Intel® Cluster Ready 1.1
Intel® Server Board S5400SF
Clustercorp* Rocks+ * V
Red Hat* Enterprise Linux 5 Update 1
Configuration C1 (Default)

Version 1.0
10/28/2008

Legal Notices

Table of Contents

1. Hardware configuration

Required Hardware Ingredients

Quantity	Item	Manufacturer	Model
5	Intel® server board S5400SF	Intel®	S5400SF
	Intel® Server Chassis	Intel®	SR1560
	Intel® HDD backplane	Intel®	ASR1500PASBP
	2 Intel® Xeon® Processors	Intel®	Quad-Core Intel® Xeon® Processor X5355 Stepping 7
	4x1GB DDR2 PC2-5300	Intel®	MT9HTF12872FY
	DVD/CDRW Slimeline SR1550/SR1560	Intel®	AXXDVDCDR
	250Gb SATA Hard Disk Drive	Seagate*	Barracuda* ST3250620AS
	Infiniband* Host Channel Adapter (HCA)	Mellanox*	MHGS18-XTC HW Version: a0 FW Version: 1.2.0
	Infiniband cabling for each system	Gore*	C4X4-2M
1	A low latency gigabit Ethernet switch, for intra-node communications	Hewlett-Packard*	ProCurve* J4904A 2848
1	48 Port 4X DDR Infiniband Switch	Flextronic®	F-X440077
1	KVM over IP Solution	Avocent	DSR1031

Table 1. Hardware Ingredients

BIOS Settings

1) The required firmware for the S5400SF server board is the following:

- BIOS version S5400.86B.06.00.0027.062520080920.
- BMC version 8
- FRU version 11
- SDR version 11

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; documentation and firmware packages are available at http://www.intel.com/support/motherboards/server.

2) BIOS configuration requires some settings deviate from the default server configuration. Execute the following steps to set the 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 Configuration->Enhanced Intel® Speedstep Technology.

- Set SATA configuration to AHCI by selecting Advanced -> SATA Controller Configuration -> Configure SATA as -> AHCI.
- Set Server Management -> Console Redirection to Serial Port B. Leave the remaining settings in their default value.
- On the frontend node set the boot order to DVD-Rom, SATA drive, IBA 600 (Intel® Boot Agent). On the compute nodes set the boot order to IBA 600 (Intel® Boot Agent), SATA drive.
- Press <F10> to save and reboot.

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

Remote console configuration / KVMIP

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

There must be a frontend node in addition to the compute nodes. The frontend node is the only node connected to the public network and acts as a gateway between the user and the compute nodes. The frontend node is connected to all the compute nodes through a private network.

4) Ethernet Port 1 (eth0) on all nodes must be connected to the private network. Used for either management, storage or messaging network. No systems outside the cluster should be attached to this network.
5) Ethernet Port 2 (eth1) on the frontend node should be connected to the public network. This is the entry point to the cluster. Port 2 should remain disconnected on all compute nodes.
6) 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

This cluster reference implementation provides the instructions for configuring Intel® servers with Clustercorp* Rocks+ V* into a copy of a certified implementation of the Intel® Cluster Ready architecture specification. The recipe is not intended to substitute the manuals of the components. For more information, definitions or acronyms refer to the proper documentation.

Required Software Ingredients

Distributed By	Description	Contact Information
Intel® Corporation	Reference Implementation Package	http://www.intel.com/go/cluster
Clustercorp*	Clustercorp* Rocks+* V Base Roll	http://www.clustercorp.com/rocksplus area51+base+cisco- ofed+dell+ganglia+hpc+java+kernel+kernel- update+service-pack+support+torque+web- server+xen-03.09.2008- 14.06.46.x86_64.disk1.iso 2a0b1773dfdb858945973b1106b15dbf
Clustercorp*	Clustercorp* Rocks+ V Intel® Cluster Ready Roll. Intel® Cluster Checker 1.2 requires Program registration.	intel-icr-5.0-1.2.x86_64.disk1.iso 651db67098aadc0b0014e71d83b26d92
Red Hat*	Red Hat* Enterprise Linux 5 Update 1	https://www.redhat.com/apps/download/ RHEL5.1-Server-20071017.0-x86_64- DVD.iso md5: 218ba37c78f1b57883d955013c4ef8a1

Collateral Documentation

- Documentation about the Intel® Cluster Ready Program is available at http://www.intel.com/go/cluster
- Documentation about Rocks* 5.0 is available at http://www.rocksclusters.org.
- Documentation about Clustercorp* Rocks+* V is available at http://www.clustercorp.com.
- Documentation about Red Hat* Entreprise Linux 5.1 is available at www.redhat.com.
- HPCC benchmark http://icl.cs.utk.edu/hpcc/index.html.

