Education

Multipath Management API : Standards Based SCSI Multipathing Administration

Hyon Kim, Sun Microsystems

Giri Basava, TLC Chair- SNIA IPSF, EMC²

SNIA Legal Notice

- The material contained in this tutorial is copyrighted by the SNIA.
- Member companies and individuals may use this material in presentations and literature under the following conditions:
  - Any slide or slides used must be reproduced without modification
  - The SNIA must be acknowledged as source of any material used in the body of any document containing material from these presentations.
- This presentation is a project of the SNIA Education Committee.

Abstract

Multipath Management: Standards Based SCSI Multipathing Administration

This session will appeal to Developers, Development Managers, and those that are seeking a fundamental understanding of the highly available storage with multiple paths and management of those multiple paths using Multipath Management API(MMA).

As storage networks have evolved, the support for SCSI multipath devices has become one of critical areas for storage administration. The session will delve into the concepts of multipath storage followed by in depth discussion of SNIA MMA(INCITS 412 -2006) which provides management interfaces to the standard capabilities defined in ANSI INCITS 408-2005(SPC-3) standards.

Contents

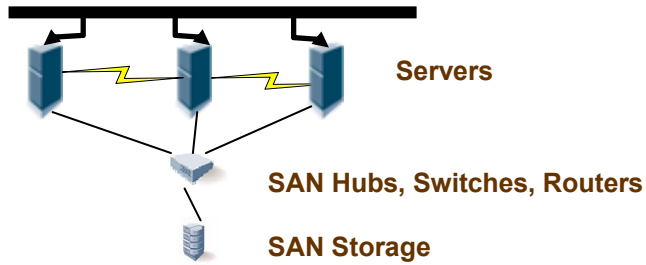
- Highly Available SAN Storage
  - Server Clustering
  - Redundancy through RAID
  - Multipath SAN Storage
- Single Path SAN Storage
- Multipath SAN Storage
  - Overview
  - Host Stack
  - Target Port Groups
  - Fibre Channel
  - iSCSI
  - Load Balancing

Contents …

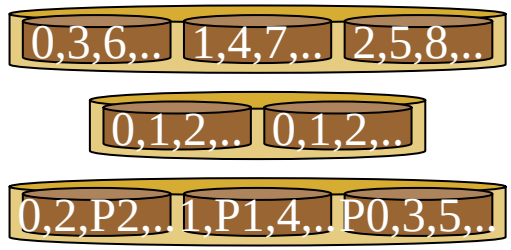
- Multipath Management API
  - Overview
  - Architecture
  - Object Model
  - SCSI Target Port Groups Support
  - Object Discovery
  - Load Balance Administration
  - Path Selection
  - Path Administration
  - Event Support
- Summary

Highly Available SAN Storage

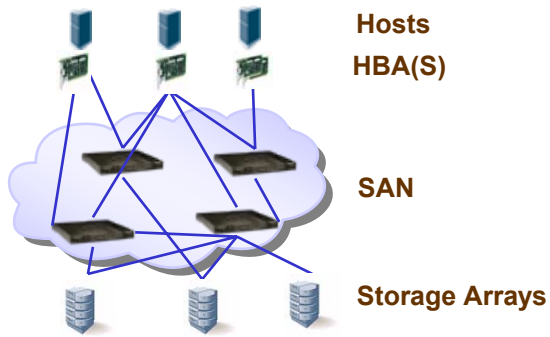
- Server Clustering
  - Server Clustering – is what looks like a single system using multiple hosts, inter connects and storage arrays.



- Redundancy through RAID
  - RAID is for Fault Tolerance & performance.
  - RAID 0: Stripes,
  - RAID 1: Mirrors, Shadows,…
  - RAID 5: Parity
  - ETC.

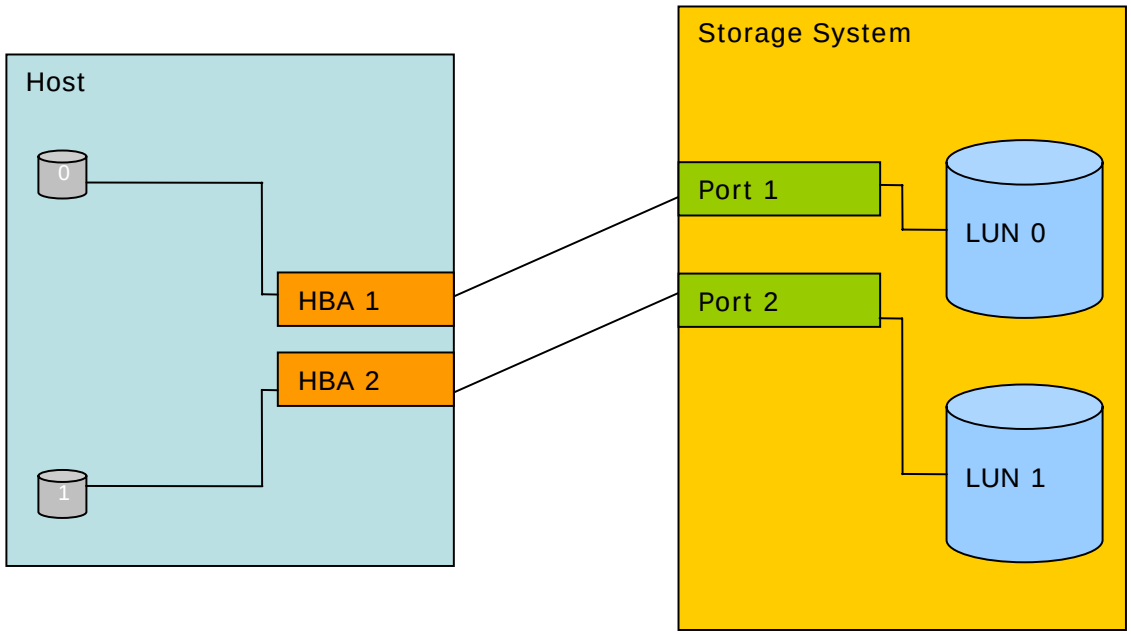


- Multipath Storage
  - Multipathing solutions use redundant physical path components–adapters, cables, and switches–to create logical "paths" between the server and the storage device.



