CLUSTERPRO for Linux Ver3.0

GFS HowTo

2004.07.30
第1版

改版履歴

版数	改版日付	内 容
1	2004/07/30	初版新規作成

最新の動作確認情報、システム構築ガイド、アップデート、トレッキングツールなどは
以下のURLに掲載されています。
システム構築前に最新版をお取り寄せください。

NECインターネット内でのご利用
　http://soreike.wsd.mt.nec.co.jp/
　[クラスタシステム]→[技術情報]→[CLUSTERPROインフォメーション]

NECインターネット外でのご利用
　http://www.ace.comp.nec.co.jp/CLUSTERPRO/
　[ダウンロード]→[Linuxに関するもの]→[ツール]

目次

1 はじめに

本書はIA-64 Linux用のCLUSTERPRO for Linux Ver3.0でGFSの二重化運用を行う際に参考となる情報を記述したものです。インストール及び構築作業前に必ずお読みください。

本書は以下のCLUSTERPROに対応しています。

= CLUSTERPRO XE for Linux Ver3.0

2 動作概要

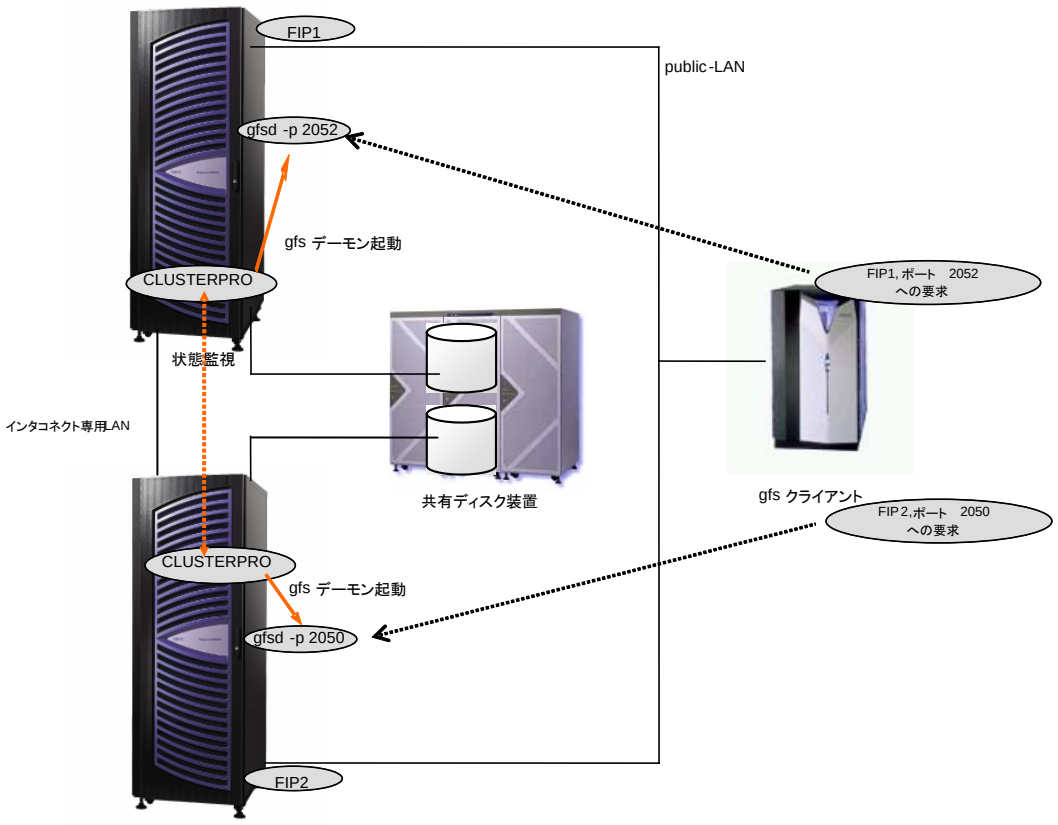
*　正常運用時

GFSのクライアント側からはCLSUTERPROが提供するフローティングIPでアクセスします。

GFSのデーモンの制御(起動/停止)[1]、フローティングIPの制御（活性化/非活性化）、および、共有ディスク上のファイルシステムの制御(マウント/アンマウント)はCLUSTERPROが内部で行います。

