Introduction

The Intel Xeon Processor E5-2600 Family Cluster Solutions Guide is an overview of supported components for the next generation of Intel Xeon® processors. The goal of this document is to give integrators and others who are enabling cluster solutions the most current information available.

By using the research and testing that has already been performed by Intel and the Intel Cluster Solutions Engineering team, you can save both time and resources. This guide does not yet provide a detailed installation recipe for Intel Xeon processor E5 solutions, but will assist you in creating new recipes.

This documentation provides relevant information concerning components and integration issues for clusters using the Intel Xeon processor E5. The guide also provides additional documentation related to Intel® Cluster Ready compliance and updates for these solutions.

How to Use this Guide

This guide is divided into two sections. The first “Support” section lists a wide variety of solution components from the HPC cluster ecosystem. Its compatibility for each of the target Linux distributions is listed, and additional notes are provided.

The second “Install Guide” section is practical instruction for implementing new solutions. It is intended to be used a starting point for creating new recipes. It will also give you additional requirements, caveats, and recommendations when migrating existing recipes.
Content Authors

Peter Hartman – Intel processors, server products, and server management
Christopher Heller – Cluster provisioning software and operating systems
Andreas Hiristius - Communication fabrics; content editor
Werner Krotz-Vogel – Developer tools
Patrick Ryan –Schedulers, resource management software, and memory configuration
Jeremy Siadal – Cluster storage; front matter
Brock Taylor – Intel Cluster Ready architecture; engineering manager

INFORMATION IN THIS DOCUMENT IS PROVIDED IN CONNECTION WITH INTEL PRODUCTS. NO LICENSE, EXPRESS OR IMPLIED, BY
ESTOPPEL OR OTHERWISE, TO ANY INTELLECTUAL PROPERTY RIGHTS IS GRANTED BY THIS DOCUMENT. EXCEPT AS PROVIDED IN
INTEL’S TERMS AND CONDITIONS OF SALE FOR SUCH PRODUCTS, INTEL ASSUMES NO LIABILITY WHATSOEVER AND INTEL DISCLAIMS
ANY EXPRESS OR IMPLIED WARRANTY, RELATING TO SALE AND/OR USE OF INTEL PRODUCTS INCLUDING LIABILITY OR WARRANTIES
RELATING TO FITNESS FOR A PARTICULAR PURPOSE, MERCHANTABILITY, OR INFRINGEMENT OF ANY PATENT, COPYRIGHT OR OTHER
INTELLECTUAL PROPERTY RIGHT.

UNLESS OTHERWISE AGREED IN WRITING BY INTEL, THE INTEL PRODUCTS ARE NOT DESIGNED NOR INTENDED FOR ANY APPLICATION IN
WHICH THE FAILURE OF THE INTEL PRODUCT COULD CREATE A SITUATION WHERE PERSONAL INJURY OR DEATH MAY OCCUR.

Intel may make changes to specifications and product descriptions at any time, without notice. Designers must not rely on the absence or characteristics
of any features or instructions marked "reserved" or "undefined." Intel reserves these for future definition and shall have no responsibility whatsoever for
conflicts or incompatibilities arising from future changes to them. The information here is subject to change without notice. Do not finalize a design with this
information.

The products described in this document may contain design defects or errors known as errata which may cause the product to deviate from published
specifications. Current characterized errata are available on request.

Contact your local Intel sales office or your distributor to obtain the latest specifications and before placing your product order.

Copies of documents which have an order number and are referenced in this document, or other Intel literature, may be obtained by calling
1-800-548-4725, or go to: http://www.intel.com/design/literature.htm

This document contains information on products in the design phase of development. All products, computer systems, dates and figures specified are
preliminary based on current expectations, and are subject to change without notice.

Intel product plans in this document do not constitute Intel plan of record product roadmaps. Please contact your Intel representative to obtain Intel’s
current plan of record product roadmaps.

Performance tests and ratings are measured using specific computer systems and/or components and reflect the approximate performance of Intel
products as measured by those tests. Any difference in system hardware or software design or configuration may affect actual performance.

Requires a system with a 64-bit enabled processor, chipset, BIOS and software. Performance will vary depending on the specific hardware and software
you use. Consult your PC manufacturer for more information. For more information, visit http://www.intel.com/info/em64t

Requires a system with Intel® Turbo Boost Technology. Intel Turbo Boost Technology and Intel Turbo Boost Technology 2.0 are only available on select
Intel® processors. Consult your PC manufacturer. Performance varies depending on hardware, software, and system configuration. For more information,
visit http://www.intel.com/go/turbo

Intel® AES-NI requires a computer system with an AES-NI enabled processor, as well as non-Intel software to execute the instructions in the correct
sequence. AES-NI is available on select Intel® processors. For availability, consult your reseller or system manufacturer. For more information, see Intel®
Advanced Encryption Standard Instructions (AES-NI).

Hyperthreading is available on select Intel® Core™ processors. Requires an Intel® HT Technology-enabled system. Consult your PC manufacturer.
Performance will vary depending on the specific hardware and software used. For more information including details on which processors support HT
Technology, visit http://www.intel.com/info/hyperthreading

See the Processor Spec Finder at http://ark.intel.com or contact your Intel representative for more information.

Intel the Intel logo, and Xeon are trademarks or registered trademarks of Intel Corporation or its subsidiaries in the United States and other countries.

*Other names and brands may be claimed as the property of others.

Copyright © 2012 Intel Corporation
**Support**

Three Linux* distributions were tracked for detailed support of Intel Xeon processor E5 targets at silicon launch. This report assumed that these three distributions represent potentially available options for commercially-supported distributions in the early Q1 timeframe, subject to changes in roadmaps and adoption.

There may be other Linux distributions may install and run correctly on Intel Xeon processors E5. Verification of other Linux distributions is not performed by this team, unless otherwise noted.

The tables included in this section represent a support matrix, where expected support level for each component is cross-referenced to each distribution. All data here is based on publicly known support data as of February 1st, 2012.

The support matrix divides the support into 4 categories.

**Green**

Fully supported. Component will be validated on the OS and formally supported by the vendor(s) at Intel Xeon processor E5 launch.

**Yellow**

Less than full support. Component receives partial or best-effort support for that OS. This also includes components that are known to work, but do not receive formal support; for example, support for a component on RHEL 6.2 that is only fully supported on Scientific Linux 6.2.

**Orange**

Delayed support. Component will not be fully supported at Intel Xeon processor E5 launch, but will be fully supported within 90 days after launch. For delays more than 90 days, the component is listed as partial or no support, as appropriate.

**Red**

No support. Component is unsupported and has known issues with the OS, may not work with the OS, or no information is available.

---

**Note**

All of the content and dates in this document are subject to change without notice. This team strives to keep the information here as current and accurate as possible. However, this is no guarantee that the information is correct or includes the latest announcements. Many of the products here may be in development or in pre-production testing, so specifics are likely to change.

*Other names and brands may be claimed as the property of others.*
Support Matrix
The support matrix provides a list of components and Linux distribution support on February 1st, 2012.

<table>
<thead>
<tr>
<th>Component</th>
<th>RHEL 5.7</th>
<th>RHEL 6.2</th>
<th>SLES 11 SP1</th>
<th>Notes</th>
</tr>
</thead>
<tbody>
<tr>
<td>EPSD Servers</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>S2600GZ</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>S2600WP</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>S2400LP</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>S2400BB</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>S2600JF</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>HW Components</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Video Driver</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>(Matrox* 200e)</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Solid State Disks</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>RAID (SATA 0/1/5/10, SAS 0/1/10)</td>
<td></td>
<td></td>
<td></td>
<td>Full SATA support starting RHEL 5.5/SLES 11 SP1, SAS support starting RHEL 6.2/SLES 11 SP2</td>
</tr>
<tr>
<td>Intel Network Drivers</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>ixgbe</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>igb</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>e1000e</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>e1000</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Intel Network Adapters</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>NE020</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>X520</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>X540</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>82580</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>I350</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Fabric Software</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>OFED</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Mellanox OFED</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Qlogic OFED</td>
<td></td>
<td></td>
<td></td>
<td>Support for Intel TrueScale InfiniBand</td>
</tr>
<tr>
<td>Cluster Storage</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Lustre</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>pNFS</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>PVFS2</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Notes:</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

