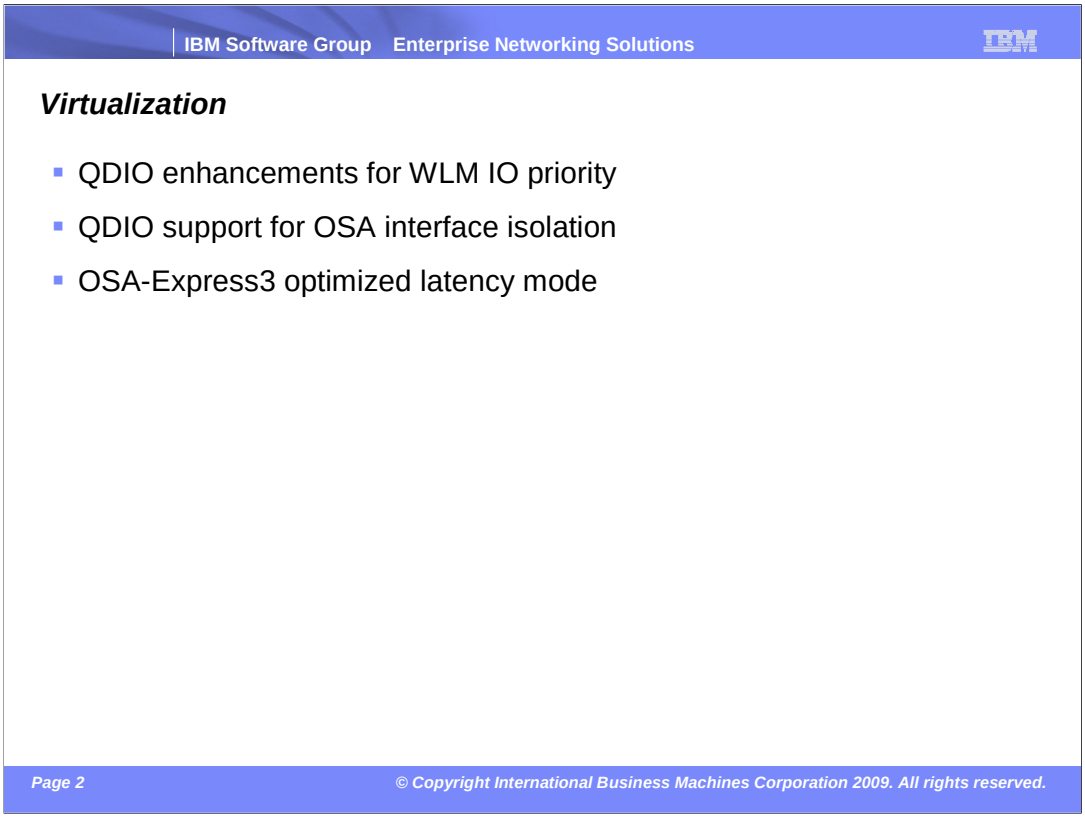
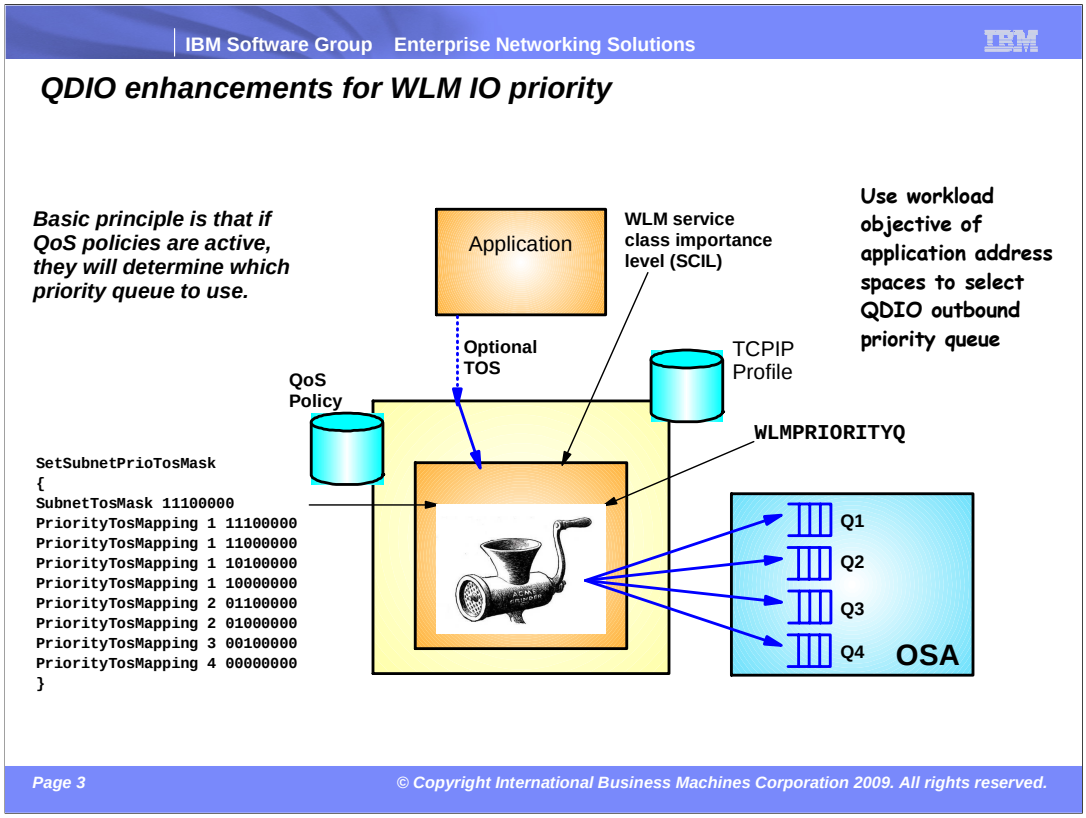
This presentation describes the enhancements to the Communications Server in z/OS V1R11 for virtualization (hardware). The virtualization theme in this release of Communications Server is related to OSA.



