Ethernet on POWER: Physical, Shared, Virtual

Steven Knudson sjknuds@us.ibm.com
IBM POWER Strategic Initiatives
20190702

Agenda

- Physical Ethernet Adapters
- Jumbo Frames
- Link Aggregation Configuration
- Shared Ethernet Adapter SEA Configuration
- VIO 2.2.3, Simplified SEA Configuration
- SEA VLAN Tagging
- VLAN awareness in SMS
- 10 Gb SEA, active – active
- ha_mode=sharing, active – active
- Dynamic VLANs on SEA
- SEA Throughput
- Virtual Switch – VEB versus VEPA mode
- AIX Virtual Ethernet adapter
- AIX IP interface
- AIX TCP settings
- AIX NFS settings
- largesend, large_receive with binary ftp for network performance
- iperf tool for network performance

Most syntax in this presentation is VIO padmin, sometimes root smitty

Physical Ethernet Adapters

- **Lets use Flow Control**
- **The 10Gb PCIe Ethernet-SR adapter uses 802.3x or “Link” Flow Control**
- **The FCoE adapter uses 802.1Qbb or Priority Flow Control. PFC requires VLAN tagging to be on (802.1q)**
- **PCIe Adapter Flow Control attribute is on by default**

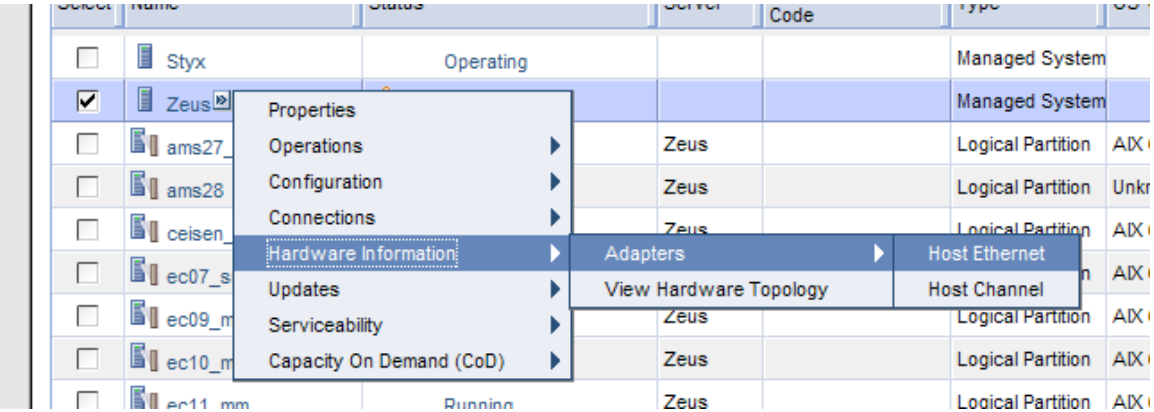
```
$ lsdev -dev ent0 -attr | grep flow
flow_ctrl      yes              Enable Transmit and Receive Flow Control
```

- **Attribute might still be disabled by switch – check status, in this case, SEA over a six link aggregation; grep syntax may vary here**

```
$ entstat -all ent14 | grep "Transmit and Receive Flow Control Status:"
Transmit and Receive Flow Control Status: Disabled
Transmit and Receive Flow Control Status: Disabled
Transmit and Receive Flow Control Status: Disabled
Transmit and Receive Flow Control Status: Disabled
Transmit and Receive Flow Control Status: Disabled
Transmit and Receive Flow Control Status: Disabled
```

Physical Ethernet Adapters

- **IVE Physical port Flow Control (802.3x, or Link) is off by default – set via HMC…**
- **These are very old now; only POWER6, and few very early POWER7**



Physical Ethernet Adapters

- **IVE - Radio Button, then Configure…**

Host Ethernet Adapters : Zeus - Mozilla Firefox: IBM Edition

ibm.com https://sahmc1.dfw.ibm.com/hmc/wcl/T11dc#tableTop_732b732b

Host Ethernet Adapters : Zeus

Select a physical port in the table below to display the port's current partition usage.

Current Status

Select	Physical Port Location Codes	Port ID	Port Type	Port Group ID	Port Group MCS Value	Connection State	Speed	Duplex
○	U78C0.001.DBJ0426-P2 - C8-T2	0	1 G	1	1	up	1 Gbps	full
◉	U78C0.001.DBJ0426-P2 - C8-T1	0	1 G	2	4	up	1 Gbps	full
○	U78C0.001.DBJ0453-P2 - C8-T4	0	1 G	1	4	down	Auto	full
○	U78C0.001.DBJ0453-P2 - C8-T3	0	1 G	2	4	down	Auto	full
○	U78C0.001.DBJ0453-P2 - C8-T2	0	1 G	1	4	up	1 Gbps	full
○	U78C0.001.DBJ0453-P2 - C8-T1	0	1 G	2	4	up	1 Gbps	full
○	U78C0.001.DBJ0426-P2 - C8-T4	0	1 G	1	4	down	Auto	full
○	U78C0.001.DBJ0426-P2 - C8-T3	0	1 G	2	4	down	Auto	full

Configure...

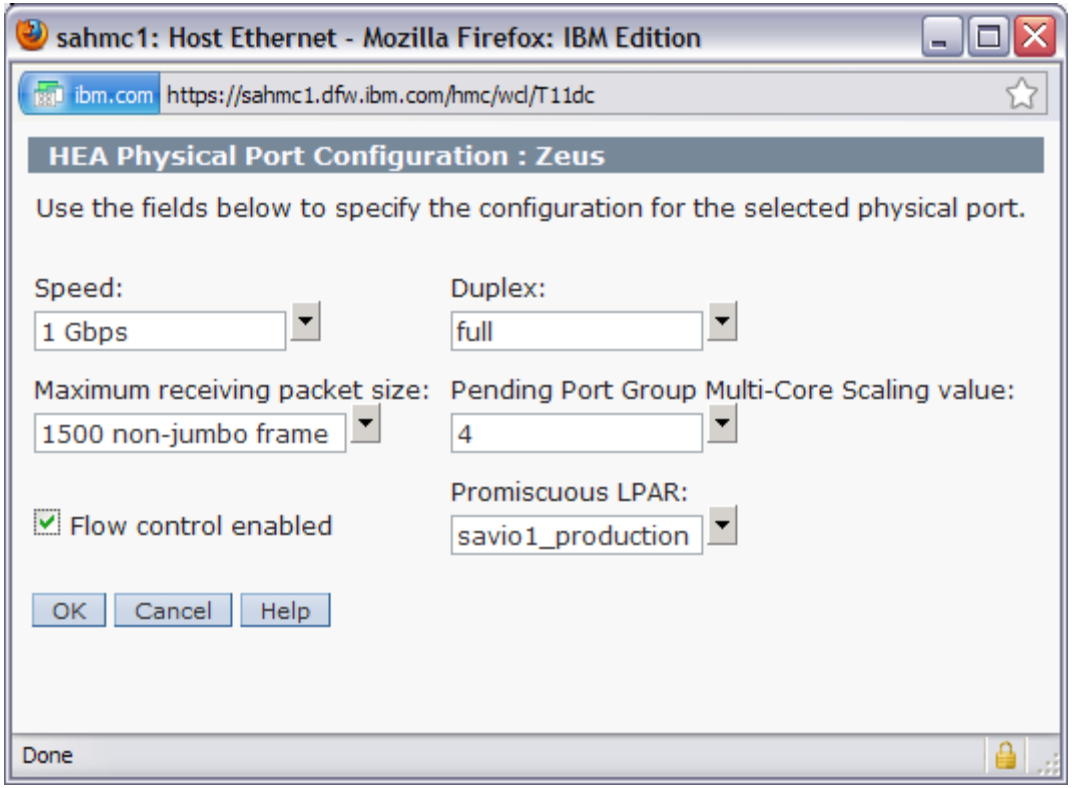
Logical Partition Usage

Logical Partition	Logical Port ID	Logical Port DRC Name	Logical Port burned-in MAC / user-defined MAC	Capability	Allowe
savio1_production	1	Port 17	00215EEB0ED0/000000000000	Base Minimum	Allow a

OK Cancel Help

Physical Ethernet Adapters

- **IVE – HEA Flow control checkbox, Promiscuous LPAR when VIO SEA will be built on this adapter**



Physical Ethernet Adapters

- **What Ethernet adapters do we have?**

```
$ lsdev -type adapter | grep ent
ent0             Available   Logical Host Ethernet Port (lp-hea)
ent1             Available   Virtual I/O Ethernet Adapter (l-lan)
ent2             Available   Virtual I/O Ethernet Adapter (l-lan)
ent3             Available   Virtual I/O Ethernet Adapter (l-lan)
ent4             Available   Shared Ethernet Adapter
```

- **What are their physical location codes?**

```
$ lsdev -type adapter -field name physloc | grep ent
ent0             U78C0.001.DBJ4725-P2-C8-T1
ent1             U9179.MHB.1026D1P-V1-C2-T1
ent2             U9179.MHB.1026D1P-V1-C3-T1
ent3             U9179.MHB.1026D1P-V1-C4-T1
ent4
```

Physical Ethernet Adapters

- **You can add VLAN tags to physical Ethernet adapters, since AIX 5.1**

```
# lsdev -Cc adapter | grep ent
ent0     Available       Logical Host Ethernet Port (lp-hea)

# smitty vlan → Add a VLAN
                                VLAN

Move cursor to desired item and press Enter.

  List All VLANs
  Add A VLAN
  Change / Show Characteristics of a VLAN
  Remove A VLAN
```

```
     +-------------------------------------------------------------------------+
     |                        Available Network Adapters                       |
     |                                                                         |
     | Move cursor to desired item and press Enter.                            |
     |                                                                         |
     |   ent0 Available  Logical Host Ethernet Port (lp-hea)                   |
     |                                                                         |
     | F1=Help                 F2=Refresh              F3=Cancel               |
     | F8=Image                F10=Exit                Enter=Do                |
F1=He| /=Find                  n=Find Next                                     |
F9=Sh+-------------------------------------------------------------------------+
```

Physical Ethernet Adapters

- ```
  Add A VLAN

  Type or select values in entry fields.
  Press Enter AFTER making all desired changes.

                                                [Entry Fields]
    VLAN Base Adapter                          ent0
  * VLAN Tag ID                                [288]                    +#
    VLAN Priority                              [3]                      +#
  ```

- VLAN Tag ID, from 2...4094
- VLAN Priority, from 1 2 0 (default best effort) 3 4 5 6 7
- Inserts VLAN Priority value in the tag header for examination by intervening switches and routers
- Enter, OK, F10 to exit smitty

- ```
  # lsdev -Cc adapter | grep ent
  ent0     Available       Logical Host Ethernet Port (lp-hea)
  ent1     Available       VLAN
  ```

- ```
  # lsattr -El ent1
  base_adapter  ent0 VLAN Base Adapter True
  vlan_priority 3    VLAN Priority     True
  vlan_tag_id   288  VLAN Tag ID       True
  ```
  ```

Physical Ethernet Adapters

- **Physical adapters should have large_send (and those that have large_receive) already set to yes**

```
$ lsdev -dev ent0 -attr |grep lar
large_receive yes              Enable receive TCP segment aggregation    True
large_send    yes              Enable hardware Transmit TCP segmentation
```

- **There is no media_speed attribute on 10Gb adapters. 1Gb adapters are usually fine with Auto_Negotiation**

```
$ lsdev -dev ent0 -attr | grep media
media_speed   Auto_Negotiation Requested media speed
```

Physical Ethernet – Jumbo Frames

- Jumbo frames is a physical setting. It is set
  - On Ethernet switch ports
  - On physical adapters
  - On the link aggregation, if used
  - On the Shared Ethernet Adapter

- Jumbo frames is **NOT** set on the virtual adapter or interface in the AIX client LPAR.

- **Do not change MTU on the AIX client LPAR interface**. We will use mtu_bypass (largesend) in AIX

- mtu_bypass – up to 64KB segments sent from AIX to SEA, resegmentation on the SEA for the physical network (1500, or 9000 as appropriate)

Link Aggregation Configuration

- **smitty etherchannel → Add An EtherChannel / Link Aggregation**

```
Telnet fahr.dfw.ibm.com [9.19.51.115]
File  Edit  Settings  Connection  Help
                    EtherChannel / IEEE 802.3ad Link Aggregation

Move cursor to desired item and press Enter.

  List All EtherChannels / Link Aggregations
  Add An EtherChannel / Link Aggregation
  Change / Show Characteristics of an EtherChannel / Link Aggregation
  Remove An EtherChannel / Link Aggregation
  Force A Failover In An EtherChannel / Link Aggregation
   ------------------------------------------------------------------------
   |                        Available Network Interfaces                  |
   |                                                                      |
   | Move cursor to desired item and press F7.                            |
   |     ONE OR MORE items can be selected.                               |
   | Press Enter AFTER making all selections.                             |
   |                                                                      |
   | > ent0                                                               |
   | > ent1                                                               |
   |   ent2                                                               |
   |                                                                      |
   | Esc+1=Help             Esc+2=Refresh          Esc+3=Cancel           |
   | F7=Select              F8=Image               F10=Exit               |
Es | Enter=Do               /=Find                 n=Find Next            |
F9 ------------------------------------------------------------------------
Connected To  fahr.dfw.ibm.com [9.19.51.115]   VT320   NUM   18, 6   11:33
```

Link Aggregation Configuration

Telnet fahr.dfw.ibm.com [9.19.51.115]

File Edit Settings Connection Help

```
                    Add An EtherChannel / Link Aggregation

Type or select values in entry fields.
Press Enter AFTER making all desired changes.

