IBM z13 10GbE RoCE Express Virtualization

Introduction and Overview of Shared RoCE (SR-IOV)

(April 2015)

Jerry Stevens (sjerry@us.ibm.com)

Agenda

- Introduction:
  System z13 10GbE RoCE Express virtualization (“shared RoCE Express”)
- System z13 Requirements (firmware / software)
- Configuring shared RoCE
- Verification
- Feedback (enablement, verification, usage, etc.)
- Backup (RoCE Express and SMC-R Overview and background information)

IBM System z13 10GbE RoCE-Express feature virtualization Introduction

Description

- Background: The IBM zEC12 (zBC12) introduced the 10GbE RoCE Express feature. The RoCE Express provides an RDMA capable network adapter providing access to RDMA over Converged Ethernet (RoCE) networks that provides optimized network interconnect for System z communications.
- IBM z13 introduces the capability to share (virtualize) the 10GbE RoCE-Express feature among multiple (up to 31) LPARs (or z/VM guest virtual machines) using PCIe standardized virtualization (SR-IOV).

Customer Value

- When exploited by z/OS using Shared Memory Communications over RDMA (SMC-R) the combined solutions provide improved transaction rates for transactional workloads due to improved network latency and lower CPU cost for workloads with larger payloads (i.e. analytics, streaming, big data, or web services) while preserving the key qualities of services required by system z clusters in enterprise data center networks.
- When customers enable SMC-R they should immediately see the benefits, and longer term the benefits will be extended as they expand their exploitation of RDMA technology on System z.
- With the ability to now share RoCE Express features:
  - Access to RoCE is extended to additional LPARs (workloads) while requiring fewer physical RoCE features
  - Both 10GbE RoCE Express physical ports can now be concurrently exploited by z/OS

Primary Audience

- Customers who have multiple CPCs in a single site with z/OS centric workloads (e.g. SYSPLEX, DB2, WAS, CICS, MQ, IMS etc.). IBM has provided the SMC Applicability Tool (SMCAT) for assisting customers with evaluating SMC applicability for their specific application workloads and environment.

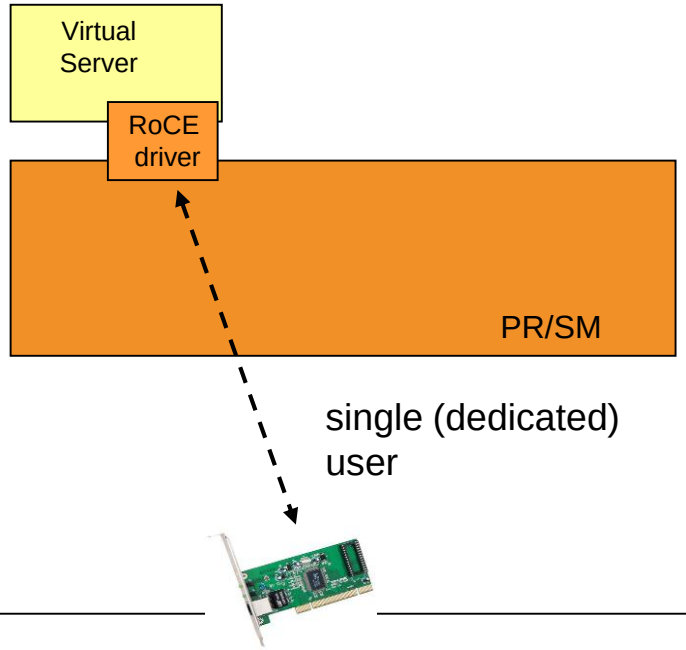
System zEC12 (zBC12) RoCE with SMC-R (existing support)

IBM zEC12: SMC-R requires a new RDMA capable NIC:

- 10GbE RoCE Express feature introduced in zEC12 GA2 and zBC12
- Support for up to 16 RoCE Express features per zCPC
- Cannot be shared across LPARS in initial deliverable (zEC12 GA2 and zBC12)
- Each LPAR requires 2 RoCE Express features for High Availability
- z/OS can only exploit a single port from each feature

10GbE RoCE Express

10GbE RoCE Express



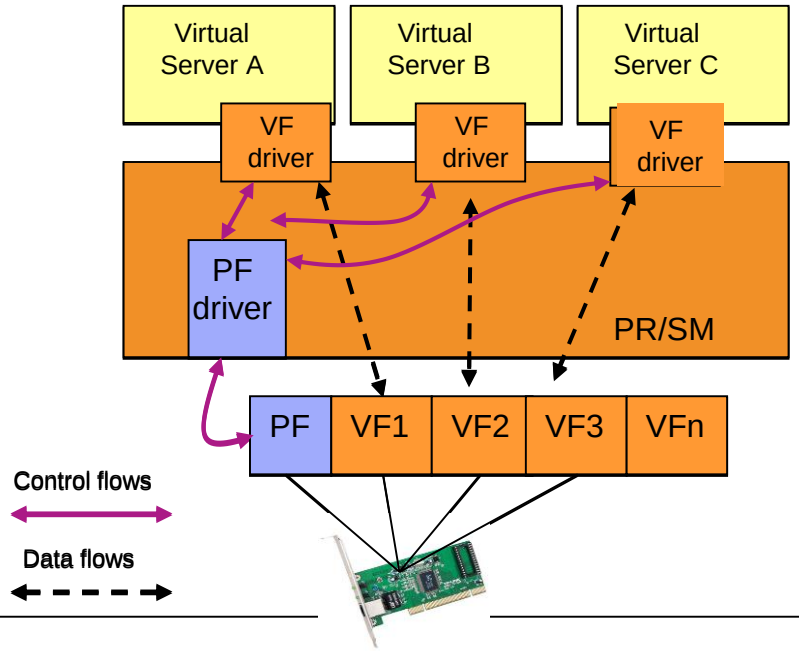
IBM z13 System: Shared RoCE with z/OS V2R1 SMC-R Support

IBM z13 System Shared RoCE support - *Available exclusively on z13:*

- Allows concurrent sharing of a RoCE Express feature by multiple virtual servers (OS instances)
- Up to 31 virtual servers (OS instances, LPARs or 2nd level guests under zVM) can share a single feature
- Each usage (OS instance) of the adapter consumes a PCIe Function ID (FID) and a Virtual Function (FIDs and VFs are configured in IOCDS)
- Support for up to 16 RoCE Express features per zCPC
- Efficient sharing (using PCIe standard SR-IOV) for an adapter (eliminates Hypervisor out of the data path)
- Enables concurrent use of both RoCE Express ports by z/OS (SMC-R)
- For High Availability each OS instance requires access to two unique (physical) features
- z/OS support will be available in z/OS V2R1 via PTF and z/OS V2R2 (base)

10GbE RoCE Express

Shared RoCE



System z13 Requirements for Shared RoCE

- System z13 firmware requirements:
  D22h Bundle 1

Software requirements:

- The following APARs / PTFs are required:
  1. VTAM: OA44576 with date of REWORK(2015029) or above
  2. TCP: PI12223 with rework date of REWORK(2015030) or above
  3. IOS: PTF UA90971 with rework date of REWORK(2015030) or above

- When testing and using RMF for RoCE: OA44524

- When testing z/OS with RoCE as a guest under z/VM: VM65577

Note. This is not a complete System z13 list. Refer to your ESP installation information.

