z/OS Communications Server: Shared Memory Communications over RDMA (SMC-R) Protocol

Gus Kassimis – kassimis@us.ibm.com
IBM Enterprise Networking Solutions

Session # 17285:
Wednesday, August 12, 2015: 10:00 AM - 11:00AM

Trademarks

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

AIX*	DB2*	HiperSockets*	MQSeries*	PowerHA*	RMF	System z*	zEnterprise*	z/VM*
BladeCenter*	DFSMS	HyperSwap	NetView*	PR/SM	Smarter Planet*	System z10*	z10	z/VSE*
CICS*	EASY Tier	IMS	OMEGAMON*	PureSystems	Storwize*	Tivoli*	z10 EC	
Cognos*	FICON*	InfiniBand*	Parallel Sysplex*	Rational*	System Storage*	WebSphere*	z/OS*	
DataPower*	GDPS*	Lotus*	POWER7*	RACF*	System x*	XIV*		

* Registered trademarks of IBM Corporation

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.

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.

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

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.

Java and all Java based trademarks and logos are trademarks or registered trademarks of Oracle and/or its affiliates.

Linear Tape-Open, LTO, the LTO Logo, Ultrium, and the Ultrium logo are trademarks of HP, IBM Corp. and Quantum in the U.S. and

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

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

OpenStack is a trademark of OpenStack LLC. The OpenStack trademark policy is available on the OpenStack website.

TEALEAF is a registered trademark of Tealeaf, an IBM Company.

Windows Server and the Windows logo are trademarks of the Microsoft group of countries.

Worklight is a trademark or registered trademark of Worklight, an IBM Company.

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

* Other product and service names might be trademarks of IBM or other 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.

This information provides only general descriptions of the types and portions of workloads that are eligible for execution on Specialty Engines (e.g, zIIPs, zAAPs, and IFLs) ("SEs"). IBM authorizes customers to use IBM SE only to execute the processing of Eligible Workloads of specific Programs expressly authorized by IBM as specified in the "Authorized Use Table for IBM Machines" provided at www.ibm.com/systems/support/machine_warranties/machine_code/aut.html ("AUT"). No other workload processing is authorized for execution on an SE. IBM offers SE at a lower price than General Processors/Central Processors because customers are authorized to use SEs only to process certain types and/or amounts of workloads as specified by IBM in the AUT.

Agenda

- RDMA and RoCE technology overview
  - zEC12 and zBC12 - 10GbE RoCE Express
  - z13 and Shared ROCE Express update
  - SMC Applicability Tool

- Shared Memory Communications over RDMA (SMC-R) Overview
  - Introduction "sockets over RDMA"
  - Key Quality of Service attributes
  - Middleware enablement (programming model)
  - Supported configurations and environment

- Why is this technology important and who benefits?

- For a deep dive on SMC-R configuration please see the proceedings from SHARE in Seattle 2015:

> *16744: z/OS Communications Server: New Shared Memory Communications over RDMA (SMC-R) Protocol - Implementation - Part 2 of 2*
> *Speaker: Dave Herr(IBM Corporation)*
>
> *https://share.confex.com/share/124/webprogram/Session16744.html*

Disclaimer: All statements regarding IBM future direction or intent, including current product plans, are subject to change or withdrawal without notice and represent goals and objectives only. All information is provided for informational purposes only, on an "as is" basis, without warranty of any kind.

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
- Key requirements:
  - A reliable "lossless" network fabric (LAN for layer 2 data center network distance)
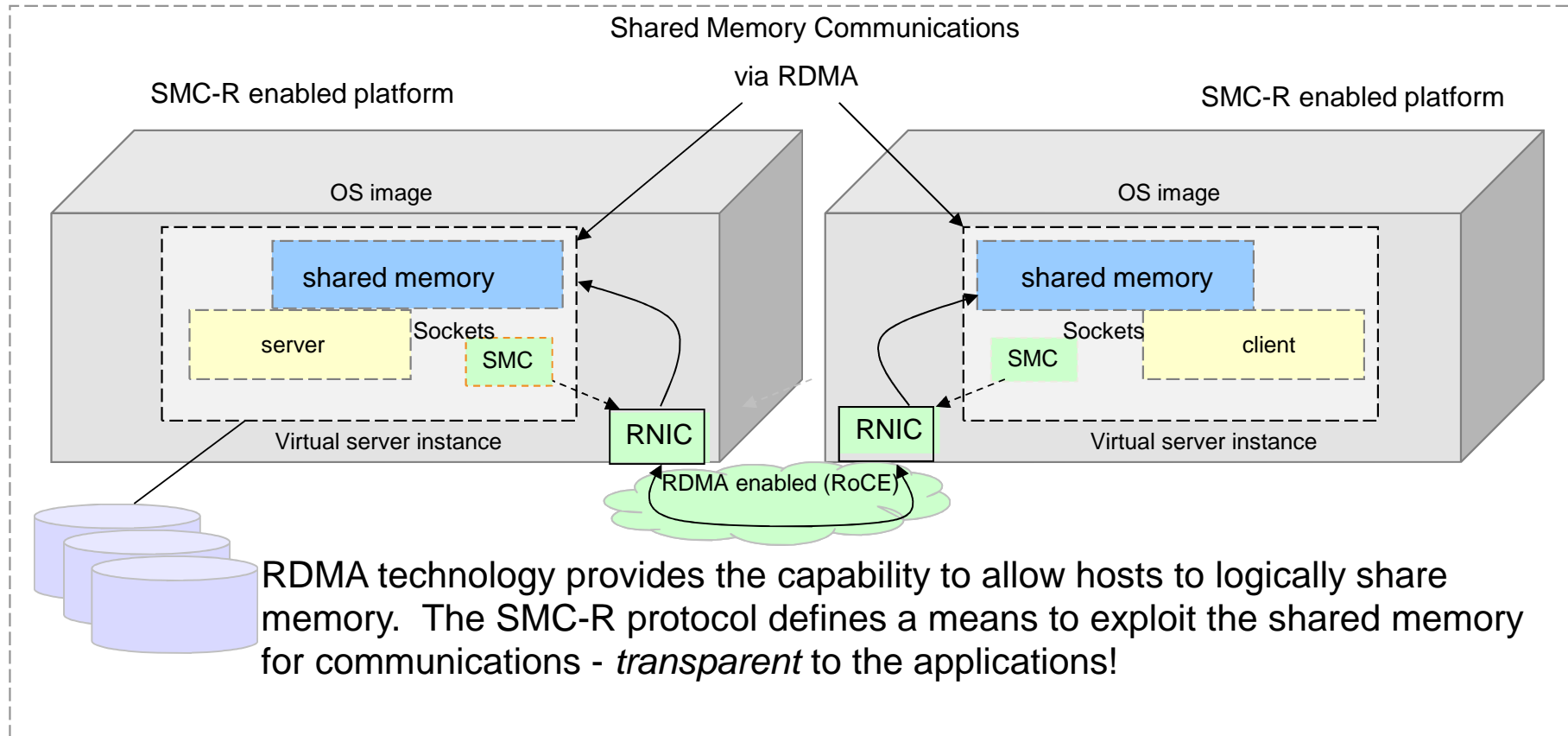 - An RDMA capable NIC (RNIC) and RDMA capable switched fabric (switches)
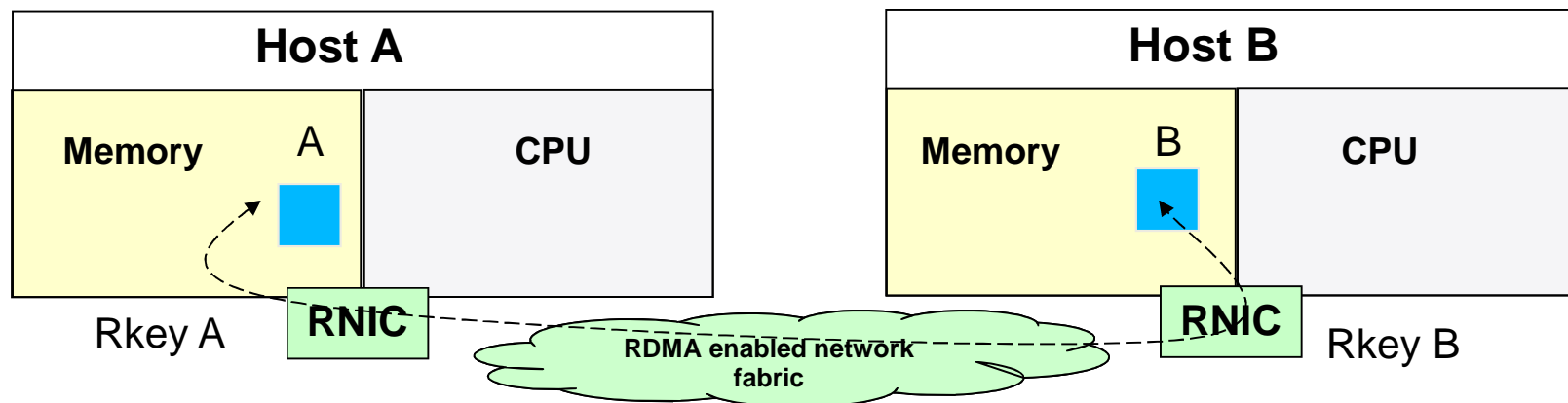
RoCE - RDMA over Converged (Enhanced) Ethernet

- 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 popular in the HPC (High Performance Computing) space

- RDMA technology is now available on Ethernet – RDMA over Converged Ethernet (RoCE)
    - RoCE uses existing Ethernet fabric but requires advanced Ethernet hardware (RDMA capable NICs and RoCE capable Ethernet switches)
    - *RoCE is a game changer!*
        - *RDMA technology becomes more affordable and prevalent in data center networks*

- Host software exploitation options fall into two general categories:
    - Native / direct application exploitation
        - Several variations, all involve deep level of expertise in RDMA and a new programming model
    - *Transparent application exploitation (e.g. sockets based)*
        - *Improve Time To Value by automatically exploiting RDMA/RoCE for sockets based TCP applications*

"Shared Memory Communications over RDMA" concepts

Clustered Systems

Shared Memory Communications via RDMA

SMC-R enabled platform

SMC-R enabled platform

OS image

OS image

shared memory

shared memory

Sockets

Sockets

server

SMC

SMC

client

Virtual server instance

RNIC

RNIC

Virtual server instance

RDMA enabled (RoCE)

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!

This solution is referred to as *SMC-R* (Shared Memory Communications over RDMA). SMC-R is an *open* sockets over RDMA protocol that provides transparent exploitation of RDMA (for TCP based applications) while preserving key functions and qualities of service from the TCP/IP ecosystem that enterprise level servers/network depend on!

Final Draft IETF (Internet Enginnering Task Force) RFC for SMC-R submitted:

https://datatracker.ietf.org/doc/draft-fox-tcpm-shared-memory-rdma/

IBM

New innovations available on zBC12 and zEC12

NEW	NEW	ENHANCED	ENHANCED	NEW
Data Compression Acceleration	**High Speed Communication Fabric**	**Flash Technology Exploitation**	**Proactive Systems Health Analytics**	**Hybrid Computing Enhancements**
Reduce CP consumption, free up storage & speed cross platform data exchange	Optimize server to server networking with reduced latency and lower CPU overhead	Improve availability and performance during critical workload transitions, now with dynamic reconfiguration; Coupling Facility exploitation (SOD)	Increase availability by detecting unusual application or system behaviors for faster problem resolution before they disrupt business	x86 blade resource optimization; New alert & notification for blade virtual servers; Latest x86 OS support; Expanding futures roadmap
zEDC Express	*10GbE RoCE Express*	*IBM Flash Express*	*IBM zAware*	*zBX Mod 003; zManager Automate; Ensembl Availability Manager; DataPower Virtual appliance SoD*

Optimize server to server networking – transparently
"HiperSockets™ -like" capability across systems

Network latency for z/OS TCP/IP based OLTP workloads *reduced* by up to **80%** **

zEC12   zBC12

Shared Memory Communications (SMC-R):

Exploit RDMA over Converged Ethernet (RoCE) to deliver superior communications performance for TCP based applications

Networking related CPU consumption for z/OS TCP/IP based workloads with streaming data patterns *reduced* by up to **60%** with a *network throughput* increase of up to **60%** ***

Typical Client Use Cases:

Help to reduce both latency and CPU resource consumption over traditional TCP/IP for communications across z/OS systems

Any z/OS TCP sockets based workload can **seamlessly** use SMC-R without requiring any application changes

NEW z/OS V2.1 SMC-R   **NEW** z/VM 6.3 support for guests   **NEW** 10GbE RoCE Express

** Based on internal IBM benchmarks in a controlled environment of modeled z/OS TCP sockets-based workloads with request/response traffic patterns using SMC-R (10GbE RoCE Express feature) vs TCP/IP (10GbE OSA Express feature). The actual response times and CPU savings any user will experience will vary.

*** Based on internal IBM benchmarks in a controlled environment of modeled z/OS TCP sockets-based workloads with streaming traffic patterns using SMC-R (10GbE RoCE Express feature) vs TCP/IP (10GbE OSA Express feature). The actual response times and CPU savings any user will experience will vary.

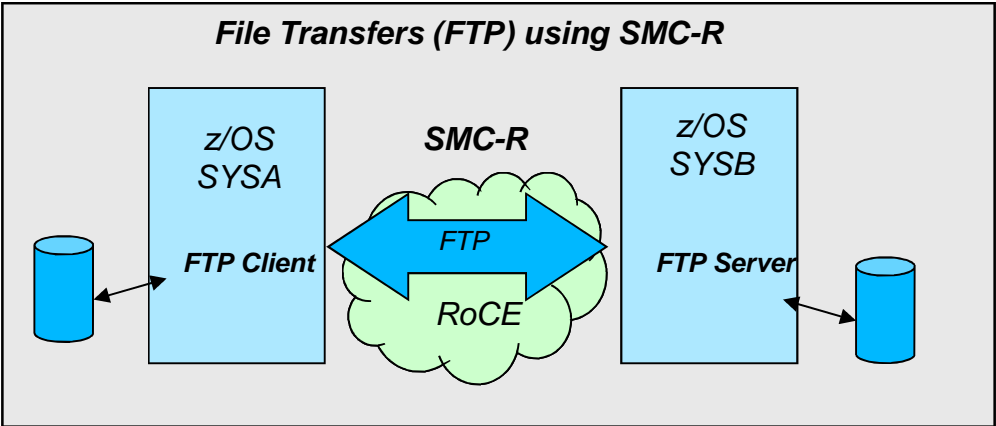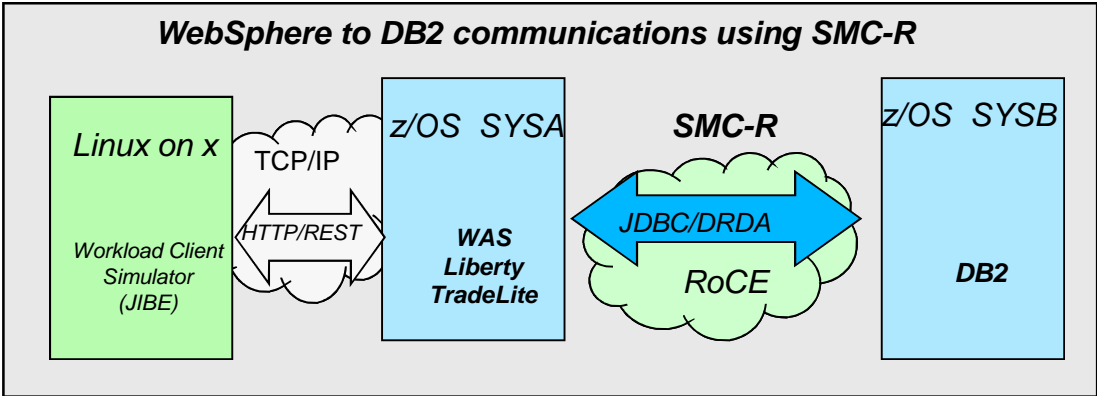# Use cases for SMC-R and 10GbE RoCE Express for z/OS to z/OS communications

Use Cases

- Application servers such as the z/OS WebSphere Application Server communicating (via TCP based communications) with CICS, IMS or DB2 – particularly when the application is network intensive and transaction oriented

- Transactional workloads that exchange larger messages (e.g. web services such as WAS to DB2 or CICS) will see benefit.

- Streaming (or bulk) application workloads (e.g. FTP) communicating z/OS to z/OS TCP will see improvements in both CPU and throughput

- Applications that use z/OS to z/OS TCP based communications using Sysplex Distributor

Plus … *Transparent to application software – no changes required!*

Performance impact of SMC-R on real z/OS workloads

40% reduction in overall transaction response time for WebSphere Application Server v8.5 Liberty profile TradeLite workload accessing z/OS DB2 in another system measured in internal benchmarks *

WebSphere to DB2 communications using SMC-R

Linux on x

Workload Client Simulator (JIBE)

TCP/IP

HTTP/REST

z/OS SYSA

WAS Liberty TradeLite

SMC-R

JDBC/DRDA

RoCE

z/OS SYSB

DB2

File Transfers (FTP) using SMC-R

z/OS SYSA

FTP Client
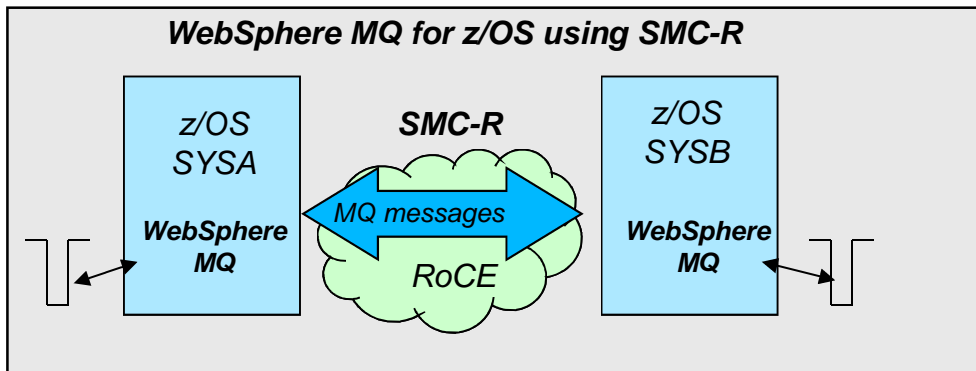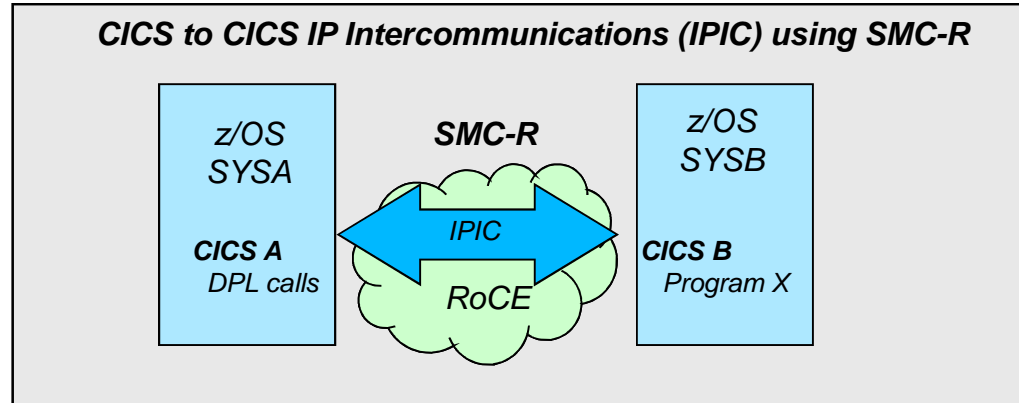
SMC-R

FTP

RoCE

z/OS SYSB

FTP Server

Up to **50% CPU savings** for FTP binary file transfers across z/OS systems when using SMC-R vs standard TCP/IP **

* Based on projections and measurements completed in a controlled environment. Results may vary by customer based on individual workload, configuration and software levels.
** Based on internal IBM benchmarks in a controlled environment using z/OS V2R1 Communications Server FTP client and FTP server, transferring a 1.2GB binary file using SMC-R (10GbE RoCE Express feature) vs standard TCP/IP (10GbE OSA Express4 feature). The actual CPU savings any user will experience may vary.

