IBM zSystems | **E**nterprise **N**etworking **S**olutions (**ENS**)

# V2R5: Z/OS COMMUNICATIONS SERVER PERFORMANCE SUMMARY REPORT

Created By: Kaji Rashad

Last Updated On: May 11th, 2022

# Table of Contents

# Trademarks, Notices, and Disclaimers

The following are trademarks of the International Business Machines Corporation in the United States and/or other countries:

| | | |
|---|---|---|
| BigInsights | HyperSwap | System z10* |
| BlueMix | IBM* | Tivoli* |
| CICS* | IBM (logo)* | UrbanCode |
| COGNOS* | IMS | WebSphere* |
| Db2* | Language Environment* | z13 |
| DFSMSdfp | MQSeries* | z14 |
| DFSMSdss | Parallel Sysplex* | z15 |
| DFSMShsm | PartnerWorld* | z16 |
| DFSORT | RACF* | zEnterprise* |
| DS6000* | Rational* | z/OS* |
| DS8000* | Redbooks* | zSecure |
| FICON* | REXX | z Systems |
| GDPS* | SmartCloud* | z/VM* |

* Registered trademarks of IBM Corporation

The following are trademarks or registered trademarks of other companies:
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 therefrom.
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.
IT Infrastructure Library is a registered trademark of the Central Computer and Telecommunications Agency which is now part of the Office of Government Commerce.
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.
Java and all Java based trademarks and logos are trademarks or registered trademarks of Oracle and/or its affiliates.
Linear Tape-Open, LTO, the LTO Logo, Ultrium, and the Ultrium logo are trademarks of HP, IBM Corp. and Quantum in the U.S.
Linux is a registered trademark of Linus Torvalds 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.
OpenStack is a trademark of OpenStack LLC. The OpenStack trademark policy is available on the OpenStack website.
Red Hat is a trademark of Red Hat Inc, an IBM Company.
TEALEAF is a registered trademark of Tealeaf, an IBM Company.
Windows Server and the Windows logo are trademarks of the Microsoft group of countries.
Worklight is a trademark or registered trademark of Worklight, an IBM Company.
UNIX is a registered trademark of The Open Group in the United States and other countries.
* Other product and service names might be trademarks of IBM or other companies.

Notes:
Performance results are 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. 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. This information provides only general descriptions of the types and portions of workloads that are eligible for execution on Specialty Engines (e.g, zIIPs, zAAPs, and IFLs) ("SEs"). IBM authorizes customers to use IBM SE only to execute the processing of Eligible Workloads of specific Programs expressly authorized by IBM as specified in the “Authorized Use Table for IBM Machines” provided at www.ibm.com/systems/support/machine_warranties/machine_code/aut.html (“AUT”). No other workload processing is authorized for execution on an SE. IBM offers SE at a lower price than General Processors/Central Processors because customers are authorized to use SEs only to process certain types and/or amounts of workloads as specified by IBM in the AUT.

# Acknowledgments

The z/OS Communications Server Performance Team would like to **thank** the following IBMers for their input on this report:

David Herr | Senior Software Engineer | z/OS Communications Server

Christopher Nyamful | Software Engineer | z/OS Communications Server

Michael Fitzpatrick | Senior Technical Staff Member | Architecture & Design for ENS

Dhananjay Patel

# Tips for Reading This Document

1) Clicking on any row in the Table of Contents will take the reader to that specific section or subsection of the document[1]
2) All hyperlinks redirect to an external webpage or internal section/sub-section

[1] PDF application must support this feature

# Preface

The performance measurements discussed in this document were collected using dedicated system environments. Results obtained in other configurations or operating system environments may vary significantly depending upon environments used. Therefore, no assurance can be given, and there is no guarantee that an individual user will achieve performance or throughput improvements equivalent to the results stated here. Readers of this document should verify the applicable data for their specific environment.

The Central Processor Unit (CPU) numbers listed includes only z/OS host networking related CPU overhead (including dispatching costs) on **general Central Processors** (CPs) from the network device driver layer up through the application socket layer. The socket applications used in the micro-benchmarks for this publication have *no application logic*, so the CPU numbers represent the total application cost which in this case equates to the network related costs. With typical production workloads, network related cost is a *small fraction* of the overall application transaction cost.

**Note:** In all benchmarks, the best practices recommended by z/OS Communications Server were utilized *when* applicable:

- ✓ `GLOBALCONFIG` ADJUSTDVipamss
- ✓ `INBPERF` DYNAMIC
  - WORKLOADQ (IWQ)
    - Client & Server Side
- ✓ `IPCONFIG` SEGMENTATIONOFFLoad (LSO)
- ✓ `IPCONFIG` QDIOACCELerator
- ✓ `TCPCONFIG` AUTODELAYAck
- ✓ MSG_WAITALL[2]
- ✓ Jumbo Frames (e.g., HOST MTU 8192)

[2] A socket read flag utilized by the application to instruct the TCP layer to delay completion of a Socket Receive or Read call until the full length of the requested data is available in the TCP receive buffer [1].

# Hardware Information

**z15**
Machine Type (Model): 8561 – T01

**Cryptographic Coprocessor Level**
Crypto Express-6S:  6.6.11
Crypto Express-7S:  7.3.26

# Workload Naming Convention

**Introduction**

You decipher the listed workloads in the following way:
[NameOfBenchmark][#OfClients](BytesSentByClient/BytesSentByServer)

For example, [RR][10](1B/100B) is interpreted as Request Response benchmark with 10 clients sending 1 byte and receiving 100 bytes from the server.

**Generic Workloads**

**RR**x(y/z): x number of clients doing **R**equest **R**esponse transactions where the client is opening a connection and performing a series of transactions sending y bytes and receiving a response of z bytes
**CRR**x(y/z): x number of clients doing **C**onnect **R**equest **R**esponse transactions where the client is performing a series of transactions opening a connection, sending y bytes, receiving a response of z bytes, and closing the connection
**STR**x(y/z): x number of clients doing **Str**eaming transactions where the client is sending y bytes and receiving a response of z bytes

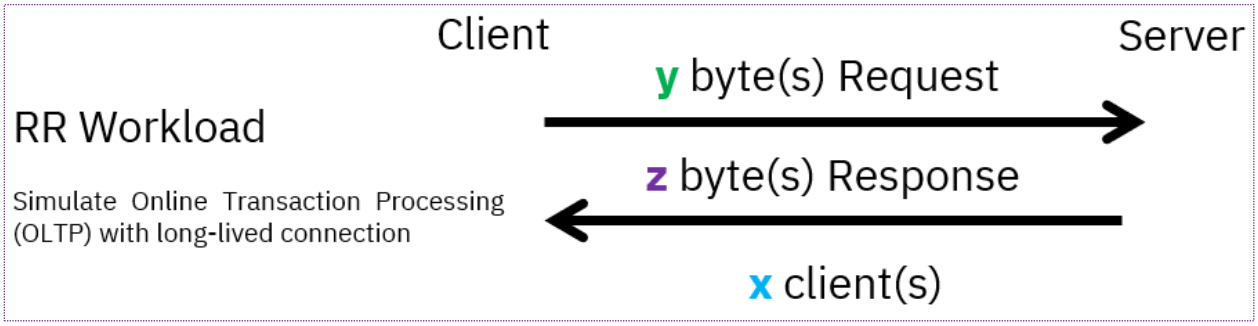


**Figure 1: Request response workload**

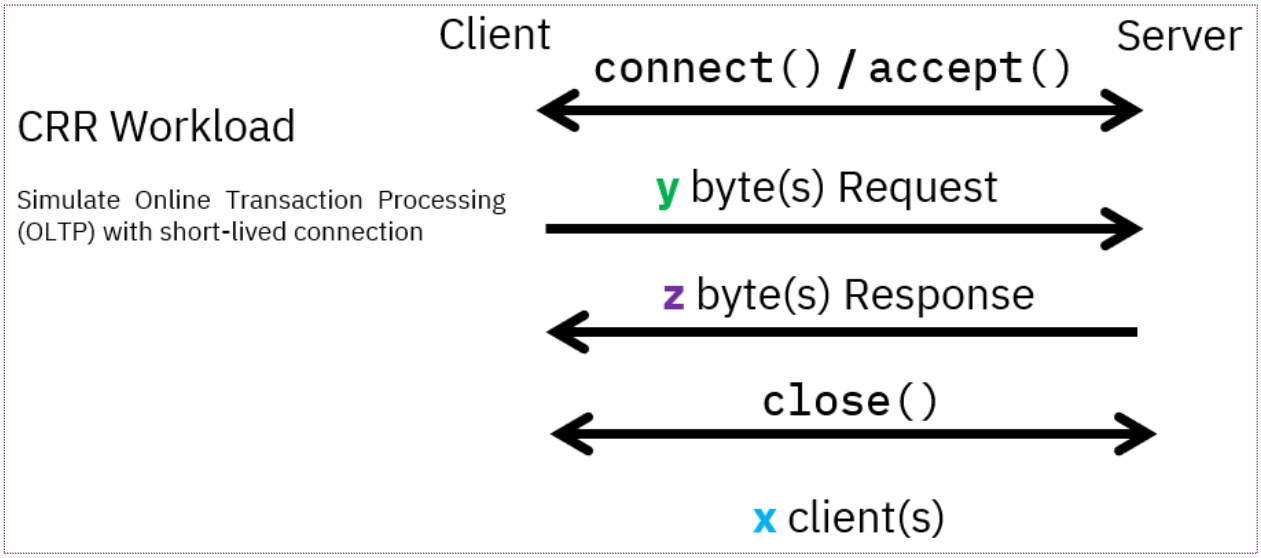


**Figure 2: Connect request response workload**

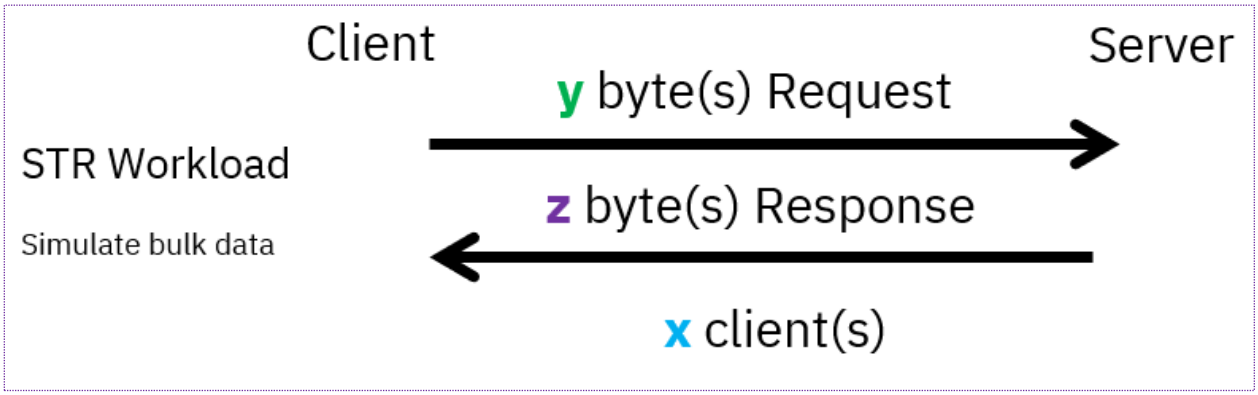


**Figure 3: Streaming workload**

**Examples**

RR40(100B/100B): In an instance of time, there are 40 clients browsing a webpage hosted on a server in which each HTTP GET request of 100 bytes contains a response of 100 bytes.

CRR9(200B/200B): In an instance of time, there are 9 clients sending a HTTP GET request containing 200 bytes and receives a response containing 200 bytes which allows them to log into their bank portal. The core difference between a RR and CRR workload is the duration of the connection. In CRR, the connection is closed after each transaction. A common use case for a bank portal is logging in to audit the balance before logging out.

STR3(1B/20MB): In an instance of time, there are 3 clients sending a 1 byte request and receiving a 20MB file in response.

**CPU Cost/Tran & Transaction [Trans/sec]**

On some graphs, the reader will observe a key legend of “CPU Cost/Tran” and “Transaction [Trans/sec]”. Our measurement uses RMF to determine the average CPU utilization. RMF results show the CPU utilization across all online CPs during a sampling interval which is taken into consideration when performing our calculations.

For example, if RMF result shows a LPAR utilization of 25% across 4 CPs then we translate this into 100% of 1 CP. If the sampling interval is 10 seconds and we are averaging 100% of 1 CP then the benchmark consumes 10 seconds of CPU during the sampling period. If there were 1 million transactions during the 10 second sampling interval then there was a transaction rate of 1,000,000 trans / 10 seconds (or 100,000 trans/sec) and a CPU Cost/Tran of 10 CPU Seconds / 1,000,000 Trans (or 10 us/tran).

# Performance Best Practices: General

**GLOBALCONFIG ADJUSTDVipamss**

A Generic Routing Encapsulation (GRE) header is added to a packet when using Sysplex Distributor with VIPAROUTE. Therefore, the packet could be fragmented as it travels from the distributor to the target stack. ADJUSTDVipamss takes the GRE length into consideration. In coding this in the distributor stack's TCP/IP profile, the GRE header is taken into the MSS value calculation. In return, a packet will not be fragmented as it travels from the distributor to the target stack. Refer to this article for more information.

**INBPERF DYNAMIC**

Processing inbound traffic for the OSA-Express interface in Queued Direct Input Output (QDIO) mode dynamically exploits an OSA hardware function called Dynamic LAN Idle. The DYNAMIC setting reacts to changes in traffic patterns and dynamically sets the interrupt-timing values to maximize throughput. Refer to this article for more information.

**QDIO Inbound Workload Queueing (WORKLOADQ)**

The core benefits of Inbound Workload Queueing (IWQ) are "finer tuning of read-side interrupt frequency to match the latency demands of the various workloads that are serviced" and "improved multiprocessor scalability as multiple OSA-Express input queues are efficiently serviced in parallel" [2]. Each queue is tailored for its specific need. For instance, the bulk queue is tailored for improved "in-order packet delivery on multiprocessor, which likely results in improvements to CPU consumption and throughput" [2]. QDIO IWQ provides benefit on both sides of the connection hence it is enabled on both sides in our test set-up when applicable. Note that WORKLOADQ requires the processing of inbound traffic for the QDIO interface to be set as DYNAMIC (e.g., INBPERF DYNAMIC WORKLOADQ). Refer to this article for more information.

**IPCONFIG SEGMENTATIONOFFLoad (LSO)**

Any large amount of data traveling over the network is broken down into smaller segments by the TCP/IP stack. This process can be CPU intensive. As an alternative, segmentation offload (i.e., Large Send Offload) is an OSA-Express feature. It reduces host CPU utilization, increases data transfer efficiency, and offloads segmentation processing to OSA [3].

**TCPCONFIG AUTODELAYAck**

Reduction in network traffic and CPU utilization can be achieved by delaying the TCP acknowledgement (ACK) *depending* on the traffic pattern. AUTODELAYAck enables the TCP stack to "automatically enable or disable a delayed ACK in a TCP connection based on the characteristic of the traffic" [4].

**IPCONFIG QDIOACCELerator**

QDIO Accelerator specifies that inbound packets that are to be forwarded by a TCP/IP stack are eligible to be routed directly between any of the following combinations of interface types: a HiperSockets interface and an OSA-Express QDIO interface, two OSA-Express QDIO interfaces, and two HiperSockets interfaces. These packets arrive at the forwarding stack, but do not traverse all the TCP/IP layers for forwarding. Therefore, valuable TCP/IP resources (storage and CPU) are not expended for purposes of routing and forwarding packets. This option also applies to packets that would be forwarded by the Sysplex Distributor. Refer to this article more information.

