3   0.3   3.0   (proc units)
PL1650R       navajo          mem  768   768   256   2048  (MB)

---- ACTIVE APPLICATION ENVIRONMENT DLPAR RESOURCES ----
SYSTEM      PARTITION       APPLICATION    MIN   DES   MIN   DES   PoD  DAYS
                                           CPU   CPU   MEM   MEM
PL1650R     navajo          app1           0     0     256   512   yes  1
                                          ----  ----  ----  ----
                               TOTAL       0     0     256   512
```

9.5. Examples

An ARF configuration with 2 LPAR nodes: lpar1 and lpar2 and 2 applications environments app1 and app2.

lpar1 is a partition of managed system pl1 managed by HMC hmc1.

lpar2 is a partition of managed system pl2 managed by HMC hmc2.

9.5.1. HMC Access Configuration

On lpar1 as user root, generate ssh key and append the public generated key file to the file authorized_keys2 (user hscroot) on hmc1 and hmc2.

On lpar2 as user root, generate ssh key and append the public generated key file to the file authorized_keys2 (user hscroot) on hmc1 and hmc2.

9.5.2. ARF Node Configuration

When defining ARF nodes, in addition to address nodes, define HMC addresses and managed system names

Node name	HMC address	Managed system name
lpar1	hmc1 dot address	pl1
lpar2	hmc2 dot address	pl2

Note ARF node names, hostnames and LPAR names must match.

9.5.3. ARF Application Activation Process

Example:

The application app1 process units requested is configured as follows:

- 0.6 (minimum)
- 1.2 (desired).

The lpar1 minimum value is 0.1, the maximum value is 1.0 and lpar1 has currently 0.1 process units.

The free spool has 1.2 process units available.

When app1 is activated on lpar1, ARF requests 0.9 process units from free spool.

(The allocated resources cannot exceed the maximum value defined for the partition).

ARF handles in the same way memory resource.

Other scenarios

The following table shows several scenarios (On/Off PoD not used):

PARTITION Description				APPLICATION resources		COMMENTS
free	alloc	min	max	min	Desired	
1.2	0.1	0.1	1.0	0.6	1.2	ARF allocates 0.9 from free spool LPAR maximum reached Minimum value satisfied Desired value not satisfied Application starts, 0.9 available for application minimum < 0.9 < desired
1.2	0.1	0.1	1.0	0.6	0.8	ARF allocates 0.8 from free spool desired value satisfied (0.8=desired) Application starts, 0.8 available for the application
0.3	1.0	0.1	1.0	0.6	1.2	ARF allocates nothing already allocated=LPAR max 0.9 already allocated is available for application Minimum value satisfied (0.6 < 0.9) Desired value not satisfied
1.2	0.1	0.1	1.0	1.2	1.5	Error: application does not start Minimum value for application > LPAR maximum
1.2 0.9	0.1 0.4	0.1	1.0	0.1 0.7(*)	0.3 (app1) 1.2 (app2)	ARF allocates 0.3 from free spool for app1 desired value satisfied (0.3=desired for app1) Application app1 starts The LPAR allocated resource is now 0.4 The free resource is now 0.9 ARF allocates 0.6 from free spool for app2 LPAR maximum reached Minimum values satisfied for app1 and app2 (0.1+0.7) Desired values not satisfied for app1 and app2 Application app2 starts

Note If minimum value for app2 is 0.9 instead of 0.7, the app2 application cannot start: minimum values for app1 and app2 are 0.1 + 0.9 = 1.0, minimum LPAR=0.1, total=1.1 > LPAR maximum.

ARF handles in the same way memory resource.

Example with use of On/off PoD

The application app1 process units requested is configured as follows:
minimum=1.0 and desired=2.0, use PoD=yes
The application app2 process units requested is configured as follows:
minimum=0.8 and desired=0.9, use PoD=yes

The LPAR values are:
Minimum=0.1
Maximum=3.0

- Current configuration
  - The currently allocated process units to the LPAR are 0.3
  - There are 1.35 process units in the free spool.
- Starting app1 application
  - To start app1 and satisfy desired value (2.0), ARF calculates process units to allocate:
    desired value + LPAR minimum – allocated value
    2.0 + 0.1 – 0.3 = 1.8 process units
    But free spool contains only 1.35 process units.
  - ARF activates 1.0 process units from On/off PoD (granularity for On/Off PoD is 1)
    So free spool contains now 2.35 process units.
  - Then ARF allocates 1.8 process units to the LPAR from the free spool.
- New configuration
  - The currently allocated process units to the LPAR are now 2.1
  - There are 0.55 process units in the free spool.
  - app1 is started on LPAR and its desired value (2.0) is satisfied.
- Starting app2 application
  - To start app2 and satisfy desired value (0.9), ARF calculates process units to allocate:
    desired values + LPAR minimum – allocated value
    2.0 + 0.9 + 0.1 – 2.1 = 0.9 process units
    But free spool contains only 0.55 process units.
  - ARF activates 1.0 process units from On/off PoD (granularity for On/Off PoD is 1)
    So free spool contains now 1.55 process units.
  - Then ARF allocates 0.9 process units to the LPAR from the free spool
- New configuration
  - The currently allocated process units to the LPAR are now 3.0 (LPAR maximum reached).
  - There are 0.65 process units in the free spool.
  - app2 is started on LPAR and its desired value (0.9) is satisfied.

The app1 and app2 applications are running on LPAR and their desired values are satisfied.
The LPAR maximum is reached so it is not possible to allocate more resources to the LPAR.
To start another application on this LPAR, its minimum cannot be greater than 1.1:
LPAR maximum – LPAR minimum -minimum app1 - minimum app2
3.0 – 0.1 – 1.0 – 0.8 = 1.1

ARF handles memory resource in the same way.

9.5.4. Releasing DLPAR and PoD Resources

When the application is stopped, ARF releases only the resources that are no longer necessary to support this application on the node. The resources are released to the free spool first and On/Off PoD resources are then deactivated if possible.

9.6. Trace and Log Files

9.6.1. Trace Files

Detailed output of acquiring and releasing resources operations is logged in trace files under /var/barf/trace directory.

The name of the trace files is built as follow:

BARF_ACQUIRE_DLPAR_RES_<node>_<application>.mmddyy_hhmmss
BARF_RELEASE_DLPAR_RES_<node>_<application >.mmddyy_hhmmss

Where :

node is the node name
application is the application name
mmddyy is the date
hhmmss is the hour

9.6.2. Log File

A summary output of acquiring and releasing resources operation is logged in barf.log file under /var/barf/log directory.

Examples:

- Activate application app1 on node navajo (barf.log file):

```
Tue Sep 5 16:36:40 DFT 2006: app1: STARTING EXEC of
barf_acquire_dlpar_res app1
Retrieve PoD resources ...
Available On/Off PoD resources for mem: 4096
********************** mem *********************************

Application Environment configuration: app1
Minimum required: 256 (from free pool and if necessary, from PoD pool)
Desired required: 512 (from free pool and if necessary, from PoD pool)

LPAR configuration: navajo
Minimum: 256
Maximum: 2048
Allocated: 256
Free: 256

Active applications resources: ""
Minimum required: 0
Desired required: 0

Calculated LPAR resources for active applications:
Resources needed to meet minimum: 256
Resources needed to meet desired: 512

Calculated values:
Will activate 1024 from the PoD pool
Will acquire  512 from the free pool
ACQUIRE 1024 mem (PoD) then 512 mem (free) ...
chhwres timeout: 5

*******************************************************************************
ACQUIRE 512 mem DONE SUCCESSFULLY
        The following On Demand resources were acquired for application -
app1.
        LPAR CPUs: 0    CoD CPUs: 0
        LPAR memory: 512 MB     CoD memory: 1024 MB
Tue Sep 5 16:37:13 DFT 2006: app1: SUCCESSFULL EXEC of
barf_acquire_dlpar_res app1
```

- Deactivate application app1 on node navajo (barf.log file):

```
Tue Sep 5 16:57:55 DFT 2006: app1: STARTING EXEC of
barf_release_dlpar_res app1
V6R1.0
Retrieve allocated PoD resources ...
********************** mem ********************************

Application Environment configuration: app1
Minimum required: 256 (from free pool)
Desired required: 512 (from free pool)

LPAR configuration: navajo
Minimum: 256
Allocated: 768

Active applications resources: ""
Minimum required: 0
Desired required: 0

Calculated LPAR resources for active applications:
Minimum needed: 256
Desired needed: 256

Calculated values:
Will release 512 to the free pool
chhwres timeout: 5

RELEASE 512 mem ...
RELEASE 512 mem DONE SUCCESSFULLY
RELEASE all PoD resources of type mem DONE SUCCESSFULLY
Application Environment app1 has no resources of type proc to release.
        The following On Demand resources were released for application -
app1.
        LPAR CPUs: 0    CoD CPUs: 0
        LPAR memory: 512 MB     CoD memory: 1024 MB
Tue Sep 5 16:58:29 DFT 2006: app1: SUCCESSFULL EXEC of
barf_release_dlpar_res app1
```

9.7. Using Custom Pre- and Post-events

Before starting an application that requires additional resources, it is possible to free some resources from partitions that are not part of the ARF configuration:
either by shutting down them or by releasing few resources from these partitions using DLPAR feature.

In the same way, after stopping an application, it is possible to re-allocate resources to partitions that are not part of the ARF configuration:
either by rebooting them, or by allocating resources to these partitions using DLPAR feature.

This can be done using ARF custom event facility. The events to customize are the following:

- ACTIVATEENV pre-event (executed just before ARF acquires resources for an application)
- DEACTIVATEENV post-event (executed just after ARF releases resources for an application).

The commands to be used in these scripts are lshwres to list partition resources and chhwres to run DLPAR operations on partitions. These commands must be called via ssh on the HMC.

Examples:

- To list information on memory for partition lpar1 on managed system pl1 managed by HMC hmc1 (output format defined by –F option):

```
ssh hscroot@hmc1 lshwres –m pl1 –r mem ––level lpar ––filter
”lpar_names=lpar1” –F curr_mem:curr_min_mem:curr_max_mem
```

  The output looks like:

```
512:256:768
```

  This means:
  current allocated memory = 512 MB
  minimum partition value = 256 MB
  maximum partition value = 768 MB

- To list information on process units for partition lpar1 on managed system pl1 managed by HMC hmc1 (output format defined by –F option):

```
ssh hscroot@hmc1 lshwres –m pl1 –r proc ––level lpar ––filter
”lpar_names=lpar1” –F
curr_proc_units:curr_min_proc_units:curr_max_proc_units
```

  The output looks like:

```
0.3:0.1:0.5
```

  This means:
  current allocated process units=0.3
  minimum partition value=0.1
  maximum partition value=0.5

- To release 256 MB of memory from partition lpar1:

```
ssh hscroot@hmc1 chhwres –m pl1 –p lpar1 –r mem –o r –q 256
```

- To allocate 256 MB of memory to partition lpar1:

```
ssh hscroot@hmc1 chhwres –m pl1 –p lpar1 –r mem –o a –q 256
```

- To release 0.1 of process units from partition lpar1:

```
ssh hscroot@hmc1 chhwres –m pl1 –p lpar1 –r proc –o r ––procunits 0.1
```

- To allocate 0.1 of process units to partition lpar1:

```
ssh hscroot@hmc1 chhwres –m pl1 –p lpar1 –r proc –o a ––procunits 0.1
```

- To activate the partition lpar1 with lpar1_normal profile:

```
ssh hscroot@hmc1 chsysstate –m pl1 –r lpar –o on –n lpar1 –f lpar1_normal
```

9.8. Useful Commands

You can invoke HMC command line interface using ssh from ARF nodes.

In addition to lshwres and chhwres commands described in previous paragraph, you can use the lscod and chcod commands:

9.8.1. lscod Command

lscod displays PoD information on managed system pl1 (managed by hmc1)

- PoD status

```
ssh hscroot@hmc1 lscod -m pl1 -t cap -c onoff -r mem
mem_onoff_state=Available,activated_onoff_mem=0,avail_mem_for_onoff=40
96,
unreturned_onoff_mem=0,onoff_request_mem_days_left=0,
onoff_mem_day_hours_left=0,onoff_mem_days_avail=999

ssh hscroot@hmc1 lscod -m pl1 -t cap -c onoff -r proc
proc_onoff_state=Available,activated_onoff_procs=0,avail_procs_for_ono
ff=6,
unreturned_onoff_procs=0,onoff_request_proc_days_left=0,
onoff_proc_day_hours_left=0,onoff_proc_days_avail=354
```

- billing information

```
ssh hscroot@129.183.12.32 lscod -m PL1650R -t bill -c onoff -r proc

sys_type=9117,sys_serial_num=10EFACD,anchor_card_ccin=528E,
anchor_card_serial_num=054000844,anchor_card_unique_id=43150434506313B
7,
resource_id=7951,sequence_num=0041,
activated_onoff_resources=0000,avail_resources_for_onoff=0006,
hist_expired_resource_days=0006,hist_unreturned_resource_days=0011,
collection_date=2006-09-04,collection_time=08:14:53,
total_sys_run_time_hours=000240,signature=AE4AFC4F69591CBB,
entry_check=E0,status=01

ssh hscroot@129.183.12.32 lscod -m PL1650R -t bill -c onoff -r mem

sys_type=9117,sys_serial_num=10EFACD,anchor_card_ccin=528E,
anchor_card_serial_num=054000844,anchor_card_unique_id=43150434506313B
7,
resource_id=7954,sequence_num=0041,
activated_onoff_resources=0000,avail_resources_for_onoff=0004,
hist_expired_resource_days=0000,hist_unreturned_resource_days=0000,
collection_date=2006-09-04,collection_time=08:15:48,
total_sys_run_time_hours=000240,signature=1FB9554DC1858096,
entry_check=A1,status=01,mem_unit_adj=1
```

- history PoD information

```
ssh hscroot@129.183.12.32 lscod -m PL1650R -t hist
```

9.8.2. chcod Command

chcod performs PoD operations on managed system pl1 (managed by hmc1)

- To activate 1GB of memory from PoD for 1day:

```
ssh hscroot@hmc1 chcod -m pl1 –o a –c onoff –r mem –q 1 –d 1
```

- To activate 1 processor of memory from PoD for 1day:

```
ssh hscroot@hmc1 chcod -m pl1 –o a –c onoff –r proc –q 1 –d 1
```

- To deactivate all On/Off PoD resources:

```
ssh hscroot@hmc1 chcod -m pl1 –o d –c onoff –r proc
ssh hscroot@hmc1 chcod -m pl1 –o d –c onoff –r mem
```

All these commands are also available through the HMC GUI interface.

Chapter 10. Configuring GLVM for ARF

10.1. GLVM Overview

The *Geographic Logical Volume Manager* (GLVM) is an AIX feature for real time geographic data mirroring over standard TCP/IP networks. GLVM can help protect your business from a disaster by mirroring your mission-critical data to a remote disaster recovery site. If a disaster, such as a fire or flood, were to destroy the data at your production site, you would already have an up-to-date copy of the data at your disaster recovery site.

GLVM builds upon the AIX Logical Volume Manager (LVM) to allow you to create a mirror copy of data at a geographically distant location. Because of its tight integration with LVM, users who are already familiar with LVM should find GLVM very easy to learn. You configure geographically distant disks as remote physical volumes and then combine those remote physical volumes with local physical volumes to form geographically mirrored volume groups. These are managed by LVM very much like ordinary volume groups.

For more details about GLVM, refer to article *Using the Geographic LVM in AIX 5L* available at URL: http://www-03.ibm.com/servers/aix/whitepapers/aix_glvm.html and click on *Full Text Whitepaper* item to see the full document.

10.2. GLVM Components

The main function of the GLVM is mirroring data of a node at a local site across an IP-based network to a node at a remote site.

GLVM uses the mirroring function of the AIX LVM and operates as a layer under the LVM.

The GLVM includes the Remote Physical Volume (RPV) device driver. The RPV is a pseudo device driver and its kernel extension that let the LVM consider the physical volume located at the remote site as another local physical volume, although the actual data I/O operations are performed on the remote site.

The local and remote sites do not have to be on the same physical network. Routers and gateways between the two sites are allowed.

TCP/IP network and the *Remote Physical Volume* (RPV) device driver are used for remote disk access.

This is shown in the following diagram:

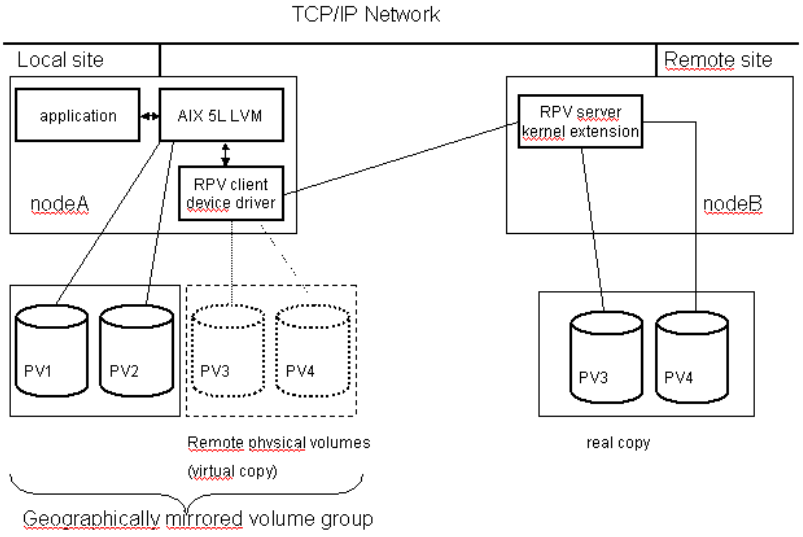


Figure 27. RPV Device Driver

The RPV device driver allows the LVM at the local site to access the disks at the remote site as if they were locally attached. This is accomplished by defining disks PV3 and PV4 to the local site as *remote physical volumes*.

The RPV device driver consists of two parts:

- The *RPV client* resides on the system where the application runs. The RPV client appears like an ordinary disk, and usually has a name such as **hdisk***n*.
- The *RPV server* resides on the system where the real disk is physically attached, and usually has a name such as **rpvserver***n*. The RPV server's job is to process I/O requests from a remote RPV client.

An RPV client and server pair works together to enable the LVM at the local site to access a disk at the remote site. There can be many RPV client and server pairs defined, one for each disk that is to be remotely accessed.

Remote physical volumes look very much like ordinary local physical volumes, except they are slower and less reliable. Remote physical volumes are slower because of the added network delay. They are less reliable because long distance networks, especially those with several routers or gateways, tend to fail more often than local disks.

The *geographically mirrored volume group* is a volume group that contains one or more physical volumes that have copies residing on Remote Physical Volumes. This is managed by LVM very much like ordinary volume group.

10.3. Configuring GLVM

Prerequisite: GLVM must be installed on each ARF node and requires **bos.rte.lvm 5.3.0.20** or higher.

GLVM is configured using SMIT interface.

The following figure shows the initial configuration:

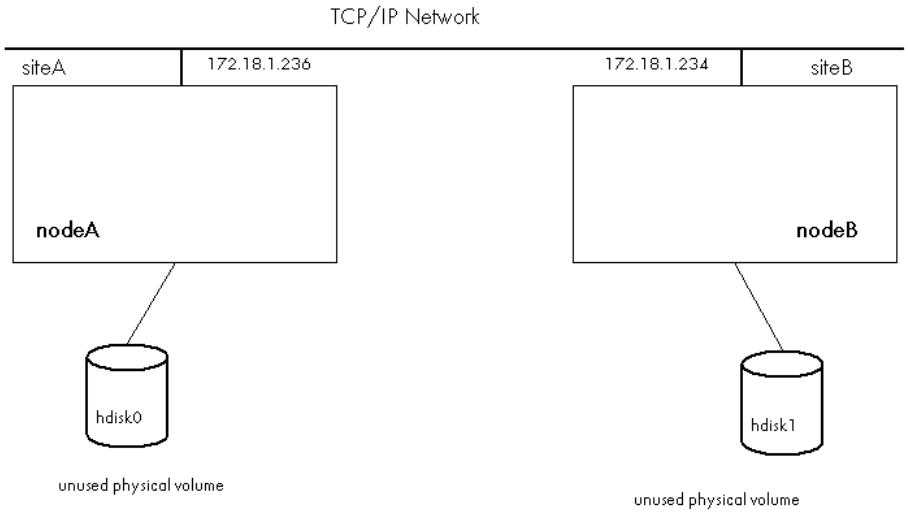


Figure 28. GLVM initial Configuration

10.3.1. Configuring RPV Server

RPV server configuration is done on each node. It consists in:

- Defining RPV server site name.
- Creating RPV servers for real disks.

10.3.1.1. Define RPV Server on Node A

To define the RPV server site name on nodeA, enter:

```
smitty rpvserver
```

1. Select **Remote Physical Volume Server Site Name Configuration.**
2. Select **Define / Change / Show Remote Physical Volume Server Site Name.**
3. Enter the RPV server site name (siteA).

To create a RPV Server for an unused real disk, enter:

```
smitty rpvserver
```

1. Select **Add Remote Physical Volumes Servers.**
2. Select unused physical volume from the list (hdisk0).

```
                        Add Remote Physical Volume Servers

Type or select values in entry fields.
Press Enter AFTER making all desired changes.

                                                        [Entry Fields]
  Physical Volume Identifiers                          00cb153c8a6b0858

* Remote Physical Volume Client Internet Address      [172.18.1.234]        +
  Configure Automatically at System Restart?          [yes]                 +
  Start New Devices Immediately?                      [yes]                 +
```

3. The **Remote Physical Volume Client Internet Address** is the RPV client's IP address. The RPV server will only accept connection requests from this IP address. Enter the node B IP address (`172.18.1.234`). This IP address must be already defined in **/etc/hosts** file.
4. **Configure Automatically at System Restart ?** answer **yes**.
5. **Start New Devices Immediately ?** answer **yes**.

The output is as follows:

```
                                COMMAND STATUS

Command: OK            stdout: yes           stderr: no

Before command completion, additional instructions may appear below.

rpvserver0 Available
```

rpvserver0 is the RPV server for real disk hdisk0 on nodeA.

Note Repeat these steps to create another RPV server for another real disk on nodeA.

To verify the configuration, enter:

```
smit rpvserver
```

1. Select **List all Remote Physical Volumes Servers.**

```
                                COMMAND STATUS

Command: OK            stdout: yes           stderr: no

Before command completion, additional instructions may appear below.

# RPV Server     Physical Volume Identifier       Physical Volume
# ------------------------------------------------------------
  rpvserver0     00cb153c8a6b0858                 hdisk0
```

You can also verify the configuration using the following commands:

```
# lsdev -Ccrpvserver
rpvserver0 Available  Remote Physical Volume Server
```

```
# lsattr -El rpvserver0
auto_online y                                Configure at System Boot     True
client_addr 172.18.1.234                     Client IP Address            True
rpvs_pvid  00cb153c8a6b08580000000000000000 Physical Volume Identifier True
```

10.3.1.2. Define RPV Server on Node B

To define the RPV server site name on nodeB, enter:

```
smitty rpvserver
```

1.. Select Remote Physical Volume Server Site Name Configuration.

2. Select Define / Change / Show Remote Physical Volume Server Site Name.

3. Enter the RPV server site name (siteB).

To create a RPV server for an unused real disk, enter:

```
smitty rpvserver
```

1. Select Add Remote Physical Volumes Servers.

2. Select unused physical volume from the list (hdisk1).

```
                    Add Remote Physical Volume Servers


Type or select values in entry fields.
Press Enter AFTER making all desired changes.

                                                      [Entry Fields]
  Physical Volume Identifiers                        00c0062a857f712a
* Remote Physical Volume Client Internet Address     [172.18.1.236]        +
  Configure Automatically at System Restart?         [yes]                 +
  Start New Devices Immediately?                     [yes]                 +
```

3. Enter the nodeA IP address (172.18.1.236). This IP address must be already defined in /etc/hosts file.

4. Configure Automatically at System Restart ? answer yes.

5. Start New Devices Immediately ? answer yes.

The output is as follows:

```
                                COMMAND STATUS

Command: OK            stdout: yes           stderr: no

Before command completion, additional instructions may appear below.

rpvserver0 Available
```

rpvserver0 is the RPV server for real disk hdisk1 on nodeB.

Note Repeat these steps to create another RPV server for another real disk on nodeB.

To verify the configuration, enter:

```
smit rpvserver
```

1. Select List all Remote Physical Volumes Servers.

```
                                COMMAND STATUS

Command: OK            stdout: yes           stderr: no

Before command completion, additional instructions may appear below.

# RPV Server     Physical Volume Identifier       Physical Volume
# ------------------------------------------------------------
  rpvserver0     00c0062a857f712a                 hdisk1
```

You can also verify the configuration using the following commands:

```
# lsdev -Ccrpvserver
rpvserver0 Available  Remote Physical Volume Server
# lsattr -El rpvserver0
auto_online y                                Configure at System Boot   True
client_addr 172.18.1.236                     Client IP Address          True
rpvs_pvid 00c0062a857f712a0000000000000000 Physical Volume Identifier True
```

At this point, RPV server is available on each node which means that it is ready to process disk I/O requests from remote RPV client. The configuration is as follows:

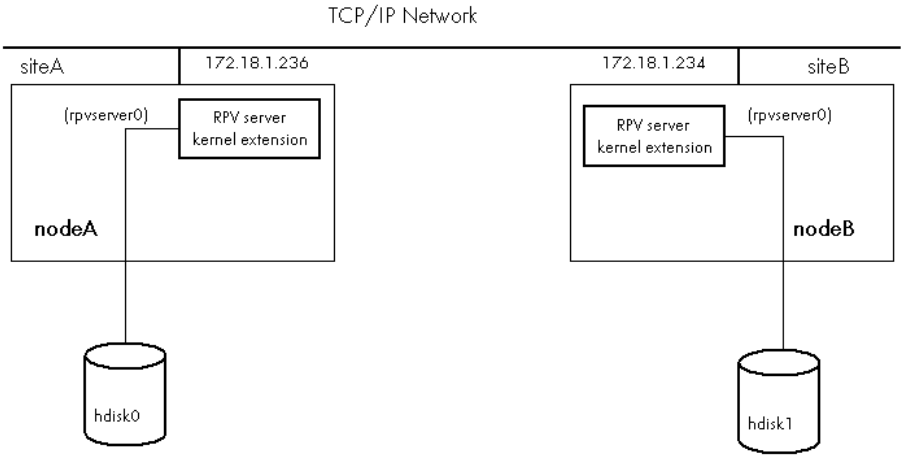


Figure 29. RPV Server Configuration

10.3.2. Configuring RPV Client

RPV client configuration is done on each node.

10.3.2.1. Define RPV client on node A

To define RPV client on nodeA, enter:

```
smitty rpvclient
```

1. Select Add Remote Physical Volume Clients. The following screen appears.

```
                    Add Remote Physical Volume Clients

Type or select a value for the entry field.
Press Enter AFTER making all desired changes.

                                                        [Entry Fields]
* Remote Physical Volume Server Internet Address     [172.18.1.234]        +
```

2. The Remote Physical Volume Server Internet Address is the RPV server's IP address. Enter the nodeB IP address (172.18.1.234). This IP address must be already defined in /etc/hosts file.
3. Then the possible values for the Remote Physical Volume Local Internet Address are determined automatically: This is the RPV client's address from which the RPV server expects to receive connection requests. Select 172.18.1.236, the nodeA IP address.
4. the SMIT menu processing contacts the RPV server in siteB, obtains a list of possible remote physical volumes and presents them in a list:

```
                  Remote Physical Volume Server Disks

  Move cursor to desired item and press F7.
  ONE OR MORE items can be selected.
  Press Enter AFTER making all selections.

    # These remote physical volumes are available
    # at site siteB:

     # (The physical volume names are as they are known
     # on the host named
     # nodeB)

     # Physical Volume        Physical Volume Identifier
     # ------------------------------------------------------
       hdisk1                 00c0062a857f712a0000000000000000
```

5. Select hdisk1. The following screen appears:

```
                    Add Remote Physical Volume Clients

Type or select values in entry fields.
Press Enter AFTER making all desired changes.

                                                      [Entry Fields]
  Remote Physical Volume Server Internet Address      172.18.1.234
  Remote Physical Volume Local Internet Address       172.18.1.236
  Physical Volume Identifiers                         00c0062a857f712a00000>
  I/O Timeout Interval (Seconds)                     [180]
 #
  Start New Devices Immediately?                     [yes]                  +
```

6. The I/O Timeout Interval determines how long the RPV client should wait for the RPV server to respond to I/O requests before giving up and assuming that the RPV server is down or unreachable. After this timeout, the RPV client fails all outstanding I/O requests.
   The LVM treats the remote physical volume as a failed disk and marks it as stale. Setting this value too low can lead to false failures, if a peak in network traffic slows response time. However, setting this value too high may cause the application to wait a very long time when an actual failure occurs.
   The default value is 180 seconds.
7. Start New Devices Immediately? Answer yes.
8. Press Enter to create the RPV client; the RPV client appears as ordinary disk device to the system administrator:

```
                                COMMAND STATUS

Command: OK            stdout: yes           stderr: no

Before command completion, additional instructions may appear below.

hdisk3 Available
```

Device hdisk3 on nodeA is the RPV client for the RPV server (rpvserver0) defined on nodeB (real disk hdisk1 on nodeB).

Note Repeat these steps to create another RPV client for another RPV server defined on nodeB (another real disk on nodeB).

To verify, enter the following commands:

```
smit rpvclient
```

1. Select List all Remote Physical Volumes Clients.

```
                                COMMAND STATUS

Command: OK            stdout: yes           stderr: no

Before command completion, additional instructions may appear below.

# RPV Client     Physical Volume Identifier       Remote Site
# ------------------------------------------------------------
  hdisk3         00c0062a857f712a                 siteB
```

```
lsdev -Ccdisk -t rpvclient
hdisk3 Available  Remote Physical Volume Client
lsattr -El hdisk3
io_timeout  180                          I/O Timeout Interval       True
local_addr  172.18.1.236                 Local IP Address           True
pvid     00c0062a857f712a0000000000000000 Physical Volume Identifier True
server_addr 172.18.1.234
```

10.3.2.2. Define RPV client on node B

To define RPV client on nodeB, enter:

```
# smitty rpvclient
```

1. Select Add Remote Physical Volume Clients. The following screen appears.

```
                    Add Remote Physical Volume Clients

Type or select a value for the entry field.
Press Enter AFTER making all desired changes.

                                                      [Entry Fields]
* Remote Physical Volume Server Internet Address     [172.18.1.234]        +
```

2. Enter the nodeA IP address (172.18.1.236). This IP address must be already defined in /etc/hosts file.
3. Select 172.18.1.234, the nodeB IP address.

4. The SMIT menu processing contacts the RPV server in siteA, obtains a list of possible remote physical volumes and presents them in a list:

```
                  Remote Physical Volume Server Disks

  Move cursor to desired item and press F7.
  ONE OR MORE items can be selected.
  Press Enter AFTER making all selections.

    # These remote physical volumes are available
    # at site siteA:

     # (The physical volume names are as they are known
     # on the host named
     # nodeA)

     # Physical Volume        Physical Volume Identifier
     # ------------------------------------------------------
       hdisk0                 00cb153c8a6b08580000000000000000
```

5. Select hdisk0. The following screen appears:

```
                     Add Remote Physical Volume Clients

Type or select values in entry fields.
Press Enter AFTER making all desired changes.

                                                      [Entry Fields]
  Remote Physical Volume Server Internet Address      172.18.1.236
  Remote Physical Volume Local Internet Address       172.18.1.234
  Physical Volume Identifiers                         00cb153c8a6b085800000>
  I/O Timeout Interval (Seconds)                     [180]
 #
  Start New Devices Immediately?                     [yes]                  +
```

6. Start New Devices Immediately? Answer yes.
7. Press Enter to create the RPV client.

```
                                COMMAND STATUS

Command: OK            stdout: yes           stderr: no

Before command completion, additional instructions may appear below.

hdisk2 Available
```

Device hdisk2 on nodeB is the RPV client for the RPV server (rpvserver0) defined on nodeA (real disk hdisk0 on nodeA).

Note Repeat these steps to create another RPV client for another RPV server defined on nodeA (another real disk on nodeA).

To verify, enter the following commands:

```
# smit rpvclient
```

1. Select List all Remote Physical Volumes Clients.

```
                                COMMAND STATUS

Command: OK            stdout: yes           stderr: no

Before command completion, additional instructions may appear below.

# RPV Client     Physical Volume Identifier       Remote Site
# ------------------------------------------------------------
  hdisk2         00cb153c8a6b0858                 siteA
```

```
# lsdev -Ccdisk -t rpvclient
hdisk2 Available  Remote Physical Volume Client
# lsattr -El hdisk2
io_timeout  180                                 I/O Timeout Interval       True
local_addr  172.18.1.234                        Local IP Address           True
pvid     00cb153c8a6b08580000000000000000 Physical Volume Identifier True
server_addr 172.18.1.236                        Server IP Address          True
```

The final configuration is as follows:

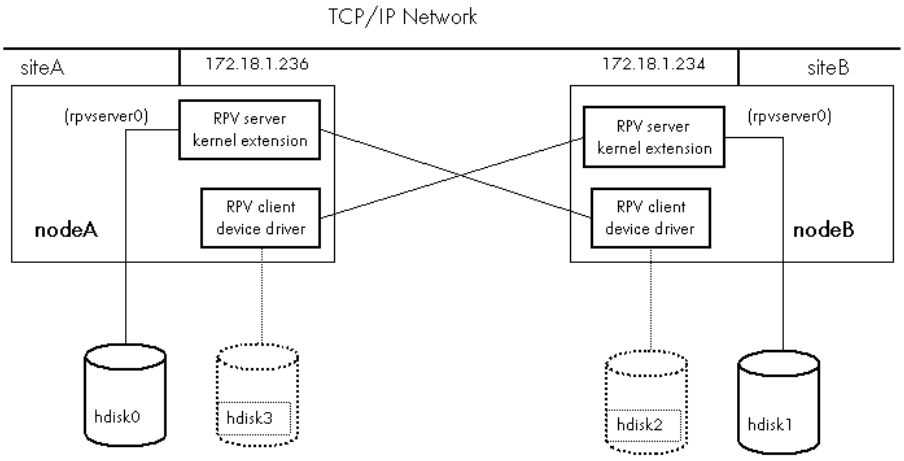


Figure 30. RPV Client and Server Configuration

In this configuration, the real disk (hdisk1) in siteB can now be accessed from siteA (hdisk3) and the real disk (hdisk0) in siteA can be accessed from siteB (hdisk2).

10.3.3. Configuring Volume Group

The disks on the 2 sites are now ready for volume group configuration.

The creation/configuration will be done on one node, then volume group will be imported on other node.

The AIX mirroring feature will be used.

On nodeA

Two physical volumes are available: **hdisk0** (real disk) and **hdisk3** (RPV client disk for **hdisk1** on nodeB).

1. Create a volume group **glvmvg** on these 2 disks:

   ```
   # mkvg  -y glvmvg hdisk0 hdisk3
   ```

2. Change characteristics of the volume group (not activated at system restart and quorum disable):

   ```
   # chvg -a n -Q n glvmvg
   ```

3. Create JFS2 log (**glvmvglog**) in the volume group (**glvmvg**) with the following characteristics:

   - position on physical volume: **center**
   - number of copies of each logical partition: **2** *(for AIX mirroring)*
   - max number of physical volumes: **2** *(for AIX mirroring)*
   - allocate each logical partition copy on a separate physical volume: **superstrict** *(for AIX mirroring)*
   - number of logical partitions:**1**,
   - physical volumes: **hdisk0 hdisk3**:

   ```
   # mklv -y glvmvglog  -t jfs2log -a c -u 2 -c 2 -s s glvmvg 1
   hdisk0 hdisk3
   ```

4. Format the logical volume created to be a jsflog:

   ```
   # logform /dev/glvmvglog (and answer y)
   ```

5. Create JFS2 logical volumes (**glvmvglv1**) for filesystem in the volume group (**glvmvg**) with the following characteristics:

   - position on physical volume: **center**
   - number of copies of each logical partition: **2** *(for AIX mirroring)*
   - max number of physical volumes: **2** *(for AIX mirroring)*
   - allocate each logical partition copy on a separate physical volume: **superstrict** *(for AIX mirroring)*
   - number of logical partitions: **size in LP number** (ex: if LP number=10 and PP size=32MB, logical partition size is 320 MB)
   - physical volumes: **hdisk0 hdisk3**:

   ```
   # mklv -y glvmvglv1  -t jfs2  -a c  -u 2  -c 2 -s s glvmvg 10
   hdisk0 hdisk3
   ```

6. Create the filesystem on the previously defined logical volume (specify mount point and not automatically mounted at system restart):

   ```
   # crfs -v jfs2 -d'glvmvglv1' -m'/glvmfs1' -A no
   ```

7. Check the configuration by mounting and umounting file system:

   ```
   # mount /glvmfs1
   # umount /glvmfs1
   ```

8. Deactivate the volume group:

   ```
   # varyoffvg glvmvg
   ```

On the other node (nodeB)

Two physical volumes are available: **hdisk1** (real disk) and **hdisk2** (RPV client disk for **hdisk0** on nodeA).

1. Check if disks related to the volume group are already seen with a PVID:

   ```
   # lspv
   ```

   If PVID is none enter:

   ```
   # chdev -l hdisk -a pv=yes
   ```

2. Import the volume group definition from one disk:

   ```
   # importvg -y glvmvg hdisk1
   ```

3. Change volume group characteristics:

   ```
   # chvg –a n -Q n glvmvg
   ```

4. Check every thing works by mounting and umounting file system:

   ```
   # mount /glvmfs1
   # umount /glvmfs1
   ```

5. Deactivate the volume group:

   ```
   # varyoffvg glvmvg
   ```

Verify GLVM configuration

To verify GLVM configuration on each node, use **smit glvm_utils** menu:

```
                Geographic Logical Volume Manager Utilities

Move cursor to desired item and press Enter.

  Geographically Mirrored Volume Groups
  Geographically Mirrored Logical Volumes
  Remote Physical Volume Clients
  Remote Physical Volume Servers
```

Select **Geographically Mirrored Volume Groups.**

Select **List Geographical Volume Group Information.**

If volume group **glvmvg** is varied on nodeA:

```
#Volume Group   Logical Volume    RPV          PVID           Site
glvmvg          glvmvglog       hdisk3       00c0062a857f712a siteB
glvmvg          glvmvglv1       hdisk3       00c0062a857f712a siteB
```

If volume group **glvmvg** is varied on nodeB:

```
# Volume Group  Logical Volume    RPV          PVID            Site
glvmvg          glvmvglog       hdisk2       00cb153c8a6b0858  siteA
glvmvg          glvmvglv1       hdisk2       00cb153c8a6b0858  siteA
```

Select **Geographically Mirrored Volume Groups.**

Select **Verify Mirror Copy Site Locations for a Volume Group.**

If volume group **glvmvg** is varied on nodeA:

```
Checking Volume Group glvmvg
#   Site      Copy Physical Volumes
#siteA      PV1 hdisk0
 siteB      PV2 hdisk3
Checking Logical Volume glvmvglog
```

```
Checking Logical Volume glvmvglv1
```

If volume group **glvmvg** is varied on **nodeB**:

```
Checking Volume Group glvmvg.
#   Site      Copy Physical Volumes
siteA      PV1 hdisk2
#siteB  PV2 hdisk1
Checking Logical Volume glvmvglog.
Checking Logical Volume glvmvglv1.
```

10.4. Configuring ARF to Use GLVM

After configuring GLVM on the 2 sites (**nodeA** and **nodeB**), it is necessary to instruct ARF to use GLVM.

1. When defining nodes, enter the node address used for RPV client and server communication in the 2 following fields:
   - Adresses List (to tell the monitoring daemon to monitor the RPV network).
   - Address(es) used for Remote Physical Volumes.

   The address must be defined in **/etc/hosts** file on each node and **root** access must be allowed for each node (**/.rhosts** file if rsh is used)).

   If **ssh** is used, refer to chapter "Installing and Configuring SSH", paragraph "Authorizing additional node addresses for SSH".

   For performance and security reasons, it is the system administrator's responsibility to use a dedicated network and to secure the RPC client-server network between the 2 sites.

```
                              Add a Managed Node

Type or select values in entry fields.
Press Enter AFTER making all desired changes.

                                                      [Entry Fields]
*   Node Name                                        [nodeA]
*   Addresses List                                  [172.16.108.52 172.18.1.236]

  TTY device for PPP hearbeat                        []
  VIO servers address list used for disk IO          []
  Address(es) used for Remote Physical Volumes       [172.18.1.236]
```

2. To make effective the ARF AIX mirroring feature (used by GLVM and configured in the previous paragraph), it is necessary to run the following SMIT menu before activating applications on the nodes:

```
smit barf
  Configuration Definition
    Disaster Recovery Configuration
```

```
                    Disaster Recovery Configuration

Type or select values in entry fields.
Press Enter AFTER making all desired changes.

                                                      [Entry Fields]
  ACTIVATE Disaster Recovery ?                       yes                  +
```

Enter **yes**.

3. Propagate ARF configuration to all nodes using the appropriate SMIT menu.
4. When activating an application, you can indicate if you want to start the application even if GLM copies are staled (**Force start when GLVM copies are staled** field).

```
                              Activate Application

Type or select values in entry fields.
Press Enter AFTER making all desired changes.

