z/OS Communications Server support for RoCE Express3

Documentation changes for TCP/IP APAR PH34117 and SNA APAR OA60855

Version 2 Release 3

Version 2 Release 4

Version 2 Release 5

z/OS Communications Server

Contents

New Function Summary

Communications Server support for RoCE Express3 features

z/OS® V2R5, V2R4, V2R3 Communications Server, with TCP/IP APAR PH34117 and SNA APAR OA60855, extends the Shared Memory Communications over Remote Direct Memory Access (SMC-R) function to support the next generation IBM® RoCE Express3® feature. The IBM RoCE Express3 feature allows TCP/IP stacks on different LPARs within the same central processor complex (CPC) to leverage the power of these state-of-the-art adapters to optimize network connectivity for mission critical workloads by using Shared Memory Communications technology.

Incompatibilities: This function does not support IPAQENET interfaces that are defined by using the DEVICE, LINK, and HOME statements. Convert your IPAQENET definitions to use the INTERFACE statement to enable this support.

Dependencies: This function requires the IBM z16™ or later systems. To enable the z/OS Communications Server support for RoCE Express3 features, complete the appropriate tasks in the following table.

Table. Task topics to enable z/OS Communications Server support for RoCE Express3 features

Task	**Reference**
Configure at least one IBM RoCE Express3 feature in HCD. For each RoCE Express3 port, configure the physical network ID (PNetID), the physical channel ID (PCHID), the Function ID (FID), the virtual function ID (VF), and the port number (PORTNUM).	z/OS Hardware Configuration Definition User's Guide
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 RoCE Express3 features to provide full redundancy support. • Do not specify PortNum for RoCE Express3 PFIDs, or specify the PORTNUM value configured in HCD for the PFID.	• 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
For z/OS V2R5 only, optionally configure or update the PFID and SMCRIPADDR parameters on the SMCR configuration statement on one or more IPAQENET OSD interface statements to enable these interfaces for SMC-Rv2.	• INTERFACE – IPAQENET OSA-Express QDIO interfaces statement in z/OS Communications Server: IP Configuration Reference • Shared Memory Communications multiple IP subnet support (SMCv2) in z/OS Communications

z/OS Communications Server

Use the FID values configured in HCD to define PFID values that represent physically different RoCE Express3 features to provide full redundancy support.	Server: IP Configuration Guide
Display information about a RoCE Express3 interface by issuing the Netstat DEvlinks/-d command and specifying the name of the RoCE Express3 interface.	Netstat DEvlinks/-d report in z/OS Communications Server: IP System Administrator's Commands

Netstat operator commands (DISPLAY TCPIP,,NETSTAT)

Table. New and changed Communications Server Netstat operator commands (DISPLAY TCPIP,,NETSTAT)

Parameter	Description
DevLinks	The card generation level and speed information are displayed for RNIC interfaces representing "RoCE Express3" features.

NETSTAT TSO commands

Table. New and changed Communications Server NETSTAT TSO commands

Parameter	Description
DEvlinks	The card generation level and speed information are displayed for RNIC interfaces representing "RoCE Express3" features.

Netstat UNIX commands

Table. New and changed Communications Server z/OS UNIX netstat commands

Parameter	Description

z/OS Communications Server

-d	The card generation level and speed information are displayed for RNIC interfaces representing "RoCE Express3" features.

TCP/IP callable NMI (EZBNMIFR)

Table. New Communications Server TCP/IP callable NMI (EZBNMIFR)

Request	Parameter/output	Description
GetRnics	NWMRnicBGen	Value of NWMRNICBGENREXP3 represents the RoCE Express3 feature.

SNA command behavior changes

Table. New and changed Communications Server commands with changed behavior

Command	Description of behavior change
DISPLAY ID	• If the resource that is being displayed is a RDMA over Converged Ethernet (RoCE) TRLE, message IST2389I contains the "RoCE Express®" generation level and the transmission speed. RoCE Express3 was added to the RoCE Express generation level. • If the resource that is being displayed is a RoCE Express3 TRLE, message IST2362I always displays the microcode level.
DISPLAY TRL	• If the TRLE operand specifies a RDMA over Converged Ethernet (RoCE) TRLE, message IST2389I contains the "RoCE Express" generation level and the transmission speed. RoCE Express3 was added to the RoCE Express generation level. • If the TRLE operand specifies a RoCE Express3 TRLE, message IST2362I always displays the microcode level.

Conventions and terminology that are used in this information

...

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®, IBM 25 GbE RoCE Express2, IBM 10 GbE RoCE Express3, and IBM 25 GbE RoCE Express3 feature capabilities. When this term is used in this information, the processing being described applies to all of these 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.
- RoCE Express3, which is a generic term representing an IBM RoCE Express3 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 Express3 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, IBM 25 GbE RoCE Express2, IBM 10 GbE RoCE Express3, or IBM 25 GbE RoCE Express3 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.

Connectivity and gateway functions

...

IBM 10 GbE RoCE Express, RoCE Express2, and RoCE Express3 feature

Enables the use of Remote Direct Memory Access (RDMA) processing by using Shared Memory Communications over RDMA (SMC-R) protocols.

...

Shared Memory Communications over Remote Direct Memory Access

...

