Supermicro ICR Recipe
For 1U Twin ™ 'Department Cluster' with Clustercorp Rocks+ 5.1

Version 1.3
6/25/2009

Table of Contents

1. System Configuration

Bill Of Materials (Hardware) The primary bill of materials, orderable from Supermicro's distributors, is shown in Table 1.   The BOM in Table 1 corresponds to that used for the cluster certification.

Quantity	Item	Manufacturer	Model
16	SuperServer 1U Twin™ with QDR InfiniBand	Supermicro	 SS6016TT-IBQF
2	24-port Layer 3 1/10-Gigabit Ethernet Switch	Supermicro	 SSE-G24_TG4
1	Gigabit Ethernet Switch Stacking Cable	Supermicro	MCC4L30-001
2	Gigabit Ethernet CX4 Stacking Module	Supermicro	AOM-SSE-X2C
1	36-port InfiniBand Switch	Mellanox	MTS3600
32	InfiniBand Cable, 1m, QSFP	Mellanox	MCC4Q30C-001
64	Intel® Xeon® Processor	Intel Corp.	Intel® Xeon® Processor X5550 (Nehalem)
192	2GB DDR-3-1066 ECC Registered Memory	Hynix	HMT125R7AFP4C-G7TB
32	3.5" SATA 250GB Hard Disk Drive	Seagate	Barracuda ES.2 ST3250310NS

Table 1: 1U Twin™ Cluster- Bill of Hardware Materials

This recipe provides a template that resellers and end-users can customize for their application specific needs. In addition to the hardware items in Table 1, the following are either required or useful during the installation process: USB keyboard and mouse, USB DVD-ROM drive, USB 4 port hub, two Ethernet cables and a laptop or desktop computer capable of running a web browser. Microsoft Internet Explorer 7 (7.0.6000.16681) was used while developing this recipe. Please refer to the SS6016TT-IBQF User Manual for instructions on assembling the unit. Although the BOM specifies the 1UTwin™ system, the 2U Twin²™ uses the same server motherboard, BIOS and Firmware and thus qualifies as 'materially identical' under the rules of the ICR program. Therefore, both 1U Twin™ and 2U Twin²™ based clusters using the X8DTT-IBQF and X8DTT-IBXF motherboards are ICR certified with Clustercorp Rocks+ 5.1.

Bill Of Materials (Software) The software bill of materials outlined in Table 2 below consists of two DVDs. The first is the Clustercorp Rocks+ 5.1 Jumbo DVD. In order to build and deploy ICR certified clusters with Rocks+, a software and support license must be purchased from Clustercorp. This can be obtained in multiple ways as described at: http://www.clustercorp.com/hardware/supermicro.html . During the certification process, an LG External Super-Multi DVD Rewriter USB DVD drive (model GSA-E60N) was used. The second DVD was a RedHat Enterprise Linux 5.3 distro DVD.

Distributed By	Description	File Name or Location
Cluster Corp	Rocks5.1 + Jumbo Distribution	Jumbo-5.1
Red Hat	RedHatEL 5.3	5.3

Table 2: 1U Twin™ Cluster- Bill of Software Materials

Bill Of Materials (Intel Cluster Ready License) The system vendor is required to run the Intel Cluster Checker tool both before it leaves the factory and after installation at the end user site. The Cluster Checker tool requires a license file to be installed on the cluster. The license file can be obtained free of charge directly from Intel at http://www.intel.com/go/cluster (registration required) if desired. Alternatively, the reseller can use Supermicro's pass through license included in the download bundle from the Supermicro web site along with this recipe.

Bill Of Materials (Download Bundle) The system vendor is required to run the Intel Cluster Checker tool both before it leaves the factory and after installation at the end user site (see Section 4 below). The Cluster Checker tool requires several files which are bundled together as a single download from the Supermicro web site along with this recipe. The bundle includes: fingerprint files, XML Config & Output files, Supermicro ICR license file (COM_*), the cluster certification certificate, a copy of this recipe document and a README file.

2. Firmware and BIOS Settings

Once the hardware and software system components have been obtained and the servers have been assembled and racked, you can begin the system configuration. Refer to the ® 1U Twin™ (SS-6016TT) System User Manual for details on the server assembly procedure.

