Release Notes

Scyld ClusterWare Release 4.9.0-490g0000

About This Release

Scyld ClusterWare Release 4.9.0-490g0000 (released April 22, 2011) is the latest update to Scyld ClusterWare 4.

Scyld ClusterWare 4.9.0 expects to execute in a Red Hat RHEL4 Update 9 or CentOS 4.9 base distribution environment, each having been updated to the latest RHEL4/CentOS4 errata (https://rhn.redhat.com/errata/rhel4as-errata.html) as of the Scyld ClusterWare 4.9.0 release date. Any compatibility issues between Scyld ClusterWare 4.9.0 and RHEL4 are documented on the Penguin Computing Support Portal at http://www.penguincomputing.com/support.

Visit http://docs.redhat.com/docs/en-US/Red_Hat_Enterprise_Linux/index.html to view the Red Hat Enterprise Linux 4 *4.9 Release Notes* and *4.9 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/files/scyld-docs/CW4/ReleaseNotes.pdf. 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 4.9.0 (see the Section called *Upgrading Earlier Release of Scyld ClusterWare to Scyld ClusterWare 4.9.0*), or how to install Scyld ClusterWare 4.9.0 as a fresh install (see the Section called *Installing Scyld ClusterWare 4.9.0 on a Non-Scyld ClusterWare System*).

Important for clusters using Infiniband

Updating from RHEL4 Update 6 (or CentOS 4.6) to Update 7 will update the Open Fabric Infiniband (OFED) software stack from version 1.2 to 1.3. Updating to Update 8 or beyond will update to OFED version 1.4. Scyld ClusterWare itself is compatible with either OFED 1.2, 1.3, or 1.4, but OFED 1.3 and 1.4 may be incompatible with MPI stacks supplied by certain ISV applications. Penguin Computing recommends that you do not upgrade Infiniband clusters beyond RHEL4 Update 6 without first verifying that your applications that use Infiniband are compatible with OFED 1.4. See the Section called *OFED 1.2 vs. OFED 1.3 Issues* in the Section called *Known Issues And Workarounds* for details.

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: 2.6.9-100.490g0000.ELsmp. 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 `490g0000`. If you find a Panasas rpm matching the to-be-installed 2.6.9-100.490g0000.ELsmp kernel, then download that rpm and continue with the Scyld ClusterWare update or install. Install the Panasas rpm after you finish installing the associated 2.6.9-100.490g0000.ELsmp kernel. If you do not find an appropriate Panasas rpm, then do not install this latest Scyld ClusterWare 4.9.0. The Panasas storage will not work with the 2.6.9-100.490g0000.ELsmp kernel without a matching Panasas kernel module.

Upgrading Earlier Release of Scyld ClusterWare to Scyld ClusterWare 4.9.0

When upgrading from an earlier Scyld ClusterWare 4 version to Scyld ClusterWare 4.9.0, 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. The latest Red Hat errata is RHEL4 Update 9. Penguin Computing recommends that Infiniband clusters do not upgrade beyond RHEL4 Update 6 (or CentOS 4.6) without first verifying that all the applications that use Infiniband are compatible with OFED 1.3 and 1.4. See the Section called *OFED 1.2 vs. OFED 1.3 Issues* in the Section called *Known Issues And Workarounds* for details. Non-Infiniband clusters should safely upgrade to the RHEL4 Update 9 (or CentOS 4.9) base distribution errata.

