Intel® Cluster Ready 1.0
Intel® Server Board S5000PAL
Platform* Open Cluster Stack* version 4.5.1
Red Hat* Enterprise Linux* 4 Update 5
Configuration C1 (with Lava)

Version 1.2.1
7/25/2008

Legal Notices

Table of Contents

1. Hardware configuration

Required Hardware

Quantity	Item	Manufacturer	Model
32	Intel® server board S5000PAL	Intel®	S5000PAL
	Intel® Server Chassis	Intel®	SR1500
	Intel® HDD backplane	Intel®	ASR1550ASBP
	2 Intel® Xeon® Processors	Intel®	Quad-Core Intel® Xeon® Processor 5355 (Stepping 7)
	4x1GB DDR2 PC2-5300	Samsung*	M395T2953CZ4-CD501
	DVD/CDRW Slim line S1500	Intel®	AXXDVDCDR
	100Gb SATA Hard Disk Drive	Seagate*	Barracuda* ST3250620AS
	Infiniband* Host Channel Adapter (HCA)	Mellanox*	AXXIBDDRPT (MT25204) HW Version: a0 FW Version: 1.2.0
	Infiniband Cables	Gore*	C4X4-2M
1	A low latency gigabit Ethernet switch, for intra-node communications	Hewlett-Packard*	HPJ4904A
1	48 Port 4X DDR IB Switch	Flextronic*	F-X440077 (MT47396)
1	KVM over IP Solution	Avocent*	DSR8035

BIOS Settings

1) The required BIOS for the S5000PAL server board is S5000.86B.10.00.0088.031420081550.

This version consists of the following firmware components:

- BIOS version 88
- BMC version 63
- FRU version 47
- SDR version 47

To verify the version installed on each server, enter the BIOS setup screen by pressing <F2> when prompted during BIOS POST, navigate to System Management->System Information tab, and find the version information on that screen.
Please refer to the BIOS installation instructions on how to update the BIOS.

2) BIOS configuration requires some settings deviate from the default server configuration.

Execute the following steps to set proper configuration:

- Enter the BIOS setup screen by pressing <F2> during BIOS POST.
- Load default BIOS settings by pressing <F9>.
- Disable graphical splash screen by setting Main -> Quiet Boot to Disabled.
- Disable processor throttling by unsetting Advanced->Processor->Intel® Speedstep Technology.
- Set Server Management -> Console Redirection to Serial Port B. Default settings are enough.
- Set the boot order to DVD-Rom, SATA drive, IBA 500 (Intel® Boot Agent)
- Press <F10> to save and reboot.

These BIOS settings must be implemented precisely the same among all nodes of the cluster.

Remote Console Configuration

3) Using a KVMIP solution is one method for supplying remote console access. Connect and configure the KVMIP solution such that there is remote console access for each node in the cluster. Note: this recipe specifies an Avocent* KVM over IP solution. Other KVMIP solutions or serial-over-LAN solutions will also satisfy this requirement.

Cluster Configuration

4) There must be a frontend node plus compute nodes. All the switching equipment must interconnect all nodes.
5) Ethernet Port 1 (eth0) on all nodes must be connected to the private network. This is the messaging network for Ethernet as well as the management and storage network. No systems outside the cluster should be attached to this network.
6) Ethernet Port 2 (eth1) on the frontend node should be connected to the public network. This is point of entry to the cluster. Port 2 should remain disconnected on all compute nodes.
7) This recipe includes an Infiniband* fabric as a second messaging network. Each node must connect port 1 of its HCA to the Infiniband switch.

2. Getting started

Introduction

8) This cluster recipe provides the instructions for configuring Intel® servers with Platform* Open Cluster Stack* into a copy of a certified implementation of the Intel® Cluster Ready architecture specification.

