



Date Effective: 12/12/06	Final Release Notes 85-0048-18	Approved:
Supersedes 85-0048-17	Version 4.1 QuickSilver Fabric Access Software for Linux	Page 1 of 14

Copyright © 2006 SilverStorm Technologies
ALL RIGHTS RESERVED

This document contains confidential and proprietary information. This information is and shall remain the property of SilverStorm Technologies and the document (and all copies thereof) must be returned to SilverStorm upon written request.

No part of this document may be distributed or reproduced in any form without prior written authorization from SilverStorm. This document is provided for evaluation purposes of SilverStorm products and services only. The information in this document is subject to change without notice and should not be construed as a commitment of any kind by SilverStorm.

SilverStorm assumes no responsibility for any errors or inaccuracies that may appear in the document. SilverStorm does not assume any liability that may occur due to use or application of the information provided herein.



Version 4.1 QuickSilver Fabric Access Software for Linux

Table of Contents

INTRODUCTION	4
PURPOSE	4
NOTES	4
INSTALLATION.....	5
SUPPORTED IN THIS RELEASE.....	5
Host Platforms	5
HCA Configurations.....	7
INSTALLATION REQUIREMENTS	7
Software and Firmware Requirements.....	7
INSTALLATION INSTRUCTIONS	8
POST-INSTALLATION TASKS	9
PRODUCT CONSTRAINTS	9
QUICKSILVER FABRIC ACCESS - SPECIFIC INFORMATION	12
FUNCTION AND OPERATION CHANGES	12
KNOWN PRODUCT LIMITATIONS	13



Version 4.1 QuickSilver Fabric Access Software for Linux

This page intentionally left blank.



Introduction

Purpose

The purpose of this document is to provide a description of Version 4.1 of the QuickSilver Fabric Access Software for Linux.

Note – There are references in this document to MPI, Scali MPI, and FastFabric. Those components are available only in the optional QuickSilver FastFabric Toolset.

Notes

Included in this release:

- QuickSilver Fabric Access™ firmware Version 4.1.0.2.2
- QuickSilver Fabric Access Drivers: Version 4.1.0.2.2
- IB Access: Version 4.1.0.2.2
- Discovery: Version 4.1.0.2.2
- Fast Fabric: Version 4.1.0.2.2
- IP over IB: Version 4.1.0.2.2
- MPI Version 4.1.0.2.2
- Virtual NIC: Version 4.1.0.2.2
- Virtual HBA: Version 4.1.0.2.2
- SDP: Version 4.1.0.2.2
- uDAPL: Version 4.1.0.2.2
- RDS: Version 4.1.0.2.2

NOTE: IP over IB in this release requires either a SilverStorm version 3.x subnet manage (SM) or a Cisco SM version BRHIBSM, Revision 0260 for the IBM BladeCenter. The SM must support IB multicast.



Installation

Supported in this Release

- PCI, PCI-X and PCI-Express
- System memory up to 64G is supported for all ULPs, up to 8G has been tested
- Environments running MPI require a minimum of 512M of Host DRAM, depending on application and number of MPI instances, larger memory may be required. A minimum of 1G per CPU is recommended.

Host Platforms

The following kernels are supported in this release for the given platforms:
AMD64 & EM64T

RHEL3: (Update 8) **2.4.21-47.EL** **2.4.21-47.ELsmp**

Note: In Redhat EL 3.0 there are different kernels for Opteron vs EM64T.
The appropriate release/kernel must be selected to match the CPU type.
Also the matching QuickSilver build must be selected for installation.

X86_64

RHEL4: (Update 3) **2.6.9-34.ELsmp**
RHEL4: (Update 4) **2.6.9-42.ELsmp**
RHEL4: (Current) **2.6.9-42.0.3.ELsmp**

SLES9 (SP3) **2.6.5-7.244-smp**
SLES9 (Current) **2.6.5-7.282-smp**

SLES10: (Initial) **2.6.16.21-0.8-smp**

IA32

RHEL3: (Update 8) **2.4.21-47.ELsmp** **2.4.21-47.ELhugemem**

RHEL4: (Update 3) **2.6.9-34.ELsmp**
RHEL4: (Update 4) **2.6.9-42.ELsmp**
RHEL4: (Current) **2.6.9-42.0.3.ELsmp**

SLES9: (SP3) **2.6.5-7.244-smp** **2.6.5-7.244-bigsmp**
SLES9: (Current) **2.6.5-7.282-smp** **2.6.5-7.282-bigsmp**

SLES10: (Initial) **2.6.16.21-0.8-smp**

- CPU model of Linux kernel can be identified by "uname -m" and /proc/cpuinfo as follows:



Version 4.1 QuickSilver Fabric Access Software for Linux

	Uname	/proc/cpuinfo
IA32:	i686	Intel Pentium 4 or later
EM64T:	x86_64	Intel CPUs
Opteron:	x86_64	AMD CPUs

Other combinations (such as i586 uname) are not supported at this time.

