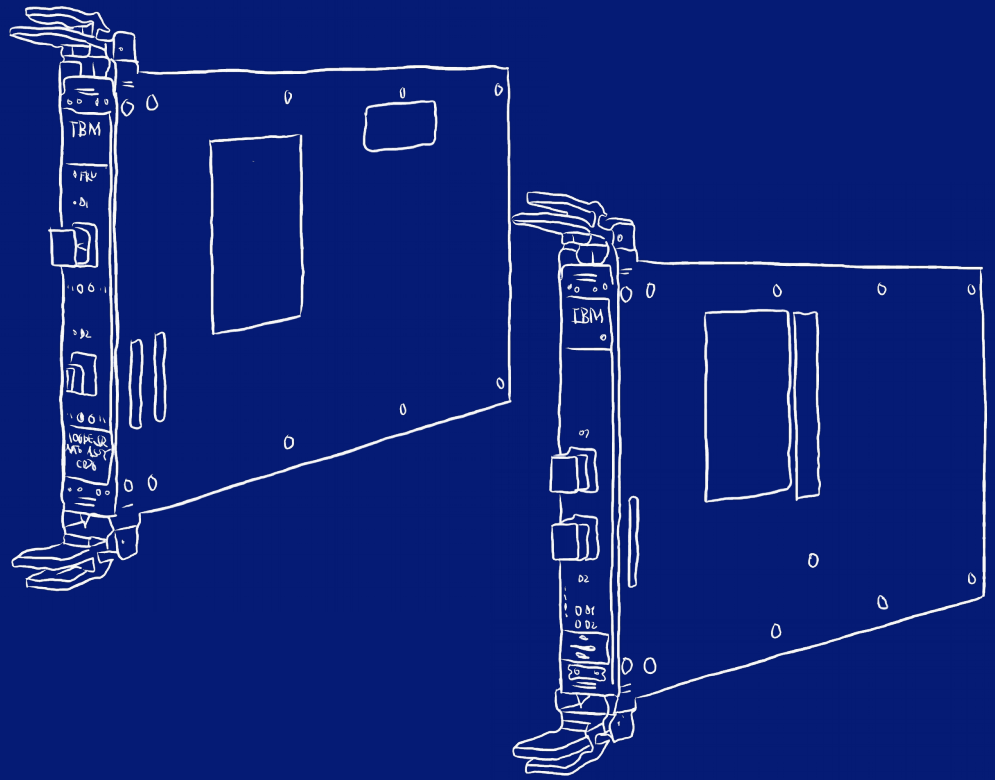
Linux on IBM Z Networking:
OSA-Express and RoCE Express Side by Side

—

Stefan Raspl

Linux on IBM Z Development



Trademarks

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

AIX*	DB2*	HiperSockets*	MQSeries*	PowerHA*	RMF	System z*	zEnterprise*	z/VM*
BladeCenter*	DFSMS	HyperSwap	NetView*	PR/SM	Smarter Planet*	System z10*	z10	z/VSE*
CICS*	EASY Tier	IMS	OMEGAMON*	PureSystems	Storwize*	Tivoli*	z10 EC	
Cognos*	FICON*	InfiniBand*	Parallel Sysplex*	Rational*	System Storage*	WebSphere*	z/OS*	
DataPower*	GDPS*	Lotus*	POWER7*	RACF*	System x*	XIV*		

* Registered trademarks of IBM Corporation

The following are trademarks or registered trademarks of other companies.

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.
Cell Broadband Engine is a trademark of Sony Computer Entertainment, Inc. in the United States, other countries, or both and is used under license therefrom.
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.
IT Infrastructure Library is a registered trademark of the Central Computer and Telecommunications Agency which is now part of the Office of Government Commerce.
ITIL is a registered trademark, and a registered community trademark of the Office of Government Commerce, and is registered in the U.S. Patent and Trademark Office.
Java and all Java based trademarks and logos are trademarks or registered trademarks of Oracle and/or its affiliates.
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
Linux is a registered trademark of Linus Torvalds in the United States, other countries, or both.
Microsoft, Windows, Windows NT, and the Windows logo are trademarks of Microsoft Corporation in the United States, other countries, or both.
OpenStack is a trademark of OpenStack LLC. The OpenStack trademark policy is available on the OpenStack website.
TEALEAF is a registered trademark of Tealeaf, an IBM Company.
Windows Server and the Windows logo are trademarks of the Microsoft group of countries.
Worklight is a trademark or registered trademark of Worklight, an IBM Company.
UNIX is a registered trademark of The Open Group in the United States and other countries.

* 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

- **The Cards**
  - Models & Features
  - Virtualization Capabilities
- **Device Drivers, Features and Commands**
- **Usage**
- **Performance**
- **Summary**
- **References**

OSA-Express

OSA-Express6S 10GbE

RoCE Express

RoCE Express2 10GbE

OSA-Express

- 1, 10 and 25GbE models with varying HW features:
  - 1GbE: Base-T or fiber optics, 2 ports
  - 10 and 25GbE: Fiber only, 1 port
- 25GbE model strictly requires 25GbE capable switch – no negotiation to 10GbE
- Considered platform's native networking card
- Supported by all operating systems on IBM Z
- Supports TCP/IP[1] traffic only

Feature	z14	z13	zEC12
OSA-Express7S	25 GbE		
OSA-Express6S	10 GbE 1 GbE 1000Base-T		
OSA-Express5S	10 GbE 1 GbE 1000Base-T	10 GbE 1 GbE 1000Base-T	10 GbE 1 GbE 1000Base-T
OSA-Express4S	10 GbE 1 GbE 1000Base-T	10 GbE 1 GbE 1000Base-T	10 GbE 1 GbE 1000Base-T
OSA-Express3			10 GbE 1 GbE 1000Base-T

RoCE Express

- Introduced with zEC12 for SMC-R
- 10 and 25GbE models, optical connectors only
- 25GbE model strictly requires 25GbE capable switch – no negotiation to 10GbE
- All models feature 2 ports
- Fiber optics only
- TCP/IP[1] or RoCE (RDMA over Converged Ethernet)
- TCP/IP functionality exploited by Linux only

Feature	z14	z13	zEC12
RoCE Express 2	25 GbE 10 GbE		
RoCE Express	10 GbE	10 GbE	10 GbE

[1] *Synonymous to any kind of "traditional" network traffic (UDP, SCTP, etc.)*

OSA-Express

- Multiple operating modes, configurable on a per-CHPID basis
- Covered in this presentation:
  - **OSD**: Queued Direct Input/Output (QDIO)
  - **OSE**: Non-Queued Direct Input/Output
- Other modes:
  - **OSM**: Required for DPM, connectivity to intra-node management network (INMN)
  - **OSX**: Connectivity to zEnterprise BladeCenter Extension
- Supported up to z13 and OSA-Express5S only:
  - **OSN**: Open Systems Adapter for Network Control Program
- Unsupported by *Linux on Z*:
  - **OSC**: OSA Integrated Console Controller

RoCE Express

- Operating mode chosen by software – can be used in parallel
  - TCP/IP
  - RDMA over Converged Ethernet (RoCE)

OSA-Express

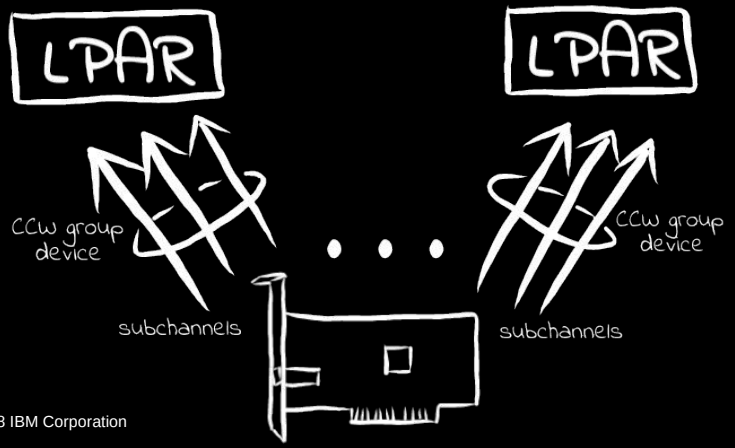
- **Features** (selection)
  - HW offloads: Checksumming, TCP segmentation offload (***TSO***)
  - Layer 2 and layer 3 mode
  - VLAN, QoS, VIPA, ARP, et al
- **RAS**
  - Extended RAS
  - Concurrent firmware updates
    95+ percent completely concurrent