*Other names and brands may be claimed as the property of others.
<table>
<thead>
<tr>
<th>Component</th>
<th>RHEL 5.7</th>
<th>RHEL 6.2</th>
<th>SLES 11 SP1</th>
<th>Notes</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Intel Developer Tools</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>C++ Composer XE</td>
<td></td>
<td></td>
<td></td>
<td>Intel* Cluster Studio XE 2012</td>
</tr>
<tr>
<td>Fortran Composer XE</td>
<td></td>
<td></td>
<td></td>
<td>Intel* Cluster Studio XE 2012</td>
</tr>
<tr>
<td>Debugger</td>
<td></td>
<td></td>
<td></td>
<td>Intel* Cluster Studio XE 2012</td>
</tr>
<tr>
<td>Cilk Plus</td>
<td></td>
<td></td>
<td></td>
<td>Intel* Cluster Studio XE 2012</td>
</tr>
<tr>
<td>MPI Library</td>
<td></td>
<td></td>
<td></td>
<td>Intel* Cluster Studio XE 2012</td>
</tr>
<tr>
<td>Math Kernel Library</td>
<td></td>
<td></td>
<td></td>
<td>Intel* Cluster Studio XE 2012</td>
</tr>
<tr>
<td>Integrated Performance Primitives</td>
<td></td>
<td></td>
<td></td>
<td>Intel* Cluster Studio XE 2012</td>
</tr>
<tr>
<td>Threading Building Blocks</td>
<td></td>
<td></td>
<td></td>
<td>Intel* Cluster Studio XE 2012</td>
</tr>
<tr>
<td>Trace Analyzer and Collector</td>
<td></td>
<td></td>
<td></td>
<td>Intel* Cluster Studio XE 2012</td>
</tr>
<tr>
<td>VTune Amplifier XE</td>
<td></td>
<td></td>
<td></td>
<td>Intel* Cluster Studio XE 2012</td>
</tr>
<tr>
<td>Inspector XE</td>
<td></td>
<td></td>
<td></td>
<td>Intel* Cluster Studio XE 2012</td>
</tr>
<tr>
<td>Parallel Advisor</td>
<td>U</td>
<td>U</td>
<td>U</td>
<td>Windows only – Linux later</td>
</tr>
<tr>
<td><strong>Cluster Managers</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Bright Cluster Manager</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Platform HPC</td>
<td></td>
<td></td>
<td>D</td>
<td>SLES 11 SP1 support added March 2012</td>
</tr>
<tr>
<td><strong>Applications</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Altair RADIOSS</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Ansys ANSYS</td>
<td>D</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Ansys Fluent</td>
<td>D</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>ESI PAM-Crash</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Flow Science Flow3D</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Landmark Nexus</td>
<td>D</td>
<td>U</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Landmark ProMAX</td>
<td>D</td>
<td>U</td>
<td></td>
<td></td>
</tr>
<tr>
<td>LSTC LS-Dyna</td>
<td>L</td>
<td>L</td>
<td>U</td>
<td>CentOS support only</td>
</tr>
<tr>
<td>Metacomput CFD++</td>
<td>L</td>
<td>L</td>
<td></td>
<td>RHEL clone support</td>
</tr>
<tr>
<td>MSC MD-Nastran</td>
<td>L</td>
<td>U</td>
<td>U</td>
<td></td>
</tr>
<tr>
<td>Simulia ABAQUS</td>
<td>D</td>
<td>D</td>
<td></td>
<td>Expected 90 days after OS release</td>
</tr>
<tr>
<td><strong>Batch/Schedulers</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Grid Engine</td>
<td></td>
<td></td>
<td>L</td>
<td>Univa GE or Open Grid Scheduler</td>
</tr>
<tr>
<td>PBS Professional</td>
<td></td>
<td></td>
<td></td>
<td>Version 8 is said to support all</td>
</tr>
<tr>
<td>Platform LSF</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>SLURM</td>
<td></td>
<td></td>
<td>L</td>
<td>V2.5.X or V3.0.X if NUMA support is needed</td>
</tr>
<tr>
<td>TORQUE</td>
<td></td>
<td></td>
<td>L</td>
<td></td>
</tr>
<tr>
<td><strong>Cluster Monitoring</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Ganglia</td>
<td>L</td>
<td>L</td>
<td>L</td>
<td></td>
</tr>
<tr>
<td>Nagios</td>
<td>L</td>
<td>L</td>
<td>L</td>
<td></td>
</tr>
</tbody>
</table>

*Other names and brands may be claimed as the property of others.*
Intel Xeon Processor E5 family

Models

Intel Xeon processor E5-2600 is a performance processor using socket R (LGA2011). It is an ideal processor for HPC cluster applications, as the design provides 2 QPI links between processors, 4 memory channels per socket, and 3 DIMMs per channel.

New Technologies

This section defines a few key technologies for Intel Xeon processor E5.

- **AVX (Intel® Advanced Vector Extensions):** A new 256-bit instruction set extension to SSE and is designed for applications that are Floating Point (FP) intensive. Intel AVX improves performance due to wider vectors, new extensible syntax, and rich functionality. This results in better management of data and general purpose applications like image, audio/video processing, scientific simulations, financial analytics and 3D modeling and analysis. See [http://software.intel.com/en-us/avx](http://software.intel.com/en-us/avx) for more information.

- **TSC deadline timer:** offers system software a low overhead per-logical-thread deadline timer in TSC units.

Intel Server Boards

The following are common components among the listed Intel server boards based on Intel Xeon processor E5.

- **Network:** Powerville 82580 Ethernet controller, supported in RHEL5.4+, RHEL6.0+ and SLES11SP1+. Windows drivers can be found at [http://downloadfinder.intel.com](http://downloadfinder.intel.com).
- **Graphics:** Matrox 200e graphics controller. Supported on all Linux systems, built in on Windows.
- **Storage:** Intel C600 Series PCH, with RSTe SATA/SAS RAID support. Drivers supporting SATA are included in RHEL5.4+, RHEL6.0+ and SLES11SP1+; SAS support is available in RHEL6.2 and SLES11SP2.
- **Full-width boards support SAS. Half-width boards do not support SAS.**
- **RAID:** Intel® RSTe 0/1/10 minimum. RAID 5 possible.
- **Chipset:** Intel C600 Series PCH, Supported in RHEL5.4+, RHEL6.0+ and SLES11SP1+.
- **Expansion:** PCI Express Gen 3 (PCIe3) expansion slots for supported adapters. Other adapters will operate at PCIe Gen 2 speed.
- **Memory Support:** ECC DDR3 UDIMM, ECC DDR3 RDIMM, ECC DDR3 LRDIMM, 800MHz to 1600MHz

### Half-width Server Boards

<table>
<thead>
<tr>
<th>Board Name</th>
<th>Processors</th>
<th>DIMMs</th>
<th>Network</th>
<th>Expansion</th>
</tr>
</thead>
<tbody>
<tr>
<td>S2600JF</td>
<td>Dual socket -EP</td>
<td>8 DIMMs over 8 channels</td>
<td>Dual 1GbE</td>
<td>Support for I/O Module, 3 PCIe3x16</td>
</tr>
<tr>
<td>S2600WP</td>
<td>Dual socket -EP (up to 130W)</td>
<td>16 DIMMs over 8 channels</td>
<td>Dual 1GbE</td>
<td>Support for I/O Module 2 PCIe3x16</td>
</tr>
</tbody>
</table>

*Other names and brands may be claimed as the property of others.*
Full-width Server Boards

<table>
<thead>
<tr>
<th>Codename</th>
<th>Processors</th>
<th>DIMMs</th>
<th>Network</th>
<th>Expansion</th>
</tr>
</thead>
<tbody>
<tr>
<td>S2600GZ</td>
<td>Dual socket -EP</td>
<td>24 or 16 DIMMs (depending on version) over 8 channels</td>
<td>Quad 1GbE</td>
<td>Support for IO Module</td>
</tr>
<tr>
<td>(Max Memory board)</td>
<td></td>
<td></td>
<td></td>
<td>2xPCIe3 x8 on 1U</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>6xPCIe3x16 or 2x16 2x8 on 2U</td>
</tr>
<tr>
<td>S2600IP</td>
<td>Dual socket –EP (up to 130W)</td>
<td>16 DIMMs over 8 channels</td>
<td>Quad 1GbE, not available in 1U model</td>
<td>Support for IO Module or SAS ROC module</td>
</tr>
<tr>
<td>(Max IO board)</td>
<td></td>
<td></td>
<td></td>
<td>Up to 4 double-width PCIe3 x16 or 8 single-width PCIe3 slots</td>
</tr>
<tr>
<td>W2600CR</td>
<td>Dual socket –EP (up to 150W)</td>
<td>16 DIMMs over 8 channels</td>
<td>Dual 1GbE</td>
<td>Support for IO Module or SAS ROC module</td>
</tr>
<tr>
<td>(Workstation board)</td>
<td></td>
<td></td>
<td></td>
<td>Up to 4 double-width PCIe3 x16 or 8 single-width PCIe3 slots</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>Includes HD audio, USB 3.0, eSATA, IEEE 1394b</td>
</tr>
</tbody>
</table>

HPC Board Qualification

The Cluster Solutions and Technology group is testing and performing Intel Cluster Ready certification on S2600GZ and S2600JF servers for the HPC Cluster Market. Intel Enabled Solutions Acceleration Alliance (Intel ESAA) program also includes S2600WP for its large memory capability, and S2400LP for the low-power, high density market.

Intel Storage

RSTe3

Intel 6-series chipsets, such as the Intel C600 chipset, include support for Intel Rapid Storage Technology enterprise (RSTe). This provides integrated SATA and SAS RAID 0, 1, and 10 directly in the chipset. RAID 5 is supported with an upgraded ROM. Using RSTe mode with single drives is likely to provide better performance with SATA drives than AHCI mode.

Features such as boot support, alerting, RAID expansion, verification, and system management are standard. Drive roaming between RSTe3 and AHCI is also supported.

The driver is available in the Linux kernel source tree starting with v3.0-rc6. The required isci driver is available in RHEL 5.7 and 6.1 or later and SLES11SP2 or later. Drivers are available separately for some earlier versions, such as SLES 11SP1.

Validated driver support for 32-bit and 64-bit Windows is available for Windows Server 2008, including R2 and SP2, Windows Vista, and Windows 7.

Server Management

There are multiple options to control and monitor modern servers. The most common management features are provided by the Intelligent Platform Management Interface (IPMI). For example, servers with embedded IPMI can be turned on or off or reset as long as power is applied to the system and a network is connected and configured. Other management interfaces can extend and enhance IPMI for even greater control.

Intelligent Power Node Manager uses the capabilities of IPMI to minimize power consumption, especially in a limited power environment. Active Management Technology (AMT) can be used to provision a host with an OS, as well as monitor hardware. Data Center Manager (DCM) can connect to multiple AMT clients to monitor vast arrays of hardware and consolidate datacenter management tasks.

*Other names and brands may be claimed as the property of others.*
Intelligent Platform Management Interface (IPMI)

Servers with embedded IPMI can be monitored and controlled via in-band or out-of-band communication. Essentially, an IP address is assigned to the Baseboard Management Controller (BMC), either statically or by means of a DHCP server. With the IP address and the access credentials configured, the IT manager has access to the hardware sensors provided by IPMI and to all control features of IPMI. If the embedded IPMI device is enhanced by a Remote Management Module (RMM) console access can be routed to another computer via LAN.

The ipmitool Utility

The utility ipmitool provides the interface for the IPMI functionality within the embedded BMC chip on the server motherboard. The general in-band ipmitool syntax is:

```
ipmitool <ipmi segment> <command>
```

For example, to get the current general settings for the chassis, such as power status, the following command can be used:

```
ipmitool chassis status
```

To get the current temperature readings across all of the temperature sensors on the server, run the command:

```
ipmitool sdr type Temperature get
```

Out-Of-Band commands are run using the syntax:

```
ipmitool -I lan -H <ip address> -U <username> -f <password> <segment> <command>
```

Where <ip address> specifies the IP address of the server that is being queried and <username> and <password> specify the credentials required to access the BMC.

Commands can use either format. There are extensive commands available via this utility. Please see the man page for full coverage of all the available options.

