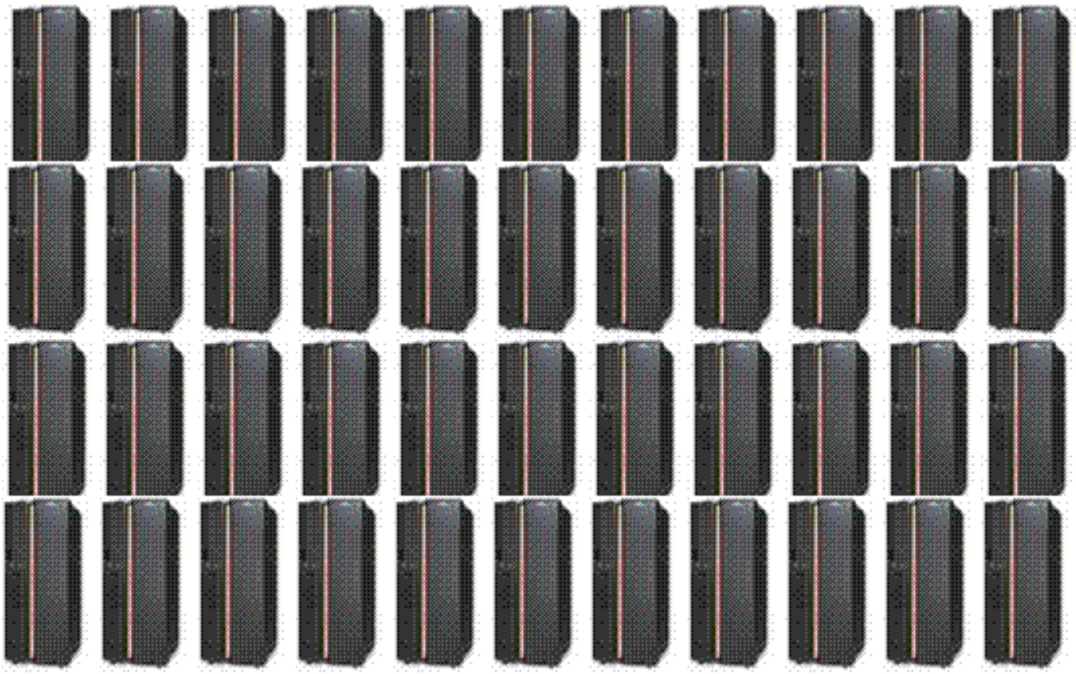
CSM Hints and Tips for Installing a Large Cluster



Version 1.0
August 2006

Contributors:
Bruce Potter
Sean Safron
Norm Nott
Linda Mellor
Connie Graff
Diana Morris
Bernie King-Smith
Scot Sakolish
Bill LePera
John Simpson
Vallard Benincosa
Al Sabawi
Josh Horton
Margaret C. Moran

IBM Corporation
Systems and Technology Group
Route 100
Somers, New York 10589

Produced in the United States of America
May 2006

This document was developed for products and/or services offered in the United States. IBM may not offer the products, features, or services discussed in this document in other countries.

The information may be subject to change without notice. Consult your local IBM business contact for information on the products, features and services available in your area.

All statements regarding IBM future directions and intent are subject to change or withdrawal without notice and represent goals and objectives only.

IBM, the IBM logo, the e-business logo, eServer, AIX 5L, Micro-Partitioning, POWER, POWER4+, POWER5+, pSeries, System p, System p5, System x, BladeCenter and Virtualization Engine are trademarks or registered trademarks of International Business Machines Corporation in the United States, or other countries, or both. A full list of U.S. trademarks owned by IBM is available at: http://www.ibm.com/legal/copytrade.shtml

Other company, product, and service names may be trademarks or service marks of others.

Linux is a registered trademark of Linus Torvalds in the United States, other countries, or both.

IBM hardware products are manufactured from new parts, or new and used parts. In some cases, the hardware product may not be new and may have been previously installed. Regardless, IBM warranty terms apply.

Photographs show engineering and design models. Changes may be incorporated in production models.

Copying or downloading the images contained in this document is expressly prohibited without the written consent of IBM.

Information concerning non-IBM products was obtained from the suppliers of these products or other public sources. Questions on the capabilities of the non-IBM products should be addressed with those suppliers.

All performance information was determined in a controlled environment. Actual results may vary. Performance information is provided “AS IS” and no warranties or guarantees are expressed or implied by IBM. Buyers should consult other sources of information, including system benchmarks, to evaluate the performance of a system they are considering buying.

When referring to storage capacity, 1TB equals total GB divided by 1000; accessible capacity may be less.

Many of the features described in this document are operating system dependent and may not be available on Linux. For more information, see:

http://www.ibm.com/servers/eserver/pseries/linux/whitepapers/linux_pseries.html.

The IBM home page is located at: http://www.ibm.com.

The IBM System p5, eServer p5 and pSeries home page is located at: http://www.ibm.com/systems/p/

Introduction

Cluster Systems Management (CSM) software provides a distributed systems management solution that allows a system administrator to set up and maintain a cluster of nodes that run the AIX® or Linux® operating systems. CSM simplifies cluster administration tasks by providing management from a single point-of-control.

This white paper is intended to be used by a System Administrator. The system administrator should:

- Be highly skilled in using AIX or Linux commands and utilities.
- Be comfortable with most basic system administration tools and processes.
- Possess a solid understanding of AIX-based or Linux-based operating systems.
- Be familiar with fundamental networking/distributed computing environment concepts.
- Be highly skilled in using CSM on AIX and/or Linux.

Some of the information in this paper could be useful in any cluster, but it is intended for a cluster of 128 nodes or greater.
For more information, see the CSM Documentation Library
http://publib.boulder.ibm.com/infocenter/clresctr/index.jsp?topic=/com.ibm.cluster.csm.doc/clusterbooks.html

This white paper is not a step-by-step procedure, but rather a collection of hints and tips on areas that may need to be changed or tuned for a large cluster. This paper assumes that an operating system and CSM have already been installed on the CSM management server.

