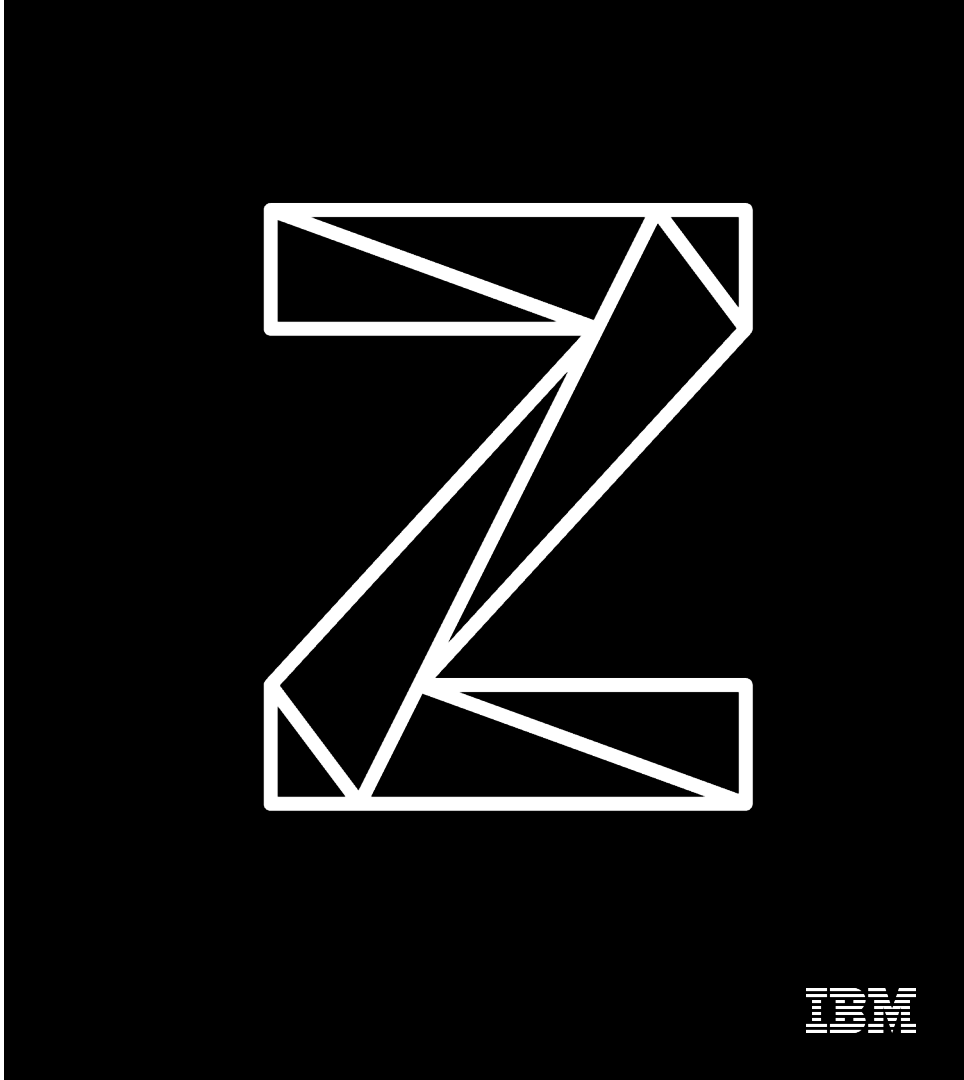
Accelerate Networking with Shared Memory Communications for IBM Z

Chelsea Jean-Mary Isaac
Washington Systems Center



Trademarks

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

CICS*	IBM*	IBM Z*	z15
Db2*	IBM (logo)*	LinuxONE	z/OS*
GDPS*	IBM Cloud Pak	WebSphere*	z/VM*
HiperSockets	ibm.com	z14*	z/VSE*

