



# OFA Developer Workshop 2014

## Shared Memory Communications over RDMA (SMC-R): Update

Jerry Stevens IBM  
sjerry@us.ibm.com



# Trademarks, copyrights and disclaimers



- 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 other IBM trademarks is available on the web at "[Copyright and trademark information](http://www.ibm.com/legal/copytrade.shtml)" at <http://www.ibm.com/legal/copytrade.shtml>
- Other company, product, or service names may be trademarks or service marks of others.
- THE INFORMATION CONTAINED IN THIS PRESENTATION IS PROVIDED FOR INFORMATIONAL PURPOSES ONLY. WHILE EFFORTS WERE MADE TO VERIFY THE COMPLETENESS AND ACCURACY OF THE INFORMATION CONTAINED IN THIS PRESENTATION, IT IS PROVIDED "AS IS" WITHOUT WARRANTY OF ANY KIND, EXPRESS OR IMPLIED. IN ADDITION, THIS INFORMATION IS BASED ON IBM'S CURRENT PRODUCT PLANS AND STRATEGY, WHICH ARE SUBJECT TO CHANGE BY IBM WITHOUT NOTICE. IBM SHALL NOT BE RESPONSIBLE FOR ANY DAMAGES ARISING OUT OF THE USE OF, OR OTHERWISE RELATED TO, THIS PRESENTATION OR ANY OTHER DOCUMENTATION. NOTHING CONTAINED IN THIS PRESENTATION IS INTENDED TO, NOR SHALL HAVE THE EFFECT OF, CREATING ANY WARRANTIES OR REPRESENTATIONS FROM IBM (OR ITS SUPPLIERS OR LICENSORS), OR ALTERING THE TERMS AND CONDITIONS OF ANY AGREEMENT OR LICENSE GOVERNING THE USE OF IBM PRODUCTS OR SOFTWARE.
- © Copyright International Business Machines Corporation 2013. All rights reserved.

# Agenda Topics

---

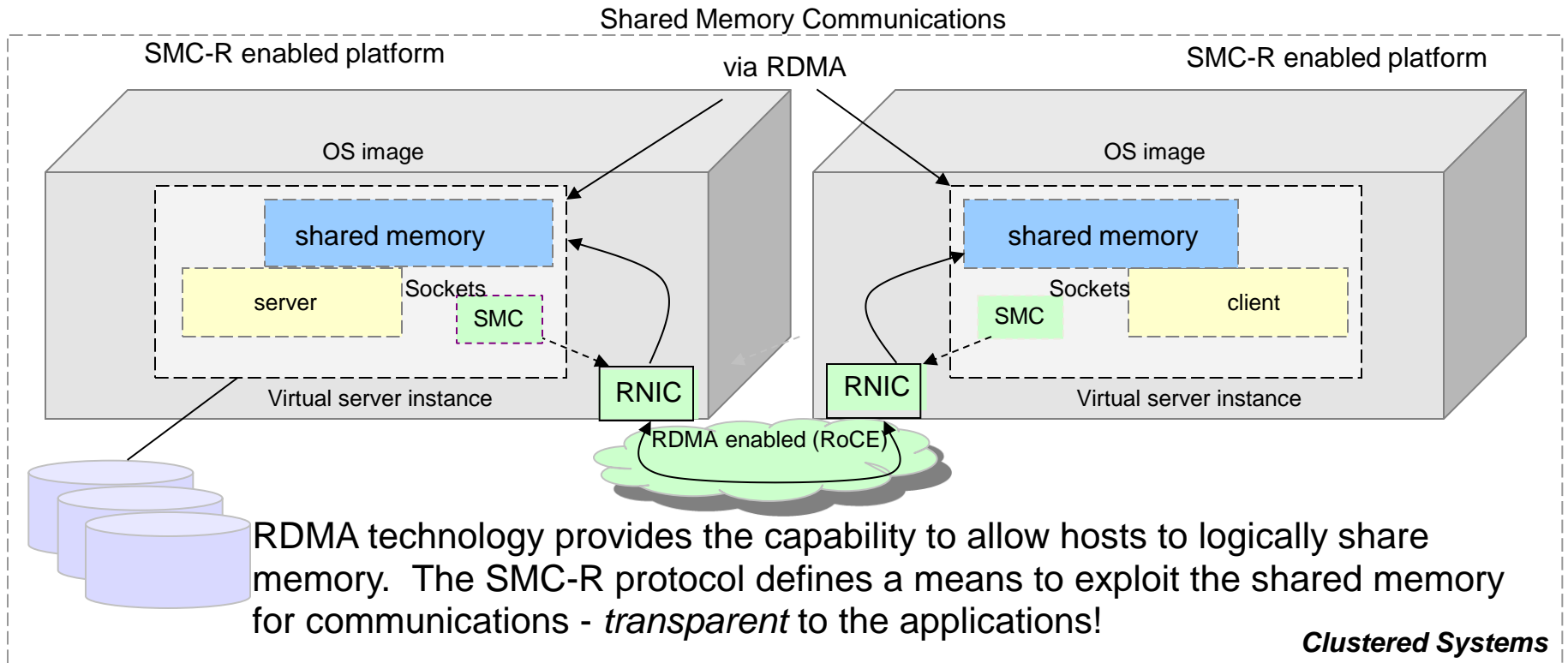
1. SMC-R Review
2. Status (what's available)
3. Performance Overview
4. Linux SMC-R Status

