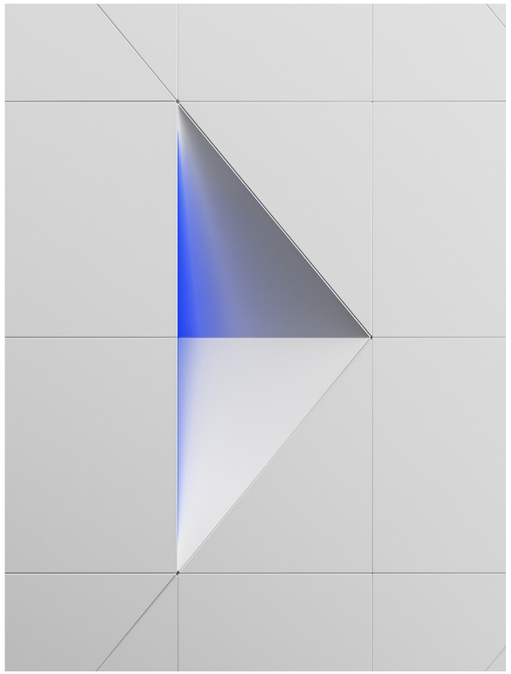
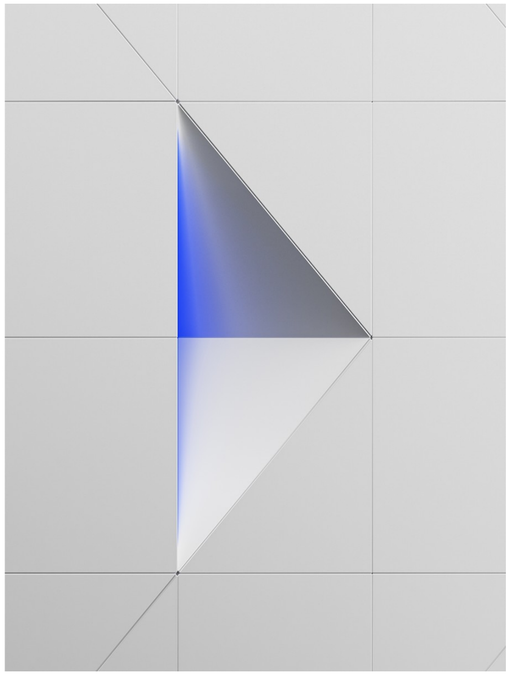
Everything You Need to Know About Networking on IBM z17 With Linux on Z

Stefan Raspl
Principal Product Manager
Linux & Virtualization on IBM Z and LinuxONE

Contents



Contents



z16 SOD: Transition to PCIe-based adapters like RoCE Express as the strategic adapter for Linux on IBM Z & LinuxONE

Statement of Direction

*"In the future, IBM plans to shift from **OSA-Express** to **PCIe-based networking devices like RoCE Express** as the target strategic adapter type for IBM Z **direct access networking connection to Linux** operating systems.*

*[...] Linux on IBM Z clients that indirectly access the OSA-Express adapter family through the **z/VM Virtual Switch (VSwitch)** will be **unaffected** by this change.*

Linux on IBM Z networking currently supports two Ethernet networking connectivity options: the OSA-Express adapter family and the RoCE Express adapter family. Use of PCIe-based networking devices as provided by the RoCE Express adapter family is aligned with the deployment model for Linux on other architectural platforms, facilitates use of broader existing Linux ecosystem tooling, and eases the effort to enable exploitation of industry hardware optimizations and integrate into industry software-defined networking models and tools, including Red Hat OpenShift Container Platform (OCP).

Clients are strongly encouraged to plan accordingly for their adoption of RoCE Express adapters for IBM Z networking connectivity.