Icon Legend – Each item in this paper is marked with one or more icons. These icons indicate which platforms are supported:

 Linux  AIX  IBM System p  IBM System x

Management Server Setup and Tuning

Useful tools available from xCSM

The xCSM package provides extra (as-is) utilities and white papers to be used with Cluster Systems Management software for Linux. The package is available for download separately

from the CSM product and can be used to extend your use of CSM. xCSM gives the CSM administrator even more capabilities to manage their clusters by providing: new commands, monitoring conditions/responses, some xCAT utilities, papers describing the use of CSM in specific environments, install customization scripts, and more.

The xCSM RPM can be installed on either Linux or AIX (it is a "noarch" RPM). It is most useful on Linux, but it has a few features that can be used on AIX as well.
Here are some commands referenced in this paper that you will find in /opt/xcsm/bin:

- **genhosts** - creates an **/etc/hosts** file for the cluster
- **helparp** – temporarily changes the ARP table for a large cluster.

Install the xCSM RPM on your management server to provide additional utilities for large clusters. The RPM, along with a description of its features, is available at:
http://www14.software.ibm.com/webapp/set2/sas/f/csm/utilities/xCSMfixhome.html

Tuning ARP (Address Resolution Protocol)

Address Resolution Protocol (ARP) is a network protocol that maps a network layer protocol address to a data link layer hardware address. For example, ARP is used to resolve an IP address to the corresponding Ethernet address.

In large networks, the ARP table can become overloaded, which can give the appearance that CSM is slow.

The following article discusses ARP "Neighbour table overflow" on Linux, and includes some suggestions for adjusting ARP related parameters:
http://www.uwsg.iu.edu/hypermail/linux/net/0307.3/0004.html

To temporarily change the parameters on Linux, run the following from the command line:

```
echo "512" >/proc/sys/net/ipv4/neigh/default/gc_thresh1
echo "2048" >/proc/sys/net/ipv4/neigh/default/gc_thresh2
echo "4096" >/proc/sys/net/ipv4/neigh/default/gc_thresh3
echo "240" >/proc/sys/net/ipv4/neigh/default/gc_stale_time
```

or

If you have installed the xCSM RPM, you can run: **/opt/xcsm/bin/helparp**

To permanently change the parameters, add the following to **/etc/sysctl.conf** and reboot the Linux server.

net.ipv4.conf.all.arp_filter = 1
net.ipv4.conf.all.rp_filter = 1
net.ipv4.neigh.default.gc_thresh1 = 512
net.ipv4.neigh.default.gc_thresh2 = 2048
net.ipv4.neigh.default.gc_thresh3 = 4096
net.ipv4.neigh.default.gc_stale_time = 240

The following tables show *arptab_nb* and *arptab_bsiz* attribute values that can be adjusted to tune ARP for AIX:

Number of Nodes	***arptab_nb* Value**
1-64	25 (system default)
65-128	64
129-256	128
257-512	256
>512	For systems larger than 512 nodes take the next higher power of 2 size, and divide by 2

Number of Interfaces	***arptab_bsiz* Value**
1-3	7 (system default)
4	8
5	10
6	12
7	14
8 or more	2 X number of interfaces

Once you have determined the appropriate values for your cluster, you can temporarily change the attributes by running the **no** command with the new values:

no –o arptab_nb=*64*
no –o arptab_bsiz=*10*

For details, see the *IBM AIX Installation Guide and Reference* for the version of AIX that you are using (AIX 5L 5.2 or 5.3).

For a permanent change, edit the **/etc/tunables/nextboot** file and add the following to the bottom of the file with the new values:

no:
    arptab_nb = "*64*"
    arptab_bsiz = "*10*"

Then run:

tunrestore –f /etc/tunables/nextboot

Tuning TCP/IP Buffer Size

In large clusters, the TCP/IP buffer size on the management server only should be increased to prevent incorrect node status from being returned. If the buffer size is too low for the number of nodes in the cluster, a node could be reported as down even though it can be reached using ping and the RMC subsystem on the node is active.

To temporarily increase the TCP/IP buffer size on Linux, run the following from the command line:
echo 524288 > /proc/sys/net/core/rmem_max
echo 262144 > /proc/sys/net/core/rmem_default

You must also recycle RMC. Run the following from the command line:
/usr/sbin/rsct/bin/rmcctrl -k
/usr/sbin/rsct/bin/rmcctrl -s

For a permanent change, add the following lines to **/etc/sysctl.conf** and reboot the Linux server:
net.core.rmem_max = 524288
net.core.rmem_default = 262144

Note: To use larger values on Linux, increase both ***rmem_max*** and ***rmem_default*** in increments of 262144. For example, increase ***rmem_max*** to 786432 and ***rmem_default*** to 524288.

To temporarily increase the TCP/IP buffer size on AIX, run the following from the command line:
no -o udp_recvspace=262144

You must also recycle RMC. Run the following from the command line:
/usr/sbin/rsct/bin/rmcctrl -k
/usr/sbin/rsct/bin/rmcctrl -s

For a permanent change on AIX 5.2, set the *pre520tune* attribute to **disable** by running the following commands:

1. **lsattr -El sys0**
2. **no -p -o udp recvspace=262144**

AIX 5.2D and later versions support storing network option changes using the **no** command and the **/etc/tunables/nextboot** file, as follows:

- The **no -r** command sets the value changes to apply after reboot.
- The **no -p** command sets the value changes immediately and to apply after reboot.

See the AIX documentation for detailed command usage information.

Note: To use larger values on AIX, increase the ***udp_recvspace*** value in increments of 262144, not to exceed the ***sb_max*** value.

no –L udp_recvspace

```
------------------------------------------------------------------------------
NAME          CUR     DEF     BOOT    MIN     MAX     UNIT     TYPE    DEPENDENCIES
------------------------------------------------------------------------------
udp_recvspace 262144  262144  262144  4K      2G-1    byte     C       sb_max
------------------------------------------------------------------------------
```

no –L sb_max

```
------------------------------------------------------------------------------
NAME         CUR    DEF    BOOT   MIN    MAX    UNIT         TYPE  DEPENDENCIES
------------------------------------------------------------------------------
sb_max       1M     1M     1M     1      2G-1   byte          D
------------------------------------------------------------------------------
```

For details, see the *CSM for AIX 5L and Linux Administration Guide*, at:
http://publib.boulder.ibm.com/infocenter/clresctr/vxrx/index.jsp?topic=/com.ibm.cluster.csm.doc/csm15/am7ad12018.html%23wq612

Tuning AIX Network Attributes

Tuning network attributes can impact cluster performance. CSM provides a sample that can be used to start tuning network switches. The instructions are provided in **/opt/csm/samples/docs/tunables.README**, which is installed with **csm.server**.