# Topic 1 SMC-R Review

---



# Shared Memory Communications over RDMA Concepts / Overview

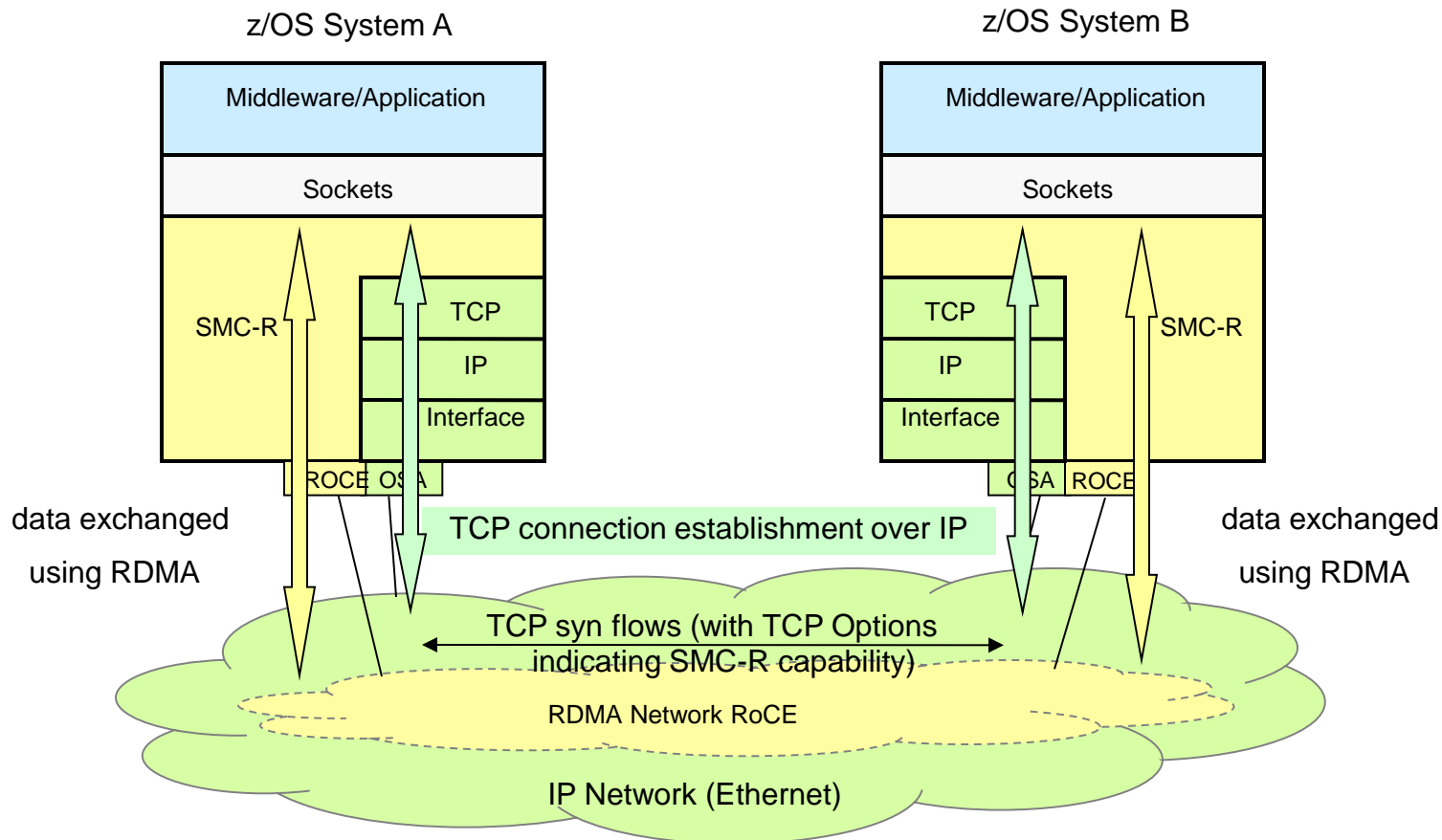


SMC-R is an *open sockets over RDMA* protocol that provides transparent exploitation of RDMA (for TCP based applications) while preserving key functions and qualities of service from the TCP/IP ecosystem that enterprise level servers/network depend on!

Draft IETF RFC for SMC-R:

<http://tools.ietf.org/html/draft-fox-tcpm-shared-memory-rdma-03>

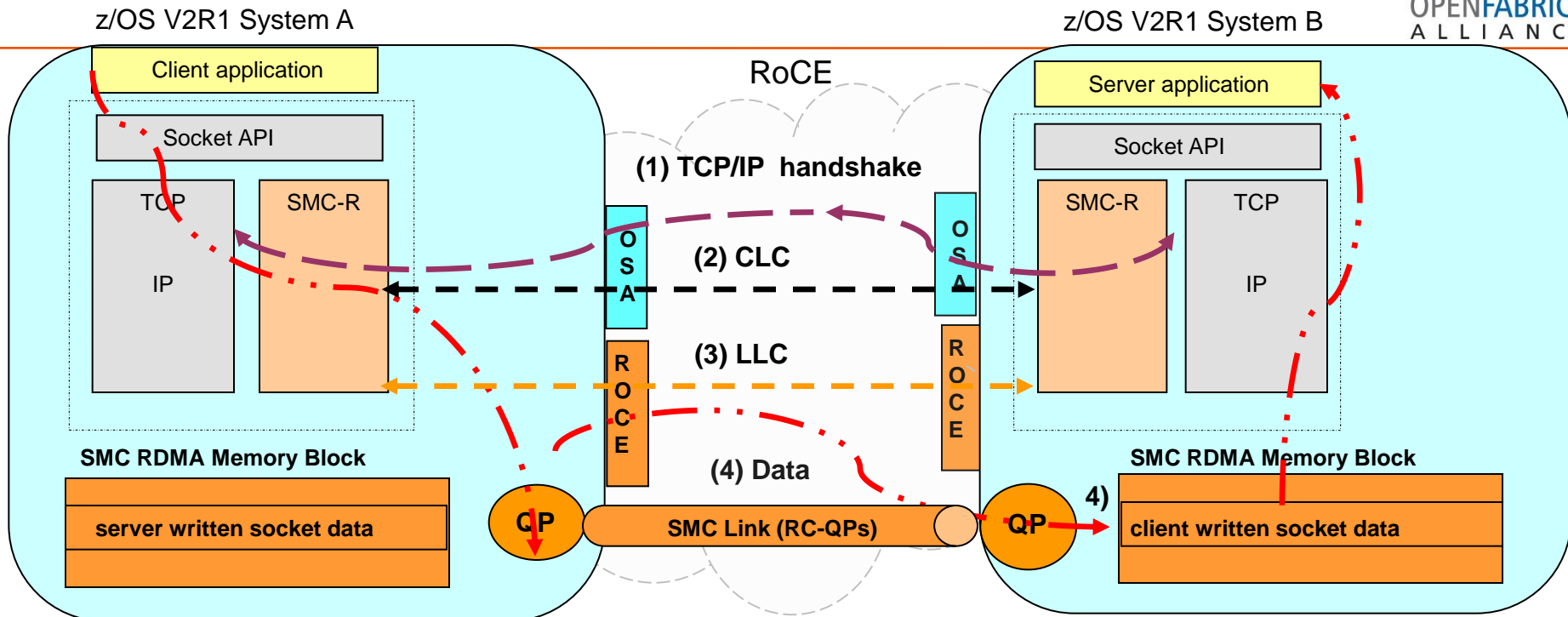
# Dynamic Transition from TCP to SMC-R



Dynamic (in-line) negotiation for SMC-R is initiated by presence of TCP Options

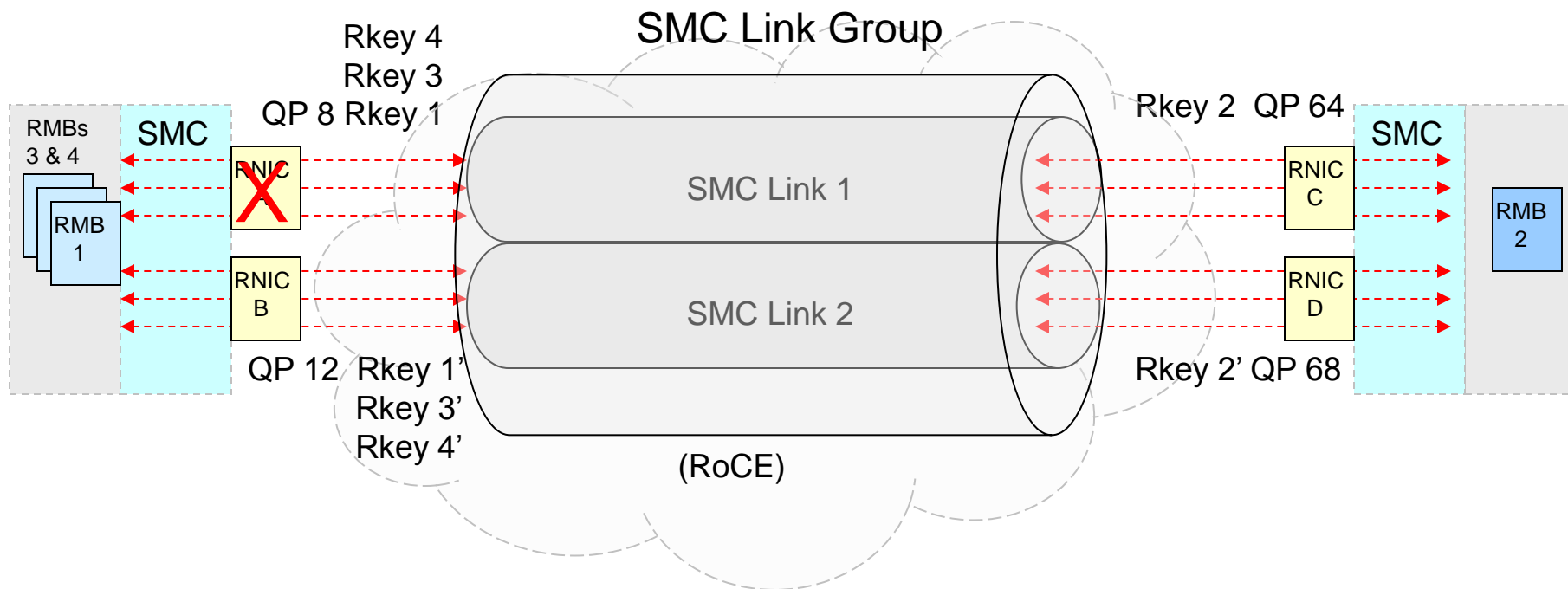
TCP connection transitions to SMC-R allowing application data to be exchanged using RDMA

