z/VM and Linux on IBM Z Networking

Malcolm Beattie

IBM

November 2018

Session CG




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

Db2*	FlashCopy*	IBM (logo)*	OMEGAMON*	z13*	z/Architecture*	zSeries*
DirMaint	FlashSystem	IBM Z*	PR/SM	z13s	zEnterprise*	z/VM*
DS8000*	GDPS*	LinuxONE*	RACF*	z14	z/OS*	z Systems*
ECKD	ibm.com	LinuxONE Emperor	System z10*	z10 BC	zSecure	
FICON*	IBM eServer	LinuxONE Rockhopper	XIV*	z10EC		

* Registered trademarks of IBM Corporation

Adobe, the Adobe logo, PostScript, and the PostScript logo are either registered trademarks or trademarks of Adobe Systems Incorporated in the United States, and/or other countries.
IT Infrastructure Library is a Registered Trade Mark of AXELOS Limited.
ITIL is a Registered Trade Mark of AXELOS Limited.
Linear Tape-Open, LTO, the LTO Logo, Ultrium, and the Ultrium logo are trademarks of HP, IBM Corp. and Quantum in the U.S. and other countries.
Intel, Intel logo, Intel Inside, Intel Inside logo, Intel Centrino, Intel Centrino logo, Celeron, Intel Xeon, Intel SpeedStep, Itanium, and Pentium are trademarks or registered trademarks of Intel Corporation or its subsidiaries in the United States and other countries.
Linux is a registered trademark of Linus Torvalds in the United States, other countries, or both.
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.
Java and all Java-based trademarks and logos are trademarks or registered trademarks of Oracle and/or its affiliates.
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.
UNIX is a registered trademark of The Open Group in the United States and other countries.
VMware, the VMware logo, VMware Cloud Foundation, VMware Cloud Foundation Service, VMware vCenter Server, and VMware vSphere are registered trademarks or trademarks of VMware, Inc. or its subsidiaries in the United States and/or other jurisdictions.
Other product and service names might be trademarks of IBM or other companies.

Notes:

Performance is in Internal Throughput Rate (ITR) ratio based on measurements and projections using standard IBM benchmarks in a controlled environment. The actual throughput that any user will experience will vary depending upon considerations such as the amount of multiprogramming in the user's job stream, the I/O configuration, the storage configuration, and the workload processed. Therefore, no assurance can be given that an individual user will achieve throughput improvements equivalent to the performance ratios stated here.
IBM hardware products are manufactured from new parts, or new and serviceable used parts. Regardless, our warranty terms apply.
All customer examples cited or described in this presentation are presented as illustrations of the manner in which some customers have used IBM products and the results they may have achieved. Actual environmental costs and performance characteristics will vary depending on individual customer configurations and conditions.
This publication was produced in the United States. IBM may not offer the products, services or features discussed in this document in other countries, and the information may be subject to change without notice. Consult your local IBM business contact for information on the product or services available in your area.
All statements regarding IBM's future direction and intent are subject to change or withdrawal without notice, and represent goals and objectives only.
Information about non-IBM products is obtained from the manufacturers of those products or their published announcements. IBM has not tested those products and cannot confirm the performance, compatibility, or any other claims related to non-IBM products. Questions on the capabilities of non-IBM products should be addressed to the suppliers of those products.
Prices subject to change without notice. Contact your IBM representative or Business Partner for the most current pricing in your geography.
This information provides only general descriptions of the types and portions of workloads that are eligible for execution on Specialty Engines (e.g, zIIPs, zAAPs, and IFLs) ("SEs"). IBM authorizes customers to use IBM SE only to execute the processing of Eligible Workloads of specific Programs expressly authorized by IBM as specified in the “Authorized Use Table for IBM Machines” provided at www.ibm.com/systems/support/machine_warranties/machine_code/aut.html (“AUT”). No other workload processing is authorized for execution on an SE. IBM offers SE at a lower price than General Processors/Central Processors because customers are authorized to use SEs only to process certain types and/or amounts of workloads as specified by IBM in the AUT.

Agenda

- Basic VSWITCH configuration
- VLANs, Access Ports and Trunk Ports
- VSWITCH port numbers
  - Port-based v. User-based VSWITCHes and their unification
  - Port attributes and authorisation: GRANT, Directory Network Authorisation (DNA) and ESMs
- Link Aggregation Groups (LAGs) and LACP
- Global VSWITCH: sharing a LAG (port group) between LPARs
- OSA-Express7S 25GbE SR
- RoCE-Express and its use for Linux
- RoCE-Express2 25 GbE

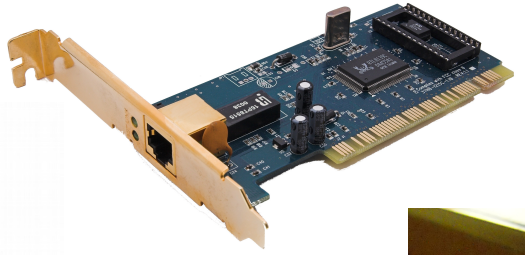
Basic VSWITCH configuration

Recipe for basic ethernet networking

- Take one basic ethernet switch

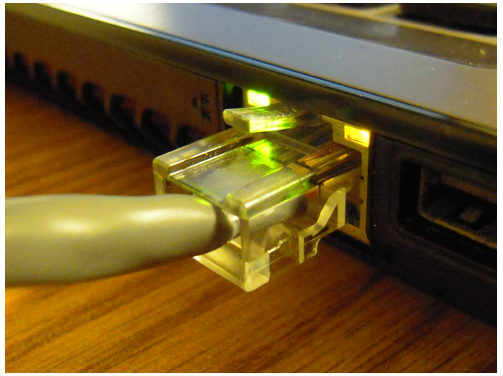
- Take one basic server...

