Voltaire InfiniBand GridStack Software Revision 4.3.5.c

Release Notes

August 22, 2007

CONTENTS

1 Introduction

This document provides the release notes for Voltaire InfiniBand Stack, Revision 4.3.5c.

Please note some sections of this document are considered to be a beta version.

2 Release Description

This Voltaire GridStack^{TM} package is an Open Fabrics EWG OFED based distribution designed to ease installation and management of the Voltaire InfiniBand Stack (or InfiniBand Host Stack) on systems that include ConnectX HCAs.

This distribution is based on the OFED 1.2.c-11 GA release.

Please refer to the Voltaire HCA 4X0 User Manual for further detail.

3 Release Contents

Item	Description
Binary package compressed as a.tar file	4.3.5c_5
Documentation	Linux GridStack™ for HCA 4X0 User Manual (399Z00121)

4 Supported HCAs

The following ConnectX HCAs are supported in this release:

1. HCA420-EX
2. HCA420-EX-D
3. HP ConnectX HCAs
4. IBM ConnectX HCAs

5 Supported HCA Firmware Versions

1. HCA420-EX, HCA420-EX-D - 2.2.0
2. HP ConnectX HCAs – 2.2.0
3. IBM ConnectX HCAs - 2.2.0

6 Supported Operating Systems

1. RHEL 4 U 4	-	kernel 2.6.9-42-EL
2. RHEL 4 UP 5	-	kernel 2.6.9.55-EL
3. RHEL 5	-	kernel 2.6.18.8-EL5
4. Suse SLES 10	-	kernel 2.6.16.21-0.8
5. Suse SLES 10 sp 1	-	kernel 2.6.16.46-0.12

7 Supported CPU Architectures

1. x86_64 (AMD & EM64t)
2. PPC64 (IBM JS21)
3. HP BladeSystem c-Class (x86_64 & ia64)
4. IBM BladeCenter HS21 & LS21

8 Performance Envelope

Measurement		Platform
Test Name	Units	x86_64 PCI-E DDR
ib_write_lat	µSec	1.18
ib_rdma_lat	µSec	1.18
ib_read_lat	µSec	2.44
ib_send_lat	µSec	1.41
ib_write_bw	Mbits/Sec	11286
ib_rdma_bw	Mbits/Sec	11360
ib_rdma_bw_bidirection	Mbits/Sec	22397
ib_read_bw	Mbits/Sec	11371
ib_send_bw	Mbits/Sec	11316
iperf ipoib-cm	Mbits/Sec	11673
iperf ipoib	Mbits/Sec	3993
netperf ipoib-cm	Mbits/Sec	6375
netperf_ipoib	Mbits/Sec	4369
voltaire_mpi_bandwidth	Mbits/Sec	11848
voltaire_mpi_latency	µSec	1.7
open_mpi_bandwidth	Mbits/Sec	11904
open_mpi_latency	µSec	2.13

9 SW Components Implementations Status

1. Core IB support – GA
2. IPoIB-UD – GA, IPoIB-CM - beta
3. Voltaire MPI – GA
4. Open MPI – beta
5. RDS – beta
6. SDP – beta
7. Bonding driver – beta
8. GVD (GridVision Daemon) - GA
9. HIS (Host Identification Service) - GA

10 Prerequisites

1. One of the supported operating systems must be running on one of the supported CPU architectures.
2. Make sure the following packages are installed on your system prior to GridStack installation: kernel sources, kernel symbols, zlib, tcl-devel, pciutils, pciutils-devel, gcc packages, and libstdc++.
3. If you are using RHEL AS4 OS, make sure to install the sysfs-utils package.
4. Before installing the GridStack package, make sure to uninstall any previously installed InfiniBand stack from your system.

To remove previously installed InfiniBand packages:

- If the ibhost-3.5.x package is installed on your system, run the following shell command:

```
rpm -e `rpm -qa |grep ibhost`
```

- If GridStack-4.1.5_x or GridStack-4.3.XXX is installed on your system, run the following shell command:

```
/usr/voltaire/uninstall.sh
```

- If OFED-1.0 or OFED-1.1 is installed on your system, run the following shell command:

```
cd <ofed install dir> ; ./uninstall.sh
```

- If the OFED-1.2 package is installed on your system, run the following shell command:

```
ofed_uninstall.sh
```

11 Installation of the Voltaire GridStack^{TM} Package

The script ./install.sh is used to compile and install GridStack, version 4.3.5.c

The ./install.sh help (–h) lists all supported options, including the switches to be used with different compilers, Voltaire MPI compile options, and the switch --make-bin-package that creates the GridStack binary package.