RoCE Express

- **Features** (selection)
  - HW offloads: Checksumming, TSO
  - RDMA over Converged Ethernet (RoCE)
  - Flow Control, Explicit Congestion Notification
  - IPoIB, uDAPL, et al
  - VLAN, QoS, et al
- **RAS**
  - Regular RAS
  - Changing optics of a single card disrupts entire PCHID
  - Firmware updates are disruptive

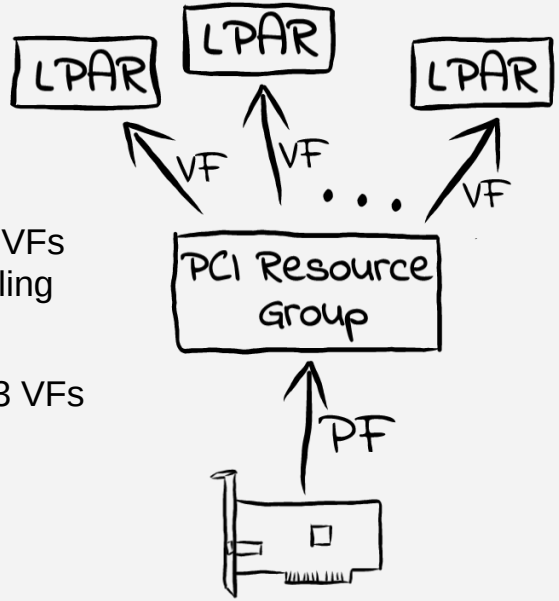
OSA-Express

- Up to 1,920 subchannels per card
  - 1,920 subchannels per card at ***one*** outbound queue
  - 480 subchannels per card at ***four*** outbound queues
- 3 subchannels form a so-called ***CCW group device*** required per stack for OSD CHPIDs
  ⇒ 160 to 640 stacks per card
- Each virtualized instance provides full functionality



RoCE Express

- Single Root I/O Virtualization (***PCI SR-IOV***)
- **Virtual Functions** (***VFs***) provide a limited subset of card functionality
- **Physical function** (***PF***) held by PCI Resource Group
  ⇒ required for certain functionalities, including firmware updates
- ***RoCE Express***: Up to 31 VFs per card(!), each VF handling both ports
- ***RoCE Express2***: Up to 63 VFs per port



OSA-Express

Model	*#Cards*	*#Ports / Card*	*Total #Ports*	*#IP Stacks / Card*[1]	Total #IP Stacks / Machine
z14	48	1-2	48-96	160-640	**3,840-30,720**
z14 ZR1	48	1-2	48-96	160-640	**3,840-30,720**
z13	48	1-2	48-96	160-640	**3,840-30,720**
z13s	48	1-2	48-96	160-640	**3,840-30,720**
zEC12	48	1-2	48-96	160-640	**3,840-30,720**
zBC12	48	1-2	48-96	160-640	**3,840-30,720**
zBC12 w/ OSA-Express3	16-32	1-2	16-64	160-640	**1,280-20,480**

[1] *Depends on #outbound queues mode*

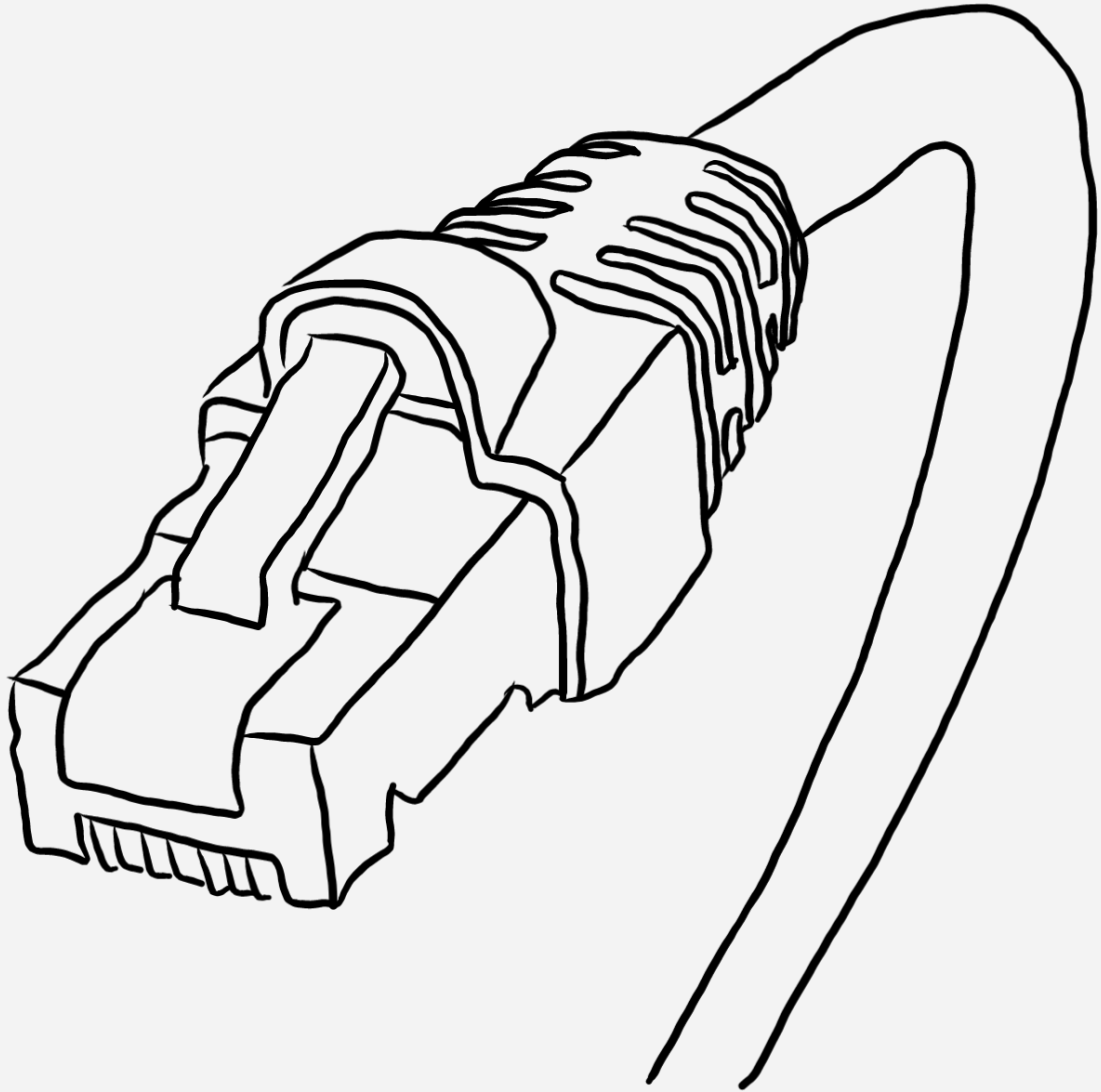
RoCE Express

Model	*#Cards*	*#Ports / Card*	*Total #Ports*	*#IP Stacks / Card*[2]	Total #IP Stacks / Machine
z14	8	2	16	31-126	**328-1008**
z14 ZR1	4	2	8	31-126	**164-504**
z13	16	2	32	31	**496**
z13s	16	2	32	31	**496**
zEC12	16	2	32	2	**64**
zBC12	16	2	32	2	**64**

[2] *Depends on card generation*

Agenda

- **The Cards**
- **Device Drivers, Features and Commands**
  - Distro Support
  - Device Drivers
  - Tools
- **Usage**
- **Performance**
- **Summary**
- **References**



OSA-Express

- **All OSA-Express models**
  - All Linux distributions in service
- **OSA-Express7S**
  - Correct link speed display requires
    - RHEL 7.7 or later
    - SLES 12 SP4 or later
    - Ubuntu 19.04 or later

RoCE Express

- **RoCE Express**
  - RHEL 7 or later
  - SLES 12 or later
  - Ubuntu 16.04 LTS or later
- **RoCE Express2**
  - RHEL 7.3 with service or later
  - SLES 12 SP3 with service or later
  - Ubuntu 16.04 LTS with service or later
- **z/VM**: v6.3 with service or later for PCI passthrough support

OSA-Express

- **`qeth`**
  - OSD CHPIDs
  - Covers all OSA-Express models
  - Subject of the remainder of this presentation
- **`lcs`** (alternative driver):
  - OSE CHPIDs
  - IP address must be set in OSA/SF
  - Utilizes regular CCW instead of QDIO mode ⇒ inferior performance

