Configuring Shared Memory Communications Version 2 (SMCv2) For Your z/OS Environment

Chelsea Jean-Mary Isaac
Client Technical Specialist

chelsea.t.jean-mary@ibm.com
IBM Washington Systems Center

IBM

Trademarks

Agenda

SMCv2 Requirements

SMC-Rv2 Requirements
Hardware Prerequisites

IBM Z Hardware Requirements
- IBM z15 or z16

RDMA over Converged Ethernet (RoCE) Express2 or RoCE Express3*
- PNetIDs assigned
  - OSA and RoCE PNetIDs must be configured and match
- PFIDs assigned
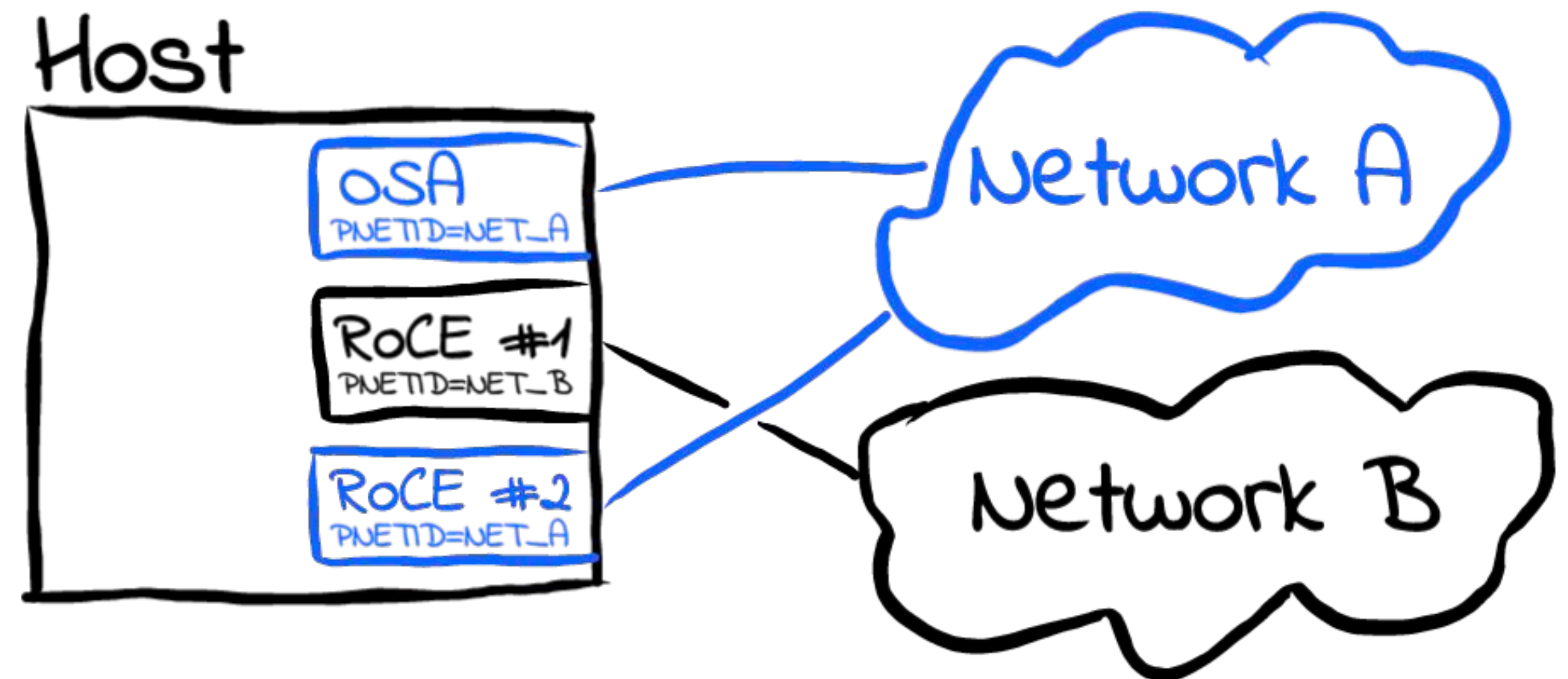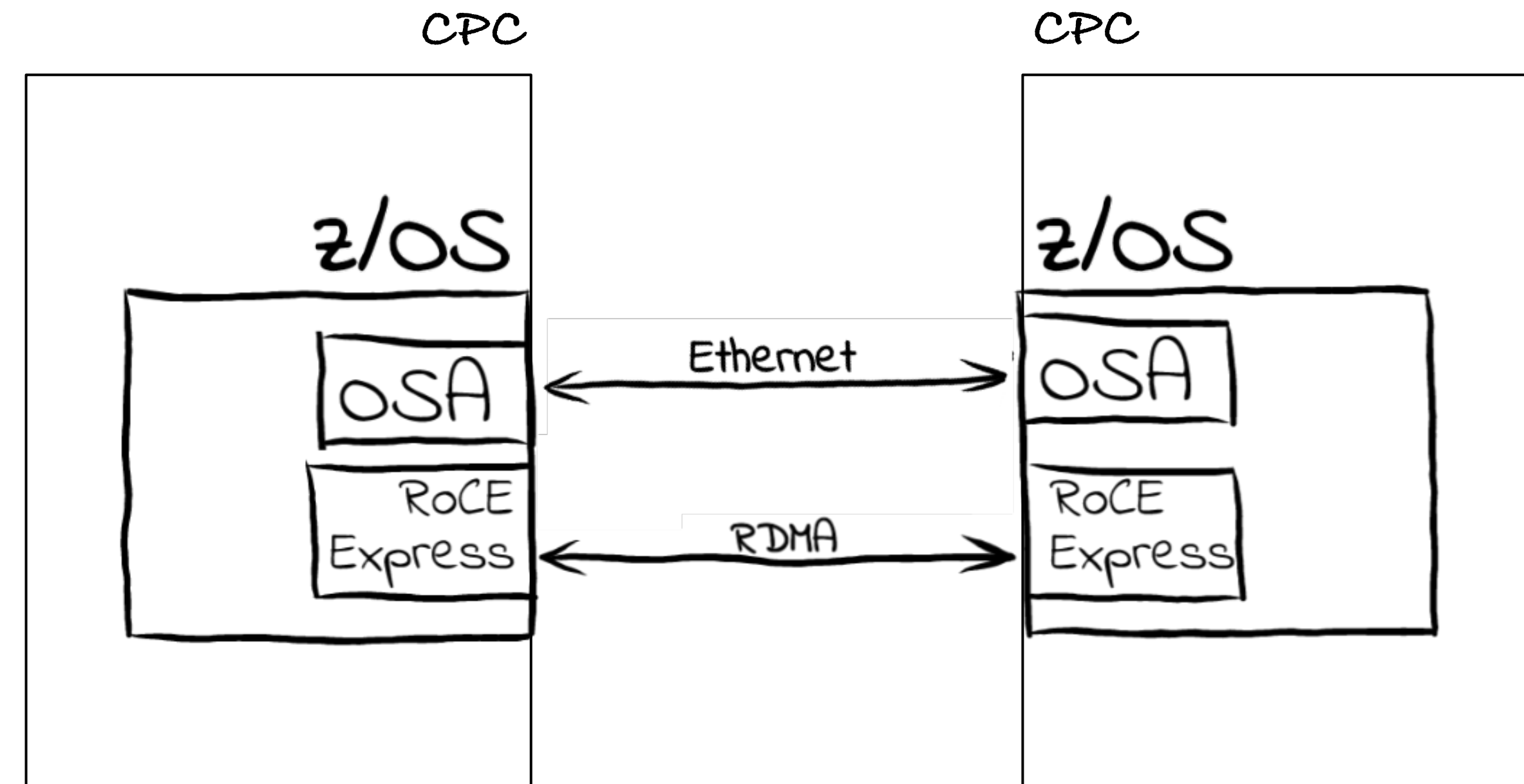
*z16 only

SMC-Rv2 Requirements

Hardware Prerequisites

Setting up the environment for Shared Memory Communications over RDMA

1. Install and configure the "RoCE Express2" or "RoCE Express3" (z16) features in the hardware configuration definition (HCD).

2. Assign PNetID values and configure the values in the HCD for both the "RoCE Express" ports and any OSA-Express devices that will use the "RoCE Express2" or "RoCE Express3" ports for SMC-R communications.

3. Configure Ethernet switches for RoCE Express features.
   - Switch port must support jumbo frames if RoCE MTU size is configured at 2K or 4K
   - Switch port must be configured in TRUNK mode and have all OSA VLAN in the access list.

SMC-Rv2 Requirements Hardware Configuration Definition

HCD Main Panel:

On the Primary Task Selection panel, select "Define configuration data".

```
                              z/OS V2.4 HCD
Command ===> ---------------------------------------------------------

                       Hardware Configuration

Select one of the following.

1      0.    Edit profile options and policies
       1.    Define, modify, or view configuration data
       2.    Activate or process configuration data
       3.    Print or compare configuration data
       4.    Create or view graphical configuration report
       5.    Migrate configuration data
       6.    Maintain I/O definition files
       7.    Query supported hardware and installed UIMs
       8.    Getting started with this dialog
       9.    What's new in this release

For options 1 to 5, specify the name of the IODF to be used.

I/O definition file . . . 'KUTCAT.IODF23'                            +
```

SMC-Rv2 Requirements Hardware Configuration Definition

HCD Main Panel:

On the resulting panel, select "Processors".

```
                        z/OS V2.4 HCD
        ─── Define, Modify, or View Configuration Data ───

     Select type of objects to define, modify, or view data.

     3_  1. Operating system configurations
                consoles
                system-defined generics
                EDTs
                    esoterics
                    user-modified generics
          2. Switches
                ports
                switch configurations
                    port matrix
          3. Processors
                channel subsystems
                    partitions
                    channel paths
                PCIe functions
          4. Control units
          5. I/O devices
          6. Discovered new and changed control units and I/O devices
     F1=Help    F2=Split    F3=Exit    F9=Swap    F12=Cancel
```

SMC-Rv2 Requirements Hardware Configuration Definition

HCD Main Panel:

On the Processor List panel, select an eligible processor and action "Work with PCIe functions".

```
    Goto   Filter   Backup   Query   Help
 -----------------------------------------------------------------------------
                            Processor List          Row 1 of 3 More:        >
 Command ===> _____ Scroll ===> CSR

 Select one or more processors, then press Enter. To add, use F11.


 / Proc. ID Type +    Model +   Mode+ Serial-# + Description
 _ WSCCECF  8561      T01       LPAR             WSC
 f WSCCECK  3931      A01       LPAR             WSC
 _ WSCCECQ  3906      M04       LPAR             WSC
 ***************************** Bottom of data ********************************
```

```
 F1=Help          F2=Split        F3=Exit        F4=Prompt     F5=Reset        F7=Backward
 F8=Forward       F9=Swap         F10=Actions    F11=Add       F12=Cancel      F13=Instruct
 F20=Right        F22=Command
```

SMC-Rv2 Requirements
Hardware Configuration
Definition

```
    Goto   Filter   Backup   Query   Help
-----------------------------------------------------------------------------
                          PCIe Function List          Row 1 of 16 More:      >
Command ===> _____ Scroll ===> CSR

Select one or more PCIe functions, then press Enter. To add, use F11.

Processor ID . . . . . : WSCCECK


/ FID    CHID+   P+   VF+   Type+         UID    Description
_ 0011   1D8     1    1     ROCE-2        ____   _____
_ 0012   1D8     1    2     ROCE-2        ____   _____
_ 0013   1D8     1    3     ROCE-2        ____   _____
_ 0014   1D8     1    4     ROCE-2        ____   _____
_ 0015   1D8     1    5     ROCE-2        ____   _____
_ 0016   1D8     1    6     ROCE-2        ____   _____
_ 0017   1D8     1    7     ROCE-2        ____   _____
_ 0018   1D8     1    8     ROCE-2        ____   _____
_ 0021   1BC     1    1     ROCE-2        ____   _____
_ 0022   1BC     1    2     ROCE-2        ____   _____
_ 0023   1BC     1    3     ROCE-2        ____   _____
_ 0024   1BC     1    4     ROCE-2        ____   _____
_ 0025   1BC     1    5     ROCE-2        ____   _____
_ 0026   1BC     1    6     ROCE-2        ____   _____
_ 0027   1BC     1    7     ROCE-2        ____   _____
_ 0028   1BC     1    8     ROCE-2        ____   _____
**************************** Bottom of data *********************************




 F1=Help        F2=Split      F3=Exit       F4=Prompt     F5=Reset      F7=Backward
 F8=Forward     F9=Swap       F10=Actions   F11=Add       F12=Cancel    F13=Instruct
F20=Right       F22=Command
```

SMC-Rv2 Requirements Hardware Configuration Definition

RoCE Express Virtualization

Use PF20=Right to scroll to the partition assignments for the displayed PCIe functions, one panel for each defined channel subsystem.

RoCE Adapter defined FID used by TCPIP config

```
     Goto   Filter   Backup   Query   Help
 ----------------------------------------------------------------------------
                              PCIe Function List        Row 1 of 16 More: <      >
 Command ===> _____  Scroll ===> CSR

 Select one or more PCIe functions, then press Enter. To add, use F11.
 Processor ID . . . . . : WSCCECK         WSC
 1=OS KOSP11      2=OS KOSP12     3=CF KOSP13     4=CO KOSP14     5=CO KOSP15
 6=CO KOSP16      7=CO KOSP17     8=CO KOSP18     9=CO KOSP19     A=CO KOSP1A
 B=OS KOSP1B      C=OS KOSP1C     D=OS KOSP1D     E=OS KOSP1E     F=CO KOSP1F
                              ------- Partitions 1x -------
 / FID     CHID    P    VF     1 2 3 4 5 6 7 8 9 A B C D E F
 _ 0011    1D8     1    1      _ _ _ _ _ _ _ _ _ _ _ _ _ _ _
 _ 0012    1D8     1    2      _ _ _ _ _ _ _ _ _ _ a _ _ _ _
 _ 0013    1D8     1    3      _ _ _ _ _ _ _ _ _ _ a _ _ _ _
 _ 0014    1D8     1    4      _ _ _ _ _ _ _ _ _ _ a _ _ _ _
 _ 0015    1D8     1    5      _ _ _ _ _ _ _ _ _ _ a _ _ _ _
 _ 0016    1D8     1    6      _ _ _ _ _ _ _ _ _ _ a _ _ _ _
 _ 0017    1D8     1    7      _ _ _ _ _ _ _ _ _ _ a _ _ _ _
 _ 0018    1D8     1    8      _ _ _ _ _ _ _ _ _ _ a _ _ _ _
 _ 0021    1BC     1    1      _ _ _ _ _ _ _ _ _ _   _ _ _ _
 _ 0022    1BC     1    2      _ _ _ _ _ _ _ _ _ _ a _ _ _ _
 _ 0023    1BC     1    3      _ _ _ _ _ _ _ _ _ _ a _ _ _ _
 _ 0024    1BC     1    4      _ _ _ _ _ _ _ _ _ _ a _ _ _ _
 _ 0025    1BC     1    5      _ _ _ _ _ _ _ _ _ _ a _ _ _ _
 _ 0026    1BC     1    6      _ _ _ _ _ _ _ _ _ _ a _ _ _ _
 _ 0027    1BC     1    7      _ _ _ _ _ _ _ _ _ _ a _ _ _ _
 _ 0028    1BC     1    8      _ _ _ _ _ _ _ _ _ _ a _ _ _ _
 ****************************** Bottom of data ******************************



   F1=Help        F2=Split        F3=Exit       F4=Prompt     F5=Reset        F7=Backward
   F8=Forward     F9=Swap         F10=Actions   F11=Add       F12=Cancel      F13=Instruct
 F19=Left         F20=Right       F22=Command
```

SMC-Rv2 Requirements Hardware Configuration Definition

Use F11=Add to define a new PCIe function

SMC-Rv2 Requirements Hardware Configuration Definition

Set PCIe Adapter Type to ROCE-2

```
       Goto    Filter   Backup   Query   Help
-----------------------------------------------------------------------
                           PCIe Function List
Command ===> _____ Scroll ===> CSR

S   ┌──────────────────── Add PCIe Function ────────────────────
    │
P   │
    │   Specify or revise the following values.
    │
/   │   Processor ID  . . . . :  WSCCECK        WSC
    │
_   │   Function ID . . . . . .  0030
_   │   Type  . . . . . . . . .  _____   +
_   │                         ┌──────────── Available PCIe Adapter Types ──────────
_   │   Channel ID . . . . .   │                                      Row 1 of 4
_   │   Port . . . . . . . .   │ Command ===> _____
_   │   Virtual Function ID    │
_   │   Number of virtual fun  │ Select one.
_   │   UID  . . . . . . . .   │
_   │                          │   PCIe Adapter Type
_   │   Description  . . . .   │   ISM
_   │                          │   NVME
_   │    F1=Help      F2=Split │ / ROCE-2
_   │    F12=Cancel            │   ZHYPERLINK
_   └──────────────────────    │ *************** Bottom of data ***************
_  0027    1BC    1    7       │
_  0028    1BC    1    8       │
***************************    │
                               │
                               │
                               │
                               │    F1=Help            F2=Split          F3=Exit
                               │    F7=Backward        F8=Forward        F9=Swap
                               │    F12=Cancel         F22=Command
    F1=Help       F2=Split     └─────────────────────────────────────────────────
    F8=Forward    F9=Swap      F10=Actions   F11=Add        F12=Cancel   F13=Instruct
```

SMC-Rv2 Requirements Hardware Configuration Definition

Function ID will match TCP/IP INTERFACE statement (for SMC-Rv2)

```
        Goto   Filter   Backup   Query   Help
-------------------------------------------------------------------------
                            PCIe Function List
Command ===> _____ Scroll ===> CSR

 S                          ── Add PCIe Function ──

 P
        Specify or revise the following values.

 /      Processor ID  . . . . . : WSCCECK        WSC

 _      Function ID . . . . . . . 0030
 _      Type  . . . . . . . . . . ROCE-2         +
 _
 _      Channel ID . . . . . . . . . . . . . 1BC   +
 _      Port . . . . . . . . . . . . . . . . 2     +
 _      Virtual Function ID  . . . . . . . 1       +
 _      Number of virtual functions  . . 1
 _      UID  . . . . . . . . . . . . . . . . ____
 _
 _      Description  . . . . . . . . . . _____
 _
 _       F1=Help      F2=Split     F3=Exit      F4=Prompt   F5=Reset    F9=Swap
 _       F12=Cancel
 _
 _  0027   1BC    1    7     ROCE-2        ____  _____
 _  0028   1BC    1    8     ROCE-2        ____  _____
 ******************************* Bottom of data *******************************




    F1=Help        F2=Split       F3=Exit       F4=Prompt    F5=Reset     F7=Backward
    F8=Forward     F9=Swap        F10=Actions   F11=Add      F12=Cancel   F13=Instruct
```

SMC-Rv2 Requirements Hardware Configuration Definition

NOTE: Each physical network ID entry correlates to a physical port on the adapter.

Associate the RoCE Express Adapter to an OSA interface with same PNetID

```
    Goto   Filter   Backup   Query   Help
-  ┌──────────────────── Add/Modify Physical Network IDs ────────────────┐
C  │                                                                      │
   │  If the Channel ID (CHID) is associated to one or more physical networks, │
S  │  specify each physical network ID corresponding to each applicable   │
   │  physical port.                                                      │
P  │                                                                      │
   │  Physical network ID 1  . .  _____                        │
   │  Physical network ID 2  . .  _____                        │
   │  Physical network ID 3  . .  _____                        │
/  │  Physical network ID 4  . .  _____                        │
   │                                                                      │
─  │                                                                      │
─  │                                                                      │
─  │    F1=Help       F2=Split      F3=Exit      F5=Reset     F9=Swap     F12=Cancel │
_  └──────────────────────────────────────────────────────────────────┘
_  0016    1D8    1   6    ROCE-2    ____  _____
_  0017    1D8    1   7    ROCE-2    ____  _____
_  0018    1D8    1   8    ROCE-2    ____  _____
_  0021    1BC    1   1    ROCE-2    ____  _____
_  0022    1BC    1   2    ROCE-2    ____  _____
_  0023    1BC    1   3    ROCE-2    ____  _____
_  0024    1BC    1   4    ROCE-2    ____  _____
_  0025    1BC    1   5    ROCE-2    ____  _____
_  0026    1BC    1   6    ROCE-2    ____  _____
_  0027    1BC    1   7    ROCE-2    ____  _____
_  0028    1BC    1   8    ROCE-2    ____  _____
******************************* Bottom of data *******************************

F1=Help         F2=Split        F3=Exit         F4=Prompt      F5=Reset       F7=Backward
F8=Forward      F9=Swap         F10=Actions     F11=Add        F12=Cancel     F13=Instruct
```

SMC-Rv2 Requirements IOCDS Configuration

Sample RoCE IOCP / IOCDS declaration

SMC-Rv2 Requirements
Software Prerequisites

Software
– z/OS
  • IBM z/OS V2R5
  • Define at least one user-defined Enterprise ID in z/OS! (disabled by default)