# SMC-R (Contact and RDMA Processing)



- 1) Application issues standard TCP Connect; Normal TCP/IP connection (3-way syn) handshake; Determine ability/desire to support SMC-Remote (based on TCP option)
- 2) When both hosts provide SMC TCP option then exchange RDMA credentials (QPs, RMBEs, GIDs, etc.) within TCP data stream (CLC messages – Connection Level Control messages) ... can still fall back to IP
- 3) **If first contact**... then establish point-to-point SMC Link via SMC LLC (Link Layer Control) commands (RDMA-Memory-Block (RMB) pair over RC-QP... the same link (QP/RMB) can be used for multiple TCP connections across same 2 peers)
- 4) Applications issue standard socket send; SMC-R performs RDMA-write into partner's RMBE slot (RMB Element); peer consumes data via standard socket read

# SMC-R Link Group Architecture



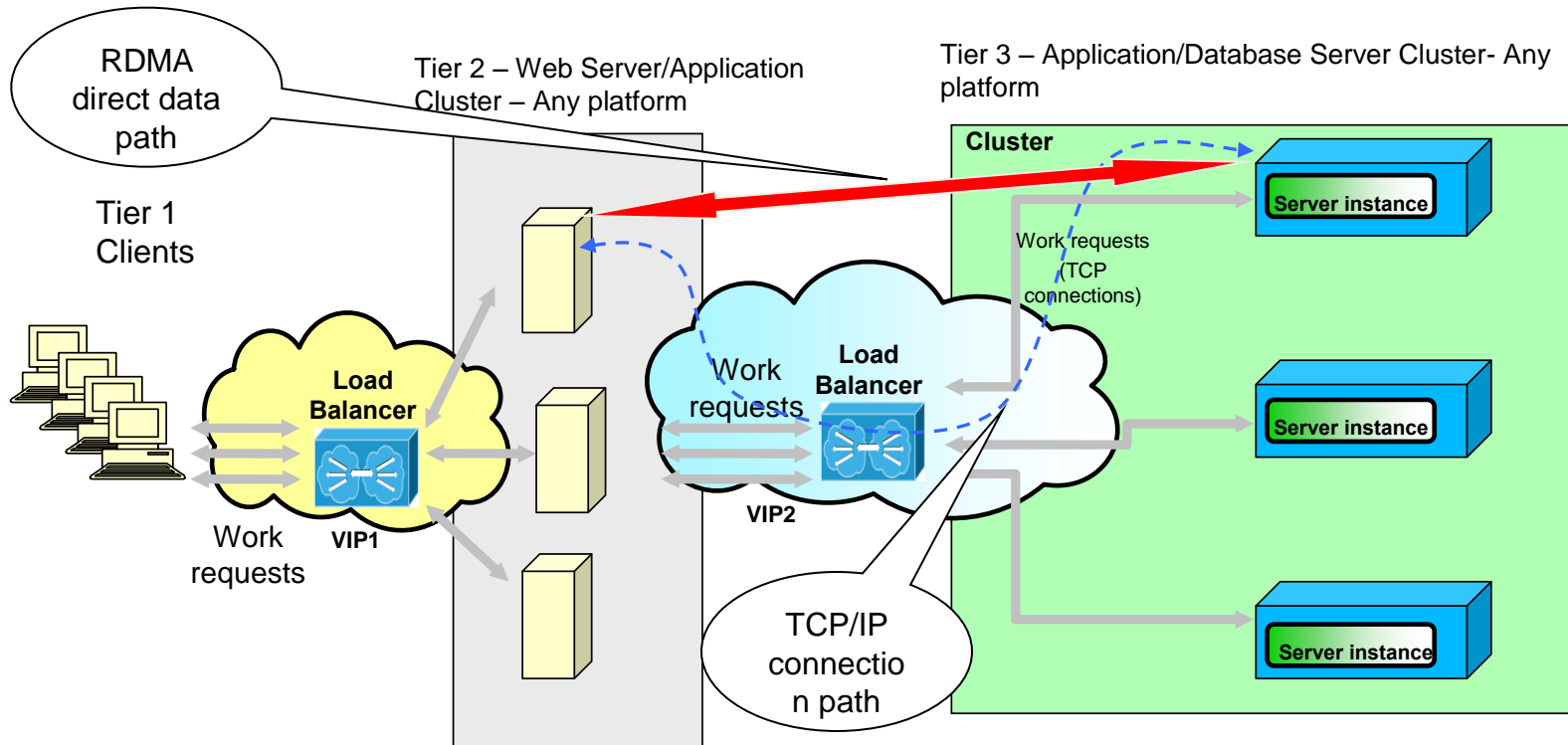
If one path (e.g. an RNIC) becomes unavailable (in this example RNIC A)... then:

- traffic on the SMC Link 1 is transparently moved to SMC Link 2 using the redundant hardware
- all application workload RDMA traffic continues without interruption... once SMC Link 1 is recovered then traffic can resume using both paths.

