

#### IBM z/OS 2.4

**SMC Version 2 Introduction** 

SMC-Dv2 and ISMv2

October 15, 2020

Randall Kunkel kunkel@us.ibm.com







## SMCv2 Agenda Topics

- 1. SMC Review
  - 1. SMC-R and RoCE Express2
  - 2. SMC-D and ISM
  - 3. SMC Performance Benefits
  - 4. SMC on other platforms
  - 5. SMC limitation single IP subnet
- 2. SMC Version 2 Introduction
  - 1. RoCEv2 (Routable RoCE)
  - 2. SMC V2 Enterprise ID SMCv2 Connection Eligibility (User Defined and System EIDs)
  - 3. SMC-Dv2 (ISMv2) Introduction
  - 4. SMC-Dv2 Enablement
  - 5. SMC-Dv2 Verification
  - 6. SMC-Dv2 NCA panels (appendix)





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

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



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 SMC-R with RoCE and within System z SMC-D with 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)

© 2020 IBM Corporation

# Dynamic Transition from TCP to SMC-R (RFC7609)



© 2020 IBM Corporation



## 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



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



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



© 2020 IBM Corporation







#### 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.



#### HiperSockets Comparison

Same benchmarks as shown on the previous chart summarized here in 3 categories:

1. transactional small

- 2. transactional large
- 3. streaming

Note. R/R (request /response) = transactional

#### 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.

#### SMC-R / SMC-D Value: Superior Performance (CPU / Latency / Throughput)

	SMC-R (vs. OSA)	SMC-D (vs. HiperSockets)	
Network Latency	Network latency for z/OS TCP/IP based OLTP workloads reduced by up to 80% <sup>1</sup>	Network latency for z/OS TCP/IP based OLTP workloads throughput increased by up to 91% with up 48% reduction in latency <sup>2</sup>	
Network Related CPU cost	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%1	Networking related CPU consumption for z/OS TCP/IP based workloads with streaming data patterns CPU reduced by up to 88% with a network throughput increase of up to 9X <sup>2</sup>	

1. SMC-R: based on internal IBM benchmarks in a controlled environment of modeled 2/OS TCP sockets-based workloads with request/response traffic patterns and bulk patterns using SMC-R (10GbE RoCE Express feature) vs TCP/IP (10GbE OSA Express feature) vs TCP/IP (10GbE OSA Express feature). The actual response times and CPU savings any user will experience will vary

2. SMC-D based on internal IBM benchmarks in a controlled environment of modeled z/OS TCP sockets-based workloads with request/response traffic patterns and bulk patterns using SMC-D (ISM) vs TCP/IP (HiperSockets). The actual response times and CPU savings any user will experience will vary.



## SMC-D Key Attributes - Summary

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



## SMC and Other Platforms

#### 1. Linux SMC (SMC-R and SMC-D) is now available on:

- 1. RHEL 8
- 2. SLES 12 SP4, SLES 15 SP1
- 3. Ubuntu 18.10

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

#### 2. AIX: System P with AIX 7.2 SMC-R

https://www-

01.ibm.com/common/ssi/cgibin/ssialias?infotype=an&subtype=ca&appname=gpateam&supplier=872&letternum=ENUSAP17-0480



#### SMC is Limited to a Single IP Subnet - RoCE is not Routable



#### Note.

SMC-R (RoCE) and SMC-D (ISM) follow the same base SMC protocol and rules. The single subnet connection eligibility (limitation) applies to both forms of SMC.





17

# Each z/OS Instance - unique IP Subnets – same CPC





#### Groups of Systems separated on Unique Subnets – Same CPC





### SMC Version 2

#### Introduction







#### RoCEv2 - "IP Routable RoCE"







# SMCv2 defines the specifications for SMC over multiple IP subnets.

#### **IBM will deliver the SMCv2 solutions in multiple steps:**

#### 1. SMC-Dv2 / ISMv2 (z/OS 2.4 PTFs and IBM z15) Today's presentation – see z/OS V2R4 3Q 2020 RFA:

https://www-01.ibm.com/common/ssi/ShowDoc.wss?docURL=/common/ssi/rep\_sm/s/897/ENUS5650-ZOS/index.html&lang=en&request\_locale=en