Useful IPMI commands

- Power the server on: power on
- Power the server off: power off
- Cycle the power on the server: power cycle
- Turn on the identification light for a time: chassis identify <interval in seconds>
- Display the active sensors: sdr list all
- Show system event log: sel list
- Boot directly into bios on next boot: chassis bootparam set bootflag force_bios

*Other names and brands may be claimed as the property of others.*
Operating Systems

Our solution guide is currently tracking 5 different publicly available Linux* distributions—the three target distributions and two additional distributions—as well as the aforementioned kernel version/changes. This may change in future versions of this guide.

Red Hat Enterprise Linux (RHEL)

RHEL 5.7 and later support the required drivers and many new features, but does not support all of the capabilities of the Intel Xeon Processor E5 platform that RHEL 6.2 supports. RHEL5.8 is available since mid-February 2012. RHEL 5.8 includes better, but not complete, support for the platform.

RHEL 6.2 was released in December 2011. It includes the following updates over RHEL 6.1 that improve support for the Intel Xeon Processor E5 Platform:

- Updated ixgbe NIC driver
- PCIe Gen 3 support
- New CPU microcode updates
- Intel Xeon Processor E5-2600 PMU (Performance Monitoring Unit)
- Performance event support
- Oprofile support for Intel Xeon Processor E5-2600
- glibc memory optimized functions
- Improved AES-NI support
- Support for Predictive Failure Analysis,
- Support for the C600 Series Chipset SAS Controller
- Improved software RAID performance

RHEL Clones

There are two major RHEL clones available, CentOS and Scientific Linux.

After a period of uncertainty around the status and progress of CentOS, its releases now follow RHEL releases closely again. It is expected that CentOS will, from now on, continue in this fashion.

The latest versions of both clones are usually available two to four months after the release of the corresponding RHEL release.

Due to the fact that Scientific Linux is targeted primarily at the scientific community, particularly High Energy Physics, the official status can be ‘beta’ or ‘test’ until it has been certified by those user communities, even though it is production ready.


CentOS is available at http://www.centos.org.

SUSE

SUSE Linux Enterprise (SLES) 11 SP1 was the target distribution used for pre-launch testing. It’s newer updates support many of the requirements and technologies needed for the new processors. Additional drivers may be required, depending on the hardware used.

SLES 11 SP2 was released in mid-February and features an enhanced 3.0 kernel, providing better support for these key areas:

- Native support for the Intel Xeon processor family E5 and chipsets
- Support for transparent hugepages, improving performance for memory intensive workloads
- Better utilization of multi-core systems

*Other names and brands may be claimed as the property of others.
The 3.0 kernel also provides numerous other performance improvements over the 2.6.32 kernel. Many of these improvements are present in the block I/O handling area and file system support structure. SLES11 SP2 also adds Btrfs support. The Btrfs file system provides the ability to snapshot file system contents, simplify volume management and enable new features such as update roll-back.

**Note regarding RHEL 5.8 and SLES 11 SP2.**

The purpose of this guide is to help implement solutions time to market (TTM) with the launch of the Intel Xeon Processor E5 2600. Since these two Linux distributions were released only very shortly before the launch, this guide continues to use RHEL 5.7 and SLES11 SP1 as target distributions for the time. However, due to enhanced support of new Intel Xeon processor features in RHEL5.8 and SLES 11SP2, this guide recommends that the newer distributions be used when they are supported by solution hardware and software.

Ubuntu

Ubuntu 11.10 is based on the kernel.org 3.0 kernel. This makes Ubuntu a good choice for early hardware testing, as it uses current kernels with newer support capabilities. It is also a good hypervisor on which to test the other distributions as they move from alpha to beta to release.

**Operating System Components**

The following versions of components are provided in tracked distributions. Note that Red Hat pulls packages and updates selectively from upstream OFED sources and calls it "RDMA".

<table>
<thead>
<tr>
<th>Component Versions Included as Part of Linux Distribution</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Item</strong></td>
</tr>
<tr>
<td>kernel</td>
</tr>
<tr>
<td>libstdc++</td>
</tr>
<tr>
<td>gcc</td>
</tr>
<tr>
<td>glibc</td>
</tr>
<tr>
<td>xfree86</td>
</tr>
<tr>
<td>OFED</td>
</tr>
<tr>
<td>Java RTE</td>
</tr>
<tr>
<td><strong>Item</strong></td>
</tr>
<tr>
<td>kernel</td>
</tr>
<tr>
<td>libstdc++</td>
</tr>
<tr>
<td>gcc</td>
</tr>
<tr>
<td>glibc</td>
</tr>
<tr>
<td>xfree86</td>
</tr>
<tr>
<td>OFED</td>
</tr>
<tr>
<td>Java RTE</td>
</tr>
</tbody>
</table>

*Other names and brands may be claimed as the property of others.*
Kernel Support for New Features

When talking about Linux operating systems, one must first address the hardware features that are most relevant to HPC and then determine what kernel version supports those hardware features. The kernel.org 2.6.38 kernel will be sufficient for basic Intel Xeon processor E5 hardware support, although the kernel.org 2.6.39 has comprehensive support for features such as Hugepage error recovery and some of the advanced power features. The kernel.org 3.0 and later kernels support essentially all features of the Intel Xeon Processor E5 platform.

Older operating system versions may run normally on Intel Xeon processor E5; however they may be limited by support for new processor technologies, and may not provide driver support for the latest chipset on Intel Xeon Processor E5 platforms. Using older kernels or Linux distributions can lower performance and not support certain new features.

Linux distributions with commercial support such as Red Hat Enterprise Linux (RHEL) or SUSE Enterprise Linux Server (SLES) use customized, branched kernels with version numbers, at times, lower than the current or required Linux kernel versions. For example, RHEL 6.2 uses a 2.6.32 kernel. However, the maintainers add, or “back-port”, important features from newer kernel.org kernels into their kernels in the distribution. The most recent updates or service packs of major Linux distributions will have proper support for the new and advanced features of the Intel Xeon processor E5 platform.

New and advanced features and technologies that are now supported are: AVX, Xsaveopt, TSC Deadline Timer, Large VT-d pages, ACPI 4.0 Dynamic Core Allocation, SAS, DIMM error reporting for Intel Xeon processor E5 (kernel, mcelog), Energy Efficient Turbo, Package Thermal support, P-states, and Core C-state tuning.

The following tables list Intel Xeon processor E5 features and kernel/library support for Linux. Unless stated otherwise, all kernel versions listed here refer to kernel.org source.

<table>
<thead>
<tr>
<th>Feature</th>
<th>Support</th>
</tr>
</thead>
<tbody>
<tr>
<td>AVX in Kernel</td>
<td>In kernel 2.6.30.</td>
</tr>
<tr>
<td>AVX in Library (Theora)</td>
<td>WIP in upstream, target 1H 2011</td>
</tr>
<tr>
<td>AVX in glibc</td>
<td>glibc 2.11</td>
</tr>
<tr>
<td>AVX in gcc</td>
<td>Gcc 4.4.0</td>
</tr>
<tr>
<td>AVX in binutils</td>
<td>Binutils 2.19.51.0.1</td>
</tr>
<tr>
<td>256bit vectorizer in gcc</td>
<td>Gcc 4.6.0</td>
</tr>
<tr>
<td>AVX in gdb</td>
<td>Gdb 7.2; Backport to 7.1 available.</td>
</tr>
<tr>
<td>MCA Recovery (core)</td>
<td>In kernel 2.6.39</td>
</tr>
<tr>
<td>Hugepage error recovery</td>
<td>In kernel 2.6.39</td>
</tr>
</tbody>
</table>

*Other names and brands may be claimed as the property of others.*
<table>
<thead>
<tr>
<th>Feature</th>
<th>Support</th>
</tr>
</thead>
<tbody>
<tr>
<td>TSC Deadline Timer</td>
<td>WIP, target kernel 2.6.38</td>
</tr>
<tr>
<td>intel_idle driver</td>
<td>2.6.35</td>
</tr>
<tr>
<td>Core C-states tuning</td>
<td>Backports to 2.6.32/35/36 are available.</td>
</tr>
<tr>
<td>Package level thermal</td>
<td>2.6.36</td>
</tr>
<tr>
<td>RAPL (Running Average Power Limit) -</td>
<td>WIP in upstream, target kernel 2.6.38.</td>
</tr>
<tr>
<td>Socket Power meter &amp; DRAM power meter</td>
<td></td>
</tr>
<tr>
<td>Energy Efficient Turbo</td>
<td>Kernel: 2.6.35 / User space tool is available but not finalized.</td>
</tr>
<tr>
<td>Power Limit notification</td>
<td>Kernel: 2.6.35</td>
</tr>
<tr>
<td>PMU support (perf_events)</td>
<td>WIP in upstream, target kernel 2.6.39.</td>
</tr>
<tr>
<td>Install/Boot support for SAS Controller Unit</td>
<td>Available in RHEL 6.2.</td>
</tr>
<tr>
<td>(SCU)</td>
<td></td>
</tr>
<tr>
<td>SAS Controller Unit (SCU) Driver and libas</td>
<td>WIP: Plan: ~ww32/Kernel 2.6.40</td>
</tr>
<tr>
<td>SMBUS driver for Intel C600 Series</td>
<td>Done – Kernel 2.6.37</td>
</tr>
<tr>
<td>Intel C600 Series DID patches (SATA, HD Audio</td>
<td>In kernel 2.6.36: HDA and AHCI/IDE SATA;</td>
</tr>
<tr>
<td>SMBUS, TCO Watchdog)</td>
<td>2.6.37: TCO Watchdog</td>
</tr>
<tr>
<td>Report SATA3 Speed</td>
<td>2.6.38: additional LPC controller</td>
</tr>
<tr>
<td>TXT enabling (LT-SX)</td>
<td>In kernel 2.6.29</td>
</tr>
<tr>
<td>TXT for Xen: In Xen 3.4</td>
<td></td>
</tr>
<tr>
<td>TXT for KVM: In kernel (version To Be Filled)</td>
<td></td>
</tr>
<tr>
<td>Tboot: OSV need to use most recent version</td>
<td></td>
</tr>
<tr>
<td>Pause Loop Exiting</td>
<td>WIP</td>
</tr>
<tr>
<td>Large VT-d pages for native kernel</td>
<td>WIP: Plan: Kernel 2.6.38</td>
</tr>
<tr>
<td>Large VT-d pages for KVM</td>
<td>Pending. Need more info</td>
</tr>
<tr>
<td>Large VT-d pages (1GB/2MB) for Xen</td>
<td>Xen 4.1</td>
</tr>
<tr>
<td>AMT support: MEI driver, LMS, ACU</td>
<td>MEI driver – Target kernel 2.6.39.</td>
</tr>
</tbody>
</table>

*Other names and brands may be claimed as the property of others.*
Cluster Managers

The area of Cluster Managers includes software solutions that provide integrated installation and configuration of a cluster operating environment, including cluster node provisioning. They also provide cluster-wide command and control, and some solutions include additional features such as monitoring and job scheduling.