3. Frontend Installation

Starting Installation

7) To obtain the required Clustercorp* rolls the vendor should be contacted at http://www.clustercorp.com/rocksplus.
8) Insert the Base Roll disk into the DVD drive attached to the frontend and boot the system.
9) When the Rocks* splash screen appears, type "frontend" at the prompt. NOTE: Refer to section 7-Known issues before proceeding.

If nothing is entered within a timeout period, the system will assume that it is a compute node installation. If this happens, the system should be rebooted to retry.

10) On the 'Configure TCP/IP' screen, select manual configuration for IPv4 and deselect IPv6 support. Select 'OK' to continue.
11) On the 'Manual TCP/IP Configuration' screen, enter the IPv4 address and its netmask of the public network. A gateway and a name server should be also entered if required.

The system has an expiration timeout. If these steps are not completed in time, the system will reboot itself and the procedure must be restarted.

Configuring Installation

12) On the 'Welcome to Rocks' screen, click on 'CD/DVD-based Roll' and select all options except the 'dell' and 'torque' rolls. Click on 'Submit'

The rolls that must be selected are area51, base, cisco-ofed, ganglia, hpc, java, kernel, kernel-update, service-pack, support, web-server, xen.

13) Back into the 'Welcome to Rocks' screen, click on 'CD/DVD-based Roll' then insert the Intel® Cluster Ready roll media and click on 'Continue' . Select the Intel® Cluster Ready roll. Manually inserting the roll after completing the installation will not complete the requirements.
14) Back into the 'Welcome to Rocks' screen, repeat previous step with the Red Hat* Media.
15) On the 'Welcome to Rocks' screen, click on 'Next'.
16) On the 'Cluster Information' screen, enter the fully qualified host name and complete the basic cluster information as required. Click on 'Next'.
17) On the 'Ethernet Configuration for eth0' screen, enter the network address and netmask of the frontend system into the private network. Click on 'Next'.

For instance, the frontend private address may be 192.168.1.1 and its netmask 255.255.255.0.

18) On the 'Ethernet Configuration for eth1' screen, enter the network address and netmask of the frontend system into the public network. Click on 'Next'.

For instance, the frontend public address may be 10.0.0.2, its netmask 255.255.255.0 and the gateway 10.0.0.1. These values must match the underlying network environment.

19) On the 'Miscellaneous Network Settings' screen, enter the gateway and DNS server values if required. Click on 'Next'.
20) On the 'Root Password' screen, enter and confirm the admin user password. Click on 'Next'.
21) On the 'Time Configuration', enter proper Time Zone and NTP server. Click on 'Next'.

Disk Partitioning

The default sizes of partitions are not enough; a more appropriate scheme will be used instead.

22) On the 'Disk Partitioning' screen, select 'Manual Partitioning'. Click on 'Next'.
23) Remove all the partitions (including an extended one if present) pressing 'Delete' and define new partitions pressing 'New' as shown below. After defining all partitions, click 'Next' to begin the partitioning.

- 100MB ext2 partition mounted at /boot, fixed size.
- 4096MB swap partition, fixed size.
- 20480MB ext3 partition mounted at /, fixed size.
- 20480MB ext3 partition mounted at /var, fixed size.
- The remaining space should be used for an ext3 partition mounted at /state/partition1, selected to 'Fill to maximum allowable size'.

24) The installer will require the selected rolls during installation. Insert the Base and ICR roll when required.

The system will be automatically rebooted when completed; the expected installation time is about 45 minutes.

4. Frontend Customization

Starting Customization

25) After the frontend reboots, log into the system as 'root'.
26) Open a console terminal and press just 'enter' to answer the questions about SSH configuration. This will enable password-less logins across the cluster.
27) The following files should be copied to the /root directory (check the Intel® Cluster Ready website). Verify Checksums before proceeding with the installation to avoid file corruption problems. In addition, verify the file's line termination format, use "dos2unix" when needed:
    a. Intel® Cluster Checker License
    b. Intel® Cluster Checker XML configuration file
    c. Checksums for Frontend and Compute nodes

Infiniband* Support

To fully enable Infiniband* support, the subnet manager and the IP over IB network support need to be configured.

