10Gb LAN and SR-IOV on Power

IBM

Power Systems

Open Innovation to Put Data to Work

Marie-Lorraine Bontron - Jean-Manuel Lenez

Network Technologies on POWER Systems

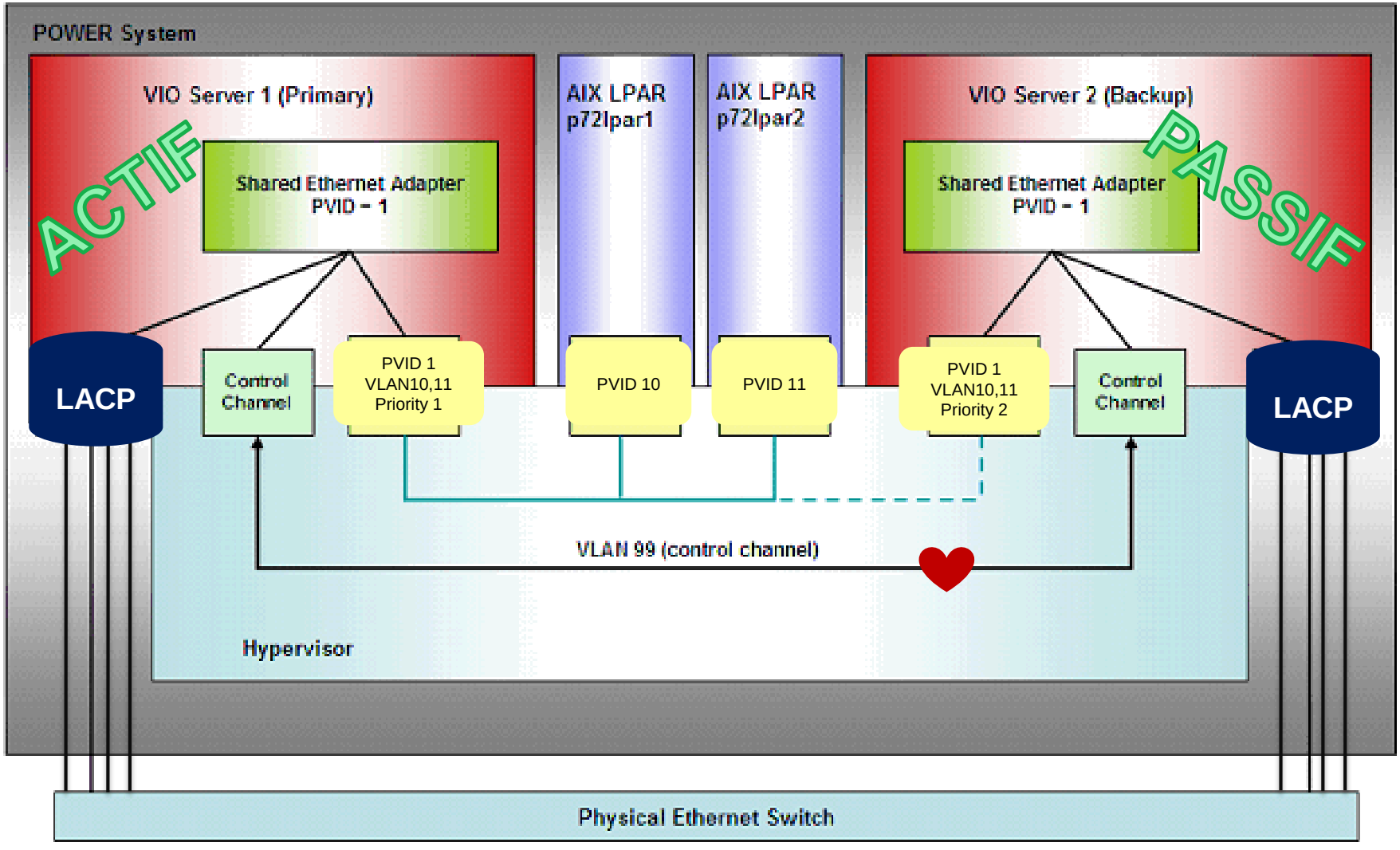
- **Dedicated Adapters**
  - Best possible performance
  - Adapter exclusively bound to particular partition; no resource sharing
- **Virtual Ethernet Adapter**
  - Hypervisor internal switching
- **VIOS Shared Ethernet Adapter**
  - Hypervisor Switch uplink through VIOS
  - Options for high availability
    - SEA failover, SEA failover w. load sharing, NIB
- **Single Root I/O Virtualization (SR-IOV) and vNIC** ← **vNIC Announced 5th October 2015**
  - SR-IOV is PCIe standard for hardware resource sharing
  - vNIC is new virtual adapter type
- **Host Ethernet Adapter (HEA)**
  - Adapter virtualization technology
  - Not available for P7+ and P8 servers

Comment obtenir le maximum des interfaces 10G sur Power

- Architecture pour répartir la charge et utiliser tous les interfaces
  - SEA load-sharing
  - NIB avec virtual switches
- Tuning des interfaces dans les VIOS et les LPARs
  - Segmentation offload et segmentation aggregation
  - Jumbo-frames

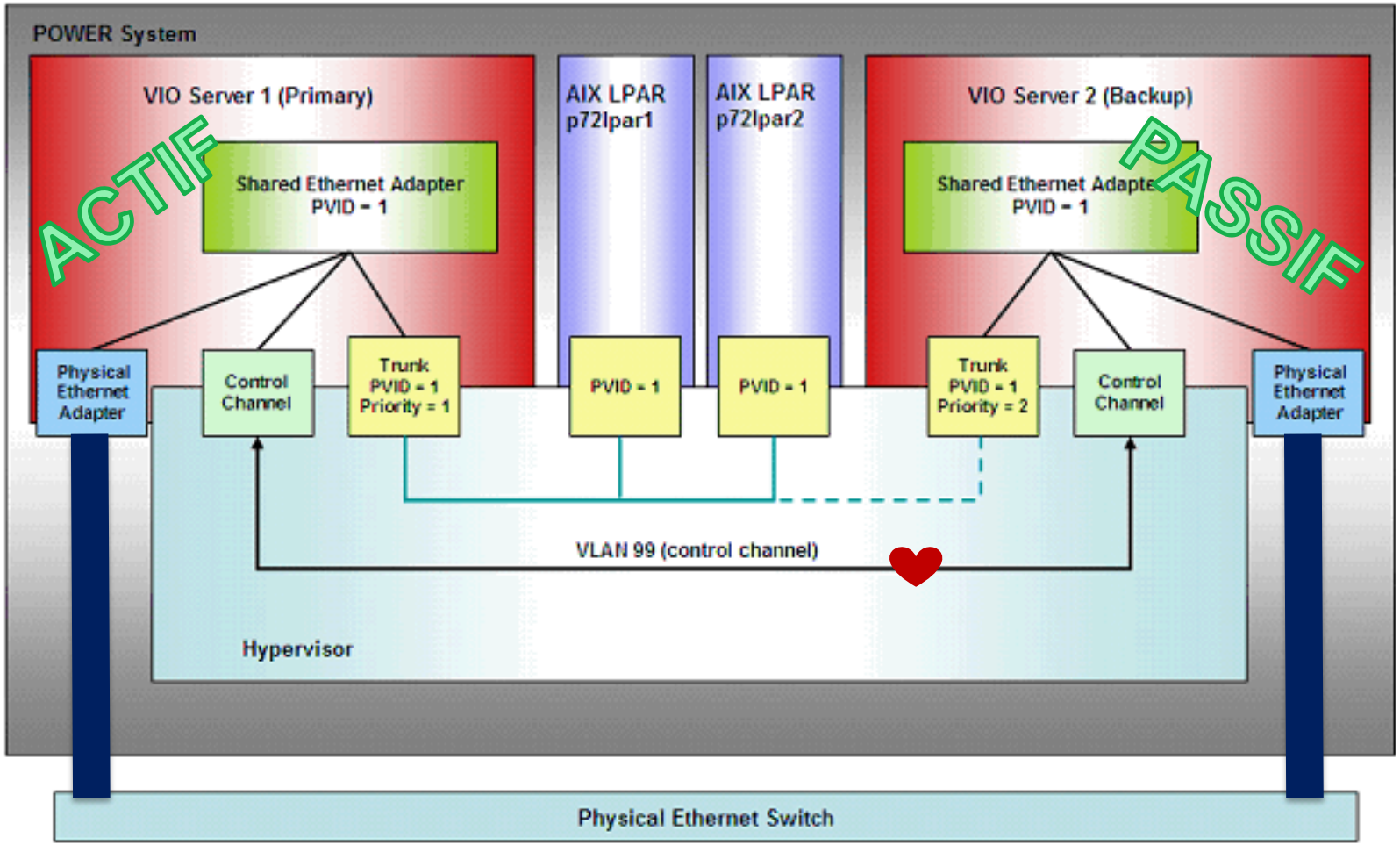
Implémentation du réseau dans les VIOS - rappel

- Architecture typique avec du 1Gb: SEA failover, aggrégation de liens pour augmenter la bande passante et support de VLANs multiples.



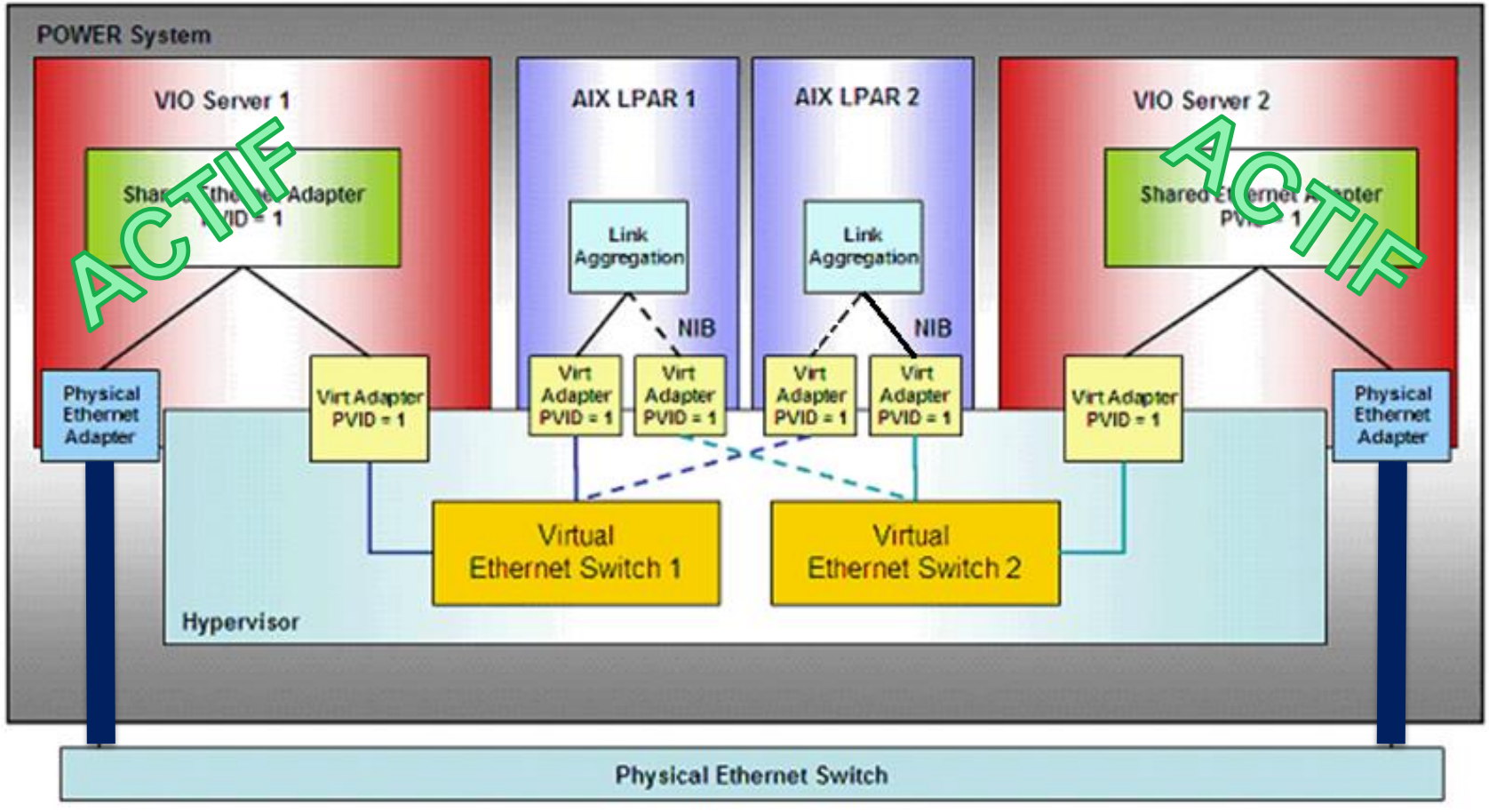
Implémentation du 10G dans les VIOS

- Avec le remplacement des cartes 1G par des cartes 10G, il n'est plus acceptable d'avoir une configuration «active-passive» avec un VIOS qui «attend»…



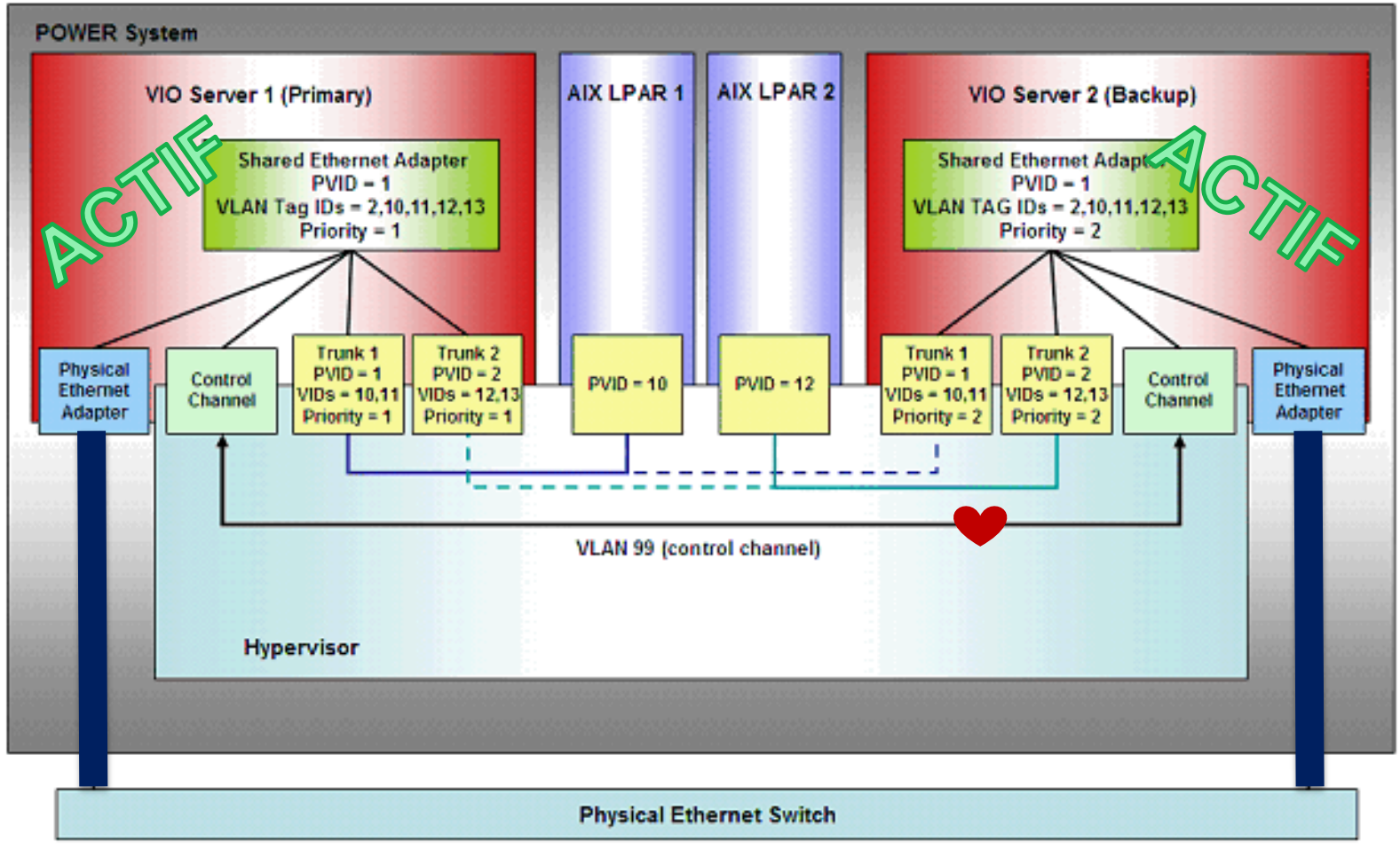
Implémentation du 10G dans les VIOS – Option 1

- Configuration avec NIB (Network Interface Backup) au niveau de la LPAR => besoin d'une @IP_to_ping pour le failover
- Répartition de la charge par LPAR
- Support d'un seul ou de multiple VLANs



Implémentation du 10G dans les VIOS – Option 2

- Configuration en SEA «Load-Sharing», basé sur les VLANs
- Implique d'avoir plusieurs VLANs
- La répartition est effectuée au niveau des VIOS



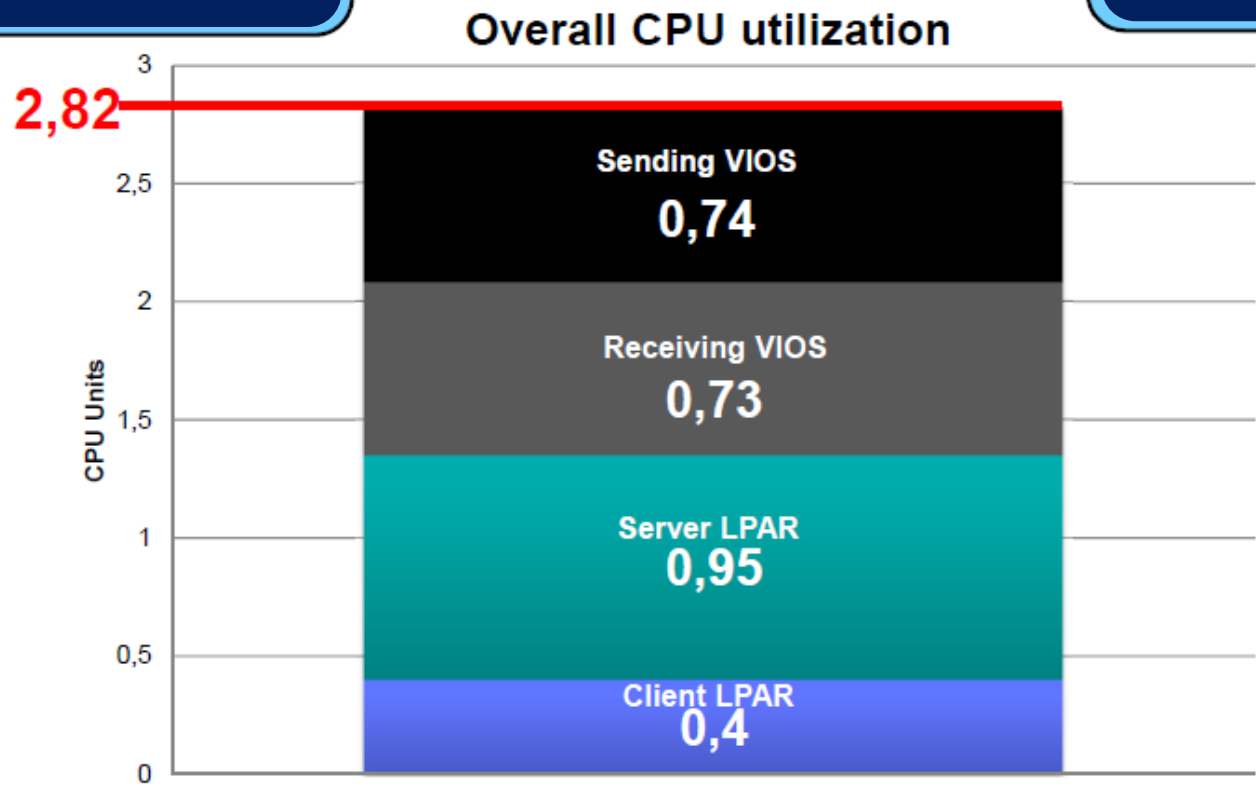
Ethernet virtuel : mesure de l'activité réseau

- L'envoi de frame par l'interface ethernet virtuel correspond à un transfert de mémoire initié par un appel à l'hyperviseur (H_SEND_LOGICAL_LAN) de l'OS du site envoyeur. Ceci implique l'utilisation de CPU...