z/OS V1R11 Communications Server enhancements provide a simple mechanism to select QDIO outbound priority queue. The new support allows installations who do not use QoS networking policies to set QDIO outbound queue priority based on Workload Manager (WLM) input. This support can be used to extend the WLM importance level of the application that sends data to the selection of a QDIO outbound queue.

Also new in z/OS V1R11 is support for OSA isolation to prevent LPARs that share an OSA port from sending data to each other without sending it out onto the network.

And finally a new OSA operational mode known as Optimized Latency Mode (OLM) is being introduced in this release to accompany new OSA microcode.



Write queues are effectively staging areas for packets while the OSA-Express is writing them. The lower the priority queue number, the more resources it will receive from the OSA-Express. The TCP/IP stack assigns the write priority, which is based upon a Type Of Service (TOS) setting that might come from the QoS policy component or the application itself (for example, EE). The device driver honors the priority by using the appropriate staging queue.

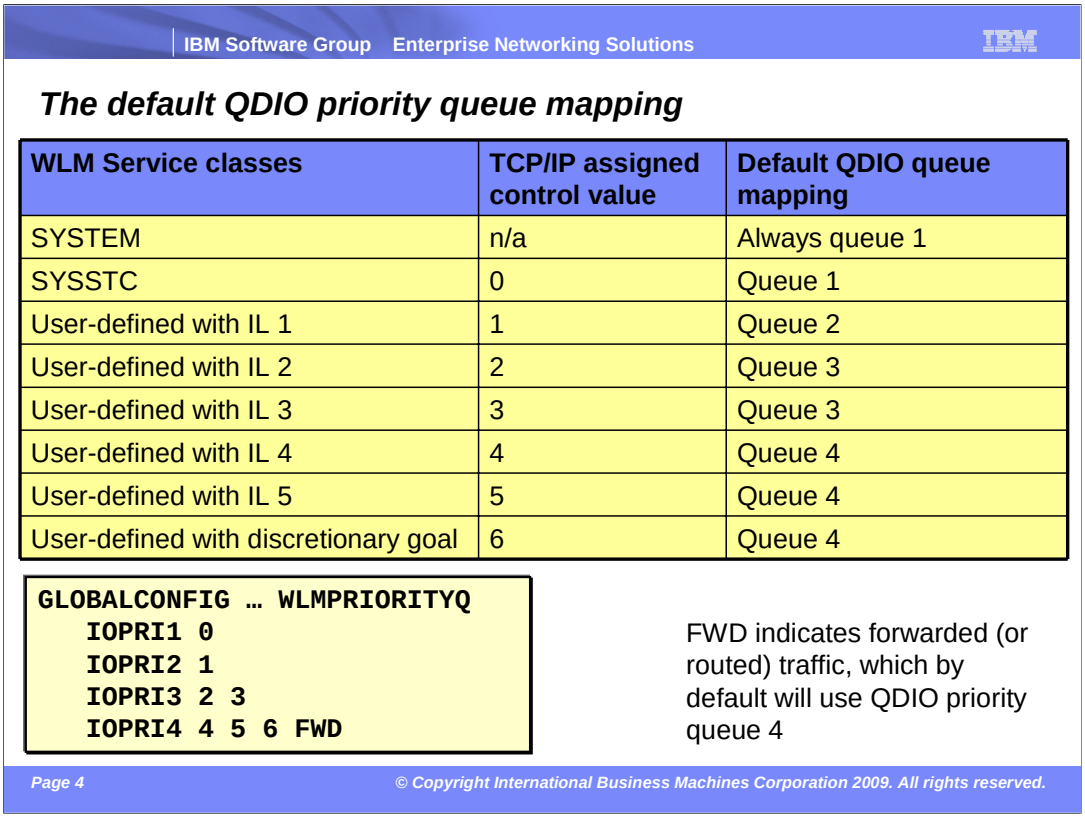
Many shops do little to prioritize their OSA-Express outbound data, missing any benefits the prioritization provides.