*** Registered trademarks of IBM Corporation**

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.
IT Infrastructure Library is a Registered Trade Mark of AXELOS Limited.
ITIL is a Registered Trade Mark of AXELOS Limited.
Linear Tape-Open, LTO, the LTO Logo, Ultrium, and the Ultrium logo are trademarks of HP, IBM Corp. and Quantum in the U.S. and other countries.
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.
Linux is a registered trademark of Linus Torvalds in the United States, other countries, or both.
Kubernetes and Container Initiative™ are registered trademark of The Linux Foundation.
Red Hat and Red Hat OpenShift are registered trademarks of Red Hat, Inc. Open
Microsoft, Windows, Windows NT, and the Windows logo are trademarks of Microsoft Corporation in the United States, other countries, or both.
Java and all Java-based trademarks and logos are trademarks or registered trademarks of Oracle and/or its affiliates.
UNIX is a registered trademark of The Open Group in the United States and other countries.
VMware, the VMware logo, VMware Cloud Foundation, VMware Cloud Foundation Service, VMware vCenter Server, and VMware vSphere are registered trademarks or trademarks of VMware, Inc. or its subsidiaries in the United States and/or other jurisdictions.
Other product and service names might be trademarks of IBM or other companies.

Notes:
Performance is in Internal Throughput Rate (ITR) ratio 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.
All customer examples cited or described in this presentation are presented as illustrations of the manner in which some customers have used IBM products and the results they may have achieved. 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.

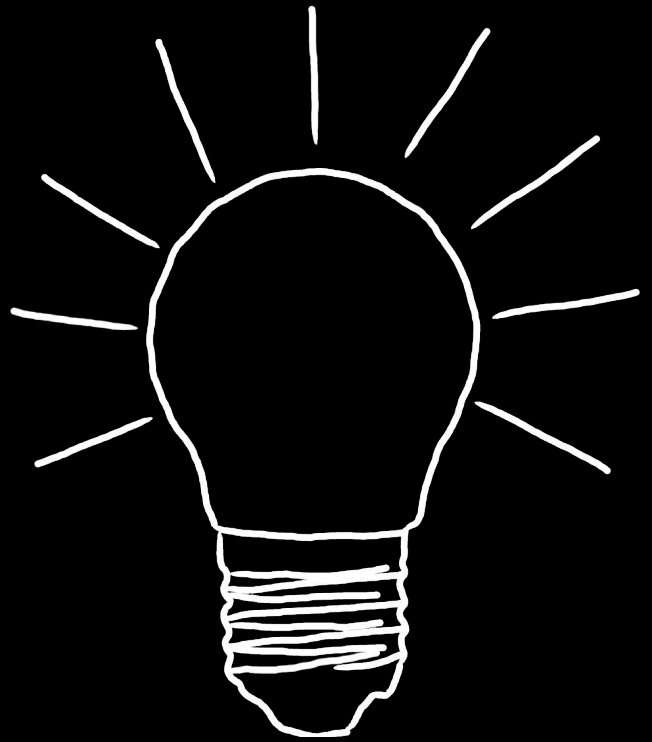
Agenda

- **SMC Basics**
  - Motivation
  - The SMC Protocol
  - Benefits
- **SMC for Z**
  - SMC-D and SMC-R
- **SMC in Action**
- **Miscellaneous**

What if we had a networking technology that could provide

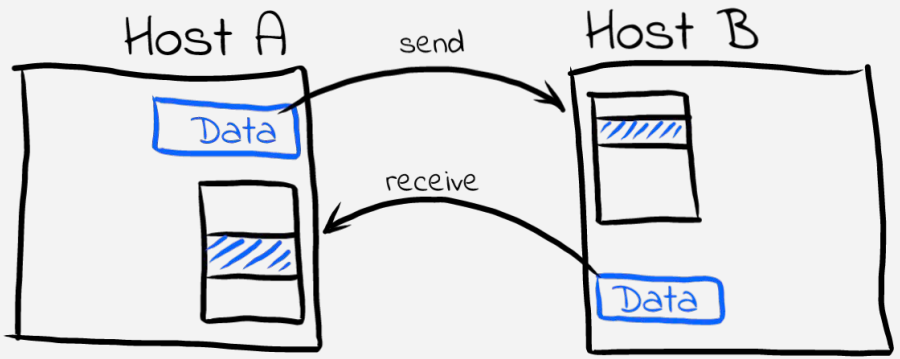
- **low latency**
- **high throughput**

and **save CPU cycles** at the same time?