Combien de CPU est nécessaire pour atteindre un débit donné

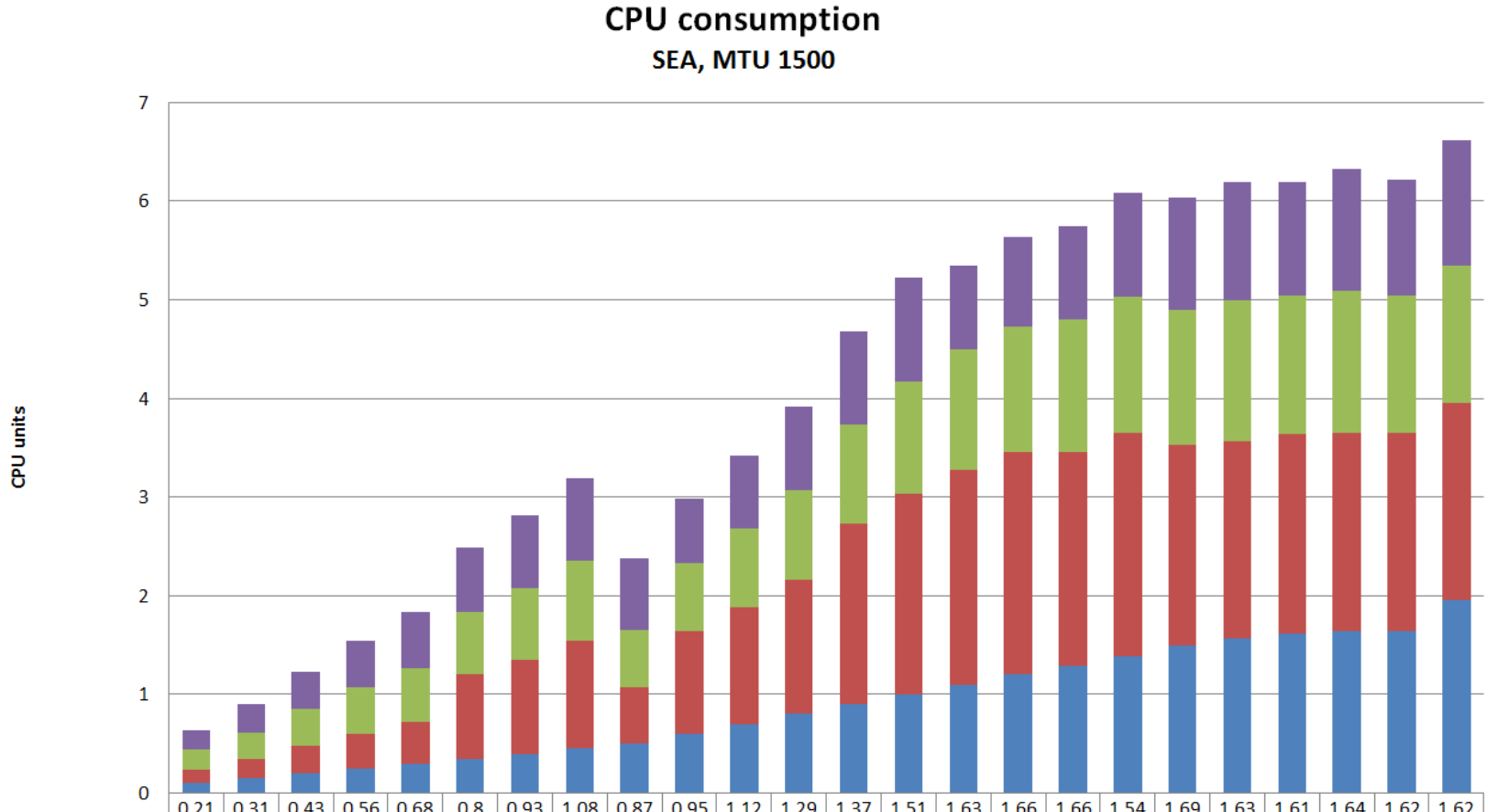
Throughput: 930 Mbit/s

D'un point de vue global, quelles instances sont impliquées dans l'activité réseau



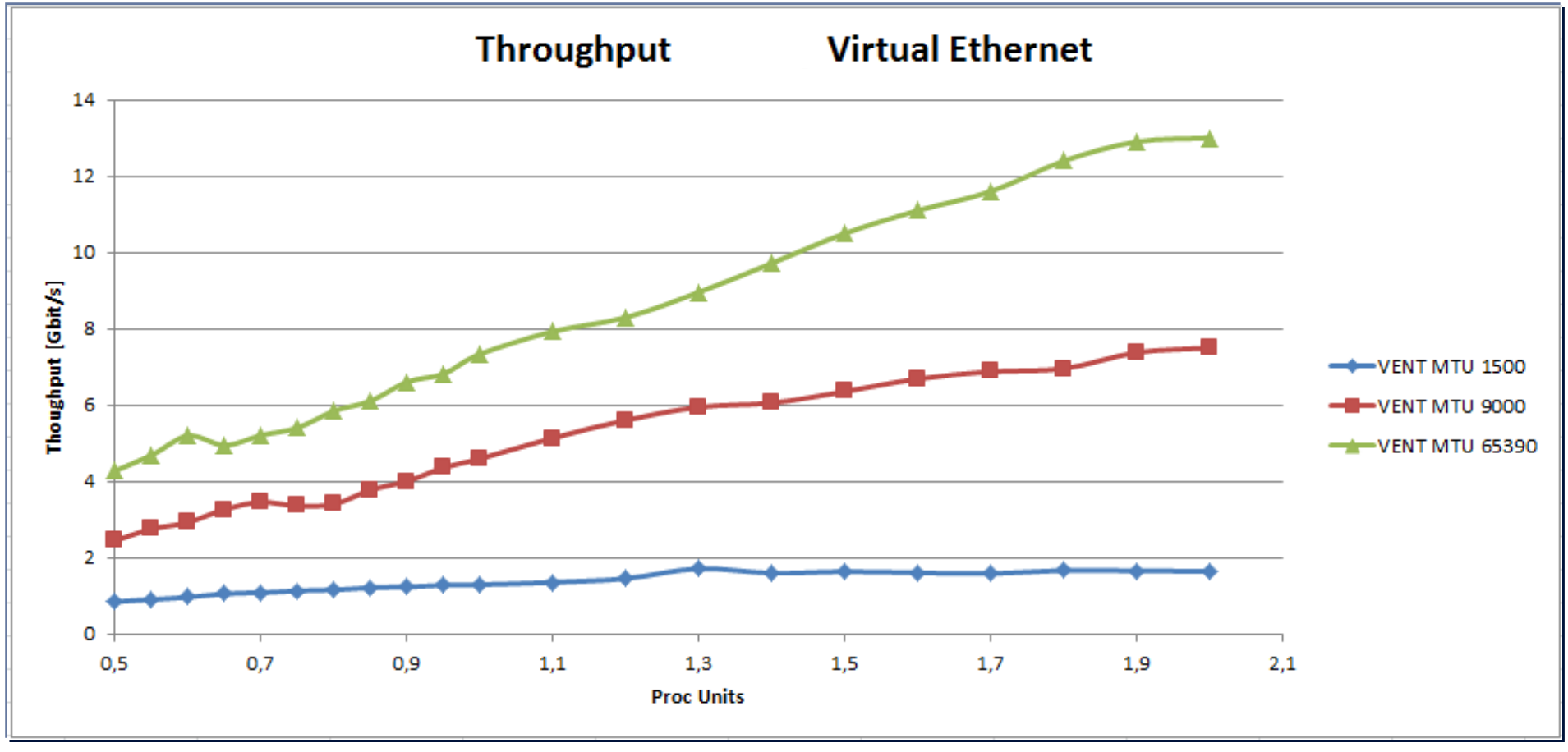
Ethernet virtuel : mesure de l'activité réseau

- Sans aucun paramétrage, on atteint un débit de 1.6Gbits/s avec plus de 6 CPUs



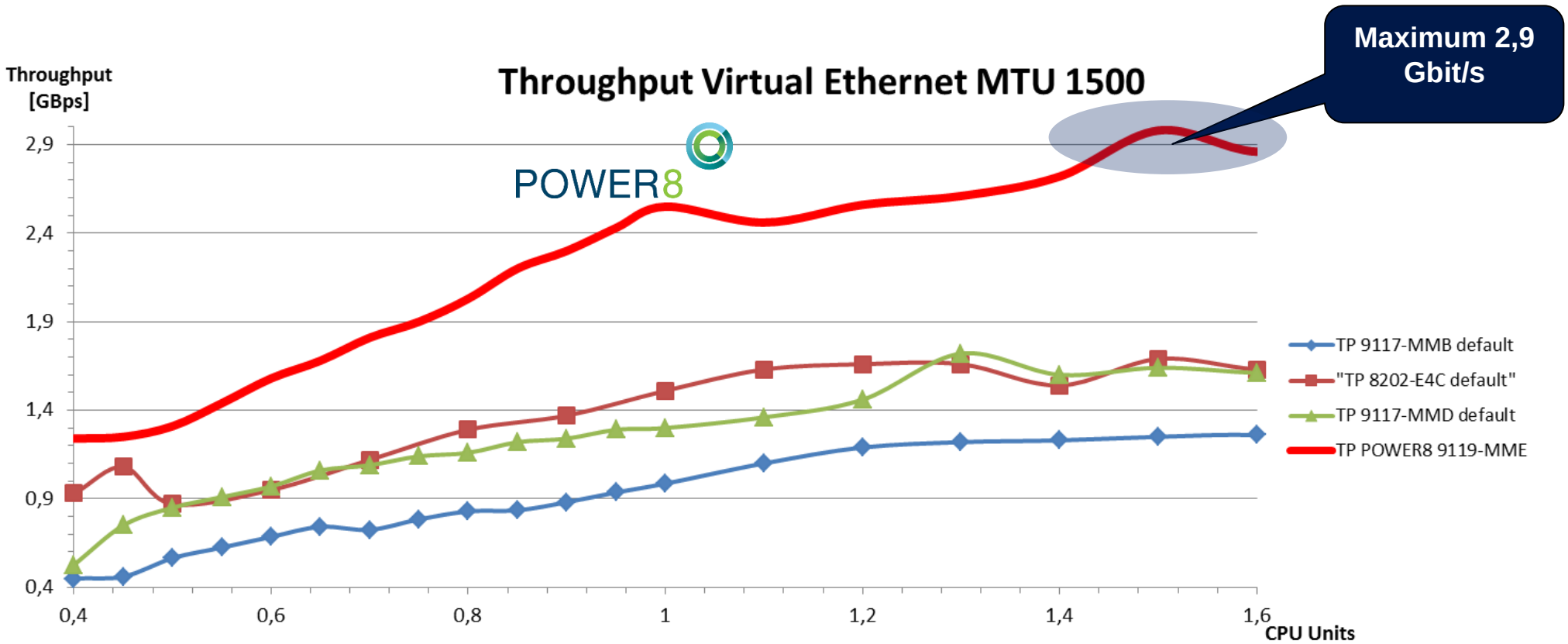
	0,21	0,31	0,43	0,56	0,68	0,8	0,93	1,08	0,87	0,95	1,12	1,29	1,37	1,51	1,63	1,66	1,66	1,54	1,69	1,63	1,61	1,64	1,62	1,62
Receiving VIOS	0,2	0,29	0,38	0,47	0,57	0,65	0,74	0,83	0,73	0,66	0,74	0,85	0,94	1,06	0,84	0,91	0,95	1,05	1,13	1,2	1,15	1,24	1,18	1,28
Sending VIOS	0,2	0,27	0,37	0,47	0,54	0,64	0,73	0,82	0,58	0,69	0,8	0,91	1,01	1,14	1,22	1,27	1,34	1,38	1,37	1,43	1,4	1,44	1,39	1,39
Server	0,14	0,19	0,28	0,35	0,42	0,85	0,95	1,09	0,57	1,04	1,18	1,36	1,83	2,03	2,18	2,26	2,17	2,26	2,04	1,99	2,02	2,01	2,01	1,99
Client	0,1	0,15	0,2	0,25	0,3	0,35	0,4	0,45	0,5	0,6	0,7	0,8	0,9	1	1,1	1,2	1,29	1,39	1,49	1,57	1,62	1,64	1,64	1,96

Ethernet virtuel : débit selon la mtu



Ethernet virtuel : débit selon le modèle de power

- **Benchmark avec 8 sessions TCP en paralléle**
- **Configuration:**
  - **LPAR cliente :**
    - **Power E870 9119-MME**
    - **AIX 7.1 TL3 SP 4**
    - **capped**
    - **2 VPs**
    - **Virtual Ethernet Adapter, MTU 1500**
  - **LPAR serveur :**
    - **Power E870 9119-MME (Same as client)**
    - **AIX 7.1 TL3 SP 4**
    - **uncapped**
    - **4 VPs**
    - **Virtual Ethernet Adapter, MTU 1500**



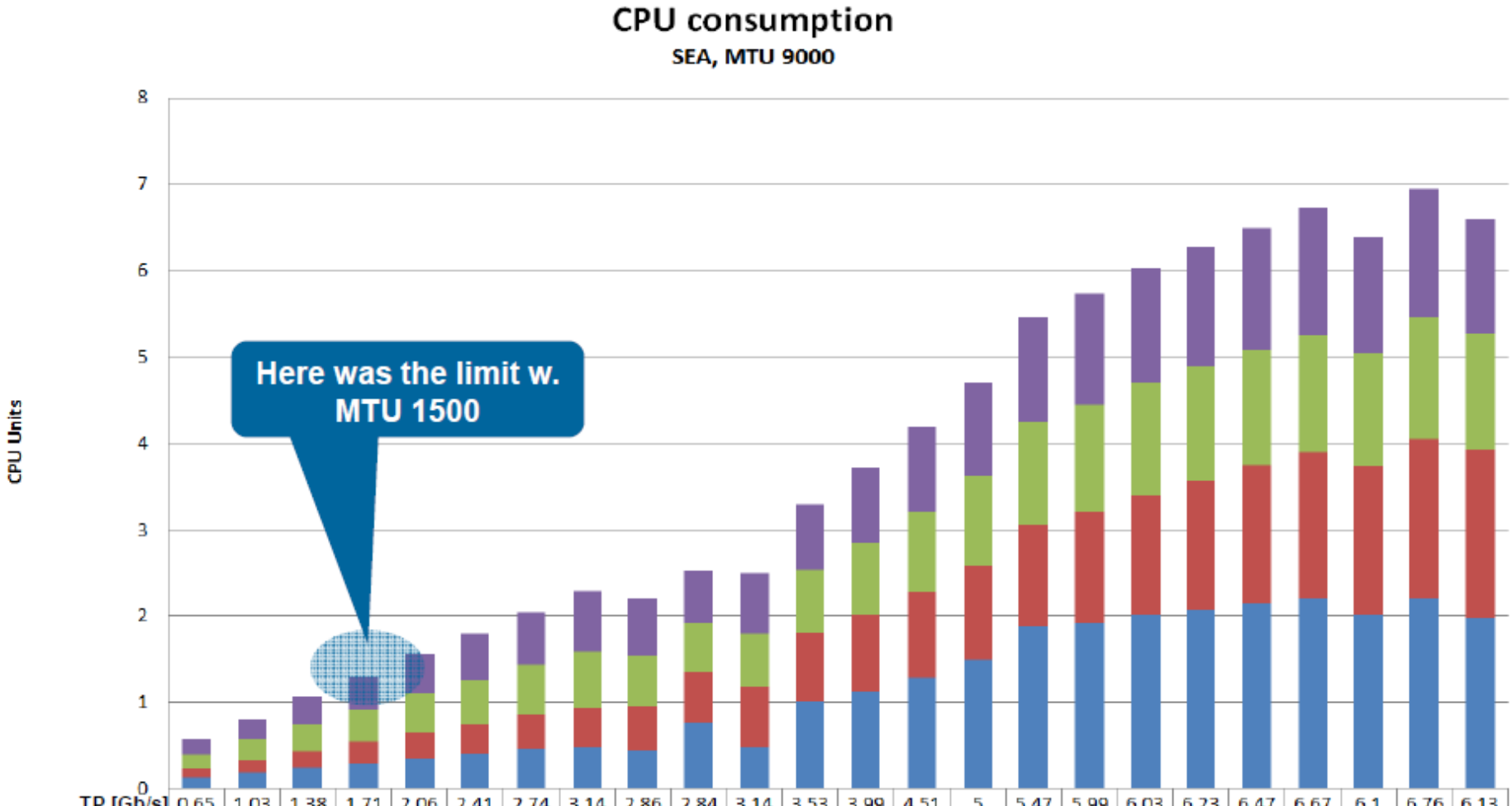
Tuning pour la virtualisation du 10Gb LAN

- Options pour augmenter le débit et réduire la consommation CPU
  - jumbo frame
  - segmentation offload – segment aggregation

- Jumbo frame
  - Doit être implémenté sur une base «end-to-end»
  - AIX supporte l'option mtu_discover qui permet de négocier le mtu en début de connexion
  - Le paramètre mtu_size peut être changé dynamiquement de 1500 to 9000 sur les interfaces réseau virtuels

Jumbo frame

- Avec mtu à 9000, environ 6Gb/s pour 6 CPUs… 3.7 x plus de débit



TP [Gb/s]	0,65	1,03	1,38	1,71	2,06	2,41	2,74	3,14	2,86	2,84	3,14	3,53	3,99	4,51	5	5,47	5,99	6,03	6,23	6,47	6,67	6,1	6,76	6,13
Receiving VIOS	0,18	0,24	0,32	0,39	0,46	0,53	0,6	0,7	0,66	0,61	0,7	0,76	0,86	0,97	1,08	1,21	1,3	1,32	1,38	1,42	1,46	1,33	1,49	1,33
Sending VIOS	0,16	0,24	0,31	0,37	0,45	0,51	0,58	0,66	0,59	0,55	0,62	0,73	0,84	0,94	1,04	1,19	1,23	1,3	1,33	1,33	1,37	1,32	1,41	1,35
Client	0,1	0,15	0,2	0,25	0,3	0,35	0,4	0,45	0,5	0,6	0,69	0,79	0,89	0,99	1,09	1,19	1,29	1,39	1,49	1,59	1,69	1,72	1,84	1,95
Server	0,13	0,18	0,24	0,29	0,35	0,4	0,46	0,48	0,45	0,76	0,48	1,01	1,12	1,28	1,49	1,87	1,92	2,01	2,07	2,15	2,2	2,01	2,21	1,97

Segmentation offload – Segment aggregation

Segmentation offload (largesend)