Connect a keyboard, mouse and monitor to each server in turn and configure the BIOS as follows. Enter the BIOS setup as described in the system User Manual. First ensure that the BIOS rev is 1.1.b01, or dated 623/09 or later. If not, download the latest BIOS from the Supermicro web site and update the BIOS as described on the BIOS download web page. Next go to the far right tab on the main menu and enter 'Load optimized defaults'. Then set each BIOS parameter as follows, where '/' indicates a submenu in the BIOS setup.

- Advanced / Processor&ClockOptions / SimultaneousMultithreading = Disabled
- Advanced / Processor&ClockOptions / Intel EIST Technology = Disabled
- Advanced / IDE-SATA-FloppyOptions / ConfigureSATA = AHCI
- Advanced / PCI-PnPConfiguration / Load Onboard LAN 1 Option ROM = Enabled
- Advanced / ACPIConfiguration / ACPIVersionFeatures = ACPI2.0
- Advanced / IPMIConfiguration / LAN Configuration / IPAddress / IPAddrSource = Static
- Boot / BootDevicePriority / 1stBootDevice = Network: IBA GE Slot

On the head node only:

- Advanced / PCI-PnPConfiguration / Load Onboard LAN 1 Option ROM = Disabled

Power off each server after the BIOS update has completed. The next step is to install RHEL 5.3 and Rocks+ 5.1 on the cluster head node.

3. Software Installation

Cluster Head Node Installation

Connect the keyboard, monitor and mouse to the cluster head node along with a USB DVD drive. Insert the Rocks+ 5.1 Jumbo DVD into the drive and power on the head node. Refer to the Rocks Installation Guide and User Guide on the DVD for details on the procedures outlined below. It may be useful to have a 2nd copy of the DVD available in the laptop or desktop mentioned above so these guides are readily available.

- When the Rocks splash screen comes up type 'frontend' and enter.
- If you see 'CDROM not found' select OK and hit enter
- Enable IPv4 / Manual and disable IPv6. Use the following: 192.168.10.5 / 255.255.255.0, 192.168.10.1, 192.168.10.5
- At CDROM not found, cdrom:/ks.cfg select OK and enter
- Tab to CD/DVD based roll & hit enter
    - Select all rolls using either the mouse or keyboard tab & space
- Tab to next and hit enter
- Tab to CD/DVD based roll & hit enter
    - Insert the RHEL5.3 DVD in the drive & hit continue
    - Select the RHEL5.3 roll & hit enter
- Set the FQDN to x8twin.supermicro.com or your preferred FQDN
- Set the cluster name to x8twin or consistent with the FQDN
- Config eth0 as: 10.1.1.1 / 255.0.0.0, leave eth1 as already configured above
- Select manual disk partition, delete any partitions on the disk (/dev/sda)
    - Tab to new, hit enter and use the following: mount point is /, file system is ext3, size is 50000; tab to ok and hit enter (the recommended minimum size is 15000)

  - Tab to new, hit enter and use the following: tab to file system type, down arrow to swap and enter, size is 8000, tab to ok and hit enter (the recommended minimum size is 4000)
  - Tab to new, hit enter and use the following: mount point is /export, file system is ext3, size is 'fill to maximum allowable', tab to ok and hit enter
- Insert DVD media as prompted, set Linux config parameters when prompted. The system will install and boot to Linux
- Log in as root, open a terminal window, hit return three times to generate ssh keys
- Enter the command 'rocks sync users'
- Start the web browser, go to Roll Docs / intel-icr / 5.0, open section 3.1 and follow the directions to install the icr license obtained above in the Bill of Materials section.
- Open section 5.1 and follow the directions to set up Infiniband with IPoIB (required). Enter the commands in section 5.1 of the Roll Docs at a command prompt as root. The steps are given approximately as:
  - Enter the command 'rocks add host interface `hostname –s` ip=172.30.0.0 subnet=ipoib name=`hostname –s` -ib'
  - Enter the command 'ifup ib0' to ensure that the ib interface is up
  - Enter the command 'ifconfig ib0' and check that the ib interface is active
  - Start the ib subnet manager
    - Enter the command 'chkconfig opensmd on'
    - Enter the command 'service opensmd start'
    - Enter the command 'ibstat' and verify status is active
  - Enter the command 'rocks list network' and verify
    - Private 10.0.0.0
    - Public 192.168.10.0
    - IPoIB 172.30.0.0
  - Enter the command 'insert-ethers', select compute and hit ok
- You are now ready to begin installing the compute nodes

