

This presentation describes the new functions related to sysplex distributor performance in z/OS V1R12 Communications Server.



Inbound traffic separation is now being supported for OSA-Express3 using multiple read queues. TCP/IP will register with OSA which traffic is to be received on each read queue and OSA-Express Data Router function routes traffic to the correct queue. With sysplex distributor traffic separated onto its own read queue, it can be efficiently accelerated or presented to the target application. Also with bulk data traffic separated onto its own read queue, TCP/IP will service the bulk data queue from a single processor and this will assist in reducing the problems with out of order packet delivery.

Support is added for a sysplex distributor to load balance IPv6 connections to DataPower appliances. By deploying the sysplex distributor as the load balancing component for both tiers of distribution, z/OS administrators can use a single load balancing solution for the z/OS workload to the DataPower and z/OS application processing tiers. This simplifies the load balancing administration.

A new sysplex distribution method, HotStandby, is now supported. One target can be designated as a preferred target receiving all new incoming connection requests, and the hot-standby servers act as backup servers should the designated preferred target become unavailable.



Outbound traffic separation (assignment to specific priority queue) on the multiple write queues can be accomplished by using Policy Agent and configuring a policy with the SetSubnetPrioTosMask statement. Beginning in z/OS V1R11 Communications Server, outbound traffic separation can also be accomplished by using the WLMPRIORITYQ parameter on the GLOBALCONFIG statement. Each priority queue is processed independently of the others. For example, one processor can be building writes on priority queue four.

All inbound traffic is received on the single read queue. This includes both batch and interactive traffic and both traffic destined for this TCP/IP stack and traffic to be forwarded by this TCP/IP stack. Multiple processes only run for inbound traffic when data is accumulating on the read queue – typically during burst periods when z/OS Communications Server is not keeping up with the OSA. A single process is used to package the data, queue it, and schedule the TCP/IP stack to process it. This same process also performs acceleration functions, such as sysplex distributor connection routing accelerator. The TCP/IP stack separates the traffic types (sysplex distributor traffic, bulk inbound data such as FTP, and son on) to be forwarded to the appropriate stack component that will process them.

For these reasons, z/OS Communications Server is becoming the bottleneck as OSA-Express3 10GbE nears line speed. z/OS Communications Server is injecting latency and increasing processor utilization. This can impede scalability.



All inbound QDIO traffic is received on a single read queue regardless of the data type.

The maximum amount of storage available for inbound traffic is limited to the read buffer size (64K read SBALs) times the maximum number of read buffers (126).

Under the current z/OS Communications Server model, another QDIO input process will eventually be driven, and another TCP/IP stack thread, thus allowing multiple threads to process the one inbound read queue. However, this is only done when the OSA detects the host is now "falling behind" using the QDIO interrupt threshold algorithm.

Multiple processes are used for inbound traffic when data is accumulating on the read queue. This can cause bulk data packets for a single TCP connection to arrive at the TCP layer out of order. Each time the TCP layer on the receiving side sees out of order data, it transmits a duplicate ACK.



With z/OS Communications Server V1R12, inbound traffic separation is supported using multiple read queues. TCP/IP will register with OSA which traffic to be received on each read queue. The OSA-Express Data Router function routes traffic to the correct queue.

Each read queue can be serviced by a separate process. The primary input queue is used for general traffic. One or more ancillary input queues (AIQs) are used for specific traffic types.



The supported traffic types are streaming bulk data and sysplex distributor. Examples of bulk data traffic are FTP, TSM, NFS, and TDMF. Both IP versions are supported for all types of traffic.

With bulk data traffic separated onto its own read queue, TCP/IP will service the bulk data queue from a single processor. Using one processor solves the out of order delivery issue – there are no more race conditions. The determination of bulk data traffic is based in detecting streaming traffic. Bulk data traffic is registered per connection (5-tuple).

With sysplex distributor traffic separated onto its own read queue, it can be efficiently accelerated or presented to the target application. The determination of sysplex distributor traffic is based on active VIPADISTRIBUTE definitions. SD traffic is registered per DVIPA address.

All other traffic is processed simultaneous with the bulk data and sysplex distributor traffic

The dynamic LAN idle timer is updated independently for each read queue. This ensures the most efficient processing of inbound traffic based on the traffic type.



The QDIO inbound workload queuing function is enabled with the INBPERF DYNAMIC WORKLOADQ setting on IPAQENET and IPAQENET6 INTERFACE statements. WORKLOADQ is not supported for INBPERF DYNAMIC on IPAQENET LINK statements. For steps to convert from IPv4 IPAQENET DEVICE, LINK, and HOME definitions to the IPv4 IPAQENET INTERFACE statement, see z/OS Communications Server: IP Configuration Guide.

The VMAC parameter can be specified with or without the MAC address.

For more information, see the IPAQENET INTERFACE and IPAQENET6 INTERFACE statements in z/OS Communication Server: IP Configuration Reference.