Command Line Length

For CSM commands to work on large AIX clusters, the ***ncargs*** system attribute must be set to a high value. The ***ncargs*** attribute controls the maximum command line length, including environment variables, on AIX nodes. The ***ncargs*** AIX system default is four 4K blocks; increasing this value to 16 4K blocks ensures that all CSM commands will work for 1024

nodes with fully-qualified domain names. For 2048 nodes, the ***ncargs*** value must be set to 32 4K blocks.

To return an AIX node's ***ncargs*** attribute value, enter:
lsattr -EH -l sys0 | grep ncargs

To change a node's ***ncargs*** attribute value, enter:
chdev -l sys0 -a ncargs=16

No reboot or refresh of daemons is required.

An alternative to setting the ***ncargs*** attribute value is to use the **xargs** command, which allows commands to exceed the command line character limit.

Tuning NFS (Network File System)

If you are planning to concurrently install more than eight Linux nodes using the **installnode** command, tune NFS before running the command, as follows:

1. Increase the NFSD count:
   - For SUSE Linux Enterprise Server, edit **/etc/sysconfig/nfs** and set **USE_KERNEL_NFSD_NUMBER** to a higher value than the default of 8. Based on the number of nodes you are installing, consider setting the value to 16 or 32.

   - For Red Hat EL, update **/etc/init.d/nfs** and change **RPCNFSDCOUNT** to a higher value than the default value of 8.

(The following two steps apply to both SUSE Linux Enterprise Server and Red Hat EL.)

2. Restart the **nfs** server to ensure that the values take effect.

3. Use multiple install servers to increase the number of concurrent CSM and Linux operating system installations.

You can use HTTP instead of NFS to serve files during the installation of the Red Hat or SUSE Linux Enterprise Server operating system. For more information, see the Install Server section in this document.

To list the current NFS values on AIX, run:
nfso –L

The default value of **3891** for *nfs_max_threads* should be sufficient.

Tuning NIM (AIX Network Installation Manager)

NIM enables a cluster administrator to centrally manage the installation and configuration of AIX and optional software on machines within a network environment.

If you are installing a large number of AIX nodes, you can adjust the following NIM settings to improve NIM scalability and performance:

1. Enable the multithread option on the NIM **nimesis** daemon using the *max_nimesis_threads* attribute. Setting this attribute value improves NIM performance when installing a large number of nodes. Specify a value between 20 and 150 that is approximately one half of the number of nodes you are installing concurrently. For example, if you are installing 100 nodes, set the *max_nimesis_threads* value to **50**.

Run the following command on the NIM master to set the value:
nim -o change -a max_nimesis_threads=50 master

2. Set the ***global_export*** attribute to **yes** on the management server (NIM master). Always set this attribute when you are simultaneously running NIM operations on many nodes.

Run the following command on the NIM master to set the value:
nim -o change -a global_export=yes master

3. Check your ***arptab*** values and change them, if necessary. See the Tuning ARP section in this document.

For details, see the *IBM AIX Installation Guide and Reference* for the version of AIX that you are using (AIX 5L 5.2 or 5.3).

Note: If your NIM master and dhcp server are on the same server, you must restrict the number of concurrent AIX node installations to 32. To install a larger number of AIX nodes concurrently, set up your NIM master on a separate CSM install server. See the *CSM for AIX 5L and Linux Planning and Installation Guide* for details on using install servers, at:

http://publib.boulder.ibm.com/infocenter/clresctr/vxrx/index.jsp?topic=/com.ibm.cluster.csm.doc/csm15/am7il12012.html

Using a mksysb image to install AIX nodes

A **mksysb** image is a system backup image created by the AIX **mksysb** command. You can use this image to install other AIX servers, or to restore the AIX server that was the source of the **mksysb**.

NIM supports the use of **mksysb** images. A NIM **mksysb** installation is faster than a NIM rte installation, and **mksysb** allows you to install additional software. You can also use a **mksysb** image to install CSM on your AIX nodes.

For a complete description of AIX **mksysb** support, see the *AIX Installation Guide and Reference*. The *CSM for AIX 5L and Linux Planning and Installation Guide* also describes a **mksysb** scenario, at:
http://publib.boulder.ibm.com/infocenter/clresctr/vxrx/index.jsp?topic=/com.ibm.cluster.csm.doc/csm15/am7il12033.html%23mksysb

Tuning Fanout Values for CSM Commands

Fanout values define the number of nodes to concurrently perform various operations on. There are a number of ways to set the fanouts for CSM commands. Many CSM fanouts can be set using environment variables in the root shell profile.

Default fanout values for the parallel CSM commands are listed in the following table:

Command	**CSM_FANOUT Default**	**CSM_FANOUT_DELAY Default**	**DSH_FANOUT Default**
installnode	16	1200 Seconds	
updatenode	32		
dsh/dcp			64
cmfupdatenode	32		
smsupdatenode	32		

In large clusters, use the following fanout values to run parallel CSM commands:

installnode

The **installnode** command uses the ***CSM_FANOUT*** and ***CSM_FANOUT_DELAY*** environment variables to control how many Linux nodes are rebooted, and thus installed, concurrently. ***CSM_FANOUT*** sets the maximum number of concurrent reboots. If this variable is not set, 16 nodes are rebooted concurrently. If it is set to 0, all nodes are rebooted concurrently. ***CSM_FANOUT_DELAY*** sets the delay in seconds between rebooting groups of nodes. If this variable is not set, the delay is 1200 seconds (20 minutes).