- ...that contains a basic ethernet Network Interface Card (NIC)



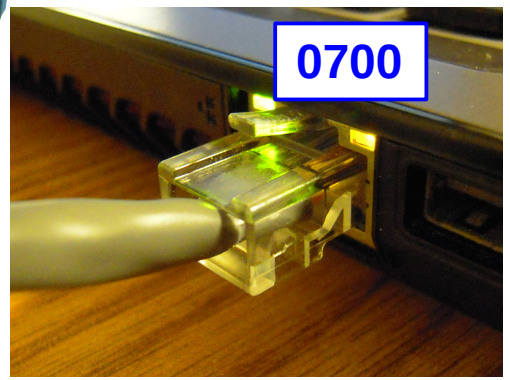
- Take a basic ethernet cable...

- and use it to connect the basic NIC to the basic switch



Basic VSWITCH Configuration

- A basic VSWITCH acts as a basic ethernet switch
  - DEFINE VSWITCH *MYNET* ETHERNET
  - As a CP command or in SYSTEM CONFIG
- A basic virtual Network Interface Card (vNIC) is given to a guest and acts as a basic ethernet card that can be plugged into that basic switch
  - with a user directory entry: NICDEF 700 TYPE QDIO
  - or dynamically from guest with CP DEFINE NIC 700 TYPE QDIO
- Plug in a vNIC connection to a VSWITCH (“coupling”)
  - at guest startup time with the user directory entry: NICDEF 700 TYPE QDIO **LAN SYSTEM *MYNET***
  - or dynamically from the guest: CP COUPLE 700 SYSTEM *MYNET*
    - can omit SYSTEM *MYNET* if CP can deduce it from the directory
- Unplug vNIC dynamically from the guest, if needed, with CP UNCOUPLE 700



Basic VSWITCH Configuration

- So far, guests with vNICs coupled to *MYNET* can talk to each other through the switch but not to anything outside

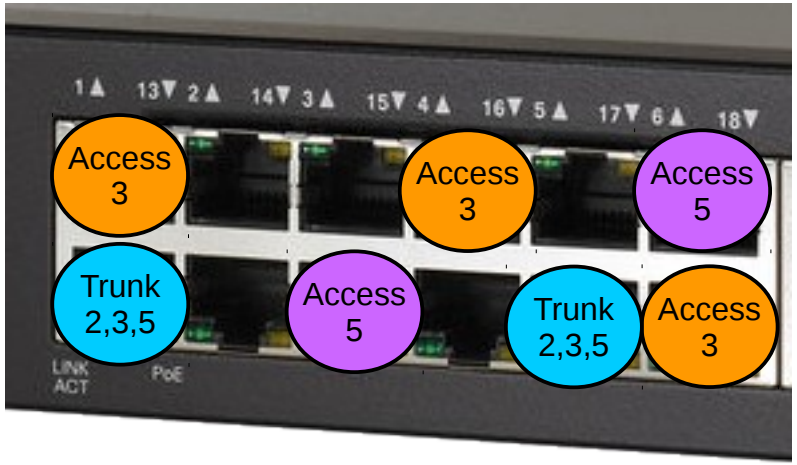
- Like higher-end physical ethernet switches, the VSWITCH can be configured with a choice of connectivity to upstream switches via “uplinks”
- To reach outside the VSWITCH, we need to assign one (or more) real physical OSA-Express ports as uplinks:
  - DEFINE VSWITCH ... **UPLINK RDEV E000.P1**

RDEVs E000-2 port 1

- Now the VSWITCH sends/receives ethernet traffic between the guests' vNICs and the physical NIC(s)
  - to/from other LPARs sharing the OSA CHPID
  - and to/from the upstream switch to which the OSA ports are connected