RoCE Express

- **`mlx4`**: RoCE Express
- **`mlx5`**: RoCE Express2

OSA-Express

- **CCW group device**
  Consists of three device numbers:
  - *Read* device (control data ⇐ OSA)
  - *Write* device (control data ⇒ OSA)
  - *Data* device (network traffic)

```
$ lszdev qeth
TYPE  ID                          ON   PERS  NAMES
qeth  0.0.1240:0.0.1241:0.0.1242  no   no
qeth  0.0.bd00:0.0.bd01:0.0.bd02  yes  yes   encbd00
```

- **Physical identifier**: Card identified by PCHID
- **Hotplugging**: Only group devices can be set online, not the individual devices.
  Use *znetconf* command:

```
# consecutive device numbers assumed if only
# one specified
$ znetconf --add 8000
Scanning for network devices...
Successfully configured device 0.0.8000 (enc8000)
```

RoCE Express

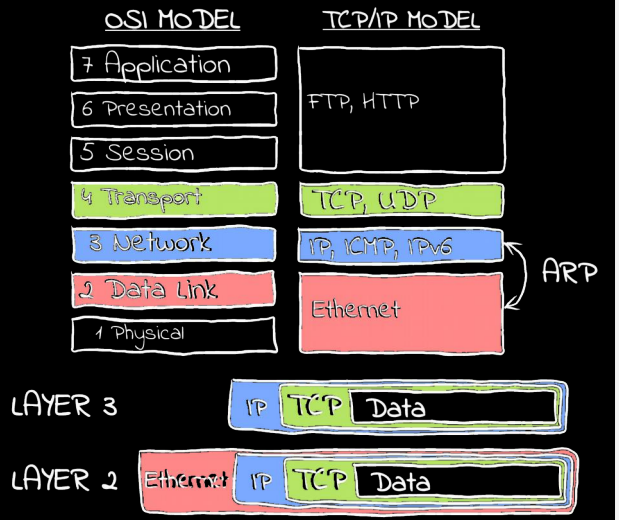
- Regular **PCI device**
- **Physical Identifier**:
  - RoCE Express: FID identifies card
  - RoCE Express2: FID identifies port
- **Hotplugging**: Requires knowledge of FID. E.g. to hotplug RoCE Express with FID 0x80, run:

```
# hotplug device
$ echo 1 > /sys/bus/pci/slots/00000080/power

# check for device's availability
$ lspci
0000:00:00.0 Ethernet controller: Mellanox
             Technologies MT27500/MT27520 Family
             [ConnectX-3/ConnectX-3 Pro Virtual
             Function]
```

OSA-Express

- Multiple layer modes:
  - ***Layer 2*** (**default**, recommended): Maximum compatibility with Linux tooling and frameworks
  - ***Layer 3***: Reduced compatibility.
    OSA handles ARP, special support for VIPA, Proxy ARP, IP Address Takeover.
- Specify layer mode as option, e.g.

```
# configure as layer 3 device (non-default)
$ znetconf --add 8000 --option layer2=0
```



RoCE Express

- Features:
  - Layer 2 only
  - ARP / TCP/IP implemented in the Linux kernel

OSA-Express

- **Specifying Ports** (starts at 0)
  - Use `portno` attribute, e.g. znetconf
    `$ znetconf --add 8000 --option portno=1`
  - Each CCW group device corresponds to one port on the card. I.e. second CCW group device required to utilize both ports.
- **Receive Packet Steering (RPS)**:
  - Direct packets to specific CPUs to take advantage of hot caches
    ⇒ Meaningful with multiple CPUs only
  - Supported.

RoCE Express

- **Specifying Ports**:
  - RoCE Express: One FID per device (two ports).
  - RoCE Express2: One FID per port
  - Note: Ethernet ports start at 0, while RDMA device ports start at 1!
- **Receive Packet Steering (RPS)**:
  - Direct packets to specific CPUs to take advantage of hot caches
    ⇒ Meaningful with multiple CPUs only
  - Can provide good performance improvements, especially with many connections and small packet sizes.

Device offloads:
- Use ethtool to configure checksum and TCP segmentation offloads.
- E.g. offload checksumming of inbound IP packages:

```
$ ethtool -K eth0 rx on
```

OSA-Express

- **`lsqeth`**: List all devices handled by the *qeth* device driver. Also includes HiperSockets.

```
$ lsqeth -p
devices      CHPID i'face  cardtype  port prio-q'ing rtr4 rtr6 lay'2 cnt
------------------------------------------------------------------------------
0.0.bd00/... x85   encbd00 OSD_10GIG 0    always_q_2 n/a  n/a  1     64
0.0.f500/... x76   encf500 OSD_1000  0    always_q_0 n/a  n/a  1     64
```

- **`lszdev`**: Like `lsqeth` when used as follows:
  `$ lszdev qeth`
- **`qethqoat`**:
  - List registered MAC or IP addresses, depending on layer mode (current operating system instance only)
  - physical and logical device information

  Use to provide additional information in case of service calls instead of OSA/SF
- Commands for layer 3 devices only:
  - **`getharp`**:
    Display ARP cache contents.
  - **`gethconf`**:
    Configure IPA, VIPA & Proxy ARP

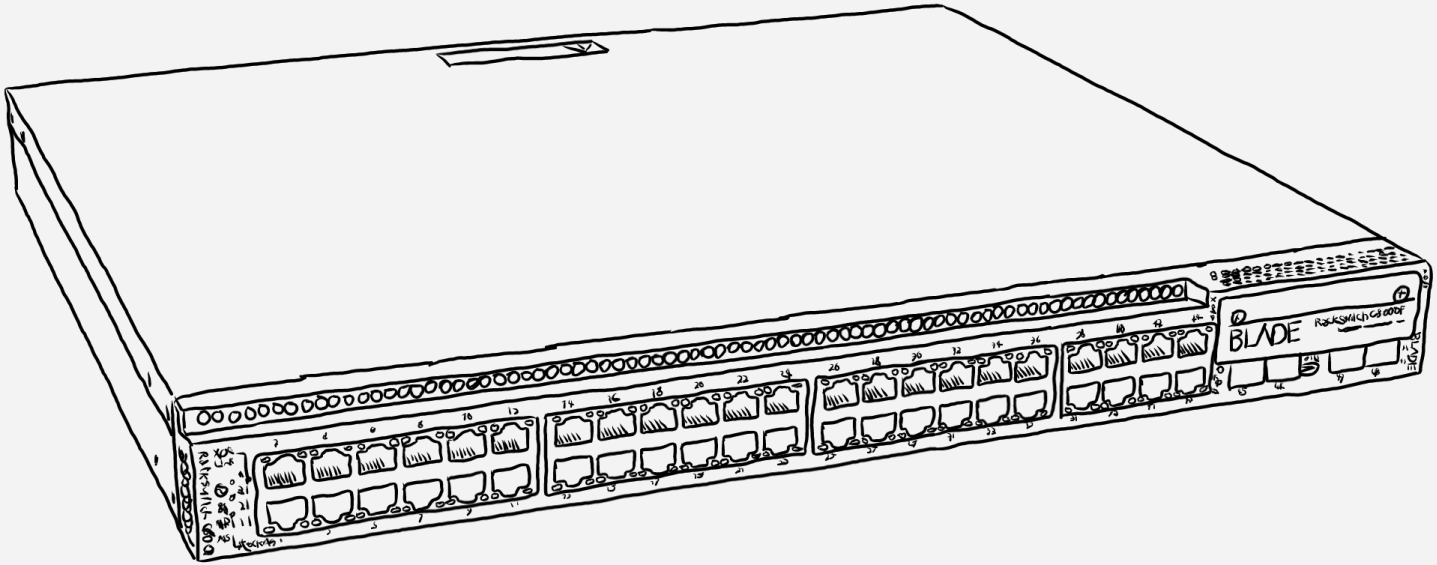
```
$ qethqoat eth2
PCHID: 0x0240
CHPID: 0x92
Manufacturer MAC address: 00:14:5e:76:ed:26
Configured MAC address: 00:00:00:00:00:00
Data device sub-channel address: 0xe202
CULA: 0x00
Unit address: 0x02
Physical port number: 0
Number of output queues: 1
Number of input queues: 1
Number of active input queues: 0
CHPID Type: OSD
Interface flags: 0x00000000
OSA Generation: OSA-Express3
Port speed/mode: 1000 Mb/s / full duplex
Port media type: copper
Jumbo frames: yes
Firmware: 0x00000085

IPv4 router: no
IPv6 router: no
IPv4 vmac router: no
IPv6 vmac router: no
Connection isolation: not active
IPv4 assists enabled: 0x00000000
IPv6 assists enabled: 0x00215c60
IPv4 outbound checksum enabled: 0x00000000
IPv6 outbound checksum enabled: 0x00000000
IPv4 inbound checksum enabled: 0x00000000
IPv6 inbound checksum enabled: 0x00000000
L2 vmac: 02:00:00:f2:e3:bb

gmac
----
33:33:00:00:00:01
01:00:5e:00:00:01
33:33:ff:f2:e3:bb
```

RoCE Express

- **`zpcictl`**:
  - Part of *s390-tools*
  - Handle defective PCI devices
- Other tools provided by the vendor would require access to the PF.

