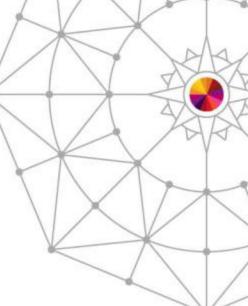




I/O Synergy – The Whole is Greater than the Parts (I/O Performance)

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

March 13, 2014 Session 15081





Introduction



- Synergy focus areas
- I/O Performance Related Topics
- For each topic:
 - Overview
 - Pointers/Tips
 - Triva/Interesting Facts

See url http://www.ibm.com/legal/copytrade.shtml for a list of trademarks



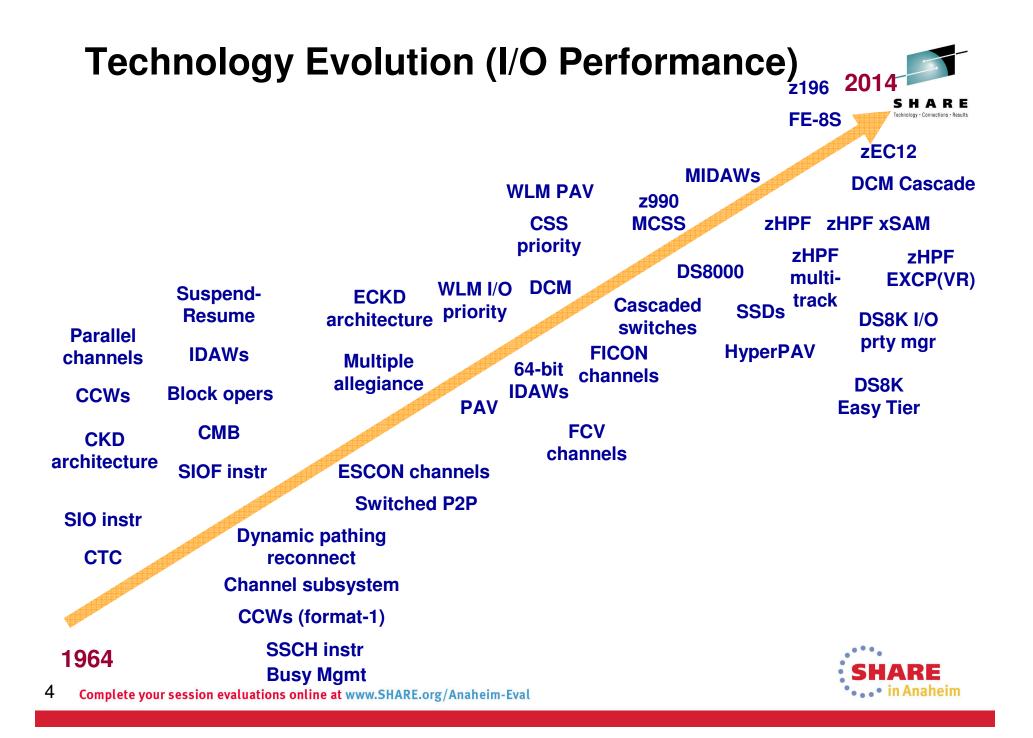
I/O and Storage Constraints

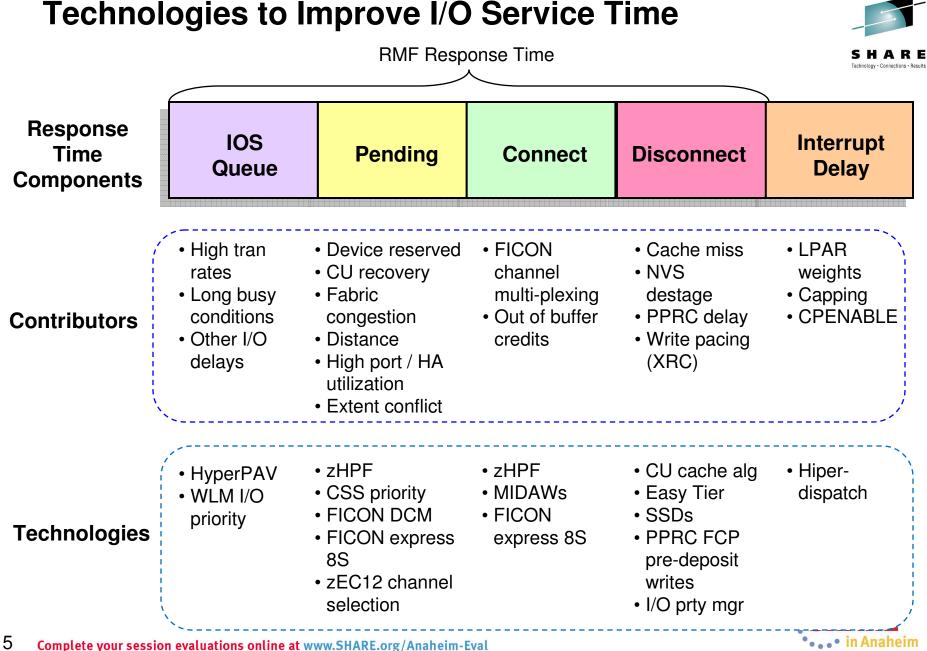


SHARE

Focus Area	Description	Client Impact
I/O Rates and Bandwidth	Ability to provide a balanced system for MIPs and I/O requests within the existing footprint and power envelope	Inability to achieve high CPU utilization because of I/O delays and concurrency limits. Inability to consolidate more physical footprints. Inability to grow the workloads on z/OS.
Latency	Fit within the existing batch window, ability to handle workload spikes	Bottlenecks will force clients to split workloads and spend resources managing around the I/O limits
Scalability	Number of devices, channels, devices per channel, aliases, I/O slots for accelerators, addressable storage capacity	Inability to add additional work and data to the existing z/OS infrastructure
Resiliency	Ability to recover from a channel path, device, storage controller, or site failure with minimal impact to the system and applications.	Application delays and outages
Complexity	I/O configuration planning, tuning, problem determination, capacity planning	Added cost and risk. Frequency of manual effort increases chance for errors, degrades resilience, and adds cost.







Technologies to Improve I/O Service Time

High Performance FICON for System z (zHPF)



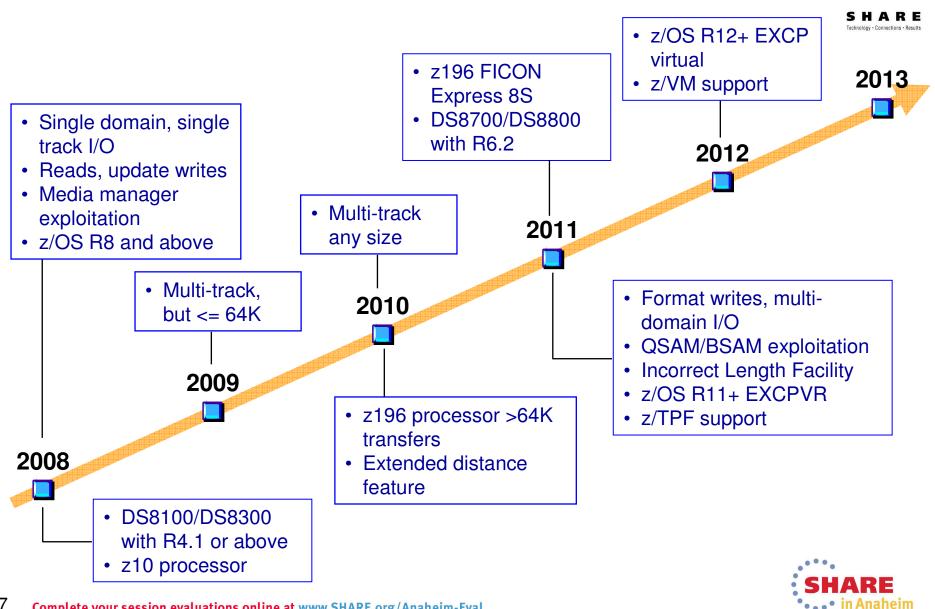
zHPF is a critical element of the System z I/O strategy

How does zHPF improve I/O performance?

