HOT TOPICS

zHyperLink Write Support

zHyperLink technology now provides update write capability for both simplex and metro mirror environments.

By: Tariq Hanif, Brian Lee, Dale Riedy
Published: November 19th, 2021
Read time: 5 minutes

Overview

I/O latency as viewed by the application is more than just I/O service time. It also includes interrupt processing time, time waiting to get dispatched on a CPU after an I/O has completed, and delays caused by having to reload the CPU cache. Traditional I/O to DASD is asynchronous, which means that the application requests a record, the operating system initiates an I/O request, and the application is suspended until the I/O completes and the application is redispatched on a CPU. zHyperLink I/O is synchronous, which means that the CPU spins waiting for the I/O request to complete. This eliminates the delays normally associated with asynchronous I/O.

I/Os that are good candidates for zHyperLink are those where the application is suspended until the I/O completes. Two examples are Db2 synchronous database reads that occur when the application requests a record that is not in the buffer pool and Db2 log writes issued at transaction commit time. I/Os that are not good candidates for zHyperLink are reads used for prefetching or read-ahead and delayed writes because both are performed asynchronous to the execution of the application.

zHyperLink write requests have an additional requirement: the write I/Os must follow a log write pattern, which means that the writes must always go forward in the data set without skipping any records. The same record may be written multiple times, but the writes never go backwards. When you are finished with the last record in the data set, you can restart at the beginning of the data set.

In 2019, zHyperLink write support was introduced for simplex and local Metro Mirror environments (see Figure 1). In Metro Mirror configurations all DS8000s must be within 150m of the Z server and connected via zHyperlink.

In 2020 for z/OS V2R3 and above, zHyperLink write support was provided for DS8000 Global Mirror (asynchronous mirroring) configurations (see Figure 2 for an example). This does not include support for Extended Remote Copy (XRC).

Db2 for z/OS currently exploits zHyperLink writes for active log data sets. Media Manager support for dual logging provides the ability for Db2 to issue zHyperLink writes to two active log data sets simultaneously[1].

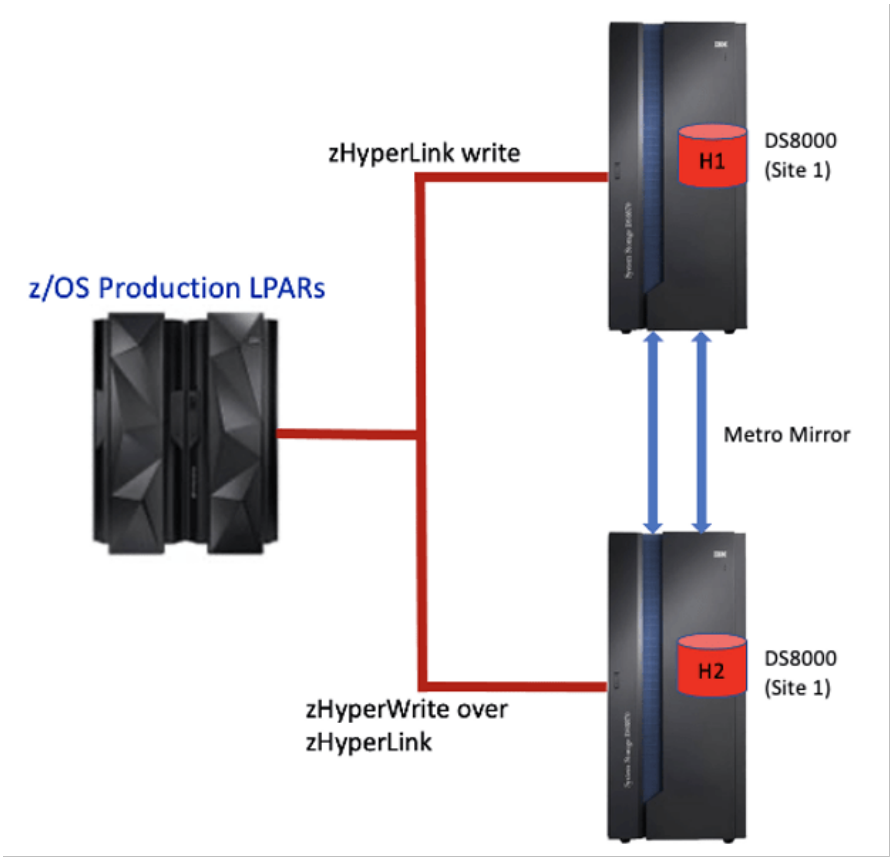

Figure 1: zHyperLink writes in a Metro Mirror environment.