Required Software Ingredients

Distributed By	Description	Contact Information
Intel® Corporation	Reference Implementation Package	http://www.intel.com/go/cluster
Intel® Corporation	Intel® Cluster Checker 1.1-018 Program registration is needed.	http://registrationcenter.intel.com/ l_clck_p_1.1.018.tgz md5 bfcc285c6102c1d7cb0c62c7be4390c7
Platform Computing*	Platform* OCS 4.5.1 Base Roll Site registration is needed.	http://my.platform.com/ ocs4.5.1_base.x86_64.disk1.rc4.iso md5 759a6a6a5f61608ea42a2a773579a30c
Platform Computing*	Platform* OCS 4.5.1 Intel Cluster Runtime Roll	http://my.platform.com/ intel_runtime-4.5.1-0.x86_64.disk1.iso md5 b4dd22345d0a02584c6f5da9b5bf2a03
Platform Computing*	Platform* OCS 4.5.1 Intel Cluster Checker Roll	http://my.platform.com/ intel_clusterchecker-1.1-008.x86_64.disk1.iso md5 40e9189b307f5a7abd82ce5e37364fab
Sun Microsystems*	Java* Runtime Environment 6u6. Linux RPM and Linux x64 RPM	http://www.java.com/en/download/manual.jsp jre-6u6-linux-i586-rpm.bin md5 4d6333d620c033bfedca9c83aed57688 jre-6u6-linux-x64-rpm.bin md5 dba89ac1bf7828380911fe5267876a0f
Mozilla*	Firefox* 2.0.0.15	http://www.mozilla.com/en-US/firefox/all-older.html firefox-2.0.0.15.tar.gz md5 f86ea1223794891d7ef893376854fc12
Open Fabrics Alliance*	Open Fabrics* Enterprise Distribution (OFED)	http://openfabrics.org OFED-1.3.1tgz md5 606787bf5a0c64091de92743c4eb7896

Collateral Documentation

- Intel Cluster Ready Program: http://www.intel.com/go/cluster
- Intel server documentation and firmware packages are available at http://www.intel.com/support/motherboards/server
- Platform® OCS documentation, available at http://www.platform.com/Products/platform-open-cluster-stack

3. Platform* OCS frontend installation

Preparation

9) Obtain Platform* OCS 4.5.1 rolls; the vendor should be contacted if needed in order to get the required media.

Install Platform* OCS on the Frontend

10) Insert the Platform* OCS installation disk into the DVD drive attached to the frontend.
11) Boot the System and configure BIOS settings if needed. At last boot from the DVD drive.
12) When the Platform* OCS splash screen appears, type frontend at the prompt. If nothing is entered within a timeout period the system will attempt to install the compute node software onto the system. The system should be rebooted to retry.
13) Accept the license to continue.

Select Platform* OCS Roll Packages

14) Some rolls can be distributed separately, and others may come altogether in the same installation disk. The list of required rolls and their versions is shown below.

The Intel® MPI Runtime roll bundled within Platform* Base roll shouldn't be installed as it is outdated as the Intel® Runtime roll provides updated versions of all needed libraries.

Also, the Intel Runtime and Intel Cluster Checker rolls don't come as part of the Platform* OCS 4.5.1 Base roll.

Roll	Version
Platform	4.5.1
Base	4.5.1
OS	4.5.1
Redhat	4.5.1
Kernel	4.5.1
Modules	4.5.1
HPC	4.5.1
NTop	4.5.1
Ganglia	4.5.1
Extras	4.5.1
Lava	1.0.0
Intel Vendor	4.5.1
Intel Runtime	4.5.1
Intel Cluster Checker	1.1

15) If any required roll is displayed as an option, it must be selected. Then, select OK to continue
16) Select "Yes" when prompted "Do you have another roll CD/DVD", and then enter the Intel Cluster Runtime Roll media in the CD/DVD drive of the frontend.
17) Repeat step 16) for the Intel Cluster Checker Roll CD-ROM.
18) At last, when prompted "Do you have another roll CD/DVD", select 'No' to continue with the installation.

Configure General Cluster Information

19) Add cluster information
- Hostname + Domain name, lowercase, e.g. frontend.localdomain
- Cluster name, automatically inserted from Hostname + Domain name entry
- Organization, automatically inserted from Hostname + Domain name entry
- Locality (city)
- State Country
- Contact
- URL
- LatLong

Set Disk Partitioning Scheme

20) Select Autopartition
21) When prompted, select OK to remove all partitions.
22) Set partition sizes as shown below
- Root: 20000
- Swap: 4000
- Var: 20000
- Export: <leave blank>

23) Select 'Fill up rest of free space'
24) Select 'No' when it asks if a partition review is needed

Configure Boot Loader

25) Select 'OK' to accept the default configuration for the boot loader.

Configure Ethernet

26) Eth0 configuration - This is the Ethernet connection to the private network. Enter the private IP and netmask here. For example, a common private IP is 192.168.1.1, with a netmask of 255.255.255.0.
27) Eth1 configuration - This is the Ethernet connection to the public network used to access the cluster. Enter the public IP and netmask here. It cannot be assigned dynamically using DHCP.
28) Miscellaneous network settings
- Gateway – add the desired gateway information.
- Primary DNS - It is required to have a valid Primary DNS entry. Errors may occur if this value is not configured.
- Secondary DNS - A secondary DNS entry is optional based on the network configuration.

