Release Notes

Scyld ClusterWare Release 5.11.2-5112g0000

About This Release

Scyld ClusterWare Release 5.11.2-5112g0000 (released January 20, 2015) is the latest update to Scyld ClusterWare 5.

Scyld ClusterWare 5.11.2 expects to execute in a Red Hat RHEL5 Update 11 or CentOS 5.11 base distribution environment, each having been updated to the latest RHEL5/CentOS5 errata (https://rhn.redhat.com/errata/rhel-server-errata.html) as of the Scyld ClusterWare 5.11.2 release date. Any compatibility issues between Scyld ClusterWare 5.11.2 and RHEL5 are documented on the Penguin Computing Support Portal at http://www.penguincomputing.com/services-support/.

Visit https://docs.redhat.com/docs/en-US/Red_Hat_Enterprise_Linux to view the Red Hat Enterprise Linux 5 *5.11 Release Notes* and *5.11 Technical Notes*.

For the most up-to-date product documentation and other helpful information, visit the Penguin Computing Support Portal.

Important

Before continuing, make sure you are reading the most recent *Release Notes*, which can be found on the Penguin Computing Support Portal at http://www.penguincomputing.com/services-support/documentation/. The most recent version will accurately reflect the current state of the Scyld ClusterWare yum repository of rpms that you are about to install. You may consult the *Installation Guide* for its more generic and expansive details about the installation process. The *Release Notes* document more specifically describes how to upgrade an earlier version of Scyld ClusterWare to Scyld ClusterWare 5.11.2 (see the Section called *Upgrading Earlier Release of Scyld ClusterWare to Scyld ClusterWare 5.11.2*), or how to install Scyld ClusterWare 5.11.2 as a fresh install (see the Section called *Installing Scyld ClusterWare 5.11.2 on a Non-Scyld ClusterWare System*).

Important for clusters using Panasas storage

If the cluster uses Panasas storage, then you must ensure that a Panasas kernel module is available that matches the Scyld ClusterWare kernel you are about to install: 2.6.18-400.1.1.el5.5112g0000. Login to your Panasas account at http://www.my.panasas.com, click on the *Downloads* tab, then click on *DirectFLOW Client for Linux* and then on *Search DirectFLOW Release*, and do a *Keyword* search for `5112g0000`. If you find a Panasas rpm matching the to-be-installed 2.6.18-400.1.1.el5.5112g0000 kernel, then download that rpm and continue with the Scyld ClusterWare update or install. Install the Panasas rpm after you finish installing the associated 2.6.18-400.1.1.el5.5112g0000 kernel. If you do not find an appropriate Panasas rpm, then do not install this latest Scyld ClusterWare 5.11.2. The Panasas storage will not work with the 2.6.18-400.1.1.el5.5112g0000 kernel without a matching Panasas kernel module.

Upgrading Earlier Release of Scyld ClusterWare to Scyld ClusterWare 5.11.2

If you wish to upgrade a RHEL4 (or CentOS4) base distribution to RHEL5/CentOS5, then we recommend you accomplish this with a full install of Release 5, rather than attempt to *update* from an earlier major release to Release 5. Visit https://docs.redhat.com/docs/en-US/Red_Hat_Enterprise_Linux for the Red Hat Enterprise Linux 5 *Installation Guide* for details. If you already have installed Scyld ClusterWare 4 on the physical hardware that you intend to convert to RHEL5/CentOS5, then we recommend that you backup your master node prior to the new installation of RHEL5/CentOS5, as some of the Scyld ClusterWare configuration files may be a useful reference for Release 5, especially files in `/etc/beowulf/`.

When upgrading from an earlier Scyld ClusterWare 5 version to Scyld ClusterWare 5.11.2, you should perform the following steps:

1. Stop the Beowulf cluster: `/sbin/service beowulf stop`
2. Clean the yum cache to a known state: `yum clean all`
3. Update the RHEL5/CentOS5 base distribution, taking care to not update or install Scyld ClusterWare or a kernel from the base distribution:

   ```
   yum --disablerepo=cw* --exclude=kernel-* update
   ```

4. Remove any **ecryptfs-utils** package older than version 44:

   ```
   rpm -q ecryptfs-utils
   ```

   to determine which version(s) are installed. If, for example, this shows you:

   ```
   ecryptfs-utils-38-8.el5
   ecryptfs-utils-38-8.el5
   ecryptfs-utils-56-8.el5
   ecryptfs-utils-56-8.el5
   ```

   then you must remove both the i386 and x86_64 version 38 packages:

   ```
   rpm -e --allmatches ecryptfs-utils-38
   ```

5. Ensure that file `/etc/yum.repos.d/clusterware.repo` specifies the yum repository *baseurl* version *5.11*.
6. If your cluster includes Infiniband hardware, then you can ensure that all the necessary Infiniband-related rpms from the base distribution are installed:

   ```
   yum groupinstall Infiniband
   ```

   where the *Infiniband* group is defined through the Scyld ClusterWare repo configuration file and refers to rpms that reside in the base distribution repository.
7. Now *groupupdate* to the newest Scyld ClusterWare packages. If you want to **retain** the existing openmpi-*x.y*-scyld packages, then follow the instructions in the Section called *Issues with OpenMPI*. Otherwise, update to the newest packages and **replace** existing openmpi-*x.y*-scyld packages by doing:

   ```
   yum groupupdate Scyld-ClusterWare
   ```

8. If **yum** fails with a complaint involving *kmod-tg3-3.86*, then you must manually remove that no longer needed package:

   ```
   rpm -qa | grep kmod-tg3-3.86 | xargs rpm -e
   ```

   and then repeat the groupupdate.
9. If **yum** fails with a *Transaction Check Error* that complains that a base distribution rpm is newer than the Scyld ClusterWare rpm that is attempting to replace it, then you must manually install the downlevel Scyld ClusterWare rpm(s). For example, if the complaint is about the **kernel** rpms, then do:

   ```
   cd /var/cache/yum/
   ls cw-*/packages/kernel-*
   ```

   and locate the newest Scyld ClusterWare 2.6.18-400.1.1.el5.5112g0000 kernel, which should reside in either `cw-core/packages/` or `cw-updates/packages/`. Then install that newest kernel and related packages:

   ```
   rpm -iv --oldpackage cw-*/packages/kernel-*5112g0000*rpm
   ```

   and then repeat the groupupdate.
10. If **yum** fails with a *Transaction Check Error* that complains that the new `kmod-task_packer` conflicts with a file from an earlier `kmod-task_packer` package, then you need to explicitly remove the earlier set of rpms:

    ```
    rpm -qa | egrep "kmod.*bproc|task_packer|filecache" | xargs rpm -e --nodeps
    ```

    and then repeat the groupupdate, which should now *Complete!* successfully.

11. **If the cluster uses Panasas storage**, then you should have already downloaded the Panasas rpm that matches the Scyld ClusterWare 5.11.2 kernel you have just installed. Now install the Panasas rpm using **rpm -i**.
12. Compare `/etc/beowulf/config`, which remains untouched by the Scyld ClusterWare update, with the new `config.rpmnew` (if that file exists), examine the differences:

    ```
    cd /etc/beowulf
    diff config config.rpmnew
    ```

    and carefully merge the `config.rpmnew` differences into `/etc/beowulf/config`. Please see the Section called *Resolve *.rpmnew and *.rpmsave configuration file differences* for details.

    Similarly, the preexisting `/etc/beowulf/fstab` may have been saved as `fstab.rpmsave` if it was locally modified. If so, merge those local changes back into `/etc/beowulf/fstab`.
13. Examine `/etc/grub.conf` to confirm that the new 2.6.18-400.1.1.el5.5112g0000 kernel is the default, then reboot your master node.
14. After rebooting the new kernel, and after installing any new kernel modules, you should rebuild the master node's list of modules and dependencies using **/sbin/depmod**. See the Section called *Issues with kernel modules* for details.

Installing Scyld ClusterWare 5.11.2 on a Non-Scyld ClusterWare System

When installing Scyld ClusterWare 5.11.2 on a system that does not yet contain Scyld ClusterWare, you should perform the following steps:

1. Clean the yum cache to a known state: `yum clean all`
2. Update the RHEL5/CentOS5 base distribution, taking care to not update or install Scyld ClusterWare or a kernel from the base distribution:

    ```
    yum --disablerepo=cw* --exclude=kernel-* update
    ```

    then remove base distribution packages that conflict with Scyld ClusterWare:

    ```
    yum remove openmpi* mvapich*
    ```
3. Remove any **ecryptfs-utils** package older than version 44:

    ```
    rpm -q ecryptfs-utils
    ```

    to determine which version(s) are installed. If, for example, this shows you:

    ```
    ecryptfs-utils-38-8.el5
    ecryptfs-utils-38-8.el5
    ecryptfs-utils-56-8.el5
    ecryptfs-utils-56-8.el5
    ```

    then you must remove both the i386 and x86_64 version 38 packages:

    ```
    rpm -e --allmatches ecryptfs-utils-38
    ```
4. Download a custom Yum repo file:

    a. Login to the Penguin Computing Support Portal at http://www.penguincomputing.com/services-support/.

    b. Click on *Download your Scyld ClusterWare 5 YUM repo file* to download the new `clusterware.repo` file and place it in the `/etc/yum.repos.d/` directory.

    c. Set the permissions: `chmod 644 /etc/yum.repos.d/clusterware.repo`

   d. The new `clusterware.repo` contains a *baseurl* entry that uses `https` by default. If your local site is configured to not support such encrypted accesses, then you must edit the repo file to instead use `http`.

5. If your cluster includes Infiniband hardware, then you can ensure that all the necessary Infiniband-related rpms from the base distribution are installed:

   ```
   yum groupinstall Infiniband
   ```

   where the *Infiniband* group is defined through the Scyld ClusterWare repo configuration file and refers to rpms that reside in the base distribution repository.

6. Install Scyld ClusterWare:

   ```
   yum groupinstall Scyld-ClusterWare
   ```

   If **yum** fails with a *Transaction Check Error* that complains that a base distribution rpm is newer than the Scyld ClusterWare rpm that is attempting to replace it, then you must manually install the downlevel Scyld ClusterWare rpm(s). For example, if the complaint is about the **kernel** rpms, then do:

   ```
   cd /var/cache/yum/
   rpm -iv --oldpackage cw-*/packages/kernel-*5112g0000*rpm
   ```

   and then repeat the groupinstall:

   ```
   yum groupinstall Scyld-ClusterWare
   ```

   which should now *Complete!* successfully.

7. **If the cluster uses Panasas storage**, then you should have already downloaded the Panasas rpm that matches the Scyld ClusterWare 5.11.2 kernel you have just installed. Now install the Panasas rpm using **rpm -i**.

8. Configure the network for Scyld ClusterWare: run **/usr/sbin/beonetconf** to specify the cluster interface, the maximum number of compute nodes, and the beginning IP address of the first compute node. See the *Installation Guide* for more details.

9. If the private cluster network switch uses Spanning Tree Protocol (STP), then either reconfigure the switch to disable STP, or if that is not feasible because of network topology, then enable *Rapid STP* or *portfast* on the compute node and edge ports. See the Section called *Issues with Spanning Tree Protocol and portfast* for details.

10. Examine `/etc/grub.conf` to confirm that the new 2.6.18-400.1.1.el5.5112g0000 kernel is the default, then reboot your master node.

11. After rebooting the new kernel, and after installing any new kernel modules, you should rebuild the master node's list of modules and dependencies using **/sbin/depmod**. See the Section called *Issues with kernel modules* for details.

12. The first time Beowulf services start, e.g., when doing **/sbin/service beowulf start** or **/etc/init.d/beowulf start**, you will be prompted to accept a Scyld ClusterWare End User License Agreement (EULA). If you answer with an affirmative *yes*, then Beowulf services start and Scyld ClusterWare functionality is available, and you will not be prompted again regarding the EULA.

    However, if you do not answer with *yes*, then Beowulf services will not start, although the master node will continue to support all non-Scyld ClusterWare functionality available from the base distribution. Any subsequent attempt to start Beowulf services will again result in a prompt for you to accept the EULA.

    Note: if Beowulf is configured to automatically start when the master node boots (i.e., **/sbin/chkconfig --list beowulf** shows Beowulf *on* for levels 3, 4, and 5), then the first reboot after installing Scyld ClusterWare will fail to start Beowulf because `/etc/init.d/beowulf` is not executed interactively and no human sees the prompt for EULA acceptance. In this event, you may start Beowulf manually, e.g., using **/sbin/service beowulf start**, and respond to the EULA prompt.

