



z/OS V2R1 CS: Shared Memory Communications - RDMA (SMC-R), Part 2

David Herr – dherr@us.ibm.com
IBM Raleigh, NC

Tuesday, August 13th, 11:00am
Session: 13628



Trademarks

The following are trademarks of the International Business Machines Corporation in the United States and/or other countries.

AIX*	DB2*	HiperSockets*	MQSeries*	PowerHA*	RMF	System z*	zEnterprise*	z/VM*
BladeCenter*	DFSMS	HyperSwap	NetView*	PR/SM	Smarter Planet*	System z10*	z10	z/VMSE*
CICS*	EASY Tier	IMS	OMEGAMON*	PureSystems	Storwize*	Tivoli*	z10 EC	
Cognos*	FICON*	InfiniBand*	Parallel Sysplex*	Rational*	System Storage*	WebSphere*	z/OS*	
DataPower*	GDPS*	Lotus*	POWER7*	RACF*	System x*	XIV*		

* Registered trademarks of IBM Corporation

The following are trademarks or registered trademarks of other companies.

Adobe, the Adobe logo, PostScript, and the PostScript logo are either registered trademarks or trademarks of Adobe Systems Incorporated in the United States, and/or other countries.

Cell Broadband Engine is a trademark of Sony Computer Entertainment, Inc. in the United States, other countries, or both and is used under license therefrom.

Intel, Intel logo, Intel Inside, Intel Inside logo, Intel Centrino, Intel Centrino logo, Celeron, Intel Xeon, Intel SpeedStep, Itanium, and Pentium are trademarks or registered trademarks of Intel Corporation or its subsidiaries in the United States and other countries.

IT Infrastructure Library is a registered trademark of the Central Computer and Telecommunications Agency which is now part of the Office of Government Commerce.

ITIL is a registered trademark, and a registered community trademark of the Office of Government Commerce, and is registered in the U.S. Patent and Trademark Office.

Java and all Java based trademarks and logos are trademarks or registered trademarks of Oracle and/or its affiliates.

Linear Tape-Open, LTO, the LTO Logo, Ultrium, and the Ultrium logo are trademarks of HP, IBM Corp. and Quantum in the U.S. and

Linux is a registered trademark of Linus Torvalds in the United States, other countries, or both.

Microsoft, Windows, Windows NT, and the Windows logo are trademarks of Microsoft Corporation in the United States, other countries, or both.

OpenStack is a trademark of OpenStack LLC. The OpenStack trademark policy is available on the [OpenStack website](#).

TEALEAF is a registered trademark of Tealeaf, an IBM Company.

Windows Server and the Windows logo are trademarks of the Microsoft group of countries.

Worklight is a trademark or registered trademark of Worklight, an IBM Company.

UNIX is a registered trademark of The Open Group in the United States and other countries.

* Other product and service names might be trademarks of IBM or other companies.

Notes:

Performance is in Internal Throughput Rate (ITR) ratio based on measurements and projections using standard IBM benchmarks in a controlled environment. The actual throughput that any user will experience will vary depending upon considerations such as the amount of multiprogramming in the user's job stream, the I/O configuration, the storage configuration, and the workload processed. Therefore, no assurance can be given that an individual user will achieve throughput improvements equivalent to the performance ratios stated here.

IBM hardware products are manufactured from new parts, or new and serviceable used parts. Regardless, our warranty terms apply.

All customer examples cited or described in this presentation are presented as illustrations of the manner in which some customers have used IBM products and the results they may have achieved. Actual environmental costs and performance characteristics will vary depending on individual customer configurations and conditions.

This publication was produced in the United States. IBM may not offer the products, services or features discussed in this document in other countries, and the information may be subject to change without notice. Consult your local IBM business contact for information on the product or services available in your area.

All statements regarding IBM's future direction and intent are subject to change or withdrawal without notice, and represent goals and objectives only.

Information about non-IBM products is obtained from the manufacturers of those products or their published announcements. IBM has not tested those products and cannot confirm the performance, compatibility, or any other claims related to non-IBM products. Questions on the capabilities of non-IBM products should be addressed to the suppliers of those products.

Prices subject to change without notice. Contact your IBM representative or Business Partner for the most current pricing in your geography.

This information provides only general descriptions of the types and portions of workloads that are eligible for execution on Specialty Engines (e.g. zIIPs, zAAPs, and IFLs) ("SEs"). IBM authorizes customers to use IBM SE only to execute the processing of Eligible Workloads of specific Programs expressly authorized by IBM as specified in the "Authorized Use Table for IBM Machines" provided at www.ibm.com/systems/support/machine_warranties/machine_code/aut.html ("AUT"). No other workload processing is authorized for execution on an SE. IBM offers SE at a lower price than General Processors/Central Processors because customers are authorized to use SEs only to process certain types and/or amounts of workloads as specified by IBM in the AUT.

Agenda



Shared Memory Communications – RDMA (SMC-R), Part 2

- SMC-R Configuration and Monitoring
- SMC-R Diagnosis
- Appendix - SMC-R Network Management Interface and SMF enhancements



13627: z/OS V2R1 CS: Shared Memory Communications - RDMA (SMC-R), Part 1
Tuesday, August 13, 2013: 9:30 AM-10:30 PM
Room 206 (Hynes Convention Center)
Speakers: [Gus Kassimis](#) (IBM Corporation) and [Dave Herr](#) (IBM Corporation)

For answers to frequently asked questions on SMC-R and RoCE please see:

<http://www-03.ibm.com/support/techdocs/atmastr.nsf/WebIndex/FQ131485>

Disclaimer: All statements regarding IBM future direction or intent, including current product plans, are subject to change or withdrawal without notice and represent goals and objectives only. All information is provided for informational purposes only, on an “as is” basis, without warranty of any kind.



SMC-R Configuration and Monitoring

SMC-R Configuration and Monitoring

- System requirements
- TCP/IP Profile changes
- Netstat report and TCP/IP display changes
- VTAM command changes
- SMF enhancements

SMC-R System Requirements

- Before using SMC-R, you must take these actions:
 - Configure these values using Hardware Configuration Definition (HCD):
 - PCIe function ID (PFID)
 - Configure two PFIDs per physical network for redundancy
 - Physical network ID (PNet ID) for OSA and RNIC interfaces
 - **NOTE: PNet IDs are required for SMC-R enabled OSD devices and RoCE adapters**
 - Configure Ethernet switches appropriately
 - Optionally define VLAN ID values to be used
 - Enable “flow control” capability

SMC-R HCD Configuration

HCD Main Panel:
1
3 (Processor list)
Then select (/) processor

Goto Filter Backup Query Help

----- Actions on selected processors

Command ==>

Select one or

/ Proc. ID Ty	8_ 1. Add like (a)
_ D76 20	2. Repeat (Copy) processor configurations (r)
_ H87 20	3. Change (c)
_ H88 20	4. *Prime serial number (i)
_ MR31 28	5. Delete (d)
_ PBUV4 28	6. View processor definition (v)
/ P88 28	7. View related CTC connections (k)
*****	8. Work with functions (f)
	9. Work with partitions (SMP) (p)
	10. Work with attached channel paths (SMP) (s)
	11. Work with attached devices . . . (SMP) (u)
	12. Copy to channel subsystem . . . (SMP) (y)
	13. Work with channel subsystems . . (XMP) (p,s)

* = requires TSA I/O Operations

F1=Help F1=Help F2=Split F3=Exit F9=Swap F12=Cancel
 F8=Forward
 F20=Right F22=Command

SMC-R HCD Configuration

8 RoCE Adapters defined
Function IDs used by
TCPIP config

```

Goto  Filter  Backup  Query  Help
-----
                                PCIe Function List          Row 1 of 8 Mo
Command ==> _____ Scroll =
Select one or more PCIe functions, then press Enter. To add, use F11.
Processor ID . . . . . : P88                zHelix P88


/ FID  PCHID  VF+  Type+  Description
a 001   380    ___  ROCE
- 002   3A0    ___  ROCE
- 003   3C0    ___  ROCE
- 004   3E0    ___  ROCE
- 005   500    ___  ROCE
- 006   520    ___  ROCE
- 007   540    ___  ROCE
- 008   560    ___  ROCE
***** Bottom of data *****

F1=Help      F2=Split      F3=Exit      F4=Prompt      F5=Backward
F8=Forward   F9=Swap       F10=Actions  F11=Add        F12=Cancel    F13=Instruct
F20=Right    F22=Command
  
```

Physical Channel IDs

SMC-R HCD Configuration – Define a PFID

```
Goto Filter Backup Query Help
-----
PCie Function List
Command ==> _____ Scroll ==> PAGE
S _____ Add PCie Function _____
P
Specify or revise the following values.
Processor ID . . . . . : P88          zHelix P88
Function ID . . . . . : ROCE
Type . . . . . : _____ +
PCHID . . . . . : 380
Virtual Function ID . . . . . : _____ +
Description . . . . . : _____
* F1=Help   F2=Split   F3=Exit   F4=Prompt   F5=Reset   F9=Swap
  F12=Cancel
F20=Right  F22=Command
```



