High Performance Ficon Demystified, Update and User Experience

Dale Riedy
IBM
riedy@us.ibm.com
8 August 2012
Session 11699
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.
Agenda

What does zHPF Do For Me?

How Does zHPF Do It?

The Effect On Exchanges

Other Improvements
4X the of FICON I/Os per Second

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

77% Increase

FICON Express4 and FICON Express2
FICON Express8
FICON Express8S

New

Complete your sessions evaluation online at SHARE.org/AntehmEval
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 4 Gbps
FICON Express4 4 Gbps
FICON Express8 8 Gbps
FICON Express8S 8 Gbps

108% Increase

zSeries
z9
z990
z890
z10
z10
z196
z196
z196
z196

1600 FICON Express8S 8 Gbps

zHPF

Complete your sessions evaluation online at SHARE.org/AnaheimEval
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

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

- z196 processor >64K transfers

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

- Format writes, multi-domain I/O
  - QSAM/BSAM exploitation
  - Incorrect Length Facility

- z/OS R11 and above, EXCPVR

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

What does zHPF Do For Me?

How Does zHPF Do It?

The Effect On Exchanges

Other Improvements
How does zHPF do it?

• Significantly reduced Channel and CU overhead

• Takes advantage of hardware assists in Fibre Channel interface chips

• Rides on top of an existing standard protocol called…. 
FCP ???

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

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

- FCP ≠ SCSI
  - It is true however, that SCSI is the single largest user of FCP
Why FCP?

- Many HBA vendors have optimized firmware and hardware to accelerate FCP I/O

- FCP protocol has less ‘Chit Chat’
## Read Comparison Summary (5 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 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

RNID
Accept

Exchanges zHPF capabilities

PLOGI
LIRR
PRLI

Responses not shown unless new/changed

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

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

- IDAWs 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
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

Complete your sessions evaluation online at SHARE.org/AnaheimEval
Command Mode Review
Transport Mode
Transport Mode

ORB → TCW
Transport Mode
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

TCCB
Transport Command & Control Block
Transport Mode

- **ORB**
- **TCW**
- **TSB** Transport Status Block

**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

**TCCB** Transport Command & Control Block
Link View of 4 Reads in Command Mode

Prefix Cmd + 64 bytes
Read (4k)
Read (4k)
Read (4k)
Read (4k)

Status Accept

Cmd Resp

Data (4k) + CRC
Data (4k) + CRC
Data (4k) + CRC
Data (4k) + CRC

Status Accept

<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)

Command phase

Data (16k) + CRC

Data phase

Status phase

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>DCW Data Count</td>
<td></td>
<td></td>
<td></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>
<th>CSS Priority</th>
<th>Reserved</th>
<th>Rsv. for Pgm.</th>
<th>Reserved</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>1</td>
<td>Key</td>
<td></td>
<td>LPM</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>0 0 0 0 0 0 0 0 0 0 B</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>2</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>3</td>
<td></td>
<td>Channel-Program Address</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>4</td>
<td>Reserved</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>5</td>
<td>Reserved</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>6</td>
<td>Reserved</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>7</td>
<td>Reserved</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

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

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

- Output-Data Address
- Input-Data Address
- Transport-Status-Block Address
- Transport-Command-Control Block Address
- Output Count
- Input Count
- Reserved
- Interrogate-TCW Address
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:
At 50,000 I/Os per Second, if 20% hit one of the above 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

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

<table>
<thead>
<tr>
<th>CHANNEL PATH ID</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></td>
<td>ID TYPE</td>
<td>G SHR</td>
<td>PART TOTAL</td>
<td>BUS</td>
<td>PART TOTAL</td>
</tr>
<tr>
<td>00 FC_S 5 Y</td>
<td>0.00</td>
<td>100.00</td>
<td>100.00</td>
<td>0.84</td>
<td>0.13</td>
</tr>
<tr>
<td>01 FC_S 5 Y</td>
<td>0.00</td>
<td>100.00</td>
<td>100.00</td>
<td>0.85</td>
<td>0.13</td>
</tr>
<tr>
<td>02 FC_S 4 Y</td>
<td>0.14</td>
<td>2.30</td>
<td>0.85</td>
<td>0.10</td>
<td>0.10</td>
</tr>
<tr>
<td>03 FC_S 4 Y</td>
<td>0.13</td>
<td>2.27</td>
<td>0.84</td>
<td>0.11</td>
<td>0.11</td>
</tr>
<tr>
<td>04 FC_S 5 Y</td>
<td>0.13</td>
<td>2.24</td>
<td>0.82</td>
<td>0.10</td>
<td>0.10</td>
</tr>
<tr>
<td>05 FC_S 5 Y</td>
<td>0.13</td>
<td>2.25</td>
<td>0.83</td>
<td>0.10</td>
<td>0.10</td>
</tr>
<tr>
<td>06 FC_S 4 Y</td>
<td>0.12</td>
<td>2.23</td>
<td>0.83</td>
<td>0.10</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

- Ficon Express-2 or above

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

- All supported releases of z/OS
  - zHPF mode has to be enabled (IECIOSxx parmlib or SETIOS command)
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 message and logrec
# MIH Message Example

<table>
<thead>
<tr>
<th>Message ID</th>
<th>Parameters</th>
<th>Status</th>
</tr>
</thead>
<tbody>
<tr>
<td>IOS071I 031B,62,<em>MASTER</em>,</td>
<td>START PENDING</td>
<td>DEVICE RESERVED BY ANOTHER SYSTEM</td>
</tr>
<tr>
<td>IOS071I 0980,40,IOSAS,</td>
<td>START PENDING</td>
<td>NO I/O OPERATION IS IN PROGRESS</td>
</tr>
<tr>
<td>IOS071I 0410,F2,WHATEVER,</td>
<td>START PENDING</td>
<td>I/O WAITING FOR EXTENT CONFLICT</td>
</tr>
<tr>
<td>IOS071I 1029,A8,JES3,</td>
<td>START PENDING</td>
<td>I/O OPERATION IS EXECUTING</td>
</tr>
</tbody>
</table>
Transport Mode

ORB

TCW

TSB
Transport Status Block

Interrogated TCW

TIDAW
Data

TIDAW
TIDAW
TIDAW
TIDAW

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

TIDAW
Data

TCCB
Transport Command & Control Block

TSB
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.
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

PRLO Process Logout
Hmm, 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/AnaheimEval
• Any Additional Questions?
Thank You For Your Time And Attention

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

Complete your sessions evaluation online at SHARE.org/AnaheimEval