#SHAREorg

SHARE
Technology · Connections · Results

Improving z/OS I/O Resiliency

Dale F. Riedy
IBM
riedy@us.ibm.com

7 August 2012
Session 11709

SHARE in Anaheim 2012

Legal Stuff

- Notice
  - IBM may have patents or pending patent applications covering subject matter described in this document. The furnishing of this document does not give you any license to these patents. You can send license inquiries, in writing to: *IBM Director of Licensing, IBM Corporation, North Castle Drive, Armonk, NY 10504-1785 U.S.A.*
  - Any references in this information to non-IBM Web sites are provied for convenience only and do not in any manner serve as an endorsement of those Web sites. The materials at those Web sites are not part of the materials for this IBM product and use of those Web sites is at your own risk.
- Trademarks
  - The following terms are trademarks of the International Business Machines Corporation in the United States, other countries, or both: FICON® IBM® Redbooks™ System z10™ z/OS® zSeries® z10™
  - Other Company, product, or service names may be trademarks or service marks of others.

Agenda

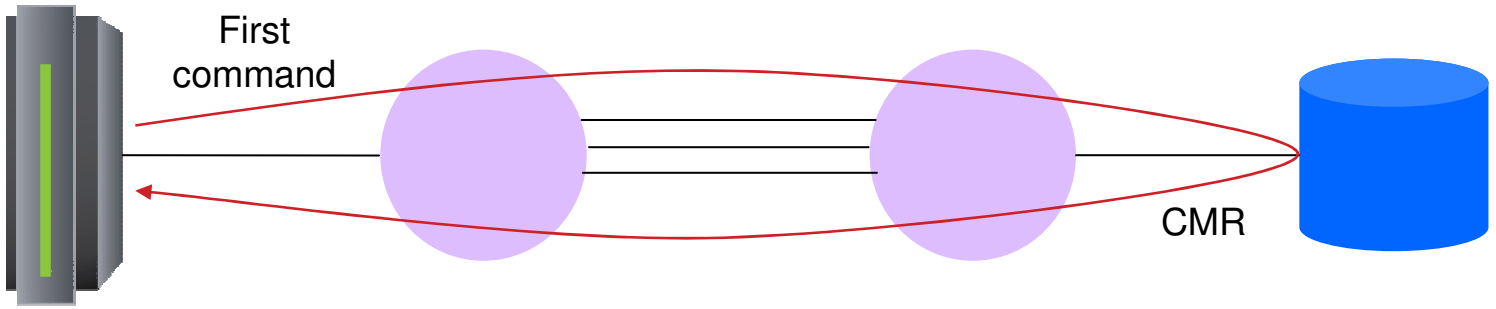
- CMR Time Health Check
- Improved Channel Path Recovery
- IPL from Alternate Subchannel Set
- IOSSPOFD Tool

Symptoms of a Path Related Problem

- Workloads are seeing unacceptable I/O service times
- RMF device activity report shows higher than normal I/O service times
- RMF I/O queuing report shows abnormally high initial command response time on a subset of the paths
- No single root cause has been identified
  - ISL failures, CU port congestion, CU HA utilization, control unit failures, wrong laser type, ports initialize at the wrong link speed, DWDM issues

What is Initial Command Response Time?

- Initial command response (CMR) time is the amount of time from when the channel sends the first command until it gets a response from the control unit
  - One round trip through the fabric
  - Good for detecting fabric congestion and other problems on a path



RMF I/O Queuing Report

z/OS V1R9 SYSTEM ID SYD1 DATE 09/17/2009 INTERVAL 09.59.990
RPT VERSION V1R8 RMF TIME 20.10.00 CYCLE 1.000 SECONDS
TOTAL SAMPLES = 600 IODF = A1 CR-DATE: 09/16/2009 CR-TIME: 15.57.12 ACT: ACTIVATE

LCU	CU	DCM GROUP MIN	MAX	DEF	CHAN PATHS	CHPID TAKEN	% DP BUSY	% CU BUSY	AVG CUB DLY	AVG CMR DLY	CONTENTION RATE	DELAY Q LNGTH	AVG CSS DLY
0008	03F0				**2B**	**0.012**	**0.00**	**0.00**	**0.0**	**6.7**			
					76	0.013	0.00	0.00	0.0	0.1			
					36	0.015	0.00	0.00	0.0	0.1			
					6C	0.013	0.00	0.00	0.0	0.3			
					B4	0.012	0.00	0.00	0.0	0.1			
					C6	0.012	0.00	0.00	0.0	0.1			
					46	**0.008**	**0.00**	**0.00**	**0.0**	**3.8**			
					47	0.008	0.00	0.00	0.0	0.2			
					*	0.093	0.00	0.00	0.0	1.3	0.000	0.00	0.1
0009	0434				**2B**	**0.007**	**0.00**	**0.00**	**0.0**	**4.2**			
					76	0.007	0.00	0.00	0.0	0.2			
					36	0.005	0.00	0.00	0.0	0.1			
					6C	0.008	0.00	0.00	0.0	0.1			
					B4	0.008	0.00	0.00	0.0	0.1			
					C6	0.010	0.00	0.00	0.0	0.1			
					46	**0.007**	**0.00**	**0.00**	**0.0**	**4.2**			
					47	0.005	0.00	0.00	0.0	0.2			
					*	0.057	0.00	0.00	0.0	1.1	0.000	0.00	0.1

CMR Health Check

- New I/O related health check that provides real time detection of mismatched CMR times, which is a symptom of fabric congestion and other problems
  - OA33367 – z/OS 1.10 and up, available in z/OS 1.13 base
  - IOS_CMRTIME_MONITOR, enabled by default
  - Default: run every 5 minutes
- Notify you when a problem is detected
- No other action taken by the health check

CMR Health Check Parameters

- Threshold
  - The path with the highest average CMR time must be greater than this value before z/OS checks for a CMR time mismatch
  - Values – 0 to 100, default = 3 (specified in ms)
- Ratio
  - The path with the highest average CMR time must be “ratio” times greater than the path with lowest CMR time before an exception is reported.
  - Values – 2 to 100, default = 5
- XCU – control unit numbers to be excluded
- XTYPE – device types to be excluded (DASD or TAPE)

Parameter Examples

Threshold	Ratio	CMR Times	Results
10	5	Path 1: 10 Path 2: 1	No exception is reported since path 1's CMR time is not higher than the threshold of 10 ms.
10	5	Path 1: 12 Path 2: 3	Although path 1 is over the threshold, no exception reported since it is not more than 5 times higher than path 2's CMR time,
10	5	Path 1: 11 Path 2: 2	Exception reported since path 1's CMR time is more than 5 times higher than path 2's CMR time.
0	5	Path 1: 5 Path 2: 1	No exception is reported since path 1's CMR time is not more than 5 times higher than path 2's CMR time.
0	5	Path 1: 5.1 Path 2: 1	Exception reported since path 1's CMR time is more than 5 times higher than path 2's CMR time.