We are currently tracking two commercial provisioning vendors that have Intel Cluster Ready-integrated provisioning solutions. Please see the Operating system section above to determine which operating systems will be ready for Intel Xeon processor E5.

Platform* HPC

http://www.platform.com/cluster-computing/platform-hpc

Platform Computing, an IBM company, released their PCM 3.1 product which will add support for RHEL6.2.

Platform HPC is a comprehensive stack incorporating provisioning, workload management, MPI and an application portal. Platform Computing states that “Platform HPC is the industry’s easiest and most complete HPC management software. Its robust cluster and workload management capabilities are accessible using the latest design in web-based interfaces - making it powerful, yet easy to use.”

Platform Cluster Manager is a Linux cluster provisioning and management solution, used to provision, run, manage and monitor clusters. It is a component of Platform HPC.

All components of Platform HPC are validated on the supported operating systems listed at http://www.platform.com/cluster-computing/platform-hpc/system-requirements

Bright* Cluster Manager

http://www.brightcomputing.com/Bright-Cluster-Manager.php

Bright Computing’s BCM 5.2 supports RHEL6.2, Scientific Linux 6.2, SLES11SP2, and Centos 6.2. Enhancements to the new Bright Cluster Manager include:

• The addition of SLURM and LSF as integrated workload management options, extending the previous choices of PBS Professional, Torque/Moab, Torque/Maui, Grid Engine.
• Expansion of cluster health management framework with an extended range of user-selectable early-warning and intervention capabilities for any metric within the cluster.
• A customizable web-based, user portal separate from the admin GUI.
• The inclusion of several dozen customer-driven enhancements, including several suggested by TOP500 system administrators.

Other Provisioning Solutions

StackIQ (formerly Clustercorp) ROCKS+ 5.4 will support RHEL 5.7 and CentOS 5.7. Rocks+ 6.0.1 will support new distributions such as RHEL 6.2.C600 Series

Rocks (rocksclusters.org) 5.4.3 is the latest revision and is built against CentOS 5.6. This does not implement the necessary hardware support. A beta implementation supporting RHEL 6.2 is available now.

Additional tests using Warewulf 3.1 and RHEL 6.2 have been successful on the latest hardware.

*Other names and brands may be claimed as the property of others.
Cluster Monitoring

Nagios Core
http://www.nagios.org/projects/nagioscore
Previously known as just "Nagios", Nagios Core is an open source service monitoring tool that is free without a commercial support plan. Plug-ins are available for specific hardware vendors such as HP and Dell. System requirements are a Linux system with a C compiler and Apache installed.

Previously, Platform Cluster Manager installed Nagios by default. Future versions of PCM are expected to use a different cluster monitoring dashboard as part of the HPC suite. Nagios 2.12 is included version in PCM 2.1.

Latest version: 3.3.1 (Jul-25-11). Updates and bug fixes have been 1-6 months apart while minor version updates have been every year.

Ganglia
http://ganglia.sourceforge.net
Ganglia is community-driven and seems to have no commercial support or enhanced suite. Plug-ins, or modules, are available for hardware/software-specific monitoring. It is installed from source GNU autoconf and has no listed requirements. Requirements are based on the modules used.

The open source ROCKS project and StackIQ have been distributing Ganglia as a roll in their latest releases. The Rocks 5.4 Ganglia roll includes version 3.1.7.

Latest version: 3.3.1 (Feb-08-12).

Scheduling and Management

Grid Engine

Sun Grid Engine
Versions 6.2u5 and earlier. SGE Open-source Created by Sun. Since Oracle’s acquisition of Sun, SGE is only available for download through Oracle. Oracle versions are not OSS.
http://gridengine.info/

Open Grid Scheduler
Latest Version: Grid Engine 2011.11 (Nov-11-14). This release includes support for hwloc, GPU integration, NFSv4 and more.

Last open source version was picked up by the community to keep it free and open. OS Support states Linux x86_64. Installed from source.
http://gridscheduler.sourceforge.net

Univa Grid Engine
Latest Version: 8.0.1 (Apr-04-2011). Another fork of Oracle Grid Engine is Grid Engine created by Univa. Univa is dedicated to working closely with the open source Open Grid Scheduler team but will offer commercial support including automated migration service from other schedulers like Platform LSF. Univa’s GE datasheet currently shows support for RHEL up to 5.6 and SLES 10 and 11.

Both Rocks and Rocks+ currently provide a roll for SGE6.2u4.

PBS Professional
http://www.pbsworks.com/Product.aspx?id=1

Altair’s commercial-grade HPC workload management solution. A single, unified system designed to easily create intelligent policies to manage computing assets. License required for use. Free 30-day temporary licenses available for up to 700 cores. Longer term trial licenses available upon request.

Latest Version: 11.2 (Jun-30-2011) Fully supported on Intel Xeon Processor hardware with targeted operating systems. New versions release twice a year with the next release expected Q2 2012.

Platform LSF
http://www.platform.com/workload-management/high-performance-computing

According to Platform Computing, Platform LSF is the most powerful workload manager for demanding, distributed and mission-critical high performance computing environments. It provides a complete set of workload management capabilities; all designed to work together. A license is required for use, but a demo license is available upon request.

Latest Version 8 (Nov-15-2010). Current version of Platform HPC includes LSF 7 update 6. Supported OS chart on their site indicates that all version since 6 have supported RHEL 5 and SLES 11. RHEL 6 support is available in 7u6 and 8.

SLURM (Simple Linux Utility for Resource Management)
https://computing.llnl.gov/linux/slurm/overview.html

SLURM is a self-contained, open source cluster management and job scheduling system which boasts its high scalability and fault tolerance. It can use MUNGE for authentication and is configurable to use with Maui for more advanced scheduling. OS support includes Linux, Solaris, AIX, and OS X. Debian and Ubuntu packages available.

SLURM is the default resource manager in Bright Cluster Manager.

Latest Version: 2.3.3 (Jan-24-2012) Added support for Cray XT and XE computers.

TORQUE
http://www.adaptivecomputing.com/resources/docs/torque/index.php

TORQUE is the open source fork of OpenPBS, which is no longer actively developed. Active community support and updates for TORQUE are provided by Cluster Resources and Adaptive Computing. Commercial technical support offered by Adaptive Computing for those who purchase Moab products.

OS Support includes Solaris and Linux. Kernel 2.6 has been supported since version 2.2.X

Version 2.5.X
Recommended version. New features and functionality expected. Considered reliable after many successful deployments. Latest is 2.5.9 (Nov-05-11). Estimated at Launch: 2.5.10 Updates have been 1-3 months apart.

*Other names and brands may be claimed as the property of others.
Version 3.0.X
This includes features of 2.5.X plus NUMA support. Also supports multiple Machine Oriented Mini-server (MOM) daemons on a single host. 3.0.X is not backwards compatible and is only recommended if NUMA or multi-mom support is needed. Latest is 3.0.4 (Jan-16-12) Estimated at launch: 3.0.5 Updates have been about 3 months apart.

Version 4.0.X
Higher scalability using job radix for each MOM to communicate with other MOMs instead of having each MOM report back to a single superior, eliminating a single stress point. Adds higher security using new authorization daemon trqauthd which runs as root. Latest is 4.0.0 (Feb-21-12)

Intel Developer Tools

Intel Cluster Studio XE 2012
Intel Cluster Studio 2012 XE package adds the Intel Vtune Amplifier XE, Intel Inspector XE for Linux in Q4 2011, with Intel Xeon processor E5 support, for all popular Linux distributions.

Intel Compilers
Intel compiler application optimization for Intel Xeon processor E5, by version is shown below.

- 11.0: limited AVX support
- 11.1: adding -xavx switch, limited optimization
- 12.1: best application optimization results for Intel Xeon processor E5 is with 12.1 U6 (Sep 2011) which is part Intel Cluster Studio XE 2012, or later

<table>
<thead>
<tr>
<th>Linux Release</th>
<th>Intel® 64 32-bit apps</th>
<th>Intel® 64 64-bit apps</th>
</tr>
</thead>
<tbody>
<tr>
<td>RedHat* EL5</td>
<td>S</td>
<td>V</td>
</tr>
<tr>
<td>RedHat EL6</td>
<td>BE</td>
<td>S</td>
</tr>
<tr>
<td>SUSE* SLES 10</td>
<td>BE</td>
<td>S</td>
</tr>
<tr>
<td>SUSE SLES 11</td>
<td>S</td>
<td>V</td>
</tr>
<tr>
<td>ICR 1.2</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Fedora* 15/16</td>
<td>BE</td>
<td></td>
</tr>
<tr>
<td>OpenSUSE 11.1</td>
<td>BE</td>
<td></td>
</tr>
<tr>
<td>Asianux* 4¥</td>
<td>BE</td>
<td></td>
</tr>
<tr>
<td>Ubuntu* 10.0.4#</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Debian* 6¥</td>
<td>BE</td>
<td></td>
</tr>
<tr>
<td>CAOS* 1€</td>
<td>BE</td>
<td></td>
</tr>
<tr>
<td>CentOS* 5.3€</td>
<td>BE</td>
<td></td>
</tr>
</tbody>
</table>

V: Validated
S: Supported
BE: Best Effort
€ Intel MPI only
¥ Regarding the use of the Intel MPI runtime library to support co-array Fortran on single node platforms running Debian, Ubuntu, and Asianux OSes, the Intel MPI runtime library is required to work.