SMC-R HCD Configuration – Add PNETID

```

Goto  Filter  Backup  Query  Help
-
- Add/Modify Physical Network
C
S  If the PCHID is associated to one or more phys
S  each physical network ID corresponding to each
P  Physical network ID 1 . . NETWORK1
P  Physical network ID 2 . . _____
P  Physical network ID 3 . . _____
P  Physical network ID 4 . . _____
/
c
-
- F1=Help      F2=Split     F3=Exit      F5=Reset     F9=Swap     F12=Cancel
-
- 006      520      ___      ROCE      _____
- 007      540      ___      ROCE      _____
- 008      560      ___      ROCE      _____
-
***** Bottom of data *****
F1=Help      F2=Split     F3=Exit      F4=Prompt     F5=
F8=Forward    F9=Swap     F10=Actions  F11=Add       F12=
F20=Right    F22=Command

```

**NOTE: Each Physical Network ID entry correlates to a physical port on the Adapter.
For RoCE Express, entry 1 correlates to port 1, entry 2 to port 2.
For OSA Express entry 1 correlates to Port 0, entry 2 to Port 1**

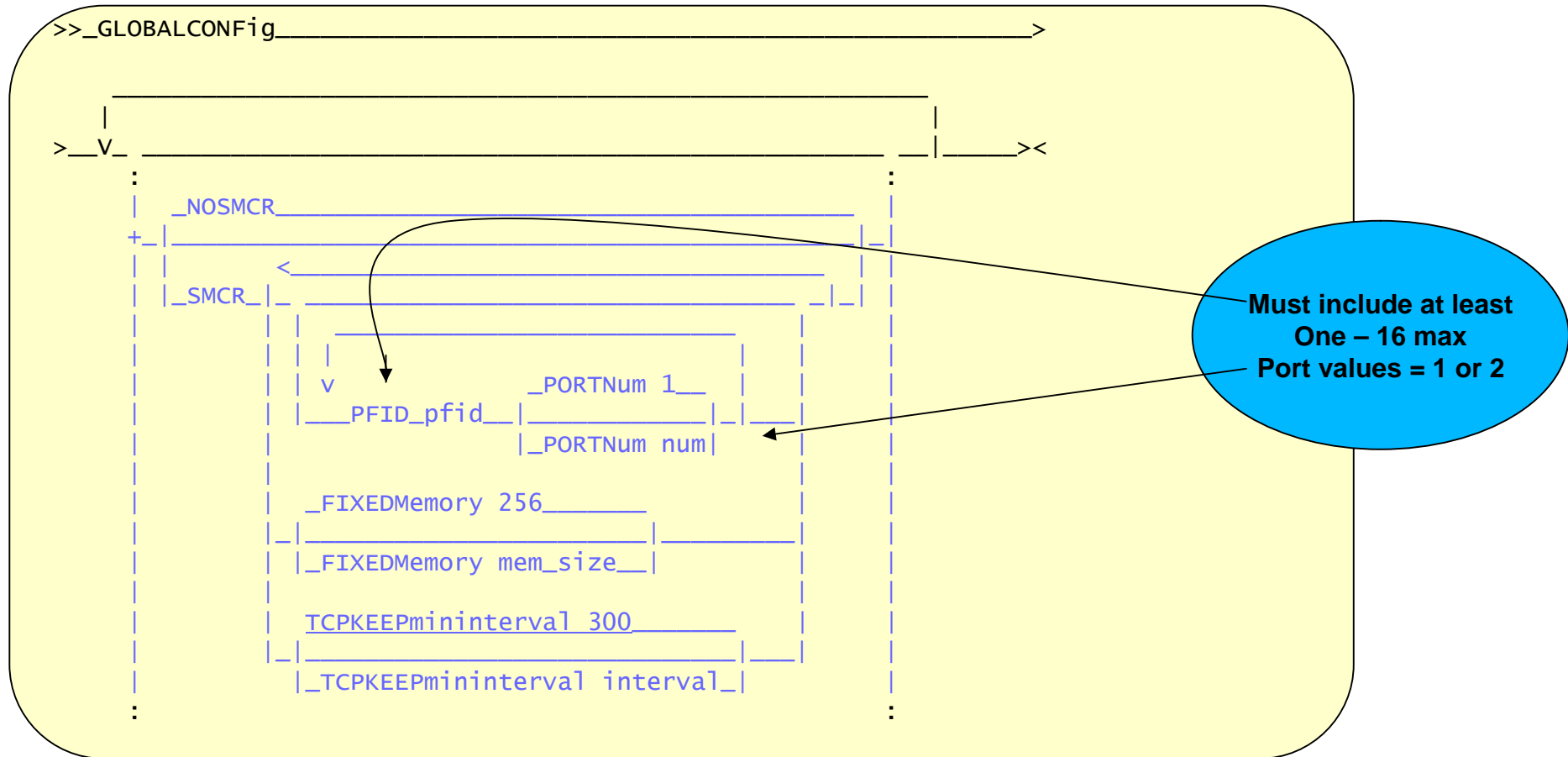
Associates this RoCE adapter to an OSA interface with Same PNETID

SMC-R TCP/IP Configuration

- GLOBALCONFIG statement – Required update
- IPAQENET INTERFACE statement
- IPAQENET6 INTERFACE statement
- PORT statement
- PORTRANGE statement

- SMFCONFIG statement
 - Details are covered with the Network Management enhancements - Appendix

SMC-R TCP/IP Configuration



EZARIUTxyyyy for RNIC interface, **IUTxyyyy** for TRLE, where **x** = PORTNUM and **yyyy** = PFID

SMC-R TCP/IP Configuration – Enable/Disable

- Switching from SMCR to NOSMCR
 - Prevents new TCP connections from using SMC-R and new SMC-R links from being started
 - Existing SMC-R links and TCP connections unaffected
 - SMC-R links are deleted when no more TCP connections are using them
- Switching from NOSMCR to SMCR
 - Existing TCP connections are unaffected, but new TCP connections are eligible to use SMC-R
 - Previous SMCR settings, if any, are used unless new values provided

SMC-R TCP/IP Configuration – Modifying PFIDs

- Full replacement of PFID values
 - PFIDs that you want to continue using must be included on GLOBALCONFIG SMCR statement
 - New PFIDs in list are automatically started, assuming an SMC-R capable OSD interface had been started previously
- Steps to delete an existing RNIC interface
 - VARY STOP the RNIC interface
 - Delete PFID value from the GLOBALCONFIG SMCR statement
 - Issue VARY OBEYFILE
 - RNIC interface is deleted when VARY OBEYFILE completes successfully

SMC-R TCP/IP Configuration – FIXEDMemory

- Maximum amount of fixed 64-bit private storage that TCP/IP can use for SMC-R processing
 - Includes RMB storage and staging buffer storage
- Valid range is 30-9999 (megabytes)
 - Defaults to FIXEDM 256 (megabytes)
- Can be changed using VARY OBEYFILE
 - If SMCR statement specified without FIXEDMemory, the limit is unchanged
 - Lowering the storage limit does not impact existing SMC-R links or TCP connections

SMC-R TCP/IP Configuration – Fixed Memory estimation

Configuration and workload assumptions:

2 PFIDs, all on the same physical network

12 SMC-R link groups expected (3 VLANs, 4 peers per VLAN)

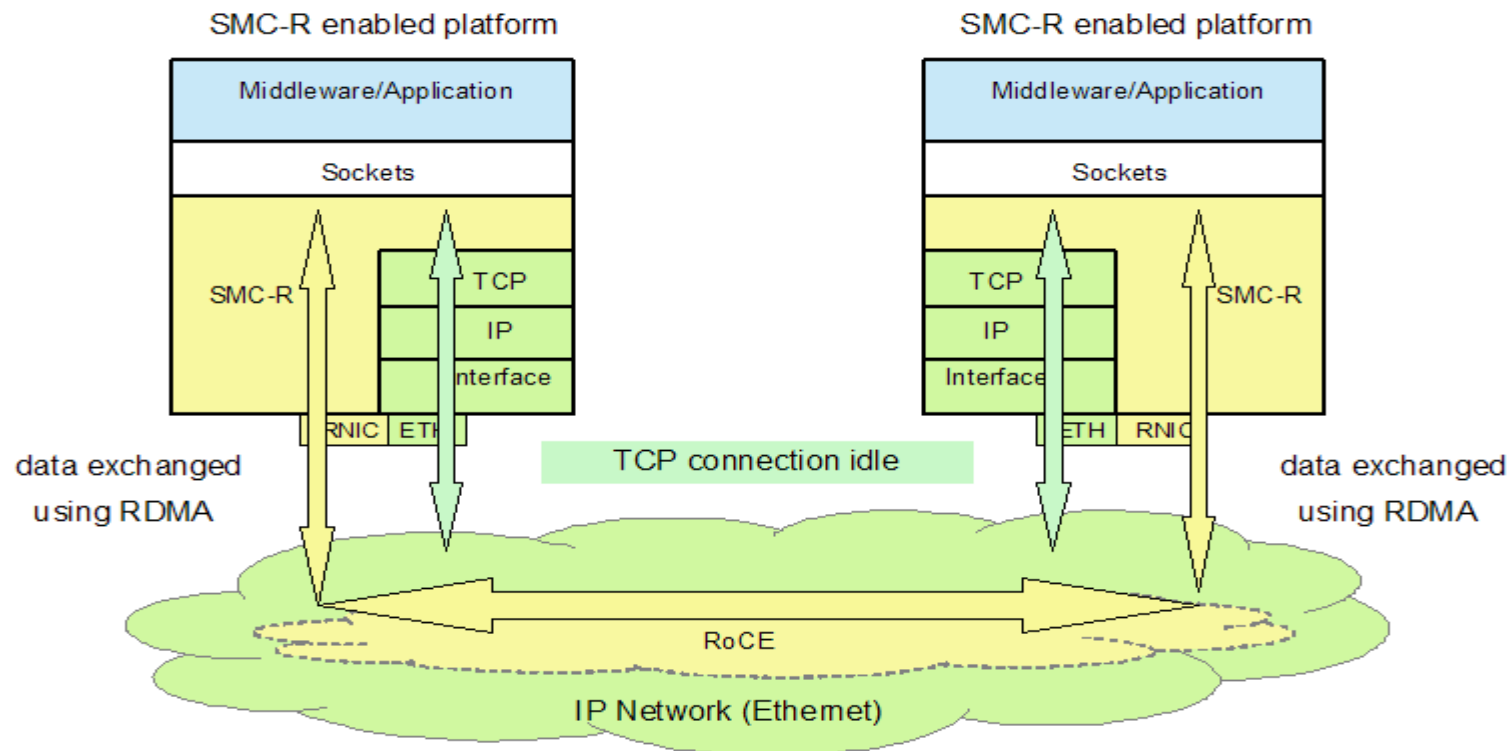
224 TCP connections will use SMC-R

"Staging buffers"	=	8M
PFIDs (2 PFIDs*1M)	=	2M
RMBs (12 link groups*3M)	=	36M
Workload ((224 connections/8)*1M)	=	<u>28M</u>
TOTAL	=	74M

**Estimation only!
Use Display TCPIP,,STOR
to determine the
right amount for you**

TCP connections using SMC-R appear idle - KEEPALIVE

- All application data flows “out-of-band” with SMC-R
- TCP connection is maintained, but just for control purposes



SMC-R TCP/IP Configuration – SMCR

TCPKEEPmininterval

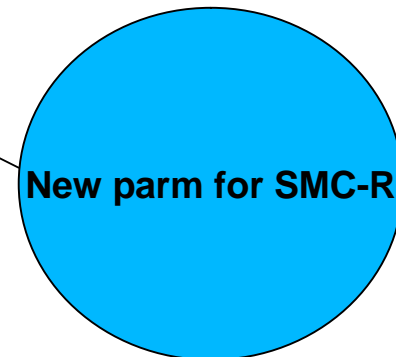
- Load balancers or firewalls use data traffic as an indication that a TCP connection is healthy
 - Might terminate the connection if no data flows within a certain period of time
- TCP keepalive processing periodically sends a packet over existing TCP connections
 - Application indicates connection is eligible for keepalive by specifying the `SO_KEEPALIVE` `setsockopt()` option
 - Time interval to use is determined by these criteria:
 - `TCP_KEEPALIVE` `setsockopt()` option, if specified
 - `TCPCONFIG INTERVAL` value, or default

SMC-R TCP/IP Configuration – *SMCR* *TCPKEEPmininterval*

- Defines, in seconds, the minimum interval that TCP keepalive packets are sent for TCP connections using SMC-R links
- Valid range is 0 – 2147460 (seconds)
 - 0 means no TCP keepalive packets are to be sent
 - Defaults to TCPKEEP 300 (seconds, or five minutes)
- Can be changed using VARY OBEYFILE
 - If SMCR statement specified without TCPKEEPmininterval, the minimum interval is unchanged
 - The changed value applies to existing SMC-R links and the TCP connections using those links

SMC-R keepalive example

- Assume these values have been specified:
 - Application specifies SO_KEEPALIVE and TCP_KEEPALIVE setsockopt() as 5 minutes
 - TCPCONFIG INTERVAL set to 10 minutes
 - GLOBALCONFIG SMCR TCPKEEP set to 25 minutes
- For TCP connections that use SMC-R:
 - TCP connection probes sent every 25 minutes
 - SMC-R link probes sent every 5 minutes
- For TCP connections that do not use SMC-R:
 - TCP connection probes sent every 5 minutes

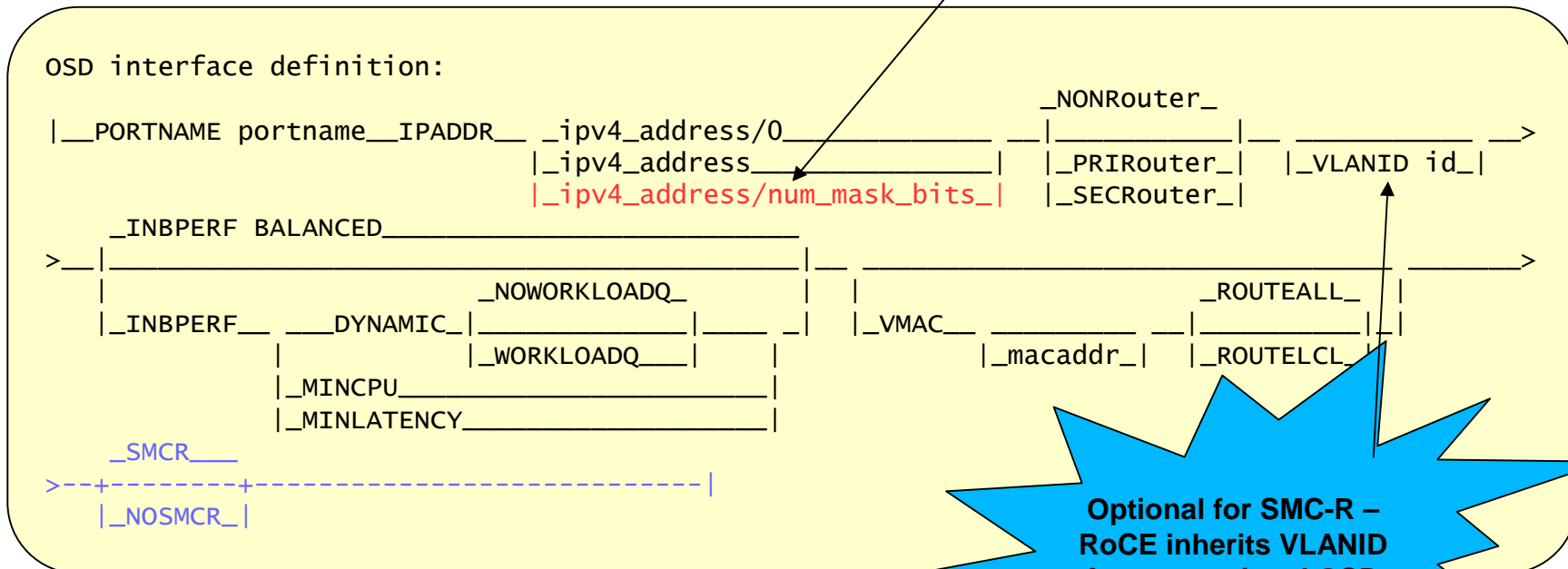


SMC-R TCP/IP Configuration – OSA Interface

- IPAQENET INTERFACE statements

- SMCR only valid for CHPIDTYPE OSD
- SMCR cannot be used with IPv4 OSD interfaces defined using DEVICE and LINK statements

Must be non-zero subnet
For SMC-R

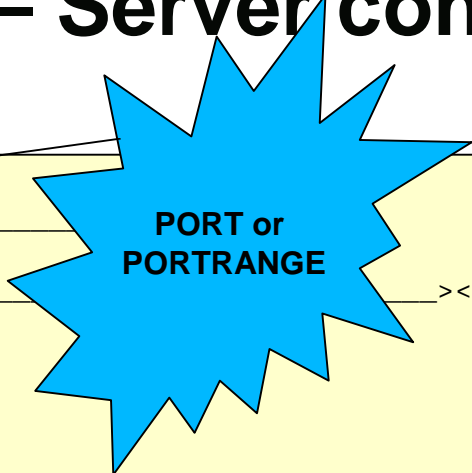


SMC-R TCP/IP Configuration – Server control

```
>> _PORT_ _V_num_ TCP _RESERVED_ ><
      |_____|
      | _jobname_ |
      | _Options_ |

Options:
      _DELAYAcks_
|_____|
| _NOAUTOLog_ | | _NODELAYAcks_ | | _SHAREPort_ | | _BIND ipaddr_ | | _SAF resname_ |
|_____| |_____| |_____| |_____|
      | _SHAREPORTWLM_ |

>> _NOSMCR_ |
```



Do not use SMC-R for applications connecting to this port(s) - ie, short-lived connections

