# Getting the most Out of Your OSA Adapter with z/OS Communication Server

Alfred B Christensen – alfredch@us.ibm.com
IBM Raleigh, NC, USA

Session: 8325
Thursday, March 3, 2011: 11:00 AM-12:00 PM

# Getting the Most Out of Your OSA Adapter with z/OS Communication Server

| | |
|---|---|
| **Session number:** | 8325 |
| **Date and time:** | Thursday, March 3, 2011: 11:00 AM-12:00 PM |
| **Location:** | Room 212A (Anaheim Convention Center) |
| **Program:** | Communications Infrastructure |
| **Project:** | Communications Server |
| **Track:** | Network Support and Management and SNA/IP Integration |
| **Classification:** | Technical |
| **Speaker:** | Alfred B Christensen, IBM |
| **Abstract:** | zOSA-Express ports are used for both SNA and IP LAN connectivity to System z operating systems. In this session we will focus on how z/OS Communications Server uses OSA for IPv4 LAN connectivity based on Queued Direct IO (QDIO).  The session will discuss both configuration and operational aspects, such as maintenance of the OSA address table (OAT), ARP processing, use of Virtual IP Addresses (VIPA), interface availability, sharing capabilities, VLAN support, Virtual MACs, OSA Direct SNMP, and other management related functions. The session will focus on the software use of OSA and will discuss hardware aspects only where such aspects are of importance to understanding how the hardware and software cooperate to deliver the desired functions. |

# Trademarks, notices, and disclaimers

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

- Advanced Peer-to-Peer Networking®
- AIX®
- alphaWorks®
- AnyNet®
- AS/400®
- BladeCenter®
- Candle®
- CICS®
- DataPower®
- DB2 Connect
- DB2®
- DRDA®
- e-business on demand®
- e-business (logo)
- e business(logo)®
- ESCON®
- FICON®

- GDDM®
- GDPS®
- Geographically Dispersed Parallel Sysplex
- HiperSockets
- HPR Channel Connectivity
- HyperSwap
- i5/OS (logo)
- i5/OS®
- IBM eServer
- IBM (logo)®
- IBM®
- IBM zEnterprise™ System
- IMS
- InfiniBand ®
- IP PrintWay
- IPDS
- iSeries
- LANDP®

- Language Environment®
- MQSeries®
- MVS
- NetView®
- OMEGAMON®
- Open Power
- OpenPower
- Operating System/2®
- Operating System/400®
- OS/2®
- OS/390®
- OS/400®
- Parallel Sysplex®
- POWER®
- POWER7®
- PowerVM
- PR/SM
- pSeries®
- RACF®

- Rational Suite®
- Rational®
- Redbooks
- Redbooks (logo)
- Sysplex Timer®
- System i5
- System p5
- System x®
- System z®
- System z9®
- System z10
- Tivoli (logo)®
- Tivoli®
- VTAM®
- WebSphere®
- xSeries®
- z9®
- z10 BC
- z10 EC

- zEnterprise
- zSeries®
- z/Architecture
- z/OS®
- z/VM®
- z/VSE

\* All other products may be trademarks or registered trademarks of their respective companies.

**The following terms are trademarks or registered trademarks of International Business Machines Corporation in the United States or other countries or both:**
- 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 there from.
- 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 trademarks of Microsoft Corporation in the United States, other countries, or both.
- InfiniBand is a trademark and service mark of the InfiniBand Trade Association.
- 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.
- UNIX is a registered trademark of The Open Group in the United States and other countries.
- Linux is a registered trademark of Linus Torvalds in the United States, other countries, or both.
- 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.
- IT Infrastructure Library is a registered trademark of the Central Computer and Telecommunications Agency, which is now part of the Office of Government Commerce.

**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.

Refer to www.ibm.com/legal/us for further legal information.

# Agenda

- Introduction

- What is QDIO?

- OSA inbound "routing"

- Overview of selected recent OSA enhancements

- AppendixA: Monitoring OSA using SNMPl

- Appendix B: OSA-Express3 dual port IOCP, TRLE, and z/OS CS Interface definitions
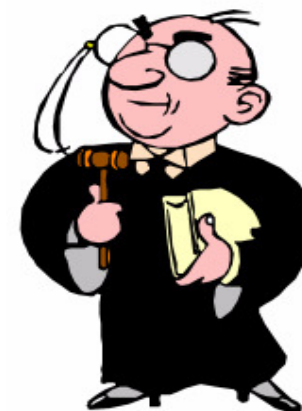
*Disclaimer: All statements regarding IBM future direction or intent, including current product plans, are subject to change or withdrawal without notice and represent goals and objectives only. All information is provided for informational purposes only, on an "as is" basis, without warranty of any kind.*

# Getting the most out of your OSA adapter with z/OS Communications Server

# Introduction

IBM ®

# OSA-Express connectivity overview - EC

| Feature | Feature Name | Ports | z900 | z990 | z9 EC | z10 EC | z196 | CHPIDs | Connectors |
|---------|-------------|-------|------|------|-------|--------|------|--------|------------|
| **OSA-Express** | | | | | | | | | |
| 2362 | 155 ATM SM | 2 | X | RPQ | N / A | N / A | N / A | OSD, OSE | SC Duplex |
| 2363 | 155 ATM MM | 2 | X | RPQ | N / A | N / A | N / A | OSD, OSE | SC Duplex |
| 2364 | GbE LX | 2 | X | C | C | N / A | N / A | OSD, L2/L3** | SC Duplex |
| 2365 | GbE SX | 2 | X | C | C | N / A | N / A | OSD, L2/L3** | SC Duplex |
| 2366 | Fast Ethernet | 2 | X | C | C | N / A | N / A | OSD, OSE | RJ-45 |
| 2367 | Token Ring | 2 | X | X | N / A | N / A | N / A | OSD, OSE | RJ-45 |
| **OSA-Express** | | | | | | | | | |
| 1364 | GbE LX | 2 | 09/04 | X | C | N / A | N / A | OSD, L2/L3** | LC Duplex |
| 1365 | GbE SX | 2 | 09/04 | X | C | N / A | N / A | OSD, L2/L3** | LC Duplex |
| 1366 | 1000BASE-T | 2 | N / A | X | C | N / A | N / A | OSC, OSD L2/L3, OSE | RJ-45 |
| **OSA-Express2** | | | | | | | | | |
| 3364 | GbE LX | 2 | N / A | 01/05 | X | wd Jun09 | C | OSD L2/L3, OSN* | LC Duplex |
| 3365 | GbE SX | 2 | N / A | 01/05 | X | wd Jun09 | C | OSD L2/L3, OSN* | LC Duplex |
| 3366 | 1000BASE-T | 2 | N / A | 05/06 | X | X | C | OSC, OSD L2/L3, OSE, OSN* | RJ-45 |
| 3368 | 10 GbE LR | 1 | N / A | 01/05 | X | wd Jun08 | C | OSD L2/L3** | SC Duplex |
| **OSA-Express3** | | | | | | | | | |
| 3362 | GbE LX | 4 | N / A | N / A | N / A | 30May08 | X | OSD L2/L3, OSN | LC Duplex |
| 3363 | GbE SX | 4 | N / A | N / A | N / A | 30May08 | X | OSD L2/L3, OSN | LC Duplex |
| 3367 | 1000BASE-T | 4 | N / A | N / A | N / A | 28Oct08 | X | OSC, OSD L2/L3, OSE, OSN*, OSM*** | RJ-45 |
| 3370 | 10 GbE LR | 2 | N / A | N / A | N / A | 30May08 | X | OSD L2/L3, OSX*** | LC Duplex |
| 3371 | 10 GbE SR | 2 | N / A | N / A | N / A | 28Oct08 | X | OSD L2/L3, OSX*** | LC Duplex |

X = Available for ordering  C = Carry forward on an upgrade
*   OSN is exclusive to z10 and z9
**   L2/L3 = Layer 2/Layer 3 which is applicable to z990 and later servers

*** OSM and OSX are exclusive to z196
SOD – z196 is the last high-end server to support OSA-Express2

# OSA-Express connectivity overview - BC

| Feature | Feature Name | Ports | z800 | z890 | z9 BC | z10 BC | CHPIDs | Connectors |
|---------|--------------|-------|------|------|-------|--------|--------|------------|
| **OSA-Express** | | | | | | | | |
| 2362 | 155 ATM SM | 2 | X | RPQ | N / A | N / A | OSD, OSE | SC Duplex |
| 2363 | 155 ATM MM | 2 | X | RPQ | N / A | N / A | OSD, OSE | SC Duplex |
| 2364 | GbE LX | 2 | X | C | C | N / A | OSD, L2/L3 ** | SC Duplex |
| 2365 | GbE SX | 2 | X | C | C | N / A | OSD, L2/L3 ** | SC Duplex |
| 2366 | Fast Ethernet | 2 | X | C | C | N / A | OSD, OSE | RJ-45 |
| 2367 | Token Ring | 2 | X | X | N / A | N / A | OSD, OSE | RJ-45 |
| **OSA-Express** | | | | | | | | |
| 1364 | GbE LX | 2 | 09/04 | X | C | N / A | OSD, L2/L3 ** | LC Duplex |
| 1365 | GbE SX | 2 | 09/04 | X | C | N / A | OSD, L2/L3 ** | LC Duplex |
| 1366 | 1000BASE-T | 2 | N / A | X | C | N / A | OSC, OSD L2/L3, OSE | RJ-45 |
| **OSA-Express2** | | | | | | | | |
| 3364 | GbE LX | 2 | N / A | 01/05 | X | wd Jun09 | OSD L2/L3, OSN * | LC Duplex |
| 3365 | GbE SX | 2 | N / A | 01/05 | X | wd Jun09 | OSD L2/L3, OSN * | LC Duplex |
| 3366 | 1000BASE-T | 2 | N / A | 05/06 | X | X | OSC, OSD L2/L3, OSE, OSN * | RJ-45 |
| 3368 | 10 GbE LR | 1 | N / A | 01/05 | X | C | OSD L2/L3 ** | SC Duplex |
| **OSA-Express3** | | | | | | | | |
| 3362 | GbE LX | 4 | N / A | N / A | N / A | X | OSD L2/L3, OSN | LC Duplex |
| 3363 | GbE SX | 4 | N / A | N / A | N / A | X | OSD L2/L3, OSN | LC Duplex |
| 3367 | 1000BASE-T | 4 | N / A | N / A | N / A | X | OSC, OSD L2/L3, OSE, OSN * | RJ-45 |
| 3369 | 2P 1000BASE-T | 2 | N / A | N / A | N / A | X | OSC, OSD L2/L3, OSE, OSN * | RJ-45 |
| 3370 | 10 GbE LR | 2 | N / A | N / A | N / A | X | OSD L2/L3 | LC Duplex |
| 3371 | 10 GbE SR | 2 | N / A | N / A | N / A | X | OSD L2/L3 | LC Duplex |
| 3373 | 2P GbE SX | 2 | N / A | N / A | N / A | X | OSD L2/L3, OSN | LC Duplex |

X = Available for ordering  C = Carry forward on an upgrade
* OSN is exclusive to z10 and z9   ** L2/L3 = Layer 2/Layer 3 which is applicable to z990 and later servers

# What are the CHPID types used for by selected System z operating systems?

| Opera-ting system | CHPID Type: Device driver operation: | OSE LCS | OSE LSA | OSD QDIO L3 | OSD QDIO L2 | OSN CDLC | OSC (CDLC) | OSM QDIO L2/L3 | OSX QDIO L3 | Comments |
|---|---|---|---|---|---|---|---|---|---|---|
| z/OS | VTAM SNA LLC2 LAN | | ✓ | | | | | | | |
| | VTAM SNA CDLC | | | | | ✓ | | | | |
| | IPv4 | ✓ | | ✓ | | | | | ✓ | |
| | IPv6 | | | ✓ | | | | ✓ | ✓ | |
| | z/OS NIP/MCS Console | | | | | | ✓ | | | |
| | z/OS non-SNA DFT VTAM | | | | | | ✓ | | | VTAM Local 3270 major node |
| z/VM | VTAM SNA LLC2 LAN | | ✓ | | | | | | | |
| | VTAM SNA CDLC | | | | | ✓ | | | | |
| | IPv4 | ✓ | | ✓ | | | | | ✓ | |
| | IPv6 | | | ✓ | | | | ✓ | ✓ | z/VM 5.1 |
| | z/VM Virtual Switch | | | ✓ | ✓ | | | | | z/VM 5.1 for layer 2 support |
| | z/VM IPL Console | | | | | | ✓ | | | |
| | z/VM 3270 terminal | | | | | | ✓ | | | |
| Linux on System z | CSL SNA LLC2 LAN | ✓ | | | ✓ | | | | | CSL 6.2.1 |
| | CCL SNA LLC2 LAN | ✓ | | | ✓ | | | | | CCL 1.2, Linux 2.6 |
| | CCL SNA CDLC | | | | | ✓ | | | | CCL 1.2, Linux 2.6 |
| | IPv4 | ✓ | | ✓ | ✓ | | | | ✓ | |
| | IPv6 | | | ✓ | ✓ | | | ✓ | ✓ | |

## Getting the most out of your OSA adapter with z/OS Communications Server

# What is QDIO?

# Queued Direct IO (QDIO) – How System z accesses a Local Area Network

**OSE: LCS or LSA OSA**

Host Memory

Channel

Channel-attached device

Control Unit

NIC

OSA

**OSD: QDIO OSA-Express(2)**

Host Memory

Store and forward

NIC

OSA

**OSD: QDIO OSA-Express3**

Host Memory

Data router

NIC

OSA

*The same way a mainframe connected to an old IBM 3172, or a Cisco CIP*

*The OSA adapter and System z can access shared memory. OSA interrupts System z using PCI, while System z "interrupts" OSA via Signaling (SIGA instruction)*

*OSA-Express3 uses direct memory access (DMA) technologies and hardware data router – allowing data to flow through OSA without slow store and forward processing.*

# OSA and QDIO work together

- **QDIO Layer 3 – for TCP/IP traffic only (IPv4 and IPv6)**
  - z/OS, z/VM, z/VSE, z/TPF, and Linux on System z
  - OSA IP Processing Assist (IPA) offloads many TCP/IP functions to the OSA adapter, of which some are:
    - Performs all ARP processing
    - Provides Multicast support
    - Builds MAC and LLC headers
    - Performs checksum processing
    - Performs outbound TCP segmentation
  - For SNA/APPN/HPR traffic with QDIO layer 3: use TN3270 and Enterprise Extender

- **QDIO Layer 2 – network protocol agnostic**
  - z/VM V5.2 with PTFs and Linux on System z
    - z/OS does not support QDIO Layer 2
  - Network protocol may be TCP/IP, SNA LLC2, NetBIOS, etc. (no IP Assist)
  - Linux uses QDIO Layer 2 for SNA LLC2 traffic in/out of Linux
    - IBM Communications Server for Linux and Communications Controller for Linux

- **OSA and QDIO are designed for high speed communication between System z and the Local Area Network**
  - Reduced TCP/IP path length
  - QDIO IP Processing Assist
  - LPAR-to-LPAR communication with port sharing
  - Direct Memory Access (DMA) Protocol
    - Memory-to-memory communication
      - I/O interrupts minimized
      - Continuous direct data exchanges
  - Dynamic customization

