IBM z/OS SMC Applicability Tool (SMC-AT)

Evaluating SMC (SMC-R and SMC-D) Applicability in your Environment

(03-10-16)

Dave Herr (dherr@us.ibm.com)

Agenda topics

- Problem:
  How to evaluate the applicability and the potential benefits of SMC in your IBM System z (customer's) environment
- There are two forms of SMC:
  - SMC-R with 10Gbe RoCE Express (cross platform) and
  - SMC-D with ISM on IBM z13™ or z13s™ (intra-CPC)
- SMC Applicability Tool (SMCAT) Introduction
- SMCAT Overview
- SMCAT Usage and Reporting Overview
- Backup
  1. SMCAT Configuration and Usage (starting / stopping) Details
  2. SMC-R and SMCAT Evaluation Considerations / References
  3. SMC-R Background
  4. SMC-D Background

Evaluating SMC applicability and benefits

As customers express interest in SMC-R and RoCE Express or SMC-D and ISM one of the initial questions asked is:

- "What benefit will SMC-R or SMC-D provide in my environment?"
  - Some users are well aware of significant traffic patterns that can benefit from SMC
  - But others are unsure of how much of their TCP traffic (in their environment) is:
    - z/OS to z/OS and
    - how much of that traffic is well suited to SMC (SMC-R or SMC-D)

- Reviewing SMF records, using Netstat displays, Ctrace analysis and reports from various Network Management products can provide these insights…

  This approach can be a time consuming activity that requires significant expertise.

SMC Applicability Tool Introduction

A new tool called SMC Applicability Tool (SMCAT) has been created that will help customers determine the value of SMC-R and SMC-D in their environment with minimal effort and minimal impact

- SMCAT is integrated within the TCP/IP stack:
  Gathers 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
  - Does not require RoCE hardware or SMC-R (SMC-D) function. No IP configuration changes are required (measures your existing TCP/IP traffic).

- Available via the service stream on existing z/OS releases:

  - V1R13 - APAR PI48309 PTF UI31050
  - V2R1, V2R2 - APAR PI48155, PTFs UI31054 (2.1) and UI31055 (2.2)

SMCAT Usage Overview

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 / SMC-D for)
  - If subnets are used, the entire subnet must be comprised of z/OS systems that are SMC-R / SMC-D 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 summary report is generated that includes:
  1. ***Percent of traffic eligible for SMC*** *(% of TCP traffic that is eligible for SMC)*
     - *All traffic that matches configured IP addresses (not using IPSec or FRCA)*
  2. ***Percent of traffic well suited for SMC*** *(your eligible traffic that is also "well suited" to SMC, 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 benefit (reduced latency, reduced CPU cost or both)*

SMCAT Usage Overview (continued)

The Summary Report includes 2 sections based on the specified IP addresses/subnets defined in SMCAT configuration file:

1. Total potential benefit:
   All eligible[1] TCP traffic that matched the specified SMCAT configuration (IP address/subnets). This total includes IP traffic (if present) that did not meet the direct IP route connectivity requirement.[2]

2. Immediate benefit:
   The eligible[1] TCP traffic that can use SMC immediately / as is (meets SMC direct route connectivity requirements).

   Detected by the tool automatically (non-routed traffic).

1. "Eligible" means TCP traffic not using IPSec or FRCA.
2. The total benefit report represents the opportunity to re-configure routing topology (single subnet) to enable SMC.

SMC Applicability Tool Sample Report (Direct Connections, part 1)

```
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
```

This portion of the report shows direct IP connections. This same report is repeated for indirect connections (different IP subnets).

How much of all of my TCP workload can benefit from SMC?

report continues…

SMC Applicability Tool Sample Report (Direct Connections, part 2)

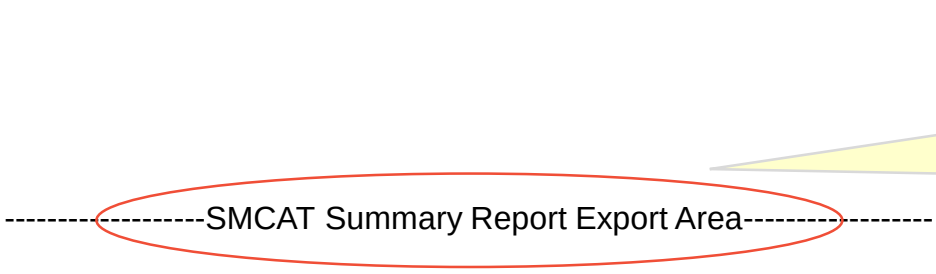
```
Application send sizes used for well-suited connections:
     Size                                  # sends      Percentage
     ----                                  -------      ----------
     1500 (<=1500):                           2000       39%
     4K (>1500 and <=4k):                     1500       29%
     8K (>4k and <= 8k):                         0        0%
     16K (>8k and <= 16k):                       0        0%
     32K (>16k and <= 32k):                      0        0%
     64K (>32k and <= 64k):                    500       10%
     256K (>64K and <= 256K):                 1090       22%
     >256K:                                     10        0%

   Application receive sizes used for well-suited connections:
     Size                                  # recvs      Percentage
     ----                                  -------      ----------
     1500 (<=1500):                           1500       32%
     4K (>1500 and <=4k):                     1250       27%
     8K (>4k and <= 8k):                         0        0%
     16K (>8k and <= 16k):                       0        0%
     32K (>16k and <= 32k):                    250        5%
     64K (>32k and <= 64k):                    500       10%
     256K (>64K and <= 256K):                 1200       25%
     >256K:                                      8        0%
```

The purpose of this information is to allow the user to better understand the characteristics of their workloads, what type of messages sizes are sent or received.

