CS z/OS zSeries and System z9 Hardware Exploitation

System z9 and zSeries hardware exploitation - agenda

- **OSA update**
- **10 Gigabit Ethernet support**
- **QDIO OSA-Express2 segmentation offload**

OSA update

OSA-Express connectivity and CHPID support - overview

Feature	Feature Name	Ports	z900	z990	z9-109	CHPIDs	Connectors
5201	OSA-2 Token Ring	2	X	N / A	N / A	OSA	Copper, RJ-45
5202	OSA-2 FDDI	1	X	N / A	N / A	OSA	Fiber, SC Duplex
2362	OSA-E 155 ATM SM	2	X	RPQ	N / A	OSD, OSE	Fiber, SC Duplex
2363	OSA-E 155 ATM MM	2	X	RPQ	N / A	OSD, OSE	Fiber, SC Duplex
2364	OSA-E GbE LX	2	X	C	C	OSD	Fiber, SC Duplex
2365	OSA-E GbE SX	2	X	C	C	OSD	Fiber, SC Duplex
2366	OSA-E Fast Ethernet	2	X	C	C	OSD, OSE	Copper, RJ-45
2367	OSA-E Token Ring	2	X	X	N / A	OSD, OSE	Copper, RJ-45
1364	OSA-E GbE LX	2	09/04	06/03	C	OSD	Fiber, LC Duplex
1365	OSA-E GbE SX	2	09/04	06/03	C	OSD	Fiber, LC Duplex
1366	OSA-E 1000BASE-T Ethernet	2	N / A	06/03	C	OSC, OSD, OSE	Copper, RJ-45
3364	OSA-E2 GbE LX	2	N / A	01/05	X	OSD, OSN *	Fiber, LC Duplex
3365	OSA-E2 GbE SX	2	N / A	01/05	X	OSD, OSN *	Fiber, LC Duplex
3366	OSA-E2 1000BASE-T Ethernet	2	N / A	N / A	X	OSC, OSD, OSE, OSN *	Copper, RJ-45
3368	OSA-E2 10 GbE LR	1	N / A	01/05	X	OSD	Fiber, SC Duplex

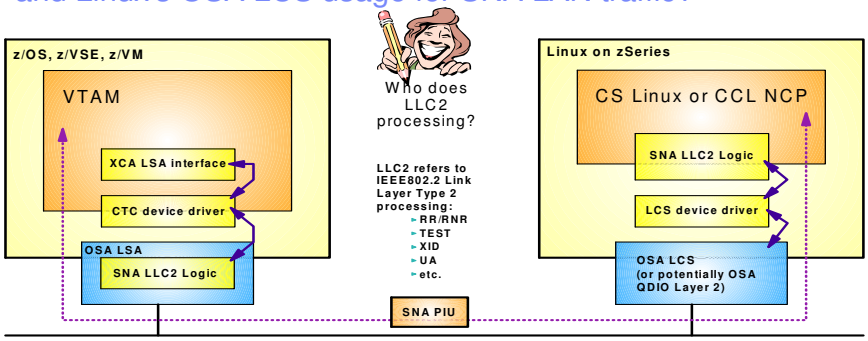
LX = Long wavelength transceiver, SX = Short wavelength transceiver, LR - Long Reach transceiver
X = Available for ordering C = Carry forward on an upgrade from z900 or z990
* OSN is exclusive to z9-109. Hardware availability is 09/16/05

What are the CHPID types used for?

