N_Port_ID Virtualization

Functional Specification

Version 1.0

1 Project Description

N_Port_ID Virtualization (NPIV) allows one FibreChannel (FC) Port to represent many physical ports, thus virtualizing the hardware. System administrators can use their existing management schemes such as LUN Masking and Switch Zoning to perform access control against these logical ports. NPIV is especially useful for virtual machine environments such as Xen or Solaris Logical Domains.

1.1 Definition

This project has several different modules. At the lowest level, the Solaris FibreChannel device driver stack (a.k.a. Leadville) must be changed to support NPIV. Specifically, the FCA (FibreChannel Adapter) drivers, which actually interface to the hardware must be changed. The interface between the FCA driver, and the generic FibreChannel port drivers (fp/fctl) must also be changed to support NPIV.

Further up, the HBA-API will be changed to manage NPIV. HBA-API is an industry standard interface designed to report status of the HBA and attached devices. Although the HBA API will be changed to support NPIV, the standard is immature, so we will be implementing a proprietary extensions, to be replaced when the standard is mature.

We will change the fcinfo(1M) utility to support NPIV. The utility will have new administration options to create and delete NPIV ports; this utility will be called fcadm, but implemented as a hardlink to fcinfo.

Containers, Xen and Logical Domains will all be supported. For each guest operating system, we will optionally map an NPIV WWN. If a volume appears on that NPIV WWN, it will be automatically mapped to the appropriate guest operating system.

Although we are planning to support Logical Domains with NPIV, with similar concepts, this work will be put back later. We will file a new PSARC case at that time. This two phase approach is due to the missing ability to dynamically add block devices to Logical Domains, which is already on the roadmap for Logical Domains, as well as resource constraints in the NPIV implementation team.

1.2 Motivation, Goals, and Requirements

Target customers are those customers deploying relatively complex SANs and use Zoning and/or LUN Masking. They may also use Solaris virtual machine technology such as Xen, Logical Domains, or Containers. Using NPIV allows customers to use existing Zoning and LUN Masking techniques.

1.3 Changes From the Previous Release

This project increases functionality of existing products. This project is a backwards compatible change to existing functionality, and should be considered a minor release.

1.4 Program Plan Overview

• 1.4.1 Development

Development is ongoing in Beijing. Many areas of the design, including FibreChannel device driver modifications, HBA API changes, and fcadm configuration tool have been prototyped and design is complete. Other areas, such as Xen and Containers have been partially prototyped.

Two major areas are incomplete in this specification, but will be completed by commitment review:

- Xen Migration support for NPIV
- Dynamically adding block devices to Zones

1.4.2 Quality Assurance/Testing

- Testing will consist of functionality, both without virtualization, but also in xVM and Container environment.
- Additionally, we will perform basic I/O performance tests, as well as SAN reconfiguration tests. We will test the maximum number of NPIV addresses on an HBA.
- We will test interoperability with Sun's switches which support NPIV. We will test against both of Sun's HBAs.

1.4.3 Documentation

This project requires updates to man pages as listed in the interface documents. It will require updates to manuals for FibreChannel, Xen and Containers, specifically:

- Solaris Fibre Channel and Storage Multipathing Administration Guide
- System Administration Guide: Solaris Containers-Resource Management and Solaris Zones
- xVM: TBD

1.4.4 Release Cycle

We plan to do an alpha release of our driver on Opensolaris approximately December

2007. We plan put back Q3 FY08.

1.4.5 Technical Support

TBD

1.4.6 Training

TBD

1.5 Related Projects

NPIV is an addition to the Solaris FibreChannel stack, commonly known as Leadville. It modifies the fcinfo CLI.

