IBM Highly Available Cluster Multiprocessing for AIX (HACMP6.1)

IBM Highly Available Cluster Multiprocessing for AIX, (HACMP) provides the ability to keep business-critical applications and systems operational 7 days per week, 24 hours per day. An HACMP solution helps avoid downtime, enables prompt recovery from any hardware, network and application failures, and also gives you the means to take down an individual server (node) for planned maintenance and upgrades without having to take down the entire cluster.

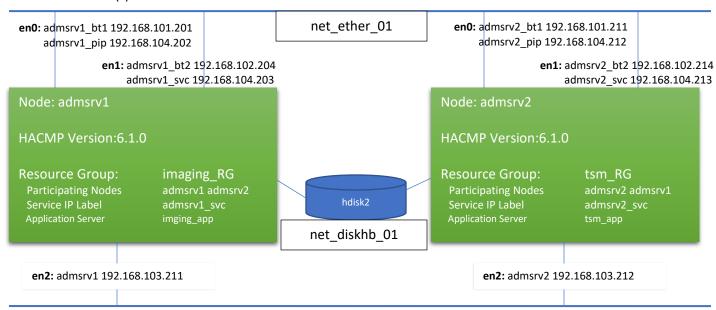
Cluster Name: adminserv_cluster

Cluster Connection Authentication Mode: Standard
Cluster Message Authentication Mode: None
Cluster Message Encryption: None
Use Persistent Labels for Communication: No

HACMP Nodes and Networks Configure Planning

General network connection consideration: An example of correct HACMP networking consists of two separate Ethernet networks, each with two network interfaces on each node. Two routers connect the networks, and route packets between the cluster and clients, but not between the two networks.

There are 2 node(s) defined



To avoid cluster partitioning, 2 networks, One TCPIP Network: net_ether_01, one Non-IP Network: netdiskhb 01.

NODE admsrv1:	NODE admsrv2:
Network net_diskhb_01	Network net_diskhb_01
admsrv1_hdisk2_01 /dev/hdisk2	admsrv2_hdisk2_01 /dev/hdisk2
Network net_ether_01	Network net_ether_01

admsrv1_svc	192.168.104.203	admsrv1_svc	192.168.104.203
admsrv2_svc	192.168.104.213	admsrv2_svc	192.168.104.213
admsrv1_bt1	192.168.101.201	admsrv2_bt1	192.168.101.211
admsrv1_bt2	192.168.102.204	admsrv2_bt2	192.168.102.214

Before using the Two-Node Cluster Configuration Assistant to configure an HACMP cluster definition, make sure that:

- The node running the Assistant has start and stop scripts for the application(s) to be made highly available.
- Both nodes have TCP/IP connectivity to each other.
- Both nodes are physically connected to all disks configured within the volume groups.
- Both nodes have the HACMP software and the same version of the RSCT software.
- Both nodes have a copy of the application that is to be highly available.
- The etc/hosts file on both nodes is configured with a service IP label/address to be specified in the Assistant.

Let's start from Configure an HACMP Cluster - > Remove an HACMP Cluster if there is any cluster exist

Prepare 2 basic network configuration/authorization files: .rhosts and /etc/hosts

```
root@admsrv1:/# cat /.rhosts
admsrv1 root
admsrv2 root
nimserver root
root@admsrv1:/ # cat /etc/hosts
            1.1 src/bos/usr/sbin/netstart/hosts, cmdnet, bos530 7/24/91 10:00:46
#IBM PROLOG BEGIN TAG
# This is an automatically generated prolog.
# bos530 src/bos/usr/sbin/netstart/hosts 1.1
# Licensed Materials - Property of IBM
# Internet Address Hostname # Comments
# 192.9.200.1 netOsample # ethernet name/address
# 128.100.0.1 tokenOsample # token ring name/address
# 10.2.0.2 x25sample # x.25 name/address
127.0.0.1
               loopback localhost # loopback (lo0) name/address
# Comunication Interfaces
192.168.104.201 admsrv1 admsrv1.livingstonintl.com #configure it on en0 of admsrv1 as alias
192.168.104.211 admsrv2 admsrv2.livingstonintl.com #configure it on en0 of admsrv2 as alias
# Management Interfaces
#192.168.103.211 admsrv1 admsrv1.livingstonintl.com
#192.168.103.212 admsrv2 admsrv2.livingstonintl.com
# HACMP ADMSRV1 Interfaces
192.168.101.201 admsrv1 bt1 admsrv1 bt1.livingstonintl.com #configure it on en0 of admsrv1
192.168.104.202 admsrv1_pip_admsrv1_pip.livingstonintl.com
192.168.103.203 admsrv1_svc admsrv1_svc.livingstonintl.com adminserv
192.168.102.204 admsrv1_bt2 admsrv1_bt2.livingstonintl.com #configure it on en1of admsrv1
# HACMP ADMSRV2 Interfaces
192.168.101.211 admsrv2 bt1 admsrv2 bt1.livingstonintl.com #configure it on en0 of admsrv2
```

