Altair®

PBS Professional®

13.1.500

Power Awareness Release Notes

for SGI Systems

PBS Professional 13.1.500 Power Awareness Release Notes, updated 6/8/16. Please send any questions or suggestions for improvements to agu@altair.com.

Contact Information

For the most recent information, go to the PBS Works website, www.pbsworks.com, select "My PBS", and log in with your site ID and password.

Altair

Altair Engineering, Inc.,
1820 E. Big Beaver Road,
Troy, MI 48083-2031 USA
www.pbsworks.com

Sales

pbssales@altair.com
248.614.2400

Technical Support

Need technical support? We are available from 8am to 5pm local times:

Location	Telephone	e-mail
Australia	+1 800 174 396	anz-pbssupport@india.altair.com
China	+86 (0)21 6117 1666	es@altair.com.cn
France	+33 (0)1 4133 0992	pbssupport@europe.altair.com
Germany	+49 (0)7031 6208 22	pbssupport@europe.altair.com
India	+91 80 66 29 4500 +1 800 425 0234 (Toll Free)	pbs-support@india.altair.com
Italy	+39 800 905595	pbssupport@europe.altair.com
Japan	+81 3 5396 2881	pbs@altairjp.co.jp
Korea	+82 70 4050 9200	pbs@altair.co.kr
Malaysia	+91 80 66 29 4500 +1 800 425 0234 (Toll Free)	pbs-support@india.altair.com
North America	+1 248 614 2425	pbssupport@altair.com
Russia	+49 7031 6208 22	pbssupport@europe.altair.com
Scandinavia	+46 (0)46 460 2828	pbssupport@europe.altair.com
Singapore	+91 80 66 29 4500 +1 800 425 0234 (Toll Free)	pbs-support@india.altair.com
South America	+55 11 3884 0414	br_support@altair.com
UK	+44 (0)1926 468 600	pbssupport@europe.altair.com

Contents

1 Supported Platforms and New Features 1
- 1.1 About These Release Notes . . . 1
- 1.2 Supported Platforms for PBS 13.1.500 . . . 1
- 1.3 Features Introduced in Previous Versions . . . 1

2 Installation 3
- 2.1 Installing PBS on the SGI ICE . . . 3

3 Power Awareness 11
- 3.1 Using Power Awareness . . . 11
- 3.2 Terminology . . . 11
- 3.3 Prerequisites . . . 12
- 3.4 Configuring PBS for Power Awareness . . . 12
- 3.5 New Attributes and Resources for Power Awareness . . . 13
- 3.6 Reading and Setting Power-related Attributes and Resources in Hooks 14
- 3.7 Submitting Jobs with Power Profiles . . . 16
- 3.8 Caveats and Restrictions for Configuring Power Awareness . . . 17

4 Issues 19
- 4.1 Open Bugs in PBS 13.1.500 . . . 19

Index 21

Contents

1 Supported Platforms and New Features

1.1 About These Release Notes

These release notes describe changes for 13.1.500, and include new installation instructions in "Installation" on page 3.

1.2 Supported Platforms for PBS 13.1.500

The following hardware:

- SGI ICE X or newer (ICE 8200 and ICE 8400 do not have power management)

The following operating systems and versions of SGI Management Center (SMC):

- Red Hat Enterprise Linux 6 and 7 with SMC 3.1 and above
- SUSE Linux Enterprise Server 11 and 12 with SMC 3.1 and above

1.3 Features Introduced in Previous Versions

Power Profiles

This version of PBS supports running each job within a pre-defined power profile (for example, to limit the job's power consumption).

Energy Accounting

This version of PBS accounts for the energy used by each job. When power awareness is enabled, you can see the energy used by each job in resources_used.energy, which is the number of kWh the job used during its run. PBS shows the energy used in the output of `qstat` and in the `pbs_server` accounting files.

2 Installation

2.1 Installing PBS on the SGI ICE

This chapter is designed to **replace** section 3.6.5 in the PBS Professional Installation & Upgrade Guide.

Make sure that you have covered the prerequisites in section 3.5, "Prerequisites for Installing on UNIX/Linux Systems", on page 45 of the PBS Professional Installation & Upgrade Guide.

2.1.1 Considerations

Most sites that are running PBS Professional on an SGI ICE will not want PBS to manage the cpusets on the machine. These sites should install the standard PBS MoM.

Placement sets improve job placement on execution nodes. See "Placement Sets" on page 229 in the PBS Professional Administrator's Guide. If you run the cpuset MoM, placement sets are generated automatically for the machines listed in "Generation of Placement Set Information" on page 972 in the PBS Professional Administrator's Guide. If you run the standard MoM, placement sets are not automatically generated, and you may wish to configure placement sets. See sectoin 2.1.0.9, "Configuring Placement Sets on the SGI ICE", on page 9 of the PBS Professional Installation & Upgrade Guide for the steps involved.