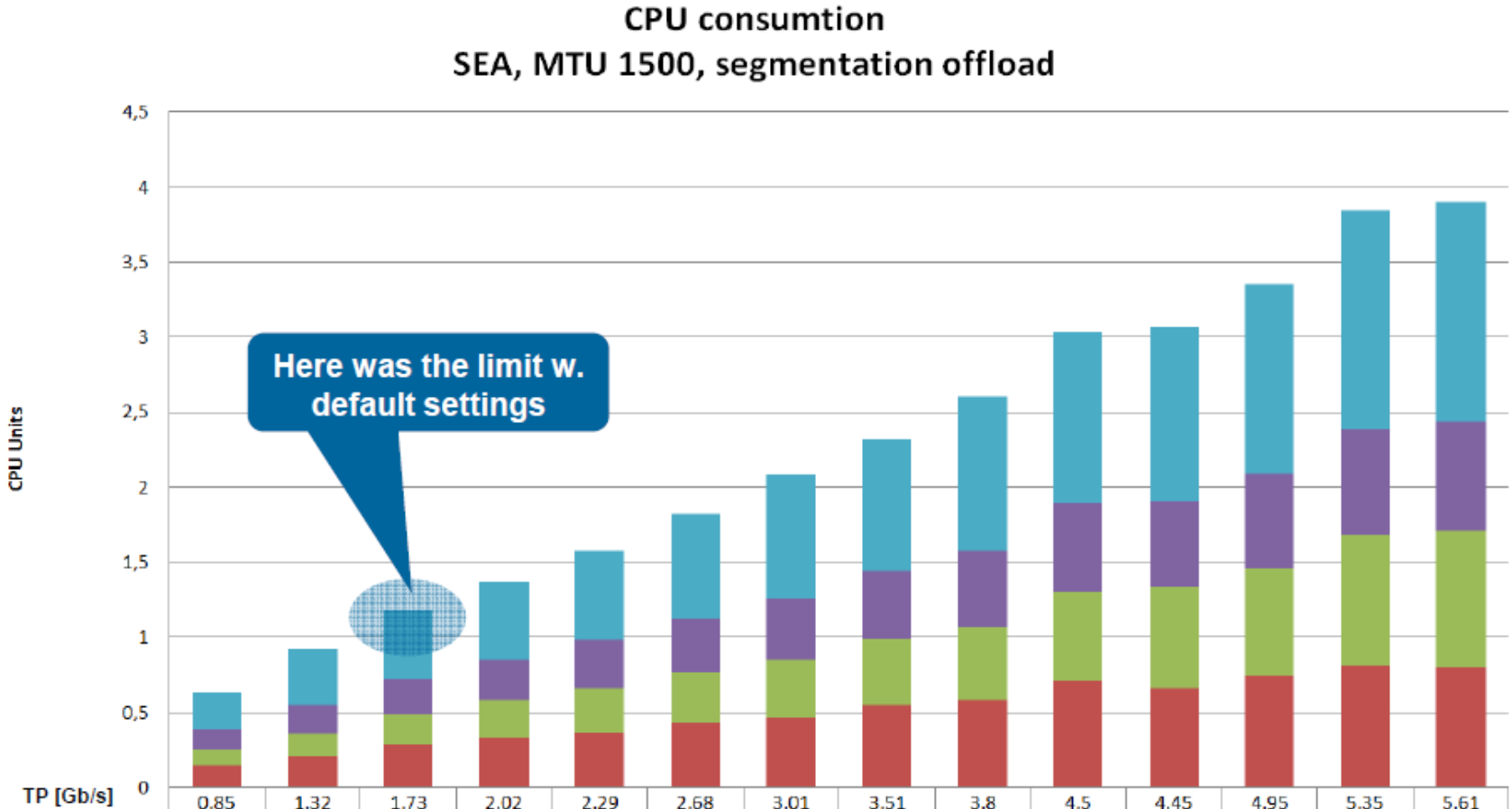
- La tâche de segmenter les données en frame de “mtu” appropriée est déportée de l’OS à la carte physique. La LPAR peut envoyer des paquets de 64K au travers des adapteurs ethernet virtuel.
- Les bénéfices directs sont les suivants:
  - Réduction significative de la consommation CPU dans la LPAR et dans le VIOS pour **l’envoi** de paquets
  - Amélioration du débit effectif **sortant** du VIOS pour les connexions réseau rapide.
- La configuration se fait uniquement sur les LPARs et le VIOS. Aucune intervention n’est requise sur les équipements réseau physique
  - Sur les LPARs : activer le “largesend” au niveau de l’interface réseau (en0)
  - Sur les VIOS : activer le “largesend” au niveau des cartes physiques et du SEA

Segment aggregation (largereceive)

- Les paquets sont bufferisés au niveau du VIOS et passé à l’adapteur Ethernet virtuel de la LPAR par bloc de 64K.
- Les bénéfices directs sont les suivants:
  - Réduction significative de la consommation CPU dans la LPAR et dans le VIOS pour **la réception** de paquets
  - Amélioration du débit effectif **entrant** du VIOS pour les connexions réseau rapide.
- La configuration se fait uniquement sur le VIOS. Aucune intervention n’est requise sur les équipements réseau physique, ni sur les LPARs
  - Sur les VIOS : activer le “largereceive” au niveau des cartes physiques et du SEA

Consommation CPU globale pour un SEA sur 10GbE avec largesend

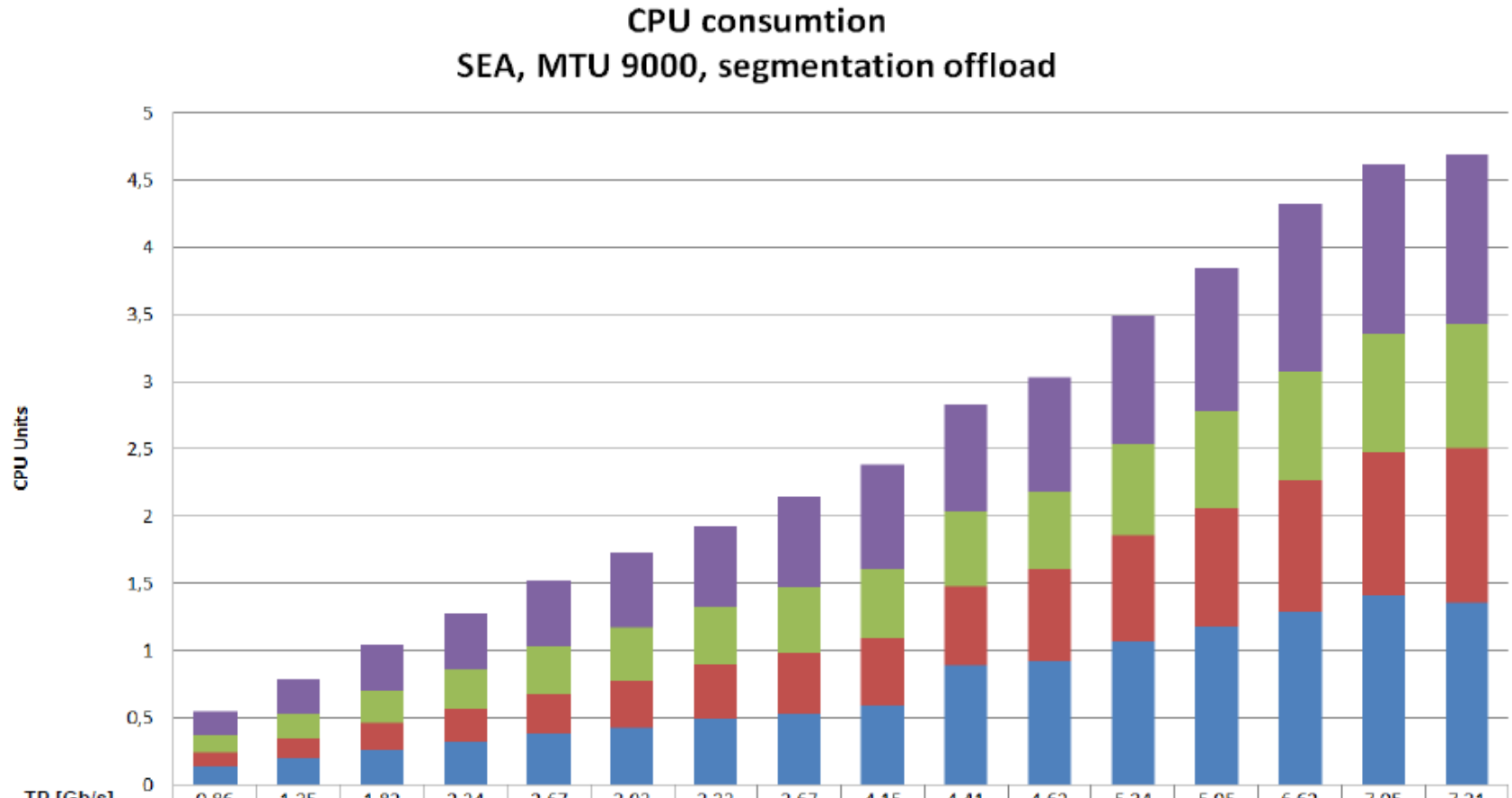
- Avec largesend, environ 5.6Gb/s pour 4 CPUs … 3.5 x plus avec 2/3 de CPU



TP [Gb/s]	0,85	1,32	1,73	2,02	2,29	2,68	3,01	3,51	3,8	4,5	4,45	4,95	5,35	5,61
Receiving VIOS	0,26	0,37	0,46	0,52	0,6	0,69	0,82	0,88	1,02	1,15	1,16	1,26	1,45	1,46
Sending VIOS	0,13	0,19	0,23	0,27	0,32	0,36	0,41	0,45	0,51	0,59	0,57	0,63	0,71	0,73
Client	0,1	0,15	0,2	0,25	0,29	0,34	0,39	0,44	0,49	0,59	0,68	0,71	0,87	0,91
Server	0,15	0,21	0,29	0,33	0,37	0,43	0,46	0,55	0,58	0,71	0,66	0,75	0,81	0,8

Consommation CPU globale pour un SEA sur 10GbE avec largesend et jumbo frame

- Avec largesend ET jumbo frames, 7.3 Gb/s pour 4.6 CPUs …
  … 4.5 x plus avec 75% de CPU…

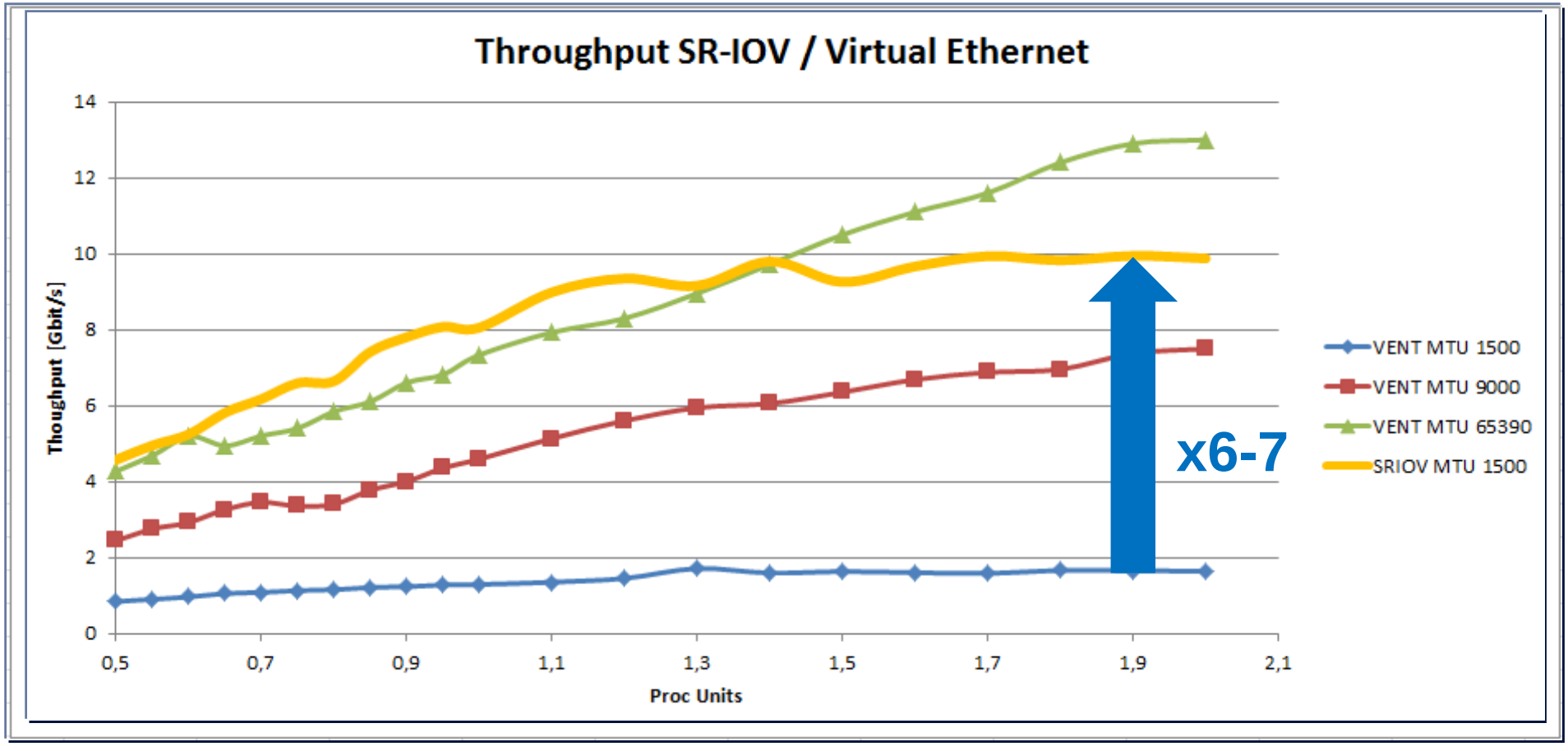


TP [Gb/s]	0,86	1,35	1,82	2,24	2,67	3,03	3,33	3,67	4,15	4,41	4,62	5,34	5,95	6,62	7,05	7,31
Receiving VIOS	0,18	0,26	0,34	0,42	0,49	0,56	0,61	0,68	0,77	0,8	0,85	0,95	1,06	1,24	1,26	1,26
Sending VIOS	0,13	0,18	0,24	0,29	0,35	0,39	0,42	0,49	0,52	0,55	0,57	0,68	0,73	0,81	0,88	0,92
Client	0,1	0,15	0,2	0,25	0,3	0,35	0,4	0,45	0,5	0,59	0,69	0,79	0,88	0,98	1,07	1,15
Server	0,14	0,2	0,26	0,32	0,38	0,43	0,5	0,53	0,59	0,89	0,92	1,07	1,18	1,29	1,41	1,36

SR-IOV PERFORMANCE

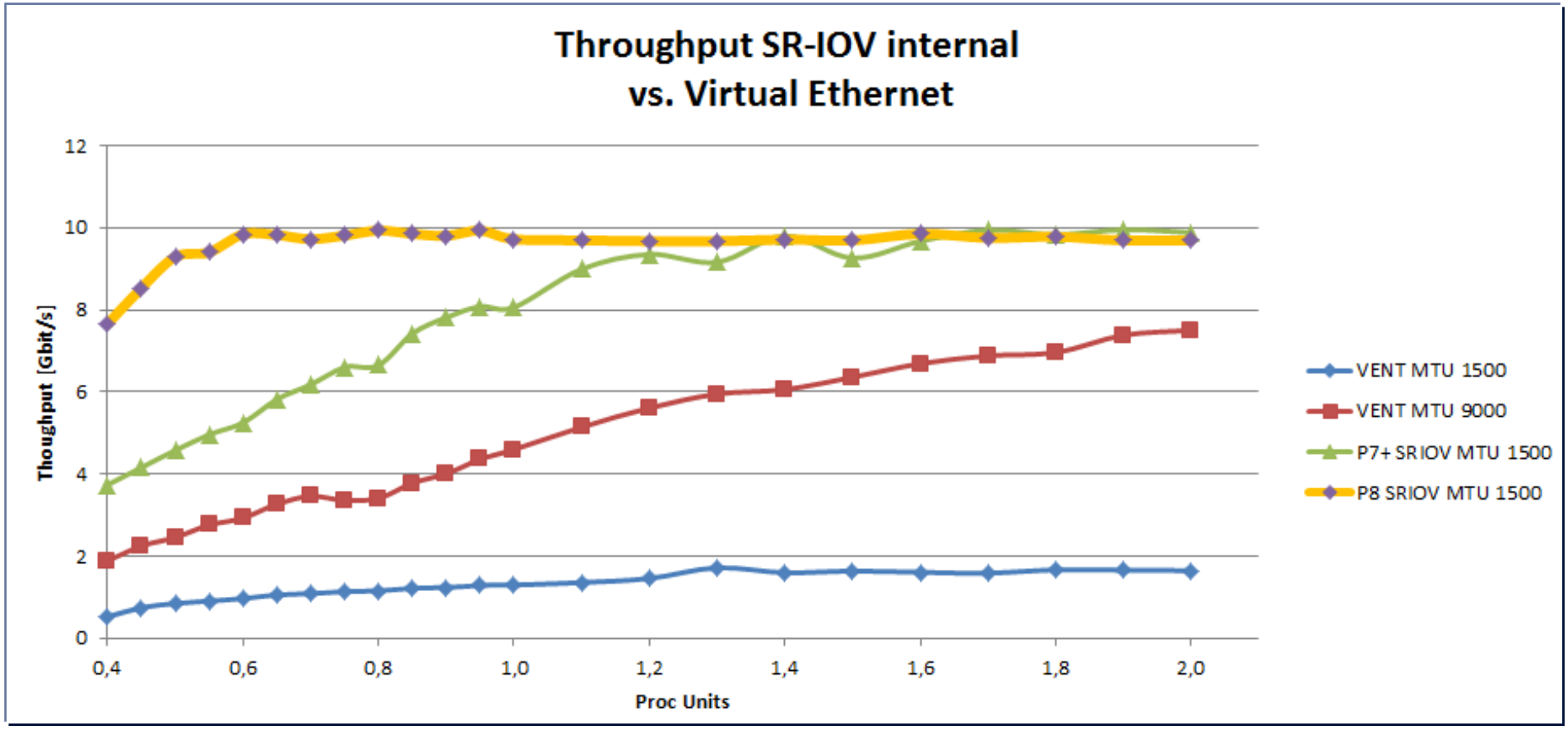
Performance tests for Single Root I/O Virtualization

- **Benchmark results 10 Gigabit Ethernet**
  - SR-IOV provides a better out-of-box performance than SEA with 10Gb (x 6-7 better TP/CPU ratio than Virtual Ethernet)
  - Much lower CPU usage at MTU 1500 or 9000
  - **CPU utilization with SR-IOV is equivalent to Virtual Ethernet switching with MTU 64K**



POWER8 SR-IOV internal switching on POWER8 S824

- POWER8 provides access to adapter line-speed with less CPU units compared to POWER7+.