   To fully update (including OFED 1.4), you must first remove some current Infiniband packages that are restructured in the latest RHEL4/CentOS4 release and will not otherwise update smoothly:

   ```
   rpm -qa | grep dapl | xargs rpm -e --nodeps --allmatches
   ```

   then update a RHEL4 base distribution using:

   ```
   up2date -u
   ```

   or a CentOS 4 distribution using:

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

   as appropriate for your cluster.
4. **Only if upgrading Scyld ClusterWare 4.2.1 or earlier**, then your existing `clusterware.repo` file contains an outdated user/password authentication and must be updated. Perform these additional steps:
   a. Login to the Penguin Computing Support Portal at http://www.penguincomputing.com/support.
   b. Click on *Download your Scyld ClusterWare 4 YUM repo file* to download the new `clusterware.repo` file and place it in the `/etc/yum.repos.d/` directory, replacing the file already there.
   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 upgrading from Scyld ClusterWare 4.8.x or earlier**, then edit file `/etc/yum.repos.d/clusterware.repo` to change the *baseurl* entries from that earlier version (e.g., *4.8*) to version *4.9*.
6. If your cluster includes Infiniband hardware, and if you have chosen to upgrade to CentOS 4.9 (and OFED 1.4), 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-scyld packages** (see the Section called *Issues with OpenMPI*):

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

   and manage the specific installation of new openmpi-scyld packages in the next step.

   Otherwise, if you want to **replace the existing openmpi-scyld packages**, then do:

   ```
   yum groupupdate 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/
ls cw-*/packages/kernel-*
```

and locate the newest Scyld ClusterWare 2.6.9-100.490g0000.ELsmp kernel, which should reside in either `cw-core/packages/` or `cw-updates/packages/`. Then install that newest kernel:

```
rpm -iv --oldpackage cw-*/packages/kernel-smp-*490g0000*rpm
```

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

8. If the previous **yum groupupdate** excluded the openmpi-scyld* packages, then you can explicit download new packages, if any are available, and install them in a way that avoids removing older openmpi-scyld* packages:

   ```
   rm -fr /tmp/ompinew
   mkdir -p /tmp/ompinew
   yumdownloader --destdir /tmp/ompinew openmpi-scyld*
   rpm -iv /tmp/ompinew/openmpi-scyld*
   ```

   Note the use of **yumdownloader** and **rpm -i**. This is necessary because doing **yum install openmpi-scyld*** would not, in fact, do a simple *install* and retain older packages. Rather, it would perform an *update* and remove the older openmpi-scyld* package(s).

9. **If the cluster uses Panasas storage**, then you should have already downloaded the Panasas rpm that matches the Scyld ClusterWare 4.9.0 kernel you have just installed. Now install the Panasas rpm using **rpm -i**.

10. 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`.

11. Disable (or delete) any old `/etc/beowulf/init.d/` scripts that may be lingering from older Scyld ClusterWare releases.

    ```
    /sbin/beochkconfig --list
    ```

    displays the list of existing Beowulf scripts. Common old scripts are `ipmi` (replaced by `20ipmi`), `torque` (replaced by `90torque`), and unnumbered infiniband scripts (replaced by `15openib` and `16ipoib`). `memdev` is no longer necessary and should be deleted. Any such old script should be disabled (e.g., **/sbin/beochkconfig ipmi off**) or deleted (e.g., **/sbin/beochkconfig --del ipmi**).

12. Examine `/etc/grub.conf` to confirm that the new 2.6.9-100.490g0000.ELsmp kernel is the default, then reboot your master node.

13. If upgrading from Scyld ClusterWare 4.2.x, 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). See step 11 in the Section called *Installing Scyld ClusterWare 4.9.0 on a Non-Scyld ClusterWare System* for details.

Installing Scyld ClusterWare 4.9.0 on a Non-Scyld ClusterWare System

When installing Scyld ClusterWare 4.9.0 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. The latest Red Hat errata is RHEL4 Update 9. Penguin Computing recommends that Infiniband clusters do not upgrade beyond RHEL4 Update 6 (or CentOS 4.6). See the Section called *OFED 1.2 vs. OFED 1.3 Issues* in the Section called *Known Issues And Workarounds* for details. Non-Infiniband clusters should be upgraded to RHEL4 Update 9 (or CentOS 4.9) base distribution errata.