		Traffic type				
CHPID type	Feature	SNA/APPN/HPR	TCP/IP	3270	NCP	OSA/SF required
OSD zSeries System z9	GbE, 10 GbE 1000BASE-T Ethernet Fast Ethernet	No (L3) Use EE or TN3270E Yes (L2)	Yes	No	No	No
OSE zSeries System z9	1000BASE-T Ethernet Fast Ethernet	Yes	Yes	No	No	Yes
OSC z990, z890 z9-109	1000BASE-T Ethernet	No	No	Yes	No	No
OSN z9-109 exclusive	1000BASE-T Ethernet GbE	No	No	No	Yes	No

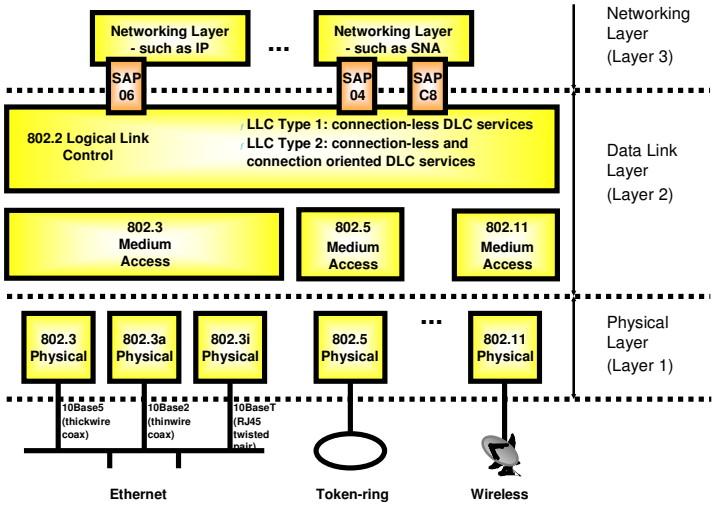
- **z/OS and Linux on zSeries support both IPv4 and IPv6 traffic over QDIO layer 3 interfaces.**
- **QDIO layer 2 mode is supported on z890, z990, and z9-109 only.**
- **Only Linux currently supports QDIO layer 2 mode.**
  - When using QDIO layer 2 mode for IP traffic, none of the OSA QDIO layer 3 IP assist functions are available
    - ARP offload, Large send segmentation offload, checksum offload, etc.

What is the difference between VTAM's OSA LSA usage and Linux's OSA LCS usage for SNA LAN traffic?



- **VTAM does not include SNA LLC2 processing logic, but has a higher-level interface defined that generally is known as LSA (Link Services Architecture).**
  - VTAM and SNA-specific LLC2 code in the OSA adapter communicates with each other using the LSA primitives
  - The actual device driver between VTAM and the OSA adapter in the case of LSA is a CTC device driver
- **Linux and the SNA software running on Linux provide full SNA LLC2 logic and are able to present fully built SNA LAN frames to the OSA adapter**
  - Can use the LCS device driver to interface with the OSA adapter (a LAN frame is a LAN frame!)
  - Also opens up for potentially using QDIO in layer 2 mode since the SNA solutions on Linux do not depend on SNA-specific capabilities in the OSA adapter

IEEE802 - the lower layers - structure



10 Gigabit Ethernet support

10 Gigabit Ethernet support - QDIO

- **CS z/OS v1R7 adds support for OSA-Express2 10 Gigabit Ethernet (Gbe) LR (LR = Long Range) feature**
- **Requires z990, z890, or z9-109**
- **Configured and managed exactly like Gigabit Ethernet**
- **Transparent except the following will reflect the actual speed:**
  1. the Speed field on the Netstat DEVLINKS/-d report output
  2. the SNMP MIB object ifHighSpeed (from the IF-MIB)

```
DevName: OTGETH1           DevType: MPCIPA
  DevStatus: Ready
  LnkName: LOTGETH1          LnkType: IPAQENET   LnkStatus: Ready
    NetNum: n/a  QueSize: n/a  Speed: 0000010000
    IpBroadcastCapability: No
.
.
DevName: OGETHD            DevType: MPCIPA
  DevStatus: Ready
  LnkName: LOGETHD           LnkType: IPAQENET   LnkStatus: Ready
   NetNum: n/a  QueSize: n/a   Speed: 0000001000
   IpBroadcastCapability: No
.
```

