ITEA

INFORMATION TECHNOLOGY

FOR EUROPEAN ADVANCEMENT

(ITEA2 09011)

Optimize HPC Applications on Heterogeneous Architectures

Deliverable: D5-1.5.2 (L2.5.2)

Bullx SCS4 R4 : System Software – first version

Version: V1.1

Date: December 2013

Authors: Bull SAS

Status: Final

Visibility: Public

HISTORY

Document version #	Date	Remarks	Author
V1.0	November 2013		D.Foueillassar
V1.1	December 2013	Introduction added plus some corrections	D.Foueillassar

TABLE OF CONTENTS

1. Introduction

This document is part of the PerfCloud project (or Work package 5 of H4H) deliverables.
The deliverable D5-1.5.2 consists of the SuperComputer Suite (SCS) Release 4 Beta software including those specified with PerfCloud D5-1.5.1 (Basic software) and D5-1.5.1 extension (Native Xeon-Phi) and the Release notes which are in this last document.
The software (beta version) and present release notes have been made available end of November.
Final version and an update of this document will be made available in January when the software will be in General Delivery.
In this document you'll find User documentation list, the supported hardware and corresponding firmware levels and the restrictions and known limitations for this beta software.

2. Delivery Contents

2.1 Designation

This Software Release Bulletin applies to
bullx supercomputer suite 4 Release 4 Delivery Media

2.2 Media delivered

76743460-101	bullx scs 4 R4 beta bullx DE for RHEL 6
76743461-101	bullx scs 4 R4 beta bullx MC for RHEL 6
76743462-001	bullx scs 4 R4 beta bullx MM for RHEL 6
76743463-001	bullx scs 4 R4 beta bullx PFS for RHEL 6
76743464-001	bullx scs 4 R4 beta bullx BM for RHEL 6
76743465-001	bullx scs 4 R4 beta bullx MPI for RHEL 6
76743466-001	bullx scs 4 R4 beta bullx EP for RHEL 6
76743467-101	bullx scs 4 R4 beta bullx HPC Foundation for RHEL 6
76743469-001	bullx scs 4 R4 beta Documentation for RHEL 6

76743399-101	RHEL 6.4 for EM64T Installation
76743399-001	RHEL 6.4 for EM64T Supplementary
76743401-101	RHEL 6.4 RPMS Sources (2 DVDs)
76743433-101	RHEL 6.4 Compute Nodes for EM64T Installation
76743433-001	RHEL 6.4 Compute Nodes for EM64T Supplementary
76743434-101	RHEL 6.4 Compute Nodes SRPMS Source

Note Check the Bull Support Web site for the most up-to-date product information, documentation, firmware updates, software fixes and service offers: http://support.bull.com

3. Documentation

bullx scs 4 AE R4 Documentation Title	Reference
bullx scs 4 R4 Software Release Bulletin (SRB)	86 A2 91FK 01
bullx scs 4 R4 Installation and Configuration Guide	86 A2 74FK 01
bullx scs 4 R4 – Extreme Pack Installation and Configuration Guide	86 A2 75FK 01
bullx MC Administration Guide	86 A2 76FK 01
bullx MC Monitoring Guide	86 A2 77FK 01
bullx MC Power Management Guide	86 A2 78FK 01
bullx MC Storage Guide	86 A2 79FK 01
bullx MC Security Guide	86 A2 81FK 01
bullx MC InfiniBand Guide	86 A2 80FK 01
bullx MC Ethernet Guide	86 A2 82FK 01
bullx PFS Administration Guide	86 A2 86FK 01
bullx DE User's Guide	86 A2 84FK 01
bullx MPI User's Guide	86 A2 83FK 01
bullx BM User's Guide	86 A2 85FK 01
bullx MM Argos User's Guide	86 A2 87FK 01
bullx EP Administration Guide	86 A2 88FK 01
Extended Offer Administration Guide	86 A2 89FK 01
bullx scs 4 R4 Documentation Overview	86 A2 90FK 01

All the documentation is installed under the directory:

/usr/share/doc/bullxscs4/AE4

All the documentation can be downloaded from the **Bull Support Web**, as described below.

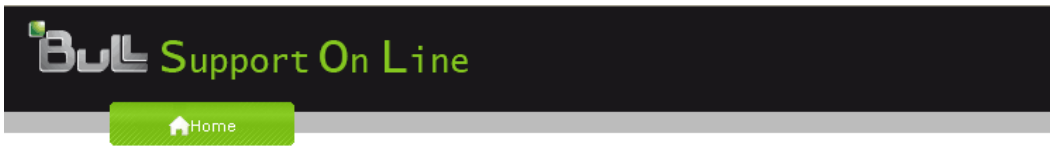
How to download the documentation from the Bull Support Web

Go to: http://support.bull.com/documentation/byproduct

Register using your customer login or your internal login (Bull representative).

To get the list of the manuals related to bullx scs 4, select:

Infrastructure > Extreme Computing > bullx supercomputer suite



Documentation Catalog

Download the manuals you need, or the Portfolio, which includes all the manuals.

