OFA Developer Workshop 2013

Shared Memory Communications over RDMA (SMC-R)

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" 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.

Sockets over RDMA. . .
What are the requirements?

In order to gain acceptance of RoCE based solutions in the enterprise data center network environment offerings must provide a complete solution consisting of the following:

- Performance! The primary value point of RDMA is performance advantages.[1]
- Key attributes - the following "Time to Market / Value" and Total Cost to Deploy / Operate attributes must also be considered:
  1. Full compatibility with existing socket based applications (no application changes required) [2]
  2. Can not regress key existing TCP/IP network operational or administrative attributes:
     1. Security (existing IP address based (IP filters) and TCP connection level security (SSL))
     2. HA (high availability and resiliency across redundant hardware - separate physical adapters)
     3. LB (server or clustering load balancing)
     4. Consumable (requirements to configure and deploy) or "Turnkey"
  3. Preserve the existing network IP topology, "IP eco-system" and administration model (minimize disruption and required configuration changes and runtime / ongoing operational cost)
- Interoperability (the need for a common protocol / solution).

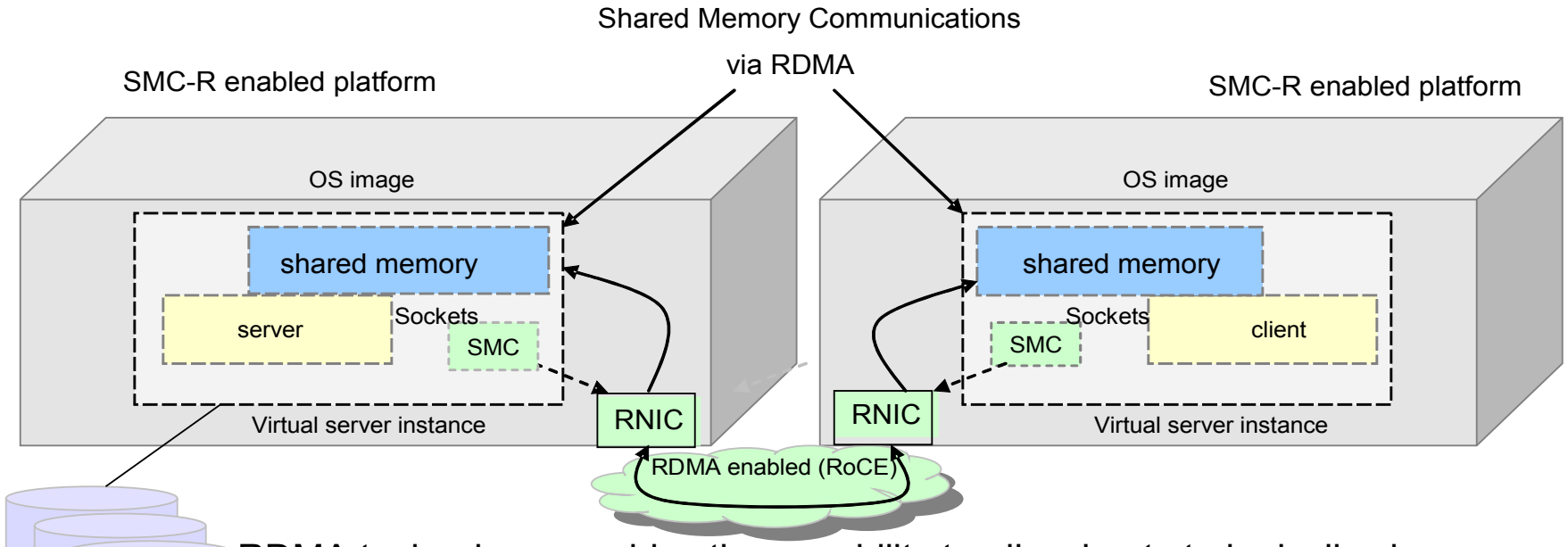
The SMC-R solution was created to meet all of the above objectives.

1. Performance is a very broad topic with many variables. Actual performance gains will vary for each environment (platform, OS, workload type, processing environment, virtualization etc.). Customers will carefully evaluate performance gains vs. cost and risk.

2. It is recognized that application specific or native RDMA solutions are also still needed and must be accounted for as part of this overall common strategy.

