Trademarks

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

AIX™
BladeCenter™
CICS™
Cognos™
DataPower™
DB2™
DFSMS
EASY Tier
IMS
FICON™
GDPS™
HiperSockets™
HyperSwap
InfinitiBand™
Lotus™
MQSeries™
NetView™
OMEGAMON™
Parallel Sysplex™
POWER7™
PowerHA™
PR/SM
PureSystems
RACF™
RMF
Smarter Planet™
Storwize™
System Storage™
System z™
zEnterprise™
z/VM™
z/VSE™
z10
z10 EC
Tivoli™
WebSphere™
x/OS™
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. zILPs, 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 A - SMC-R Network Management Interface and SMF enhancements
  - Appendix B - SMC-R dedicated

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 References


- SMC-R Overview
  - Overview with audio (youtube)
- SMC-R Implementation
  - With audio (youtube)
- Shared Memory Communications over RDMA: Performance Considerations (White Paper)
- Performance information
- FAQ
- Diagnosing Problems with SMC-R
- RFC
- SMC-R and Security Considerations White Paper
- SMC-R over distance
- SMC-R VLAN Configuration Considerations
- SMC-R Linux Overview
Shared Memory Communications – Remote (SMC-R)

SMC-R Background

Both TCP and SMC-R “connections” remain active

TCP connection establishment over IP
TCP conn flows (with TCP Options indicating SMC-R capability) + SMC setup flow (rendezvous)

Dynamic (in-line) negotiation for SMC-R is initiated by presence of TCP Options

TCP connection transitions to SMC-R allowing application data to be exchanged using RDMA
RDMA

Key attributes of RDMA

- Enables a host to read or write directly from/to a remote host’s memory
  without involving the remote host’s CPU
  - By registering specific memory for RDMA partner use
  - Avoids TCP “packetizing” and ACK processing
  - Interrupts still required for notification (i.e. CPU cycles are not completely eliminated)

- Reduced networking stack overhead by using streamlined, low level, RMDA interfaces

Key requirements:

- A reliable “lossless” network fabric (LAN for layer 2 data center network distance)
- An RDMA capable NIC (RNIC) and RDMA capable switched fabric (switches)
Shared Memory Communications over RDMA (SMC-R) is a protocol that allows *TCP sockets* applications to transparently exploit RDMA (RoCE).

SMC-R is a “hybrid” solution that:

- Uses TCP connection (3-way handshake) to establish SMC-R connection.
- Each TCP end point exchanges TCP options that indicate whether it supports the SMC-R protocol.
- SMC-R “rendezvous” (RDMA attributes) information is then exchanged within the TCP data stream (similar to SSL handshake).
- Socket application data is exchanged via RDMA (write operations).
- TCP connection remains active (controls SMC-R connection).
- This model preserves many critical existing operational and network management features of TCP/IP.
RoCE Express features can be shared across LPARs
  - Up to 31 operating system instances can share one feature
- Both RoCE Express ports can be used simultaneously
- No additional RNIC definitions at z/OS Comm Server
  - PFID values are still defined on TCP/IP GLOBALCONFIG statement
  - PFID value must be unique if the RoCE Express feature is being shared by multiple TCP/IP stacks
- No change in RNIC activation
  - RoCE Express features are still activated when the first SMC-R capable OSA interface is activated
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/IOCDS):
    - 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
  - Virtual Function ID (VF)
  - Configure Ethernet switches appropriately
    - Optionally define VLAN ID values to be used
    - Enable “flow control” capability
### SMC-R HCD Configuration

**HCD Main Panel:**
1. Select (Processor list)
2. Then select (/) processor

#### Actions on selected processors

Select by number or action code and press Enter.

<table>
<thead>
<tr>
<th>Number</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Add like... [a]</td>
</tr>
<tr>
<td>2</td>
<td>Repeat (Copy) processor configurations [r]</td>
</tr>
<tr>
<td>3</td>
<td>Change... [c]</td>
</tr>
<tr>
<td>4</td>
<td>*Prime serial number... [i]</td>
</tr>
<tr>
<td>5</td>
<td>Delete... [d]</td>
</tr>
<tr>
<td>6</td>
<td>View processor definition... [v]</td>
</tr>
<tr>
<td>7</td>
<td>View related CTC connections... [k]</td>
</tr>
<tr>
<td>8</td>
<td>Work with PCIe functions... [f]</td>
</tr>
<tr>
<td>9</td>
<td>Work with partitions... (SMP) [p]</td>
</tr>
<tr>
<td>10</td>
<td>Work with attached channel paths (SMP) [s]</td>
</tr>
<tr>
<td>11</td>
<td>Work with attached devices... (SMP) [u]</td>
</tr>
<tr>
<td>12</td>
<td>Copy to channel subsystem... (SMP) [y]</td>
</tr>
<tr>
<td>13</td>
<td>Work with channel subsystems... (XMP) [p,s]</td>
</tr>
</tbody>
</table>

* = requires TSA I/O Operations

F1=Help  F2=Split  F3=Exit  F9=Swap  F12=Cancel

F1=Help  F20=Right  F22=Command
SMC-R HCD Configuration

Select one or more PCIe functions, then press Enter. To add, use F11.