Cluster Compute Node Installation

Attach the monitor and keyboard to the first compute node and then power the node up. Verify that it begins to PXE boot and install Rocks. Move the monitor and keyboard back to the head node and verify that the 'insert-ethers' utility GUI shows 'found appliance compute-0-0'. Wait 10 seconds and power up the second compute node. Continue as above until all the compute nodes have started their install. The install will take about 30 minutes. After all the compute nodes have finished installing enter F8 at the insert-ethers GUI to exit and save cluster state.

Fix /etc/resolv.conf on the compute nodes as follows:

- On the head node, open /etc/resolv.conf in an editor and remove 'supermicro.com', save and exit
- Enter the command 'scp –p /etc/resolv.conf compute-0-0:/etc/resolv.conf'
- Copy /resolv.conf to each compute node in turn

The copy commands may take awhile to execute. Please be patient as the above modification to resolv.conf will fix that. For large clusters the scp command is easily scripted.

Reboot the compute nodes and verify minimal functionality as follows.

- Enter the command ‘tentakel shutdown –r now’ from a root terminal window on the head node.
- After the compute nodes have rebooted enter the command ‘tentakel uptime’. This should finish within a second or two. If it does not, verify the resolv.conf on each compute node by entering ‘tentakel cat /etc/resolv.conf’.

4. Verify a Correct Cluster Build

Responsibilities of Reseller

Under the terms of the pass through certification clause of the ICR program agreement (between Supermicro and Intel), the reseller shall run the cluster checker tool against the fingerprint files & XML configuration file provided by Supermicro along with this recipe. The cluster must pass the tests listed below before it leaves the reseller’s factory. In addition, the reseller must make provisions to re-run the tests once the cluster is installed at the end user site. This helps ensure that the system is functional after shipping (loose cables for example). The cluster checker tests are not burdensome. They could typically be incorporated into the reseller’s system burn in procedure for example.

Cluster Validation with ICR

You are now ready to validate extended cluster functionality with the ICR Cluster Checker. Refer to the ICR documentation and Cluster Checker manual for details on the procedures outlined below. The ICR documents can be found at http://www.intel.com/go/cluster. Begin the cluster validation process as follows:

- Log in to the head node as root, open a terminal window, and hit return three times to generate SSH keys if so prompted.
- Enter the command ‘su – icr’ and cd ~, which will take you to the preinstalled ICR account & home directory. All the required environment variables should now be set
- Open a web browser and access ‘Roll Docs’. The home page should default to the Rocks+ documentation already installed on the head node. If it does not come up, select ‘localhost’ as the URL which will bring up the Clustercorp Rocks+ 5.1 documentation. Go to the Roll Docs submenu from there.
- In the terminal window copy the Download Bundle from the Bill of Materials above in Section 1 to the icr account home directory. The Download Bundle should be extracted from the WinZip archive before copying to the head node.
- Copy the fingerprint files to /tmp. These files are named compute-0-0*.list and x8twin*.list.
- Next make a subdirectory /temp in the icr account home directory. Move all of the *.out and *.xml files to that subdirectory. These are saved here for later reference.
- Next copy /temp/icr_config_ib.xml to the icr account home directory.
- You are now ready to verify that the cluster configuration is compliant, functional and performant. The detailed procedure can be found in the Roll Docs, Section 6.1, Running the Cluster Checker. There are however some

important differences. There is no need to generate the fingerprint files. The fingerprint files from the download bundled copied to /tmp will be used instead. This will ensure that the installed packages and rolls match those certified by Supermicro. For the convenience of the reseller, the key commands are given below.