双方向スタンバイの場合にはインスタンス毎にGFSのポート番号を変更するのでGFSクライアント側からは該当するインスタンスのポート番号を指定してアクセスをしてください。

■正常運用時



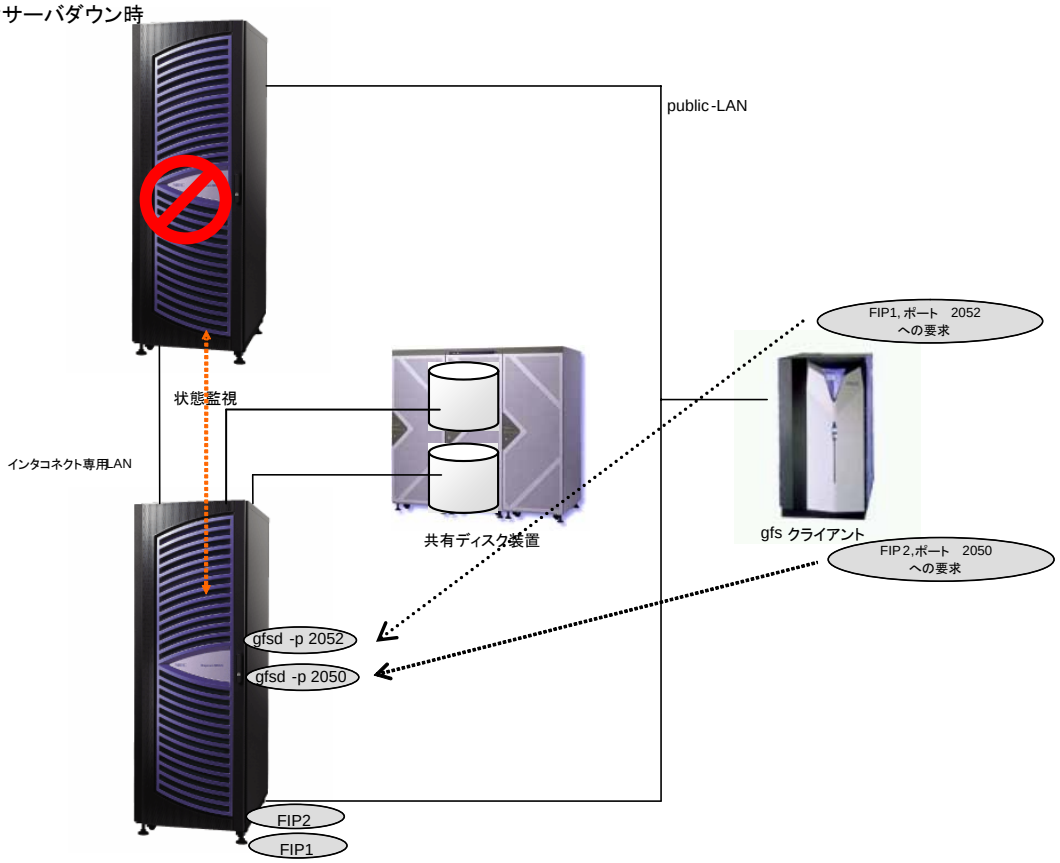
[1] CLUSTERPROで制御するのはgfsdのみです。gfsiod_s、gfsctl_sは後述のように/etc/rc.d/init.d/gfsにて制御(起動/停止)してください。GFSサーバ上のディスク装置とGFSクライアント側のディスク装置の関連付けは/etc/gfstab_sに記述してください。