192.168.104.212 admsrv2_pip_admsrv2_pip.livingstonintl.com

Prepare/define/configure Communication Networks/Interfaces/Devices

smitty tcpip - > Minimum Configuration for Boot IPs define (Further Configuration for Alias address define)

IP Label	Network	Туре	Node	Address	Interface	Netmask
admsrv1_hdisk2_01	net_diskhb_01	diskhb	admsrv1	/dev/hdisk2	hdisk2	
admsrv1_svc	net_ether_01	ether	admsrv1	192.168.103.203	dynamic	255.255.255.0/24
admsrv1_pip	net_ether_01	ether	admsrv1	192.168.104.202	dynamic	255.255.255.0/24
admsrv1_bt1	net_ether_01	ether	admsrv1	192.168.101.201	en0	255.255.255.0/24
admsrv1_bt2	net_ether_01	ether	admsrv1	192.168.102.204	en1	255.255.255.0/24
admsrv2_hdisk2_01	net_diskhb_01	diskhb	admsrv2	/dev/hdisk2	hdisk2	
admsrv2_svc	net_ether_01	ether	admsrv2	192.168.103.213	dynamic	255.255.255.0/24
admsrv2_pip	net_ether_01	ether	admsrv2	192.168.104.212	dynamic	255.255.255.0/24
admsrv2_bt1	net_ether_01	ether	admsrv2	192.168.101.211	en0	255.255.255.0/24
admsrv2_bt2	net_ether_01	ether	admsrv2	192.168.102.214	en1	255.255.255.0/24

Four subnets involved: 192.168.101.0; 192.168.102.0;192.168.103.0; 192.168.104.0

- All boot addresses must be defined on different subnets(101 for bt1, 102 for bt2).
- Service addressesmust be on a different subnet(103) from all boot addresses(101,102) and persistent addresses(104).

A clinfo.rc file is installed on each node in the cluster, containing the IP addresses of several client machines.

Service interface (admsrv1_svc, admsrv2_svc):

A service interface is a communications interface configured with an HACMP service IP label. This interface serves as each node's primary HACMP connection to each network. The service IP label is used in the following ways:

- By clients to access application programs
- For HACMP heartbeat traffic.

With IPAT via aliasing, during fallover, the service IP label is aliased onto the boot interface along with the heartbeat alias. (Another IP Address Takeover method is IPAT via replacement, during fallover, the service IP label is swapped with the boot time address, not with the heartbeating alias IP address).

Boot interface (admsrv1_bt1, admsrv1_bt2; admsrv2_bt1, admsrv2_bt2):

A boot interface is a communications interface with an HACMP boot IP label that backs up a service interface. All client traffic is carried over the service interface. Boot interfaces are hidden from client applications and carry only internal HACMP traffic. If a service interface fails, HACMP can move the service IP label onto a boot interface. Using a boot interface eliminates a network interface as a single point of failure.

In addition, if a node fails, the cluster can use a boot interface on another cluster node as a location for its service IP label when performing a resource group fallover.

A node can have from zero to seven boot interfaces for each network to which it connects. Your software configuration and hardware constraints determine the actual number of boot interfaces that a node can support.

