bullx cluster suite XR 5v3.1U1

Software Release Bulletin

extreme computing

REFERENCE
86 A2 73EJ 00

extreme computing

bullx cluster suite XR 5v3.1U1

Software Release Bulletin

Subject: Release Notes for Bull extreme computing offer

Special Instructions: This document should be read first

Software supported: bullx cluster suite XR 5v3.1U1

August 2009

BULL CEDOC
357 AVENUE PATTON
B.P.20845
49008 ANGERS CEDEX 01
FRANCE

REFERENCE
86 A2 73EJ 00

Printed in France

Trademarks and Acknowledgements

Table of Contents

Chapter 1. Introduction

1.1 About this Software Release Bulletin

This Software Release Bulletin must be read first so that:

- Your CD/DVDs and documentation delivery package can be verified.
- The **bullx cluster suite** components may be understood.
- Information on new functionalities and licensing aspects is taken into account.
- Any known problems or restrictions in the software are noted.

1.2 bullx cluster suite Software Overview

The Bull extreme computing software suite named **bullx cluster suite XR 5v3.1U1** consists of:

			bullx cluster suite contents Open Source & Bull proprietary	Optional features Commercial Products (under license)
APPLICATION DEVELOPMENT	Scientific Libraries		See Bull Scientific Studio table	Cluster MKL (Intel®) 10.1 with blas, lapack and fft
	Parallel Libraries	Mpi	MPIBull2 1.3.9	
	Compilers (C, C++, Fortran)		GNU Fortran 4.1.2, C/C++ 4.1.2	Intel Fortran 11.0.069, C/C++ 11.0.069
TOOLS	Operation	Batch Mgr. Scheduler		PBS Pro 10.0(Altair®)
		Debuggers	gdb 6.5, Electric Fence 2.2.2	Totalview (Etnus ®) idb (Intel®), DDT (Allinea®)
		Profiling	oprofile 0.9.2	
		Performance Analysis	PAPI 3.6.2, perfctr 2.6.38, HPC Toolkit 4.9.0_1520 Hpcsnap-0.2.5	Intel® Trace tools
	Cluster Administration	Distributed shell	pdsh 2.8.1	
		Deployment / installation	ClusterDB 20.5.0, KSIS 3.0.13	
		Control & monitoring	Bull System Manager HPC Edition, Conman 0.2.1, Ganglia 3.0.5	SLURM 2.0.1 (InfiniBand & GBEthernet)
		Back-up	Bull System Backup Restore	
		Disk subsystems	Bull System Manager HPC Edition Nec_admin commands for FDA, Ddn commands for DDN, Xyratec commands for Optima, DGC commands for EMC.	
	Dump crash analysis		kdump / crash	
OS	System Administration		Bull System Manager -HPC Edition	
	File system		Ext3, NFSv3,NFSv4, Lustre 1.6.7, gfs and gfs2	
	Operating System		Linux kernel 2.6.18-128 with RHEL5 v3 u1	Bull modified Linux kernel based on Red Hat 2.6.18-128 kernel for Lustre and HPC Toolkit

1.3 Bull Scientific Studio Libraries

Open Source Libraries	bullx cluster suite XR 5v3.1U1 Version
blacs_mpibull2	blacs-1.1_p3-mpibull2_1.3.9_12.Bull
BlockSolve95 mpibull2	BlockSolve95-3.0-mpibull2_1.3.9_12.Bull
fftw2 mpibull2	fftw2-2.1.5-mpibull2_1.3.9_12.Bull
fftw3	fftw3-3.2.1-Bull.12
GlobalArray (ga)	ga-4.1.1-mpibull2_1.3.9_Bull.4
gmp (gmp_sci XBAS)	gmp_sci-4.3.0-Bull.12
gsl	GSL-1.12-Bull.12
lapack(lapack_sci)	lapack_sci-3.2.1-Bull.12
MPFR	MPFR-2.4.1-Bull.12
netCDF	netCDF-4.0-mpibull2_1.3.9_Bull.12
OpenS	OpenS-1.0-Bull.5 OpenS_shelf-1.0-Bull.5
ParMetis	ParMETIS-3.1-mpibull2_1.3.9_12.Bull
PETSc	PETSc-2.3.3_p15-mpibull2_1.3.9_Bull.12
pgapack	pgapack-1.0.0.1_3-mpibull2_1.3.9_Bull.12
pHDF5_mpibull2	pHDF5-1.8.2-mpibull2_1.3.9_Bull.12
pNetCDF	pNetCDF-1.0.3-mpibull2_1.3.9_Bull.12
ScaLAPACK	ScaLAPACK-1.8.0-mpibull2_1.3.9_Bull.12
sciport	sciport-1.0-12.Bull
SciStudio_shelf	SciStudio-1.0-Bull.13 SciStudio_shelf-1.0-Bull.13
sHDF5	sHDF5-1.8.2-Bull.12
SuperLU_DIST	SuperLU_DIST-2.3-mpibull2_1.3.9_Bull.12
SuperLU_MT	SuperLU_MT-2.0-Bull.12
SuperLU_SEQ-3	SuperLU_SEQ-3.1-Bull.12
Valgrind	valgrind_opens-3.3.1-Bull.4

Table 1-1. Scientific Studio Libraries

Chapter 2. What's New in bullx cluster suite XR 5v3.1U1

2.1 Hardware

2.1.1 bullx blade system

bullx blade system chassis
The **bullx** chassis contains compute blades, a first level interconnect, a management unit and all components necessary to power and cool the blades, the interconnect, and the management unit.

It can host up to 18 compute blades in 7U. The interconnect switch integrated in the chassis uses **InfiniBand** technology. The management unit **CMM** (Chassis Management Module) controls the drawer and communicates with the cluster management infrastructure through **Ethernet** ports.

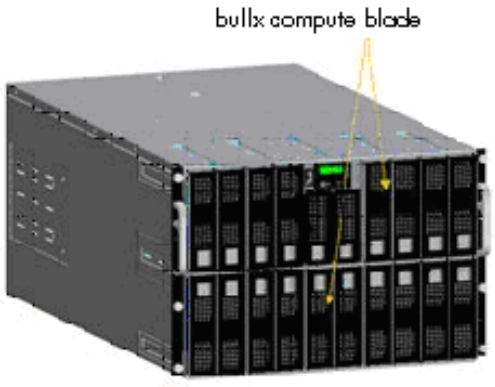

Figure 2-1. bullx chassis system

bullx compute blade
Each compute blade contains:

- 2 quad core **Intel® Xeon® 5500** series processors
- 12 memory DIMMs
- 1 HDD / SSD or diskless
- A built-in high performing interconnect per blade: **QDR InfiniBand**

A three level management hierarchy has been designed for the **bullx blade system**.

The first level of management is represented by the **Baseboard Management Controller** (**BMC**) embedded on each compute blade. The **BMC** of each compute blade is connected to the **Chassis Management Controller** (**CMC**) to manage installation, maintenance, monitoring, and power management of the corresponding compute blades.

The **CMC** ensures the second level of management. It deals essentially with the different components of the chassis: power supplies, fans, etc. The interconnect is also connected to the **CMC**.

The 24 port 1 Gb Ethernet switch of the **CMM** has 3 ports dedicated to external connections. This design allows a **bullx blade chassis** to be integrated into a large cluster with the third level of management ensured via Service Nodes.

2.1.2 Broadcom Switches

Support for **Broadcom Ethernet** switches for **bullx blade systems**.

2.1.3 NovaScale R422 E2 machines

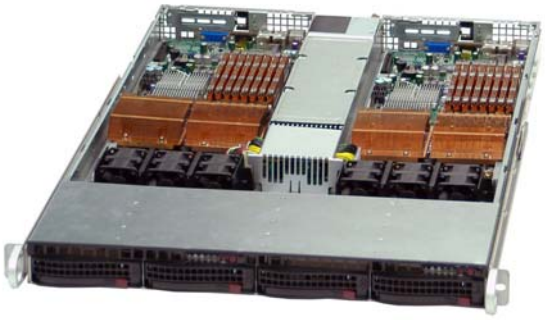

Figure 2-2. NovaScale R422 E2 machine

Bull NovaScale R422 E2 2 x 2-socket rack servers, based on **Intel® Xeon®** processors, are ideally suited for use as Compute Nodes.