For the well suited connections, this portion of the report shows the **send message** sizes (broken into “bucket sizes”)

For the well suited connections, this portion of the report shows the **receive message** sizes (broken into “bucket sizes”)

SMC Applicability Tool Sample Report (All Connections, part 3)



At the end of the report all send / receive sizes (buckets) are copied into this unformatted export area to allow for easy exporting

The first two rows show the application send and receive sizes for the indirect connections. The second two rows show the data for the directly connected connections

If your report indicates your workloads exchanged large messages (message sizes => 8k) then you might realize CPU savings in addition to the latency savings.

Note. If you would like to better understand your potential CPU savings, then copy the entire report (including this export area) and send the data to IBM (dherr@us.ibm.com).

IBM will produce an estimate based on the data from your report(s).

Backup

Backup Topics:

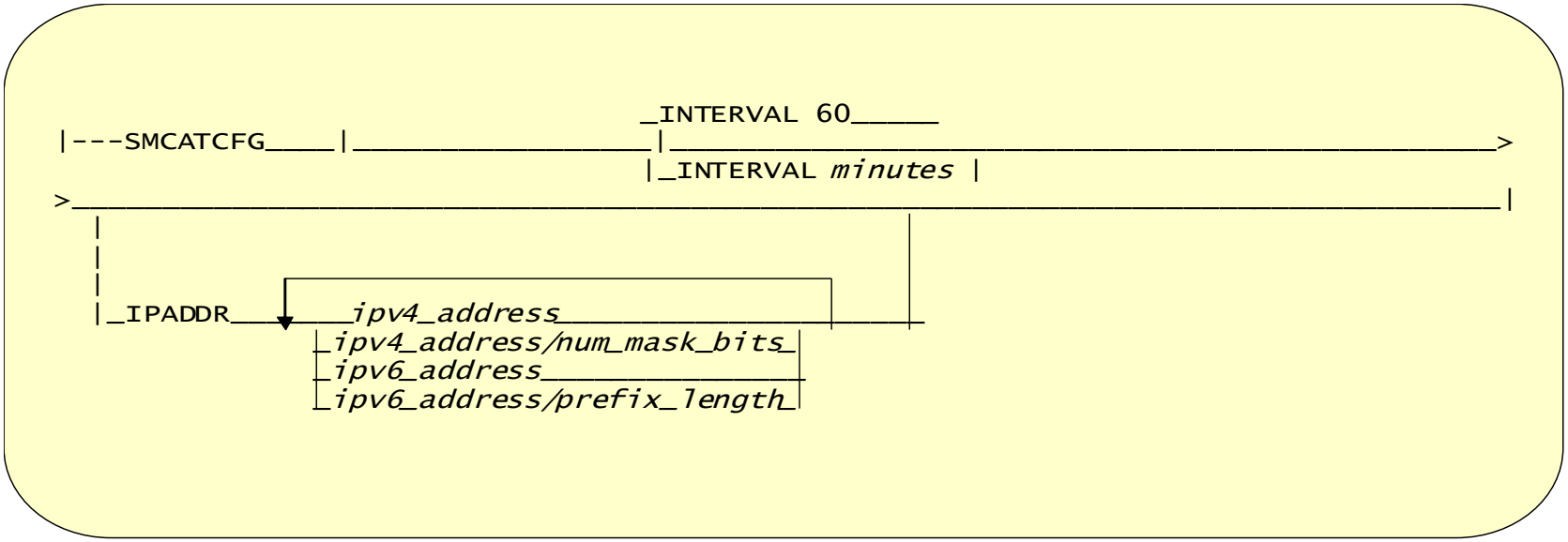
1. SMCAT Configuration and Usage (starting / stopping) Details
2. SMC-R and SMCAT Evaluation Considerations / References
3. SMC-R Background
4. SMC-D Background

Backup Topic 1 (Configuration and Usage Details)

Configuring the SMCAT Dataset

SMCAT data set configuration

- Interval defaults to 60 minutes
  Max interval is 1440 minutes (24 hours)
- IPADDR is a list of IPv4 and Ipv6 addresses and subnets
  256 max combination of addresses and subnets



SMCAT Dataset Example

```
SMCATCFG INTERVAL 120
IPADDR
C5::1:2:3:4/126
9.67.113.61
```

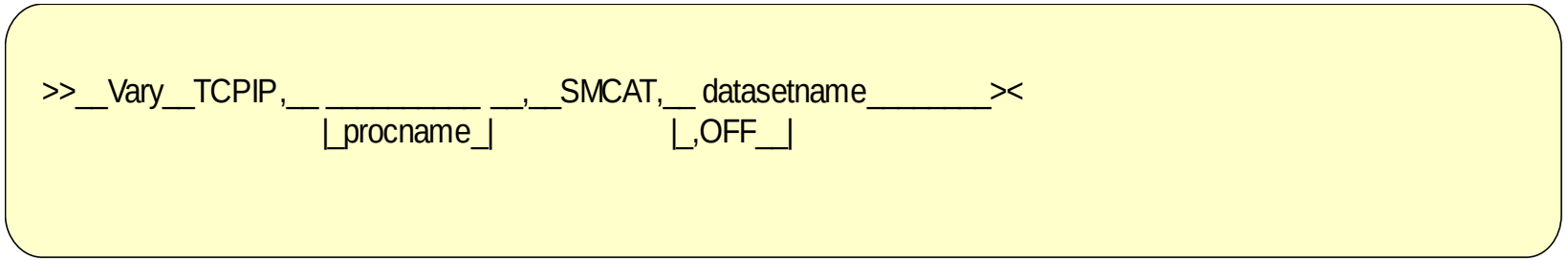
When SMCAT is started using this SMCAT configuration data set it will:

- Monitor TCP traffic for 2 hours for:
  - IPv6 prefix C5::1:2:3:4/126 and
  - IPv4 address 9.67.113.61

Starting and Stopping SMCAT

Vary TCPIP,,SMCAT command starts and stops the monitoring tool:

- datasetname value indicates that SMCAT is being turned on
- datasetname contains the SMCATCFG statement that specifies monitoring interval and IP addresses or subnets to be monitored
- OFF will stop SMCAT monitoring and generate report




```
VARY TCPIP,TCPPROC,SMCAT,USER99.TCPIP.SMCAT1
```

SMCAT Operator Messages

Key messages – Operator console:

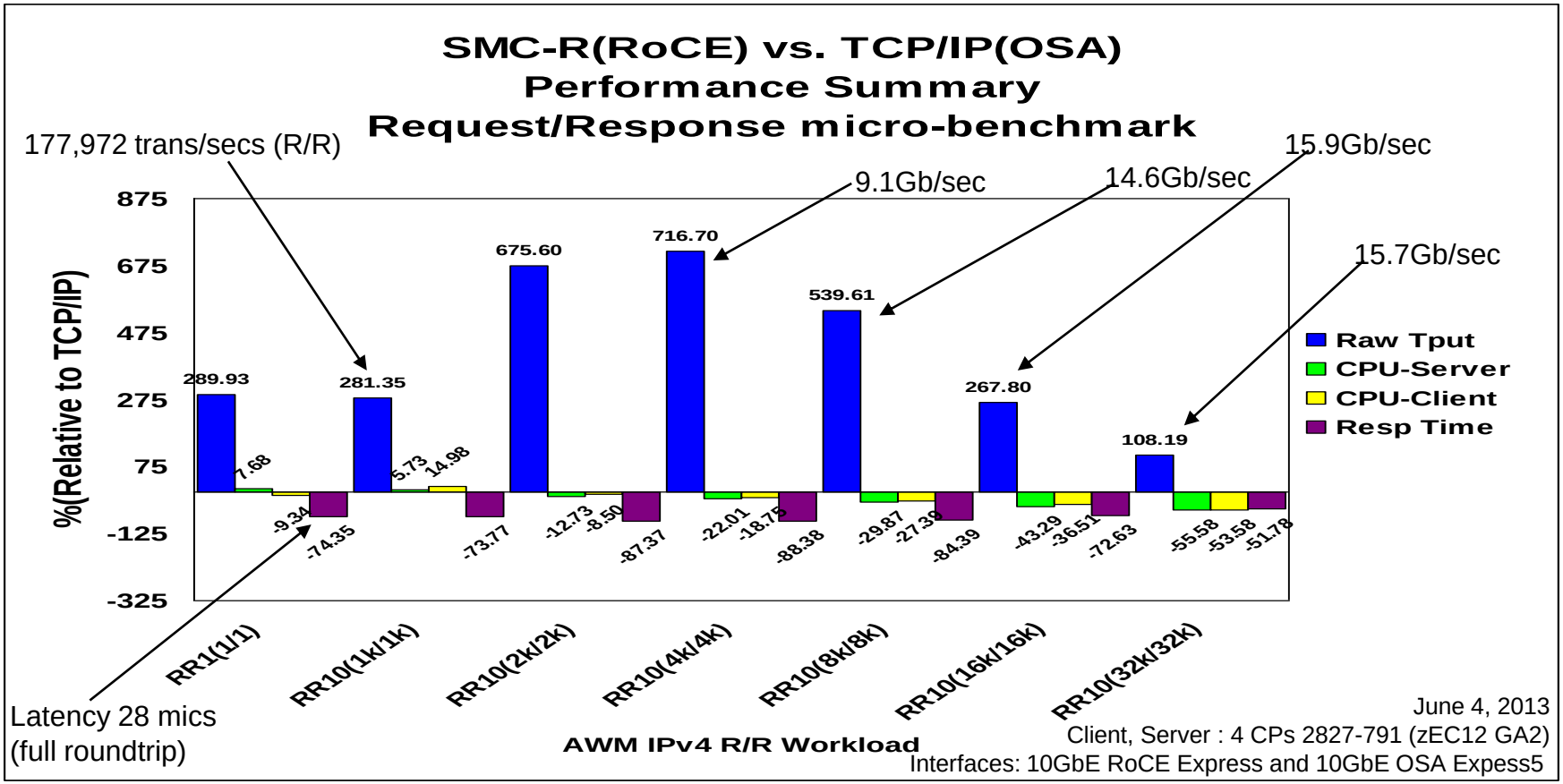
- EZD2031I SMC APPLICABILITY TOOL HAS STARTED COLLECTING DATA
- EZD2032I SMC APPLICABILITY TOOL HAS STOPPED COLLECTING DATA

Configuration information and the SMCAT report are sent to the system log

```
STC06578    EZD2040I TCP/IP CS V2R2 TCPIP Name: TCPIP
080         SMC Applicability Configuration Parameters - 02/04/2015, 10:09:49.08
080         Interval: 3 minutes
080         IP addresses/subnets being monitored
080
080         9.67.113.61
080         C5::1:2:3:4/126
080         End of configuration parameters
```

Backup Topic 2 (SMC-R and SMCAT Considerations)

Backup **SMC-R – Micro benchmark performance results**



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 and SMCAT Considerations

SMCAT Report Terminology:

- Eligible connections:
  TCP connections that match SMCAT configured IP addresses/subnets (IPSEC and FRCA connections are excluded)
- Well-suited connections:
  Eligible connections that have workloads ideally suited to exploit SMC-R (not short lived)