# 2. SMC-Rv2: see SoD (next chart and) in the z/OS V2R4 3Q 2020 RFA



Statement of Direction: Shared Memory Communications v2 (SMCv2) - RDMA over Converged Ethernet v2 (RoCEv2) and Linux(R) support (Issued September 22, 2020)

Today, SMC for both SMC-R and SMC-D is limited to communications for hosts attached to a common IP subnet. SMCv2 defines SMC over multiple IP subnets. The SMCv2 multiple IP subnet support extends SMC capability to additional application workloads that were previously ineligible for SMC. z/OS V2.4 delivers SMCv2 multiple IP subnet capability for SMC-D (SMC-Dv2). IBM plans to make SMCv2 multiple IP subnet capability available for SMC- R exploiting "routable RoCE" (RoCEv2) in a future z/OS deliverable. IBM is working with Linux distribution partners to provide SMCv2 support for Linux on IBM Z and **IBM** LinuxONE.



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



#### Software: z/OS 2.4 APARs

- 1. PH22695 (TCP/IP)
- 2. OA59152 (VTAM)
- 3. PH25898 (NCA)
- 4. OA59235 (z/OS IOS)

Note. "Down-level systems" (systems without SMC-Dv2 support) require an SMCv2 toleration update:

- z/OS toleration: PH17556
- Linux toleration (compatibility): https://linux-on-z.blogspot.com/p/smc-for-linux-on-ibm-z.html

#### Hardware:

The new ISMv2 capability is available on IBM z15.

- For IBM z15 T01, refer to the MCL number P46601.067 driver D41C.
- The ISMv2 support is in the base of the IBM z15 T02.

Note. There are no configuration changes required in HCD to define or to enable ISMv2.



## SMC V2 Enterprise ID (EID)

SMCv2 defines the SMC specifications for multiple IP subnet support (applicable to SMC-R and SMC-D).

IP Subnet is no longer the SMC connection criteria, SMC V2 connections can span multiple IP subnets

SMCv2 (will include) RoCEv2 is designed for the "Enterprise Data Center" – not for the WAN

SMC V2 connection eligibility:

- Systems are formed into a logical group by the administrator defining a common ID in each system the **SMC V2 Enterprise ID (EID)** defines V2 eligibility
- $\,\circ\,$  EIDs are dynamically exchanged during SMC connection setup
- $_{\odot}\,$  Systems with a common EID can connect using SMC V2  $^{1}$
- The logical "group" could be smaller groups (business units, system roles, etc.), the entire data center or multiple ("close proximity") data centers etc.
  - 1. Defining an EID enables SMCv2. Multiple EIDs can be configured and dynamically exchanged with the peer host. Other connection criteria must also be met (see the SMC-Dv2 connection eligibility on chart 30).



#### **EID Concepts**





SMC CLC Confirm (EID-B)

hosts can dynamically determine connection eligibility - When EIDs match – the connection is V2 eligible

## EID Example 1 – Centers at Close Proximity (Single EID)



## EID Example 2 – Centers at Distance (multiple EIDs)





SMC-D Version 1 (ISMv1) – Single Subnet (review)



The ISM architecture mirrors your network topology (the "physical Layer 2 network"). All SMC connectivity is based on same subnet. ISM can not connect ("route") across subnets (shown as VLANs here).



### Enabling SMC-Dv2

- 1. Defining an SMCv2 **Enterprise ID (EID)** enables the stack for SMCv2.
- 2. There are two types of EIDs:
  - 1. User defined EIDs allows user defined naming conventions
  - 2. System EID one per CPC system, auto defined applies to SMC-Dv2 only
- 3. Example:

GlobalConfig Statement: SMCGlobal: smceid EID-A ! Specify my user defined EID-A

... and / or ...

GlobalConfig: SMCD: SystemEID

! Enable System EID! Auto defined EID for this CPC for SMC-Dv2 only

Note. NCA support is available for SMC-Dv2 (see appendix)



## SMC-Dv2 Connection Eligibility

The SMC-Dv2 and ISMv2 multiple IP subnet connection eligibility requirements are summarized as follows, both the TCP client and server LPARs (guests) must:

- 1. Be updated to support SMC-Dv2
- 2. Execute on the same IBM Z CPC that supports ISMv2 (z15)
- 3. Must have access to a common ISM VCHID
- 4. Must be SMCv2 enabled, defined with the same EID (user defined EID or System EID)
- 5. Must not be associated with a TCP/IP connection that requires IPSec encryption

Note.