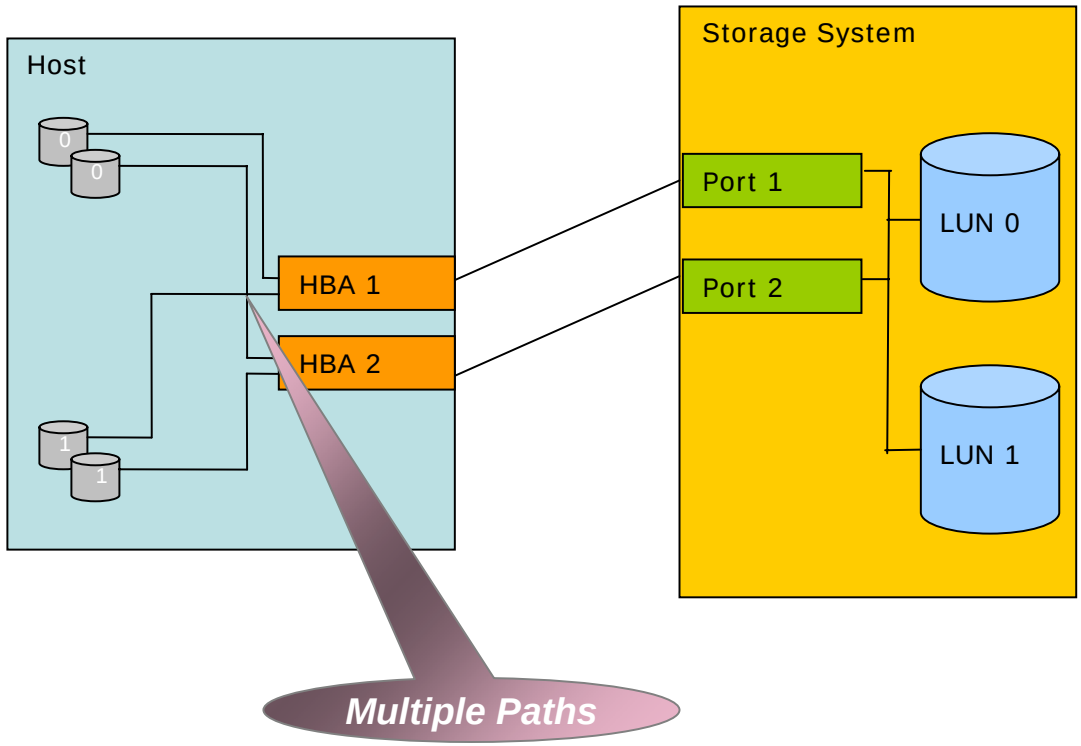
Single Path SAN Storage

- Single path configuration.
  - One path to each logical device/unit.
  - Single point of failure.



