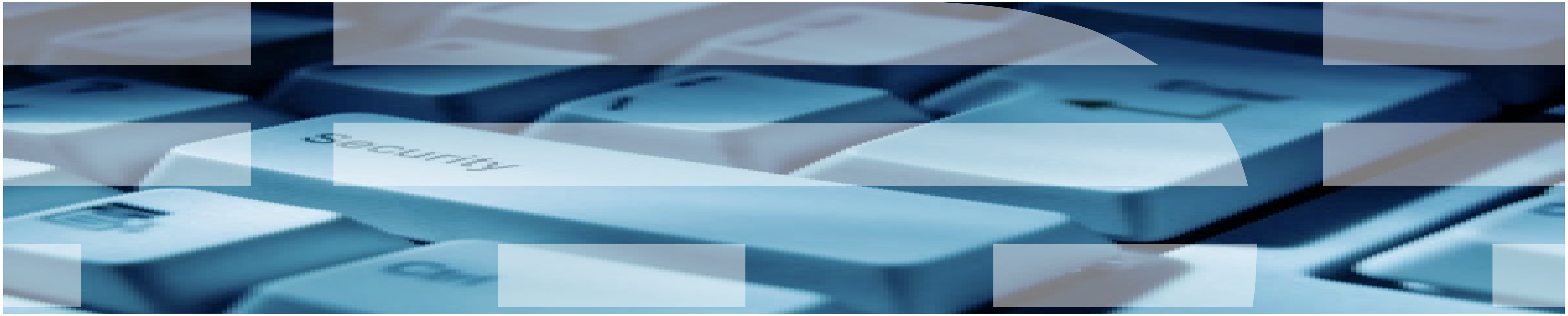
The z Exchange – September 15, 2020

IBM z/OS Communications Server
Shared Memory Communications (SMC)

Randall Kunkel (kunkel@us.ibm.com)



Please note

IBM’s statements regarding its plans, directions, and intent are subject to change or withdrawal without notice and at IBM’s sole discretion.

Information regarding potential future products is intended to outline our general product direction and it should not be relied on in making a purchasing decision.

The information mentioned regarding potential future products is not a commitment, promise, or legal obligation to deliver any material, code or functionality. Information about potential future products may not be incorporated into any contract.

The development, release, and timing of any future features or functionality described for our products remains at our sole discretion.

Performance is based on measurements and projections using standard IBM benchmarks in a controlled environment. The actual throughput or performance that any user will experience will vary depending upon many factors, including 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 results similar to those stated here.

Agenda

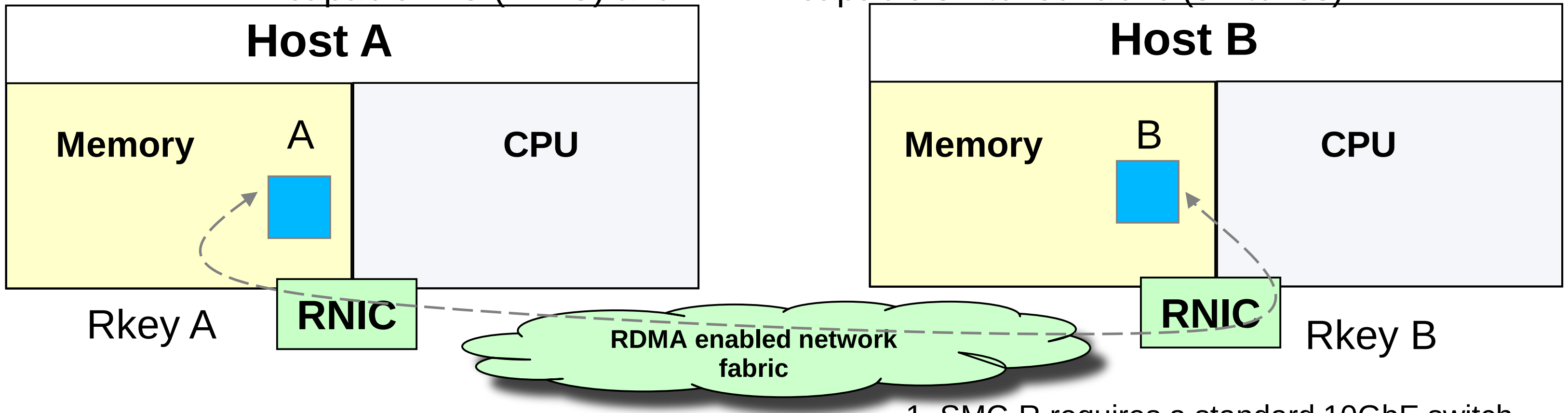
1. Review Shared Memory Communications over RDMA (SMC-R)
2. Shared Memory Communications – Direct Memory Access (SMC-D):
   - Introduction: SMC-D: Summary of SMC-D and ISM functions
   - Objectives / Value of SMC-D and ISM (performance overview)
3. IBM z System z13 Internal Shared Memory (ISM) virtual PCI function
4. Getting started: Setup requirements for enabling SMC-D:
   - ISM System definitions (defining FIDs in HCD)
   - z/OS Communications Server configuration requirements (enabling SMC-D)
5. SMC Applicability Tool (SMC-AT)

Shared Memory Communications over RDMA (SMC-R)

Review: RDMA (Remote Direct Memory Access) Technology Overview

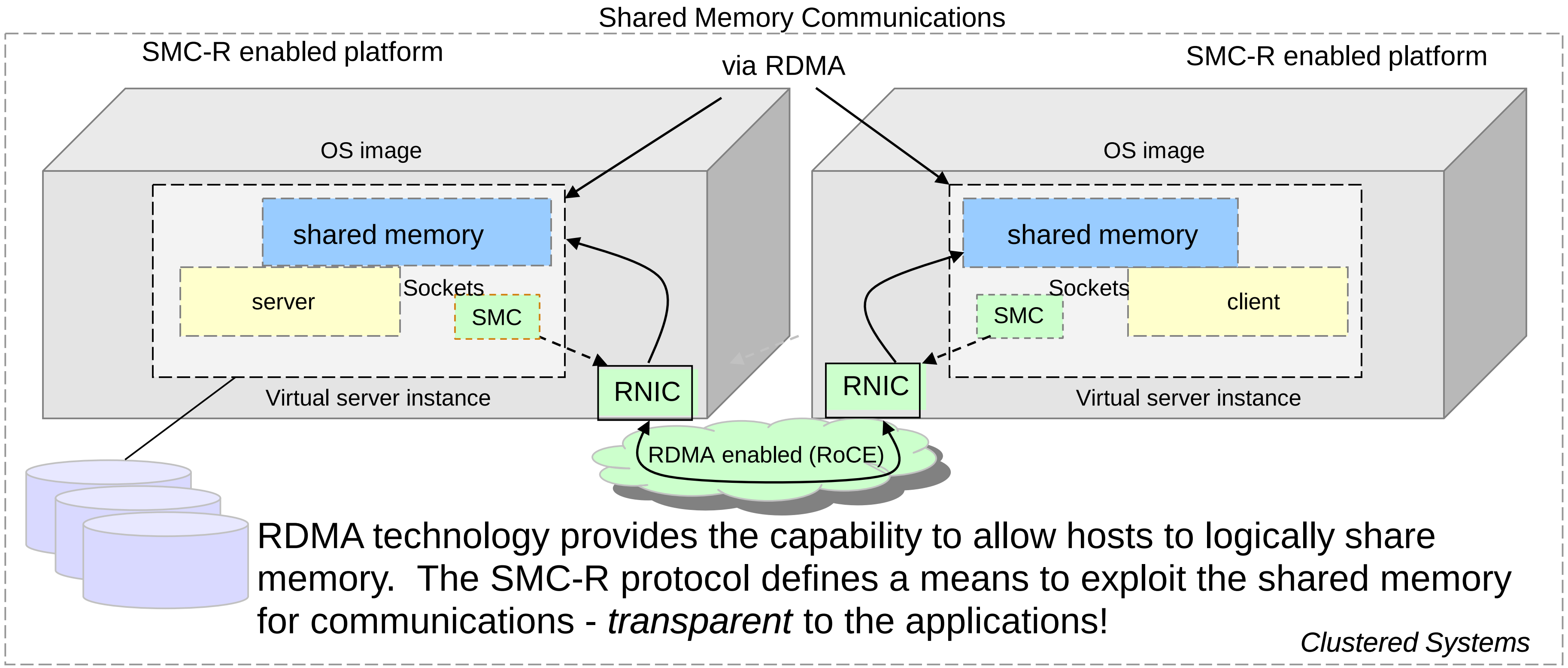
Key attributes of RDMA

- 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

Review: Shared Memory Communications over RDMA (SMC-R)



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 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!

