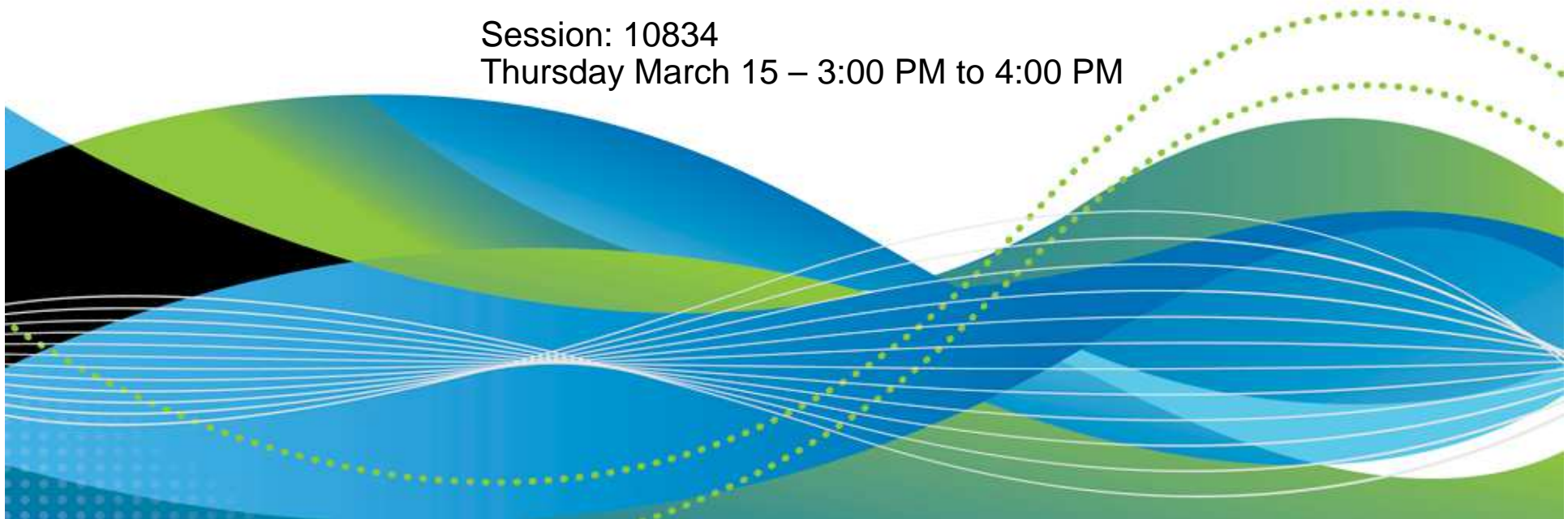


Sysplex and Network topology considerations

Gus Kassimis - kassimis@us.ibm.com

IBM Enterprise Networking Solutions
Raleigh, NC, USA

Session: 10834
Thursday March 15 – 3:00 PM to 4:00 PM



Sysplex and Network topology considerations

Session number:	10834
Date and time:	Thursday, March 15, 2012: 3:00 PM-4:00 PM
Location:	Hickory (Omni Hotel CNN Center)
Program:	Communications Infrastructure
Project:	Communications Server
Track:	Tracks: Network Support and Management
Classification:	Technical
Speaker:	Gus Kassimis, IBM
Abstract:	<p>In this session we will look at how to provide optimized network access to the images and between the images of a z/OS Sysplex. Various newer networking technologies will be reviewed and explained, such as multiple virtual LAN support by OSA, the use of virtual MAC addresses, QDIO acceleration, using WLM service level information to choose QDIO outbound priority queue, Multiple Inbound Queueing, etc. The session will also review various Sysplex access network topology examples and discuss aspects of each. In particular, the session will discuss the recent high availability features, techniques for achieving network interface resilience in a flat network and the impacts of using state-full firewalls in the Sysplex access network.</p>

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® | • GDDM® | • Language Environment® | • Rational Suite® | • zEnterprise |
| • AIX® | • GDPS® | • MQSeries® | • Rational® | • zSeries® |
| • alphaWorks® | • Geographically Dispersed Parallel Sysplex | • MVS | • Redbooks | • z/Architecture |
| • AnyNet® | • HiperSockets | • NetView® | • Redbooks (logo) | • z/OS® |
| • AS/400® | • HPR Channel Connectivity | • OMEGAMON® | • Sysplex Timer® | • z/VM® |
| • BladeCenter® | • HyperSwap | • Open Power | • System i5 | • z/VSE |
| • Candle® | • i5/OS (logo) | • OpenPower | • System p5 | |
| • CICS® | • i5/OS® | • Operating System/2® | • System x® | |
| • DataPower® | • IBM eServer | • Operating System/400® | • System z® | |
| • DB2 Connect | • IBM (logo)® | • OS/2® | • System z9® | |
| • DB2® | • IBM® | • OS/390® | • System z10 | |
| • DRDA® | • IBM zEnterprise™ System | • OS/400® | • Tivoli (logo)® | |
| • e-business on demand® | • IMS | • Parallel Sysplex® | • Tivoli® | |
| • e-business (logo) | • InfiniBand® | • POWER® | • VTAM® | |
| • e-business (logo)® | • IP PrintWay | • POWER7® | • WebSphere® | |
| • ESCON® | • IPDS | • PowerVM | • xSeries® | |
| • FICON® | • iSeries | • PR/SM | • z9® | |
| | • LANDP® | • pSeries® | • z10 BC | |
| | | • RACF® | • z10 EC | |
- * 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



- Intra-Sysplex connectivity
- Networking Sysplex availability
- Using OSA for network connectivity
- Network availability in a flat network environment (No dynamic routing updates)
- Network Subplex support



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.

Sysplex and Network Topology Considerations

Intra-Sysplex connectivity



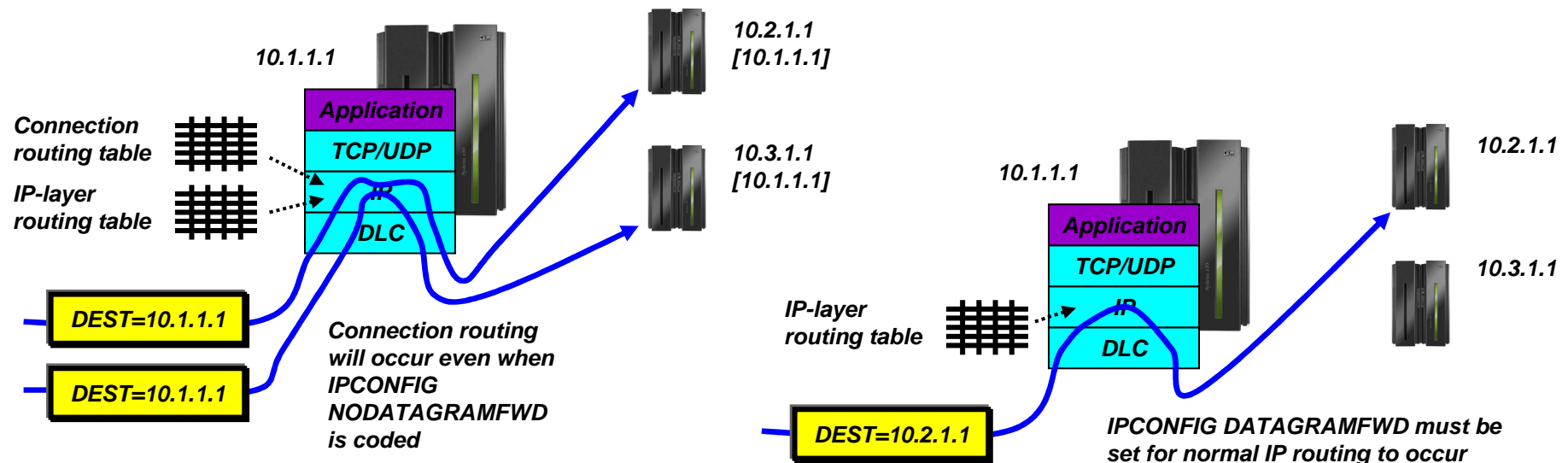
Two types of intra-Sysplex/subplex routing

▪ Connection routing

- IP routing decision based upon connection routing table (CRT), destination IP address and specific connection (4-tuple)
 - Packets to the same IP address, but belonging to two different connections, may go to two different targets
- Used by Sysplex Distributor
- Used by movable Dynamic VIPA support
- Not subject to the setting of IPCONFIG DATAGRAMFWD/NODATAGRAMFWD

▪ Normal IP routing

- IP routing decision based upon IP-layer routing table and destination IP address
 - All packets to the same IP address are treated the same
- Forwarding to z/OS TCP/IP stacks through another z/OS TCP/IP stack
- Subject to the setting of the IPCONFIG DATAGRAMFWD/NODATAGRAMFWD option



The role of XCF, ISTIQDIO HiperSockets, and external LAN interfaces in a z/OS Sysplex/subplex

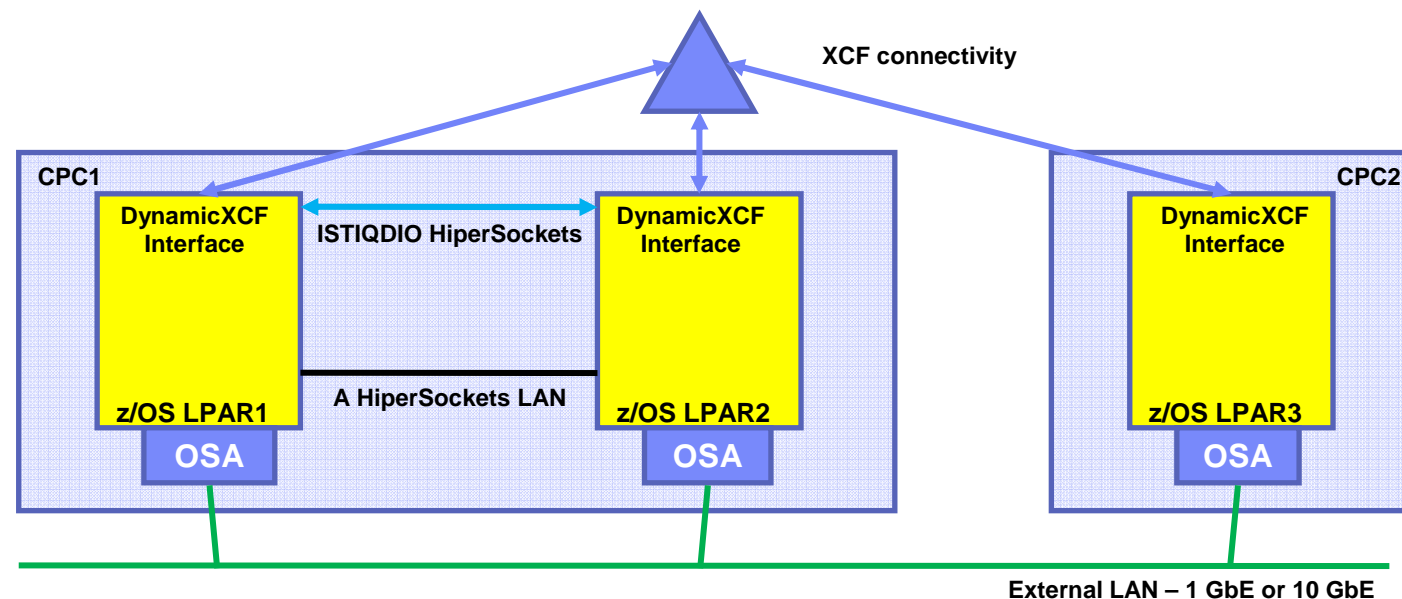
- **XCF**
 - All XCF control messaging between stacks in a Sysplex (DVIPA availability, etc.) – always go via XCF messages
 - DynamicXCF SD connection routing (but only if no VIPAROUTE defined)
 - If so configured through static or dynamic routing, normal IP routing between LPARs – normally not recommended

- **ISTIQDIO HiperSockets**
 - If ISTIQDIO is defined (in VTAM) DynamicXCF SD connection routing between LPARs on same CPC goes this way instead of XCF
 - Considered part of the DynamicXCF network interface – no separate DEVICE/LINK or INTERFACE definitions

- **External LAN or a manually defined HiperSockets LAN**
 - If VIPAROUTE defined, then used for SD connection routing between LPARs
 - VIPAROUTE is generally recommended
 - Normal IP routing

Only define DynamicXCF interfaces as OSPF interfaces, if you want to be able to use XCF as a last-resort connectivity between z/OS stacks.

If you have “enough” redundancy built into your OSA adapters, data center switches, and switch connectivity, you may not need to ever use XCF for normal IP routing.



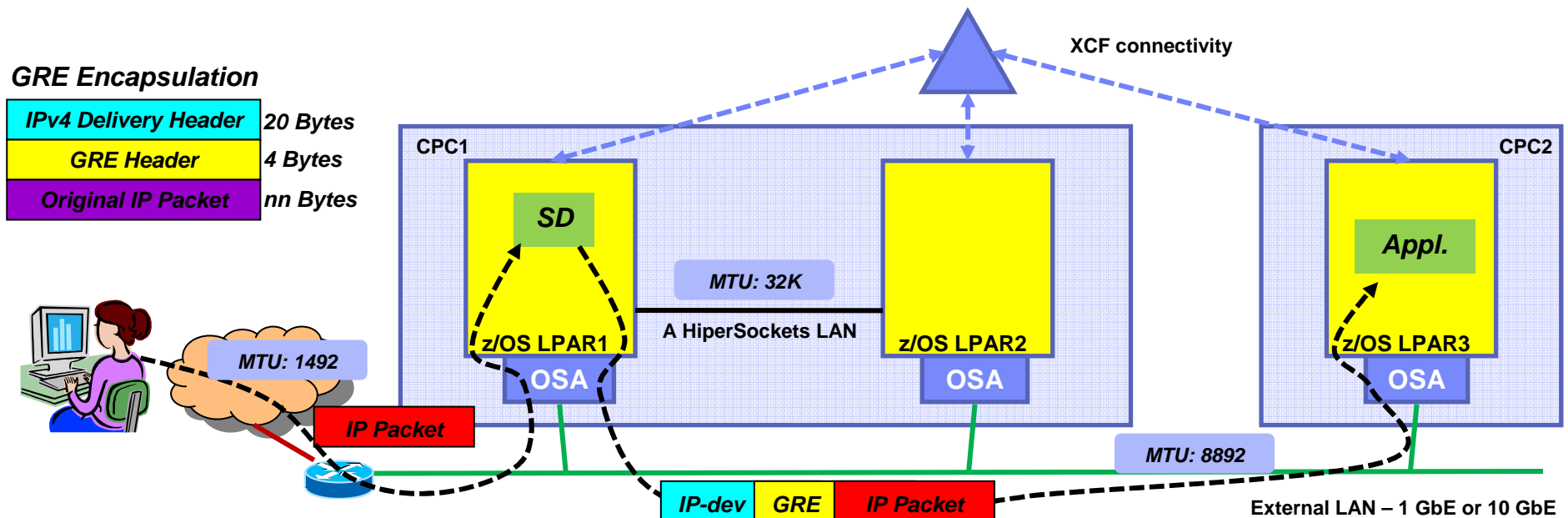
So what should I use for what type of routing?