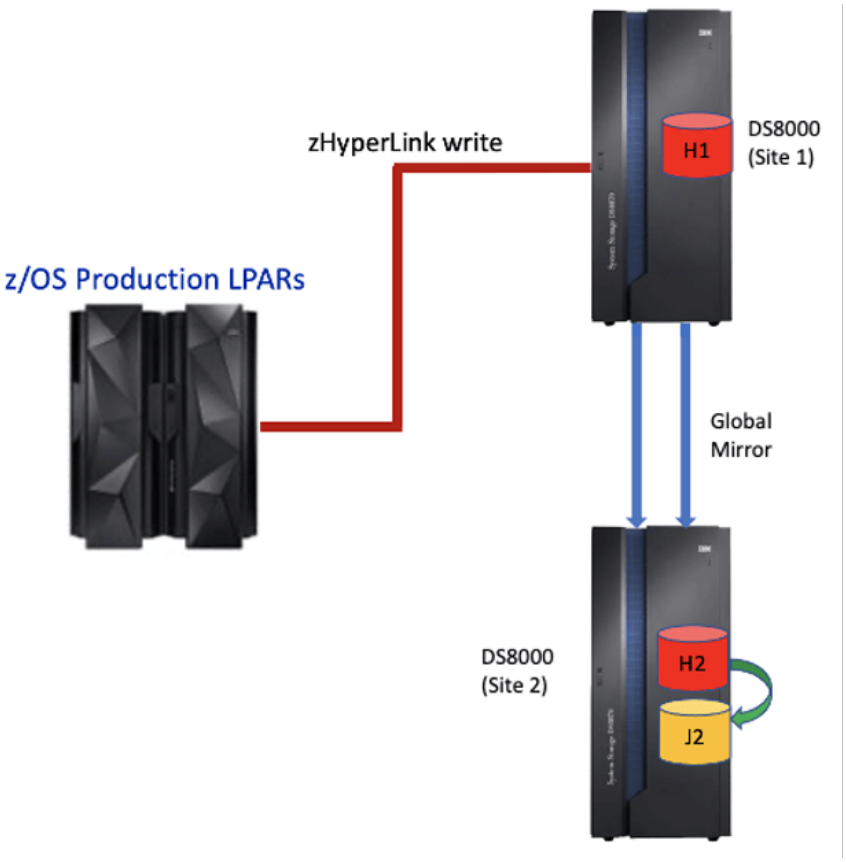

Figure 2: zHyperLink writes in a Global Mirror environment.

Hardware and I/O Configuration Requirements

zHyperLink is a point-to-point connection between the IBM Z processor and a DS8000. The maximum distance is 150 meters. To use zHyperLink, you need one or more zHyperLink Express features (cards) in the IBM Z processor. Each feature supports up to 2 ports and up to 16 features (32 ports) can be installed.

On the DS8000 side, each zHyperLink connection plugs directly into a port in the I/O enclosure; a special host adapter card is not required. Up to 2 connections per I/O enclosure is supported. The maximum number of connections supported is model dependent and is based on the number of enclosures and cores. zHyperLink connectivity is also required to the secondary for write support in Metro Mirror

environments, which means both primary and secondary must be within 150m of the IBM Z processor.

FICON connectivity to the DS8000 is still required for initialization for performing I/Os that are not eligible for zHyperLink and as a fallback if the zHyperLink I/O fails. Using zHyperLink does not require you to reduce the maximum number of FICON channel paths configured for a device.

In the IBM Z I/O configuration, define the PCIe function definitions (PFIDs) for the zHyperLink ports. The association between the zHyperLink port and DS8000 is discovered dynamically. Define 4 PFIDs per zHyperLink port for each logical partition that is sharing the DS8000.

Software Requirements

Find the PTFs for zHyperLink by searching for Fix Category IBM.Function.zHyperLink, APAR keyword HYPERL/K.

Enabling zHyperLink Write on z/OS

On a z/OS LPAR, zHyperLink is disabled by default. To enable zHyperLink write processing during IPL time, include the following in the IECIOSxx parmlib member[2]:

ZHPF=YES, if zHPF is not already enabled on the LPAR.

ZHYPERLINK,OPER=WRITE or OPER=ALL. If ALL is defined, both read and write processing is enabled.

zHyperLink write can also be enabled dynamically on a z/OS LPAR using these z/OS commands[3]:

SETIOS ZHPF=YES, if zHPF is not already enabled on the LPAR.

SETIOS ZHYPERLINK,OPER=WRITE or OPER=ALL.

To verify if zHyperLink support is enabled, issue:

- DISPLAY IOS,ZHYPERLINK to display whether zHyperLink is enabled for writes for the system (see Figure 3).
- DISPLAY M=DEV(devno) to display whether zHyperLink is enabled for writes for a device (see Figure 4).
- DISPLAY M=DEV(devno),ZHYPERLINK to display whether zHyperLink is enabled for writes for a device. If zHyperLink is disabled, the reasons why zHyperLink is disabled are displayed.
- DISPLAY M=CU(cuno) to show whether zHyperLink is enabled for writes at the control unit level and to show the state of the PFIDs (see Figure 5).

```
D IOS,ZHYPERLINK
IOS634I 16.12.16 IOS SYSTEM OPTION 061
ZHYPERLINK IS ENABLED FOR READ AND WRITE OPERATIONS
```

Figure 3: DISPLAY IOS,ZHYPERLINK command output.

```
D M=DEV(0A70F)
IEE174I 16.18.33 DISPLAY M 296
DEVICE 0A70F    STATUS=ONLINE
CHP                          50    52    69    6B
ENTRY LINK ADDRESS           35    1235  02    02
DEST LINK ADDRESS            07    1207  08    08
PATH ONLINE                  Y     Y     Y     Y
CHP PHYSICALLY ONLINE        Y     Y     Y     Y
PATH OPERATIONAL             Y     Y     Y     Y
MANAGED                      N     N     N     N
CU NUMBER                    A701  A701  A701  A701
INTERFACE ID                 0230  0033  0031  0231
MAXIMUM MANAGED CHPID(S) ALLOWED:   0
DESTINATION CU LOGICAL ADDRESS = 67
SCP CU ND            - 002107.996.IBM.75.0000000LBL41.0230
SCP TOKEN NED        - 002107.900.IBM.75.0000000LBL41.6700
SCP DEVICE NED       - 002107.900.IBM.75.0000000LBL41.670F
WWNN                 - 500507630AFFD03A
HYPERPAV ALIASES CONFIGURED = 16
ZHYPERLINKS AVAILABLE = 4
FUNCTIONS ENABLED = MIDAW, ZHPF, HS, XPAV, ZHYPERLINK
```