Post-Installation Configuration Issues

Following a successful update or install of Scyld ClusterWare, you may need to make one or more configuration changes, depending upon the local requirements of your cluster. Larger cluster configurations have additional issues to consider; see the Section called *Post-Installation Configuration Issues For Large Clusters*.

Resolve **.rpmnew* and **.rpmsave* configuration file differences

As with every Scyld ClusterWare upgrade, after the upgrade you should locate any Scyld ClusterWare `*.rpmsave` and `*.rpmnew` files and perform merges, as appropriate, to carry forward the local changes. Sometimes an upgrade will save the locally modified version as `*.rpmsave` and overwrite the basic file with a new version. Other times the upgrade will keep the locally modified version untouched, installing the new version as `*.rpmnew`.

For example,

```
cd /etc/beowulf
find . -name \*rpmnew
find . -name \*rpmsave
```

and examine each such file to understand how it differs from the configuration file that existed prior to the update. You may need to merge new lines from the newer `*.rpmnew` file into the existing file, or perhaps replace existing lines with new modifications. For instance, this is commonly done with `/etc/beowulf/config` and `config.rpmnew`. Or you may need to merge older local modifications in `*.rpmsave` into the newly installed pristine version of the file. For instance, this is occasionally done with `/etc/beowulf/fstab.rpmsave`.

Generally speaking, be careful when making changes to `/etc/beowulf/config`, as mistakes may leave your cluster in a non-working state. For example, in general do not manually change the existing keyword entries for *interface*, *nodes*, *iprange*, and *nodeassign*, as those are more accurately manipulated by the **/usr/sbin/beonetconf** command. The *kernelimage* and *node* entries are automatically managed by Beowulf services and should not be merged.

The remaining differences are candidates for careful merging. Pay special attention to merge additions to the *bootmodule*, *modarg*, *server*, *libraries*, and *prestage* keyword entries. New *nodename* entries for *infiniband* or *ipmi* are offsets to each node's IP address on the private cluster network, and these offsets may need to be altered to be compatible with your local network subnet. Also, be sure to merge differences in `config.rpmnew` comments, as those are important documentation information for future reference.

Contact Scyld Customer Support if you are unsure about how to resolve particular differences, especially with `/etc/beowulf/config`.

Disable SELinux

Scyld ClusterWare execution currently requires that SELinux be disabled. Edit `/etc/sysconfig/selinux` and ensure that *SELINUX=disabled* is set. If SELinux was not already set to *disabled*, then the master node must be rebooted for this change to take effect.

Disable NetworkManager

Scyld ClusterWare execution currently requires that NetworkManager be disabled. If NetworkManager is installed, then:

```
/sbin/service NetworkManager stop
/sbin/chkconfig NetworkManager off
```

if the service is enabled.

Disable library prelinking

Scyld ClusterWare migration between cluster nodes requires stable dynamic libraries. Edit `/etc/sysconfig/prelink` and ensure that *PRELINKING=no* is set. This will permanently block subsequent (usually daily) **/usr/sbin/prelink** operations. In addition, to immediately undo prelinking:

```
/usr/sbin/prelink --undo -all
```

See the *Administrator's Guide* for more details.

Optionally reduce size of /usr/lib/locale/locale-archive

Glibc applications silently open the file `/usr/lib/locale/locale-archive`, which means it gets downloaded by each compute node early in a node's startup sequence. The default RHEL5 `locale-archive` is 54 MBytes in size. This consumes significant network bandwidth and causes serialization delays if numerous compute nodes attempt to concurrently boot, and thereafter this large file consumes significant RAM filesystem space on the node. It is likely that a cluster's users and applications do not require all the international locale data that is present in the default file. With care, the cluster administrator may choose to rebuild `locale-archive` with a greatly reduced set of locales and thus create a significantly smaller file. See the Administrator's Guide for details.

Optionally enable TORQUE

If you wish to run TORQUE, enable it on the master node:

```
/sbin/chkconfig torque on
```

After you successfully start the cluster compute nodes for the first time, enable the `/etc/beowulf/init.d/torque` script:

```
/sbin/beochkconfig 90torque on
```

then restart TORQUE and restart the compute nodes:

```
/sbin/service torque restart
/usr/sbin/bpctl -S all -R
```

See the *Administrator's Guide* for more details about TORQUE configuration, and the *User's Guide* for details about how to use TORQUE.

Optionally enable TORQUE scheduler

Scyld ClusterWare installs by default both the TORQUE resource manager and the Maui job scheduler. The Maui installation can coexist with an optionally licensed Moab job scheduler installation, although after the initial installation of either of these job schedulers, the cluster administrator needs to make a one-time choice of which job scheduler to employ.

If Moab is not installed, then simply activate Maui by moving into place two global profile files that execute **module load maui** and then start the *maui* service:

```
cp /opt/scyld/maui/scyld.maui.{csh,sh} /etc/profile.d
/sbin/chkconfig maui on
/sbin/service maui start
```

If Moab was previously installed, is currently active, and is the preferred job scheduler, then the cluster administrator can ignore the Maui installation (and any subsequent Maui updates) because Maui installs in a deactivated state and will not affect Moab.

If Maui is active and the cluster administrator subsequently installs Moab, or chooses to use an already installed Moab as the default scheduler, then deactivate Maui so as to not affect Moab:

```
rm /etc/profile.d/scyld.maui.*
/sbin/chkconfig maui off
/sbin/service maui stop
```

and then activate Moab as appropriate for the cluster.

Optionally enable Scyld Integrated Management Framework (IMF)

Scyld IMF is used by a cluster administrator to monitor and administer the cluster using a Web browser. It requires Apache on the master node (service *httpd*) and is access-protected with a Web application-specific username *admin* and password combination.

To enable the Scyld IMF interface, perform the following steps on the master node:

1. Enable the *httpd* service, if it is not already enabled:

   ```
   /sbin/chkconfig httpd on
   /sbin/service httpd start
   ```

2. Initialize the username *admin* by assigning it a unique password:

   ```
   /usr/bin/htpasswd /etc/httpd/scyld-imf/htpasswd-users admin
   ```

To use Scyld IMF, point your Web browser at the URL http://MasterNode/scyld-imf, where *MasterNode* is the name or IP address of the master node, whereupon you are prompted for a valid username (i.e., *admin*) and password (which was initialized as described above). See the *Administrator's Guide* for more details.

Optionally enable Ganglia monitoring tool

To enable the Ganglia cluster monitoring tool,

```
/sbin/chkconfig beostat on
/sbin/chkconfig xinetd on
/sbin/chkconfig httpd on
/sbin/chkconfig gmetad on
```

then either reboot the master node, which automatically restarts these system services; or without rebooting, manually restart *xinetd* then start the remaining services that are not already running:

```
/sbin/service xinetd restart
/sbin/service httpd start
/sbin/service gmetad start
```

See the *Administrator's Guide* for more details.

Optionally enable beoweb service

The beoweb service facilitates remote job submission and cluster monitoring (e.g., used by POD Tools). Enable and start beoweb:

```
/sbin/chkconfig beoweb on
/sbin/service beoweb start
```

See the *Administrator's Guide* for more details.

Optionally enable NFS locking

If you wish to use cluster-wide NFS locking, then you must enable locking on the master node and on the compute nodes. First ensure that NFS locking is enabled and running on the master:

```
/sbin/chkconfig nfslock on
/sbin/service nfslock start
```

Then for each NFS mount point for which you need the locking functionality, you must edit `/etc/beowulf/fstab` (or the appropriate node-specific `/etc/beowulf/fstab.`*N* file(s)) to remove the default option *nolock*. See the *Administrator's Guide* for more details.

Optionally adjust the size limit for locked memory

OpenIB, MVAPICH, and MVAPICH2 require an override to the limit of how much memory can be locked.

Scyld ClusterWare adds a *memlock* override entry to `/etc/security/limits.conf` during a Scyld ClusterWare upgrade (if the override entry does not already exist in that file), regardless of whether or not Infiniband is present in the cluster. The new override line,

```
*    -    memlock    unlimited
```

raises the limit to *unlimited*. If Infiniband is not present, then this new override line is unnecessary and may be deleted. If Infiniband is present, we recommend leaving the new *unlimited* line in place. If you choose to experiment with a smaller discrete value, then understand that Scyld ClusterWare MVAPICH requires a minimum of 16,384 KBytes, which means changing *unlimited* to *16384*. If your new discrete value is too small, then MVAPICH reports a "CQ Creation" or "QP Creation" error.

Optionally enable automatic CPU frequency management

If you wish to enable automatic CPU frequency management, you must have the base distribution's **cpuspeed** package installed, and then enable the Scyld ClusterWare script:

```
/sbin/beochkconfig 30cpuspeed on
```

You may optionally create a configuration file `/etc/beowulf/conf.d/cpuspeed.conf` (or node-specific `cpuspeed.conf.`*N*), ostensibly derived from the master node's `/etc/sysconfig/cpuspeed`, to override default behavior. See **man cpuspeed** for details.

Optionally enable SSHD on compute nodes

If you wish to allow users to execute MVAPICH2 applications, or to use **/usr/bin/ssh** or **/usr/bin/scp** from the master to a compute node, or from one compute node to another compute node, then you must enable **sshd** on compute nodes by enabling the script:

```
/sbin/beochkconfig 81sshd on
```

The cluster is preconfigured to allow user *root* ssh access to compute nodes. The cluster administrator may wish to configure the cluster to allow ssh access for non-root users. See the *Administrator's Guide* for details.

Optionally allow IP Forwarding

By default, the master node does not allow IP Forwarding from compute nodes on the private cluster network to external IP addresses on the public network. If IP Forwarding is desired, then edit `/etc/beowulf/config` to enable the directive *ipforward yes*, and ensure that the file `/etc/sysconfig/iptables` eliminates or comments-out the default entry:

```
-A FORWARD -j REJECT --reject-with icmp-host-prohibited
```

Optionally increase the ip_conntrack table size

Certain workloads doing IP forwarding may trigger a syslog message *ip_conntrack: table full, dropping packet*. If IP forwarding is enabled, then at cluster startup time Scyld ClusterWare insures a max max table size of at least 524,288 and a related table hashsize of 65,536 (maintaining the default 8-to-1 ratio for performance reasons). However, this max value may still be inadequate for local workloads, and the *table full, dropping packet* syslog messages may still occur. Use:

```
/sbin/sysctl net.ipv4.ip_conntrack_max
```

to see the current max size, then keep manually increasing the max until the syslog messages stop occurring, e.g., use:

```
/sbin/sysctl -w net.ipv4.ip_conntrack_max=Nmax
```

to try new *Nmax* values. An effective *Nmax* also determines an effective *Nhash* hashsize: 1/8th the *Nmax* value. Make these values persist across master node reboots by adding:

```
options ip_conntrack hashsize=Nhash
```

to `/etc/modprobe.conf`, and adding:

```
net.ipv4.ip_conntrack_max=Nmax
```

to `/etc/sysctl.conf`.

Optionally reconfigure node names

You may declare site-specific alternative node names for cluster nodes by adding entries to `/etc/beowulf/config`. The syntax for a node name entry is:

```
nodename format-string [IPv4offset] [netgroup]
```

For example,

```
nodename node%N
```

allows the user to refer to node 4 using the traditional *.4* name, or alternatively using names like *node4* or *node004*. See **man beowulf-config** and the *Administrator's Guide* for details.

Post-Installation Configuration Issues For Large Clusters

Larger clusters have additional issues that may require post-installation adjustments.

Optionally increase the number of nfsd threads

The default count of 8 **nfsd** NFS daemons may be insufficient for large clusters. One symptom of an insufficiency is a syslog message, most commonly seen when you currently boot all the cluster nodes:

```
nfsd: too many open TCP sockets, consider increasing the number of nfsd threads
```

Scyld ClusterWare automatically increases the nfsd thread count to at least one thread per compute node, with a lowerbound of eight (for <=8 nodes) and an upperbound of 64 (for >=64 nodes). If this increase is insufficient, then increase the thread count (e.g., to 16) by executing:

```
echo 16 > /proc/fs/nfsd/threads
```

Ideally, the chosen thread count should be sufficient to eliminate the syslog complaints, but not significantly higher, as that would unnecessarily consume system resources. One approach is to repeatedly double the thread count until the syslog error messages stop occurring, then make the satisfactory value *N* persistent across master node reboots by creating the file `/etc/sysconfig/nfs`, if it does not already exist, and adding to it an entry of the form:

```
RPCNFSDCOUNT=N
```

A value *N* of 1.5x to 2x the number of nodes is probably adequate, although perhaps excessive. See the *Administrator's Guide* for a more detailed discussion of NFS configuration.