WLM requires the system administrator to assign each job a WLM service class, which includes a WLM importance level indicating how important each application or job is to the business.

With z/OS V1R11 Communications Server, a new GLOBALCONFIG profile statement establishes a mapping of WLM service class importance levels to outbound QDIO priorities.

Using the new WLMPRIORITYQ parameter allows you to map outbound OSA-Express data with an IPv4 Type of Service (ToS) byte or IPv6 Traffic Class of zeros.

Using already established importance levels eases the extension of this prioritization through z/OS and z/OS Communications Server, using OSA-Express and onto the LAN. In addition, WLMPRIORITYQ allows the outbound priority to be applied to forwarded packets containing a ToS or Traffic Class of zeros.



Packets from jobs with a WLM SYSTEM service class are *always* written on OSA-Express priority queue 1 when WLMPRIORITYQ is enabled and the ToS/Traffic class is zero. This is not displayed on a netstat config report nor is it customizable.

IOPRI1 0 OSA-Express priority queue 1 is used for packets from jobs with a *control value* 0 (SYSSTC).

IOPRI2 1 OSA-Express priority queue 2 is used for packets from jobs with a *control value* 1 (services classes with importance level 1).

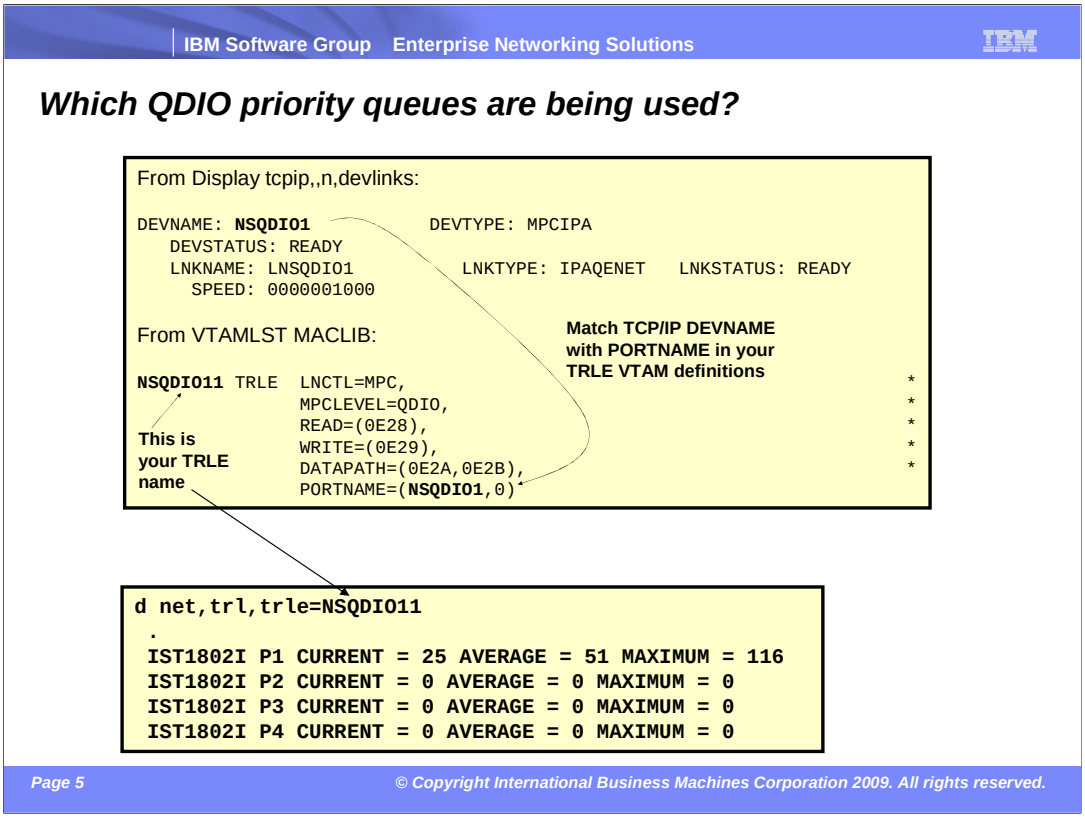
IOPRI3 2 3 OSA-Express priority queue 3 is used for packets from jobs with *control values* 2 and 3 (services classes with importance levels 2 and 3).

