z/OS Communications Server Performance Improvements

Gus Kassimis – kassimis@us.ibm.com
IBM Raleigh, NC, USA

Session: 8318
Friday, March 4, 2011: 8:00 AM-09:00 AM

Trademarks, notices, and disclaimers

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

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

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

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

- Adobe, the Adobe logo, PostScript, and the PostScript logo are either registered trademarks or trademarks of Adobe Systems Incorporated in the United States, and/or other countries.
- Cell Broadband Engine is a trademark of Sony Computer Entertainment, Inc. in the United States, other countries, or both and is used under license there from.
- Java and all Java-based trademarks are trademarks of Sun Microsystems, Inc. in the United States, other countries, or both.
- Microsoft, Windows, Windows NT, and the Windows logo are trademarks of Microsoft Corporation in the United States, other countries, or both.
- InfiniBand is a trademark and service mark of the InfiniBand Trade Association.
- Intel, Intel logo, Intel Inside, Intel Inside logo, Intel Centrino, Intel Centrino logo, Celeron, Intel Xeon, Intel SpeedStep, Itanium, and Pentium are trademarks or registered trademarks of Intel Corporation or its subsidiaries in the United States and other countries.
- UNIX is a registered trademark of The Open Group in the United States and other countries.
- Linux is a registered trademark of Linus Torvalds in the United States, other countries, or both.
- ITIL is a registered trademark, and a registered community trademark of the Office of Government Commerce, and is registered in the U.S. Patent and Trademark Office.
- IT Infrastructure Library is a registered trademark of the Central Computer and Telecommunications Agency, which is now part of the Office of Government Commerce.

Notes:

- Performance is in Internal Throughput Rate (ITR) ratio based on measurements and projections using standard IBM benchmarks in a controlled environment. The actual throughput that any user will experience will vary depending upon considerations such as the amount of multiprogramming in the user's job stream, the I/O configuration, the storage configuration, and the workload processed. Therefore, no assurance can be given that an individual user will achieve throughput improvements equivalent to the performance ratios stated here.
- IBM hardware products are manufactured from new parts, or new and serviceable used parts. Regardless, our warranty terms apply.
- All customer examples cited or described in this presentation are presented as illustrations of the manner in which some customers have used IBM products and the results they may have achieved. Actual environmental costs and performance characteristics will vary depending on individual customer configurations and conditions.
- This publication was produced in the United States. IBM may not offer the products, services or features discussed in this document in other countries, and the information may be subject to change without notice. Consult your local IBM business contact for information on the product or services available in your area.
- All statements regarding IBM's future direction and intent are subject to change or withdrawal without notice, and represent goals and objectives only.
- Information about non-IBM products is obtained from the manufacturers of those products or their published announcements. IBM has not tested those products and cannot confirm the performance, compatibility, or any other claims related to non-IBM products. Questions on the capabilities of non-IBM products should be addressed to the suppliers of those products.
- Prices subject to change without notice. Contact your IBM representative or Business Partner for the most current pricing in your geography.

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

Agenda

- What is one of the most important factors in determining TCP/IP performance over OSA-Express?
  - Why is inbound packet processing key to TCP/IP performance
- Optimizing the inbound path
  - Evolution of optimizations
- The latest optimizations
  - Optimized Latency Mode
  - Inbound Workload Queuing
- How about outbound packet processing?
  - Segmentation offload
  - WLM priority queuing

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.

© 2011 SHARE and IBM Corporation

Optimizing inbound communications using OSA-Express

Special thanks to Tom Moore, Senior Performance Analyst for the z/OS Communications Server, for contributing much of the content of this presentation!

© 2011 SHARE and IBM Corporation

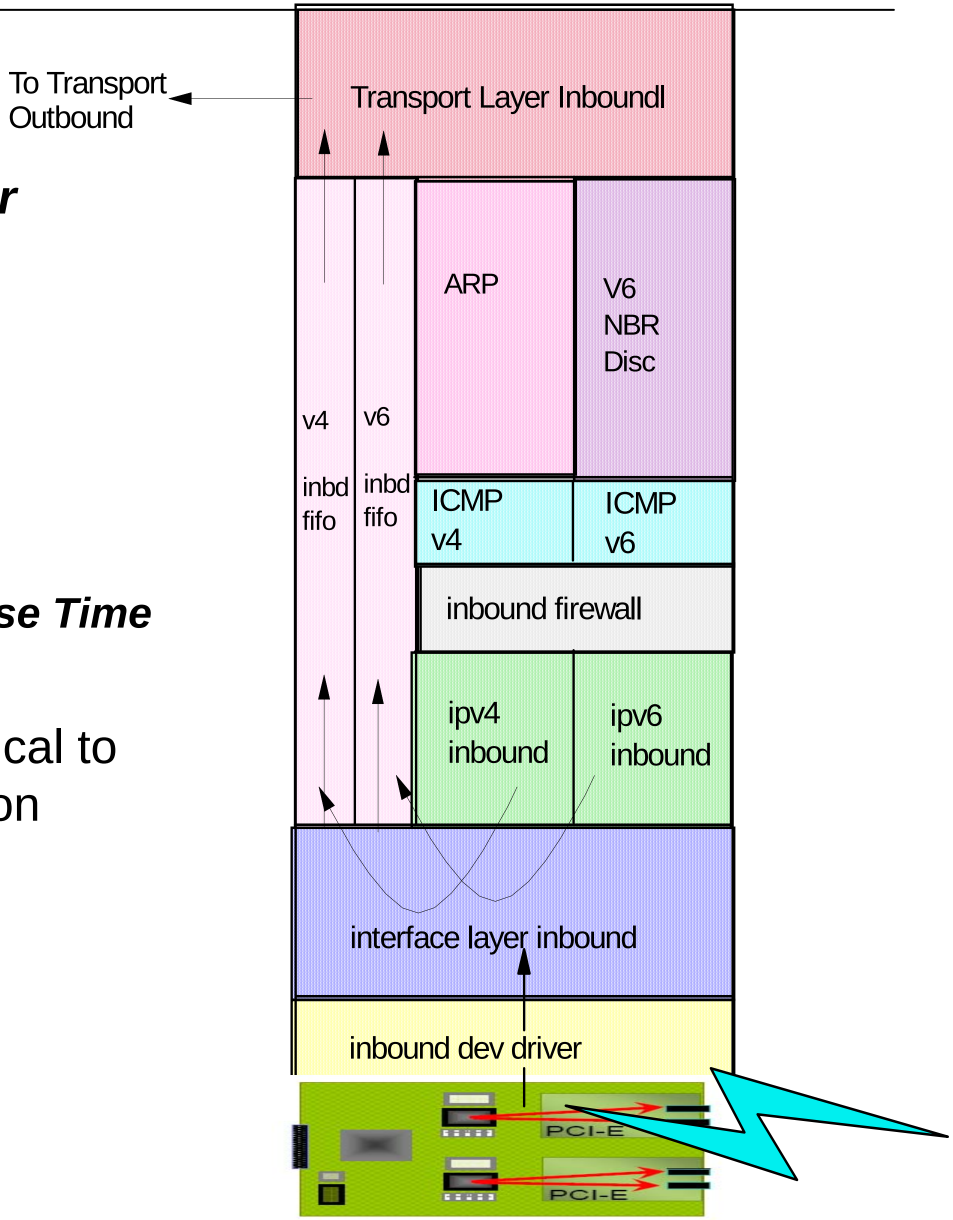
Introduction

- More than any other factor, the ***behavior of the inbound (receiving) communications adapter*** influences overall performance* of z/OS Communications Server.

  Key performance characteristics: CPU consumption, Throughput, and Response Time

- Because this inbound behavior is so critical to performance of the overall communication stack, this presentation focuses heavily on this area.

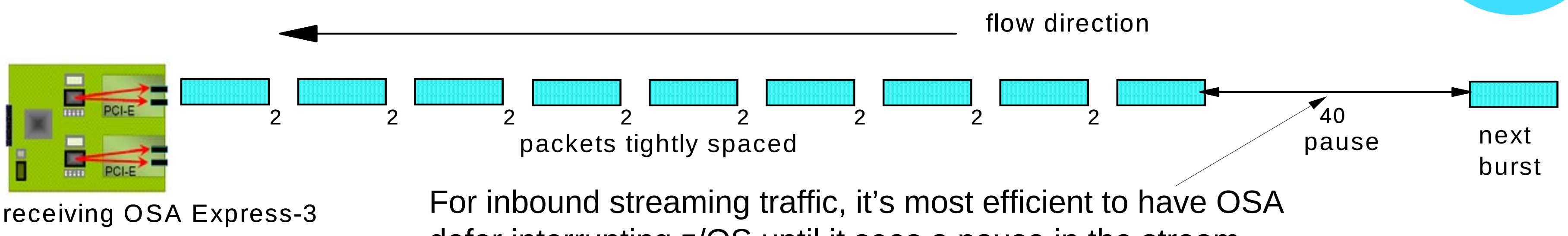
- So… let's get started by looking at two common network traffic patterns….



© 2011 SHARE and IBM Corporation

Timing Considerations for Various Inbound Loads…

Inbound Streaming Traffic Pattern

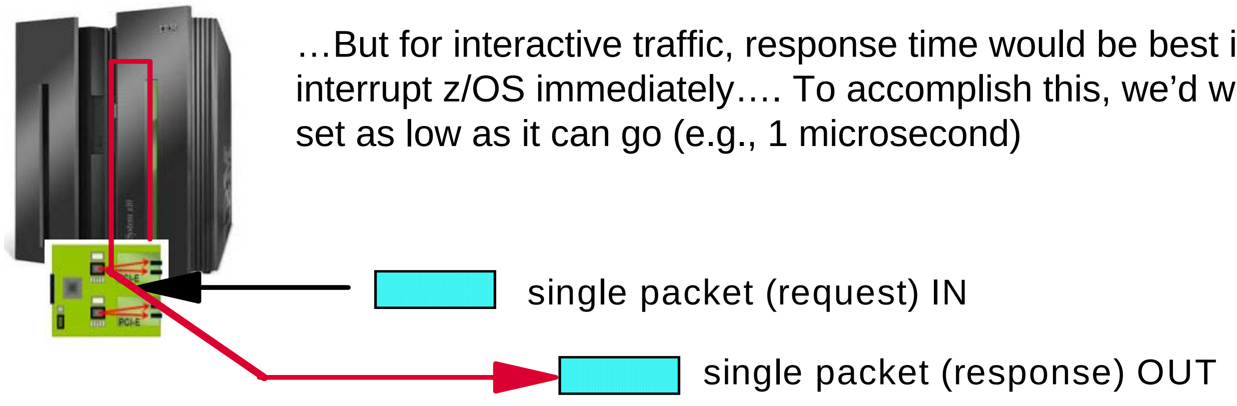


For inbound streaming traffic, it’s most efficient to have OSA defer interrupting z/OS until it sees a pause in the stream…….