- **VIPAROUTE** is often the best choice for connection routing
 - Exploits network redundancy
 - Often as fast or faster than XCF
 - Does not use Coupling Facility CPU cycles, which often is a limited resource

	Exchange control messages between stacks in a Sysplex or Subplex	Sysplex Distributor connection routing (forwarding inbound packets for distributed connections)	General IP routing between stacks in a Sysplex or Subplex
XCF messaging	Always	Yes - If no VIPAROUTE specified (or for traffic associated with SWSA and MLS)	Can be used (not recommended)
ISTIQDIO (Dedicated HiperSockets LAN)	Never	Yes - If defined in VTAM start options and no VIPAROUTE defined. Used for connection routing to LPARs on same CPC only.	Can be used (not recommended since XCF will be used for LPARs on other CPCs)
All other connectivity options between stacks in a Sysplex or Subplex (OSA, HiperSockets, Channel links, etc.)	Never	Yes - If VIPAROUTE is defined	Always

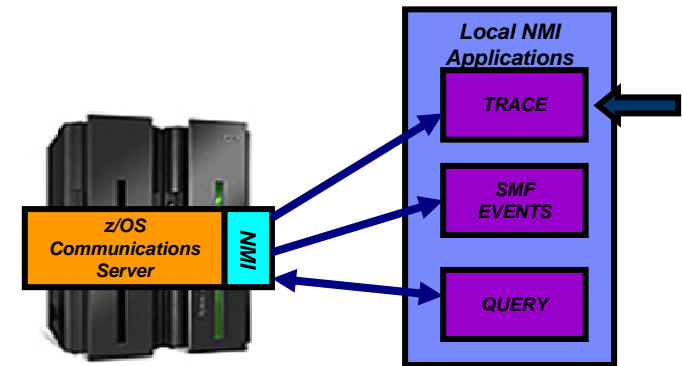
VIPAROUTE and MTU size considerations

- When VIPAROUTE is used, the distributing stack adds a GRE header to the original IP packet before forwarding to the target stack
- Two ways to avoid fragmentation between distributing and target stacks:
 - Have clients use path MTU discovery
 - z/OS will factor in the GRE header size (24 bytes) when responding with next-hop MTU size
 - Not always possible to control distributed nodes' settings from the data center
 - Use jumbo-frames on the data center network
 - The access network will typically be limited to Ethernet MTU size (1492 bytes), while the data center network will be able to use jumbo frame MTU size (8892 bytes)
 - Adding the GRE header will not cause fragmentation in this scenario



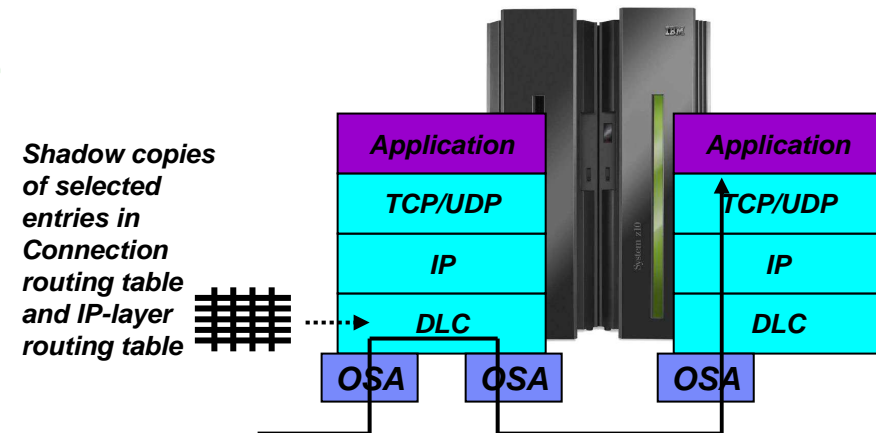
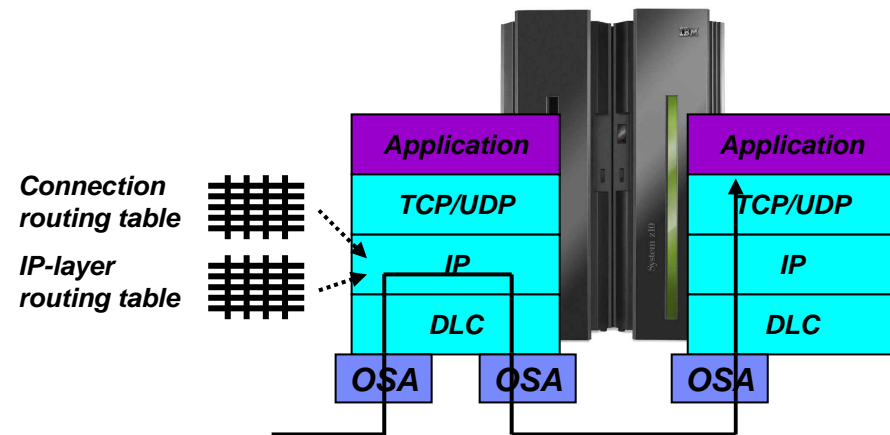
z/OS V1R12 - Enhance Packet Trace for Sysplex Distributor VIPAROUTE traffic

- Apply Packet Trace filters to Sysplex Distributor VIPAROUTE traffic
 - Sysplex Distributor encapsulates VIPAROUTE traffic with GRE header, for IPv4 traffic, or an IPv6 header, for IPv6 traffic
 - Existing filter support only operates on the outer packet header, not the encapsulated packet
 - Packet Trace can now filter on the destination DVIPA address and/or the ports located inside the encapsulated packet
- In addition, the next hop address is now included in the packet trace

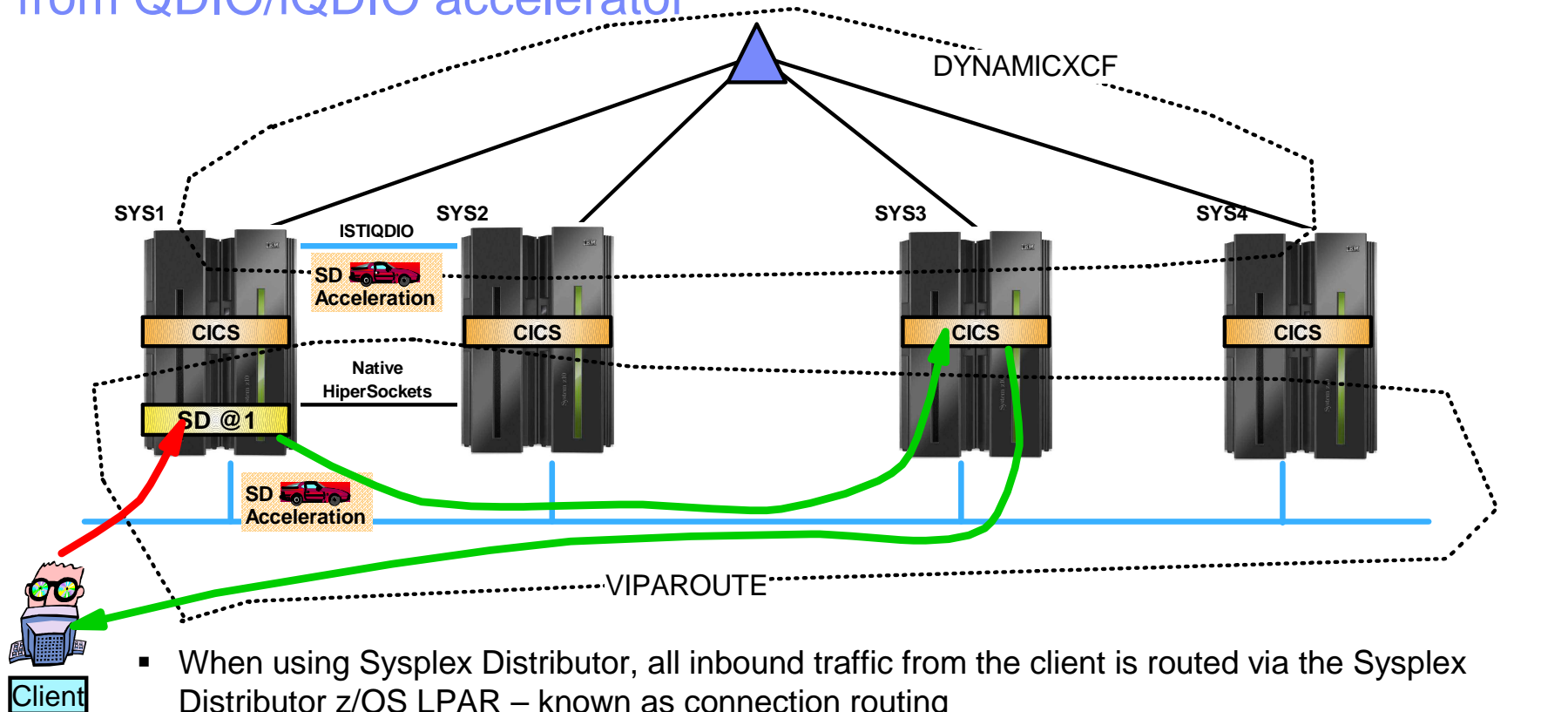


z/OS V1R11 QDIO and iQDIO routing accelerator

- **Provides fast path IP forwarding for these DLC combinations:**
 - Inbound OSA-E QDIO → Outbound OSA-E QDIO or HiperSockets
 - Inbound HiperSockets → Outbound OSA-E QDIO or HiperSockets
- **Adds Sysplex Distributor (SD) acceleration**
 - Inbound packets over HiperSockets or OSA-E QDIO
 - When SD gets to the target stack using either:
 - Dynamic XCF connectivity over HiperSockets
 - VIPAROUTE over OSA-E QDIO
- **Improves performance and reduces processor usage for such workloads**
- **Restrictions:**
 - QDIO routing accelerator is IPv4 only
 - Mutually exclusive with IPSECURITY
 - Requires IP Forwarding to be enabled (for non-SD acceleration)
 - No acceleration for:
 - Traffic which requires fragmentation in order to be forwarded
 - VIPAROUTE over HiperSockets
 - Incoming fragments for an SD connection
 - Interfaces using optimized latency mode (OLM)



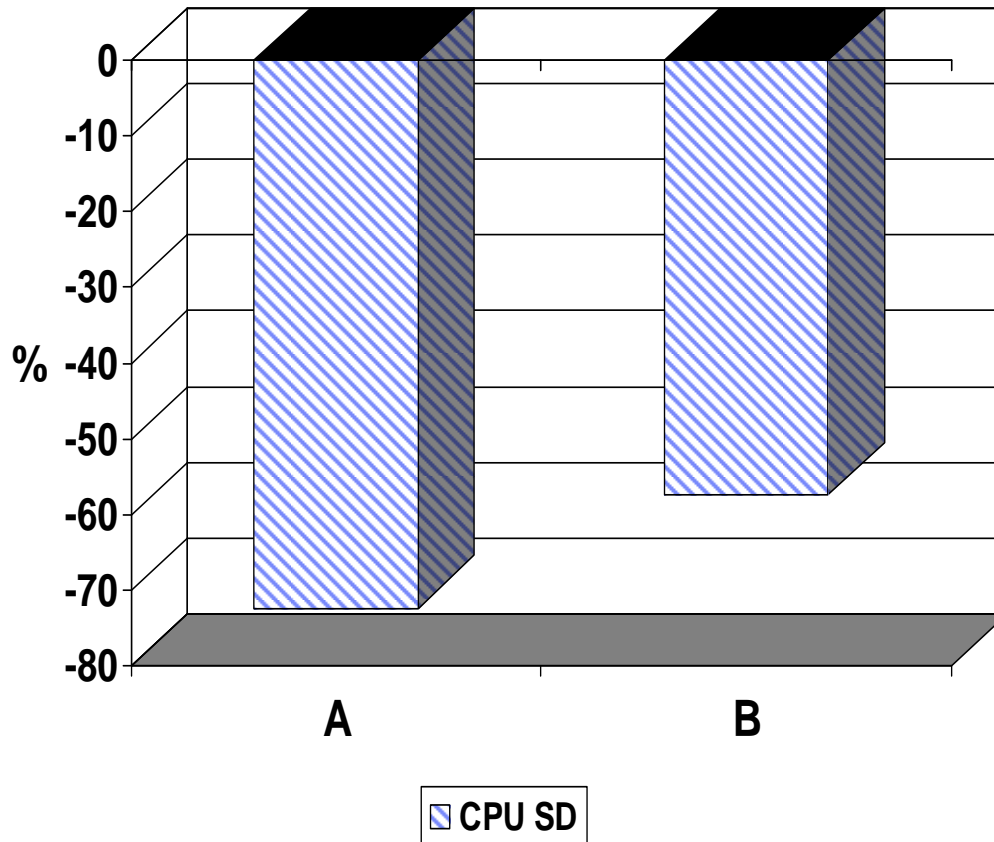
z/OS V1R11: Sysplex Distributor connection routing may benefit from QDIO/iQDIO accelerator



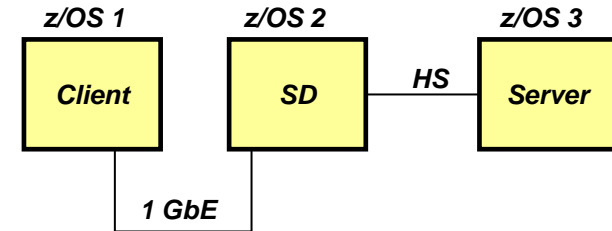
- When using Sysplex Distributor, all inbound traffic from the client is routed via the Sysplex Distributor z/OS LPAR – known as connection routing
 - Outbound traffic goes directly back to the client
- When inbound packets to Sysplex Distributor is over QDIO or iQDIO (HiperSockets), Sysplex Distributor will perform accelerated connection routing when outbound is a DYNAMICXCF iQDIO interface - or when the outbound interface is a QDIO network interface
 - Helping reduce CPU overhead and latency in the Sysplex Distributor LPAR (SYS1)

Sysplex Distributor accelerator performance

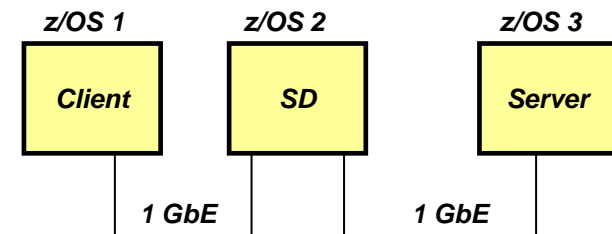
- ✓ Intended to benefit all existing Sysplex Distributor users
- ✓ Measurements with Interactive workload (RR20)
- ✓ Small data sizes (100 in, 800 out)
- ✓ Percentages relative to no acceleration



Configuration A – Three z10 LPARs with OSA Express 3 cards and HiperSockets between SD and server LPARs



Configuration B – Three z10 LPARs with OSA Express 3 cards



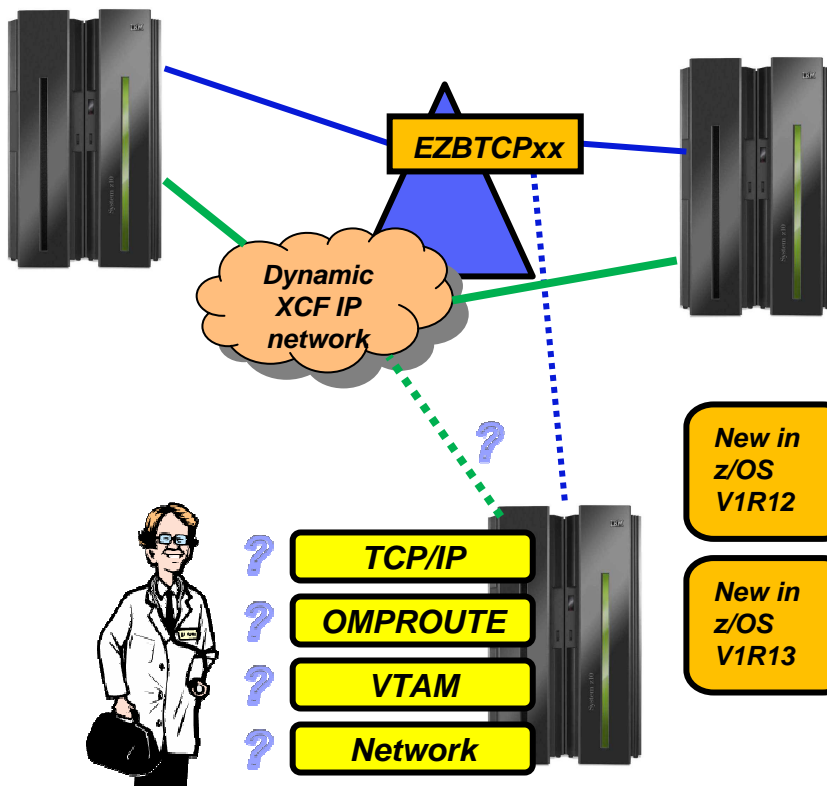
Note: The performance measurements discussed in this presentation are preliminary z/OS V1R11 Communications Server numbers and were collected using a dedicated system environment. The results obtained in other configurations or operating system environments may vary.