RoCE Express3 feature environment

A RoCE Express3 feature operates in a shared RoCE environment. In general, the same rules and guidelines for defining and using a RoCE Express2 feature apply to a RoCE Express3 feature. Each RoCE Express3 port can be shared by 31 or 63 operating system instances or TCP/IP stacks across the same CPC, depending on the z16 model's capability. See Shared RoCE environment for a description of operating in a shared RoCE environment.

Shared Memory Communications terms

The following terms apply to Shared Memory Communications (SMC). You can use this list as needed for brief descriptions when you are using other SMC information.

...

IBM 10 GbE RoCE Express feature, RoCE Express2, RoCE Express3 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.

IBM RoCE Express3 interface

An interface that is dynamically created by TCP/IP that uses a particular port of a 10 GbE or 25 GbE RoCE Express3 feature.

...

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. RoCE Express2 and RoCE Express3 features always operate 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.

z/OS Communications Server

RoCE Express

Generic term for either IBM 10 GbE RoCE Express, IBM 10 GbE RoCE Express2, IBM 25 GbE RoCE Express2, IBM 10 GbE RoCE Express3, or IBM 25 GbE RoCE Express3 feature.

RoCE Express2

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

RoCE Express3

Generic term for either IBM 10 GbE RoCE Express3 or IBM 25 GbE RoCE Express3 feature.

Using Shared Memory Communications

Configuration considerations for Shared Memory Communications

SMC-R VLANID usage for the RoCE Express2 feature or RoCE Express3 feature

Whether SMC-R communications use virtual LANs depends on the definition of the SMC-R capable OSD interfaces that are extended to the associated RoCE Express2 or RoCE Express3 interfaces. The RoCE Express2/RoCE Express3 feature can be shared by TCP/IP stacks that are configured to use different VLAN capabilities for the RoCE Express2/RoCE Express3 feature. There is no limit on the number of unique VLANIDs that you can use per RoCE Express2 or RoCE Express3 port.

SMC-R physical network considerations

One TCP/IP stack can define up to 16 Peripheral Component Interconnect Express (PCIe) function ID (PFID) values. Each PFID value must match an FID value configured in the hardware configuration definition (HCD).

- In a dedicated RoCE environment, each PFID represents a unique PCHID definition of a 10 GbE RoCE Express feature, and only one of the two 10 GbE RoCE Express ports for the feature can be used at a time.
- In a shared RoCE environment, each PFID represents a virtual function (VF) usage of a 10 GbE RoCE Express feature, a RoCE Express2 feature, or a RoCE Express3 feature, and multiple PFID values can be associated with the same physical feature and port.

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 requirements are:

z/OS Communications Server

- z/OS Version 2 Release 1 with APARs OA51949 and PI75199 applied
- z/OS Version 2 Release 2 with APARs OA51950 and PI75200 applied

To use SMC-R with RoCE Express3 features, the minimum software requirements are:

- z/OS Version 2 Release 3 and z/OS Version 2 Release 4 with APARs OA60855 and PH34117 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
- If you use RoCE Express3 features, you must have IBM z16 or later systems
- You must have one or more IBM 10 GbE RoCE Express, RoCE Express2, or RoCE Express3 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.
- If you configure more than 24 Peripheral Component Interconnect Express (PCIe) devices, you must configure the IEASYSxx LFAREA parameter. The 24 PCIe devices include all z/OS Communications Server PCIe devices and other z/OS PCIe devices. In z/OS Communications Server, you can configure the following PCIe devices:
  - IBM 10 GbE RoCE Express features
  - RoCE Express2 features
  - RoCE Express3 features
  - Internal shared memory (ISM) devices

...

System requirements for SMC-D

...

If you configure more than 24 Peripheral Component Interconnect Express (PCIe) devices, you must configure the IEASYSxx LFAREA parameter. The 24 PCIe devices include all z/OS Communications Server PCIe devices and other z/OS PCIe devices. In z/OS Communications Server, you can configure the following PCIe devices:

- 10 GbE RoCE Express features
- RoCE Express2 features
- RoCE Express3 features
- Internal shared memory (ISM) devices

...

z/OS Communications Server

Configuring Shared Memory Communications over RDMA

Procedure

...

2. Configure the SMCR parameter on the GLOBALCONFIG statement in the TCP/IP profile. The SMCR parameter includes the following subparameters:

- PFID specifies the PCI Express (PCIe) function ID (PFID) value for a "RoCE Express" feature that this stack uses.

  You must code at least one PFID subparameter for this stack to use SMC-R, and two PFIDs per PNetID per stack for redundancy.

  – When the 10 GbE RoCE Express features operate in a dedicated RoCE environment, each 10 GbE RoCE Express feature must have a unique PFID value, but each TCP/IP stack that uses the 10 GbE RoCE Express feature specifies the same PFID value.

  – When a 10 GbE RoCE Express, a RoCE Express2, or a RoCE Express3 feature operates in a shared RoCE environment, each TCP/IP stack that uses the same "RoCE Express" feature must have a unique PFID value, even if the TCP/IP stacks are defined on different LPARs.

- PORTNUM specifies the 10 GbE RoCE Express port number to use for each PFID.

Guideline: You do not have to code PORTNUM for a PFID representing a RoCE Express2 or RoCE Express3 feature. The port number is defined for the PFID in the HCD, and VTAM and the TCP/IP stack learns the port number during PFID activation.