1.5.1 Dependencies on Other Sun Projects

Existing completed projects:

PSARC 1997/385 Fibre Channel Driver Re-architecture (Leadville)

PSARC 2004/291 Fibre Channel HBA Port Utility

1.5.2 Dependencies on Non-Sun Projects

None

1.5.3 Sun Projects Depending on this Project

None

1.5.4 Projects Rendered Obsolete by this Project

None

1.5.5 Related Active Projects [Describe the relationship.]

None

1.5.6 Suggested Projects to Enhance this Program

Integrate SCSI passthrough for Xen. SCSI pass through allows SCSI devices to be presented as SCSI devices to a DOMU (not block devices, as this project supports). This is especially valuable for non disk SCSI devices such as tape drives.

Fujitsu is working on this support for Linux/Xen, and it will likely be included in Xen 3.2 release. NPIV is complimentary because one NPIV port's devices can be mapped to a DOMU.

1.6 Competitive Analysis

The other major virtualization environments (Xen/Linux, Microsoft Virtualization and VMWare) are all implementing NPIV. At this point, limited is support is available from the HBA vendors with special drivers, although we anticipate that support for VMWare, and likely Microsoft virtualization will be included in the operating system. We are working with members of the Xen community on the operating system independent portion of NPIV support.

2 Technical Description

2.1 Architecture

The NPIV project encompasses many different layers. This is summary; details are included in section 4 and beyond.

- At the lowest level level is the Leadville device driver. The change to the Leadville device driver is relatively simple. The current architecture of the Leadville stack (PSARC 1997/385) encompasses several device drivers. At the lowest level is the FCA (or FibreChannel Adapter) which actually communicates with the FibreChannel Adapter hardware/firmware. It is able to send/receive encapsulated FibreChannel commands. The FCA driver communicates with the fp and fctl drivers. The fp driver currently has one instance per physical and is responsible for per port handling (for example reconfiguration). There is one fctl instance per system and it is responsible for maintaining knowledge of all ports and upper layer protocols.

  The architectural change for adding NPIV is to change the current one to one relationship between the FCA instance and fp instance to a one to many relationship.
- Currently Leadville is managed using the HBA API for status reporting as well as Solaris specific project private interface. There is ongoing work in T11, the standards body responsible for HBA API, to update the interface to support NPIV; however this is still at a conceptual phase. For now, we will implement proprietary, backward compatible extensions to the HBA API to support NPIV.
- Currently we have a CLI called fcinfo which is used to report status of HBAs and devices attached to the HBA. We will extend this tool to support NPIV. We will create a hardlink to the tool called fcadm, used for administering NPIV. This change will not break compatibility with fcinfo.
- For each Virtual machine environment (xVM and Containers) we will do the necessary work to bind a NPIV port to a guest operating instance. In the case of xVM, we will migrate a DOMU's WWN when the DOMU migrates.
- We will not support Logical Domains in the initial release but will support them in the future.

2.2 Interfaces

2.2.1 Exported Interfaces

Interface Name	Proposed Stability Classification	Specified in What Document?	Former Stability Classification or Other Comments
FCA	Project private	NPIV_FCA_Interface_Doc-0.2.pdf	Base document is specified FibreChannel Case
Proprietary HBA-API extensions for NPIV	Project Private	HBA-API_ext.man	When standard NPIV extensions to HBA API are approved, this will be depricated
fcadm command line	Committed	fcadm.man	
fcadm output	Uncommitted	fcadm.man	
fcinfo output	Uncommitted	fcadm.man	
/etc/cfg/fp/NPIV format	Project Private		Storage for NPIV WWN
xenstore	Volatile	Xenstore.man	Changes for existing Xen database
xm input/output	Volatile	xm.man	Xen CLI
/usr/lib/xennpivd file	Project Private		Xen reconfiguration daemon
xnpivd file	Project Private		Xen reconfiguration driver

zonecfg input	Uncommitted	zonecfg.man	Zone configuration

2.2.2 Imported Interfaces

Interface Name	Proposed Stability Classification	Specified in What Document?	Former Stability Classification or Other Comments
T11 NPIV Specs	Standard	T11 web site	See Appendix A for exact titles
libhbaapi(3LIB)	Standard	Solaris Man pages	Derived from T11 FC-MI
Libxml	Volatile	Solaris Man pages	Used for storing WWN configuration
libzonecfg	Contracted Project Private		Necessary for integrating NPIV support with Zones

2.3 User Interface

See section 4 and beyond for details of user interface.