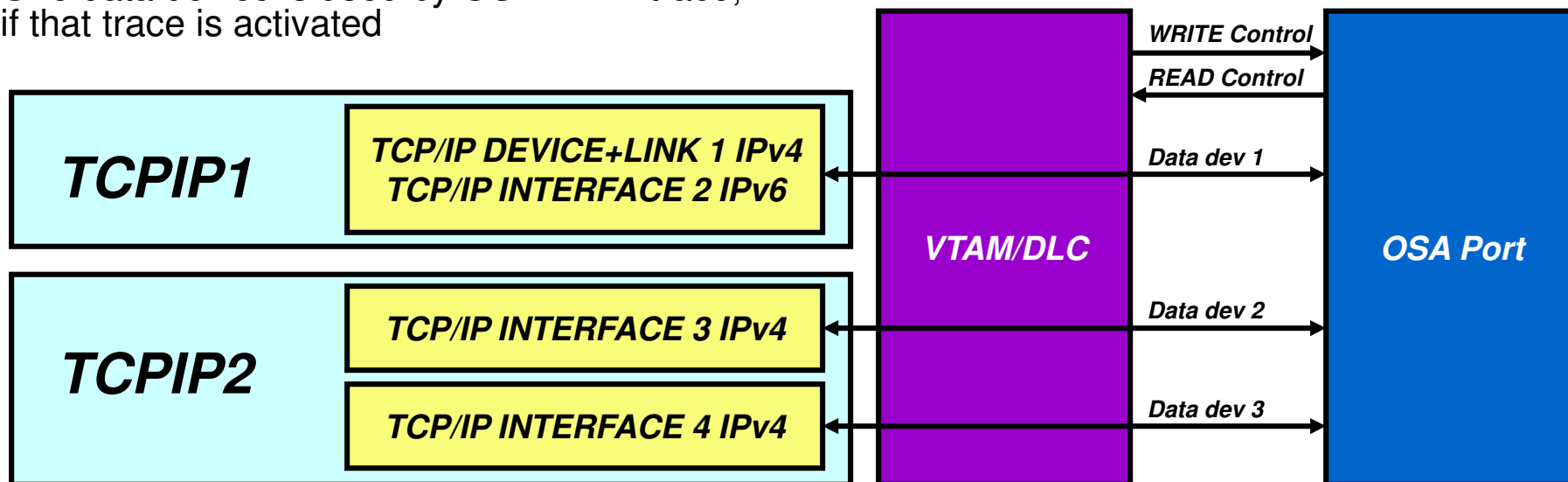- **10 Gigabit Ethernet, Gigabit Ethernet, 1000BASE-T Ethernet (1000, 100, and 10 Mbit)**

*Host Memory*

*Data router*

*NIC*

*OSA*
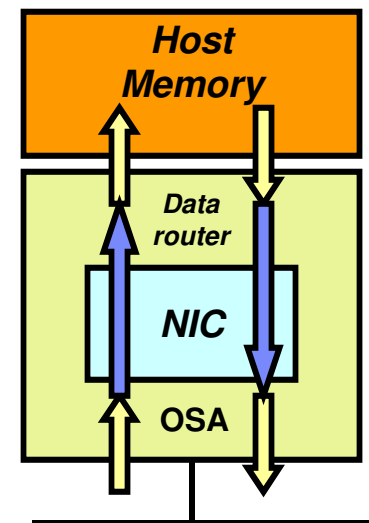
# A little more on QDIO and devices

- Control data between z/OS CS and the OSA port is exchanged over one read device and one write device (READ even address, WRITE odd address)

- The READ and WRITE device pair is shared among all the IP interface definitions in an LPAR that use the same OSA port

- Data is exchanged over data devices

- One data device is used per:
  – A set of exactly one IPv4 and one IPv6 interface
    • IPv4: DEVICE+LINK, IPv6: INTERFACE stmt.
  – One IPv4 interface defined with INTERFACE stmt.
  – One IPv6 interface

- One data device is used by OSAENTA trace, if that trace is activated

**TRLE**

```
LNCTL=MPC,
READ=(2FD6),
WRITE=(2FD7),
MPCLEVEL=QDIO,
DATAPATH=(2FD8,2FD9,2FDA,2FDB),
PORTNAME=(O3ETHB1P),
PORTNUM=(1)
```

**TCPIP1**

TCP/IP DEVICE+LINK 1 IPv4
TCP/IP INTERFACE 2 IPv6

**TCPIP2**

TCP/IP INTERFACE 3 IPv4

TCP/IP INTERFACE 4 IPv4

**VTAM/DLC**

WRITE Control
READ Control
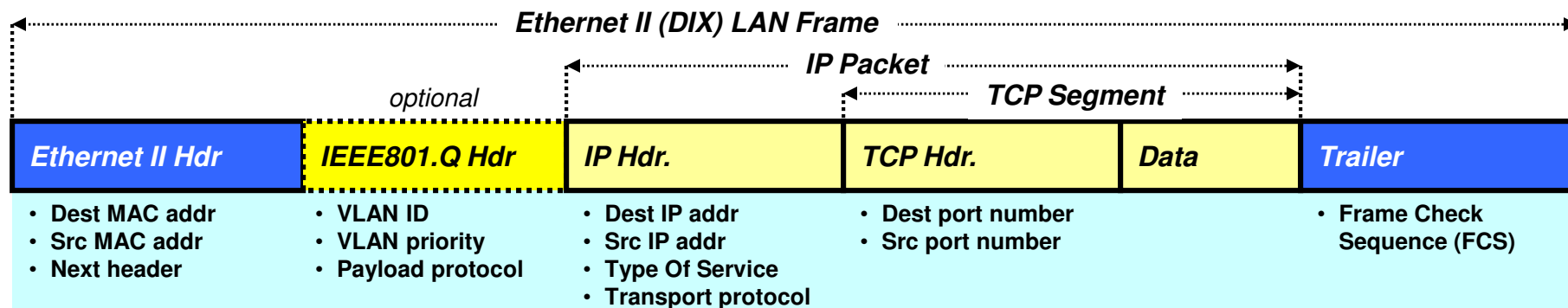
Data dev 1

Data dev 2

Data dev 3

**OSA Port**

# Getting the most out of your OSA adapter with z/OS Communications Server

# OSA inbound "routing"
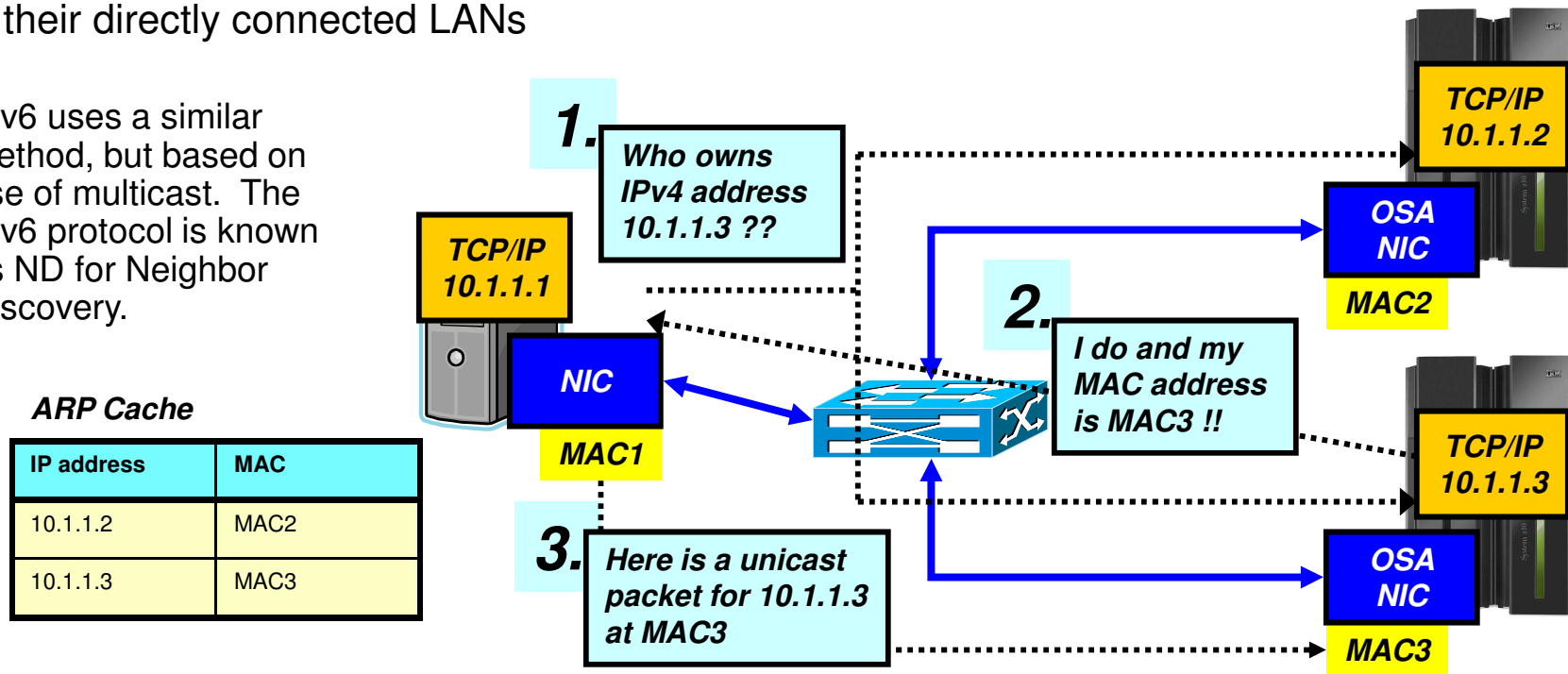
IBM®

# Some basic LAN technology overview

- The LAN infrastructure transports "Frames" between Network Interface Cards (NICs) that are attached to the LAN media (Copper or fiber optic)
  - Ethernet II (DIX) – MTU 1500 and jumbo frame MTU 9000 (most common)
  - IEEE802 – MTU 1492 and jumbo frame MTU 8992

- Each NIC has a physical hardware address
  - A Media Access Control (MAC) address
    - Burned in (world-wide unique by vendors) or alternatively locally administered

- Every frame comes from a MAC and goes to a MAC
  - There are special MAC values for broadcast and multicast frames

- Every frame belongs to the physical LAN or to one of multiple Virtual LANs (VLAN) on the physical LAN
  - A VLAN ID is in the IEEE801.Q header if VLAN technologies are in use

- A frame carries a payload of a specified protocol type, such as ARP, IPv4, IPv6, SNA LLC2, etc.

| | | Ethernet II (DIX) LAN Frame | | | |
|---|---|---|---|---|---|
| | optional | IP Packet | | | |
| | | | TCP Segment | | |
| **Ethernet II Hdr** | **IEEE801.Q Hdr** | **IP Hdr.** | **TCP Hdr.** | **Data** | **Trailer** |
| • Dest MAC addr<br>• Src MAC addr<br>• Next header | • VLAN ID<br>• VLAN priority<br>• Payload protocol | • Dest IP addr<br>• Src IP addr<br>• Type Of Service<br>• Transport protocol | • Dest port number<br>• Src port number | | • Frame Check Sequence (FCS) |

# Correlation of IPv4 addresses and MAC addresses on a LAN – Address Resolution Protocol (ARP)

- An IPv4 node uses the ARP protocol to discover the MAC address of another IPv4 address that belongs to the same IPv4 subnet as it does itself.

- ARP requests are broadcasted to all NICs on the LAN

- The one NIC that has a TCP/IP stack with the requested IPv4 address responds directly back to the IPv4 node that sent out the broadcast

- Each IPv4 node maintains a cache of IPv4 addresses and associated MAC addresses on their directly connected LANs

IPv6 uses a similar method, but based on use of multicast. The IPv6 protocol is known as ND for Neighbor Discovery.

**1.** *Who owns IPv4 address 10.1.1.3 ??*

*TCP/IP 10.1.1.1*

**NIC** MAC1

**2.** *I do and my MAC address is MAC3 !!*

*TCP/IP 10.1.1.2*

*OSA NIC* MAC2

*TCP/IP 10.1.1.3*

*OSA NIC* MAC3

**3.** *Here is a unicast packet for 10.1.1.3 at MAC3*

### ARP Cache

| IP address | MAC |
|------------|------|
| 10.1.1.2   | MAC2 |
| 10.1.1.3   | MAC3 |

# An example – formatted by Wireshark (http://www.wireshark.org)

- **A sample Address Resolution Request frame**

```
Ethernet II, Src: 02:00:00:13:89:01 (02:00:00:13:89:01), Dst: Broadcast (ff:ff:ff:ff:ff:ff)
  Destination: Broadcast (ff:ff:ff:ff:ff:ff)
    Address: Broadcast (ff:ff:ff:ff:ff:ff)
    .... ...1 .... .... .... .... = IG bit: Group address (multicast/broadcast)
    .... ..1. .... .... .... .... = LG bit: Locally administered address (this is NOT the factory default)
  Source: 02:00:00:13:89:01 (02:00:00:13:89:01)
    Address: 02:00:00:13:89:01 (02:00:00:13:89:01)
    .... ...0 .... .... .... .... = IG bit: Individual address (unicast)
    .... ..1. .... .... .... .... = LG bit: Locally administered address (this is NOT the factory default)
  Type: 802.1Q Virtual LAN (0x8100)
802.1Q Virtual LAN
  000. .... .... .... = Priority: 0
  ...0 .... .... .... = CFI: 0
  .... 0001 0101 1111 = ID: 351
  Type: ARP (0x0806)
Address Resolution Protocol (request)
  Hardware type: Ethernet (0x0001)
  Protocol type: IP (0x0800)
  Hardware size: 6
  Protocol size: 4
  Opcode: request (0x0001)
  Sender MAC address: 02:00:00:13:89:01 (02:00:00:13:89:01)
  Sender IP address: 10.3.51.16 (10.3.51.16)
  Target MAC address: 00:00:00_00:00:00 (00:00:00:00:00:00)
  Target IP address: 10.3.51.17 (10.3.51.17)
```

# So how does this work in a System z environment with shared (virtualized) OSA adapters?

**10.0.4.0**

**10.0.3.0**

| LPAR1 VIPA 192.168.1.1 OSA 10.0.1.1 | LPAR2 VIPA 192.168.2.1 OSA 10.0.1.2 |
|---|---|

OSA Default Router

EMIF (PR/SM)

**CSA**

Physical MAC: MAC1

**10.0.1.0**

If the destination IP address in a frame that arrives at the OSA adapter doesn't belong to any of the LPARs that share the adapter, that IP packet is given to the LPAR that acts as the Default OSA router. There are separate IPv4 and IPv6 default OSA router specifications.