<table>
<thead>
<tr>
<th>FID</th>
<th>PCHID</th>
<th>VF+</th>
<th>Type+</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>001</td>
<td>108</td>
<td>1</td>
<td>ROCE</td>
<td>S36</td>
</tr>
<tr>
<td>002</td>
<td>108</td>
<td>2</td>
<td>ROCE</td>
<td>S36</td>
</tr>
<tr>
<td>003</td>
<td>108</td>
<td>3</td>
<td>ROCE</td>
<td>S36</td>
</tr>
<tr>
<td>004</td>
<td>108</td>
<td>4</td>
<td>ROCE</td>
<td>S36</td>
</tr>
<tr>
<td>005</td>
<td>108</td>
<td>5</td>
<td>ROCE</td>
<td>S36</td>
</tr>
<tr>
<td>006</td>
<td>108</td>
<td>6</td>
<td>ROCE</td>
<td>S37</td>
</tr>
<tr>
<td>007</td>
<td>108</td>
<td>7</td>
<td>ROCE</td>
<td>S37</td>
</tr>
<tr>
<td>008</td>
<td>108</td>
<td>8</td>
<td>ROCE</td>
<td>S37</td>
</tr>
<tr>
<td>009</td>
<td>108</td>
<td>9</td>
<td>ROCE</td>
<td>S37</td>
</tr>
</tbody>
</table>

9 RoCE Adapters defined
FIDs used by TCPIP config

RoCE Express
Virtualization
SMC-R HCD Configuration – Define a PFID

**PCle Function List**

<table>
<thead>
<tr>
<th>Command ===</th>
<th>Scroll ===</th>
</tr>
</thead>
</table>

**Add PCle Function**

Specify or revise the following values.

<table>
<thead>
<tr>
<th>Processor ID</th>
<th>S88</th>
<th>zSphinx S88</th>
</tr>
</thead>
<tbody>
<tr>
<td>Function ID</td>
<td>010</td>
<td></td>
</tr>
<tr>
<td>Type</td>
<td>ROCE</td>
<td>+</td>
</tr>
<tr>
<td>PCHID</td>
<td>108</td>
<td></td>
</tr>
<tr>
<td>Virtual Function ID</td>
<td>10</td>
<td>+</td>
</tr>
<tr>
<td>Description</td>
<td>Roce FID 10</td>
<td></td>
</tr>
</tbody>
</table>

F1=Help  F2=Split  F3=Exit  F4=Prompt  F5=Reset  F9=Swap  F12=Cancel  
F20=Right  F22=Command

Function ID will match TCPIP GLOBALCONFIG value.
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 the same PNETID.
SMC-R IOCDS Configuration example

sample RoCE IOCP / IOCDS declaration:

FUNCTION FID=100,PCHID=400,VF=1,PART=((LPA),(LPB,LPC)),PNETID=(NET1,NET2),TYPE=ROCE
FUNCTION FID=101,PCHID=400,VF=2,PART=((LPB),(LPA,LPC)),PNETID=(NET1,NET2),TYPE=ROCE
FUNCTION FID=102,PCHID=400,VF=3,PART=((LPC),(LPA,LPB)),PNETID=(NET1,NET2),TYPE=ROCE
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

Must include at least
One – 16 max
Port values = 1 or 2

EZARIUTxyyyy for RNIC interface, IUTxyyyy for TRLE, where x = PORTNUM and yyy = PFID
The PFID value on the TCPIP GLOBALCONFIG SMCR statement has a slightly different meaning in a shared environment:

- In a dedicated environment, the PFID directly identifies the RoCE Express feature, and all TCP/IP stacks sharing the feature use the same PFID.

- In a shared environment, each TCP/IP stack has its own unique PFID value to represent the RoCE Express feature.

- You define in HCD which RoCE Express feature is represented by a given PFID value and virtual function (VF) number.