*IBM plans to continue to work toward **common networking adapters for all operating systems** on IBM Z, IBM LinuxONE, and Linux on IBM Z."*

***Source**: IBM z16 Announcement Letter*

Benefits of PCI Networking Devices for Linux

- PCI-based networking devices make for a better user experience in the Linux on IBM Z ecosystem:
  - Eliminates need for zSystems-specific tooling
  - Smoother transition from other platforms
  - Less training effort for personnel
  - Better integration with Linux on zSystems distributions
  - Native adapter virtualization
    - ⇛ Better integration with Linux ecosystem
    - ⇛ Facilitates use of operators designed for full cloud-native SDN experience

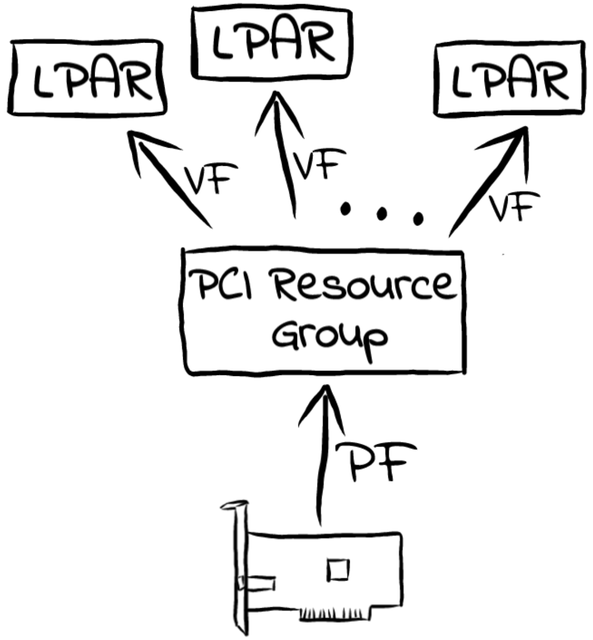


Fig.1: *SR-IOV as implemented in z16 and prior*

OSA- and RoCE Express
Strengths and Strengths

Channel Device

OSA-Express7S

PCI Device

RoCE Express3

Virtualisation

- OSA
  - Up to 480 IP stacks (vNICs) per port
  - z/VM VSWITCH
  - KVM Open vSwitch
  - VNIC characteristics
- RoCE Express 3
  - Up to 63 VFs per port
  - Excellent for shared traffic
  - No promiscuous mode

OSA	RoCE	Aspect
-	++	RDMA and SMC-R
++	+	RAS
+	++	Latency
++	+	Virtualization
++	+(-)	TCP/IP usage

Having to choose…
…is not always a nice thing!

Channel Device

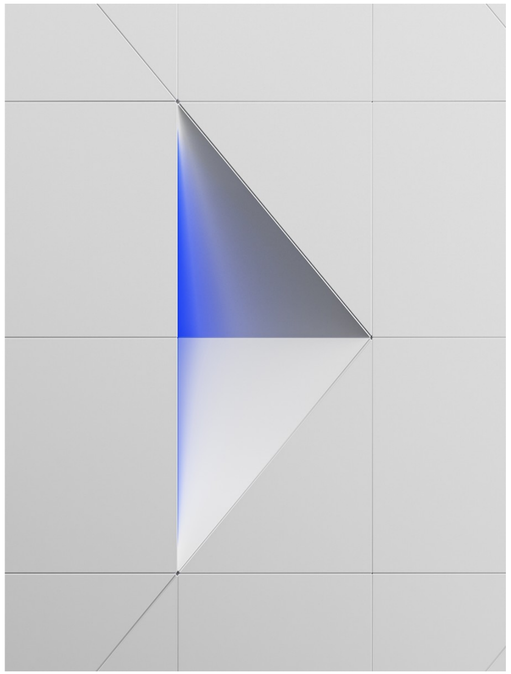
OSA-Express7S

PCI Device

RoCE Express3

- Previously, while Linux could utilize both networking devices as general purpose networking (TCP/IP, Ethernet), all other operating systems had to use a channel device (OSA-Express) for general purpose networking
- Clients wanting to take advantage of RoCE Express adapters with Linux mostly had to have them as extra, because other use cases required OSA.
  - More investment, more administrative effort and less efficient use of resources

Contents



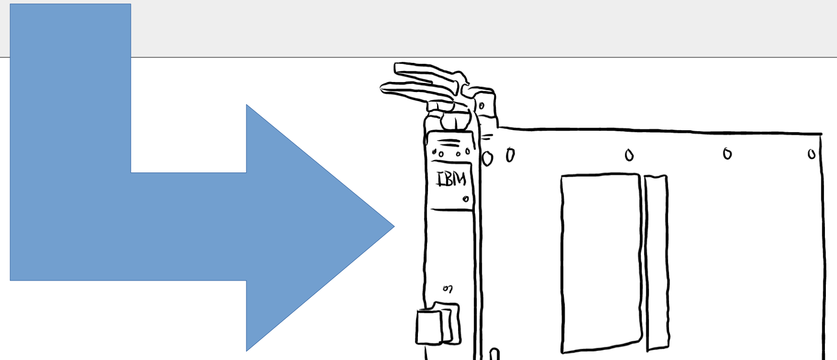
Convergence! Enter “Network Express”!