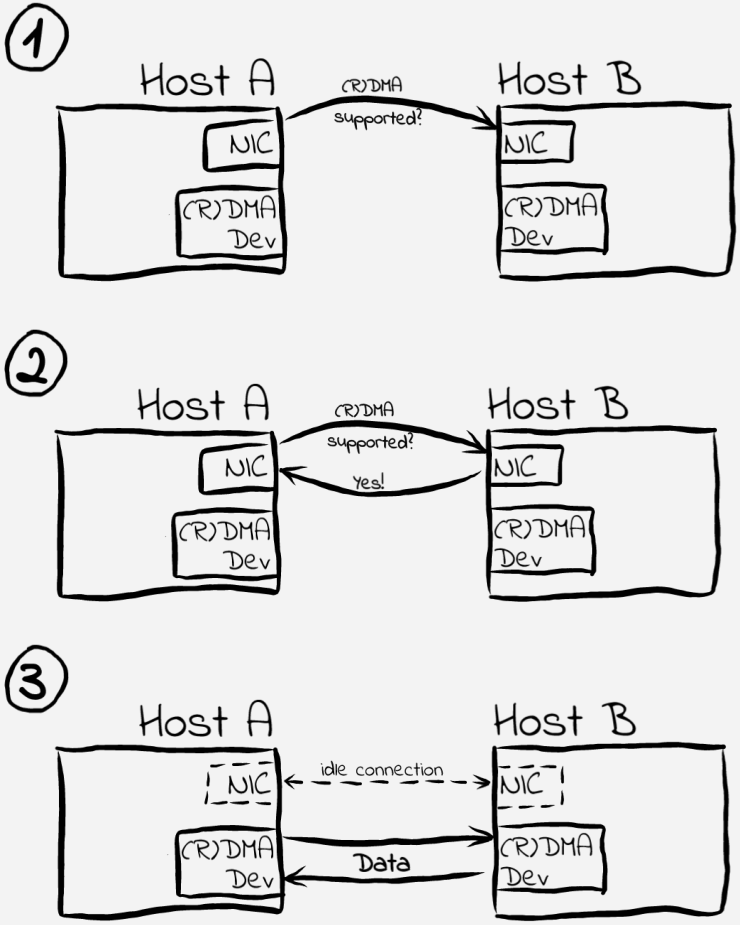
The RDMA Approach

- RDMA (***R****emote* ***D****irect* ***M****emory* ***A****ccess*) based technology originating from Infiniband (IB)
- Enables a host to read or write directly from/to a remote hosts's memory with drastically reduced use of remote hosts's CPU (interrupts required for notification only)
- Native / direct application exploitation requires rewrite of network-related program logic, deep level of expertise in RDMA and a new programming model
- Therefore, provide a transparent approach:
  - **SMC-R**: Use *RDMA* over Converged Ethernet (RoCE) technology
    - Unlike IB, RoCE does not require unique network components (host adapters, switches, security controls, etc.)
  - **SMC-D**: Use *DMA* when both hosts are within a Z system via virtual PCI device



