z/OS Thru-V2R1 Communications Server Performance Functions Update - Session 15512

David Herr – dherr@us.ibm.com

IBM Raleigh, NC

Trademarks

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

AIX*
BladeCenter*
CICS*
Cognos*
DataPower*

DB2*
DFSMS
EASY Tier
FICON*
GDPS*

HiperSockets*
HyperSwap
IMS
InfiniBand*
Lotus*

MQSeries*
NetView*
OMEGAMON*
Parallel Sysplex*
POWER7*

PowerHA*
PR/SM
PureSystems
Rational*
RACF*

RMF
Smarter Planet*
Storwize*
System Storage*
System x*

System z*
System z10*
Tivoli*
WebSphere*
XIV*

zEnterprise*
z10
z10 EC
z/OS*

z/VM*
z/VSE*

* Registered trademarks of IBM Corporation

The following are trademarks or registered trademarks of other companies.

Adobe, the Adobe logo, PostScript, and the PostScript logo are either registered trademarks or trademarks of Adobe Systems Incorporated in the United States, and/or other countries.
Cell Broadband Engine is a trademark of Sony Computer Entertainment, Inc. in the United States, other countries, or both and is used under license therefrom.
Intel, Intel logo, Intel Inside, Intel Inside logo, Intel Centrino, Intel Centrino logo, Celeron, Intel Xeon, Intel SpeedStep, Itanium, and Pentium are trademarks or registered trademarks of Intel Corporation or its subsidiaries in the United States and other countries.
IT Infrastructure Library is a registered trademark of the Central Computer and Telecommunications Agency which is now part of the Office of Government Commerce.
ITIL is a registered trademark, and a registered community trademark of the Office of Government Commerce, and is registered in the U.S. Patent and Trademark Office.
Java and all Java based trademarks and logos are trademarks or registered trademarks of Oracle and/or its affiliates.
Linear Tape-Open, LTO, the LTO Logo, Ultrium, and the Ultrium logo are trademarks of HP, IBM Corp. and Quantum in the U.S. and
Linux is a registered trademark of Linus Torvalds in the United States, other countries, or both.
Microsoft, Windows, Windows NT, and the Windows logo are trademarks of Microsoft Corporation in the United States, other countries, or both.
OpenStack is a trademark of OpenStack LLC. The OpenStack trademark policy is available on the OpenStack website.
TEALEAF is a registered trademark of Tealeaf, an IBM Company.
Windows Server and the Windows logo are trademarks of the Microsoft group of countries.
Worklight is a trademark or registered trademark of Worklight, an IBM Company.
UNIX is a registered trademark of The Open Group in the United States and other countries.
* Other product and service names might be trademarks of IBM or other companies.

Notes:

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

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

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

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

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

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

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

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

Agenda

- V2R1 Performance Enhancements
- Optimizing inbound communications using OSA-Express
- Optimizing outbound communications using OSA-Express
- OSA-Express4
- z/OS Communications Server Performance Summaries

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.

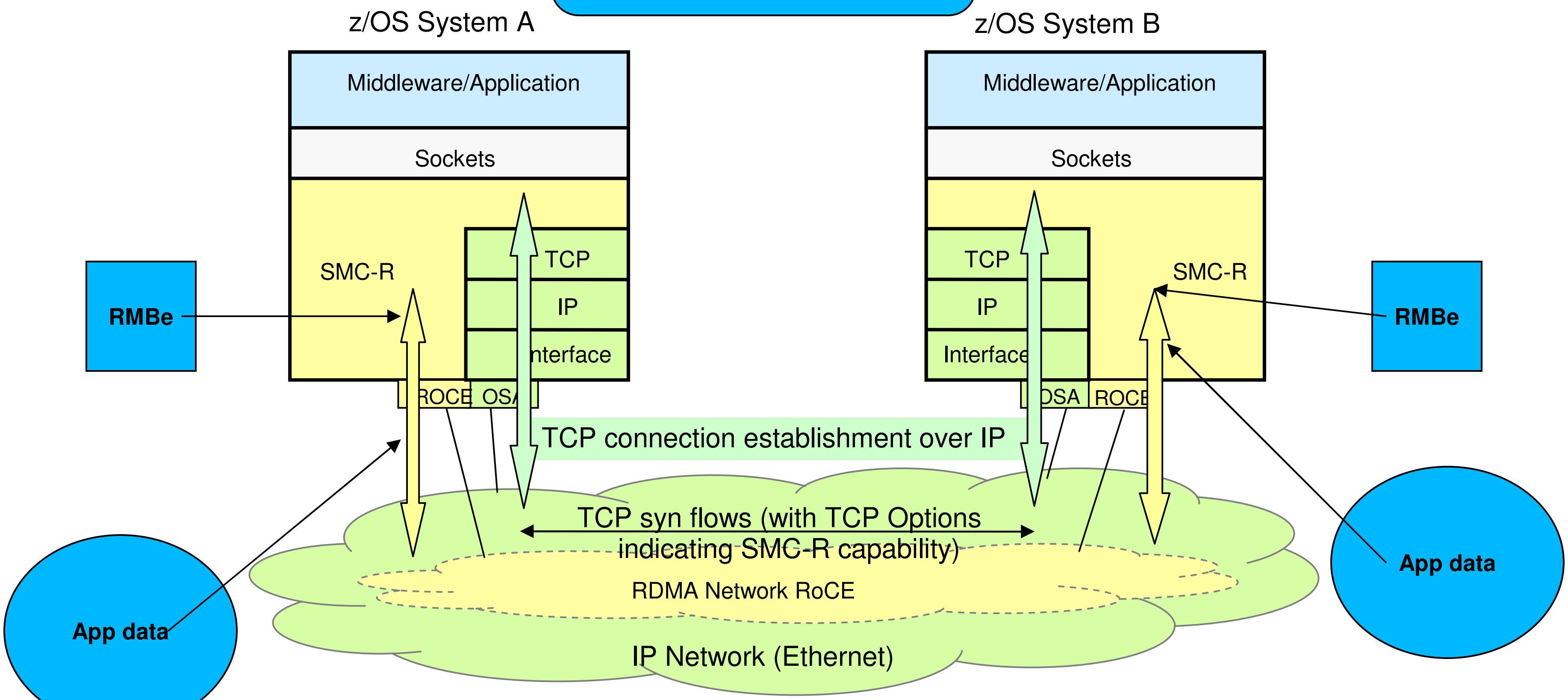
V2R1 Performance Enhancements

Shared Memory Communications – Remote (SMC-R)

V2R1

SMC-R Background

Both TCP and SMC-R “connections” remain active



TCP connection establishment over IP

TCP syn flows (with TCP Options indicating SMC-R capability)

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

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

SMC-R - RDMA

V2R1

- Key attributes of RDMA
  - Enables a host to read or write directly from/to a remote host's memory ***without*** involving the remote host's CPU
    - By registering specific memory for RDMA partner use
    - **Interrupts still required for notification (i.e. CPU cycles are not completely eliminated)**
  - Reduced networking stack overhead by using streamlined, low level, RMDA interfaces
  - Key requirements:
    - A reliable "lossless" network fabric (LAN for layer 2 data center network distance)
    - An RDMA capable NIC (RNIC) and RDMA capable switch

SMC-R - Solution

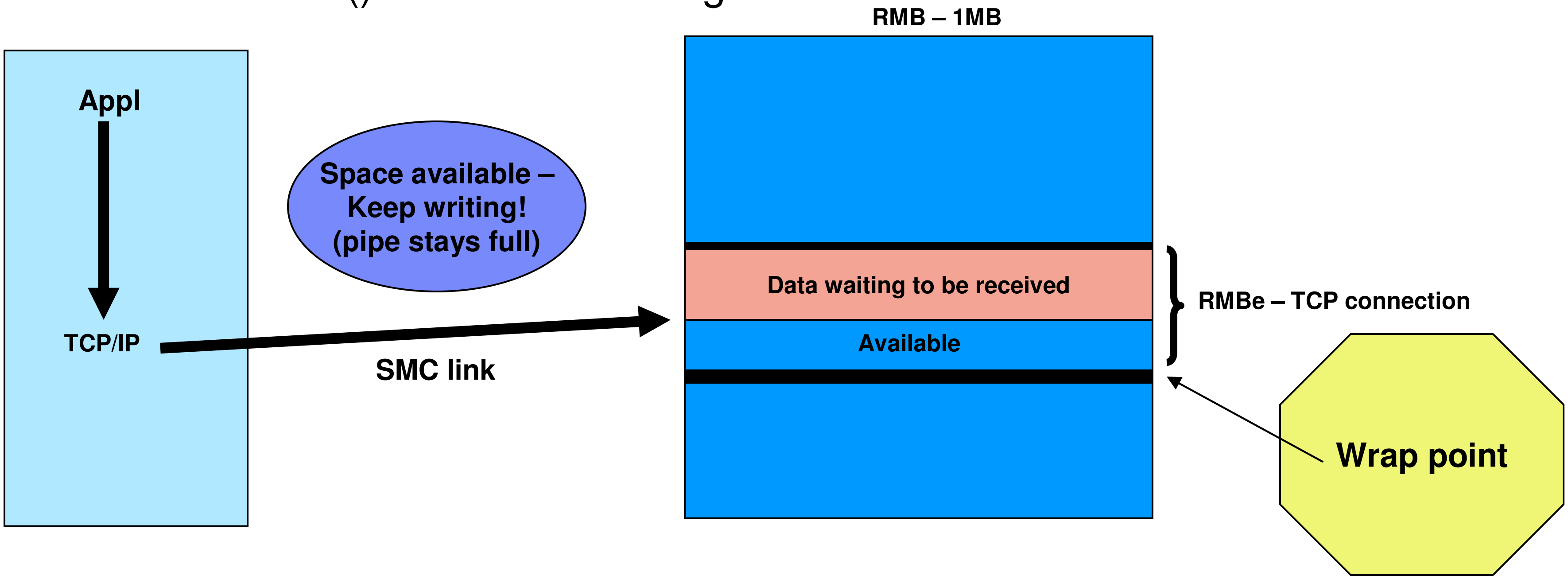
V2R1

- Shared Memory Communications over RDMA (SMC-R) is a protocol that allows *TCP sockets* applications to transparently exploit RDMA (RoCE)
- SMC-R is a “hybrid” solution that:
  - Uses TCP connection (3-way handshake) to establish SMC-R connection
  - Each TCP end point exchanges TCP options that indicate whether it supports the SMC-R protocol
  - SMC-R “rendezvous” (RDMA attributes) information is then exchanged within the TCP data stream (similar to SSL handshake)
  - Socket application data is exchanged via RDMA (write operations)
  - TCP connection remains active (controls SMC-R connection)
  - This model preserves many critical existing operational and network management features of TCP/IP

SMC-R – Role of the RMBe (buffer size)

V2R1

- The RMBe is a slot in the RMB buffer for a specific TCP connection
    - Based on TCPRCVBufrsize – NOT equal to
    - Can be controlled by application using setsockopt() SO_RCVBUF
    - 5 sizes – 32K, 64K, 128K, 256K and 1024K (1MB)
    - Depending on the workload, a larger RMBe can improve performance
        - Streaming (bulk) workloads
        - Less wrapping of the RMBe = less RDMA writes
        - Less frequent “acknowledgement” interrupts to sending side
        - Less write() blocks on sending side


RMB – 1MB
Appl
TCP/IP
Space available – Keep writing! (pipe stays full)
SMC link
Data waiting to be received
Available
RMBe – TCP connection
Wrap point

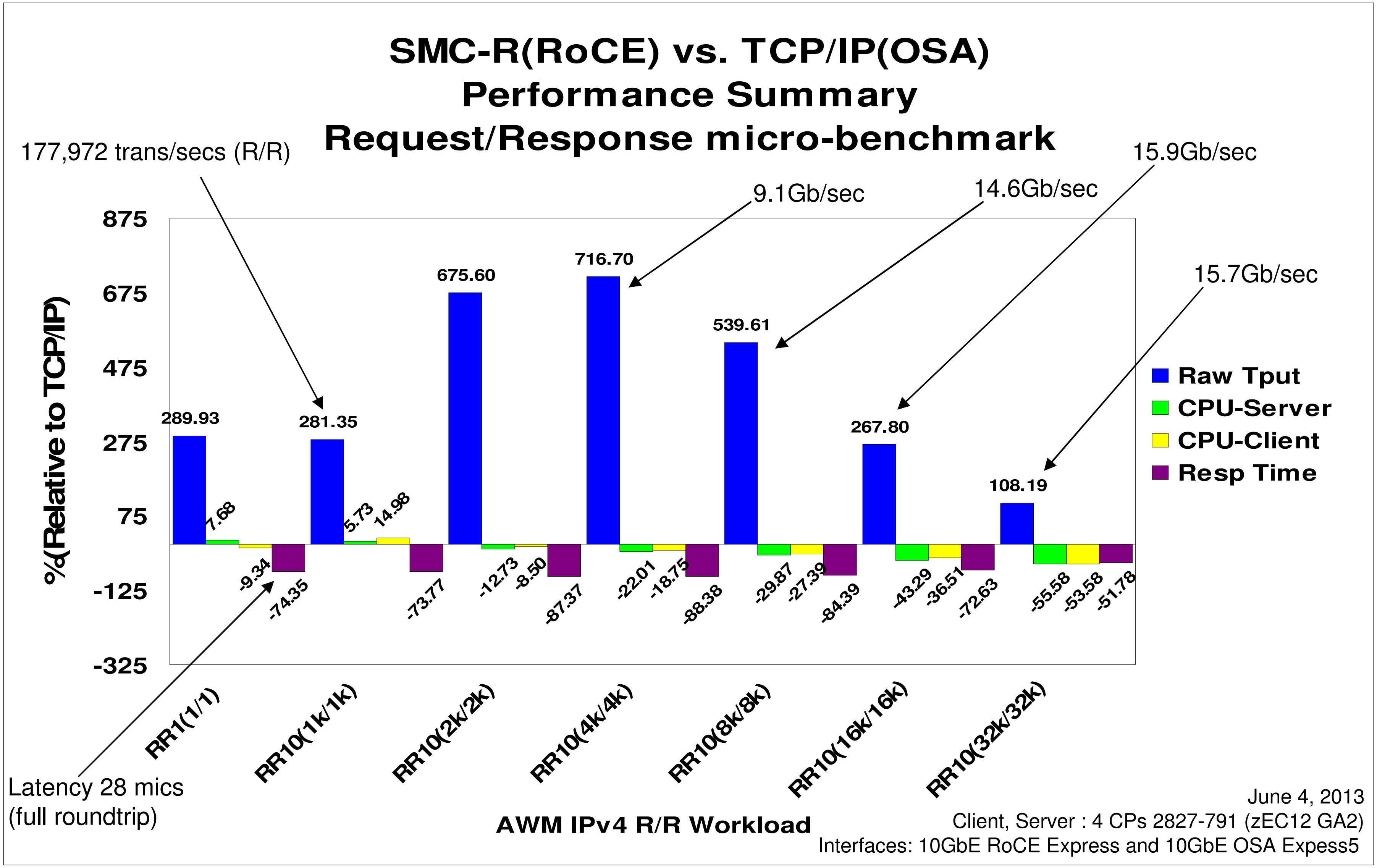
SMC-R – Micro benchmark performance results

V2R1

- Response time/Throughput and CPU improvements
- Workload:
  - Using AWM (Application Workload Modeler) to model “socket to socket” performance using SMC-R
    - AWM very lightweight - contains no application/business logic
      - Stresses and measures the networking infrastructure
      - Real workload benefits ***will be smaller*** than the improvements seen in AWM benchmarks!
  - MTU: RoCE (1K and 2K) OSA (1500 and 8000)
  - Large Send enabled for some of the TCP/IP streaming runs
  - RR1(1/1): Single interactive session with 1 byte request and 1 byte reply
  - RR10: 10 concurrent connections with various message sizes
  - STR1(1/20M): Single Streaming session with 1 byte request (Client) and 20,000,000 bytes reply (Server)
  - Used large RMBs – 1MB

SMC-R – Micro benchmark performance results

V2R1



SMC-R(RoCE) vs. TCP/IP(OSA)
Performance Summary
Request/Response micro-benchmark

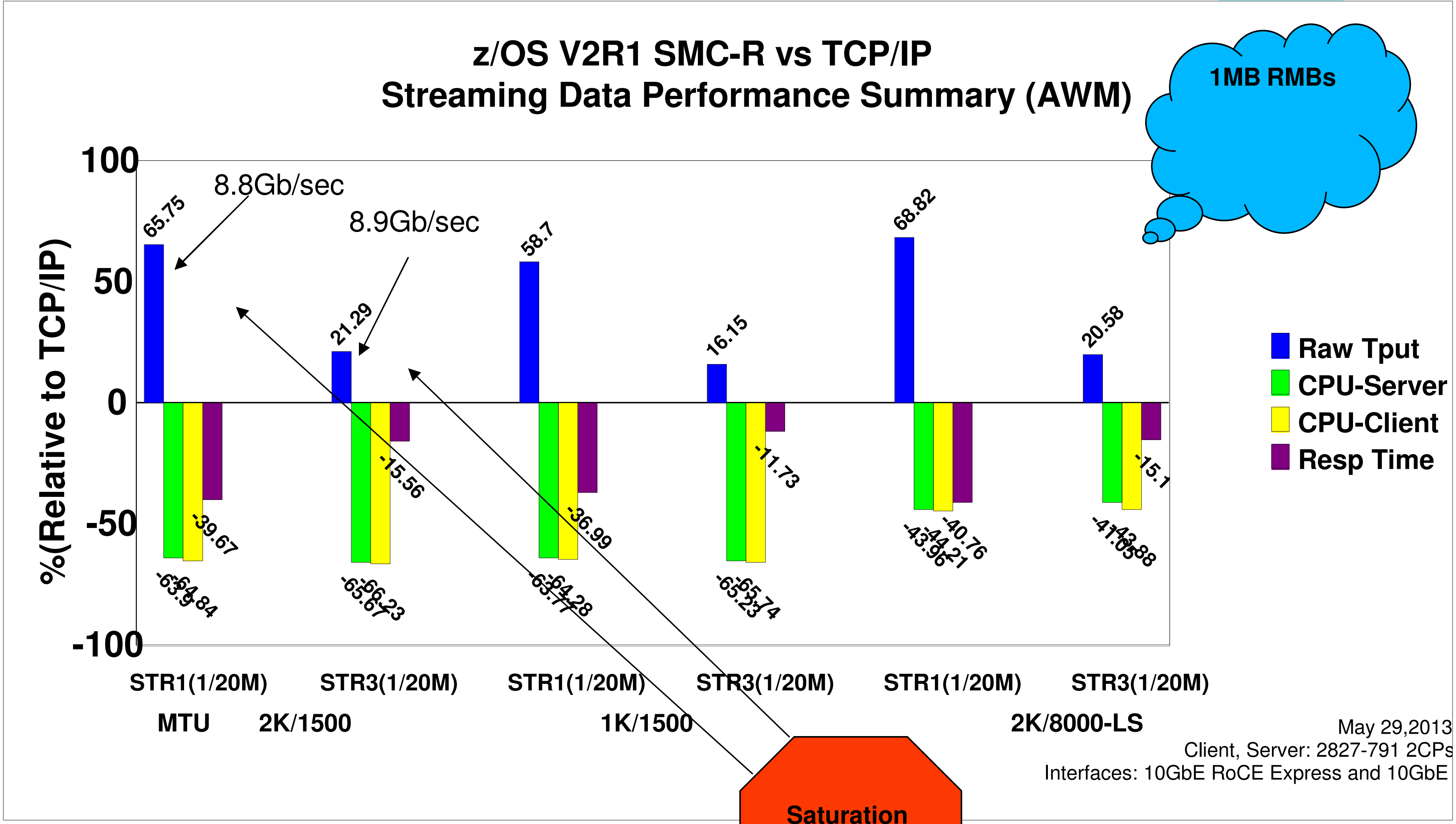
June 4, 2013
Client, Server : 4 CPs 2827-791 (zEC12 GA2)
Interfaces: 10GbE RoCE Express and 10GbE OSA Expess5