The RoCE feature can support sending data in three different MTU sizes: 1024, 2048 and 4096.
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) = 28M
  TOTAL = 74M

Estimation only! Use Display TCPIP,,STOR to determine the right amount for you
SMC-R Autonomics configuration

New for V2R2!
SMC-R TCP/IP Configuration – SMCGLOBAL AUTOCACHE

- Cache IP destinations with persistent SMC-R establishment failures
  - Cached after three consecutive failures to use SMC-R
  - While cached, connections will use TCP
    - Avoids unnecessary overhead (rendezvous) trying to use SMC-R
  - Cache cleared every interval (20 minutes) or by disabling AUTOCACHE function

New for V2R2!
SMC-R TCP/IP Configuration – SMCGLOBAL AUTOSMC

- Enables the stack to monitor incoming TCP connections to dynamically determine whether SMC-R is beneficial for a local TCP server application (Well-Suited!!! – Remember this)
  - Identifies short-lived connections exchanging little data
  - Results of this monitoring influences whether TCP connections to a particular server (port) use SMC-R

New for V2R2!
New PORT/PORTRANGE SMC configuration option added
- NOSMC was added in z/OS V2R1 (NOSMCR)
- PORT/PORTRANGE configuration overrides AUTOSMC monitoring

Example:
- User configures SMC option on PORT statement for port 50
- User also codes SMCGlobal AUTOSMC (default)
- AUTOSMC monitoring suggests TCP is a better choice for workload to port 50
- However, the PORT SMC option overrides AUTOSMC and connections to port 50 will use SMC-R

New for V2R2!
SMC-R TCP/IP Configuration – Server control

- New PORT/PORTRANGE SMC configuration option added
- NOSMCR changed to NOSMC (both will work)

New for V2R2!

PORT or PORTRANGE

Options:
- DELAYAcks
- NOAUTolog
- NODELAYAcks
- SHAREPort
- BIND ipaddr
- SAF resname
- SHAREPORTWLM

New PORT/PORTRANGE SMC configuration option added
NOSMCR changed to NOSMC (both will work)
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

OSD interface definition:

```
| __PORTNAME portname__IPADDR__ ipv4_address/0_____________ __|___________|__ ___________ __|> |
| ipv4_address_______________|  |_PRIRouter_  |_VLANID id_  |
| ipv4_address/num_mask_bits_|
| __SECRouter_ |
|_INBPERF BALANCED__________________________ |> __|___________________________________________|__ _________________________________ _______>
|                       _NOWORKLOADQ_       |  |                     _ROUTEALL_  |
|_INBPERF__ ___DYNAMIC_|_____________|____ _|_VMAC__ _________ __|__________|_
|           |_WORKLOADQ___|    |            |_macaddr_  |_ROUTEELCL |
|_MINCPU_______________________| |_MINLATENCY___________________|
|__SMCR___

Must be non-zero subnet
For SMC-R

Optional for SMC-R – RoCE inherits VLANID from associated OSD
SMC-R TCP/IP Configuration – SYNTAXCHECK

- Can be used to validate SMCR profile changes before use
- Run against profile data set

```
11.53.38  v tcpip,tcpcs3,syntaxcheck,dsn=user.tcpparms(tcpcs3)
11.53.38  EZZ0060I PROCESSING COMMAND: VARY TCPIP,TCPCS3,SYNTAXCHECK,
           DSN=USER.TCPPARMS(TCPCS3)
11.53.38  EZZ0061I VARY SYNTAXCHECK COMMAND BEGINNING
11.53.38  EZZ0300I OPENED SYNTAXCHECK FILE 'USER.TCPPARMS(TCPCS3)'
11.53.38  EZZ0309I PROFILE PROCESSING BEGINNING FOR
           'USER.TCPPARMS(TCPCS3)'
11.53.38  EZZ0316I PROFILE PROCESSING COMPLETE FOR FILE
           'USER.TCPPARMS(TCPCS3)'
11.53.38  EZZ0062I VARY SYNTAXCHECK FOUND NO ERRORS
11.53.38  EZZ0065I VARY SYNTAXCHECK COMMAND COMPLETE

