

User Experience with Communications Server OSA Express Inbound Workload Queuing

John McLaughlin
Bank of America

August 14 2013
Session Number13913



AGENDA

- Inbound Workload Queuing
 - Requirements and workload mix
 - IBM Announcement for Inbound Workload Queuing
- Bank of America Benchmark Process
- Benchmark Results
- IBM Follow-up
- Conclusions

High Level View of our Workload Distribution

- Day Shift
 - Online Transactions to multiple regions running on multiple SYSPLEX Configurations
 - Online has priority
 - High volume online LPARS do not run Day Shift File transfers
 - However OSAs are shared across CECs.
- Night Shift
 - Batch Processing
 - File Transfers
- Generally speaking the “Natural” flow of work does not present major conflicts between streaming and interactive workloads

Typical CEC OSA Configuration

CEC			
LPAR			
LPAR			
LPAR			
LPAR			
LPAR			
LPAR			
LPAR			
LPAR			
LPAR			
LPAR			
LPAR			
LPAR			
10 Gig OSA	10 Gig OSA	10 Gig OSA	10 Gig OSA

IBM Rationale for Inbound Workload Queuing*

Inbound Processing Overview Prior to V1R12

- QDIO uses multiple write queues for traffic separation
- QDIO uses only one read queue
 - Multiple CPs are used only when data is accumulating on the queue
 - Single Process for initial interrupt and read buffer packaging
 - TCPIP stack performs inbound data separation for SD traffic, bulk data, EE traffic, etc.
- z/OS Comm. Server is becoming the bottleneck as OSA nears 10GbE line Speed
 - Inject latency
 - Increase processor utilization
 - Impede scalability

*(From IBM Overview of V1R12 Communications Server Features)

IBM Announcement

- **Performance improvements for streaming bulk data** – Processing for OSA-Express in QDIO mode supports inbound workload queuing. Inbound workload queuing uses multiple input queues for each QDIO data device (sub channel device) to improve TCP/IP stack scalability and general network optimization.
- You implement the performance improvements for streaming bulk data by enabling inbound workload queueing to process streaming bulk data traffic concurrently with other types of inbound QDIO traffic.
- You enable these improvements for a QDIO interface, inbound traffic for connections that exhibit streaming bulk data behavior is processed on an ancillary input queue (AIQ)
- All other inbound traffic is processed on the primary input queue or on an ancillary input queue for sysplex distributor connection routing.

Single Read Queue

- Single Read Queue has negative impact for performance of bulk data
- Single read queue for all inbound QDIO traffic regardless of data type
- Single process for initial interrupt and read buffer processing
- TCP/IP stack performs inbound data separation
- Multiple processes run when data is accumulating on read queue

