Intel® Cluster Ready 1.1
Intel® Server Board S5520UR
Clustercorp* Rocks+ * V.I
Red Hat* Enterprise Linux 5 Update 3
Configuration C1 (Default)

Version 1.2
1/27/2010

Legal Notices

Table of Contents

1. Hardware configuration

Required Hardware ingredients

Quantity	Item	Manufacturer	Model
33	Intel® server board Urbanna	Intel®	S5520UR
	Intel® server chassis	Intel®	SR1600URSASBPP
	Intel® HDD backplane	Intel®	ASR1600PASBP
	2 Intel® Xeon® Processors	Intel®	X5570 @ 2.93GHz Stepping: 4
	6x 2GB DDR3 PC3-10600	Micron	MT18JSF25672PDZ-1G4D1
	500Gb SATA Hard Disk Drive 3 Gbs	Seagate*	Barracuda* ST3500320NS
	DVD/CDRW – Slimeline SR1550/SR1560	Intel®	AXXDVDCDR
	ConnectX IB - Dual-Port InfiniBand Adapter Card	Mellanox*	MHGH29-XTC Hw Revision: a0 Fw Version: 2.6.000
1	A low latency gigabit Ethernet switch.	Hewlett-Packard*	ProCurve* J4904A
1	144 Port DDR InfiniBand Switch.	Flextronic®	F-X440044 Base on MT47396 Infiniscale-III Mellanox Technologies Fw Version: 1.0.0
33	Infiniband cabling	Gore*	C4X4-2M
1	KVM over IP Solution	Avocent	DSR8035

BIOS Settings

1) The required firmware (S5500.86B.01.00.0042.090420091227) components for the S5520UR server board are:

- BIOS version 42
- BMC version 0.43
- ME version 1.10
- FRU/SDR version 0.18

To verify the version installed on each server, enter the BIOS setup screen by pressing <F2> when prompted during BIOS POST, then 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 deviated 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->Intel® Speedstep Technology.
- Select Server Management -> Console Redirection. Select Console Redirection and choose "Serial Port A". Default settings are enough.
- Installer node:

    a. Set the boot order to SATA3:TSSTcorp CDDVDW, #0500 ID01 LUN ATA, IBA GE Slot0100 (Intel® Boot Agent)
- Compute nodes:
    a. Set the boot order to IBA GE Slot0100, #0500 ID01 LUN ATA drive.
- Press <F10> to save and reboot.

These BIOS settings must be set identically on 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 to each node in the cluster. Note: this recipe requires an Avocent KVM over IP solution. However, any other KVMIP or serial-over-LAN solutions will satisfy this requirement.

Cluster configuration

4) There must be a separate head node plus the 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 connected to this network.
6) Ethernet Port 2 (eth1) on the Node Installer should be connected to the public network. This is the entry point 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 (ib0) of its HCA to the Infiniband switch.
8) Serial port on the compute nodes should be connected to a terminal server if the terminal server is used for remote console logging on compute nodes.

2. Getting started

Introduction

This cluster reference implementation provides the instructions for configuring Intel® servers with Clustercorp* Rocks+ V.I* 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 S5520UR-ICR1.1-ROCKSPLUS5.1-RH5.3-C1-v1.2-config.xml ICR1.1-ROCKSPLUS5.1-RH5.3-C1-v1.2-head.list ICR1.1-ROCKSPLUS5.1-RH5.3-C1-v1.2-compute-node.list ICR1.1-ROCKSPLUS5.1-RH5.3-C1-v1.2.tar.gz ICR1.1-ROCKSPLUS5.1-RH5.3-C1-v1.2.tar.gz.md5
Clustercorp*	Rocks+ V.I* version 5.1 Base Roll Ganglia Roll HPC Roll Intel-ICR Roll Java Roll Kernel Roll Mellanox OFED Roll Support Roll Torque Roll Web Server Roll Inside Intel-ICR Roll: - Intel® Cluster Checker 1.3u2 - Intel® Cluster Runtime 2.1-2 Program registration is needed.	http://www.clustercorp.com/rocksplus base+ganglia+hpc+intel-icr+intel-icr-mfg+java+kernel+mlnx-ofed+support+torque+web-server-10.09.2009-14.10.42.x86_64.disk1.iso md5: 9ce2c1a3709bd614806d11b1644af008
Clustercorp*	Intel® Driver Roll version 5.1	http://www.clustercorp.com/rocksplus intel-driver-5.1-1.1.x86_64.disk1.iso md5: 0a4c78358004c3b24f1ca36586775cc0
Red Hat*	Red Hat* Enterprise Linux 5 Update 3	https://www.redhat.com/apps/download/ md5: c5ed6b284410f4d8212cafc78fd7a8c5

Collateral Documentation

- Intel® Cluster Ready Program: http://www.intel.com/go/cluster
- Intel® Server Boards: http://www.intel.com/products/server/motherboard
- Clustercorp* Rocks+ V.I* Provisioning System Website: http://www.clustercorp.com.
- Red Hat* Operating System Website: http://www.redhat.com.