SMC-R Monitoring - PCIe

- Activation of first SMC-R capable OSD causes PFIDs to be activated
- Use DISPLAY PCIe command to display defined PFIDs

```
D PCIe
IQP022I 10.35.27 DISPLAY PCIe
PCIe      0011 ACTIVE
PFID      DEVICE TYPE NAME      STATUS ASID  JOBNAME PCHID
001C      10GbE RoCE                   CNFG  ← 0578
0015      10GbE RoCE                   STNBY ← 052C
0018      10GbE RoCE                   CNFG  ← 0540
```

```
V TCPIP,TCPCS1,START,OSD1
EZZ0060I PROCESSING COMMAND: VARY TCPIP,TCPCS1,START,OSD1
EZZ0053I COMMAND VARY START COMPLETED SUCCESSFULLY
EZZ4340I INITIALIZATION COMPLETE FOR INTERFACE OSD1
EZZ4340I INITIALIZATION COMPLETE FOR INTERFACE EZARIUT1001C
EZZ4340I INITIALIZATION COMPLETE FOR INTERFACE EZARIUT20018
```

```
D PCIe
IQP022I 10.38.17 DISPLAY PCIe
PCIe      0011 ACTIVE
PFID      DEVICE TYPE NAME      STATUS ASID  JOBNAME PCHID
001C      10GbE RoCE                   ALLC  0039 VTAMCS 0578
0015      10GbE RoCE                   STNBY ← 052C
0018      10GbE RoCE                   ALLC  0039 VTAMCS 0540
```

Two RoCEs ready for use on this LPAR

First SMC-R enabled OSD activated – All CNFG RoCEs activated

SMC-R Monitoring – RoCE statistics

- *DISPLAY TRL,TRLE,DEVSTATS* output
- Statistics represent **adapter** activity
- DEVSTATS new and only valid for RoCE devices

```
D NET,TRL,TRLE=IUT1001C,DEVSTATS  
IST097I DISPLAY ACCEPTED
```

```
...  
IST314I END
```

```
IST2396I RNIC STATISTICS FOR IUT1001C
```

IST2397I DESCRIPTION	OVERFLOW	COUNT
IST924I -----		
IST2398I INBOUND RDMA FRAMES	0	65535
IST2398I INBOUND RDMA OCTETS	2	4294967295
IST2398I INBOUND FRAME ERRORS	0	0
IST2398I INBOUND DROPPED FRAMES	0	0
IST2398I OUTBOUND RDMA FRAMES	0	65160
IST2398I OUTBOUND RDMA OCTETS	2	4414812756
IST2398I OUTBOUND FRAME ERRORS	0	0
IST2398I OUTBOUND DROPPED FRAMES	0	0
IST314I END		

SMC-R Monitoring – Netstat changes

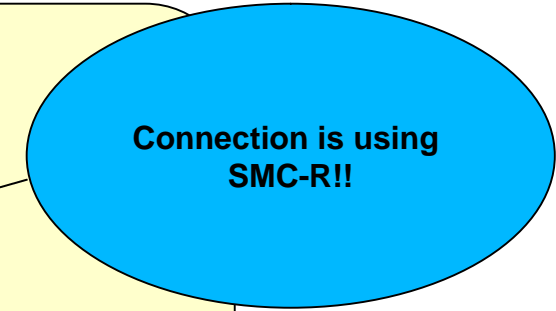
- Netstat ALL/-A report
- Netstat ALLConn/-a report
- Netstat CONFIG/-f report
- Netstat COnn/-C report
- Netstat DEvlinks/-d report
- Netstat PORTList/-o report
- Netstat STATS/-S report
- D TCPIP,,STOR command

SMC-R Monitoring – Netstat changes

- Netstat ALL/-A report provides SMC-R information about TCP connections when SMC-R is enabled
 - If TCP connection uses SMC-R, provides SMC link ID and link group ID information
 - If TCP connection does not use SMC-R, provides reason code
- All three connection reports support SMCID/-U filter
 - Reports only those connections using the specified SMC-R link or link group
 - Can specify * on the filter to report all connections using SMC-R

SMC-R Monitoring – Netstat ALL changes

```
D TCPIP,TCPCS1,NETSTAT,ALL,IPPORT=10.1.1.14+21
EZD0101I NETSTAT CS V2R1 TCPCS1
CLIENT NAME: FTPDOE34          CLIENT ID: 0000003B
LOCAL SOCKET: ::FFFF:10.1.1.14..21
FOREIGN SOCKET: ::FFFF:10.1.1.24..1024
...
SMC INFORMATION:
SMCSTATUS:      ACTIVE          SMCGROUPID:      2D8F0100
LOCALSMCLINKID: 2D8F0101        REMOTESMCLINKID: 729D0101
-----
1 OF 1 RECORDS DISPLAYED
END OF THE REPORT
```



- SMCID filter – Show only SMC-R connections with xxxxxxxx link or group ID
 - Asterisk (*) can be specified to show all SMC-R connections
 - Works on CONN and ALLCONN commands

SMC-R Monitoring – Netstat ALL changes

```
D TCPIP,TCPCS1,NETSTAT,ALL,IPPORT=10.1.1.14+21
EZD0101I NETSTAT CS V2R1 TCPCS1
CLIENT NAME: FTPDOE34                CLIENT ID: 0000003B
LOCAL SOCKET: ::FFFF:10.1.1.14..21
FOREIGN SOCKET: ::FFFF:10.1.1.24..1024
...
SMC INFORMATION:
SMCSTATUS:          INACTIVE
SMCREASON:          00005301 - PEER DID NOT ACCEPT SMC-R REQUEST
-----
1 OF 1 RECORDS DISPLAYED
END OF THE REPORT
```

Connection is not using SMC-R!!

```
D TCPIP,TCPCS1,NETSTAT,ALL,IPPORT=10.1.1.14+21
EZD0101I NETSTAT CS V2R1 TCPCS1
CLIENT NAME: FTPDOE34                CLIENT ID: 0000003B
LOCAL SOCKET: ::FFFF:10.1.1.14..21
FOREIGN SOCKET: ::FFFF:10.1.1.24..1024
...
SMC INFORMATION:
SMCSTATUS:          INACTIVE
SMCREASON:          00008888 - *Peer generated*
-----
1 OF 1 RECORDS DISPLAYED
END OF THE REPORT
```

Connection is not using SMC-R – Reason generated by Peer

SMC-R Monitoring – Netstat ALL

- For SMC-R connections:
 - BytesIn and BytesOut equal data sent/received on this SMC-R link for this connection
 - SegmentsIn and SegmentsOut are count of RDMA read/write operations
 - Other fields reflect the TCP component of the connection

```
D TCPIP,TCPCS1,NETSTAT,ALL,IPPORT=10.1.1.14+21
```

```
MVS TCP/IP NETSTAT CS V2R1          TCPIP Name: TCPCS          21:42:39
Client Name: FTPD1                   Client Id: 000000F9
Local Socket: 9.42.104.43..21        Foreign Socket: 9.42.103.165..1035
BytesIn:          0000000035          BytesOut:          0000000265
SegmentsIn:       0000000017          SegmentsOut:       0000000014
Last Touched:    21:41:20             State:             Establish
RcvNxt:          0214444666           SndNxt:           0216505563
ClientRcvNxt:    0214443596           ClientsndNxt:     0216504670
InitRcvSeqNum:   0214443560           InitSndSeqNum:    0216504404
```

Data over SMC-R link

SMC-R Monitoring – Netstat CONFIG changes

```
GLOBAL CONFIGURATION INFORMATION:
TCPIPSTATS: YES  ECSALIMIT: 2096128K  POOLLIMIT: 2096128K
MLSCHKTERM: NO  XCFGRPID: 11          IQDVLANID: 27

SYSPLEXWLMPOLK: 060  MAXRECS: 100
EXPLICITBINDPORTRANGE: 05000-06023  IQDMULTIWRITE: YES
WLMRIORITYQ: YES
  IOPRI1 0 1
  IOPRI2 2
  IOPRI3 3 4
  IOPRI4 5 6 FWD
SYSPLEX MONITOR:
  TIMERSECS: 0060  RECOVERY: YES  DELAYJOIN: NO  AUTOREJOIN: YES
  MONINTF:  YES  DYNROUTE: YES  JOIN:  YES
ZIIP:
  IPSECURITY: YES  IQDIOMULTIWRITE: YES
SMCR: YES
  FIXEDMEMORY: 200M  TCPKEEPMININT: 00000300
  PFID: 001C  PORTNUM: 1
  PFID: 0015  PORTNUM: 2
```