...

If PFID 0018 and PFID 0019 represent RoCE Express2 or RoCE Express3 features, you do not need to specify PORTNUM on the GLOBALCONFIG definition.

```
GLOBALCONFIG SMCR

      PFID 0018

      PFID 0019

      FIXEDMEMORY 200
```

...

VTAM displays and tuning statistics

When a "RoCE Express" or internal shared memory (ISM) interface is first started, VTAM dynamically creates a transport resource list element (TRLE) to represent it.

- For a RoCE Express2 or RoCE Express3 interface, the TRLE name is in the form of IUT*pffff*, where p is the port number and *ffff* is the PCI-Express function ID (PFID).

– For a 10 GbE RoCE Express feature, if you specify GLOBALCONFIG SMCR PFID 0018 PORTNUM 1, the TRLE name is IUT10018.

– For a RoCE Express2 or RoCE Express3 feature, if you specify GLOBALCONFIG SMCR PFID 0055, the TRLE name will be IUT10055 or IUT20055. It depends on the port number defined for PFID 55 in the Hardware Configuration Definition (HCD). VTAM learns the port number when the PFID is activated.

...

Tip: For a 10 GbE RoCE Express TRLE, use the presence of a virtual function number (VFN) in the DISPLAY TRL or DISPLAY ID command output to determine whether the "RoCE Express" feature operates in a shared RoCE environment. A VFN is present in a shared environment and absent in a dedicated environment. RoCE Express2 or RoCE Express3 TRLE always has a VFN and always operates in a shared environment.

PORTNUM *num*

Specifies the "RoCE Express" port number to use for a particular PFID. Configure each PFID to use only a single port. The port number can be 1 or 2; 1 is the default value.

Rules:

- You do not need to configure PORTNUM for IBM RoCE Express2 or RoCE Express3 features in the TCP/IP profile. The correct port number for these features is configured in the Hardware Configuration Definition (HCD) and is learned by VTAM and the TCP/IP stack during PFID activation. VTAM ignores the GLOBALCONFIG SMCR PORTNUM value if it differs from the port number configured in the HCD for the IBM RoCE Express2/RoCE Express3 feature.
- If the 10 GbE RoCE Express feature operates in a dedicated RoCE environment, you can activate either port 1 or port 2 but not both simultaneously for an individual PFID value. If PORTNUM 1 and PORTNUM 2 definitions for the same PFID value are created, the port that is first activated is used.
- If the 10 GbE RoCE Express feature operates in a shared RoCE environment, you can use both port 1 and port 2 on an individual RNIC adapter, but the PFID value that is associated with each port must be different. You cannot simultaneously activate PORTNUM 1 and PORTNUM 2 definitions for the same PFID value.

  For example, if PFID 0013 and PFID 0014 are both defined in HCD to represent the RNIC adapter with PCHID value 0140, you can configure PFID 0013 PORT 1 PFID 0014 PORT 2 to use both ports on the RNIC adapter. However, if you specify PFID 0013 PORT 1 PFID 0013 PORT 2, only the first port that is activated is used.

Examples

The following examples shows the use of the SMCR parameter to define two "RoCE Express" features that use PFID values 0018 and 0019 and port numbers 1 and 2, and to limit the stack to 500 megabytes of 64-bit storage for SMC-R communications. The first example represents 10 GbE RoCE Express features, and the second example represents RoCE Express2/RoCE Express3 features.

```
GLOBALCONFIG SMCR PFID 0018 PORTNUM 1 PFID 0019 PORTNUM 2 FIXEDMEMORY 500

GLOBALCONFIG SMCR PFID 0018 PFID 0019 FIXEDMEMORY 500
```

TCP/IP profile record Global configuration section

Table. TCP/IP profile record Global configuration section

Offset	Name	Length	Format	Description
52(X'34')	NMTP_GBCFPFs(16)	96	Binary	SMCR PFID array that contains up to 16 entries. Each entry contains the following information: • PFID (2-byte hexadecimal value) • PortNum • MTU value **Note:** When PFID represents a RoCE Express2 feature or RoCE Express3 feature, the PortNum value is the port number configured for the PFID in the Hardware Configuration Definition (HCD). This port number is learned by VTAM and TCP/IP during activation of the PFID and might be different from the value coded for PORTNUM for this PFID on the GLOBALCONFIG SMCR statement.

RDMA network interface card (RNIC) interface statistics record (subtype 44)

Table. RNIC interface statistics specific section

Offset	Name	Length	Format	Description
85(X'55')	SMF119SM_RSGen	1	Binary	"RoCE Express" feature generation level **X'01'** IBM 10 GbE RoCE Express feature **X'02'** RoCE Express2 feature **X'03'** RoCE Express3 feature

Report field descriptions

SMCR

Indicates whether this stack supports Shared Memory Communications over Remote Direct Memory Access (SMC-R) for external data network communications. This field can have the following values:

Yes

Indicates that this stack can communicate with other stacks on the external data network by using SMC-R. The SMCR parameter was specified on the GLOBALCONFIG profile statement. When the SMCR field has the value Yes, the following information is displayed:

FixedMemory

Indicates the maximum amount, in megabytes, of 64-bit private storage that the stack can use for the send and receive buffers that are required for SMC-R communications. The fixed memory value was defined by using the SMCR FIXEDMEMORY parameter on the GLOBALCONFIG. If the SMCR FIXEDMEMORY parameter was not specified on the GLOBALCONFIG statement, the default value of 256 is displayed.