Agenda

- **The Cards**
- **Device Drivers, Features and Commands**
- **Usage**
  - LPAR
  - z/VM
  - SMC-R
- **Performance**
- **Summary**
- **References**



OSA-Express

- **`znetconf`**: List, configure and remove *qeth*-based network devices
  - All changes are not persistent
  - Example:

```
$ znetconf --unconfigured
Scanning for network devices...
Device IDs                Type    Card Type  CHPID Drv.
------------------------------------------------------------
0.0.b100,0.0.b101,0.0.b102 1731/01 OSA (QDIO)    8a qeth

$ znetconf --add b100 -o isolation=drop
Scanning for network devices...
Successfully configured device 0.0.b100 (encb100)

$ znetconf --remove b100
Remove network device 0.0.b100
(0.0.b100,0.0.b101,0.0.b102)?
Warning: this may affect network connectivity!
Do you want to continue (y/n)?y
Successfully removed device 0.0.b100 (encb100)
```

- **`chzdev`**: (Persistently) configure and remove (*qeth*) devices on Z
  - Example:

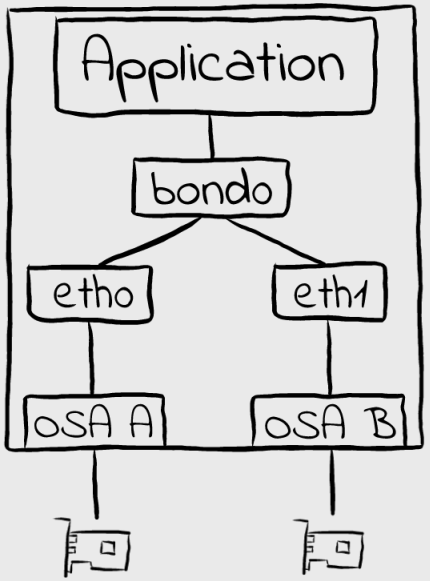
```
$ chzdev -e qeth 0.0.8000 isolation=drop
QETH device 0.0.b100:0.0.b101:0.0.b102 configured

$ chzdev -d qeth b100
QETH device 0.0.b100:0.0.b101:0.0.b102 deconfigured
```

RoCE Express

- PCI devices remain configured / unconfigured according to last state change

Linux `bonding` Driver

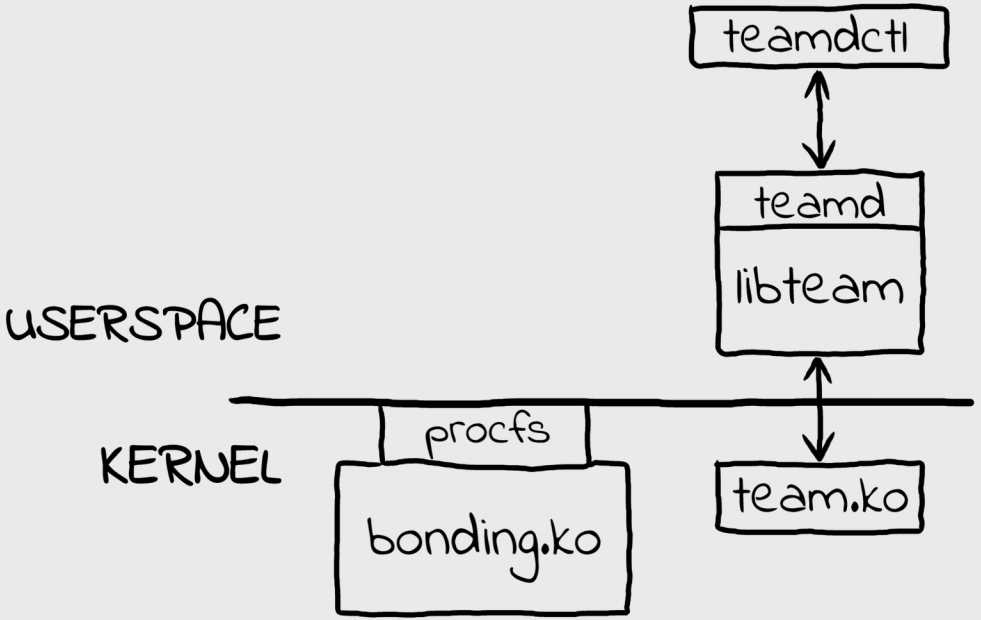
- Use Linux **`bonding`** driver to aggregate multiple network interfaces into a single logical “bonded” interface.
- Recommended driver for channel bonding
- Works with both, OSA-Express and RoCE Express cards
  - However: For OSA, layer 2 devices only!
- Various modes available, providing HA or load-balancing functionality.
- See white paper
  Linux Channel Bonding Best Practices and Recommendations
  for further details.




```
# load bonding module with miimon
# option (enables link monitoring)
$ modprobe bonding miimon=100 mode=balance-rr

# add MAC addresses to slave devices eth0 & eth1
# (not necessary for VSWITCH)
$ ip link set dev eth0 address 00:06:29:55:2A:01
$ ip link set dev eth1 address 00:05:27:54:21:04

# activate the bonding device bond0
$ ip addr add 10.1.1.1/24 dev bond0

# connect slave devices eth0 & eth1 to
# bonding device bond0
$ ifenslave bond0 eth0 eth1
```

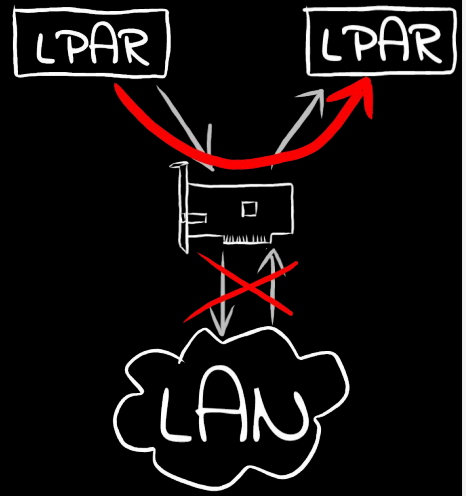
Teaming Driver

- Alternative to Linux kernel's "*bonding*" module:
  "*Solve the same problem using a different approach*"
  ⇒ *comparable functionality*
- Works with both, OSA-Express and RoCE Express cards
  - OSA: Layer 2 devices only
- Different architecture, relying on userspace components
- Different terminology as compared to bonding driver:
  - "*team*" vs "*bond*" device
  - "*ports*" vs "*slaves*"
  - "*runners*" vs "*bonding modes*"
- Various programming APIs
- See http://libteam.org/ for further details




```
# start teaming daemon in background,
# creates instance team0 in round-robin mode
$ teamd -d

# add ports (=slaves)
$ teamdctl team0 port add eth1
$ teamdctl team0 port add eth2

# add IP address and activate
$ ip addr add 192.168.3.37 dev team0
$ ip link set team0 up
```

OSA-Express



- Shortcut within device
- No extra configuration required
- Will not work with TSO enabled
- Works with all operating system images on Z
- **Controlling shared traffic**:
  - VEPA (*Virtual Edge Port Aggregator*) mode: Send all traffic to adjacent switch for consistent enforcement of security policy. Requires reflective relay mode in switch! Example:
    ```
    $ echo forward >
    /sys/devices/qeth/0.0.e200/isolation
    ```
  - Drop any traffic intended for other OS image sharing the same OSA device:
    ```
    $ echo drop >
    /sys/devices/qeth/0.0.e200/isolation
    ```

