



IBM Inside Sales

International Technical Support Organization Global Content Services

ITSO – z System Hardware Workshop

www.ibm.com/redbooks

Part 4 – Native PCIe Adapters – zEDC and RoCE



Trademarks

The following are trademarks of the International Business Machines Corporation in the United States, other countries, or both.

Not all common law marks used by IBM are listed on this page. Failure of a mark to appear does not mean that IBM does not use the mark nor does it mean that the product is not actively marketed or is not significant within its relevant market.

Those trademarks followed by ® are registered trademarks of IBM in the United States; all others are trademarks or common law marks of IBM in the United States.

For a complete list of IBM Trademarks, see www.ibm.com/legal/copytrade.shtml :

BladeCenter®, DB2®, e business(logo)®, DataPower®, ESCON, eServer, FICON, IBM®, IBM (logo)®, MVS, OS/390®, POWER6®, POWER6+, POWER7®, Power Architecture®, PowerVM®, /390®, System p®, System p5, System x®, z Systems®, System z9®, System z10®, WebSphere®, X-Architecture®, zEnterprise®, z9®, z10®, z196®, z114®, zEnterprise System z196®, zEnterprise System z114®, zEnterprise System zEC12®, zEnterprise System zBC12®, z13®, z/Architecture®, z/OS®, z/VM®, z/VSE®, zSeries®

The following are trademarks or registered trademarks of other companies.

Adobe, the Adobe logo, PostScript, and the PostScript logo are either registered trademarks or trademarks of Adobe Systems Incorporated in the United States, and/or other countries. Cell Broadband Engine is a trademark of Sony Computer Entertainment, Inc. in the United States, other countries, or both and is used under license therefrom.

Java and all Java-based trademarks are trademarks of Sun Microsystems, Inc. in the United States, other countries, or both.

Microsoft, Windows, Windows NT, and the Windows logo are registered trademarks of Microsoft Corporation in the United States, other countries, or both.

Intel, Intel logo, Intel Inside, Intel Inside logo, Intel Centrino, Intel Centrino logo, Celeron, Intel Xeon, Intel SpeedStep, Itanium, and Pentium are trademarks or registered trademarks of Intel Corporation or its subsidiaries in the United States and other countries.

UNIX is a registered trademark of The Open Group in the United States and other countries.

Linux is a registered trademark of Linus Torvalds in the United States, other countries, or both.

ITIL is a registered trademark, and a registered community trademark of the Office of Government Commerce, and is registered in the U.S. Patent and Trademark Office.

IT Infrastructure Library is a registered trademark of the Central Computer and Telecommunications Agency, which is now part of the Office of Government Commerce.

* All other products may be trademarks or registered trademarks of their respective companies.

Notes:

Performance is in Internal Throughput Rate (ITR) ratio based on measurements and projections using standard IBM benchmarks in a controlled environment. The actual throughput that any user will experience will vary depending upon considerations such as the amount of multiprogramming in the user's job stream, the I/O configuration, the storage configuration, and the workload processed. Therefore, no assurance can be given that an individual user will achieve throughput improvements equivalent to the performance ratios stated here.

IBM hardware products are manufactured from new parts, or new and serviceable used parts. Regardless, our warranty terms apply.

All customer examples cited or described in this presentation are presented as illustrations of the manner in which some customers have used IBM products and the results they may have achieved. Actual environmental costs and performance characteristics will vary depending on individual customer configurations and conditions.

This publication was produced in the United States. IBM may not offer the products, services or features discussed in this document in other countries, and the information may be subject to change without notice. Consult your local IBM business contact for information on the product or services available in your area.

All statements regarding IBM's future direction and intent are subject to change or withdrawal without notice, and represent goals and objectives only.

Information about non-IBM products is obtained from the manufacturers of those products or their published announcements. IBM has not tested those products and cannot confirm the performance, compatibility, or any other claims related to non-IBM products. Questions on the capabilities of non-IBM products should be addressed to the suppliers of those products.

Prices subject to change without notice. Contact your IBM representative or Business Partner for the most current pricing in your geography.

Important information about today's workshop

- The ITSO z hardware team created 7 IBM z13 presentations to be delivered today
 - Part 1 – IBM z13 – Positioning / introduction
 - Part 2 – z13 CPC Details Capacity and Performance
 - Part 3 – z13 I/O Subsystem
 - **Part 4 – Native PCIe Adapters – zEDC and RoCE (what's new only)**
 - Part 5 – HMC, CoD, RAS and zAware
 - Part 6 – Installation Planning
 - Part 7 – Software Support

- The main references for the presentations today are:
 - IBM z13 Technical Guide – Redbook – SG24-8251
 - IBM z13 Technical Introduction – Redbook - SG24-8250

- **Part of the available material may not be presented..** 😞
 - Even if we don't cover the presentations entirely,
 - The material can be download from:
 - <http://www.redbooks.ibm.com/Redbooks.nsf/pages/addmats>

- **The material being presented may not fully match the copied version you have**

- **You can always get the latest version .. If you want it, just ask !** 😊

- **Please ask questions, make comments and share your own experiences at any time**

- **Thank You !**

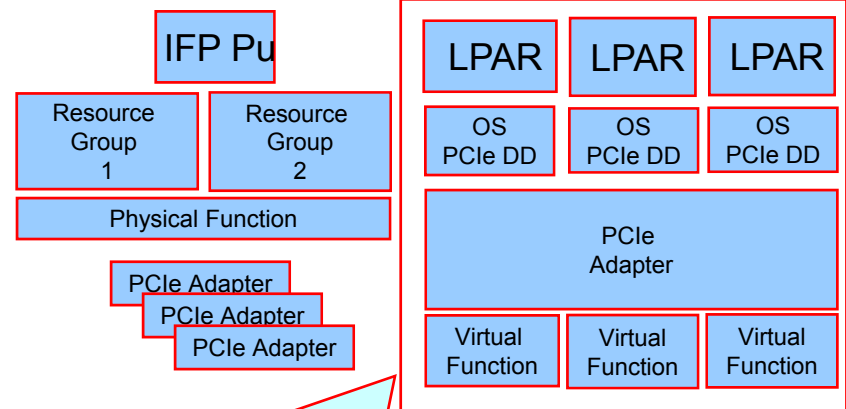
Native PCIe Adapter

Prior servers all have features (FICON, OSA, ISC, Crypto, Flash Express) that utilize the I/O infrastructure. They have an ASIC on the feature that handles virtualization, recovery, diagnostics, failover, and concurrent firmware updates



These features also have an adaptation layer that maps the z System I/O architecture to a PCIe adapter
The IBM ASIC, in addition to system firmware, provides the adaptation layer that maps traditional z System I/O instructions such as Start Subchannel to the unique internal architecture of the feature. The adaptation layer changes, as the physical hardware and functions are updated. Each Operating System (OS) is not required to provide changes (such as updates to unique device driver per OS) and can maintain the same I/O instructions

With zEC12 at GA2, zBC12 and z13, z System introduced two features, 10GbE RoCE Express and zEDC Express, with industry standard PCIe adapters. The adaptation layer and the associated ASIC is no longer needed. With the elimination of the adaptation layer, these features are designed to offer significant performance improvements. These features with native PCIe adapters physically plug into a mother card that provides Vital Product Data (VPD), hot plug capability, etc. The features then plug into the PCIe I/O drawer.

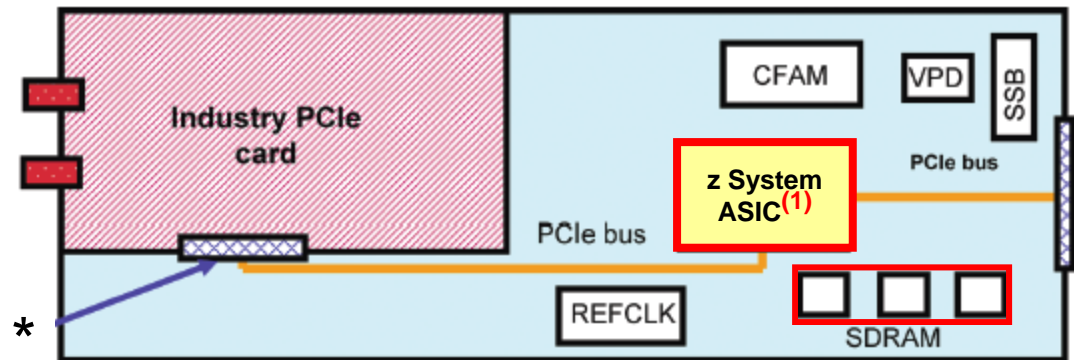


The adapter management functions (such as diagnostics, recovery and firmware updates) are handled differently by the system. There is a new dedicated, non-customer PU, called the **Integrated Firmware Processor (IFP)** which will perform these and other physical functions in a management firmware LPAR. The logical location of the management firmware LPAR is called the **Resource Group (RG)**. For resiliency, there are two RGs per system sharing the IFP.