Persistent node IP label (admsrv1 pip, admsrv2 pip):

A persistent node IP label is an IP alias that can be assigned to a specific node on a cluster network. A persistent node IP label:

- Can be fallover to another NIC of the same node when old NIC failed, but Always stays on the same node (is node-bound) even all NICs on this node fail, it lost then.
- Coexists on a NIC that already has a service or boot IP label defined
- Does not require installing an additional physical NIC on that node
- Is not part of any resource group.

Assigning a persistent node IP label provides a node-bound address that you can use for administrative purposes, because a connection to a persistent node IP label always goes to a specific node in the cluster.

You can have one persistent node IP label per network per node.

Non-IP point-to-point heartbeating network (net_diskhb_01), called a disk heartbeating network, over any shared disk in an enhanced concurrent mode volume group. A heartbeating network ensures that each node always has a communication path to the other nodes - even if a network fails. This prevents your cluster from becoming partitioned. Otherwise, a network failure may cause nodes to attempt to take over resource groups that are still active on other nodes. In this situation, if you have set a forced varyon setting, you may experience data loss or divergence.

In a disk heartbeating network, two nodes connected to the disk periodically write heartbeat messages and read heartbeat messages (written by the other node) on a small, non-data portion of the disk. A disk heartbeating network, like the other non-IP heartbeating networks, connects only two nodes. In clusters with more than two nodes, use multiple disks for heartbeating. Each node should have a non-IP heartbeat path to at least one other node. If the disk heartbeating path is severed, at least one node cannot access the shared disk.

You have two different ways for configuring a disk heartbeating network in a cluster:

- Create an enhanced concurrent volume group shared by multiple nodes in your cluster. Then use the HACMP Extended Configuration SMIT path to configure a point-to-point pair of discovered communication devices.
- Creating a cluster disk heartbeating network(admsrv1_hdisk2_01), and then add devices to it using the Add Pre-Defined Communication Interfaces(admsrv1_hdisk2_01, admsrv2_hdisk2_01) and Devices(/dev/hdisk2) panel in SMIT.

The HACMP cluster verification utility verifies that the disk heartbeating networks are properly configured. # smitty hacmp -> configure HACMP Networks - > Manage Concurrent Volume Group for Muti-Node Disk Heartbeat - > Create a new Volume Group and Logic Volume for Muti-Node Disk Heartbeat choose hdisk2

DNS

To ensure that cluster event completes successfully and quickly, HACMP disables NIS or DNS hostname resolution by setting the following AIX environment variable during service IP label swapping: NSORDER = local

As a result, the /etc/hosts file of each cluster node must contain all HACMP defined IP labels for all cluster nodes.