*　片サーバダウン時

　相手サーバのCLUSTERPROから応答がなくなることをトリガとしてFIPと共有ディスク上のファイルシステムを正常なサーバ側へ移動します。

　さらにダウンしたサーバで動作していたポート番号のgfsdとFIPを起動します。GFSクライアント側からは正常時と同じFIPとポート番号でアクセスできます。

■片サーバダウン時



3 事前準備

3.1 GFS serverのインストール

GFSのリファレンスを参照してGFSサーバのインストールを行います。

3.2 CLUSTERPROのインストール

構築ガイド「クラスタ生成編」を参照してCLUSTERPROのインストール、構築を行います。

GFSに依存した設定は、CLUSTERPROトレッキングツールでクラスタ情報を作成する時に行います。

4 構築手順

4.1 GFSリソースの設計

GFSをクラスタリングする場合には、GFSに依存した以下のリソースの計画を立ててください。

切替えパーティション	GFSでshareするファイルシステムを構築するパーティションデバイスのスペシャルファイル名。 1つのフェイルオーバグループに複数のパーティションを登録することが可能です。 CLUSTERPROのdiskリソース、および、disk monitorリソースにおいて指定します。
マウントポイント （GFSサーバ側）	上記ファイルシステムをマウントするマウントポイント。登録した全てのディスクリソース数分必要です。
マウントポイント （GFSクライアント側）	クライアント上で、GFSサーバにてエクスポートされたファイルシステムをマウントするマウントポイント。
ポート番号	gfsdが使用するポート番号。 CLUSTERPROのexecリソース中の実行スクリプト(start.sh,stop.sh)において指定します。
フローティングIPアドレス	GFSクライアントからシームレスにアクセスするために使用するIPアドレス。(サーバの実IPと同じセグメントで他のIPアドレスと重複しないIPアドレスが必要) CLUSTERPROのfloating IPリソースにおいて指定します。

またGFSの設定に際して、上記以外に以下の情報が必要です。

ディスクリソース	GFSサーバ側、および、クライアント側から見たディスク装置のスペシャルファイル名。 GFSサーバ上のディスク装置と、GFSクライアント側のディスク装置の関連付けを行う時に指定します。 例) GFSサーバ側から見たディスク装置のスペシャルファイル名: /dev/scsi/host2/bus0/target16/lun0/disc GFSクライアント側から見たディスク装置のスペシャルファイル名: /dev/scd/FC101
ポート番号	gfsiodが使用するポート番号。 /etc/rc.d/init.d/gfsにおいて指定します。gfsdが使用するポート番号とは異なるものを指定してください。