The Netstat DEvlinks/-d report has been updated for IPAQENET and IPAQENET6 interfaces to include a WorkloadQueueing field when InbPerf is Dynamic to indicate whether the QDIO inbound workload queuing function is enabled. The possible values are Yes indicating the function is enabled, No indicating the function is not enabled, and Unsupported indicating that the OSA-Express does not support the function. This field is not displayed when InbPerf is not Dynamic or for IPAQENET DEVICE/LINK definitions. This information can also be retrieved with a network management application that has been updated to use information returned by the GetIfs callable NMI.

In addition, the InbPerf field has been moved from the right column to the left column.

For more information, see the Netstat DEvlinks/-d report in z/OS Communications Server: IP System Administrator's Commands.



The output for the Display ID=trlename and Display TRL,TRLE=trlename commands has been updated to indicate whether the QDIO inbound workload queuing function is in use for the QDIO interface. For each input queue, it includes the queue ID and queue type in addition to the read storage. The queue type is PRIMARY for the primary input queue, BULKDATA for the bulk data AIQ, and SYSDIST for the sysplex distributor connection routing AIQ. The queue type value N/A indicates that the queue is initialized but is not currently in use by the TCP/IP stack. The queue ID and queue type can be used to correlate with VTAM tuning statistics, packet trace, and OSA-Express Network Traffic Analyzer (OSAENTA) trace output for the QDIO interface. For more information, see the DISPLAY ID command and DISPLAY TRL command in z/OS Communications Server: SNA Operation.



The Netstat ALL/-A report has been updated to include the interface name for bulk data TCP connections that are using the QDIO inbound workload queuing function. For all other sockets, the ancillary input queue will display as N/A and the BULKDATAINTFNAME statement is not displayed. For more information, see the Netstat ALL/-A report in z/OS Communications Server: IP System Administrator's Commands. This information can also be retrieved with a network management application that has been updated to use information returned by the GetConnectionDetail callable NMI.



The Netstat STATS/-S report has been updated to include the total number of segments received for all connections from the bulk data AIQ of the OSA-Express QDIO inbound workload queuing function. For more information, see the Netstat STATS/-S report in z/OS Communications Server: IP System Administrator's Commands. This information can also be retrieved with a network management application that has been updated to use information returned by the GetGlobalStats callable NMI.



IBM WebSphere DataPower appliances are often used as a front-end processing tier to z/OS, providing transparent web services enablement or support of z/OS applications by providing acceleration and more efficient handling of web services functions.

In many of these environments, multiple DataPower instances are typically used in a cluster to allow for higher availability and scalability. This requires an external Tier 1 load balancer to distribute incoming web services connection requests to the DataPower appliance cluster.

Once DataPower has performed its portion of the processing of the incoming request, it will route a request to a z/OS application tier, such as CICS, IMS, WebSphere, or DB2. In a sysplex environment, the z/OS application tier will likely require a 2nd tier of load balancing to be performed.

Tier 2 load balancing can use DataPower's built-in load balancing support, which by default uses a round robin distribution. Another option, shown in the diagram, is to use the z/OS sysplex distributor which receives real time z/OS Workload Manager recommendations regarding current capacity and performance of each application and LPAR instance.

Sysplex distributor can be used as a load balancer for both tiers allowing one load balancing solution; this simplifies network management.

Support added in sysplex distributor and DataPower allows DataPower to communicate with sysplex distributor over an IPv4 TCP control connection which provides dynamic load balancing feedback:

Processor capacity information provided by the DataPower appliance allows sysplex distributor to optimize its load balancing decisions to avoid overloaded DataPower appliances by directing new requests to DataPower instances that are less used.

Connection state information provides server availability. It is used during a takeover. All connections are provided during takeover allowing Non-disruptive takeover by a backup sysplex distributor (planned or unplanned).

The return path from the DataPower appliance to a client can bypass sysplex distributor. Since Generic Resource Encapsulation (GRE) is used to forward inbound distributed packets to DataPower, the DataPower appliance can use the inner IP header to send outbound packets directly to client. Since DataPower sends connection termination information on a control connection, sysplex distributor does not need to monitor connection traffic to determine when a connection has ended.



In V1R12, support is added for a sysplex distributor to load balance IPv6 connections to DataPower appliances. By deploying the sysplex distributor as the load balancing component for both tiers of distribution, z/OS administrators can use a single load balancing solution for the z/OS workload to the DataPower and z/OS application processing tiers. This simplifies the load balancing administration.

Similar to IPv4 support, support added in sysplex distributor and DataPower allows DataPower to communicate with sysplex distributor over an IPv6 TCP control connection which provides dynamic load balancing feedback:

CPU capacity and Connection state information are received from the DataPower appliance on the TCP control connection.

Different from IPv4 support, IPv6 Encapsulation is used to forward inbound distributed packets to DataPower instead of GRE.



In V1R12, IPv6 support is added. The ENCAP parameter replaces the GRE parameter. ENCAP means that the Tier 1 targets are IPv6 non-z/OS targets, and IPv6 Encapsulation is used to send packets on a connection. IPv6 DVIPAs are supported to DataPower targets.