4. Supported Platforms

The following hardware is supported by **bullx scs 4 R4.**

4.1 Platforms

4.1.1 Management Nodes

R-Series:

1. bullx R423-E2T2
2. bullx R424-E2
3. bullx R423-E3I
4. bullx R423-E3
5. bullx R424-E3
6. bullx R424-F3

S-Series:

7. bullx S6030 (Nehalem EX)
8. bullx S6030 32 DIMM (Westmere EX)

4.1.2 Compute Nodes

R-Series:

9. bullx R422-E2
10. bullx R423-E2T2
11. bullx R424-E2
12. bullx R425-E2
13. bullx R423-E3I
14. bullx R423-E3
15. bullx R424-E3
16. bullx R425-E3 (Nvidia K20c, Xeon Phi 3120A) (new)
17. bullx R424-F3
18. bullx R428-E3
19. bullx R421-E3 (Nvidia K20m, Nvidia K20Xm, Xeon Phi 5110P, Xeon Phi 3120P, Xeon Phi 7120P) (new)

B-Series:

20. bullx B500
21. bullx B505 (M2050, M2070, M2090)
22. bullx B510
23. bullx B510-IBF

24. bullx B710
25. bullx B715(Xeon Phi 7120X) (new)
26. bullx B515 (Nvidia K20m, Nvidia K20Xm, Xeon Phi 5110P)

S-Series:

27. S6010 (Nehalem EX)
28. S6030 (Nehalem EX)
29. S6030 32 DIMM (Westmere EX)
30. S6010 (Westmere EX)
31. bi/quadri module S6010 (Nehalem EX)
32. bi/quadri module S6010 (Westmere EX)

4.1.3 Service Nodes

R-Series:

33. bullx R423-E2T2
34. bullx R424-E2
35. bullx R423-E3I
36. bullx R423-E3
37. bullx R424-E3
38. bullx R424-F3

S-Series:

39. bullx S6030 (Nehalem EX)
40. bullx S6030 (Westmere EX)

4.2 Interconnect

Ethernet Foundry FastIron

41. FLS624/648
42. FLS624/648-STK
43. FCX624/648-E
44. FCX624/648-S
45. NetIron MLX
46. BigIron RX

Ethernet CISCO

47. Catalyst 3560
48. Nexus 7000
49. Catalyst 3750
50. Catalyst 3750X

51. SG300-28
52. SG300-52
53. Catalyst 2960S-(24TS & 48TS)-L

InfiniBand Voltaire Grid

54. Director 4036
55. Director 4700

InfiniBand Mellanox

56. MTS 3600 36p
57. IS5030
58. MSX6036F-1SFR-QSFP
59. MSX6536-10R
60. MSX6025

4.3 Storage

StoreWay

61. Optima1500

EMC-CLARiiON

62. CX4 M120
63. CX4 M240
64. CX4 M480

DDN

65. SFA10K
66. SFA10KT
67. 9900

NetApp (LSI)