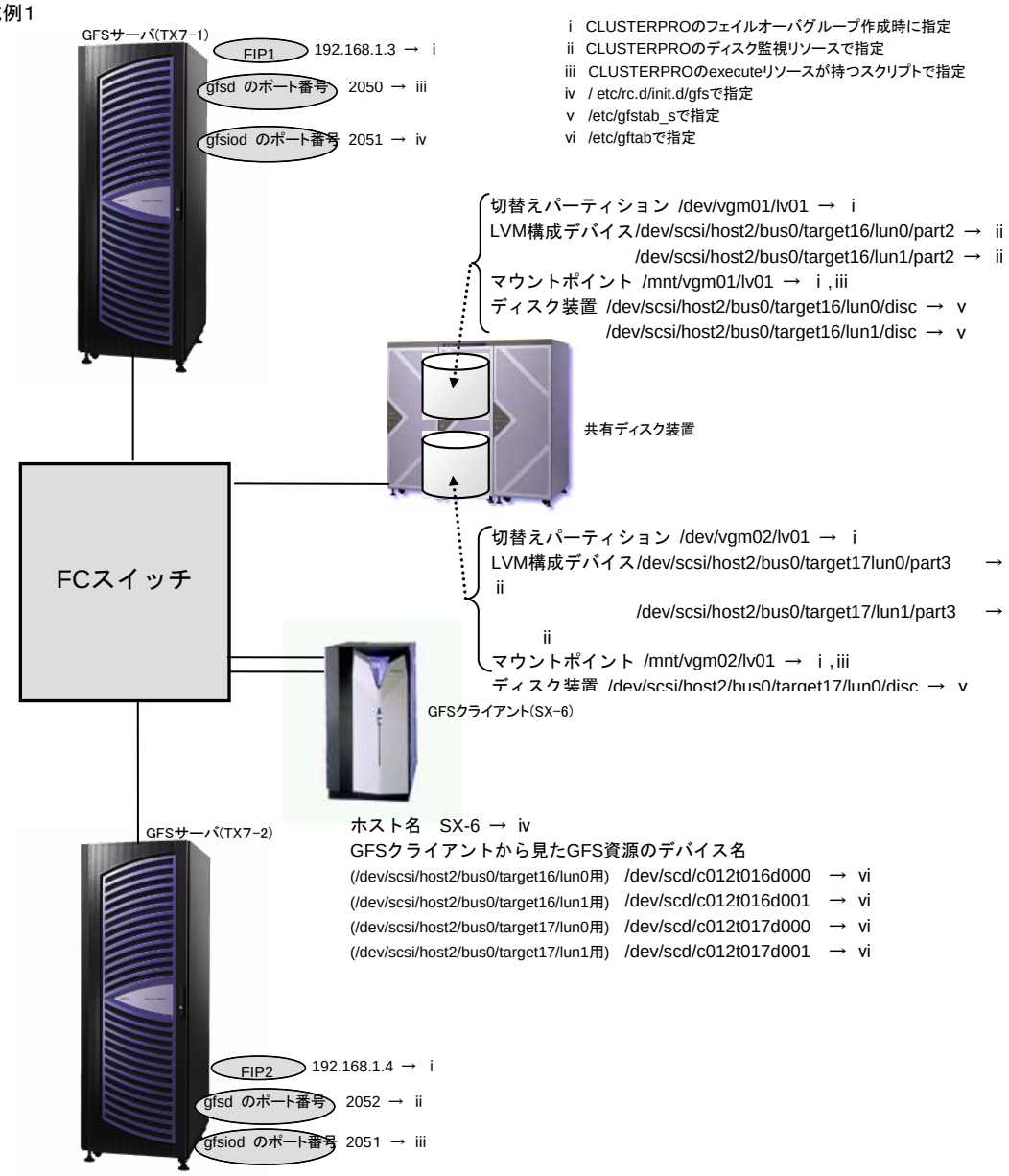
4.2 クラスタ情報の作成

トレッキングツールでクラスタ情報を作成します。クラスタ情報の作成の手順は「クラスタ生成編(共有ディスク)」と「トレッキングツール編」を参照してください。

既にクラスタを構築している場合には、クラスタの設定を変更します。クラスタ情報の変更の手順は「メンテナンス編」と「トレッキングツール編」を参照してください。

ここでは、下記構成でGFSを運用する場合を例として説明します。

■構成例1



(1) グループを作成する

GFS用のグループを作成します。

構成例の情報よりグループ、FG1を作成。

パラメータ名	設定値	備考
名前	FG1	既に作成したグループと重複しないこと。 マネージャ上に表示されるのでGFS用のグループであることが判りやすい名称にすることを推奨します。
コメント	任意	グループの特長を補足するコメントを設定してください。 ブランクのままでもかまいません。
起動サーバ	TX7-1 TX7-2	現用系サーバが最高プライオリティになるように設定してくださ。
グループ起動属性	自動起動	
フェイルオーバ排他属性[2]	排他なし	
自動フェイルバック属性	手動フェイルバック	

[2] 「フェイルオーバ排他属性」の詳細については、「CLUSTERPRO for Linux Ver3.0 トレッキングツール編」 5 パラメータ詳細 の章をご参照ください。

(2) グループにdiskリソースを割り当てる

（1）で作成したグループにdiskリソースを作成します。

パラメータ名		設定値	備考
名前		disk1	既に作成したリソースと重複しないこと。
コメント		任意	ブランクでもOK
デバイス名		/dev/vgm01/lv01	
マウントポイント		/mnt/lvgm01/lv01	
ファイルシステム		xfs	
活性異常時の復旧動作	活性リトライ	0回	
	フェイルオーバしきい値	1回	
	最終動作	グループ停止	
非活性異常時の復旧動作	非活性リトライしきい値	0回	
	最終動作	クラスタデーモン停止とOSシャットダウン	
依存関係		設定しない	「規定の依存関係に従う」のチェックをはずす