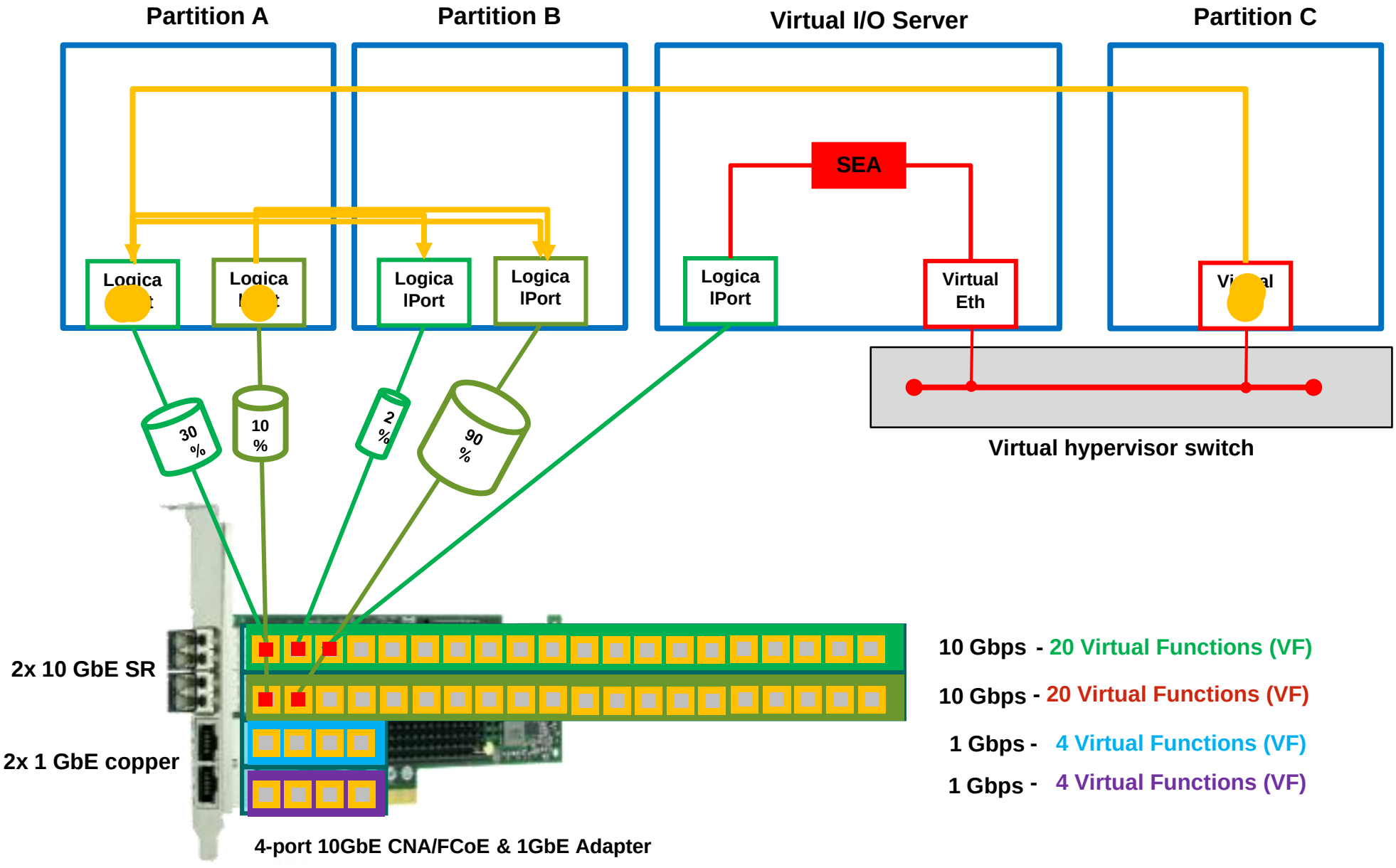


CPU units consumption: SEA / SR-IOV

Throughput:	2.5 Gbit/s	5 Gbit/s	5 Gbit/s	10 Gbit/s
CPU consumtion	5.8	2.8	0.8	1.6
	E870 SEA default	E870 SEA + LSO (large segment offload)	S824 SR-IOV	S824 SR-IOV

SR-IOV ARCHITECTURE

SR-IOV architecture



SR-IOV ADAPTERS AND PLACEMENT

Systems with SR-IOV Support

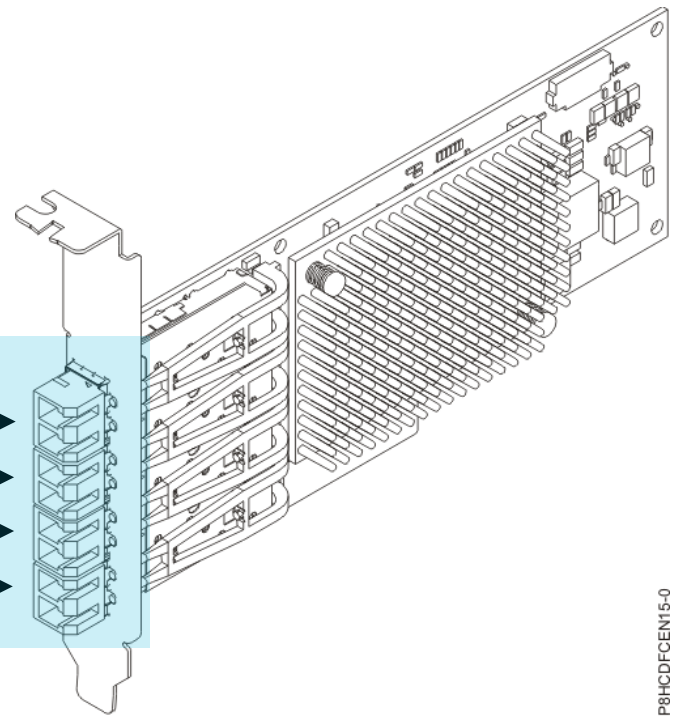
- **4/2014 GA**
  - **9117-MMD (IBM Power 770), 9179-MHD (IBM Power 780), 8412-EAD (IBM Power ESE)**
    - **System node PCIe slots**
- **3/2015 GA**
  - **9119-MME (IBM Power System E870), 9119-MHE (IBM Power System E880)**
    - **System node PCIe slots**
- **6/2015 GA**
  - **Power scale-out servers, expanded options for Power E870 and E880, and Power E850**
  - **PCIe Gen3 I/O expansion drawer.**
  - **The following POWER8 PCIe slots are SR-IOV capable:**
    - **All Power E870/E880 and Power E850 system node slots.**
    - **Slots C6, C7, C10, and C12 of a Power S814 (1S 4U) or S812L (1S 2U) server.**
    - **Slots C2, C3, C4, C5, C6, C7, C10, and C12 of a S824 or S824L server (2-socket, 4U) with both sockets populated. If only one socket is populated, then C6, C7, C10, and C12.**
    - **Slots C2, C3, C5, C6, C7, C10, and C12 of a S822 or S822L server (2-socket, 2U) with both sockets populated. If only one socket is populated, then C6, C7, C10, and C12.**
    - **Slots C1 and C4 of the 6-slot Fan-out Module in a PCIe Gen3 I/O drawer. If system memory is less than 128GB only slot C1 of a Fan-out Module is SR-IOV capable.**

New PCIe Gen3 adapters

- **PCIe3 4-port 10Gb FCoE SR optical or copper twinax**
- **#EN15: Optical 10GbE SR, Full Hight**
- **#EN16: Optical 10GbE SR, Low Profile**
- **#EN17: Copper 10GbE SR, Full Hight**
- **#EN18: Copper 10GbE SR, Low Profile**

16 VFs per port

4 ports 10GbE CNA



=

64 VFs per adapter

	Announced	GA
Supported in PCIe Gen3 I/O drawer and the 4U scale-out system units and in the Power E850/E870/E880 system node slots.	28. April 2015	June 2015

SR-IOV capable slots on E870/E880 systems

- **9119-MHE or 9119-MME system nodes**
- **All slot positions are SR-IOV capable**
- **See adapter placement rules**
  http://www-01.ibm.com/support/knowledgecenter/9119-MME/p8eab/p8eab_87x_88x_slot_details.htm?lang=en

Slot Position	Slot	Device Feature or CCIN	Device Description	SR-IOV capable	Used by Partition / Profile
1	P1-C1	EN0J	PCIe2 4-port (10Gb FCoE & 1GbE) SR&RJ45 Adapter	yes	vios1
2	P1-C2		8 Gigabit PCI Express Dual Port Fibre Channel Adapter	yes	lpar1
3	P1-C3		8 Gigabit PCI Express Dual Port Fibre Channel Adapter	yes	vios1
4	P1-C4	EN0J	PCIe2 4-port (10Gb FCoE & 1GbE) SR&RJ45 Adapter	yes	vios2
5	P1-C5		8 Gigabit PCI Express Dual Port Fibre Channel Adapter	yes	vios2
6	P1-C6		SAS RAID Controller, PCIe3 x8, Quad-port 6Gb	yes	
7	P1-C7		8 Gigabit PCI Express Dual Port Fibre Channel Adapter	yes	vios2
8	P1-C8		Ethernet controller	yes	

PCIe3, x16 Slot with direct Processor Module connection
PCIe3, x8 Slot with direct Processor Module connection

SR-IOV capable slots on PCIe Gen3 I/O expansion drawer

- **Two slot positions per Fan-out module are SR-IOV capable**
- **See adapter placement rules**
  https://www-01.ibm.com/support/knowledgecenter/9119-MHE/p8eab/p8eab_emx0_slot_details.htm

Slot Position	Slot	Device Feature or CCIN	Device Description	SR-IOV capable	Used by Partition / Profile
1	P1-C1	EN0H	PCIe2 4-port (10Gb FCoE & 1GbE) SR&RJ45 Adapter	yes	vios1
2	P1-C2		8 Gigabit PCI Express Dual Port Fibre Channel Adapter	no	vios1
3	P1-C3		8 Gigabit PCI Express Dual Port Fibre Channel Adapter	no	vios2
4	P1-C4	EN0H	PCIe2 4-port (10Gb FCoE & 1GbE) SR&RJ45 Adapter	yes	vios1
5	P1-C5		8 Gigabit PCI Express Dual Port Fibre Channel Adapter	no	vios2
6	P1-C6		SAS RAID Controller, PCIe3 x8, Quad-port 6Gb	no	
7	P2-C1		8 Gigabit PCI Express Dual Port Fibre Channel Adapter	yes	vios2
8	P2-C2		Ethernet controller	no	
9	P2-C3			no	
10	P2-C4			yes	
11	P2-C5			no	
12	P2-C6			no	

PCIe3, x16 Slot PCIe3, x8 Slot

SR-IOV capable slots on S824 systems

Slot Position	Slot	Device Feature or CCIN	Device Description	SR-IOV capable	Used by Partition / Profile
1	C2	EN0H	PCIe2 4-port (10Gb FCoE & 1GbE) SR&RJ45 Adapter	yes	vios1
2	C3		8 Gigabit PCI Express Dual Port Fibre Channel Adapter	yes	vios1
3	C4		8 Gigabit PCI Express Dual Port Fibre Channel Adapter	yes	vios1
4	C5	EN0H	PCIe2 4-port (10Gb FCoE & 1GbE) SR&RJ45 Adapter	yes	vios2
5	C6		8 Gigabit PCI Express Dual Port Fibre Channel Adapter	yes	vios2
6	C7		SAS RAID Controller, PCIe3 x8, Quad-port 6Gb	yes	
7	C8		8 Gigabit PCI Express Dual Port Fibre Channel Adapter	no	vios2
8	C9		Ethernet controller	no	
9	C10	EN0H	PCIe2 4-port (10Gb FCoE & 1GbE) SR&RJ45 Adapter	yes	Hypervisor
10	C11		Empty Slot	no	
11	C12	EN0H	PCIe2 4-port (10Gb FCoE & 1GbE) SR&RJ45 Adapter	yes	Hypervisor
	C14		IBM PCIe3 x8 Cache SAS RAID Internal Adapter 6Gb	yes	
	C15		IBM PCIe3 x8 Cache SAS RAID Internal Adapter 6Gb	yes	

PCIe3, x16 Slot with direct Processor Module connection
PCIe3, x8 Slot with direct Processor Module connection
PCIe3, x8 Slot on Switch 1, 32-lane
PCIe3, x8 Slot on Switch 2, 48-lane
For SAS/RAID Adapter

QUALITY OF SERVICE

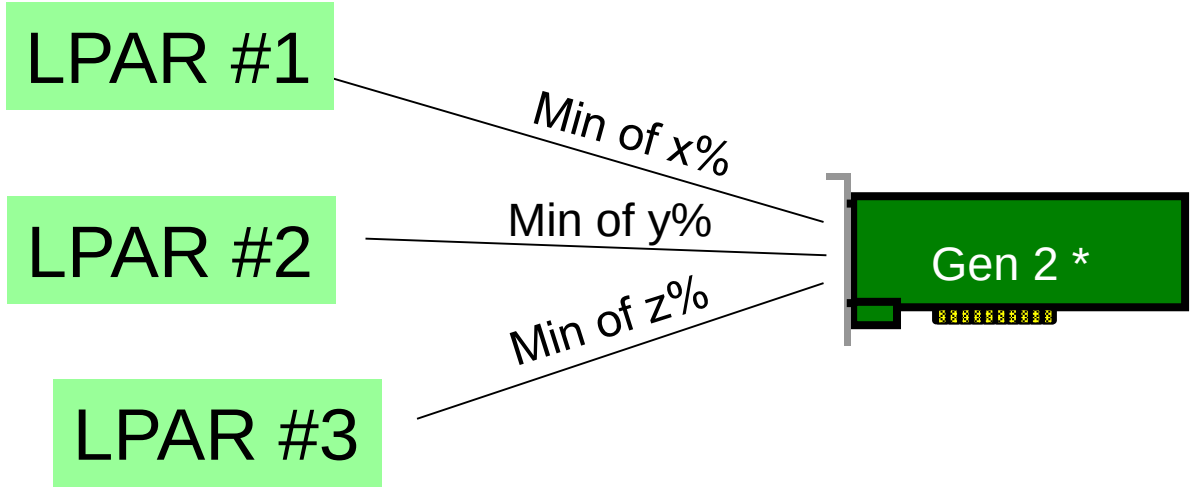
SR-IOV desired bandwidth

- Capacity setting controls adapter and system resource levels, including desired bandwidth
- New logical port must be created to assign new capacity to partitions.
- Capacity can not be changed dynamically.
- Therefore, it may be desired to leave some capacity.
- Capacity setting must be a multiple of the default (2%).

Physical Port ID	Location Code	Port Type	Available Logical Ports	Configured Logical Ports
0	U78CB.001.WIH0002-P1-C7-T1	Converged Ethernet	19	1

Logical port capacity (%) 2 Available physical port capacity: 98.0

SR-IOV desired bandwidth

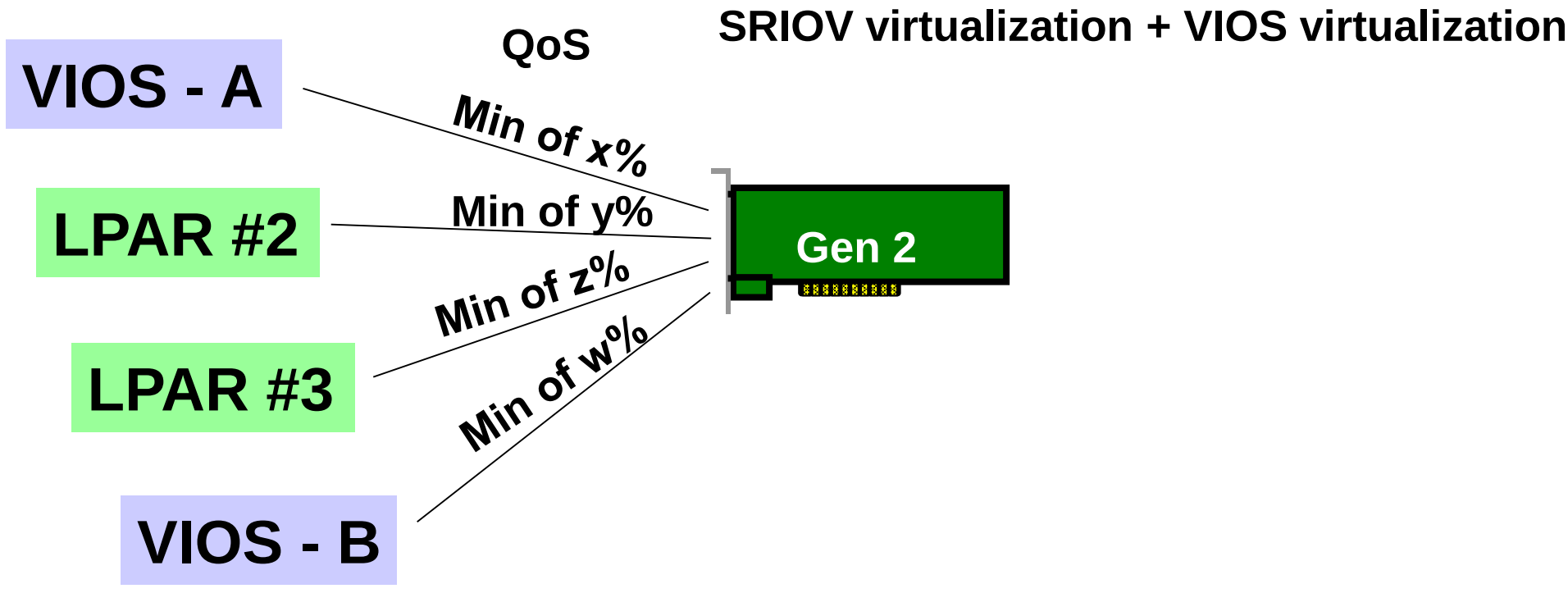


- If LPAR #1 needs it, will have x% desired outgoing bandwidth.
- If LPAR #2 needs it, will have y% desired outgoing bandwidth.
- If LPAR #3 needs it, will have z% desired outgoing bandwidth.
- If additional outgoing bandwidth available, any partition can use it.
- If a partition doesn't need its minimum, that bandwidth is available to other partitions until the owning partition needs it.
- Capacity settings don't have any influence on the incoming bandwidth.

