Xsigo Host Drivers
and
WinIB Stack Change Proposal

Hal Rosenstock
James Yang

Agenda

- Host Drivers
  - VNIC
  - VHBA
  - Linux Architecture
    - vHBA
    - TLAs
    - Xsigo Core (xscore)
      - Message Sequencing
  - Status, OFED 1.4 Plan, and Beyond
- WinIB Stack Changes - Proposal for changing IBbus driver
  - Purpose
  - Flexible Child PDO
  - Build Issues
  - Special Mode Support
  - Concerns

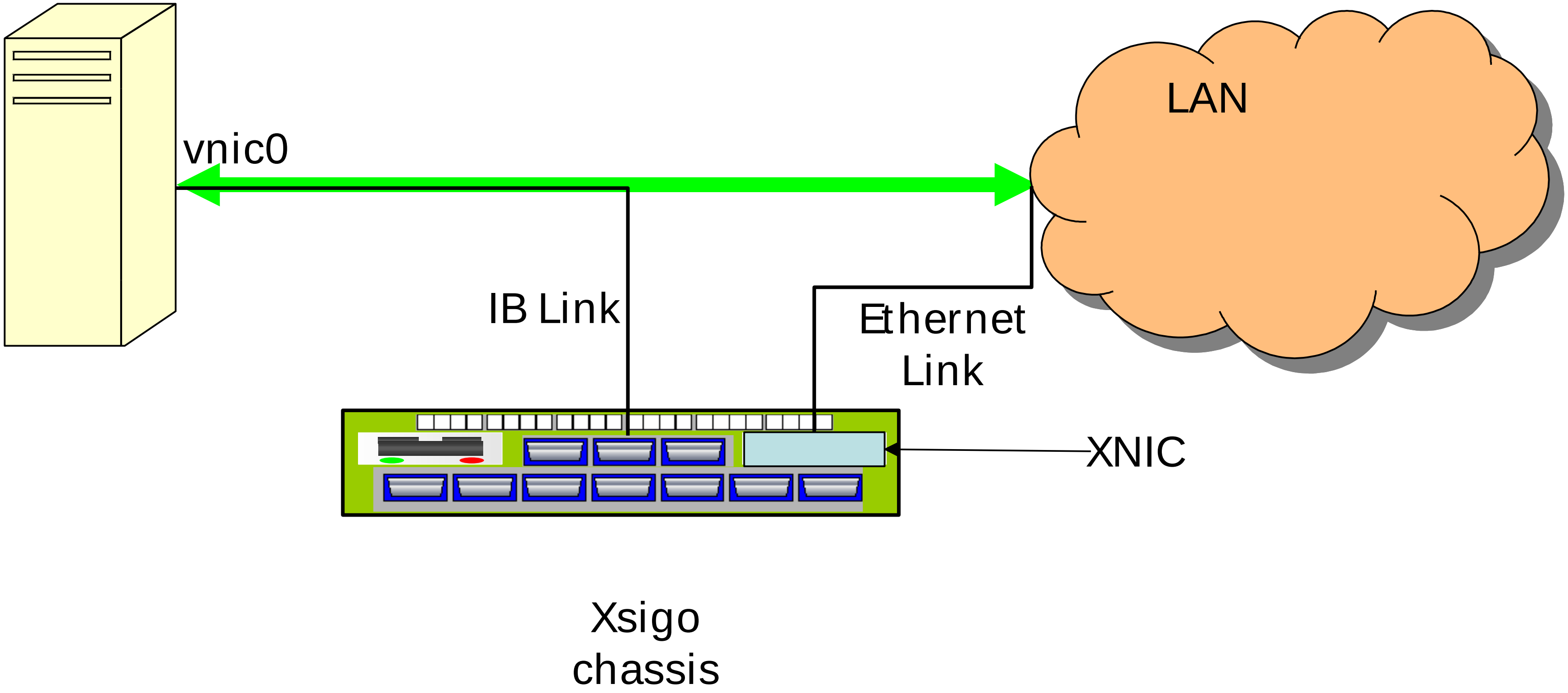
Host Drivers

- V* drivers
  - Kernel drivers for virtual adapters
    - VNIC
    - VHBA
    - Others possible
- VNIC
  - Virtual NIC
- VHBA
  - Virtual HBA
- OpenFabrics Environments
  - Linux
  - Windows

