z/OS Communications Server

Communications Server support for 25 GbE RoCE Express2 features

Version 2 Release 1

Note:
Links to related publications are from original documents and might not work. The links to publications are included for reference purposes only.

Contents

Tables

Conventions and terminology that are used in this information

Commands in this information that can be used in both TSO and z/OS® UNIX environments use the following conventions:

- When describing how to use the command in a TSO environment, the command is presented in uppercase (for example, NETSTAT).
- When describing how to use the command in a z/OS UNIX environment, the command is presented in bold lowercase (for example, **netstat**).
- When referring to the command in a general way in text, the command is presented with an initial capital letter (for example, Netstat).

All the exit routines described in this information are *installation-wide exit routines*. The installation-wide exit routines also called installation-wide exits, exit routines, and exits throughout this information.

The TPF logon manager, although included with VTAM®, is an application program; therefore, the logon manager is documented separately from VTAM.

Samples used in this information might not be updated for each release. Evaluate a sample carefully before applying it to your system.

Note: In this information, you might see the following Shared Memory Communications over Remote Direct Memory Access (SMC-R) terminology:

- RoCE Express®, which is a generic term representing IBM® 10 GbE RoCE Express, IBM 10 GbE RoCE Express2, and IBM 25 GbE RoCE Express2 feature capabilities. When this term is used in this information, the processing being described applies to both features. If processing is applicable to only one feature, the full terminology, for instance, IBM 10 GbE RoCE Express will be used.
- RoCE Express2, which is a generic term representing an IBM RoCE Express2® feature that might operate in either 10 GbE or 25 GbE link speed. When this term is used in this information, the processing being described applies to either link speed. If processing is applicable to only one link speed, the full terminology, for instance, IBM 25 GbE RoCE Express2 will be used.
- RDMA network interface card (RNIC), which is used to refer to the IBM® 10 GbE RoCE Express, IBM® 10 GbE RoCE Express2, or IBM 25 GbE RoCE Express2 feature.
- Shared RoCE environment, which means that the "RoCE Express" feature can be used concurrently, or shared, by multiple operating system instances. The feature is considered to operate in a shared RoCE environment even if you use it with a single operating system instance.

Clarification of notes

Information traditionally qualified as Notes is further qualified as follows:

Note Supplemental detail

Tip Offers shortcuts or alternative ways of performing an action; a hint

Guideline
Customary way to perform a procedure

Rule Something you must do; limitations on your actions

Restriction
Indicates certain conditions are not supported; limitations on a product or facility

Requirement
Dependencies, prerequisites

Result Indicates the outcome

Chapter 1. New Function Summary

Communications Server support for 25 GbE RoCE Express2 features

z/OS Communications Server V2R1 is enhanced to support IBM 25 GbE RoCE Express2 features.

To enable the z/OS Communications Server support for 25 GbE RoCE Express2 features, complete the appropriate tasks in Table 1.

Table 1. Task topics to enable z/OS Communications Server support for 25 GbE RoCE Express2 features

Task	Reference
Configure at least one IBM 25 GbE RoCE Express2 feature in HCD. For each IBM RoCE Express2 port, configure the physical network Identifier (PNetID), the physical channel identifier (PCHID), the function ID (FID), the virtual function ID (VF), and the port number (PORTNUM).	z/OS Hardware Configuration Definition (HCD) Reference Summary
Configure or update the GLOBALCONFIG SMCR statement in the TCP/IP profile. • Use the FID values configured in HCD to define the PFID values that represent physically different IBM 25 GbE RoCE Express2 features to provide full redundancy support. Do not specify PortNum for IBM RoCE Express2 PFIDs.	• GLOBALCONFIG statement in z/OS Communications Server: IP Configuration Reference • Shared Memory Communications over Remote Direct Memory Access in z/OS Communications Server: IP Configuration Guide
Display information about a RoCE Express2 interface, including the interface speed, by issuing the Netstat DEvlinks/-d command and specifying the RoCE Express2 interface name.	Netstat DEvlinks/-d report in z/OS Communications Server: IP System Administrator's Commands

To find all related topics about Communications Server support for 25 GbE RoCE Express2 features, see Table 2.