(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)



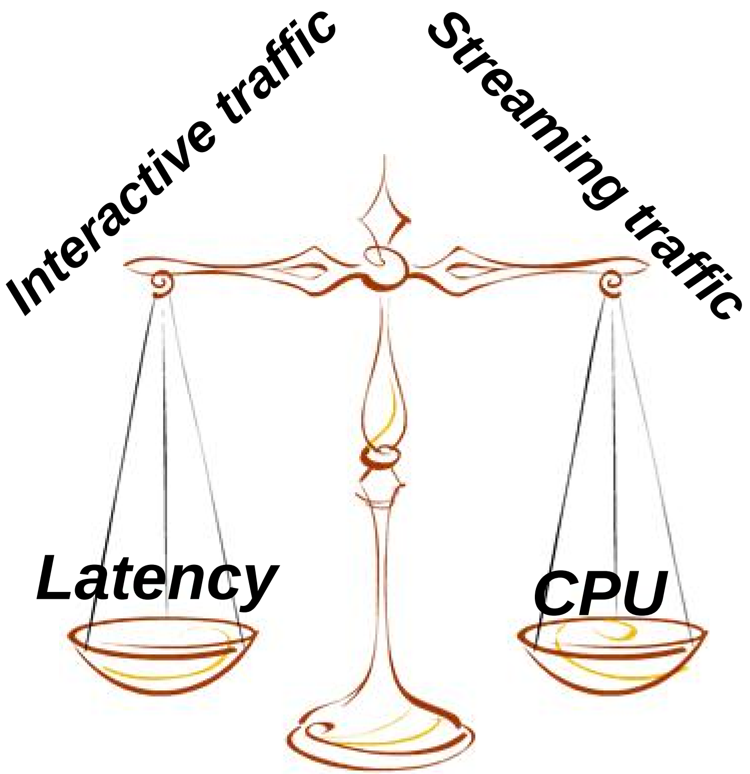
Read-Side interrupt frequency is all about the LAN-Idle timer!

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

© 2011 SHARE and IBM Corporation

Setting the Lan-Idle timer – A balancing act…

- Lowering the Lan-Idle timer:
  - Helps optimize latency for interactive traffic
  - But can increase CPU usage (more interrupts to process, more dispatches, etc.)
  - And what about streaming traffic?



- Increasing the the Lan-Idle timer:
  - Helps optimize CPU usage (less interrupts, dispatches)
  - Optimal for streaming traffic
  - But what latency for interactive traffic?

© 2011 SHARE and IBM Corporation

Setting the LAN Idle Timer – pre z/OS V1R9

- Prior to z/OS V1R9, Communications Server supported only static LAN-Idle timer settings

- On these earlier releases, you'd configure INBPERF on the INTERFACE or LINK statements

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

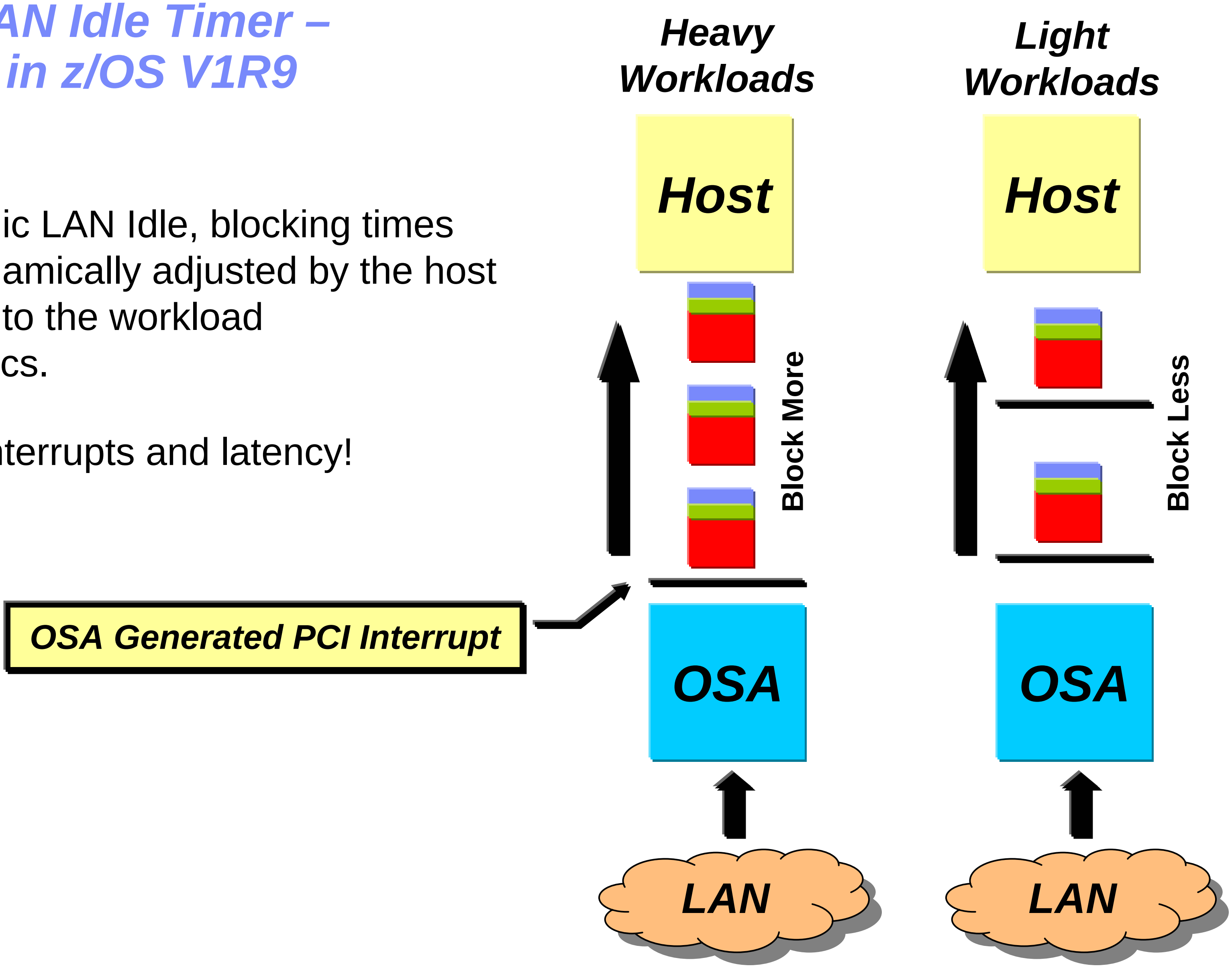
- **BALANCED** (default) - a static interrupt-timing value, selected to achieve reasonably high throughput and reasonably low CPU
- **MINCPU** - a static interrupt-timing value, selected to minimize host interrupts without regard to throughput
- **MINLATENCY** - a static interrupt-timing value, selected to minimize latency

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

© 2011 SHARE and IBM Corporation

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!



© 2011 SHARE and IBM Corporation

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!***

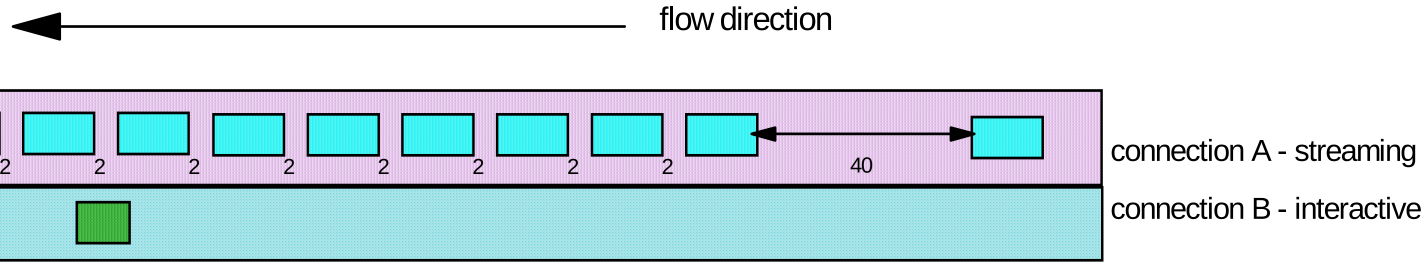
  – MINCPU - a static interrupt-timing value, selected to minimize host interrupts without regard to throughput

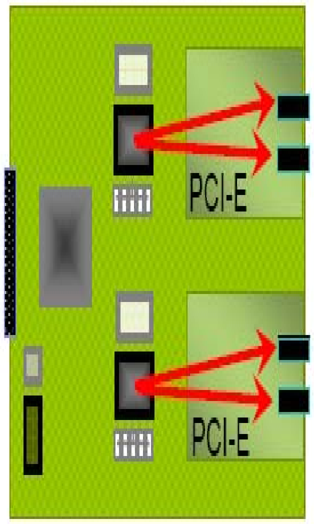
  – MINLATENCY - a static interrupt-timing value, selected to minimize latency

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

© 2011 SHARE and IBM Corporation

Dynamic LAN Idle Timer: But what about mixed workloads?





receiving OSA Express-3

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…..

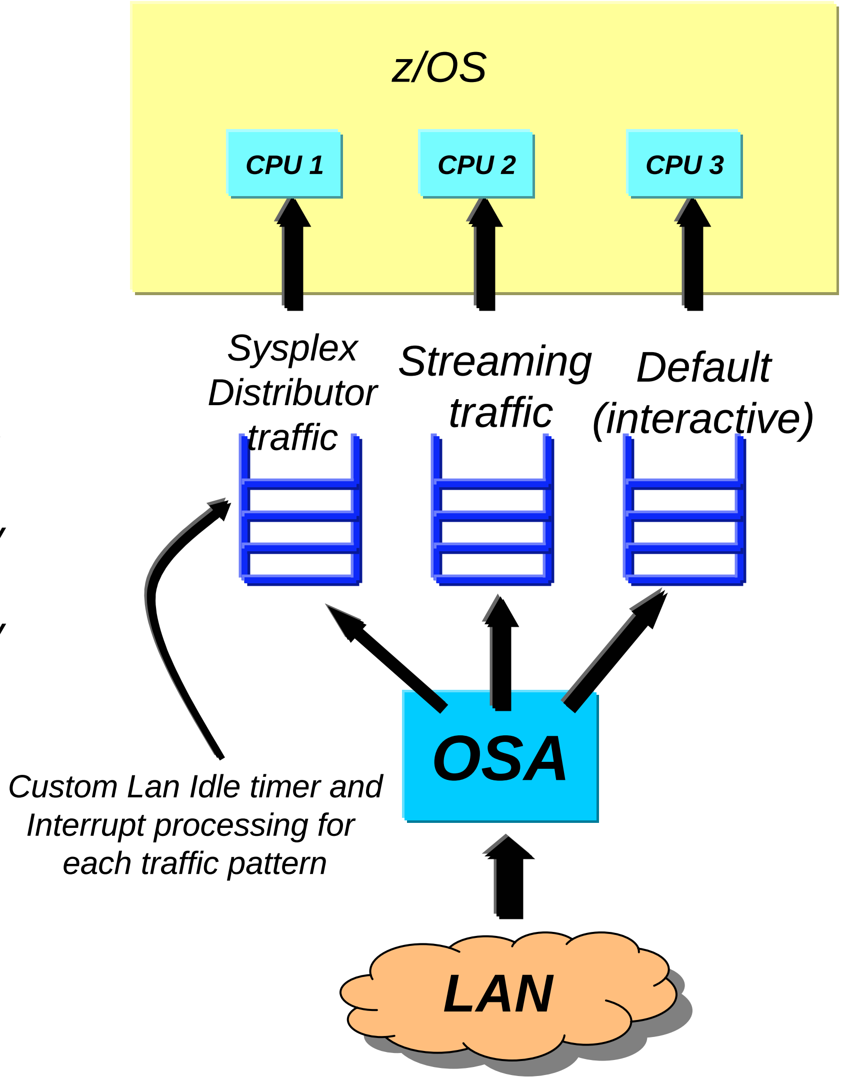
© 2011 SHARE and IBM Corporation

Extending Dynamic LAN Idle Timer: Inbound Workload Queueing (OSA-Express3 IWQ and z/OS V1R12)

With OSA-Express3 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.