Configuring shared RoCE Express (IO Configuration)

IO Configuration Requirements:

- **FID**:
  In IOCDS (or HCD) configure multiple FIDs (PCI Function IDs) for each physical adapter (PCHID) based on your anticipated needs for sharing each adapter:
  Example:
  Configure a unique FID for each physical feature for each z/OS TCP/IP stack that requires access to the feature (card/port), up to 31 FIDs per feature

  Note. Considerations for calculating the number of RoCE FIDs:
  When calculating the number of FIDs required for your environment there are two configuration aspects that can increase the number of required FIDs:
  - Multiple TCP/IP stacks: z/OS supports multiple stacks. If you run with more than one stack per z/OS instance you will need to configure a unique FID per stack.
  - Multiple RoCE ports: Each TCP/IP stack can access either physical port (1 or 2) or both ports. Each port will require a unique FID (i.e. unique FID per physical port).

- **VF**:
  Each unique FID also requires a corresponding VF number (ID)

- **PNETID**:
  Physical Network IDs (PNETIDs) must be configured (in IOCDS/HCD) for **both RoCE and OSA port**.

  Note 1. When exploiting z/OS SMC-R, RoCE and the related OSA adapters / ports must be associated. This association is created when adapters are connected to the same physical network (i.e. when adapters (ports) are configured with matching **Physical Network IDs**).

  Note 2. Port numbers are not configured in IOCDS/HCD. Port numbers are only configured in the OS (TCPIP configuration).

RoCE SR-IOV (virtualization) IOCDS Example

- HCD / IOCDS (configure VFs):

  sample RoCE IOCP / IOCDS declaration:

```
FUNCTION FID=100,PCHID=400,VF=1,PART=((LPA),(LPB,LPC)),PNETID=(NET1,NET2),TYPE=ROCE
FUNCTION FID=101,PCHID=400,VF=2,PART=((LPB),(LPA,LPC)),PNETID=(NET1,NET2),TYPE=ROCE
FUNCTION FID=102,PCHID=400,VF=3,PART=((LPC),(LPA,LPB)),PNETID=(NET1,NET2),TYPE=ROCE
```

Note. TCP/IP profile (coordination):

GLOBALCONFIG SMCR PFID (port num mtu) definition:
- Must match PFIDs defined in IOCDS
- Must specify the appropriate port (defaults to 1) and optional MTU (defaults to 1k)
- Should follow best practices for HA (provision 2 unique PCHIDs with unique PFIPs)

Configuring shared RoCE Express (TCPIP Configuration)

z/OS TCP/IP Configuration Requirements:

1. RoCE PCIe Function IDs (FIDs) must be configured in the TCPIP profile Globalconfig statement (SMCR parameter).

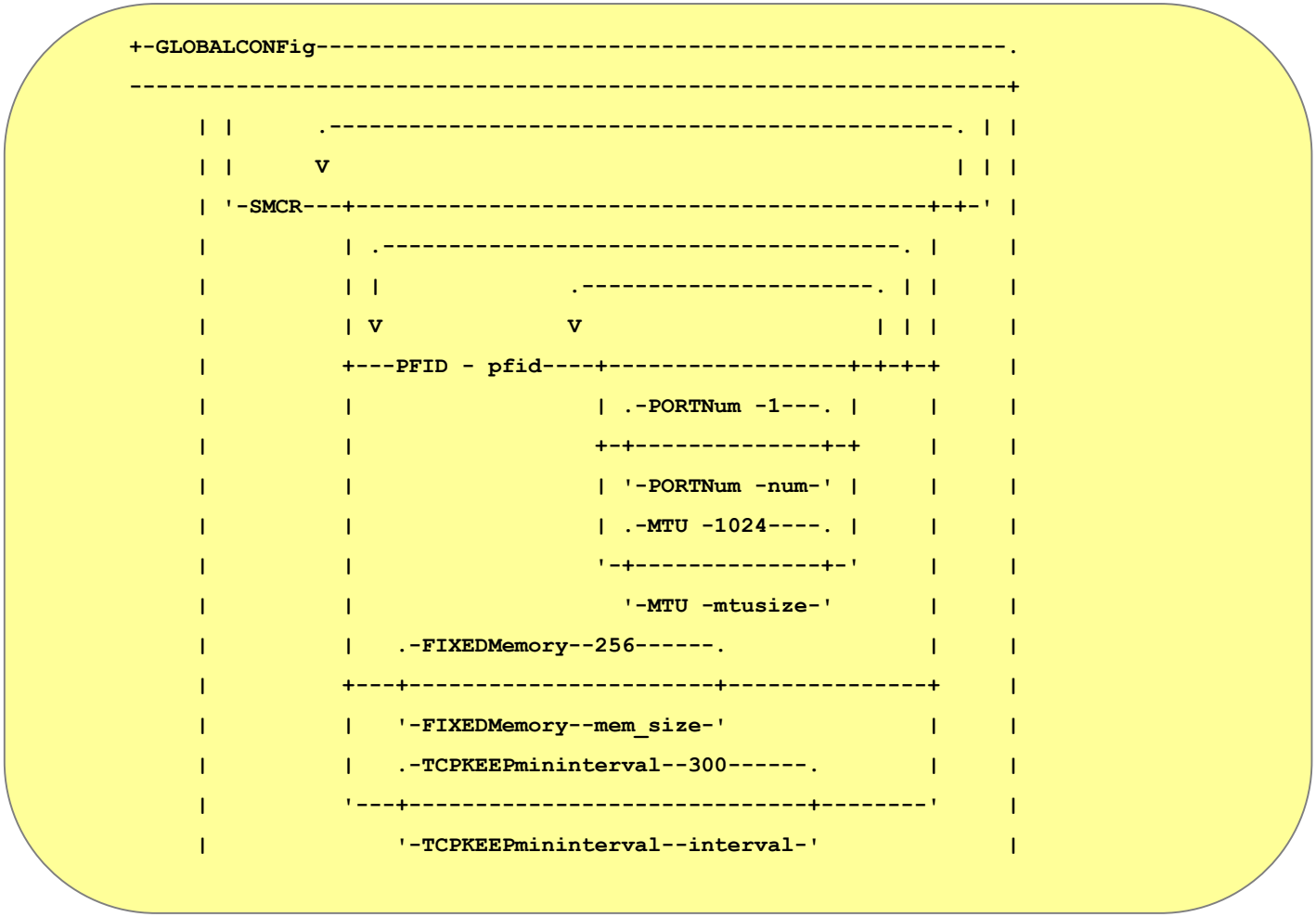
   The FID values must be coordinated with your IOCDS.

2. OSA Interface statement requirements:
   Generally there are no changes required to your OSA interface statements.
   Review the next 2 charts (GLOBALCONFIG and IPAQENET INTERFACE).