Table 2. All related topics about z/OS Communications Server support for 25 GbE RoCE Express2 features

Book name	Topics
IP Configuration Guide	• Shared Memory Communications terms • SMC-R link groups • System requirements for SMC-R in a shared RoCE environment
IP System Administrator's Commands	• Netstat DEvlinks/-d report
SNA Messages	• IST2361I
z/OS Hardware Configuration Definition (HCD) Reference Summary	N/A

Chapter 2. IP Configuration Guide

SMC-R link groups

A Shared Memory Communications over RDMA (SMC-R) link group is a logical grouping of SMC-R links between two communicating peers, as shown in Figure 1 on page 4. An SMC-R link group is formed when the initial SMC-R link is established between two peers.

All SMC-R links in an SMC-R link group must be *equal* links. SMC-R links are considered to be equal when all of the following conditions are true:

- The links provide access to the same RDMA memory buffers at the remote peer virtual servers.
- The links have the same VLAN ID, or they do not use a VLAN ID.
- The links have the same TCP server and TCP client roles or relationship.

  A peer that is acting as the TCP connection server has different responsibilities for establishing and maintaining SMC-R communications than a peer that is acting as the TCP connection client. Unique SMC-R link groups are established between two peers when the peers act as both servers and clients for TCP connections.

When the initial SMC-R link is established and a second "RoCE Express" interface is available, Communications Server establishes an equal SMC-R link between the peers. The "RoCE Express" interfaces are shown as RNICs in Figure 1 on page 4.

Adding a second SMC-R link to the SMC-R link group provides the following benefits:

- High availability

  To maintain high availability, you need two SMC-R links between SMC-R peers. If a failure occurs with one SMC-R link, TCP connections that are using the failing SMC-R link are switched to the other active link in the link group and disruptions to application workloads are avoided. For more information, see High availability considerations.
- Workload balancing

  TCP connections are distributed across the SMC-R links in a link group, increasing bandwidth and avoiding bottlenecks.

  Rule: Workload balancing within an SMC-R link group occurs only when both the local and the remote peers have two "RoCE Express" interfaces, and thus two SMC-R links are established in the link group.

  Guideline: When you are provisioning PFIDs for a specific PNetID for each TCP/IP stack, you should assign "RoCE Express" PFIDs with the same link speed (i.e. 10 GbE, 25 GbE). If this guideline is not followed, this will result in SMC-R link groups being created with SMC-R links with unequal throughput capacities. During error recovery scenarios, fail over processing might overload the lower capacity SMC-R link that could result in recovery failures.