There will be several extensions to existing CLI user interfaces, and one new user interface:

- fcadm will allow creating and deleting NPIV ports. fcadm is intended to be used in non virtualized environments. Fcadm shares functionality with fcinfo and will likely be implmented as a hardlink to fcinfo.
- We will extend the fcinfo command to support NPIV.
- We will extend the Xen xm command to support NPIV. This will support creating/deleting and reporting status NPIV virtual ports associated with a Xen DOMU.
- We will extend the Container zonecfg command to support NPIV.

2.4 Compatibility and Interoperability

2.4.1 Standards Conformance

We are implementing N_Port_ID Virtualization. NPIV is a T11 standard specified in T11 FC-LS/FC-DA/FC-GS-4. See Appendix A for full titles.

2.4.2 Operating System and Platform Compatibility

Solaris Nevada on both X86 and SPARC platforms will be supported. Both Emulex and Qlogic HBAs will be supported. Both Emulex and Qlogic support NPIV in their 4Gb products; previous products (1Gb/S and 2Gb/S) will not support NPIV.

SCSI 3 PGR (persistent group reservations). NPIV will impact PGRs – a reservation will be based on NPIV WWN. However, this should be beneficial for users; today a guest operating system using SCSI 3 PGR will share the device with any other guest operating system sharing that HBA; with NPIV using a dedicated WWN, the device will only be dedicated to that affiliated guest operating system.

2.4.3 Interoperability with Sun Projects/Products

On the SAN, NPIV will interoperate with 4Gb/S FibreChannel switches, and all target devices.

- NPIV will support Xen, Logical Domains and Containers (as well as being usable without virtualization).

2.4.4 Interoperability with External Products

NPIV will interoperate with any Fibrechannel switch which supports NPIV – generally vendors' 4Gb/S products support NPIV. If a user attempts to use an NPIV capable HBA with a switch that is not capable, the HBA will function with one port, and no virtual ports. Legacy HBAs (2 Gb/S and 1Gb/S) will continue to operate in non NPIV mode.

- All FibreChannel targets will transparently support NPIV HBAs (NPIV functionality will be invisible to targets).

2.4.5 Coexistence with Similar Functionality

No known overlap in functionality.

Today, for Emulex and Qlogic HBAs, end users may select SCSA architecture ("native") drivers supplied by the respective vendors, or use Leadville. Emulex and Qlogic have agreed not to supply SCSA architecture drivers supporting NPIV for Solaris Nevada. They may support NPIV in Solaris 10 and prior releases. We have discussed ways of providing transparent upgrade to Solaris Nevada.

2.4.6 Support for Multiple Concurrent Instances

Only one instance of FibreChannel device drivers may be present on the machine. This includes Logical Domains, Containers and Xen.

2.4.7 Compatibility with Earlier and Future Releases

This is backwards compatible with all existing interfaces.

2.5 Performance and Scalability

2.5.1 Performance Goals

We expect runtime CPU requirements for adding one NPIV port to be similar to adding an additional HBA. This CPU load will be principally felt at reconfiguration time, when events occur on the SAN, and Leadville must handle reconfiguration.

2.5.2 Performance Measurement

TBD.

2.5.3 Scalability Limits and Potential Bottlenecks

HBA Vendors have hardware limitations due to their HBA's physical resource limitations (on board memory). This limits the number of active NPIV ports bound to a physical HBA. This number is fairly large (~100), and varies with each model of HBA.

2.5.4 Static System Behavior

Database configuration file will be small, approximately 100 bytes per NPIV WWN.

2.5.5 Dynamic System Behavior

System behavior will be similar to non NPIV case, with the exception of reconfiguration noted above. During reconfiguration, every virtual HBA will perform reconfiguration. This is by design.