11.54.32  v tcpip,tcpcs3,syntaxcheck,dsn=user.tcpparms(tcpcs3)
11.54.32  EZZ0060I PROCESSING COMMAND: VARY TCPIP,TCPCS3,SYNTAXCHECK,
           DSN=USER.TCPPARMS(TCPCS3)
11.54.32  EZZ0061I VARY SYNTAXCHECK COMMAND BEGINNING
11.54.32  EZZ0300I OPENED SYNTAXCHECK FILE 'USER.TCPPARMS(TCPCS3)'
11.54.32  EZZ0309I PROFILE PROCESSING BEGINNING FOR
           'USER.TCPPARMS(TCPCS3)'
11.54.32  EZZ0316I PROFILE PROCESSING COMPLETE FOR FILE
           'USER.TCPPARMS(TCPCS3)'
11.54.32  EZZ0064I VARY SYNTAXCHECK FOUND ERRORS: SEE PREVIOUS MESSAGES
11.54.32  EZZ0065I VARY SYNTAXCHECK COMMAND COMPLETE
```
SMC-R Monitoring
SMC-R Monitoring – PCIe UPDATE

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

Two RoCEs ready for use on this LPAR

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

Two RoCEs ready for use on this LPAR
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 ALL changes

- Netstat ALL/-A report provides SMC-R information about TCP servers and connections when SMC-R is enabled
  - For servers (STATE: LISTEN) display provides SMC-R statistics
  - For SMC-R connections display 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

- View current server details
  - 90% of monitored connections over last interval had ideal workload for SMC-R
  - AutoSMC% must be >= 50% for UseSMC to be YES

<table>
<thead>
<tr>
<th>CLIENT NAME: USER19</th>
<th>CLIENT ID: 00000052</th>
</tr>
</thead>
<tbody>
<tr>
<td>LOCAL SOCKET: 0.0.0.0..4206</td>
<td></td>
</tr>
<tr>
<td>FOREIGN SOCKET: 0.0.0.0..0</td>
<td></td>
</tr>
<tr>
<td>BYTESIN:</td>
<td>00000000000000000000</td>
</tr>
<tr>
<td>BYTESOUT:</td>
<td>00000000000000000000</td>
</tr>
<tr>
<td>SEGMENTSIN:</td>
<td>00000000000000000000</td>
</tr>
<tr>
<td>SEGMENTSOUT:</td>
<td>00000000000000000000</td>
</tr>
<tr>
<td>STARTDATE:</td>
<td>01/30/2015</td>
</tr>
<tr>
<td>LAST TOUCHED:</td>
<td>19:02:05</td>
</tr>
<tr>
<td>STATE:</td>
<td>LISTEN</td>
</tr>
<tr>
<td>CONNECTIONSIN:</td>
<td>00000000200</td>
</tr>
<tr>
<td>CONNECTIONSDROPPED:</td>
<td>00000000000</td>
</tr>
<tr>
<td>MAXIMUMBACKLOG:</td>
<td>000000000010</td>
</tr>
<tr>
<td>CONNECTIONFLOOD:</td>
<td>NO</td>
</tr>
<tr>
<td>CURRENTBACKLOG:</td>
<td>0000000000</td>
</tr>
<tr>
<td>SERVERBACKLOG:</td>
<td>00000000000</td>
</tr>
<tr>
<td>FRCABACKLOG:</td>
<td>00000000000</td>
</tr>
<tr>
<td>CURRENTCONNECTIONS:</td>
<td>00000000050</td>
</tr>
<tr>
<td>SEF:</td>
<td>100</td>
</tr>
<tr>
<td>QUIESCED:</td>
<td>NO</td>
</tr>
<tr>
<td>SMC INFORMATION:</td>
<td></td>
</tr>
<tr>
<td>SMCRCURRCONNS:</td>
<td>000000000025</td>
</tr>
<tr>
<td>SMCRTOTALCONNS:</td>
<td>000000000100</td>
</tr>
<tr>
<td>UseSMC:</td>
<td>Yes</td>
</tr>
<tr>
<td>Source:</td>
<td>AutoSMC</td>
</tr>
<tr>
<td>AutoSMC%:</td>
<td>090</td>
</tr>
</tbody>
</table>
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
  LOCALSMCRCVBUF:   64K            REMOTESMCRCVBUF:  64K

---- 1 OF 1 RECORDS DISPLAYED
END OF THE REPORT

• SMCID filter – Show only SMC-R connections with xxxxxxxxx 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

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 – This destination addr has been cached (*)

New for V2R2!

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

<table>
<thead>
<tr>
<th>MVS</th>
<th>TCP/IP</th>
<th>NETSTAT CS V2R1</th>
<th>TCPIP Name: TCPCS</th>
<th>21:42:39</th>
</tr>
</thead>
<tbody>
<tr>
<td>Client Name: FTPD1</td>
<td>BytesIn: 0000000035</td>
<td>BytesOut: 0000000265</td>
<td>0000000014</td>
<td></td>
</tr>
<tr>
<td>Local Socket: 9.42.104.43..21</td>
<td>SegmentsIn: 0000000017</td>
<td>SegmentsOut: 0000000014</td>
<td>0216505563</td>
<td></td>
</tr>
<tr>
<td>Foreign Socket: 9.42.103.165..1035</td>
<td>Last Touched: 21:41:20</td>
<td>State: Establsh</td>
<td>0216504670</td>
<td></td>
</tr>
<tr>
<td>RcvNxt: 0214444666</td>
<td>SndNxt: 0214443596</td>