                                                      [Entry Fields]
* Application List                                    app1                +
* Starting Node                                       nodeA               +
  Mode of Take-over                                   automatic           +
  List of Take-over nodes                            []                   +
  Start Application ?                                [yes]                +
  Activate Resources ?                               [yes]                +
  Force start when GLVM copies are staled            [no]                 +
```

Chapter 11. Configuring MirrorView for ARF

11.1. MirrorView Overview

MirrorView applies to EMC CLARiiON disk subsystems.

MirrorView controls that enable access to shared disks from both the primary site hosts and the secondary site hosts are automated through *Application Roll-over Facility* pre-event scripts.

To help you to configure the MirrorView functionality with an *Application Roll-over Facility* cluster, an example of the configuration of two nodes is given below:

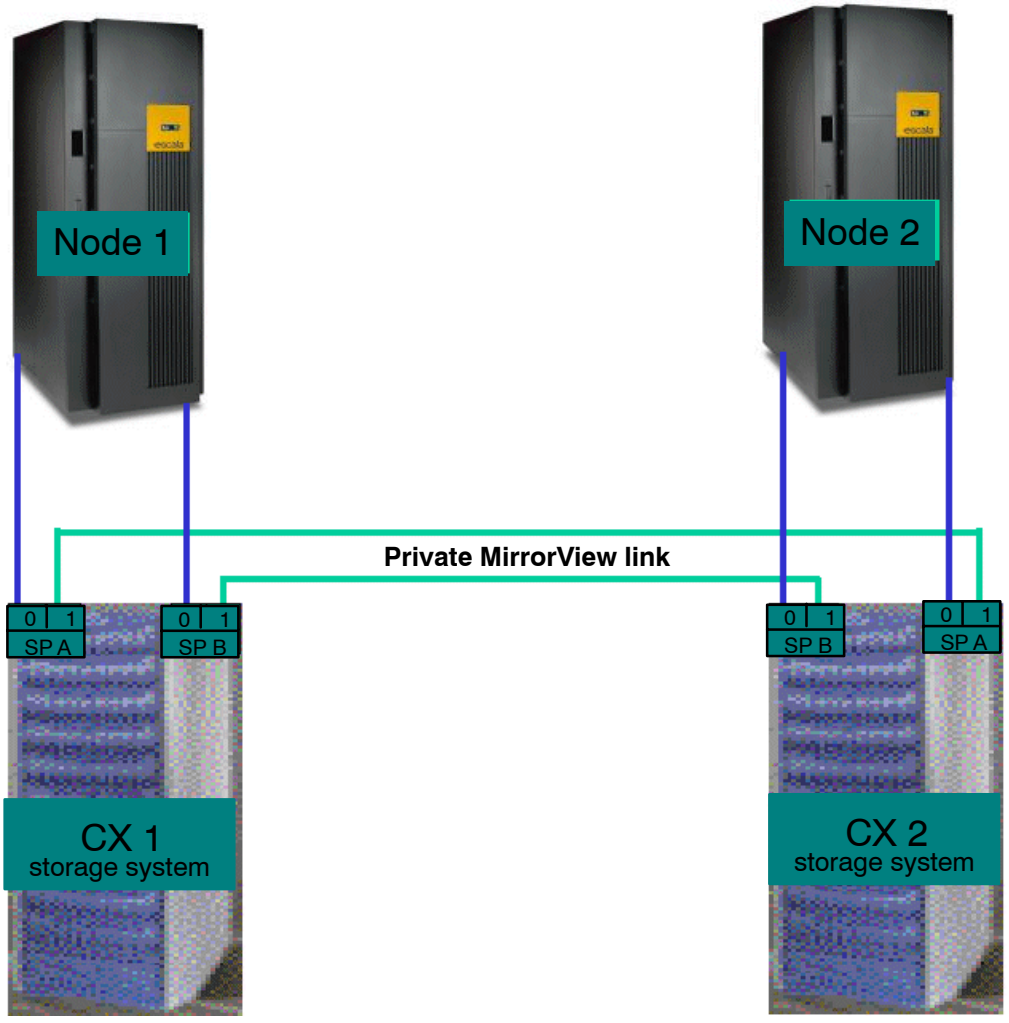


Figure 31. MirrorView Configuration Example

11.2. Initializing MirrorView with Application Roll-over Facility

11.2.1. Initializing Navisphere Use by Cluster Nodes

Link java 1.4

Define a symbolic link so that **/usr/java/bin/java** relates to java 1.4:

```
ln -s /usr/java14 /usr/java
```

Configure a Navisphere security file for root user

For each CX storage system, create a root user (using Navisphere) and make it visible on each node.

For each node, configure a Navisphere Security file for root user, as in the following example:

```
cd /usr/lpp/NAVICLI
java -jar navicli.jar -h cx401_spa -password password -scope 0
-AddUserSecurity
```

Before initializing MirrorView, the CX storage systems must have been configured with Navisphere Manager (initialize Navisphere Agents, create Raid groups, create Storage Group, bind Lun's, etc ...).

For more information, refer to Navisphere Manager or Supervisor manuals.

11.2.2. Initializing MirrorView

For a correct behavior of MirrorView with *Application Roll-over Facility*, it is recommended to create only one Storage Group reserved for an *Application Roll-over Facility* use, for each cluster node. (For the configuration example above, you have to define one Storage Group for cluster `Node 1` and another one for cluster `Node 2`).

1. **Initialize Mirrors:**

   Using Navisphere Manager follow these steps for each CX storage system:

   a. Create and allocate the Write Intent Logs (two private LUN's, one for each SP).

   b. Connect the CX for MirrorView.

2. **Create Mirrors:**

   To configure the *Application Roll-over Facility* cluster, you have to access all the primary images on one cluster node, and then to synchronize, from this cluster node, the *Application Roll-over Facility* configuration on all cluster nodes.

   So, it is recommended to create all the mirrors on one CX storage system belonging to one unique cluster node (`CX 1` belonging to cluster `Node 1` in the given example).

3. **Create the Remote Mirrror.**

4. **Add a secondary image to the Remote mirror**

   with **recovery policy** parameter set to **automatic** and **synchronization rate** parameter set to **high**.

   The mirror changes to the ACTIVE state.

   The secondary image will change from *Out_of_Sync* state to *Synchronizing* state and then to *In_Sync* state.

5. Integrate Primary images in Storage Group:

   This step is mandatory to obtain the hdisk visibility under AIX.

   It is recommended to create all the mirrors on one CX storage system connected to one unique cluster node.

   It is recommended to integrate all the Primary Images in the Storage Group reserved for an *Application Roll-over Facility* use of the CX storage system connected to this node (Storage Group of `CX 1` connected to cluster `Node 1` in the given example).

   Once the storage group created with all the primary images, run the **cfgmgr** command on the cluster node to discover and to access the hdisks.

6. Create Volume Groups:

   On one cluster node (`Node 1` for the given example), using **SMIT**, create the Volume Groups and if needed, the relative Logical Volumes and FileSystems.

7. Import Volume Groups to the other cluster nodes:

   a. Remove Primary images from Storage Group

   It is recommended to remove all the Primary Images in the Storage Group reserved for an *Application Roll-over Facility* use (Storage Group of `CX 1` connected to cluster `Node 1` in the given example).

   b. Promote Secondary Images

   This step makes the Secondary Images becomes the Primary.This step is necessary to discover the relative hdisks and have acces on them.

   c. Integrate Primary images in Storage Group

   This step is mandatory to obtain the hdisk visibility under AIX.

   The Primary Images must be integrated in the Storage Group of the CX storage system belonging to the other node(s) associated for the import (Storage Group of `CX 2` belonging to cluster `Node 2` in the given example).

   To discover and to access the hdisks, execute the **cfgmgr** command on the cluster node for which the CX storage system is connected to (cluster `Node 2` in the given example).

   On the cluster node, for which the CX storage system is connected to, execute the **importvg** command relatively to each volume group of the cluster node (cluster `Node 2` in the given example).

   d. Change the recovery Policy

   This step is necessary if you want to have the automatic synchronization.

   Otherwise, the synchronization will be manual.

8. Configure the *Application Roll-over Facility* cluster:

On one cluster node (`Node 1` for the given example), configure the cluster topology and the cluster resources and synchronize, from this cluster node, the *Application Roll-over Facility* configuration on all cluster nodes.

11.3. Configuring MirrorView Environment

To make MirrorView properly work in an *Application Roll-over Facility* environment you have to perform the following actions on your cluster configuration:

1. Define the storage systems connected to the nodes.
2. Define the mirrored volume groups.
3. Synchronize the MirrorView environment to all nodes.
4. Make active the MirrorView environment in the cluster configuration.

Once the MirrorView environment is active, volume groups management will be automatically taken into account according to the *Application Roll-over Facility* event by promoting secondary images on the appropriate node and by managing Storage groups.

WARNING

In case of damage on one storage system, automatic take over will occur (halt –q on the node connected to it) but recovery (after changing the sub-system) will need manual intervention (see the procedure described in *Recovery from Hardware Problems*, on page 11-7) before restarting *Application Roll-over Facility* on the node which was stopped due to the damage.

The following paragraphs describe how to configure and activate the MirrorView environment in an *Application Roll-over Facility* cluster from a unique node.

1. Define the storage systems connected to a node:

   For each CX sub-system connected to the nodes, use the following smit menus:

```
#smit barf
   Configuration Definition
      Configure MirrorView environment
         Define Storage Systems
             Add a Storage System
```

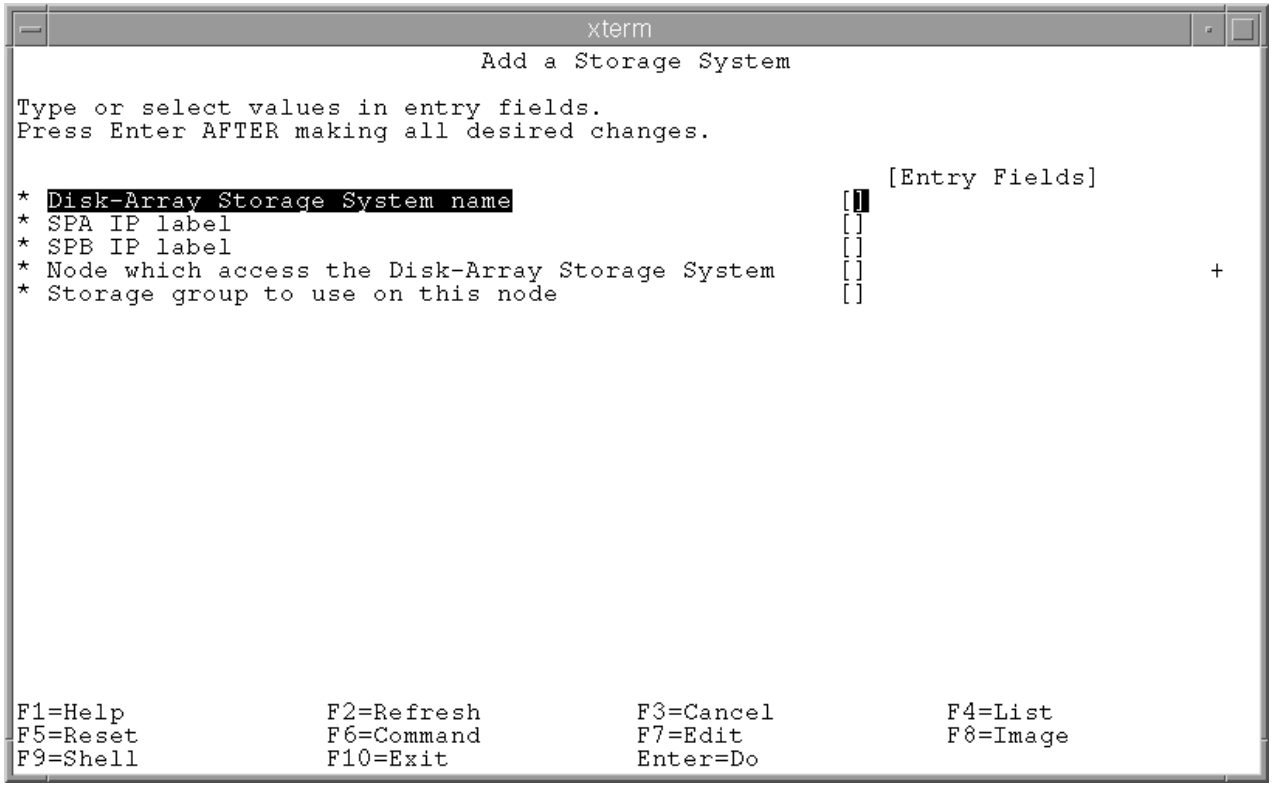


Figure 32. Add a Storage System screen

In the **Disk-Array Storage System name** field, enter a name that will uniquely identify the sub-system.

In the **SPA IP label** field enter the IP label of the first SP as defined in the **/etc/hosts** file.

In the **SPB IP label** field enter the IP label of the second SP as defined in the **/etc/hosts** file.

In the **Node which access the Disk-Array Storage System** field enter the name (as defined in the *Application Roll-over Facility* configuration) of the node that accesses the sub-system.

In the **Storage group to use on this node** field enter the name of the storage group (as defined in the Navisphere configuration) to be used to manage disks access on the node previously defined.

2. **Define the mirrored volume groups:**

   For each mirrored volume group used as resource in the *Application Roll-over Facility* configuration, use the following smit menus:

```
#smit barf
   Configure Definition
      Configure MirrorView environment
         Define mirrored Volume groups
             Add a mirrored Volume group
```

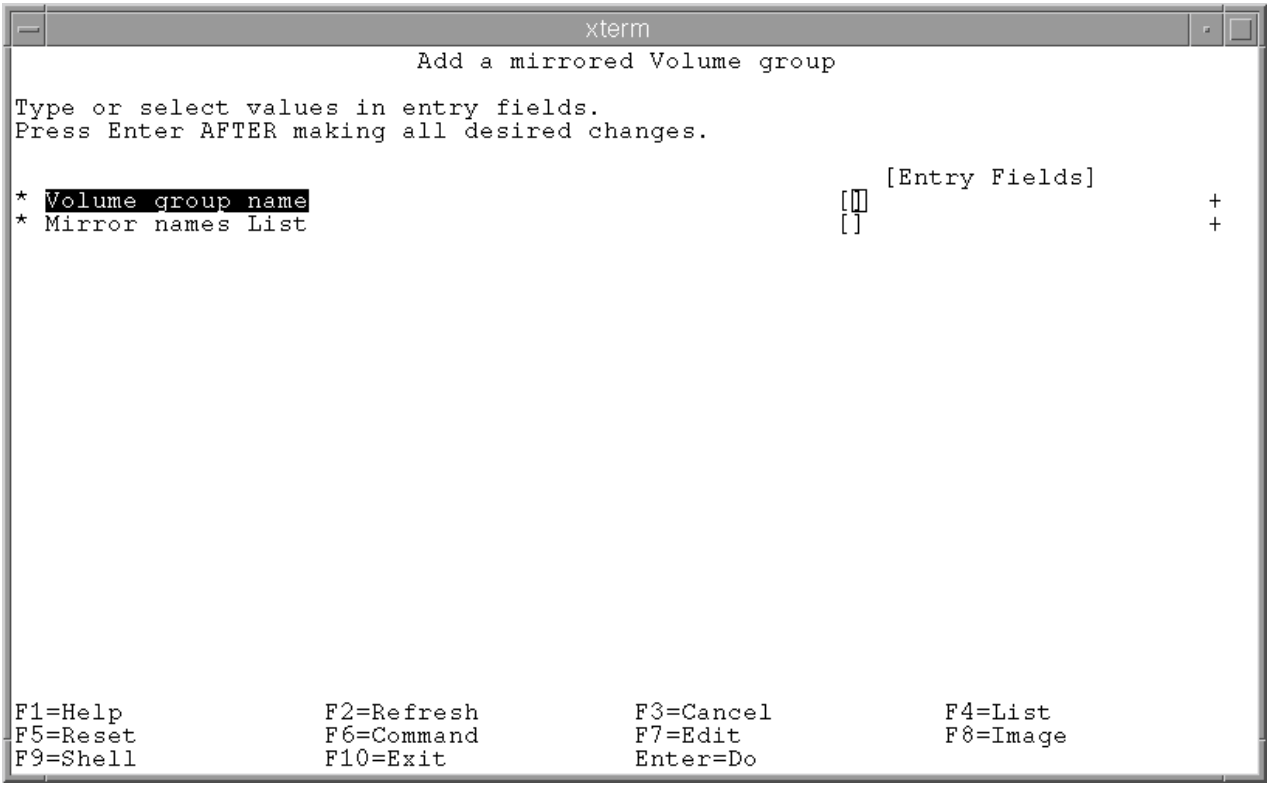


Figure 33. Add a mirrored Volume group screen

In the **Volume group name** field enter the name of a mirrored volume group.

In the **Mirror names List** field enter a coma separated list of the mirrors which compose the volume group (as defined in the Navisphere configuration) or select them using the F4 key.

3. Synchronize the MirrorView environment to all nodes:

   After having defined all storage systems and volume groups used in the cluster configuration on one node, use the following smit menus to synchronize the MirrorView environment to all cluster nodes:

   ```
   #smit barf
      Configuration Definition
         Configure MirrorView environment
            Synchronize MirrorView environment
   ```

4. Make Active the MirrorView environment:

   To make effective the MirrorView environment, use the following smit menus before starting *Application Roll-over Facility* on the cluster nodes:

   ```
   #smit barf
      Configuration Definition
         Configure MirrorView environment
            Activate/Deactivate Mirrorview
   ```

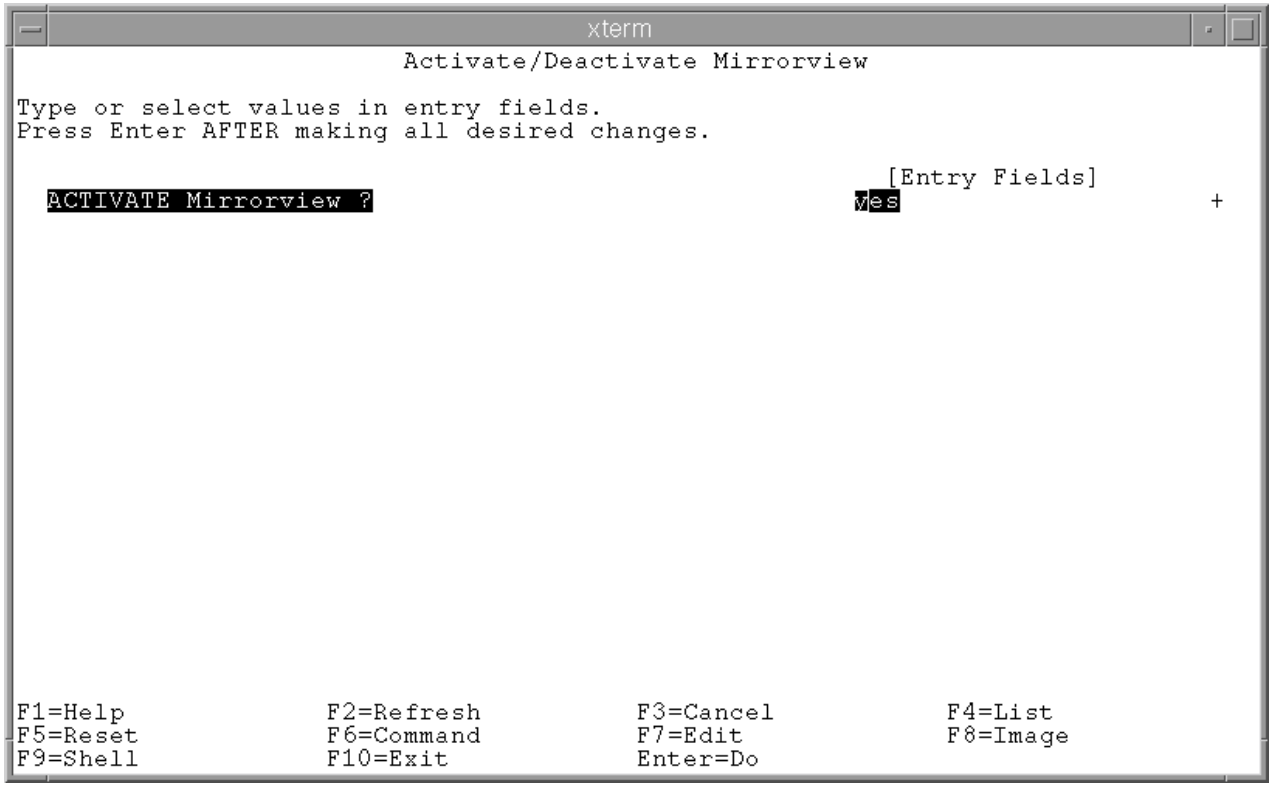


Figure 34. Activate/Deactivate Mirrorview screen

In the ACTIVATE Mirrorview ? field select “yes".

Once the MirrorView environment has been made active, *Application Roll-over Facility* may be started on the nodes.

11.4. Maintaining the MirrorView Environment

In addition to the menus allowing to define storage systems and volume groups used in the MirrorView environment, menus are provided to modify or remove their definitions.

They can be found under:

```
# smit barf
   Configuration Definition
       Configure MirrorView environment
          Define Storage Systems
              Add a Storage System
```

and:

```
# smit barf
   Configuration Definition
      Configure MirrorView environment
         Define mirrored Volume groups
```

If you make some changes to your MirrorView environment, don't forget to synchronize the new environment to all cluster nodes. See *Synchronizing the Application Roll-over Facility Configuration*, on page 6-21 for more information.

11.4.1. Recovery from Hardware Problems

In case of a CX storage system failure, if it becomes again available after repair, the recovery is automatically taken into account using the following menus on the node connected to the CX that was operational:

```
# smit barf
   Configuration Definition
      Configure MirrorView environment
         Rebuild Secundary images
```

But, in case of major natural disasters (for example, water flooding of an Information System, through fire disaster, up to earthquake etc), or for any reason, it could be necessary to replace the failing CX storage system.

The following procedure describes the steps required to restore manually the original MirrorView configuration in case of a CX storage system replacement.

Don't forget that you have to configure the new CX storage system with Navisphere Manager (Initialize Navisphere Agents, create Raid groups, create Storage Group, bind Lun's, etc).

1. Initialize Mirrors for the new CX storage system:

   Using Navisphere Manager, create the Write Intent Log. Two private LUN's (one for each SP) will be bound automatically.

   Remember: Don't forget to bind new LUN's in the Raid Groups.

2. Add the Secondary Images.

11.4.2. Remarks About ARF Behavior Using MirrorView

After having stopped or replaced a CX storage system, you have to start *Application Roll-over Facility* on the node on which the CX is connected to.

Do not start *Application Roll-over Facility* if the state *Synchronizing* is displayed on the Navisphere Manager.

Wait that the state becomes *In_Sync* or *Consistent*.

If the state becomes *Out_Of_Sync*, you have to manually restart the synchronization phase of the mirrors.

Chapter 12. Configuring NetApp MetroCluster for ARF

12.1. NetApp MetroCluster Overview

NetApp MetroCluster is a unique synchronous replication solution protecting your critical data against site disasters. It provides the capability to force a failover when an entire NetApp storage system (including the controllers and storage) is destroyed or unavailable.

There are two configurations:

- Stretch MetroCluster for distances between two sites up to 500m
- Fabric MetroCluster for distances greater than 500m(max 100km) between two sites

In a MetroCluster configuration, each disk shelf on a storage controller has a mirror shelf on its partner.

Stretch or Fabric MetroCluster provides data mirroring and the additional ability to initiate a failover if an entire site becomes lost or unavailable.

Stretch or Fabric MetroCluster contains two complete copies of the specified data volumes or file systems that you indicated as being mirrored volumes. These copies are called plexes and are continually and synchronously updated every time Data ONTAP writes data to the disks. Plexes are physically separated from each other across different groupings of disks.

12.2. MetroCluster in Stretch mode

The Stretch MetroCluster configuration includes the following connections:

- Connections from each controller to the user network.
- The MetroCluster interconnect between the two controllers.
- Connections from each controller to its own storage:
  - Controller A to vol X
  - Controller B to vol Y
- Connections from each controller to its partner's storage:
  - Controller A to vol Y
  - Controller B to vol X
- Connections from each controller to the mirrors of its storage:
  - Controller A to vol X' (X-mirror)
  - Controller B to vol Y' (Y-mirror)

The following figure illustrates the stretch MetroCluster configuration.

Note This simplified figure does not show disk shelf-to-disk shelf connections.

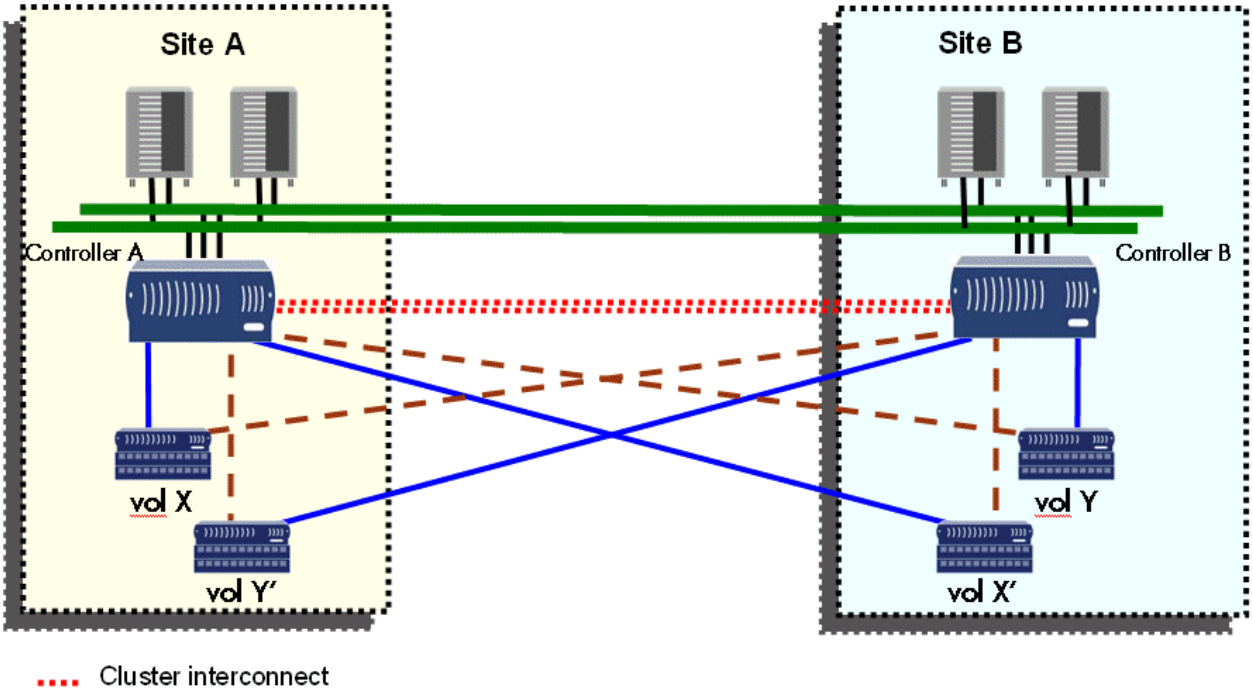


Figure 35. MetroCluster in Stretch Mode

12.3. MetroCluster in Fabric mode

Fabric MetroCluster provides data mirroring and the failover abilities of a stretch MetroCluster at distances greater than 500 meters.

The following figure illustrates the Fabric MetroCluster configuration.

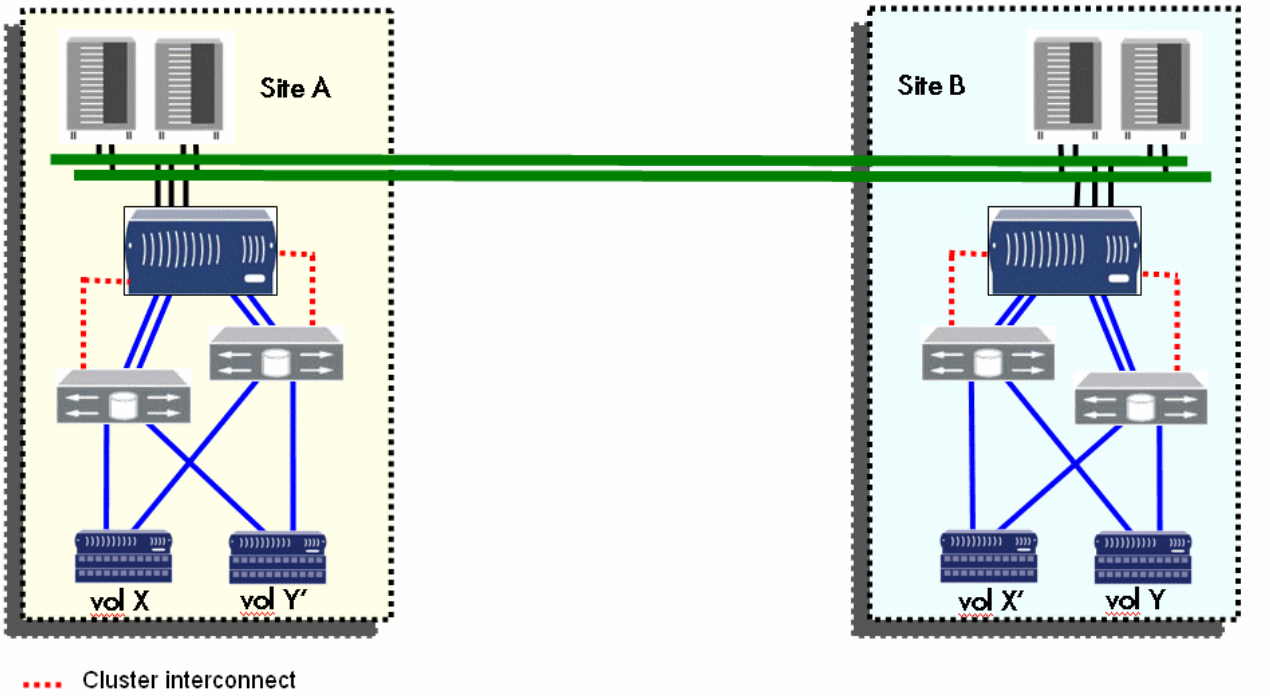


Figure 36. MetroCluster in Fabric Mode

12.4. Configuring NetApp MetroCluster Environnement for ARF

To make NetApp MetroCluster properly work in an Application Roll-over Facility environment, you have to perform the following actions on your cluster configuration:

1. Configure the NetApp Metrocluster Sites
2. Define a MetroCluster User
3. Make active the NetApp Metrocluster environment
4. Synchronize the configuration

1. Configure Netapp MetroCluster Sites

Define every site name corresponding to a storage system (2 controllers max) connected to a list of ARF nodes.

```
#smit barf
   Configuration Definition
     NetApp Metrocluster Configuration
        Define MetroCluster Site
```

- In the **MetroCluster site name** field, enter a name that identifies the site.
- In the **Nodes List** field, enter the ARF nodes list of the site name.
- In the **Filer addresses List** field, enter the list of the IP addresses of the storage system (controller).

2. Define a MetroCluster User

To specify a MetroCluster User configured in the NetApp subsystem use the following menu:

```
#smit barf
   Configuration Definition
     NetApp Metrocluster Configuration
        Define MetroCluster User
```

- In the **User Name to Execute commands on the Filers** field, enter a name authorized to execute the following commands on the filers: **cf** , **lun**, **vol**, **aggr**.

It is necessary to authorize the root user of both ARF nodes to access the NetApp subsystem, via ssh, with the MetroCluster user login defined above. This is configured on the NetApp subsystem.

---

3. Make active the Netapp Metrocluster environment:

To make effective the NetApp Metrocluster environment, use the following smit menus before starting Applications on the cluster nodes:

```
#smit barf
   Configuration Definition
     NetApp Metrocluster Configuration
        Activate/Deactivate MetroCluster
```

- In the **ACTIVATE Metrocluster ?** field select "yes"

4. Synchronize the configuration

To synchronize the configuration on all nodes, use the following smit menus before starting Applications on the cluster nodes:

```
#smit barf
   Configuration Definition
     Propagate Configuration
```

Once the NetApp Metrocluster environment has been made active, Applications may be started on the nodes

12.5. Maintaining the NetApp MetroCluster environnement for ARF

In addition to the menus allowing to define the storage systems used in the NetApp Metrocluster environment, the two following menus are provided to modify or remove their definitions:

```
#smit barf
   Configuration Definition
     NetApp Metrocluster Configuration
       Change/Show a MetroCluster Site Definition
```

and:

```
#smit barf
   Configuration Definition
     NetApp Metrocluster Configuration
       Remove a MetroCluster Site Definition
```

If you make some changes in your configuration NetApp Metrocluster environment, do not forget to synchronize the new environment to all cluster nodes.

12.6. Recovery from Hardware Problems

In case of a NetApp Metrocluster storage system failure, when it becomes again available after repair, the recovery is not performed automatically. You must use the giveback procedure described by NetApp to restore manually the original NetApp Metrocluster Configuration.

12.7. ARF Behavior Using NetApp MetroCluster

The following scenarios of storage system failure may occur:

- Site Failure:

  If the applications are launched by ARF in automatic take over mode, the heartbeat mechanism detects the failed node and the applications are taken over by the node defined as “take over node”.

  ARF decides to make a recovery by running a specific data ONTAP command on the surviving controller. The applications are launched on the take over node and will automatically access to the copy of data of the surviving storage system. The mirroring of data is disable.

- Node Failure:

  If the applications are launched by ARF in automatic take over mode, the heartbeat mechanism detects the failed node and the applications are taken over by the node defined as “take over node”. The take over node will automatically access to the same copy of data as the failed node was.

- Storage system Failure (controller + disk failure):

  The node where the applications are launched by ARF detects an error on the storage system failure. ARF decides to make a recovery by running a specific data ONTAP command on the surviving controller. The applications running on the node will be freeze

during few seconds before the copy of data of the surviving storage system will be available. The mirroring of data is disabled.

- Controller failure or Disk shelf failure:

  The recovery command is automatic. ARF has nothing to do.

- Interconnect failure:

  The mirroring of data is disabled and no failover is done. The two sites will be running independently.

12.8. ARF Recovery Procedure

12.8.1. Giveback procedure

After a Storage system Failure and after repair, you must run the following procedure on the surviving controller, assuming the following:

- Site A is the takeover site – Controller name is FilerA
- Site B is the disaster site – Controller name is FilerB
- Aggregate mirrored: aggr0

On the surviving controller FilerA, run the following steps:

1. Verify that the disaster recovery is done and the mirror is degraded for aggregates:

   On FilerA:

   - Execute the **cf status** command to verify that the surviving controller (FilerA) **has taken over** the failed controller (FilerB).
   - Execute the **aggr status –v** command to verify that the aggregate **aggr0 is online and mirror degraded**.

2. Power on disk shelf without controller on failed site (FilerB) and recreate aggregate mirror:

   On FilerA:

   - Execute the **aggr status –v** command to verify that the aggregate **aggr0 is online and resyncing**.
   - Wait awhile for all aggregates to be **mirrored** and then verify with the **aggr status –v** command.
   - Execute the **partner** command to go on the FilerB hosted by FilerA.
   - Execute the **aggr status –v** to see **aggr0 online and aggr0(1) failed and out-of-date**.
   - Execute the **aggr mirror aggr0 –v aggr0(1)** command to recreate aggregate mirror.
   - Wait awhile for all aggregates to be **mirrored** and verify with the **aggr status aggr0 –r** command.
   - Execute the **partner** command to go on the FilerA.

3. After resynchronization is done, power on the NetApp Controller on site B (FilerB):

   - Execute the **aggr status –v** command to verify that the aggregate **aggr0 is online and mirrored**.
   - Execute the **cf status** command to verify that the NetApp controller on site B(FilerB) is **ready for giveback**.

4. Giveback on the controller of the site A (FilerA):

   On FilerA:

   - Execute the **cf giveback** command.
   - Wait awhile and verify with the **cf status** command that the FilerA is **up and cluster enabled**.

   On the controller of the site B (FilerB):

   - Execute the **cf status** command to verify that the **FilerB is up and cluster enabled**.

12.8.2. Roll-over procedure

Roll-over the ARF Applications as follows:

```
#smit barf
   Manage Application Environment
     Roll-over Application Environment
```

12.8.3. Example of a giveback procedure after repair of a Storage system Failure

We assume the following:

- Site A is the takeover site – Controller name is FilerA
- Site B is the disaster site – Controller name is FilerB
- Aggregate mirrored: aggr0

1. Verify that the disaster recovery is done and the mirror is degraded for aggregates:

   FilerA (takeover)> **cf status**

   FilerA **has taken over** FilerB

   FilerA(takeover)>**partner**

   FilerB/FilerA> **cf status**

   FilerB **has been taken over by** FilerA.

   FilerA(takeover) > **aggr status –v**

```
  Aggr State      Status            Options
aggr0 online      raid4, aggr       root, diskroot, nosnap=off,
                  mirror degraded   raidtype=raid4, raidsize=8,
                                    ignore_inconsistent=off,
                                    snapmirrored=off,
                                    resyncsnaptime=60,
                                    fs_size_fixed=off,
                                    snapshot_autodelete=on,
                                    lost_write_protect=on

      Volumes: vol0, arfm3, arfm4
      Plex /aggr0/plex0: online, normal, active
          RAID group /aggr0/plex0/rg0: normal
      Plex /aggr0/plex2: offline, failed, inactive
```

FilerA(takeover)>**partner**

FilerB/FilerA> **aggr status -v**

```
 Aggr State      Status            Options
aggr0 online     raid4, aggr       root, diskroot, nosnap=off,
                                   raidtype=raid4, raidsize=8,
                                   ignore_inconsistent=off,
                                   snapmirrored=off,
                                   resyncsnaptime=60,
                                   fs_size_fixed=off,
                                   snapshot_autodelete=on,
                                   lost_write_protect=on

        Volumes: vol0, arfm1, arfm2

        Plex /aggr0/plex2: online, normal, active
            RAID group /aggr0/plex2/rg0: normal
```

2. Power on disk shelf without controller on site B and Recreate aggregate mirror:

   FilerA(takeover)> **aggr status –v**

```
 Aggr State      Status            Options
aggr0 online     raid4, aggr       root, diskroot, nosnap=off,
                 resyncing         raidtype=raid4, raidsize=8,
                                   ignore_inconsistent=off,
                                   snapmirrored=off,
                                   resyncsnaptime=60,
                                   fs_size_fixed=off,
                                   snapshot_autodelete=on,
                                   lost_write_protect=on

        Volumes: vol0, arfm3, arfm4

        Plex /aggr0/plex0: online, normal, active
            RAID group /aggr0/plex0/rg0: normal

        Plex /aggr0/plex2: online, normal, resyncing
            RAID group /aggr0/plex2/rg0: recomputing parity 0% completed
```

Wait awhile for all aggregates to be mirrored and then verify all aggregates.

FilerA(takeover)> **aggr status –v**

```
 Aggr State      Status            Options
aggr0 online     raid4, aggr       root, diskroot, nosnap=off,
                 mirrored          raidtype=raid4, raidsize=8,
                                   ignore_inconsistent=off,
                                   snapmirrored=off,
                                   resyncsnaptime=60,
                                   fs_size_fixed=off,
                                   snapshot_autodelete=on,
                                   lost_write_protect=on

        Volumes: vol0, arfm3, arfm4

        Plex /aggr0/plex0: online, normal, active
            RAID group /aggr0/plex0/rg0: normal

        Plex /aggr0/plex2: online, normal, active
            RAID group /aggr0/plex2/rg0: normal
```

Recreate aggregate mirror for each one:

FilerA(takeover)> **partner**

FilerB/FilerA> **aggr status -v**

```
Aggr State      Status            Options
aggr0 online    raid4, aggr       root, diskroot, nosnap=off,
                                  raidtype=raid4, raidsize=8,
                                  ignore_inconsistent=off,
                                  snapmirrored=off,
                                  resyncsnaptime=60,
                                  fs_size_fixed=off,
                                  snapshot_autodelete=on,
                                  lost_write_protect=on

        Volumes: vol0, arfm1, arfm2

        Plex /aggr0/plex2: online, normal, active
            RAID group /aggr0/plex2/rg0: normal

aggr0(1) failed     raid4, aggr       root, diskroot, raidtype=raid4,
                    out-of-date       raidsize=8, resyncsnaptime=60
        Volumes: <none>

        Plex /aggr0(1)/plex2: offline, failed, out-of-date

        Plex /aggr0(1)/plex6: offline, normal, out-of-date
            RAID group /aggr0(1)/plex6/rg0: normal
```

FilerB/FilerA> aggr mirror aggr0 –v aggr0(1)

```
This will destroy the contents of aggr0(1).  Are you sure? yes
```

FilerB/FilerA > aggr status aggr0 -r

```
Aggregate aggr0 (online, raid4, resyncing) (block checksums)
  Plex /aggr0/plex2 (online, normal, active, pool1)
    RAID group /aggr0/plex2/rg0 (normal)

      RAID Disk Device  HA  SHELF BAY CHAN Pool Type  RPM  Used (MB/blks)    Phys (MB/blks)
      --------- ------  ------------ ---- ---- ---- ----- --------------    --------------
      parity    0a.17   0a    1   1   FC:A   1  FCAL 15000 136000/278528000 137104/280790184
      data      0a.26   0a    1   10  FC:A  1  FCAL 15000 136000/278528000 137104/280790184

  Plex /aggr0/plex7 (online, normal, resyncing 10% completed, pool0)
    RAID group /aggr0/plex7/rg0 (normal)

      RAID Disk Device  HA  SHELF BAY CHAN Pool Type  RPM  Used (MB/blks)    Phys (MB/blks)
      --------- ------  ------------ ---- ---- ---- ----- --------------    --------------
      parity    0b.17   0b    1   1   FC:B   0  FCAL 15000 136000/278528000  137104/280790184
      data      0b.18   0b    1   2   FC:B   0  FCAL 15000 136000/278528000  137104/280790184
```

Wait a while for all aggregates to be synchronized:

Wait for this message on FilerA:

```
  [FilerA (takeover): raid.rg.resync.done:notice]: partner:/aggr0/plex7/rg0: resyn-
chronization completed in 14:22.87
 [FilerA (takeover): raid.mirror.resync.done:notice]: /aggr0: resynchronization
completed in 14:22.99
```

Verify aggregate:

FilerB/FilerA> aggr status aggr0 -r

```
Aggregate aggr0 (online, raid4, mirrored) (block checksums)
  Plex /aggr0/plex2 (online, normal, active, pool1)
    RAID group /aggr0/plex2/rg0 (normal)

      RAID Disk Device  HA  SHELF BAY CHAN Pool Type  RPM  Used (MB/blks)    Phys (MB/blks)
      --------- ------  ------------ ---- ---- ---- ----- --------------    --------------
      parity    0a.17   0a    1   1   FC:A   1  FCAL 15000 136000/278528000 137104/280790184
      data      0a.26   0a    1   10  FC:A   1  FCAL 15000 136000/278528000 137104/280790184

  Plex /aggr0/plex7 (online, normal, active, pool0)
    RAID group /aggr0/plex7/rg0 (normal)

      RAID Disk Device  HA  SHELF BAY CHAN Pool Type  RPM  Used (MB/blks)    Phys (MB/blks)
      --------- ------  ------------ ---- ---- ---- ----- --------------    --------------
      parity    0b.17   0b    1   1   FC:B   0  FCAL 15000 136000/278528000 137104/280790184
      data      0b.18   0b    1   2   FC:B   0  FCAL 15000 136000/278528000 137104/280790184
```

3. After resynchronization is done, power on the NetApp controller on site B :

   FilerA(takeover)> aggr status –v

```
 Aggr State      Status            Options
aggr0 online     raid4, aggr       root, diskroot, nosnap=off,
                 mirrored          raidtype=raid4, raidsize=8,
                                   ignore_inconsistent=off,
                                   snapmirrored=off,
                                   resyncsnaptime=60,
                                   fs_size_fixed=off,
                                   snapshot_autodelete=on,
                                   lost_write_protect=on

        Volumes: vol0, arfm3, arfm4

        Plex /aggr0/plex0: online, normal, active
            RAID group /aggr0/plex0/rg0: normal

        Plex /aggr0/plex2: online, normal, active
            RAID group /aggr0/plex2/rg0: normal
```

   FilerA(takeover)> partner

   FilerB/FilerA> aggr status -v

```
 Aggr State      Status            Options
aggr0 online     raid4, aggr       root, diskroot, nosnap=off,
                 mirrored          raidtype=raid4, raidsize=8,
                                   ignore_inconsistent=off,
                                   snapmirrored=off,
                                   resyncsnaptime=60,
                                   fs_size_fixed=off,
                                   snapshot_autodelete=on,
                                   lost_write_protect=on

        Volumes: vol0, arfm1, arfm2

        Plex /aggr0/plex2: online, normal, active
            RAID group /aggr0/plex2/rg0: normal

        Plex /aggr0/plex7: online, normal, active
            RAID group /aggr0/plex7/rg0: normal
```

   Wait for the message: "Waiting for giveback" on NetApp controller on site B

4. On site A, execute the command cf giveback:

FilerA (takeover)> cf status

```
FilerA has taken over FilerB.
FilerB is ready for giveback.
```

FilerA (takeover)>cf giveback

```
please make sure you have rejoined your aggregates before giveback.
Do you wish to continue [y/n] ? y
```

FilerA>cf status

```
Cluster enabled, FilerB is up.
Negotiated failover enabled (network_interface).
```

FilerB>cf status

```
Cluster enabled, FilerA is up.
Negotiated failover enabled (network_interface).
```

Chapter 13. Configuring RDR/RDR-CO for ARF

13.1. RDR/RDR-CO Overview

RDR/RDR-CO applies to StoreWay FDA materials. Using RDR/RDR-CO in an ARF configuration enables to automate resources and applications failover from a production site to a backup site either automatically in case of:

- disaster
- node failure
- Disk Array failure

or manually by using ARF menus (i.e for maintenance reasons).

The following scheme shows an example of a disaster recovery configuration with one node and one FDA per site.

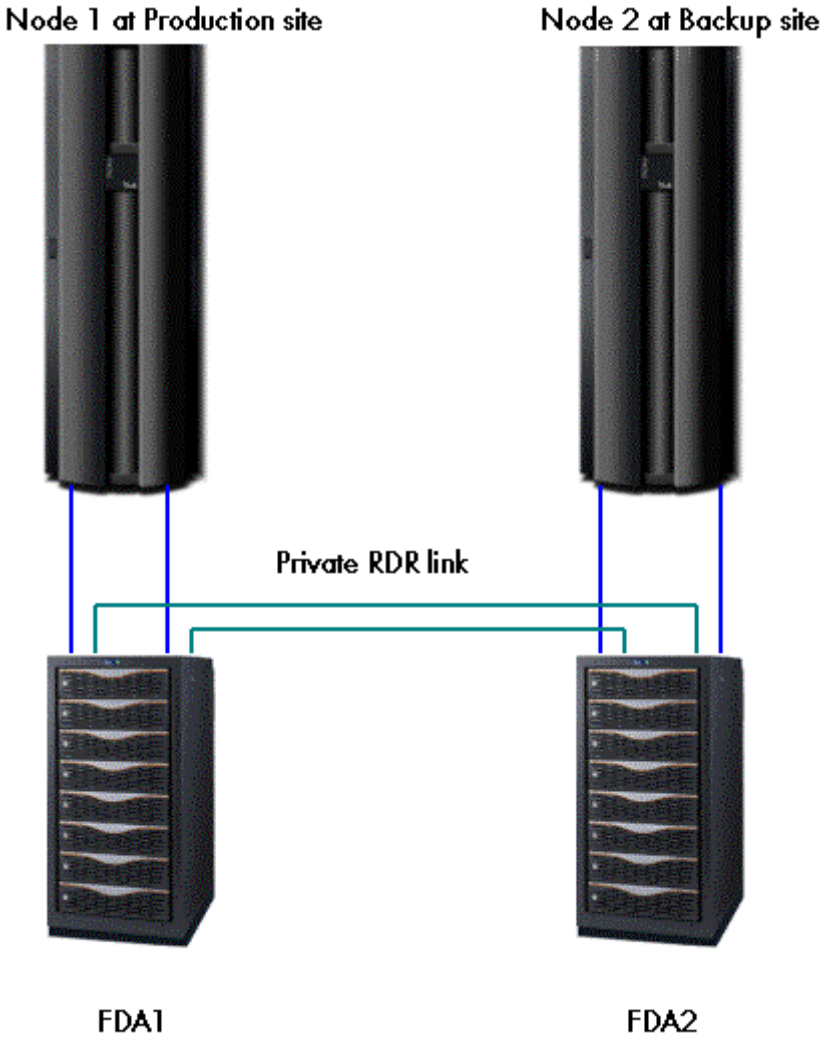


Figure 37. Disaster Recovery Configuration

13.2. Principle

With RDR or RDR-CO, luns on the production site called Master Volumes (MV) are paired with luns on the backup site called Replicate Volumes (RV). In normal situation, datas on MVs are replicated on paired RVs. In that case, RVs are not reachable. For enabling the RVs to be reachable and writable, MVs and RVs must be separated.

With RDR-CO, several pairs can be grouped to insure data consistency between these pairs. These groups are called Atomic Groups (ATG).

Three modes can be used to replicate data:

- **synchronous mode.** In this mode, each IO on a MV is immediately copied to the paired RV. Response time depends on the line speed betweens FDAs and on the distance between the two sites.
- **semi synchronous mode.** In this mode IOs on the MVs are not immediately synchronized on the RVs, but the order of writes is guaranteed on the RVs preserving data consistency.
- **background mode.** IOs on the RVs are done asynchronously. In this mode, there is no guaranty of data consistency.

In a RDR configuration, it is mandatory to create Control Volumes (one Control Volume by Disk Array and by node), used for replication.

The control volume is a lun of minimum size (200 Mb) to add in LD Set of the node. This means that it will be seen as a hdisk on the server but it cannot be used in a volume group.

13.3. Initializing RDR/RDR-CO with ARF

On the FDAs, using iSM manager tool :

1. Create luns for datas and for Control volumes.
2. Create LD Sets (one for each server which accesses the FDA).
3. For each LD Set, add the luns that can be accessed by each server.
4. Configure FDAs for using replication.
5. Create the pairs of MV/RV.
6. If needed, create ATGs for using with RDR-CO (not available on all FDA models).

On the AIX servers:

1. Run **cfgmgr**.
2. Install **fdacli.rte** software from *Storeway FDA software replication products* CD.
3. Run **iSMfill_tgtvol** command.
4. Create the **/etc/iSMrpl/ctlvol.conf** file to specify the control volumes.
5. Run **iSMvollist –r** command.
6. Run **iSMvollist –ax** command to check that the disks indicated in the **/etc/iSMrpl/ctlvol.conf** file are correct. If not, modify this file.

13.4. Creating Volume Group

Preambule:

- If ATG cannot be used, use only one pair per volume group for integrity reason,.
- If ATG can be used, create one volume group per ATG. You must use the same name for the Atomic Group and the Volume Group.

On the production site (where MVs are accessible) :

- Once the pair (RDR) or ATG (RDR-CO) are in replicate state, create the volume group, logical volumes and file systems.
- Once the volume group is created, separate the pair or ATG. This makes RVs writable on the backup site.

On the backup site:

- Import the volume group

Note If the PVId of the hdisk is `none', create it and run the commands **iSMvollist –r.**

On the production site, replicate again the pair or ATG.

13.5. Configuring the Application Roll-over Facility Cluster

On one cluster node, configure the cluster topology and the Applications environment, then allow use of RDR using the menu:

```
# smit barf
     Configuration Definition
         FDA Storage Data Replication Configuration
               Use FDA Storage Data Replication ?          yes
```

Then propagate configuration to other cluster nodes.

13.6. Behavior of ARF with RDR/RDR-CO

With use of RDR/RDR-CO configured, ARF allows only to start an application with pairs in a rpl/sync state or ATG in Atomic (Rpl/sync).

If for some reason, you need to start an application with pair or ATG in another state (i.e separate), then unconfigure the use of RDR/RDR-CO before starting the application. In that case, data integrity is under user responsability. Once problems have been solved, re-configure the use of RDR/RDR-CO.

13.7. Using CLI on the AIX servers

The commands that are listed below can be usefull to display information about the FDA connected to a node using AIX CLI commands (man pages exist for all these commands).

iSMvollist — gives information about FDAs connected to the server from which the command is launched

iSMvollist –d — displays the list of the disk arrays and the number of logical disks in each disk array registered in the Volume List.

iSMvollist –ax — gives the correspondence between the LUNs and the hdisks for each FDA connected to the server.

iSMrc_sense — acquires and displays the specific volume name or AT group (also gives the correspondence between the LUN and the hdisk for a given LUN, vg or AT group)

```
iSMrc_sense –vol ldname –volflg ld
iSMrc_sense –vol vgname –volflg vg
iSMrc_sense –atg atgname for atomic groups
```

iSMrc_arrayinfo — acquires and displays information about the specified disk array's replication function (gives the paths state for each link between the FDA connected to the server and the replicated FDA).