Channel Device

OSA-Express7S

PCI Device

RoCE Express3

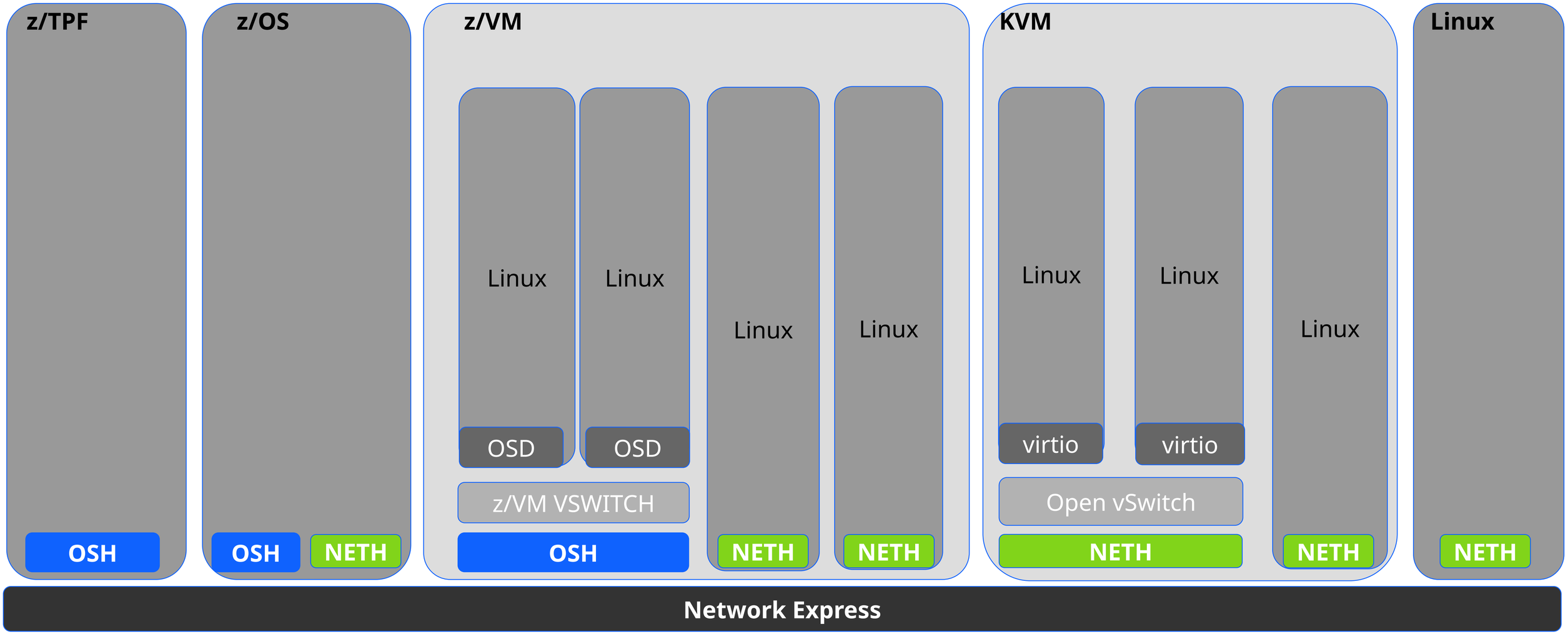


Network Express

- The new Network Express adapter can be configured as both at the same time:
  - NETH FID (PCI device) for Linux
  - OSH channel device for z/OS and z/VM VSWITCH using Enhanced QDIO (EQDIO)
- Supports 10 GbE and 25 GbE

- IOCP example:
  FUNCTION PCHID=100,FID=(2100,2),PART=LP2,TYPE=NETH,VF=8
  FUNCTION PCHID=100,FID=2200,PART=LP2,TYPE=**NETH**,VF=40,**FIDPARM**=01
  CHPID PCHID=100,TYPE=**OSH**,PART=(LP1,LP2,LP3),PATH=80,SHARED