PCIe I/O Features – Introducing new “Native” PCIe zEDC Express and 10GbE RoCE Express

Traditional z System I/O PCIe Feature

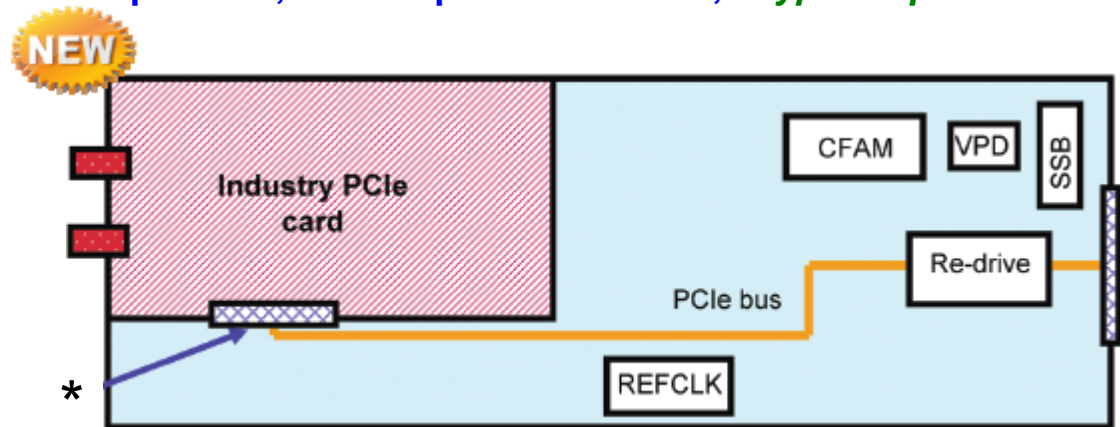
- One z System ASIC per Channel/PCHID
- Definition and LPAR Assignment
 - HCD/IOCP CHPID definition or
 - Alternate definition outside HCD/IOCP is possible for some. For example: *Crypto Express4S is not defined as a CHPID*
- Virtualization and support by Channel Subsystem LIC on System Assist Processors (SAPs)



Traditional z System I/O PCIe Features: FICON Express8S, OSA-Express5S and 4S, *Crypto Express4S*

Native PCIe Features

- z System ASIC role moved to the new z System I/O Controller (zIOC) in the PCIe I/O fanout and to the PCI Support Partition.
- Definition and LPAR Assignment
 - HCD/IOCP FUNCTION definition similar to CHPID definition but with different rules or
 - Alternate definition outside HCD/IOCP is possible for some. For example: *Flash Express is not defined with FUNCTIONS*
- Virtualization and support by the zIOC and Redundancy Group LIC running on the Integrated Firmware Processor (IFP) (*Note: NOT applicable to Flash Express*)

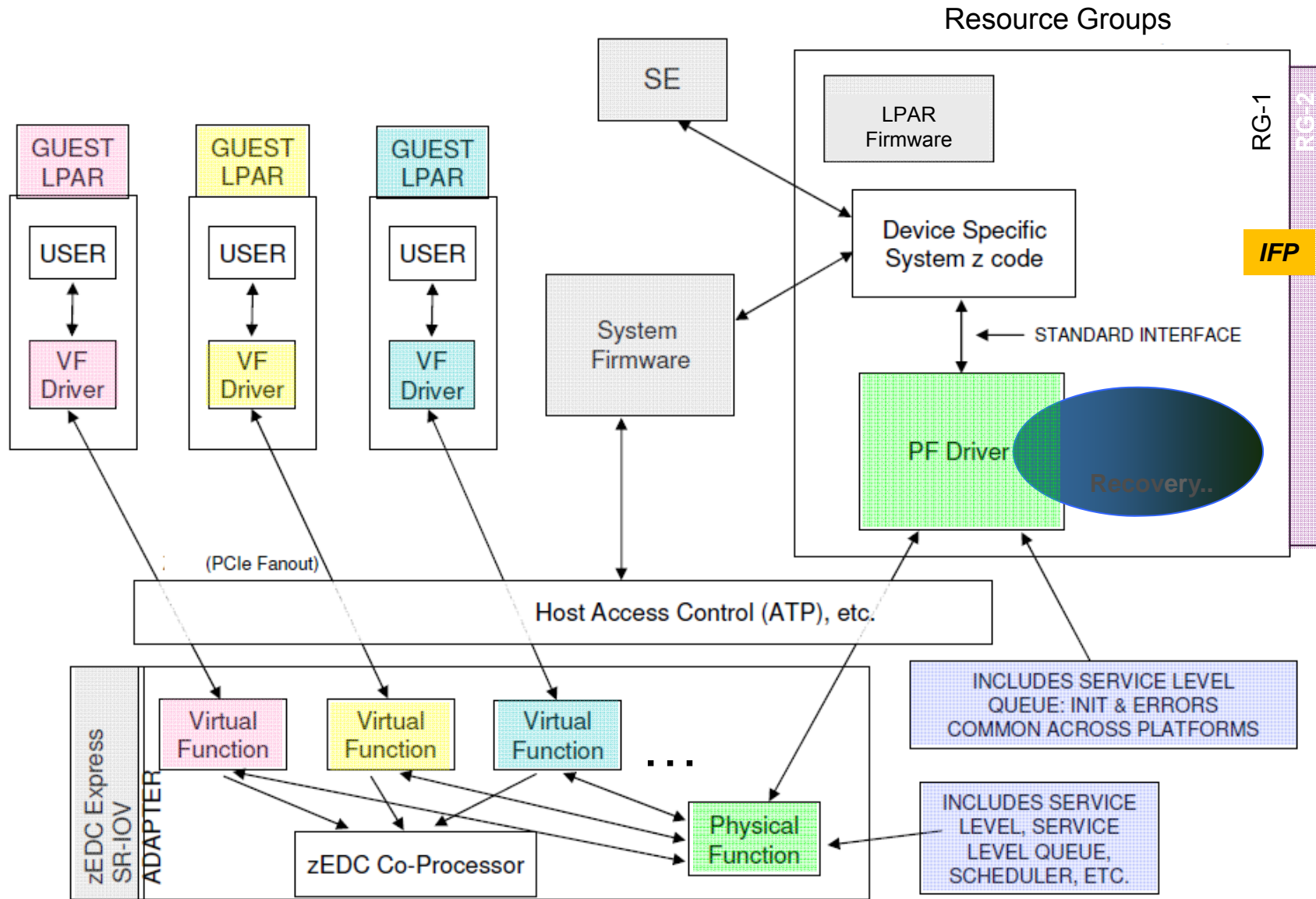


Native PCIe Feature: zEDC Express, 10GbE RoCE Express, and *Flash Express*


*PCIe Adapter Connector

(1) z System – ASIC Functions: Virtualization, Recovery, Diagnostics, Failover, Concurrent FW updates, etc.