Sysplex and Network Topology Considerations

Networking Sysplex availability



Sysplex autonomics extended with CSM storage constrained monitoring



Monitoring:

- Monitor CS health indicators
 - Storage usage critical condition (>90%) - CSM, TCPIP Private & ECSA
 - For more than TIMERSECS seconds
- Monitor dependent networking functions
 - OMPROUTE availability
 - VTAM availability
 - XCF links available
- Monitor for abends in Sysplex-related stack components
 - Selected internal components that are vital to Sysplex processing
 - Does not include "all" components
- Selected network interface availability and routing
- **Monitor for repetitive internal abends in non-Sysplex related stack components**
 - **5 times in less than 1 minute**
- **Detect when CSM FIXED or CSM ECSA has been constrained (>80% utilization) for multiple monitoring intervals**
 - **For 3 times the TIMERSECS value**

Actions:

- Remove the stack from the IP Sysplex (manual or automatic)
 - Retain the current Sysplex configuration data in an inactive state when a stack leaves the Sysplex
- Reactivate the currently inactive Sysplex configuration when a stack rejoins the Sysplex (manual or automatic)



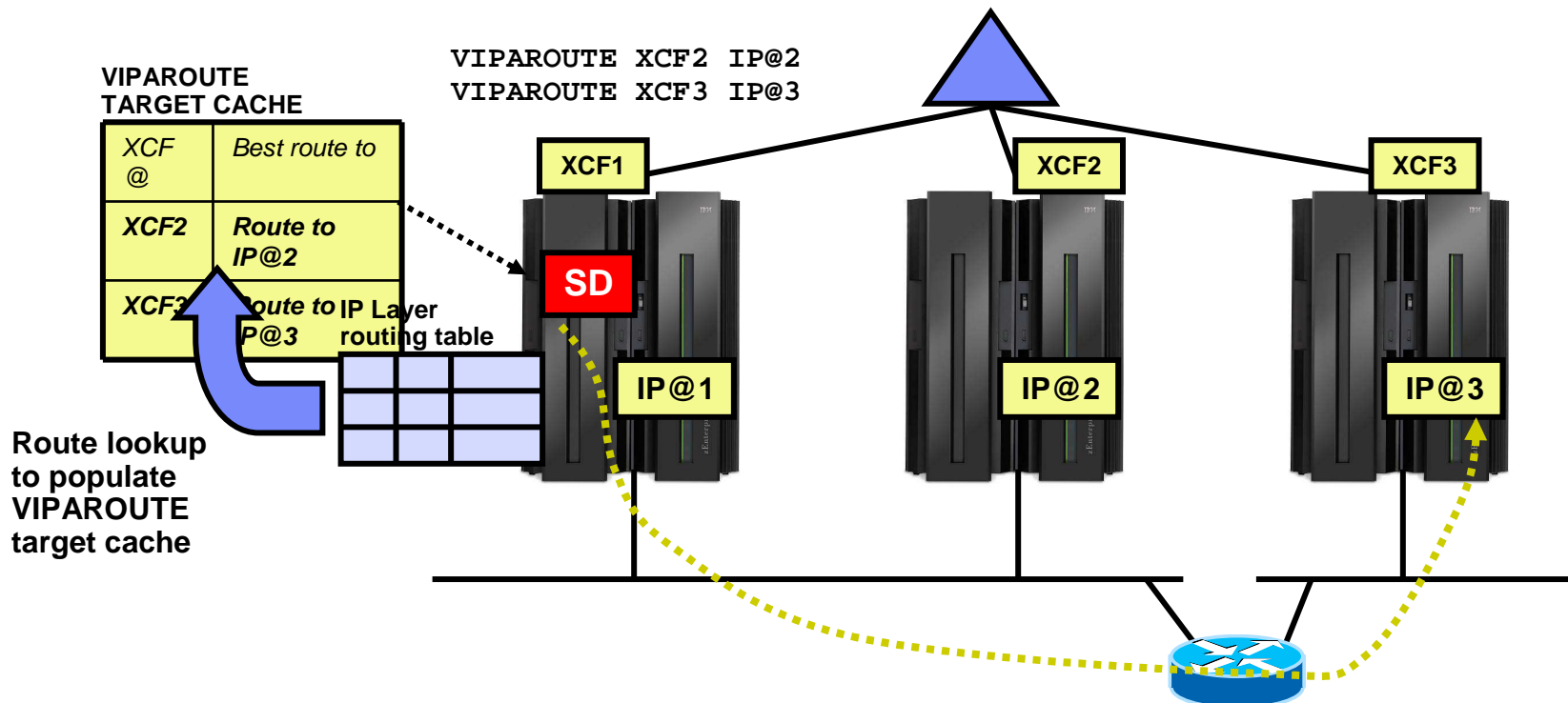
Sick? Better remove myself from the IP Sysplex!



Feeling better? Maybe it's time to rejoin the IP Sysplex

VIPAROUTE target cache update during initialization

- When using VIPAROUTE, a VIPAROUTE target cache is used to minimize the time it takes to route a Sysplex Distributor packet
- The target cache is updated every 60 seconds, which in some cases have caused delays during a primary stack's take-back of a distributed DVIPA
- z/OS V1R13 shortens the interval for VIPAROUTE route lookups in situations where the stack joins a Sysplex, or OMPROUTE is restarted
 - Will now start with 5 seconds, and gradually increase to 60 seconds



Basing automated operations on TCP/IP start-up messages

```
GlobalConfig SysplexMonitor ; Enable Sysplex autonomics
              TimerSecs 60 ; Check interval (default)
              AutoRejoin ; Rejoin automatically
              DelayJoin ; Delay joining till OMPROUTE is up
              DynRoute ; Interface mon w. dyn routes
              MonInterface ; Interface monitoring
              Recovery ; Remove myself automatically
;
DEVICE OSAQDIO4 MPCIPA
LINK QDIO4 IPAQENET OSAQDIO4 MONSYSPLEX
;
INTERFACE QDIO6
  DEFINE IPAQENET6
  PORTNAME OSAQDIO4
  MONSYSPLEX
```

This set of SysplexMonitor definitions will automatically leave and join the Sysplex based on the availability and health of selected Sysplex TCP/IP resources.

When starting TCP/IP, the stack will not join the Sysplex until OMPROUTE is up and running and has learned dynamic routes over at least one monitored network interface (those coded with the MONSYSPLEX keyword)

```
*EZD1166E TCPCS DELAYING SYSPLEX PROFILE PROCESSING - OMPROUTE IS NOT
ACTIVE
```

```
*EZD1211E TCPCS DELAYING SYSPLEX PROFILE PROCESSING - ALL MONITORED
INTERFACES WERE NOT ACTIVE
```

```
*EZD1212E TCPCS DELAYING SYSPLEX PROFILE PROCESSING - NO DYNAMIC ROUTES
OVER MONITORED INTERFACES WERE FOUND
```

```
EZD1176I TCPCS HAS SUCCESSFULLY JOINED THE TCP/IP SYSPLEX GROUP EZBTCPCS
EZD1214I INITIAL DYNAMIC VIPA PROCESSING HAS COMPLETED FOR TCPCS ←
```

Some applications do resolver calls when they start up. If they are started after TCPIP is up, but OMPROUTE has not learned the needed routes, resolver calls that need to use the DNS may fail. So, there is a need to start these applications after a route has been learned. If you are not using AUTOLOG with DELAYSTART DVIPA to start server address spaces, let your automation software kick off on the EZD1214I message.

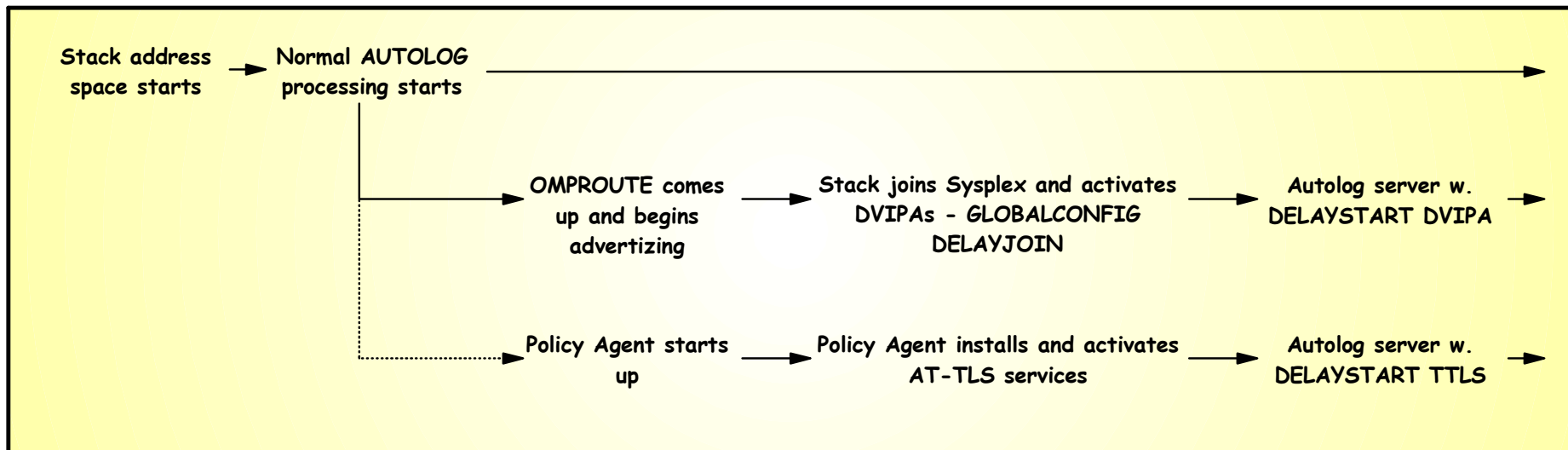
z/OS V1R10 implemented improved AUTOLOG sequencing of the TCP/IP start-up process

- The pre-V1R10 AUTOLOG DELAYSTART option delays application start until DVIPAs are configured and active
- z/OS V1R10 adds another option to AUTOLOG DELAYSTART that can be used to delay start of an application until AT-TLS services are available

```

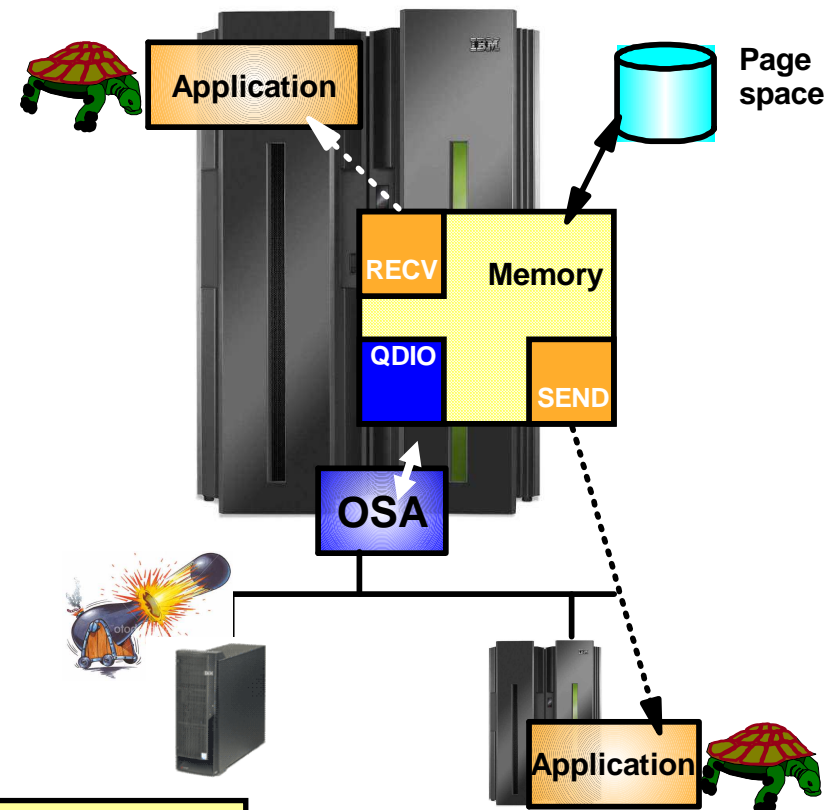
+-DVIPA--+
DELAYSTART-+-----+-----
+-TTLS---+
    
```

If DELAYSTART is specified without a sub-option, it defaults to DVIPA



z/OS V1R11 improved responsiveness to storage shortage conditions

- Improved OMPROUTE tolerance for storage shortage situations
- Improved handling of situations where “slow” applications use excessive amounts of storage buffers at the transport protocol layer
- Throttle amount of parallel QDIO operations
- Data-link control (DLC) level discard of QDIO input buffers to relieve inbound overrun

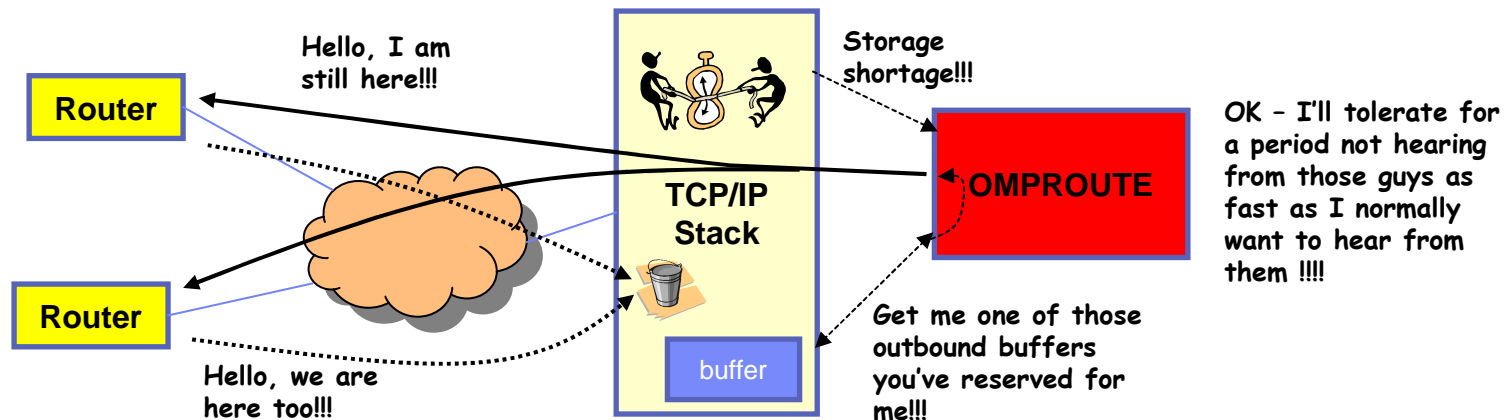


When storage shortage occurs:

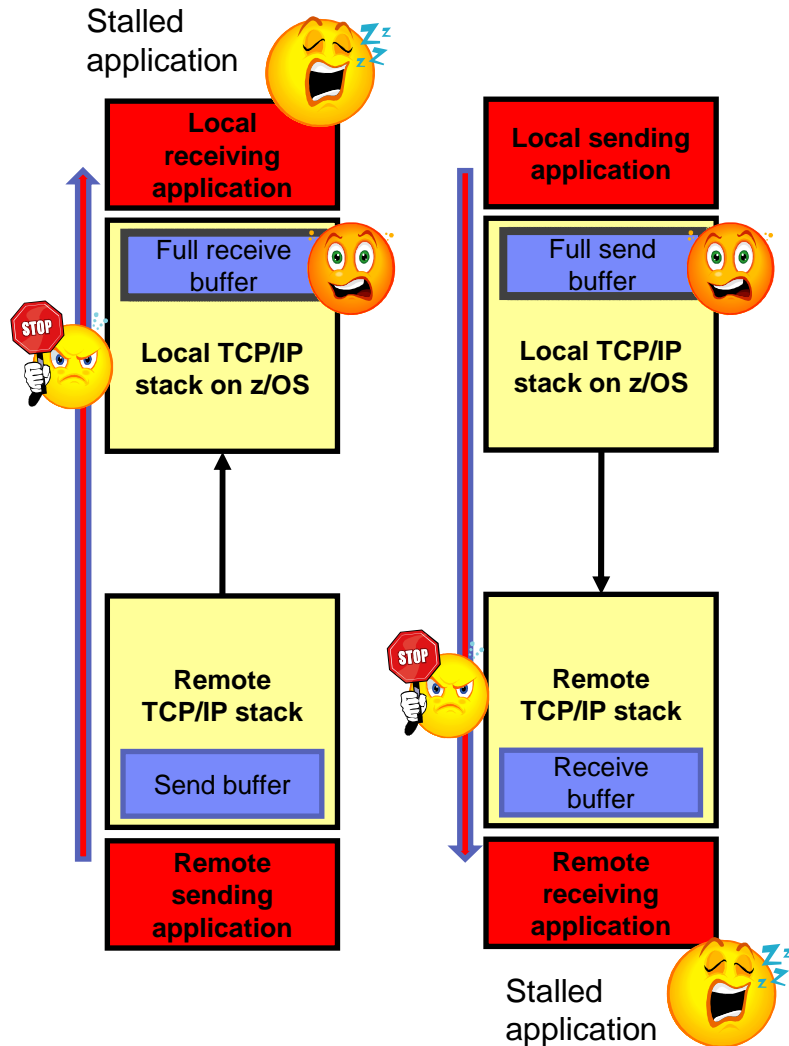
- ✓ *Stay up!*
- ✓ *Throttle workload at the source*
- ✓ *Prevent network spikes from monopolizing common z/OS storage*
- ✓ *Report which connections use excessive amounts of storage*

z/OS V1R11 storage shortages and OMPROUTE availability

- OMPROUTE and the TCP/IP stack work together to make OMPROUTE more tolerant of storage shortage conditions:
 - TCP/IP stack informs OMPROUTE of stack storage shortage conditions
 - During a storage shortage, OMPROUTE temporarily suspends requirement for periodic routing updates from neighbor routers
 - TCP/IP stack ensures that dispatchable units for OMPROUTE can always obtain the control blocks that they require
 - TCP/IP stack satisfies storage requests for OMPROUTE as long as storage remains available
- Temporarily keeps OMPROUTE from timing out routes due to lack of routing updates from neighbor routers during a storage shortage
- Decreases likelihood of OMPROUTE exiting or failing to send routing updates to neighbor routers



z/OS V1R11 storage shortages and slow or stalled applications



- Data in a send buffer is page fixed – awaiting IO operations to be initiated
 - When application is not making progress or fixed storage is constrained
 - All new data added to TCP send queue is marked as page-able
 - When storage becomes constrained, all unsent data on send queues for all non-local TCP connections is marked as page-able
 - Before data is sent to remote stack it is changed back to fixed, as required by the DLC

- It was very difficult to identify which local applications caused excessive amounts of space to be used on the send or receive queues
 - Alerts issued to indicate TCP queue in constrained state
 - Indicate old data on send or receive queue
 - Identify connection (connection id, job name, addresses, ports)
 - Constrained state entry and exit indicated
 - Issued to syslogd using TRMD

z/OS V1R11 Storage shortages and slow or stalled applications...

- This feature is *automatically enabled*
- Look for these messages in syslogd

EZZ8662I TRMD TCP receive queue constrained entry logged: *date time* , connid= *connid* ,
jobname= *jobname* , lipaddr= *lipaddr* , lport= *lport* , ripaddr= *ripaddr* , rport= *rport* ,
correlator= *correlator* , probeid= *probeid* , sensorhostname= *sensorhostname*

EZZ8663I TRMD TCP receive queue constrained exit logged: *date time* , connid= *connid* ,
jobname= *jobname* , lipaddr= *lipaddr* , lport= *lport* , ripaddr= *ripaddr* , rport= *rport* ,
correlator= *correlator* , duration= *duration* , probeid= *probeid* , sensorhostname= *sensorhostname*

EZZ8664I TRMD TCP send queue constrained entry logged: *date time* , connid= *connid* ,
jobname= *jobname* , lipaddr= *lipaddr* , lport= *lport* , ripaddr= *ripaddr* , rport= *rport* ,
correlator= *correlator* , probeid= *probeid* , sensorhostname= *sensorhostname*

EZZ8665I TRMD TCP send queue constrained exit logged: *date time* , connid= *connid* ,
jobname= *jobname* , lipaddr= *lipaddr* , lport= *lport* , ripaddr= *ripaddr* , rport= *rport* ,
correlator= *correlator* , duration= *duration* , probeid= *probeid* , sensorhostname= *sensorhostname*

Use the *correlator* value to identify pairs of entry/exit messages

An entry message with no corresponding exit message indicates queue is still constrained

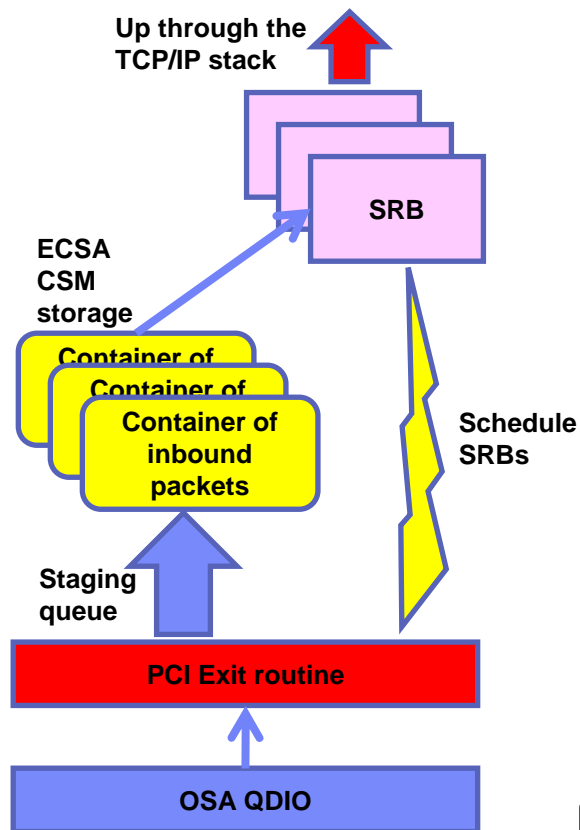
Storage shortages and slow or stalled applications...

- **TCP Queue Size – New IDS policy controls in z/OS V1R13**
 - Helps you manage the amount of storage TCP can take up for the queues used for holding sent and received data
 - For example, out of order packets awaiting re-sequencing
 - Provides user control over storage constraint availability improvements added in z/OS V1R11
 - Helps avoid TCP causing storage constraint situations
 - IDS policy enables dropping of connections that exceed specified limits

z/OS V1R11 storage shortages and QDIO device driver actions

Before z/OS V1R11, there was no limits on

1. Number of SRBs
2. Number of containers on the staging queue



- **Number of parallel SRBs is now limited to:**

- For 1 Gigabit Ethernet:
 - Maximum execution threads per QDIO data device = 4
- For 10 Gigabit Ethernet and HiperSockets:
 - Maximum execution threads per QDIO data device = $\text{Min}(\text{LPAR CPUs} + 1, 4) * 2$

- **Use of CSM storage for containers on the staging queue is also being limited:**

- Gigabit speed OSA-Express
 - Two Meg if CSM critical/constrained
 - Four Meg if CSM not critical/constrained
- Ten-Gigabit speed OSA-Express or HiperSockets
 - Four Meg if CSM critical/constrained
 - Six Meg if CSM not critical/constrained

- **If more data arrives than the current limit allows, inbound packets will be discarded**

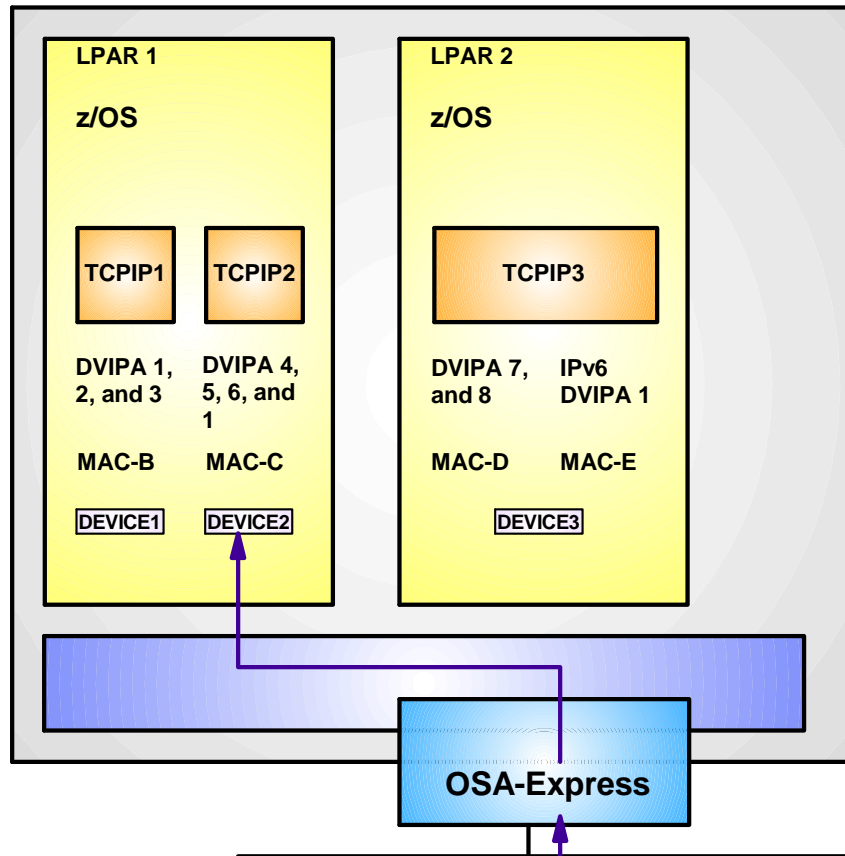
IST2273E PACKETS DISCARDED FOR jobname - READ QUEUE CONGESTION

Sysplex and Network Topology Considerations

Using OSA for network connectivity



Virtualizing the OSA adapter - based on virtual MACs per IP network interface that share the adapter



OSA "routing" logic for inbound packets:

1. Destination MAC address
2. VLAN ID
3. IPv4 or IPv6 address

DestMAC=MAC-C, DestIP=DVIPA1

- **Enables first hop routers and load balancers to use dispatch mode (MAC-level) forwarding**
 - Avoids use of GRE
 - Enables use of dispatch mode by devices that do not support GRE (Cisco CSM and CSS or F5's BigIP)
 - Enables use of dispatch mode for IPv6 for which GRE isn't defined
 - Removes the need for using NAT instead of dispatch mode forwarding
 - NAT requires strict control of outbound path to handle NAT on outbound flows

- **Makes System z LPARs look more like "normal" TCP/IP nodes on a LAN**
 - Simplifies network infrastructure
 - Avoids the whole PRIROUTER/SECROUTER setup issue when sharing a port between multiple LPARs
 - Layer-2 visibility into final source/destination of LAN traffic

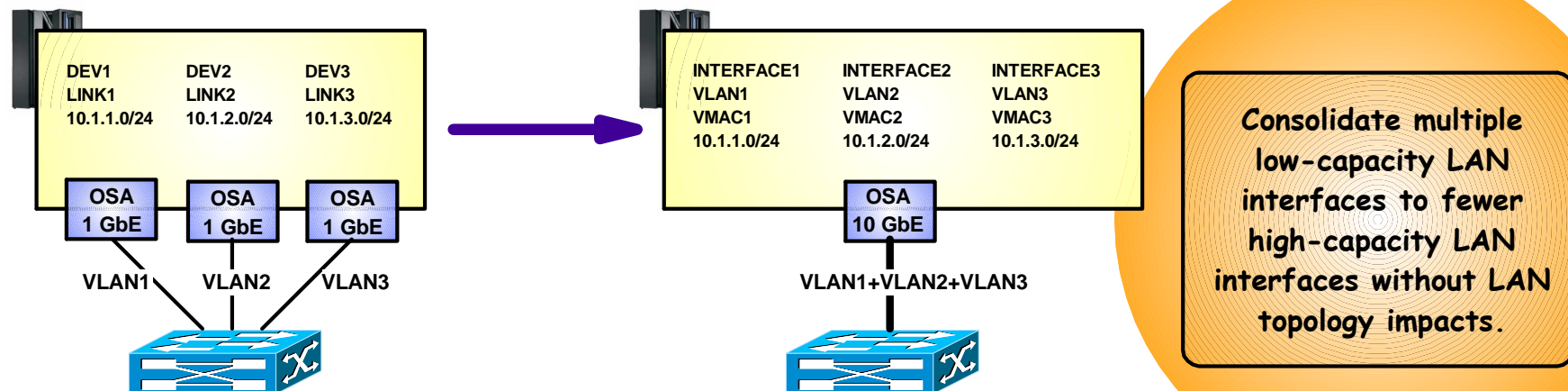
OSA-Express virtual MAC while operating in QDIO layer-3 mode

- **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
- **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 and z10**
- **Prior to z/OS V1R10, each stack may define one VLAN per protocol (IPv4 or IPv6) for each OSA port**
 - One VLAN/VMAC for the LINK statement
 - One VLAN/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
- **z/OS V1R13 will raise limit from 8 to 32 VLANs per stack per OSA port**
 - No impact to OSA
 - Driven by emphasis on VLANs for network isolation in the zEnterprise IEDN



Multiple VLANs per OSA port per stack per IP protocol version

- **Per OSA port, a z/OS TCP/IP stack prior to z/OS V1R10 supported at a maximum two VLANs:**
 - one VLAN for IPv4
 - one VLAN for IPv6
- **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 multiple 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 a new IPv4-enabled version of the INTERFACE configuration statement
 - IPv4 INTERFACE statement only supports QDIO interfaces
 - Start converting your IPv4 QDIO DEVICE/LINK/HOME stmts. to INTERFACE stmts.
 - Each VLAN must use layer-3 virtual MAC addresses and each VLAN must have a unique MAC address

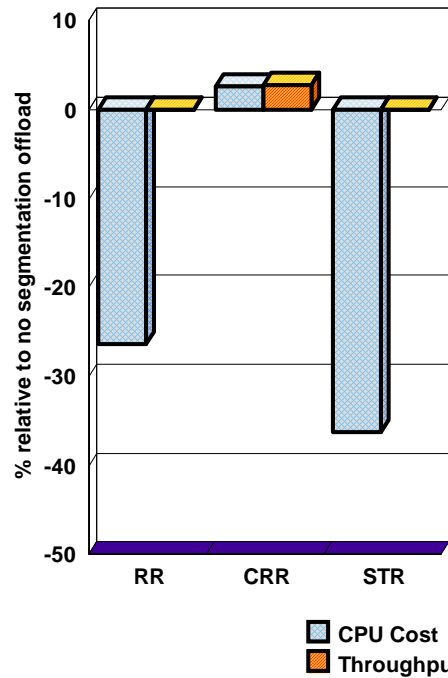


z/OS V1R10 segmentation offload performance measurements on a z10

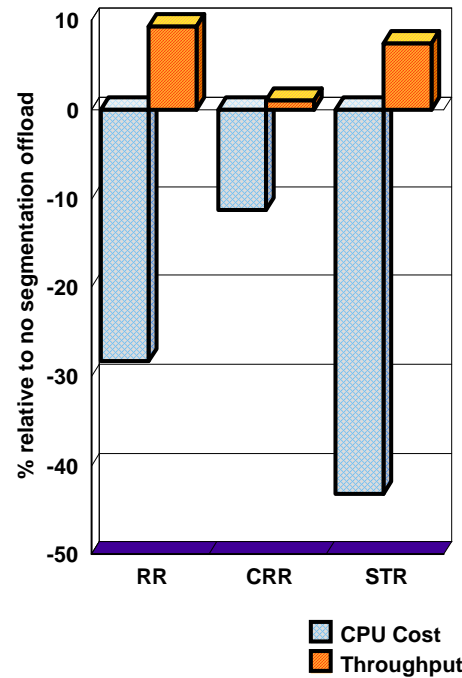


Note: The performance measurements discussed in this presentation were collected using a dedicated system environment. The results obtained in other configurations or operating system

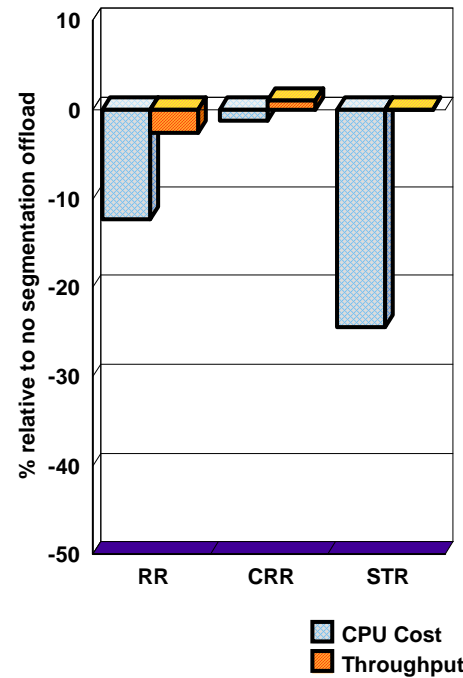
OSA Express3 1Gb



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.