                                                        [Entry Field
  EtherChannel / Link Aggregation Adapters              ent0,ent1
  Enable Alternate Address                              no                +
  Alternate Address                                    []
  Enable Gigabit Ethernet Jumbo Frames                  no
  Mode                                                  8023ad
  IEEE 802.3ad Interval                                 long
  Hash Mode                                             src_dst_port
  Backup Adapter                                        ent               +
       Automatically Recover to Main Channel            yes               +
       Perform Lossless Failover After Ping Failure     yes
  Internet Address to Ping                             []
  Number of Retries                                    []
  Retry Timeout (sec)                                  []

Esc+1=Help          Esc+2=Refresh       Esc+3=Cancel        Esc+4=
Esc+5=Reset         F6=Command          F7=Edit             F8=
F9=Shell            F10=Exit            Enter=Do
```

Connected To fahr.dfw.ibm.com [9.19.51.115] VT320

Mode 8023ad when network admin configures LACP on switch ports

Use source and destination port numbers to hash over the links

Would NOT use backup adapter if configuring SEA failover

If you will configure SEA over this aggregation, do NOT configure Address to Ping

Link Aggregation Configuration

- **Mode – standard if network admin explicitly configures switch ports in a channel group for our server**
- **Mode – 8023ad if network admin configures LACP switch ports for our server. ad = Autodetect – if our server approaches switch with one adapter, switch sees one adapter. If our server approaches switch with a Link Aggregation, switch auto detects that.** For 10Gb, we should be LACP/8023ad**.**
- **Hash Mode – default is by IP address, good fan out for one server to many clients. But will transmit to a given IP peer on only one adapter**
- **Hash Mode – src_dst_port, uses source and destination port numbers in hash. Multiple connections between two peers likely hash over different adapters. Best opportunity for multiadapter bandwidth between two peers. Whichever mode used, we prefer hash_mode=src_dst_port**
- **Backup adapter – optional, standby, single adapter to same network on a different switch. Would not use this for link aggregations underneath SEA Failover configuration. Also would likely not use on a large switch, where active adapters are connected to different, isolated "halves" of a large "logical" switch.**
- **Address to ping – Not typically used. Aids detection for failover to backup adapter. Needs to be a reliable address, but perhaps not the default gateway. Do not use this on the Link Aggregation, if SEA will be built on top of it. Instead use netaddr attribute on SEA, and put VIO IP address on SEA interface.**
- **Using mode and hash_mode, AIX readily transmits on all adapters. You may find switch delivers receives on only adapter – switches must enable hash_mode setting as well.**

Link Aggregation Configuration

- **$ mkvdev –lnagg ent0,ent1 -attr mode=8023ad hash_mode=src_dst_port**
  ent8 available
  en8
  et8

- **There is no largesend, large_send attribute on a link aggregation**

Shared Ethernet Adapter SEA Configuration

- **Bridged virtual Ethernet adapters in VIO, before configuring SEA**
  # chdev -l ent0 -a dcbflush_local=yes
  ent0 changed

- **Create SEA**
- **If you are using netaddr "address to ping," you must have VIO IP on the SEA interface**
- **netaddr not typically needed**
- **With SEA, VIO local IP config is often on a "side" virtual adapter**

- **$ mkvdev -sea ent8 -vadapter entN -default entN -defaultid Y -attr ha_mode=auto ctl_chan=entK **
  netaddr=<reliable_ip_to_ping_outside_the_server> largesend=1 large_receive=yes
  ent10 available
  en10
  et10

- **You want largesend on the SEA, and mtu_bypass (largesend) on AIX LPAR ip interfaces. mtu_bypass on AIX ip interfaces boosts thruput LPAR to LPAR within the machine, with no additional cpu utilization. Along with that, largesend on the SEA will LOWER sending AIX LPAR cpu, and sending VIO cpu, when transferring to a peer outside the machine.**

Shared Ethernet Adapter SEA Failover switch port settings

- One vendor's suggestions on portfast, and bpdu-guard
  http://www.cisco.com/en/US/docs/switches/lan/catalyst4000/7.4/configuration/guide/stp_enha.html
- PortFast causes a switch or trunk port to enter the spanning tree forwarding state immediately, bypassing the listening and learning states. (Faster SEA Failover)
- Caution multiple times in the article - You can use PortFast to connect a single end station or a switch port to a switch port. If you enable PortFast on a port connected to another Layer 2 device, such as a switch, you might create network loops.