Figure 4: DISPLAY M=DEV(devno) command output.

```
D M=CU(A701)
IEE174I 16.10.19 DISPLAY M 055
CONTROL UNIT A701
CHP                        70    72    69    6B
ENTRY LINK ADDRESS         36    36    05    05
DEST LINK ADDRESS          07    07    08    08
CHP PHYSICALLY ONLINE      Y     Y     Y     Y
PATH VALIDATED             Y     Y     Y     Y
MANAGED                    N     N     N     N
ZHPF - CHPID               Y     Y     Y     Y
ZHPF - CU INTERFACE        Y     Y     Y     Y
INTERFACE ID               0230  0033  0031  0231
MAXIMUM MANAGED CHPID(S) ALLOWED - 0
DESTINATION CU LOGICAL ADDRESS - 67
CU ND                - 002107.996.IBM.75.0000000LBL41.0230
CU NED               - 002107.996.IBM.75.0000000LBL41.6700
TOKEN NED            - 002107.900.IBM.75.0000000LBL41.6700
WWNN                 - 500507630AFFD03A
FUNCTIONS ENABLED    - ZHPF, ZHYPERLINK, XPAV
XPAV CU PEERS        - 0300, 0900, 0B00, 1101, 1403, 1800, 1A01, 1D01,
                       1F01, 2301, 2501, 2801, 2A01, 3101, 3200, 3400,
                       5600, 7300, 7700, 7900, 7B00, 7D00, 8100, 8300,
                       8700, 8901, 9C00, A201, A701
DEFINED DEVICES
  0A700-0A70F
DEFINED PAV ALIASES
  0A710-0A71F
USABLE HYPERPAV ALIASES - 16
ZHYPERLINKS
  PFID      PCHID  Port  LinkId  S/W St    Port St
  00002430  016C   01    0080    Alloc     Oper
  00002431  016C   01    0080    Alloc     Oper
  00002432  016C   01    0080    Alloc     Oper
  00002433  016C   01    0080    Alloc     Oper
  00002434  0190   01    0180    Alloc     Oper
  00002435  0190   01    0180    Alloc     Oper
  00002436  0190   01    0180    Alloc     Oper
  00002437  0190   01    0180    Alloc     Oper
  00002438  01CC   02    0280    Alloc     Oper
  00002439  01CC   02    0280    Alloc     Oper
  0000243A  01CC   02    0280    Alloc     Oper
  0000243B  01CC   02    0280    Alloc     Oper
  0000243C  01D0   01    0380    Alloc     Oper
  0000243D  01D0   01    0380    Alloc     Oper
  0000243E  01D0   01    0380    Alloc     Oper
  0000243F  01D0   01    0380    Alloc     Oper
```

Figure 5: DISPLAY M=CU(cuno) command output.

Enabling zHyperLink Write on Db2

– zHyperLink write support is available on Db2 12 for z/OS.

To enable zHyperLink, use the Interactive Storage Management Facility (ISMF) interface to change the zHLink Write parameter of the storage class used to allocate the Db2 active log data sets to YES from the default of NO. Use the VARY SMS command to override the storage class setting for SMS managed data sets or to allow zHyperLink writes to be used for non-SMS managed data sets. Note: The VARY SMS command change does not persist across an IPL.

To exploit zHyperLink support for Db2 active log datasets, the ZHYPERLINK Db2 subsystem parameter must be set to either ACTIVELOG or ENABLE. This can be dynamically updated while Db2 is running by issuing the Db2 SET SYSPARM command[4].

Enabling zHyperLink Write on DS8000

- zHyperLink write support for synchronous mirroring is available on the DS8880 and DS8900.
- zHyperLink write support for asynchronous mirroring is available only on the DS8900.

To enable zHyperLink, logon to DS8000 HMC as an administrator, select Settings and then System to see the list of supported licensed functions. The zHyperLink Function shows the options to enable I/O Write and I/O Read (see Figure 6). Enabling zHyperlink writes on the DS8000 will result in a quiesce/resume of the DS8000 servers similar to a code load, and so while this is non-disruptive, it would typically be scheduled for a quiet time.