IOPRI4 4 5 6 FWD OSA-Express priority queue 4 is used for packets from jobs with *control values* 4, 5, 6 and all non-accelerated forwarded packets. Control values 4, 5, and 6 are services classes with importance levels 4 and 5 and discretionary.

You can instead create your own mapping. Unspecified control values default to QDIO priority 4.

The WLMPRIORITYQ settings can be changed dynamically using an OBEYFILE command.

You can either use the VTAM® Display trlename command or VTAM trlename tuning statistics to measure the effects of QDIO enhancements for WLM IO priority.



DEVNAME is **NSQDIO1** which matches the portname of the TRLE named **NSQDIO11.** *trlename* is therefore **NSQDIO11**.

Use VTAM command Display NET,TRL,TRLE=*trlename*. Look for message number IST1802I which reflects the distribution of work across the four QDIO priority queues.

As shown in this example, QDIO priority queue 1 is the only queue with activity. This is a good example of underutilized prioritization. Traffic prioritization in this case is first in, first out.

Analyzing tuning statistics for QDIO priority queues

F vtam,tnstat,trle=NSQDIO11,cnsl,time=1

```
IST924I -------------------------------------------
IST1233I DEV      = 0E2A       DIR      = WR/1
IST1755I SBALMAX  =          0 SBALAVG  =          0
IST1756I QDPTHMAX =          0 QDPTHAVG =          0
IST1723I SIGACNTO =          0 SIGACNT  =          0
IST1721I SBALCNTO =          0 SBALCNT  =          0
IST1722I PACKCNTO =          0 PACKCNT  =          0
IST2242I SIGMCNTO =          0 SIGMCNT  =          0
IST1236I BYTECNTO =          0 BYTECNT  =          0
IST1810I PKTIQDO  =          0 PKTIQD   =          0
IST1811I BYTIQDO  =          0 BYTIQD   =          0
IST924I -------------------------------------------
IST1233I DEV      = 0E2A       DIR      = WR/2
```


© Copyright International Business Machines Corporation 2009. All rights reserved.

Continuing with the example from the previous slide, issue the F vtam,tnstat,trle=NSQDIO11,cnsl,time=1 command to see VTAM tuning statistics.

VTAM tuning statistics accumulates counters for a specified interval. In this case time=1 so the interval is 1 minute. At the end of the interval, VTAM displays the counters at the console (due to the CNSL option) and resets the counters for the next interval. These counters will show a more accurate distribution of the workload across the 4 data device write priority queues (identified by WR/1, WR/2, WR/3, and WR/4).

These counters are very specific to device driver internal processing, but the BYTECNTO and BYTECNT counters show how many bytes have flowed across this priority queue in the last interval. BYTECNT is the number of bytes that have been written on this priority queue in that interval. BYTECNTO is incremented by 1 every time BYTECNT overflows (is incremented beyond 2^{32}).

Use the ‘NOTNSTAT’ command to disable tuning statistics and prevent console flooding.

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
```

Page 7 © Copyright International Business Machines Corporation 2009. All rights reserved.

This slide uses a F NET,TNSTAT,TRLE=OSAQ4,CNSL,TIME=1 command to verify the effect of enabling WLMPRIORITYQ on a sample z/OS system.

Without WLMPRIORITYQ almost all outbound traffic goes to queue 4, while the traffic is more evenly spread over the priority queues with WLMPRIORITYQ enabled.

This example shows an easy way to verify that WLMPRIORITYQ works.

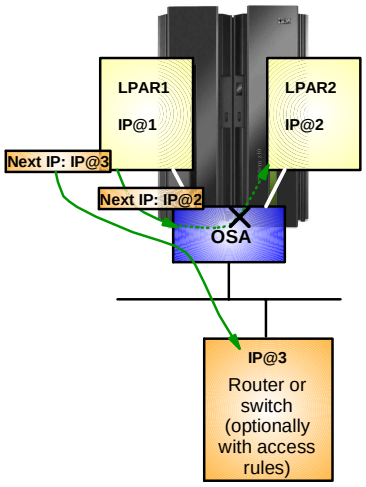
A D NET,TRL,TRLE=OSAQ4 command shows the same pattern:

```
IST1802I P1 CURRENT = 0 AVERAGE = 1 MAXIMUM = 3
IST1802I P2 CURRENT = 0 AVERAGE = 2 MAXIMUM = 2
IST1802I P3 CURRENT = 0 AVERAGE = 0 MAXIMUM = 0
IST1802I P4 CURRENT = 0 AVERAGE = 2 MAXIMUM = 3
```

OSA interface isolation

- New function added to the OSA adapter
  - z/OS Communications Server adds support for this new function in z/OS V1R11
- Allow customers to disable shared OSA local routing functions
  - ISOLATE/NOISOLATE option on QDIO network interface definition
- OSA local routing can in some scenarios be seen as a security exposure
- Depends on OSA MCL update

Be carefull using ISOLATE if you use OSPF and share a subnet between stacks that share an OSA port.



If you enable ISOLATE, packets with a nexthop IP address of another stack that share the OSA port, will be discarded.

© Copyright International Business Machines Corporation 2009. All rights reserved.

In some environments where strict control over routing between IP nodes must be enforced, the loop-back feature of a shared OSA port can prevent such rules from being enforced. For example, you send an IP packet from LPAR1 to a home IP address of LPAR2 (without VLAN tagging or attached to the same VLAN). In that case OSA will send that packet up to LPAR2 directly without sending it out to the switch. If the switch is there to enforce access rules, that behavior is an issue.

OSA-Express connection isolation provides a way for a stack using an OSA-Express to prevent packets from flowing directly between two stacks sharing the OSA. When connection isolation is in effect, OSA-Express will discard any packets when the next hop address was registered by a sharing stack. OSA-Express requires that both stacks sharing the port be non-isolated for direct routing to occur.

OSA-Express connection isolation is only supported for OSA-Express features in QDIO mode.

OSA-Express connection isolation is not supported when the OSA-Express is defined using a DEVICE and LINK statement.

OSA interface isolation notes

- Isolation does not prevent traffic on other interfaces between stacks
- Dynamic routing
- Static routing
- Alternatives to the ISOLATE function
  - OSA internal routing
  - VLAN
- OSA-Express2 or OSA-Express3 Ethernet features in QDIO mode
- Minimum IBM System z9® EC or BC

© Copyright International Business Machines Corporation 2009. All rights reserved.

Isolation only prevents direct routing over OSA-Express QDIO. It does not prevent traffic from flowing between these stacks over another interface, such as HiperSockets™, an MPCPTP connection, or an XCF connection.

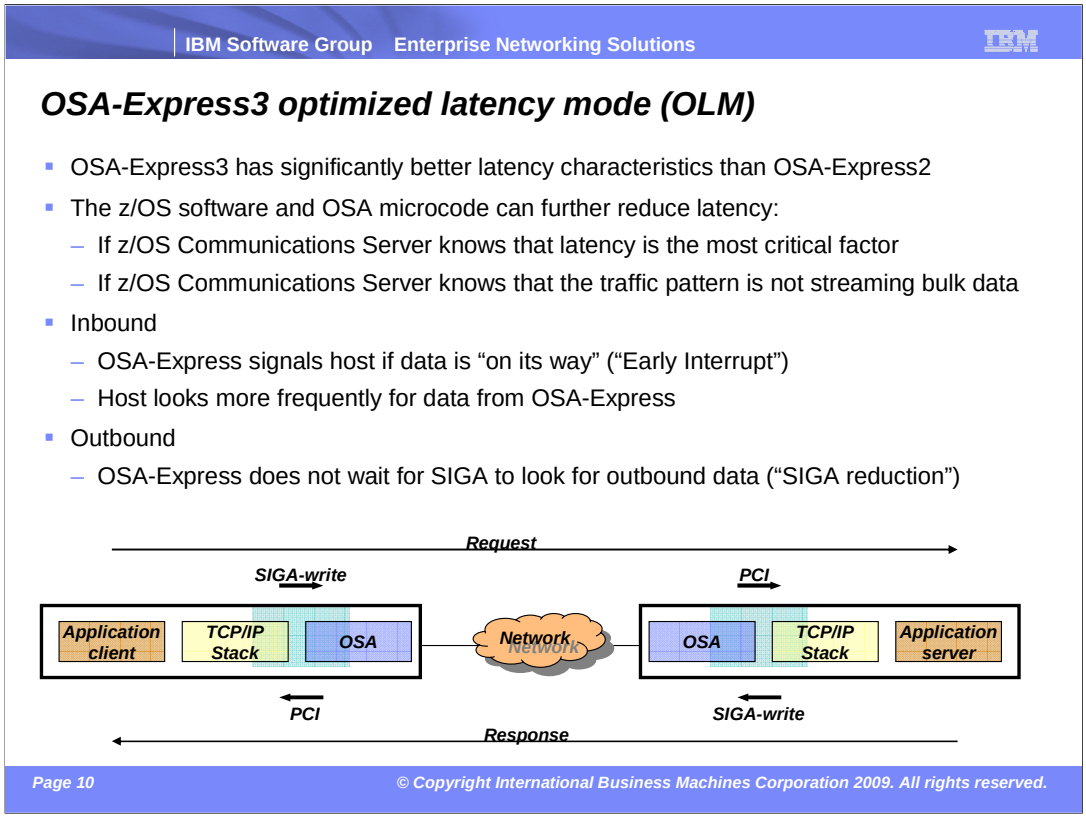
When isolation is in effect, dynamic routing will not learn a route over the OSA between stacks which share the OSA port.

It also does not preclude traffic between the stacks over the OSA adapter using a static route with the next hop address of a router on the LAN. However, this can result in excessive ICMP redirect packets from the router to the originating host. If you use such a static route technique, you should turn off receipt of ICMP redirects on the sharing hosts. If possible, you should also configure the router to not send ICMP redirects.