Optionally increase the max number of processID values

The kernel defaults to using a maximum of 32,768 processID values. Scyld ClusterWare automatically increases this default to 98,304 [= 3*32768], which likely is adequate for small- to medium-size clusters and which keeps pid values at a familiar 5-column width maximum. Because BProc manages a common process space across the cluster, even the increase to 98,304 may be insufficient for very large clusters and/or workloads that create large numbers of concurrent processes. The cluster administrator can increase the value further by using the **sysctl** command, e.g.,

```
/sbin/sysctl -w kernel.pid_max=N
```

directs the kernel to use pid values up to *N*. The kernel (and BProc) supports an upperbound of 4,194,304 [= (4*1024*1024)]. To set a value *N* that persists across master node reboots, add an entry

```
kernel.pid_max=N
```

to `/etc/sysctl.conf`.

Optionally increase the max number of open files

RHEL5 defaults to a maximum of 1024 concurrently open files. This value may be too low for large clusters. The cluster administrator can add a *nofile* override entry to `/etc/security/limits.conf` to specify a larger value. Caution: for *nofile*, use only a numeric upperbound value, never *unlimited*, as that will result in being unable to login.

Issues with Ganglia

The Ganglia cluster monitoring tool may fail for large clusters. If the `/var/log/httpd/error_log` shows a fatal error of the form *PHP Fatal error: Allowed memory size of 8388608 bytes exhausted*, then edit the file `/etc/php.ini` to increase the *memory_limit* parameter. The default is *memory_limit = 8M* can be safely doubled and re-doubled until the error goes away.

Post-Installation Release of Updated Packages

From time to time, Penguin Computing may release updated Scyld ClusterWare 5.11 rpms to track Red Hat kernel security or bug fix errata or to fix critical Scyld ClusterWare problems. You can check for the availability of updated Scyld ClusterWare rpms by doing:

```
yum list updates --disablerepo=* --enablerepo=cw*
```

If updates are available, you should first download the latest version of the Scyld ClusterWare 5 *Release Notes* from the Penguin Computing Support Portal (http://www.penguincomputing.com/services-support/documentation/) to ensure you have the latest guidance before updating your cluster. In general, if you choose to update Scyld ClusterWare packages, then you should update all available packages.

If your cluster uses Panasas storage, then before updating Scyld ClusterWare you must ensure that a Panasas kernel module is available that matches the Scyld ClusterWare kernel that will be installed. See the section called *Important for clusters using Panasas storage* in the *About This Release* introduction for more information.

Notable Feature Enhancements And Bug Fixes Beyond Scyld ClusterWare 5.2.0

New in Scyld ClusterWare 5.11.2 - Scyld Release 5112g0000 - January 20, 2015

1. The base kernel updates to 2.6.18-400.1.1. See https://rhn.redhat.com/errata/RHSA-2014-2008.html for details.
2. The igb Ethernet driver updates to version 5.2.15, derived from http://sourceforge.net/projects/e1000/files/.
3. The optional e1000e Ethernet driver updates to version 3.1.0.2, derived from http://sourceforge.net/projects/e1000/files/.
4. The openmpi-1.8-scyld packages update to version 1.8.4, which by default update and replace only earlier version 1.8 packages and do not affect OpenMPI version 1.7, 1.6, or 1.5 packages. See the Section called *Issues with Open-MPI* for general issues about supporting multiple concurrent versions. The libraries were built with Gnu version 4.1.2, Intel version 2013_sp1.3.174, and PGI version 14.6 compiler families. Scyld releases of OpenMPI are derived from http://www.open-mpi.org/. See the *User's Guide* Appendix C, *OpenMPI Release Information* for details.

New in Scyld ClusterWare 5.11.1 - Scyld Release 5111g0000 - December 15, 2014

1. The base kernel updates to 2.6.18-400. See https://rhn.redhat.com/errata/RHSA-2014-1959.html for details.
2. Fix a compute node **bpslave** "soft lockup" that would occasionally occur during node boot.
3. The bproc *filecache* functionality now downloads to a compute node a mirror image of the master node's symlinks that follow a path to the final leaf file. For example, opening `/lib64/libcrypt.so.1` creates the symlink `/lib64/libcrypt.so.1` and downloads the leaf file `/lib64/libcrypt-2.5.so`. Previously, bproc *filecache* downloaded only the final leaf file and named it `/lib64/libcrypt.so.1`. This requires a coordinated update to the *beoserv*, *beoclient3*, and *bproc* packages.
4. The beonss **kickbackproxy** daemon that executes on each compute node now throttles its attempts to reconnect to the master node **kickbackdaemon** server if that connection has been lost. Previously, the **kickbackproxy** would rapidly attempt to reconnect, thereby keeping an otherwise idle orphaned compute node busy and thus constraining a run-to-completion reboot.

New in Scyld ClusterWare 5.11.0 - Scyld Release 5110g0000 - October 19, 2014

1. The base kernel updates to 2.6.18-398. See https://rhn.redhat.com/errata/RHBA-2014-1196.html for details.
2. Populate each compute node at boot time by pushing the master node's file `/etc/beowulf/conf.d/limits.conf` to the the node as `/etc/security/limits.conf`. This master node's file is initially a concatenation of the master node's `/etc/security/limits.conf` and the files in the directory `/etc/security/limits.d/`. The cluster administrator may edit `/etc/beowulf/conf.d/limits.conf` as desired.
3. Fix a compute node hang that can occur when attempting to link an application that references a nonexistent library file.
4. Support bproc *filecache* pathnames that include embedded `/../` strings. Previously, these were rejected without resolving the true pathname.
5. Fix a rare bug that exhibits itself as a compute node that continually retries an unsuccessful boot, complaining that the communication **bpslave**-**bpmaster** communication (which defaults to port 932) cannot be established.
6. TORQUE updates to version 4.2.9, from www.adaptivecomputing.com/support/download-center/torque-download/.
7. The openmpi-1.8-scyld packages update to version 1.8.3, which by default update and replace only earlier version 1.8 packages and do not affect OpenMPI version 1.7, 1.6, or 1.5 packages. See the Section called *Issues with Open-MPI* for general issues about supporting multiple concurrent versions. The libraries were built with Gnu version 4.1.2, Intel version 2013_sp1.3.174, and PGI version 14.6 compiler families. Scyld releases of OpenMPI are derived from http://www.open-mpi.org/. See the *User's Guide* Appendix C, *OpenMPI Release Information* for details.
8. The MPICH3 mpich-scyld release updates to version 3.1.3, derived from http://www.mpich.org/. The libraries were built with Gnu version 4.1.2, Intel version 2013_sp1.3.174, and PGI version 14.6 compiler families. See the *User's Guide* Appendix F, *MPICH-3 Release Information* for details.
9. Fix a problem in PVM that results in a hung application with unkillable threads.
10. NVIDIA K40 GPU now executes in *persistance* mode for quicker startup of GPU operations.

New in Scyld ClusterWare 5.10.6 - Scyld Release 5106g0000 - August 20, 2014

1. The base kernel updates to 2.6.18-371.11.1. See https://rhn.redhat.com/errata/RHSA-2014-0926.html for details.

2. The MPICH3 mpich-scyld release updates to version 3.1.2, derived from http://www.mpich.org/. The libraries were built with Gnu version 4.1.2, Intel version 2013_sp1.3.174, and PGI version 14.6 compiler families. See the *User's Guide* Appendix F, *MPICH-3 Release Information* for details.

New in Scyld ClusterWare 5.10.5 - Scyld Release 5105g0000 - July 23, 2014

1. The base kernel updates to 2.6.18-371.9.1. See https://rhn.redhat.com/errata/RHSA-2014-0740.html for details.
2. **/sbin/service beowulf reload** now re-reads the `/etc/beowulf/config` *libraries* entries and rebuilds the list of libraries managed by the bproc *filecache* functionality for the master node and all the *up* compute nodes.
3. TORQUE updates to version 4.2.8, from www.adaptivecomputing.com/support/download-center/torque-download/.
4. MVAPICH2 updates to version 2.0, derived from http://mvapich.cse.ohio-state.edu/. See the *User's Guide* Appendix E, *MVAPICH2 Release Information* for details.
5. The MPICH3 mpich-scyld release updates to version 3.1.1, derived from http://www.mpich.org/. See the *User's Guide* Appendix F, *MPICH-3 Release Information* for details.
6. The various MPI library suites (OpenMPI, MPICH, MPICH2, MVAPICH2, MPICH3) have been rebuilt with newer versions of the Intel (v2013_sp1.3.174) and PGI (v14.6) compiler families.

New in Scyld ClusterWare 5.10.4 - Scyld Release 5104g0000 - June 2, 2014

1. The base kernel updates to 2.6.18-371.8.1. See https://rhn.redhat.com/errata/RHSA-2014-0433.html for details.
2. The igb Ethernet driver updates to version 5.2.5, derived from http://sourceforge.net/projects/e1000/files/.
3. Scyld ClusterWare now distributes openmpi-1.8-scyld packages, which are a redistribution of OpenMPI version 1.8 and derived from http://www.open-mpi.org/. See the *User's Guide* Appendix C, *OpenMPI Release Information* for details.

New in Scyld ClusterWare 5.10.3 - Scyld Release 5103g0000 - April 3, 2014

1. The base kernel updates to 2.6.18-371.6.1. See https://rhn.redhat.com/errata/RHSA-2014-0285.html for details.
2. The igb Ethernet driver updates to version 5.1.2, derived from http://sourceforge.net/projects/e1000/files/.
3. Scyld ClusterWare now distributes an optional e1000e Ethernet driver, version 3.0.4, derived from http://sourceforge.net/projects/e1000/files/.
4. TORQUE updates to version 4.2.7, from www.adaptivecomputing.com/support/download-center/torque-download/.
5. The openmpi-1.7-scyld packages update to version 1.7.5, which by default update and replace only earlier version 1.7 packages and do not affect OpenMPI version 1.6 or 1.5 packages. See the Section called *Issues with OpenMPI* for general issues about supporting multiple concurrent versions. Scyld releases of OpenMPI are derived from http://www.open-mpi.org/. See the *User's Guide* Appendix C, *OpenMPI Release Information* for details.
6. The mpich-scyld release updates to version 3.1, derived from http://www.mpich.org/. See the *User's Guide* Appendix F, *MPICH-3 Release Information* for details.
7. **/sbin/service beowulf start** and **restart** now check the size of `/usr/lib/locale/locale-archive` and issue a warning if the file is huge and thus would impact cluster performance. See the Administrator's Guide for details.

New in Scyld ClusterWare 5.10.2 - Scyld Release 5102g0000 - February 14, 2014

1. The base kernel updates to 2.6.18-371.4.1. See https://rhn.redhat.com/errata/RHSA-2014-0108.html for details.
2. Distribute a basic `/etc/beowulf/init.d/89opensm` script that optionally starts OpenSM on compute nodes. The script is initially disabled.
3. Eliminate the unnecessary requirement that TORQUE Python libraries be installed in order for **beostatus** filtering to work.

New in Scyld ClusterWare 5.10.1 - Scyld Release 5101g0000 - December 18, 2013

1. The base kernel updates to 2.6.18-371.3.1. See https://rhn.redhat.com/errata/RHSA-2013-1790.html for details.
2. TORQUE updates to version 4.2.6.1, from www.adaptivecomputing.com/support/download-center/torque-download/.
3. Improve the synchronization between the **beoserv** and **bpmaster** daemons with respect to what port number the latter wishes to use to communicate with the compute node **bpslave** daemons.

New in Scyld ClusterWare 5.10.0 - Scyld Release 5100g0000 - November 18, 2013

1. The base kernel is updated to 2.6.18-371.1.2. See https://rhn.redhat.com/errata/RHSA-2013-1348.html and https://rhn.redhat.com/errata/RHSA-2013-1449.html for details.
2. The igb Ethernet driver updates to version 5.0.6, derived from http://sourceforge.net/projects/e1000/files/.
3. The openmpi-1.7-scyld packages are updated to version 1.7.3, which by default update and replace only earlier version 1.7 packages and do not affect OpenMPI version 1.6 or 1.5 packages. See the Section called *Issues with Open-MPI* for general issues about supporting multiple concurrent versions. Scyld releases of OpenMPI are derived from http://www.open-mpi.org/. See the *User's Guide* Appendix C, *OpenMPI Release Information* for details.
4. The MVAPICH2 release is updated to version 2.0b, derived from http://mvapich.cse.ohio-state.edu/. See the *User's Guide* Appendix E, *MVAPICH2 Release Information* for details.
5. The **sendstats** daemon more reliably avoids being started more than once per compute node.