Environments
– LPAR   yes
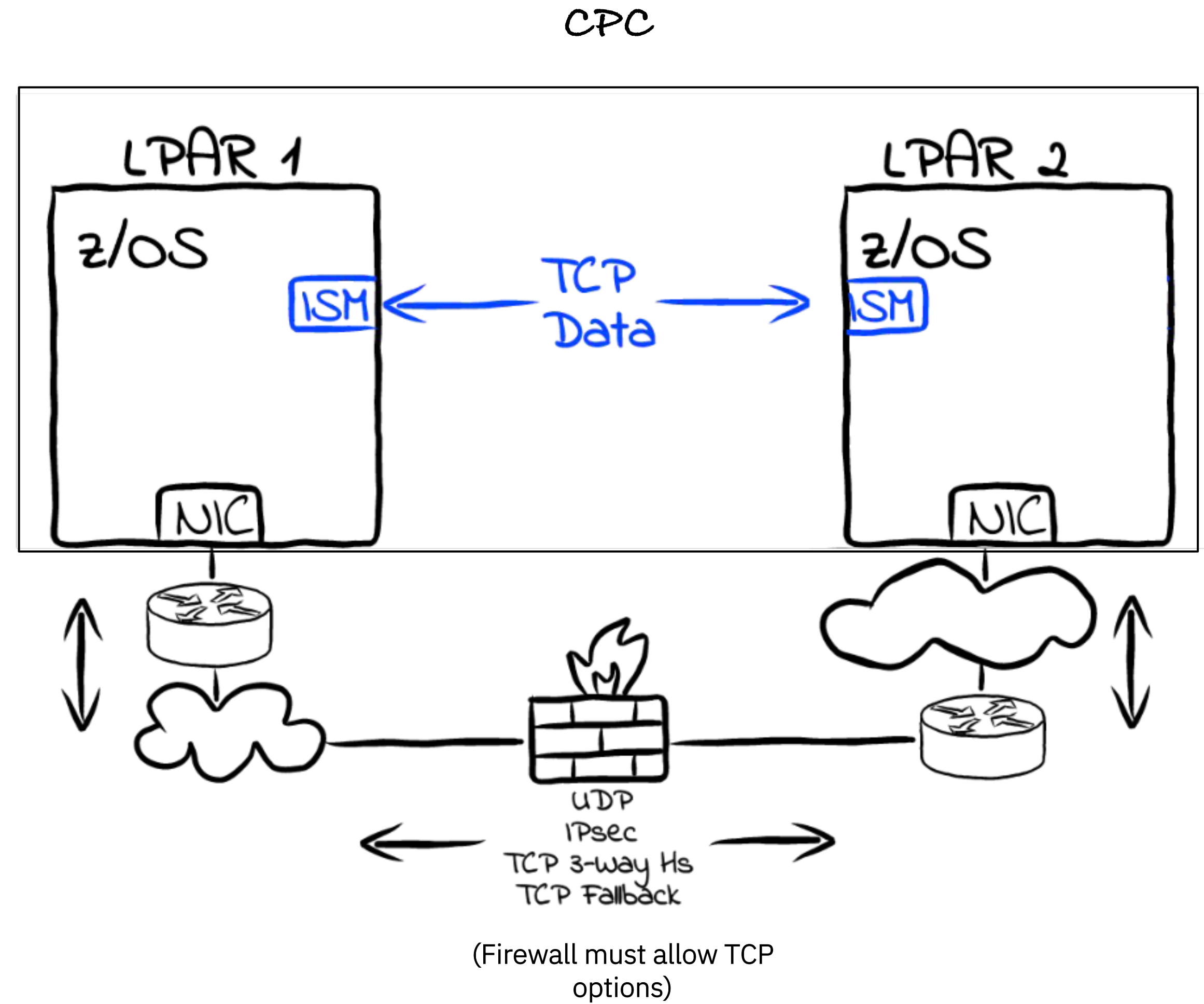– z/VM guests   yes

SMC-Dv2 Requirements
Hardware Prerequisites

IBM Z hardware
requirements
– IBM z15 or z16

ISMv2 devices
– Virtual PCI network
   adapter of VCHID
   type ISM

CPC

LPAR 1

z/OS

ISM ← TCP Data → ISM

LPAR 2

z/OS

NIC

NIC

UDP
IPsec
TCP 3-way Hs
TCP Fallback

(Firewall must allow TCP
options)

SMC-Dv2 Requirements

Hardware Prerequisites

Setting up the environment for Shared Memory Communications – Direct Memory Access

Define ISM VCHIDs:

- (Recommended) Without a PNetID in the HCD

- With a PNetID in the hardware configuration definition (HCD) that matches the PNetID definitions for any OSA or HiperSockets devices that will use the ISM device for SMC-D (SMC-Dv1 compatibility)

SMC-Dv2 Requirements Hardware Configuration Definition

Use F11=Add to define a new PCIe function

```
    Goto   Filter   Backup   Query   Help
 --------------------------------------------------------------------
                            PCIe Function List      Row 1 of 16 More:        >
 Command ===> _____    Scroll ===> CSR

 Select one or more PCIe functions, then press Enter. To add, use F11.

 Processor ID . . . . . : WSCCECK


 / FID    CHID+   P+   VF+   Type+          UID    Description
 _ 0011   1D8     1    1     ROCE-2         ____   _____
 _ 0012   1D8     1    2     ROCE-2         ____   _____
 _ 0013   1D8     1    3     ROCE-2         ____   _____
 _ 0014   1D8     1    4     ROCE-2         ____   _____
 _ 0015   1D8     1    5     ROCE-2         ____   _____
 _ 0016   1D8     1    6     ROCE-2         ____   _____
 _ 0017   1D8     1    7     ROCE-2         ____   _____
 _ 0018   1D8     1    8     ROCE-2         ____   _____
 _ 0021   1BC     1    1     ROCE-2         ____   _____
 _ 0022   1BC     1    2     ROCE-2         ____   _____
 _ 0023   1BC     1    3     ROCE-2         ____   _____
 _ 0024   1BC     1    4     ROCE-2         ____   _____
 _ 0025   1BC     1    5     ROCE-2         ____   _____
 _ 0026   1BC     1    6     ROCE-2         ____   _____
 _ 0027   1BC     1    7     ROCE-2         ____   _____
 _ 0028   1BC     1    8     ROCE-2         ____   _____
 ***************************** Bottom of data *******************************




 F1=Help          F2=Split        F3=Exit         F4=Prompt      F5=Reset        F7=Backward
 F8=Forward       F9=Swap         F10=Actions     F11=Add        F12=Cancel      F13=Instruct
 F20=Right        F22=Command
```

SMC-Dv2 Requirements Hardware Configuration Definition

Add a virtual channel ID

SMC-Dv2 Requirements Hardware Configuration Definition

Use F11=Add to define a new PCIe function

```
   Goto   Filter   Backup   Query   Help
---------------------------------------------------------------------------
                          PCIe Function List        Row 1 of 18 More: <      >
Command ===> _____  Scroll ===> CSR

Select one or more PCIe functions, then press Enter. To add, use F11.
Processor ID . . . . : WSCCECK
1=OS KOSP01      2=OS KOSP02     3=CF KOSP03     4=CO KOSP04     5=CO KOSP05
6=CO KOSP06      7=CO KOSP07     8=CO KOSP08     9=OS KOSP09     A=OS KOSP0A
B=CO KOSP0B      C=CO KOSP0C     D=CO KOSP0D     E=OS KOSP0E     F=OS KOSP0F
                          ------- Partitions 0x -------
/ FID     CHID    P    VF    1 2 3 4 5 6 7 8 9 A B C D E F
_ 0011    1D8     1    1     a _ _ _ _ _ _ _ _ _ _ _ _ _ _
_ 0012    1D8     1    2     _ _ _ _ _ _ _ _ _ _ _ _ _ _ _
_ 0013    1D8     1    3     _ _ _ _ _ _ _ _ _ _ _ _ _ _ _
_ 0014    1D8     1    4     _ _ _ _ _ _ _ _ _ _ _ _ _ _ _
_ 0015    1D8     1    5     _ _ _ _ _ _ _ _ _ _ _ _ _ _ _
_ 0016    1D8     1    6     _ _ _ _ _ _ _ _ _ _ _ _ _ _ _
_ 0017    1D8     1    7     _ _ _ _ _ _ _ _ _ _ _ _ _ _ _
_ 0018    1D8     1    8     _ _ _ _ _ _ _ _ _ _ _ _ _ _ _
_ 0021    1BC     1    1     a _ _ _ _ _ _ _ _ _ _ _ _ _ _
_ 0022    1BC     1    2     _ _ _ _ _ _ _ _ _ _ _ _ _ _ _
_ 0023    1BC     1    3     _ _ _ _ _ _ _ _ _ _ _ _ _ _ _
_ 0024    1BC     1    4     _ _ _ _ _ _ _ _ _ _ _ _ _ _ _
_ 0025    1BC     1    5     _ _ _ _ _ _ _ _ _ _ _ _ _ _ _
_ 0026    1BC     1    6     _ _ _ _ _ _ _ _ _ _ _ _ _ _ _
_ 0027    1BC     1    7     _ _ _ _ _ _ _ _ _ _ _ _ _ _ _
_ 0028    1BC     1    8     _ _ _ _ _ _ _ _ _ _ _ _ _ _ _
_ 0030    1BC     2    1     a _ _ _ _ _ _ _ _ _ _ _ _ _ _
_ 0031    7C2          1     a _ _ _ _ _ _ _ _ _ _ _ _ _ _
****************************** Bottom of data ******************************



 F1=Help        F2=Split       F3=Exit      F4=Prompt     F5=Reset      F7=Backward
 F8=Forward     F9=Swap        F10=Actions  F11=Add       F12=Cancel    F13=Instruct
F19=Left        F20=Right      F22=Command
```

SMC-Dv2 Requirements IOCDS Configuration

Sample RoCE IOCP / IOCDS declaration

SMC-Dv2 Requirements
Software Prerequisites

Software
- z/OS
  - IBM z/OS V2R54 (via APAR) or later
  - Define at least one user-defined Enterprise ID in z/OS! (disabled by default)

Environments
- LPAR   yes
- z/VM guests   yes

SMCv2 z/OS TCP/IP Profile Configuration
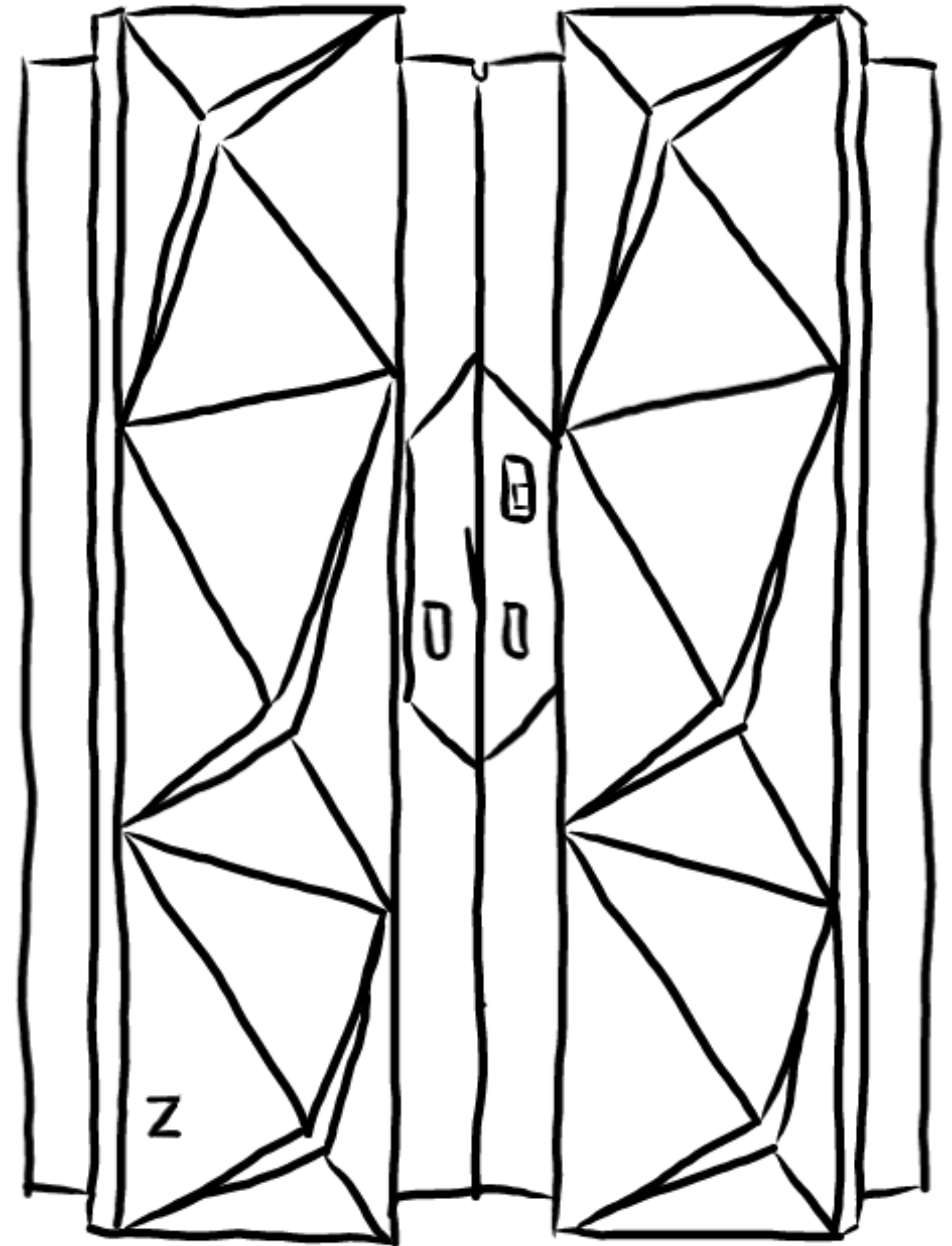
SMCv2 Setup & Verification
SMC-Rv2 Setup

z/OS TCP/IP profile configuration requirements
– Configure GLOBALCONFIG **SMCR** statement
– Add **EID** on SMCGLOBAL statement
– Define **SMCR, RoCEv2 IP Address (SMCRIPADDR)** and **PFID** (configured on OSA INTERFACE statement)

Start the interfaces (RNIC is automatically activated)

```
GLOBALCONFIG
    SMCGLOBAL SMCEID user_EID_values ENDSMCEID
    SMCR
…
INTERFACE intf_name DEFINE IPAQENET CHPIDTYPE OSD
    PORTNAME port_name IPADDR ipv4/num_mask_bits
    SMCR PFID pfid SMCRIPADDR ipv4addr SMCRMTU mtusize
```

SMCRIPADDR must be in the same subnet as IP address of the OSA-Express QDIO INTERFACE statement

SMCRMTU is optional

PFID must be RoCE Express2 or RoCE Express3 (z16)

SMCv2 Setup & Verification
SMC-Dv2 Setup

z/OS TCP/IP profile configuration requirements

– Configure GLOBALCONFIG **SMCD** statement

– Add **SYSTEMEID** or **User-defined EID** on SMCGLOBAL statement
  • SystemEID is specified on GLOBALCONFIG SMCD statement
  • User-defined EIDs are specified on SMCGLOBAL

– Define **SMCD** and **IP Address (IPADDR)** with a subnet mask
  • Using HiperSockets: configured on IPAQIDIO INTERFACE STATEMENT
  • Using OSA: configured on IPAQENET INTERFACE statement. Also include **CHPIDTYPE OSD**

Start the OSA and HiperSocket interfaces (ISM interfaces are automatically started)

SMCV2 Setup & Verification
SMC-Dv2 Setup

z/OS TCP/IP profile
configuration
requirements

```
GLOBALCONFIG SMCD
    SMCGLOBAL SMCEID user_EID_values ENDSMCEID
or
GLOBALCONFIG SMCD SYSTEMEID
…
INTERFACE intf_name DEFINE IPAQIDIO SMCD
      IPADDR ipv4addr/num_mask_bits
…
INTERFACE intf_name DEFINE IPAQENET SMCD
      CHPIDTYPE OSD
      IPADDR ipv4addr/num_mask_bits
```

Either a SYSTEMEID or user-defined EID enables SMC-Dv2 for this TCP/IP stack

For HiperSocket interfaces

For OSD interfaces

SMCv2 Setup & Verification
Verification

z/OS TCP/IP profile configuration SYNTAXCHECK
- Can be used to validate SMCR profile changes before use
- Run against profile data set

```
Format:
V TCPIP,procname,SYNTAXCHECK,datasetname

Example:
V TCPIP,TCPIPB,SYNTAXCHECK,DSN=SYS1.TCPPARMS(PROFSYSB)
EZZ0060I PROCESSING COMMAND: VARY TCPIP,TCPIPB,SYNTAXCHECK,DSN=SYS1.TCPP
 ARMS(PROFSYSB)
 EZZ0061I VARY SYNTAXCHECK COMMAND BEGINNING
…
 EZZ0316I PROFILE PROCESSING COMPLETE FOR FILE 'SYS1.TCPPARMS(PROFSYSB)'
 EZZ0062I VARY SYNTAXCHECK FOUND NO ERRORS
 EZZ0065I VARY SYNTAXCHECK COMMAND COMPLETE
```

SMCv2 Setup & Verification
Verification

SMC-R Monitoring – DISPLAY PCIE
- Activation of first SMC-R capable OSD causes PFIDs to be activated
- Use DISPLAY PCIE to display defined PFIDs

```
Format:
D PCIE


Example:
D PCIE
IQP022I 19.06.25 DISPLAY PCIE 545
PCIE       0013 ACTIVE
PFID      DEVICE TYPE NAME            STATUS   ASID   JOBNAME   CHID VFN  PN
00000111  ISM                        ALLC     002A   NET       07C2 0001
00000011  RoCE Express3              ALLC     002A   NET       01D8 0001 1
00000021  RoCE Express3              CNFG                      01BC 0001 1
```

RoCE ready for use on this LPAR

If the devices are in standby, issue a system command "CF PFID(pfid),ONLINE"

SMCV2 Setup & Verification

Verification

Netstat Devlinks/-d for OSD report

– SMC V1 and V2 capabilities
– If V2 capable:
  • Displays configured PFID, SMCRIPADDR, and SMCRMTU

```
Format:
D TCPIP,procname,NETSTAT,DEVlinks

Example:
D TCPIP,TCPIPB,NETSTAT,DEVLINKS
…
IntfName: GIG3 IntfType: IPAQENET IntfStatus: Ready
   PortName: GIG2E Datapath: OBE2 DatapathStatus: Ready
   CHPIDType: OSD      SMCR: Yes
   PNetID: IBMTRUNKSET1  SMCD: Yes
...
SMCR Capability: V2
   Associated Multi-subnet RNIC interface: EZARIUT10011
      RoCE PFID: 0011 SMCRMTU: 1024
      SMCRIPADDR: 9.82.24.157
   Unassociated ISM interfaces: EZAISMU1
IntfName: EZARIUT10011   IntfType: RNIC IntfStatus: READY
```

SMCV2 Setup & Verification
Verification

Netstat CONFIG/-f report
– Displays:
  - User-defined EIDs
  - SMC exclude and permit lists

```
Format:
D TCPIP,procname,NETSTAT,CONFIG

Example:
D TCPIP,TCPIPB,NETSTAT,CONFIG
…
Global Configuration Information:
SMCGlobal:
    AutoCache: YES AutoSMC: YES
    SMCEID: IBM-COMMSERVER-MULTIPASS
    SMC Exclude List: NONE
    SMC Permit List: ALL
SMCR: YES
    FixedMemory: 0000256M TcpKeepMinInt: 00000300
SMCD: YES
    FixedMemory: 0000256M TcpKeepMinInt: 00000300
    SMCDSYSEID: YES (IBM-SYSZ-ISMSEID0000000023B83931)
```

SMCv2 Setup & Verification
Verification

Netstat ALL/-a report
– Displays SMC version and EID

```
D TCPIP,TCPIPD,N,ALL,IPPORT=9.82.24.151+21
EZD0101I NETSTAT CS V2R4 TCPIPD 331
CLIENT NAME: FTPSERVE                          CLIENT ID: 00000CC2
   LOCAL SOCKET: ::FFFF:9.82.24.153..21
   FOREIGN SOCKET: ::FFFF:9.82.24.151..1242
…
  SMC INFORMATION:
       SMCDSTATUS:        ACTIVE               SMCDVERSION:         2
          LOCALSMCLINKID: 02010001             REMOTESMCLINKID:     01010001
          LOCALSMCRCVBUF: 64K                  REMOTESMCRCVBUF:     64K
          SMCEID:         IBM-SYSZ-ISMSEID0000000023B83931
```
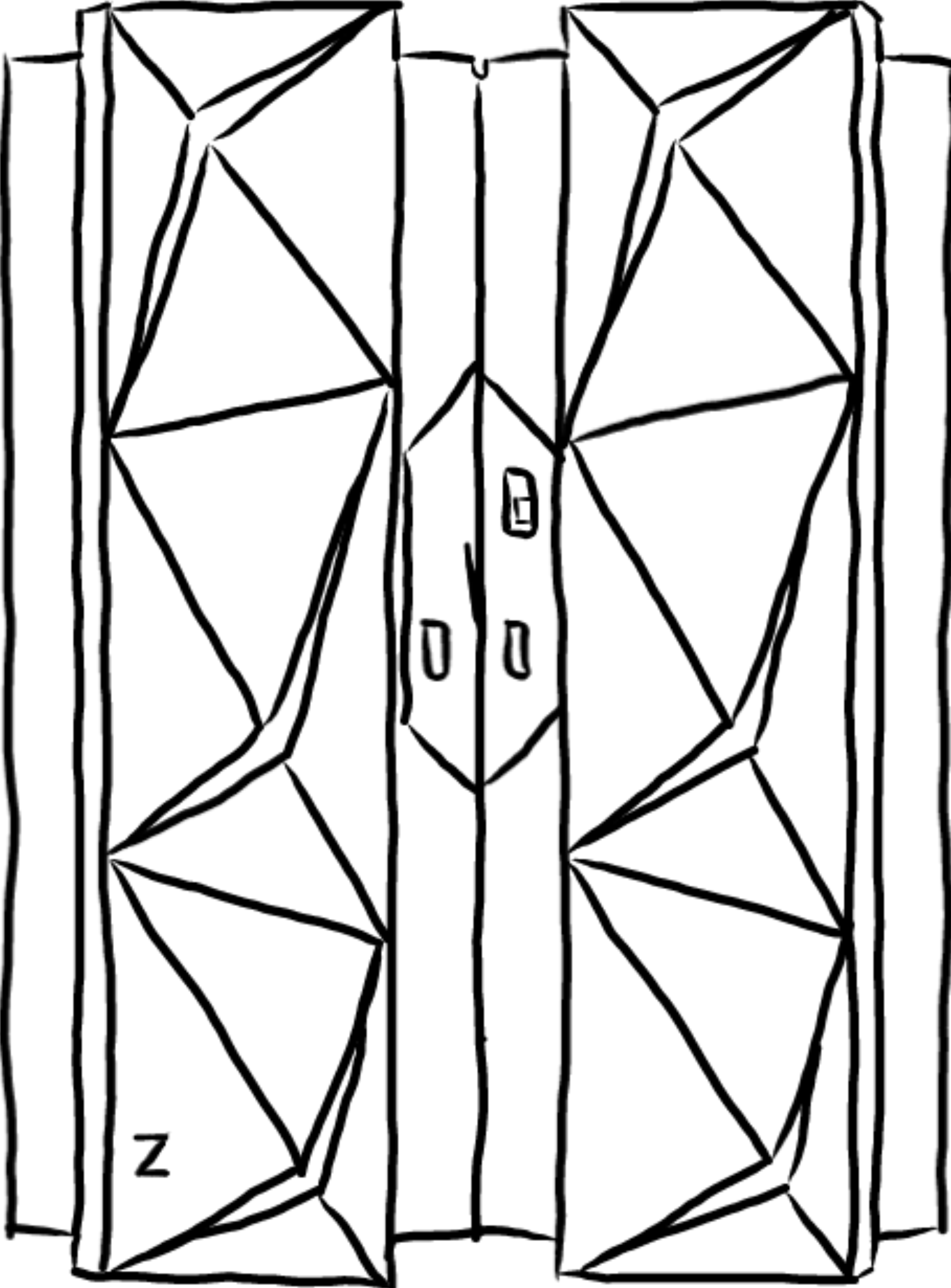
Connection is using SMC-Dv2!