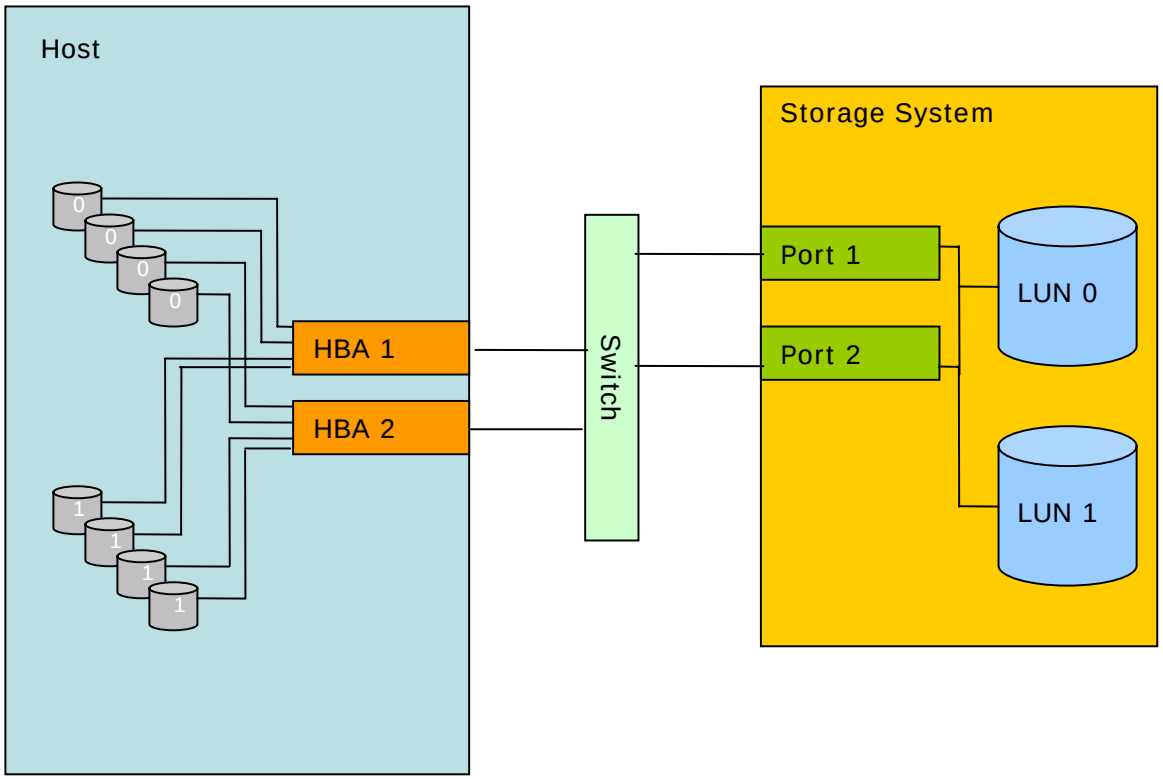
Multipath SAN Storage - Overview

- Simple multiple path configuration
  - Each LUN has two paths
  - Application IO can survive loss of one path.



Multipath SAN Storage – Overview…

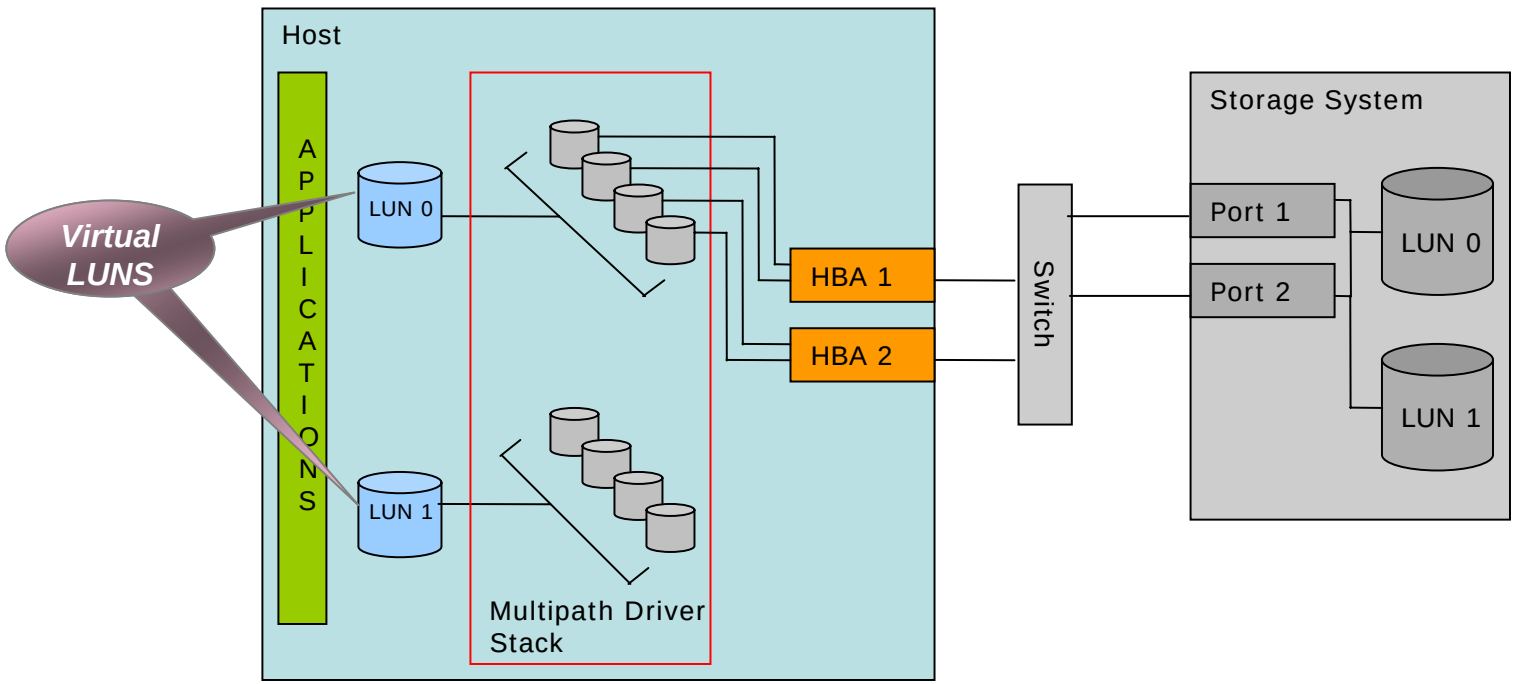
- More paths…
  - Each LUN has four paths.
  - Application IO can survive loss of three paths.