You might not need the ISOLATE function. You might prefer the reduction in latency and LAN traffic that OSA internal routing provides. Or you can use VLAN to achieve some measure of isolation. The ISOLATE function is intended for customers who want to isolate the stacks and prevent any traffic from flowing between the stacks over the OSA.

ISOLATE is limited to OSA-Express2 or OSA-Express3 Ethernet features in QDIO mode (CHPID type OSD) and running at least an IBM System z9 Enterprise Class (EC) or z9 Business Class (BC). See the 2094DEVICE, 2096DEVICE, 2097DEVICE, or 2098DEVICE Preventive Service Planning (PSP) bucket for more information.

If you want traffic to flow between two stacks that share an OSA-Express port but ensure that the traffic flows over an external LAN, you should configure each stack on a separate virtual LAN. The ISOLATE function can protect against a configuration error which might accidentally allow some traffic to bypass the LAN.

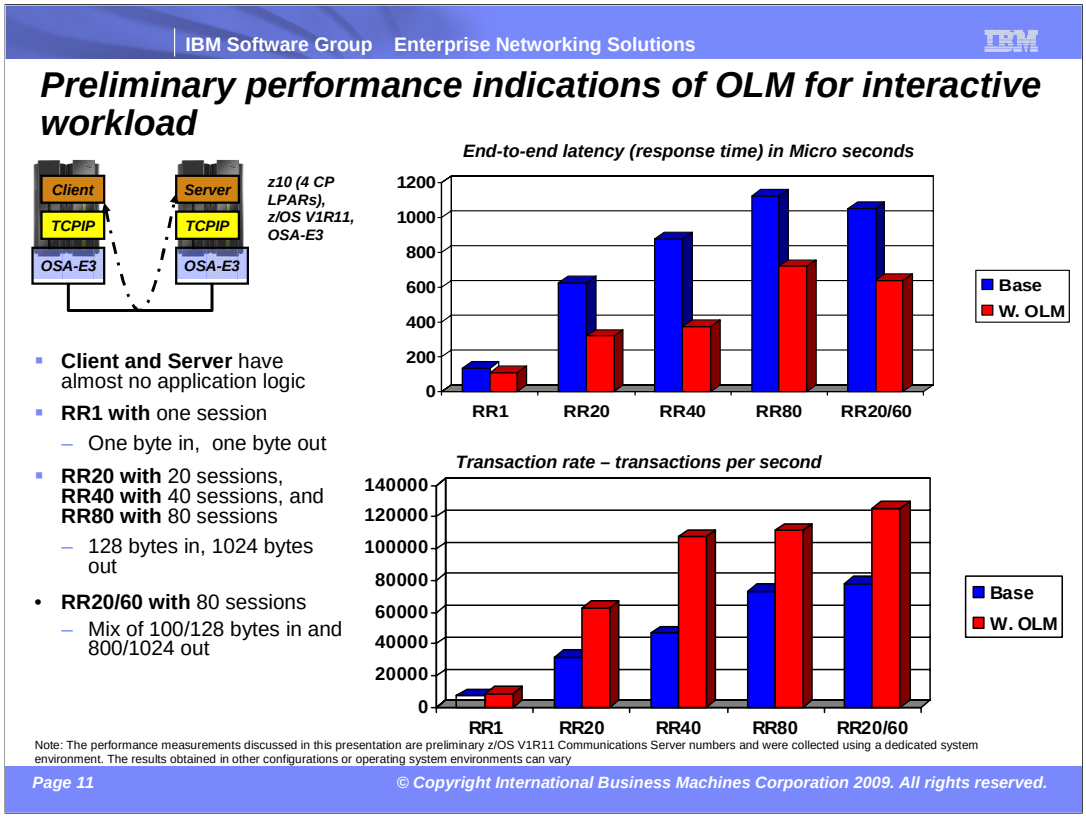


OSA-Express3 and z/OS V1R11 Communications Server have a new mode of operation for an OSA-Express3 in QDIO mode: optimized latency mode (OLM). OLM has several processing improvements.

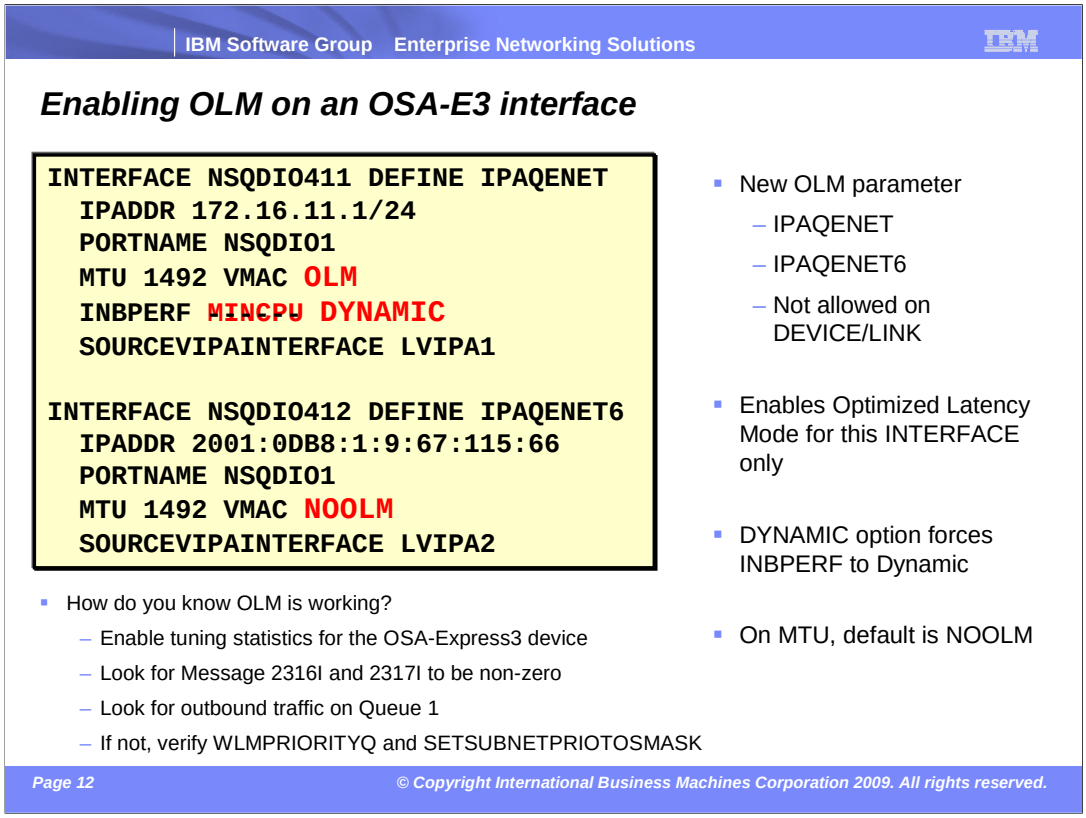
For inbound processing, OSA signals the host when data is "on its way". On inbound processing, z/OS V1R11 Communications Server looks more frequently for available data to process, ensuring any new data is read from the OSA-Express3 without requiring another PCI.

For outbound processing, OSA-Express3 also looks more frequently for available data to process, thus not requiring a Signal Adapter (SIGA) instruction to always know more data is available. OSA-Express3, as with previous generations of OSA-Express, supports four write priority queues. This additional scan for available data is only done for write priority queue 1.

You define an OSA-Express3 for QDIO mode using the INTERFACE statement to allow the OSA-Express3 to operate in optimized latency mode. You also install the correct OSA-Express Licensed Internal Code (LIC) level. See the appropriate PSP bucket for your level of z10 for the level that supports optimized latency mode.



You should only configure an OSA-Express3 to operate in OLM mode when the traffic over this OSA-Express3 demands the best latency possible. This is typically a high volume, interactive workload. In this type of workload, OLM can significantly improve both latency time per transaction and throughput. Preliminary performance runs, though not officially verified, show total end to end time for one transaction reduced by 17%. More significantly, in a typical customer environment with 20 simultaneous interactive workloads being processed at any given time, average latency was almost cut in half, and throughput improved 95%.