- Because PortFast can be enabled on nontrunking ports connecting two switches, spanning tree loops can occur because BPDUs are still being transmitted and received on those ports. (Remember, we have a virtual switch in our hypervisor)
- Console> (enable) set spantree portfast bpdu-guard 6/1 enable
- Bpdu-guard is not a panacea; it is disabled if you are VLAN tagging. When you are configuring SEA Failover, if you have any doubt about configuration, review it with Support Line to avoid BPDU storm.
- **Current generations of VIO (3Q 2016, VIO 2.2.4) have added capability to detect and prevent BPDU storm. This is not the hazard it used to be.**

Shared Ethernet Adapter SEA Configuration

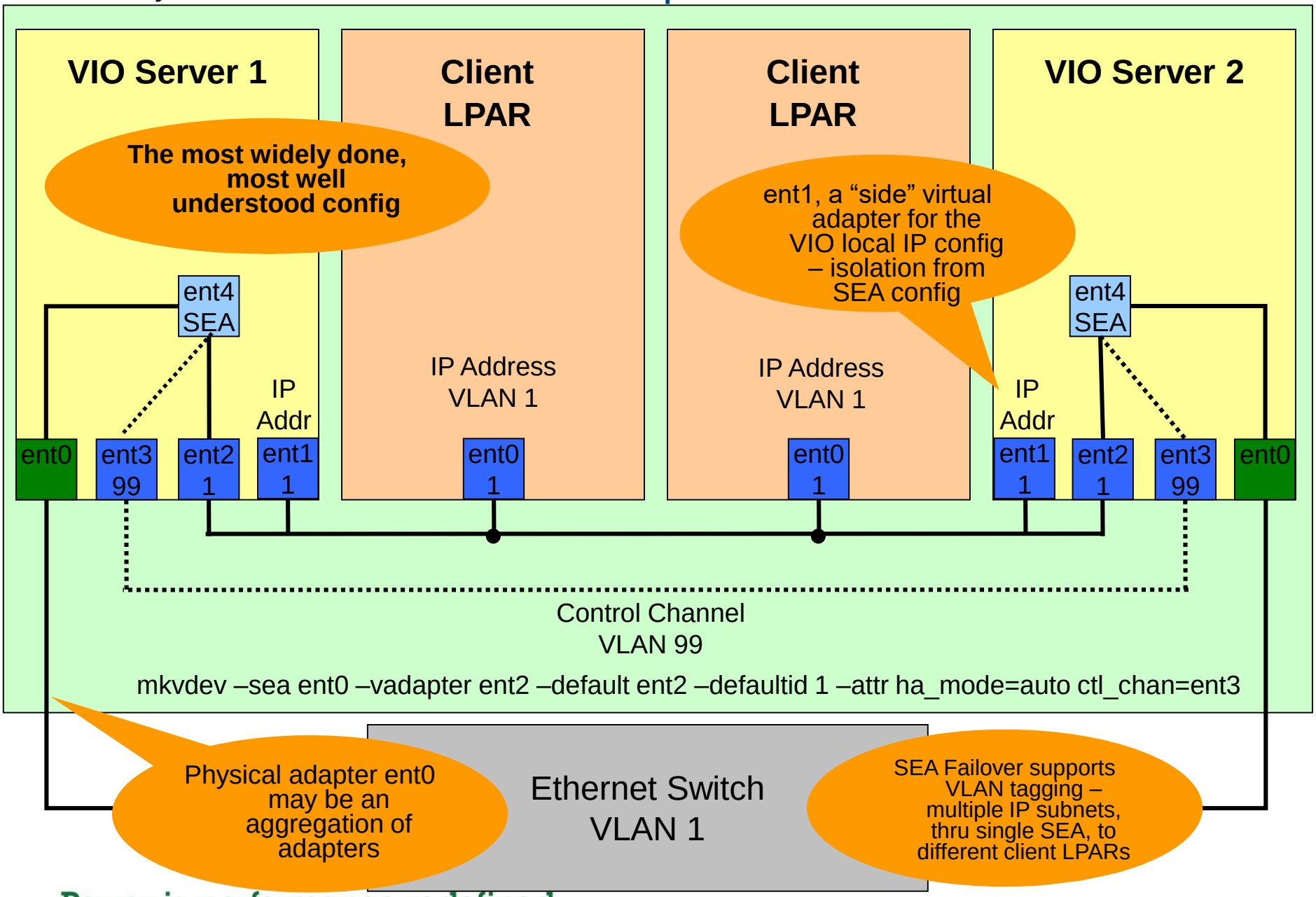
- **VIO local IP config, on SEA IP interface**

  $ mktcpip (no flags, gives a helpful usage message)
  **$ mktcipip -hostname hostname -inetaddr ip_addr -interface en10 -netmask 255.255.255.0 **
  -gateway gateway_ip -nsrvaddr dns_ip -nsrvdomain your.domain.com –start

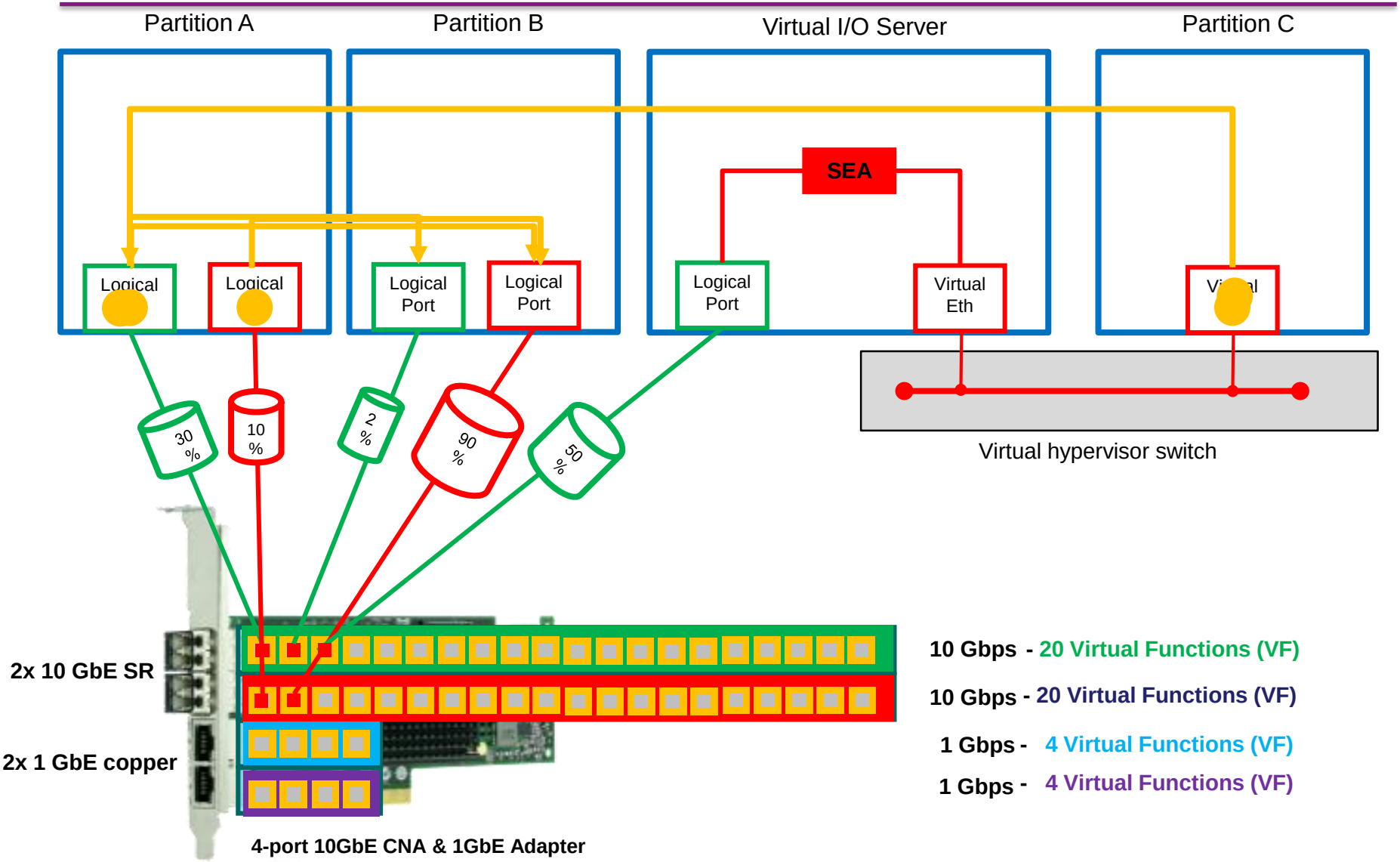
```
$ netstat -state -num
Name  Mtu   Network     Address           Ipkts Ierrs    Opkts Oerrs  Coll
en10  1500  link#2      42.d4.90.0.f0.4  52052352     0 12046192     0     0
en10  1500  9.19.98     9.19.98.41       52052352     0 12046192     0     0
lo0   16896 link#1                        6724868     0  6724868     0     0
lo0   16896 127         127.0.0.1         6724868     0  6724868     0     0
lo0   16896 ::1%1                         6724868     0  6724868     0     0
```

- **If you have mtu_bypass attribute on SEA interface, you will want set it on for bulky traffic to and from VIO local IP address. Most bulky traffic thru SEA, is NOT destined for VIO local IP. What traffic is? Live Partition Mobility, transferring memory state of the moving LPAR is done VIO to VIO.**
- `$ lsdev -dev en10 -attr | grep mtu_`
  `mtu_bypass  off  Enable/Disable largesend for virtual Ethernet`
- `$ chdev -dev en10 -attr mtu_bypass=on`
  `en10 changed`
- **mtu_bypass observed at ioslevel 2.2.1.1, and oslevel –s 6100-04-05-1015. Earlier than this, use root command line**
  `# ifconfig en10 largesend ; echo "ifconfig en10 largesend" >>/etc/rc.net`

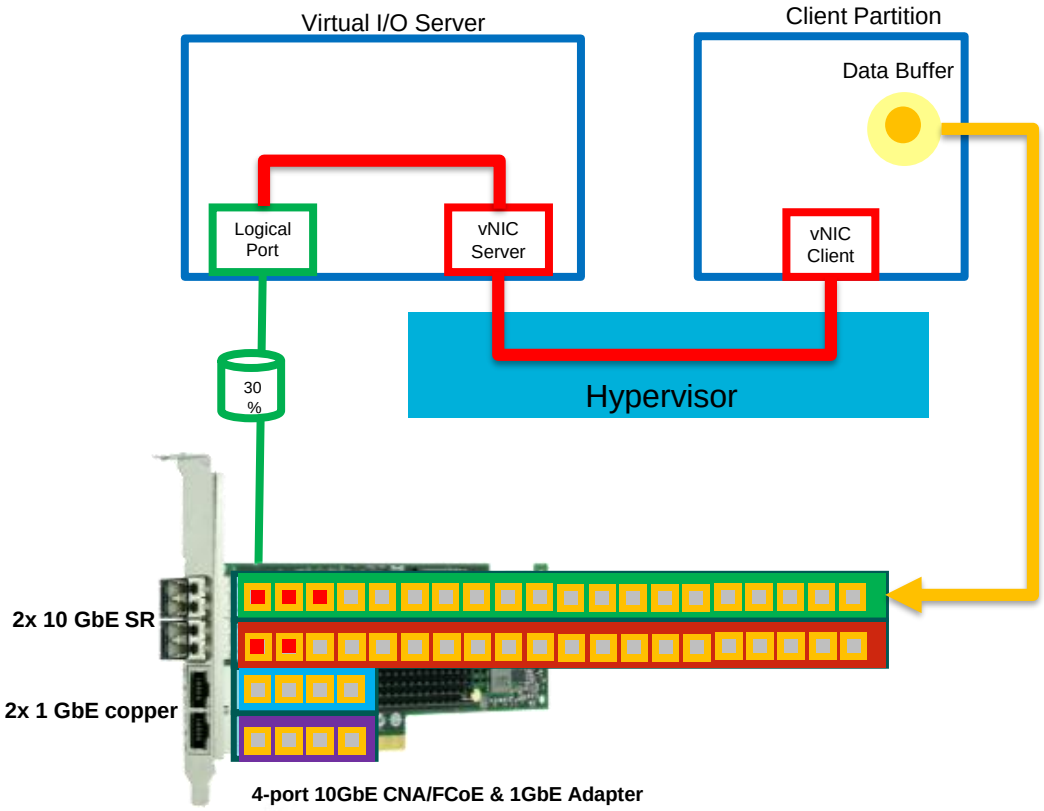
Shared Ethernet Adapter Failover



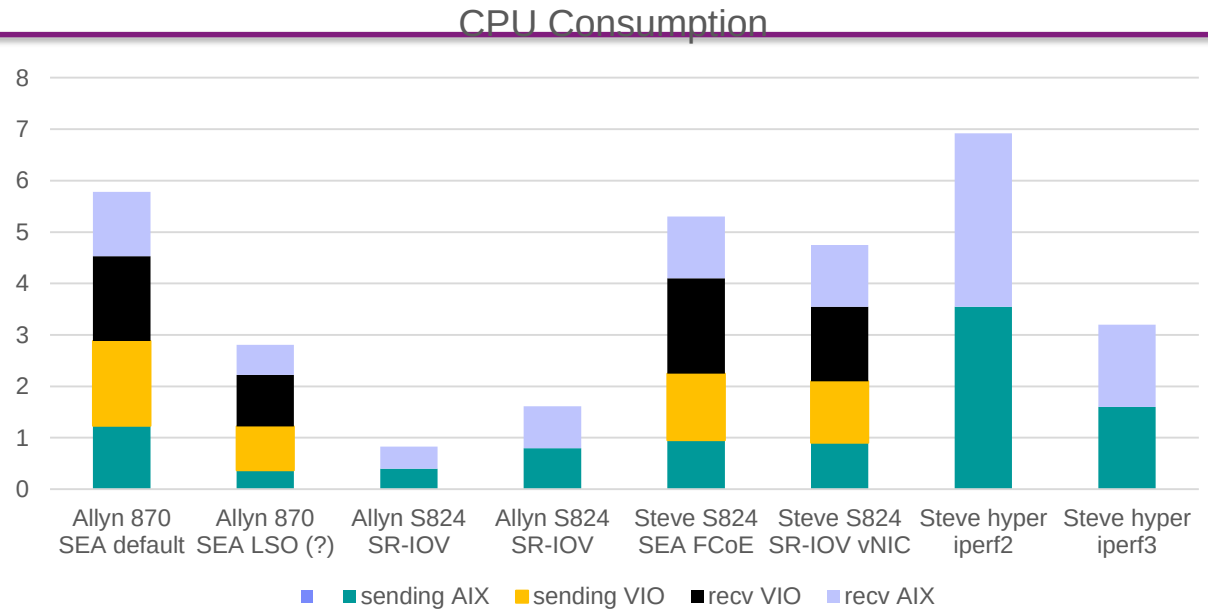
SR-IOV Architecture Internal Switching in conjunction with SEA

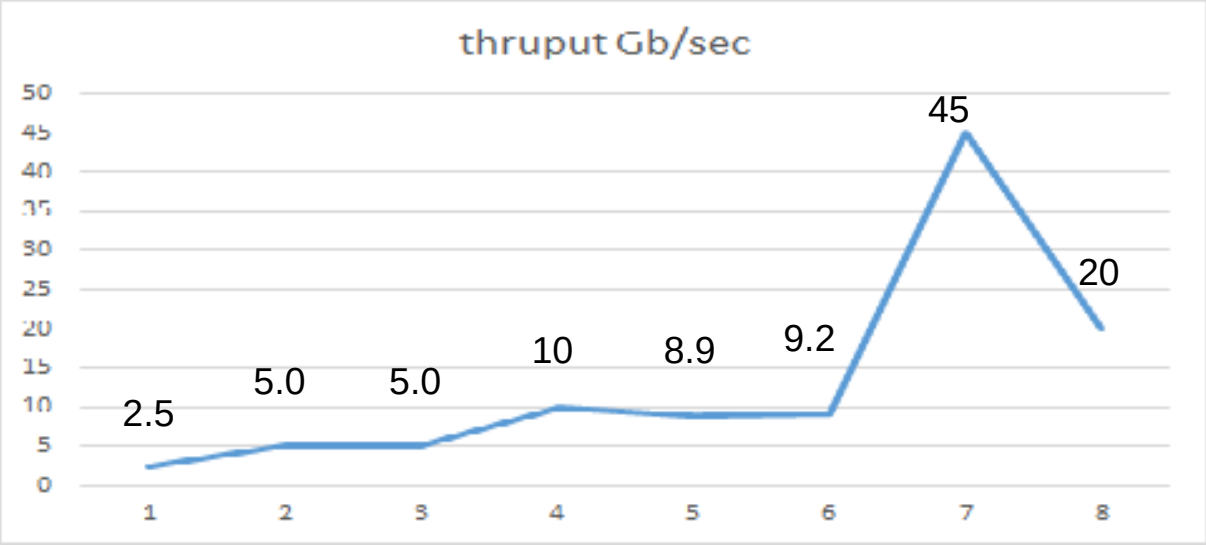


vNIC Architecture



SEA, SR-IOV (VF), and Power8 phyp virtual Ethernet



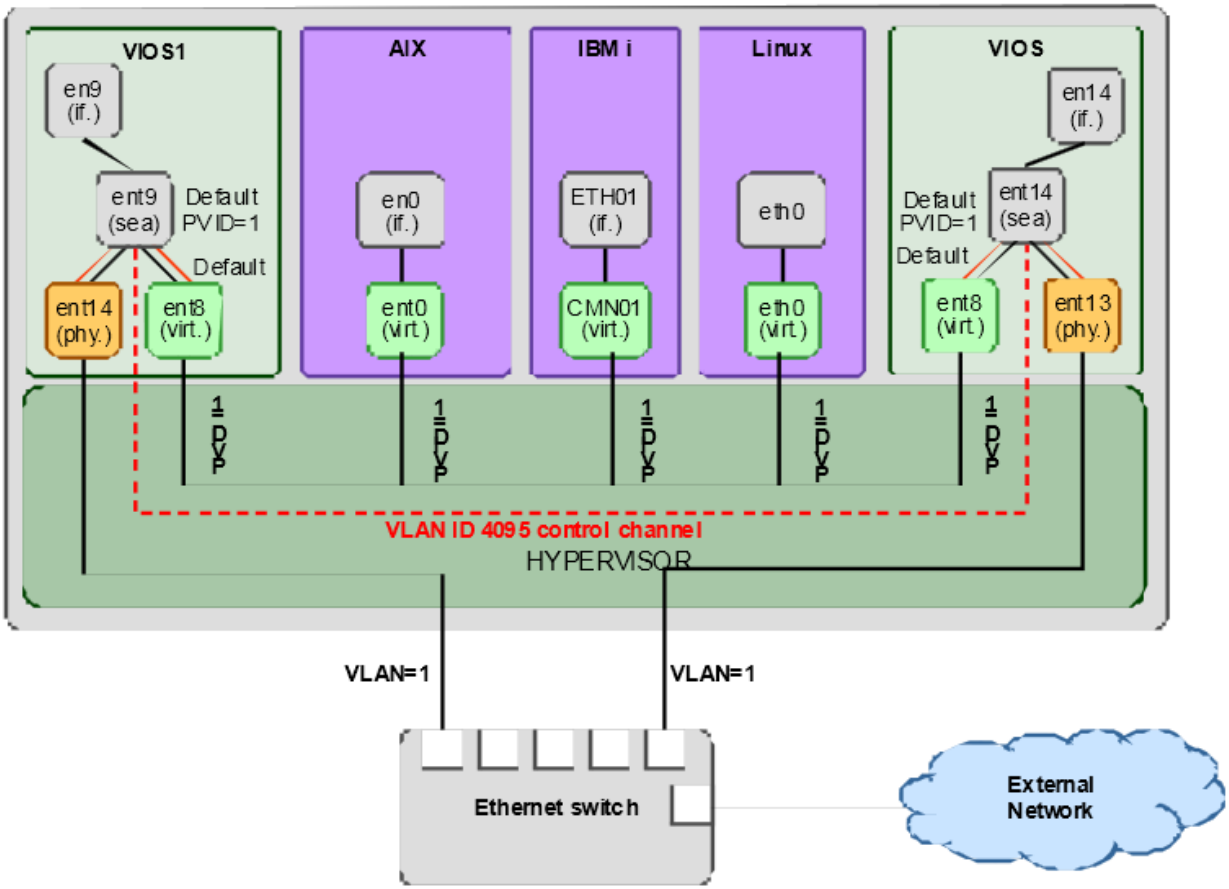


VIO 2.2.3

- **Simplified SEA Failover setup**
- **If you do not specify a control channel (ctl_chan) when running mkvdev –sea … The system will "discover" SEA Failover "partner" adapter in the other VIO**
- **Discovery protocol uses VLAN id 4095. If this is one of your actual tagged VLANs, you must continue to use control channel on both sides**
- **Multiple SEA pairs in the machine can share VLAN 4095 for discovery**
- **This is still SEA failover, and we still set priority 1 or 2 on the trunked virtual adapter in the SEA**
- **VIO server 2.2.3, HMC 7.7.8, Firmware 780. Not supported on MMB or MHB at this time.**
- **Perhaps we stay consistent with our current Power7 practices, and use this for new Power8 machines.**

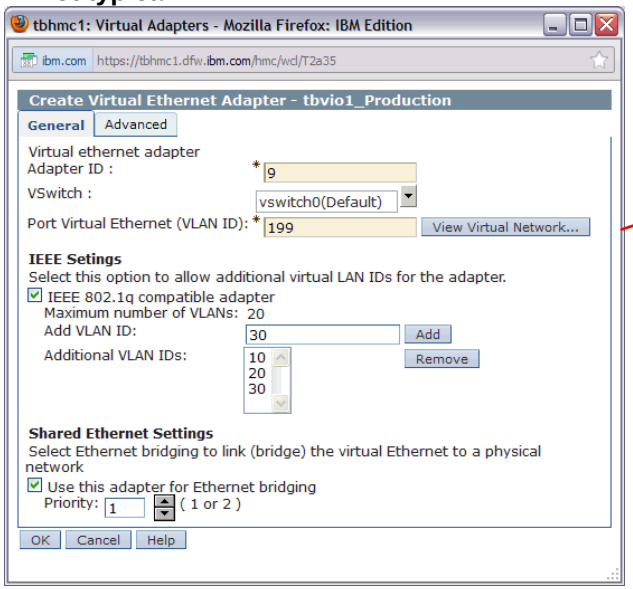
VIO 2.2.3

- **Simplified SEA Failover setup –no control channel adapter in either SEA**



SEA Configuration, VLAN tagged configuration

- **10Gb is a large pipe, and many start to consider VLAN tagging, to consolidate networks onto one adapter.**
- **Lets stay with the original config, as shown in Section 3.6, Fig 3-8 in redp4194. http://www.redbooks.ibm.com/abstracts/redp4194.html**
- **Trunked virtual adapter, ent1 in VIO, is on an unused PVID, 199 in example.**
- **Communication VLANs are added as 802.1q "additional VLANs" 10, 20, 30**
- **SEA Failover, dual VIOs supported here, but not shown**
- **Every VLAN device on top of SEA not required, unless VIO requires a local IP on each subnet – not typical.**



The following virtual LAN IDs are used:

- 10
- 20
- 30

In addition, the default virtual LAN ID 199 is also used. This default virtual LAN ID must be unique and not used by any clients in the network or physical Ethernet switch ports. For further details, see 3.6.4, "Ensuring VLAN tags are not stripped on the Virtual I/O Server" on page 68.

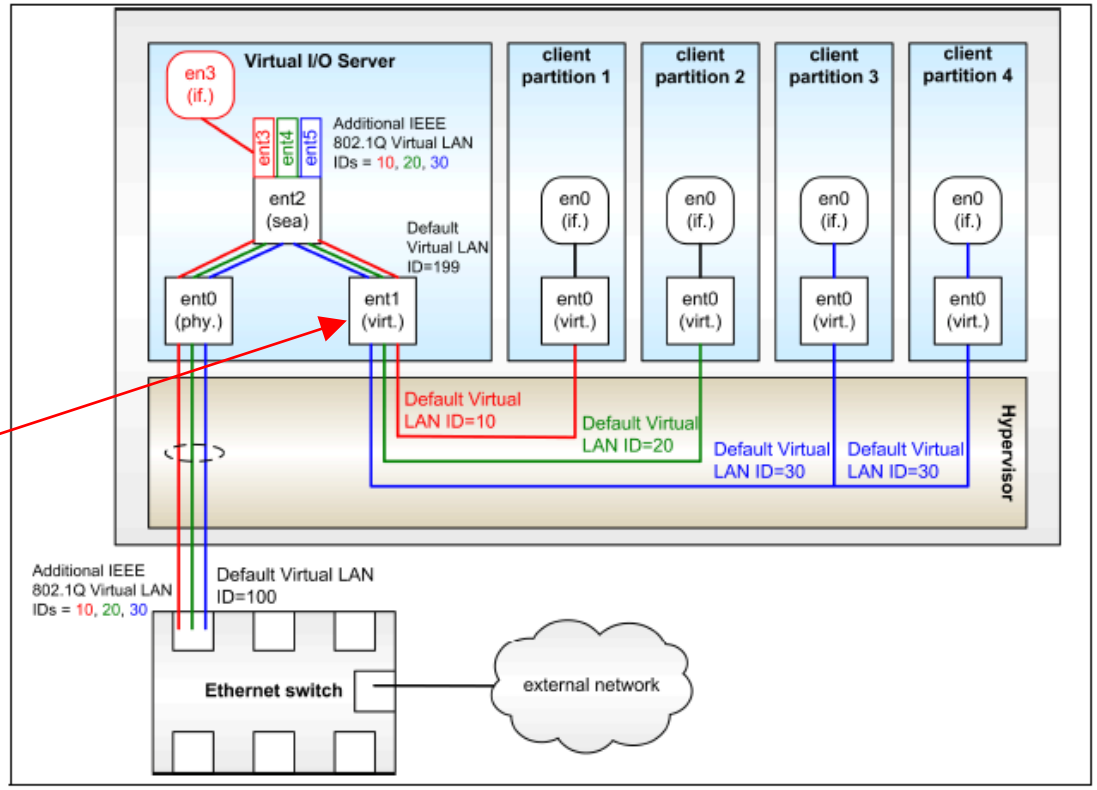


Figure 3-8 VLAN configuration scenario

Tagged configuration – VLAN awareness in SMS

- Your network admin might notify you that your switch port is configured as follows. They seem to be moving away from "access" ports, to "trunk" ports.