ディスクリソース調整プロパティの値は必要に応じて変更してください。規定値で特に問題がない場合は変更する必要はございません。

(3) グループにexecリソースを割り当てる

（1）で作成したグループにexecリソースを作成します。

パラメータ名		設定値	備考
名前		exec1	既に作成したリソースと重複しないこと。
コメント		任意	
スクリプトの種類		トレッキングツールで作成したスクリプト	
Start script		サンプルを次章の通り修正したもの	「置換」で置き換えるか「編集」で修正してください
Stop script		サンプルを次章の通り修正したもの	
活性異常時の復旧動作	活性リトライ	0回	
	フェイルオーバしきい値	1回	
	最終動作	グループの停止	
非活性異常時の復旧動作	非活性リトライしきい値	0回	
	最終動作	クラスタデーモン停止とOSシャットダウン	
依存関係		(2)で作成したdiskリソースに依存するように設定	「規定の依存関係に従う」のチェックをはずす

execリソースの調整プロパティ

パラメータ名		設定値	備考
開始スクリプト	同期設定	同期	
	タイムアウト	1800秒	
停止スクリプト	同期設定	同期	
	タイムアウト	1800秒	
ログ出力先		任意	ログ出力先のディレクトリが存在しない場合はグループの起動時にエラーとなります

execリソース調整プロパティの値は必要に応じて変更してください。規定値で特に問題がない場合は変更する必要はございません。

(4) グループにfloating IPリソースを割り当てる

(1)で作成したグループにfloating IPリソースを追加します。

パラメータ名		設定値	備考
名前		fip1	既に作成したリソースと重複しないこと。
コメント		任意	
IPアドレス		192.168.1.3	
活性異常時の復旧動作	活性リトライ	0回	
	フェイルオーバしきい値	1回	
	最終動作	グループの停止	
非活性異常時の復旧動作	非活性リトライしきい値	0回	
	最終動作	クラスタデーモン停止とOSシャットダウン	
依存関係		（3）で作成したexecリソースに依存するように設定	「規定の依存関係に従う」のチェックをはずす

フローティングIPリソース調整プロパティの値は必要に応じて変更してください。規定値で特に問題がない場合は変更する必要はございません。

(5) Active-Active構成を組むときは 必要なグループ数分(1)～(4)を繰り返す

(1)の「起動サーバ」の設定で現用系サーバを変更して(1)～(4)を繰り返します。

(6) ディスク監視リソースを作成する

（2）で作成したディスクリソースを監視するリソースを作成します。
構成例1のケースで、デバイス”/dev/vgm01/lv01”を監視するためには以下の2つのディスク監視リソースを作成します。

diskw1

パラメータ名	設定値	備考
名前	diskw1	既に作成したリソースと重複しないこと。
コメント	任意	ブランクでもOK
監視デバイス名	/dev/scsi/host2/bus0/target16/lun0/part2	注意点1
監視方法	TUR	注意点2
回復対象	FG1	
再活性化しきい値	3	
フェイルオーバしきい値	1	
最終動作	グループ停止	

diskw2

パラメータ名	設定値	備考
名前	diskw2	既に作成したリソースと重複しないこと。
コメント	任意	ブランクでもOK
監視デバイス名	/dev/scsi/host2/bus0/target16/lun1/part2	注意点1
監視方法	TUR	注意点2
回復対象	FG1	
再活性化しきい値	3	
フェイルオーバしきい値	1	
最終動作	グループ停止	

注意点

1. 監視デバイス名にLVMのデバイスファイル名を指定することはできません。
   LVMの構成要素のパーティションデバイス名を指定してください。
2. 監視方法にDummy readを使用することはできません。TURを指定してください。

(7) IP監視リソースを作成する

クラスタを構成するサーバが外部と正常に通信できることを監視するためにIP監視リソースを作成します。

ここでは、192.168.1.0/24ネットワークに2台のゲートウェイ（192.168.1.252,192.168.1.253）が存在するとして設定例を示します。