2.6 Failure and Recovery

2.6.1 Resource Exhaustion

If the HBA runs out of hardware resources to support NPIV, this will be reported when the port is attempted to be created.

2.6.2 Software Failures

We don't have planned software failures.

2.6.3 Network Failures

SAN Failures are managed as today. Cables failures between the adapter and switch are

handled by disabling the link. Failures between switch and device create RSCN (State Change Notification) messages to the NPIV port. The NPIV port can then act on the message by determining the nature of change.

2.6.4 Data Integrity

n.a.

2.6.5 State and Checkpointing

n.a.

2.6.6 Fault Detection

n.a.

2.6.7 Fault Recovery (or Cleanup after Failure)

n.a.

2.7 Security

NPIV is an important part of controlling access to devices on the SAN. Storage Administrators use it to segregate devices on the SAN. However, it's not intended as a secure protocol. FC-SP (FibreChannel Security Protocol, a part of the T11 FibreChanel standard) is intended as a secure protocol. FC-SP includes authentication of nodes and encryption. FC-SP is not widely implemented in the industry and is not implemented in the Solaris FibreChannel stack.

With NPIV, it is possible to do spoofing of existing devices on the SAN. However, this capability is already present in non NPIV device drivers for Linux and possibly other operating systems. Other NPIV implementations will also allow spoofing NPIV addresses.

In short, although NPIV can allow the user to do spoofing, this capability is already a problem on the SAN. Although there is an industry standard addressing this problem, there has been insufficient demand to implement the solution.

2.8 Software Engineering and Usability

2.8.1 Namespace Management

We introduce no new packages or names.

2.8.2 Dependencies on non-Standard System Interfaces

No non standard/stable interfaces are being used.

• 2.8.3 Year 2000 Compliance

NPIV does not use dates or time, so is Y2K compliant.

2.8.4 Internationalization (I18N)

No change in existing levels of I1ILN. fcadm will follow same conventions as fcinfo.

2.8.5 64-bit Issues

Existing FibreChannel device drivers are 64 bit clean and will remain this way.

2.8.6 Porting to other Platforms

Some portions of Xen code are platform independent, and needs to be portable to other operating systems (e.g. Linux). Most of this project however is specific to Solaris and would require significant effort to port.

2.8.7 Accessibility

No new interfaces, just additions to existing interfaces. All interfaces are CLIs.

3 Release Information

Note: Some of the packaging and installation details may not be available at design time. Describe expected solutions, and augment the description as the details are decided.

3.1 Product Packaging

Bundled with Solaris

3.1.1 Package Overview

No new packages will be created. The fcadm command will be added to SUNWfcprt package. This package currently contains fcinfo.

New Xen files, xennpivd and xnpivd changes will be in existing package SUNWxpvu.

3.1.2 (Default) Installation Locations

NPIV configuration file in /etc/cfg/fp/NPIV. Note /etc/cfg/fp is currently used for cfgadm_fp(1M).

Fcadm will be installed in /usr/sbin/
Xen NPIV deamon will be installed in /usr/svc/method/
Xen NPIV device driver will be installed in /platform/i86xpv/kernel/ directory structure.

3.1.3 Effect on External Environment

None

3.2 Installation

3.2.1 Installation procedure

Installed during standard Solaris installation. Unchanged from current.

3.2.2 Effects on System Files

None

3.2.3 Boot-Time Requirements

fcadm will be called at boot time (from SMF) to enable NPIV ports. Enabling NPIV ports may slow boot time, again similarly to adding additional physical HBAs.

3.2.4 Licensing

Standard Solaris/OpenSolaris.

3.2.5 Upgrade

FCA interface is versioned; because Leadville components are bundled and this error is not expected to occur, error recovery will be to fail loading Leadville.

HBA API interface is versioned, so that necessary level will be checked for proprietary extensions.