CMR Health Check Report Example

```
CHECK(IBMIOS,IOS_CMRTIME_MONITOR)
START TIME: 12/10/2011 16:34:03.455536
CHECK DATE: 20100501  CHECK SEVERITY: MEDIUM
CHECK PARM: THRESHOLD(3),RATIO(5),XTYPE(),XCU()

IOSHC113I Command Response Time Report

The following control units show inconsistent average command response
(CMR) time based on these parameters:
THRESHOLD = 3
RATIO = 5

CMR TIME EXCEPTION DETECTED AT: 12/10/2011 16:29:24.212239
CONTROL UNIT = 25C0
ND = 002107.941.IBM.75.0000000WH391

          ENTRY EXIT  CU    I/O     AVG
  CHPID   LINK  LINK  INTF  RATE    CMR
   81     2C51  2DC4  0030  72.330  9.21
   22     3C1B  3DC2  0031  71.651  9.47
   82     2C52  2DC0  0032  72.333  8.70
   84     2C54  2DCC  0100  71.810  1.92
   21     3C19  3DD2  0231  72.122  1.79
* Medium Severity Exception *

IOSHC112E Analysis of command response (CMR) time detected one or
more control units with an exception.
```

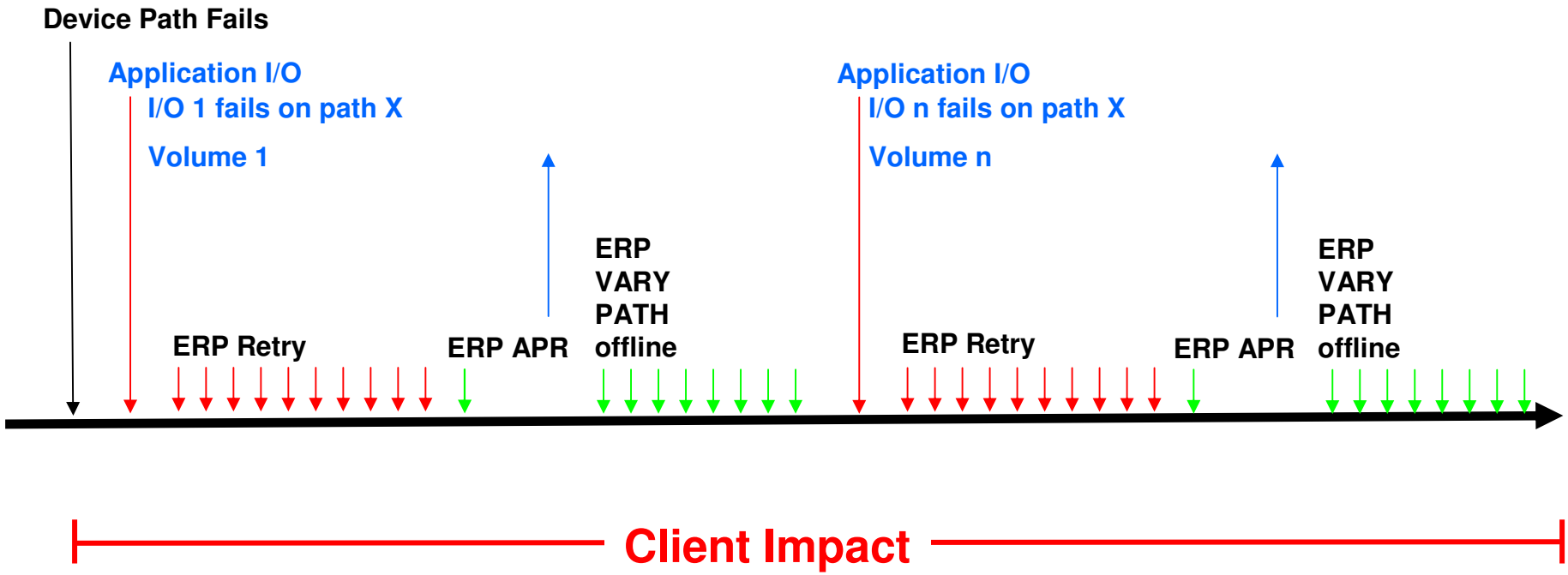
These are the exception paths

Exception message appears in system log

Agenda

- CMR Time Health Check
- Improved Channel Path Recovery
- IPL from Alternate Subchannel Set
- IOSSPOFD Tool