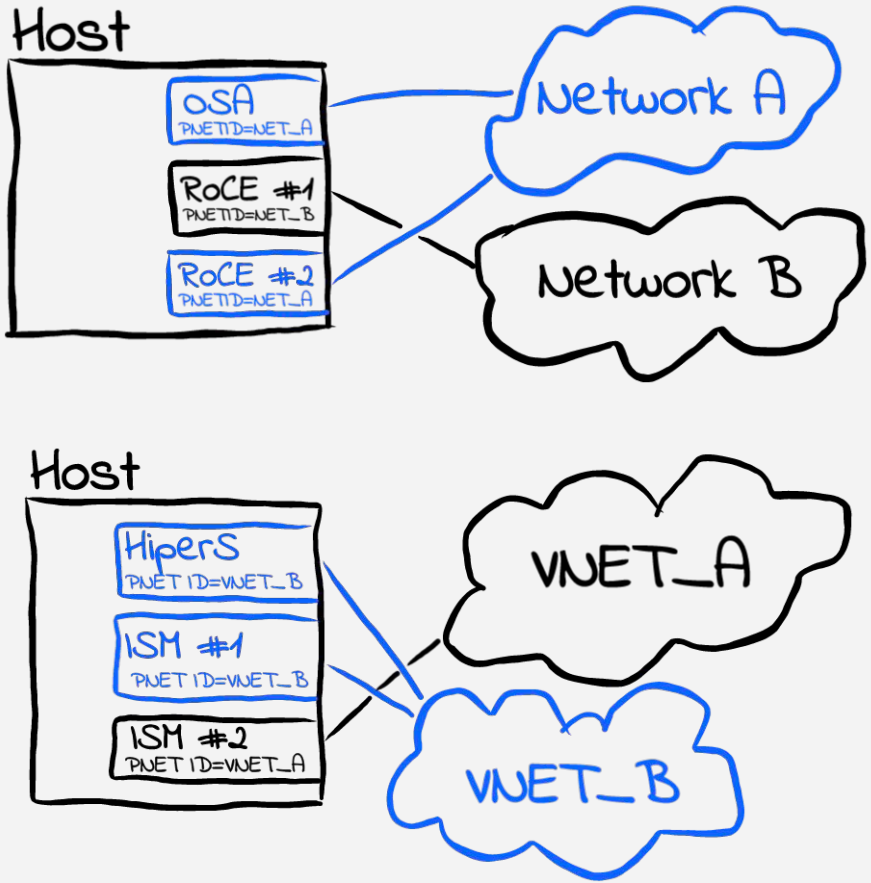
Overview

- For each new TCP connection:
  - Start out with a regular TCP/IP connection, advertising (R)DMA capabilities
  - If peer confirms, negotiate details about the (R)DMA capabilities & connectivity
  - Switch over to an (R)DMA device for actual traffic depending on the peers' capabilities
  - Regular TCP connection through NICs remains active but idle



PNET IDs

- **PNET ID:** *Physical network identifier*
- Customer-defined value to logically group NICs and RDMA adapters connected to the same physical network within a host.
- Defined in
  - IOCDS for any of OSA, RoCE, HiperSockets or ISM
- *Typically* associate
  - OSA and RoCE cards, or
  - HiperSockets and ISM devices
- **Note:** PNET IDs help to locate a suitable (R)DMA device for a given NIC ***within a host***. The peer can use totally different PNET IDs (as long as the right devices are grouped)

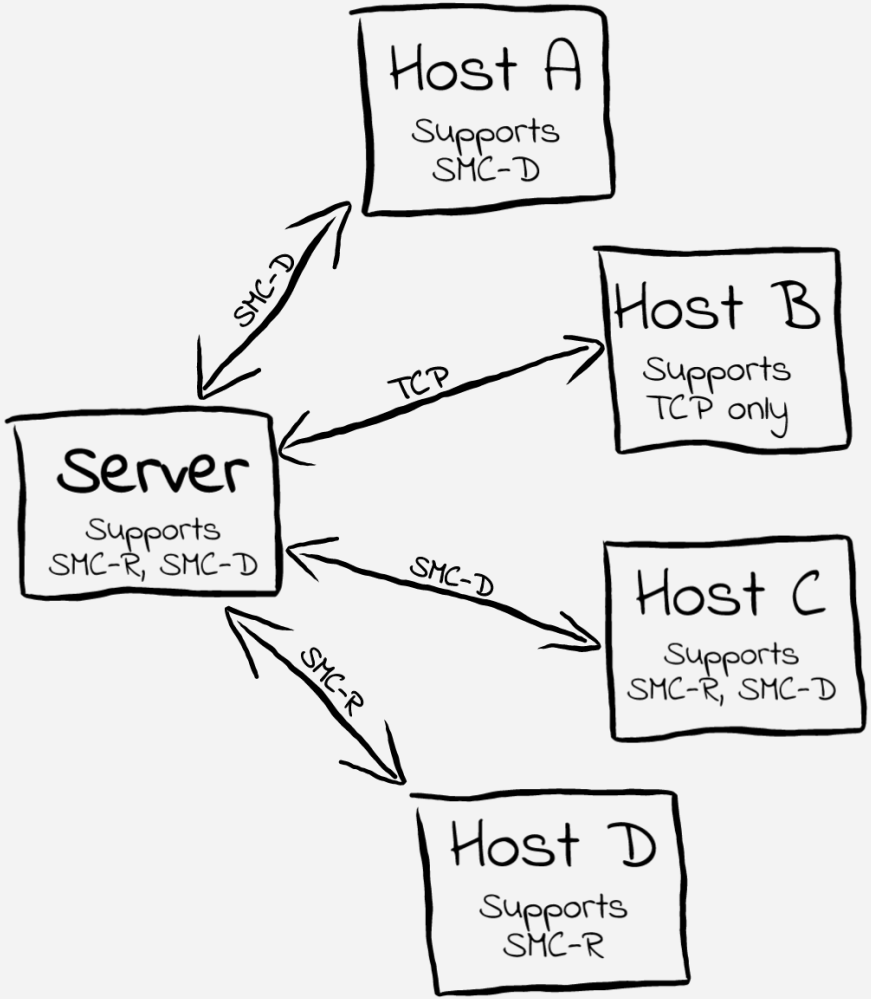


Benefits

- Less latency
- Lower CPU usage
- Run your applications unmodified
  - SMC is transparent to existing applications – no changes required