Significant Latency reduction across all data sizes (52-88%)
Reduced CPU cost as payload increases (up to 56% CPU savings)
Impressive throughput gains across all data sizes (Up to +717%)

Note: vs typical OSA customer configuration
MTU (1500), Large Send disabled
RoCE MTU: 1K

SMC-R – Micro benchmark performance results

V2R1



May 29,2013
Client, Server: 2827-791 2CPs
Interfaces: 10GbE RoCE Express and 10GbE

Saturation reached

Notes:

- ***Significant throughput benefits and CPU reduction benefits***
  - ***Up to 69% throuput improvement***
  - ***Up to 66% reduction in CPU costs***
- ***2K RoCE MTU does yield throughput advantages***
- **LS – Large Send enabled (Segmentation offload)**

SMC-R – Micro benchmark performance results

V2R1

- Summary –
  - Network latency for z/OS TCP/IP based OLTP (request/response) workloads reduced by up to 80%*
    - Networking related CPU consumption reduction for z/OS TCP/IP based OLTP (request/response) workloads increases as payload size increases
  - Networking related CPU consumption for z/OS TCP/IP based workloads with streaming data patterns reduced by up to 60% with a network throughput increase of up to 60%**
  - CPU consumption can be further optimized by using larger RMBe sizes
    - Less data consumed processing
    - Less data wrapping
    - Less data queuing

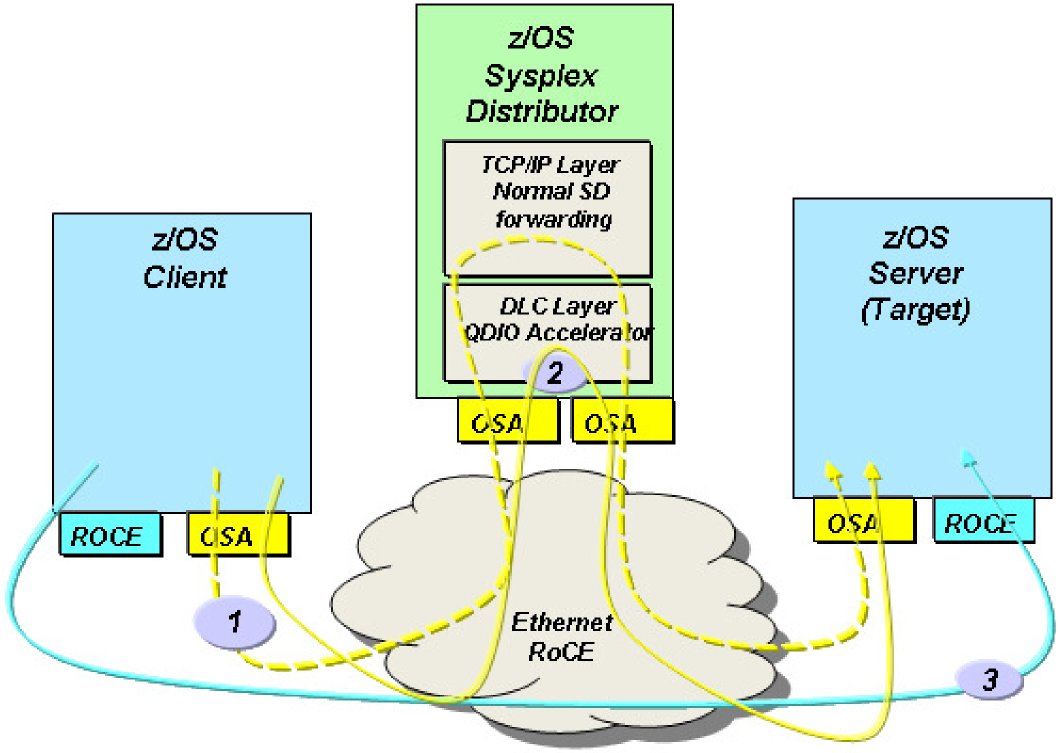
* Based on benchmarks of modeled z/OS TCP sockets based workloads with request/response traffic patterns using SMC-R vs. TCP/IP. The actual response times and CPU savings any user will experience will vary.

** Based on benchmarks of modeled z/OS TCP sockets based workloads with streaming data patterns using SMC-R vs. TCP/IP. The benefits any user will experience will vary

SMC-R – Sysplex Distributor performance results

V2R1

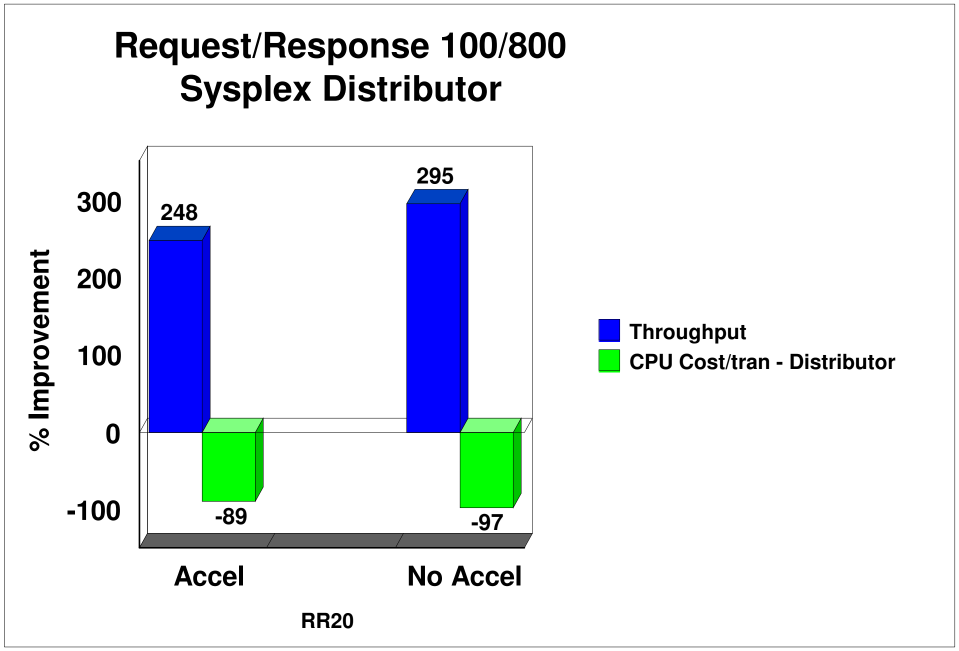
With SMC-R the distributing stack is bypassed for inbound data. Connection setup and SMC-R rendezvous packets will be the only inbound traffic going through the distributing stack.
Remember that all outbound traffic bypasses the distributing stack for all scenarios.



Line 1 - TCP/IP distributed connections without QDIO Accelerator
Line 2 - TCP/IP distributed connections utilizing QDIO Accelerator
Line 3 - SMC-R distributed connections

SMC-R – Sysplex Distributor performance results

V2R1



Results from Sysplex distributing Stack perspective

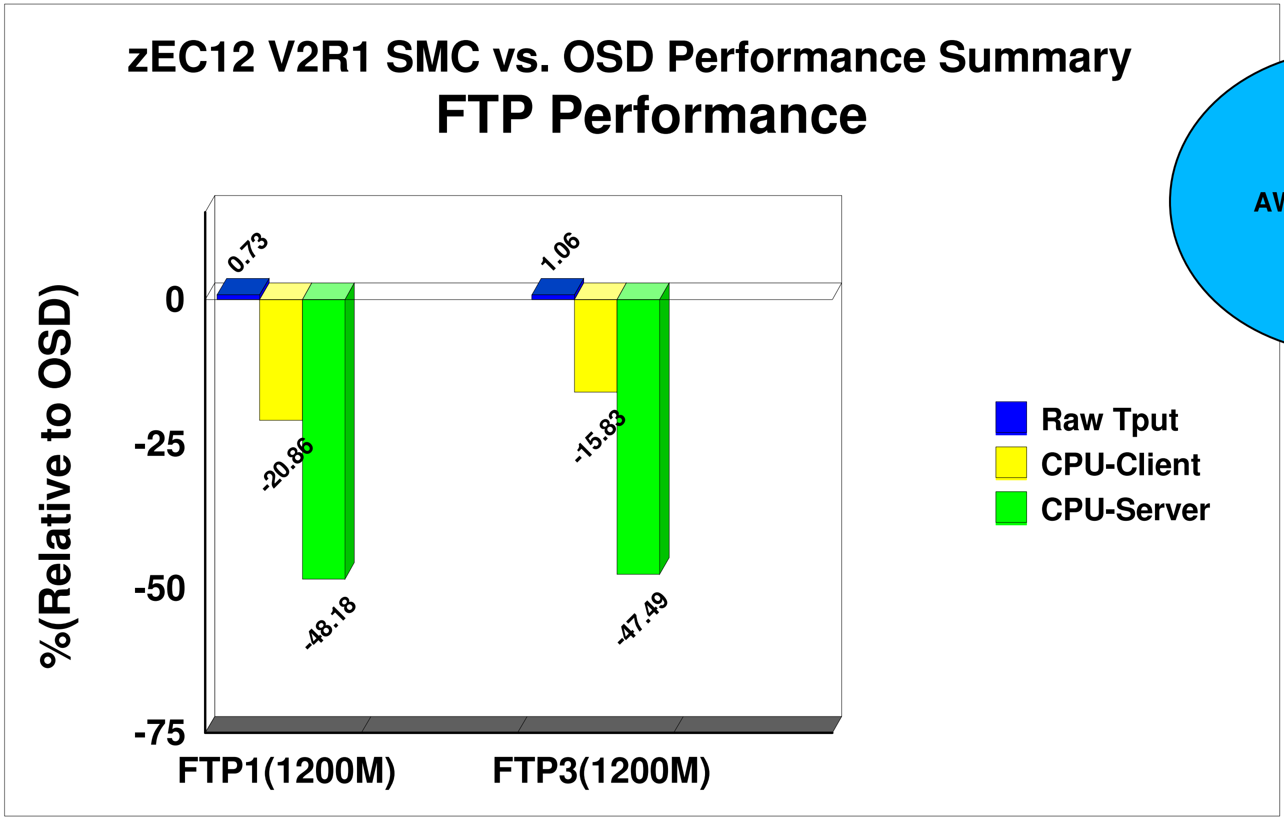
SMC-R removes virtually all CP processing on Distributing stack

250%+ throughput improvement

Workload – 20 simultaneous request/response connections sending 100 and receiving 800 bytes. Large data workloads would yield even bigger performance improvements.

SMC-R – FTP performance summary

V2R1



AWM FTP client

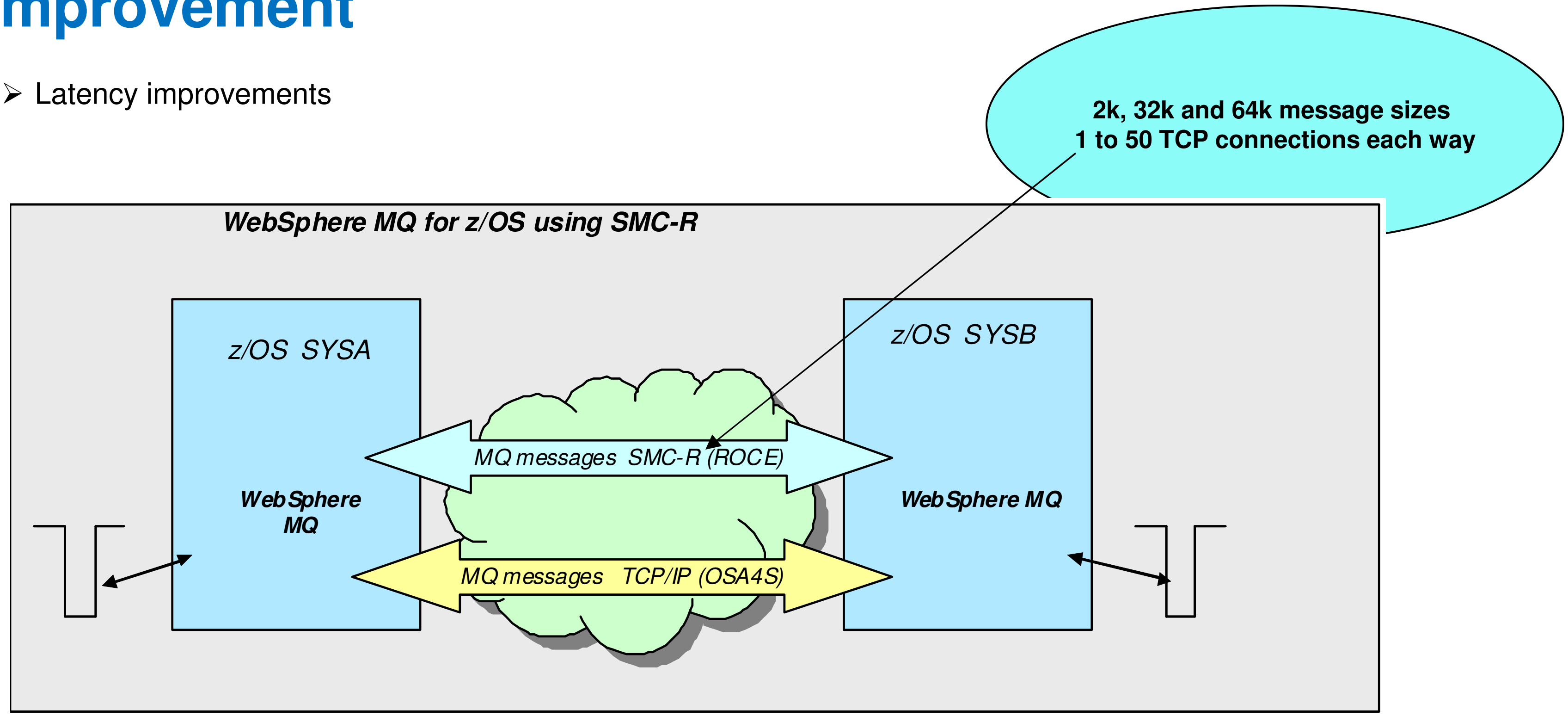
- ➢FTP binary PUTs to z/OS FTP server, 1 and 3 sessions, transferring 1200 MB data
- ➢OSD – OSA Express4 10Gb interface
- ➢Reading from and writing to DASD datasets – Limits throughput

The performance measurements discussed in this document were collected using a dedicated system environment. The results obtained in other configurations or operating system environments may vary significantly depending upon environments used.

SMC-R - WebSphere MQ for z/OS performance improvement

V2R1

- ➢ Latency improvements


2k, 32k and 64k message sizes
1 to 50 TCP connections each way
WebSphere MQ for z/OS using SMC-R
z/OS SYSA
WebSphere MQ
z/OS SYSB
WebSphere MQ
MQ messages SMC-R (ROCE)
MQ messages TCP/IP (OSA4S)

- ▪WebSphere MQ for z/OS realizes *up to a 3x increase* in messages per second it can deliver across z/OS systems when using SMC-R vs standard TCP/IP for 64K messages over 1 channel *

*Based on internal IBM benchmarks using a modeled WebSphere MQ for z/OS workload driving non-persistent messages across z/OS systems in a request/response pattern. The benchmarks included various data sizes and number of channel pairs. The actual throughput and CPU savings users will experience may vary based on the user workload and configuration.