Migrating to SMCv2

SMCv2 Migration
Migrating from SMC-Rv1 to SMC-Rv2

Key things:
- SMC-Rv1 communications are disabled when the RoCE PFID is not configured on the GLOBALCONFIG.

- PFIDs are required on the OSA INTERFACE statement under the SMCR parameter.

- When an EID is enabled on GLOBALCONFIG, it enables SMC-Rv2 for the TCP/IP stack.

- The SMC-Rv2 enabled peers must share a common EID.

If you must continue to communicate with down-level SMC-R peer hosts which only support SMC-Rv1 (over the same IP subnet)
1. Configure the same FID value defined on the OSA INTERFACE SMCR PFID parameter on the GLOBALCONFIG SMCR statement
2. The SMC-Rv2 enabled peers must be attached to the same IP subnet.

SMCv2 Migration
Migrating from SMC-Rv1 to SMC-Rv2

z/OS TCP/IP profile configuration requirements
- Removing **PFID** on GLOBALCONFIG **SMCR** statement disables SMC-Rv1

- Add a user-defined EID on the the SMCGlobal parameter of GLOBALCONFIG

- Defining **PFID** and **SMCRIPADDR** on the OSA INTERFACE statement under the SMCR parameter is required for SMC-Rv2

```
GLOBALCONFIG
    SMCGLOBAL SMCEID user_EID_values ENDSMCEID
    SMCR
…
INTERFACE intf_name DEFINE IPAQENET CHPIDTYPE OSD
    PORTNAME port_name IPADDR ipv4/num_mask_bits
    SMCR PFID pfid SMCRIPADDR ipv4addr SMCRMTU mtusize
```

If communicating with SMC-Rv1 hosts, define a PFID on GLOBALCONFIG statement.
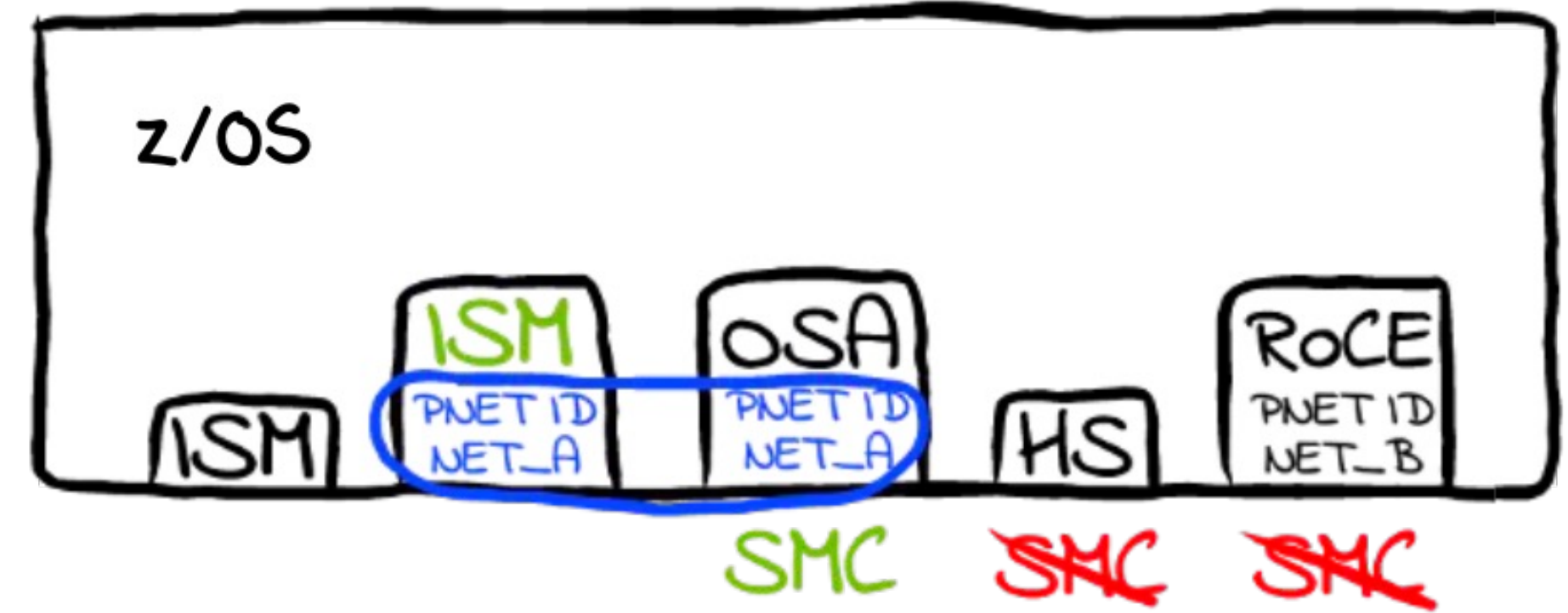
SMCv2 Migration
Migrating from SMC-Dv1 to SMC-Dv2

Define ISM VCHIDs without a PNetID in HCD

If you must continue to communicate with down-level SMC-D peer hosts which only support SMC-Dv1

1. Define ISM VCHIDs with a PNetID that matches your OSA or HiperSockets PNetID definitions.

2. The SMC-D enabled peers must be attached to the same IP subnet.

SMC-Dv1: Only interfaces with matching PNetIDs are enabled for SMC-D.
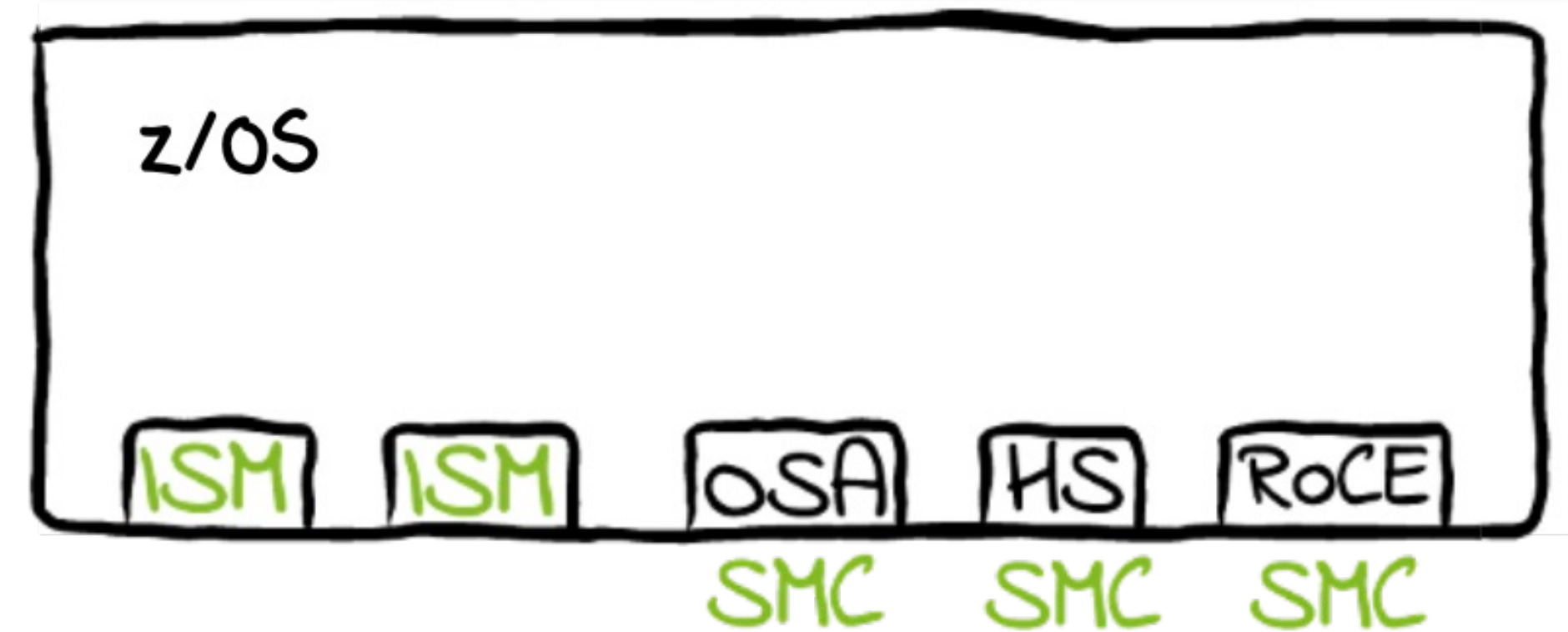
SMC-Dv2: Any interface is enabled for SMC-D – no PNetID required!

SMCv2 Migration
Migrating from SMC-Dv1 to SMC-Dv2

z/OS TCP/IP profile configuration requirements
- Add **SYSTEMEID** or a **User-defined EID** on SMCGLOBAL statement
- Use DISPLAY PCIE to display defined PFIDs
  - SYSTEMEID is specified on GLOBALCONFIG SMCD statement
  - User-defined EIDs are specified on SMCGLOBAL

```
GLOBALCONFIG SMCD SYSTEMEID
or
GLOBALCONFIG SMCD
    SMCGLOBAL SMCEID user_EID_values ENDSMCEID
…
INTERFACE intf_name DEFINE IPAQIDIO SMCD
       IPADDR ipv4addr/num_mask_bits
…
INTERFACE intf_name DEFINE IPAQENET SMCD
       CHPIDTYPE OSD
       IPADDR ipv4addr/num_mask_bits
```

Defining EID enables SMC-Dv2

SMCD parameters on INTERFACE statements remains the same

Use Cases
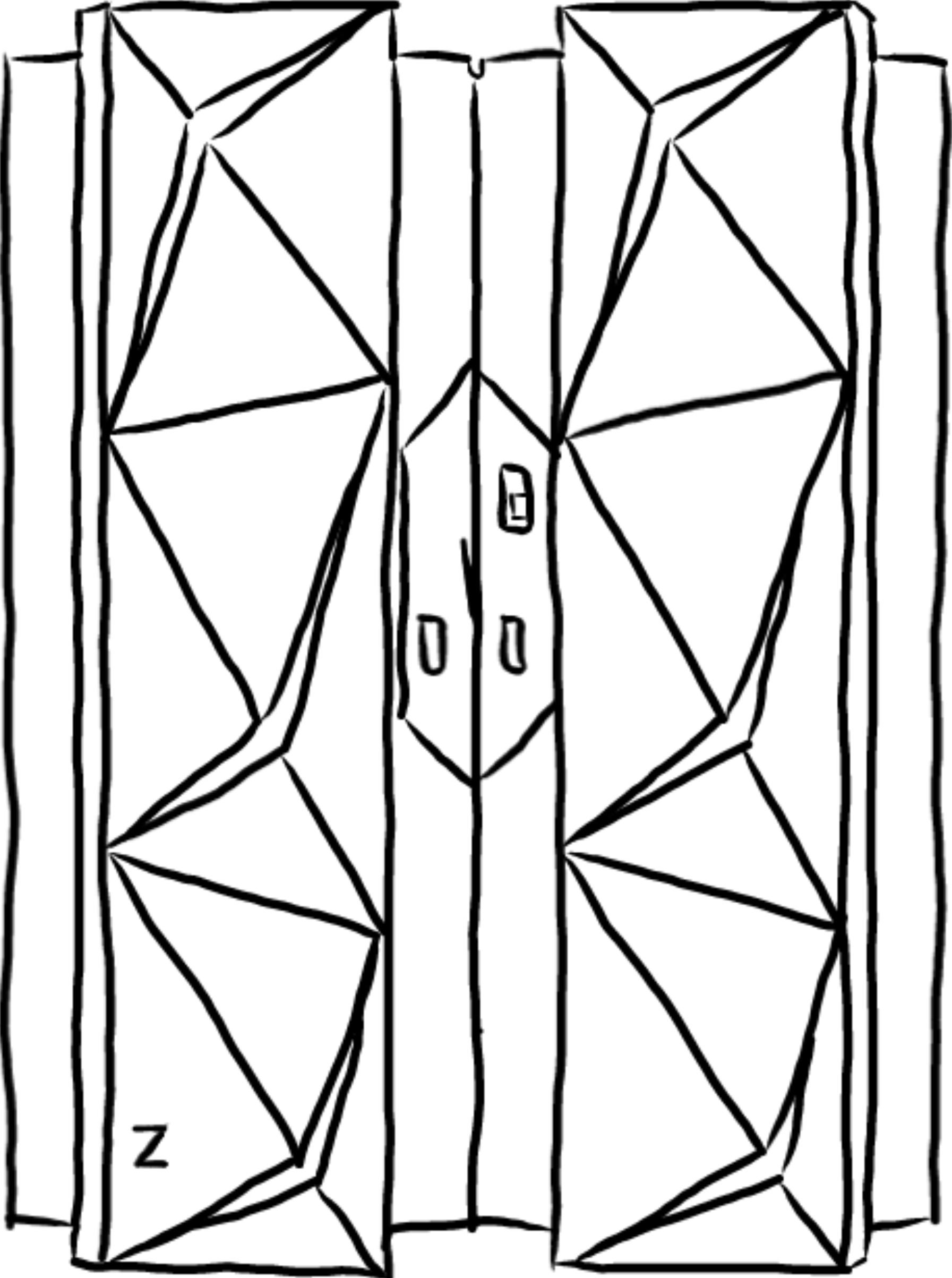
Configuring with both v1 and v2 peer hosts

Use Cases
Configuring with SMC-Rv1 and SMC-Rv2 peer hosts

WSC (SMC-Rv1)

ATS1 (SMC-Rv2)
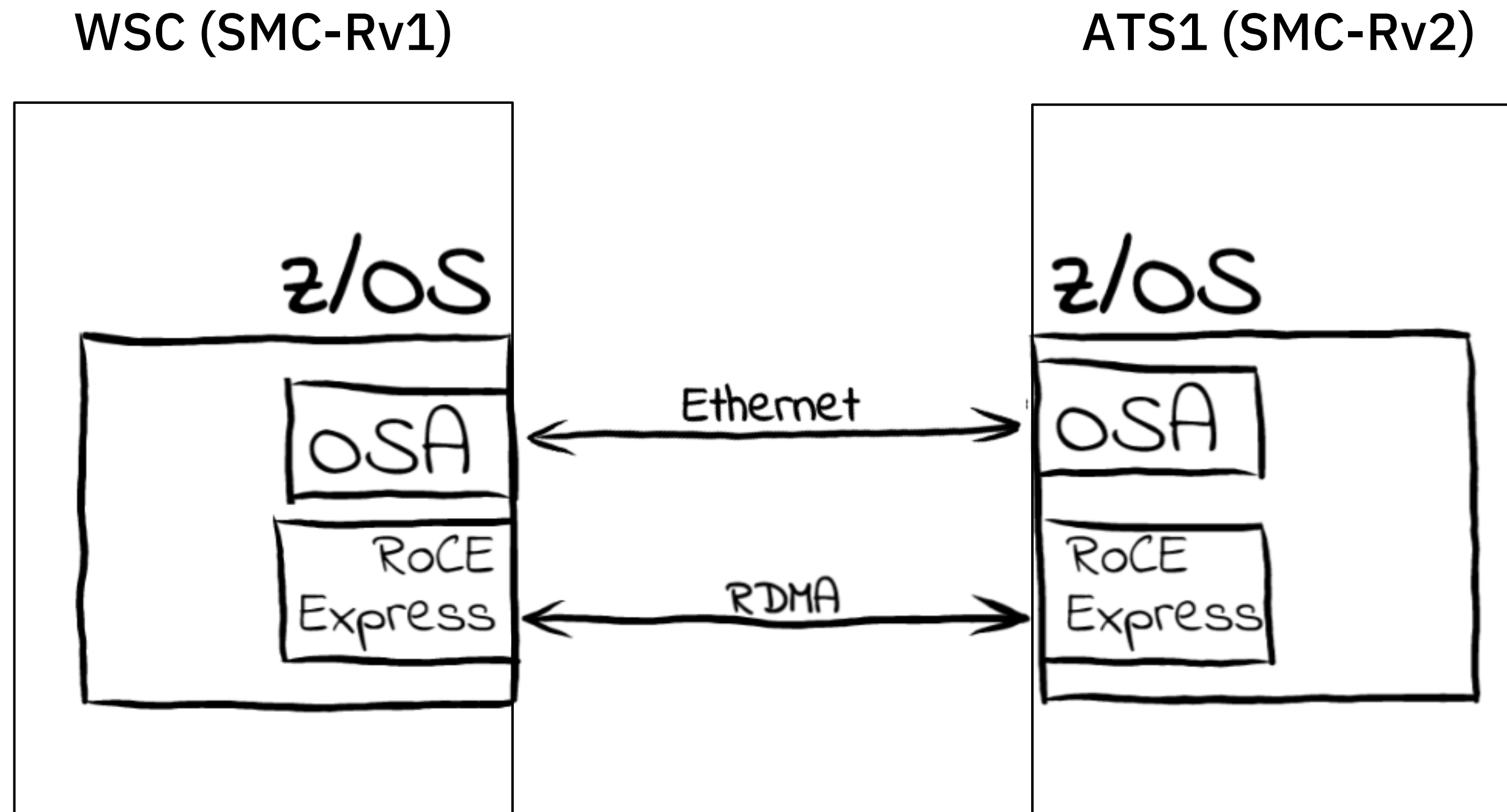
- – Both peers must be on the same IP subnet

- – Include PFIDs on GLOBALCONFIG statement

OSA IPADDR: 9.82.24.151/24
SMCRIPADDR: 9.82.24.157
PNetID: IBMTRUNKSET1
PFID: 00000011

OSA IPADDR: 9.82.24.201/24 (same subnet)
SMCRIPADDR: 9.82.24.207 (same subnet as OSA)
PNetID: NETID1
PFID: 00000101

Use Cases
Configuring with SMC-Rv1 and SMC-Rv2 peer hosts

WSC z/OS LPAR TCP/IP Profile Configuration (SMC-Rv1):

```
GLOBALCONFIG
    SMCR PFID 00000011 PORTNum 1
…
INTERFACE GIG3 DEFINE IPAQENET CHPIDTYPE OSD
    PORTNAME GIG2E IPADDR 9.82.24.151/24
    SMCR
```

WSC z/OS LPAR
Hardware Configuration:

```
OSA PNetID: IBMTRUNKSET1
RoCE PNetID: IBMTRUNKSET1
ROCE PFID: 00000011
```

ATS1 z/OS LPAR TCP/IP Profile Configuration (SMC-Rv2):

```
GLOBALCONFIG
    SMCGLOBAL SMCEID IBM-COMMSERVER-MULTIPASS ENDSMCEID
    SMCR PFID 00000101 PORTNUM 1
…
INTERFACE OSAQDIO24 DEFINE IPAQENET CHPIDTYPE OSD
    PORTNAME OSAQDIO2 IPADDR 9.82.24.201/24
    SMCR PFID 00000101 SMCRIPADDR 9.82.24.207 SMCRMTU 1024
```

ATS1 z/OS LPAR
Hardware Configuration:

```
OSA PNetID: NETID1
RoCE PNetID: NETID1
ROCE PFID: 00000101
```

Use Cases
Configuring with SMC-Dv1 and SMC-Dv2 peer hosts
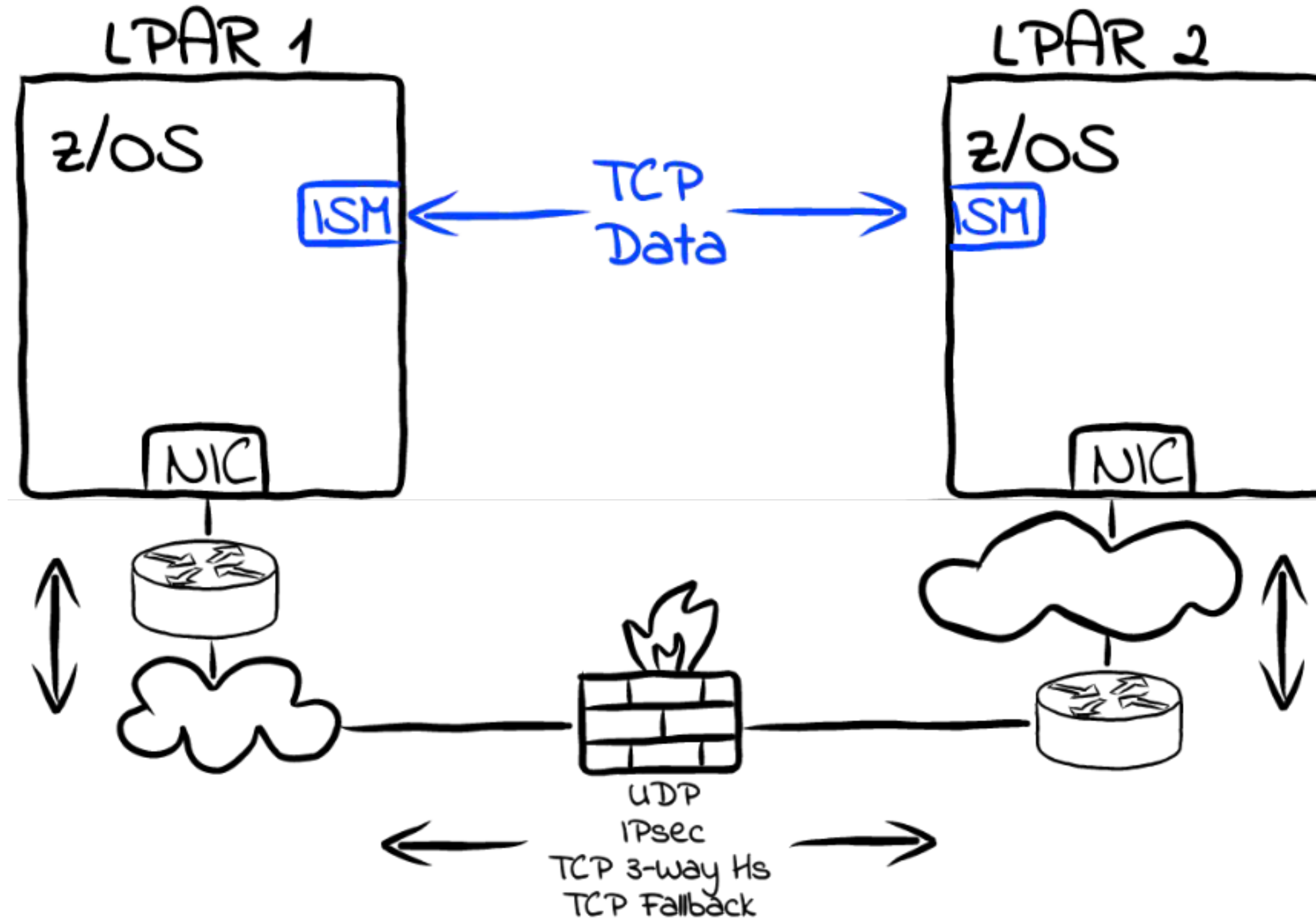
- PNetIDs must be assigned to both OSA (or HiperSockets) and ISM device

- The SMC peers must be within the same IP subnet

Note: OSA, HiperSockets, and ISM devices (all three) should not be assigned the same PNetID.