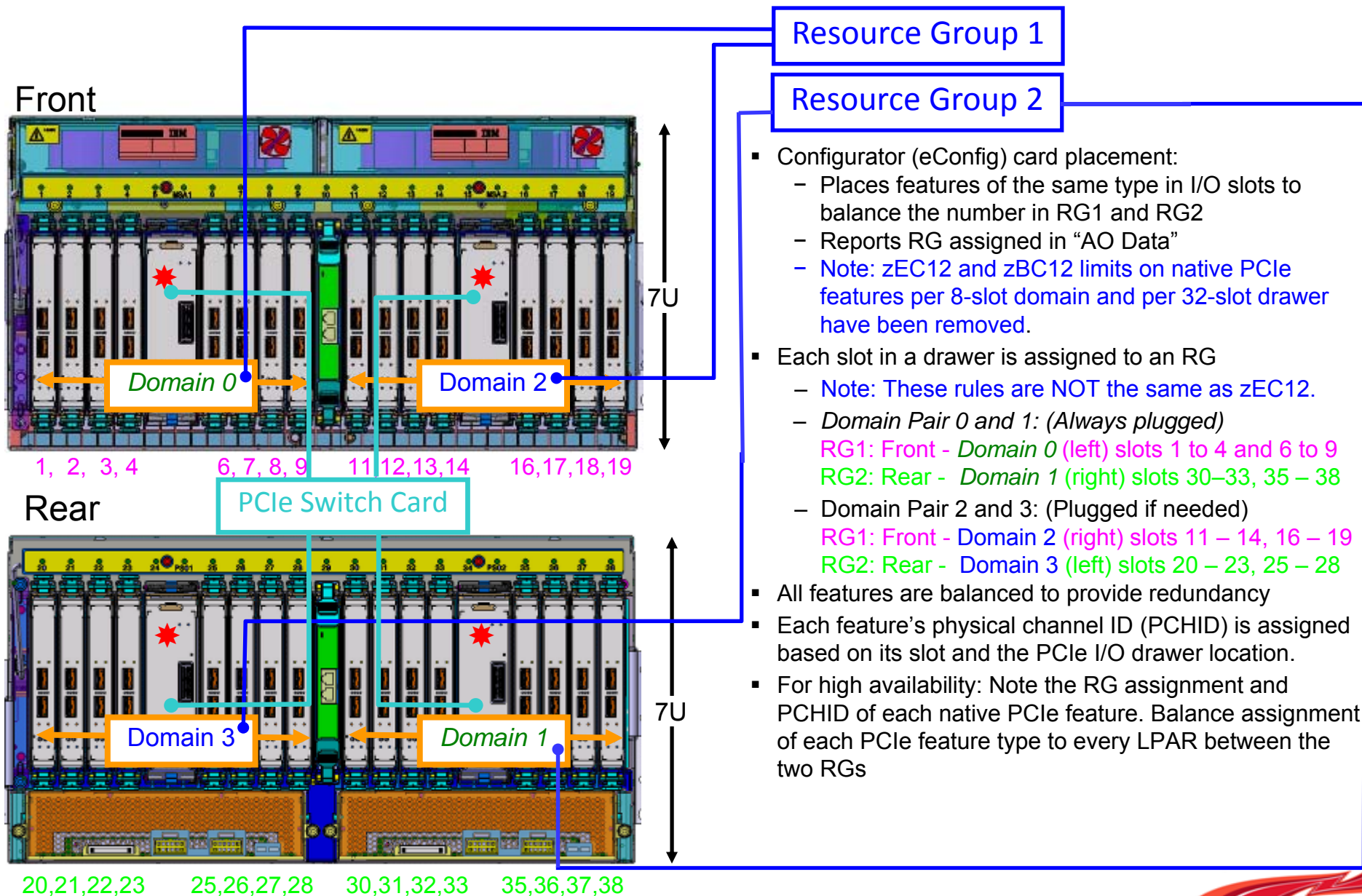
“Native” PCIe Infrastructure details



Native PCIe FUNCTION definition, assignment and mapping

- Conceptually similar to channel (**CHPID**) or I/O device definition with different rules
- **FUNCTION Definition in HCD or HCM to create IOCP input**
 - Uniquely identified by a hexadecimal **FUNCTION Identifier (FID)** in the range **000 – FFF** 
 - **NOT** assigned to a Channel Subsystem so ANY LPAR can be defined to a FUNCTION.
 - Has a **PARTITION** parameter that dedicates it to **ONE** LPAR or allows reconfiguration among a group of LPARs. (**A FUNCTION can NOT be defined as shared.**)
 - If the intended PCIe hardware supports multiple partitions, has a decimal **Virtual Function Identifier (VF=)** in the range 1 – n, where n is the maximum number of partitions the PCIe feature supports. Examples: A RoCE feature supports up to 31 partitions, a zEDC Express feature supports up to 15
 - May have other parameters specific to the PCIe feature. For Example, 10GbE RoCE Express requires a **Physical Network Identifier (PNETID=)**.
- **FUNCTION Mapping to hardware**
 - Assign a Physical Channel Identifier (**PCHID=**) to identify the hardware feature in a specific PCIe I/O drawer and slot to be used for the defined FUNCTION.
 - Methods:
 - Manually using the configurator (**eCONFIG**) “AO Data” report
 - With assistance using the CHPID Mapping tool with eConfig Configuration Report File (**CFR**) input
 - **Note: Unlike CHPIDs, multiple FUNCTIONS can be mapped to the SAME PCHID.** This is conceptually similar to mapping multiple InfiniBand coupling CHPIDs to the same adapter and port.

Native PCIe feature Plugging and Resource Groups (RGs)



eConfig “AO” Data – PCHIDs and Resource Groups Changed Rules for z13

Source	Drwr	Slot	F/C	PCHID/Ports or AID	Resource Group	Comment
A19/LG15/J01	Z22B	03	0420	108	RG1	Resource Group
A19/LG15/J01	Z22B	04	0420	10C	RG1	
A19/LG15/J01	Z22B	06	0411	110/D1D2	RG1	
A19/LG15/J01	Z22B	07	0411	114/D1D2	RG1	
A15/LG02/J01	Z22B	35	0411	170/D1D2	RG2	
A15/LG02/J01	Z22B	36	0411	174/D1D2	RG2	
A15/LG02/J01	Z22B	37	0420	178	RG2	
A15/LG02/J01	Z22B	38	0420	17C	RG2	

(*Note: F/C 0420 = zEDC Express, F/C 0411 = 10GbE RoCE Express)

eConfig places (plugs) hardware into slots to maximize serviceability and availability.

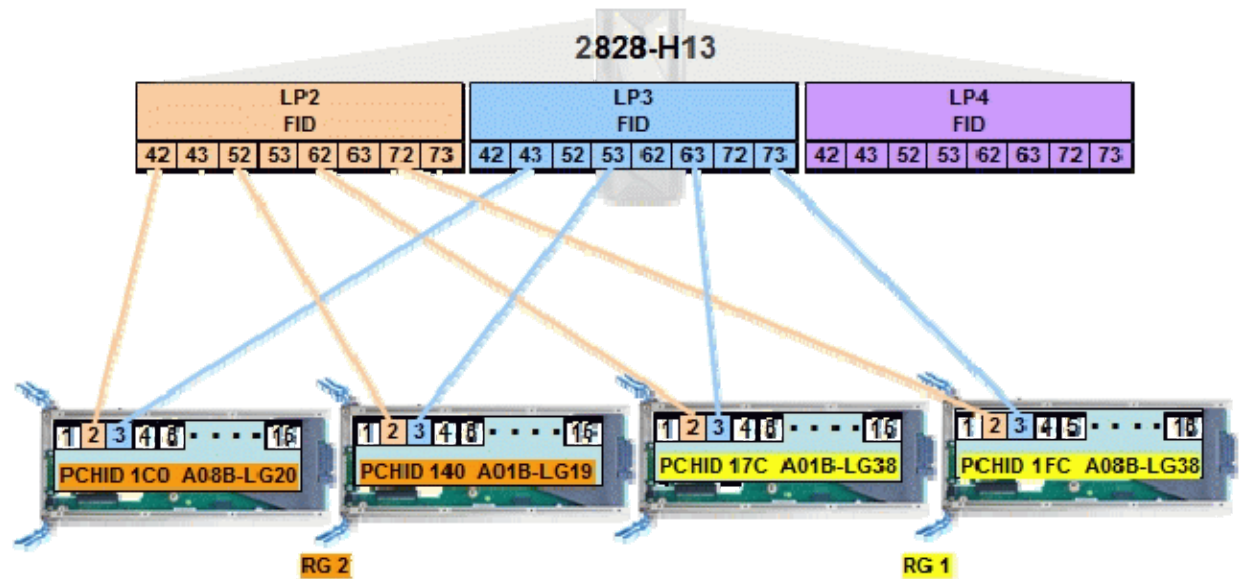
New for Native PCIe: Balanced plugging between two Resource Groups (RG1 and RG2)

Objective – Ensure that an LPAR has access to hardware in both Resource Groups for best availability.

Action - Define FUNCTIONS and assign to LPARs in HCD/IOCP. Map assigned FUNCTIONS to Physical Channel IDs (PCHIDs) in both RGs using the CHPID Mapping Tool or manually in HCD/IOCP.

Sample IOCP FUNCTION statements AFTER Mapping to PCHIDs*

- A **FUNCTION** is identified by a FUNCTION Identifier (**FID**): hexadecimal 000 – FFF
Suggestion: Use ODD FIDs for RG1 hardware, EVEN FIDs for RG2 hardware (z13 doesn't care, you may)
- TYPE** is new for z13 and is required (Human factors)
- FUNCTIONs can be **dedicated** or reconfigurable, not shared – One LPAR in the Access List
- To “share” a feature among LPARs, define a FUNCTION for each LPAR. These functions must have different Function IDs (FIDs) and different **Virtual Function ID (VF): decimal 1 - n, the maximum number LPARs the feature supports**
- Physical Network Identifier (PNETID)** is positional to identify the network names: (Port D1 - top, Port D2 – bottom)
- Note that LP14 and LP15 have access to each type of hardware in both Resource Groups and, on RoCE, to both networks