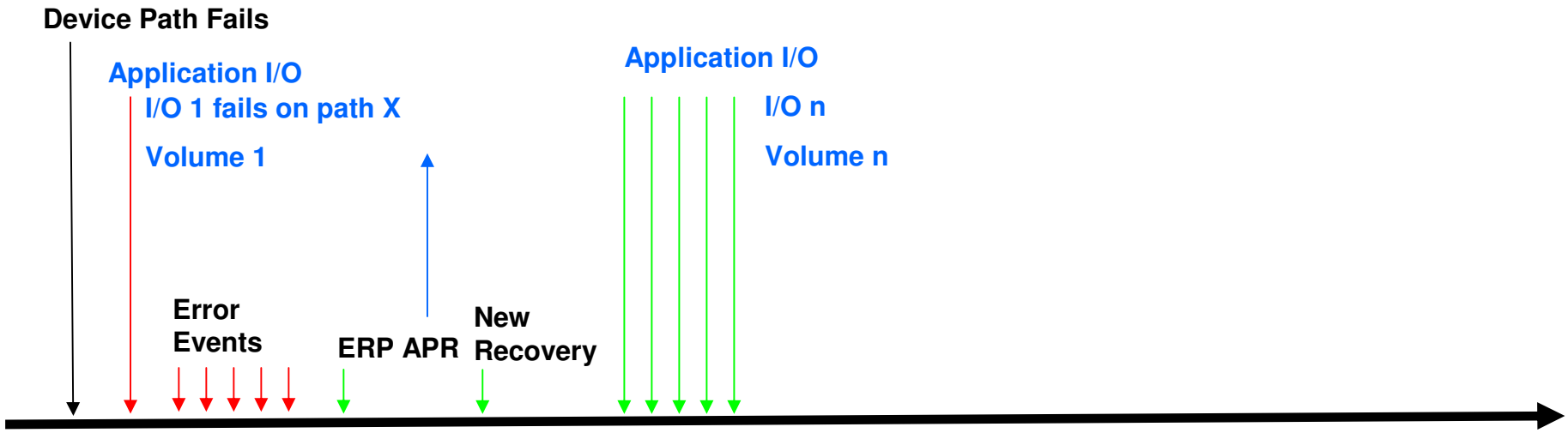
I/O Recovery for Failing Path - Before



Accelerated Device Path Recovery

- Improved system resilience for H/W errors
- Clients would rather see path taken offline than continue to cause problems (e.g., link thresholding support on z9)
  - IOS recovery delays application I/O even when there are other paths
  - Avoid needing to manually take paths offline or via automation
- In particular:
  - IFCC and other path error thresholding
  - Proactively removing a path from all devices in an LCU
- DASD and tape only

I/O Recovery for Failing Path - After



Parmlib and Command Changes

- New IECIOSxx parmlib and SETIOS commands to enable the new function

```
RECOVERY,PATH_SCOPE={DEVICE|CU}
         [,PATH_INTERVAL=nn]
         [,PATH_THRESHOLD=nnn]
```

  - New display IOS command to display the status:

```
D IOS,RECOVERY
IOS103I hh.mm.ss RECOVERY OPTIONS
LIMITED RECOVERY FUNCTION IS DISABLED
PATH RECOVERY SCOPE IS BY CU
PATH RECOVERY INTERVAL IS nn MINUTES
PATH RECOVERY THRESHOLD IS nnn ERRORS
```

IFCC Thresholding

- Remove path for intermittent errors
- Default: at least 10 IFCCs per minute (PATH_THRESHOLD) over a 10 minute period (PATH_INTERVAL)
- Remove the path from all devices in the LCU
- ERP path related error monitoring

IOS050I CHANNEL DETECTED ERROR ON dddd,yy,op,stat,
PCHID=pppp

IOS210I PATH RECOVERY INITIATED FOR PATH pp ON CU cccc,
REASON=PATH ERROR THRESHOLD REACHED

Proactively Removing Paths – Dynamic Pathing Validation

- Dynamic Pathing Validation issues I/Os down each path to test state of the path group
- If error occurs, path is removed from device
- Each device trips over the error
- If PATH_SCOPE=CU, do all devices in LCU

IOS051I INTERFACE TIMEOUT DETECTED ON ON dddd,yy,op,stat,
PCHID=pppp
IOS071I dddd,cc,jjjjjjjj, START PENDING

IOS450E dddd, cc NOT OPERATIONAL PATH TAKEN OFFLINE

IOS210I PATH RECOVERY INITIATED FOR PATH pp ON CU cccc,
REASON=DYNAMIC PATHING ERROR

Proactively Removing Paths – Link Threshold Exceeded

- Each device trips over the link threshold condition
- Stray I/O may interfere recovery after customer fixes the problem
- If PATH_SCOPE=CU, do all devices in LCU

```
IOS001E dddd,INOPERATIVE PATHS pp pp pp
IOS2001I dddd,INOPERATIVE PATHS
      STATUS FOR PATH(S) pp,pp,pp....
      LOGICAL PATH IS REMOVED OR NOT ESTABLISHED (A0)
      LINK RECOVERY THRESHOLD EXCEEDED FOR LOGICAL PATH (06)
IOS210I PATH RECOVERY INITIATED FOR PATH pp ON CU cccc,
      REASON=LINK THRESHOLD EXCEEDED
```

D M=DEV(devno,(chp))

```
D M=DEV(410,(48))
IEE174I hh.mm.ss DISPLAY M idr
 DEVICE 0410    STATUS=ONLINE
 CHP                      48
 ENTRY LINK ADDRESS       22
 DEST LINK ADDRESS        E0
 PATH ONLINE              N
 CHP PHYSICALLY ONLINE    Y
          ...
 PATH OFFLINE DUE TO THE FOLLOWING REASON(S):
              [PATH RECOVERY ERROR]
              [BY OPERATOR]
              [CONTROL UNIT INITIATED RECOVERY]
              [CONFIGURATION MANAGER]
```

Identifying Detecting H/W Components

- When an error occurs, it is difficult to determine where the failing or misbehaving component is:
  - Channel, switch(es), CU interface, links
- Identify detecting component based on H/W logout data
- Not controlled by PATH_SCOPE option