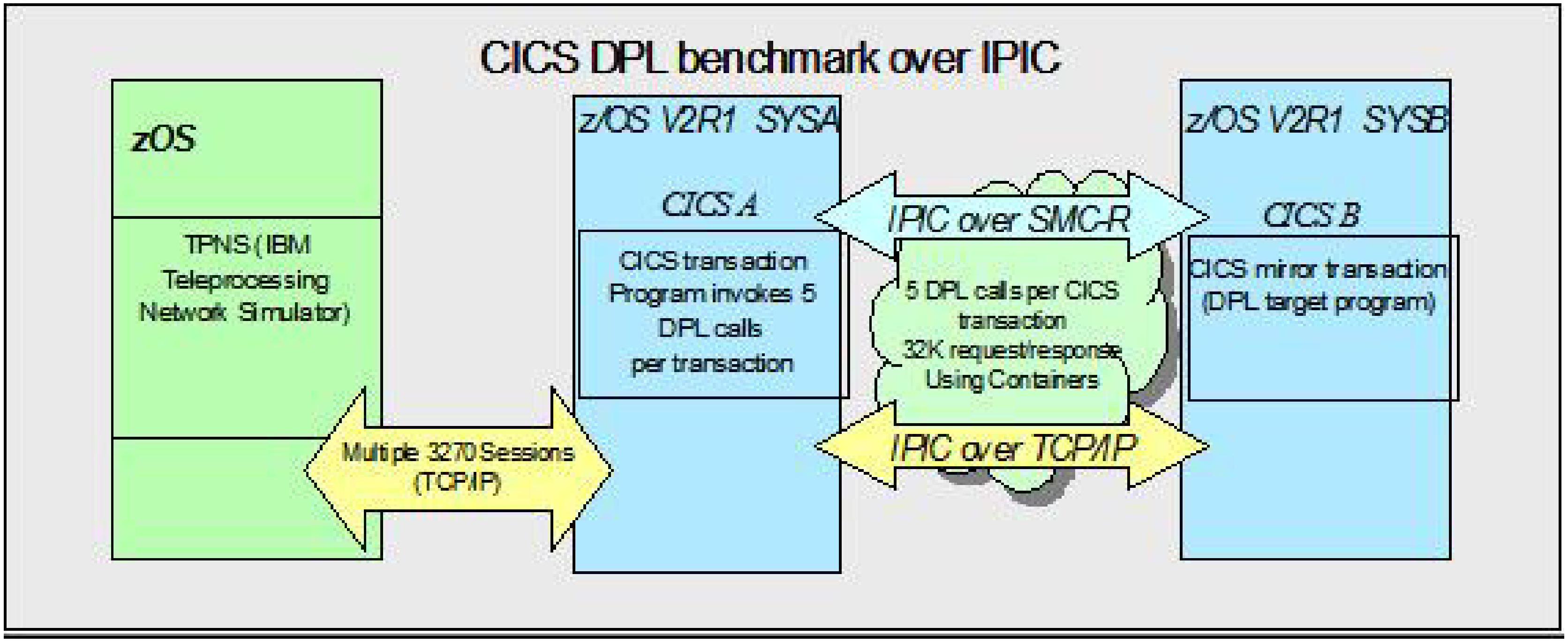
SMC-R - WebSphere MQ for z/OS performance improvement

V2R1

- Latency improvements
- Workload
  - Measurements using WebSphere MQ V7.1.0
  - MQ between 2 LPARs on zEC12 machine (10 processors each)
  - Request/Response workload
  - On each LPAR, a queue manager was started and configured with 50 outbound sender channels and 50 inbound receiver channels, with default options for the channel definitions (100 TCP connections)
  - Each configuration was run with message sizes of 2KB, 32KB and 64KB where all messages were non-persistent
    - Results were consistent across all three message sizes

SMC-R – CICS performance improvement

V2R1


CICS DPL benchmark over IPIC
zOS
TPNS (IBM Teleprocessing Network Simulator)
Multiple 3270 Sessions (TCP/IP)
z/OS V2R1 SYSA
CICS A
CICS transaction Program invokes 5 DPL calls per transaction
IPIC over SMC-R
5 DPL calls per CICS transaction
32K request/response
Using Containers
IPIC over TCP/IP
z/OS V2R1 SYSB
CICS B
CICS mirror transaction (DPL target program)

- Benchmarks run on z/OS V2R1 with latest zEC12 and new 10GbE RoCE Express feature
  – Compared use of SMC-R (10GbE RoCE Express) vs standard TCP/IP (10GbE OSA Express4S) with CICS IPIC communications for DPL (Distributed Program Link) processing
  – ***Up to 48% improvement in CICS transaction response time*** as measured on CICS system issuing the DPL calls (CICS A)
  – ***Up to 10% decrease in overall z/OS CPU consumption*** on the CICS systems

SMC-R – CICS performance improvement

V2R1

- Response time and CPU utilization improvements
- Workload - Each transaction
  - Makes 5 DPL (Distributed Program Link) requests over an IPIC connection
  - Sends 32K container on each request
  - Server program Receives the data and Send back 32K
  - Receives back a 32K container for each request

IPIC - IP Interconnectivity

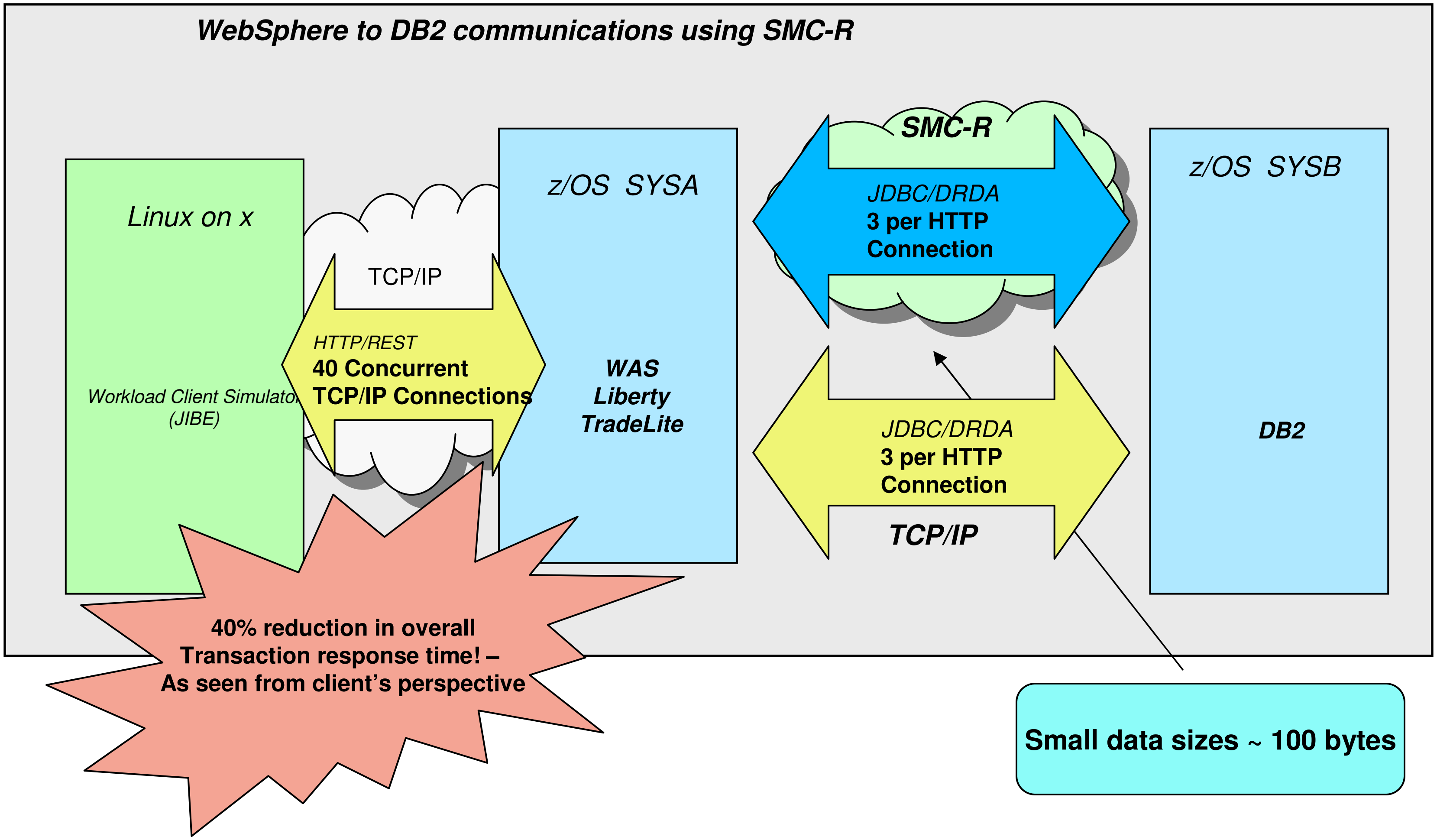
• Introduced in CICS TS 3.2/TG 7.1
•TCP/IP based communications
•Alternative to LU6.2/SNA for Distributed program calls

Note: Results based on internal IBM benchmarks using a modeled CICS workload driving a CICS transaction that performs 5 DPL calls to a CICS region on a remote z/OS system, using 32K input/output containers. Response times and CPU savings measured on z/OS system initiating the DPL calls. The actual response times and CPU savings any user will experience will vary.

SMC-R – Websphere to DB2 communications performance improvement

- Response time improvements

V2R1



WebSphere to DB2 communications using SMC-R

Linux on x

Workload Client Simulator (JIBE)

TCP/IP

HTTP/REST
40 Concurrent TCP/IP Connections

z/OS SYSA

WAS Liberty TradeLite

SMC-R

JDBC/DRDA
3 per HTTP Connection

JDBC/DRDA
3 per HTTP Connection

TCP/IP

z/OS SYSB

DB2

40% reduction in overall Transaction response time! – As seen from client's perspective

Small data sizes ~ 100 bytes

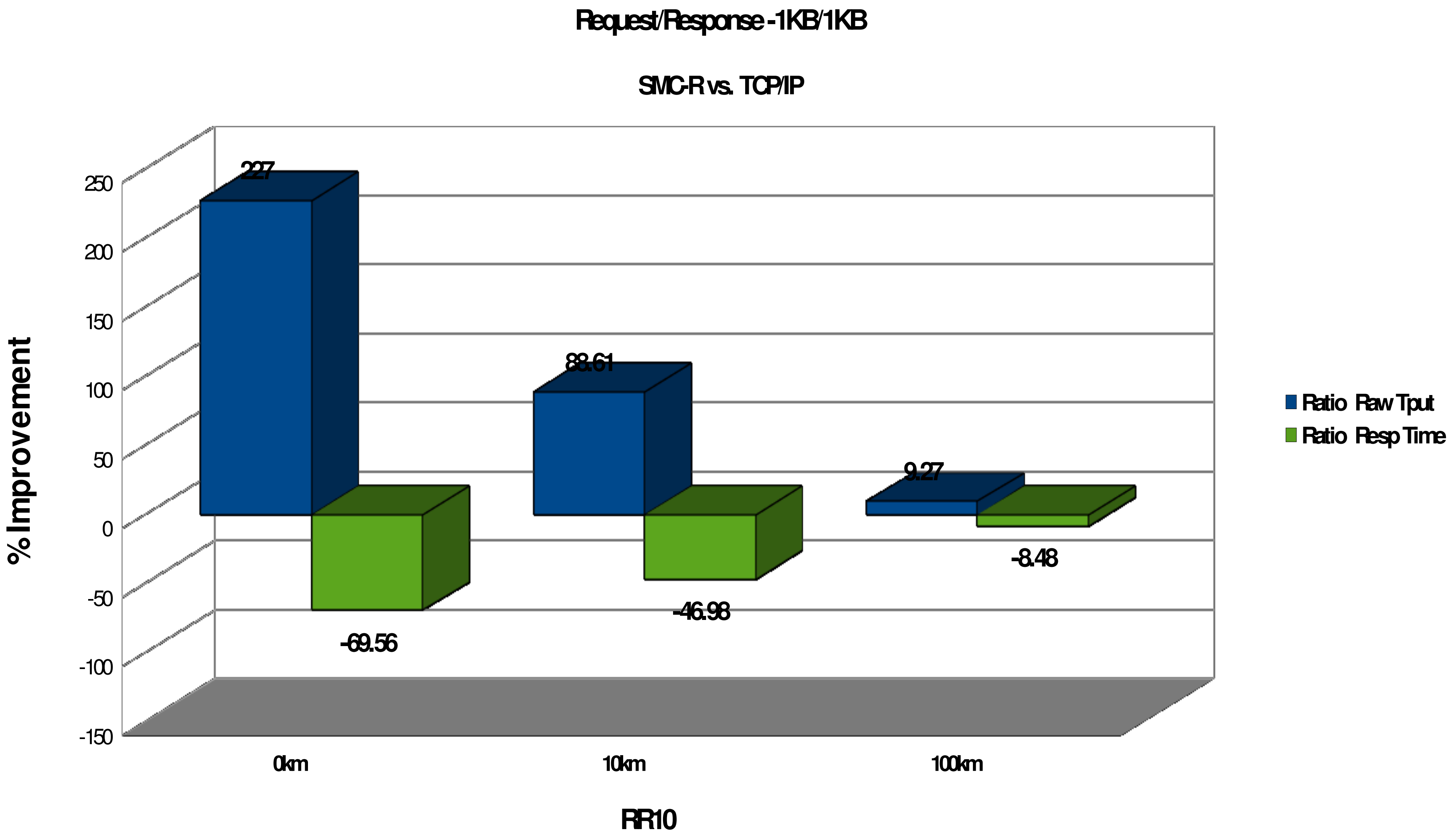
Based on projections and measurements completed in a controlled environment. Results may vary by customer based on individual workload, configuration and software levels.

SMC-R and RoCE performance benchmarks at distance

- Initial statement of support for SMC-R and RoCE Express
  - 300 meters maximum distance from RoCE Express port to 10GbE switch port using OM3 fiber cable
    - 600 meters maximum when sharing the same switch across 2 RoCE Express features
    - Distance can be extended across multiple cascaded switches
    - All initial performance benchmarks focused on short distances (i.e. same site)
- Updated testing for RoCE and SMC-R over long distances
  - IBM System z™ Qualified Wavelength Division Multiplexer (WDM) products for Multi-site Sysplex and GDPS ® solutions qualification testing updated to include RoCE and SMC-R. Two vendors already certified their DWDM solution for SMC-R and RoCE Express:

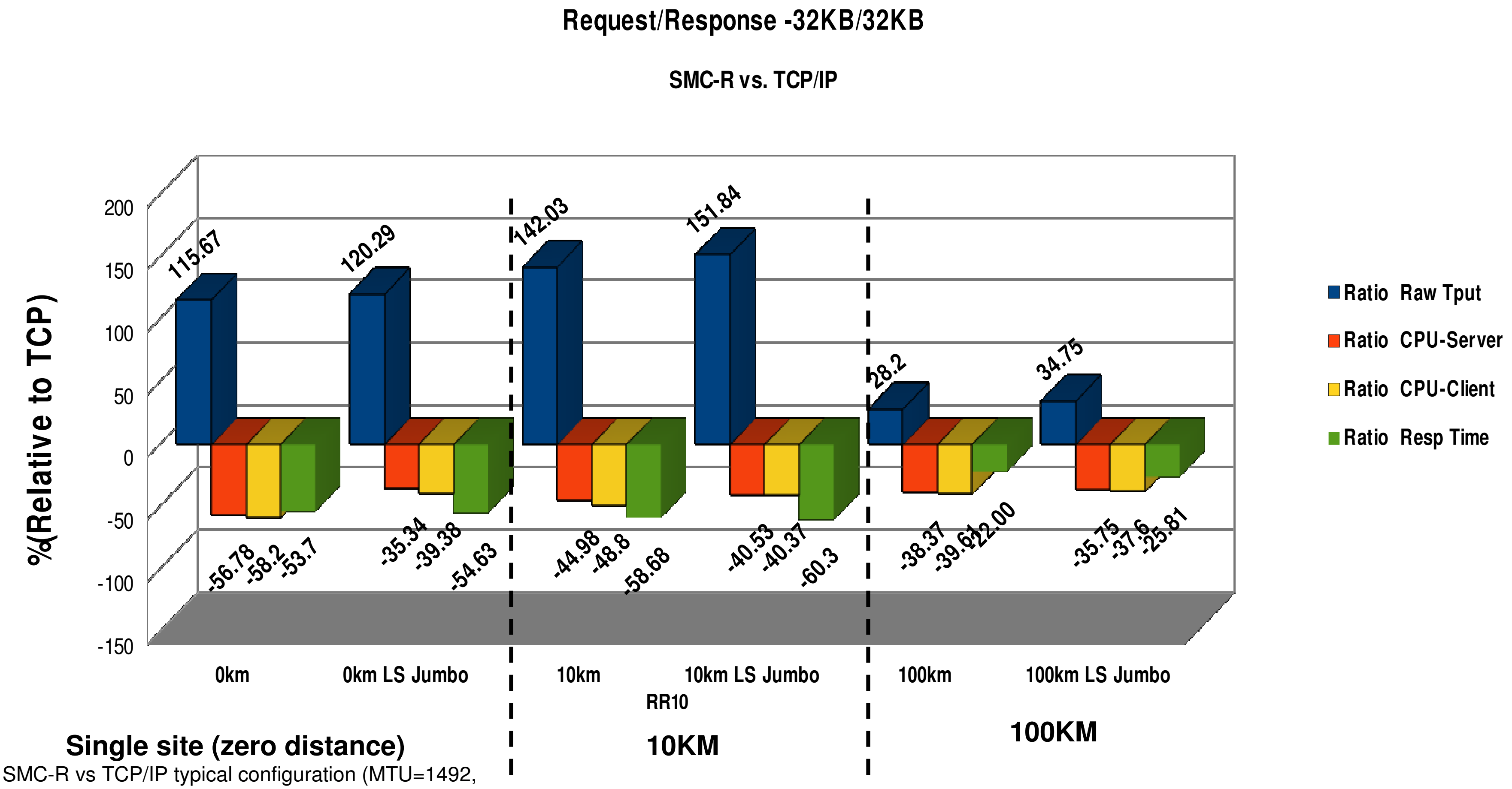
    1. Fibernet DUSAC 4800 Release 2.2b - on two client cards, the FTX-n and the FTX-10C (both cards are single port transponders). The qualification letter for this release can be found at the following link:
       - https://www-304.ibm.com/servers/resourcelink/lib03020.nsf/pages/FibernetSL?OpenDocument&pathID=
    2. Cisco 15454 Release 9.6.0.5 - on the 10 x 10G client card (15454-M-10x10G-LC) in 5:5 transponder mode.The qualification letter for this release can be found at the following link:
       - https://www-304.ibm.com/servers/resourcelink/lib03020.nsf/pages/ciscoSystemsInc?OpenDocument&pathID=
- *But how does SMC-R and RoCE perform at distance?*

SMC-R RoCE performance at distance - Request/Response Pattern (small data)



- Notes:
  - RR10(1K/1K): 10 persistent TCP connections simulating request/response data pattern, client sends 1KB request, server responds with 1KB
  - ***Substantial response time (i.e. latency) improvements at 10KM, benefits drop off at 100km***

SMC-R RoCE performance at distance – Request/Response Pattern



Single site (zero distance)

10KM

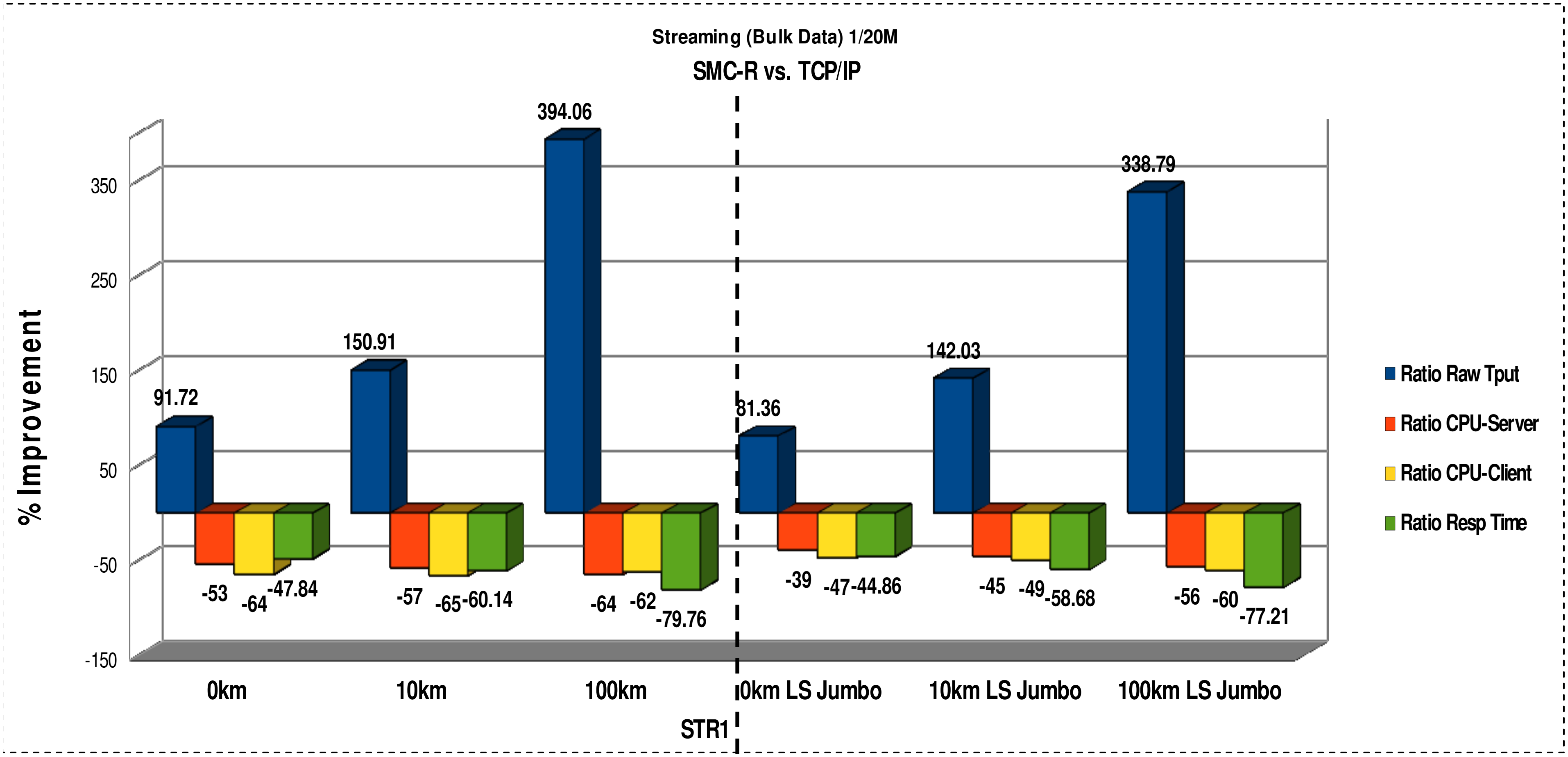
100KM

1) SMC-R vs TCP/IP typical configuration (MTU=1492, Large Send Disabled)
2) SMC-R vs TCP/IP optimal configuration optimal TCP/IP configuration (MTU=8000, Large Send Enabled)