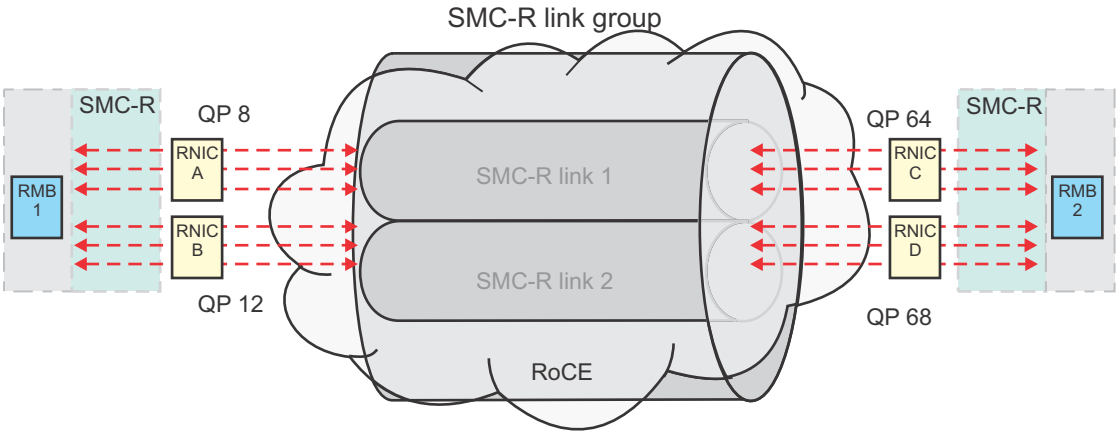

Figure 1. SMC-R link group

Because SMC-R links within a link group are considered equal, TCP connections can be assigned to any SMC-R link within the group. Furthermore, the client and the server can choose to assign the TCP connection to different SMC-R links within the group, and can move the TCP connections from one SMC-R link to another within the group. For example, in Figure 1, client traffic might flow over one SMC-R link (between RNICs A and C) and server traffic might flow over the other SMC-R link (between RNICs B and D). The peers do not have to exchange knowledge of which physical "RoCE Express" interface is being used for data transmission, and the recipient is only aware that data was placed into the RDMA memory buffer.

An SMC-R link group remains active for up to 10 minutes after the last TCP connection that is using the link group is stopped.

System requirements for SMC-R in a shared RoCE environment

You need to ensure that your system meets the requirements to use Shared Memory Communications over RDMA (SMC-R) with "RoCE Express" features operating in a shared RoCE environment.

To use SMC-R with 10 GbE RoCE Express features operating in a shared RoCE environment, the minimum software requirement must be z/OS Version 2 Release 1 with APARs OA44576 and PI12223 applied.

To use SMC-R with RoCE Express2 features, the minimum software requirement is:

- z/OS Version 2 Release 1 with APARs OA51949 and PI75199 applied

SMC-R requires RDMA over Converged Ethernet (RoCE) hardware and firmware support. The following minimum hardware requirements must be met to use SMC-R:

- If you use 10 GbE RoCE Express features, you must have IBM z13® (z13) or later systems.
- If you use RoCE Express2 features, you must have IBM z14™ or later systems
- You must have one or more IBM 10 GbE RoCE Express or RoCE Express2 features.

  "RoCE Express" features are dual ports with short range (SR) optics and can be shared across multiple operating systems images or TCP/IP stacks in a central processor complex (CPC).

Guideline: Provide two "RoCE Express" features per unique physical network. For more information, see “RoCE network high availability.”

- You must have System z® OSA-Express for traditional Ethernet LAN connectivity.

  SMC-R does not impose any specific OSA requirements.
- You must have standard 10 GbE or 25 GbE switches.

RoCE network configuration requirements

z/OS Communications Server supports connectivity to multiple, distinct layer 2 networks through unique physical LANs. Each unique physical network is identified by existing Ethernet standards that are based on the physical layer 2 broadcast domain. You can define a physical network ID (PNet ID) for each physical network. For more information, see Physical network considerations.

For hosts to communicate by using SMC-R, they must connect directly to the same Ethernet layer 2 LAN network. If VLANs are in use, each host must also have access to the same VLAN. For more information, see VLANID considerations.

There are restrictions on the physical distances that can be used to route RDMA frames. To understand these distance specifications and limitations, see *IBM z Systems® Planning for Fiber Optic Links*.

RoCE network high availability

Because RoCE connections do not use IP routing and the RDMA connections to remote hosts are direct point-to-point connections that use reliable connected queue pairs (RC QPs), there is no concept of an alternative IP route to the peer. SMC-R connectivity is possible with a single 10 GbE RoCE Express feature, but a loss in that single feature means that the associated TCP connections and workloads are disrupted. Therefore, redundant 10 GbE RoCE Express features on both the local and remote hosts are required to achieve network high availability with SMC-R. If your TCP workloads require high availability, redundant 10 GbE RoCE Express features and redundant Ethernet switches are required. The SMC-R protocol actively uses both features, rather than using one feature with the other in standby mode. For more information, see High availability considerations.

IBM 10 GbE RoCE Express features also have redundant internal PCIe support structures, or PCIe internal paths, as described in Redundancy levels. To avoid another single point of failure, install each 10 GbE RoCE Express feature that is managed by the same operating system with a unique internal path. For more information about how to install a 10 GbE RoCE Express feature to achieve full redundancy, see your IBM service representative.