**OSA Address Table (OAT)**

| IP address | LPAR / Device |
|---|---|
| 192.168.1.1 | LPAR1/Dx |
| 192.168.2.1 | LPAR2/Dy |
| 10.0.1.1 | LPAR1/Dx |
| 10.0.1.2 | LPAR2/Dy |
| Primary | LPAR2/Dy |
| Secondary | LPAR1/Dx |

**LAN frames**

| Dest_MAC | Dest_IPv4 | Who gets it? | Why? |
|---|---|---|---|
| MAC1 | 10.0.1.1 | LPAR1 | Registered |
| MAC1 | 192.168.1.1 | LPAR1 | Registered |
| MAC1 | 10.0.1.2 | LPAR2 | Registered |
| MAC1 | 192.168.2.1 | LPAR2 | Registered |
| MAC1 | 10.0.3.5 | LPAR2 | Default router |
| MAC1 | 10.0.4.6 | LPAR2 | Default router |

- An OSA NIC has a physical MAC address, just like all NICs
- An OSA NIC is often used by multiple LPARs and TCP/IP stacks
  - The NIC is virtualized so it functions as the NIC for multiple TCP/IP stacks, each with their own IP address
- If someone ARPs for 10.0.1.1 in LPAR1, OSA will return MAC1
- If someone ARPs for 10.0.1.2 in LPAR2, OSA will also return MAC1
- So what does OSA then do when a unicast frame arrives with a destination MAC address of MAC1?
  - It peeks into the IP header inside the frame, and consults a table known as the OSA Address Table (OAT) to see which LPAR the IP address inside the frame belongs to
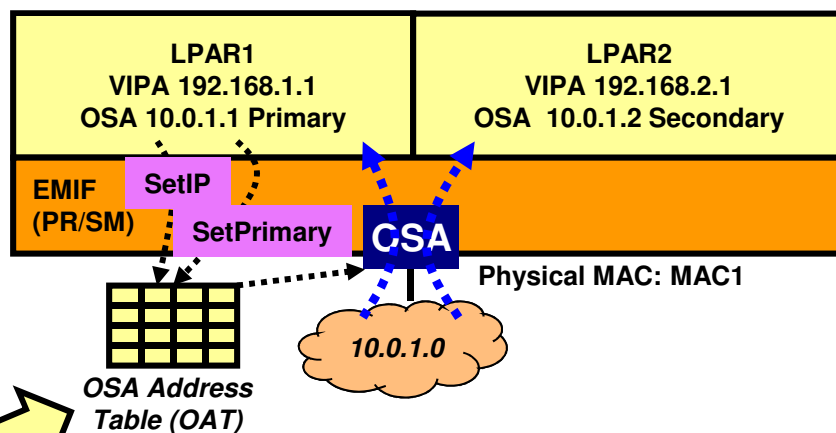
# Setting a few facts straight about OSA and "routing"

- **OSA is not a full-function IP router.**
  - OSA can analyze a destination IP address in a LAN frame to decide which LPAR an incoming IP packet belongs to.
  - OSA soes not participate in dynamic routing updates.
  - OSA does not act as a general IP router
    - Depends on the System z TCP/IP stacks to do that and handle all IP routing issues.

- **Originally, the OSA NIC has a single physical MAC address that is shared among all the LPARs that share the OSA port.**
  - All IP packets to all LPARs can be destined for one and the same MAC address
  - In that case, OSA selects stack based on destination IP address and the OSA Address Table

- **If z/OS acts as an IP router to IP networks behind z/OS, the destination address may be any IP address on those networks behind z/OS.**
  - LPAR designated as default router will receive such packets
  - Can become extremely cumbersome to set up if LPARs that share an OSA port are connected to different back-end networks

- **When a port is defined as an OSE/LCS port, the content of the OAT must be maintained *manually* through OSA/SF interaction.**

- **When a port is defined as a OSD/QDIO port, the content of the OAT is maintained *automatically* by the sharing TCP/IP stacks.**

# QDIO layer-3: less administration, more dynamics – but OAT remains an important element of OSA inbound routing

**QDIO Layer 3 - the OSA adapter is IP-aware and offloads certain functions from the TCP/IP stacks – check-summing, ARP processing, TCP segmentation, etc.**
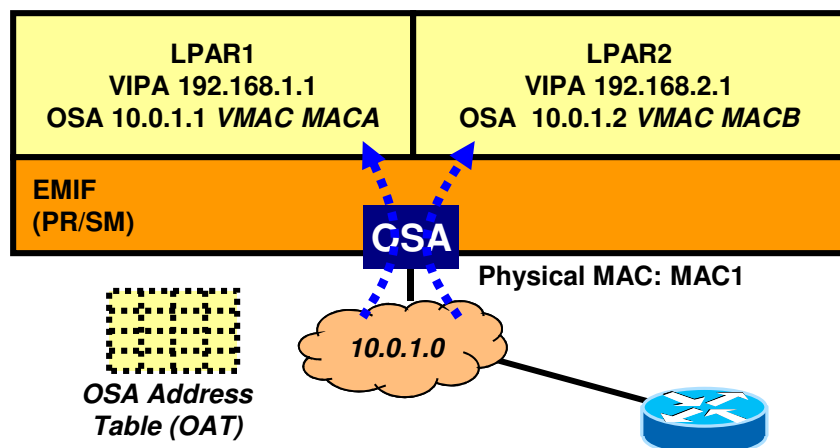
| | LPAR1<br>VIPA 192.168.1.1<br>OSA 10.0.1.1 Primary | LPAR2<br>VIPA 192.168.2.1<br>OSA  10.0.1.2 Secondary |
|---|---|---|

**EMIF (PR/SM)** · **SetIP** · **SetPrimary** · **CSA**

**Physical MAC: MAC1**

*10.0.1.0*

*OSA Address Table (OAT)*

**OAT Updates:**
- **LCS: operator intervention via OSA/SF commands**
- **QDIO: dynamic by software via QDIO control channel**

***OAT is maintained dynamically by TCP/IP***

| IP address | LPAR / Device |
|---|---|
| 192.168.1.1 | LPAR1/Dx |
| 192.168.2.1 | LPAR2/Dy |
| 10.0.1.1 | LPAR1/Dx |
| 10.0.1.2 | LPAR2/Dy |
| Primary | LPAR1/Dx |
| Secondary | LPAR2/Dy |

- QDIO device definitions in the TCP/IP stacks are used to dynamically establish the stack as the OAT default router, secondary router, or non-router.
- Whenever a QDIO device is activated or the TCP/IP home list is modified (through OBEYFILE command processing or through dynamic changes, such as dynamic VIPA takeover), the TCP/IP stack updates the OAT configuration dynamically with the HOME list IP addresses of the stack.
- The OAT includes all (non-LOOPBACK) HOME IP addresses of all the stacks that share the OSA adapter.
- The fact that the OSA microcode is IP address-aware (as it is in this scenario) is the reason for referring to this as QDIO layer 3 processing (layer 3 is generally the networking layer in an OSI model - the IP networking layer when using TCP/IP)

# Wouldn't it be so much easier with a MAC address per z/OS TCP/IP network interface?

| LPAR1<br>VIPA 192.168.1.1<br>OSA 10.0.1.1 *VMAC MACA* | LPAR2<br>VIPA 192.168.2.1<br>OSA  10.0.1.2 *VMAC MACB* |
|---|---|
| **EMIF**<br>**(PR/SM)** | **CSA** |

**YES !!**

**Physical MAC: MAC1**

**10.0.1.0**

***OSA Address***
***Table (OAT)***

**Router's ARP cache**

| IP Address | MAC Address |
|---|---|
| 10.0.1.1 | MACA |
| 10.0.1.2 | MACB |

***The whole mess***
***with PRIROUTER***
***and SECROUTER***
***is gone !!***

*It also removes issues with*
*external load balancers that*
*use MAC-level forwarding.*
*It makes a system z LPAR look*
*like a "normal" TCP/IP node on*
*a LAN.*

- A packet for 192.168.1.1 arrives at the router from the downstream network
  - Router determines the destination is not directly attached to the router
  - Router looks into its routing table and determines next-hop IP address for this destination is 10.0.1.1, which is on a network that is directly attached to the router
  - Router looks into its ARP cache and determines the associated MAC address is MACA
  - Router forwards the IP packet in a LAN frame to MACA

- Frame arrives in OSA
  - OSA determines which LPAR it belongs to based on the virtual MAC address
  - The OAT may optionally still play a role in inbound routing decisions by the OSA adapter:
    - If the destination address (192.168.1.1) were not in the home list of this LPAR (and hence not in the OAT for this LPAR), should OSA still send it up to the LPAR or not?
    - You decide via a configuration option
      - ROUTEALL: send all packets with my VMAC to me
      - ROUTELCL: send only packets to me if they are in my HOME list
  - The OAT remains in use for other functions, such as ARP ownership, etc.

# OSA-Express virtual MAC while operating in QDIO layer-3 mode

- **OSA MAC sharing problems do not exist if each stack has its own MAC**
  - "virtual" MAC
  - To the network, each stack appears to have a dedicated OSA port (NIC)

- **MAC address selection**
  - Coded in the TCP/IP profile
  - Generated and assigned by the OSA adapter

- **All IP addresses for a stack are advertised with the virtual MAC**
  - by OSA using ARP for IPv4
  - by the stack using ND for IPv6

- **All external routers now forward frames to the virtual MAC**
  - OSA will "route" to an LPAR/Stack by virtual MAC instead of IP address
  - All stacks can be "routing" stacks instead of 1 PRIROUTER stack

- **Simplifies configuration greatly**
  - No PRIROUTER/SECROUTER!

- **Supported on System z9, z10, and z196**

- **Prior to z/OS V1R10, each stack may define one VLAN per protocol (IPv4 or IPv6) for each OSA port**
  - One VMAC for the LINK statement
  - One VMAC for the INTERFACE statement

- **z/OS V1R10 allows each stack to define up to 8 VLANS per protocol, each with its own VMAC for each OSA port**
  - A total of 8 IPv4 and 8 IPv6 logical network interfaces per stack
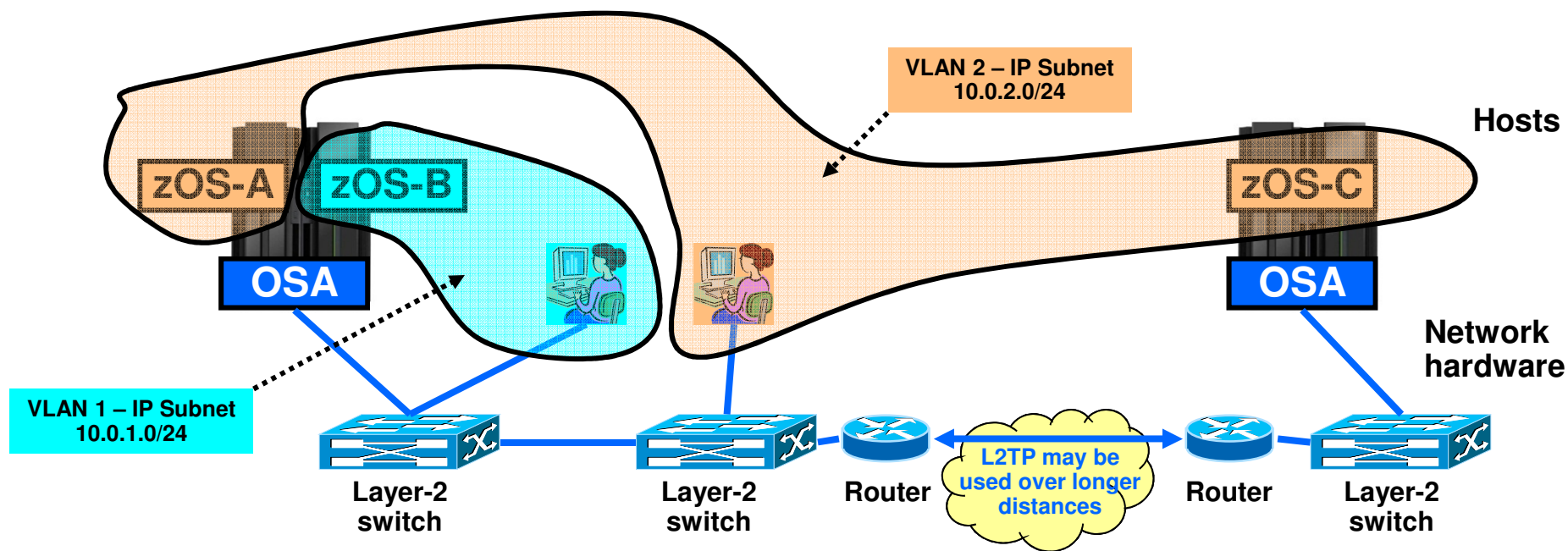
*I see no general reasons to not use VMACs. Use them!!!*
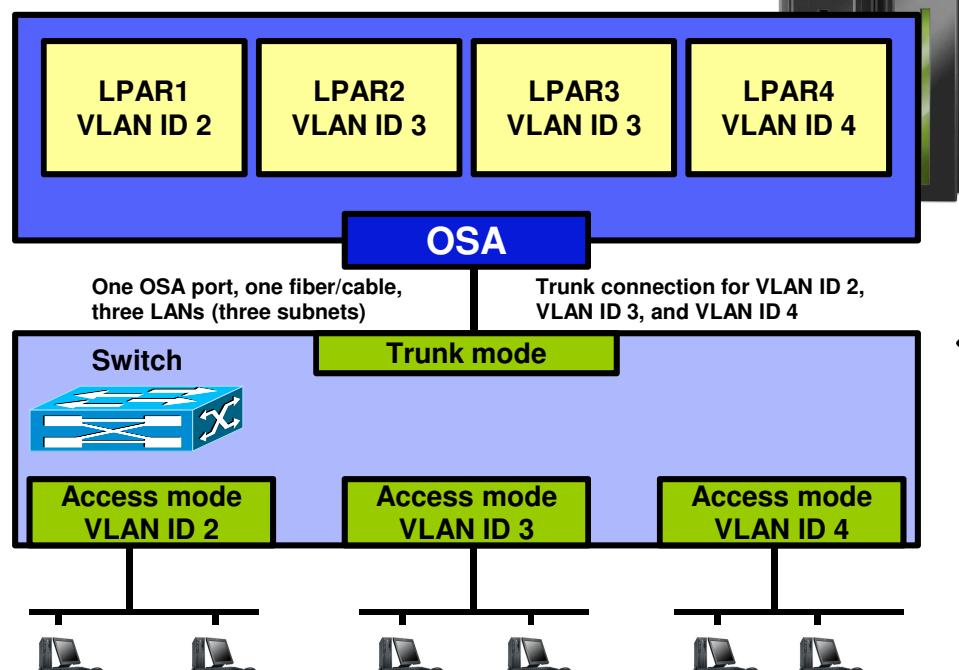
# What is a Virtual LAN (a VLAN)?

**Wikipedia:**

- A **virtual LAN**, commonly known as a **VLAN**, is a *group of hosts* with a common set of requirements that communicate as if they were *attached* to the same *broadcast domain*, regardless of their physical location.

- A VLAN has the same attributes as a physical LAN, but it allows for end stations to be grouped together even if they are not located on the same network switch.

- Network reconfiguration can be done through software instead of physically relocating devices.