3. Clustercorp* Rocks+ Frontend Installation

Preparation

Install the BIOS package according to the instructions supplied with the package, and ensure proper BIOS settings.
Obtain Clustercorp* Rocks+* V.I distribution. Please contact the vendor in order to get the required media. Prepare and be ready to provide the following data before continuing the deployment below - full domain name, fixed Head Ip/netmask, private IP/netmask, gateway, DNS, NTP.

Install Clustercorp* Rocks+* on the frontend

9) Insert the Base Roll disk into the DVD drive attached to the frontend and boot the system.
10) When the Clustercorp* Rocks+* splash screen appears, type "build" at the prompt and press <enter>.

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 restart the frontend node installation.

Configuring Installation

11) On the 'Configure TCP/IP' screen

Select manual configuration for IPv4 and deselect IPv6 support, press <OK> to continue.

12) 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.

13) On the 'Welcome to Rocks' screen

Press <CD/DVD-based Roll> and select the following rolls, then press <Submit>

- base
- ganglia
- hpc
- intel-icr
- java
- kernel
- mlnx-ofed
- support
- torque
- web-server

The following rolls must be added manually:

- Red Hat OS
- Intel Driver

14) Back into the 'Welcome to Rocks' screen

Press on <CD/DVD-based Roll>, then insert the Intel® Driver roll media and press on <Continue>.
Select the Intel® Driver roll and press <Next>. Manually inserting the roll after completing the installation will not satisfy the requirements.

15) Back into the 'Welcome to Rocks' screen, repeat previous step with the Red Hat* Media.

Finally, the following list should be displayed on the roll window.

- base
- ganglia
- hpc
- intel-icr
- java
- kernel
- mlnx-ofed
- support
- torque
- web-server
- Red_Hat_Enterprise_Linux_5
- intel-driver

16) On the 'Welcome to Rocks' screen

Press <Next>.

17) On the 'Cluster Information' screen

Enter the fully qualified host name and complete the basic cluster information as required, press <Next>.

18) On the 'Ethernet Configuration for eth0' screen

Enter the network address and netmask for the frontend private network, then press <Next>.
For instance, the frontend private address may be 192.168.1.1 and its netmask 255.255.255.0.

19) On the 'Ethernet Configuration for eth1' screen

Enter the network address and netmask for the frontend public network, then press <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.

20) On the 'Miscellaneous Network Settings' screen

Enter the gateway and DNS server values if required, then press <Next>.

21) On the 'Root Password' screen

Enter and confirm the administrator user password, then press <Next>.

22) On the 'Time Configuration' screen

Enter proper Time Zone and NTP server, then press <Next>.

Disk Partitioning

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

23) On the 'Disk Partitioning' screen

Select 'Manual Partitioning' and then press <Next>.

Remove all the partitions (including an extended one if present) pressing <Delete> and define new partitions pressing <New> as shown below. After all partitions were defined, press <Next> to begin the partitioning.

- 100MB ext2 partition mounted at /boot, fixed size.
- 4096MB swap partition, fixed size.
- 51200MB 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 intel-driver roll and the RedHat* media when required.

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

4. Frontend Customization

Starting Customization

During the first boot, the frontend will compile and install the Mellanox* OFED packages. The expected completion time is about 15 minutes.

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) Make sure that the following files are located at /root directory

- Intel® Cluster Checker License (check the Intel® Cluster Ready website)
- Intel® Cluster Ready Reference Package (ZIP file)

Enabling Intel® Cluster Ready Software

28) Create a icr folder for scripts

```
mkdir -p /opt/intel/icr/sbin
```

29) Decompress the recipe package into the /opt/intel/icr/sbin directory and change to it.

```
unzip Intel_Cluster_Ready_Reference_Recipe_S5520UR-ICR1.1-
ROCKSPLUS5.1-RH5.3-C1-v1.2_20100127.zip -d /opt/intel/icr/sbin
cd /opt/intel/icr/sbin
```

30) The next step will verify the md5sum of the scripts tarball, then untar the Intel® Reference Recipe Script package. To achieve this execute

```
md5sum -c ICR1.1-ROCKSPLUS5.1-RH5.3-C1-v1.2.tar.gz.md5
tar -zxf ICR1.1-ROCKSPLUS5.1-RH5.3-C1-v1.2.tar.gz
```

Run the installation script

The installation script will generate the configuration file necessary for the recipe and will launch the configuration scripts. Also, the script will configure the IB network with the ip and netmask entered in the command line. If the ip and netmask are not specified the script will assign to ib0 the next subnet available from the provisioning network. The IP address 172.20.99.1 and the netmask 255.255.255.0 will be used for this configuration.