To increase the number of nodes you can install concurrently, use install servers; see the *CSM for AIX 5L and Linux Planning and Installation Guide,* which is available at http://publib.boulder.ibm.com/infocenter/clresctr/vxrx/index.jsp?topic=/com.ibm.cluster.csm.doc/csm15/am7il12012.html

If you are using Linux install servers, you can increase the ***CSM_FANOUT*** variable by approximately 16 nodes per Linux install server. For example, if you have 10 Linux install servers, you can install 160 nodes concurrently.

If your nodes take less than 20 minutes to install, you can also decrease the ***CSM_FANOUT_DELAY*** to the amount of time it takes one Linux node to install.

If you are experiencing poor **NFS** performance, node installation is slow, or you are getting install timeouts, set ***CSM_FANOUT*** to **4** and ***CSMFANOUT_DELAY*** to **2400**.

updatenode


The number of AIX or Linux servers updated in parallel by the **updatenode** command is controlled by the ***CSM_FANOUT*** environment variable. If this variable is not set, **updatenode** will run up to 32 nodes in parallel.

If you are using install servers, the same ***CSM_FANOUT*** value can be used for both the **updatenode** and **installnode** commands. For example, a ***CSM_FANOUT*** value of 160 is appropriate for 10 Linux install servers. Because the **cfmupdatenode** command is called by **updatenode**, and should not be used with a high fanout, you must specify a lower ***CSM_FANOUT*** for **updatenode** by running the command with ***--cfmoptions -M 32***, where 32 is a sample ***CSM_FANOUT***.

dsh and dcp


The **dsh** and **dcp** commands use the same fanout setting. You can specify the fanout with the ***DSH_FANOUT*** environment variable, or on the command line with the **-f** flag. If neither is set, the default value is 64.

The **dsh** and **dcp** fanout is only restrained by the number of remote shell commands that can be run in parallel. Experiment with the ***DSH_FANOUT*** on your **management server** to see if higher values are appropriate.

The **dshbak** command formats output from the **dsh** command. For more information, see the *CSM for AIX 5L and Linux Command and Technical Reference*, at:

http://publib.boulder.ibm.com/infocenter/clresctr/vxrx/index.jsp?topic=/com.ibm.cluster.csm.doc/csm15/am7cm12054.html

cfmupdatenode

The **cfmupdatenode** command uses the ***CSM_FANOUT*** environment variable or the -M flag to control fanout. If neither of these is set, the default fanout is 32.

CFM does not utilize install servers, so you should not increase its fanout to accommodate them.

If you are setting the ***CSM_FANOUT*** in a shell profile, you can always call **cfmupdatenode** with the -M flag to ensure that it does not inherit high settings designed for **updatenode** or **installnode**.

smsupdatenode

The number of Linux nodes updated in parallel by the **smsupdatenode** command is controlled by the ***CSM_FANOUT*** environment variable. If this variable is not set, **smsupdatenode** will run to 32 nodes in parallel by default.

SMS files reside on install servers, so if you are using install servers, the same ***CSM_FANOUT*** value can be used by **smsupdatenode**, **updatenode** and **installnode**. For example, a ***CSM_FANOUT*** value of 160 is appropriate for 10 Linux install servers.

Tuning RSCT Heartbeat Attributes

The *Status* attribute value for Managed nodes is defined by the underlying cluster infrastructure (RSCT). In a large cluster environment, the heartbeat mechanism can be tuned for efficiency based on the size of the cluster, network configurations for bandwidth, performance, and traffic, and administrator preference.

To see the current values for ***HeartbeatFrequency*** and ***HeartbeatSensitivity***, run **csmconfig** on your management server:

- ***HeartbeatFrequency*** is the number of seconds between heartbeat messages sent to the nodes. The default value is 12.
- ***HeartbeatSensitivity*** is the number of missed heartbeat messages sent to a node to declare the node unreachable (its *Status* would then be **0**). The default value is 8.

To reduce network traffic, increase the ***HeartbeatFrequency*** value (ie making the Heartbeat Less Frequent) and decrease the ***HeartbeatSensitivity*** value. Use the **csmconfig** command to change these values. For details, see the CSM *for AIX 5L and Linux Command and Technical Reference*, at:
http://publib.boulder.ibm.com/infocenter/clresctr/vxrx/index.jsp?topic=/com.ibm.cluster.csm.doc/csm15/am7cm12034.html

Tuning CSM Hardware Control


If your network appears slow, or a hardware control point is having problems and is slow, or you are getting any of the following error messages, you may need to increase hardware control timeout values.

Before increasing any timeout values, check your network settings and your hardware control attributes. Run **rpower query** to a single node, and then to multiple nodes to see if errors are returned.

- 2651-675 Error connecting to hardware control point
  Increase the ***HC_CONNECT_TIMEOUT*** (Java daemon to hardware control point) value by 60 seconds at a time. The default value is 60 seconds.

- 2651-627 Daemon socket read/write timeout or 2651-653 Node not found
  Increase the ***HC_SOCKET_TIMEOUT*** (Library to Java Daemon)
  value by 60 seconds at a time. The default value is 120 seconds.

- 2651-670 Error connecting to hardware control daemon (for BMC)
  Increase the ***HC_FILE_DESC_MAX*** (number of open sockets between library and Java daemon) value by 500 at a time. The default value is 4096. The **ulimit –a** command returns the cluster limit.

- 2651-670 Error connecting to hardware control daemon
  You may be out of memory; run the following command to enable tracing:
  startsrc -s IBM.HWCTRLRM -eHC_JAVA_VERBOSE=/tmp/java.txt
  Check the **/tmp/java.txt** file for the following text: "Signaling in VM: OutOfMemoryError" - If this message is present, continue as follows:

  HC_JAVA_HEAP - defaults to 256MB for Java 1.3x and 900MB for Java 1.4x. The maximum heap is configurable through the environment variable ***HC_JAVA_HEAP***=n, where n is the size in MB. Memory allocation problems occur routinely in the JVM, which consumes memory without performing any garbage collection until an allocation request fails. At that point, the JVM issues an Allocation Failure (AF) message such as:

<AF[390]: Allocation Failure. need 65552 bytes, 1800523 ms since last AF>
<AF[390]: managing allocation failure, action=2 (158540704/896334336)>

NOTES:

- Once memory has been allocated on the Java heap, it is not returned to the O/S. Therefore, keep the maximum heap (–Xmx) as low as possible (in the low-20% free-space range) to force the garbage collector to manage the heap by running and compacting more frequently.