VNIC

- Host device driver
  - Kernel module in Linux
- Presents various ethernet iocards in Xsigo IO directors as “true” ethernet device in host
  - IB fabric is “IO bus” for transfer of control and data
    - Separate RC connections

VNIC



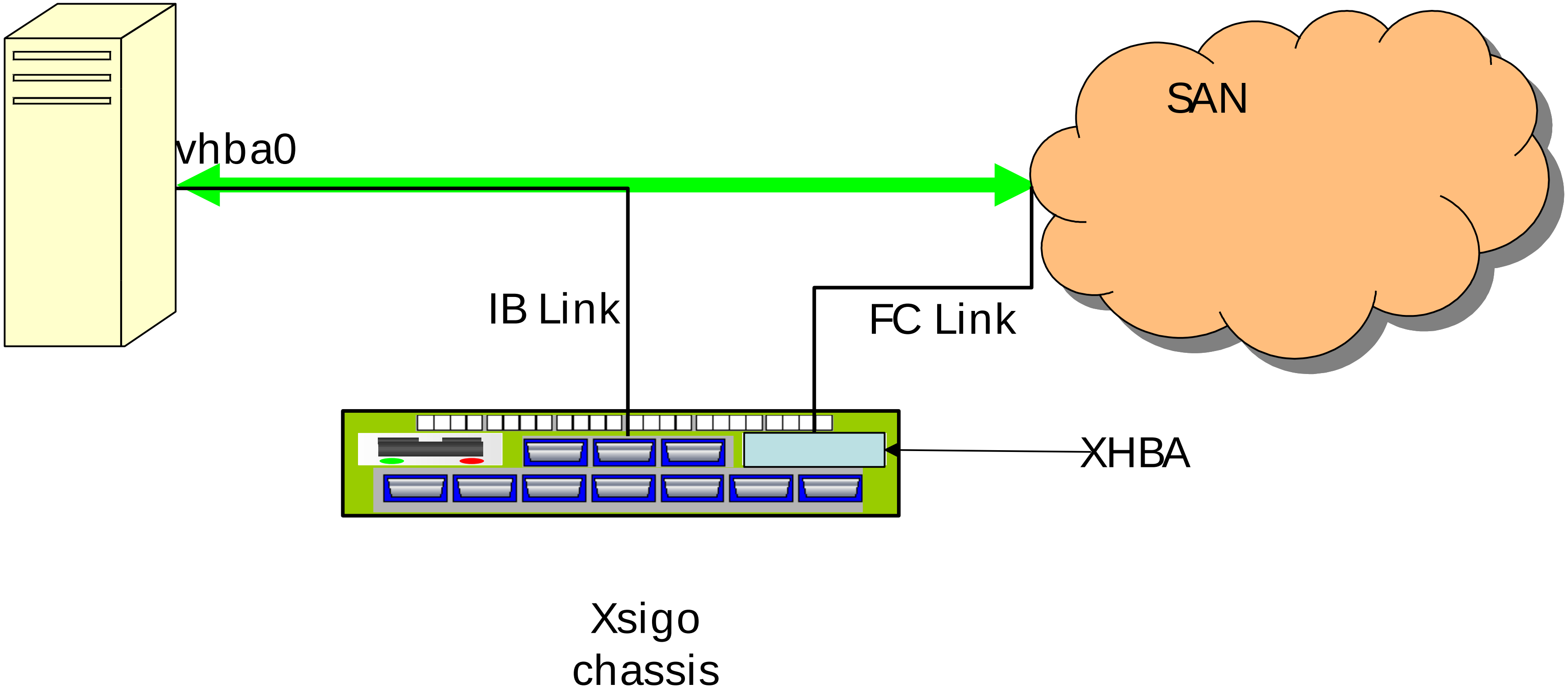
VNIC Operation

- User loads the xsvnic kernel driver
- User configuration/ discovery of virtual ethernet interface mapping of interface name in host to iocard/ port
- User can configure the virtual ethernet interface like any other interface
  - ifconfig, vconfig for VLANs, etc.
  - Supports IPv4 and IPv6 protocols
- The virtual ethernet interface operates like any other ethernet interface
  - ping, ssh, scp, netperf, ...