IETF RFC for SMC-R:

http://www.rfc-editor.org/rfc/rfc7609.txt

Innovations available on zBC12 and zEC12

Data Compression Acceleration	**High Speed Communication Fabric**	ENHANCED **Flash Technology Exploitation**	ENHANCED **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%*****

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%******

Shared Memory Communications (SMC-R):

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

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

z/OS V2.1 SMC-R

z/VM 6.3 support for guests

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

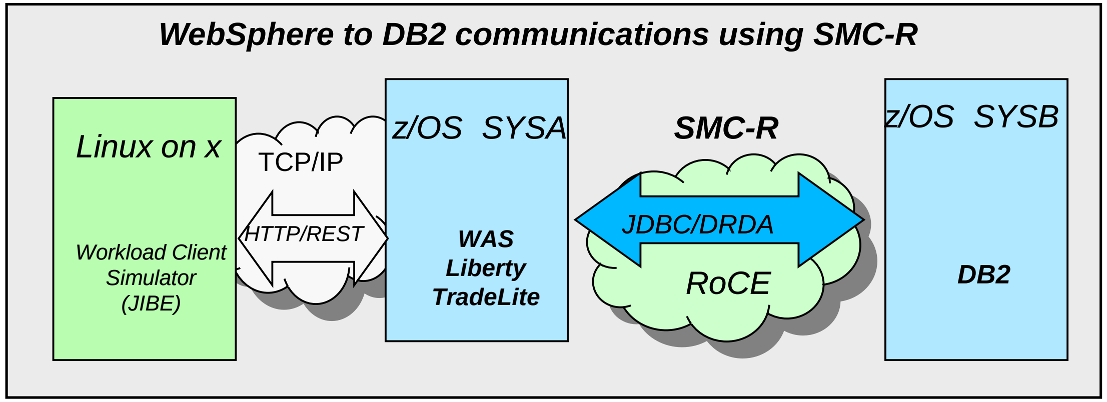
Use Cases

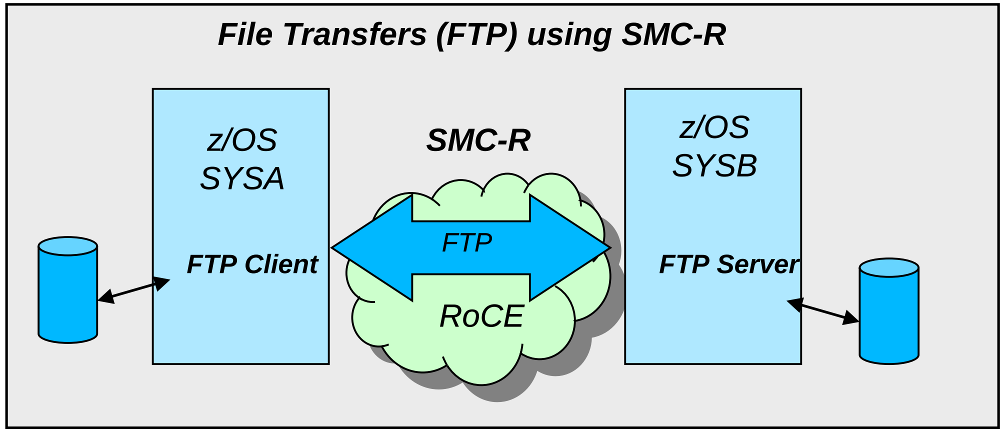
- Application servers such as the z/OS WebSphere Application Server communicating (via TCP based communications) with CICS, IMS or DB2 – particularly when the application is network intensive and transaction oriented
- Transactional workloads that exchange larger messages (e.g. web services such as WAS 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 *



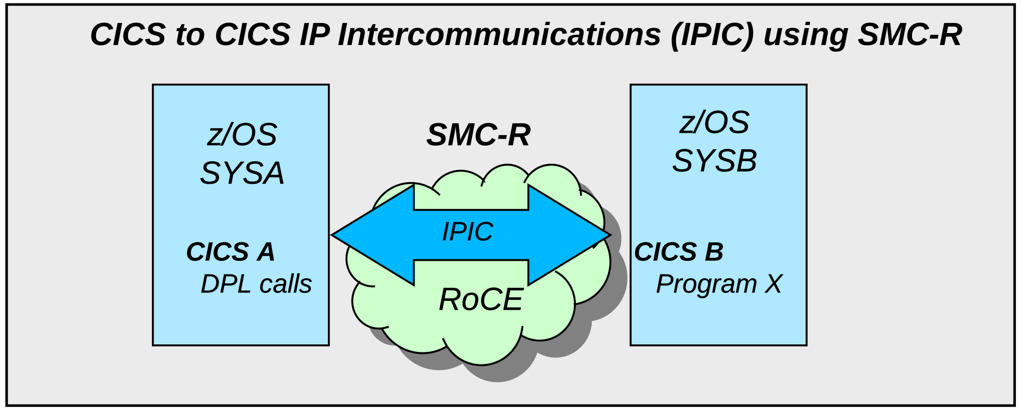


Up to ***50% CPU savings*** for FTP binary file transfers across z/OS systems when using SMC-R vs standard TCP/IP **

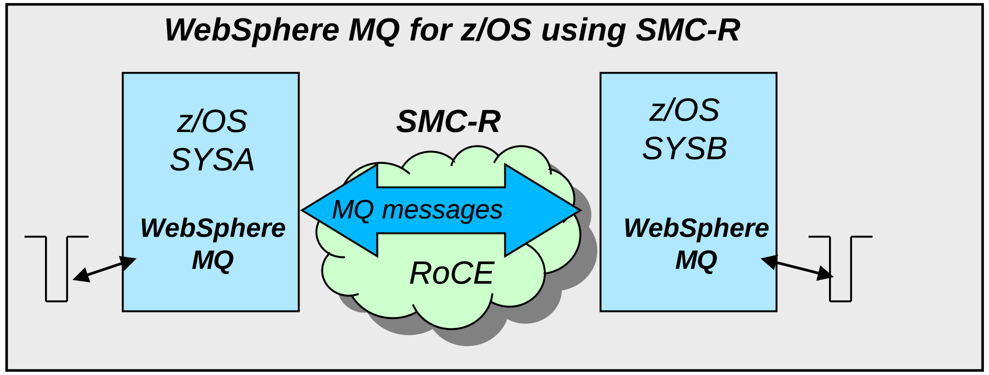
* Based on projections and measurements completed in a controlled environment. Results may vary by customer based on individual workload, configuration and software levels.
** Based on internal IBM benchmarks in a controlled environment using z/OS V2R1 Communications Server FTP client and FTP server, 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



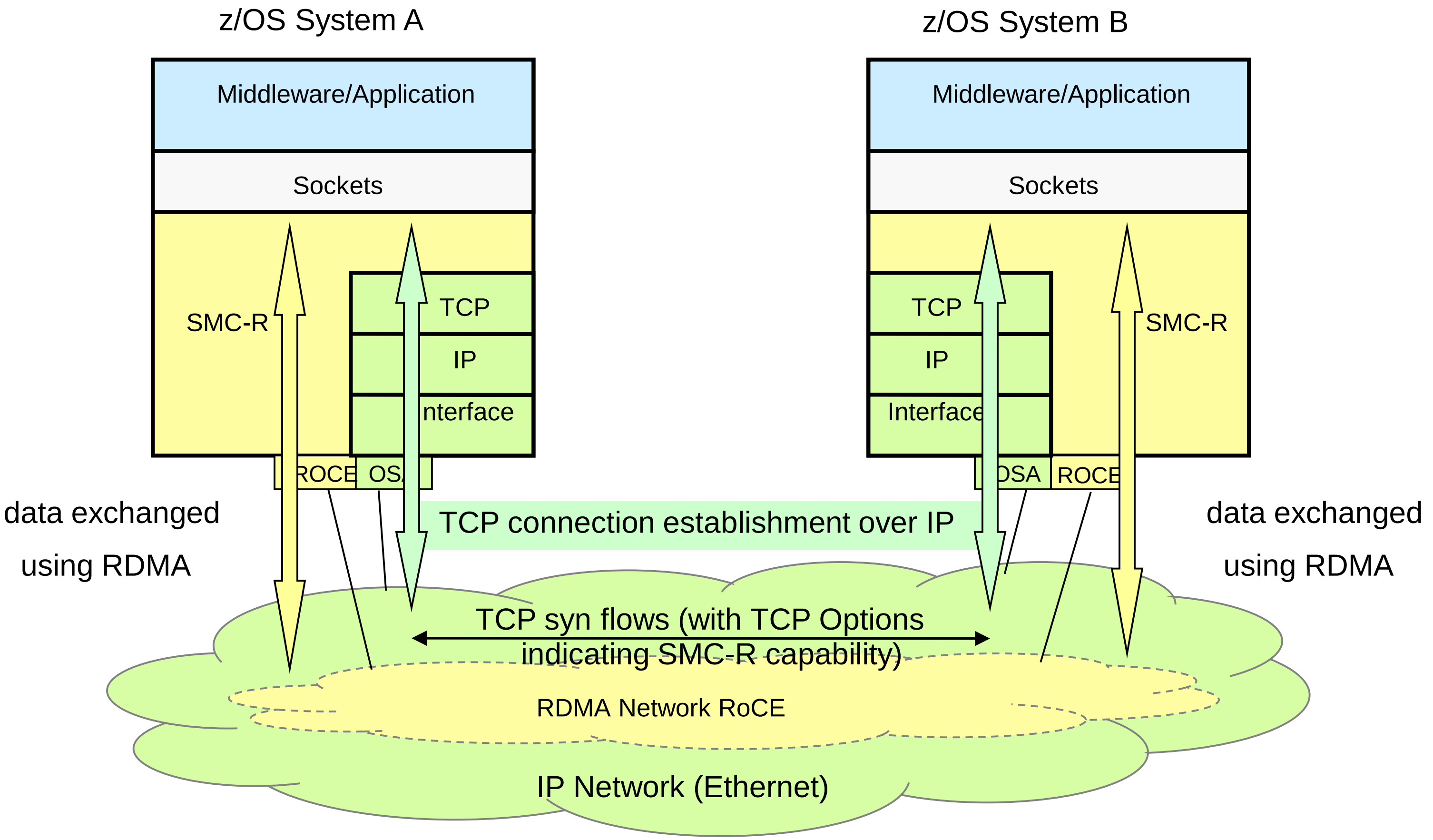
WebSphere MQ for z/OS using SMC-R

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.