<td>ClientSndNxt: 0216504404</td>
<td>0216504404</td>
<td></td>
</tr>
<tr>
<td>ClientRcvNxt: 0214443596</td>
<td>InitSndSeqNum: 0214444666</td>
<td>InitSndSeqNum: 0214443596</td>
<td>0216504404</td>
<td></td>
</tr>
</tbody>
</table>

**Data over SMC-R link**
SMC-R Monitoring – Netstat CONFIG changes

GLOBAL CONFIGURATION INFORMATION:
TCP/IPSTATS: YES  ECSALIMIT: 2096128K  POOLLIMIT: 2096128K
MLSCHKTERM: NO  XCFGRPID: 11  IQDVLANID: 27

SYSPLEXWLMPOLK: 060  MAXRECS: 100
EXPLICITBINDPORTRANGE: 05000-06023  IQDMULTIWRITE: YES
WLMPRIORITYQ: 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
SMCGLOBAL:
  AUTOCACHE: YES  AUTOSMC: YES
SMCR: YES
  FIXEDMEMORY: 200M  TCPKEEPMININT: 00000300
  PFID: 001C  PORTNUM: 1  MTU: 1024
  PFID: 0015  PORTNUM: 2  MTU: 4096

New for V2R2!

Netstat CONFIG/-f report includes new GLOBALCONFIG SMCGLOBAL
and 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

Two RoCEs associated with this OSD – Same PNETID
RoCEs will use same VLANID

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
  SMLINKGROUPID: 2D8F0100 VLANID: 100 MTU: 1024
  LOCALGID: FE80::200:FF:FE35:F740
    LOCALMACADDR: 02000035F740 LOCALQP: 000040
  REMOTEGID: FE80::200:FF:FE35:F740
    REMOTEMACADDR: 02000135F740 REMOTEQP: 000041
  SMLINKBYTESIN: 160
  SMLINKINOPERATIONS: 5
  SMLINKBYTESOUT: 344
  SMLINKOUTOPERATIONS: 11
  TCP CONNECTIONS: 1
  LINK RECEIVE BUFFER INUSE: 64K
    64K BUFFER INUSE: 64K
SMC-R Monitoring – Netstat changes

<table>
<thead>
<tr>
<th>SMCLINKGROUPID</th>
<th>PNETID</th>
<th>REDUNDANCY</th>
<th>LINK GROUP RECEIVE BUFFER</th>
<th>LOCALSMCLINKID</th>
<th>REMOTESMCLINKID</th>
</tr>
</thead>
<tbody>
<tr>
<td>2D8F0100</td>
<td>ZOSNET</td>
<td>FULL</td>
<td>TOTAL: 3M 64K BUFFER TOTAL: 1M</td>
<td>2D8F0101</td>
<td>729D0101</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>2D8F0102</td>
<td>729D0102</td>
</tr>
</tbody>
</table>

Two SMC-R links in this group

Full redundancy – this is the ideal setting

Total RMB storage for this group and type in use

END OF THE REPORT
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 RoCE Express adapters at each peer
  - Follow installation guidelines to assure each (PCIe) adapter has a unique system internal path:
    - RoCE Express adapter I/O drawer plugging / provisioning: Each LPAR is provisioned 2 adapters where:
      * 1st card is plugged into the left side and
      * 2nd card is plugged into the right side of the I/O drawer
        (using hardware configuration tools: eConfig, HCD and CMT)
    - 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

No workload balancing if not full redundancy
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
SMC R 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

Netstat STATS/-S report shows SMC-R connection stats with PROTOCOL=TCP
### SMC-R Monitoring – Display STOR command

The command `D TCPIP,TCPCS1,STOR` is used to display the storage information for different components. The output includes details for ECSA, private, modules, HVCOMMON, and HVPRIVATE, among others.

**Current, Maximum, Limit (in bytes):**
- **ECSA:** 2925K, 3972K, NO LIMIT
- **PRIVATE:** 8980K, 8980K, NO LIMIT
- **MODULES:** 10073K, 10073K, NO LIMIT
- **HVCOMMON:** 1M, 1M, NO LIMIT
- **HVPRIVATE:** 2M, 2M, NO LIMIT
- **TRACE HVCOMMON:** 2579M, 2579M, 2579M
- **SMC-R FIXEDMEMORY:** 11M, 11M, 256M
- **SMC-R SEND MEMORY:** 4M, 4M
- **SMC-R RECV MEMORY:** 3M, 3M

**Status:**
- `EZZ8459I DISPLAY TCPIP STOR COMPLETED SUCCESSFULLY`

**Total Fixedmemory includes control blocks used in outbound operations**

**New connections will “fallback” to TCP/IP if limit reached**
SMC-R Monitoring – VTAM commands
SMC-R Monitoring – VTAM Commands

• List only those TRLEs that are dynamically created to represent RNIC interfaces
SMC-R Monitoring – VTAM Commands

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