- 2 x 2 quad core Intel® Xeon® processors (5500 series, up to 2.93 GHz)
- **Intel QuickPath** Interconnect (QPI) technology – providing point-to-point high-speed links to distributed shared memory
- Up to 2 x 96 GB DDR3 1066 MHz memory
- Double data rate and lower power consumption of DDR3 memory
- High-end connectivity, with 2 x PCI-Express Gen 2 (16 x) slots
- Embedded InfiniBand DDR or QDR adapters
- Optimal internal storage capacity with up to 2 x 2 SATA2 disks (up to 2 x 1000 GB per node)
- Scalable remote management features, with IPMI2.

2.1.4 NovaScale R425 E1 and R425 E2 machines

Bull **NovaScale R425** servers are double socket, dual or quad core machines and include a powerful PSU to support internal **NVIDIA Tesla C1060** accelerator cards. This accelerated Compute Node is able to manage 2 **C1060** boards (directly attached implementation).

Figure 2-3. NovaScale R425 E1 machine

Figure 2-4. NovaScale R425 E2 machine

2.1.5 Storage Devices

StoreWay Optima1500 Storage systems

Developed for Fibre Channel standards for server connections and Serial Attached SCSI (SAS) standards for disk connections, these systems can support high-performance disks and high-capacity disks (SATA) in the same subsystem. These systems include 2 x 4 Gb/s FC host ports per controller (optionally 6 per controller). They offer up to 64 TB SAS or 144 TB SATAII capacity in 2U 12 disk drawers.

DDN S2A 9550 Storage systems

The S2A 9550 Storage Appliance is specifically designed for high-performance, high-capacity network storage applications. Delivering up to 3 GB/s large file performance from a single appliance and scaling to 960 TBs in a single storage system.

2.2 Software

From this distribution onwards the **BAS5 for Xeon** software suite has been renamed as **bullx cluster suite (bullx CS)**. Existing **BAS5 for Xeon** distributions can be upgraded to **bullx cluster suite XR 5v3.1U1. bullx cluster suite** is used for the management of all the nodes of a Bull extreme computing cluster.

See Chapter 2 in the **bullx cluster suite XR 5v3.1U1** *Installation and Configuration Guide* for full details of the upgrade procedure for **BAS5 for Xeon V1.1**, **V1.2** and **V3.1** clusters.

2.2.1 bullx cluster suite XR

bullx cluster suite XR covers two separate distributions:

- **bullx cluster suite XR SN** based on the **Red Hat Enterprise Linux 5.3** operating system.
- **bullx cluster suite XR CN** based on the **bullx cluster suite XR CN-OS** operating system. This distribution is compatible with **Red Hat Enterprise Linux 5.3** and is only available for the Compute Nodes.

The table below shows the two **bullx cluster suite XR** installation options for the cluster nodes.

Cluster Type	Service Node Distribution	Compute Node Distribution
1	**bullx cluster suite XR SN**	**bullx cluster suite XR SN**
2	**bullx cluster suite XR SN**	**bullx cluster suite XR CN**

Table 2-1. **bullx cluster suite XR** installation types

2.2.2 InfiniBand

- Introduction of **OFED 1.4.1** for **bullx cluster suite** clusters.
- Implementation of the **OpenSM** service to provide subnet management for networks with non-managed InfiniBand switches.

See The *InfiniBand Guide* for more information.

2.2.3 Bull Scientific Studio

Bull Scientific Studio is included in the **bullx cluster suite** delivery, and includes a range of Open Source libraries that can be used to facilitate the development and execution of a wide range of applications.

See The *Table 1.1* in this document for details of the Bull **Scientific Studio** libraries included in this release.

2.2.4 SLURM 2.0.1

bullx cluster suite XR 5v3.1U1 supports **SLURM** version **2.0.1**

See The **bullx cluster suite** *Administrator's Guide* and the *User's Guide* for more information on **SLURM**.

With this **BAS** release, the base SLURM version has changed from **SLURM 1.3.10** to **SLURM 2.0.1.**

2.2.5 nsfirm command

The **nsfirm** command is used for various maintenance operations, such as obtaining the **BIOS** or **BMC** version, upgrading the firmware, flashing the BIOS, etc.

See **bullx cluster suite** *Maintenance Guide* for more information.

2.2.6 Hardware Discovery tools

The new **initClusterDB, swtDiscover** and **nodeDiscover** commands can be used to discover and to add cluster hardware to the Cluster Database. Some of this hardware, including new **Ethernet** switches and hardware management cards, will also be configured by these tools. These tools may be used when installing the Bull distribution for the first time, or when adding hardware to extend a cluster.

See *Chapter 2* in the **bullx cluster suite** *Maintenance Guide* for more information.

2.2.7 PBS Professional GridWorks Analytics Add-on

The **GridWorks Analytics** feature uses a parser to collect information from the PBS Professional server node (normally this is the cluster Management Node). The Application Server installed on a Login Node shows the information stored in the analytics database, either graphically or in the form of tables. These reports can be used to analyse and improve the performance of **PBS Professional** on the cluster, and to troubleshoot configuration problems.

2.2.8 stordepha heuristic function

Quorum Disk for **Cluster Suite** is not supported for this release. It is recommended to use the **heuristic** functionality when configuring with the **stordepha** tool (option **-H**).

2.2.9 Documentation

The introduction of 2 new cross-system manuals, the *InfiniBand Guide* and the *LDAP Authentication Guide*.

See	*Chapter 4* for the full list of manuals delivered with the **bullx cluster suite XR 5v3.1U1** distribution.

Chapter 3. bullx cluster suite XR 5v3.1U1 Features

3.1 Nodes typology

bullx cluster suite XR 5v3.1U1 nodes are defined as follows:

3.1.1 Service Nodes

1. Management Service Node –>
   - Supports the Cluster Management Utility stack
   - Operates using RHEL5.3
2. Login and NFS I/O Service Nodes –>
   - Provides both NFS server and development environment
   - According to the cluster topology implemented, the NFS server or Login function is not configured, giving a pure NFS server node or a pure Login node as the case may be
   - When the Login function is activated, the Login node becomes the user's access point to the cluster
   - Operates using RHEL5.3
3. Lustre I/O Service Nodes –>
   - I/O nodes that support the OSS and MDS server functions of the Lustre file system
   - Strictly dedicated to the **Lustre** file system
   - Connected to the cluster storage arrays which may be FDA, DDN or EMC
   - Operates using RHEL5.3 with the Bull modified kernel

3.1.2 Compute Nodes

1. Compute Nodes –>
   - Used for parallel computing
   - Two types of Compute Nodes are possible. The COMPUTE node is performance oriented and provides a minimum environment. The extended COMPUTEX node is more complete, and provides the necessary environment for running most ISV applications and also those which need the Intel Cluster Ready environment
   - Operates using the **RHEL5.3** operating system or the **bullx cluster suite XR CN-OS** operating system.

For small clusters (up to 24 nodes), Cluster Management, NFS and Login functions can be concentrated on a single node.

3.2 Linux kernel and distribution

bullx cluster suite XR 5v3.1U1 operates using a standard **RHEL5.3** Linux Red Hat distribution.

- Linux kernel version 2.6.18-128

The Bull kernel is at the level of the **Red Hat** 2.6.18-128 standard kernel. It includes patches for the **Lustre** file system, and for Bull **HPC Toolkit** performance analysis tools.

The Bull kernel is fully compatible with the **RHEL5.3** distribution **BUT is Bull, and not Red Hat,** maintained. When the **XLUSTRE** and/or the **XTOOLKIT** products are required on a cluster, the Bull modified kernel is installed **on all the cluster nodes.**

Note If **Intel® VTune Performance Analyzer for Linux** is to be installed on the cluster, the HPC Toolkit **(XTOOLKIT)** product must be installed - see Chapter 3 in the *Installation and Configuration Guide*.