RoCE bandwidth

The 10 GbE RoCE Express features provide 10 GbE ports. When redundant features are defined, SMC-R link groups can be formed by using both features, resulting in a 20 GbE logical pipe to each physical network. z/OS Communications Server uses only two features within a link group at any particular time.

Shared Memory Communications over RDMA terms and concepts

The following terms and concepts apply to Shared Memory Communications over Remote Direct Memory Access (SMC-R). You can use this list as needed for brief descriptions when you are using other SMC-R information.

Associated RNIC interface
An IBM 10 GbE RoCE Express interface that is associated with an SMC-R capable interface that has the same physical network ID.

IBM 10 GbE RoCE Express feature or RoCE Express2 feature
A feature that enables Remote Direct Memory Access by managing low-level functions that the TCP/IP stack typically handles.

IBM 10 GbE RoCE Express interface
An interface that is dynamically created by TCP/IP that uses a particular port of an IBM 10 GbE RoCE Express feature.

IBM RoCE Express2 interface
An interface that is dynamically created by TCP/IP that uses a particular port of a 10 GbE or 25 GbE RoCE Express2 feature.

Internal path
The System z internal PCIe infrastructure for "RoCE Express" features. The internal path of a "RoCE Express" feature is determined based on how the feature is plugged into the System z I/O drawers.

Operating system images
Logical partitions (LPARs) or guest virtual machines that operate in the same central processor complex (CPC).

Physical channel ID (PCHID)
A 2-byte hexadecimal value that is used to uniquely define a RoCE Express feature.

PCIe function ID (PFID)
A value that is configured on the SMCR parameter of the GLOBALCONFIG statement in the TCP/IP profile to identify a "RoCE Express" feature. The PFID represents a physical "RoCE Express" feature and must match a FID value configured in the hardware configuration definition (HCD) for the PCHID value that identifies the feature. When the "RoCE Express" feature is installed on a System z that supports a shared RoCE environment, the same physical feature can be shared with other operating system images, and multiple PFID values specified on the same GLOBALCONFIG statement can represent different ports on the same physical "RoCE Express" feature.

Peripheral Component Interconnect Express (PCI Express, or PCIe)
A local bus that provides the high-speed data path between the processor and an attached "RoCE Express" feature.

Physical network ID (PNet ID)
A value that is defined to uniquely identify your physical layer 2 LAN fabric or physical broadcast domain. You can use this value to logically associate the System z features, adapters, and ports to be physically connected to your network. You specify the PNet ID in a single step within the hardware configuration definition (HCD), and all operating systems of all associated central processor complexes (CPCs) can dynamically learn and use this definition.

RDMA network interface card (RNIC)
A RoCE Express feature that enables Remote Direct Memory Access by managing low-level functions that are typically handled by the TCP/IP stack.

RDMA over Converged Ethernet (RoCE)
An InfiniBand Trade Association (IBTA) standard that enables Remote Direct Memory Access over Converged Ethernet.

Redundancy level
For an SMC-R link group, this value indicates the level to which z/OS Communications Server can provide dynamic failover processing if there is a failure of an underlying "RoCE Express" interface or the associated network hardware.

Reliable connected queue pair (RC QP)
A logical connection between two virtual servers that enables that specific pair of servers to use RDMA communications between themselves.

Remote Direct Memory Access (RDMA)
A high-speed, low-latency network communications protocol in which data is transferred directly to the memory of a remote host with no involvement from the remote host processors or operating system.

Remote memory buffer (RMB)
Local memory that is used to receive inbound data over an SMC-R link. The remote peer places TCP socket application data directly into the RMB that the local peer assigns to receive data for the TCP connection. The local peer then copies the data from the RMB into the receive buffer of the receiving socket application.

Rendezvous processing
The sequence of TCP connection management flows that are required to establish SMC-R communications between two peers.

RMB element (RMBE)
The specific portion of an RMB that is associated with a specific TCP connection. Each RMB is partitioned into RMBEs.

RoCE environments
Depending on the level of hardware that is used, the 10 GbE RoCE Express feature operates in either a shared or a dedicated RoCE environment. A RoCE Express2 feature always operates in a shared RoCE environment.