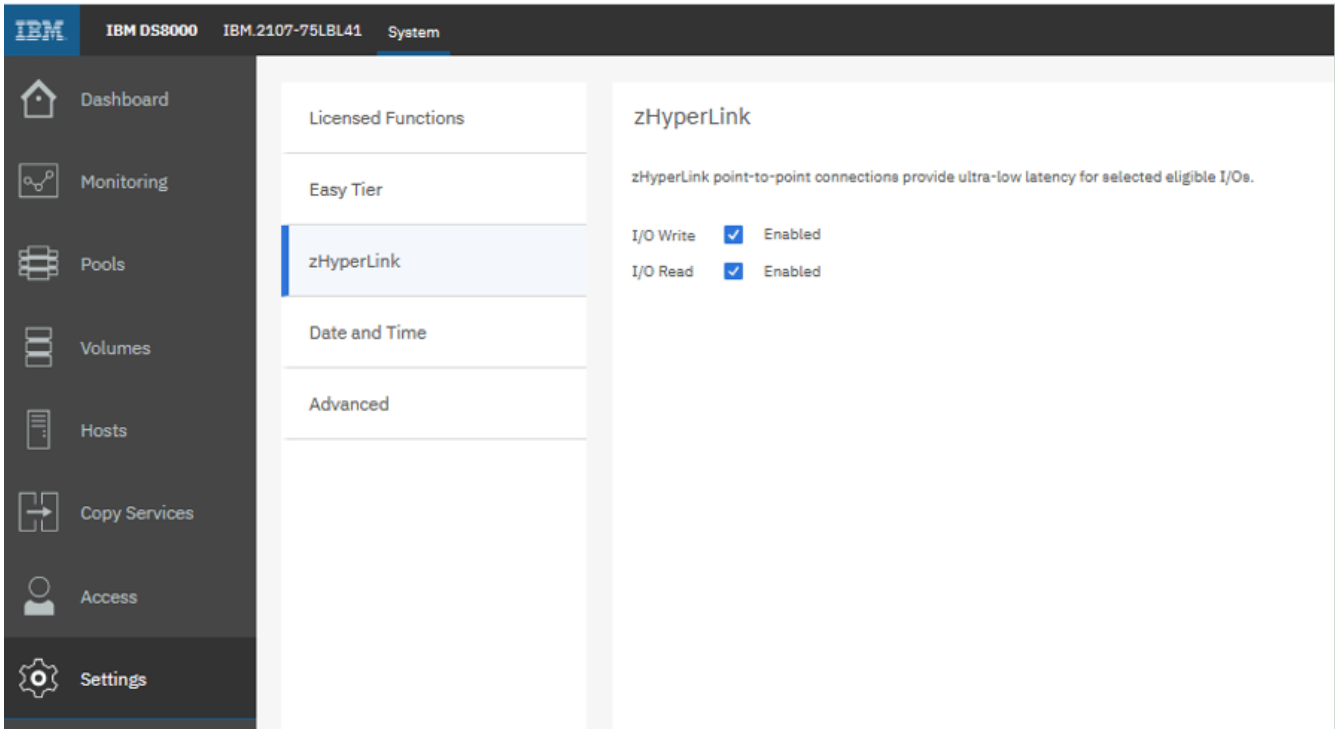

Figure 6: Enabling zHyperLink on DS8000.

For Metro Mirror environments, the devices on which the Db2 log data sets reside must be in the duplex state, in a HyperSwap configuration, and zHyperWrite must be enabled. In the IBM test lab, the solution was tested when the devices were in a simplex and the following mirroring configurations:

- Single site
  - Simplex

- 2-site
  - Metro Mirror (MM) (see Figure 7)
  - Global Mirror (GM) (see Figure 8)

- 3-site
  - Multi-Target (MT) MM/MM
  - Multi-Target MM/GM (see Figure 9)
  - MGM (see Figure 9)

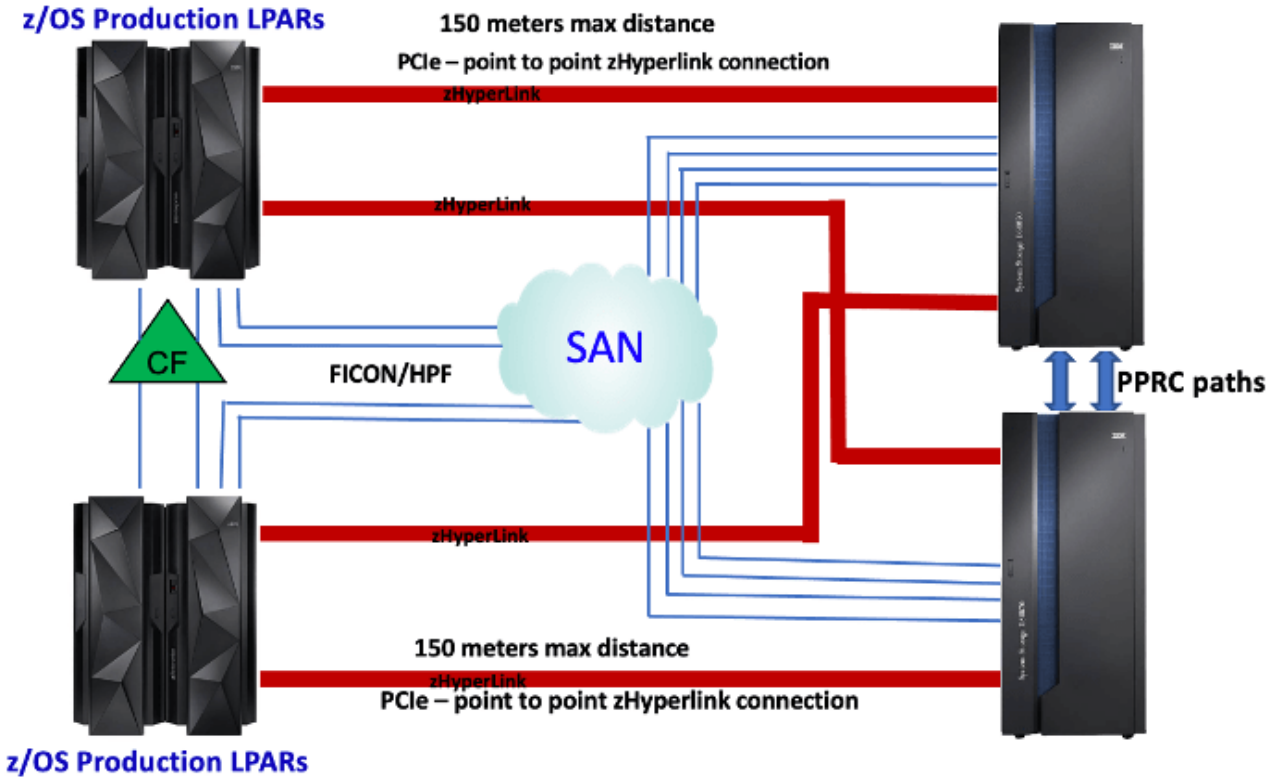

Figure 7: Two Site Metro Mirror environment.