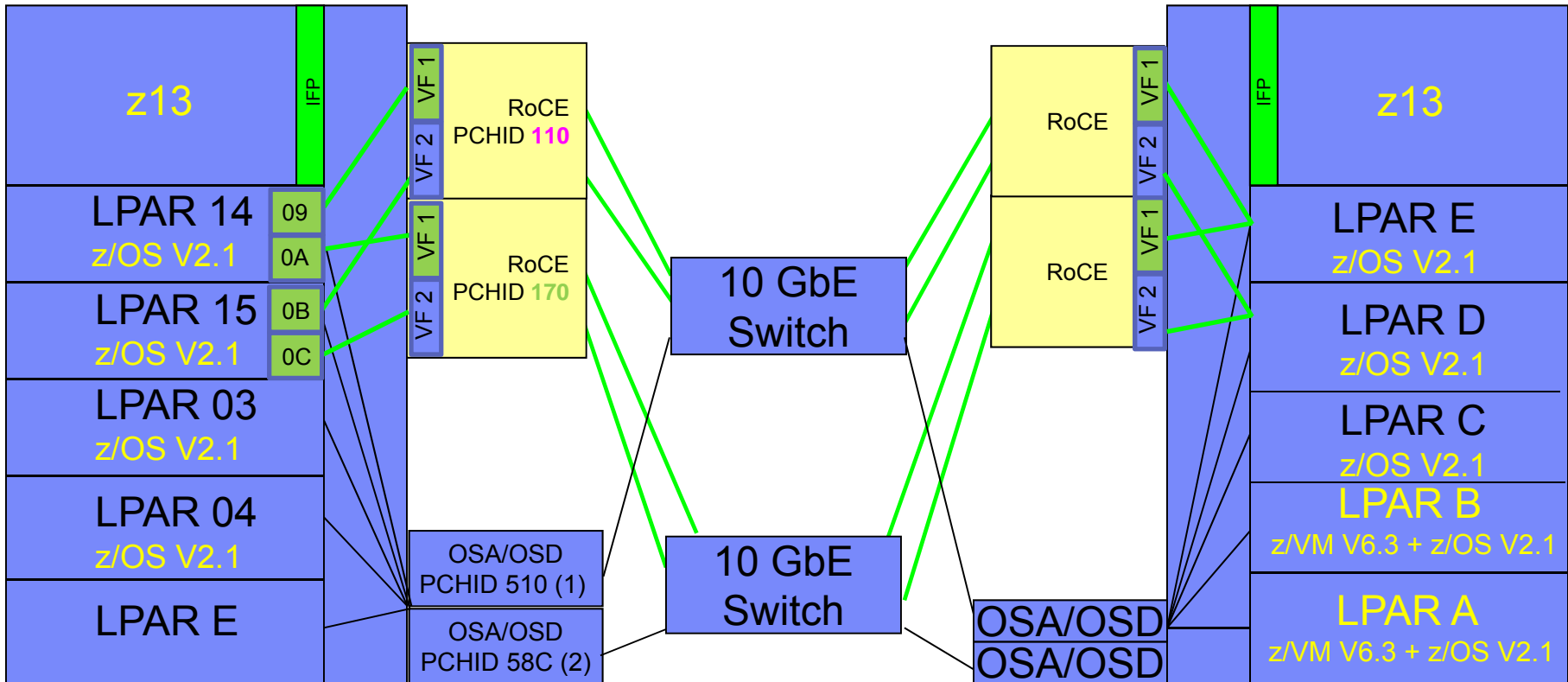


```

ID          MSG1='zEDC',MSG2='SYS1.IODF18 - 2015-01-15 11:30',          *
            SYSTEM=(2828,1),LSYSTEM=SCZP501,                          *
            TOK=('SCZP501',00800001B8D72827113028690113227F00000000,  *
            00000000,'13-08-15','11:30:28','SYS1','IODF18') RESOURCE   *
RESOURCE    PARTITION=((CSS(0),(LP1,1),(LP2,2),(LP3,3),(LP4,4),(LP5,5),(*,6), *
            (*,7),(*,8),(*,9),(*,A),(*,B),(*,C),(*,D),(*,E),(*,F)),(CSS(1),(*,1), *
            (*,2),(*,3),(*,4),(*,5),(*,6),(**,7),(*,8),(*,9),(*,A),(*,B),(*,C),(*,D),(*,E),(*,F)))
FUNCTION    FID=042,VF=2,PCHID=1C0,PART=((LP2),(LP3,LP4))
FUNCTION    FID=043,VF=3,PCHID=1C0,PART=((LP3),(LP2,LP4))
FUNCTION    FID=052,VF=2,PCHID=140,PART=((LP2),(LP3,LP4))
FUNCTION    FID=053,VF=3,PCHID=140,PART=((LP3),(LP2,LP4))
FUNCTION    FID=062,VF=2,PCHID=17C,PART=((LP2),(LP3,LP4))
FUNCTION    FID=063,VF=3,PCHID=17C,PART=((LP3),(LP2,LP4))
FUNCTION    FID=072,VF=2,PCHID=1FC,PART=((LP2),(LP3,LP4))
FUNCTION    FID=073,VF=3,PCHID=1FC,PART=((LP3),(LP2,LP4))
    
```

* IOCP statements in this example are for zBC12 zEDC adapters

Sample IOCP FUNCTION statements – cont ..



10GbE RoCE Express Functions for LPAR LP14, Reconfigurable to LP03 or LP04

```
FUNCTION FID=9, VF=1, PART=((LP14), (LP03, LP04)), PNETID=(NET1, NET2), *
    TYPE=ROCE, PCHID=110
```

```
FUNCTION FID=A, VF=1, PART=((LP14), (LP03, LP04)), PNETID=(NET1, NET2), *
    TYPE=ROCE, PCHID=170
```

10GbE RoCE Express Functions for LPAR LP15, Reconfigurable to LP03 or LP04

```
FUNCTION FID=B, VF=2, PART=((LP15), (LP03, LP04)), PNETID=(NET1, NET2), *
    TYPE=ROCE, PCHID=110
```

```
FUNCTION FID=C, VF=2, PART=((LP15), (LP03, LP04)), PNETID=(NET1, NET2), *
    TYPE=ROCE, PCHID=170
```

```
CHPID PCHID=510, PATH=(CSS(0,1,2,3), D0), TYPE=OSD, SHARED, PNETID=NET1
CHPID PCHID=58C, PATH=(CSS(0,1,2,3), D1), TYPE=OSD, SHARED, PNETID=NET2
```

Physical Network Identifier

(PNETID) is positional to identify the network names: (Port D1 - top, Port D2 - bottom)

Note that LP14 and LP15 have access to each type of hardware in both Resource Groups and, on RoCE, to both networks

zEDC and RoCE - What changed with z13



Considerations about zEDC and RoCE z13 topics

- For completeness and technical reference, this presentation contains a full set of slides around zEDC and RoCE PCI Native Adapters
- Only the slides containing changes since zEC12 will be presented today

Nothing is Changing for zEDC for z13

- Almost none of the source pool for z13 currently has Flash Express
 - BSAM/QSAM/HSM use case requires Pervasive zEDC
- z/OS Clients Migrating to z13 can upgrade their entire zEC12/zBC12/z13 infrastructure with zEDC compression as part of the migration to z13
 - Gain File Transfer elapsed time (This is critical mission availability to many customers)
 - Cross platform and z/OS to z/OS
 - Gain Disk and Tape Savings
 - Media Savings (disks, tapes, VTS storage)
 - Reduce replication costs

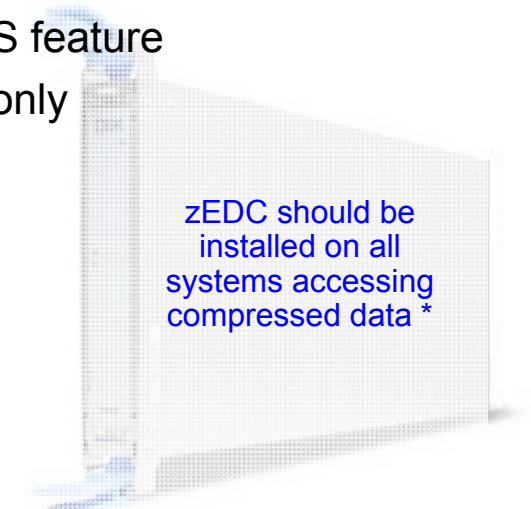
Deploying zEDC - Review

▪ Operating system requirements

- Requires z/OS 2.1 (with PTFs) and the zEDC Express for z/OS feature
- z/OS V1.13 and V1.12 offer software decompression support only
- z/VM V6.3 support for z/OS V2.1 guest:

▪ Server requirements

- Available on zEC12, zBC12 and z13
- zEDC Express feature for PCIe I/O drawer (FC#0420)
 - Each feature can be shared across up to 15 LPARs
 - Up to 8 features available on zEC12/zBC12/z13
- Recommended high availability configuration per server is four features
 - This provides up to 4 GB/s of compression/decompression
 - Provides high availability during concurrent update (half devices unavailable during update)
 - Recommended minimum configuration per server is two features
- Steps for installing zEDC Express in an existing zEC12/zBC12/z13
 - Apply z/OS Service; Hot plug a zEDC Express adapter; update your IODF, and Dynamic Activate



* For the full zEDC benefit, zEDC should be active on ALL systems that might access or share compressed format data sets. This eliminates instances where software inflation would be used when zEDC is not available.

IBM Offers Multiple Compression Technologies for the Mainframe

Type	Optimized for:	Performance Overhead	Supported data	Frequency of access post compression
CMPSC compression on z System co-processor chip	Optimal for DB2 or select DFSMS files	On Chip, relatively little CPU overhead and less I/O, Fast	<ul style="list-style-type: none"> DB2 - Optimized for row-wise access to data is required DFSMS files – for VSAM and non-VSAM extended format data sets 	Often
Other software compression (zlib, or similar)	Most compression uses industry std today. Used by many file types	Higher CPU - software instructions executed. Note: if Java then eligible for zIIP	Any. De facto standard for almost any type of data	Often
Tape HW compression	Tape compression – optimized for use with large files, archival purposes	Performed by the tape subsystem	Any	Often – Rare (application dependent)
Archival / Backup	Archive data and data to backup/copy	CPU overhead, longer wall clock time	DFSMSHsm, DFSMSdss	Often – Rare (application dependent)
Real time compression	IBM NAS storage	No performance degradation	SVC	Designed for active primary data.
zEDC Express	Active, for cross platform data exchange. Enables compression of active and inactive data	Processing on zEDC Express- expect minimal CPU overhead, low latency	<ul style="list-style-type: none"> SMF though logger zlib compatible Java BSAM/QSAM Extended format SOD DFSMSHsm /dss Encryption Facility 	Frequent access required. Useful for files that previously used software compression as well

ISVs exploit zEDC

zEDC was expressly created using industry standard APIs to encourage ISVs to leverage its high-speed compression value in applications ISVs create. With access to zEDC, ISV applications are more valuable to end users.