**MSG_WAITALL**

MSG_WAITALL is beneficial in streaming workloads. The flag bit decreases the frequency of interrupts occurring for the application receiving data as less interrupts can result in improvements to CPU consumption and throughput. The receiving application is interrupted only when all requested data can be returned. To avoid blocking the application indefinitely, the flag bit should only be set in scenarios where the application expects to receive enough data to fill its buffer or the connection will terminate.

**Jumbo Frames**

When a client and host communicate with each other over a network, it is possible to utilize a higher Maximum Transmission Unit (MTU) size if the *entire network path* supports it. A higher MTU size can reduce the amount of segmentation that occurs for larger payloads, which may result in a higher throughput and reduced CPU cycles [5]. If Jumbo Frames is configured, then enable path MTU discovery.

# Performance Best Practice: Security

**HEAPPOOLS64**

Application Transparent Transport Layer Security (AT-TLS) creates System SSL environments using the z/OS Language Environment (LE). These System SSL environments use the LE runtime default options or those specified in the CEEPRMxx parmlib member. The default LE runtime does not have HEAPPOOLS64 enabled. For large AT-TLS configurations, running without HEAPPOOLS64 enabled could result in additional contention for user heap storage across the different System SSL environments. This could lead to slow-downs or timeouts processing TLS handshakes. By enabling the HEAPPOOLS64 runtime option, this contention for user heap storage can be eliminated. The following measurements (e.g., Figure 4 & 5) were gathered under the same z/OS hardware and software environment used for the AT-TLS section of this report. As evident by Figure 4, HEAPPOOLS64 has a positive impact on connections using AT-TLS. In our measurements, the positive impact was measured across all TLSv1.2 and TLSv1.3 ciphers.

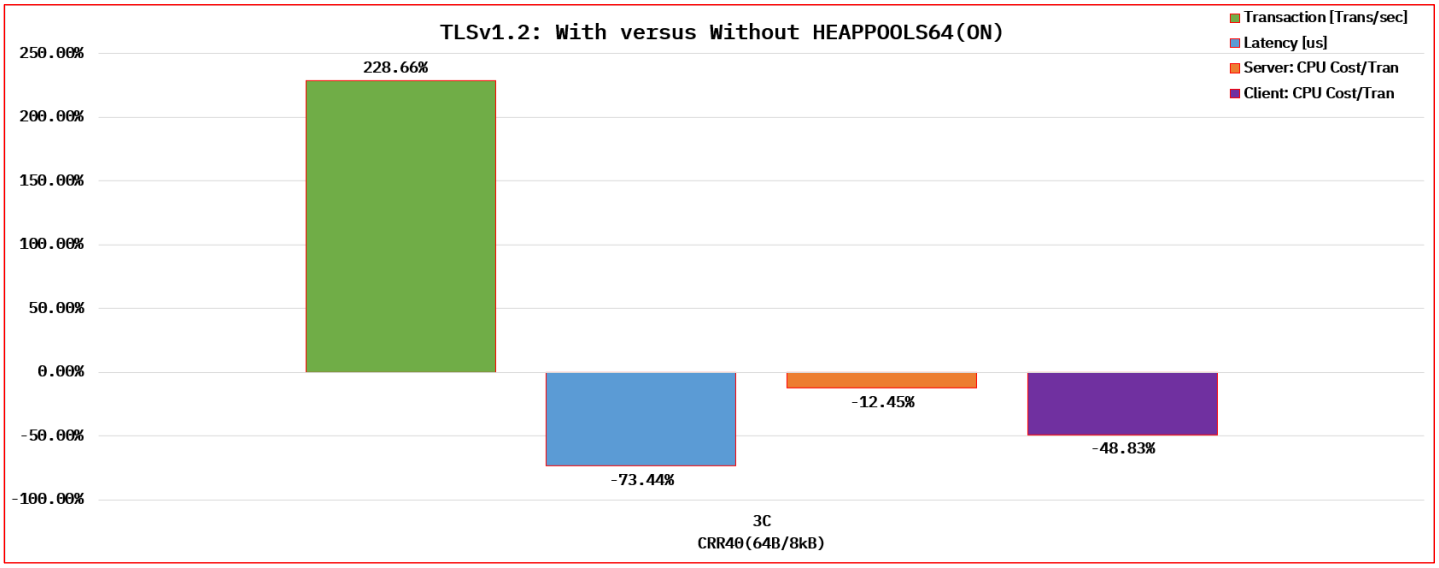


**Figure 4**

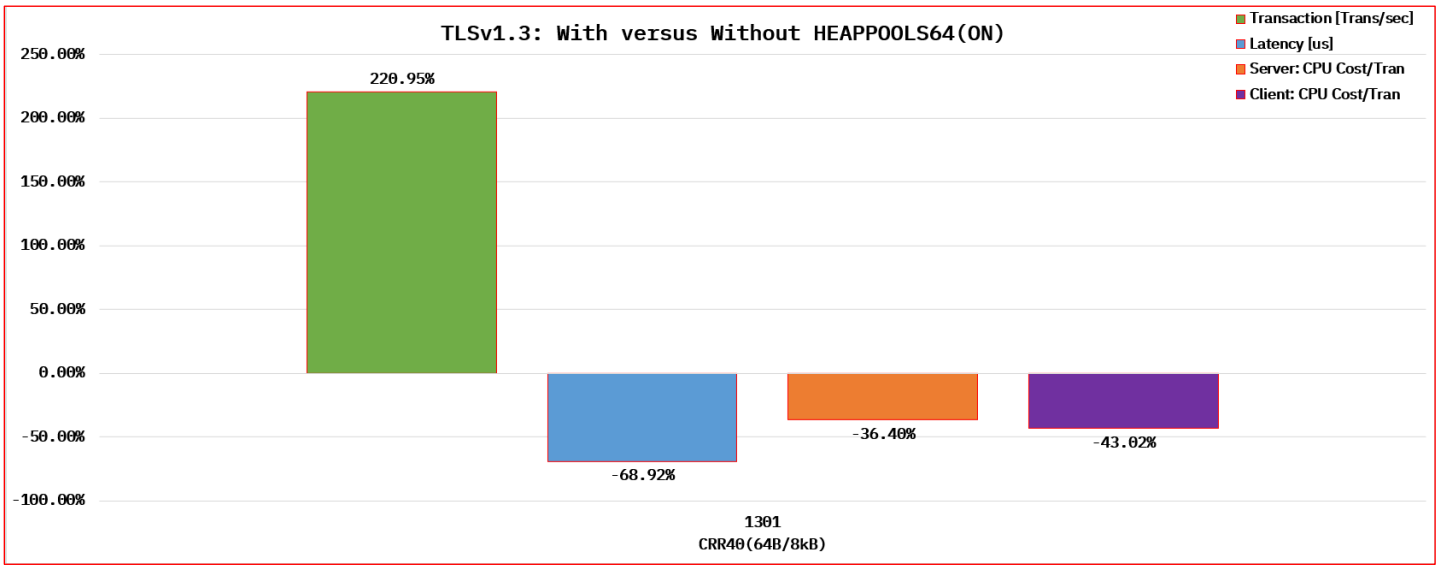


**Figure 5**

# V2R5: New Function

## SMCv2

### General Background

In the initial release of Shared Memory Communication (SMC), routing a packet from one subnet to another was not allowed. This prevented some clients from utilizing SMC. To ease the adoption rate, SMCv2 lifts the subnet limitation.

### Background: SMC-D

Shared Memory Communications Direct Memory Access (SMC-D) uses Internal Shared Memory (ISM) in allowing two SMC capable peers to communicate intra-CPC. During the TCP connection handshake, the capable peers dynamically detect SMC eligibility before using it. The protocol boosts workload performance by providing a low latency and high bandwidth solution. Refer to this article for more information.

### Test Environment: SMC-Dv2 versus SMC-Dv1 Over Same Network Topology

In the following sections, the focus is on comparing SMC-D v2 against v1. The goal is to verify whether SMC-Dv2 performance is inline with SMC-Dv1. The below two diagrams shows one of the differences between SMC-D v2 and v1.

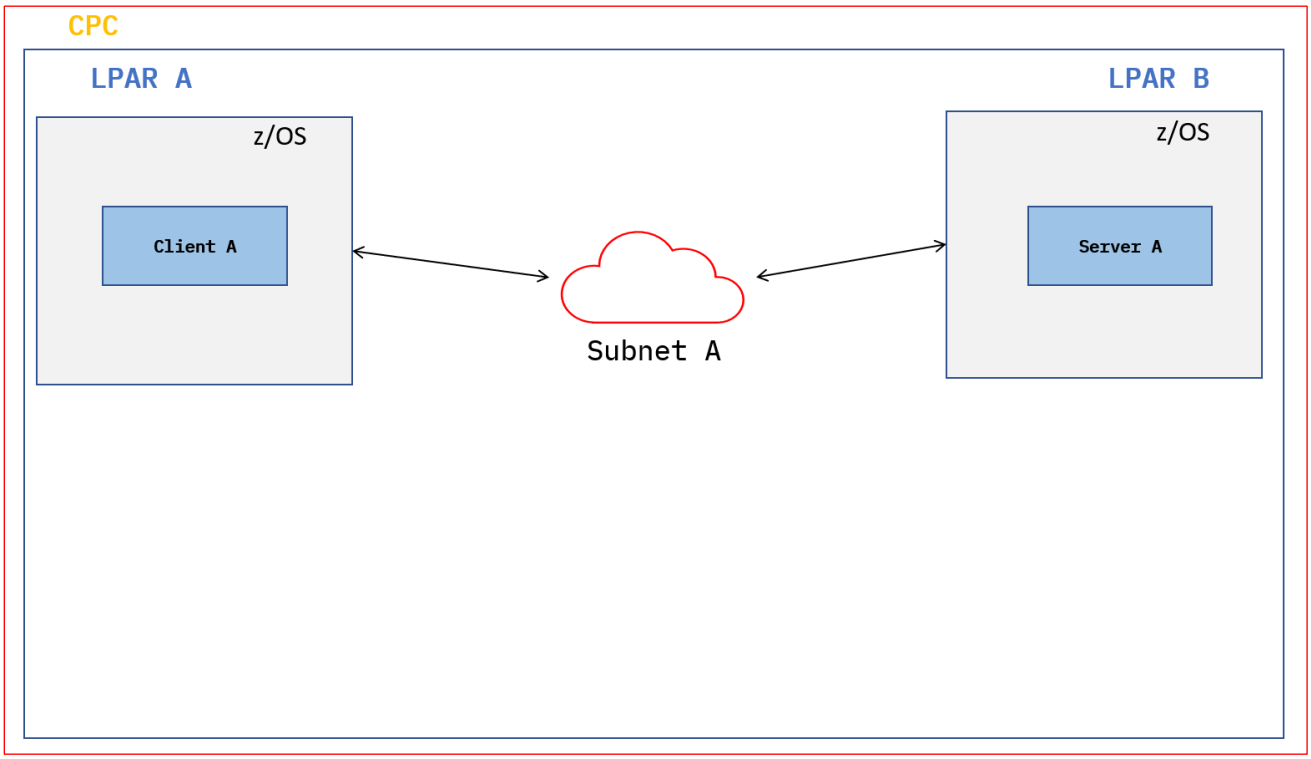


**Figure 6: SMC-Dv1 set-up where packets cannot traverse multiple subnets within the CPC**

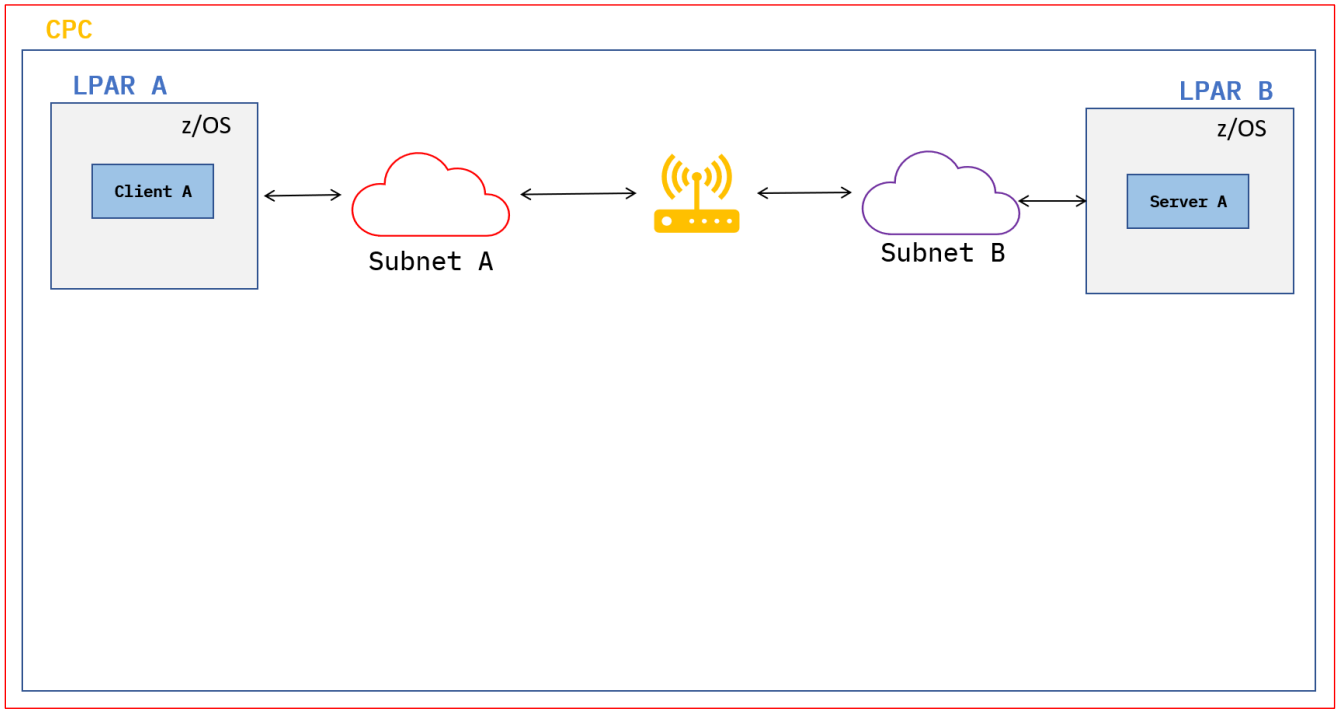


**Figure 7: SMC-Dv2 set-up where packets can traverse multiple subnets within the CPC**

## z/OS Environment Configuration: SMC-Dv2 versus SMC-Dv1 Over Same Network Topology

Below is the environment configuration in which the data was collected:

- Central Processor Complex (CPC): z15
- Release: V2R5
- Number of CPUs: 4 (Dedicated) per LPAR
- Interface: ISMv1 and ISMv2
- Workloads
  - RR10(4kB/4kB)
  - STR3(1B/20MB) (i.e., GET)

### RR Observation

Based on studying the workload results, SMC-Dv2 performs as equivalent to SMC-Dv1.