LPAR #1 OSA PNetID: IBMTRUNKSET2
ISM 1 PNetID: IBMTRUNKSET2
ISM 1 VCHID: 07C2

LPAR #2 OSA PNetID: IBMTRUNKSET2
ISM 2 PNetID: IBMTRUNKSET2
ISM 2 VCHID: 07C2

Use Cases
Configuring with SMC-Dv1 and SMC-Dv2 peer hosts

LPAR #1 TCP/IP Profile Configuration (SMC-Dv1):

```
GLOBALCONFIG SMCD
…
INTERFACE GIG3 DEFINE IPAQENET CHPIDTYPE OSD
    PORTNAME GIG2E IPADDR 9.82.24.158/24
    SMCD
```

LPAR #1
Hardware Configuration:

```
OSA PNetID: IBMTRUNKSET2
ISM PNetID: IBMTRUNKSET2
ISM VCHID: 07C2
```

LPAR #2 TCP/IP Profile Configuration (SMC-Dv2):

```
GLOBALCONFIG SMCD SYSTEMEID
…
INTERFACE GIG4 DEFINE IPAQENET CHPIDTYPE OSD
    PORTNAME GIG3E IPADDR 9.82.24.160/24
    SMCD
```

LPAR #2
Hardware Configuration:

```
OSA PNetID: IBMTRUNKSET2
ISM PNetID: IBMTRUNKSET2
ISM VCHID: 07C2
```

Configuring with v2 peer hosts

Use Cases
Configuring with SMC-Rv2 peer hosts

WSC

ATS1
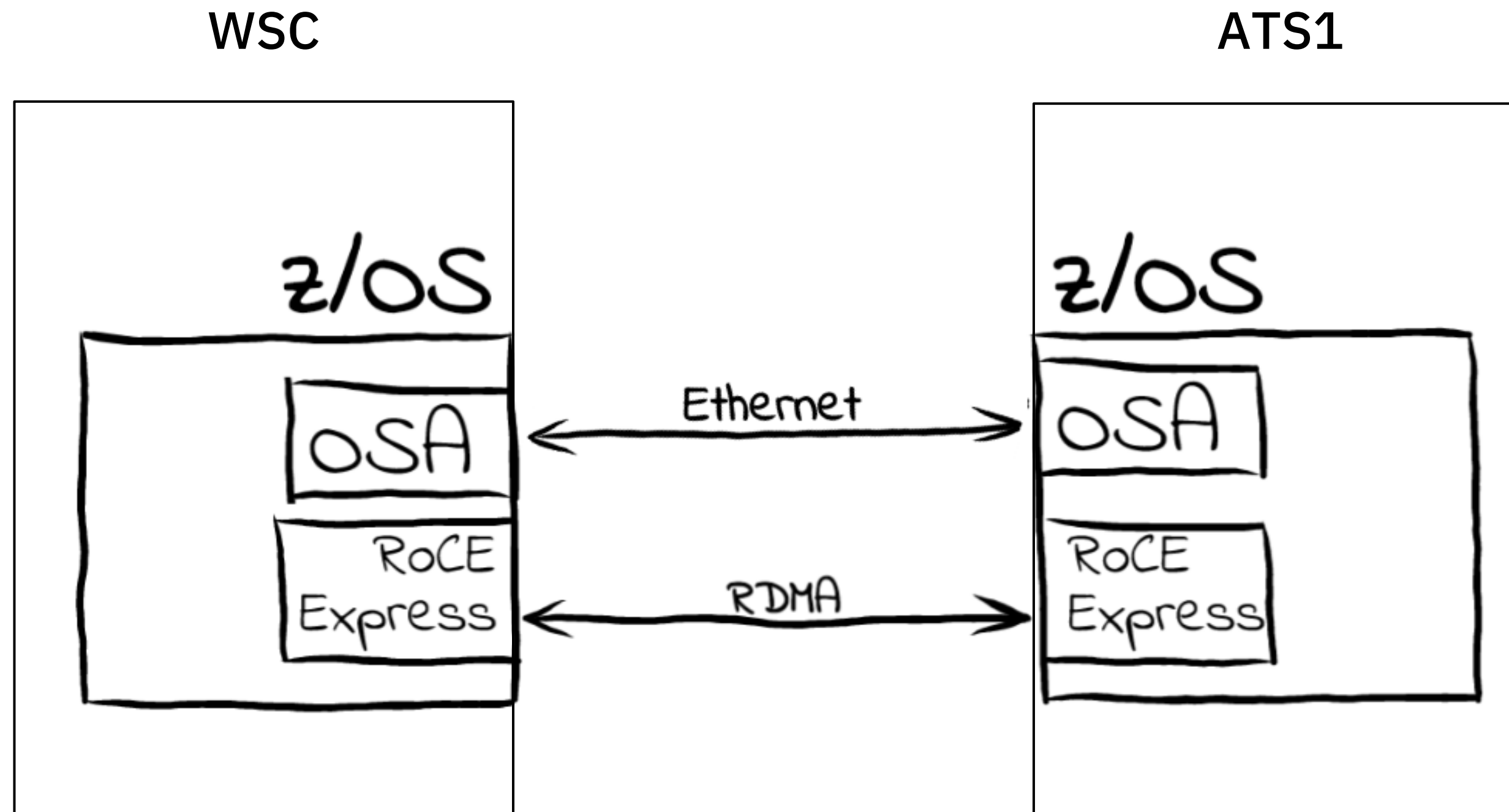
- Peers can be on different IP subnets

- Include PFIDs on INTERFACE statement

- Peers must configure the same EID

OSA IPADDR: 9.82.24.151/24
SMCRIPADDR: 9.82.24.157
PNetID: IBMTRUNKSET1
PFID: 00000011

OSA IPADDR: 9.82.17.30/24 (different subnet)
SMCRIPADDR: 9.82.17.15 (same subnet as OSA)
PNetID: NETID1
PFID: 00000101

Use Cases
Configuring with SMC-Rv2 peer hosts

WSC z/OS LPAR TCP/IP Profile Configuration:

```
GLOBALCONFIG
    SMCGLOBAL SMCEID IBM-COMMSERVER-MULTIPASS ENDSMCEID
    SMCR
…
INTERFACE GIG3 DEFINE IPAQENET CHPIDTYPE OSD
    PORTNAME GIG2E IPADDR 9.82.24.151/24
    SMCR PFID 00000011 SMCRIPADDR 9.82.24.157 SMCRMTU 1024
```

ATS1 z/OS LPAR TCP/IP Profile Configuration:

```
GLOBALCONFIG
    SMCGLOBAL SMCEID IBM-COMMSERVER-MULTIPASS ENDSMCEID
    SMCR
…
INTERFACE OSAQDIO24 DEFINE IPAQENET CHPIDTYPE OSD
    PORTNAME OSAQDIO2 IPADDR 9.82.17.30/24
    SMCR PFID 00000101 SMCRIPADDR 9.82.17.15 SMCRMTU 1024
```

WSC z/OS LPAR
Hardware Configuration:

```
OSA PNetID: IBMTRUNKSET1
RoCE PNetID: IBMTRUNKSET1
ROCE PFID: 00000011
```

ATS1 z/OS LPAR
Hardware Configuration:

```
OSA PNetID: NETID1
RoCE PNetID: NETID1
ROCE PFID: 00000101
```

Use Cases
Configuring with SMC-Dv2 peer hosts
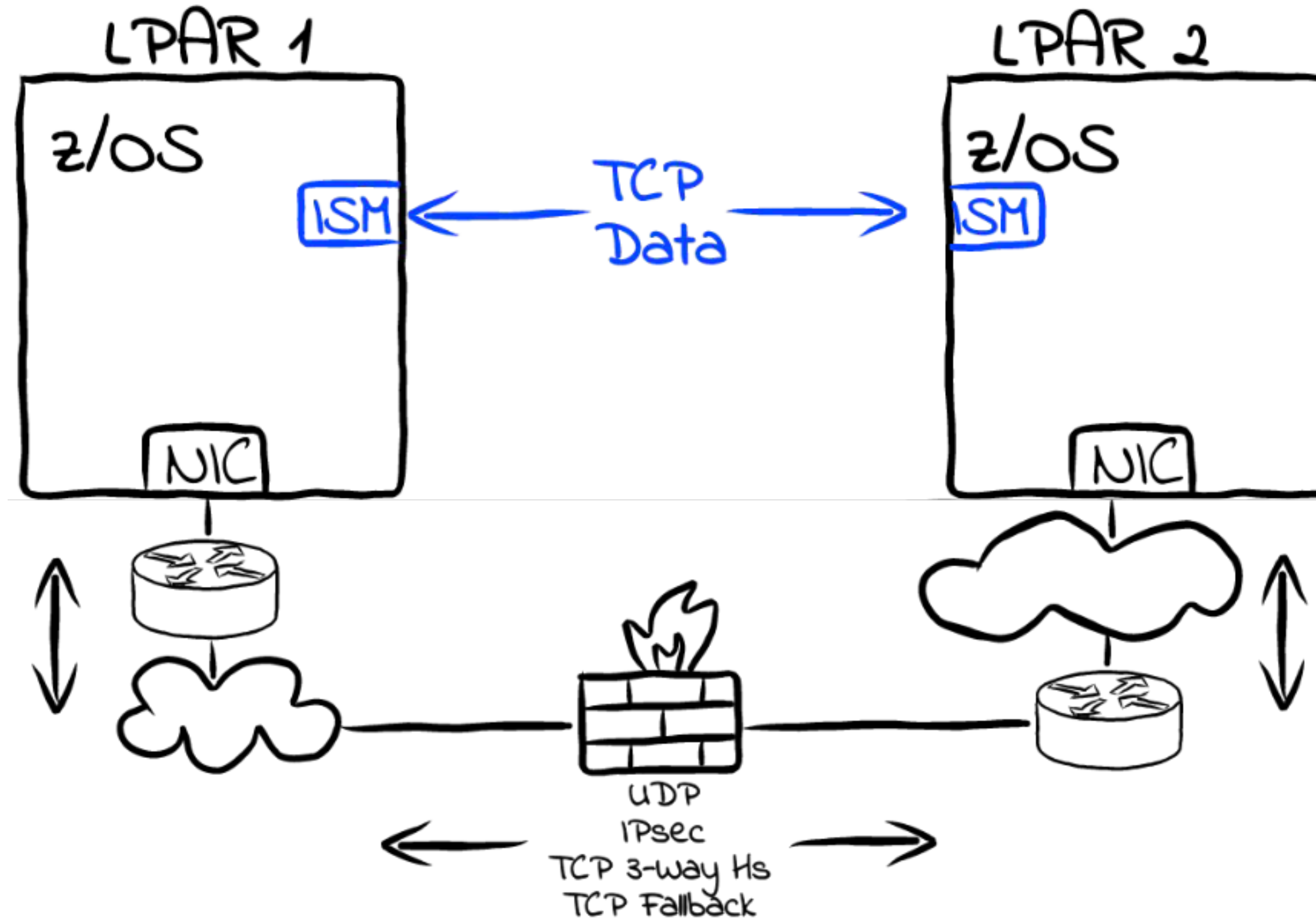
- PNetIDs are not required or defined

- The SMC peers can be on different IP subnets

Note: OSA, HiperSockets, and ISM devices (all three) should not be assigned the same PNetID.

LPAR #1 OSA PNetID: IBMTRUNKSET2
ISM 1 PNetID: NONE
ISM 1 VCHID: 07C2

LPAR #2 OSA PNetID: IBMTRUNKSET2
ISM 2 PNetID: NONE
ISM 2 VCHID: 07C2

Use Cases
Configuring with SMC-Dv2 peer hosts

LPAR #1 TCP/IP Profile Configuration

LPAR #1
Hardware Configuration:

```
GLOBALCONFIG SMCD SYSTEMEID
…
INTERFACE GIG3 DEFINE IPAQENET CHPIDTYPE OSD
    PORTNAME GIG2E IPADDR 9.82.24.151/24
    SMCD
```

```
OSA PNetID: IBMTRUNKSET2
ISM PNetID: NONE
ISM VCHID: 07C2
```

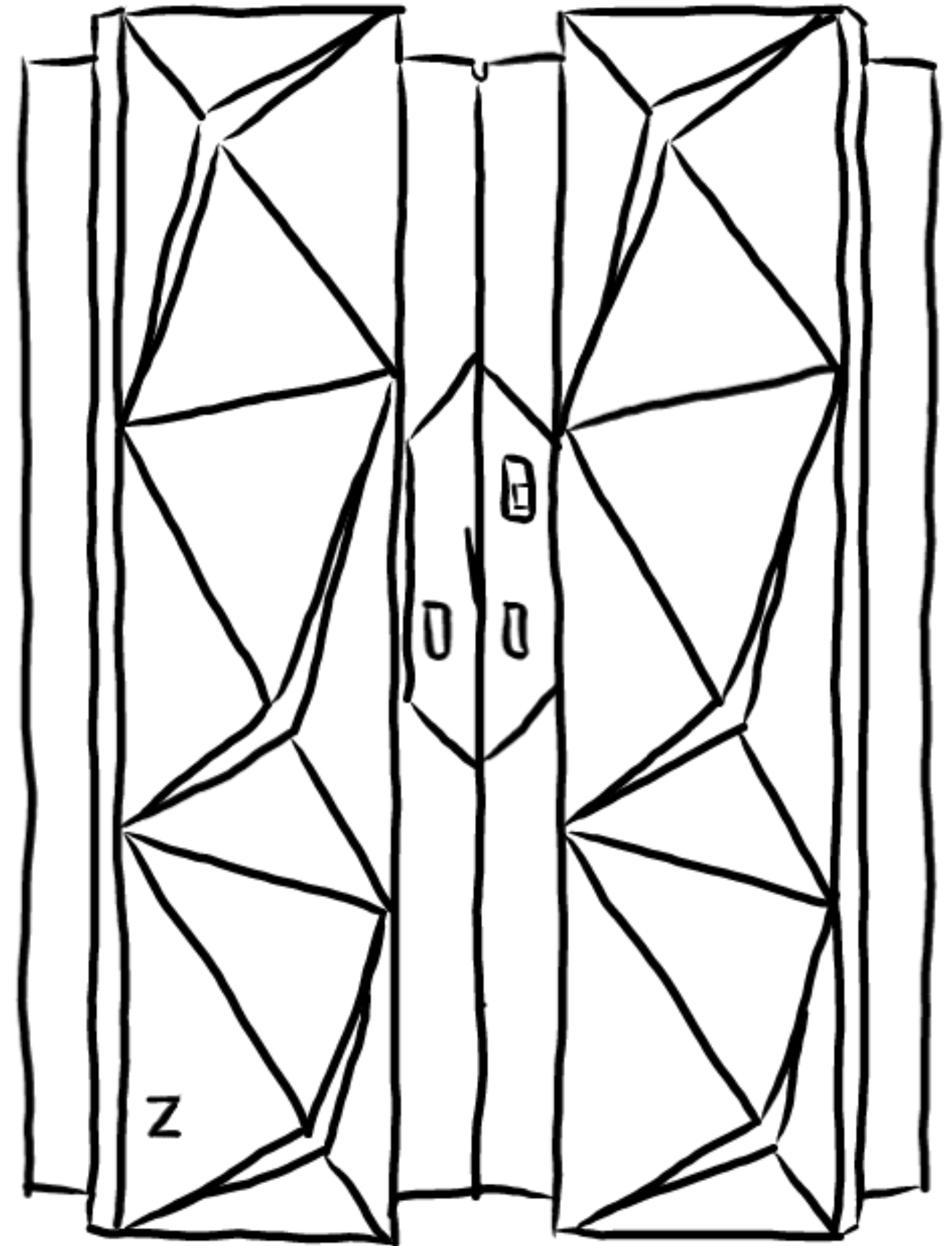LPAR #2 TCP/IP Profile Configuration

LPAR #2
Hardware Configuration:

```
GLOBALCONFIG SMCD SYSTEMEID
…
INTERFACE GIG4 DEFINE IPAQENET CHPIDTYPE OSD
    PORTNAME GIG3E IPADDR 9.82.24.153/24
    SMCD
```

```
OSA PNetID: IBMTRUNKSET2
ISM PNetID: NONE
ISM PFID: 07C2
```

Controlling which IP addresses are allowed to use SMC with filters

Controlling Which IP Addresses Are Allowed to Use SMC with Filters

Ideal for users: *who need more granular control over the SMC connection eligibility*
– SMC filters apply to all versions and types of SMC (SMC-Dv1, SMC-Dv2, SMC-Rv1, SMC-Rv2)
– Applied to remote host's IP address or IP subnet (not remote RoCEv2 IP address)
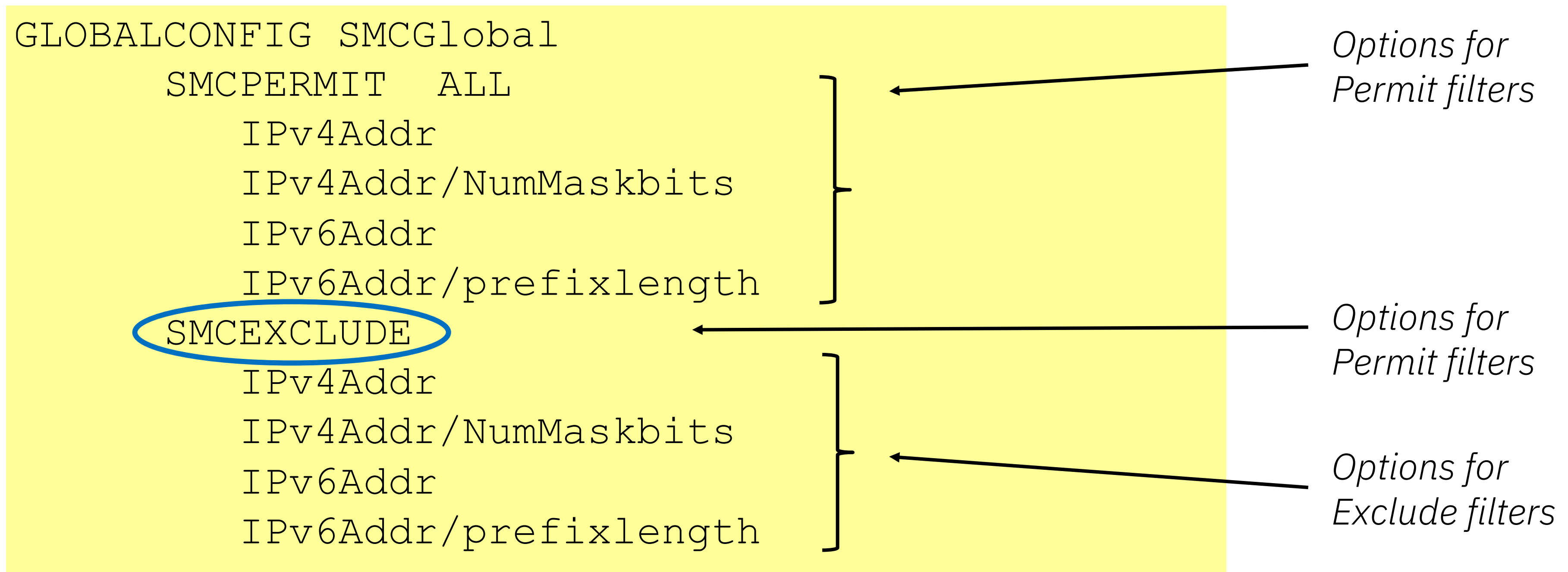– Use of SMC filters is optional

Considerations
– Defaults to "Permit All". Once filter is added, only peer hosts that match the filter are allowed to use SMC
– Exclude filters override Permit filter
– Minimize number of SMC filters to achieve required level of SMC connection eligibility control
– Define using the z/OS Network Configuration Assistant (NCA) TCP/IP profile technology to share definitions

Controlling Which IP Addresses
Are Allowed to Use SMC with
Filters

z/OS SMC Filters Setup
- IP addresses on SMC filters are for TCP/IP peer hosts IP address
- **SMCGlobal:** Options are defined on the GLOBALCONFIG statement

```
GLOBALCONFIG SMCGlobal
     SMCPERMIT   ALL
         IPv4Addr
         IPv4Addr/NumMaskbits
         IPv6Addr
         IPv6Addr/prefixlength
     SMCEXCLUDE
         IPv4Addr
         IPv4Addr/NumMaskbits
         IPv6Addr
         IPv6Addr/prefixlength
```

*Options for
Permit filters*

*Options for
Permit filters*

*Options for
Exclude filters*

Default is all if Permit filters not defined!