```
iSMrc_arrayinfo –arrayname fdaname –linfo
```

The different possible states can be : Ready, Link Check, Fault or Offline

iSMrc_query — acquires and displays the copy state of the specified pairs. In case of AT group, it acquires and displays the AT group and all the RDR pairs belonging to specified AT group (the Actvity state, the RV Access (not ready or read write) and the copy difference (in KB) for a given group volume (vg), lun (ld) or atomic group (atg).

```
iSMrc_query –mv vgname –mvflg vg
iSMrc_query –mv ldname –mvflg ld
iSMrc_query –atg atgname for atomic groups
```

Chapter 14. Configuring SRDF for ARF

14.1. SRDF Overview

The Symmetrix Remote Data Facility (SRDF®) is a business continuance solution that maintains a mirror image of data at the device level in Symmetrix® arrays located in physically separate sites. The Solutions Enabler SRDF component extends the basic SYMCLI command set to include SRDF commands that allow you to perform control operations on remotely located RDF devices. SRDF provides a recovery solution for component or site failures between remotely mirrored devices, as shown in the following figure. SRDF mirroring reduces backup and recovery costs and significantly reduces recovery time after a disaster.

For more information, refer to *EMC Solutions Enabler Symmetrix SRDF Family CLI* documentation.

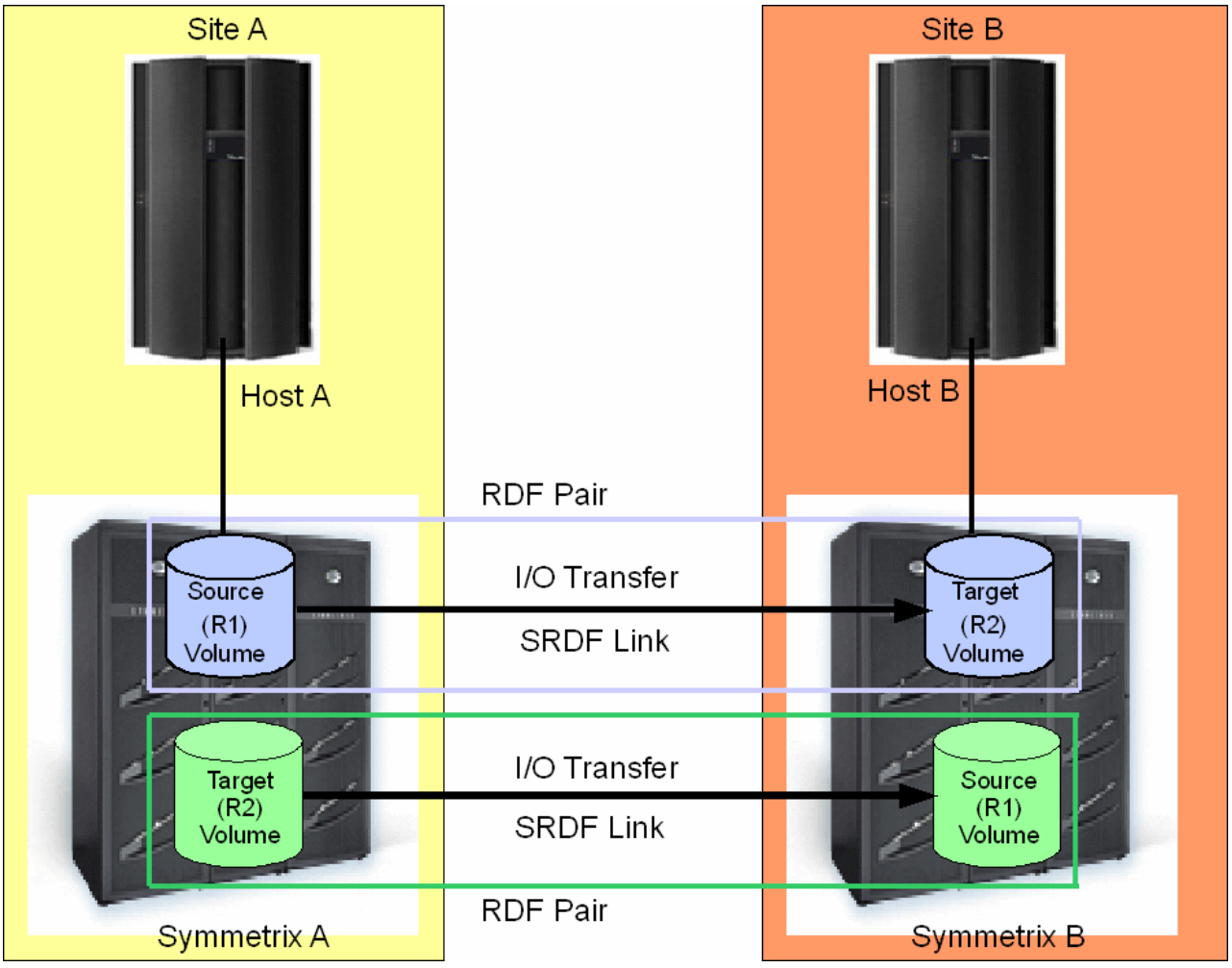


Figure 38. SRDF bidirectional configuration

14.2. Initializing SRDF with Application Roll-over Facility

Before initializing SRDF, the Symmetrix storage systems must have been configured (create Lun's, bind Lun's etc).

14.2.1. Installing Symmetrix Command line lpp

If you plan to use the Fibre adapter on the VIO Server , install the Symmetrix Command line lpp (SYMCLI.SYMCLI.rte) on the VIO Server.

If you plan to use the Fibre adapter directly on the Node partition, install the Symmetrix Command line lpp (SYMCLI.SYMCLI.rte) on the Node.

14.2.2. Finding SRDF devices

Configuration and status information can be viewed for each device on every Symmetrix array containing SRDF devices.

Using SYMCLI, you can find all SRDF devices on a Symmetrix array and view their physical (host) and Symmetrix device names. In addition, you can display details about the SRDF devices, the number of invalid tracks for both the SRDF source device and the target device, and the various SRDF device states.

You can find all Symmetrix arrays that are reachable through the SRDF links. For example, to view how Symmetrix arrays are attached to your host, enter:

```
symcfg list
```

14.2.3. Listing SRDF devices

The **symrdf list** command lists the SRDF devices that are visible to your host, or SRDF devices that are configured on a given Symmetrix array.

For example, to list the SRDF devices that are visible to your host, enter:

```
symrdf list pd
```

The results provide details about the SRDF devices, source (R1) and target (R2), and Cascaded RDF devices (R21) .

14.2.4. Creating RDF group

To configure SRDF devices, use the basic SYMCLI command set. For details see the man pages for all these commands.

Initially, you must explicitly create an empty RDF group that can be populated with devices.

- To create Device Group, enter:

  ```
  symdg create <DgName> [-type REGULAR | RDF1 | RDF2 | RDF21]
  ```

- To create Consistency Group, enter

  ```
  symcg create <CgName> [-type REGULAR | RDF1 | RDF2 | RDF21]
  ```

Note In these commands the type can be specified either in the form “RDFx” or “Rx”.

14.2.4.1. Adding SRDF devices in Device Group

```
symld -g <DgName> add dev <SymDevName>
```

Note The Device Group must be in Read/Write access.

14.2.4.2. Adding SRDF devices in Consistency Group

```
symcg -cg <CgName> add dev <SymDevName>
```

Note The Consistency Group must be in Read/Write access.

14.2.5. Creating Volume Group

Using SMIT, create the Volume Groups and if needed, the relative Logical Volumes and FileSystems.

14.3. Configuration example

We assume the following:

- Primary Site A (Local Site):

  Node A – Symmetrix A

- Secondary Site B (Remote Site):

  Node B Symmetrix B

1. On the Node A:

   a. List SRDF devices:

   ```
   symrdf list pd
   ```

   From the result, retrieve the list of the devices:

   ```
   Sym Dev     Rdev        Type        State
      0124     0144        R1:1         RW
      012B     0151        R2:1         WD
      01EA     0444        R1:1         RW
      01F1     0451        R2:1         WD
   ```

   b. Configure RDF Groups:

   ```
   symdg create DG_VG1 -type RDF1
   symcg create CG_VG2 -type RDF2
   ```

   c. Add devices in RDF Groups

   . For Device Group DG_VG1:

   ```
   symld -g DG_VG1 add dev 0124
   symld -g DG_VG1 add dev 01EA
   ```

   . For Consistency Group CG_VG2:

   ```
   symcg -cg CG_VG2 add dev 012B
   symcg -cg CG_VG2 add dev 01F1
   ```

2. On the Node B:

   a. List SRDF devices:

   ```
   symrdf list pd
   ```

   b. Configure RDF Groups:

   ```
   symdg create DG_VG1 -type RDF2
   symcg create CG_VG2 -type RDF1
   ```

c. Add devices in RDF Groups

. For Device Group DG_VG1:

```
symld -g DG_VG1 add dev 0144
symld -g DG_VG1 add dev 0444
```

. For Consistency Group CG_VG2:

```
symcg -cg CG_VG2 add dev 0151
symcg -cg CG_VG2 add dev 0451
```

3. On the Node A, run the following SRDF failover command to have all RDF groups in RW access mode:

```
symrdf -cg CG_VG2 -noprompt failover
```

4. On one cluster node (Node A for the given example), using SMIT, create the Volume Groups and if needed, the relative Logical Volumes and FileSystems.
5. On the node B, run the following SRDF failover commands to have all RDF groups in RW access mode:

```
symrdf -g DG_VG1 -noprompt failover
symrdf -cg CG_CG2 -noprompt failback
```

6. Import Volume Groups to the other cluster nodes (Node B in the given example) then execute the following SRDF failover command to return to the initial configuration:

```
symrdf -g DG_VG1 -noprompt failback
```

7. Check on all cluster nodes that the RDF Groups are in Synchronized state.

14.3.1. Configure the Application Roll-over Facility cluster

On one cluster node (Node A for the given example), configure the cluster topology and the cluster resources and synchronize, from this cluster node, the Application Roll-over Facility configuration on all cluster nodes.

14.4. Configuring SRDF Environment

To make SRDF properly work in an Application Roll-over Facility environment you have to perform the following actions on your cluster configuration:

1. Make Active the SRDF environment
2. Add RDF Groups
3. Synchronize the configuration

1. Make Active the SRDF environment:

To make effective the SRDF environment, use the following smit menus before starting Application Roll-over Facility on the cluster nodes:

```
#smit barf
      Configuration Definition
            EMC SRDF Configuration
                  Activate/Deactivate SRDF
```

Figure 39. Activate/Deactivate SRDF screen

Set the ACTIVATE SRDF ? field to "yes".

2. Add a RDF Group

For each mirrored volume group used as resource in the Application Roll-over Facility configuration, use the following smit menus:

```
#smit barf
      Configuration Definition
            EMC SRDF Configuration
                  Add a RDF Group
```

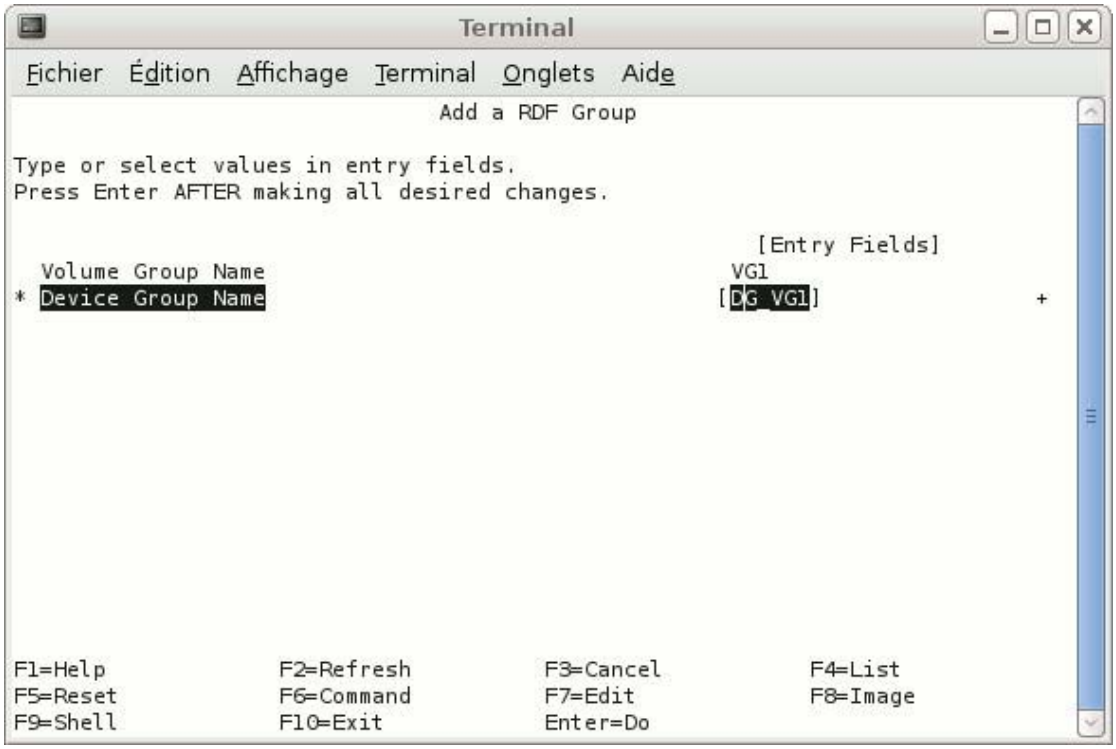


Figure 40. Add a RDF Group

You can add one RDF Group for one volume group:

1. Select a Volume Group previously configured.
2. In the RDF Group type, select either:
   - Device Group
   - Consistency Group
3. Select the name of the Device or Consistency Group.

3. Synchronize the configuration

To synchronize the configuration on all nodes, use the following smit menus before starting Applications on the cluster nodes:

```
#smit barf
      Configuration Definition
            Propagate Configuration
```

Then, applications may be started on the nodes .

14.5. Maintaining the SRDF Environment

In addition to the menus allowing to add RDF Groups used in the SRDF environment, menus are provided to modify or remove their definitions. These menus can be accessed as follows:

```
#smit barf
      Configuration Definition
            EMC SRDF Configuration
                  Change/Show a RDF Group definition
```

and:

```
#smit barf
Configuration Definition
      EMC SRDF Configuration
            Remove a RDF Group
```

If you make some changes to your SRDF environment, do not forget to synchronize the new environment to all cluster nodes. See *Synchronizing the Application Roll-over Facility Configuration*, on page 6-21.

14.6. ARF Recovery Procedure

14.6.1. Giveback procedure

After a site failure (site or storage), the copies of the data are not synchronized, and the data of R2 could be more up-to-date than R1 copies.
If the interruption between the two sites was short, ARF can update the data during the application Roll-over, providing that the operation will not take more than 5 min.
If you think that the update could take more than 5 min, it is recommended to proceed manually.

The following procedure has to be launched on the repaired server (Fibre channel owner) and after storage and link have been repaired in the case of a storage failure:

- Check the state of the R1 copy
- If necessary change R1 copy to the write-disable mode,
- Update the R1 copy with the data of the R2 copy.

These steps are detailed below.

1. Check if a R1 copy in Read-Write mode exists on the storage system where the disaster occurred:

```
# symrdf list pd

Symmetrix ID: 000190300357
                        Local Device View
-------------------------------------------------------------------------
                     STATUS     MODES                      RDF  S T A T E S
Sym         RDF     ---------  -----  R1 Inv   R2 Inv -------------------
Dev  RDev  Typ:G  SA RA LNK  MDAT  Tracks  Tracks Dev RDev Pair
---- ---- -------  ---------  ----- -------  ------- --- ---- ---------

0124 0144   R1:1  RW RW NR   S..1    2102     1      RW  RW   Split
012B 0151   R2:1  RW WD RW   S..2       0     0      WD  RW   Synchronized
01EA 0444   R1:1  RW RW NR   S..1      30     0      RW  RW   Split
01F1 0451   R2:1  RW WD RW   S..2       0     0      WD  RW   Synchronized

Total            -------- --------
  Track(s)              0        1
  MB(s)               0.0      0.1
```

```
Legend for MODES:
M(ode of Operation): A = Async, S = Sync, E = Semi-sync, C = Adaptive Copy
D(omino)           : X = Enabled, . = Disabled
A(daptive Copy)    : D = Disk Mode, W = WP Mode, . = ACp off
(Mirror) T(ype)    : 1 = R1, 2 = R2
```

```
# symdg list
                D E V I C E      G R O U P S
                                   Number of
  Name      Type      Valid   Symmetrix ID  Devs   GKs  BCVs  VDEVs  TGTs
  DG_VG1    RDF1      Yes     000190300357     2     0     0      0     0
```

2. Change R1 copy to write-disable (on the disaster site storage system)

```
# symrdf –g|cg <RDF group_name> -noprompt write_disable R1 -force

An RDF 'Write Disable R1' operation execution is
in progress for device group '<RDF group_name> '. Please wait...
    Write Disable device(s) on SA at source (R1).............Done.
The RDF 'Write Disable R1' operation successfully executed for
device group '<RDF group_name> '.
```

```
# symrdf list pd

Symmetrix ID: 000190300357
                          Local Device View
-------------------------------------------------------------------------
                 STATUS     MODES                      RDF  S T A T E S
Sym    RDF       ---------  -----  R1 Inv   R2 Inv ----------------------
Dev  RDev  Typ:G SA RA LNK MDAT  Tracks  Tracks Dev RDev Pair
---- ---- --------  ---------  ----- -------  ------- --- ---- ---------

0124 0144  R1:1  RW RW NR   S..1   2102      1    WD  RW    Split
012B 0151  R2:1  RW WD RW   S..2      0      0    WD  RW    Synchronized
01EA 0444  R1:1  RW RW NR   S..1     30      0    WD  RW    Split
01F1 0451  R2:1  RW WD RW   S..2      0      0    WD  RW    Synchronized

Total                                -------- --------
  Track(s)                                  0        1
  MB(s)                                   0.0      0.1

Legend for MODES:

M(ode of Operation): A = Async, S = Sync, E = Semi-sync, C = Adaptive Copy
D(omino)           : X = Enabled, . = Disabled
A(daptive Copy)    : D = Disk Mode, W = WP Mode, . = ACp off
Mirror) T(ype)     : 1 = R1, 2 = R2
```

3. Update the RDF group (R1 copy):

```
# symrdf {–g|cg}  <RDF group_name> -noprompt update [-until
<InvalidTracks>] [-force]

An RDF 'Update R1' operation execution is
in progress for device group<RDF group_name> . Please wait...
    Suspend RDF link(s).....................................Done.
```

```
    Mark source (R1) devices to refresh from target (R2)......Started.
    Device: 0124 in (0357,001)............................... Marked.
    Mark source (R1) devices to refresh from target (R2)......Done.
    Merge device track tables between source and target.......Started.
    Devices: 0124, 01EA in (0357,001)........................ Merged.
    Merge device track tables between source and target.......Done.
    Resume RDF link(s)........................................Started.
    Resume RDF link(s)........................................Done.
The RDF 'Update R1' operation successfully initiated for
device group<RDF group_name>
```

When the RDF group is in the "R1 Updated" state , the roll-over of the application can be launched.

```
# symrdf list pd


Symmetrix ID: 000190300357
                    Local Device View

-------------------------------------------------------------------------
                 STATUS     MODES                      RDF  S T A T E S
Sym      RDF     ---------  -----  R1 Inv   R2 Inv ----------------------
Dev  RDev  Typ:G    SA RA LNK  MDAT  Tracks  Tracks Dev RDev Pair
---- ---- --------  ---------  ----- -------  ------- --- ---- ---------
0124 0144   R1:1    WD RW RW   S..1       0        0 WD  RW   R1 Updated
012B 0151   R2:1    RW WD RW   S..2       0        0 WD  RW   Synchronized
01EA 0444   R1:1    WD RW RW   S..1       0        0 WD  RW   R1 Updated
01F1 0451   R2:1    RW WD RW   S..2       0        0 WD  RW   Synchronized

Total                                -------- --------
  Track(s)                                  0        0
  MB(s)                                   0.0      0.0

Legend for MODES:

M(ode of Operation): A = Async, S = Sync, E = Semi-sync, C = Adaptive Copy
D(omino)           : X = Enabled, . = Disabled
A(daptive Copy)    : D = Disk Mode, W = WP Mode, . = ACp off
(Mirror) T(ype)    : 1 = R1, 2 = R2
```

14.7. Application Roll-over Facility Behavior Using SRDF

The following scenarios of storage system failure may occur:

- Site Failure:

  If the applications are launched by ARF in automatic take over mode, the heartbeat mechanism detects the failed node and the applications are taken over by the node defined as "take over node".

  ARF decides to make a recovery by running a specific SRDF command on the controller of the surviving site. The applications are launched on the take over node and will automatically access to the copy of data of the surviving storage system. The mirroring of data is disable.

- Node Failure:

  If the applications are launched by ARF in automatic take over mode, the heartbeat mechanism detects the failed node and the applications are taken over by the node defined as "take over node". The take over node will automatically access to the same copy of data as the failed node was.

- Storage system Failure:

  The node where the applications are launched by ARF detects an error on the storage system failure. ARF decides to make a recovery by running a specific SRDF command on the surviving controller. The applications are launched on the take over node and will automatically access to the copy of data of the surviving storage system. The mirroring of data is disable.

- Controller failure or Disk shelf failure:

  The recovery command is automatic. ARF has nothing to do.

- Link SRDF failure:

  The mirroring of data is disabled and no failover is done. The two sites will be running independently.

Chapter 15. Configuring Enhanced Remote Mirror (ERM) for ARF

15.1. ERM Overview

The Enhanced Remote Mirroring (ERM) is an option of the IBM DS4000 Storage Manager software and is used for replication data between DS4000 Storage Subsystem over a remote distance.

Note *Application Roll-over Facility* is compatible with the IBM DS5020 storage subsystems, which are supported by ERM.

In the event of disaster or unrecoverable error at one storage system, Application Roll-over Facility promotes automatically the second storage system to take over responsibility for normal I/O operations.

Three modes can be used with ERM: Metro Mirroring, Global Copy, Global Mirroring.

Only Metro Mirroring mode will be used with ARF.

The Metro Mirroring mode is a synchronous mirroring mode. Any host write request is written to the primary storage system and then transferred to the secondary storage system.

One of the LUN will have a primary role on the primary storage system. It will be mapped on the host of the Primary site.

The other LUN will have a secondary role on the secondary storage system. It will be mapped on the host of the Secondary site.

In case of disaster, the Secondary site will attribute the primary role to its associate LUN and the Secondary site will become Primary site.

The secondary logical drive is always read-only as long as the mirror is active.

The IBM DS4000 Storage Manager software provides two methods for managing storage systems:

- Host agent (in-band) management method
  Using this method method, you manage the storage systems through the Fibre Channel I/O path to the host.
- Direct (out-band) management method
  Using this method,you manage the storage systems directly over the network through a TCP/IP Ethernet connection to each controller. You must define the IP address and host name for each controller.

MultiPath Function

The MultiPath driver for AIX when attached to a DS4000 storage subsystem is the AIX FCP disk ARAY driver (RDAC) or the AIX Multi Path I/O (MPIO).

The RDAC is an IBM Multipath device driver that provides controller-failover support when a failure occurs anywhere along the Fibre Channel I/O path. The AIX FCP disk array driver uses LUN-level failover and supports four paths to the storage.

Configuration Example

To help you to configure the ERM functionality with an Application Roll-over Facility cluster, an example of configuration of two nodes is given below.

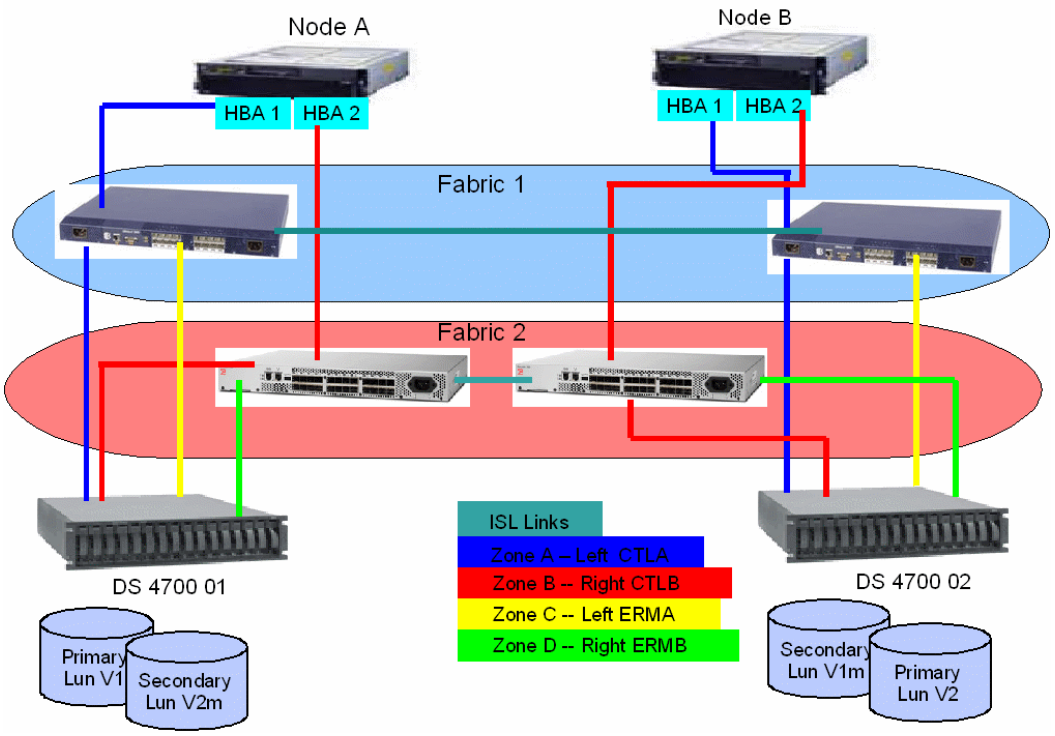


Figure 41. Enhanced Remote Mirror (ERM) configuration

Note You can also use only one fabric.

ISL Links is provided by the Fibre Channel Switch Links.

Data writing is done using zones A and B.

Data replication is done using zones ERMA and ERMB.

For more information refer to *IBM System Storage DS4000 and Storage Manager V10* documentation.

15.2. Initializing ERM with Application Roll-over Facility

15.2.1. Installing the DS4000 Storage Manager v10 on each ARF node

For each node of the ARF cluster, install the DS4000 Storage Manager v10 and create arrays and logical drives.

The following packages must be installed on ARF nodes:

- SMclient software package:

  SMclient is a Java-based graphical user interface for managing storage systems through the Ethernet network or from the host computer.

  The command-line interface (SMcli) is also packaged with the SMclient and provides command line access to perform all management functions. It is used by ARF.

- SMruntime software package:

  SMruntime is a java runtime environment required for SMclient. It must be installed before SMclient.

- SMesm software package:

  SMesm is required for automatic ESM firmware synchronization. It must be installed before SMclient.

- SMagent software package:

  SMagent is an optional component that allows in-band management of the DS4000 storage server. It is required only for in-band management that is through Fibre Channel.

- For MultiPath, it is recommended to use the AIX FCP disk ARAY driver (RDAC) with ARF.

For more information refer to *DS400 Storage Manager Version 10 Installation and Host Support Guide for AIX, HP-UX, Solaris and Linux on POWER*.

15.2.2. Overview of DS4000 Storage Management (SMclient)

For more information refer to *IBM System Storage DS4000 and Storage Manager V10* documentation.

Use the DS4000 Storage Management client (SMclient) to:

- Add a Storage Subsystem
- Manage a Storage Subsystem (create arrays and logical drives, assign logical drives into storage partitions)
- Enable the Enhanced Remote Mirroring (ERM).
- Create, Remove, Suspend Mirror Relationship

Creating Mirrors:

1. Activate the ERM feature (or check that it is activated).
   This activation must be performed only once on each storage system, regardless the number of mirrors to create.
2. Create the LUNs:
   Create a LUN on the first storage system, and an equivalent LUN on the second storage system. Check also that the Fibre links between the two storage systems are correctly configured (a dedicated link is required for ERM).
3. Create the mirror between the two volumes:
   Create an ERM link between the two LUNs and mirror them. One of the LUN will have a primary role on the primary storage system. It will be mapped on the host of the Primary site. The other LUN will have a secondary role on the secondary storage system. It will be mapped on the host of the Secondary site.

15.2.3. Check ERM LUN accessibility from each ARF node or VIO Server

On each ARF node, you have to perform the following actions :

1 - Verify that the hdisks and the logical drives of the associated Storage subsystem are visible by the ARF node.

2 - Verify that the name of the two storage subsystems and the ip address of the corresponding controller are defined on your ARF node.

3 - Verify that the Mirror Relationship of your different mirrors is accessible from your ARF node.

These tasks are detailed below.

1. Verify hdisk and Logical drives associated:

   If you use RDAC driver, use the fget_config –Av command.

   If you use MPIO driver, use the mpio_get_config –A command.

2. Verify that the name of the two storage subsystems and the ip address of the corresponding controller are defined:

   Run the following SMcli command:

   ```
   SMcli –i –d
   ```

   If it is not OK, run the following SMcli command for each DS4000 storage subsystem:

   ```
   SMcli –A <addrip_ctrlA>  <addrip_ctrlB>
   ```

   <addrip_ctrlA> : ip address of the controller A of the DS4000 storage subsystem

   <addrip_ctrlB> : c ip address of the controller B of the DS4000 storage subsystem

3. Verify the Mirror Relationship of your different mirrors:

   Run the following SMcli command for each Storage subsystem:

   ```
   SMcli –n <nameofstoragesubsystem> -c "show storagesubsystem;"
   ```

   or

   ```
   SMcli -n <nameofstoragesubsystem> –c  "show remoteMirror
   localLogicalDrive ["name of the LUN"] synchronizationProgress;"
   ```

15.2.4. Create Volume Group

With ERM, the primary logical drive is read-write and the secondary logical drive is always read-only as long as the mirror is active.

So, to create Volume groups, you must run the following steps:

- On the ARF node associated with the Storage System that contains the primary LUN:
  - Using SMIT, create the volume groups, logical volumes, filesystems.
  - Using SMclient, reverse the role of the LUNs mirrors (primary to secondary) on the Storage System.
- On the other ARF node: Import the volume group.

The No reservation option must be set on the disk.

15.3. Configuring ARF for ERM Environment

Before configuring, verify that ARF V7 or later is installed and especially that the **Bull.approllf.erm** lpp is installed.

To make ERM properly work in an Application Roll-over Facility environment you have to perform the following tasks on your cluster configuration:

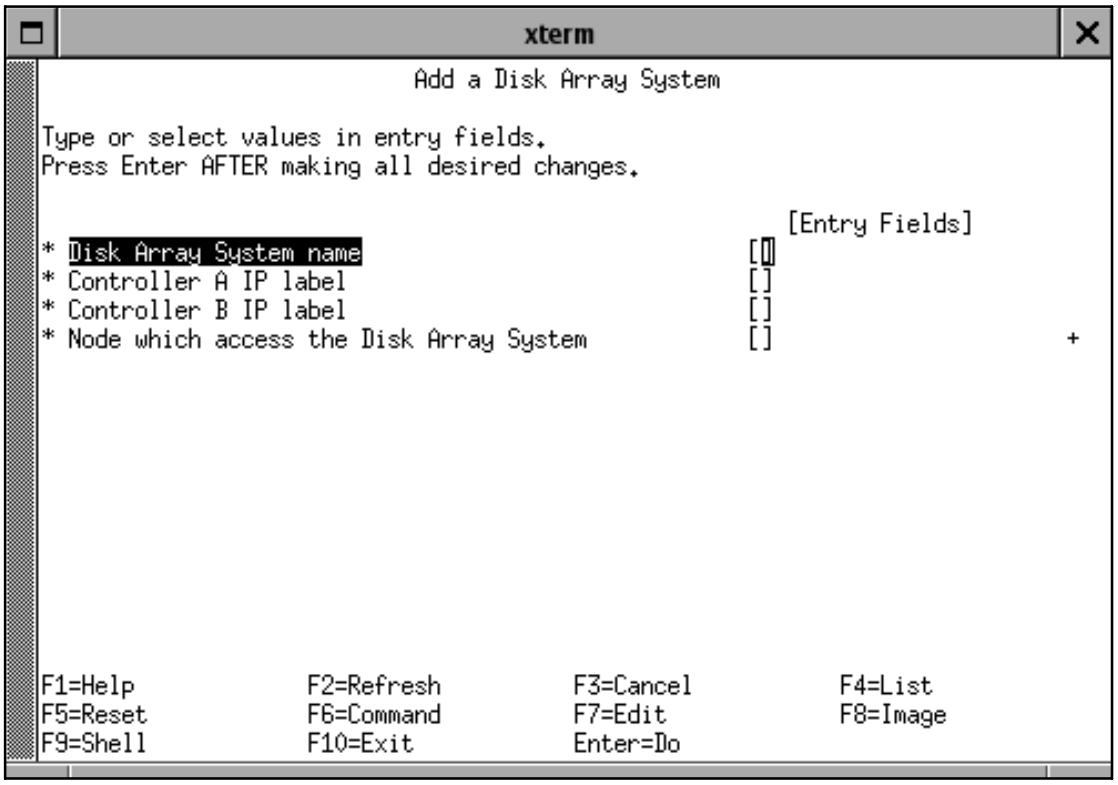
1. Define Disk Array System
2. Activate the ERM environment
3. Synchronize ERM environment

These tasks are detailed below.

1. **Define Disk Array System:**

   For each DS4700 Storage subsystem connected to the nodes, use the following menus:

```
# smit barf
      Configuration Definition
            Configure ERM environment
                  Define Disk Array Systems
                        Add a Disk Array System
```



   In the **Disk Array System name** field, enter a name that will uniquely identify the storage subsystem. This name is identical to the name defined in the SMclient software.

   In the **Controller A IP label** field enter the IP address of the controller A of the Storage Subsystem as defined in the **/etc/hosts** file.

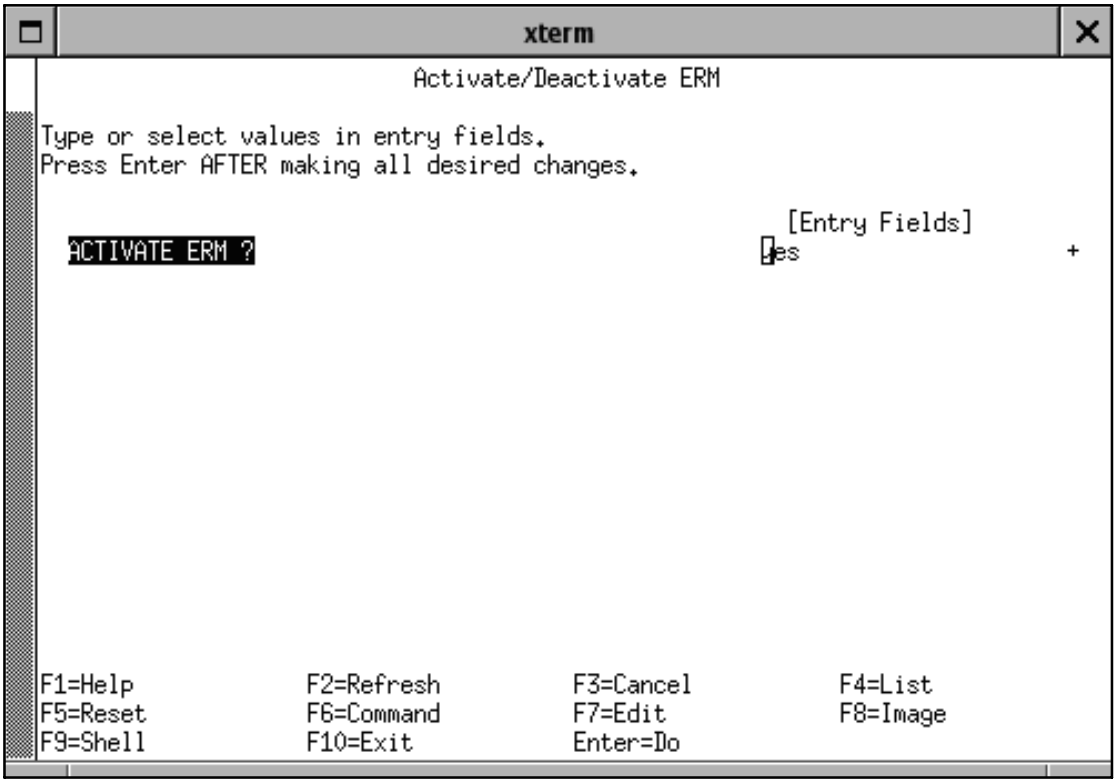
   In the **Controller B IP label** field enter the IP address of the controller B of the Storage Subsystem as defined in the **/etc/hosts** file.

   Enter the appropriate node name in the **Node which access the Disk Array System** field.

2. Activate ERM environment:

   To make effective ERM environment, use the following menus before starting Applications on the cluster nodes:

   ```
   # smit barf
         Configuration Definition
               Configure ERM environment
                     Activate/Deactivate ERM
   ```



   In the ACTIVATE ERM ? field select yes.

3. Synchronize ERM environment:

   To synchronize the ERM configuration on all nodes, use the following smit menus before starting Applications on the cluster nodes:

   ```
   # smit barf
         Configuration Definition
               Configure ERM environment
                     Synchronize ERM environment
   ```

15.4. Maintaining the ERM environment

In addition to the menus allowing to add Disk Array System in the ERM environment, menus are provided to modify or remove their definitions. These menus can be accessed as follow:

```
# smit barf
      Configuration Definition
            Configure ERM environment
                  Change/Show a Disk Array System definition
```

And :

```
# smit barf
      Configuration Definition
            Configure ERM environment
                  Remove a Disk Array System
```

If you make some changes to your ERM environment, do not forget to synchronize the new environment to all cluster nodes:

```
   # smit barf
         Configuration Definition
               Configure ERM environment
                     Synchronize ERM environment
```

15.5. Application Roll-over Facility Behavior Using ERM

The following scenarios of storage system failure may occur:

- Site Failure:

  If the applications are launched by ARF in automatic take over mode, the heartbeat mechanism detects the failed node and the applications are taken over by the node defined as “take over node”.

  ARF decides to make a recovery on the surviving controller. It consists to reverse the role of the secondary to primary logical drives on the surviving controller before launching applications on the take over node. The applications launched on the take over node will automatically access to the primary copy of data of the surviving storage system. The mirror data state becomes “unsynchronized”.

- Node Failure:

  If the applications are launched by ARF in automatic take over mode, the heartbeat mechanism detects the failed node and the applications are taken over by the node defined as “take over node”.

  ARF decides on the take over node to reverse role of the secondary to primary logical drives on the surviving controller. The applications are launched on the take over node and will automatically access to the primary copy of data of the surviving storage system. The mirror data state is “synchronized”.

- Storage system failure:

  The node where the applications are launched by ARF detects an error on the storage system failure. If the application is launched in automatic mode, ARF decides to take over node and the local node kills it.

  If the application is in automatic mode, ARF on the take over node decides to reverse role of the secondary to primary logical drives on the surviving controller. The mirroring data becomes “unsynchronized”.

- Controller failure:

  The recovery command is automatic. ARF has nothing to do.

- Interconnect failure:

  The mirror data state becomes “unsynchronized” and no failover is authorized by ARF. The applications running on the two sites still run.

15.6. ERM Recovery procedure

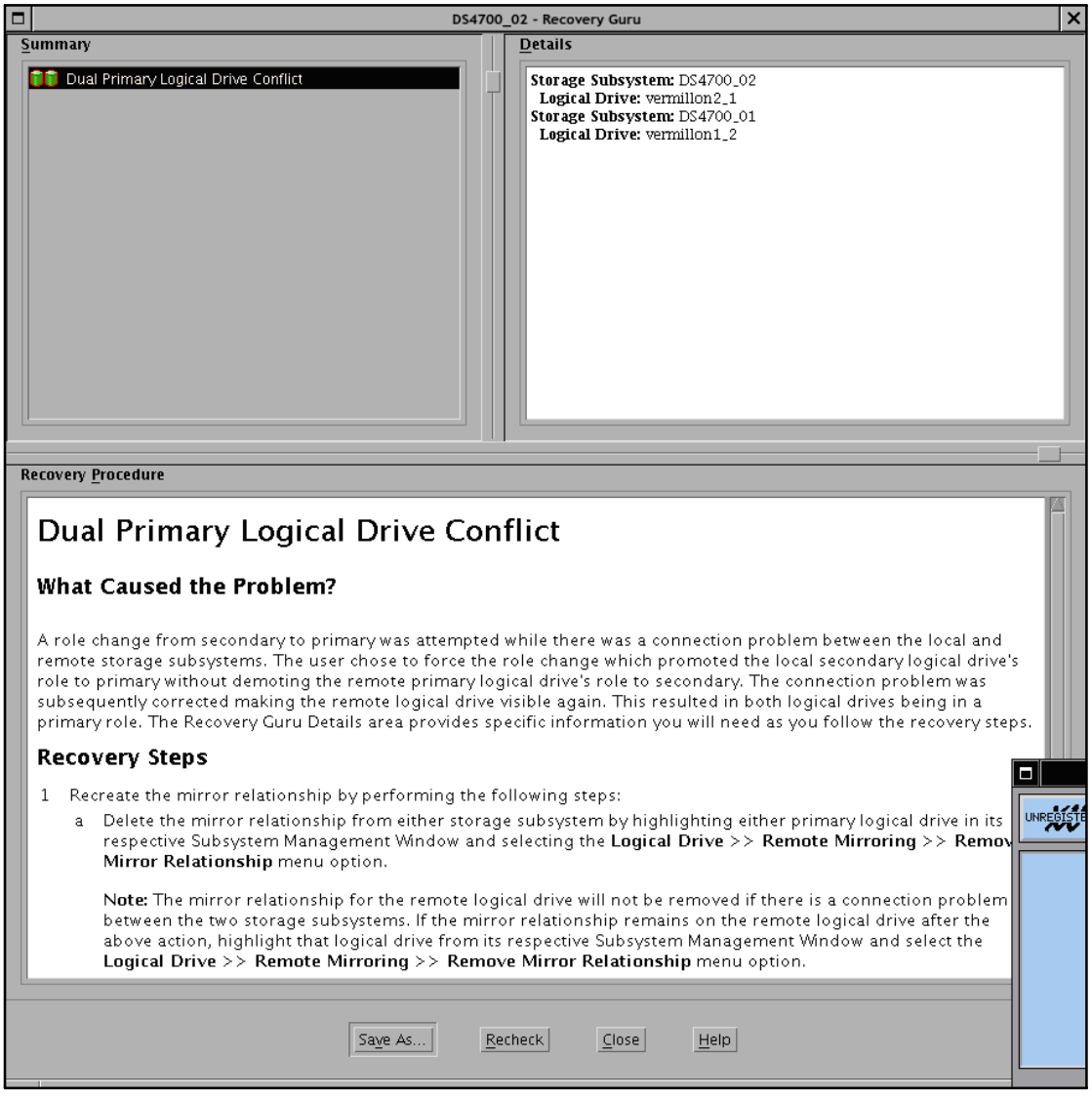
- Node Failure:

  After a Node Failure and repair, the mirroring of data is synchronized. The roll-over of the application can be launched. To roll-over the ARF Applications do as follows:

  ```
  # smit barf
        Manage Application Environment
              Roll-over Application Environment
  ```

- Storage system failure or Site Failure:

  After a Storage Subsystem Failure and after repair, there will be a Dual Primary error condition. Click **Recovery Guru** to see the problem and resolve him.



The Recovery Guru will instruct you to remove and re-create the mirror relationship.

You must run the following steps using the **SMcli** command or the SMclient Interface :

- Remove Mirror Relationship
- Re-Create Mirror Relationship

- **Interconnect failure:**

  After an Interconnect Failure and after repair, there will be a **Mirror Communication Error** condition. Click **Recovery Guru** to see the problem and resolve it.

  When connectivity is restored between the controller owner of the primary logical drive and the controller owner of the secondary logical drive, depending on the configured resynchronization method, either an automatic resynchronization takes place or manual resynchronization must be performed. Only the data in changed blocks will be transferred during the resynchronisation process. The status of the mirrored pair will be transferred during the resynchronization process. The status of the mirrored pair changes from an Unsynchronized state to a Synchronization-in-Progress state. Then, wait Synchronized state before launching application.

15.7. ERM Example

The following example describes how to use DS4000 Storage Management Client to create ERM Relationship:

1. Activate ERM on a storage system:
   Go to the **Storage Subsystem** menu, then **Remote Mirroring** and **Activate…**
2. Select the LUN with the primary role and select the **Create Remote Mirror…** function.

- **Primary Site A** (Local Site):

  Node A – Name of the primary Storage Subsystem: DS4700_01

  ERM1_1 : Primary LUN (mirror is ERM2_1 on DS4700_02)

  ERM4_2 : Primary LUN (mirror is ERM4_2 on DS4700_02)

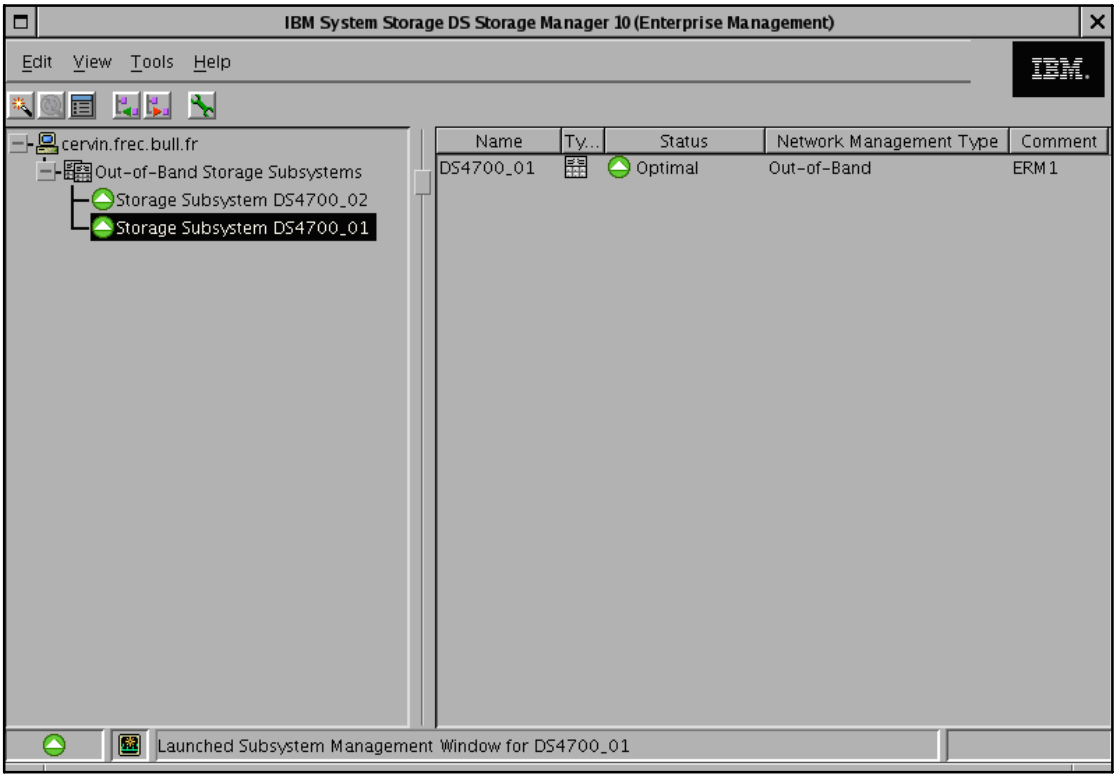
- **Secondary Site** (Remote Site):

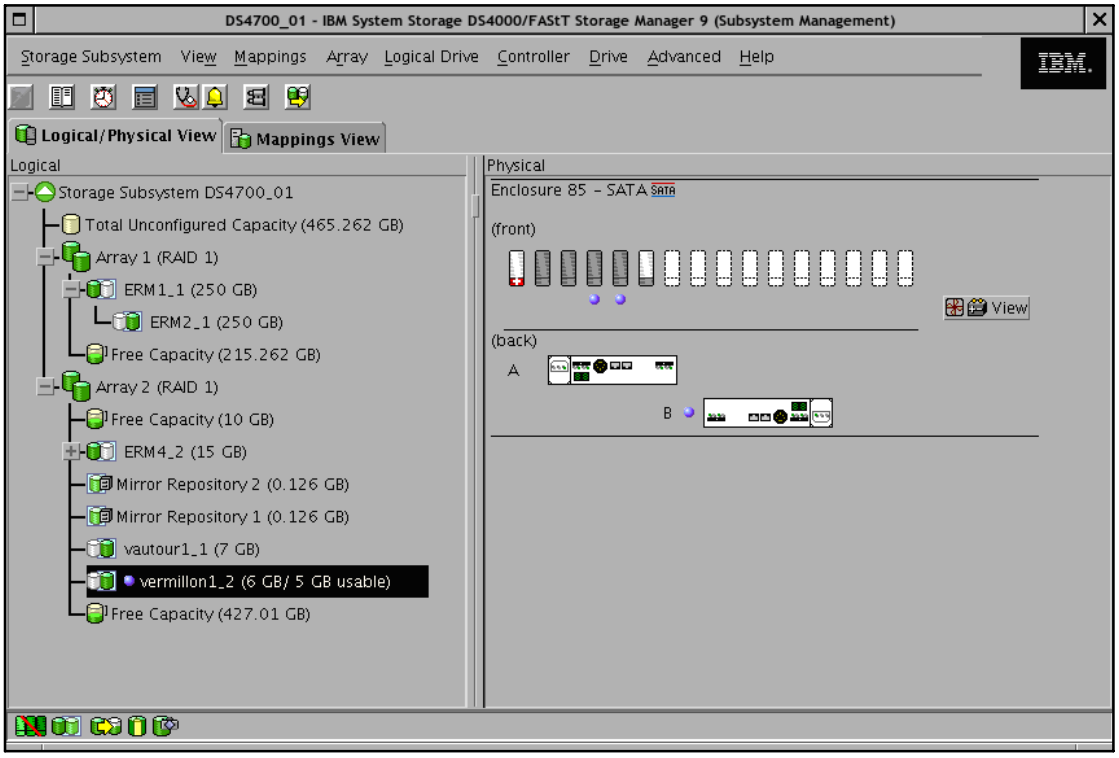
  Node B – Name of the secondary Storage Subsystem: DS4700_02

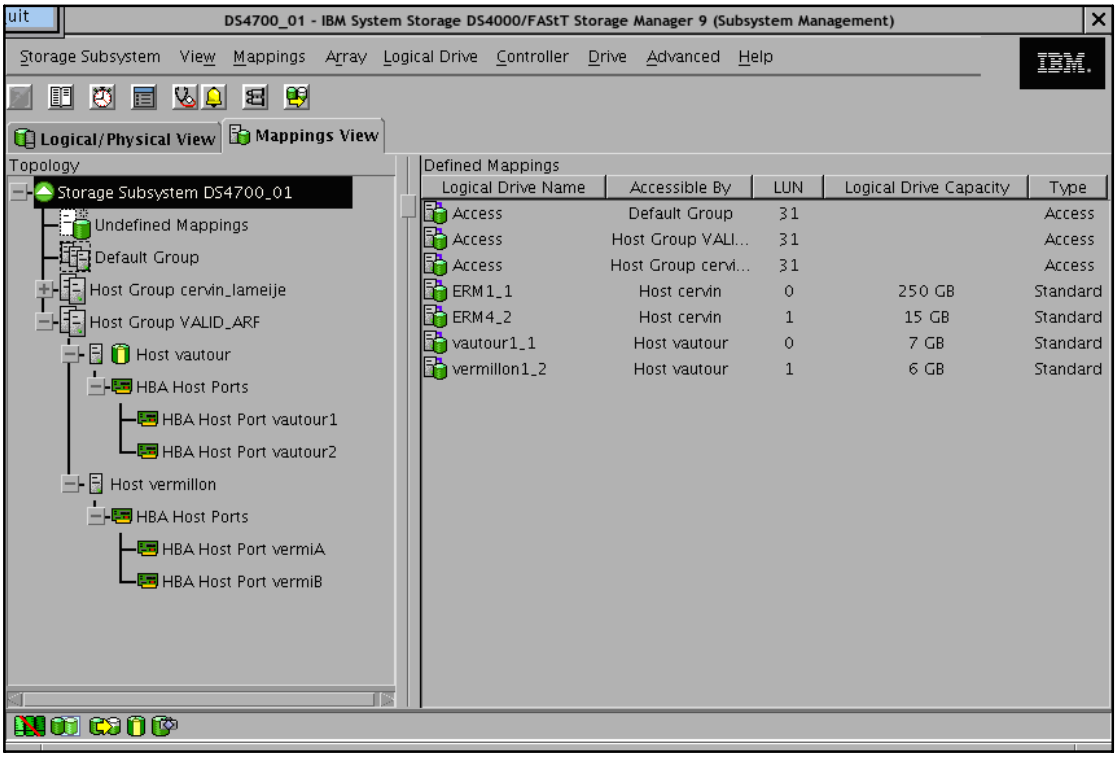
  ERM2_1 : Secondary LUN

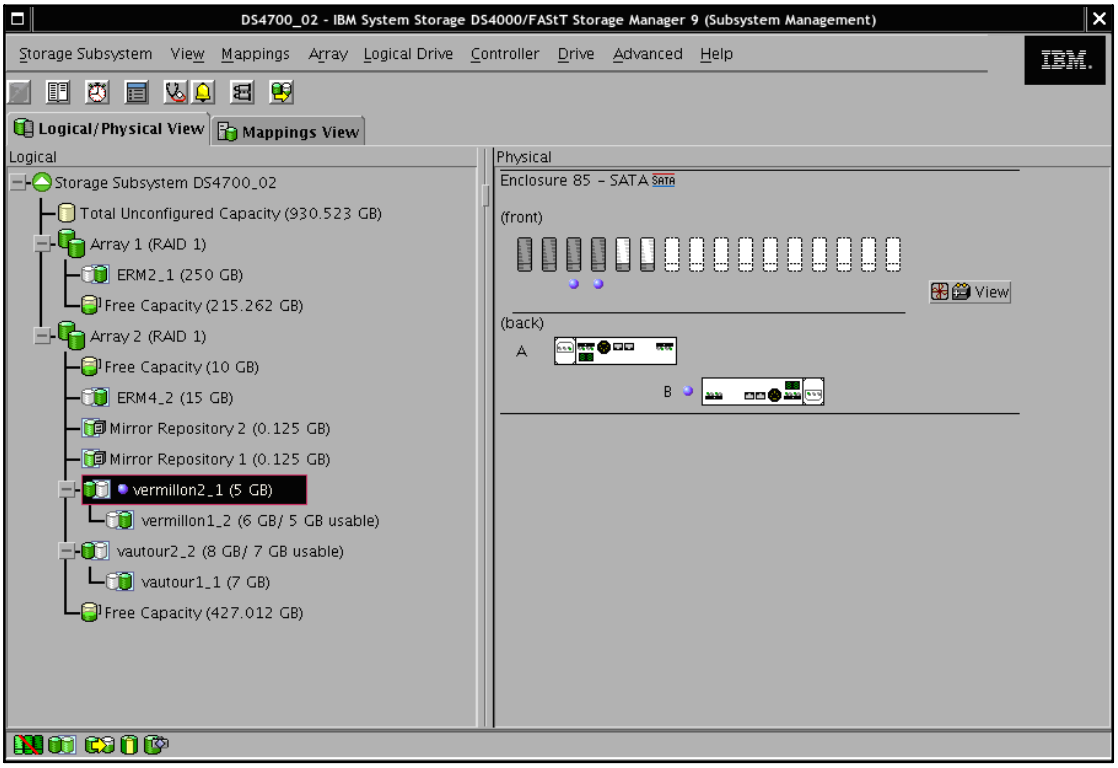
  ERM4_2: Secondary LUN

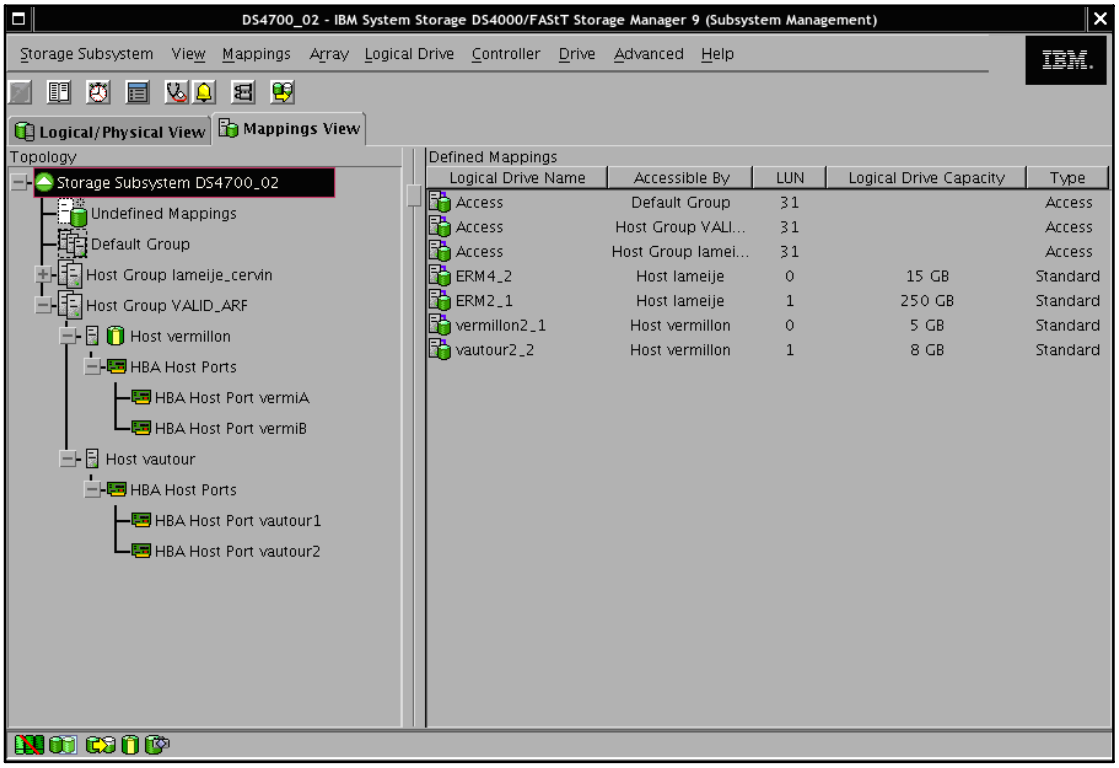
Following is an SMclient view after Storage System configuration, logical drive creation and mapping view:


IBM System Storage DS Storage Manager 10 (Enterprise Management)
Edit View Tools Help
IBM.
cervin.frec.bull.fr
Out-of-Band Storage Subsystems
Storage Subsystem DS4700_02
Storage Subsystem DS4700_01
Name
Ty...
Status
Network Management Type
Comment
DS4700_01
Optimal
Out-of-Band
ERM1
Launched Subsystem Management Window for DS4700_01


DS4700_01 - IBM System Storage DS4000/FAStT Storage Manager 9 (Subsystem Management)
Storage Subsystem View Mappings Array Logical Drive Controller Drive Advanced Help
IBM.
Logical/Physical View
Mappings View
Logical
Storage Subsystem DS4700_01
Total Unconfigured Capacity (465.262 GB)
Array 1 (RAID 1)
ERM1_1 (250 GB)
ERM2_1 (250 GB)
Free Capacity (215.262 GB)
Array 2 (RAID 1)
Free Capacity (10 GB)
ERM4_2 (15 GB)
Mirror Repository 2 (0.126 GB)
Mirror Repository 1 (0.126 GB)
vautour1_1 (7 GB)
vermillon1_2 (6 GB/ 5 GB usable)
Free Capacity (427.01 GB)
Physical
Enclosure 85 - SATA
(front)
View
(back)
A
B


uit
DS4700_01 - IBM System Storage DS4000/FAStT Storage Manager 9 (Subsystem Management)
Storage Subsystem View Mappings Array Logical Drive Controller Drive Advanced Help
IBM.
Logical/Physical View
Mappings View
Topology
Storage Subsystem DS4700_01
Undefined Mappings
Default Group
Host Group cervin_lameije
Host Group VALID_ARF
Host vautour
HBA Host Ports
HBA Host Port vautour1
HBA Host Port vautour2
Host vermillon
HBA Host Ports
HBA Host Port vermiA
HBA Host Port vermiB
Defined Mappings
Logical Drive Name
Accessible By
LUN
Logical Drive Capacity
Type
Access Default Group 31 Access
Access Host Group VALI... 31 Access
Access Host Group cervi... 31 Access
ERM1_1 Host cervin 0 250 GB Standard
ERM4_2 Host cervin 1 15 GB Standard
vautour1_1 Host vautour 0 7 GB Standard
vermillon1_2 Host vautour 1 6 GB Standard


DS4700_02 - IBM System Storage DS4000/FAStT Storage Manager 9 (Subsystem Management)
Storage Subsystem View Mappings Array Logical Drive Controller Drive Advanced Help
IBM.
Logical/Physical View
Mappings View
Logical
Storage Subsystem DS4700_02
Total Unconfigured Capacity (930.523 GB)
Array 1 (RAID 1)
ERM2_1 (250 GB)
Free Capacity (215.262 GB)
Array 2 (RAID 1)
Free Capacity (10 GB)
ERM4_2 (15 GB)
Mirror Repository 2 (0.125 GB)
Mirror Repository 1 (0.125 GB)
vermillon2_1 (5 GB)
vermillon1_2 (6 GB/ 5 GB usable)
vautour2_2 (8 GB/ 7 GB usable)
vautour1_1 (7 GB)
Free Capacity (427.012 GB)
Physical
Enclosure 85 - SATA
(front)
View
(back)
A
B



On Node A:

```
Node A# SMcli -i -d
```

```
DS4700_02          100.10.100.2      100.10.100.4
DS4700_01          100.10.100.1     100.10.100.2
```

- If you use RDAC driver:

```
Node A# fget_config -Av
```

```
---dar1---
User array name = 'DS4700_01'
dac3 ACTIVE dac1 ACTIVE
Disk      DAC    LUN Logical Drive
utm               31
hdisk26   dac1     2
hdisk24   dac3     0 ERM1_1
hdisk25   dac3     1 ERM4_2
```

- If you use MPIO driver, run:

```
mpio_get_config -A
```

- Check the configuration:

```
Node A# SMcli –n DS4700_01 -c "show storagesubsystem;"
```

```
Performing syntax check...
Syntax check complete.
Executing script...
Storage Subsystem profile
PROFILE FOR STORAGE SUBSYSTEM: DS4700_01 (6/5/09 8:55:51 AM)
SUMMARY------------------------------
.......
MIRRORED PAIRS-----------------------------
SUMMARY
   Number of mirrored pairs: 2 of 32 used
   Write consistency group: 2
   P = Primary logical drive
   S = Secondary logical drive
   LOCAL LOGICAL DRIVE  REMOTE      STATUS
   ERM1_1 (P)           ERM2_1 (S)  Synchronized
   ERM4_2 (P)           ERM4_2 (S)   Synchronized
DETAILS
   MIRRORED PAIR: Mirrored pair: ERM1_1 and ERM2_1
      Mirror status:    Synchronized
      Write mode:       Synchronous
      Synchronization priority:  Highest
      Resynchronization:         Manual
      Local logical drive:  ERM1_1
         Role:              Primary
         Logical Drive ID:
60:0a:0b:80:00:26:59:2a:00:00:06:cb:49:9b:a9:30
         Capacity:          250 GB
      Remote logical drive:         ERM2_1
         Role:                      Secondary
         Logical Drive ID:
60:0a:0b:80:00:42:4a:d2:00:00:06:9a:49:9c:24:0c
         Capacity:                  250 GB
         Remote storage subsystem:  Not Available
```

```
Node A# SMcli –n DS4700_01 -c "show remoteMirror localLogicalDrive
    ["ERM1_1"] synchronizationProgress;"
```