IOS050I CHANNEL DETECTED ERROR ON ddddd,yy,op,stat,
PCHID=pppp

IOS054I ddddd,pp ERRORS DETECTED BY comp, comp,…

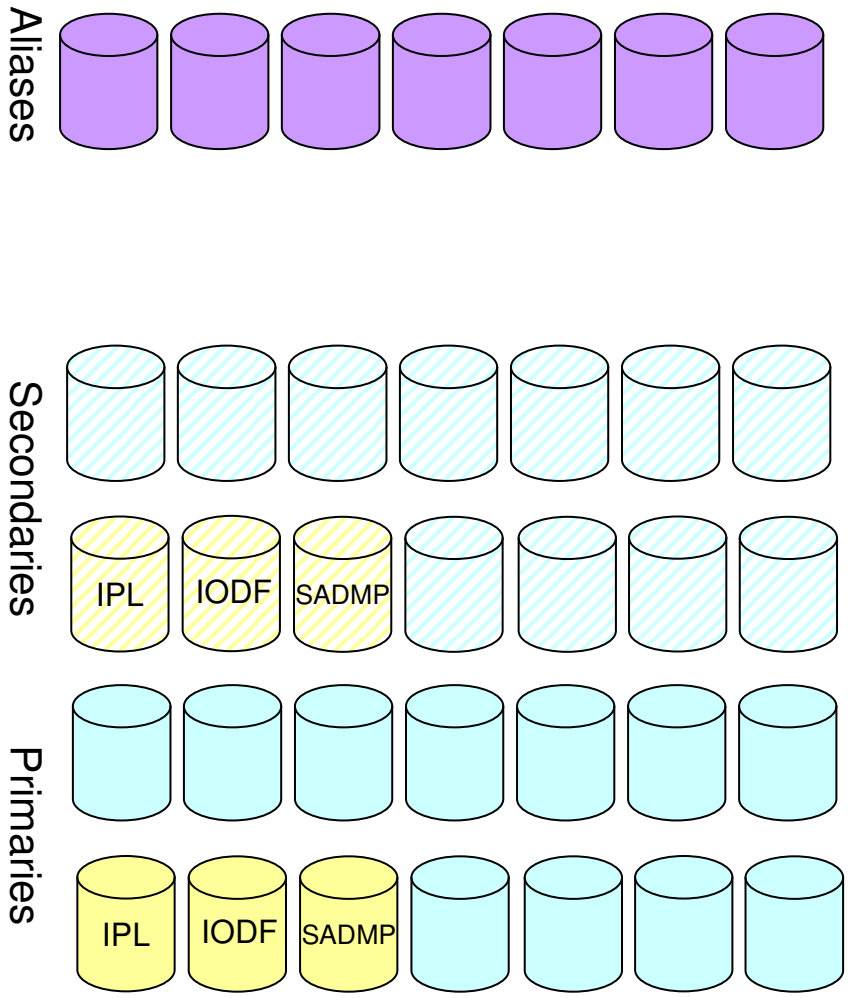
Where *comp* is one or more of the following:

CHANNEL, CHAN SWITCH PORT, CU SWITCH PORT, CONTROL UNIT

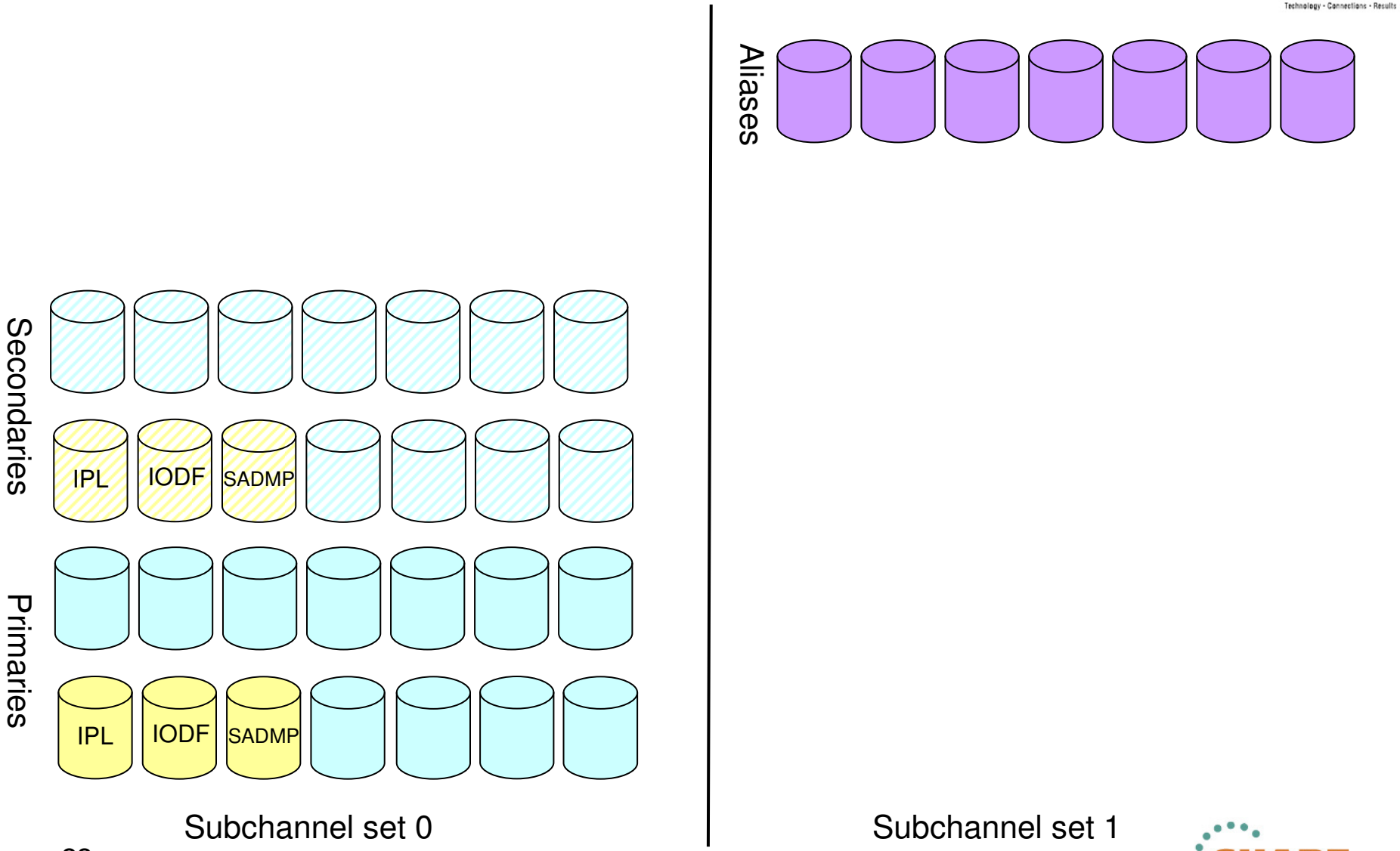
Agenda

- CMR Time Health Check
- Improved Channel Path Recovery
- IPL from Alternate Subchannel Set
- IOSSPOFD Tool