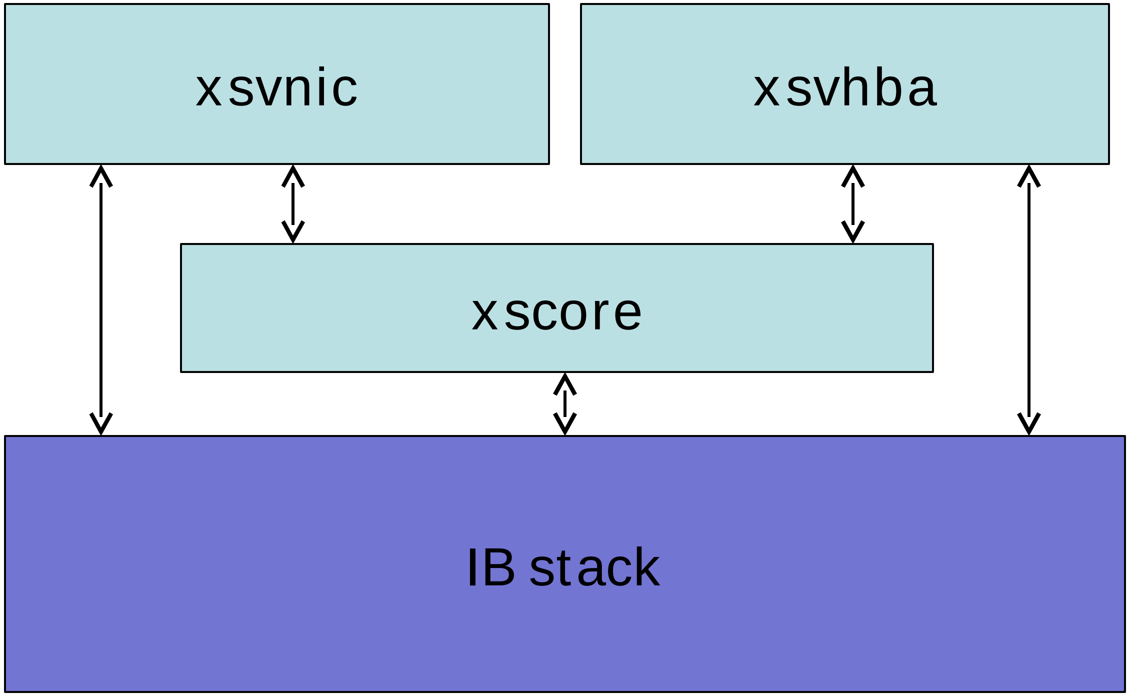
VHBA

- Host device driver
  - Kernel module in Linux
  - SCSI initiator
- Presents various FC iocards as true HBA device in host
  - IB fabric is “IO bus” for transfer of control and data
    - Separate RC connections
      - Data connection uses RDMA

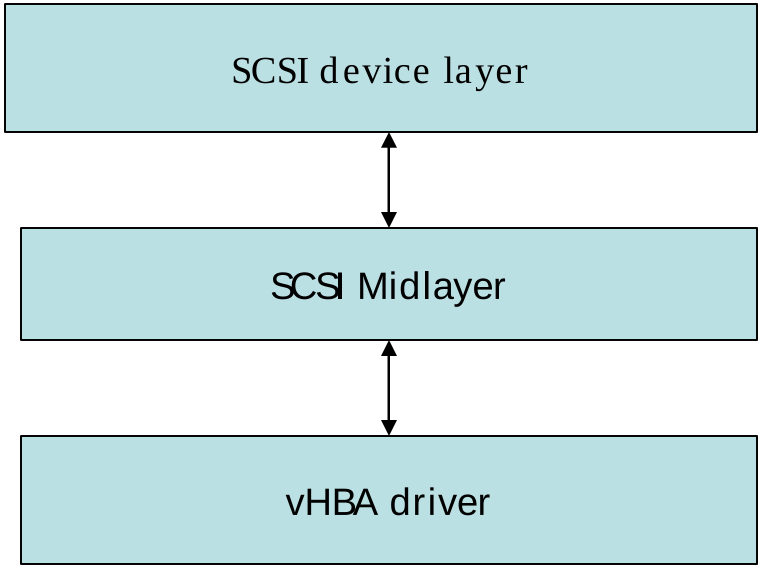
VHBA



Linux Architecture



vHBA



SCSI device commands/ data

LUN mgmt, queuing, error recovery

IOCB/ data

TLAs

- XDS – Xsigo Directory Services
- XCPM – Xsigo Configuration Protocol
- XSMP – Xsigo Session Management Protocol
- XCM – Xsigo Configuration Manager