3.3 Cluster Management

All the cluster nodes are controlled and monitored from the Management Node. Cluster Management uses the cluster administration tools, which are centralized on the Management Node:

- **ClusterDB** contains the data that is required for the cluster management tools.
- **pdsh**, a distributed shell, is used to run commands in parallel on all the nodes, or a group of nodes, of the cluster.
- **KSIS** is used to produce and deploy node images.
- **Bull System Manager - HPC Edition** is used to monitor the cluster and its activity.
- **Syslog-ng** is used to centralize the **/var/log/**messages and the **/var/log/syslog** files, for each node of the cluster on the Management Node. This allows global events to be monitored.
- **Conman** enables access to all the consoles of the cluster nodes at the same time.

3.4 High-speed interconnect

InfiniBand

The **InfinBand** network support relies on the **OFED 1.4.1 OpenFabrics** software stack.

Ethernet

High Speed Gigabit Ethernet.

3.5 Storage

Storage appliance monitoring is fully integrated within **Bull System Manager - HPC Edition**.

Storage Management includes configuration tools for all types of storage systems that are supported:

- StoreWay **Optima1250** (I/O NFS only – no HA)
- StoreWay **Optima1500**
- **EMC/Clariion AX4-5** (I/O NFS only – no HA)
- Bull **FDA 2500 & 2900**
- **DataDirect Networks 9550**
- **EMC/Clariion CX3-40F, CX300, CX4-120, CX4-480**

These last two storage systems are oriented to support **Lustre** and are High Availability capable.

3.6 Lustre

Introduction of **Lustre 1.6.7**, shared file system from **Sun CFS**:

- **InfiniBand** (OFED 1.4.1) and **Gigabit Ethernet** interconnect support
- **FDA, EMC** and **DDN** storage systems fibre channel connection support
- The automatic configuration and deployment of **Lustre** is tightly coupled with the storage appliance configuration and deployment using storage deployment models
- Centralized management and monitoring toolset from Bull
- High-Availability for **OSS** and **MDS** nodes

3.7 Parallel computing

Parallel computing is ensured by MPI libraries:

- The optimized **MPIBull2-1.3.9** library provides **MPI1** and **MPI2** levels for both **GigaBit Ethernet** and **InfiniBand** interconnects thanks to its flexible architecture.

The **MPI** environment switching tool allows users to run the application in the right environment when there are different **MPI** libraries on the cluster.

3.8 SLURM / Batch Scheduler

Two possibilities are available for cluster job resource management:

- The **SLURM** version **2.0.1.** Resource Manager for both **InfiniBand** and **GigaBit Ethernet** interconnects.
- The **PBS Pro 10.0** Batch Scheduler from Altair for both **InfiniBand** and **GigaBit Ethernet** interconnects which interface directly with the MPI libraries.

Important

PBS Professional and **SLURM** are exclusive and cannot both be installed on the same cluster.

The installation and configuration of **SLURM version 2.0.1** using the **configurator.html** tool and the **setup.sh** script is described in the **Bullx cluster suite XR 5v3.1U1** *Installation and Configuration Guide*, ref **86 A2 19FA 02.**

3.9 Development environment

The development environment provides both standard **gcc,** and optimized **Intel C/C++** and **Fortran** compilers. The **Intel C/C++** compiler supports **OpenMP** primitives.

Bullx cluster suite XR 5v3.1U1 has been validated with **Intel C/C++** and **Fortran** 11.0.069 compilers. The compatibility with later **11.x** compiler releases is assured provided that the Bull **intelruntime-11.0.069 RPM** is not installed, and the compiler environment is made available on all the Compute Nodes.

Some scientific libraries, for example the **Electric Fence** debugger, are provided with the standard **RHEL5** distribution.

MPIBull2 optimized libraries

Bull delivers enhanced, recompiled **MPIBull2** optimized versions of Open Source libraries within Bull **Scientific Studio**.

See Table 1.2 for details of the **Scientific Studio** libraries delivered with **Bullx cluster suite XR 5v3.1U1.**

NVIDIA Mathematical and Scientific Libraries

NVIDIA CUDA™ Toolkit and **Software Development Kit** are installed automatically on the **LOGIN**, **COMPUTE** and **COMPUTEX** reference nodes for clusters which include **Tesla** graphic accelerators, so that the **NVIDIA** compilers and the **NVIDIA** mathematical and scientific libraries are in place for the application.

3.10 Performance tools

HPC Toolkit is an Open Source suite of multi-platform tools for profile-based performance analysis of applications, and is used to:

1. Collect raw profile information
2. Convert various types of profiling information into platform independent XML formats
3. Synthesize browsable representations that correlate performance metrics gathered from multiple sources with program source code.

3.11 High Availability

High Availability functions using HA **Cluster Suite 5** software.

Important for Bullx cluster suite XR 5v3.1U1

- Lustre I/O High Availability is fully supported for this release.
- NFS I/O High Availability is supported in this release for NFS3 in a reduced active/passive mode only.
- Highly Available I/O NFS nodes must be deployed using a dedicated KSIS I/O node image that strictly excludes Lustre.
- PBS Professional High Availability is supported for this release.
- LSF High Availability is supported for this release.
- SLURM High Availability is supported for this release.
- Management Node High Availability is supported for this release.

Chapter 4. bullx cluster suite XR 5V3.1U1 Software and Documentation

4.1 bullx cluster suite XR 5V3.1U1 Delivery contents

4.1.1 Linux XHPC

BAS5 for Xeon V3.1 - XHPC Bull Linux HPC for Xeon	DVD ref : 76742478-101
bullx cluster suite XR 5v3.1 U1	DVD ref : 76742561-101
bullx cluster suite XR SN-OS Errata for RHEL5.x (EM64T)	DVD ref : 76742562-101
bullx cluster suite XR 5V3.1U1 documentation	CD ref: 86 A2 12FB 02

4.1.2 InfiniBand Software - optional

BAS5 for Xeon V3.1 - XIB InfiniBand Software	CD ref : 76742479-001

4.1.3 BAS5 for Xeon V3.1 - XLustre V1 Lustre Software - optional

BAS5 for Xeon V3.1 - XLustre Lustre Software	CD ref : 76742480-001

4.1.4 bullx cluster suite XR CN-OS Operating System - optional

bullx cluster suite XR CN-OS for RHEL5.x (EM64T)	DVD ref : 76742563-101
bullx cluster suite XR CN-OS Errata for RHEL 5.x (EM64T)	DVD ref : 76742564-101

4.1.5 Bull HPC for Xeon - PBS Pro V10.0 - optional

Bull HPC for Xeon PBS-Pro v10.0	CD ref : 76742477-001

4.1.6 Bull HPC for Xeon - LSF V7.04. - optional

Bull HPC for Xeon V3.1 LSF V7.04	CD ref : 76742481-001

This CD includes the Bull **BAS5 for Xeon** *LSF Installation and Configuration Guide* reference **86 A2 39FB 01**.

4.1.7 bullx cluster suite XR 5V3.1U1Documentation

bullx cluster suite XR 5V3.1U1 Software Documentation	Reference
bullx cluster suite *User's Guide*	86 A2 22FA rev 02
bullx cluster suite *Administrator's Guide*	86 A2 20FA rev 02
bullx cluster suite XR 5V3.1U1 *Installation and Configuration Guide*	86 A2 19FA rev 02
bullx cluster suite *Maintenance Guide*	86 A2 24FA rev 02
bullx cluster suite *Application Tuning Guide*	86 A2 23FA rev 02
bullx cluster suite *High Availability Guide*	86 A2 25FA rev 02
Cross-System Documentation	Reference
InfiniBand Guide	86 A2 42FD rev 00
LDAP Authentication Guide	86 A2 41FD rev 00

These guides are delivered on the **bullx cluster suite XR 5V3.1U1 Documentation CD (ref : 86 A2 12FB 02)**.

4.2 Other Software - not included in the BAS5 for Xeon V3.1 Delivery

4.2.1 Bull Extension Pack

The **Bull Extension Pack CD** for **bullx cluster suite XR 5V3.1U** machines is the one below. This is part of the *Red Hat Enterprise Linux 5.3 EM64T Media and Documentation* delivery.

Bull Extension Pack for NovaScale Universal Rack-Optimized & Tower Series with RHEL5.3	CD ref : 76742499-001

This CD-ROM contains the RPMs required for **Bull System Backup Restore**, and for the **LSI MegaRAID 8408E** Adapter installed on **NovaScale R440** and **R460** machines.