root@admsrv1# cat /etc/netsvc.conf #@(#)43 1.1 src/tcpip/etc/netsvc.conf, tcpip, tcpip530 4/3/02 22:12:29 #IBM PROLOG BEGIN TAG

```
# This is an automatically generated prolog.
# tcpip530 src/tcpip/etc/netsvc.conf 1.1
# Licensed Materials - Property of IBM
# Restricted Materials of IBM
# (C) COPYRIGHT International Business Machines Corp. 2002
# All Rights Reserved
# The /etc/netsvc.conf file is used to specify the ordering of name resolution for
# gethostbyname, gethostbyname2, gethostbyaddr routines
# and alias resolution for the sendmail command.
# Any host setting defined in /etc/netsvc.conf file will override the default host order
# and the host setting given in the /etc/irs.conf file.
# The environment variable NSORDER overrides the host settings in the /etc/netsvc.conf file,
# which in turn overrides the host settings in the /etc/irs.conf file.
# To specify the host ordering, enter the following:
# hosts = value [, value]
# Use one or more of the following values for the hosts keyword:
# auth
           Designates the specified server as authoritative. A resolver doesnot continue searching
        for host names further than an authoritative server. For example,
#
        when two services are given as values for the host keyword and the first service
#
        is made authoritative, and if the resolver cannot find the host name in the
        authoritative service, then the resolver terminates its search. However,
#
        the auth option has no effect if the resolver is unable to contact
#
        the authoritative server; in this case, the resolver continues to search the next
#
        service given in the same entry.
#
        Indicate that the specified service is authoritative by followingit by an = and then auth.
#
        The auth option is only valid when used in conjunction with a service value for the hosts keyword.
# bind
          Uses BIND/DNS services for resolving names
# local
          Searches the local /etc/hosts file for resolving names
# nis
         Uses NIS services for resolving names. NIS must be running if you specify this option
# nis+
          Uses NIS plus services for resolving names. NIS plus must be running if you specify this option
          Uses LDAP services for resolving names
# Idap
# Idap_nis Uses LDAP NIS services for resolving names
# hind4
           Uses BIND/DNS services for resolving only IPv4 addresses
# bind6
           Uses BIND/DNS services for resolving only IPv6 addresses
# local4 Searches the local /etc/hosts file for resolving only IPv4 addresses
# local6 Searches the local /etc/hosts file for resolving only IPv6 addresses
# nis4
          Uses NIS services for resolving only IPv4 addresses
# nis6
          Uses NIS services for resolving only IPv6 addresses
# nis+4
          Uses NIS plus services for resolving only IPv4 addresses
# nis+6
          Uses NIS plus services for resolving only IPv6 addresses
           Uses LDAP services for resolving only IPv4 addresses
# Idap4
# Idap6
           Uses LDAP services for resolving only IPv6 addresses
# Idap_nis4 Uses NIS LDAP services for resolving only IPv4 addresses
# Idap_nis6 Uses NIS LDAP services for resolving only IPv6 addresses
# Any value a user specified other than the ones above, is considered as a user option.
#The user option format is: <key>[none|4|6], where<key> is the name of the dynamic loadable module
# that is going to reside under /usr/lib/netsvc/dynload/ directory.
```

```
# Following the <key> can be nothing, 4, or 6 which represents the address family.
# If we configure a user option as "dave4", then there should be a dynamic loadable module
# "dave.so" under /usr/lib/netsvc/dynload/ directory.
# "dave4" user option tells the resolver to use /usr/lib/netsvc/dynload/dave.sodynamic loadable module
# to do the name resolution and resolve only IPv4 addresses.
# Example:
# hosts = nis=auth, bind6, dave4
# The sendmail command searches the local /etc/aliases file, or uses NIS if specified for
# resolving aliases. You can override the default by specifying how to resolve aliases
# in the /etc/netsvc.conf file.
# To specify alias ordering to the sendmail command, enter the following:
# alias = value [, value]
# Use one or more of the following values for the alias keyword:
# files Searches the local /etc/aliases file for the alias
# nis Uses NIS services for resolving alias
# Example:
# aliases = nis, files
# hosts=local,bind
hosts=local
```

The length of the <key> can be between 1 to 8 characters.

Define Shared Volume Group and Filesystems Storage distribution (layout):

VG	Shared/	Node	Disks	LV Name	File System
name	Major Number				
rootvg	No	admsrv1	hdisk0 (mirror)	hd5	N/A
		admsrv2	hdisk1 (mirror)	hd6	N/A
				hd8	N/A
				hd4	1
				hd2	/usr
				hd9var	/var
				hd3	/tmp
				hd1	/home
				hd10opt	/opt
				lg_dumplv	N/A
				paging00	N/A
				ibmlv	/ibm
				admsrvlv	/admsrv
				livedump	/var/adm/ras/livedump
				hd11admin	/admin
havg		admsrv1&admsrv2	hdisk2	hahblv	/hahb
appvg	enhanced	admsrv1&admsrv2	hdisk18	jfs2logapp	N/A
	concurrent		hdisk19	cmapplv	/cmapp
				tsmha1lv	/tsmha1
cmvg	enhanced	admsrv1&admsrv2	hdisk3	jfs2logcm	N/A
	concurrent		hdisk4	db2inst1lv	/home/db2inst1
			hdisk5	db2fenc1lv	/home/db2fenc1
			hdisk6	db2inst2lv	/home/db2inst2

