SCSI - 3 Device Mulitpathing

MPxIO

PSARC 1999 / 647

Presented By

Network Storage

Background & Motivation

Storage Pools

Increased availability and I/O bandwidth

Multiple HCIs connected to the Storage pool

Solaris identifies a separate and independent device instance

Many varying multipathing solutions for Solaris

Current Limitations

Wasteful of system resources

Spec nodes, dev_info_t, driver soft state and inodes in the root file system.

Device configuration management

prtconf, DR and multiple logical names

Stateful drivers

Tape drivers in multipath configurations

Failover / Error management

Limited error status information with the buf(9s) structure

Multiple user interfaces

Testing nightmare

Goals

Define a generic instance scheme for representing multipathed devices within Solaris

Support for booting, DR and power management

Common architecture for I/O path management

Automatic failover to alternate paths on transport failures

Tunable load balancing for improved I/O performance

Co-existance with other multipathing solutions

Software Architecture

Client Device Drivers

Disassociation of multipath devices from physical paths

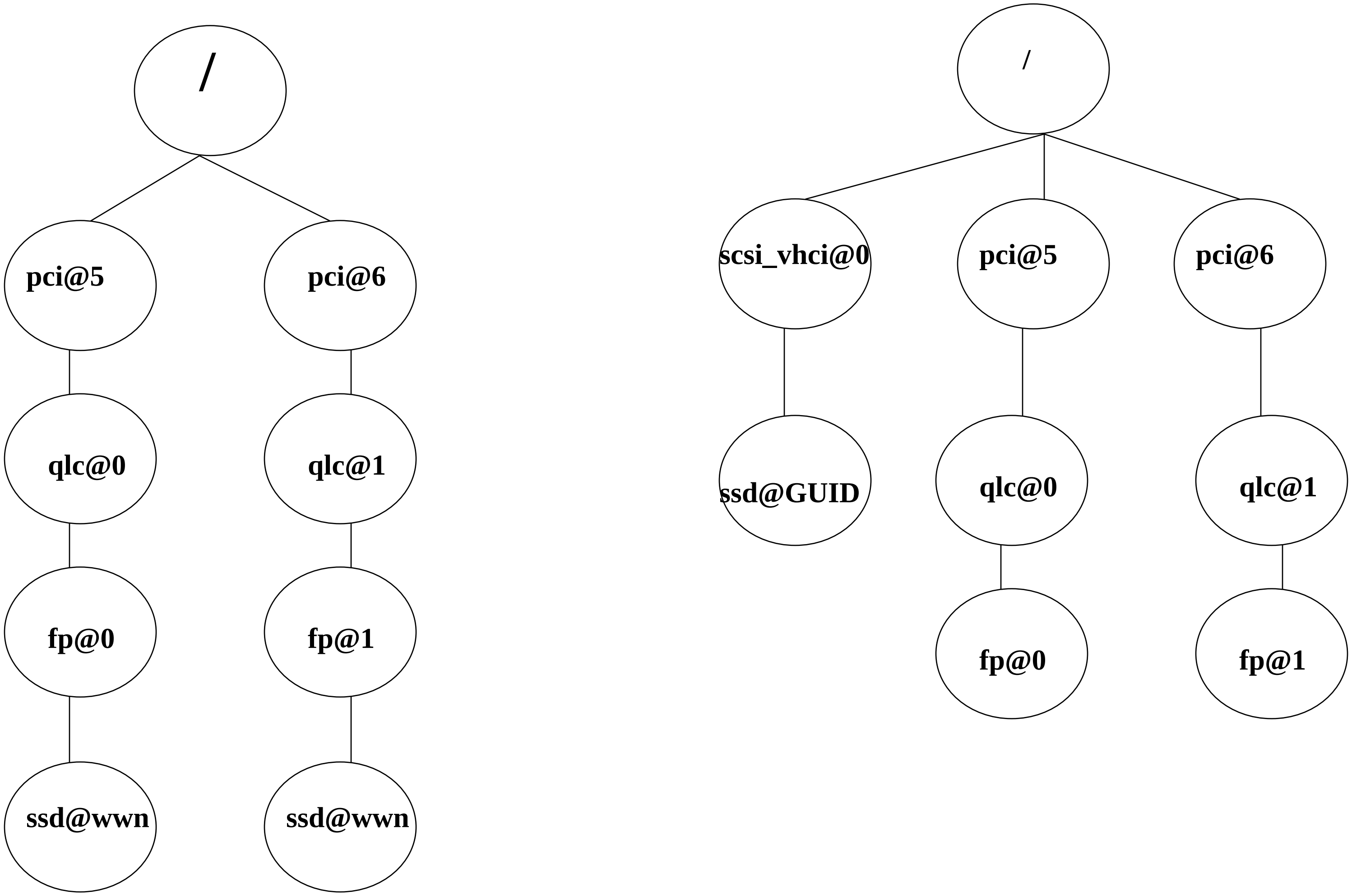
vHCI Drivers (Virtual Host Controller Interface)