New in Scyld ClusterWare 5.9.6 - Scyld Release 596g0000 - October 16, 2013

1. The base kernel is updated to 2.6.18-348.18.1. See https://rhn.redhat.com/errata/RHSA-2013-1292.html for details.

New in Scyld ClusterWare 5.9.5 - Scyld Release 595g0002 - September 27, 2013

1. TORQUE updates to version 4.2.5, from www.adaptivecomputing.com/support/download-center/torque-download/. This distribution also fixes problems that were introduced by the recent Scyld ClusterWare inclusion of the *Maui* scheduler, where the installation of *Maui* perturbed an optionally preexisting installation of the *Moab* scheduler. Both *Maui* and *Moab* can now coexist as installed packages, although the local cluster administrator must perform a one-time

selection of which scheduler to use, if both are installed. See the Section called *Optionally enable TORQUE scheduler* for details.

New in Scyld ClusterWare 5.9.5 - Scyld Release 595g0001 - September 6, 2013

1. TORQUE updates to version 4.2.4.1, from www.adaptivecomputing.com/support/download-center/torque-download/. This TORQUE also fixes a *pbs_mom* security vulnerability that was announced by Adaptive Computing on Sept. 6, 2013, that afflicts all TORQUE releases to date.

New in Scyld ClusterWare 5.9.5 - Scyld Release 595g0000 - August 30, 2013

1. The base kernel is updated to 2.6.18-348.16.1. See https://rhn.redhat.com/errata/RHSA-2013-1166.html for details.
2. TORQUE updates to version 4.2.4, from www.adaptivecomputing.com/support/download-center/torque-download/. This Scyld ClusterWare distribution changes the default job scheduler from the problematic built-in *pbs_sched* to Adaptive Computing's *Maui*, currently version 3.3.1. Maui distributes as a separate rpm and is required by TORQUE 4.2.4. See the *User's Guide* Appendix B, *TORQUE and Maui Release Information* for details.
3. The MVAPICH2 release is updated to version 2.0a, derived from http://mvapich.cse.ohio-state.edu/. See the *User's Guide* Appendix E, *MVAPICH2 Release Information* for details.
4. The fftw, mpich2, mvapich2, mpich-scyld, and openmpi-1.5, -1.6, and -1.7 packages have been rebuilt using newer Intel and PGI compiler suites: Intel composer_xe 2013.5.192 vs. composerxe-2011.4.191, and PGI 13.6 vs. 11.9.

New in Scyld ClusterWare 5.9.4 - Scyld Release 594g0000 - July 19, 2013

1. The base kernel is updated to 2.6.18-348.12.1. See https://rhn.redhat.com/errata/RHSA-2013-0847.html and https://rhn.redhat.com/errata/RHSA-2013-1034.html for details.
2. The Scyld ClusterWare packaging for OpenMPI has changed in order to more easily install and retain multiple co-existing versions on the master node. See the Section called *Issues with OpenMPI* for general issues about supporting multiple concurrent versions. This Scyld release installs openmpi-1.7-scyld rpms containing OpenMPI version 1.7.2 and openmpi-1.6-scyld rpms containing OpenMPI version 1.6.5. These openmpi-1.6-scyld rpms will only update (and replace) earlier openmpi-1.6-scyld rpms and will not update any existing (and now deprecated) openmpi-scyld rpms, which will likely be version 1.6.4. Scyld releases of OpenMPI are derived from http://www.open-mpi.org/. See the *User's Guide* Appendix C, *OpenMPI Release Information* for details.
3. The env-modules package now properly recognizes **module load** defaults that are declared in `.version` files found in the `/opt/scyld/modulefiles/` subdirectories.
4. Fix a rare deadlock of a master or compute node BProc I/O Daemon that can occur under very high workloads.
5. Suppress various redundant BProc syslog messages, e.g., a flurry of redundant ECONNREFUSED warnings.

New in Scyld ClusterWare 5.9.2 - Scyld Release 592g0000 - May 10, 2013

1. The base kernel is updated to 2.6.18-348.4.1. See https://rhn.redhat.com/errata/RHSA-2013-0747.html for details.

2. The MVAPICH2 release is updated to version 1.9.0, derived from http://mvapich.cse.ohio-state.edu/. See the *User's Guide* Appendix E, *MVAPICH2 Release Information* for details.
3. The mpich-scyld release is updated to version 3.0.4, derived from http://www.mpich.org/. See the *User's Guide* Appendix F, *MPICH-3 Release Information* for details.
4. A **/sbin/service beowulf start** (or **restart**) and **reload** now saves timestamped backups of various `/etc/beowulf/` configuration files, e.g., `config` and `fstab`, to assist a cluster administrator to recover a working configuration after an invalid edit.
5. Supports forwarding compute node log messages to an alternative **syslogd** server other than to the default master node server. See the *Administrator's Guide* for details.
6. Improve (again) the run-to-completion algorithm for determining if a compute node is effectively idle (and thus can be rebooted).

New in Scyld ClusterWare 5.9.1 - Scyld Release 591g0001 - March 27, 2013

1. The base kernel is updated to 2.6.18-348.3.1. See https://rhn.redhat.com/errata/RHSA-2013-0621.html for details.
2. TORQUE updates to version 4.2.2, from www.adaptivecomputing.com/resources/downloads/torque/. See the *User's Guide* Appendix B, *TORQUE Release Information* for details.

New in Scyld ClusterWare 5.9.1 - Scyld Release 591g0000 - March 21, 2013

1. The base kernel is updated to 2.6.18-348.2.1. See https://rhn.redhat.com/errata/RHSA-2013-0594.html for details.
2. TORQUE updates to version 4.2.1, from http://www.adaptivecomputing.com/resources/downloads/torque/. See the *User's Guide* Appendix B, *TORQUE Release Information* for details.
3. MVAPICH2 is updated to version 1.9b, derived from http://mvapich.cse.ohio-state.edu/. See the *User's Guide* Appendix E, *MVAPICH2 Release Information* for details.
4. OpenMPI is updated to version 1.6.4, derived from http://www.open-mpi.org/. See the *User's Guide* Appendix C, *Open-MPI Release Information* for details.
5. Includes the first release of mpich-scyld, which is the Scyld ClusterWare distribution of mpich-3, version 3.0.2, derived from http://www.mpich.org/. See the *User's Guide* Appendix F, *MPICH-3 Release Information* for details.

New in Scyld ClusterWare 5.9.0 - Scyld Release 590g0000 - February 11, 2013

1. The base kernel is updated to 2.6.18-348.1.1. See https://rhn.redhat.com/errata/RHBA-2013-0006.html and https://rhn.redhat.com/errata/RHSA-2013-0168.html for details.
2. The Scyld ClusterWare igb Ethernet driver is version 4.1.2, derived from http://sourceforge.net/projects/e1000/.
3. FFTW is version 3.3.2.
4. Improve the run-to-completion algorithm for determining if a compute node is effectively idle (and thus can be rebooted).

New in Scyld ClusterWare 5.8.5 - Scyld Release 585g0000 - January 14, 2013

1. The base kernel is updated to 2.6.18-308.24.1. See https://rhn.redhat.com/errata/RHSA-2012-1540.html for details.
2. The TORQUE release is updated to version 4.2.0, derived from http://www.adaptivecomputing.com/resources/downloads/torque/. See the *User's Guide* Appendix B, *TORQUE Release Information* for details.

New in Scyld ClusterWare 5.8.4 - Scyld Release 584g0001 - November 5, 2012

1. Fix a BProc problem that exhibits itself as the **bpmaster** daemon consuming 100% of a master node CPU.
2. The OpenMPI release is updated to version 1.6.3, derived from http://www.open-mpi.org/. See the *User's Guide* Appendix C, *OpenMPI Release Information* for details.

New in Scyld ClusterWare 5.8.4 - Scyld Release 584g0000 - November 1, 2012

1. The base kernel is updated to 2.6.18-308.16.1. See https://rhn.redhat.com/errata/RHSA-2012-1323.html for details.
2. The OpenMPI release is updated to version 1.6.2, derived from http://www.open-mpi.org/. See the *User's Guide* Appendix C, *OpenMPI Release Information* for details.
3. The MPICH2 release is updated to version 1.5, derived from http://www.mcs.anl.gov/research/projects/mpich2/. See the *User's Guide* Appendix D, *MPICH2 Release Information* for details.
4. The MVAPICH2 release is updated to version 1.8.1, derived from http://mvapich.cse.ohio-state.edu/. See the *User's Guide* Appendix E, *MVAPICH2 Release Information* for details. Since the Scyld ClusterWare 5 MVAPICH2 transport mechanism is *ssh*, the cluster administrator will likely need to configure the cluster to allow *ssh* access for non-root users. See the Section called *Optionally enable SSHD on compute nodes* and the *Administrator's Guide* for details.

New in Scyld ClusterWare 5.8.3 - Scyld Release 583g0000 - September 20, 2012

1. The base kernel is updated to 2.6.18-308.13.1. See https://rhn.redhat.com/errata/RHSA-2012-1174.html for details.
2. Fix a BProc problem that left a "lingering ghost" process on the master node that was not associated with any process on a compute node.
3. Support the *nonfatal* mount option for harddrive entries specified in `/etc/beowulf/fstab` to more gracefully handle clusters that have some nodes with harddrives and some nodes without, thus perhaps avoiding needing node-specific `/etc/beowulf/fstab.`*N* file(s).
4. The OpenMPI release is updated to version 1.6.1, derived from http://www.open-mpi.org/. See the *User's Guide* Appendix C, *OpenMPI Release Information* for details.

New in Scyld ClusterWare 5.8.2 - Scyld Release 582g0000 - August 14, 2012

1. The base kernel is updated to 2.6.18-308.11.1. See https://rhn.redhat.com/errata/RHSA-2012-0721.html and https://rhn.redhat.com/errata/RHSA-2012-1061.html for details.

2. The Scyld ClusterWare igb Ethernet driver is version 3.4.8, derived from http://sourceforge.net/projects/e1000/. Most noticeably, this newer driver eliminates the "HBO bit set" syslogged messages that were introduced by version 3.4.7.
3. The OpenMPI environment modules now define *MPI_HOME*, *MPI_LIB*, *MPI_INCLUDE*, and *MPI_SYSCONFIG*.
4. The file `/etc/beowulf/conf.d/sysctl.conf` now gets copied at boot time to every compute node as `/etc/sysctl.conf` to establish basic **/sbin/sysctl** values. The master node's `/etc/sysctl.conf` serves as the initial contents of `/etc/beowulf/conf.d/sysctl.conf`.
5. Replicate the master node's **ulimit** values for *filelimit* and *memlock* for ssh sessions on a compute node.

New in Scyld ClusterWare 5.8.1 - Scyld Release 581g0000 - June 22, 2012

1. The base kernel is updated to 2.6.18-308.8.1. See https://rhn.redhat.com/errata/RHSA-2012-0690.html for details.
2. The Scyld ClusterWare Adaptec aacraid driver is version 1.1.7-29100, useable for the 6805 controller, and is derived from http://www.adaptec.com/en-us/support/raid/sas_raid/sas-6805/.
3. Scyld ClusterWare MVAPICH2 now overrides the default MVAPICH2 CPU affinity management scheme. See the Section called *Scyld ClusterWare MVAPICH/MVAPICH2 CPU affinity management* for details.

New in Scyld ClusterWare 5.8.0 - Scyld Release 580g0000 - June 11, 2012

1. The base kernel is updated to 2.6.18-308.4.1. See https://rhn.redhat.com/errata/RHSA-2012-0150.html and https://rhn.redhat.com/errata/RHSA-2012-0480.html for details. This kernel requires a base distribution of RHEL5-U8 or CentOS 5.8.
2. The Scyld ClusterWare igb Ethernet driver is version 3.4.7, derived from http://sourceforge.net/projects/e1000/.
3. The Scyld ClusterWare Adaptec aacraid driver is version 1.1.7-28801, useable for the 6805 controller, and is derived from http://www.adaptec.com/en-us/support/raid/sas_raid/sas-6805/.
4. The TORQUE release is version 2.5.10, derived from http://www.adaptivecomputing.com/resources/downloads/torque/. See the *User's Guide* Appendix B, *TORQUE Release Information* for details.
5. The OpenMPI release is version 1.6, derived from http://www.open-mpi.org/. See the *User's Guide* Appendix C, *Open-MPI Release Information* for details.
6. The MPICH2 release is version 1.4.1p1, derived from http://www.mcs.anl.gov/research/projects/mpich2/. See the *User's Guide* Appendix D, *MPICH2 Release Information* for details.
7. The MVAPICH2 release is version 1.8, derived from http://mvapich.cse.ohio-state.edu/. See the *User's Guide* Appendix E, *MVAPICH2 Release Information* for details. Since the Scyld ClusterWare 5 MVAPICH2 transport mechanism is *ssh*, the cluster administrator will likely need to configure the cluster to allow *ssh* access for non-root users. See the Section called *Optionally enable SSHD on compute nodes* and the *Administrator's Guide* for details.

