Shared Memory Communications – Direct Memory Access (SMC-D) Frequently Asked Questions

Contents

1. What is SMC-D?

Question:

1.1. What is Shared Memory Communications – Direct Memory Access (SMC-D)?

Answer:

Shared Memory Communications – Direct Memory Access (SMC-D) leverages the existing Shared Memory Communications protocol used over RDMA (SMC-R) to provide highly optimized inter-system operating system communications. Instead of using RoCE, SMC-D uses Internal Shared Memory (ISM) technology within the system. ISM provides adapter virtualization (virtual functions) to facilitate the internal communications. SMC-D does not require any additional physical hardware (provided with firmware level 27 within IBM z13 and base for IBM z13s) – no adapters, card slots, switches, fabric management, PCI infrastructure – so there are cost savings from using the ISM capability. SMC-D can be enabled in z/OS V2.2 (PTF) with a single TCP/IP profile keyword[1]. ISM is provided with the IBM z13s and IBM z13 (firmware level 27).

Question:

1.2. It sounds like you just described HiperSockets – does SMC-D replace HiperSockets?

Answer:

SMC-D (like SMC-R) still requires an IP network, a TCP/IP connection and SMC-D only applies to connections using TCP sockets. The IP network required for SMC-D can be provided with an external LAN (e.g. OSA-Express) or provided by HiperSockets. SMC-D is only supported by z/OS. For these reasons, HiperSockets will still be needed (i.e. for Linux on z Systems environments, other network protocols, etc.). The SMC protocol bypasses TCP/IP processing (related to exchanging user data) for providing a direct optimized form of communications. The performance benefits of SMC-D is superior to (in some cases significantly better than) HiperSockets. See the reference page at the end of this document for benchmark data.

[1] ISM VCHID and PFIDs must be defined in HCD (IOCDS).

2. Benefit of SMC-D

Question:

2.1. What is the benefit of Internal Shared Memory (ISM)?

Answer:

Internal Shared Memory (ISM) provides adapter virtualization with high scalability that allows logical partitions (or guest virtual machines) to logically share virtual memory. ISM enables SMC-D (see next question). ISM provides both cost savings in physical resources and processor cost (see performance benefits).

ISM is based on PCIe architecture, which leverages the existing z system PCI eco-system, assets such as HCD, IOCDS, IOCP and memory I/O virtualization and translation. ISM is provided by a VCHID (similar to HiperSockets). There can be up to 32 ISM VCHIDs per CPC and up to 255 virtual functions (VFs) per VCHID. That means there can be up to 8160 VFs per CPC.

Question:

2.2. What is the value of SMC-D?

Answer:

The value of SMC-D is about providing significantly improved performance for intra-system communications. The value is related to co-located workloads. The improvement attributes are typically described as latency, throughput and CPU cost. Improvements can potentially improve (increase) application workload transaction rates while reducing your CPU cost. The type and amount of improvement is based on the workload characteristics (i.e. OLTP vs. streaming).

SMC-D is transparent to socket applications, requires no IP topology changes and preserves connection level security. SMC-D does not expose data to the external network. It requires no additional hardware (such as adapters, card slots, switches, and fabric management). It provides higher scalability, bandwidth and virtualization.

See the reference page at the end of this document for performance benchmark data.

Question:

2.3. Can I use SMC-D to talk to another z13s or z13 system?

Answer:

No, SMC-D is for intra-CPC communications only (two LPARs or z/VM guests on a single CPC).

For communication between CPCs SMC-R and RoCE is an option.

Question:

2.4. The term “Shared Memory Communications” suggests that special memory might be required in the operating system. Are there any new or special memory requirements or memory considerations for exploiting SMC-D?

Answer:
No, there are no new or special memory hardware or operating system memory requirements for the exploitation of SMC-D. The memory used for SMC-D to exchange application data is similar to SMC-R and is managed by the z/OS Communications Server. The vast majority of the memory is managed by and owned by the TCP/IP stack and is allocated within TCP (ASID) 64-bit private, on behalf of socket application middleware. This storage is not CSM managed. Application middleware memory management is not affected or changed. Exploiting SMC-D does require using fixed memory. The majority of this memory is a variable amount managed by the TCP/IP stack. The amount of fixed memory used for SMC-D can be planned for, monitored and controlled by the administrator (limited by the FIXEDMEMORY definition on IP GLOBALCONFIG SMCD statement).