See The **BAS5 for Xeon** *Maintenance Guide* for more details on **Bull System Backup Restore**.

4.2.2 Proprietary Software

Note The **Intel Compilers** are not delivered with **BAS5 for Xeon V3.1**. But this release has been built to work with Intel compilers Version **11.0.069**, and has been validated with these versions, and should also be compatible with later versions.

Chapter 5. Licensing

Bull is committed to be in conformance with Free Software Foundation, Inc. and recommendations from other standardization organizations.

Most of the software is Public or **GPL,** and consequently the source files may be distributed. The license terms are included on the distribution CD in the **GPL** directory.

Some products have their own licenses and others are proprietary; below is the list of proprietary software:

Proprietary Bull

Bull System Manager

MPI_Bull libraries

MPI_Analyser

Storageadmin

Proprietary Emulex

Lptools

Proprietary Altair

PBS_Pro

Proprietary Intel

Intel C, C++, Fortran
Intel Trace Tools
MKL library

Proprietary Allinea

DDT

Proprietary ETNUS

Totalview

Proprietary Platform

LSF

5.1 License keys

License name	License type	Link to	FLEXlm	Comment
Intel® C++ Compiler for Linux	Nodelock counted	License server Host-name & MAC address	Yes	License Server and license keys should be on the same system
	Nodelock uncounted	MAC address	No	
	Floating	License server Host-name & MAC address	Yes	License keys anywhere on the network Concurrent users limitation
Intel® Fortran Compiler for Linux	Nodelock	Same as Intel® C++	Yes	
	Nodelock uncounted	MAC address	No	
	Floating	Same as Intel® C++	Yes	Concurrent users limitation
Intel® Cluster MKL Math.Kernel Lib	Small, medium or large cluster		Yes	License control at product installation only
DDT (Allinea) Cluster license Package (Linux 64 bits)	Nodelock	DDT License Server Hostname or IP address License server serial number	No	Nb cpus/cores & Nb users
	Floating	DDT License Server Hostname or IP address, port and LS MAC address	No	Nb cpus/cores & Nb users
ETNUS Totalview		License server Host-name & MAC address	Yes	License server can be local or remote. Max concurrent TotalView users Max simultaneous CPUs
Altair PBS Professional	Floating	License server Host-name & MAC address	Yes	License server can be local or remote. Max concurrent core users Max simultaneous cores
LSF	Floating	License server Host-name & MAC address	Yes	License server can be local or remote.

Note Host -ID is equivalent to the MAC address and is obtained by using the **lmhostid** command.

Chapter 6. bullx cluster suite XR 5v3.1U1 Software Installation

Read this chapter carefully first before carrying out an upgrade to **bullx cluster suite XR 5v3.1U1** from an existing **BAS5 for Xeon V1.1**, **V1.2** or **V3.1** cluster, as described in *Chapter 2* in the BAS5 for Xeon *Installation and Configuration Guide* OR before installing **bullx cluster suite XR 5v3.1U1** from scratch, as described in *Chapter 3* in the BAS5 for Xeon *Installation and Configuration Guide*.

6.1 installnfs - upgrade mode

The **upgrade** mode provided by the **installnfs** script (See section 2.2.5.2 in the *Installation and Configuration Guide)* must ONLY be used to upgrade **BAS5 for Xeon V1.1** and **V1.2** clusters to **bullx cluster suite XR 5v3.1U1**.

Do not use the installnfs upgrade mode if bullx cluster suite XR 5v3.1U1 has already been installed.

6.2 Console Redirection for NovaScale R423E2T2 and R425 platforms

On **NovaScale R423E2T2** and **R425** platforms, following the installation of **bullx cluster suite XR 5v3.1U1** on the reference node change the **console=ttyS1** setting in the **/boot/grub/menu.lst** file to **console=ttyS2**, as shown in the example below. This should be done before the nodes are deployed:

BEFORE:

```
kernel /boot/vmlinuz-2.6.18-128.1.6.el5.Bull.1 ro root=LABEL=/
console=tty0 console=ttyS1,115200 nmi_watchdog=0
kernel /boot/vmlinuz-2.6.18-128.el5 ro root=LABEL=/   console=tty0
console=ttyS1,115200 nmi_watchdog=0
```

AFTER:

```
kernel /boot/vmlinuz-2.6.18-128.1.6.el5.Bull.1 ro root=LABEL=/
console=tty0 console=ttyS2,115200 nmi_watchdog=0
kernel /boot/vmlinuz-2.6.18-128.el5 ro root=LABEL=/   console=tty0
console=ttyS2,115200 nmi_watchdog=0
```

The same change must be made in the **/etc/securetty** and **/etc/inittab** files. Run the **init q** command once the change has been made.

Alternatively, the **console=ttyS1** setting must be changed in these 3 files on each node individually, followed by a reboot.

6.3 Time Zone Settings for the Installation

By default all nodes installed using **NFS** are configured with the **US** keyboard setting (**KEYTABLE** option in **/etc/sysconfig/keyboard**) and the time zone configured for the Management Node.

Do not use the following time zones when installing the Management Node:
Brazil/*, CET, CST6CDT, Chile/*, Cuba, EET, EST, EST5EDT, Egypt, Eire, Etc/*, Factory, GB, GB-Eire, GMT, GMT+0, GMT-0, GMT0, Greenwich, HST, Hongkong, Iceland, Iran, Israel, Jamaica, Kwajalein, Libya, MET, MST, MST7MDT, Mexico/*, NZ, NZ-CHAT, Navajo, PRC, PST8PDT, Poland, Portugal, ROC, ROK, Singapore, Turkey, UCT, UTC, Universal, W-SU, WET, Zulu

See https://bugzilla.redhat.com/show_bug.cgi?id=481617 for more information on this problem.

See Section 3.1.6 in the **bullx cluster suite XR 5v3.1U1** *Installation and Configuration Guide* for more information on setting the Time Zone for the Management Node

Once the installation process has been completed for all the nodes of the cluster, it is then possible to modify the time zone settings if you wish to use one of the forbidden time zones above. This is done as follows:

1. On the Management Node run the command:

```
system-config-date
```

2. On the other nodes use the **pdsh** command to modify the **/etc/sysconfig/clock** file.

6.4 Important Notes regarding the installation

6.4.1 Cluster DB

When a pre-installed cluster is delivered, be sure to save the **Cluster DB**. This is specific to your configuration.

6.4.2 Ksis deployment

To use Ksis to deploy your nodes it is mandatory to put the / file system on sda.

Disk Space

Ksis uses all the disk space available on the deployed nodes for the swap partition. So if you wish the swap partition to be the same size on both the Reference Node and on the deployed nodes, create dummy partitions on the Reference Node before deploying it.

6.4.3 BIOS update

Do not update your **BIOS** unless you are sure that it is necessary. Contact Bull Technical Support for more information.

6.4.4 SSD Devices and small capacity disks

If the Disk Device is less than 73 GBs in size (66 GBs of free space) then manual partitioning must be used for the installation, as described in the third point of *Section 3.4.4* of the bullx cluster suite *Installation and Configuration Guide* (ref 86 A2 19FA 02).

6.4.5 Partitioning Problems when installing RHEL5.3

For example, on **NovaScale R440** and **R460** platforms, re-installing a **RHEL5.3** system on a disk previously installed with **LVM** partitioning may lead to the following errors:

1. Cannot boot with grub error 15
2. Error message: Specified nonexistent disk xxx ….

To overcome this problem, enter the command:

```
cat /proc/partitions
```

Output example

```
major minor  #blocks  name
   8     0   71687402 sda
   8     1     104391 sda1
   8     2   70204050 sda2
 253     0   68157440 dm-0
```

```
253     1    2031616 dm-1
```

3. Check your hardware
   If you have partitions names starting with DM, then you may have this problem. This occurs when **RAID** hardware is validated.

   Check that the jumper that validates RAID hardware management on the motherboard of the machine is correctly set to the **no RAID** position, and then manually reinstall the system from the **RHEL5.3** DVD.

Please, refer to the **NovaScale R440** and **R460** hardware documentation for information about setting **RAID** hardware on local disks.