11.1 Viewing the Installation Options

To view the install.sh options, run the following command:

```
./install.sh --help
```

The following details the install.sh options.

Option	Description
--gs-config <config-name>	Defines alternative configuration for GridStack. If not specified, the default configuration will be used. The default configuration file is located under: **GridStack-4.3.5.c-5/install/vlt-gs.conf.default** After installation, the GridStack configuration details are specified in **/usr/voltaire/gs.conf**
--ofed-config <config-name>	Defines alternative configuration for OFED. If not specified, the default configuration will be used. The default configuration file is located in **GridStack-4.3.5.c-5/install/ofed.conf.default** After the installation the GS configuration details are specified in **/usr/voltaire/ofed.conf**
--ipoib-config <conf-file>	Defines the IPoIB interface configuration file. See example for this configuration file in **GridStack-4.3.5.c-5/install/ipoib.conf**
--no-v-mpi	Prevents compilation and installation of Voltaire MPI
--no-o-mpi	Prevents compilation and installation of Open MPI
--v-mpi-compiler-ifc	Defines the path to Intel IFC compiler
--v-mpi-compiler-icc	Defines the path to Intel ICC compiler
--v-mpi-compiler-path	Defines the path to Pathscale compiler

Option	Description
--v-mpi-compiler-pgi	Defines the path to PGI compiler
--make-bin-package <full|slim|mini>	Creates a GridStack-4.3.5.c binary package. The created binary package name reflects the architecture (ARCH), distribution, kernel version, and the package type: full, slim, or mini. For example the file **../ GridStack-4.3.5.c_5-redhat-k2.6.9-55.EL-ppc64-full.tar.bz2** is created based on the following components: - RedHat Enterprise 4 Update 5 (distribution) - PPC64 (architecture) - k2.6.9-55.EL (kernel version) - full (package type) Please see explanation for packages types in the GridStack User Manual.
--custom <cust-name>	Builds a customized version
--dont-patch	Prevents applying patches to the source code
--unlock	Removes install lock
--fix-symvers	Fixes IB kernel symbol versions This option is relevant only for users that develop modules using GridStack modules.
--nocolor	Prevents printing of colored output
--set-timeofday	Synchronizes the system time and date according to the install.sh script
--no-stop	Prevents unload of running stack

11.2 Performing GridStack Installation

You can install the GridStack package either from the source files or from the binary package, as described below.

Compile and install GridStack-4.3.5.c from source using the following commands:

```
tar -zxf GridStack-4.3.5.c-5.tgz
cd  GridStack-4.3.5.c-5
./install.sh --ofed-config mlx4 <build and compile options>
```

Install GridStack-4.3.5.c from the binary package the using following commands:

```
tar -jxf  GridStack-4.3.5.c-5-redhat-k2.6.9-55.ELsmp-x86_64-
full.tar.bz2
cd  GridStack-4.3.5.c-5-redhat-k2.6.9-55.ELsmp-x86_64-full
./install.sh --ofed-config mlx4 <build and compile options>.
```

For information on how to install and configure the Voltaire GridStack, please refer to the Voltaire Linux GridStack for HCA 4X0 User Manual (listed in Section 3).

12 Limitations

- This package only supports ConnectX HCAs.
  For other types of HCAs you must install GridStack-4.3.0.
- Supports DHCP server with InfiniBand related changes [Internet Systems Consortium (ISC) DHCP Server V3.0.5] provided by Voltaire.