Dynamic Transition from TCP to SMC-R



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

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!*

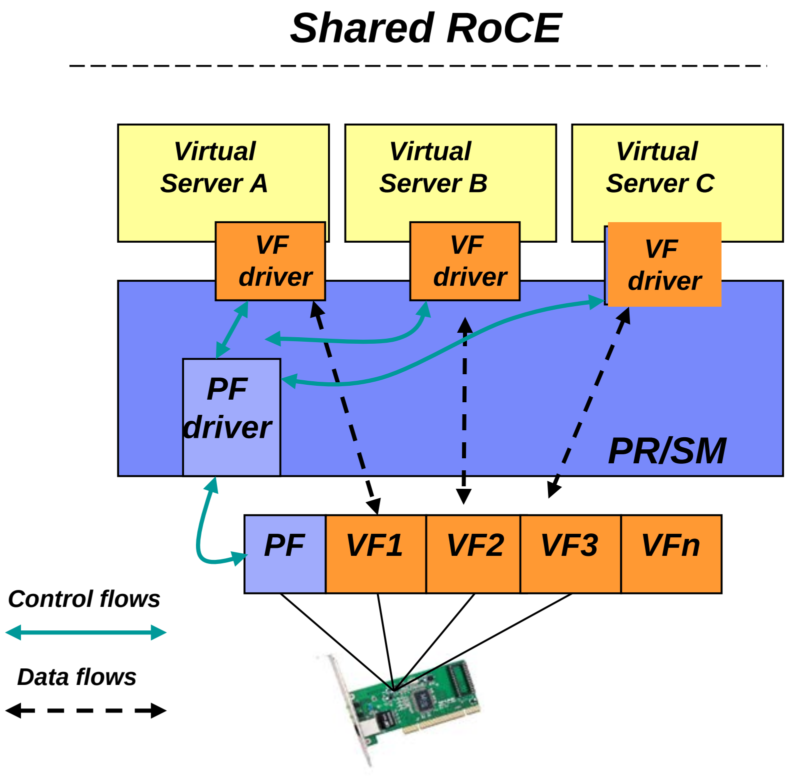
SMC-R and 10GbE RoCE Express Requirements

- **Operating system requirements**
  - Requires z/OS 2.2 which supports the SMC-R protocol
- **Server requirements**
  - 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)

SMC-R and RoCE Express Support – IBM z13, z14 and z15 Systems

- Shared RoCE Express - Available on IBM z13 System and higher
  - 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 is available on z/OS V2R2 (base) and on z/OS V2R1 via APAR/PTF
    - z/OS V2R1: UI28823 / PI38739 and UI28842 / PI38739 and UA77894 / OA47561
- z14 introduces RoCE Express2 10GbE
  - Can be shared across 63 Virtual Functions (VFs) per physical port – earlier version supported 31 VFs
  - Supported by default on z/OS V2R3, requires the following APARs on prior releases:
    V2R1: OA51949 / PI75199
    V2R2: OA51950 / PI75200
- z14 GA2 introduces RoCE Express2 25GbE
  - No software changes required (see backup)

10GbE
RoCE
Express



SMC-R TCP Connection Eligibility

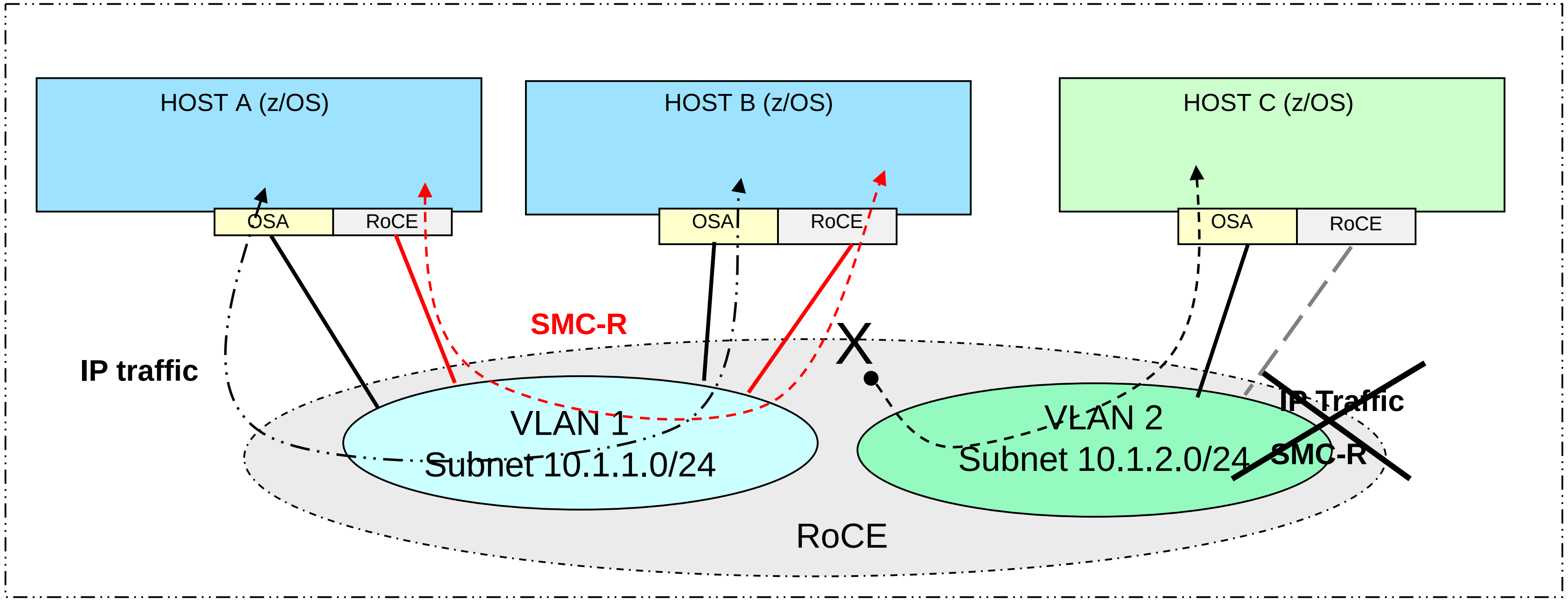
Rules… All eligible hosts must:

1. be **SMCR enabled** (z/OS V2R2 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 Key Attributes - Summary

- ✓ Optimized Network Performance (leveraging RDMA technology)
- ✓ Transparent to (TCP socket based) application software
- ✓ Leverages existing Ethernet infrastructure (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 and Other Platforms

1. SMC-R is an open protocol designed to meet the key objectives (discussed earlier in this presentation) required by our customers. SMC-R is described in the IETF with an information RFC 7609.

2. SMC Status for:

   1. Linux:
      SMC-R and SMC-D has been accepted upstream in the Linux kernel and is now available on RHEL 8, SLES 12 SP4, SLES 15 SP1 and Ubuntu 18.10.

   2. For additional details about Linux SMC:

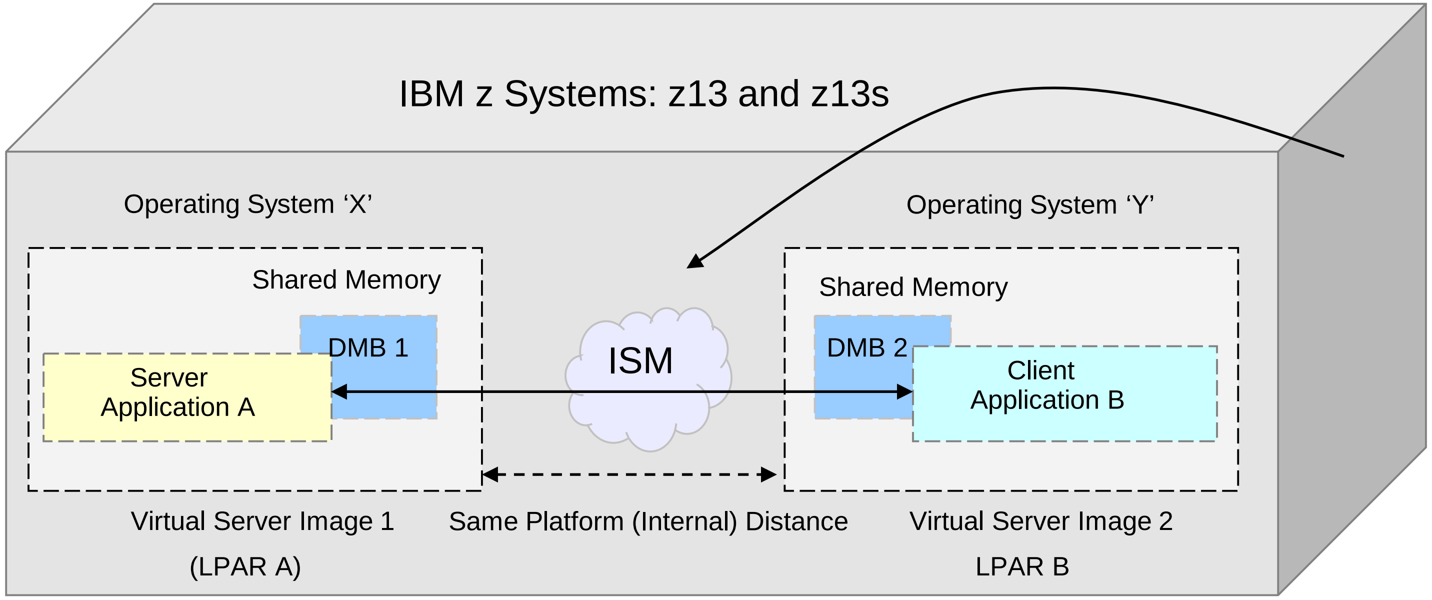
      - https://linux-on-z.blogspot.com/p/smc-for-linux-on-ibm-z.html

   3. AIX:
      System P with AIX 7.2 **SMC-R is now GA**! (Oct 27, 2017). See the details here:

      https://www-01.ibm.com/common/ssi/cgi-bin/ssialias?infotype=an&subtype=ca&appname=gpateam&supplier=872&letternum=ENUSAP17-0480

Shared Memory Communications – Direct Memory Access (SMC-D Introduction)

Shared Memory Communications-Direct Memory Access (SMC-D) over Internal Shared Memory (ISM)