4. Installation option.
   After checking your hardware, use the **nodmraid** option when installing RHEL.

6.4.6 OpenSM Subnet Manager

The **OpenSM** subnet manager RPMs are included in the **bullx cluster suite XR 5v3.1U1** delivery, and following the installation are to be found in the **UPGRADES** directory. These must be installed manually using the command below:

```
yum install opensm*
```

The following packages will be installed for **bullx cluster suite XR 5v3.1U1.**

opensm-3.1.11-0.1.ofed1.3.2.x86_64.rpm
opensm-devel-3.1.11-0.1.ofed1.3.2.x86_64.rpm
opensm-libs-3.1.11-0.1.ofed1.3.2.x86_64.rpm
opensm-static-3.1.11-0.1.ofed1.3.2.x86_64.rpm

Once the packages have been installed use the command below to launch the **opensmd** service:

```
service opensmd start
```

Check that the **InfiniBand** network has been detected by the **OpenSM** Subnet Manager by using the command:

```
ibnetdiscover
```

See The *InfiniBand Guide* for more information on the configuration of the **OpenSM** subnet manager and using the **ibnetdiscover** command.

6.4.7 LDAP Authentication Protocol

Starting the ldap-auth service automatically

After **LDAP** has been installed and configured the **ldap-auth** service must be configured, using the command below, so that it starts automatically on the Management Node:

```
# chkconfig --levels 35 ldap-auth on
```

Run the command below to check the service state:

```
# chkconfig --list ldap-auth
```

If the service is not active then start it with the command below:

```
# service ldap-auth start
```

The service will restart automatically when the init is at run level 3 or 5.

Name Service Consistency

After the installation and configuration of the **LDAP** or **NIS** authentication protocols, it is recommended to restart the **nscd** service on all Service Nodes to avoid consistency problems in the Name Service (NS) cache content. To do this, run the following command:

```
# pdsh -w node[a-z] service nscd restart
```

6.4.8 Bonus RPM Installation

The BONUS packages must be installed manually. Search for the latest version of these RPMs in the sub-directories of the **/release** directory on the Management Node and then install them on the node by using the command:

```
yum localinstall xx* xx*
```

6.4.9 Ethernet 10 Gigabit cards

WARNING
These cards are supported for enterprise backbone network connections only. They are not supported for administration or interconnect networks.

If you use **Myrinet 10** Gigabit cards, the **myri10ge-linux-<version> driver** for standard **RHEL5.3** and for **Bull** kernels are delivered on the **BAS5 for Xeon V3.1 XHPC** DVD. These are installed as follows:

6.4.9.1 RHEL5.3 kernel

RPM Location

XHPC/modules-rhel/myri10ge-linux-<version>.el5.x86_64.rpm

Installation command

```
yum install myri10ge-linux --enablerepo=xhpc-modules-rhel
```

6.4.9.2 Bull kernel

RPM Location

XHPC/modules-bull/myri10ge-linux-<version>.el5.Bull.1.x86_64.rpm

Installation command

```
yum install myri10ge-linux --enablerepo=xhpc-modules-bull
```

6.4.10 Intel Compilers and Runtime Libraries

See Chapter 7 in the **bullx cluster suite XR 5v3.1U1** *Installation and Configuration Guide*

6.4.11 Bull System Backup Restore

The installation warning, below, for the **Bull System Backup Restore** rpm does not have functional consequences.

```
==================================================
WARNING: it seems that webmin isn't installed on that system.If you
install it later, don't forget to add mkcdrec to the list of modules in
/etc/webmin/webmin.acl to make mkcdrec's webmin module available
==================================================
```

Webmin is not mandatory for **BSBR**.

6.4.12 NovaScale R421/R422 DHCP reboot

To prevent difficulties when rebooting **NovaScale R421/R422** machines via **DHCP**:

a. Stop the node using the command below:

```
ipmitool -U <bmcuser> -P <bmcpasswd> -I lan -H <bmcip> chassis power off
```

b. Reset the **BMC** from cold:

```
ipmitool -U <bmcuser> -P <bmcpasswd> -I lan -H <bmcip> bmc reset cold
```

This can take a while (up to 1 minute)

c. Restart the machine when the **BMC** is available:

```
ipmitool -U <bmcuser> -P <bmcpasswd> -I lan -H <bmcip> chassis power on
```

6.5 Upgrading to bullx cluster suite XR 5v3.1U1

6.5.1 Updating from BAS5 for Xeon V3.1

NVIDIA CUDA™ Software Development Kit

To avoid the error message;

```
cutil.cpp:47:27: error: builtin_types.h: No such file or directory
```

the **NIVIDIA CUDA** Software Development Kit must be installed as described below:

1. Remove the **BAS5 for Xeon V3.1 CUDA Toolkit** RPM:

```
# rpm -e cuda_sdk
```

2. Install the **bullx cluster suite XR 5v3.1U1 CUDA Toolkit** RPM as follows:

   a. Add the following line to the **/etc/yum.repos.d/xhpc-common.repo** file.

```
exclude=cuda_sdk-2.10-Bull.2
```

   b. Run the command below to install the **cuda-sdk** package

```
# yum install cuda_sdk
```

Note Ignore the warning messages which appear.

6.5.2 Upgrading from BAS5 for Xeon V1.1 and V1.2

PBS Professional clusters

Verify that the PBS license file is mirrored (in the same directory on both servers).

NFS High Availability clusters

Before **bullx cluster suite XR 5v3.1U1** is installed, save the existing **BAS5 for Xeon V1.1** or **V1.2 /etc/storageadmin/haionfs.conf** file on the Management Node on an external back-up device.

Following the upgrade of the Management Node software to **bullx cluster suite XR 5v3.1U1**, copy the saved **haionfs.conf** file to **/etc/storageadmin/ha/hafsnfs.conf** file on the Management Node.

6.5.3 Upgrading from BAS5 for Xeon V1.1

gmond.conf

The **gmond.conf** file is saved with your cluster details in it after upgrading from **BAS5 for Xeon V1.1** to **bullx cluster suite XR 5v3.1U1**. Check that the file in the **/usr/share/doc/..../template** folder does not present new metrics for the cluster. If it does, customise the template again as described in Chapter 3 in the **BAS5 for Xeon** *Installation and Configuration Guide* for both the Management Node and Reference Nodes.

6.6 Storage

6.6.1 Software for StoreWay Optima1250 storage systems

In order that the **xyr_admin** commands work correctly the software versions for **StoreWay Optima1250** storage systems should be:

Firmware: 3.3.27 minimum
StoreWay Master: 3.07.02 minimum

6.6.2 EMC management

See Chapter 4 in the *Installation and Configuration Guide* for more information.

6.7 hpcsnap

If you need to contact Bull HPC Support please use the **hpcsnap** tool to record the details of your **bullx cluster suite** installation.

See Section 3.4 in the **bullx cluster suite** *Maintenance Guide* for more information on the **hpcsnap** tool.

Chapter 7. High Availability

7.1 Configuring NTP on Nodes for clusters with Management Node High Availability

The configuration of **NTP** for all cluster nodes has to be modified when Management Node High Availability is implemented. The **NTP** service on the nodes has to include the IP addresses of both Management Nodes.

1. Disable the post deployment configuration of **NTP**. In the example which follows the Primary Management Node has an IP address of 10.0.0.1 and the Secondary Management Node has an IP address of 10.0.0.2. The command below disables the **ksis** post deployment configuration of the **NTP** service and has to be launched on the Primary Management Node:

```
ksis postconfig disable CONF_20_NTP
ksis postconfig buildconf
rsync -a --del /etc/systemimager/ root@10.0.0.2:/etc/systemimager
```

2. Deploy a node reference image to one node of each type e.g. Login, I/O, COMPUTE(X).

```
ksis deploy < Node_image> node
```

3. Configure **NTP** on each deployed node type by adding two server lines for the Management Nodes in the **/etc/ntp.conf** file, as shown in the example below:

```
server 10.0.0.1
server 10.0.0.2
driftfile driftfile /var/lib/ntp/drift
```

4. Make an image of each node type:

```
ksis create <New_ Node_image> <New_ Node>
```

5. Deploy the new image to all the nodes of that type in the cluster.

```
ksis deploy <New_Node_image> node[1-x]
```

6. Run **postconfig.**

7.2 Starting Nagios on the Secondary Node for Mixed Management Node Installs

For clusters which include a Primary Management Node which has been upgraded from **BAS5 for Xeon V1.1** and **V1.2** to **bullx cluster suite XR 5v3.1U1** and a Secondary Management Node which is installed from 'scratch', **Nagios** will not start on the Secondary Management Node as the **UIDs** are not the same.