- Leverages optimizations made in the HBA for FCP
- Uses a protocol that has less chit-chat between the channel and device
- Packages entire command stream into one or two sequences
- Packages data more efficiently into frames (less buffer credits used)
- H/W data router on FICON express 8S provides additional acceleration



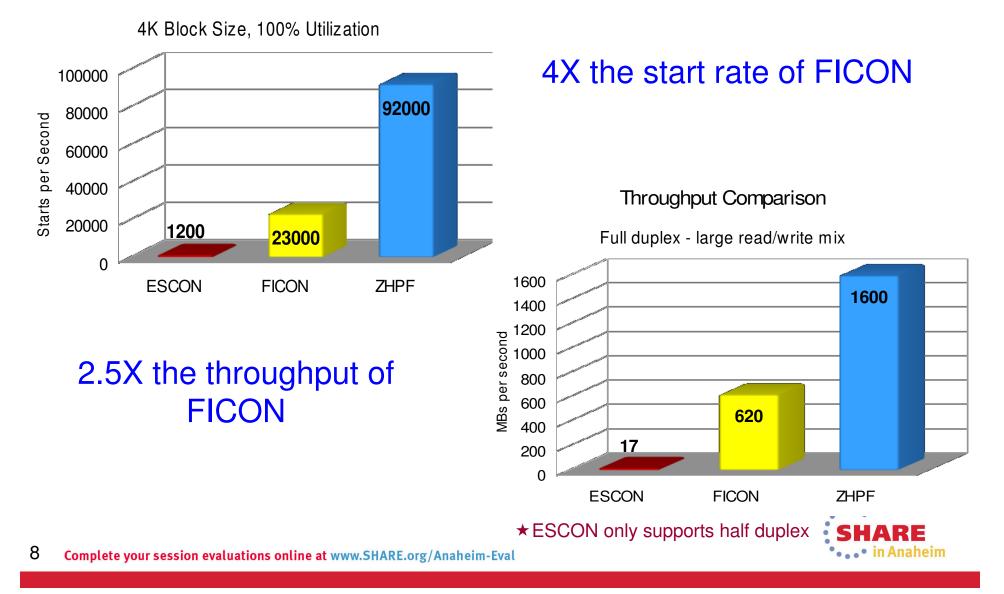
zHPF Evolution



zHPF/FICON/ESCON Comparison



Start Rate Comparison

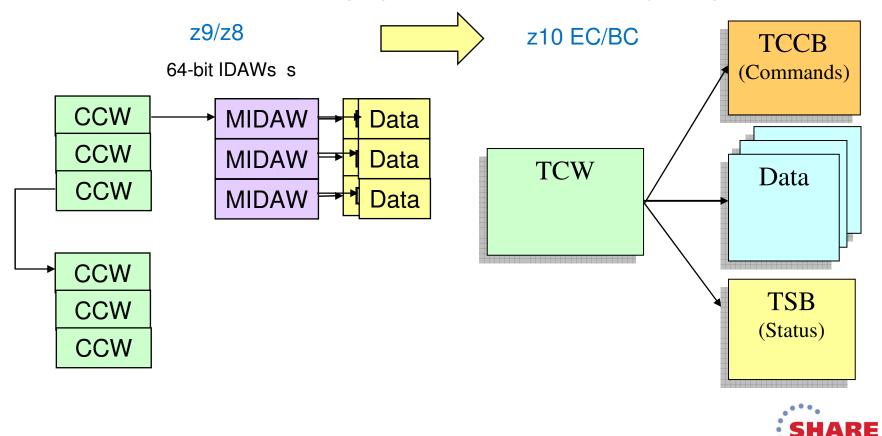


zHPF Trivia



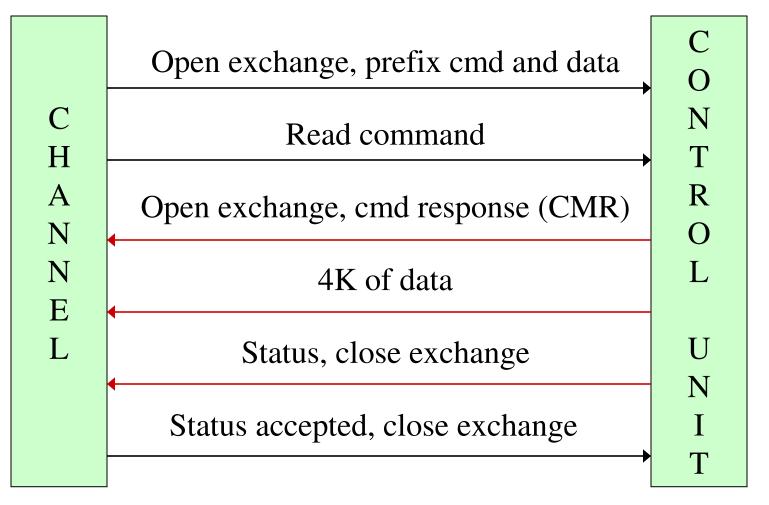
in Anaheim

zHPF is the first time we have dramatically changed the format of a channel program since System 360. Changes have been made to basic CCW channel program (e.g., format-1 CCWs, MIDAWs), and major changes have been made to the underlying architecture (XA, ESCON, FICON), but the channel program format was not radically changed.



FC Link Protocol for FICON





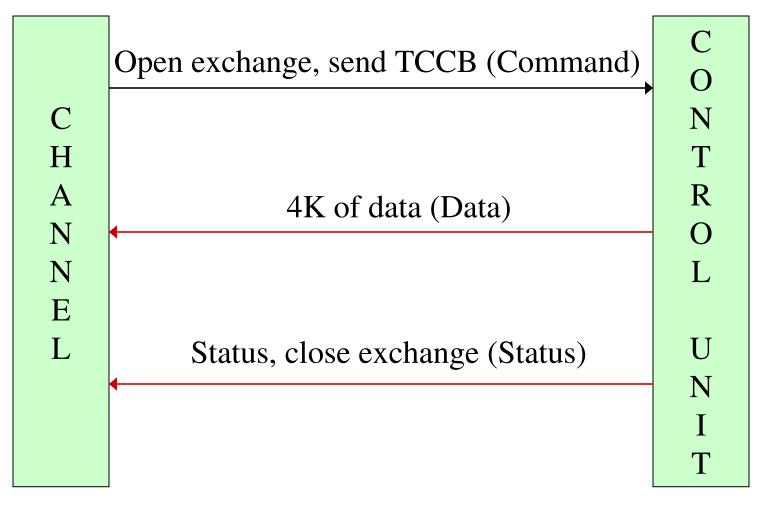
Example of a 4K read I/O request (6 sequences, 8 frames)



FC Link Protocol for zHPF (and FCP)



•••• in Anaheim



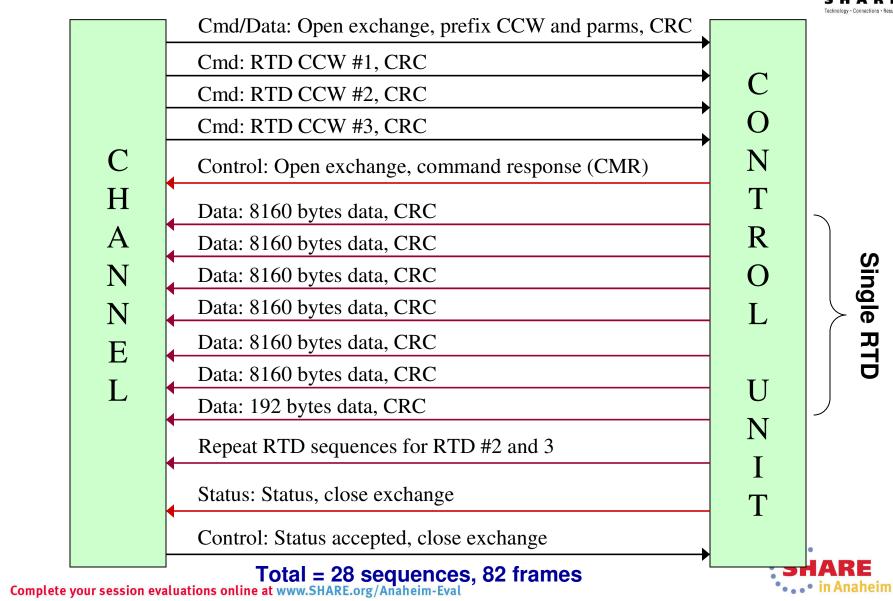
Example of a 4K read I/O request (3 sequences, 5 frames)