Note that all paths (SMC Links) have equal access to all RMBs!



# TCP Connection Load Balancing with SMC-R



- Server clustering is a prevalent deployment pattern for Enterprise class servers
  - Provide High Available, eliminate single points of failure, ability to grow/shrink capacity dynamically, ability to perform non-disruptive planned maintenance, etc.
- TCP connection load balancing is a key solution for load balancing within a cluster environment
  - External or Internal load balancers provide this capability
- Existing TCP connection load balancing solutions are **not compatible with other RDMA solutions**
  - They are not aware of the RDMA protocol **AND** RDMA flows **can not** flow through intermediate nodes
- The SMC-R protocol allows existing TCP load balancing solutions to be deployed **with no changes**
  - TCP Connection load balancing for SMC-R connections is actually more efficient than normal TCP/IP connections
    - Load balancer selects optimal back end server, data flows can then bypass the load balancer

# SMC-R Overview



- Shared Memory Communications over RDMA (SMC-R) is a protocol that allows *TCP sockets* applications to transparently exploit RDMA (RoCE)
- SMC-R is a “hybrid” solution that:
  - Uses TCP connection (3-way handshake) to establish SMC-R connection
  - Each TCP end point exchanges TCP options that indicate whether it supports the SMC-R protocol
  - SMC-R “rendezvous” (RDMA attributes) information is then exchanged within the TCP data stream (similar to SSL handshake)
  - Socket application data is exchanged via RDMA (write operations)
  - TCP connection remains active (controls SMC-R connection)
  - This model preserves many critical existing operational and network management features of TCP/IP

# SMC-R Key Attributes - Summary




- ✓ Optimized Network Performance (leveraging RDMA technology)
- ✓ Transparent to (TCP socket based) application software
- ✓ Leverages existing Ethernet infrastructure (RoCE)
- ✓ Preserves existing network security model
- ✓ Resiliency (dynamic failover to redundant hardware)
- ✓ Transparent to Load Balancers
- ✓ Preserves existing IP topology and network administrative and operational model

# Topic 2 SMC-R Availability

---



# New innovations available on IBM zBC12 and zEC12



<b>NEW</b>	<b>NEW</b>	<b>ENHANCED</b>	<b>ENHANCED</b>	<b>NEW</b>
<b>Data Compression Acceleration</b>	<b>High Speed Communication Fabric</b>	<b>Flash Technology Exploitation</b>	<b>Proactive Systems Health Analytics</b>	<b>Hybrid Computing Enhancements</b>
Reduce CP consumption, free up storage & speed cross platform data exchange	Optimize server to server networking with reduced latency and lower CPU overhead	Improve availability and performance during critical workload transitions, now with dynamic reconfiguration; Coupling Facility exploitation (SOD)	Increase availability by detecting unusual application or system behaviors for faster problem resolution before they disrupt business	x86 blade resource optimization; New alert & notification for blade virtual servers; Latest x86 OS support; Expanding futures roadmap
<i>zEDC Express</i>	<i>10GbE RoCE Express</i>	<i>IBM Flash Express</i>	<i>IBM zAware</i>	<i>zBX Mod 003; zManager Automate; Ensemble Availability Manager; DataPower Virtual appliance SoD</i>

# 10GbE RoCE Express with SMC-R: Transparent optimized server to server networking!

**Network latency** for z/OS TCP/IP based OLTP workloads **reduced** by up to **80%**<sup>1</sup>

**Networking related CPU consumption** for z/OS TCP/IP based workloads with streaming data patterns **reduced** by up to **60%**<sup>2</sup> with a **network throughput** increase of up to **60%**<sup>2</sup>



## **Shared Memory Communications (SMC-R):**

Exploit RDMA over Converged Ethernet (RoCE) to deliver superior communications performance for TCP based applications

## **Typical Client Use Cases:**

Help to reduce both latency and CPU resource consumption over traditional TCP/IP for communications across z/OS systems

**Any** z/OS TCP sockets based workload can **seamlessly** use SMC-R without requiring any application changes

**NEW** z/OS V2.1  
SMC-R

**NEW** z/VM 6.3 support  
for guests

**NEW** 10GbE RoCE  
Express

<sup>1</sup> Based on internal IBM benchmarks in a controlled environment of modeled z/OS TCP sockets-based workloads with request/response traffic patterns using SMC-R (10GbE RoCE Express feature) vs TCP/IP (10GbE OSA Express feature). The actual response times and CPU savings any user will experience will vary.

<sup>2</sup> Based on internal IBM benchmarks in a controlled environment of modeled z/OS TCP sockets-based workloads with streaming traffic patterns using SMC-R (10GbE RoCE Express feature) vs TCP/IP (10GbE OSA Express feature). The actual response times and CPU savings any user will experience will vary.

