IBM z/OS Communications Server and OSA-Express Best Practices

Version 1.2

Jerry Stevens (sjerry@us.ibm.com)

August 08, 2023

Table of Contents

IBM z/OS Communications Server OSA-Express QDIO Performance Considerations and Best Practices

Document Purpose

This document is provided for the purpose of aiding IBM z/OS customers by providing a general set of considerations (a checklist) for guidance focused on configuring OSA-Express for optimizing network performance. This document does not replace the formal IBM publications such as the IBM z/OS Communications Server IP Configuration Guide or Configuration Reference. The formal publications should be used for details about additional guidance and "how to" information.

This document is intended to be a "living document" that could be updated and replaced as considerations change (i.e. new hardware, enhancements, features, performance-oriented field maintenance etc.).

Document History

V1.0 represents the initial version of this document originally published in March of 2022 intended for use by IBM z/OS Communications Server customers.

V1.2 was published in August of 2023 to provide updated performance recommendations for OSA sharing and the introduction of OSA-Express Enhanced Inbound Blocking (EIB).

Background

1. The information contained in this document is focused on z/OS Communications Server best practices for configuring and using OSA-Express in QDIO mode for optimizing your network performance. All users must perform an analysis and assessment of their specific and unique environment prior to adopting any suggested changes.
2. The main focus of this topic is around optimizing the OSA configuration for optimal performance, scale, and throughput for your z/OS application workloads and for the avoidance of packet loss.
3. Inbound Packet Loss that occurs at z/OS or OSA can occur for various reasons but often occurs during periods of peak traffic rates or sudden bursts of TCP/IP or UDP/IP traffic. TCP is known to be a somewhat "bursty" protocol that mitigates or manages packet loss with both TCP and IP congestion management protocols (e.g. DRS) and TCP retransmission. In most cases the TCP/IP

retransmission protocols will handle / recover from packet loss. However, persistent higher levels of packet loss can impact performance and TCP connection recovery. Packet loss can occur in the network and at the network endpoints. Lossy or congested networks / bottlenecks (routers and switches) can also cause similar performance issues.

Cases in which valid packets (i.e. packets that are not malformed or contain any other protocol or CHKSUM errors preventing delivery) are dropped or discarded by OSA (hardware or firmware) are often caused by z/OS's inability to keep up with the arrival rate of inbound network traffic, which can occur in sudden bursts. Here, "keep up with" means the consumption of or the processing and replacing of OSA read storage (QDIO SBALs) at a pace that exceeds the network traffic arrival rate which can come at very large bursts at line speed (10 GbE or 25 GbE).

When this occurs, it is usually related to resource constraints in z/OS. The most common types of resource constraints are:

a. Lack of CPU resources (e.g. high CPU utilization, CPU virtualization or sharing contention (CP priority, High, Med, or Low), dispatching delays caused by CPU constraints, internal software lock contention all can cause device driver or stack processing delays etc.).
b. Lack of system storage, such as CSM, HV Common or available system fixed (real) memory.
   Note: If the z/OS processing falls behind the network packet arrival rate causing OSA to deplete input (SBAL) read storage or memory, packets can be dropped. VTAM Tuning statistics can be used to monitor for this condition. TNSTAT counter "No Reads" tracks this condition. This counter should be zero (or very low). When this counter is frequently nonzero it indicates there is a z/OS resource issue.
c. OSA virtualization (sharing or provisioning). A physical OSA can be shared among many Logical partitions (LPARs) or guests. The entire OSA usage must be taken into consideration. Over sharing of a physical OSA can contribute to or cause packet loss or performance degradation.

4. Resource constraints are often avoidable through the tuning of the system configuration by following a set of best practices which includes the provisioning of resources for both the peak and steady state workloads.

This paper provides a summary of "best practices" for z/OS Communications Server and the configuration of OSA-Express in QDIO mode (IPAQENET INTERFACE) for optimal performance.

Note:
This paper is organized in a “primary list” of considerations followed by a more “general list” of considerations. Some items are sequence sensitive or interdependent. For this reason, it is necessary to review this entire document prior to making any changes.

OSA Configuration (primary) Considerations:

OSA QDIO Read Storage

1. **Set OSA QDIO (IPAQENET INTERFACE) Read Storage Parameter to Max (126)**
   The ReadStorage sub-parameter on the IPAQENET INTERFACE statement defines how much storage (number of read buffers or SBALs) is used for each OSA input queue. Input queues are used by OSA to copy inbound packets to host memory. Inbound Workload Queuing (IWQ) is described in the next topic. Enabling IWQ impacts storage usage. Without IWQ a single input queue is used per OSA INTERFACE. With IWQ up to 6 input queues can be used. With IWQ the Read Storage value is applied to all applicable input queues for the INTERFACE.

   VTAM start option QDIOSTG (QDIO Storage) provides a z/OS global configuration value allowing all INTERFACEs to use a single system wide (global) setting.