Check the **UIDs** and **GUIDs** are the same on both Management Nodes, and if they are not the same, change the **UID** and **GUID** on the Primary Node to match those in place on the Secondary Node.

7.3 Cluster Suite status display

If you use the Cluster Suite 5 **clustat** command to display the status of the Cluster Suite defined services for that node, be aware that the text display mode of **clustat** truncates service names to 12 characters.

For example, for two services respectively named **lustre_xena10** and **lustre_xena11**, the **clustat** command will display:

```
service:lustre_xena1          disabled
service:lustre_xena1          disabled
```

To ensure that the services are displayed correctly, use the XML display mode for the **clustat** command: **clustat –x**, the services will be displayed as below:

```
name="service:lustre_xena10" last-owner="xena10"
name="service:lustre_xena11" last-owner="xena11"
```

7.4 PBS Professional High Availability

Following an upgrade from **BAS5 for Xeon v1.1** and **v1.2** ensure that the PBS license file is mirrored (in the same directory on both servers).

7.5 NovaScale R422 E2, R423 E2 and R423 E2T2 High Availability pairs

The **stordepha dump** option is not supported for these machines. Use the **reboot** option instead.

7.6 ClusterDB and ldap Mountpoint labels

Use the **HA_MGMT:cdb** label instead of the **HA_MGMT:clusterdb** for the **/var/lib/pgsl/data** mount point, and the **HA_MGMT:dblustre** label instead of **HA_MGMT:ldaplustre** for the **/var/lib/ldap** mount point.

Note This information replaces the labels indicated in *Table 3.1* and *Chapter 3* in the bullx cluster suite *High Availability Guide* for these mount points.

7.7 Problems relocating the HA_MGMT service with LDAP

Restart the **nscd** service on all Service Nodes to avoid relocation problems for the **HA_MGMT** service for clusters which use the **LDAP** authentication protocol. Run the following command to do this:

```
# pdsh -w node[a-z] service nscd restart
```

Chapter 8. Restrictions and Known Problems

8.1 X Windows display on the Management Node

Problem Description: Bad X Windows definition. The bottom line does not display correctly and appears cut following a reboot after the RHEL Server **Congratulations the installation is complete** screen appears on X11 systems.

Solution:

a. Hold down the **Ctrl Alt F2** keys to go to the shell prompt for console 2

b. Save the **xorg.conf** file by using the commands below:

```
cd /mnt/sysimage/etc/X11
cp -p xorg.conf xorg.conf.orig
```

c. Edit the **xorg.conf** file by using the command below:

```
vi /mnt/sysimage/etc/X11/xorg.conf
```

d. Go to the **Screen** section, subsection **Display** and after the `Depth 24 line` add the following line.

```
-------------------------------------------------------------------
  Modes   "1024x768" "832x624"
-------------------------------------------------------------------
```

e. Save the file and exit **vi**

f. Confirm that the modifications have been registered by running the command:

```
diff xorg.conf.orig xorg.conf
```

This will give output similar to that below:

```
-------------------------------------------------------------------
27a28
>               Modes   "1024x768" "832x624"
-------------------------------------------------------------------
```

g. Check the screen appearance is OK by holding down the **Ctrl Alt F6** keys

h. Click on the **Reboot** button

Note The screen resolution can be changed at any time by holding down **Ctrl Alt -** or **Ctrl Alt +** on the keyboard.

Do not run the **system-config-display** command.

8.2 Ethernet Management Network

Problem Description: Poor performance for the Management Network

Solution: 1. Delete the following lines from the **/etc/sysctl.conf** file on all **bullx cluster suite XR 5v3.1U1** Service and Compute Nodes using **pdsh**:

```
## MLX4_EN tuning parameters ##
net.ipv4.tcp_timestamps = 0
net.ipv4.tcp_sack = 0
net.core.netdev_max_backlog = 250000
net.core.rmem_max = 16777216
net.core.wmem_max = 16777216
net.core.rmem_default = 16777216
net.core.wmem_default = 16777216
net.core.optmem_max = 16777216
net.ipv4.tcp_mem = 16777216 16777216 16777216
net.ipv4.tcp_rmem = 4096 87380 16777216
net.ipv4.tcp_wmem = 4096 65536 16777216
## END MLX4_EN ##
```

2. Run the command below as root:

```
sysctl -p
/sbin/ib_ipoib_sysctl unload
mv /sbin/ib_ipoib_sysctl /sbin/ib_ipoib_sysctl.orig
```

3. Create the **sysctl_param** file with the default parameters, as below.

```
#cat >/tmp/sysctl_param <<EOF
net.ipv4.tcp_timestamps = 1
net.ipv4.tcp_sack = 1
net.core.netdev_max_backlog = 1000
net.core.rmem_max = 131071
net.core.wmem_max = 131071
net.core.rmem_default = 126976
net.core.wmem_default = 126976
net.core.optmem_max = 20480
net.ipv4.tcp_mem = 196608 262144 393216
net.ipv4.tcp_rmem = 4096 87380 4194304
net.ipv4.tcp_wmem = 4096 16384 4194304
EOF
```

4. Load the **sysctl** settings from the **sysctl_param** by running the command below:

```
#sysctl -p /tmp/sysctl_param
```

8.3 pdsh and nsctrl commands

The **pdsh** and **nsctrl** commands do not work correctly if there are any files in the directory where the command is launched with file names that include the name of a node, or a list of nodes, e.g. n[1-4].

8.4 HPC Toolkit

HPC Toolkit does not work in an MPIBull2 environment.

8.5 Lustre

8.5.1 OSTs out of space

Solution:

There are a few cases where the Lustre **OST**s might become unbalanced in this way. The primary reason is if a very large single file was created on the **OST** and is consuming a large amount of space.

Another possibility, when the **OST** appears as the first **OST** in the **index 0** file system, is that **lfs setstripe** is incorrectly setting the starting **OST** index to 0 for a large number of files. The starting **OST** index should be set to -1, instead. This will then start on the next available **OST**, round robin.

The **MDS** will avoid allocating objects on **OST**s with less than 0.1% of space available, but depending on the workload and average file size this figure may not be large enough. The workaround is to disable object allocation to these **OST**s manually by running on the **MDS** node. An example of device use follows:

```
mds# lctl

lctl> device_list
```

```
1 UP mdt MDT MDT_UUID 3
2 UP mds mds1 mds1_UUID 5
3 UP lov lov_mds1 604a2454-f571-4f0e-866a-2ab888b7c977 4
4 UP osc OSC_localhost_ost1_mds1 604a2454-f571-4f0e-866a-2ab888b7c977
5
5 UP osc OSC_localhost_ost2_mds1 604a2454-f571-4f0e-866a-2ab888b7c977
5
```

```
lctl> device %OSC_localhost_ost1_mds1 # could also use "device 4" here
```

```
Name OSC_localhost_ost1_mds1 is device 4
```

```
lctl> deactivate
```

```
(console) Apr 12 15:00:46 localhost kernel: Lustre:
20470:30034:(recover.c:322:ptlrpc_set_import_active()) setting import
ost1_UUID INACTIVE by administrator request
```

```
lctl> quit
```

The **MDS** will not allocate any new objects to **ost1**, but as this **OST** is not deactivated on the clients they can still carry out read/write/unlink operations on the files there. When the files are unlinked from this **OST**, or files are created on the new OSTs the space usage will be balanced and the OST can be reactivated:

```
mds# lctl --device %OSC_localhost_ost1_mds1 recover
```

```
(syslog) Lustre: OSC_localhost_ost1_mds1: Connection restored to
service ost1 using nid 0@lo
```

See the link below for a description and evolution of the problem on the **CFS** website: https://bugzilla.lustre.org/show_bug.cgi?id=12162

8.5.2 Compatibility with MPI/IO

Solution: To have both services, mount the **Lustre** File system using either the **localflock** option or the **flock** option, depending on the need for either local or global file locking capabilities.