Performance impact of SMC-R on real z/OS workloads (cont)

Up to 48% reduction in response time and up to 10% CPU savings for CICS transactions using DPL (Distributed Program Link) to invoke programs in remote CICS regions in another z/OS system via CICS IP interconnectivity (IPIC) when using SMC-R vs standard TCP/IP *

CICS to CICS IP Intercommunications (IPIC) using SMC-R

z/OS
SYSA

SMC-R

CICS A
DPL calls

IPIC

RoCE

z/OS
SYSB

CICS B
Program X

WebSphere MQ for z/OS using SMC-R

z/OS
SYSA

SMC-R

**WebSphere
MQ**

MQ messages

RoCE

z/OS
SYSB

**WebSphere
MQ**

WebSphere MQ for z/OS *realizes up to 200% increase in messages per second* it can deliver across z/OS systems when using SMC-R vs standard TCP/IP **

* Based on internal IBM benchmarks using a modeled CICS workload driving a CICS transaction that performs 5 DPL (Distributed Program Link) calls to a CICS region on a remote z/OS system via CICS IP interconnectivity (IPIC), using 32K input/output containers. Response times and CPU savings measured on z/OS system initiating the DPL calls. The actual response times and CPU savings any user will experience will vary.

** Based on internal IBM benchmarks using a modeled WebSphere MQ for z/OS workload driving non-persistent messages across z/OS systems in a request/response pattern. The benchmarks included various data sizes and number of channel pairs The actual throughput and CPU savings users will experience may vary based on the user workload and configuration.
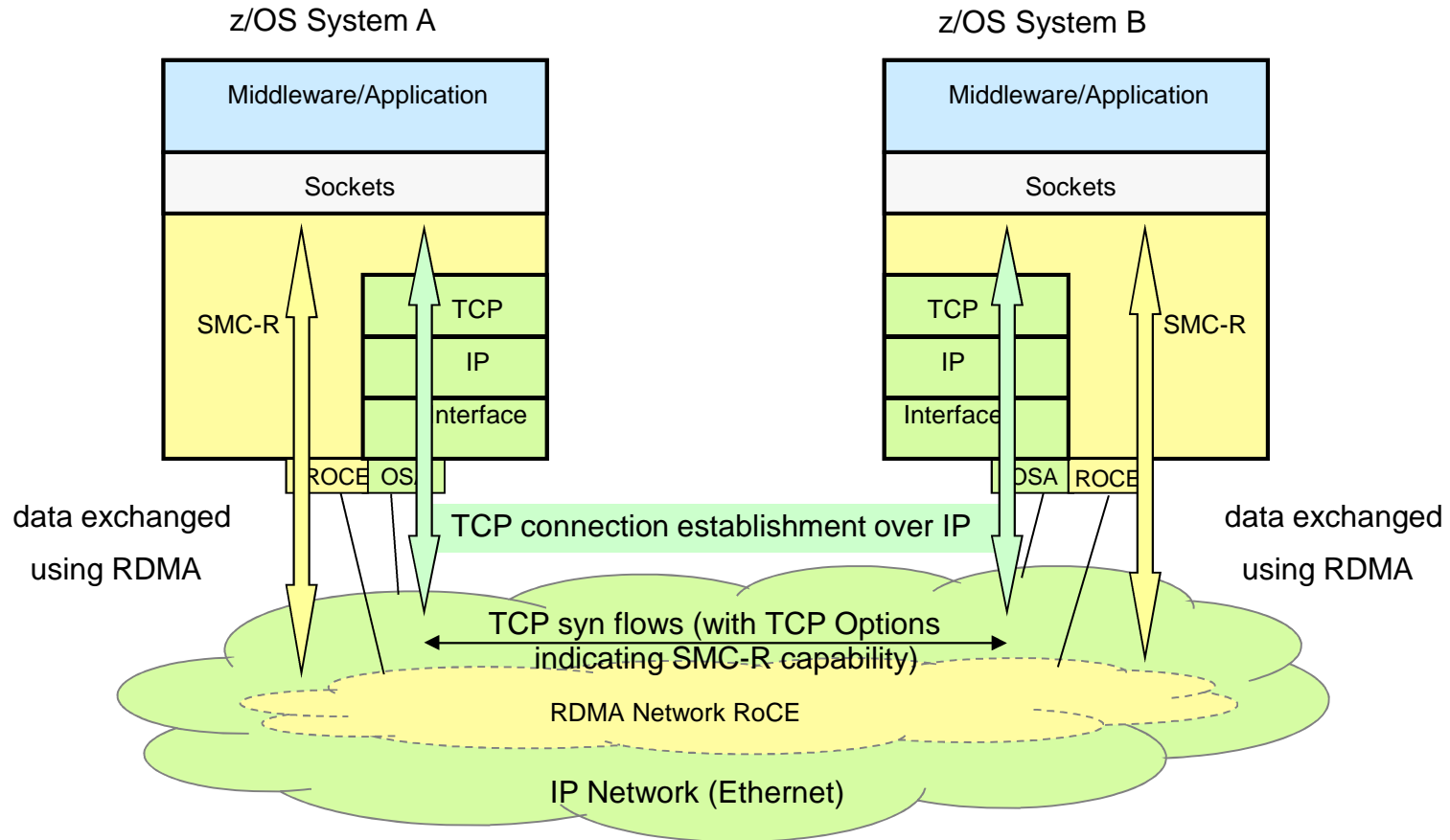
For additional SMC-R performance information

17291: z/OS Communications Server Performance: Updates and Recommendations
Wednesday, August 12, 2015: 11:15 AM-12:15 PM
Asia 2 (Dolphin)
Speaker: Dave Herr (IBM Corporation)

Dynamic Transition from TCP to SMC-R

z/OS System A

z/OS System B

Middleware/Application

Sockets

SMC-R

TCP

IP

Interface

ROCE OSA

Middleware/Application

Sockets

TCP

IP

Interface

OSA ROCE

SMC-R

data exchanged
using RDMA

data exchanged
using RDMA

TCP connection establishment over IP

TCP syn flows (with TCP Options
indicating SMC-R capability)

RDMA Network RoCE

IP Network (Ethernet)

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

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

SMC-R Overview

- Shared Memory Communications over RDMA (SMC-R) is a protocol that allows *TCP sockets* applications to transparently exploit RDMA (RoCE)

- SMC-R is a "hybrid" solution that:
  - Uses TCP connection (3-way handshake) to establish SMC-R connection
  - Each TCP end point exchanges TCP options that indicate whether it supports the SMC-R protocol
  - SMC-R "rendezvous" (RDMA attributes) information is then exchanged within the TCP data stream (similar to SSL handshake)
  - Socket application data is exchanged via RDMA (write operations)
  - TCP connection remains active (controls SMC-R connection)
  - This model preserves many critical existing operational and network management features of TCP/IP

Why a "Hybrid Protocol"? (Why TCP/IP + SMC-R?)

- The Hybrid model of SMC-R leverages key existing attributes:

  – Follows standard TCP/IP connection setup
  – Dynamically switches to RDMA (SMC-R)
  – TCP connection remains active (idle) and is used to control the SMC-R connection
  – Preserves critical operational and network management TCP/IP features such as:
    - Minimal (or zero) IP topology changes
    - Compatibility with TCP connection level load balancers (e.g Sysplex Distributor)
    - Preserves existing IP security model (e.g. IP filters, policy, VLANs, SSL etc.)
  – Minimal network admin / management changes

- *Significant reduction in Time to Value!*

SMC-R and 10GbE RoCE Express Requirements

- **Operating system requirements**
  - Requires z/OS 2.1 which supports the SMC-R protocol

- **Server requirements**
  - Exclusive to zEC12 (with Driver 15E) and zBC12
  - New 10 GbE RoCE Express feature for PCIe I/O drawer (FC#0411)
    - Single port enabled for use by SMC-R
    - Each feature must be dedicated to one LPAR
    - "RNIC" and "RoCE Express" terms in this presentation are synonyms
  - Recommended minimum configuration two features per LPAR for redundancy
    - Up to 16 features supported
  - OSA Express – either 1 GbE or 10 GbE
    - Configured in QDIO mode (OSD CHPIDs only, not OSX)
    - Does not need to be dedicated to the LPAR
  - Standard 10GbE Switch or point to point configuration supported
    - Does not need to be CEE capable
    - Switch must support and have enabled Global pause frame (a standard Ethernet switch feature for Ethernet flow control described in the IEEE 802.3x standard)

IBM

SMC-R and Shared ROCE Support – IBM z13 System

- Shared RoCE support - Available exclusively on new IBM z13 System

  - Allows concurrent sharing of a RoCE Express feature by multiple virtual servers (OS instances)

    - Efficient sharing for an adapter (getting the Hypervisor out of the data path)
    - Up to 31 virtual servers (LPARs or 2nd level guests under zVM)
    - Will also enable use of both RoCE Express ports by z/OS

  - z/OS support will be available in z/OS V2R2 (base) and on z/OS V2R1 via APAR/PTF

    - z/OS V2R1: UI28823 / PI38739 and UI28842 / PI38739 and UA77894 / OA47561

10GbE RoCE Express

Shared RoCE
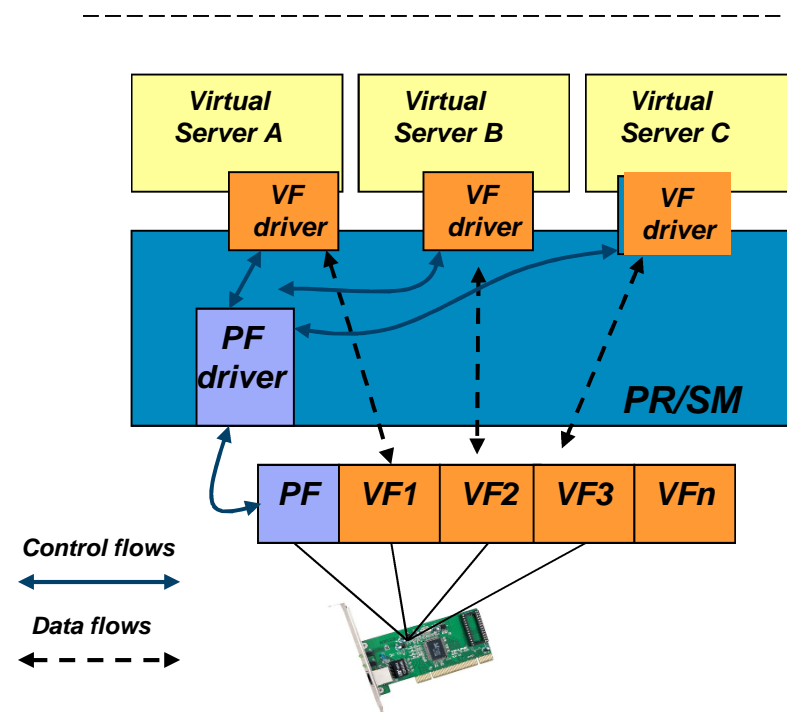
SMC-R adapter virtualization: Overview

• Multiple PFIDs (PCIe Function IDs) with unique Virtual Function IDs are configured for each physical adapter (PCHID) in HCD (IOCDS)
  • Up to 31 PFIDs (VFs) 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)
  • Up to 16 physical adapters per CPC (no change)
• Adapter virtualization is transparent to application software

SMC-R adapter virtualization: TCP/IP Configuration

- No (minor) changes in TCP/IP configuration
  - Global Configuration Statement SMCR option continues to define PFID / port
- VFs are almost transparent to CommServer
- TCP/IP requires (consumes) a single PFID (VF) per port
  - All VLANs (if multiple) uses a single PFID (VF)
- Both physical 10GbE ports can be exploited - another PFID (VF) must be configured (per port)
- Multiple TCP/IP stacks (CINET) in a single system exploiting SMC-R:
  - Each TCP/IP stack requires a unique PFID (VF)
  - Potential migration consideration (prior to Shared RoCE support, multiple stacks could use the same PFID for SMC-R – each TCP/IP stack now needs a unique PFID)

SMC-R adapter virtualization: Product externals (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. For best practices (HA) the objective is to provision 2 FIDs (PCHIDs) having unique paths (opposite values 0 and 1). This display (PFIP) can be used to validate your configuration for HA.

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

SMC-R adapter virtualization: z/OS DISPLAY PCIE

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

Same PCHID indicates same physical RoCE
Express Feature

```
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
```

Unique PFID for each Shared RoCE Express
Unique VFN (within same PCHID) for each Shared
RoCE Express instance

SMC-R TCP Connection Eligibility

Rules… All eligible hosts must:

1. be **SMCR enabled** (z/OS V2R1 and having SMC-R enabled with RoCE Express cards allocated)

2. *Physical Connectivity:*

   – Direct Ethernet (OSA Express) and RoCE connectivity to the same physical Layer 2 network

3. *IP Connectivity:*

   – (on a per PNet basis) have direct access to the ***same IP subnet and VLAN*** (i.e. no IP routing or firewalls)

     Note. VLANs are optional for customer networks (i.e. on a per PNet ID basis either define a single IP interface with an optional VLAN ID or if multiple IP interfaces are required then all must have a VLAN ID)

4. **not** require IPSec (SSL is supported)

… then during the traditional TCP/IP connection setup the above criteria is dynamically assessed (via SMCR rendezvous process)… where all socket based TCP connections among the eligible hosts that connect over the IP fabric will automatically and transparently exploit SMCR

IP/Ethernet VLAN topology - Implications on SMC-R communications
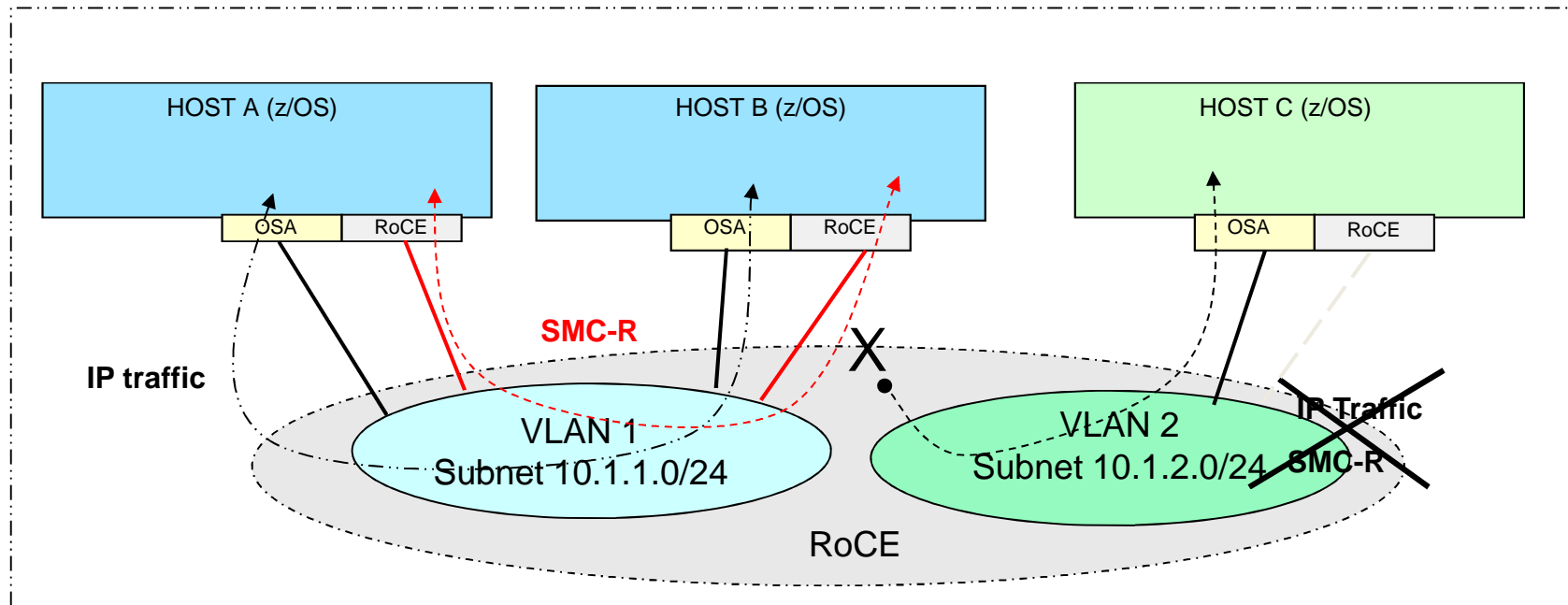
- SMC-R requires both hosts to be on the **same layer 2 network** (physical LAN or VLAN) and in the **same IP subnet** when communicating via TCP/IP (i.e. have a direct communication path without the need to traverse IP routers)
- VLANs allows users to subdivide a LAN into isolated "virtual networks" isolating servers to a specific authorized group. VLANs are optional.
- Since SMC-R connection processing leverages your existing IP topology (TCP/IP connection setup) SMC-R connections transparently "inherit" the same VLAN and IP Subnet connection eligibility attributes of the associated TCP connection. When VLANs are in use, SMC-R connections then become VLAN qualified.