Note.
The TCPIP configuration for shared RoCE is very similar to the zEC12 RoCE TCPIP configuration. The only differences for Shared RoCE are:

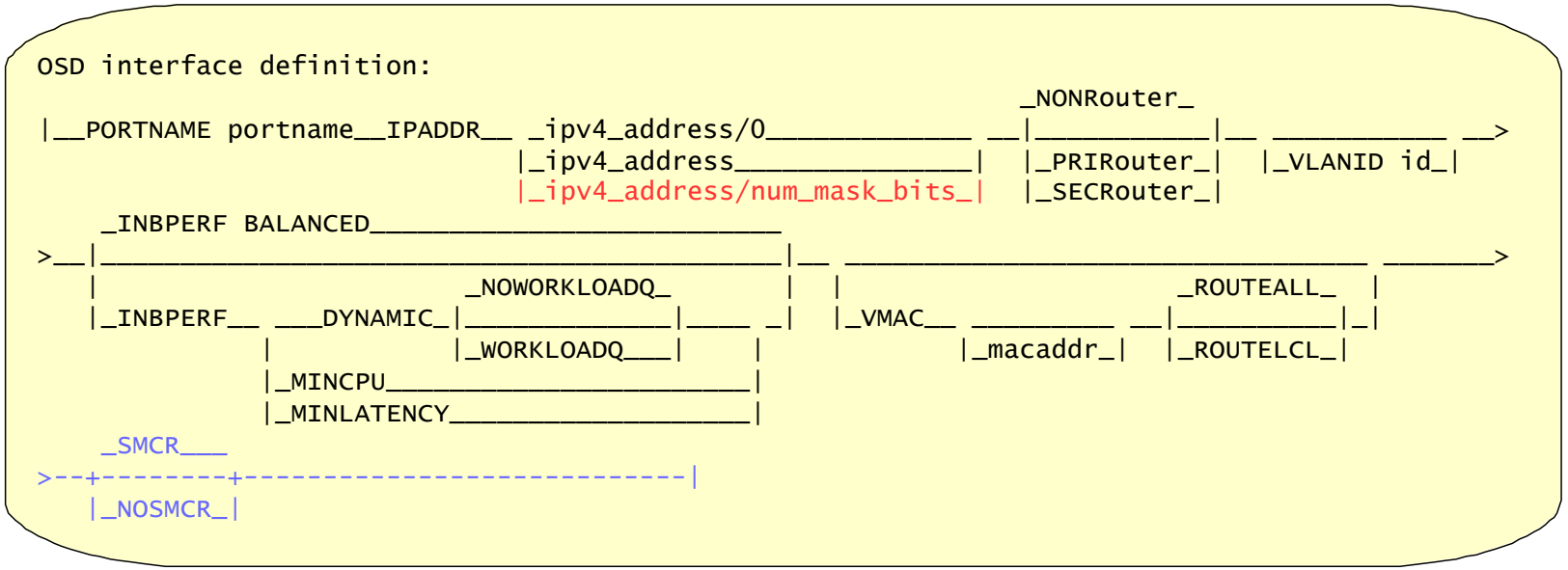
- When using multiple stacks within the same z/OS instance, each stack must configure unique FID values.
- Configuring multiple ports: each unique port requires configuring a unique FID / Portnumber

TCPIP GLOBALCONFIG Statement



Function externals: IPAQENET INTERFACE syntax

- IPAQENET INTERFACE statements
  - SMCR only valid for CHPIDTYPE OSD
  - SMCR cannot be used with IPv4 OSD interfaces defined using DEVICE and LINK statements
  - Must specify a non-zero subnet mask



Verification: z/OS D PCIE (RoCE VFs are defined)

z/OS PCIE display is updated to display the RoCE VF number

```
D PCIE
IQP022I 13.22.43 DISPLAY PCIE 056
PCIE 0010 ACTIVE
PFID   DEVICE TYPE NAME           STATUS ASID JOBNAME PCHID VFN
0011   10GbE RoCE Express         CNFG                0140  0001
0012   10GbE RoCE Express         CNFG                0140  0002
0061   10GbE RoCE Express         CNFG                0154  0001
0062   10GbE RoCE Express         CNFG                0154  0002
0071   10GbE RoCE Express         CNFG                0158  0001
0072   10GbE RoCE Express         CNFG                0158  0002
```

CommServer Device Verification (VF number / PFIP)

RoCE virtualization is dynamically detected (first RoCE activation) and is fundamentally transparent … minor change in CommServer product externals

```
D NET,TRL,TRLE=IUT10011
IST097I DISPLAY ACCEPTED
NAME = IUT10011, TYPE = TRLE 341
IST1954I TRL MAJOR NODE = ISTTRL
IST486I STATUS= ACTIV, DESIRED STATE= ACTIV
IST087I TYPE = *NA* , CONTROL = ROCE, HPDT = *NA*
IST2361I SMCR PFID = 0011 PCHID = 0140 PNETID = PNETID1
IST2362I PORTNUM = 1 RNIC CODE LEVEL = **N/A**
IST2389I PFIP = 01000300
IST2417I VF = 0001
IST924I -------------------------------------------
IST1717I ULPID = TCPIP2 ULP INTERFACE = EZARIUT10011
IST1724I I/O TRACE = OFF TRACE LENGTH = *NA*
IST314I END
```

The **PFIP** (PCIE Firmware Internal Path) is displayed for each FID (PCHID). The first byte (value 0 or 1) denotes the path. The PFIP can be used to validate HA and SP administrative actions.

When RoCE virtualization support is present VTAM TRLE display for RoCE PFIDs will include new MSG IST2417I displaying the VF number (ID) for this PFID

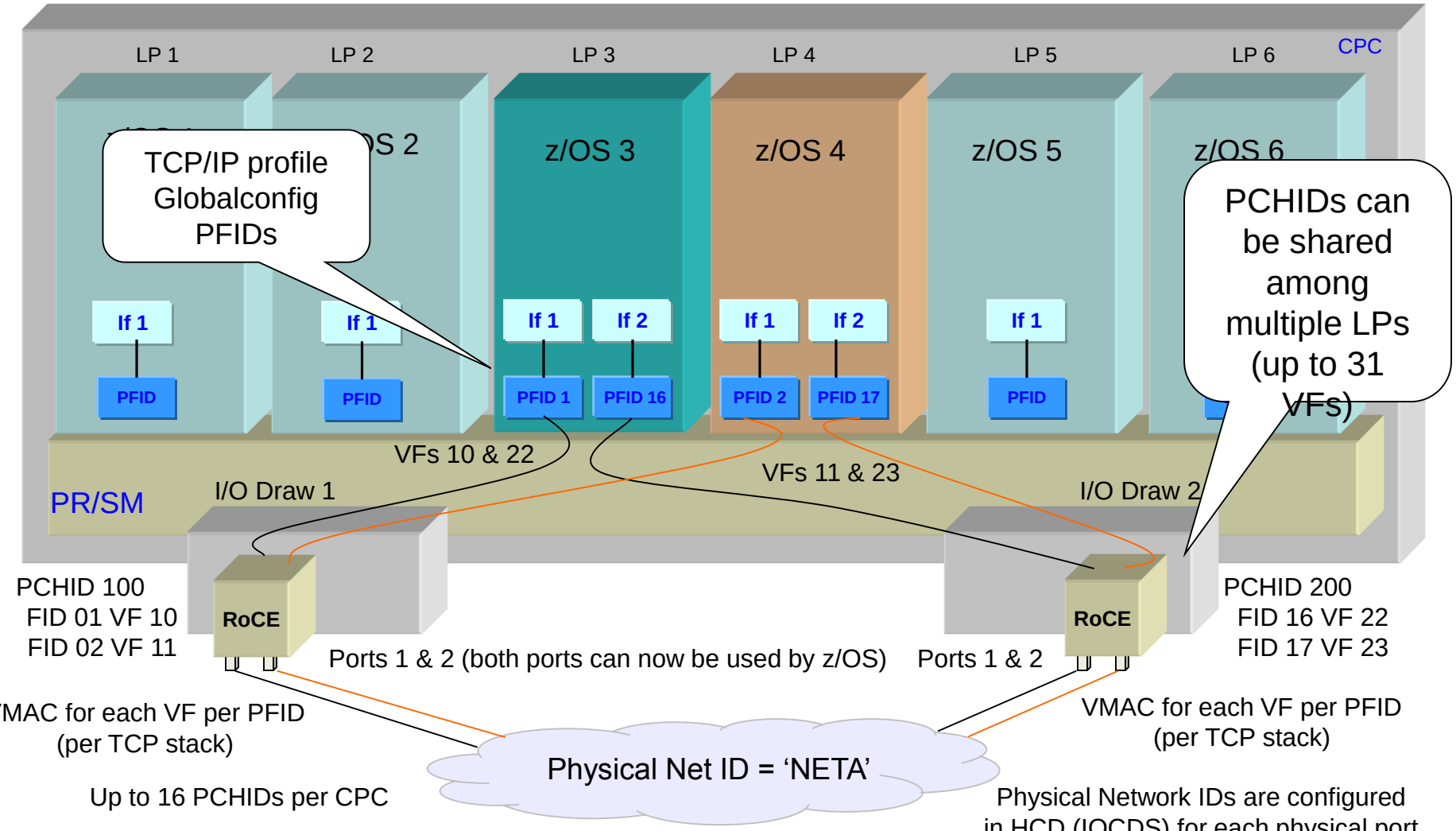
Function externals: Sample Netstat ALL/-A report