“Shared Memory”

across unique OS instances within the same CPC

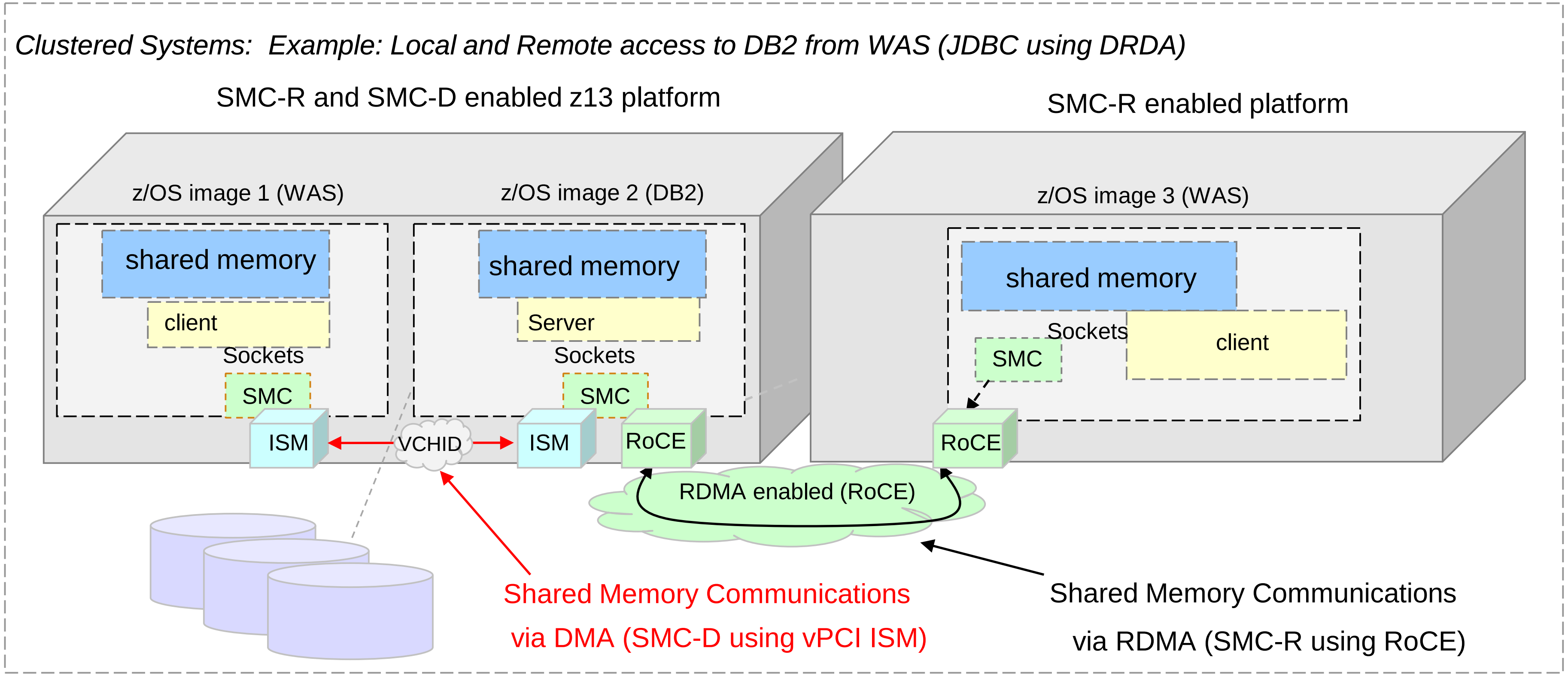
SMC-D (over ISM) extends the value of the Shared Memory Communications architecture by enabling SMC for direct LPAR to LPAR communications. SMC-D is very similar to SMC-R (over RoCE) extending the benefits of SMC-R to same CPC operating system instances without requiring physical resources (RoCE adapters, PCI bandwidth, NIC ports, I/O slots, network resources, 10GbE switches etc.).

Note 1. The performance benefits of SMC-R (cross CPC) and HiperSockets (within CPC) are similar to each other.

SMC-D / ISM provides significantly improved performance benefits above both within the CPC.

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

Shared Memory Communications within the enterprise data center (RoCE) and within System z (ISM)

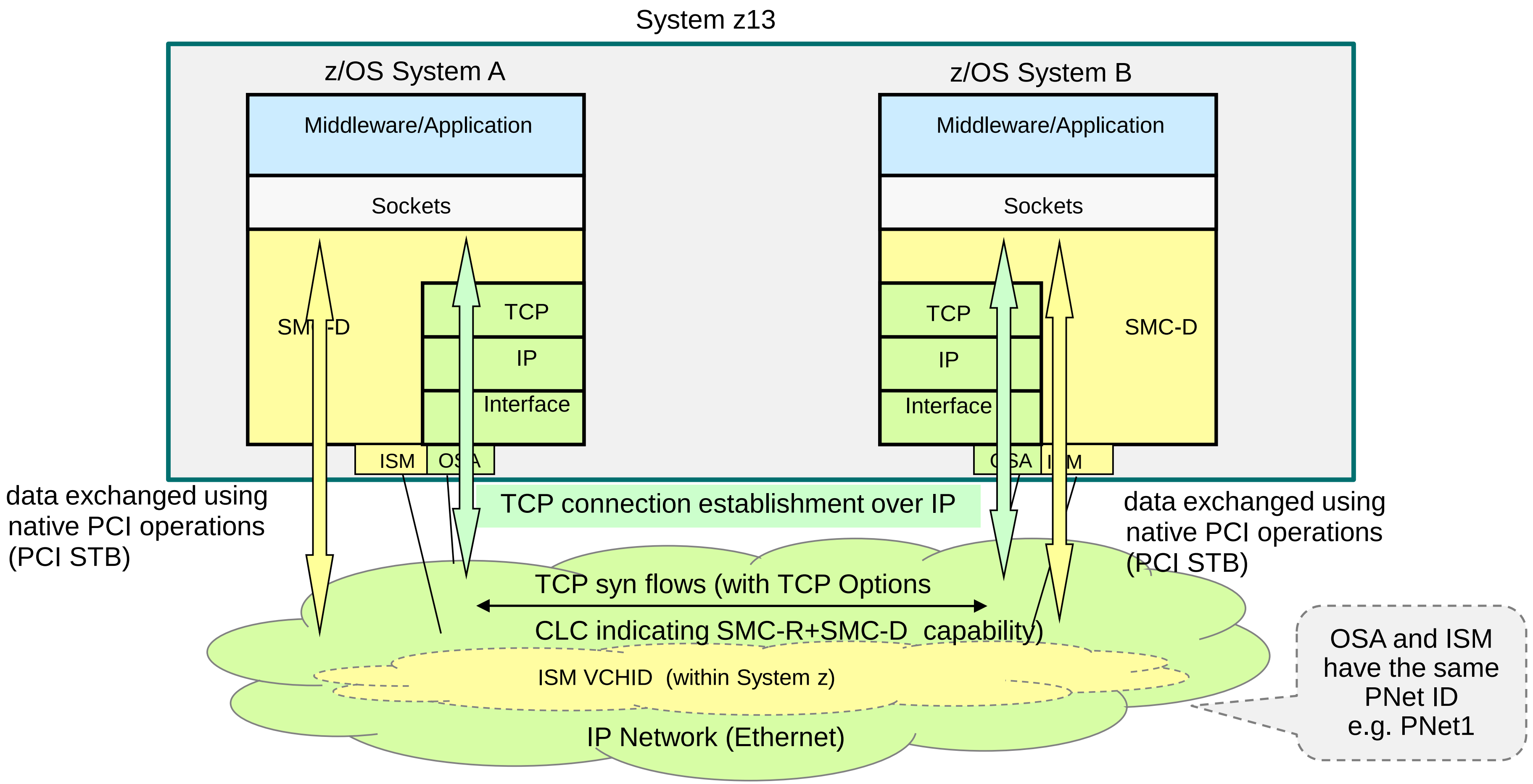

Clustered Systems: Example: Local and Remote access to DB2 from WAS (JDBC using DRDA)
SMC-R and SMC-D enabled z13 platform
SMC-R enabled platform
z/OS image 1 (WAS)
z/OS image 2 (DB2)
z/OS image 3 (WAS)
shared memory
shared memory
shared memory
client
Server
client
Sockets
Sockets
Sockets
SMC
SMC
SMC
ISM
VCHID
ISM
RoCE
RoCE
RDMA enabled (RoCE)
Shared Memory Communications via DMA (SMC-D using vPCI ISM)
Shared Memory Communications via RDMA (SMC-R using RoCE)

Both forms of SMC can be used concurrently combining to provide a highly optimized solution.