Configure System Clock

29) Set time zone
30) Set NTP server – This can be a hostname or an IP address

Set Root Password

31) Enter the root password that will be shared to each node in the cluster.

Software Installation on the Frontend

32) The expected installation time is on the order of 30 minutes. During the install, you will see the Red Hat* installation screen, including status bars and time remaining. Due to the way Platform* OCS calculates remaining time and bytes, these status marks will go into negative numbers. This is normal and installation will complete properly. Allow the software installation to complete on the frontend node. The system will reboot automatically when the installation completes.
33) After the frontend reboots, log in to the system as root.
34) Upon first login the system will prompt for ssh configuration setup. Accept the default responses for each ssh setup question.

Any missing roll can be added before customization using the 'rollops --add' command after inserting the media.

All the software listed on the software bill of materials is assumed to be present on the /root directory. It is strongly recommended to double check software checksums using the md5sum command; this will avoid delays due to transfer errors.

Remote Console Configuration

To fulfill remote console requirements serial consoles may be used.

Changes on the kernel boot parameters and the init scripts are required; this will enable console output during the boot process and also establish a serial terminal for remote connections.

During the commands shown below specific serial port settings are assumed, these can vary according to the actual hardware being used.

35) Configure the console through the serial port on the frontend.

```
sed –i –e 's/\(.*\quiet\)/\1 console=ttyS1,115200/'
/boot/grub/grub.conf
```

36) Configure the console through the serial port in the kernel parameters of the nodes. Init scripts will be modified automatically during provisioning.

```
sed –i –e 's/\(append.*\)/\1 console=ttyS1,115200/'
/tftpboot/pxelinux/pxelinux.cfg/default
```

Infiniband* Installation

37) Temporarily fix conflicting dependencies, the packages will be reinstalled after OFED* deployment.

```
rpm –e openmpi_pcc hpl
```

38) Extract the OFED* package to a temporary location

```
tar -zxf OFED-1.3.1.tgz -C /tmp
cd /tmp/OFED-1.3.1
```

Due to dependency issues appearing during compute node provisioning; only the HPC subset of packages will be installed on the compute nodes. The full set of packages will be then installed only on the frontend node.

39) Compile and install the High Performance Computing subset of OFED* packages

```
./install.pl --without-depcheck --hpc
```

40) Add OFED* packages to the compute node distribution

```
cd RPMS/redhat-release-4AS-6.1/x86_64/
for i in `ls *.rpm`; do rocks-compute -a -p $i; done
```

41) Rebuild the compute node distribution bundle

```
cd /home/install; rocks-dist dist
```

42) To complete the OFED* installation of the frontend, the debug packages needs to be excluded. Generate the ofed.conf file and remove the debuginfo packages.

```
cd /tmp/OFED-1.3.1
./install.pl --print-available
sed -i -e 's/.*debuginfo=y//' ofed.conf
./install.pl --without-depcheck --config ofed.conf
cd
```

43) Clean up after installation, temporary files should be removed.

```
rm -fr /tmp/OFED-1.3.1
rpm -ivh /home/install/rocks-dist/lan/x86_64/RedHat/RPMS/openmpi_pcc-
1.2.4-1.x86_64.rpm --nodeps
rpm -ivh /home/install/rocks-dist/lan/x86_64/RedHat/RPMS/hpl-ia32e-
1.0a-0.ia32e.rpm
```

Infiniband* Configuration

44) Configure the subnet manager to launch on boot

```
chkconfig --level 35 opensmd on
```

45) Edit the file /opt/rocks/etc/ofed.cfg to add a base address and netmask for the Infiniband network.

```
echo "IBnetwork0 = 192.168.2.0" >> /opt/rocks/etc/ofed.cfg
echo "PrivateNetmask = 255.255.255.0" >> /opt/rocks/etc/ofed.cfg
```

For example, to configure port 1 on your card, the following lines are required.

```
IBnetwork0 = 192.168.2.0
PrivateNetmask = 255.255.255.0
```

46) Restart openib and opensm services, and run the configuration script for Infiniband

```
/etc/init.d/openibd restart
/etc/init.d/opensmd restart
ofed-ipoib-config -a
```

Infiniband* Device Configuration

To enable full compatibility with pre existent applications, the /etc/dat.conf file describing OFED* interfaces should be customized. Only one device description should remain.