- This TCP connection uses SMC-R communications

```
D TCPIP,TCPCS1,NETSTAT,ALL,IPPORT=10.1.1.14+21
EZD0101I NETSTAT CS V2R1 TCPCS1
CLIENT NAME: FTPDOE34                  CLIENT ID: 0000003B
  LOCAL SOCKET: ::FFFF:10.1.1.14..21
  FOREIGN SOCKET: ::FFFF:10.1.1.24..1024
...
    SMC INFORMATION:
      SMCSTATUS:        ACTIVE          SMCGROUPID:         2D8F0100
      LOCALSMCLINKID:   2D8F0101        REMOTESMCLINKID:    729D0101

----
1 OF 1 RECORDS DISPLAYED
END OF THE REPORT
```

z/OS Shared RoCE (virtualization via SR-IOV) System View



Feedback

Your Feedback is very important to us:

- Installation (hardware, firmware or software)
- Configuration and Enablement (HCD/IOCDS, TCPIP, other)
- Verification and monitoring (status, usage, etc.)
- Performance observations
- Exploitation, RAS, etc.

Backup (Review and reference materials)

- Additional SMC-R reference information:

  http://www-01.ibm.com/software/network/commserver/SMCR/

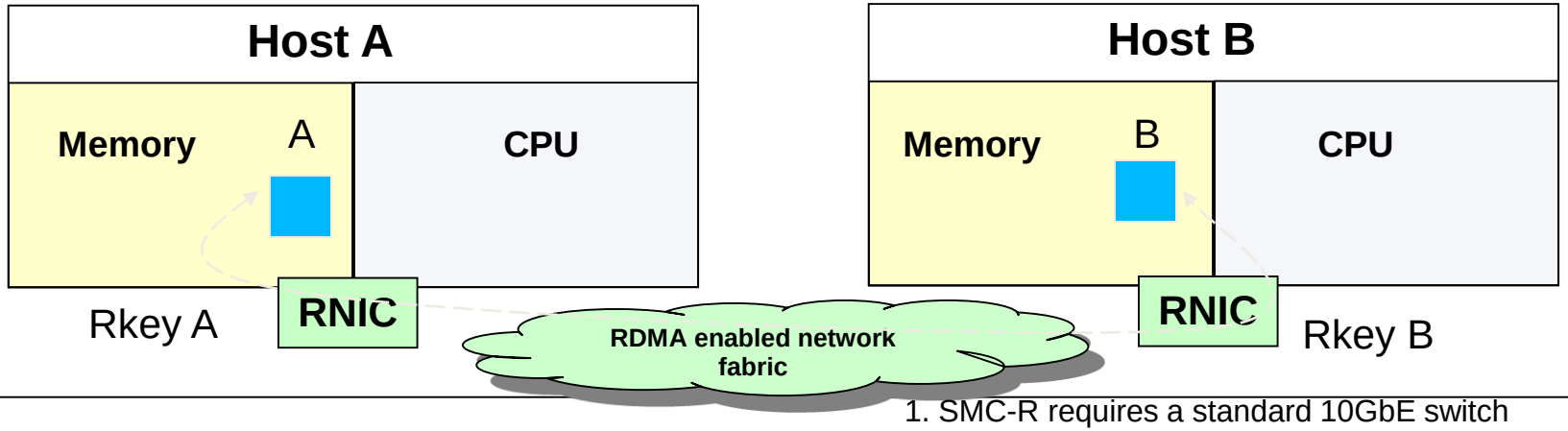
IBM 10GbE RoCE-Express feature Introduction

- Implemented as a PCIe device (PCIe Function ID or PFID)
- Provides an RDMA capable network interface card (RNIC) that provides access to RDMA over Converged Ethernet (RoCE) networks using standard 10GbE switches that provides optimized System z network interconnect for CPC to CPC communications.
- Each physical adapter (feature) can now be shared by multiple z/OS instances (LPARs or z/OS guest virtual machines when running z/VM).
- Best practices (for redundancy and high availability) requires access to 2 unique physical adapters per z/OS instance (LPAR redundancy).
- Maximum of 16 adapters per CPC
- z/OS Communications Server SMC-R support in V2R1
  - Transparent application middleware exploitation
  - Includes new SMCR support with traditional statistical data in NetStat and NMI.

RDMA (Remote Direct Memory Access) Technology Overview

Key attributes of RDMA

- Enables a host to read or write directly from/to a remote host's memory ***without*** involving the remote host's CPU
  - By registering specific memory for RDMA partner use
  - Interrupts ***still required*** for notification (i.e. CPU cycles are not completely eliminated)
- Reduced networking stack overhead by using streamlined, low level, RMDA interfaces
  - Low level APIs such as uDAPL, MPI or RDMA verbs allow optimized exploitation
    - > *For applications/middleware willing to exploit these interfaces*
- Key requirements:
  - A reliable "lossless" network fabric (LAN for layer 2 data center network distance)
  - An RDMA capable NIC (RNIC) and RDMA capable switched fabric (switches)[1]