28) Start the OFED* subnet manager, also setup the service to start at boot time.

```
service opensmd start
chkconfig opensmd on
```

29) Setup the IP over IB interface on the frontend to be enabled at boot time.

```
sed -i -e 's/ONBOOT=no/ONBOOT=yes/' /etc/sysconfig/network-
scripts/ifcfg-ib0
```

30) Complete the IP over IB network configuration; in the example below, the subnet address follows the default settings already bundled inside the Base roll. Harmless warnings will be shown regarding missing default values on the plug-in implementation.

```
rocks add network ipoib subnet=172.30.0.0 netmask=255.255.0.0
```

Enabling Intel® Cluster Ready Software

31) Verify the Intel® Cluster Ready specification compliance. Expected output CLUSTER_READY_VERSION=1.1

```
cat /etc/intel/icr
```

32) Copy Intel® Cluster Checker license to the default Intel® licenses directory.

```
cp /root/*.lic /opt/intel/licenses
chmod a+r /opt/intel/licenses/*.lic
```

33) Verify the version of Intel® Cluster Checker tool. The expected result is 1.2

```
/opt/intel/clck/1.2/cluster-check --version
```

34) Create a directory in the /etc/intel folder to hold Intel® Cluster Checker configuration files and give permissions for the 'icr' user to access it.

```
mkdir /etc/intel/clck
chmod 777 /etc/intel/clck
```

35) Place the Intel® Cluster Checker configuration files (XML and checksums) from the Reference Implementation Package in above created directory and give permissions for the 'icr' user to access them.

```
cp S5400SF-ICR1.1-ROCKS+-RH-C1-config.xml /etc/intel/clck
cp S5400SF-ICR1.1-ROCKS+-RH-C1-frontend.chk /etc/intel/clck
cp S5400SF-ICR1.1-ROCKS+-RH-C1-compute-node.chk /etc/intel/clck
chmod 666 /etc/intel/clck/*
```

36) Set a password for the 'icr' user account and switch to enable ssh password-less logins just pressing 'enter'. It will be used to run Intel® Cluster Checker.

```
passwd icr
su - icr
```

37) Enable user execution of OFED* command tools.

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

38) Create a directory to store configuration files and Intel® Cluster Checker logs

```
mkdir clck_results
```

Frontend Compliance Validation

Before provisioning the nodes, only the frontend will be validated. Note that localhost is added to nodelist file like a fake node to let Intel® Cluster Checker validate minimum functional requirements.

39) Create a temporary nodelist file. Example contents and commands on how to generate this file are shown below.

```
server # head
localhost
```

```
echo -n $HOSTNAME | cut -d. -f1 > /etc/intel/clck/nodelist
sed -i -e 's/\(.*\)/\1 # head/' /etc/intel/clck/nodelist
echo localhost >> /etc/intel/clck/nodelist
```

40) Launch Intel® Cluster Checker as user to validate minimum functional requirements. Expect the 'cluster_size' test to fail.

```
cd clck_results
source /opt/intel/clck/1.2/clckvars.sh
/opt/intel/clck/1.2/cluster-check /etc/intel/clck/S5400SF-ICR1.1-
ROCKS+-RH-C1-config.xml --compliance 1.1
```

41) Logout from the 'icr' account.

```
exit
```

42) Launch Intel® Cluster Checker as 'root' to validate minimum system requirements. Expect the cluster_size module to fail. The hdparm module may fail also as the same server is acting as both frontend and compute node.

```
cd ~icr/clck_results
source /opt/intel/clck/1.2/clckvars.sh
/opt/intel/clck/1.2/cluster-check /etc/intel/clck/S5400SF-ICR1.1-
ROCKS+-RH-C1-config.xml --compliance 1.1
cd
```

5. Compute Node Installation

Compute Node Provisioning

43) Start the provisioning service on the frontend. A status screen will appear to show detected compute nodes. Note: the command below must be executed at a Linux terminal on graphic environment. If other console is used (serial, ssh, etc) the process will hang .