RoCE Distance (SMC-R) imposes limits on the distance between exploiting hosts:

- Typically deployed within a datacenter or site
- Can be deployed across sites for a distance up to 100 KM

  IBM System z® Qualified Wavelength Division Multiplexer (WDM) products for Multi-site Sysplex and GDPS® solutions

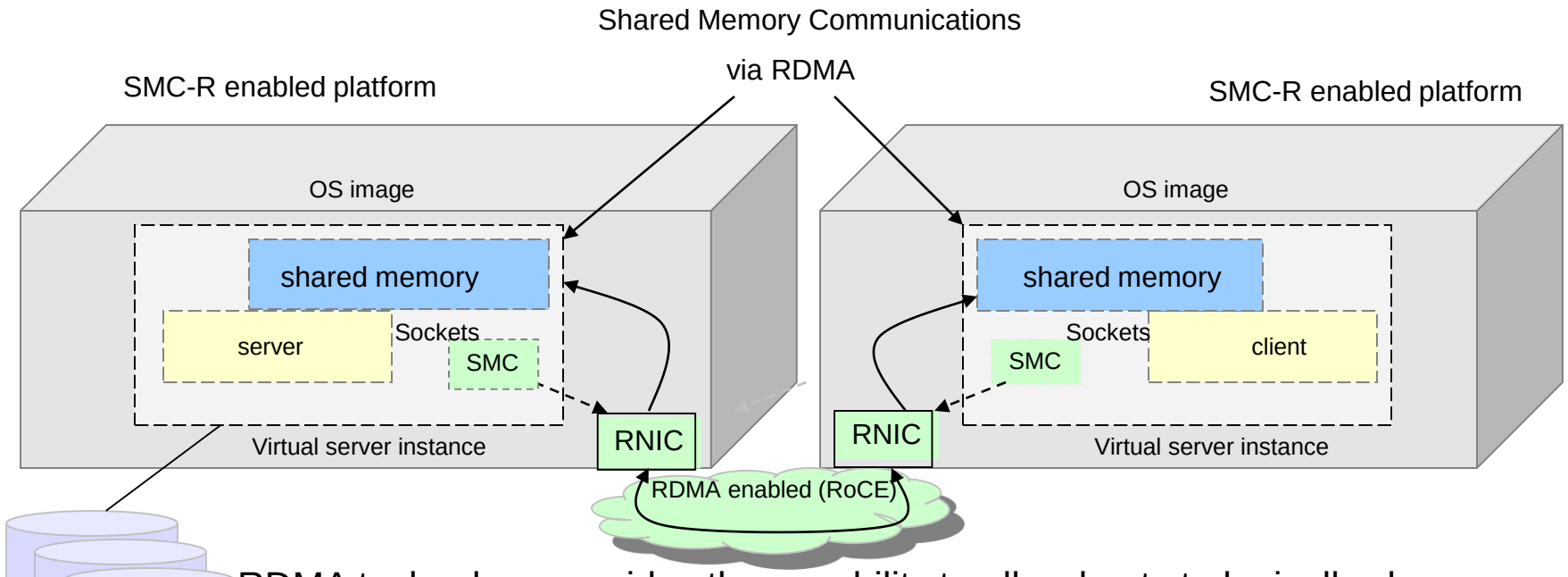
SMC-R and SMCAT Considerations

- Consider configuring a subset of IP addresses for similar workloads/similar servers and then extrapolate the SMCAT results of your sample across your enterprise data center
- SMCAT does not take into consideration your type of network attachment (e.g. OSA, XCF, HiperSockets, etc.). SMC-R only applies to RoCE (Ethernet). In order to exploit SMC-R only the TCP connections that use OSA are capable of using SMC-R.
- Authorize users to invoke the command by permitting user IDs for CONTROL access to the RACF profile name MVS.VARY.TCPIP.SMCAT
- The configuration data set must be cataloged and fully-qualified
  - Specified without any quotation marks
  - Can be either a sequential data set or a member in a PDS
  - Cannot be a z/OS UNIX file
- Communications Server: IP System Administrator's Commands
  - VARY TCPIP,,SMCAT section
- SMC-R references:
  http://www-01.ibm.com/software/network/commserver/SMCR/

Backup Topic 3 (SMC-R Background)