1. SMC-R requires a standard 10GbE switch

RoCE - RDMA over Converged (Enhanced) Ethernet

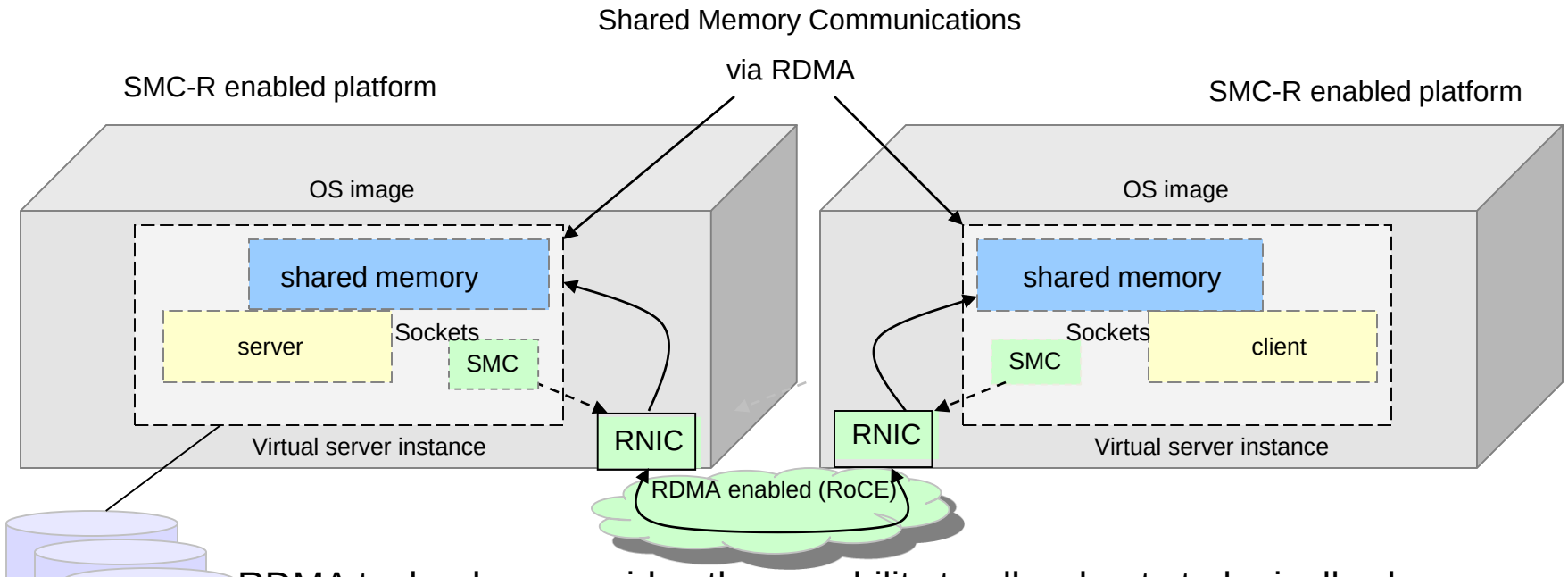
- RDMA based technology has been available in the industry for many years – primarily based on Infiniband (IB)
- IB requires a completely unique network eco system (unique hardware such as host adapters, switches, host application software, system management software/firmware, security controls, etc.) – IB is common in the HPC market
- RDMA technology is now available on Ethernet – RDMA over Converged Ethernet (RoCE)
  - RoCE uses existing Ethernet fabric (including standard Ethernet switches) but requires advanced RDMA capable NICs (RNICs or host adapters)
  - *Game changer: makes RDMA technology affordable for datacenter networks*
- Host software exploitation options fall into two general categories:
  - Native / direct application exploitation (many variations)
  - Transparent application exploitation (e.g. sockets based)

Shared Memory Communications over RDMA Introduction

- A new sockets based communications protocol is introduced on System z
- SMC-R (Shared Memory Communications over RDMA) provides the access and exploitation of the new IBM 10GbE RoCE-Express feature hardware on System z and to the RoCE network fabric.
- The following charts provide an introduction to the new technology…

"Shared Memory Communications over RDMA" concepts

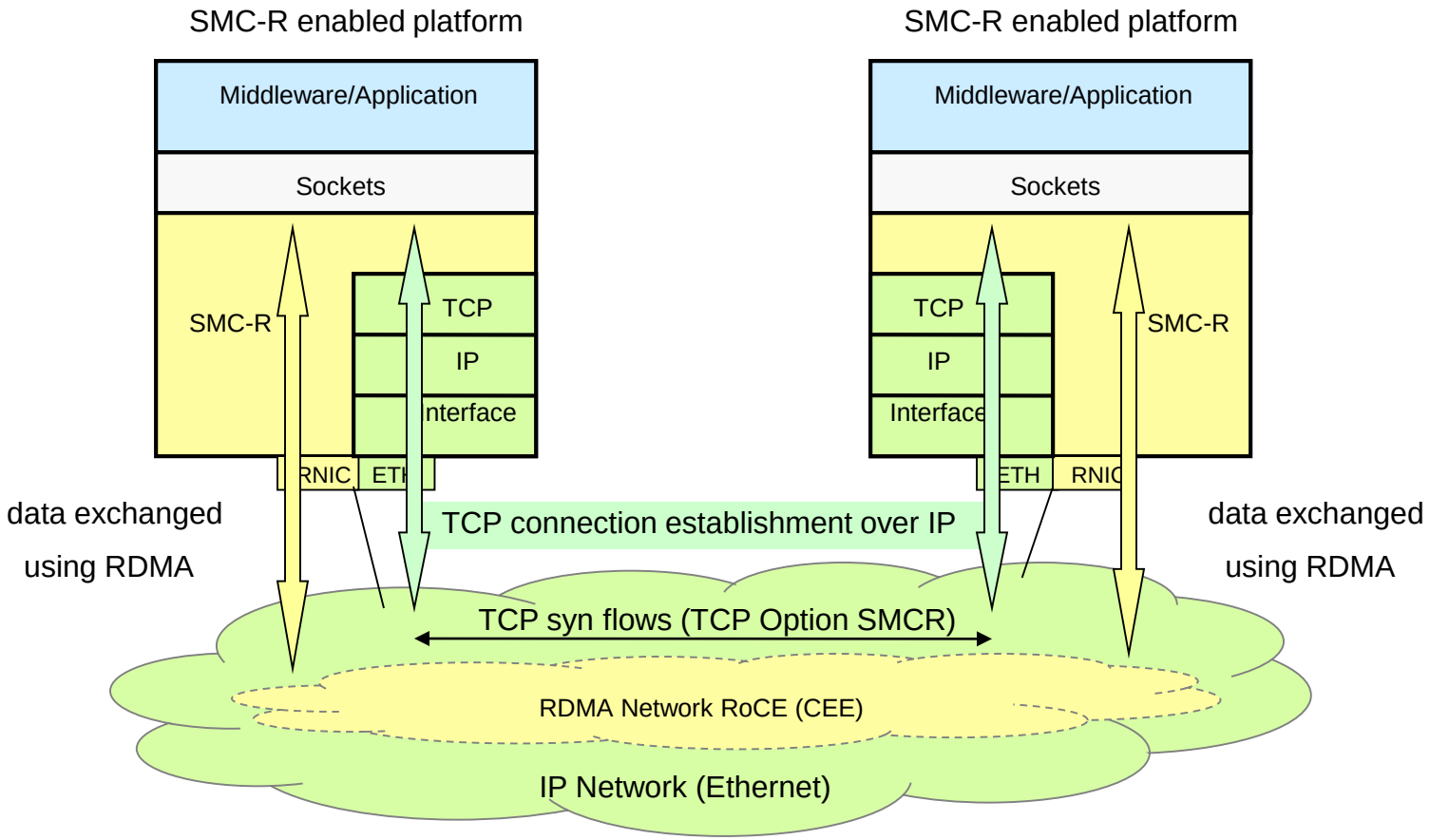
Clustered Systems



RDMA technology provides the capability to allow hosts to logically share memory. The SMC-R protocol defines a means to exploit the shared memory for communications - *transparent* to the applications! SMC-R RFC: http://tools.ietf.org/html/draft-fox-tcpm-shared-memory-rdma-01

This solution is referred to as *SMC-R* (Shared Memory Communications over RDMA). SMC-R represents a sockets over RDMA protocol that provides a foundation for a complete solution meeting all of the described objectives. SMC-R is an RDMA model exploiting RDMA-writes (only) for all data movement.