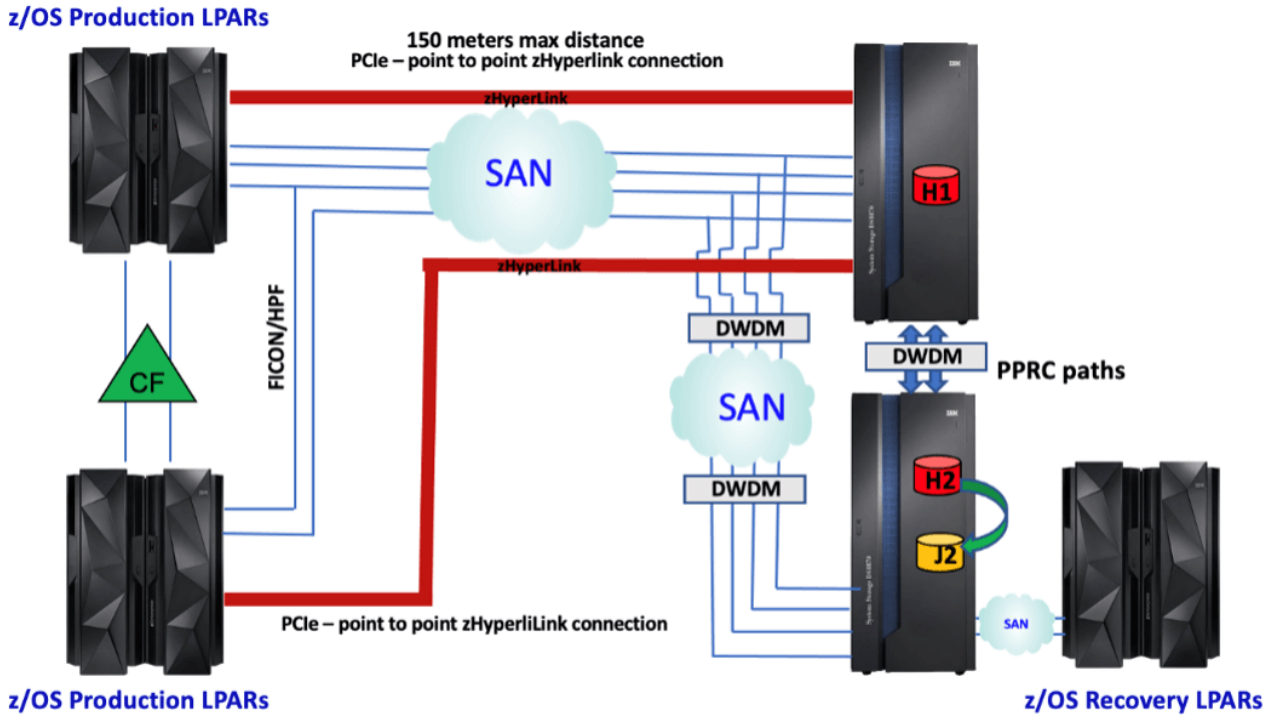

Figure 8: Two Site Global Mirror environment.

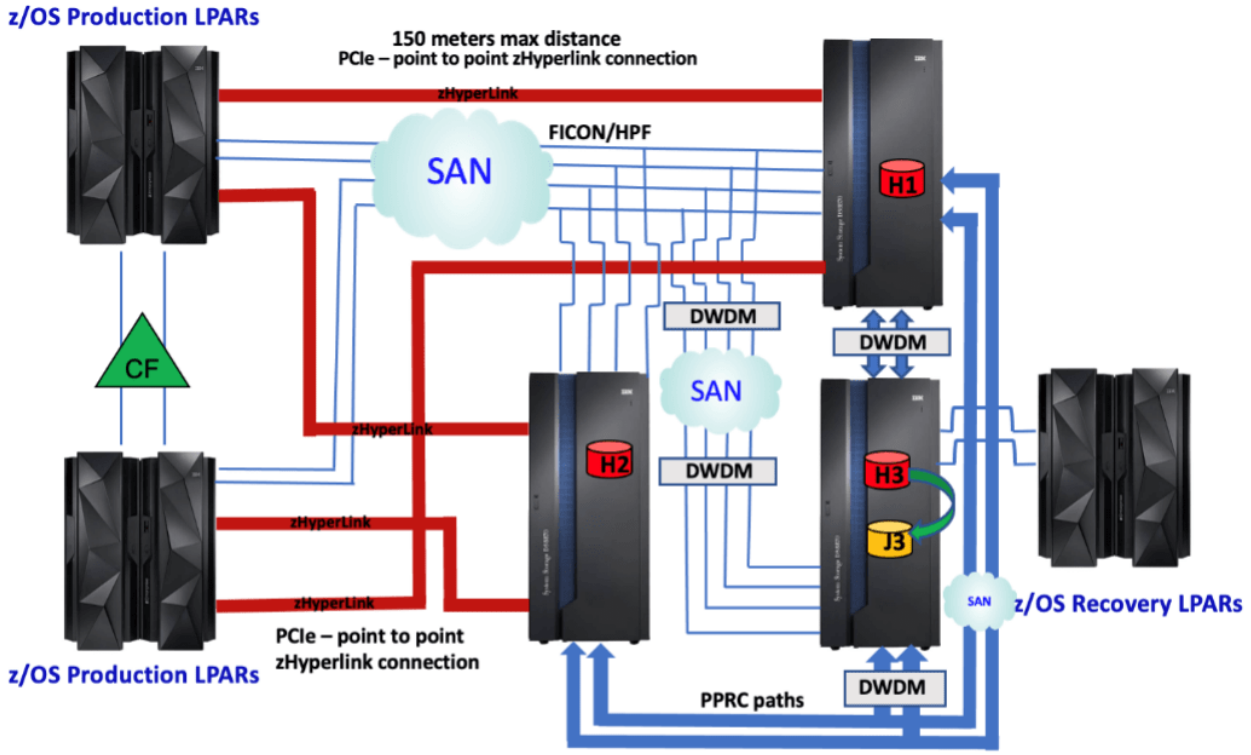

Figure 9: Three Site Multi-Target MM/GM or MGM environment.