Usage Summary



More Network Express highlights

Fig.1: *Network Express*

- Network Express and the DPU
  - Protocol support (e.g. OSH, FCP) moved across the PCI bus into the Z CP chip, and so did PF support but from the PSP (PCI Resource Group)

    ⇛ convergence enabler
  - Single port serviceability

    ⇛ enhanced RAS compared to RoCE
    ⇛ PCHID per port
  - Shared by PU chips and cores in the drawer

    ⇛ enhanced RAS
- Network Express capacity and scaling
  - max. 48 cards, 2 ports per card, max. 123 NETH FIDs per port
  - PCIe Gen 4
  - promiscuous mode for NETH
  - excellent performance for “shared” traffic

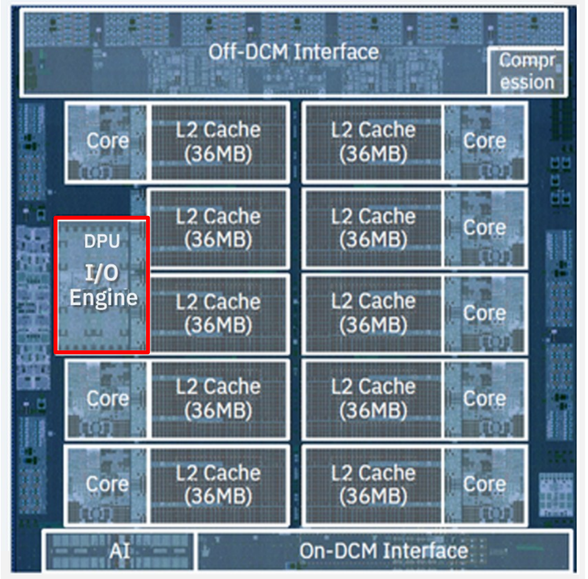


Fig.2: *z17 DPU*

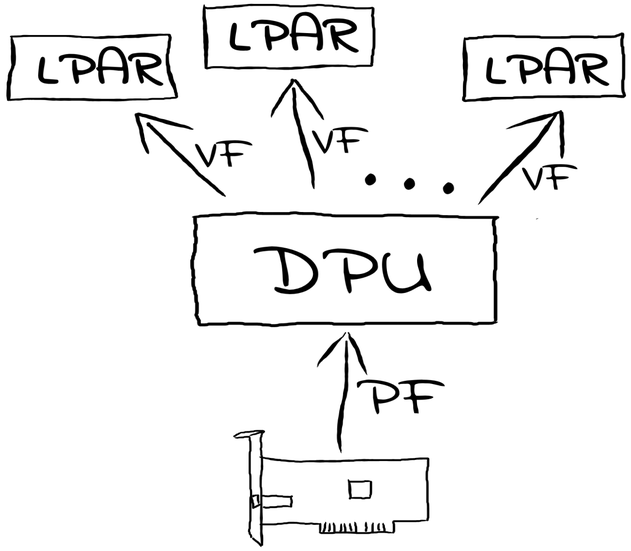


Fig.3: *SR-IOV as implemented in z17*

z17: Adapters and I/O Features
New Builds

Network Express:

- Network Express LR 25G
- Network Express SR 25G
- Network Express LR 10G
- Network Express SR 10G

Note: 25GbE model strictly requires a 25GbE switch port – no auto-negotiation down to 10GbE

OSA-Express7S 1.2:

- OSA-Express7S 1.2 25GbE LR
- OSA-Express7S 1.2 25GbE SR
- OSA-Express7S 1.2 10GbE LR
- OSA-Express7S 1.2 10GbE SR
- *OSA-Express7S 1.2 GbE LX*
- *OSA-Express7S 1.2 GbE SX*

z17: Adapters and I/O Features
New Builds – Number of Ports

Network Express:

- Network Express LR 25G
- Network Express SR 25G
- Network Express LR 10G
- Network Express SR 10G

OSA-Express7S 1.2:

- OSA-Express7S 1.2 25GbE LR
- OSA-Express7S 1.2 25GbE SR
- OSA-Express7S 1.2 10GbE LR
- OSA-Express7S 1.2 10GbE SR
- *OSA-Express7S 1.2 GbE LX*
- *OSA-Express7S 1.2 GbE SX*