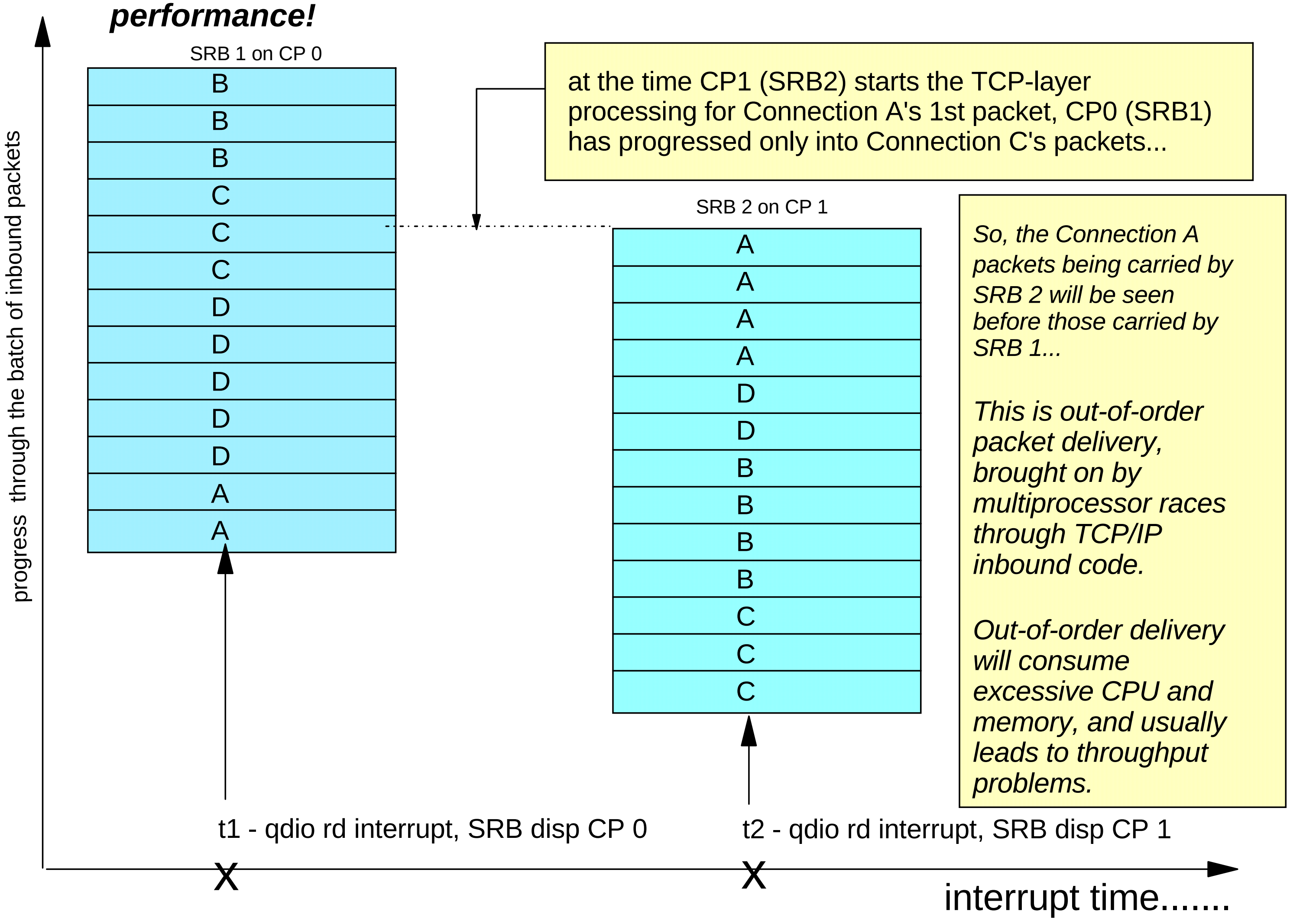
The separation of streaming traffic away from interactive also enables new streaming traffic efficiencies in Communications Server. This results in improved in-order delivery (better throughput and CPU consumption).


z/OS
CPU 1
CPU 2
CPU 3
Sysplex Distributor traffic
Streaming traffic
Default (interactive)
Custom Lan Idle timer and Interrupt processing for each traffic pattern
OSA
LAN

© 2011 SHARE and IBM Corporation

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.

© 2011 SHARE and IBM Corporation

QDIO Inbound Workload Queueing - Configuration

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

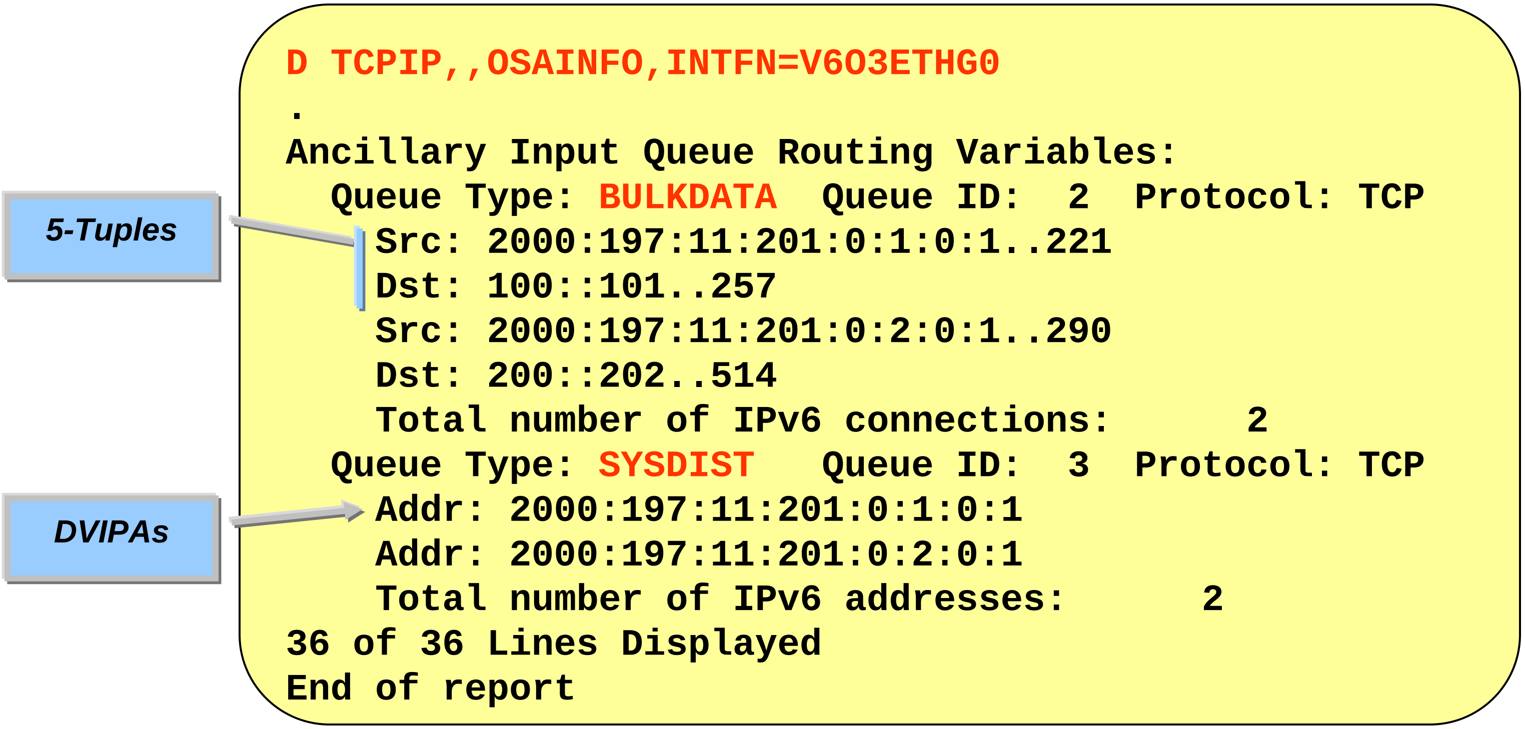
```
>>-INTERFace--intf_name------------------------------------------>
.
.-INBPERF BALANCED--------------------.
 >--+-------------------------------------+-->
    |                   .-NOWORKLOADQ-.   |
    ‘-INBPERF-+-DYNAMIC-+-------------+-+-’
              |         ‘-WORKLOADQ---’ |
          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

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

© 2011 SHARE and IBM Corporation

QDIO Inbound Workload Queueing

- 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 DVIPAs with OSA

© 2011 SHARE and IBM Corporation

QDIO Inbound Workload Queueing: Netstat DEvlinks/-d

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

```
D TCPIP,TCPCS1,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
```

© 2011 SHARE and IBM Corporation

QDIO Inbound Workload Queueing: 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
```

© 2011 SHARE and IBM Corporation

QDIO Inbound Workload Queueing: Netstat ALL/-A

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

```
D TCPIP,TCPCS1,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
```

© 2011 SHARE and IBM Corporation

QDIO Inbound Workload Queueing: Netstat STATS/-S

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

```
D TCPIP,TCPCS1,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
```

© 2011 SHARE and IBM Corporation

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.

- The new (in z/OS V1R12) 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. We'll list some usage considerations a bit later, ***but we do recommend IWQ mode.***

- So let's now move onto the one remaining timing-related OSA performance option: ***Optimized Latency Mode.***

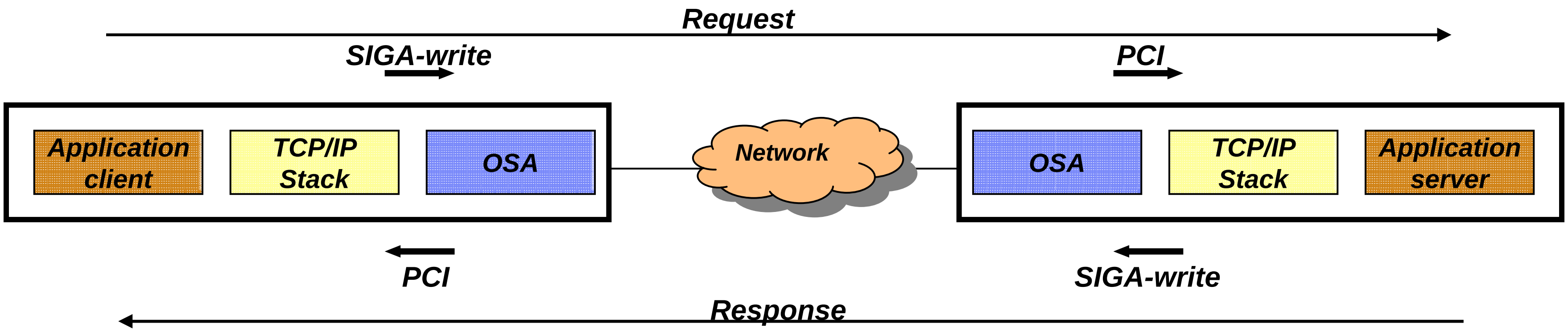
© 2011 SHARE and IBM Corporation

Optimized Latency Mode (OLM) – added in z/OS V1R11

- OSA-Express3's latency characteristics are much improved over OSA-Express2. Even so, z/OS software and OSA-Express3 microcode can further reduce latency via some aggressive processing changes (enabled via the OLM keyword on the INTERFACE statement):
  - Inbound
    - OSA-Express signals host if data is "on its way" ("Early Interrupt")
    - Host may spin for a while, if the early interrupt is fielded before the inbound data is "ready"
  - Outbound
    - OSA-Express does not wait for SIGA to look for outbound data ("SIGA reduction")
    - OSA-Express microprocessor may spin for a while, looking for new outbound data to transmit