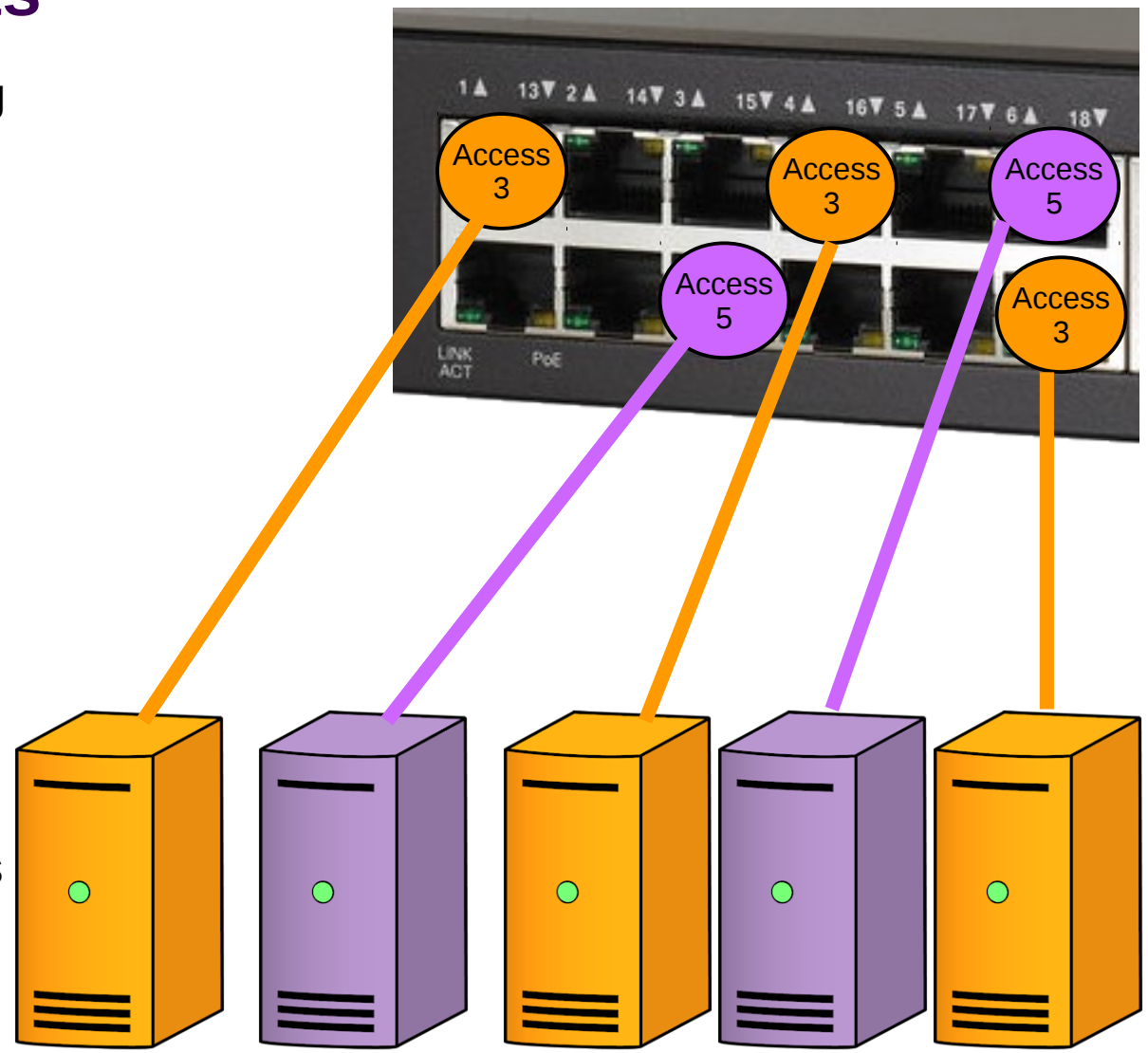
VLANs

- A “VLAN aware” ethernet switch can also deal with ethernet frames that have a VLAN (IEEE 802.1Q) tag
  - a VLAN tag is a number between 1 and 4094
  - These can be used to make a collection of switches implement a logically isolated ethernet network segment for each VLAN number, regardless of physical connections
- For such a switch, each port can be configured to be either
  - an ***Access*** port permitted to one specific VLAN number; or
  - a ***Trunk*** port permitted to a list of VLAN numbers
    - typically more than one VLAN number, often all VLAN numbers
- Once “inside” the switch, all frames can be considered to have a tag
- When the switch considers frames for forwarding from one port to another (e.g. looking at MAC addresses), the different VLAN numbers are treated as isolated, entirely independent networks



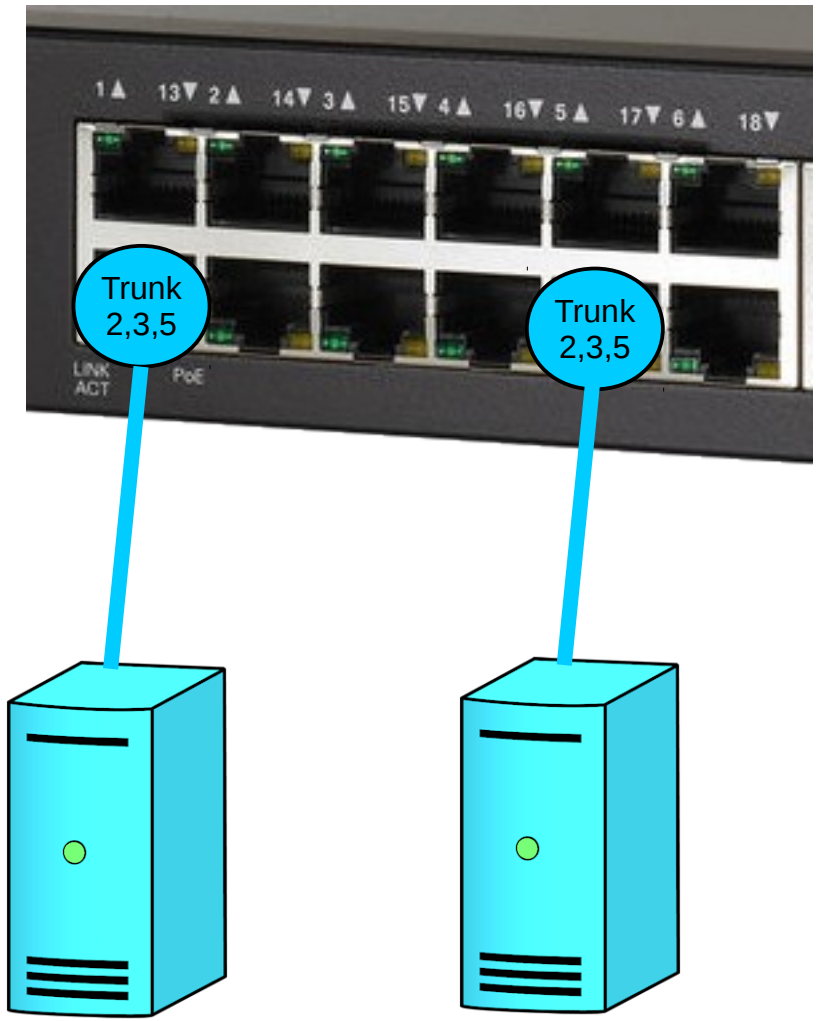
VLAN aware switch: Access Ports

- For a NIC connected to an Access port, the operating system sends and receives ordinary untagged ethernet frames
- When an untagged frame enters the switch through an Access port, the switch tags the frame with the single VLAN number of that Access port
  - in the (uncommon) case that the operating system sends a tagged frame, it is allowed through if and only if the tag matches the tag of the Access port
- When the switch inspects the frame for forwarding, it only considers sending it out of an Access port if the tag of the frame matches the tag of the Access port
- When the switch does send a frame out of an Access port, it strips the tag out, leaving an ordinary untagged ethernet frame to deliver to the NIC