SRIOV …. A Potential Technology of Interest from session aWN06

Virtualization with VIOS example

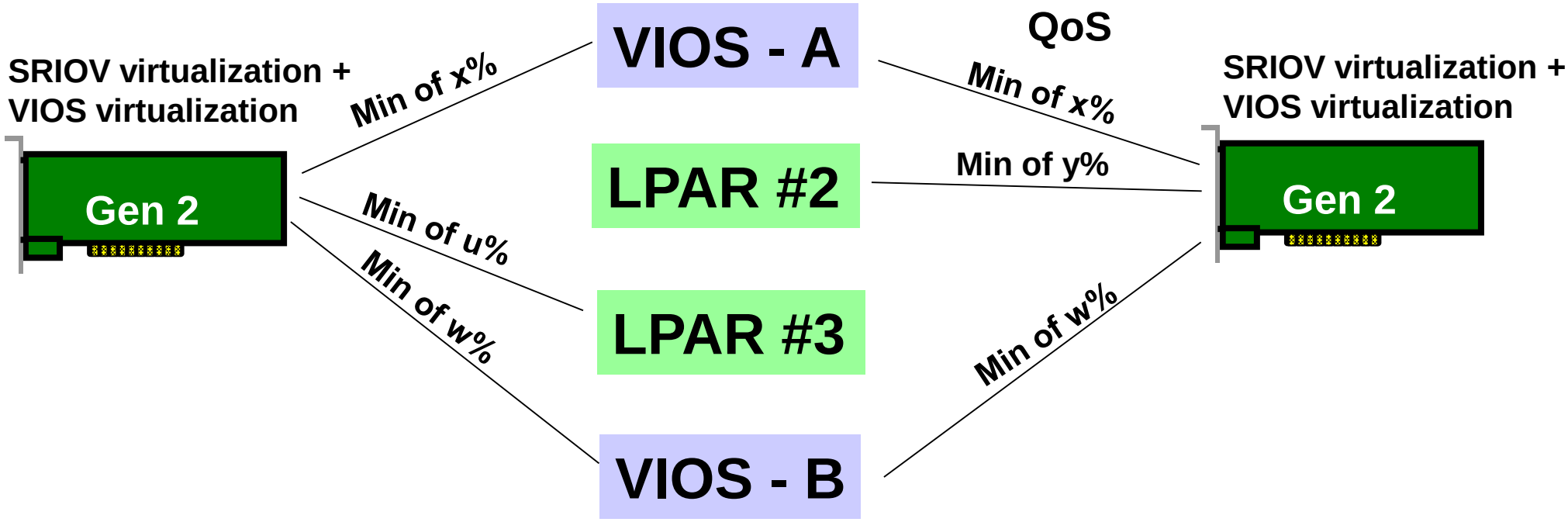
- **Redundant VIOS with one hardware resource**
- **Minimum amount of bandwidth for Quality of Service (QoS)**



SRIOV …. A Potential Technology of Interest from session aWN06

Virtualization with VIOS example

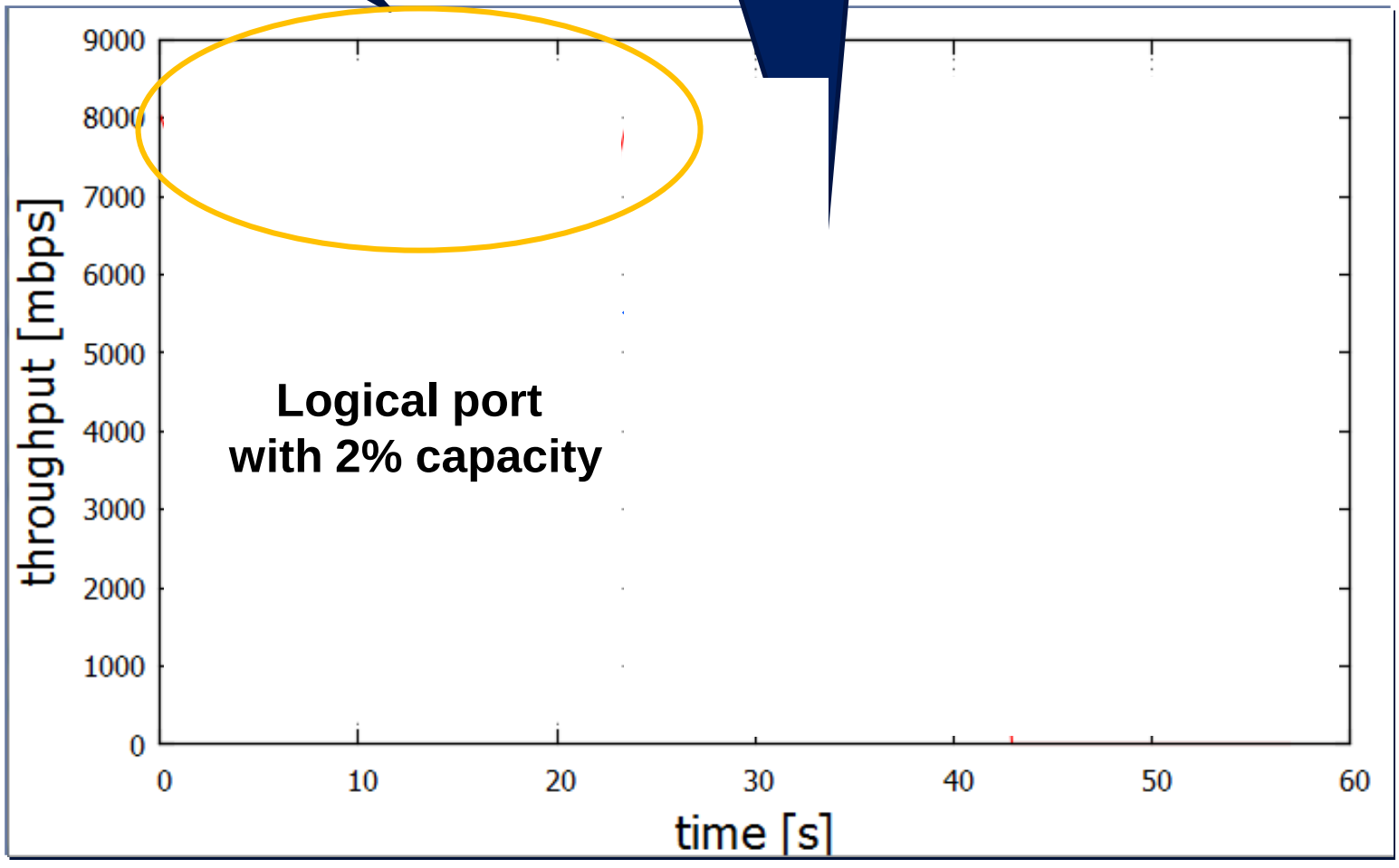
- **Redundant VIOS with redundant hardware resource**
- **Minimum amount of bandwidth for Quality of Service (QoS)**



SR-IOV desired bandwidth

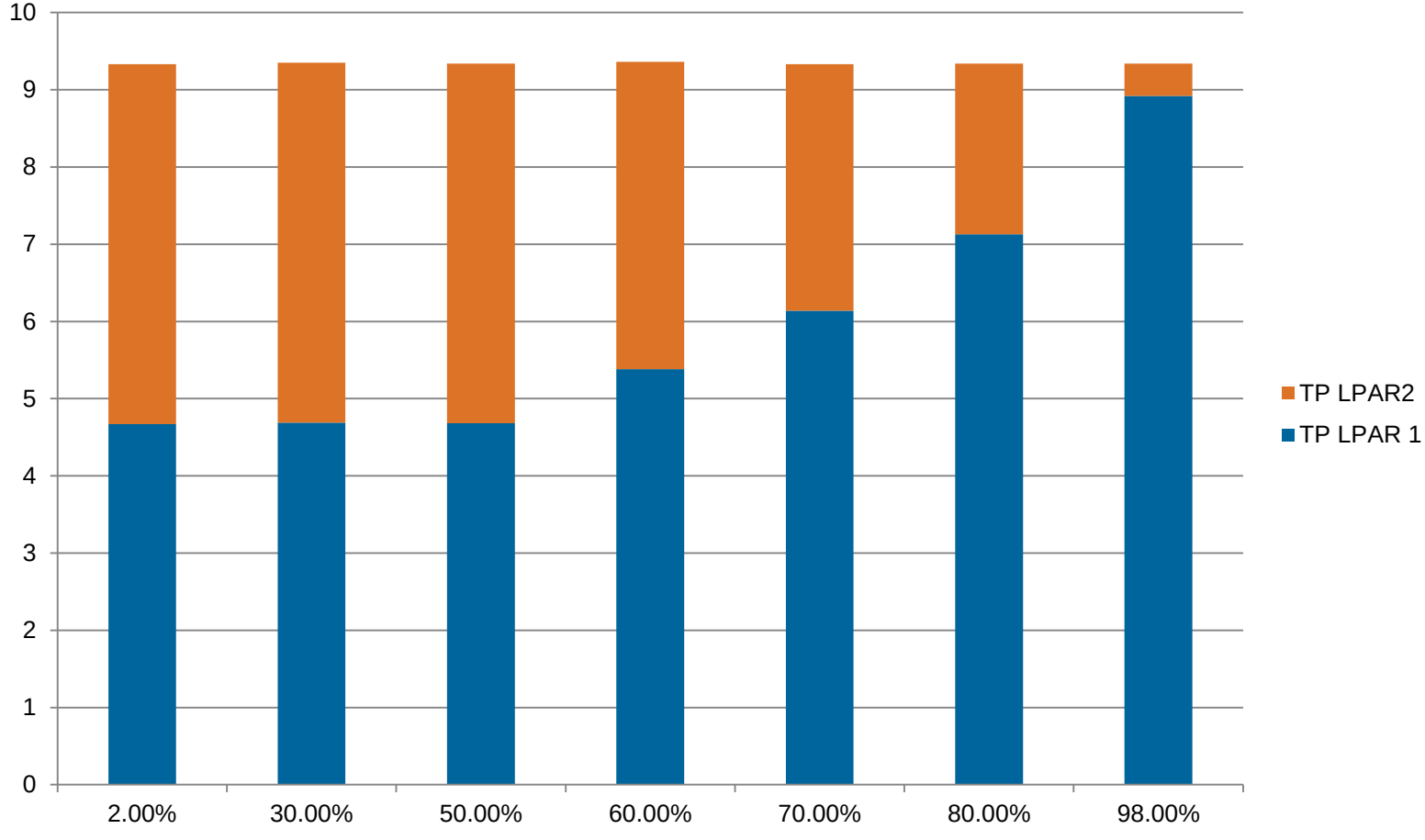
Logical ports is using physical port exclusively

10 seconds competitive situation with other ports



QoS Capacity Tests - Summary

- Bandwidth test with two competitive LPARs
- LPAR1 (blue) with different capacity settings (2% - 98%)
- LPAR2 (red) with fixed capacity of 2%

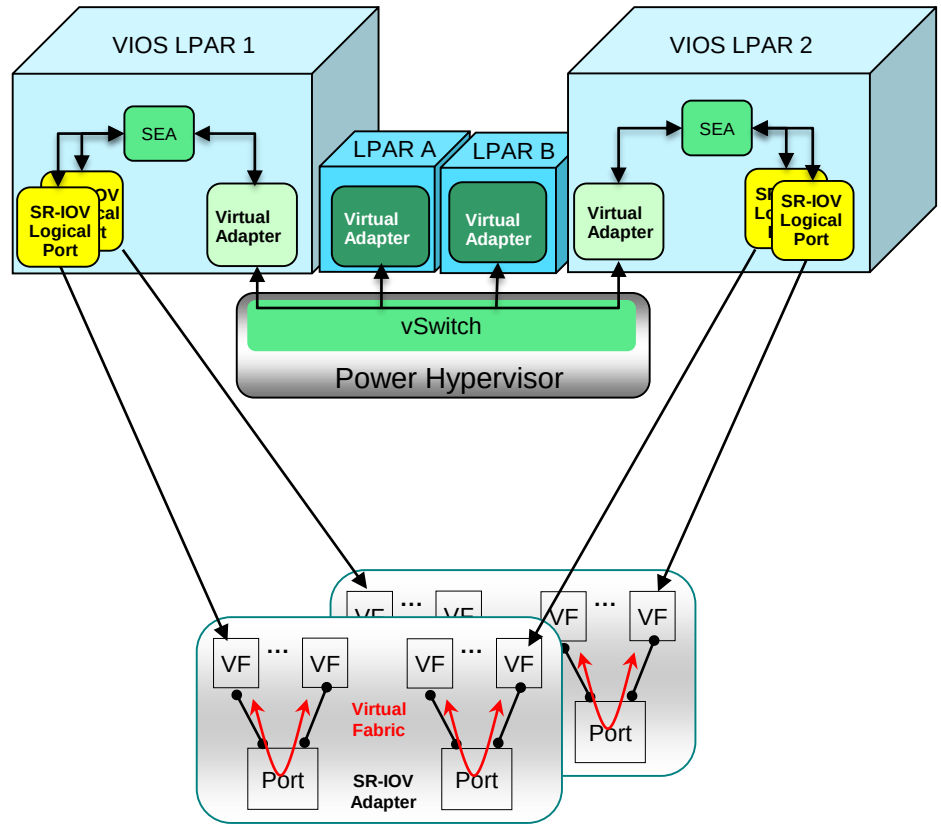


LIVE PARTITION MOBILITY

Live Partition Mobility Options with SR-IOV

Virtual Ethernet configuration

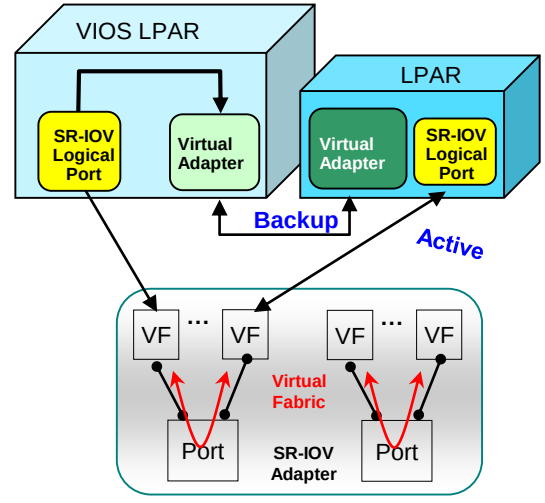
- **Use current Virtual Ethernet support with logical ports as Shared Ethernet Adapter (SEA) physical connections to the network**
- **Does not receive performance benefits provided with SR-IOV Direct Access**
- **Benefits:**
  - **LPM Capability**
  - **Adapter/port sharing to reduce number of adapters**



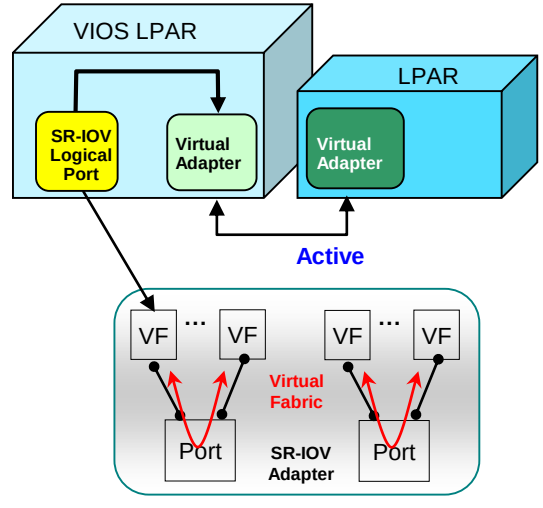
Live Partition Mobility Options with SR-IOV

Active-backup configuration

- Configure SR-IOV logical port as Active connection and Virtual Ethernet adapter as Backup
- Prior to migration, use dynamic LPAR operation to remove SR-IOV logical port
- Virtual Ethernet becomes Active connection
- Migrate the partition
- On target system, configure SR-IOV logical port as Active connection
- Options for AIX and Linux
  - Physical I/O can not be assigned (even temporarily) to an IBM i LPM capable partitions



Normal Active-Backup configuration



Prior to migration, remove Logical Port via DLPAR remove operation

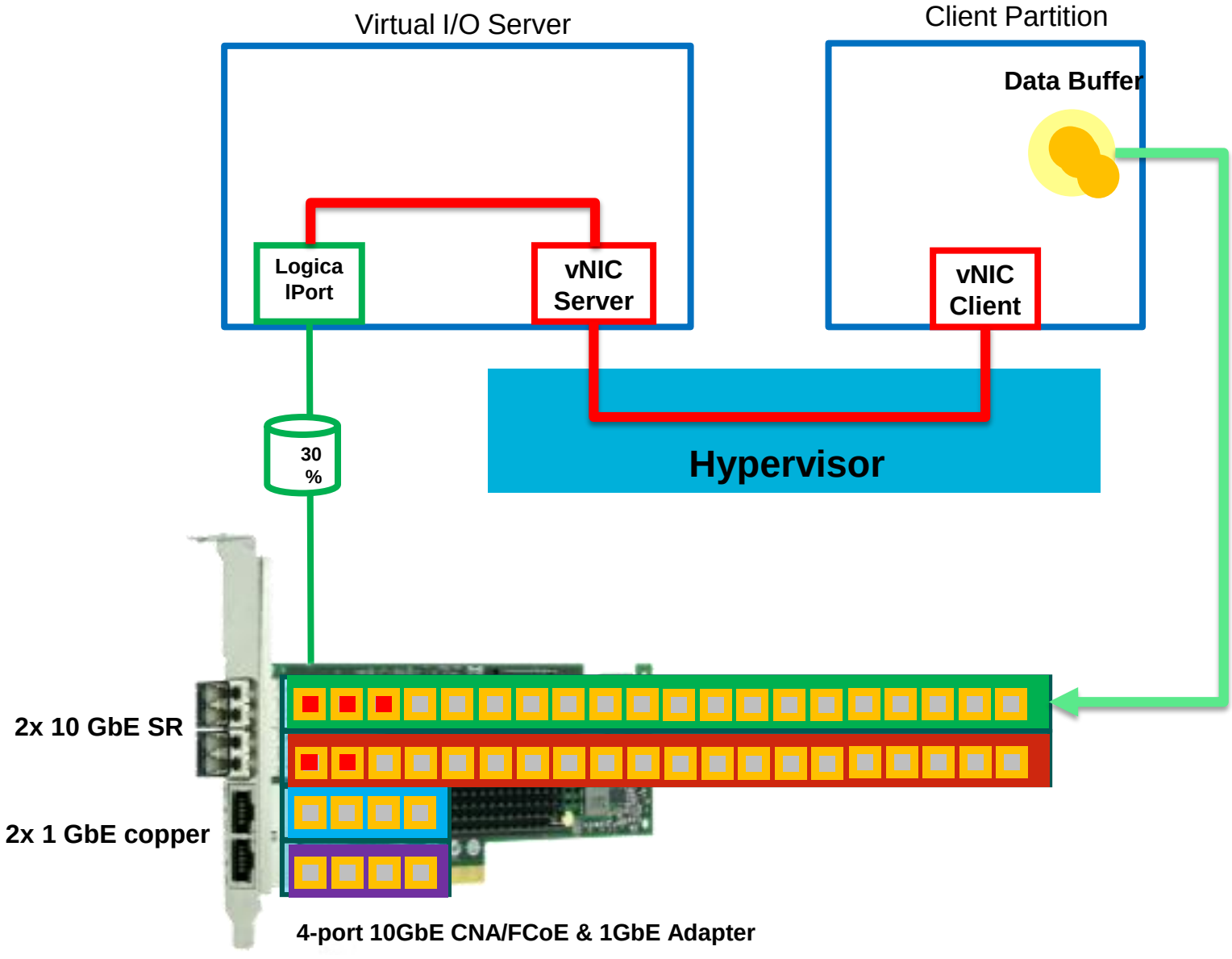
Live Partition Mobility Options with SR-IOV

Virtual Network Interface Controller (vNIC)

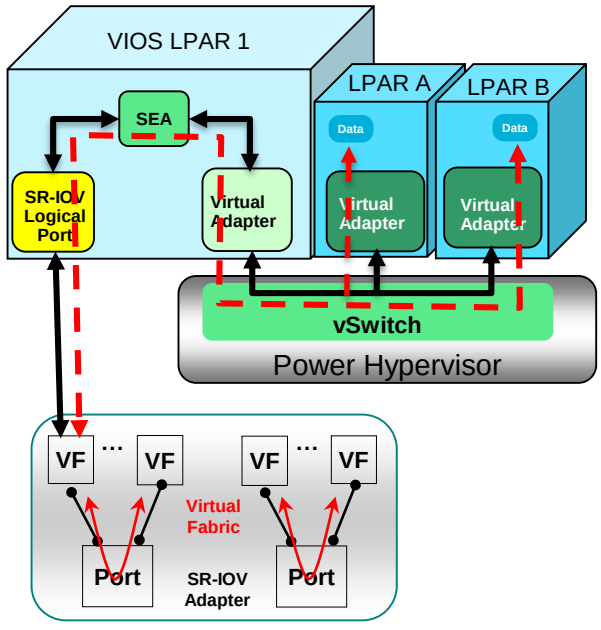
- vNIC is new virtual adapter type
- vNIC leverages SR-IOV to provide a performance optimized virtual NIC solution
- vNIC enables advanced virtualization features such as live partition mobility with SR-IOV adapter sharing
- Leverages SR-IOV Capacity value (QoS)
- Announced October 2015 for AIX & IBM i
  - Linux not announced
  - E850 support not announced
- Pre-req
  - AIX 7.1 TL4 or later or AIX 7.2 or later
  - IBM i 7.1 TR10 or later or 7.2 TR3 or later
  - VIOS 2.2.4, or later
  - Firmware level 8.4, or later