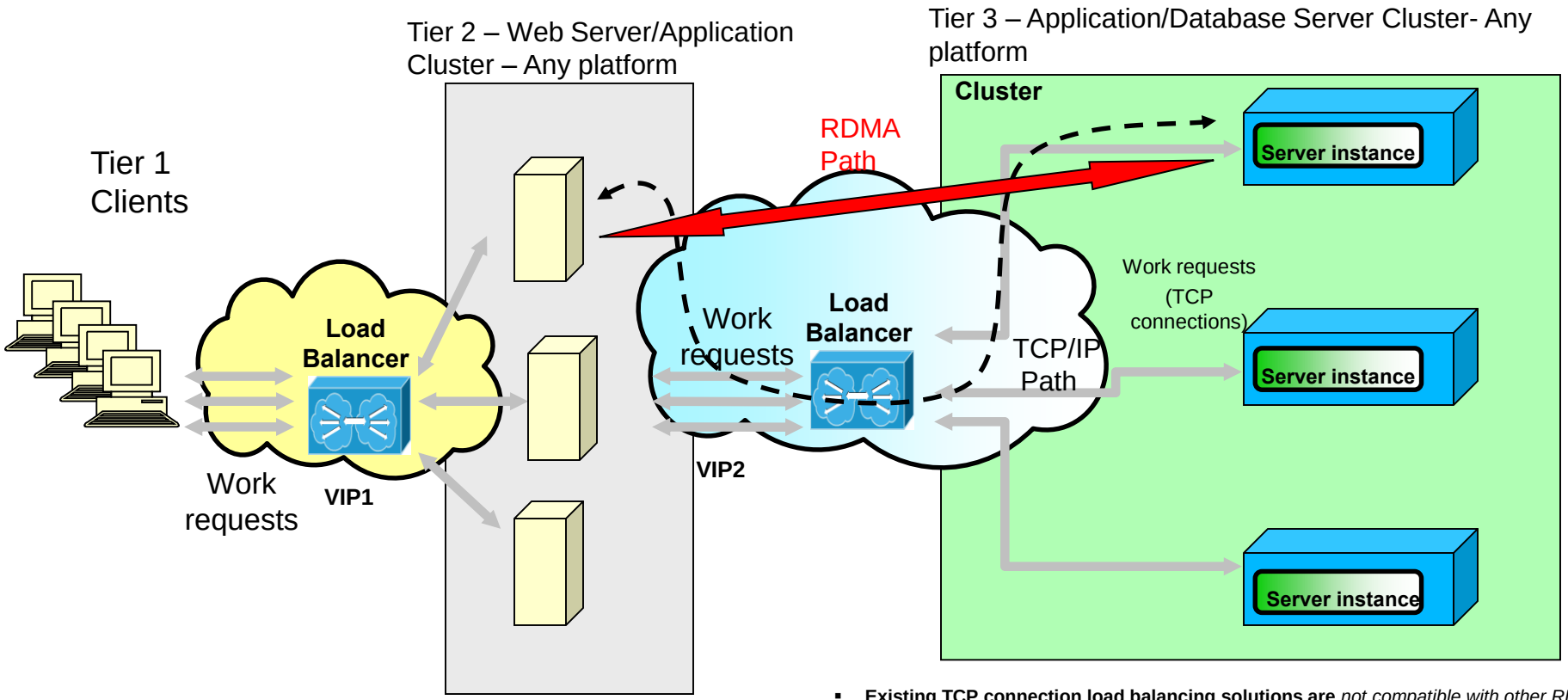
Dynamic Transition from TCP to SMC-R



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

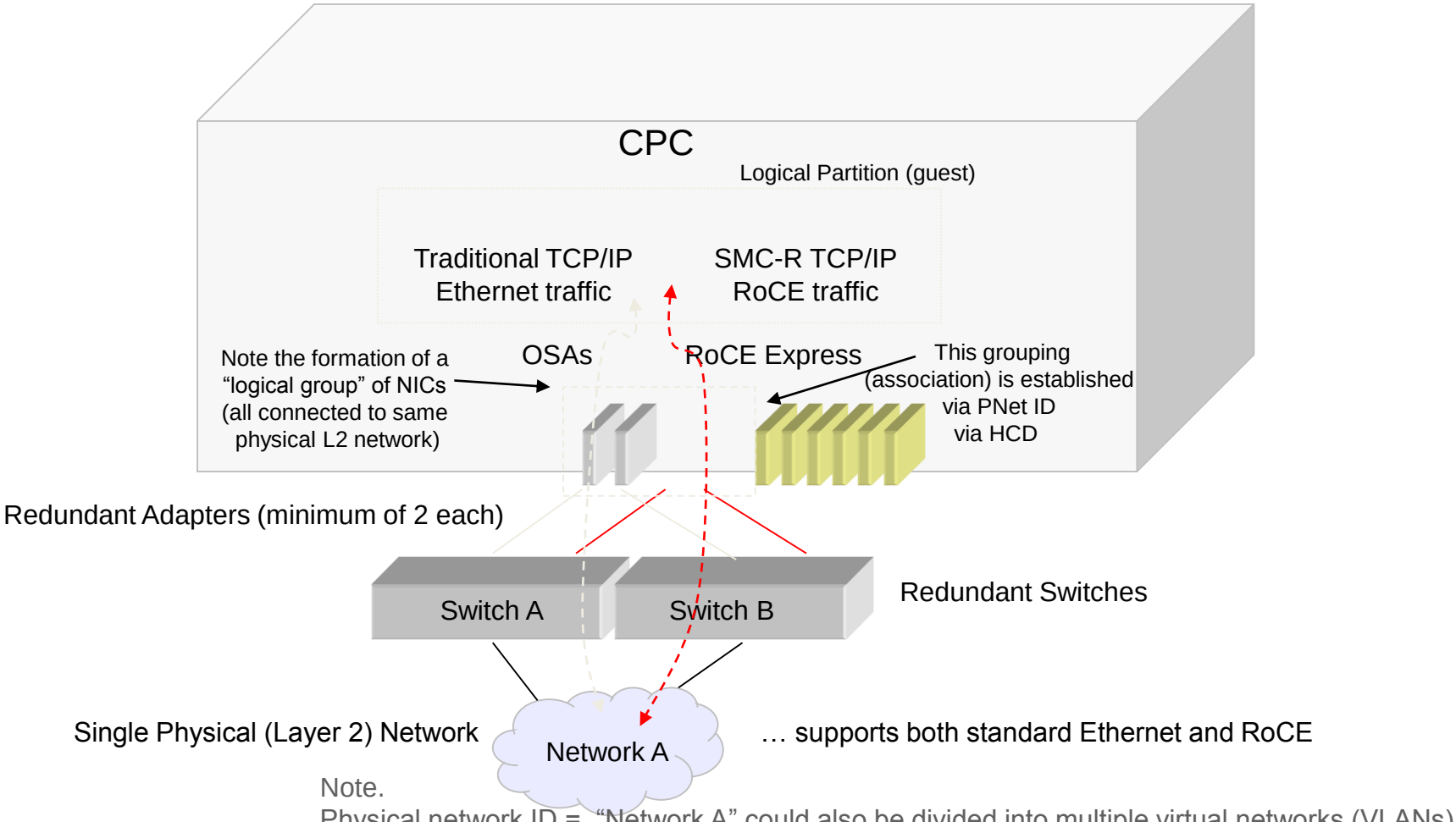
Server clustering and TCP connection load balancing