11 Complete your session evaluations online at www.SHARE.org/Anaheim-Eval



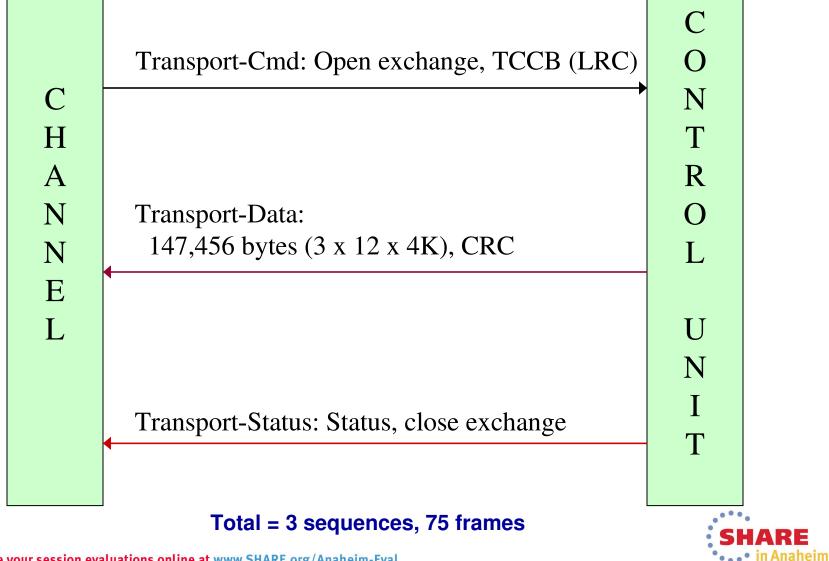
FICON – Read 3 Contiguous Tracks (12x4K Records)

12



zHPF – Read 3 Contiguous Tracks (12x4K Records)



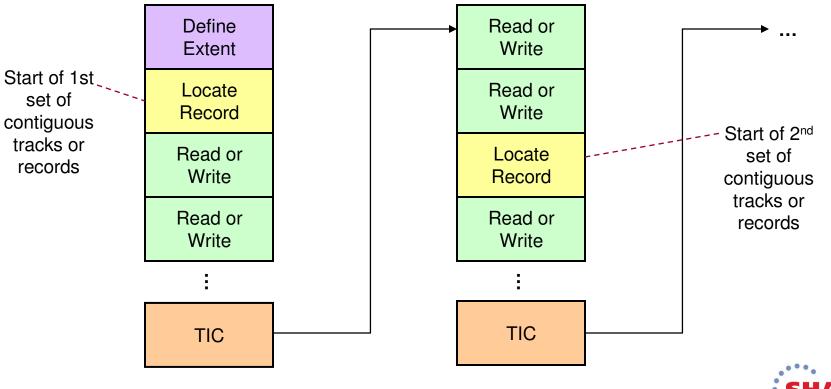


13 Complete your session evaluations online at www.SHARE.org/Anaheim-Eval

CCW Channel Programs – Virtually Unlimited in Size



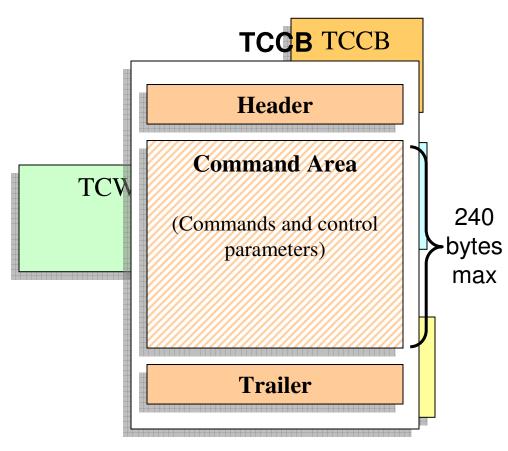
A single CCW channel program is virtually unlimited in size since CCWs can be chained together implicitly or explicitly. The size of a channel program is only limited by the amount of 31-bit virtual and real storage available.



zHPF Channel Programs – Limited by Architecture



The command block for zHPF is basically a SCSI CDB, which limits the size to 240 bytes. How can we pack enough information into a channel program so that it can be used for things like DB2 list prefetch and QSAM/BSAM I/O?



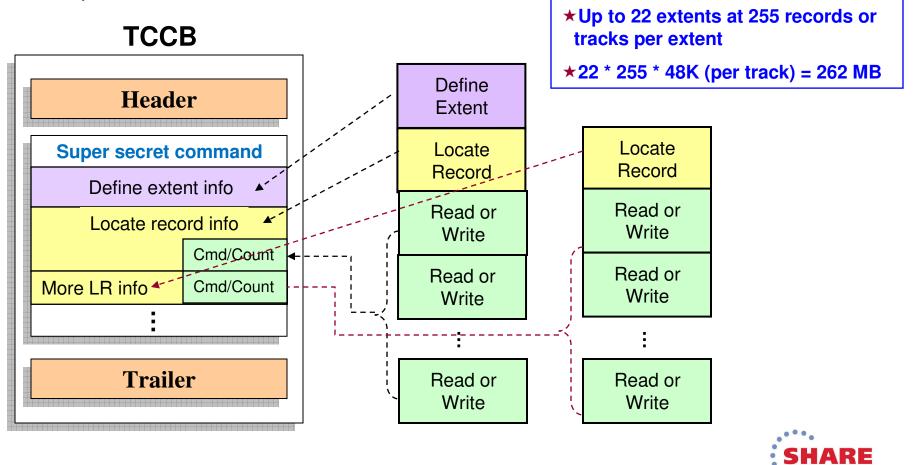


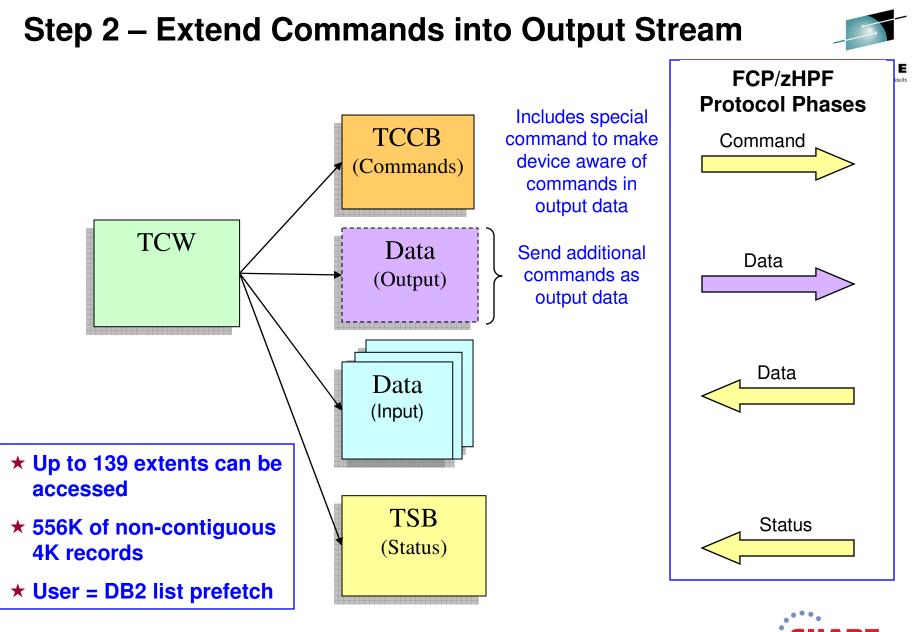
Step 1 – Collapse Lots of CCWs



In Anaheim

For CCW channel programs, z/OS collapses the define extent and locate record CCWs into a single CCW. For zHPF, we can also collapse the read/write CCWs as well as the subsequent locate records into a smaller area.





17 Complete your session evaluations online at www.SHARE.org/Anaheim-Eval

SHARE in Anaheim

zHPF Tips – How to Tell if zHPF is Used



- Channel path activity report shows FICON vs zHPF activity
 - This is the only place where zHPF usage is recorded

CHANNEL PATH UTILIZATION (%)			READ (MB/SEC)		WRITE (MB/SEC)		FICON OPERATIONS			ZHPF OPERATIONS				
ID TYPE	G SHR	PART	TOTAL	BUS	PART	TOTAL	PART	TOTAL	RATE	ACTIVE	DEFER	RATE	ACTIVE	DEFER
00 FC_S	5 Y	100.00	100.00	0.84	0.13	2.15	0.17	2.68	61.5	1.7	0.0	4.6	1.0	0.0
01 FC_S	5 Y	100.00	100.00	0.85	0.13	2.21	0.13	2.69	61.3	1.8	0.0	4.7	1.0	0.0
02 FC_S	4 Y	0.14	2.30	0.85	0.10	2.17	0.13	2.70	61.3	1.3	0.0	4.6	1.0	0.0
03 FC_S	4 Y	0.13	2.27	0.84	0.11	2.14	0.13	2.66	60.0	1.3	0.0	4.4	1.0	0.0
04 FC_S	5 Y	0.13	2.24	0.82	0.10	2.07	0.13	2.63	59.4	1.7	0.0	4.4	1.0	0.0
05 FC_S	5 Y	0.13	2.25	0.83	0.10	2.11	0.12	2.66	59.1	1.7	0.0	4.2	1.0	0.0
06 FC_S	4 Y	0.12	2.23	0.83	0.10	2.09	0.13	2.68	58.7	1.3	0.0	4.2	1.0	0.0

zHPF I/Os per second (physical channel)

Average number of active I/Os (open exchanges) -

Average number of deferred I/Os (no available open exchanges) -



zHPF Tips – Is zHPF Enabled?