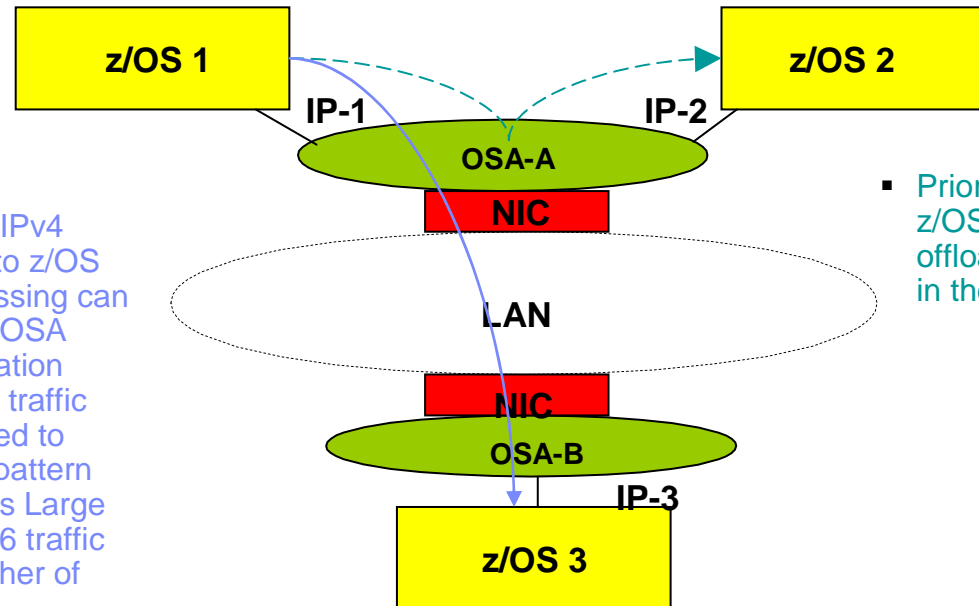
Send buffer size: 180K for streaming workload

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



Proceed with caution !

OSA-Express4S – Support of new features



- Prior to V1R13, for IPv4 traffic from z/OS 1 to z/OS 3, checksum processing can be offloaded to the OSA NIC. TCP segmentation processing for IPv4 traffic can also be offloaded to OSA on this traffic pattern (using OSA-Express Large Send support). IPv6 traffic could not exploit either of these features.

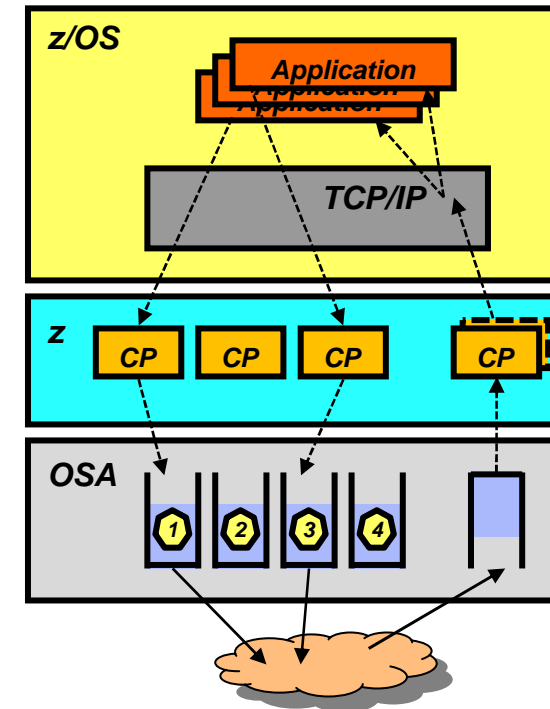
- Prior to V1R13, traffic from z/OS 1 to z/OS 2 could not exploit checksum offload. This processing would occur in the TCP/IP stack layer.

- New OSA-Express4S features exploited by z/OS V1R13 Communications Server
 - Additional checksum offload support
 - For IPv6 traffic
 - For LPAR-to-LPAR traffic (IPv4 and IPv6)
 - Large Send support for IPv6 traffic (aka “TCP segmentation offload”)
 - Note: LPAR-to-LPAR traffic (IPv4 or IPv6) cannot exploit Large Send support.
- OSA-Express4S
 - New OSA-Express, smaller form factor, exploits new I/O drawer enabled for PCIe Gen2 (increased capacity, granularity and bandwidth)
- For more details, refer to IBM US Hardware Announcement 111-136, dated July 12, 2011
http://www-01.ibm.com/common/ssi/rep_ca/6/897/ENUS111-136/ENUS111-136.PDF

Pre V1R12 OSA inbound/outbound processing overview

- Queued Direct IO (QDIO) uses multiple write queues for outbound traffic separation
 - Outbound traffic is separated by priority (policy or WLM)
 - Multiple CPs can be used to manage the write queues

- QDIO uses only one read queue
 - All inbound traffic is received on the single read queue
 - Multiple CPs are used only when data is accumulating on the queue
 - During bursts of inbound data
 - Single process for initial interrupt and read buffer packaging
 - TCP/IP stack performs inbound data separation
 - Sysplex distributor traffic
 - Bulk inbound, such as FTP
 - IPv4/IPv6
 - EE traffic
 - Etc.
 - z/OS Communications Server is becoming the bottleneck as OSA nears 10GbE line speed
 - Inject latency
 - Increase processor utilization
 - Impede scalability

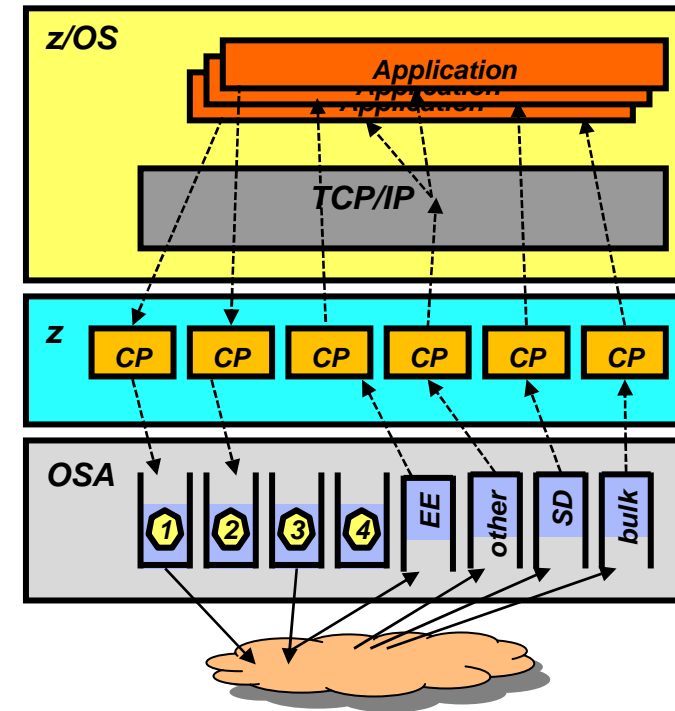


Performance problems observed for bulk inbound traffic:

- Multiple processes run when data is accumulating on the read queue
- Inbound data for a single TCP connection can arrive at the TCP layer out of order
- TCP transmits a duplicate ACK every time it sees out of order data
- Sending side enters fast retransmit recovery

OSA Inbound Workload Queueing (IWQ): Improved performance for mixed traffic patterns

- 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
 - Dynamic LAN idle timer updated per queue
- Supported traffic types (z/OS V1R12)
 - 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
- **New for z/OS V1R13 – Unique inbound queue for Enterprise Extender traffic**
 - Improved performance for EE traffic
 - Supported on OSA-Express3 and new OSA-Express4S (CHPID type OSD or OSX)
- Significant performance improvement for mixed workloads/traffic patterns – for more details see:
 - http://www-01.ibm.com/common/ssi/rep_ca/6/897/ENUS111-136/ENUS111-136.PDF



TCP/IP defines and assigns traffic to queues dynamically based on local IP address and port

Bulk traffic

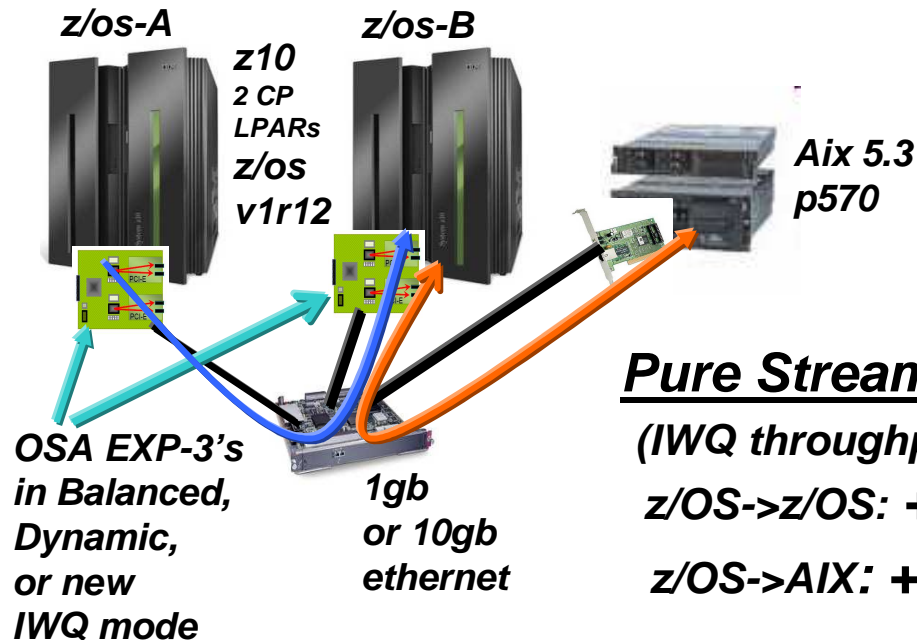
- Application sets send or receive buffer to at least 180K
- Registered per connection (5-tuple)

SD traffic

- Based on active VIPADISTRIBUTE definitions
- Registered on DVIPA address

Inbound Workload Queueing: Performance Data

Performance Test Configuration:



Your mileage may vary. Performance notes: For z/OS outbound streaming to another platform, degree of performance boost (due to IWQ) is relative to receiving platform's sensitivity to out-of-order packet delivery; for streaming INTO z/OS, IWQ will be especially beneficial when transmission is over "lossy" links.

Pure Streaming Workloads:

(IWQ throughput boost relative to INBPERF DYNAMIC)

z/OS->z/OS: +30% (23 to 41%)

z/OS->AIX: +40% (39 to 41%)

Mixed Interactive+Streaming Workload:

(workload is: interactive request/response workload running between z/OS-B and AIX, while z/OS-B is also receiving streaming traffic from z/OS-A over the same 1Gb OSA-3 handling the R/R traffic. We compare z/OS-B's OSA-3 running in IWQ mode, vs Dynamic Mode. IWQ throughput and response time improvements are relative to INBPERF DYNAMIC.)

z/OS<->AIX R/R Throughput improved 55% (Response Time improved 36%)

Streaming Throughput also improved in this test: +5%

Inbound Workload Queuing - Requirements

- IWQ requires OSA-Express3 in QDIO mode running on IBM System z10 or zEnterprise 196. For z10: minimum OSA-Express3 microcode level: Driver 79, EC N24398, MCL003. For zEnterprise 196: the current field level recommended for OSA Express 3 IWQ is 0.0F
- IWQ must be configured using the INTERFACE statement (not DEVICE/LINK)
- IWQ is not supported when z/OS is running as a z/VM guest with simulated devices (VSWITCH or guest LAN)
- Please apply z/OS V1R12 PTF UK61028 (APAR PM20056) for added streaming throughput boost with IWQ

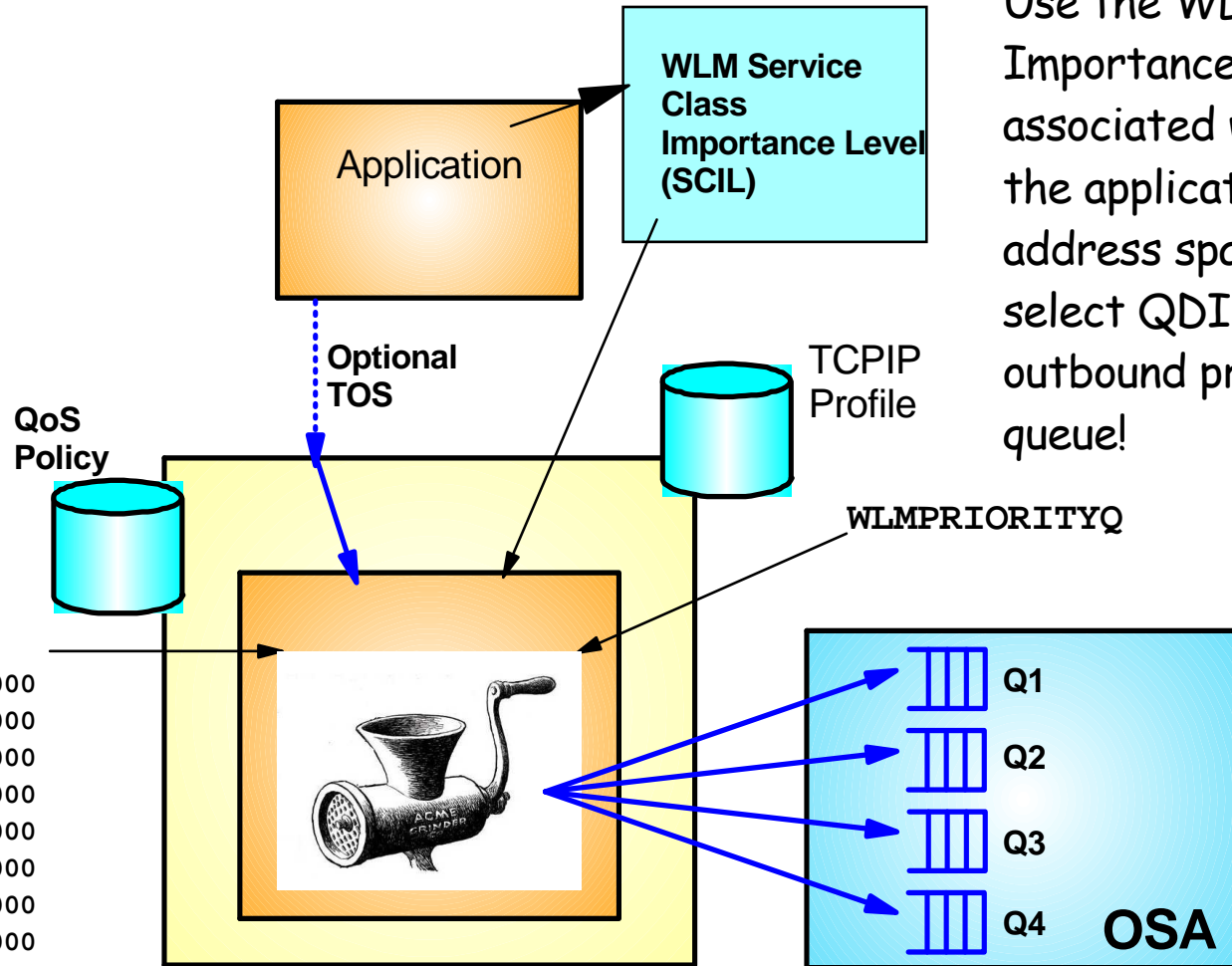
z/OS V1R11 - Extending WLM priorities to Outbound Network I/O (OSA Express)