- Increasing the maximum heap beyond the minimum can have the unintended affect of allowing the garbage collector to consume memory from the free-space rather than reallocating reclaimed memory. Garbage collector activities are logged to **/tmp/gc.txt**, for example, if IBM.HWCTRLRM were started with **HC_JAVA_VERBOSE=/tmp/gc.txt**.

  For example, to increase the maximum heap, enter:
  startsrc -s IBM.HWCTRLRM –e "HC_JAVA_VERBOSE=/tmp/gc.txt HC_JAVA_HEAP=450"

Planning for Cluster-Ready Hardware Server

Cluster Ready Hardware Server (CRHS) is a software set that enhances control over HMC-attached System p™ nodes (System p, System p5™, and OpenPower™ servers). CRHS enables access to multiple HMCs and other functions that simplify communication to the hardware service network. These include:

- Hardware discovery. See the information on Hardware discovery in the *CSM for AIX 5L and Linux Planning and Installation Guide*, at:
http://publib.boulder.ibm.com/infocenter/clresctr/vxrx/index.jsp?topic=/com.ibm.cluster.csm.doc/csm15/am7il12006.html%23hwdisco
- A central database (repository) for cluster hardware information. See the information on Shared repository in the *CSM for AIX 5L and Linux Planning and Installation Guide,* at:
http://publib.boulder.ibm.com/infocenter/clresctr/vxrx/index.jsp?topic=/com.ibm.cluster.csm.doc/csm15/am7il12006.html%23sharedrep
- An updated hardware server daemon. See the information on Hardware server daemon in the *CSM for AIX 5L and Linux Planning and Installation Guide*, at:
http://publib.boulder.ibm.com/infocenter/clresctr/vxrx/index.jsp?topic=/com.ibm.cluster.csm.doc/csm15/am7il12006.html%23hwservd
- Installation and password setup for the managed servers. See the information on Password management in the *CSM for AIX 5L and Linux Planning and Installation Guide*, at:
http://publib.boulder.ibm.com/infocenter/clresctr/vxrx/index.jsp?topic=/com.ibm.cluster.csm.doc/csm15/am7il12006.html%23pwdman

When should you use Cluster Ready Hardware Server (CRHS)?

CRHS is required if your cluster includes an IBM High Performance Switch (HPS). You can also use CRHS in environments that do not have an HPS. CRHS provides the following advantages:

- Consolidation of the service networks into a maximum of two subnets for redundancy.
- Reduced HMC requirements for recovery.
- Automated discovery of System p servers and HMCs, initialization of hardware registration, and updates when hardware configuration changes.
- Automated association of System p 575, 590, 595 servers with their frames.
- Ease of movement of System p servers from one HMC to another HMC.
- Shared repository of System p cluster hardware information.
- Reduced number of DHCP servers.

Limitations:
CRHS is supported on System p and OpenPower servers, and on HMCs running the GA5 service level. Other servers, System p servers with firmware prior to GA5 service level, and HMCs with a service level earlier than GA5 are not supported. Configurations with POWER4™ servers must keep these servers on an HMC that is separate from the CRHS configuration. POWER4 servers can still be managed by the CSM management server, but are not included in the enhanced CRHS support that is provided for System p servers.

Flexible Service Processor (FSP) hardware control power method

The Flexible Service Processor (FSP) power method allows direct hardware control of POWER5 nodes that are not attached to or controlled by a Hardware Management Console (HMC).

The following functions still require an HMC and become more critical as your cluster increases in size:

- Service Focal Point
- LPAR creation
- Dynamic LPAR
- Virtual I/O

FSP support can be used on POWER CSM management servers, but it cannot be used on x-86 based management servers, and is not intended to be used in conjunction with Cluster Ready Hardware Server (CRHS). This function requires a firmware level of GA6 on the System p nodes.

Some of these functions become more critical as a cluster increases in size. You must determine whether any of these functions are needed and to order HMCs for the cluster to provide these capabilities.

Remote Hardware Inventory and Maintenance

CSM provides commands that perform hardware inventory scans and maintenance activities on cluster nodes from the CSM management server. Performance of these commands in a large cluster environment depends upon other CSM components that have been discussed in this white paper.

Depending on the target machine types and the desired task, CSM can invoke the **dsh** or **dcp** commands, or initiate a network boot to perform that task.

The **rfwcfg** command updates CMOS settings on x86-based servers and blade servers. If the target system has an operating system installed and is managed by CSM, the **rfwcfg** command invokes the **dsh** and **dcp** commands to update the settings. See the scaling recommendations for the **dsh** and **dcp** commands contained in this document for tips on improving large cluster performance. If the target system has no operating system installed, the **rfwcfg** command initiates a network boot from the install server. The settings are updated within the network boot environment. See the scaling recommendations in the "Defining Nodes and Node Groups", "Install Servers", and "Setting up for Node Installation" sections in this document for tips on improving large cluster performance. To perform a network boot, the node boot order must be set with the Network boot device listed before the Hard Disk.

The **rfwflash** command updates system BIOS on x86-based servers and blade servers, and microcode on BladeCenter® (JS20/JS21 blade servers) and HMC-attached POWER5™ (and later) servers.

For x86-based servers and blade servers, if the target system has an operating system installed and is managed by CSM, the **rfwflash** command invokes the **dsh** and **dcp** commands to update the BIOS. See the scaling recommendations for the **dsh** and **dcp** commands contained in this document for tips on improving large cluster performance. If the target system has no operating system installed, the **rfwflash** command initiates a network boot from the install server. The BIOS is updated within the network boot environment. See the scaling recommendations listed in the "Defining Nodes and Node Groups", "Install Servers", and "Setting up for Node Installation" sections in this document for tips on improving large cluster performance. To perform a network boot, the node boot order must be set with the Network boot device listed before the Hard Disk.

BladeCenter JS20 and JS21 blade servers must have an operating system installed and be managed by CSM to update the microcode. The **rfwflash** command invokes the **dsh** and **dcp** commands to update the microcode. See the scaling recommendations for the **dsh** and **dcp** commands contained in this document for tips on improving large cluster performance.