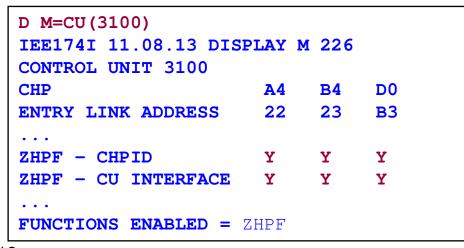


D IOS, ZHPF

IOS630I 13.35.33 ZHPF FACILITY 742 HIGH PERFORMANCE FICON FACILITY IS ENABLED

D M=DEV(3100)

IEE174I 11.32.54 DIS	PLAY M	965	
DEVICE 3100 STATUS	S=ONLIN	E	
CHP	A4	B4	D0
ENTRY LINK ADDRESS	22	2323	B3
DEST LINK ADDRESS	42	23B2	42
PATH ONLINE	Y	Y	Y
FUNCTIONS ENABLED =	MIDAW,	ZHPF	



- Enables zHPF at a system level, default is NO
- Device could be disabled for any number of reasons including CU LIC feature
- Certain I/O errors can disable zHPF at the device or CU level (these are usually a sign of a microcode problem)
- zHPF CHPID/CU interface should always be Y unless it is connected to an ESCON channel/interface or FICON express
- If zHPF doesn't appear in functions enabled, then either LIC feature not installed or microcode problem



19 Complete your session evaluations online at www.SHARE.org/Anaheim-Eval

zHPF Tips – Is zHPF Enabled?



D SMS, OPTIONS IGD002I 13:36:10 DISPLAY SMS 750 ACDS = SYS1SMS.DR5.ACDS COMMDS = SYS1SMS.DR5.COMMDS ACDS LEVEL = UNAVAIL SMS PARMLIB MEMBER NAME = IGDSMSDR INTERVAL = 15 DINTERVAL = 150 ... OAMTASK = PDSE_SYSEVENT_DONTSWAP = NO DB2SSID = SAM_USE_HPF = YES ...

 IGDSMSxx parmlib member indicates whether QSAM/BSAM should use zHPF.

DEVSERV, QDASD, 7000, RDFEATS IEE459I 11.57.32 DEVSERV QDASD 848 UNIT VOLSER SCUTYPE DEVTYPE CYL SSID... 07000 ----- 2107951 2107900 1113 7000... FEATURE CODES AT V020A0AF8 00070006001F00A6 DDFFFDE164400000 FD8000000000 C3D000BB0000000 E785C000000000 AEC0000000000

 zHPF capability/feature bits (+x'28'). At a minimum, bits 0 & 2 should be on. Other bits may also need to be on for other exploitation (e.g, bit 5 for QSAM/BSAM). Contact IBM support for specific questions.



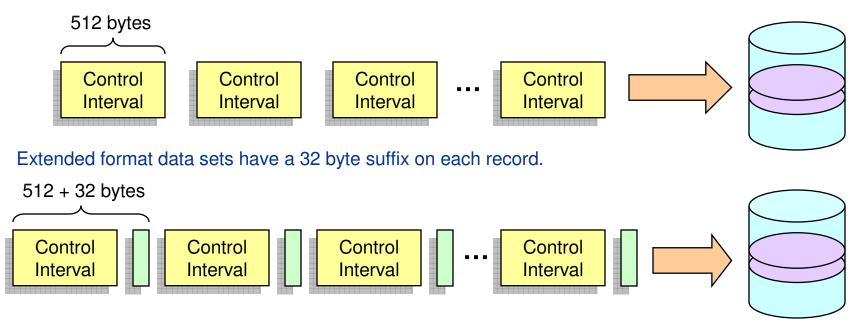
zHPF Trivia



Although the channel supported a maximum transfer size of 2M in the initial zHPF support in the z10 processor, for performance reasons, the software limited the amount of data transferred to 68K. Where did this number come from?

Around 64K is where performance issues occurred.

512 bytes is the minimum VSAM control interval size. 128 of these fit into 64K (2 tracks are required).



C Therefore, we need to add (64K / 512) * 32 = 4096 to limit to allow 64K transfers for extended format data sets to eligible for zHPF.



21 Complete your session evaluations online at www.SHARE.org/Anaheim-Eval

FICON Express 8S Channel

- zHPF protocol implemented in the hardware via a data router
- 76% higher start rates compared to zHPF on FICON express 8
- 107% higher throughput compared to zHPF on FICON express 8
- 100% of DB2 I/O can be converted to zHPF

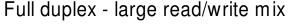
zHPF Start Rate Comparison

4K Block Size, 100% Utilization

100000 1600 1600 1400 92000 80000 I/Os per Second 1200 per second 60000 1000 800 52000 40000 770 600 MBs 400 20000 200 0 0 FICON Express 8 FICON Express 8S FICON Express 8 FICON Express 8S



zHPF Throughput Comparison



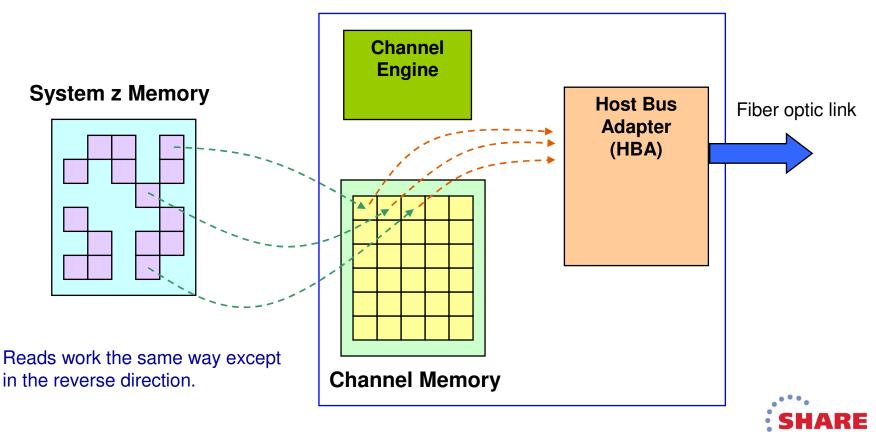
• in Anaheim

Older FICON Channels – Store and Forward



Anaheim

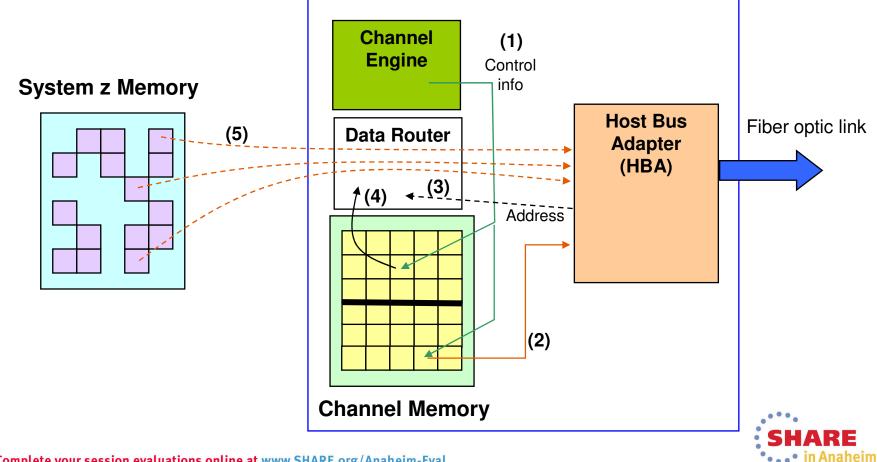
- ★ In older channels, the HBA did not access System z memory directly. A store and forward technique was used where the channel firmware would fetch information from System z memory into channel memory, which would then be accessed by the HBA.
- ★ A pipeline is established which allowed the channel to work in parallel with the HBA.