Single instance representation of multipath devices
Configuration management
I/O request routing and policy-based load balancing
Failover support

pHCI Drivers (Physical Host Controller Interface)

Device Enumeration
Physical transmission of data

Solaris Device Tree



format Example

Non mpxio case:

```
# format
Searching for disks...done


AVAILABLE DISK SELECTIONS:
       1. c2t112d0 <SUN9.0G cyl 4924 alt 2 hd 27 sec 133>
          /pci@1f,4000/pci@4/SUNW,qlc@4/fp@0,0/ssd@w2200002037070539,0
       2. c3t112d0 <SUN9.0G cyl 4924 alt 2 hd 27 sec 133>
          /pci@1f,2000/pci@4/SUNW,qlc@4/fp@1,0/ssd@w2100002037070539,0
       3. c2t114d0 <SUN9.0G cyl 4924 alt 2 hd 27 sec 133>
          /pci@1f,4000/pci@4/SUNW,qlc@4/fp@0,0/ssd@w22000020370704cc,0
       4. c3t114d0 <SUN9.0G cyl 4924 alt 2 hd 27 sec 133>
          /pci@1f,2000/pci@4/SUNW,qlc@4/fp@1,0/ssd@w21000020370704cc,0
```

mpxio case:

```
# format
Searching for disks...done

AVAILABLE DISK SELECTIONS:
       1. c4t2000002037070539d0 <SUN9.0G cyl 4924 alt 2 hd 27 sec 133>
         /scsi_vhci/ssd@g2000002037070539,0
       2. c4t20000020370704ccd0 <SUN9.0G cyl 4924 alt 2 hd 27 sec 133>
         /scsi_vhci/ssd@g20000020370704cc,0
```

luxadm Example

Non mpxio case:

```
# luxadm display /dev/rdsk/c1t21020f200000249d0s2

DEVICE PROPERTIES for disk: /dev/rdsk/c1t21020f200000249d0s2
  Status(Port A):       O.K.
  Status(Port B):       O.K.
  Vendor:               SUN
  Product ID:           T300
  WWN(Node):            20020f2000000249
  WWN(Port A):          21020f2300000249
  WWN(Port B):          22020f2300000249
  Revision:             0114
  Serial Num:           0000028411
  Unformatted capacity: 136588.000 MBytes
  Write Cache:          Enabled
  Read Cache:           Enabled
    Minimum prefetch:   0x0
    Maximum prefetch:   0x0
  Device Type:          Disk device
 Path(s):
 /dev/rdsk/c1t21020f2000000249d0s2
 /devices/pci@1f,4000/pci@4/SUNW,qlc@5/fp@0,0/ssd@w21020f2300000249,0:c,raw
 /dev/rdsk/c2t22020f2000000249d0s2
 /devices/pci@1f,4000/pci@4/SUNW,qlc@4/fp@0,0/ssd@w22020f2300000249,0:c,raw
```

luxadm Example (contd.)