TcpKeepMinInt

Indicates the minimum supported TCP keepalive interval for SMC-R links. Use the SMCR TCPKEEPMININTERVAL parameter on the GLOBALCONFIG statement to define the interval. For applications that are using the TCP_KEEPALIVE setsockopt() option, this interval indicates the minimum interval that TCP keepalive packets are sent on the TCP path of an SMC-R link. The range is 0 - 2147460 seconds. If the interval value is set to 0, TCP keepalive probe packets on the TCP path of an SMC-R link are disabled. If the SMCR TCPKEEPMININTERVAL parameter was not specified on the GLOBALCONFIG statement, then the default interval value of 300 is displayed.

PFID

Indicates the Peripheral Component Interconnect Express (PCIe) function ID (PFID) value that was defined using SMCR PFID parameter on the GLOBALCONFIG statement. The combination of PFID and port number uniquely identifies an "RoCE Express". The stack uses "RoCE Express" for SMC-R communications with other stacks on the external data network. The PFID is a 2-byte hexadecimal value.

PortNum

Indicates the "RoCE Express" port number that is used for the associated PFID. The PortNum value was specified with the PFID value on the SMCR parameter of the GLOBALCONFIG statement in the TCP/IP profile. The port number can be 1 or 2; the default port is 1.

Note: When PFID represents a RoCE Express2 feature or RoCE Express3 feature, the PortNum value is the port number configured for the PFID in the Hardware Configuration Definition (HCD). The port number is learned by VTAM during activation of the PFID and might be different from the value coded for PORTNUM for this PFID on the GLOBALCONFIG SMCR statement.

MTU

Indicates the configured maximum transmission unit (MTU) value that is used for the associated PFID. The MTU value can be 1024 or 2048 and the default MTU value is 1024.

No

Indicates that this stack cannot communicate with other stacks on the external data network by using SMC-R communications. The NOSMCR parameter was specified on the GLOBALCONFIG profile statement or the value was set by default.

Netstat DEvlinks/-d report

Report field descriptions

PortNum

Specifies the port number that is used for the associated PFID.

- When PFID represents an IBM 10 GbE RoCE Express feature, the PortNum value is specified with the PFID value on the SMCR parameter of the GLOBALCONFIG statement in the TCP/IP profile.
- When PFID represents a RoCE Express2 feature or RoCE Express3 feature, the PortNum value is the port number configured for the PFID in the Hardware Configuration Definition (HCD). The port number is learned by VTAM during activation of the PFID and might be different from the value coded for PORTNUM for this PFID on the GLOBALCONFIG SMCR statement.

...

Gen

Indicates the generation level for the "RoCE Express" feature, and is significant only if the interface is active. Possible values are:

RoCE Express

The feature is an IBM 10 GbE RoCE Express feature.

RoCE Express2

The feature is a RoCE Express2 feature.

RoCE Express3

The feature is a RoCE Express3 feature.

VTAM message IST2444I seen during PFID activation

You should not code the PORTNUM operand on the GLOBALCONFIG SMCR statement when the PFID represents a RoCE Express2 or RoCE Express3 feature. The correct port number is configured in the Hardware Configuration Definition (HCD) and is learned by VTAM® and the TCP/IP stack during PFID activation.

If you code a value for PORTNUM on GLOBALCONFIG, and it is not the same as the HCD port number value, VTAM issues message IST2444I during PFID activation. For instance, if you configured GLOBALCONFIG SMCR PFID 51 PORTNUM 2, but the correct port number is port 1, the following message is generated when VTAM activates PFID 51:

```
IST2444I PORTNUM 2 IGNORED FOR SMC-R PFID 0051, ACTIVATION CONTINUES
```

This is just an informational message, so you do not have to take any action. If you want to avoid getting this message during any subsequent PFID activation attempts, you can remove the PORTNUM value from the GLOBALCONFIG SMCR statement.

Any VARY OBEYFILE processing involving PFID 51 should work regardless of whether you correct the incorrect PORTNUM value.

Resulting display

The resources that are displayed depend on their relationship within the hierarchy that is specified on the ID operand. The following lists show what resources are displayed for each major node or minor node.

Note: Independent LUs that are defined under a PU do not always appear in this output. Only independent LUs that are currently using the PU as a boundary function for multiple concurrent sessions are displayed.

A DISPLAY ID command issued at an APPN node might show a resource name appearing in several networks even though the resource actually exists in only one network. This can happen if intermediate SSCPs are pre-V4R1 and they pass only the 8-character resource name. The real network ID is therefore lost and other network IDs might be subsequently assumed.

For a DISPLAY ID command with IDTYPE=RESOURCE or IDTYPE=DIRECTRY, if the resource type that is displayed is EN, the node might actually be a network node, end node, or SSCP. This is because in a mixed APPN and subarea network, CPs, and SSCPs that are found in or through a subarea network are represented in this host (the host where you are issuing this command) as end nodes which are served by the interchange node through which the resource was found.