VLAN 2 – IP Subnet 10.0.2.0/24

Hosts

zOS-A   zOS-B   zOS-C

OSA   OSA

Network hardware

VLAN 1 – IP Subnet 10.0.1.0/24

Layer-2 switch   Layer-2 switch   Router   L2TP may be used over longer distances   Router   Layer-2 switch

# z/OS and VLANs

## Physical network diagram

| LPAR1 VLAN ID 2 | LPAR2 VLAN ID 3 | LPAR3 VLAN ID 3 | LPAR4 VLAN ID 4 |

**OSA**

One OSA port, one fiber/cable, three LANs (three subnets)

Trunk connection for VLAN ID 2, VLAN ID 3, and VLAN ID 4

**Trunk mode**

**Switch**

| Access mode VLAN ID 2 | Access mode VLAN ID 3 | Access mode VLAN ID 4 |

- **Depending on switch configuration, the switch may interconnect the VLANs using a layer-3 IP router function.**

- **The subnets may belong to different routing domains or OSPF areas:**
  - Test, production, demo

- **The subnets may belong to different security zones:**
  - Intranet, DMZ

## Logical network diagram

LPAR1    LPAR2    LPAR3    LPAR4
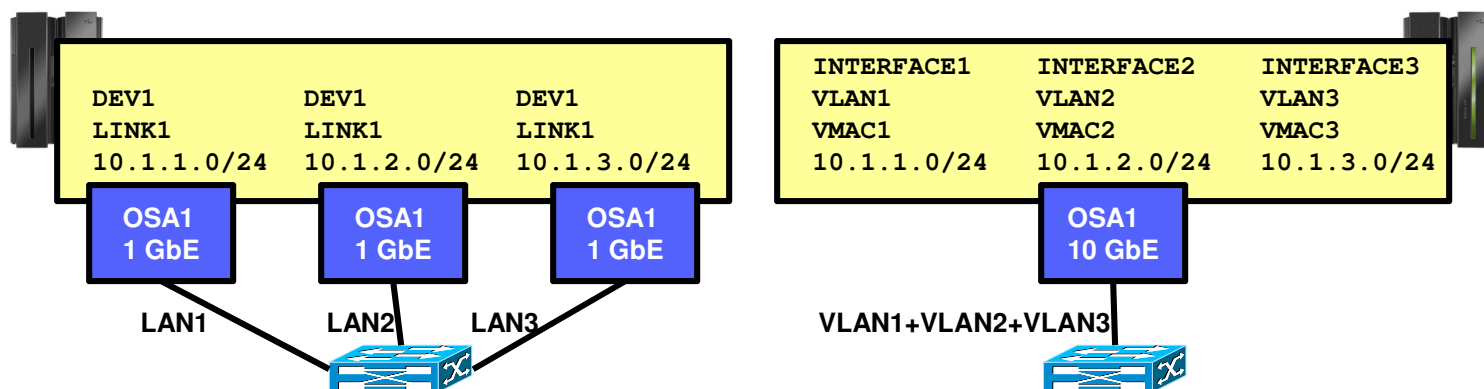
*VLAN ID 2*    *VLAN ID 3*    *VLAN ID 4*
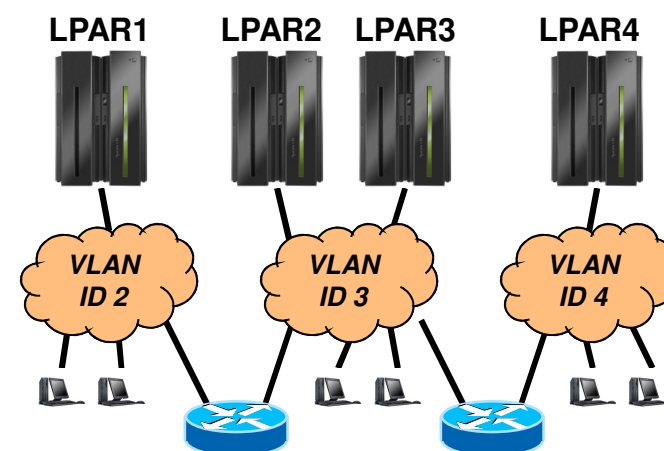
- **Each frame on the trunk mode connection carries a VLAN ID in the IEEE802.3 header that allows the network equipment to clearly identify which virtual LAN each frame belongs to.**
- **On an access mode connection, the switch will transport frames belonging to the configured VLAN ID for that access mode connection only.**

> **VLAN is a LAN media virtualization technology that allows multiple independent IP networks (IP subnets) to share one physical media, such as a cable, an adapter, or a layer-2 switch. Connectivity between VLANs is under control of IP routers.**

# Multiple network interfaces (VLANs) per OSA port per stack per IP protocol version

- As installations consolidate multiple OSA Gigabit Ethernet ports to a smaller number of 10 Gigabit Ethernet ports this limitation has become too restrictive:
  - Not possible to retain existing network interface and IP subnet topology
  - Consolidating multiple LANs to one LAN requires IP renumbering

- z/OS V1R10 added support for eight VLANs per IP protocol per OSA port:
  - Each VLAN on the same OSA port must use unique, non-overlapping IP subnets or prefixes
    - Will be enforced by the TCP/IP stack
  - Each VLAN must be defined using theIPv4-enabled version of the INTERFACE configuration statement
    - IPv4 INTERFACE statement only supports QDIO interfaces
  - Each VLAN must use layer-3 virtual MAC addresses with ROUTEALL, and each VLAN must have a unique MAC address

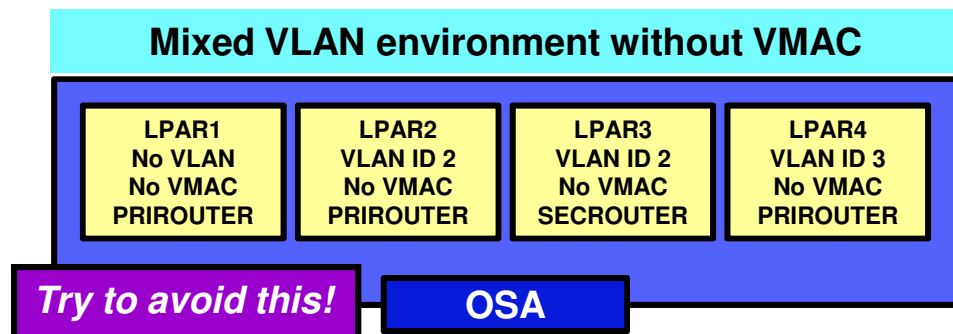- z/OS V1R13 raises the number of VLANs to 32 per IP protocol per OSA port

| DEV1 | DEV1 | DEV1 |
|------|------|------|
| LINK1 | LINK1 | LINK1 |
| 10.1.1.0/24 | 10.1.2.0/24 | 10.1.3.0/24 |

| OSA1 | OSA1 | OSA1 |
|------|------|------|
| 1 GbE | 1 GbE | 1 GbE |

LAN1    LAN2    LAN3

| INTERFACE1 | INTERFACE2 | INTERFACE3 |
|------------|------------|------------|
| VLAN1 | VLAN2 | VLAN3 |
| VMAC1 | VMAC2 | VMAC3 |
| 10.1.1.0/24 | 10.1.2.0/24 | 10.1.3.0/24 |

OSA1
10 GbE

VLAN1+VLAN2+VLAN3

**Consolidate multiple low-capacity LAN interfaces to fewer high-capacity LAN interfaces without network topology impact.**

# A few more words on VLANs with z/OS CS

- z/OS Communications Server supports one VLAN ID per defined network interface
  - This is known as a global VLAN ID
    - Remember: z/OS CS now supports up to eight interfaces per OSA port – each with its own VLAN ID
  - OSA performs inbound routing based on VLAN IDs in the incoming frames and only sends frames to z/OS with the global VLAN ID z/OS has registered

- Linux on System z supports multiple VLAN IDs per interface
  - OSA sends multiple VLAN IDs up to Linux and Linux then de-multiplexes the different virtual LANs

- Some interfaces that share an OSA port may select to not specify a VLAN ID, while others do specify a VLAN ID (a mixed VLAN environment)
  - **Warning**: be very careful - mixed VLAN environments are not recommended due to complexities with OSA default router selection and other functional issues.  Use VLAN IDs on all TCP/IP interfaces that share an OSA port - or not at all.
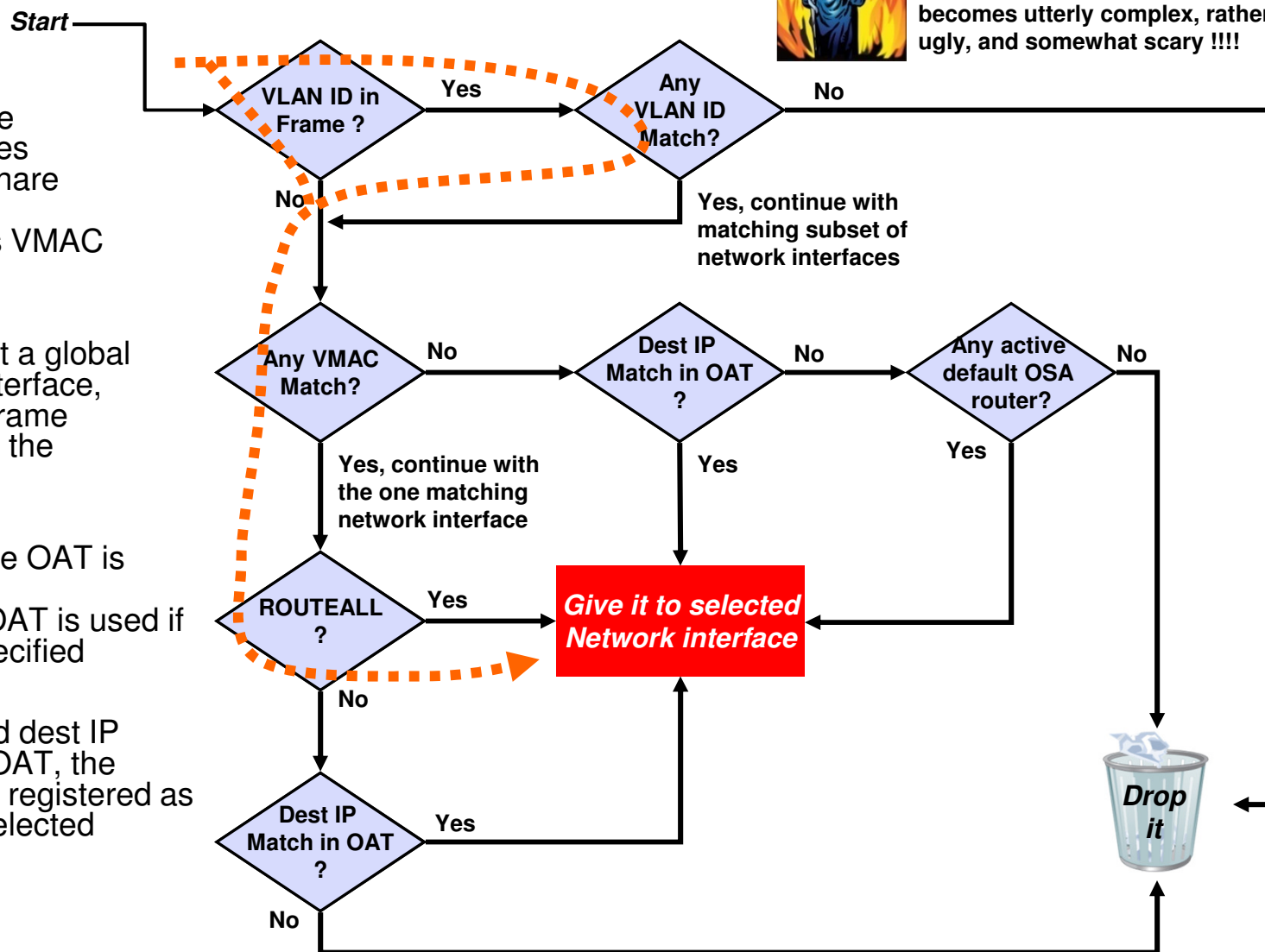
## Mixed VLAN environment without VMAC

| LPAR1<br>No VLAN<br>No VMAC<br>PRIROUTER | LPAR2<br>VLAN ID 2<br>No VMAC<br>PRIROUTER | LPAR3<br>VLAN ID 2<br>No VMAC<br>SECROUTER | LPAR4<br>VLAN ID 3<br>No VMAC<br>PRIROUTER |
|---|---|---|---|

*Try to avoid this!* — **OSA**

- **LPAR1 acts as PriRouter for:**
  - **Untagged frames**
  - **Frames tagged with a null VLAN ID**
  - **Frames tagged with an unregistered (anything but VLAN 2 or 3) VLAN ID**
  - **Frames tagged with a registered VLAN ID, an unknown destination IP address but no VLAN Pri/SecRrouter**
- **LPAR2 acts as PriRouter for frames tagged with VLAN 2 while LPAR3 acts as SecRouter for that same VLAN.**
- **LPAR4 acts as PriRouter for frames tagged with VLAN 3**
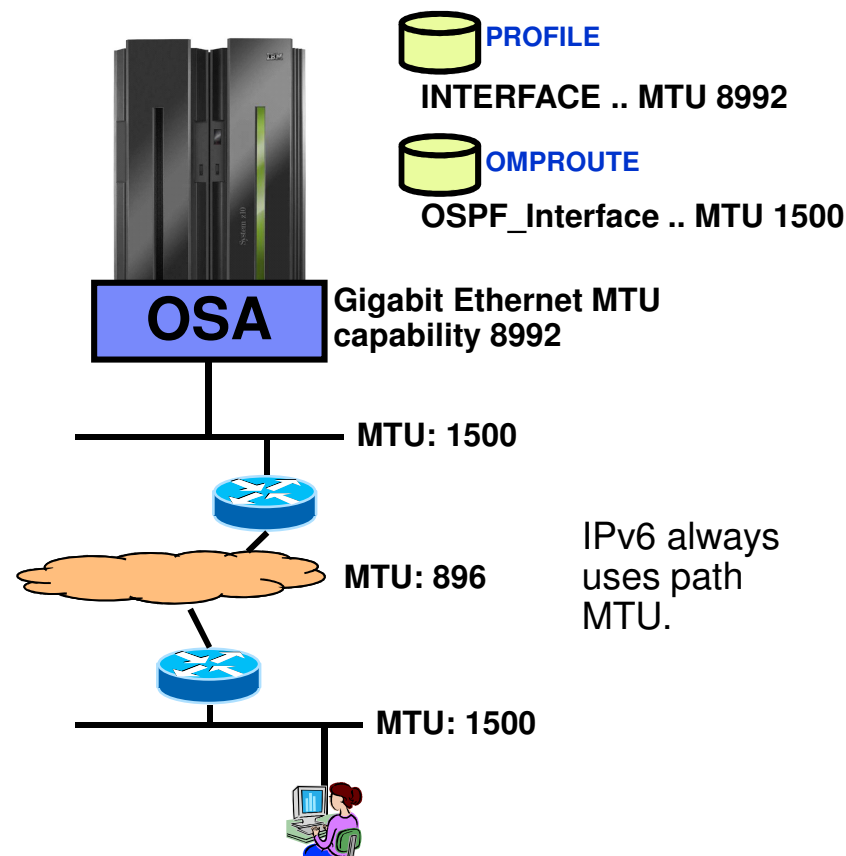
# OSA inbound routing as of early 2010

This is correct if all interfaces that share the OSA port use VLAN IDs or none of them do. If some do and others do not, this becomes utterly complex, rather ugly, and somewhat scary !!!!