```
D NET,TRL,TRLE=IUT1001C
IST097I DISPLAY ACCEPTED
IST075I NAME = IUT1001C, TYPE = TRLE
IST1954I TRL MAJOR NODE = ISTTRL
IST486I STATUS= ACTIV, DESIRED STATE= ACTIV
IST087I TYPE = *NA*, CONTROL = ROCE, HPDT = *NA*
IST2361I SMCR PFID = 001C PCHID = 0130 PNETID = ZOSNET
IST2362I PORTNUM = 1 RNIC CODE LEVEL = ***NA***
IST2389I PFIP = 02112103
IST2417I VFN = 001
IST924I -------------------------------
IST1717I ULPID = TCPCS1 ULP INTERFACE = EZARIUT1001C
IST1724I I/O TRACE = OFF TRACE LENGTH = *NA*
IST314I END
```
SMC-R Monitoring – VTAM Commands

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

D NET,TRL,TRL=QDIO101
IST097I DISPLAY ACCEPTED
IST075I NAME = QDIO101, TYPE = TRLE
IST1954I TRL MAJOR NODE = TRLCS
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
IST2331I 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 – RoCE statistics

- *DISPLAY TRL, TRLE, DEVSTATS* output
- Statistics represent adapter-per-stack’s 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
IST2398I LINK OUTAGE NOTIFICATIONS 0 2
IST314I END
```
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 POLLEQE= 0           POLLEQE =  25
IST924I
IST2368I ULP_ID = TCPCS1
IST2369I POLLCQO = 0          POLLCQ =  250
IST2370I POLLCQUO = 0         POLLCQU =  50
IST2371I POLLCQE = 0          POLLCQE = 1800
IST2372I SRBSCHDO = 0         SRBSCHD =  15
IST2373I SRBRSCHO = 0         SRBRSCHD =  0
IST2374I INBBYTELO = 0        INBBYTEL =  176
IST2375I INBBYMTO = 0         INBBYTEM =  0
IST2376I INBBYTN0 = 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 OUTBYTM0 = 0         OUTBYTEM =  0
IST2383I OUTBYTN0 = 0         OUTBYTEN = 58950
```

Stats for TCP/IP stack – TCPCS1

Bytes received over SMC-R

Request from TCP/IP to send data
SMC-R Monitoring – 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

Please refer to the appendix for more details
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

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

Subtype 41 – Reports info about SMC-R links and link Groups (ie, RMB usage)

Subtype 42/43 – Reports SMC-R link start (subtype 42) and Link stop (subtype 43)
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 (**rendezvous**)
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 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: 0000201
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 ......

Formatted to appear like TCP packets!

Outbound SMC-R data
VTAM Internal Trace (VIT) And TCP/IP Ctrace Support as well
Application data
SMC-R Configuration & Monitoring Summary

✓ 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
✓ 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
✓ If you currently have multiple TCP/IP stacks sharing a RoCE Express feature in a dedicated RoCE environment, you must:
  ✓ Define unique FID values in HCD for the stacks to use as PFIDs on the TCPIP GLOBALCONFIG SMCR statement
  ✓ Assign VF values in HCD for each FID
SMC Applicability Tool
Evaluating SMC-R applicability and benefits

As customers express interest in SMC-R and RoCE Express one of the initial questions asked is:

- “What benefit will SMC-R provide in my environment?”
  - Some users are well aware of significant traffic patterns that can benefit from SMC-R
  - But others are unsure of how much of their TCP traffic (in their environment) is:
    - z/OS to z/OS and
    - how much of that traffic is well suited to SMC-R

- Reviewing SMF records, using Netstat displays, Ctrace analysis and reports from various Network Management products can provide these insights...

This approach can be a time consuming activity that requires significant expertise.
SMC Applicability Tool Introduction

A new tool called SMC Applicability Tool (SMCAT) has been created that will help customers determine the potential value of SMC-R in their environment with minimal effort and minimal impact.

- SMCAT is integrated within the TCP/IP stack: Gather new statistics that are used to project SMC-R applicability and benefits for the current system
  - Minimal system overhead, no changes in TCP/IP network flows
  - Produces reports on potential benefits of enabling SMC-R

- Available via the service stream on existing z/OS releases as well
  - V1R13 - Apar PI27252/PTF UI24872
  - V2R1 - Apar PI29165/PTFs UI24762 and UI24763
  - Does not require:
    - SMC-R code or RoCE hardware to use