Multipath SAN Storage – Host Stack

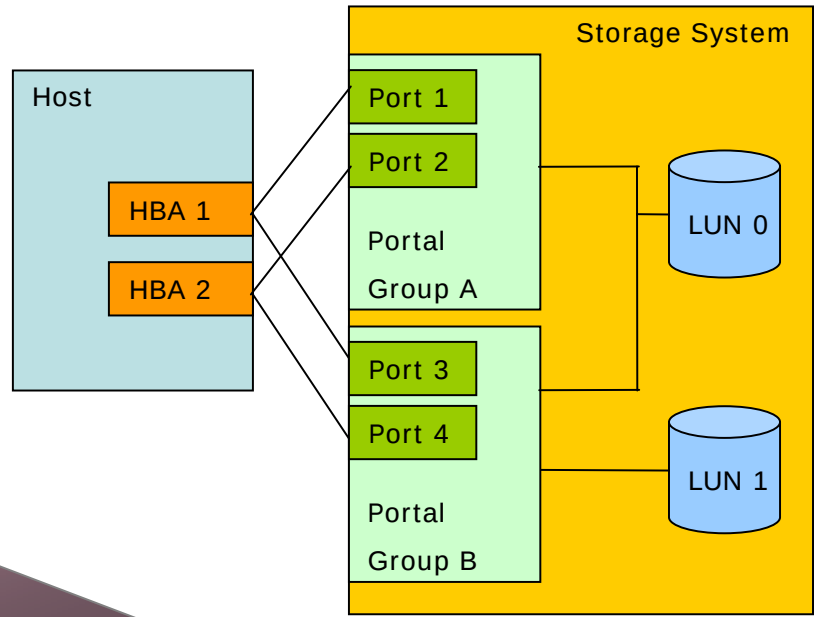
Host Multipath Stack

- Presents a single instance of LUN to the applications.
- Handles failover in case of losing one or more paths.
- May perform load balancing across the paths.



Multipath SAN Storage - Target Port Groups

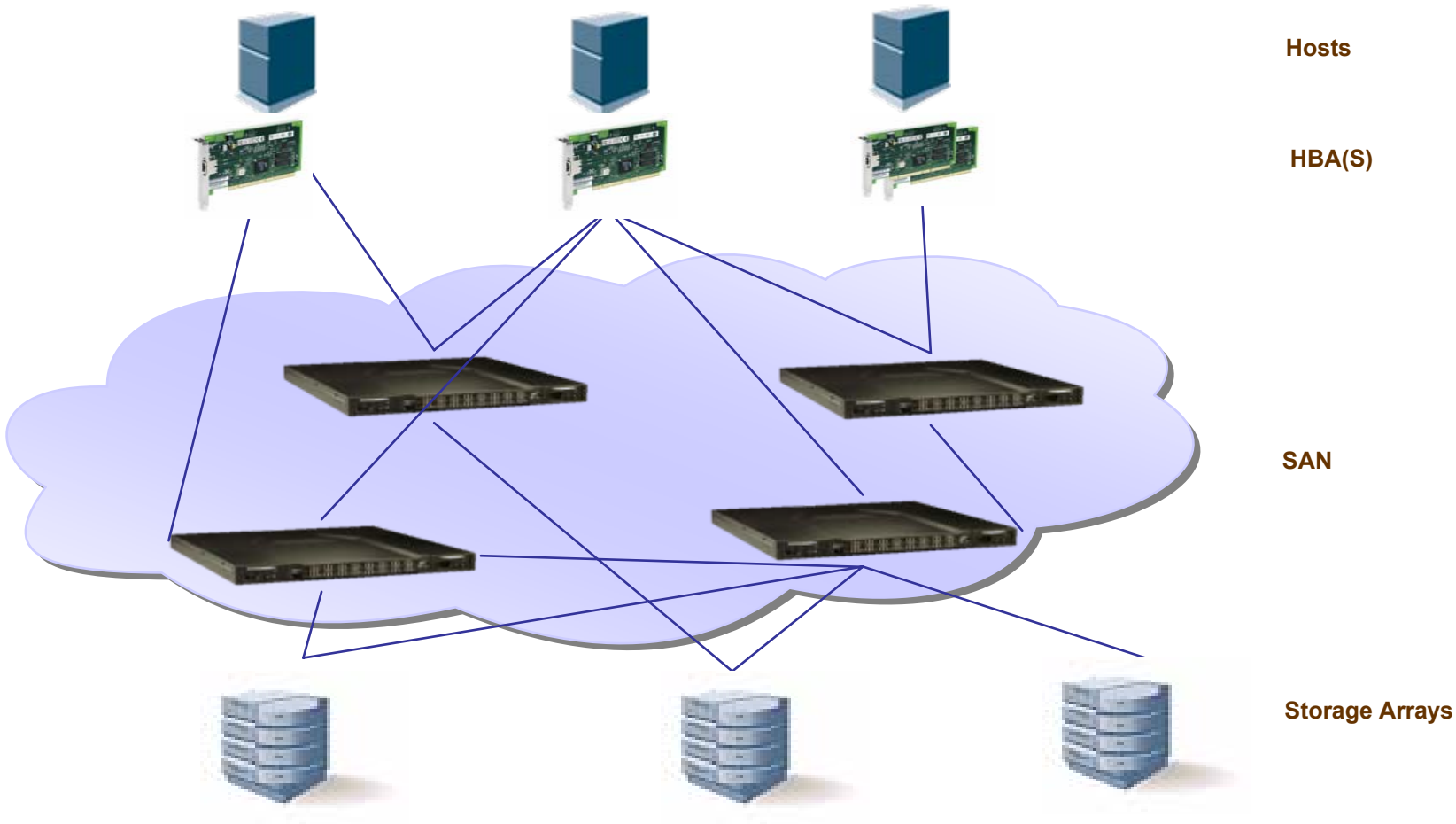
- Target Port Groups(TPG)
  - Storage systems offer another level of redundancy using multiple storage processors or portal groups(a logical group of ports).
  - Failure of one storage processor or portal group will not result in application downtime.
  - TPG's are reported per LUN.
  - Defined in SPC-3.



- TPG Access States
  - **Asymmetric** : Each TPG will have one of the following access states
    - Active/Optimized
    - Active/non-optimized
    - Standby(or passive)
    - Unavailable
  - **Symmetric**: Same state (active/optimized) will be in effect for all TPG's.