68. XBB2
69. E5400 (FC attachment only)
70. E2600
71. E5500 (SAS attachment only)

4.4 Board Support

GPU Nvidia

72. Tesla C2050
73. Quadro 6000
74. K20c for R425E3

Ethernet

75. Myricom HBA 10Gb
76. Emulex OCE 11102 10Gb

4.5 Hardware Miscellaneous

Bull Cool Cabinet Door

Power Distribution Unit from APC

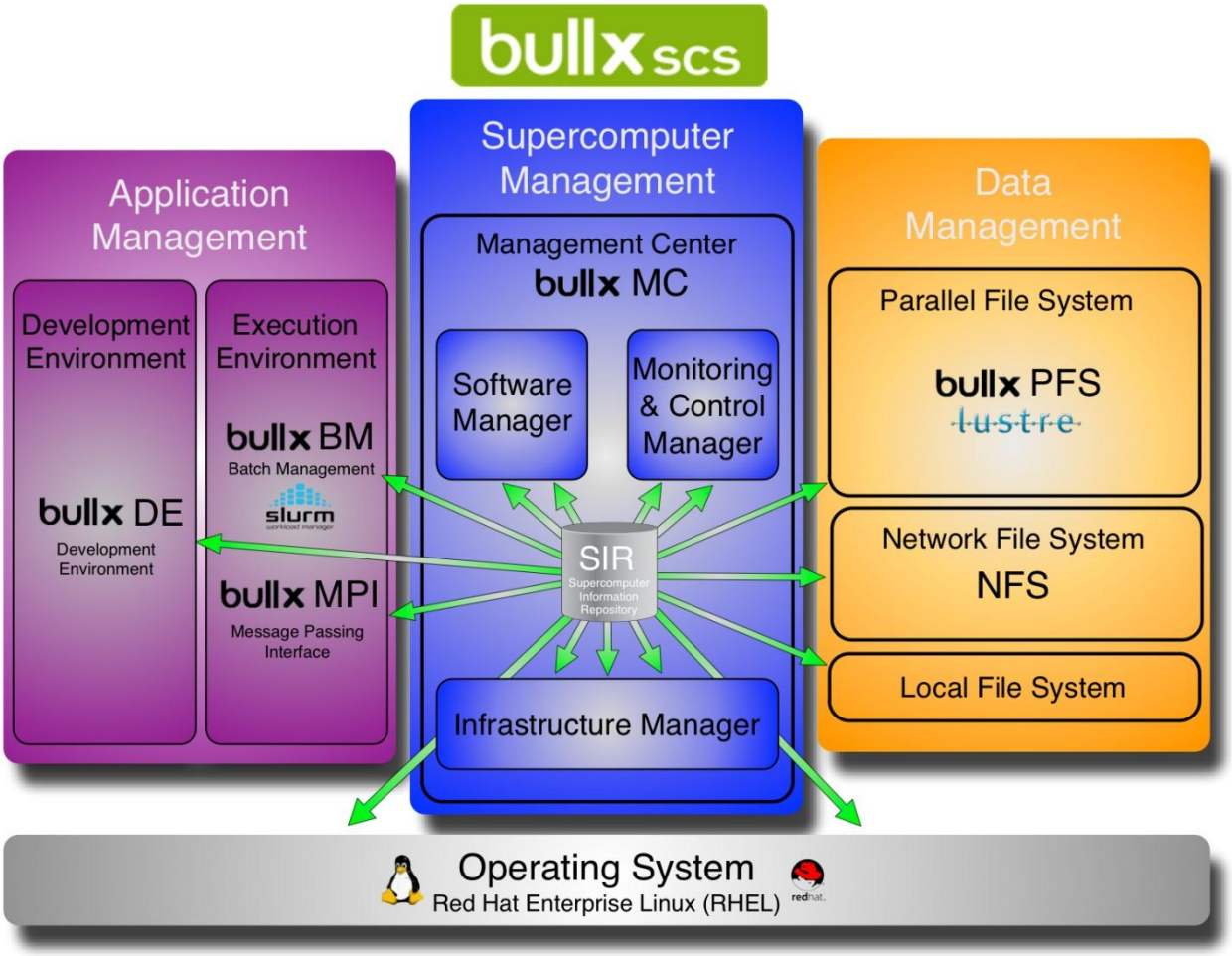
77. AP7921,
78. AP7922

5. Release Description and Features

This chapter describes the **bullx scs 4 R4** features.

5.1 Introduction

bullx supercomputer suite is a complete HPC dedicated software suite that helps customers to implement, administrate and fully operate their HPC clusters in an easy, reliable and efficient way.



bullx supercomputer suite runs on Red Hat Enterprise Linux. The scalability, manageability, security and high availability features of RHEL make it the ideal platform to run HPC applications. The longstanding technical relationship between Bull and Red Hat, with experts on both sides working closely together, ensures bullx scs leverages from the fine-tuned and completely optimized RHEL software. This effort aims to provide the environment required to deploy and administrate very large supercomputers and to run the most demanding HPC applications.

With Bull's years of experience in deploying large scale HPC supercomputers and applications, combined with Red Hat deep knowledge in Enterprise class Linux, Bull - Red Hat joint customers have access to professional class worldwide support services.

5.2 bullx scs 4 R4 Core Offer

The software included in the core offer includes the following products, detailed below:

- Red Hat 6.4
- bullx Management Center (bullx MC)
- bullx Maintenance Manager (bullx MM)
- bullx Parallel File System (bullx PFS)
- bullx Development Environment (bullx DE)
- bullx Batch Manager (bullx BM)
- bullx MPI
- bullx Extreme Pack (bullx EP)

bullx supercomputer suite is highly modular. Most software components can be used as standalone products.

These products are detailed below.

5.3 Red Hat 6.4

Red Hat 6.4 comprises two parts:

Red Hat 6.4 official base distribution used "as is".

Red Hat 6.4 HPC foundation containing:

RPMs that are not available on the Red Hat 6.4 media but are mandatory for one or more HPC applications

the official Red Hat 6.4 fixes that are mandatory to run bullx products

5.4 bullx Management Center (bullx MC)

bullx MC contains management tools for the following:

Cluster installation & configuration

Cluster control

Cluster monitoring

Cluster maintenance

Power management

Error management

IDC framework

Batch management integration

Basic security features

5.5 bullx Maintenance Manager (bullx MM)

bullx MM is an option of the bullx MC and cannot be used alone. It contains:

Argos maintenance software

Metrology sample code to be adapted to Customer requirements

5.6 bullx Parallel File System (bullx PFS)

bullx PFS contains:

Lustre

shine (Lustre Management tool)

HA integration scripts

Robinhood backup

5.7 bullx Development Environment (bullx DE)

bullx DE provides all the tools necessary to develop and optimize applications.