```
Performing syntax check...
Syntax check complete.
Executing script...
Mirror pair (primary ERM1_1/secondary ERM2_1): Synchronized
Script execution complete.
SMcli completed successfully.
```

Chapter 16. Setting Up and Using ARF Watch

This chapter explains how to set up and use the ARF Watch monitoring application.

ARF Watch is intended for system administrators and operators responsible for monitoring *Application Roll-over Facility* configurations.

16.1. ARF Watch Overview

ARF Watch is a facility for monitoring *Application Roll-over Facility* operations. With ARF Watch, the system administrators instantly have access to all the critical information they need to manage *Application Roll-over Facility* configurations and maximize up-time.

ARF Watch is a web-based application, providing *Application Roll-over Facility* information to the administrators and making monitoring *Application Roll-over Facility* intuitive and simple.

At a glance, the user can grasp the status of the main resources involved in the *Application Roll-over Facility* configurations. Abnormal status conditions are highlighted so that problem areas are easily identified. In addition to status information, ARF Watch provides information related to the *Application Roll-over Facility* configuration.

Each information page shows at once information on all the nodes of the monitored *Application Roll-over Facility* configuration. This makes comparing their respective status and configuration easy.

16.2. ARF Watch Components

ARF Watch is a client/server application.

16.2.1. Client component

The client simply consists of a **Web browser** It is the interface through which the user can access ARF Watch features and request information.

The users can run their Web browser with **Flash Player** plug-in from any host or station on the network, as long as the *Application Roll-over Facility* nodes are known and reachable through the network.

16.2.2. Server components

The server is in charge of gathering and delivering information requested by the user through the Web browser. It relies on two components: the **ARF Watch core software** and a **Web Server**.

ARF Watch Core Software

The ARF Watch core software mainly consists of programs (**Flex** application and **php** scripts) that perform the computation needed to build the information pages that are displayed through the Web browser.
This core software must be installed on the different nodes that will be monitored by *Application Roll-over Facility*.

ARF Web Server

The ARF Web Server (based on **Apache2 HTTP** Server) includes an HTTP server that handles the communications between the client and server. ARF Watch, for its own use, relies on the mechanisms of this HTTP server (including its password-related features).

ARF Watch relies on mechanisms provided by the Web Server, which implements a protection scheme based on password. For ARF Watch, the HTTP server knows one user: **arfw**.

This Web Server must be installed on the different nodes that will be monitored by *Application Roll-over Facility*.

16.3. Software Installation Concerns

The fileset related to ARF Watch is **Bull.approllf.arfw** (7.2.0.0 or higher) delivered on the **Application Roll-over Facility** CD-ROM.

Prerequisites

Read the *Application Roll-over Facility Software Release Bulletin* (SRB) that comes with the software. This document gives the procedure to install the software and the environment requirements and restrictions.

ARF Watch requires:

- A set of RPM (delivered on AIX Toolbox for Linux CD-ROM)
  - GCC compiler dynamic runtime library (**libgcc** rpm)
  - A library for manipulating JPEG image format files (**libjpeg** rpm)
  - A library of functions for manipulating PNG image format files (**libpng** rpm)
  - A library providing XML and HTML support (**libxml2** rpm)
  - A free and portable TrueType font rendering engine (**freetype2** rpm)
  - A pixmap library for the X Window System (**xpm** rpm)
- Application Roll-over Facility rte: **Bull.approllf.rte** fileset
- A HTTP server (based on Apache2) provided by **Bull.approllf.arfw.webserver** fileset.

Before installing **Bull.approllf.arfw** fileset, check that there is enough free space in the following directories:

/usr: 48 MB free for Apache2

/tmp: 60 MB free for Apache2

16.4. Setting-up ARF Watch

This section explains how to set up the environment for ARF Watch and how to get started.

In case of trouble refer to *ARF Watch Administration and Troubleshooting*, on page 16-23.

16.4.1. Setting Up the arfw User Password

The **arfw** AIX user is created at ARF Watch installation time (default password: **arfw**); this user will be used by the http server to execute php scripts on the *Application Roll-over Facility* nodes.

To set up the arfw password:

1. Login as **root**
2. Use standard AIX method to set a AIX password for **arfw** user
3. Logout
4. Login as **arfw** user and set again the password
5. Logout.

Repeat these steps on each ARF node.

16.4.2. Updating .rhosts File for arfw User (only if rsh is used)

We assume that you have already updated **/etc/hosts** and **.rhosts** files for **root** user as explained in chapter *Tailoring AIX for Application Roll-over Facility*, on page 5-1.

The *Application Roll-over Facility* nodes must be able to access each other in order to run the commands and php scripts needed to gather *Application Roll-over Facility* information. Consequently, you have to update the **.rhosts** file (located in home directory **/home/arfw**) for **arfw** user:

- Login as **arfw**
- Edit the **.rhosts** file:

  Add entries for all IP addresses of all nodes and application environments, with access rights granted to **arfw**.
- Check the access between nodes using the command **rsh <addr> -l arfw date**.

Repeat theses steps on all *Application Roll-over Facility* nodes.

16.4.3. Configuring ssh (only if sshis used)

If you plan to use **ssh** instead of **rsh** in your ARF configuration, you have to authorize ssh access to the **arfw** user for all nodes. To do this, refer to *Generating the SSH keys*, on page 3-1. In step 1, log on as **arfw** user instead of **root** user.

16.4.4. Checking the Web Server

The Web server will be started at installation time with correct configuration for ARF Watch.

Run the **ps** command to check that the **/usr/local/apache2/bin/httpd** process is running.

In case of problem, refer to *ARF Watch Administration and Troubleshooting*, on page 16-23.

16.5. Starting ARF Watch

To access ARF Watch, start a Web browser and specify the appropriate **URL** .

Start a Web browser

You can run your Web browser from any host or station on the network, as long as the *Application Roll-over Facility* nodes are known and reachable through the network. The **Flash Player** plug-in is required.

Specify URL

Specify an URL with the following format:

http://IP-spec/arfw

In this URL, *IP-spec* is an **IP label** or an **IP address** that is valid and reachable to connect to one *Application Roll-over Facility* node.

Example: to access ARF Watch on the **foo** *Application Roll-over Facility* node, whose IP label is **foo** and IP address is **192.9.200.1**, enter either:

http://foo/arfw
or
http://192.9.200.1/arfw

The corresponding Web server is contacted on the involved node and you are prompted to enter a **User ID** and a **Password**.

You must always specify **arfw** for **User ID**. The default password is **arfw**. (To change the password, see *Setting-up ARF Watch*, on page 16-22 .)

Once the correct password is entered, ARF Watch programs are run on this node and on the other *Application Roll-over Facility* nodes to gather information. Then the **Main View** page for this *Application Roll-over Facility* configuration is returned to the Web browser.

16.6. ARF Watch Main View

Note The colors in the ARF Watch pages have the following general meaning:

- GREEN indicates that everything is OK (applications are running, nodes are up...)
- ORANGE indicates a warning (some applications are not running, addresses are down...)
- RED indicates an error, a potential problem or incoherence.

The illustration below shows the **Main View** information page.

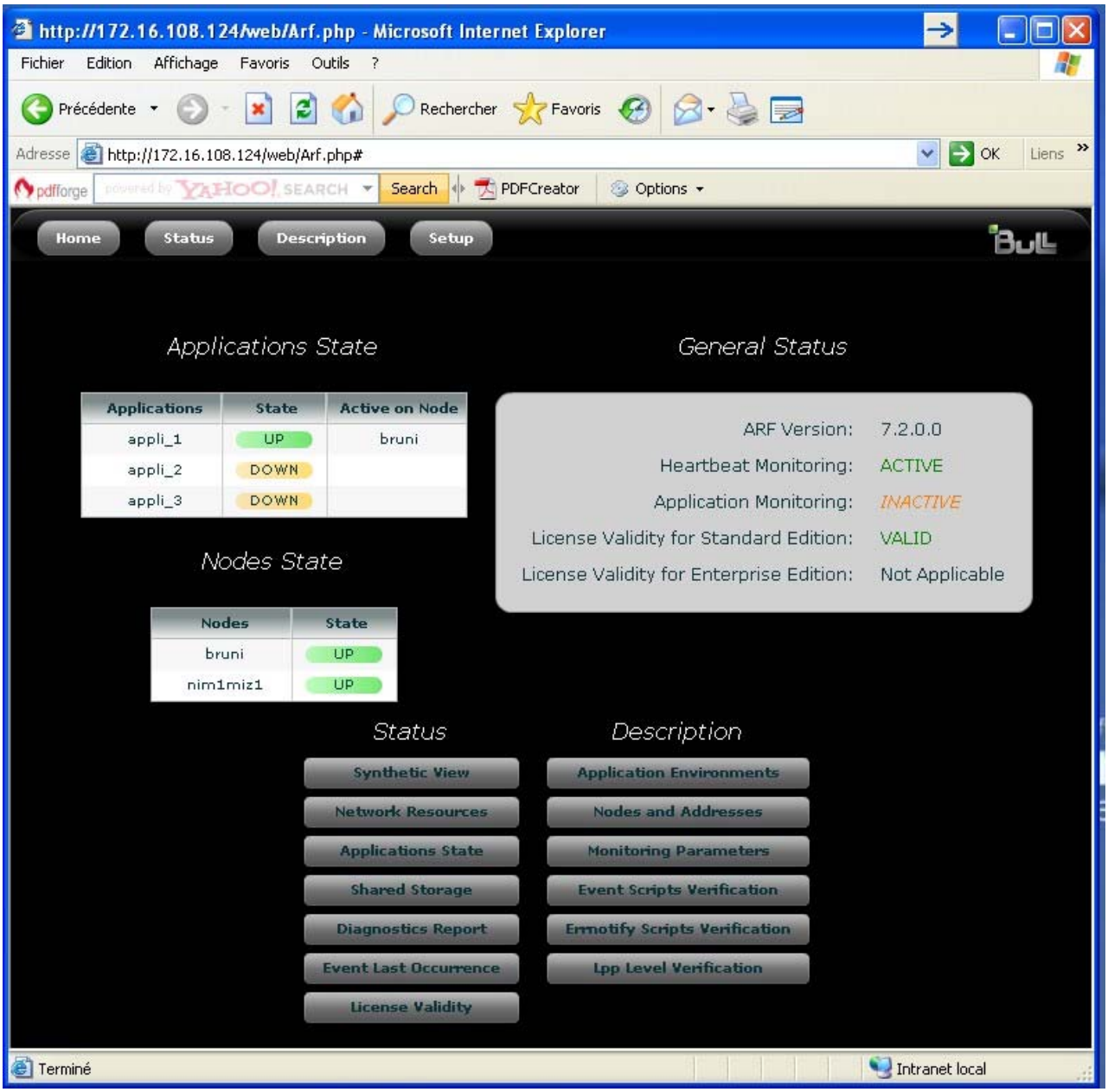


Figure 42. ARF Watch - Main View page

The information displayed in this **Main View** is:

- **Applications State**:

  Show the state of each application. The color depends on its state :

  - ORANGE if application is DOWN (not running).
  - GREEN if application is UP (running).

- **Nodes State**:

  Show the state of each node. A node icon can have the following color depending on its state:

  - RED if it is UNREACHABLE: the node can not be accessed.
  - GREEN if it is UP: the node can be accessed.

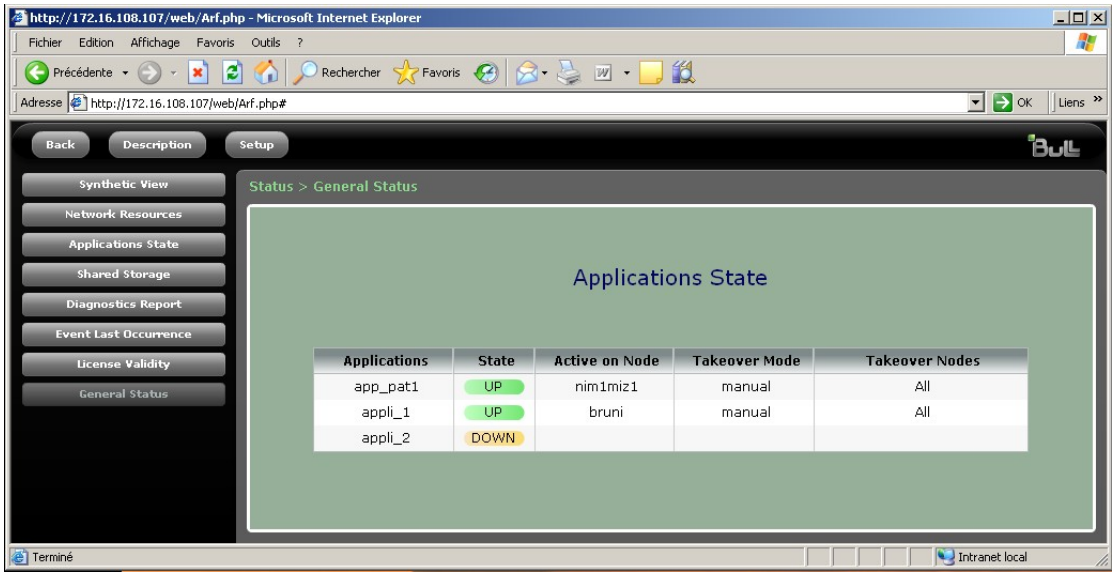
- **General Status**:

  - **ARF Version**: version number of Application Roll-over Facility fileset on each node.
  - **Heartbeat Monitoring**: state of heartbeat monitoring daemon on each node (ACTIVE or INACTIVE).
  - **Application Monitoring**: state of application monitoring daemon on each node (ACTIVE or INACTIVE).
  - **License Validity For Standard Edition**: validity of the license on each node (VALID or INVALID).
  - **License Validity For Enterprise Edition**: validity of the license on each node (VALID or INVALID)

- Direct links to various **Status** and **Description** information are available.

- **Status** button:

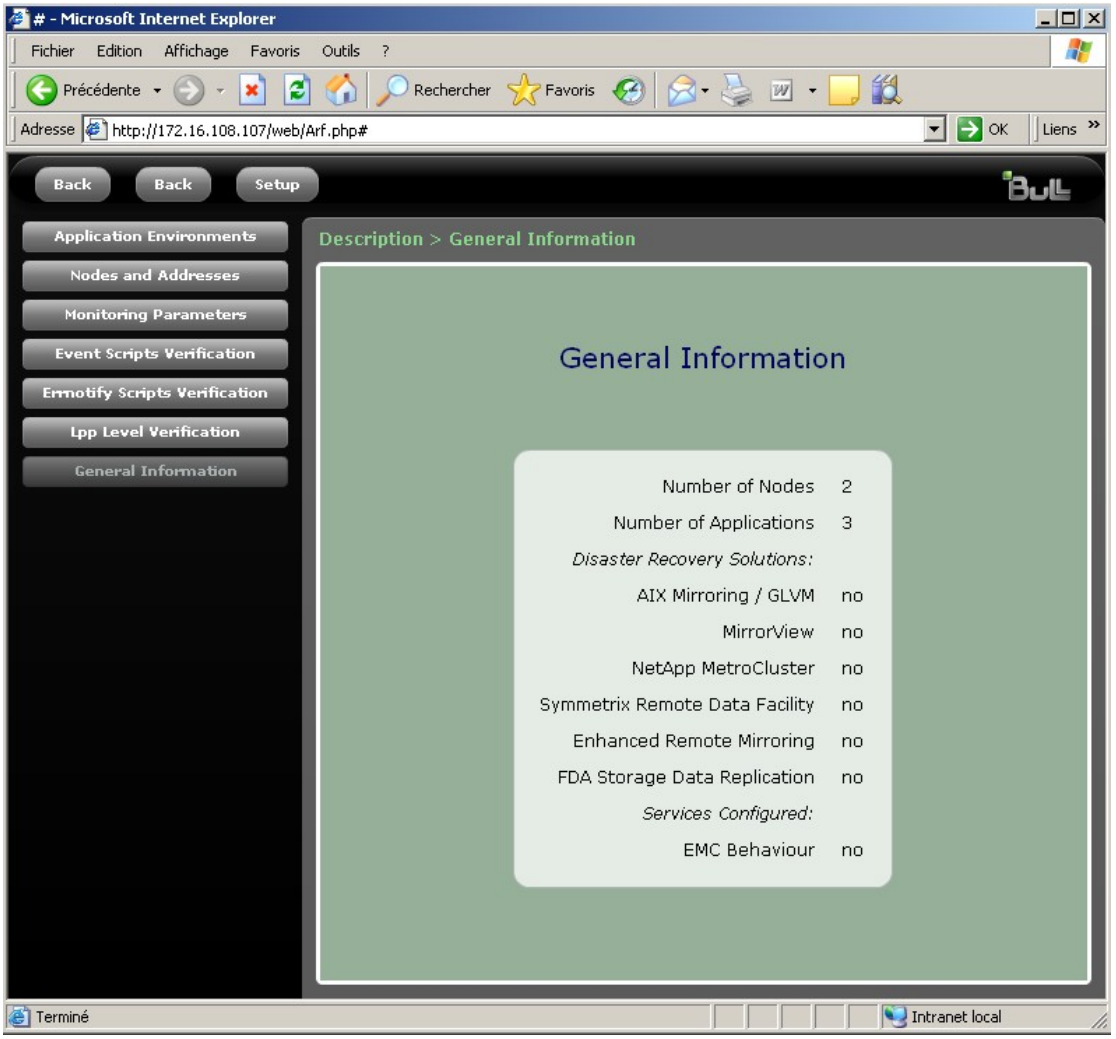
  If you click the **Status** button, the following page showing applications state is displayed:



See *Status Information*, on page 16-8 for details.

- **Description** button:

  If you click the **Description** button, the following page showing configuration information (number of nodes, number of applications, Disaster Recovery Solutions settings). is displayed:



See *Description Information*, on page 16-16 for details.

- **Setup** button:
  The **Setup** button, provides access to the **Password Management** function.
  See *Setup Function*, on page 16-22 for details

16.7. Status Information

16.7.1. Synthetic View

The illustration below shows the **Synthetic View** information page.

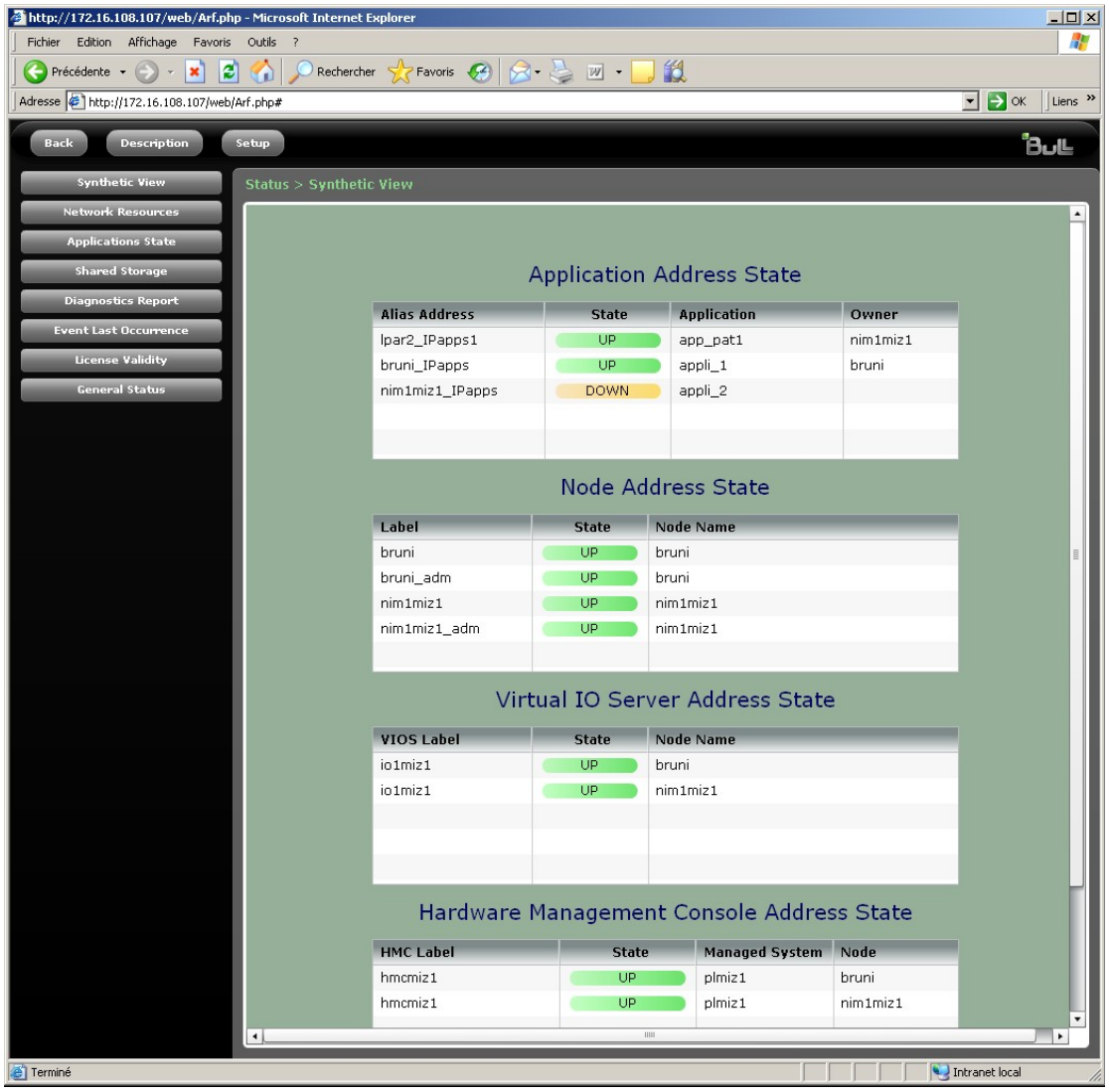


Figure 43. Synthetic View page

Synthetic View page displays several arrays showing:

- Application Address State
- Node Address State
- If virtual IO servers are defined in the *Application Roll-over Facility* configuration, the state of their interfaces is displayed.
- If HMC(s) is/are defined in the *Application Roll-over Facility* configuration, the state of their interfaces and the managed systems is displayed.
- If PP is used, the Serial Network State is displayed. If the connection is established on both sides, the state is UP.

16.7.2. Network Resource

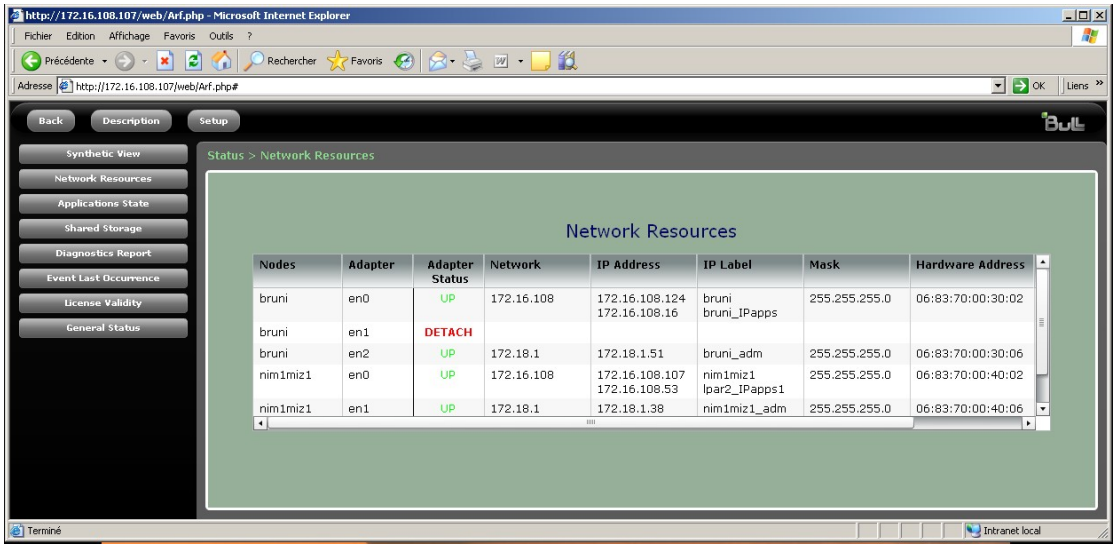


Figure 44. Network Resource page

This page provides dynamic information about the network adapters state. It shows for each node, IP address, IP label, network and netmask configured for the adapters.

The last column contains specific information for PPP connection and Etherchannel.

The **Adapter Status** field is written depending on the current state of the interface: GREEN if it is UP, ORANGE and in italic if it is DOWN, RED uppercase if it is detached.

16.7.3. Applications State

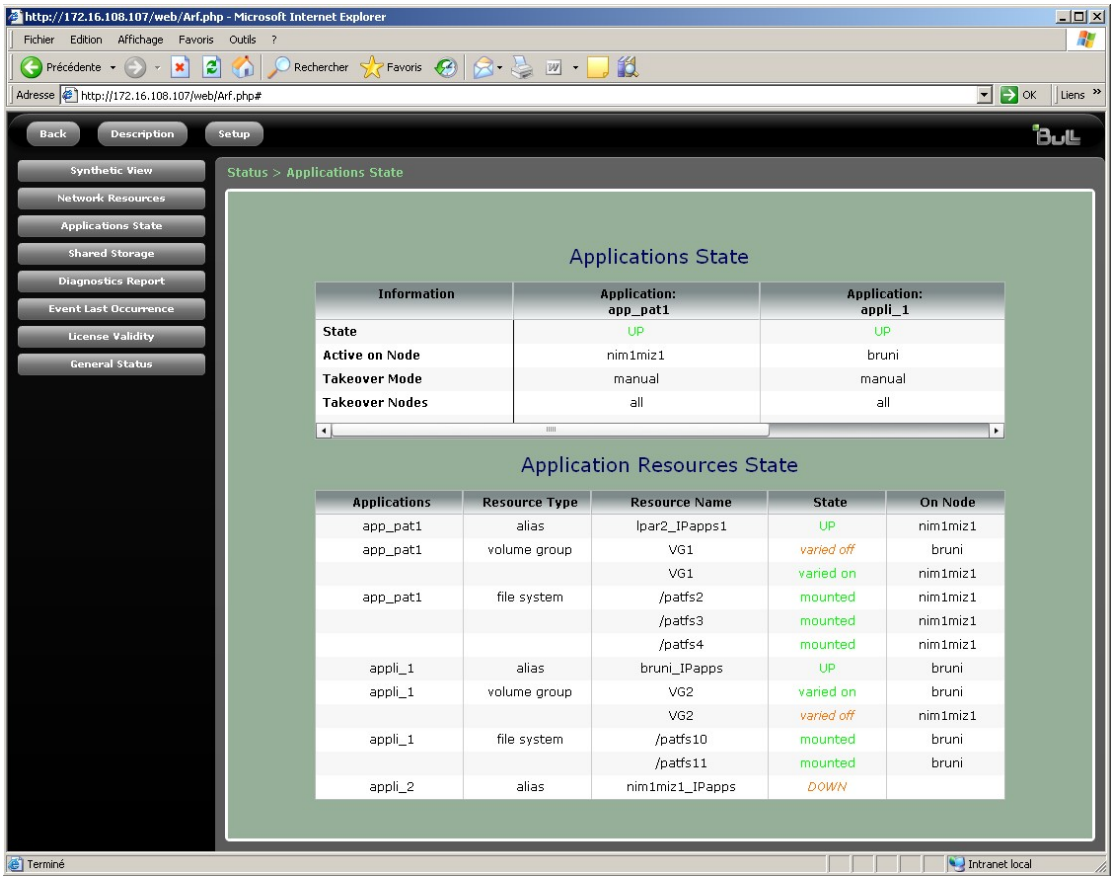


Figure 45. Application Environments State page

This page shows the state of each application, the node where the application is running, the takeover mode and takeover nodes if any.

The color depends on its state:

- ORANGE and in italic if application is DOWN (not running)
- GREEN if application is UP (running)

The state of the application resources (vg, fs, alias) is also displayed.

16.7.4. Shared Storage

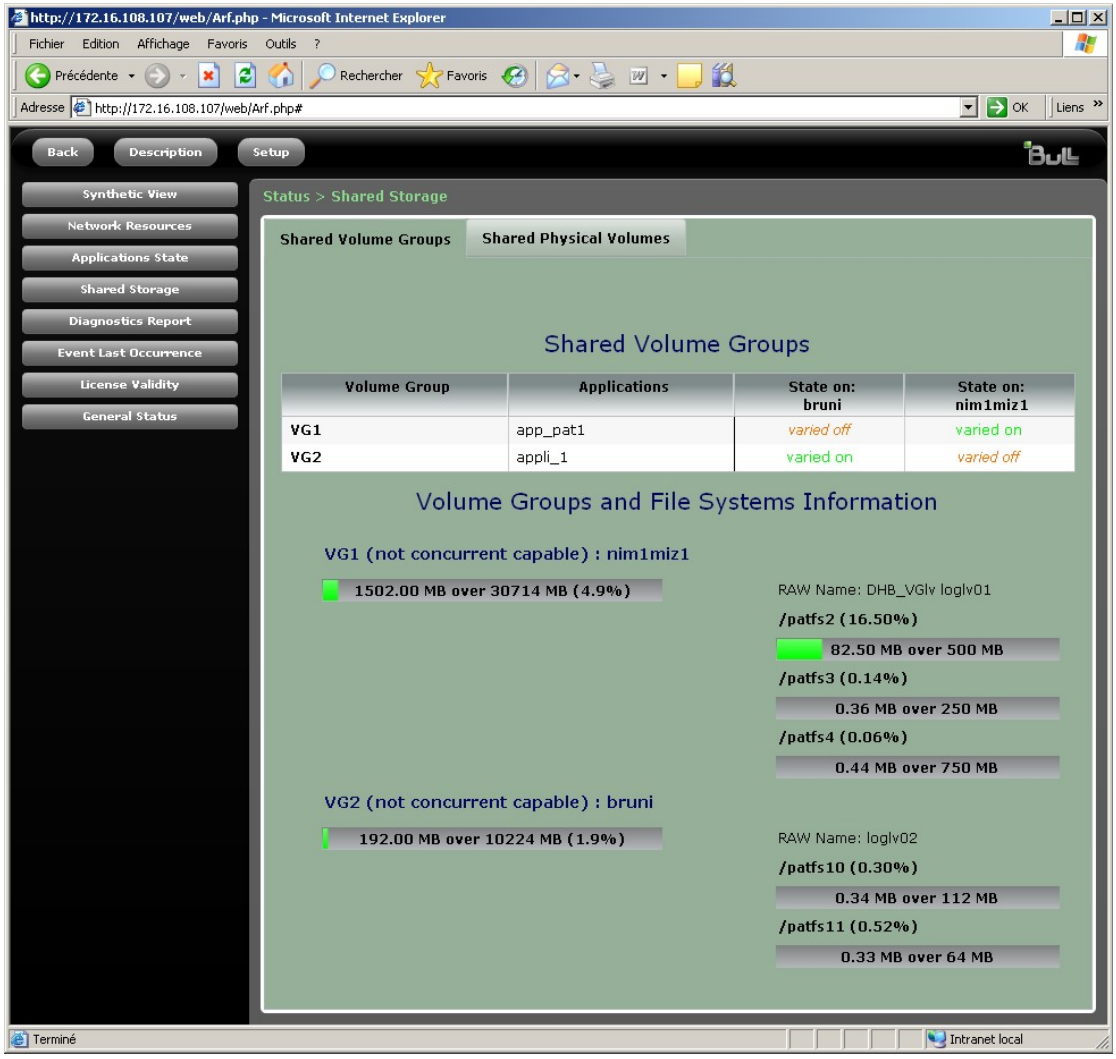
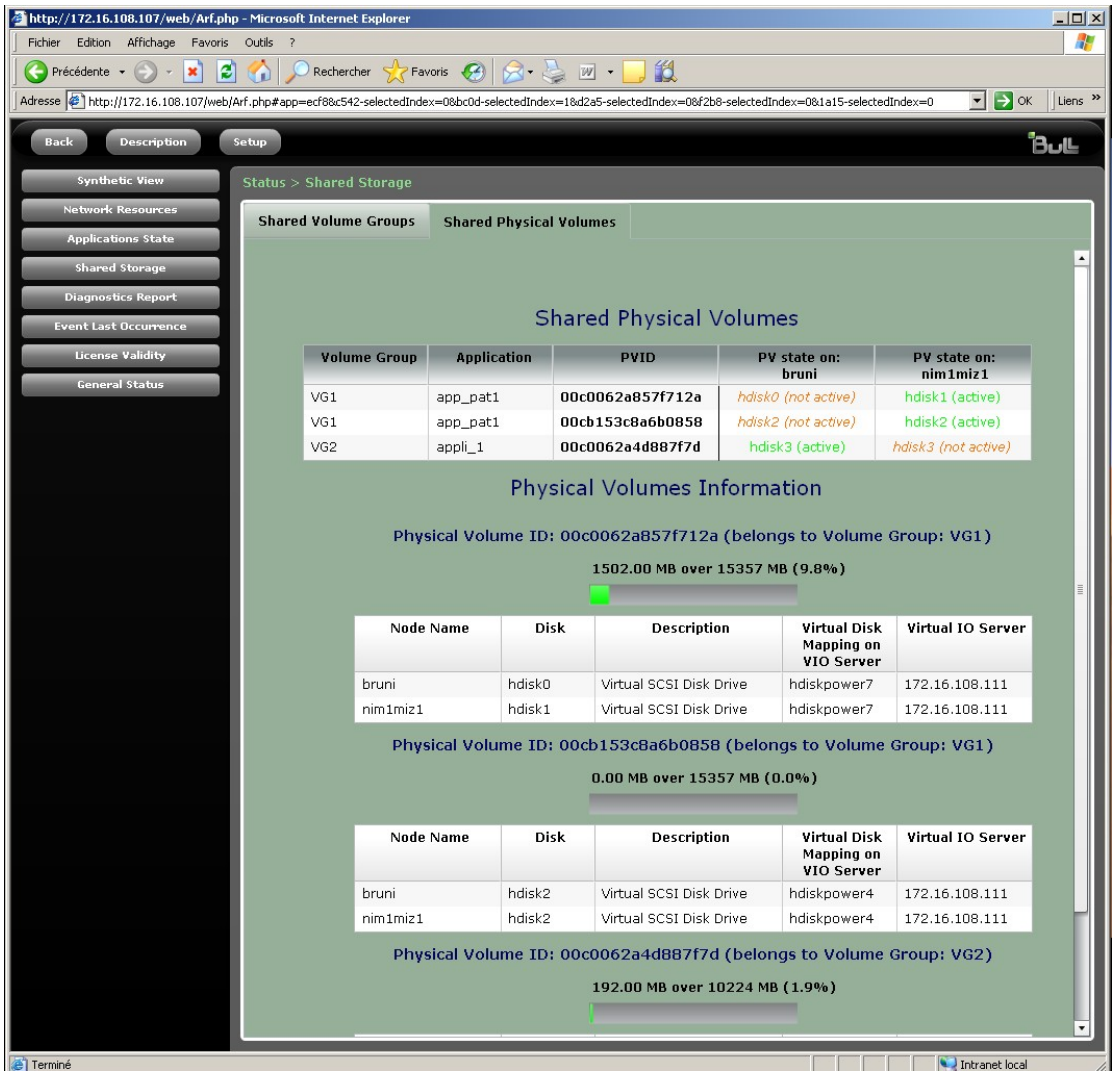


Figure 46. Shared Storage page

The information displayed is an array that indicates, for each shared volume group (VG), the application environment to which it belongs and its state on each node. A VG can be ON (varyonvg) or OFF (varyoffvg).



For each volume group and file system, total size and occupation rate are displayed.

For each physical volume, total size, occupation rate, PVID and virtual disk mapping (if Virtual IO servers are defined) is displayed.

The occupation rate for volume group, file system and physical volume appear in color:

- RED if occupation rate > 90%
- ORANGE if 80% < occupation rate <90%
- GREEN if occupation rate < 80%

16.7.5. Diagnostics Report

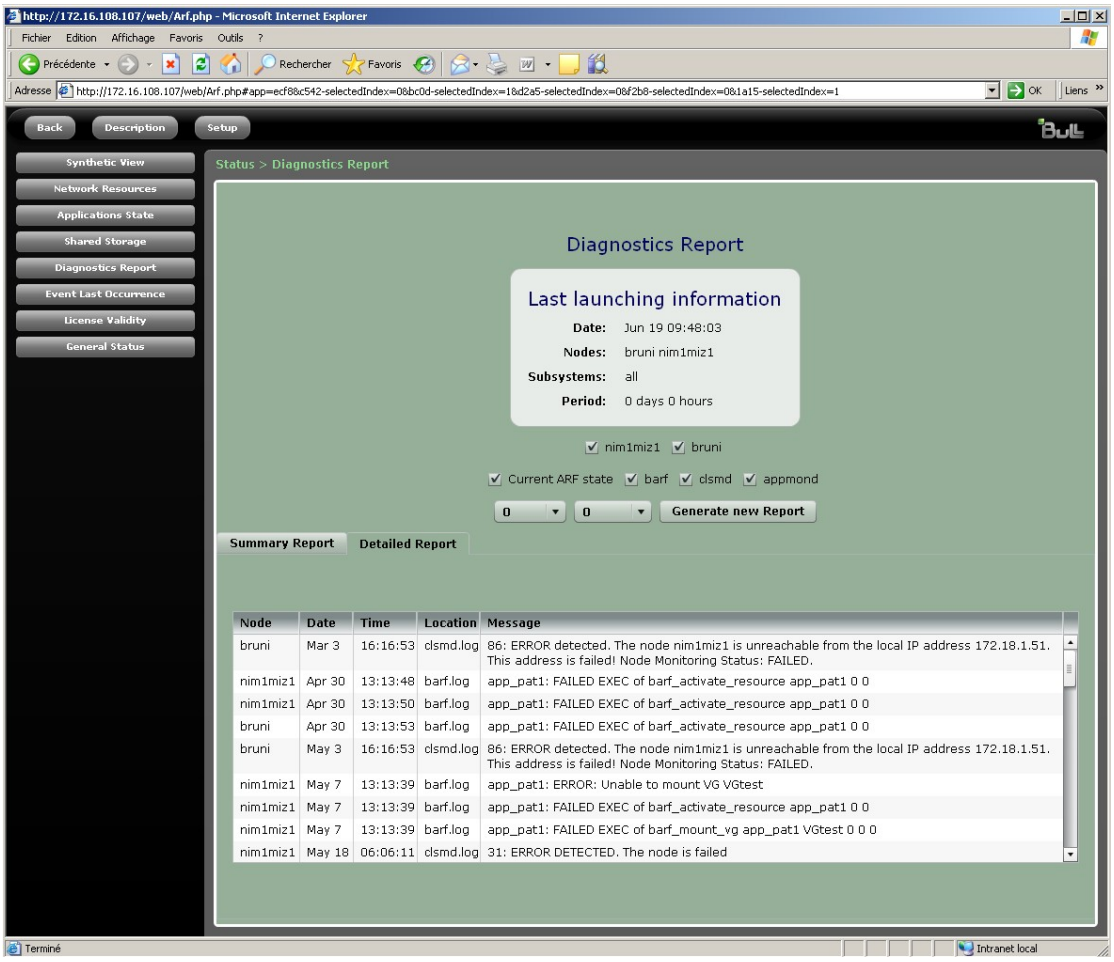


Figure 47. Diagnostic page - summary report

This page displays the diagnostic when the tool is launched with its default values.

See *Running the diagnostic tool*, on page 21-1 for more details.

The following information is reported for all the nodes:

- The origin of the problem (inconsistency state or error message in log files).
- You can launch the diagnostic tool with specific parameters. The user can customize the following items:
  - The nodes on which the diagnostic tool will be launched. By default all nodes are selected. Uncheck the nodes you do not want to diagnose.
  - The subsystems which will be scanned (aix, arf, smd). By default all subsystems are selected. Uncheck the subsystems you do not want to diagnose.
  - The period. If P is the chosen period and T the current time, alarms will be displayed for the interval [T – P; T]. The default period is 1 day. You have to select a number of days and a number of hours. The default values are 0 hours ans 0 days.

  The **Generate New Report** button launches the diagnostic tool with the specified parameters and displays the report in the same page.

16.7.6. Event Last Occurrence

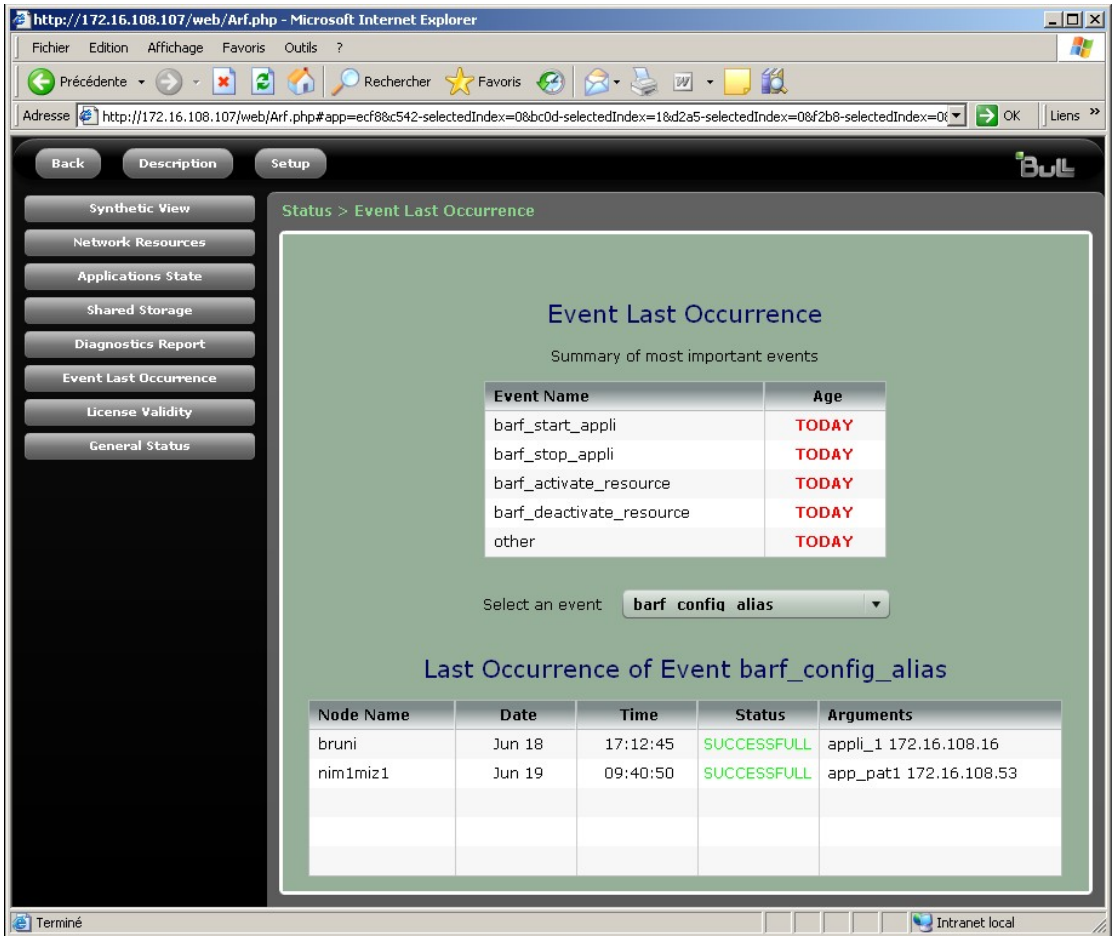


Figure 48. Event Last Occurrence page

This page is composed of two parts. The first part displays very important *Application Roll-over Facility* events state , the second part concerns a specific event.

- The part related to important events, indicates the events Age:
  - TODAY: if the event occurred during those last 24 hours. It is then written in RED and uppercases.
  - RECENT: if the event occurred between 24 and 72 hours earlier. It is then written in ORANGE and italic.
  - OLD: if this event occurred at least 72 hours earlier. It is written in GREEN.
- The part related to a specific event indicates for each node, the date and time of the last occurrence of this event, the event exit status (successfull or not) and the argument associated to the event. Information is displayed only if an event has been chosen.

  To choose an event use the drop-down menu. The *Application Roll-over Facility* events are listed. The research of information is launched by clicking on the "**Refresh**" button.

16.7.7. License Validity

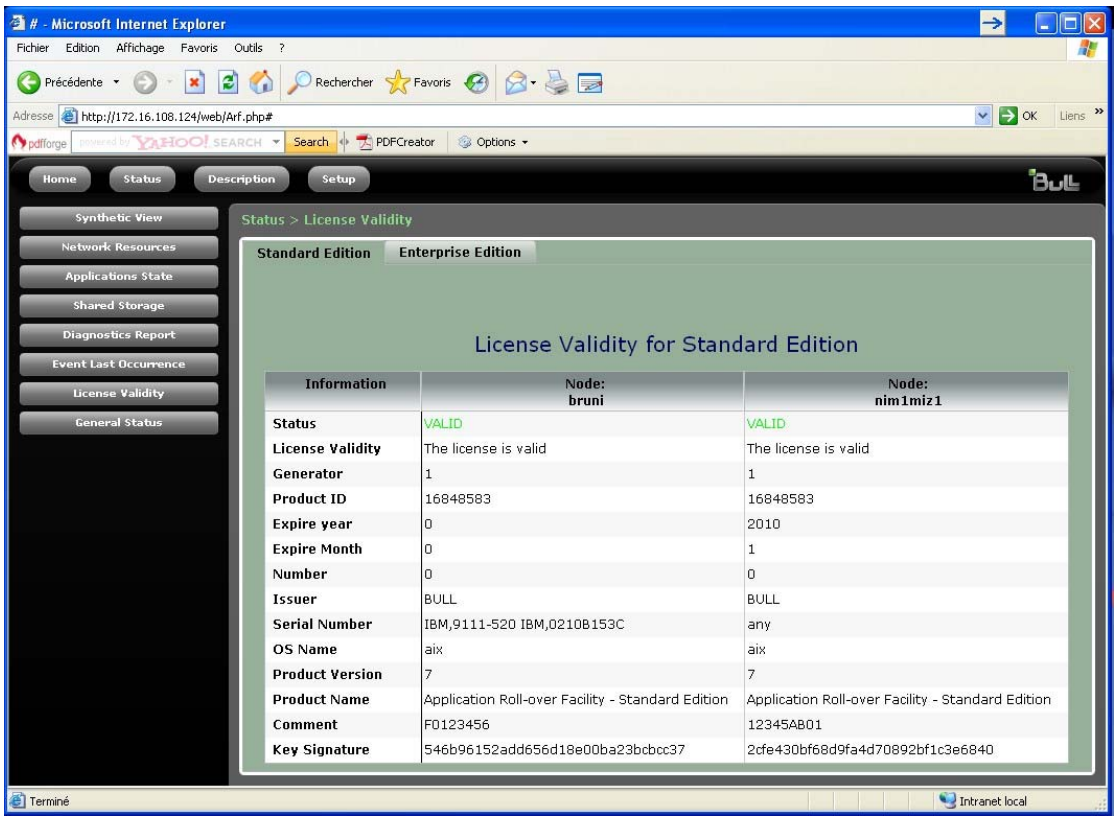


Figure 49. License Validity page

Note The **License Validity for Enterprise Edition** tab appears only if *ApplicationRoll-over Facility Enterprise Edition* is installed and configured.

For each node, the License Validity page displays information such as expiration date, system id for which the license is valid, version for which license is valid, key signature, ...

16.8. Description Information

16.8.1. Application Environments

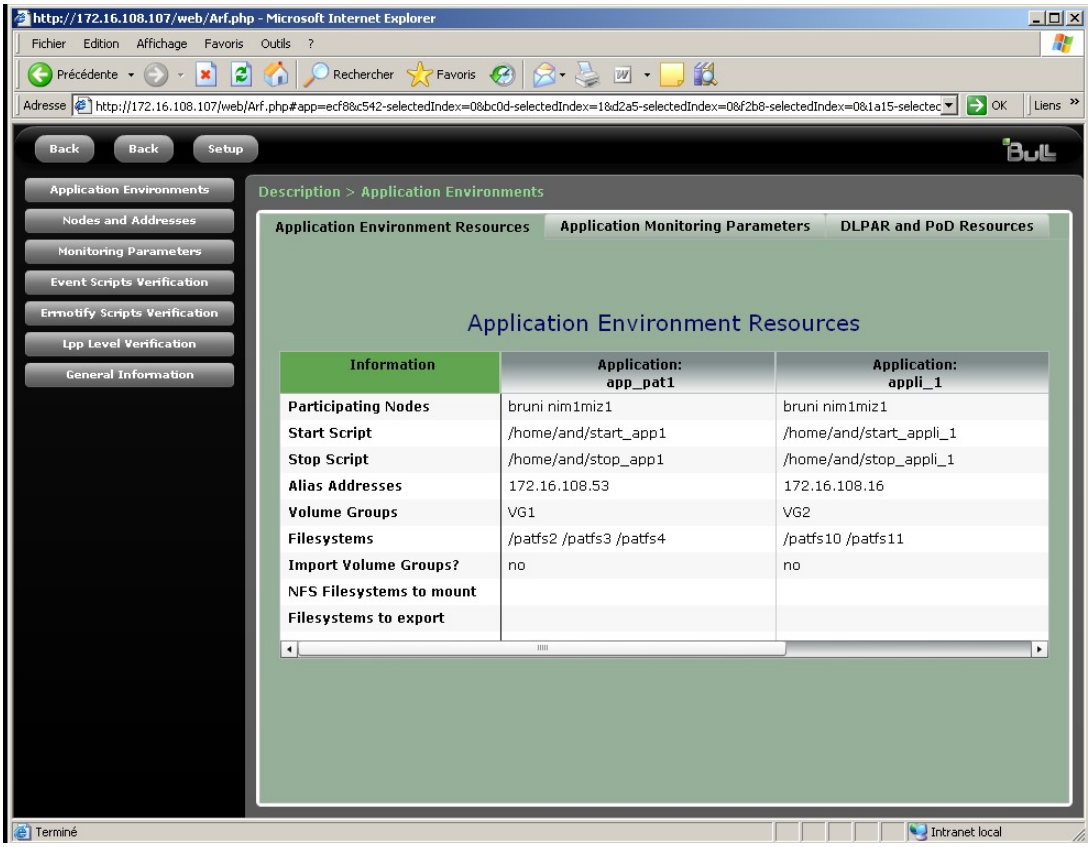


Figure 50. Application Environments and Monitoring Parameters pages

These pages display detailed configuration of application environment resources, application monitoring parameters and DLPAR and PoD resouces.

16.8.2. Nodes & Addresses

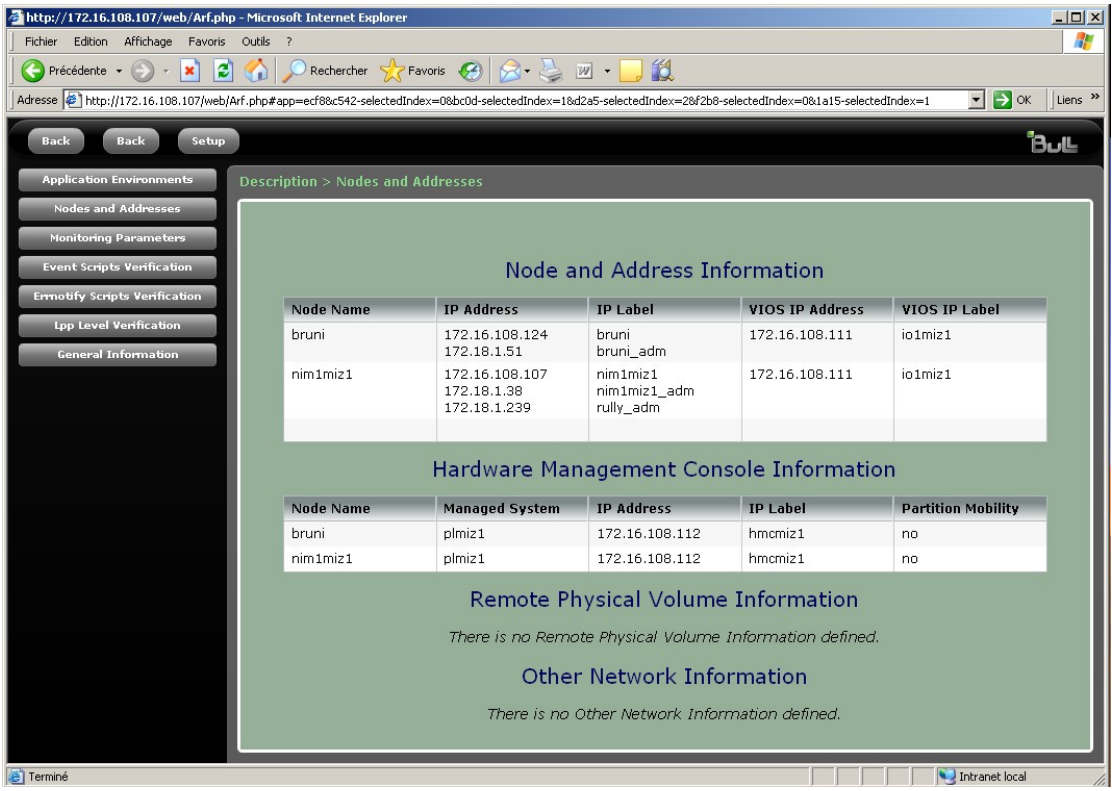


Figure 51. Nodes & Addresses page

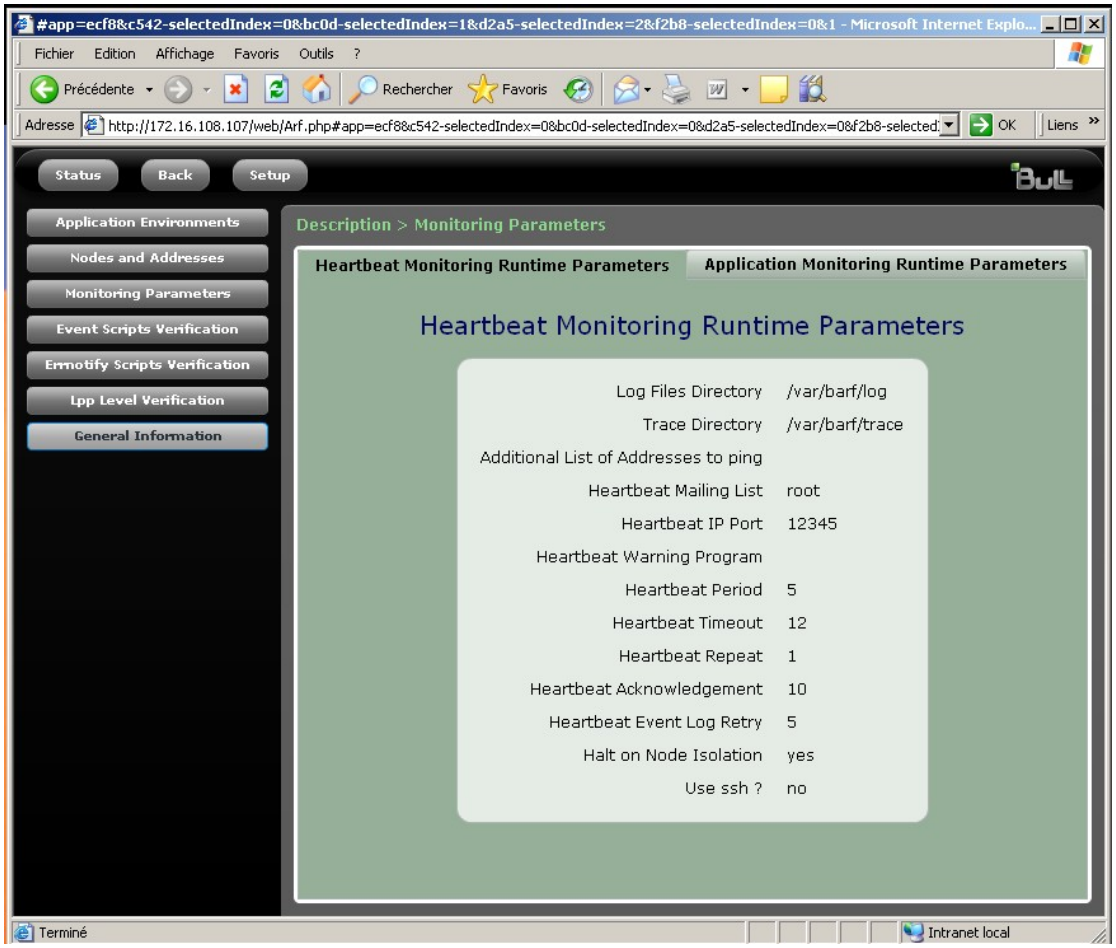
The first array shows the IP addresses and the IP labels configured in *Application Roll-over Facility* for each node.

If Virtual IO Servers are defined in the configuration, their IP addresses and IP labels are displayed.

If HMCs are defined in the configuration, their IP addresses and IP labels are displayed

An optional array gives information on serial network if PPP is used: the tty device for each node.

16.8.3. Monitoring Runtime Parameters



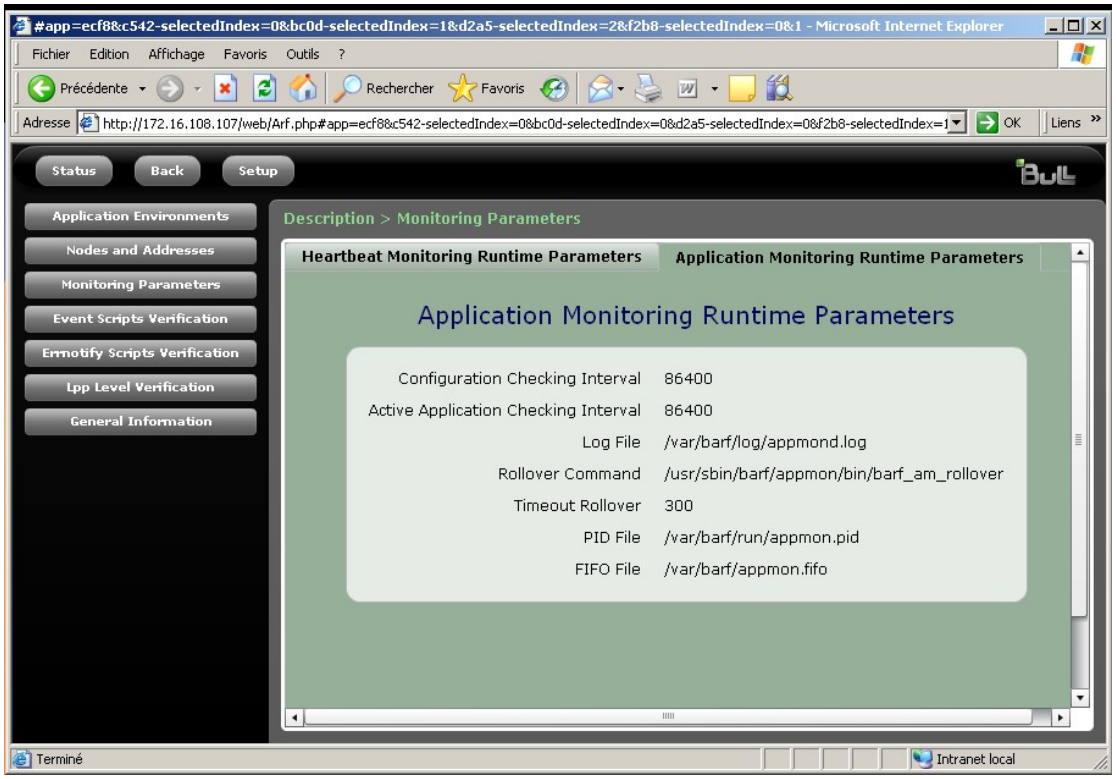


Figure 52. Heartbeat and Application Monitoring Runtime Parameters

This pages display the values of heartbeat runtime parameters (trace and log files, heartbeat parameters, etc) and application monitoring runtime parameters (checking interval, log file, rollover command, etc).

16.8.4. Event Script Verification

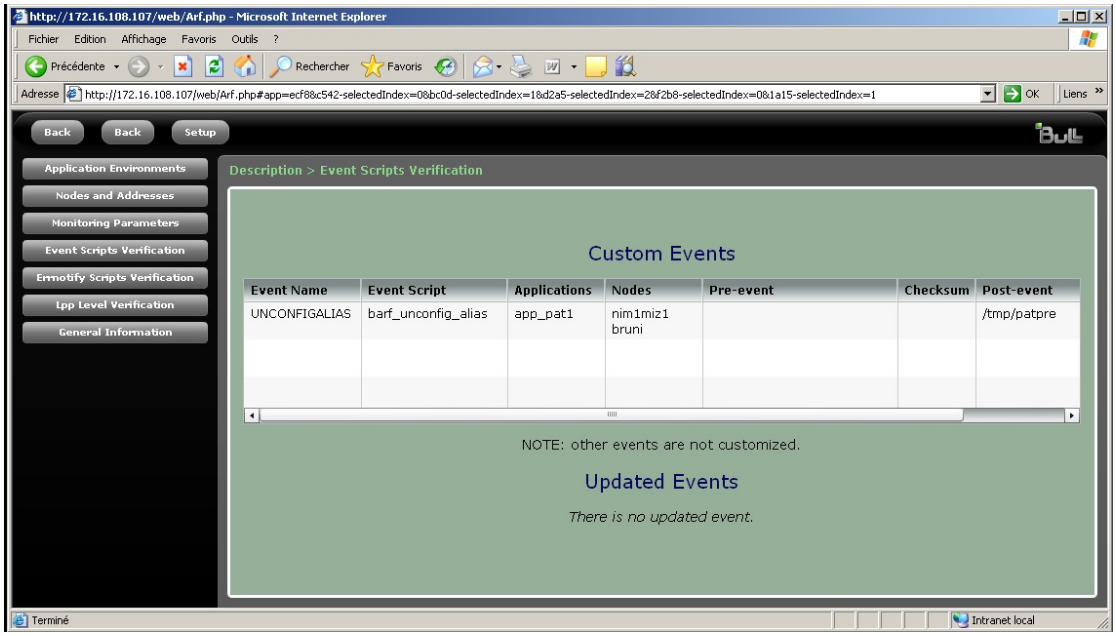


Figure 53. Event Script Verification page

The first array displays customized events. An event script is customized if at least one of the following condition is true:

- A pre-event has been added
- A post-event has been added.

The second array displays updated events. A script event is updated if the script has been modified since its installation. The result is a list of customized/modified events. This list includes the following information:

- The event name
- The event script
- The nodes that support this customization
- The application concerned
- The path of the pre-event and its checksum
- The path of the post-event and its checksum.

16.8.5. Errnotify Scripts Verification

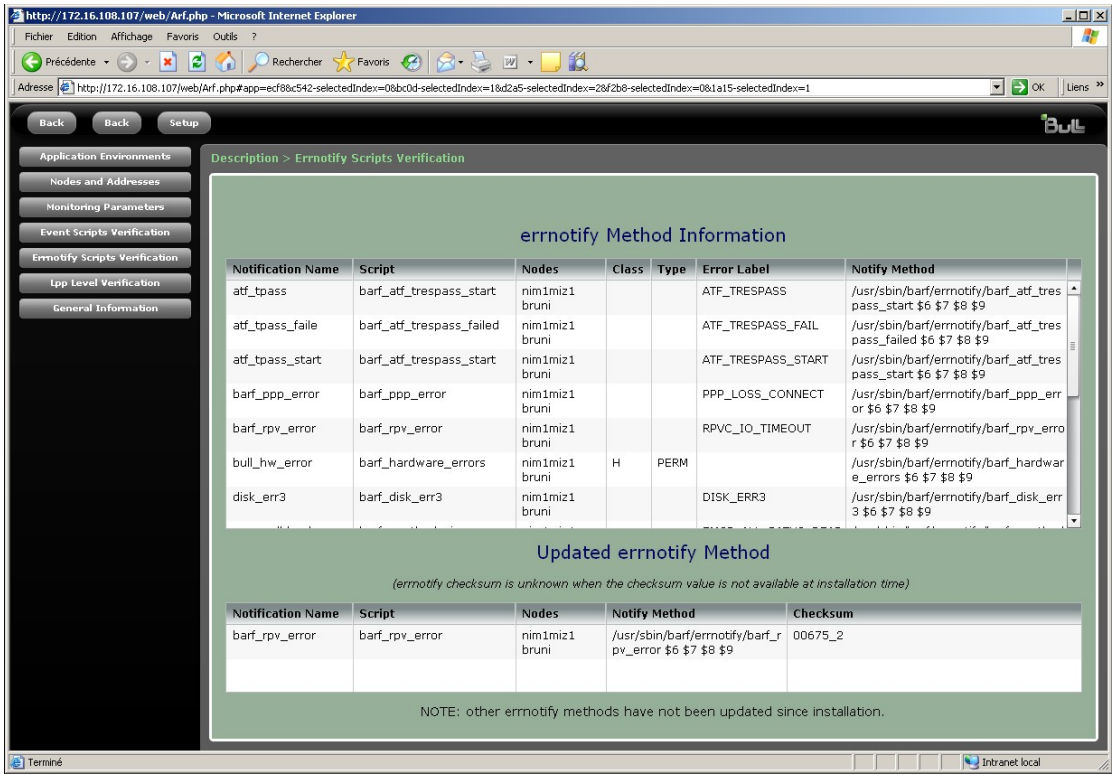


Figure 54. Errnotify Script Verification page

The first array displays standard errnotify methods and their characteristics:

- The errnotify method name
- The errnotify script
- The nodes which have this errnotify method
- The attributes of the errnotify method: Persistent flag, Process ID, Error Class, Error Type
- The Error Label
- The path and parameters of the errnotify method.

The second array displays updated errnotify methods since their installation. The result is a list of modified errnotify methods. This list includes the following information:

- The errnotify name
- The errnotify script
- The nodes where the method has been modified
- The errnotify path and checksum.

16.8.6. LPP Level Verification

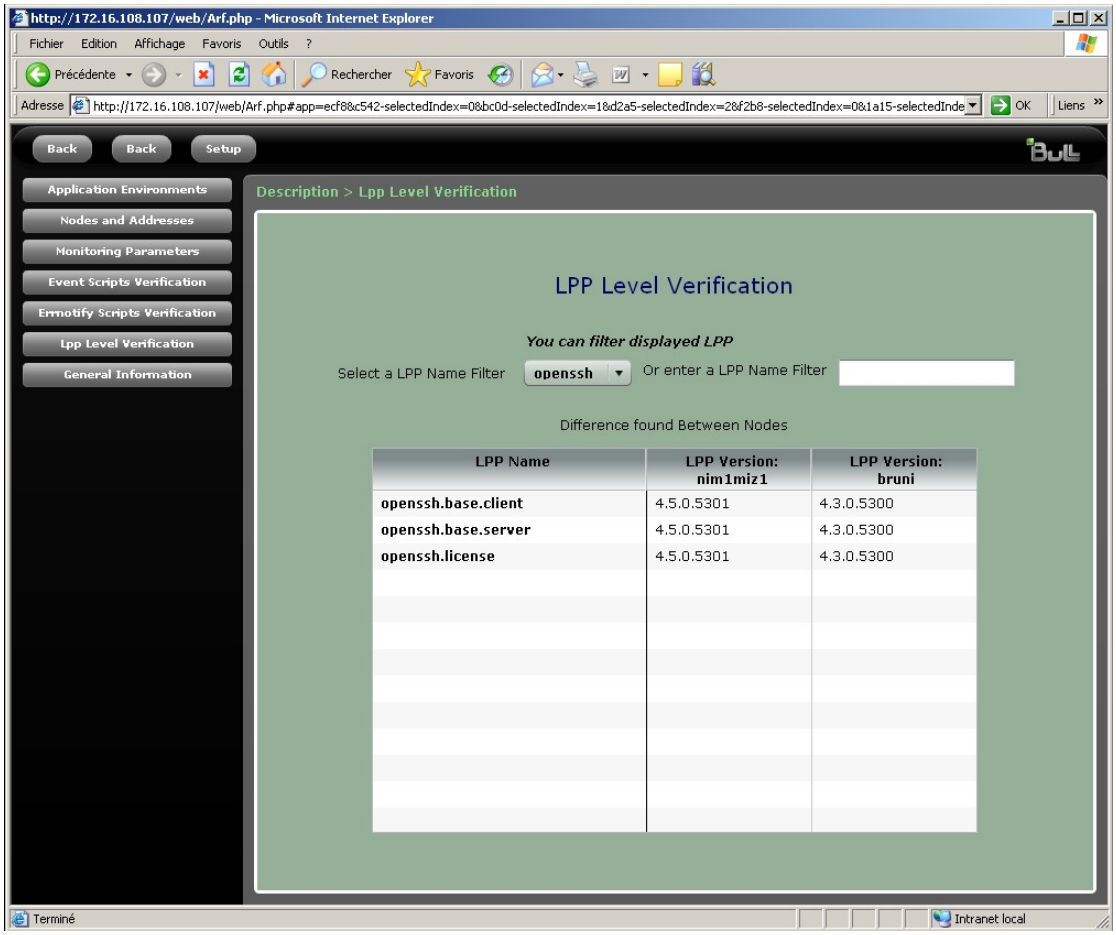


Figure 55. LPP Level Verification page

The array displays lpp for which different versions are installed on the *Application Roll-over Facility* nodes. For each lpp and for each node, the version number is displayed if the lpp is installed.

For each node, the lpp information is the result of the **lslpp** command. Then, a form offers the possibility of choosing a filter. This is a combo box proposing a list of current filters used for lpp selection.

16.9. Setup Function

16.9.1. Password Management

If you want to modify the default password (**arfw**), enter the new one twice in the form and click the **Submit** button. The ARF Watch password will be updated on the local *Application Roll-over Facility* node.

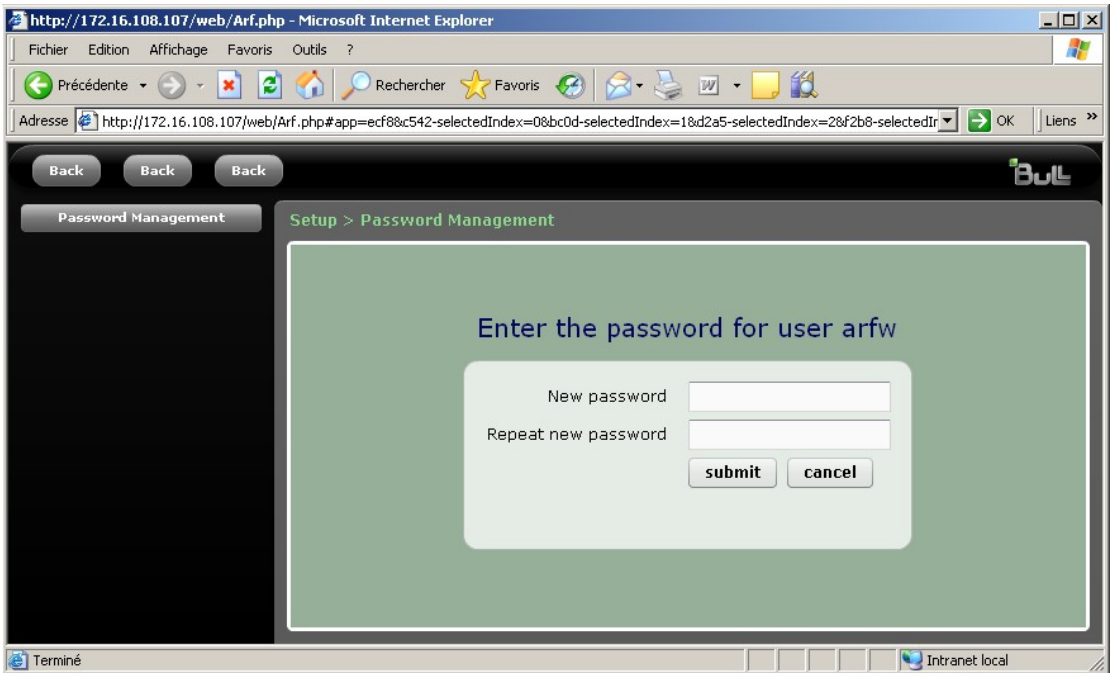


Figure 56. Password Management page

16.10. ARF Watch Administration and Troubleshooting

16.10.1. ARF Watch Password

Using the **ARF Watch Setup – Password Management** page, you can modify ARF Watch password. The default password set at installation time is **arfw**.

If you want to suppress the password prompt at ARF Watch startup, edit the Web server configuration file (**/usr/local/apache2/conf/extra/httpd-arfw.conf**) and make the following changes:

1. Comment the line:

   ```
   AllowOverride AuthConfig
   ```

2. Add the line:

   ```
   AllowOverride None
   ```

3. Restart the web server:

   ```
   /usr/local/apache2/bin/apachectl restart
   ```

User Authentification

ARF Watch relies on the authentication scheme of the Web Server (based on Apache2). In this scheme, one user is defined (**arfw** user). The Web Server is configured automatically for user authentication when ARF Watch is installed.

The **arfw** user is authorized to perform all actions on ARF Watch. When you access ARF Watch and you are prompted for a User ID and a Password, you must always specify **arfw** as the User ID. **arfw** has a specific password.

The password is saved in a file named **.mdp** in **/usr/sbin/barf/arfw/gui/web**. The **htpasswd** command is used to create the password file and to add or modify users:
htpasswd -bc /usr/sbin/barf/arfw/gui/web/.mdp usersname password

This is done automatically for **arfw** user when updating password in ARF Watch page.

16.10.2. Dealing with the Web Server

Web Server start/stop

The Web Server must be running on *Application Roll-over Facility* nodes.

- Check that the Web Server is running: run the **ps** command, and check the process **/usr/local/apache2/bin/httpd**.
- If the Web server is not running, you can start it by the command:

  ```
  /usr/local/apache2/bin/apachectl start
  ```

- If the Web server is running but you have modified its configuration file, you have to restart it by the command:

  ```
  /usr/local/apache2/bin/apachectl restart
  ```

- If the Web Server does not start or if you have problems with some ARF Watch pages, you can examine the Web Server log file **/usr/local/apache2/logs/error_log**, and try to correct the problem.
- To stop the server, run the command:

  ```
  /usr/local/apache2/bin/apachectl stop
  ```

Web Server configuration for ARF Watch

When ARF Watch software is installed, the Web Server is installed and restarted to take into account specific configuration for ARF Watch (user, directory index, alias, password management ...).

The **httpd-arfw.conf** file, under **/usr/lpp/Bull.approllf.arfw** directory, contains all needed information for Web Server configuration. This file is automatically copied in **/usr/local/apache2/conf/extra** at installatin time.

This file is included automatically in the standard Web configuration file (**httpd.conf**) under **/usr/local/apache2/conf** directory
(directive **Include /usr/local/apache2/conf/extra/httpd-arfw.conf**).

- Content of **httpd-arfw.conf**:

```
User arfw
Group nobody
DocumentRoot "/usr/sbin/barf/arfw/gui"
<Directory />
    Options FollowSymLinks
    AllowOverride AuthConfig
    Order deny,allow
    #Deny from all
</Directory>
Alias /arfw /usr/sbin/barf/arfw/gui/index.html
Alias /testphp /usr/sbin/barf/arfw/gui/web/phpinf.php
```

Chapter 17. Viewing the Configuration

This chapter describes the tasks that enable the administrator to:

- display the configuration
- show the applications status
- show the monitoring status.

17.1. Displaying Configuration

To view the configuration, perform the following procedure :

1. Type **smit barf** and select the following menus: **Configuration Definition> Display Configuration**.
2. When you press Enter, SMIT displays the following screen:

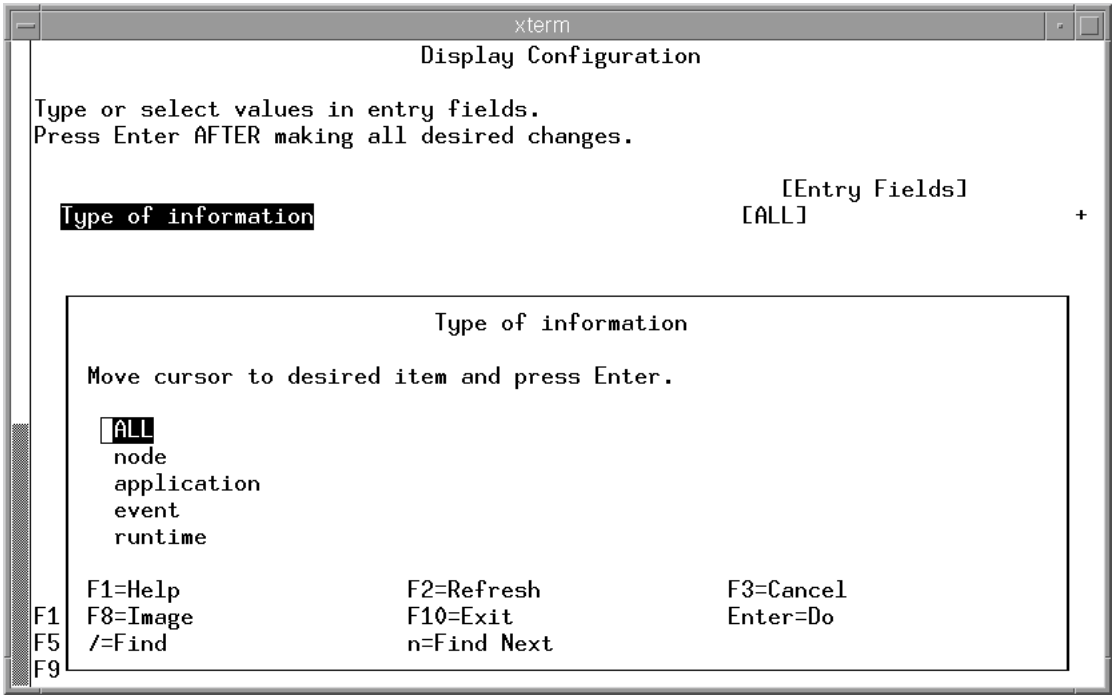


Figure 57. Display Configuration Menu

Type of information — enables you to choose which view of the cluster you want to see. Select one of these options:

ALL	to display complete information about the configuration including the nodes, their IP address, the applications, the runtime parameter and the events.
node	to display the name of all the machines included in the configuration and their IP adresses.
application	to display information about the different applications included in the configuration.
event	to display information about the different custom events configured.
runtime	to display information about the runtime parameters.

17.2. Showing Application Status

To display the status of the different applications, perform the following procedure :

1. Type **smit barf** and select the following menus : **Manage Application Environment > Show Application Status** or use the **smit barf_show_appstate** fastpath.
2. When you press Enter, SMIT displays the following screen:

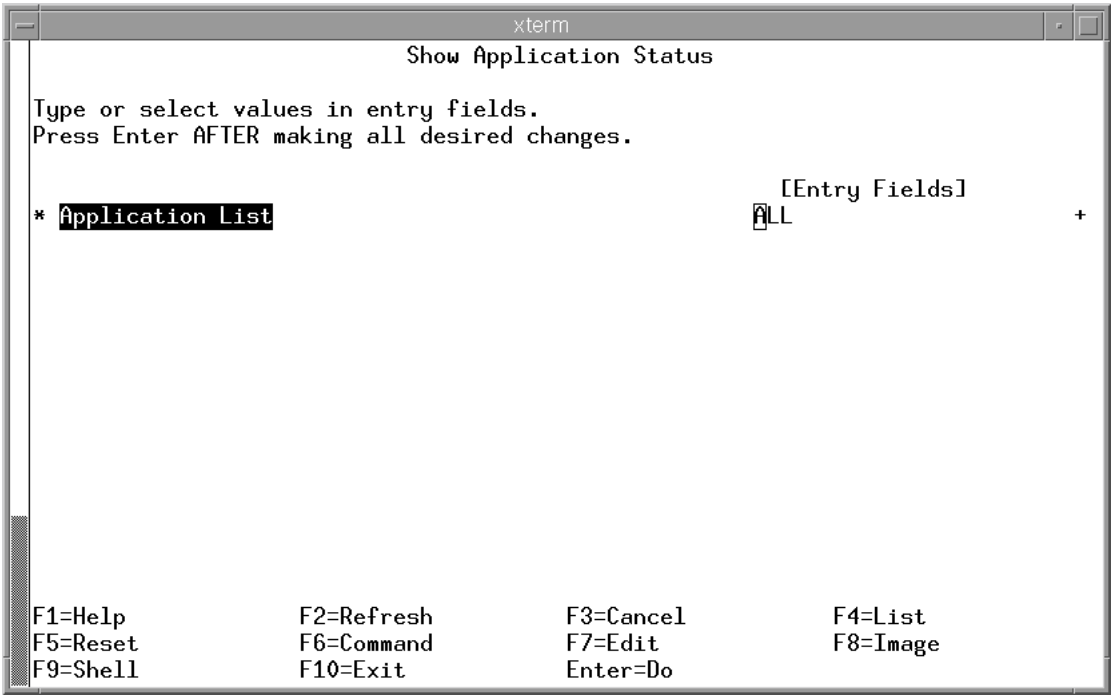


Figure 58. Show Application Status screen

Application List enter `ALL` or the name of one or more applications.

3. Press Enter to validate your selection. The following information is displayed:
   - The application status: its name, if it is running or not, on which node it runs, the take-over mode and the list of take-over nodes.
   - The application resources: IP address, the name of the volume group and file systems, and their status (varyon or mounted).

The two following screens illustrate how the application status is displayed.

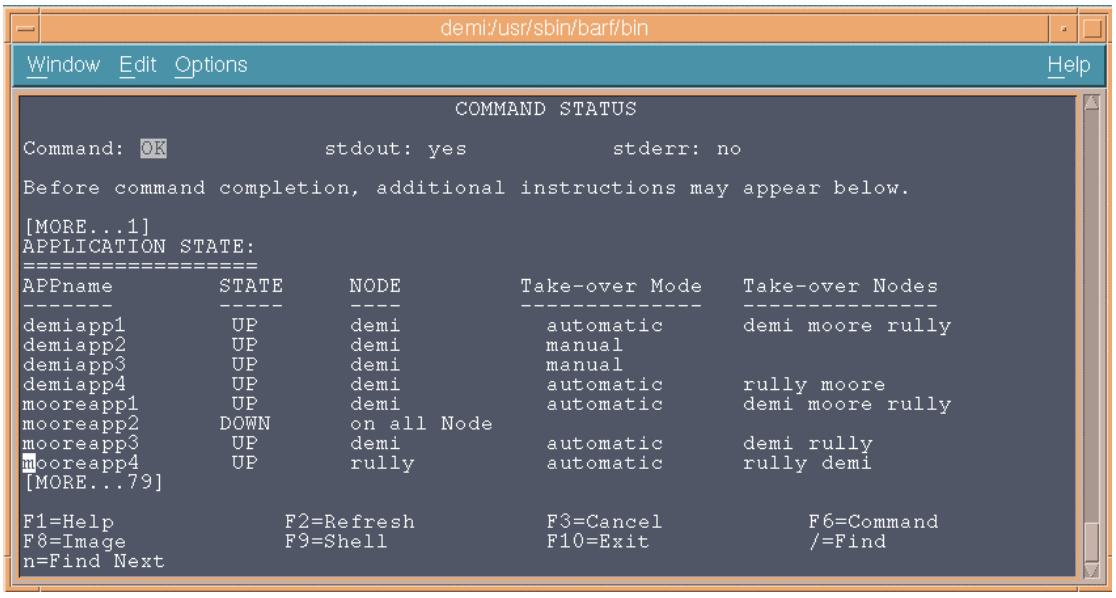


Figure 59. First example of application status

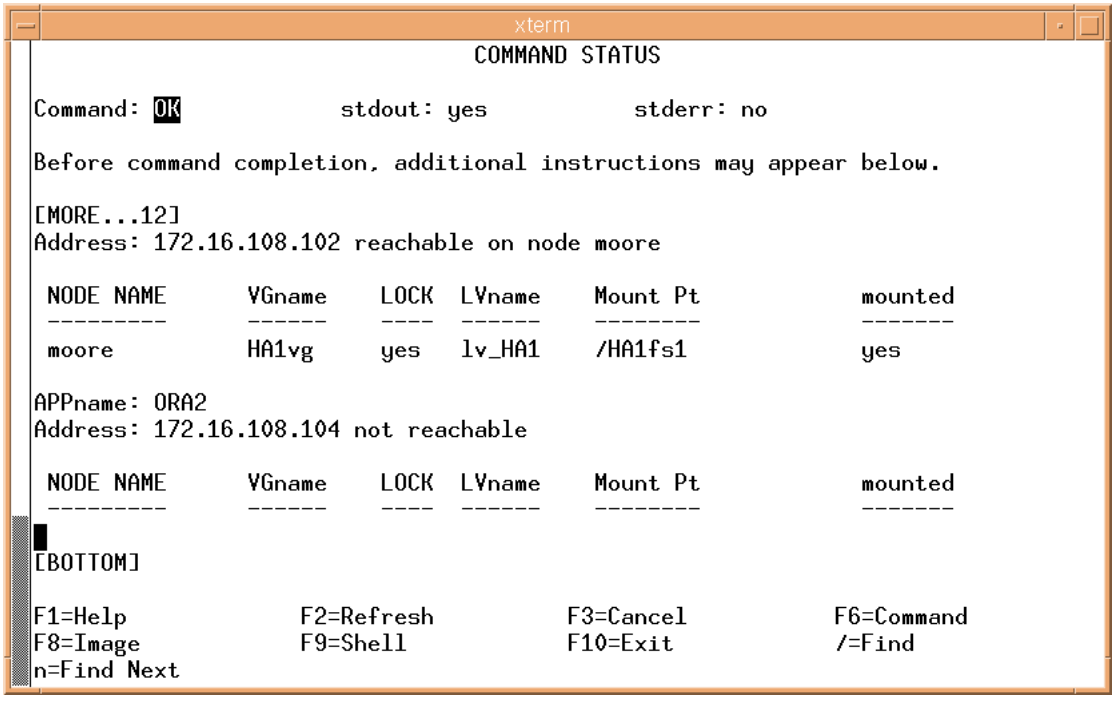


Figure 60. Second example of application status

17.3. Showing Monitoring Status

To know if the node monitoring daemon is running or not, perform the following procedure :

1. Type **smit barf** and select the following menus: **Manage Monitoring > Show Monitoring Status** or use the **smit barf_monitoring_menu** fastpath.
2. When you press Enter, SMIT displays the status of the node monitoring daemon (active or not active).

Chapter 18. Maintaining the ARF Environment

This chapter describes the tasks you must perform to maintain an *Application Roll-over Facility* environment:

18.1. Starting and Stopping ARF Services

Starting *Application Roll-over Facility* services means starting applications and their resources, and activating node monitoring (the node monitoring enables the coordination required between nodes). Starting *Application Roll-over Facility* services on a node also triggers the execution of certain *Application Roll-over Facility* scripts.

Stopping *Application Roll-over Facility* services means stopping applications and their resources, and de-activating node monitoring between the nodes.

Before starting or stopping *Application Roll-over Facility* services, you must have created an *Application Roll-over Facility* configuration.

WARNING

You must have a valid license key on all the managed nodes. If the license is not valid, the Application Roll-over Facility Services won't start.

18.1.1. Activate / De-activate Node Monitoring Services

You can activate the node monitoring even if no application is running. System administrator can start this daemon at each reboot, using SMIT menus or, better, automatically at system boot. The following line will be added to the **/etc/inittab** file:

```
arf_monitor:2:once:/usr/sbin/barf/bin/hbarf_activate_monitoring
```

To exchange heartbeats, the node monitoring daemons use each of the IP networks defined in Managed nodes configuration.

If a disaster is diagnosed by the node monitoring daemon, an appropriate alert is sent to the system administrator.

WARNING

Before activating the node monitoring, you must have a valid license key on all the managed nodes. If the license key is not valid, the node monitoring daemon won't start.

Start Node Monitoring daemon

To activate node monitoring on a node, perform the following procedure:

1. Type **smit barf** and select the following options: **Manage Monitoring > Activate Monitoring** or use the **smit hbarf_activate_monitoring** fast path. When you press Enter, the following screen appears:

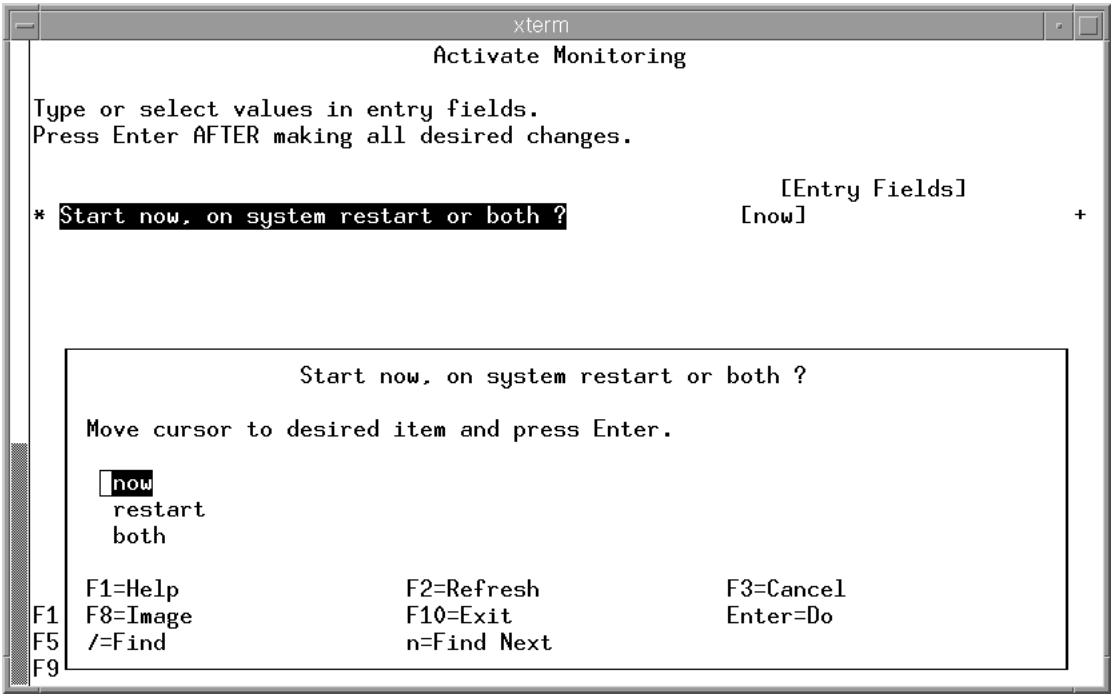


Figure 61. Activate Monitoring screen

Start now, on system restart or both ?
It is recommanded to choose '**restart**' or '**both**': the node monitoring daemon will be started after each system reboot. The following line will be added to the **/etc/inittab** file:

```
arf_monitor:2:once:/usr/sbin/barf/bin/hbarf_activate_monitoring
```

2. Press Enter to validate your selection.
3. Press F10 to exit SMIT.

Stop Node Monitoring daemon

To de-activate node monitoring on a node, perform the following procedure:

1. Type **smit barf** and select the following options: **Manage Monitoring > De-Activate Monitoring** or use the **smit hbarf_deactivate_monitoring** fast path. When you press Enter, the following screen appears:

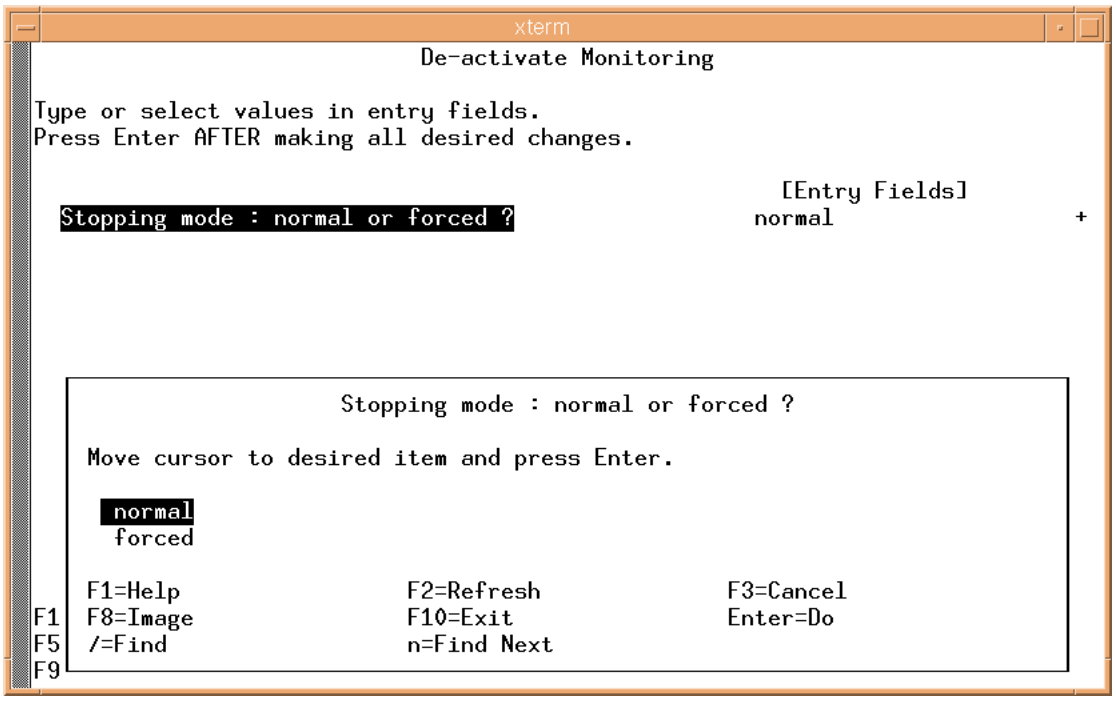


Figure 62. De-activate Monitoring screen

Stopping mode Enables you to choose the way the node monitoring is stopped. Select 'forced' when the configuration is in a critical state, in other cases select 'normal'.

2. Press Enter to validate your selection.
3. Press F10 to exit SMIT

18.1.2. Log for Node Monitoring

This log is an event history about the heartbeat mechanism. Its default pathname is:

/var/barf/log/clsmd.log

The structure of the event history file is as follows:

```
clsmd: date hour:minute:second year: daemon status: message text
```

For example:

```
clsmd: Wed Aug 23 10:43:13 2000: 10: node1: daemon monitoring
node2 started.
```

The daemon status are described in *Troubleshooting*, on page 22-1.

To change the log file directory, use the **Change/Show Runtime Parameters** SMIT menu (see *CustomizingLog and Trace Files*, on page 6-9).

For more information about the **clsmd.log** file, see *Understanding the clsmd.log file,* on page 22-2.

18.1.3. Activate and De-Activate Applications on Nodes

As explained in Chapter 1 *Concepts and Components Overview* the **Application Relocation** function is used to move applications from one node to another, either automatically (in case of node failure) or manually under administrator control.

You can activate the applications even if the node monitoring is not running, but in this case, only the manual take-over mode is authorized.

WARNING

Before activating the applications, you must have a valid license key on all the managed nodes. If the license key is not valid, the activation fails.

Activate Applications

When an application is activated on a node, the file /var/barf/ barf_activate_appli.status is updated with the name of the application.

It is possible to activate a list of applications, simultaneously, on a same node.

To activate applications and their resources on a node, perform the following procedure:

1. Type **smit barf** and select the following options: **Manage Application Environment > Activate Application** or use the **smit barf_activate_appli**. fast path. When you press Enter, SMIT displays the following screen:

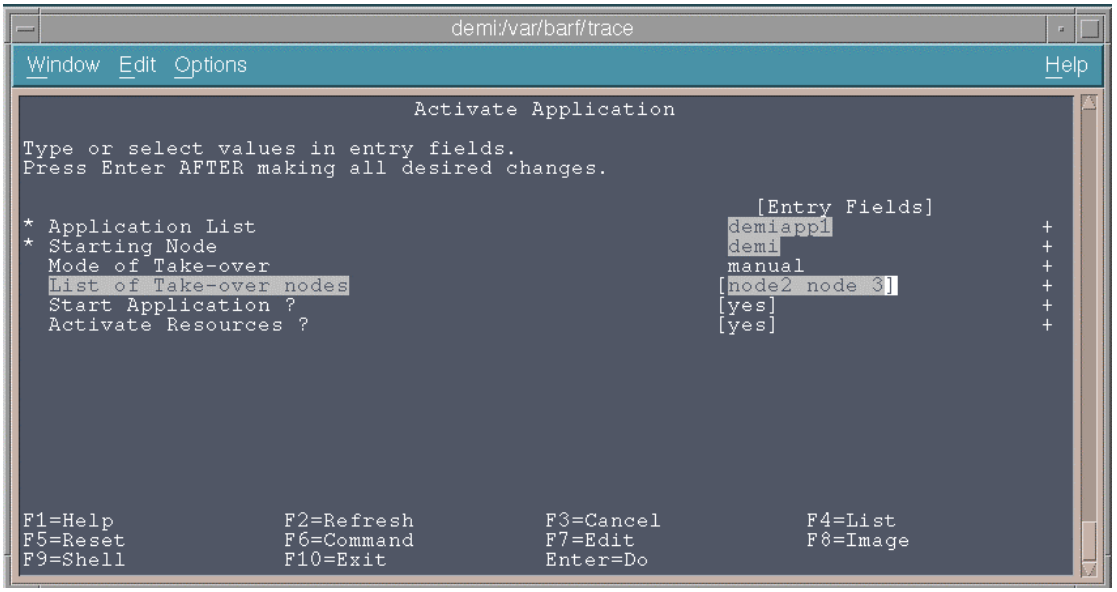


Figure 63. Activate Application screen

Application List	List of all the applications you want to start.
Starting Node	Node on which you want to start these applications.
Mode of Take-over	Mode to recover the application: either automatic or manual. Default Value: manual.
List of Take-over nodes	Ordered list of nodes on which the application(s) have to be recovered. Each node is separated by a blank. If no node is set, the default list is the Nodes list defined in the Add or Change/Show Application Environment menu (list of nodes on which the application can be running). It is sorted in alphabetical order.
Start Application ?	If 'yes', the starting script of the applications will be run.
Activate Resources ?	if 'yes', all the resources of all the applications (ex. volume group, file system) will be up.

2. Press Enter to validate your selection.
3. Press F10 to exit SMIT

De-activate Applications

You can de-activate an application only on the node where the application is running.

To de-activate applications on the local node, perform the following procedure:

1. Type **smit barf** and select the following options: **Manage Application Environment > De-Activate Application** or use the **smit barf_de_activate_appli** fast path. When you press Enter, SMIT displays the following screen:

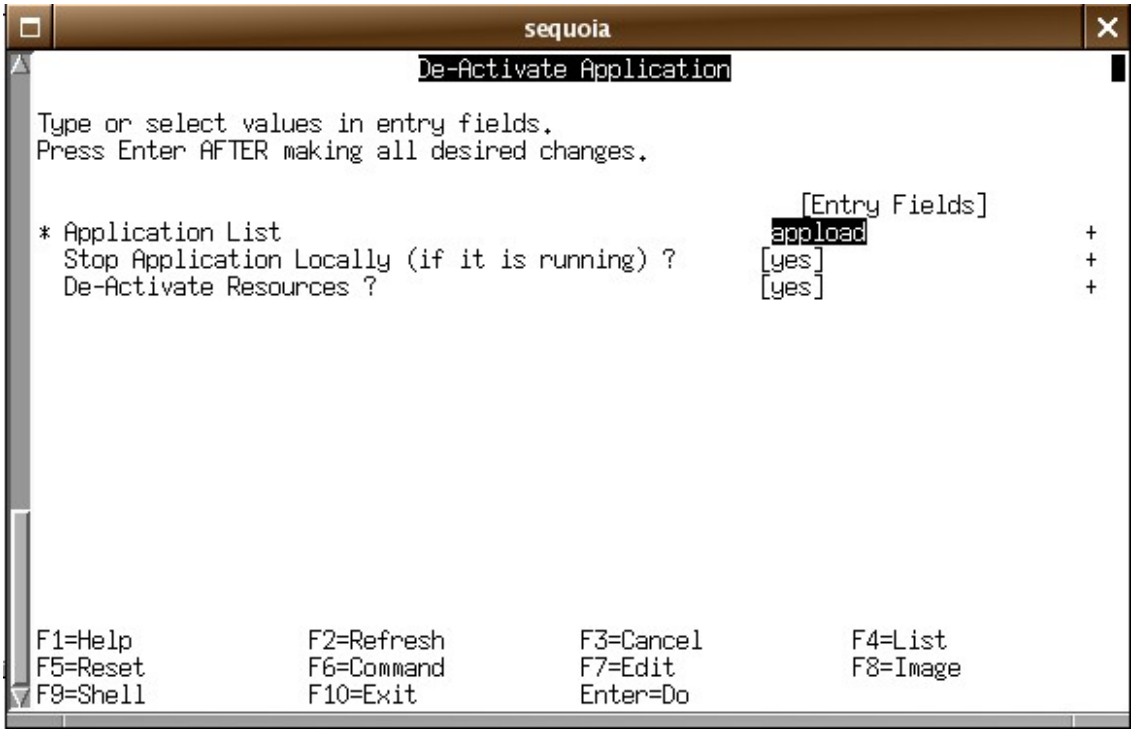


Figure 64. De-activate Application screen

Application List	List of all the applications you want to stop.
Stop Application ?	If 'yes', the stopping script of the application will be run.
De-Activate Resources ?	If 'yes', all the resources of all the applications (volume group, file system,...) will be down.

2. Press Enter to validate your selection.
3. Press F10 to exit SMIT

Roll-over Applications

You can move applications from one node to another node either automatically or manually under the administrator's control. You can move an application to another node only from the node where the application is running.
To roll-over an application from the node where the application is running to another node, run the following procedure:

1. Type **smit barf** and select the following options: **Manage Application Environment > Roll-over Application Environment** ou use the **smit barf_deactivate_appli** fast path. When you press Enter, SMIT displays the following screen:

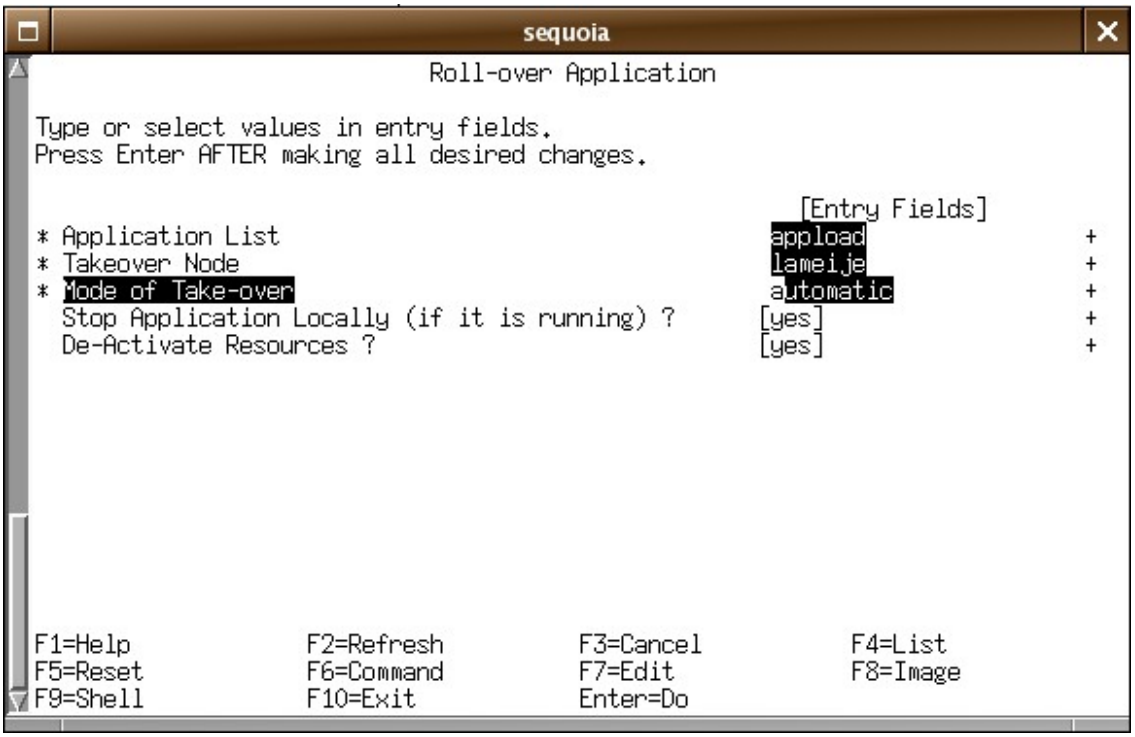


Figure 65. Roll-over Application screen

Application List	The name of applications to move to an other node.
Takeover Node	Node on which to re-start these applications.
Mode of Take-over	Mode to recover the application: either **automatic** or **manual**. Default value: **manual.**
Stop Application Locally?	If 'yes', the application is stopped.
De-Activate Resources?	If 'yes', all the resources of all the applications(volume group, filesystem...) will be down.

2. Press Enter to validate the selection.
3. Press F10 to exit SMIT.

18.2. Maintaining Shared Volume Groups on Different Nodes

It is possible to import a volume group on a node even if this VG is already used on another node of the configuration. Follow these steps:

1. On the node that currently owns the volume group, make the necessary changes to the volume group. For example add or remove a logical volume or a filesystem.
2. Remove the reserve on the shared volume group:

   ```
   varyonvg -b -u SHAREDVG
   ```

3. On the other nodes, in the resource group that currently does not own the volume group, update ODM and other files:
   - if the volume group is already imported on the node:

     ```
     importvg -L SHAREDVG hdiskY
     ```
     (or `hdiskpowerY`)
   - if it is a new volume group (never imported to the node):

     ```
     importvg -y <VGname> -n -F hdisk<x>
     ```
     (or `hdiskpower<x>`)

   This command will not activate the volume Group. Do not use smit importvg.

   ```
   varyonvg -u <VGname>
   chvg -an _Qn <VGname>
   varyoffvg <VGname>
   ```

4. On the node that currently owns the volume group, restore the reserve on the initial node:

   ```
   varyonvg SHAREDVG
   ```

18.3. Managing a Node Failure

In the configuration illustrated by Figure 66, Application 1 runs on Node1.

Automatic failover mode has been activated by running the following menu (from any node):

```
smit barf > Manage Application Environment > Activate Application
Environment