パラメータ名	設定値	備考
名前	ipw1	既に作成したリソースと重複しないこと。
コメント	任意	ブランクでもOK
IPアドレス	192.168.1.252	注意点1 注意点2
	192.168.1.253	
回復対象	FG1	
再活性化しきい値	3	
フェイルオーバしきい値	1	
最終動作	グループ停止	

注意点

1. この監視機能は登録されたIPアドレスが有効かどうかを監視するものではなく、登録されたアドレスに通信できるかを監視するものです。このセクションにCLUSTERPROで提供するfloating IP アドレスを登録した場合、グループを非活性化したときに異常が検出されてしまいます。
2. このセクションに登録された全てのアドレスに対して通信が行えない場合に、監視リソースの状態が異常となります。

4.3 GFSサーバの設定

クラスタを構成する環境に合わせて、以下のGFSのスクリプトを編集してください。
ここでは前節の「構成例1」を参考に説明します。

(1) /etc/rc.d/init.d/gfs

Linux側のOSのGFSのinitスクリプトのstart( ), stop( )のルーチンを以下のように修正してください。
start( )では"/usr/sbin/gfsctl_s -a"の実行、および、/usr/sbin/gfsiod_s の起動を行い、/usr/sbin/gfsdの起動は行わないように設定してください。
stop( )ではgfsiod_sのみ停止するようにします。下記の修正例を参考にしてください。

(オリジナルのGFSのinitスクリプトは起動しないようにして、オリジナルの/etc/rc.d/init.d/gfsを元にコピーを作成して、使用するLinuxのランレベルの/etc/rc.d/rc.X/から新たにシンボリックリンクを貼ることをお奨めします。)

* サンプルスクリプト
下記のstart( ),stop( )ルーチンは***あくまでサンプルですので実環境に合わせて作成***をしてください。
ここではgfsiod_sのデーモンにポート番号2051を指定しています。

```
start() {
      if [ $GFSSRV = 1 ]; then
            # Start daemons.
            action $"Starting Global Filesystem services (Server): " "/usr/sbin/gfsctl_s -a"
            echo
#           echo -n $"Starting GFS daemon: "
#           daemon /usr/sbin/gfsd ${GFSDCOUNT}
#           echo
            echo -n $"Starting GFS IO daemon (server)     : "
            daemon /usr/sbin/gfsiod_s -s ${GFSDSCOUNT}
            echo
            sleep 1
            echo -n $"Starting GFS IO daemon (receiver) : "
            daemon /usr/sbin/gfsiod_s -r –p 2051 ${GFSDSCOUNT}
            echo
      fi

      touch /var/lock/subsys/gfs
}
stop() {
            # Stop daemons.
            echo -n $"Shutting down GFS daemon: "
            PID=`ps -e | egrep 'gfsiod' | sed -e 's/^   *//' -e 's/ .*//'`
            if [ "$PID" != "" ]; then
                  kill -9 ${PID}
            fi
            echo
}
```

(2) /etc/gfstab_s

GFSサーバ上のディスク装置と、GFSクライアント側のディスク装置を関連付ける設定を記述します。ディスクリソースが所属するフェイルオーバグループの優先起動サーバが自サーバであるか否かに関わらず、GFSで使用する全てのディスクを記述する必要があります。

構成例1では以下のようになっているので、/etc/gfstab_sにサンプルスクリプトのように4行追加します。

（GFSクライアントのホスト名）
SX6

（GFSサーバ上のディスク装置のスペシャルデバイス名）
/dev/scsi/host2/bus0/target16/lun0/disc
/dev/scsi/host2/bus0/target16/lun1/disc
/dev/scsi/host2/bus0/target17/lun0/disc
/dev/scsi/host2/bus0/target17/lun1/disc

（GFSクライアント側のディスク装置のスペシャルデバイス名）
/dev/scd/c012t016d000
/dev/scd/c012t016d001
/dev/scd/c012t017d000
/dev/scd/c012t017d001