5.8 bullx Batch Manager (bullx BM)

bullx BM offers a full integration of Slurm in bullx SCS (bullx MPI, bullx MC, bullx DE)

5.9 bullx MPI

bullx MPI contains the Bull MPI library

5.10 bullx Extreme Pack (bullx EP)

bullx EP is an option of the Management Center and cannot be used alone. It provides extended management capabilities for very large Distributed Architecture clusters (up to 800 nodes).

5.11 bullx scs 4 R4 Extended Offer

The **extended offer** enriches the core offer and is based on commercial products that are only delivered by Bull. These products include:

Batch Managers:

- PBS Professional® Batch Manager
- IBM® Platform© Computing LSF Batch Manager

Intel® Products:

- C++ and Fortran Professional Compilers for Linux
- VTune™ Performance Analyzer for Linux
- IDB Debugger

DDT Debugger

TotalView® Debugger

5.12 New Features for bullx scs 4 R4

This section lists the new features for **bullx scs 4 R4.**

6. bullx scs 4 R4 Core Offer

6.1 bullx scs 4 R4 includes the following products:

- bullx MC 4.4
- bullx MM 4.4
- bullx PFS 2.4
- bullx DE 2.1
- bullx BM 2.60.0
- bullx MPI 1.2.5
- bullx EP 4.4

6.2 bullx scs 4 R4 includes the following new features:

6.2.1 Base distribution/HPC Foundation

Red Hat 6.4

6.2.2 bullx MC

Extended HA features

- Dual failure support on 4 nodes lustre I/O cells
- Monitoring: New tools to help the administrator to take the right actions in the event of failure.

Xeon Phi

- Native mode support
- Secure authentication for the execution of a job

NVIDIA GPU : GPU direct version V1 supported

NFS servers

- XFS filesystem support (up to 100 terabytes)
- Following configurations HA NFS N + 1 are supported:
- Configuration 2 active nodes + 1 standby node
- Configuration 3 active nodes + 1 standby node

InfiniBand

- Better scalability and supervision of InfiniBand networks thanks to the new IBMS features
- Advanced InfiniBand monitoring

Tools improvement to complete the dual boot feature

cdbm commands: finalization and robustness of these commands

Monitoring

- Improve the monitoring of the equipment
- Monitoring local disks: collect counters hard to analyze occurrences of errors, and make statistics

Allow decoding DIMMs to recognize the memory card down

Soft raid level 0 and 1 supported by the installation and deployment tools

New command to monitor the installation and configuration of Ethernet switches (Cisco only)

Bfup (Bull Firmware Update) improvements (connected with clusterdb, ...)

6.2.3 Bullx PFS

Lustre 2.4

- Performance Optimizations
- Imperative Recovery
- Metadata Cluster
- Lustre Clients Activity Monitoring
- Lustre HA Monitoring
- NFS server in user Space in Technology Preview mode[1]

Extension of Lustre monitoring to allow:

- To know the bandwidth consumed by Lustre I/O for each job
- To follow the overall I/O activity

6.2.4 bullx BM

slurm 2.6

- Fine grain accounting data (IO local, IO Lustre, Network, Temperature, Power consumption)
- Accounting per group of users and jobs

Enhanced launching performance (PMI2 supported)

6.2.5 bullx MPI

Based on OpenMPI 1.6.4

gfortran & ifort compilers support

GHC (Global Hierarchical collectives) : barrier & reduce optimization

Xeon Phi support

Enhanced performance

[1] Technology Preview features will enable customers to test the functionalities and feedback to Bull in order to enhance them before including them as generally available features in future releases of bullx scs. In the mean time Technology Preview features may not be functionally complete or enough robust for production use. However Bull will provide best effort to resolve issues related to the use of those features.
Although the normal life cycle of Technology Preview features is to become, after several possible iterations, generally available and fully supported features however this is not guaranteed. Neither is guaranteed the compatibility of such features from one release to a next one.

6.2.6 bullx DE

Up to date version for all included tools

bullx prof

multi-domain, / non intrusive, /multi-environment profiling

Including Timing, Counters & IO

I/O and MPI profiling

Xeon Phi counters access through PAPI

Two versions of the tools: MPI Analyzer, Darshan Scalasca, xPMPI, OpenSpeedShop
- A version from bullx MPI
- A version from Intel MPI

6.2.7 Bullx MM

Argos
- Show effect of action event without need to save the intervention.
- Option in command agsint for grouping or not an intervention.
- Use nodeset in filter field.
- Add open event in a closed intervention.
- Fix and memorize column size.

6.3 bullx scs 4 R4 Extended Offer

PBS Pro 13,

LSF 9.1

Intel MPI v 4.1

TotalView 8.11 ,

Allinea DDT/MAP 4.0,

Intel compilers v14

7. Firmware

Here is the list of firmware validated with bullx scs4 R4.