13 Patches Documentation

On top of the original OFED version 1.2.c, Voltaire added the following patches to this release:

Patch File Name (patches/Common/fixes/kernel/)	**Patch Description**
zzz_0010_mlx4_reset_msleep.patch	**Subject:** GridStack reset on ConnectX HCAs **Bug:** System hangs with some chipsets **Fix:** Configured a 500 msec delay after resetting the device before attempting to run config cycles on it.
zzz_0010_cm_sidr_1.patch	**Subject**: Duplicate SIDR REQs. **Bug**: The system sent reject messages if a duplicate was detected. **Fix:** Duplicates are now simply discarded.
zzz_0020_cm_sidr_2.patch	**Subject**: SIDR REQ not matching a listen. **Bug**: Dropped through to the default case of status 2 (Rejected by Service Provider). **Fix**: Replies with status value 1 (Service ID not Supported). This also fixes a bug where the cm_id_priv is removed from the remote_sidr_table twice.
zzz_0030_vlt_cma_tavor_quirk.diff	**Subject**: For Tavor based HCAs. **Goal**: To increase performance. **Fix**: Override the MTU that is returned by the SM in an answer to a Path Query with 1024.
zzz_0400_ib_find__partial_pkey.diff	**Subject**: P_Key lookup. **Goal**: For matching full and partial membership keys of the same partition. **Fix**: IPoIB sets the P_Key membership bit of limited membership P_Keys when creating a child interface. After that IPoIB looks for the full membership P_Key in the table to make the interface "RUNNING". This patch fixes the pkey lookup in order to match full and partial membership keys that belong to the same partition.

Patch File Name (patches/Common/fixes/kernel/)	**Patch Description**
zzz_0010-move-open-iscsi-crypto-functions-to-kernel_addons-R.patch & zzz_0020-open-iscsi-Change-scatterlist-len-in-crypto_digest_.patch	**Subject:** Using data digest in open-iscsi over TCP **Bug:** Scatterlist length for data digest was not calculated correctly. **Fix:** Scatterlist length should be always 1

Patch Name (patches/Common/fixes/user/)	**x86 with PCI-X**
zzz_0020_use_pf_sdp.diff	**Subject**: Fallback of SDP connection. **Goal**: Solve issue of libsdp not supporting a non-blocking connect. **Fix**: Since libsdp does not support a non-blocking connect in "both" mode, change the fallback of SDP connection from both to SDP.
zzz_0010_mstflint2.diff	**Subject**: mstflint screen output for PowerPC architectures. **Fix**: Fixed the screen output.

14 Known Issues

This section contains information on known limitations of the current version.

#	Subject	Description	Workaround
1.	Previous/Native InfiniBand modules backup	If a previous local installation of Open-IB is present on a host when installing the Voltaire GridStack$^{\text{TM}}$ package, it will be overridden by the new set of modules. When GridStack is removed, the original modules will become loadable again.	The command modinfo <module_name> tells which of the InfiniBand modules will be loaded when running modprobe (ro insmod). Voltaire modules are installed under /lib/modules/$(uname –r) /updates/kernel/drivers/infiniband
2.	Compiling 3rd party kernel modules that use InfiniBand symbols	The installation of GridStack does not update the file Module.symvers with the new signature of the InfiniBand symbols (the ones that come with the new InfiniBand kernel modules). The result is that kernel modules that use these symbols and compiled against the newer modules will not be usable since modprobe/insmod will fail loading them into the kernel.	install GridStack with the option --fix-symvers (Builder only)
3.	Using ib-bonding for Ethernet and InfiniBand together	When using bonding persistent configuration for InfiniBand slaves and having another bonding interface for Ethernet slaves, the operation of stopping GridStack cause all the bonding masters to go down, including the one that enslaves the Ethernet devices.	None

#	Subject	Description	Workaround
4.	Using bonding and GVD at the same time	If bonding uses child (P_Key) interfaces as slaves, then GVD auto-onfiguration of child interfaces might disturb the correct operation of ib-bonding.	Turn off GVD auto configuration by editing the file **/etc/gvd.conf**
5.	Bonding slaves and IP configuration	The bonding master interface should be configured with an IP address (via the ib-bond utility but not with a network script). So when using an IPoIB interface as a slave of a bonding interface, it MUST NOT have an IP address (i.e. configuration scripts with IP addresses).	None
6.	Limitations of memory pinning from user mode InfiniBand applications	Memory registration by user is limited according to the administrator settings.	Memory locking is managed by the kernel on a per-user basis. Regular users (as opposed to root users) have a limited number of pages that they may pin, where the limit is pre-set by the administrator. Registering memory for IB verbs requires pinning memory, thus an application cannot register more memory than it is allowed to pin. The user can change the system per-process memory lock limit by adding the following two lines to the file /etc/security/limits.conf: * soft memlock <number> * hard memlock <number> where <number> denotes the number of KBytes that may be locked by a user process.