Note: If model application program definitions are included in the display, any dynamic application programs built from those models that have been deactivated are not displayed. This is because dynamic application programs cannot exist in an inactive state. When a dynamic application program is deactivated and CLOSE macro processing is complete for the dynamic application program, the definition of the dynamic application program is deleted. The dynamic application program is no longer known by VTAM and will not appear in the output of any DISPLAY commands.

- Major nodes:

...

- Minor nodes:

...

– For ID=*transport_resource_list_entry*:

- Names of the Communications Server z/OS upper-layer protocols (ULPs) using this TRLE

- For a dynamic TCP TRLE, an exclusively owned TRLE, or an internal shared memory (ISM) TRLE, only one message with a ULP ID is issued because only one ULP can use each of these TRLEs. For an OSA-Express adapter or a HiperSockets Converged Interface, one message with a ULP ID is issued for each datapath channel address that a ULP uses. For other TRLEs, more than one ULP ID message can be issued, depending on how many ULPs are using the TRLE.

Rule: Only one message with a ULP ID is generated for a RoCE Express2 feature or RoCE Express3 feature, or a 10 GbE RoCE Express feature that operates in a shared RoCE environment.

- The ULP ID will be the jobname for TCP/IP ULPs, the SNA PU name for ANNC ULPs, and the XCA Major Node name for ATM or EE ULPs.

Examples

...

Displaying a RoCE Express3 TRLE:

```
D NET,TRL,TRLE=IUT100F2
IST097I DISPLAY ACCEPTED
IST075I NAME = IUT100F2, TYPE = TRLE 991
IST1954I TRL MAJOR NODE = ISTTRL
IST486I STATUS= ACTIV, DESIRED STATE= ACTIV
IST087I TYPE = *NA*               , CONTROL = ROCE, HPDT = *NA*
IST2361I SMCR PFID = 00F2  PCHID = 0238  PNETID = PLEX1
IST2362I PORTNUM = 1  RNIC CODE LEVEL = 22.27.1016
IST2389I PFIP = 01000B00   GEN = ROCE EXPRESS3   SPEED = 25GE
IST2417I VFN = 0003
IST924I -------------------------------------------------------------
IST1717I ULPID = TCPSVT ULP INTERFACE = EZARIUT100F2
IST1724I I/O TRACE = OFF  TRACE LENGTH = *NA*
IST1866I TRLE = IUT100F2   INOPDUMP = ON
IST314I END
```

...

DISPLAY TRL command

Resulting display

The resulting display shows:

- The name and status of all TRLEs in the active TRL major nodes if the TRLE operand is not specified.
- The name and status of the TRLE specified on the TRLE operand. If the status is active and the TRLE is not associated with a "RoCE Express" feature or an ISM device, the display also includes the address and operational status of the READ, WRITE, and (OSA-Express and HiperSockets only) DATA subchannels. In addition, the following information may be displayed:

  – MPC level and usage (MPC header size, maximum MPC data size, inbound data storage medium)

  – Name of the CS z/OS upper-layer protocols (ULPs) using this TRLE

  – OSA portname, OSA adapter number, and OSA microcode level

  – OSA or HiperSockets channel path id (chpid) type and number

  – Physical channel ID (PCHID) for the "RoCE Express" feature

– Virtual channel ID (VCHID) for the ISM device

– Physical network ID (PNetID) for the "RoCE Express" feature and ISM device and HiperSockets Converged Interface

– Peripheral Component Interconnect Express (PCIe) function ID (PFID) for the "RoCE Express" feature and ISM device

– Microcode level for a 10 GbE RoCE Express feature operating in a dedicated RoCE environment, a RoCE Express2 feature or a RoCE Express3 feature

– Virtual function number (VFN) for an ISM device, a 10 GbE RoCE Express feature that operates in a shared RoCE environment, a RoCE Express2 feature or a RoCE Express3 feature

– Generation level for a "RoCE Express" feature

– Transmission speed for a "RoCE Express" feature

– I/O trace status

– The capability of the connection to perform channel I/O directly to or from communications storage manager (CSM) buffers

– Storage information about the inbound and outbound queues associated with the DATA subchannels

- For a dynamic TCP TRLE, an exclusively owned TRLE, or an ISM TRLE, only one message with a ULP ID is issued because only one ULP can use each of these TRLEs. For an OSA-Express adapter or a HiperSockets Converged Interface, one message with a ULP ID is issued for each datapath channel address that a ULP uses. For other TRLEs, more than one ULP ID message can be issued, depending on how many ULPs are using the TRLE.

  Rule: Only one message with a ULP ID is generated for a 10 GbE RoCE Express feature, a RoCE Express2 feature, or a RoCE Express3 feature that operates in a shared RoCE environment.

  ...

Examples

...

Displaying a 10 GbE RoCE Express3 TRLE:

```
d net,id=iut20120
IST097I DISPLAY ACCEPTED
IST075I NAME = IUT20120, TYPE = TRLE
IST1954I TRL MAJOR NODE = ISTTRL
IST486I STATUS= ACTIV, DESIRED STATE= ACTIV
IST087I TYPE = *NA*               , CONTROL = ROCE, HPDT = *NA*
IST2361I SMCR PFID = 0120  PCHID = 02DC  PNETID = PLEX2
IST2362I PORTNUM = 1  RNIC CODE LEVEL = 14.25.1020
IST2389I PFIP = 02001000  GEN = ROCE EXPRESS3  SPEED = 10GE
IST2417I VFN = 0001
IST924I -------------------------------------------------------------
IST1717I ULPID = TCPIP2 ULP INTERFACE = EZARIUT20120
IST1724I I/O TRACE = OFF  TRACE LENGTH = *NA*
IST314I END
```

Operands

...