7.1 Nodes

Node name	Technical State (TS)
B-Series (bullx B5XX)	TS 54.02
S-Series (S60XX)	TS 39.02
B-Series (bullx B7XX)	TS 15.02
R423E3i	R3iE3X11
R423E3	R23E3X31
R425E3	R25E3X31
R424E3	R24E3X41
R424F3	R24F3X41
R428E3	R28E3X32

7.2 Switch Ethernet

Switch Ethernet name	Firmware
Cisco 2960-S	c2960s-universalk9-mz.122-55.SE5
Cisco 3560	c3560-ipbasek9-mz.122-25.SED.bin
Cisco 3750	c3750e-universalk9-mz.122-55.SE3
Cisco SG300	SG300-1.1.1.8.bin
Broadcom BCM-56224 (CMM/ESM)	5.2.0.4
Broadcom BCM-56524 (TSM)	6.2.0.4
Foundry FCX-E	7300a
Foundry FCX-S	7300a
Foundry NetIron MLX	2.7.02n
Foundry BigIrong RX	2.7.02j

7.3 Switch IB

Firmwares are delivered by ibms tools. Use the command ibms_fw to update the cluster.

7.4 Storage

https://10.223.29.139/~hpc-rd-storage/notes/

file: NetApp_FW_Matrix.pdf

8. Restrictions, Limitations, Recommendations

This chapter lists the restrictions, limitations and recommendations for this delivery.

8.1 High Availability

The following restrictions apply to High Availability for this delivery.

8.1.1 HA Heartbeat Networks

The 2 heartbeat networks must be 2 distinct Ethernet networks,

A network used for heartbeat must never be attached to an Ethernet bridge interface.

A network used for heartbeat must not be attached to an IPoIB interface.

A network used for heartbeat must not be attached to a bonding interface.

For a HA group of 2 nodes, one point to point network is mandatory. Two point to point networks is recommended. If it is not the case, the second network can be a switched network.

For a HA group of more than 2 nodes, one dedicated switched network is mandatory. Two dedicated switched networks are recommended. If it is not the case, the second network can be the management network for example.

8.1.2 Mounting Shared ext3/ext4 FS

Important **You must never mount a shared ext3/ext4 File System on both nodes of a HA group at the same time, otherwise the data will be corrupted. So, never use the mount command manually for the shared ext3/ext4 FS. These FS must always be mounted using the haresl start command.**

8.1.3 NFS User's Home not managed in HA

The NFS User's Home partitions are not managed by the HA configuration. So if the NFS resource switches to another node, you will need to manually mount the NFS User's Home on this node.

8.2 bullx MC

The following restrictions apply to bullx Management Center for this delivery.

8.3 ClusterDB

The **hard_id** field must not be modified as, for example, this impacts the management of cluster extensions. Equipment must be renamed using the command:

```
cdbm-equipment rename equipment_old_name[x,x+n]
equipment_new_name[y,y+n]
```

8.4 Cdbm commands

[NF0016082] cdbm-equipment show FAILS if columns are "cpu_total,memory_total"

In cdbm-commands, two items from the same set cannot be used simultaneously in the --columns option.

Available sets are:

Set1 : *_status (nagios_status, mode_status ...)

Set2 : memory_available, cpu_available

Set3 : memory_total, cpu_total

Example:

cdbm-equipment show --columns node,cpu_available,cpu_total ==> OK (cpu_available and cpu_total are from different sets)

cdbm-equipment show --columns node,cpu_available,memory_available ==> KO (cpu_available and memory_available are from the same set)

8.5 lc command

[NF0012065] lc is not RFC compliant for hostname.

The lc command doesn't support hyphen or dash character « - » in the basename.

For lc command, « - » is treated as minus operator and exclude nodes from a nodeset.

For example :

lc bmc[0-1]-bmc0 => bmc1

lc bmc-node[0-1]=> bcm

lc bmc-node[0-1]-bmc-node0 => ""

8.6 equipmentRecord command

The following restrictions apply to the **equipmentRecord** command for this delivery.

8.7 bullx S6010 multi-module

During discovery with **--cluster** option (whole cluster), or with the **--type node** option or
--type hwmanager but without the **--name** option, **equipmentRecord** will try to discover the MAC addresses of all the modules of the bullx S6010 multi-module.

With options **--type node --name <module_name>**, **equipmentRecord** will try to detect the MAC address of the specified module (--**name** option), but not the other modules (the **fusion_node_id** field is not used by **equipmentRecord**).

The same applies with options **--type hwmanager --name <module_hwm_name>**.

If some Ethernet ports are disabled in the BIOS, their MAC address cannot be discovered using **equipmentRecord**.

8.8 bullx R42x Cabled on One Common Port (node + BMC)

Before executing **equipmentRecord**, it is necessary to switch on the machines manually (not only the BMCs, but also the nodes) so that both MAC addresses become detectable.

8.9 Argos

[NF0014908] Some accesses are not available

The following accesses are not available due to an incompatibility with the installed **XULRunner** version:

Access to Bullforge (through the **Problems base** tab or through **BullForge** link of management views) from Intervention and Maintenance form perspectives

Access to Wiki (through the **Knowledge base** tab or through the Wiki link of management view) from Event perspective

The following error message is sent in these cases (see log for more details):

```
Creation of internet browser not successful
```

[NF0014906] The option “show preview” is not available

The previsualization of exported data from Equipment, Event, Intervention, Maintenance form and Metrology perspectives ("Show preview" button in the corresponding views) is not available due to an incompatibility with the installed XULRunner version.