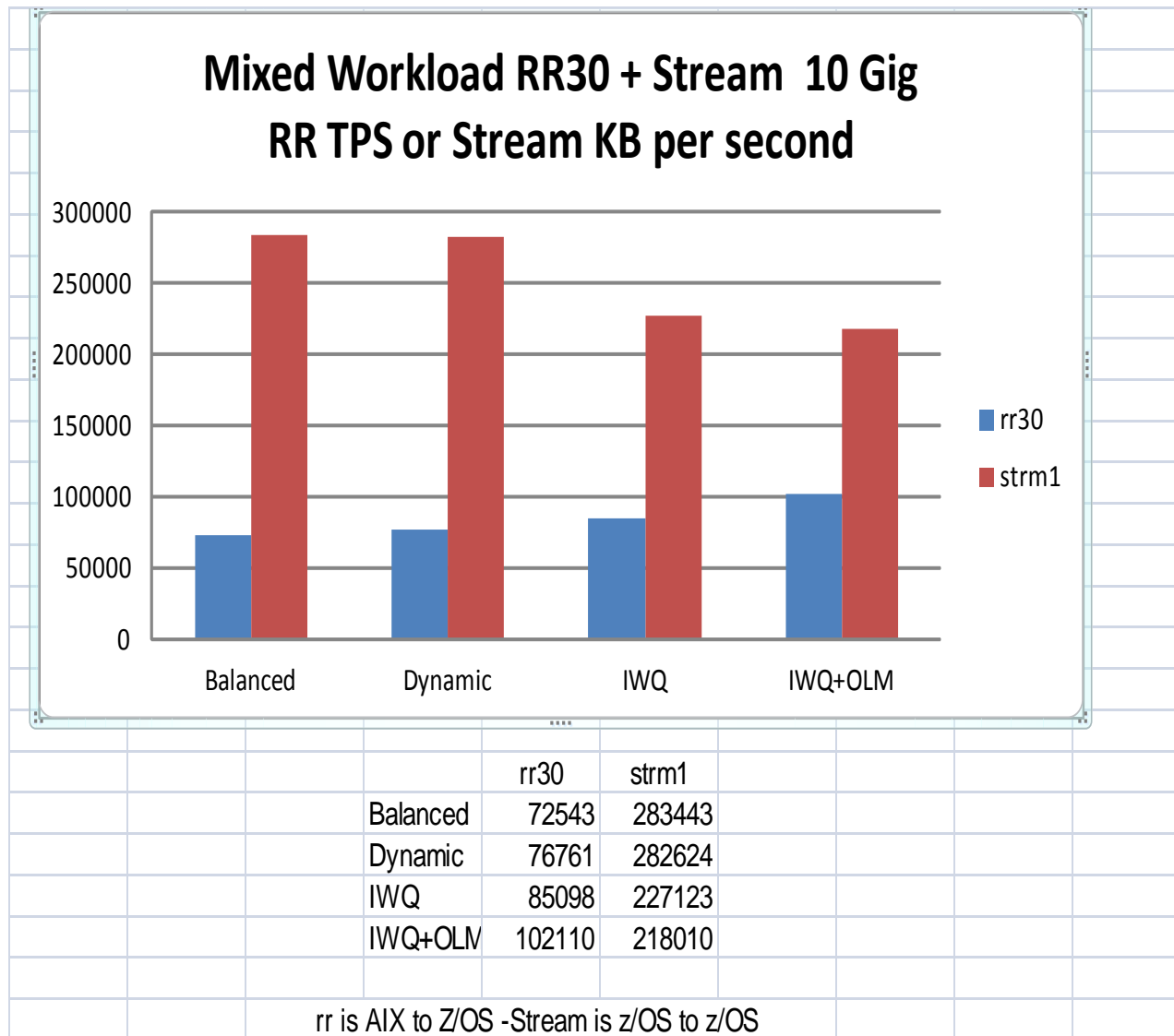
Multiple Inbound Queue Support

- OSA multiple inbound queue support allows inbound QDIO traffic separation by supporting multiple read queues
 - “Register” with OSA which traffic goes to which queue
 - OSA-Express Data Router function routes to the correct queue
Each input queue can be serviced by a separate process
 - Primary input queue for general traffic
 - One or more ancillary input queues (AIQs) for specific traffic types

Research

- **z/OS® V1R12**
- **Communications Server**
 - Performance Study:
 - *OSA-Express3*
 - *Inbound Workload Queueing*
 - Tom Moore -tdmoore@us.ibm.com
 - Patrick Brown - patbrown@us.ibm.com
- Summary of Mixed workload results page 22
 - “Up to ~45% interactive response time improvement with less than 3% increase in normalized network cpu “

Research Summary Mixed Workloads



Benchmark Configuration

3.8
GIG



CEC 5			
LPAR			
LPAR			
LPAR			
LPAR			
LPAR			
LPAR			
LPAR 1			
1 G OSA	1 G OSA	1 G OSA	1 G OSA

250
Concurrent
TN3270
Sessions

SW1

SW 2

CEC 6			
LPAR			
LPAR			
LPAR			
LPAR			
LPAR			
LPAR			
LPAR			
LPAR			
LPAR			
LPAR			
LPAR 9-2			
LPAR 9			
10 G OSA	10 G OSA	10 G OSA	10 G OSA

LPAR 9 TCPIP

**Primary
Input
Queue
(RD/1)**

**Bulkdata
Queue
(RD/2)**

**SysDist
Queue
(RD/3)**

**EE Queue
(RD/4)**

10 Gig OSA Express Type 3

Benchmark Description

- NDM a 3.8 Gigabit file from LPAR1 to LPAR9 evaluating inbound performance parameters while interactive traffic as well as test batch is processing.
 - Balanced a fixed setting trying to be the best of both worlds for interactive and bulk data (default)
 - Min CPU optimize CPU utilization
 - Min Latency Optimize throughput for interactive traffic
 - Dynamic
 - Dynamic with Inbound Workload Queuing

Measurements

- Elapsed Time from NDM
- VTAM CPU Time from RMF Monitor 2 Delta mode measuring at 10 second intervals
- TCPIP CPU Time from RMF Monitor 2 Delta mode measuring at 10 second intervals
- TN3270 Concurrent Users from SMF

TCPIP Summary Table

V1.13 from LPAR1 To LPAR9 3.8 GIG NDM Timings INBPERF	Elapsed Time Secs	TCPIP CPU Time PER 10 SECS	Kbytes/ Second	AVG TCPIP 10 Sec CPU	Avg Elapsed	CPU Delta compared to MIN CPU	Elapsed Delta Compared to Dynamic IWQ	Bytes in 10 seconds	Million Bytes in 10 secs	Million Bytes per TCPIP CPU Second	SYSNAME	LINK NAME	Hour	INBOUND BYTES	OUTBOUND BYTES
LPAR9 Min Latency	35	1.33	109,418								LPAR1	S0140000	15	18,247,524	7,199,167,639
LPAR9 Min Latency	37	1.13	103,504								LPAR1	S0240000	15	10,003,422	3,680,403,244
LPAR9 Min Latency	43	1.28	89,061	1.25	38.33	1.11	0.04	999,036,434	999.04	802.26	LPAR1	S0340000	15	13,546,122	3,596,753,643
LPAR9 Dynamic	37	1.06	103,504								LPAR1	S0440000	15	51,925,338	3,598,738,021
LPAR9 Dynamic	36	1.02	106,379								LPAR9	S1140000	15	1,012,981,728	545,103,133
LPAR9 Dynamic	36	1.20	106,379								LPAR9-2	S1140000	15	76,027,934	198,331,394
LPAR9 Dynamic	43	0.96	89,061	1.06	38.00	0.79	0.03	1,007,799,912	1007.80	951.88	LPAR9K-1	S1240000	15	21,342,886	37,121,219
LPAR9 Min CPU	39	0.68	98,196								LPAR9	S1240000	15	83,311,981	284,615,982
LPAR9 Min CPU	39	0.59	98,196								LPAR9-2	S1240000	15	85,229,724	145,399,468
LPAR9 Min CPU	39	0.51	98,196	0.59	39.00	0.00	0.05	981,958,888	981.96	1,657.31	LPAR9-4	S1240000	15	49,281,574	118,453,779
LPAR9 Default Balanced	42	0.69	91,182								LPAR9K-1	S1340000	15	12,016,610	20,932,567
LPAR9 Default Balanced	45	0.60	85,103								LPAR9	S1340000	15	18,179,085,510	472,147,106
LPAR9 Default Balanced	52	0.59	73,647	0.63	46.33	0.06	0.25	826,540,935	826.54	1,321.88	LPAR9-2	S1340000	15	109,530,980	153,237,641
LPAR9 Default Balanced Sep 28	37	0.60	103,504								LPAR9K-1	S1440000	15	35,550,324	37,527,502
LPAR9 Default Balanced Sep 28	39	0.47	98,196								LPAR9	S1440000	15	8,586,742,965	1,017,281,117
LPAR9 Default Balanced Sep 28	41	0.51	93,406								LPAR9-2	S1440000	15	77,430,118	268,026,242
LPAR9 Default Balanced Sep 28	42	0.64	91,182	0.56	39.75	0.00	0.07	963,431,362	963.43	1,735.91					
LPAR9 Dynamic IWQ	41	1.04	93,406												
LPAR9 Dynamic IWQ	34	1.55	112,636												
LPAR9 Dynamic IWQ	36	1.45	106,379	1.35	37	1.28	0.00	1,035,037,747	1035.04	768.59					
LPAR9 Dynamic IWQ April 3-4	40	1.43	95,741												
LPAR9 Dynamic IWQ	43	1.27	89,061												
LPAR9 Dynamic IWQ	40	1.29	95,741												
LPAR9 Dynamic IWQ	39	1.45	98,196	1.36	40.5	1.30		945,590,040							
3,829,639,684 (bytes)					3,829,639,664										