RoCE Express

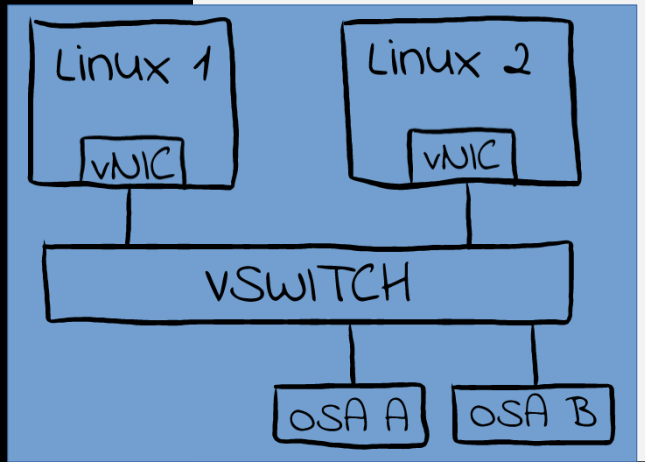
- Excellent throughput
- Shared TCP/IP traffic works with Linux images only due to lack of support in other operating systems. I.e. no shared Ethernet traffic with
  - z/OS
  - z/VSE
  - z/VM
- Shared RDMA traffic (SMC-R) with z/OS works
- No controls for control shared traffic

OSA-Express

- **Passthrough Of Real Devices**
  - Attach OSA device to Linux guest:
    `#CP ATTACH <devno_range> to <guest>`
  - Configure in guest just like in LPAR case (includes channel bonding)
- **VSWITCH**
  - Provides high availability and link aggregation
  - Supports both, layer 2 (keyword `ETHERNET`) and layer 3 (keyword `IP`) devices
  - Sample z/VM configuration:
    ```
    #CP DEFINE VSWITCH <name> ... ETHERNET …
    #CP DEFINE NIC <vdev> QDIO
    #CP COUPLE <vdev> <guest> <name>
    ```
  - Configure vNIC just like a regular device in LPAR case
  - **Note**: vNIC's layer mode must match the one of the VSWITCH

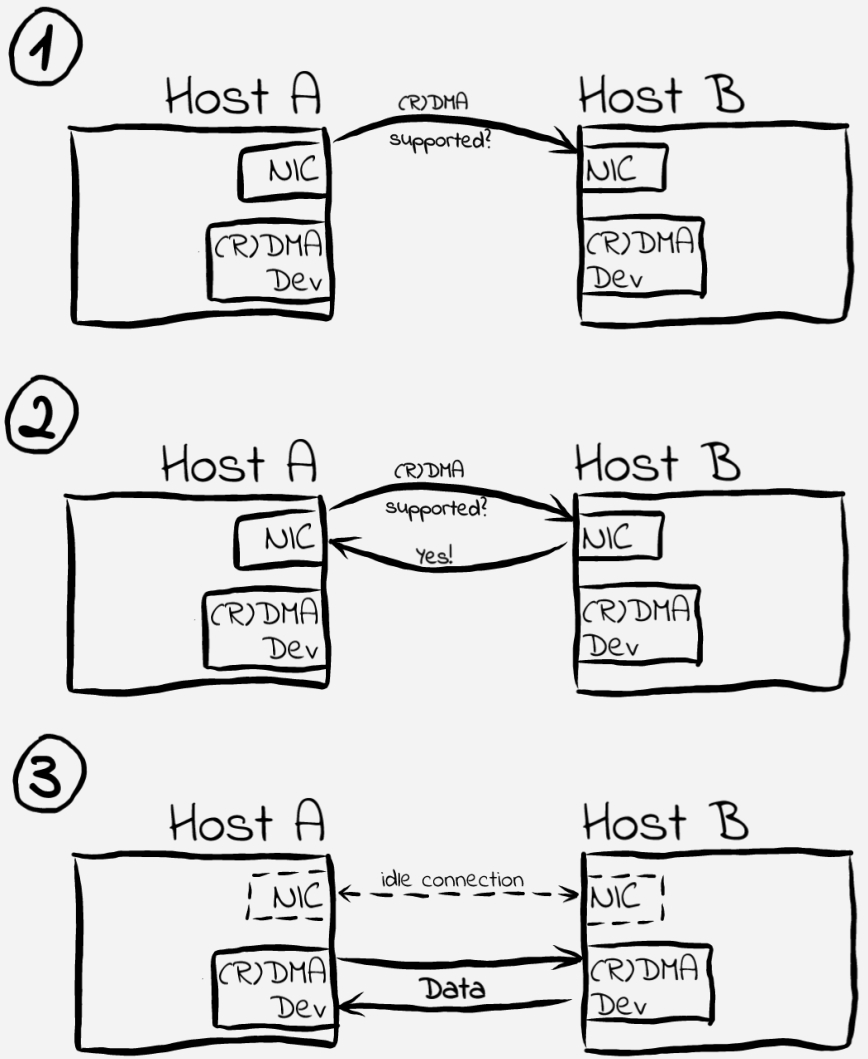
RoCE Express

- **Passthrough Of Real Devices**
  - Attach PCI FID to Linux guest:
    `#CP ATTACH PCIFUNCTION <FID> to <guest>`
  - Configure in guest just like in LPAR case (includes channel bonding)
- **VSWITCH**
  - Not supported



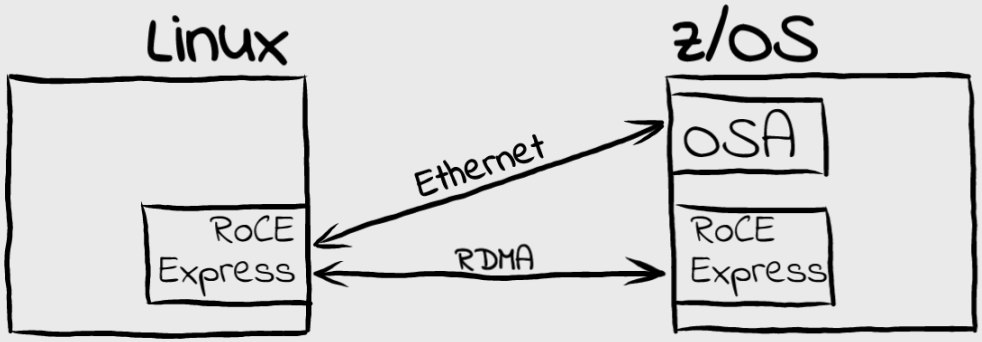
Overview

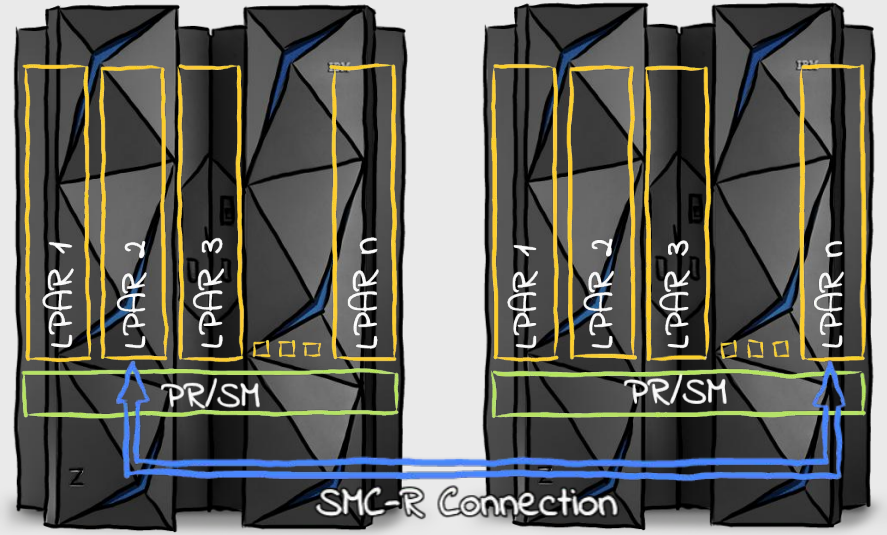
- For each new TCP connection:
  - Start out with a regular TCP/IP connection, advertising (R)DMA capabilities
  - If peer confirms, negotiate details about the (R)DMA capabilities & connectivity
  - Switch over to an (R)DMA device for actual traffic depending on the peers' capabilities
  - Regular TCP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