zHyperLink Write Performance Results

zHyperLink writes provide lower latency than zHPF in all replication configurations. Figure 10 compares the performance of zHyperLink writes to zHPF writes with no mirroring and with different mirroring environments (Metro Mirror, Global Copy, Global Mirror). These results were achieved with 26-meter zHyperLink connections using a workload that simulates Db2 log writes that were 4K in size. Note: In this test, zHyperWrite was not enabled when using zHPF in a Metro Mirror environment so the response time shown is higher than it would have been with zHyperWrite enabled. With zHyperWrite, the response time for Metro Mirror would be similar to no mirroring.

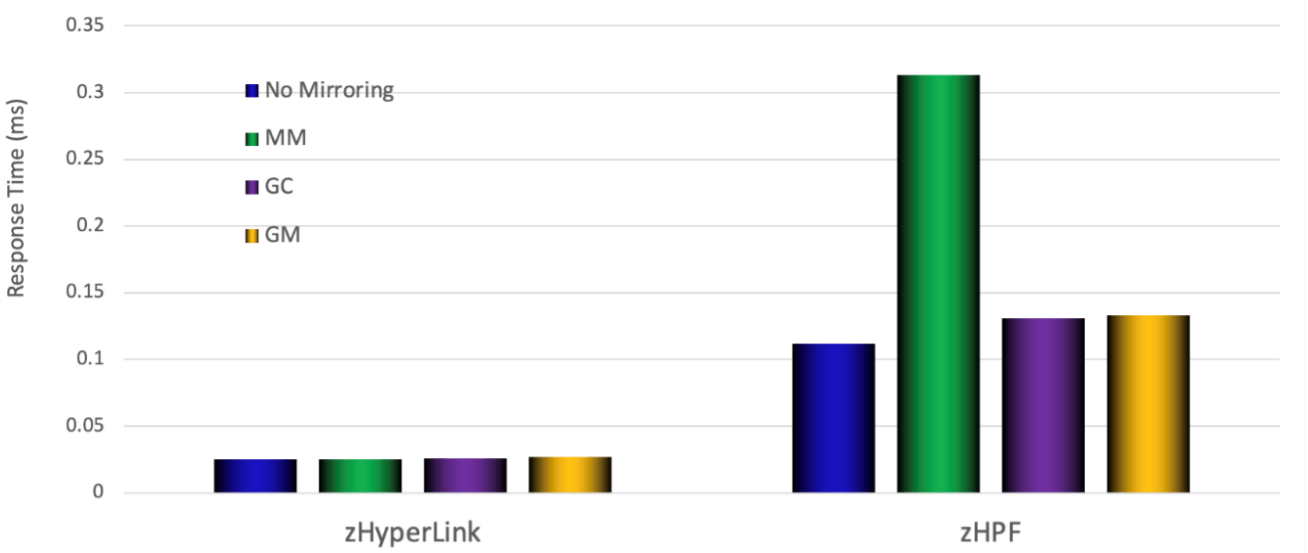

Figure 10: zHyperLink vs zHPF response time with different mirroring configurations.

Figure 11 compares zHPF to zHyperLink performance in a Global Mirror environment with a mixed random/sequential read and write workload and using a program that emulates Db2 log writes. zHyperLink reads were not enabled for this test. There was a 4x reduction in log write response time with zHyperLink. Overall, response time showed only a small improvement because reads made up a majority of the I/Os. In addition, this test showed that the success rate for zHyperLink writes was unaffected by Global Mirror and the consistency group formation time, which affects the Recovery Point Objective (RPO), was not affected by zHyperLink writes (1.1 seconds with and without zHyperLink).