*Other names and brands may be claimed as the property of others.
Intel Compiler Support for Intel 64

<table>
<thead>
<tr>
<th>Intel Compilers</th>
<th>Intel® 64 32-bit apps</th>
<th>Intel® 64 64-bit apps</th>
</tr>
</thead>
<tbody>
<tr>
<td>12.1</td>
<td>S</td>
<td>V</td>
</tr>
<tr>
<td>12.0</td>
<td>BE</td>
<td>S</td>
</tr>
<tr>
<td>11.1</td>
<td>BE</td>
<td>S</td>
</tr>
</tbody>
</table>

Intel MPI

The current Intel MPI Library release, which is available from the Intel Registration Center, is Version 4.0 (Update 3), which was released 15 Sep 2011.

Intel MPI Library will automatically recognize execution on an E5-2600 processor, so that architecture-specific optimization can be applied on the fly. Current E5-2600-specific enhancements are designed to optimize performance with the use of shared memory fabric and collective operations.

When using the Intel MPI Library with RHEL 6.x and its included RDMA, be aware of special installation requirements. These are listed in the Install Guide section.

Intel Runtime Libraries

The current Intel Runtime Library package, which is available for download from Intel Registration Center under the Intel Cluster Ready product, is version 3.3-1. Version 3.3-1 was released in late September.

This new version contains the runtime files of the following components:

- Intel C++ Compiler XE version 12.0 Update 4 for Linux
- Intel Fortran Compiler XE version 12.0 Update 4 for Linux
- Intel Math Kernel Library 10.3 Update 4 for Linux
- Intel MPI Library 4.0 Update 3 for Linux
- Intel Threading Building Blocks 3.0 Update 7 for Linux

Communication Fabrics

Open Fabrics Enterprise Distribution (OFED)

The latest OFED version provides full support for InfiniBand*, iWARP and RDMA over Converged Ethernet (RoCE).

Version 1.5.4.1 is the current version that was launched in early February 2012. OFED 1.5.4.1 includes a large number of bug fixes and improvements. FDR Infiniband is supported since OFED-1.5.4. It is highly recommended to use this OFED version.

Under certain circumstances it might be required to use an older OFED version, but do not use version 1.5.3.0, released in February 2011, as it contains a critical bug affecting QLogic* InfiniBand* products.

It is important to note that the libibverbs included in OFED 1.5.X.X is currently not fully compatible with InfiniBand kernel modules included with the Linux distributions. For example, XRC support requires the full OFED 1.5.3.2 distribution. The ofa-kernel components will need to be compiled in order to support an updated OFED 1.5.3.2. Another issue is that OFED distributions included in some Linux distributions are incomplete; components such as InfiniBand debug utilities may be missing. In such cases it is recommended to download and recompile the complete OFED distribution.

*Other names and brands may be claimed as the property of others.
RDMA over Converged Ethernet (RoCE) is enabled in 1.5.3.2 and later versions, but it requires a supported adapter. Mellanox currently offers these adapters. There is no software RoCE support included in the OFED release at the moment, but is available through third-party software.

Intel continues to work with the Interoperability Lab (IOL) at University of New Hampshire to help support compatibility testing and new driver development.

The next major release, OFED 2.0, which will use a new Linux kernel base, is planned for 2012.

InfiniBand Technology

Any information provided in this document is reproduced from these public websites:


http://www.mellanox.com

A new InfiniBand technology is available on the market. FDR (Fourteen Data Rate) upgrades the usable bandwidth of a 4x connection to 56 Gbps. Active optical cables for FDR have already been demonstrated.

All currently available FDR products are dependent on availability of the require a PCI Express 3.0 interface. FDR products using standard x8 PCI Express 2.x interfaces do not exist and cannot sustain full bandwidth.

The next technology upgrade will be EDR (Enhanced Data Rate) with 100Gbps usable bandwidth on a 4x connection. No EDR products are expected until at least mid-2012.

There are new QDR, FDR, 40GbE and 10GbE products expected to be available and supported under OFED 1.5.3.1 and newer.

New Intel servers are expected to have Mellanox InfiniBand mezzanine options. While not manufactured directly onto the motherboard, Intel server product teams treat these as onboard options.

The InfiniBand roadmap can be found here:

http://www.infinibandta.org/content/pages.php?pg=technology_overview

Cabling options for InfiniBand

Four cabling options are currently available for InfiniBand cabling: active and passive copper cables, active optical cables and optical transceivers. The performance in terms of throughput and latency is very similar across all four cabling options. The main differences lie in deployment, maintenance and overall cost.
Cabling options for InfiniBand fabric

<table>
<thead>
<tr>
<th>Cabling type</th>
<th>Copper cables</th>
<th>Active optical cables</th>
<th>Optical Transceivers</th>
</tr>
</thead>
<tbody>
<tr>
<td>Maximum length</td>
<td>~5-7m with passive cables up to 20m with active cables and both sides active (QDR)</td>
<td>~100-300m</td>
<td>Up to 10km</td>
</tr>
<tr>
<td>Relative power consumption</td>
<td>High</td>
<td>Low</td>
<td>Low</td>
</tr>
<tr>
<td>Relative total price</td>
<td>Low</td>
<td>Medium</td>
<td>High</td>
</tr>
</tbody>
</table>

**Advantages**
- cheapest cabling option
- longer reach than copper
- lower weight cables than copper
- tighter bending radius than copper
- improved airflow
- cheaper than solutions based on optical transceivers
- no optical connectors so no cleanliness issues
- longer reach than copper
- lower cable weight than copper
- tighter bending radius than copper
- improved airflow
- can use structured cabling infrastructure/patch panels
- easiest option to deploy, esp. under space constraint conditions or with challenging cable routing
- simplest maintenance

**Disadvantages**
- very short reach of passive copper cables
- relatively high power consumption
- potential cooling, weight and cabling issues due to bulky cables
- active cables require at least one IB port to supply power
- higher price than copper
- requires IB port to supply power
- more complicated maintenance compared to optical transceivers
- most expensive option
- many components
- optical connector so potential cleanliness issues

Intel Network Adapter Drivers

There are three drivers which support new adapters and on-motherboard devices. The e1000e driver will be used for entry-level and desktop adapters. The igb driver is used for 1 GbE PCI Express server adapters. The ixgbe driver is used for 10 GbE adapters.

<table>
<thead>
<tr>
<th>Intel Network Adapter Drivers</th>
<th>Speed</th>
<th>Driver - Latest driver version</th>
</tr>
</thead>
<tbody>
<tr>
<td>Intel X520 (82599)</td>
<td>10GbE</td>
<td>ixgbe - 3.8.21</td>
</tr>
<tr>
<td>Intel X540</td>
<td>10GbE</td>
<td>ixgbe - 3.8.21</td>
</tr>
<tr>
<td>Intel 82580</td>
<td>1GbE</td>
<td>igb - 3.3.6</td>
</tr>
<tr>
<td>Intel I350</td>
<td>1GbE</td>
<td>igb - 3.3.6</td>
</tr>
<tr>
<td>Intel 82574</td>
<td>1GbE</td>
<td>e1000e - 1.9.5</td>
</tr>
</tbody>
</table>

The source code for the drivers is available through Intel at http://downloadcenter.intel.com and through SourceForge at http://sourceforge.net/projects/e1000. The drivers are expected to support all target operating systems. Complete driver binaries for new platforms are available with the Intel Network Adapter Drivers Package 16.8.1.

*Other names and brands may be claimed as the property of others.*

**Intel 10GbE Adapters**

The Intel X540 10GbE Adapter complements the 82599 controller as a 10GBASE-T (RJ45 connector) 10 GbE solution for Intel products. Intel X540-based adapters will be available through an I/O module option on EPSD servers. Major OEM servers will also offer X540 controllers as onboard 10 GbE interfaces. This is in addition to standard add-in adapter cards.

The LAN-On-Motherboard option will provide a pre-boot environment which will enable PXE boot, iSCSI boot, FCoE boot and UFEI boot.

**Intel 1GbE Adapters**

Intel I350 is expected to be the standard 1 GbE network adapter chipset for Intel Xeon processor E5 platform launch. The Intel 82574L 1GbE adapter is designed for entry-level systems and may not be able to maintain the full bandwidth of MPI applications.

**Intel NetEffect iWARP Adapters**

The Intel NetEffect NE020 iWarp adapter remains the standard iWarp platform through next year. Full support for iWarp with the NE020 adapter is now provided as part of OFED. The source distribution from [http://www.openfrabrics.org](http://www.openfrabrics.org) should be used, as releases from other vendors may remove support required for iWARP. Intel provides these OFED packages, along with a readme file and a setup wrapper on Intel’s Download Center, [http://downloadcenter.intel.com](http://downloadcenter.intel.com). Default NE020 drivers included in RHEL 5.x distributions and earlier are old versions which may not fully support desired features.

Support for booting via the NE020 adapter and provisioning kernels will require a minimal OFED implementation to be included with the boot kernel.

**Cabling options for 10GbE**

The Intel 10GbE Adapters offer a variety of cabling options. There are, essentially, three options: Twisted Pair, SFP+ Direct Attach (DA) cables and Optical cables, either on the NIC or as SFP+ pluggable modules. The CX4 standard is considered obsolete and is superseded by the CR (SFP+ DA) standard. Each of the cabling options offer certain advantages, but also have disadvantages.
### Cabling options for 10GbE

<table>
<thead>
<tr>
<th>SFP+ DA</th>
<th>Twisted Pair</th>
<th>(SFP+) Optics</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Maximum length</strong></td>
<td>7m (10m specified but not used)</td>
<td>100m with Cat6a (37-55m with Cat6)</td>
</tr>
<tr>
<td><strong>Relative power consumption</strong></td>
<td>Lowest</td>
<td>High</td>
</tr>
<tr>
<td><strong>Relative latency</strong></td>
<td>Lowest</td>
<td>Medium</td>
</tr>
<tr>
<td><strong>Relative total price</strong> (NIC + cable + optics, if applicable)</td>
<td>Low</td>
<td>Lowest</td>
</tr>
</tbody>
</table>

#### Advantages
- Lowest power consumption
- Switches with <200ns port-to-port latency available
- Large selection of equipment vendors
- Can be offered as LOM option
- Switches can easily be equipped with SFP+ optics modules for covering longer distances
- Offers upgrade path to 40GbE via QSFP+-to-SFP+_DA patch cables

#### Disadvantages
- Limited cable length
- More complicated cabling compared to Twisted Pair Cat6a cabling
- SFP+ DA cables have to be pulled as cable assembly and cannot be terminated in place
- High power consumption, factor 2-3 higher than SFP+ DA, expected to be reduced with future generation PHYs
- Higher latency compared to SFP+ DA
- Cat5e officially NOT supported
- Cat6 supported, but should not be used in an HPC cluster environment

#### Other Network Adapters

Additional information regarding network or high performance fabric components from other vendors may be added in future versions of this document.