8.5.3 Performance Loss

Problem Description: If the Lustre **stripe_size** parameter was set to a value lower than **1MB** with **4KB** pages, performance loss may result after updating Lustre to version **1.6.6**. This is due to the fact that for the previous **Lustre** version, the **stripe_size** parameter was automatically (and silently) adjusted regarding the page size: 1MB minimum on 4KB page size kernels.

Solution: The recommended solution is to comment the **stripe_size** line in the **Lustre** model file corresponding to your filesystem, and run the command **lustre_util update -f <path to .lmf file>**.

8.5.4 e2fsprogs Error Message

Problem Description: The following error message appears when you try to install or reinstall the **mgs** service using the **service mgs install** or **service mgs reinstall** commands.

```
WARNING: The e2fsprogs package currently installed on your system does
not support "uninit_bg" feature.
```

```
Please install the latest version of e2fsprogs from
http://downloads.lustre.org/public/tools/e2fsprogs/ to enable this feature.
Feature will not be enabled until e2fsprogs is updated and 'tune2fs -O
uninit_bg %{device}' is run.
```

Solution: Update the **PATH** variable declaration in the **/etc/init.d/mgs** script by adding the **/usr/lib/lustre** path, as shown below.

Before update:

```
# System path, if script launched by another daemon
PATH=/usr/sbin:/usr/bin:/sbin:/bin
```

After update:

```
# System path, if script launched by another daemon
PATH=/usr/lib/lustre:/usr/sbin:/usr/bin:/sbin:/bin
```

Following the script modification, the WARNING message should not appear.

8.5.5 Tuning phase appears to start early on the MDS Nodes

Problem Description: For large **Lustre** file system configurations with dozens of nodes or **OST**s, error messages, similar to that below, may appear stating that there were problems setting the file system tuning parameters following the launch of the **lustre_util start** command:

```
lustre_util start -f t7cell
```

```
Checking devices on node18...
Checking devices on node19...
Checking devices on node23...
.......
*  Error - No file matching /proc/fs/lustre/osc/*t7cell-OST001b*/max_rpcs_in_flight
 Error - No file matching /proc/fs/lustre/osc/*t7cell-OST000d*/max_rpcs_in_flight*
 Ok - "8" written in /proc/fs/lustre/osc/t7cell-OST00de-osc/max_rpcs_in_flight
 Ok - "8" written in /proc/fs/lustre/osc/t7cell-OST00d0-osc/max_rpcs_in_flight
.......
 Ok - "8" written in /proc/fs/lustre/osc/t7cell-OST0036-osc/max_rpcs_in_flight
 *Error - No file matching /proc/fs/lustre/osc/*t7cell-OST0028*/max_rpcs_in_flight*
.......
 Ok - "8" written in /proc/fs/lustre/osc/t7cell-OST001c-osc/max_rpcs_in_flight
 Ok - "8" written in /proc/fs/lustre/osc/t7cell-OST000e-osc/max_rpcs_in_flight
 Ok - "8" written in /proc/fs/lustre/osc/t7cell-OST0000-osc/max_rpcs_in_flight
---
FILESYSTEMS STATUS
+-----------+----------+--------+--------+----------------+----------+
| filesystem|  config  |running | number |   migration    |Available |
|           |  status  | status |of clts |                |  space   |
+-----------+----------+--------+--------+----------------+----------+
|t7cell     |installed |online  |0       |0 OSTs migrated |319.2 TB  |
+-----------+----------+--------+--------+----------------+----------+
---
Ok - Devices successfully checked for filesystem t7cell
Ok - I/O schedulers successfully set for filesystem t7cell
Ok - Filesystem t7cell successfully started
*Error - No file matching /proc/fs/lustre/osc/*t7cell-OST001b*/max_rpcs_in_flight (on
node13)
```

```
Error - No file matching /proc/fs/lustre/osc/*t7cell-OST000d*/max_rpcs_in_flight (on
node13)
Error - No file matching /proc/fs/lustre/osc/*t7cell-OST0028*/max_rpcs_in_flight (on
node13)*
.......
Error - Problems setting filesystem t7cell tuning parameters on node13
```

Solution: Run the **lustre_util tune_servers** command on your file system to complete its tuning process:

```
lustre_util tune_servers -f t7cell
```

8.6 MPIBull2

8.6.1 MPI_PUBLISH_NAME

Problem Description: This service does not work when the **MPD** launching system is combined with **PBS Professional**, nor when it is combined with **SLURM**. The **MPI_PUBLISH_NAME** service only works on clusters that combine the **MPD** launching system with the **LSF** Batch Manager.

Solution: No solution for this release.

8.6.2 Oshm device and One-Sided communications

Problem Description: If the **oshm** device is used with the **MPI_Accumulate** function together with **MPI_Win_lock** and **MPI_Win_unlock** functions may generate a dead lock.

Solution: No solution for this release.

8.6.3 Segmentation Faults with mlx4_1 devices and MPIBull2

Solution: Change the **Current value** setting for the **mpibull2_ibmr_number_hcas** parameter from **0** to **1**.

8.6.4 MPI and NFS

Problem Description: Bad synchronisation between nodes for I/O Operations

Solution: To use **MPI** and **NFS** together, the shared NFS directory must be mounted with the no attribute caching (**noac**) option added. To do this, edit the **/etc/fstab** file for the **NFS** directories on each client machine (in a multi-host **MPI** environment). If the performance for I/O Operations is impacted, see the solution below.

Note All the commands below must be carried out as root.

Run the command below on the **NFS** client machines:

```
grep nfs_noac /etc/fstab
```

Note **nfs_noac** is the name of the mount point.

The **fstab** entry for **/nfs_noac** should appear as below:

```
/nfs_noac /nfs_noac nfs bg,intr,noac 0 0
```

If the **noac** option is not present, add it and then remount the **NFS** directory on each machine using the commands below.

```
umount /nfs_noac
mount /nfs_noac
```

Problem Description: Poor performance for I/O Operations

Solution: To improve performance, export the **NFS** directory from the **NFS** server with the **async** option.

This is done by editing the **/etc/exports** file on the **NFS** server to include the **async** option, as below.

Example: The following is an example of an export entry that includes the **async** option for **/nfs_noac**:

```
grep nfs_noac /etc/exports
```

```
 /nfs_noac       *(rw,async)
```

If the **async** option is not present, add it and export the new value:

```
exportfs -a
```

8.7 SLURM

8.7.1 srun does not work

Problem Description: **srun** does not work and displays the following error message:

```
srun: error: Unable to create job step: Error generating job
credential
```

Solution: Check the **SLURM** version 2.0.1 configuration file. Either the **A.** OR the **B.** parameter setting combinations must appear in the file.

A.

```
AuthType=auth/munge
CryptoType=crypto/munge
```

Note The **munge** setting, used with the **munge** service, is recommended by **Bull** for security reasons. See *Chapter 3* in the bullx cluster suite XR 5v3.1U1*Installation and Configuration Guide* (Ref 86 A2 19FA 02) for details on installing and configuring **SLURM** and **Munge**.