# Controlling Which IP Addresses Are Allowed to Use SMC with Filters

Permit Options
- New SMC users
- Existing SMC user
- All users
- SMCv2 users who move z/OS systems to different sites
- Users who want to slowly expand usage of SMC

Exclude Options
- A single or set of IP addresses in a large subnet
- An IP address that exhibits undesirable behavior
- An IP address on a temporary basis
- Overrides Permit filters

Tips & Tricks
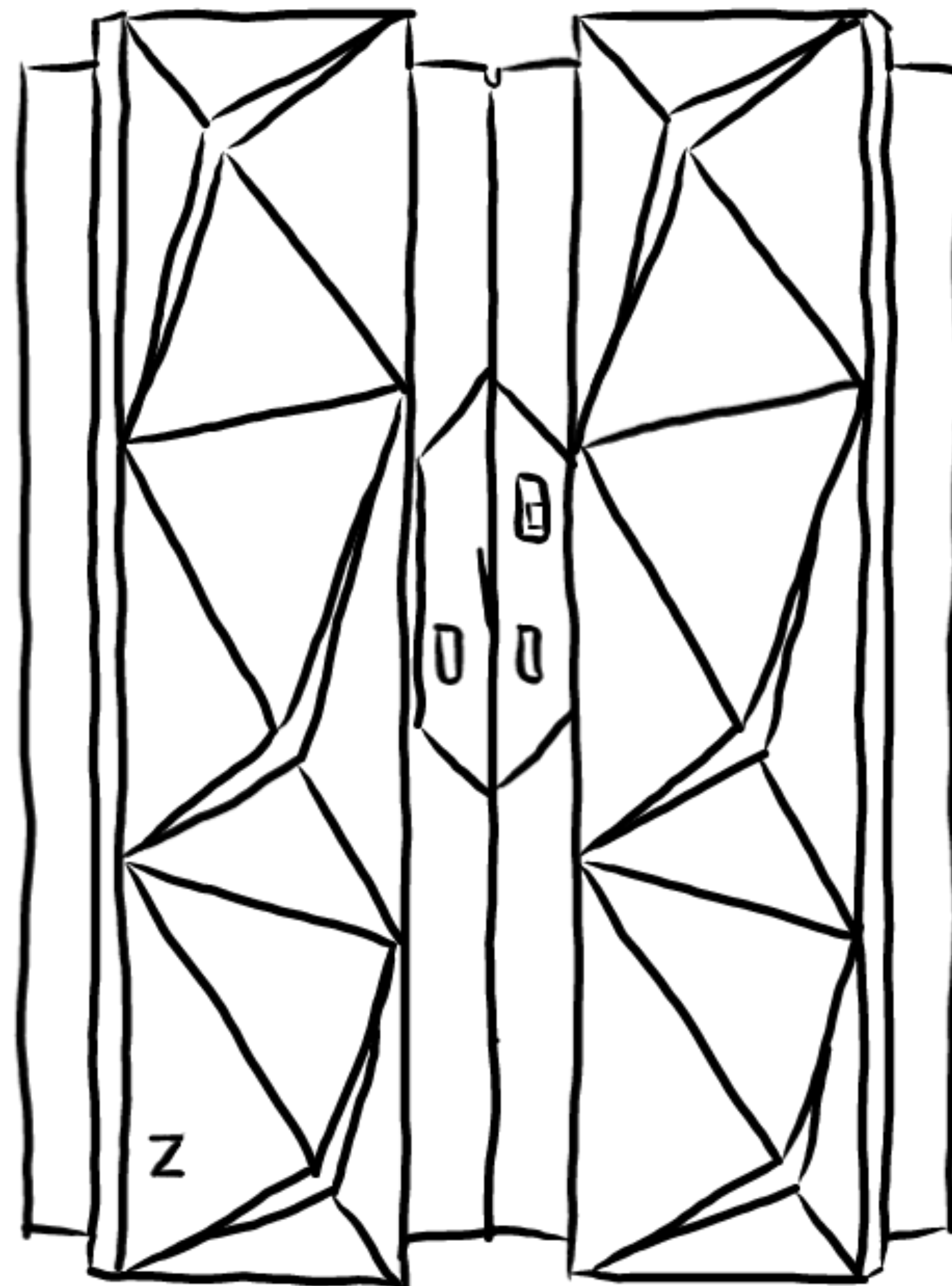
Tips & Tricks

For SMCv2, verify your firewalls permit TCP option 254 to flow.

For SMC-Rv2 with RoCEv2, verify that UDP port 4791 (RoCEv2) is open in your firewalls.

Provide physical redundancy for high-availability when using SMC-R

Deployment Considerations

Verify that your workload is applicable to SMC. Specifically, TCP only – e.g. Oracle RAC is known to predominantly use UDP, hence benefit will be small

Take into account that SMC-D/R might not apply to all traffic. Again, UDP traffic would continue to flow through regular NICs

Utilize the IBM SMC Applicability Tool (SMCAT) to monitor you current TCP/IP traffic for an IP address, group of IP addresses or IP subnets and then produce a summary report that will provide information about how much traffic (percentage) to/from those IP addresses is eligible and well suited for SMC-R and SMC-D
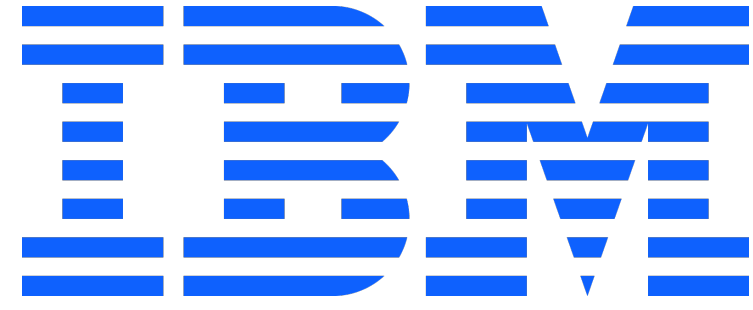
Contact

Chelsea Isaac
Client Technical Specialist
—

Chelsea.t.jean-mary@ibm.com

Appendix

References

[z/OS Communications Server: z/OS V2R5.0 Communications Server: IP Configuration Guide](#)

[z/OS 2.5 Hardware Configuration Definition User's Guide](#)

[IBM Enterprise Networking Solutions – Shared Memory Communications version 2](#)

[IBM Shared Memory Communications Version 2 (SMCv2)](#) (video)

[IBM z/OS Shared Memory Communications Version 2 (SMCv2)](#)

References

[Shared Memory Communications – Direct Memory Access (SMC-D) Frequently Asked Questions](#)

[Accelerate Networking with Shared Memory Communications for Linux on IBM Z](#)

[Configuring SMC-Dv2 for your z/OS environment](#)

[Configuring SMC-Rv2 for your z/OS environment](#)

SMC-D and SMC-R
Comparison

Feature	SMC-Dv2	SMC-Dv1	SMC-Rv2	SMC-Rv1
Intra-CEC	yes	yes	yes	yes
Cross-CEC	no	no	yes	yes
Cross-IP subnet	**yes**	no	**yes**	no
(R)DMA Device	ISM	ISM	RoCE	RoCE
PNET ID Definition	**Not required**	IOCDS or HCD	IOCDS or HCD	IOCDS or HCD
Failover	N/A	N/A	yes	yes

Dynamic Selection of SMC version and SMC type

Order of preference:
1. SMC-Dv2
2. SMC-Dv1
3. SMC-Rv2
4. SMC-Rv2

Result:
SMCv2 is used when client and server are both enabled for SMCv2

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


64
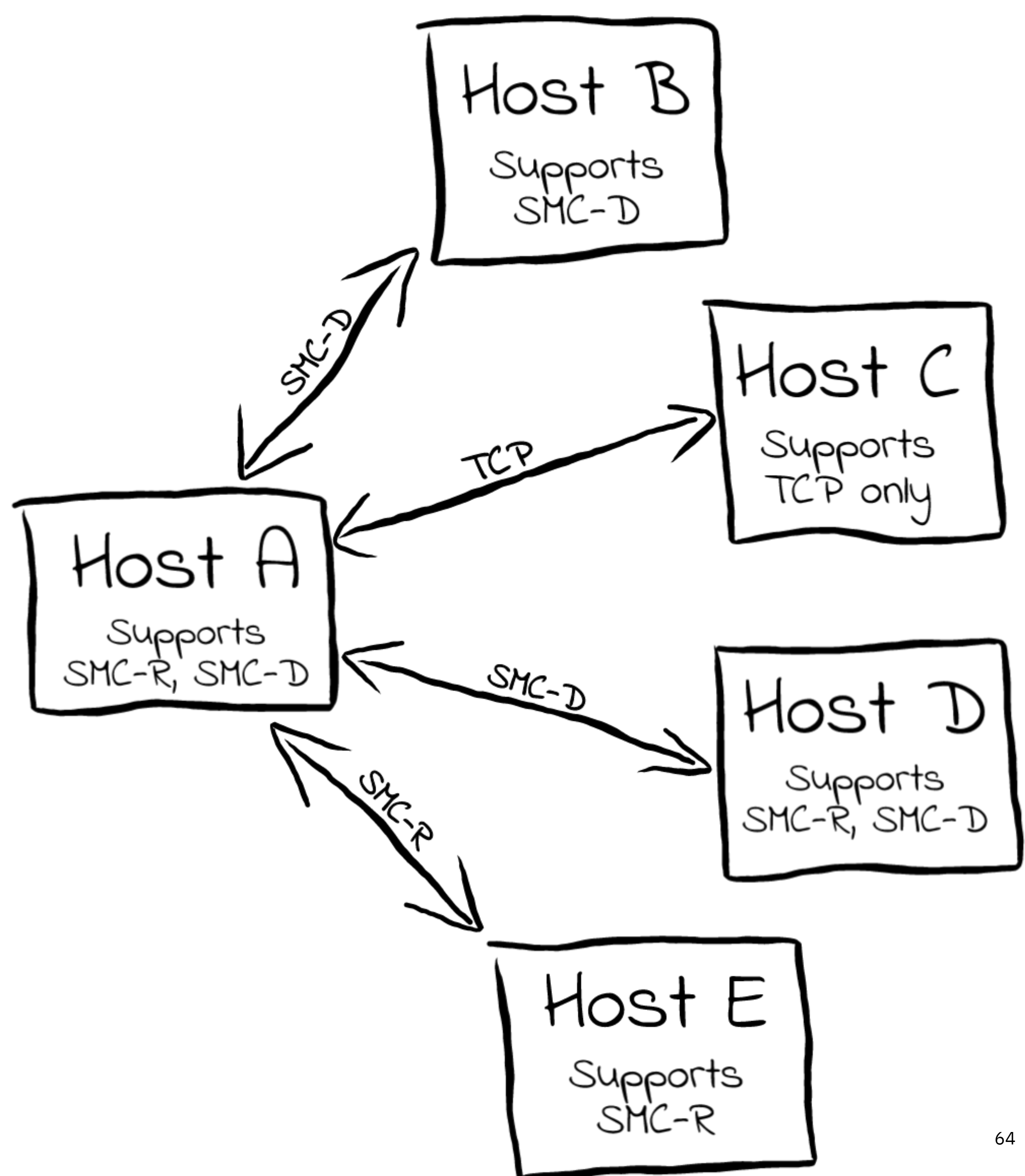
SMC-R Link Groups

After two SMC-R peers connect, a logical sMC-R link is established between the stacks over the RoCE fabric
SMC-R link groups provide for high availability, load balancing and recovery
Full redundancy requires:
– Two or more RoCE Express adapters at each peer
– Unique physical RoCE switches
Partial redundancy still possible in the absence of one or more of these conditions
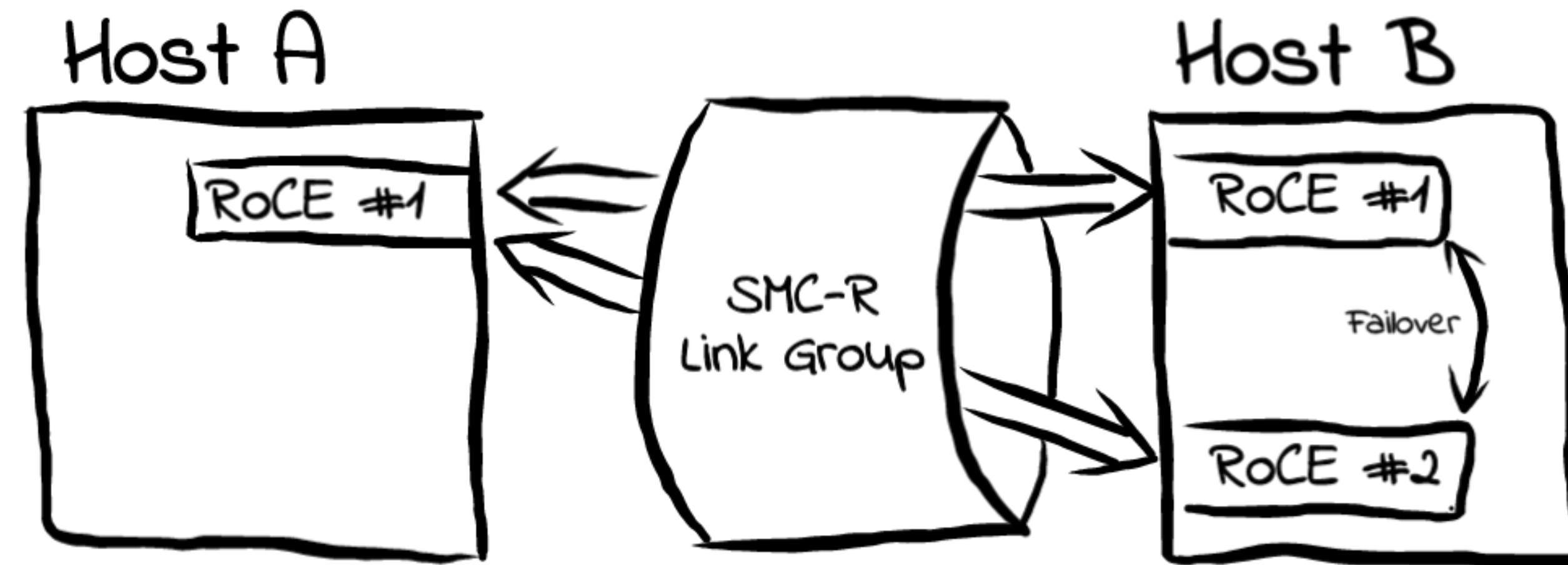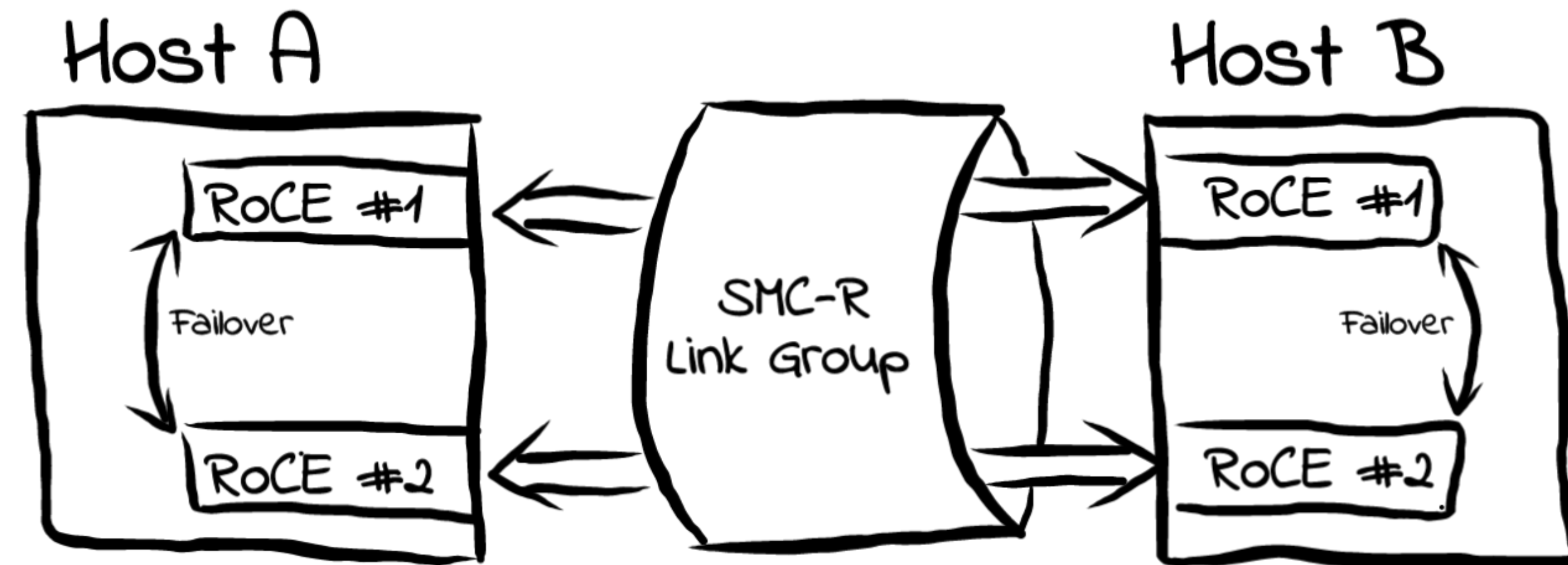
Enterprise ID

Each group of systems would be defined with a unique EID

- Fixed length 32-byte character data
- EID is a single ID representing an SMC group of systems
- Applies to the entire TCP/IP stack for IPv4 and IPv6
- Not associated to IP interfaces, networks or subnetworks
- A software attribute associated with the SMCv2 protocol. Not a hardware attribute
- Exchanged within the SMCv2 connection protocol where the EIDs are evaluated during the connection setup
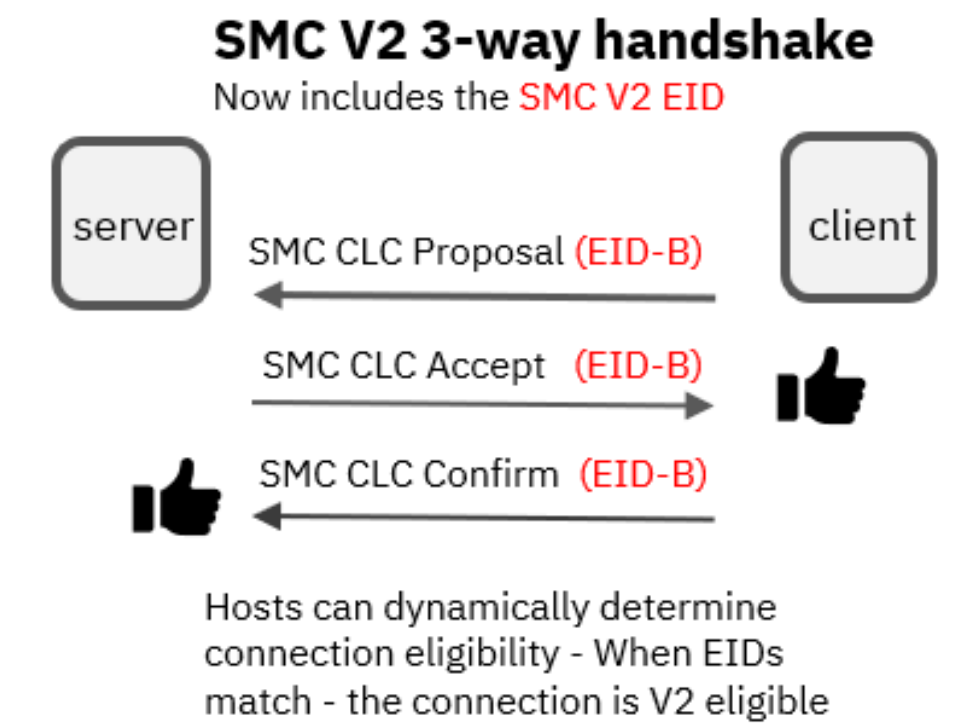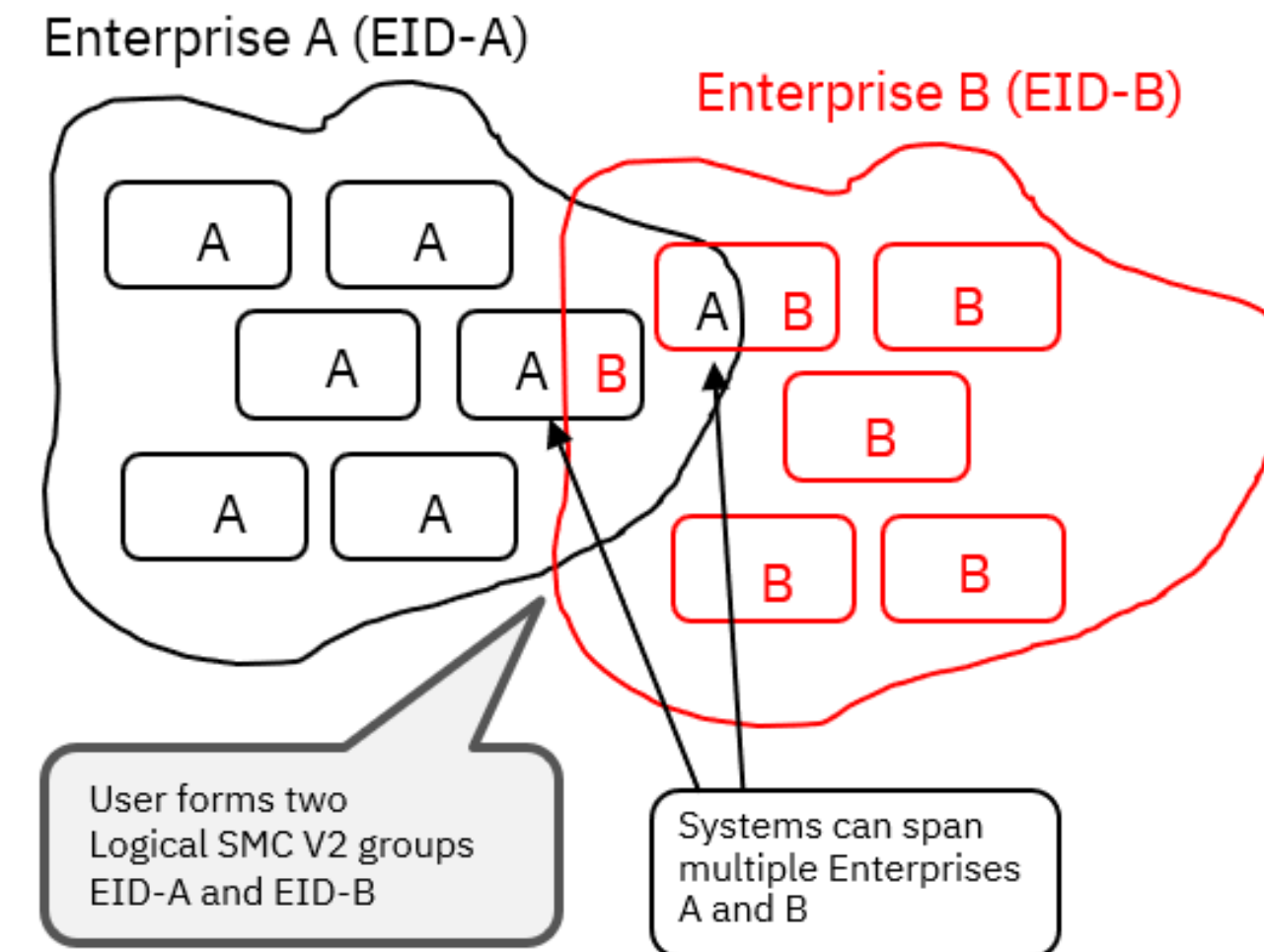
Enterprise ID format