    - Any changes in IP configuration (i.e. captures your normal TCP/IP workloads)
SMCAT Usage Overview

Activated by Operator command

*(Vary TCPIP,,SMCAT,dsn(smcatconfig)* – Input dataset contains:

- **Interval Duration**, list of IP addresses or IP subnets of peer z/OS systems ((i.e. systems that we can use SMC-R for)
  - If subnets are used, the entire subnet must be comprised of z/OS systems that are SMC-R eligible
  - It is important that all the IP addresses used for establishing TCP connections are specified (including DVIPAs, etc.)

- **At the end of the interval a summary report is generated that includes:**
  1. **Percent of traffic eligible for SMC-R** (% of TCP traffic that is eligible for SMC-R)
     - All traffic that matches configured IP addresses (not using IPSec or FRCA)
  2. **Percent of traffic well suited for SMC-R** (your eligible traffic that is also “well suited” to SMC-R, excludes workloads with very short lived TCP connections that have trivial payloads)
     - Includes break out of TCP traffic send sizes (i.e. how large is the payload of each send request)
     - Helps users quantify SMC-R benefit (reduced latency vs reduced CPU cost)
SMCAT Usage Overview (continued)

The Summary Report includes 2 sections based on the specified IP addresses/subnets defined in SMCAT configuration file:

1. Potential benefit:
   All TCP traffic that matches the configuration - Includes TCP traffic that could not use SMC-R without changes (TCP traffic that does not meet the direct IP route connectivity requirement)
   
   This represents the opportunity of re-configuring routing topology to enable SMC-R

2. Immediate benefit:
   The TCP traffic that can use SMC-R immediately / as is (meets SMC-R direct route connectivity requirements). Subset of section 1.
   
   Detected by the tool automatically (non-routed traffic)
SMC Applicability Tool Sample Report (Part 1. Direct Connections)

TCP SMC-R traffic analysis for matching direct connections
---------------------------------------------
Connections meeting direct connectivity requirements

50% of connections can use SMC-R (eligible)
67% of eligible connections are well-suited for SMC-R
79% of total traffic (segments) is well-suited for SMC-R
81% of outbound traffic (segments) is well-suited for SMC-R
75% of inbound traffic (segments) is well-suited for SMC-R

Interval Details:
Total TCP Connections: 6
Total SMC-R eligible connections: 3
Total SMC-R well-suited connections: 2
Total outbound traffic (in segments): 274
SMC-R well-suited outbound traffic (in segments): 222
Total inbound traffic (in segments): 211
SMC-R well-suited inbound traffic (in segments): 159

Application send sizes used for well-suited connections:

<table>
<thead>
<tr>
<th>Size</th>
<th># sends</th>
<th>Percentage</th>
</tr>
</thead>
<tbody>
<tr>
<td>1500 (&lt;=1500):</td>
<td>1</td>
<td>20%</td>
</tr>
<tr>
<td>4K (&gt;1500 and &lt;=4k):</td>
<td>1</td>
<td>20%</td>
</tr>
<tr>
<td>8K (&gt;4k and &lt;= 8k):</td>
<td>0</td>
<td>0%</td>
</tr>
<tr>
<td>16K (&gt;8k and &lt;= 16k):</td>
<td>0</td>
<td>0%</td>
</tr>
<tr>
<td>32K (&gt;16k and &lt;= 32k):</td>
<td>0</td>
<td>0%</td>
</tr>
<tr>
<td>64K (&gt;32k and &lt;= 64k):</td>
<td>1</td>
<td>20%</td>
</tr>
<tr>
<td>256K (&gt;64K and &lt;= 256K):</td>
<td>2</td>
<td>40%</td>
</tr>
<tr>
<td>&gt;256K:</td>
<td>0</td>
<td>0%</td>
</tr>
</tbody>
</table>

End of report

How much of my TCP workload can benefit from SMC-R?

What kind of CPU savings can I expect from SMC-R?

This report is repeated for indirect IP connections
Configuring the SMCAT Dataset

SMCAT data set configuration

- Interval defaults to 60 minutes
  Max interval is 1440 minutes (24 hours)
- IPADDR is a list of IPv4 and Ipv6 addresses and subnets
  256 max combination of addresses and subnets
SMCAT Dataset Example

When SMCAT is started using this SMCAT configuration data set it will:

- Monitor TCP traffic for 2 hours for:
  - IPv6 prefix C5::1:2:3:4/126 and
  - IPv4 address 9.67.113.61
Starting and Stopping SMCAT

Vary TCPIP,,SMCAT command starts and stops the monitoring tool:
- datasetname value indicates that SMCAT is being turned on
- datasetname contains the SMCATCFG statement that specifies monitoring interval and IP addresses or subnets to be monitored
- OFF will stop SMCAT monitoring and generate report