Note. RDMA is not routable (i.e. cannot be routed using IP routers/firewalls)

SMC-R and RoCE performance benchmarks 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)

SMC-R and RoCE performance benchmarks at distance

- IBM System z™ Qualified Wavelength Division Multiplexer (WDM) products for Multi-site Sysplex and GDPS ® solutions qualification testing updated to include RoCE and SMC-R. Vendors who have 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-304.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-304.ibm.com/servers/resourcelink/lib03020.nsf/pages/ciscoSystemsInc?OpenDocument&pathID=

    3. Huawei OptiX OSN 8800 and 6800 DWDM – Release 5.51.08.38, TN11LOA: supports PS-IFB and 10GbE and is certified for RoCE

        https://www-304.ibm.com/servers/resourcelink/lib03020.nsf/pages/HuaweiTechnology?OpenDocument&pathID=

    - *To monitor the latest products qualified refer to:*
        https://www-304.ibm.com/servers/resourcelink/lib03020.nsf/pages/systemzQualifiedWdmProductsForGdpsSolutions?OpenDocument

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

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

- **For more details:**
  ftp://public.dhe.ibm.com/software/os/systemz/pdf/SMCR_and_RoCE_Performance_at_distance_26sept14.pdf

Determining SMC-R benefits – SMC Applicability Tool

- Several customers have expressed interest in SMC-R
  - One of the first questions that is raised is "What benefit will SMC-R provide in my environment?"
    - Some users are well aware of significant traffic patterns that can benefit from SMC-R
    - But others are unsure on how much of their traffic is z/OS to z/OS and how much of that traffic is well suited to SMC-R
  - Reviewing SMF records, using Netstat displays, Ctrace analysis and reports from various Network Management products can provide these insights
    - **But it can be a time consuming activity that requires significant expertise**

SMC Applicability Tool

- A tool that will help customers determine the value of SMC-R in their environment with minimal effort and minimal impact
  - Part of the TCP/IP stack: Gather new statistics that are used to project SMC-R applicability and benefits for the current system
    - Minimal system overhead, no changes in TCP/IP network flows
    - Produces reports on potential benefits of enabling SMC-R
  - Also available *now* on existing z/OS releases via the following maintenance (includes latest enhancements):
    - z/OS V2R1 - APAR PI39612, PTF UI28867
    - z/OS V1R13 - APAR PI41713, PTF UI29684
  - Does not require SMC-R to be enabled
  - Does not require RoCE Express Features or any specific System z processor
  - Can be used for determining potential benefits prior to moving to latest software and hardware levels

SMC Applicability Tool …

- Activated by Operator command - ***Vary TCPIP,,SMCAT,dsn(smcatconfig)*** – Input dataset contains:
  - Interval Duration, list of IP addresses or IP subnets of peer z/OS systems ((i.e. systems that we can use SMC-R for)
    - If subnets are used, the entire subnet must be comprised of z/OS systems that are SMC-R eligible
    - It is important that all the IP addresses used for establishing TCP connections are specified (including DVIPAs, etc.)
  - At the end of the interval a report is generated that includes:
    1. % of TCP traffic that is eligible for SMC-R (SMC-R Eligible Traffic)
       - All traffic that matches configured IP addresses
    2. % of SMC-R Eligible Traffic that is well suited to SMC-R (excludes workloads with very short lived TCP connections that have trivial payloads)
       - Includes break out of TCP traffic send sizes (i.e. how large is the payload of each send request)
       - Helps users quantify SMC-R benefit (reduced latency vs reduced CPU cost)

SMC Applicability Tool …

- Report includes 2 sections
  - Configured TCP traffic that could not use SMC-R without changes (does not meet direct route connectivity requirements)
    - This represents the opportunity of re-configuring routing topology to enable SMC-R
  - Configured TCP traffic that can use SMC-R as is immediately (meets SMC-R direct route connectivity requirements)
    - Detected by the tool automatically (non-routed traffic)

SMC Applicability Tool Sample Report (Part 1. Direct Connections)

```
TCP SMC traffic analysis for matching direct connections
-----------------------------------------------------------
  Connections meeting direct connectivity requirements

    20% of all TCP connections can use SMC (eligible)
      90% of eligible connections are well-suited for SMC
    18% of all TCP traffic (segments) is well-suited for SMC
      20% of outbound traffic (segments) is well-suited for SMC
      16% of inbound traffic (segments) is well-suited for SMC


  Interval Details:
    Total TCP Connections:                          120
    Total SMC eligible connections:                  24
        Total SMC well-suited connections:           22
    Total outbound traffic (in segments)         110000
        SMC well-suited outbound traffic (in segments)   22000
    Total inbound traffic (in segments)          100000
        SMC well-suited inbound traffic (in segments)    16000


  Application send sizes used for well-suited connections:
    Size                            # sends      Percentage
    ----                            -------      ----------
    1500 (<=1500):                      200        39%
    4K (>1500 and <=4k):                150        29%
    8K (>4k and <= 8k):                   0         0%
    16K (>8k and <= 16k):                 0         0%
    32K (>16k and <= 32k):                0         0%
    64K (>32k and <= 64k):               50        10%
    256K (>64K and <= 256K):            109        22%
    >256K:                                0         0%
```
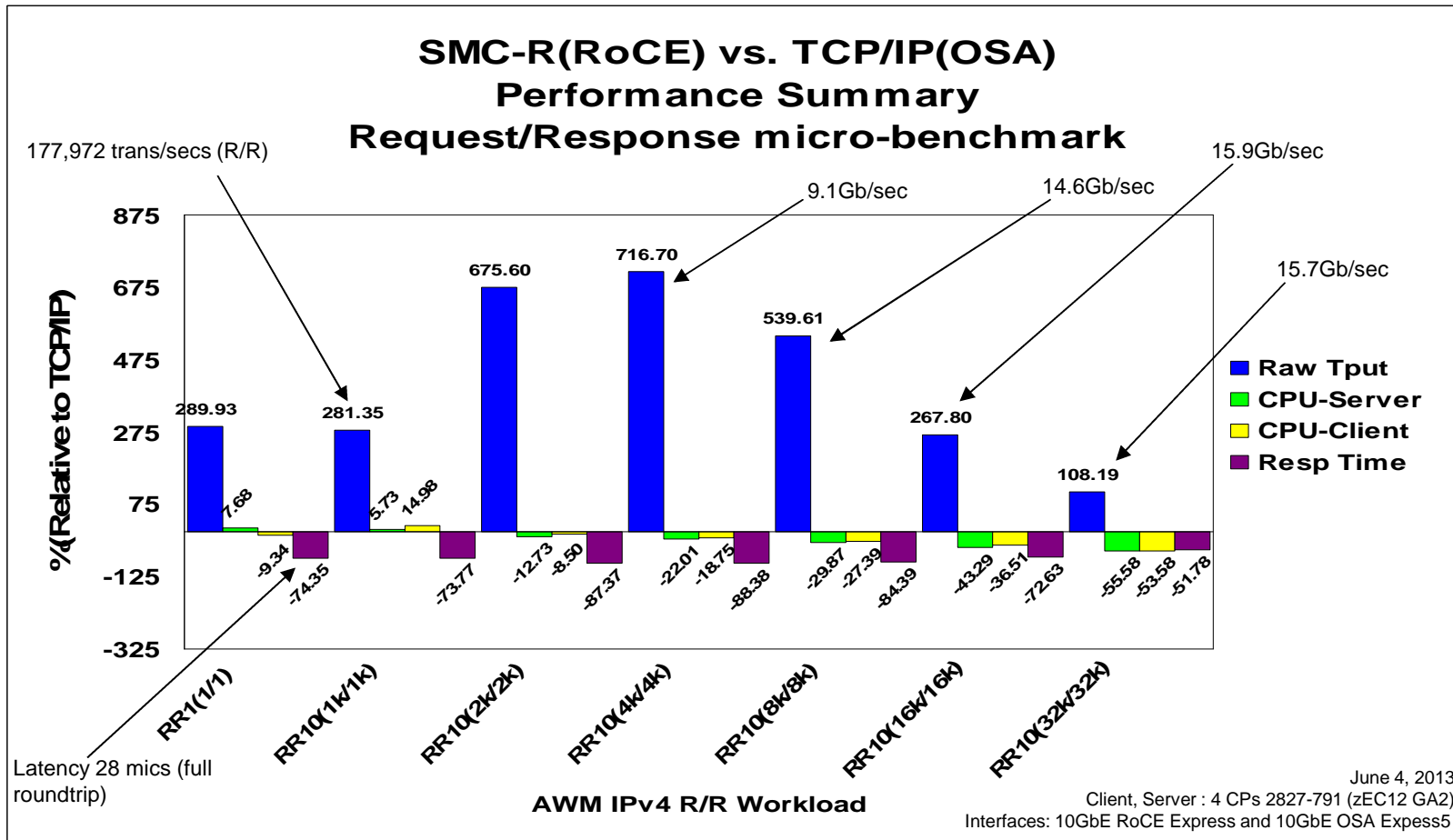
> How much of my TCP workload can benefit from SMC?

> What kind of CPU savings can I expect from SMC?

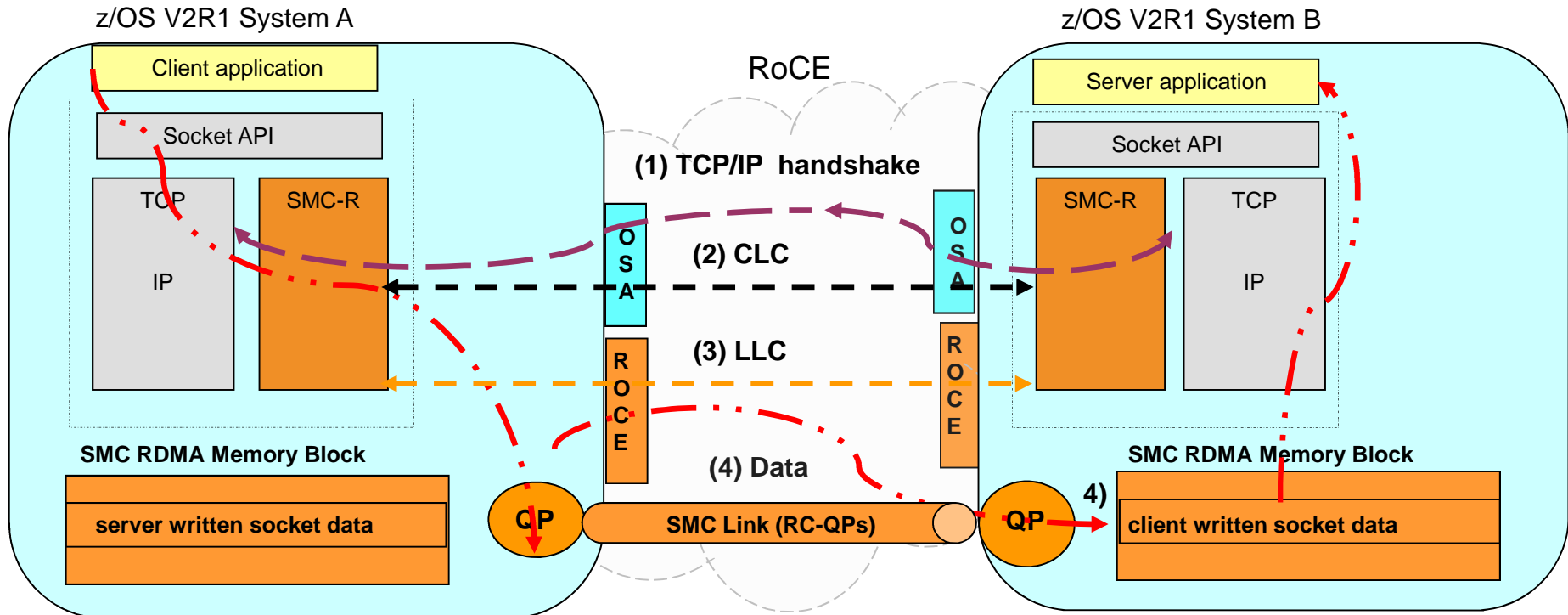> This report is repeated for indirect IP connections (different subnets)

```
End of report
```

SMC-R(RoCE) vs. TCP/IP(OSA)
Performance Summary
Request/Response micro-benchmark

177,972 trans/secs (R/R)

9.1Gb/sec

14.6Gb/sec

15.9Gb/sec

15.7Gb/sec

% (Relative to TCP/IP)

875

675

475

275

75

-125

-325

289.93

7.68
-9.34
-74.35

281.35

5.73   14.98
-73.77

675.60

-12.73 -8.50
-87.37

716.70

-22.01 -18.75
-88.38

539.61

-29.87 -27.39
-84.39

267.80

-43.29 -36.51
-72.63

108.19

-55.58 -53.58 -51.78

Legend:
- ■ Raw Tput
- ■ CPU-Server
- ■ CPU-Client
- ■ Resp Time

RR1(1/1)   RR10(1k/1k)   RR10(2k/2k)   RR10(4k/4k)   RR10(8k/8k)   RR10(16k/16k)   RR10(32k/32k)

Latency 28 mics (full roundtrip)

AWM IPv4 R/R Workload

June 4, 2013
Client, Server : 4 CPs 2827-791 (zEC12 GA2)
Interfaces: 10GbE RoCE Express and 10GbE OSA Expess5

Significant Latency reduction across all data sizes (52-88%)
Reduced CPU cost as payload increases (up to 56% CPU savings)
Impressive throughput gains across all data sizes (Up to +717%)

Note: vs typical OSA customer configuration
MTU (1500), Large Send disabled
RoCE MTU: 1K

SMC-R enhancements – z/OS V2R2

- SMC-R Autonomics
  - Automatically cache SMC-R negative set-up attempts
    - Avoid future attempts to negotiate SMC-R with the specific peer
  - Automatically determine whether SMC-R is suitable for a given z/OS TCP Server
    - Workloads with very short lived connections and very small payloads may see no benefit from SMC-R
    - Automatically disables SMC-R negotiations for that server port

- Support 4K MTU for RoCE
  - In addition to existing 1K and 2K MTU

- Enhancements in reporting of SMC-R connection local and remote buffer sizes
  - Provided on Network Management Interfaces (NMI) and TCP/IP SMF records
    - NMI GetConnectionDetail API
    - SMF Record (Type 119)

SMC-R (Contact and RDMA Processing - Concepts)

z/OS V2R1 System A

z/OS V2R1 System B

RoCE

Client application

Server application

Socket API

Socket API

(1) TCP/IP handshake

TCP

SMC-R

SMC-R

TCP

IP

O S A

O S A

(2) CLC

IP

R O C E

(3) LLC

R O C E

SMC RDMA Memory Block

(4) Data

SMC RDMA Memory Block

server written socket data

QP

SMC Link (RC-QPs)

QP

4)

client written socket data

1) Application issues standard TCP Connect; Normal TCP/IP connection (3-way syn) handshake; Determine ability/desire to support SMC-Remote (based on TCP option)

2) When both hosts provide SMC TCP option then exchange RDMA credentials (QPs, RMBEs, GIDs, etc.) within TCP data stream (CLC messages – Connection Level Control messages) … can still fall back to IP

3) **If first contact**…. then establish point-to-point SMC Link via SMC LLC (Link Layer Control) commands (RDMA-Memory-Block (RMB) pair over RC-QP… the same link (QP/RMB) can be used for multiple TCP connections across same 2 peers)

4) Applications issue standard socket send; SMC-R performs RDMA-write into partner's RMBE slot (RMB Element); peer consumes data via standard socket read

© 2015 IBM Corporation

SMC-R terms and concepts

- SMC-R link

- SMC-R link group

- Queue pair

- Remote memory buffer

- Staging buffer

SMC-R link introduction

- Purpose of rendezvous processing is to select the proper SMC-R link for this TCP connection to use, or create a new link if necessary
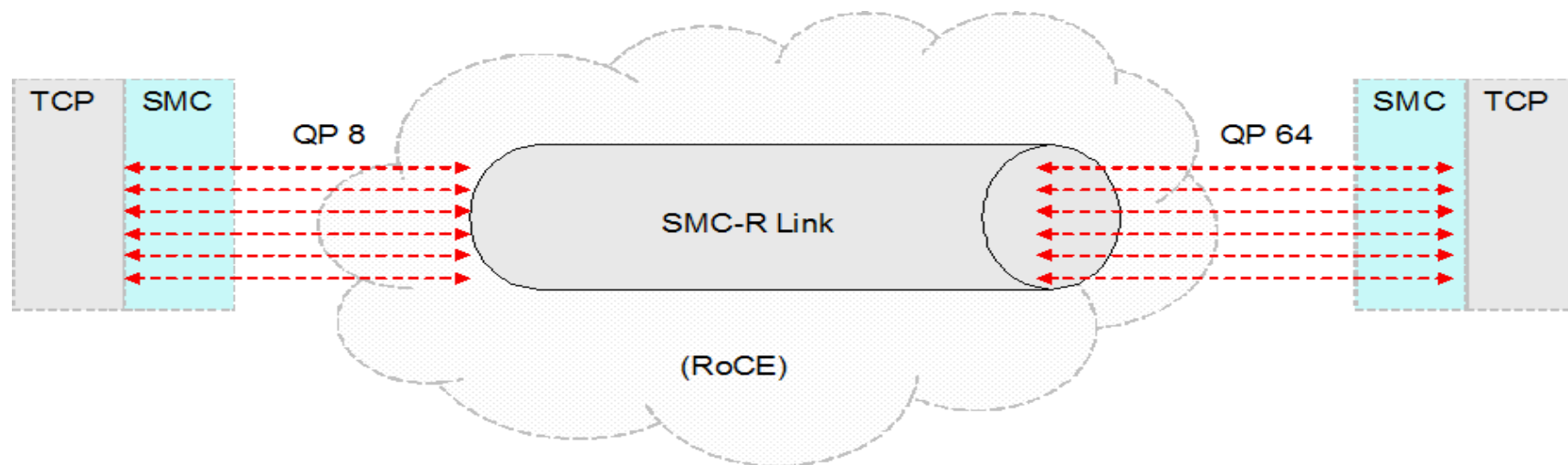
- SMC-R link is a logical point-to-point RDMA connection between two peers
  - First TCP connection between the peers that uses SMC-R causes the SMC-R link to be established
  - SMC-R link can be re-used by subsequent TCP connections
- Multiple SMC-R links can exist between two peers
  - Different SMC-R links are created if server and client roles are reversed between peers
  - Different virtual LANs, when used, require different links as well

SMC-R link definition

- SMC-R link is identified by the combination of:
  - Remote and local virtual MAC (VMAC)
  - Remote and local Global ID (GID)
    - IPv6 link-local address derived from VMAC
  - Remote and local queue pair (QP)
  - Virtual LAN (VLAN), when used to differentiate LAN traffic

- Peers assign and exchange 4-byte link IDs for easier correlation

QP 8

RNIC A

SMC-R Link   (VLAN X)

VMAC A (GID A)   VMAC C (GID C)

QP 64

RNIC C

RoCE
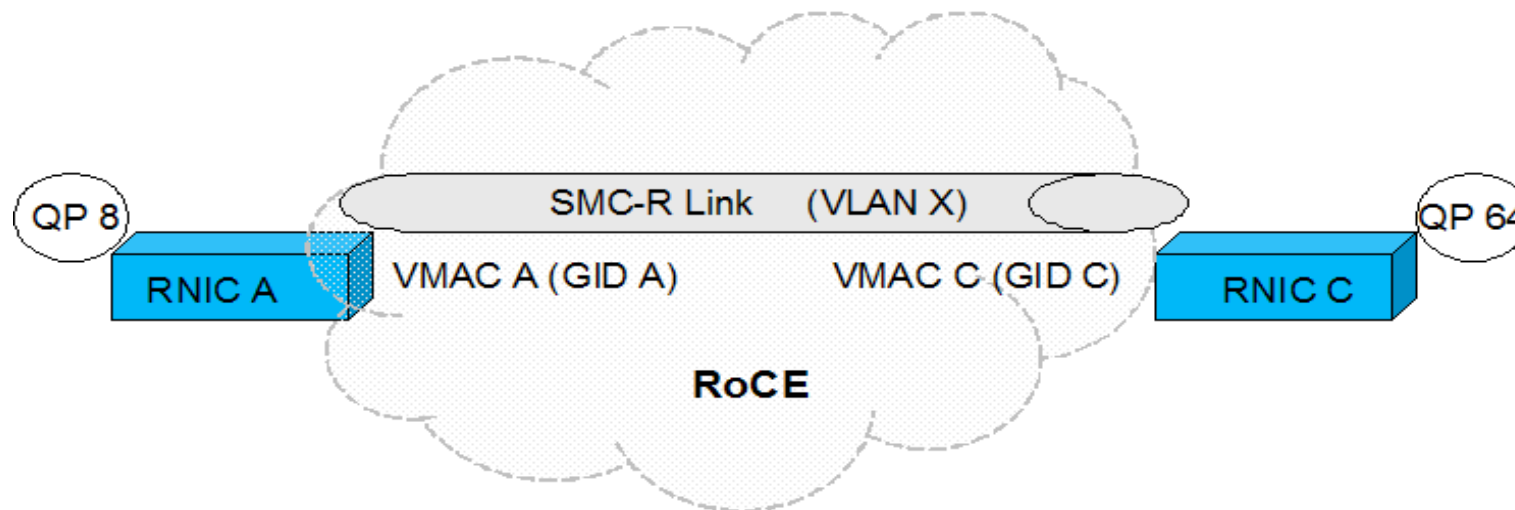
What are queue pairs?