"Shared Memory Communications over RDMA" concepts

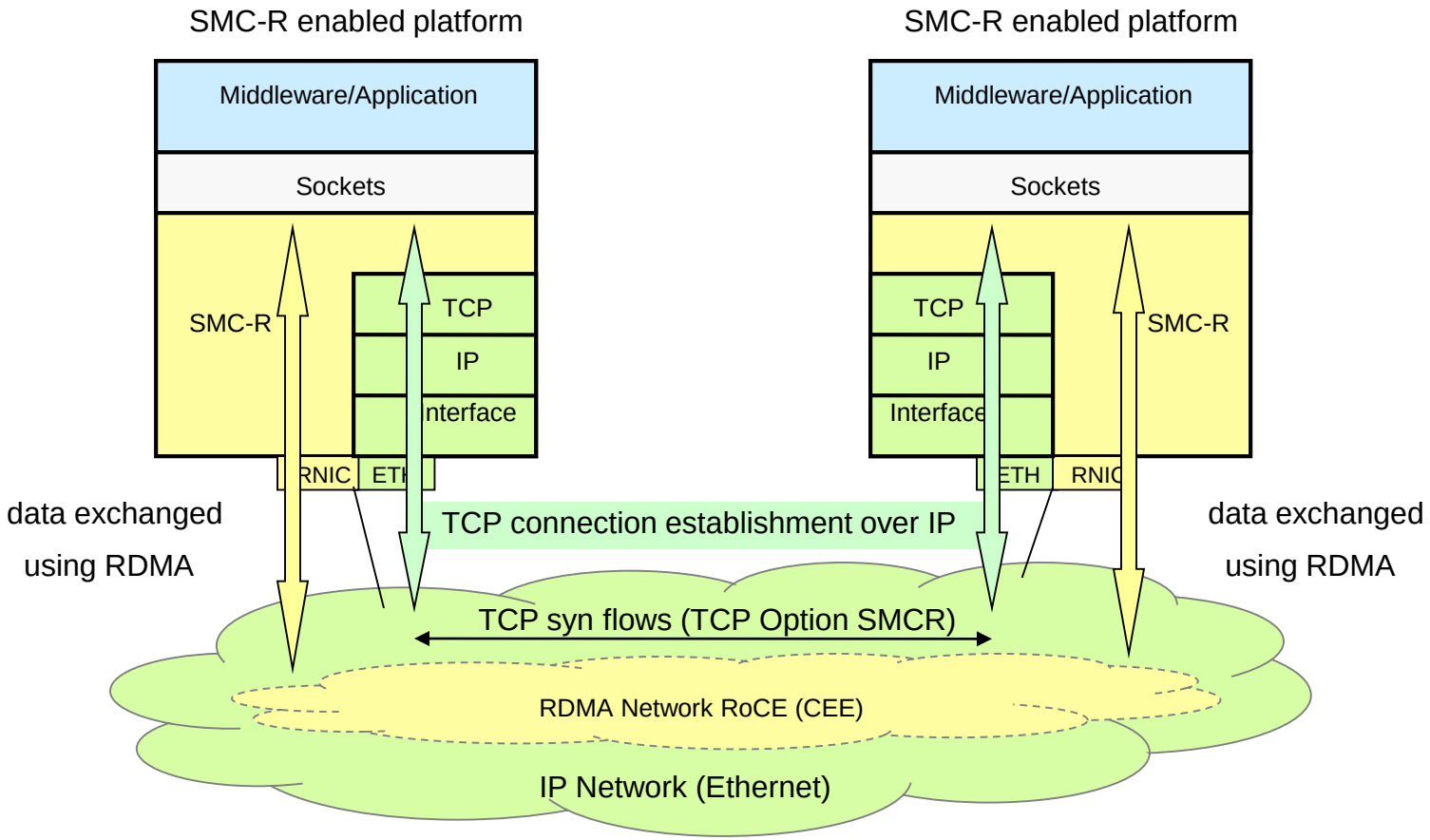
Clustered Systems



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

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

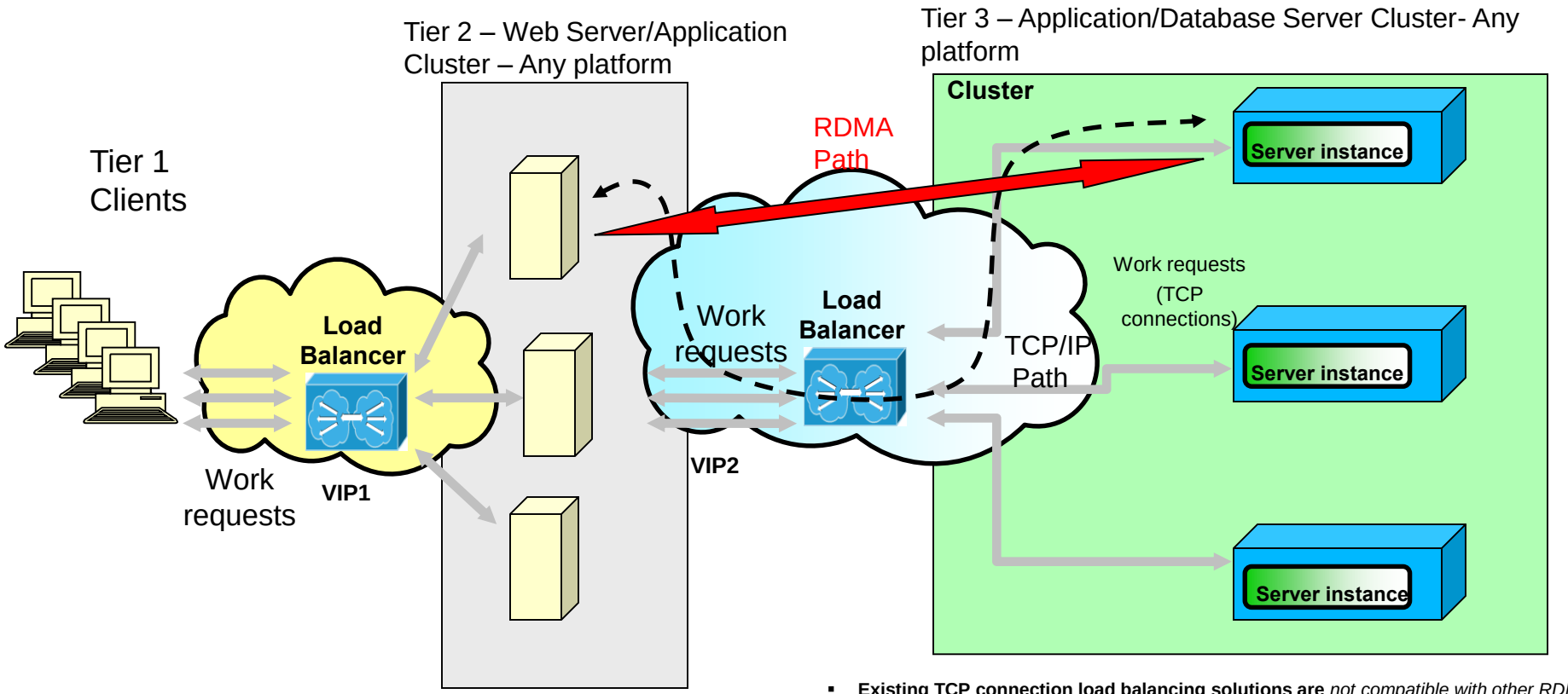