New in Scyld ClusterWare 5.7.3 - Scyld Release 573g0000 - December 16, 2011

1. The base kernel is updated to 2.6.18-275.12.1.el5.573g0000. See https://rhn.redhat.com/errata/RHSA-2011-1479.html for details.
2. Fixes a master node kernel panic caused by a lock deadlock.

3. Executing **/sbin/service beowulf start** or **restart** now verifies that the `/etc/beowulf/fstab`-specified NFS mounts which reference a $MASTER NFS server are currently exported by the master node; also, verifies that SELinux is disabled. A **start** or **restart** when the *beowulf* service is already running now exits with status 0, instead of with a nonzero status.

New in Scyld ClusterWare 5.7.2 - Scyld Release 572g0000 - November 21, 2011

1. The base kernel is updated to 2.6.18-275.7.1.el5.572g0000. See https://rhn.redhat.com/errata/RHSA-2011-1386.html for details.
2. TORQUE is updated to version 2.5.9. See the *User's Guide* Appendix B, *TORQUE Release Information*, for details.
3. OpenMPI is updated to version 1.5.4. See the *User's Guide* Appendix C, *OpenMPI Release Information*, for details.
4. MPICH2 is updated to version 1.4.1p1.
5. MVAPICH2 is updated to version 1.7.

New in Scyld ClusterWare 5.7.1 - Scyld Release 571g0000 - November 4, 2011

1. The base kernel is updated to 2.6.18-275.3.1.el5.571g0000. See https://rhn.redhat.com/errata/RHSA-2011-1212.html for details.
2. Enhances **bpsh**, **bpctl**, and **bpstat** to support a list of node name (or node number) ranges. Moreover, now a node name can be any name that is recognized by the Name Service, i.e., recognized by **getent hosts**.
3. Fixes a bug in the **insmod** on compute nodes of `/etc/beowulf/config` *bootmodule* kernel modules that have module dependencies that must be loaded in a particular order.
4. Enhances the **beoserv** daemon to support gPXE clients.
5. Fixes TORQUE so that it upgrades without altering the current contents of various configuration files, e.g., `/var/spool/torque/server_priv/nodes`.
6. Adds an NFS mount of `/usr/local/bin` to the default `/etc/beowulf/config`. Manually add this pathname to `/etc/exports` as appropriate for your local cluster.
7. The **/sbin/service beowulf start** now checks that each NFS mount in `/etc/beowulf/fstab` appears in `/etc/exports` and is actively exported.

New in Scyld ClusterWare 5.7.0 - Scyld Release 570g0000 - August 26, 2011

1. The base kernel is updated to 2.6.18-274.570g0000.el5. See https://rhn.redhat.com/errata/RHSA-2011-1065.html for details.
2. TORQUE is updated to version 2.5.7. See the *User's Guide* Appendix B, *TORQUE Release Information*, for details.

New in Scyld ClusterWare 5.6.3 - Scyld Release 563g0000 - August 12, 2011

1. The base kernel is updated to 2.6.18-238.19.1.563g0000.el5. See https://rhn.redhat.com/errata/RHSA-2011-0927.html for details.
2. The Scyld ClusterWare igb Ethernet driver is updated to version 3.1.16. This driver derives from source found at http://sourceforge.net/projects/e1000/files/. We recommend using this Scyld ClusterWare driver instead of the native RHEL5-U6 driver (version 2.1.0-k2-1).
3. TORQUE is updated to version 2.5.6. See the *User's Guide* Appendix B, *TORQUE Release Information*, for details.
4. Fixes a BProc bug that exhibited itself as an endless kernel *soft lockup* condition on a compute node, with a stack backtrace that involved *masq_wait*.
5. Further improvements to the **beosi** dump script.

New in Scyld ClusterWare 5.6.2 - Scyld Release 562g0000 - July 7, 2011

1. The base kernel is updated to 2.6.18-238.12.1.562g0000.el5. See https://rhn.redhat.com/errata/RHSA-2011-0833.html for details.
2. Fixes a BProc bug that exhibited itself as **gdb** hanging on a compute node shortly after entering a *run* directive.
3. Fix flaws in the **beosi** dump script, greatly reduce the size of the captured data for compute nodes, and enhance the capture of Infiniband statistics to include controllers other than mthca0.
4. Improves the time-of-day synchronization between the master node and the compute nodes by eliminating the syslog messages that announce each small time adjustment. Instead, each node's **bpslave** daemon now only syslogs a 24-hour summary of those small adjustments.
5. Fixes a file descriptor leak when doing *gethostbyname(".-1")* on the master node.
6. For TCP/IP over Infiniband (IPoIB), override the default *datagram* (or *Unreliable connected*) transport mode to instead use the optional *connected* mode, which supports substantially larger MTUs and other additional functionality for those peers that support *connected* mode.
7. **beoweb** is updated to version 1.5. This version enables SSL/TLS encryption for all traffic. New installations have SSL enabled by default, while old installations require modifications to the `/opt/scyld/beoweb/beoweb.ini` config file. Also new in version 1.5 is the capability to do recursive directory staging and improvements to group reporting support. Additional enhancements improve performance with the latest **podtools**.

New in Scyld ClusterWare 5.6.1 - Scyld Release 561g0000 - May 12, 2011

1. The base kernel is updated to 2.6.18-238.12.1.562g0000.el5. See https://rhn.redhat.com/errata/RHSA-2011-0429.html for details.
2. TORQUE is updated to version 2.5.5. In addition to bug fixes and feature enhancements that Adaptive Computing introduces in every TORQUE release (see the *User's Guide* Appendix B, *TORQUE Release Information*, for details), this Penguin Computing distribution incorporates a fix to a problem observed in version 2.5.3 that led to the withdrawal of that rpm from the Penguin Computing yum repo. See the Section called *Issues with TORQUE* for details.
3. Enhances beostat's **recvstats** and **beonss** to correctly recognize a compute node that is explicitly named in `/etc/hosts` to override the default hostname and aliases that otherwise would be be assigned to that IP address by beonss. Although

such overrides are discouraged because they are prone to error (e.g., using an incorrect IP address), they are now at least properly handled by Beowulf.

4. Improves the time-of-day (*TOD*) synchronization between the master node and the compute nodes by using the **adjtime** intrinsic to gradually close small time deltas, rather than always using **settimeofday** that makes abrupt changes which may cause a node's time to jump backwards (and forwards). Also, fixes flaws in the implementation of the `/etc/beowulf/config` *slavetimesync* directive, which allows the cluster administrator to turn off master-slave time synchronization in the event that the administrator is using another mechanism (e.g., **ntpd**) to sync the times.
5. Increases the BProc **bpmaster**- **bpslave** default *pingtimeout* value from 32 seconds to 64 seconds to avoid timeouts that may be triggered by transient network stalls.

New in Scyld ClusterWare 5.6.0 - Scyld Release 560g0000 - April 15, 2011

1. The base kernel is updated to 2.6.18-238.5.1.560g0002.el5. See https://rhn.redhat.com/errata/RHSA-2011-0017.html, https://rhn.redhat.com/errata/RHSA-2011-0163.html and https://rhn.redhat.com/errata/RHSA-2011-0303.html for details.
2. Introduces a Scyld ClusterWare distribution of the Adaptec **aacraid** driver, version 1.1.7-28000, that supports the new Adaptec 6805 controller. The native aacraid driver in the RHEL5u6 2.6.18-238.5.1 kernel does not yet support that controller.
3. Fixes various BProc **bpmaster** bugs that either caused a silent death of the daemon or a segfault, most commonly associated with the death or rebooting of a compute node.
4. Fixes a **beosi** bug which used **/sbin/sysctl** to revert the kernel tunable variables back to their default state, which might disable IP forwarding on the private cluster network.
5. Enhances the **beoserv** daemon to increase the max number of supported network interfaces per compute node from seven to 16.
6. OpenMPI is updated to version 1.5.3. See the Section called *Issues with OpenMPI* for general issues about supporting multiple concurrent versions. Existing applications that were built against version 1.4.3 or earlier must be rebuilt against this new version.
7. Includes a new **mpich2-scyld** package, which is a repackaging of the Open Source MPICH2 version 1.3.2 from http://www.mcs.anl.gov/research/projects/mpich2/, and a new **mvapich2-scyld** package, which is a repackaging of the Open Source MVAPICH2 version 1.6 from http://mvapich.cse.ohio-state.edu/. These Scyld ClusterWare distributions employs environment modules to manage building and linking applications to a specific compiler family, plus package-specific manpages. Use **module avail** to see the available module choices. Note: the **mpirun** command syntax differs from the mpirun used for **mpich-1.2.7p1** and **mvapich-scyld-0.9.9**. Users are encouraged to load the appropriate environment module, then use **man mpirun** to review the syntax.

New in Scyld ClusterWare 5.5.2 Update - Scyld Release 552g0004 - March 1, 2011

1. The base kernel is updated to 2.6.18-194.32.1.el5.552g0003. The base distribution components are the same as the earlier Scyld Release 552g0002 kernel, 2.6.18-194.32.1.el.552g0002, but with a change to the BProc modifications that are applied to the kernel source code that necessitated a rebuild of the kernel.
2. Fixes a BProc bug that exhibited itself as a multithreaded application (e.g., Fluent, Java), executing on a compute node, hanging during exit. This fix necessitated the release of a modified kernel, noted above.

3. Fixes **beorun** bug that incorrectly rejected as invalid a *--map* node list which included *-1*, the master node.
4. Fixes a **beoserv** segfault that occurred when booting a new node when the `/etc/beowulf/config` file uses *node* entries with no additional arguments. This bug was introduced in Scyld Release 552g0002.
5. The latest TORQUE version 2.5.3 (introduced in Scyld Release 552g0000) has seemingly introduced various problems running TORQUE jobs, and it is being withdrawn. See the Section called *Issues with TORQUE* for details.

New in Scyld ClusterWare 5.5.2 Update - Scyld Release 552g0003 - January 28, 2011

1. Fixes a BProc **bpmaster** bug that exhibited itself as a verbose stream of syslog messages of the form *EPOLLHUP, not CONN_DEAD*.

New in Scyld ClusterWare 5.5.2 Update - Scyld Release 552g0002 - January 20, 2011

1. The base kernel is updated to 2.6.18-194.32.1 (552g0002). See https://rhn.redhat.com/errata/RHSA-2011-0004.html for details.
2. The Scyld ClusterWare igb Ethernet driver is updated to version 2.4.12. This driver derives from source found at http://sourceforge.net/projects/e1000/files/. We recommend using this Scyld ClusterWare driver instead of the native RHEL5-U5 driver (version 2.1.0-k2).
3. Fixes an **rcmdd** security flaw which permitted a non-root user to gain root access using **rsh** to a compute node.
4. When booting a cluster with ipforwarding enabled, Scyld ClusterWare silently increases the ip_conntrack max table size to 524,288 to try to avoid *ip_conntrack: table full, dropping packet* syslog messages. See the Section called *Optionally increase the ip_conntrack table size* for details.
5. OpenMPI is updated to version 1.5.1. This release yet again restructures the locations of the compiler-specific libraries, executable binaries, manpages, and environment modules, but now each new release of OpenMPI can gracefully coexist with earlier releases, and existing applications that were built against an earlier version do not need to be immediately rebuilt against this new version. See the Section called *Issues with OpenMPI* for general issues about supporting multiple concurrent versions. Existing applications that were built against version 1.4.3 or earlier must be rebuilt against this new version.

New in Scyld ClusterWare 5.5.2 Update - Scyld Release 552g0001 - December 1, 2010

1. Fixes a BProc **bpmaster** bug that exhibited itself as the bpmaster daemon consuming 100% of a master node CPU, which paralyzed the cluster and drove the kernel into a *soft lockup* condition.
2. Fixes a BProc bug that exhibited itself as a kernel *soft lockup* condition that was reported on a compute node's console as the **bpslave** daemon executing *__write_lock_failed*.