For HMC-attached System p (and later) servers, the HMC must be defined in CSM as a managed device to update the microcode for its attached POWER5 (and later) servers. It is not necessary for the target system to have an operating system installed or be managed by CSM. The **rfwflash** command invokes the **dsh** and **dcp** commands to update the microcode. See the scaling recommendations for the **dsh** and **dcp** commands contained in this document for tips on improving large cluster performance.

The **rfwscan** command gathers BIOS levels and selected vital product data for x86-based servers and blade servers, and microcode levels for HMC-attached System p (and later) servers.

For x86-based servers and blade servers, if the target system has an operating system installed and is managed by CSM, the **rfwscan** command invokes the **dsh** and **dcp** commands to gather the information. See the scaling recommendations for the **dsh** and **dcp** commands contained in this document for tips on improving large cluster performance. If the target system has no operating system installed, the **rfwscan** command initiates a network boot from the install server. The information is gathered within the network boot environment. See the scaling recommendations listed in the "Defining Nodes and Node Groups", "Install Servers", and "Setting up for Node Installation" sections contained in this document for tips on improving large cluster performance. To perform a network boot, the node boot order must be set with the Network boot device listed before the Hard Disk.

For HMC-attached System p (and later) servers, the HMC must be defined in CSM as a managed device to collect the microcode levels for its attached POWER5 (and later) servers. It is not necessary for the target system to have an operating system installed or be managed by CSM. The **rfwscan** command invokes the **dsh** and **dcp** commands to collect the microcode levels. See the scaling recommendations for the **dsh** and **dcp** commands contained in this document for tips on improving large cluster performance.

For more information on the CSM Remote Hardware Inventory and Maintenance features, see the *CSM AIX 5L and Linux Administration Guide*.
http://publib.boulder.ibm.com/infocenter/clresctr/vxrx/index.jsp?topic=/com.ibm.cluster.csm.doc/csm15/am7ad12009.html

Service Processor naming and configuration

In large clusters, many different naming conventions can be used. For example, when naming and configuring service processors:

- BladeCenter - Use slot number as part of the *Hostname* value.
- System p - Set LPAR NAME to the short *Hostname* value.
- BMCs - define the node with the *HWControlNodeId* value.

See the Defining a Large Number of Cluster Nodes section in this document.

Defining Nodes and Node Groups (and preparing to define nodes)

Automatically adding hostnames to /etc/hosts

Set up hostname resolution before defining nodes, using either DNS or **/etc/hosts**. For example, using **/etc/hosts** with the xCSM RPM installed on your management server:

1. Modify the sample **/opt/xcsm/bin/genhosts** file to match your network.
2. Run **/opt/xcsm/bin/genhosts** to generate the entries.
3. Paste those entries into your **/etc/hosts** file.

The /etc/hosts format is the same on both AIX and Linux.

Networking and Node Naming Conventions

The following examples show large cluster node naming Conventions examples

Example 1:
A cluster consists of 9 frames. Each frame contains 6 BladeCenter chassis, and each BladeCenter chassis contains 14 blade servers. The naming convention is *z*[*frame #*]c[*1-6*]s[*1-14*], where *z* is the first letter of the cluster name. Make sure the frames and chassis themselves are labeled - you can determine a single blade or BladeCenter chassis location based on the name.
For example, **z9c4s11** is located in frame 9, chassis 4, slot 11.

Example 2:

For HMC-attached nodes:

- For a CEC, use the Cluster Name or Part of the Cluster Name, Frame Number, and part of the CEC Serial number.
  For example: c123f1cec456
- For a Partition, use the Cluster Name or Part of the Cluster Name, Frame Number, and Partition number.
  For example: c123f1p1

Defining a Large Number of Cluster Nodes

Depending on the type of hardware control being used in your cluster, there are different strategies for defining all of your cluster nodes at once.

To define your nodes using hostname mapping files, set the node short host name to the identifier associated with the blade server or LPAR, for example. Then use the

lshwinfo -s command to create your map files.

The following example configurations require different methods for defining nodes:

Case 1: Each hardware control point manages a group of nodes. This is the case with an HMC or with a BladeCenter management module.

Case 2: Each hardware control point manages a single node. This is the case with a BMC or with stand-alone FSP nodes (not controlled by an HMC).

Case 1: Defining nodes in a cluster with each hardware control point controlling a group of nodes

Run the **lshwinfo** command for each hardware control point to gather information about the nodes it controls. Then, this information can be passed to the **definenode** command.

For HMC-attached nodes (***PowerMethod*=hmc**) and BladeCenter nodes (***PowerMethod*=blade**), the **definenode** command automatically determines the console server information based on the hardware control information.

Step 1: Run the **lshwinf**o command to collect node information from one or more hardware control points:

lshwinfo -p <*PowerMethod*> -c <*ipaddr-list*> -s -o /tmp/nodeinfo

Where:
-p specifies the power method (**hmc**, **blade**, **xseries**)
-c specifies a list of hardware control point IP addresses
-o specifies an output file name
-s Sets the hostname field with the *HWControlNodeId* value, if resolvable.

For example:

lshwinfo -p hmc -c c209hmc -s -o /tmp/nodeinfo

HMC output is similar to:

Hostname::PowerMethod::HWControlPoint::HWControlNodeId::LParID::HWType::HWModel::HWSerialNum::DeviceType
(Make sure these are fake node names/IPs, etc.: security)
c209f1n01::hmc::c209hmc.ppd.pok.ibm.com::c209f1n01::002::7040::681::4332175::::
c209f1n02::hmc::c209hmc.ppd.pok.ibm.com::c209f1n02::001::7040::681::4332175::::
c209f1n03::hmc::c209hmc.ppd.pok.ibm.com::c209f1n03::003::7040::681::4332175::::
c209f1n04::hmc::c209hmc.ppd.pok.ibm.com::c209f1n04::004::7040::681::4332175::::

Step 2: After running **lshwinfo**, enter your node hostnames in the **/tmp/nodeinfo** file. Using the –s flag in Step 1 sets the hostname field to the *HWControlNodeId* value, if resolvable.