```
interface Ethernet1/18
 switchport mode trunk
 switchport trunk allowed vlan 10,20,30
 spanning-tree port type edge trunk
```

- If VIO is already running, SEA will be configured with a physical adapter, and a bridged virtual adapter, with 802.1q VLANs 10,20,30, just as seen on previous slide

- Since 2001, if you had AIX 5.1 running, and you were putting IP directly on a physical adapter, we could add VLAN devices on top the physical for 10,20,30 (smitty vlan), and configure IPs on those subnets. **We have handled VLANs in the operating system for a long time.**

- What do we lack? **There has been no way to specify a VLAN tag on the physical adapter in SMS.** I want to network boot a physical adapter, on VLAN 20, and install the first VIO server on the machine.

- Some workarounds
  - Network boot VIO on a different physical adapter, plugged to an access port
  - Install VIO1 from DVD media, configure tagged SEA, and network install VIO2 on virtual adapter, thru VIO1 SEA
  - You might have success adding a "native" VLAN specification on the switch port

```
interface Ethernet1/18
 switchport mode trunk
 switchport trunk native vlan 20
 switchport trunk allowed vlan 10,20,30
 spanning-tree port type edge trunk
```

  This might affect the use of "unused" VLAN id on the bridged virtual adapter in SEA; you'll have some experimentation here

- **POWER Firmware stream 760 adds VLAN awareness; the ability to specify a VLAN tag on an Ethernet adapter in SMS, for network boot**

- Observed on a 780D model, firmware AM760_051

Tagged configuration – VLAN awareness in SMS

- Version AM760_051
  SMS 1.7 (c) Copyright IBM Corp. 2000,2008 All rights reserved.

  Network Parameters
  Port 1 - IBM 2 PORT PCIe 10/100/1000 Base-TX Adapter: U2C4E.001.DBJ8765-P2-C4-T1
  1. IP Parameters
  2. Adapter Configuration
  3. Ping Test
  4. Advanced Setup: BOOTP

New option on menu at Firmware AM760_051

Navigation keys:
M = return to Main Menu
ESC key = return to previous screen X = eXit System Management Services
---
Type menu item number and press Enter or select Navigation key:

Tagged configuration – VLAN awareness in SMS

- Version AM760_051
  SMS 1.7 (c) Copyright IBM Corp. 2000,2008 All rights reserved.

  Advanced Setup: BOOTP
  Port 1 - IBM 2 PORT PCIe 10/100/1000 Base-TX Adapter: U2C4E.001.DBJ8765-P2-C4-T1
  1. Bootp Retries 5
  2. Bootp Blocksize 512
  3. TFTP Retries 5
  4. VLAN Priority 0
  5. VLAN ID 0 (default - not configured)

Specify your VLAN tag here, then escape to perform 3. ping test

  ---
  Navigation keys:
  M = return to Main Menu
  ESC key = return to previous screen X = eXit System Management Services
  ---
  Type menu item number and press Enter or select Navigation key:

Tagged configuration – VLAN awareness

- Suppose you are running AIX, and you want to kick off a network boot and reinstall from the command line. Yes, you can specify VLAN tag on the bootlist command (AIX 6100-08 or 7100-02):

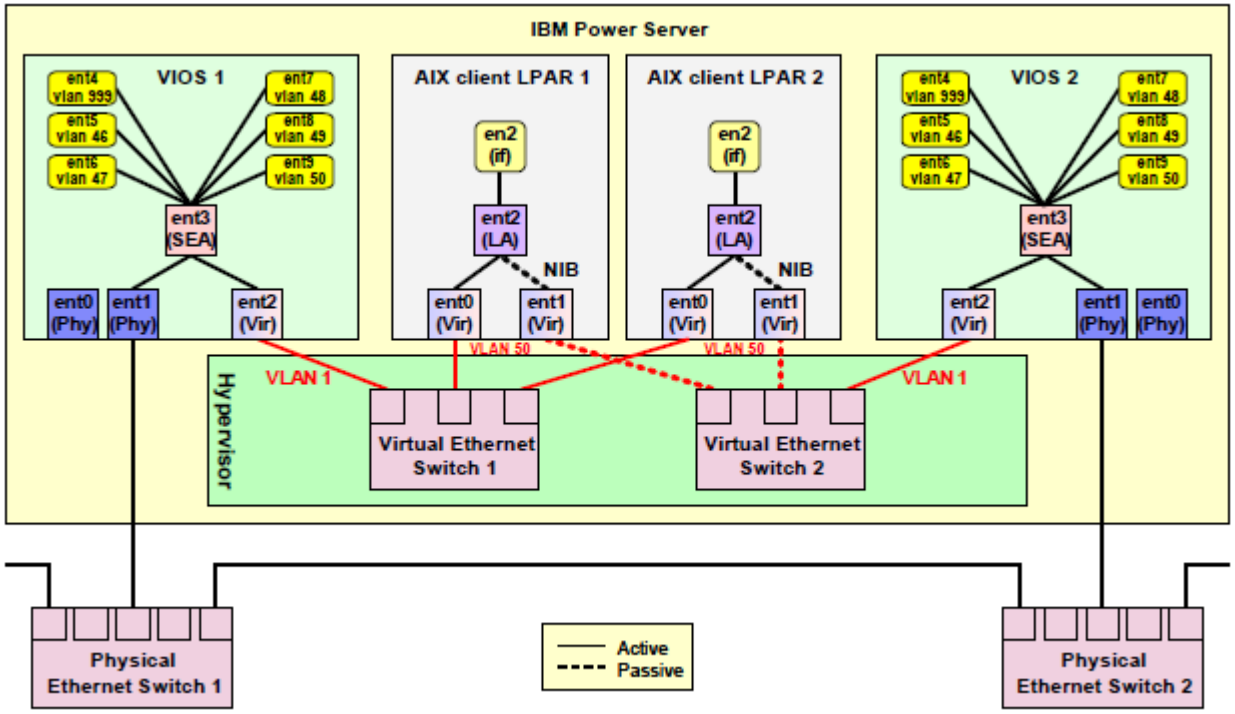
```
# bootlist -m normal ent0 client=<client_ip> bserver=<master_ip>
gateway=<client_gw> vlan_tag=<vlan_tag> [vlan_pri=<vlan_pri> ] hdisk0 hdisk1
```

- At HMC V7R7.7.0.2 it is also in lpar_netboot

```
lpar_netboot -M -n [-v] [-x] [-f] [-i] [-E environment [-E ...]]
                [-A] -t ent [-T {on|off}] [-D -s speed -d duplex
                -S server -G gateway -C client [-K subnetmask]
                [-V vlan_tag] [-Y vlan_priority]]
                partition-name partition-profile managed-system
```

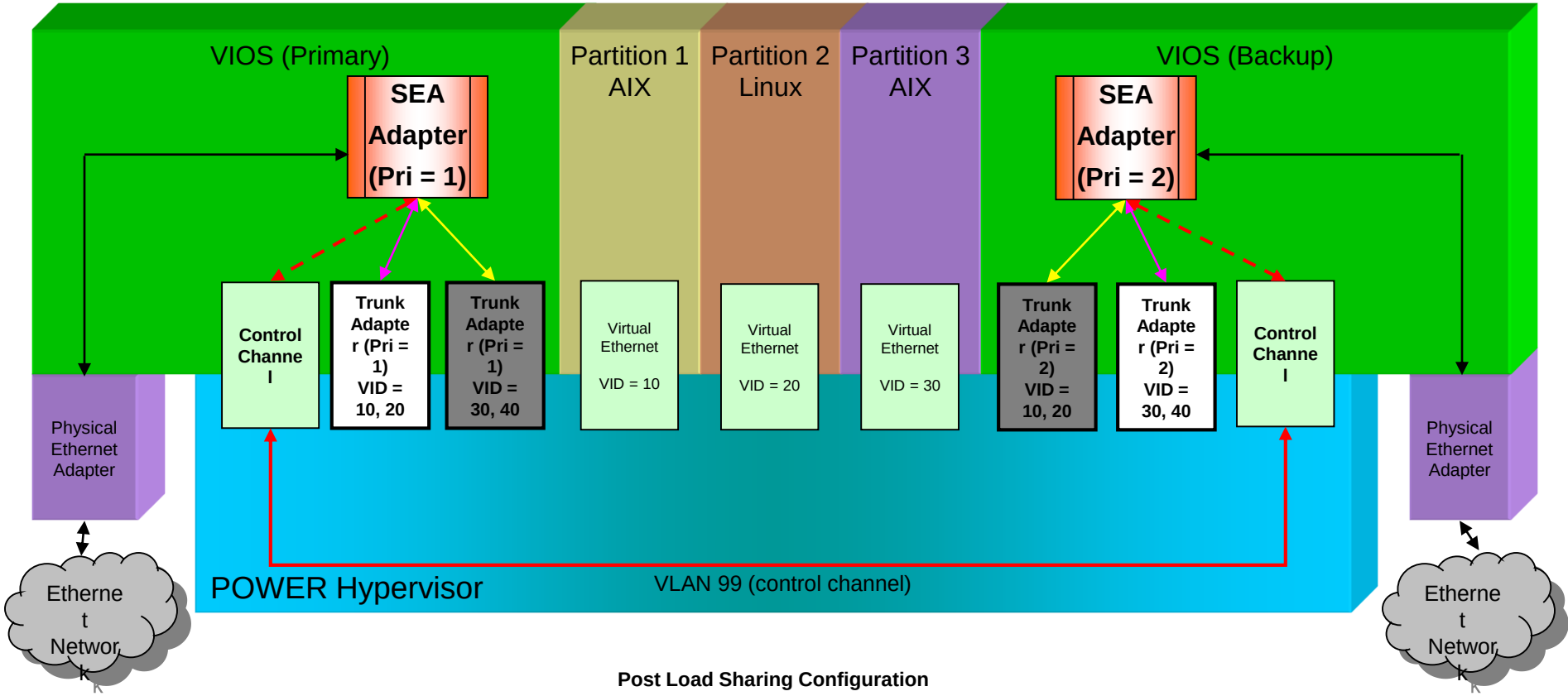
10Gb SEA Configuration, both sides active

- **Early field developed solution for shops not satisfied with idle SEA standby 10Gb adapter and switch port.**
- **Independent SEAs configured in each VIO, on same PVIDs, tagged**
- **How do they avoid BPDU Loop storm? Different Virtual Switches, and NIB in the client LPAR**
- **http://www.wmduszyk.com/wp-content/uploads/2012/01/PowerVM-VirtualSwitches-091010.pdf**
  (google "vio sea 10gb miller" look for pdf titled "Using Virtual Switches in PowerVM to Drive Maximum Value of 10Gb")
  Disclaimer - This article has expired in IBM Techdocs, and the state of author support (Miller and Speetjens) is uncertain.



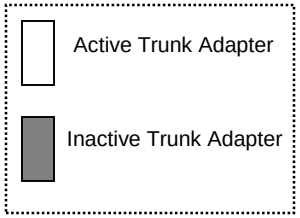
SEA Configuration, ha_mode=sharing

VIO development response to active-active requirement



Post Load Sharing Configuration

VIO client 1 & 2 are bridged by primary VIOS, client 3 is bridged by backup VIOS



SEA Configuration ha_mode=sharing

- **VIO 2.2.1.1 required**
- **Still a single SEA Failover configuration – single ctl_chan**
- **At least 2 (up to 16) trunked virtual adapters joined into each SEA**
- **Previous slide shows trunked virtual for VLAN 10,20, and a trunked virtual for VLAN 30,40, in each SEA**
- **Previous slide is tagged example. May be untagged as well.**
- **Both trunked adapters in SEA must have external access checkbox, and same trunk priority (e.g. both are 1 in vio1, and both are 2 in vio2)**
- **Set ha_mode=sharing on Primary SEA first, then Secondary**
  $ chdev –dev entX –attr ha_mode=sharing
- **Secondary offers sharing to Primary**
- **Client LPARs do not require NIB configuration**
- **POWER Admin balances placement of LPARs on VLANs**

SEA Configuration ha_mode=sharing Sample config

- **tbvio1**
  adapter 9 (ent10)
  PVID 160
  802.1q 162 164
  Pri 1

  adapter 10 (ent11)
  PVID 170
  802.1q 172 174
  Pri 1

  adapter 11 (ent12)
  PVID 199

- **tbvio2**
  adapter 10 (ent10)
  PVID 160
  802.1q 162 164
  Pri 2

  adapter 12 (ent11)
  PVID 170
  802.1q 172 174
  Pri 2

  adapter 13 (ent12)
  PVID 199

- **In both VIOs, physical ent6 is one port on FCoE adapter 5708**
  $ mkvdev –sea ent6 –vadapter ent10,ent11 –default ent10 –defaultid 160 –attr ha_mode=sharing largesend=1 large_receive=yes ctl_chan=ent12
  ent9 available

SEA Configuration ha_mode=sharing Sample config

- **entstat command on SEA shows a number of things. First, tbvio1:**

```
$ entstat -all ent9 | more
...
VLAN Ids :
   ent11: 170 172 174
   ent10: 160 162 164
...
   VID shared: 160 162 164
   Number of Times Server became Backup: 0
   Number of Times Server became Primary: 1
   High Availability Mode: Sharing
   Priority: 1
```

- **And now in tbvio2**

```
...
VLAN Ids :
   ent11: 170 172 174
   ent10: 160 162 164
...
   VID shared: 170 172 174
   Number of Times Server became Backup: 1
   Number of Times Server became Primary: 0
   High Availability Mode: Sharing
   Priority: 2
```

SEA Configuration ha_mode=sharing Sample config

- **Just a quick check, that I put all virtual adapters on the correct virtual switch:**
  $ entstat -all ent9 | grep "^Switch ID:“
  Switch ID: vswitch1
  Switch ID: vswitch1
  Switch ID: vswitch1

- **Above, how do you match adapter ID with ent name?**

- $ lsdev -type adapter -field name physloc | grep ent