New in Scyld ClusterWare 5.5.2 - Scyld Release 552g0000

1. The base kernel is updated to 2.6.18-194.26.1 (552g0001). See https://rhn.redhat.com/errata/RHSA-2010-0723.html, https://rhn.redhat.com/errata/RHSA-2010-0792.html and https://rhn.redhat.com/errata/RHSA-2010-0839.html for de-

tails.

2. The Scyld ClusterWare igb Ethernet driver is updated to version 2.3.4. This driver derives from source found at http://sourceforge.net/projects/e1000/files/. We recommend using this Scyld ClusterWare driver instead of the native RHEL5-U5 driver (version 2.1.0-k2).
3. Introduces **beoweb**, a web server that runs on a cluster's master node to facilitate remote job submission and cluster monitoring. Beoweb is an optional service (distributed as **/sbin/chkconfig beoweb off**) that must be enabled and started prior to use.
4. Introduces **POD Tools**, which contains a command-line interface called **POD Shell (podsh)**, that can be installed on Scyld and non-Scyld systems. POD Shell interfaces with the new beoweb service to provide for remote job submission and monitoring. See the *User's Guide* for details.
5. Introduces **python-scyld**, which is derived from Open Source Python version 2.6.5, and **pylons-scyld**, both of which provide a foundation framework for beoweb and POD Tools.
6. Introduces **net-snmp-scyld**. The Open Source **net-snmp** project includes various SNMP (Simple Network Management Protocol) tools: an extensible agent, an SNMP library, tools for requesting or setting information from SNMP agents, tools for generating and handling SNMP traps, a version of the **netstat** command which uses SNMP, and a Tk/Perl MIB browser. This package also contains the **snmpd** and **snmptrapd** daemons and documentation. The new net-snmp-scyld package is net-snmp with Scyld MIB module extensions built into the daemon. The Scyld MIB module implements Scyld ClusterWare node monitoring of CPU, memory, and disk usages; the enabling/disabling of memory and disk usage traps; and getting/setting trap thresholds. See the *Administrator's Guide* for details.
7. Enhances the **beostatus** command to support remote master node monitoring (utilizing beoweb on the remote node) and various options to filter the displayed information. See **man beostatus** and the *Administrator's Guide* and *User's Guide* for details.
8. Improves the performance of **beonss** *kickback* name resolution from compute nodes.
9. Fixes a rare failure to PXEboot nodes that employ the igb Gigabit Ethernet driver.
10. Fixes the Scyld ClusterWare **Ganglia** "proc_run" graph that showed an incorrect and excessively large number of running processes.
11. Enhances **beoserv** DHCP to better handle non-Scyld non-Linux compute nodes that request a DNS IP address.
12. OpenMPI is updated to version 1.5. This release restructures the locations of the compiler-specific libraries, binaries, and manpages, and changes the contents of the environment modules. Existing applications that were built against version 1.4.3 or earlier must be rebuilt against this new version.
13. TORQUE is updated to version 2.5.3.
14. Taskmaster version 5.4.1 is now available (under a separate license).

New in Scyld ClusterWare 5.5.0 Update - Scyld Release 550g0007 - October 11, 2010

1. The base kernel is updated to 2.6.18-194.11.4 (550g0005). See https://rhn.redhat.com/errata/RHSA-2010-0704.html for details.
2. Fixes a **bpcp -p** bug where the *mode* was not properly preserved across the copy.
3. Fixes a Scyld ClusterWare **Ganglia** bug where the network bytes/second data rates were being misreported.
4. OpenMPI is updated to version 1.4.3.

New in Scyld ClusterWare 5.5.0 Update - Scyld Release 550g0006

1. The base kernel is updated to 2.6.18-194.11.3. See https://rhn.redhat.com/errata/RHSA-2010-0661.html for details.
2. The igb network driver has integrated an upstream fix to more frequently update `/proc/net/dev` statistics, which means that **beostat** and **IMF** more accurately report network usage for chipsets that use that driver.
3. **beostatus** no longer requires that TORQUE be installed.

New in Scyld ClusterWare 5.5.0 Update - Scyld Release 550g0005

1. The base kernel is updated to 2.6.18-194.11.1. See https://rhn.redhat.com/errata/RHSA-2010-0610.html for details.

New in Scyld ClusterWare 5.5.0 Update - Scyld Release 550g0004

1. The base kernel is updated to 2.6.18-194.8.1. See https://rhn.redhat.com/errata/RHSA-2010-0504.html for details.
2. The Scyld ClusterWare kernel supports *perfctr*, which supports access to Intel and AMD processor performance counters. This functionality is contributed by the Performance Application Programming Interface (*PAPI*) project, version 4.1.0. See http://icl.cs.utk.edu/papi/index.html for details.
3. Scyld ClusterWare now includes an updated Gigabit Ethernet igb driver (version 2.2.9) that fixes various problems seen with the Scyld ClusterWare driver (version 1.3.28.5) that was first introduced in CW5.4.1 as an improvement over the native RHEL5-U4 driver (version 1.3.16-k2). We recommend using this newest Scyld ClusterWare driver over the native RHEL5-U5 driver (version 2.1.0-k2).
4. Fixes a problem that exhibits itself as a compute node needing an excessively long time to reboot (e.g., 15 minutes, vs. the more common two minutes, approximately).
5. Avoids the most common port number conflicts (**beoserv**'s *beofs2/tcp* port and BProc's *bproc* port) by starting with the default port numbers (possibly overridden by `config` file *server* directives), and flexibly incrementing these port numbers as needed to find an available port. See the Section called *Issues with port numbers* for details.
6. Fixes a **bpmaster** daemon segfault that occasionally occurs when performing a concurrent reboot (e.g., **/usr/sbin/bpctl -S all -R**) of a large number of nodes.

New in Scyld ClusterWare 5.5.0 Update - Scyld Release 550g0003

1. Fixes a BProc bug that caused a TORQUE fork failure, reported in a syslog message "pbs_mom: LOG_ERROR::Resource temporarily unavailable (11) in fork_me, fork failed" and leaving lingering TORQUE jobs.
2. The cluster administrator may restrict compute node access to the master node, in much the same way as an admin can assign access permissions to individual compute nodes. For example, **/usr/sbin/bpctl -M -m 0110** disallows process migrations from a compute node to the master, including migrations using **bpsh** and **bpcp**. Additionally, a new `config` file keyword, *nodeaccess*, provides the ability to make these master node and compute node access restrictions persistent across cluster reboots. See the `config` file comments and the *Administrator's Guide* for details.
3. **/usr/bin/bpcp -p** now replicates the source file's UID and GID for the target file. Previously, even when using the **-p** option, the target file was owned by root.

4. **/usr/bin/bpcp** now guarantees that the target file exists when **bpcp** exits. Previously, **bpcp** may have exited with a successful status before the target was created.

New in Scyld ClusterWare 5.5.0 Update - Scyld Release 550g0002

1. The base kernel is updated to 2.6.18-194.3.1. See https://rhn.redhat.com/errata/RHSA-2010-0398.html for details.
2. The Scyld ClusterWare kernel now supports *perfctr*, which supports access to Intel and AMD processor performance counters. This functionality is contributed by the Performance Application Programming Interface (*PAPI*) project, version 4.0.0. See http://icl.cs.utk.edu/papi/index.html for details.

New in Scyld ClusterWare 5.5.0 Update - Scyld Release 550g0001

1. Fixes a bug where doing a ctrl-c or a kill of certain workloads might leave a "lingering ghost" process on the master node: a process that was associated with the real process that had been executing on a compute node and which was properly terminated by the ctrl-c or kill. Additionally, previously a "lingering ghost" process could not be manually killed, and it would only get cleaned up when the cluster rebooted. Now "lingering ghosts" should not appear. If any does appear, it can now be killed using **/bin/kill** or **/usr/bin/killall**, as appropriate.
2. Fixes an infrequent BProc bug which exhibited itself most commonly as a kernel panic due to a segfault in *ghost_put* or to a "Kernel BUG at spinlock:119" called from *bproc_purge_requests*.

New in Scyld ClusterWare 5.5.0 - Scyld Release 550g0000 - April 15, 2010

1. The base kernel is updated to 2.6.18-194. See https://rhn.redhat.com/errata/RHSA-2010-0178.html for details.
2. The *beostat* service that supplies cluster performance statistics to **Scyld IMF**, **beostatus**, **Ganglia** and other cluster status visualization utilities now understands bonded network devices. Previously, network statistics were double-reported: counting both the aggregated bonded pseudo-device and the individual devices that comprise the bonded device.
3. Eliminates a bogus **recvstats** syslog message of the form "Received stats from IP addr" that occasionally appeared as a compute node starts up.
4. Eliminates a bogus BProc syslog message "proc.exe not null".

New in Scyld ClusterWare 5.4.1

1. The base kernel is updated to 2.6.18-164.15.1. See https://rhn.redhat.com/errata/RHSA-2009-1548.html, https://rhn.redhat.com/errata/RHSA-2009-1670.html, https://rhn.redhat.com/errata/RHSA-2010-0019.html, https://rhn.redhat.com/errata/RHSA-2010-0046.html, and https://rhn.redhat.com/errata/RHSA-2010-0147.html for details.
2. Scyld ClusterWare includes a Gigabit Ethernet igb driver (version 1.3.28.5) that fixes various problems seen with the RHEL5-U4 igb driver (version 1.3.16-k2).

3. Scyld ClusterWare 5.4.1 no longer includes an add-on forcedeth driver (version 0.61) that was distributed in Scyld ClusterWare 5.4.0. The CW5.4.1 kernel now contains a superior RHEL5-U4 forcedeth driver (version 0.62) that supports the **/sbin/vconfig** command to configure VLAN devices.
4. Scyld ClusterWare now supports non-Scyld nodes as compute nodes in the cluster, in addition to the traditional Scyld nodes that integrate into the Scyld unified process management environment. An example of a non-Scyld compute node is a server that executes a full distribution of Red Hat Enterprise Linux (RHEL) or CentOS and which boots from a local harddrive. See the *Administrator's Guide* for details.
5. Supports two new `/etc/beowulf/config` keywords, *host* and *hostrange*. The `config` file may contain zero or more of each. A *host* directive pairs a unique client MAC address with the unique IP address to be delivered to that client, together with an optional name for the client, for use if and when that client makes a DHCP request to the master node. A *hostrange* directive specifies a unique range of IP addresses that does not collide with the *iprange* addresses used for cluster compute nodes, nor with the IP address(es) used for master node(s). Every *host* IP address must fall within one of the *hostrange* ranges. These clients are typically some device or node on the cluster private network other than a compute node, such as a managed switch or some other device that uses DHCP to obtain an IP address. See the *Administrator's Guide* for details.
6. Fixes a bug in BProc where certain workloads would cause a master node kernel panic, most commonly a segfault in the routine *ghost_put*.
7. Fixes a bug in node startup which ignored a fatal mount failure and allowed the node to transition to the *up* state. Proper behavior is to abort the node startup and to leave the node in *error* state.
8. Fixes a bug where certain workloads would generate many thousands of sockets sitting in TIME_WAIT limbo, which is at best inefficient and at worst would lead to temporary socket exhaustion.
9. Fixes a bug where cluster startup would leave temporary files in `/tmp/`. These are now properly deleted.
10. The **beoserv** tftp server, which executes on the master node, now only listens on the private cluster interface. Previously it listened on all interfaces for tftp requests. Additionally, previously tftp requests only retrieved files that resided in the `/var/beowulf/boot/` directory. Now it treats a requested filename as being a pathname relative to that base directory, i.e., the file may reside in a subdirectory of `/var/beowulf/boot/`.
11. The **beoserv** daemon now automatically removes duplicate MAC addresses from file `/etc/beowulf/unknown_addresses`.
12. The master node's `/etc/ofed/dat.conf` is now copied to each node as `/etc/dat.conf` where various MPI implementations (e.g., HP-MPI included with Fluent 12 and some versions of Intel MPI) expect to find it.
13. The `/etc/beowulf/config` *prestage* directive now supports prestaging any master node file to compute nodes at cluster startup. Previously, *prestage* was limited to files that reside in one of the *libraries* directories.
14. Introduces cleaner support for the Infiniband RDMA Protocol (*SRP*) with a new startup script, `/etc/beowulf/init.d/20srp`. To use SRP, you must install the optional **srptools** rpm from the base distribution, enable the `20srp` script (e.g., using **/sbin/beochkconfig**), and reboot the cluster nodes.
15. OpenMPI is updated to version 1.4.1.
16. TORQUE is updated to version 2.3.10.
17. When starting Beowulf services (`/etc/init.d/beowulf`), Scyld ClusterWare now automatically increases some system resource parameters to better handle the demands of small- to medium-sized clusters:
    - Increase the number of available pids to a minimum of 98,304. See the Section called *Optionally increase the max number of processID values* for more information.
    - Increase the number of **nfsd** threads to at least one thread per compute node, with a lowerbound of eight (the Red Hat default) and an upperbound of 64. See the Section called *Optionally increase the number of nfsd threads* for more

information.