FICON Express 8S – H/W Data Router



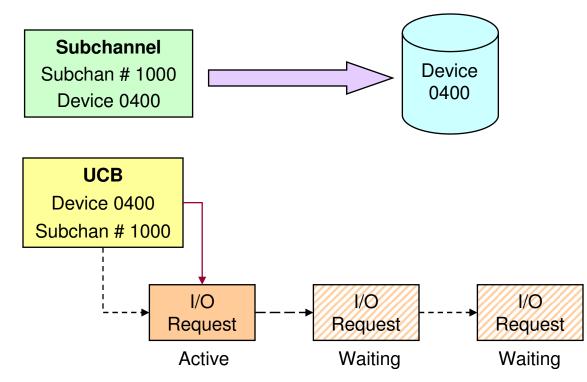
- ★ With the FICON express 8S channel, the HBA accesses System z memory directly through and under the control of a H/W data router.
- ★ The channel firmware is involved only in setting up the storage areas that need to be accessed in channel memory.



Parallel Access Volumes (PAV) – Before PAV



Although, storage controllers allow multiple I/Os to be issued to a single device from multiple systems (multiple allegiance), System z allows only one I/O to be issued to a device (subchannel in HSA) at a time for an LPAR.



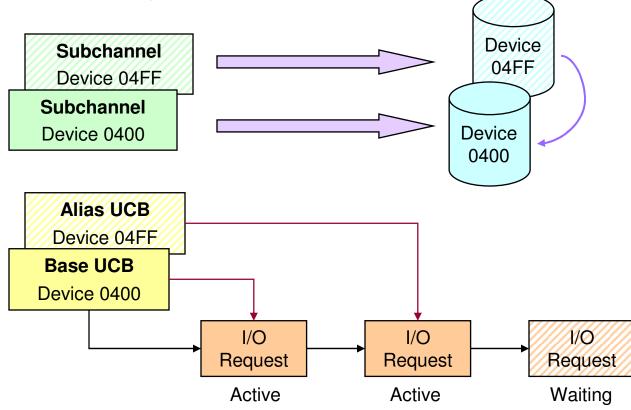
If multiple I/Os arrive for a device at the same time, the remaining I/O's will be queued internally in z/OS while waiting for the prior I/Os to complete (IOSQ time).



Static PAV



Static PAV allowed multiple *alias* devices to be defined for a *base* device. This allowed multiple I/Os to be done to a particular volume in parallel since a separate subchannel was assigned for each alias. However, the alias to base assignment was static, so this required careful planning to ensure that a base had adequate aliases defined.

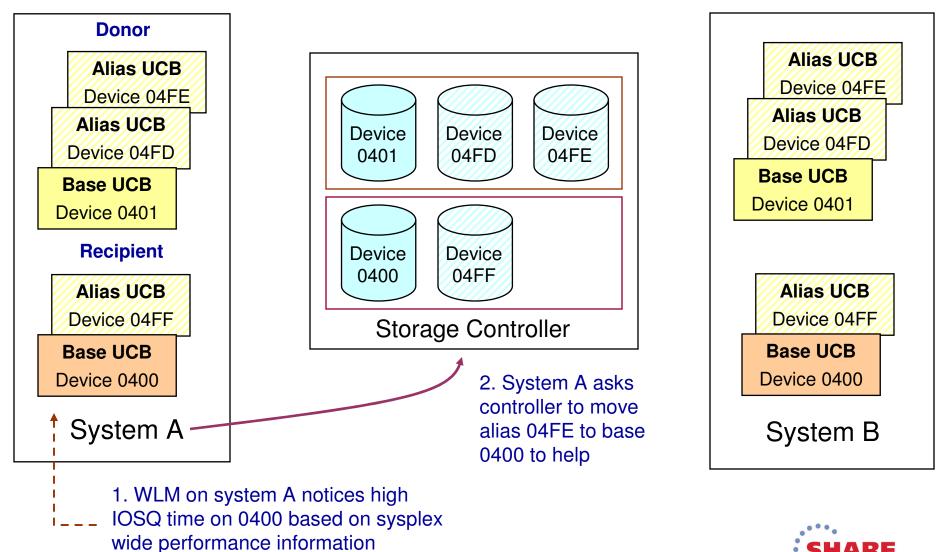




Dynamic (WLM Managed) PAV



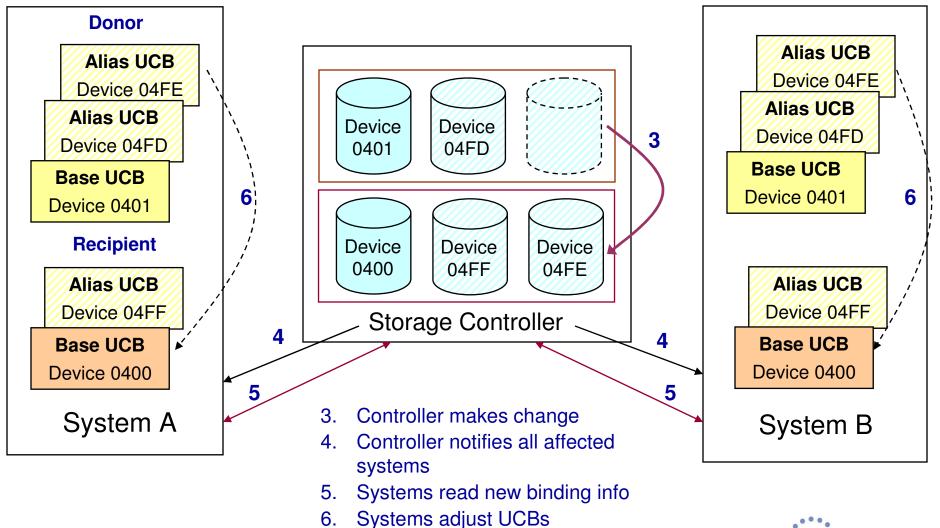
in Anaheim



27 Complete your session evaluations online at www.SHARE.org/Anaheim-Eval

Dynamic (WLM Managed) PAV



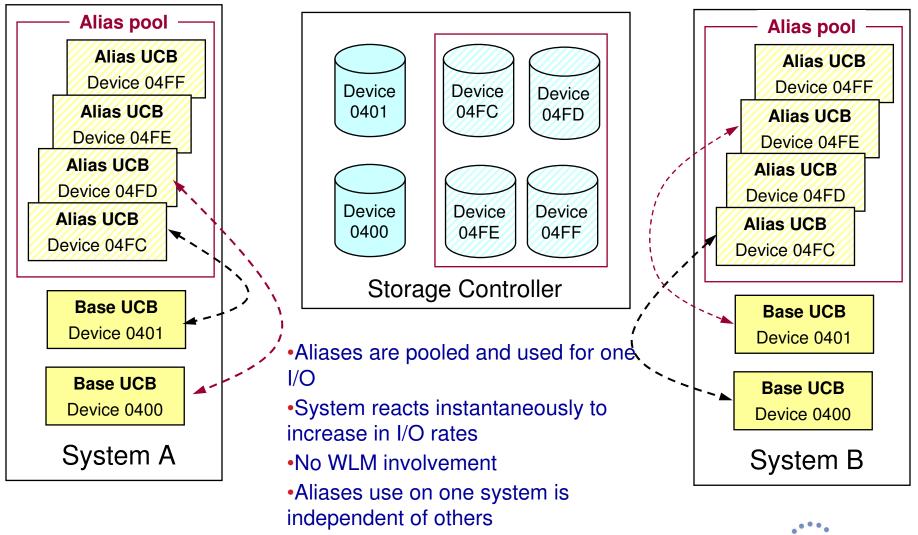




HyperPAV



in Anaheim





HyperPAV Tips – How Many Aliases?



- Little's Law:
 - Number of active requests = I/O rate * I/O service time
- Applied to HyperPAV:
 - No. of aliases = Peak avg. I/O rate * service time * multiplier
- Multiplier added since peak <u>average</u> service time doesn't account for workload spikes
 - Multiplier should be at least 2 since peak concurrent I/Os will be higher than the average
 - Monitor for increases in IOSQ time or HPAV wait delay (see next page)



HyperPAV Tips - Monitoring



			DI	RECT	A	ССЕ	SS DI	EVI	СЕ	AC!	r I V	ITY	C .		
	Z/	/OS V1R:	13	CONVER			O TRX2 OS V2R1 RI	MF		ATE 1: IME 1:					
STORAGE	DEV	DEVICE	NUMBER	VOLUME	PAV	LCU	ACTIVITY	RESP	IOSQ	CMR	DB	INT	PEND	DISC	CONN
GROUP	NUM	TYPE	OF CYL	SERIAL			RATE	TIME	TIME	DLY	DLY	DLY	TIME	TIME	TIME
XTEST	2209	33903	3339	TRXSXA	1	0032	24.901	6.54	2.27	.202	.000	.020	.212	2.96	1.10
	220A	33909	10017	TRXT01	1	0032	20.085	6.12	2.06	.207	.000	.019	.217	2.80	1.04
	220B	33909	10017	TRXT02	1	0032	30.702	7.18	2.28	.210	.000	.013	.220	3.73	0.95
	220C	33909	10017	TRXT03	1	0032	27.103	6.93	2.03	.201	.000	.016	.211	3.67	1.02
	220D	33909	10017	TRXT04	1	0032	31.040	6.30	2.25	.208	.000	.022	.218	2.85	0.98
	220E	33909	10017	TRXT05	1	0032	33.010	6.62	2.31	.203	.000	.017	.213	3.08	1.01

First sign that I may not have enough aliases defined for this LCU is high IOSQ time.

