High Availability: Why are my filesystems no longer exported from my head node after setting up HA?

Why are my filesystems no longer exported from my head node after setting up HA?

Short answer: It is done by Bright to protect the filesystems from corruption.

The longer answer, and using a workaround, is more involved. We will start with some background:

Background: Why not just export the filesystem anyway?

Bright Cluster Manager automatically disables export over NFS if a shared storage is not installed when a High Availability (HA) setup is created. The reason behind this is that a filesystem that is exported to the regular nodes, but is not part of shared block device, will display stale NFS handle errors on failover. Stale NFS handle errors occur because the file or directory that was opened by an NFS client is not reachable any more. The best way to enable the filesystem exports is to make them a part of a shared block device.

I want to export the filesystem anyway. What's a workaround to this protection?

This workaround is not recommended unless you are aware of all the risks and consequences. But you can go ahead as follows:

- **Basic principle:** If both the mount and unmount scripts (`mountscript`, `umountscript`) under `cmsh`:`partition`->`failover`->`base`, are set to `/bin/true` then cmdaemon will think that a shared storage has been installed. The exports are then no longer disabled in `/etc/exports`.

- **Handy script hint:** There is a script which is called on each node right after a failover has taken place. The script is: `/cm/local/apps/cmd/scripts/nodereconnectcheck.sh`.

  This script was used in the past for checking that NFS filesystems were mounted properly on the compute nodes after a failover, and remounting them if necessary. By default everything is commented out in this.

  If the mount script and unmount script are set to `/bin/true` to allow both head nodes to export the filesystems, then in the event of a failover, stale NFS handles will occur unless the `nodereconnectcheck.sh` script is used to unmount the old NFS mount using a lazy unmount (umount -l), and to remount it from the new active head node.

High Availability: Why are my filesystems no longer exported from my head node after setting up HA?

- **No NFS shift by default:** Bright by default does not use the `nodereconnectcheck.sh` script to unmount the old NFS mount and remount it from the new active head node. This is because if a compute job is actively doing I/O on e.g. `/home/foobar/datafile` while a failover is taking place, then the read/write operations may fail and the job may crash. This is not acceptable for HPC users.

Unique solution ID: #1114
Author: mohamed adel
Last update: 2013-05-01 22:32