-

HCA Configurations

-
- The following models of the SilverStorm HCA 9000 PCI Express HCAs are supported:
 - Dual-port, low-profile, full-bracket PCIExpress SDR with memory (7104-HCA-128LPX)
 - Dual-port, low-profile, full-bracket PCIExpress DDR with memory (7104-HCA-128LPX-DDR)
 - Dual-port, low-profile, low-profile bracket PCIExpress 8x SDR with memory (7104-HCA-128LPXB)
 - Dual-port, low-profile, low-profile bracket PCIExpress 8x DDR with memory (7104-HCA-128LPXB-DDR)
 - Dual-port, full-bracket PCIExpress 8x SDR without memory (7104-HCA-LPX2P)
 - Dual-port, full-bracket PCIExpress 8x DDR without memory (7104-HCA-LPX2P-DDR)
 - Dual-port, low-profile bracket PCIExpress 8x SDR without memory (7104-HCA-LPX2PB)
 - Dual-port, low-profile bracket PCIExpress 8x DDR without memory (7104-HCA-LPX2PB-DDR)
 - Single-port, full-bracket PCIExpress 8x SDR without memory (7104-HCA-LPX1P)
 - Single-port, full-bracket PCIExpress 8x DDR without memory (7104-HCA-LPX1P-DDR)
 - Single-port, low-profile bracket PCIExpress SDR 8x without memory (7104-HCA-LPX1PB)
 - Single-port, low-profile bracket PCIExpress 8x DDR without memory (7104-HCA-LPX1PB-DDR)
- The following models of the SilverStorm HCA 7000 PCI-X HCAs are supported:
 - QuickSilver Model 7104-HCA-128LP PCI-X
- The following non-SilverStorm HCAs are supported:
 - Blade Center H Mezzanine Card HCA

Installation Requirements

Software and Firmware Requirements

QuickSilver Fabric Access™ firmware Version 4.1.0.2.2

QuickSilver Fabric Access Drivers: Version 4.1.0.2.2

IB Access: Version 4.1.0.2.2

Discovery: Version 4.1.0.2.2

Fast Fabric: Version 4.1.0.2.2

IP over IB: Version 4.1.0.2.2

MPI Version 4.1.0.2.2

Virtual NIC: Version 4.1.0.2.2

Virtual HBA: Version 4.1.0.2.2

SDP: Version 4.1.0.2.2



Version 4.1 QuickSilver Fabric Access Software for Linux

uDAPL: Version 4.1.0.2.2

RDS: Version 4.1.0.2.2

NOTE: All QuickSilver Fabric Access software on a given node must be at the same release level.

Installation Instructions

The `/etc/modules.conf` (`/etc/modprobe.conf` for 2.6 kernels) file now includes additional comments and samples. Unless you have special edits, it is recommended during the **INSTALL** that you select **n** at the "**Do you want to keep `/etc/modules.conf`?**" prompt.

NOTE: When using non-SilverStorm HCAs, HCA firmware update will require manual selection of the firmware file to use. Fast Fabric and command line **iba_config** or **INSTALL** options will not update HCA firmware. See the **README** files in `/etc/sysconfig/iba/mt*` for HCA identification instructions.

NOTE: On 2.6.10 and later kernels, after performing the QuickSilver install, a system reboot is required.

NOTE: If you are downgrading from a newer QuickSilver Fabric Access release to an older QuickSilver Fabric Access release, the newer QuickSilver Fabric Access release must be stopped and uninstalled, prior to installing the older QuickSilver Fabric Access release. The newer QuickSilver Fabric Access must be uninstalled using the **iba_config** tool or the **INSTALL** tool provided with the newer QuickSilver Fabric Access release.

NOTE: If running a 32-bit operating system on a X86_64 (AMD Opteron) system the i686 kernel should be installed. In most distributions the default install will be an Athlon kernel and the i686 kernel will need to be installed instead.

NOTE: 2.6 distributions no longer have `/etc/modules.conf`. The same attributes are contained in `/etc/modprobe.conf`.

NOTE: The **rmmod** and **insmod** commands are deprecated in 2.6 distributions. Use **modprobe** in all cases.