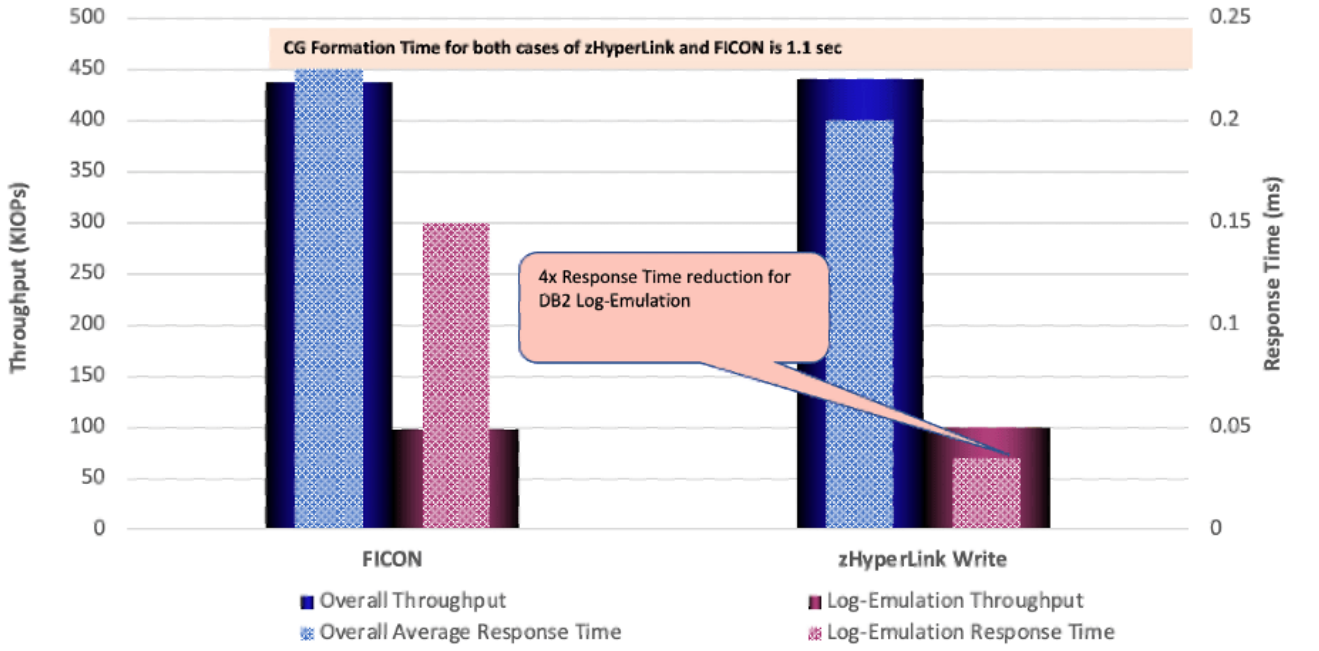

Figure 11: zHyperLink write vs FICON with Global Mirror consistency group.

zHyperLink Write Statistics and Reports

RMF shows synchronous I/O performance data at the DASD device and PCIE function level. Figure 12 shows an example of the RMF Synchronous I/O Device Activity report, which contains the zHyperLink read and write activity as well as the asynchronous I/O activity.

```
                                S Y N C H R O N O U S   I/O   D E V I C E   A C T I V I T Y

            z/OS V2R4              SYSTEM ID SY1B            START DATE 10/30/2020-14.19.00     INTERVAL 000.05.59
                                   RPT VERSION V2R4 RMF      END   DATE 10/30/2020-14.25.00     CYCLE 1.000 SECONDS

TOTAL SAMPLES =    360  IODF = 1B   CR-DATE: 07/16/2020  CR-TIME: 07.22.21     ACT: ACTIVATE
                                  - DEVICE ACTIVITY RATE -  -- AVG RESP TIME --  AVG SYNCH I/O  %        %    %     %
STORAGE  DEV   DEVICE  VOLUME  LCU -- SYNCH I/O --   ASYNCH  -SYNCH I/O - ASYNCH  TRANSFER RATE  REQ      LINK CACHE --REJECTS--
 GROUP   NUM   TYPE    SERIAL        READ   WRITE      I/O   READ  WRITE    I/O   READ   WRITE  SUCCESS  BUSY  MISS  READ WRITE
DC1NLOG  0B024 3390A   TDB024  0050  0.000  5482.85    0.547 0.000 0.064  0.786 0.000  109.9    100.0  0.00  0.00  0.00 0.00
                        LCU    0050  0.000  5482.85    0.547 0.000 0.064  0.786 0.000  109.9    100.0  0.00  0.00  0.00 0.00
DC1NLOG  0B124 3390A   TDB124  0051  0.000  5482.85    0.558 0.000 0.072  0.815 0.000  109.9    100.0  0.00  0.00  0.00 0.00
                        LCU    0051  0.000  5482.85    0.558 0.000 0.072  0.815 0.000  109.9    100.0  0.00  0.00  0.00 0.00
```

Figure 12: RMF synchronous I/O device activity report.

The DISPLAY LOG command for Db2 for z/OS shows if zHyperLink writes is enabled for

the active logs and whether the last log write for the active log data sets used zHyperLink (see Figure 13).

```
-DAP0 DIS LOG
DSNJ370I  -DAP0 DSNJC00A LOG DISPLAY 188
CURRENT COPY1 LOG - DBA1.DAP0.LOGCOPY1.DS03 is 82% FULL
CURRENT COPY2 LOG - DBA1.DAP0.LOGCOPY2.DS03 is 82% FULL
          H/W RBA - 000000002C14FE356EE8
          H/O RBA - 000000002C149640AFFF
          FULL LOGS TO OFFLOAD - 0 OF 6
          OFFLOAD TASK IS (AVAILABLE)
          SOFTWARE ACCELERATION IS ENABLED
          ZHYPERLINK WRITE IS ENABLED
            LAST LOG WRITE FOR COPY1 USED ZHYPERLINK
            LAST LOG WRITE FOR COPY2 USED ZHYPERLINK
```

Figure 13: Db2 DISPLAY LOG command output.

Alternatively, DFSMS collects detailed zHyperLink write statistics and makes them available with the SMF type 42 subtype 6 records and the DISPLAY SMS,DSNAME(dsn),STATS(ZHLWRITE) command (see Figure 14). The statistics are provided at both a Db2 request level as well as the device level.

```
D SMS,DSNAME(DBA1.DAP0.LOGCOPY1.DS02),STATS(ZHLWRITE)
IGW289I 419
D SMS,DSNAME,STATS(ZHLWRITE) Start of Report
DATA SET DBA1.DAP0.LOGCOPY1.DS02.DATA
STATISTICS Since 05/20/2021 13:51:39.142166
SUMMARY
         TOTAL      %SYNC ------------%ASYNC--------------
  WRITE REQUESTS WRITES    SKIP LNKBSY    ^EST    MISC DISABL
          751121  99.89    0.03  <0.01   <0.01   <0.01   0.00
                         -------%ASYNC-------
                           MISS   DELAY    DUAL
                          <0.01    0.04    0.01
DEVICE STATISTICS
              TOTAL   %SYNC ---------------%ASYNC--------------------
SSID DEVNO WRITES  WRITES    SKIP LNKBSY    ^EST    MISC    MISS   DELAY
1241 0030F 106e04   99.96    0.01  <0.01   <0.01   <0.01   <0.01   0.01
12E1 08C64 106e04   99.95    0.01  <0.01   <0.01   <0.01   <0.01   0.01
1241 00305 106e04   99.94    0.01  <0.01   <0.01   <0.01   <0.01   0.01
D SMS,DSNAME,STATS(ZHLWRITE) End of Report
```

Figure 14: DISPLAY SMS command output.