Contact these, or your, ISV providers directly for additional information for their future plans



- **Alebra – Parallel Data Mover (PDM)**

- Uses zEDC compression in lieu of Software-based compression to provide excellent qualities of service.



- **ASE – OMCS**

- Takes SLIKZIP and SLIKSFTP performance to a whole new level



- **PKWARE – PKZIP and SecureZIP v15**

- Accelerated deflate compression and automatic detection of zEDC



- **Software AG – Entire Net-Work**

- High performance transaction processing

Shared Memory Communication – RDMA utilizing the 10GbE RoCE Express



Shared Memory Communications

– Remote Direct Memory Access (SMC-R) Definition

- Shared Memory Communications – Remote Direct Memory Access (SMC-R) is a new communication protocol aimed at providing transparent acceleration for sockets-based TCP/IP applications and middleware
 - Remote Direct Memory Access (RDMA) technology provides low latency, high bandwidth, high throughput, low processor utilization attachment between hosts
 - SMC-R utilizes RDMA over Converged Ethernet (RoCE) as the physical transport layer
- SMC-R is built on the following concepts:
 - RDMA enablement of the communications fabric
 - Partitioning a set of OS host real memory into buffers and using RDMA technology to access this memory
 - Establishing an ‘out of band’ connection over which data is passed to the partner peer using RDMA writes and signaling

Single Root I/O Virtualization (SR-IOV) Implementation

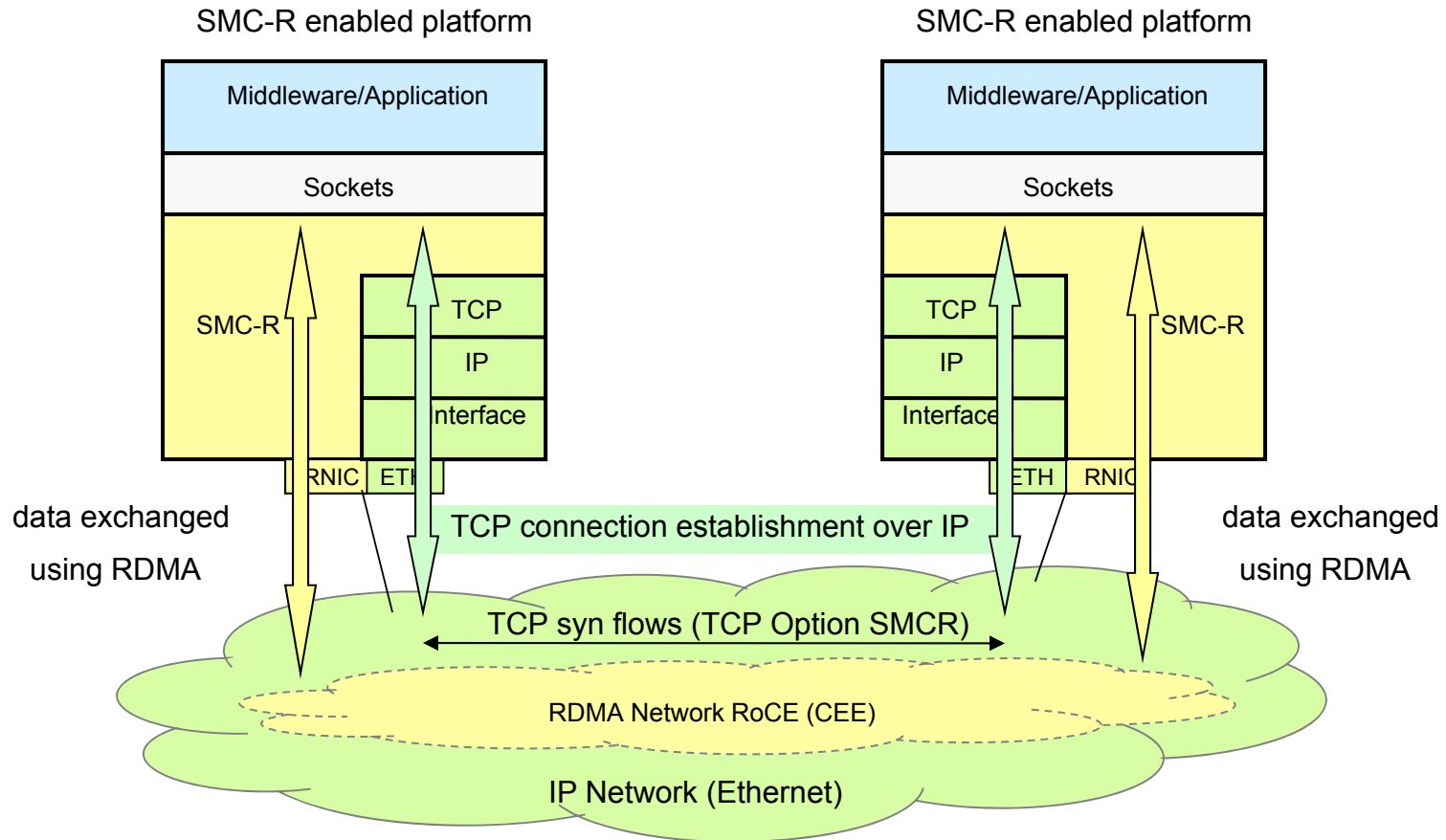
- SR-IOV is designed to provide the capability for use by multiple LPARs
 - Maximum number of LPARs (Virtual Functions) supported by zEDC Express is 15
- SR-IOV is designed to provide isolation of virtual functions within the PCIe adapter
 - One LPAR cannot cause errors visible to other virtual functions or LPARs
- Each OS has its own application queue in its own memory space
- Physical Function (PF) Driver
 - Communicates to the Physical Function in the PCIe adapter
 - Relatively limited function: Manages resource allocation, hardware error handling, code updates, recovery and diagnostics
- Device-specific z System LIC
 - Connects PF Driver to Support Element (SE) and limited system level firmware required services

z13 Shared RoCE (SR-IOV) Overview



- 10GbE RoCE Express feature becomes sharable among multiple z/OS instances (LPARs or z/VM guest virtual machines)
- Multiple RoCE PFIDs (PCIe Function IDs) with unique Virtual Function IDs are configured for each physical adapter (PCHID) in HCD (IOCDS)
- Up to 31 PFIDs supported per physical adapter
- Each z/OS instance (LP or z/VM guest) sharing the adapter consumes a unique (at least one) PFID (each PFID has a corresponding Virtual Function ID / number)
Note. An OS instance and VFs are generally (but not always) one to one
- Up to 16 physical adapters per CPC (no change)
- Adapter virtualization is transparent to upper layers (stack (transport layer) and application software)
- Dedicated RoCE (zEC12 and zBC12) interoperates with (transparent to) SR-IOV RoCE (z13)... shared RoCE environment is transparent to peer hosts