Dedicated RoCE environment
A dedicated RoCE environment applies to an IBM zEnterprise® EC12 (zEC12) with driver 15, or an IBM zEnterprise BC12 (zBC12). In this environment, only a single operating system instance can use a physical 10 GbE RoCE Express feature. Multiple operating system instances cannot concurrently share the feature.

Shared RoCE environment
A shared RoCE environment applies to an IBM z13 (z13) or later system. In this environment, multiple operating system instances can concurrently use or share the same physical RoCE Express feature. With IBM z13 (z13) or later systems, the RoCE Express feature operates in a shared environment even if only one operating system instance is configured to use the feature.

RoCE Express
Generic term for either IBM 10 GbE RoCE Express, IBM 10 GbE RoCE Express2, or IBM 25 GbE RoCE Express2 feature.

RoCE Express2
Generic term for either IBM 10 GbE RoCE Express2 or IBM 25 GbE RoCE Express2 feature.

SMC-R link
A logical point-to-point link between two virtual servers that is used for SMC-R communications.

SMC-R link group
A logical grouping of equal SMC-R links between two communicating peers.

Staging buffer
Memory that the TCP/IP stack allocates for outbound SMC-R data. Staging buffers are not associated with specific SMC-R links or link groups, and are used by all TCP connections that traverse SMC-R links on this stack. Only local applications access the staging buffer storage.

Chapter 3. SNA Messages

IST2361I SMCR PFID = *pfid* PCHID = *pchid* PNETID = *network_id*

Explanation: VTAM issues this message as part of a message group in response to a DISPLAY ID or DISPLAY TRL command for a TRLE that is associated with a "RoCE Express" interface.

VTAM also issues this message as part of a group of messages generated by the adapter interrupt monitoring function. The first message in the group is IST2419I. See message IST2419I for a complete description.

A complete description of the message group follows the example:

```
IST075I NAME = nodename, TYPE = TRLE
IST1954I TRL MAJOR NODE = trl_major_node_name
IST486I STATUS= current_status, DESIRED STATE= desired_state
IST087I TYPE = *NA* , CONTROL = ROCE , HPDT = *NA*
IST2361I SMCR PFID = pfid  PCHID = pchid  PNETID = network_id
IST2362I PORTNUM = port  RNIC CODE LEVEL = code_level
IST2389I PFIP = pci_path   GEN = generation   SPEED = speed
[IST2417I VFN = virtual_function_number]
IST924I -------------------------------------------
IST1717I ULPID = ulp_id ULP INTERFACE = ulp_interface
IST1724I I/O TRACE = iotrc TRACE LENGTH = length
[IST924I -------------------------------------------]
[IST1717I ULPID = ulp_id ULP INTERFACE = ulp_interface]
[IST1724I I/O TRACE = iotrc TRACE LENGTH = length]
```

IST075I

This message displays the resource name and resource type.

nodename
The name of the resource that was entered on the DISPLAY command.

nodetype
The resource type of the major or minor node. The *nodetype* value is always **TRLE** for this message group.

IST087I

This message displays line information associated with *nodename*.

line_type
The *line_type* value is always ***NA*** for this message group.

line_control
The *line_control* value is always **ROCE** (RDMA over Converged Ethernet) for this message group.

hpdtvalue
The *hpdtvalue* is always ***NA*** for this message group.

IST486I

This message displays status information for *nodename*.

current_status
The current status of the node. See the z/OS Communications Server: IP and SNA Codes for status information.

desired_state
The node state that is desired. See the z/OS Communications Server: IP and SNA Codes for status information. If VTAM cannot determine the desired state, *desiredstate* is ***NA***.

IST1717I

IST2361I

This message is displayed for all TRLEs that are currently being used by at least one Upper-layer Protocol (ULP). A separate IST1717I message is displayed for each ULP that is using the "RoCE Express" TRLE.

ulp_id The name of a z/OS Communications Server ULP that is using the "RoCE Express" TRLE. In this message group, the *ulp_id* value is always the TCP/IP job name.

ulp_interface
The name of the interface associated with the "RoCE Express" TRLE.