47) Create a script called ~/ofed.sh and add the following to it.

```
#!/bin/bash
echo 'OpenIB-cma u1.2 nonthreadsafe default libdaplcma.so dapl.1.2 "ib0
0" ""' > /etc/dat.conf
```

48) Execute the script on the frontend node

```
source ~/ofed.sh
```

49) Add the script to the compute node image.

```
rocks-compute -a -s ~/ofed.sh
```

Security Options

50) In order for RDMA based connections, such as InfiniBand*, to work, the memlock limits have to be increased from the default. The script below increases the memory limits to 2GB for hard and soft memlock limits. However, the hard memlock should never be greater than the amount of physical memory on the system, and the soft memlock limit should never be greater than the hard limit. Create a script called ~/limits.sh and add the following to it:

```
#!/bin/bash
echo "*     soft   memlock 2000000" >> /etc/security/limits.conf
echo "*     hard   memlock 2000000" >> /etc/security/limits.conf
```

51) Run the script on the frontend node.

```
source ~/limits.sh
```

52) Add the script to the compute node image.

```
rocks-compute -a -s ~/limits.sh
```

Java* and Java-enabled Web Browser Installation

A web browser with a Java plugin installed is required on the frontend. In addition, all nodes must install the x86_64 Java Runtime Environment. The already installed version will be replaced for enabling plug-in compatibility.

53) Install the 32-bit version of Firefox, by untarring the package into the /usr/lib directory.

```
cd ; tar -C /usr/lib/ -xzf firefox-2.0.0.15.tar.gz
```

54) Set the new Firefox as default, keeping the old one available

```
mv /usr/bin/firefox /usr/bin/firefox_old
ln -s /usr/lib/firefox/firefox /usr/bin/firefox
```

55) Install the 32-bit version of Java Runtime Environment 6u6. Execute it only to extract the file; uninstall the package to avoid conflicts with the 64 bit version. The extracted 32 bit version will be installed in a different path.

```
./jre-6u6-linux-i586-rpm.bin
rpm -e jre
```

56) Install the 64-bit version of Java Runtime Environment 6u6. Accept the license to extract the RPM and to install it on the frontend node.

```
./jre-6u6-linux-x64-rpm.bin
```

57) Install the 32-bit version of Java avoiding conflicts with the 64 bit version of the package.

```
rpm -ivh --relocate /usr/java=/usr/java/i386 jre-6u6-linux-i586.rpm
```

58) Create a symbolic link for the Firefox plugin

```
ln -s /usr/java/i386/jre1.6.0_06/plugin/i386/ns7/libjavaplugin_oji.so
/usr/lib/firefox/plugins
```

59) The extracted rpm name must be modified for the OCS installation procedures in order to work.

```
mv ~/jre-6u6-linux-amd64.rpm ~/jre-1.6.0_06-fcs.x86_64.rpm
```

60) Add the 64-bit version of Java to the distribution image for clients. Hit enter to rebuild the distribution when prompted.

```
rocks-compute -a -p jre-1.6.0_06-fcs.x86_64.rpm
```

61) Create a script called ~/java.sh and add the following to it:

```
#!/bin/bash
ln -s /usr/java/jre1.6.0_06/bin/java /usr/local/bin/java
```

62) Run the script on the frontend node.

```
source ~/java.sh
```

63) Add the script to the distribution image.

```
rocks-compute -a -s ~/java.sh
```

Intel® Cluster Ready Version Information

This cluster implementation must contain a text file specifying that this implementation complies with the Intel® Cluster Ready specification version 1.0.

64) Create a script called ~/icr.sh with the following content:

```
#!/bin/bash
mkdir /etc/intel
echo "CLUSTER_READY_VERSION=1.0" >  /etc/intel/icr
```

65) Run the script on the frontend node.

```
source ~/icr.sh
```

66) Add the script to the distribution image. Hit enter to rebuild the distribution when prompted.

```
rocks-compute -a -s ~/icr.sh
```

67) Force the rebuild of the compute node image

```
cd /home/install; rocks-dist dist; cd
```

E1000 driver options

Turning off interrupt control decreases latency and speeds up synchronization. Also, the IRQ balancing daemon can be disabled to minimize interference.

As the driver module is prepackaged into the system's initial disk image; this image will be regenerated.

These changes are only needed on the private Ethernet device.

68) Create a script called ~/e1000.sh and add the following to it.

```
#!/bin/bash
chkconfig irqbalance off
echo "options e1000 InterruptThrottleRate=0" >> /etc/modprobe.conf
mkinitrd --with=e1000 -f /boot/initrd-2.6.9-55.0.12.ELsmp.img 2.6.9-
55.0.12.ELsmp
modprobe -r e1000; modprobe e1000
```

69) Run the script on the frontend node.

```
source ~/e1000.sh
```

70) Add the script to the compute node image.

```
rocks-compute -a -s ~/e1000.sh -b
```

71) Hit enter to rebuild the distribution when prompted.