A fixed length 32-byte character data that is a user defined ID
- Valid ASCII characters allowed for the EID are upper case A-Z, 0-9, hyphen, and dot
- First character cannot be a special character
- EID cannot contain consecutive dots

System EID (SEID)

SMC-Dv2 supports a system generated EID (SEID)
Automatically defined by the OS representing the CPC
– Encoded with system serial/type
Not defined by the user
A single unique SEID per CPC
Applicable to entire Z CPC
Applicable to SMC-Dv2 and ISMv2 only
Enables all SMC-Dv2 capable hosts within the same CPC
that also specify the SEID to communicate using SMC-
Dv2

Benefits:
• Reduces the SMC-Dv2 deployment time
• Enable and start SMC-Dv2 within a CPC with
  minimal software configuration

User Defined EID (UEID)

Allows administrators to create meaningful user defined Enterprise IDs that are easily recognizable

- Fixed length 32-byte character data
- Can establish unique enterprise-wide EID naming conventions (e.g. A400—RODCLUSTER4, CC29-DEVCLUSTER18)
- Hyphen and dot allows users to compose EIDs using elements of their IT infrastructure
  - Company-Loc-Group, BusinessLine-Group, Plex-Cluster-Type, etc.
- Recommended all 32 characters are defined by the user

Multiple EIDs

A single EID should be sufficient for your enterprise data center
- There are use cases when a single z/OS system needs to reside within multiple groups supporting multiple EIDs
- z/OS Comm Server allows each TCP/IP stack to define up to 4 user-defined EIDs

Valid use cases:
- Users want to separate and control usage based on business lines, security zones, test vs production

Relationship between VCHIDs, FIDs, and connections

- **ISM devices**

  - *Virtual* PCI network adapter of VCHID type ISM

  - 32 ISM VCHIDs per CPC, 255 FIDs per VCHID
    ⇒ 8K FIDs per CPC total)

  - I.e. maximum of 255 virtual servers
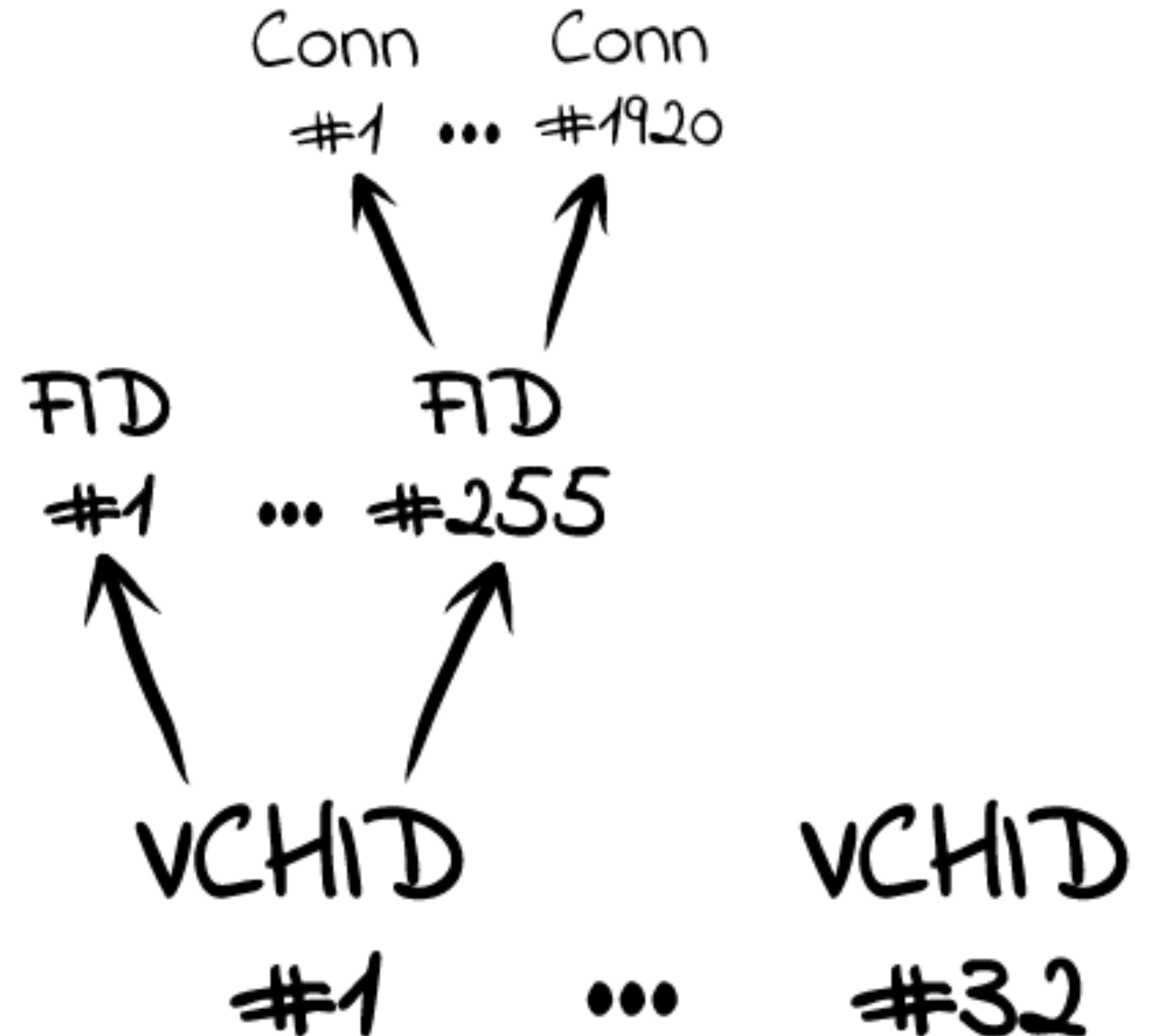   communicating over same ISM VCHID

Conn   Conn
#1 ... #1920

FID   FID
#1 ... #255

VCHID   VCHID
#1   ...   #32

Fig.1: *Relationship between VCHIDs, FIDs and connections*

Configuration considerations for Shared Memory Communications

1) Decide whether to use VLANS
    – VLANID operand is optional on SMC capable IPAQENET INTERFACE statements with OSD channel path ID type or IPAQIDIO INTERFACE statements (SMC-D only)
2) Identify physical connections between stacks
    – Use a physical network ID (PNetID) to represent the ID or name of your physical layer 2 LAN fabric
    – For SMC-R:
    – You can use two RoCE Express ports to connect to a physical network but you can use as many OSA adapters as necessary
    – Once TCP/IP stack can define up to 16 PFID values, each must match a FID value in the HCD.

Configuration considerations for Shared Memory Communications

- Define a PNetID value for **both** the RoCE Express interface (physical port) and the corresponding OSA adapters (physical port) in the HCD to match the RoCE Express features with the right OSA SMC-R adapters
- For SMC-D:
- Define a PNetID value for both the ISM device and the corresponding OSA or HiperSocket in the HCD
- The same physical network can't be used for OSA and HiperSockets. If you code the same PNetID for OSA and HiperSockets adapter, the first interface to become active will be enabled for SMC-D. The second interface will not be eligible for SMC-D communications

Configuration considerations for Shared Memory Communications

3) Provide physical redundancy for high availability when using SMC-R
   - After a TCP connection switches to SMC—R, it cannot revert to standard TCP/IP communications. So it is critical to provide redundant physical network connectivity.
   - Two RoCE Express features must be defined and active.
   - In a shared RoCE environment, ensure that the PFID values used by a given TCP/IP stack represent physically different "RoCE Express" features (i.e. configured with different PCHID values).
   - Each SMC-R peer should have multiple active RoCE Express interfaces
   - **High availability requirements between two SMC-R peers:**
     o Two unique physical RoCE Express features that use unique PCHIDS
     o Unique system PCIe support infrastructures for the two features
     o Unique physical RoCE switches

Configuration considerations for Shared Memory Communications

4) Verify the system and network requirements

SMC-Dv1 Overview

Intra-CEC connectivity using *Internal Shared Memory* (ISM) devices

IBM Z hardware requirements minimum :
– z13 or later
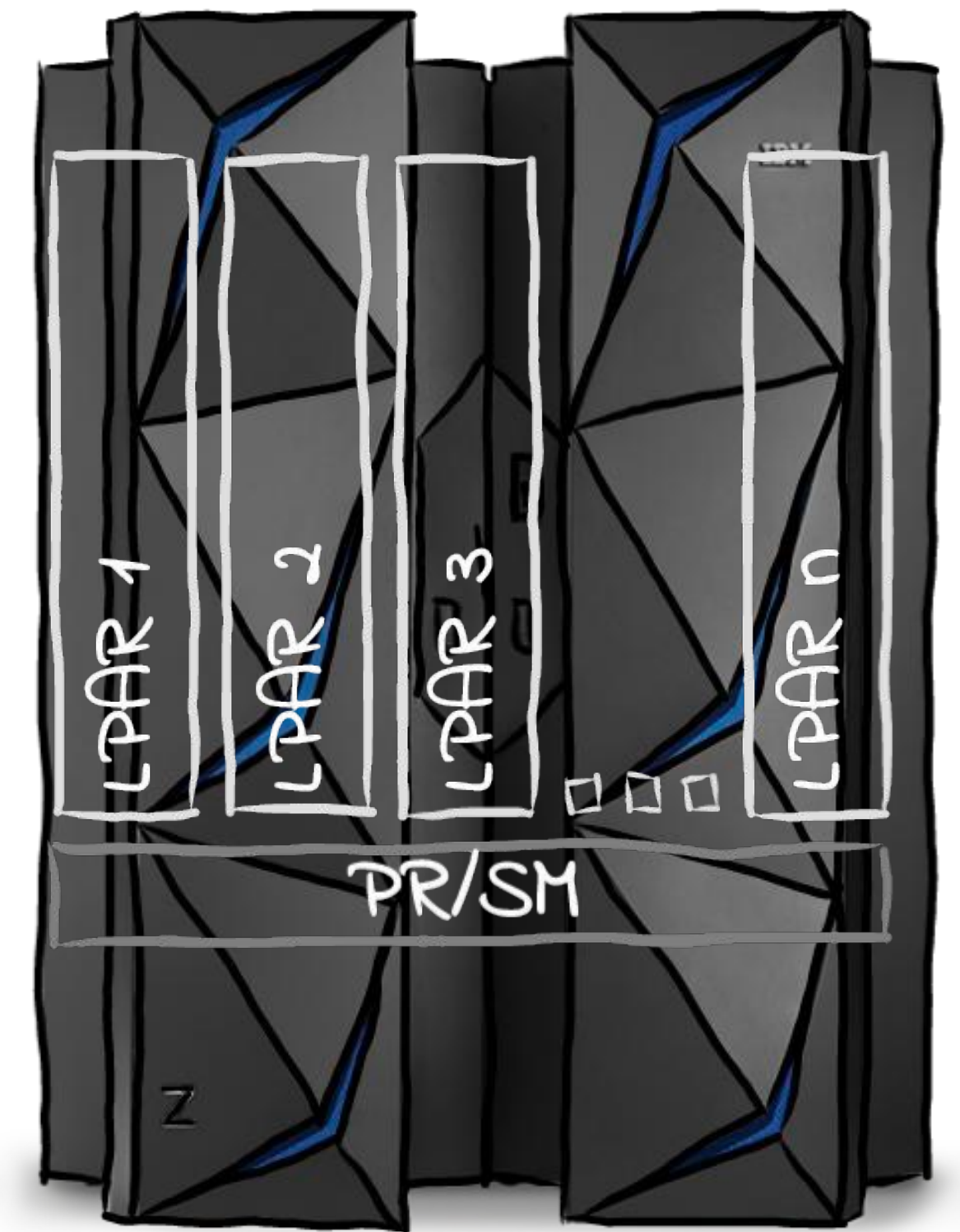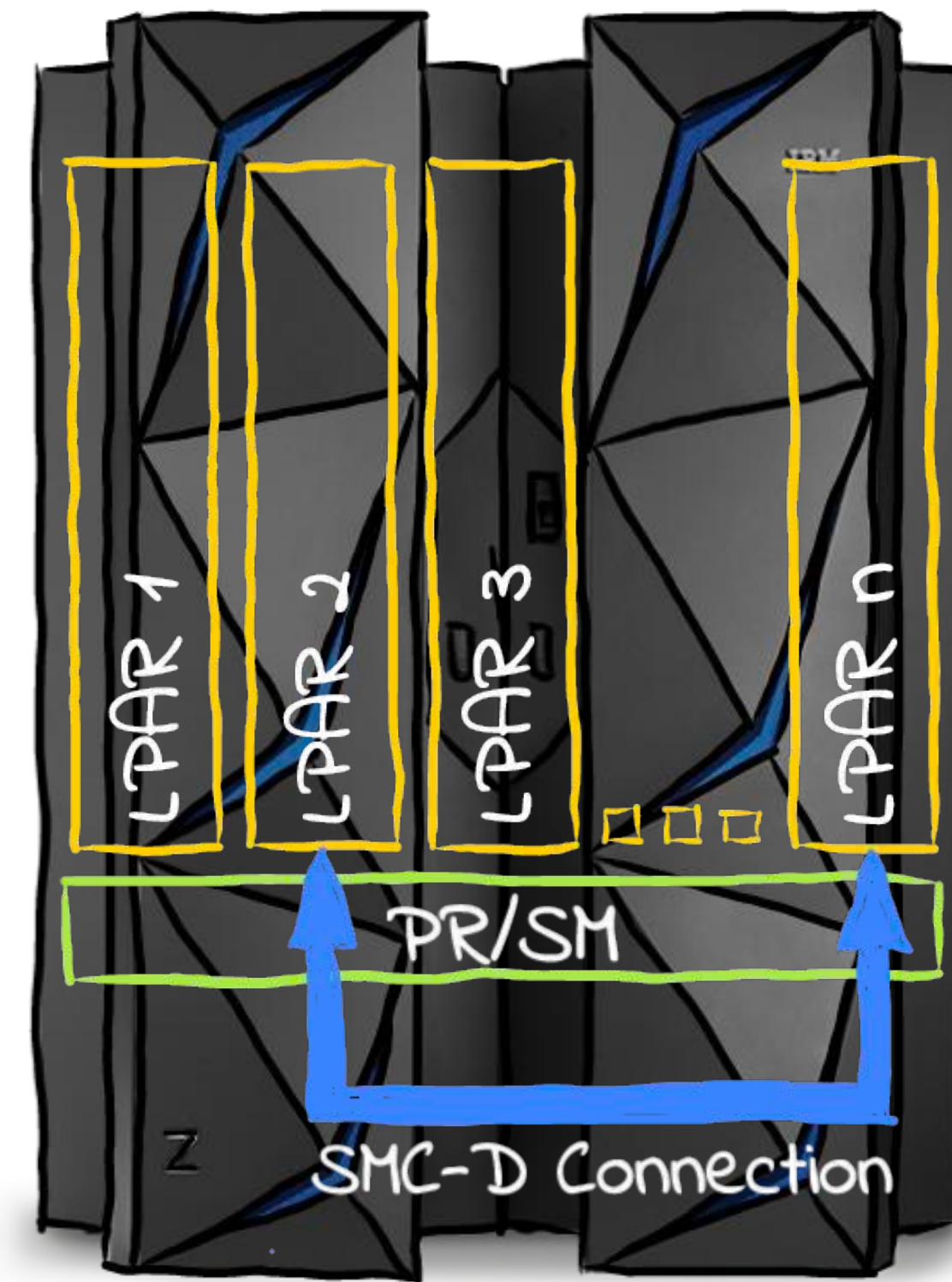
IBM Z software requirements minimum :
– z/OS Version 2 Release 2 (APARs OA48411 and PI45028 applied)

ISM devices:
– Virtual PCI network adapter
– Provides access to memory shared between LPARS
– 32 ISM VCHIDs per CPC, 255
– Each ISM VCHID represents a unique (isolated) internal network, each having a unique Physical Network ID

PNET ID Configuration
– HCD only
– Use HiperSockets, OSA or RoCE cards for regular connectivity

SMC-Rv1 Overview

Cross-CEC connectivity using *RoCE Express* cards

IBM Z hardware requirements minimum :

– IBM zEnterprise EC12 (zEC12) with driver 15, zBC12 or later

IBM Z software requirements minimum:

– z/OS Version 2 Release 1 (APARs OA51949 and PI75199 applied)

– z/OS Version 2 Release 2 (APARs OA51950 and PI75200 applied)
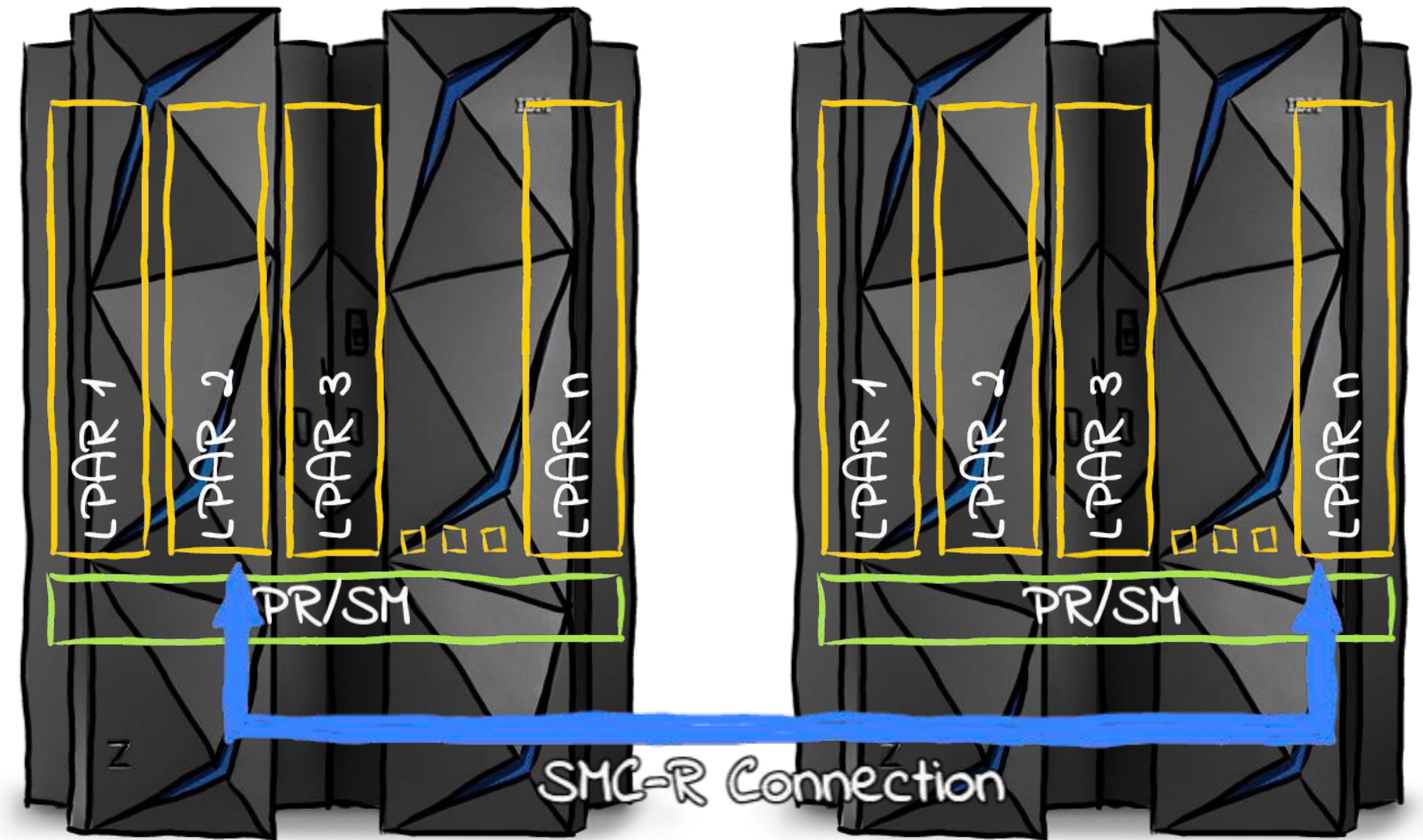
RoCE Express cards:

– Virtual PCI network adapter

– RoCE Express & RoCE Express2 cards supported

PNET ID Configuration

– HCD only

– Use HiperSockets, OSA or RoCE cards for regular connectivity

SMC-R Overview

For each new TCP connection:
- Start out with a regular TCPIP 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
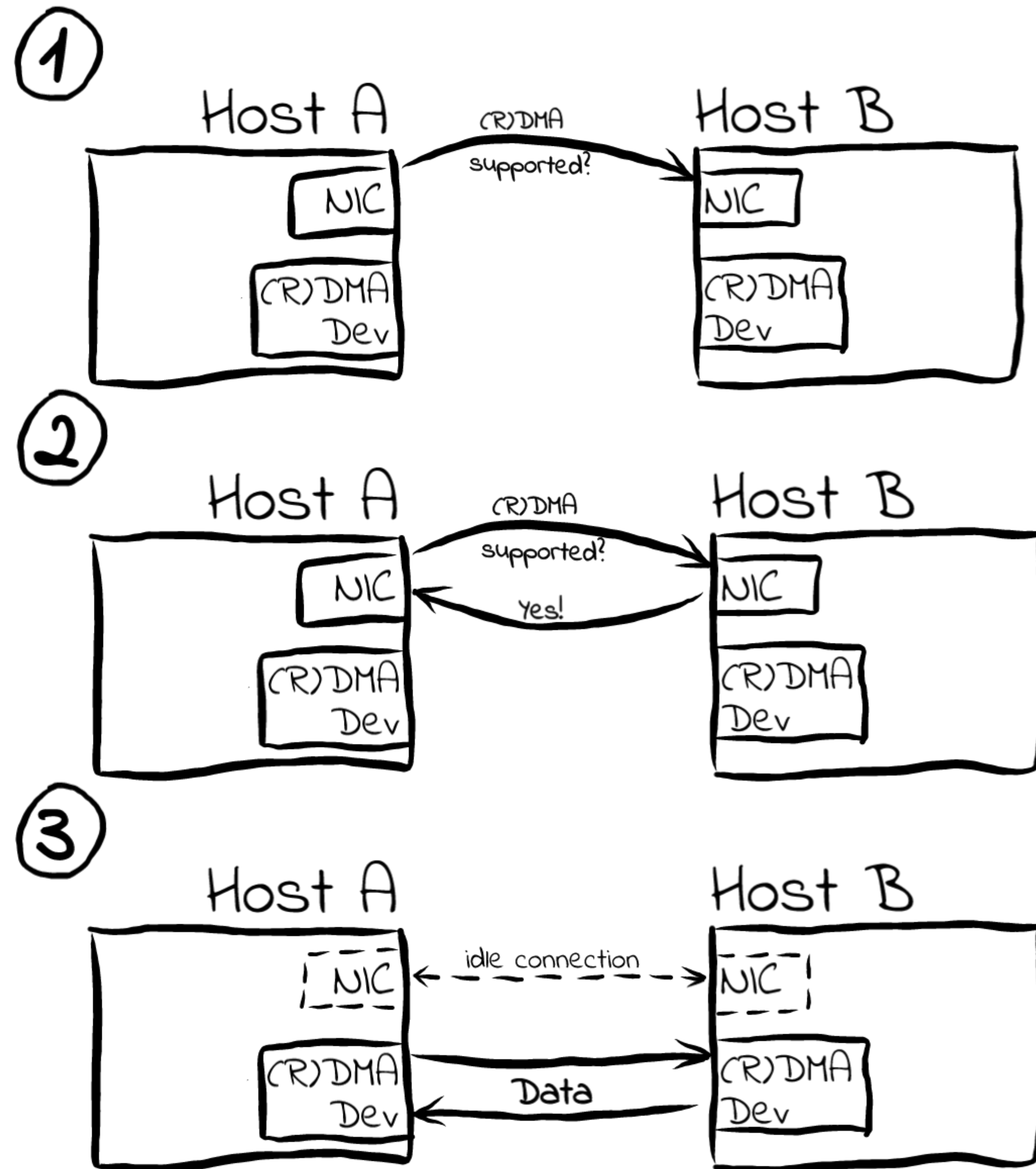
PNET IDs