10 Gigabit Ethernet support PTFed back to z/OS V1R4

- **PTFs supplied for 10 Gigabit Ethernet Support to current releases**
  - VTAM (APAR OA09759)
    - V1R4 - UA15927
    - V1R5 - UA15928
    - V1R6 - UA15929
  - TCP (APAR PQ96769)
    - V1R4 - UQ95921
    - V1R5 - UQ95922
    - V1R6 - UQ95923

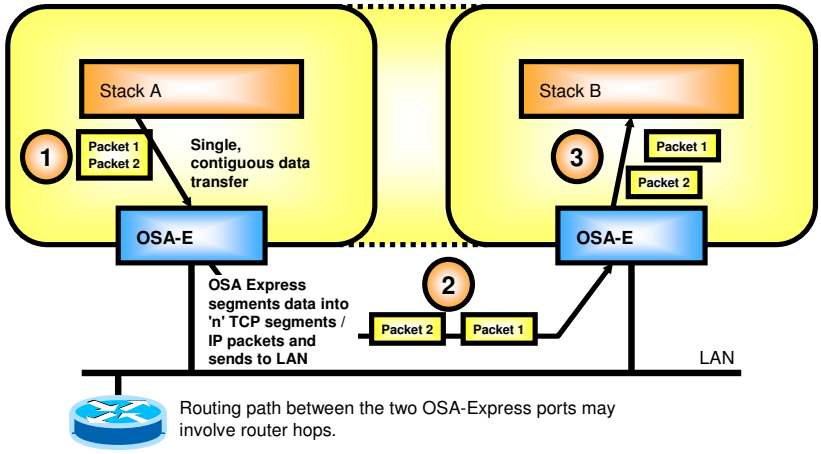
QDIO OSA-Express2 segmentation offload

TCP segmentation offload support

- **Segmenting consumes (high cost) host CPU cycles in the TCP stack**
- **Non-optimal use of Direct Memory Access**
- **CS z/OS V1R7 adds support for new OSA-Express feature (segmentation offload also referred to as 'Large Send')**
  - Offload most IPv4 TCP segmentation processing to OSA-Express in QDIO mode
  - Decrease host CPU utilization
  - Increase data transfer efficiency for IPv4 packets
- **Support automatically enabled when available in adapter**
  - Similar to existing checksum offload function
  - Checksum is offloaded whenever segmentation is offloaded
  - No configuration controls in TCP/IP
- **Applies to the OSA-Express2 features Gigabit Ethernet SX and LX, 10 Gigabit Ethernet LR**
  - Supports QDIO mode only (CHPID type OSD), and is exclusive to z990, z890, and z9-109
- **Segmentation offload support is available for z/OS V1R6.0 Communications Server.**
  - Solution was PTFed back to z/OS V1R6

Segmentation offload when next hop is reached via LAN

- **Segmentation can be offloaded to the OSA.**

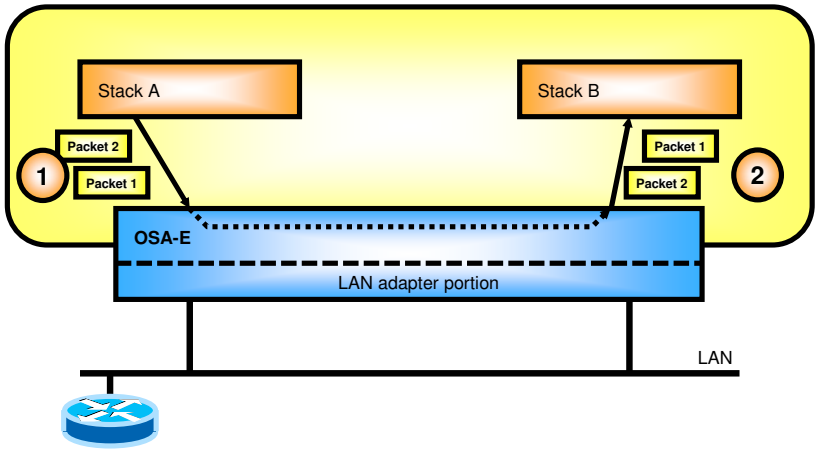


Restrictions

- **IPv4 only**
- **TCP transport only**
- **Outbound packets only**
- **Packets written to the LAN only (not to another stack sharing the OSA)**
- **Packets larger than MSS only**
- **For multipath, only when all devices in the multipath group support segoffload**
- **No IPSEC packets**