```
ent0      U78C0.001.DBJ4725-P2-C8-T1
ent1      U9179.MHB.1026D1P-V1-C2-T1
ent2      U9179.MHB.1026D1P-V1-C3-T1
ent3      U9179.MHB.1026D1P-V1-C4-T1
ent4
ent5      U9179.MHB.1026D1P-V1-C7-T1
ent6      U78C0.001.DBJ4725-P2-C6-T1
ent7      U78C0.001.DBJ4725-P2-C6-T2
ent8      U9179.MHB.1026D1P-V1-C8-T1
ent9
ent10      U9179.MHB.1026D1P-V1-C9-T1
ent11      U9179.MHB.1026D1P-V1-C10-T1
ent12      U9179.MHB.1026D1P-V1-C11-T1
```

Another 10Gb Performance Reference

- https://www.ibm.com/developerworks/wikis/download/attachments/153124943/7_PowerVM_10Gbit_Ethernet.pdf?version=1

  Gareth Coates, IBM UK Advanced Technical Support suggests higher thruput may be obtained by more trunked virtual adapters in the SEA.

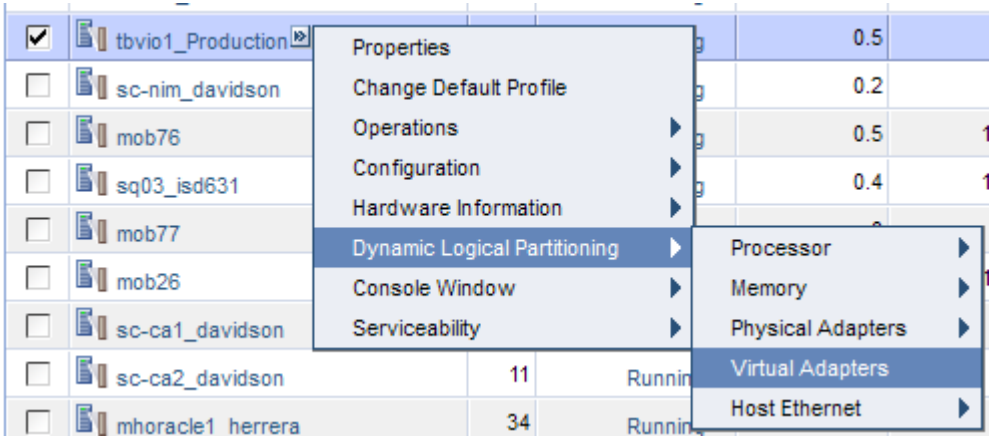
  ha_mode=sharing requires at least 2.

  In a tagged environment, perhaps you would use 4, for four different 802.1q "additional VLANs," one per trunked virtual adapter.

Dynamic VLANs

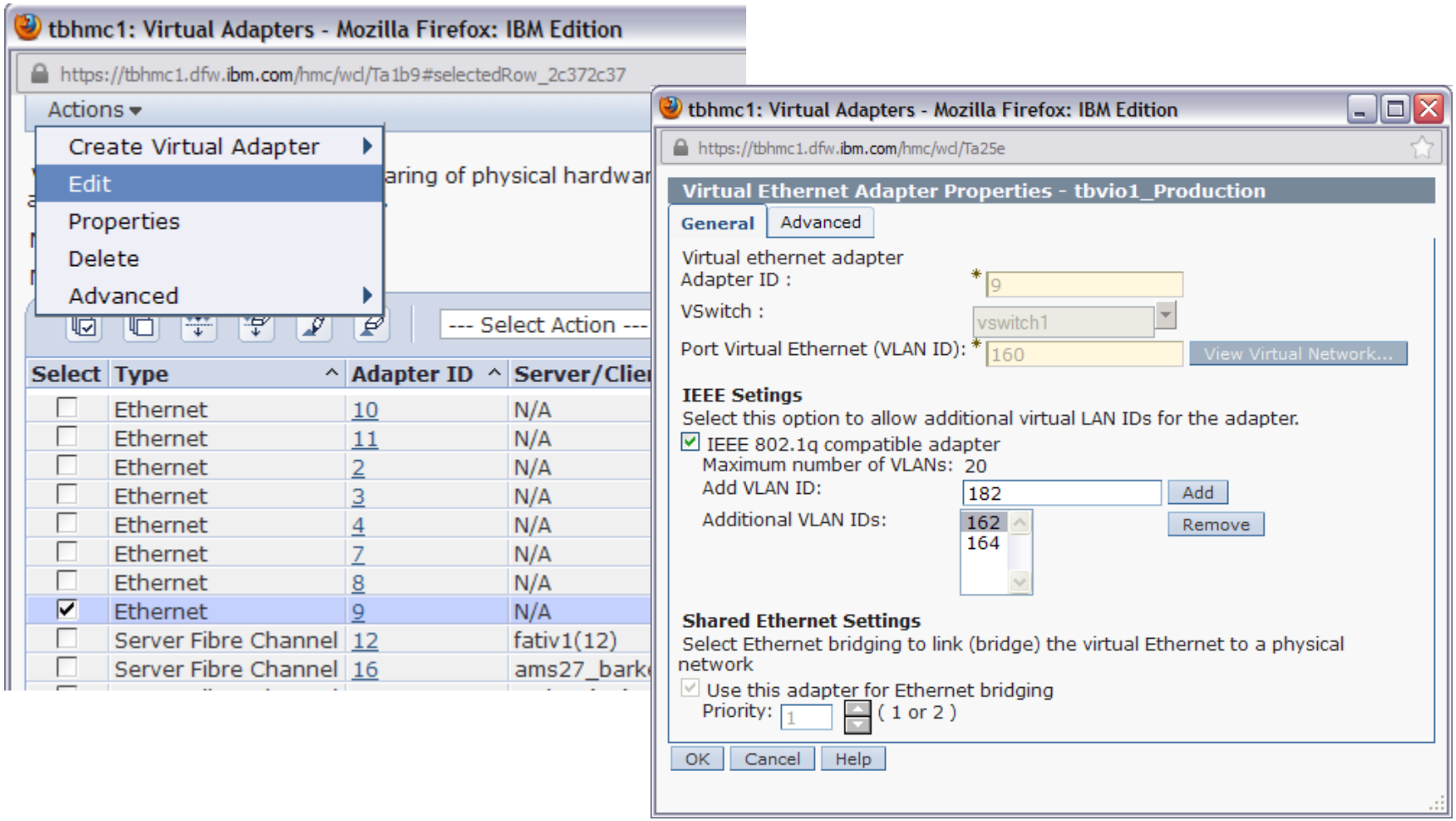
- Perhaps you have a running configuration, and you need to add an additional VLAN.
- First, what is running in VIO?
  $ entstat -all ent9 | more
  ...
  VLAN Ids :
  ent11: 170 172 174
  ent10: 160 162 164
  ...
  VID shared: 160 162 164

- DLPAR, and “edit” the adapter



Dynamic VLANs

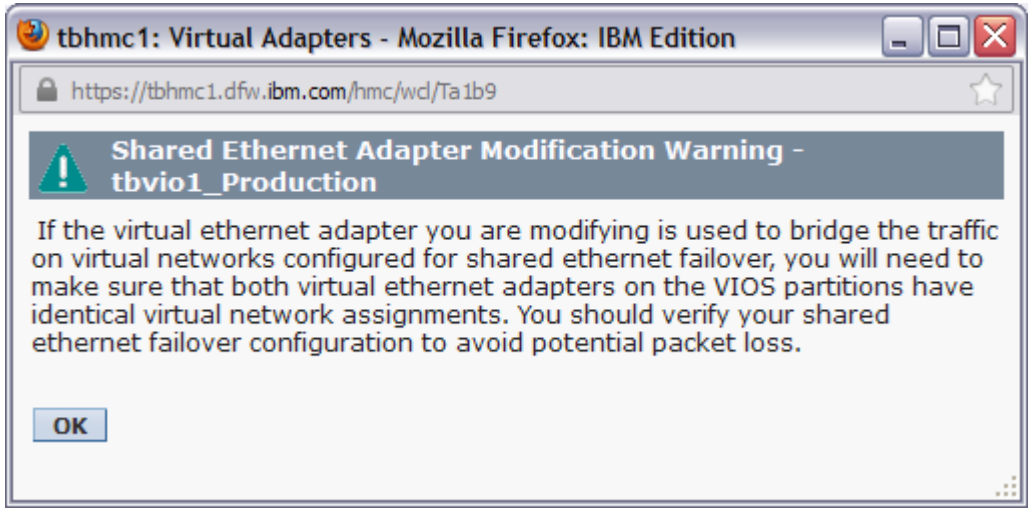
- Checkbox the adapter, and actions -> edit



Type in new VLAN id, hit **Add**, hit **OK**, hit **OK**