Question:

2.5. Can I use both SMC-D and SMC-R at the same time?

Answer:
Yes, both SMC-R and SMC-D can be used at the same time. When the TCP connection is established the SMC protocol will dynamically determine which variations are supported by both hosts. When both peer hosts are enabled for SMC-D, then the protocol will determine if both peers are eligible for SMC-D connectivity (i.e. both hosts are on the same CPC, have access to the same ISM VCHID and VLAN (when applicable)) and then select SMC-D when possible. When SMC-D connectivity is not possible, then SMC-R eligibility is evaluated.

Question:

2.6. I don't know if I have the type of or enough of the workloads with the network traffic patterns that could benefit from SMC-D. How can I determine if SMC-D applies to my environment and what level of benefit I might expect?

Answer:
IBM has provided a new tool called the SMC Applicability Tool (SMCAT). The SMCAT has been provided for z/OS V1R13 and V2R1 (via PTFs), and is in the base support of V2R2. The tool does not have any dependency on RoCE, SMC-R, or SMC-D. Instead SMCAT will monitor your current TCP/IP traffic for a specified IP address, group of IP addresses or IP subnets and then produce a summary report describing the network traffic associated with the monitored addresses. The summary report will provide information about how much traffic (percentage) to/from those IP addresses is eligible for and well suited for SMC-R and SMC-D.

SMCAT PTFs:
V1R13: APAR PI27252 PTF UI24872
V2R1: APAR PI29165 PTFs UI24762 and UI24763

SMCAT Overview document is available on the SMC-R reference material web site:
http://www.ibm.com/software/network/commserver/SMCR/

3. Configuration

Question:

3.1. Do I need to have a 10 GbE RoCE Express feature?

Answer:

No, there is no requirement for RoCE (RDMA over Converged Ethernet).

A TCP/IP connection over QDIO OSA or HiperSockets (internal QDIO (iQDIO)) is required.

Question:

3.2. Do I need to have an OSA or HiperSockets connection to take advantage of SMC-D?

Answer:

Yes, a TCP/IP connection over QDIO OSA or HiperSockets (internal QDIO (iQDIO)) is required.

The OSA connection may be a 1 GbE or 10 GbE connection but it must be defined as CHPID type of OSD (QDIO). A HiperSockets connection is defined as CHPID type IQD.

The SMC-D protocol leverages your existing IP topology and the TCP/IP connection to manage your SMC-D connectivity. Therefore you will still need to establish a standard TCP/IP connection to the peer. SMC-D uses the TCP connection to determine eligibility to exploit the Internal Shared Memory (ISM) function. TCP/IP is used not only to establish the TCP/IP and SMC-D connections, but also for Keepalive functions and to terminate the TCP and the associated SMC-D connections. The standard TCP/IP path need not be dedicated to SMC-D usage; it can be used simultaneously for other, TCP/IP traffic.

Question:

3.3. What is the feature code for SMC-D?

Answer:

There is no orderable feature code.

ISM will come standard with the IBM z13s and with the microcode level 27 for the z13.

SMC-D will come standard with z/OS V2.2 (PTF).

Question:

3.4. What operating systems support SMC-D technology?

Answer:

z/OS V2.2 requires the following APARs:

1. OA47913 PTF xxxxxxx (IOS)
2. PI45028 PTF xxxxxxx (TCP/IP)
3. OA48411 PTF xxxxxxx (VTAM)

z/VM V6.3 guest support (requires APAR VM65716).

HCD ISM support requires APAR OA4610.

Note: Linux on z Systems - IBM is working with its Linux distribution partners to include

support in future distribution releases.

Question:

3.5. How is the Internal Shared Memory (ISM) and Shared Memory Communications – Direct (SMC-D) defined?

Answer:

1. ISM is configured using a new “Internal Shared Memory” (ISM) static VCHID Type. ISM VCHID concepts are similar to IQD (HiperSockets) VCHID.
   Adapter virtualization (Virtual Functions) is provided 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)
   - Additional ISM notes:
     - ISM supports Dynamic I/O.
     - 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).
     - MACs (VMACs), MTU, ports (port number) and Frame size are all N/A.
     - ISM is supported by z/VM for z/OS guests (minor changes to support new PCI function).
2. SMC-D is enabled in z/OS within the TCP/IP profile (similar to SMC-R) on the GLOBALCONFIG statement with a single SMCD parameter.

   ISM FIDs **are not** configured in the TCP/IP profile (GLOBALCONFIG statement). Instead, ISM FIDs are dynamically discovered by TCP/IP (based on your HCD configuration). ISM FIDs (VCHIDs) are associated with your IP network adapter (OSA or HiperSockets) based on defining matching PNet IDs. A unique PNet ID may be defined on HiperSockets or QDIO OSA to match ISM. An ISM VCHID is associated with OSA or HiperSockets (based on matching PNet IDs). ISM cannot be associated with both OSA and HiperSockets (all 3 cannot have the same PNet ID). Partner LPARs (or z/OS z/VM guests) must have IP connectivity (HiperSockets or QDIO OSA) on the same IP subnet (IP layer 2) and VLAN ID (when VLANs are applicable).