CONTROLPORT means that the sysplex distributor will bring up an IPv6 TCP control connection to the DataPower sysplex client listening on that port.

TARGCONTROLLED means the distribution is controlled by the target using the weights that it sends. The distributor uses a weighted round-robin distribution to the DataPower targets using the normalized DataPower weights received on the control connection. Weighted Active and Round Robin distribution can also be used.



IPv6 Multi-Site Support is added in V1R12. It works the same as IPv4 support.

When DataPower is used in a multi-site environment, the load balancing configuration can be further enhanced by using multi-tier support of sysplex distributor in combination with load balancing to DataPower.

With this type of multi-CPC (multi-site) environment, each Tier 2 sysplex distributor is configured to load balance and route service requests to z/OS server application instances on its CPC. The Tier 2 distributors send a Site WLM weight of the group of target application instances in their CPC site to the Tier 1 distributor. The Tier 1 distributor makes its load balancing decision using a normalized weight based on the combined weight of each DataPower target along with the Site weight of its Tier 2 distribution targets. For DataPower DATAP1 and DATAP2, the weights of WAS1 and WAS2 are included in the combined weight. For DataPower DATAP3 and DATAP4, the weights of WAS3 and WAS4 are included in the combined weight.

For example, the combined Tier1/Tier2 weight of DATAP1 = (WAS1 weight 2 + WAS2 weight 10) * DP Weight 500 = 6000.

The combined weights of all the DataPower appliances are then normalized against each other.

PerformanceSD.ppt



For IPv6 Multi-Site support, each Tier 2 sysplex distributor is configured to load balance and route service requests to a set of IPv6 z/OS server application instances on its CPC. Since the Tier 2 DVIPA, DVIPA2, is defined with CPCSCOPE, the Tier 2 distributor will only load balance to targets on its CPC, CPC1 (even though DESTIP ALL is configured).

A CPCSCOPE DVIPA definition is used to define a CPC subnet. A CPCSCOPE DVIPA definition on one of the stacks in Site 1 (CPC1) is used to define the CPC1 subnet of 2001:0DB8::81:10:0/112. Another CPCSCOPE DVIPA definition on one of the stacks in Site 2 (CPC2) is used to define the CPC2 subnet of 2001:0DB8::81:20:0/112).

The Tier 1 distributor determines the CPC1 and CPC2 subnet as it receives the CPCSCOPE DVIPA definitions from the stacks that are on those CPCs. It determines that DATAP1 and DATAP2 have IP addresses that match the CPC1 subnet, 2001:0DB8::81:10:0/112. It determines that DATAP3 and DATAP4 have IP addresses that match the CPC2 subnet. It creates a combined Tier1/Tier2 weight for a DataPower appliance using the matching group weight that it receives from a distributor on that CPC with the appliance's weight.

PerformanceSD.ppt

	IBM
VIPADEFINE and VIPABACKUP statement syntax changes	
 Addition of an IPv6 prefix length to each IPv6 address parameter 	
6	2010 IBM Corporation

The VIPADEFINE and VIPABACKUP statements have a new parameter – the IPv6 prefix length on each IPv6 address parameter.

This length is used with CPCSCOPE to define a site's DataPower subnet.



The Netstat VIPADCFG/-F display will show the IPv6 prefix length when displaying information about the VIPADEFINE or VIPABACKUP statement.

The Netstat VIPADCFG/-F DETAIL display will show the IPv6 routing type of ENCAP when displaying information about a VIPADISTRIBUTE statement for IPv6 DataPower targets.



Customers have some applications that do not support database sharing or run more efficiently without database sharing, but they still want to have backup "hot-standby" targets for availability. The distributor keeps all of the workload on one active target as long as it is available. It switches to one of the backup targets if the active target is no longer available.



A new distribution method, HotStandby, is supported. You have a single target server to receive all new connection requests. Other target servers are active but not receiving any new connection requests. The stack will automatically route traffic to a backup target server when the active target server is not available.

Using the new HOTSTANDBY distribution method, one target can be designated as a preferred target. If AUTOSWITCHBACK is configured, the distributor will switch to the preferred target when it is available unless the switch from the preferred target occurs because of health problems.

Designate one or more targets as backup targets. Use a backup target if the preferred target is not available. A target is not available if it is not ready, the route from distributor to target is inactive, or if HEALTHSWITCH is configured and the target is not healthy. A target is not healthy if the TSR is 0%, the abnormal termination rate is 1000 (out of 1000 transactions), or the server reported health is 0%.

	IBM
Feedback	
Your feedback is valuable	
You can help improve the quality of IBM Education Assistant content to better meet yo needs by providing feedback.	ur
Did you find this module useful?	
Did it help you solve a problem or answer a question?	
Do you have suggestions for improvements?	
Click to send email feedback:	
mailteries@us.ibm.com?cubicgt=Ecodback_about_BorformanceSD.ppt	
This module is also available in PDF format at:/PerformanceSD.pdf	
20 Performance enhancements related to sysplex distributor © 201	10 IBM Corporation

You can help improve the quality of IBM Education Assistant content by providing feedback.