Dynamic VLANs

- Note the warning to make the same change on SEA in the other VIO, hit OK



Check entstat again for new VLAN id

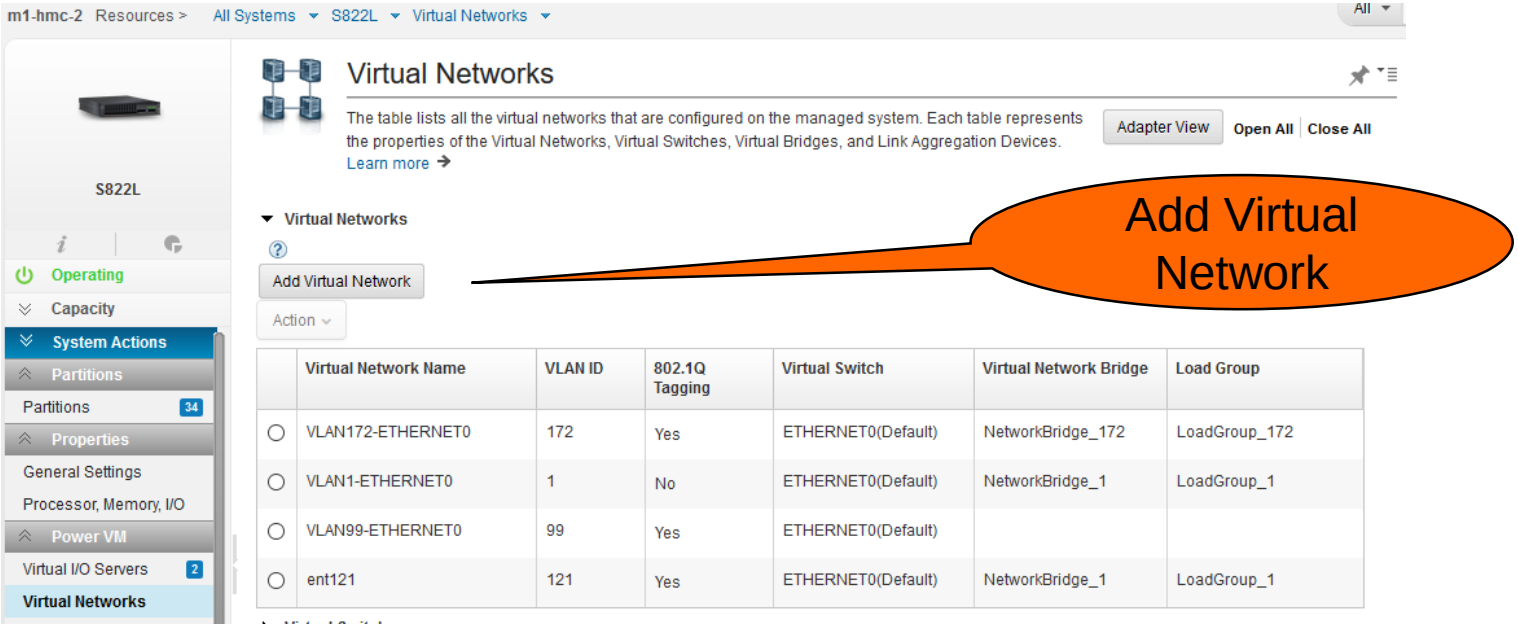
```
 $ entstat -all ent9 | more
...
VLAN Ids :
    ent11: 170 172 174
    ent10: 160 162 164 182
...
```

VID shared: 160 162 164 182

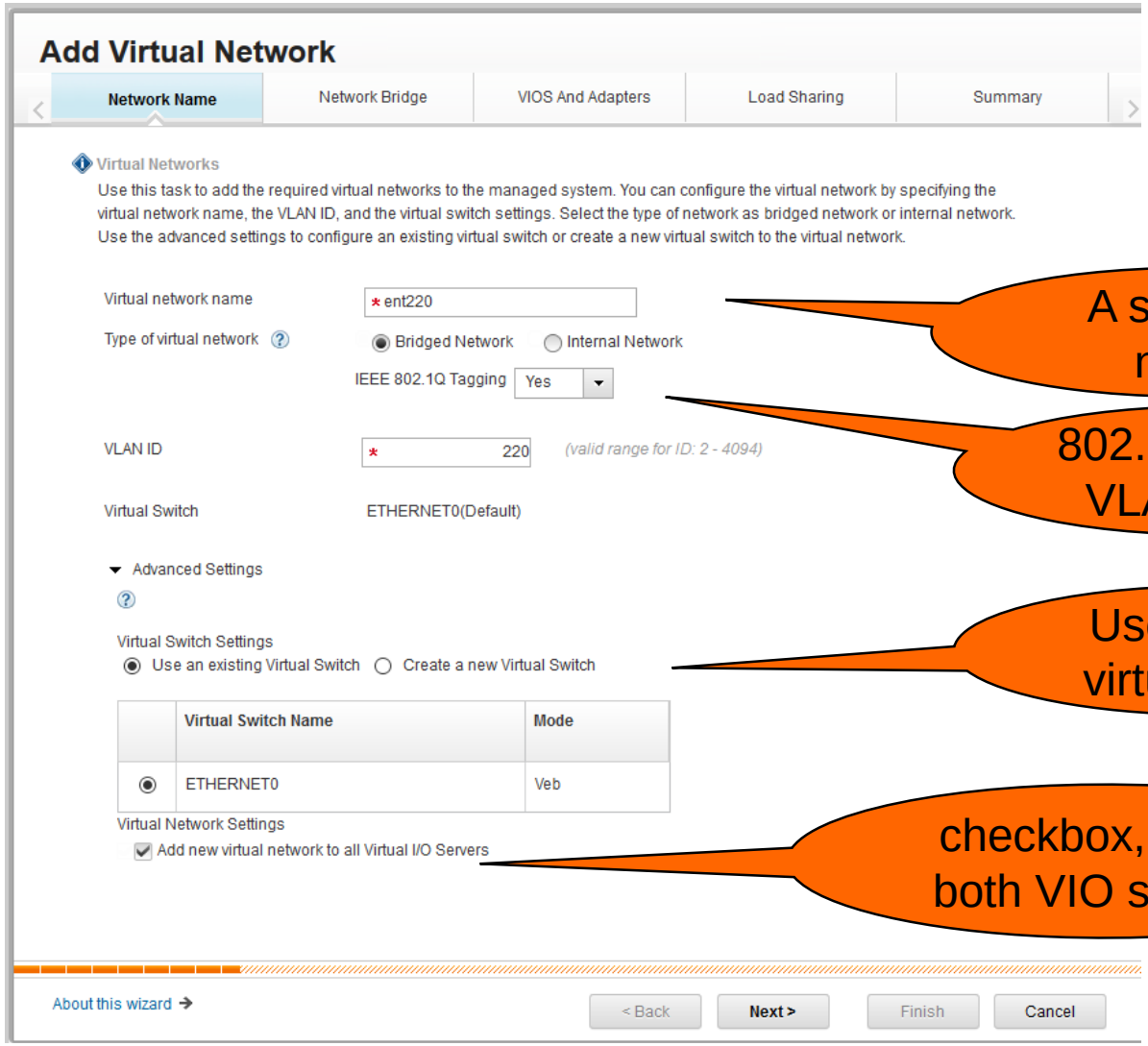
Dynamic VLANs - HMC enhanced mode

- Wizard process - "Add Virtual Network"

Started with PVID 1, and 802.1q VLAN id 121...
$ entstat -all ent7 | more
...
VLAN Ids :
ent4: 1 121
...
Looking at the Server S822L "Virtual Networks" then hit "Add Virtual Network"



Dynamic VLANs - HMC enhanced mode



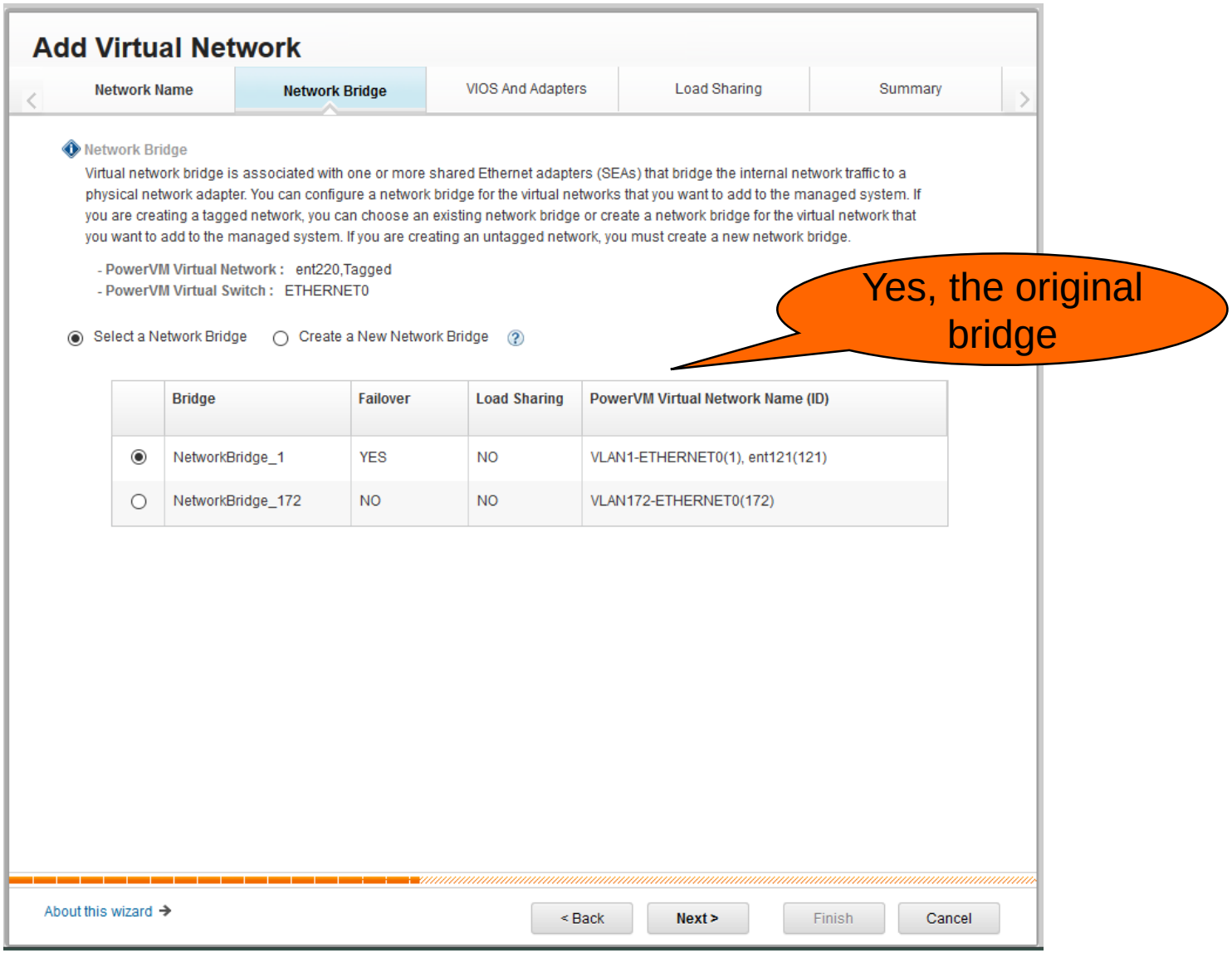
A sensible name

802.1q tag yes, VLAN id 220

Use existing virtual switch

checkbox, put in both VIO servers

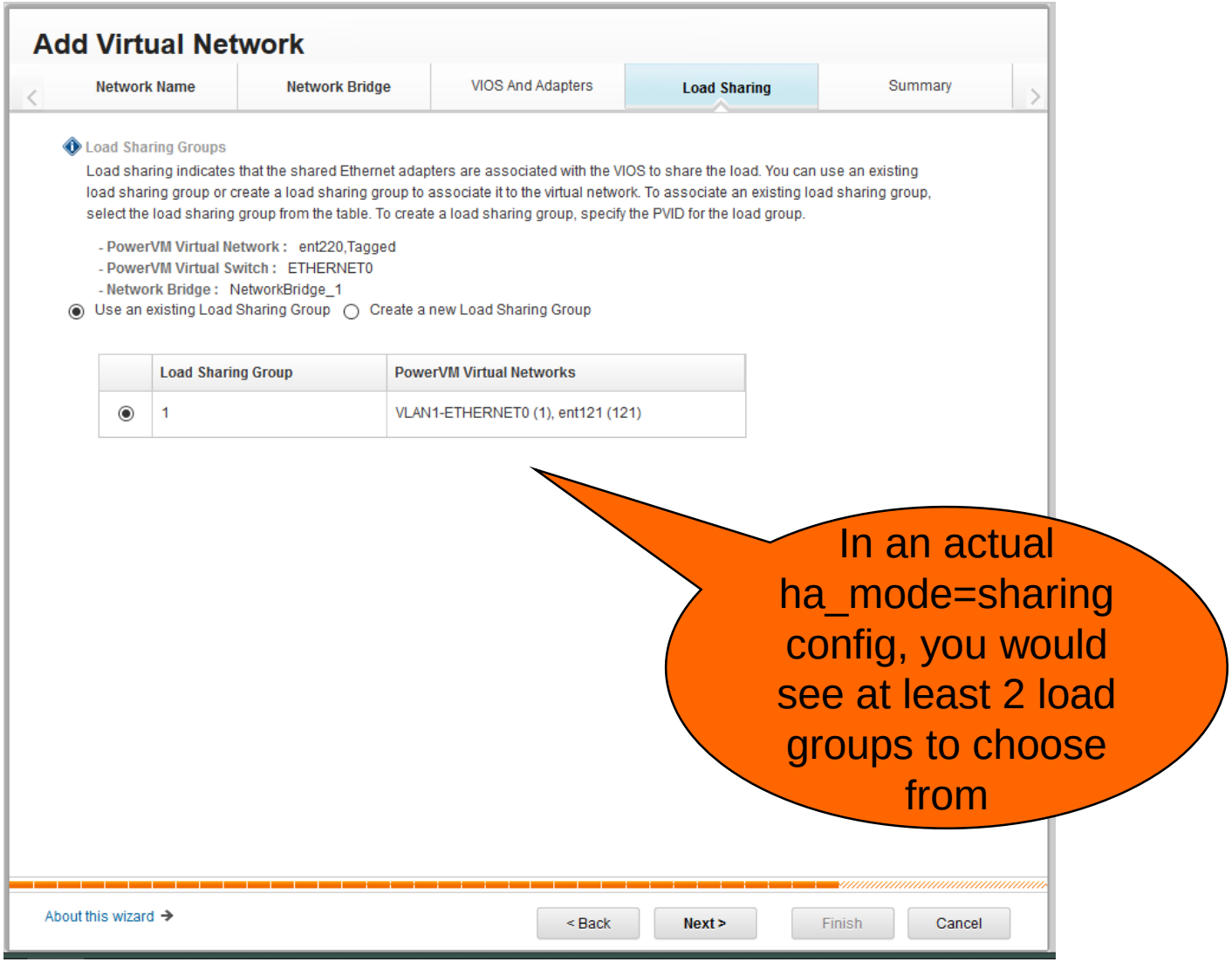
Dynamic VLANs - HMC enhanced mode



Yes, the original bridge

Power is performance redefined

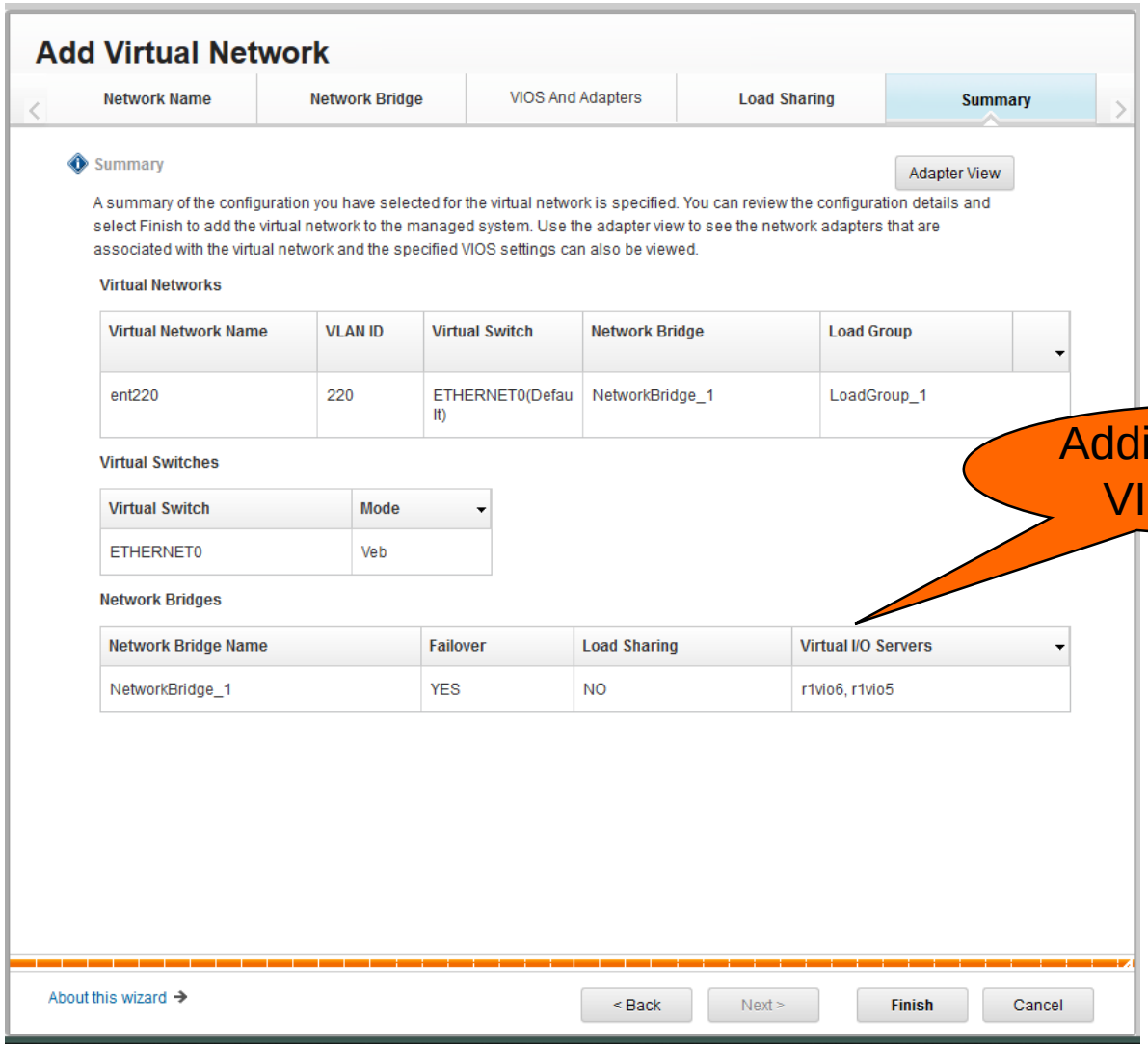
Dynamic VLANs - HMC enhanced mode



In an actual ha_mode=sharing config, you would see at least 2 load groups to choose from

Power is performance redefined

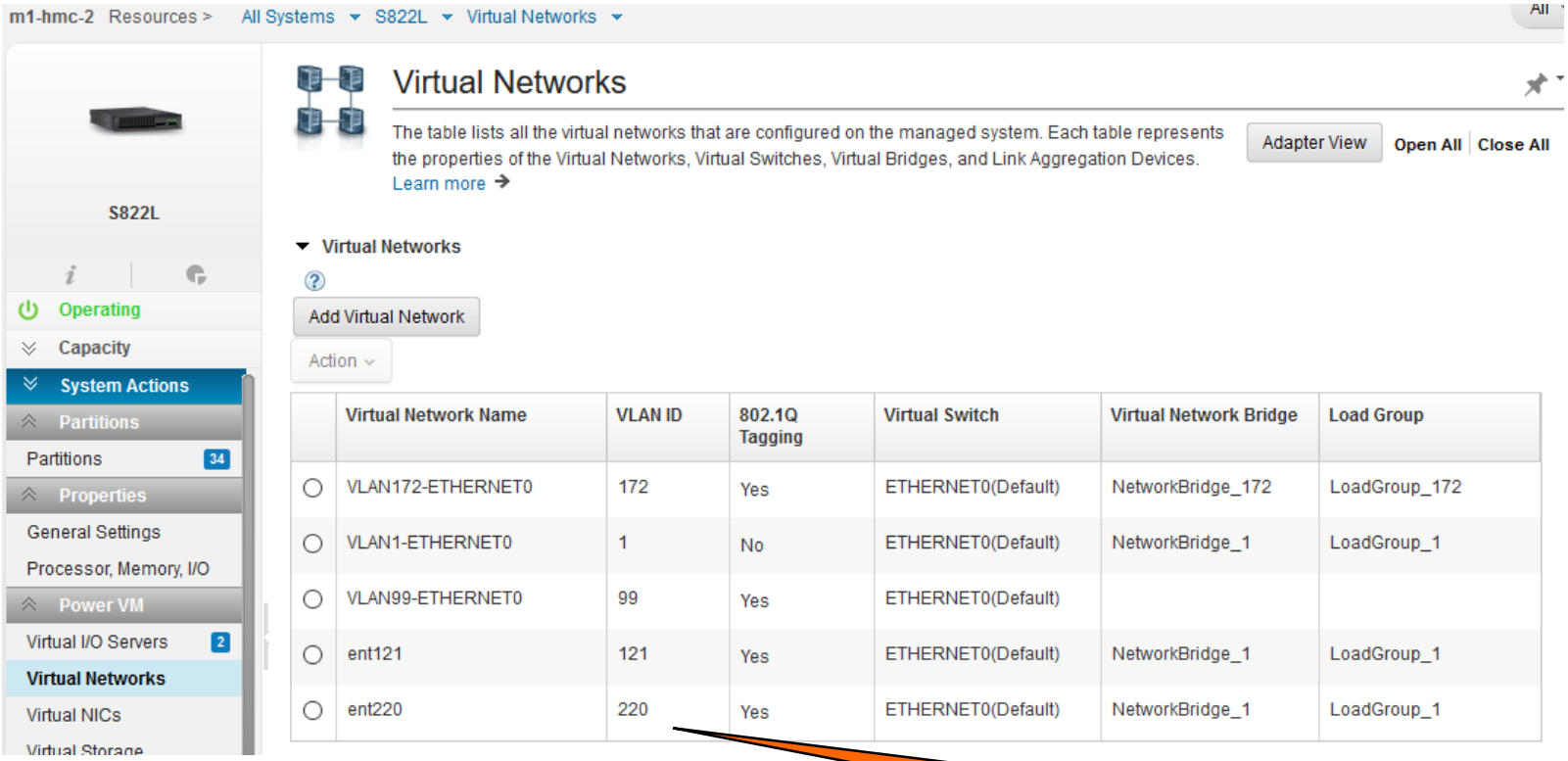
Dynamic VLANs - HMC enhanced mode



Adding it in both VIO servers

Power is performance redefined

Dynamic VLANs - HMC enhanced mode



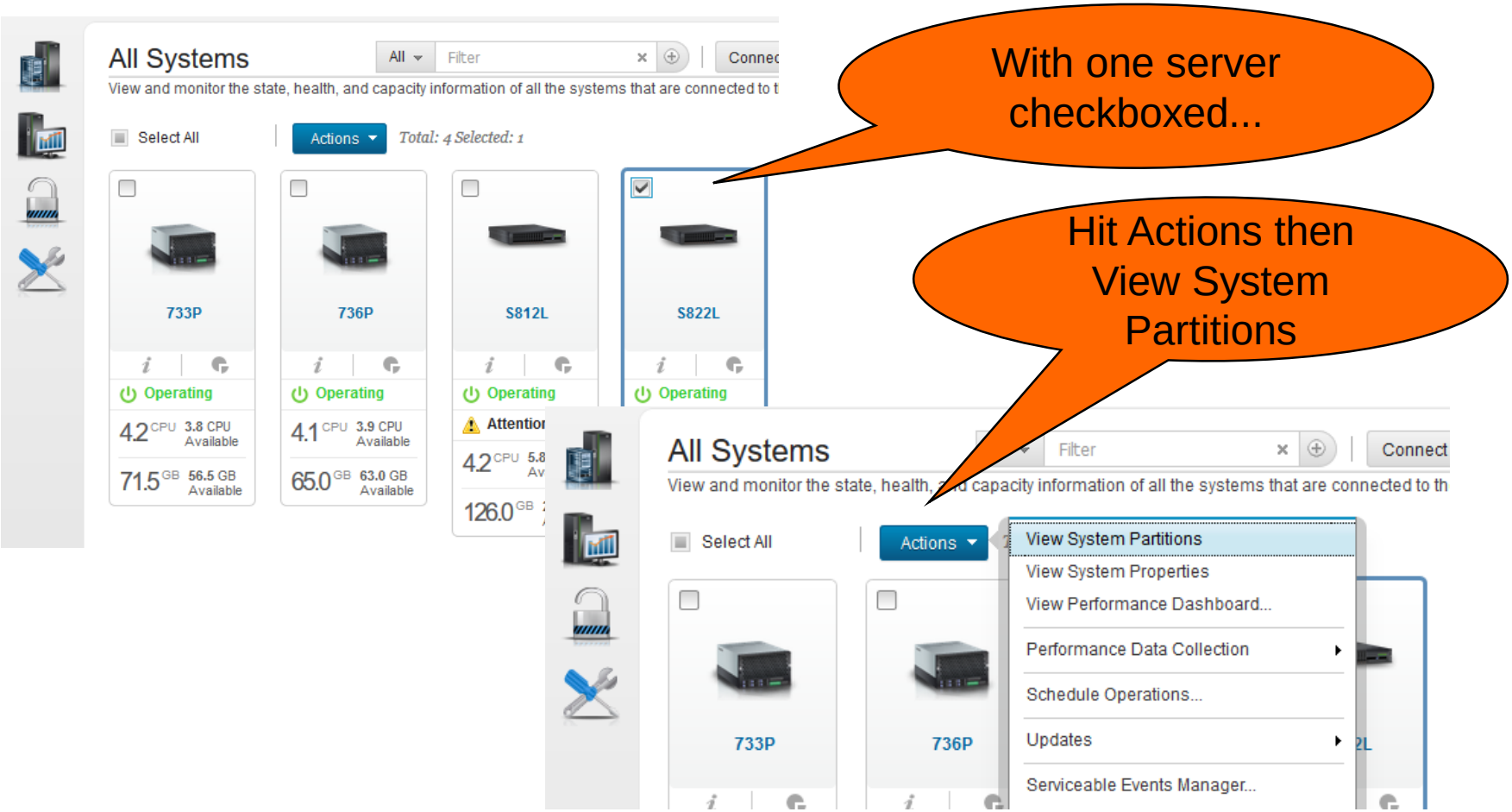
Now see 220 in the enhanced interface

Again, back in VIO...