- OLM is intended for workloads that have demanding QoS requirements for response time (transaction rate)
  - high volume interactive workloads (traffic is predominantly transaction oriented versus streaming)

- The latency-reduction techniques employed by OLM will limit the degree to which the OSA can be shared among partitions, and may also drive up z/OS CPU consumption



© 2011 SHARE and IBM Corporation

Optimized Latency Mode (OLM): How to configure

```
INTERFACE NSQDIO411 DEFINE IPAQENET
  IPADDR 172.16.11.1/24
  PORTNAME NSQDIO1
  MTU 1492 VMAC OLM
  INBPERF DYNAMIC
  SOURCEVIPAINTERFACE LVIPA1
```

- New OLM parameter
  - IPAQENET/IPAQENET6
  - **Not** allowed on DEVICE/LINK
- Enables Optimized Latency Mode for this INTERFACE only
- Forces INBPERF to DYNAMIC
- Default NOOLM

- Use Netstat DEvlinks/-d to see current OLM configuration

```
d tcpip,tcpcs,netstat,devlinks,intfname=lnsqdio1
JOB    6  EZD0101I NETSTAT CS V1R11 TCPCS
 INTFNAME: LNSQDIO1       INTFTYPE: IPAQENET   INTFSTATUS: READY
     .
     READSTORAGE: GLOBAL (4096K)      INBPERF: DYNAMIC
     .
     ISOLATE: NO                      OPTLATENCYMODE: YES
```

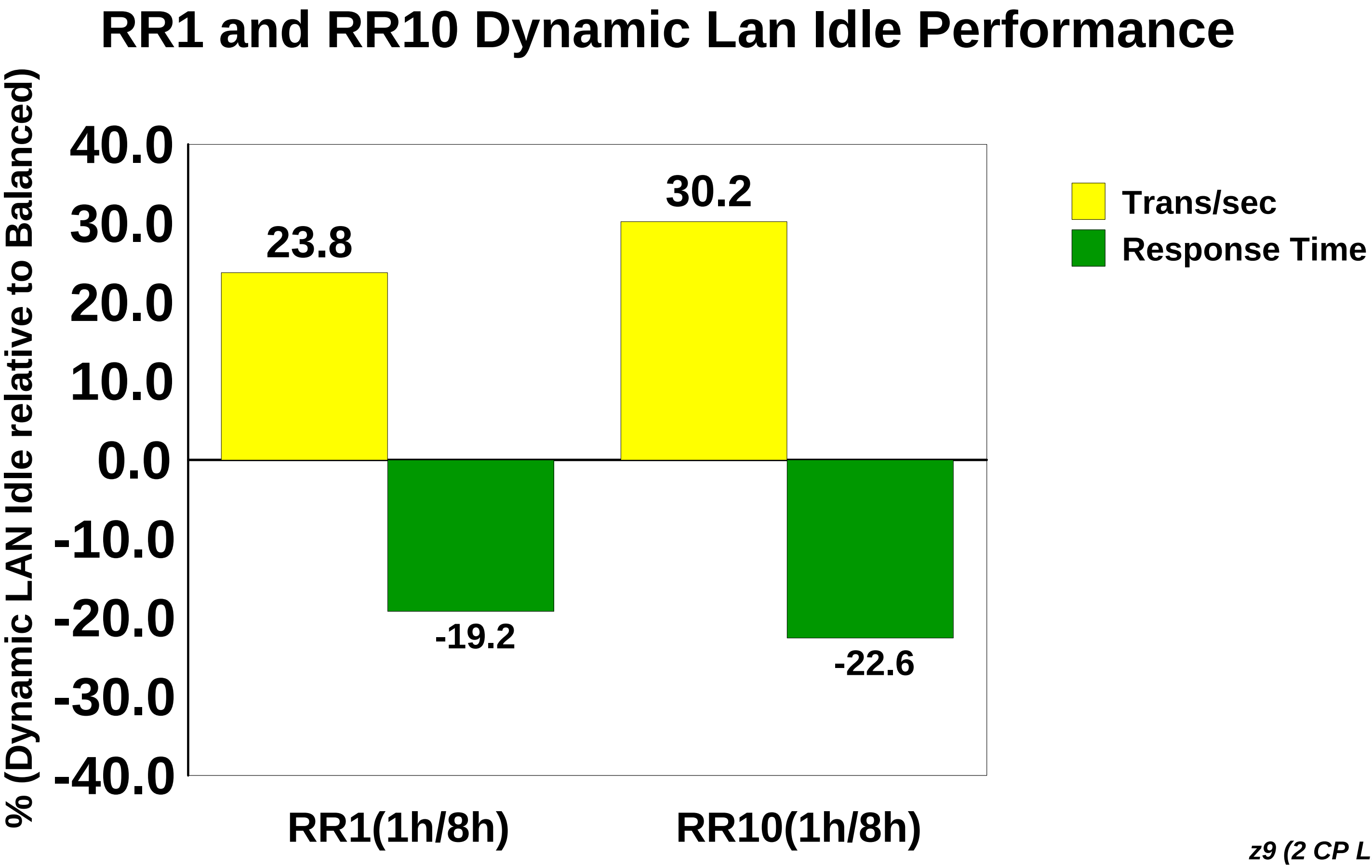
© 2011 SHARE and IBM Corporation

Performance Data

© 2011 SHARE and IBM Corporation

Dynamic LAN Idle Timer: Performance Data

Dynamic LAN Idle improved RR1 TPS 24% and RR10 TPS by 30%. Response Time for these workloads is improved 19% and 23%, respectively.

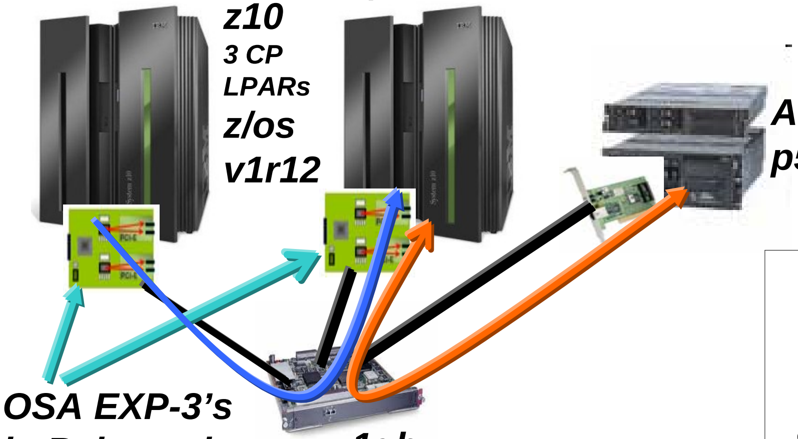


1h/8h indicates 100 bytes in and 800 bytes out

z9 (2 CP LPARs), z/OS V1R9, OSA-E2 1Gbe

© 2011 SHARE and IBM Corporation

Inbound Workload Queueing: Performance Data



z/os-A

z10
3 CP
LPARs
z/os
v1r12

z/os-B

Aix 5.3
p570

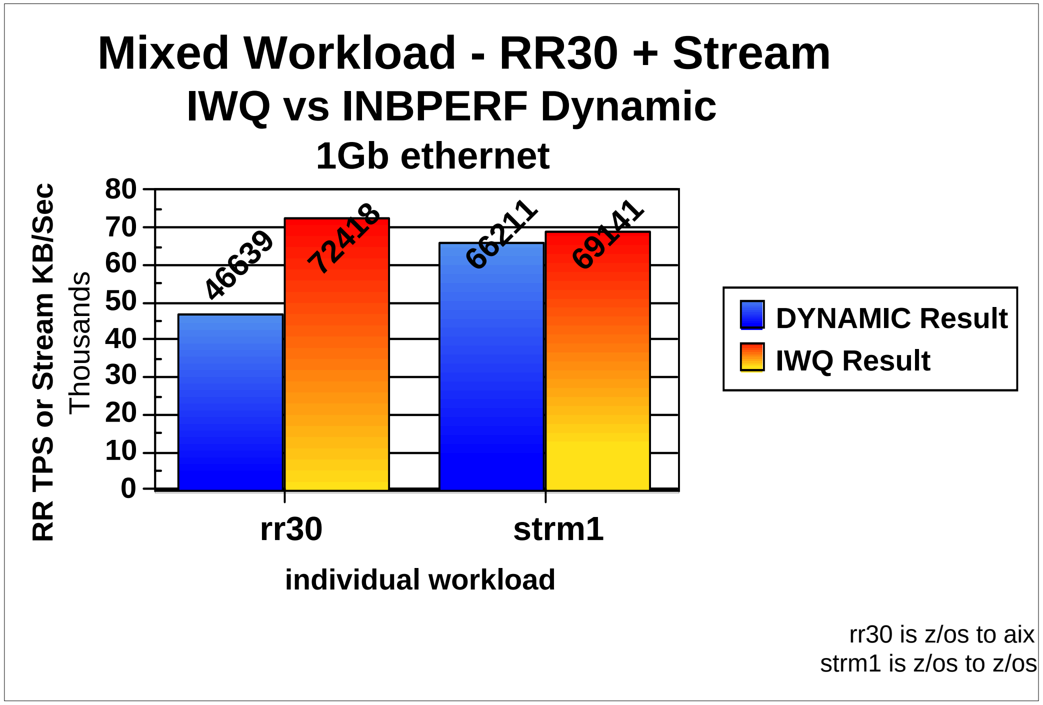
OSA EXP-3's in Balanced, Dynamic, or new IWQ mode

1gb or 10gb ethernet

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.

IWQ: Mixed Workload Results vs DYNAMIC:

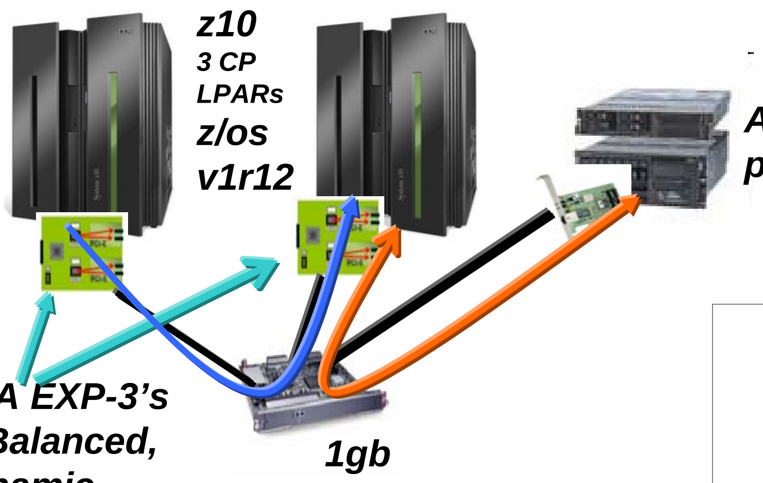
- ***z/OS<->AIX R/R Throughput improved 55% (Response Time improved 36%)***
- ***Streaming Throughput also improved in this test: +5%***