Xsigo Core (xscore)

- Common services for v* drivers
  - XDS, XCPM, etc.
  - XSMP
    - Control messaging
      - RC based
  - Xsigo IB
    - Shared connection management
    - Common IB access

Message sequencing for xscore

- Server queries SA for XDS (via ServiceRecord)
- Server queries XDS for XCMs via vendor MADs
- XDS returns list of XCMs assigned to the server on this port
- Server now creates "links" for this port
  - A link is an XSMP session with an XCM
- Server obtains configuration via XSMP

Status, OFED 1.4 Plan, and Beyond

- Currently undergoing testing at customer sites
  - x86 and x86_64 (Opteron) architectures
  - RHEL4, RHEL5, SLES10, and some other 2.6.x kernel variants
- Initial RFC sent on ewg
  - 2.6.25-rc8 based
- Experimental/technology preview
- Xsigo core and vnic for Linux for OFED 1.4
  - xscore
  - xsvnic
- To upstream kernel later in OFED 1.4 cycle
- Other drivers to follow beyond OFED 1.4
  - xsvhba
  - Windows

WinIB Stack Changes

- Proposal for changing IBbus driver

Purpose

- Find a way to provide the flexibility to create different child for different usage model
  - Current IBBus driver creates only one fixed child IPoIB
- Build issues
  - Prefast error
- Support crashdump/ hibernation/ boot mode

Flexible child PDO

- Use INF file to read the child setting/IDs, example:

```
;HKR,"Parameters","StaticChild",%REG_SZ%,"IPoIB"
HKR,"Parameters","StaticChild",%REG_SZ%,"XsigoBus"

HKR,"Parameters\IPoIB","DeviceId",%REG_SZ%,"IBA\IPoIB"
HKR,"Parameters\IPoIB","CompatibleId",%REG_SZ%,"IBA\SID_1000066a00020000"
HKR,"Parameters\IPoIB","HardwareId",%REG_SZ%,"IBA\IPoIB"
HKR,"Parameters\IPoIB","Description",%REG_SZ%,"OpenIB IPoIB Adapter"

HKR,"Parameters\XsigoBus","DeviceId",%REG_SZ%,"IBA\XsigoBus"
HKR,"Parameters\XsigoBus","CompatibleId",%REG_SZ%,"IBA\SID_0000000002139702"
HKR,"Parameters\XsigoBus","HardwareId",%REG_SZ%,"IBA\XsigoBus"
HKR,"Parameters\XsigoBus","Description",%REG_SZ%,"Xsigo Virtual Bus"
```

Flexible child PDO

- File changes:
  - bus_driver.c
  - bus_port_mgr.c
  - bus_port_mgr.h
  - ibbus.inf

Build Issues

- Use #pragma to suppress warnings

```
   // Supressing Prefast Warning details are:
   // warning 8103 : Leaking the resource stored in 'SpinLock:p_spinlock->lock'.
   // Path includes 7 statements on the following lines:
   #pragma prefast(suppress:8103, "Suprressing next line for Prefast warning for
reason mentioned above in comments")
KeAcquireSpinLockAtDpcLevel( &p_spinlock->lock );
```

- Notation xxULL for 64 bit constant won’t work for Prefast

```
CL_HTON64( 0xFFFFFFFFFFFF0000ULL ) ) == CL_HTON64(
0xFEC0000000000000ULL ) );   →

CL_HTON64( CL_CONST64( 0xFFFFFFFFFFFF0000 ) ) ) == CL_HTON64(
CL_CONST64( 0xFEC0000000000000 ) ) );
```

Special Mode Support

- Crashdump
  - Can’t call spinlock for crashdump at high IRQL
  - Change hw/mthca/kernel/mt_spinlock.h to skip spinlock when in IRQL > DISPATCH
- Boot device
  - Change mthca.inf to make it boot device

Concerns

- Can IB stack support all different children?
  - Boot device hba path?
  - Hibernation support?
  - Text mode setup support?

Thank You