- A queue pair (QP) represents one end of the SMC-R link

- Reliably connected QPs (RC-QPs) form a logical point-to-point connection
  - Allows exactly one pair of RDMA peers to send and receive RDMA messages between themselves

- RNIC adapter associates units of work to a specific QP
  - Adapter notifies the device driver when a unit of work is available to be processed
    - Data has been delivered to the peer, or data has been received from the peer
  - The device driver directs the units of work to the proper TCP/IP stack
  - TCP/IP stack (SMC layer) directs the unit of work to the proper TCP connection
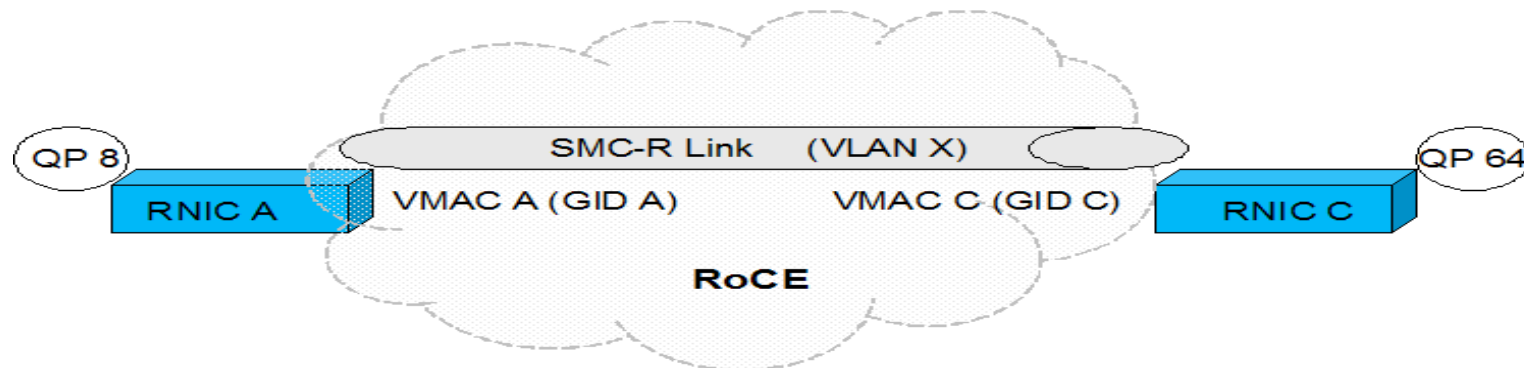
SMC-R link group introduction

z/OS Comm Server will create an SMC-R link group for redundancy and load balancing purposes
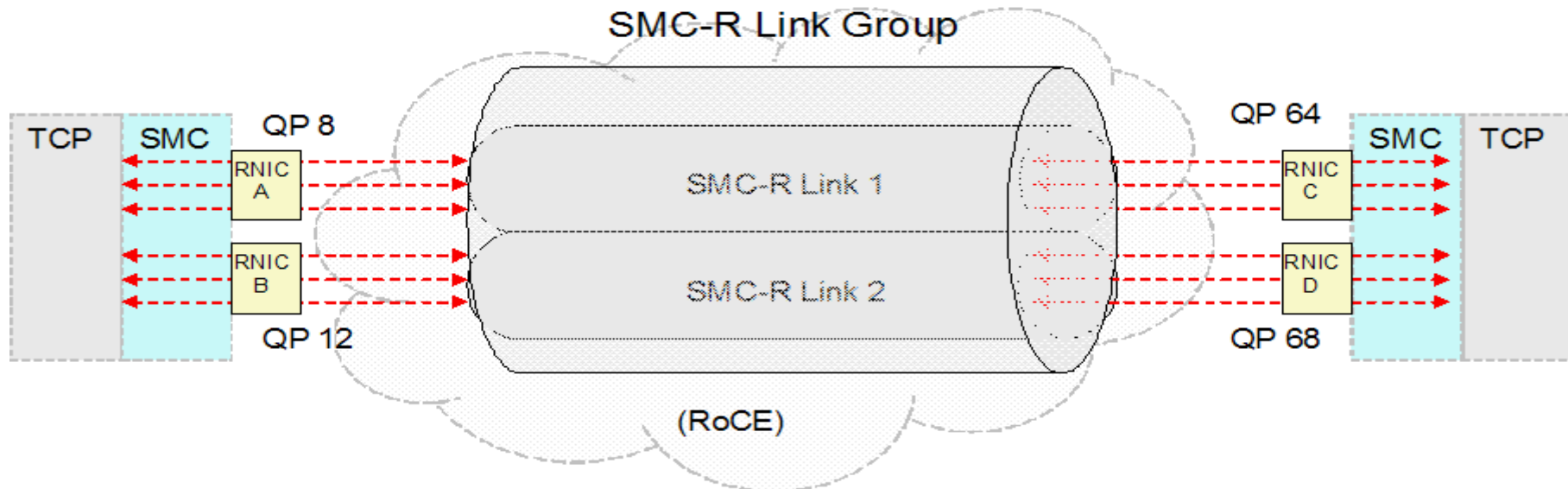
- SMC-R link group is a logical grouping of two SMC-R links between two RDMA peers
- Links within the link group are considered to be *equal*
  - The links have the same TCP server and TCP client roles
  - The links use the same VLAN, or do not use VLAN at all
  - The links have access to the same remote memory buffers (RMBs)
- Because the links are equal:
  - TCP connections can be assigned to either link
  - TCP connections can be moved from one link to the other
- SMC-R link remains active for 10 minutes after last TCP connection ends to save set-up costs

What are remote memory buffers (RMBs)?

- Remote memory buffers (RMBs) are fixed 64-bit memory used for receiving RDMA data from a peer
  - Each peer allocates memory that serves as the RMB for the remote peer
  - The sending peer's operating system places the TCP socket application data directly into the RMB
  - The receiving peer copies the data from the RMB into the TCP socket application's receive buffer

- The RMB is partitioned into different elements (RMBE)
  - All elements in a given RMB are the same size
  - Every TCP connection has its own separate RMBE

SMC-R link groups and RMBs

- Each SMC-R link within the link group has access to the RMB
  - RNIC adapter provides an RKEY to represent the physical storage
  - Multiple RKEYs can be assigned to the same storage
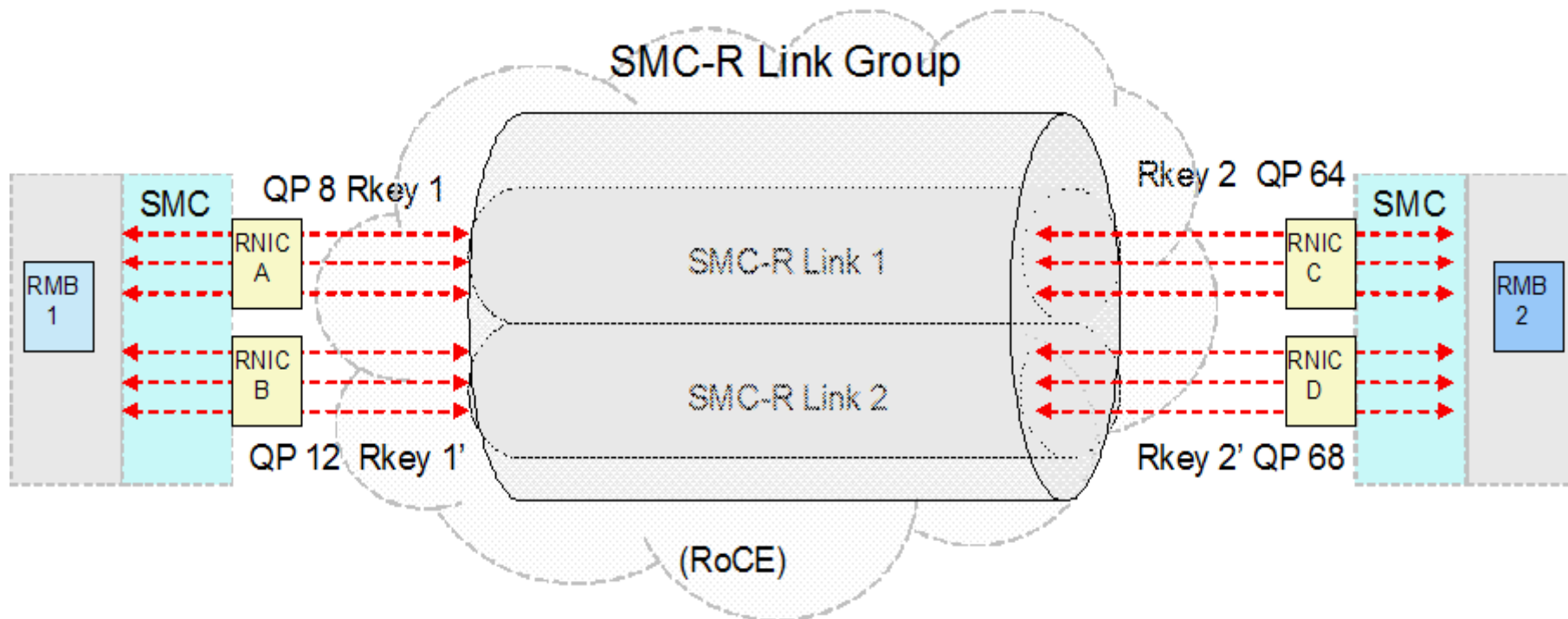
z/OS Comm Server RMB specifics

- z/OS Comm Server allocates RMBs in 1M increments
  - Three RMBs are allocated when SMC-R link group is created
  - Initial RMBs are partitioned into RMBEs when TCP connections are established and require an element

- RMB element size can range from 32K to greater than 256K
  - Based on the application buffer size specified on SETSOCKOPT( )
    - TCPCONFIG TCPRCVBUFRSIZE used if SETSOCKOPT( ) not performed

- Additional RMBs are allocated when all storage is used, or no RMBE of the proper size can be assigned from an existing RMB
  - RMBs with no RMBEs in use are freed, but at least three RMBs remain allocated

SMC-R link groups and multiple RMBs

▪ Each SMC-R link within the link group has access to **all** RMBs associated with the link group

▪ Peers can use different number of RMBs per link group
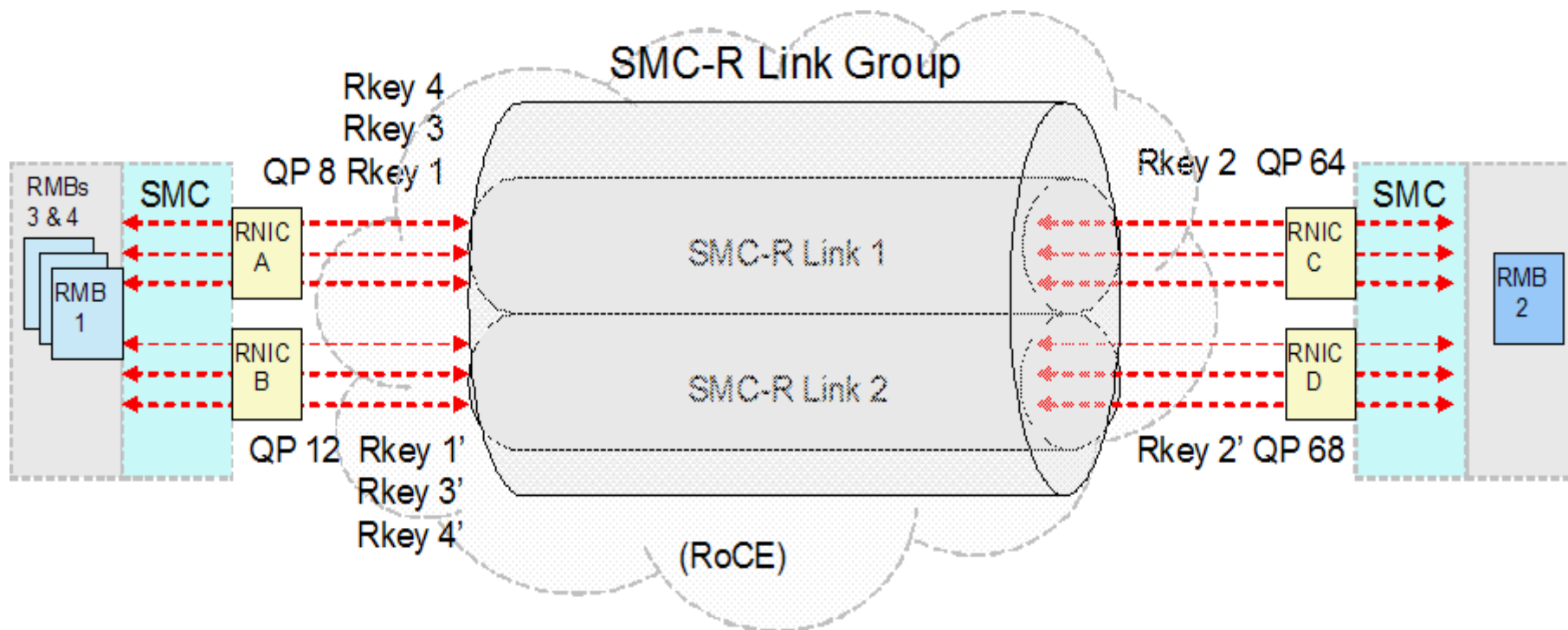
What are staging buffers?

- Staging buffers are fixed 64-bit memory used for sending RDMA data to a peer
  - Staging buffers are allocated on a per stack basis, and shared by all SMC-R link groups on this stack
  - Allocated in 1M increments
    - Stack allocates 4M of staging buffers when the first RNIC interface is started
    - Expansion and contraction of the number of buffers occurs based on volume of outbound data
  - Data is maintained in the staging buffer until RNIC adapter indicates that the data has been stored into the peer's RMB

SMC-R configuration considerations

- Physical network ID (PNet ID)

- Virtual LANs (VLANs)