2.1.2 SGI ICE Components

An ICE system consists of one Admin node, one or more Service (login) nodes, and a set of one or more compute racks. Each compute rack consists of one or more IRU nodes and one or more compute nodes per IRU. The racks are diskless. The root file system of the IRU and compute nodes are mounted read-only from a NAS managed by the Admin node. There is a single image of the root file system for all of the compute nodes and a separate image for all of the IRU nodes. Tempo node management commands are used to publish the image to the various nodes in a process that involves powering down the nodes, pushing a new image, and re-powering the nodes.

In a typical configuration, user home file systems are mounted from NAS, and each node has a separately mounted file system for `/var/spool`.

SGI follows a naming convention when preparing a system for shipment. Service nodes are named “service0”, “service1”, … Compute nodes are named rRiLnN where 'R' is the rack number starting with 1; 'L' is the IRU node number within a rack starting with 0 in each rack; N is the node number, starting with 0, under the specific Rack Leader. For example, two racks with 2 IRUs per rack and 4 nodes per IRU are named:

Table 2-1: Node Names

IRU	Rack 1	Rack 2
IRU 0	*r1i0n0*	*r2i0n0*
	r1i0n1	*r2i0n1*
	r1i0n2	*r2i0n2*
	r1i0n3	*r2i0n3*
IRU 1	*r1i1n0*	*r2i1n0*
	r1i1n1	*r2i1n1*
	r1i1n2	*r2i1n2*
	r1i1n3	*r2i1n3*

2.1.3 Requirements for the SGI ICE with Performance Suite

- In order to run PBS on the SGI ICE with Performance Suite, SGI's Tempo node management tools must already be installed. You will be using the following Tempo commands:

Table 2-2: Tempo Commands

Tempo command	Description
`cnodes --compute`	List the compute node names; useful in scripting operations
`cpower --down NODE`	Powers down
`cpower --up NODE`	Powers up the named nodes
`cimage --…`	Manages the file system image for the various nodes

- You must use the correct names for the Admin and Service nodes in any commands.
- If you will use PBS to manage the cpusets on the ICE, the file `/etc/sgi-compute-node-release` must be present in order for the PBS init.d script to create vnode definitions and for the `pbs_habitat` script to set up the CPUSET flags.

2.1.4 Choosing Whether PBS Will Manage Cpusets with SGI ICE Running Performance Suite

You can choose whether or not to use PBS to manage the cpusets on the SGI ICE. Most sites will not want PBS to manage the cpusets. To have PBS manage the cpusets, install `pbs_mom.cpuset`; otherwise, install `pbs_mom.standard`.

2.1.4.1 Using PBS to Manage cpusets on the SGI ICE

If you are running the cpuset MoM, the `init.d/pbs` script will configure one vnode per MoM. This enables cpusets and sets sharing to *default_shared*.

To provide the maximum number of available CPUs on a small node, make sure that the file `/etc/sgi-compute-node-release` is present. This way, on installation the `pbs_habitat` script will add a "`cpuset_create_flags 0`" to MoM's config file.

In order to exclude CPU 0, change the MoM configuration file line to

```
cpuset_create_flags CPUSET_CPU_EXCLUSIVE
```

This flag controls only whether CPU 0 is included in the PBS cpuset.

There is only one logical memory pool available per node on the SGI ICE. If, at startup, MoM finds:

- any CPU in an existing, non-root, non-PBS cpuset
- CPU 0 has been excluded as above

MoM will

- Exclude that CPU from the top set `/dev/cpuset/PBSPro`
- Create the top set with `mem_exclusive` set to false

Otherwise, the top set is created using all CPUs and with mem_exclusive set to *True*.

2.1.5 Installation of the PBS Server, Scheduler, and Communication Daemons

The PBS server, Scheduler, communication daemon, and commands are installed on a single service node; here we assume this node is "service0". To install the complete PBS package:

1. Log on to service0 as root.
2. Unzip and untar the install package.
3. Change to the directory created when the package was untarred.
4. Run the INSTALL script: accept the default locations, and select the number *1* (Server, execution, communication, and commands) option.
5. When asked whether you wish to start PBS, answer "*no*".
6. Modify `/etc/pbs.conf` by changing the line "`PBS_START_MOM=1`" to "`PBS_START_MOM=0`".

2.1.6 Installation of the PBS MoM

You install and configure MoM on each of the compute nodes. The installation and configuration is only done once on the root file system; this is then pushed to all of the compute nodes.