"Shared Memory Communications over RDMA" concepts

Clustered Systems



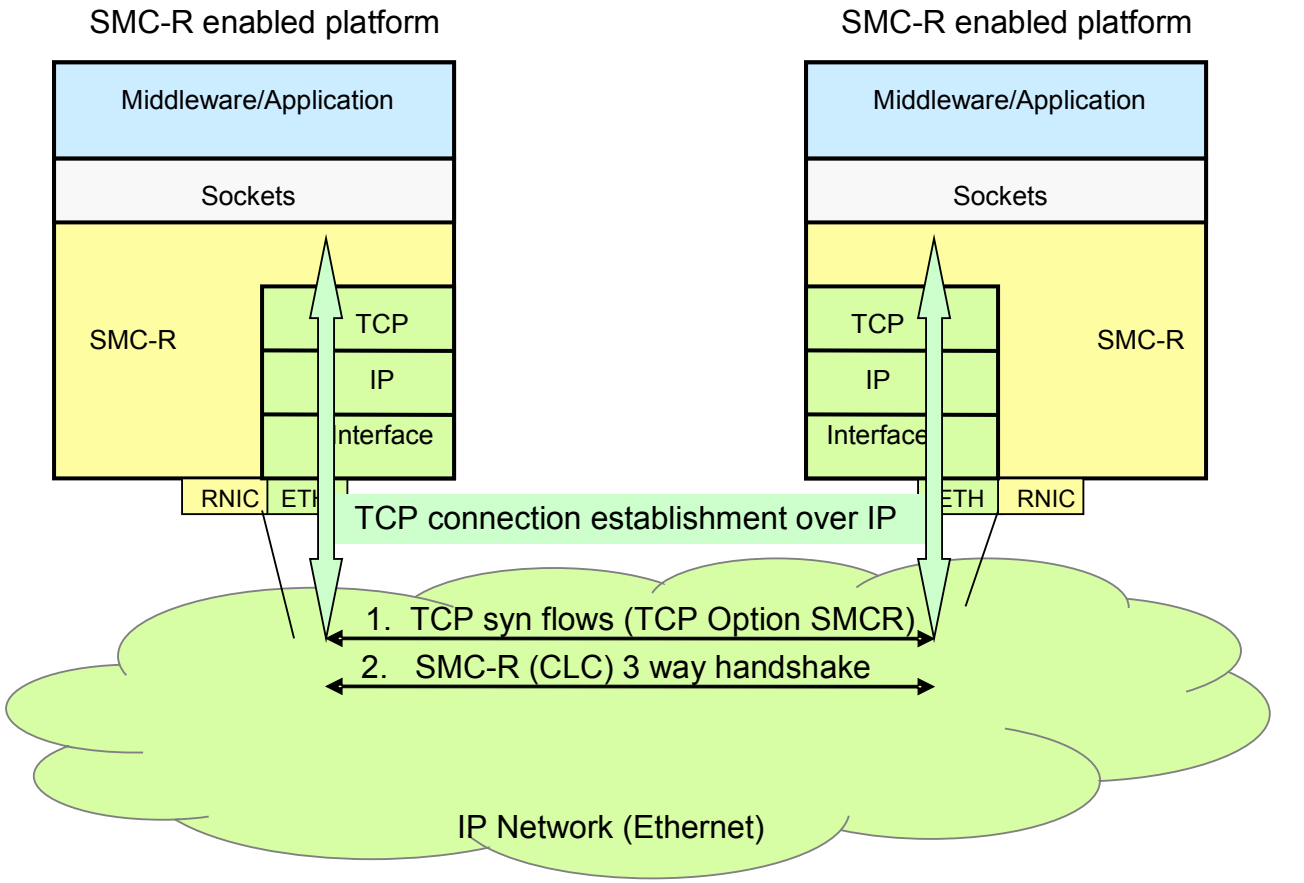
RDMA technology provides the capability to allow hosts to logically share memory. The SMC-R protocol defines a means to exploit the shared memory for communications - *transparent* to the applications! SMCR RFC: http://tools.ietf.org/html/draft-fox-tcpm-shared-memory-rdma-01

This solution is referred to as *SMC-R* (Shared Memory Communications over RDMA). SMC-R represents a sockets over RDMA protocol that provides a foundation for a complete solution meeting all of the described objectives. SMC-R is an RDMA model exploiting RDMA-writes (only) for all data movement.