Dynamic Transition from TCP to SMC-R



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

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

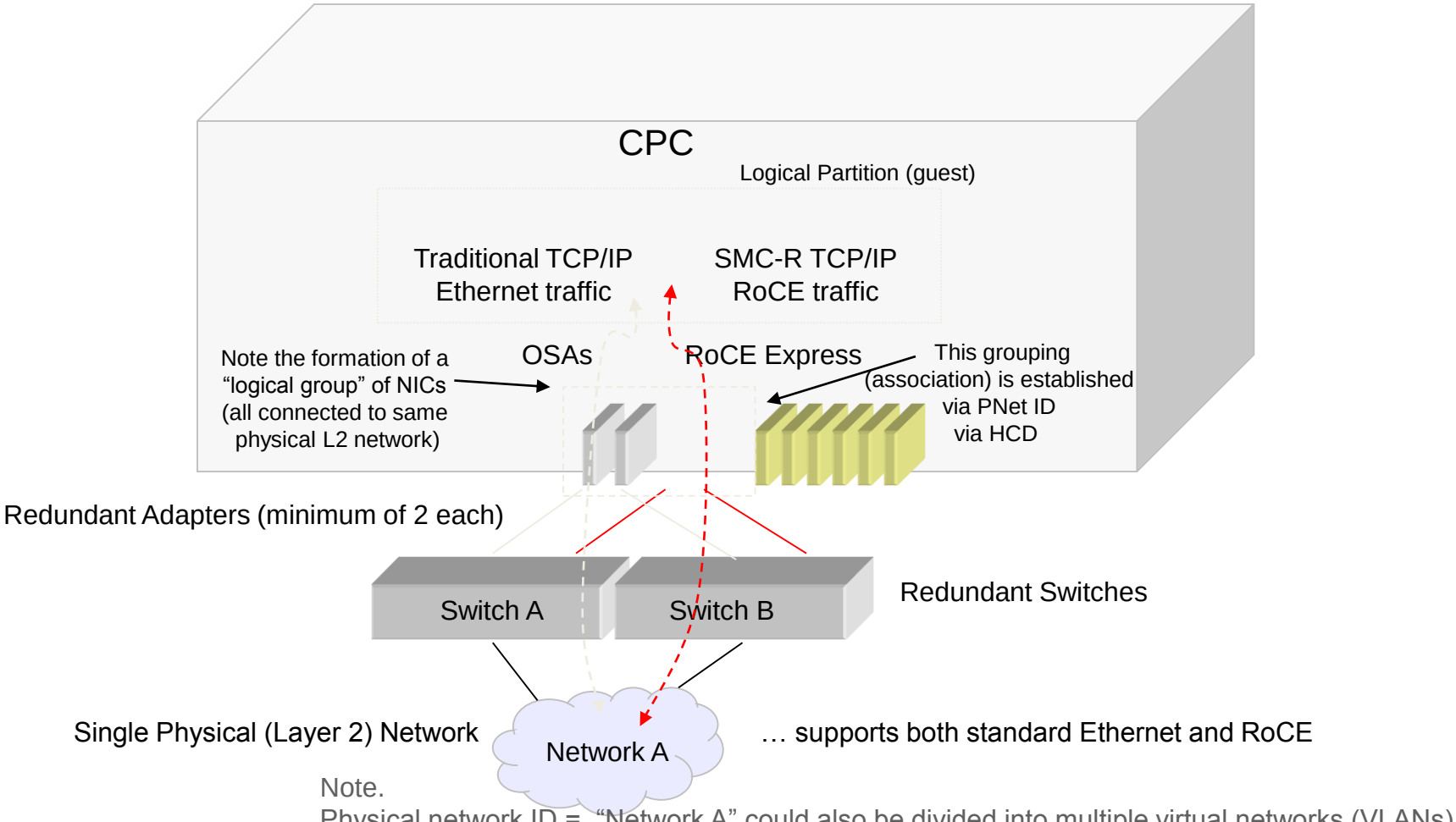
Server clustering and TCP connection load balancing