Shared Memory Communications: via System z PCI architecture:

1. RDMA (SMC-R for cross platforms via RoCE)
2. DMA (SMC-D for same CPC via ISM)

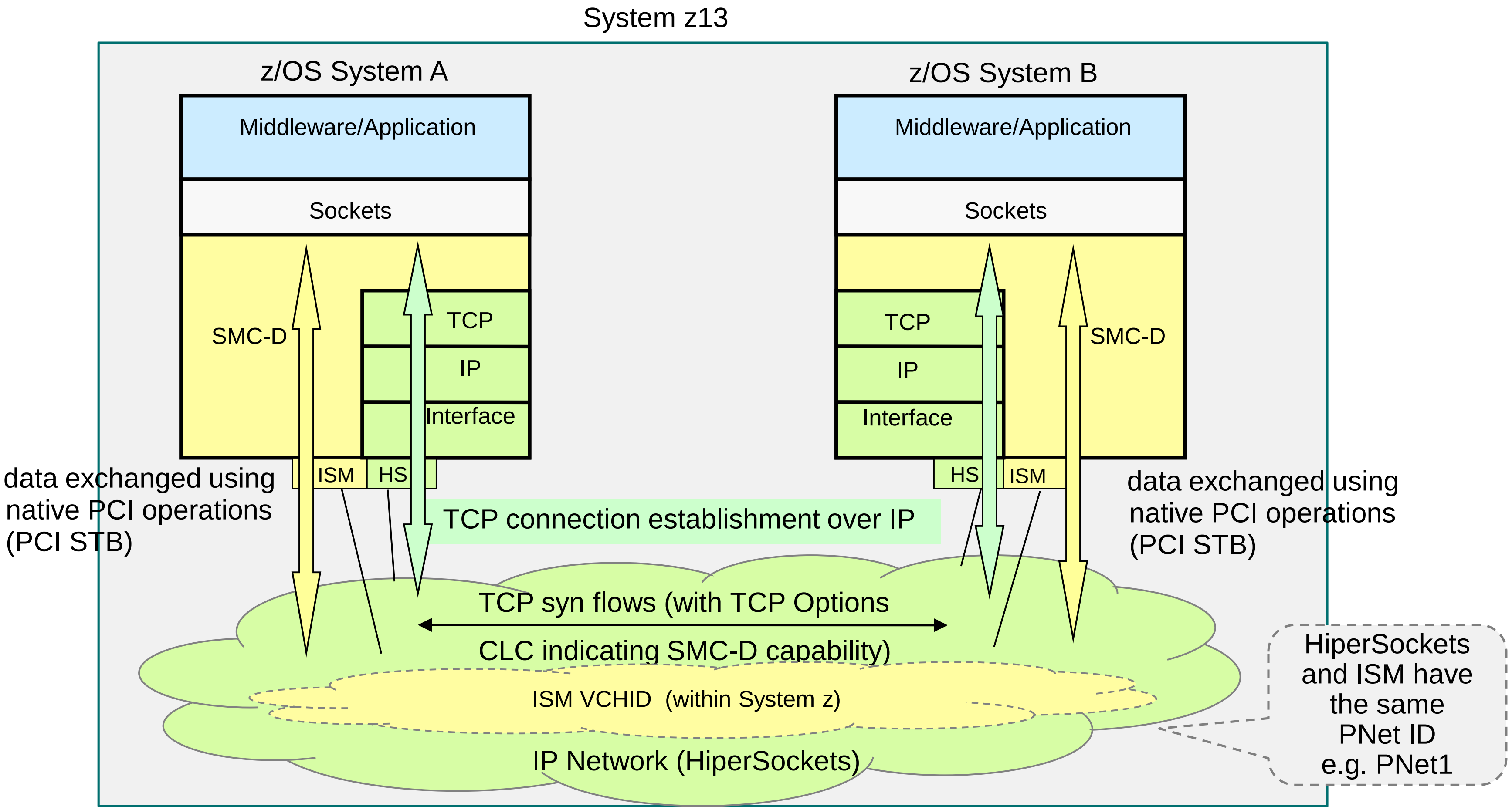
Dynamic Transition from TCP to SMC-D (OSA/LAN IP network)


System z13
z/OS System A
Middleware/Application
Sockets
SMC-D
TCP
IP
Interface
ISM
OSA
z/OS System B
Middleware/Application
Sockets
TCP
IP
Interface
SMC-D
OSA
ISM
data exchanged using native PCI operations (PCI STB)
TCP connection establishment over IP
data exchanged using native PCI operations (PCI STB)
TCP syn flows (with TCP Options
CLC indicating SMC-R+SMC-D capability)
ISM VCHID (within System z)
IP Network (Ethernet)
OSA and ISM have the same PNet ID e.g. PNet1

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

TCP connection transitions to SMC-D allowing application data to be exchanged using Direct Memory Access (LPAR to LPAR)

Dynamic Transition from TCP to SMC-D – (HiperSockets IP Network)



System z13
z/OS System A
Middleware/Application
Sockets
SMC-D
TCP
IP
Interface
ISM
HS
z/OS System B
Middleware/Application
Sockets
TCP
IP
Interface
HS
ISM
SMC-D
data exchanged using native PCI operations (PCI STB)
data exchanged using native PCI operations (PCI STB)
TCP connection establishment over IP
TCP syn flows (with TCP Options
CLC indicating SMC-D capability)
ISM VCHID (within System z)
IP Network (HiperSockets)
HiperSockets and ISM have the same PNet ID e.g. PNet1

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

TCP connection transitions to SMC-D allowing application data to be exchanged using Direct Memory Access (LPAR to LPAR)

Shared Memory Communications architecture

Faster communications that preserve TCP/IP qualities of service

- Shared Memory Communications – Direct Memory Access (SMC-D) optimizes z/OS for improved performance in ‘***within-the-box’*** communications versus standard TCP/IP over HiperSockets or Open System Adapter

Typical Client Use Cases:

- Valuable for multi-tiered work co-located onto a single z Systems server without requiring extra hardware
- Any z/OS TCP sockets based workload can seamlessly use SMC-D without requiring any application changes

SMC Applicability Tool (SMCAT) is available to assist in gaining additional insight into the applicability of SMC-D (and SMC-R) for your environment

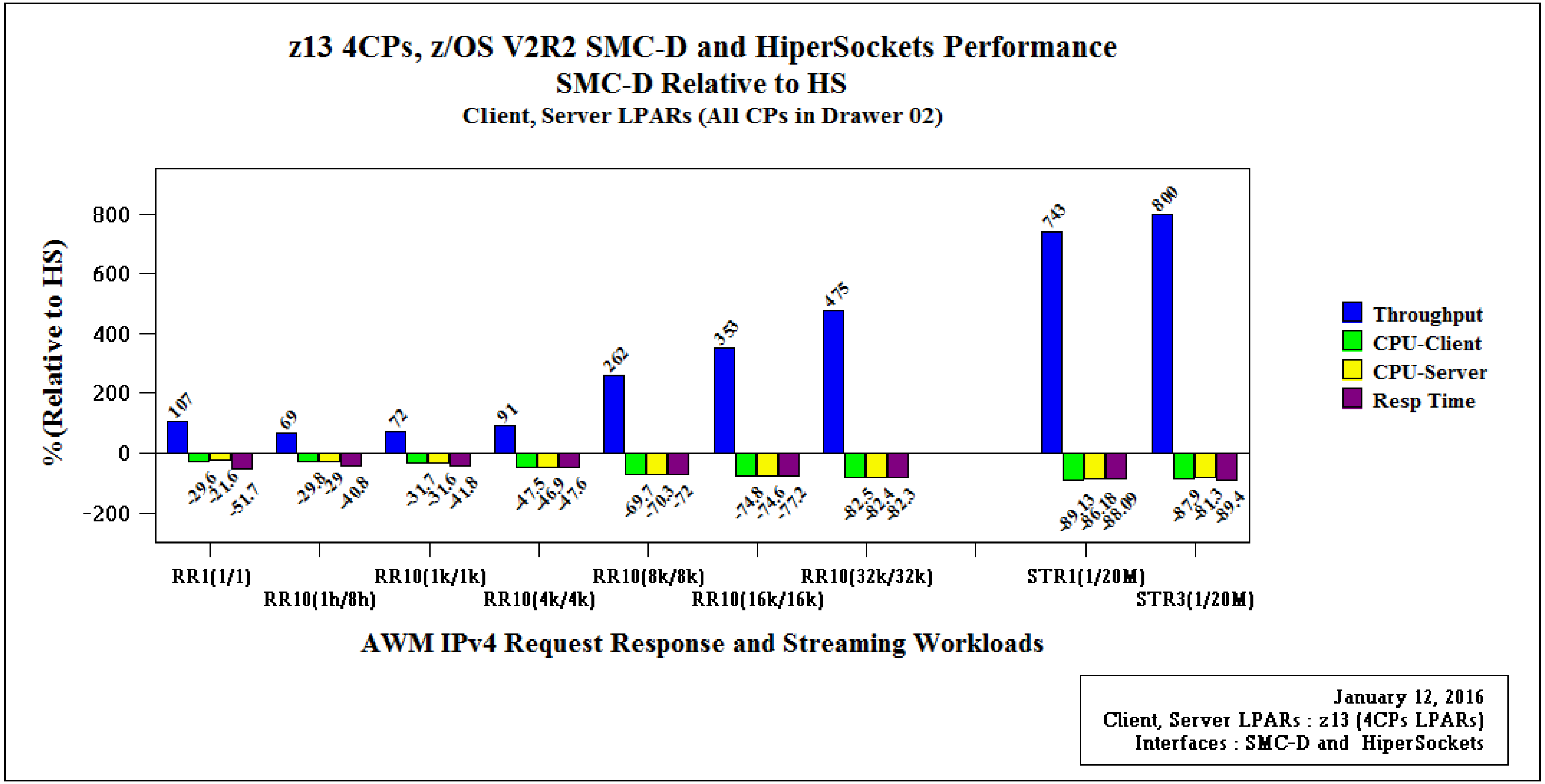
*Up to **61%** CPU savings for FTP file transfers across z/OS systems versus HiperSockets**

*Up to **9x** improvement in throughput with more than a **88%** decrease in CPU consumption and a **90%** decrease in response time for streaming workloads versus using HiperSockets**

*Up to **91%** improvement in throughput and up to **48%** improvement in response time for interactive workloads versus using HiperSockets**

* All performance information was determined in a controlled environment. Actual results may vary. Performance information is provided “AS IS” and no warranties or guarantees are expressed or implied by IBM.

HiperSockets Comparison


z13 4CPs, z/OS V2R2 SMC-D and HiperSockets Performance
SMC-D Relative to HS
Client, Server LPARs (All CPs in Drawer 02)

January 12, 2016
Client, Server LPARs : z13 (4CPs LPARs)
Interfaces : SMC-D and HiperSockets

Up to 9x the throughput! See breakout summary on next chart.

* All performance information was determined in a controlled environment. Actual results may vary. Performance information is provided “AS IS” and no warranties or guarantees are expressed or implied by IBM.