SMC-R Overview

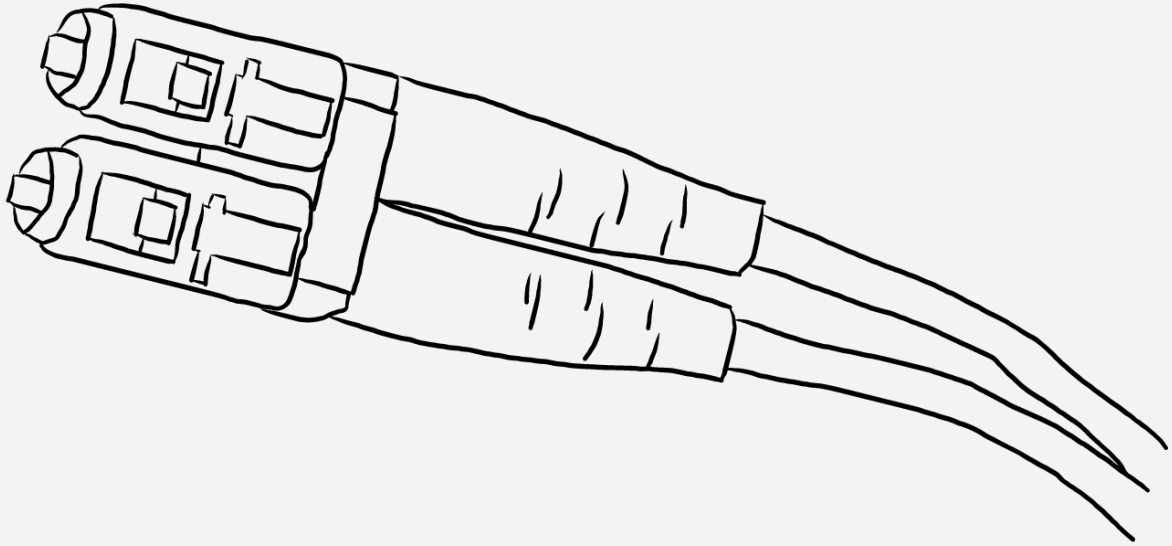
- Cross-CEC connectivity using ***RoCE Express*** cards
- Use OSA or RoCE card for regular connectivity
- Linux on Z can use a single RoCE card for regular and RDMA traffic!





Agenda

- **The Cards**
- **Device Drivers, Features and Commands**
- **Usage**
- **Performance**
  - 10 GbE Cards
  - 10 vs 25GbE Cards
- **Summary**
- **References**



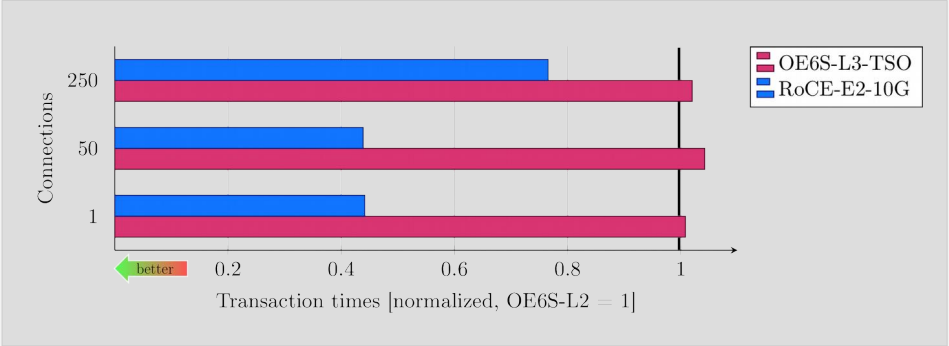
Setup

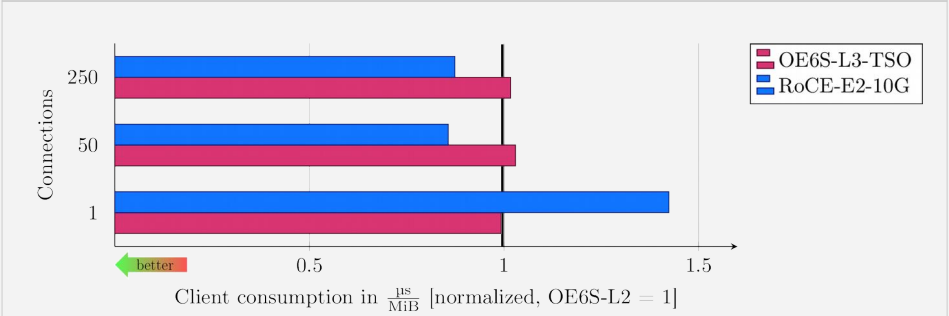
- **Machine**: IBM z14
- **Configuration**: 2 LPARs, each with
  - 4 IFLs – SMT-2
  - Memory per LPAR: 4GB
- **Linux Distro**: SLES 15 GA
- **MTU Size**: Default MTUs
  - OE6S-L3-TSO: 1492 Bytes
  - All others: 1500 Bytes
- **Benchmark**: *uperf*, see https://github.com/uperf/uperf

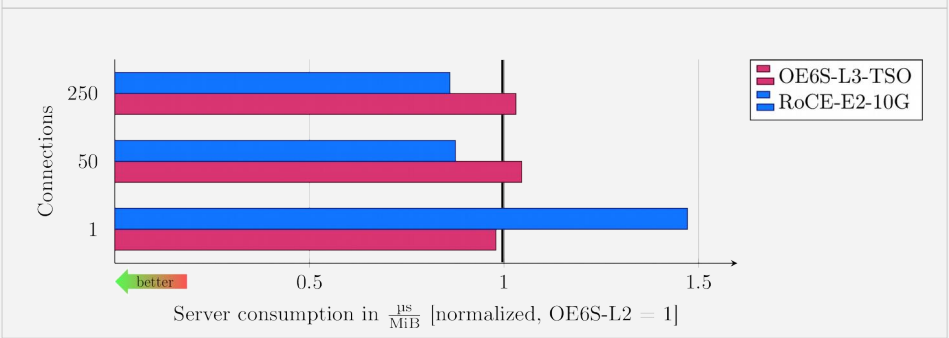
Offload	OE6S-L2	OE6S-L3-TSO	RoCE Express2 10GbE
Layer	2	3	2
rx-checksumming	off	**on**	**on**
tx-checksumming	off	**on**	**on**
tcp-segmentation-offload (TSO)	off	**on**	**on**
receive-packet-steering (RPS)	off	off	**on**

10GbE Cards

- **Note**: All measurements normalized to OE6S-L2
- **Workload**: `rr1c-200x1000`
  - Client sends 200 Bytes to server, gets back 1K
  - Latency-sensitive workload
- **OSA-Express6S**
  - no difference when using TSO (as expected for this kind of workload)
- **RoCE Express2 10Gb**
  - Significantly lower transaction times compared to OE6S
  - Notably higher processor consumption for the single connection case
  - Lower processor consumption for multi connection cases (50, 250 connections)