1. Log on to the Admin node as root.
2. Determine which image file is being used on the compute nodes. To list the nodes on the rack 1:

   cimage --list-nodes r1

   It will show output in the form "*node: image_name kernel*" similar to

   r1i0n0: compute-sles10sp1 2.6.26.46-0.12-smp

   Thus node r1i0n0 is running the image "`compute-sles10sp1`" and the kernel version "2.6.26.46-0.12-smp". For the remaining steps, it is assumed that those are the images and kernel available.
3. List the available images:

   cimage --list-images

   which will list the images available for the compute nodes. Each image may have multiple kernels.
4. Unless you are experienced in managing the image files, it is suggested that you create a copy of the image in use and install PBS in that copy. To copy an image:

   cinstallman --create-image --clone --source compute-sles10sp1 --image compute-sles10sp1pbs
5. The image file lives in the directory `/var/lib/systemimager/images`, so change into the `tmp` directory found in the new image just cloned:

   cd /var/lib/systemimager/images/
   compute-sles10sp1pbs/tmp
6. Copy the PBS install package into this `tmp` directory, unzip and untar it.
7. Change back to the root home directory and chroot to the new image file:

   chroot /var/lib/systemimager/images/
   compute-sles10sp1pbs /bin/sh

   The new root is in effect.
8. Change directory into the PBS directory which was created below `tmp` when the PBS package was untarred. This is `/tmp/PBSPro_<version>`, which was created for the

image when the PBS package was untarred on the Admin node:

```
cd /tmp/PBSPro_<version>
```

9. Run the INSTALL script to install the PBS programs into the normal execution directory `/opt/pbs/M.N.P.S` in this system image.

   ```
   ./INSTALL
   ```

   a. Accept the default directories.

   b. Choose install option *2* (execution).

   c. Answer "*No*" when you are asked whether you wish to start PBS now.

10. The default MoM does not use cpusets. To use cpusets, replace the default `pbs_mom` executable with `PBS_EXEC/sbin/pbs_mom.cpuset`.

    ```
    cp PBS_EXEC/sbin/pbs_mom.cpuset PBS_EXEC/sbin/pbs_mom
    ```

    `PBS_EXEC` is the path to the location where the PBS binaries were installed.

11. Exit from the chroot shell and return to root's normal home directory.

12. Power down each rack of compute nodes:

    ```
    for n in `cnodes --compute` ; do
        cpower node off $n
    done
    ```

13. Publish the new system image to the compute nodes:

    ```
    cimage --push-rack compute-sles10sp1pbs r\*
    ```

    This instruction will take several minutes to finish.

14. Set the new image and kernel to be booted. This set need not be done if: (1) rather than cloning a new image, you have installed PBS into the image already running on the compute nodes; or (2) you are using an image that was already pushed to the nodes.

    ```
    cimage --set compute-sles10sp1pbs 2.6.26.46-0.12-smp r\*i\*n\*
    ```

15. Power up the compute nodes:

    ```
    for n in `cnodes --compute` ; do
        cpower node on $n
    done
    ```

    It will take several minutes for the compute nodes to reboot.

2.1.7 Adding Compute Nodes

1. Log on to the service node as root
2. On the Service node, start the PBS server, scheduler, and communication daemons:

   ```
   /etc/init.d/pbs start
   ```

3. Using `qmgr`, add the compute nodes to the PBS configuration:

   ```
   for N in `cnodes --compute`
   do
       qmgr -c "create node $N"
   done
   ```

2.1.8 Configuring Placement Sets on the SGI ICE

If you are not already using PBS to manage the placement sets on the ICE (you are running the standard MoM), you may wish to configure placement sets. This will improve job placement on execution nodes. See "Placement Sets" on page 229 in the PBS Professional Administrator's Guide.

Placement sets can be defined only after you have defined the compute nodes as in the previous section.

1. Shut down the server.
2. Add a resource named "router" to the `PBS_HOME/server_priv/resourcedef` file, by adding the following line:

   ```
   router type=string_array, flag=h
   ```

3. Restart the server
4. Run the placement set generation script:

   ```
   PBS_EXEC/lib/init.d/sgiICEplacement.sh
   ```

5. Verify the result:
   a. Run the `pbsnodes -a` command
   b. Look for the line "`resources_available.router`" in each node. The value assigned to the "router" resource should be in the form "*r#,r#i#*", where *r* identifies the rack number and *i* identifies the IRU number.

