High Performance FICON Demystified, Update and User Experience

Dale Riedy
IBM
riedy@us.ibm.com
5 February 2013
Session 13056
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 provided 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.
4X the of FICON I/Os per Second

I/Os per second – I/O driver benchmark
4k block size, channel 100% utilized

77% Increase

New

FICON Express
FICON Express
FICON Express
FICON Express
FICON Express
FICON Express
FICON Express
FICON Express

z990 z990 z990 z990 z10 z10 z196 z196 z196 z196 z196 z196 zEC12 zEC12
More than 2X FICON Throughput

I/O driver benchmark
MegaBytes per second

Full-duplex
Large sequential read/write mix

FICON Express
2 Gbps

FICON Express2
2 Gbps

FICON Express4
4 Gbps

FICON Express8
8 Gbps

FICON Express8S
8 Gbps

zSeries
z9
z990
z890

z10
z10

z196
z196

z114
z114

108% Increase

1600
FICON Express8S
8 Gbps

New

Complete your sessions evaluation online at SHARE.org/SFEval
Response Time Improves Too

Single FICON Express8S channel: zHPF vs FICON READ 4k bytes/IO
Total I/O Response Time vs IO/sec
zHPF Evolution

Single domain, single track I/O
- Reads, update writes
- Media manager exploitation
- z/OS R8 and above

Multi-track, but <= 64K

Multi-track any size

Format writes, multi-domain I/O
- QSAM/BSAM exploitation
- Incorrect Length Facility
- z/OS R11+ EXCPVR

100% of DB2 I/O is now converted to zHPF

2008

DS8100/DS8300 with R4.1 or above
- z10 processor

2009

z196 processor >64K transfers

2010

2011

z196 FICON Express 8S
- DS8700/DS8800 with R6.2

2012

z/OS R12+ EXCP virtual
Agenda

- What does zHPF Do For Me?
- How Does zHPF Do It?
- The Effect On Exchanges
- Other Improvements
How does zHPF do it?

- Rides on top of an existing standard protocol called….

F. C. P.
FCP ???

- Does zHPF convert my I/O to SCSI ???????
  NO !

- FCP is a generic method to transfer commands, data, and status

- FCP \(\neq\) SCSI
  - It is true however, that SCSI is the single largest user of FCP
Why FCP?

- FCP protocol has less ‘Chit Chat’
- Many HBA vendors have optimized firmware and hardware to accelerate FCP I/O
Read Comparison Summary (4 4K Reads)

<table>
<thead>
<tr>
<th></th>
<th>Channel to CU in Ficon Mode</th>
<th>CU to Channel in Ficon Mode</th>
<th>Total</th>
<th>Channel To CU in zHPF Mode</th>
<th>CU to Channel in zHPF Mode</th>
<th>Total</th>
<th>% Reduction in zHPF Mode¹</th>
</tr>
</thead>
<tbody>
<tr>
<td>Exchanges</td>
<td>1</td>
<td>1</td>
<td>2</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>50</td>
</tr>
<tr>
<td>Sequences</td>
<td>6</td>
<td>6</td>
<td>12</td>
<td>1</td>
<td>2</td>
<td>3</td>
<td>75</td>
</tr>
<tr>
<td>Frames</td>
<td>6</td>
<td>14</td>
<td>20</td>
<td>1</td>
<td>10</td>
<td>11</td>
<td>45</td>
</tr>
<tr>
<td>CRC Gen / Check</td>
<td>5</td>
<td>5</td>
<td>10</td>
<td>1</td>
<td>1</td>
<td>2</td>
<td>80</td>
</tr>
</tbody>
</table>

¹Except for exchanges, as the number of reads in a single I/O increase, the % reduction in Transport Mode increases
Let’s look under the hood

Image Credit: Flikr user aka Razz
Creative Commons License
No I/O Definition changes

- zHPF coexists with FICON

- Channel is STILL type=FC

- No I/O configuration (IOCDS/IODF) changes for zHPF capable channels or control units
Initialization Changes

Switch

BLACK = Unchanged from Ficon
RED = Changed from Ficon
GREEN = NEW in zHPF

FLOGI
PLOGI (Fabric Ctrlr)
QSA
SCR
RNID
PLOGI (Mgmt Srvr)