*Other names and brands may be claimed as the property of others.*
Cluster File Systems

This section covers a high-level overview of several current parallel or cluster storage filesystems. While there are many options available, four different filesystems are covered in this document. For the purposes of this document, a cluster file system includes the software stack, design, and operation of the cluster storage solution. This is normally separate from the physical hardware used.

The fundamental difference between standard file systems, such as NFS, and cluster file systems is the decoupling of the data (such as files) and the metadata (how these files are stored). These designs implement a separate metadata service or server that holds information such as location, structure, and attributes. This provides capabilities such as the distribution of files and directories across multiple storage sites.

The cluster file systems that are currently part of research for this document, and are available free for commercial use include:

- Parallel Virtual File System version 2 (PVFS2)
- Lustre
- GlusterFS
- Parallel NFS (pNFS)
- Ceph
- Fraunhofer File System (FhGFS)
- Hadoop Distributed File System (HDFS)

From the current research, information is provided here about 4 of these file systems. In addition, research includes commercial file systems such as GPFS and CXFS, as well as commercial storage solutions, including those from Panasas, IBRIX (HP), Isilon, and others.

GlusterFS

http://www.gluster.org

Gluster is both a community-based project and a commercial product. The Gluster file system (GlusterFS) aggregates distributed storage resources into a single, global namespace. This namespace is available on all storage systems and data can be accessed over TCP/IP or RDMA. GlusterFS is a userspace implementation. A complete GlusterFS server implementation is available for download, and is also available as a virtual appliance. Gluster was recently acquired by Red Hat.

Current version: The current version of Gluster is 3.2.5. CentOS or RHEL 5.1 or greater and Ubuntu 8.04 or greater are recommended. Source code is available for any OS.

Lustre

http://www.lustre.org

The Lustre project was fostered by Cluster File Systems and is the most widely-used open source file system for clusters, and is used on many of the Top500 (including the top 10) clusters. The project is currently owned by Oracle. Whamcloud develops Lustre solutions to address known issues and supports current distributions, as well as provides commercial management solutions for Lustre.

A Lustre client requests data from a metadata target (MDT), which refers the client to an object storage server (OSS), which accesses the data on one or more object storage targets (OST). Lustre isolates clients from direct access to the files on the object storage targets, which simplifies features such as distributed file locking and file system repair, and allows for simplified upgrades of storage.
Current version: The current versions of Lustre are 1.8.5 and 2.0.0 and are supported on Oracle Enterprise Linux 5, RHEL 5, SLES 10 and 11, Scientific Linux 5, and Fedora 11. Upgrades from 1.8.x to 2.0.x are not supported. Supported Linux versions should be the same through 2012. Support for Lustre 1.8.x will end mid-2012.

Newer 1.8.7-wc1 and 2.2 versions of Lustre can be downloaded from Whamcloud. These include support for RHEL 6.

Parallel NFS

http://www.pnfs.com

Parallel NFS (pNFS) is now part of the NFS 4.1 standard, and provides distributed file access at the file, block, and object levels. Unlike other file systems, it does not define how communication between metadata servers and data stores is performed. The pluggable client obtains layout information from metadata servers to directly access data stored on the network. pNFS support is provided by a broad group of industry vendors.

Current version: pNFS initial releases are available for download. It is expected to be standard in Linux releases in 2H'11. RHEL 6.2 includes a tech preview of pNFS.

PVFS2

http://www.pvfs.org

The Parallel Virtual File System version 2 is an open source joint project developed and funded through U.S. national labs, government agencies, and universities. The original version was released in the early 1990’s and the current PVFS2 file system was initially launched in 2003. PVFS2 can be compiled on current Linux distributions without patches. PVFS2 also supports MPI-IO (ROMIO) and IB Verbs.

Current version: The OrangeFS branch, which has become the current standard and offers commercial support, is updated to 2.8.4. Microsoft Windows support is also available.

Comparison

The following table compares several of the free file systems available, based on both use and the most current documentation available on January 31st, 2012. In many cases, a feature is available only through a beta or 3rd-party add-on. These items are marked “no”. The following definitions are used:

- Direct file access: Clients are able to access the file system directly at the file level
- Direct block access: Clients are able to access the file system directly at the block or device level
- FUSE: File system in USEr space. File system is loadable, configurable, and mountable in user space, using a the FUSE kernel module as a bridge
- Optimized Data Paths: File system is able to direct clients to closest data, if multiple copies exist.
- Data balancing: Supports redistributing data across storage systems.
- Distributed Locks: Supports locking mechanism allowing multiple clients simultaneous access to a file.
- Distributed hard links: Supports creating hard links on different storage systems
- MPI-IO Optimization: File system provides integrated support for MPI-IO or ROMIO

*Other names and brands may be claimed as the property of others.
<table>
<thead>
<tr>
<th>Comparison of Parallel and Distributed File Systems</th>
<th>Lustre</th>
<th>GlusterFS</th>
<th>PVFS2</th>
<th>pNFS</th>
<th>HDFS</th>
<th>FhGFS</th>
<th>Ceph</th>
</tr>
</thead>
<tbody>
<tr>
<td>Source Available</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
<td>No</td>
<td>Yes</td>
</tr>
<tr>
<td>Commercial</td>
<td>Yes/No</td>
<td>Yes</td>
<td>No</td>
<td>No</td>
<td>No</td>
<td>Yes</td>
<td>No</td>
</tr>
<tr>
<td>Free</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
</tr>
<tr>
<td>License</td>
<td>GNU GPL</td>
<td>GNU GPL</td>
<td>GNU GPL</td>
<td>GNU GPL</td>
<td>Apache Prop</td>
<td>GNU GPL</td>
<td></td>
</tr>
<tr>
<td>Current Version</td>
<td>2.1.1</td>
<td>3.2.5*</td>
<td>2.8.41</td>
<td>4.12</td>
<td>1.0.0</td>
<td>2011-04r15</td>
<td>0.43</td>
</tr>
<tr>
<td>Feature Support</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Distributed metadata servers</td>
<td>Yes</td>
<td>No</td>
<td>Yes</td>
<td>No4</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
</tr>
<tr>
<td>Client Access Options</td>
<td>Native</td>
<td>Native, NFS</td>
<td>Native, pNFS</td>
<td>NFS</td>
<td>Native</td>
<td>Native</td>
<td>Native</td>
</tr>
<tr>
<td>Direct file access</td>
<td>No</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
<td>No</td>
<td>Yes</td>
<td>Yes</td>
</tr>
<tr>
<td>Direct block access</td>
<td>No</td>
<td>No</td>
<td>No</td>
<td>Yes</td>
<td>No</td>
<td>No</td>
<td>RBD driver</td>
</tr>
<tr>
<td>Kernel Patch Required</td>
<td>Yes</td>
<td>No</td>
<td>No</td>
<td>No5</td>
<td>No</td>
<td>No</td>
<td>Included</td>
</tr>
<tr>
<td>FUSE</td>
<td>Yes</td>
<td>Required</td>
<td>Yes</td>
<td>No</td>
<td>No</td>
<td>No</td>
<td>Yes</td>
</tr>
<tr>
<td>Disk file format</td>
<td>Mod ext4</td>
<td>Any3</td>
<td>No</td>
<td>Any4</td>
<td>Native</td>
<td>Any</td>
<td>BTRFS (Any)</td>
</tr>
<tr>
<td>Metadata Server Failover</td>
<td>Yes</td>
<td>Yes</td>
<td>No</td>
<td>No4</td>
<td>Yes</td>
<td>No</td>
<td>No</td>
</tr>
<tr>
<td>Quotas</td>
<td>Yes</td>
<td>Yes</td>
<td>No</td>
<td>No4</td>
<td>Yes</td>
<td>No</td>
<td>No</td>
</tr>
<tr>
<td>Load balancing</td>
<td>No</td>
<td>Yes</td>
<td>No</td>
<td>No4</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
</tr>
<tr>
<td>Stripping</td>
<td>Yes</td>
<td>Yes</td>
<td>No</td>
<td>No4</td>
<td>No</td>
<td>Yes</td>
<td>0 only</td>
</tr>
<tr>
<td>Mirroring</td>
<td>No</td>
<td>Yes</td>
<td>No</td>
<td>No4</td>
<td>Yes</td>
<td>No</td>
<td>No</td>
</tr>
<tr>
<td>Data replication</td>
<td>Remote</td>
<td>Yes</td>
<td>No</td>
<td>No4</td>
<td>Yes</td>
<td>No</td>
<td>Yes</td>
</tr>
<tr>
<td>Optimized Data Paths</td>
<td>No</td>
<td>No</td>
<td>No</td>
<td>No4</td>
<td>Yes</td>
<td>Yes</td>
<td>No</td>
</tr>
<tr>
<td>Data balancing</td>
<td>No</td>
<td>No</td>
<td>No</td>
<td>No4</td>
<td>Analyzer</td>
<td>No</td>
<td>No</td>
</tr>
<tr>
<td>Distributed Locks</td>
<td>Yes</td>
<td>No</td>
<td>No</td>
<td>No4</td>
<td>No</td>
<td>No</td>
<td>No</td>
</tr>
<tr>
<td>Distributed hard links</td>
<td>Yes</td>
<td>No</td>
<td>No</td>
<td>No4</td>
<td>Yes</td>
<td>No</td>
<td>No</td>
</tr>
<tr>
<td>MPI-IO Optimization</td>
<td>Yes</td>
<td>No</td>
<td>Yes</td>
<td>No</td>
<td>No</td>
<td>No</td>
<td>No</td>
</tr>
<tr>
<td>IB Verbs Support</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
<td>No</td>
<td>No</td>
<td>Yes</td>
<td>No</td>
</tr>
<tr>
<td>Metadata transaction logs</td>
<td>Yes</td>
<td>No</td>
<td>No</td>
<td>No4</td>
<td>No</td>
<td>No</td>
<td>No</td>
</tr>
</tbody>
</table>

1 Released under OrangeFS
2 Included in NFS.
3 Ext3 or Ext4 recommended. XFS recommended for >100 GB files.
4 Depends upon backend solution used.
5 May be required for older kernels.