```
>>__Vary__TCPIP,__,SMCAT__,datasetname____><
   |procname|,OFF|

VARY TCPIP,TCPPROC,SMCAT,USER99.TCPIP.SMCAT1
```
SMCAT Operator Messages

Key messages – Operator console:

- EZD2031I SMC APPLICABILITY TOOL HAS STARTED COLLECTING DATA
- EZD2032I SMC APPLICABILITY TOOL HAS STOPPED COLLECTING DATA

Configuration information and the SMCAT report are sent to the system log:

STC06578  EZD2040I TCP/IP CS V2R2 TCPIP Name: TCPIP
080  SMC Applicability Configuration Parameters - 02/04/2015, 10:09:49.08
080  Interval: 3 minutes
080  IP addresses/subnets being monitored
080  9.67.113.61
080  C5::1:2:3:4/126
080  End of configuration parameters
Thank you!
Appendix A: 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

- **GetIfs**
  - 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 GetIfs, GetIfStats and GetIfStatsExtended for RNIC interfaces
- One record per RNIC interface
- Same RNIC information as reported in GetIfs for RNIC interface
  - Provided to correlate this record with GetIfs information
  - PNet ID can be used to correlate this record with GetSmcLinks information
- RNIC stack statistics (GetIfStats) always provided, even if RNIC interface is not active
- VTAM tuning statistics (GetIfStatsExtended) 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
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
SMCRGRPSTATS: 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
Appendix B: SMC-R dedicated
Select by number or action code and press Enter.

1. Add like .................................. (a)
2. Repeat (Copy) processor configurations (r)
3. Change ................................... (c)
4. *Prime serial number ............... (i)
5. Delete ................................. (d)
6. View processor definition ........... (v)
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 F2=Split F3=Exit F9=Swap F12=Cancel

F8=Forward F20=Right F22=Command
SMC-R HCD Configuration

Select one or more PCIe functions, then press Enter. To add, use F11.

Processor ID . . . . : P88       zHelix P88

<table>
<thead>
<tr>
<th>FID</th>
<th>PCHID</th>
<th>VF+</th>
<th>Type+</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>001</td>
<td>380</td>
<td></td>
<td>ROCE</td>
<td></td>
</tr>
<tr>
<td>002</td>
<td>3A0</td>
<td></td>
<td>ROCE</td>
<td></td>
</tr>
<tr>
<td>003</td>
<td>3C0</td>
<td></td>
<td>ROCE</td>
<td></td>
</tr>
<tr>
<td>004</td>
<td>3E0</td>
<td></td>
<td>ROCE</td>
<td></td>
</tr>
<tr>
<td>005</td>
<td>500</td>
<td></td>
<td>ROCE</td>
<td></td>
</tr>
<tr>
<td>006</td>
<td>520</td>
<td></td>
<td>ROCE</td>
<td></td>
</tr>
<tr>
<td>007</td>
<td>540</td>
<td></td>
<td>ROCE</td>
<td></td>
</tr>
<tr>
<td>008</td>
<td>560</td>
<td></td>
<td>ROCE</td>
<td></td>
</tr>
</tbody>
</table>

8 RoCE Adapters defined Function IDs used by TCPIP config

Physical Channel IDs
SMC-R HCD Configuration – Define a PFID

Specify or revise the following values.

- Processor ID: P88
- Function ID: __
- Type: ROCE
- PCHID: 380
- Virtual Function ID: __
- Description: 

You fill in – Function ID will match TCPIP GLOBALCONFIG value.
SMC-R HCD Configuration – Add PNETID

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 – Enable/Disable SMC-R

- **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
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
Load balancers might use data traffic as an indication that a TCP connection is healthy

Could 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

New parm for SMC-R