SMC-R Overview

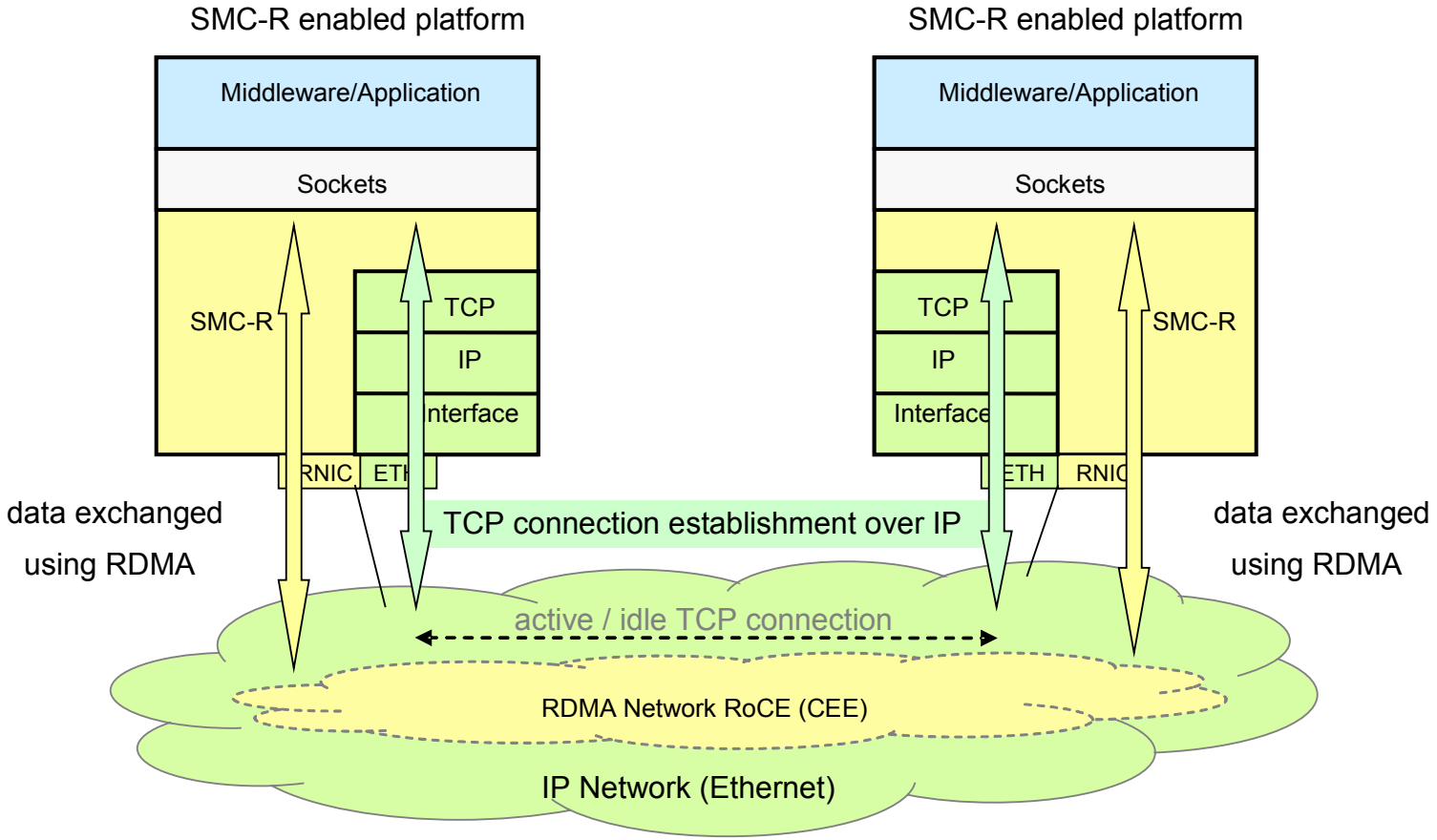
- Shared Memory Communications over RDMA (SMC-R) is a protocol that allows TCP sockets applications to transparently exploit RDMA (RoCE)
- SMC-R is a “hybrid” solution that:
  - Uses TCP connection (3-way handshake) to establish SMC-R connection
  - Switching from TCP to “out of band” SMC-R is controlled by a TCP Option (Experimental Option “magic number”)
  - SMC-R “rendezvous” (RDMA attributes) information is then exchanged within the TCP data stream
  - 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 (see backup charts)

Dynamic Transition from TCP to SMC-R (part 1)



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

Dynamic Transition from TCP to SMC-R (part 2)