Typical TCP/IP configuration
MTU=1492, Large Send disabled

- Notes:
  - RR10(32K/32K): 10 persistent TCP connections simulating request/response data pattern, client sends 32KB request, server responds with 32KB .
  - ***CPU benefits of SMC-R for streaming connections unaffected by distance (and in several cases better at longer distances)***
  - ***Significant response time improvement***

SMC-R RoCE performance at distance – Streaming/Bulk Data (1 session)



Typical TCP/IP configuration
MTU=1492, Large Send disabled

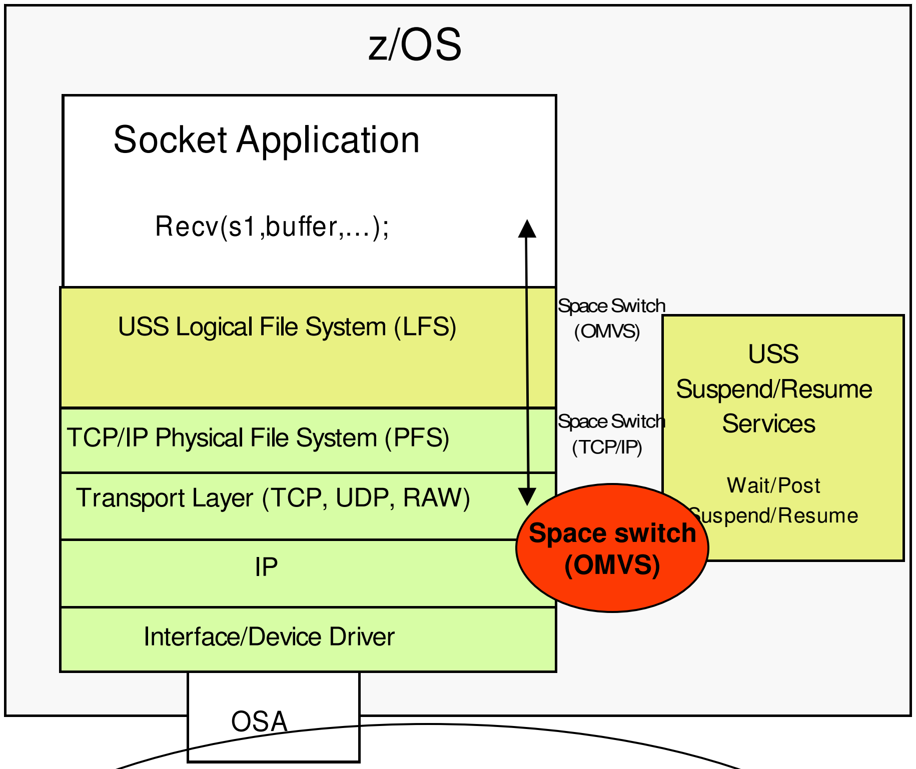
Optimal TCP/IP configuration
MTU=8000, Large Send Enabled

- Notes:
  - STR1: Single TCP connection simulating streaming data pattern, client sends 1 byte, server responds with 20MB of data.
  - ***CPU benefits of SMC-R for streaming connections unaffected by distance (and in several cases better at longer distances)***
  - ***Significant throughput improvements at distance (improving overall response time significantly)***

TCP/IP Enhanced Fast Path Sockets

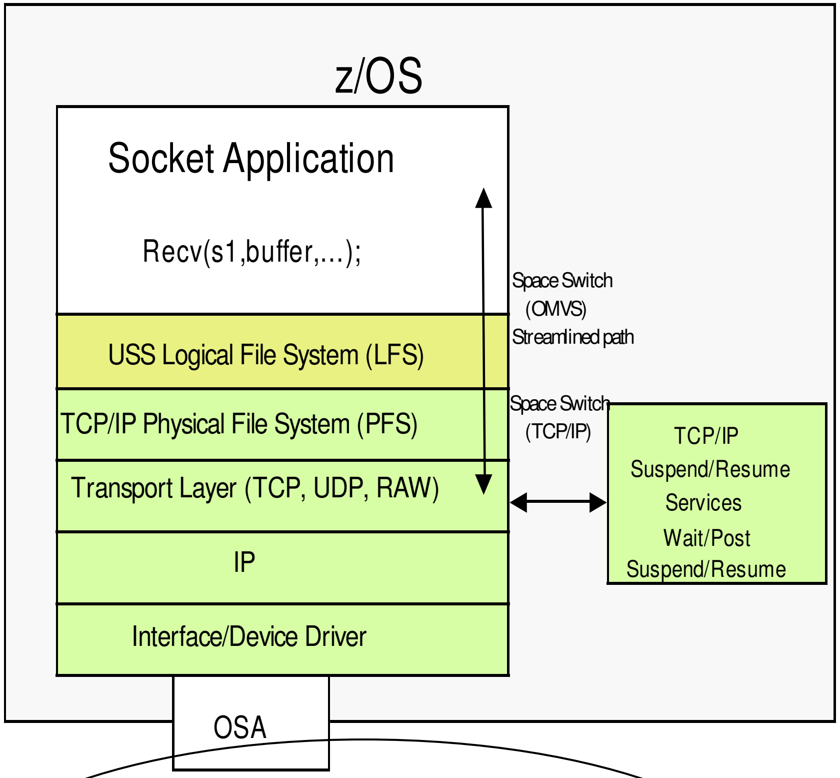
V2R1

TCP/IP sockets (normal path)



- ➢Full function support for sockets, including support for Unix signals, POSIX compliance
- ➢When TCP/IP needs to suspend a thread waiting for network flows, USS suspend/resume services are invoked

TCP/IP fast path sockets (Pre-V2R1)



- ➢Streamlined path through USS LFS for selected socket APIs
- ➢TCP/IP performs the wait/post or suspend/resume inline using its own services
- ➢Significant reduction in path length

TCP/IP Enhanced Fast Path Sockets

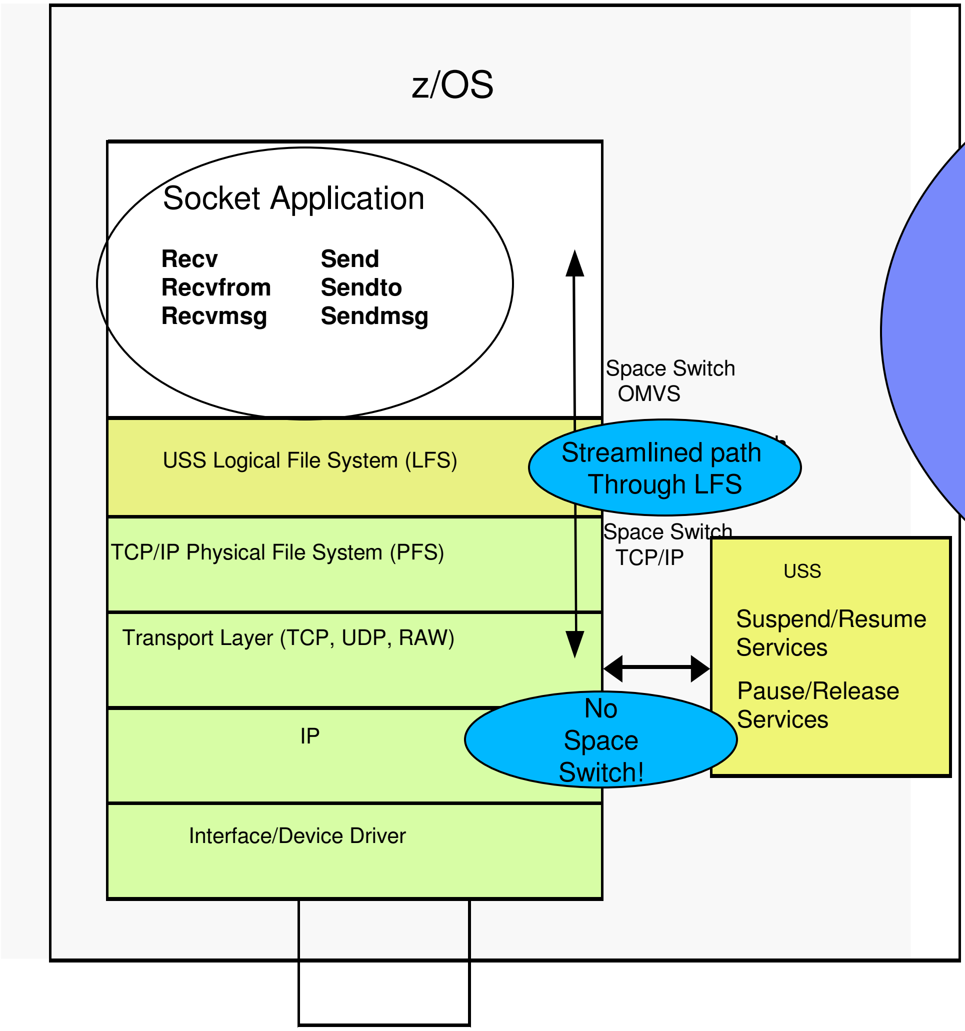
V2R1

Pre-V2R1 fast path provided CPU savings but not widely adopted:

- ➢ No support for Unix signals (other than SIGTERM)
  - ➢Only useful to applications that have no requirement for signal support
- ➢ No DBX support (debugger)
- ➢ Must be explicitly enabled!
  - ➢BPXK_INET_FASTPATH environment variable
  - ➢Iocc#FastPath IOCTL
- ➢ Only supported for UNIX System Services socket API or the z/OS XL C/C++ Run-time Library functions

TCP/IP Enhanced Fast Path Sockets

V2R1



Fast path sockets performance without all the conditions!:

- Enabled by default
- Full POSIX compliance, signals support and DBX support
- Valid for **ALL** socket APIs (with the exception of the Pascal API

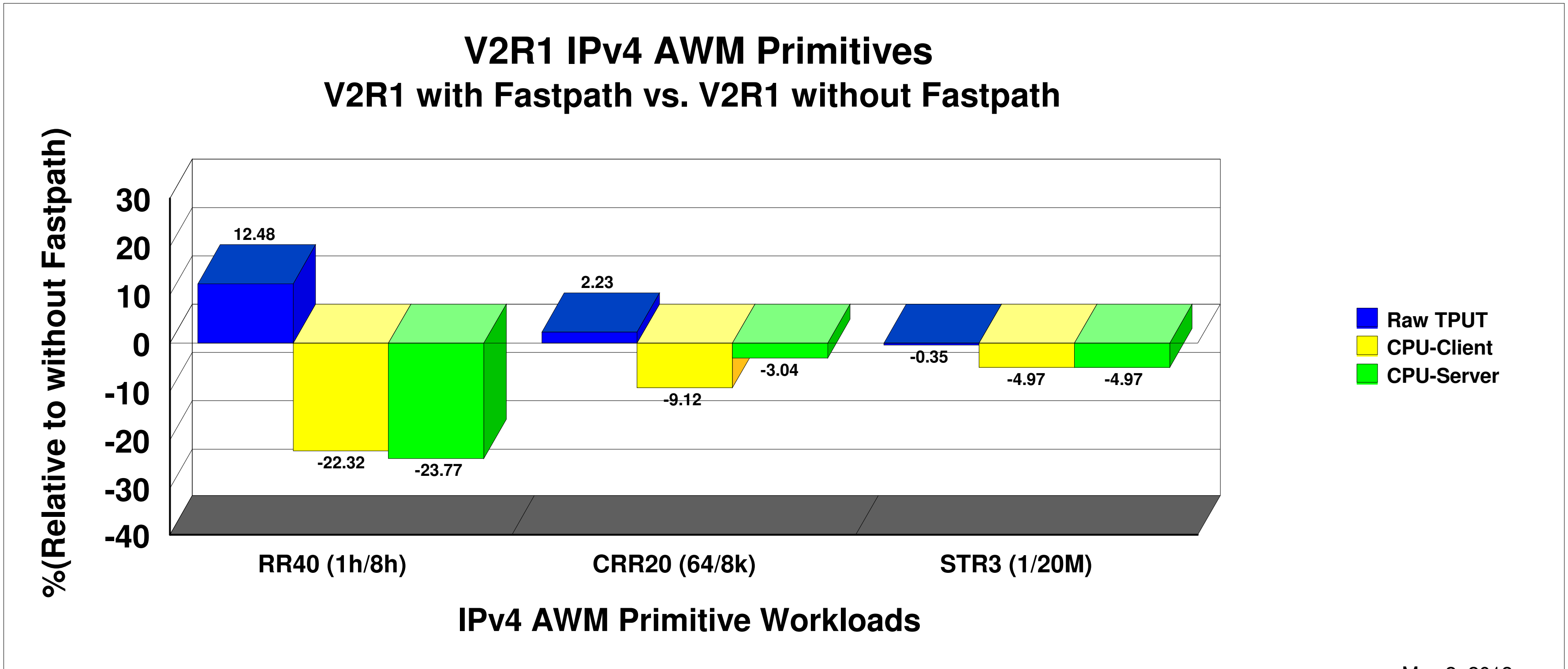
TCP/IP Enhanced Fast Path Sockets

V2R1

- No new externals
- Still supports “activating Fast path explicitly” to avoid migration issues
  - Provides performance benefits of enhanced Fast Path sockets
  - Keeps the following restrictions:
    - Does not support POSIX signals (blocked by z/OS UNIX)
    - Cannot use dbx debugger

TCP/IP Enhanced Fast Path Sockets

V2R1



May 2, 2013
Client and server LPARs: zEC12 with 6 CPs per LPAR
Interface: OSA-E4 10 GbE

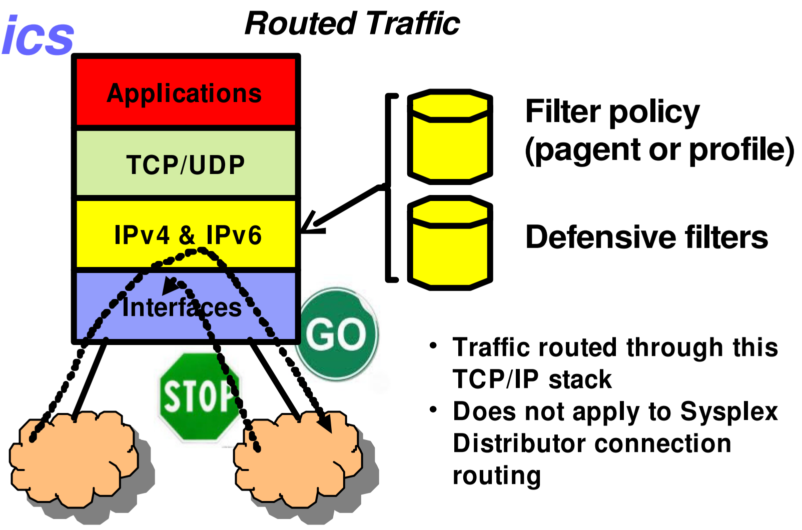
Note: The performance measurements discussed in this presentation are z/OS V2R1 Communications Server numbers and were collected using a dedicated system environment. The results obtained in other configurations or operating system environments may vary.