New bit indicates support for PRLI

Control Unit

Accept

Accept

RNID

Exchanges zHPF capabilities

PLOGI
LIRR
PRLI

Responses not shown unless new/changed

Complete your sessions evaluation online at SHARE.org/SFEval
Totally New I/O Structures

- CCWs no longer exist in zHPF (They live on happily in Ficon)
  - Replaced by Device Control Words (DCWs)

- IDAWs and MIDAWs no longer exist in zHPF (They too are alive and well in FICON)
  - Replaced by Transport Indirect Data Address Words (TIDAW)

- New structures added
  - Transport Control Word (TCW)
  - Transport Status Block (TSB)
  - Transport Command & Control Block (TCCB)
Command Mode Review
Command Mode Review

ORB

CCW
Command Mode Review

- ORB
- CCW
- IDAW / MIDAW
Command Mode Review

ORB → CCW → IDAW / MIDAW → Data
Command Mode Review

ORB

CCW

IDAW / MIDAW

Data

CCW

IDAW / MIDAW

Data
Command Mode Review

ORB

CCW

CCW

CCW

CCW

CCW

Data

IDAW / MIDAW

Data

IDAW / MIDAW

Data

IDAW / MIDAW

Data

IDAW / MIDAW

Data
Transport Mode
Transport Mode

ORB -> TCW
Transport Mode

ORB -> TCW

Data

Data

Data

TIDAW
TIDAW
TIDAW
TIDAW
Transport Mode

ORB \rightarrow TCW

Data \rightarrow TIDAW

Data \rightarrow TIDAW

Data \rightarrow TIDAW

Addressing & Control Information
- Device Command Word
- Device Command Word
- Device Command Word
- Device Command Word
- Device Command Word
- Checking Block Code
- Count

TCCB
- Transport Command & Control Block
Transport Mode

**ORB**
**Data**
**Data**
**Data**

**TCW**
**TIDAW**
**TIDAW**
**TIDAW**

**TSB**
Transport Status Block

**TCCB**
Transport Command & Control Block

- Addressing & Control Information
  - Device Command Word
  - Device Command Word
  - Device Command Word
  - Device Command Word
  - Device Command Word
  - Checking Block
  - Count
Link View of 4 Reads in Command Mode

Prefix Cmd + 64 bytes

Read (4k) → Cmd Resp

Read (4k) ← Data (4k) + CRC

Read (4k) ← Data (4k) + CRC

Read (4k) ← Data (4k) + CRC

Status Accept → Status

<table>
<thead>
<tr>
<th></th>
<th>Channel to Control Unit</th>
<th>Control Unit to Channel</th>
</tr>
</thead>
<tbody>
<tr>
<td>Total Commands</td>
<td>5</td>
<td>N/A</td>
</tr>
<tr>
<td>Exchanges</td>
<td>2</td>
<td>2</td>
</tr>
<tr>
<td>Sequences</td>
<td>6</td>
<td>6</td>
</tr>
<tr>
<td>Frames</td>
<td>6</td>
<td>14</td>
</tr>
<tr>
<td>CRC Generate / Check</td>
<td>5</td>
<td>5</td>
</tr>
</tbody>
</table>
Link View of 4 Reads in Transport (zHPF) Mode

Prefix  + 64 bytes of prefix data +
Read (4k) + Read (4k) + Read (4k) +
Read (4k)

Data (16k) + CRC

Status

<table>
<thead>
<tr>
<th></th>
<th>Channel to Control Unit</th>
<th>Control Unit to Channel</th>
</tr>
</thead>
<tbody>
<tr>
<td>Total Commands</td>
<td>5</td>
<td>N/A</td>
</tr>
<tr>
<td>Exchanges</td>
<td>1</td>
<td>1 (same one)</td>
</tr>
<tr>
<td>Sequences</td>
<td>1</td>
<td>2</td>
</tr>
<tr>
<td>Frames</td>
<td>1</td>
<td>10</td>
</tr>
<tr>
<td>CRC Generate / Check</td>
<td>1</td>
<td>1</td>
</tr>
</tbody>
</table>
Device Control Word (DCW)