SMC-D supports IP connectivity over OSA or HS (no change for SMC-Dv2). TCP/IP connections over multiple IP subnets will typically be external connections over OSA.



## SMC-Dv2 with ISMv2 (z15) – SMC-Dv2 over multiple subnets



In this example LPs 1 and 4 can form TCP/IP connections over multiple subnets (shown as VLAN1 and VLAN 2) and the TCP connections are now eligible for SMC-Dv2 (multiple subnets). The multiple subnet connections will use the ISMv2 internal connection. But... how could LPs 4 and 5 connect over ISM?



# ISMv2 – Enables Multi-Subnet Internal Connectivity

SMC-Dv2 and ISMv2 Introduces two new ISMv2 Concepts:

 New type of Internal (ISMv2) Connectivity not associated with Layer 2 (ISM connections not related to any VLANs or subnets – "Internal ISMv2 -Layer 3" connections)

and...

- 2. ISM device association: how ISM devices are associated with an IP device by the OS (software) a difference in the SMC-Dv2 protocol:
  - Associated ISM VCHIDs ISM VCHIDs associated with an IP device (OSA or HS) based on matching PNet IDs (existing SMCv1 and ISMv1 support also applies to SMC-Dv2):

#### 2. Unassociated ISM VCHIDs:

ISM VCHIDs unassociated with any IP device – ISM VCHIDs defined without PNet ID (**New** SMCv2 and ISMv2 support)



# SMC-Dv2 (ISMv2) – Unassociated ISM VCHID (No PNet ID)









#### SMC Version 2

#### Enablement - Samples







# Enabling SMC-Dv2 – Global Config Example (Define 4 EIDs and System EID)



# Enabling SMC-Dv2 – Verify Global Config Settings (netstat config)

SMCGLOBAL: AUTOCACHE: NO AUTOSMC: YES SMCEID: A1.E5BCDEFGHIJKLMNOPQRSTUVWXYZ32 SMCEID: B-USE.THIS.ONE. SMCEID: C3.G7 SMCEID: D4-H890 SMCR: YES FIXEDMEMORY: 0000256M TCPKEEPMININT: 00000300 PFID: A003 PORTNUM: 1 MTU: 1024 PFID: A004 PORTNUM: 1 MTU: 1024 SMCD: YFS FIXEDMEMORY: 0000256M TCPKEEPMININT: 00000300 SMCDSYSEID: YES (IBM-SYSZ-ISMSEID00000000CEB88561)



In this example: four unassociated ISM VCHIDs and 1 associated ISM VCHID are shown Note that z/OS supports up to four unassociated ISM VCHIDs.



•

•

# Display an active TCP connection using SMC-Dv2

D TCPIP,TCPCS3,N,ALL,PORT=21 is from SMCv2,

LOCAL SOCKET: 172.16.1.3..1042 FOREIGN SOCKET: 172.16.1.5..21

SMC INFORMATION:

SMCDSTATUS:ACTIVESMCDVERSION:2LOCALSMCLINKID:42010002REMOTESMCLINKID:42020002LOCALSMCRCVBUF:64KREMOTESMCRCVBUF:64KSMCEID:B-USE.THIS.ONE.64K



Display ISMv2 Connection (Netstat DEVL, SMC) – Part 1

D TCPIP, TCPCS3, N, DEVL, SMC			
EZD0101I NETSTAT CS V2R4 TCPC	CS3 134		
INTFNAME: EZAISM01 I	INTFTYPE: ISM	INTFSTATUS:	READY
PFID: 0051 TRLE: IUT00051	PFIDSTATUS: READY		
PNETID: NETID1			
GIDADDR: 8003000100010051			
INTERFACE STATISTICS:			
BYTESIN	= 2434361		
INBOUND OPERATIONS	= 48		
BYTESOUT	= 2433740		
OUTBOUND OPERATIONS	= 52		
SMC LINKS	= 1		
TCP CONNECTIONS	= 1		
INTF RECEIVE BUFFER INUSE	= 64K		
DEVICE INTERRUPTS	= 48		

Continues next page . . .