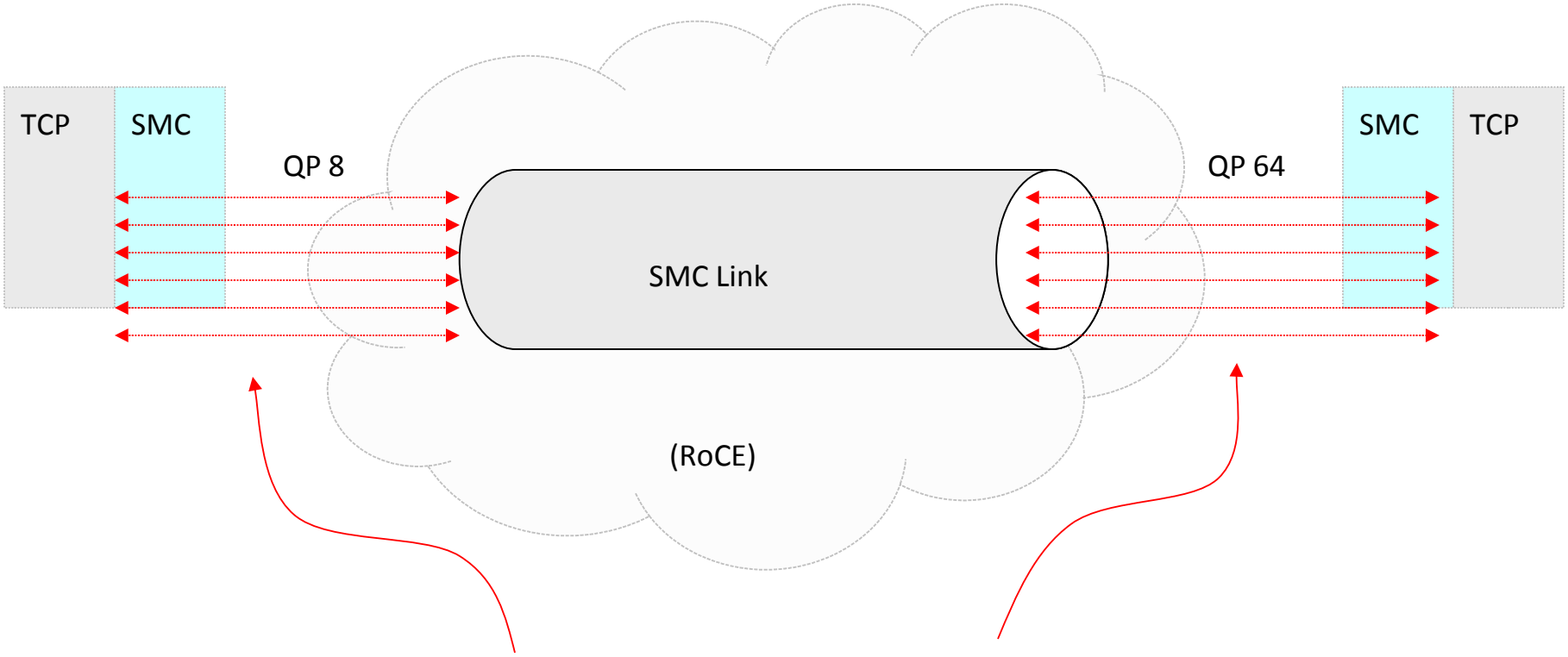


TCP connection transitions to SMC-R allowing application data to be exchanged using RDMA

Why a Hybrid Protocol?
(TCP + SMC-R)

- Follows standard TCP/IP connection setup
- Dynamically switches to RDMA (SMC-R)
- TCP connection remains active (idle) and is used to control the SMC-R connection
- Preserves critical operational and network management TCP/IP features such as:
  - No IP topology changes
  - Preserves existing IP security model (e.g. IP filters, policy, VLANs, SSL etc.)
  - Compatibility with TCP connection level load balancers
  - Minimal network admin / management changes

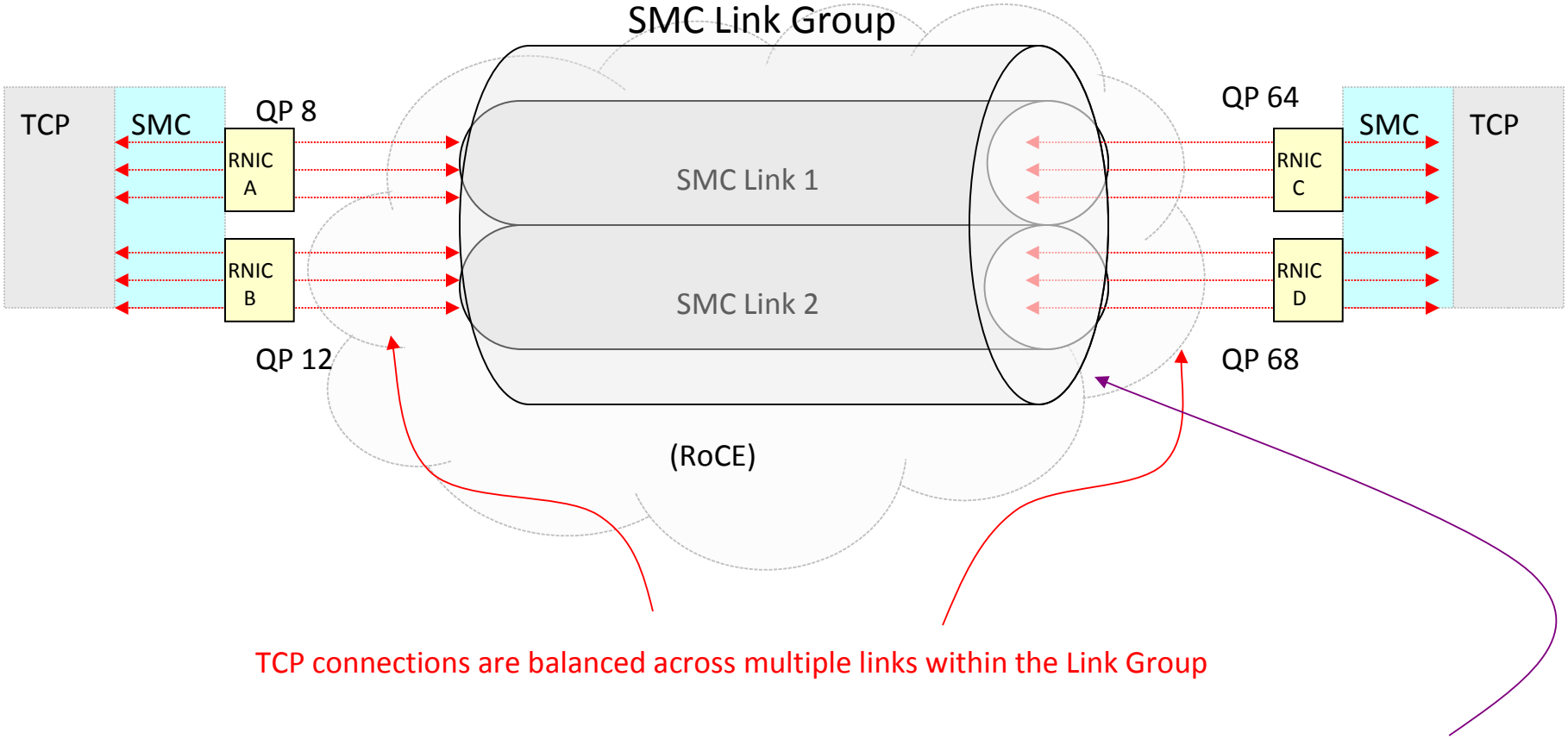
SMC-R Link Architecture (RC-QPs)