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

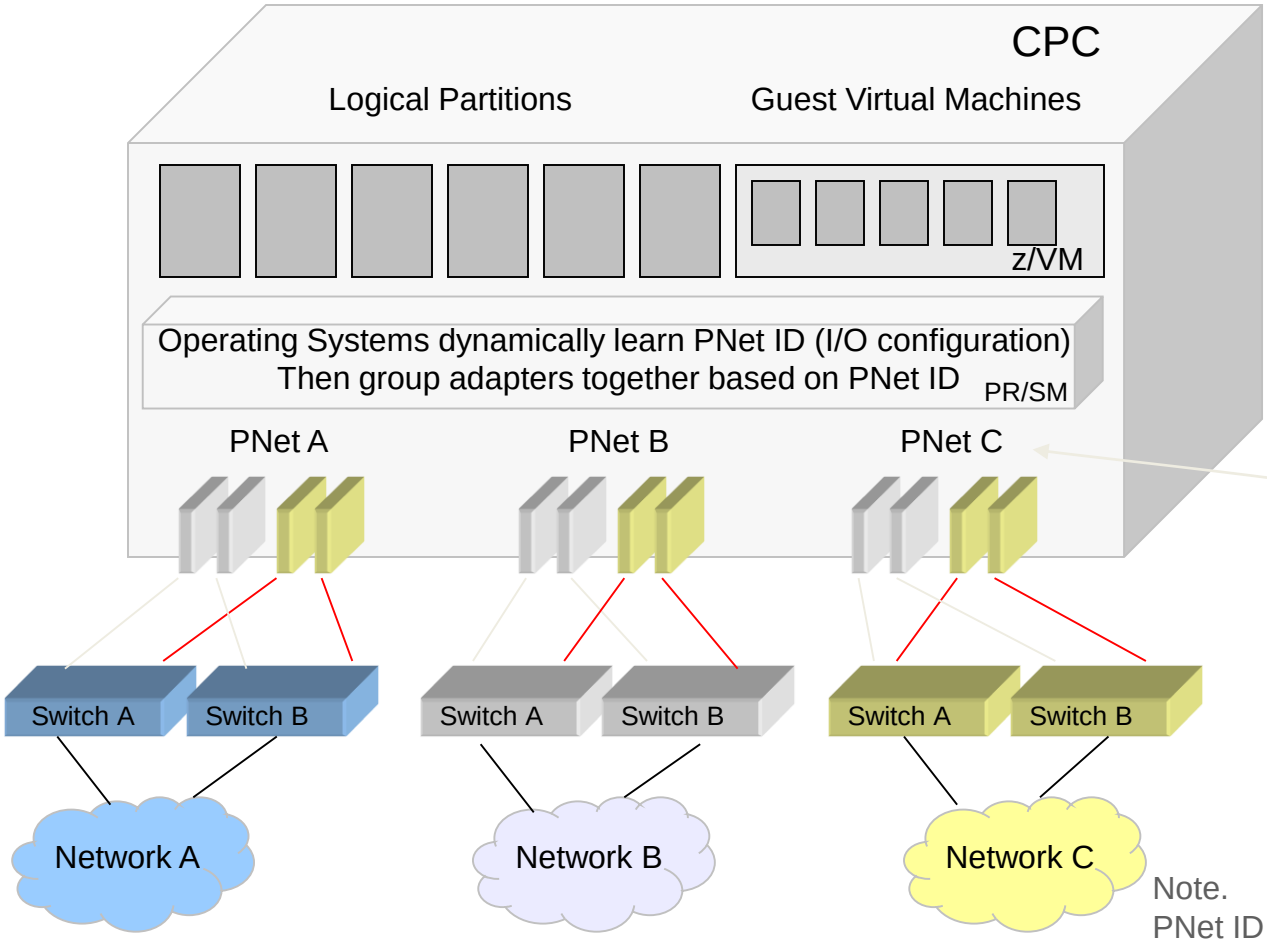
- **Existing TCP connection load balancing solutions are** *not compatible with other RDMA solutions*
- They are not aware of the RDMA protocol ***AND*** RDMA flows ***can not*** flow through intermediate nodes
- **The SMC-R protocol allows existing TCP load balancing solutions to be deployed** *with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

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



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

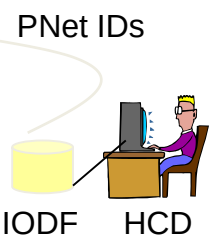
Multiple Physical RoCE Networks (Physical Network IDs)
Associating 10GbE RoCE Express features with their physical networks



Administrative model. . .

How to configure PNet ID?

(Operating Systems learn)



PCHID
PNet ID

Note.
PNet ID is a customer defined ID

("IEDN" is a reserved ID)

Three unique (physically isolated) Layer 2 Physical Networks

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

Summary

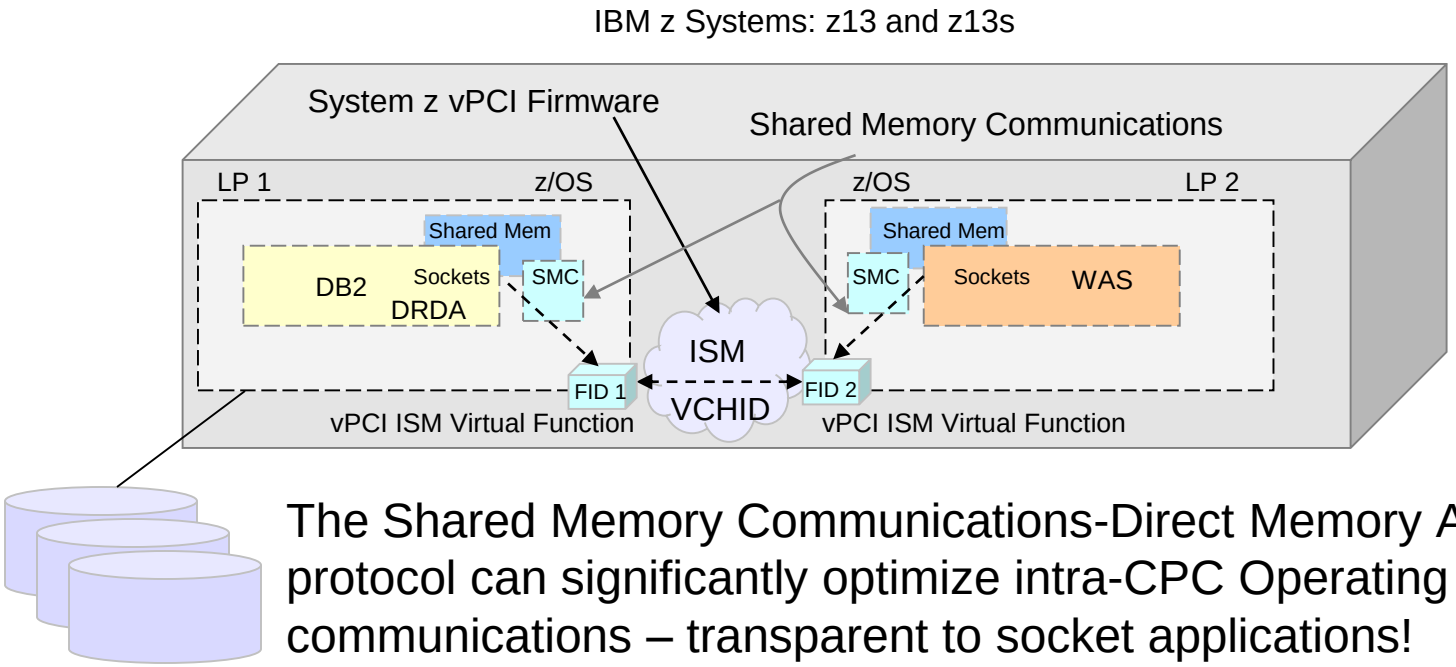
- z/OS application workloads transparently exploit 10GbE RoCE Express feature using z/OS V2R1 SMC-R. When SMC-R is enabled:
  - transactional workloads (WAS, CICS, IMS, MQ, etc.) can potentially see an increase in their overall transaction rate (i.e. transactions per second) with a slight savings in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