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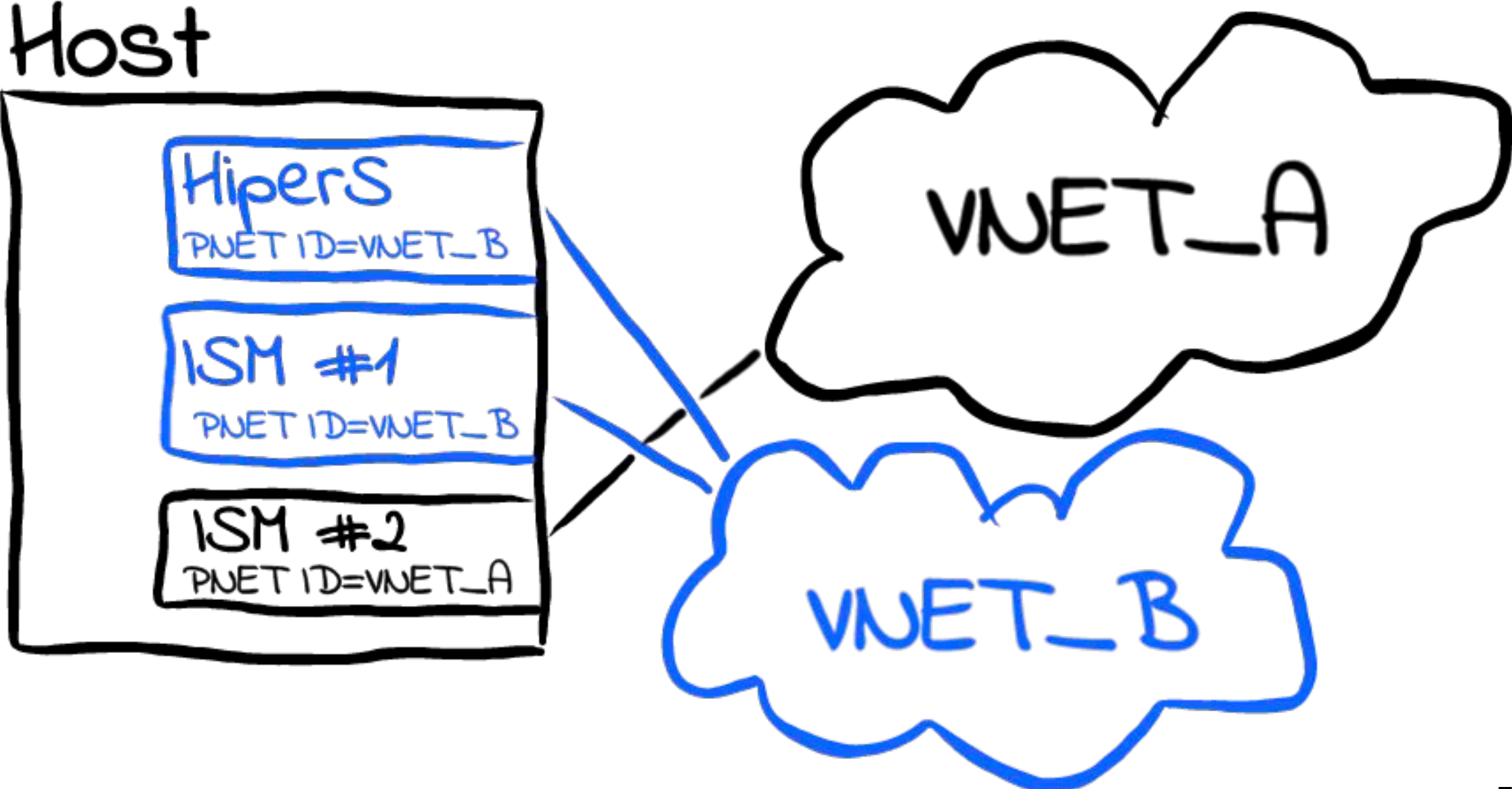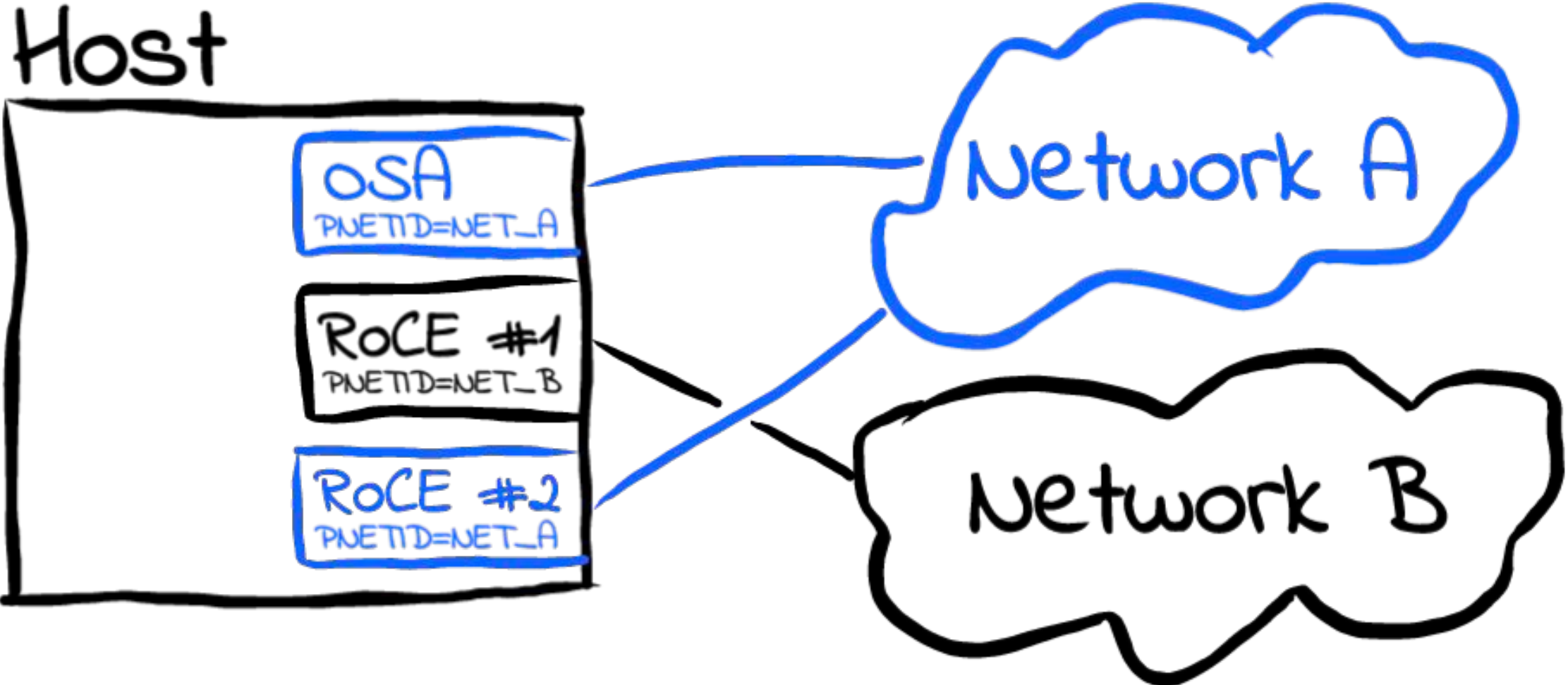
Defined in:
– Hardware configuration definition (HCD)
Typically associate:
– OSA and RoCE cards, or
– HiperSockets and ISM devices
Note: PNET IDs help to locate a suitable (R)DMA device for a given NIC *within a host*. The peer can use totally different PNET IDs (as long as the right devices are grouped)

Physical Network ID configuration issues

In the HCD definitions, the same PNetID values have different meaning for different type of devices.

Device type	PNetID 1	PNetID 2	PNetID 3	PNetID 4
HiperSockets	Represents the device	Not used	Not used	Not used
ISM	Represents the device	Not used	Not used	Not used
OSD adapter	Represents port 0	Represents port 1	Not used	Not used
RoCE Express	Represents port 1	Represents port 2	Not used	Not used
RoCE Express2	Represents the device	Not used	Not used	Not used
RoCE Express3	Represents the device	Not used	Not used	Not used

SMC-Rv2 OSA and RoCE relationship

Figure 1. SMC-Rv2 OSA and RoCE relationship

PFIDs

PFID: *Peripheral Component Interconnect Express (PCIe) function identifier*
A value that represents the SMC device

For SMC-R, represents a physical "RoCE Express" feature
For SMC-D, represents an ISM (internal shared memory) device

Defined in:
– Hardware configuration definition (HCD)
– Configured on the SMCR parameter of the GLOBALCONFIG statement in the TCP/IP profile to identify a "RoCE Express" feature
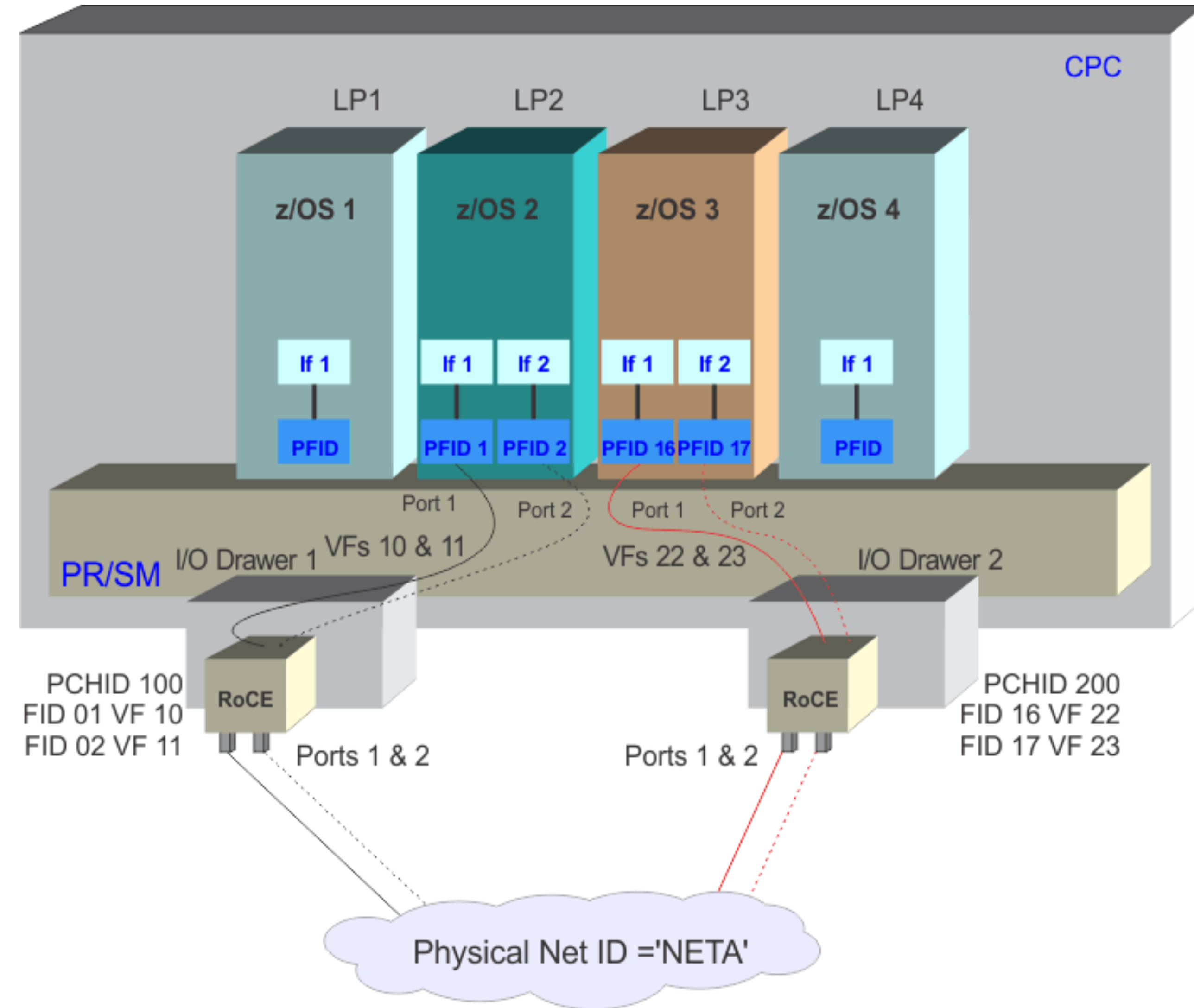
RDMA

RDMA was first codified in November 1993 by Hewlett-Packard engineers.
If you give networked systems a way to access each other's main memory without interrupting the processor or operating system, you can drive down latency, drive up throughput and simplify life for everybody.
This eliminated back-and-forth traffic between systems that was slowing computers.

The Infiniband Trade Association defined an initial version of RDMA over Converged Ethernet (RoCE, pronounced "rocky") in 2010 and today's more complete version that supports routing in 2014.

Now, RDMA is nearly everywhere, it's becoming invisible. RDMA is a function running on embedded processor cores on smart network-interface cards, aka smart NICs.

RDMA

Direct Memory Access (DMA) technology – the ability to access host memory directly without CPU intervention

Remote Direct Memory Access (RDMA) – allows for accessing memory data from one host to another. It greatly improves throughout and performance while lowering latency because less CPU cycles are needed to process the network packets

In a traditional network data path, an application needs to go through the buffers using the sockets API. With RDMA, the kernel is bypassed. Network latency is lowered while data throughout is increased because there's a lot less CPU cycles involved.
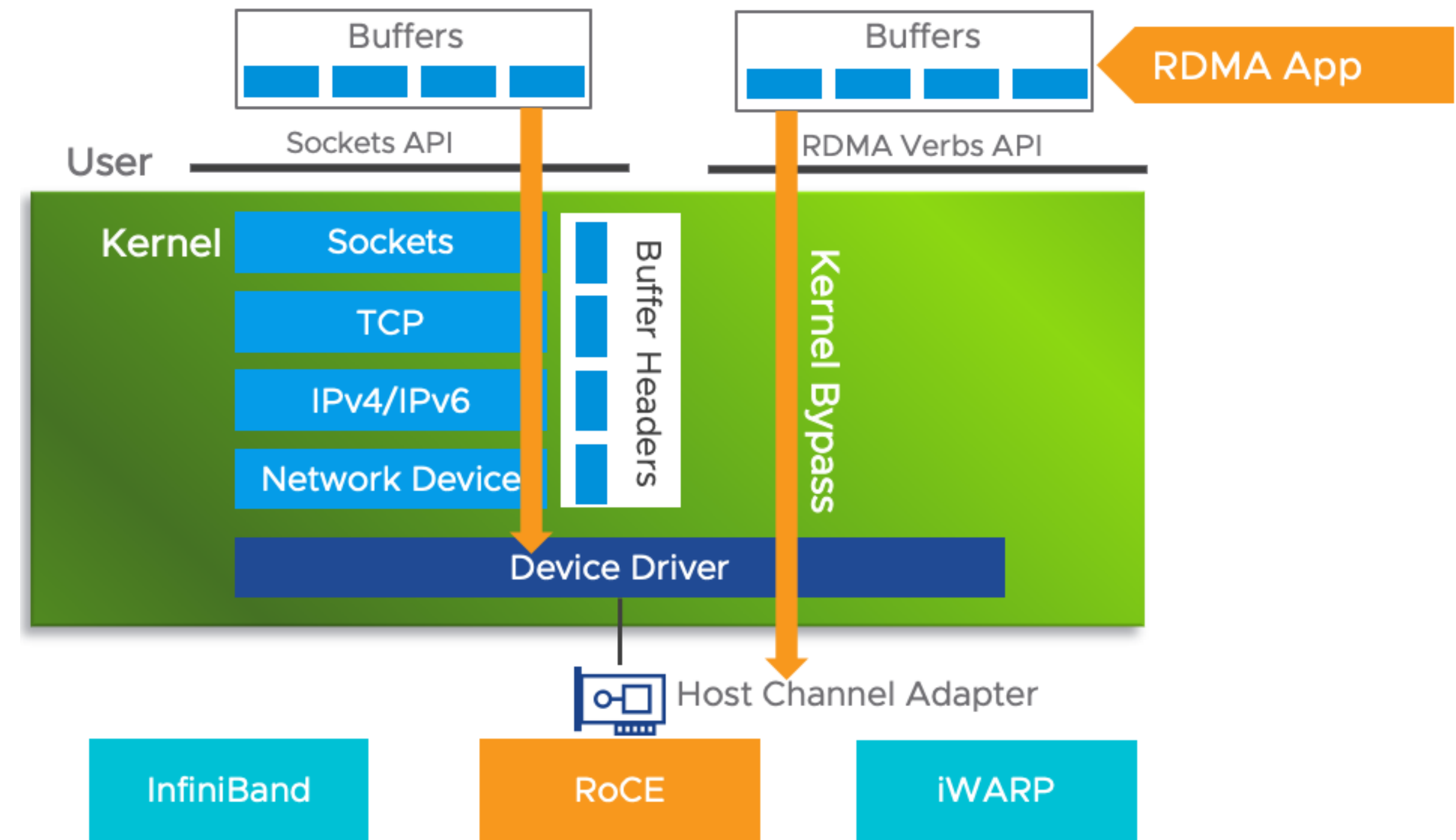
Deployment Scenario

- **Could be any networking topology as long as both LPARs are located on the same CEC**
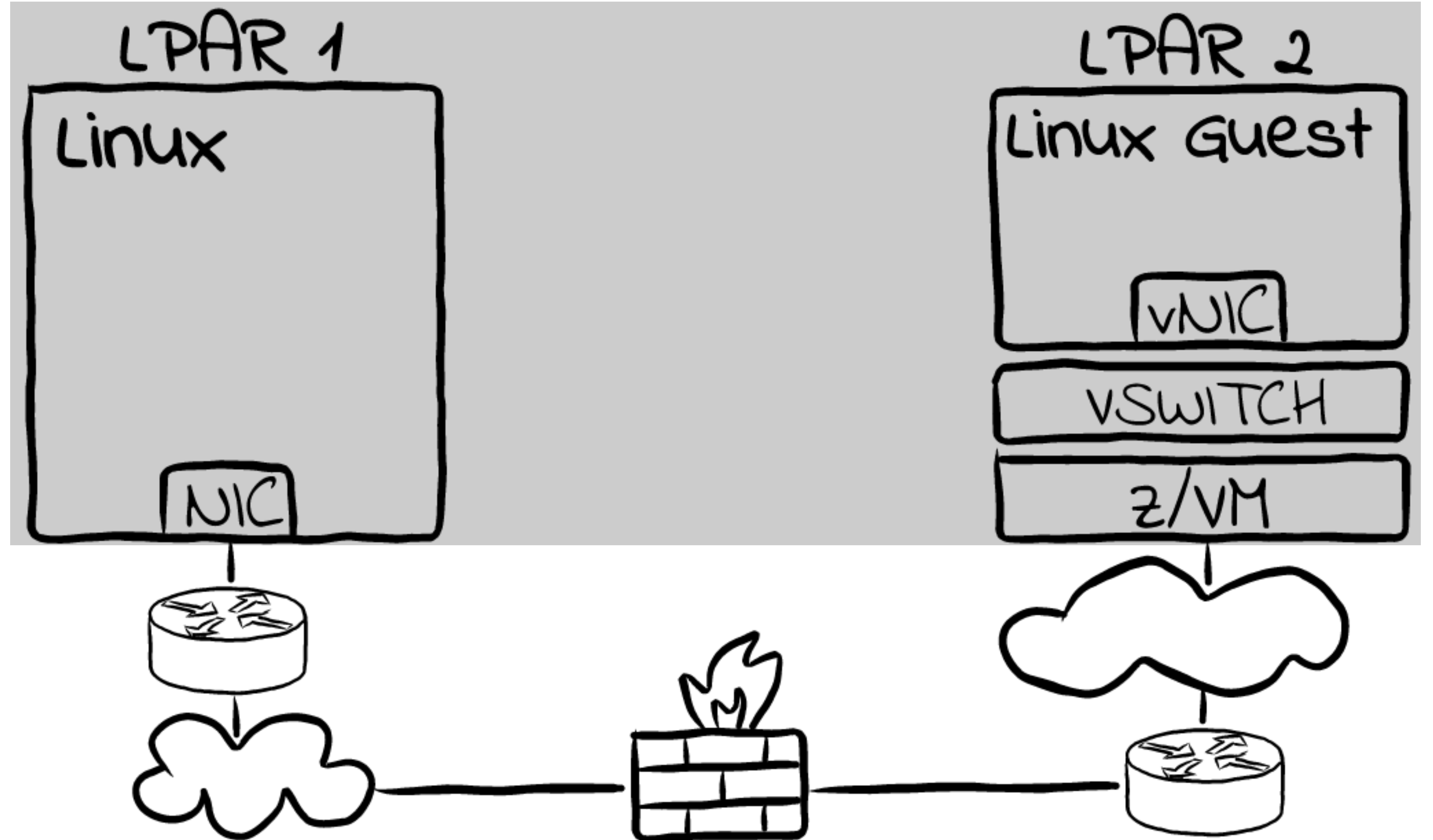
Traffic Flows