Multiple TCP (via SMC) Connections share the same SMC Link

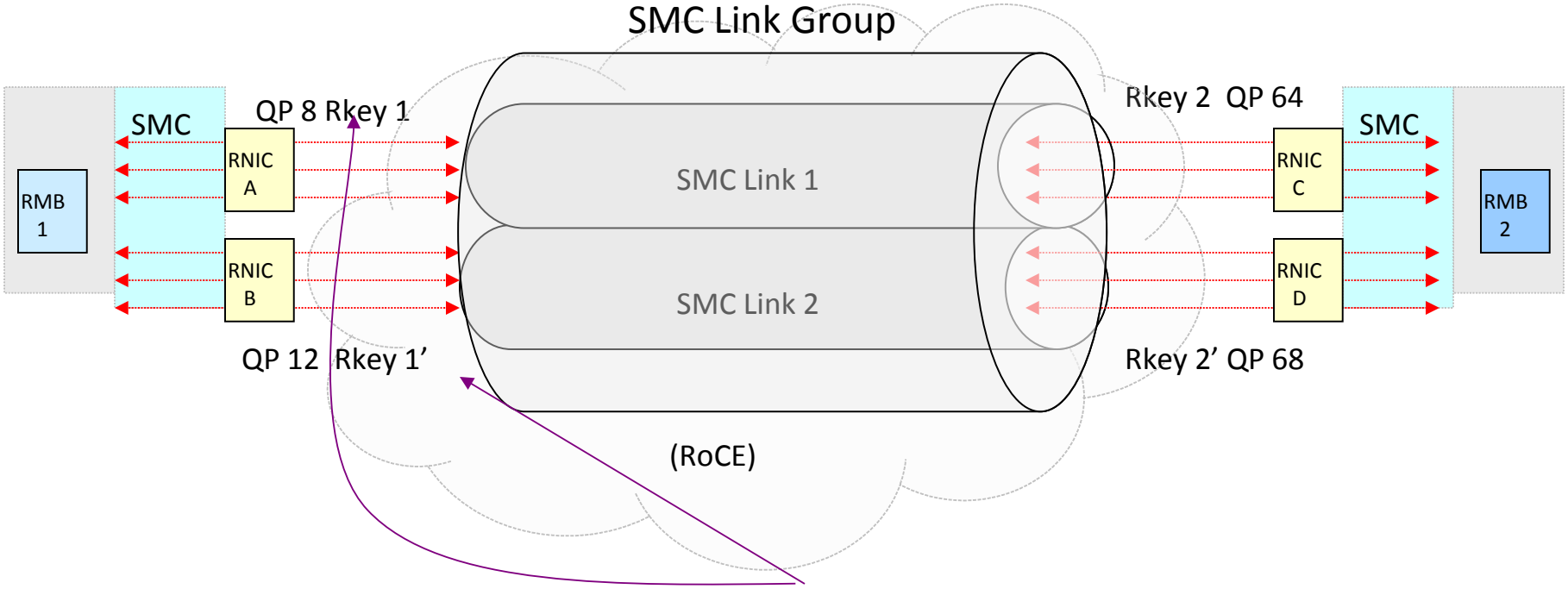
SMC-R Link Groups

(Provides resiliency, link level load balancing and additional bandwidth)