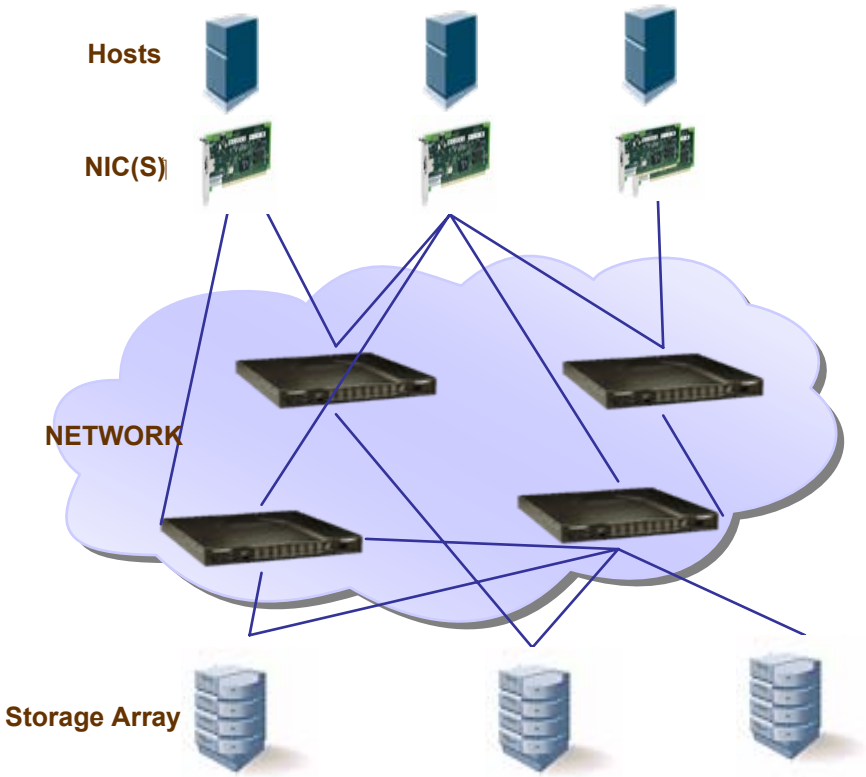
ALUA(Asymmetric Logical Unit Access): When supported by the storage arrays, host based multipath implementations can manage the portal group states.

Multipath SAN Storage - Fibre Channel



Multipath SAN Storage – iSCSI

- Storage system with multiple iSCSI Nodes (targets).
- Multiple network portals per iSCSI Node.
- Multiple sessions to a network portal.
- Multi Connection Sessions (MCS)*
- Link Aggregation (Teaming or Trunking)*



* MMA is oblivious to the multiple paths in these configurations.

Multipath SAN Storage – Load Balancing

- Well known algorithms
  - Round Robin: Take a turn among all active paths.
  - Least blocks: Select a path based on number of outstanding blocks to read/write
  - Least IO (Queue Depth): Select a path based on number of outstanding IO's.
  - LBA region: Performs load balance using sequential stream detection.
  - Failover only: Only one path is active at a time.
  - Etc.
- Device Specific
  - Depends on the multipath device
  - Vendor specific implementation and optimizations.

MMA - Overview

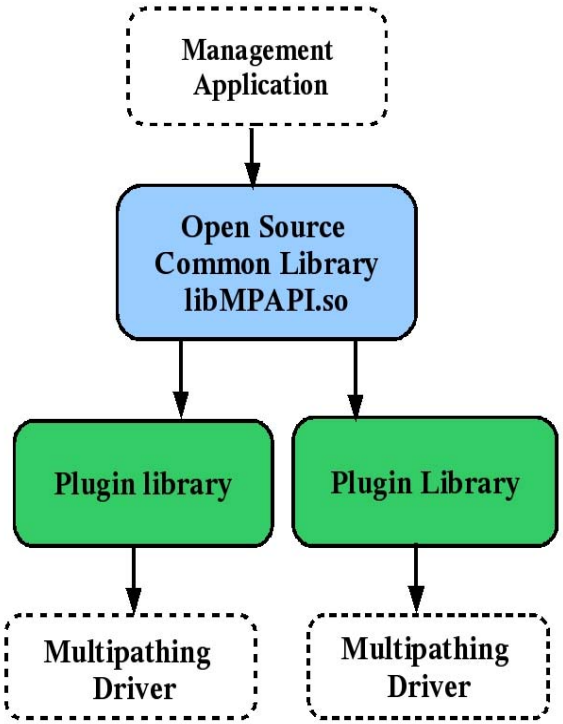
- ANSI standard API for SCSI Multipathing administration
  - Provides host view of SCSI layer multipathing.
  - Can be mapped to in SMI-S SCSI Multipath Management Subprofile for enterprise-wide support.
  - Two tier API architecture.
- Multipath device administration
  - Load balance and path selection.
  - failover and failback.
- Follows SCSI Target Port Groups model
  - Defines corresponding Target Port Group object.
  - Asymmetric access state is reflected in TPG object state.