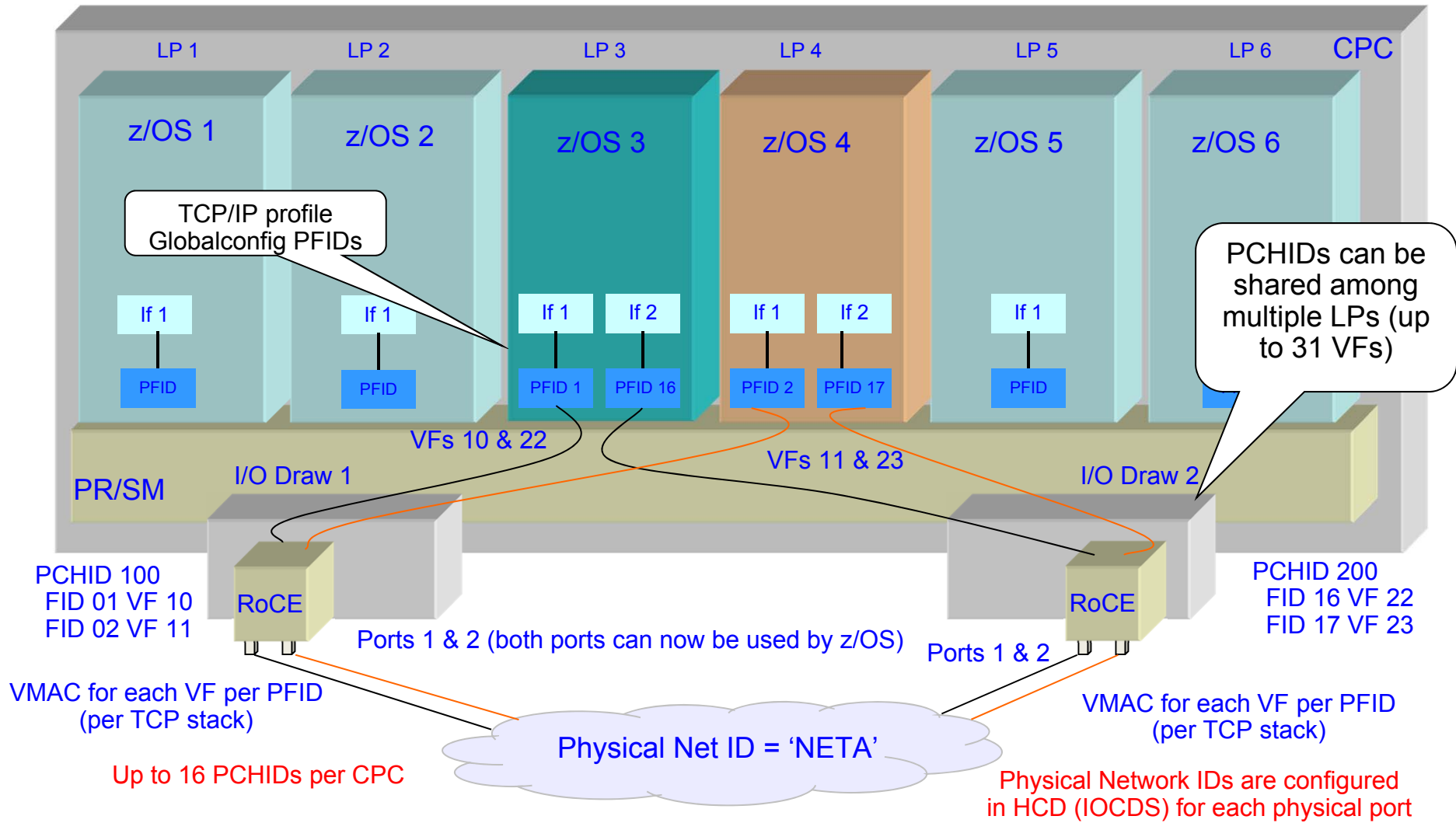
Dynamic Transition from TCP to SMC-R - Review



Dynamic (in-line) negotiation for SMC-R is initiated by presence of TCP Option (SMCR)

TCP connection transitions to SMC-R allowing application data to be exchanged using RDMA

z/OS Shared RoCE System View



IOCP Statements for a 'Shareable' RoCE Express Adapter

- The following is a sample IOCP configuration for defining a RoCE Express adapter shared between LPARs
FUNCTION FID= *3-digit hex value*, PCHID = *3-digit hex value naming the physical slot location* , PART = ((*initial access*), (*candidate list*)), VF = *2-digit hex value*, TYPE = ROCE, PNETID = (*port0, port1,,*)

PNETID array identifies the network the ports are associated with, so all FIDs on a RoCE adapter i.e. PCHID, must have the same PNETID for each port

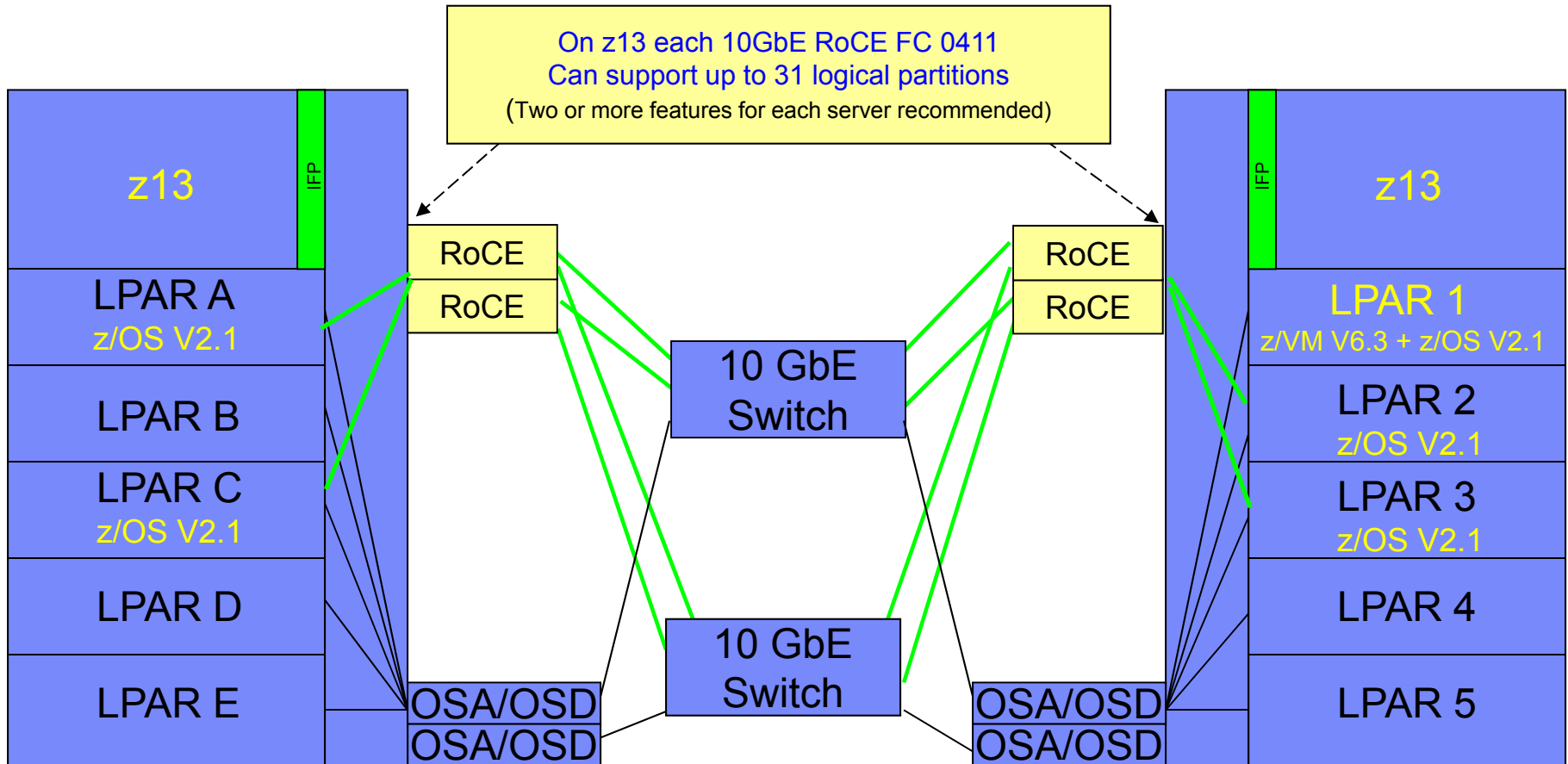
10GbE RoCE Express Functions for LPAR 08, Reconfigurable to LP09 with access to two networks

```
FUNCTION FID=05, PCHID=100, PART=((LP08),(LP09)), VF=1, TYPE=ROCE, PNETID=(NET1,,,)
FUNCTION FID=06, PCHID=12C, PART=((LP08),(LP09)), VF=1, TYPE=ROCE, PNETID=(,NET2,,)
```

10GbE RoCE Express Functions for LPAR 12, Reconfigurable to LP06 with access to two networks

```
FUNCTION FID=07, PCHID=100, PART=((LP12),(LP06)), VF=2, TYPE=ROCE, PNETID=(NET1,,,)
FUNCTION FID=08, PCHID=12C, PART=((LP12),(LP06)), VF=2, TYPE=ROCE, PNETID=(,NET2,,)
```