mpxio case:

```
# luxadm display /dev/rdsk/c1t60020f200000033939a2c2b60008d4aed0s2
DEVICE PROPERTIES for disk:
/dev/rdsk/c1t60020f200000033939a2c2b60008d4aed0s2
  Status(Port A):       O.K.
  Status(Port B):       O.K.
  Vendor:               SUN
  Product ID:           T300
  WWN(Node):            20020f2000000249
  WWN(Port A):          21020f2300000249
  WWN(Port B):          22020f2300000249
  Revision:             0114
  Serial Num:           0000028411
  Unformatted capacity: 136588.000 MBytes
  Write Cache:          Enabled
  Read Cache:           Enabled
    Minimum prefetch:   0x0
    Maximum prefetch:   0x0
  Device Type:          Disk device
 Path(s):
  /dev/rdsk/c1t60020f200000033939a2c2b60008d4aed0s2
  /devices/scsi_vhci/ssd@g60020f200000033939a2c2b60008d4ae:c,raw
  /devices/scsi_vhci/ssd@g60020f200000033939a2c2b60008d4ae:c,raw
+   Controller          /devices/pci@1f,4000/pci@4/SUNW,qlc@5/fp@0,0
+     Device address    21020f2300000249,0
+     Class             primary
+     State             online
+   Controller          /devices/pci@1f,4000/pci@4/SUNW,qlc@4/fp@0,0
+     Device address    22020f2300000249,0
+     Class             secondary
+     State             standby
```

Coexistance With Other Multipathing/Volume Management Software

- mpxio may be disabled globally or on a per pHCI basis
- mpxio creates single instance of a multipathed device; no conflict with veritas DMP or AP.
- Volume management software will need to deal with long names

System Configuration Changes

- libdevinfo(3)
  Enahnced to provide multipathing info
- prtconf(1M)

```
ssd, instance #10
            Driver properties:
                 <...>
            Hardware properties:
                name <mpxio-component> length <7>
                    value ’client’
                name <client-guid> length <33>
                    value ’60020f200000024a39afb79f000a18fe’
            Paths from multipath bus adapters:
                fp#0 (standby)
                    name <node-wwn> length <8>
                        value <0x50020f20000003d9>.
                    name <port-wwn> length <8>
                        value <0x50020f23000003d9>.
                    name <target> length <4>
                        value <0x00000002>.
                    name <lun> length <4>
                        value <0x00000002>.
                     name <path-class> length <8>
                         value ’primary’
                fp#1 (online)
```

```
                    name <node-wwn> length <8>
                        value <0x50020f200000024a>.
                    name <port-wwn> length <8>
                        value <0x50020f230000024a>.
                    name <target> length <4>
                        value <0x00000001>.
                    name <lun> length <4>
                        value <0x00000002>.
                    name <path-class> length <10>
                        value ’secondary’
```

- Example of prtconf output with pHCI disabled.

```
SUNW,qlc, instance #0
            System properties:
                name <mpxio-disable> length <4>
                    value ’yes’
                name <unit-address> length <2>
                    value ’2’
                name <hba0-enable-adapter-hard-loop-ID> length <4>
                    value <0x00000000>.
                name <hba0-adapter-hard-loop-ID> length <4>
                    value <0x00000000>.
```