MMA - Architecture

- Two tier API layers (Common and Plugin) accommodate both common interfaces and vendor specific solution.
  - Storage management application can call the same interfaces across different platforms.
  - Vendors export their own implementation of multipathing support through commonly-defined interfaces.
  - Common library is open sourced with SNIA license:

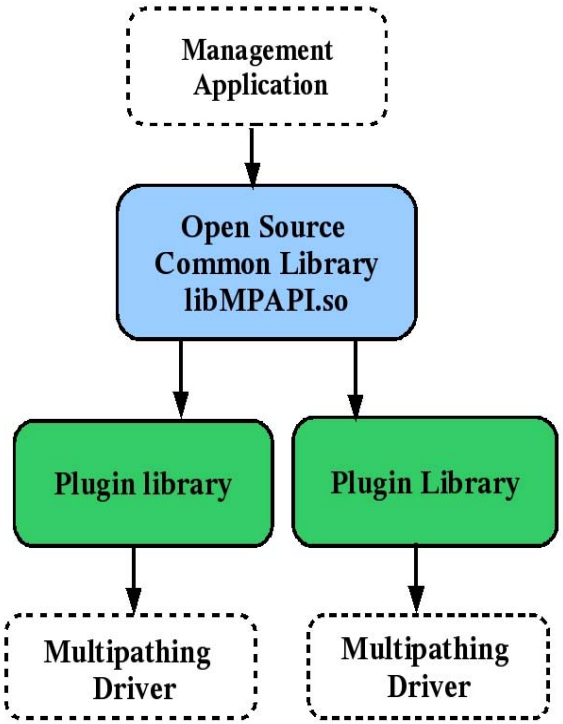
    http://sourceforge.net/projects/mp-mgmt-api
  - Plugin library implementation example: http://cvs.opensolaris.org/source/xref/nwsc/src/sun_nws/mpapi_svplugin/

MMA – Architecture



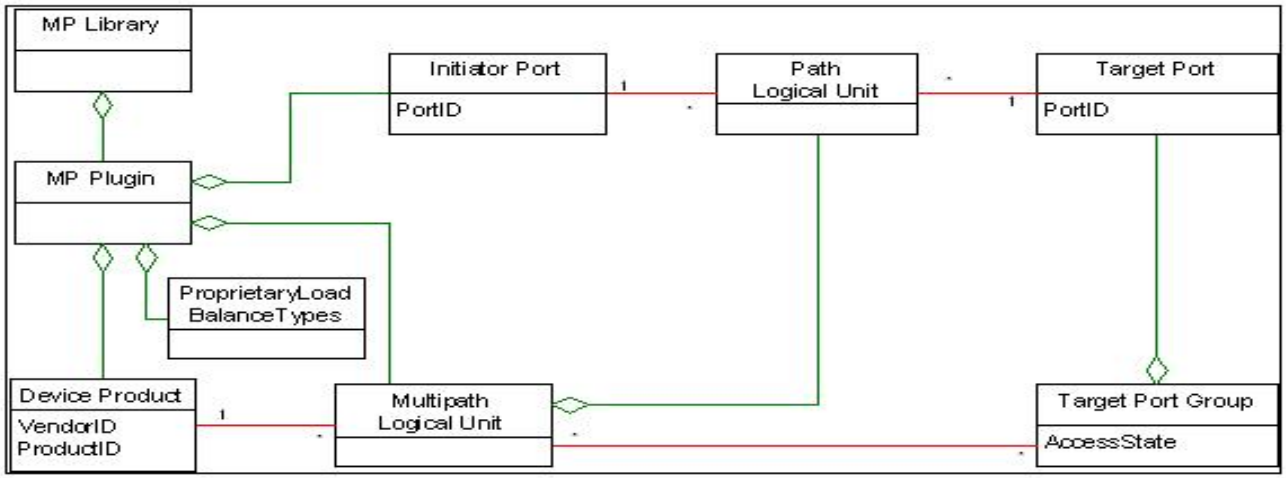
- Common Library
  - Provides interfaces that applications invoke to administer multipath device.
  - Defines a set of entry points that a plugin needs to implement.
  - Loads plugin(s) and dispatches application requests to appropriate plugin(s)

MMA - Architecture



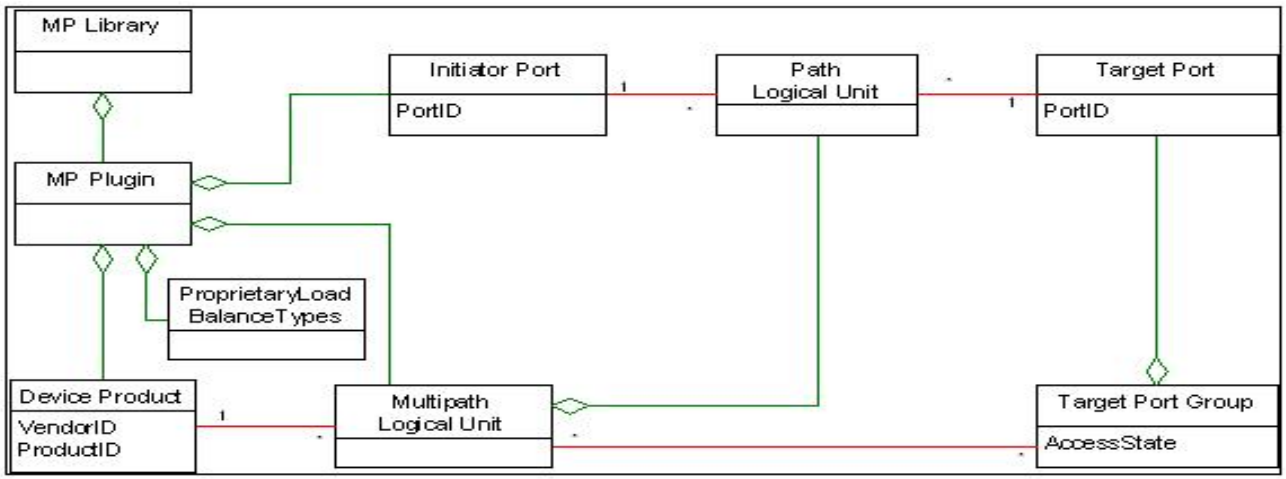
- Plugin library
  - Provided by a vendor to support MP API for the specific set of targets that its multipathing driver supports.
  - Registers with the common library at the installation time.
  - Vendor specific way to communicate with the multipathing driver.