<table>
<thead>
<tr>
<th>Command</th>
<th>Control Flags</th>
<th>Reserved</th>
<th>CD Count</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td></td>
<td>DCW Data Count</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>Optional Control Parameters (e.g., define extent and locate record parameters)</td>
</tr>
</tbody>
</table>
Device Control Word (DCW)

- Control Flags
  - CC (Chain Command)
    - Another command follows. If the command completes “normally” the next command is to be executed

- CD Count
  - Number of bytes that follow the DCW that contain Control Data for the DCW

- Data Count
  - Number of bytes of data to be transferred in the data phase for this DCW not including any Pad and CRC
ORB

<table>
<thead>
<tr>
<th>Word</th>
<th>Interruption Parameter</th>
<th>Channel-Program Address</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td></td>
<td></td>
</tr>
<tr>
<td>1</td>
<td>Key</td>
<td>0000000000</td>
</tr>
<tr>
<td>2</td>
<td>CSS Priority</td>
<td>Reserved</td>
</tr>
<tr>
<td>3</td>
<td>Reserved</td>
<td>Reserved</td>
</tr>
<tr>
<td>4</td>
<td>Reserved</td>
<td>Reserved</td>
</tr>
<tr>
<td>5</td>
<td>Reserved</td>
<td>Reserved</td>
</tr>
<tr>
<td>6</td>
<td>Reserved</td>
<td>Reserved</td>
</tr>
<tr>
<td>7</td>
<td>Reserved</td>
<td>Reserved</td>
</tr>
</tbody>
</table>

Specifies Transport (zHPF) Mode
## TCW (Transport Control Word)

<table>
<thead>
<tr>
<th>Word</th>
<th>Flags</th>
<th>Reserved</th>
<th>TCCBL</th>
<th>RW</th>
<th>Reserved</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>F 0 0 0 0 0 0</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>1</td>
<td></td>
<td>Reserved</td>
<td>TCCBL</td>
<td>RW</td>
<td>Reserved</td>
</tr>
<tr>
<td>2</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>3</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>4</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>5</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>6</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>7</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>8</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>9</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>10</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>11</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>12</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>13</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>14</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>15</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

- 0: Interrogate-TCW Address
- 1: Output-Data Address
- 2: Input-Data Address
- 3: Transport-Status-Block Address
- 4: Transport-Command-Control Block Address
- 5: Output Count
- 6: Input Count
- 7: Reserved
- 8-12, 14-15: Reserved
Agenda

- What does zHPF Do For Me?
- How Does zHPF Do It?
- The Effect On Exchanges
- Other Improvements
How does zHPF affect EXCHANGES?

- Little’s Law states:
  - The number of “things” in a system can be determined by multiplying the average arrival rate of those “things” by the average time each “thing” stays in the system.

- Applied to zHPF:
  - The average number of Exchanges active at any given time = Average I/O rate \* Average response time

  - Example: 30000 Ficon I/Os / Second on a given channel with .3ms service time\(^1\) uses 9 Active Exchanges at any given time

\(^1\) The amount of time the I/O is active in the channel
How does zHPF affect EXCHANGES?

- The CU holds on to the Exchange even if the device:
  - Is reserved
  - Detects an Extent Conflict
  - Cache Miss
  - etc
- Drives requirement for higher number of possible open Exchanges

Example:
Assume we are doing 50,000 I/Os per second with an average service time of 0.5 ms. If 20% hit one of the above conditions and if each of those conditions lasts for 10ms, then:

- 100 Exchanges are needed for Busies
- 20 Exchanges are needed for the rest
How does zHPF affect EXCHANGES?

- CU can dynamically adjust the number of open Exchanges any one channel can open to THAT CU (physical link)

- Channel maintains a Exchange count and Exchange Limit for each physical link to a control unit
New RMF Fields for zHPF