RNICTRLE

Specifies that a diagnostic dump of a "RoCE Express" feature needs to be taken under certain conditions. The RNICTRLE operand can be used only with the MESSAGE trigger or as part of an immediate dump.

RNICTRLE=MSGVALUE

MSGVALUE is valid only when the MESSAGE operand is used and specifies either message IST2391I, IST2406I or IST2419I. Specifying the MSGVALUE keyword allows VTAM to collect diagnostic dump information for the "RoCE Express" feature that is identified in these messages.

RNICTRLE=RNIC_TRLEName

The format of *RNIC_TRLEName* must be IUT*yxxxx*, where *xxxx* is the Peripheral Component Interconnect Express (PCIe) function ID (PFID) that identifies the "RoCE Express" feature, and

y is the port number that is used on the "RoCE Express" interface. The value of *y* can be 1 or 2.

Usage

The "RoCE Express" diagnostic dump is taken in addition to any other dumps that CSDUMP produces. After the "RoCE Express" diagnostic dump is produced, recovery of the "RoCE Express" feature is attempted.

Notes: No "RoCE Express" diagnostic dump is taken in either of the following cases:

- The "RoCE Express" TRLE is not active when CSDUMP produces the dump.
- A specific RNIC_TRLEName is specified for RNICTRLE but the TRLE is not an RDMA over Converged Ethernet (RoCE) TRLE.

Rules:

- When the 10 GbE RoCE Express feature operates in a dedicated RoCE environment, the diagnostic dump deactivates the 10 GbE RoCE Express feature, and causes an inoperative condition for all users.
- When RNICTRLE represents a RoCE Express2 feature, a RoCE Express3 feature, or a 10 GbE RoCE Express feature operating in a shared RoCE environment, the diagnostic dump only affects the TCP/IP stack that configured the PFID value included in the value of RNIC_TRLEName. Other TCP/IP stacks that use the same feature are not affected.

Guideline: Ensure that multiple "RoCE Express" interfaces are active with the same physical network ID to avoid loss of connections during a CSDUMP operation. For more information, see High availability considerations in z/OS Communications Server: IP Configuration Guide.

...

IUT*npfid*

This TRLE is created when TCP/IP activates an IPAQENET or IPAQENET6 interface with CHPIDTYPE OSD with Shared Memory Communications - RDMA (SMC-R) specified or taken as the default, and SMC-R is enabled on the system.

- For an IBM 10 GbE RoCE Express TRLE, the *npfid* value is derived from the PORTNUM and PFID values on the SMCR parameter of the GLOBALCONFIG statement in the TCP/IP profile. For example, IUT20018 indicates that the PORTNUM value is 2 and the PFID value is 0018. If PORTNUM is not specified, the default value is 1.
- For an IBM RoCE Express2 or RoCE Express3 TRLE, the *npfid* value is derived from the PFID value on the SMCR parameter of the GLOBALCONFIG statement or the IPAQENET INTERFACE statement in the TCP/IP profile and the port number that VTAM learns during activation of the RoCE Express2 or RoCE Express3 feature. The learned port number is used instead of any PORTNUM value specified on the GLOBALCONFIG SMCR statement. For example, IUT20153 indicates that the PFID value is 153 and that the learned port number is 2.

No subchannels are associated with this TRLE.

SNA Diagnosis Volume 2: FFST Dumps and the VIT

HCQ entry for invoking a RoCE HCQ operation (Part 1)

Entry:

HCQ

VIT option:

CIA

Event:

Invocation of a Remote Direct Memory Access (RDMA) over Converged Ethernet (RoCE) HCA Command Queue (HCQ) operation with an IBM RoCE Express2 feature or IBM RoCE Express3 feature, as part of Shared Memory Communications over Remote Direct Memory Access (SMC-R) processing.

HCQ2 entry for invoking a RoCE HCQ operation (Part 2)

Entry:

HCQ2

VIT option:

CIA

Event:

Invocation of a Remote Direct Memory Access (RDMA) over Converged Ethernet (RoCE) HCA Command Queue (HCQ) operation with an IBM RoCE Express2 feature or IBM RoCE Express3 feature, as part of Shared Memory Communications over Remote Direct Memory Access (SMC-R) processing.

HCQ3 entry for invoking a RoCE HCQ operation (Part 3)

Entry:

HCQ3

VIT option:

CIA

Event:

Invocation of a Remote Direct Memory Access (RDMA) over Converged Ethernet (RoCE) HCA Command Queue (HCQ) operation with an IBM RoCE Express2 feature or RoCE Express3 feature, as part of Shared Memory Communications over Remote Direct Memory Access (SMC-R) processing.

HCQ4 entry for invoking a RoCE HCQ operation (Part 4)

Entry:

HCQ4

z/OS Communications Server

VIT option:

CIA

Event:

Invocation of a Remote Direct Memory Access (RDMA) over Converged Ethernet (RoCE) HCA Command Queue (HCQ) operation with an IBM RoCE Express2 feature or IBM RoCE Express3 feature, as part of Shared Memory Communications over Remote Direct Memory Access (SMC-R) processing.