TN3270
Sessions Hour Bytes In Bytes Out
270 15 1595896 55191994

Comparison of TCPIP Performance Data

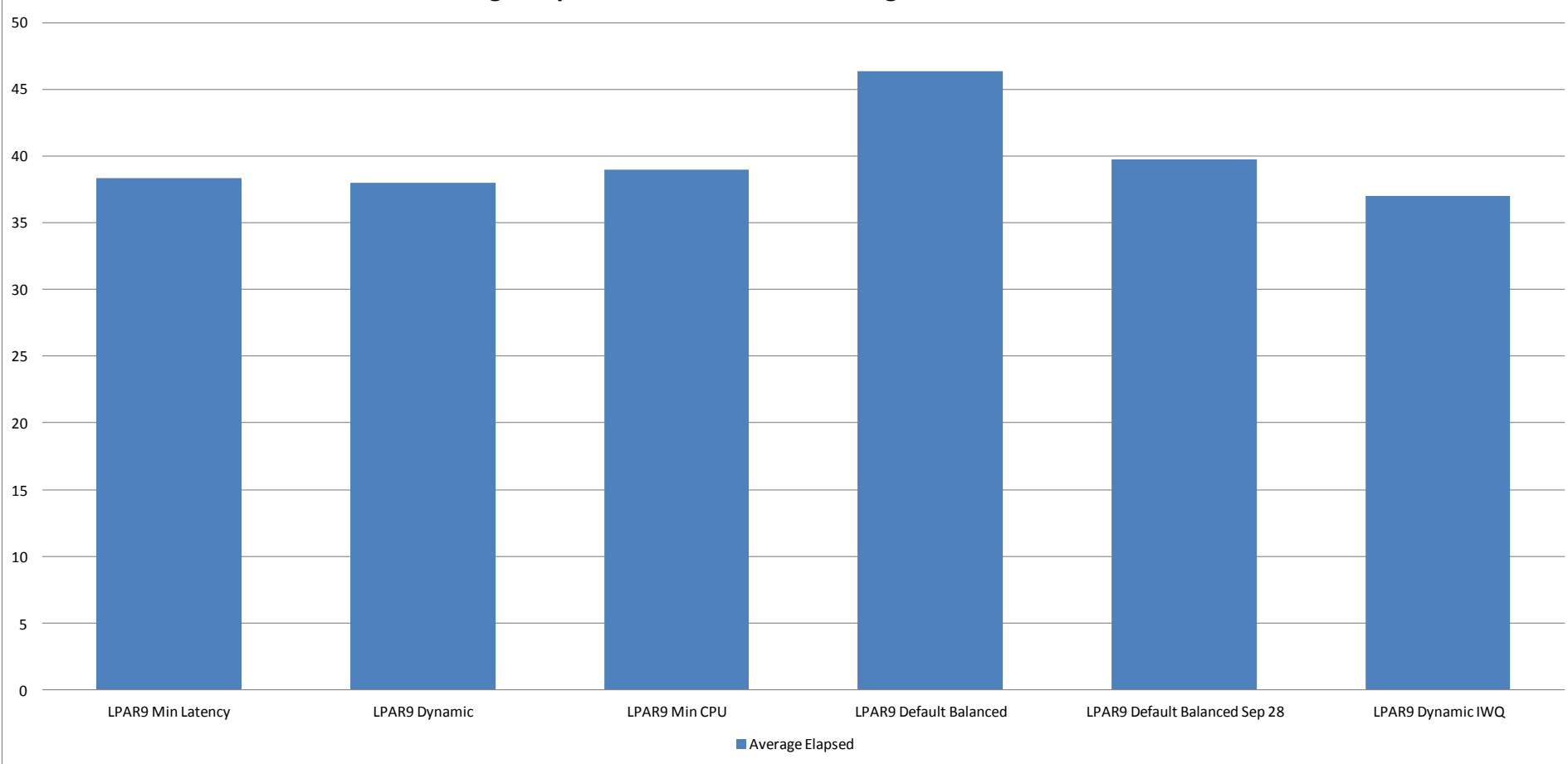
V1.13 from LPAR1 To LPAR9 3.8 GIG NDM Timings INBPERF	AVG TCPIP 10 Sec CPU	Avg Elapsed	CPU Delta compared to MIN CPU	Elapsed Delta Compared to Dynamic IWQ	Bytes in 10 seconds	Million Bytes in 10 secs	Million Bytes per CPU Second
9S01 Min Latency							
9S01 Min Latency							
9S01 Min Latency	1.25	38.33	1.11	0.04	999,036,434	999.04	802.26
9S01 Dynamic							
9S01 Dynamic							
9S01 Dynamic							
9S01 Dynamic	1.06	38.00	0.79	0.03	1,007,799,912	1007.80	951.88
9S01 Min CPU							
9S01 Min CPU							
9S01 Min CPU	0.59	39.00	0.00	0.05	981,958,888	981.96	1,657.31
9S01 Default Balanced							
9S01 Default Balanced							
9S01 Default Balanced	0.63	46.33	0.06	0.25	826,540,935	826.54	1,321.88
9S01 Default Balanced Sep 28							
9S01 Default Balanced Sep 28							
9S01 Default Balanced Sep 28							
9S01 Default Balanced Sep 28	0.56	39.75	0.00	0.07	963,431,362	963.43	1,735.91
9S01 Dynamic IWQ							
9S01 Dynamic IWQ							
9S01 Dynamic IWQ	1.35	37	1.28	0.00	1,035,037,747	1035.04	768.59
9S01 Dynamic IWQ April 3-4							
9S01 Dynamic IWQ							
9S01 Dynamic IWQ							
9S01 Dynamic IWQ	1.36	40.5	1.30		945,590,040		
NDM (bytes)		3,829,639,664					