   To fully update (including OFED 1.4), you must first remove some current Infiniband packages that are restructured in the latest RHEL4/CentOS4 release and will not otherwise update smoothly:

   ```
   rpm -qa | grep dapl | xargs rpm -e --nodeps --allmatches
   ```

   then update a RHEL4 base distribution using:

   ```
   up2date -u
   ```

   or CentOS using:

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

   as appropriate for your cluster.
3. Remove base distribution packages that conflict with Scyld ClusterWare 4:

   ```
   yum remove openmpi* lam*
   ```
4. Download a custom Yum repo file:

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

   b. Click on *Download your Scyld ClusterWare 4 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, and if have chosen to upgrade to CentOS 4.9 with its OFED 1.4, 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:

   ```
   rpm -iv --oldpackage cw-*/packages/kernel-smp-*490g0000*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 4.9.0 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 2.6.9-100.490g0000.ELsmp kernel is the default, then reboot your master node.
11. 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** or **/usr/sbin/beosetup** commands. 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.

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 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 *admin* account 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 three system services; or without rebooting, manually restart *xinetd*, and 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

To enable the beoweb service that facilitates remote job submission and cluster monitoring (e.g., used by POD Tools):

```
/sbin/chkconfig beoweb on
```

then either reboot the master node, which automatically starts beoweb; or without rebooting, manually start beoweb:

```
/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 16384 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 **kernel-utils** 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/cpuspeed.conf`, to override default behavior. See **man cpuspeed** for details.