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

Use the WLM Importance Level associated with the application address spaces to select QDIO outbound priority queue!

```

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
}
    
```



The default QDIO priority queue mapping

WLM Service classes	TCP/IP assigned control value	Default QDIO queue mapping
SYSTEM	n/a	Always queue 1
SYSSTC	0	Queue 1
User-defined with IL 1	1	Queue 2
User-defined with IL 2	2	Queue 3
User-defined with IL 3	3	Queue 3
User-defined with IL 4	4	Queue 4
User-defined with IL 5	5	Queue 4
User-defined with discretionary goal	6	Queue 4

```
GLOBALCONFIG ... WLM PRIORITYQ
```

```
IOPRI1 0
```

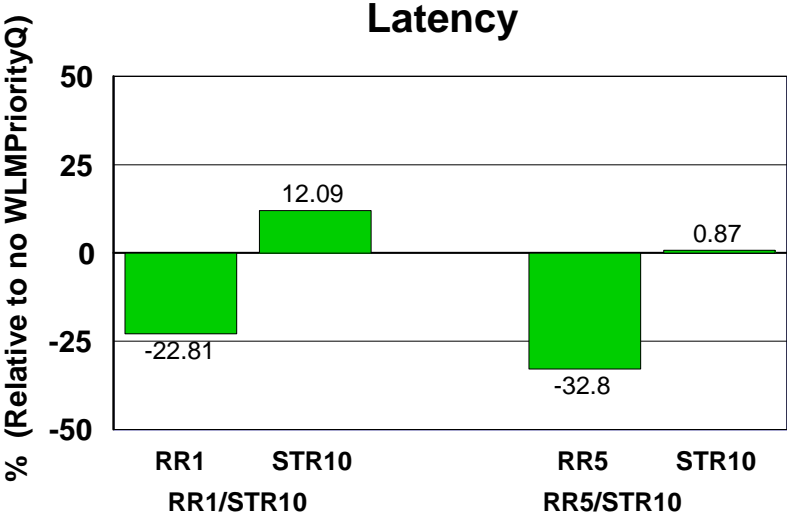
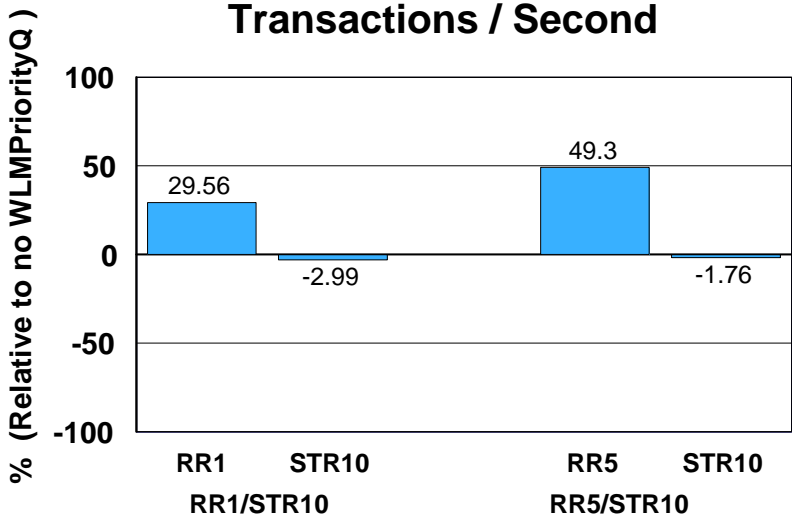
```
IOPRI2 1
```

```
IOPRI3 2 3
```

```
IOPRI4 4 5 6 FWD
```

FWD indicates forwarded (or routed) traffic, which by default will use QDIO priority queue 4

OSA Express (QDIO) WLM Outbound Priority Queuing



- ▶ 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
- ▶ WLM PRIORITYQ 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.

Sysplex and Network Topology Considerations

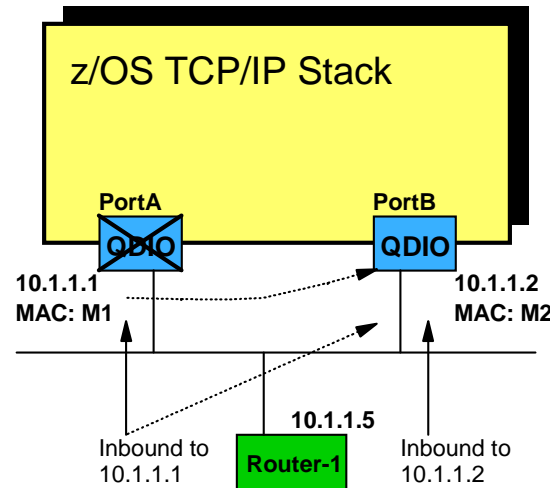
**Network availability in a flat
network environment
(No dynamic routing updates)**



Interface resilience without dynamic routing

Requirement for this feature to function properly:

- ▶ At least two adapters attached to the same network (broadcast media) - referred to as a LAN group.
- ▶ Adapters must use either LCS or QDIO
- ▶ The two adapters should be two physical adapters for real availability benefits



10.x.y.0/24

Router's initial ARP Cache

IP address	Mac address
10.1.1.1	M1
10.1.1.2	M2

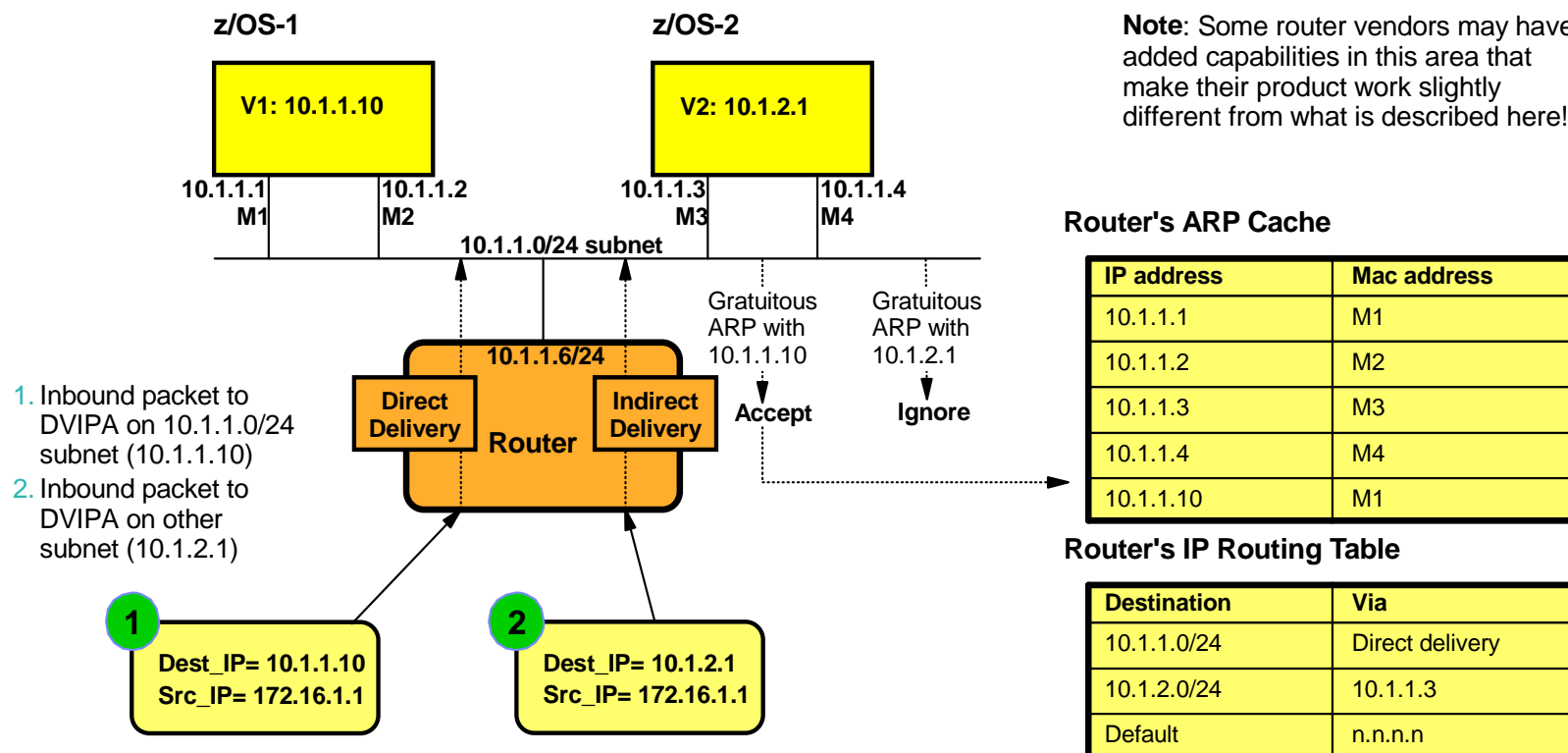
Router's ARP Cache after movement of 10.1.1.1 to PortB

IP address	Mac address
10.1.1.1	M2
10.1.1.2	M2

Example: PortA fails or is shut down

1. The z/OS TCP/IP stack moves address 10.1.1.1 to the other QDIO adapter (PortB), which is on the same network (same network prefix) as PortA was.
2. The z/OS TCP/IP stack issues a gratuitous ARP for IP address 10.1.1.1 with the MAC address of PortB (M2) over the PortB adapter
3. Downstream TCP/IP nodes on the same subnet with that IP address in their ARP cache, will update their ARP caches to point to M2 for IP address 10.1.1.1 and will thereafter send inbound packets for both 10.1.1.1 and 10.1.1.2 to MAC address M2

Some (restricted) support of dynamic VIPA without dynamic routing



z/OS VIPA addresses in a flat network configuration without dynamic routing must be allocated out of the same subnet as the directly attached network that all members of the Sysplex are attached to - in this example, the 10.1.1.0/24 subnet.

All LPARS must be attached to one and the same IP subnet via OSA ports. Network interfaces belonging to other IP subnets cannot be used for re-routing around failed OSA ports. Availability of the network to which the OSA ports are attached becomes of outmost importance and must generally be based on what is known as Layer-2 availability functions in the switches.

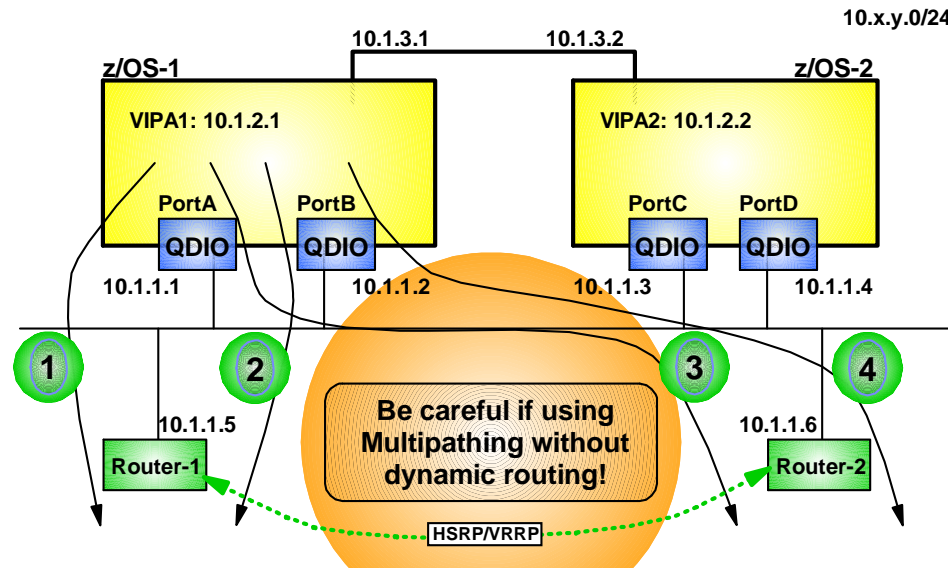
Outbound IP traffic load-balancing over multiple network interfaces and first-hop routers (MULTIPATH)

IPCONFIG MultiPath [PerConnection or PerPacket]

z/OS-1's IP Routing Table (extract)

Destination	Via
10.1.1.0/24	Direct delivery
Default	10.1.1.5 / PortA 1
Default	10.1.1.5 / PortB 2
Default	10.1.1.6 / Port A 3
Default	10.1.1.6 / Port B 4

z/OS V1R5 raised the number of multipath routes from 4 to 16.



▪ Static route definitions on z/OS:

- If an adapter fails in such a way that z/OS TCP/IP gets informed, it will skip over the corresponding entries from the routing table
- If one of the first-hop routers loses its connection to the backbone network or if it "dies" - z/OS TCP/IP doesn't know anything about it since it doesn't participate in dynamic routing updates - and it will continue to attempt to use the corresponding routing table entries - connections will time out, UDP packets will be lost, etc.
- If the two routers deploy VRRP or HSRP between them on the interfaces towards the z/OS systems, then the fact that one of them turns into a black hole can be hidden for z/OS - z/OS continues to send packets to both first-hop addresses, they are just both serviced by the one surviving router

▪ Dynamic routing updates:

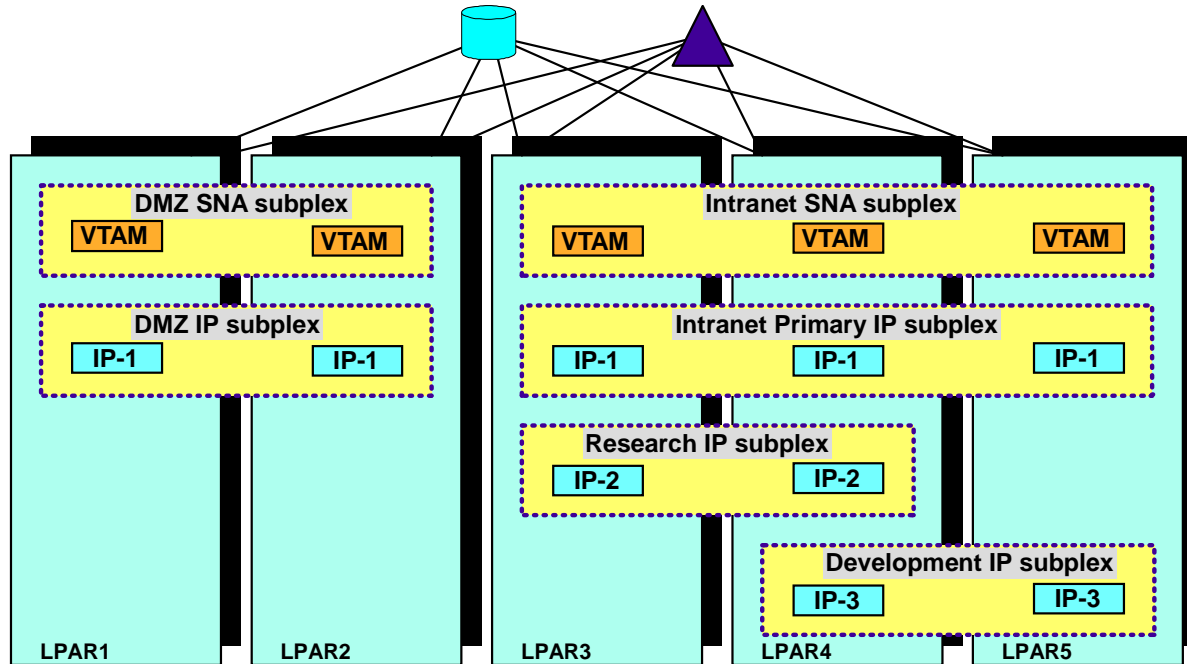
- z/OS TCP/IP will know both if the adapter itself fails or if the first-hop router fails - and dynamically update the routing table entries and recover from the router outage.
- Only OSPF supports multiple routes to the same destination - RIPv1 and RIPv2 do not.