# Topic 3 SMC-R Performance

---

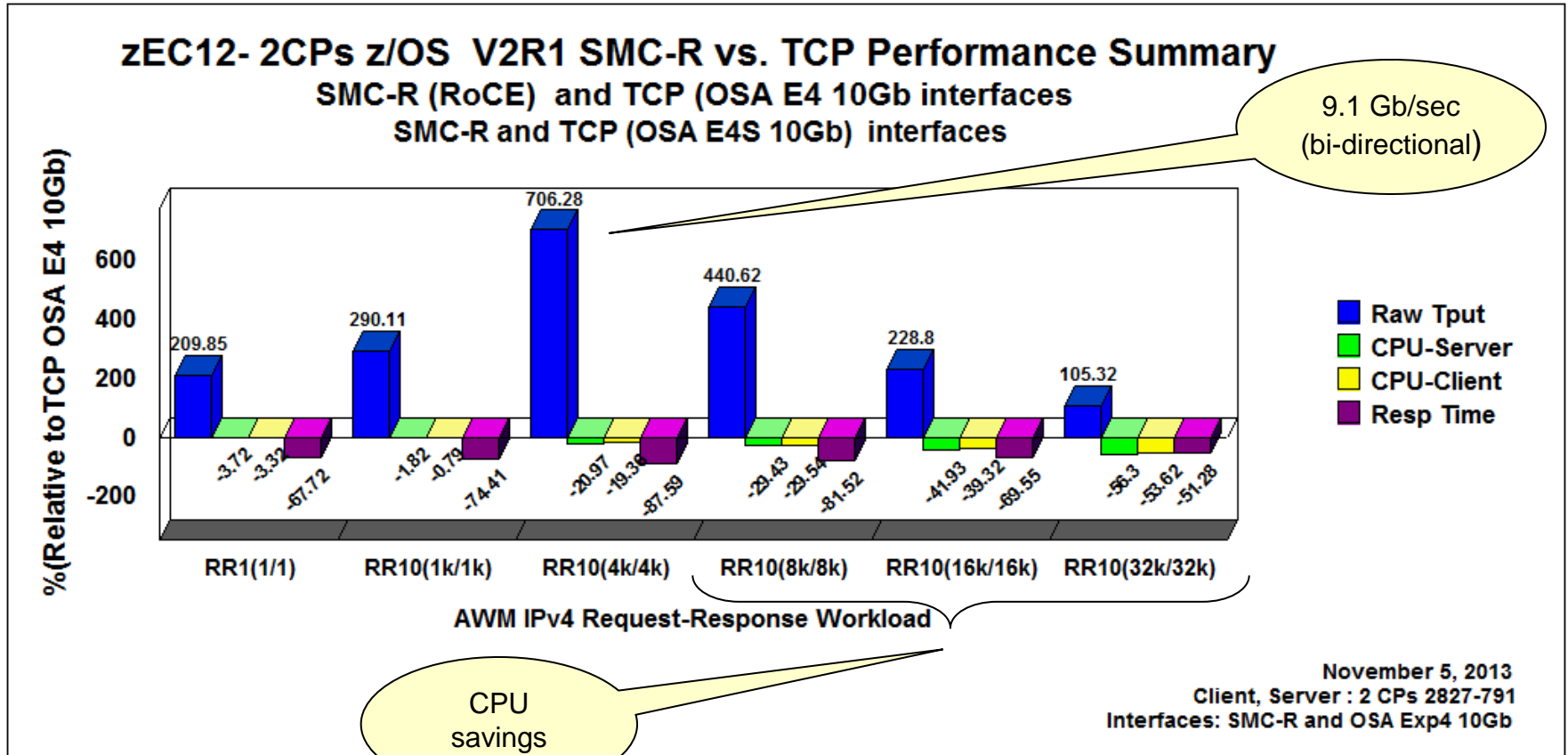


# z/OS to z/OS SMC-R Performance

(micro benchmarks)



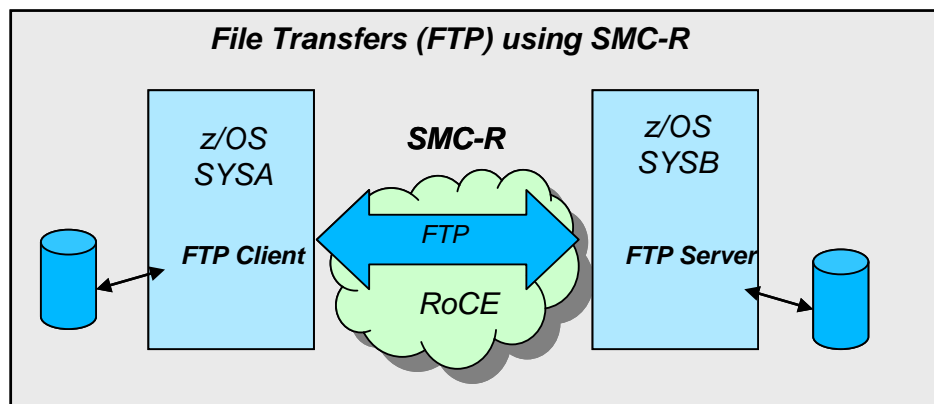
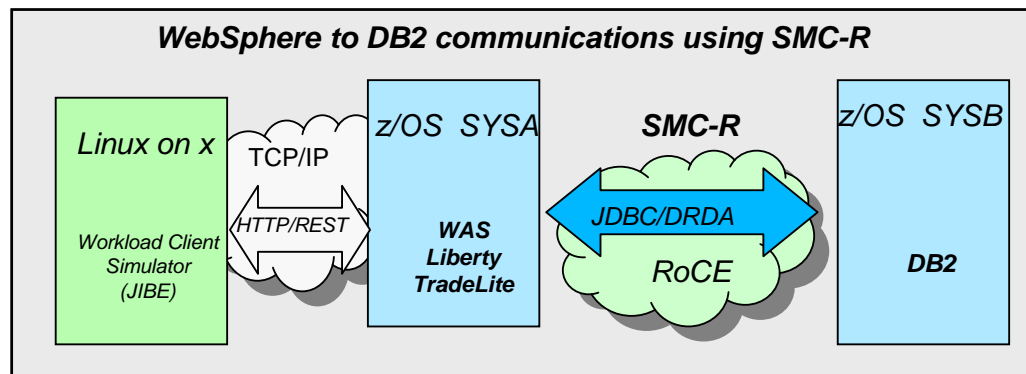
OPENFABRICS  
ALLIANCE