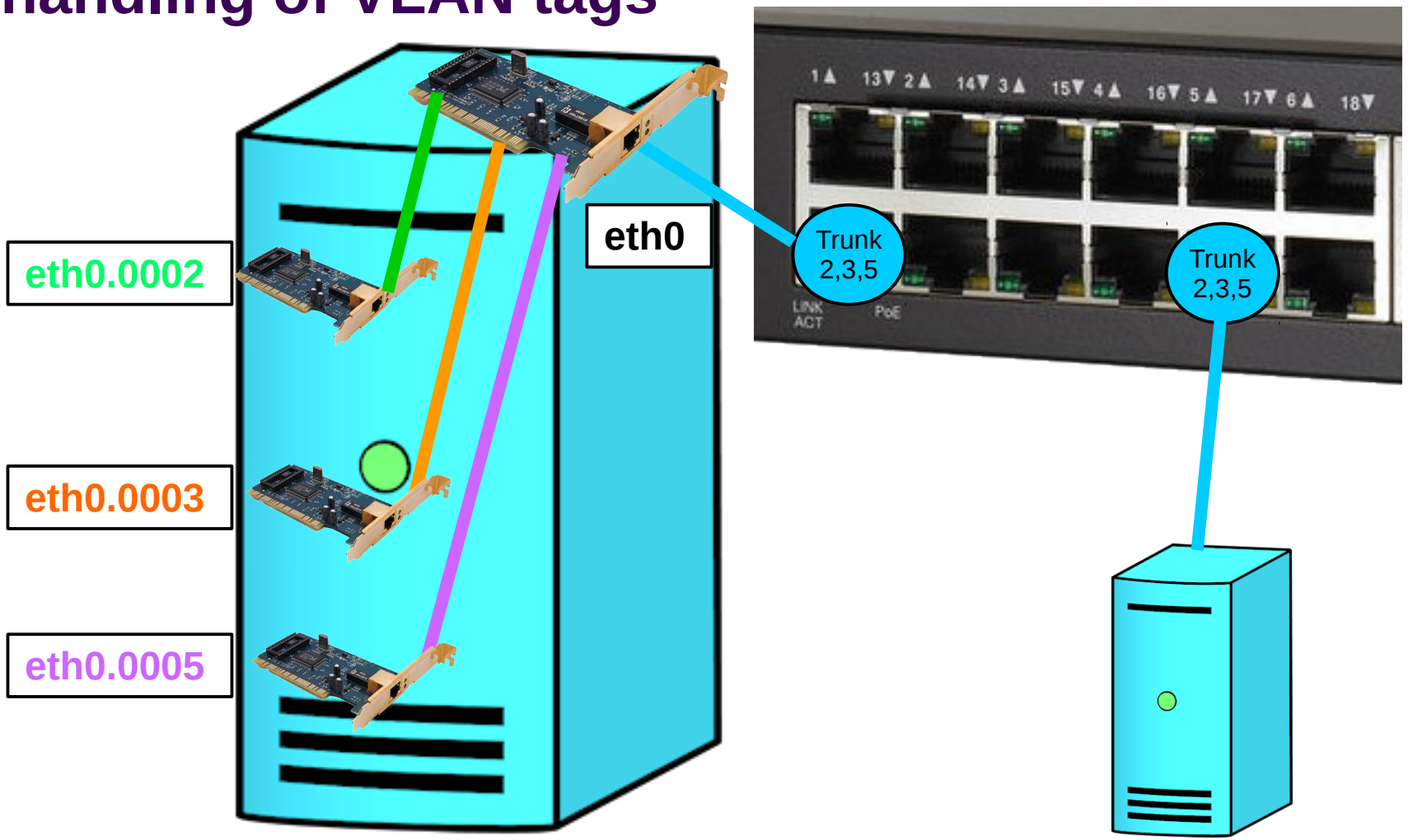
VLAN aware switch: Trunk Ports

- For a NIC connected to an Trunk port, the operating system (usually*) sends and receives tagged ethernet frames
- When a tagged frame enters the switch through a Trunk port, the switch allows it in if and only if the tag of the frame is on the list of allowed tags for that Trunk port
- When the switch inspects the frame for forwarding, it only considers sending it out of a Trunk port if the tag of the frame is on the list of allowed tags for that Trunk port
- When the switch does send a frame out of a Trunk port, it leaves the tag on the frame*
- * Corner case: A VLAN-aware switch has a “Native VLAN number” (nearly always = 1)
  - Inbound untagged frames get tagged with the Native VLAN number.
  - Outbound frames whose tag is the Native VLAN number get stripped of the tag and sent out untagged.
  - This allows the switch to treat all frames internally as tagged even though outside, some are tagged and some are untagged



Operating System handling of VLAN tags

- Operating systems that support VLAN tagging usually do so by allowing the creation of a VLAN network interface for any given VLAN number
  - Frames sent out of that interface are tagged by the O/S with that VLAN number
  - Tagged frames received by the physical NIC are delivered by the O/S as though they arrived via the VLAN interface with that number