			hdisk7	db2lslogginglv	/db2lslogging
				db2rmlogginglv	/db2rmlogging
		1 22 1 1	1 1: 100	db2fenc2lv	/home/db2fenc2
tsmvg	enhanced concurrent	admsrv2&admsrv1	hdisk20	jfs2logtsm	N/A
	107		hdisk21	tsmsrvlv	/tsmsrv
			hdisk22	tsmstglv	/tsmstg
			hdisk23	tsmha2	/tsmha2
docvg	enhanced concurrent	admsrv1&admsrv2	hdisk8	jfs2logdoc	N/A
			hdisk9	ubosstglv	/ubosstg
			hdisk10	lbosdata01lv	/lbosdata01
			hdisk11	lbosdata02lv	/lbosdata02
			hdisk12	lbosdata03lv	/lbosdata03
			hdisk13	lbosdata04lv	/Ibosdata04
			hdisk14	lbosdata05lv	/lbosdata05
			hdisk15	lbosdata06lv	/lbosdata06
			hdisk16	lbosdata07lv	/Ibosdata07
			hdisk17	lbosdata08lv	/lbosdata08
				lbosdata09lv	/lbosdata09
				arstmplv	/arstmp
docvg2	enhanced	admsrv1&admsrv2	hdisk24	jfs2logdoc2	N/A
J	concurrent		hdisk25	lbosdata10lv	/lbosdata10
			hdisk26	lbosdata11lv	/lbosdata11
			hdisk27	lbosdata12lv	/lbosdata12
			hdisk28	lbosdata13lv	/lbosdata13
				lbosdata14lv	/lbosdata14
				lbosdata15lv	/lbosdata15
docvg3	enhanced	admsrv1&admsrv2	hdisk32	jfs2logdoc3	N/A
	concurrent		hdisk33	lbosdata16lv	/lbosdata16
			hdisk34	lbosdata17lv	/lbosdata17
			hdisk35	lbosdata18lv	/lbosdata18
			hdisk36	lbosdata19lv	/lbosdata19
			hdisk37	lbosdata20lv	/lbosdata20
			hdisk38	lbosdata21lv	/lbosdata21
			hdisk39	lbosdata22lv	/lbosdata22
				lbosdata23lv	/Ibosdata23
				lbosdata24lv	/lbosdata24

HACMP and **VG**'s active and passive varyon

HACMP correctly varies on the volume group in active state on the node that owns the resource group, and changes active and passive states appropriately as the state and location of the resource group changes.

Upon cluster startup:

- On the node that owns the resource group, HACMP activates the volume group in active state. Note that HACMP activates a volume group in active state only on one node at a time.
- HACMP activates the volume group in passive state on all other nodes in the cluster.

Upon fallover:

- If a node releases a resource group, or, if the resource group is being moved to another node for any other reason, HACMP switches the varyon state for the volume group from active to passive on the node that releases the resource group, and activates the volume group in active state on the node that acquires the resource group.
- The volume group remains in passive state on all other nodes in the cluster.

Upon node reintegration, HACMP does the following:

- Changes the varyon state of the volume group from active to passive on the node that releases the resource group
- Varies on the volume group in active state on the joining node
- Activates his volume group in passive state on all other nodes in the cluster.

Note: The switch between active and passive states is necessary to prevent mounting file systems on more than one node at a time.

Note: Passive state varyon allows only a limited number of read-only operations on the volume group:

- LVM read-only access to the volume group's special file
- LVM read-only access to the first 4K of all logical volumes that are owned by the volume group.

Note: The following operations are not allowed when a volume group is varied on in passive state:

- Operations on file systems, such as file systems mounting
- Any operations on logical volumes, such as having logical volumes open
- Synchronizing volume groups.