rr30 is z/os to aix
strm1 is z/os to z/os

© 2011 SHARE and IBM Corporation

Inbound Workload Queueing: Performance Data



z/os-A

z/os-B

z10
3 CP
LPARs
z/os
v1r12

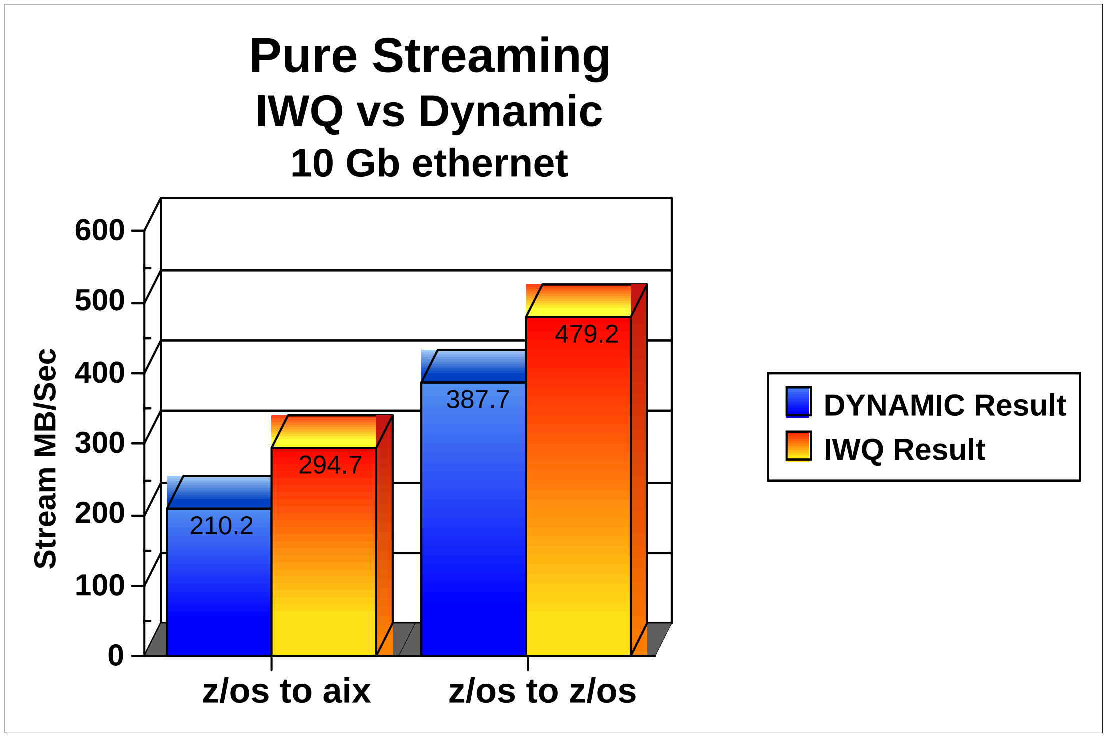
Aix 5.3
p570

OSA EXP-3's
in Balanced,
Dynamic,
or new
IWQ mode

1gb
or 10gb
ethernet

IWQ: Pure Streaming Results vs DYNAMIC:

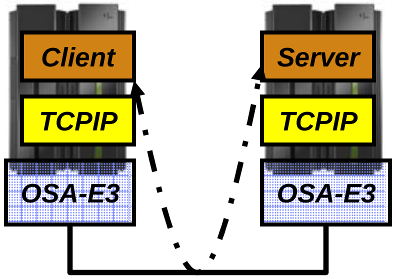
- ***z/OS<->AIX Streaming Throughput improved 40%***
- ***Z/OS<->z/OS Streaming Throughput improved 24%***



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.

© 2011 SHARE and IBM Corporation

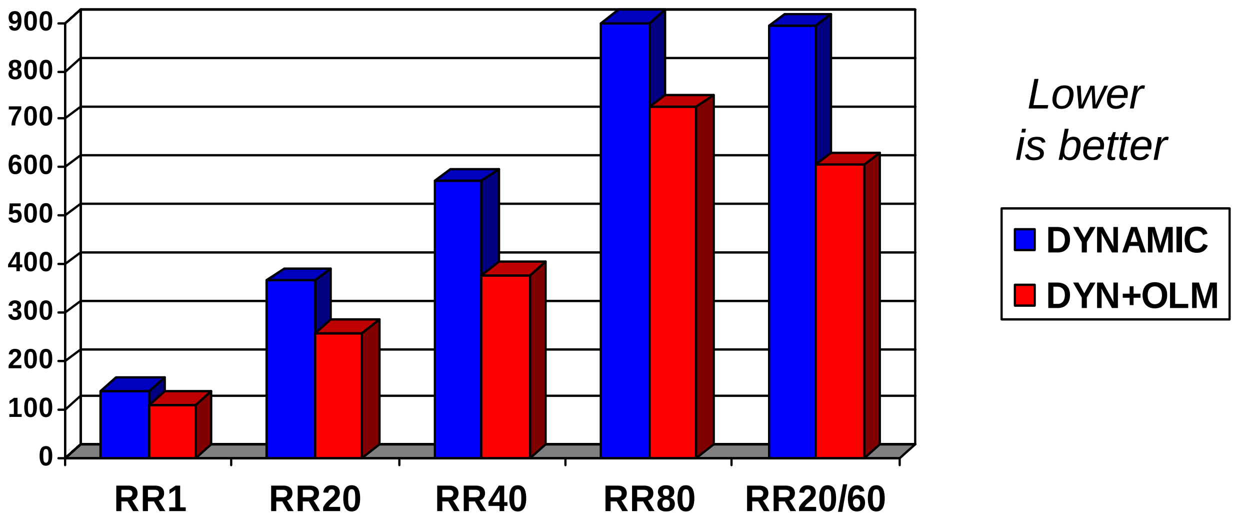
Optimized Latency Mode (OLM): Performance Data



z10 (4 CP LPARs), z/OS V1R11, OSA-E3 1Gbe

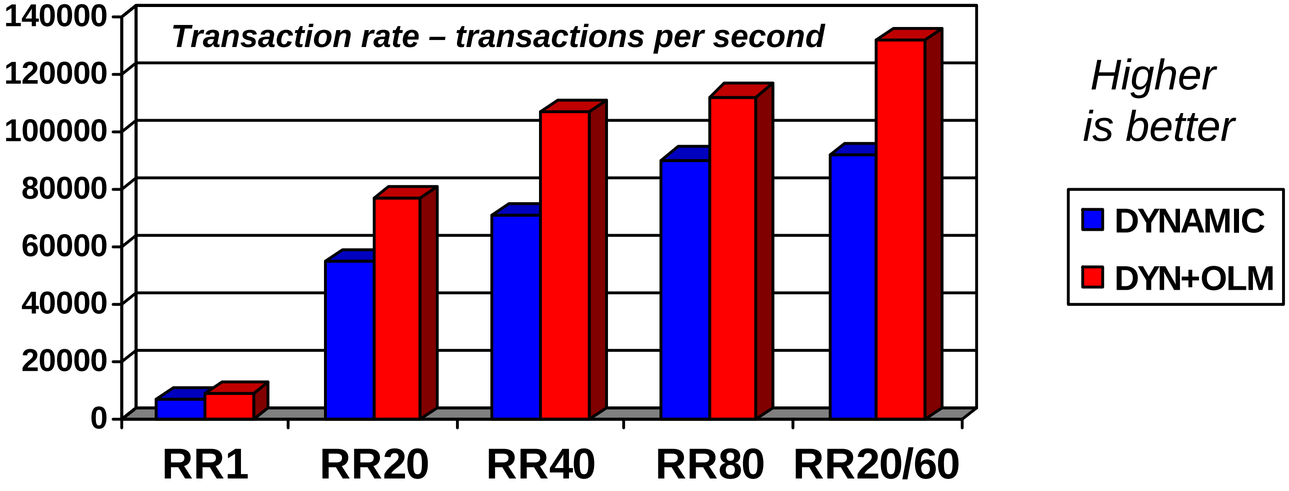
- **Client and Server**
  - Has close to no application logic
- **RR1**
  - 1 session
  - 1 byte in 1 byte out
- **RR20**
  - 20 sessions
  - 128 bytes in, 1024 bytes out
- **RR40**
  - 40 sessions
  - 128 bytes in, 1024 bytes out
- **RR80**
  - 80 sessions
  - 128 bytes in, 1024 bytes out
- **RR20/60**
  - 80 sessions
  - Mix of 100/128 bytes in and 800/1024 out

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



Lower is better

Transaction rate – transactions per second



Higher is better

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.

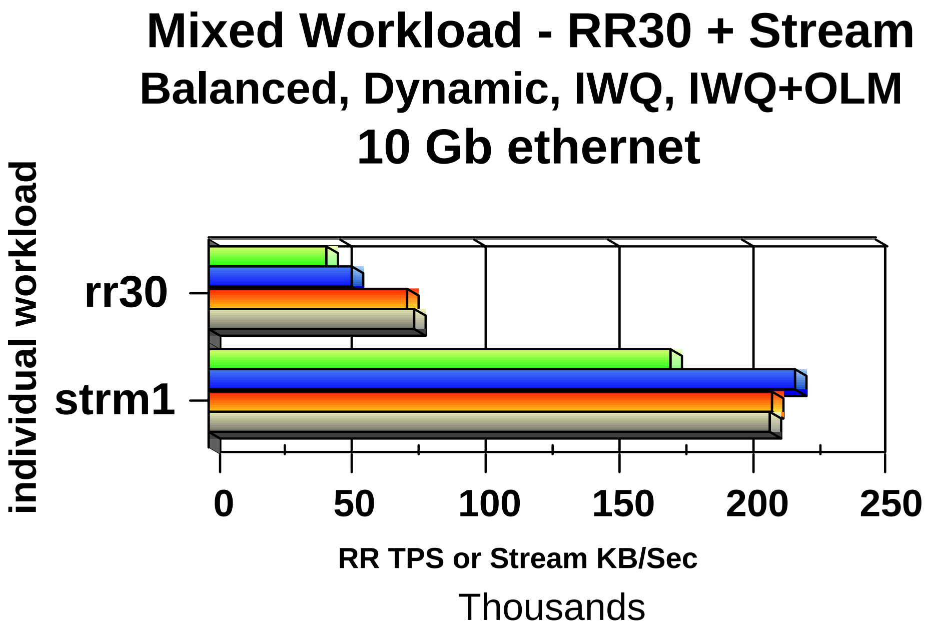
© 2011 SHARE and IBM Corporation

Combined IWQ + OLM: Performance Data for Mixed Workload

In z/OS V1R11, OLM usage was discouraged on z/OS images expected to be handling large amounts of streaming traffic. (OLM's 'early-interrupt' mechanism could significantly drive up CPU consumption for streaming workloads, while not providing any streaming throughput improvement.)

With the z/OS V1R12 IWQ design, OLM does not engage (nor would we want it to engage) on the streaming traffic queue. So the IWQ+OLM combination is not exposed to the CPU consumption increases that might be seen with OLM by itself.