- Redundancy

SMC-R Link Architecture (RC-QPs)
Multiple TCP connections per SMC Link

| TCP | SMC | | | SMC | TCP |

QP 8

QP 64

SMC Link

(RoCE)
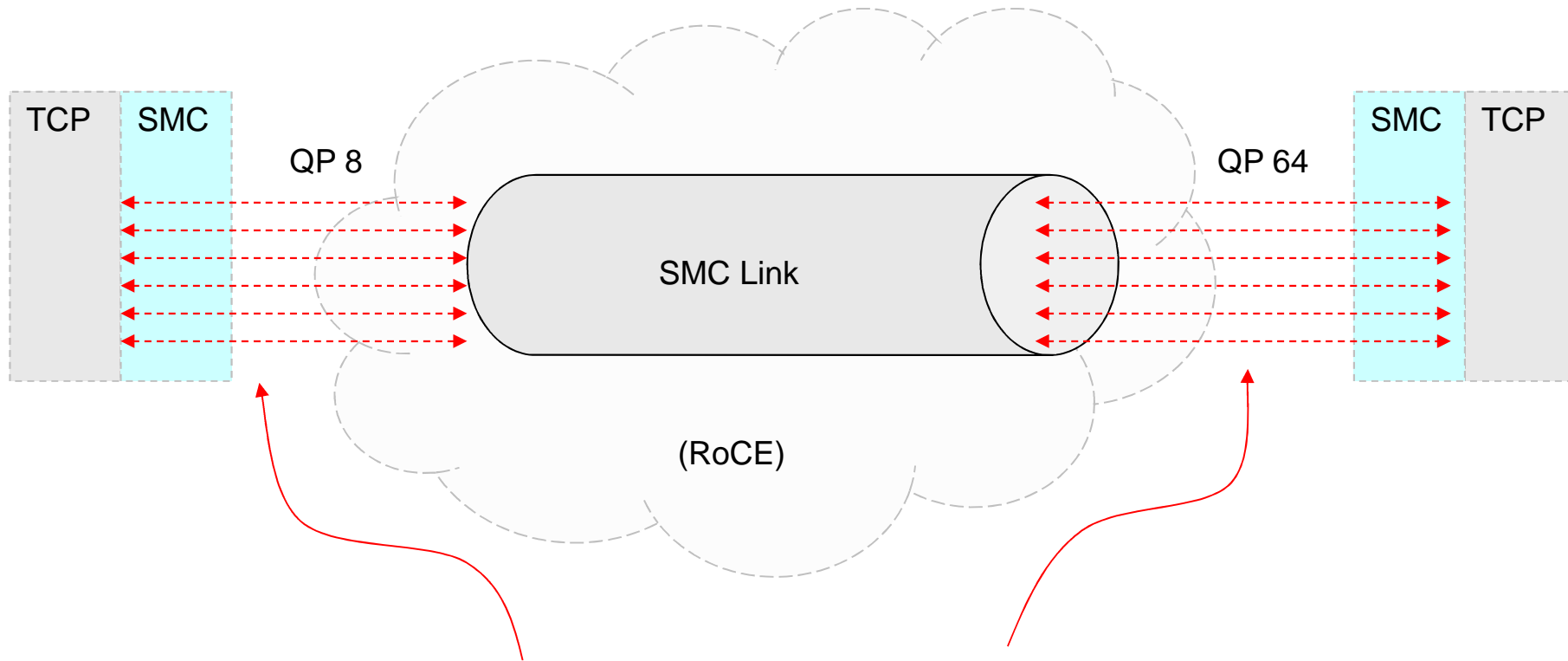
Multiple TCP (via SMC) Connections share the same SMC Link
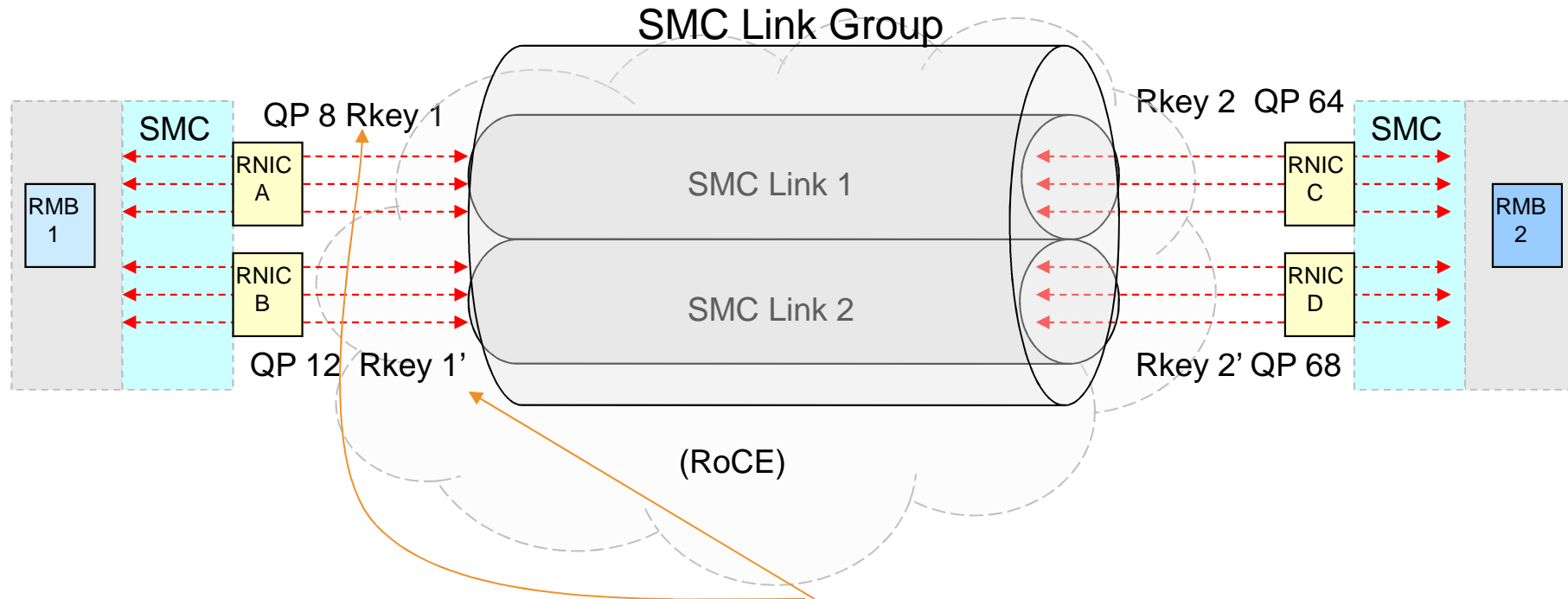
SMC-R Link Groups – Multiple SMC Links
(Provides resiliency, link level load balancing and additional bandwidth)
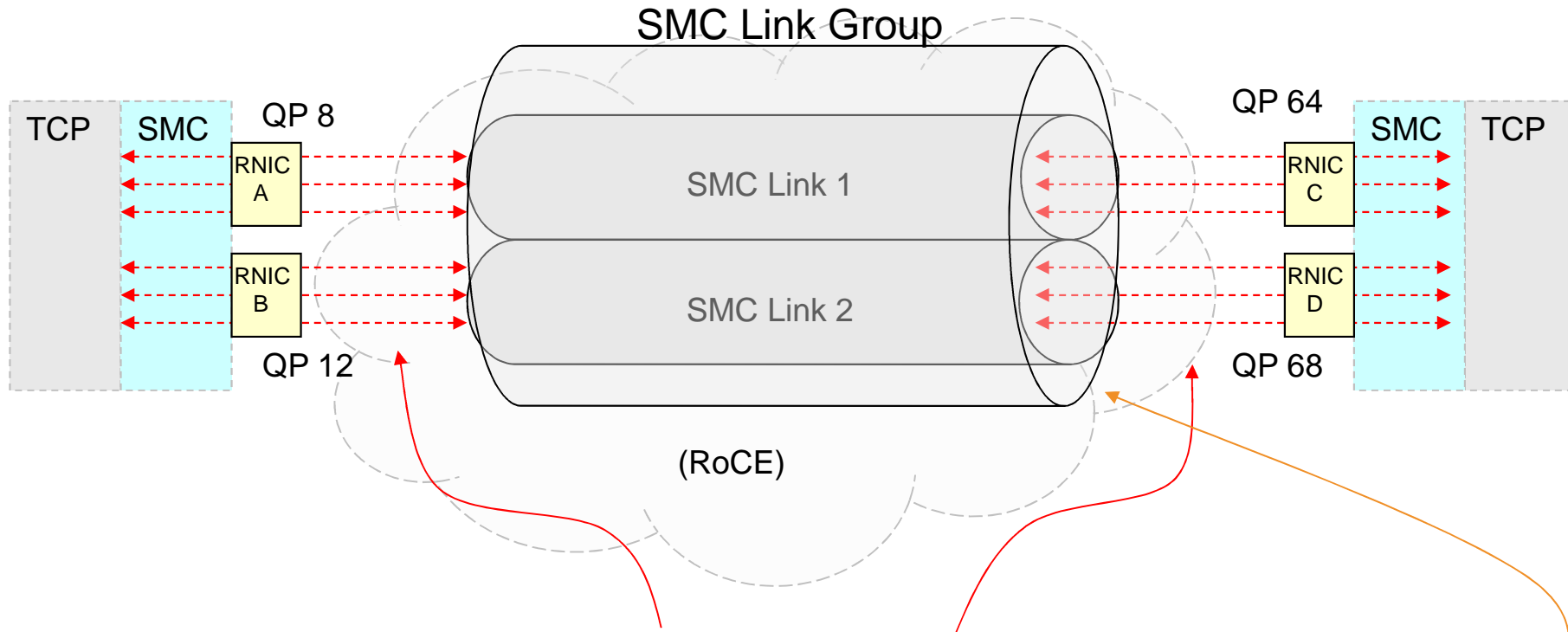
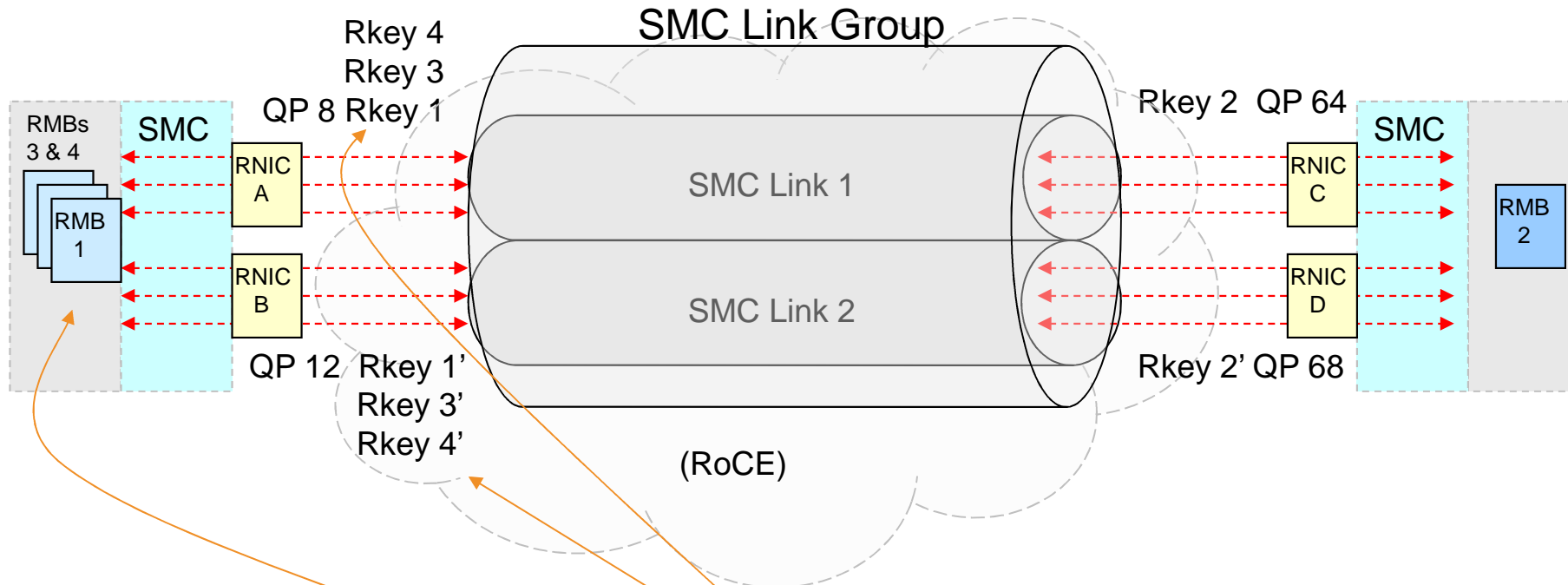TCP connections are balanced across multiple links within the Link Group

Multiple SMC Links across unique physical RNICs are grouped together to form a single SMC Link Group

SMC-R Memory Architecture (Part 1)

Each SMC Link has equal access (unique Rkey) to the peer's memory or RMB(s)

SMC-R Memory Architecture (Part 2)

SMC link groups also support multiple RMBs. Each peer can independently manage (add or remove) RMBs based on the needs of the link group, workload, and OS unique memory management requirements. Again, all SMC links continue to have equal access (Rkeys) to all RMBs.

SMC-R Memory Architecture (High Availability)
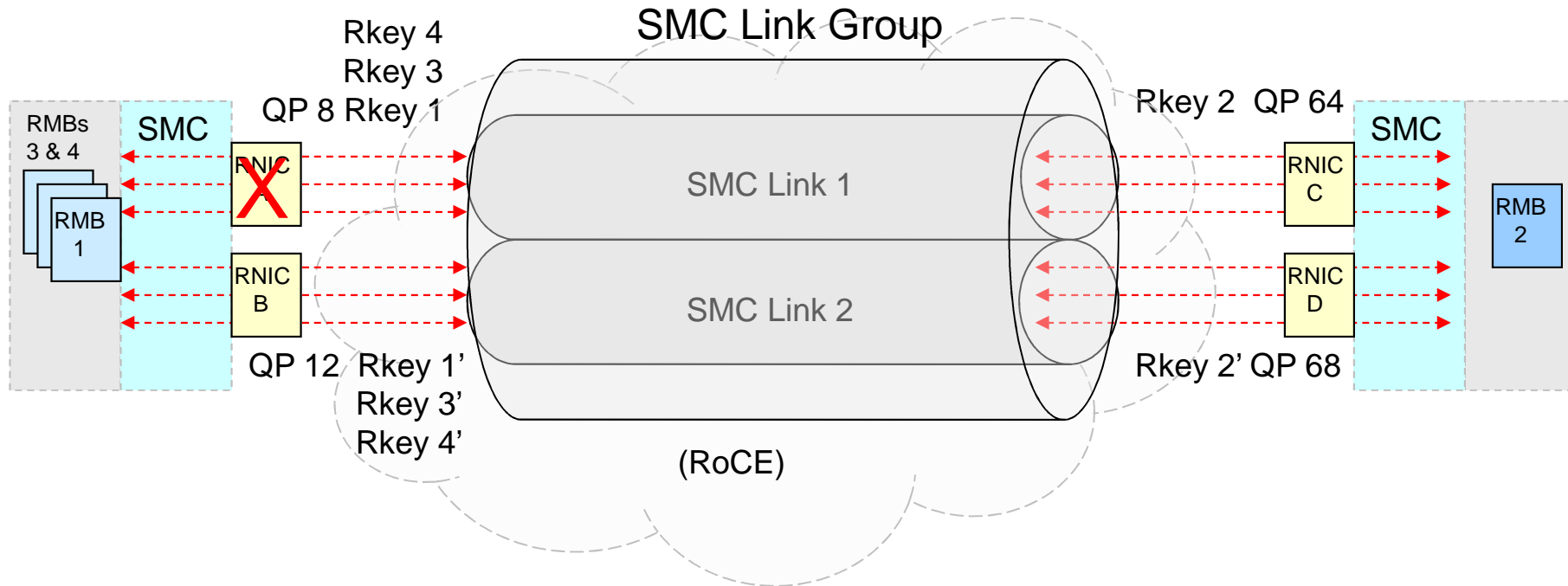
If one path (e.g. an RNIC) becomes unavailable (in this example RNIC A)… then:
• traffic on the SMC Link 1 is transparently moved to SMC Link 2 using the redundant hardware
• all application workload RDMA traffic continues without interruption…
  once SMC Link 1 is recovered then traffic can resume using both paths.

  Note that all paths (SMC Links) have equal access to all RMBs!

# Enabling SMC-R support in z/OS CommServer

- Specify GLOBALCONFIG SMCR parameter
    - Must specify at least one PCIe function ID (PFID) value
        - A PFID represents a specific RDMA network interface card (RNIC) adapter
        - Maximum of 16 PFID values can be coded
    - Up to eight TCP/IP stacks can share the same PFID in a given LPAR

- Start IPAQENET or IPAQENET6 INTERFACE with CHPIDTYPE OSD
    - SMC-R is enabled by default for these interface types
    - SMC-R is not supported on any other interface types

- SMC-R function is now enabled!

High-level SMC-R operations

- Start the first SMC-R capable OSD interface

  - All PFIDs are activated and grouped according to physical network

- Start TCP connection that traverses OSD interface
  - Rendezvous processing determines if TCP connection can use SMC-R
  - If necessary, SMC-R link and link group created

- Terminate last TCP connection that is using SMC-R link
  - SMC-R link remains active for 10 minutes to save setup costs

- Stop last SMC-R capable OSD interface
  - RNIC interfaces remain active

RNIC and OSD interaction

- RNIC activation is initiated as part of OSD interface activation
 - Assuming OSD defined using INTERFACE statement
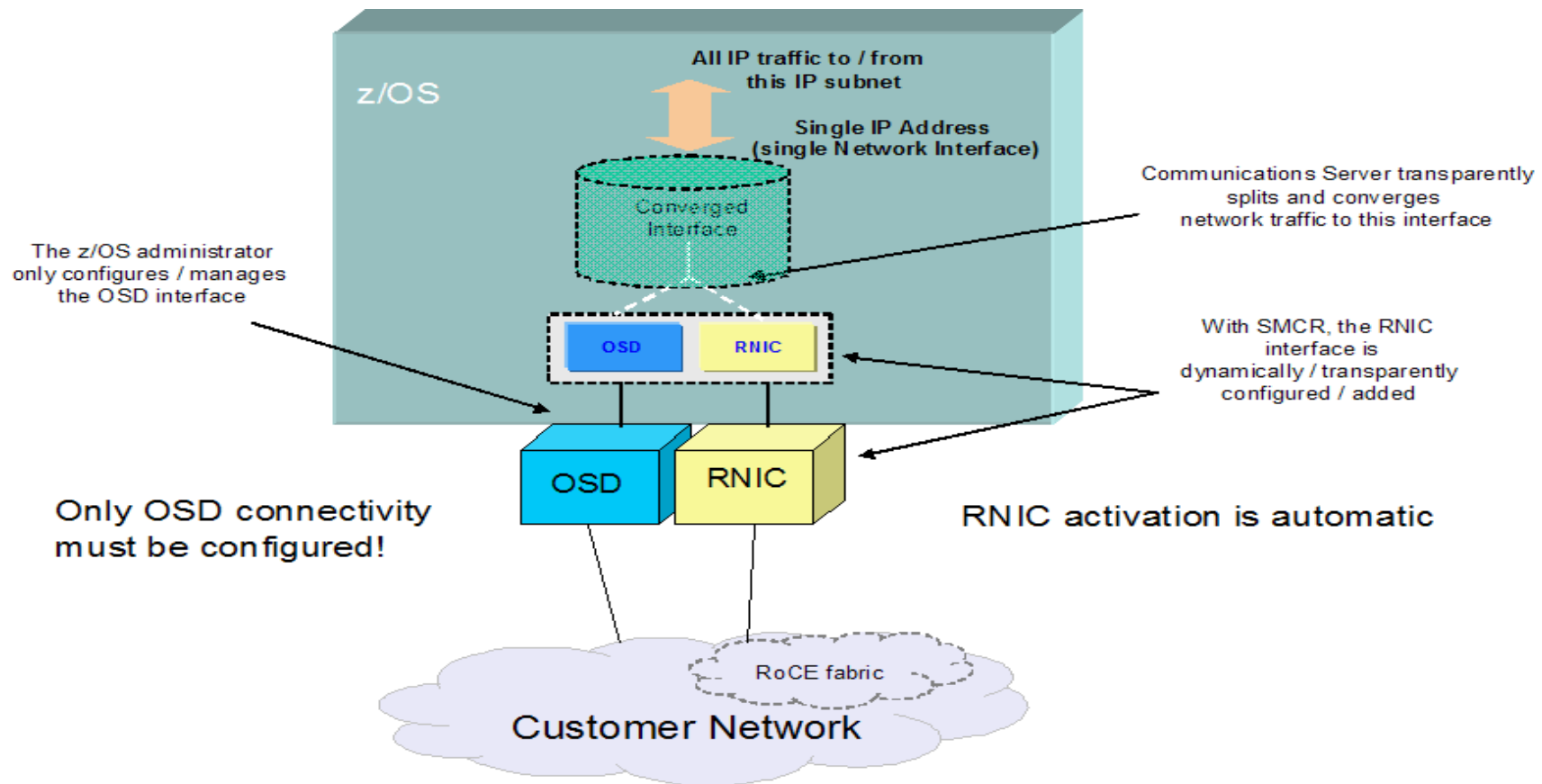
RNIC interface

- An RNIC interface is dynamically created for each PFID defined on the GLOBALCONFIG SMCR parameter
  - Created and activated when first SMC-R capable OSD interface is started
  - Associated VTAM TRLE is dynamically created as well

- Remains active even after all SMC-R capable OSD interfaces are stopped, unless manually stopped as well
  - Ideally, the operator should only need to manage the OSD interfaces
  - If RNIC interface is stopped by the operator, it must be manually restarted by the operator before it is used again
    - Starting an SMC-R capable OSD interface has no effect here