- **Server clustering is a prevalent deployment pattern for Enterprise class servers**
  - Provide High Available, eliminate single points of failure, ability to grow/shrink capacity dynamically, ability to perform non-disruptive planned maintenance, etc.
- **TCP connection load balancing is a key solution for load balancing within a cluster environment**
  - External or Internal load balancers provide this capability

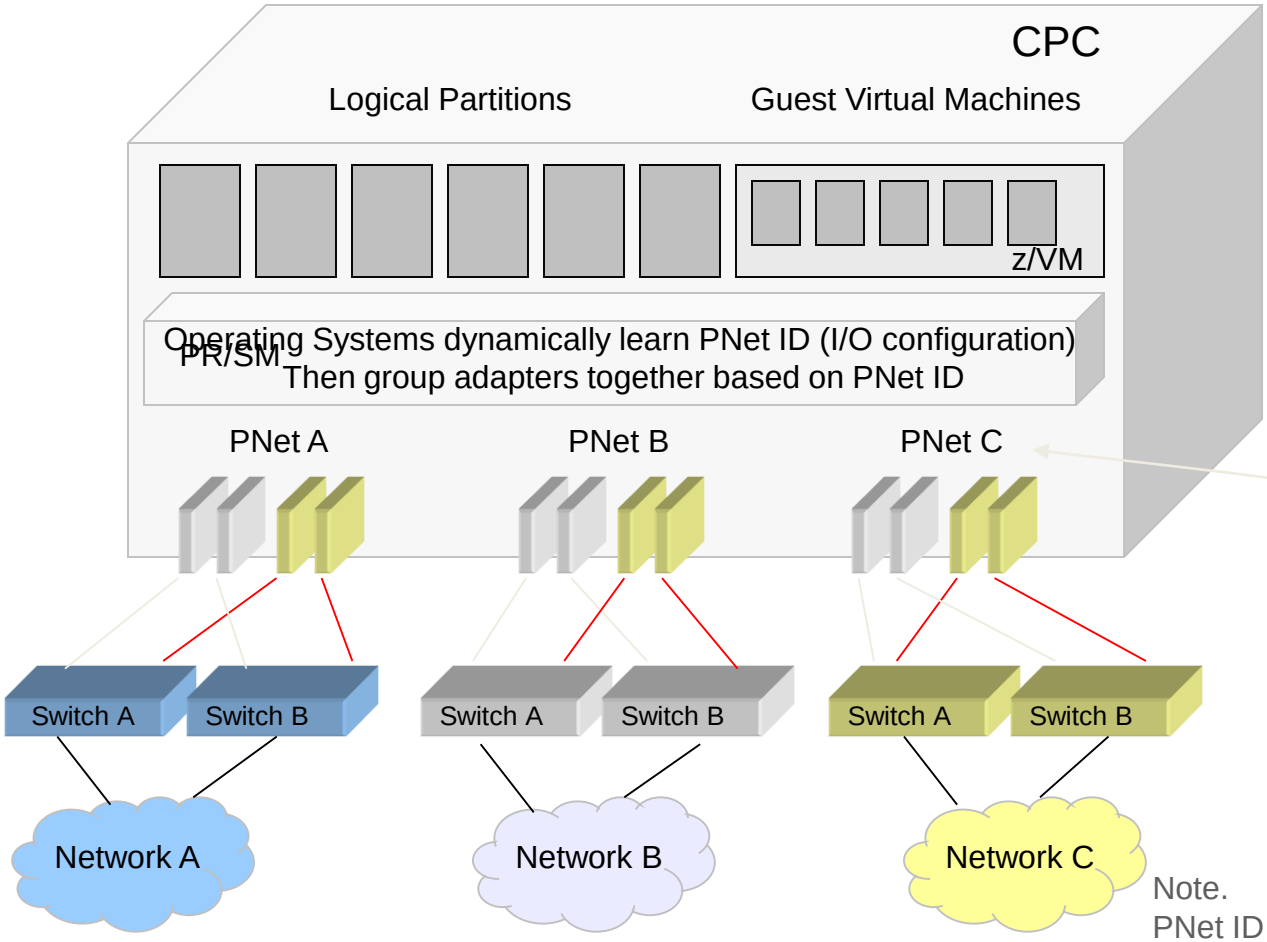
- **Existing TCP connection load balancing solutions are** *not compatible with other RDMA solutions*
- They are not aware of the RDMA protocol ***AND*** RDMA flows ***can not*** flow through intermediate nodes
- **The SMC-R protocol allows existing TCP load balancing solutions to be deployed** *with no changes*
  - *TCP Connection load balancing for SMC-R connections is actually more efficient than normal TCP/IP connections*
    - *Load balancer selects optimal back end server, data flows can then bypass the load balancer*

Network Physical Connectivity for the IBM 10Gbe RoCE Express feature
System z network adapters (Requires OSAs + RoCE Express)



Note.
Physical network ID = “Network A” could also be divided into multiple virtual networks (VLANs)

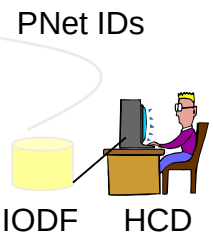
Multiple Physical RoCE Networks (Physical Network IDs)
Associating 10GbE RoCE Express features with their physical networks



Administrative model. . .

How to configure PNet ID?

(Operating Systems learn)