### RR Result

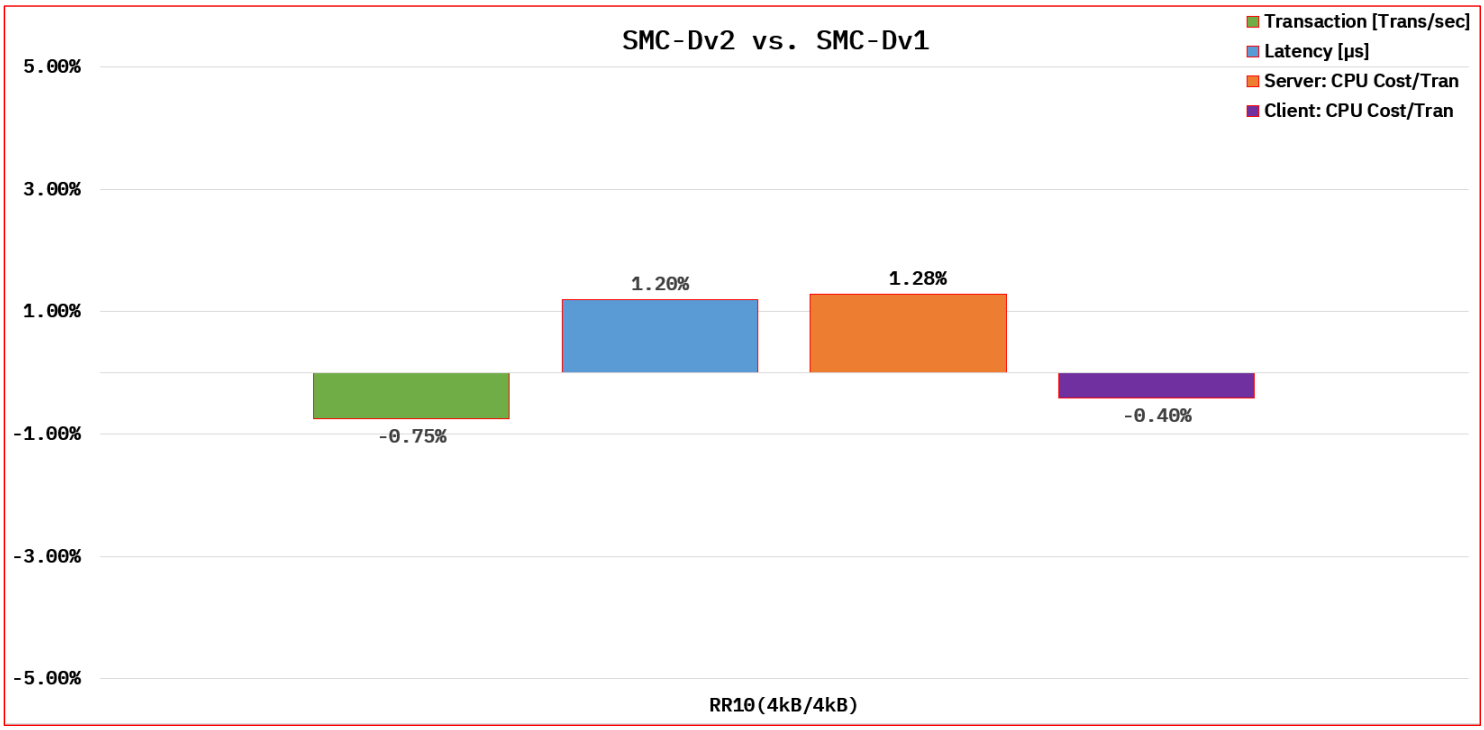


**Figure 8: SMC-Dv2 performance is as equivalent to SMC-Dv1**

### STR Observation

A GET through SCM-Dv2 performs as well as SMC-Dv1.

**STR Result**

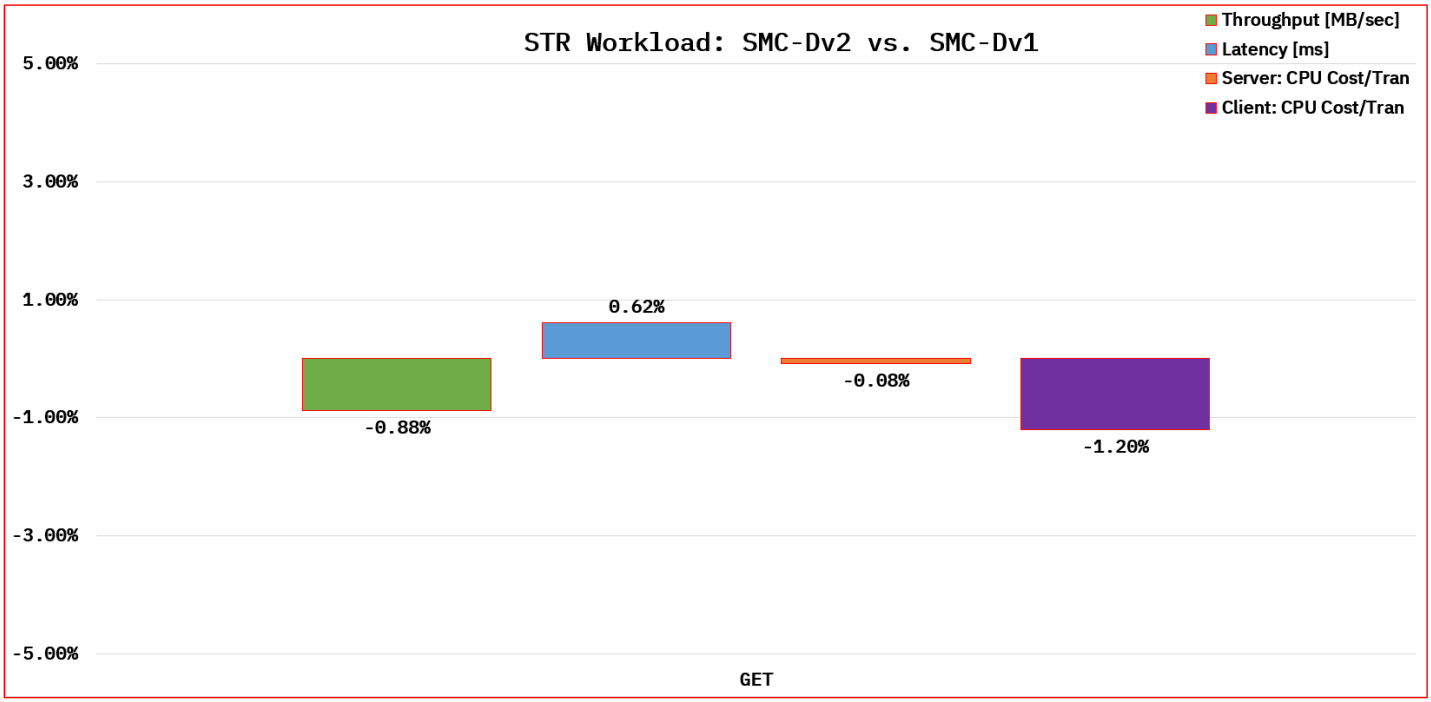


**Figure 9: Doing a GET through SMC-Dv2 performs as well as SMC-Dv1**

## Test Environment: SMC-Dv2 versus HiperSockets Over Same Network Topology

In the following sections, the focus is on comparing SMC-Dv2 against HiperSockets. Both are intra-CPC memory to memory communication solution. However, the former (e.g., SMC-D) does not require any TCP/IP processing for memory-to-memory data movements hence it is faster.

### z/OS Environment Configuration: SMC-Dv2 versus HiperSockets Over Same Network Topology

Below is the environment configuration in which the data was collected:

- CPC: z15
- Release: V2R5
- Number of CPUs: 4 (Dedicated) Per LPAR
- Interface
  - ISMv2
  - IUTIQDIO
    - Maximum Frame Size (MFS): 16 [kB] & 64 [kB]
- Workload
  - RR10(4kB/4kB)
  - STR3(20MB/1B) (i.e., PUT)

## RR & STR Observation

As expected, SMC-Dv2 significantly outperforms HiperSockets in all aspects: transaction rate, throughput, latency, and CPU cost.

## RR Result

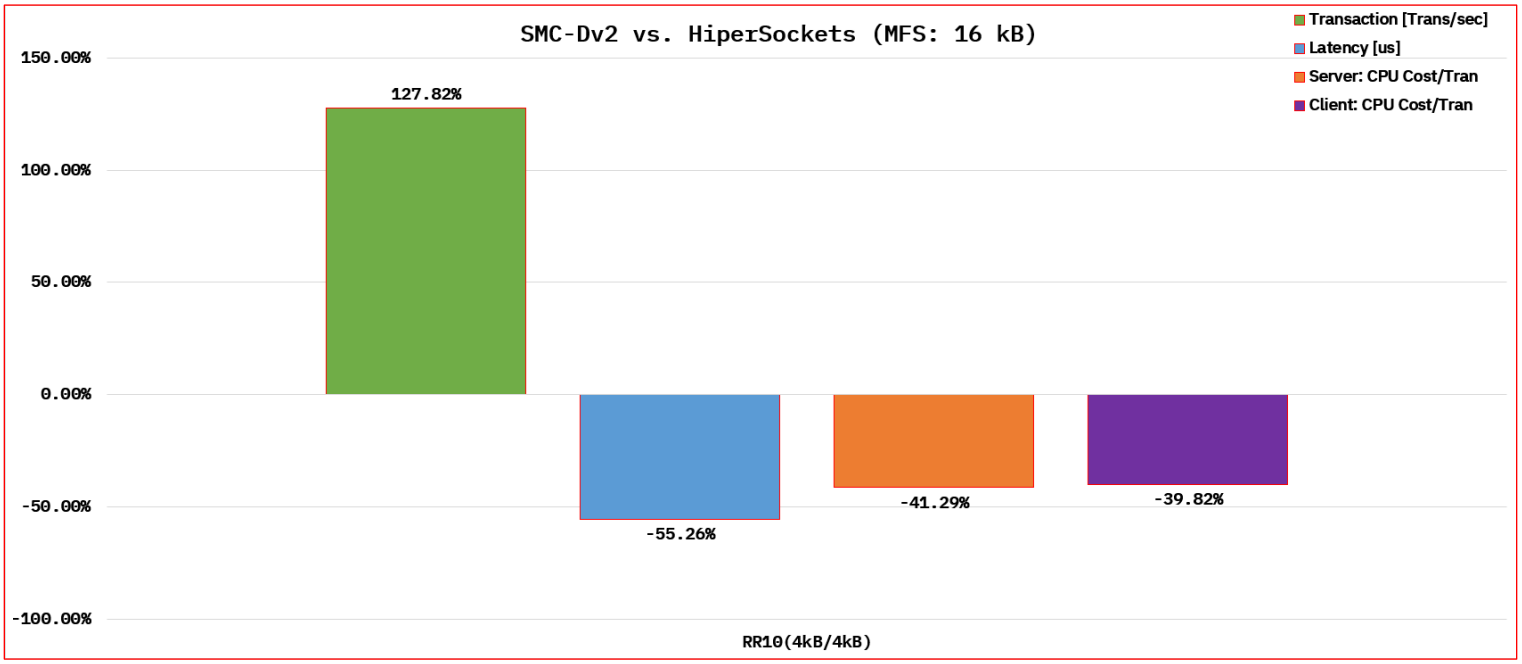


**Figure 10: SMC-Dv2 achieves much higher transaction rate with lesser delay and CPU cost in comparison to HiperSockets**