Optionally enable SSHD on compute nodes

If you wish to allow users to **/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
```

See the *Administrator's Guide* for details.

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 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, 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

The kernel 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 4.9 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 4 *Release Notes* from the Penguin Computing Support Portal (http://www.penguincomputing.com/files/scyld-docs/CW4/ReleaseNotes.pdf) 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 4.3.1

New in Scyld ClusterWare 4.9.0 - Scyld Release 490g0000

1. The base kernel is upgraded to 2.6.9-100.490g0000.ELsmp. See https://rhn.redhat.com/errata/RHSA-2011-0263.html for details.
2. 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.
3. 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.
4. Enhances the **beoserv** daemon to increase the max number of supported network interfaces per compute node from seven to 16.
5. OpenMPI is upgraded to version 1.5.3. See the Section called *Upgrading Earlier Release of Scyld ClusterWare to Scyld ClusterWare 4.9.0* step 7 and step 8 for how to install co-existing openmpi-scyld releases, and the Section called *Issues with OpenMPI* for details. Existing applications that were built against version 1.4.3 or earlier must be rebuilt against this new version.
6. 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 4.8.2 Update - Scyld Release 482g0005

1. The base kernel is upgraded to 2.6.9-89.35.1.482g0003. The base distribution components are the same as the earlier Scyld Release 482g0004 kernel, 2.6.9-89.35.1.482g0002, 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 482g0004.
5. The latest TORQUE version 2.5.3 (introduced in Scyld Release 482g0000) has seemingly introduced various problems running TORQUE jobs, and it is being withdrawn. See the Section called *Issues with TORQUE version 2.5.3* for details.

New in Scyld ClusterWare 4.8.2 Update - Scyld Release 482g0004

1. The base kernel is upgraded to 2.6.9-89.35.1. See https://rhn.redhat.com/errata/RHSA-2011-0162.html for details.

2. OpenMPI is upgraded 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 *Upgrading Earlier Release of Scyld ClusterWare to Scyld ClusterWare 4.9.0* step 7 and step 8 for how to install co-existing openmpi-scyld releases, and the Section called *Issues with OpenMPI* for details. Existing applications that were built against version 1.4.3 or earlier must be rebuilt against this new version.

New in Scyld ClusterWare 4.8.2 Update - Scyld Release 482g0003

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 4.8.2 Update - Scyld Release 482g0002

1. The base kernel is upgraded to 2.6.9-89.33.1. See https://rhn.redhat.com/errata/RHSA-2010-0936.html for details.
2. The Scyld ClusterWare igb Ethernet driver is upgraded 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 RHEL4-U8 driver (version 1.2.45-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.

New in Scyld ClusterWare 4.8.2 Update - Scyld Release 482g0001

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 4.8.2 - Scyld Release 482g0000

1. The base kernel is upgraded to 2.6.9-89.31.1 (482g0000). See https://rhn.redhat.com/errata/RHSA-2010-0779.html for details.
2. The Scyld ClusterWare igb Ethernet driver is upgraded 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 RHEL4-U8 driver (version 1.2.45-k2).

3. The Scyld ClusterWare e1000e Ethernet driver is upgraded to version 1.2.17. This driver derives from source found at http://sourceforge.net/projects/e1000/files/. We recommend using this Scyld ClusterWare driver instead of the native RHEL4-U8 driver (version 0.3.3.3-k6).
4. 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.
5. 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.
6. 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.
7. 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.
8. 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.
9. Improves the performance of **beonss** *kickback* name resolution from compute nodes.
10. Fixes a rare failure to PXEboot nodes that employ the igb Gigabit Ethernet driver.
11. Fixes the Scyld ClusterWare **Ganglia** "proc_run" graph that showed an incorrect and excessively large number of running processes.
12. **openmpi-scyld** is upgraded 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 upgraded to version 2.5.3.
14. Taskmaster version 5.4.1 is now available (under a separate license).

New in Scyld ClusterWare 4.8.1 Update - Scyld Release 481g0008

1. The base kernel is upgraded to 2.6.9-89.29.1 (481g0005). See https://rhn.redhat.com/errata/RHSA-2010-0676.html for details.
2. OpenMPI is upgraded to version 1.4.3.
3. Fixes a problem that rarely exhibited itself as a *waitpid()* failure in programs executing in `/etc/beowulf/init.d/` scripts during node bootup.
4. Enhances **beoserv** DHCP to better handle non-Scyld non-Linux compute nodes that request a DNS IP address.

New in Scyld ClusterWare 4.8.1 Update - Scyld Release 481g0007

1. The base kernel is upgraded to 2.6.9-89.0.29 (481g0004). See https://rhn.redhat.com/errata/RHSA-2010-0718.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.
4. Fixes a **bpcp -p** bug where the *mode* was not properly preserved across the copy.
5. Fixes a Scyld ClusterWare **ganglia** bug where the network bytes/second data rates were being misreported.

New in Scyld ClusterWare 4.8.1 Update - Scyld Release 481g0006

1. The base kernel is upgraded to 2.6.9-89.0.28 (481g0003). See https://rhn.redhat.com/errata/RHSA-2010-0606.html for details.
2. Scyld ClusterWare now includes an e1000e Ethernet driver, version 1.2.8 from http://sourceforge.net/projects/e1000/, replacing the e1000e (version 0.5.18.3) that was introduced in CW4.3.1 as an improvement over the native RHEL4-U3 driver.
3. Scyld ClusterWare now includes an igb Ethernet driver, version 2.2.9, from http://sourceforge.net/projects/e1000/, replacing the igb (version 1.3.8.6) that was introduced in CW4.3.1 as an improvement over the native RHEL4-U3 driver.
4. 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.
5. 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 4.8.1 Update - Scyld Release 481g0005

1. The base kernel is upgraded to 2.6.9-89.0.26. See https://rhn.redhat.com/errata/RHSA-2010-0474.html for details.
2. 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).
3. 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.
4. **/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.
5. **/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.

6. 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.
7. Eliminates a bogus **recvstats** syslog message of the form "Received stats from IP addr" that occasionally appeared as a compute node starts up.

New in Scyld ClusterWare 4.8.1 Update - Scyld Release 481g0004

1. The base kernel is upgraded to 2.6.9-89.0.25. See https://rhn.redhat.com/errata/RHSA-2010-0394.html for details.
2. 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.
3. 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*.
4. Eliminates a bogus BProc syslog message "proc.exe not null".

New in Scyld ClusterWare 4.8.1

1. The base kernel is upgraded to 2.6.9-89.0.23. See https://rhn.redhat.com/errata/RHSA-2009-1541.html, https://rhn.redhat.com/errata/RHSA-2009-1671.html, https://rhn.redhat.com/errata/RHSA-2010-0020.html, https://rhn.redhat.com/errata/RHSA-2010-0076.html, and https://rhn.redhat.com/errata/RHSA-2010-0146.html for details.
2. 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.
3. 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.
4. Fixes a bug in BProc where certain workloads would cause a master node kernel panic, most commonly a segfault in the routine *ghost_put*.
5. 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.
6. 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.

7. Fixes a bug where cluster startup would leave temporary files in `/tmp/`. These are now properly deleted.
8. 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/`.
9. The **beoserv** daemon now automatically removes duplicate MAC addresses from file `/etc/beowulf/unknown_addresses`.
10. 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.
11. 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.
12. 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.
13. OpenMPI is upgraded to version 1.4.1.
14. TORQUE is upgraded to version 2.3.10.
15. 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 4.8.0

1. The initial CW4.8.0 release included a kernel that was based upon RHEL4 2.6.9-89.0.15. The current CW4.8.0 yum repository contains a newer kernel that is based upon RHEL4 2.6.9-89.0.18. See https://rhn.redhat.com/errata/RHSA-2009-1024.html, https://rhn.redhat.com/errata/RHSA-2009-1132.html, https://rhn.redhat.com/errata/RHSA-2009-1211.html, https://rhn.redhat.com/errata/RHSA-2009-1223.html, https://rhn.redhat.com/errata/RHSA-2009-1438.html, and https://rhn.redhat.com/errata/RHSA-2009-1522.html for details about Red Hat kernel changes between CW4.3.1 and 2.6.9-89.0.15. See https://rhn.redhat.com/errata/RHSA-2009-1541.html and https://rhn.redhat.com/errata/RHSA-2009-1671.html for details about subsequent changes through 2.6.9-89.0.18.
2. The initial CW4.8.0 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 CW4.3.1 and called **ClusterAdmin**. The latest CW4.8.0 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**.
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. OpenMPI is upgraded to version 1.3.3.
13. TORQUE is upgraded to version 2.3.7.
14. 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.
15. 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.

Known Issues And Workarounds

The following are known issues of significance with the latest version of Scyld ClusterWare 4.9.0 and suggested workarounds.

Issues with TORQUE version 2.5.3

TORQUE version 2.5.3, which was introduced in Scyld Release 482g0000 and withdrawn in Scyld Release 482g0005, has seemingly introduced various problems running TORQUE jobs. If your cluster has experienced new TORQUE problems that you believe appeared after an upgrade to version 2.5.3, then we suggest reverting to an earlier version:

```
rpm -qa | grep torque-2.5.3 | xargs rpm -e --nodeps
yum install torque
```

will install the preferred TORQUE version.

OFED 1.2 vs. OFED 1.3 Issues

Updating from RHEL4 Update 6 (or CentOS 4.6) to Update 7 also updates the Open Fabric Infiniband (OFED) software stack from version 1.2 to 1.3. Updating to Update 8 updates to OFED version 1.4. Scyld ClusterWare itself is compatible with either OFED 1.2, 1.3, or 1.4, but OFED 1.3 and 1.4 may be incompatible with MPI stacks supplied by certain ISV applications.

For example, OFED 1.2 includes DAPL 1.0, with configurations found in `/etc/ofed/dat64.conf` and `dat32.conf`. OFED 1.3 includes DAPL 2.0, with configurations found in `/etc/ofed/dat.conf`, and thus some DAPL applications will fail to find their intended DAPL libraries. One solution might be to reconfigure a DAPL application to use the IBV (OpenIB) transport instead of the DAPL transport. However, not all MPI stacks support the IBV transport.

No OFED 1.3 problems have been observed for applications based upon **MVAPICH** or **OpenMPI**.

You may want to continue to use OFED 1.2 and to avoid an upgrade to OFED 1.3 or 1.4. This may be accomplished by doing an update to the base distribution and excluding the Infiniband-related rpms, thus retaining whatever Infiniband-related rpms (presumably OFED 1.2) are already installed on the master node. For RHEL4:

```
up2date -u --exclude={*dapl*,ib*,infiniband*}  \
           --exclude={libib*,libcxgb3*,libehca*,libmlx4*}  \
           --exclude={libmthca*,libnes*,librdmacm*,libsdp*}  \
           --exclude={ofed*,openib*,opensm*,qlvnictools*,qperf*,srp*}
```

or for CentOS:

```
yum update --exclude={*dapl*,ib*,infiniband*}  \
           --exclude={libib*,libcxgb3*,libehca*,libmlx4*}  \
           --exclude={libmthca*,libnes*,librdmacm*,libsdp*}  \
           --exclude={ofed*,openib*,opensm*,qlvnictools*,qperf*,srp*}
```

The full list of Infiniband-related rpms is:

```
compat-dapl-1.2.5
compat-dapl-devel-1.2.5
compat-dapl-static-1.2.5
dapl
dapl-devel
dapl-static
dapl-utils
ibsim
ibutils
infiniband-diags
libcxgb3
libcxgb3-devel
```

```
libcxgb3-static
libehca
libibcm
libibcm-devel
libibcm-static
libibcommon
libibcommon-devel
libibcommon-static
libibmad
libibmad-devel
libibmad-static
libibumad
libibumad-devel
libibumad-static
libibverbs
libibverbs-devel
libibverbs-static
libibverbs-utils
libipathverbs
libmlx4
libmthca
libmthca-devel
libmthca-static
libnes
librdmacm
librdmacm-devel
librdmacm-static
librdmacm-utils
libsdp
ofed-docs
openib
openib-diags
openib-mstflint
openib-perftest
openib-srptools
openib-tvflash
opensm
opensm-devel
opensm-libs
opensm-static
qlvnictools
qperf
srptools
udapl
udapl-devel
```

The most recent information is available on the Penguin Computing Support Portal, http://www.penguincomputing.com/support.

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.

Caution using ethtool

ethtool -G may be used to set the network interface's ring buffer size. Performing this action on an interface that uses the **forcedeth** driver will cause that interface to stop working. Use **ethtool -i** *interface-name* to view the *interface-name* and driver pairing.

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 alternative 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 Scyld-repackaged release of the Open Source OpenMPI (http://www.open-mpi.org/: the **openmpi-scyld** base package, plus several compiler-environment-specific packages: **openmpi-scyld-gnu**, **openmpi-scyld-intel**, and **openmpi-scyld-pgi**. Each openmpi-scyld rpm carries a version number that matches the originating Open Source OpenMPI version.

Beginning with openmpi-scyld version 1.5.1, Scyld ClusterWare installs the files into version-specific directories, which allows multiple openmpi-scyld versions to co-exist on the master node. (See the Section called *Upgrading Earlier Release of Scyld ClusterWare to Scyld ClusterWare 4.9.0* step 7 and step 8 for how to install co-existing openmpi-scyld releases.) The `/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.

Supporting only one openmpi-scyld version on the master node tends to cause problems because OpenMPI applications linked against an earlier version may break when a new version updates and replaces the old, until the applications are rebuilt against the new version. If a new openmpi-scyld version is installed to co-exist with previous version, vs. updated to replace earlier versions, then (for example) the default **module load openmpi/gnu** references the newest version, and the version-specific **module load openmpi/gnu/1.5** references the older openmpi-scyld version 1.5, thereby allowing for a grace period for users to continue to execute applications that are linked to an older version, without needing to immediately rebuild every application to work with the new version. Version 1.5.1 (and beyond) can co-exist with earlier versions, although co-existence is problematic:

- Version 1.5 libraries, executable binaries, and manpages reside in directories `/opt/scyld/openmpi/gnu/`, etc., whereas version 1.5.1 files reside in directories `/opt/scyld/openmpi/1.5.1/gnu/`, etc. This causes no problems with co-existence.

  However, version 1.5 modulefiles reside in (for example) file `/opt/scyld/modulefiles/openmpi/gnu`, whereas version 1.5.1 (and beyond) modulefiles use those same pathnames as directories, and use version-specific names for the modulefiles themselves, e.g., `/opt/scyld/modulefiles/openmpi/gnu/1.5.1`. If and when version 1.5.1 installs to co-exist with version 1.5, it silently transforms the preexisting version 1.5 modulefiles into the same directory and file paradigm employed by version 1.5.1. In other words, the co-existing gnu modulefiles become:

  ```
  /opt/scyld/modulefiles/openmpi/gnu/1.5
  /opt/scyld/modulefiles/openmpi/gnu/1.5.1
  ```

- Version 1.4.3 (and earlier) uses libraries located in `/usr/lib64/OMPI/`*compiler*`/`, executable binaries located in `/usr/openmpi/`*compiler*`/bin/`, and manpages located in `/usr/openmpi/man/`. While version 1.5.1 can co-exist with 1.4.3, in order to employ **module load openmpi/gnu/1.4.3** the cluster administrator must manually create version 1.4.3 modulefiles as (for example) files:

  ```
  /opt/scyld/modulefiles/openmpi/gnu/1.4.3
  ```

  following the same paradigm employed by the 1.5.1 modulefiles.

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 CPU affinity management

CW4.2.0 (and later releases) support Infiniband via Open Source kernel drivers, OpenIB, OFED, and a Scyld ClusterWare-enhanced MVAPICH. The CW4.2.0 MVAPICH default behavior is to assign threads of each multithreaded job to specific CPUs in each node, starting with cpu0 and incrementing upward. While keeping threads pinned to a specific CPU may be an optimal NUMA and CPU cache strategy for nodes that are dedicated solely to a single job, it is usually suboptimal if multiple multithreaded jobs share a node, as each job's threads get permanently assigned to the same low-numbered CPUs. The CW4.2.1 (and beyond) default behavior is to not impose strict CPU affinity assignments, which allows the kernel CPU scheduler to migrate threads as it sees fit to load-balance the node's CPUs as workloads change over time.

However, the user may override this default using:

```
export VIADEV_ENABLE_AFFINITY=1
```

Conflicts with base distribution of openmpi and lam

Scyld ClusterWare 4.9.0 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** and **lam** packages from the base distribution, then run the command:

```
yum remove openmpi* lam*
```

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

Reducing the size of /usr/lib/locale/locale-archive

Glibc applications silently open the file `/usr/lib/locale/locale-archive`, which means it gets pulled into a node's libcache early in a node's startup sequence. The default `locale-archive` is commonly many dozens of megabytes in size. This consumes significant network bandwidth to move it from the master to each node, and thereafter consumes significant

RAM filesystem space on the node. It is likely that your 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. For example, on a quiescent master node:

```
cd /usr/lib
mv locale locale.orig
mkdir locale
cp -a locale.orig/en_US* locale
/usr/sbin/build-locale-archive
```

rebuilds `/usr/lib/locale/locale-archive` as a 2MB file with only the locale data for United States English, and preserves the original `/usr/lib/locale` directory as a backup.

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.

Master node or compute node reports "instable time" in syslog

The Linux 2.6.9 kernel's x86_64 timer code is overly aggressive in declaring "lost ticks" and for some workloads produces a spurious warning, such as "Many lost ticks. Your time source seems instable or some driver is hogging interrupts." This message can be safely ignored.

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*.

Issues with ptrace

ptrace-based debuggers may fail to re-attach.

Details: You may use a **ptrace**-based debugger, such as **gdb**, running on a master node to debug a process that is running on a compute node. However, if the debugger exits without cleanly detaching from the traced process using **ptrace_detach**, then any further attempts to do a **ptrace_attach** to the target process will fail. For example, if **gdb** is tracing a process on a compute node and **gdb** is killed by a signal, then further tracing of that target process will not be possible.