Omegamon XE for Db2 Performance Expert shows the average elapsed time for class 3 suspensions for synchronous I/O log writes[5]. The TIME/EVENT for SYNCHRONOUS LOG WRITE I/O shows the benefit of using zHyperLink writes for Db2 active logs when compared to reports prior to enabling zHyperLink writes (see Figure 15).

```
   LOCATION: DSNC1N0                  OMEGAMON XE FOR DB2 PERFORMANCE EXPERT (V5R4M0)                  PAGE: 1-1
      GROUP: DSNC1N0                        ACCOUNTING REPORT - LONG                  REQUESTED FROM: 10/30/20 14:19:00.00
     MEMBER: DC1N                                                                              TO: 10/30/20 14:25:01.01
  SUBSYSTEM: DC1N                             ORDER: PRIMAUTH-PLANNAME                 INTERVAL FROM: 10/30/20 14:19:00.00
DB2 VERSION: V12                                 SCOPE: MEMBER                                 TO: 10/30/20 14:25:01.00

PRIMAUTH: USRT031  PLANNAME: db2jcc_a
ELAPSED TIME DISTRIBUTION                                        CLASS 2 TIME DISTRIBUTION
-------------------------------------------------------------    ------------------------------------------------------------------
APPL   |============================> 58%                       CPU    |========> 17%
DB2    |=====> 10%                                               SECPU  |
SUSP   |================> 32%                                    NOTACC |===> 6%
                                                                 SUSP   |====================================> 77%

AVERAGE       APPL(CL.1)  DB2 (CL.2)    CLASS 3 SUSPENSIONS   AVERAGE TIME  AV.EVENT  TIME/EVENT    HIGHLIGHTS
------------  ----------  ----------    --------------------  ------------  --------  ----------    --------------------------
ELAPSED TIME    0.001226    0.000510    LOCK/LATCH(DB2+IRLM)      0.000023      0.20    0.000116    #OCCURRENCES    :  7841320
 NONNESTED      0.001226    0.000510     IRLM LOCK+LATCH          0.000008      0.02    0.000347    #ALLIEDS        :        0
 STORED PROC    0.000000    0.000000     DB2 LATCH                0.000015      0.17    0.000085    #ALLIEDS DISTRIB:        0
 UDF            0.000000    0.000000    SYNCHRON. I/O             0.000368      1.00    0.000368    #DBATS          :  7841320
 TRIGGER        0.000000    0.000000     DATABASE I/O             0.000000      0.00    0.000488    #DBATS DISTRIB. :        0
                                          READ CACHE HIT          0.000000      0.00    0.000488    #NO PROGRAM DATA:        0
CP CPU TIME     0.000103    0.000087     LOG WRITE I/O            0.000368      1.00    0.000368    #NORMAL TERMINAT:  7841320
 AGENT          0.000103    0.000087    OTHER READ I/O            0.000000      0.00         N/C      #ROLLUP TRAN    :  7841320
  NONNESTED     0.000103    0.000087    OTHER WRTE I/O            0.000000      0.00    0.001171    #DDFRRSAF ROLLUP:   784132
  STORED PRC    0.000000    0.000000    SER.TASK SWTCH            0.000001      0.00    0.072448    #ABNORMAL TERMIN:        0
  UDF           0.000000    0.000000     UPDATE COMMIT            0.000000      0.00         N/C    #CP/X PARALLEL. :        0
  TRIGGER       0.000000    0.000000     OPEN/CLOSE               0.000000      0.00         N/C    #UTIL PARALLEL. :        0
 PAR.TASKS      0.000000    0.000000     SYSLGRNG REC             0.000000      0.00         N/C    #IO PARALLELISM :        0
                                         EXT/DEL/DEF              0.000001      0.00    0.088166    #PCA RUP COUNT  :        0
SE CPU TIME     0.000000    0.000000    OTHER SERVICE             0.000000      0.00    0.000288    #RUP AUTONOM. PR:        0
```

Figure 15: Omegamon XE for Db2 Performance accounting report – long.

About the authors

Tariq Hanif is a Senior Software Engineer in the z/OS I/O Supervisor Team. In z/OS, his focus areas of testing are Parallel Sysplex, Storage Replication Solutions and Disaster Recovery.

Brian Lee is a Software Developer for z/OS DFSMS Media Manager.

Dale Riedy is a Senior Technical Staff Member with z/OS I/O Supervisor Design and Development.

We would like to thank Nick Clayton, Distinguished Engineer - Enterprise Storage Development, for reviewing and providing comments on this article.

Rita Beisel contributed to the editorial review of this article.

---

1. Faster Db2 Active Log Writes with Media Manager Parallel Write Support using zHyperLink

   ↩

2. z/OS MVS Initialization and Tuning Reference. Members of SYS1.PARMLIB. IECIOSxx (I/O related parameters).

   ↩

3. z/OS MVS System Commands. MVS system commands reference. SETIOS command.

   ↩

4. Db2 Commands. -SET SYSPARM (Db2).↩

5. OMEGAMON for Db2 Performance Expert on z/OS 5.4.0. Accounting Report – Long.