72) Reboot the head node to fully enable all the customization changes

4. Compute node installation

Build Compute Nodes

Please refer to the Platform* OCS documentation for detailed instructions on how to run the 'insert-ethers' command. Please also refer to the board documentation on how to boot into PXE mode.

73) From a root prompt on the frontend, run

```
insert-ethers --appliance compute
```

74) For each compute node,

- Boot the machine
- Press <F2> when prompted for the setup menu
- Select Boot Manager, then the IBA device for network boot.
- Watch the frontend console for indication that the compute node installation has begun (an asterisk will appear). When the asterisk appears, begin installing the next compute node. If finished, allow all the compute nodes to complete installation and proceed to the next step.

When an asterisk appears next to each compute node in the insert-ethers list on the frontend, exit the insert-ethers program by pressing <F9>. Allow the compute node software to be installed on all nodes. Estimated time of completion is 15 minutes per server.

Due to PXE limitations, the provisioning of more than 8 nodes in parallel is not recommended. If the procedure fails, rebooting the machine and selecting network boot again will restart the process.

5. Verify a Correct Cluster Build

Please see the documentation that comes with the Intel® Cluster Checker Tool for detailed instructions on configuring and running the tool.

The old version of Cluster Checker tool will be replaced with the latest in order to take advantage of its new features. The installation of the Intel Cluster Checker roll can not be avoided because it installs as a side effect dependencies required by the Intel Cluster Ready specification.

Intel® Cluster Checker Installation

75) Place a proper license at /opt/intel/licenses. The license file will be requested to allow tool execution.

```
cp ~/*.lic /opt/intel/licenses
chmod o+r /opt/intel/licenses/*.lic
```

76) Unpack and install Intel Cluster Checker. Choose '1' to uninstall the previous version; latest version will be automatically installed.

```
tar -xzf l_clck_p_1.1.018.tgz -C /tmp
cd /tmp/l_clck_p_1.1.018
./install.sh
```

77) Clean up temporary files

```
cd
rm -fr /tmp/l_clck_p_1.1.018
```

Intel® Cluster Checker Configuration

78) Create a user account on the frontend to run the tool.

```
useradd clck
passwd clck
```

79) Switch from root to the clck user

```
su - clck
```

80) Create the mpd configuration file required to use the Intel® MPI Library and assign the proper permissions to the file.

```
echo secretword=anything > ~/.mpd.conf
echo password=anything >> ~/.mpd.conf
chmod 600 ~/.mpd.conf
```

81) Enable user execution of OFED* command tools

```
echo 'export PATH=$PATH:/usr/sbin' >> ~/.bashrc
```

82) Create a directory to store configuration files and run the Intel® Cluster Checker tool

```
mkdir clck_results
```

83) Place the Intel Cluster Checker configuration files (XML and checksums) from the Reference Implementation Package in the clck_results directory.

84) Create a node list file and place it in the clck_results directory. This is a file containing the list of nodes to validate. Name the file "nodelist". The first line must be the hostname of the frontend and specified as the frontend node using "# head".

```
frontend-hostname # head
compute-0-0
compute-0-1
…
compute-0-N
```

85) Edit the user XML configuration file. Edit the <nodefile> entry to reflect the location of the nodelist file created before.

86) Edit the root XML configuration file. Edit the <compute_node> and <head_node> entries for the <copy_exactly> test module to reflect the location of the checksum files saved before. Edit the <nodefile> entry to reflect the location of the nodelist file created before.

87) Exit out of the clck account

```
exit
```

88) Synchronize the configuration files in the compute nodes

```
cluster-fork 411get --all
```

Intel® Cluster Checker Execution

89) Run Intel® Cluster Checker as root

```
source /opt/intel/clck/1.1/clckvars.sh
/opt/intel/clck/1.1/cluster-check S5000PAL-ICR1.0-POCS-RH4-C1-root.xml
```

90) Switch to the clck account.

```
su - clck
```

91) Change to the clck_results directory.

```
cd clck_results
```

92) Run Intel® Cluster Checker as user clck. The output will be displayed on screen as well as saved to the current directory

```
source /opt/intel/clck/1.1/clckvars.sh
/opt/intel/clck/1.1/cluster-check S5000PAL-ICR1.0-POCS-RH4-C1-user.xml
```

Output Analysis

93) Verify Intel® Cluster Checker reports successful cluster build. Look at the final line in the output, either on screen or in the created file. The last line should read "Check has Succeeded."

6. Release Notes

- The OFED 1.3.1 package has a known issue in which a harmless warning is displayed when installing the kernel-ib package. This warning is caused by a missing dependency against the pci-utils package used to locate available ports on the Infiniband devices.