- In the icr account window, run the command 'cluster-check --certification=1.1 icr_config_ib.xml'. This will verify that the cluster installation is compliant with the ICR specification. Note that the dashes in the command are double dashes. This check will take quite a bit of time (30 minutes – 1 hour) to complete as it checks each compute node's CPU, memory, network and disk drive subsystems for correctness and performance. It also runs the HPCC benchmark as an application level test.
- Next open another terminal as root on the head node. In the root account, cd to the icr account home directory. Run the following commands: 'cluster-check –compliance=1.1 icr_config_ib.xml'. This test should complete within a few minutes.
- Either of the above procedures may fail. Screen output from the cluster-check tool is saved into a file *.out. Detailed output from the cluster-check run is saved in a corresponding *.xml file.

If the cluster-check tool flags an error for any of the above commands, the output must be examined and fixed in order to certify the cluster as compliant. Some failures are expected & normal. They do not imply that the cluster has failed certification. Check the corresponding *.out file in the /temp directory for similar errors. If a matching type of error is found in the /temp *.out file, it may be safely ignored. If a matching type of error is not found in the /temp *.out file, it must be fixed to certify the cluster as compliant.

If the error occurs during the –compliance checks, it is often easiest to start by re-installing the compute nodes. This is easily done under Rocks+ as root : 'rocks set host pxeboot compute-0-0 action=install' followed by 'ssh compute-0-0 "shutdown –r now" '. Repeat for each compute node in turn, compute-0-1, compute-0-2 etc. If the errors persist after re-install, contact Supermicro support.

If the cluster installation is flagged as functionally incorrect or failing subsystem performance, then examine the cluster checker output to determine which subsystem(s) is failing. Some typical failures might include:

- Node Failures
  - Stream test fails – reseat DIMMs, ensure all channels are populated.
  - DGEMM test fails – check scaled performance vs that of the certified system (2.66 GHz Xeon 5550) to determine if this is an actual failure.
  - HDParm test fails – swap out the failing hard drives and rerun the test
- Network Failures
  - IB – A typical failure is that one or more of the Infiniband links is non-functional. Determine which nodes have failing links by examining the 'all to all' output. Reseat the Infiniband cables on those nodes. The Rocks+ 5.1 installation procedure automatically checks the IB firmware and updates it if needed. Note however that the Mellanox 2.6.0 rev

firmware or later is required. If they continue to fail reset the IB switch module. If they continue to fail swap out the board.

Note that performance failures may be reasonable and expected. A different processor may not pass the DGEMM test, so scaling the result (in the *.xml file) to the actual processor frequency is needed. Similarly the Stream test can fail if for example the user specified a cluster where not all memory channels are populated. HDParm failures are very common. Different drive model or RAID configurations will give different performance. This is expected. The reseller must examine the *.xml out and determine if the measured performance is reasonable for the given configuration.

Power and Cooling Requirements The cluster configuration described in the BOM was measured as consuming 11 kW at 208V AC running Linpack on each node. This is expected to be highly configuration dependant, for example depending on the CPU power rating, number of DIMMs installed and so on. Therefore this measurement should be used only as a reference point.

Permissible Configuration Variants The pass through cluster certification is valid for certain variations on the configuration detailed in this recipe. Different processor types, memory manufacturer, density and number of DIMMs, disk drive manufacturer, capacity and count are permitted. Different models of server motherboard and Rocks software stack are not permitted. In that case the reseller may complete the full certification procedure themselves and apply to the ICR program web site for certification certificate. Important minimum configurations limits include

- Memory – 1 GB per processor core
- Disk – the head node must have 65GB of available storage. Available storage is formatted capacity minus swap.

5. Contacts

The reseller is responsible for first level product support with the end user. A comprehensive 2nd / 3rd level support package (for the reseller) is available as follows:

- Initial support requests can be made through the Supermicro support center via phone, email or web. It is essential to have the unit serial number on hand in order to process the request. The Supermicro support center can be found at http://www.supermicro.com/support/. Resellers may use the SuperServer support hotline.
- The Supermicro support center will make a determination as to whether the problem is hardware or software related. Hardware related issues will be handled by the Supermicro support center in the usual fashion.
- Software related issues will be redirected to Clustercorp. It is essential to have purchased Rocks+ from Clustercorp to be licensed to use the software and receive support.

6. Release Notes

- When creating user accounts, in order for the auto-mounter to work correctly, the home directory must be specified as /export/home/*. After the user account is created on the head node, issue the command 'rocks sync users' to replicate across the compute nodes.