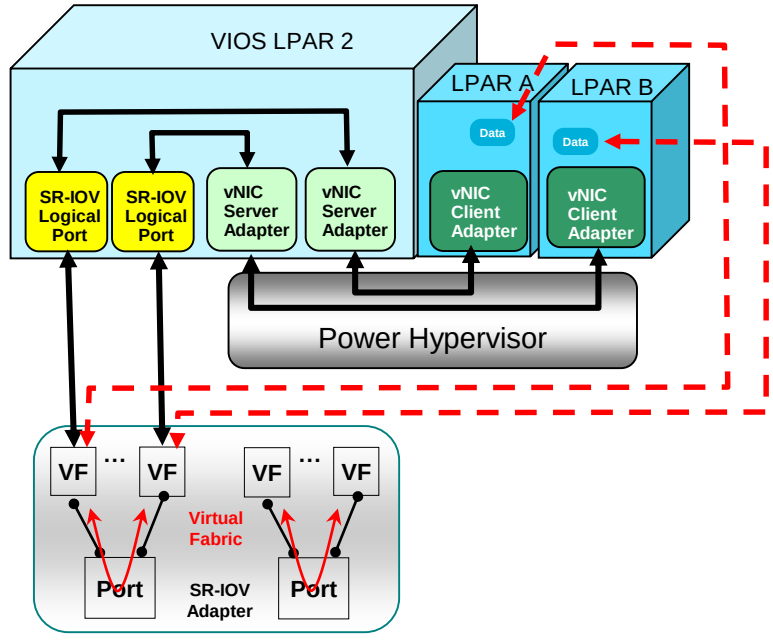
VIRTUAL NETWORK INTERFACE CONTROLLER (VNIC)

vNIC Architecture



Comparison of Virtual Enet & vNIC





Virtual Ethernet (current)

- Multiple copies of data
- Many-to-one relationship between virtual adapters and physical adapter
- QoS based on VLAN tag PCP bits (i.e. 8 traffic classes)

Dedicated vNIC (target 4Q2015)

- SR-IOV with advanced virtualization features (e.g. LPM)
- Improved performance
  - Eliminates data copies
    - Data flows between client partition memory and adapter
  - Optimized control flow, no overhead from the vSwitch or SEA
  - Multiple queue support
- Efficient
  - Lower CPU and Memory usage (no data copy)
  - Leverages adapter offload capabilities
    - LPAR to LPAR communication
- Deterministic QoS
  - One-to-one relationship between vNIC client adapter and SR-IOV logical port
  - Extends logical port QoS

SR-IOV LINK AGGREGATION

SR-IOV and Network Port Aggregation Technologies

Introduction

- There are a number of network port aggregation technologies
- This presentation covers key port aggregation technologies when using SR-IOV
- Main benefits of port aggregation are increased network bandwidth and fault-tolerant of link failures
- Use care when considering switch aware port aggregation technologies with SR-IOV enabled adapters

Network Port Aggregation

AIX

- **EtherChannel or Static Link Aggregation**
- **IEEE 802.3ad/802.1ax Link Aggregation Control Protocol (LACP)**
- **Network Interface Backup (NIB)**

IBM i

- **EtherChannel or Static Link Aggregation**
- **IEEE 802.3ad/802.1ax Link Aggregation Control Protocol (LACP)**
- **Virtual IP Address (VIPA)**

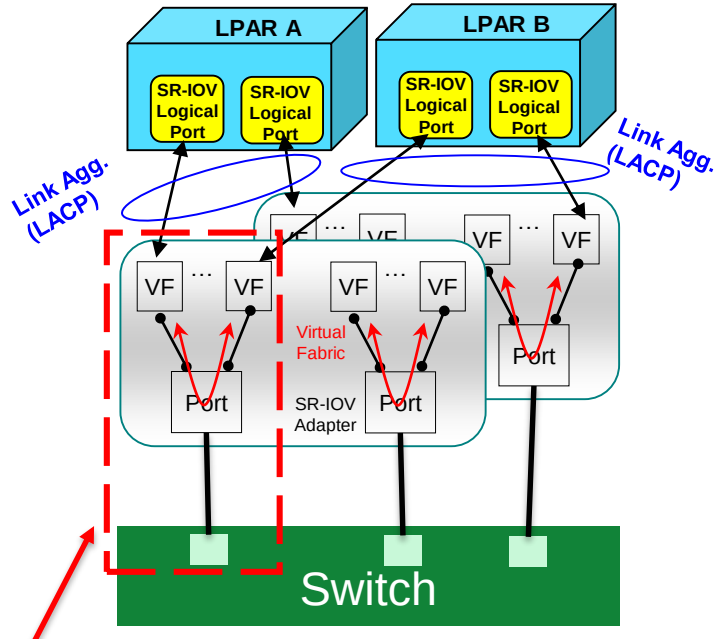
Linux

- **Several Bonding/port trunking modes including LACP and Active-Backup**

Link Aggregation Using LACP

Issue

- **Link Aggregation (LACP) will not function properly with multiple logical ports using the same physical port**
- **Switch expects a single partner (MAC – physical layer) on a link**
- **Multiple SR-IOV logical ports on the same physical port creates multiple partners on the link**



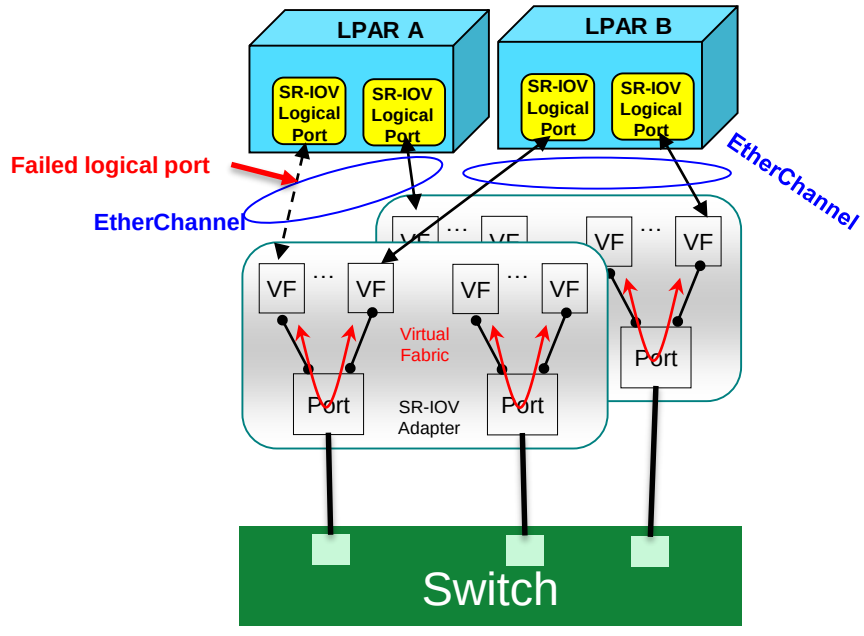
Invalid Link Aggregation configuration. Two logical ports assigned to one physical port.

VF = Virtual Function

Static Link Aggregation

Issue

- SR-IOV logical ports may go down while the physical link remains up
- Switch port failover occurs when physical link goes down
- Switch does not recognize a logical port going down and will continue to send traffic on the physical port
- Static Link Aggregation is not recommended for an SR-IOV configuration

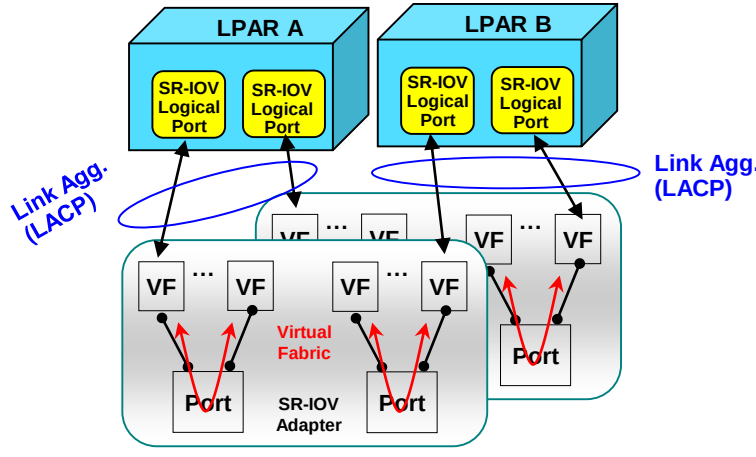


Not recommended – Switch will not detect if logical link fails

Network Port Aggregation Recommendations

If you require bandwidth greater than a single link's bandwidth with failover?

- **Use Link Aggregation (LACP) with one logical port per physical port.**
- **Provides greater bandwidth than a single link with failover**
- **Other adapter ports may be shared or used in a LACP configuration**
- **Best Practice**
  - **Assign 100% capacity to each SR-IOV logical port in the Link Aggregation Group to prevent accidental assignment of another SR-IOV logical port to the same physical port**

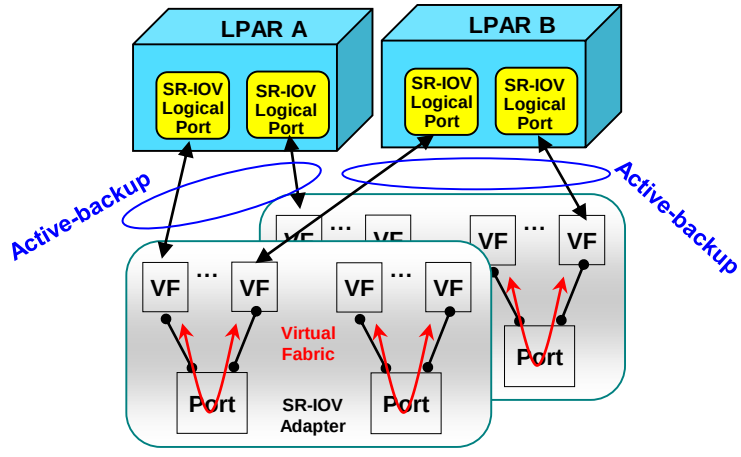


Link aggregation with one logical port assigned to each physical port.

Network Port Aggregation Recommendations

If you require bandwidth less than a single link's bandwidth and failover?

- **Use Active-Backup approach (e.g. AIX NIB, IBM i VIPA, or Linux bonding driver active-backup)**
- **Allows sharing of the physical port by multiple partitions**
- **When an SR-IOV logical port is configured in an active-backup configuration, it must be configured with the capability to detect when to failover from the primary to the backup adapter.**
  - **For AIX, configure network interface backup with IP address to ping.**
  - **For IBM i with VIPA, options include Routing Information Protocol (RIP), Open Shortest Path First (OSPF) or customer monitor script.**
  - **For Linux, use the bonding support to configure monitoring to detect network failures.**



Active backup configuration (i.e. no switch configuration required) allows sharing of physical port

SR-IOV HW/SW minimums POWER8 (GA June/2015)

- IBM Power System E870 (9119-MME), IBM Power System E880 (9119-MHE), IBM Power System E850 (8408-E8E)
  IBM Power System S824 (8286-42A), IBM Power System S814(8286-41A), IBM Power System S822(8284-22A), IBM Power System S824L(8247-42L), IBM Power System S822L (8247-22L), IBM Power System S812L(8247-21L)
- Added SR-IOV support for PCIe Gen3 I/O expansion drawer (2 SR-IOV slots per Fan-out Module)
- HMC required for SR-IOV
- Server firmware 830
- PowerVM standard or enterprise edition
  - PowerVM express edition allows only one partition to use the SR-IOV logical ports per adapter
- Minimum client operation systems:
  - AIX 6.1 TL9 SP5 and APAR IV68443, or later
  - AIX 7.1 TL3 SP5 and APAR IV68444, or later
  - IBM i 7.1 TR10, or later
  - IBM i 7.2 TR2, or later
  - Red Hat Enterprise Linux 6.5, or later
  - Red Hat Enterprise Linux 7, or later
  - SUSE Linux Enterprise Server 11 SP3, or later
  - SUSE Linux Enterprise Server 12, or later
  - Ubuntu 15.04, or later
  - SR-IOV logical ports assigned to the VIOS requires VIOS 2.2.3.51, or later

Thank you

Thank you

Fin - résumé des options

Besoin en bande passante		
High: > 10 G	SR-IOV ou adapter dédié en LACP	Pas de support LPM
Médium (5Gbits/s)	SR-IOV avec vNIC	Attendre support du VIOS failover (fin 2016)
Low	SEA ou SR-IOV ou …	

BACKUP SLIDES

PowerVM Single Root I/O Virtualization
Fundamentals, Design and Configuration

SR-IOV CONFIGURATION

SR-IOV configuration checklist

- **Managed Systems needs PCIe adapter(s) with SR-IOV support.**
- **Each SR-IOV adapter needs to be placed into a SR-IOV capable PCIe slot.**
  (Review adapter placement guide in the IBM Knowledge Center)
- **Supported level of System Firmware, HMC and Operating Systems**
  http://www-01.ibm.com/common/ssi/printableversion.wss?docURL=/common/ssi/rep_ca/2/899/ENUSLG15-0062/index.html
- **Configure adapter(s) into SR-IOV shared mode on HMC**
  Classic GUI: systems management -> server -> properties -> select I/O tab
  New GUI: Select Managed System -> Hardware Virtualized I/O
  Cli: (**`chhwres -r sriov`**)
- **Configure Logical Ports**
  During partition creation
  DLPAR: Dynamic partitioning → SR-IOV Logical Ports
  In Partition Profile

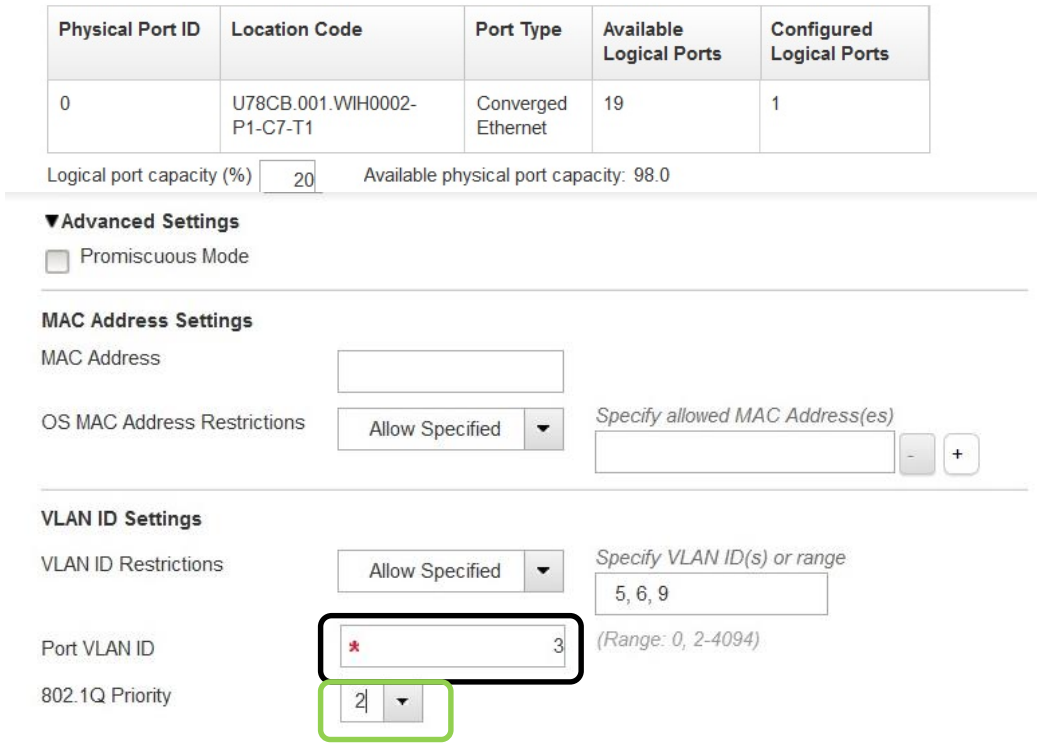
Logical Port Properties – Advanced Settings

Port VLAN ID – Set a non zero PVID to have the adapter add a VLAN tag with this VLAN to all untagged transmit packets and strip the VLAN tag from receive packets with this VLAN.

- Received packets that have a match for this VLAN id will be received by the OS as untagged packets

Port VLAN ID (PVID) Priority – A value between 0-7 can be set for the PVID priority. This value only applies if the PVID is set to a non zero value

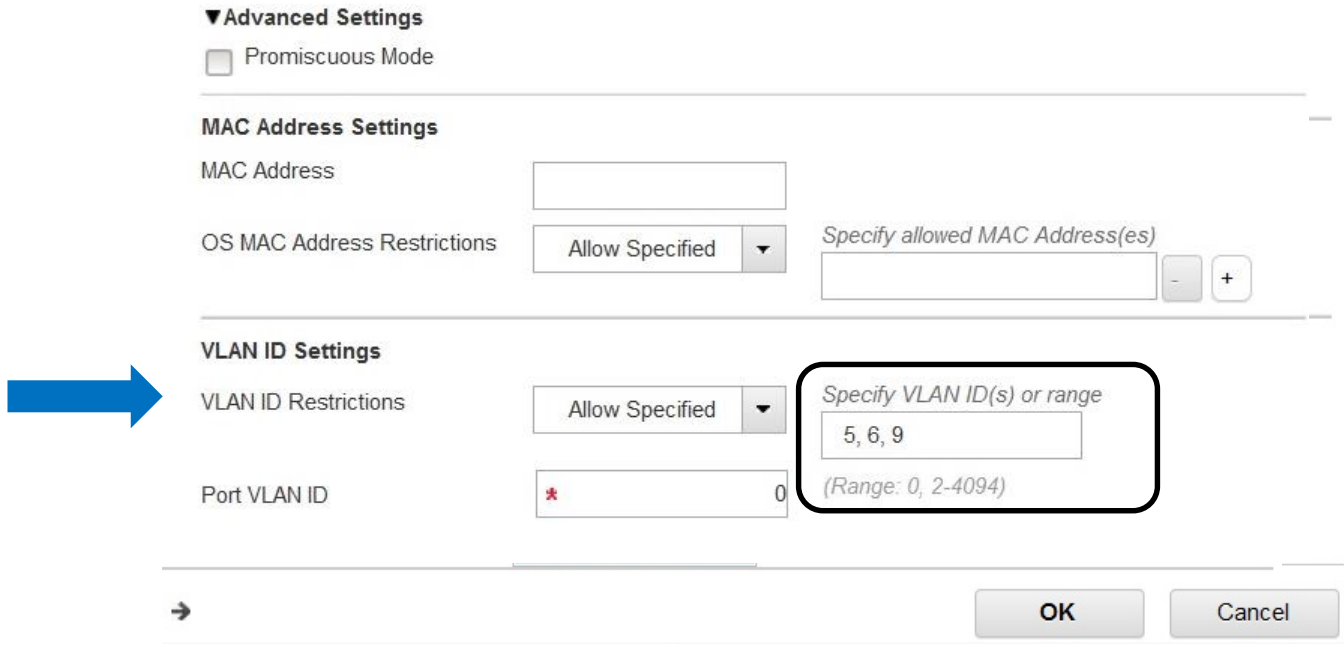
Click OK when you are done configuring all the logical port settings



Logical Port Properties – Advanced Settings

VLAN restrictions – Allows for restrictions on VLANs that the logical port device driver can use.