No segmentation offload when next hop is not reached via LAN

- **Segmentation cannot be offloaded when packets loop through the adapter to another target stack that shares the adapter.**



Netstat devicelink report

Netstat DEVLINKS/-d enhancement

```
DevName: OGETHD             DevType: MPCIPA
  DevStatus: Ready
  LnkName: LOGETHD            LnkType: IPAQENET   LnkStatus: Ready
    NetNum: n/a  QueSize: n/a  Speed: 0000001000
   IpBroadcastCapability: No
   CfgRouter: Pri                    ActRouter: Pri
   ArpOffload: Yes                   ArpOffloadInfo: Yes
   ActMtu: 8992
   VLANid: 3                         VLANpriority: Enabled
   ReadStorage: GLOBAL (4096K)       InbPerf: Balanced
   ChecksumOffload: Yes              SegmentationOffload: Yes
   SecClass: 255
 BSD Routing Parameters:
.
.
.
```

SNMP MIB support

NOTES

- ibmMvsIfFlag MIB object now contains the tcpSegOffloadEnabled(6) bit.

tcpSegOffloadEnabled(6) bit (0x02)

```
# snmp -v walk ibmMvsIfFlag
ibmMvsIfFlag.2 = '20'h
ibmMvsIfFlag.3 = '20'h
ibmMvsIfFlag.5 = 'a4'h
ibmMvsIfFlag.7 = 'aa'h <-- x'02' - on - Segmentation Offload Enabled
```

- Details about the ibmMvsIfFlag MIB object can be found in the notes pages.

Packet trace when segmentation offload is used

NOTES

- Packet trace enhancements

CTRACE COMP(SYSTCPDA) SUB((TCPSVT)) SHORT

```
-------------------------------------------------------------------------------
1352 SWEDEN    PACKET    00000004 19:38:01.875792 Packet Trace
 To Interface      : LOGETHC          Device: QDIO Ethernet     Full=2708
  Tod Clock        : 2005/01/18 19:38:01.875792                 Intfx: 7
  Sequence #       : 0                Flags: Pkt Out Offl
 IpHeader: Version : 4                Header Length: 20
  Tos              : 60               QOS: Interactive2 Normal Service
  Offload Length   : 2708             ID Numbers: E4E6-E4E8
  Fragment         : DontFragment     Offset: 0
  TTL              : 64               Protocol: TCP             CheckSum: 0000
  Source           : 197.11.107.1
  Destination      : 197.11.105.1

 TCP
  Source Port      : 1662  ()         Destination Port: 50030 ()
  Sequence Number  : 931515623        Ack Number: 1441765436
  Header Length    : 32               Flags: Ack Psh
  Window Size      : 32768            CheckSum: 5E20 0000 Urgent Data Pointer:
  Offload Segments : 3                Length: 1248              Last: 160
   Option          : NOP
   Option          : NOP
➢    Option          : Timestamp        Len: 10 Value: C23B12F5 Echo: C23B12F4
```