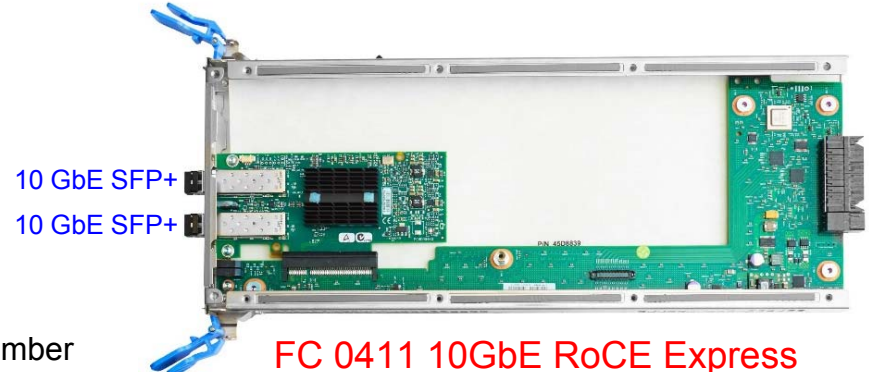

z13: 10GbE RoCE Express Sample Configuration



- This configuration allows redundant SMC-R connectivity among LPAR A, LPAR C, LPAR 2, and LPAR 3
- Both 10 GbE
- LPAR to LPAR OSD connections are required to establish the SMC-R communications
 - 1 GbE OSD connections can be used instead of 10 GbE
 - OSD connections can flow through the same 10 GbE switches or different switches
 - z13 exclusive: Simultaneous use of both 10 GbE ports on 10 GbE RoCE Express features

z13 - 10GbE RoCE Express Feature

- Designed to support high performance system interconnect
 - Shared Memory Communication (SMC) over Remote Direct Memory Access (RDMA) (SMC-R) Architecture exploits RDMA over Converged Ethernet (CE) - RoCE
 - Shares memory between peers
 - Read/write access to the same memory buffers without application changes
 - Designed to increase transaction rates greatly with low latency and reduced CPU cost
- Configuration
 - *z13 - Both 10 GbE SFP+ ports enabled*
 - *z13 - Support for up to 31 Logical Partitions*
 - A switched connection requires an enterprise-class 10 GbE switch with SR Optics, *Global Pause* enabled & *Priority Flow Control (PFC)* disabled
 - Point-to-point connection is supported
 - *Either connection supported to z13, zEC12 and zBC12*
 - Not defined as a CHPID and does not consume a CHPID number
 - Up to 16 features supported on a zBC12/zEC12
 - Link distance up to 300 meters over OM3 50 micron multimode fiber
- Exploitation and Compatibility
 - z/OS V2.1
 - *IBM SDK for z/OS Java Technology Edition, Version 7.1 (February 24, 2014)*
 - *z/VM V6.3 support for z/OS V2.1 guest exploitation (June 27, 2014)*
 - Linux on z System – IBM is working with Linux distribution partners to include support in future releases*



*Note: All statements regarding IBM's plans, directions, and intent are subject to change or withdrawal without notice. Any reliance on these Statements of General Direction is at the relying party's sole risk and will not create liability or obligation for IBM.

SMC-R and RoCE performance at distance

▪ Initial statement of support for SMC-R and RoCE Express

- 300 meters maximum distance from RoCE Express port to 10GbE switch port using OM3 fiber cable
 - 600 meters maximum when sharing the same switch across 2 RoCE Express features
 - Distance can be extended across multiple cascaded switches
 - All initial performance benchmarks focused on short distances (i.e. same site)

▪ Updated testing for RoCE and SMC-R over long distances

- IBM z System™ Qualified Wavelength Division Multiplexer (WDM) products for Multi-site Sysplex and GDPS® solutions qualification testing updated to include RoCE and SMC-R. Two vendors already certified their DWDM solution for SMC-R and RoCE Express:

1. Fibernet DUSAC 4800 Release 2.2b - on two client cards, the FTX-n and the FTX-10C (both cards are single port transponders). The qualification letter for this release can be found at the following link:

– <https://www.ibm.com/servers/resourcelink/lib03020.nsf/pages/FibernetSL?OpenDocument&pathID=>

2. Cisco 15454 Release 9.6.0.5 - on the 10 x 10G client card (15454-M-10x10G-LC) in 5:5 transponder mode. The qualification letter for this release can be found at the following link:

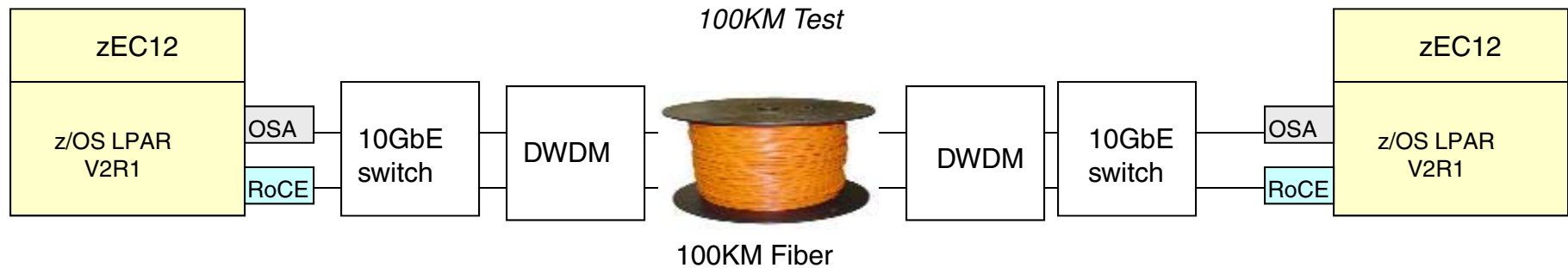
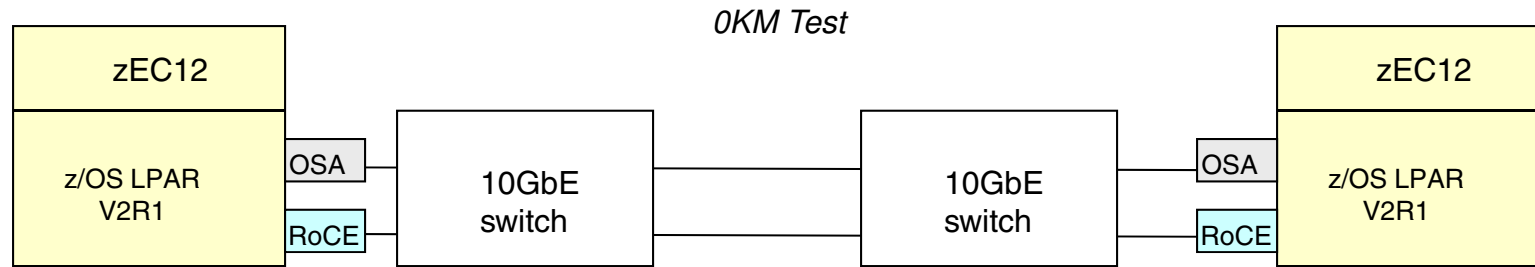
– <https://www.ibm.com/servers/resourcelink/lib03020.nsf/pages/ciscoSystemsInc?OpenDocument&pathID=>

▪ But how does SMC-R and RoCE perform at distance?

Internal IBM testing with SMC-R and RoCE at distance

- **Performance benchmarks performed using the IBM Application Workload Modeler (AWM) tool**
- **Micro-benchmarks: Tests included AWM in client and server mode on separate z/OS LPARs generating TCP socket traffic**
 - No business logic in AWM (simply sends/receives data)
 - Does exercise full TCP/IP API and protocol stack layers
- **Environment: 2 z/OS LPARs on zEC12 with 2 dedicated CPUs each with following connectivity**
 - OSA Express 5S (For TCP/IP benchmarks)
 - RoCE Express (For SMC-R benchmarks)
- **NOTE:** Based on internal IBM benchmarks using a modeled socket workload in a controlled laboratory environment using micro benchmarks. Your results may vary based on your configuration, workloads and environment.

Internal IBM testing with SMC-R and RoCE at distance - Configurations



Summary of performance benchmarks of SMC-R at distance

- **Micro-benchmarks performed at 10km (native ethernet) and 100km (with DWDM) distances**
 - At 10km
 - Request/Response workloads (1K/1K payloads): up to 47% lower latency and up to 88% higher throughput than TCP/IP
 - Request/Response workloads (32K/32K payloads): up to 60% lower latency and up to 150% higher throughput than TCP/IP
 - Streaming workloads (20M in one direction): Up to 60% improvement in latency and up to 150% throughput improvement vs TCP/IP
 - At 100km
 - Request/Response workloads (1K/1K payloads): up to 9% lower latency and up to 9% higher throughput than TCP/IP
 - Request/Response workloads (32K/32K payloads): up to 25% lower latency and up to 35% higher throughput than TCP/IP
 - Streaming workloads (20M in one direction): Over 80% improvement in latency and 394% throughput improvement vs TCP/IP (single connection)
 - CPU benefits of SMC-R for larger payloads consistent across all distances
- **NOTE:** Based on internal IBM benchmarks using a modeled socket workload in a controlled laboratory environment using micro benchmarks. Your results may vary based on your configuration, workloads and environment.

Summary of performance benchmarks of SMC-R at distance (cont)

■ Performance summary

- Technology viable even at 100km distances with DWDM
- At 10km: Retain significant latency reduction and increased throughput
- At 100km: Large savings in latency and significant throughput benefits for larger payloads, modest savings in latency for smaller payloads
- CPU benefits of SMC-R for larger payloads consistent across all distances

■ Use cases for SMC-R at distance

- TCP Workloads deployed on Parallel Sysplex spanning sites
- Software based replication (i.e. TCP based) across sites (Disaster Recovery)
 - e.g. InfoSphere Data Replication suite for z/OS
- File transfers across z/OS systems in different site
 - FTP, Connect:Direct, SFTP, etc.
- Opportunity: Lower CPU cost for sending/receiving data while boosting throughput and lowering latency



■ Questions ?