3.2.6 Software Removal

Normal package tools

3.3 System Administration

System administration is based on several different tools. Typically the user will select one tool to use for administration depending on his or her environment. For non virtualized environments, the user will use the fcadm command to create NPIV ports. For xVM, the xm command can be used. For zones, the zonecfg command will be used.

- The administrator will use fcinfo to report status, regardless of the virtualization environment.
- For this initial release of software, the system administrator must administer the WWN namespace. This involves creating a unique 64 bit value for each NPIV port.

Documentation will guide the user in creating a correctly formatted WWN.

- Future HBAs will have the ability to create WWNs based upon the HBA's permanent WWN, and importantly store in non volatile memory which generated WWN are in use. It is important to be able to store in use WWN so that WWNs are not reused.

4 FibreChannel Driver NPIV Architecture

4.1 Description

The major architectural change is to change the relationship between the FCA (Fibre Channel Adapter, one instance per FC port) and the fp (supports a generic fibrechannel port) driver. Currently the relationship between the FCA and fp driver instances is 1:1, this project changes the relationship to 1:many.

Additionally, the interface between the FCA and fp driver changed to support NPIV. The FCA Driver must returns information about support for NPIV and also return specific errors related to NPIV. The fp driver can alsoo create and delete NPIV ports through this interfaces

The public user interface to NPIV is based on extensions to the HBA-API covered in the next section. The HBA API communicates to the driver through new ioctl calls. These are project private interfaces to the driver.

4.2 Interfaces

4.2.1 User-visible

No public interfaces.

4.2.2 Internal (optional for ARC review)

Updated FCA interfaces are included in file NPIV_FCA_Interface_Doc-0.2.pdf

New ioctl interfaces to support HBA API are project private.

4.3 Operation

The high level operation of the FCA interface remains unchanged. The port's permanent WWN (the one that the adapter currently uses, found on the HBA's NV memory), will continue to be used automatically. The permanent WWN will never be disabled.

NPIV ports are created and deleted by private ioctl calls into the fp driver. This causes the fp driver to call the fca driver's `fca_bind_port` function to enable the port in the FCA driver. In response, the FCA driver then creates a new fp instance.

Neither the Emulex and Qlogic HBAs not have enough on board hardware resources to support fcip (IP over FC) on NPIV ports. We will not attach the fcip driver for NPIV ports. Current non NPIV ports will continue to attach fcip, so there is no problem with upgrade.

5 HBA API Architecture

5.1 Description

HBA API library is extended to support NPIV. The current implementation of HBA-API is specified in PSARC 2002/644 FibreChannel HBA API V2. HBA-API is an industry standard interface.

5.2 Interfaces

5.2.1 User-visible

See section 2.2

5.2.2 Internal (optional for ARC review)

HBA_API_ext.man

5.3 Operation

HBA-API interface is a SNIA/T11 industry standard specification that allows accessing information about FibreChannel HBAs from user level libraries. There is work in T11 to add support for NPIV to HBA-API, but at this point it as at a conceptual level, and not possible to implement.
For this reason, we have developed our own extensions to HBA-API to support NPIV. The are backwards compatible with the standard HBA-API. We plan on implementing the T11 standard when it becomes mature.

6 fcadm and fcinfo CLI

6.1 Description

fcadm is a new CLI command used to administer NPIV in a standalone environment. Fcadm's syntax and implementation is based upon fcinfo. It will be a simple command which allows NPIV ports to be created and deleted.

Additionally, fcinfo will be extended to accommodate NPIV. fcinfo will be able to list relationship between NPIV ports and physical adapter. It will also list error states, for example when a switch is not able to support HBA API.

6.2 Interfaces

6.2.1 User-visible

See proposed fcadm man page, and additions for fcinfo.

6.2.2 Internal (optional for ARC review)

Fcadm will use the HBA-API extensions described above.

6.3 Operation

Users will create and delete NPIV ports using fcadm. fcinfo will be used to return status of NPIV ports.