In z/OS V1R11 Communications Server, you configure the OSA-Express to operate in OLM using the new OLM parameter on the TCP/IP INTERFACE statement. No VTAM configuration is required. The default parameter is NOOLM which means do not use OLM.

This new parameter is supported for both IPv4 and IPv6, but only on the INTERFACE statement, only for QDIO devices (type IPAQENET and IPAQENET6), and is not supported on a DEVICE/LINK pair. So if you want to implement OLM, you define your OSA-Express3 with an INTERFACE statement.

Hopefully, if you follow the publications and this presentation, you will have no problems with using OLM on an OSA-Express3. If you do encounter problems, perform these steps.

Display the OSA-Express3 using D TCPIP,NETSTAT,DEVLINKS. Ensure OptLatencyMode displays as Yes. If it does not, make sure OLM is configured on the INTERFACE statement. If OLM is configured, look at the console for messages EZD0045I or EZD0046I. If you see EZD0045I, issue a D TRL,TRLE= for this OSA-Express3 and verify the LIC level. Look at the 2097DEVICE or 2098DEVICE PSP to see which level is required for OLM, and make sure the displayed LIC level is at least an OLM level. If you see EZD0046I, ensure there are no more than 4 concurrent interfaces to this OSA-Express3. See the “Restrictions for use of OLM” slide for more information on this restriction.

If OptLatencyMode is Yes on the NETSTAT display, enable tuning statistics for this OSA-Express3, using the F vtamproc,TNSTAT,TRLE= command. With traffic flowing over this OSA-Express3, look for non-zero counts for message IST2316I and IST2317I. If you see those, OLM is operating. If not, verify you have either enabled WLMPRIORITYQ or have used SETSUBNETPRIOTOSMASK to ensure traffic is directed to queues 1, 2, or 3.

Restrictions for use of OLM

- Concurrent interfaces to an OSA-Express port using OLM is limited to four
  - If one or more interfaces operate OLM, only four total interfaces allowed
  - All four interfaces can operate in OLM
  - An interface can be:
    - Another LPAR using the OSA-Express port
    - Another VLAN defined for this OSA-Express port
    - Another protocol (IPv4 or IPv6) interface defined for this OSA-Express port
    - Another stack on the same LPAR using the OSA-Express port
    - Any stack activating the OSA-Express Network Traffic Analyzer (OSAENTA)
- QDIO Accelerator or HiperSockets Accelerator will not accelerate traffic to or from an OSA-Express operating in OLM

© Copyright International Business Machines Corporation 2009. All rights reserved.

There are two significant restrictions for an OSA-Express3 operating in OLM. Both restrictions are to ensure OLM is effective.

One restriction is that there can only be four concurrent interfaces to the same OSA express port when one of the interfaces is operating in OLM. This restriction is necessary since multiple users cause a loss of effectiveness of OLM. For example, if the OSA has to look for data from many users without a SIGA and service many users, then the latency gain for one user is lost while another user is being serviced. To achieve the best latency possible, configure only one user for the OSA-Express3 operating in OLM.