In this 10Gb test, IWQ provided a 38% interactive throughput boost versus the dynamic setting. And the IWQ+OLM combination outperformed dynamic by 43%.



	rr30	strm1
🟩 Balanced	44461	173261
🟦 Dynamic	54071	220467
🟧 IWQ	74856	211866
▨ IWQ+OLM	77309	210944

z10 2 CP LPARs, OSA-E3, 10Gbe, and AIX 5.3 p570 client

rr30 is z/os to z/os
strm1 is aix to z/os

© 2011 SHARE and IBM Corporation

Detailed Usage Considerations for IWQ and OLM

© 2011 SHARE and IBM Corporation

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 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

© 2011 SHARE and IBM Corporation

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)

© 2011 SHARE and IBM Corporation

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-Express3. From a CPU-consumption perspective, OLM is therefore a more attractive option when combined with IWQ than without IWQ

- Only supported on OSA-Express3 with the INTERFACE statement

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

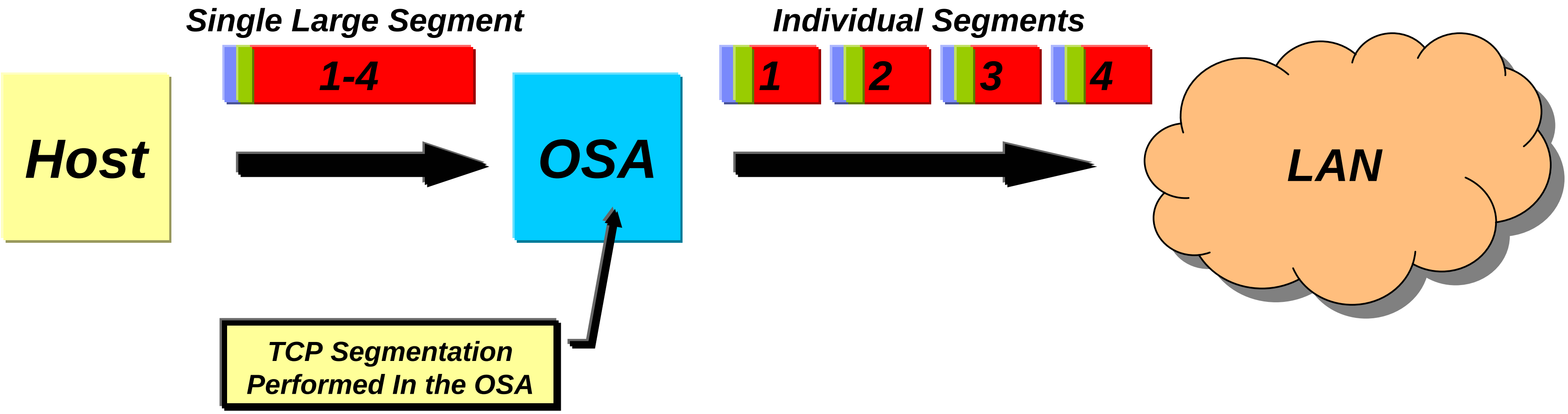
© 2011 SHARE and IBM Corporation

Optimizing outbound communications using OSA-Express

© 2011 SHARE and IBM Corporation

TCP Segmentation Offload

- Segmentation consumes (high cost) host CPU cycles in the TCP stack
- V1R7 (PTFed to V1R6) offered new OSA-Express (QDIO mode) feature Segmentation Offload (also referred to as "Large Send")
  - Offload most IPv4 TCP segmentation processing to OSA
  - Decrease host CPU utilization
  - Increase data transfer efficiency for IPv4 packets

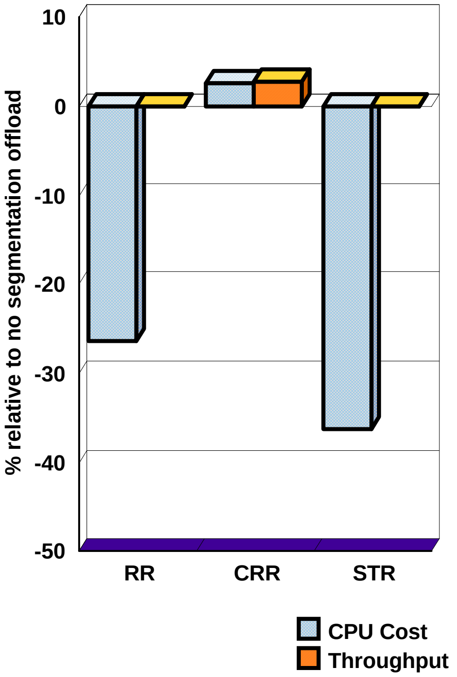


© 2011 SHARE and IBM Corporation

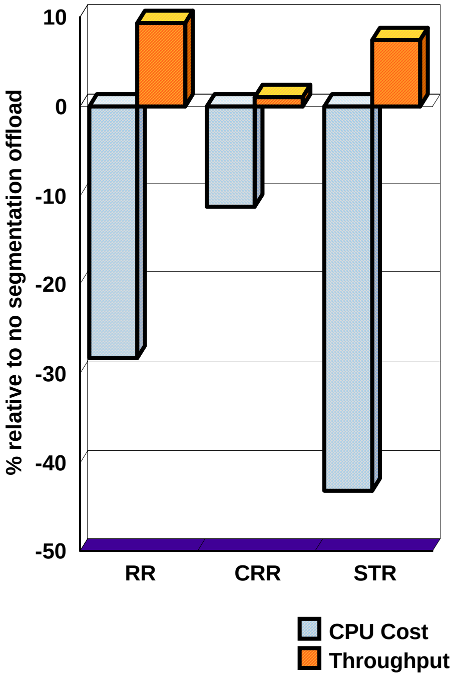
z/OS V1R10 segmentation offload performance measurements on a z10

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

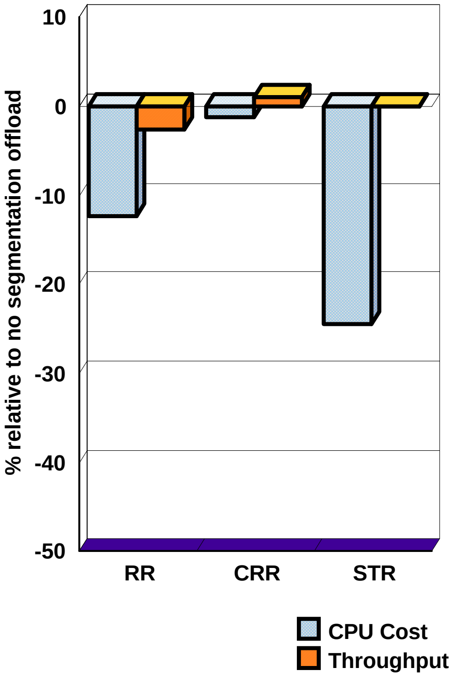
OSA Express3 1Gb



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 !

© 2011 SHARE and IBM Corporation

TCP Segmentation Offload: Configuration

- Enabled with GLOBALCONFIG SEGMENTATIONOFFLOAD

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

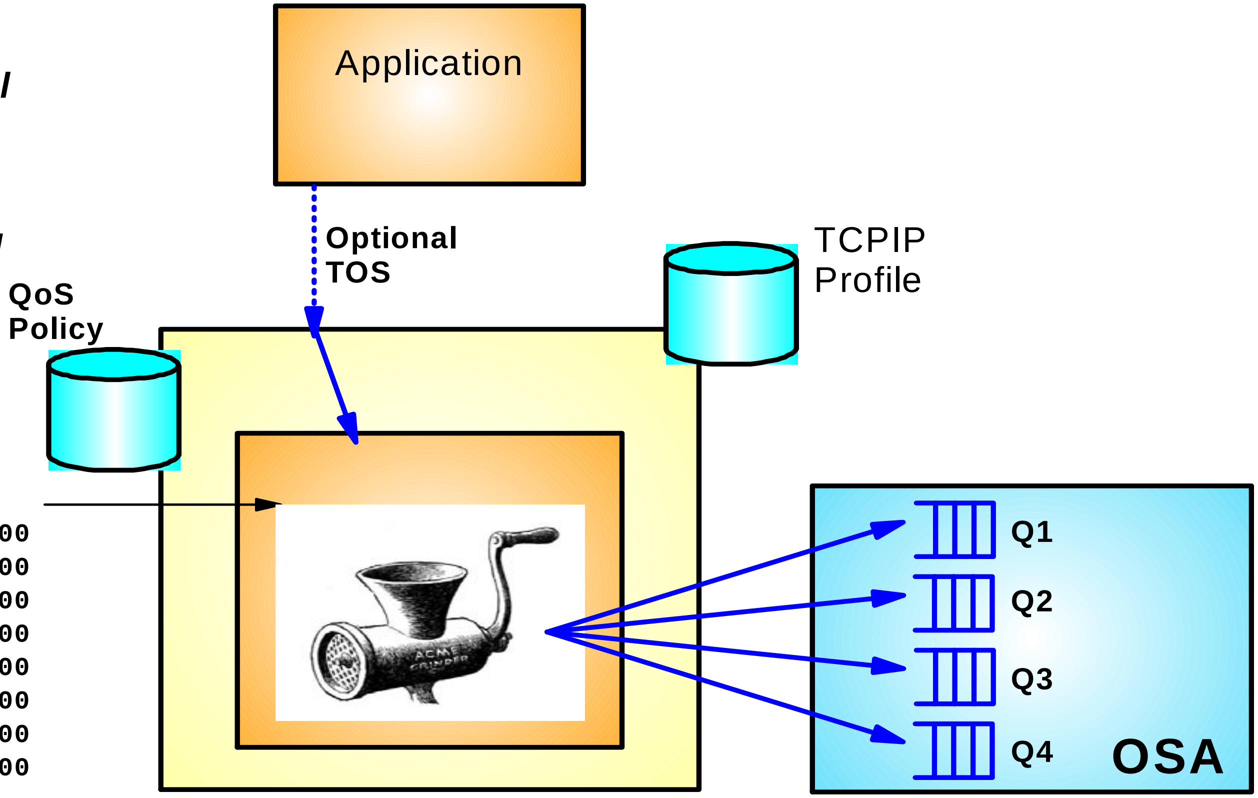
- Disabled by default
- TCP/IP stack will still do segmentation for
  - Packets going LPAR to LPAR
  - IPSec encapsulated packets
  - When multipath is in effect (unless all interfaces in the multipath group support segmentation offload)

© 2011 SHARE and IBM Corporation

OSA Express Outbound priority queuing

Prior to z/OS V1R11 you have the ability to control which outbound priority queue is used for your network traffic using TCP/IP configuration and QoS policies (PagenT)

```
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
}
```

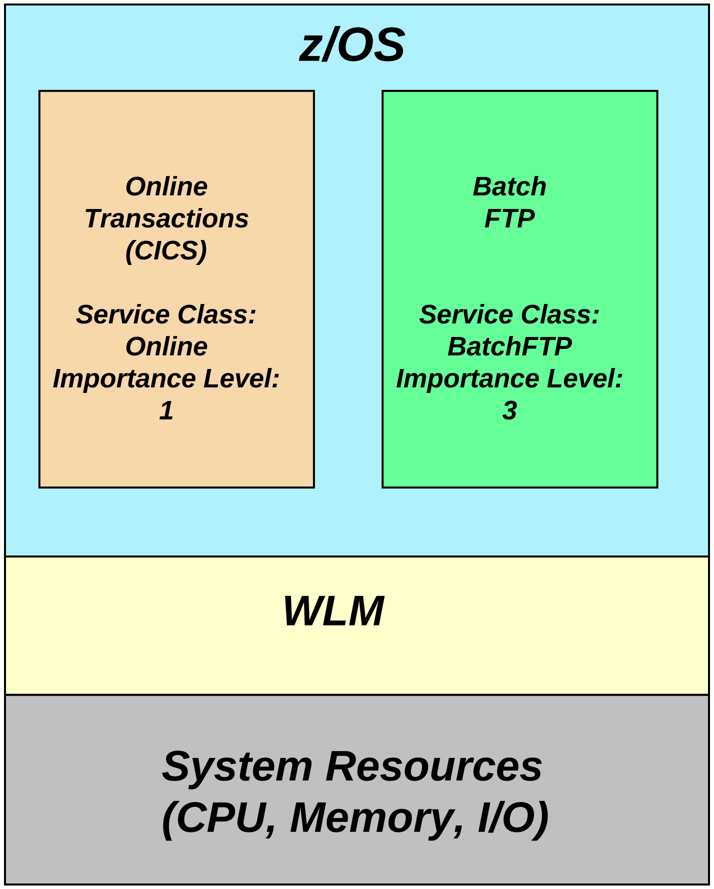


- While this feature allows for very flexible means of prioritizing outbound network traffic it has not been widely exploited by users
  - *How can we simplify its exploitation?*