2 ports / card 1 ports / card

z17: Adapters and I/O Features Carry Forward

New build or carry forward:

- OSA-Express7S 1.2 25GbE LR
- OSA-Express7S 1.2 25GbE SR
- OSA-Express7S 1.2 10GbE LR
- OSA-Express7S 1.2 10GbE SR
- OSA-Express7S 1.2 GbE LX
- OSA-Express7S 1.2 GbE SX

- ~~RoCE Express~~

Carry forward only from z15:

- OSA-Express7S 10 Gigabit Ethernet LR
- OSA-Express7S 10 Gigabit Ethernet SR
- OSA-Express7S Gigabit Ethernet LX
- OSA-Express7S Gigabit Ethernet SX
- OSA-Express7S 1000BASE-T Ethernet

- ~~RoCE Express~~

Virtualization Capabilities
Network Express vs RoCE Express

Network Express:

Model	*#Cards*	*#Ports / Card*	*Total #Ports*	*#VFs / Port*
z17	48	2	96	123

RoCE Express:

Model	*#Cards*	*#Ports / Card*	*Total #Ports*	*#VFs / Port*
z16	16	2	32	63
z15	16	2	32	63
z14	8	2	16	63
z14 ZR1	4	2	8	63

Promiscuous mode possible --> Open vSwitch support
FIDPARM for allowing/disallowing promiscuous mode

No promiscuous mode --> no Open vSwitch

Migration Considerations

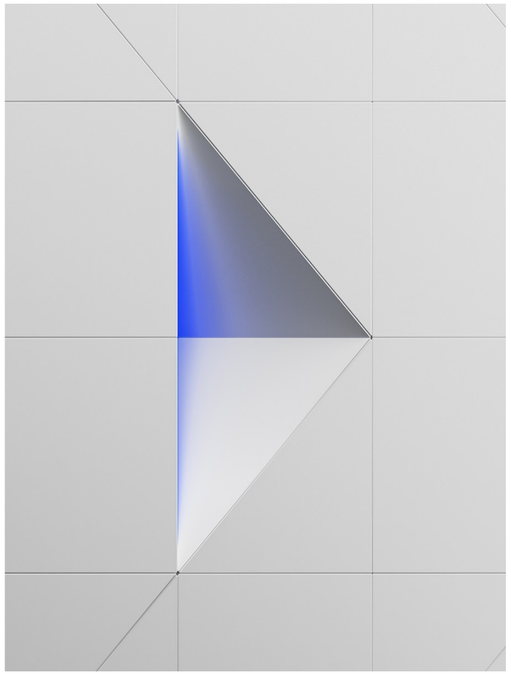
- For the overwhelming majority of cases:
  ROCE|ROC2 → NETH; OSD (Linux) → NETH; OSD (non Linux) → OSH

- Remember the z16 SOD: moving between PCI-based adapters is assumed to be a very smooth and straightforward migration

- Hints moving form Channel I/O based Networking to PCI-based:

	Usage	**Migration Option z17 / *z16***
Linux in LPAR with OSA	Direct-attached OSA	Direct-attached Network Express (NETH) / *RoCE Express*
z/VM	VSWITCH with OSA	VSWITCH with OSH (OSD) / *OSA*
	Direct-attached OSA	Direct-attached Network Express (NETH) / *RoCE Express*
	SSI with direct-attached OSA	None / *None* (no LGR for PCI devices) => Use VSWITCH attachment or direct-attached OSA
KVM	Open vSwitch with OSA	Open vSwitch with Network Express (NETH) / *Remain ‘as is’ (OSA)*

Contents



Operating the Equipment in Linux
or “Device Drivers, Features and Command”

Network Express:

OSA-Express:

z17 Minimum Distribution Requirements:

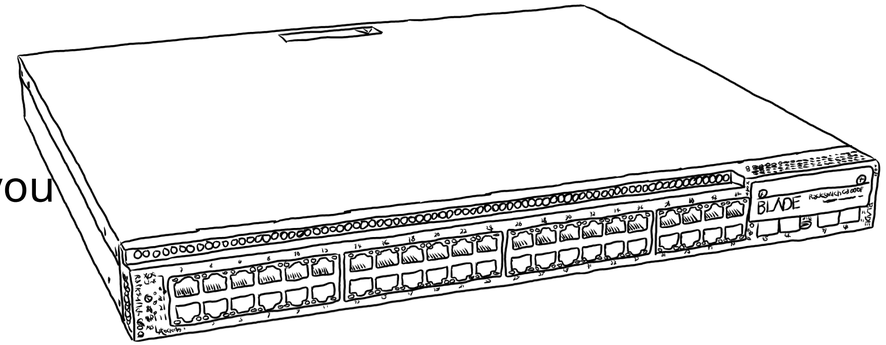
- SUSE SLES 16.0
- SUSE SLES 15.6
- SUSE SLES 12.5
- Red Hat RHEL 10.0
- Red Hat RHEL 9.4
- Red Hat RHEL 8.10
- Canonical Ubuntu 24.04 LTS
- Canonical Ubuntu 22.04 LTS

Network Express NETH Promiscuous Mode

- FIDPARM currently solely used for “promiscuous mode allowed”
  - IOCDs/IOCP: if not set, not allowed
  - In Linux: read only sysfs attribute:

```
$ cat /sys/bus/pci/devices/<pci_dev_id>/fidparm
1
```

- Enabler for things like Open vSwitch

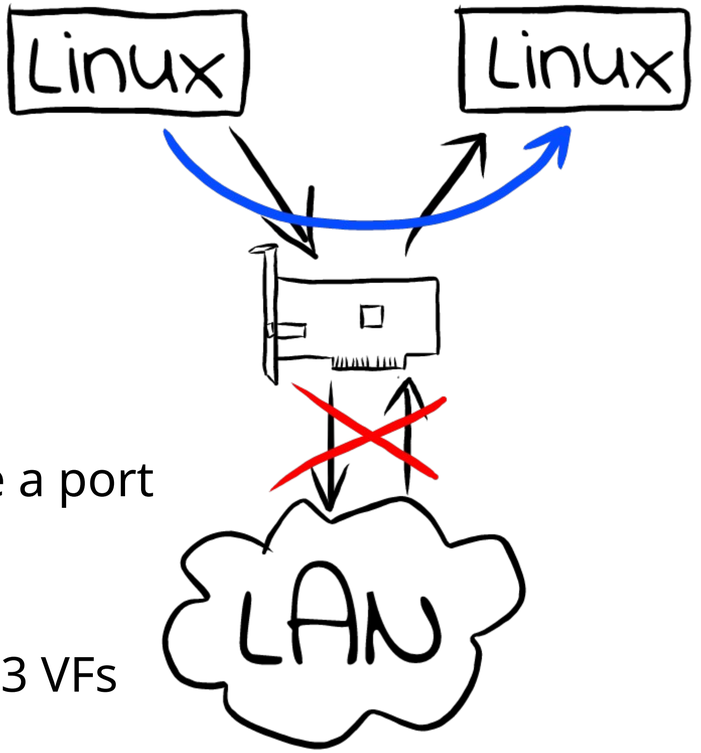


- Linux exploitation works as usual
  - Open vSwitch does it auto-magically for you
  - And so does tcpdump via AF_PACKET
  - Manually e.g. via:

```
$ ip link set eth1 promisc on
```

Network Express
NETH ↔️ OSH
Shared Trafic

- “vNICs sharing a same physical port can talk to each-other directly”
  - Shortcut: latency and throughput
  - In contrast VEPA mandates going out to the LAN
    - Can be desirable for policy reasons
- In the past: OSA-Express could do shared traffic and RoCE Express could do it even better
  - In certain scenarios: shared traffic between RoCE Express 3 VFs preferable to HiperSockets
- Network Express: NETH and OSH can now share a port
  - NETH ↔️ OSH (zOS) shared traffic possible
  - Great performance as between RoCE Express 3 VFs



Refresher: PCI devices activation

- Candidate List and Access List (Classic mode)
- FID X can be on the candidate list for multiple LPARs
- LPARs may try to configure and deconfigure a FID
  - First come first serve

- Linux view:
  - Not on Candidate List or configured to another LPAR
    ⇒ not available
    ⇒ invisible to Linux
    ⇒ no pci slot
    ⇒ no device
  - On Candidate List and not configured to another LPAR
    ⇒ visible to Linux
    ⇒ pci slot → power
    attribute