Step 3: Run the **definenode** command to define the nodes:
definenode -M /tmp/nodeinfo

After defining the nodes, back up your node definitions using the **lsnode -F > /store/nodedef** command. You can also use the **csmbackup** command to back up all management server information.

Case 2: Defining nodes in a cluster with each hardware control point controlling a single node

For BMC nodes and stand-alone FSP nodes, there is one hardware control point for each node. The **lshwinfo** command is not practical for this environment, because it requires running the command once for each node. Instead, set up your IP addresses sequentially, as follows:

For BMC nodes:

Step 1: Initially, define all nodes with the default attribute values and no hardware control or console information. For example, enter:

definenode -n n0001-n0117

Step 2: Assign the hardware control and console information to one frame at a time. In the following example, **F34** is a node group containing all nodes in frame 34. The example assumes that all nodes in the frame are attached to a single console server, **ts34**. It also assumes that the hardware control points (BMCs) are assigned sequential IP addresses, starting at 172.29.123.1. For each frame, run a command similar to:

P=1; for i in $(lsnode -N F34); do chnode -n $i ConsolePortNum=$P ConsoleServerName=ts34 ConsoleMethod=mrv PowerMethod=bmc HWControlNodeId=$i HWControlPoint=172.29.123.$P; $P=$(echo "$P + 1" | bc - 1); done

After defining the nodes, back up the node definitions using the **lsnode -F > /store/nodedef** command. You can also use the **csmbackup** command to back up the management server information.

For FSP nodes:

Step 1: Run the **hwsda** command**:**

hwsda –f –o /tmp/nodeinfo

The **–f** flag lists hardware information for each SP discovered on the network, in **lshwinfo** format.

Step 2: After running **hwsda**, enter each node hostname in the **/tmp/nodeinfo** file.

Step 3: Run the **definenode** command to define the nodes:

definenode -M /tmp/nodeinfo

After defining the nodes, back up the node definitions using the **lsnode -F > /store/nodedef** command. You can also use the **csmbackup** command to back up the management server information.

Useful node groups

Creating the following node groups can facilitate cluster management:

F1,F2,…F*n*
A node group based on Frame numbers. For example, if your Frame has 39 nodes, create an **F1** node group that includes all nodes:

nodegrp –n n0001-n0039 F1

Install Server Nodes
If the management server is the install server for ALL separate install servers, and no other nodes use the management server as an install server:

nodegrp –w "InstallServer =''" InstallServers

For example, to create a node group of all nodes installed by install servers, in each frame where the first node is the install server:

nodegrp –w "Hostname like'%s1'" InstallServersS1

NoIsvrF1,NoIsvrF2,…NoIsvrF*n*

For example, to create a node group of all nodes in a frame except for the install servers:

nodegrp –a zf1s2+37 NoIsvrF1 (+37 is the number of nodes in the frame -2)

To create a node group of install servers in all cluster frames:

nodegrp –w "InstallServer = 'Name Server'" InstallServer

To create a node group of all cluster nodes except for install servers:

nodegrp -a 'nodegrp -d "," -S AllNodes InstallServers' NoIsvrgrp

You can also create dynamic node groups for each *HWType* and *HWModel*, to track system information and check configuration by hardware type.

The following example is for a cluster of 9 frames, with each frame containing 6 BladeCenter chassis, and each chassis containing 14 blade servers, using the naming convention *z*[*frame #*]c[*1-6*]s[*1-14]* (*z* is the first letter of the cluster name). For example, z9c4s11 is located in frame 9, chassis 4, slot 11.

You could create the following dynamic node groups:

- **nodegrp -w "Hostname like 'z9c%'" z9** - includes all blade servers in frame 9.
- **nodegrp -w "Hostname like 'z9c4s%'" z9c4** - includes all blade servers in frame 9, BladeCenter chassis 4.
- **nodegrp -w "Hostname like '%s7.cluster.com'" s7** – includes all blade servers in slot 7, which includes all frames and all blade servers.

Improved scaling of MAC address collection

For System x™ you can collect either the UUID or the MAC Address. It is recommended that you use UUID.

For example, collect the MAC Address if multiple network adapters on a node are connected to subnets shared with the management server, because you might want to use a specific adapter or subnet to do the installation. Because the UUID is common to the server, you cannot use it to specify a specific install adapter.

For System p Linux and AIX you must always collect the MAC Address.

The **csmsetupinstall**, **csmsetupyast,** or c**smsetupks** command first attempts to collect the node UUID, which requires minimal time. If that fails, the command attempts to collect the MAC address. Collecting the MAC address on System x nodes requires rebooting the nodes and using a serial terminal to read the MAC address from the display. MAC address collection is more time-consuming and more error-prone than UUID collection.

To collect MAC addresses or UUIDs on SLES using the **InstallServers** node group, enter:

csmsetupyast -x –N InstallServers

To collect MAC addresses or UUIDs on AIX using the **InstallServer** node group:

1. Run **getadapters -D -t ent -s 100 -d full –N InstallServer -z mystanzafile**
2. Verify that the **getadapters** command returns the information you expect; make any required changes manually to the stanza file.
3. Run **getadapters –w –f mystanzafile**

Configuring cluster network adapter interfaces

CSM supports secondary network adapter configuration during a node installation or update using the **updatenode -c** command. Secondary adapters are additional adapters not used for remote network node installation. On AIX nodes, CSM supports Ethernet, HPS (switch), and Multilink adapter configuration. On Linux nodes, CSM supports Ethernet adapter configuration.

Support for HPS and Multilink adapters on AIX requires the nodes to be installed with a minimum operating system level of AIX 5L V5.3 with Recommended Maintenance Package 5300-03, or AIX 5L V5.2 with Recommended Maintenance Package 5200-07.

CRHS is required if your cluster uses an IBM High Performance Switch (HPS).