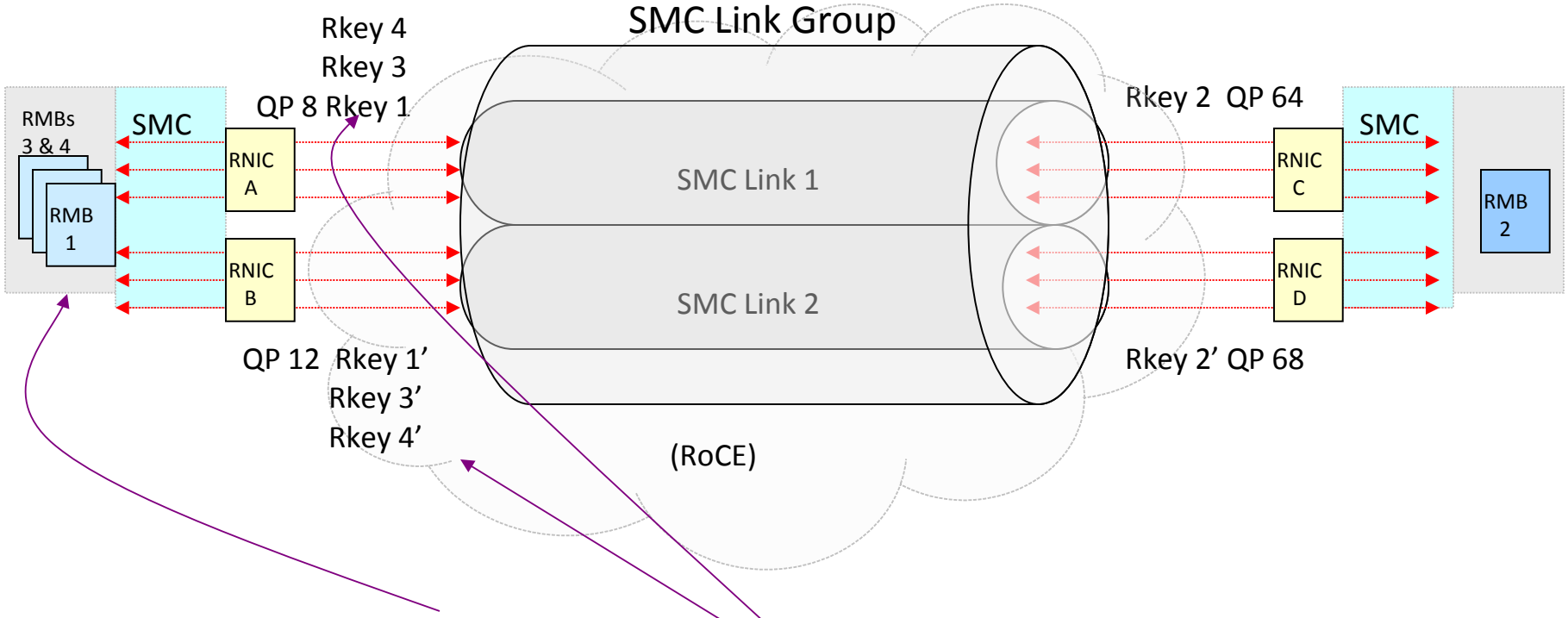
Multiple SMC Links across unique physical RNICs are grouped together to form a single SMC Link Group

SMC-R Memory Architecture (Part 1)



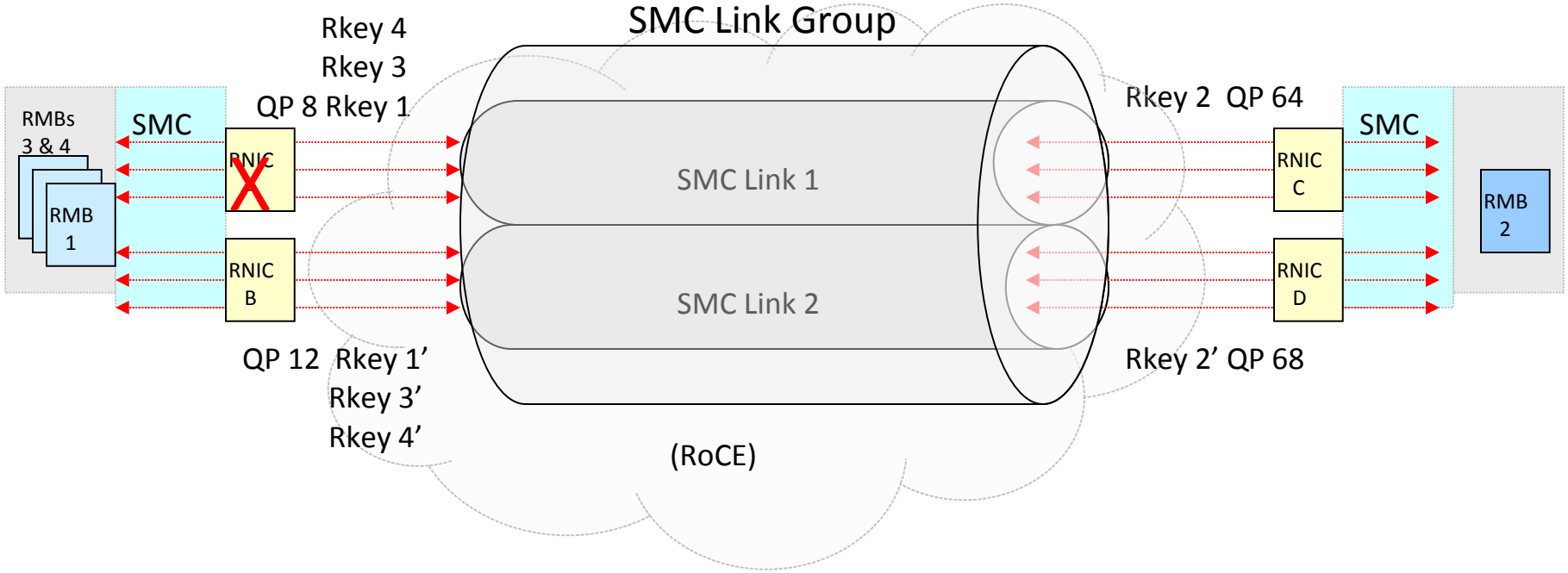
Each SMC Link has equal access (unique Rkey) to the peer's memory or RMB(s)