VTAM CPU Stats

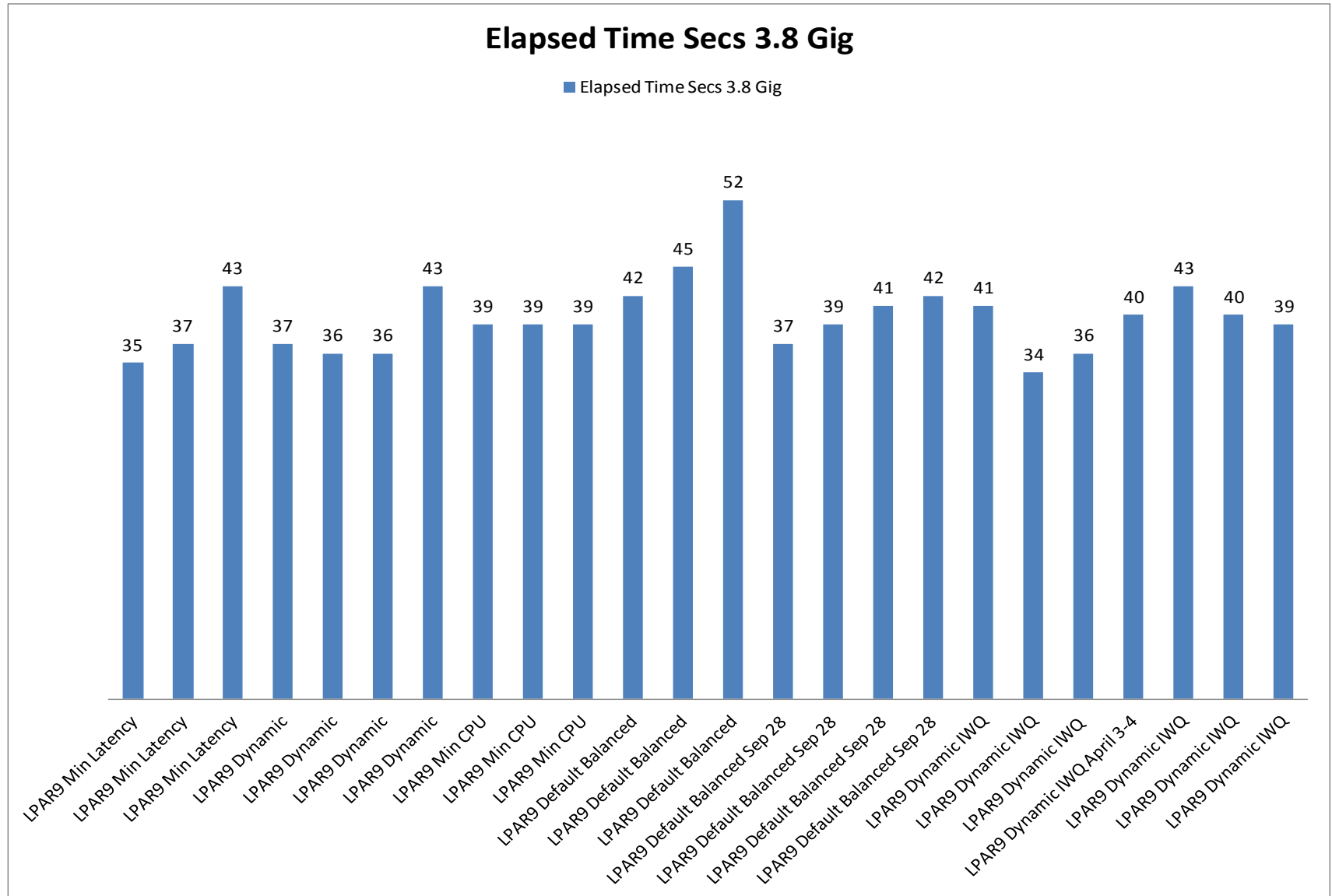
V1.12 TS02946. from LPAR1 to LPAR9 3.8 GIG NDM Timings for INBPERF parameters	Elapsed Time Secs	VTAM RMF Monitor II CPU Time PER 10 SECOND S	Kbytes/ Second	VTAM AVG CPU Seconds Per 10 Seconds	Avg Elapsed Seconds	CPU Delta compared to MIN CPU	Elapsed Delta Compared to Dynamic IWQ
9S01 Min Latency	38	0.04	100,780				
9S01 Min Latency	37	0.05	103,504				
9S01 Min Latency	40	0.04	95,741	0.04	38.33	0.00	0.00
9S01 Min CPU	42	0.04	91,182				
9S01 Min CPU	41	0.05	93,406				
9S01 Min CPU	45	0.04	85,103	0.04	42.67	0.00	0.11
9S01 Default Balanced	38	0.04	100,780				
9S01 Default Balanced	40	0.04	95,741				
9S01 Default Balanced	41	0.04	93,406	0.04	39.67	0.00	0.00
9S01 Dynamic IWQ	39	0.05	98,196				
9S01 Dynamic IWQ	40	0.06	95,741				
9S01 Dynamic IWQ	38	0.05	100,780				
9S01 Dynamic IWQ	39	0.05	98,196	0.05	39.00	0.21	0.00
NDM (bytes)					3,829,639,664		