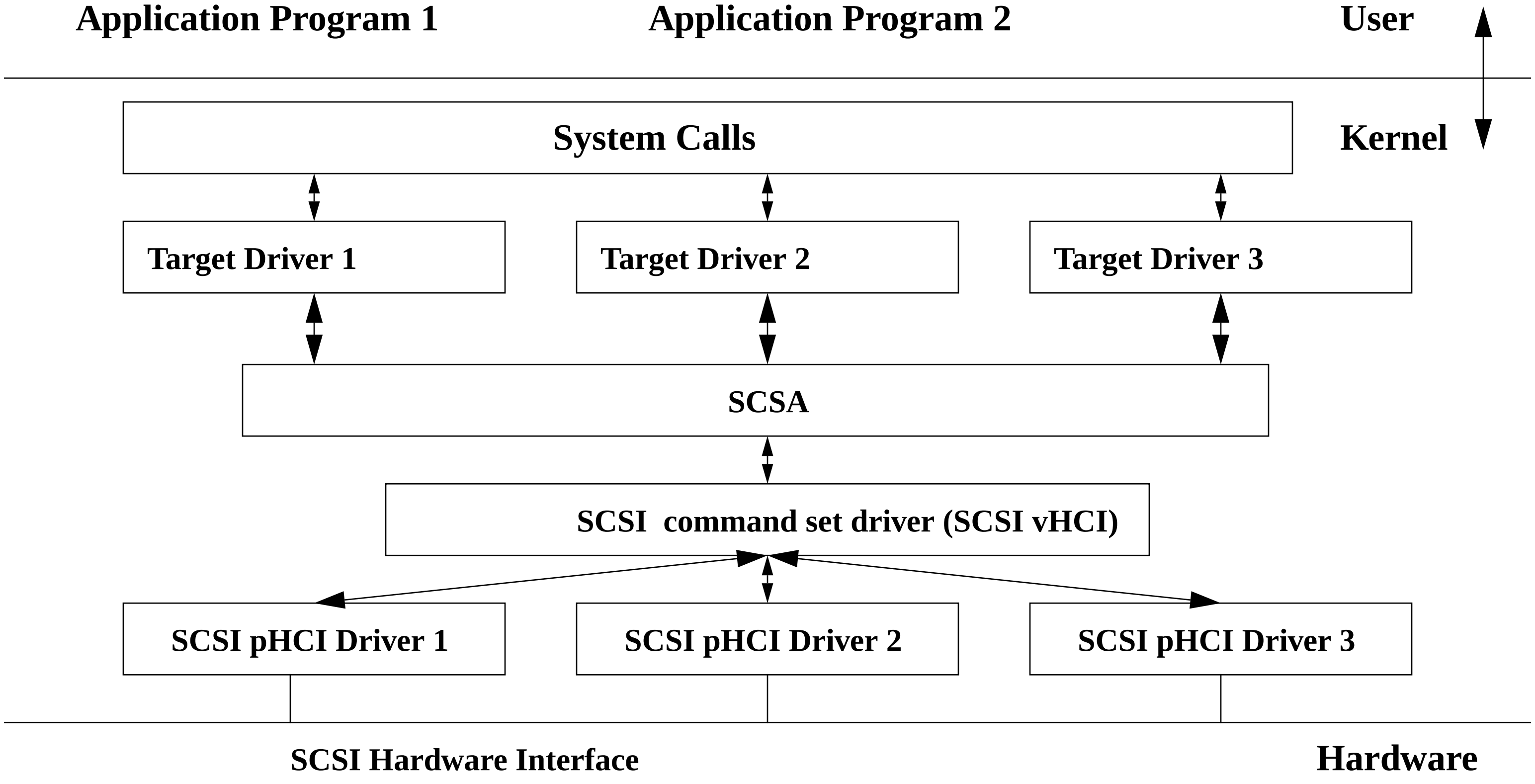
iostat(1M) Changes

```
iostat -xX
                extended device statistics
  device      r/s    w/s   kr/s   kw/s wait actv  svc_t  %w  %b
  ssd0       1.2    0.6   10.0    2.6  0.0  0.1   72.8   0   1
  ssd0.fp0    0.6    0.3    5.1    1.2  0.0  0.1   72.8   0   1
  ssd0.fp1    0.6    0.3    4.9    1.4  0.0  0.0   72.8   0   0
```

Extended statistics with path and error stats included:

```
iostat -xXe
                extended device statistics                    ---- errors ---
  device      r/s    w/s   kr/s   kw/s wait actv  svc_t  %w  %b s/w h/w trn tot
  ssd0       1.2    0.6   10.0    2.6  0.0  0.1   72.8   0   1   5   1   1   7
  ssd0.fp0    0.6    0.3    5.1    1.2  0.0  0.1   72.8   0   1   0   0   0   0
  ssd0.fp1    0.6    0.3    4.9    1.4  0.0  0.0   72.8   0   0   0   1   1   2
```