## STR Results

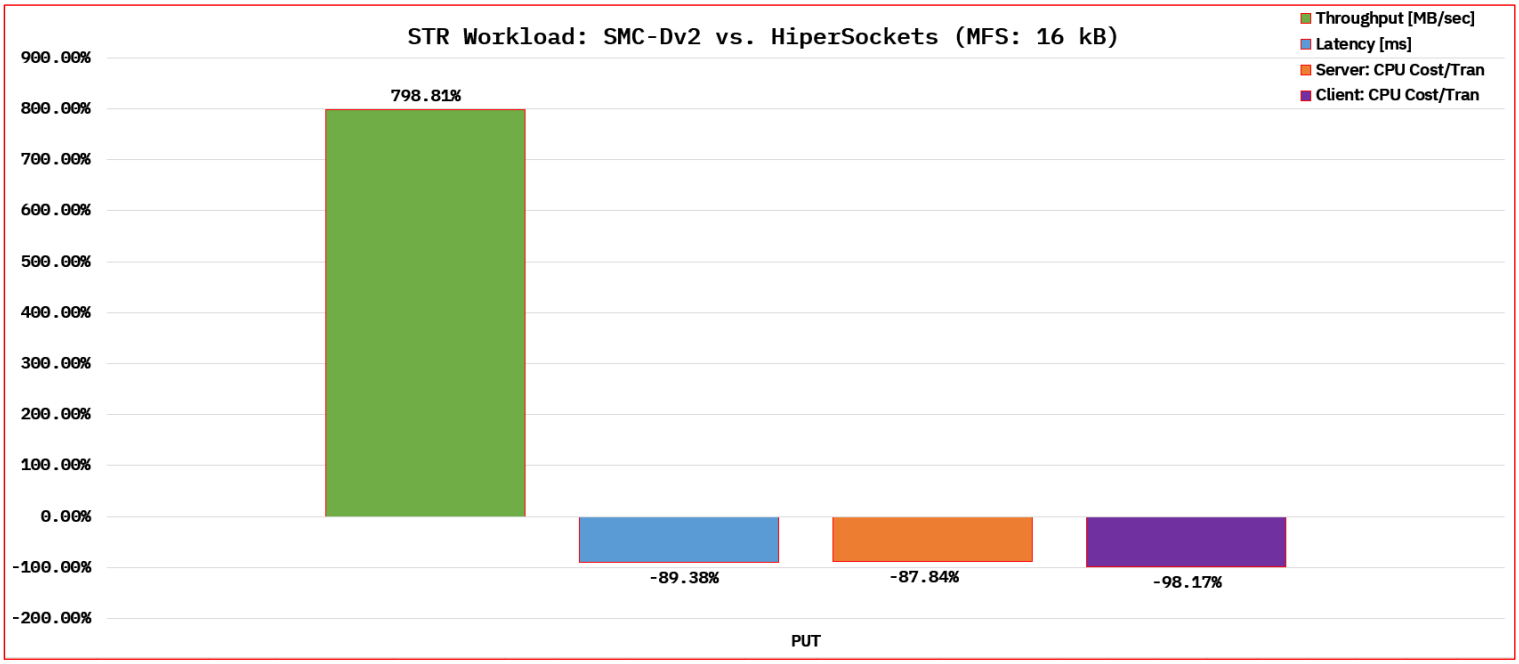


**Figure 11: SMC-Dv2 achieves much higher throughput with lower latency and CPU cost in comparison to HiperSockets**

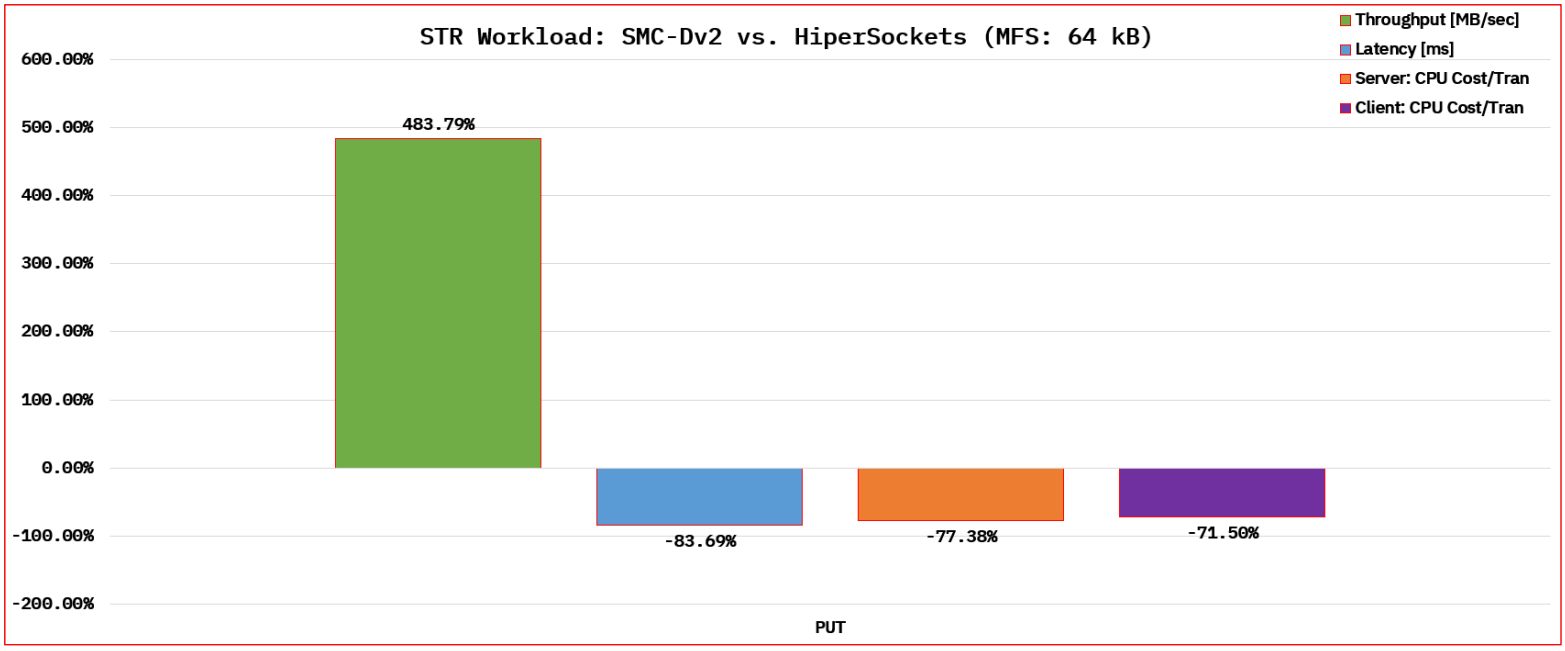


**Figure 12: SMC-Dv2 still significantly outperforms HiperSockets while the highest MFS is in usage**

## Test Environment: SMC-Rv2 versus SMC-Rv1 Over a Single Subnet via 25GbE RoCE Express3

In the following sections, the focus is on comparing **SMC-Rv2 against SMC-Rv1** via 25GbE RoCE Express3. Both CPCs share a common subnet. In other words, this is **not** a multi-hop experiment. It is possible to use SMC-Rv2 with a single subnet network topology (refer to figure 13). The goal is to verify whether SMC-Rv2 performance is inline with SMC-Rv1 performance. The below two diagrams shows one of the major differences between SMC-Rv2 versus SMC-Rv1.

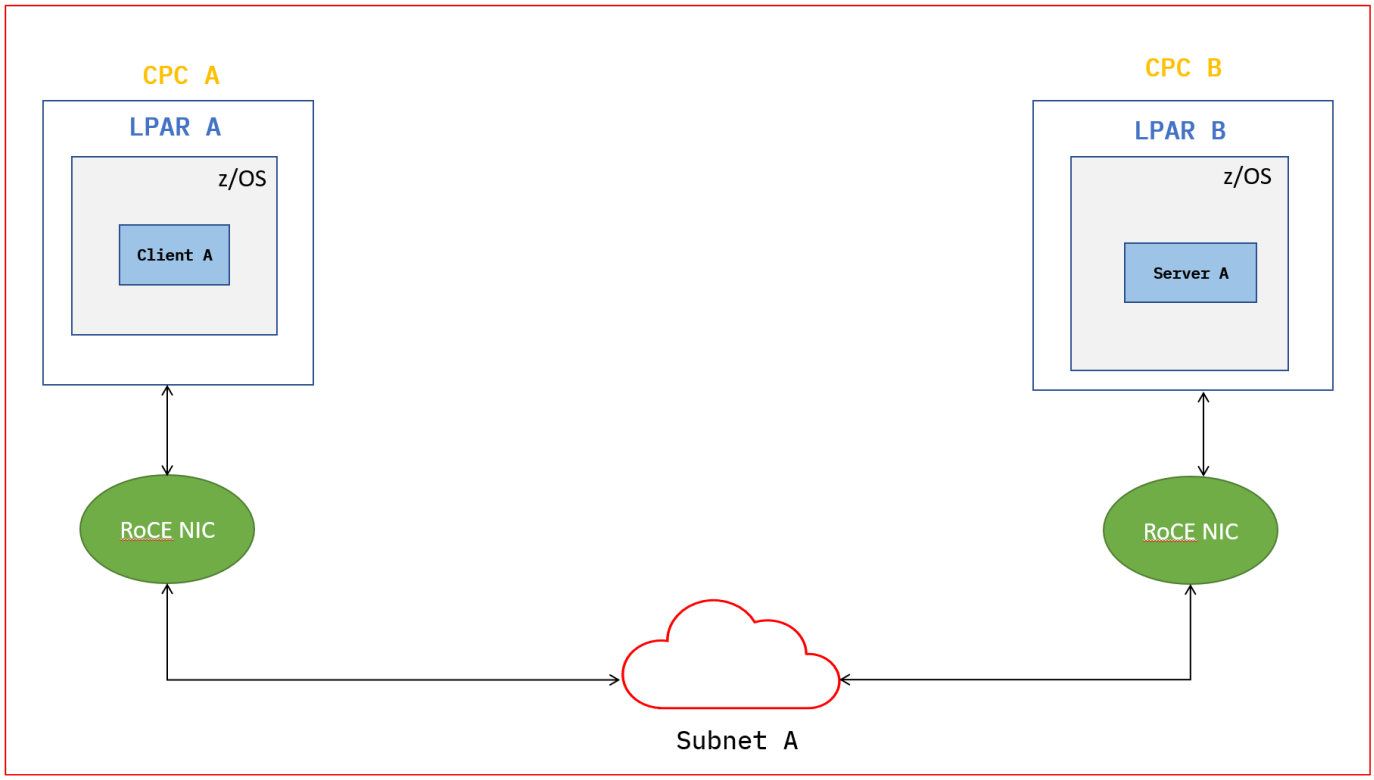


**Figure 13: SMC-Rv1 does not allow the traversal of multiple subnets**

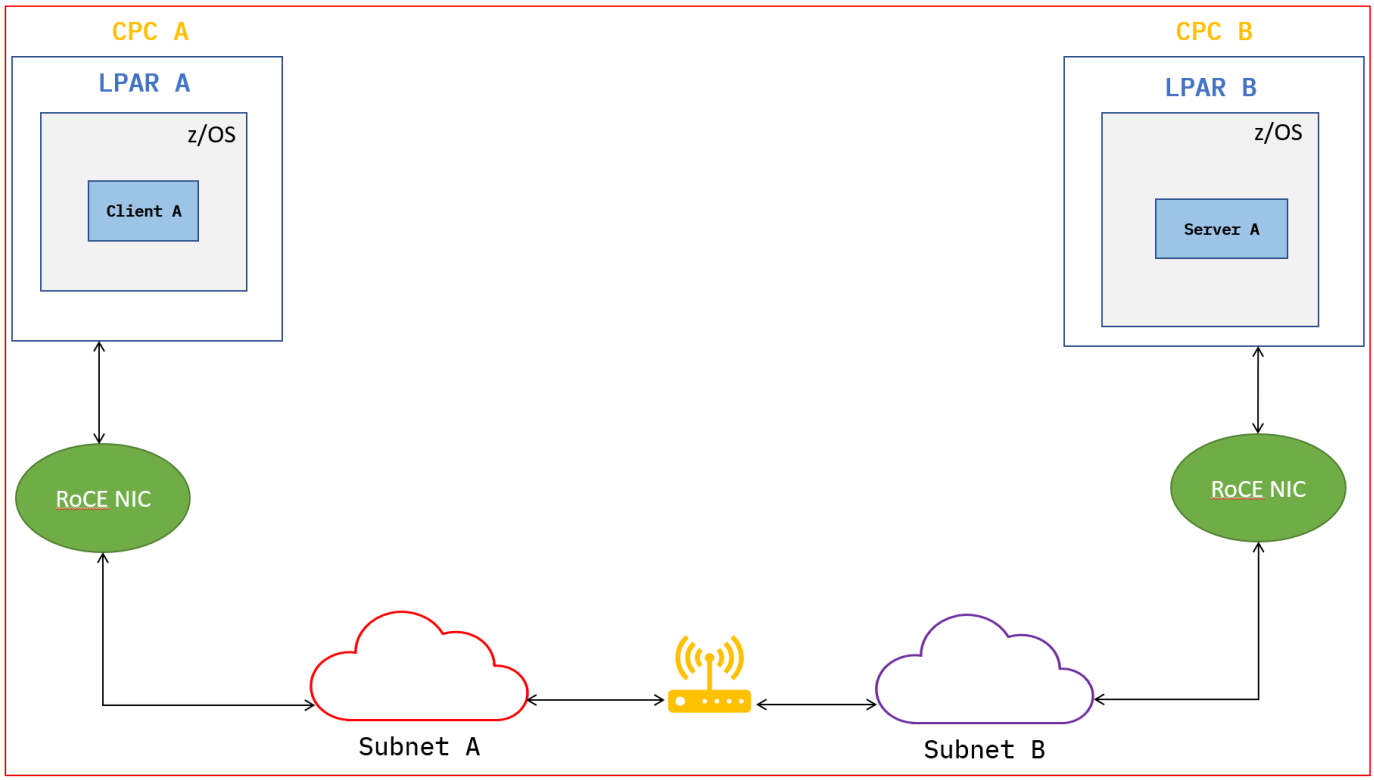


**Figure 14: SMC-Rv2 does allow the traversal of multiple subnets**

**z/OS Environment Configuration: SMC-Rv2 (Directly Attached) versus SMC-Rv1 via 25GbE RoCE Express3 (Single Subnet)**

Below is the environment configuration in which the data was collected:

- CPC: z15
- Release: V2R5
- Number of CPUs: 4 (Dedicated) per LPAR
- Interface: 25GbE RoCE Express3
- Workloads
  - RR60(4kB/4kB)
  - STR3(1B/20MB)

**RR Observation**

For request response workload, SMC-Rv2 (directly attached) performs as equivalent to SMC-Rv1.

## RR Result


**Figure 15: SMC-Rv2 performs equivalent to SMC-Rv1 in a single subnet**

## STR Observation

For streaming workload, there were minimal change in throughput when comparing SMC-Rv2 (directly attached) against SMC-Rv1.

**STR Result**

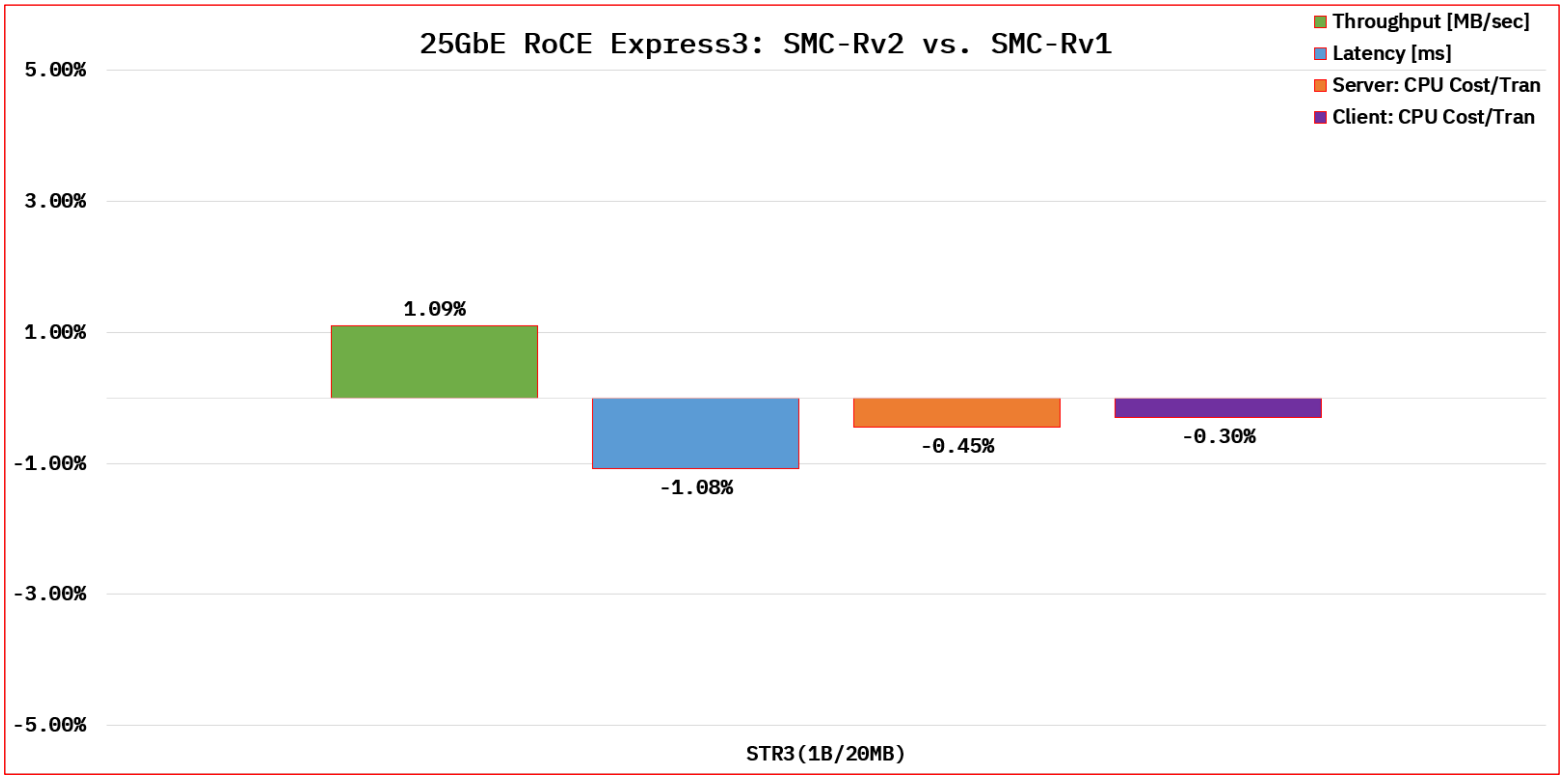


**Figure 16: SMC-Rv2's STR throughput is just as good as SMC-Rv1's throughput**

## Test Environment: SMC-Rv2 (Indirect) via 25GbE RoCE Express2 versus TCP/IP via OSA-Express 7S 25GbE (Multiple Subnets)

In the following sections, the focus is on comparing SMC-Rv2 via 25GbE RoCE Express2 against TCP/IP via OSA-Express 7S 25GbE in multi-hop environment. Both CPCs do **not** share a common subnet. In other words, this is a multi-hop experiment (refer to figure 14). The goal is to verify whether SMC-Rv2 is superior to TCP/IP in a multi-hop environment.

**z/OS Environment Configuration: SMC-Rv2 (Indirect) via 25GbE RoCE Express2 versus TCP/IP via OSA-Express 7S 25GbE (Multiple Subnets)**

Below is the environment configuration in which the data was collected:

- CPC: z15
- Release: V2R5
- Number of CPUs: 4 (Dedicated) per LPAR
- Interface: 25GbE RoCE Express2 & OSA-Express 7S 25GbE
- L3 Switch Speed: 25GbE
- Number of Hop(s): 1
- Workloads
  - RR60(4kB/4kB)
  - STR3(1B/20MB)

**RR Observations**

SMC-Rv2, compared to TCP/IP in a multi-hop environment, is more CPU conservative. Typical TCP processing of packets (e.g., segmentation, flow control, congestion control, etc.) is not being performed in the stack rather on the RNIC when using the SMC-Rv2 (i.e., RoCEv2) protocol. The RNIC processes the packet in an efficient manner. Due to this, less CPU cycles are used to achieve the same amount of transaction rate.