NDM 3.8 Gig Elapsed Times LPAR1 to LPAR9

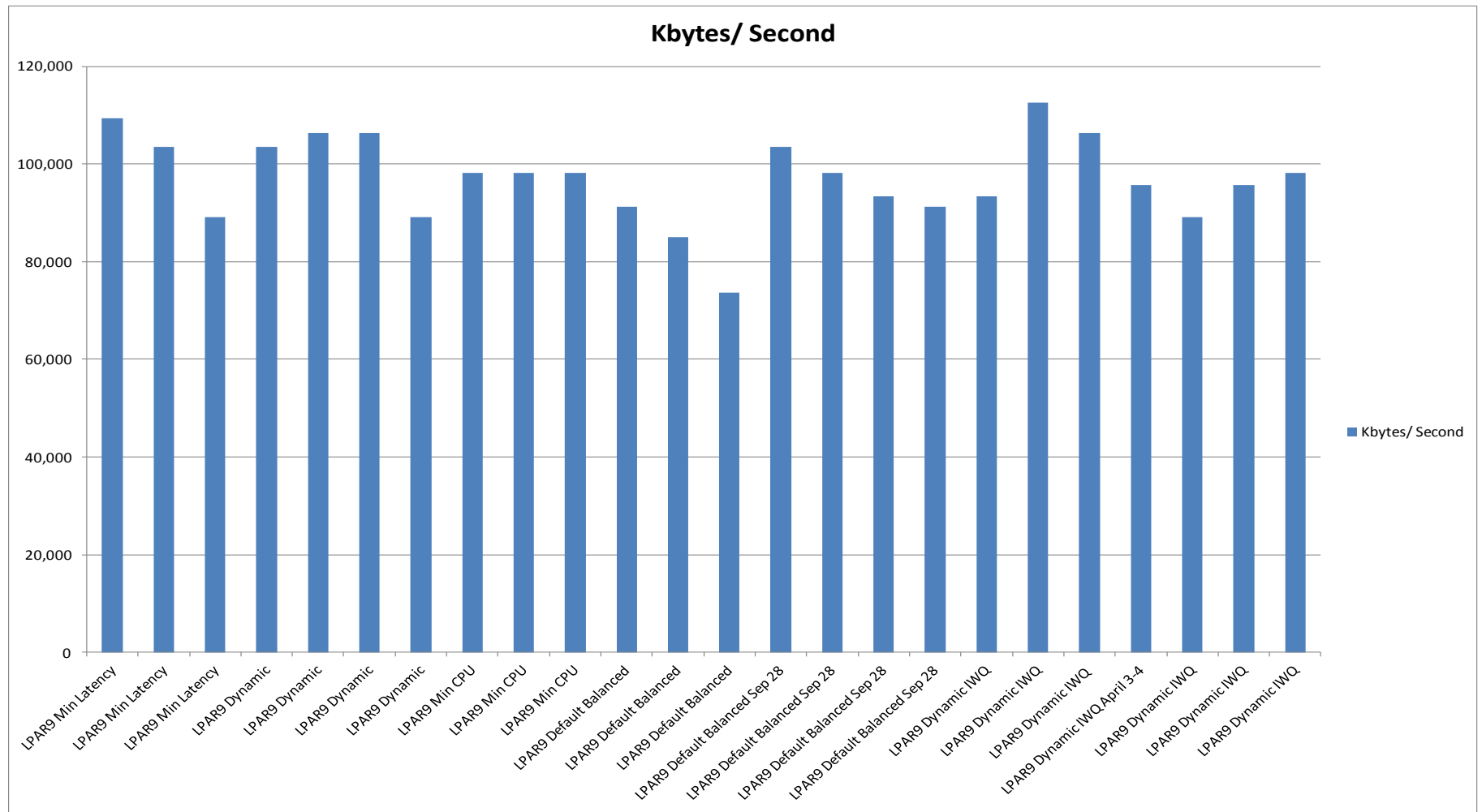
Average Elapsed Seconds to NDM 3.8 Gigs from LPAR1 to LPAR9