The second restriction is that QDIO Accelerator or HiperSockets Accelerator will not accelerate traffic either to or from an OSA-Express operating in OLM. This is because while traffic is accelerated from an OLM OSA-Express to another device, latency time is being increased. This is contrary to optimizing latency, the intent of this mode. Either QDIOACCELERATOR or IQDIOROUTING can still be configured on the IPCONFIG statement; QDIOACCELERATOR still enables QDIO Accelerator, and IQDIOROUTING still enables HiperSockets Accelerator. However, acceleration will only be performed using OSA-Express devices not operating in OLM.

Trademarks, copyrights, and disclaimers

IBM, the IBM logo, ibm.com, and the following terms are trademarks or registered trademarks of International Business Machines Corporation in the United States, other countries, or both:

HiperSockets System z9 VTAM z/OS

If these and other IBM trademarked terms are marked on their first occurrence in this information with a trademark symbol (® or ™), these symbols indicate U.S. registered or common law trademarks owned by IBM at the time this information was published. Such trademarks may also be registered or common law trademarks in other countries. A current list of other IBM trademarks is available on the Web at "Copyright and trademark information" at http://www.ibm.com/legal/copytrade.shtml

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

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

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

IBM makes no representations or warranties, express or implied, regarding non-IBM products and services.

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

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

Performance is based on measurements and projections using standard IBM benchmarks in a controlled environment. All customer examples described are presented as illustrations of how those customers have used IBM products and the results they may have achieved. The actual throughput or performance that any user will experience will vary depending upon considerations such as the amount of multiprogramming in the user's job stream, the I/O configuration, the storage configuration, and the workload processed. Therefore, no assurance can be given that an individual user will achieve throughput or performance improvements equivalent to the ratios stated here.

© Copyright International Business Machines Corporation 2009. All rights reserved.

Note to U.S. Government Users - Documentation related to restricted rights-Use, duplication or disclosure is subject to restrictions set forth in GSA ADP Schedule Contract and IBM Corp.

© Copyright International Business Machines Corporation 2009. All rights reserved.