Mixing SMC Usage

- Both variants of SMC can be used concurrently to provide an optimized solution
- Enable SMC independent of peers' capabilities; i.e. no commonality in SMC support on all peers required
- Use:
  - SMC-D for local connections
  - SMC-R for remote connections
  - Fall-back to regular TCP where neither SMC variant is supported



Agenda

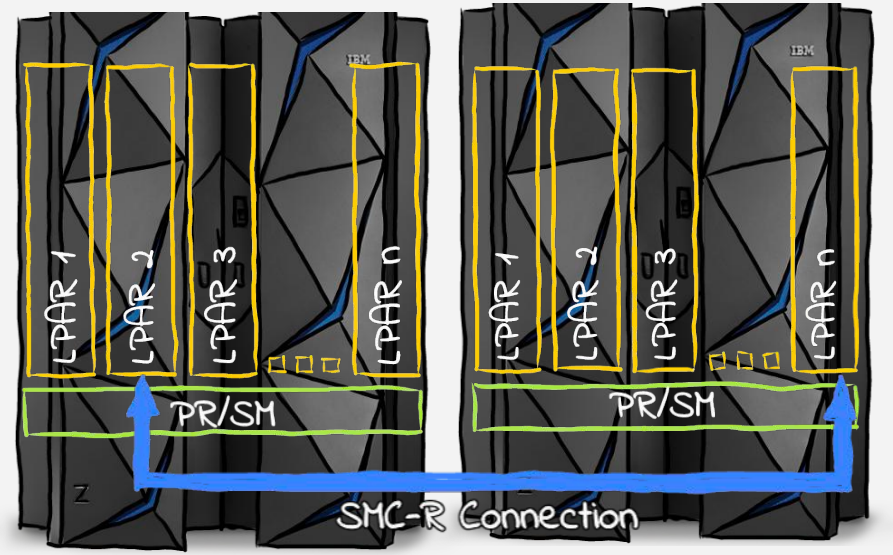
- **SMC Basics**
  - Motivation
  - The SMC Protocol
  - Benefits
- **SMC for Z**
  - SMC-D and SMC-R
- **SMC in Action**
- **Miscellaneous**

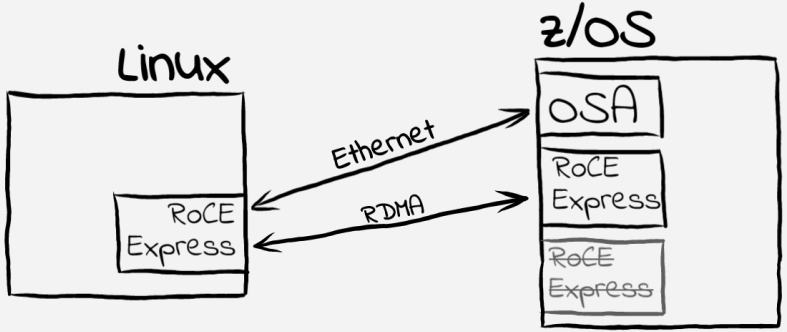
SMC-D Overview

- Intra-CEC connectivity using ***Internal Shared Memory*** (ISM) devices
- IBM Z hardware requirements
  - IBM z13 (requires driver level 27 (GA2)) and z13s, or later
  - LinuxONE Emperor and LinuxONE Rockhopper, or later
  - Classic mode only (i.e. DPM not supported)
- ISM devices
  - *Virtual* PCI network adapter of new VCHID type ISM
  - Provides access to memory shared between LPARs
  - 32 ISM VCHIDs per CPC, 255 VFs per VCHID (8K VFs per CPC total) i.e. the maximum no. of virtual servers that can communicate over the same ISM VHID is 255
  - Each ISM VCHID represents a unique (isolated) internal network, each having a unique Physical Network ID
    - SMC-D version 2 allows for IP devices without a PNetID
- PNET ID configuration
  - IOCDS only
  - Use HiperSockets, OSA or RoCE cards for regular connectivity