NPIV port data will be stored in the file /etc/fp/cfg/NPIV_WWN. Format will use XML. This file will be read at boot type by fcadm, called from SMF.

7 Xen Architecture

Xen modifications consist of several parts:

- Bind an NPIV port to a DOMU
- When a device comes online that is associated with a DOMU's NPIV port, automatically map that device to the DOMU
- When a DOMU migrates between two physical servers, first disable the NPIV port on the original server, and enable the NPIV port on the destination server at appropriate points in the migration process.

These topics are covered in more details in the following sections.

7.1 Description

We modify the Xen's standard CLI, xm, to accept NPIV WWN and associate with a DOMU and store to Xenstore. This is the sole part of the change that is platform independent, and is suitable to contribute to the Xen community.

Management will be handled by a new daemon, xennpivd. It will be started at boot time from smf(5). The main function is to wait for changes on the SAN using the HBA API interface. It also is responsible for creating and deleting NPIV ports. The HBA API interface provides for callbacks when the SAN changes. If the daemon determines that a FibreChannel target has been added or deleted, it determines if it is from an NPIV port that is mapped to a DOMU. If it is from an NPIV port, it will perform the appropriate Xen commands to add or delete the device to the SAN. More details are in section 7.3

We introduce a small device driver, xnpivd. Due to Xen architecture, we need a device driver to register for Xenstore change callbacks. Its sole function is to callback to xennpivd when Xenstore is updated.

Note, we will also support migration of DOMUs across physical servers, by disabling an NPIV port on the origin server, and enabling the NPIV port on the destination server. This work is not complete, but will be by commitment review.