Value

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

Backup Topic 4 (SMC-D Background)

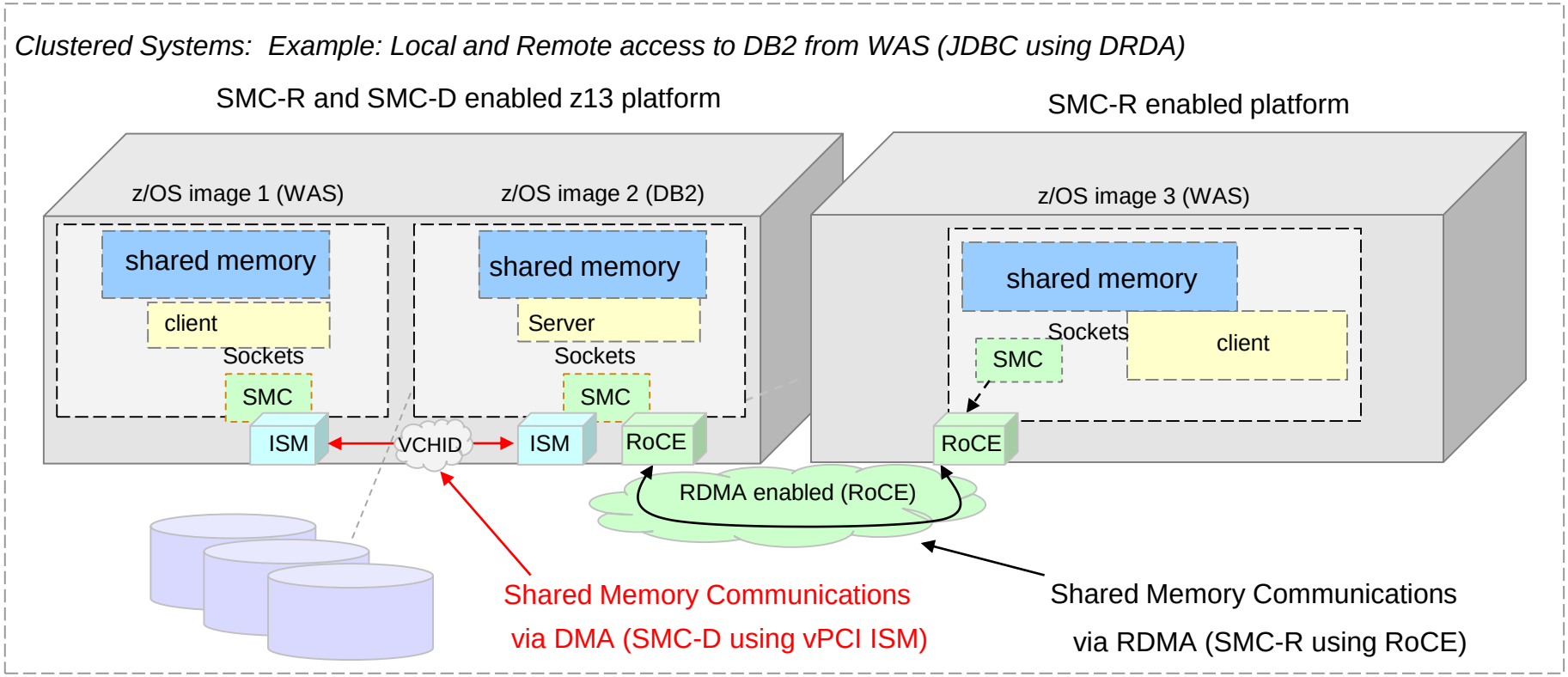
SMC-D over ISM:
Internal Shared Memory vPCI Function with ISM VCHIDs



The Shared Memory Communications-Direct Memory Access (SMC-D) protocol can significantly optimize intra-CPC Operating Systems communications – transparent to socket applications!

- Tightly couples socket API communications / memory within the CPC.
- Eliminates TCP/IP processing in the data path.
- ISM is a z System firmware solution (leveraging existing OS virtual memory and does require additional hardware).

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



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)

SMC-AT for SMC-D

- SMC-AT can also be used to evaluate the benefits of SMC-D for intra-CEC network traffic.
- Additional information about SMC-D and ISM can be found on the SMC website.

Thank You