QDIO Accelerator coexistence with IP Filtering

V2R1

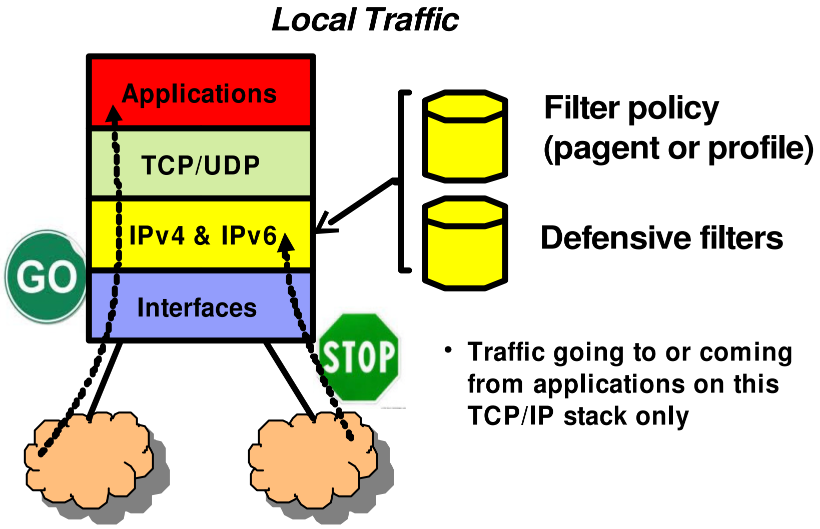
Background information: IP filtering basics

- IP filtering at the z/OS IP Layer
  - Filter rules defined based on relevant attributes
  - Used to control routed and local traffic
  - Defined actions taken when a filter rule is matched
- IP filter rules are defined in three ways:
  - TCPIP profile
  - Policy Agent
  - Defense Manager Daemon
- "Not Filtering" routed traffic means that all routed traffic is permitted by the effective filter rules



Routed Traffic

- Traffic routed through this TCP/IP stack
- Does not apply to Sysplex Distributor connection routing



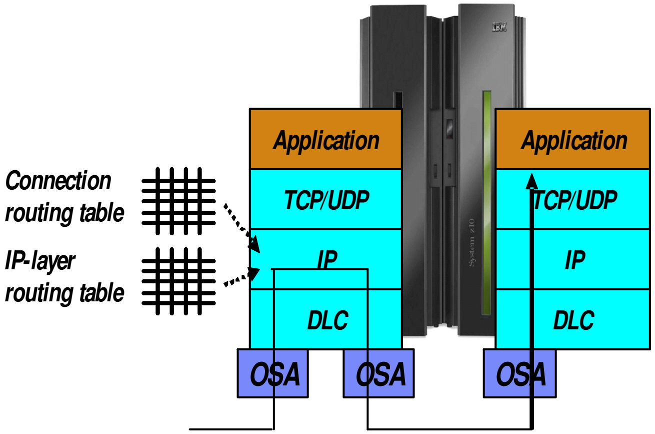
Local Traffic

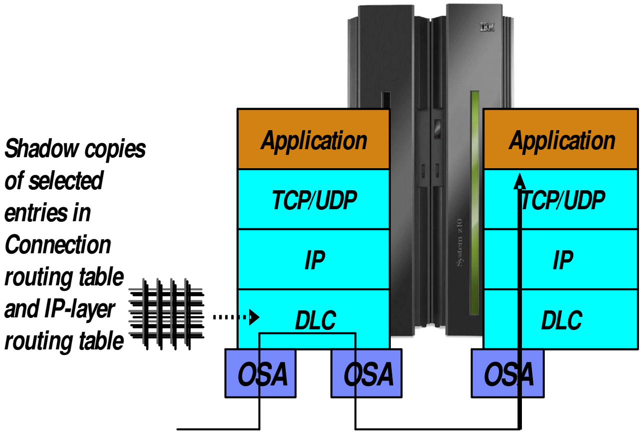
- Traffic going to or coming from applications on this TCP/IP stack only

Background information: QDIO Accelerator

V1R11

- Provides fast path IP forwarding for these DLC combinations
  - Inbound QDIO, outbound QDIO or HiperSockets
  - Inbound HiperSockets, outbound QDIO or HiperSockets
- 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




Shadow copies of selected entries in Connection routing table and IP-layer routing table
Application
TCP/UDP
IP
DLC
OSA
OSA
Application
TCP/UDP
IP
DLC
OSA

QDIO Accelerator coexistence with IP Filtering

- Problem
  - No support for acceleration when IP security enabled
    - Even if stack processing is not needed for forwarded traffic

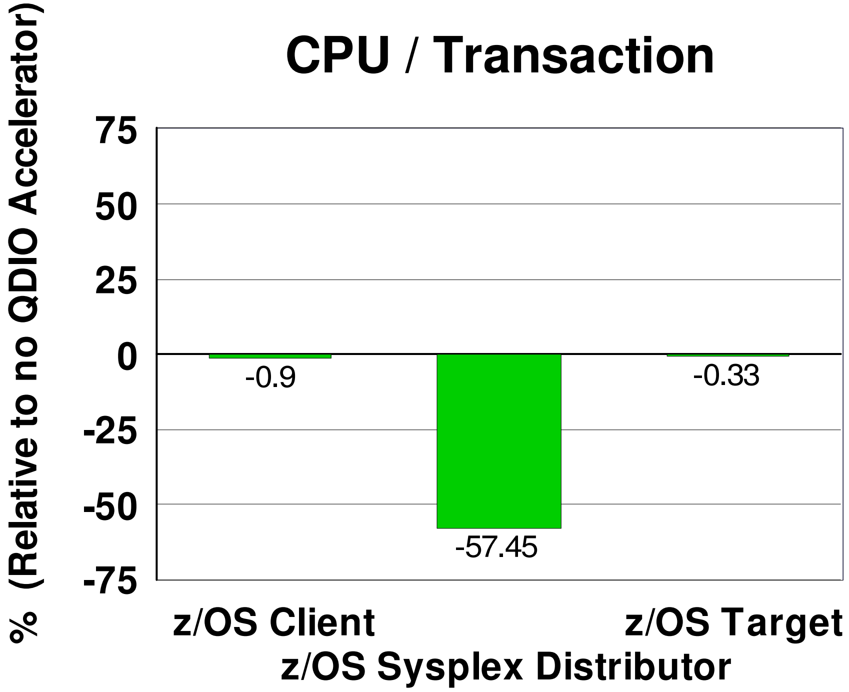
- Solution
  - Allow QDIO accelerator for routed traffic when IPCONFIG IPSECURITY and IPCONFIG QDIOACCELERATOR configured
    - QDIO accelerator for non-Sysplex Distributor traffic requires that the acceleration stack must not filter or log routed traffic
    - Always allow QDIO accelerator for Sysplex Distributor traffic
  - Not supported for HiperSockets accelerator

QDIO Accelerator: IPCONFIG syntax and performance results

…

```
|  _NOQDIOACCELerator________________________          |
|_|__________________________________________|_________|
| |                   _QDIOPriority 1________|         |
| |_QDIOACCELerator__|_______________________|         |
|                    |_QDIOPriority priority_|         |
```

…



Request-Response workload
RR20: 20 sessions, 100 / 800

FTP using zHPF – Improving throughput

- There are many factors that influence the transfer rates for z/OS FTP connections. Some of the more significant ones are (in order of impact):
  - **DASD read/write access**
  - Data transfer type (Binary, ASCII..)
  - Dataset characteristics (e.g., fixed block or variable)

  *Note the network (Hipersockets, OSA, 10Gb, SMC-R) characteristics have very little impact when reading from, and writing to, DASD as you will see in our results section.

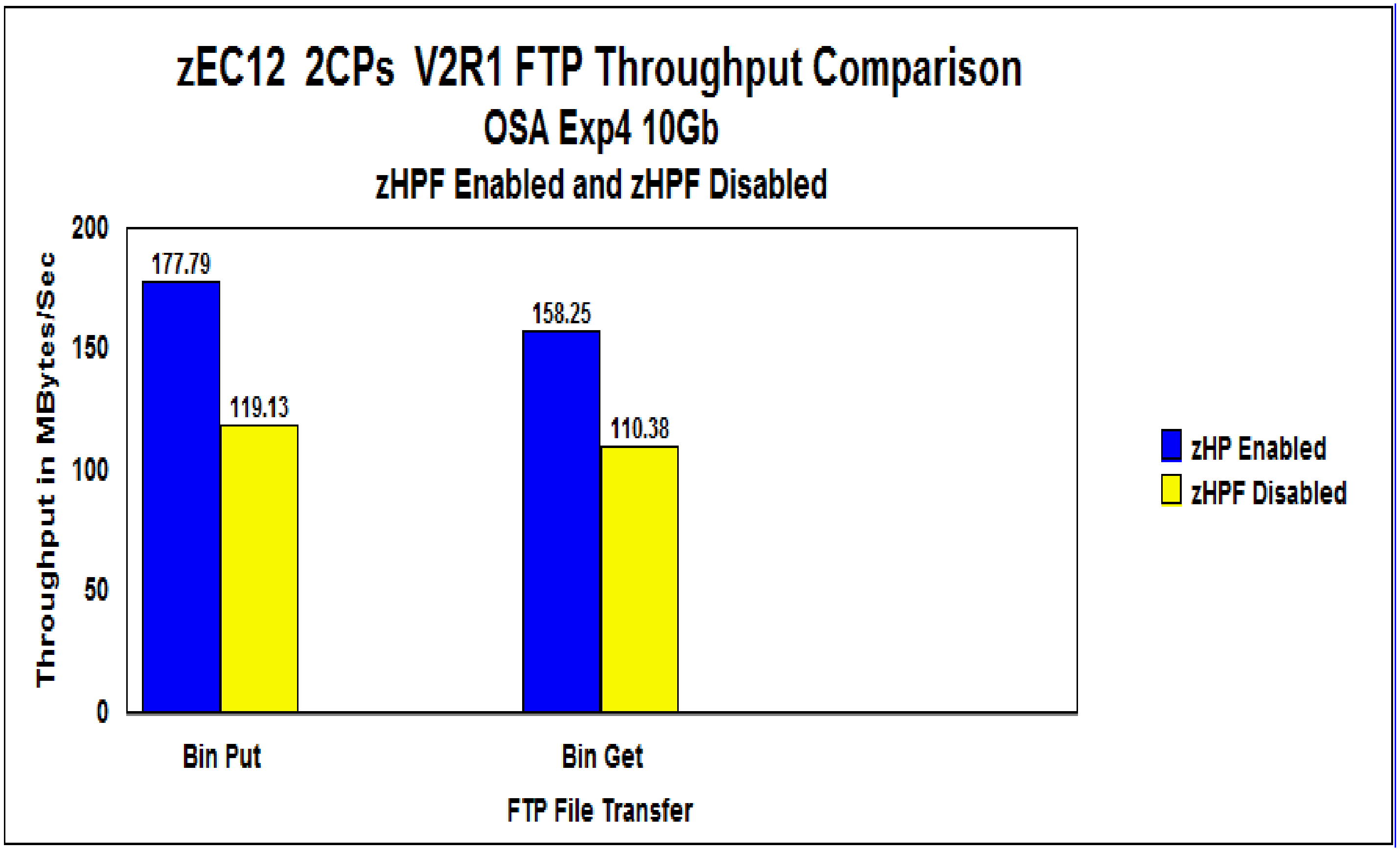
- zHPF FAQ link
  - http://www-03.ibm.com/support/techdocs/atsmastr.nsf/WebIndex/FQ127122
  - Works with DS8000 storage systems

FTP using zHPF – Improving throughput

- FTP Workload
  - z/OS FTP client GET or PUT 1200 MB data set from or to z/OS FTP server
  - DASD to DASD (read from or write to)
  - zHPF enabled/disabled
  - Single file transfer
  - Used Variable block data set for the test
    - Organization .... PS
    - Record Format ...VB
    - Record Length …6140
    - Block size 23424
  - For Hipersocket
    - Configure GLOBALCONFIG IQDMULTIWRITE

FTP using zHPF – Improving throughput

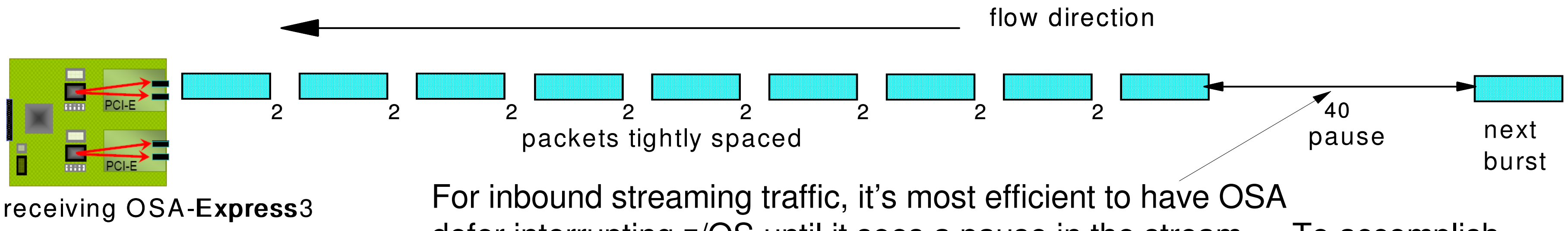
Throughput is improved by 43-49% with Enabling zHPF



Optimizing inbound communications using OSA-Express

Timing Considerations for Various Inbound workloads…

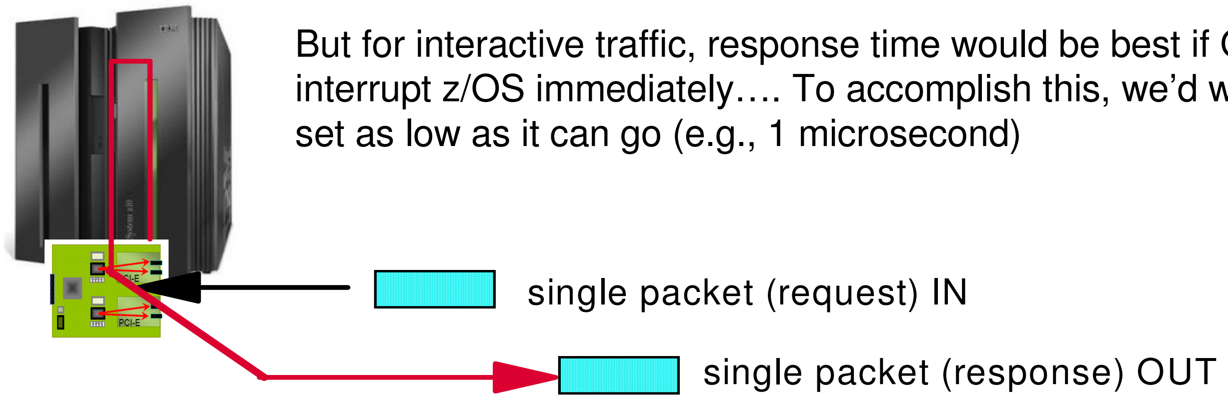
Inbound Streaming Traffic Pattern



For inbound streaming traffic, it's most efficient to have OSA defer interrupting z/OS until it sees a pause in the stream…. To accomplish this, we'd want the OSA **LAN-Idle timer** set fairly high (e.g., don't interrupt unless there's a traffic pause of at least 20 microseconds)

Interactive Traffic Pattern

But for interactive traffic, response time would be best if OSA would interrupt z/OS immediately…. To accomplish this, we'd want the OSA **LAN-Idle timer** set as low as it can go (e.g., 1 microsecond)

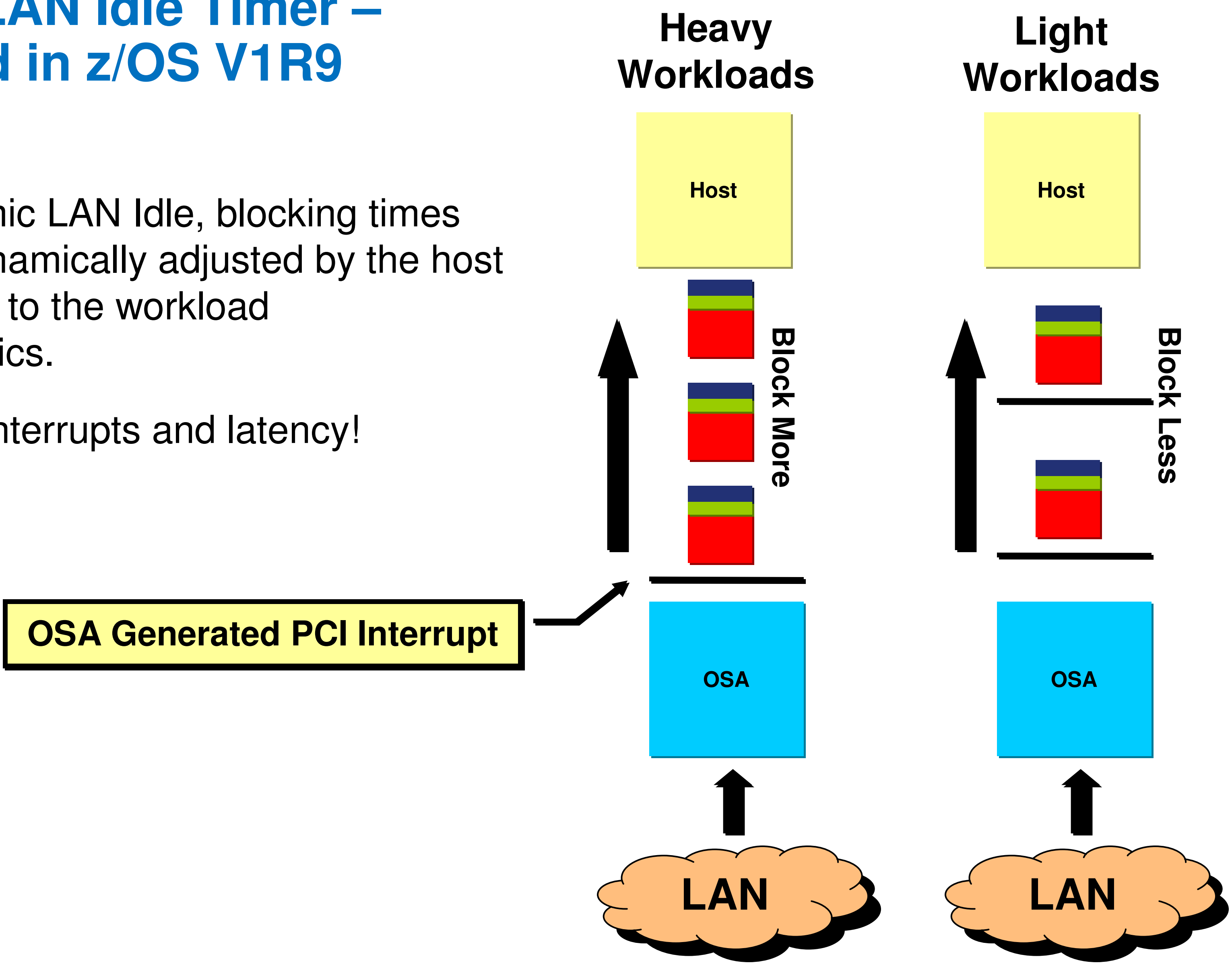


Read-Side interrupt frequency is all about the LAN-Idle timer!

For detailed discussion on inbound interrupt timing, please see Part 1 of "z/OS Communications Server V1R12 Performance Study: OSA-Express3 Inbound Workload Queueing". http://www-01.ibm.com/support/docview.wss?uid=swg27005524

Dynamic LAN Idle Timer – Introduced in z/OS V1R9

- With Dynamic LAN Idle, blocking times are now dynamically adjusted by the host in response to the workload characteristics.
- Optimizes interrupts and latency!