```
 $ entstat -all ent7 | more
...
VLAN Ids :
    ent4: 1 121 220
...
```

HMC Enhanced+ Mode

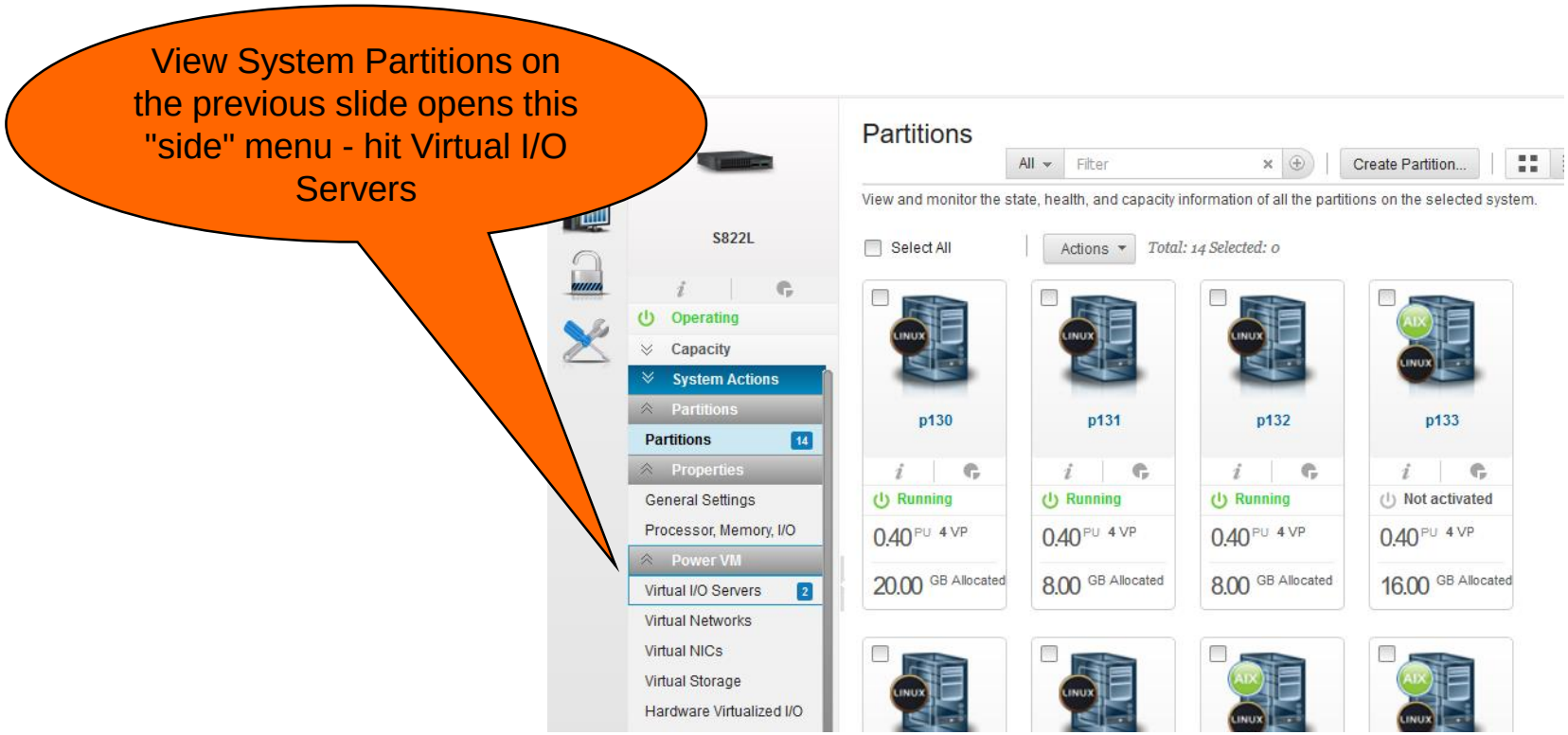
- HMC 8.8.6, last version with both Classic and Enhanced modes
- It wasn't clear how to reach "profiles" for VIO servers, to configure virtual adapters, but I found it



Power is performance redefined

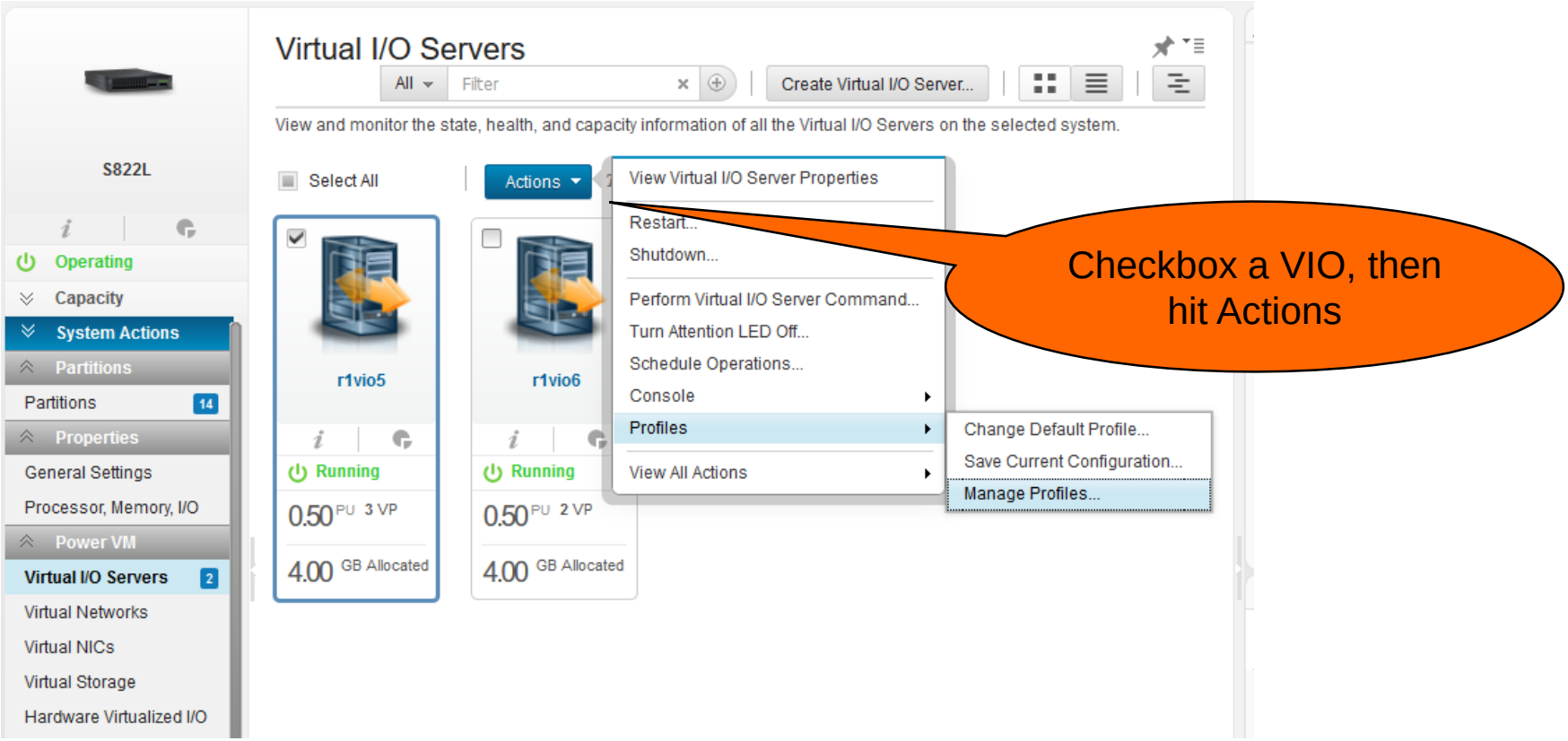
HMC Enhanced+ Mode

- HMC 8.8.6, last version with both Classic and Enhanced modes



HMC Enhanced+ Mode

- HMC 8.8.6, last version with both Classic and Enhanced modes



Power is performance redefined

SEA qos_mode

- **With VLAN tagging, SEA can participate in Quality of Service Selection**
- ```
  # lsdev -Cc adapter | grep ent
  ent0     Available       Virtual I/O Ethernet Adapter (l-lan)
  ent1     Available       Logical Host Ethernet Port (lp-hea)
  ent2     Available       Virtual I/O Ethernet Adapter (l-lan)
  ent3     Available       Shared Ethernet Adapter
  ```
- ```
  $ lsdev -dev ent3 -range qos_mode
  disabled
  strict
  loose
  ```
- strict - More important traffic is sent preferentially over less important traffic. This mode provides better performance and more bandwidth to more important traffic; however, it can result in substantial delays for less important traffic.
- Lets start with loose
  ```
  $ chdev –dev ent3 –attr qos_mode=loose
  ```

SEA Throughput

- **$ seastat –d ent5 (In VIO, which LPARs are getting how much traffic thru SEA?)**