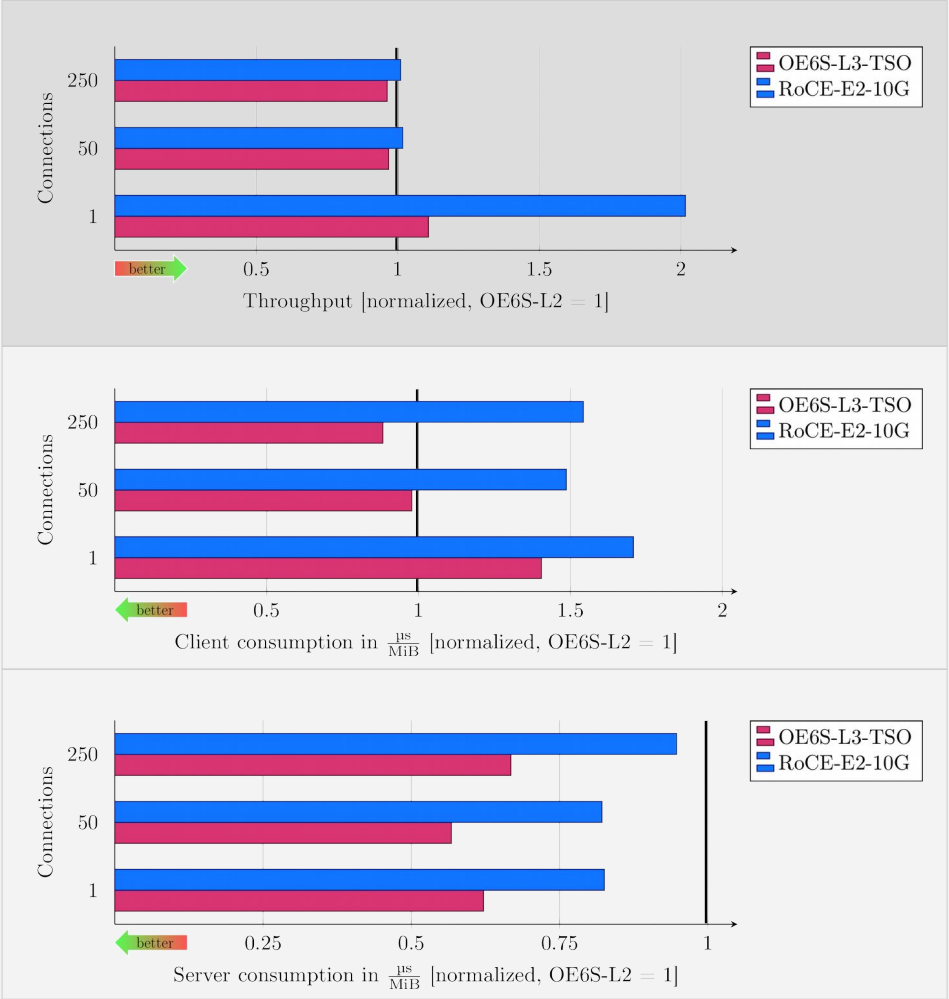






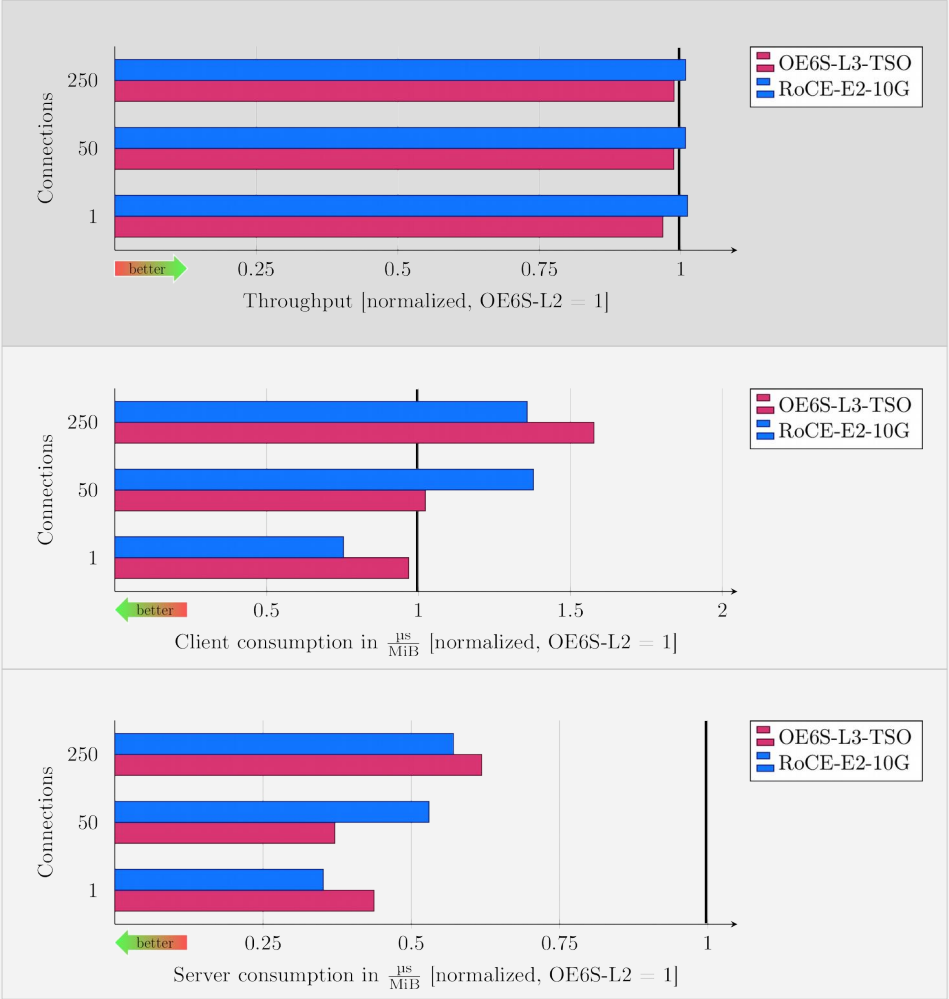
10GbE Cards

- **Note**: All measurements normalized to OE6S-L2
- **Workload**: `rr1c-200x30k`
  - Client sends 200 Bytes to server, gets back 30K
  - Typical transactional client-server workload
- **OSA-Express6S**
  - reaching line speed for 50 and 250 connections case
  - Single connection case slightly benefits from TSO
  - As expected, TSO is highly beneficial on server side
- **RoCE Express2 10Gb**
  - reaches line speed for 50 and 250 connections case
  - outperforms OE6S for single connection case
  - As expected, higher processor consumption compared to OE6S(-L3-TSO)



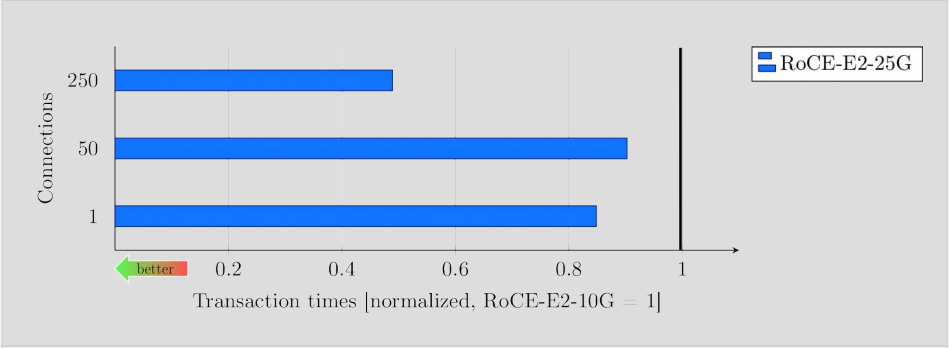
10GbE Cards

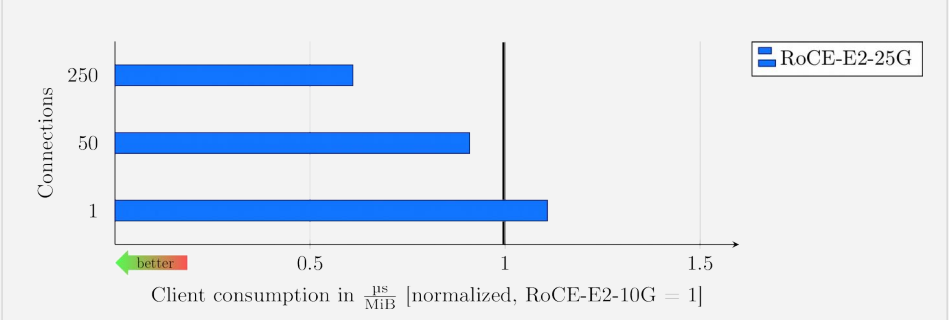
- **Note**: All measurements normalized to OE6S-L2
- **Workload**: `str-readx30k`
  - Server continuously sends 30K blocks of data
  - Typical streaming workload
- **OSA-Express6S**
  - reaches line speed for all three workloads (1, 50, 250 connections)
  - significant processor consumption savings when using TSO (+ checksum offload)
- **RoCE Express2 10Gb**
  - reaches line speed for all three workloads (1, 50, 250 connections)
  - significant processor consumption savings when compared to OE6S-L2 due to HW offloads (TSO, checksum)

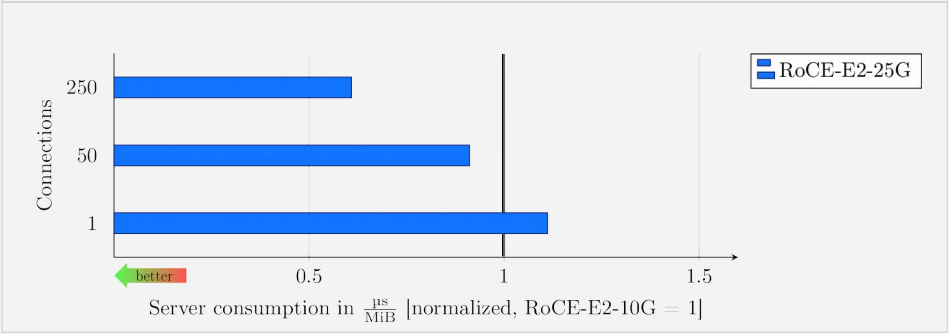


10 vs 25GbE RoCE Express2

- **Note**: All measurements normalized to RoCE-E2-10G
- **Workload**: `rr1c-200x1000`
  - Client sends 200 Bytes to server, gets back 1K
  - Latency-sensitive workload
- **RoCE Express2 25Gb**
  - Significantly lower transaction times and processor consumption for 250 parallel connections
  - Slight improvements for 50 parallel connections (~10% faster and lower processor consumption)
  - Mixed picture for single connection case