Dynamic LAN Idle Timer: Configuration

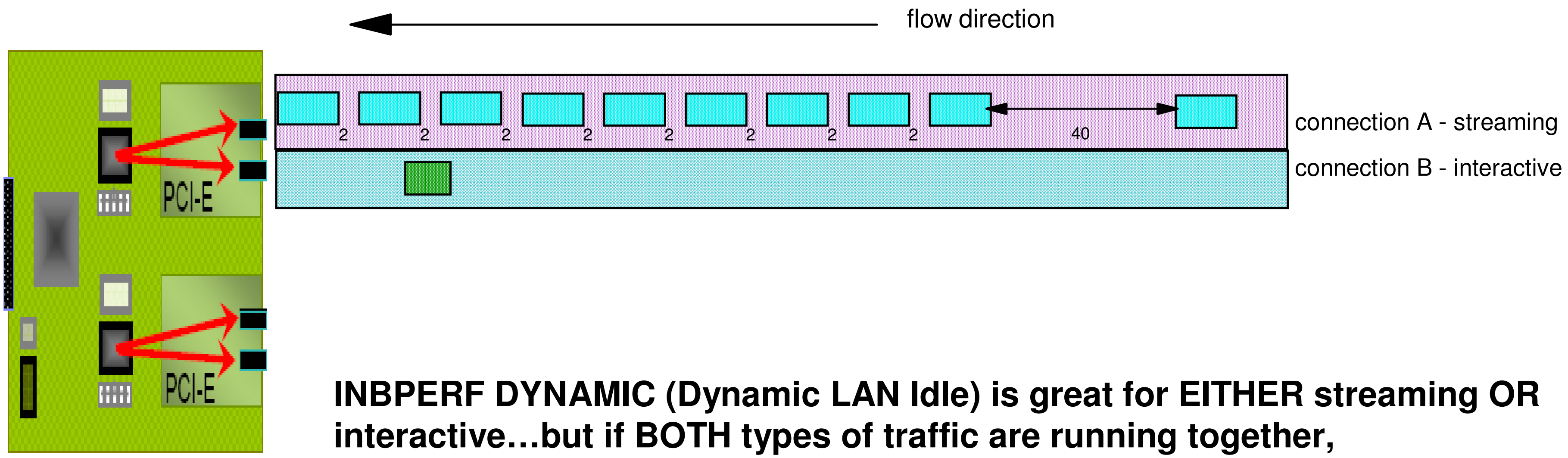
- Configure INBPERF DYNAMIC on the INTERFACE statement

```
>>-INTERFace--intf_name-------------------------------------------->
 .
   .-INBPERF BALANCED--------.
>--+-------------------------+--------->
   '-INBPERF--+-DYNAMIC----+-'
              +-MINCPU-----+
              '-MINLATENCY-'
 .
```

- BALANCED (default) - a static interrupt-timing value, selected to achieve reasonably high throughput and reasonably low CPU
- **DYNAMIC** - a dynamic interrupt-timing value that changes based on current inbound workload conditions ← **Generally Recommended!**
- 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

Note: These values cannot be changed without stopping and restarting the interface

Dynamic LAN Idle Timer: But what about mixed workloads?



receiving OSA-Express3

INBPERF DYNAMIC (Dynamic LAN Idle) is great for EITHER streaming OR interactive…but if BOTH types of traffic are running together, DYNAMIC mode will tend toward CPU conservation (elongating the LAN-Idle timer). So in a mixed (streaming + interactive) workload, the interactive flows will be delayed, waiting for the OSA to detect a pause in the stream…..

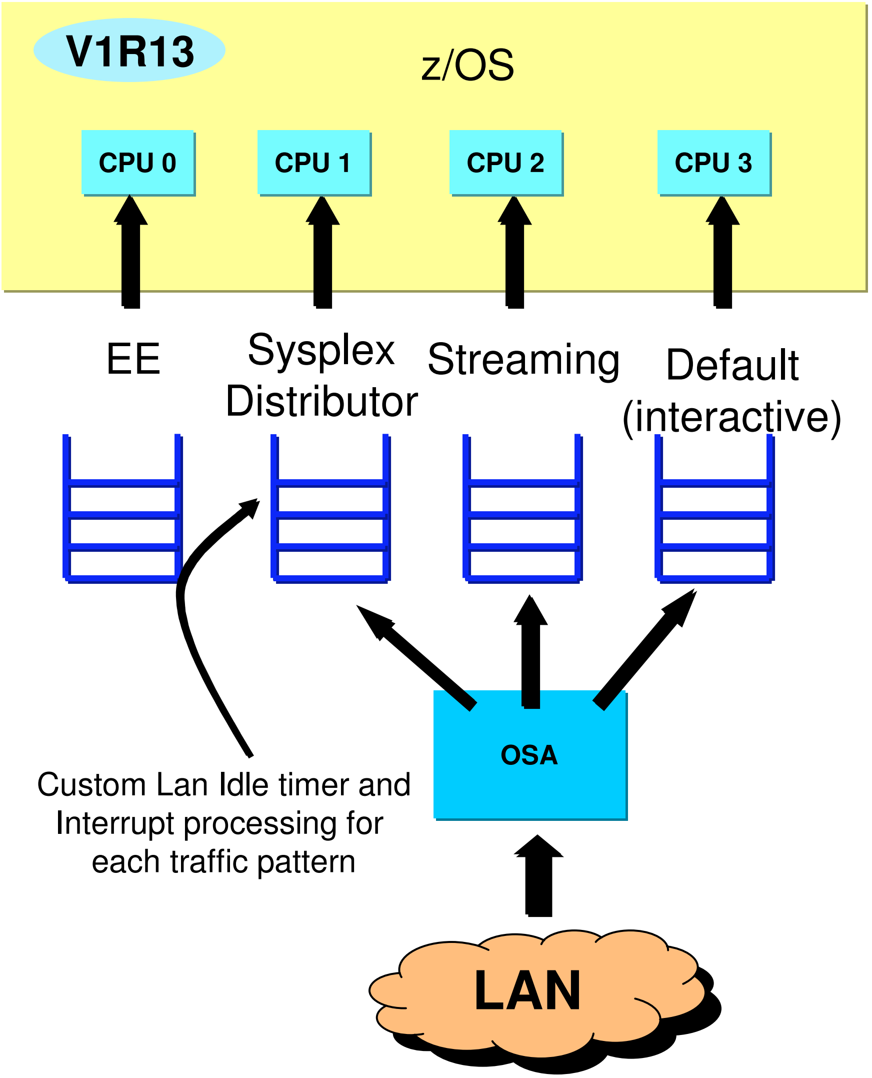
Inbound Workload Queuing

V1R12

Starting with OSA-Express3S IWQ and z/OS V1R12, OSA now directs streaming traffic onto its own input queue – transparently separating the streaming traffic away from the more latency-sensitive interactive flows…

And each input queue has its own LAN-Idle timer, so the Dynamic LAN Idle function can now tune the streaming (bulk) queue to conserve CPU (high LAN-idle timer setting), while generally allowing the primary queue to operate with very low latency (minimizing its LAN-idle timer setting). So interactive traffic (on the primary input queue) may see significantly improved response time.

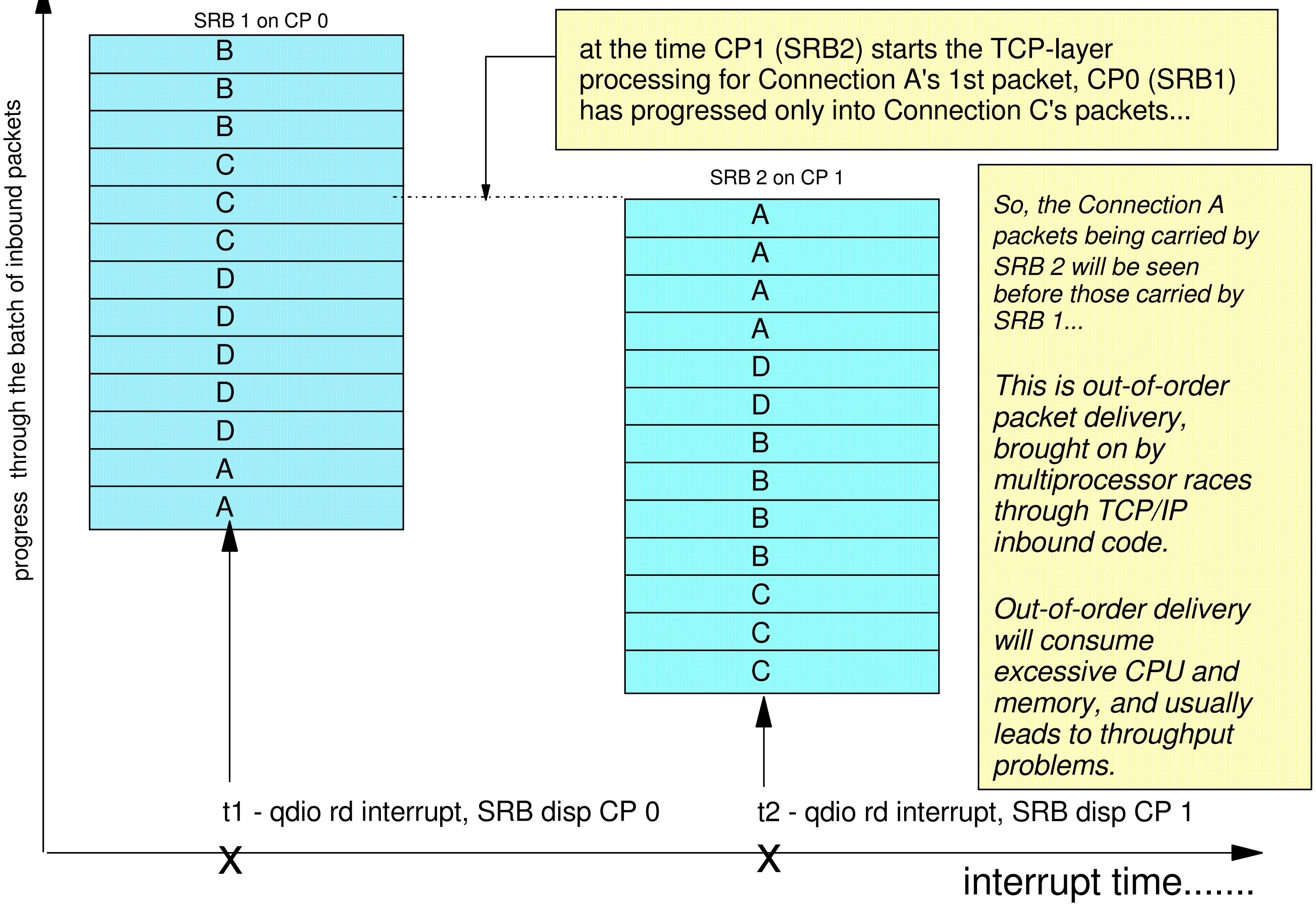
The separation of streaming traffic away from interactive also enables new streaming traffic efficiencies in Communications Server. This results in improved in-order delivery (better throughput and CPU consumption).



Custom Lan Idle timer and Interrupt processing for each traffic pattern

Improved Streaming Traffic Efficiency With IWQ

Before we had IWQ, Multiprocessor races would degrade streaming performance!


progress through the batch of inbound packets
SRB 1 on CP 0
B
B
B
C
C
C
D
D
D
D
D
A
A
SRB 2 on CP 1
A
A
A
A
D
D
B
B
B
B
C
C
C
t1 - qdio rd interrupt, SRB disp CP 0
t2 - qdio rd interrupt, SRB disp CP 1
X
X
interrupt time.......

at the time CP1 (SRB2) starts the TCP-layer processing for Connection A's 1st packet, CP0 (SRB1) has progressed only into Connection C's packets...

So, the Connection A packets being carried by SRB 2 will be seen before those carried by SRB 1...

This is out-of-order packet delivery, brought on by multiprocessor races through TCP/IP inbound code.

Out-of-order delivery will consume excessive CPU and memory, and usually leads to throughput problems.

IWQ does away with MP-race-induced ordering problems!

With streaming traffic sorted onto its own queue, it is now convenient to service streaming traffic from a single CP (i.e., using a single SRB).

So with IWQ, we no longer have inbound SRB races for streaming data.

QDIO Inbound Workload Queuing – Configuration

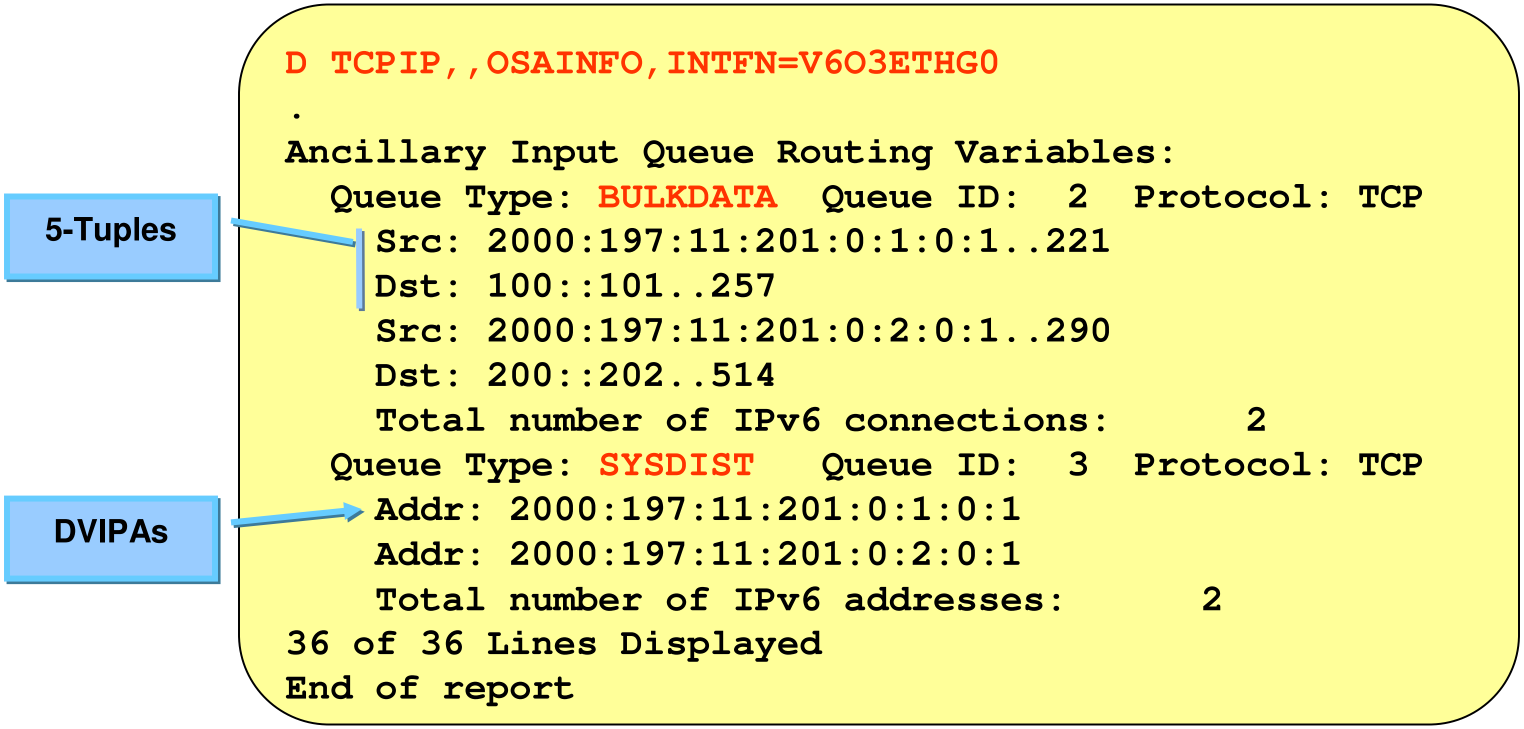
- INBPERF DYNAMIC WORKLOADQ enables QDIO Inbound Workload Queuing (IWQ)

```
>>-INTERFace--intf_name------------------------------------------>
.
.-INBPERF BALANCED--------------------.
 >--+--------------------------------------+-->
    |                   .-NOWORKLOADQ-.    |
    `-INBPERF-+-DYNAMIC-+-------------+--+-'
              |         `-WORKLOADQ---'  |
              +-MINCPU-------------------+
              `-MINLATENCY---------------'
```

  - INTERFACE statements only - no support for DEVICE/LINK definitions
  - QDIO Inbound Workload Queuing requires VMAC

QDIO Inbound Workload Queuing

- Display OSAINFO command (V1R12) shows you what's registered in OSA



```
D TCPIP,,OSAINFO,INTFN=V6O3ETHG0
.
Ancillary Input Queue Routing Variables:
  Queue Type: BULKDATA  Queue ID:  2  Protocol: TCP
    Src: 2000:197:11:201:0:1:0:1..221
    Dst: 100::101..257
    Src: 2000:197:11:201:0:2:0:1..290
    Dst: 200::202..514
    Total number of IPv6 connections:      2
  Queue Type: SYSDIST   Queue ID:  3  Protocol: TCP
    Addr: 2000:197:11:201:0:1:0:1
    Addr: 2000:197:11:201:0:2:0:1
    Total number of IPv6 addresses:      2
36 of 36 Lines Displayed
End of report
```

- BULKDATA queue registers 5-tuples with OSA (streaming connections)
- SYSDIST queue registers Distributable DVIPAs with OSA

QDIO Inbound Workload Queuing: Netstat DEvlinks/-d

- Display TCPIP,,Netstat,DEvlinks to see whether QDIO inbound workload queueing is enabled for a QDIO interface

```
D TCPIP,,NETSTAT,DEVLINKS,INTFNAME=QDIO4101L
EZD0101I NETSTAT CS V1R12 TCPCS1
INTFNAME: QDIO4101L         INTFTYPE: IPAQENET   INTFSTATUS: READY
    PORTNAME: QDIO4101  DATAPATH: 0E2A     DATAPATHSTATUS: READY
    CHPIDTYPE: OSD
    SPEED: 0000001000
...
    READSTORAGE: GLOBAL (4096K)
    INBPERF: DYNAMIC
      WORKLOADQUEUEING: YES
    CHECKSUMOFFLOAD: YES
    SECCLASS: 255                    MONSYSPLEX: NO
    ISOLATE: NO                      OPTLATENCYMODE: NO
...
1 OF 1 RECORDS DISPLAYED
END OF THE REPORT
```

QDIO Inbound Workload Queuing: Display TRLE

- Display NET,TRL,TRLE=trlename to see whether QDIO inbound workload queueing is in use for a QDIO interface

```
D NET,TRL,TRLE=QDIO101
IST097I DISPLAY ACCEPTED
...
IST2263I PORTNAME = QDIO4101   PORTNUM =   0   OSA CODE LEVEL = ABCD
...
IST1221I DATA  DEV = 0E2A STATUS = ACTIVE     STATE = N/A
IST1724I I/O TRACE = OFF  TRACE LENGTH = *NA*
IST1717I ULPID = TCPCS1
IST2310I ACCELERATED ROUTING DISABLED
IST2331I QUEUE   QUEUE     READ
IST2332I ID      TYPE      STORAGE
IST2205I ------  --------  ---------------
IST2333I RD/1    PRIMARY   4.0M(64 SBALS)
IST2333I RD/2    BULKDATA  4.0M(64 SBALS)
IST2333I RD/3    SYSDIST   4.0M(64 SBALS)
...
IST924I -------------------------------------------------------------
IST314I END
```

QDIO Inbound Workload Queuing: Netstat ALL/-A

- Display TCPIP,,Netstat,ALL to see whether QDIO inbound workload BULKDATA queueing is in use for a given connection

```
D TCPIP,,NETSTAT,ALL,CLIENT=USER1
EZD0101I NETSTAT CS V1R12 TCPCS1
CLIENT NAME: USER1                    CLIENT ID: 00000046
  LOCAL SOCKET: ::FFFF:172.16.1.1..20
  FOREIGN SOCKET: ::FFFF:172.16.1.5..1030
    BYTESIN:            00000000000023316386
    BYTESOUT:           00000000000000000000
    SEGMENTSIN:         00000000000000016246
    SEGMENTSOUT:        00000000000000000922
    LAST TOUCHED:       21:38:53         STATE:              ESTABLSH
...
Ancillary Input Queue: Yes
    BulkDataIntfName: QDIO4101L
...
    APPLICATION DATA:   EZAFTP0S D USER1     C      PSSS
----
1 OF 1 RECORDS DISPLAYED
END OF THE REPORT
```

QDIO Inbound Workload Queuing: Netstat STATS/-S

- Display TCPIP,,Netstat,STATS to see the total number of TCP segments received on BULKDATA queues

```
D TCPIP,,NETSTAT,STATS,PROTOCOL=TCP
EZD0101I NETSTAT CS V1R12 TCPCS1
TCP STATISTICS
  CURRENT ESTABLISHED CONNECTIONS      = 6
  ACTIVE CONNECTIONS OPENED            = 1
  PASSIVE CONNECTIONS OPENED           = 5
  CONNECTIONS CLOSED                   = 5
  ESTABLISHED CONNECTIONS DROPPED      = 0
  CONNECTION ATTEMPTS DROPPED          = 0
  CONNECTION ATTEMPTS DISCARDED        = 0
  TIMEWAIT CONNECTIONS REUSED          = 0
  SEGMENTS RECEIVED                    = 38611
...
  SEGMENTS RECEIVED ON OSA BULK QUEUES= 2169
  SEGMENTS SENT                        = 2254
...
END OF THE REPORT
```

Quick INBPERF Review Before We Push On….