Define Cluster with Two_node Cluster Configuration Assistant

cluster Name: Ims

Node Names(primary Node: admsrv1/Secondary Node: admsrv2)

Resource Group(Name, Service IP, Application Server Name, & start/stop script)

Resource Group & Application Server Configuration

Resource Group Name	imaging_RG	
Participating Nodes (Default Node Priority)	admsrv1 admsrv2	
Startup Policy	Online On Home Node Only	
Fallover Policy	Fallover To Next Priority Node In The List	
Fallback Policy	Fallback To Higher Priority Node In The List	
Fallback Timer Policy (empty is immediate)	D ,	+
Service IP Labels/Addresses	[admsrv1_svc]	+
Application Servers	[imaging_app]	+
Volume Groups	[cmvg docvg appvg docvg2 docvg3]	+
Use forced varyon of volume groups, if necessary	false	+
Automatically Import Volume Groups	false	+
Filesystems (empty is ALL for VGs specified)	[]	+
Filesystems Consistency Check	fsck	+
Filesystems Recovery Method	sequential	+
Filesystems mounted before IP configured	false	+
Filesystems/Directories to Export (NFSv2/3)		+
		+
Filesystems/Directories to NFS Mount		
Network For NFS Mount		+

Resource Group Name

Participating Nodes (Default Node Priority)

Startup Policy Fallover Policy Fallback Policy

Fallback Timer Policy (empty is immediate)

Service IP Labels/Addresses Application Servers

Volume Groups

Use forced varyon of volume groups, if necessary Automatically Import Volume Groups

 $\label{eq:Filesystems} \textit{Filesystems (empty is ALL for VGs specified)}$

Filesystems Consistency Check Filesystems Recovery Method

Filesystems mounted before IP configured

Filesystems/Directories to Export (NFSv2/3)

Filesystems/Directories to NFS Mount

Network For NFS Mount

Application Server Name

New Server Name Start Script

Stop Script

Application Monitor Name(s)

Application Server Name

New Server Name Start Script Stop Script Application Monitor Name(s) tsm_RG

admsrv2 admsrv1

Online On Home Node Only

Fallover To Next Priority Node In The List Fallback To Higher Priority Node In The List

[admsrv2_svc] [tsm_app]

[tsmvg] + false + fsck + sequential + false + false + false

+

imaging_app

[imaging_app]

[]

[]

[]

[/admsrv/hacmp/imaging_start_app.ksh] [/admsrv/hacmp/imaging_stop_app.ksh]

tsm_app
[tsm_app]

[/admsrv/hacmp/tsm_start_app.ksh] [/admsrv/hacmp/tsm_stop_app.ksh]

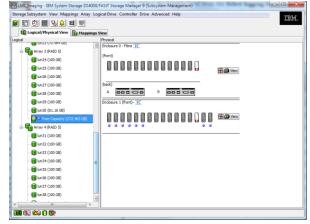
Start Hacmp service on both nodes simuteniously

#smitty hacmp -> System Management(C-SPOC) - > HACMP Service - > Start Cluster Services choose both admsrv1 & admsrv2 by pressing F7

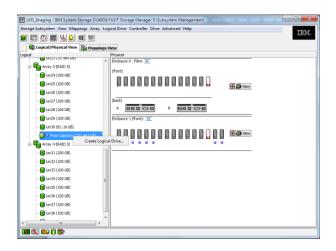
Test Resource Group takeover: #smitty hacmp - > System Management - > Resource Groups and Applications - > Move a Resource Group to another Node/Site

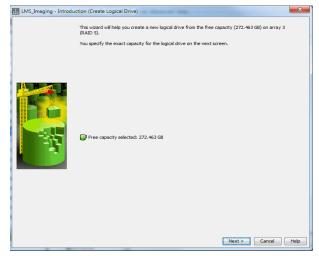
Check: ifconfig -a, find service IP move to new node, lsvg -o, find volume groups varyon on new node.