Note that there may be other performance problems that are causing high IOSQ queue. For example, high disconnect time due to staging data into cache. Adding aliases won't necessarily help in these cases.





	z	/OS V1R13		SYSTEM	ID TEST	1		START				
				RPT VERSION V1R13 RMF				END				
TOTAL	SAMPLES	= 901 1	IODF = 72	CR-DAT	CR-TIME:							
							AVG	AVG		AVG		
LCU	CU	DCM GROUP	CHAN	CHPID	% DP	% CU	CUB	CMR		CSS	HPA	V
		MIN MAX DEF	PATHS	TAKEN	BUSY	BUSY	DLY	DLY		DLY	WAIT	MAX
0 A 00	2600		0C	0.003	0.00	0.00	0.0	0.1				
			OF	0.007	0.00	0.00	0.0	0.0				
			42	0.007	0.00	0.00	0.0	0.2				
			45	0.007	0.00	0.00	0.0	0.6				
			00	0.007	0.00	0.00	0.0	1.2				
			03	0.007	0.00	0.00	0.0	0.5				
			06	0.003	0.00	0.00	0.0	1.3				
			09	0.003	0.00	0.00	0.0	2.2			\frown	\frown
			*	0.043	0.00	0.00	0.0	0.6		0.1	0.000	10
ava	ilable	mber of I/O r	·							ic		
		umber of alia		-			ecar		.11 (11	15 —	SH	ARE



```
d m = dev(0710)
SY1
SY1
     IEE174I 23.35.49 DISPLAY M 835
DEVICE 0710
              STATUS=ONLINE
                      10
                                     40
CHP
                           20
                                30
DEST LINK ADDRESS
                      10
                           20
                                30
                                     40
PATH ONLINE
                           Y
                      Y
                                Y
                                     Y
CHP PHYSICALLY ONLINE Y
                           Y Y
                                     Y
                           Y
                                Y
                                     Y
PATH OPERATIONAL
                      Y
MANAGED
                           Ν
                                N
                                     Ν
                      Ν
                      0700 0700 0700 0700
CU NUMBER
MAXIMUM MANAGED CHPID(S) ALLOWED:
                                   0
DESTINATION CU LOGICAL ADDRESS = 07
SCP CU ND
                  = 002107.000.IBM.TC.03069A000007.00FF
                  = 002107.900.IBM.TC.03069A000007.0700
SCP TOKEN NED
SCP DEVICE NED
                  = 002107.900.IBM.TC.03069A000007.0710
HYPERPAV ALIASES IN POOL 32
```

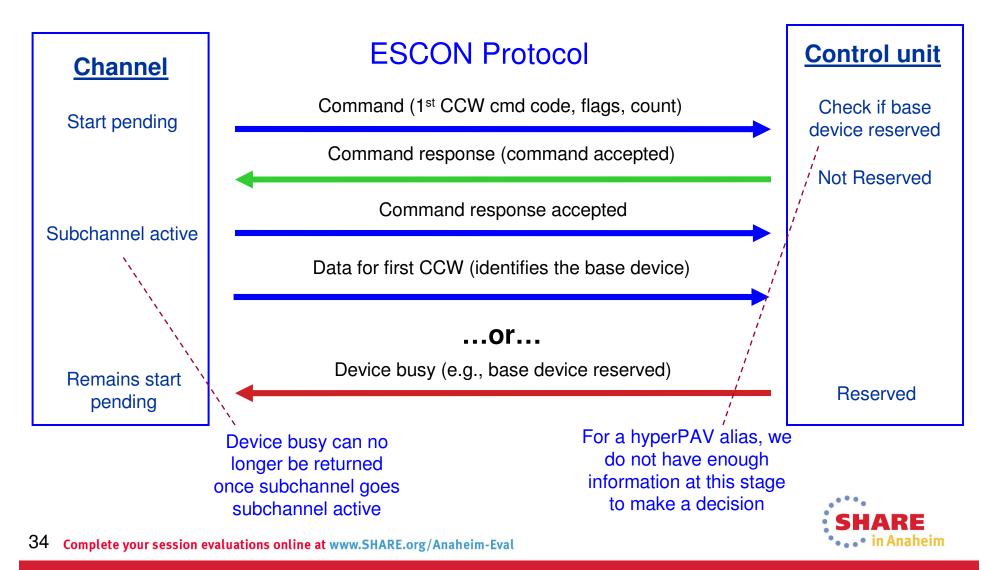
Sometimes an I/O configuration or z/OS issue will cause aliases not to be available for use. You should compare this number with what you have defined for your LCU.



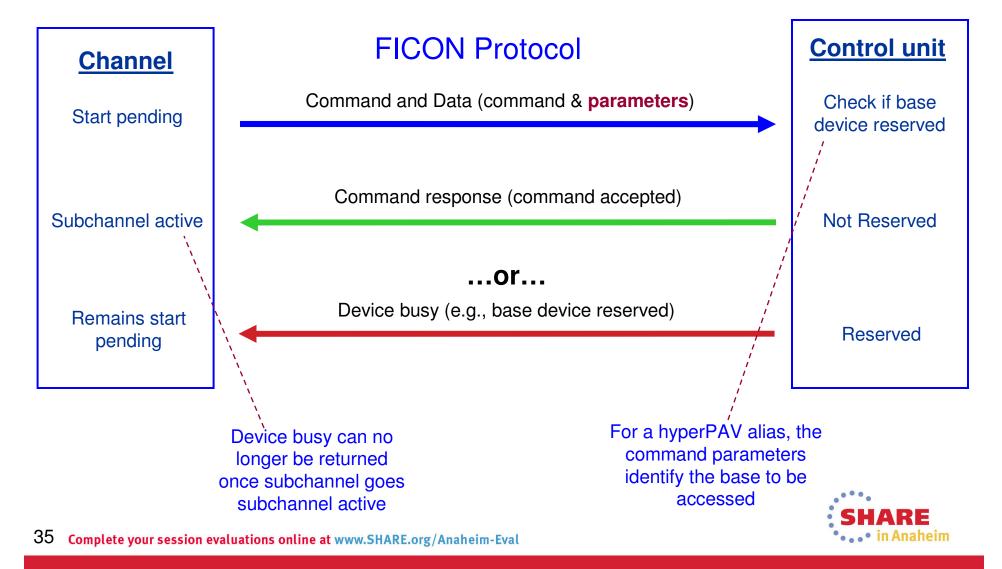
PAV Trivia



Why is HyperPAV supported only for FICON channels?



Why is HyperPAV supported only for FICON channels?