# Performance Benefits: Actual Workloads

**40% reduction in overall transaction response time** for WebSphere Application Server v8.5 Liberty profile TradeLite workload accessing z/OS DB2 in another system measured in internal benchmarks <sup>1</sup>



Up to **50% CPU savings** for FTP binary file transfers across z/OS systems when using SMC-R vs standard TCP/IP <sup>2</sup>

1. Based on projections and measurements completed in a controlled environment. Results may vary by customer based on individual workload, configuration and software levels.  
 2. Based on internal IBM benchmarks in a controlled environment using z/OS V2R1 Communications Server FTP client and FTP server, transferring a 1.2GB binary file using SMC-R (10GbE RoCE Express feature) vs standard TCP/IP (10GbE OSA Express4 feature). The actual CPU savings any user will experience may vary.

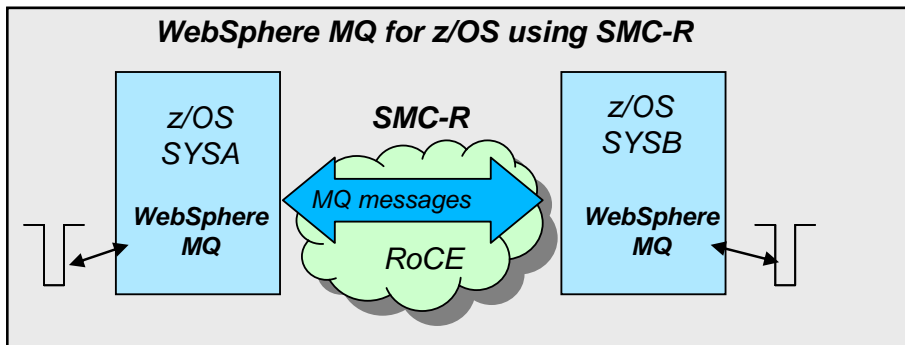
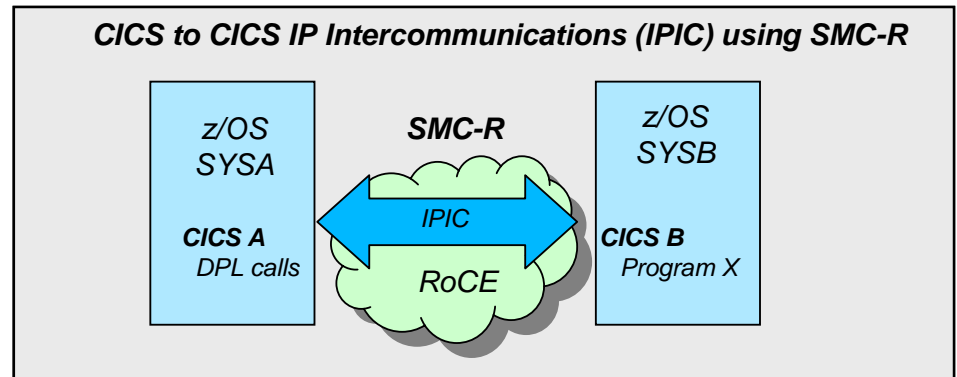
# Performance Benefits: Actual Workloads

(continued)



OPENFABRICS  
ALLIANCE

**Up to 48% reduction in response time and up to 10% CPU savings** for CICS transactions using DPL (Distributed Program Link) to invoke programs in remote CICS regions in another z/OS system via CICS IP interconnectivity (IPIC) when using SMC-R vs standard TCP/IP <sup>3</sup>



WebSphere MQ for z/OS **realizes up to 200% increase in messages per second** it can deliver across z/OS systems when using SMC-R vs standard TCP/IP <sup>4</sup>

<sup>3</sup> Based on internal IBM benchmarks using a modeled CICS workload driving a CICS transaction that performs 5 DPL (Distributed Program Link) calls to a CICS region on a remote z/OS system via CICS IP interconnectivity (IPIC), using 32K input/output containers. Response times and CPU savings measured on z/OS system initiating the DPL calls. The actual response times and CPU savings any user will experience will vary.

<sup>4</sup> Based on internal IBM benchmarks using a modeled WebSphere MQ for z/OS workload driving non-persistent messages across z/OS systems in a request/response pattern. The benchmarks included various data sizes and number of channel pairs. The actual throughput and CPU savings users will experience may vary based on the user workload and configuration.