3 Power Awareness

3.1 Using Power Awareness

On SGI systems with the required SMC software, PBS Professional can monitor and control job power usage. PBS provides information about job energy usage in the output of the `qstat` command, and records energy usage in the accounting logs. You can optionally associate a power profile with each job at submission time to control the job's power draw.

You can allow jobs to request power profiles. Each job can request a power profile by requesting a value for the eoe resource. SMC provides a list of power profiles which PBS uses to set each vnode's resources_available.eoe. When a job requests a power profile, it is sent to vnodes that have this profile available. When the job runs, the vnodes where the job runs are set to the power profile requested by the job. The vnode's current_eoe attribute shows this profile.

PBS collects energy consumption information using SMC, and records it in the job's resources_used.energy value.

Once a job is submitted, the process of setting power profiles on vnodes is handled automatically by PBS. The default power setting for a node is no power capping. If a job runs on a node, the node uses the requested power profile, but when the job finishes, the node goes back to the default setting.

3.2 Terminology

Activate a power profile
: To set a power profile on a node, for example, to set a node's power profile to match the specifications for "low".

Deactivate a power profile
: To reset the power profile of the node to its default setting, which is no power capping.

3.3 Prerequisites

- Install PBS as usual. See the PBS Professional Installation & Upgrade Guide.
- You must run this version of PBS on one of the supported versions listed in "Supported Platforms for PBS 13.1.500".
- A prologue script will not run when power_provisioning is set to *True*. Any prologue script must be converted to an execjob_prologue hook.

3.4 Configuring PBS for Power Awareness

1. Set the server's power_provisioning attribute to *True*:

   qmgr -c "set server power_provisioning=True"

2. Set the power_enable vnode attribute to *True* on each vnode where jobs will use power profiles:

   For a specific vnode:

   qmgr -c "set node r0i0n0 power_enable=True"

   For all vnodes:

   qmgr -c "set node @default power_enable=True"

3. Restart or send a SIGHUP to each MoM.

4. The resources_available.eoe on each vnode should show the configured power profiles. It may take up to two minutes for resources_available.eoe to be updated.

   You can check the values for resources_available.eoe:

   pbsnodes -a | grep resources_available.eoe

5. If, after two minutes, resources_available.eoe is not updated, restart or HUP each MoM again. If vnodes do not show the expected values for resources_available.eoe, we recommend checking the pbs_mom log file for errors, and double-checking your SMC configuration.

6. Make sure that each vnode you intend to use for running jobs under power profiles shows the expected values for resources_available.eoe.

3.5 New Attributes and Resources for Power Awareness

power_provisioning

Server attribute. When set to *True*, PBS can use power awareness.

Format: *Boolean*

Default: *unset*

Readable by all, settable by PBS Manager.

Python type: bool

power_enable

Vnode attribute. When set to *True*, this vnode can use power profiles.

Visible to all, settable by the administrator.

Format: *Boolean*

Default: *unset*

Readable by all, settable by PBS Manager.

Python type: bool

current_eoe

Vnode attribute. Shows the current value of eoe on the vnode.

Visible to all. Settable by manager. We do not recommend setting this attribute manually.

Format: *string*

Default: *unset*

Readable by all, settable by PBS Manager.

Python type: str

energy

Resource. Consumable. PBS records the job's energy usage in the job's resources_used.energy.

Format: *float*

Units: *kWh*

eoe

Resource. Stands for "Energy Operational Environment". Non-consumable. When set on a vnode in resources_available.eoe, contains the list of available power profiles. When set for a job in Resource_List.eoe, can contain at most one power profile. (A job can request only one power profile.)

Default value for resources_available.eoe: *unset*

Format: *string_array*

3.6 Reading and Setting Power-related Attributes and Resources in Hooks

3.6.1 Reading & Setting Vnode Attributes in Hooks

The following table lists the power-specific vnode attributes that can be read or set when the vnode object is retrieved via an event. An "*r*" indicates read, an "*s*" indicates set.

Vnode Attribute	queuejob	modifyjob (before run)	movejob	runjob	execjob_begin	execjob_attach	execjob_prologue	execjob_launch	execjob_end	execjob_epilogue	execjob_preterm	exechost_startup	exechost_periodic	provision
current_eoe	---	---	---	---	r, s	r	r, s	r	r, s	r, s	r, s	r, s	r, s	---
power_enable	---	---	---	---	r, s	r	r, s	r	r, s	r, s	r, s	r, s	r, s	---

3.6.2 Reading & Setting Job Resources in Hooks

3.6.2.1 Reading and Setting Resources Requested by Job

The following table shows the power-specific members of the job's Resource_List attribute that can be read or set in each type of hook, when retrieving the object through an event. An "*r*" indicates read, an "*s*" indicates set.