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

44) Boot the compute nodes on groups of 8 nodes (at maximum) simultaneously to avoid network booting conflicts. Ensure that network boot is the first boot option in the BIOS configuration of each compute node. On the frontend wait until and * symbol appears next to each node entry and then press F8.

Important: Rocks+ only allows graphical installation. So, redirecting the console to the serial port will result in errors during the compute nodes provisioning.

45) Wait for the provisioning to finish, the expected total installation time is about 30 minutes. Execute the command below and expect each node hostname as the output.

```
cluster-fork hostname
```

The provisioning procedure may fail if all the compute nodes are booted at the same time. Rebooting the affected system will re-start the process if required.

6. Cluster Validation

Compute Node Validation

46) Force user account distribution. All compute nodes must report 'stat:0'.

```
rocks sync users
```

47) Log in as the 'icr' user. Double check the access to the compute nodes.

```
su - icr
cluster-fork hostname
```

48) Update the "nodelist" file with the nodes file generated by "insert-ehters".

```
cp /tmp/nodes /etc/intel/clck/nodelist
```

49) Change to the clck_results directory.

```
cd clck_results
```

50) Launch a validation using Intel® Cluster Checker as "icr" user. All tests should pass.

```
source /opt/intel/clck/1.2/clckvars.sh
/opt/intel/clck/1.2/cluster-check /etc/intel/clck/S5400SF-ICR1.1-
ROCKS+-RH-C1-config.xml --compliance=1.1
```

51) Run Intel® Cluster Checker in wellness mode as "icr" user.
Important: In order to obtain the performance reference values that appear in the Intel® Cluster Checker XML configuration file, it is necessary to modify its default configuration. See the section 8-Performance Considerations.

```
source /opt/intel/clck/1.2/clckvars.sh
/opt/intel/clck/1.2/cluster-check /etc/intel/clck/S5400SF-ICR1.1-
ROCKS+-RH-C1-config.xml --level 5 --exclude copy_exactly
```

52) Log out of the clck account

```
exit
```

53) Change to the clck_results directory

```
cd /home/icr/clck_results
```

54) Run Intel® Cluster Checker in wellness mode as root. All tests should pass

```
source /opt/intel/clck/1.2/clckvars.sh
/opt/intel/clck/1.2/cluster-check /etc/intel/clck/S5400SF-ICR1.1-
ROCKS+-RH-C1-config.xml --level 4 --include only copy exactly --
include_only dmidecode --include_only hdparm
```

55) Verify that 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 say "Check has Succeeded"

7. Known Issues

Booting from an external USB device may cause installer crash

Due to a race condition between the USB device settlement and the USB driver discovery time, using external USB devices may cause the frontend installation to break. A workaround for this issue is shown below.

56) At boot time, provide an invalid device to wait for USB settling and force device detection.

```
frontend ks=invalid
```

57) At the 'Error downloading kickstart file' screen, re-start the provisioning process.

```
cdrom:/ks.cfg
```

8. Performance Considerations

Default values of Ns and NBs factors for HPCC benchmark

Intel® Cluster Checker installation does not set the HPCC benchmark configuration factors for any specific platform. It is rather a generic installation with default values. Therefore, we recommend to set the configuration factors to better suit the hardware used in this recipe. Login as root to make the changes.

58) Change the default values.

```
sed -i -e 's/^[0-9]*[ ]*Ns$/18500\t\tNs/'
/opt/intel/clck/1.2/external/hpccinf.txt
sed -i -e 's/^[0-9]*[ ]*NBs$/168\t\tNBs/'
/opt/intel/clck/1.2/external/hpccinf.txt
```

Default configuration on Ethernet latency

The Ethernet latency can be decreased disabling the firewall and the IRQ balancing daemon in systems with many processors. This tuning will also minimize latency variations, enhancing the compute node responsiveness when handling small synchronization messages.

59) Disable the firewall on all nodes.

```
chkconfig iptables off
cluster-fork 'chkconfig iptables off'
```

60) Disable the IRQ balancing daemon on all nodes.

```
chkconfig irqbalance off
cluster-fork 'chkconfig irqbalance off'
```

61) Restart the frontend and the compute nodes. Just stopping those services won't clear the network stack properly.

```
cluster-fork reboot
reboot
```