# Topic 4 Linux SMC-R Status

---



# SMC-R for Linux: Overview



- IBM is in the process of developing an SMC-R solution for Linux
- kernel based solution (working towards upstream kernel acceptance)
- No special license requirements (GPL)
- Introduces a new AF\_SMC socket family and a preload library to transparently run AF\_INET socket applications for SMC-R (no application changes required)

# SMC-R for Linux: Key Functions

- Uses BSD sockets interface
- Conforms to SMC-R specifications (RFC)
- Provide key QoS (SSL, load balancer compatibility, HCA fail-over etc.)
- Kernel solution enables key functions (e.g. QP sharing, memory mgt, fork(), link groups etc.)
- Will be Java enabled
- Usability: minimal configuration steps required (i.e. zero IP topology changes)

# SMC-R for Linux: Project Status



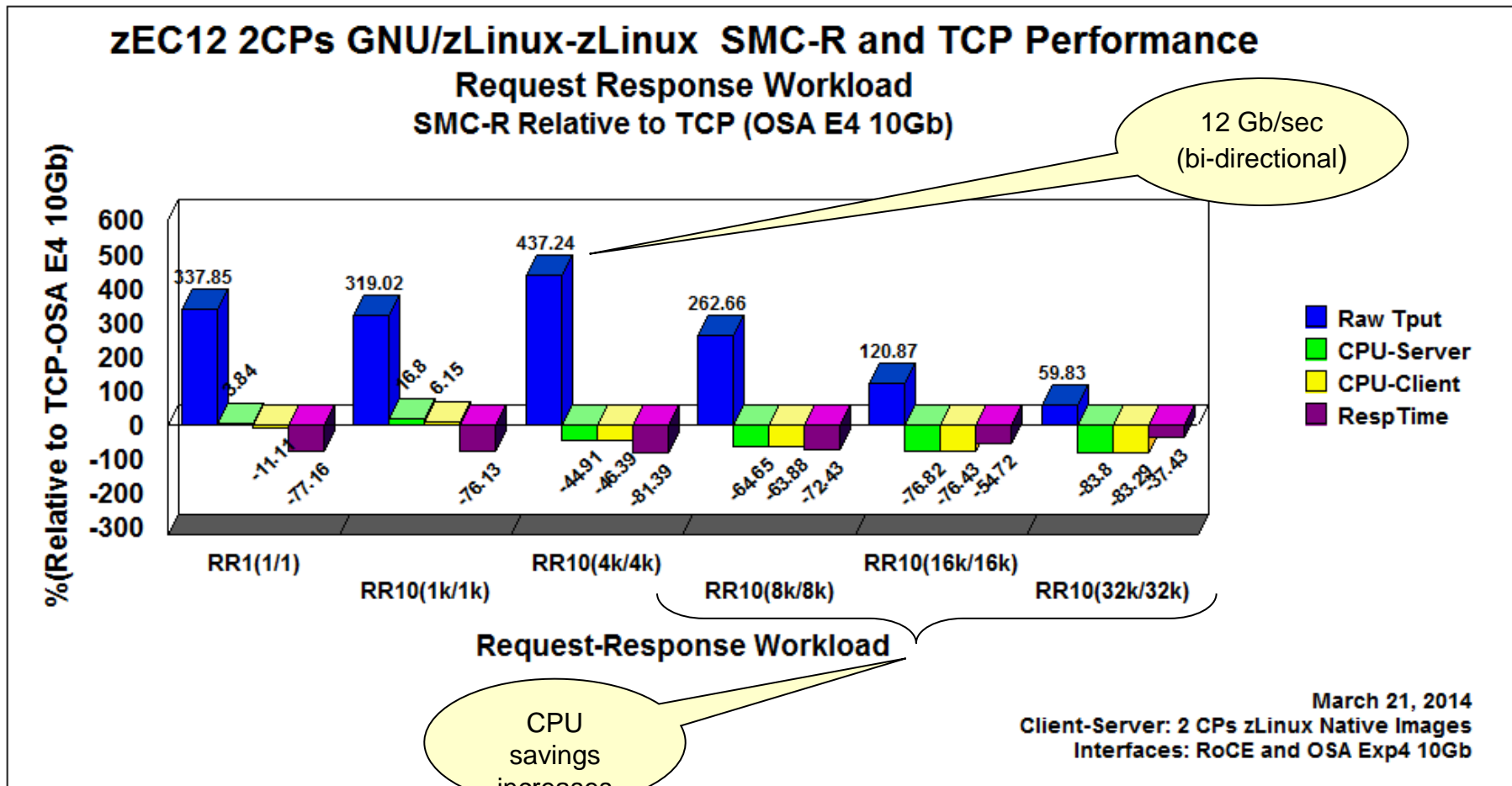
- Linux internal testing is in progress (significant progress with various levels of testing)
- Posting code: objective is summer 2014

# Linux to Linux SMC-R Performance

(micro benchmarks)



OPENFABRICS  
ALLIANCE



Note. Results are preliminary and are subject to change.

# Linux Hardware Software Performance Configuration

- Hardware Software Configuration used for measurements
  - zEC12 Client, Server Linux on System z images with dedicated 2 CPs each
  - Linux version (~ Red Hat 6.0)
    - Linux SMC-R code is preliminary code
  - z /OS V2R1
  - Network interfaces used
    - 10GbE RoCE Express
    - OSA Express4s 10Gb
  - For TCP (Linux on System z) : Running Layer 2 Mode (Checksum Offload and Segmentation Offload not applicable)



# SMC-R References

---



<http://www-01.ibm.com/software/network/commserver/SMCR/>



Thank You



#OFADevWorkshop