- Increase the ARP cache capacity from the default threshold values of 128, 512, and 1024 to new values of 512, 2048, and 4096, respectively, and increase the gc_interval from 30 seconds to 240 seconds. See **man 7 arp** for more details.

New in Scyld ClusterWare 5.4.0 - Scyld Release 540g0000 - October 23, 2009

1. The initial CW5.4.0 release included a kernel that was based upon RHEL5 2.6.18-164.2.1. The current CW5.4.0 yum repository contains a newer kernel that is based upon RHEL5 2.6.18-164.9.1. See https://rhn.redhat.com/errata/RHSA-2009-1243.html and https://rhn.redhat.com/errata/RHSA-2009-1455.html for details about the Red Hat kernel changes between CW5.3.0 and 2.6.18-164.2.1. See https://rhn.redhat.com/errata/RHSA-2009-1548.html and https://rhn.redhat.com/errata/RHSA-2009-1670.html for details about subsequent changes through 2.6.18-164.9.1. This kernel is compatible with a base distribution of RHEL5-U4 or CentOS 5.4.
2. The initial CW5.4.0 release included the **Scyld Integrated Management Framework (IMF)** with some enhancements that were only available as separately licensed modules, versus the unrestricted full Scyld IMF that was bundled into CW5.3.0 and called **ClusterAdmin**. The latest CW5.4.0 yum repository once again contains the fully functional Scyld IMF and is distributed under the Scyld ClusterWare license.
3. Scyld ClusterWare includes the **env-modules** environment-modules package, which enables the dynamic modification of a user's environment via modulefiles. Each modulefile contains the information needed to configure the shell for an application, allowing a user to easily switch between applications with a simple **module switch** command that resets environment variables like PATH and LD_LIBRARY_PATH. A number of modules are already installed configuring application builds and execution with OpenMPI, including jobs submitted through TORQUE. For more information on these modules, see the *Programmer's Guide* for details. For more information about creating your own modules, see http://modules.sourceforge.net, or view the manpages **man module** and **man modulefile**.
4. Scyld ClusterWare now includes **pacct**, a utility to generate simple reports from the verbose TORQUE log files. There are two types of log files: the *event log*, which record events from each TORQUE daemon, and the *accounting logs*. The accounting log files reside by default in the `/var/spool/torque/server_priv/accounting/` directory. See http://www.nacad.ufrj.br/~bino/pbs_acct-e.html for more information about this tool. Note: the Scyld ClusterWare version of **pacct** reports total core hours, rather than total node hours.
5. The **mvapich** package has been renamed to **mvapich-scyld** in order to better distinguish it from the RHEL5 **mvapich** package that carries a newer version number. Installing or upgrading Scyld ClusterWare will install **mvapich-scyld** and silently remove the base distribution's **mvapich** package.
6. The Pathscale compiler is no longer supported. Accordingly, the **mpich**, **mvapich-scyld**, and **openmpi-scyld** packages no longer include Pathscale libraries that previously resided in `/usr/lib64/MPICH/p4/path/`, `/usr/lib64/MPICH/vapi/path/`, and `/usr/openmpi/path/share/`, respectively.
7. The **mpich** and **mvapich-scyld** libraries now explicitly limit an application to a maximum of 1000 threads. This is not a reduction of a previous capability; it is, in fact, a bounds check that recognizes and enforces an existing limitation in the implementation.
8. In some instances, **mpirun -machine vapi** was not properly linking the application to the MVAPICH libraries on a compute node, and was instead mistakenly linking with the default gnu p4 (Ethernet) libraries. This has now been fixed, in part by replicating the master node's `/usr/lib64/MPICH/` directory structure on each compute node at node startup. The libraries themselves are only pulled to a compute node if and when they are actually needed.
9. Fixes a bug with **MVAPICH** (Infiniband) applications which improperly left lingering application threads running after the application was supposedly killed by a TORQUE **qdel**, or after some, but not all, the application's threads died

because they were explicitly killed (e.g., using **/usr/bin/kill**) or abnormally terminated (e.g., with a segmentation violation).

10. The **beonss** name space functionality has improved robustness, error reporting via the syslog, and a modest performance improvement for compute-to-master *kickback* communication.
11. **bpsh** (and process migration, in general) now communicates the current **umask** specification to the compute nodes. Previously, the **umask** was ignored, and files created on a compute node defaulted to world-writeable permissions.
12. Scyld ClusterWare includes a Gigabit Ethernet igb driver (version 1.3.28.5) that fixes various problems seen with the RHEL5-U4 igb driver (version 1.3.16-k2).
13. Scyld ClusterWare 5.4.0 no longer includes an add-on tg3 driver (version 3.86, from RHEL5-U2) that was distributed in Scyld ClusterWare 5.3.0. The CW5.4.0 kernel now contains a fixed RHEL5-U4 tg3 driver (version 3.96-1).
14. OpenMPI is updated to version 1.3.3.
15. TORQUE is updated to version 2.3.7.
16. Scyld ClusterWare's default port numbers can now be overridden using the *server* directive in `/etc/beowulf/config`. See the Section called *Issues with port numbers* for details.
17. Scyld ClusterWare's **beoserv** daemon now responds to any DHCP request that arrives on the cluster private network. Previously, **beoserv** only functioned as a DHCP server for Scyld nodes.

New in Scyld ClusterWare 5.3.0 - Scyld Release 530g0000 - April 9, 2009

1. The base kernel is updated to 2.6.18-128.1.1. See https://rhn.redhat.com/errata/RHSA-2009-0225.html and https://rhn.redhat.com/errata/RHSA-2009-0264.html for details. This kernel requires a base distribution of RHEL5-U3 or CentOS 5.3.
2. Includes the RHEL5-U2 tg3 driver (version 3.86) to supersede the flawed RHEL5-U3 tg3 (version 3.93) driver that occasionally misbehaves for Broadcom NetXtreme or NetLink controllers. The misbehavior consists of initially linking at the expected 1000 Mbps/Full, then disconnecting and relinking at a much slower 10 Mbps/Half.
3. Includes the new **Scyld Integrated Management Framework (IMF)** cluster monitoring tool. See the Section called *Optionally enable Scyld Integrated Management Framework (IMF)* and the *Administrator's Guide* for details.
4. Includes a simplified method to enable cluster-wide NFS locking. See the Section called *Optionally enable NFS locking* and the *Administrator's Guide* for details.
5. Includes compiler-specific FFTW libraries in `/usr/lib64/FFTW/`.
6. OpenMPI is updated to 1.2.9. This new package also includes per-compiler built binaries, which required relocating some OpenMPI files in order to accommodate this enhancement. Previously, there was only one set of OpenMPI binaries available, regardless of the compiler toolchain used, located in `/usr/openmpi/bin/`. Now the binaries reside in `/usr/openmpi/bin/`*compilerName*`/`, consistent with how they were built. The OpenMPI include files have also been moved from their old location of `/usr/openmpi/include` into the new location of `/usr/include/openmpi/`. The library locations have not changed.
7. TORQUE is updated to version 2.3.6.
8. Ganglia is updated to version 3.0.7.
9. Fixes a bug in the Scyld ClusterWare Name Service which leaked a file descriptor in a software thread's compute node environment for each Name Service request, leading to the thread being stuck in a loop, consuming 100% of its CPU and making no forward progress. This was observed with the TORQUE **pbs_mom** daemon and in any other user program which issues more than 1024 Name Service requests.

10. Fixes a bug where if the Sun Grid Engine (SGE) was configured with an *admin_user* other than root, then **/usr/bin/qdel** could not delete jobs that run on a compute node.
11. Fixes a shortcoming in Ganglia: it was not displaying pie chart node metrics.
12. Fixes a bug where BProc gets confused about a process' true node residency, incorrectly believing the process is executing on the master node, even though the process is in fact correctly executing on the intended compute node. This bug is due to a timing window and tends to occur (if at all) in circumstances of high rates of process creation and/or destruction. The effect of this bug is that these "bad bookkeeping" processes are outside BProc's unified process space and are thus immune to normal process management performed on the master node, e.g., **/bin/kill** or *signal()*.

Known Issues And Workarounds

The following are known issues of significance with the latest version of Scyld ClusterWare 5.11.2 and suggested workarounds.

Issues with TORQUE

Scyld ClusterWare repackages the TORQUE resource manager available from Adaptive Computing, http://www.adaptivecomputing.com/support/download-center/torque-download/. TORQUE version 2.5.3 was introduced in Scyld Release 552g0000 as **torque-2.5.3**, which was an update the previous **torque-2.3.10**. Version 2.5.3 introduced various problems running TORQUE jobs and was subsequently withdrawn in Scyld Release 552g0004. In particular, TORQUE 2.5.3 rejects the syntax of some previously-acceptable Matlab jobs. Scyld ClusterWare currently distributes **torque-2.5.5**, which fixes the problems running Matlab. In every new TORQUE release, the Adaptive Computing developers fix bugs, add new features, and on occasion change configuration and scripting options. View the Adaptive Computing's TORQUE Release Notes and Changelog in the Scyld ClusterWare *User's Guide Appendix B. TORQUE Release Information*.

If you believe that a newly updated TORQUE has introduced new TORQUE problems, then you may choose to revert to an earlier version:

```
rpm -e --nodeps torque-2.5.5
yum install torque --exclude=torque-2.5.5
```

As always, after installing a different TORQUE, you must restart torque and restart the cluster:

```
/sbin/service torque restart
/sbin/service beowulf restart
```

Issues with IP Forwarding

If the *beowulf* service has started and IP forwarding is in use, then a subsequent **/sbin/service iptables stop** (or **restart**) will hang because it attempts to unload the ipt_MASQUERADE kernel module while the *beowulf* service is using (and not releasing) that module. For a workaround, edit `/etc/sysconfig/iptables-config` to change:

```
    IPTABLES_MODULES_UNLOAD="yes"
to: IPTABLES_MODULES_UNLOAD="no"
```

Issues with kernel modules

The **/sbin/modprobe** command uses `/usr/lib/`\`uname -r\`/`modules.dep` to determine the pathnames of the specified kernel module and that module's dependencies. The **/sbin/depmod** command builds the human-readable `modules.dep` file, and it should be executed *on the master node* after installing any new kernel module.

Executing **modprobe** on a compute node requires additional caution. The first use of **modprobe** retrieves the current `modules.dep` from the master node using bproc's *filecache* functionality. Since any subsequent **depmod** on the master node rebuilds `modules.dep`, then a subsequent **modprobe** on a compute node will only see the new `modules.dep` if that file is copied to the node using **bpcp**, or if the node is rebooted and thereby silently retrieves the new file.

In general, you should not execute **depmod** on a compute node, since that command will only see those few kernel modules that have previously been retrieved from the master node, which means the node's newly built `modules.dep` will only be a sparse subset of the master node's full `module.dep`. Bproc's *filecache* functionality will always properly retrieve a kernel module from the master node, as long as the node's `module.dep` properly specifies the pathname of that module, so the key is to have the node's `module.dep` be a current copy of the master's file.

Issues with ptrace

Cluster-wide **ptrace** functionality is not yet supported in ClusterWare 5. For example, you cannot use a debugger running on the master node to observe or manipulate a process that is executing on a compute node, e.g., using **gdb -p procID**, where *procID* is a processID of a compute node process. **strace** does function in its basic form, although you cannot use the **-f** or **-F** options to trace forked children if those children move away from the parent's node.

Issues with xpvm

xpvm is not currently supported in ClusterWare 5.

Caution using beosetup

The **/usr/sbin/beosetup** utility is deprecated. At this time, we do not recommend using **beosetup** for observing or altering the cluster state while new compute nodes are booting.