Packet trace when segmentation offload is used *(continued)*

NOTES

- Offload in session report

CTRACE COMP(SYSTCPDA) SUB((TCPSVT)) SHORT OPTIONS((SESSION))

```
TcpHdr    IO F         Seq         Ack RcvWnd  Data Delta Time       TimeStamp ...
 o AP      O . 3179919890 2925854225  32768  3540   0.000000 22:05:39.105695 ...
   A       I w 2925854225 3179862774  31646     0   0.000287 22:05:39.105982 ...
 o A       O . 3179923430 2925854225  32768 57116   0.000494 22:05:39.106476 ...
 o AP      O . 3179980546 2925854225  32768  7336   0.000061 22:05:39.106537 ...
```

```
Data Segment Stats:                  Inbound,        Outbound
 Number of data segments:                  0,            4176
 Maximum segment size:                  1460,            1460
 Largest segment size:                     0,           16544
 Average segment size:                     0,           16044
 Smallest segment size:                    0,             160
 Segments/window:                        0.0,             1.0
 Average bytes/window:                     0,           16047
 Most bytes/window:                        0,           16544
 Offload Sends:                                          4174          (99.95%)
 Offload Segments:                                      57217
 Offload Bytes:                                      67002936          (99.99%)
 Total Packets(normal + offload):                       57227          ( 7.31%)
```

VTAM TNSTATS to view how segmentation offload is applied

```
IST924I -------------------------------------------------------------
IST1233I DEV      = 0E2A       DIR      = WR/4
IST1755I SBALMAX  = 2          SBALAVG  = 1
IST1756I QDPTHMAX = 0          QDPTHAVG = 0
IST1723I SIGACNTO = 0          SIGACNT  = 6
IST1721I SBALCNTO = 0          SBALCNT  = 6
IST1722I PACKCNTO = 0          PACKCNT  = 21
IST1236I BYTECNTO = 0          BYTECNT  = 184984
IST1810I PKTIQDO  = 0          PKTIQD   = 0
IST1811I BYTIQDO  = 0          BYTIQD   = 0

IST924I -------------------------------------------------------------
IST1233I DEV      = 0E2E       DIR      = READ
IST1719I PCIREALO = 0          PCIREAL  = 6
IST1720I PCIVIRTO = 0          PCIVIRT  = 0
IST1750I PCITHRSO = 0          PCITHRSH = 0
IST1751I PCIUNPRO = 0          PCIUNPRD = 0
IST1752I RPROCDEO = 0          RPROCDEF = 0
IST1753I RREPLDEO = 0          RREPLDEF = 0
IST1754I NOREADSO = 0          NOREADS  = 0
IST1721I SBALCNTO = 0          SBALCNT  = 6
IST1722I PACKCNTO = 0          PACKCNT  = 21
IST1236I BYTECNTO = 0          BYTECNT  = 186084
IST1810I PKTIQDO  = 0          PKTIQD   = 0
IST1811I BYTIQDO  = 0          BYTIQD   = 0
```

- There are 2 VTAM TNSTAT responses on this page.
- The TNSTATs were gathered from a single data transfer operation (180KB transmission).
- The top half of the page shows TNSTATs from the sending host and the bottom section shows TNSTATs from the receiving host.
- The packet count is the same but the bytecount is different. The reason the bytecount is different is because the OSA-Express2 generated headers for each segment.

Things to think about

- **Big send buffer (up to 56KB) maximizes offloading. Configure the TCP send buffer size using the following existing mechanisms...**
  - TCPSENDBfrsize on TCPCONFIG statement sets default for all applications
  - SETSOCKOPT (SO_SNDBUF) by the application overrides default
- **Send buffer size also limited by receive buffer size at other end of connection.**
  - TCPRCVBufrsize on TCPCONFIG statement sets default for all applications
  - SETSOCKOPT (SO_RCVBUF) by the application overrides default
- **APARs supplied for QDIO OSA-Express2 Segmentation Offload for z/OS V1R**
  - TCP APAR: PK02490 - PTF: UK04060 and UK04061
  - VTAM APAR: OA11148 - PTF: UA18116
- **There is no interdependency between the VTAM code and the TCP/IP code. The VTAM code can be applied without the TCP/IP code and vice versa. However, segmentation offload is not enabled unless both pieces are applied.**