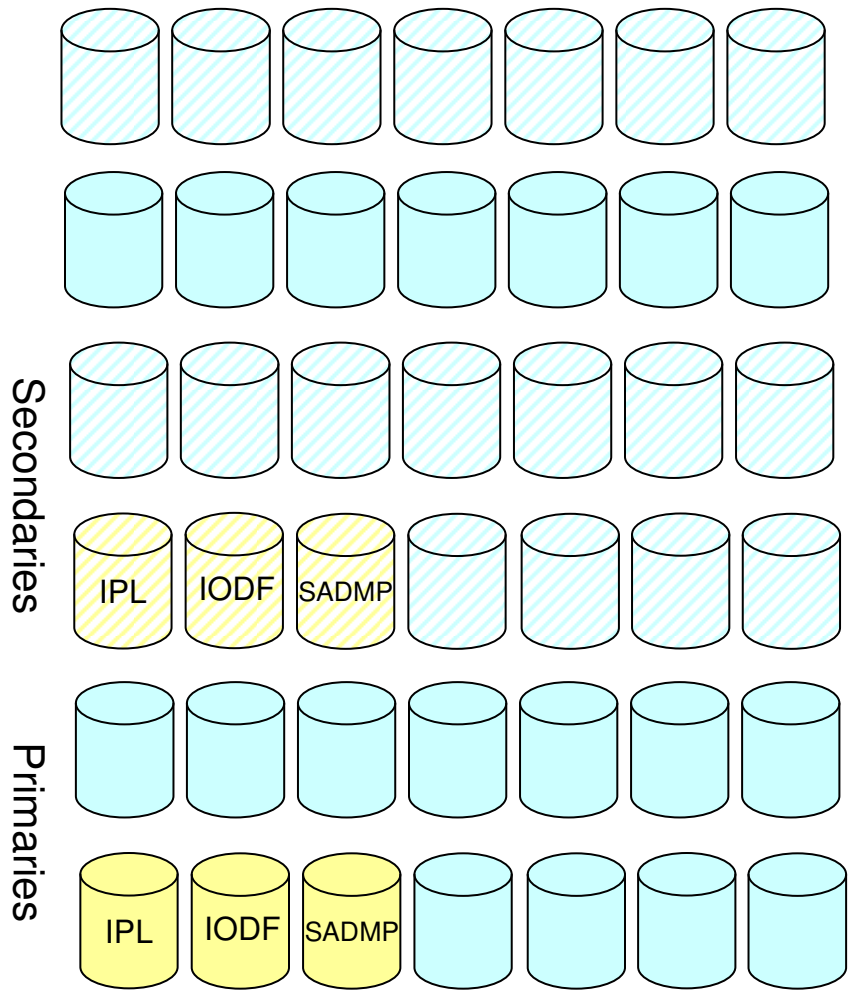
Device Number Constraint Relief

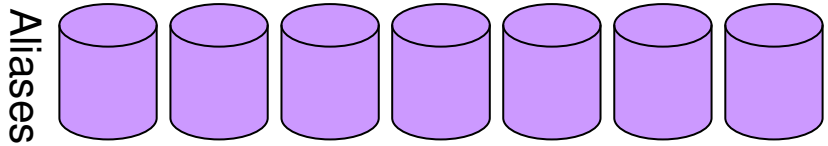


Device Number Constraint Relief



Device Number Constraint Relief

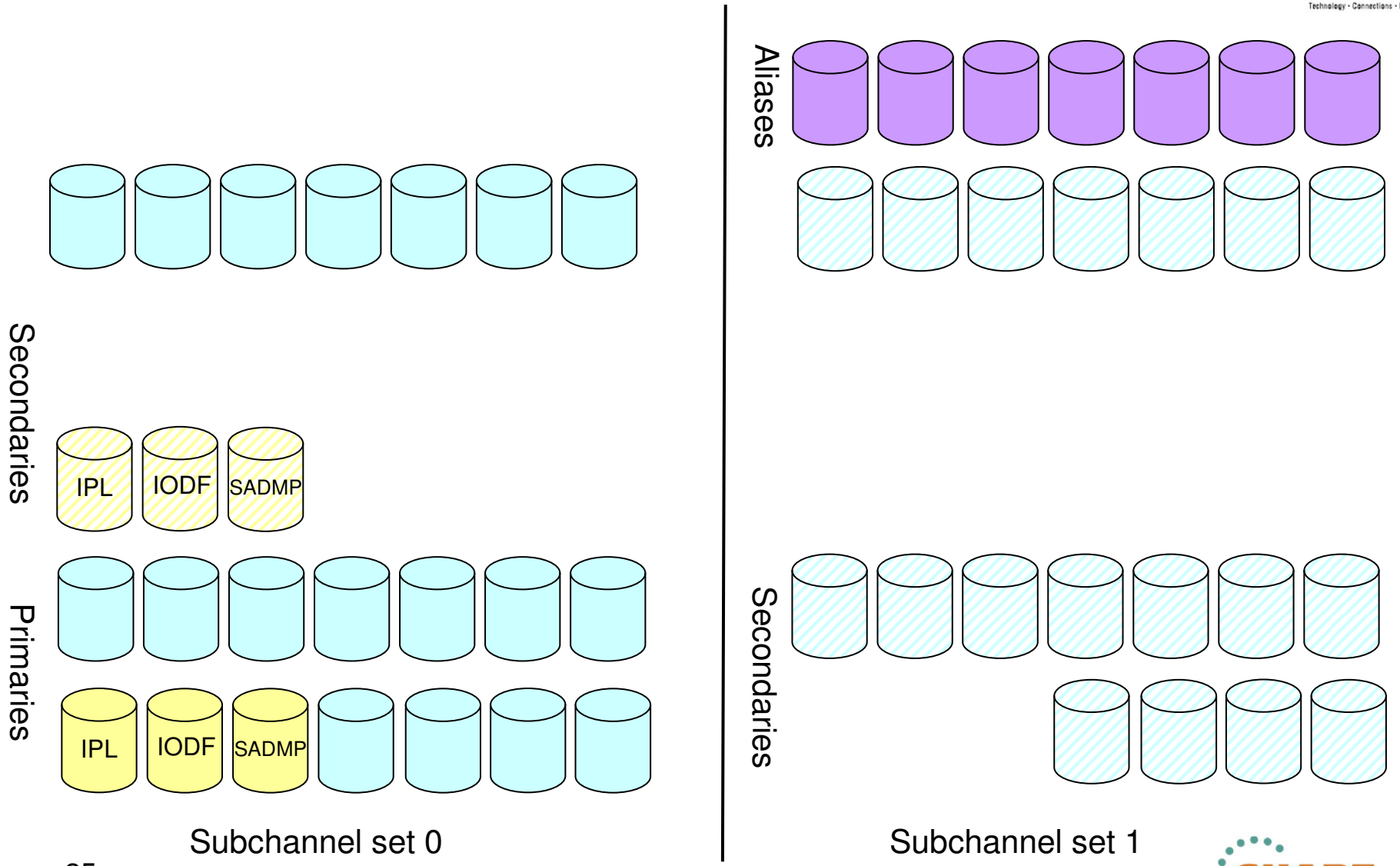




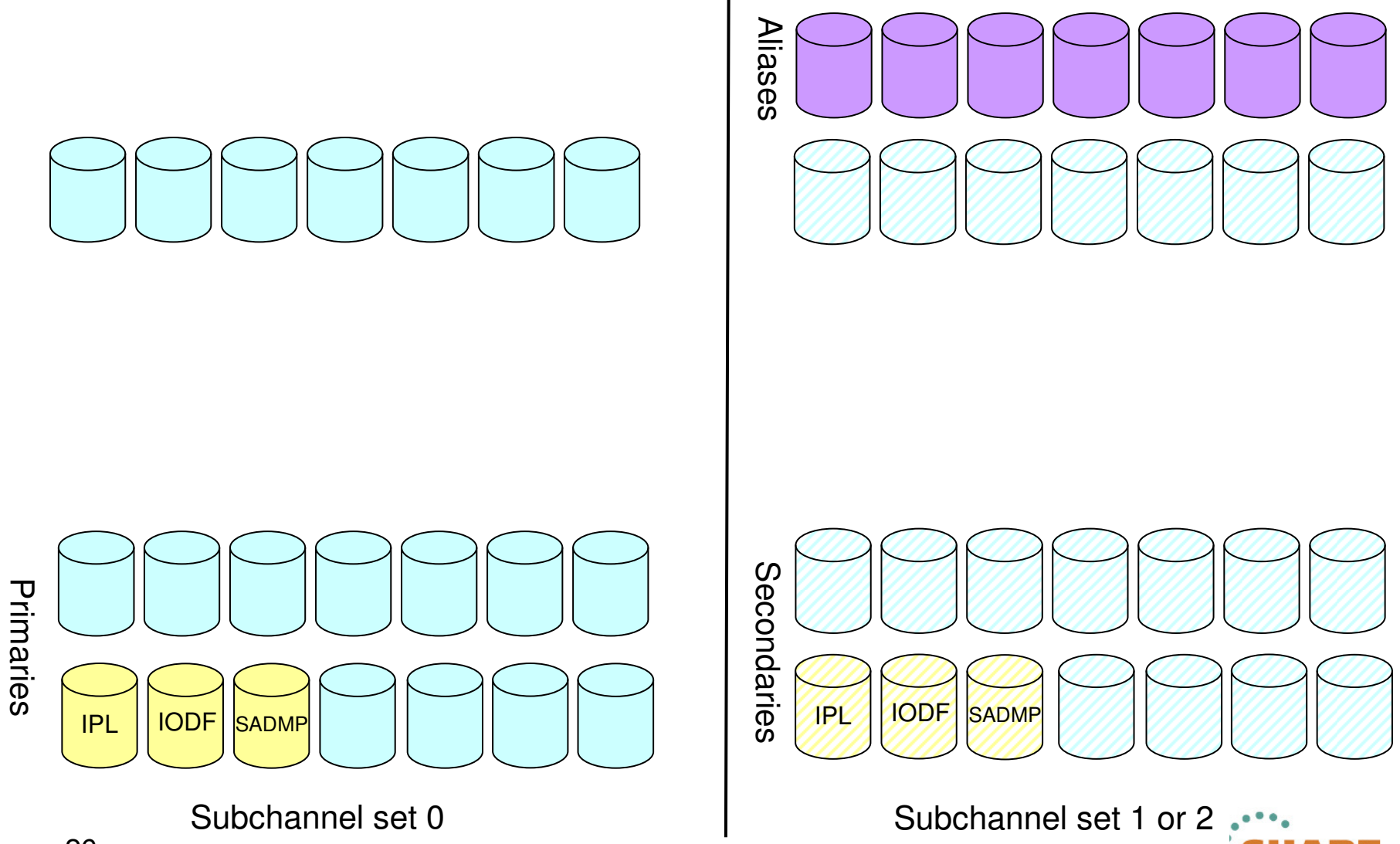
Subchannel set 0

Subchannel set 1

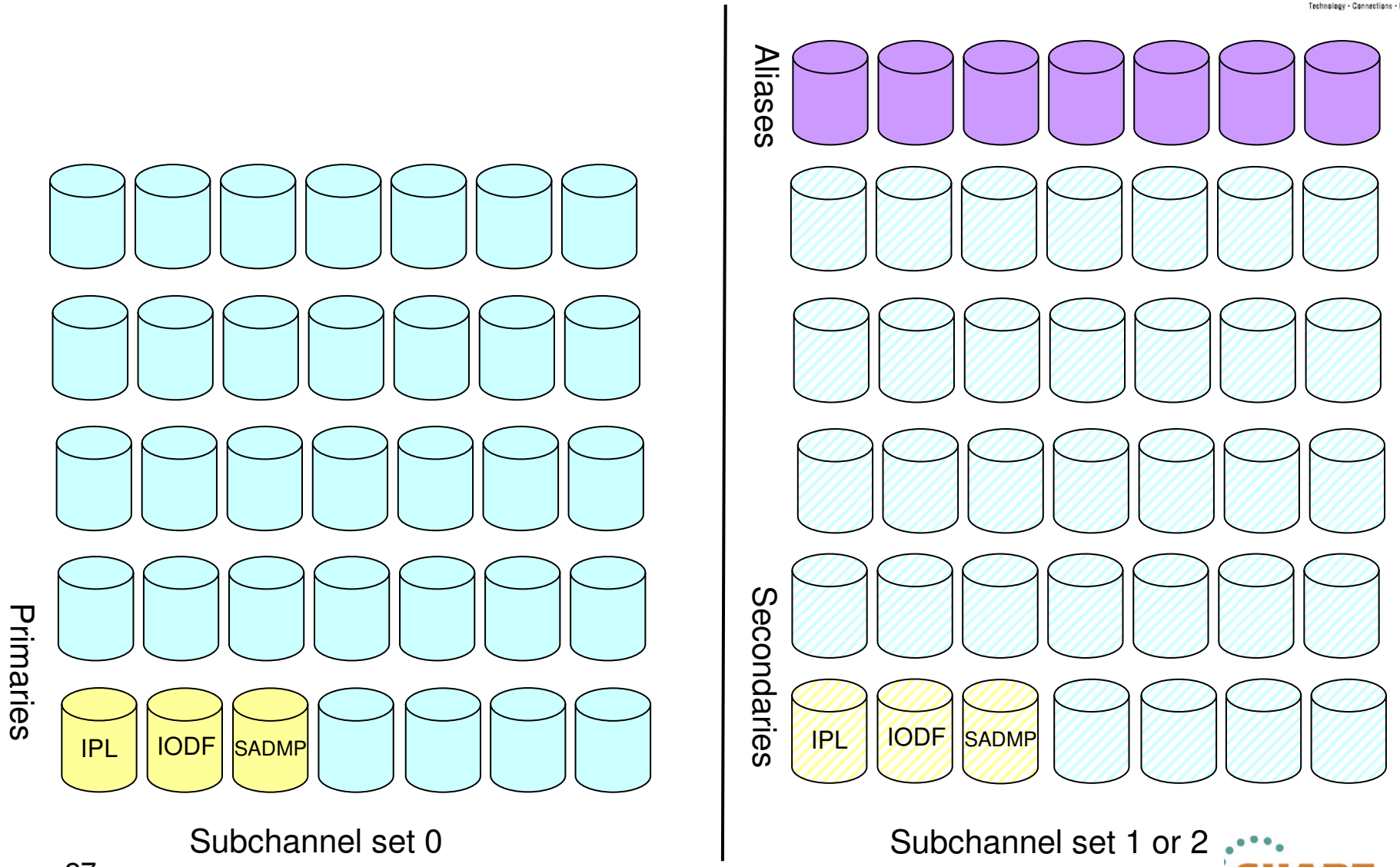
Device Number Constraint Relief



Device Number Constraint Relief



Device Number Constraint Relief



Using the Alternate Subchannel Set for Secondary Devices

- z/OS 1.10 and APAR OA24142 introduced the ability to define your secondary PPRC devices in the alternate subchannel set
- Benefits:
  - Makes room for more primary devices in subchannel set zero