Sysplex and Network Topology Considerations

Network Subplex support



Networking sub-plexing within a z/OS Sysplex



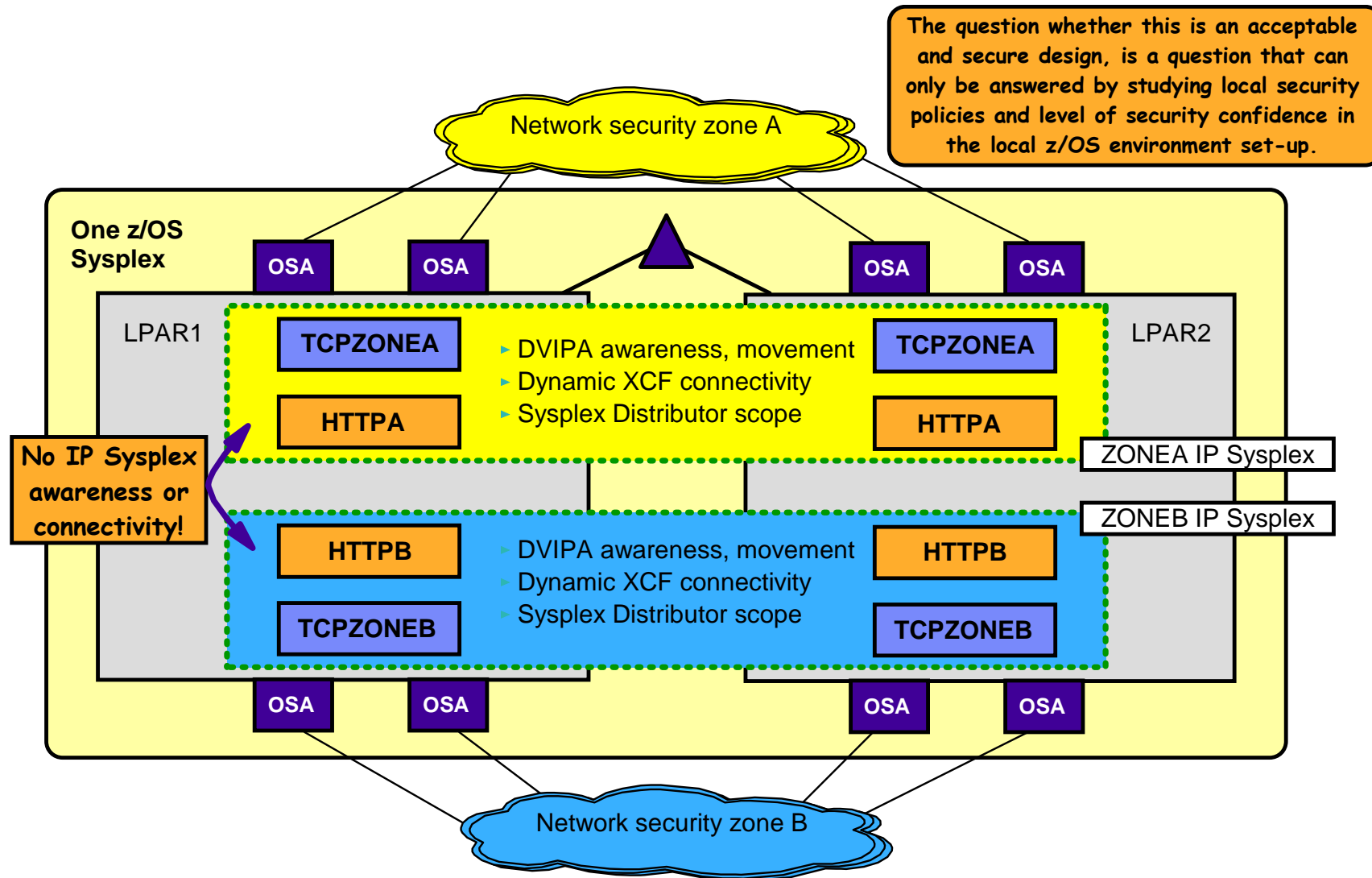
Networking Subplexing
-
Sysplex partitioning from a network perspective

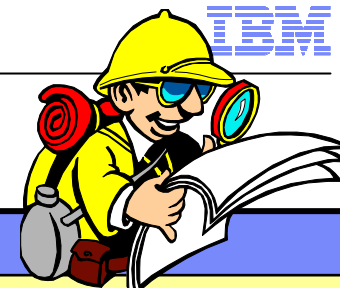
➤ Networking subplex scope:

- ▶ VTAM Generic Resources (GR) and Multi-Node Persistent Session (MNPS) resources
- ▶ Automatic connectivity - IP connectivity and VTAM connectivity over XCF (including dynamic IUTSAMEH and dynamic HiperSockets based on Dynamic XCF for IP)
 - HiperSockets VLANID support also added as part of this support
- ▶ IP stack IP address (including dynamic VIPA) awareness and visibility
- ▶ Dynamic VIPA movement candidates
- ▶ Sysplex Distributor target candidates



- One SNA subplex per LPAR
- An IP subplex cannot span multiple SNA subplexes
- Different IP stacks in an LPAR may belong to different IP subplexes
- Standard RACF controls for stack access and application access to z/OS resources need to be in place.

An example of sub-plexing within a z/OS Sysplex





For more information

URL	Content
http://www.twitter.com/IBM_Commserver 	IBM z/OS Communications Server Twitter Feed
http://www.facebook.com/IBMCommserver 	IBM z/OS Communications Server Facebook Page
https://www.ibm.com/developerworks/mydeveloperworks/blogs/IBMCommserver/?lang=en	IBM z/OS Communications Server Blog
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.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/atmastr.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
http://www.ibm.com/developerworks/rfe/?PROD_ID=498	RFE Community for z/OS Communications Server
https://www.ibm.com/developerworks/rfe/execute?use_case=tutorials	RFE Community Tutorials

For pleasant reading

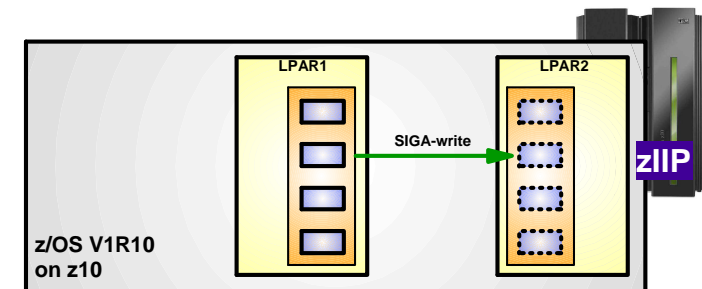
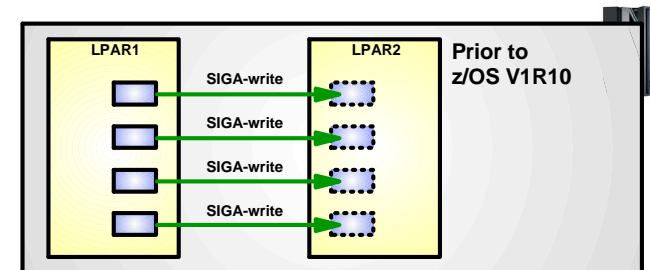
Sysplex and Network Topology Considerations

Appendix



z10 HiperSockets processing enhancements for large messages

- HiperSockets is a unique System z intra-CEC connectivity technology
- Hipersockets can now move multiple output data buffers in one write operation
 - Reduces CPU utilization (up to 10% reduction)
 - For large outbound messages (> 32K)
 - Enabled via
 - GLOBALCONFIG IQDMULTIWRITE
- For more CPU savings, the processing for these large writes can be offloaded to a zIIP
 - The zIIP Assisted HiperSockets for Large Messages capability helps lower processor utilization for handling of large outbound messages, and helps make new workload traffic from XML, JAVA, and other languages, as well as general bulk data transfers more attractive on the platform.
 - Enabled via
 - GLOBALCONFIG ZIIP IQDIOMULTIWRITE
- Netstat devlinks report will show the status per HiperSockets interface
- HiperSockets Multiple Write support (not including zIIP-Assist) is being retrofit to V1R9 with APARs
 - PK64880 & OA24882

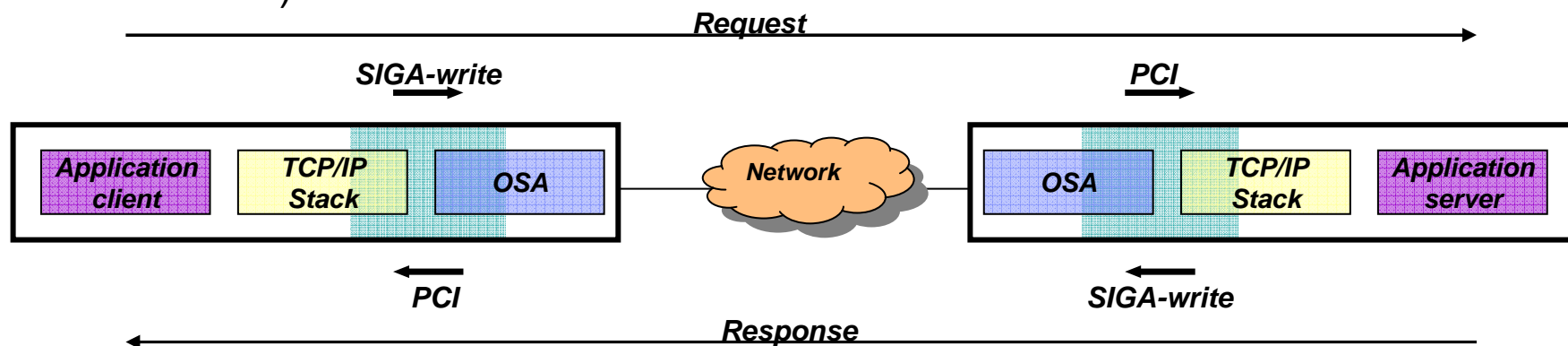


z/OS CS V1R10 on a z10 supports transfer of multiple buffers with a single SIGA-write instruction. Throughput improvements for streaming workloads.

Up to 30% general CP CPU usage reduction.

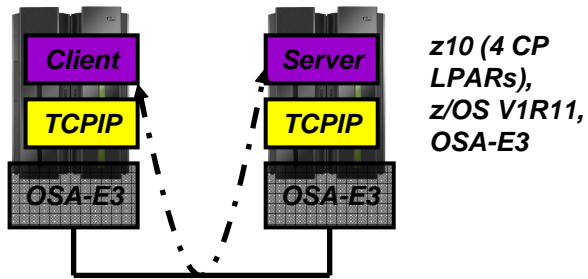
z/OS V1R11 OSA-Express optimized latency mode (OLM)

- 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”)



Note: requires new coming OSA microcode support – Refer to Preventive Service Planning (PSP) bucket for 2097DEVICE and 2098DEVICE

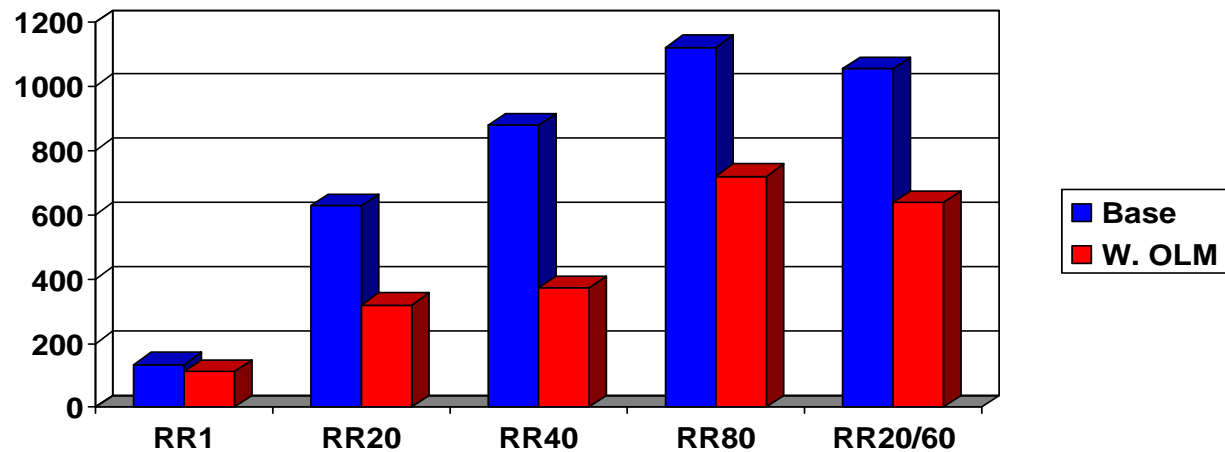
Preliminary performance indications of OLM for interactive workload



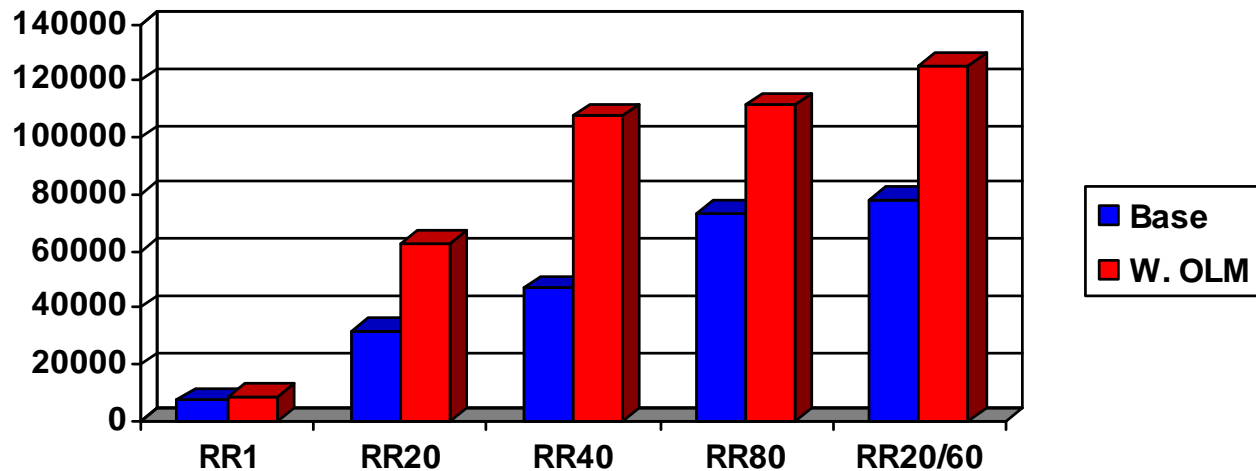
z10 (4 CP LPARs),
z/OS V1R11,
OSA-E3

- **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

End-to-end latency (response time) in Micro seconds



Transaction rate – transactions per second



Note: The performance measurements discussed in this presentation are preliminary z/OS V1R11 Communications Server numbers and were collected using a dedicated system environment with prototype code. The final results obtained in other configurations or operating system environments may vary.