**RR Result**

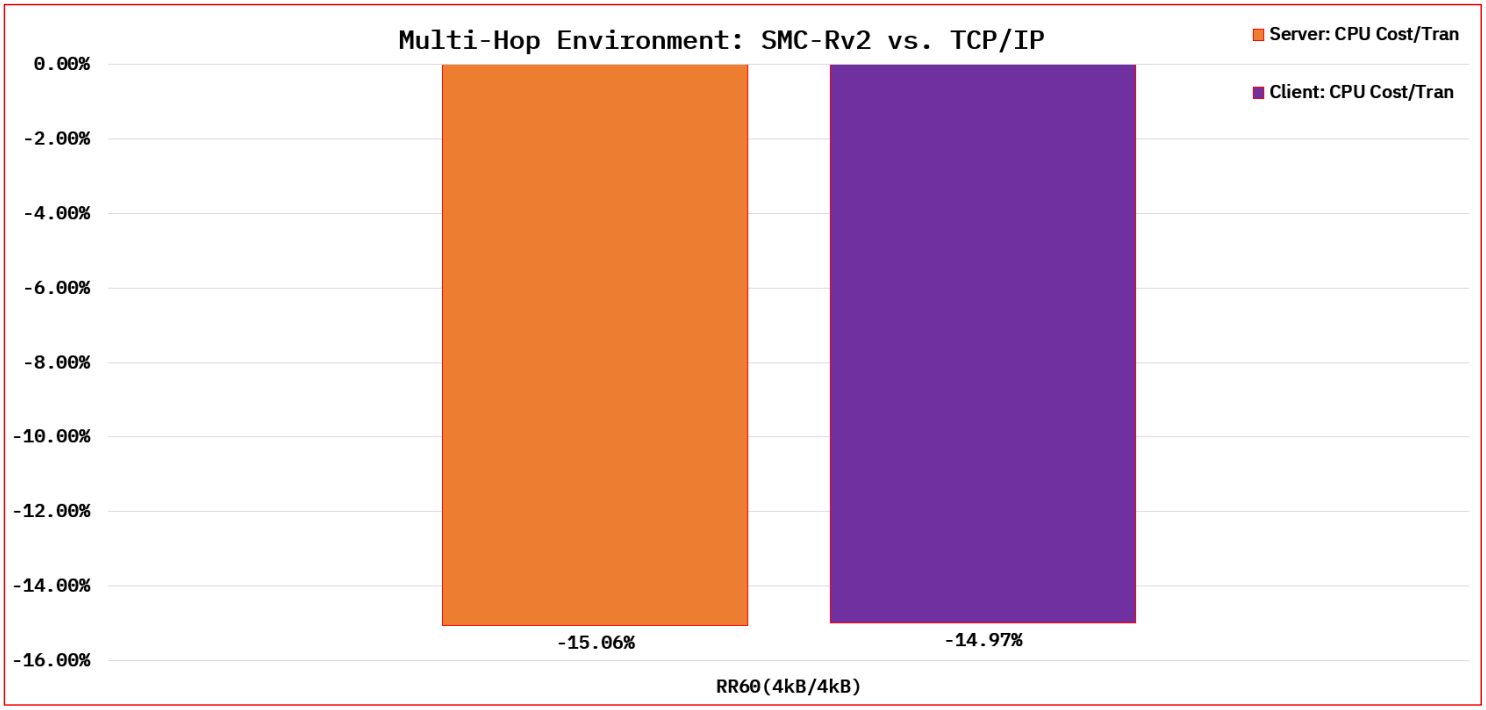


**Figure 17: SMC-Rv2 outperforming TCP/IP in a multi-hop environment for RR (i.e., OLTP) workloads**

## STR Observations

The observations for long lived flows (i.e., streaming) are the same as request response (e.g., OLTP) workloads (i.e., SMC-Rv2 outperformed TCP/IP in a multi-hop environment). For long lived flows, there is a higher CPU cost savings because the bulk data processing is completed by the RNIC. This was evident by the client's CPU cost reduction of ~ 85% and server's CPU cost reduction of ~ 51%.

## STR Result

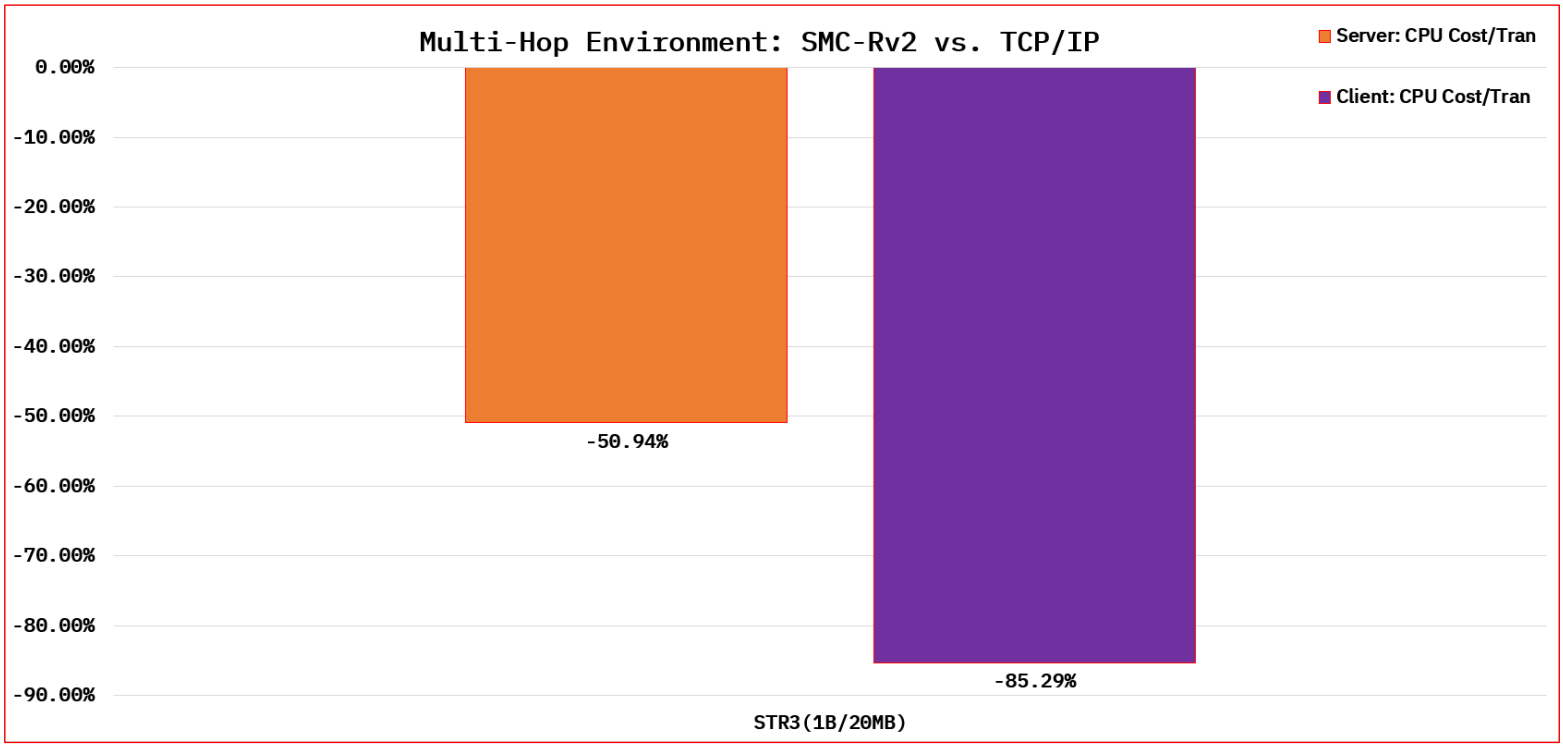


**Figure 18: SMC-Rv2 outperforming TCP/IP in a multi-hop environment for long lived (i.e., streaming) flows**

# zERT Policy-based Enforcement

## Background

The z/OS Encryption Readiness Technology (zERT) provide clients the ability to make their TCP/IP stack a central collection point for cryptographic protection attributes for TLS/SSL, IPsec, SSH, and No Recognized Protection (NRP)[3]. zERT Enforcement extends the ability by giving clients the flexibility to define policies. An end user may define a rule within a policy that describes the acceptable or unacceptable cryptographic protection attributes associated with a given TCP/IP connection. When the rule is matched, the policy allows the TCP/IP stack to take an action or multiple actions. In simple terms, zERT Enforcement enables enterprises to ensure traffic in their network environment adheres to company's policy.

## Test Environment: zERT Enforcement Actions

zERT Enforcement actions are used to act on a connection when a rule within a policy is matched. Such actions include logging to the console, logging to Syslog, writing to System Management Facility (SMF), and resetting the connection. From a performance perspective, the focus of this section of the report was in understanding the impact on network performance when a connection is evaluated against different rules then an action is taken. In our measurements, the below actions were of interest:

1. AuditRecord – Writing SMF type 119 subtype 11 records
2. LogConsole – Writing log to the console and TCP/IP joblog
3. LogSyslog – Writing log to Syslog with the highest log level (e.g., LogLevel 7)

The below figure gives a visualization of the test environment.

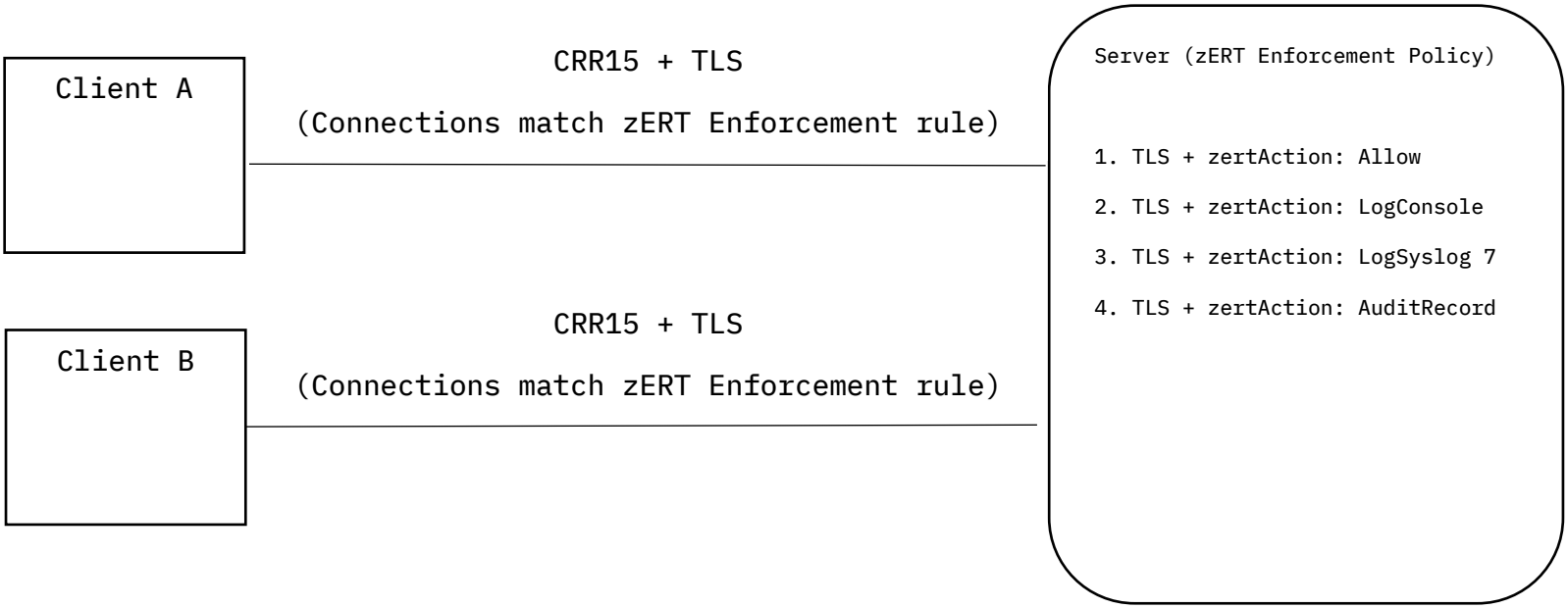


**Figure 19: Two client instance where each open 15 connections to a single server instance**

[3] TLS/SSL, SSH, IPsec, and NRP support is available for **solely** TCP-based connections

## z/OS Environment Configuration: zERT Enforcement Actions

Below is the environment configuration in which the data was collected:

- CPC: z15
- Release: V2R5
- Number of CPUs: 4 (Dedicated) per LPAR
- Interface: OSA-Express 7S 10GbE
- Workload
    - CRR30(200B/200B)

## CRR Observations

In our measurement, the server had active zERT Enforcement policies such that all incoming connections match a zERT rule. The base case consisted of not taking any action when a rule is matched whereas this was compared to other cases that consisted of taking an action. From the server's perspective, evaluating zERT Enforcement policies then taking a certain action has minimal performance impact as evident by the below figure.

## CRR Results

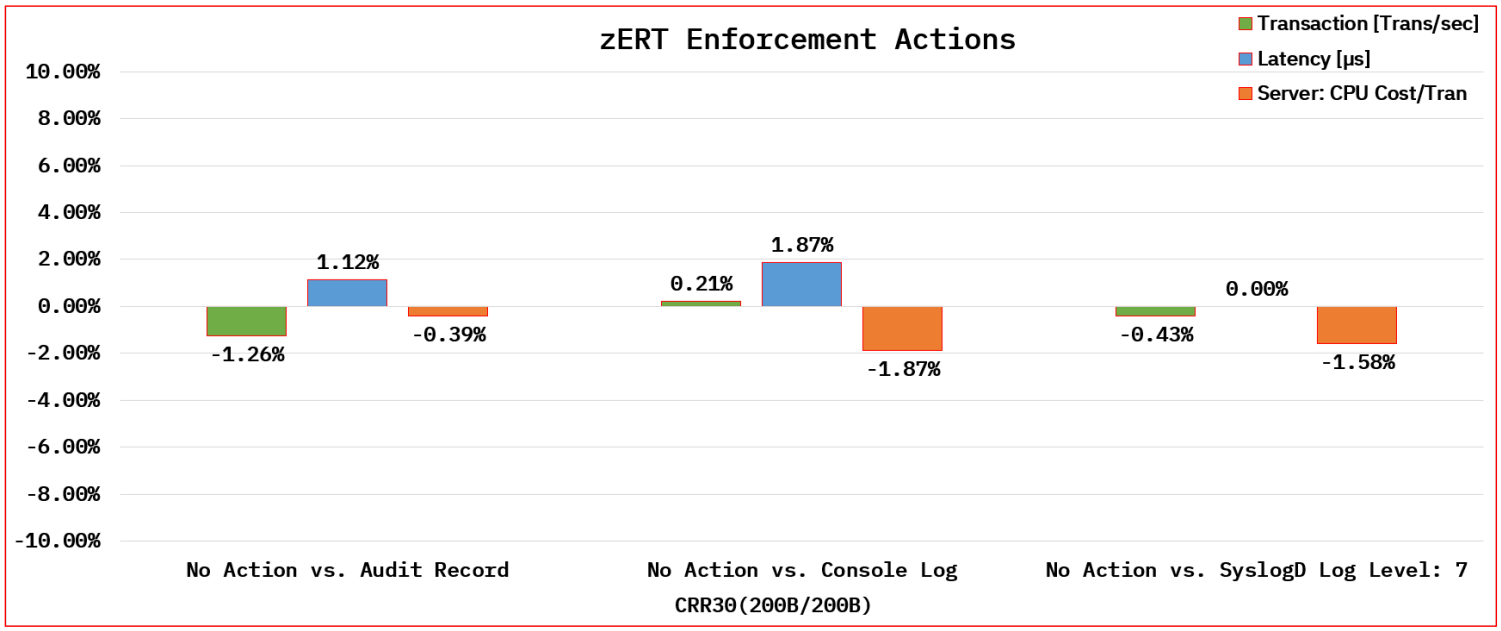


**Figure 20: A z/OS server evaluating zERT Enforcement policies do not suffer from any performance degradation**[4]

[4] The client & server were running z/OS

# V2R5: Hardware Performance

# z15: SMC-Rv1 25GbE RoCE Express3 versus Express2

## Background

RDMA over Converged Ethernet (RoCE) enables existing datacenters running on ethernet topology at layer two in taking advantage of the Remote Direct Memory Access (RDMA) feature. RDMA decreases CPU usage and latency. It decreases CPU usage because it bypasses the TCP protocol processing. It reduces latency because it has less network layers to traverse plus the processing is completed on an application specific integrated circuit (ASIC). The SMC-R protocol is based on RDMA. To use SMC-R, a RoCE network interface controller (RNIC) is required.