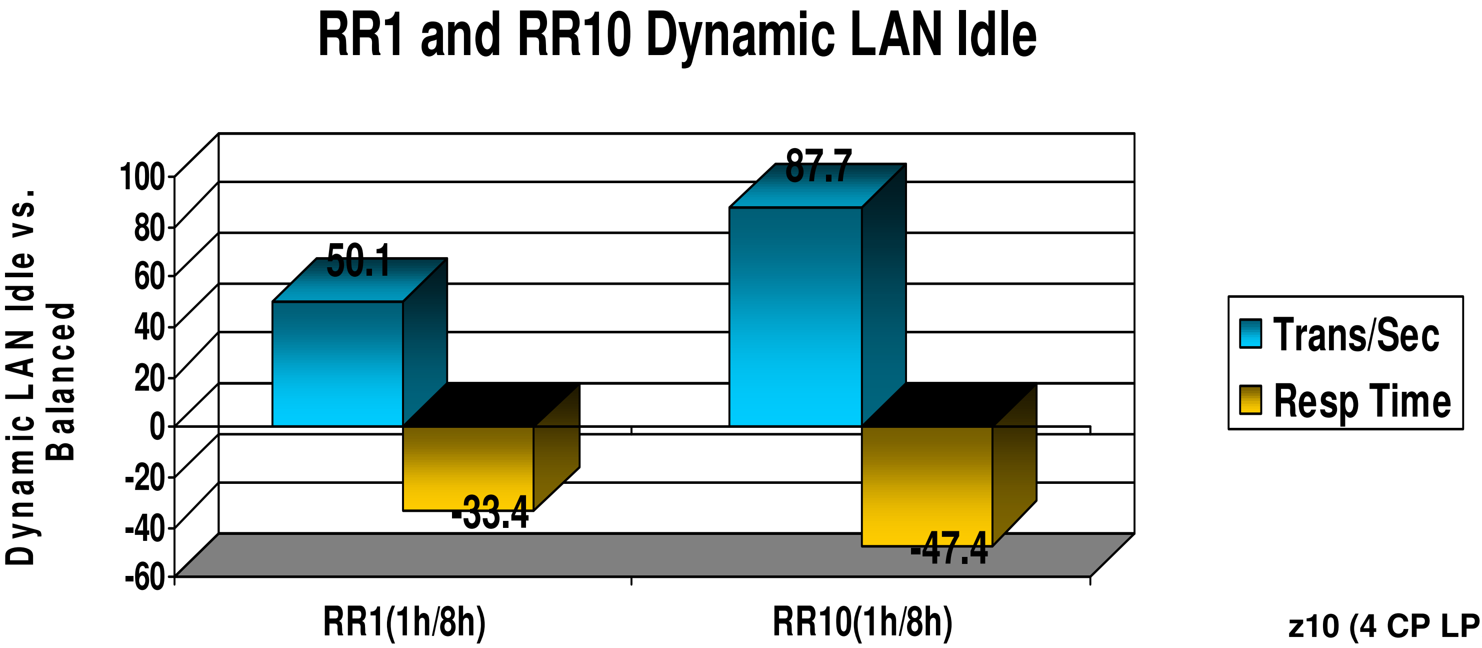
- The original static INBPERF settings (MINCPU, MINLATENCY, BALANCED) provide sub-optimal performance for workloads that tend to shift between request/response and streaming modes.

- We therefore **recommend customers specify INBPERF DYNAMIC**, since it self-tunes, to provide excellent performance even when inbound traffic patterns shift.

- Inbound Workload Queueing (IWQ) mode is an extension to the Dynamic LAN Idle function. IWQ improves upon the DYNAMIC setting, in part because it provides finer interrupt-timing control for mixed (interactive + streaming) workloads.

Dynamic LAN Idle Timer: Performance Data

Dynamic LAN Idle improved RR1 TPS 50% and RR10 TPS by 33%. Response Time for these workloads is improved 33% and 47%, respectively.

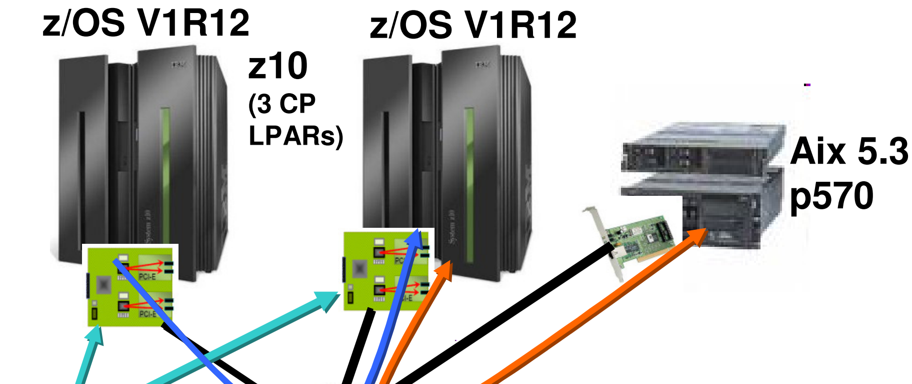


z10 (4 CP LPARs), z/OS V1R13, OSA-E3 1Gbe

1h/8h indicates 100 bytes in and 800 bytes out

Note: The performance measurements discussed in this presentation are z/OS V1R13 Communications Server numbers and were collected using a dedicated system environment. The results obtained in other configurations or operating system environments may vary.

Inbound Workload Queuing: Performance Data



z/OS V1R12

z10

(3 CP LPARs)

z/OS V1R12

Aix 5.3

p570

OSA-Express3's
In Dynamic
or IWQ mode

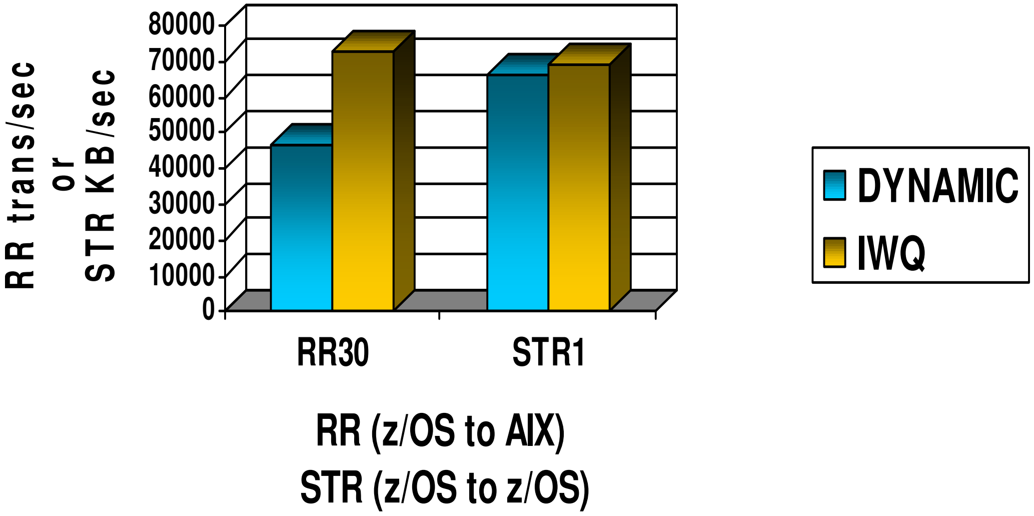
1GBe
or 10GBe
network

IWQ: Mixed Workload Results vs DYNAMIC:

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

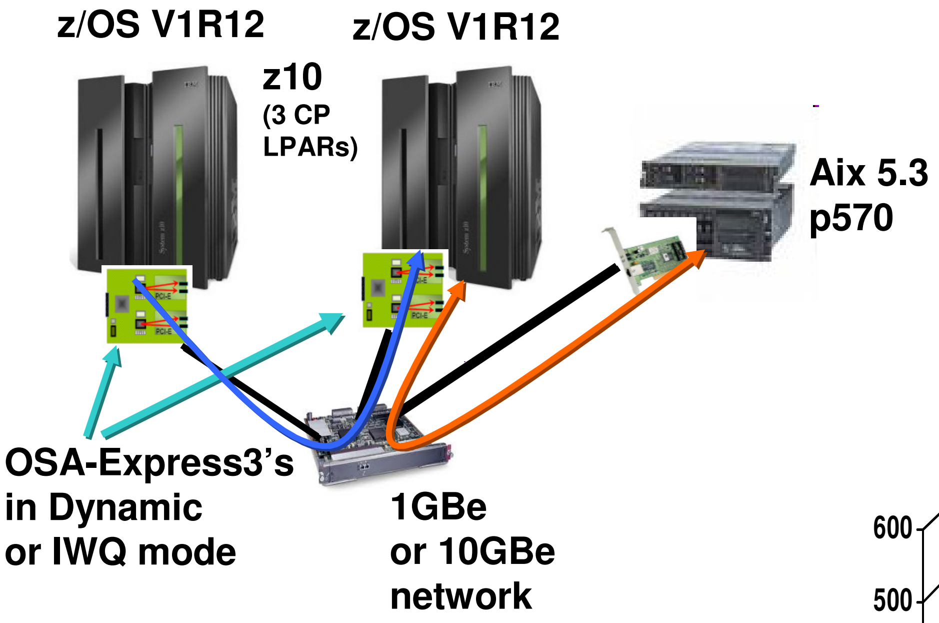
–Streaming Throughput also improved in this test: +5%

Mixed Workload (IWQ vs Dynamic)



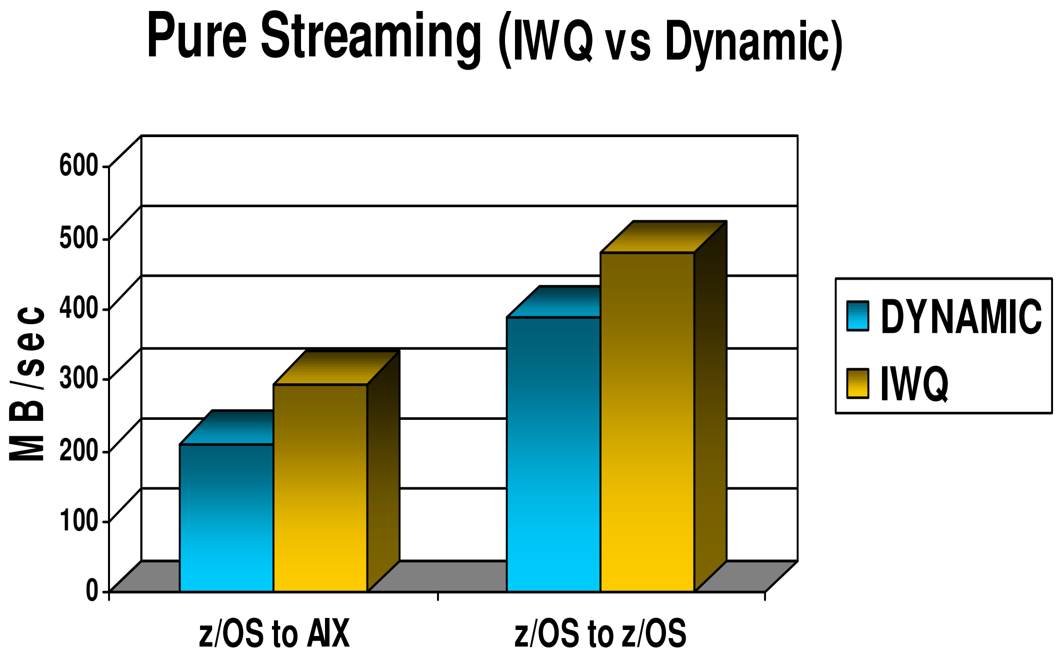
For z/OS outbound streaming to another platform, the degree of performance boost (due to IWQ) is relative to receiving platform's sensitivity to out-of-order packet delivery. For streaming INTO z/OS, IWQ will be especially beneficial for multi-CP configurations.

Inbound Workload Queuing: Performance Data



IWQ: Pure Streaming Results vs DYNAMIC:

- –z/OS<->AIX Streaming Throughput improved 40%
- –z/OS<->z/OS Streaming Throughput improved 24%



For z/OS outbound streaming to another platform, the 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 for multi-CP configurations.

IWQ Usage Considerations:

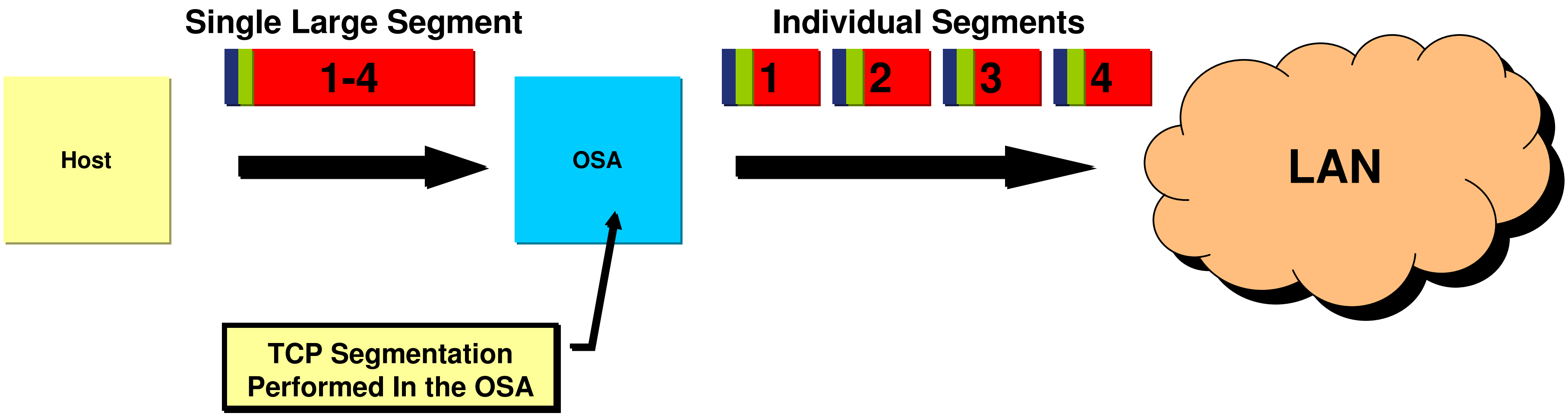
- Minor ECSA Usage increase: IWQ will grow ECSA usage by 72KBytes (per OSA interface) if Sysplex Distributor (SD) or EE is in use; 36KBytes if SD and EE are not in use
- IWQ requires OSA-Express3/OSA-Express4/OSA-Express5 in QDIO mode running on zEnterprise 196/ zEC12(for OSAE5).
- IWQ must be configured 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)

Optimizing outbound communications using OSA-Express

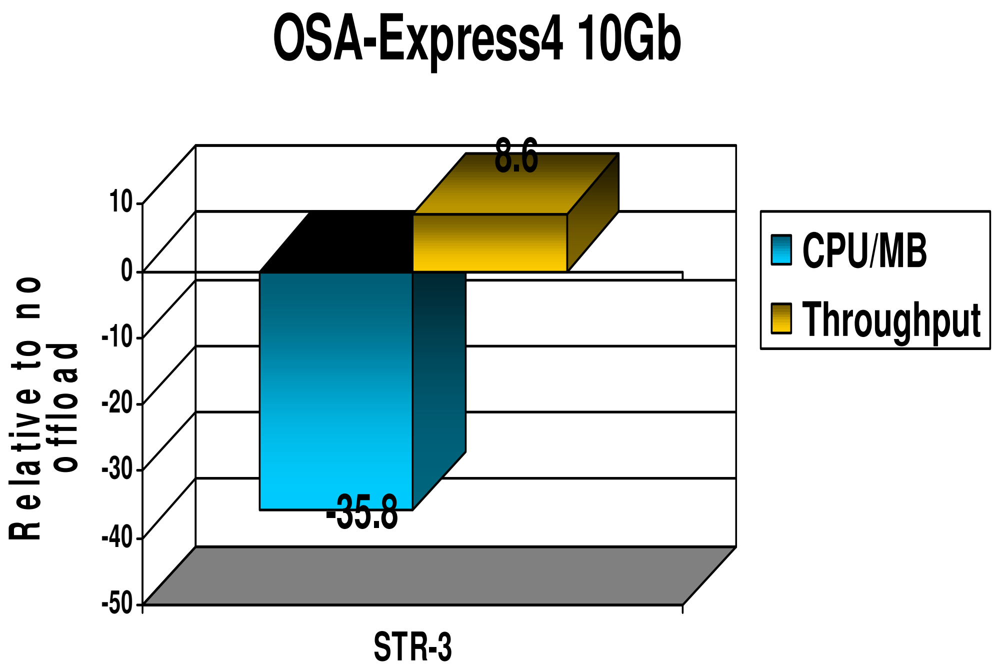
TCP Segmentation Offload

V1R13

- Segmentation consumes (high cost) host CPU cycles in the TCP stack
- Segmentation Offload (also referred to as “Large Send”)
  - Offload most IPv4 and/or IPv6 TCP segmentation processing to OSA
  - Decrease host CPU utilization
  - Increase data transfer efficiency
  - Checksum offload also added for IPv6



z/OS Segmentation Offload performance measurements



Send buffer size: 180K for streaming workloads

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

Note: The performance measurements discussed in this presentation are z/OS V1R13 Communications Server numbers and were collected using a dedicated system environment. The results obtained in other configurations or operating system environments may vary.