z/VM VLAN aware VSWITCH Creation

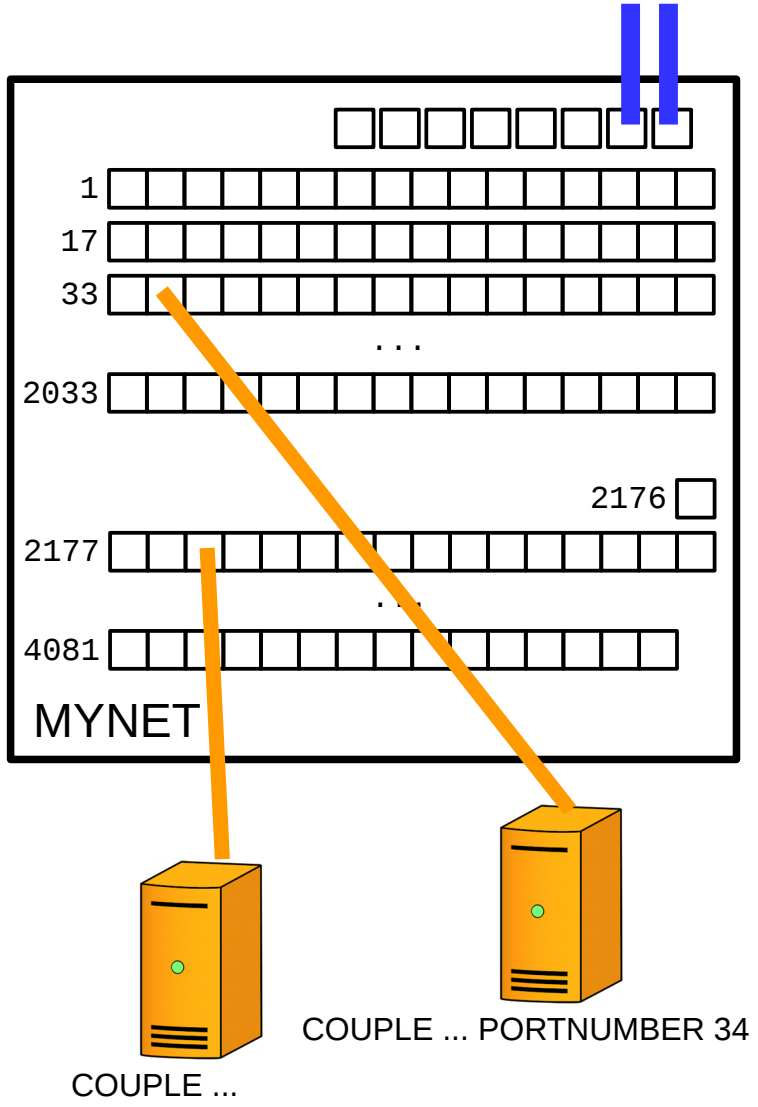
- Create a VLAN-aware VSWITCH with

  DEFINE VSWITCH *MYNET* ETHERNET ... **VLAN AWARE NATIVE NONE**

- Best practices
  - Use VLAN AWARE (rather than VLAN *defvid*) so that a guest that has not been given access will get errors
  - Use NATIVE NONE so that there is no chance of untagged frames escaping from z/VM

z/VM VSWITCH port numbers

- (We will assume from now an up-to-date z/VM so that it supports DNA and the Network Security Enhancements that allow the unification of behaviour of USERBASED and PORTBASED VSWITCHes)
- A z/VM VSWITCH models a physical ethernet switch with about 4000 ports
- When a port number is specified or implied at COUPLE time, the vNIC is plugged into that port number (must be 1-2048)
  - Port number can be specified via COUPLE ... PORTNUMBER *portnum*
  - or implied/enforced by a clause in the user directory entry:
    NICDEF vdev TYPE QDIO LAN SYSTEM *MYNET* **PORTNUMBER *portnum***
- When no port number is specified or implied at COUPLE time, CP chooses a port number in the range 2176-4095



z/VM VSWITCH port attributes and authorisation

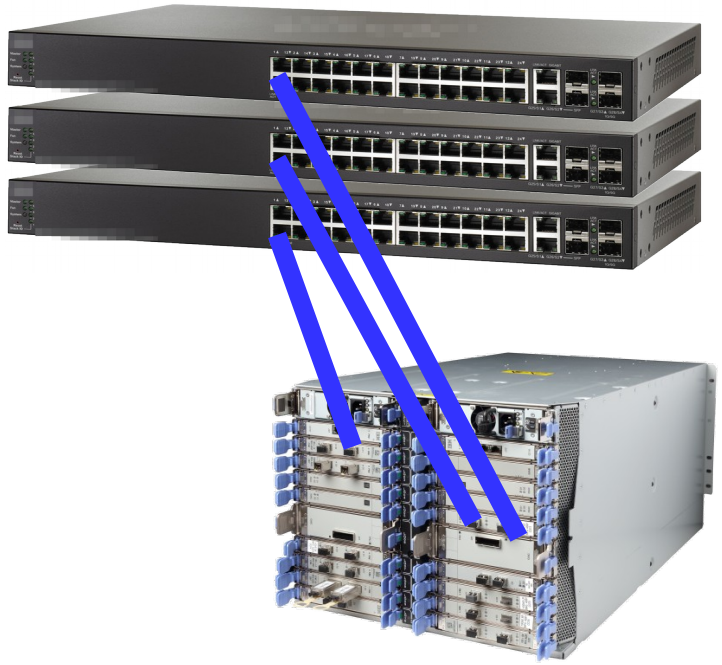
- The attributes of a port include
  - userid of the guest that can couple a vNIC to this port
  - whether the port type is Access or Trunk
  - what the permitted VLAN number(s) is/are
  - whether the port will allow promiscuous sniffing of frames on the VSWITCH (default is no, of course)
- When a vNIC couples a vNIC by port number (1-2048) the attributes of the port number must either
  - have been specified by a command CP SET VSWITCH *MYNET* PORTNUMBER *portnum* **attributes...**
  - be available in the user directory NICDEF statement for that vNIC
- When a vNIC couples with no specified or implicit port number (i.e. the VSWITCH it behaving as a USERBASED VSWITCH), the attributes of the port number must either
  - have been specified by a command CP SET VSWITCH *MYNET* GRANT *userid* **attributes...**
  - be available in the user directory NICDEF statement for that vNIC

z/VM VSWITCH authorisation with an External Security Manager

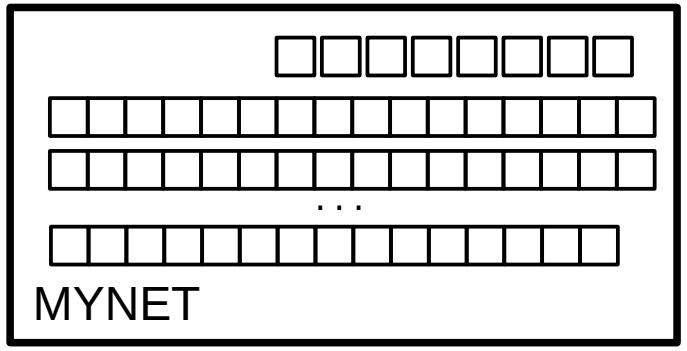
- When an ESM is used to manage a VSWITCH
  - the ESM is the ultimate authority on granting access and authorising vNIC characteristics
    - provided it supports DNA, it can find sufficient information for this in the user directory
- Best practices:
  - Use ESM and groups to manage VLAN assignments
    - Simplifies VLAN changes
    - Overrides VLAN specification on NICDEF
    - CP will use NICDEF if ESM defers