- If Promiscuous is selected neither VLANs or MAC addresses can be restricted
- Allow All– No restrictions on which VLANs can be used. This option can only be used if there are also no restrictions on the OS MAC Addresses
- Deny All– OS can not configure a VLAN ID. The OS will only receive packets that are untagged
- Allow Specified – Set a list of VLAN IDs that the OS is allowed to use for the logical port
- VLANs still need to be configured in the OS



HMC CLI: Configure Adapter into shared mode

```
$ lshwres -m p8-E870-9119-MME-SNXXXXXXX -r sriov
 --rsubtype adapter
adapter_id=null,slot_id=21040124,adapter_max_logical_ports=null,
 config_state=dedicated,functional_state=1,logical_ports=null,
 phys_loc=U78CD.001.FZH0469-P1-C4,phys_ports=null,
 sriov_status=null,alternate_config=0

$ chhwres -r sriov -m p8-E870-9119-MME-SNXXXXXXX \
 -rsubtype  adapter -o a -a slot_id=21040124

$ lshwres -m p8-E870-9119-MME-SNXXXXXXX -r sriov \
 --rsubtype adapter
adapter_id=1,slot_id=21010101,adapter_max_logical_ports=48,
 config_state=sriov,functional_state=1,logical_ports=48,
 phys_loc=U78CD.001.FZH0860-P1-C1,
 phys_ports=4,sriov_status=running,alternate_config=0
```

PowerVM Single Root I/O Virtualization
Fundamentals, Design and Configuration

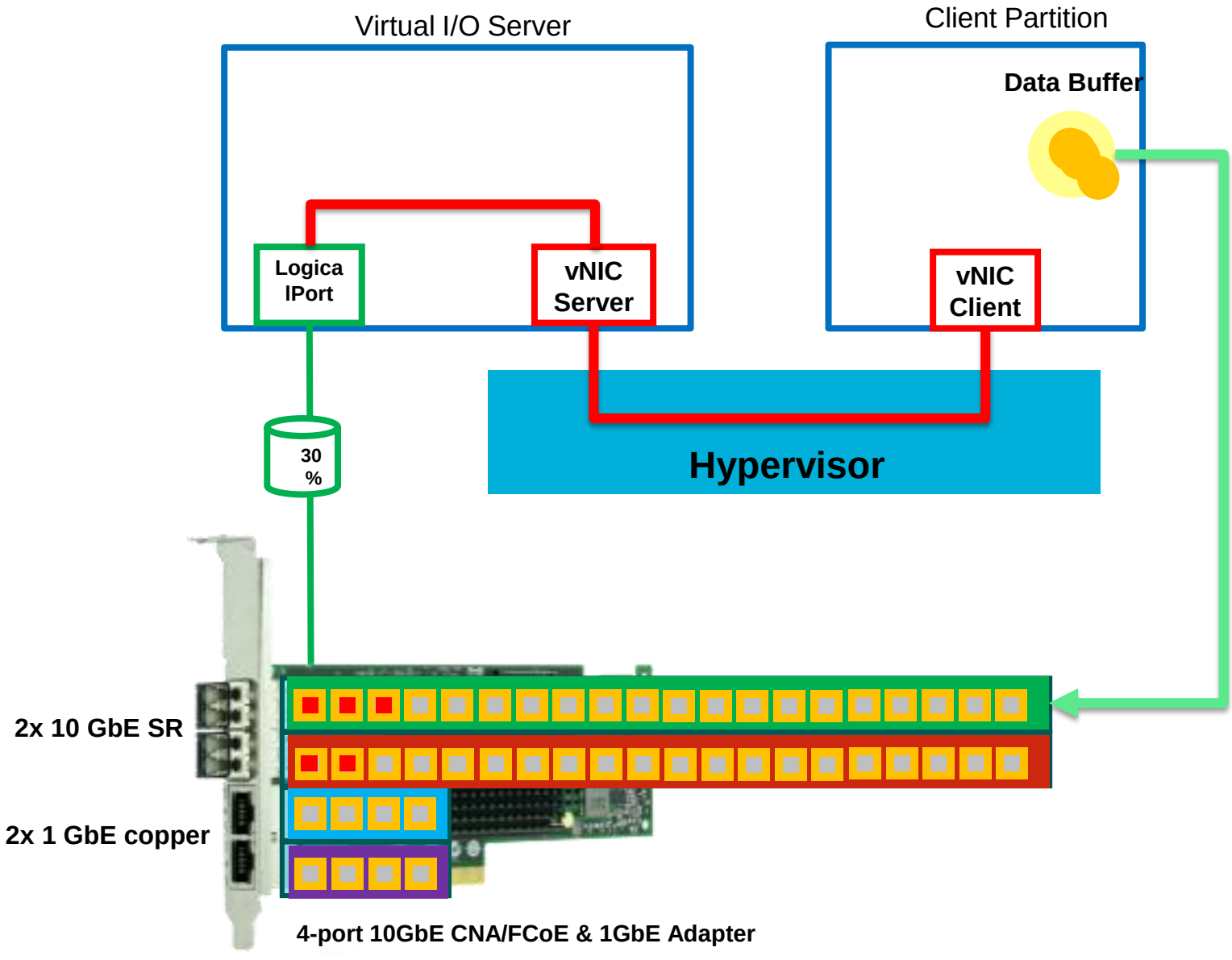
VIRTUAL NETWORK INTERFACE CONTROLLER (VNIC) - NEW

Live Partition Mobility Options with SR-IOV

Virtual Network Interface Controller (vNIC)

- Becomes available in December 2015
- vNIC is new virtual adapter type
- vNIC leverages SR-IOV to provide a performance optimized virtual NIC solution
- vNIC enables advanced virtualization features such as live partition mobility with SR-IOV adapter sharing
- Leverages SR-IOV Capacity value (QoS)
- Announced October 2015 for AIX & IBM i
  - Linux not announced
  - E850 support not announced
- Pre-req
  - AIX 7.1 TL4 or later or AIX 7.2 or later
  - IBM i 7.1 TR10 or later or 7.2 TR3 or later
  - VIOS 2.2.4, or later
  - FW 840 or later
  - HMC V8R8.4.0 or later

vNIC Architecture



PowerVM Single Root I/O Virtualization
Fundamentals, Design and Configuration

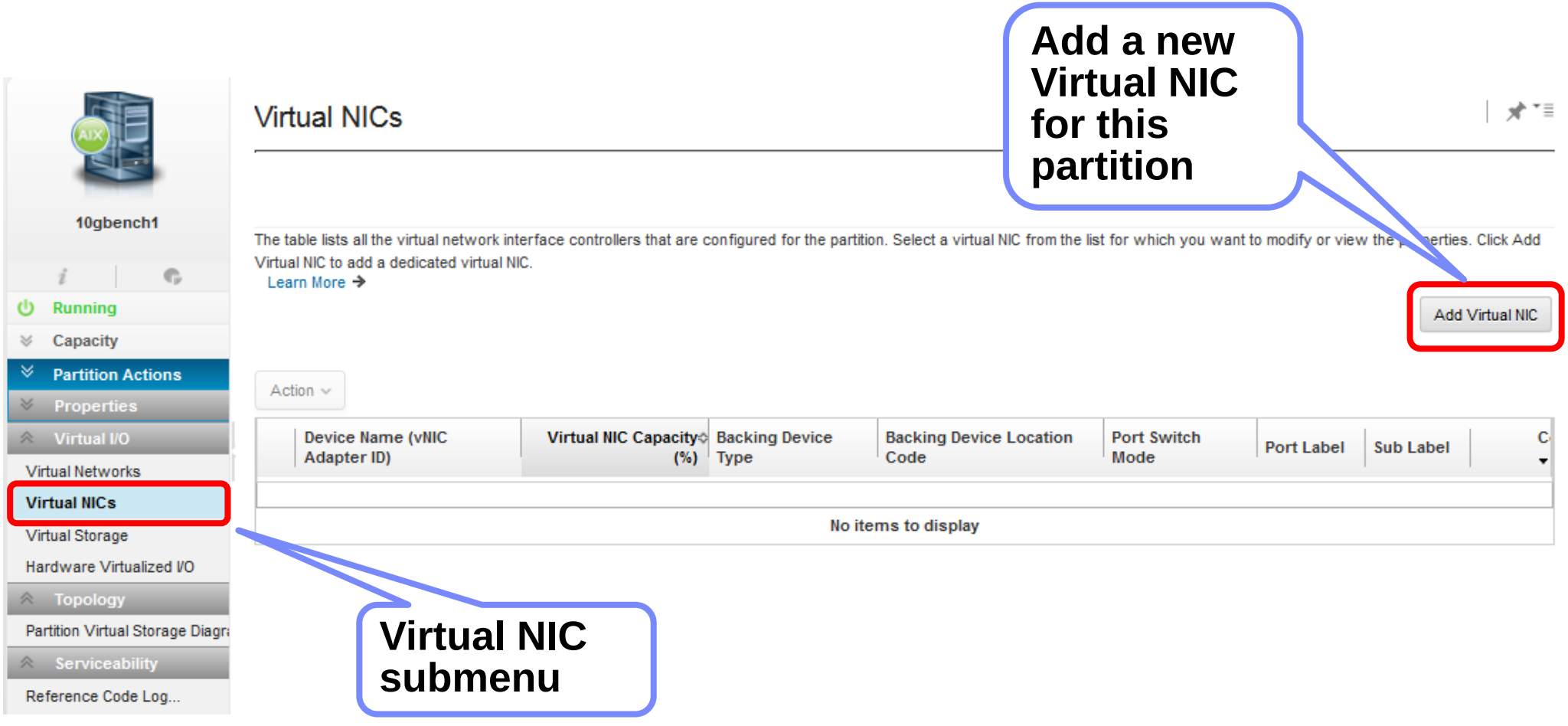
VNIC CONFIGURATION

Add virtual network interface controller

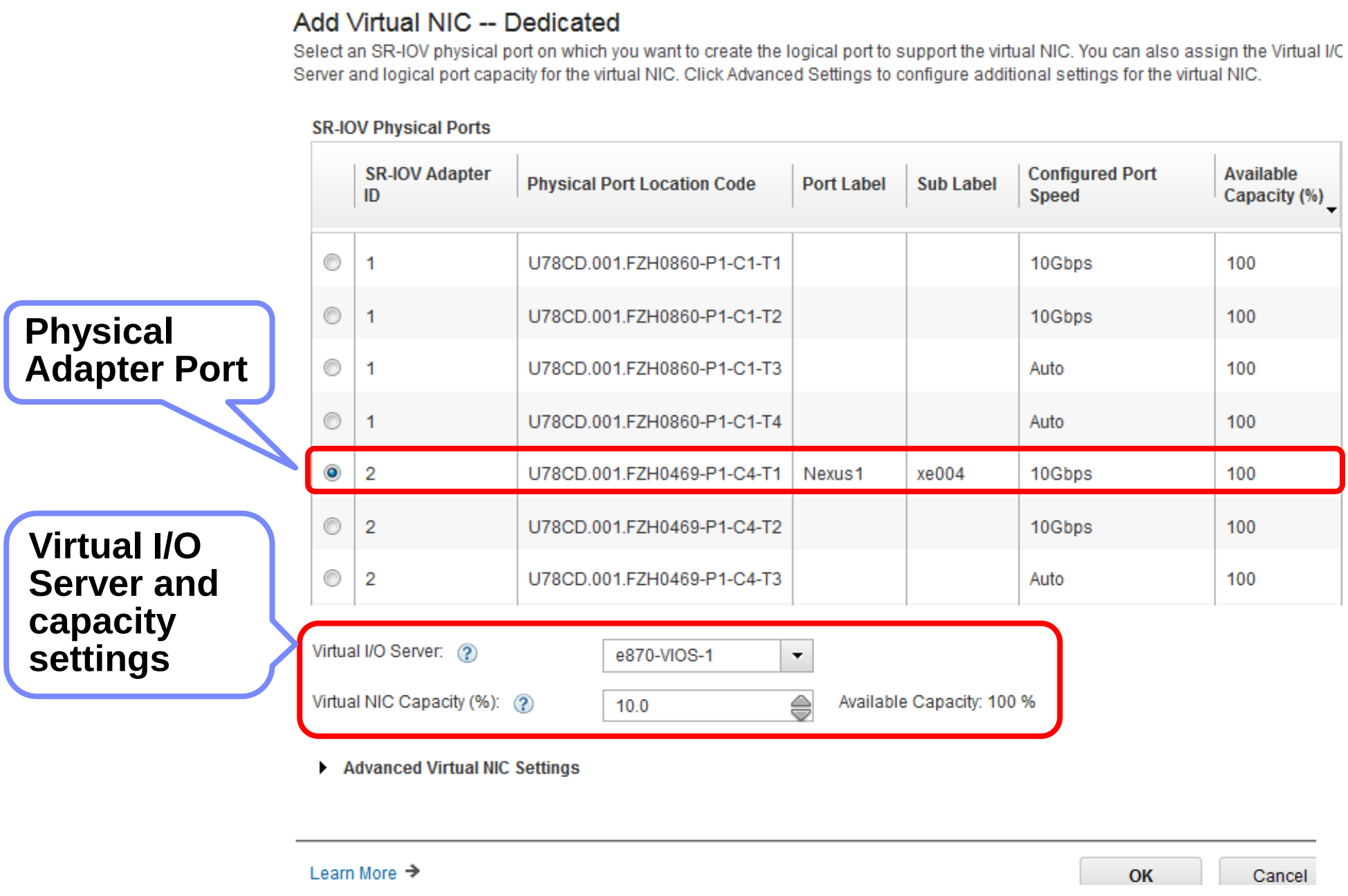
- Exclusively supported on **new HMC GUI** and **CLI**
- Prerequisites for a **running** client partition:
  - The Virtual I/O Server (VIOS) that hosts the virtual NIC is running with an active Resource Monitoring and Control (RMC) connection.
  - The client partition has an active RMC connection.
- Prerequisites for a **inactive** client partition:
  - The Virtual I/O Server (VIOS) that hosts the virtual NIC is running with an active RMC connection or is shutdown.
- With a new virtual NIC adapter, you can specify the following settings
  - The Virtual I/O Server that hosts the virtual NIC
  - SR-IOV physical port on a running SR-IOV adapter in the shared mode
  - Virtual NIC capacity
  - Default VLAN ID (Advanced settings)
  - Tagged VLANs allowed (Advanced settings)
  - MAC addresses allowed (Advanced settings)

Add virtual network interface controller

- From the LPAR context, click on Virtual I/O -> Virtual NICs
- Select Add Virtual NIC to create a new adapter



Add virtual network interface controller



Add virtual network interface controller Advanced Settings

Virtual client slot ID can be specified

Virtual NIC Adapter ID: Next available

MAC address restrictions can be configured

MAC Address Settings

MAC Address: HMC-assigned

OS MAC Address Restrictions: Allow all

Default VLAN ID and VLAN filtering can be configured

VLAN ID Settings

VLAN ID Restrictions: Allow all

Range: 0, 2 - 4,094

Port VLAN ID: 0

Virtual network interface controller

Virtual I/O Server:

```
padmin@vios1:$ lsmap -all -vnic
Name          Physloc                          ClntID ClntName   ClntOS
------------- -------------------------------- ------ ---------- -------
vnicserver0   U9119.MME.0647C9R-V1-C32897         32 10gbench1      AIX

Backing device:ent19
Status:Available
Physloc:U78CD.001.FZH0469-P1-C4-T1-S1
Client device name:ent3
Client device physloc:U9119.MME.0647C9R-V32-C6
```

Client Partition perspective (AIX):

```
root@10gbench1:/> lsdev | grep ent
ent0       Available            Virtual I/O Ethernet Adapter (l-lan)
ent1       Available            Virtual I/O Ethernet Adapter (l-lan)
ent2       Available            Virtual I/O Ethernet Adapter (l-lan)
ent3       Available            Virtual NIC Client Adapter (vnic)
ent4       Available            Virtual NIC Client Adapter (vnic)
vscsi0     Available            Virtual SCSI Client Adapter
vscsi1     Available            Virtual SCSI Client Adapter
```

PowerVM Single Root I/O Virtualization
Fundamentals, Design and Configuration

FAULT-TOLERANT SETUP WITH SR-IOV AND VNIC

SR-IOV fault-tolerant configuration

- **OS takes care of network failure detection and recovery**
  - Network Interface Backup (NIB) for AIX
  - Virtual IP Address (VIPA) for IBM i
  - Active-Backup Bonding for Linux

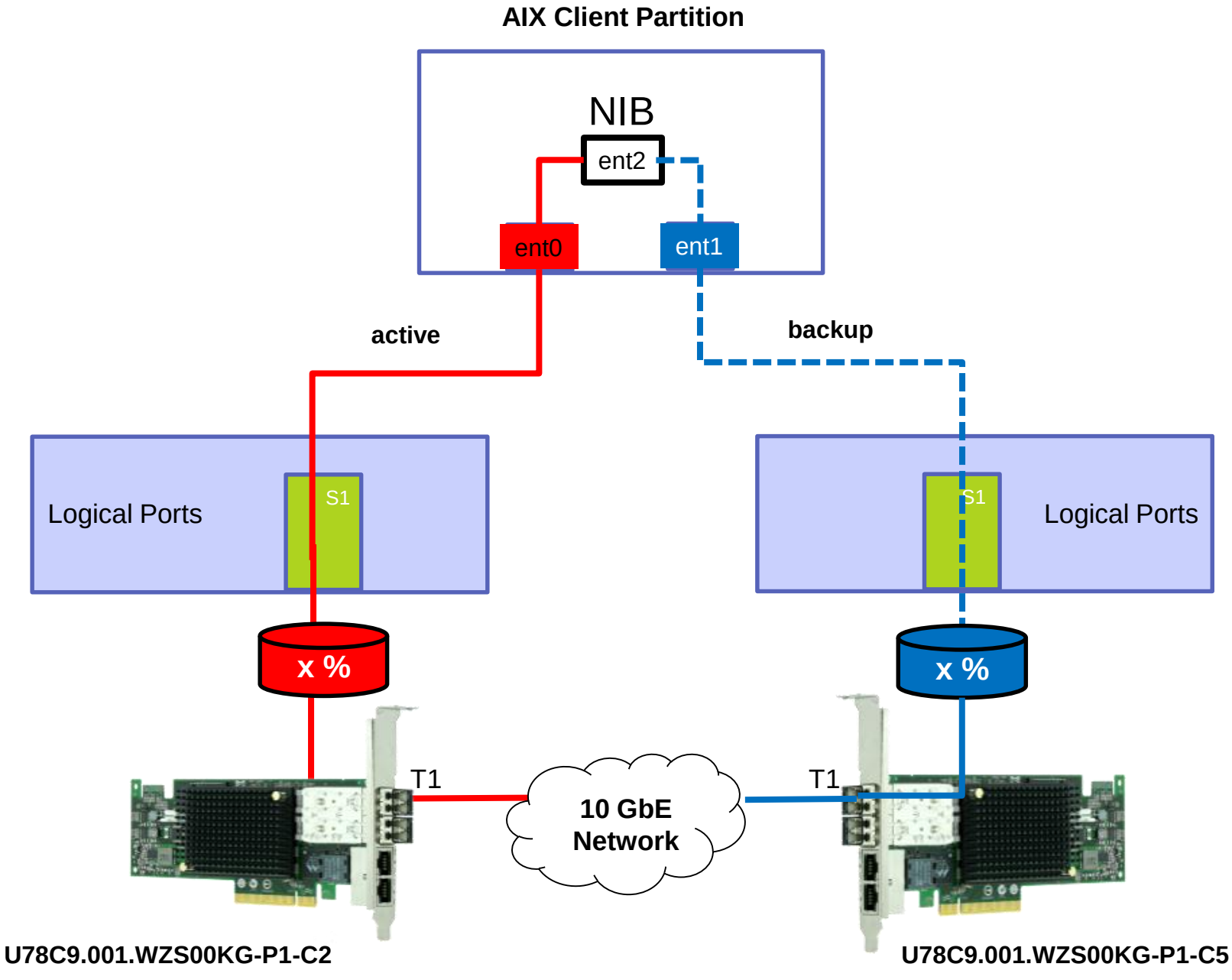
SR-IOV (active) / SR-IOV (backup)

- Two SR-IOV logical ports per network connection
- Same performance, even in failure conditions.
- Load distribution is controlled on a client level, by choosing the active/backup adapter role.
- Load distribution can be modified dynamically by switching the active/backup adapter role.