#	Subject	Description	Workaround
			The above change to **/etc/security/limits.conf** will allow any user process in the system to lock up to <number> KBytes of memory. On some systems, it may be possible to use "unlimited" for the size to disable these limits entirely.
7.	ifconfig reports wrong HW address		linux:/root # ip address show dev ib0 5: ib0: <BROADCAST,MULTICAST,UP> mtu 2044 qdisc pfifo_fast qlen 128 link/infiniband 00:00:04:04:fe:80:00:00:00:00:00:00:00:08:f1:04:03:96:08:79 brd 00:ff:ff:ff:ff:12:40:1b:ff:ff:00:00:00:00:00:00:ff:ff:ff:ff inet 193.168.70.16/24 brd 193.168.70.255 scope global ib0 inet6 fe80::208:f104:396:879/64 scope link valid_lft forever preferred_lft forever
8.	Using fork() with InfiniBand applications	Using fork() in a program that uses InfiniBand is limited to the following conditions: 1. Parent process may continue running without any limitations on memory access 2. Child process gets a SISEGV signal (segmentation fault) when trying to access a memory that was registered by the parent	Fork support from kernel 2.6.16 and above is available provided: that applications do not use threads. The fork() is supported as long as the parent process does not run before the child exits or calls exec(). The former can be achieved by calling wait(childpid), and the latter can be achieved by application specific means. The Posix system() call is supported.

#	Subject	Description	Workaround
9.	Host vs. VSM HIS interoperability	A host with more than one InfiniBand interface may appear on switch tables with the same IP address for its different ports.	This should be fixed in future switch software releases.
10.	ib1 gets the same configuration as ib0 on SLES10	Under SLES10, when configuring ib0, ib1 gets the same configuration. This happens even if /etc/sysconfig/network/ifcfg-ib1 does not exist.	Create a network configuration file for ib1 even if it is not connected to the network. In process of fixing that with Novell.
11.	HCA IRQ line may block on AMD with old BIOS installed	HCA may stop functioning and Kernel log (printed by dmesg) may show: "irq 169: nobody cared (try booting with the "irqpoll" option)". Also, most IB commands (like ibv_devinfo) may cause the shell to hang. This seems to happen only on AMD machines with old BIOS.	Upgrade the BIOS or add the noirqdebug option to the kernel boot line (in grub.conf).
12.	SLES10 YAST does not support IPoIB configuration.	On SuSE Linux distribution, the YAST setup tool does not recognize the InfiniBand interface. Therefore, it cannot be used to configure the interface.	Edit the network script manually or use the supplied utility: ib-config-as-eth. In process of fixing that with Novell.

#	Subject	Description	Workaround
13.	SLES10:Ping Broadcast/Multicast blocked by default	Ping broadcast might not work because the kernel might block the stack to reply to the ping request.	To enable ping broadcast reply, run the following command from your shell: echo 0 > /proc/sys/net/ipv4/icmp_echo_ignore_broadcasts or edit /etc/sysctl.conf file and add the line: net.ipv4.icmp_echo_ignore_ broadcasts = 0
14.	On PowerPC: Removing RDS module after RDS traffic causes a kernel crash		Close applications using RDS before unloading the RDS module.
15.	uDAPL - Connection fail between client and server while changing the client data size and parameters on the next C/S session	The cause is that ib1on the servers is also up and replies to ARPs from the client. Sometimes port 2 replies to the ARP, and the client tries to connect to this port rather than to port 1 (ib0). This causes the error.	Change this by sysctl net.ipv4.conf.all.arp_ignore=1.
16.	RDS support is missing in SLES9 up3	RDS is not supported on sles9sp3.	