For more information, See the *CSM AIX 5L and Linux Planning and Installation Guide* at http://publib.boulder.ibm.com/infocenter/clresctr/index.jsp?topic=/com.ibm.cluster.csm.doc/csm15/am7il12020.html

Install Servers

CSM install servers can be used to as a file server for node updates, to manage network boot requests, and to drive operating system installations. For Linux network installations, an install server responds to PXE (System x) and **bootp** (System p) requests. For AIX network installations, an install server is the AIX Network Installation Manager (NIM) master that responds to **bootp** requests. CSM supports multiple install servers, which can be managed by a Linux or an AIX management server.

Install servers provide several advantages for cluster administrators:

- Creating separate install servers for each operating system provides full node installation capability from a single management server, regardless of the management server operating system.
- Multiple install servers improve installation scaling and performance.
- Install servers at remote locations facilitate full installation and improve installation performance.
- Distributing files to install servers can improve performance when updating nodes with the **updatenode** command.

The CSM management server or selected cluster nodes can be used as install servers. The type of install servers required depends upon your particular cluster environment. For example, if your cluster contains AIX nodes only, your AIX management server can be used as the install server for the entire cluster. If your cluster includes a mix of AIX and Linux nodes, then at least one install server is required for each operating system.

When using an install server that is not also your management server, the install server must be defined, installed, and added to your cluster as a Managed node before using it to define or install nodes.

An install server can be used to install different operating system levels but not different operating system distributions. A separate install server is required for each operating system distribution. For example, a Red Hat EL install server is required to install Red Hat EL nodes, and a SLES install server is required to install SLES nodes.

Using NFS or HTTP for node installation

CSM supports NFS and HTTP for installing the Linux operating system on your cluster nodes. By default, CSM uses NFS to do node installations.

For more information on how to decide see the *CSM AIX 5L and Linux Planning and Installation Guide*, Section – Decide whether to use NFS or HTTP for Installs, at:

http://publib.boulder.ibm.com/infocenter/clresctr/vxrx/index.jsp?topic=/com.ibm.cluster.csm.doc/csm15/am7il12011.html%23dwnfshttp

If you choose HTTP for node installation, CSM provides automatic setup of the Apache 2 HTTP server on the install server by running the **csmconfig** command with attribute value pairs.

Attributes for the Network Install Protocol: Two **csmconfig** attributes dictate how CSM will set up your install server for network installations:

- *NetworkInstallProtocol*
- *SetupNetworkInstallProtocol*

For details on configuring Apache for improved scaling, see the *CSM AIX 5L and Linux Planning and Installation Guide*, Section - Setting up an HTTP server manually.

http://publib.boulder.ibm.com/infocenter/clresctr/index.jsp?topic=/com.ibm.cluster.csm.doc/csm15/am7il12041.html

.

Additional Node Configuration

Redirecting system logs to the management server via syslog

To redirect all system logs to a Linux management server, configure the management server and create a post-install script, as follows:

Configure the Linux management server:

1. Edit **/etc/sysconfig/syslog** by adding the –r flag to accept remote logging: **SYSLOGD_OPTIONS="-r “**
2. Restart the syslog

Create a post-install script on each Managed node:

1. Create a file: **/csminstall/csm/scripts/installprereboot/0001CSM_syslog**

The syslog file contents should be similar to:

```
#!/bin/ksh

mv -f /etc/syslog.conf /etc/syslog.conf.ORIG
echo "*.*       $MGMTSVR_IP" >/etc/syslog.conf

case $DISTRO_NAME in
      SLES*)
            if grep 'SYSLOGD_PARAMS="-m0' /etc/sysconfig/syslog >/dev/null 2>&1
            then
                  :
```

```
            else
                  perl -pi -e 's/SYSLOGD_PARAMS="/SYSLOGD_PARAMS="-m0 /'
/etc/sysconfig/syslog
            fi
            /etc/init.d/syslog restart
            ;;
      RedHat*)
            /etc/rc.d/init.d/syslog start
            ;;
esac
exit 0
```

Enabling node-to-node ssh in a Cluster

The following setup is needed for GPFS and MPI among other things.

1. If root's **$HOME/.ssh** directory (typically **/root/.ssh** on Linux, and **/.ssh** on AIX) does not exist on the management server, create the directory:

mkdir -m700 $HOME/.ssh

2. Run the following command on the management server to create a symbolic link:
Sym link **$HOME/.ssh** into **/cfmroot/$HOME/.ssh**.

For example, on Linux:

```
mkdir /cfmroot/root
ln -s /root/.ssh /cfmroot/root/.ssh
```

For example, on AIX:

```
ln -s /.ssh /cfmroot/.ssh
```

To verify that ssh is setup correctly, install your nodes and **ss**h to one of them. Then make sure that node can ssh to the remaining installed nodes without being prompted for a password (or to enter "yes").

Network time synchronization

Though not required for CSM, there are important advantages to running a network time synchronization system in a CSM cluster; if you are not running a time service you are encouraged to install one. Some optional CSM supported subsystems require host clock synchronization. KRB5 for example will not run if the host clocks are more than 5 minutes out of sync. NFS can generate errors if the hosts on the network are not time synchronized.

See the **NTP** setup information in **/opt/csm/samples/docs/ntp.README.**

To link the **/etc hosts**, **passwd, shadow**, and **group** files in the **/cfmroot/etc/** directory for all nodes, see the **CFM** information, in the *CSM AIX 5L and Linux Administration Guide*
http://publib.boulder.ibm.com/infocenter/clresctr/vxrx/index.jsp?topic=/com.ibm.cluster.csm.doc/csm15/am7ad12004.html%23cfengine

CSM product information

The CSM library, including the latest documentation updates. See
http://publib.boulder.ibm.com/infocenter/clresctr/vxrx/index.jsp.
.
The CSM errata file for the latest documentation updates. See
http://publib.boulder.ibm.com/infocenter/clresctr/vxrx/index.jsp?topic=/com.ibm.cluster.csm.doc/clusterbooks.html.

The CSM support web site. See
http://www14.software.ibm.com/webapp/set2/sas/f/csm/download/home.html.

The CSM FAQ (Frequently Asked Questions). See
http://www.ibm.com/developerworks/forums/dw_thread.jsp?forum=907&thread=128386&cat=53&treeDisplayType=threadmode1

The xCSM Utilities Software. See
http://www14.software.ibm.com/webapp/set2/sas/f/csm/utilities/xCSMfixhome.html