  - Eliminates the need to have a separate OS config in the IODF depending on which set of devices you are using
- Secondaries are defined as “special” 3390D devices
  - Secondary device must have the same 4 digit device number as the primary device
  - Subchannel set is transparent to device allocation, most operator commands, and parmlib
  - Mirroring must be going in the same direction (e.g., 0->1 or 1->0)

Defining Special Secondary Devices

```
                              Add Device

Specify or revise the following values.

Device number  . . . . . . . . 1400  + (0000 - FFFF)
Number of devices  . . . . . . 1___
Device type  . . . . . . . . . 3390d_______  +

Serial number  . . . . . . . . __________
Description  . . . . . . . . . ________________________________

Volume serial number . . . . . ______  (for DASD)

PPRC usage . . . . . . . . . . _  + (for DASD)

Connected to CUs . . ____  ____  ____  ____  ____  ____  ____  ____  +
```

```
                        Specify Subchannel Set ID

Specify the ID of the subchannel set into which devices are placed,
then press Enter.

Configuration ID . : GENFT          AQFT
Device number  . . : 1400           Number of devices  : 1
Device type  . . . : 3390D

Subchannel Set ID    1  +
```

Specifying the Subchannel Set to Use

LOADxx Member

```
*          IODF Suffix
*          |   IODF HLQ OS cfg         Schset
*          |   |          |              |
*          V   V          V              V
*---+----1----+----2----+----3----+----4----+----5--
IODF       8C IOSTOOL   CONFIG01 00 Y 0
```