## Test Environment: SMC-Rv1 25GbE RoCE Express3 versus Express2

In the following sections, the focus is on comparing the newer RNIC (e.g., RoCE Express3) against the older version (e.g., RoCE Express2).

### z/OS Environment Configuration: SMC-Rv1 25GbE RoCE Express3 versus Express2

Below is the environment configuration in which the data was collected:

- CPC: z15[5]
- Release: V2R5
- Number of CPUs: 4 (Dedicated) per LPAR
- Interfaces: 25GbE RoCE Express3 and 25GbE RoCE Express2
- Workloads
    - RR1(1B/1B)
    - STR3(1B/20MB)

### RR Observation

The newer RNIC does provide a slightly better transaction rate with a lesser delay for request response workloads.

[5] The IBM RoCE Express3 family is exclusive to the IBM z16 family [6].

**RR Result**

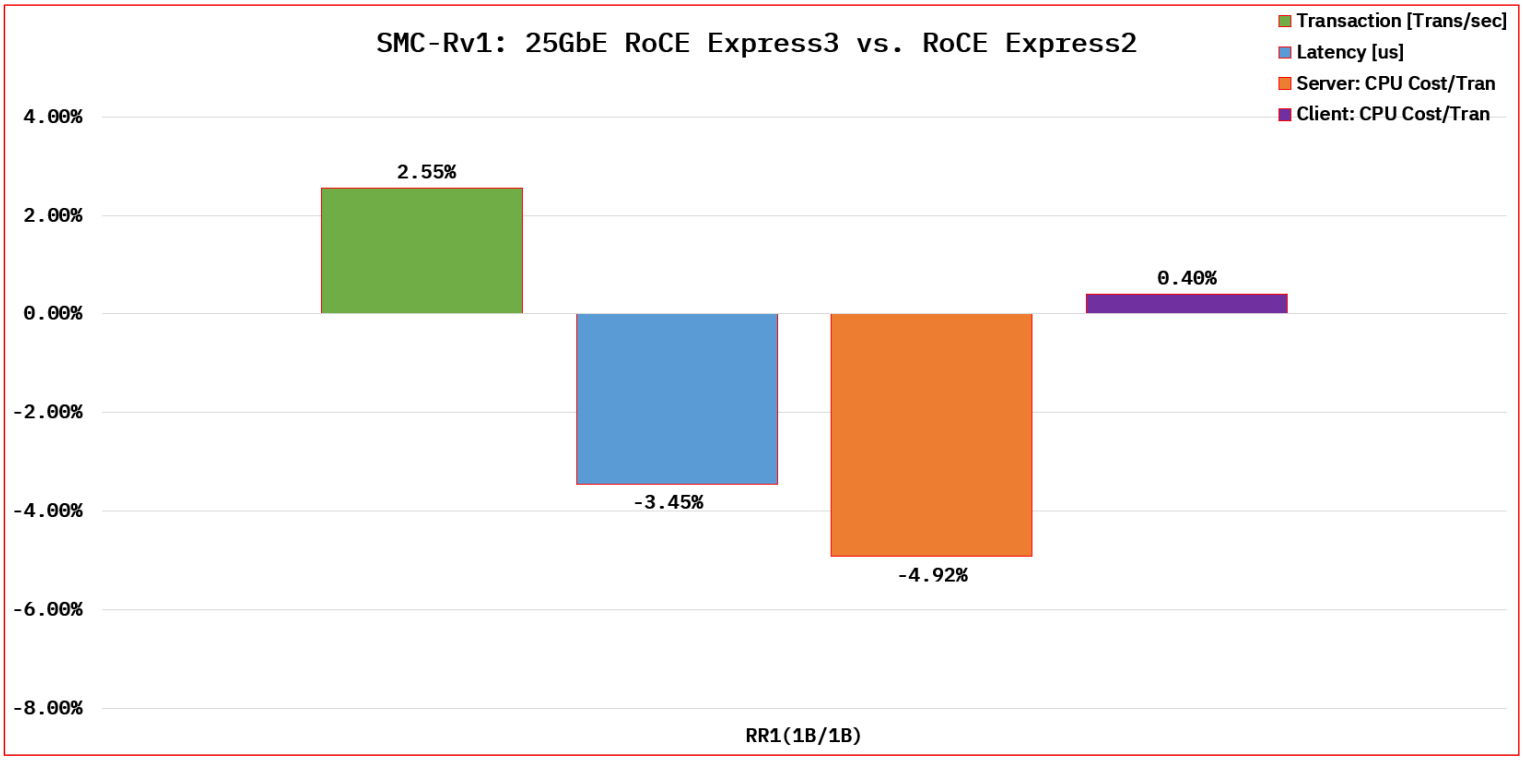


**Figure 21: 25GbE RoCE Express3 does have transaction rate and lesser delay perks in contrast to RoCE Express2**

**STR Observation**

For streaming workloads, we did not see any significant delta between the different adapters.

# SMC Applicability Tool

Many clients express interest in Shared Memory Communications (SMC). However, they are not quite sure of SMC's full potential in their environment. With expertise and significant time commitment, one can determine their environment traffic patterns that can take advantage of SMC.

SMC Applicability Tool (SMCAT) alleviates a customer's significant time commitment by monitoring and evaluating their TCP/IP network traffic. A system administrator can utilize the tool's evaluation to determine the applicability of SMC in their ecosystem. To enable SMCAT, refer here.

# AT-TLS

## Background

Application Transparent Transport Layer Security (AT-TLS) is a function within TCP/IP (i.e., stack) that allows for transparent implementation of TLS protection to TCP traffic through policies. It is essentially a System SSL wrapper that lives in the stack. The optimized integration between System SSL and AT-TLS ensures no extra overhead in the AT-TLS path versus direct calls to System SSL.

Our AT-TLS measurements focus primarily on recent cryptographic optimizations and their effect on TLSv1.2 and TLSv1.3 handshake performance. We gather AT-TLS measurements to benchmark new cipher suites and provide release-to-release comparisons.

## TLS Session Reuse: Abbreviated versus Full Handshake

Throughout this section, we use the following terms:

`Long Handshake`, which means a full TLS handshake with no TLS session reuse or session caching.

`Short Handshake`, which means an abbreviated TLS handshake that reuses attributes from a previous TLS session that was originally negotiated with a full TLS handshake. Short handshakes often require less processing than full handshakes and thus require less CPU cycles [7].

The below two sections summarize the hardware and software used for all measurements in the subsequent sections.

## z/OS Environment Configuration: Hardware

- CPC: z15
- Number of CPUs: 4 (Dedicated) per LPAR
- Interface: OSA-Express 7S 10GbE
- Cryptographic Coprocessor: Crypto Express-7S

## z/OS Environment Configuration: Software

- Release: V2R4 and V2R5
- ICSF FMID
  - HCR77D0 (V2R4)
    - PTF: UJ01386
  - HCR77D2 (V2R5)
    - PTF: UJ06231
- TLS Protocol Version: TLSv1.2, TLSv1.3
- TLSv1.2 Ciphers
  - 3C: TLS_RSA_WITH_AES_128_CBC_SHA256
  - 9C: TLS_RSA_WITH_AES_128_GCM_SHA256
  - C027: TLS_ECDHE_RSA_WITH_AES_128_CBC_SHA256
  - C02F: TLS_ECDHE_RSA_WITH_AES_128_GCM_SHA256
- TLSv1.3 Ciphers
  - 1301: TLS_AES_128_GCM_SHA256
  - 1302: TLS_AES_256_GCM_SHA384
- Server certificate with RSA 2048 bit key

## Test Environment: TLSv1.2 RSA_xxx Ciphers versus ECDHE_RSA_xxx Ciphers

In the following section, the goal was to illustrate the cost of using ephemeral elliptic curve Diffie-Hellman (TLS_ECDHE) key exchange with z15 CPACF ECC support versus fixed RSA (TLS_RSA) in terms of transaction rate, latency, and CPU cost. ECDHE cipher suites are becoming popular because they provide perfect forward secrecy, unlike fixed RSA suites.

### CRR Observations

In Elliptic Curve Diffie-Hellman Ephemeral (ECDHE) key agreement, a new key pair is generated for every long handshake (i.e., non-cache) to guarantee *perfect forward secrecy*. This provides significant security benefits, but at a cost as evident by the following results. Therefore, we recommend session caching when possible.

## CRR Results

As a reminder, here are the TLSv1.2 Ciphers:

- 3C: TLS_RSA_WITH_AES_128_CBC_SHA256
- 9C: TLS_RSA_WITH_AES_128_GCM_SHA256
- C027: TLS_ECDHE_RSA_WITH_AES_128_CBC_SHA256
- C02F: TLS_ECDHE_RSA_WITH_AES_128_GCM_SHA256

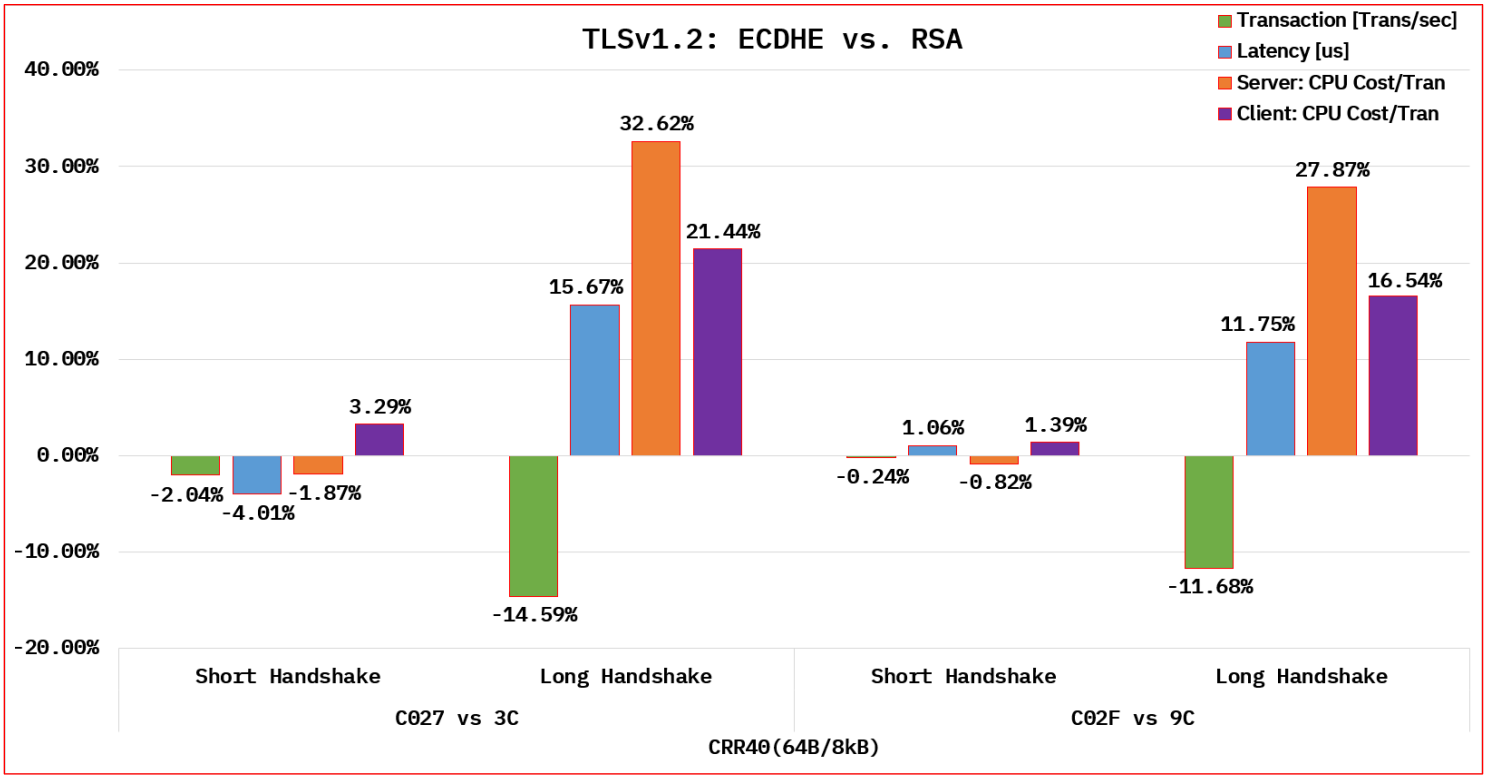


**Figure 22: Comparison of TLSv1.2 ECDHE and RSA key exchange & agreement performance**

### Test Environment: ICSF CPACF ECC Support On V2R5

In the following test scenario, the goal was to measure the impact of z15's CPACF support for Elliptic Curve Cryptography (ECC). This support significantly improves the performance of TLS_ECDHE cipher suites on V2R5 as well as on V2R4 with ICSF PTF UJ01386 applied to HCR77D1. The comparison is between V2R4 without the support and V2R5 with the support.

## CRR Observations

CPACF ECC support dramatically increases the number of handshakes per second (i.e., transaction rate) while significantly reducing CPU usage for workloads using ECDHE key exchange. Therefore, ECDHE-based handshakes become quite affordable on z15. The short handshakes also benefits, to a lesser extent, as evident by the CPU savings on the server side.