7.2 Interfaces

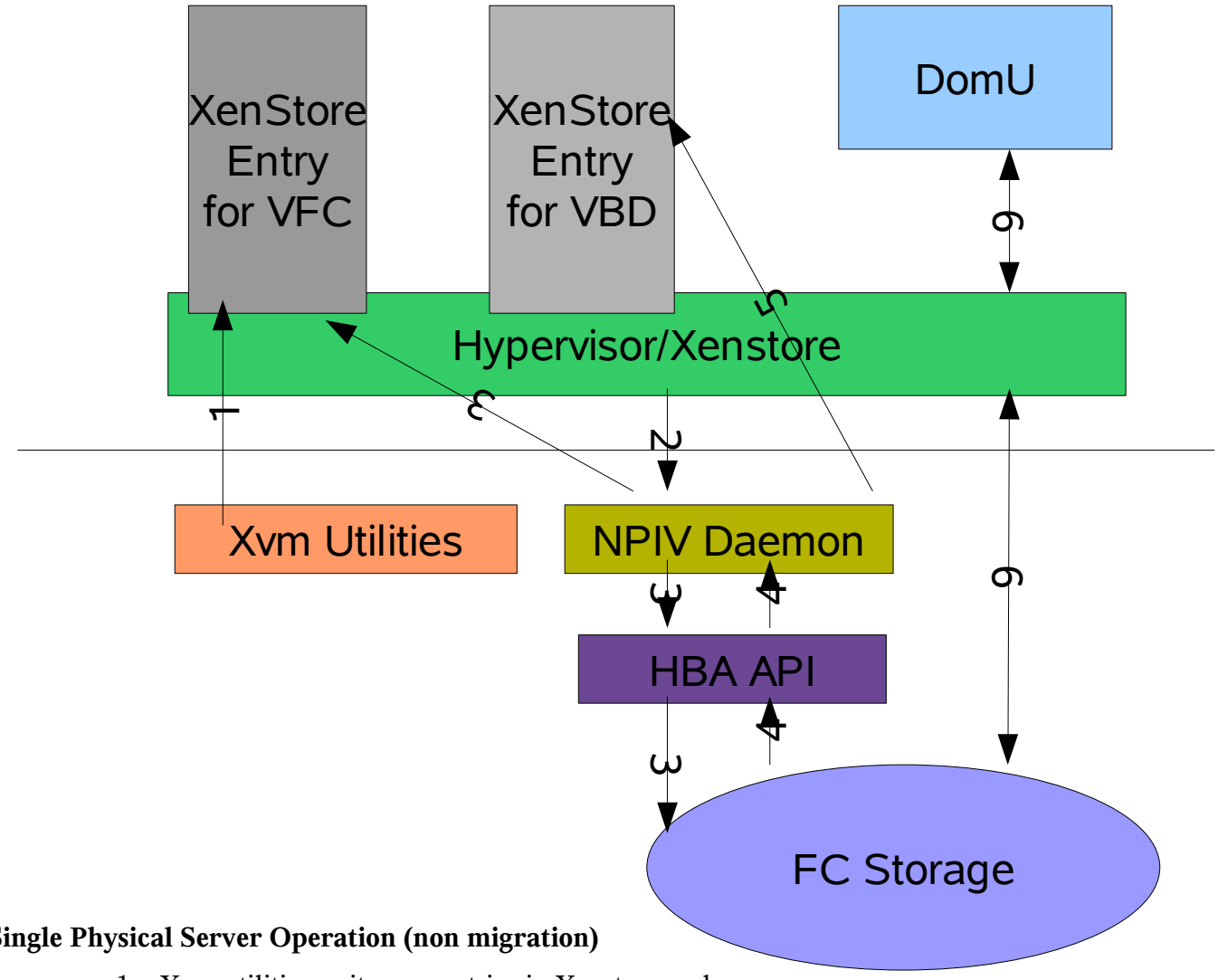
7.2.1 User-visible

xm will be modified. See section 2.1

/usr/lib/xennpivd daemon file location.

7.2.2 Internal (optional for ARC review)

7.3 Operation



Single Physical Server Operation (non migration)

1. Xvm utilities write new entries in Xenstore under /local/domain/0/backend/vfc/$DOMUID/.

2. Xnpivd, the backend driver of FC port, is triggered by the new entry, then it calls the NPIV daemon (xennpivd) through a door.
3. NPIV daemon calls HBAAPI to create the Virtual port and then updates the newly created entry with 'connected' if succeeded. It also registers devices event via HBAAPI.
4. Devices discovered from the new FC port trigger events which are captured by the NPIV daemon.
5. NPIV daemon calls xvm utilities/Xen API to attach new devices to DomU.
6. The DomU is now able to access the storage via the virtual port.

8 Containers Architecture

8.1 Description

Containers changes. Changes for containers are similar to Xen in result, although different due to implementation. Specifically, we're able to associate an NPIV virtual port with a non global zone, and optionally automatically map devices from that port to that zone.

8.2 Interfaces

8.2.1 User-visible

User visible changes (zonecfg) are described in Section 2

8.2.2 Internal (optional for ARC review)

8.3 Operation

Currently, operation is limited to devices discovered at when adding NPIV port information. This will be expanded to be dynamic in the future.

When user adds NPIV port information via zonecfg, we use extended HBA API commands to create the port on the HBA. We then scan for devices matching this controller's HBA, and add the devices using libzonecfg interface.

We expect to do this dynamically by running a daemon in the global zone, and waiting for callbacks from the HBA API. This part of design will be completed for commitment review.

Appendix A: Standards Supported

References

T11 FC -LS FibreChannel Link Services, T11/05-345v1

- T11 FC-DA, FibreChannel Device Attach T11/04-202vA
- T11 FC-GS-4 FibreChannel Generic Services (T11/04-031v2)

R.1 Related Projects

1997/385 Fibre Channel Driver Re-architecture (Leadville)

2006/260 Solaris on Xen

2002/174 Virtualization and Namespace Isolation in Solaris (Zones)

2002/644 FibreChannel HBA API V2

R.2 Background Information for this Project or its Product

See OpenSolaris project page:

http://opensolaris.org/os/project/npiv/

R.3 Interface Specifications

None

R.4 Project Details

None