The following warning message is sent in these cases:

```
The preview isn't available.
```

Workaround

For these two defects, the workaround is:

1. Donwload XULRunner version 1.9 as a tar file on the node(s) on which Argos GUI is installed.

Untar this file in a directory <XULRunnerPath> (ex: /opt/argos/xulrunner).

Add the following line to the **/opt/argos/gui/argos.ini** file :

```
-Dorg.eclipse.swt.browser.XULRunnerPath=<XULRunnerPath>
```

Launch Argos GUI.

8.10 Ksis and R425-E2 and R423-E2T2 Nodes

The images ksis taken from the machines R425-E2 and R423-E2T2 are not compatible with all other models.

8.11 Peer to peer transfers with NVIDIA GPU Direct.

In hardware equipped with NVIDIA GPU (bullx R425-E3, bullx R421-E3, bullx B505, bullx B515, bullx B715), the direct peer to peer data transfer between two GPUs have to be disabled.

8.12 bullx PFS (Lustre)

The following restrictions apply to bullx Parallel File System for this delivery.

8.13 OOM Situation on Lustre/LNET Routers

Lustre routers must NOT be used as **Lustre** clients as this could lead the machine to an out-of-memory situation.

8.14 Using InfiniBand Multirail

To use Lustre with **InfiniBand multirail**, it is necessary to configure IPoIB interfaces as follows:

On the server side, where there are two InfiniBand interfaces:
It is required to use two different subnets, one for each interface.

On the client side, where there is only one InfiniBand interface:
An IP alias has to be created so that the client node has one logical interface on each subnet, matching the subnets defined on the server side.

8.15 Checksum should be Disabled on Lustre Clients

Client checksum is enabled by default, which could lead to poor performance on clients.

Workaround

Check that the **/etc/shine/tuning.conf** file contains the two following lines:

After "alias declaration section":

```
alias checksums=/proc/fs/lustre/osc/*${ost}*/checksums
```

And after "Tuning Parameters":

```
"0" checksums CLT
```

To check if this client checksum function is properly disabled, enter:

```
cat /proc/fs/lustre/osc/*/checksums
```

Every file should display the '0' value.

8.16 Filesystem Status Using Shine in an HA Context

[NF0013153] - HA Lustre: After failover, wrong status for target

When a filesystem is configured with HA capability, the **shine status** command displays inaccurate states on targets that are migrated to failover nodes.

To display the target's actual status with **shine**, you must explicitly specify the failover node on which the target was migrated, using option **-f**
Migrated targets can be retrieved with the **hashine monitorfs** command.

8.17 Example with a 3 OSS HA group (machu[7-9])

Let's assume `machu8` has been fenced for some reason.

```
# hashine monitorfs -f fs3
```

```
+---------------------------+---------+--------+--------------+--------------------+
|                           | Status  | Node   | Configured ? | Migration status   |
+---------------------------+---------+--------+--------------+--------------------+
| mdt_machu6ldn.cx480.mdtfs3 | Started | machu9 | Yes | Running on primary node    |
| ost_machu9ldn.da0.d0fs3    | Started | machu9 | Yes | Running on primary node    |
| ost_machu9ldn.da0.d1fs3    | Started | machu9 | Yes | Running on primary node    |
| ost_machu9ldn.da0.d2fs3    | Started | machu9 | Yes | Running on primary node    |
| ost_machu9ldn.da0.d3fs3    | Started | machu9 | Yes | Running on primary node    |
```

```
| ost_machu8ldn.da0.d4fs3    | Started | machu9 | Yes | Running on failover node
                                                        #01 (prefered node machu8) |
| ost_machu8ldn.da0.d5fs3    | Started | machu7 | Yes | Running on failover node
                                                        #02 (prefered node machu8) |
| ost_machu8ldn.da0.d6fs3    | Started | machu9 | Yes | Running on failover node
                                                        #01 (prefered node machu8) |
| ost_machu8ldn.da0.d7fs3    | Started | machu7 | Yes | Running on failover node
                                                        #02 (prefered node machu8) |
+----------------------------+---------+--------+-------------+--------------------+
```

```
# shine status -f fs3 -V target
```

```
FILESYSTEM TARGETS (fs3test)
+-------------------------+-----+----+-----------+--------------------+----------+
|        target id        |type |idx |  nodes    |      device        | status   |
+-------------------------+-----+----+-----------+--------------------+----------+
|MGS                      |MGT  | 0  |machu6     |None                |external  |
|mdt_machu6ldn.da0.mdtfs3 |MDT  | 0  |machu6     |/dev/ldn.da0.mdtfs3 |online    |
|ost_machu9ldn.da0.d0fs3  |OST  | 0  |machu[7-9] |/dev/ldn.da0.d0fs3  |online    |
|ost_machu9ldn.da0.d1fs3  |OST  | 1  |machu[7-9] |/dev/ldn.da0.d1fs3  |online    |
|ost_machu9ldn.da0.d2fs3  |OST  | 2  |machu[7-9] |/dev/ldn.da0.d2fs3  |online    |
|ost_machu9ldn.da0.d3fs3  |OST  | 3  |machu[7-9] |/dev/ldn.da0.d3fs3  |online    |
|ost_machu8ldn.da0.d4fs3  |OST  | 4  |machu[7-9] |/dev/ldn.da0.d4fs3  |offline   |
|ost_machu8ldn.da0.d5fs3  |OST  | 5  |machu[7-9] |/dev/ldn.da0.d5fs3  |offline   |
|ost_machu8ldn.da0.d6fs3  |OST  | 6  |machu[7-9] |/dev/ldn.da0.d6fs3  |offline   |
|ost_machu8ldn.da0.d7fs3  |OST  | 7  |machu[7-9] |/dev/ldn.da0.d7fs3  |offline   |
+-------------------------+-----+----+-----------+--------------------+----------+
```