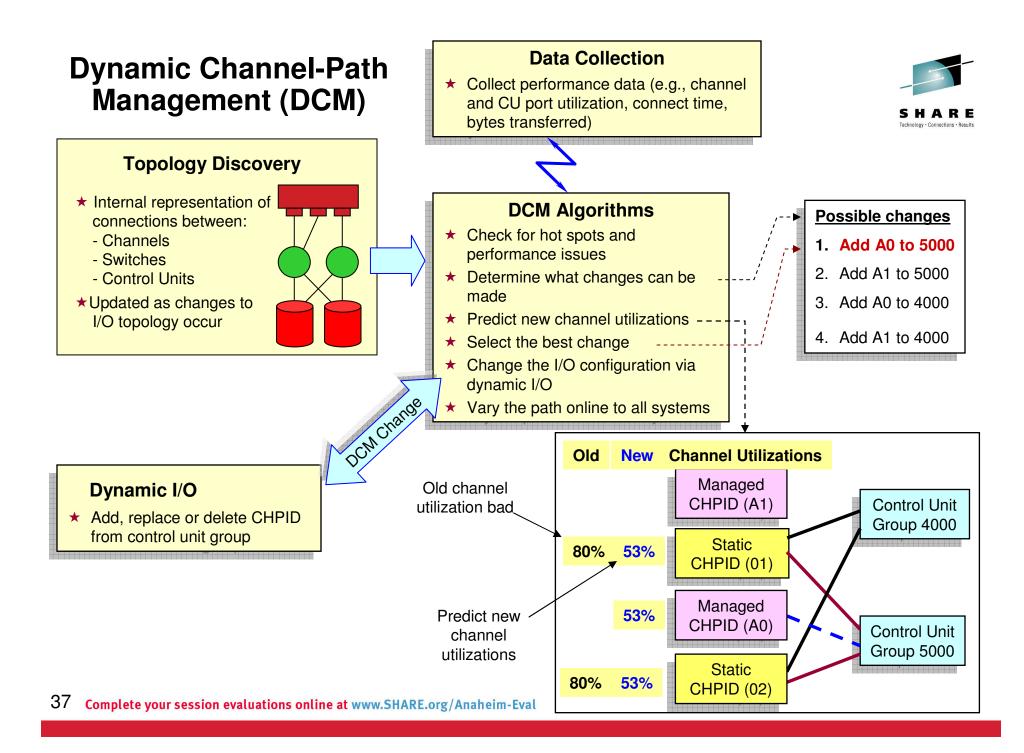
PAV Trivia

Dynamic Channel-Path Management (DCM)



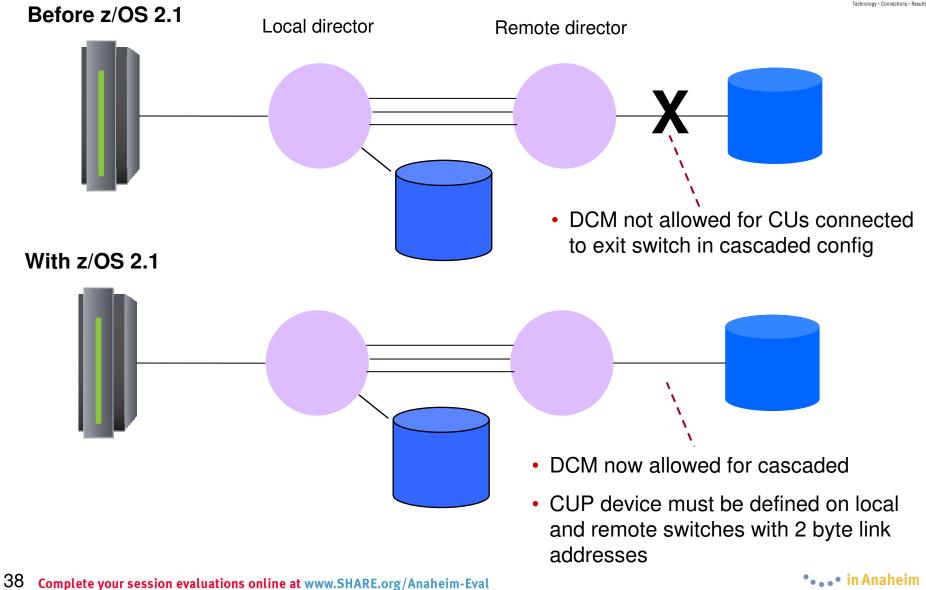
- Goals
 - Simplify I/O configuration management
 - Improve performance management
 - Enhance RAS
- System manages a pool of channels that are dynamically added to control units as needed
 - Managed channel
 - Managed control unit
- System will add or move channels based on:
 - Channel utilization hot spots
 - Port utilization hot spots
 - Availability need down to your last channel path





DCM Multi-Switch Cascade





zEC12 Enhanced Channel Path Selection



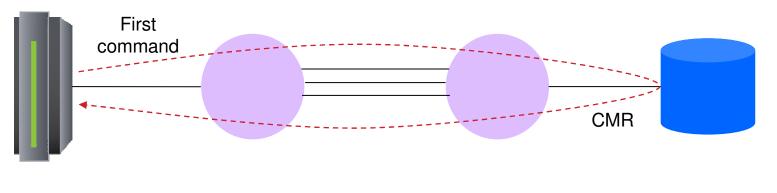
- Overall I/O performance can be affected by poor performance on one channel path
 - See RMF I/O queuing report
- There are many possible causes:
 - Multi-system workload spikes
 - SAN congestion or destination port congestion
 - Control unit host adapter or port congestion
 - Firmware failures in the SAN, channel extenders, control units
 - Hardware failures (link speeds did not initialize correctly)
 - Mis-configuration and cabling errors
- Channel subsystem will now balance paths based on initial command response time



What is Initial Command Response Time?



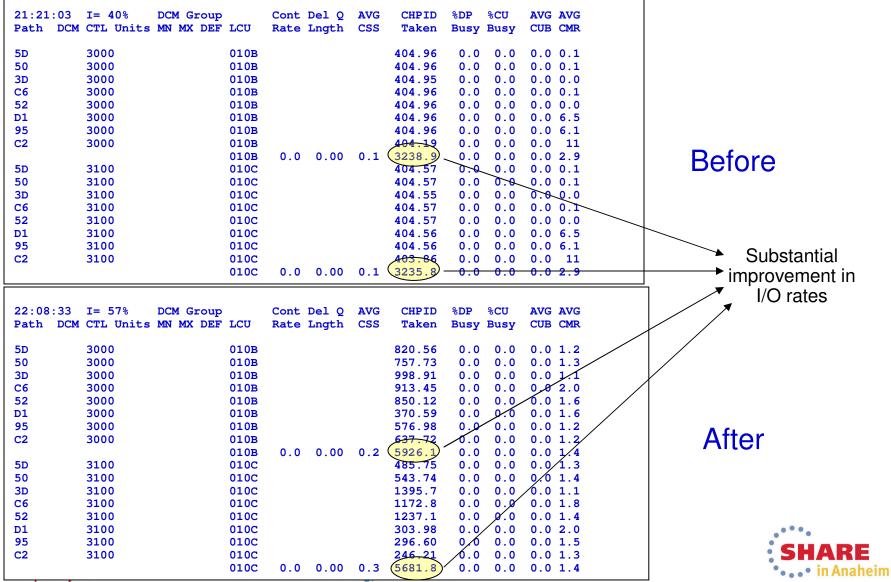
- Initial command response (CMR) time is the amount of time from when the channel sends the first command until it gets a response from the control unit
 - One round trip through the fabric
 - Originally created to distinguish between IFCCs that occur because of congestion vs lost frames
 - Good for detecting fabric congestion and other problems on a path



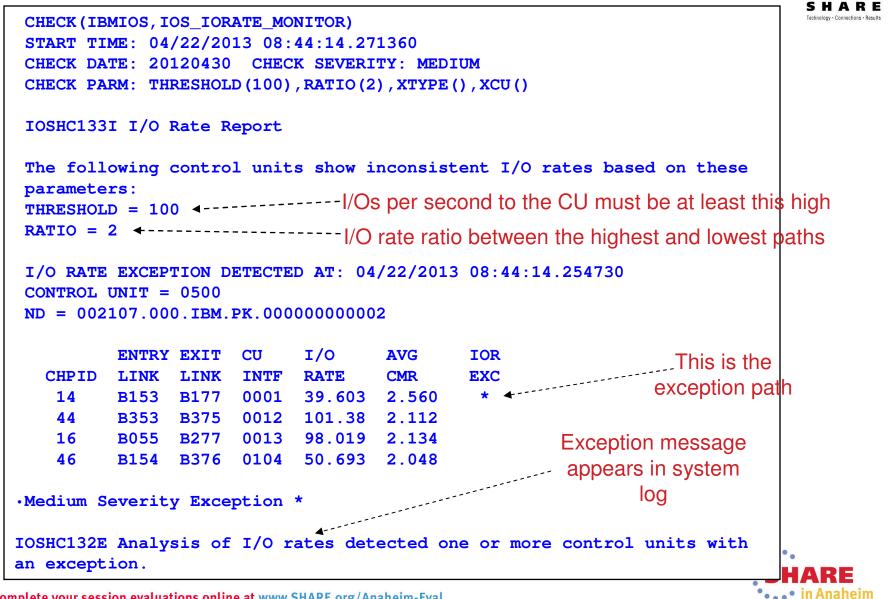


zEC12 Enhanced Channel Path Selection





I/O Rate Health Check (OA40508)