Resource in Resource_List	queuejob	modifyjob (before run)	movejob	runjob (on reject)	runjob (on accept)	resvsub	execjob_begin	execjob_attach	execjob_prologue	execjob_launch	execjob_end	execjob_epilogue	execjob_preterm	exechost_startup	exechost_periodic	provision
energy	r, s	r, s	r	r, s	r	---	r	r	r	r	r	r	r	---	r	---
eoe	r, s	r, s	r	r, s	r	---	r	r	r	r	r	r	r	---	r	---

3.6.2.2 Reading and Setting Resources Used by Job

The following table shows the power-specific members of the job's resources_used attribute that can be read or set in each type of hook, when retrieving the object through an event. An "*r*" indicates read, an "*s*" indicates set.

Resource in resources_used	queuejob	modifyjob (before run)	movejob	runjob (on reject)	runjob (on accept)	resvsub	execjob_begin	execjob_attach	execjob_prologue	execjob_launch	execjob_end	execjob_epilogue	execjob_preterm	exechost_startup	exechost_periodic	provision
energy	---	---	---	---	---	---	r, s	r, s	r, s	r, s	r, s	r, s	r, s	---	r, s	---
eoe	---	---	---	---	---	---	r, s	r, s	r, s	r, s	r, s	r, s	r, s	---	r, s	---

3.6.2.2.i Caveat for Setting Values for Used Resources

You can set a value for resources_used.<resource> in an execjob_end hook, but that value has no effect because the resource has already been reported to the server.

3.6.2.3 Reading Server Attributes in Hooks

All hooks have the same access to server attributes. You can read any server attribute that is not at its default value. If a server attibute is at its default value, an attempt to read the value returns "None". This is the same behavior as for `qstat -Bf`.

No hooks can set any server attributes.

3.7 Submitting Jobs with Power Profiles

3.7.1 Viewing Available Power Profiles

The admissible values for eoe are reported in the resources_available.eoe vnode resources. These values are only shown after PBS and SMC are successfully configured. To view available power profiles.

```
pbsnodes -a | grep -e "^\w" -e eoe
```

3.7.2 Submitting a Job Requesting a Power Profile

To submit a job requesting a specific power profile, request a value for eoe inside a select statement. For example:

```
qsub -lselect=4:eoe=high:ncpus=8 zoomjob
```

3.7.3 Caveats for Submitting Jobs Requesting Power Profiles

- A job can request at most one power profile.
- A job requesting a power profile must request an existing profile that is set on the number of required vnodes. If your job requests a non-existent profile, or requests more vnodes for the profile than are available, the job will remain queued. If your job unexpectedly

remains queued, you can get more information about why by asking `qstat` to show the job's comments:

```
qstat -s
```

For a complete description of the `qstat` command, see page 210 of the PBS Professional 13.0 Reference Guide, section 2.58, "qstat".

3.8 Caveats and Restrictions for Configuring Power Awareness

- You cannot use power profiles on any vnodes where the PBS server and/or scheduler are running.
- The scheduler will not preempt via suspension or checkpoint any job requesting a value for eoe. However, it will preempt these jobs via requeue.
- If you set power_provisioning to *False* while a job is running, the vnodes where the job runs do not have their profile deactivated when the job finishes, and the job's resources_used.energy value is not set at the end of the job.
- If a job does not request a value for eoe, there is no activation of a power profile on a node, but the job's resources_used.energy is still calculated.
- Energy consumption is measured and power profiles are set on a per-host basis. We recommend that you give each job exclusive use of its host(s) in order to get accurate energy use measurements and to avoid resetting the power profile of a host that is already running another job. For instructions on how to give a job exclusive use of a host, see page 92 of the PBS Professional 13.0 User's Guide, section 5.7, "Specifying Job Placement".

4 Issues

4.1 Open Bugs in PBS 13.1.500

The following bugs are open for this version of PBS:

Table 4-1: Open Bugs in PBS 13.1.500

Bug	Issue	Title
PBS-13522	341575	unset of power_provisioning does not update MOM PBS_power hook

Index

A

activate a power profile 11

C

current_eoe
- definition 14
- reading and setting in hooks 14

D

deactivate a power profile 11

E

energy
- definition 13
- reading and setting in hooks 15

eoe
- definition 14
- reading and setting in hooks 15

P

power profile
- activate 11
- deactivate 11

power_enable 12
- definition 13
- reading and setting in hooks 14

power_provisioning 12
- definition 13

prologue 12

pstate 16

R

Resource_List.eoe 14

resources_available.eoe 12, 14

resources_used.energy 13, 17

S

server attributes 16

Index