Targets normally running on node `machu8` are seen as **offline** with shine output whereas they are actually started on the other nodes of the HA group.

To display the actual state of target `ost_machu8ldn.da0.d5fs3`, enter:

```
# shine status -f fs3 -l fs3-OST0005 -V target -F machu7
```

```
FILESYSTEM TARGETS (seb)
+-------------------------+-----+----+------------+--------------------+--------------------------
+
|       target id         |type |idx | nodes      | device             |status
|
+-------------------------+-----+----+------------+--------------------+--------------------------
+
| ost_machu8ldn.da0.d5fs3 |OST  | 5  | machu[7-9] |/dev/ldn.da0.d5fs3  |recovering for 269s (0/1)
|
+-------------------------+-----+----+------------+--------------------+--------------------------
+
```

8.18 Shine show Command

[NF0013150] shine show info - option -f no more accepted

Shine version 0.912 does not accept the **-f** option with the **shine show info** command.

Workaround

Use the following command without the **-f** option:

```
shine show info
```

8.19 bullx BM (SLURM)

The following restrictions apply to bullx Batch Manager for this delivery.

8.19.1 Energy Accounting Plugin Configuration and Usage

Administrators that configure the new **acct_gather_energy/rapl** plugin need to configure **AcctGatherNodeFreq<200** and also **JobAcctGatherFrequency<200** in order to be sure that the values collected from the RAPL sensors do not overflow, which might give false results.

Current implementation of the plugin does not allow the correct management of energy measurements on clusters where some of the nodes have processors that do not support RAPL (RAPL is supported on Sandy-Bridge and later architectures).
Hence, it is preferable to configure the plugin for energy accounting on clusters where all the nodes support RAPL (homogeneous clusters). For clusters where some of the nodes have RAPL support and some do not (heterogeneous clusters), only the nodes that have RAPL support will give the correct measures. The nodes without RAPL will report errors (without slurmd aborting). If all the nodes allocated to a job have RAPL support, the energy consumption reported for the job will be correct. If any of the nodes allocated to a job does not support RAPL, the energy consumption reported for the job will be incorrect. On heterogeneous clusters, the Slurm administrator can use partitioning to isolate the nodes that do not support RAPL.

Current implementation of the plugin allows the reporting of the per step/job energy consumption in case of exclusive job allocation. This means that jobs need to use the nodes exclusively to report correct energy consumption. In case of node sharing the reported consumed energy will not have a correct value.

8.19.2 NF0014005 - Sview allows a non root user to switch to admin mode

When using **sview** a normal user may change to **Admin** mode but cannot do anything to harm the cluster. The only thing that he can do in Admin mode is to see jobs of other users except their own, without being able to influence their execution.

This will be corrected in a future version.

8.19.3 HA SLURM

[NF0014026] With HA, how can we start slurm with option -c or -D ?

If HA-SLURM is enabled, the start of the slurm services does not allow the use of some parameters. These parameters not allowed are:

For **slurmctld** controller:

startclean|restart|reconfig|condrestart|test

For **slurmdbd** controller:

restart|condrestart|reconfig

If one of the above modes of **service start** is needed the best procedure is the following:

1. Deactivate the management of slurm service from Pacemaker by setting this service as 'unmanaged':

```
crm resource unmanage slurm
```

Make the start of the service with the needed parameters, make the observations for as long as it is needed and in the end stop again the service.

Reactivate the management of slurm service within Pacemaker:

```
crm resource manage slurm
```

Important: During the phase where slurm service is passed 'unmanaged', there is no HA support for slurm so the Administrators need to be careful.

8.20 Accounting is Not Available for sbatch without Step

[NF0013999] error with **sstat** when a job is launched with **sbatch**

If the accounting is configured, the values are available only for steps.