42 Complete your session evaluations online at www.SHARE.org/Anaheim-Eval

Additional Reading



- Evolution of the System z Channel Presentation given at Share August 2011 by Patty Driever and Howard Johnson
- IBM zEnterprise 196 and IBM zEnterprise 114 I/O and FICON Express8S Channel Performance
- FICON DCM for System Programmers Whitepaper on Techdocs
- IBM zEnterprise Storage I/O Advancements IBM Journal of Research and Development, Vol 56





Thank you





44 Complete your session evaluations online at www.SHARE.org/Anaheim-Eval



Appendix

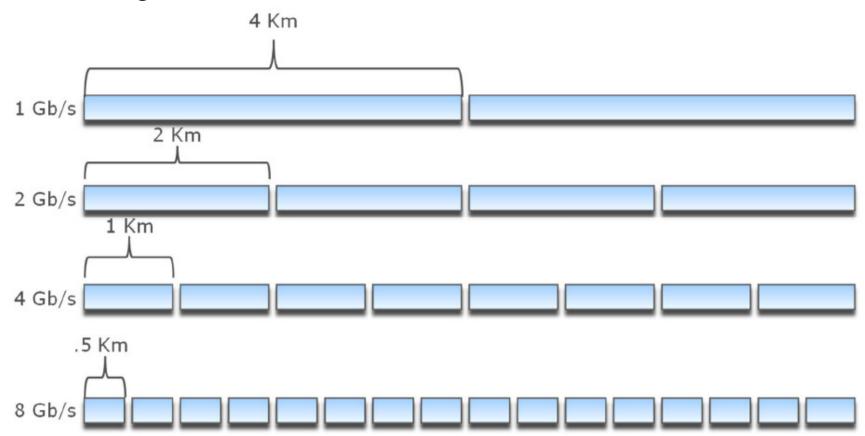


45 Complete your session evaluations online at www.SHARE.org/Anaheim-Eval

FICON Trivia



How long is a fiber channel frame?

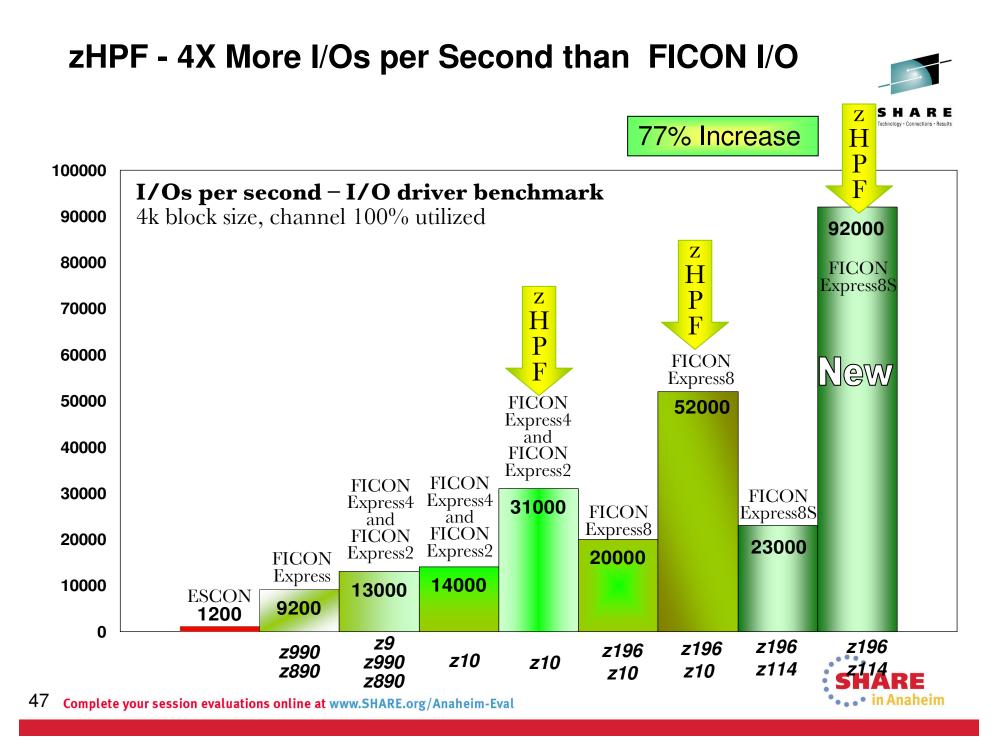


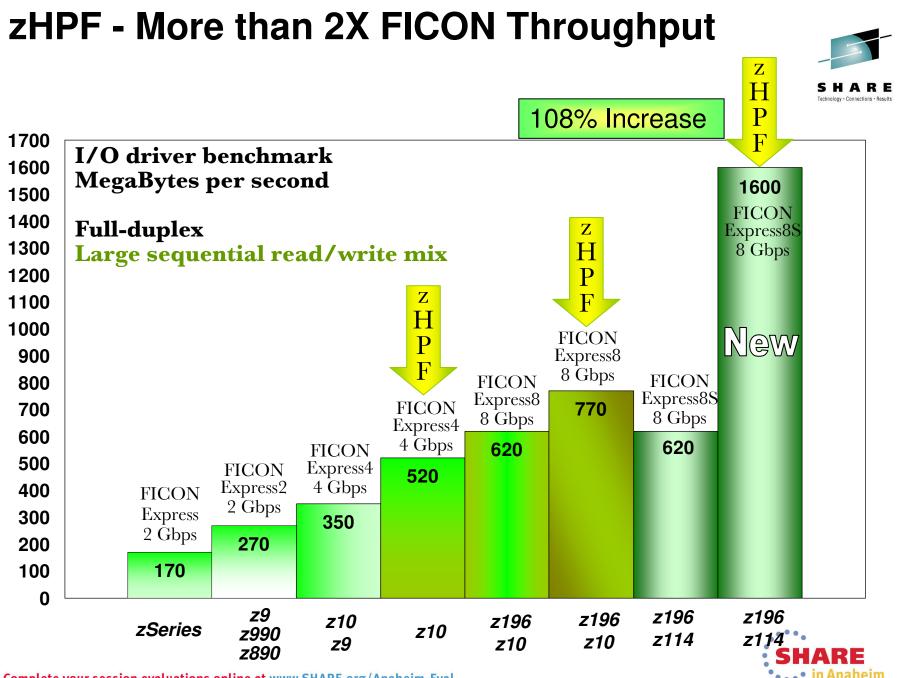
★ The fibre channel frame is the basic building block in FICON. A frame can have a payload up to 2112 bytes. Each byte is represented by 10 bits (8b/10b encoding).



46 Complete your session evaluations online at www.SHARE.org/Anaheim-Eval

Slide by Howard Johnson





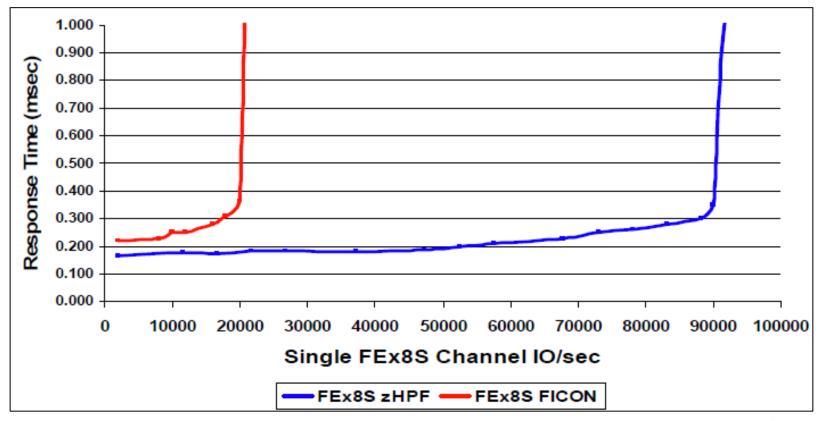
48 Complete your session evaluations online at www.SHARE.org/Anaheim-Eval

•••• in Anaheim

zHPF - Response Time Improves Too



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





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.75.000000WH391
         ENTRY EXIT
                    CU
                           I/O
                                   AVG
                                                       These are the
  CHPID LINK LINK INTF RATE
                                   CMR
                                             exception paths
         2C51 2DC4 0030 72.330 9.21 ←
   81
         3C1B 3DC2 0031 71.651 9.47 -
   22
                                                         Exception
         2C52 2DC0 0032 72.333 8.70
   82
               2DCC 0100 71.810 1.92
   84
         2C54
                                                     message appears
   21
         3C19
               3DD2 0231 72.122 1.79
                                                        in system log
* Medium Severity Exception *
 IOSHC112E Analysis of command response (CMR) time detected one or
                                                                         ....
more control units with an exception.
                                                                         • • • in Anaheim
```