VSWITCH Uplink High Availability: active/passive failover

- When defining a VSWITCH you can specify up to three uplink ports
  - CP DEFINE VSWITCH *MYNET* ... **UPLINK RDEV *rdev*$_1$.P*n*$_1$ *rdev*$_2$.P*n*$_2$ *rdev*$_3$.P*n*$_3$**
- As usual for OSA-Express device numbers
  - each *rdev* specifies the first of a triple of consecutive real device numbers
  - each port P*n* specifies which of the physical ports on the OSA-Express card is used
    - P0 or P1 for current generations of OSA-Express
    - Note that this meaning of “port number” is completely different from that we have been using elsewhere in this presentation to number ports of an ethernet switch, whether a physical one or a VSWITCH
- All these physical uplinks must be connected to the same ethernet network
  - the same broadcastable “Layer 2” and
  - the target physical switch ports configured as Trunk ports if the VSWITCH is VLAN AWARE
- CP only uses one of these uplink ports at a time
  - if one fails (or its cable or the target switch or target port) then CP uses one of the others
  - CP SET VSWITCH *MYNET* SWITCHOVER ... can be used to trigger a failover/failback



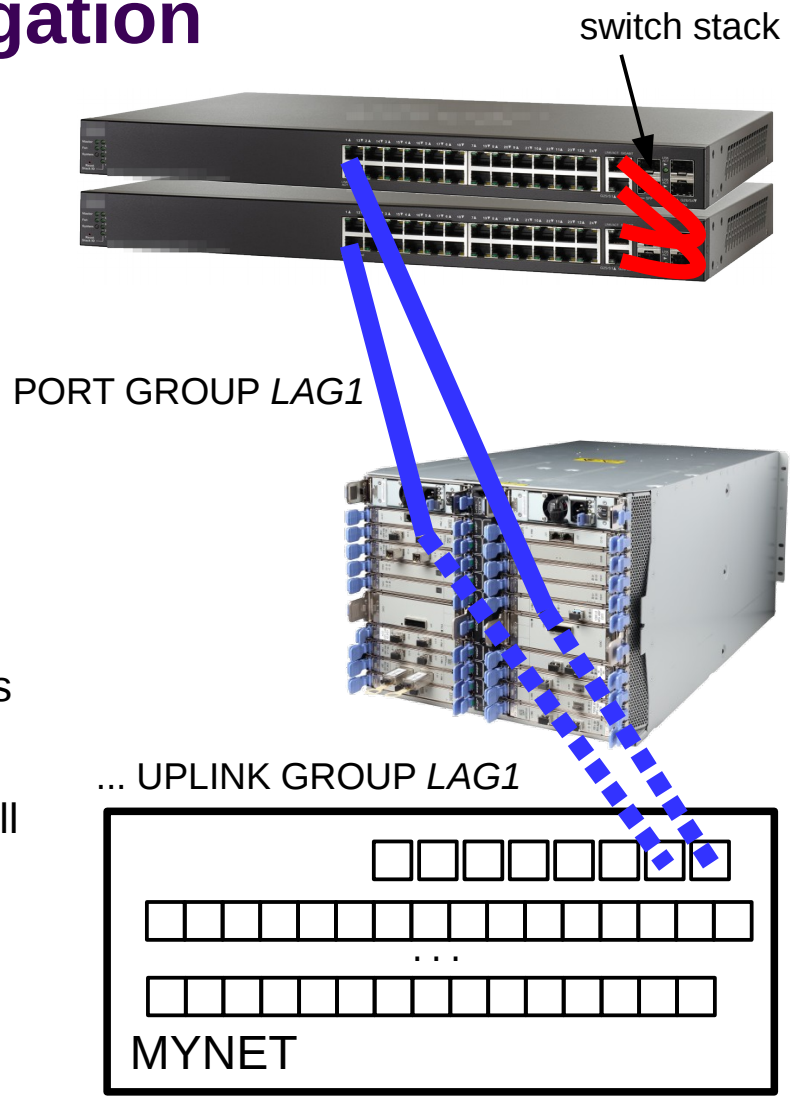