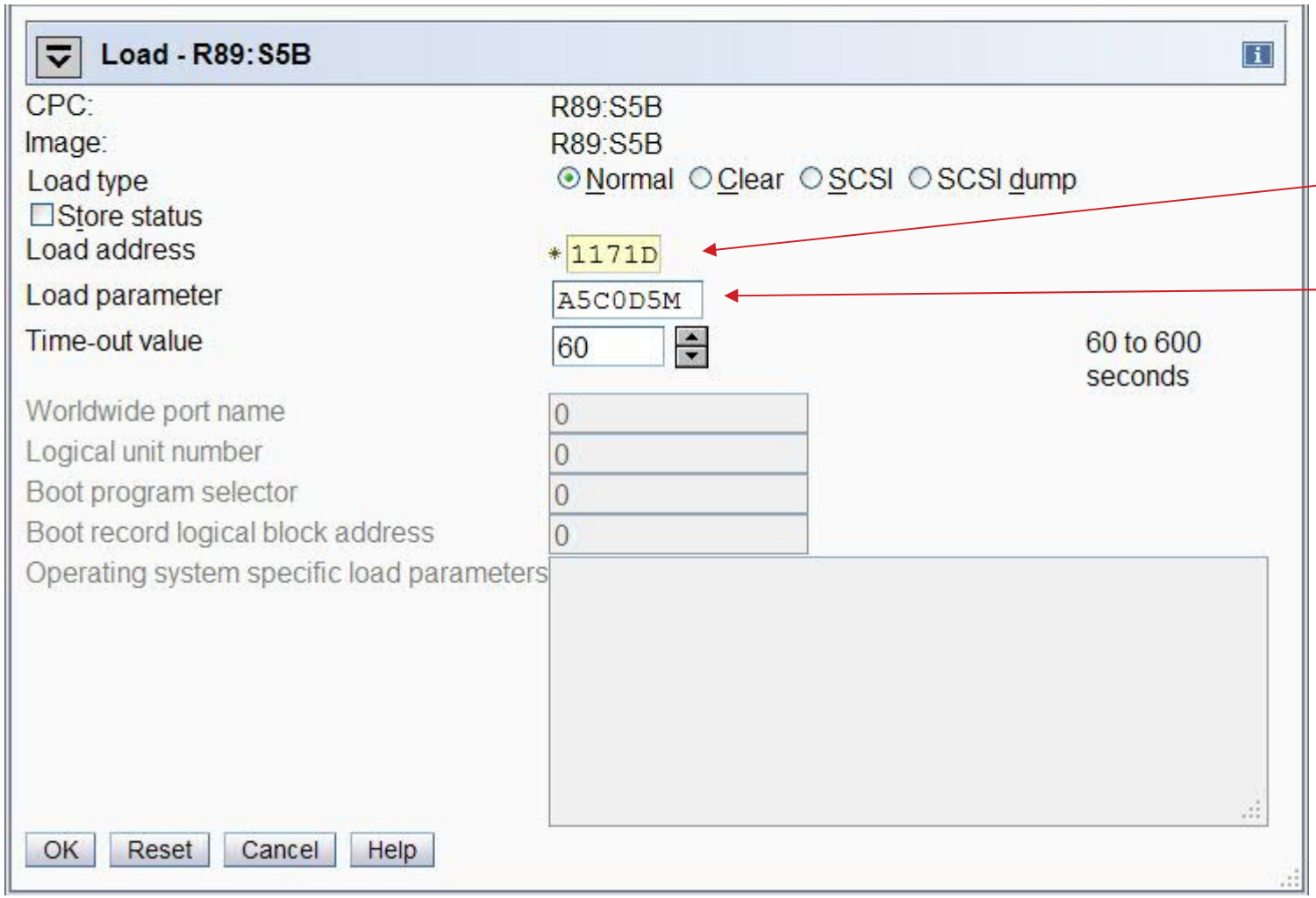
…or…

IEA111D SPECIFY SUBCHANNEL SET TO BE USED FOR DEVICES THAT ARE ACCESSIBLE FROM MULTIPLE SUBCHANNEL SETS - REPLY SCHSET=X

IPL from Alternate Subchannel Set

- Issues
  - The original support did not include the ability to put the PPRC secondaries for IPL (SYSRES) and IODF devices in the alternate subchannel set
    - The secondary devices still had to be in subchannel set 0
- Solution
  - z196 GA2 allows a 5 digit number to be specified for the load device on the HMC
    - z/OS 1.13 base
    - z/OS 1.11 and 1.12 with APARs OA35135, OA35136, OA35137, OA35139 and OA35140

HMC Image Profile – Load Information



5 digit IPL device

4 digit IODF device number

LOADxx Changes

```
*          IODF Suffix
*          |   IODF HLQ OS cfg          Schset
*          |   |          |               |
*          V   V          V               V
*---+-----1-----+-----2----+----3----+----4----+----5--
IODF       8C  IOSTOOL   CONFIG01 00  Y  *
```

Indicates to use subchannel set id of IPL device for other devices

AutoIPL/DIAGxx Changes

- DIAGxx AUTOIPL statement allows an "*" to prefix the device numbers specified for SADMP and IPL devices. The asterisk signifies that the currently active subchannel set should be used.
  - *AUTOIPL SADMP(*0180,SP03E0) MVS(*0181,0181MG)*
  - *AUTOIPL MVS(LAST) is unchanged*
- D DIAG/IGV007I
  - Asterisk shown if specified for SADMP or IPL device
    - *AUTOIPL SADMP(*0180,SP03E0) MVS(*0181,0181MG)*
  - If MVS(LAST) specified, device number of currently active IPL volume is shown, prefixed with asterisk
    - AUTOIPL SADMP(NONE) MVS(*0980,0181MG)

Standalone Dump

- SADMP IPL and output devices can be in an alternate subchannel set
  - SADMP generation not updated to allow 5 digit device numbers for output data set
  - Subchannel set id is inherited from the IPL device, for DASD only
  - If no output device in the IPL device subchannel set, use subchannel set 0
- Advantages:
  - Assuming PPRC is used for SADMP devices, only have to generate one copy of the SADMP program and output data sets

Standalone Dump

- SADMP start up message was changed to display the subchannel set id used
- Other SADMP messages were not changed to show the subchannel set id

AMD083I STAND-ALONE DUMP INITIALIZED. SCHSET: s IPLDEV: dddd
LOADP: pppppppp

Agenda

- CMR Time Health Check
- Improved Channel Path Recovery
- IPL from Alternate Subchannel Set
- IOSSPOFD Tool

z/OS Single Point of Failure Service

- z/OS 1.10 introduced IOSSPOF service which allows you to check for single points of failure (SPOFs)
  - Check for SPOFs for a specific device
  - Check for common SPOFs between two devices
    - E.g., primary and backup XCF couple data sets
- Examples:
  - Only one online path to the device
  - All online paths go through the same switch
  - All online paths are connected to the same port or host adapter card on the control unit

z/OS Single Point of Failure Service

- SPOF messages written to the programmer/job log or included as part of a health check
  - XCF_CDS_SPOF – Check XCF couple data sets for SPOFs

IOSPF251I Volumes WLMPKP (0485) and WLMPKA (0486) share a logical subsystem.

IOSPF203I Volume WLMPKP (0485) has only one online path

IOSPF253I Volumes LOGPKP (0487) and LOGPKA (0488) share the same physical control unit.

IOSPF253I Volumes FDSPKP (0489) and FDSPKA (048A) have all paths share the same switch.

IOSSPOFD Tool

- Allows you to check for single points of failure in your own configuration
- Run as a batch job, invoked from a program, CLIST or REXX exec
- Input is a list of device numbers, volsers, or data set names
- Uses the IOSSPOF service to check for single points of failure and generate messages
- Available at z/OS tools and toys website
  - http://www-03.ibm.com/systems/z/os/zos/features/unix/bpxa1ty2.html

IOSSPOFD Input (SYSIN DD)

- Checking individual devices for single points of failure
  - DEVLIST(410,411,980-9A0)
  - VOLLIST(SYSRES,WORK*,TEST01)
  - DSNLIST(SYS1.NUCLEUS,SYS1.LINKLIB,DB2.DATABASE)

- Checking pairs of devices for single points of failure between them
  - DEVN1(0410) DEVN2(1410)
  - VOLSER1(RACFPM) VOLSER2(RACFAL)
  - DSN1(SYS1.RACF.PRIMARY) DSN2(SYS1.RACF.ALT)
  - IND_CHECKS(YES|NO)

Sample Output

```
Input: DSNLIST(SYS1.NUCLEUS,SYS1.LINKLIB,DB2.DATABASE)

IOSPF303I Volume SYSRES (0980) with SYS1.NUCLEUS has only
one online path.
IOSPF303I Volume SYSRES (0980) with SYS1.LINKLIB has only
one online path.
IOSPF301I Volume *NONE* with DB2.DATABASE could not be
found
+SPOFD001I RTC: 00000008  RSN: 00000000
```

```
Input: VOLSER1(PRMRY) VOLSER2(ALT) IND_CHECKS(YES)

IOSPF253I Volumes PRMRY (0980) and ALT (0981) share the
same physical control unit.

IOSPF203I Volume PRMRY has only one online path.

+SPOFD001I RTC: 00000008  RSN: 00000000
```

Thank you