- **HiperSockets might be an obvious choice, but security policies often mandate a traffic to pass an external firewall**
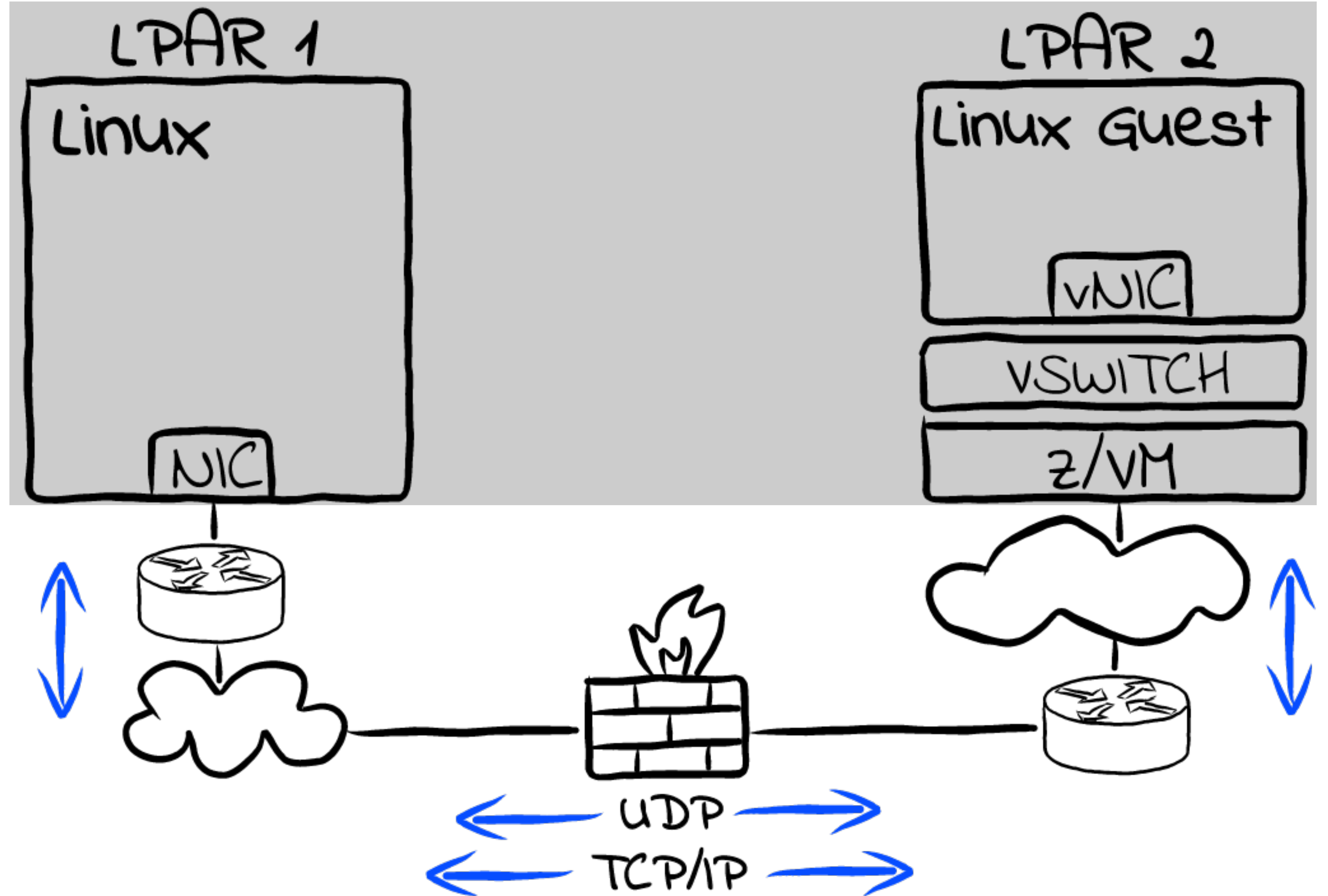
Establishing Connections

- **TCP 3-way handshake is followed by an extra SMC-specific handshake for each new connection**

- **Honors firewall rules!**

- **Overhead is minor, but to be considered for short-lived connections**
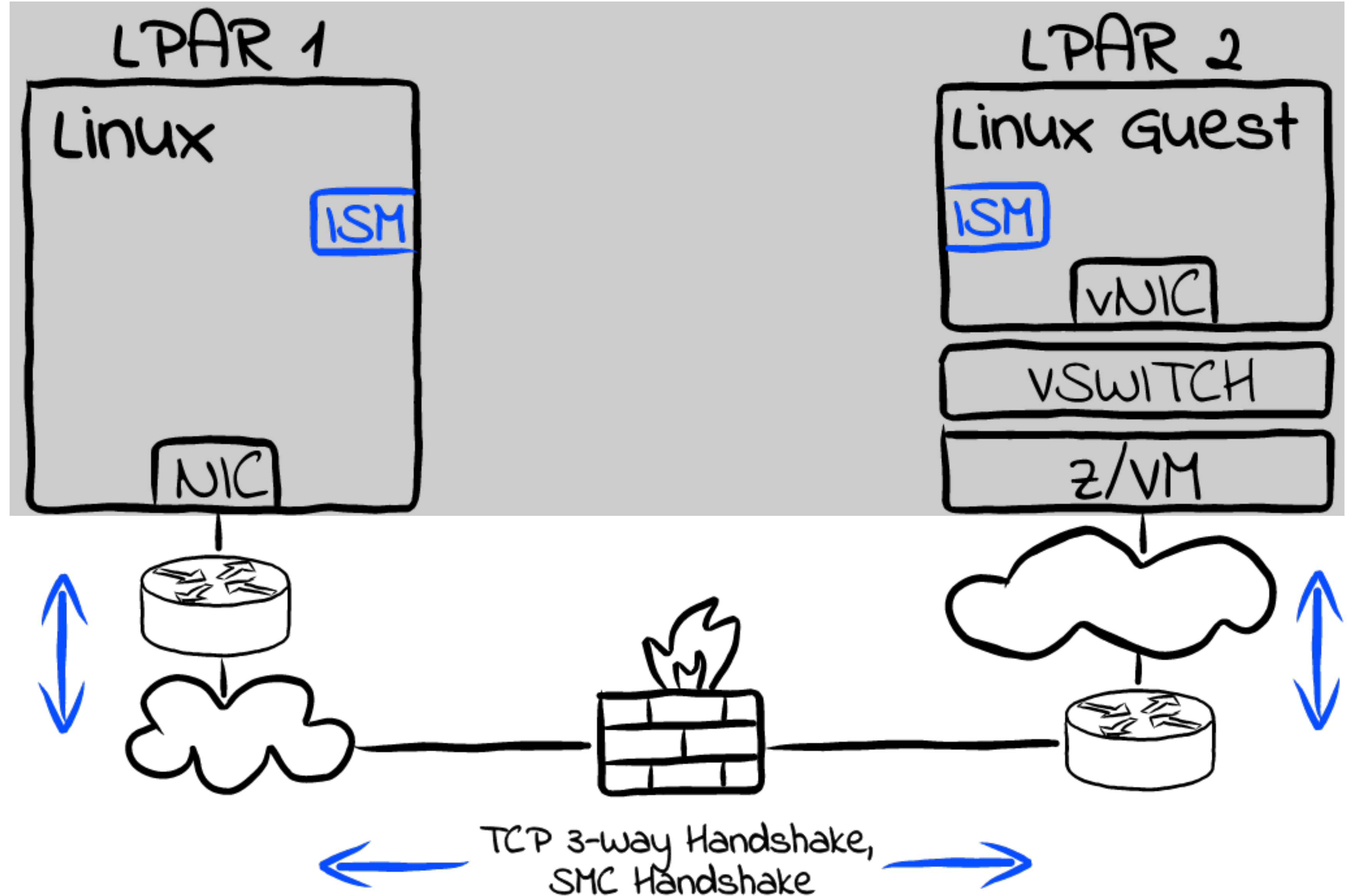
Data Flow

- **Once established, TCP data is transmitted through memory-to-memory copy via ISM devices**

- **Uneligible traffic takes the "*detour*" ⇒ Regular connectivity still needed**

SMCv2 3-Way Handshake

Server (IP subnet X)   IP network   client (P subnet Y)

TCP 3-way hand shake (SYN Flows with TCP option 254 = SMCR)

A ← - → A

SMC V2 CLC **Proposal** (SMC-R V2 GID y + V2 GID (IP addr B) EID)   1
← -

2   SMC V2 CLC **Accept** (SMC-R V2 GID (IP addr A) EID RMB info, etc)
- →

SMC V2 CLC **Confirm** (SMC-R V2 GID (IP addr B) EID RMB info, etc.)   3
← -
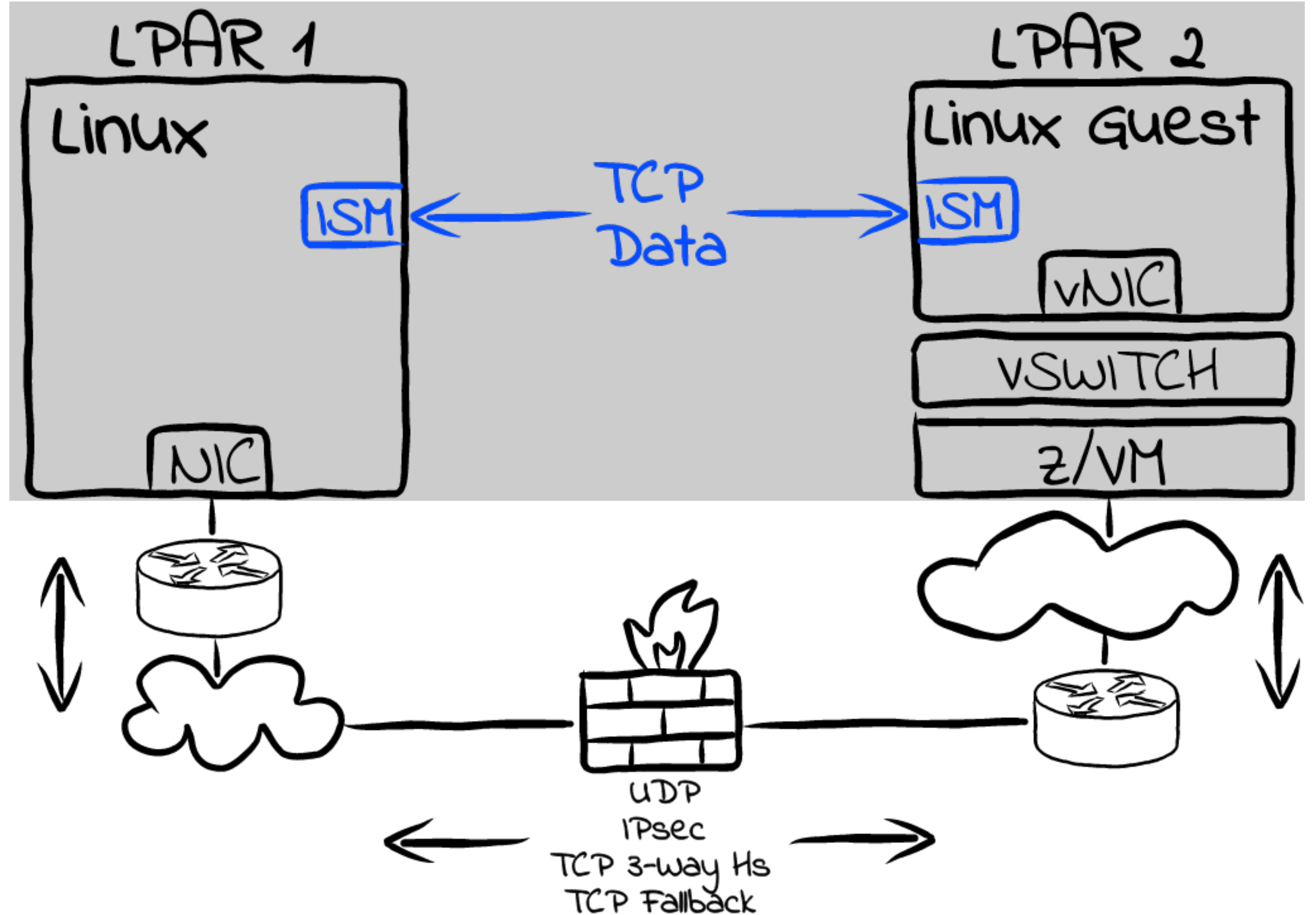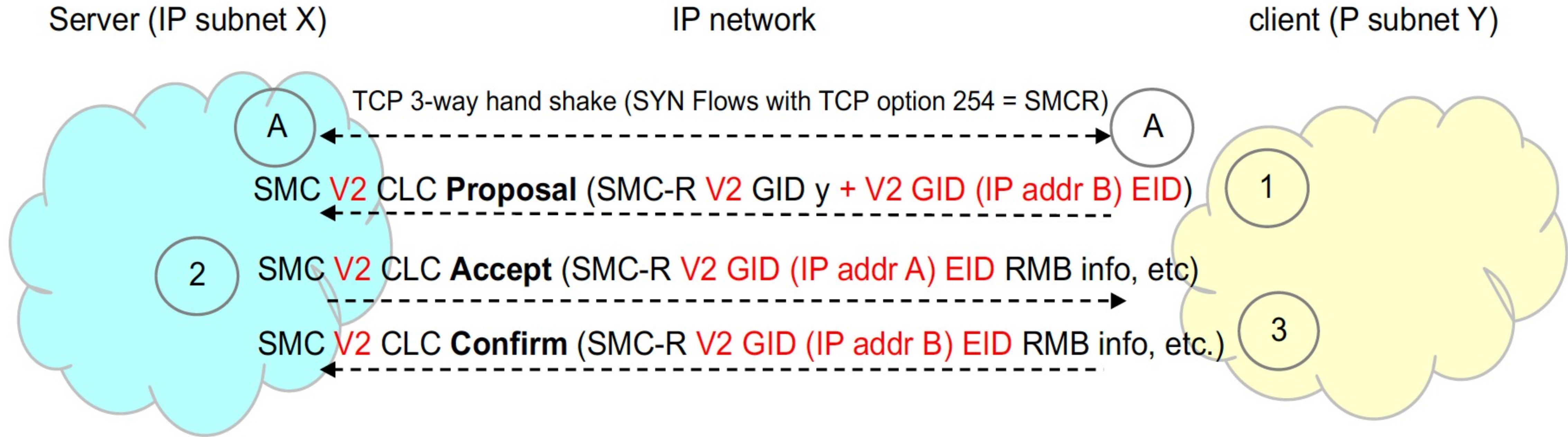
A.   The TCP 3-way handshake precedes the SMC CLC handshake.

1-3. After SMC CLC 3-way handshake completes the TCP connection

transitions to SMCRv2 using RoCEv2 (standard LLC flows continue)

SMCv2 3-Way Handshake

RoCEv2 – "IP Routable RoCE"

Ether Type indicates that packet is RoCE (i.e. next hdr is IB GRH)

RoCE

| Eth L2 Hdr | Eth Type | IB GRH | IB BTH + (L4 hdr) | IB Payload | ICRC | FCS |

RoCEv2

| Eth L2 Hdr | Eth Type | IP Hdr | Prot # | UDP Hdr | Port # | IB BTH + (L4 hdr) | IB Payload | ICRC | FCS |

Ether Type indicates that packet is IP (i.e. next hdr is IP)

IP Protocol # Indicates packet is UDP

UDP Port # Indicates next header is IB.BTH

UDP Port number 4791 is reserved for RoCEv2 👍

12

SMC-Rv2 with RoCEv2 ("Routable RoCE")

CPC -A

z/OS images

CPC -B

z/OS images

Layer 2 networks

Layer 2 networks

IP subnet A

IP subnet B

SMC V2 connections are no longer restricted the same IP subnet

RoCEv2 is no longer limited to a LAN

Layer 3 networks

SMC-Rv2 / RoCEv2 traffic now crosses IP routers - encapsulated in UDP/IP packets – IP routable

RoCEv2 uses UDP Port 4791 (must be open)

13

© 2021 IBM Corporation

91
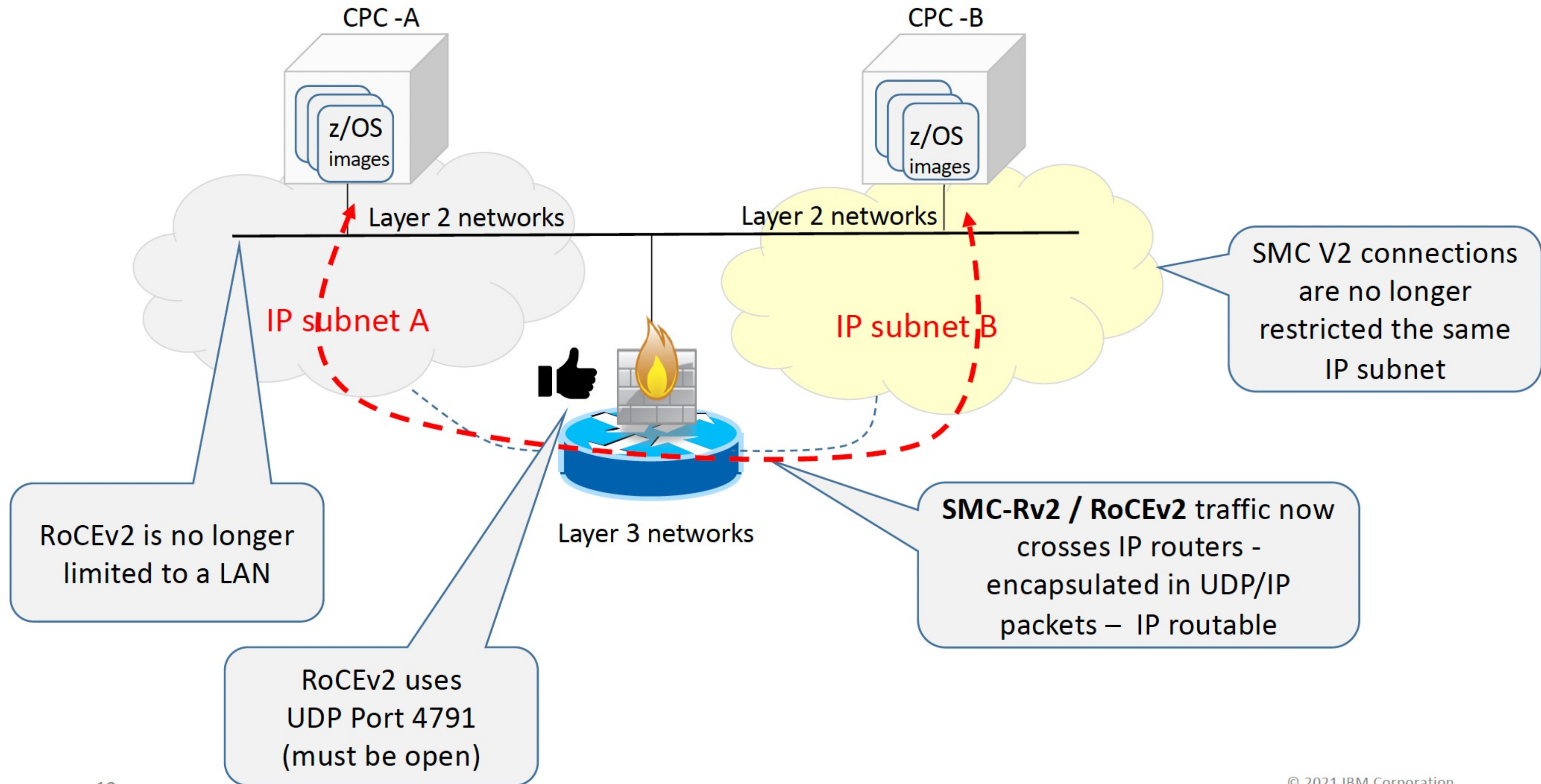
SMC-Dv2 with ISMv2 (multiple subnets)

CPC -A

z/OS . . . z/OS z/OS ISMv2 Linux . . . Linux Linux z/VM

Layer 2 networks   Layer 2 networks

IP subnet A   IP subnet B

SMC-Dv2 connections are no longer restricted to hosts within the same IP subnet

ISMv2 enables SMC-Dv2 allowing ISM communications across IP subnets.

Layer 3 networks

TCP/IP connections that setup across multiple IP Subnets are also now eligible to use **SMC-Dv2 / ISMv2**

In this example, the z/OS Systems (subnet A) are now eligible to use SMC-Dv2 with the Linux systems (subnet B)!

Summary

- **SMC-D accelerates TCP LPAR-to-LPAR traffic by using memory-to-memory copies, bypassing**

    1) the TCP/IP stack

    2) the connecting networking fabric

- **At the same time,**

    – SMC-Dv2 works for *any* network topology

    – SMC-D honors security policies

SMC-D

Key Attributes

- Easy HW setup: Add ISM device – done!

- Massive performance benefits as compared to HiperSockets – gets even better when compared to regular NICs!

- Transparent to (TCP socket based) applications

- 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

Typical Workloads To Benefit

- *Transaction-oriented / latency-sensitive*

- *bulk data streaming*, e.g. when running backups

- Huge amounts of concurrent connections

SMC-R

Key Attributes

- Massive performance benefits

- Transparent to (TCP socket based) applications

- 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

Typical Workloads To Benefit

- *Transaction-oriented/latency sensitive workloads*

- *Transactional workloads* exchanging large messages (WAS Type 4 Connections to Db2 or CICS)

- *bulk data streaming*, e.g. FTP or when running backups

- Applications where data must be transferred between two locations in the data center

- z/OS Sysplex Distributor (SD) clients with a RoCE connection to SD using VIPARoute

- Any IT organizations requiring fast file transfers between processors running z/OS

Overview

Enables a host to read or write directly from/to a remote host's memory with drastically reduced use of remote host's CPU

SMC improves throughput, lowers latency and cost, and maintains existing functions

There are two types of Shared Memory Communications:

– SMC-R: Use (*R*emote *D*irect *M*emory *A*ccess) RDMA over Converged Ethernet (RoCE) technology
  • Does not require unique network components (host adapters, switches, security controls, etc.)
  • Utilizes Ethernet fabric with RDMA-capable NICs and switches

– SMC-D: Use DMA when both hosts are within a Z system via virtual PCI device