IST1724I

This message displays trace information for *nodename*.

iotrc Specifies whether I/O Trace is active for this "RoCE Express" TRLE (ON or OFF).

length Specifies the number of bytes being recorded for I/O Trace for this "RoCE Express" TRLE.

IST1954I

This message displays the TRL major node name.

trl_major_node_name
The name of the TRL major node defining the "RoCE Express" TRLE.

IST2361I

This message provides configuration information for the "RoCE Express" feature associated with *nodename*.

pfid The 2-byte hexadecimal Peripheral Component Interconnect Express (PCIe) function ID for the "RoCE Express" feature associated with *nodename*.

pchid The 2-byte hexadecimal physical channel ID (PCHID) for the "RoCE Express" feature associated with *nodename*.

network_id
The physical network identifier for the "RoCE Express" interface associated with *nodename*.

IST2362I

This message provides configuration and operational information about the "RoCE Express" feature associated with *nodename*.

port A decimal representation of the "RoCE Express" port number associated with *nodename*.

code_level
The processor code level of the "RoCE Express" feature. The code level is in the form **xxxxx.yyyyy.zzzzz** if the 10 GbE RoCE Express feature is operating in a dedicated RoCE environment, or if this is a 10 GbE RoCE Express2 feature.

xxxxx Major version.

yyyyy Minor version.

zzzzz Subminor version.

The code level is **NA** if the 10 GbE RoCE Express feature is operating in a shared RoCE environment.

IST2389I

This message displays additional configuration information for the "RoCE Express" feature associated with *nodename*.

pci_path
The PCI-function internal path (PFIP) value for the "RoCE Express" feature associated with *nodename*.

generation
The generation level of the "RoCE Express" feature. Possible values are:

ROCE EXPRESS
The TRLE represents an IBM 10 GbE RoCE Express feature.

ROCE EXPRESS2
The TRLE represents an IBM 10 GbE RoCE Express2 or IBM 25 GbE RoCE Express2 feature.

speed The throughput speed of the "RoCE Express" feature. Possible values are:

10GE The IBM 10 GbE RoCE Express or IBM 10 GbE RoCE Express2 feature uses 10 gigabit Ethernet ports.

25GE The IBM 25 GbE RoCE Express2 feature uses 25 gigabit Ethernet ports.

IST2417I

This message displays the virtual function number (VFN) that is associated with *nodename*. This message is displayed only when the "RoCE Express" feature operates in a shared RoCE environment.

virtual_function_number
The VFN value for the "RoCE Express" feature that is associated with *nodename*.

System action: Processing continues.

Operator response: None.

System programmer response: None.

User response: None.

Problem determination: Not applicable.

Source: z/OS Communications Server SNA

Module: Use the modifiable VTAM start option MSGMOD=YES (*f procname,vtamopts,msgmod*=yes or *f procname,msgmod*=yes) to display the issuing module when a message is issued. See z/OS Communications Server: SNA Operation and z/OS Communications Server: SNA Resource Definition Reference for more information about start options.

Routing code: 2

Descriptor code: 5

Automation: This message is not a candidate for automation.

Example:

```
IST097I DISPLAY ACCEPTED
IST075I NAME = IUT2001D, TYPE = TRLE
IST1954I TRL MAJOR NODE = ISTTRL
IST486I STATUS= ACTIV, DESIRED STATE= ACTIV
IST087I TYPE = *NA*               , CONTROL = ROCE, HPDT = *NA*
IST2361I SMCR PFID = 001D  PCHID = 0138  PNETID = NETWORK1
IST2362I PORTNUM = 2 RNIC CODE LEVEL = *NA*
IST2389I PFIP = 08040101   GEN = ROCE EXPRESS   SPEED = 10GE
IST2417I VFN = 0002
IST924I -------------------------------------------------------------
IST1717I ULPID = TCPCS ULP INTERFACE = EZARIUT2001D
IST1724I I/O TRACE = OFF  TRACE LENGTH = *NA*
IST314I END
```

Index

L

S

Printed in USA