UPLINK RDEV E000.P1 E000.P0 E200.P0



Ethernet Switch High Availability: Link Aggregation

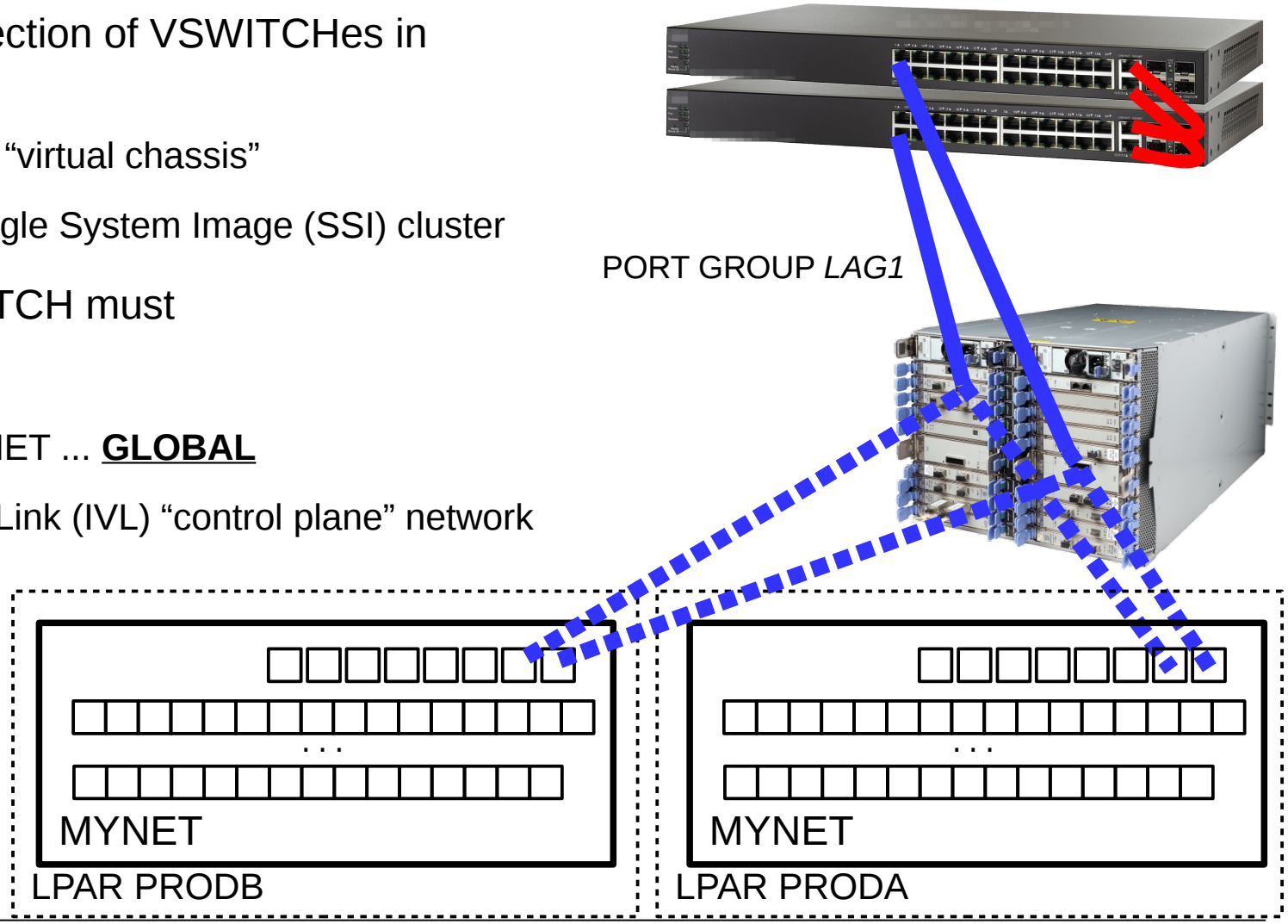
- Link Aggregation (LAG) as implemented by the Link Aggregation Control Protocol (LACP, IEEE 802.3ad) is a way to combine multiple uplink ports into a *port group* that
  - supports all uplink connections to be used concurrently with load balancing between them
  - handles failure of uplink connections transparently (except for the last of the group, of course)
- It requires the switches at both ends of the connections to support LACP
  - z/VM VSWITCH supports this, as do non-low-end physical ethernet switches
  - the uplink OSA ports of a port group must be plugged into the same switch (or cluster/stack of switches) so that the switch LACP software can control all the ports in the port group
- OSA ports used for Link Aggregation are **dedicated** to that port group and can **not** be shared in the traditional way with other LPARS

SET PORT GROUP *LAG1* JOIN E000 E200
DEFINE VSWITCH MYNET ETHERNET **UPLINK GROUP *LAG1***



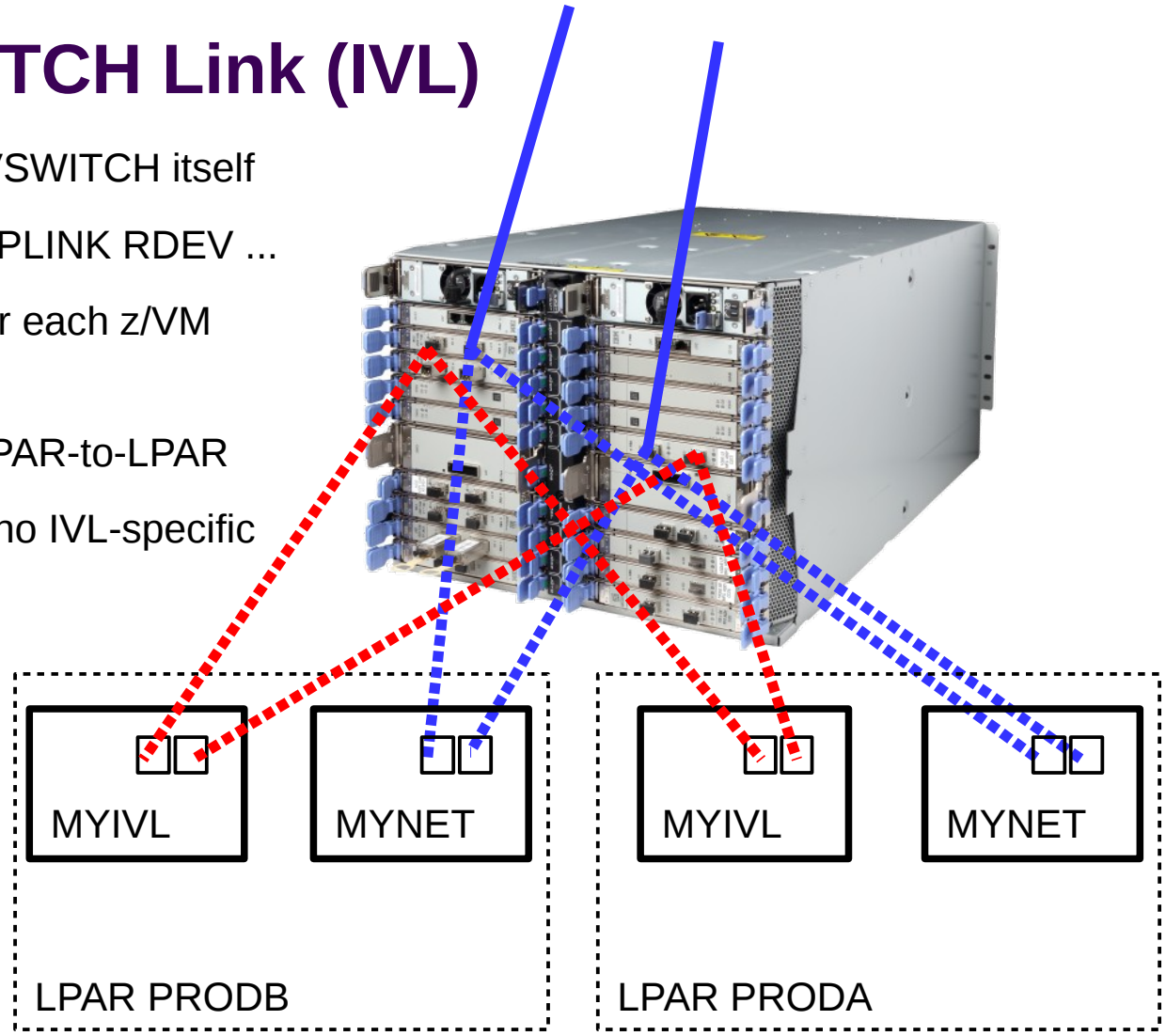
Global VSWITCH: Sharing a Port Group Between LPARs