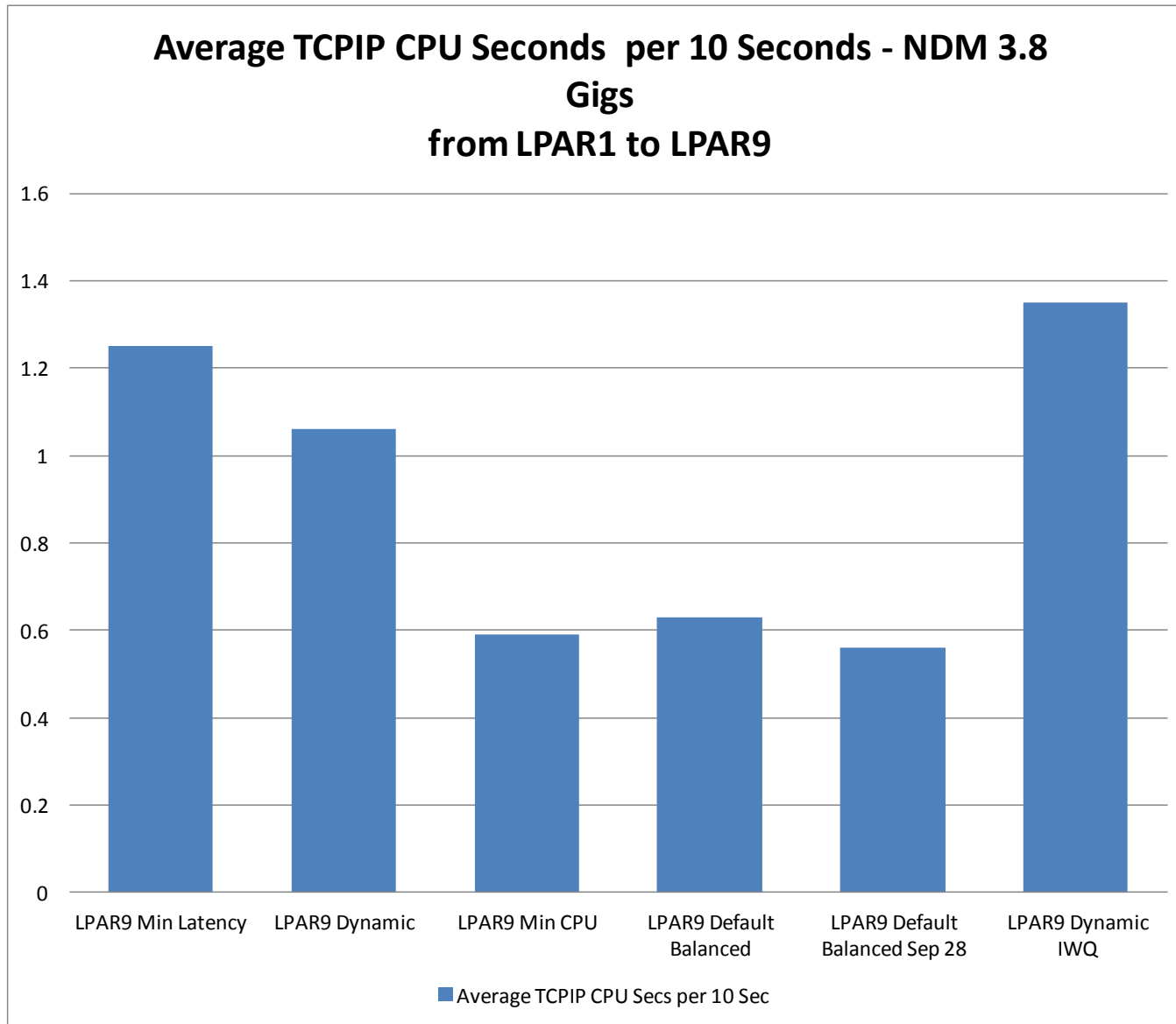
Elapsed Times Detail NDM 3.8 Gig LPAR1 to LPAR9