31) Execute the following command:

```
chmod +x ./rocksplus_icr_install.sh
./rocksplus_icr_install.sh -r ICR1.1-ROCKSPLUS5.1-RH5.3-C1-v1.2 -ip
172.20.99.1 -nm 255.255.255.0
```

32) Reload the environment to reflect the changes made by the installer

```
source ~/.bashrc
```

5. Compute Node Installation

Please refer to Clustercorp* Rocks+* documentation for detailed instructions on how to run 'insert-ethers' tool. Please refer to the board documentation on how to boot into PXE mode.

Compute Node Provisioning

33) Populate the rack number for the compute nodes hostname. The number is the one in "##" compute-##-XX

```
rackNumber="<popule with the Rack number>"
```

34) Start the provisioning service on the frontend. A status screen will appear to show detected compute nodes.

```
insert-ethers --appliance compute --cabinet=$rackNumber
```

Boot the compute nodes with a 1 minute between each one to avoid network booting conflicts. Ensure that all compute-nodes BIOS are configured to boot over the network as the first boot option. On the frontend wait until an '*' symbol appears next to each node entry and then press 'F8'.

Note 1:
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. If an '*' symbol did not appear after rebooting the node you should press 'F9', type the following commands and reboot the failing node:

```
rocks remove host <compute_node-name>
insert-ethers --update
insert-ethers --appliance compute --cabinet=$rackNumber --
rank=<compute_node-rank>
```

Note 2:
An errant network device (e.g. local IP switch broadcasting a DHCP request) can be detected and registered as a compute node during this installation phase. This must be corrected to replace the associated compute node name for a real compute node registration. The following steps will replace the registered compute node:

```
insert-ethers --replace=<compute_node-name>
```

35) Check that the compute nodes were installed. Type the command below. A list of nodes will be displayed. If all nodes were installed successfully, an "os" label must appear next to the node name. Wait until all the nodes are installed, the expected total installation time is about 30 minutes.

```
rocks list host pxeboot
```

36) Check that the compute nodes have booted successfully. Use the 'tentakel' command to verify that all the nodes are available; expect each node hostname as the output. Please wait until all the nodes are available.

```
tentakel hostname
```

Note:
If a compute node reports 'down' execute the following steps to reinstall the failing node and after that manually reboot the failing node. If this does not resolve the problem refer to Note 1 on step 34.

```
rocks set host pxeboot <compute_node-name> action=install
```

37) Log in as 'icr' and press just <enter> to answer the questions about SSH configuration. This will enable password-less logins across the cluster.

```
su – icr –c "exit"
```

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

```
rocks sync users
```

39) Reload the 'igb' module to use the new configuration.

```
tentakel 'modprobe –r –f igb ; modprobe igb'
modprobe –r –f igb ; modprobe igb
```

Warning: if not properly executed this command could leave your cluster without network connectivity and a manual restart should be done.

6. Verify the Cluster

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

40) Setting the global Recipe environments variables

```
source $CONF_BOARD
```

Run Intel® Cluster Checker tool automatically with icr_passthrough script

41) The icr_passthrough script executes Intel® Cluster Checker with the deployment parameter.

```
chmod a+x $PACKAGES_FOLDER/icr_passthrough.sh
$PACKAGES_FOLDER/icr_passthrough.sh
```

NOTE: You must execute this script at least one to complete Intel® Cluster Ready installation. After that, you may manually execute Intel® Cluster Checker in deployment mode, please refer to the Appendix Section for more instructions.

Output Analysis

42) Verify that Intel® Cluster Checker reports a successful cluster build. Look at the final line in the output, either on screen or in the associated output file. The last line should read "Check has succeeded."

File_tree exceptions

Some files may fail and it is accepted if:
▪ The file differs due to the use of prelink (or similar utility) by the Linux* distribution. The checksum of the original, unmodified file must be identical on all nodes.
▪ The file contains inline version control system information. The file must be identical on all nodes other than the inline version information.
▪ The file contains node-specific identification or configuration data. The file must be identical on all nodes other than the node-specific data

The following may fail during file_tree module test:

/opt/mlnx-ofed/src/OFED-1.4-mlnx8/RPMS/redhat-release-5Server-5.3.0.3/x86_64/*
/usr/java/jdk1.6.0_07/register*
/usr/java/jdk1.6.0_07/jre/lib/servicetag/registration*
/opt/torque/mom_logs/*
/opt/torque/lib64/xpbs/tclIndex
/opt/torque/lib64/xpbsmon/tclIndex
/opt/rocks/lib/graphviz/config

7. Appendix

Enable SOL

To enable SOL on your compute nodes you should execute the following command:

```
rocks set host bootflags compute flags="console=ttyS0,115200 text
noipv6 kssendmac"
```

Run Intel® Cluster Checker tool by manual execution

43) Switch to the icr account.

```
su – icr
```

44) Run Intel® Cluster Checker in deployment mode

```
cluster-check --deployment
```

Note: It is expected that some files fail the file_tree module. These files difference won't affect the cluster functionality. A list of these files is provided at 'Verify the cluster' section.

45) Run Intel® Cluster Checker in deployment mode as root.

```
exit
cluster-check --deployment
```