SMC-D / ISM to HiperSockets Summary Highlights

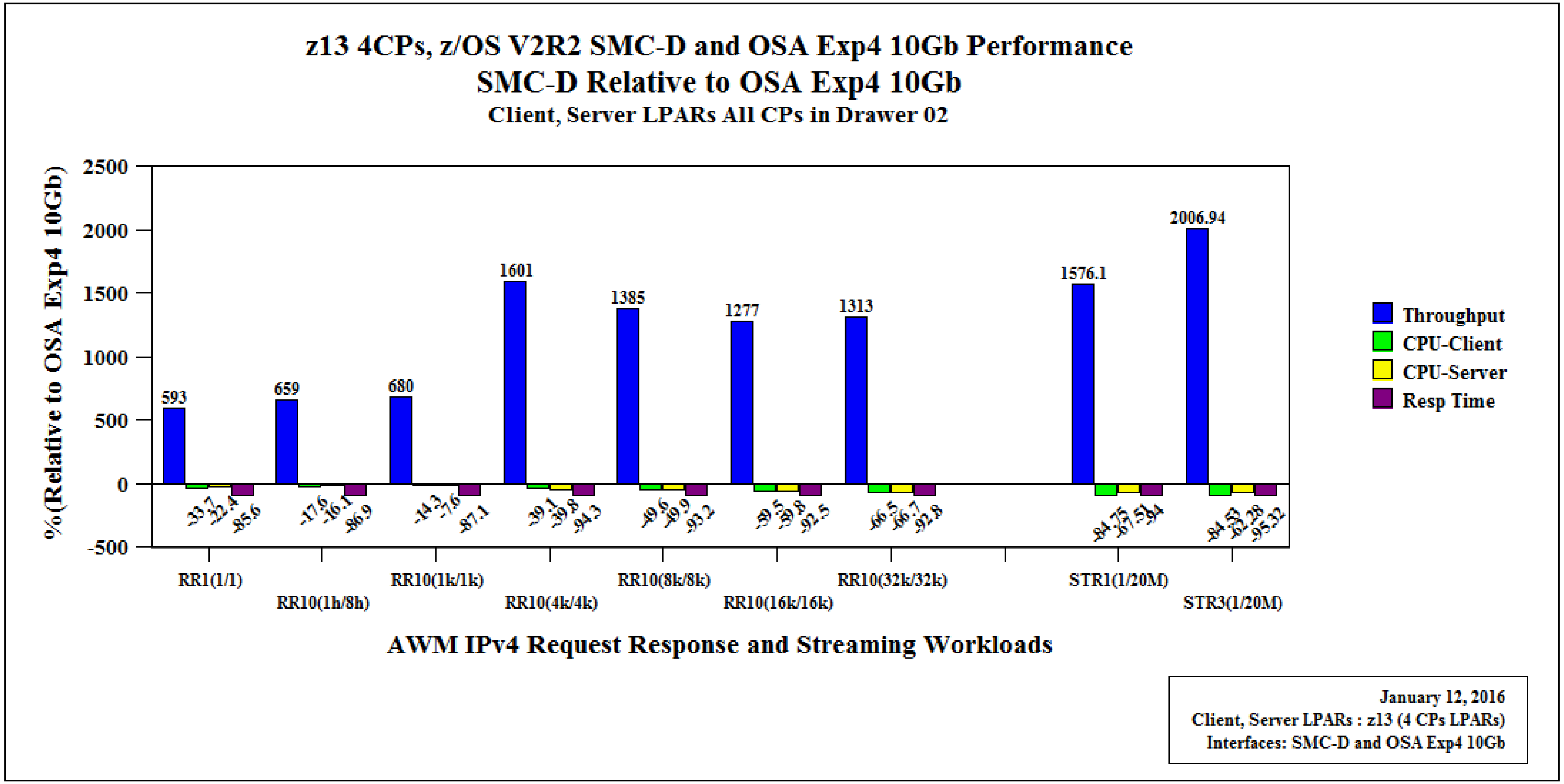
- **Request/Response Summary for Workloads with 1k/1k – 4k/4k Payloads:**
  - **Latency: Up to 48% reduction in latency**
  - **Throughput: Up to 91% increase in throughput**
  - **CPU cost: Up to 47% reduction in network related CPU cost**

- **Request/Response Summary for Workloads with 8k/8k – 32k/32k Payloads:**
  - **Latency: Up to 82% reduction in latency**
  - **Throughput: Up to 475% (~6x) increase in throughput**
  - **CPU cost: Up to 82% reduction in network related CPU cost**

- **Streaming Workload**:
  - **Latency: Up to 89% reduction in latency**
  - **Throughput: Up to 800% (~9x) increase in throughput**
  - **CPU cost: Up to 89% reduction in network related CPU cost**

* All performance information was determined in a controlled environment. Actual results may vary. Performance information is provided “AS IS” and no warranties or guarantees are expressed or implied by IBM.

OSA Comparison



January 12, 2016
Client, Server LPARs : z13 (4 CPs LPARs)
Interfaces: SMC-D and OSA Exp4 10Gb

Up to 21x the throughput! See breakout summary on next chart.

* All performance information was determined in a controlled environment. Actual results may vary. Performance information is provided “AS IS” and no warranties or guarantees are expressed or implied by IBM.

SMC-D / ISM to OSA Summary Highlights

- **Request/Response Summary for Workloads with 1k/1k – 4k/4k Payloads:**
  - **Latency: Up to 94% reduction in latency**
  - **Throughput: Up to 1601% (~17x) increase in throughput**
  - **CPU cost: Up to 40% reduction in network related CPU cost**

- **Request/Response Summary for Workloads with 8k/8k – 32k/32k Payloads:**
  - **Latency: Up to 93% reduction in latency**
  - **Throughput: Up to 1313% (~14x) increase in throughput**
  - **CPU cost: Up to 67% reduction in network related CPU cost**

- **Streaming Workload**:
  - **Latency: Up to 95% reduction in latency**
  - **Throughput: Up to 2001% (~21x) increase in throughput**
  - **CPU cost: Up to 85% reduction in network related CPU cost**
- **FTP:**
  - **For Binary Get and Put:**
    - **Up to 58% lower (receive side) CPU cost and**
    - **Up to 26% lower (send side) CPU cost and equivalent throughput**

* All performance information was determined in a controlled environment. Actual results may vary. Performance information is provided “AS IS” and no warranties or guarantees are expressed or implied by IBM.

SMC-D and ISM (vPCI) Overall Value Points

*Provides **Highly optimized:** improved throughput, reduced latency and CPU cost for intra-CPC communications along with:*

- ✓ Provides the same list of key SMC-R value points:
  - ✓ Transparent to socket applications, no IP topology changes, preserves connection level security, VLAN isolation, transparent with load balancers, etc.
- ✓ …without requiring hardware (adapters, card slots, switches, PCI infrastructure, fabric management, etc.)… cost savings
- ✓ Provides superior resiliency / High Availability (no hardware failures)
- ✓ Provides high scalability, bandwidth and virtualization (i.e. 8k virtual functions)
- ✓ Preserves security (connection level security + secure internal communications)
- ✓ Preserves value of z Systems co-location of workloads (e.g. highly optimized internal communications)
- ✓ Enabled in z/OS with a single TCP/IP profile keyword [1]

Note 1. ISM VCHID and FIDs must be defined in HCD (IOCDS)

Internal Shared Memory (ISM) Introduction

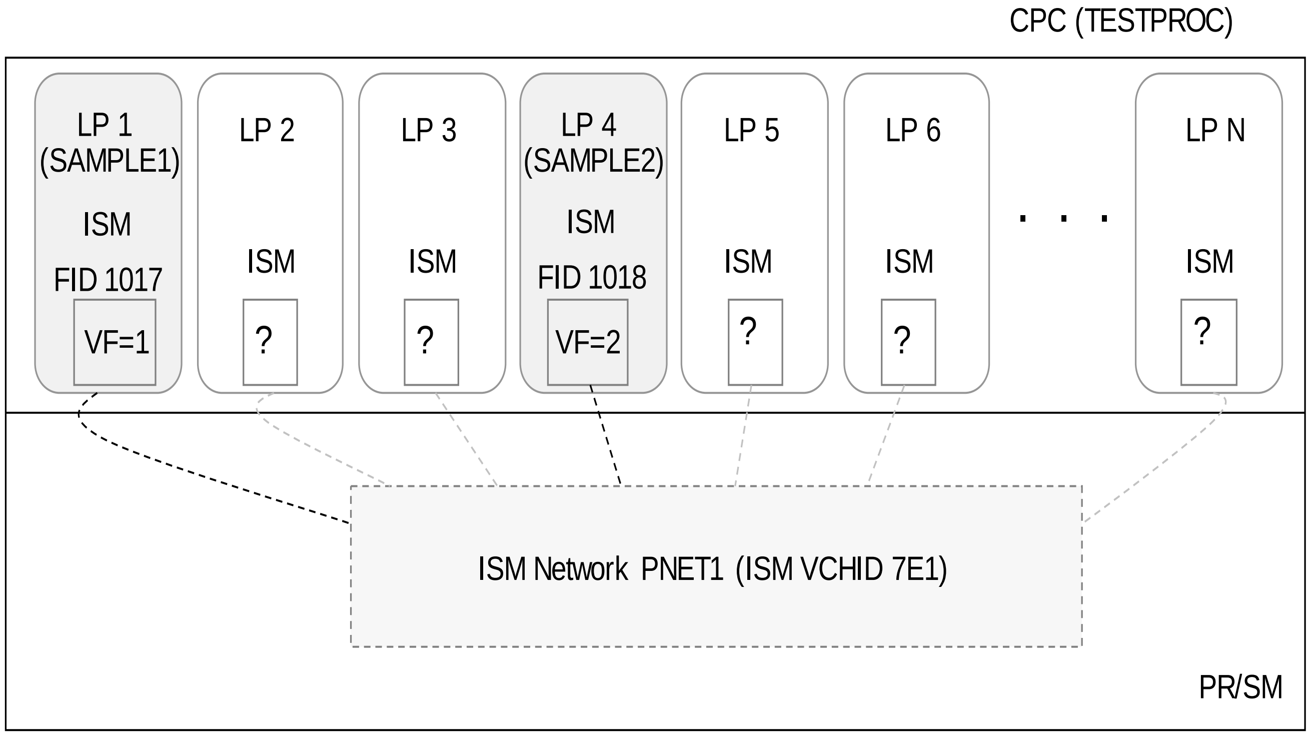
- ISM enables the ability for Operating Systems (LPARs) to share virtual memory (similar to RDMA)
- New “Internal Shared Memory” (ISM) VCHID Type
  (ISM VCHID concepts are similar to IQD (HiperSockets) VCHID)
- ISM is based on existing z System’s PCI architecture (i.e. virtual PCI Function / adapter)
- Introduces a new PCI Function type (ISM virtual PCI function)
- System admin / configuration / operations follows the same process (HCD/IOCDS) as existing PCI functions (e.g. RoCE Express, zEDC Express, etc.)
- ISM supports Dynamic I/O
- continued…