## CRR Results

As a reminder, here are the TLSv1.2 Ciphers:

- C027: TLS_ECDHE_RSA_WITH_AES_128_CBC_SHA256
- C02F: TLS_ECDHE_RSA_WITH_AES_128_GCM_SHA256

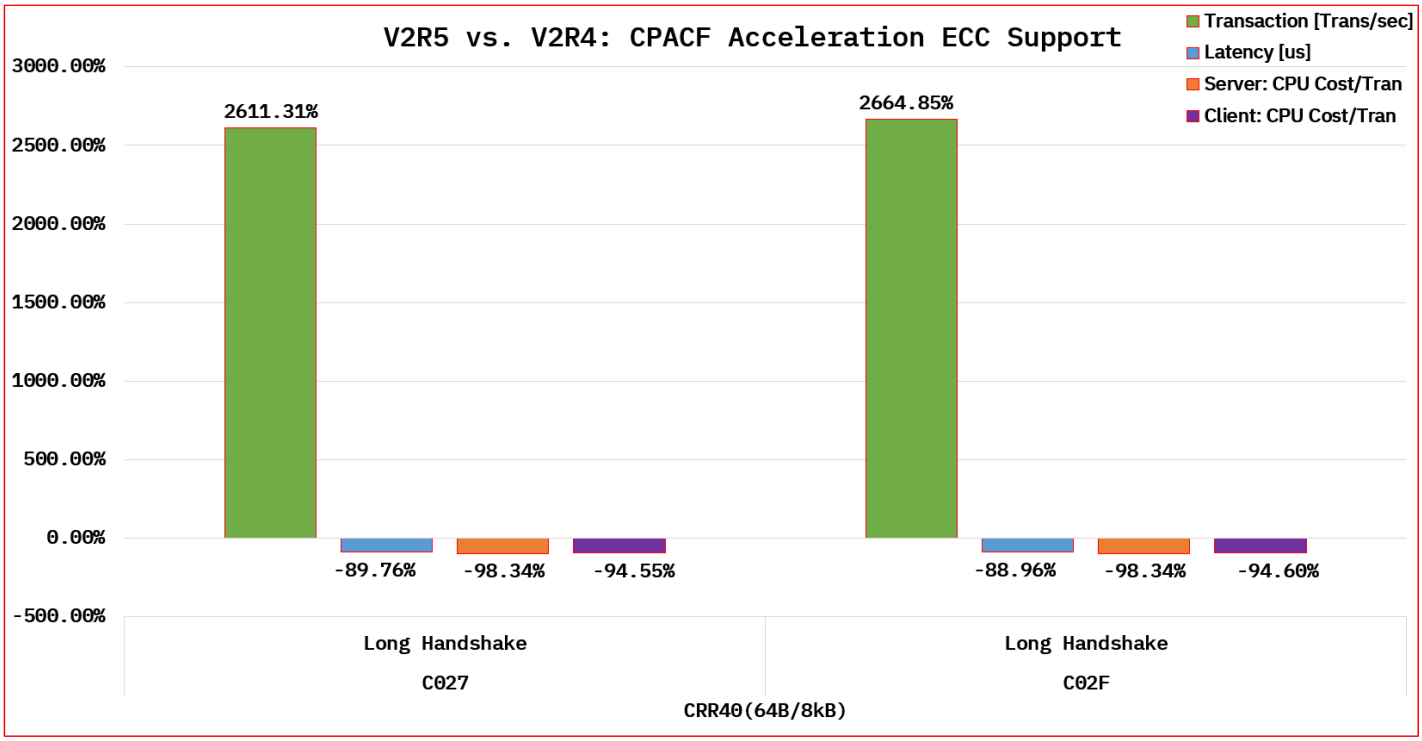


**Figure 23: Comparison of using and not using CPACF Acceleration ECC Support for ECDHE key exchange cryptographic long handshake operations**

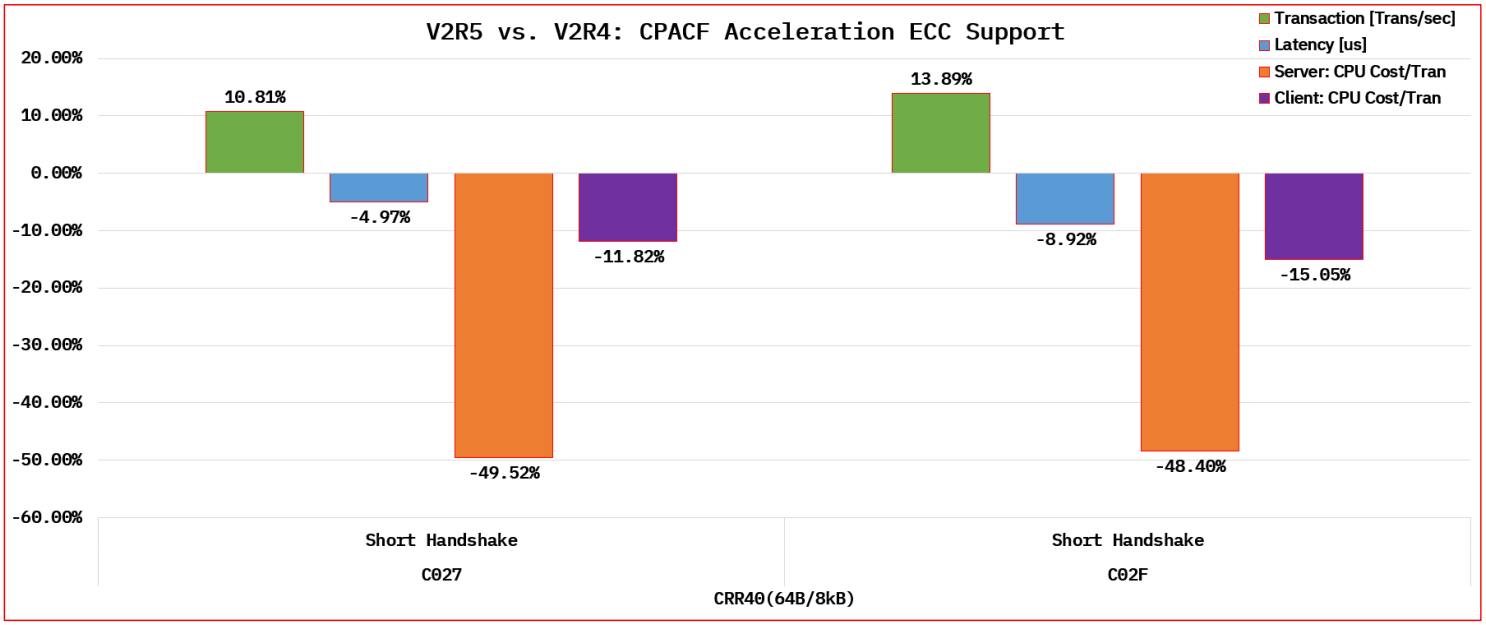


**Figure 24: Comparison of using and not using CPACF Acceleration ECC Support for ECDHE key exchange cryptographic short handshake operations**

## Test Environment: TLSv1.3 ICSF RSASSA-PSS Support (V2R5 versus V2R4)

This test focuses on measuring TLSv1.3 handshakes using ICSF feature to use Crypto Express with cleartext RSA keys. The improvements on V2R5 combines acceleration of the Elliptic Curve Cryptography (ECC) CPACF support and the RSASSA-PSS cleartext support under TLSv1.3.

### CRR Observations

The RSASSA-PSS usage with cleartext RSA keys becomes affordable from a transaction rate and CPU cost perspective. Significant performance improvement (~ 1265%) is shown more with long handshake operations. Short handshake still benefits from this support as evident by the server CPU cost savings.

### CRR Results

As a reminder, here are the TLSv1.3 Ciphers:

- 1301: TLS_AES_128_GCM_SHA256
- 1302: TLS_AES_256_GCM_SHA384

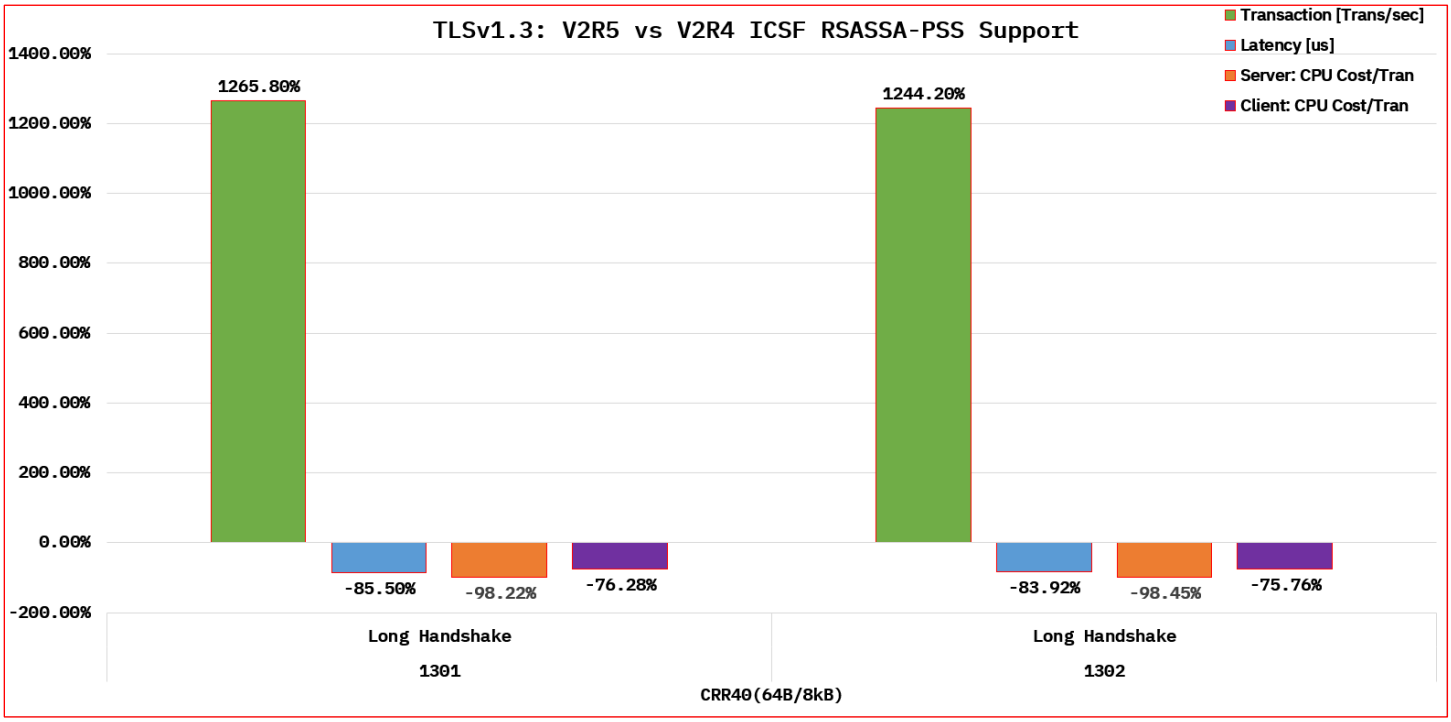


**Figure 25: TLSv1.3 ICSF RSASSA-PSS Support on V2R5 shows significant improvement in transaction rate, delay, and CPU savings for long handshakes**

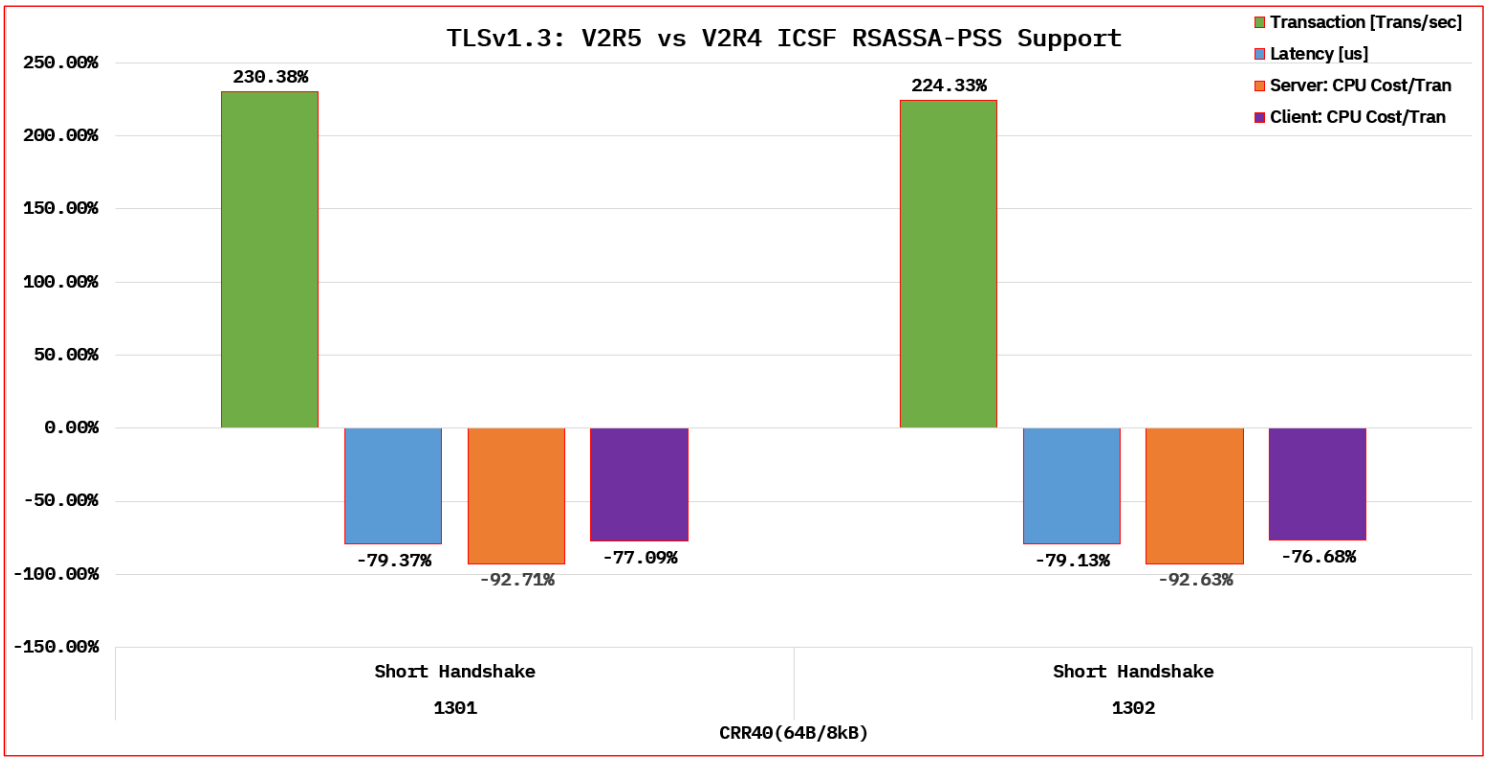


**Figure 26: TLSv1.3 ICSF RSASSA-PSS Support on V2R5 shows significant improvement in transaction rate, delay, and CPU savings for short handshakes**

# General Hardware Performance

## OSA-Express 7S 25GbE

### Background

Open Systems Adapter (OSA) Express (OSA-Express) continues to release new models with additional features and hardware updates. The adapter is a network controller that you can install in a mainframe I/O cage. It is the strategic communications device for the mainframe architecture.

### Reminder

The z/OS Communications Server team wants to remind the audience that a larger network interface controller (NIC) speed does not solely equate to higher bandwidth. It also equates to lesser delay as evident by Figure 27.

### Test Environment: OSA-Express 7S 25GbE Versus OSA-Express 7S 10GbE

In the following sections, the goal is to show how a higher bandwidth NIC offers higher throughput *plus* lesser delay.

**z/OS Environment Configuration: OSA-Express 7S 25GbE Versus OSA-Express 7S 10GbE**

Below is the environment configuration in which the data was collected:

- CPC: z15
- Release: V2R4
- Number of CPUs: 4 (Dedicated) per LPAR
- Interfaces: OSA-Express 7S 25GbE & OSA-Express 7S 10GbE
- Workloads
    - RR40(100B/100B)
    - STR3(20MB/1B)

**RR Observation (i.e., Lesser Delay)**

In addition to providing higher throughput, a higher bandwidth NIC also provides lower network latency.