- Consequences:
  - Cooperative sharing
  - FIDs “start” as not configured
- DPM: exclusive ownership model but FIDs still “start” not configured
- z/VM and KVM passed through devices are passed through as configured

- Configuration state persistent across IMLs
  ⇒ once configured not much to worry about
- In LPAR you may have to mind initially
- Options:
  - Access List
  - Support Element
  - Linux auto-activation (DPM only, distro)
  - `pci/slots/<FID>/power`

Refresher: Identifying PCI devices/functions

- Function Identifier (FID)
  - Mandatory for every VF/device
  - Unique per CPC
  - Not portable - makes migration complicated!
- User Identifier (UID)/ Unique UID (UUID)
  - Optional – if not set in IOCDS, UIDs are assigned automatically, but not persistent!
  - Unique* within guest/partition only
  - Well suited for migrations
  - Assign UIDs based on conventions, e.g. always use UID 1 in each partition for main networking interface
  - Reflected as Domain in PCI ID:

    0001:00:00.0

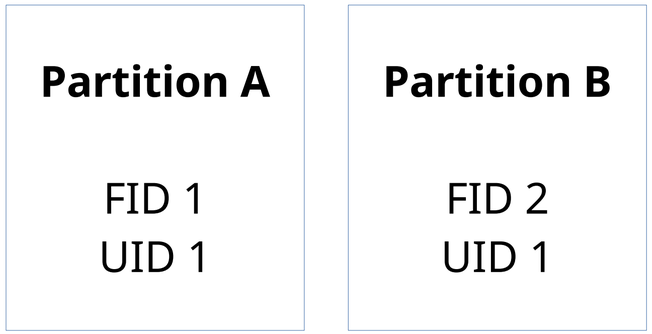


Fig.1: *FIDs are unique within a CPC, UIDs are unique within a partition*

*) The per guest uniqueness of UIDs (which are then also called UUIDs) is enforced by DPM, z/VM and KVM, but for classic mode LPARs unless specified in the IOCDS (IOCP statement: UUID; example `UUID PART=(LP1,LP3,LP4)`).

Refresher: Predictable Interface Names

- A priori knowledge of interface name required for RoCE-only LPAR installs with some Linux distributions

```
ro ramdisk_size=60000 cio_ignore=all,!condev,!1900,!1940
ip=192.168.91.48::192.168.91.49:16:<lpar>:vlan210:none
nameserver=192.168.91.49 domain=<domain>
vlan=vlan210:eno1 inst.repo=... inst.ssh=1
sshpassword=xxx inst.vnc=1 inst.vncpassword=xxx
```

Fig.1: *Interface name specified as part of PRM file on RHEL*

- With *Linux kernel 5.13* and *Systemd v249*:
  - **With unique UIDs enabled (recommended):**
    ⇒ eno<UID_in_dec>
  - With unique UIDs disabled:
    ⇒ ens<FID_in_dec>
  - Beware: This is UID/FID in decimal!
- Before interface names could be any of:
  - ens<FID>
  - enP<UID>s<FID>
  - enP<UID>p0s
  - enP<UID>p0np0
  - enP<UID>p0s0np0
  - enP<UID>p0

Distro support:

RHEL 9, Ubuntu 22.04

Selection criteria would depend upon, among others: UID value, FID being all numeric or not, kernel level, et al.

Trick for installation:

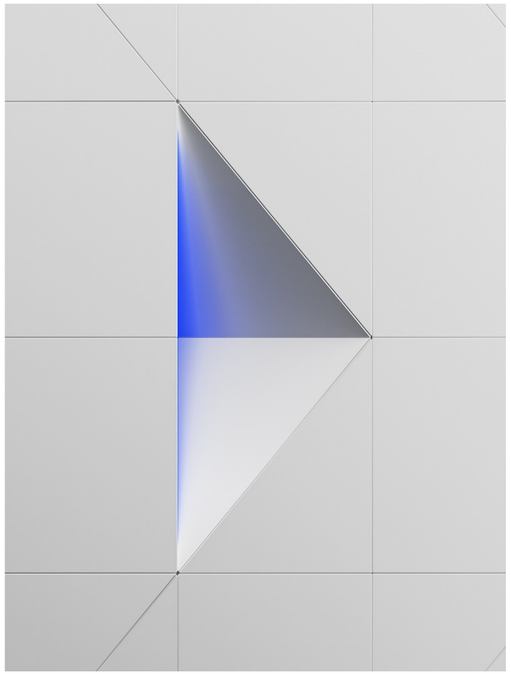
- Install with some name
- Use output on the HMC console to figure out right name
- Install with right name