* Application List              [Application1]
* Starting Node                 [Node1]
Mode of Take-over               [automatic]
List of Take-over nodes         [Node2]
Start Application ?             [yes]
Activate Resources ?            [yes]
```

Let's assume that Node1 fails:

Since automatic failover mode was acivated, the applications will be taken over by the next node in the take-over list (Node2).

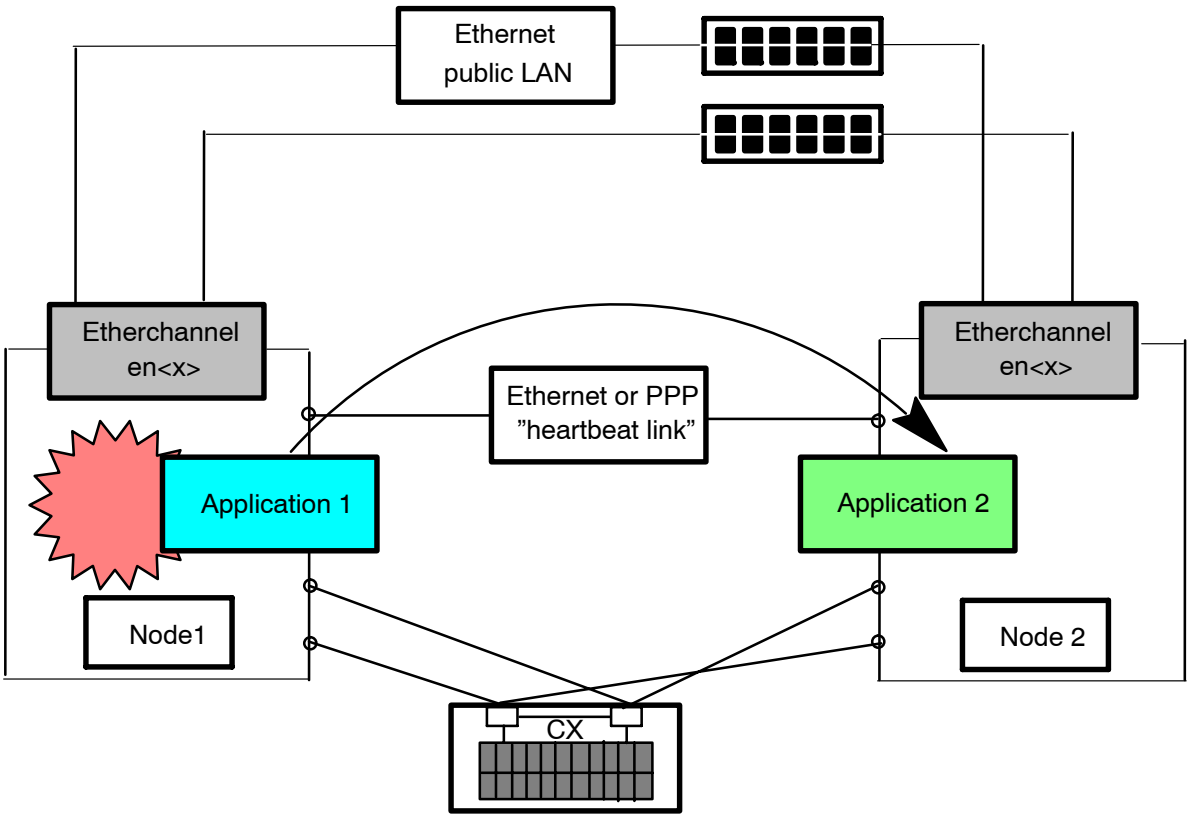


Figure 66. Application take over in case of Node failure

Let's assume that Node 1 is repaired then restarted (monitoring is active).

The applications normally belonging to Node 1 (for example Application 1) must be MANUALLY stopped on Node2:

From Node2 run the following menu:

```
smit barf > Manage Application Environment > De-activate
Application Environment
```

Then roll-over the application on Node1 by running the following menu from Node2:

```
smit barf > Manage Application Environment > Roll-over
Application Environment

* Application List                                    [Application1]
Takeover Node                                        [Node1]
Mode of Take-over                                    [automatic]
Stop Application  Locally (if it is running) ?       [yes]
De-Activate Resources ?                              [yes]
```

Note In case of AIX mirrored VGs (*Application Roll-over Facility* 3.0), the content of VGs belonging to Application2 will be also re-synchronized.

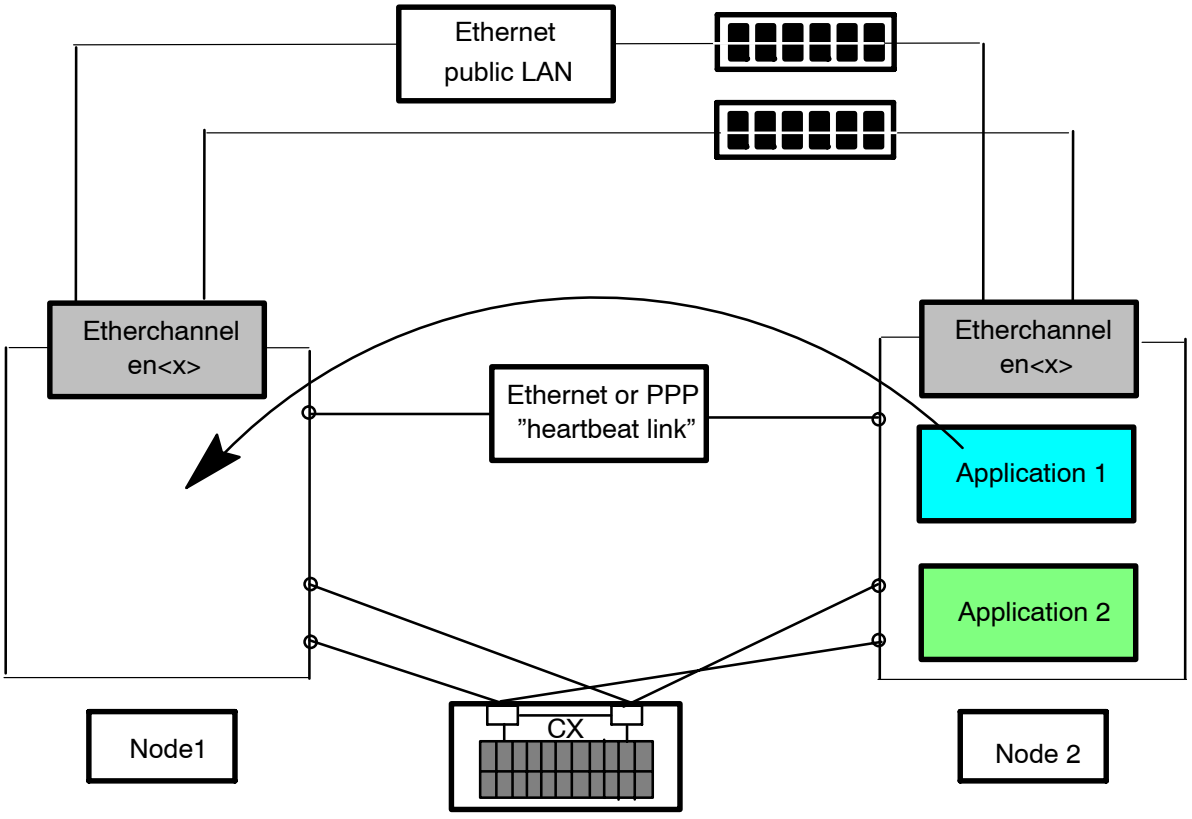


Figure 67. Application recovery after node repair

Note In this case, the Takeover Node is a Destination Node. Specifying 'automatic' means that if Destination Node (Node1) fails, Application1 will fail over to Node2 which was specified as Takeover Node when Application1 was initially started on Node1.

18.4. Managing an Ethernet Network Failure

18.4.1. Failure of Ethernet Network Monitored by ARF

Check on each node that the monitoring is active:

```
vi /var/barf/log/clsmd.log

clsmd: Tue Jan 20 17:01:18 2004: 2: marie: 1.0.0.5: node daemon
monitoring galante: 1.0.0.20
```

On a node, disconnect the Ethernet cable (2 cables if Etherchannel) for the network monitored by *Application Roll-over Facility*.

Check that the resources remain in place: VGs, FS, IP aliases and applications.

Check that the defect is registered in the /var/barf/log/clsmd.log file of each node, and that it corresponds to the failed IP address of the network (address of the unplugged Ethernet cable) declared in *Application Roll-over Facility* configuration.

Expected results

- For example, following is an extract of the log file one of the nodes:

```
vi /var/barf/log/clsmd.log

clsmd: Thu Jan 22 10:19:48 2004: 85: marie: ERROR detected. A
network problem occurs between the local IP address 1.0.0.5 and
the remote IP address 1.0.0.20
```

- Check with the `netstat –i` command that the failing network corresponds to the network registered in the /var/barf/log/clsmd.log file.

18.4.2. Repair of Ethernet Network Monitored by ARF

Plug back the cable you previously unplugged.

The monitoring of the previously failed network will automatically start again as follows:

```
clsmd: Fri Jul 30 13:39:04 2004: 41: marie: 1.0.0.5 is receiving
again heartbeats from galante: 1.0.0.20
```

18.5. Changing the ARF Topology

Changing the topology means modifying information about the nodes of the configuration. You can change the application configuration or migrate application to other nodes dynamically using SMIT menus.

To change the topology you must stop and restart *Application Roll-over Facility* services. You must propagate the configuration across all the nodes after each change.

When you configure an *Application Roll-over Facility* cluster, configuration data is stored in specific object classes in the ODM. The AIX ODM object classes are stored in the default system configuration directory: **/etc/objrepos**.

The *Application Roll-over Facility* system administrator may need to perform any of the following tasks relating to nodes:

- Adding one or more nodes
- Removing a node
- Changing the attributes of a node.

18.5.1. Adding a Node

You can add a node to an active *Application Roll-over Facility* configuration. You do not need to stop and restart *Application Roll-over Facility* services for the node to become part of the configuration.

To add a node, refer to the procedure described in *Defining Nodes*, on page 6-4.

18.5.2. Removing a Node

You cannot remove a node from an active configuration. Before removing a node from the configuration, you must follow these steps:

1. Stop the *Application Roll-over Facility* services on the node to be removed. It means to stop the applications and the monitoring if they are running.
2. Remove the node from any application in which it participates. See *Changing the Application Environment*, on page 18-13 for more information.
3. On the local node, remove the node from the topology definition.
4. From the main *Application Roll-over Facility* SMIT screen, select the following options: **Configuration Definition > Managed Node > Remove a Managed Node** or use the **smit barf_conf_manage_menu** fast path. SMIT displays the list of all *Application Roll-over Facility* nodes.
5. Select the node you want to remove and press Enter. SMIT displays a popup message asking if you are sure you want to proceed. Press Enter again to remove the node from the topology definition.
6. Synchronize the topology and the applications using SMIT menus. When the synchronization completes, the node is removed from the topology definition.

18.5.3. Changing the IP Address List of a Node

You cannot change the IP address of a node from an active configuration. You must first stop the *Application Roll-over Facility* services on the node; it means stopping the applications and the monitoring if they are running.

To change the IP address of a node, perform the following procedure:

1. From the **Configuration Definition** menu select the **Manage Nodes** option and press Enter.

2. Select Change/Show a Managed Node option and press Enter. SMIT displays a pick list of nodes.

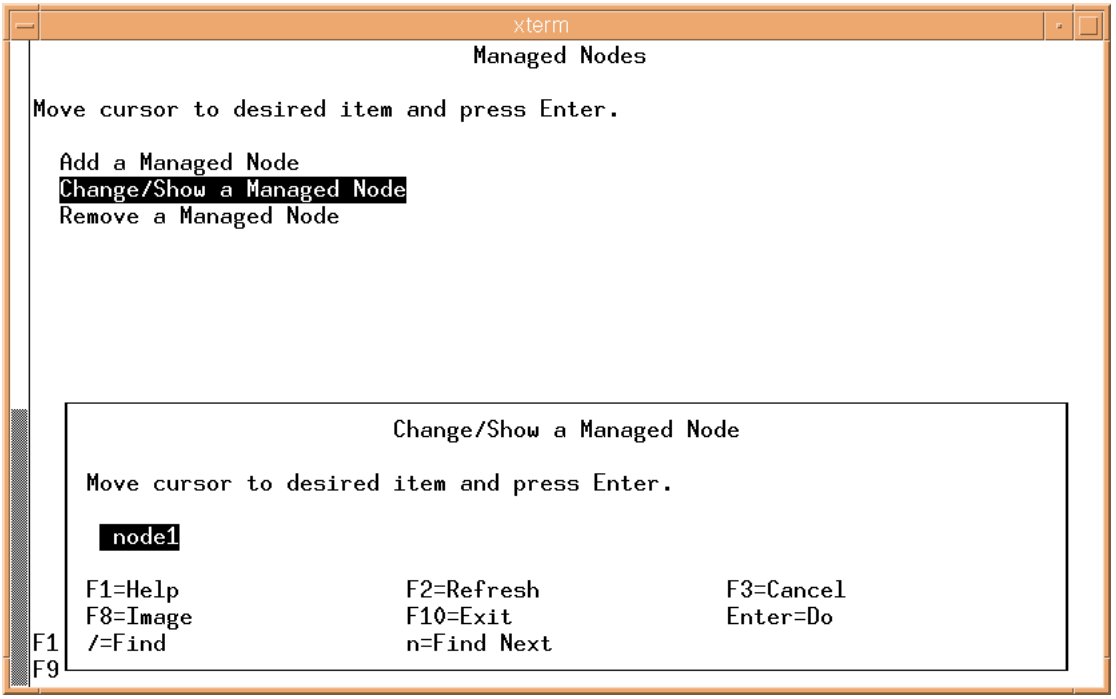


Figure 68. Managed Nodes screen

3. Select a node and press Enter. SMIT displays the following screen :

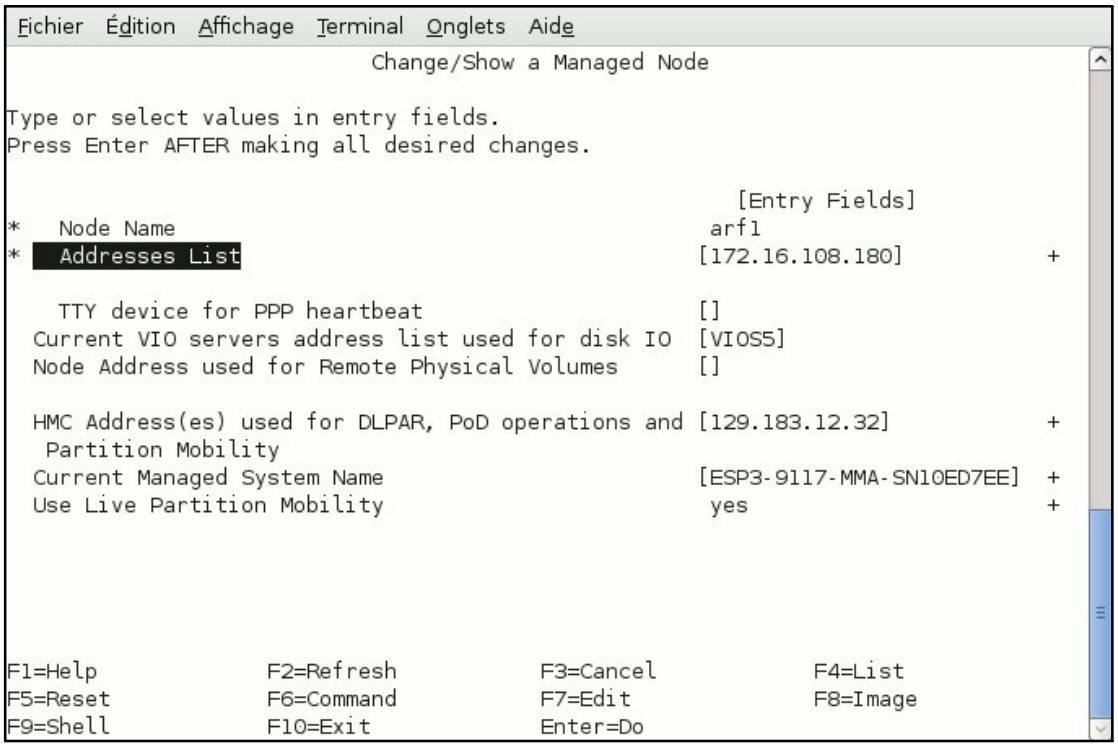


Figure 69. Change/Show a Managed Node screen

Enter the new adresses (or complete those existing) for the node in the **Addresses List** field. Then press Enter to validate the new addresses.

4. After the command completes you must synchronize the topology using SMIT menus.

18.6. Changing the Application Environment

The system administrator may need to perform any of the following tasks relating to applications:

- Adding one or more application
- Removing an application
- Changing the resources of an application

18.6.1. Adding an Application

You can add an application to an active configuration. You do not need to stop and restart *Application Roll-over Facility* services.

To add a new application, refer to *Adding Applications*, on page 6-4.

When a new application is added you must synchronize the topology using SMIT menus.

18.6.2. Changing an Application

You cannot make changes to a running application. You must first stop the application on the node where it is running.

Note See *Showing Application Status*, on page 17-2 to know if an application is running.

Then, to change an application, perform the following steps:

1. Type **smit barf** and select the following options: **Configuration Definition> Application Environment > Change/Show Application Environment** or use the **smit barf_ch_appli** fasthpath. When you press Enter, SMIT displays the following screen:

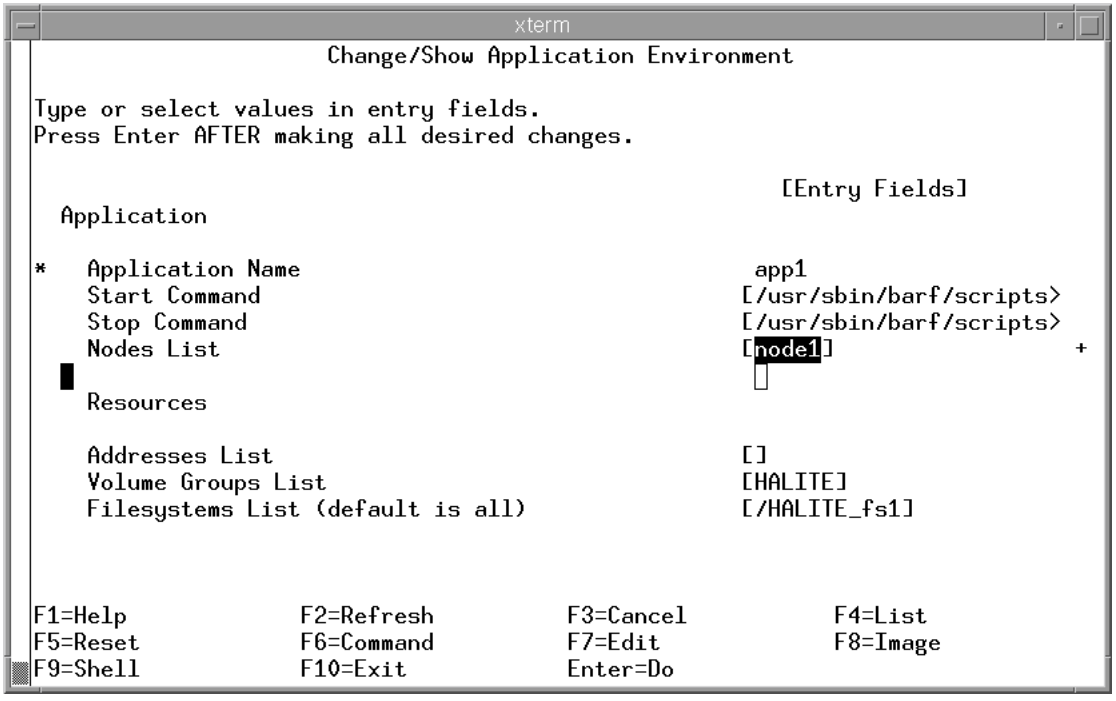


Figure 70. Change/Show Application Environment screen

Application Name — Enter an ASCII text string that refers the application. This name can include alphabetic and numeric characters and underscores. Use no more than 31 characters.

Start Command	Enter the pathname of the script (followed by arguments) called by the *Application Roll-over Facility* event scripts to start the application. This script must be in the same location on each node that might start the application. The contents of the script, however, may differ.
Stop Command	Enter the pathname of the script called by the *Application Roll-over Facility* event scripts to stop the application. This script must be in the same location on each node that may start the server. The contents of the script, however, may differ.
Nodes List	List of the nodes on which the application can be running.
Addresses List	List of the addresses associated at this application.
Volume Groups List	List of all the volume groups that will be used by the application, and will be varied on.
Filesystems List	List of all the filesystems used by the application, and mounted.

2. Press Enter to add this information to the ODM on the local node. Press the F3 key to return to previous SMIT screen and to add other application.
3. After the command completes you must synchronize the topology using SMIT menus.

18.6.3. Removing an Application

You cannot remove a running application. You must first stop the application on the node where it is running.

Note See *Showing Application Status*, on page 17-2 to know if an application is running.

Then, take the following steps to remove an application :

1. From the main *Application Roll-over Facility* SMIT screen, select the following options: **Configuration Definition > Application Environment> Remove Application Environment** or use the **smit barf_conf_app_menu** fast path. SMIT displays a list of all *Application Roll-over Facility* applications.
2. Select the application you want to remove and press Enter. SMIT displays a popup message asking if you are sure you want to proceed. Press Enter again to remove the application from the topology definition.
3. On the local node, return to the SMIT **Configuration Definition** menu and select the **Propagate Topology** option to synchronize the topology. When the synchronization completes, the application is removed from the topology definition.
4. After the command completes you must synchronize the topology using SMIT menus.

18.7. Changing the Custom Pre/Post Events

The *Application Roll-over Facility* system administrator may need to perform any of the following tasks relating to applications :

- Add a Custom Pre/Post-event
- Change/Show Custom Pre/Post-event
- Remove Custom Pre/Post-event.

These tasks can be done dynamically, but if an application is running the changes will be effective the next time the application will be activated.

18.7.1. Adding a Custom Pre/Post Event

You can add custom events to an active configuration, You do not need to stop and restart ARF services, but the changes will be effective the next time the application will be activated. To add a new custom event, refer to *Configuring Custom Events*, on page 6-7.

After the task is performed you must synchronize the topology using SMIT menus.

18.7.2. Change/Show a Custom Pre/Post-event

You can only change and show the script of a custom event. Take the following steps:

1. Type **smit barf** and select the following options: **Configuration Definition> Application Environment > Change/Show Custom Pre/Post-event** or use the **smit barf_ch_custom** fast path.
2. Select the application name and then the event name. When you press Enter, SMIT displays the following screen:

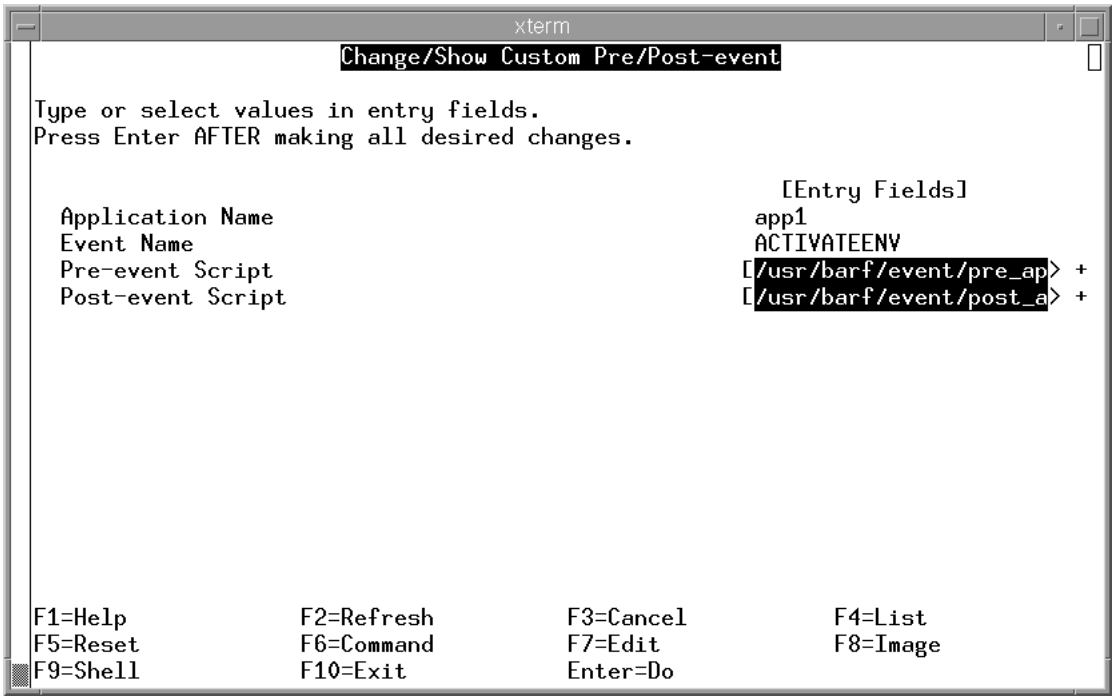


Figure 71. Change/Show Custom Pre/Post-event screen

Pre-Event Script Filename — (optional) If you have defined custom events press F4 for the list. Or enter the name of a custom-defined event to run before the event command executes. This command provides pre-processing before an event occurs. Remember that

	Application Roll-over Facility will not process the event until this pre-event script or command has completed.
Post-Event Script Filename	(optional) If you have defined custom events press F4 for the list. Or enter the name of the custom event to run after the event command executes successfully. This script provides post-processing after an event.

3. Press Enter to change the information to *Application Roll-over Facility* in the local ODM.
4. Repeat this operation for all the post and pre-events you want to change.
5. After the task is performed you must synchronize the topology using SMIT menus.

Note Synchronizing does not propagate the actual new or changed scripts; you must add these to each node manually.

18.7.3. Remove a Custom Event

You can remove a custom event dynamically, but the changes will be effective the next time the application will be activated.

Take the following steps to remove a custom event:

1. From the main *Application Roll-over Facility* SMIT screen, select the following options: **Configuration Definition > Application Environment> Remove Custom Pre/Post-event** or use the **smit barf_conf_app_menu** fast path. SMIT displays a list of all *Application Roll-over Facility* custom events.
2. Select the application you want to remove and press Enter. SMIT displays a popup message asking if you are sure you want to proceed. Press Enter again to remove the custom event from the topology definition.
3. After the task is performed you must synchronize the topology using SMIT menus.

18.8. Verifying the Configuration

Verifying the configuration assures you that all resources used by *Application Roll-over Facility* are validly configured, and that ownership and takeover of those resources are defined and in agreement across all nodes. You should verify the configuration after making changes to a node.

Refer to *Verifying Application Roll-over Facility Environment*, on page 6-20.

18.9. Changing the EMC Takeover Parameter

To change the “EMC split bus behaviour Configuration" parameter on a node, perform the following procedure:

1. Enter the **smit barf** command. Then select the following options: **Configuration Definition> EMC split bus behavior Configuration** or use the **smit barf_emc_menu** fast path. SMIT displays the following screen:

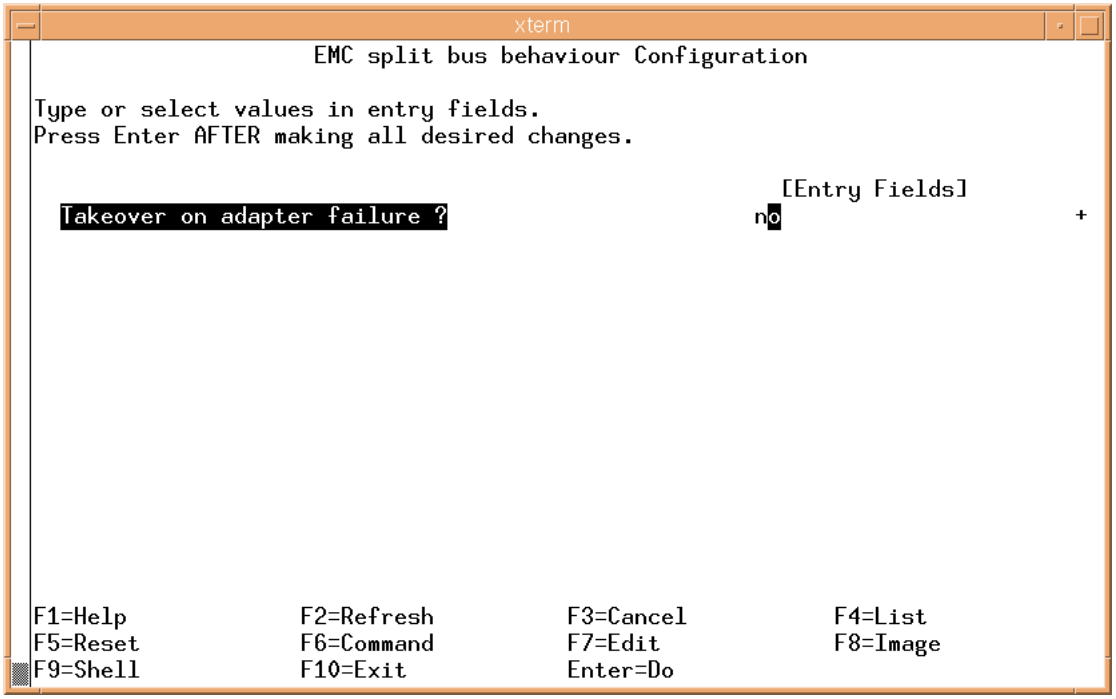


Figure 72. EMC split bus behaviour Configuration screen

Takeover on adapter failure ? yes or no

2. Change the takeover parameter, if necessary, and press Enter.

Chapter 19. Running Application Monitoring

The Application Monitoring allows the administrator to periodically check the status of running applications. This chapter describes how to run the Application Monitoring and provides an example of configuration.

19.1. Starting/Stopping the daemon

19.1.1. Starting the daemon

To start the *Application Monitoring* daemon:

1. Type **smit barf** and select the following options: **Manage Application Monitoring> Activate Application Monitoring** or use the **smit barf_appmondmn** fast path. The following menu appears:

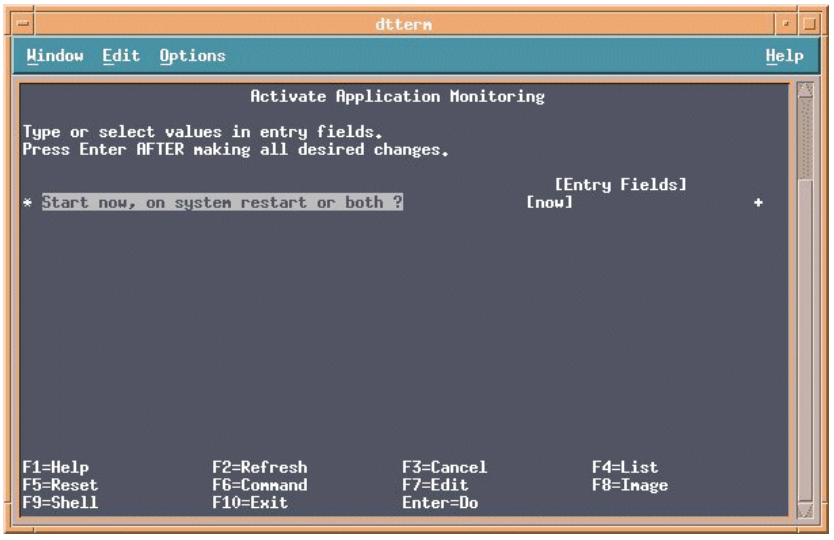
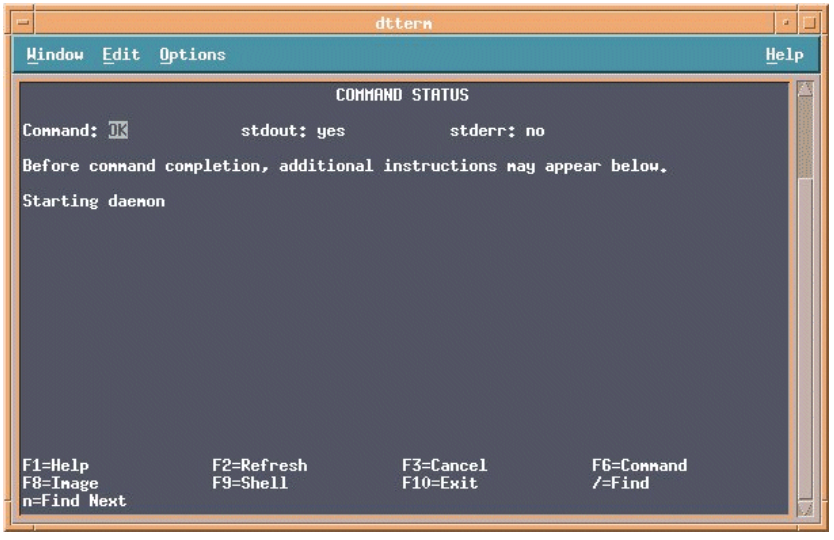


Figure 73. Activate Application Monitoring

2. Choose '**now**' to start immediately, '**restart**' to start at boot time or '**both**' for both.
3. Press Enter, the following menu appears:



4. Press Enter to confirm and return to the Manage Application Monitoring.
5. Press F10 to exit SMIT.

19.1.2. Stopping the daemon

To stop the **Application Monitoring** daemon:

1. Type **smit barf** and select the following options: **Manage Application Monitoring> De-activate Application Monitoring** or use the **smit barf_appmondmn** fast path. The following menu appears:

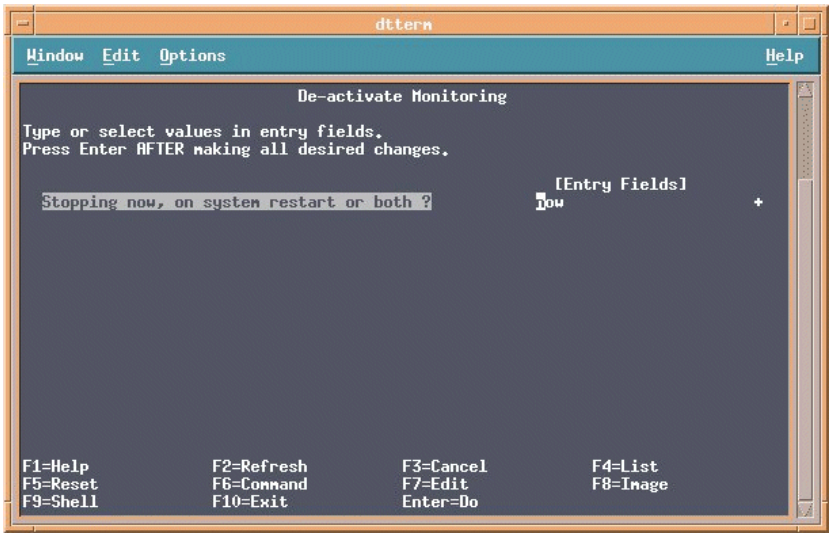
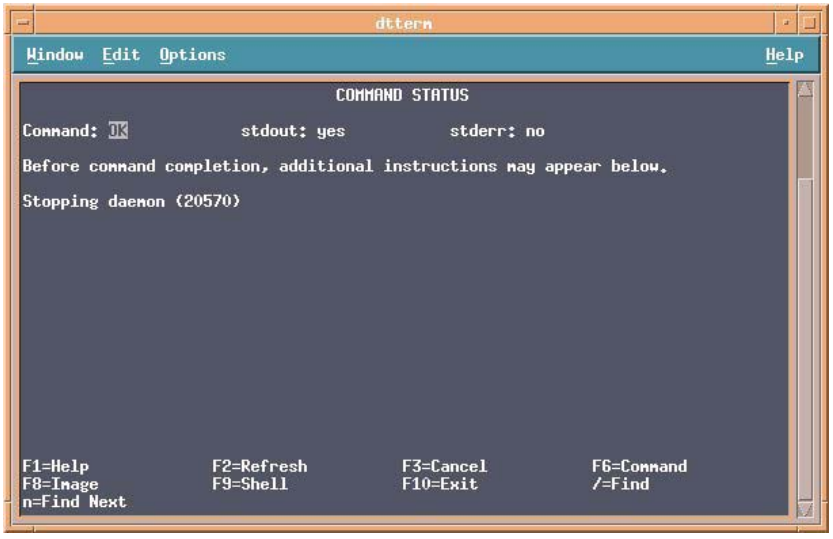


Figure 74. De-activate Application Monitoring

2. Choose '**now**' to stop immediately, '**restart**' to stop at boot time or '**both**' for both.
3. Press Enter, the following menu appears:



4. Press Enter to return to the Manage Application Monitoring.
5. Press F10 to exit SMIT.

19.1.3. Restart the running daemon

To restart the **Application Monitoring** running daemon:

1. Type **smit barf** and select the following options: **Manage Application Monitoring> Restart Application Monitoring** or use the **smit barf_appmondmn** fast path. The following menu appears:

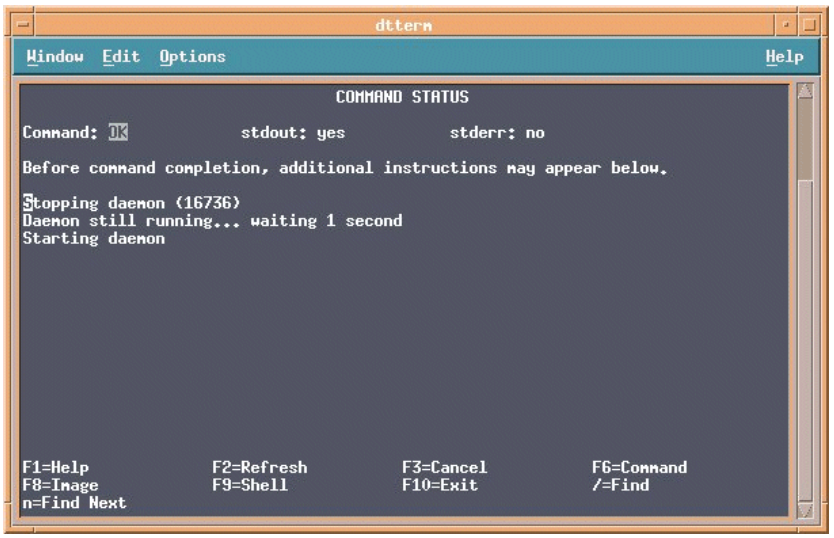


Figure 75. De-activate Application Monitoring

2. Press Enter to return to the Manage Application Monitoring.
3. Press F10 to exit SMIT.

19.2. Showing the running daemon status

To display the status of the **Application Monitoring** running daemon:

1. Type **smit barf** and select the following options: **Manage Application Monitoring> Show Application Monitoring Status** or use the **smit barf_appmondmn** fast path. If the daemon is stopped, the following menu appears:

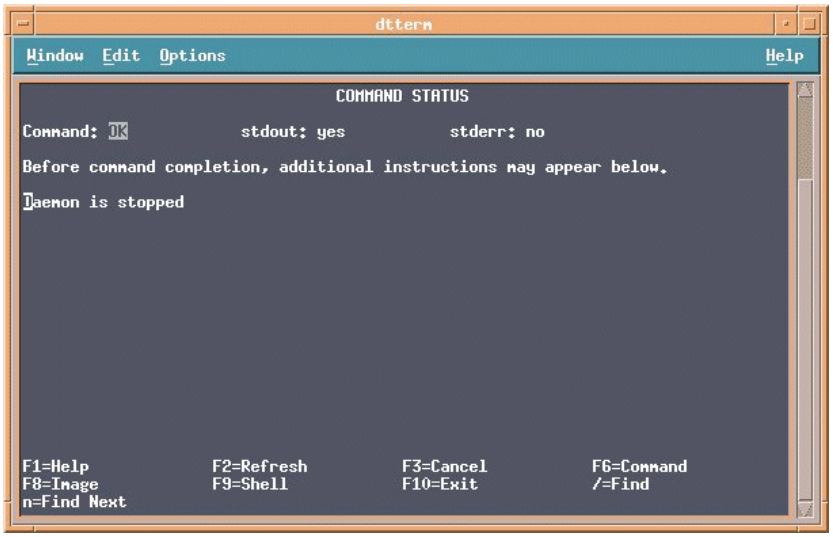
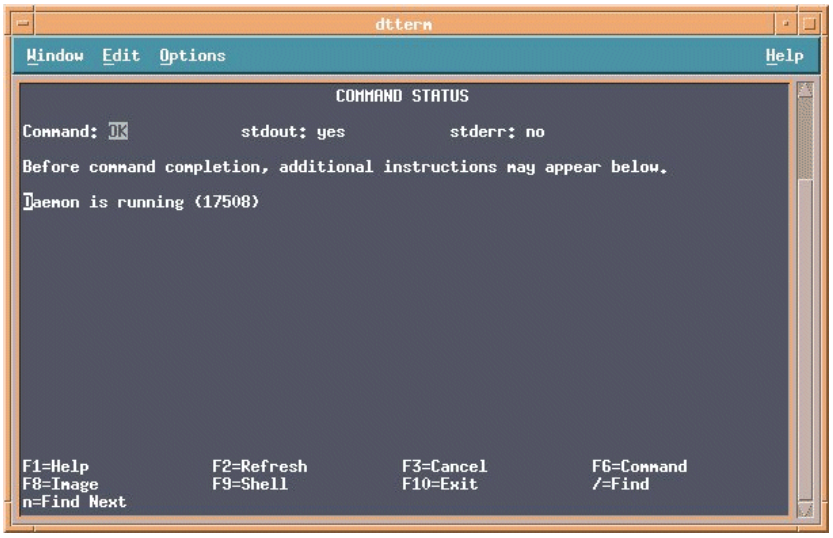


Figure 76. Show Application Monitoring Daemon Status (stopped)

If the daemon is running, the following menu appears:



2. Press Enter to return to the Manage Application Monitoring.
3. Press F10 to exit SMIT.

Chapter 20. Saving and Restoring Configurations

This chapter explains how to use the **snapshot** utility to save and restore *Application Roll-over Facility* configurations.

20.1. Overview

When a topology and resources configuration is completed, you can save it. This configuration can later be restored if necessary.

The **snapshot** utility saves in a file a record of all the data that define a particular configuration. This facility enables you to recreate a configuration (a process called applying a snapshot), provided the cluster is configured with the requisite hardware and software to support this configuration.

WARNING
The snapshot utility cannot be used to re-create a configuration during the *Application Roll-over Facility* migration from one version to a new one.

Because the snapshots are simple ASCII files that can be sent via e-mail, they can make remote problem determination easier.

20.1.1. Information Saved in a Snapshot

The primary information saved in a snapshot is the data stored in the *Application Roll-over Facility* ODM classes. This is the information used to recreate the configuration when a snapshot is applied.

The snapshot does not save any user-customized scripts, applications, or other non-*Application Roll-over Facility* configuration parameters. For example, the name of an application and the location of its start and stop scripts are stored in the **BARFapp** ODM object class.

However, the scripts themselves as well as any application they may call are not saved.

The snapshot also does not save any device- or configuration-specific data which is outside the scope of *Application Roll-over Facility*. For instance, the facility saves the names of shared file systems and volume groups; however, other details, such as NFS options or LVM mirroring configuration are not saved.

20.1.2. Format of a Snapshot

The *Application Roll-over Facility* snapshot utility stores the data in the **/usr/sbin/barf/snap** directory, in the two following files:

snap_xxx.odm	This file contains all the data stored in the *Application Roll-over Facility* ODM object classes. Because the ODM information must be largely the same on every node, the snapshot saves the values from one node only. Refer to *Snapshot ODM Data File*, on page 22-4 for more information.
snap_xxx.info	This file contains the output from standard AIX and *Application Roll-over Facility* system management commands. Output from any custom snapshot methods is appended to this file.

Note **snap_xxx** is a name of your choice.

20.2. Creating a Snapshot

You can initiate snapshot creation from any node. You can create a snapshot on a running node, and you can create multiple snapshots. The snapshot facility retrieves information from each node. Accessibility to all nodes is required.

Because of the large amount of data which must be retrieved when creating the snapshot, the time and memory consumed may be substantial, especially when the number of nodes is high. The snapshot files typically require approximately 5 Kb per node.

To create a cluster snapshot perform the following steps:

1. Enter **smit barf** and select **Configuration Definition > Snapshot > Create a Snapshot**. The following screen appears :

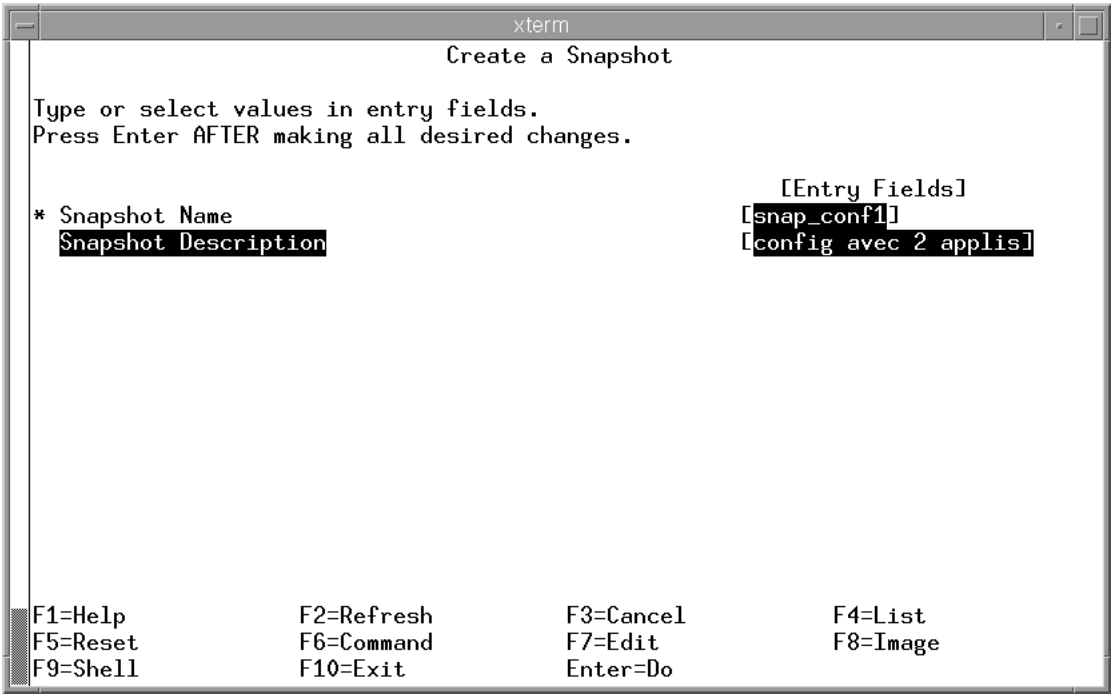


Figure 77. Create a Snapshot screen

Snapshot Name	Enter the name of your choice for the basename for the snapshot files. The directory path for storage and retrieval of the snapshot is **/usr/sbin/barf/snap**. You cannot create a snapshot using an existing snapshot file name.
Snapshot Description	Enter a descriptive text that you want to be inserted into the snapshot. You can specify any text string up to 255 characters in length.

2. Press Enter to validate your selection.

20.3. Applying an Application Roll-over Facility Snapshot

Applying an *Application Roll-over Facility* snapshot overwrites the data in the existing ODM classes on all nodes with the new ODM data contained in the snapshot. You can apply a snapshot from any node. Applying a snapshot may affect both AIX and *Application Roll-over Facility* ODM objects and system files as well as user-defined files.

To apply a cluster snapshot using SMIT, perform the following steps.

1. Enter **smit barf** and select **Snapshot > Apply a Snapshot**. SMIT displays the **Snapshot to Apply** screen containing a list of all the snapshots that exist in the **/usr/sbin/barf/snap** directory.
2. Select the snapshot that you want to apply and press Enter. SMIT displays the Apply a Snapshot screen.

Undoing an Applied Snapshot

Before the new configuration is applied, the snapshot facility saves the current configuration in a file called **~barf_snap.n.odm**, where n is either 1, 2, or 3. The saved snapshots are cycled so that only three generations of snapshots exist. If the apply process fails, you can re-apply the previous configuration. These saved snapshot are stored in the **/usr/sbin/barf/snap** directory.

20.4. Removing an Application Roll-over Facility Snapshot

Removing a snapshot deletes both of the ASCII files that define the snapshot from the snapshots directory (**/usr/sbin/barf/snap**). You must remove the two files manually by using the following commands:

rm /usr/sbin/barf/snap/snap_file.odm

rm /usr/sbin/barf/snap/snap_file.info

Chapter 21. Diagnosing the ARF Resources

21.1. Overview

The diagnostic tool allows you to generate diagnostic information. An administrator may want to use this tool on a regular basis (for example daily) to check the health of the Application Roll-over Facility configuration.

This tool analyzes the state and consistency of ARF resources. Optionally, in accordance with the flags you specify, it performs further subsystem-specific diagnostics by scanning log files on the managed nodes for messages that occurred in the specified period (default 24 hours).

- Resources consistency analysis

  The diagnostic tool analyzes the state and consistency of ARF resources. It looks for the resources definition and checks the state of these resources. After analysis, it reports any potential problem or abnormal condition.

- Subsystem-specific diagnostics

  In accordance with the flags you specify, the diagnostic tool is able to generate diagnostics related to these three subsystems: AIX, ARF, Monitoring Daemon.

  - AIX

    The diagnostic tool looks for hardware errors logged in the AIX error lof file (**errlog**) of each managed node. Any logged hardware error is reported.

  - ARF

    The diagnostic tool scans error messages logged by ARF in the **/var/barf/log/barf.log** file and reports messages that contain **error** or **failed** strings.

  - Monitoring Daemon

    The diagnostic tool scans error messages logged by the Monitoring Daemon in the **/var/barf/log/clsmd.log** file and reports messages that contain **error**, **LAN failure** or **disaster** strings.

- Diagnostics infomation consolidation

  The diagnostic tool gathers the information extracted from the involved log files of the different nodes. Then it consolidates the collected data, taking into account timestamps and summarizing the information of repeated errors.

21.2. Using the Diagnostic Tool

1. Type **smit barf** and select **Show Diagnostics** menu, or use the **smit barfdiag** fast path. The following pop-up menu appears:

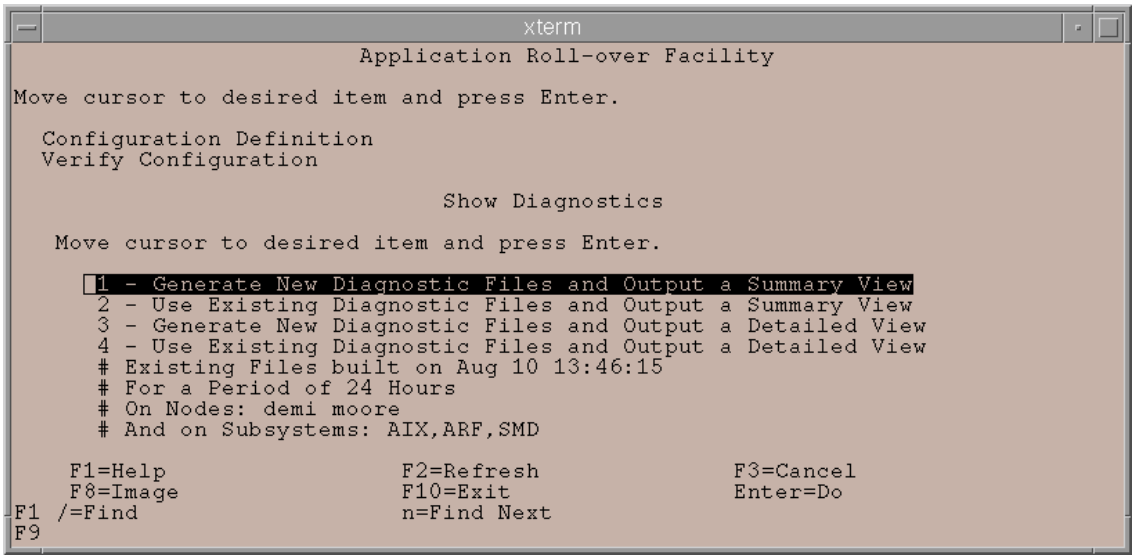


Figure 78. Show Diagnostics menu

2. Select the type of report you want to view.

   You can choose to generate a new report to obtain recent diagnostic information or to use the information that is stored in an already existing report. In bothe cases you can view a summary report or a detailed report

   Note that the lines beginning with "#", below the four menu choices, give information on the conditions (date, nodes, ...) in which the last report has been generated.

- If you select **1 – Generate New Diagnostic files and Output a Summary View**, the following dialog is displayed:

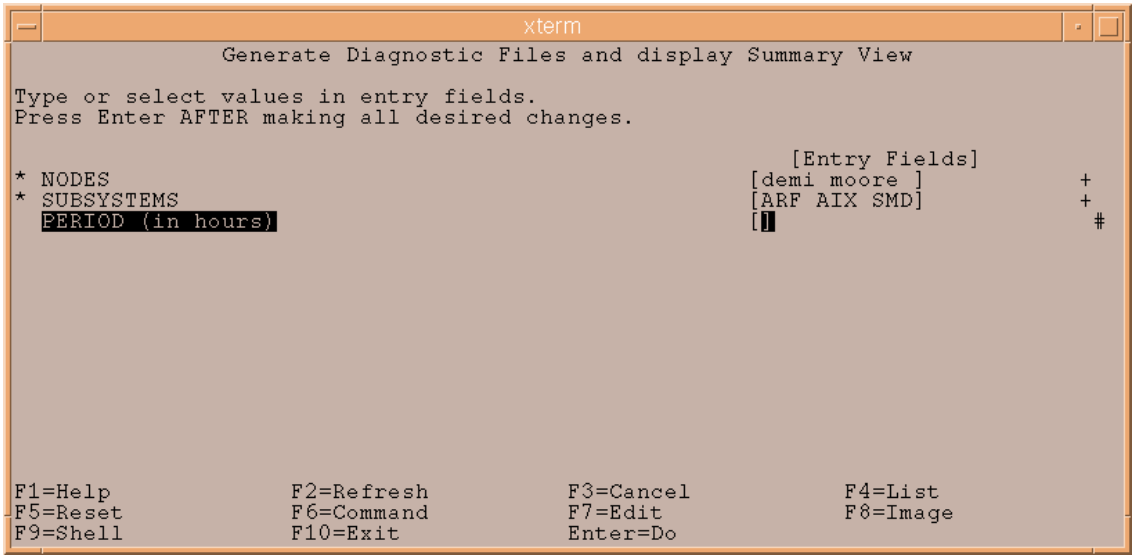


Figure 79. Generate Diagnostic Files and display Summary View

NODES — By default, the diagnostic tool is executed for all the managed nodes. You can restrict its execution to some nodes selected from the list.

SUBSYSTEMS By default, the diagnostic tool is executed for all the subsystems. You can restrict its execution to some subsystems selected from the list.

PERIOD By default, the period is 24 hours (error messages older than one day are ignored). You can enter another period in hours. A period of 0 means infinite.

The state and consistency of resources are always analyzed.

- If you select 3 – Generate New Diagnostic files and Output a Detailed View, the following dialog is displayed:

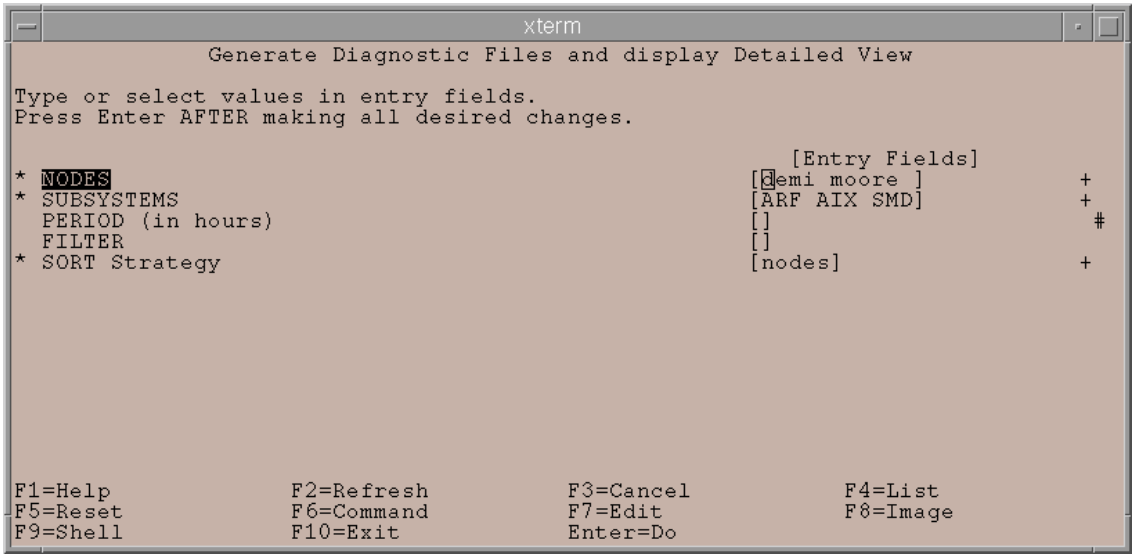


Figure 80. Generate Diagnostic Files and display Detailed View

Two additinal parameters are displayed:

FILTER By default, no filter is applied. You can specify a specific string to search.

SORT Strategy By default nodes is selected. You can select date or subsystem.

- If you select 2 – Use Existing Diagnostic files and Output a Summary View or 4 - Use Existing Diagnostic files and Output a Detailed View, the diagnostic tool is not executed and the reports are displayed using previously generated files.

3. Press Enter. The diagnostic tool is executed in the case of choice 1 or 3 and the report is displayed.
4. Press F10 to exit SMIT.

21.3. Diagnostics files

The /var/barf/diag directory is used to store the various log files generated by the tool. The files are overwritten each time the tool is executed.

The arf_global_report file contains the information that was displayed the last time the tool was executed.

The arf_detailed_report file contains more detailed information and can be of interest for troubleshooting purpose.

Chapter 22. Troubleshooting

This chapter describes the different tools that help you to diagnose configuration problems.

22.1. Understanding the log Files

Your first approach to diagnose a problem affecting a configuration should be to search for messages output by the *Application Roll-over facility* subsystems in the *Application Roll-over facility* log files. These messages can provide invaluable information toward understanding the current state of an application.

Besides *Application Roll-over facility* log files, AIX errlog provides useful information:
If a node halts, reboot it then check AIX errlog with **errpt -a** command and look for error labels like:
DISK_ERR3, SC_DISK_ERR3, ATF_TRESPASS_FAILED, PPATH_DEVICE_GONE, EMCP_VOL_DEAD or EMCP_ALL_PATHS_DEAD.
ARF Automatic Error Notification mechanism halts the node (with halt -q command) when there is no more access to disks.
See *Automatic Error Notification*, on page 5-2.

The following sections describe the types of messages output by the *Application Roll-over facility* software and the log files into which the system writes these messages.

To change the log files to another destination, refer to the section *Customizing Log and Trace Files*, on page 6-9. Log and Trace destination directories changes will take effect when you synchronize resources and the next time *Application Roll-over facility* services are restarted.

Note Existing log and trace files will not be moved to the new location.

22.1.1. barf.log File

The **/var/barf/log/barf.log** file is a standard text file. When checking this file, first find the most recent error message associated with your problem. Then read back through the log file to the first message relating to that problem. Many error messages cascade from an initial error that usually indicates the problem source.

When scripts start, complete, or encounter error conditions, the *Application Roll-over facility* software generates a message. For example, the following fragment from the log file illustrates the start and completion messages for several *Application Roll-over facility* scripts. The messages include any parameters passed to the script.