## RR Result

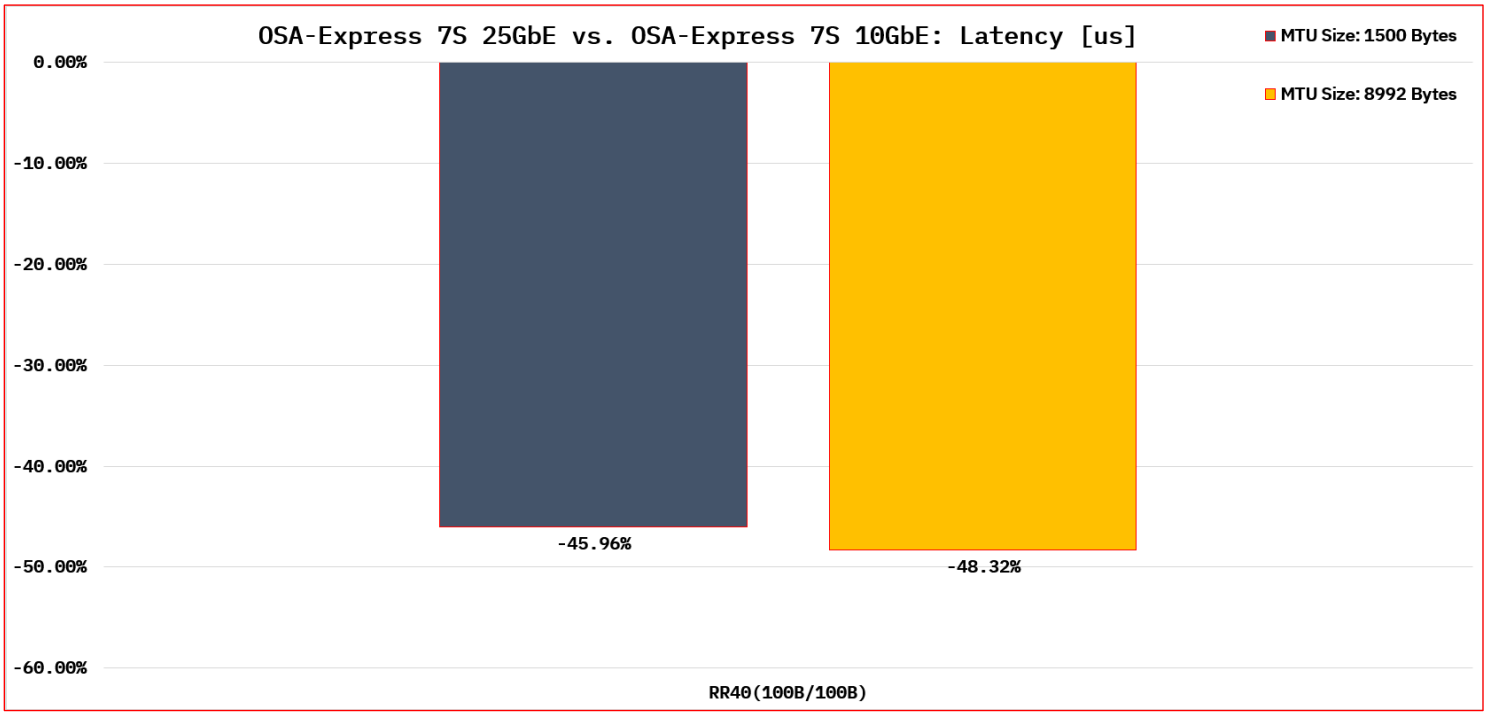


**Figure 27: OSA-Express 7S 25GbE decreases latency significantly**

### STR Observation (i.e., Higher Bandwidth)

The streaming workload allowed us to verify that OSA-Express 7S 25GbE indeed provides a higher throughput as expected.

## STR Result

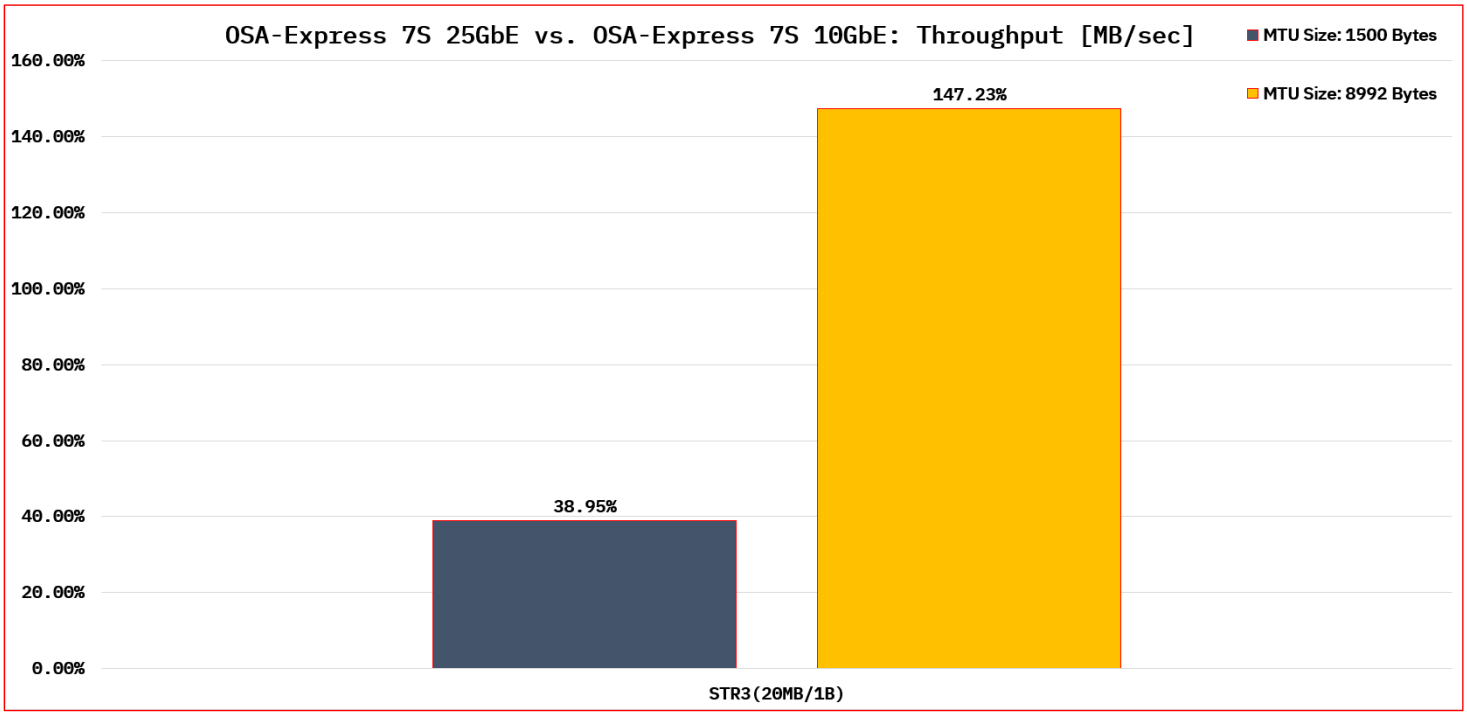


**Figure 28: OSA-Express 7S 25GbE increases throughput tremendously**

# V2R5 vs. V2R4: Release to Release Performance Comparison

## V2R5 vs. V2R4

### Introduction

In this sub-section, the pure focus was on benchmarking the latest release, V2R5, against the previous release, V2R4.

### z/OS Environment Configuration

Below is the environment configuration in which the data was collected:

- CPC: z15
- Release: V2R5 & V2R4
- Number of CPUs: 2 (Dedicated) per LPAR
- Interface: OSA-Express 7S 10GbE
- Workloads
    - RR60(4kB/4kB)
    - CRR40(64B/8kB)
    - STR3(1B/20MB)
    - STR3(20MB/1B)

### Synopsis

Performance of V2R5, which consists of new functions and improved existing functions, is on par with V2R4.

## CICS Sockets

### Background

CICS is a mixed language application server. It reduces complexity by providing APIs to developers who implements different applications consisting of interface and business logic written in different languages (e.g., Java, C, COBOL, etc.). It is a z/OS middleware that can interact with other middleware such as Db2. Intercommunication within the same CPC reduces delay. You can refer here and here for more information.

### z/OS Environment Configuration

Below is the environment configuration in which the data was collected:

- CPC: z15
- Release: V2R5 & V2R4
- Number of CPUs: 4 (Dedicated) per LPAR
- Interface: OSA-Express 7S 10GbE
- Workloads
  - RR40(100B/100B)
  - RR40(1kB/1kB)
  - CRR20(2kB/2kB)

### Synopsis

In our CICS Sockets measurements, the performance on V2R5 is level with V2R4.

## Enterprise Extender

### Background

Enterprise Extender (EE) is a standard that is documented in RFC 2353. It extends SNA High Performance Routing (HPR) traffic of any logical unit over an IP infrastructure transparently without any required infrastructure changes. Such offering makes EE attractive to end users (i.e., clients). For more information, refer here.

### z/OS Environment Configuration

Below is the environment configuration in which the data was collected:

- CPC: z15
- Release: V2R5 & V2R4
- Number of CPUs: 4 (Dedicated) per LPAR
- Interface: OSA-Express 7S 10GbE
- Workloads
  - RR20(100B/800B)
  - STR3(1B/20MB)

### Synopsis

In our EE measurements, the performance on V2R5 is level with V2R4.

# FTPD

## Background

The File Transfer Protocol (FTP) is a popular application for transferring files between computers. On z/OS, there is a FTP server application (e.g., FTPD). It is broken down into two processes: daemon and server. The daemon process idles for an incoming connection. Upon establishing a new connection, it creates a new server process. For each login session, there is a control and data connection. The former is used to exchange command request and replies whereas the latter is used to exchange data (e.g., files). You can refer here for more information.

## z/OS Environment Configuration

Below is the environment configuration in which the data was collected:

- CPC: z15
- Release: V2R5 & V2R4
- Number of CPUs: 2 (Dedicated) per LPAR
- Interface: OSA-Express 7S 25GbE
- Workloads
    - BIN: Put & Get
    - ASCII: Put & Get
    - Large MVS File Transfer

## Synopsis

In our FTPD measurements, the performance on V2R5 is level with V2R4.

# HiperSockets

## Background

HiperSockets is a hardware feature that provides high speed LPAR to LPAR communication within the same CPC. It is a processor to memory architecture rather than processor to I/O. Due to the intra-traffic characteristic, the following perks are offered at a higher level: network availability, security, simplicity, performance, and cost reduction [8].

### z/OS Environment Configuration

Below is the environment configuration in which the data was collected:

- CPC: z15
- Release: V2R5 & V2R4
- Number of CPUs: 4 (Dedicated) per LPAR
- Interface: IUTIQDIO
- Maximum Frame Size (MFS): 64 [kB]
- Workloads
  - RR60(1kB/1kB), RR60(4kB/4kB), RR60(8kB/8kB), RR60(16kB/16kB), RR60(32kB/32kB), RR60(64kB/64kB)
  - CRR40(200B/200B), CRR40(64B/8kB), CRR40(64B/32kB)
  - STR1(1B/20MB), STR3(1B/20MB), STR3(1B/1GB)
  - STR1(20MB/1B), STR3(20MB/1B), STR3(1GB/1B)

### Synopsis

In our HiperSockets measurements, the performance on V2R5 is level with V2R4.

# IPsec

### Background

Internet Protocol security (IPsec) is a collection of protocols that offers end-to-end or payload encryption [9]. The former is known as *tunnel mode* whereas the latter is known as *transport mode*. In the common scenario, IPsec is used for host-to-host communication [10]. In this common use case, tunnel mode is favored because it encrypts the IP header in addition to the payload [10]. z/OS Communications Server supports IP filtering, IPsec, and Internet Key Exchange (IKE). It supports IKEv1 and IKEv2 [11].

## z/OS Environment Configuration

Below is the environment configuration in which the data was collected:

- CPC: z15
- Release: V2R5 & V2R4
- V2R4 ICSF FMID: HCR77D0
- V2R5 ICSF FMID: HCR77D2
- Number of CPUs: 4 (Dedicated) per LPAR
- Interface: OSA-Express 7S 10GbE
- Cryptographic Coprocessor: Crypto Express-6S & Crypto Express-7S
- Tunnel Mode
- MTU: 8992 Bytes
- Encryption: AES_CBC KeyLength 128, AES_GCM_16 KeyLength 128
- Authentication: HMAC_SHA2_256_128, ESP Null
- Workloads
    - RR60(4kB/4kB)
    - CRR40(64B/8kB)

## Test Environment: V2R4

On V2R4, performance metrics were gathered on the following combinations:

1. Plaintext
2. IPsec + AES_CBC + HCR77D0 + Crypto Express-6S
3. IPsec + AES_GCM + HCR77D0 + Crypto Express-6S

## Test Environment: V2R5

On V2R5, performance metrics were gathered on the following combinations:

1. Plaintext
2. IPsec + AES_CBC + HCR77D2 + Crypto Express-7S
3. IPsec + AES_GCM + HCR77D2 + Crypto Express-7S

### Synopsis

In comparison to V2R4, the performance was equivalent based on the following table:

| Test Case | V2R4 | V2R5 | V2R5 vs. V2R4 |
|---|---|---|---|
| #1 | Plaintext | Plaintext | Equivalent |
| #2 | IPsec + AES_CBC + HCR77D0 + Crypto Express-6S | IPsec + AES_CBC + HCR77D2 + Crypto Express-7S | Equivalent |
| #3 | IPsec + AES_GCM + HCR77D0 + Crypto Express-6S | IPsec + AES_GCM + HCR77D2 + Crypto Express-7S | Equivalent |

# TN3270E

### Background

TN3270 Enhanced (TN3270E) is a Telnet server enabling users to remotely access their host application. It provides access to z/OS VTAM SNA applications on the z/OS host. For more information, refer here.

### z/OS Environment Configuration

Below is the environment configuration in which the data was collected:

- CPC: z15
- Release: V2R5 & V2R4
- Number of CPUs: 2 (Dedicated) per LPAR
- Interface: OSA-Express 7S 10GbE

### Synopsis

In our TN3270E measurements, the performance on V2R5 is level with V2R4.

# References

## z/OS Communications Server Performance Index

The following index contains all z/OS Communications Server Performance related publications. The posted materials are updated as necessary.

URL: https://www.ibm.com/support/pages/node/317829

## Additional References

[1] C. Rufus, *IBM z/OS V2R1 Communications Server TCP/IP Implementation Volume 3: High Availability, Scalability, and Performance,* 1st ed. USA: IBM, 2013, pp. 292 - 293

[2] "*QDIO inbound workload queueing",* IBM. Accessed On: Mar. 31, 2020. [Online]. Available: Here

[3] D. Herr, *Getting the most out of your OSA (Open Systems Adapter) with z/OS Comm Server*. RTP: IBM, 2013, [Online]. Available: share.confex.com/share/121/webprogram/Handout/Session13222/SHARE%20OSA_Boston.pdf.

[4] B. White, O. Ferreira, T. Missawa, T. Sudewo, *IBM z/OS V2R2 Communications Server TCP/IP Implementation: Volume 3 High Availability, Scalability, and Performance.* USA: IBM, 2016, pp: 305.

[5] "*Fragmentation consideration*", IBM. Accessed On: Mar. 31, 2020. [Online]. Available: Here

[6] "*IBM z16 puts innovation to work while unlocking the potential of your hybrid cloud transformation",* IBM. Accessed on: April 22nd, 2022. [Online]. Available: Here

[7] "What is AT-TLS session caching?" IBM. Access On: Mar. 15, 2022. [Online]. Available: Here

[8] "What is a HiperSocket?" IBM. Access On: Mar. 15, 2022. [Online]. Available: Here

[9] "What is IPsec? | How IPsec VPNs work." Cloudflare. Access On: July 6, 2021. [Online]. Available: Here

[10] Palo Alto Networks LIVECommunity, What is IPSec? (Sep. 2nd, 2016). Accessed: July 6, 2021. [Online Video]. Available: Here

[11] "Overview of using IP security." IBM. Accessed: July 6, 2021. [Online]. Available: Here