Internal Shared Memory (ISM) Introduction (part 2)

- Provides adapter virtualization (Virtual Functions) with high scalability:
  - 32 ISM VCHIDs per CPC (each VCHID represents a unique internal shared memory network each with a unique Physical Network ID)
  - 255 VFs per VCHID (8k VFs per CPC)
    (i.e. the maximum no. of virtual servers that can communicate over the same ISM VCHID is 255)
- Each ISM VCHID represents a unique (isolated) internal network, each having a unique Physical Network ID (PNet IDs are configured in HCD/IOCDS)
- ISM VCHIDs support VLANs (i.e. can be sub-divided into VLANs)
- ISM provides a GID ("Global ID" internally generated by firmware) that corresponds with each ISM FID. The GID is used to locate / address a host on an ISM network (VCHID)
- MACs (VMACs), MTU, physical ports[1] and Frame size are all N/A

[1] Note 1. ISM VCHIDs provide support for a single logical port (also see PNet ID topic)

Introduction: IBM z13 / z13s
Internal Shared Memory (ISM) virtual PCI Function



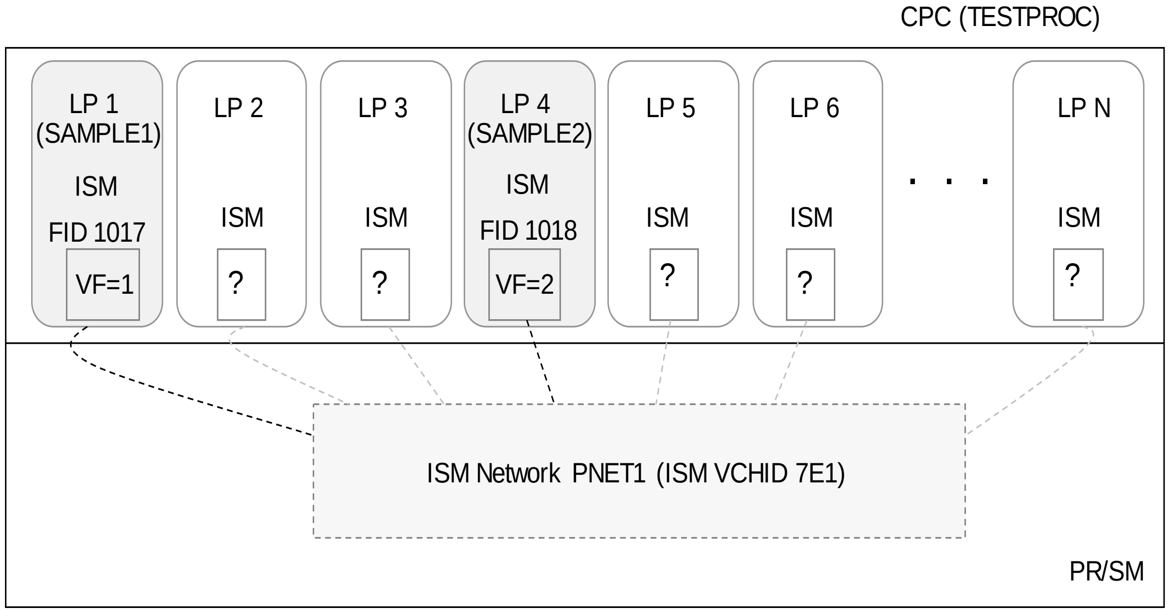
```
FUNCTION FID=1017,PCHID=7E1,VF=1,PART=((SAMPLE1),(SAMPLE1,SAMPLE2)),PNETID=(PNET1),TYPE=ISM
FUNCTION FID=1018,PCHID=7E1,VF=2,PART=((SAMPLE2),(SAMPLE1,SAMPLE2)),PNETID=(PNET1),TYPE=ISM
```

Install / Configure / Enable: ISM and SMC-D Overview

Associating ISM with your IP Network devices (OSA or HS)

- ISM Functions must also be associated with another Channel (CHID), either:
  1. A HiperSockets channel (IQD) or
  2. An OSA Express (OSD) channels
     Note. A single ISM VCHID can not be associated with both (IQD and OSD)
- The association of an ISM VCHID (Function IDs) to the channel(s) is created by defining (HCD) matching **Physical Network IDs** (PNet IDs)
- ISM supports one PNet ID per ISM VCHID (a single “logical port”)
- PNet IDs are dynamically discovered by z/OS (from HCD config)
- The channel devices (OSD or IQD) provide IP connectivity and are associated with ISM based on having matching PNet IDs

ISM Configuration Example (see the following HCD charts)

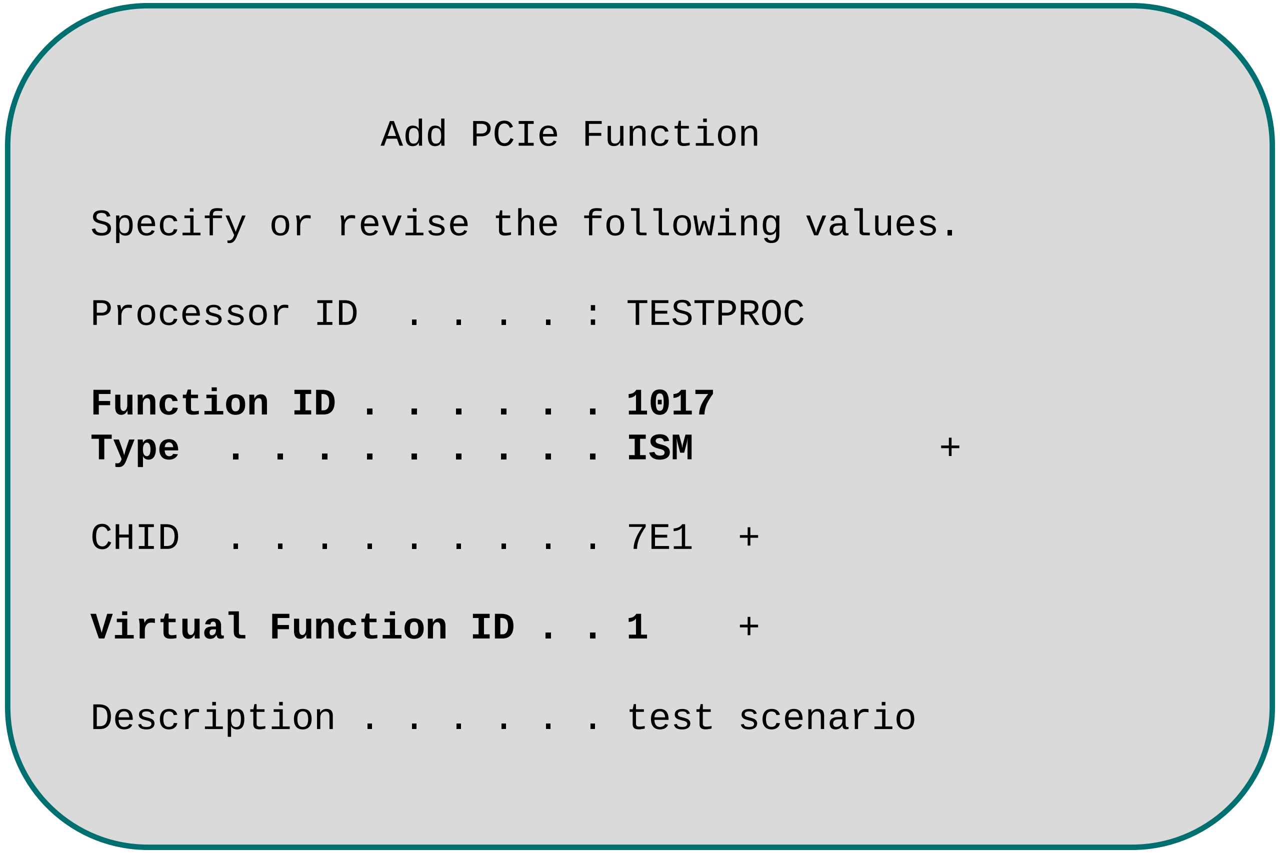


```
FUNCTION FID=1017,PCHID=7E1,VF=1,PART=((SAMPLE1),(SAMPLE1,SAMPLE2)),PNETID=(PNET1),TYPE=ISM
FUNCTION FID=1018,PCHID=7E1,VF=2,PART=((SAMPLE2),(SAMPLE1,SAMPLE2)),PNETID=(PNET1),TYPE=ISM
```

Add PCIe Function

Define the ISM function:

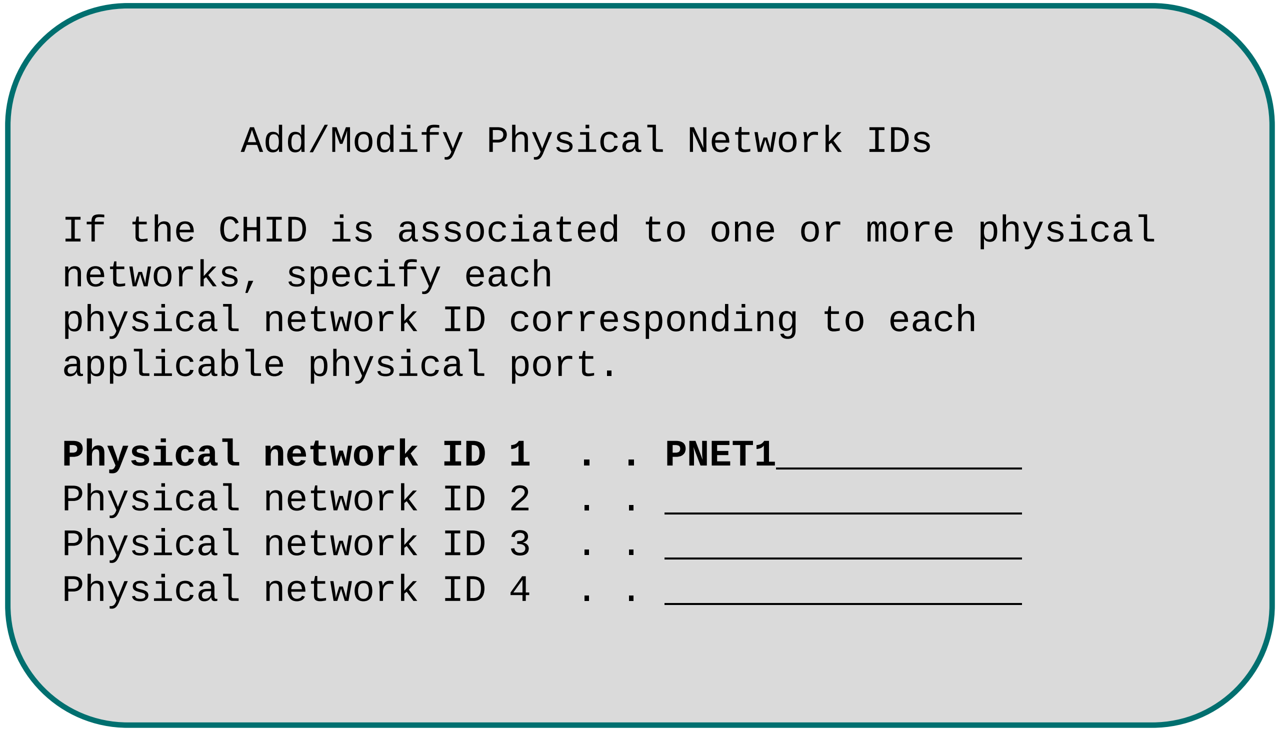
1. action f on processor to see the PCIe function list
2. action add on function list (PF11 or line command add like)
   Note the ISM VCHID 7E1



```
                Add PCIe Function

 Specify or revise the following values.

 Processor ID . . . . : TESTPROC

 Function ID . . . . . . 1017
 Type  . . . . . . . . . ISM             +

 CHID  . . . . . . . . . 7E1  +

 Virtual Function ID . . 1    +

 Description . . . . . . test scenario
```

Press Enter

Add/Modify ISM PNet ID



```
        Add/Modify Physical Network IDs

If the CHID is associated to one or more physical
networks, specify each
physical network ID corresponding to each
applicable physical port.

Physical network ID 1  . . PNET1___________
Physical network ID 2  . . ________________
Physical network ID 3  . . ________________
Physical network ID 4  . . ________________
```

Press Enter

PNet ID Notes.

1. ISM supports a single PNet ID per ISM VCHID
2. ISM PNet IDs must be unique among other ISM VCHIDs for this System
3. ISM PNet IDs must match a corresponding IQD VCHID or OSD Channel(s)

z/OS CommServer Exploitation of Internal Shared Memory (ISM)

- ISM enables Shared Memory Communications-Direct Memory Access (SMC-D)
- Once the ISM HCD configuration is complete, SMC-D can be enabled in z/OS with a single TCP/IP parameter (**GLOBALCONFIG SMCD**).
- Notes:
  - ISM FIDs **are not** defined in TCP/IP. ISM FIDs are dynamically discovered.
  - An OS can be enabled for both SMC-R and SMC-D. SMC-D is used when both peers are within the same CPC (and using the ISM VCHID and VLAN).
  - ISM FIDs (VCHIDs) must be associated with an IP network. The association is accomplished by defining matching PNet IDs (e.g. HS and ISM).
    Notes:
    - Your OSA (or IQD channel) must have a PNet ID defined (and must match your ISM FID)
    - The OSA or IQD INTERFACE statement must have IPSubnet defined
  - Host virtual memory is managed by each OS (similar to SMC-R, logically shared memory) following existing z System’s PCI I/O translation architecture (i.e. only minor changes required for z/VM guests). There are no required configuration changes.

TCPIP Profile GlobalConfig SMCD

- SMCD parameter on GLOBALCONFIG (similar to SMCR)
- **Single Keyword! SMCD** is the only required setting to enable SMC-D
- Key difference from SMCR parameter: ISM PFIDs are not defined in TCPIP (ISM FIDs are auto discovered based on matching PNETID associated with the OSD or HiperSockets)

```
>>-GLOBALCONFig--------------------------------------------------------->

   .--------------------------------------------------------------.
   V                                                              |
>-----+--------------------------------------------------------+--+--><
     :                                                          :
      | .-NOSMCD---------------------------------------------. |
      | |                                                    | |
      +-+----------------------------------------------------+-+
      | |      .-------------------------------------------. | |
      | |      V                                           | | |
      | '-SMCD---+---------------------------------------+-+-' |
      |          |   .-FIXEDMemory--256------.           |     |
      |          +---+-----------------------+-----------+     |
      |          |   '-FIXEDMemory--mem_size-'           |     |
      |          |   .-TCPKEEPmininterval--300------.    |     |
      |          '---+------------------------------+----'     |
      |              '-TCPKEEPmininterval--interval-'          |
```

SMC Applicability Tool (SMC-AT)

Determining SMC benefits – SMC Applicability Tool

- Several customers have expressed interest in SMC-R and SMC-D
  - One of the first questions that is raised is “What benefit will SMC provide in my environment?”
    - Some users are well aware of significant traffic patterns that can benefit from SMC
    - 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
  - 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 or SMC-D in their environment with minimal effort and minimal impact
  - Part of the TCP/IP stack: Gather new statistics that are used to project SMC 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 maintenance (z/OS V1R13, z/OS V2R1)
  - Does not require SMC-R or SMC-D 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
    - For both SMC-R and SMC-D

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 for)
    - If subnets are used, ***the entire subnet must be comprised of z/OS systems*** that are SMC 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 (SMC-R Eligible Traffic)
    2. % of SMC-R Eligible Traffic that is well suited to SMC (excludes workloads with very short lived TCP connections that have trivial payloads)

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

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

     15% of all TCP connections can use SMC (eligible)
       93% of eligible connections are well-suited for SMC
     15% of all TCP traffic (segments) is well-suited for SMC
       15% of outbound traffic (segments) is well-suited for SMC
       14% of inbound traffic (segments) is well-suited for SMC

  Interval Details:
    Total TCP Connections:                                      100
    Total SMC eligible connections:                              15
          Total SMC well-suited connections:                     14
    Total outbound traffic (in segments)                       1000
          SMC well-suited outbound traffic (in segments)        150
    Total inbound traffic (in segments)                         500
          SMC well-suited inbound traffic (in segments)          70


        . . .
```

How much of my TCP workload can benefit from SMC-R?

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

```
    . . .

    Application send sizes used for well-suited connections:
      Size                                 # sends     Percentage
      ----                                 -------     ----------
      1500 (<=1500):                            15      37%
      4K (>1500 and <=4k):                       7      17%
      8K (>4k and <= 8k):                        3       7%
      16K (>8k and <= 16k):                      4      10%
      32K (>16k and <= 32k):                     8      20%
      64K (>32k and <= 64k):                     3       7%
      256K (>64K and <= 256K):                   1       2%
      >256K:                                     0       0%

    Application receive sizes used for well-suited connections:
      Size                                 # recvs     Percentage
      ----                                 -------     ----------
      1500 (<=1500):                             8      38%
      4K (>1500 and <=4k):                       3      14%
      8K (>4k and <= 8k):                        2      10%
      16K (>8k and <= 16k):                      2      10%
      32K (>16k and <= 32k):                     4      20%
      64K (>32k and <= 64k):                     1       5%
      256K (>64K and <= 256K):                   1       5%
      >256K:                                     0       0%

    -------------------SMCAT Summary Report Export Area------------------

    20,10,4,5,10,5,2,0
    10,5,3,3,5,2,2,0
    15,7,3,4,8,3,1,0
    8,3,2,2,4,1,1,0

    -------------------End Export Area-----------------------------------

End of report
```

What kind of CPU savings and/or latency savings can I expect from SMC-R for outbound traffic?

What about inbound traffic?

This report is repeated for indirect IP connections

SMC References

- SMC One-Stop Shopping Web Page within IBM Knowledge Center (Includes latest links to ALL other SMC References):

  https://www.ibm.com/support/knowledgecenter/SSLTBW_2.3.0/com.ibm.zos.v2r3.halz002/smc_reference.htm

Shared Memory Communications Reference Information

Shared Memory Communications Reference Information

This site provides a comprehensive list of reference material for solutions related to IBM Shared Memory Communications: SMC-R, SMC-D and SMC-AT.

THANK YOU