4. Application Questions

Question:

4.1. What changes do I need to make to the application to take advantage of SMC-D?

Answer:

There are no application changes required. The application is not involved in the decision to use SMC-D or not. This is handled inside of z/OS Communications Server configuration.

Question:

4.2. Can I enable SMC-D but exclude some applications from using it?

Answer:

Yes, there are 2 options for excluding specific applications:

1. Static solution: For applications that act as the TCP server you can configure NOSMC on the PORT and PORTRANGE statements. Disables both SMC-R and SMC-D.
2. Dynamic solution: AUTOSMC is a parameter on the GLOBALCONFIG SMCGLOBAL statement that will enable SMC autonomics that dynamically disables application workloads that are not “well-suited” for SMC (i.e. application workloads that predominantly uses short lived connections).

Question:

4.3. What application traffic is supported using SMC-D?

Answer:

With the exception of IPsec traffic, all other TCP/IP traffic using TCP sockets is eligible for SMC-D. Other network connectivity requirements must also be met. The IP connection must flow between QDIO OSA adapter ports or HiperSockets on two z/OS V2R2+ (PTF) partners that are enabled for SMC-D. TCP connection eligibility is dynamically discovered during the initial TCP setup. Connection level security such as SSL or AT-SSL can be used with SMC-D. If the connection is eligible it will use SMC-D (ISM) and there are no required changes needed for applications.

Question:

4.4. What are the SMC-D protocol connection eligibility requirements for two peer hosts to connect using SMC-D?

Answer:

In order to be eligible to connect using SMC-D the following must be true about the two peer hosts physical and network connectivity between two peer hosts.
Both hosts must meet the following:

1. Physical connectivity:
    a. On the same z13/z13s CPC
    b. Configured to use the same ISM VCHID.
    c. The ISM VCHID (FIDs) must be configured with the same PNet ID as your IP

network.

d. Configured to have direct network access to the same physical Layer-2 network (same physical external LAN or internal LAN using HiperSockets), Physical networks are defined by PNet ID (HCD).

2. IP connectivity:
   a. Direct network interface configured with the same IP Subnet (for IPv4) or prefix (for IPv6). INTERFACE statements are required for both OSA and HiperSockets (DEVICE/LINK is not supported).
   b. Access to the same VLAN ID (if VLAN is applicable)

Both hosts must also be enabled for SMC-D. The TCP connection cannot use IPSec.

Question:

4.5. Can two hosts that are in unique IP security zones that must traverse a network firewall exploit SMC-D?

Answer:

No, SMC-D does not support IP Layer 3 routing. Both peer hosts must be within the same Layer-2 IP network (same subnet).

Question:

4.6. What are the steps required to use SMC-D between two partner servers?

Answer:

Prerequisites for the establishment of an SMC-D connection:

1. The communication partners have the required software SMC-D support. See Question 3.4.
2. The TCP route selected at both ends of the connection must lead over an interface for a QDIO OSA port or HiperSockets LAN that has not been disabled for SMC-D (network interfaces are enabled for SMC-D by default, when defined with INTERFACE statements).
3. The QDIO OSA adapters or HiperSockets LAN and the ISM VCHID at an end of the connection must have been associated with each other using the same Physical Network IDs (PNet IDs). PNet IDs are configured in HCD (or IOCDS).
4. SMC-D must be enabled on the GLOBALCONFIG statement in the PROFILE.TCPIP file on both partners. This is the only required TCP/IP configuration step. ISM FIDs are not configured in the TCPIP profile. ISM FIDs are dynamically discovered once the FIDs are configured online.
5. The partnering OSA adapter ports or HiperSockets LAN at both ends of the connection must be using the same IP subnet and VLAN ID if one has been assigned.
   Notes:
   a. VLAN IDs are not defined on any SMC-D or ISM definition; the VLAN IDs are dynamically learned and inherited from the QDIO OSA or HiperSockets definition.)

b. Subnet mask (IPv4) must be configured on the OSA or HiperSockets interface that is used for SMC-D.
c. DYNXCF HiperSockets is also eligible for SMC-D. It is enabled by configuring a PNet ID on the IQD VCHID used for DYNXCF that matches your ISM VCHID. The DYNAMICXCF statement also provides a SMCD | NOSMCD option.