   The default value of QDIOSTG = MAX (4 MB for 1 GbE or 10 GbE and 8 MB for 25 GbE). QDIOSTG also supports defining a specific number of read buffers, up to 126 (8 MB).

   The stack (TCPIP Profile) INTERFACE default value for ReadStorage = Global (use the VTAM QDIOSTG setting).

   Users should consider defining QDIOSTG = 126 (8 MB) and ReadStorage = Global.

   Note: In cases in which there are a large no. of INTERFACE statements (e.g. > 4) per stack, users could consider reducing the storage requirements by setting ReadStorage to a lower value (e.g. ReadStorage = MAX (4 MB)) on selected INTERFACE statements (i.e. Interfaces with less network traffic).

System Storage

2. **System Storage: review other related system storage definitions:**
   a. CSM FIXED MAX in IVTPRM00 (consider at least the default of 512 MB)
   b. CSM HVCOMM in IVTPRM00 (consider using default of 2 G)
   c. The actual real storage available to this z/OS system (see D M=STOR or D M=HIGH).

Inbound Workload Queuing (IWQ)

3. **Use z/OS Communications Server Inbound Workload Queuing (IWQ)**
   OSA-Express Inbound Workload Queuing (IWQ) uses a form of "packet steering" to distribute inbound packets (network traffic) to unique QDIO input queues. Each unique input queue is optimized and provisioned with resources based on the type of workload.

   The separation of network traffic allows z/OS to optimize the parallel processing creating higher levels of z/OS performance and scalability. IWQ also avoids resource constraints that can lead to packet loss. The primary input queue represents inbound IP network traffic that is transactional (interactive vs. bulk or streaming) and destined for applications within this z/OS instance. The other 5 input queues (Ancillary Input Queues) are designed to minimize the impact of processing related to the primary input queue by handling other unique types of network traffic as follows: 1. Bulk (streaming traffic) 2. Sysplex Distributor 3. Enterprise Extender 4. IPSec encrypted traffic and 5. zCX. The primary and the bulk input queues are always populated with read storage. The other 4 ancillary input queues are only populated with read storage when the related function is actually used. The bulk queue is also designed to minimize out of order packets that can occur within z/OS due to the multiple thread (SRBs) processing for large amounts of data for the same TCP connection. Out of order packets can lead to packet loss, retransmission, and poor performance. This optimization applies to both inbound and outbound streaming workloads for bulk data.

   Prior to enabling IWQ, the previous two suggested steps for defining storage should be reviewed and completed.

   IWQ is defined in the TCP/IP Profile on the OSA IPAQENET INTERFACE Statement (OSD CHPID) on the INBOUND Performance sub-parameter **(INBPERF DYNAMIC WORKLOADQ).**

   Refer to the IP Config Guide (QDIO Inbound Workload Queuing in Chapter 2) and the IP Config Reference (IPAQENET INTERFACE Statement) for additional information.

   Note: The IWQ bulk queue registration function has been improved with APAR PH43504. APAR PH43504 is recommended.

QDIO Accelerator

4. **Use QDIO Accelerator when a z/OS instance is defined as a Sysplex Distributor host.**
   QDIO Accelerator can reduce both CPU and storage costs related to the processing of Sysplex distributed (forwarded) network traffic. The savings are maximized when QDIO accelerator is used in combination with IWQ. QDIO Accelerator is also beneficial when this z/OS instance is using **DATAGRAMFWD**.

   Accelerator is defined on the IPConfiguration statement in the TCP/IP profile. Refer to the IP Config Guide and IP Config Reference for additional information.

   Note: When using QDIO Accelerator, it is suggested that users review and enable both:
   a. **VIPARoute** (VIPADYNAMIC) and
   b. **AdjustDVIPAMSS** (Global Config statement).
      Refer to the IP Config Guide and IP Config Reference for both topics.

General Considerations

Miscellaneous OSA and TCP/IP Performance Considerations

1. Miscellaneous OSA and TCP/IP stack performance considerations:

Consider the following general set of OSA related best practices:

1. Enable segmentation offload with **GlobalConfig SEGMENTATIONOFFLOAD**
2. Utilize outbound priority queuing with **GlobalConfig WLMPRIORITYQ**
3. When possible, **utilize jumbo frames with Path MTU Discovery (and ADJUSTDVIPAMSS for VIPAROUTE,** see item 4 above).
   Note that jumbo frames can be specified in multiple places. Refer to topic "Determining Maximum Transmission Unit" in the IP Config Guide
4. Use **INBPERF setting of DYNAMIC** (see item 3 above)
5. Use **TCPCONFIG AUTODELAYACKS**
6. The default **TCPCONFIG TCPRCVBufrsize is 64 K,** this allows DRS to enable. We suggest that users don't use a lower value.
7. This item is an application programming consideration.
   Include the **MSG_WAITALL flag** on socket recv calls for applications receiving large/bulk data (socket API solution - refer to IP Sockets Application Programming Interface Guide and Reference topic RECVFROM).