OSA OLM usage

- **New OLM parameter for OSA-E3 interfaces (will only have effect when OSA microcode is at the correct level)**
 - IPAQENET
 - IPAQENET6
 - Not allowed on DEVICE/LINK
- **Enables Optimized Latency Mode for this INTERFACE only**
 - Forces INBPERF to Dynamic
 - Default is NOOLM
- **Concurrent interfaces to an OSA-Express port using OLM is limited to four**
 - If one or more interfaces operate OLM, only four total interfaces allowed
 - All four interfaces can operate in OLM
 - An interface can be:
 - Another LPAR using the OSA-Express port
 - Another VLAN defined for this OSA-Express port
 - Another protocol (IPv4 or IPv6) interface defined for this OSA-Express port
 - Another stack on the same LPAR using the OSA-Express port
 - Any stack activating the OSA-Express Network Traffic Analyzer (OSAENTA)
- **QDIO Accelerator or HiperSockets Accelerator will not accelerate traffic to or from an OSA-Express operating in OLM**
- **How do you know OLM is working?**
 - Enable tuning statistics for the OSA-Express3 device
 - Look for Message 2316I and 2317I to be non-zero
 - Look for outbound traffic on Queue 1
 - If not, verify WLMRIORITYQ and SETSUBNETPRIOTOSMASK



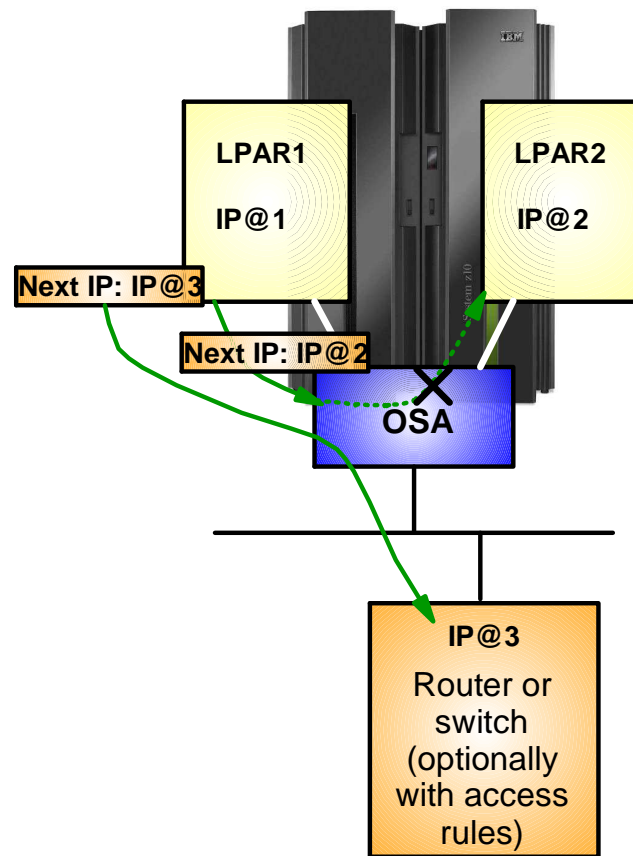
z/OS V1R11 OSA interface isolation

- 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
 - Would typically be combined by VLAN use to achieve network isolation
- OSA local routing can in some scenarios be seen as a security exposure

- Depends on OSA MCL update
 - Refer to Preventive Service Planning (PSP) buckets for latest information
 - 2094DEVICE, 2096DEVICE, 2097DEVICE, or 2098DEVICE

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.

Sysplex and Network Topology Considerations

Fast Response Cache Accelerator (FRCA)



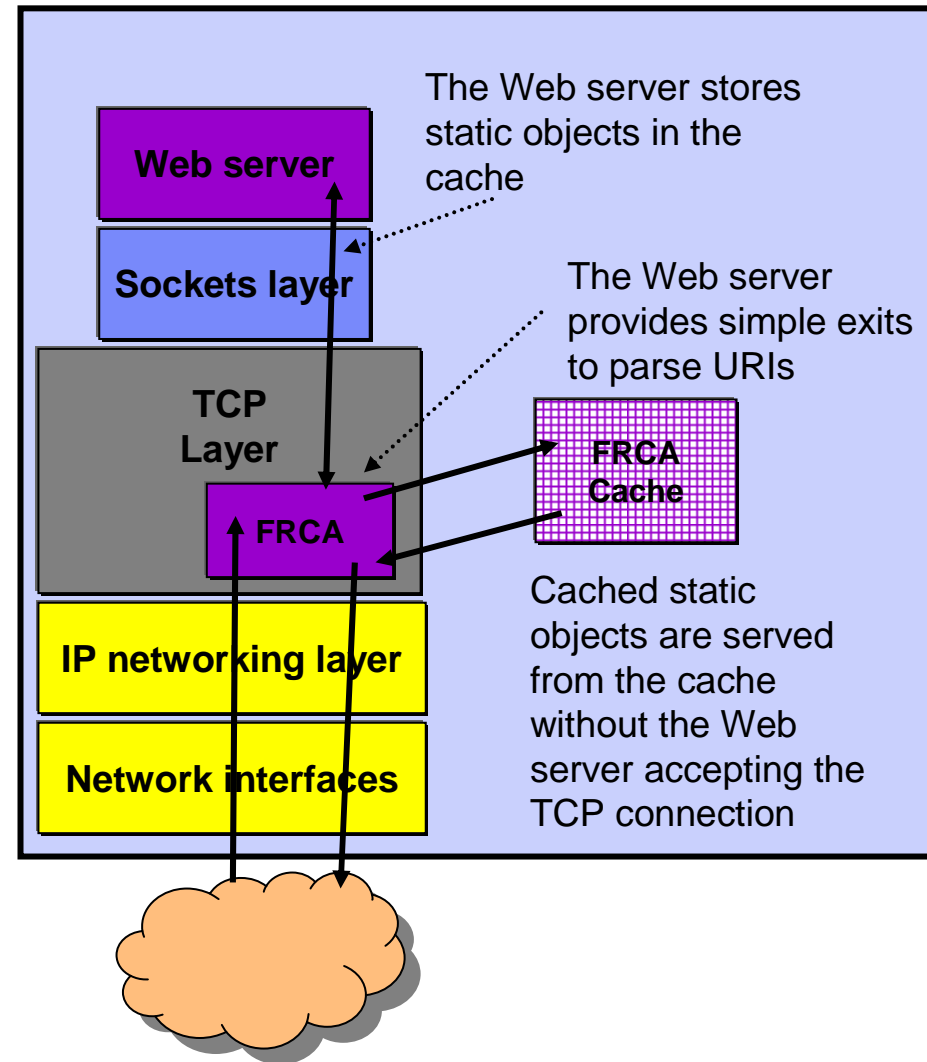
Fast Response Cache Accelerator - FRCA

- **FRCA provides a hybrid Web server environment**
 - Partly Web server
 - Partly TCP/IP stack

- **The Web server loads a set of TCP/IP stack exits to parse received data**
 - Allows FRCA to work with any protocol, not just HTTP

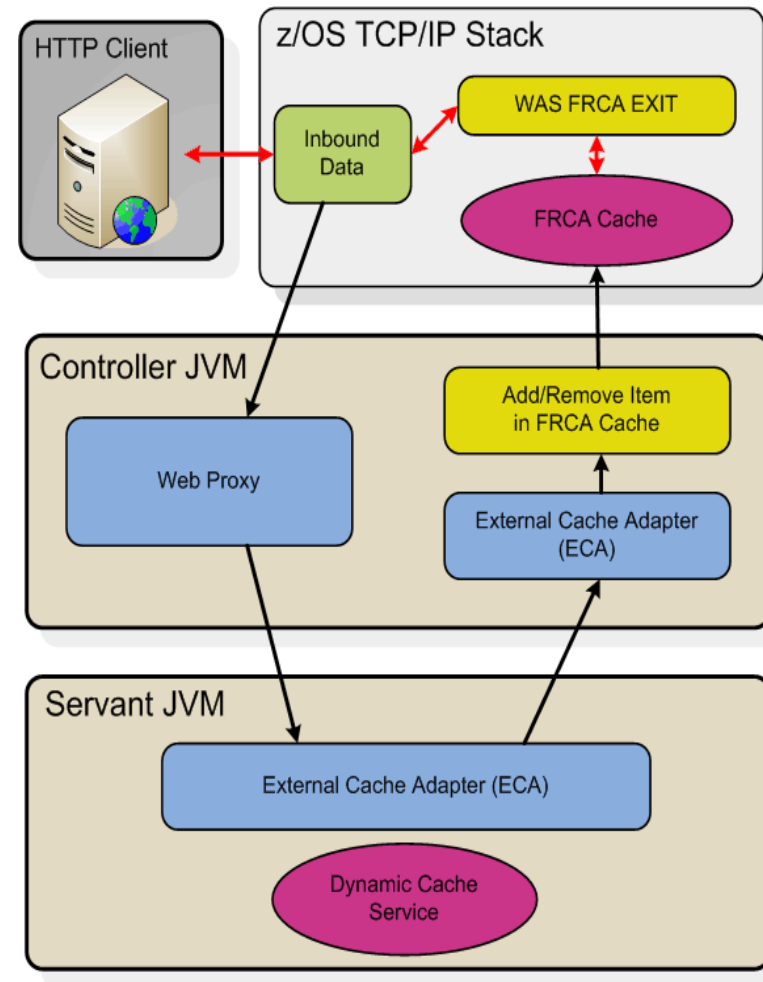
- **Web pages are cached within the TCP/IP stack**
 - Requests are handled without traversing the full protocol stack up to the Web server
 - Significant performance improvements when compared to the Web server handling all requests

- **Currently used by**
 - z/OS HTTP server
 - WebSphere Application Server for z/OS V7



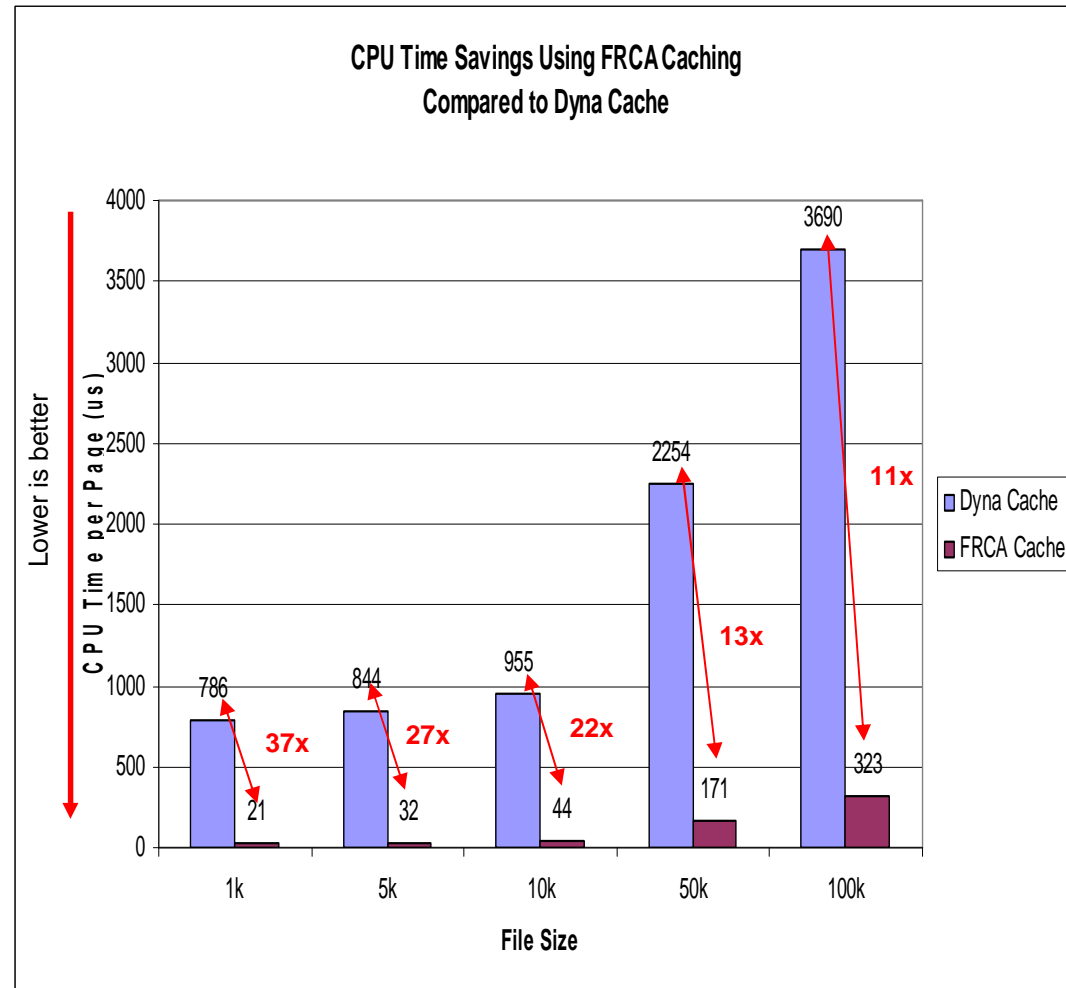
Fast Response Cache Accelerator (FRCA) use by WebSphere Application Server - overview

- **The FRCA cache is an HTTP cache that is maintained by TCP/IP**
- **Cached responses can be served with high performance using a minimal amount of CPU cycles**
 - Serve static requests from the FRCA cache
 - Provide equivalent performance on WAS as is possible with the FRCA cache on the web server
 - Serve dynamic content from the FRCA cache
 - Serve the same content that the Dynamic Cache serves but serve it from the FRCA cache
 - Record HTTP Access Log entries for requests served from the FRCA cache
- **WebSphere Application Server V7 exploits the FRCA cache**



z/OS: CPU Cost Savings Using FRCA Cache Compared to Dyna Cache

- The amount of CPU time needed to process a request is dramatically reduced using FRCA as compared to Dyna Caching
 - Dyna Cache is 37 times more costly than FRCA caching for 1k file sizes
 - Larger file sizes, 5k to 100k, Dyna Cache is 27 to 11 times more costly



System Configuration
 Workload: Simple File Server App
 SUT: IBM z10 Processor (model 2097 – 720) 4 x 4.4 GHz, 32 GB Real
 Driver: x/335 model 8676-21X, 4x3.06 GHz, 2 GB RAM

Note: The performance measurements discussed in this presentation were collected using a dedicated system environment. The results obtained in other configurations or operating system environments may vary.

z/OS V1R11: improved FRCA backlog visibility

- Display TCPIP,NETSTAT,ALL to see the new ServerBacklog and FRCABacklog fields

```

MVS TCP/IP NETSTAT CS V1R11          TCPIP Name: TCPCS          16:49:28
Client Name: BPXOINIT                Client Id: 00000021
  Local Socket: 0.0.0.0..10007
  Foreign Socket: 0.0.0.0..0
...
...
MaximumBacklog:      0000000010
CurrentBacklog:      0000000008
  ServerBacklog:    0000000003      FRCABacklog: 0000000002
CurrentConnections: 0000000000      SEF:                100
  
```

Established: Full three-way handshake has completed.

Half-open: You have received SYN and responded with a SYN+ACK, but are waiting for the final ACK from the client

z/OS V1R11 also improves the SD server backlog health monitoring to only include the ServerBacklog when calculating TCP/IP health for the actual server