*Other names and brands may be claimed as the property of others.*
HPC Applications

Many HPC applications are not expected to have support for RHEL 6.x versions and their clones in the time frame of Intel Xeon processor E5 launch. This does not mean that an application does not work correctly, but official support may not be available from the ISV. Lack of official support may mean no support at all, best effort support, or no advertised support. Most vendors are expected to offer support for RHEL 6.x sometime in 2012.

Some applications do not formally support Red Hat distributions, instead supporting common source distributions such as CentOS or Scientific Linux.

Intel Cluster Ready

Intel Cluster Ready specification 1.2 is recommended for Intel Xeon processor E5 solutions. Instead of reproducing the specification, which is available for download, this table lists the most significant changes from version 1.1.

<table>
<thead>
<tr>
<th>Guide to Changes in Intel Cluster Ready Specification 1.2</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>ICR v1.2 Requirement</strong></td>
</tr>
<tr>
<td>4.1 Base Cluster Requirements</td>
</tr>
<tr>
<td>4.2 Base Software Requirements</td>
</tr>
<tr>
<td></td>
</tr>
<tr>
<td>6. Cluster Functions</td>
</tr>
<tr>
<td>6.2 OS Interface and Basic Runtime Environment</td>
</tr>
<tr>
<td>6.3 Command System and Tools</td>
</tr>
<tr>
<td></td>
</tr>
<tr>
<td>6.4 Message Passing Layer</td>
</tr>
<tr>
<td>6.6 Message Fabric</td>
</tr>
<tr>
<td>6.7 Remote Execution Environment</td>
</tr>
</tbody>
</table>

*Other names and brands may be claimed as the property of others.*
**Intel Cluster Checker**

Features specific to Intel Xeon processor E5 and Romley-based products were introduced in Intel Cluster Checker version 1.8, which was released in August. Version 1.8 of the Intel Cluster Checker for Linux provides these new and improved features:

- Intel® Cluster Ready Specification version 1.2 verification
- Support for second generation Intel Core(TM) processors
- Updated prepackaged benchmarks
- Simplified Test Module List
- Expanded *Intel Cluster Checker Test Modules Reference Guide*

*Other names and brands may be claimed as the property of others.*
Install Guide

The following section describes changes or recommendations for installing Intel Xeon processor E5 platform systems. Included here are specific steps or configuration options needed to successfully install new systems.

Intel Cluster Ready Reference Implementations

Intel will provide updated reference implementations for Intel Xeon processor E5-based platforms. These reference implementations include written recipes, scripts, and preconfigured Intel Cluster Checker input files, which makes creating a solution very easy. They also include InfiniBand support. Reference implementations include:

- Bright* Cluster Manager 5.2, RHEL 6.2
- Bright* Cluster Manager 5.2, RHEL 5.7
- Platform* Cluster Manager 3.1, RHEL 6.2, with OFED

Reference implementations may be downloaded here:

Intel Processor Settings

Turbo mode

This new mode in Intel* server boards helps workloads that are CPU bound to speed up by several percent by increasing the peak processing bounds past the thermal design power for a brief time. Turbo mode takes the capabilities of processors used, the number of active cores, the temperature of the processors, and the estimated power consumption of the processors into account, and then calculates how far over the normal frequency each core can be pushed.

For HPC applications, most CPU bound workloads benefit from having Turbo mode enabled by several percentage points.

It is advised to set Turbo mode ON.

Hyperthreading/Simultaneous Multi Processing (HT/SMT)

HT/SMT is Intel's way of scheduling poorly-threaded applications so that a single core can function like two virtual cores. Use of HT/SMT helps very few workloads. In a few workloads, having HT/SMT actually hurts performance. Performance increases in those workloads that HT/SMT helps is generally low.

It is advised to set HT/SMT OFF.

Enhanced Intel SpeedStep Technology (EIST)

EIST dynamically adjusts processor voltage and core frequency, to reduce heat production and power consumption during times when an individual core is idle. There is a brief moment during the start of an HPC workload execution when the core frequency is low where the workload will not have the full power of the processor available. However, EIST, and the associated P-states, is needed to access Turbo mode.

It is advised to turn EIST OFF when testing an HPC application's performance, and ON at all other times.
Non-Uniform Memory Access (NUMA)

NUMA-enabled memory architecture causes memory local to a processor socket to be accessed significantly faster than memory attached to a different processor socket. NUMA-aware software can take advantage of this to increase performance for some workloads. However, non-NUMA-aware software can take a performance hit from enabling NUMA in BIOS.

Typically, it is advised to have run “NUMA Enabled” in BIOS for applications that are mainly running processes, such as serial code or MPI type codes. If an application is mainly running internal threads, such as OpenMP codes, the application engineer should experiment whether their individual code has higher performance with NUMA either enabled or disabled, as the solution isn’t necessarily obvious.

Linux Installation

Red Hat Enterprise Linux (RHEL) 5.7, 5.8, 6.1 and 6.2 have all of the required drivers necessary for installation. SUSE Linux Enterprise Server 11 SP1 has all of the required drivers necessary for installation excluding SAS drivers. SUSE is planning a Maintenance Release of SLES11SP1 that includes the necessary SAS drivers.

Follow the RHEL5.7, 5.8, 6.1, 6.2, or SLES 11SP1, SLES 11SP2 instructions that come with each distribution to install on Intel Xeon processor E5-based servers. There are no additional instructions for installation of Linux on these servers beyond attaining the necessary distribution version level as detailed in the Support Matrix and Operating Systems sections of this guide.

Intel MPI with RDMA

There is a known issue and workaround for use of Intel MPI the default RDMA on RHEL 5.7, 5.8, 6.1, and 6.2.

- Do a customize install (via the RHEL install GUI) and manually select all packages under the InfiniBand support category.
- The dapl-devel package is the one that creates the symbolic links for libdat.so and libdat2.so that are needed by Intel MPI. Red Hat does not install any of the “devel” packages by default. The dapl-devel rpm is not installed by the GUI even if you select everything under the infiniband support category. You will have to go to the installation media and install the RPM manually.
- Red Hat does not start the OFA drivers or opensm by default. You have to manually start them or use `chkconfig` to change their startup configuration.
- Note that the startup script was renamed from /etc/init.d/openib to /etc/init.d/rdma
- In RHEL 6.1 you must use `/sbin/ifconfig` for the IPoIB interface using the normal method, i.e., adding `/etc/sysconfig/network-scripts/ifcfg-ib0`. It appears that those scripts want to use the `ifconfig` command to configure the device and in EL 6.1, `ifconfig` does not work for InfiniBand. You have to use `/sbin/ifconfig` instead.

RSTe3 drivers

Although the isci driver is included with RHEL6.1, it should be updated to a newer version. This driver should also be included in any initrd used for initial node boot. Provisioning on Intel Xeon Processor E5-based compute nodes may fail without this driver. The latest updates can be found at [http://sourceforge.net/projects/intel-sas/files](http://sourceforge.net/projects/intel-sas/files).

This will bring the isci driver on par with upstream 3.0 kernel driver. It does not contain any libsas and libata updates, due to required changes needed in libata which is built into the kernel. A Z-stream update from Red Hat is required for libsas and libata and customers need to request a Driver Update Process (DUP) directly from Red Hat to resolve the issue. The lack of libsas/libata update severely impacts SAS expander stability and SGPIO support for RHEL 6.1.

*Other names and brands may be claimed as the property of others.*
Intel Server Platforms

Installation requirements are provided at individual product support pages.

Installing IPMI Tools

Tools and utilities necessary to use IPMI are included with most Linux distributions.

RHEL5.7 and later

RedHat Enterprise Linux 5.7 automatically installs the necessary OpenIPMI libraries and utilities needed for configuring and using IPMI on Intel Xeon Processor servers.

RHEL6.1 and later

RedHat Enterprise Linux 6.1 ships the necessary OpenIPMI and ipmitool packages with the ISO, but does not install them automatically.

1. Install the OpenIPMI and ipmitool rpm packages from the RHEL6.1 installation package.
2. Turn on the OpenIPMI devices to start automatically
   
   chkconfig ipmi on

3. Start the OpenIPMI service
   
   service ipmi start

SLES11SP1

Suse Linux Enterprise Server 11SP1 automatically installs the necessary OpenIPMI libraries and ipmitool.
Setting up Out-Of-Band IPMI for Static IP Access

For channel numbers, use “1” for an integrated channel. When setting the user name, also assign an integer user number to the name. The privilege level for administrators is “4”.

1. Ensure In-Band IPMI functionality is working.
2. Enable static IP addresses
   ```bash
   ipmitool lan set \<channel number\> ipsrc static
   ```
3. Set an IP address to the BMC from the server being set up
   ```bash
   ipmitool lan set \<channel number\> ipaddr \<ip address\>
   ```
4. Set the Netmask for that IP address
   ```bash
   ipmitool lan set \<channel number\> netmask \<netmask\>
   ```
5. Set the gateway for that IP address
   ```bash
   ipmitool lan set \<channel number\> defgw ipaddr \<gateway IP address\>
   ```
6. Enable proper ARP responses
   ```bash
   ipmitool lan set \<channel number\> arp respond on
   ```
7. Set Admin authorization on the integrated port
   ```bash
   ipmitool lan set \<channel number\> auth ADMIN NONE
   ```
8. Set a user name to access IPMI functionality over lan
   ```bash
   ipmitool user set name \<user number\> \<username\>
   ```
9. Set a password for that user
   ```bash
   ipmitool user set password \<user number\> \<password\>
   ```
10. Set user privileges
    ```bash
        ipmitool user priv \<user number\> \<privilege level\> \<channel number\>
    ```
11. Enable user
    ```bash
        ipmitool user enable \<user number\>
    ```
12. Install OpenIPMI and ipmitool packages on a separate server, if necessary.
13. Test the Out-Of-Band functionality on the separate server
    ```bash
        ipmitool -I lan -H \<ip address\> -U \<username\> -f \<password\> chassis status
    ```

*Other names and brands may be claimed as the property of others.*
Compiling OFED

There are several sources for the OFED software stack. The “official” OFED stack can be obtained from http://www.openfabrics.org. There are also OFED stacks provided and customized by vendors that use OFED.