- **Dest MAC**
  - VMACs are unique across all interfaces and VLANs that share an OSA port
  - z/OS CS registers VMAC with the OSA port

- **VLAN ID**
  - If z/OS CS has set a global VLAN ID for an interface, OSA verifies the frame VLAN ID matches the global VLAN ID

- **Dest IP address**
  - Without VMAC, the OAT is always used
  - With VMAC, the OAT is used if ROUTELCL is specified

- **Default OSA router**
  - If OAT is used and dest IP address is not in OAT, the interface currently registered as default router is selected

Start

**VLAN ID in Frame ?** — Yes → **Any VLAN ID Match?** — No →

**Any VLAN ID Match?** — Yes, continue with matching subset of network interfaces

**VLAN ID in Frame ?** — No →

**Any VMAC Match?** — No → **Dest IP Match in OAT ?** — No → **Any active default OSA router?** — No → Drop it

**Any VMAC Match?** — Yes, continue with the one matching network interface

**Dest IP Match in OAT ?** — Yes →

**Any active default OSA router?** — Yes →

**ROUTEALL ?** — Yes → **Give it to selected Network interface**

**ROUTEALL ?** — No →

**Dest IP Match in OAT ?** — Yes → Give it to selected Network interface

**Dest IP Match in OAT ?** — No → Drop it

# This elusive Maximum Transmission Unit – what is the size really?

- **Interface MTU**
  - Configured on the INTERFACE statement or learned from the device
    - For IPv4, OSA reports 1492/8992 – even when you use Ethernet II frames
    - For IPv6, OSA reports 1500/9000 – because IPv6 always uses Ethernet II frames
  - Reported as ActMtu: in a DEVLINKS netstat report

- **Configured route MTU**
  - For static routes: the value coded in the BEGINROUTES section for the destination
  - For dynamic routes: the MTU value from the OMPROUTE interface over which the route was learned
  - If no MTU size defined in OMPROUTE, a default of 576 will be used

- **Actual route MTU**
  - The lowest of the interface MTU and the configured route MTU
  - If path MTU is not enabled, this is the MTU that will be used for that route

- **Path MTU**
  - With path MTU enabled, TCP/IP starts out trying to use the actual route MTU and may then lower the MTU to what is discovered
    - With path MTU enabled, all frames are sent with the Do Not Fragment bit
    - Path MTU values are cached, but will timeout after a certain amount of time to accommodate topology changes that allows a higher MTU

**PROFILE**

**INTERFACE .. MTU 8992**

**OMPROUTE**

**OSPF_Interface .. MTU 1500**

**OSA**  Gigabit Ethernet MTU capability 8992

**MTU: 1500**

**MTU: 896**

**MTU: 1500**

IPv6 always uses path MTU.

| MTU type | MTU value |
|----------|-----------|
| Interface MTU | 8992 |
| Configured route MTU | 1500 |
| Actual route MTU | 1500 |
| Path MTU to this destination | 896 |

# Some examples on MTU specifications

- MTU 1500 configured on an IPv4 INTERFACE stmt: (1000Base-T)
  - **CfgMtu: 1500                                    ActMtu: 1500**
  - OSA reports 8992 in this case, but our configured MTU overrides that and determines the actual MTU

- No MTU configured on an IPv4 INTERFACE stmt: (1000Base-T)
  - **CfgMtu: None                                    ActMtu: 8992**
  - We have no configured MTU to override what OSA reports, so the actual MTU ends up being 8992 as reported by OSA for an IPv4 interface

- No MTU configured on an IPv6 INTERFACE stmt: (1000Base-T)
  - **CfgMtu: None                                    ActMtu: 9000**
  - Since this is an IPv6 interface, OSA reports the full jumbo frame size of 9000 bytes as the MTU it supports.  Since we did not override that with a configured MTU, the 9000 ends up being the actual MTU size.

- In my sample setup, I have MTU 1500 coded in the OMPROUTE configuration file, so all actual route MTUs end up being 1500
  - **Default                     9.42.105.65        UGO              0000000002 QDIO4A**
  - **      Metric: 00000001   MTU: 1500**

**Getting the most out of your OSA adapter with z/OS Communications Server**

# Overview of selected recent OSA enhancements

# Inbound OSA/QDIO performance: Dynamic LAN idle timer

- Dynamic LAN idle timer is designed to reduce latency and improve network performance by dynamically adjusting the inbound blocking algorithm.
  - When enabled, the z/OS TCP/IP Stack is designed to adjust the inbound blocking algorithm to best match the application requirements.

- For latency sensitive applications, the blocking algorithm is modified to be "latency sensitive." For streaming (throughput sensitive) applications, the blocking algorithm is adjusted to maximize throughput. In all cases, the z/OS TCP/IP stack dynamically detects the application requirements, making the necessary adjustments to the blocking algorithm.
  - The monitoring of the application and the blocking algorithm adjustments are made in real-time, dynamically adjusting the application's LAN performance.

- System administrators can authorize the z/OS TCP/IP stack to enable a dynamic setting, which was previously a static setting.
  - The z/OS TCP/IP stack is designed to dynamically determine the best setting for the current running application based on system configuration, inbound workload volume, CPU utilization, and traffic patterns.

- For an OSA Express2 or Express3 port in OSD/QDIO mode the z/OS TCP/IP profile **INBPERF** parameter can be specified with one of the following options:
  - **MINCPU** - a static interrupt-timing value, selected to minimize host interrupts without regard to throughput
  - **MINLATENCY** - a static interrupt-timing value, selected to minimize latency
  - **BALANCED** (default) - a static interrupt-timing value, selected to achieve reasonably high throughput and reasonably low CPU
  - **DYNAMIC** - implements the new dynamic LAN idle algorithm ← generally recommended!

© 2011 SHARE and IBM Corporation

# OSA-Express3 dual port support by z/OS CS

- Transparent error handling
  - Prior to this support, if one packet within a QDIO read operation was in error, all packets within that entire read operation were discarded as part of error handling.
  - With the new transparent error handling of OSA-Express3, individually packets with errors are marked as invalid.
  - The z/OS Communications Server later discards the packets which are marked as invalid.
  - This new efficiency reduces the number of unnecessary retransmissions which improves overall I/O efficiency.
  - To support the new "transparent error handling" function:
    - Available with z/OS V1R10
    - z/OS V1R9: PTF UA37872
    - z/OS V1R8: PTF UA37871

- Without specific z/OS CS support, only the two port 0 ports on an OSA-E3 adapter can be used
  - One port 0 per CHPID

- Support for both port 0 and port 1 (the ability to configure PORTNUM in the TRLE)
  - Available with z/OS V1R10
  - z/OS V1R9: APAR OA25548 (PTF: UA42717)
  - z/OS V1R8: APAR OA25548 (PTF: UA42716)

### OSA-Express3



OSA LED
A0

CHPID X
Ports 0 and 1

OSA LED
A1

SC/LC Duplex Connectors

J00
J01
FRU
J01
J00

OSA LED
A1

CHPID X+1
Ports 0 and 1

OSA LED
A0

SC/LC Duplex Connectors
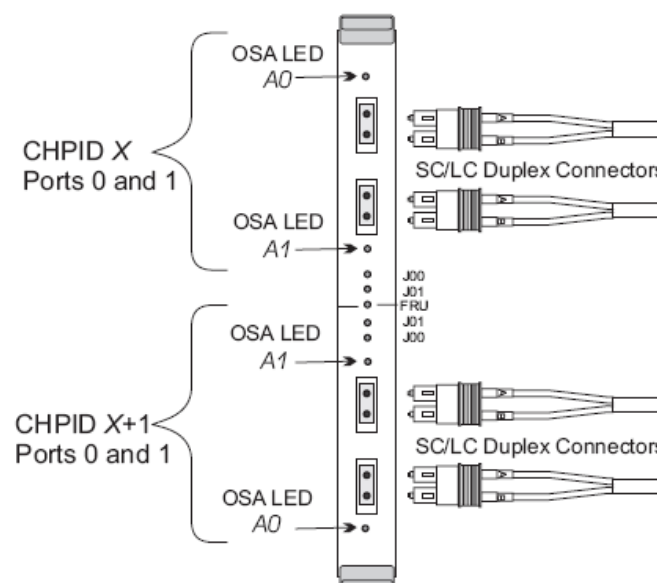
**See appendix for sample IOCP, TRLE, and INTERFACE definitions**

© 2011 SHARE and IBM Corporation

# OSA-Express Network Traffic Analyzer (OSAENTA) function

- Diagnosing OSA-Express QDIO problems can be somewhat difficult
  - TCP/IP stack (CTRACE and/or packet trace)
  - VTAM (VIT)
  - OSA (hardware trace)
  - Network (sniffer trace)

- Often not clear where the problem is and which trace(s) to collect
  - Offloaded functions and shared OSAs can complicate the diagnosis
    - ARP frames
    - Segmentation offload
    - Etc.

- Supported with OSA-Express2 and with OSA-Express3

- Allows z/OS Communications Server to collect Ethernet data frames from the OSA adapter
  - Not a sniffer trace (but similar in some aspects)
  - No promiscuous mode
  - Minimizes the need to collect and coordinate multiple traces for diagnosis
  - Minimizes the need for traces from the OSA Hardware Management Console (HMC)
  - Formatted with the standard z/OS Communications Server packet trace formatter
    - Including the ability to convert to Sniffer/Wireshark format

**Which trace should I use?**

*Sniffer*

*HMC Hardware Trace*

OSA-E2

LAN

z/OS

*VIT* | VTAM

z/VM, Linux, z/VSE, z/TPF

TCPIPA | TCPIPB | TCPIPC | TCPIPD

*CTRACE*

*Packet Trace*
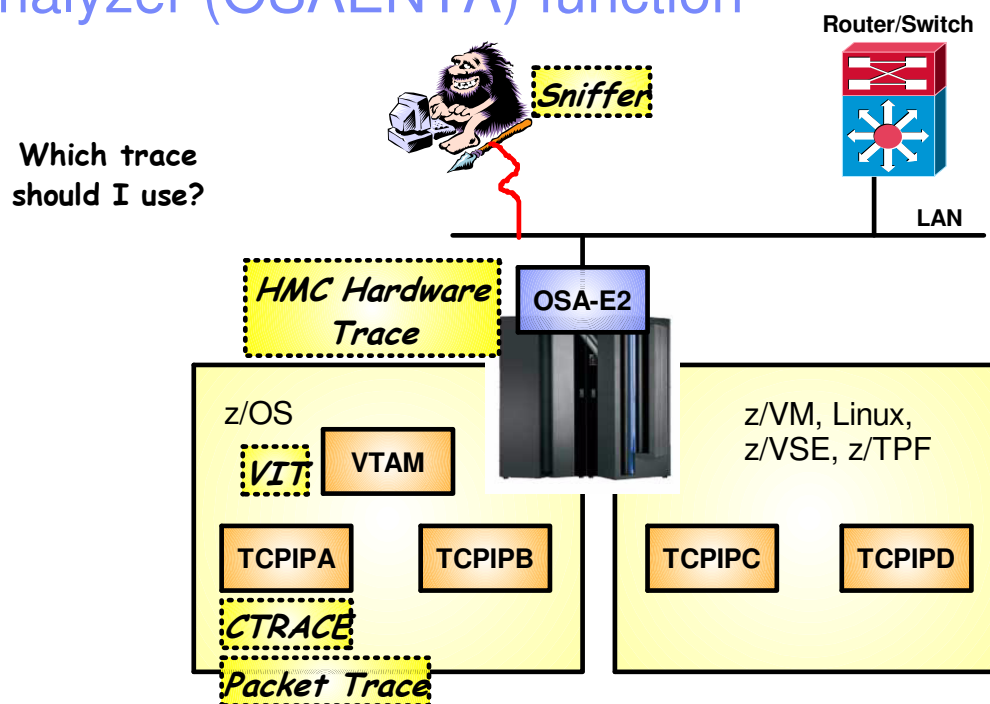
# QDIO priority queue selection based upon WLM importance level

*Basic principle is that if QoS policies are active, they will determine which priority queue to use.*

**Application**

**WLM service class importance level (SCIL)**

**Use workload objective of application address spaces to select QDIO outbound priority queue**

**Optional TOS**

**QoS Policy**

**TCPIP Profile**

**WLMPRIORITYQ**

```
SetSubnetPrioTosMask
{
SubnetTosMask 11100000
PriorityTosMapping 1 11100000
PriorityTosMapping 1 11000000
PriorityTosMapping 1 10100000
PriorityTosMapping 1 10000000
PriorityTosMapping 2 01100000
PriorityTosMapping 2 01000000
PriorityTosMapping 3 00100000
PriorityTosMapping 4 00000000
}
```

**Q1**

**Q2**

**Q3**

**Q4    OSA**
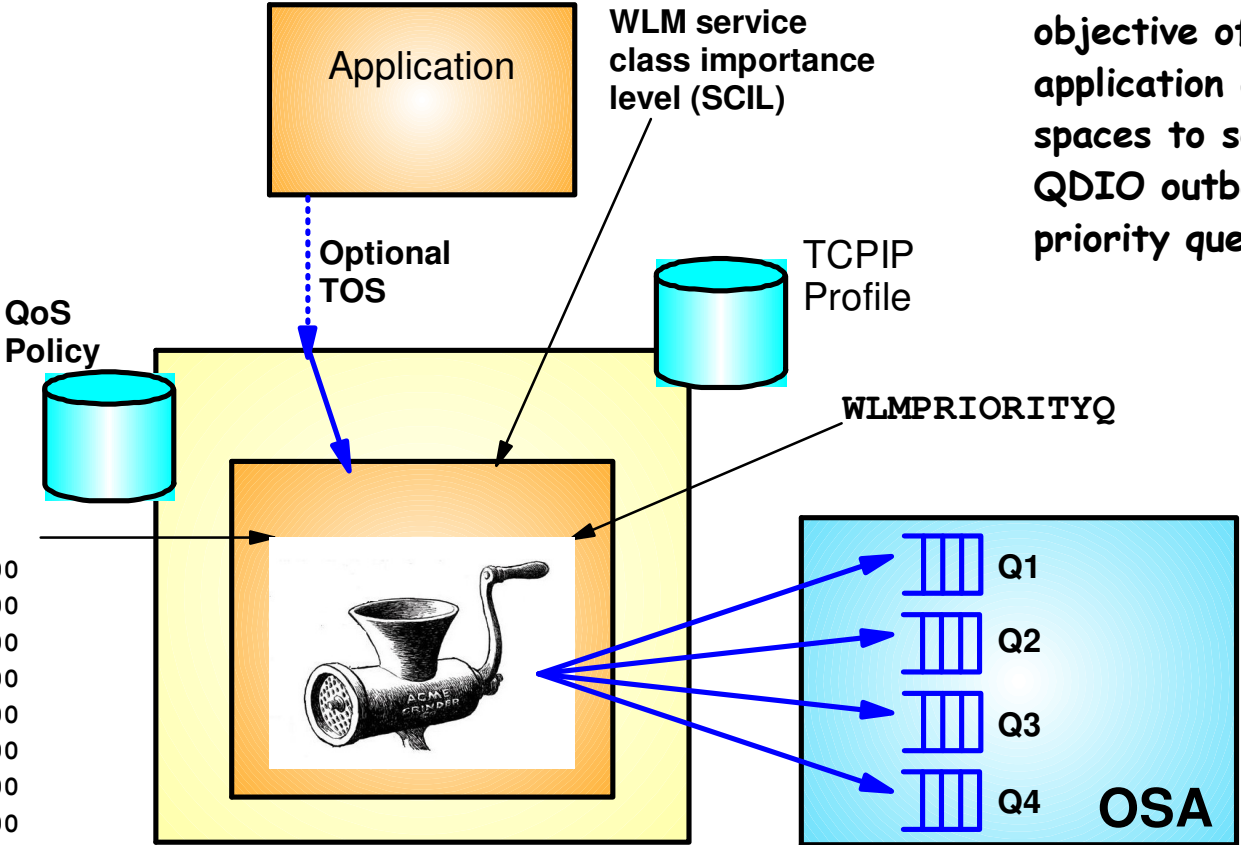
**An easy way to take advantage of QDIO's four outbound priority queues!**