B.

```
AuthType=auth/none
CryptoType=crypto/openssl
```

8.7.2 SLURM Man Pages

Problem Description: Incorrect details in the **srun** man page:

1. The **--comment** option is listed twice. Ignore the text below:

```
 --comment=<string> An arbitrary comment.
```

2. The **contiguous** option appears twice, the first description, shown below, is correct.

```
--contiguous
If set, then the allocated nodes must form a contiguous set. Not
honored with the topology/tree or topology/3d_torus plugins, both of
which can modify the node ordering. Not honored for a job step's
allocation.
```

3. The **parallel run** options listed are incorrect. The corrected options, from **-c** to **-p,** are shown below:

```
Usage: srun [OPTIONS...] executable [args...]
 -c, --cpus-per-task=ncpus number of cpus required per task
     --checkpoint=time     job step checkpoint interval
     --checkpoint-dir=dir  directory to store job step checkpoint image
files
     --comment=name        arbitrary comment
-d, --slurmd-debug=level   slurmd debug level
-D, --chdir=path           change remote current working directory
-e, --error=err            location of stderr redirection
    --epilog=program       run "program" after launching job step
-E, --preserve-env         env vars for node and task counts override
command-line flags
    --get-user-env         used by Moab.  See srun man page.
-H, --hold                 submit job in held state
-i, --input=in             location of stdin redirection
-I, --immediate[=secs]     exit if resources not available in "secs"
    --jobid=id             run under already allocated job
-J, --job-name=jobname     name of job
-k, --no-kill              do not kill job on node failure
-K, --kill-on-bad-exit     kill the job if any task terminates with a
 non-zero exit code
-l, --label                prepend task number to lines of stdout/err
-L, --licenses=names       required license, comma separated
-m, --distribution=type    distribution method for processes to nodes
(type = block|cyclic|arbitrary)
    --mail-type=type       notify on state change: BEGIN, END, FAIL or
ALL
    --mail-user=user       who to send email notification for job state
changes
    --mpi=type             type of MPI being used
    --multi-prog           if set the program name specified is the
configuration specification for multiple programs
-n, --ntasks=ntasks        number of tasks to run
    --nice[=value]          decrease secheduling priority by value
   --ntasks-per-node=n     number of tasks to invoke on each node
-N, --nodes=N              number of nodes on which to run (N = min[-
max])
-o, --output=out           location of stdout redirection
-O, --overcommit           overcommit resources
-p, --partition=partition  partition requested
    --prolog=program       run "program" before launching job step
    --propagate[=rlimits]  propagate all [or specific list of] rlimits
    --pty                  run task zero in pseudo terminal
```

Problem Description: The **scontrol** man page for the **update** command refers to **MinMemory=<megabytees>**. This should read **MinMemoryNode=<megabytes>**

8.8 InfiniBand Switches

The **ibsw_fw_update** command to update the firmware for **InfiniBand** switches is only supported for **Mellanox** switches. Please contact Bull technical support regarding **WIPRO** switches.

8.9 syslog-ng

Problem Description: For a large cluster, if the number of maximum simultaneous connections is above the default values, logs are dropped without any warning.

Solution: Increase the **max-connections** parameter for the **TCP** and **unix-stream** connections in the **syslog-ng** configuration file on the Management Node.

Example

Increase the default value of 500 to 3000 as shown below for the **TCP** line:

```
{ tcp("X.X.X.X") port(5000); keep-alive(yes) max-connections(3000)); } ;
```

and for the **unix-stream** line:

```
{ unix-stream ("/dev/log" max-connections(3000)); internal(); };
```

8.10 Bull System Manager

8.10.1 Map view refresh

Problem Description: The **Refresh** button in **Map** view for the **BSM** console updates the **Map** view incorrectly.

Solution: Do not use the **Back** button in the top left hand corner of the menu bar when navigating between **Map** views, use the **Back to previous map** link in the top right hand corner instead. The **Refresh** button will then update correctly.

8.10.2 Display of Interconnect and Lustre Performance Graphics

Problem Description: Interconnect and **Lustre** Graphics are not displayed in the **Global Cluster Performance** Window.

Solution: 1. Customize the **/etc/gmond.conf** file for the **Lustre** metrics. See the instructions in the file.

2. Edit the **/etc/php.ini** configuration file

3. Change the **register_long_arrays** parameter from **Off** to **On**

4. Reboot the **httpd** service.

```
service httpd restart
```

Note There is no solution for the display of the Interconnect Performance Graphics for this release.

8.10.3 Nagios and PBS Professional

Problem Description: **Nagios** does not monitor any activities relating to **PBS Professional** following the upgrade from **BAS5 for Xeon V1.2** to **bullx cluster suite XR 5v3.1U1**.

Solution: No solution available for this release.

8.11 Intel Tools

8.11.1 Intel Vtune Performance Analyzer for Linux

Problem Description: The following message appears when running **Intel Vtune**:

```
Loading the VTune analyzer sampling driver:
FATAL: Module vtune_drv not found.

Error:  unable to find device "vtune" in /proc/devices !

FATAL: Error running install command for vtune_drv
```

Solution: This is not a fatal error but only a warning and results when the kernel loaded is not the Bull modified kernel.

See Chapter *7* in the **bullx cluster suite XR 5v3.1U1** *Installation and Configuration Guide* for more information.

8.11.2 Intel Fortran version 11 compilers with the Fortran 90 standard.

Problem Description: **Intel** Fortran version 11 compilers produce errors with the **Fortran 90 standard** using the **r16** option on clusters with MPIBull2. The technical description of this problem follows:

*The Intel® C compiler does not support a 16byte (i.e. 128 bit) definition of long double. By default on Linux the Intel® C compiler defines long double as a 10byte (80 bit) representation. Intel® C compiler supports the data types of the native C++ Compiler (in this case gcc). The Intel Fortran compiler with **-r16** is defining the floating point number to by 16bytes (128bits). This is why the values can not be compared or printed when you use the **-r16** switch.*

Solution: There is no solution for this release.

8.12 Storage

8.12.1 I/O device aliases on nodes

Problem Description: If I/O aliases have been deployed on a node (using the **stordepmap** command), these aliases are recreated at boot time only if Fibre Channel connections are present. If a FC cable is reconnected after the boot phase, the remote I/O devices for this link will be available to the operating system (**/dev/sdxxx**) but not the aliases (**/dev/ldn.xxx**).

Solution: Force the creation of I/O aliases by running the command:

```
service stormapping start
```

8.13 IBS Tools

8.13.1 IBS tool for InfiniBand Diagnostics

Problem Description: Error message *"…..could not find ParserDetails.ini"* appears when running IBS.

Solution Run the command below as root

```
perl -MXML::SAX -e "XML::SAX->add_parser(q(XML::SAX::PurePerl))->save_parsers()"
```

8.14 Electric Fence

Component: Electric Fence

Problem Description: Debugging does not work on **InfiniBand** clusters with **MPI**

Solution No Solution

8.15 Conman and IPMI Tools

8.15.1 ipmitools on NovaScale R440 and R460 Platforms

Problem Description: The **ipmitool chassis status** and/or **nsctrl** commands indicate that the machine is **off** when the machine is **on**.

Solution: Run the **ipmitool bmc reset warm** command locally on the machine, or remotely via the LAN connection.

8.16 Additional Ethernet Cards for NovaScale R460 Machines

Problem Description: The **installnfs** command will not run for the **installation if there is an additional Ethernet Card on the NovaScale R460** machine. The slotting information in section *G.3.3* in the **BAS5 for Xeon V3.1** *Installation and Configuration Guide* is incorrect

Solution Additional Ethernet boards can only be added on **NovaScale R460** machines which are used as Management Nodes. Additional Ethernet boards cannot be installed on Service Nodes that are included in the deployment.

8.17 Bull System Backup Restore

Problem Description: A backup CD created with **BSBR** on a **NovaScale R423** machine will not mount when inserted into a **NovaScale R480** machine.

Solution: The problem is that the **NovaScale R423** machines use **/dev/hdb** for the CD-ROM drive, whereas **NovaScale R480** machines use **/dev/hda.** Before making the backup create the **/dev/hda** device on the **NovaScale R423** machine for the CD-ROM drive.

Chapter 9. Default Logins for different Cluster Elements

Element	Login	Password	Comments
Baseboard Management Controller	administrator	administrator	
InfiniBand switches	enable	voltaire	Equivalent to root → used for configuration switch
	admin	123456	Read only
Ethernet switches	admin	admin	
	admin	admin	Same login and password for root
DDN Storage subsystems	admin	password	The same logins are defined in the **/etc/storageadmin/ddn_admin.conf** file.
NEC Storage subsystems	iSM	iSM	Change to **admin** and **password** to match logins defined in the **/etc/storageadmin/nec_admin.conf** file.
Xyratex Optima 1200 Storage subsystems	admin	password	The same logins are defined in the **/etc/storageadmin/xyr_admin.conf** file.
EMC/DGC CX3 or CX4 Series Storage systems	User defined at the first connection	User defined at the first connection	It is recommended to use **admin** and **password** in the same way as for other systems.

BULL CEDOC
357 AVENUE PATTON
B.P.20845
49008 ANGERS CEDEX 01
FRANCE

REFERENCE
96 A2 73EJ 00