HCQ5 entry for invoking a RoCE HCQ operation (Part 5)

Entry:

HCQ5

VIT option:

CIA

Event:

Invocation of a Remote Direct Memory Access (RDMA) over Converged Ethernet (RoCE) HCA Command Queue (HCQ) operation with an IBM RoCE Express2 feature or IBM RoCE Express3 feature, as part of Shared Memory Communications over Remote Direct Memory Access (SMC-R) processing.

HCQ6 entry for invoking a RoCE HCQ operation (Part 6)

Entry:

HCQ6

VIT option:

CIA

Event:

Invocation of a Remote Direct Memory Access (RDMA) over Converged Ethernet (RoCE) HCA Command Queue (HCQ) operation with an IBM RoCE Express2 feature or IBM RoCE Express3 feature, as part of Shared Memory Communications over Remote Direct Memory Access (SMC-R) processing.

RNICTRLE

Specifies that a diagnostic dump of a "RoCE Express" feature needs to be taken under certain conditions. The RNICTRLE operand can be used only with the MESSAGE trigger.

RNICTRLE=MSGVALUE

MSGVALUE is valid only when the MESSAGE operand is used, and MESSAGE specifies either message IST2391I, IST2406I or IST2419I. Specifying the MSGVALUE keyword allows VTAM to collect diagnostic dump information for the "RoCE Express" feature identified in these messages.

RNICTRLE=RNIC_TRLEName

The format of RNIC_TRLEName must be IUT*yxxxx*, where *xxxx* is the Peripheral Component Interconnect® Express (PCIe) function ID (PFID) that identifies the "RoCE Express" feature, and y is the port number used on the "RoCE Express" interface. The value of y can be 1 or 2.

Usage

The "RoCE Express" diagnostic dump is taken in addition to any other dumps that CSDUMP produces. After the "RoCE Express" diagnostic dump is produced, recovery of the "RoCE Express" feature is attempted.

Notes: No "RoCE Express" diagnostic dump is taken in either of the following cases:

- The TRLE is not active when CSDUMP produces the dump.
- A specific TRLE value is coded for RNICTRLE but the TRLE is not an RDMA over Converged Ethernet (RoCE) TRLE.

Rules:

- When the "RoCE Express" operates in a dedicated RoCE environment, the diagnostic dump deactivates the 10 GbE RoCE Express feature, and causes an inoperative condition for all users.
- When RNICTRLE represents a RoCE Express2 feature, a RoCE Express3 feature, or when the 10 GbE RoCE Express feature operates in a shared RoCE environment, the diagnostic dump only affects the TCP/IP stack that configured the PFID value included in the value of RNIC_TRLEName. Other TCP/IP stacks that use the same feature are not affected.

Guideline: Ensure that multiple "RoCE Express" interfaces are active with the same physical network ID to avoid loss of connections during a CSDUMP operation. For more information, see High availability considerations in z/OS Communications Server: IP Configuration Guide.

Table. Bytes 2 and 3 (completion code) of the DLC status code (continued)

Hexadecimal Code	Meaning
X'5116'	Command operation timeout **Explanation:** One of the following cases happened: • The 10 GbE RoCE Express device driver issued a Host channel adapter (HCA) configuration register (HCR) command to the RoCE hardware, but the hardware did not complete the operation within the internally specified timeout threshold. • The RoCE Express2 or RoCE Express3 device driver issued an HCA command queue (HCQ) command to the RoCE hardware, but the hardware did not complete the operation within the internally specified timeout threshold. The "RoCE Express" device driver initiates INOP processing to recover from the error.
X'5117'	PCIe load operation failure **Explanation:** During the processing of an HCR operation for a 10 GbE RoCE Express feature or an HCQ operation for a RoCE Express2 feature or RoCE Express3 feature, the "RoCE Express" device driver received an error in response to a PCIe store operation. The "RoCE Express" device driver might initiate INOP processing to recover from this error.
X'5118'	PCIe store operation failure **Explanation:** During the processing of an HCR operation for a 10 GbE RoCE Express feature or an HCQ operation for a RoCE Express2 feature or RoCE Express3 feature, the "RoCE Express" device driver received an error in response to a PCIe store operation. The "RoCE Express" device driver might initiate INOP processing to recover from this error.
X'5121'	Command operation failure **Explanation:** One of the following cases happened: • The 10 GbE RoCE Express device driver issued an HCR command to the RoCE hardware, but the hardware rejected the operation with a specific status code. • The RoCE Express2 or RoCE Express3 device driver issued an HCQ command to the RoCE hardware, but the hardware rejected the operation with a specific status code. The specific operation failed.

IP Messages: Volume 4 (EZZ, SNM)

EZZ4336I

EZZ4336I ERROR DURING *link_control_function* INTERFACE *interface_name* –

CODE *error_code* DIAGNOSTIC CODE *internal_diagnostic_code*

Explanation

The Link Layer detected an error during activation of the interface.

link_control_function is the function that is being performed on the interface.

interface_name is the name of the interface.

error_code is the Data Link Control (DLC) status code for the link layer.

internal_diagnostic_code is an internal diagnostic code for use by IBM.