# OSA Express (QDIO) WLM Outbound Priority Queuing

**Transactions / Second**

**Latency**
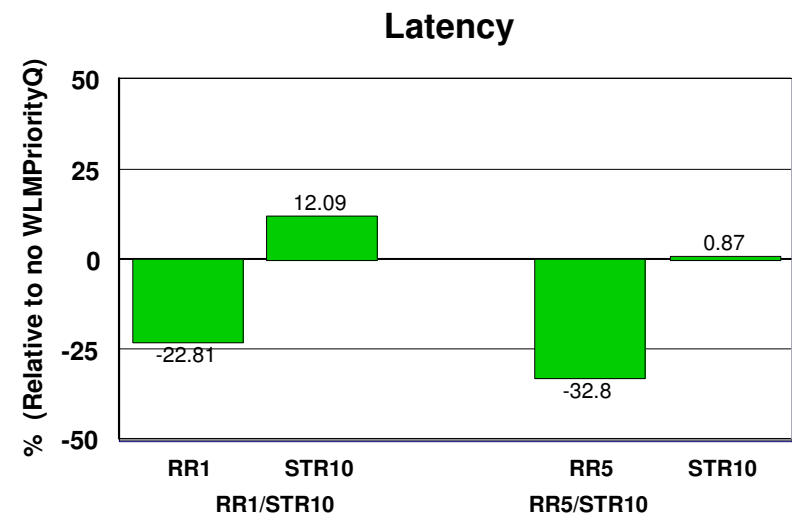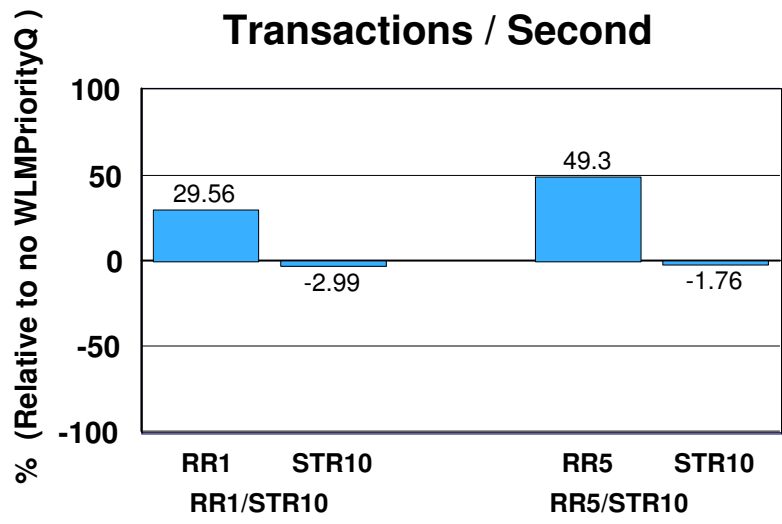


- **Request-Response and Streaming mixed workload**
- **RR1/STR10:  1 RR session, 100 / 800 and 10 STR sessions, 1 / 20 MB**
- **RR5/STR10:  5 RR sessions, 100 / 800 and 10 STR sessions, 1 / 20 MB**
- **WLMPRIORITYQ assigned importance level 2 to interactive workloads and level 3 to streaming workloads**
- **The z/OS Workload Manager (WLM) system administrator assigns each job a WLM service class**
- **Hardware:  z10 using OSA-E2 (1 GbE)**
- **Software:  z/OS V1R11**
- **z/OS V1R11 with WLM I/O Priority provides 29.56 to 49.3% higher throughput for interactive workloads compared to V1R11 without WLM I/O Priority (Avg= 39.43% higher).**
- **z/OS V1R11 with WLM I/O Priority provides 22.81 to 32.8% lower latency compared to V1R11 without WLM I/O Priority (Avg= 27.80% lower).**

*Note: The performance measurements discussed in this presentation are preliminary z/OS V1R12 Communications Server numbers and were collected using a dedicated system environment.  The results obtained in other configurations or operating system environments may vary.*

# OSA-Express optimized latency mode (OLM)

- OSA-Express3 has significantly better latency characteristics than OSA-Express2

- The z/OS software and OSA microcode can further reduce latency:
  - If z/OS Communications Server knows that latency is the most critical factor
  - If z/OS Communications Server knows that the traffic pattern is not streaming bulk data

- Inbound
  - OSA-Express signals host if data is "on its way" ("Early Interrupt")
  - Host looks more frequently for data from OSA-Express

- Outbound
  - OSA-Express does not wait for SIGA to look for outbound data ("SIGA reduction")

- Enabled via PTFs for z/OS V1R11
  - PK90205 (PTF UK49041) and OA29634 (UA49172).

© 2011 SHARE and IBM Corporation
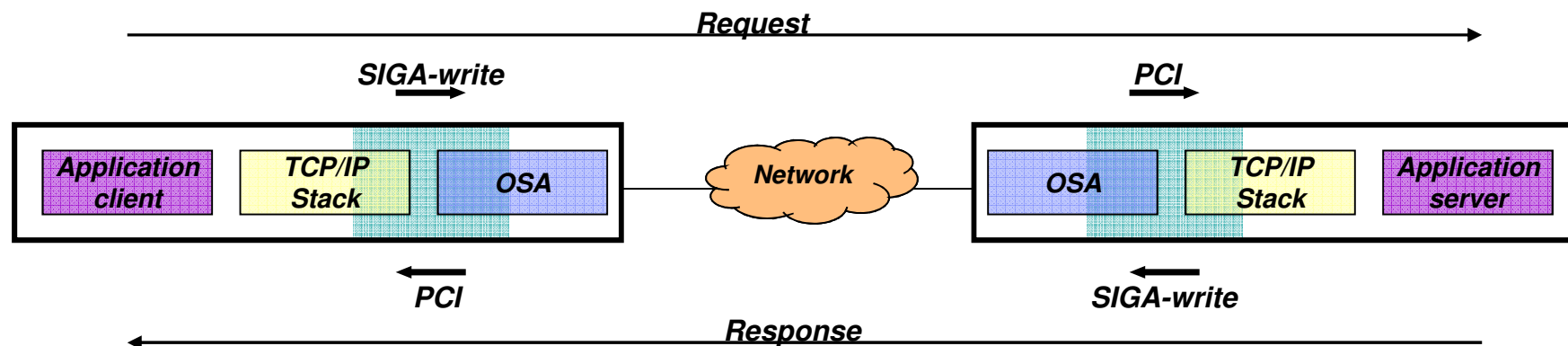
# Performance indications of OLM for interactive workload – INPERF=DYNAMIC vs. INBPERF=DYNAMIC + OLM
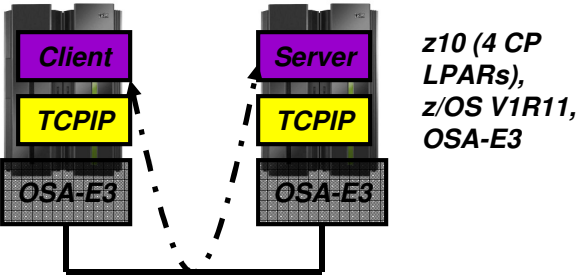
**z10 (4 CP LPARs), z/OS V1R11, OSA-E3**

- **Client and Server**
  - Has close to no application logic

- **RR1**
  - 1 session
  - 1 byte in 1 byte out

- **RR20**
  - 20 sessions
  - 128 bytes in, 1024 bytes out

- **RR40**
  - 40 sessions
  - 128 bytes in, 1024 bytes out

- **RR80**
  - 80 sessions
  - 128 bytes in, 1024 bytes out

- **RR20/60**
  - 80 sessions
  - Mix of 100/128 bytes in and 800/1024 out
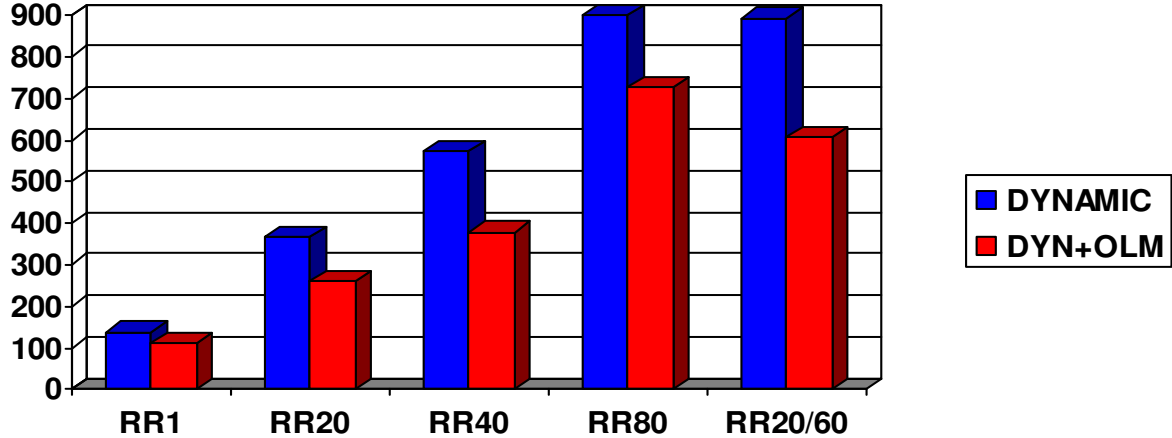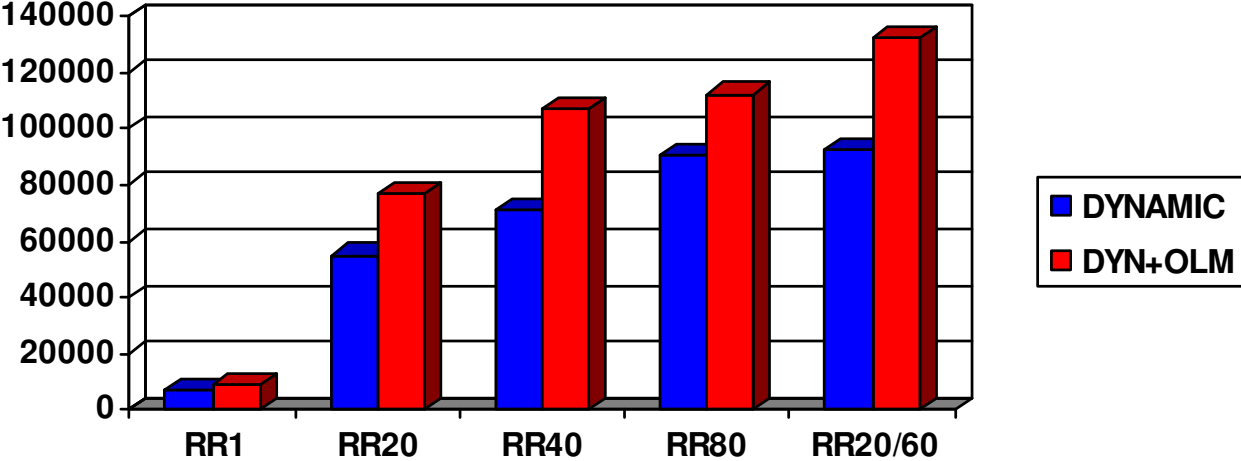
*Note: The performance measurements discussed in this presentation are preliminary z/OS V1R11 Communications Server numbers and were collected using a dedicated system environment. The results obtained in other configurations or operating system environments may vary.*

# OSA interface isolation

- New function added to the OSA adapter
  - z/OS Communications Server adds support for this new function in z/OS V1R11


- Allow customers to disable shared OSA local routing functions
  - ISOLATE/NOISOLATE option on QDIO network interface definition


- OSA local routing can in some scenarios be seen as a security exposure


- Depends on OSA MCL update

Be careful using ISOLATE if you use OSPF and share a subnet between stacks that share an OSA port.
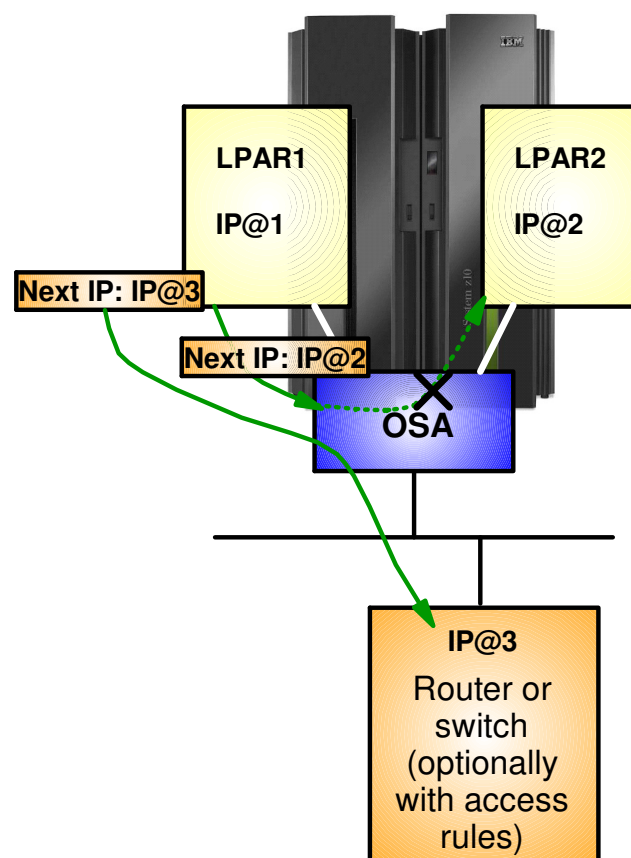
If you enable ISOLATE, packets with a nexthop IP address of another stack that share the OSA port, will be discarded.

LPAR1

IP@1

LPAR2

IP@2

Next IP: IP@3

Next IP: IP@2

OSA

IP@3
Router or switch (optionally with access rules)

# OSA multiple inbound queue support: improved bulk transfer and Sysplex Distributor connection routing performance

- Allow inbound QDIO traffic separation by supporting multiple read queues
  - "Register" with OSA which traffic goes to which queue
  - OSA-Express Data Router function routes to the correct queue

- Each input queue can be serviced by a separate process
  - Primary input queue for general traffic
  - One or more ancillary input queues (AIQs) for specific traffic types

- Supported traffic types
  - Bulk data traffic queue
    - Serviced from a single process - eliminates the out of order delivery issue
  - Sysplex distributor traffic queue
    - SD traffic efficiently accelerated or presented to target application
  - All other traffic not backed up behind bulk data or SD traffic

- Dynamic LAN idle timer updated per queue

- Early performance data (**Note**: your mileage will vary)
  - Request/Response transaction rate improvements (up to 55%)
  - Streaming workload throughput improvements (up to 40%)
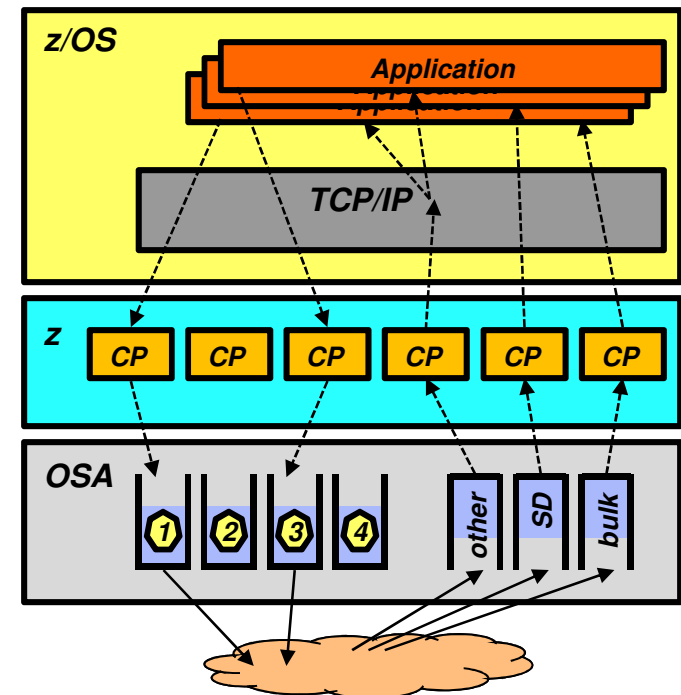
*TCP/IP defines and assigns traffic to queues dynamically based on local IP address and port*

*Bulk traffic*
- *Application sets send or receive buffer to at least 180K*
- *Registered per connection (5-tuple)*

*SD traffic*
- *Based on active VIPADISTRIBUTE definitions*
- *Registered on DVIPA address*

© 2011 SHARE and IBM Corporation

# Operator command to query and display OSA information

- OSA/SF has been used for years to configure OSA and display the configuration. OSA/SF has played a more central role for OSE devices (pre-QDIO) than for today's OSD devices (QDIO).