TCP Segmentation Offload: Configuration

V1R13

- Enabled with IPCONFIG/IPCONFIG6 SEGMENTATIONOFFLOAD

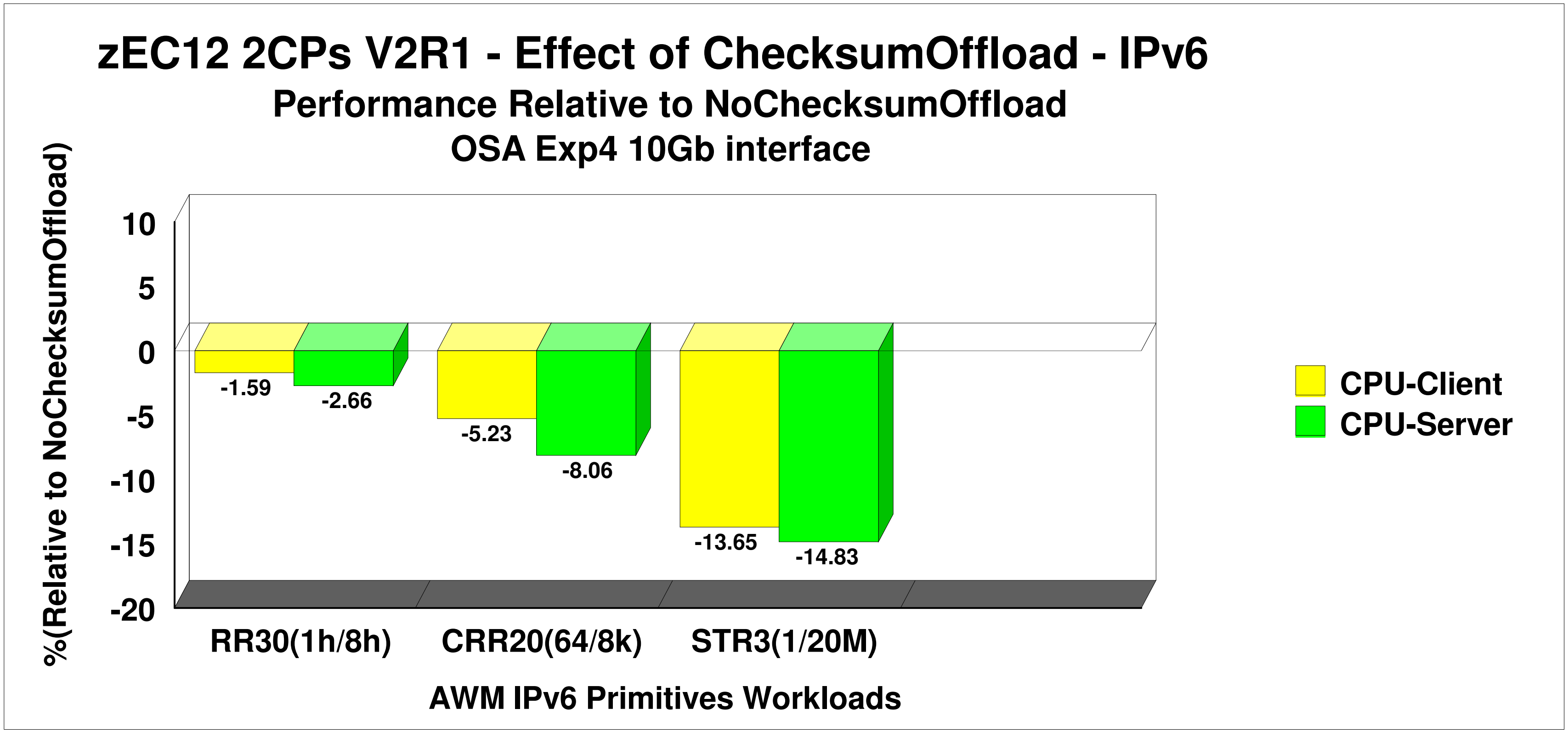
```
>>-IPCONFIG-------------------------------------------------->
.
.
>----+-----------------------------------------------+-+-------><
     | .-NOSEGMENTATIONOFFLoad-.                     |
     +-+-----------------------+---------------------+
     | '-SEGMENTATIONOFFLoad---'                     |
```

- Disabled by default
- Previously enabled via GLOBALCONFIG
- Segmentation cannot be offloaded for
  - Packets to another stack sharing OSA port
  - IPSec encapsulated packets
  - When multipath is in effect (unless all interfaces in the multipath group support segmentation offload)

Reminder! Checksum Offload enabled by default

z/OS Checksum Offload performance measurements

V1R13

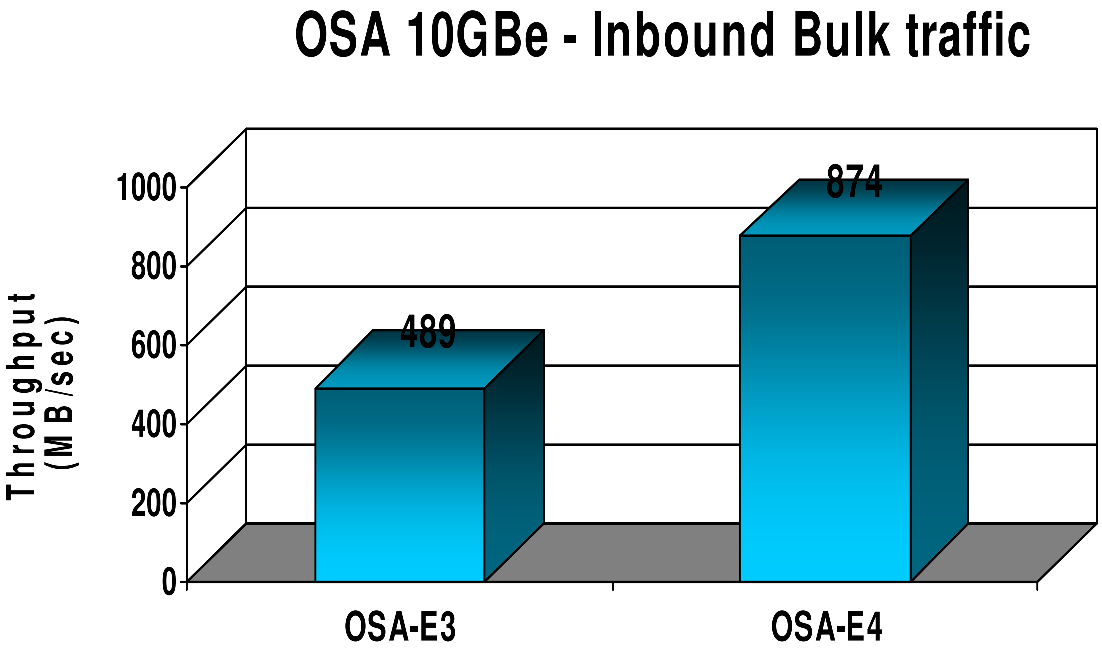


Note: The performance measurements discussed in this presentation are z/OS V2R1 Communications Server numbers and were collected using a dedicated system environment. The results obtained in other configurations or operating system environments may vary.

OSA-Express4/5

OSA-Express4/5 Enhancements – 10GB improvements

- Improved on-card processor speed and memory bus provides better utilization of 10GB network

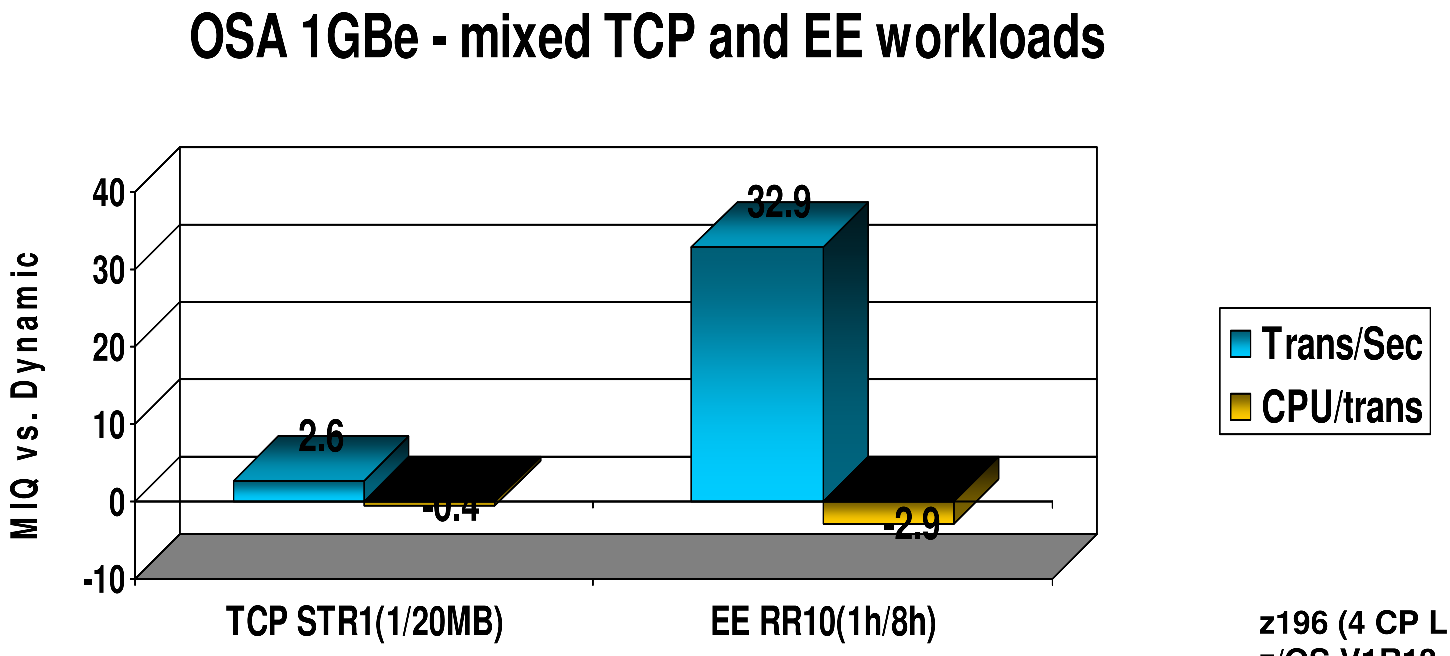


z196 (4 CP LPARs), z/OS V1R13, OSA-E3/OSA-E4 10Gbe

Note: The performance measurements discussed in this presentation are z/OS V1R13 Communications Server numbers and were collected using a dedicated system environment. The results obtained in other configurations or operating system environments may vary.

OSA-Express4 Enhancements – EE Inbound Queue

- Enterprise Extender queue provides internal optimizations
  - EE traffic processed quicker
  - Avoids memory copy of data



z196 (4 CP LPARs), z/OS V1R13, OSA-E3/OSA-E4 1Gbe

Note: The performance measurements discussed in this presentation are z/OS V1R13 Communications Server numbers and were collected using a dedicated system environment. The results obtained in other configurations or operating system environments may vary.

OSA-Express4 Enhancements – Other improvements

- Checksum Offload support for IPv6 traffic
- Segmentation Offload support for IPv6 traffic

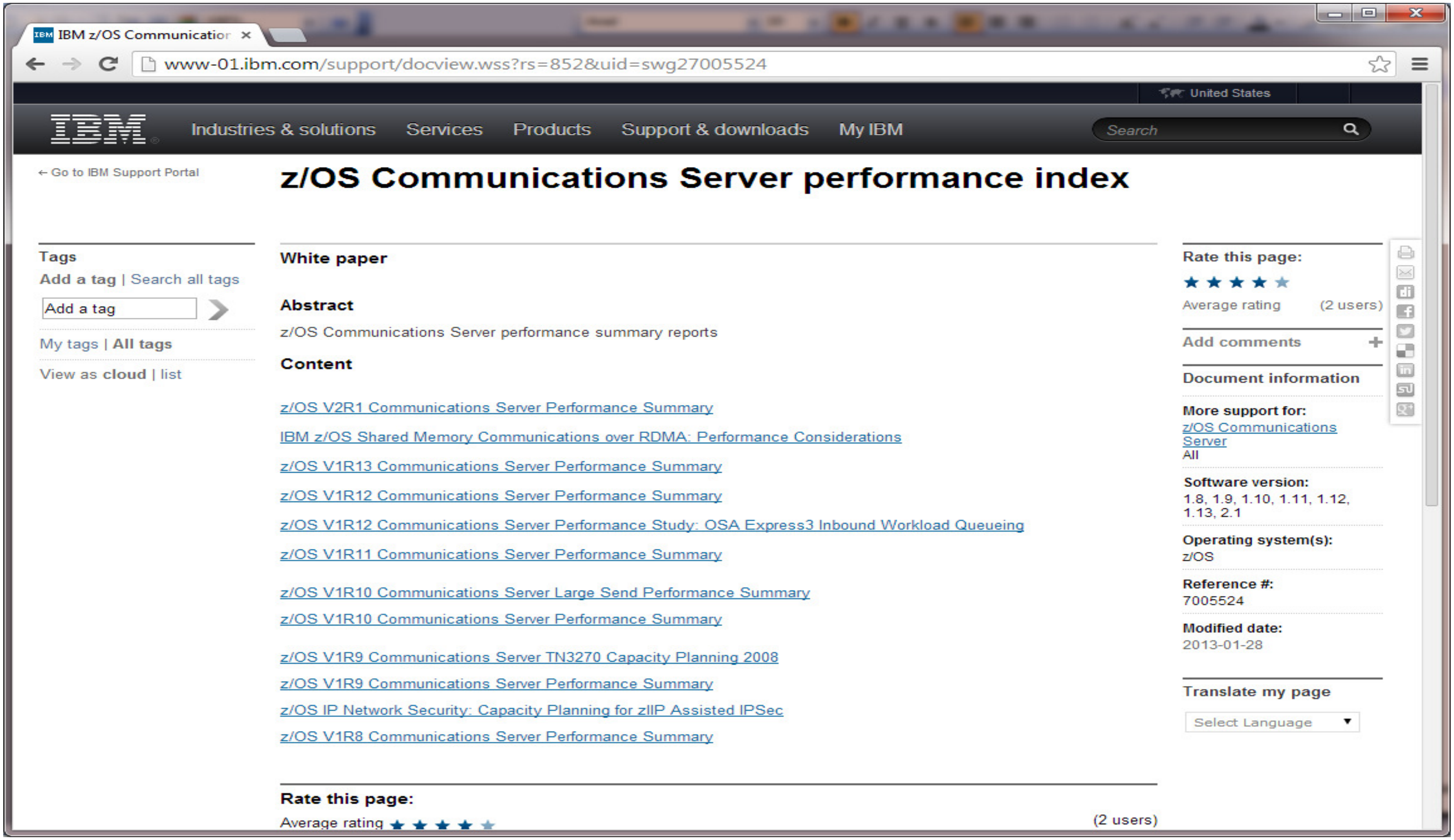
z/OS Communications Server Performance Summaries

z/OS Communications Server Performance Summaries

- Performance of each z/OS Communications Server release is studied by an internal performance team
- Summaries are created and published online
  - http://www-01.ibm.com/support/docview.wss?rs=852&uid=swg27005524
- Recently added:
  - The z/OS VR1 Communications Server Performance Summary
    - Release to release comparisons
    - Capacity planning information
  - IBM z/OS Shared Memory Communications over RDMA: Performance Considerations - Whitepaper

z/OS Communications Server Performance Website

www-01.ibm.com/support/docview.wss?uid=swg27005524



← Go to IBM Support Portal

z/OS Communications Server performance index

Tags

Add a tag | Search all tags

Add a tag

My tags | All tags

View as cloud | list

White paper

Abstract

z/OS Communications Server performance summary reports

Content

- z/OS V2R1 Communications Server Performance Summary
- IBM z/OS Shared Memory Communications over RDMA: Performance Considerations
- z/OS V1R13 Communications Server Performance Summary
- z/OS V1R12 Communications Server Performance Summary
- z/OS V1R12 Communications Server Performance Study: OSA Express3 Inbound Workload Queueing
- z/OS V1R11 Communications Server Performance Summary
- z/OS V1R10 Communications Server Large Send Performance Summary
- z/OS V1R10 Communications Server Performance Summary
- z/OS V1R9 Communications Server TN3270 Capacity Planning 2008
- z/OS V1R9 Communications Server Performance Summary
- z/OS IP Network Security: Capacity Planning for zIIP Assisted IPSec
- z/OS V1R8 Communications Server Performance Summary

Rate this page:

Average rating ★ ★ ★ ★ ★ (2 users)

Rate this page:

★ ★ ★ ★ ★

Average rating (2 users)

Add comments +

Document information

More support for:
z/OS Communications Server
All

Software version:
1.8, 1.9, 1.10, 1.11, 1.12, 1.13, 2.1

Operating system(s):
z/OS

Reference #:
7005524

Modified date:
2013-01-28

Translate my page

Select Language ▾

Please fill out your session evaluation

- z/OS CS Performance Improvements
- QR Code:

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

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

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

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

Detailed Usage Considerations for IWQ and OLM

IWQ Usage Considerations:

- Minor ECSA Usage increase: IWQ will grow ECSA usage by 72KBytes (per OSA interface) if Sysplex Distributor (SD) is in use; 36KBytes if SD is not in use

- IWQ requires OSA-Express3 in QDIO mode running on IBM System z10 or OSA-Express3/OSA-Express4 in QDIO mode running on zEnterprise 196.
  - For z10: the minimum field level recommended for OSA-Express3 is microcode level- Driver 79, EC N24398, MCL006
  - For z196 GA1: the minimum field level recommended for OSA-Express3 is microcode level- Driver 86, EC N28792, MCL009
  - For z196 GA2: the minimum field level recommended for OSA-Express3 is microcode level- Driver 93, EC N48158, MCL009
  - For z196 GA2: the minimum field level recommended for OSA-Express4 is microcode level- Driver 93, EC N48121, MCL010

- 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)
- Make sure to apply z/OS V1R12 PTF UK61028 (APAR PM20056) for added streaming throughput boost with IWQ

OLM Usage Considerations(1): OSA Sharing

- Concurrent interfaces to an OSA-Express port using OLM is limited.
  - If one or more interfaces operate OLM on a given port,
    - Only four total interfaces allowed to that single port
    - Only eight total interfaces allowed to that CHPID
  - All four interfaces can operate in OLM
  - An interface can be:
    - Another interface (e.g. IPv6) defined for this OSA-Express port
    - Another stack on the same LPAR using the OSA-Express port
    - Another LPAR using the OSA-Express port
    - Another VLAN defined for this OSA-Express port
    - Any stack activating the OSA-Express Network Traffic Analyzer (OSAENTA)

OLM Usage Considerations (2):

- QDIO Accelerator or HiperSockets Accelerator will not accelerate traffic to or from an OSA-Express operating in OLM

- OLM usage may increase z/OS CPU consumption (due to “early interrupt”)
  - Usage of OLM is therefore not recommended on z/OS images expected to normally be running at extremely high utilization levels
  - OLM does not apply to the bulk-data input queue of an IWQ-mode OSA. From a CPU-consumption perspective, OLM is therefore a more attractive option when combined with IWQ than without IWQ

- Only supported on OSA-Express3 and above with the INTERFACE statement

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