<table>
<thead>
<tr>
<th>CHANNEL PATH</th>
<th>UTILIZATION(%)</th>
<th>READ(MB/SEC)</th>
<th>WRITE(MB/SEC)</th>
<th>FICON OPERATIONS</th>
<th>ZHPF OPERATIONS</th>
</tr>
</thead>
<tbody>
<tr>
<td>ID TYPE</td>
<td>G SHR</td>
<td>PART TOTAL</td>
<td>BUS</td>
<td>PART TOTAL</td>
<td>PART TOTAL</td>
</tr>
<tr>
<td>00 FC_S</td>
<td>5 Y</td>
<td>100.00</td>
<td>100.00</td>
<td>0.84</td>
<td>0.13</td>
</tr>
<tr>
<td>01 FC_S</td>
<td>5 Y</td>
<td>100.00</td>
<td>100.00</td>
<td>0.85</td>
<td>0.13</td>
</tr>
<tr>
<td>02 FC_S</td>
<td>4 Y</td>
<td>0.14</td>
<td>2.30</td>
<td>0.85</td>
<td>0.10</td>
</tr>
<tr>
<td>03 FC_S</td>
<td>4 Y</td>
<td>0.13</td>
<td>2.27</td>
<td>0.84</td>
<td>0.11</td>
</tr>
<tr>
<td>04 FC_S</td>
<td>5 Y</td>
<td>0.13</td>
<td>2.24</td>
<td>0.82</td>
<td>0.10</td>
</tr>
<tr>
<td>05 FC_S</td>
<td>5 Y</td>
<td>0.13</td>
<td>2.25</td>
<td>0.83</td>
<td>0.10</td>
</tr>
<tr>
<td>06 FC_S</td>
<td>4 Y</td>
<td>0.12</td>
<td>2.23</td>
<td>0.83</td>
<td>0.10</td>
</tr>
</tbody>
</table>
What Do I Need to Exploit zHPF?

- z10 at Driver 76 or higher
  - Power On Reset is REQUIRED to activate zHPF
- z196, z114, zEC12

- FICON Express-2 or above, FICON Express 8S for full exploitation

- Control Unit that supports zHPF
  - Check with your vendor for appropriate code and/or hardware levels
  - Enable the LIC feature

- All supported releases of z/OS
  - zHPF mode has to be enabled (IECIOSxx parmlib or SETIOS command)
  - SAM_USE_HPF=YES in IGDSMSxx (QSAM/BSAM support)
Agenda

What does zHPF Do For Me?

How Does zHPF Do It?

The Effect On Exchanges

Other Improvements
MIH

- Reduced False Missing Interrupt for reserves
  -Avoids “Go to the end of the line” penalty for MIH due to reserves
  -zHPF allows the OS to interrogate the state of an existing I/O operation
- Enhanced MIH messages and logrec
MIH Message Example

IOS071I 031B,62,*MASTER*, START PENDING
STATUS: DEVICE RESERVED BY ANOTHER SYSTEM

IOS071I 0980,40,IOSAS, START PENDING
STATUS: NO I/O OPERATION IS IN PROGRESS

IOS071I 0410,F2,WHATEVER,START PENDING
STATUS: I/O WAITING FOR EXTENT CONFLICT

IOS071I 1029,A8,JES3,START PENDING
STATUS: I/O OPERATION IS EXECUTING
Transport Mode

ORB

TCW

Data

TIDAW
TIDAW
TIDAW
TIDAW

Addressing & Control Information
Device Command Word
Device Command Word
Device Command Word
Device Command Word
Device Command Word
Checking Block
Count

Data

TIDAW

TSB
Transport Status Block

Interrogate TCW

TIDAW
TCCB
TSB

Data

TCCB
Transport Command & Control Block
Temporary Logout

- CU firmware updates can be “cleaner” with zHPF support
- zHPF introduces a “temporary logout” concept
  - CU tells channel that it is ‘going away’
I'm going to go away for 3 seconds

PRLO
Process Logout
Hmmm, I better stop sending him new work

PRLO
Process Logout

I’m going to go away for 3 seconds
Hmmm, I better stop sending him new work

Ok, All existing work has completed. I'm ready

I'm going to go away for 3 seconds

PRLO Process Logout
I’m going to go away for 3 seconds

Hmmm, I better stop sending him new work

Ok, All existing work has completed. I’m ready

I’m going to go away for 3 seconds

3 Seconds later
Hmmm, I better stop sending him new work.

Ok, All existing work has completed. I’m ready.

Time’s up. Let’s get going.

I’m going to go away for 3 seconds.

3 Seconds later.

Complete your sessions evaluation online at SHARE.org/SFEval
Thank You For Your Time And Attention

Feel free to e-mail me with any zHPF or FICON questions