© 2011 SHARE and IBM Corporation

z/OS Workload Manager (WLM)
Managing workloads of different business priorities

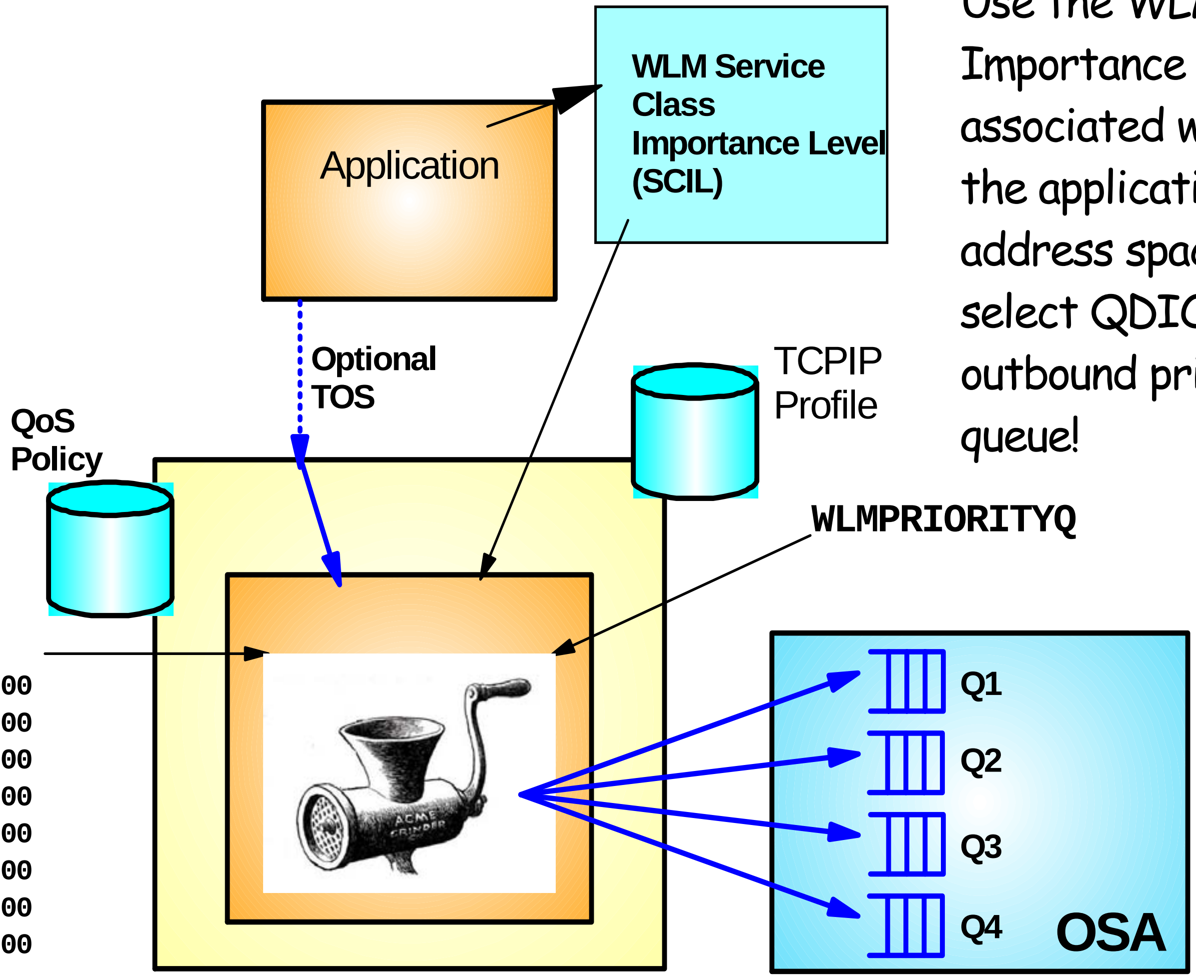
- WLM policy allows users to specify the business goals and priorities for all their z/OS workloads
  - Sysplex-wide goals
  - WLM manages key system resources (memory, CPU) to help workloads achieve their specified goals
  - What happens when resources are over-committed?
    - WLM begins prioritizing access to system resources based on the specified Importance Level of each Service Class associated with the workloads currently executing
      - Emphasis is placed on meeting the goals for the more important workloads
  - Over time WLM resource priority management has been expanded to also include I/O prorities (DASD and Tape)
    - But what about Network I/O priority?



© 2011 SHARE and IBM Corporation

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
}
```

© 2011 SHARE and IBM Corporation

The default QDIO priority queue mapping

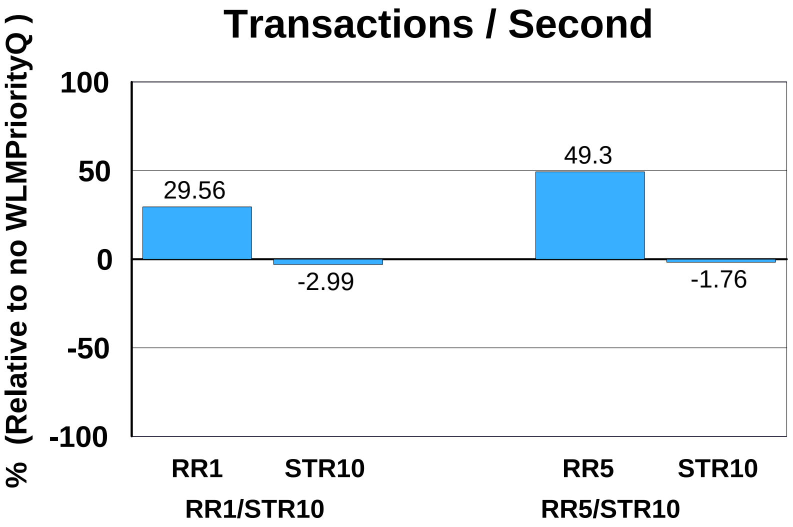
WLM Service classes	TCP/IP assigned	Default QDIO queue mapping
SYSTEM	n/a	Always queue 1
SYSSTC	0	Queue 1
User-defined with IL 1	1	Queue 2
User-defined with IL 2	2	Queue 3
User-defined with IL 3	3	Queue 3
User-defined with IL 4	4	Queue 4
User-defined with IL 5	5	Queue 4
User-defined with discretionary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

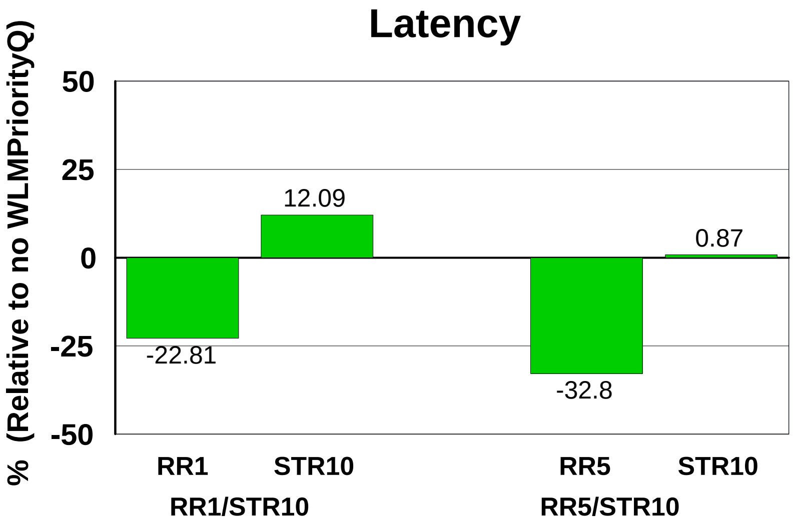
```
GLOBALCONFIG … WLMPRIORITYQ
   IOPRI1 0
   IOPRI2 1
   IOPRI3 2 3
   IOPRI4 4 5 6 FWD
```

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

© 2011 SHARE and IBM Corporation

OSA Express (QDIO) WLM Outbound Priority Queuing





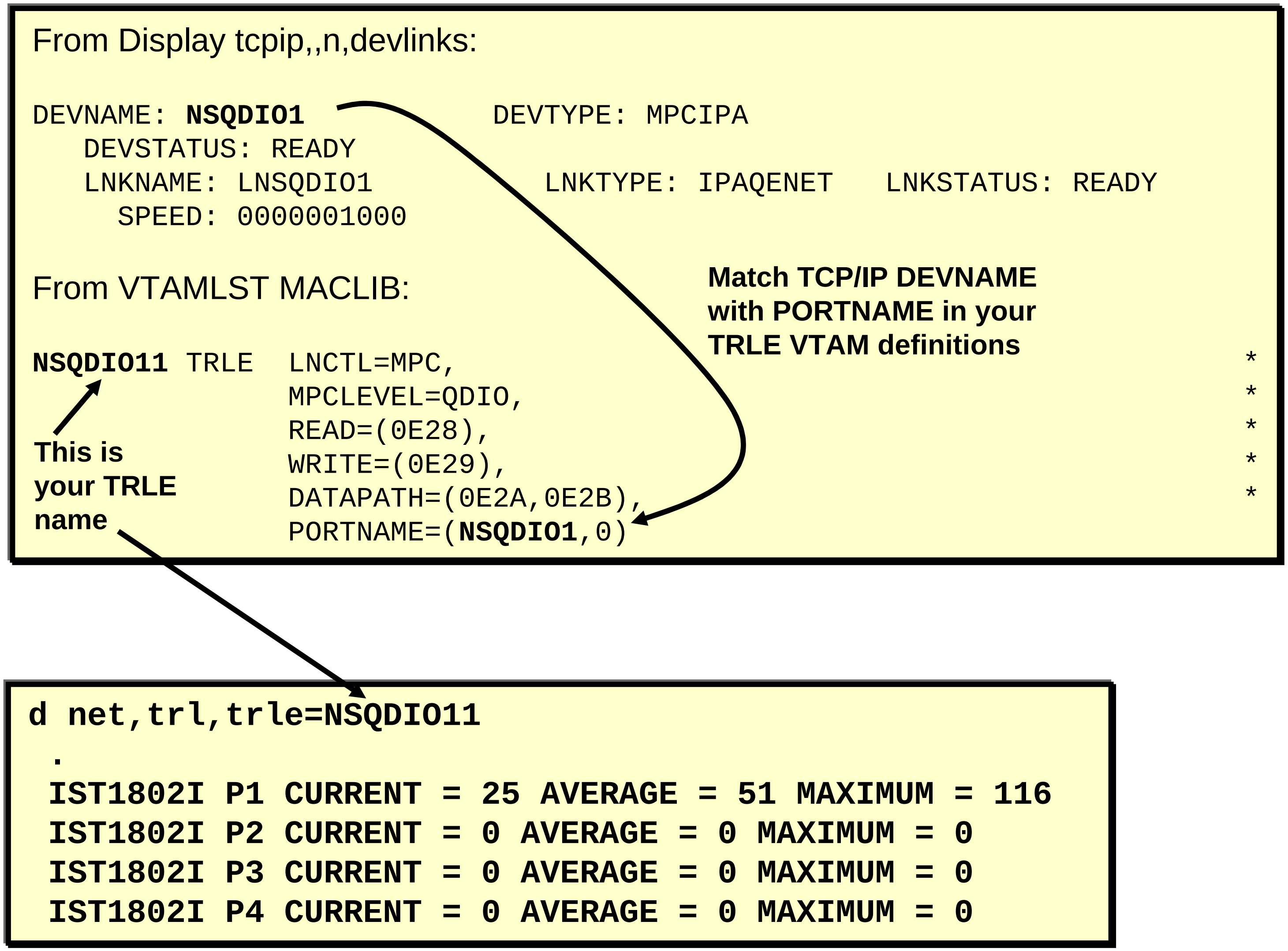
- Request-Response and Streaming mixed workload
- RR1/STR10: 1 RR session, 100 / 800 and 10 STR sessions, 1 / 20 MB
- RR5/STR10: 5 RR sessions, 100 / 800 and 10 STR sessions, 1 / 20 MB
- WLMPRIORITYQ assigned importance level 2 to interactive workloads and level 3 to streaming workloads
- The z/OS Workload Manager (WLM) system administrator assigns each job a WLM service class
- Hardware: z10 using OSA-E2 (1 GbE)
- Software: z/OS V1R11

- z/OS V1R11 with WLM I/O Priority provides 29.56 to 49.3% higher throughput for interactive workloads compared to V1R11 without WLM I/O Priority (Avg= 39.43% higher).
- z/OS V1R11 with WLM I/O Priority provides 22.81 to 32.8% lower latency compared to V1R11 without WLM I/O Priority (Avg= 27.80% lower).

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

© 2011 SHARE and IBM Corporation

Which QDIO priority queues are being used?



From Display tcpip,,n,devlinks:

```
DEVNAME: NSQDIO1          DEVTYPE: MPCIPA
   DEVSTATUS: READY
   LNKNAME: LNSQDIO1         LNKTYPE: IPAQENET   LNKSTATUS: READY
     SPEED: 0000001000
```

From VTAMLST MACLIB:

```
NSQDIO11 TRLE  LNCTL=MPC,                                               *
               MPCLEVEL=QDIO,                                           *
               READ=(0E28),                                             *
               WRITE=(0E29),                                            *
               DATAPATH=(0E2A,0E2B),                                    *
               PORTNAME=(NSQDIO1,0)
```

Match TCP/IP DEVNAME with PORTNAME in your TRLE VTAM definitions

This is your TRLE name

```
d net,trl,trle=NSQDIO11
 .
 IST1802I P1 CURRENT = 25 AVERAGE = 51 MAXIMUM = 116
 IST1802I P2 CURRENT = 0 AVERAGE = 0 MAXIMUM = 0
 IST1802I P3 CURRENT = 0 AVERAGE = 0 MAXIMUM = 0
 IST1802I P4 CURRENT = 0 AVERAGE = 0 MAXIMUM = 0
```

© 2011 SHARE and IBM Corporation

Example of enabling WLMPRIORITYQ

VTAM TNSTATS before enabling WLMPRIORITYQ

```
IST1233I DEV      = 2E02       DIR      = WR/1
..
IST1236I BYTECNTO =          0 BYTECNT  =         72
IST1810I PKTIQDO  =          0 PKTIQD   =          0
IST1811I BYTIQDO  =          0 BYTIQD   =          0
IST924I ---------------------------------------------
-
IST1233I DEV      = 2E02       DIR      = WR/2
..
IST1236I BYTECNTO =          0 BYTECNT  =          0
IST1810I PKTIQDO  =          0 PKTIQD   =          0
IST1811I BYTIQDO  =          0 BYTIQD   =          0
IST924I ---------------------------------------------
-
IST1233I DEV      = 2E02       DIR      = WR/3
..
IST1236I BYTECNTO =          0 BYTECNT  =          0
IST1810I PKTIQDO  =          0 PKTIQD   =          0
IST1811I BYTIQDO  =          0 BYTIQD   =          0
IST924I ---------------------------------------------
-
IST1233I DEV      = 2E02       DIR      = WR/4
..
IST1236I BYTECNTO =          0 BYTECNT  =      34738
IST1810I PKTIQDO  =          0 PKTIQD   =          0
IST1811I BYTIQDO  =          0 BYTIQD   =          0
```

VTAM TNSTATS after enabling WLMPRIORITYQ with defaults

```
IST1233I DEV      = 2E02       DIR      = WR/1
..
IST1236I BYTECNTO =          0 BYTECNT  =       1552
IST1810I PKTIQDO  =          0 PKTIQD   =          0
IST1811I BYTIQDO  =          0 BYTIQD   =          0
IST924I ---------------------------------------------
-
IST1233I DEV      = 2E02       DIR      = WR/2
..
IST1236I BYTECNTO =          0 BYTECNT  =      55421
IST1810I PKTIQDO  =          0 PKTIQD   =          0
IST1811I BYTIQDO  =          0 BYTIQD   =          0
IST924I ---------------------------------------------
-
IST1233I DEV      = 2E02       DIR      = WR/3
..
IST1236I BYTECNTO =          0 BYTECNT  =          0
IST1810I PKTIQDO  =          0 PKTIQD   =          0
IST1811I BYTIQDO  =          0 BYTIQD   =          0
IST924I ---------------------------------------------
-
IST1233I DEV      = 2E02       DIR      = WR/4
..
IST1236I BYTECNTO =          0 BYTECNT  =      90411
IST1810I PKTIQDO  =          0 PKTIQD   =          0
IST1811I BYTIQDO  =          0 BYTIQD   =          0
```

© 2011 SHARE and IBM Corporation

z/OS Communications Server Performance Summaries

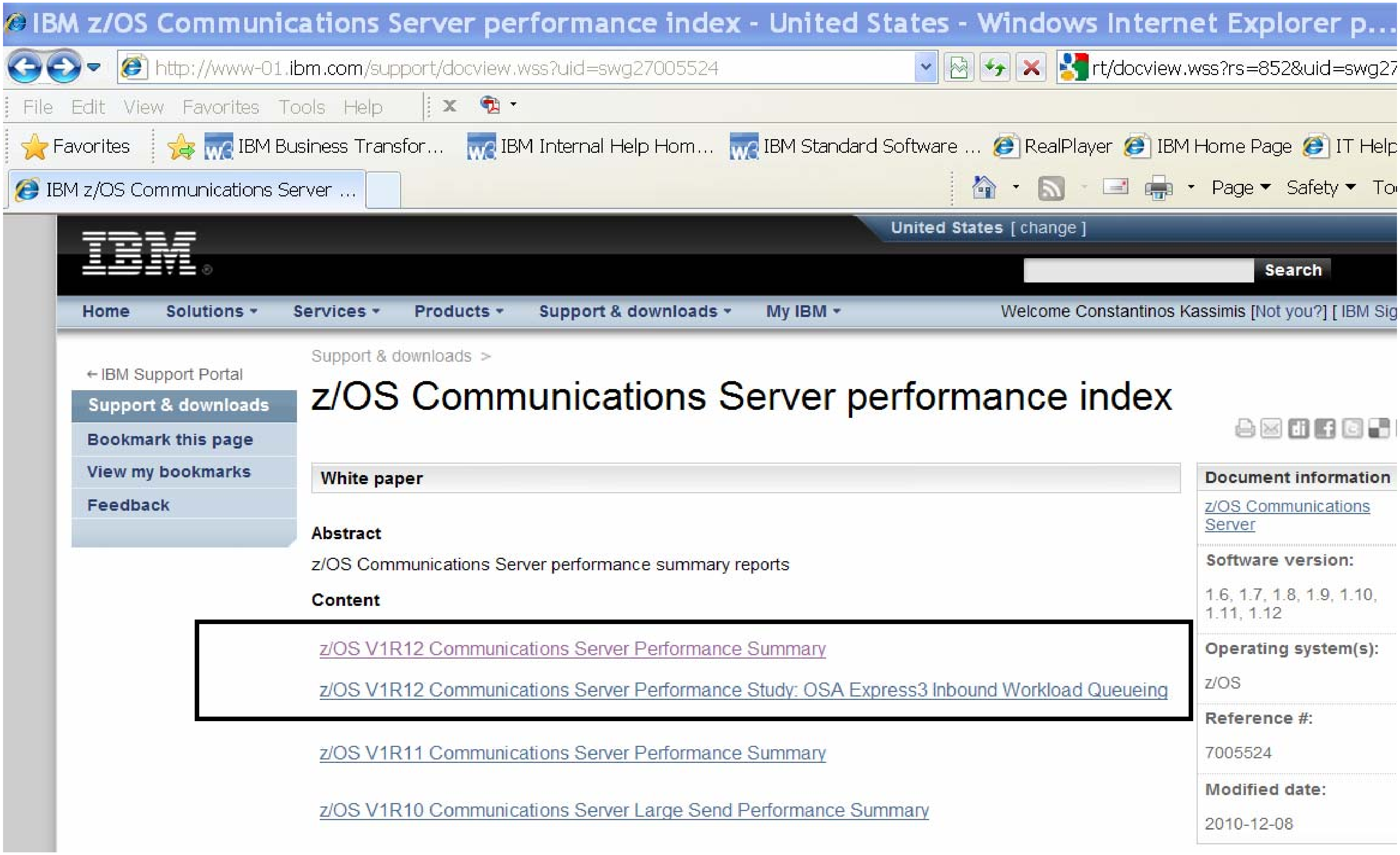
© 2011 SHARE and IBM Corporation

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 on line
  - http://www-01.ibm.com/support/docview.wss?rs=852&uid=swg27005524
- Ex: The z/OS V1R12 Communications Server Performance Summary includes:
  - The z/OS V1R12 Communications Server performance summary includes:
    - Performance of z/OS V1R12 Communications Server line items
    - Release to release performance comparisons (z/OS V1R12 Communications Server versus z/OS V1R11 Communications Server)
    - Capacity planning performance for:
      - TN3270 (Clear Text, AT-TLS, and IPSec)
      - FTP (Clear Text, AT-TLS, and IPSec)
      - CICS Sockets performance
    - CSM usage
    - VTAM buffer usage

© 2011 SHARE and IBM Corporation

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


IBM z/OS Communications Server performance index - United States - Windows Internet Explorer p...
http://www-01.ibm.com/support/docview.wss?uid=swg27005524
rt/docview.wss?rs=852&uid=swg27
File Edit View Favorites Tools Help
Favorites
IBM Business Transfor...
IBM Internal Help Hom...
IBM Standard Software ...
RealPlayer
IBM Home Page
IT Help
IBM z/OS Communications Server ...
Page
Safety
To
IBM
United States [change]
Search
Home
Solutions
Services
Products
Support & downloads
My IBM
Welcome Constantinos Kassimis [Not you?] [IBM Sig
← IBM Support Portal
Support & downloads
Bookmark this page
View my bookmarks
Feedback
Support & downloads >

z/OS Communications Server performance index

White paper

Abstract

z/OS Communications Server performance summary reports

Content

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

Document information

z/OS Communications Server

Software version:

1.6, 1.7, 1.8, 1.9, 1.10, 1.11, 1.12

Operating system(s):

z/OS

Reference #:

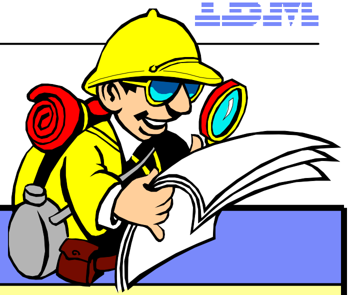
7005524

Modified date:

2010-12-08

© 2011 SHARE and IBM Corporation

For more information



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

For pleasant reading ….

© 2011 SHARE and IBM Corporation