NOTE: The "IB Development" component also includes VAPI header files and libraries. The header files will be installed to `/usr/include/iba/vapi` and the libraries will be installed to `/lib` on 32-bit operating systems and `/lib64` on 64 bit operating systems.

NOTE: When performing HPL benchmarks it is recommended to obtain and use the latest **libgoto** BLAS math library. An older version of this library is included in

`/usr/local/src/mpi_apps/goto`. This should be updated with the latest version (licensing restrictions limit the redistribution of this newer library). The README in `/usr/local/src/mpi_apps/goto` provides more information and the web site for `libgoto`.

For detailed installation instructions, please see the *Fabric Access Software Users Guide*.

Post-Installation Tasks

None.

Product Constraints

- On 2.6.13 and later kernels, autostart for SDP should be disabled by using the `iba_config` command.
- When using the default MPI run command (`mpirun_rsh`) the user may need to modify the SSH configuration on each node to support enough incoming SSH connections. Edit `/etc/ssh/sshd_config` and ensure that **MaxStartups** is greater than the number of CPUs on the machine. Alternately, use the MPD job management tool instead of `mpirun_rsh`. MPD is recommended for clusters with more than 256 cpus in total.
- SDP Zcopy has been tested up to 77 connections. Bcopy has been tested up to 100 connections.
- By default, MPI is compiled for machines with no more than 32 CPUs. To use MPI on larger systems, install the MPI source and edit `/usr/local/src/InfiniServ/mpich/mpid/vapi/mpid_smpi.h`. Change the `SMPI_MAX_NUMLOCALNODES` definition, then rebuild the MPI source with the `do_build` script.
- All nodes running a single instance of an MPI application must be running the same release of Fabric Access software.
- To fix an alignment bug a minor change has been made to the `ib_sd.h` header file which affects the interface to the Subnet Data interface of `IbAccess`. This change is fully source compatible with previous releases. However existing applications/drivers which use the Subnet Data interface to `IbAccess` directly will need to be recompiled. Note the Subnet Data interface is the API for making direct SA queries to the SM/SA. This change does not affect applications using `IPoIB`, `uDAPL` or other mechanisms to resolve paths in the fabric.
- When running Oracle 10g (or `crload`) using `DAPL`, the `/etc/modules.conf` must be edited to enable the "options ibt" and "options mt23108vpd" lines indicated for Oracle. If you intend to run oracle, during the install you should select to `n` at the "Do you want to keep `/etc/modules.conf`?" prompt. The recommended `modules.conf` settings for Oracle 10g are:
 - `options ibt CmMaxDgrms=15000 CmReuseQP=1 GsiMinRecvPerPort=1000 GsiPreallocRecvPerPort=1200 CmMinTurnaroundTime=1000`
 - `options mt23108vpd MaxVLs=5 use_mellanox_locking=1`

- When running more than 2000 IB connections on a node (for example a 256 process MPI job with 8 processes per node), the CmMaxDgrms value must be increased on the options ibt line. For example:

```
options ibt CmMaxDgrms=3000
```

- IPoIB performance can depend on the application used. The default value for IpoibRecvBuffers has been changed from 50 to 2000. This provides better performance for interconnect intensive applications, such as Oracle and Lustre over IPoIB.
- When running more than 2000 IB connections on a node (for example a 256 process MPI job with 8 processes per node), the CmMaxDgrms value must be increased on the options ibt line. For example:
 - `options ibt CmMaxDgrms=3000`

-

- The following limitations exist with fork when using InfiniBand.
 1. The child process will have holes in his address space for all memory in the parent that is registered with the HCA. If the child attempts to access this memory, he will receive a segmentation violation. All memory registration is on page boundaries, so it is advised that the application page align all memory regions. Otherwise, the child will may lose access to other data structures which reside in the same page as a memory region.
 2. The child process will not be able to exit. If the child attempts to exit, he will receive a segmentation violation. The child can exec without issues.
- A disk array may return a medium error which the scsi mid-layer in certain kernels does not handle gracefully. It will eventually right itself and issue a message that the disk cannot be read. This is what could be expected to be returned from a device which has never been initialized or performing a fdisk on a device which is not valid for this purpose (e.g. /dev/sdf1 instead of /dev/sdf) and can therefore not be read from correctly. The sector requested is not valid in either case from the hardware devices perspective.

For a disk that has never been written, an fdisk /dev/<base devname> should be issued to write a correct partition table.

Avoid issuing a fdisk to a device which cannot be partitioned (e.g./dev/sdf1).