SMC-R Overview

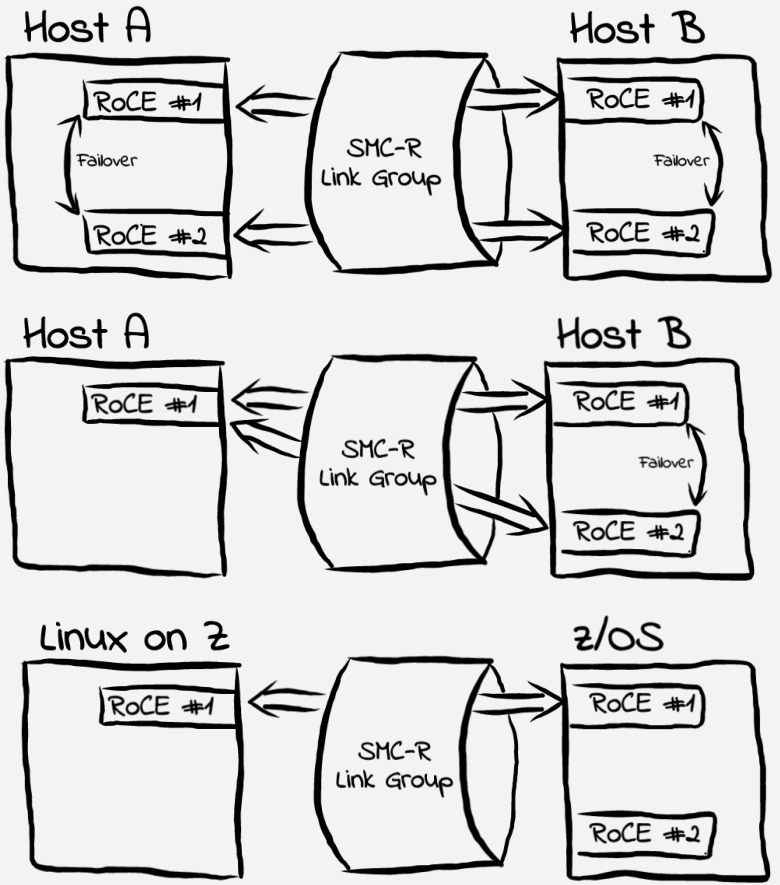
- Cross-CEC connectivity using ***RoCE Express*** cards
- IBM Z hardware requirements
  - IBM z12EC and z12BC or later
  - LinuxONE Emperor and LinuxONE Rockhopper or later
  - Classic and DPM mode supported
- RoCE Express cards
  - RoCE Express & RoCE Express2 cards supported
  - Switches need to support and enable *Global Pause* (standard Ethernet switch flow control feature as described in IEEE 802.3x)
- PNET ID configuration
  - IOCDS only
  - Use OSA or RoCE card for regular connectivity
- **Note:**
  - Linux on Z can use a single RoCE card for regular and RDMA traffic!
  - No link failover!





SMC-R Link Groups

- SMC-R ***link groups*** provide for load balancing and recovery
  - New TCP connection is assigned to the SMC-R link with the fewest TCP connections
  - Load balancing only performed when multiple RoCE Express adapters are available at each peer
- **Full redundancy** requires:
  - Two or more RoCE Express adapters at each peer
  - Unique system internal paths for the RoCE Express adapters
  - Unique physical RoCE switches
- **Partial redundancy** still possible in the absence of one or more of these conditions
- Linux on Z:
  - No failover support (yet)



Agenda

- **SMC Basics**
  - Motivation
  - The SMC Protocol
  - Benefits
- **SMC for Z**
  - SMC-D and SMC-R
- **SMC in Action**
- **Miscellaneous**