Segmentation offload performance details

- **OSAE-2, 1 GbE (versus no segmentation offload):**

Workload	Trans/Sec Delta %	CPU/Tran Delta %
RR 60	+ 1.3 %	- 0.7 %
CRR 9	+ 2 %	- 0.1 %
STR (1/20M):		
64K(send)/32K(recv)	Equal	- 28.9 %
180K(send)/64K(recv)	Equal	- 36.3 %
256K(send)/64K(recv)	Equal	- 39.2 %

- **OSAE-2, 10 GbE (versus no segmentation offload):**

Workload	Trans/Sec Delta %	CPU/Tran Delta %
RR 60	+ 1.7 %	- 2 %
CRR 60	+ 5.2 %	- 1 %
STR (1/20M):		
64K(send)/32K(recv)	+ 1.1 %	- 33.4 %
180K(send)/64K(recv)	+ 1.5 %	- 41.5 %
256K(send)/64K(recv)	+ 0.4 %	- 44.9 %

Trademarks, Copyrights and Disclaimers

The following terms are trademarks or registered trademarks of International Business Machines Corporation in the United States, other countries, or both:

IBM	CICS	IMS	MQSeries	Tivoli
IBM(logo)	Cloudscape	Informix	OS/390	WebSphere
e(logo)business	DB2	iSeries	OS/400	xSeries
AIX	DB2 Universal Database	Lotus	pSeries	zSeries

Java and all Java-based trademarks are trademarks of Sun Microsystems, Inc. in the United States, other countries, or both.

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

Intel, ActionMedia, LANDesk, MMX, Pentium and ProShare are trademarks of Intel Corporation in the United States, other countries, or both.

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

Linux is a registered trademark of Linus Torvalds.

Other company, product and service names may be trademarks or service marks of others.

Product data has been reviewed for accuracy as of the date of initial publication. Product data is subject to change without notice. This document could include technical inaccuracies or typographical errors. IBM may make improvements and/or changes in the product(s) and/or program(s) described herein at any time without notice. Any statements regarding IBM's future direction and intent are subject to change or withdrawal without notice, and represent goals and objectives only. References in this document to IBM products, programs, or services does not imply that IBM intends to make such products, programs or services available in all countries in which IBM operates or does business. Any reference to an IBM Program Product in this document is not intended to state or imply that only that program product may be used. Any functionally equivalent program, that does not infringe IBM's intellectual property rights, may be used instead.

Information is provided "AS IS" without warranty of any kind. THE INFORMATION PROVIDED IN THIS DOCUMENT IS DISTRIBUTED "AS IS" WITHOUT ANY WARRANTY, EITHER EXPRESS OR IMPLIED. IBM EXPRESSLY DISCLAIMS ANY WARRANTIES OF MERCHANTABILITY, FITNESS FOR A PARTICULAR PURPOSE OR NONINFRINGEMENT. IBM shall have no responsibility to update this information. IBM products are warranted, if at all, according to the terms and conditions of the agreements (e.g., IBM Customer Agreement, Statement of Limited Warranty, International Program License Agreement, etc.) under which they are provided. Information concerning non-IBM products was obtained from the suppliers of those products, their published announcements or other publicly available sources. IBM has not tested those products in connection with this publication and cannot confirm the accuracy of performance, compatibility or any other claims related to non-IBM products. IBM makes no representations or warranties, express or implied, regarding non-IBM products and services.

The provision of the information contained herein is not intended to, and does not, grant any right or license under any IBM patents or copyrights. Inquiries regarding patent or copyright licenses should be made, in writing, to:

IBM Director of Licensing
IBM Corporation
North Castle Drive
Armonk, NY 10504-1785
U.S.A.

Performance is based on measurements and projections using standard IBM benchmarks in a controlled environment. All customer examples described are presented as illustrations of how those customers have used IBM products and the results they may have achieved. The actual throughput or performance 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 or performance improvements equivalent to the ratios stated here.