- A Global VSWITCH is formed from a collection of VSWITCHes in **different** z/VM LPARs
  - similar to a physical ethernet “stack” or “virtual chassis”
  - may (but need not) be LPARs in an Single System Image (SSI) cluster
- The VSWITCHes forming a Global VSWITCH must
  - share the same name and type
  - be defined with DEFINE SWITCH MYNET ... **GLOBAL**
  - communicate over an Inter-VSWITCH-Link (IVL) “control plane” network
- The VSWITCHes can then share a port group



Global VSWITCH: The Inter-VSWITCH Link (IVL)

- The Inter-VSWITCH Link (IVL) “control-plane” network is a VSWITCH itself
  - DEFINE VSWITCH MYIVL ETHERNET **TYPE IVL** UPLINK RDEV ...
  - Only one is needed (and only one can be defined) for each z/VM system
- The IVL's uplink devices are only needed to communicate LPAR-to-LPAR
  - “bouncing off” the inside of the OSA can be used so no IVL-specific separate network is needed
  - Multiple uplink ports on separate OSA cards is recommended for High Availability
  - The communication uses multicast and there is a way to have multiple isolated IVL “domains”
    - e.g. for separating test/prod Global VSWITCHes
    - TYPE IVL DOMAIN ***d*** (where ***d*** is A, B, ..., H)



OSA-Express7S 25GbE SR - FC 0429

- The recent GA2 announcement introduces a new generation of OSA
  OSA-Express7S 25GbE
- Introduces updates to OSA Express hardware
- Provides one 25GbE physical port (one 25GbE port per feature)
- Initially offered as 25GbE only
- Requires 25GbE optics and Ethernet switch 25GbE support
- Short Reach (SR) only
- OSD CHPID Type
- **Auto negotiate not supported**
- New NIC module type
- Up to 48 features on z14 M0x/ZR1 or LinuxONE Emperor II / Rockhopper II
- Possible OSA consolidation (10 Gb to 25 Gb)
- Software requirements: For z/VM 6.4 APAR PI99085; for Linux on Z, update required for all supported distros

RoCE-Express

- Linux can exploit RoCE Express and RoCE Express2 as a standard NIC for Ethernet
- Note that these cards have **two** ports whereas OSA Express cards only have **one** port
- However they are driven differently by operating systems compared to OSA Express cards
  - They do not participate in any VSWITCH virtualisation
  - They do not use traditional CHPIDs and device numbers for their configuration
- A specific Linux distribution level is required (reference PSP bucket for additional details)
  - SLES 11 SP4, SLES 12 SP2
  - RHEL 6.8, RHEL 7.3
  - Ubuntu 16.04 (+ additional patches)
  - **Note:** Linux does not currently support SMC-R

RoCE-Express

- For RoCE-Express technology, the I/O Configuration (IOCDS) needs to define
  - A real PCIFUNCTION id (*rpfid*, a.k.a. real FID) for the desired number of Virtual Functions (VFs) on the ports of a RoCE-Express (ROCE) or RoCE-Express2 (ROC2) PCHID...
  - ...and allocate each to an LPAR...
    - ...or more than one but the FID is only **reconfigurable** not **shared** so only one LPAR can bring it online at once
    - RoCE Express2 supports 63 VFs per physical port; original RoCE Express supports 31 VFs per port
- Each *rpfid* (corresponding to a VF) can then be handed out to an operating system images to use the port
  - For a driving operating system in an LPAR (z/OS or Linux), only one VF is necessary...
    - Note that z/VM cannot drive a RoCE-Express NIC for its own use - it can only hand out VFs to virtual machines
  - ...but for virtual machines to be able to drive a RoCE-Express NIC, one VF is needed for each virtual machine
- z/VM gives a VF to a virtual machine with:
  ATTACH PCIFUNCTION *rpfid* TO *userid* [AS *vpfid*]

RoCE-Express

- There are a number of virtualisation constraints and these become the responsibility of the systems programmer
  - Limitations on numbers of virtual machines (one VF each) and some other IOCDS numeric constraints
  - Management of *rpfid* numbers to be allocated to virtual machines
  - Memory management requires planning
    - pinned memory used by operating systems for driving PCIFUNCTIONs is not pageable when in use
- In a z/VM environment some benefits of functionality and auto-tuning are no longer available
  - No Live Guest Relocation permitted for guests using PCIFUNCIONs
  - For memory management of the pinned memory needed by PCIFUNCTIONs, configuration in SYSTEM CONFIG via STORAGE IOAT ... needs planning, monitoring and adjusting as needed
  - There is no corresponding virtualisation to VSWITCH for switching, measuring, securing or isolating networks

25GbE RoCE-Express2 - FC 0430

- The recent GA2 announcement introduces 25 GbE RoCE Express2
  - Based on the existing RoCE Express2 generation hardware
  - Provides two 25GbE physical ports
  - Requires 25GbE optics and Ethernet switch 25GbE support
- The same base card as the 10 GbE RoCE Express2 (z14 GA1)
- Better performance and throughput than 10GbE RoCE Express2
  - The actual bandwidth achieved is based on your application workloads and the characteristics of the specific application itself
- No software updates or migration actions

IBM 25GbE RoCE Express2

10 GbE RoCE-Express2 and 25GbE RoCE-Express2

Machine	Max Features	Ports per Feature	Max VFs per Feature
z14 M01-M05 and LinuxONE Emperor II	8	2	2 x 63 = 126
z14 ZR1 and LinuxONE Rockhopper II	4	2	2 x 31 = 62

And finally...

- Thank you
- Questions?
- Contact details:

Malcolm Beattie
beattiem@uk.ibm.com

We want your feedback!

- Please submit your feedback online at ….
  - http://conferences.gse.org.uk/2018/feedback/CG

- Paper feedback forms are also available from the Chair person

- This session is CG