2 RoCE adapters defined

Netstat CONFIG/-f report includes new GLOBALCONFIG SMCR settings

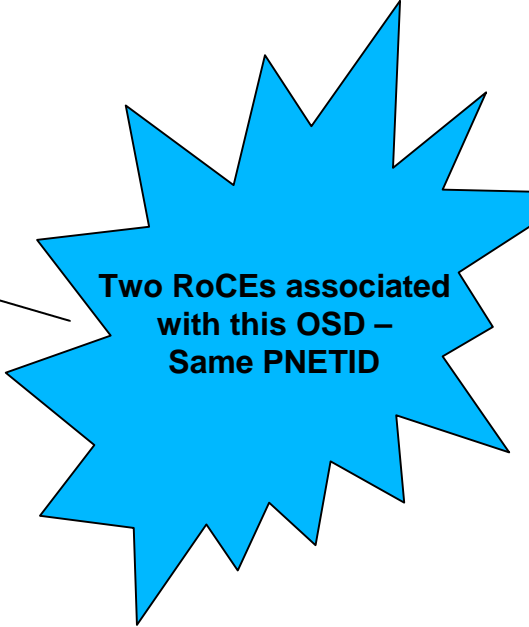
SMC-R Monitoring – Netstat DEVLINKS changes

```
D TCPIP,TCPCS1,NETSTAT,DEVLINKS,INTFNAME=OSD1
EZD0101I NETSTAT CS V2R1 TCPCS1
INTFNAME: OSD1          INTFTYPE: IPAQENET  INTFSTATUS: READY
PORTNAME: OSDP00B     DATAPATH: 0B22    DATAPATHSTATUS: READY
CHPIDTYPE: OSD        SMCR: YES
PNETID: ZOSNET
...
VLANID: 100          VLANPRIORITY: DISABLED
...
ASSOCIATED RNIC INTERFACE: EZARIUT1001C
ASSOCIATED RNIC INTERFACE: EZARIUT20015

D TCPIP,TCPCS1,NETSTAT,DEVLINKS,INTFNAME=OSXC9INT1
EZD0101I NETSTAT CS V2R1 TCPCS1
INTFNAME: OSXC9INT1    INTFTYPE: IPAQENET  INTFSTATUS: READY
PORTNAME: IUTXP0C9    DATAPATH: 0E56    DATAPATHSTATUS: READY
CHPIDTYPE: OSX        CHPID: C9
PNETID: IEDN
```



RoCEs will use same VLANID



Two RoCEs associated with this OSD – Same PNETID

Netstat DEvlinks/-d report shows associated RoCE adapters

SMC-R Monitoring – Netstat DEVLINKS, SMC changes

```
D TCPIP,TCPCS1,NETSTAT,DEVLINKS,SMC
EZD0101I NETSTAT CS V2R1 TCPCS1
INTFNAME: EZARIUT1001C      INTFTYPE: RNIC      INTFSTATUS: READY
PFID: 001C  PORTNUM: 1  TRLE: IUT1001C
PNETID: ZOSNET
VMACADDR: 02000035F740
GIDADDR:  FE80::200:FF:FE35:F740
INTERFACE STATISTICS:
  BYTESIN                = 160
  INBOUND OPERATIONS    = 5
  BYTESOUT               = 344
  OUTBOUND OPERATIONS   = 11
  SMC LINKS              = 1
  TCP CONNECTIONS       = 1
  INTF RECEIVE BUFFER INUSE = 64K
SMC LINK INFORMATION:
LOCALSMCLINKID: 2D8F0101  REMOTESMCLINKID: 729D0101
SMCLINKGROUPID: 2D8F0100  VLANID: 100  MTU: 1024
LOCALGID:  FE80::200:FF:FE35:F740
  LOCALMACADDR: 02000035F740  LOCALQP: 000040
REMOTEGID:  FE80::200:1FF:FE35:F740
  REMOTEMACADDR: 02000135F740  REMOTEQP: 000041
SMCLINKBYTESIN:          160
SMCLINKINOPERATIONS:     5
SMCLINKBYTESOUT:         344
SMCLINKOUTOPERATIONS:    11
TCP CONNECTIONS:         1
LINK RECEIVE BUFFER INUSE: 64K
  64K  BUFFER INUSE:     64K
```

One SMC link over this RoCE adapter

What size RMB is this connection using

SMC-R Monitoring – Netstat changes

```
SMC LINK GROUP INFORMATION:  
SMCLINKGROUPID: 2D8F0100 PNETID: ZOSNET  
REDUNDANCY: FULL  
LINK GROUP RECEIVE BUFFER TOTAL: 3M  
64K ← BUFFER TOTAL: 1M  
LOCALSMCLINKID REMOTESMCLINKID  
-----  
2D8F0101 729D0101  
2D8F0102 729D0102  
2 OF 2 RECORDS DISPLAYED  
END OF THE REPORT
```

Total RMB storage for this group and type in use

Two SMC-R links in this group

Full redundancy – this is the ideal setting



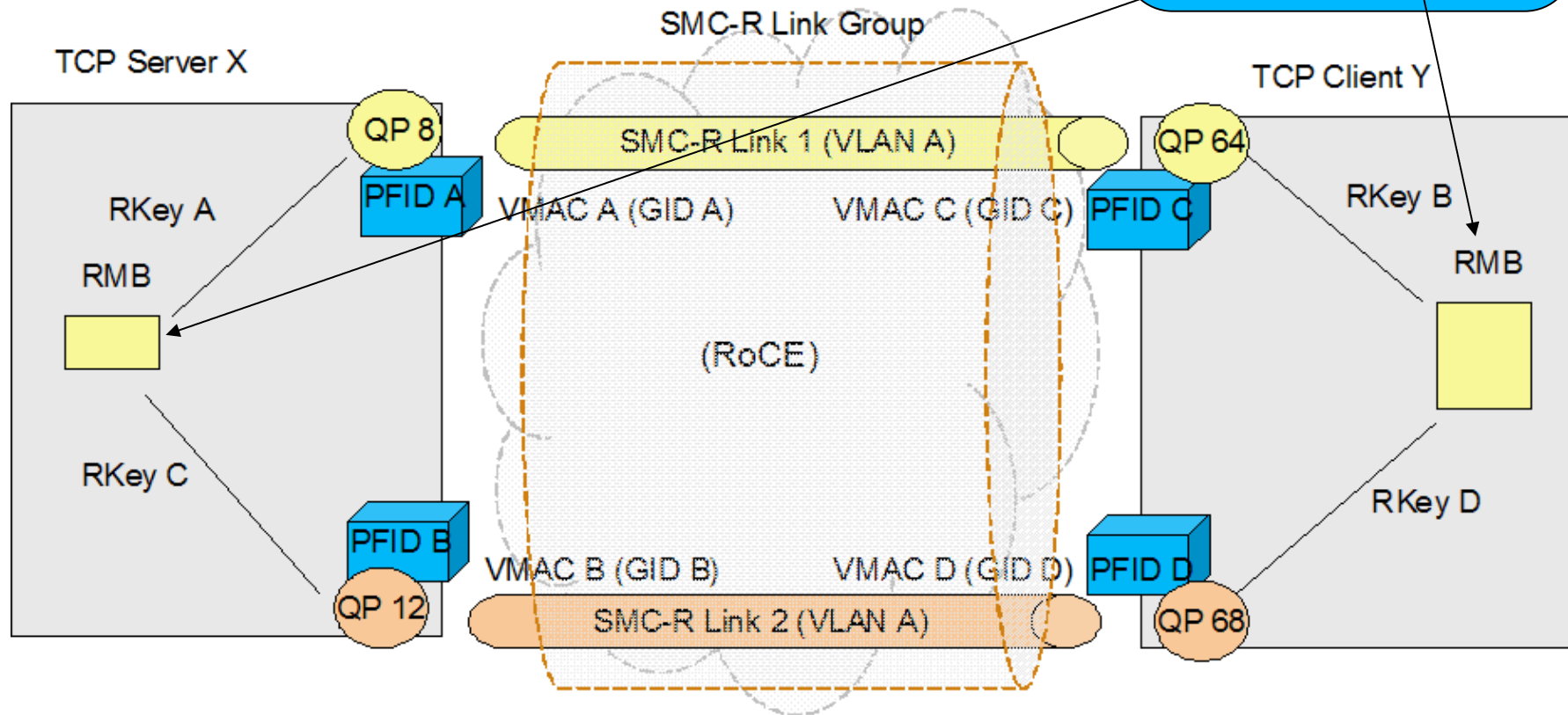
A bit more about redundancy

- SMC-R link groups provide for load balancing and recovery
 - New TCP connection is assigned to the SMC-R link with the fewest TCP connections
 - Load balancing only performed when multiple RNIC adapters are available at each peer
- Full redundancy requires:
 - Two or more RNIC adapters at each peer
 - Unique system internal paths for the RNIC adapters
 - Unique physical RoCE switches
- Partial redundancy still possible in the absence of one or more of these conditions