MMA – Object Model



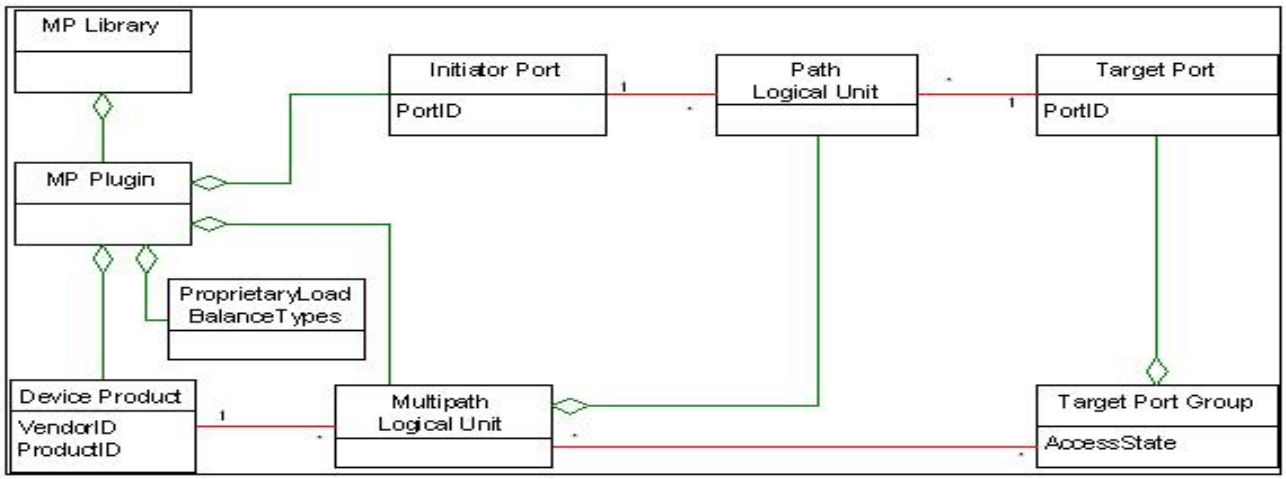
- Library
  - Represents a common library.
  - Properties: version, built time, vendor
- Plugin
  - Vendor provided plugin library.
  - Identified by the vendor name and file name.
  - Properties: underneath driver info, load balance support, autofailback, autoprobe state.

MMA – Object Model



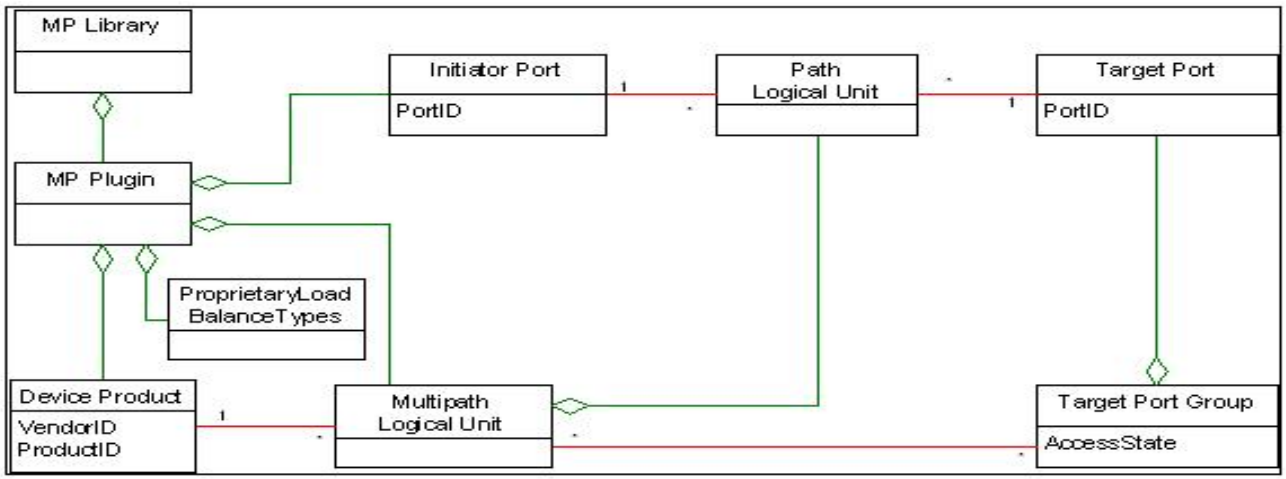
- Device Product
  - Target devices that a plugin supports.
  - Properties: vendor, supported load balance algorithm
- Proprietary Load Balance Type
  - Represents vendor specific load balance algorithm.
  - Identified by a name and proprietary properties that a vendor provides.