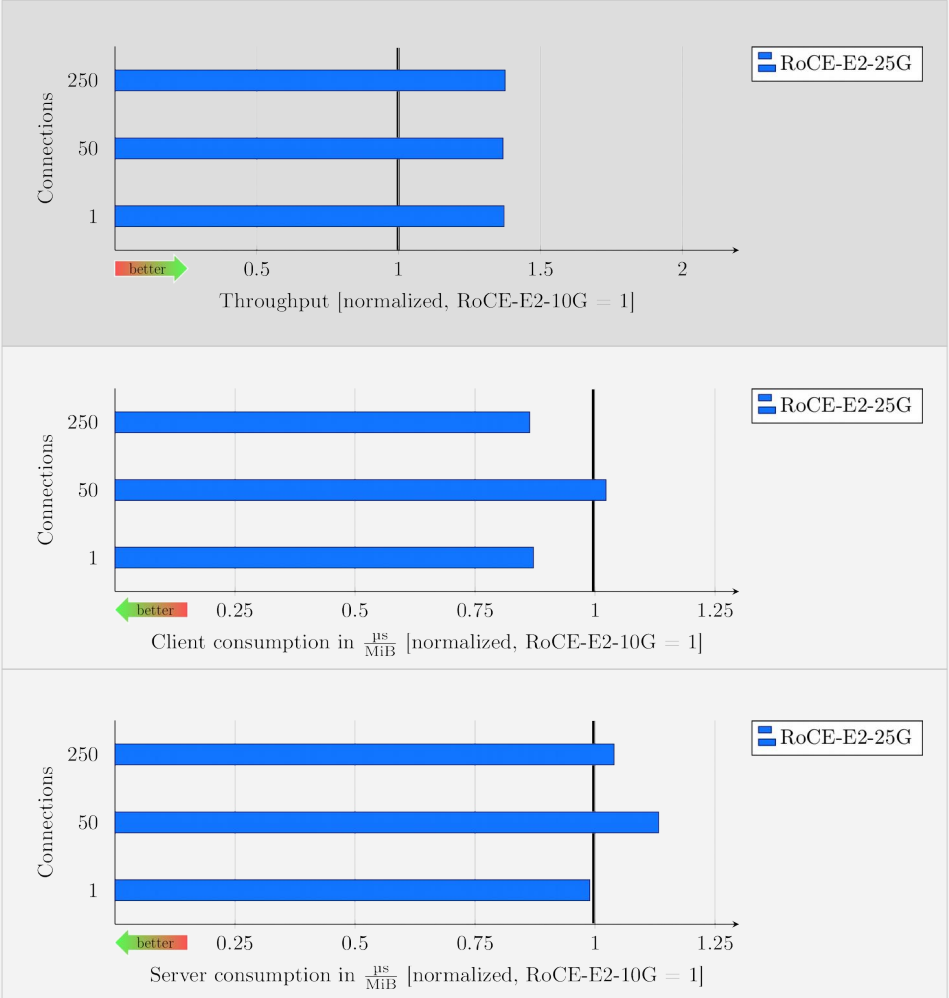






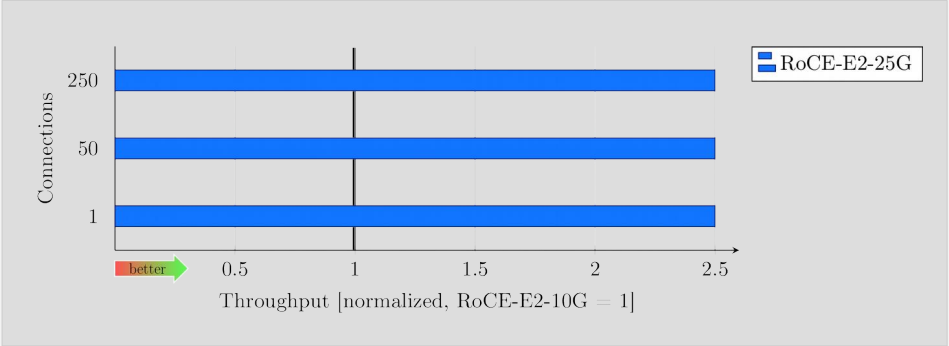
10 vs 25GbE RoCE Express2

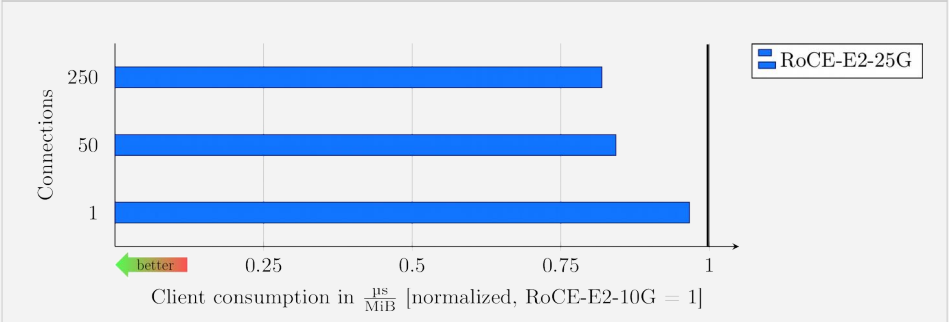
- **Note**: All measurements normalized to RoCE-E2-10G
- **Workload**: `rr1c-200x30k`
  - Client sends 200 Bytes to server, gets back 30K
  - Typical transactional client-server workload
- **RoCE Express2 25Gb**
  - significant (>35%) increase in throughput for all cases
  - Mixed results for consumption (keeping the level of the 10Gb version)

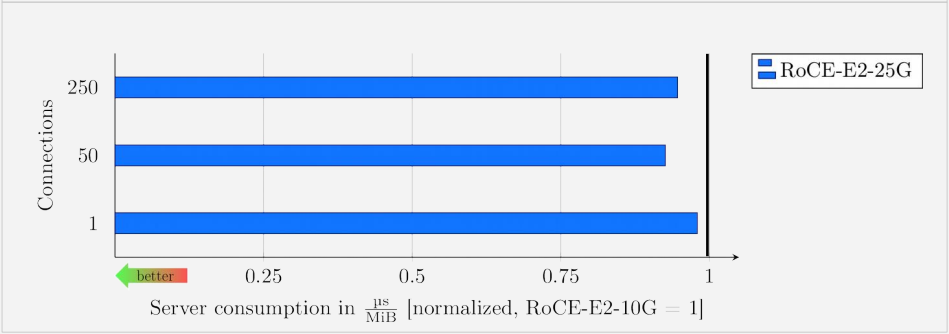


10 vs 25GbE RoCE Express2

- **Note**: All measurements normalized to RoCE-E2-10G
- **Workload**: `str-readx30k`
  - Server continuously sends 30K blocks of data
  - Typical streaming workload
- **RoCE Express2 25Gb**
  - reaching line speed for all test cases
  - thus, factor 2.5 higher throughput than 10Gb version (also reaching line speed for all cases)
  - slight improvements in processor consumption, but almost on same level (as expected)







OSA-Express

- Vast virtualization capabilities
- Supported by all IBM Z operating systems
- Supported by z/VM VSWITCH
- Excellent RAS capabilities
- Pseudo-promiscuous mode available
- SMC-R requires add’l RoCE Express card
- Performance: Scalable & economic CPU usage

RoCE Express

- 2 Ports on all models
- Limited virtualization
- Shared network traffic:
  - Excellent performance
  - Non-RDMA traffic for Linux-to-Linux only
- z/VM VSWITCH not supported
- Run SMC-R with a single device
- Performance:
  - Low latency
  - Mixed CPU consumption

References

- ***smc-tools* Homepage**
  https://www.ibm.com/developerworks/linux/linux390/smc-tools.html
- **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:**
  https://linux-on-z.blogspot.com/p/smc-for-linux-on-ibm-z.html
- **Blogs**
  - **Linux On Z Distributions News**
    https://linuxmain.blogspot.com/
  - **Linux On Z Latest Development News**
    https://linux-on-z.blogspot.com/
  - **Containers on Z, primarily *Docker***
    https://containersonibmz.com/