A bit more about redundancy

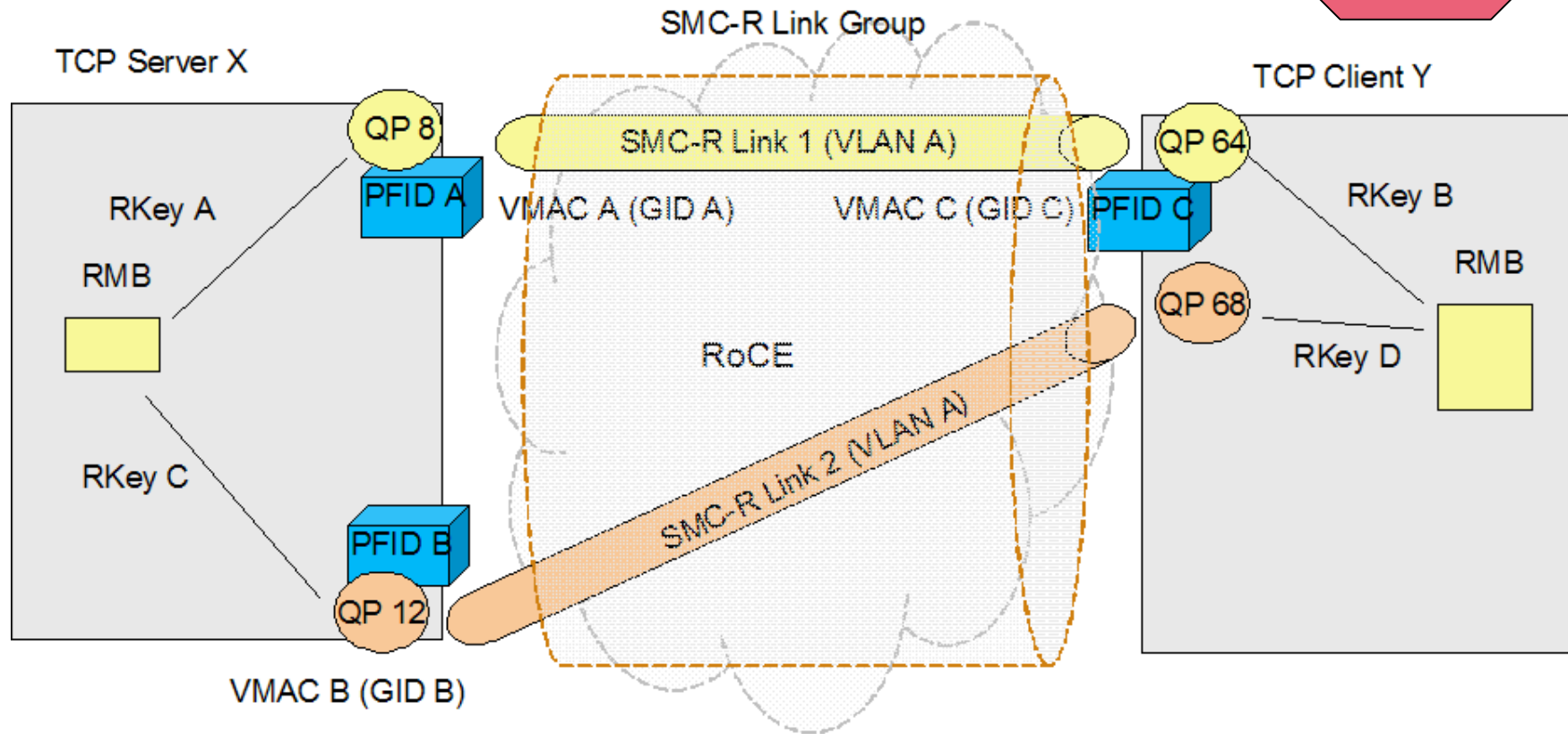
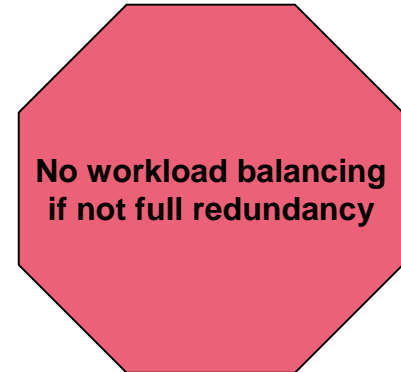
- Full failover capability exists at both server and client
 - Recommended configuration

Note the redundant links share the same RMB(s)



A bit more about redundancy

- Partial redundancy
 - Failover recovery is possible at the TCP server, but not at the TCP client



SMC-R Monitoring – Netstat STATS changes

```
D TCPIP,TCPCS1,NETSTAT,STATS,PROTOCOL=TCP
EZD0101I NETSTAT CS V2R1 TCPCS1
TCP STATISTICS
  CURRENT ESTABLISHED CONNECTIONS      = 2
  ...
  CONNECTIONS DROPPED BY KEEPALIVE     = 0
  CONNECTIONS DROPPED BY FINWAIT2     = 0
SMCR STATISTICS
  CURRENT ESTABLISHED SMC LINKS        = 2
  SMC LINK ACTIVATION TIME OUTS        = 0
  ACTIVE SMC LINKS OPENED              = 0
  PASSIVE SMC LINKS OPENED             = 2
  SMC LINKS CLOSED                     = 0
  CURRENT ESTABLISHED CONNECTIONS      = 2
  ACTIVE CONNECTIONS OPENED            = 0
  PASSIVE CONNECTIONS OPENED           = 2
  CONNECTIONS CLOSED                   = 0
  SEGMENTS RECEIVED                    = 8
  SEGMENTS SENT                        = 284
  RESETS SENT                          = 0
  RESETS RECEIVED                      = 0
END OF REPORT
```



Blue – SMC-R only stats
Red - Subset of TCP statistics

Netstat STATS/-S report shows SMC-R connection stats with PROTOCOL=TCP

SMC-R Monitoring – Display STOR command

```

D TCPIP,TCPCS1,STOR
EZZ8453I TCPIP STORAGE
EZZ8454I TCPCS1 STORAGE CURRENT MAXIMUM LIMIT
EZZ8455I ECSA 2925K 3972K NOLIMIT
EZZ8455I PRIVATE 8980K 8980K NOLIMIT
EZZ8455I ECSA MODULES 10073K 10073K NOLIMIT
EZZ8455I HVCOMMON 1M 1M NOLIMIT
EZZ8455I HVPRIVATE 2M 2M NOLIMIT
EZZ8455I TRACE HVCOMMON 2579M 2579M 2579M
EZZ8455I SMC-R FIXEDMEMORY 11M 11M 256M
EZZ8455I SMC-R SEND MEMORY 4M 4M
EZZ8455I SMC-R RECV MEMORY 3M 3M
EZZ8459I DISPLAY TCPIP STOR COMPLETED SUCCESSFULLY
    
```

New connections will “fallback” to TCP/IP if limit reached

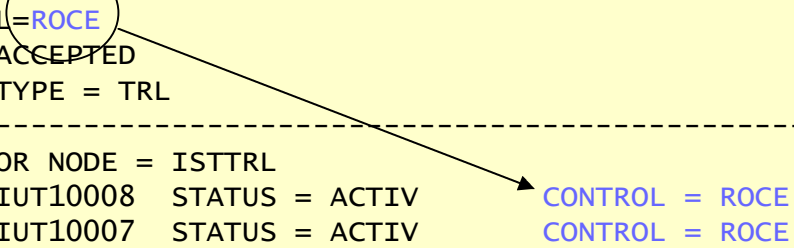
RMBs

Total Fixedmemory includes control blocks used in outbound operations

SMC-R Monitoring – VTAM Commands

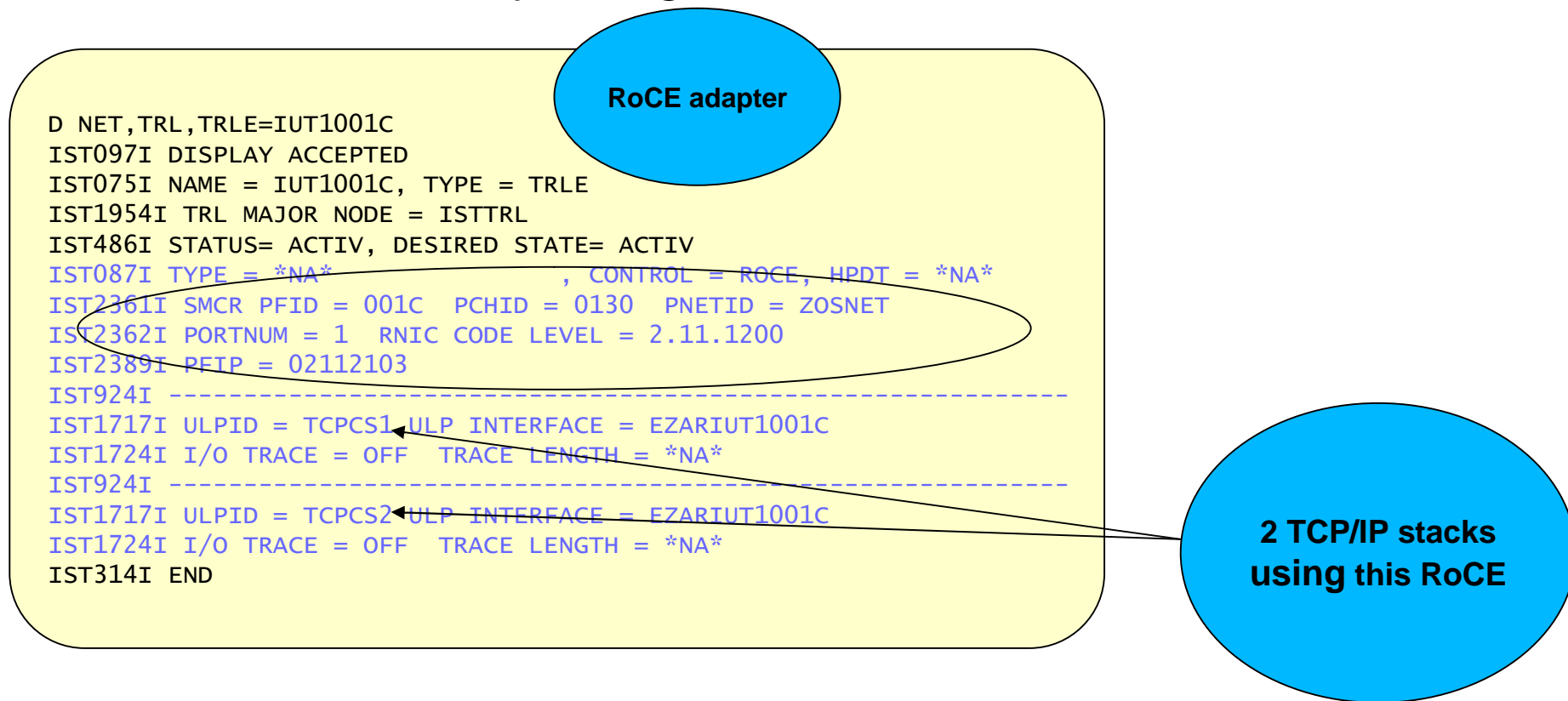
List only those TRLEs that are dynamically created to represent RNIC interfaces

```
D NET,TRL,CONTROL=ROCE
IST097I DISPLAY ACCEPTED
IST350I DISPLAY TYPE = TRL
IST924I -----
IST1954I TRL MAJOR NODE = ISTTRL
IST1314I TRLE = IUT10008 STATUS = ACTIV CONTROL = ROCE
IST1314I TRLE = IUT10007 STATUS = ACTIV CONTROL = ROCE
IST1454I 2 TRLE(S) DISPLAYED
IST924I -----
IST1954I TRL MAJOR NODE = TRLALLBV
IST1454I 0 TRLE(S) DISPLAYED
IST314I END
```