# Display ISMv2 Connection (Netstat DEVL, SMC) – Part 2

SMCD LINK INFORMATION: LOCALSMCLINKID: 42010001 REMOTESMCLINKID: 42020001 SMCDVERSION: 2 SMCEID: B-USE.THIS.ONE. VLANID: N/A LOCALGID: 8003000100010051 REMOTEGID: 8002000200020052 REMOTEHOSTTYPE: Z/OS **REMOTEHOSTNAME:** TCPCS5 SMCLINKBYTESIN: 2434361 SMCLINKINOPERATIONS: 48 2433740 SMCLINKBYTESOUT: SMCLINKOUTOPERATIONS: 52 TCP CONNECTIONS: 1 LINK RECEIVE BUFFER INUSE: 64K 64K 64K BUFFER INUSE:



•

## Display ISMv2 Connection (Unassociated ISM) – Part 1

D TCPIP, TCPCS3, N, DEVL, SMC EZD01011 NETSTAT CS V2R4 TCPCS3 134

INTFNAME: EZAISMU1 INTFTYPE: ISM INTFSTATUS: READY PFID: 0071 TRLE: IUT00071 PFIDSTATUS: READY PNETID: \*NONE\* GIDADDR: 8003001900190071 INTERFACE STATISTICS: BYTESIN = 4867594INBOUND OPERATIONS = 71 = 4867179BYTESOUT = 78 OUTBOUND OPERATIONS SMC LINKS = 1 TCP CONNECTIONS = 1 INTF RECEIVE BUFFER INUSE = 64KDEVICE INTERRUPTS = 71

Continues next page ...



SMCD LINK INFORMATION: LOCALSMCLINKID: 42010002 REMOTESMCLINKID: 42020002 SMCDVERSION: 2 SMCEID: B-USE.THIS.ONE. VLANID: N/A LOCALGID: 8003001900190071 REMOTEGTD: 8002001A001A0072 **REMOTEHOSTTYPE:** Z/OS REMOTEHOSTNAME: TCPCS5 SMCLINKBYTESIN: 4867594 SMCLINKINOPERATIONS: 71 4867179 SMCLINKBYTESOUT: SMCLINKOUTOPERATIONS: 78 TCP CONNECTIONS: 1 LINK RECEIVE BUFFER INUSE: 64K 64K 64K BUFFER TNUSE:



## Displaying the ISM TRLE



# Displaying OSA or HS Interfaces (Specific OSA Interface)

D TCPIP, TCPCS3, N, DEVL, INTFN=QDIO4103L EZD01011 NETSTAT CS V2R4 TCPCS3 240 INTFNAME: QDIO4103L INTFTYPE: IPAQENET INTFSTATUS: READY

ASSOCIATED ISM INTERFACE: EZAISM01 UNASSOCIATED ISM INTERFACES: EZAISMU1 EZAISMU2 EZAISMU3 EZAISMU4

.... now includes both associated and unassociated ISM interfaces



SMC-Dv2 Miscellaneous Information

- 1. NMI and SMF updates: Reference the IP Programming Manual (minor updates)
- 2. IBM SMCv2 Overview: SMC-Dv2 / ISMv2 White Paper:

https://www.ibm.com/support/knowledgecenter/SSLTBW\_2.4.0/com.ibm.zos.v2r4.halz002/smc\_reference.htm

3. SMC-AT: Using your existing output... The 2.4 APAR includes minor changes in the SMC-AT report to clarify what could be achieved with SMCv2.



#### Questions – Comments?

Q1. What about SMC-Dv2 performance?

A1. No changes.

SMC-Dv2 changes the connection protocol but the ISM data transfer technology is identical.

Q2. What about network security?

A2. No changes.

SMC-D is based on your TCP/IP connection authentication and existing encryption (such as TLS). After connection setup, the user data transfer remains within the CPC.