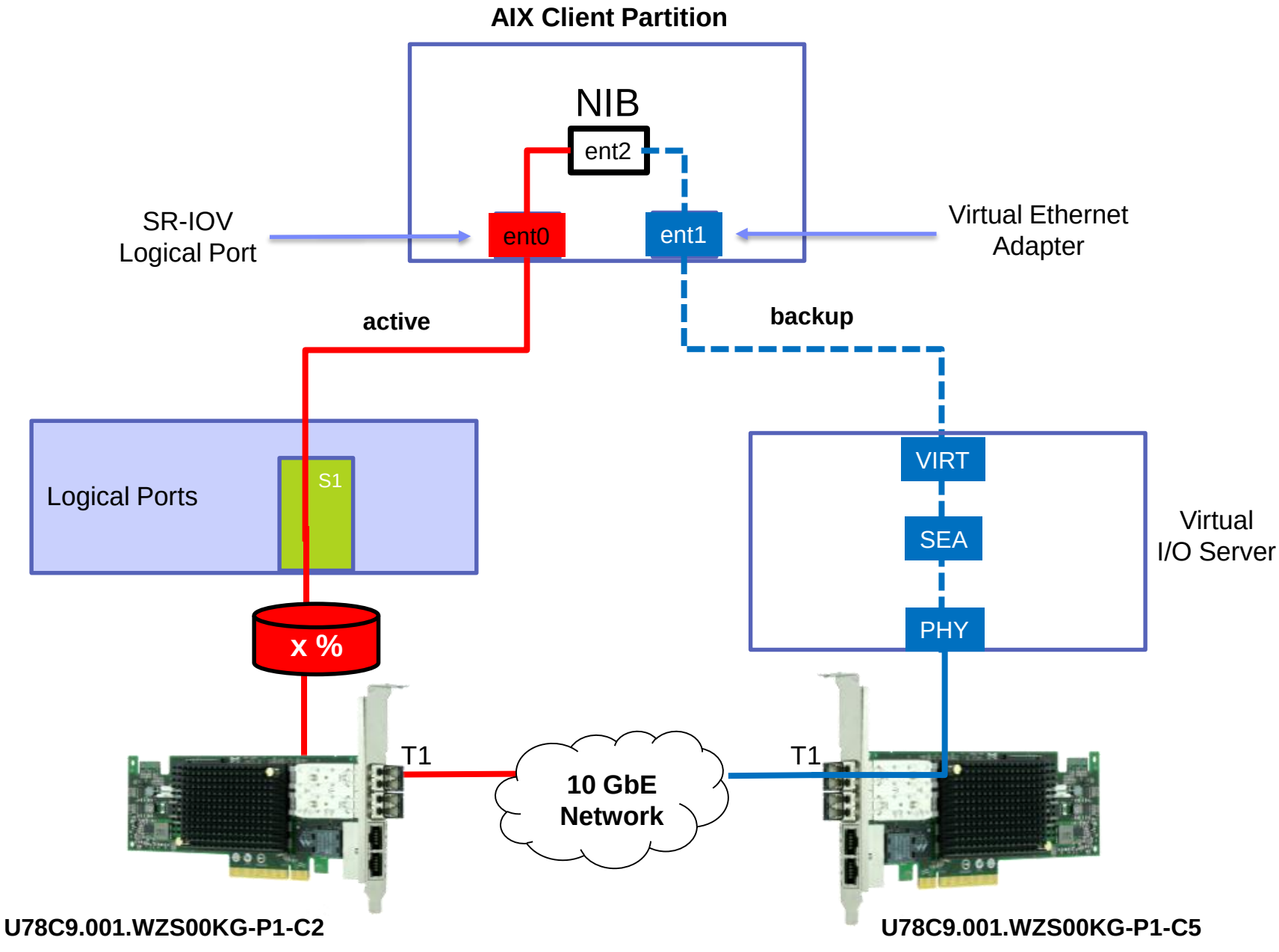
SR-IOV (active) / SEA (backup)

- One SR-IOV logical port plus additional vitual ethernet adapter per network connection.
- Possibility of using non-SR-IOV capable network adapters for backup.
- SR-IOV logical port should be the preferred active path for optimal performance.
- Potential performance decrease in a failure condition (switching to Virtual Ethernet Adapter).

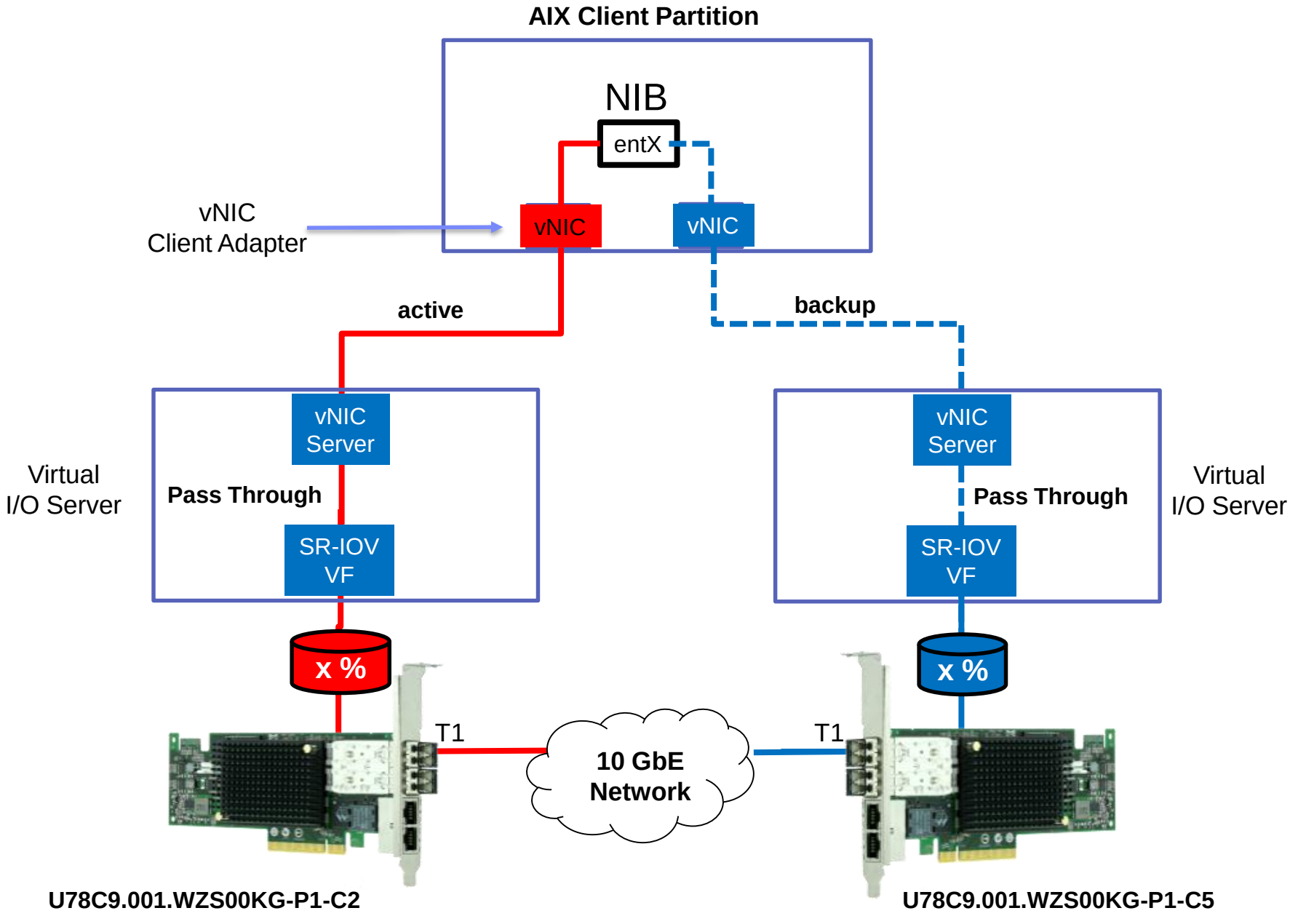
SR-IOV Adapter Redundancy – Example #1



SR-IOV Adapter Redundancy – Example #2



vNIC and Adapter Redundancy – Example



PowerVM Single Root I/O Virtualization
Fundamentals, Design and Configuration

MAINTENANCE AND MONITORING

Physical Adapter Replacement

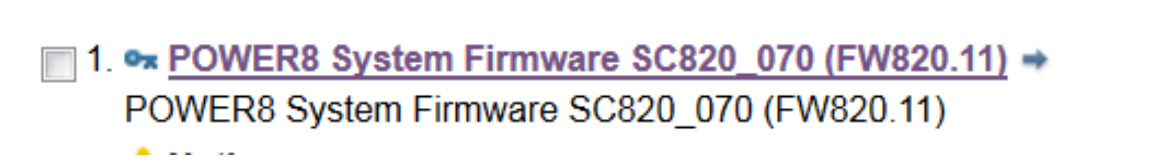
- SR-IOV adapters can be added, removed, and replaced without disrupting the system or shutting down the partitions.
- For adapter replacement, all the logical ports must be deconfigured.
- The HMC provides a GUI for adapter concurrent maintenance operations. (**Serviceability → Hardware → MES Tasks → Exchange FRU)**
- New adapter must have the same capabilities (same type/feature code).
- When new adapter is plugged into the same slot as the original adapter, the hypervisor will automatically associate the old adapter's configuration with the new adapter.
- If the new adapter is plugged in to a different slot, the **`chhwres`** command is needed to associate the original adapter configuration with the new adapter.

```
$ chhwres -m Server1 -r sriov –rsubtype adapter -o m -a \
"slot_id=2101020b,target_slot_id=21010208"
```

SR-IOV Adapter Firmware Upgrade

- There are 2 pieces of firmware for SR-IOV that are built into system firmware
  - Adapter driver firmware – The driver code that configures the adapter and logical ports
  - Adapter firmware – The firmware that runs on the adapter

- Both levels of firmware are automatically updated to the levels included in the active system firmware in the following cases
  - System boot/reboot
  - Adapter transitioned into SR-IOV mode
  - Adapter level concurrent maintenance

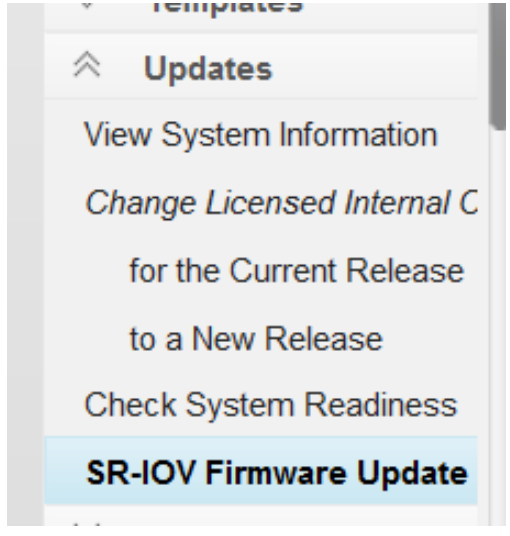


SR-IOV Adapter Firmware Upgrade

- When system firmware is updated concurrently, the SR-IOV levels on currently configured SR-IOV adapters are not automatically updated
  - Updating the SR-IOV levels will cause a temporary network outage on the logical ports on the affected adapter

- Starting with system firmware level FW830, the SR-IOV firmware levels can be viewed and updated using the HMC GUI

- On the HMC enhanced+ GUI select the Server -> Actions -> SR-IOV Firmware Update

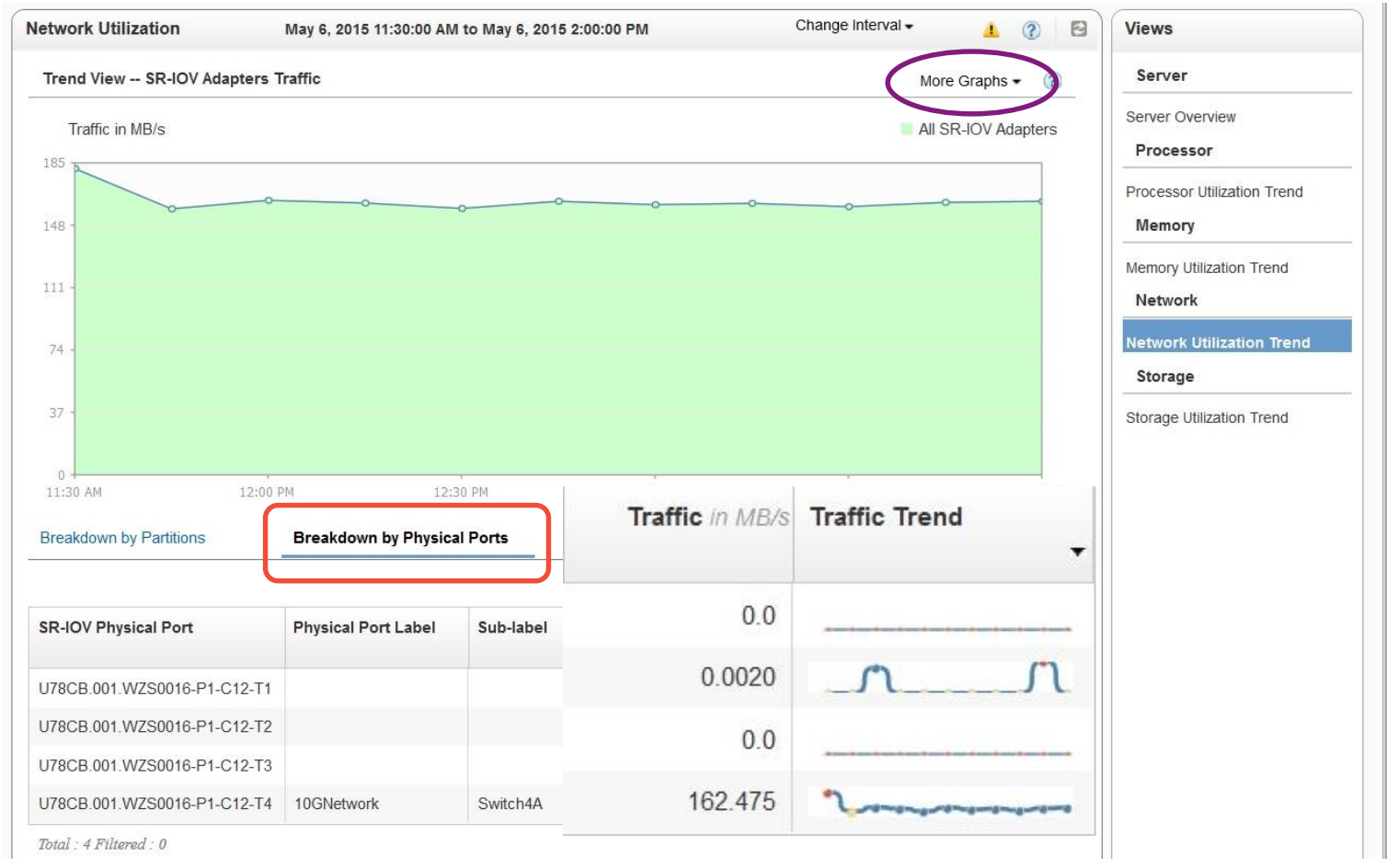


SR-IOV promiscuous logical port mode

- A promiscuous logical port receives all unicast traffic.
- The destination MAC address does not necessarily has to match the logical port's address.
- Number of promiscuous logical ports per physical port is limited to one to minimize potential performance impact.
- Limitation is valid for all configured, active or shutdown partitions.
- The management console indicates the number of logical ports on the physical ports that are allowed to have a promiscuous permission setting.

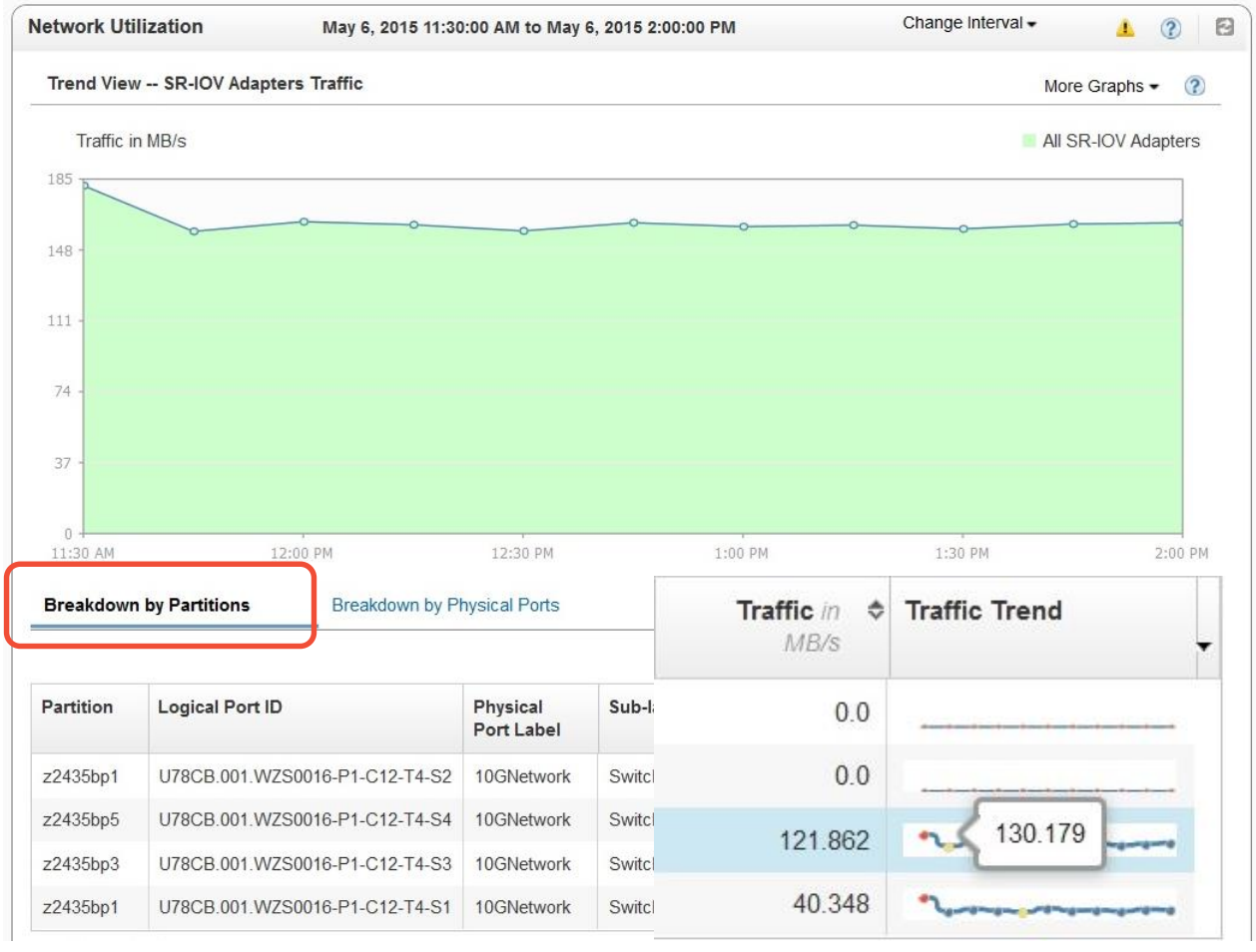
Performance Monitor for SR-IOV

- From the performance monitor screen click Network Utilization Trend -> More Graphs -> SR-IOV adapters
- The breakdown by physical ports shows how heavily utilized a physical port is and can be used to determine whether or not there is additional bandwidth available



Performance Monitor for SR-IOV

- The breakdown by partitions shows each logical port individually and which LPAR owns it
- This can be used to determine which logical ports are using the physical ports bandwidth



Additional Ressources

- **Redpaper:**

IBM Power Systems SR-IOV: Technical Overview and Introduction

- **10 Gigabit Ethernet Performance for IBM Power Systems**

AIX Virtual User Group (VUG) USA
Session replay: http://www.youtube.com/watch?v=QlNjcO_B1PI
Presentation: www.tinyurl.com/ibmaixvug

- **SRIOV - Allyn Walsh & Steve Nasypany**

AIX Virtual User Group (VUG) USA
Session replay: http://youtu.be/65wyBrr2Vrc
Presentation: www.tinyurl.com/ibmaixvug

- **April 2015 Announcement**

http://www-01.ibm.com/common/ssi/rep_ca/1/897/ENUS115-021/ENUS115-021.PDF

- **SR-IOV specification**

http://www.pcisig.com/specifications/iov/