OSA QDIO INTERFACE Configuration

2. OSA INTERFACE configuration optimization suggestions:

Consider the following best practices related to OSA INTERFACE configuration:

1. Convert to **INTERFACE statement** from **DEVICE/LINK (must)**
2. On INTERFACE use**:**
   i. **OSA VMAC (OSA-generated)**
   ii. Configure an **IP subnet mask** on the OSA INTERFACE statement
   iii. When using **VLAN ID** configure your switch port in **Trunk Mode**
3. **High Availability:** Configure and deploy multiple physical OSAs **(multiple OSA Interfaces)** for **High Availability** by using dynamic routing and multipath (per connection).
4. **Adapter Interrupt Monitor (AIMON)** is VTAM start option that controls a function that monitors for and recovers from lost (missing) interrupts. AIMON defaults to off. AIMON supports OSA, HiperSockets, RoCE and ISM

devices. Although lost interrupts are **very rare** if an interrupt is lost for a critical network interface it can be disruptive and can be difficult to diagnose. **Consideration should be given to enabling AIMON (for all devices or for your critical OSA (QDIO) interfaces**

OSA-Express Sharing Performance Considerations

3. OSA-Express Sharing Performance Considerations:

When multiple LPARs share the same physical OSA-Express port there are some performance aspects to consider:

1. When most or all the network traffic among the LPARs sharing the OSA-Express is external to the LAN, then this configuration should perform as expected.
   This assumes the bandwidth requirements of all LPARs sharing the OSA does not exceed the bandwidth capability (e.g 10 or 25 GbE) of the shared OSA.

2. When a significant amount of network traffic (e.g. streaming workloads) is internal, between any two LPARs sharing an OSA, then this type of workload might experience performance issues. With heavy network traffic, packet loss can occur within the OSA internal LP to LP forwarding processing (due to OSA internal storage constraints). When this occurs, streaming workloads can experience a reduction in throughput due to retransmissions (packet loss).

   For this case, if OSA sharing is required, then it is recommended that the streaming traffic is forced external (out and back) to the LAN by configuring the two INTERFACE statements on to separate VLANs. This also means unique IP subnets, which requires an IP route that forces traffic over an external router and back into the target LPAR.

   Note: Users should NOT attempt to use the QDIO ISOLATE function to resolve this performance issue. ISOLATE is designed to **prohibit any traffic** among sharing LPARs. When ISOLATE is configured by either LPAR sharing an OSA, OSA will **not allow any communications** between the LPARs causing packets to be dropped if internal traffic is attempted.

   The following information is copied from the IP Configuration Reference under IPAQENET INTERFACE ISOLATE:

Prevent OSA-Express from routing packets directly to another TCP/IP stack that is sharing the OSA adapter. In this mode, OSA-Express adapter discards any packets when the next hop address was registered by another stack that is sharing the OSA adapter. Packets can flow between two stacks that share the OSA only by first going through a router on the LAN. For more details, see the OSA-Express connection isolation information in z/OS Communications Server: IP Configuration Guide.

OSA-Express7S Enhanced Inbound Blocking (EIB)

4. **OSA-Express7S Enhanced Inbound Blocking (EIB) Function**
   A new performance function called ***Enhanced Inbound Blocking*** or EIB is available for OSA-Express7S on the IBM z16. The z/OS 2.4 and 2.5 CommServer support for the EIB function was made available in TCP/IP APAR PH44281 and VTAM APAR OA62831.
   Using EIB is recommended for z/OS instances that have heavy or frequent inbound streaming workloads, which applies to most z/OS use cases.

   QDIOEIB = ENABLED | DISABLED

   OSA-Express in QDIO mode provides inbound blocking that optimizes the transfer of packets to host memory by controlling the blocking rate of packets per QDIO SBAL. The EIB function provides enhancements to the current OSA QDIO inbound blocking that dynamically adjusts the OSA blocking rate based on the current availability of host read storage (SBALs). If the z/OS read storage were to become low the OSA blocking factor will dynamically increase attempting to avoid any packet loss.

   Requirements:
   1. The interface must use 126 SBALs (8 M)
   2. Interface definition must specify INBPERF DYNAMIC

   Note: The current setting of EIB can be displayed using the z/OS Display OSAINFO command. Also note that in addition to the EIB function, an OSA hardware packet drop counter (formerly provided at the HMC/SE) is now provided in the z/OS Display OSAINFO information. This counter is a single hardware counter for the entire OSA physical port applicable to all LPARs using this OSA. Any packet loss occurring noted by this counter indicates the packets were never seen or processed by software.

   See the APARs listed for additional EIB usage information.

Trademarks

IBM, the IBM logo, and ibm.com are trademarks or registered trademarks of International Business Machines Corp., registered in many jurisdictions worldwide. Other product and service names might be trademarks of IBM or other companies. A current list of IBM trademarks is available on the Web at Copyright and Trademark information (http://www.ibm.com/legal/copytrade.shtml).