- OSD devices exclusively use IPA signals exchanged with the host to enable and configure features and register IP addresses to OSA.

- However, there has so far been no mechanism to display the information directly from OSA without OSA/SF.

- z/OS V1R12 implements a new D TCPIP,,OSAINFO command for use with OSA Express3:
  - Base OSA information
  - OSA address table information
  - Information related to the new multiple inbound queues
  - Etc.

```
D TCPIP,,OSAINFO,INTFN=V6O3ETHG0,MAX=100


EZZ0053I COMMAND DISPLAY TCPIP,,OSAINFO COMPLETED SUCCESSFULLY
EZD0031I TCP/IP CS V1R12   TCPIP Name: TCPSVT      15:39:52
Display OSAINFO results for Interface: V6O3ETHG0
PortName: O3ETHG0P  PortNum: 00  DevAddr: 2D64    RealAddr: 0004
PCHID: 0270          CHPID: D6    CHPID Type: OSD  OSA code level: 5D76
Gen: OSA-E3          Active speed/mode: 10 gigabit full duplex
Media: Singlemode Fiber          Jumbo frames: Yes  Isolate: No
PhysicalMACAddr: 001A643B887C  LocallyCfgMACAddr: 000000000000
Queues defined Out: 4  In: 3    Ancillary queues in use: 2
Connection Mode: Layer 3         IPv4: No    IPv6: Yes
SAPSup: 00010293
…
```

# OSA Express segmentation offload (large send) - update

FYI

- For 2097 and 2098 systems (z10)
  - For OSA Express 2 customers required code levels are:
    - Driver-73:  Segmentation Offload Not currently supported
      - or
    - Driver-76  EC Level N10953  (HYNET3X) MCL001
  - For OSA Express 3 customers
    - Driver-73: EC Level F85897 (HYNETST)  MCL010
      - or
    - Driver-76:EC Level  N10959 ( HYNETST) MCL001


- For 2094 and 2096 systems (z9)
  - OSA Express2 Cards
    - Driver-67  EC Level G40946 (HYNET3X) MCL007


- For 2084 and 2086 systems (z990 and z890)
  - OSA Express2 Cards
    - Driver-55  EC Level J13476 (HYNET3X) MCL023

- z/OS V1R9 Communications Server
  - APAR PK47376 - PTF UK26977
  - APAR PK56723 - PTF UK32713
  - APAR PK64756 - PTF UK37435

- z/OS V1R10 Communications Server
  - APAR PK64756 - PTF UK37433

- z/OS V1R11 Communications Server
  - No additional PTFs

**CAUTION**
**TRIPPING HAZARD**

**Proceed with caution !**

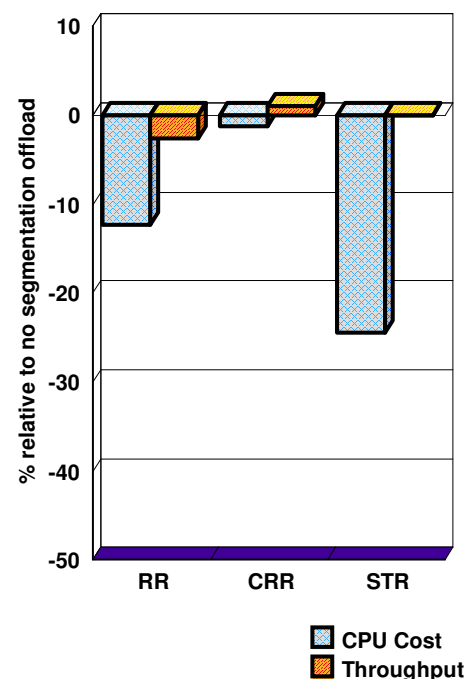# z/OS V1R10 segmentation offload performance measurements on a z10
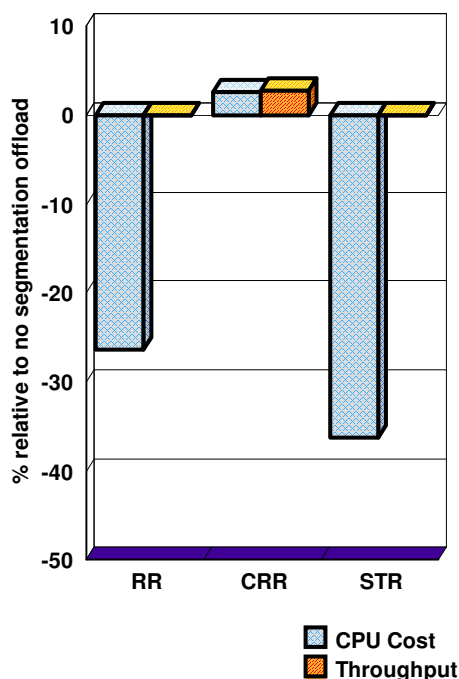
**OSA Express3 1Gb**

**OSA Express3 10Gb**
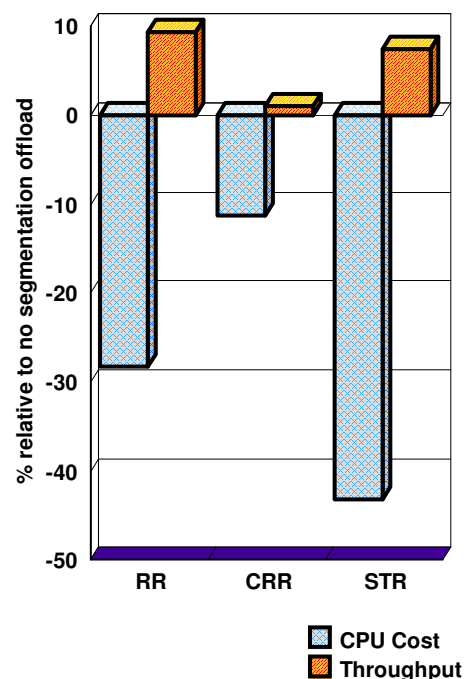
**OSA Express2 1Gb**

*Segmentation offload is generally considered safe to enable at this point in time. Please always check latest PSP buckets for OSA driver levels.*

Send buffer size: 180K for streaming workload

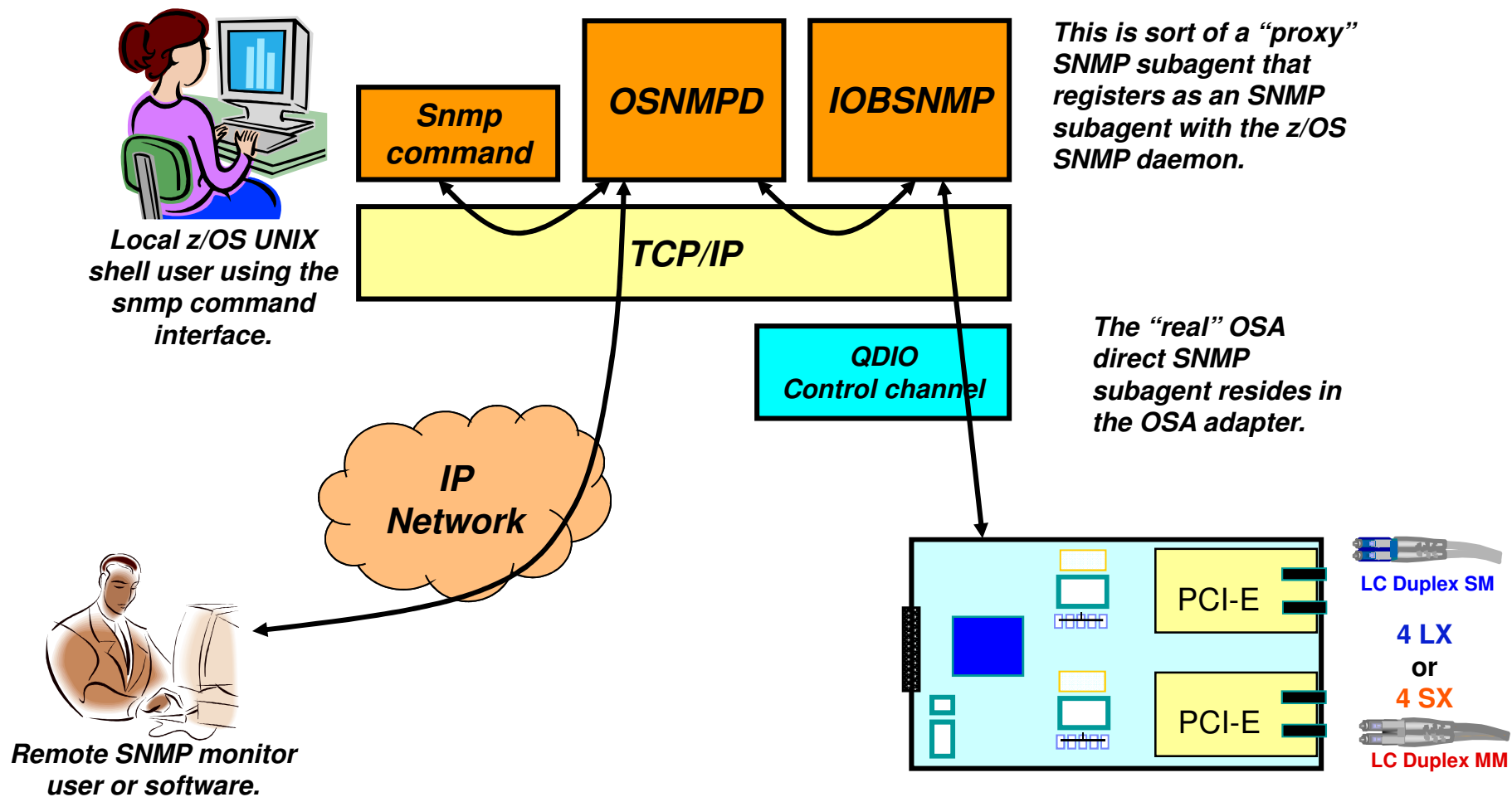**Segmentation offload may significantly reduce CPU cycles when sending bulk data from z/OS**

© 2011 SHARE and IBM Corporation

# Getting the most out of your OSA adapter with z/OS Communications Server

# Appendix A:
# Monitoring OSA using SNMP

# OSA Direct SNMP component overview

**Snmp command**

**OSNMPD**

**IOBSNMP**

*This is sort of a "proxy" SNMP subagent that registers as an SNMP subagent with the z/OS SNMP daemon.*

*Local z/OS UNIX shell user using the snmp command interface.*

**TCP/IP**

**QDIO Control channel**

*The "real" OSA direct SNMP subagent resides in the OSA adapter.*

**IP Network**

PCI-E

PCI-E

**LC Duplex SM**

**4 LX**
**or**
**4 SX**

**LC Duplex MM**

*Remote SNMP monitor user or software.*

# OSA Direct SNMP MIB definitions

- Management Information Base (MIB) variables for the IBM OSA-Express Direct SNMP solution is described:
  - MIB ASN.1 description (and documentation): osaexp3.mib
  - MIB Object ID (OID) to MIB variable name mapping: osaexp3.mibs.data

- These files need to be downloaded from the Web:
  - Point your browser to IBM Resourcelink and log in:
    https://www.ibm.com/servers/resourcelink/svc03100.nsf?Opendatabase
    - a registration is required to use IBM resourcelink
  - Select "Library" on the left side
  - Then look to the right side under "Library short cuts" and select "Open System Adapter (OSA) library"
  - At this point you can select the "OSA-Express Direct SNMP MIB module"
  - You might consider signing up to monitor this site to receive notification when changes occur
  - The osaexp3.mibs.data file need to be installed into /etc/mibs.data or appended to an existing /etc/mibs.data file in order to use the z/OS snmp command with names instead of the (cryptic) OID

# SNMP files to download from IBM Resourcelink

# How to enable the OSA Direct SNMP support

- Configure and start the z/OS CS SNMP daemon, if not already done so
  – See the z/OS CS documentation for doing this

- Download the relevant OSA mibs.data file and append it to your /etc/mibs.data file

- Configure and start the IOBSNMP address space
  – Sample JCL in hlq.SEZAINST(IOBSNMP)

```
//IOBSNMP   PROC P='-s TCPCS'
//IOBSNMP   EXEC PGM=IOBSNMP,TIME=1440,REGION=4096K,DYNAMNBR=5,
//          PARM='&P.'
//SYSPRINT DD  SYSOUT=*
//SYSUDUMP DD  SYSOUT=*
//*
//* The options available for the PARMS referenced by P are
//* -d level   - turn on specified debugging
//*              0 - no tracing
//*              1 - minimal tracing
//*              2 - maximum tracing
//*             >2 - maximum tracing plus SNMP traces
//* -c community  - use specified community name
//* -p port number - use specified port number
//* -s stack      - send request to specified stack
//*
//* Defaults:  -d 0 -c public  -p 161  -s Default_stack
```

**This is the SNMPv1 community name (password). Default is 'public'. Can be set in your SNMP daemon PWSRC file.**

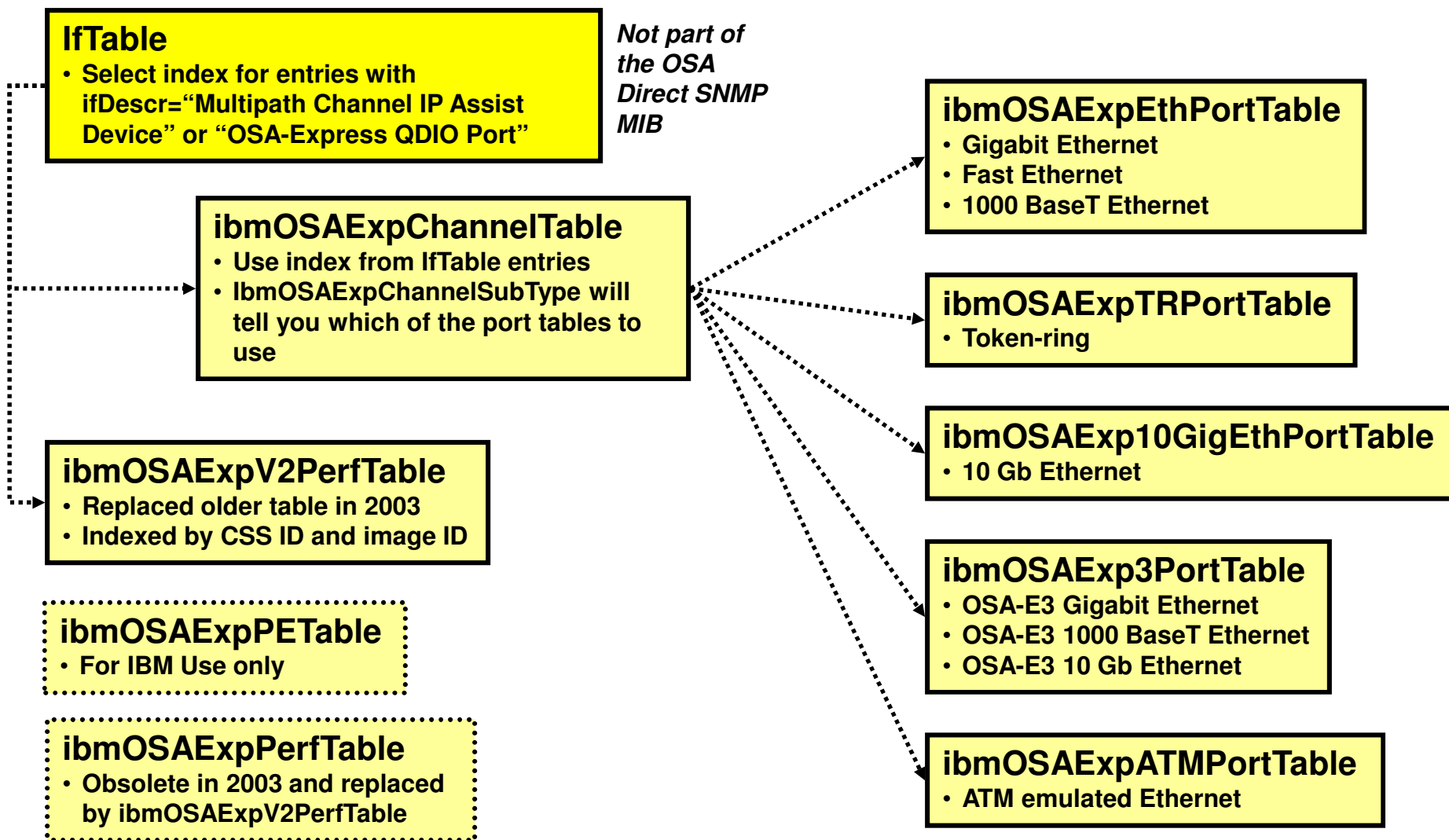**This is the SNMP daemon UDP port number to send SNMP requests to. Default is 161.**

```
OSNMPD   00000022 UDP
  Local Socket:   ::..161
  Foreign Socket: *..*
```