MMA – Object Model



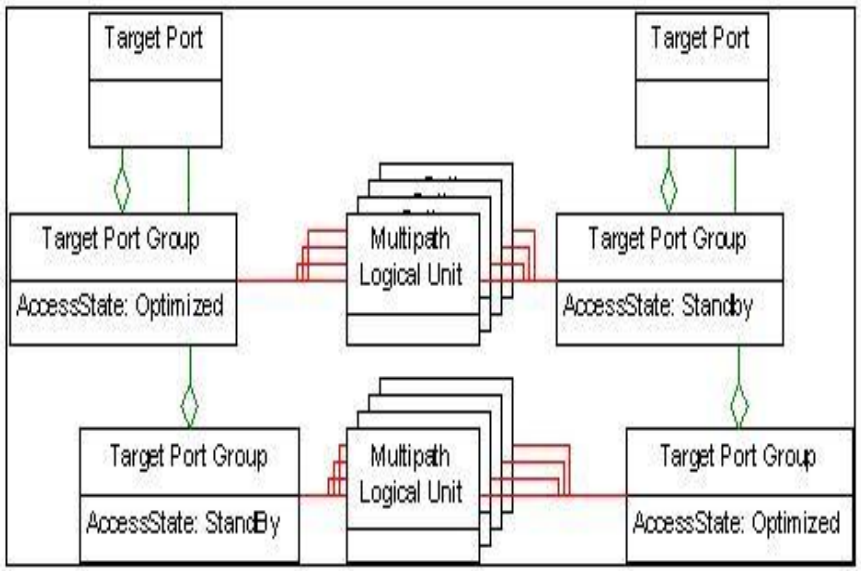
- Initiator Port: Identified by name and transport type(iSCSI, FC, SAS).
  - Ex: 210000e08b028a5c(FC port WWN)
    - iqn.1986-03.com.sun:01:e00000000000.4665b07d,4000002a00ff(iSCSI iqn name)
- Target Port: Identified by name and SCSI Inquiry VPD page 83 type 4h identifier.
  - A SCSI device may have multiple target ports.
- Target Port Group: Identified by Target Port Group ID field in SCSI Report TPGs cmd.
  - Group of target ports with the same access state relative to a SCSI Logical Unit.
  - Applies to both symmetric and asymmetric devices.

MMA – Object Model



- Path Logical Unit
  - Consists of an initiator port, a target port and a SCSI logical unit.
  - Properties: state(okay, error), administratively disabled?, path weight..
- Multipath Logical Unit
  - Virtual device that aggregates all paths to the same device (SCSI Logical Unit).
  - Identified by the SCSI INQUIRY VPD page 83h.
    - Ex. 600C0FF00000000000036223AE73EB705
  - Properties: asymmetric?, load balance info, auto failback enabled, auto probe enabled?, override path.

MMA - SCSI TPG Support



- Response from SCSI Report Target Port Groups command is reflected in API Target Port Group object and its properties.
  - TPG object state shows Asymmetric Access State: active optimized, active nonoptimised, standby, unavailable, transitioning.
  - Status code reflected in explicitFailover property.
  - Preferred bit corresponds to PreferredLUPath property.
- Failover occurs when explicit state state change is made.
- A plugin library/multipathing driver synthesizes TPG for legacy devices.

MMA - Object Discovery

- Allows to discover attributes of API objects
  - initiator port, target port and multipath device.
  - path and TPG associations.
- APIs
  - MP_Get<object>OidList()
  - MP_GetAssociated{path|TPG}OidList()
  - MP_Get<object>Properties()

MMA - Load Balance Administration

- Common load balance algorithms are built into API
  - round robin, least block, least IO, LBA region, failover only.
  - Supported load balance(s) are exposed through Plugin/Device Product property supportedLoadBalanceTypes.
  - Specific load balance type can be set at a Plugin layer or an individual multipath device layer as allowed by the underneath driver.
- Allows a proprietary algorithm
  - Vendor's private load balance algorithm is exposed through proprietary load balance type.
  - Not actually interpreted by API.
- APIs
  - MP_SetPluginLoadBalanceType()
  - MP_SetLogicalUnitLoadBalanceType()
  - MP_GetProprietaryLoadBalanceOidList()
  - MP_GetProprietaryLoadBalanceTypeProperties()

MMA - Path Selection

- Allows to administratively influence path selection for load balance.
  - disable a path to remove it from a list of the selectable paths by the driver.
  - Make a path override underneath path selection schemes.
  - Assign a weight among paths to influence path selection by the driver.
- APIs
  - MP_EnablePath()/MP_DisablePath()
  - MP_SetOverridePath()
  - MP_SetPathWeight()

MMA - Path Adminstration

- Path state can be explicitly set, if allowed by underneath driver/device
  - Target Port Group object state can be changed to trigger failover for an asymmetric device.
  - Autofailback behavior can be configurable in a plugin and/or an individual multipath device.
  - Autoprobe to validate operational paths that are not currently used.
- APIs
  - MP_SetTPGAccess()
  - MP_EnableAutoFailback()/MP_DisableAutoFailback
  - MP_EnableAutoProbe()/MP_DisableAutoProbe

MMA – Event Support

- Long running applications can be dynamically updated on any state change/removal/addition of objects.
  - Plugin keeps track of the changes through OS event driven communication.
  - The application provided functions are executed as an event occurs.
- APIs
  - MP_RegisterForObjectVisibilityChanges()
  - MP_DeregisterForObjectVisibilityChanges()
  - MP_RegisterForObjectPropertyChanges()
  - MP_DeregisterForObjectPropertyChanges()

Summary

- Multipathing… a high availability solution, is the ability to use more than one read/write path to a storage device.
- Multipathing… provides fault tolerance against single-point-of failure in hardware components.
- Multipathing… provides load balancing of I/O traffic, thereby improving system and application performance
- MMA… is ANSI standard API for SCSI Multipathing administration
- MMA… allows vendor specific multipathing solutions to be configured and managed through common interfaces across different platforms.
- MMA... MultipathManagementAPI under SNIA HWG document

Q&A / Feedback

- Please send any questions or comments on this presentation to SNIA: trackvirtualization@snia.org

Many thanks to the following individuals for their contributions to this tutorial.

SNIA Education Committee

Robert Peglar
David Butchart
Niraj Jaiswal
Pual von Behren