NDM Kilobytes per Second for 3.8 Gig LPAR1 to LPAR9



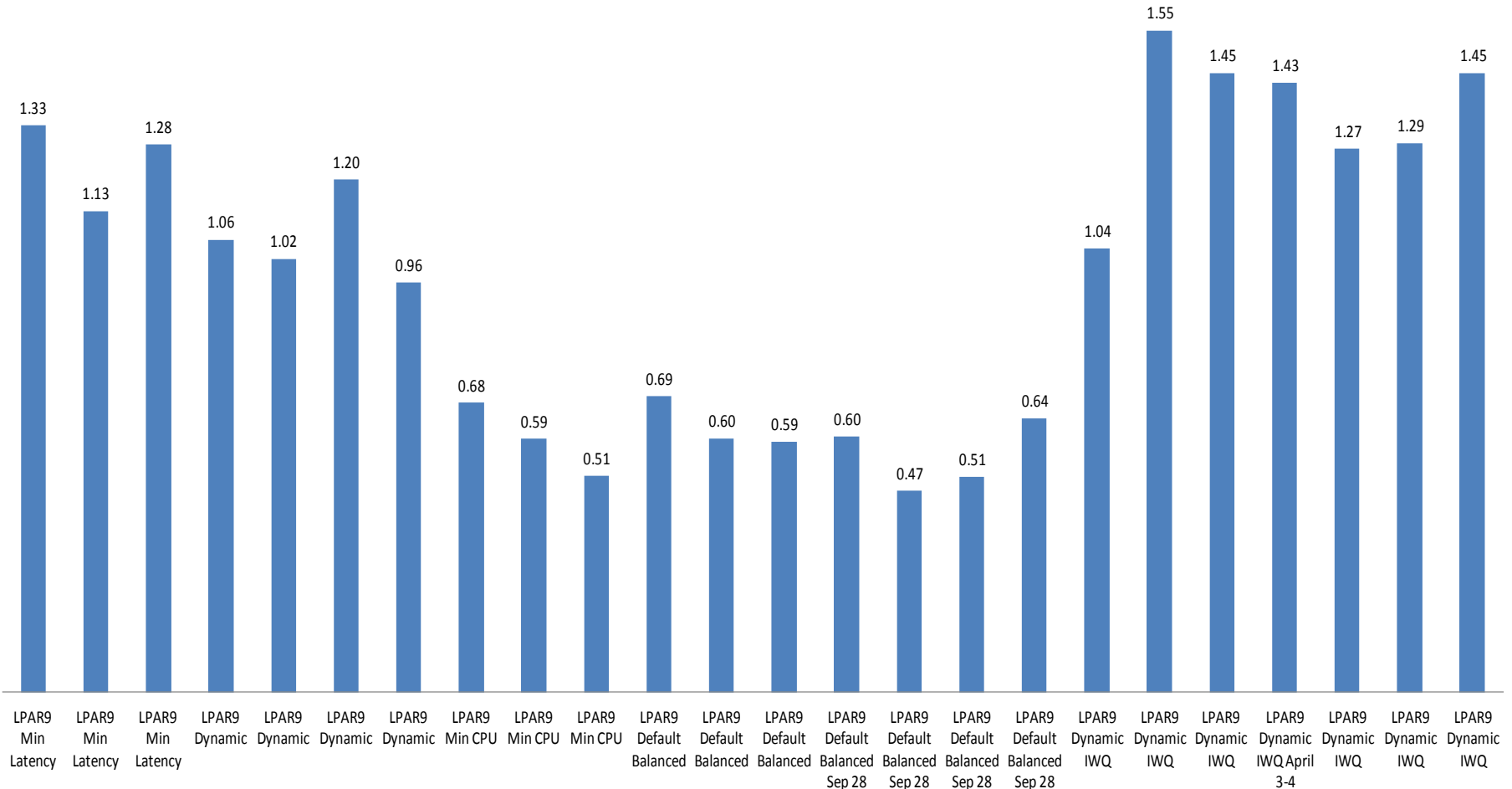
TCPIP CPU Time per 10 Seconds



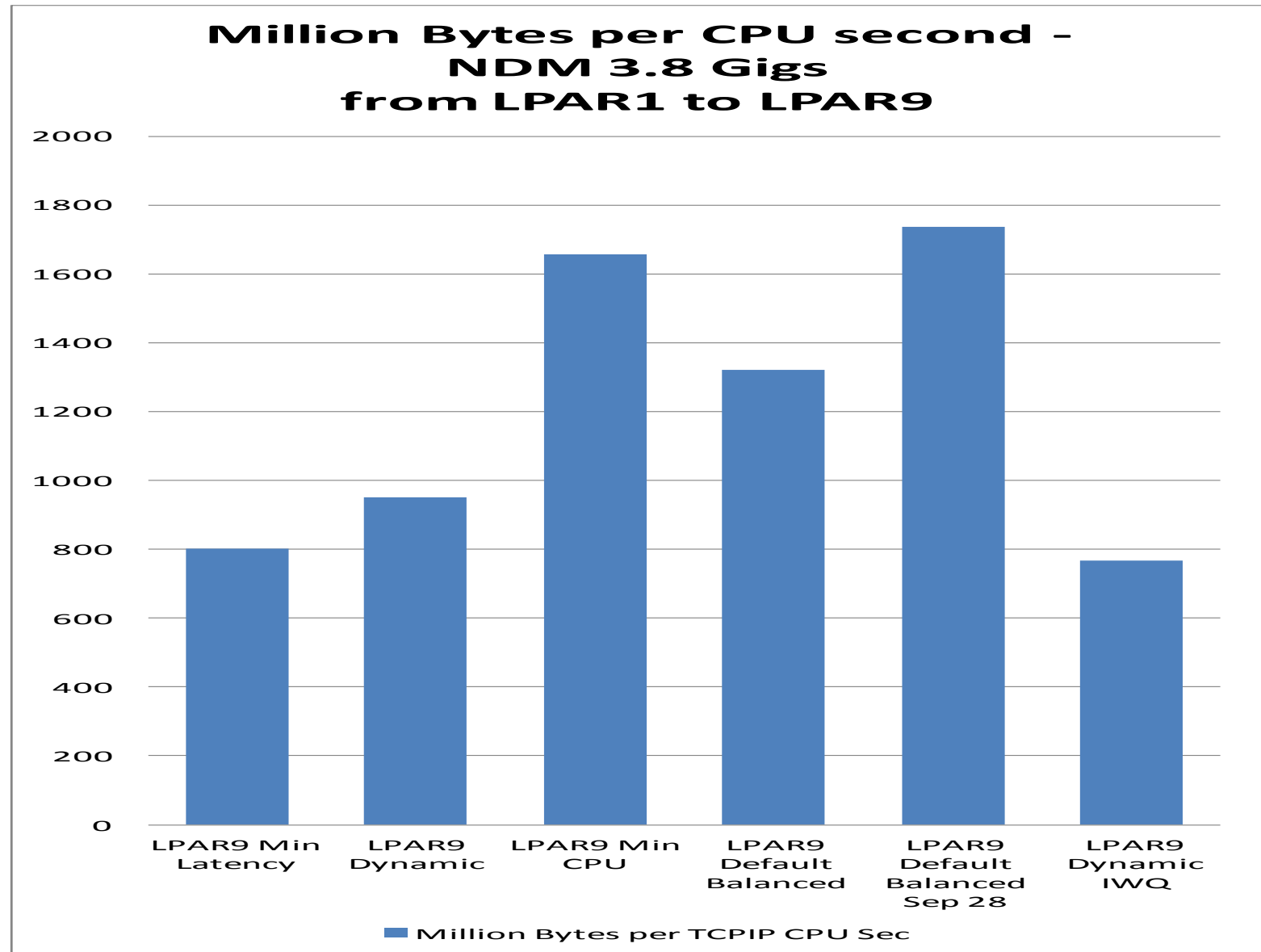
TCPIP CPU Detail

TCPIP CPU Time PER 10 SECONDS

■ TCPIP CPU Time PER 10 SECONDS



Million Bytes per TCPIP CPU Second



Multiple LPARS sharing OSAs

- IBM benchmark used dedicated OSA

SYSNAME	LINK NAME	HOUR	INBOUND BYTES	OUTBOUND BYTES
LPAR1	S0140000	15	18,247,524	7,199,167,639
LPAR1	S0240000	15	10,003,422	3,680,403,244
LPAR1	S0340000	15	13,546,122	3,596,753,643
LPAR1	S0440000	15	51,925,338	3,598,738,021
LPAR9	S1140000	15	1,012,981,728	545,103,133
LPAR9-2	S1140000	15	76,027,934	198,331,394
LPAR9-K	S1240000	15	21,342,886	37,121,219
LPAR9	S1240000	15	83,311,981	284,615,982
LPAR9-2	S1240000	15	85,229,724	145,399,468
LPAR9-4	S1240000	15	49,281,574	118,453,779
LPAR9-K	S1340000	15	12,016,610	20,932,567
LPAR9	S1340000	15	18,179,085,510	472,147,106
LPAR9-2	S1340000	15	109,530,980	153,237,641
LPAR9-K	S1440000	15	35,550,324	37,527,502
LPAR9	S1440000	15	8,586,742,965	1,017,281,117
LPAR9-2	S1440000	15	77,430,118	268,026,242
	TN3270			
	Sessions	Hour	Bytes In	Bytes Out
	270	15	1595896	55191994