* サンプルスクリプト

```
/dev/scsi/host2/bus0/target16/lun0/disc SX6 /dev/scd/c012t016d000
/dev/scsi/host2/bus0/target16/lun1/disc SX6 /dev/scd/c012t016d001
/dev/scsi/host2/bus0/target17/lun0/disc SX6 /dev/scd/c012t017d000
/dev/scsi/host2/bus0/target17/lun1/disc SX6 /dev/scd/c012t017d001
```

4.4 NFSサーバに関する設定

GFSサーバが動作するためにはNFSサーバが動作している必要があります。以下の設定を確認してください。

(1) システム起動時にnfsdが起動すること（CLUSTERPROにおいて開始/停止の制御を行わないでください）。

(2) システム起動時に以下のコマンドラインが実行されること。

```
# echo 1 > /proc/sys/nec/nfs-syncmode
```

上記はGFSサーバのフェイルオーバを実現するための設定です。

5 スクリプト

GFSの設定にあわせてexecリソースに登録するスクリプトを編集します。
ここでは4.2節の「構成例1」を参考に説明します。

5.1 start.sh

start.shではGFSに依存して下記の記述が必要です。

+ GFS デーモンの起動 (gfsd)
+ クライアントがマウントするディレクトリのエクスポート(exportfs)
  ※exportfsのオプションとして以下の2点を指定する必要があります。
  1. sync
  2. no_wdelay

* サンプルスクリプトの説明

次ページのstart.shは***あくまでサンプルですので実環境に合わせて作成***をしてください。

/usr/sbin/gfsd から /bin/chmod までがGFS依存の起動処理です。下記を環境とフェイルオーバグループによって修正してください。

i. gfsdのポート番号（*2050*）
ii. 共有ディスク上のファイルシステムのマウントポイント(*/mnt/vgm01/lv01*)

次ページのstart.sh内の上記値は、前章の4.2節で作成したフェイルオーバグループを想定した値になっています。別のフェイルオーバグループ用スクリプトを作成するときにはフェイルオーバグループ間で値が重複しないように予め計画をしてください

```
#! /bin/sh
#****************************************
#*                  start.sh                  *
#****************************************

if [ "$CLP_EVENT" = "START" ]
then
      if [ "$CLP_DISK" = "SUCCESS" ]
      then
            echo "NORMAL1"


###########BEGIN############

            /usr/sbin/gfsd -p 2050 4
            /usr/sbin/exportfs -o rw,sync,no_wdelay,insecure,no_root_squash  SX6:/mnt/vgm01/lv01
            /bin/chmod 777 /mnt/vgm01/lv01

############END#############


            if [ "$CLP_SERVER" = "HOME" ]
            then
                  echo "NORMAL2"
            else
                  echo "ON_OTHER1"
            fi
      else
            echo "ERROR_DISK from START"
      fi
elif [ "$CLP_EVENT" = "FAILOVER" ]
then
      if [ "$CLP_DISK" = "SUCCESS" ]
      then
            echo "FAILOVER1"


###########BEGIN############

            /usr/sbin/gfsd -p 2050 4
            /usr/sbin/exportfs -o rw,sync,no_wdelay,insecure,no_root_squash  SX6: /mnt/vgm01/lv01
            /bin/chmod 777 /mnt/vgm01/lv01

############END#############


            if [ "$CLP_SERVER" = "HOME" ]
            then
                  echo "FAILOVER2"
            else
                  echo "ON_OTHER2"
            fi
      else
            echo "ERROR_DISK from FAILOVER"
      fi
else
      echo "NO_CLP"
fi
echo "EXIT"
exit 0
```

5.2 stop.sh

stop.shではGFSに依存して下記の記述が必要です。

+ GFS デーモンの停止 (gfsd)
+ クライアントがマウントするディリクトリのアンエクスポート（exportfs）

* サンプルスクリプトの説明

次ページのstop.shは***あくまでサンプルですので実環境に合わせて作成***をしてください。

/usr/sbin/gfsd から /bin/chmod までがGFS依存の停止処理です。最後の”sleep”共有ファイルシステムを使用しているアプリケーションが停止するのを待つための処理です。下記を環境とフェイルオーバグループによって修正してください。

i. gfsdのポート番号（*2050*）
ii. 共有ディスク上のファイルシステムのマウントポイント(*/mnt/vgm01/lv01*)