Q3. Is it really that simple – just define an EID and go? A3. Depends... see next Q&A

- Q4. What about ISM VCHIDs (associated or unassociated)? How do I decide?
- A4. It depends on where you're coming from... existing SMC-D user or a new user, the level of the Operating Systems you need to support, what type of isolation is required etc. The IP Config Guide provides some things to think about.

Q5... other?



# Thank You!



© 2020 IBM Corporation



#### Appendix – NCA Panels (SMC-Dv2)

### Locating SMC settings in the NCA TCP/IP stack configuration

Network Configuration Assistant (Home) > TCP/IP Profile > TCP/IP Profile : PLEX.IMAGE3.STACK3

#### TCP/IP Profile for Group PLEX, System Image IMAGE3, Stack STACK3

#### Configure

Use the following links to create and modify TCP/IP resources to define this stack's profile configuration.

TCP/IP Stack Resources	Status
Interfaces: Attach to networks	Configured
Routes: Connect to other systems	Configured
Ports: Reserve ports for TCP/IP applications	Not configured
Security: Control network access to and from the System	Not configured
Source IP Addressing: Control outbound connection source IP addressing	Configured
Performance and Protocol: Tune your TCP/IP stack	Configured
Management and Traces: Enable TCP/IP stack systems management and diagnosis	Not configured

As part of this update, all SMC configuration has been consolidated into one link on this panel. See next slide for more

#### Network Configuration Assistant (Home) > TCP/IP Profile > TCP/IP Profile : PLEX.IMAGE3.STACK3 > Pe

#### **Configure Performance and Protocol Properties**

Global Settings and Configuration Status

Settings	Status
TCP protocol settings	Not configured
UDP protocol settings	Not configured
IP protocol settings	Not configured
Shared Memory Communications Settings	Configured
Global network device layer optimization settings	Not configured
TCP/IP storage tuning	Not configured



		And then you ena	ble SMCv2 and set	its parameters in
You turn on SMC-D in this tab		this tab		
Network Configuration Assistant (Home)  TCP Memory Communications Settings Shared Memory Communications S	'IP Profile ► TCP/IP Profile : Settings	PLEX.INAGE3.STACK3   Performa	ance and Protocol ▶ Shared Help	
General settings SMC-D settings	SMC-R settings SMCv2	settings		
Shared Memory Communications - Dire Global Property Setting: Customize the following property. Configurat ✓ Enable SMC-D Maximum 64-bit fixed memory used for buf (n) Minimum interval to send TCP keep-alive p (s)	ct Memory Access (SMC- on will be generated to enabl iers negabytes) Range is 30-999 ackets over the TCP path seconds) Range is 0 - 21474	<ul> <li>D) settings</li> <li>e or disable this property.</li> <li>P). Default is 256.</li> <li>160. Default is 300.</li> </ul>		See next slide for configuration on this tab
				© 2020 IBM Corporation



OK

Cancel

Network Configuration Assistant (Home) > TCP/IP Profile > TCP/IP Profile : PLEX.IMAGE3.STACK3 > Performance and Protocol > Shared Memory Communications Settings Help Shared Memory Communications Settings General settings SMC-D settings SMC-R settings SMCv2 settings Shared Memory Communications version 2 (SMCv2) settings configure Enterprise IDs (EIDs) to enable SMCv2, which enables SMC across multiple subnets (z/OS V2R4 or later) User-defined enterprise ID (EIDs) Actions -No filter applied **Origin Reusable Configuration EID Value** Description Filter Filter Filter There is no data to display. Total: a Selected: a Global Property Setting Systomize the following property. Cenfiguration will be generated to enable or disable this property. Generate a system Enterprise ID (EID) for use with SMC-Dv2 Tip: The system EID is an internal EID that is built by the SMCv2 software stack, which represents the CPC that the operating system is operating on, Generating it guarantees all LPARs on the same CPC that support SMC-Dv2 and generate a system EID will be able to communicate using that CPC's system EID.

If you want to code userdefined EIDs, select this item to enable the User-Defined EIDs table below

Then use Actions->New to create user-defined EIDs. You can also modify, delete, copy, etc from the Actions menu on this table

To create a system-defined EID for SMC-Dv2, select "Customize the following property" and then select "Generate a system Enterprise ID..." as shown here.



		۱.	
			/
		-	
		۲	