Supported Environments

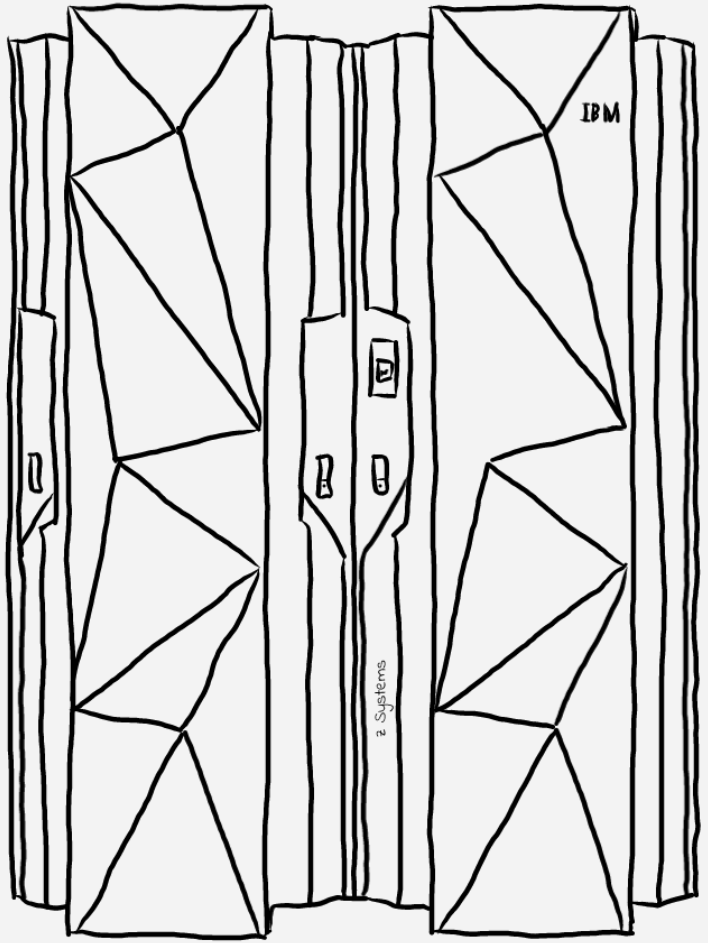
- **Operating Systems:**
  - ***SMC-D***
    - **Linux on Z**
    - **z/OS**: IBM z/OS V2R2 (via APAR) or later
  - ***SMC-R***
    - **Linux on Z**
    - **z/OS**: IBM z/OS V2R1 (via APAR) or later
    - **AIS**: System P with AIX 7.2
- **Linux on Z Environments**
  - **LPAR** yes
  - **z/VM guests** yes (z/VM 6.3 or later)
  - **KVM guests** in progress
  - **Docker** limited support, to be improved

Prerequisites

- ***Direct connectivity*** over same IP subnet. i.e. no routed traffic, no peers in different IP subnets
- (R)DMA device(s) attached and configured
- PNET IDs assigned
- ***TCP only***, i.e. no UDP
- No IPsec (SSL/TLS works)
- No NAT (violates same subnet prerequisite)
- (SMC-D only): Classic mode required (i.e. no DPM support)

Agenda

- **SMC Basics**
  - Motivation
  - The SMC Protocol
  - Benefits
- **SMC for Z**
  - SMC-D and SMC-R
- **SMC in Action**
- **Miscellaneous**



Summary

Key Attributes

- Leverages existing Ethernet infrastructure (SMC-R)
- Transparent to (TCP socket based) application software
- Preserves existing network addressing-based security models
- Preserves existing IP topology and network administrative and operational model
- Transparent to network components such as channel bonding and load balancers
- Built-in failover capabilities (SMC-R)

Typical Workloads To Benefit

- *Transaction-oriented,*
- *Latency-sensitive,* and
- *Bulk data streaming*, e.g. when running backups.

References

- ***smc-tools* Homepage**
  https://www.ibm.com/developerworks/linux/linux390/smc-tools.html
- **Whitepaper: Performance Evaluation of SMC-D with SAP Banking on IBM Z**
  http://www-03.ibm.com/support/techdocs/atsmastr.nsf/WebIndex/WP102792
- **RFC7609 (SMC-R)**
  https://tools.ietf.org/html/rfc7609
- **Linux on Z (technical):**
  https://www.ibm.com/developerworks/linux/linux390/
- **SMC for Linux on Z:**
  http://linux-on-z.blogspot.com/p/smc-for-linux-on-ibm-z.html
- **Webcasts**
  https://developer.ibm.com/tv/linux-ibm-z/
- **Blogs**
  - **Linux on z distributions new** http://linuxmain.blogspot.com/
  - **Linux on Z latest development news** http://linux-on-z.blogspot.com/
  - **KVM on Z** http://kvmonz.blogspot.com/
  - **Containers on Z, primarily *Docker*** http://containersonibmz.com/