System action

The TCP/IP stack takes one of the following actions:

- If the *interface_name* value represents a RoCE or an internal shared memory (ISM) interface and the *link_control_function* value is ENABLE CALLS TO, an error occurred while TCP/IP was registering a VLAN ID with the IBM 10 GbE RoCE Express feature or ISM device. The registration failure might be a device issue or result from the fact that the registration request exceeded the maximum number of VLAN IDs that can be registered with the device. The interface remains active, but any TCP connections that are established across this VLAN ID might not use Shared Memory Communications (SMC) processing.
- If *the interface_name* value represents a RoCE Express2 or RoCE Express3 interface and the *link_control_function* value is DURING ACTIVATION OF, an error occurred during activation processing of the underlying RoCE Express2 or RoCE Express3 feature. Despite the error, a RoCE Express2/RoCE Express3 interface was created and remains active. The interface name is created dynamically by using the form EZARIUT*ppppp*, where p is the configured or defaulted port number from the GLOBALCONFIG SMCR profile statement and *ffff* is the PFID value that represents the RoCE Express2 or RoCE Express3 feature. This dynamically created *interface_name* value might not use the actual port number configured in HCD for the RoCE Express2 or RoCE Express3 feature because VTAM and the TCP/IP stack learn the port number configured in HCD for the feature during activation of the interface. After activation, VTAM and the TCP/IP stack use the learned port number instead of any configured port number for the value of p in the *interface_name* value.

  For more information about the GLOBALCONFIG SMCR operand, see GLOBALCONFIG statement in z/OS Communications Server: IP Configuration Reference.

- In all other cases, TCP/IP deactivates the interface.

Operator response

- If the *interface_name* value represents a RoCE Express2 or RoCE Express3 interface and the *link_control_function* value is DURING ACTIVATION OF, contact the system programmer to resolve the underlying problem with the activation of the RoCE Express2 or RoCE Express3 feature. After the problem is resolved, attempt to activate the interface again.

  After activation, the name of the interface might change if VTAM and the TCP/IP stack determine that a different port number has been configured for the RoCE Express2 or RoCE Express3 feature.

- If the last 4 digits of the error code are X'3016', the most likely reason for the error is that the TRLE definition for the interface is not active. In this case, activate the TRLE and restart the interface. Otherwise, inform the system programmer about the error.

System programmer response

See the z/OS Communications Server: IP and SNA Codes for information about Data Link Control (DLC) status codes for the link layer and perform the action described for the indicated status code. If applicable, correct the hardware problem and restart the interface.

Module

TCPIP

Procedure name

EZBIFIUT

SNA Messages

IST2361I

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.

...

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 RoCE Express2 or RoCE Express3 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 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.

ROCE EXPRESS3

The TRLE represents an IBM 10 GbE RoCE Express3 or IBM 25 GbE RoCE Express3 feature.

speed

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

10GE

The RoCE Express feature uses 10 gigabit Ethernet ports.

25GE

The RoCE Express feature uses 25 gigabit Ethernet ports.

...

IST2396I

IST2396I **RNIC STATISTICS FOR *trlename***

Explanation

VTAM issues this group of messages in response to a DISPLAY TRL,TRLE=*trlename*,DEVSTATS command when *trlename* represents a "RoCE Express" interface.

...

IST2398I

This message displays the number of occurrences for the statistic described by *description*. The possible combinations of *overflow*, *count*, and *description* are:

- **INBOUND FRAMES DROPPED** = *overflow count*

Represents the number of inbound frames that were dropped on this "RoCE Express" interface.

Restriction: This value is always 0 for RoCE Express2 and RoCE Express3 adapters.

- **INBOUND FRAME ERRORS** = *overflow count*

Represents the number of inbound frames that had errors on this "RoCE Express" interface.

Restriction: This value is always 0 when the 10 GbE RoCE Express adapter operates in a shared RoCE environment.

- **INBOUND RDMA FRAMES** = *overflow count*

Represents the number of RoCE frames received on this "RoCE Express" interface.

- **INBOUND RDMA OCTETS** = *overflow count*

Represents the number of RoCE octets received on this "RoCE Express" interface.

- **OUTBOUND FRAMES DROPPED** = *overflow count*

Represents the number of outbound frames that were dropped on this "RoCE Express" interface.

Restriction: This value is always 0 for RoCE Express2 and RoCE Express3 adapters.

...

IST2444I

IST2444I **PORTNUM *port_value* IGNORED FOR SMC-R PFID *pfid_value*, ACTIVATION CONTINUES**

Explanation

VTAM issues this message when it attempts to activate an RDMA network interface card (RNIC) interface and the port number configured on the TCP/IP GLOBALCONFIG SMCR statement for this PCIe function ID (PFID) does not match the port number value configured for this PFID in the hardware configuration definition (HCD). The port number configured in the HCD is used and the port number configured on the GLOBALCONFIG statement is ignored.

The RoCE Express2/RoCE Express3 feature requires that the port number be defined for a given PFID in the HCD, so the PORTNUM operand on a GLOBALCONFIG SMCR statement that defines a RoCE Express2/RoCE Express3 PFID is not necessary. VTAM learns the port number during RNIC interface activation and provides the real port number to the TCP/IP stack.

...

Trademarks