Add shared volume groups to HACMP resource group



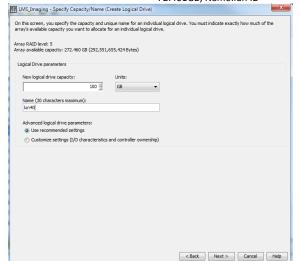
Create Logic Drive

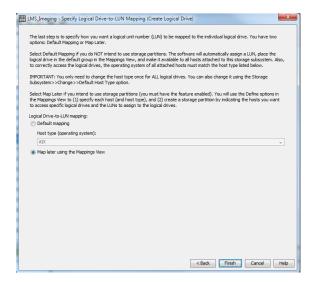


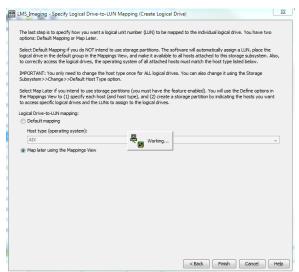


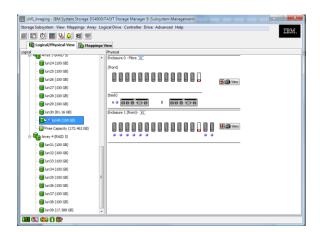
New Logic Drive capacity: 100GB, Name:lun40; 100GB, Name: lun41;

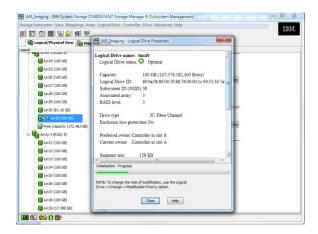
72.461GB, Name:lun42



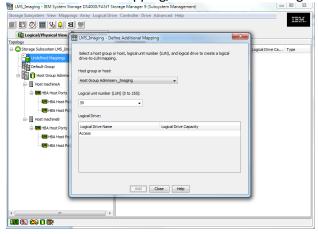








Define Additional Mapping, adds new created Logic Drives to Host Group: Adminserv_Imaging



On Admsrv1 and Admsrv2:

cfgmgr

Ispv

hdisk40	none	None
hdisk41	none	None
hdisk42	none	None

for i in 40 41 42

> do

> chdev -l hdisk\$i -a pv=yes

> done

hdisk40 changed

hdisk41 changed

hdisk42 changed

Ispv

hdisk40	00ce35aa2788a941	None
hdisk41	00ce35aa278eb6e9	None
hdisk42	00ce35aa278eb896	None

After migrated to CGI, LMS servers (admsrv1/2) are logic partitions on IBM MODEL 8205-E6D(p740), PVs are assigned to LMS from HITASH storage via 2 vscsi(2/3) visual cards connected to 2 vios servers

```
root@admsrv1:/># lspv |grep None
               00f86f6237c0f7f1
                                                   None
root@admsrv1:/># lsattr -El hdisk58
PCM PCM/friend/vscsi
algorithm fail_over
hcheck_cmd test_unit_rdy
                                                                       False
True
                                               Path Control Module
                                               Algorithm
               test_unit_rdy
                                                Health Check Command
                                                                           True
                                                Health Check Interval
hcheck interval 20
                                                                           True
hcheck mode nonactive
                                                Health Check Mode
                                                                           True
                                                Maximum TRANSFER Size
               0×40000
max transfer
                                                                          True
               00f86f6237c0f7f1000000000000000 Physical volume identifier False
pvid
queue_depth 8
                                                Queue DEPTH
                                                                           True
reserve policy no reserve
                                                Reserve Policy
root@admsrv1:/># chpath -1 hdisk58 -p vscsi2 -a priority=2
path Changed
root@admsrv1:/># chpath -1 hdisk58 -p vscsi3 -a priority=1
path Changed
root@admsrv2:/># lspv |grep None
hdisk58
              00f86f6237c0f7f1
                                                   None
root@admsrv2:/># lsattr -El hdisk58
PCM PCM/friend/vscsi algorithm fail over
                                              Path Control Module
                                                                         False
algorithm fail_over
hcheck_cmd test_unit_rdy
                                                Algorithm
                                                                           True
                                                Health Check Command
                                                                           True
                                                Health Check Interval
hcheck interval 20
                                                                          True
hcheck_mode nonactive
                                                Maximum TRANSFER Size
                                                                           True
                                                Health Check Mode
pvid 00f86f6237c0f7f100000000000000 Physical volume identifier False queue_depth 8
reserve policy no_reserve
                                                Reserve Policy
                                                                           True
root@admsrv2:/>#
root@admsrv2:/># chpath -1 hdisk58 -p vscsi2 -a priority=2
path Changed
root@admsrv2:/># chpath -1 hdisk58 -p vscsi3 -a priority=1
```