#	Subject	Description	Workaround
17.	Warning messages in kernel log after loading the bonding	After loading the bonding module the kernel log (viewed with dmesg) shows the following: bonding: Warning: either miimon or arp_interval and arp_ip_target module parameters must be specified, otherwise bonding will not detect link failures! See bonding.txt for details . bonding: bond0: Warning: failed to get speed and duplex from ib0, assumed to be 100Mb/sec and Full. bonding: bond0: Warning: The first slave device you specified does not support setting the MAC address. This bond MAC address would be that of the active slave.	These messages should be ignored.
18.	ib-config and ib-config-as-eth script does not check correctness of their input parameters	It is possible to pass illegal values to the scripts that will be accepted by them without warning or rejection.	

#	Subject	Description	Workaround
19.	Output warning during the running of some IB utilities	The warning looks as follows: libibverbs: Warning: RLIMIT_MEMLOCK is 32768 bytes .This will severely limit memory registrations May appear for non root users when running some InfiniBand utilities that use libibverbs.	Increase the maximum size that may be locked into memory by ulimit –l <value> This operation requires root permissions.
20.	FW image for HP and IBM Mezzanine cards	FW images for HP and IBM Mezzanine cards were created by Voltaire.	None
21.	Warning messages when burning firmware	When burning firmware, it is possible that the following message will appear during the operation: You are about to replace current PSID in the image file - "XXXXXX" with a different PSID - "YYYYYY." Note: It is highly recommended not to change the image PSID. And after that: You are about to replace current PSID on flash - "YYYYYY" with a different PSID - "HP_XXXXXX". Note: It is highly recommended not to change the PSID.	Ignore the recommendation and answer with 'y' to complete the operation.

#	Subject	Description	Workaround
22.	Burning firmware to an HCA that was not purchased from Voltaire	**ib-burn** identifies the type of HCA by reading an identification string from the HCA flash memory. If the HCA was not purchased from Voltaire then ib-burn will not be able to identify the HCA type and decide which firmware to burn.	The exact type of HCA should be passed to ib-burn (use option -i). Alternatively, firmware image file may be given to ib-burn (using option -a).
23.	IPoIB Interface order might change after adding an additional HCA	It is not guaranteed that a specific InfiniBand port will have the same name (ib0, ib1, etc.) after adding an HCA. This may cause, for example, a hardware address change for an IPoIB interface.	None
24.	Bonding module for Ethernet	The bonding module that comes with GridStack replaces the original bonding driver for Ethernet that comes with the OS. It is possible to use the new bonding driver for Ethernet interfaces but it was not tested to work with it.	None
25.	On SLES10SP1 ip multicast not being sent through ib1	If a sles10sp1 host tries to route traffic of some IP multicast subnet through interface ib1, the traffic will not be sent.	Bring ib0 down manually (ifdown ib0), or manually give it a different IP than the IP of ib1.
26.	Performance	Performance: TFTP and iSCSI data transport speed can be improved	Fixed version was provided after the QA start
27	IPoIB Connected Mode on IBM Blades (HS21)	On IBM blades (HS21) RHEL 5 & SLES10 (SP1). Heavy traffic over IPoIB Connected Mode may cause a system crash.	None
28.	Open MPI and ibutils are not supported under ppc64 with SLES10 sp1		None

#	Subject	Description	Workaround
29.	Open MPI does not support gen1 VAPI driver	Although the *mvapi* component exists in Open MPI, it is not being maintained for more than a year now.	None
30.	Open MPI CPU affinity and NUMA awareness	Open MPI does not support CPU affinity and NUMA awareness for memory allocation. We are currently working to add this to the Open MPI components.	None
31.	Open MPI threads safe	Current Open MPI 1.2 does not fully support threads safe.	None
32.	Open MPI fault tolerance	Open MPI in version 1.2 does not support “fault tolerance” or “check-point restart”	None
33.	Open MPI multi HCAs or multi ports	The support for nodes with multi HCAs (or ports) has only limited flexibility and we are currently working to add more features to the next version of Open MPI.	None
34.	Open MPI TCP and IB interconnect working together	**Do not** run Open MPI using both TCP and IB communication (in IB fabric use TCP for debug and problem isolation only).	None
35.	Open MPI progress thread	Open MPI currently does not support the use of progress thread for overlapping of computation and communication.	None
36.	Open MPI RDMA in collective operations	Open MPI currently does not use RDMA capabilities in collective operations.	None