```
ven 14 jun 10:25:38 DFT 2002: app1: STARTING EXEC of barf_activate_resource app1
ven 14 jun 10:25:39 DFT 2002: app1: STARTING EXEC of barf_make_disk_available hdisk2
ven 14 jun 10:25:42 DFT 2002: app1: SUCCESSFULL EXEC of barf_make_disk_available hdisk2
ven 14 jun 10:25:42 DFT 2002: app1: STARTING EXEC of barf_mount_vg app1 HALITE
ven 14 jun 10:25:53 DFT 2002: app1: SUCCESSFULL EXEC of barf_mount_vg app1 HALITE
ven 14 jun 10:25:53 DFT 2002: app1: STARTING EXEC of barf_mount_fs app1 /HALITE_fs1
ven 14 jun 10:25:54 DFT 2002: app1: SUCCESSFULL EXEC of barf_mount_fs app1 /HALITE_fs1
ven 14 jun 10:25:54 DFT 2002: app1: SUCCESSFULL EXEC of barf_activate_resource app1
ven 14 jun 10:25:56 DFT 2002: app1: STARTING EXEC of barf_start_appli -n app1
ven 14 jun 10:25:57 DFT 2002: app1: SUCCESSFULL LAUNCH of barf_start_appli -n app1 in
background
ven 14 jun 10:25:57 DFT 2002: app1: INFO: check if application app1 is already started
on remote node castor ...
```

22.1.2. Trace Files

In addition to the start, completion, and error messages generated by scripts, the *Application Roll-over facility* software generates a detailed report of each step of script processing. In verbose mode, which is the default, the shell generates a message for each command executed in the script, including the values of all arguments to these commands.

This file is created in the directory **/var/barf/trace** and its name is composed of the name of the application and the date/hour of the operation (example: `appli_name.trace.MMDDYY_HHMMSS`).

Note This trace corresponds to operation executed on the node itself. For example, in case of application moving from one node to another node, de-activation is traced on one node, re-activation is traced on the other node.

The following fragment from a trace file illustrates the verbose output of the start aplication script:

```
ven 14 jun 10:25:38 DFT 2002: STARTING EXEC of barf_activate_resource app1
[21] . /usr/sbin/barf/lib/barf_lib
[21] [21] basename barf_activate_resource
PROG=barf_activate_resource
barf_activate_resource[21] export PS4=$PROG[$LINENO]
barf_activate_resource[21] typeset -r PATH_BARF_BIN=/usr/sbin/barf/bin
barf_activate_resource[21] typeset -r PATH_BARF_LIB=/usr/sbin/barf/lib
barf_activate_resource[21] typeset -r PATH_BARF_UTILS=/usr/sbin/barf/utils
barf_activate_resource[21] typeset -r PATH_BARF_DATA=/usr/sbin/barf/data
barf_activate_resource[21] typeset -r
FILE_BARF_STATUS=/var/barf/barf_activate_appli.status
barf_activate_resource[24] barf_activate_resource[24] basename
barf_activate_resource
PROGRAM=barf_activate_resource
barf_activate_resource[25] CATALOG=barf.cat
barf_activate_resource[26] ODMDIR=/etc/objrepos
barf_activate_resource[27] BARFBIN=/usr/sbin/barf/bin
barf_activate_resource[28] BARFUTILS=/usr/sbin/barf/utils
barf_activate_resource[29]
PATH=/usr/bin:/etc:/usr/sbin:/usr/ucb:/usr/bin/X11:/s
bin:/usr/java130/jre/bin:/usr/java130/bin:/usr/sbin/barf/bin:/usr/sbin/barf
/util
s:/usr/sbin/barf/bin
barf_activate_resource[31] APP=app1
barf_activate_resource[64] barf_activate_resource[64] retrieve_resource
addr
.............
```

22.1.3. clsmd.log File

When the *Application Roll-over facility* node monitoring is started, stopped or when its state changes, it generates messages. These messages can be informational, such as a warning message, or they can report a fatal error.

The structure of the event history file is as follows:

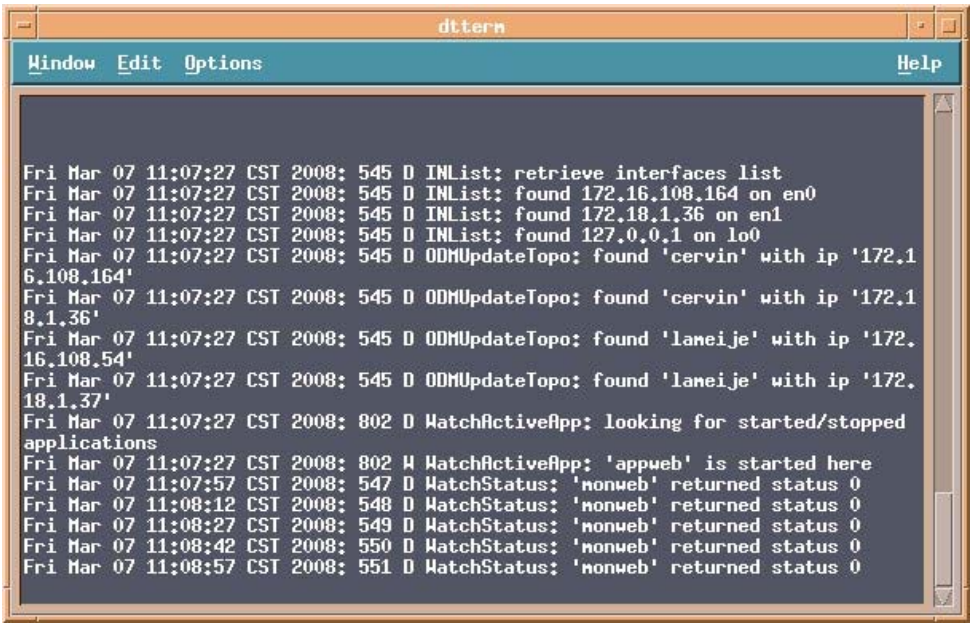
```
clsmd: date hour:minute:second year: daemon status: message text
```

The following example illustrates heartbeat messages output by the heartbeat daemon.

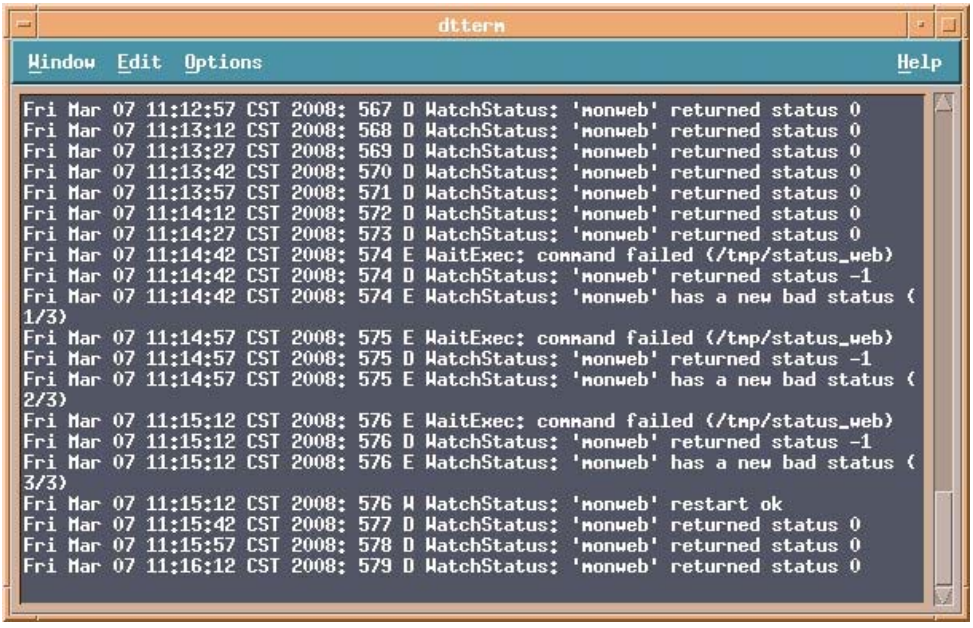
```
clsmd: Tue Jun 18 14:15:53 2002: 0: pollux: 120.0.1.20 daemon monitoring
castor: 120.0.1.50 -started.
clsmd: Tue Jun 18 14:15:53 2002: 1: pollux: 120.0.1.20 node daemon on
standby: waiting for heartbeats from castor: 120.0.1.50.
clsmd: Tue Jun 18 14:15:59 2002: 22: pollux: 120.0.1.200 Node castor :
120.0.1.50 ready.
clsmd: Tue Jun 18 14:15:59 2002: 2: pollux: 120.0.1.20 node daemon
monitoring castor: 120.0.1.50.
```

22.1.4. Example of Monitoring Application Log File

The Monitoring Application Log File is present on each node running the application in the /var/barf/log/appmond.log file.
The following figure shows an example of a Monitoring Application Log File for a running application.



If the monitored process is killed, the status command returns 1(bad status) and the application monitoring restarts the application.
The following figure shows an example of a Monitoring Log File for a restarted application.



22.1.5. mail Message

When the Node Monitoring daemon detects a node failure it sends a mail to the administrator according to the Heartbeat Mailing list declared in the runtime parameters.

The following example illustrates the mail message:

```
Message 1:
From root Wed Oct 9 13:50:13 2002
Date: Wed, 9 Oct 2002 13:50:13 +0200
From root
To: root
Subject: CAUTION: DISASTER DETECTED

clsmd: Wed Oct 9 13:49:12 2002: 84: rully: ERROR detected. The node moore
is failed.
```

22.1.6. Console System Message

When the Node Monitoring daemon detects a node failure it sends a message to the system console.

The following example illustrates the system console message:

```
CAUTION: DISASTER DETECTED!
clsmd: Thu Oct 10 11:05:45 2002: 84: demi: 120.0.2.30 ERROR detected. The
node moore: 120.0.2.60 is failed.
```

22.2. Understanding ODM Database

22.2.1. ODM Classes

The *Application Roll-over Facility* configuration is saved in BARF* ODM classes in the **/etc/objrepos** directory. Some of these files are listed below:

BARFtopo	Node information including names of participating nodes, and addresses list corresponding to each node.
BARFapp	Application information including names, and the depending resources.
BARFevent	Event information including the name, description, and names of pre- and post-processing scripts.
BARFrtp	Runtime information including logs and heartbeat parameters.
BARFservice	Service information including the value of the Inactive takeover attribute
BARFlogs	Log information including trace file for each operation on an application.
BARFactiveapp	Active Application information including names, list of take-over nodes and current node on which the application is currently running.
BARFsync	Synchronization information.

You can use the **odmget** command to check the contents of the database.

22.2.2. Snapshot ODM Data File

The snapshot ODM data file is an ASCII text file divided into three delimited sections:

Version section	This section identifies the version of the cluster snapshot. The characters <VER and </VER delimit this section. The version number is set by the snapshot software.
Description section	This section contains user-defined text that describes the cluster snapshot. You can specify up to 255 characters of descriptive text. The characters <DSC and </DSC delimit this section.

ODM data section — This section contains the ARF ODM object classes in generic AIX ODM stanza format. The characters <ODM and </ODM delimit this section.

Following is an excerpt from a sample snapshot ODM data file showing the various ODM stanzas that are saved.

```
<VER
1.0
</VER
<DSC
My Snapshot
</DSC
<ODM
BARFtopo:
  nodename = ”node1”
  nodeaddr = ”120.1.10.1”
  mobility = 0
BARFapp:
  appname = ”app1”
  restype = ”node”
  resvalue = ”node1”
BARFapp:
  appname = ”app1”
  restype = ”addr”
  resvalue = ”120.1.11.1”
BARFapp:
  appname = ”app1”
  restype = ”vg”
  resvalue = ”HALITE”
  .
  .
</ODM
```

22.3. Daemon Status for Node Monitoring

Following are the different daemon status that can be displayed when you run the node monitoring daemon:

Status number	Message name	Message text	Comment
0	Node started	<local node>: daemon monitoring <remote node> started.	The monitoring node daemon is started on the local node.
1	Node standby mode	<local node>: node daemon on standby: waiting for heartbeats from <remote node>.	The monitoring node daemon is waiting for heartbeats from the other remote nodes to begin the monitoring.
2	Node normal mode	<local node>: node daemon monitoring <remote node>.	The local node daemon is monitoring the remote node.
4	Node stopped	<local node>: node daemon monitoring <remote node> stopped.	The monitoring daemon running on the node has been stopped.
20	Monitoring begins	<local node> monitored by <remote node>.	The node is monitored by a daemon on the remote node.
21	Local node ready	<local node>: local node ready.	The monitoring daemon has been started on the local node.
22	Remote node ready	Node on <remote node> ready.	The daemon on the remote node is started.
24	No more monitoring	<local node> no more monitored by <remote node>.	The daemon on the other node is stopped.
33	Forced stop successful	<local node>: stop successful. Warning: <remote node> was not informed.	The daemon has been stopped but the daemon on the remote node does not know it.
34	Stop operation canceled	<local node>: stop operation canceled because <remote node> is unreachable.	The daemon to be stopped cannot contact the daemon on the remote node.
40	No more heartbeat	<local node>: no more heartbeat received from <remote node>.	The two nodes were exchanging heartbeats. A heartbeat timeout occurred because no more heartbeat was received.
41	Heartbeat received	<local node> is receiving again heartbeats from <remote node>.	After a heartbeat timeout has occurred, the heartbeats are received again.
50	LAN failure	<local node>: LAN failure.	The node daemon cannot ping any other host.
80	Remote node unreachable	<local node>: node on <remote node> unreachable.	The monitoring daemon on the node cannot communicate with the daemon running on the remote node.
84	Node failed	<local node>: ERROR detected. The node <remote node> is failed	A node failure has been diagnosed: the monitoring daemon on the remote node has been unreachable by network too many times.
85	Network Problem	<local node>: ERROR detected. The node <remote node> is unreachable. A network problem occurs between the local IP address and the remote IP address.	A network problem has been diagnosed.
86	Local IP Address Failed	<local node>: ERROR detected. The node <remote node> is unreachable from the local IP address. This address is failed.	A Local IP Address Failed has been diagnosed.
87	Daemon no more running	<local node>: ERROR detected. The monitoring node daemon is no more running on node <remote node>.	The Monitoring Node Daemon is no more running.

Appendix A.Defining Shared LVM Components

This appendix describes how to define the LVM (Logical Volume Group) components shared by the nodes in an *Application Roll-over Facility* configuration.

A.1. Overview

Creating the volume groups, logical volumes, and file systems shared by the nodes in an *Application Roll-over Facility* configuration requires that you perform some steps on all nodes. In general, you define the components on one node (referred to in the text as the “source node") and then manually import the volume group onto each of the nodes in the configuration (referred to as “destination nodes"). This ensures that the ODM definitions of the shared components are the same on all nodes. It can be done automatically during the propagation of the configuration. *Application Roll-over Facility* imports the volume group during the synchronization of the application resources.

While configuring file systems for a resource group, you can select one or more individual file systems to be mounted in the particular resource group.

A.2. Creating a Shared Volume Group on Source Node

Use the **smit mkvg** fast path to create a shared volume group. Use the default field values unless your site has other requirements, or unless you are specifically instructed otherwise here:

VOLUME GROUP name
: The name of the shared volume group should be unique within the *Application Roll-over Facility* configuration.

Activate volume group AUTOMATICALLY at system restart?
: Set to 'no' so that the volume group can be activated as appropriate by the *Application Roll-over Facility* event scripts.

Volume Group MAJOR NUMBER
: It is not necessary to import Volume Groups with the same major number on all nodes, except if Volume Groups contain NFS exported File Systems. NFS uses volume group major numbers to help uniquely identify exported filesystems. Therefore, if NFS is used, all nodes using NFS exported filesystem must have imported the volume group on which the filesystem resides with the same major number.
However, Application Rollover Facility will use a major number free and available on all nodes during the propagation of resources. You can use the lvlstmajor command on each node to determine a free major number common to all nodes.

A.3. Creating a Shared File System on Source Node

Use the **smit crjfs** fast path to create the shared file system on the source node. When you create a journaled file system, AIX creates the corresponding logical volume. Therefore, you do not need to define a logical volume. You do, however, need to later rename both the logical volume and the log logical volume for the file system and volume group.

Mount AUTOMATICALLY at system restart?	Make sure this field is set to 'no'.
Start Disk Accounting	Make sure this field is set to 'no'.

A.4. Renaming JFS Logs and Logical Volumes on Source Node

AIX assigns a logical volume name to each logical volume it creates. Examples of logical volume names are `/dev/lv00` and `/dev/lv01`. Within an *Application Roll-over Facility* configuration, the name of any shared logical volume must be unique. Also, the journaled file system log (**jfslog** or **jfs2log**) is a logical volume that requires a unique name in the *Application Roll-over Facility* configuration.

To make sure that logical volumes have unique names, rename the logical volume associated with the file system and the corresponding **jfslog** or **jfs2log** logical volume. Use a naming scheme that indicates the logical volume is associated with a certain file system. For example, `lvsharefs` could name a logical volume for the `/sharefs` file system.

1. Use the **lsvg -l volume_group_name** command to determine the name of the logical volume and the log logical volume (**jfslog** or **jfs2log**) associated with the shared volume groups. In the resulting display, look for the logical volume name whose type is **jfs** or **jfs2**. Then look for the logical volume names whose type are **jfslog** or **jfs2log**. They are the log logical volumes.
2. Use the **smit chlv** fast path to rename the logical volume and the log logical volume.

After renaming the **jfslog** or **jfs2log** or a logical volume, check the **/etc/filesystems** file to make sure the **dev** and **log** attributes reflect the change. Check the **log** attribute for each file system in the volume group and make sure that it has the new **jfslog** or **jfs2log** name. Check the **dev** attribute for the logical volume you renamed and make sure that it has the new logical volume name.

A.5. Adding Copies to Logical Volume on Source Node

To add logical volume copies on a source node:

1. Use the **smit mklvcopy** fast path to add copies to a logical volume. Add copies to both the **jfslog** log logical volume and the logical volumes in the shared file systems. To avoid space problems, first mirror the **jfslog** log logical volume and then the shared logical volumes.

   The copies should reside on separate disks that are controlled by different disk adapters and are located in separate drawers or units, if possible.

Note These steps do not apply to Disk Arrays, which provide their own mirroring of logical volumes. Continue with *Testing a File System*.

2. Verify the number of logical volume copies. Enter:

   ```
   lsvg -l volume_group_name
   ```

   In the resulting display, locate the line for the logical volume for which you just added copies. Notice that the number in the physical partitions column is x times the number in the logical partitions column, where x is the number of copies.
3. To verify the placement of logical volume copies, enter:

   ```
   lspv -l hdiskx
   ```

   where `hdiskx` is the name of each disk to which you assigned copies. That is, you enter this command for each disk. In the resulting display, locate the line for the logical volume for which you just added copies. For copies placed on separate disks, the numbers in the logical partitions column and the physical partitions column should be equal. Otherwise, the copies were placed on the same disk and the mirrored copies will not protect against disk failure.

A.6. Verifying the File Systems

To run a consistency check each file system information:

1. Enter:

   ```
   fsck /filesystem_name
   ```

2. Verify that you can mount the file system by entering:

   ```
   mount /filesystem_name
   ```

3. Verify that you can unmount the file system by entering:

   ```
   umount /filesystem_name
   ```

A.7. Varying Off a Volume Group on the Source Node

After completing the previous tasks, use the **varyoffvg** command to deactivate the shared volume group. You vary off the volume group so that it can be properly imported onto the destination node and activated as appropriate by the *Application Roll-over Facility* event scripts. Enter the following command:

```
varyoffvg volume_group_name
```

A.8. Importing a Volume Group onto Destination Nodes

Importing the volume group onto the destination nodes synchronizes the ODM definition of the volume group on each node on which it is imported.

When adding a volume group to the resource group, you may choose to manually import a volume group onto the destination nodes or you may choose to automatically import it onto all the destination nodes in the resource group.

To avoid the following error message: `516-567 volume group`, run the **lspv** command to check if the disks related to the VG are already seen with a PVID. If the PVID is "none", read it from the disk, using the command: **chdev -l hdisk<x> -a pv=yes**

A.8.1. Importing a Volume Group Automatically

Automatic import of a volume group may be set in SMIT under the *Application Roll-over Facility* Resources menu. It enables *Application Roll-over Facility* to automatically import shareable volume groups onto all the destination nodes in the resource group. Automatic import allows you to create a volume group and then add it to the resource group immediately, without manually importing it onto each of the destination nodes in the resource group.

Note Each volume group is assigned a major number when it is created. When *Application Roll-over Facility* automatically imports a volume group, the major number already assigned to the volume group will be used if it is available on all the destination nodes. Otherwise, any free major number will be used.

Prerequisites and Notes

To import available volume groups, make sure that the following conditions are met:

- Volume group names must be the same across nodes, and unique to the configuration.
- Logical volumes and file systems must have unique names.
- All physical disks must be known to AIX and have PVIDs assigned.
- A volume group is available for auto import onto other nodes only if the physical disks on which it resides are available to all of the nodes in the resource group.

A.8.2. Importing a Volume Group Manually

If you do not want *Application Roll-over Facility* to import your volume group automatically upon adding it to the resource group, make sure that the **Automatically Import Volume Groups** flag is set to False (this is the default value). Use the **smit importvg** fast path.

VOLUME GROUP name
: Enter the name of the volume group that you are importing. Make sure the volume group name is the same name that you used on the source node.

PHYSICAL VOLUME name
: Enter the name of a physical volume that resides in the volume group. Note that a disk may have a different physical name on different nodes. Make sure that you use the disk name as it is defined on the destination node.

ACTIVATE volume group after it is imported?
: Set the field to 'yes'.

Volume Group MAJOR NUMBER
: Use the same major number on all nodes. Use the **lvlstmajor** command on each node to determine a free major number common to all nodes.

A.8.3. Changing a Volume Group Startup Status

By default, a volume group that has just been imported is configured to automatically become active at system restart. In an *Application Roll-over Facility* environment, a volume group should be varied on as appropriate by the *Application Roll-over Facility* event scripts. Therefore, after importing a volume group, use the Change a Volume Group screen to reconfigure the volume group so that it is not activated automatically at system restart.

Use the **smit chvg** fast path to change the characteristics of a volume group.

Activate volume group Automatically at system restart?
: Set this field to 'no'.

A QUORUM of disks required to keep the volume group on-line?
: Set this field to 'no' if AIX mirroring is used for this Volume Group.

A.8.4. Vary Off Volume Group on Destination Nodes

Use the **varyoffvg** command to deactivate the shared volume group so that it can be imported onto another destination node or activated as appropriate by the *Application Roll-over Facility* event scripts. Enter:

```
varyoffvg volume_group_name
```

Appendix B. Customizing the Warning Program File

The Heartbeat Warn Program is a user-defined program to be run by the ARF Monitoring daemon when the following events are detected:

1. A Node Failure has been diagnosed.
   Status number: 84.
   The monitoring daemon on the remote node has been unreachable by network too many times.
2. A Network Problem has been diagnosed.
   Status number: 85.
   A Network problem occurs between the local IP Address and the remote IP Address
3. A Local IP Address Failed has been diagnosed.
   Status number: 86.
   A failure has been detected on the LAN adapter.
4. A Local Network Problem has been diagnosed.
   Status number: 88.
   The network from the local IP address is unreachable.

For this ARF Monitoring events previously described, you can customize the Warning Program File to execute some actions like send email, send messages, …

To use this user-defined script, it is mandatory to register its full pathname in the Heartbeat Warn Program field of the Runtime Parameters.
Refer to 6.7.*Configuring the Node Monitoring*, on page 13.

An example of this script is given here after:

```
#!/bin/bash
#
############################################################
#
# Heartbeat Warn Program: This script is a user-defined program to be run by the
# ARF Monitoring daemon when the following events are detected:
#
# - Node Failed
# - Network Problem
# - Local IP Address Failed
# - Local Network Problem
#
# The full pathname of this script must be registered in the
# Heartbeat Warn Program field of the Runtime Parameters.
#
#
# Event Monitoring Description Param #1: Status – Param #2 – Param #3
#
# Node Failed 84 LocalNodeName RemoteNodeName
# Network Problem 85 LocalIPAddr RemoteIPAddr
# Local IP Address Failed 86 LocalNodeName LocalIPAddr
# Local Network Problem 88 LocalNodeName LocalIPAddr
#
typeset Status=$1
```

```
typeset LocalNodeName=""
typeset RemoteNodeName=""
typeset LocalIPAddr=""
typeset RemoteIPAddr=""


case $Status in
84)
        #
        # Node Failed. A node failure has been diagnosed.
        # The monitoring daemon on the remote node has been
        # unreachable by network too many times.
        #
        LocalNodeName=$2
        RemoteNodeName=$3
        #
        # Action to do in case of a Node Failed. (must be customized)
        # For example, send email, console message, call script, ...
        #
        # For test purpose
        #
        # echo $Status $LocalNodeName $RemoteNodeName at $(date) >>
        /var/barf/log/res_warn_program

;;
85)
        #
        # A Network Problem has been diagnosed.
        # A Network problem occurs between the local IP Address
        # and the remote IP Address.
        #
        LocalIPAddr=$2
        RemoteIPAddr=$3
        #
        # Action to do in case of a Network Problem. (must be customized)
        # For example, send email, console message, call script, ...
        #
        # For test purpose
        #
        # # echo $Status $LocalIPAddr $LocalIPAddr at $(date) >>
        /var/barf/log/res_warn_program

;;
86)
    #
    # A Local IP Address Failed has been diagnosed.
    # A failure is detected on the LAN adapter
    #
    LocalNodeName=$2
    LocalIPAddr=$3
    #
    # Action to do in case of a local IP Address Failed. (must be customized)
    # For example, send email, console message, call script, ...
    #
    #
    # For test purpose
    #
    # echo $Status $LocalNodeName $LocalIPAddr at $(date) >>
```

```
    /var/barf/log/res_warn_program

;;
88)
        #
        # A Local Network Problem has been diagnosed.
        #
        LocalNodeName=$2
        LocalIPAddr=$3
        #
        # Action to do in case of a Local Network Problem. (must be customized)
        # For example, send email, console message, call script, ...
        #
        # For test purpose
        #
        # echo $Status $LocalNodeName $LocalIPAddr at $(date) >>
        /var/barf/log/res_warn_program
        ;;
esac
exit
```

Appendix C. Configuring Fibre Channel Network for ARF

This chapter explains how to configure Fibre Channel Network for heartbeat monitoring.

This configuration may save additional Ethernet adapters or RS232 Serial Line adapters for heartbeat monitoring.

C.1. Configuring FC switches

If zoning is configured at FC switch level, zones must be created to allow communication between FC adapters:

- a first zone, named **broadcast** (this name is mandatory) containing the WWNs of all FC adapters with IP configured
- and other zones containing each FC adapter on one node and the corresponding adapter of the other node(s).

If you create TWO Fibre Channel IP Networks for heartbeat, select IP addresses on different subnets. Example:

- 1.0.0.1 on fcs0 on first node and 1.0.0.2 on fcs0 on second node for first network
- 2.0.0.1 on fcs1 on first node and 2.0.0.2 on fcs1 on second node for second network

C.2. Enable a FC Network Device On all Nodes and FC Adapters

Initially the *fcnet<x>* (FC Network Protocol Device) associated with *fcs<x>* adapter is in **defined** state.

1. Change *fcnet<x>* to **enabled** state:

```
smit devices
      -> FC Adapter
            -> FC Network Protocol Device
                  -> Enable a FC Network Device
```

   (Or fastpath= **smit fcnpd**).

   Select each *fcs<x>* adapter on the node in order to change the state of *fcnet<x>* device.

   This will allow *fcnet<x>* device to be configured as **available** on next **cfgmgr** or **reboot**.

2. Run **cfgmgr** to configure the *fcnet<x>* Fiber Channel Network Device.
3. Check the device state:

```
lsdev -C | grep fc
```

```
fc0    Defined      2R-08-02     Fibre Channel Network Interface
fcnet0 Available    2R-08-02     Fibre Channel Network Protocol Device
fcs0   Available    2R-08        FC Adapter
```

Note *fc<x>* network interface may not exist in defined state.

C.3. Configure FC Network Interface and IP Address

On each node, if *fc<x>* FC Network device does not exist:

- Create it,
- Specify its IP address (e.g.: 1.0.0.1 on the first node and 1.0.0.2 on the second node) and network mask.
- Activate the interface:

```
smit mkinet1fc
```

```
INTERNET ADDRESS (dotted decimal)                  [1.0.0.1]
Network MASK (hexadecimal or dotted decimal)       [255.255.0.0]
Network Interface Name                             fc0
ACTIVATE the Interface after Creating it           yes
```

On each node, if *fc<x>* FC Network device exists in *defined* state:

- Specify its IP address and network mask.
- Activate the interface:

```
smit chinet
```

```
Network Interface Name                             fc0
INTERNET ADDRESS (dotted decimal)                  [1.0.0.1]
Network MASK (hexadecimal or dotted decimal)       [255.255.0.0]
Current STATE                                      up
```

```
lsdev -C | grep fc
```

```
fc0    Available    2R-08-02     Fibre Channel Network Interface
fcnet0 Available    2R-08-02     Fibre Channel Network Protocol Device
fcs0   Available    2R-08 FC     Adapter
```

C.4. Modify /etc/hosts and /.rhosts on EACH node

Register IP addresses in the /etc/hosts and /.rhosts files:

Example:

- For /etc/hosts:

```
1.0.0.1 node1_ipfc
1.0.0.2 node2_ipfc
```

- For /.rhosts and /home/arfw/.rhosts:

```
node1_ipfc
node2_ipfc
```

```
netstat -i
```

This command shows *fc<x>* interfaces:

```
fc0 65280    link#3       0.0.c9.37.55.11    17639 0 17636 0 0
fc0 65280    link#3       node1_ipfc         17639 0 17636 0 0
```

```
ifconfig fc0
```

```
fc0: flags=e000843<UP,BROADCAST,RUNNING,SIMPLEX,GROUPRT,64BIT>
inet 10.0.50.20 netmask 0xffffff00 broadcast 10.0.50.255
```

C.5. Check Fibre Channel Network Connection

It is possible to ping node1 from node2:

```
ping node1_ipfc
```

and conversely, it is possible to ping node2 from node1:

```
ping node2_ipfc
```

Note Very first ping may start after a delay of about 30 seconds. Please wait ...

From this point, the addresses of the Fibre Channel Network can be configured in the *Application Roll-over Facility* Environment, as described in *Defining a Topology*, on page 6-1 and in *Changing the IP Address List of a Node*, on page 18-11.

Restriction:

The network availability cannot be ensured, as Etherchannel functionality is not available on such adapters.

Index

H

I

J

K

L

M

N

O

P

R

S

T

U

V

W

Bull Cedoc
357 avenue Patton
BP 20845
49008 Angers Cedex 01
FRANCE

REFERENCE
86 A2 95EF 14