Physical network (PNet) ID concepts

- Customer-defined value for logically grouping OSD interfaces and RNIC adapters based on physical connectivity
  - Customer defines PNet ID values for both OSA and RNIC interfaces in HCD
  - z/OS CommServer gets the information dynamically
    - Learns the definitions during activation of the interfaces
    - Associates the OSD interfaces with the RNIC interfaces that have matching PNet ID values

- If you do not configure a PNet ID for the RNIC adapter, activation fails

- If you do not configure a PNet ID for the OSA adapter, activation succeeds, but the interface is not eligible to use SMC-R

Physical network ID example

- Three physically separate networks defined by customer
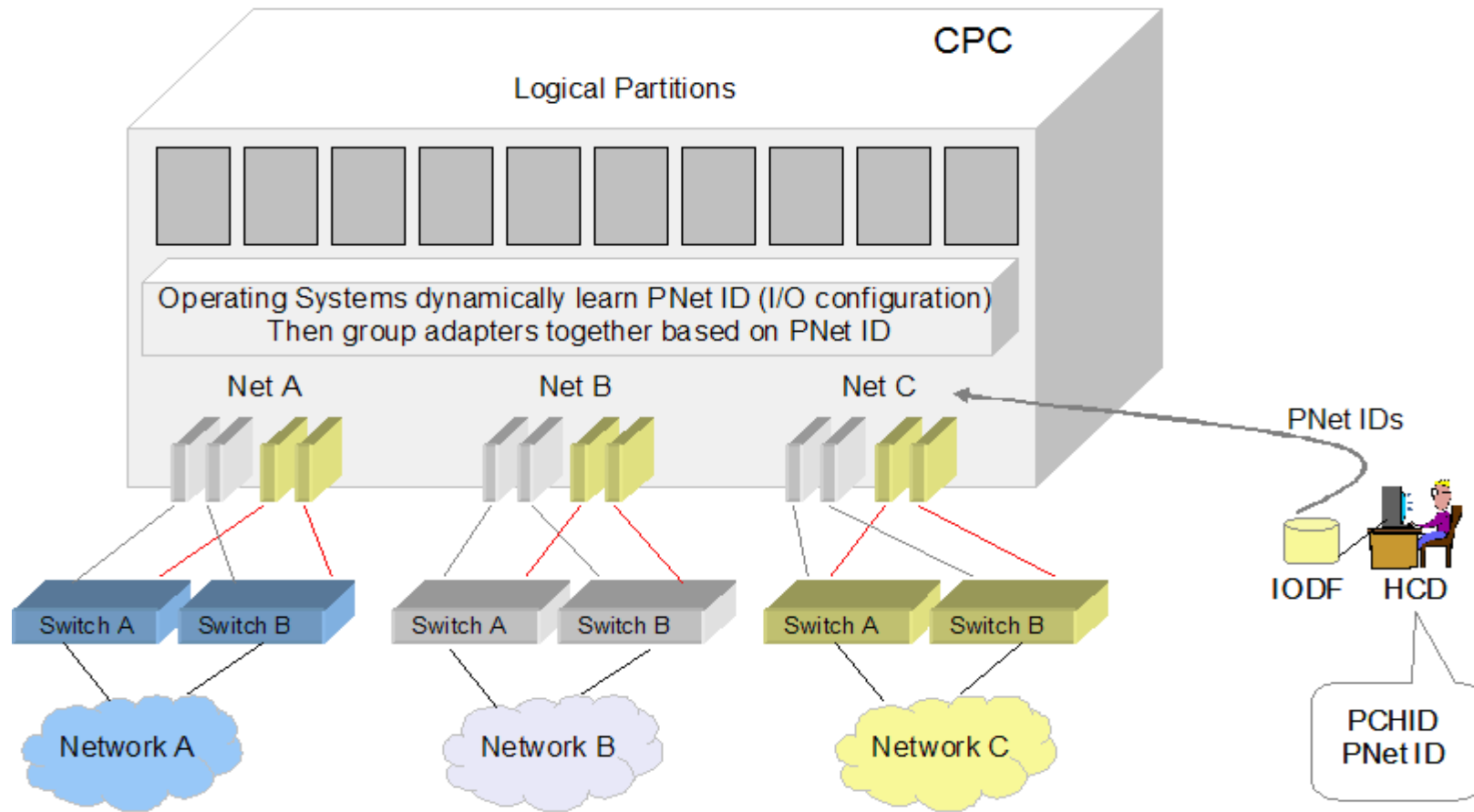
PNet ID example, configuration

- Define PFIDs and PNet ID values in HCD

- Define PFIDs and OSD INTERFACEs in TCP/IP profile

	TCP/IP Configuration		VTAM Configuration

GLOBALCONFIG

PFID 100 PortNum 1

PFID 200 PortNum 1

PFID 300

PFID 400

PFID 500 PortNum 1

PFID 600 PortNum 2

INTERFACE 1 OSD
PortName 1 SMCR

INTERFACE 2 OSD
PortName 2 SMCR

INTERFACE 3 OSD
PortName 3 SMCR

INTERFACE 4 OSD
PortName 4 SMCR

TRLE 1 OSD
PortName 1

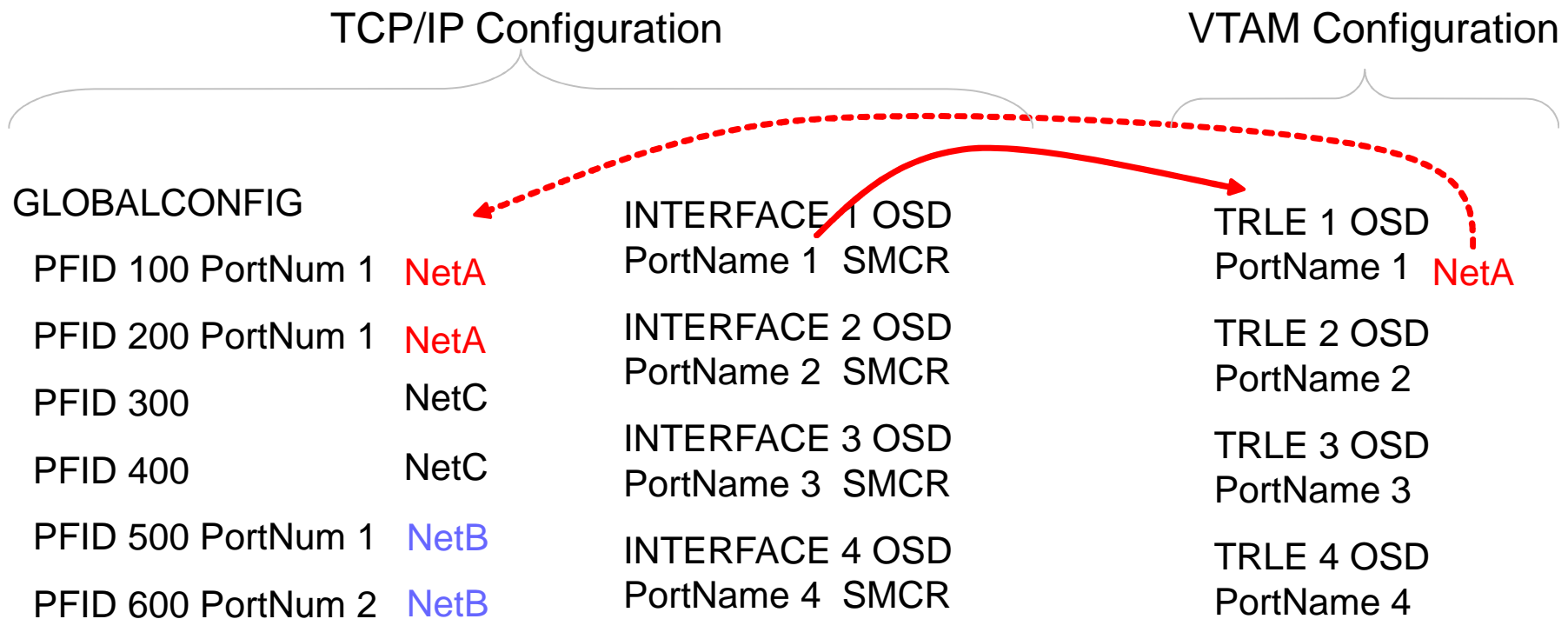TRLE 2 OSD
PortName 2

TRLE 3 OSD
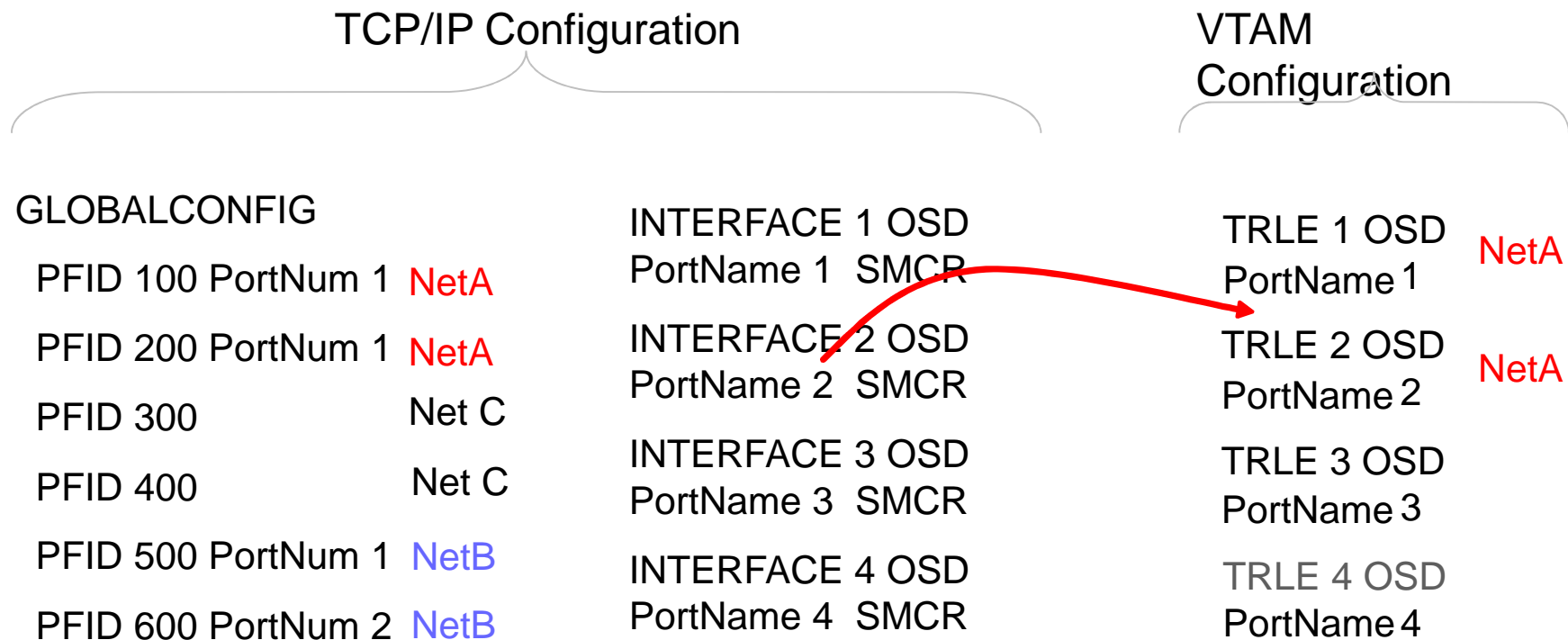PortName 3

TRLE 4 OSD
PortName 4

PNet ID example, activate first OSD

- Activation of first SMC-R capable OSD starts all RNIC interfaces
  - PNet ID values discovered for OSD and all PFIDs

TCP/IP Configuration   VTAM Configuration

GLOBALCONFIG

PFID 100 PortNum 1 NetA

PFID 200 PortNum 1 NetA

PFID 300   NetC

PFID 400   NetC

PFID 500 PortNum 1 NetB

PFID 600 PortNum 2 NetB

INTERFACE 1 OSD
PortName 1 SMCR

INTERFACE 2 OSD
PortName 2 SMCR

INTERFACE 3 OSD
PortName 3 SMCR

INTERFACE 4 OSD
PortName 4 SMCR

TRLE 1 OSD
PortName 1 NetA
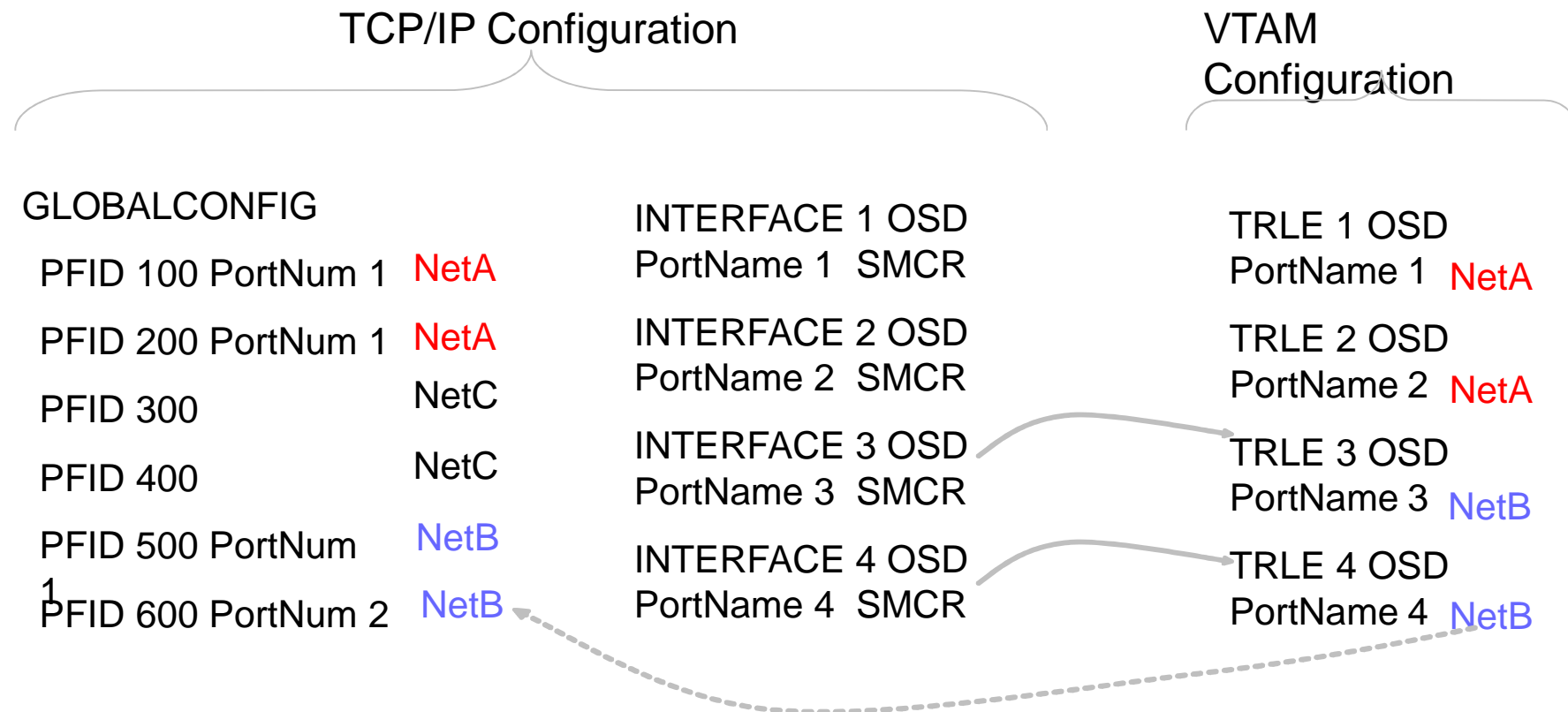
TRLE 2 OSD
PortName 2

TRLE 3 OSD
PortName 3

TRLE 4 OSD
PortName 4

PNet ID example, activate second OSD

- Second OSD has same PNet ID, so just associate with same set of PFIDs as INTERFACE 1

TCP/IP Configuration | VTAM Configuration

GLOBALCONFIG

PFID 100 PortNum 1 NetA

PFID 200 PortNum 1 NetA

PFID 300   Net C

PFID 400   Net C

PFID 500 PortNum 1 NetB

PFID 600 PortNum 2 NetB

INTERFACE 1 OSD
PortName 1 SMCR

INTERFACE 2 OSD
PortName 2 SMCR

INTERFACE 3 OSD
PortName 3 SMCR

INTERFACE 4 OSD
PortName 4 SMCR

TRLE 1 OSD
PortName 1   NetA

TRLE 2 OSD
PortName 2   NetA

TRLE 3 OSD
PortName 3

TRLE 4 OSD
PortName 4

PNet ID example, second PNet ID

- Subsequent OSD interfaces were assigned different PNet ID, so they are associated with different set of PFIDs

TCP/IP Configuration

VTAM Configuration

GLOBALCONFIG

PFID 100 PortNum 1 NetA

PFID 200 PortNum 1 NetA

PFID 300   NetC

PFID 400   NetC

PFID 500 PortNum 1 NetB

PFID 600 PortNum 2 NetB

INTERFACE 1 OSD
PortName 1 SMCR

INTERFACE 2 OSD
PortName 2 SMCR

INTERFACE 3 OSD
PortName 3 SMCR

INTERFACE 4 OSD
PortName 4 SMCR

TRLE 1 OSD
PortName 1 NetA

TRLE 2 OSD
PortName 2 NetA

TRLE 3 OSD
PortName 3 NetB

TRLE 4 OSD
PortName 4 NetB

SMC-R and VLAN Configuration Rules

- The following rules apply when using VLANs with SMC-R:
  1. The Ethernet switch port VLAN mode must be consistent between the OSA Express Ethernet ports and their associated RoCE Express RDMA ports
     - If the OSA Express Ethernet switch ports are configured in trunk mode, their associated RoCE Express RDMA switch ports must also be configured in trunk mode
     - If the OSA Express Ethernet switch ports are configured in access mode, their associated RoCE Express RDMA switch ports must also be configured in access mode
  3. The VLAN mode must be consistent between all of the hosts that will communicate over a LAN fabric (PNET) using SMC-R
     - You can't mix access and trunk modes among hosts on the same PNET if you are using SMC-R
  4. The RoCE Express features must be on the same VLAN to communicate
     - If you are using **access mode**, the switch ports that are serving the RoCE Express features on a PNET must all be configured with the same VLAN ID. The RoCE VLAN ID is not required to match the VLAN ID of associated OSA Express features.
     - If you are using **trunk mode**, the RoCE Express features switch ports must be configured to allow the same VLAN IDs as the OSA Express features that they are associated with

- For more details on SMC-R and VLANs refer to the following:
  ftp://public.dhe.ibm.com/software/os/systemz/pdf/SMCR_RoCE_VLAN_Requirements_25sept14.pdf