次ページのstop.sh内の上記値は、前章の4.2節で作成したフェイルオーバグループを想定した値になっています。別のフェイルオーバグループ用スクリプトを作成するときにはフェイルオーバグループ間で値が重複しないように予め計画をしてください

```
#! /bin/sh
#****************************************
#*                  stop.sh                    *
#****************************************

if [ "$CLP_EVENT" = "START" ]
then
      if [ "$CLP_DISK" = "SUCCESS" ]
      then
            echo "NORMAL1"


############BEGIN#############

            PID=`ps -e | egrep 'gfsd' | egrep '2050' | awk '{print $1}'`
            if [ "$PID" != "" ]; then
            kill -9 ${PID}
            fi

            rval=1

            while [ $rval -ne 0 ]
            do
              # timeout
                  sleep 1
            ps -e | awk '
            BEGIN {
            found = 0;
                  }
            {
            if ($4 == "gfsd" && $6 == "2050") {
                        found++;
                  }
                  }
            END {
                     exit found;
                   }'
              rval=$?
            done
            /usr/sbin/exportfs -u SX6:/mnt/vmg01/lv01
            sleep 30

#############END##############
```

```
        if [ "$CLP_SERVER" = "HOME" ]
        then
            echo "NORMAL2"
        else
            echo "ON_OTHER1"
        fi

    else
        echo "ERROR_DISK from START"
    fi
elif [ "$CLP_EVENT" = "FAILOVER" ]
then
    if [ "$CLP_DISK" = "SUCCESS" ]
    then
        echo "FAILOVER1"



###########BEGIN#############

        PID=`ps -e | egrep 'gfsd' | egrep '2050' | awk '{print $1}'`
        if [ "$PID" != "" ]; then
        kill -9 ${PID}
        fi

        rval=1

        while [ $rval -ne 0 ]
        do
          # timeout
            sleep 1
        ps -e | awk '
        BEGIN {
        found = 0;
            }
        {
        if ($4 == "gfsd" && $6 == "2050") {
                found++;
            }
            }
        END {
              exit found;
             }'
          rval=$?
        done
        /usr/sbin/exportfs -u SX6:/mnt/vmg01/lv01
        sleep 30

############END#############



        if [ "$CLP_SERVER" = "HOME" ]
        then
            echo "FAILOVER2"
        else
            echo "ON_OTHER2"
        fi

    else
        echo "ERROR_DISK from FAILOVER"
    fi
else
    echo "NO_CLP"
fi
echo "EXIT"
exit 0
```

6 運用上の注意点

GFSサーバ機能を二重化する際に留意して頂きたい事項です。

* 本書で記載されているGFSサーバの二重化運用は、サーバダウン時にフェイルオーバを実現することを目的としております。サーバダウンが発生した場合、その際にGFSクライアントから発行され実行中であったI/Oは継続できます。しかし、GFSサーバとディスクの間の経路障害はフェイルオーバされません。

* GFSのセッションを維持するために同一クラスタ内のすべてのサーバで共有ディスクについて下記が必要です。
  + 同一デバイスで見えること
  + 同一メジャー/マイナ番号で見えること

  サーバにより内蔵ディスクの構成が異なる場合には、上記の構成が実現できない場合があります。導入前に構成を充分確認ください。

* /etc/exportfsにフェイルオーバ対象となるファイルシステムは記述しないでください。

* フェイルオーバ発生時にクライアントからのNFS経由のアクセスがタイムアウトする可能性があります。
  タイムアウトが発生してもクライアント側から再マウントの必要はありません。

* フェイルオーバ発生時にクライアントからのGFS経由のアクセスがタイムアウトする場合にはGFSのタイムアウト値を変更してください。
  フェイルオーバに必要な時間はユーザ環境によって異なりますので必ず実環境でフェイルオーバを発生させて問題がないことを確認してください。