This guide will provide guidelines for installing the original OFED stack from http://www.openfabrics.org and for the OFED stack provided by Intel at http://downloadcenter.intel.com.

According to the README.txt provided by OFED, certain OS software packages are required for compiling the OFED kernel modules and various optional components.

### OS software packages required to compile OFED kernel modules

<table>
<thead>
<tr>
<th>Linux Distribution</th>
<th>Required Packages</th>
</tr>
</thead>
<tbody>
<tr>
<td>All distributions</td>
<td>gcc, glib, glibc-devel, glibc, glibc-devel, zlib-devel, libstdc++-devel; in order to compile the 32-bit libraries on a 64-bit platform the 32-bit version of glibc-devel has to be installed</td>
</tr>
<tr>
<td>RHEL, Fedora</td>
<td>kernel-devel, rpm-build, redhat-rpm-config</td>
</tr>
<tr>
<td>SLES, openSUSE</td>
<td>kernel-source, rpm</td>
</tr>
</tbody>
</table>

### OS software packages required to compile optional OFED components

<table>
<thead>
<tr>
<th>Component</th>
<th>Required Packages</th>
</tr>
</thead>
<tbody>
<tr>
<td>MVAPICH</td>
<td>a Fortran Compiler (such as gcc-g77)</td>
</tr>
<tr>
<td>MVAPICH2</td>
<td>libsysfs-devel</td>
</tr>
<tr>
<td>OpenMPI</td>
<td>libsysfs-devel</td>
</tr>
<tr>
<td>ibutils</td>
<td>tcl-8.4, tcl-devel-8.4, tk, libstdc++-devel</td>
</tr>
<tr>
<td>mstflint</td>
<td>libstdc++-devel (32-bit on ppc64), gcc-c++</td>
</tr>
<tr>
<td>rnfs-utils</td>
<td>RHEL, Fedora: krb5-devel, krb5-libs, libevent-devel, nfs-utils-lib-devel, openldap-devel, e2fsprogs-devel</td>
</tr>
<tr>
<td></td>
<td>SLES11: krb5-devel, libevent-devel, nfsidmap-devel, libopenssl-devel, libblkid-devel</td>
</tr>
</tbody>
</table>

*Other names and brands may be claimed as the property of others.*
Compiling the default OFED stack

It is fairly simple to compile the OFED stack from openfabrics.org.
1. Download the latest source code from http://www.openfabrics.org/
2. Unpack the tar-ball
3. Change into the OFED-x.x.x.x directory
4. Run the install script, named ./install.pl
   a. A menu will be presented and the appropriate options can be selected
   b. Using the “--all” command line option will install all components without prompting
   c. Using the “--hpc” command line option will install components typically needed for HPC without prompting
   d. In order to also compile 32bit libraries on 64bit platforms use the “--build32” command line option

For a detailed description of all available options, please refer to the README.txt file.

Compiling OFED for the Intel NetEffect Ethernet Server Cluster Adapter

The OFED stack and install package provided by Intel at http://downloadcenter.intel.com/ is targeted specifically at the 10 Gb iWarp-enabled server adapter.
1. Download the latest source code and setup package from http://downloadcenter.intel.com/
2. The OFED-1.5.x.x.tgz package
3. The OFED-1.5.x.x_setup.tgz package
4. Unpack the OFED-1.5.x.x_setup.tgz package
5. Run the setup.sh script
   a. ./setup.sh <path to .tgz package> <install path>
   b. A menu will be presented and the appropriate options can be selected

For a detailed description of all available options, please refer to the README.txt file.

Fix an RPM incompatibility

There is an incompatibility with OFED 1.5.3.1 and newer RPM releases, such as those found in RHEL 6.1. This affects the ibutils package. To resolve this issue, follow these steps.
1. Open the file /usr/lib/rpm/macros for editing.
2. Locate the line starting with "%_check_files".
3. Comment it out using a “#” at the beginning of the line.
4. Save and exit.
Optimizing Ethernet Performance

Interrupt Throttling

For fine-grained operations that require low latency, which is typically the case with HPC applications, disable interrupt throttling. This can be controlled through an optional parameter to the kernel driver module. The driver included with the Linux distribution may not correctly support this feature, so it is recommended that you update to the latest version of the driver available from Intel at http://downloadcenter.intel.com or from SourceForge at http://sourceforge.net/projects/e1000

In order to force the processor to respond immediately to all interrupts from network adapters, the interrupt throttling feature can be disabled by setting the following option in the /etc/modprobe.conf file.

```
option <driver> InterruptThrottleRate=0
```

Replace `<driver>` with e1000, e1000e, igb or ixgbe, as appropriate.

Configure multiple interfaces differently by simply adding values separated by a comma. For example, to enable interrupt throttling in `dynamic conservative` mode on the first supported interface, e.g. eth0, and disable it on the second supported interface, e.g. eth1, use the following option.

```
option <driver> InterruptThrottleRate=3,0
```

The ixgbe driver only supports dynamic throttling mode (set value to 1).

Low Latency Interrupts

In cases where completely disabling the interrupt throttling results in an overall performance degradation due to significantly increased processor overhead, Intel network adapters using the igb and ixgbe drivers have a feature called Low Latency Interrupts (LLI). This allows for immediate generation of an interrupt upon processing receive (RX) packets that match certain criteria as set by kernel driver parameters.
Memory Configuration

For highest bandwidth, use 1333MHz or 1600MHz dual-rank ECC DDR3 1.35V-1.65V RDIMMs. Fill only one slot per channel. Best bandwidth is attained if all RDIMMs are made by the same manufacturer and are the same model.

For largest capacity, use 800MHz or 1333MHz quad-rank or dual-rank ECC DDR3 1.35V LR-DIMMS. Highest capacity RDIMMs are typically quad-rank. Fill all slots to capacity. It is recommended to use all the same manufacturer and model for all RDIMMs.

- Max Bandwidth – 1600MHz 8 GB 2Rx4 RDIMMs
- Max Capacity – 1066MHz 32GB 4Rx4 RDIMMs or 1066MHz/1333MHz 32GB 4Rx4 LRDIMMS

### Memory Configuration for Intel HPC Serverboards

<table>
<thead>
<tr>
<th>Sockets</th>
<th>CPS</th>
<th>SPC</th>
<th>DPC</th>
<th>Total Capacity</th>
<th>DPC</th>
<th>Total Capacity</th>
</tr>
</thead>
<tbody>
<tr>
<td>S2600JF</td>
<td>2</td>
<td>4</td>
<td>1</td>
<td>64 GB</td>
<td>1</td>
<td>256 GB</td>
</tr>
<tr>
<td>S2400LP</td>
<td>2</td>
<td>3</td>
<td>2</td>
<td>48 GB</td>
<td>2</td>
<td>384 GB†</td>
</tr>
<tr>
<td>S2600WP</td>
<td>2</td>
<td>4</td>
<td>2</td>
<td>64 GB</td>
<td>2</td>
<td>512 GB†</td>
</tr>
<tr>
<td>S2600GZ (Max Memory board)</td>
<td>2</td>
<td>4</td>
<td>1</td>
<td>64 GB</td>
<td>2</td>
<td>512 GB†</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>1U: 2</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>2U: 3</td>
<td></td>
<td>768 GB†</td>
</tr>
<tr>
<td>S2600IP (Max IO board)</td>
<td>2</td>
<td>4</td>
<td>2</td>
<td>64 GB</td>
<td>2</td>
<td>512 GB†</td>
</tr>
<tr>
<td>W2600CR (Workstation board)</td>
<td>2</td>
<td>4</td>
<td>2</td>
<td>64 GB</td>
<td>2</td>
<td>512 GB†</td>
</tr>
</tbody>
</table>

CPS=Channels per Socket, SPC= Slots per Channel, DPC = DIMMs per Channel
† If using 4Rx4 RDIMMs, 2 DPC will downclock to 800 MHZ
‡ For 3 DPC using 4Rx4 32GB modules, LRDIMMs required. 1333MHz will downclock to 1066 MHz

These are guidelines for best performance, for both maximum bandwidth and maximum capacity scenarios.

- Populate each socket with the same amount of memory.
- Populate all channels with the same amount of memory.
- Ensure each channel has an even number of ranks, always the case if using 2R or 4R.
- Populate the DIMMs in the correct order as indicated by the documentation. Typically, the first slot in each channel is furthest from the socket.

*Other names and brands may be claimed as the property of others.
References

The Intel® Cluster Ready and Cluster Checker home page can be found at

http://www.intel.com/go/cluster

General information on Intel® product-support offerings may be obtained at:

http://www.intel.com/software/products/support

*Other names and brands may be claimed as the property of others.
 Obtaining Support

Your feedback is very important to us. If you would like to submit comments or have questions about this document, you can email us at icr-support@intel.com.

To receive technical support for the tools provided in this product and technical information including FAQ’s and product updates, you need to register for an Intel(R) Premier Support account at the Registration Center by registering your product serial number at

http://www.intel.com/software/products/registrationcenter

Intel ® Cluster Ready and Intel ® Cluster Checker are supported by Intel ® Premier Support. Direct customer support requests can be submitted at

https://premier.intel.com

Submitting Issues

Once you have registered your serial number(s) to obtain product support, use these steps to submit a new request.

2. Log in to the site. Note that your username and password are case-sensitive.
3. Click on the “Submit Issue” link in the left navigation bar.
4. Choose “Development Environment (tools, SDV, EAP)” from the “Product Type” drop-down list.
5. Depending on the issue, choose the “Intel® Cluster Ready ” or “Intel ® Cluster Checker” or from the “Product Name” drop-down list.
6. Enter your question and complete the fields in the windows that follow to successfully submit the issue.

Solution Guide Updates

The latest copy of this solution guide can be obtained by emailing a request to icr-support@intel.com.