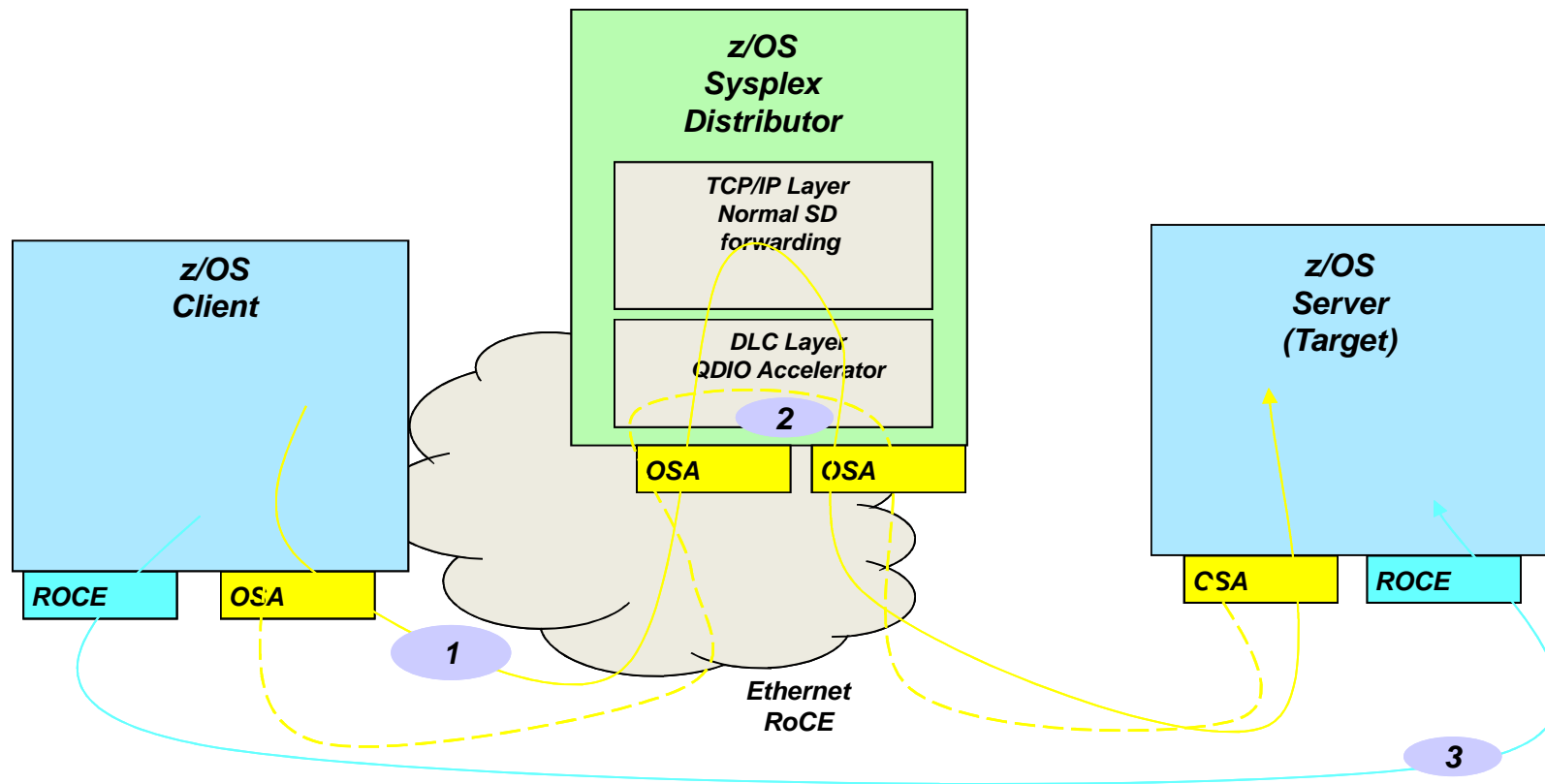# SMC-R interactions with TCP/IP functions

- Sysplex Distributor

- Security functions
  - Application Transparent Transport Layer Security (AT-TLS)
  - Intrusion Detection Services (IDS)
  - IP Security (IPSec)
  - Multilevel Security (MLS)

- TCP application sockets compatibility

- Fast Response Cache Accelerator (FRCA)
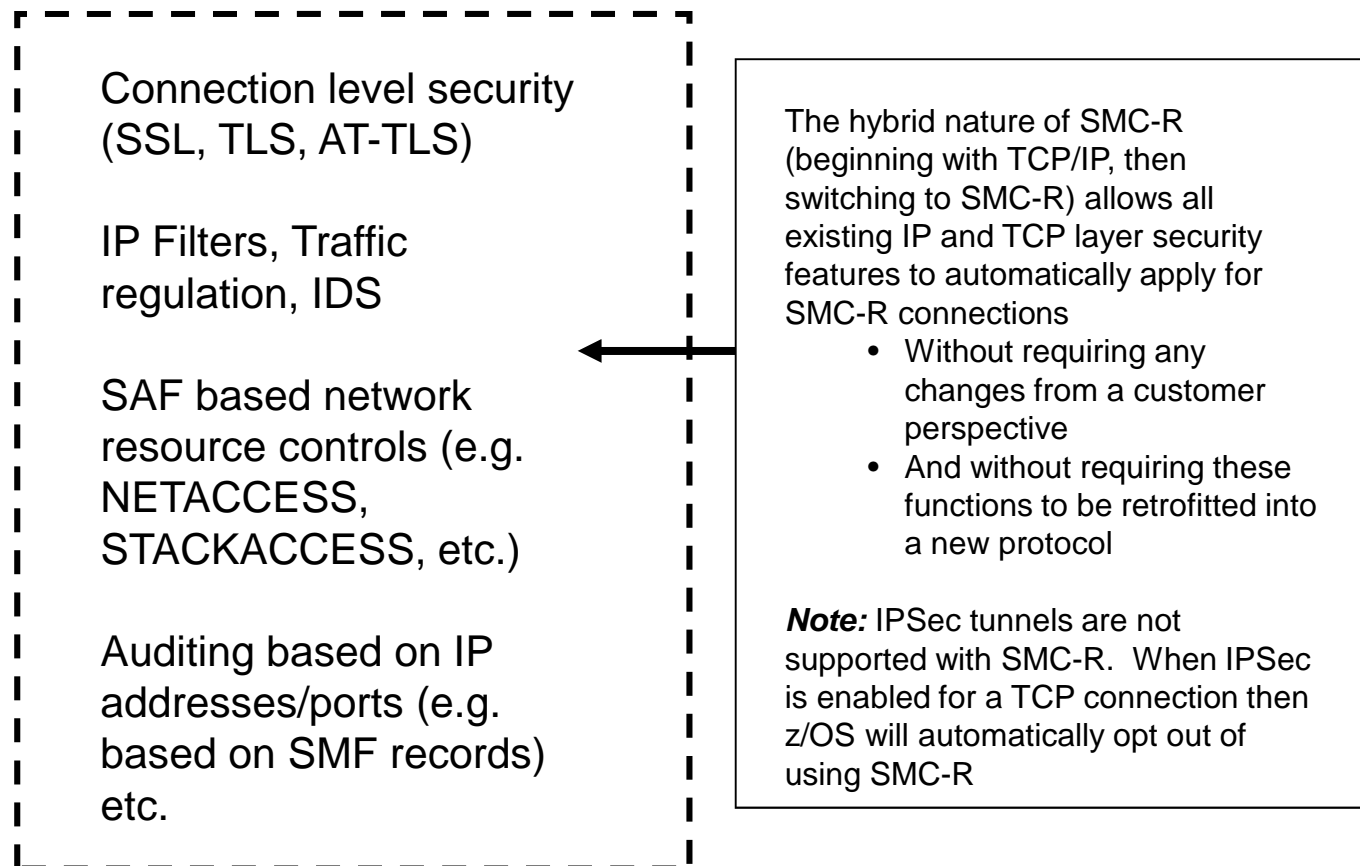
Sysplex Distributor and SMC-R

- Sysplex Distributor can be deployed with SMC-R with no additional configuration updates
  - *If the client application resides on z/OS host* that is enabled for SMC-R and meets SMC-R criteria for connecting to the target z/OS system
    - Note: Sysplex Distributor clients not running on z/OS platform continue to work using normal TCP/IP flows even if SMC-R is enabled on z/OS
  - TCP connections set up using rendezvous processing
  - Once TCP connection is established, rendezvous processing determines correct SMC-R link to use with server application
  - Application data does not flow through the Sysplex Distributor at all after SMC-R link is created or selected
    - Represents a performance improvement compared to normal Sysplex Distributor flows

Sysplex Distributor example with SMC-R

- SMC-R link establishment and socket traffic completely bypasses the Sysplex Distributor stack and DLC layers

SMC-R preserves existing security model

Connection level security (SSL, TLS, AT-TLS)

IP Filters, Traffic regulation, IDS

SAF based network resource controls (e.g. NETACCESS, STACKACCESS, etc.)

Auditing based on IP addresses/ports (e.g. based on SMF records) etc.

The hybrid nature of SMC-R (beginning with TCP/IP, then switching to SMC-R) allows all existing IP and TCP layer security features to automatically apply for SMC-R connections
- Without requiring any changes from a customer perspective
- And without requiring these functions to be retrofitted into a new protocol

Note: IPSec tunnels are not supported with SMC-R. When IPSec is enabled for a TCP connection then z/OS will automatically opt out of using SMC-R

AT-TLS and SMC-R

- AT-TLS can be deployed with SMC-R
    - AT-TLS negotiations take place after rendezvous exchange
        - Negotiation takes place over SMC-R link
    - Encryption and decryption of data occurs normally

- Same is true for applications/middleware implementing TLS or SSL directly

Intrusion Detection Services and SMC-R

- IDS functions that involve checks during TCP connection set up have no special interaction with SMC-R (i.e. these all apply to the TCP connections that end up using SMC-R)
  - Scan detection and reporting
  - Traffic regulation of TCP connections

- Detection, reporting and prevention of attacks related to socket data apply to TCP connections using SMC-R
  - TCP constrained queue events
    - Normal constrained conditions apply
    - TCP connection also considered to be constrained when data was stored into peer RMB but was not acknowledged for 30 seconds
  - Global TCP stall events

Solution: Security functions that cannot exploit SMC-R

- Other security functions do not always interoperate with SMC-R
  - Might require TCP/IP to examine TCP packet data, but SMC-R does not convert the application data into TCP packets
    - IPSec when tunneling is required
    - IPSec filters when the filter denies a packet for a TCP connection
    - MLS when packet tagging is required
  - In these scenarios, z/OS will opt out of using SMC-R for the affected TCP connections
  - If functions are activated dynamically, affected TCP connections that are using SMC-R are terminated

TCP application sockets capability and FRCA

- SMC-R protocol is intended to be transparent to, and fully compatible with, TCP socket applications

  – Support for all Socket APIs except for PASCAL sockets

- Use of SMC-R should not impact application use of socket API functions such as:
  – MSGWAITALL
  – MSG_PEEK
  – Urgent Data
  – Accept and Receive (ANR)

- SMC-R cannot be used with Fast Response Cache Accelerator (FRCA)

  – TCP/IP automatically opts out of using SMC-R

SMC-R Key Attributes - Summary

✓ Optimized Network Performance (leveraging RDMA technology)

✓ Transparent to (TCP socket based) application software

✓ Leverages existing 10GbE technology (RoCE)

✓ Preserves existing network security model

✓ Resiliency (dynamic failover to redundant hardware)

✓ Transparent to Load Balancers

✓ Preserves existing IP topology and network administrative and operational model

SMC-R References

- **SMC-R One Stop Shopping Web Page (Includes latest links to ALL other SMC-R References):**

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

IBM
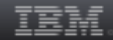
This page provides a comprehensive list of reference material related to SMC-R:

- SMC-R Overview
- SMC-R Overview (with audio) (01:09:07)
- SMC-R Implementation
- SMC-R Implementation (with audio) (00:58:04)
- Shared Memory Communications over RDMA: Performance Considerations (white paper)
- SMC-R performance information
- SMC-R FAQ
- Diagnosing Problems with SMC-R
- SMC-R and Security Considerations (white paper)
- SMC-R informational RFC
- SMC-R performance over distance (894KB)
- SMC-R VLAN configuration considerations (1.1MB)
- Linux SMC-R Overview (Note: The IBM SMC-R Linux source code is pending evaluation by the Linux open source community) (397KB)
- SMC-R Applicability Tool (SMCAT) Overview (429KB)
- SMC-R and IBM System z13 10GbE RoCE Express Virtualization (SR-IOV) Overview (683KB)

THANK YOU

Please complete your session evaluation

- z/OSOS Communications Server: Shared Memory Communications over RDMA (SMC-R) Protocol
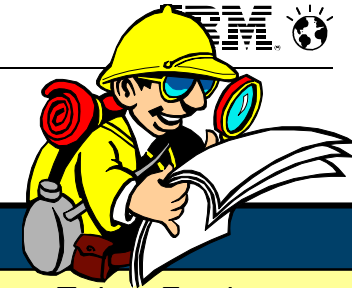
- Session # 17285

- QR Code:

IBM Comm Server

Find us on Facebook at
http://www.facebook.com/IBMCommserver

Follow us on Twitter at
http://www.twitter.com/IBM_Commserver

Read the z/OS Communications Server blog at
http://tinyurl.com/zoscsblog

Visit the z/OS CS YouTube channel at
http://www.youtube.com/user/zOSCommServer

For more information

URL	Content
http://www.twitter.com/IBM_Commserver	IBM z/OS Communications Server Twitter Feed
http://www.facebook.com/IBMCommserver	IBM z/OS Communications Server Facebook Page
https://www.ibm.com/developerworks/mydeveloperworks/blogs/IBMCommserver/?lang=en	IBM z/OS Communications Server Blog
http://www.ibm.com/systems/z/	IBM System z in general
http://www.ibm.com/systems/z/hardware/networking/	IBM Mainframe System z networking
http://www.ibm.com/software/network/commserver/	IBM Software Communications Server products
http://www.ibm.com/software/network/commserver/zos/	IBM z/OS Communications Server
http://www.redbooks.ibm.com	ITSO Redbooks
http://www.ibm.com/software/network/commserver/zos/support/	IBM z/OS Communications Server technical Support – including TechNotes from service
http://www.ibm.com/support/techdocs/atsmastr.nsf/Web/TechDocs	Technical support documentation from Washington Systems Center (techdocs, flashes, presentations, white papers, etc.)
http://www.rfc-editor.org/rfcsearch.html	Request For Comments (RFC)
http://www.ibm.com/systems/z/os/zos/bkserv/	IBM z/OS Internet library – PDF files of all z/OS manuals including Communications Server
http://www.ibm.com/developerworks/rfe/?PROD_ID=498	RFE Community for z/OS Communications Server
https://www.ibm.com/developerworks/rfe/execute?use_case=tutorials	RFE Community Tutorials

For pleasant reading ….

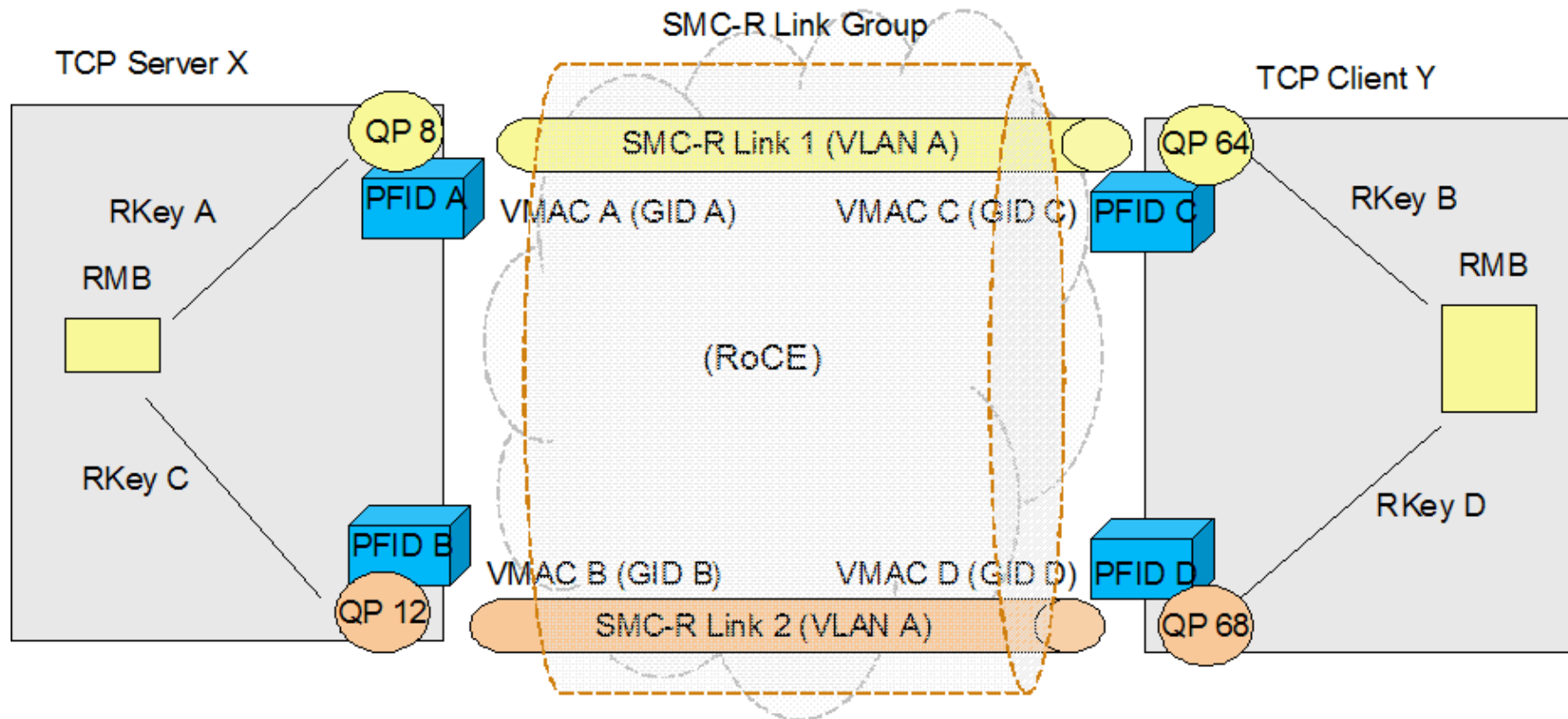# Backup charts on SMC-R and 10GbE RoCE Express

Redundancy

- SMC-R link groups provide for load balancing and recovery
  - New TCP connection is assigned to the SMC-R link with the fewest TCP connections
  - Load balancing only performed when multiple RNIC adapters are available at each peer

- Full redundancy requires:
  - Two or more RNIC adapters at each peer
  - Unique system internal paths for the RNIC adapters
  - Unique physical RoCE switches

- Partial redundancy still possible in the absence of one or more of these conditions