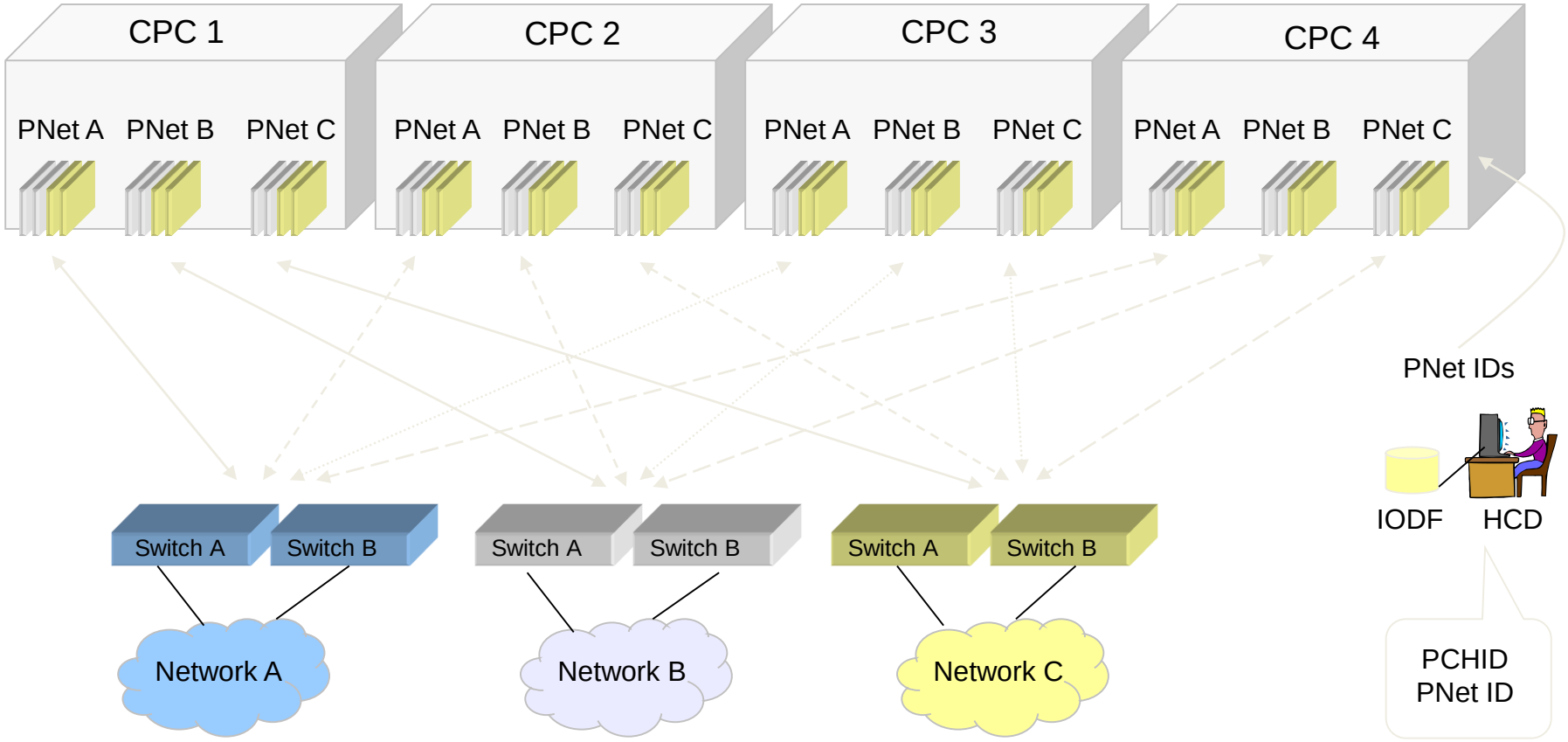
PCHID
PNet ID

Note.
PNet ID is a customer defined ID

("IEDN" is a reserved ID)

Three unique (physically isolated) Layer 2 Physical Networks

Multiple Systems with Multiple Physical Networks
Associating Physical Adapters with their Physical Networks



Three unique (physically isolated) Layer 2 Physical Networks

Value Summary: IBM 10GbE RoCE Express with SMC-R

Summary

- z/OS application workloads transparently exploit 10GbE RoCE Express feature using z/OS V2R1 SMC-R. When SMC-R is enabled:
  - transactional workloads (WAS, CICS, IMS, MQ, etc.) can potentially see an increase in their overall transaction rate (i.e. transactions per second) with a slight savings in CPU.
  - Streaming workloads (e.g. FTP) will see a CPU savings and a throughput improvement.
  - SD workloads (where the client and server both are within the SYSPLEX (i.e. WAS to DB2)) will benefit by eliminating SD in the network path (reduces cost in the SD host and reduces latency by avoiding the trip through the SD host)

Value

- Reduced network latency resulting in improved transaction rates
- Reduce CPU cost for bulk or streaming workloads or transaction workloads with larger messages (e.g. web services protocols)
- Time to value…. optimized network performance without:
  - requiring application changes
  - requiring network IP topology or security changes
  - sacrificing existing TCP/IP qualities of services (e.g. network resiliency)
  - significant operational (day to day administrative) changes

Evaluating the Benefits of RoCE-Express with SMC-R (part 1)

Physical Planning / Setup Requirements

1. z13 installation
2. z/OS V2R1 installation
3. 2 z/OS LPARs required (single CPC sufficient for IESP testing)
4. Install 2 10GbE RoCE-Express features[1]
5. RoCE Express features physical cabling:
   1. On each RoCE Express feature select one port (either port can be used)
   2. RoCE features can be connected:
      - through standard 10GbE Ethernet switch or
      - both adapters can be directly cabled together

1. SMC-R also requires standard Ethernet connectivity through OSA-Express configured in OSD mode:
   - OSA (OSD) is also required (for exploitation of SMC-R and comparison of RDMA to standard Ethernet)
   - OSA can be shared by the 2 LPARs or each LPAR can use a separate OSA

Evaluating the Benefits of RoCE-Express with SMC-R (part 2)

System software planning and Setup Requirements

1. Define RoCE-Express PFIDs (FIDs and VFs) in HCD (or IOCDS)
2. Define matching PNet IDs for OSA (OSD) and RoCE Express (physical ports) in HCD or IOCDS
3. In TCP/IP profile on Global Config:
    1. Enable SMC-R
    2. Define PFID(s) (need at least one PFID per stack)
4. Define OSA interface statements (required) on the same IP subnet (subnet mask is required and VLAN ID is optional)
5. Start OSA (RoCE Express operations are transparently controlled by OSA interface)
6. Execute test cases . . .

Evaluating the Benefits of RoCE-Express with SMC-R (part 3)

Application workloads planning and Setup Requirements

1. Application workloads can be a test tool (iPerf) or (ideally) real application (IBM middleware) workloads (or both)
2. Ideally plan to execute two types of workloads:
   - Transactional (Interactive) workloads (e.g. WAS to CICS or DB2)
   - Streaming (bulk) workloads (e.g. FTP)

Test Scenarios

- Compare application workload performance with and without SMC-R enabled:
  - Run workloads using traditional Ethernet with OSA (OSD)
  - Repeat workloads using RoCE-Express with SMC-R enabled

  compare transaction, throughput and CPU cost with existing benchmarks or sample transactions
- Feedback requested:
  - Installation / deployment (how easy to deploy and verify)
  - Performance differences (wall clock, transaction rate, CPU, throughput, etc.)

Netstat DEvlinks/-d with RNICs

```
D TCPIP,TCPCS1,NETSTAT,DEVLINKS,SMC
EZD0101I NETSTAT CS V2R1 TCPCS1
INTFNAME: EZARIUT1001C      INTFTYPE: RNIC       INTFSTATUS: READY
  PFID: 001C  PORTNUM: 1  TRLE: IUT1001C
  PNETID: ZOSNET
  VMACADDR: 02000035F740
  GIDADDR:  FE80::200:FF:FE35:F740
  INTERFACE STATISTICS:
    BYTESIN                          = 160
    INBOUND OPERATIONS               = 5
    BYTESOUT                         = 344
    OUTBOUND OPERATIONS              = 11
    SMC LINKS                        = 1
    TCP CONNECTIONS                  = 1
    INTF RECEIVE BUFFER INUSE        = 64K
  SMC LINK INFORMATION:
    LOCALSMCLINKID: 2D8F0101  REMOTESMCLINKID: 729D0101
      SMCLINKGROUPID: 2D8F0100  VLANID: 100  MTU: 1024
      LOCALGID:  FE80::200:FF:FE35:F740
        LOCALMACADDR:  02000035F740  LOCALQP:  000040
      REMOTEGID: FE80::200:1FF:FE35:F740
        REMOTEMACADDR: 02000135F740  REMOTEQP: 000041
      SMCLINKBYTESIN:           160
      SMCLINKINOPERATIONS:      5
      SMCLINKBYTESOUT:          344
      SMCLINKOUTOPERATIONS:     11
      TCP CONNECTIONS:          1
      LINK RECEIVE BUFFER INUSE: 64K
        64K   BUFFER INUSE:     64K
```

Thank You