Sample Type 50 Report with IWQ

			Sum							
			OSA- EXPRESS READS EXHAUSTED	OSA-EXP REPLENISHMENT	OSA- EXPRESS PROC DEFER	OSA- EXPRESS PCIR FOR	OSA- EXPRESS PCIT FOR	OSA- EXPRESS PCIU FOR	OSA- EXPRESS PCIV FOR	OSA- EXPRESS SBAL COUNT
SYSTEM* ID	SUBCHANNEL POLARITY	DEVICE ADDRESS	FOR READ	DEFERRALS FOR READ	REQD QUEUE	READ QUEUE	READ QUEUE	READ QUEUE	READ QUEUE	FOR READ
LPAR9	RD/1	DEVA	33	6,438	14,932	32,606,368	55,513	145,069	9,791,934	31,065,061
		DEVE	6	9,471	636	23,247,921	37,153	112,117	5,457,601	23,259,525
		DEV2	1,617	59,434	1,160	35,447,474	273,278	526,151	14,386,433	57,295,466
		DEV6	573	58,440	2,430	32,776,243	127,673	455,996	12,480,568	46,883,576
	RD/2	DEVA	118	0	0	0	0	0	0	15,971,061
		DEVE	5	0	0	0	0	0	0	4,962,317
		DEV2	105	0	0	0	0	0	0	7,325,585
		DEV6	121	0	0	0	0	0	0	8,239,819
	RD/4	DEVA	0	0	0	0	0	0	0	2,550,730
		DEVE	0	0	0	0	0	0	0	4,673,429
		DEV2	0	0	0	0	0	0	0	198,404
		DEV6	0	0	0	0	0	0	0	437,277

IBM Research Follow-up

- Presented results to Patrick Brown the author of Inbound Workload Queuing whitepaper
- We hypothesized that the lack of “think time” between interactive traffic in the IBM benchmark could account for the differences

Summary of Results

- IWQ with Dynamic INBPERF had best throughput but required more than double the CPU of Min CPU and Balanced (default)
- Min CPU had lowest CPU with only 5 % degradation of throughput
- Our heaviest NDM traffic is in the middle of the night with little interactive traffic.
- Balanced (default) meets our current objectives however new requirements for additional throughput and combined workloads that previously were isolated cause us to continue to evaluate and consider IWQ .

Description of IWQ from IBM Announcement

- OSA-Express3 or later features can perform a degree of traffic sorting by placing inbound packets for differing workload types on separate processing queues. This function is called QDIO inbound workload queuing (IWQ). With the inbound traffic stream already sorted by the OSA-Express feature, z/OS Communications Server provides the following performance optimizations:
- Finer tuning of read-side interrupt frequency to match the latency demands of the various workloads that are serviced
- Improved multiprocessor scalability, because the multiple OSA-Express input queues are now efficiently serviced in parallel
- When QDIO IWQ is enabled, z/OS Communications Server and the OSA-Express feature establish a primary input queue and one or more ancillary input queues (AIQs), each with a unique read queue identifier (QID) for inbound traffic. z/OS Communications Server and the OSA-Express feature cooperatively use the multiple queues in the following way:
 - The OSA-Express feature directs an inbound packet (received on this interface) that is to be forwarded by the sysplex distributor to the sysplex distributor AIQ. z/OS Communications Server then tailors its processing for the sysplex distributor queue, notably by using the multiprocessor to service sysplex distributor traffic in parallel with traffic on the other queues.
 - The TCP layer automatically detects connections operating in a bulk-data fashion (such as the FTP data connection), and these connections are registered to the receiving OSA-Express feature as bulk-mode connections. The OSA-Express feature then directs an inbound packet (received on this interface) for any registered bulk-mode connection to the TCP bulk-data AIQ. z/OS Communications Server tailors its processing for the bulk queue, notably by improving in-order packet delivery on multiprocessors, which likely results in improvements to CPU consumption and throughput. Like other AIQs, processing for data on the bulk queue can be in parallel with traffic on the other queues.
 - The OSA-Express feature directs an inbound Enterprise Extender packet (received on this interface) to the Enterprise Extender AIQ. This allows z/OS Communications Server to process inbound traffic on the Enterprise Extender queue in parallel with inbound traffic on the other queues for this interface.

IWQ Restrictions from IBM Announcement

- QDIO IWQ is not supported for IPAQENET interfaces defined with the DEVICE, LINK, and HOME statements. You must convert your IPAQENET definitions to use the INTERFACE statement to enable this support.
- QDIO IWQ is not supported for a z/OS guest on z/VM® using simulated (virtual) devices, such as virtual switch (VSWITCH) or guest LAN.
- Bulk-mode TCP connection registration is supported only in configurations in which a single inbound interface is servicing the bulk-mode TCP connection. If a bulk-mode TCP connection detects that it is receiving data over multiple interfaces, QDIO IWQ is disabled for the TCP connection and inbound data from that point forward is delivered to the primary input queue.
- QDIO IWQ does not apply for traffic that is sent over an OSA port that is shared by the receiving TCP/IP stack when an indirect route (where the next hop and destination IP address are different) is being used; this traffic is placed on the primary input queue. QDIO IWQ does apply when traffic on the shared OSA path uses a direct route (where the next hop and destination IP address are the same).