**The name of the TCP/IP stack address space to use for communication with the SNMP daemon and the OSA adapter.**

# OSA Direct SNMP MIB structure

**IfTable**
- Select index for entries with ifDescr="Multipath Channel IP Assist Device" or "OSA-Express QDIO Port"

*Not part of the OSA Direct SNMP MIB*

**ibmOSAExpChannelTable**
- Use index from IfTable entries
- IbmOSAExpChannelSubType will tell you which of the port tables to use

**ibmOSAExpV2PerfTable**
- Replaced older table in 2003
- Indexed by CSS ID and image ID

**ibmOSAExpPETable**
- For IBM Use only

**ibmOSAExpPerfTable**
- Obsolete in 2003 and replaced by ibmOSAExpV2PerfTable

**ibmOSAExpEthPortTable**
- Gigabit Ethernet
- Fast Ethernet
- 1000 BaseT Ethernet

**ibmOSAExpTRPortTable**
- Token-ring

**ibmOSAExp10GigEthPortTable**
- 10 Gb Ethernet

**ibmOSAExp3PortTable**
- OSA-E3 Gigabit Ethernet
- OSA-E3 1000 BaseT Ethernet
- OSA-E3 10 Gb Ethernet

**ibmOSAExpATMPortTable**
- ATM emulated Ethernet

# Navigating to port-specific information

**1** **snmp –v walk ifTable**

```
. . . . .
ifIndex.19 = 19
ifIndex.20 = 20
ifDescr.1 = Loopback Device
ifDescr.2 = Loopback
ifDescr.3 = Loopback IPv6
ifDescr.5 = Virtual IP Address IPv6
ifDescr.6 = Virtual IP Address Device
ifDescr.7 = Virtual IP Address Link
ifDescr.8 = Multipath Channel Point-to-Point Device
ifDescr.9 = Multipath Channel Point-to-Point
ifDescr.10 = LCS Device
ifDescr.11 = Token Ring
ifDescr.12 = Multipath Channel IP Assist Device
ifDescr.13 = IP Assist QDIO Ethernet
ifDescr.14 = IP Assist QDIO Ethernet IPv6
ifDescr.17 = Virtual IP Address Device
ifDescr.18 = Virtual IP Address Link
ifDescr.19 = Virtual IP Address Device
ifDescr.20 = Virtual IP Address Link
ifType.1 = 53
ifType.2 = 24
. . . . .
```

**3** **snmp –v walk ibmOSAExpChannelTable**

```
ibmOSAExpChannelNumber.12 = '00d0'h
ibmOSAExpChannelType.12 = 17
ibmOSAExpChannelHdwLevel.12 = 5
ibmOSAExpChannelSubType.12 = 177
ibmOSAExpChannelShared.12 = 1
ibmOSAExpChannelNodeDesc.12 =
'200006d0f0f0f1f7f3f0f0f0f5c9c2d4f0f2f0f0f0f
0f0f0f0c2c5f1f5c5d000'h
ibmOSAExpChannelProcCodeLevel.12 = '0751'h
ibmOSAExpChannelPCIBusUtil1Min.12 = 0
ibmOSAExpChannelProcUtil1Min.12 = 0
ibmOSAExpChannelPCIBusUtil5Min.12 = 0
ibmOSAExpChannelProcUtil5Min.12 = 0
ibmOSAExpChannelPCIBusUtilHour.12 = 0
ibmOSAExpChannelProcUtilHour.12 = 0
```

*The PCI Bus and Processor utilization data are for the entire CHPID; not just for this LPAR.*

*The ibmOSAExpChannelSubType value of 177 tells us to use the OSA-Express3 port table (ibmOSAExp3PortTable)*

**2** **snmp –v get ifIndex.12**

```
ifIndex.12 = 12
```

**Interface index 12 from the TCP/IP stack is reused to select entries in the ibmOSAExpChannelTable and later in the ibmOSAExp3PortTable.**

# OSA-E3 port-specific information

**snmp –v walk ibmOSAExp3PortTable**

```
ibmOsaExp3PortNumber.12 = 0
ibmOsaExp3PortType.12 = 177
ibmOsaExp3LanTrafficState.12 = 4
ibmOsaExp3ServiceMode.12 = 0
ibmOsaExp3DisabledStatus.12 = '0000'h
ibmOsaExp3ConfigName.12 = IBM Default Configuration File
ibmOsaExp3ConfigSpeedMode.12 = 0
ibmOsaExp3ActiveSpeedMode.12 = 6
ibmOsaExp3MacAddrActive.12 = '00145e74e0f0'h
ibmOsaExp3MacAddrBurntIn.12 = '00145e74e0f0'h
ibmOsaExp3PortName.12 = OSAQDIO4
ibmOsaExp3TotalPacketsXmit.12 = 1148636
ibmOsaExp3TotalPacketsRecv.12 = 1073478
ibmOsaExp3GoodPacketsXmit.12 = 1148636
ibmOsaExp3GoodPacketsRecv.12 = 1073478
ibmOsaExp3Packet64Xmit.12 = 165369
ibmOsaExp3Packet65to127Xmit.12 = 144365
ibmOsaExp3Packet128to255Xmit.12 = 51352
ibmOsaExp3Packet256to511Xmit.12 = 8292
. . . . . .
```

*These traffic numbers represent the aggregate workload of all the LPARs and TCP/IP stacks that share this OSA port.*

*A single TCP/IP stack's traffic numbers can be found in a netstat devlinks report.*

# What can you do with all this information?

You can obviously buy a MIB browser to look at it, but you need something that is intelligent enough to translate the many enumerations into meaningful text.

There are intelligent software solutions that can look at this data for you.

You can also, relatively easily write some simple REXX logic to go and pull it up and print it out the way you want it.

This is a small example of such an approach.

```
OSA SNMP data for: OSAQDIO4 on channel: 00d0 w. MAC address: 00145e74e0f0

Channel      - Number .............. 00d0
=======      - Type ................ OSD
             - Subtype ............. OSA-E3 1000BaseT Ethernet
             - Hardware level ...... OSA-Express3 4.00
             - Share mode .......... Shared channel
Processor    - Code level .......... 7.51
=========    - Util last minute ....   0.00%
             - Util last 5 minutes .   0.00%
             - Util last hour ......   0.00%
PCI bus      - Util last minute ....   3.00%
=======      - Util last 5 minutes .   3.00%
             - Util last hour ......   3.00%
Port         - Port number ......... 0
====         - Port type ........... OSA-E3 1000BaseT Ethernet
             - Traffic state ....... Enabled
             - Service mode ........ Not in service mode
             - Configuration name .. IBM DEFAULT CONFIGURATION FILE
             - Configured speed .... Auto negotiate
             - Active speed ........ 1000 Mb Full Duplex
             - Burned-in MAC ....... 00145e74e0f0
             - Active MAC .......... 00145e74e0f0
             - Port name ........... OSAQDIO4
```

# Detailed traffic data for an OSA port

```
Transmit  – Total packets............   11368821 100.00%
          – Good packets.............   11368821 100.00%
          – Broadcast packets.......        1331   0.01%
          – Multicast packets........      365247   3.21%
          – Size: 64 bytes or less...    2860015  25.16%
          – Size: 65 to 127 bytes....    2290829  20.15%
          – Size: 128 to 255 bytes...     554968   4.88%
          – Size: 256 to 511 bytes...      75555   0.66%
          – Size: 512 to 1023 bytes..     164339   1.45%
          – Size: 1024 bytes or more.    5423115  47.70%
Receive   – Total packets............   15265498 100.00%
          – Good packets.............   15265498 100.00%
          – Broadcast packets.......      140888   0.92%
          – Multicast packets........     1716694  11.25%
          – Size: 64 bytes or less...    1907217  12.49%
          – Size: 65 to 127 bytes....    3549608  23.25%
          – Size: 128 to 255 bytes...     174752   1.14%
          – Size: 256 to 511 bytes...     284678   1.86%
          – Size: 512 to 1023 bytes..      79139   0.52%
          – Size: 1024 bytes or more.    9270104  60.73%
```

# Detailed LAN error statistics for an OSA port

```
Errors      - Alignment errors.........          0
            - CRC errors..............          0
            - Missed packets..........          0
            - Single collisions.......          0
            - Multiple collisions......         0
            - Excessive collisions.....         0
            - Late collisions.........          0
            - Deferred................          0
            - Sequence errors.........          0
            - No receive buffer.......          0
            - Receive length error.....         0
            - XON transmitted.........          0
            - XON received............          0
            - XON transmitted.........          0
            - XOFF received...........          0
            - Receive jabber..........          0
            - Receive undersize.......          0
            - Receive oversize........          0
            - Receive fragment........          0
```

**Getting the most out of your OSA adapter with z/OS Communications Server**

# Appendix B:
# OSA-Express3 dual port IOCP, TRLE, and z/OS CS Interface definitions

IBM®

# OSA-Express3 IOCP definitions

```
CHPID PATH=(CSS(0,1),FD),SHARED,                                   *
      PCHID=581,TYPE=OSD


CNTLUNIT CUNUMBR=2FD0,PATH=((CSS(0),FD),(CSS(1),FD)),UNIT=OSA


IODEVICE  ADDRESS=(2FD0,14),CUNUMBR=2FD0,UNIT=OSA,                 *
      UNITADD=00
IODEVICE  ADDRESS=(2FDE,1),CUNUMBR=2FD0,UNIT=OSAD,                 *
      UNITADD=FE
```

```
10.37.39  d m=chp(fd)
      10.37.39  IEE174I 10.37.39 DISPLAY M 825                               C
      CHPID FD:  TYPE=11, DESC=OSA DIRECT EXPRESS, ONLINE
      DEVICE STATUS FOR CHANNEL PATH FD
           0  1  2  3  4  5  6  7  8  9  A  B  C  D  E  F
      02FD +  +  +  +  +  +  +  +  +  +  +  +  +  +  +  .
      SWITCH DEVICE NUMBER = NONE
      PHYSICAL CHANNEL ID = 0581
      ********************** SYMBOL EXPLANATIONS **********************
      + ONLINE     @ PATH NOT VALIDATED   – OFFLINE    . DOES NOT EXIST
      * PHYSICALLY ONLINE   $ PATH NOT OPERATIONAL
```

# OSA Express3 TRLE definitions

```
O3ETHB0T TRLE  LNCTL=MPC,                              X
               READ=(2FD0),                            X
               WRITE=(2FD1),                           X
               MPCLEVEL=QDIO,                          X
               DATAPATH=(2FD2,2FD3,2FD4,2FD5),         X
               PORTNAME=(O3ETHB0P),                    X
               PORTNUM=(0)

O3ETHB1T TRLE  LNCTL=MPC,                              X
               READ=(2FD6),                            X
               WRITE=(2FD7),                           X
               MPCLEVEL=QDIO,                          X
               DATAPATH=(2FD8,2FD9,2FDA,2FDB),         X
               PORTNAME=(O3ETHB1P),                    X
               PORTNUM=(1)
```
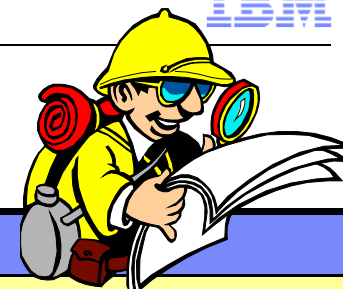
# OSA Express3 TCP/IP Interface definitions

```
INTERFACE O3ETHB0 DEFINE IPAQENET
 PORTNAME O3ETHB0P
 IPADDR 16.11.16.105/20
 SOURCEVIPAINT LGEVIPA1
 MTU 1500
 VLANID 3
 READSTORAGE GLOBAL
 INBPERF BALANCED
 IPBCAST
 MONSYSPLEX
 DYNVLANREG
 OLM
 VMAC 0204100B1069 ROUTEALL

INTERFACE O3ETHB1 DEFINE IPAQENET
 PORTNAME O3ETHB1P
 IPADDR 16.11.17.105/20
 SOURCEVIPAINT LGEVIPA1
 MTU 1500
 VLANID 3
 READSTORAGE GLOBAL
 INBPERF BALANCED
 IPBCAST
 MONSYSPLEX
 DYNVLANREG
 OLM
 VMAC 0204100B1169 ROUTEALL
```

# For more information

| URL | Content |
|-----|---------|
| http://www.twitter.com/IBM_Commserver | IBM Communications Server Twitter Feed |
| http://www.facebook.com/IBMCommserver | IBM Communications Server Facebook Fan Page |
| http://www.ibm.com/systems/z/ | IBM System z in general |
| http://www.ibm.com/systems/z/hardware/networking/ | IBM Mainframe System z networking |
| http://www.ibm.com/software/network/commserver/ | IBM Software Communications Server products |
| http://www.ibm.com/software/network/commserver/zos/ | IBM z/OS Communications Server |
| http://www.ibm.com/software/network/commserver/z_lin/ | IBM Communications Server for Linux on System z |
| http://www.ibm.com/software/network/ccl/ | IBM Communication Controller for Linux on System z |
| http://www.ibm.com/software/network/commserver/library/ | IBM Communications Server library |
| http://www.redbooks.ibm.com | ITSO Redbooks |
| http://www.ibm.com/software/network/commserver/zos/support/ | IBM z/OS Communications Server technical Support – including TechNotes from service |
| http://www.ibm.com/support/techdocs/atsmastr.nsf/Web/TechDocs | Technical support documentation from Washington Systems Center (techdocs, flashes, presentations, white papers, etc.) |
| http://www.rfc-editor.org/rfcsearch.html | Request For Comments (RFC) |
| http://www.ibm.com/systems/z/os/zos/bkserv/ | IBM z/OS Internet library – PDF files of all z/OS manuals including Communications Server |

*For pleasant reading ….*