SMC-R Memory Architecture (Part 2)



SMC link groups also support multiple RMBs. Each peer can independently manage (add or remove) RMBs based on the needs of the link group, workload, and OS unique memory management requirements. Again, all SMC links continue to have equal access (Rkeys) to all RMBs.

SMC-R Architecture (High Availability)

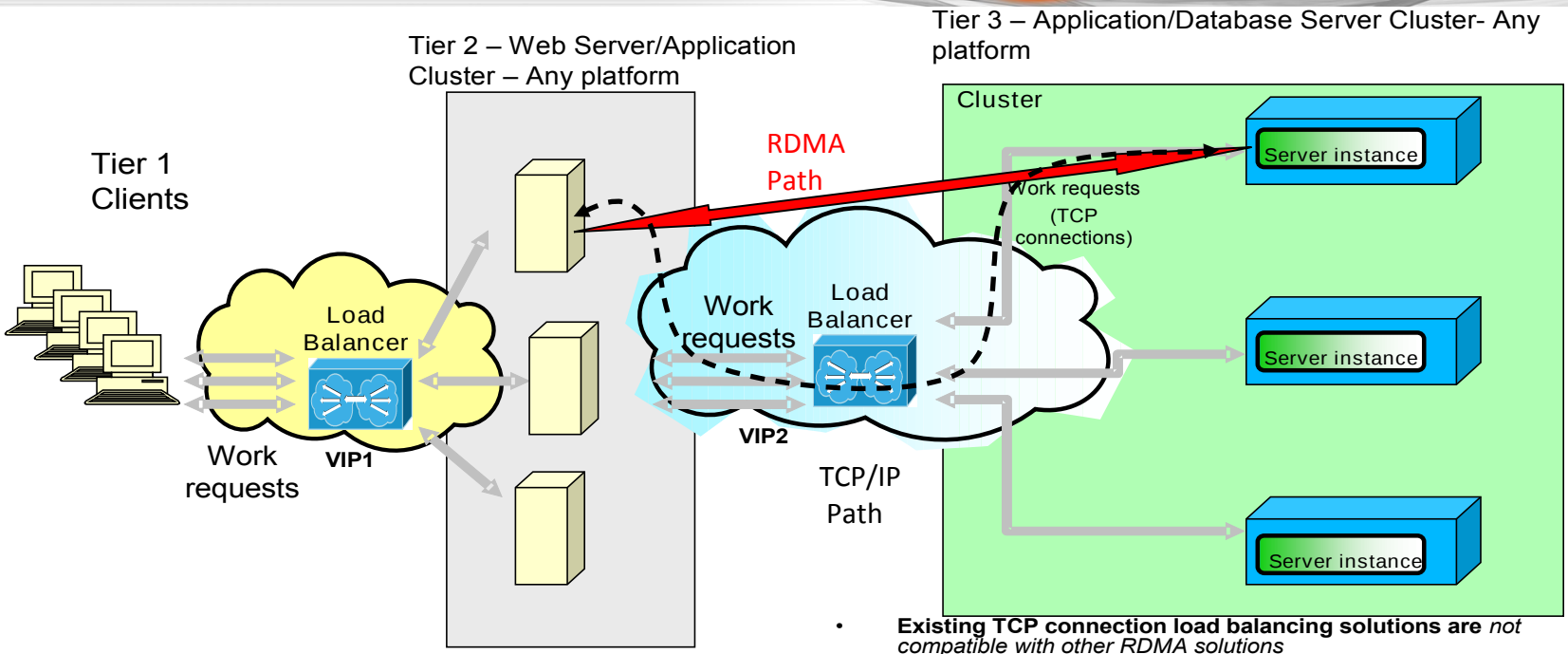


If one path (e.g. an RNIC) becomes unavailable (in this example RNIC A)… then:

- 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!