Issues with port numbers

Scyld ClusterWare employs several daemons that execute in cooperating pairs: a server daemon that executes on the master node, and a client daemon that executes on compute nodes. Each daemon pair communicates using tcp or udp through a presumably unique port number. By default, Scyld ClusterWare uses ports 932 (*beofs2*), 933 (*bproc*), 3045 (*beonss*), and 5545 (*beostats*). In the event that one or more of these port numbers collides with a non-Scyld ClusterWare daemon using the same port number, the cluster administrator can override Scyld ClusterWare default port numbers to use different, non-colliding unused ports using the `/etc/beowulf/config` file's *server* directive. See **man beowulf-config** and `/etc/beowulf/config` for a discussion of the *server* directive.

The official list of assigned ports and their associated services is http://www.iana.org/assignments/port-numbers, and `/etc/services` is a list shipped with your base distribution. However, the absence in either list of a specific port number is no guarantee that the port will not be used by some software on your cluster. Use **lsof -i :***portNumber* to determine if a particular port number is in active use.

A common collision is with *beofs2* port 932 or *bproc* port 933, since the **rpc.statd** or **rpc.mountd** daemons may randomly grab either of those ports before Beowulf can grab them. However, Beowulf recognizes the conflict and tries al-

ternative ports until it finds an unused port. If this flexible search causes problems with other daemons, you can edit `/etc/beowulf/config` to specify a tentative override value using the *server beofs2* or *server bproc* directive, as appropriate.

Less common are collisions with *beonss* port 3045 or *beostats* port 5545. The *server beonss* and *server beostats* override values are used as-specified and not adjusted by Beowulf at runtime.

Issues with OpenMPI

Scyld ClusterWare distributes a repackaged release of the Open Source OpenMPI, derived from http://www.open-mpi.org/. The Scyld ClusterWare distribution consists of the `openmpi-`*x.y*`-scyld` base package for the latest OpenMPI version *x.y.z*, plus plus several compiler-environment-specific packages for `gnu`, `intel`, and `pgi`. For example, the distribution of OpenMPI version 1.7.1 consists of the base rpm `openmpi-1.7-scyld-1.7.1` and the various compiler-specific rpms: `openmpi-1.7-scyld-gnu-1.7.1`, `openmpi-1.7-scyld-intel-1.7.1`, and `openmpi-1.7-scyld-pgi-1.7.1`.

By default, **yum groupupdate Scyld-ClusterWare** only updates `openmpi-`*x.y*`-scyld-`* packages with a newer *x.y.z* version. For example, `openmpi-1.7-scyld-1.7.2` updates `openmpi-1.7-scyld-1.7.1`, but it does not affect any installed `openmpi-1.6-scyld` or `openmpi-1.5-scyld` packages.

Scyld ClusterWare installs the files into *x.y.z* version-specific directories, which allows multiple versions to co-exist on the master node. Each `/opt/scyld/openmpi/`*version* directory contains *compiler* subdirectories `gnu`, `intel`, and `pgi`, each of which contain libraries, executable binaries, and manpages associated with that particular compiler. The directory `/opt/scyld/openmpi/`*version*`/examples` contains source code examples.

The modulefiles have pathnames `/opt/scyld/modulefiles/openmpi/`*compiler*`/`*version*, where *version* is a file that amends $PATH, $LD_LIBRARY_PATH, and $MANPATH with pathnames that point into the associated compiler-specific `/opt/scyld/openmpi/`*version/compiler/* subdirectories.

Many customers support multiple OpenMPI versions because some applications may only work properly when linked to specific OpenMPI versions. Sometimes an application needs only to be recompiled and relinked against a newer version of the libraries. Other applications may have a dependency upon a particular OpenMPI version that a simple recompilation won't fix. The cluster administrator can specify which compiler and version is the default by manipulating the contents of the various `.version` files in the `/opt/scyld/modulefiles/openmpi/` subdirectories. For example, a default **module load openmpi** could reference version 1.6.4 of the gnu libraries, while a version-specific **module load openmpi/gnu/1.7.1** or **module load openmpi/intel/1.7.1** references version 1.7.1 of the gnu or intel libraries.

Scyld ClusterWare supports several OpenMPI versions by default: the latest Open Source releases of versions 1.5.x, 1.6.x, 1.7.x, and 1.8.x If the cluster administrator wishes to retain additional *x.y.z* releases within a *x.y* family, then instead of doing a default:

```
yum groupupdate Scyld-ClusterWare
```

the administrator should do:

```
yum groupupdate Scyld-ClusterWare --exclude=openmpi*scyld-*
```

and then download specific rpms from the yum repo as desired, then install (not update) them manually. For example:

```
mkdir -p /tmp/ompinew
yumdownloader --destdir /tmp/ompinew openmpi-1.5-scyld-*1.5.5*
rpm -iv /tmp/ompinew/openmpi-1.5-scyld-*1.5.5*
```

Note the use of **yumdownloader** and **rpm -i**. This is necessary because doing **yum install openmpi-1.5-scyld-*1.5.5*** will not, in fact, execute a simple *install* and retain older packages. Rather, it actually executes an *update* and removes any and all older installed versions of `openmpi-1.5-scyld-*` rpms.

Prior to Scyld ClusterWare 5.9.3, the OpenMPI packages were distributed as `openmpi-scyld-*`, and by default the **yum groupupdate Scyld-ClusterWare** always updated to the latest version *x.y.z* and removed all older versions. The cluster administrator can do nothing, and the master node will continue to retain any installed `openmpi-scyld` distribution(s), or the administrator can choose to convert to the new `openmpi-`*x.y*`-scyld` naming scheme by removing the older distribution, then installing the newer equivalent distribution:

```
yum remove openmpi-scyld-*
yum install openmpi-1.6-scyld-*
```

Issues with Spanning Tree Protocol and portfast

Network switches with Spanning Tree Protocol (STP) enabled will block packets received on a port for the first 30 seconds after the port comes online, giving the switch and the Spanning Tree algorithm time to determine if the device on the new link is a switch, and to determine if Spanning Tree will block or forward packets from this port. This is done to prevent "loops" which can cause packets to be endlessly repeated at a high rate and consume all network bandwidth. Each time the link goes down and comes back up, another 30-second blocking delay occurs. This delay can prevent PXE/DHCP from obtaining an IP address, or can prevent the node's initial kernel from downloading its initial root filesystem, which results in the node endlessly iterating in the early boot sequence, or can delay the node's ongoing *filecache* provisioning of libraries to the node.

We recommend disabling STP if feasible. If not feasible, then we recommend reconfiguring the switch to use *Rapid STP* or *portfast*, which avoids the 30-second delay, or employing some other port mode that will forward packets as a port comes up. There is no generic procedure for enabling these options. For Cisco switches, see http://www.cisco.com/en/US/products/hw/switches/ps700/products_tech_note09186a00800b1500.shtml. For other switch models, see the model-specific documentation.

If that reconfiguration is also not possible, you may need to increase the default Scyld ClusterWare timeout used by the node to a value safely greater than the STP delay: e.g., add *rootfs_timeout=120 getfile_timeout=120* to the `/etc/beowulf/config` *kernelcommandline* entry to increase the timeouts to 120 seconds.

Issues with Gdk

If you access a cluster master node using **ssh -X** from a workstation, some graphical commands or program may fail with:

```
Gdk-ERROR **: BadMatch (invalid parameter attributes)
  serial 798 error_code 8 request_code 72 minor_code 0
Gdk-ERROR **: BadMatch (invalid parameter attributes)
  serial 802 error_code 8 request_code 72 minor_code 0
```

Remedy this by doing:

```
export XLIB_SKIP_ARGB_VISUALS=1
```

prior to running the failing program. If this workaround is successful, then consider adding this line to `/etc/bashrc` or to `~/.bashrc`. See https://bugs.launchpad.net/ubuntu/+source/xmms/+bug/58192 for details.

Caution when modifying Scyld ClusterWare scripts

Scyld ClusterWare installs various scripts in `/etc/beowulf/init.d/` that **node_up** executes when booting each node in the cluster. Any site-local modification to one of these scripts will be lost when a subsequent Scyld ClusterWare update

overwrites the file with a newer version. If a cluster administrator believes a local modification is necessary, we suggest:

1. Copy the to-be-edited original script to a file with a unique name, e.g.:

   ```
   cd /etc/beowulf/init.d
   cp 20ipmi 20ipmi_local
   ```

2. Remove the executable state of the original:

   ```
   /sbin/beochkconfig 20ipmi off
   ```

3. Edit `20ipmi_local` as desired.

4. Thereafter, subsequent Scyld ClusterWare updates may install a new `20ipmi`, but that update will not re-enable the non-executable state of that script. The locally modified `20ipmi_local` remains untouched. However, keep in mind that the newer Scyld ClusterWare version of `20ipmi` may contain fixes or other changes that need to be reflected in `20ipmi_local` because that edited file was based upon an older Scyld ClusterWare version.

Caution using tools that modify config files touched by Scyld ClusterWare

Software tools exist that might make modifications to various system configuration files that Scyld ClusterWare also modifies. These tools do not have knowledge of the Scyld ClusterWare specific changes and therefore may undo or cause damage to the changes or configuration. Care must be taken when using such tools. One such example is **/usr/sbin/authconfig**, which manipulates `/etc/nsswitch.conf`.

Scyld ClusterWare modifies these system configuration files at install time:

```
/etc/exports
/etc/nsswitch.conf
/etc/security/limits.conf
/etc/sysconfig/syslog
```

Additionally, Scyld ClusterWare uses **/sbin/chkconfig** to enable *nfs*.

Running nscd service on master node may cause kickbackdaemon to misbehave

The **nscd** (Name Service Cache Daemon) service executes by default on each compute node. However, if this service is also enabled on the master node, then it may cause the Scyld ClusterWare name service **kickbackdaemon** to misbehave.

Workaround: when Beowulf starts, if it detects that **nscd** is running on the master node, then Beowulf automatically stops **nscd** and reports that it has done so. Beowulf does not invoke **/sbin/chkconfig nscd off** to permanently turn off the service.

Note: even after stopping **nscd** on the master node,

```
/sbin/service nscd status
```

will report that **nscd** is running because the daemon continues to execute on each compute node, as controlled by `/etc/beowulf/init.d/09nscd`.

Scyld ClusterWare MVAPICH/MVAPICH2 CPU affinity management

The default MVAPICH/MVAPICH2 behavior is to assign threads of each multithreaded application to specific CPUs in each node, beginning with cpu0 and incrementing upward. Keeping threads pinned to specific, otherwise idle CPUs is usually an optimal NUMA and CPU cache strategy for nodes that are dedicated to a single application. However, it is usually

suboptimal if multiple MVAPICH/MVAPICH2 applications share the same node, since every MVAPICH/MVAPICH2 application's threads get permanently assigned to the same low-numbered CPUs and thus contend for the same CPU cycles and CPU caches. The Scyld ClusterWare MVAPICH/MVAPICH2 distribution reverses this default behavior and does not impose strict CPU affinity assignments, which allows the kernel CPU scheduler to migrate threads within a node as it sees fit to load-balance the node's CPUs as workloads change over time.

However, the user may override this Scyld ClusterWare default for MVAPICH using:

```
export VIADEV_ENABLE_AFFINITY=1
```

and for MVAPICH2 using:

```
export MV2_ENABLE_AFFINITY=1
```

Conflicts with base distribution of OpenMPI

Scyld ClusterWare 5.11.2 includes MPI-related packages that conflict with certain packages in the Red Hat or CentOS base distribution.

If **yum** informs you that it cannot install or update Scyld ClusterWare because various **mpich** and **mpiexec** packages conflict with various **openmpi** packages from the base distribution, then run the command:

```
yum remove openmpi* mvapich*
```

to remove the conflicting base distribution packages, then retry the *groupupdate* of Scyld-ClusterWare.

Beofdisk does not support local disks without partition tables

Currently, **beofdisk** only supports disks that already have partition tables, even if those tables are empty. Compute nodes with preconfigured hardware RAID, where partition tables have been created on the LUNs, should be configurable. Contact Customer Service for assistance with a disk without partition tables.

Issues with bproc and the getpid() syscall

BProc interaction with *getpid()* may return incorrect processID values.

Details: The Red Hat's glibc implements the *getpid()* syscall by asking the kernel once for the current processID value, then caching that value for subsequent calls to *getpid()*. If a program calls *getpid()* before calling *bproc_rfork()* or *bproc_vrfork()*, then bproc silently changes the child's processID, but a subsequent *getpid()* continues to return the former cached processID value.

Workaround: do not call *getpid()* prior to calling *bproc_[v]rfork*.