– **Ewerson Palacio**

bird@br.ibm.com

– **Frank Packheiser**

F.Packheiser@de.ibm.com

– **Parwez Hamid**

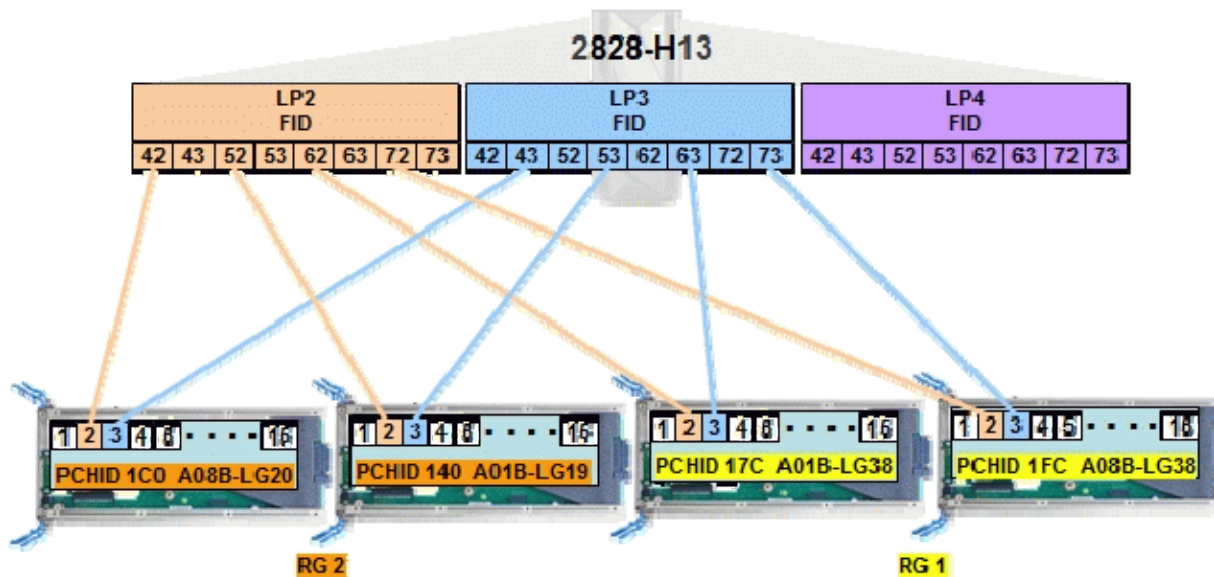
pnh@us.ibm.com

IBM z13

Reinventing enterprise IT
for digital business

zEDC – Implementation Example

- PCIe Function ID (PFID)**
 - A 3 hex digit value defined in HCD
 - Initially, a PCIe function can be assigned to one LPAR only, but can be reassigned to another LP using candidate list
 - PCIe functions behave similarly as reconfigurable CHPIDs
 - PFIDs can be genned prior to H/W installed (for 2827 w/GA2 support or 2828)
- Virtual Function ID (VF ID)**
 - 1-15 value which identifies the LPAR
- Sample Function IOCP statements for zBC12 in the diagram shown**



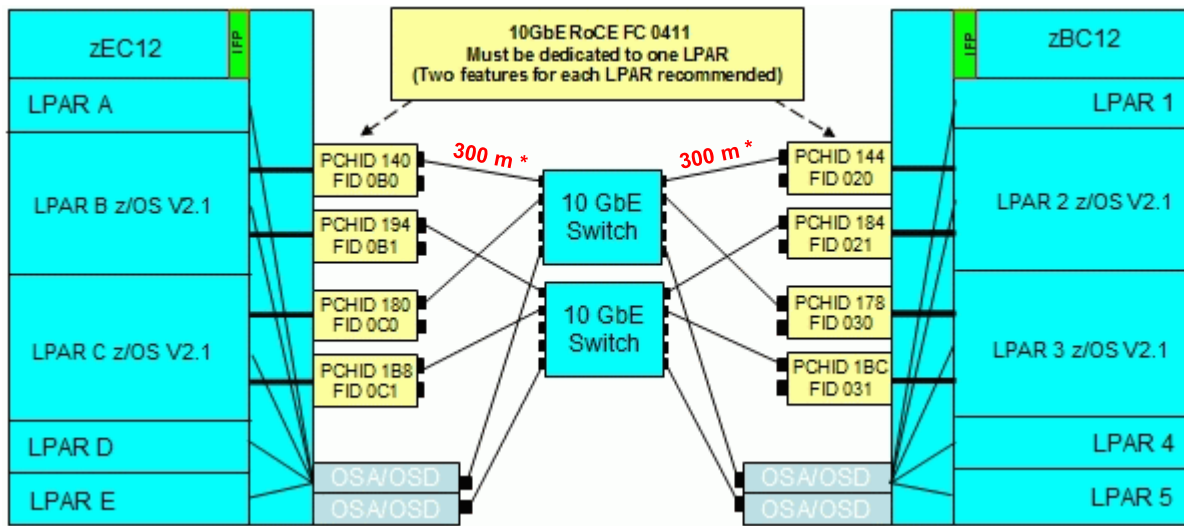
```

ID          MSG1='zEDC',MSG2='SYS1.IODF18 - 2013-08-15 11:30',
            SYSTEM=(2828,1),LSYSTEM=SCZP501,
            TOK=('SCZP501',00800001B8D72827113028690113227F00000000,
            00000000,'13-08-15','11:30:28','SYS1','IODF18') RESOURCE
RESOURCE    PARTITION=((CSS(0),(LP1,1),(LP2,2),(LP3,3),(LP4,4),(LP5,5),(*,6),
            (*,7),(*,8),(*,9),(*,A),(*,B),(*,C),(*,D),(*,E),(*,F)),(CSS(1),(*,1),
            (*,2),(*,3),(*,4),(*,5),(*,6),(**,7),(*,8),(*,9),(*,A),(*,B),(*,C),(*,D),(*,E),(*,F)))
FUNCTION    FID=042,VF=2,PCHID=1C0,PART=((LP2),(LP3,LP4))
FUNCTION    FID=043,VF=3,PCHID=1C0,PART=((LP3),(LP2,LP4))
FUNCTION    FID=052,VF=2,PCHID=140,PART=((LP2),(LP3,LP4))
FUNCTION    FID=053,VF=3,PCHID=140,PART=((LP3),(LP2,LP4))
FUNCTION    FID=062,VF=2,PCHID=17C,PART=((LP2),(LP3,LP4))
FUNCTION    FID=063,VF=3,PCHID=17C,PART=((LP3),(LP2,LP4))
FUNCTION    FID=072,VF=2,PCHID=1FC,PART=((LP2),(LP3,LP4))
FUNCTION    FID=073,VF=3,PCHID=1FC,PART=((LP3),(LP2,LP4))
    
```

RoCE – Implementation Example

■ PCIe Function ID (PFID)

- A 3 hex digit value defined in HCD and on the GLOBALCONFIG statement in the TCP/IP Profile to uniquely identify an RDMA network interface card (RNIC) adapter
- Initially, a PCIe function can be assigned to one LPAR only but can be reassigned to another LP using candidate list
- PCIe functions behave similarly as reconfigurable CHPIDs
- PFIDs can be genned prior to H/W installed (for 2827 w/GA2 support or 2828)



■ Physical Network ID (PNet ID)

- A value that uniquely identify your physical layer 2 LAN fabric. This value is used to logically associate the z System adapters and ports to be physically connected to your network
- PNet ID is only defined in HCD
- Operating Systems dynamically learn and use this definition
- If a PNet ID is not configured for the RNIC adapter, activation will fail

■ Sample Function IOCP statements for zBC12 in the diagram shown

```

ID          MSG1='ROCE',MSG2='SYS1.IODF18 - 2013-08-15 11:30',
           SYSTEM=(2828,1),LSYSTEM=SCZP501,
           TOK=('SCZP501',00800001B8D72827113028690113227F00000000,
           00000000,'13-08-15','11:30:28','SYS1','IODF18')
RESOURCE   PARTITION=((CSS(0),(LP1,1),(LP2,2),(LP3,3),(LP4,4),(LP5,5),
           (*,6),(*,7),(*,8),(*,9),(*,A),(*,B),(*,C),(*,D),(*,E),(*,F)),
           (CSS(1),(*,1),(*,2),(*,3),(*,4),(*,5),(*,6),(*,7),(*,8),(*,9),
           (*,A),(*,B),(*,C),(*,D),(*,E),(*,F)))
FUNCTION   FID=020,PCHID=144,PNETID=NETA,PART=((LP2),(LP3))
FUNCTION   FID=021,PCHID=184,PNETID=NETA,PART=((LP2),(LP3))
FUNCTION   FID=030,PCHID=178,PNETID=NETB,PART=((LP3),(LP2))
FUNCTION   FID=031,PCHID=1BC,PNETID=NETB,PART=((LP3),(LP2))....

CHPID     PCHID=510,PATH=(CSS(0,1,2,3),D0),TYPE=OSD,SHARED,PNETID=NETA
CHPID     PCHID=58C,PATH=(CSS(0,1,2,3),D1),TYPE=OSD,SHARED,PNETID=NETB
    
```

300 m * - with OM3 fiber cable. The latency advantages of RDMA are diminished when travelling long distances, and so RDMA performs best when used within datacenter distances.(600 m total w/ a switch).