SMC-R Monitoring – VTAM Commands

- DISPLAY ID=*RNIC_trlename* generates the same output
- Provides RNIC adapter information, including which TCP/IP stacks are currently using the RNIC TRLE



SMC-R Monitoring – VTAM Commands

- Provides PNet ID value, if available
 - A value of ***NA*** is displayed if no PNet ID was configured

OSD adapter

```
D NET,TRL,TRLE=QDIO101
IST097I DISPLAY ACCEPTED
IST075I NAME = QDIO101, TYPE = TRLE
IST1954I TRL MAJOR NODE = TR LCS
IST486I STATUS= ACTIV, DESIRED STATE= ACTIV
IST087I TYPE = LEASED , CONTROL = MPC , HPDT = YES
IST1715I MPCLEVEL = QDIO MPCUSAGE = SHARE
IST2263I PORTNAME = QDIO4101 PORTNUM = 0 OSA CODE LEVEL = ABCD
IST2337I CHPID TYPE = OSD CHPID = C1 PNETID = ZOSNET
IST2184I QDIOSYNC = ALLINOUT = SYNCID = QDIO101 - SAVED = NO
IST1577I HEADER SIZE = 4096 DATA SIZE = 0 STORAGE = ***NA***
. . .
```

**OSD device –
Can link to RoCE adapters
by PNETID (new)**

SMC-R Monitoring – TNSTAT changes

- MODIFY TNSTAT,TRLE=RNIC_trlename
- RNIC-wide statistics and user-specific statistics provided

```

IST1230I TIME      = 18051835   DATE      = 09182      ID   = IUT1001C
IST1719I PCIREALO = 0          PCIREAL   = 5
IST1751I PCIUNPRO = 0          PCIUNPRD  = 3
IST2366I POLLEQO  = 0          POLLEQ    = 15
IST2367I POLLEQEO = 0          POLLEQE   = 25
IST924I -----
IST2368I ULP_ID    = TCPCS1
IST2369I POLLCQO  = 0          POLLCQ    = 250
IST2370I POLLCQUO = 0          POLLCQU   = 50
IST2371I POLLCQEO = 0          POLLCQE   = 1800
IST2372I SRBSCHDO = 0          SRBSCHD   = 15
IST2373I SRBRSCHO = 0          SRBRSCHD  = 0
IST2374I INBBYTLO = 0          INBBYTEL  = 176
IST2375I INBBYTMO = 0          INBBYTEM  = 0
IST2376I INBBYTNO = 0          INBBYTEN  = 305306
IST2377I DATAREQO = 0          DATAREQ   = 60
IST2378I POSTO    = 0          POST      = 70
IST2379I POSTEO   = 0          POSTELEM  = 178
IST2380I POSTQUEO = 0          POSTQUED  = 10
IST2381I OUTBYTLO = 0          OUTBYTEL  = 176
IST2382I OUTBYTMO = 0          OUTBYTEM  = 0
IST2383I OUTBYTNO = 0          OUTBYTEN  = 58950
    
```

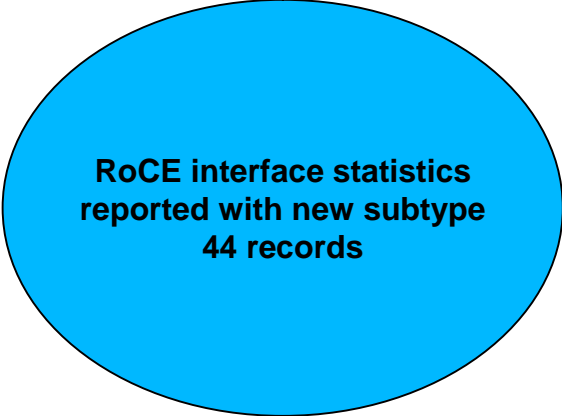
SMC-R Monitoring – Network Management Enhancements

Please refer to the appendix for more details

- Network Management Interface (NMI)
 - Updates to some existing callable NMI reports
 - Creation of two new SMC-R specific callable NMI reports
- System Management Facilities (SMF)
 - Updates to some existing SMF Type 119 records
 - Creation of four new SMC-R specific records
- Minor Simple Network Management Protocol (SNMP) changes

SMC-R Monitoring – SMF Enhancements

- **Support added to existing SMF 119 records**
 - TCP Termination (subtype 2)
 - Report SMC-R capability of the TCP connection, if applicable
 - TCPIP Profile (subtype 4)
 - Report SMC-R configuration settings
 - TCP Statistics (subtype 5)
 - Report SMC-R statistics and storage usage
 - Interface statistics (subtype 6)
 - OSD – PNet ID, SMC-R capability



**RoCE interface statistics
reported with new subtype
44 records**

SMC-R Monitoring – SMF Enhancements

- New SMF 119, subtype 44 interval record
 - Controlled by **existing** SMFCONFIG IFStatistics and NOIFStatistics parameters
- One record generated per RNIC interface
 - SMC-R link and TCP connection usage statistics
 - Storage statistics
 - PNet ID for correlation with SMC-R link groups
- Close-out record generated if recording stopped or TCP/IP stack terminates
- No close-out record if RNIC interface is stopped during interval

SMC-R Diagnosis

SMC-R Diagnosis – Traces

- Even though TCP does not create traditional packets for SMC-R data, data is formatted as packet trace data
 - Trace enabled same as for TCP/IP connections (protocol, port, IP addr..)
 - Application traffic
 - Connection Layer Control (CLC) and Link Layer Control (LLC) flows

- Full support for TCP/IP component trace (CTRACE), Data trace and VTAM Internal Trace (VIT) – No additional config necessary

SMC-R Diagnosis – CLC Packet trace

- Example of Connection Layer Control (CLC) Proposal request sent over TCP connection

```
TCP
Source Port      : 4005  ()          Destination Port: 4005  ()
Sequence Number  : 4142276139       Ack Number: 2044371877
Header Length    : 32               Flags: Ack Psh
Window Size     : 4096              CheckSum: D329 FFFF Urgent Data Pointer: 0000
Option          : NOP
Option          : NOP
Option          : Timestamp         Len: 10 Value: F1DE7250 Echo: F1DE724E
```

```
SMCR
Eyec            : SMCR              MsgType: Proposal
Length         : 52                Version: 1
Flag1          :                   PeerId: 7718D4C1C302A003
RNICaddr       : fe80::d4c1:c302:a003
Macaddr        : D4C1C3-02A003
Subnet Mask    : 172.16.1.5        Mask bits: 16
```

```
Ip Header       : 20                IP: 10.81.5.5, 10.81.3.3 Offset: 0
000000 45000068 002F0000 40065DB8 0A510505 0A510303
```

```
Protocol Header : 32                Port: 4005, 4005          Offset: 14
000000 0FA50FA5 F6E61E2B 79DAA3A5 80181000 D3290000 0101080A F1DE7250 F1DE724E
```



**Request to use
SMC-R**

SMC-R Diagnosis – Data packet trace

```

CTE:      84 178FA25E(0x25e)  212 00000008 2013/02/27 11:23:20.339971

      84 MVS030  SMC      00000008 11:23:20.339971 SMC Trace
To Interface   : EZARIUT1A003  Device: RNIC      Full=22
Tod Clock      : 2013/02/27 11:23:20.339965  Intfx: 23
Segment #     : 0              Flags: Out SMC
Source        : 10.81.3.3
Destination   : 10.81.5.5
Source Port   : 4005           Dest Port: 4005

SMC          Protocol: TCP
Payload      : 22             VlanId: 0
Local Conn Id : C3770101     Remote Conn Id: AC830101
Local Gid     : fe80::d4c1:c301:a003
Remote Gid    : fe80::d4c1:c302:a003
Local Qid     : 000001       Remote Qid: 000002
Local Conn Index : 1         Remote Conn Index: 1
Local RMB RKey : 00000101   Remote RMB RKey: 00000201
Producer Cursor : 22        Sequence Number: 0
Producer Flags  : 00 ( )
Consumer Cursor : 0         Sequence Number: 0
Connection State : 00 ( )

Data          : 22           Data Length: 22           Offset: 0
000000 6E6E6E40 A3839740 8481A381 40A39640 |>>> tcp data to nnn@...@....@...@|
000010 83938985 95A3         |client .....|
  
```

Outbound SMC-R data

**VTAM Internal Trace (VIT)
And TCP/IP Ctrace
Support as well**

Application data

Formatted to appear like TCP packets!