If a user want to have accounting on a **sbatch** script, he must create steps using **srun** commands in his script.

sacct will not print any value if there is not step, and **sstat** will return the "sstat: error: no steps running for job n"error message .

8.21 bullx MPI

The following restrictions apply to bullx MPI for this delivery.

8.21.1 MPI Cancel is Supported for Receive Requests only

bullx MPI conforms to the **MPI-2** standard. Certain aspects of this standard are not yet implemented. In particular **MPI_Cancel** is supported only for receive requests.

8.21.2 Bullxmpi-mic and mic runtime

The bullxmpi-mic can only use one intel mic runtime.

This runtime is loaded in slurm prolog script, as desbribed in slurm documentation “2.9.2 MIC prolog/epilog”.

8.22 Base Distribution and HPC Foundation

The following restrictions apply to the Base Distribution and HPC Foundation for this delivery.

8.22.1 x2apic Mode on bullx S6010 and bullx S6030 Multi-modules Platforms

[NF0010602] – The kernel boot **x2apic_phys** parameter must be added in **/boot/grub/menu.lst** to have the **x2apic** initialized in physical mode.

8.22.2 Power Management Issues on bullx S6010 multi-module

[NF0010562] Hang in C-states initialization code at boot time with Intel(R) Xeon(R) CPU E7- 4830 (Westmere) processors:

The deep C-states must be disabled on these platforms to prevent a system hang at boot time. To prevent the problem, the BIOS may be set to:

CPU C State	<Enable>
Package C State limit	<C1 state>
C1 Auto Demotion	<Disable>
C3 Auto Demotion	<Disable>
NHM C3 report	<Disable>
C1E	<Enable>
Monitor/Mwait	<Enable>

[NF0010563] Mesca BCS: Performance States Not working

This problem is fixed with BIOS version >= BIOSX 2.14.1

8.23 Numa I/O Extensions not supported by Redhat kernel on S6010 and S6030 Platforms

The Redhat kernel has no support for Numa I/O extensions available on the S6010 and S6030 platforms for bullx system.

The BIOS **NUMA IO extensions** option must be set to Disabled on these platforms when the RedHat kernel is installed to fallback to a minimal Numa I/O support.

8.24 Check Firmware Version for Emulex Adapter LPE 12002-M8

On the I/O nodes, check the firmware version of the adapters.

```
ioadm show adapter | grep -i lpfc | awk {'print $8'}
```

If the version is equal to, or greater than **2.00A4** , you do not need to upgrade the firmware.

Else update the firmware as described below.

To update firmware version for Emulex adapter LPE 12002-M8

1. Download the last version of Application kit CLI version here:
   http://www.emulex.com/downloads/emulex/linux/rhel61/management-and-utilities.html

Unzip elxocmcore-rhel5-rhel6-6.0.21.1-1.tgz:

```
tar -xvf elxocmcore-rhel5-rhel6-6.0.21.1-1.tgz
```

Choose the folder corresponding to your architecture and Linux version

Install the rpm:

```
rpm -ivh elxocmcore-6.0.9.1-1.x86_64.rpm elxocmlibhbaapi-6.0.9.1-
1.x86_64.rpm
```

Check that the tool works properly:

```
hbacmd ListHba
```

Download the last version of firmware and boot code here:

http://www.emulex.com/downloads/emulex/firmware-and-boot-code/lpe12002-firmware-and-bootcode/firmware-and-boot-code.html

Note: Choose "**Universal Boot**" for boot code version.

Unzip the 2 files.

Upgrade the firmware adapters:

```
for i in `/usr/sbin/hbacmd listhbas | awk '/Port WWN/{print $4}'`
do /usr/sbin/hbacmd Download ${i} /<Firmware Path File>
done
```

8.25 Dump with B515 and Xeon Phi

Kernel dump functionnality Kdump through sysrq (with the command "echo c > /proc/sysrq-trigger") will not work on B515 with Xeon Phi cards.

***Workaround*:**

Use kdump with NMI (normal use) : "power diag" or "nsctrl dump".

8.26 LSI/NetApp controllers clock is not synchronized to a time server

[NF0015554] NetAPP 5500 clock is not synchronized to a time server.

NTP configuration on controllers is not currently supported by NetApp

***Workaround**:*

It is recommended to setup a cron to synchronize the NetApp storage array clocks with the management node once a day.

```
Create a file on the management node (both nodes in case of HA MNGT)
under :
/etc/cron.d/storlsi.cron
containing a cron line for each NetApp storage array running the
following command :
storadm cli -da <da_name> "set storageArray time;"
```

Example :

```
cat /etc/cron.d/storlsi.cron
# Synchronize LSI/NetApp storage array time once a day at 4am
# minute hour dom month dow user cmd
0 4 * * * root PATH=$PATH:/usr/bin:/bin:/sbin /usr/bin/storadm cli -da da0 "set
storageArray time;" >/dev/null
0 4 * * * root PATH=$PATH:/usr/bin:/bin:/sbin /usr/bin/storadm cli -da da3 "set
storageArray time;" >/dev/null
```