- When running older versions of LSTC LS-DYNA (e.g., those that are linked against mpich archive libraries), the parameter **VIADDEV_INLINE_ADJUST=8** must be added to the **mpi.param** file. This works around a bug in this older mpich library (which caused LS-DYNA to fail with an Invalid Parameter error in **VAPI_post_sr** in **viapriv.c** line 607). The parameter is not necessary when running against applications linked with the **libmpich.so** shared library.
- SuSe 2.6 releases seem to occasionally have issues loading **ics_sdp** and RDS. Those releases have a parallel startup/shutdown process that may cause either SDP or RDS to get an "Unknown symbol" error for symbols supplied in **ics_offload**. Detours include:
 - Do not autostart rds
 - For SLES9 load **module-init-tools-3.0_pre10-37.32.x86_64.rpm**. This version has a locking fixed that may solve the problem.
 - edit **/etc/sysconfig/boot** and change the **RUN_PARALLEL** attribute to **no**.
- SuSE Linux Enterprise Server 10 (SLES 10) has the following limitations:



Version 4.1 QuickSilver Fabric Access Software for Linux

-
- udapl running on SLES 10 is a Beta feature in this release
- SDP running on SLES 10 is a Beta feature in this release
- Scali running on SLES 10 is a Beta feature in this release



QuickSilver Fabric Access - Specific Information

Function and Operation Changes

- On Opteron and EM64T systems with 64 bit OSs, QuickSilver Fabric Access Software libraries are now installed to /lib64. MPI libraries are installed to /usr/local/lib64. Customer makefiles may need to be adjusted accordingly. Makefiles already using mpicc, mpif77, etc will need no changes. On IA32 systems, the libraries are still installed into /lib and /usr/local/lib.

■

Known Product Limitations

This section describes the known product limitations for this release.

- The libgoto BLAS library included in the MPI sample applications for used by HPL does not support arbitrary combinations of OS and CPU types. It has been observed that mixing Redhat ELS3 with other linux distributions will not reliably work. These effects are due to optimizations in libgoto which may change the on wire MPI protocol used by libgoto for different CPU and OS types.
- Some versions of MVAPICH2 may fail with errors in modify QP to RTR. To work around these errors the mt23108vpd driver can be configured with ResponderResources=8. The issue is that MVAPICH2 hardcodes QP parameters and does not check the HCA's supported parameters.
- MPD in mpich2 does not interoperate with RHEL3. When this combination is run, MPD does not always detect process termination, so MPD does not release the console. To resolve this issue, it is recommended to use a newer OS release, such as RHEL4.
- When building mpi applications, you must not explicitly link mpich or libc. Instead, use mpicc, mpif77 or mpif90 to link your application. These commands will include mpich and libc and Infiniband libraries in the proper order.
- **Unable to force DDR HCA to run SDR link**

Description: The **iba_portconfig** command permits the port parameters to be changed. This can be used to enable/disable various speeds and link widths. On a SilverStorm 9000 SDR memory-free HCA running the latest 3.3 software, it was observed the **iba_portconfig** could properly change the HCA link from 4x to 1x and back. In each case the link retrained and became Active. However, when the same operations (both link width and link speed changes) were attempted on a SilverStorm 9000 DDR HCA, the link stayed in down/polling and never came back up.

For customers with a DDR HCA that have either a cable that is too long or an unreliable cable, it is desirable to be able to force the link to SDR.

Impact/Workaround: For customers with a DDR HCA that have either a cable that is too long or an unreliable cable, it is desirable to be able to force the link to SDR. They are unable to do this.
- **The output of 'cat /proc/driver/ics_srp/driver' does not display all IOCs in a large fabric**

Description: Issuing 'cat /proc/driver/ics_srp/driver' on a host in a fabric with many srp IOCs causes some of the output to be truncated. In some cases IOCs are not displayed.



Version 4.1 QuickSilver Fabric Access Software for Linux

- *The output of '/proc/driver/ics_srp/roundrobin_0' does not display all IOCs in a large fabric*
Description: The output of the commands '/proc/driver/ics_srp/roundrobin_0' or '/proc/driver/ics_srp/failover_0' is truncated when there are more than 4 entries in the file.
- *Segmentation fault with iba_capture when dumping a SDP proc file*
Description: Executed iba_capture on Host26, which is a SLES10, 32-bit host. The command failed. The host produced a corrupted archive, even after a reboot.
Impact/Workaround: Disable SDP autostart on SLES10 systems by using the iba_config command.