Refresher: Performance Tuning

- Most applicable “classic” hardware offloads like rx/tx checksumming or TCP segmentation offload enabled per default
- Receive Packet Steering (RPS)
  - ly across specified CPUs
  - improvements, especially with many connections and small packet sizes
- Receive Flow Steering (RFS)
  - Similar to RPS, but considers location of userspace process consuming inbound traffic
- Striding RQ
  - More efficient use of inbound buffers
- Interrupt Moderation
  - Modify waiting behavior on inbound packets before sending an interrupt
  - Can save CPU cycles at the cost of latency and throughput, or improve latency at the cost of CPU

NOTE: As some of these options can have counter-productive effects in corner cases, ***always*** have a **representative benchmark** ready to verify that you are indeed improving performance!

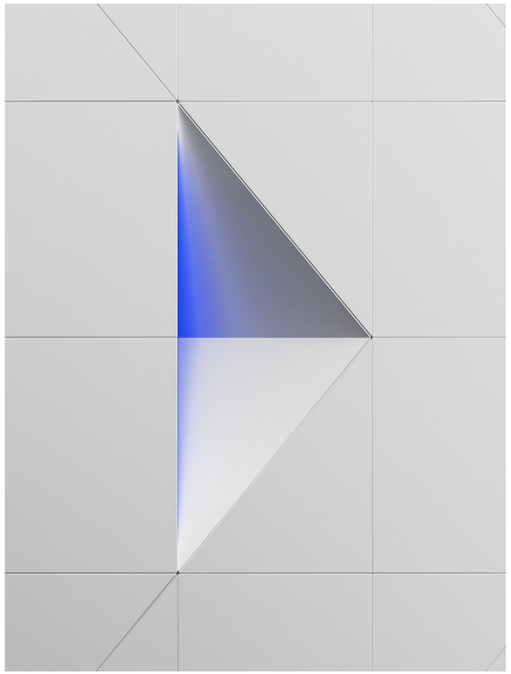
Contents



Summary

- Network Express is best of both worlds (RoCE Express and OSA-Express),and the next level PCI accessed Networking Adapter
- Enhancements over RoCE Express 3
  - Enhanced RAS
  - Promiscuous mode
  - NETH ↔️ OSH shared traffic
  - 123 VFs per port
- Cost reduction and simplification
- For Linux Network Express NETH looks and feels like RoCE Express!
  - Same tools same procedures
  - Same performance tuning options
  - Added capabilities integrate seamlessly

Contents



References

- “IBM z17 (9175) Technical Guide"
  https://www.redbooks.ibm.com/redpieces/abstracts/sg248579.html

“Device Drivers, Features, and Commands”
https://www.ibm.com/docs/en/linux-on-systems?topic=configuration-device-drivers-features-commands

- “Networking with PCI Adapters and Functions”
  https://www.ibm.com/docs/en/linux-on-systems?topic=express-networking-pci-adapters-functions
- Performance of Linux on Z Networking Adapters
  https://ibm.biz/BdPFbw
- Linux on Z Documentation
  https://www.ibm.com/docs/en/linux-on-systems?topic=linux-z-linuxone
- Webcasts
  http://ibm.biz/Linux-on-IBMZ-LinuxONE-Webcasts

Your feedback is important!

SHARE
EDUCATE ▸ NETWORK ▸ INFLUENCE

Submit a session evaluation for each session you attend:

www.share.org/evaluation

Experience more with IBM

Visit us at the IBM Booth #113

After a full day of technical sessions, take a break with us!

Connect with our experts, snap a photo with the z17 Plexi or the latest Telum II, and get an up-close look at our Spyre Accelerator.

Come back each day for fresh topics and demos at our expert stations.

Think 2026

Join 5000+ senior business and technology leaders who are seizing the AI revolution to unlock unprecedented growth and productivity at **Think 2026**.

Find out more information using the QR code below.

IBM Digital Asset Haven

IBM Digital Asset Haven is the operational backbone for financial institutions and regulated enterprises entering the digital asset economy.

Find out more information using the QR code below.