SCSI Functional Block Diagram



Phase I Features

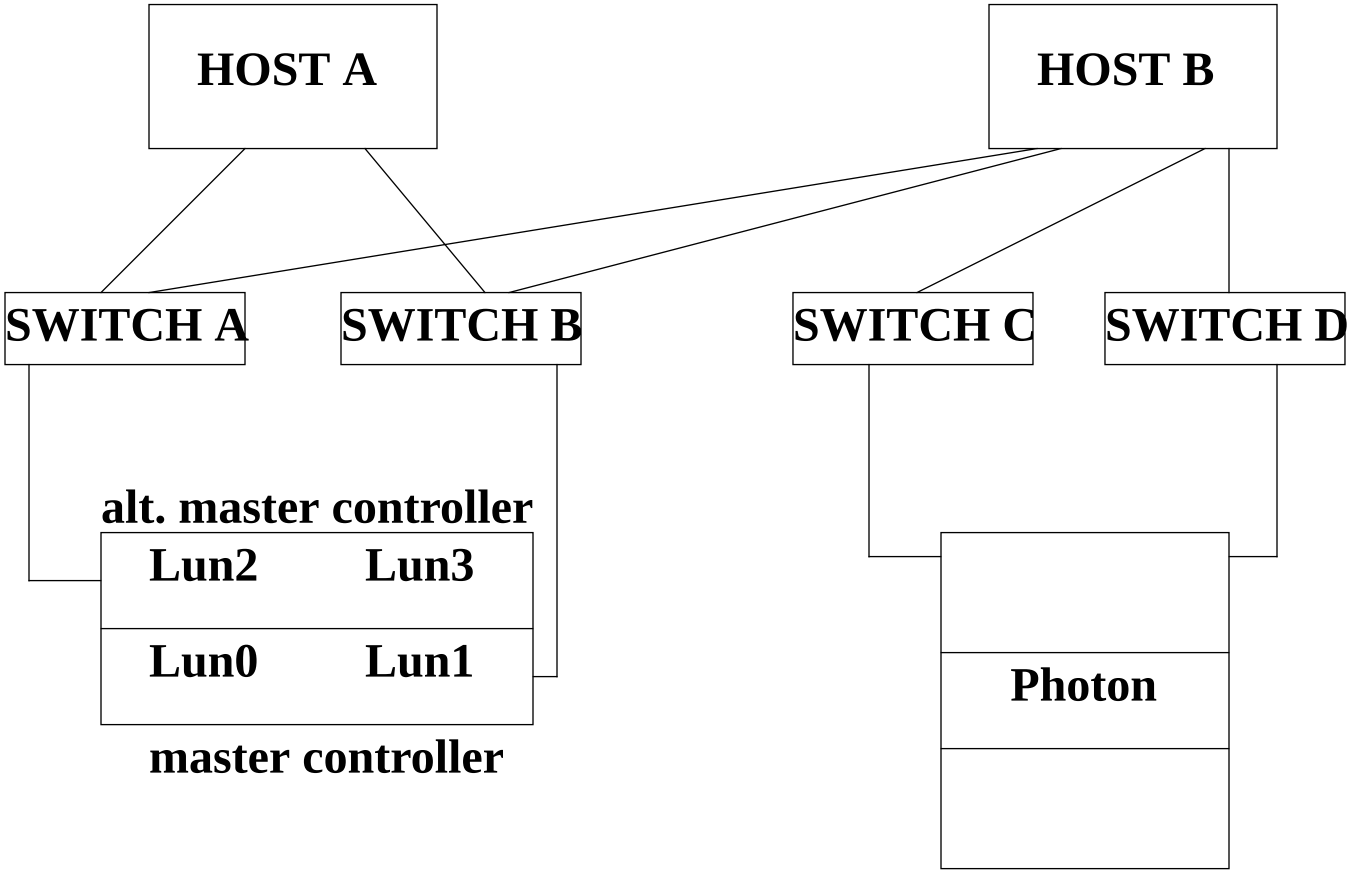
- Dynamic N-path multipathing with automatic discovery of new paths
- Support for T300 and A5k
- Automatic failover
- Target drivers: ssd and ses
- Enable/disable globally or per HBA
- libdevinfo(3), prtconf(1M) changes
- luxadm changes to display multipathing info and manual failover/failback
- cfgadm(1M) support (Tapestry)
- DR