5. Announcements, RAS (Reliability, Accessibility, Scalability), and Security

Question:

5.1. What are the Announcement Letters referring to the SMC-D support?

Answer:

IBM z13s… Announcement Letter 116-002, 02/016/2016 (availability date 03/10/2016)

Question:

5.2. What security is available for SMC-D? How do I know the data traveling across the memory is protected?

Answer:

Similar to SMC-R, SMC-D supports connection level security features such as SSL or AT-TLS. IPSec is not supported. SMC-D preserves all other existing network security features, IP topology, and network administrative and operational models available in z/OS Communications Server. For example, the data that travels across the memory between operating systems can be protected with encryption, data integrity controls, IP filters, authentication, access controls and so on.

See White Paper on security considerations http://www.ibm.com/common/ssi/cgi-bin/ssialias?subtype=WH&infotype=SA&appname=STGE_ZS_SW_USEN&htmlfid=ZSW03255USEN&attachment=ZSW03255USEN.PDF

6. Positioning of RoCE with SMC-R versus OSA-Express and HiperSockets

Question:

6.1. What differences exist between HiperSockets, SMC-R (RoCE) and SMC-D ((ISM) in terms of the communication fabric and the underlying technology?

Answer:

Shared Memory Communications is a highly optimized communications protocol designed for Enterprise Data Center multi-tiered application workloads. SMC places an emphasis on co-location (data center and within CPC). SMC allows applications using TCP sockets to directly communicate by bypassing the TCP/IP (data and packet level related) processing, yet preserves all of the TCP/IP security and administrative controls. With this architecture, SMC provides several significant performance advantages.

SMC can now be exploited:

1. within a CPC using Shared Memory Communications – Direct Memory Access with ISM (without requiring any additional physical network hardware) and
2. across CPCs using Shared Memory Communications – Remote Direct Memory Access with IBM 10GbE RoCE Express feature (RoCE fabric).

Both forms of SMC can be used concurrently or independently of each other. z/OS will dynamically determine which form of SMC can be used for each TCP connection and select the most optimal option (i.e. SMC-D when both hosts are within the same CPC).

HiperSockets and ISM are technologies based on virtual adapters. While there are some similarities, there are also some key differences. Here is a summary comparison of the two technologies.

1. HiperSockets™ summary:
    a. Represents a virtual connection technology that requires neither a network adapter card nor special cabling (no network hardware).
    b. Is implemented in system firmware (IQD) as an internal LAN (e.g. supporting broadcast, multi-cast and unicast protocols).
    c. Is a z Systems proprietary link type, based on QDIO architecture and can be viewed as an internal NIC represented with an IPv4 or IPv6 address and a Virtual MAC.
    d. Is supported by z/OS, Linux on z Systems, z/VM, and z/VSE.
    e. In z/OS, it is included as part of a Dynamic XCF network.
    f. Supports various upper layer protocols (e.g. SNA, TCP, UDP, IPv4 IPv6, etc.).
    g. Relies on standard TCP/IP routing and therefore automatic backup to OSA relies on TCP/IP dynamic routing.
    h. Data movement is provided with memory-to-memory data moves, while executing

the communication layers of the TCP/IP layers on standard CPUs. zIIP offload can be exploited to minimize this cost.

2. Internal Shared Memory (ISM) summary:
    a. Represents a virtual connection technology that requires neither a network adapter card nor special cabling (no networking hardware).
    b. Is implemented in system firmware (ISM) as an internal shared memory fabric providing a direct approach to Shared Memory Communications
    c. Is a z Systems proprietary technology and is based on PCIe architecture.
    d. Is currently supported by z/OS only.
    e. Can be included as part of a Dynamic XCF network (when associated with the XCF HiperSockets IP network)
    f. ISM is exploited by SMC-D only (which only supports TCP connections).
    g. ISM is not an internal NIC and therefore does not require an IPv4, IPv6 address or a Virtual MAC.
    h. Relies on a separate TCP/IP network and an existing TCP connection over that network.
    i. ISM is not affected by IP routing and therefore dynamic routing, fail-over and automatic backup to another fabric are N/A.
    j. Data movement is provided by a direct memory-to-memory data move model providing a direct application move of TCP sockets data. TCP/IP processing is N/A to data movement.

7. Pricing

Question:

7.1. What is the price of SMC-D?

Answer:

SMC-D is part of z/OS Communications Server base cost (see Question 3.4 for requirements) and there is no additional charge. ISM is also a no cost feature provided with z13s and z13 firmware level 27.

For more information about SMC-R / RoCE and SMC-D / ISM please see
http://www.ibm.com/support/techdocs/atsmastr.nsf/WebIndex/FQ131485

SMC-R / SMC-D / SMCAT reference information:

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