Redundancy levels

- Various levels of redundancy possible
  - Full redundancy
  - Partial redundancy
    - Partial redundancy only at the remote host
    - Partial redundancy only at the local host
    - Partial redundancy due to non-unique local internal path
  - No redundancy (single remote and local RNIC adapters)

Full redundancy example

- Full failover capability exists at both server and client
 - Recommended configuration

SMC-R Link Group

TCP Server X

TCP Client Y

QP 8   SMC-R Link 1 (VLAN A)   QP 64

RKey A   PFID A   VMAC A (GID A)   VMAC C (GID C)   PFID C   RKey B

RMB   RMB

(RoCE)

RKey C   RKey D

PFID B   VMAC B (GID B)   VMAC D (GID D)   PFID D

QP 12   SMC-R Link 2 (VLAN A)   QP 68

PCIe function internal path

- Full redundancy requires external and internal redundancy
  - "External redundancy" requires multiple RNIC adapters and unique RoCE switches
  - "Internal redundancy" requires unique PCIe support structures
    - Support partitions
    - System/z I/O drawers

- z/OS CommServer discovers **local** "internal redundancy" level during RNIC adapter activations
  - Referred to as PCIe function internal path (PFIP)

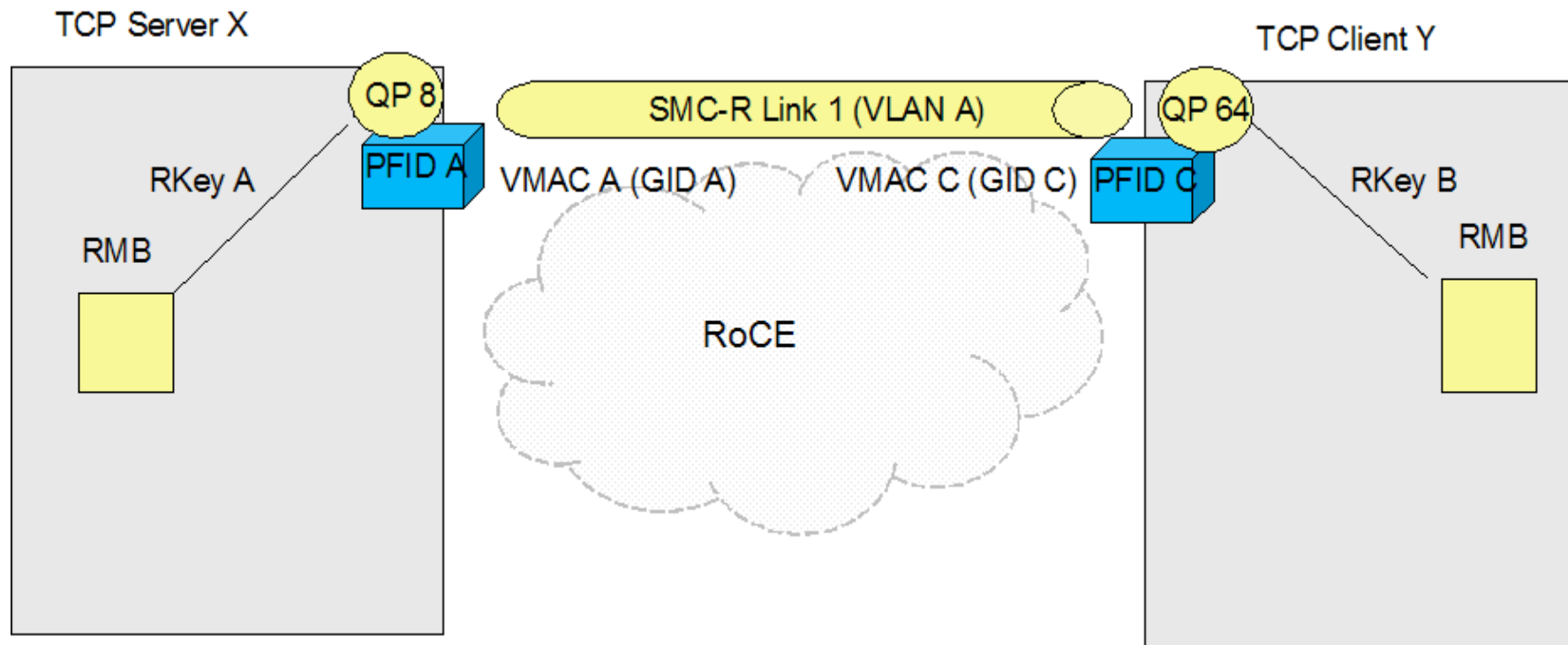- z/OS CommServer does not learn the **remote** "internal redundancy" level

Partial redundancy example

- Failover recovery is possible at the TCP server, but not at the TCP client
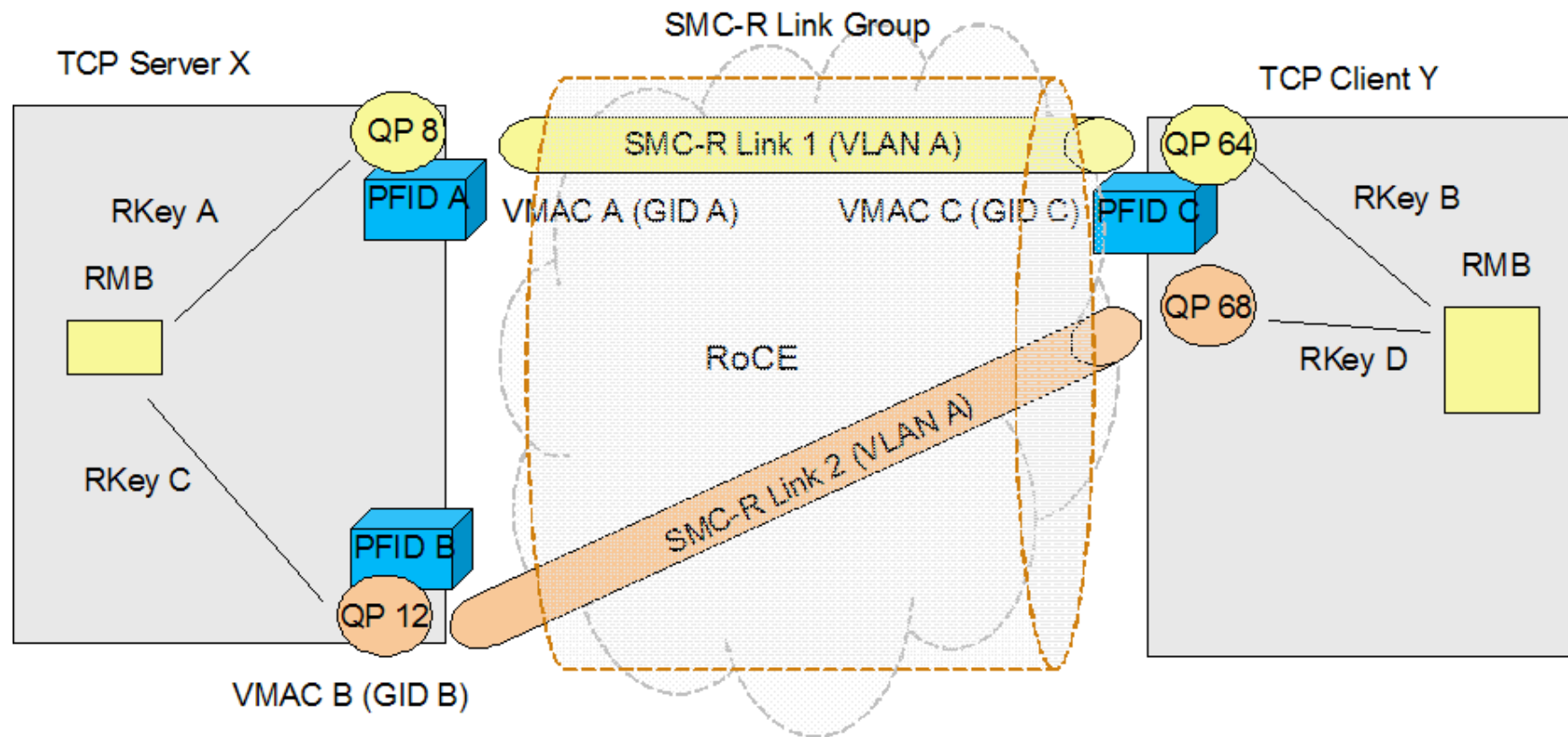
No redundancy example

- No failover capability for either the server or the client in this configuration, since only one RNIC adapter available at each peer
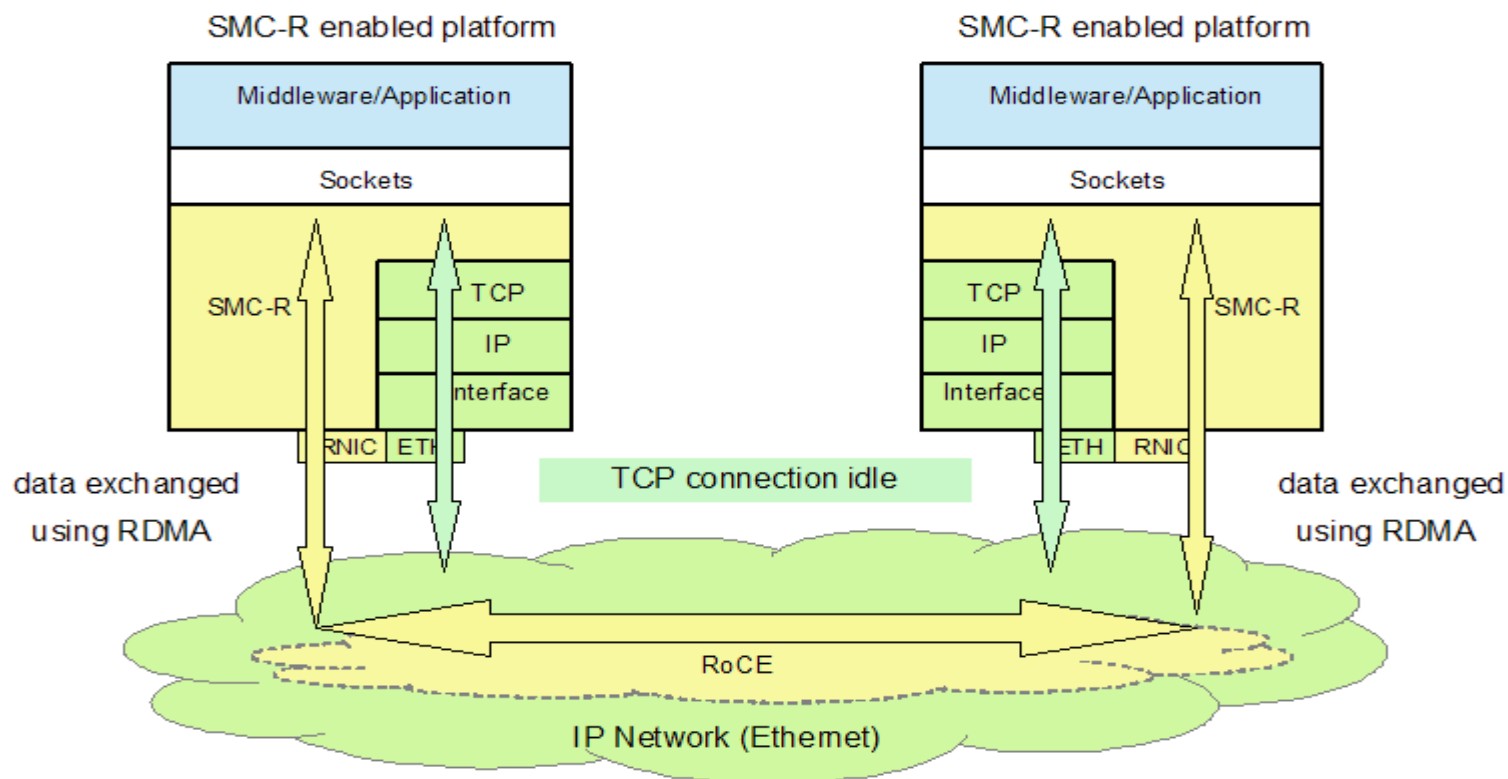
TCP keepalive and SMC-R

- Load balancers or firewalls use data traffic as an indication that a TCP connection is healthy
  - Might terminate the connection if no data flows within a certain period of time

- TCP keepalive processing periodically sends a packet over existing TCP connections
  - Application indicates connection is eligible for keepalive by specifying the SO_KEEPALIVE setsockopt( ) option
  - Time interval to use is determined by these criteria:
    - TCP_KEEPALIVE setsockopt( ) option, if specified
    - TCPCONFIG INTERVAL value, or default

TCP connections using SMC-R appear idle

- All application data flows "out-of-band" with SMC-R

- TCP connection is maintained, but just for control purposes

SMC-R keepalive processing

- By definition, TCP connections using SMC-R can look idle to the IP network
  - Sending repeated keepalive packets might generate excessive and unwanted traffic
  - Ensuring that the SMC-R link is still active fits better with the intent of keepalive processing

- SMC-R keepalive processing handles both the SMC-R link and the TCP connection
  - Existing keepalive settings determine time interval for SMC-R link keepalive probes
  - New algorithm for determining time interval for TCP connection keepalive probes

Keepalive algorithm for TCP connections using SMC-R

- TCP connection probe interval set to the largest of three values:
  - TCP_KEEPALIVE setsockopt( )
  - TCPCONFIG INTERVAL
  - GLOBALCONFIG SMCR TCPKEEPMININTERVAL
    - New optional configuration statement
    - Defaults to five minutes

- SMC-R link probe interval set based on existing keepalive algorithm

- Application must still specify SO_KEEPALIVE to enable processing for the TCP connection

SMC-R keepalive example

- Assume these values have been specified:
  - Application specifies SO_KEEPALIVE and TCP_KEEPALIVE setsockopt( ) as 5 minutes
  - TCPCONFIG INTERVAL set to 10 minutes
  - GLOBALCONFIG SMCR TCPKEEP set to 25 minutes

- For TCP connections that use SMC-R:
  - TCP connection probes sent every 25 minutes
  - SMC-R link probes sent every 5 minutes

- For TCP connections that do not use SMC-R:
  - TCP connection probes sent every 5 minutes

VLAN interaction with SMC-R

- OSD interfaces support VLAN usage, but it is not required

- VLAN mode propogated to RNIC interfaces associated with the OSD interfaces
  - RNIC adapters can operate in VLAN or "no-VLAN" mode
  - SMC-R communications use or do not use VLAN depending on the OSD definitions

- RNIC adapter supports VLAN IDs in the range of 1 to 4094

Physical networks, VLANs, and subnets

- VLANs, or subnets, can still be used to achieve logical separation of the physical network
  - A given physical network can include multiple VLANs or subnets
  - Each VLAN or subnet must be part of a single physical network

- z/OS CommServer cannot enforce the logical separation
  - Assumes that you have correctly configured the networks

- IPv4 OSD interfaces MUST have subnet mask configured
  - Can be started, but is not eligible for SMC-R

- At least one IPv6 address between peers MUST have a prefix value in order to use SMC-R

Setting the VLAN mode for an RNIC adapter

- RNIC adapters can be shared by up to eight TCP/IP stacks

- The VLAN mode of the RNIC is defined by the VLAN mode defined for the first OSD interface that is started
  - Applies across all stacks sharing the RNIC
  - SMC-R capable OSD interfaces with different VLAN modes start, but cannot use this RNIC interface
    - Can use other RNIC interfaces that have the proper VLAN mode

- **Use the same VLAN mode for all SMC-R capable OSD interfaces accessing the same physical network**

SMC-R failover example

- Application traffic switches transparently to second link after RNIC failure
  - Partial redundancy might be available for one host

- SMC-R link is recovered after RNICs are restarted, but TCP connections are not switched back
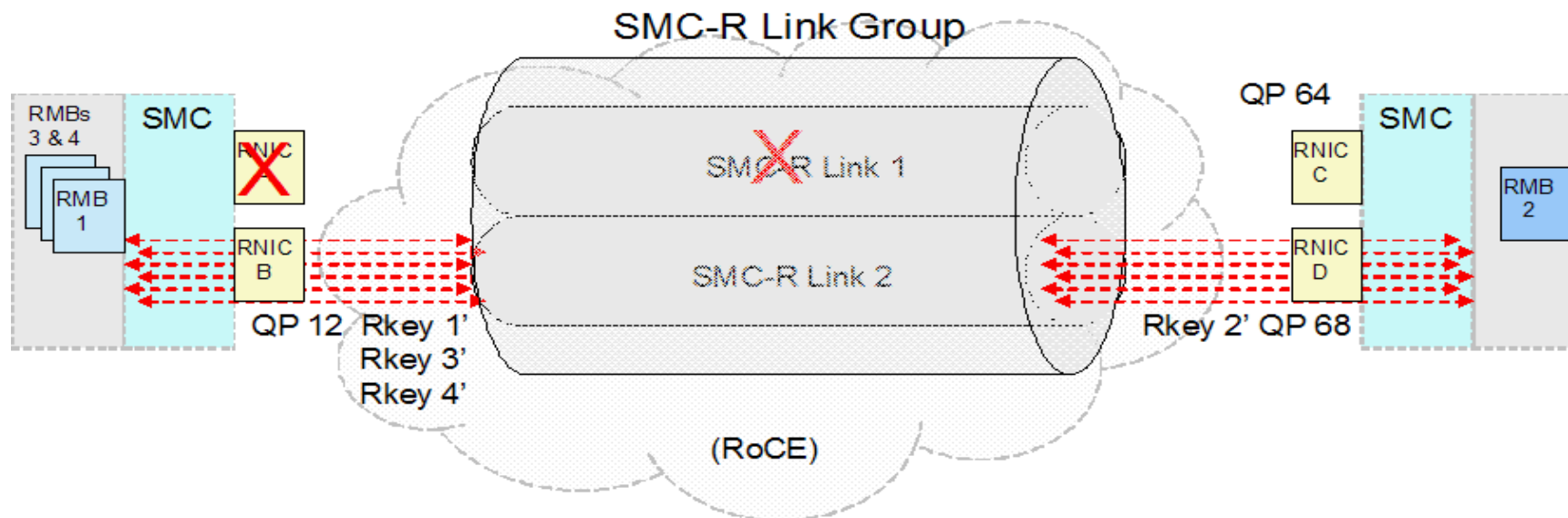
Configuring more than two RNIC adapters

- More than two PFIDs can be configured with the same PNet ID
  - z/OS CommServer will create no more than two SMC-R links within an SMC-R link group
    - Priority given to RNIC adapters with unique PFIP values
    - RNIC interfaces that are being used in the SMC-R link group are called the "associated RNICs"
    - Remaining PFIDs are reserved for backup purposes only
  - If one of the associated RNICs fails, a new SMC-R link is created over one of the "backup" PFIDs
    - When failing RNIC recovers, it is now used as backup for the associated RNICs
  - Allows for an orderly approach for planned outages of the RNIC adapters