Server clustering and TCP connection load balancing (compatibility 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 Key Attributes

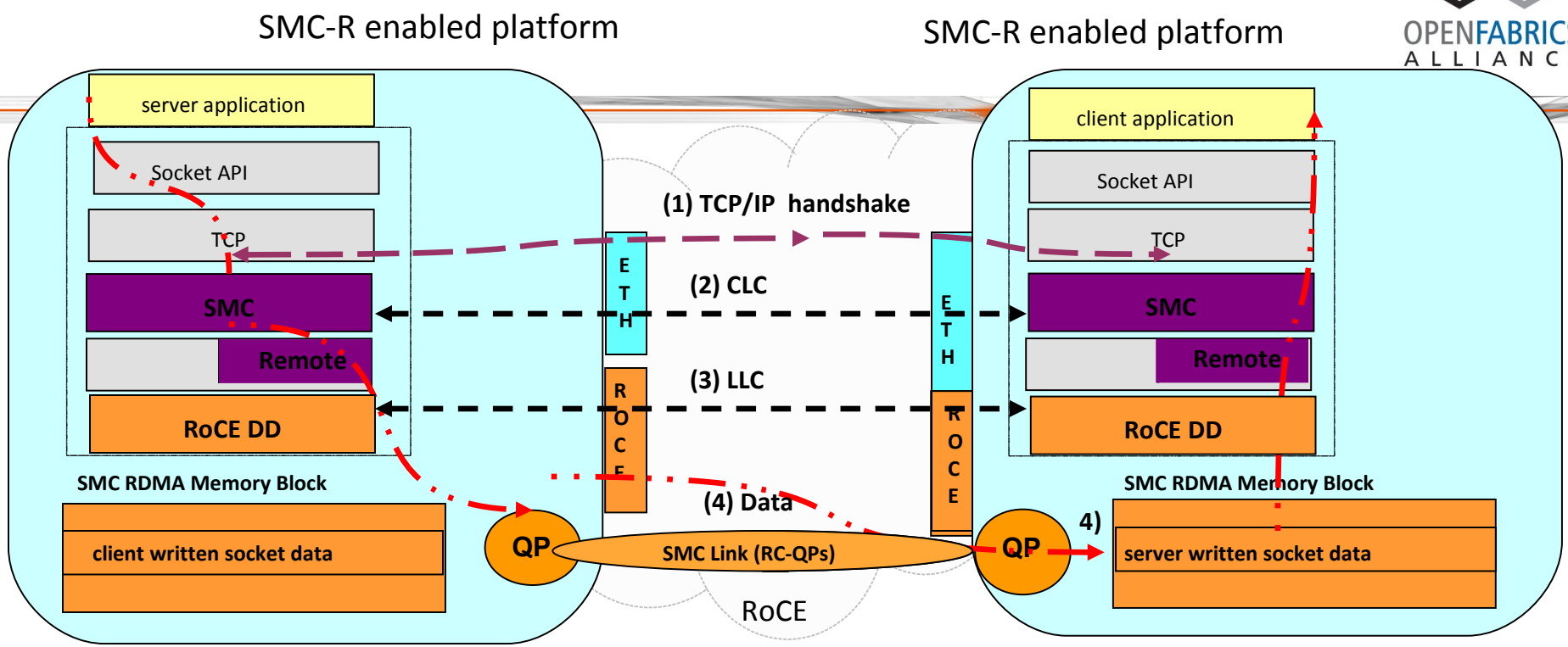
1. **Performance – significant improvements over standard TCP**
   - Transaction rate / latency (for large number of connections)
   - Throughput (streaming) + CPU savings
   - Scalability
2. **Socket transparency and compliance to Sockets API**
   - e.g. Urgent data, Fork() support, etc.
3. **RoCE compatible (exploits existing Ethernet)**
4. **IP and connection related security features (IP Filters, SSL, etc.)**
5. **High-availability: non-disruptive failover to redundant hardware**
6. **Preserves IP "eco-system" (i.e. no IP topology changes with dynamic discovery)**
7. **Transparent compatibility with existing Load-balancing**
8. **Resource sharing / scalability (memory, QPs, VFs, etc.)**
9. **Preserve existing application task threading model (non-blocking)**
10. **Minimal changes for network admin, management and operations**

Feedback

- This solution is targeting the enterprise class of customers. It is acknowledged that different categories of users will have unique values and priorities.
- Assuming the performance attributes of SMC-R are “significant” what’s your view of this approach?
- Are the key attributes (security, HA, LB, etc.) described here important to you or to your customers’ environment?
- What’s missing? Questions or comments?

Backup

SMC-R (Contact and RDMA Processing - Concepts)



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 attributes (QPs, RMBEs, GIDs, etc.) within TCP data stream (CLC messages) … can still fall back to IP

3) **If first contact**…. then establish point-to-point SMC Link via SMC LLC 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; peer consumes data via standard socket read

Thank You

OPENFABRICS ALLIANCE