scsi_vhci Tunable Parameters

- mpxio parameters may be configured via /kernel/drv/scsi_vhci.conf

```
#
# Copyright (c) 2000 by Sun Microsystems, Inc.
# All rights reserved.
#
#pragma ident   "@(#)scsi_vhci.conf     1.1     00/12/18 SMI"
#
name="scsi_vhci" class="root";
#
# mpxio Global enable/disable configuration
# possible values are mpxio-disable="no" or mpxio-disable="yes"
#
mpxio-disable="no";
# mpxio-disable="yes";
#
# Load Balancing global configuration
# possible values are load-balance="none" or load-balance="round-robin"
#
load-balance="round-robin";
```

Example Configuration



Concepts

Primary Path: Path to LUN through Controller that it resides on.

Secondary Path: Path to LUN through Alternate Controller.

Path States:

ONLINE: Path is available and will be used for I/O.
STANDBY: Path is available but will not be used for I/O.
OFFLINE: Path is unavailable.

Failover: Switch to STANDBY paths.

Failover

Automatic Failover: Happens when the last ONLINE path fails.

One way to initiate this type of failover is through a cable pull. The systems reports the cable pull by an OFFLINE event in /var/adm/messages.

After 90 seconds, the OFFLINE timeout occurs and initiates the automatic failovers, by switching to available STANDBY paths.

Manual Failover: User initiated via luxadm:

```
#luxadm failover primary /dev/rdsk/c6t60020F2000000023538B2952D0001BDA7d0s2
```

External Failover Detection: In a multihost environment, need to switch paths when failover initiated by some other host is detected.

Load Balancing

Round Robin:

Use all currently ONLINE paths in a round robin fashion

For Photons, all available paths will be ONLINE and will be used for load balancing

For T300s, only a subset of all the paths may be ONLINE

None:

Always use same path until it fails.

Tapestry

Tapestry project provides the Fiber Channel specific plug-in for cfgadm(1M)

Enumeration is based on Port WWN. In the mpxio environment, one can think of it as enumerating a “path”.

Scenario #1: Device has previously been enumerated through another path. mpxio framework will just add a new path to the existing device. If the newly enumerated path is ONLINE it may be used for I/Os. If path is STANDBY it may be used as a failover destination when required. The `luxadm display /dev/rdsk/c?t<guid>d0s?` command may be used to confirm the addition of the new path.

Scenario #2: Device has not been previously enumerated. This corresponds to enumerating the first path to a device. The mpxio device gets created as /dev/{r}dsk/c?t<guid>d0s? The `luxadm display <WWN>` command may be used to dislplay the device and path details

Troubleshooting

Check installation and configuration files:

- /kernel/drv/{sparcv9}/scsi_vhci and /kernel/misc/{sparcv9}/mpxio
- /etc/name_to_major must have a unique entry for scsi_vhci
- /kernel/drv/scsi_vhci.conf, /kernel/drv/qlc.conf, and /kernel/drv/ssd.conf

Ensure use of ambers and crystals+ only with firmware versions 1.10 and higher.
(else luxadm would not support it.)

Crystal+ Port 1 issue: Bugid#4438711 fails to issue offline event for cablepulls.
This causes failovers to fail. Shall be fixed by RR.

Purple must have firmware 1.17 and be configured for mpxio
“sys mp_support mpxio”

iostat and prtconf: These enhancements will not be part of our s8 patches for now.

The complete guide can be found at
/net/drv60.ebay/export/projects/PYTHON/doc/MPxIO_FAQ

References

PSARC materials: /net/sac.eng/export/SDF/sac/PSARC/1999/647/commit.materials

Unit Test Plan: http://webhome.ebay/hssdd/fcdrv/lab/testplans/mpxio.tstplan.html