SMC-R Configuration & Monitoring Summary

- ✓ Consider the NOSMCR option on the PORT/ PORTRANGE statements for short-lived connections
- ✓ Consider using larger TCP RECEIVE buffer sizes for streaming/bulk connections
- ✓ At least two RNIC adapters per peer per physical network are highly recommended for reliability and load balancing
- ✓ Combination of PFID and PORTNUM values on GLOBALCONFIG SMCR statement define a given RNIC adapter
- ✓ You can specify either PORTNUM 1 or PORTNUM 2 for a given PFID, but you cannot use both
- ✓ Client and server must be in the same physical network (and VLAN)
- ✓ SMC-R enabled OSD interfaces must have non-zero subnets or prefix (IPv6)
- ✓ Use DISPLAY TCPIP,,STOR to monitor storage as workloads increase

Please fill out your session evaluation

- z/OS V2R1 CS: Shared Memory Communications - RDMA (SMC-R), Part 2
- Session # 13628
- QR Code:







Find us on Facebook at
<http://www.facebook.com/IBMCommserver>



Follow us on Twitter at
http://www.twitter.com/IBM_Commserver



Read the z/OS Communications Server blog at
<http://tinyurl.com/zoscsblog>



Visit the z/OS CS YouTube channel at
<http://www.youtube.com/user/zOSCommServer>

Appendix: SMC-R Network Management Interface and SMF enhancements

Function externals: Network Management enhancements

- Network Management Interface (NMI)
 - Updates to some existing callable NMI reports
 - Creation of two new SMC-R specific callable NMI reports
- System Management Facilities (SMF)
 - Updates to some existing SMF Type 119 records
 - Creation of four new SMC-R specific records
- Minor Simple Network Management Protocol (SNMP) changes

Function externals: Updates to callable NMI reports

- Getlfs
 - Report SMC-R capability and PNet ID for OSD interfaces
 - Use PNet ID to associate OSD with RNIC interfaces
 - Report PNet ID for OSX interfaces
 - Report minimal information about RNIC interfaces
- GetProfile
 - Report GLOBALCONFIG SMCR and SMFCONFIG settings
 - Report SMC-R information from INTERFACE, PORT and PORTRANGE statements

Function externals: Updates to callable NMI reports, part 2

- GetConnectionDetail (when SMC-R is enabled)
 - Report local SMC-R link group ID, and remote and local SMC-R link IDs if TCP connection is using SMC-R
 - Report reason code if TCP connection is not using SMC-R
- GetGlobalStats
 - Report SMC-R specific statistics and TCP statistics which might include SMC-R related statistics
- GetStorageStatistics
 - Report SMC-R storage usage
- No changes to GetIfStats and GetIfStatsExtended

Function externals: New GetRnics callable NMI

- Provides combination of Getlfs, GetlfStats and GetlfStatsExtended for RNIC interfaces
 - One record per RNIC interface
 - Same RNIC information as reported in Getlfs for RNIC interface
 - Provided to correlate this record with Getlfs information
 - PNet ID can be used to correlate this record with GetSmcLinks information
 - RNIC stack statistics (GetlfStats) always provided, even if RNIC interface is not active
 - VTAM tuning statistics (GetlfStatsExtended) only provided for active RNIC interface
- No filters supported on this NMI

Function externals: New GetSmcLinks callable NMI

- Provides SMC-R link and link group information
 - One record per SMC-R link group
 - One section for SMC-R link group statistics
 - One or more sections of SMC-R link statistics
 - *One section for each SMC-R link that is part of the link group*
 - PNet ID associated with the SMC-R link group can be used to correlate the group with RNIC interfaces
- No filters supported on this NMI

Function externals: Updates to existing SMF records

- TCP Termination (subtype 2)
 - Report SMC-R capability of the TCP connection, if applicable
 - If using SMC-R, remote and local SMC-R link ID and local SMC-R link group ID
 - If not using SMC-R, reason code
- TCPIP Profile (subtype 4)
 - Report GLOBALCONFIG SMCR and SMFCONFIG settings
 - Report INTERFACE SMCR settings
 - Report PORT and PORTRANGE NOSMCR settings

Function externals: Updates to existing SMF records, part 2

- TCP Statistics (subtype 5)
 - Report SMC-R specific statistics
 - Report TCP statistics which might include SMC-R related statistics
 - Report SMC-R storage usage
- Interface statistics (subtype 6)
 - Report PNet ID for OSX interfaces
 - Report SMC-R capability and PNet ID for OSD interfaces
 - Use PNet ID to associate OSD with RNIC interfaces
 - **RNIC interfaces are reported using new subtype 44 records**

Function externals: SMFCONFIG updates

- Two new options for controlling new SMC-R specific settings
 - SMFCONFIG TYPE119 IFStatistics controls subtype 44 now as well

```

>>-SMFCONFIG-----+-----+----->
                    '-| Type 118 options |-'

-----
v                                     |
>-----+-----+-----+-----<<
      +-TYPE118--| Type 118 Options |+-+
      '-TYPE119--| Type 119 options |-'

Type 119 Options
-----
v                                     |
|-----+-----+-----+-----|
|  -NOSMCRGROUPStatistics-  |
| +-+-----+-----+-----+ |
|  '-SMCRGROUPStatistics---' |
|  -NOSMCRLINKEvent-----  |
| +-+-----+-----+-----+ |
|  '-SMCRLINKEvent-----'  |

```

Function externals: Netstat CONFIG/-f report, SMFCONFIG

- Netstat CONFIG/-f report includes new SMFCONFIG settings

```
SMF PARAMETERS:  
TYPE 118:  
  TCPINIT:      00   TCPTERM:      02   FTPCLIENT:    03  
  TN3270CLIENT: 04   TCPIPSTATS:   05  
TYPE 119:  
  TCPINIT:      YES  TCPTERM:      YES  FTPCLIENT:    YES  
  TCPIPSTATS:   YES  IFSTATS:     YES  PORTSTATS:    YES  
  STACK:        YES  UDPTERM:     YES  TN3270CLIENT: YES  
  IPSECURITY:   NO   PROFILE:     YES  DVIPA:        YES  
  SMCGRPSTATS: YES  SMCRLNKEVENT: YES
```

Function externals: New SMCR Link Group Statistics record

- New SMF 119, subtype 41 interval record
 - Controlled by SMFCONFIG SMCRGROUPStatistics and NOSMCRGROUPStatistics parameters
- One record generated for all SMC-R link groups
 - One section for each active SMC-R link group
 - Includes RMB usage statistics
 - One section for each active SMC-R link
 - SMC-R link section includes SMC-R link group ID for correlation
- Close-out record generated if recording stopped or TCP/IP stack terminates
- No close-out record if SMC-R link group terminates during interval

Function externals: New SMCR Link State Start record

- New SMF 119, subtype 42 event record
 - Controlled by SMFCONFIG SMCRLINKEvent and NOSMCRLINKEvent parameters
- One record generated when SMC-R link starts
 - Provides minimal information about the link
 - SMC-R link and link group ID values
 - Link identification (7-tuple) information

Function externals: New SMCR Link State End record

- New SMF 119, subtype 43 event record
 - Controlled by SMFCONFIG SMCRLINKEvent and NOSMCRLINKEvent parameters
- One record generated when SMC-R link terminates
 - Provides same information as SMC-R Link State Start record
 - Provides statistical information related to SMC-R link
 - Storage statistics
 - TCP connection usage statistics

Function externals: New RNIC Interface Statistics record

- New SMF 119, subtype 44 interval record
 - Controlled by SMFCONFIG IFStatistics and NOIFStatistics parameters
- One record generated per RNIC interface
 - SMC-R link and TCP connection usage statistics
 - Storage statistics
 - PNet ID for correlation with SMC-R link groups
- Close-out record generated if recording stopped or TCP/IP stack terminates
- No close-out record if RNIC interface is stopped during interval

Function externals: SNMP updates

- Provide configured SMCR value for OSD interfaces
- Provide PNet ID information for OSD and OSX interfaces
 - Provided regardless of whether SMC-R is enabled or not
- Provide minimal information for RNIC interfaces
 - PNet ID value
 - Associated TRLE name
- Provide information about ports that are restricted, at the server, from using SMC-R