```
================================================================================
Advanced Statistics for SEA
Device Name: ent5
================================================================================
MAC: 32:43:23:7A:A3:02
----------------------
VLAN: None
VLAN Priority: None
Hostname: mob76.dfw.ibm.com
IP: 9.19.51.76
Transmit Statistics:          Receive Statistics:
--------------------          -------------------
Packets: 9253924              Packets: 11275899
Bytes: 10899446310            Bytes: 6451956041
================================================================================
MAC: 32:43:23:7A:A3:02
----------------------
VLAN: None
VLAN Priority: None
Transmit Statistics:          Receive Statistics:
--------------------          -------------------
Packets: 36787                Packets: 3492188
Bytes: 2175234                Bytes: 272207726
================================================================================
MAC: 32:43:2B:33:8A:02
----------------------
VLAN: None
VLAN Priority: None
Hostname: sharesvc1.dfw.ibm.com
IP: 9.19.51.239
Transmit Statistics:          Receive Statistics:
--------------------          -------------------
Packets: 10                   Packets: 644762
Bytes: 420                    Bytes: 484764292
```

SEA Throughput

- **chdev –dev ent7 –attr accounting=enabled**
- **VIO topas, then uppercase E**

```
Topas Monitor for host:    mdvio1      Interval:   2   Wed Apr  3 12:15:55 2013
  ==============================================================================
  Network                                  KBPS   I-Pack   O-Pack    KB-In   KB-Out
  ent7 (SEA PRIM)                        4825.6   3100.1   3099.6   2412.8   2412.8
    |\--ent5 (PHYS)                      2412.9   1794.3   1306.8   2293.5    119.4
    |\--ent2 (VETH)                      2412.7   1305.8   1792.8    119.3   2293.4
     \--ent4 (VETH CTRL)                    1.9      0.0      5.5      0.0      1.9
  lo0                                       0.0      0.0      0.0      0.0      0.0
```

To see SEA traffic in VIO topas, you must have IP address on the SEA interface (en7 here), and not on a “side” virtual adapter

SEA Throughput

- # ./sk_sea (what is total aggregate packet count on SEA?
  In VIO, as root, after $ oem_setup_env)

  sk_sea -i interval -a adapter

  -i interval (seconds)
  -a adapter
  -h or -? Usage

- # ./sk_sea -i 10 -a ent5

  net to SEA--> 341656869 SEA to virt--> 341656842 250416752 <--to net from SEA 250416752 <--to SEA from virt

  net to SEA--> 1089 SEA to virt--> 1089 535 <--to net from SEA 535 <--to SEA from virt

  net to SEA--> 804 SEA to virt--> 804 523 <--to net from SEA 523 <--to SEA from virt

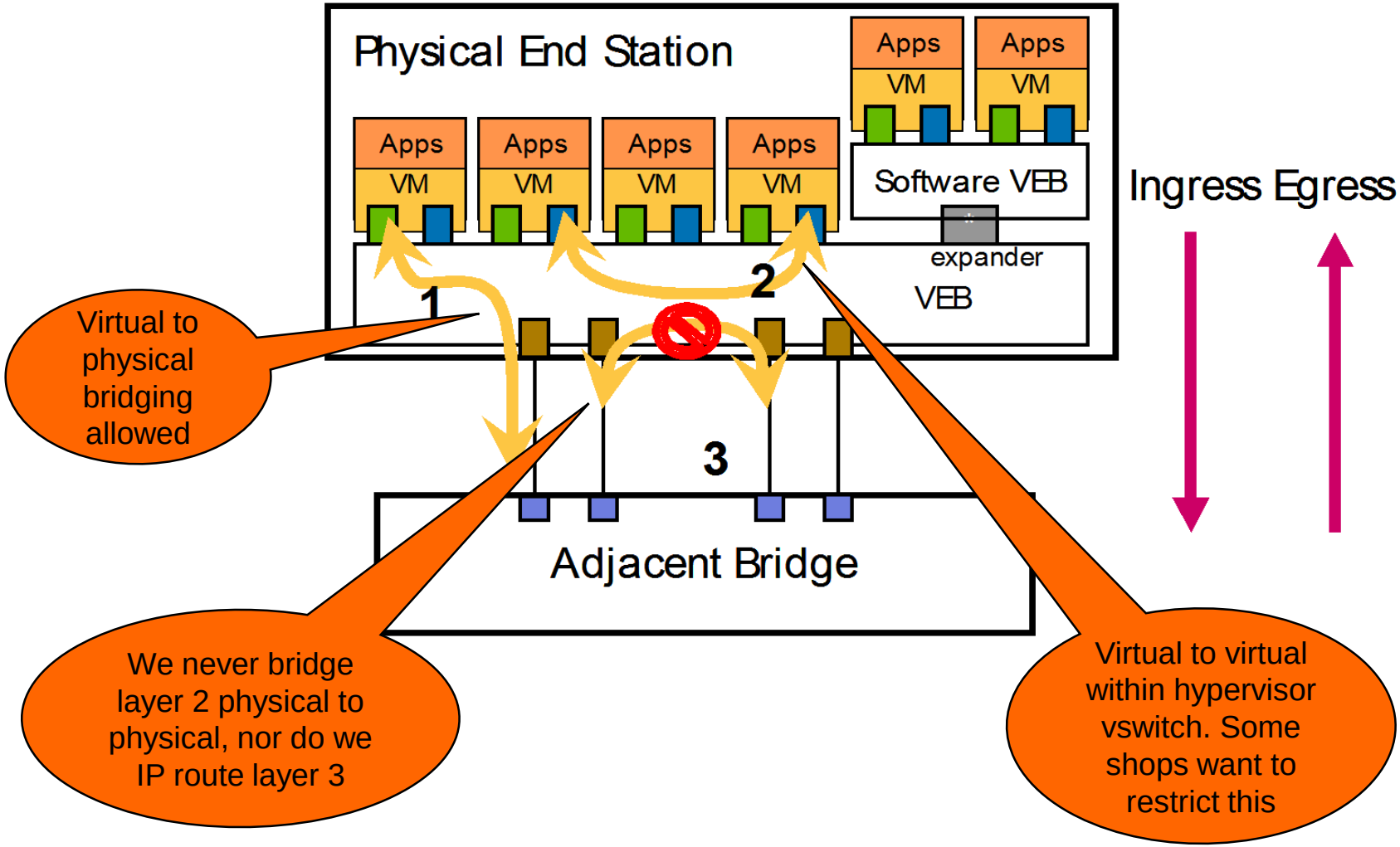
  net to SEA--> 902 SEA to virt--> 902 537 <--to net from SEA 537 <--to SEA from virt

  net to SEA--> 1125 SEA to virt--> 1125 620 <--to net from SEA 620 <--to SEA from virt

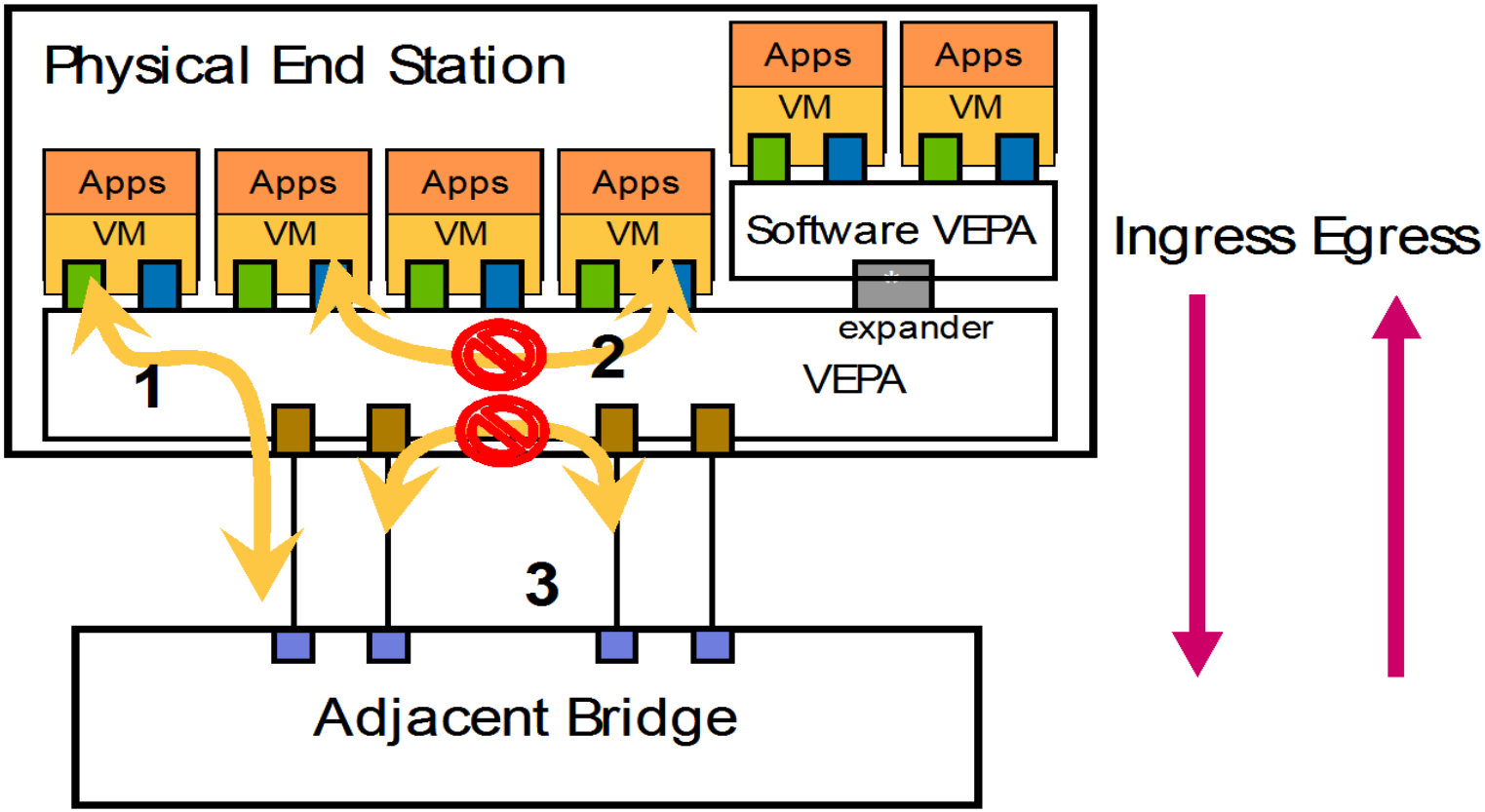
Virtual Switch – VEB versus VEPA mode

- Virtual Ethernet Bridging, VEB mode (what we've always done)
- Virtual Ethernet Port Aggregator, VEPA mode, the ability to isolate LPARs that are on the same subnet. These LPARs can reach peers beyond the SEA, out on the physical subnet, but there is no LPAR to LPAR traffic between these peers, within the hypervisor.
- At HMC 777, and POWER firmware stream 770, we now can specify that a virtual switch is VEB or VEPA.
- You may also see the acronym VSN, Virtual Server Networking. VSI, Virtual Station Interface configuration not required.
- VEPA fits a scenario where different departments in a single machine, are isolated on different VLANs, but ALL departments must join a particular VLAN for, perhaps backup. They should reach peers outside the machine, but must not reach each other within the machine.
- Early development plan was 802.1Qbg, but it was realized that Cisco was implementing 802.1BR. No interoperation.
- POWER development realized we could still implement VEPA in the PowerVM vswitch even though we did not negotiate VEPA / VSI with the physical switch.

Virtual Switch in Virtual Ethernet Bridging (VEB) mode

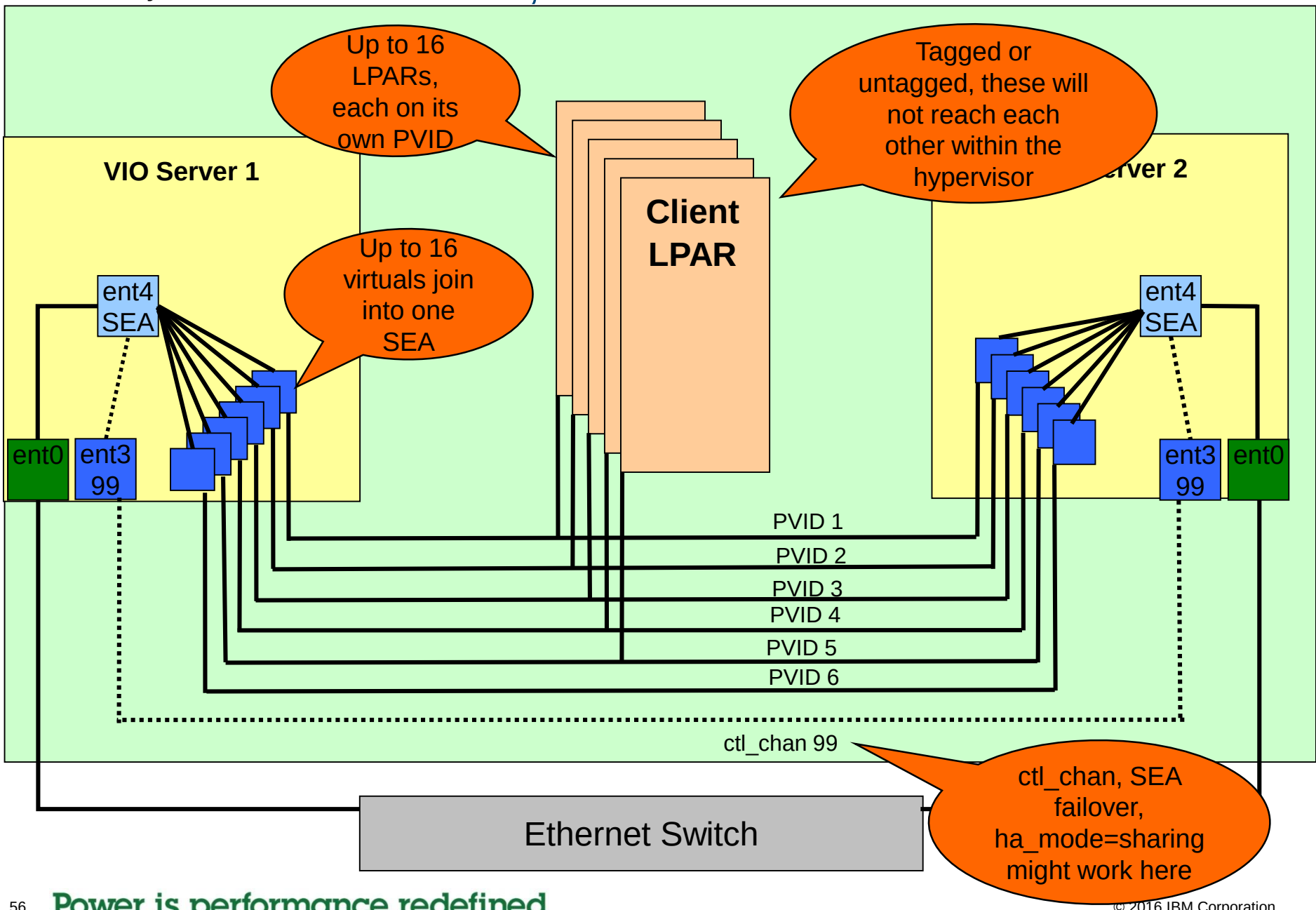


Virtual Switch in Virtual Ethernet Port Aggregation (VEPA) mode



Virtual switch in VEPA mode

Before VEPA, Isolation with VEB mode



VEPA – Server must be VSN Phase 2 Capable

- **hmca62:~ # lssyscfg -r sys -m wiz -F name,state,ipaddr, type_model,serial_num,vsn_phase2_capable,vsi_on_veth_capable**

  wiz,Operating,10.33.5.110,8231-E2B,108854P,1,1

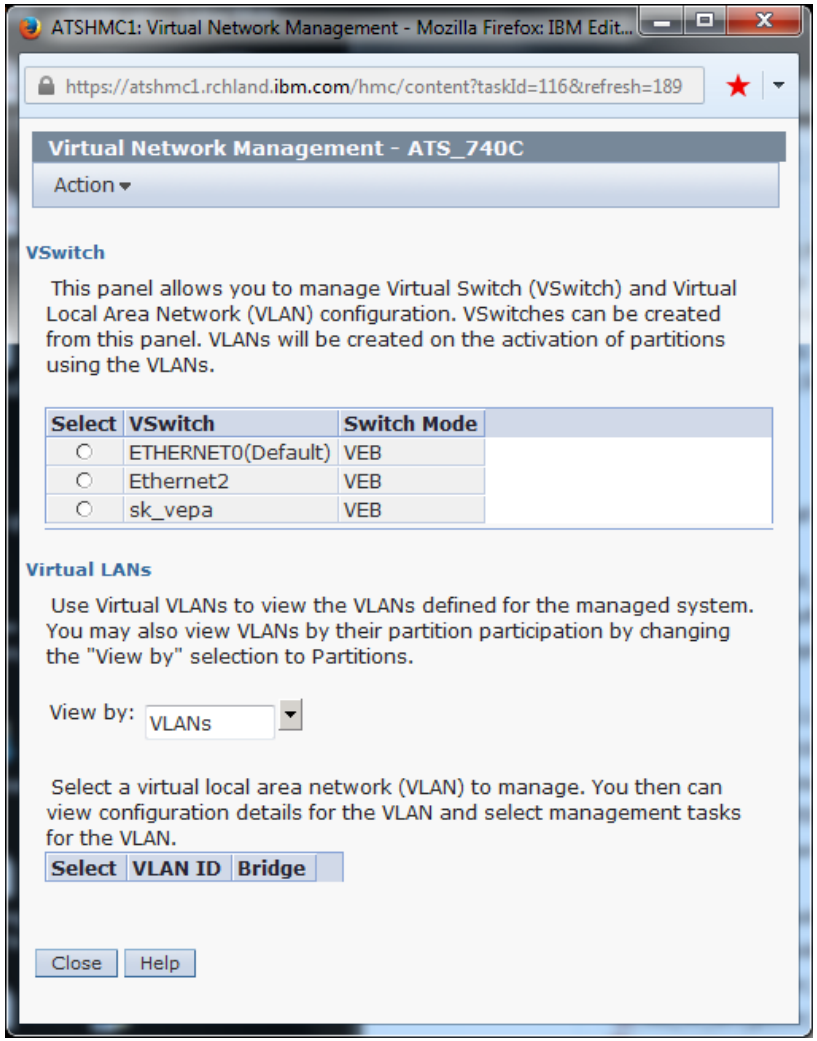
HMC command line or HMC browser GUI

General | Processors | Memory | I/O | Migration | Power-On Parameters | **Capabilities** | Advanced

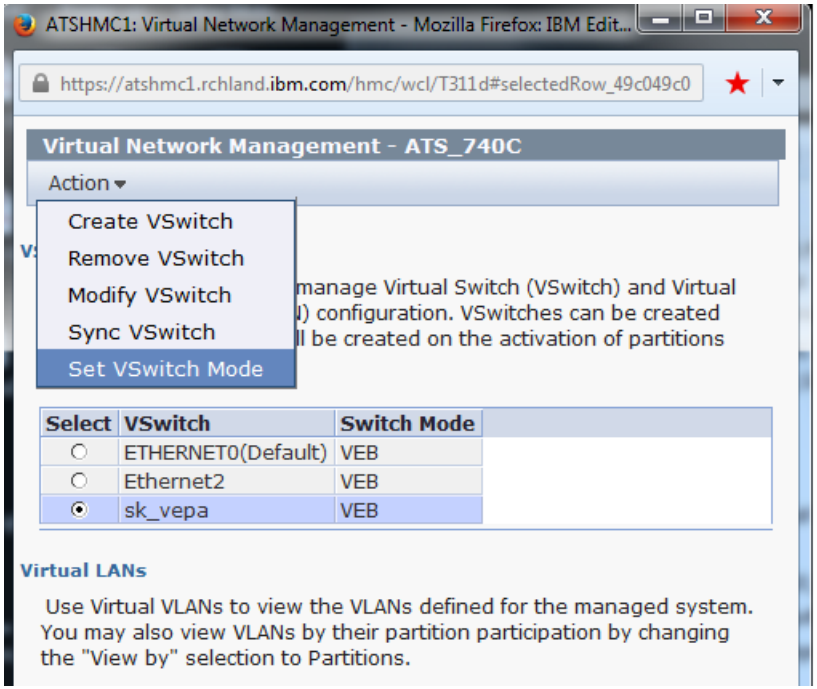
Capability	Value
Hardware Discovery Capable	True
Active Partition Mobility Capable	True
Inactive Partition Mobility Capable	True
IBM i Partition Mobility Capable	True
Partition Processor Compatibility Mode Capable	True
Partition Availability Priority Capable	True
Electronic Error Reporting Capable	True
Active Partition Processor Sharing Capable	True
Firmware Power Saver Capable	True
Hardware Power Saver Capable	True
Virtual Switch Capable	True
Virtual Fibre Channel Capable	True
Active Memory Expansion Capable	True
Partition Suspend Capable	True
Partition Remote Restart Capable	True
Virtual Trusted Platform Module Capable	True
Dynamic Platform Optimization Capable	False
Hardware Memory Encryption Capable	True
Hardware Memory Compression Capable	True
Virtual Server Network Phase 2 Capable	True

OK | Cancel | Help

VEPA – Set Vswitch Mode



- Switches are created in VEB mode. Set VSwitch mode after SEAs are configured

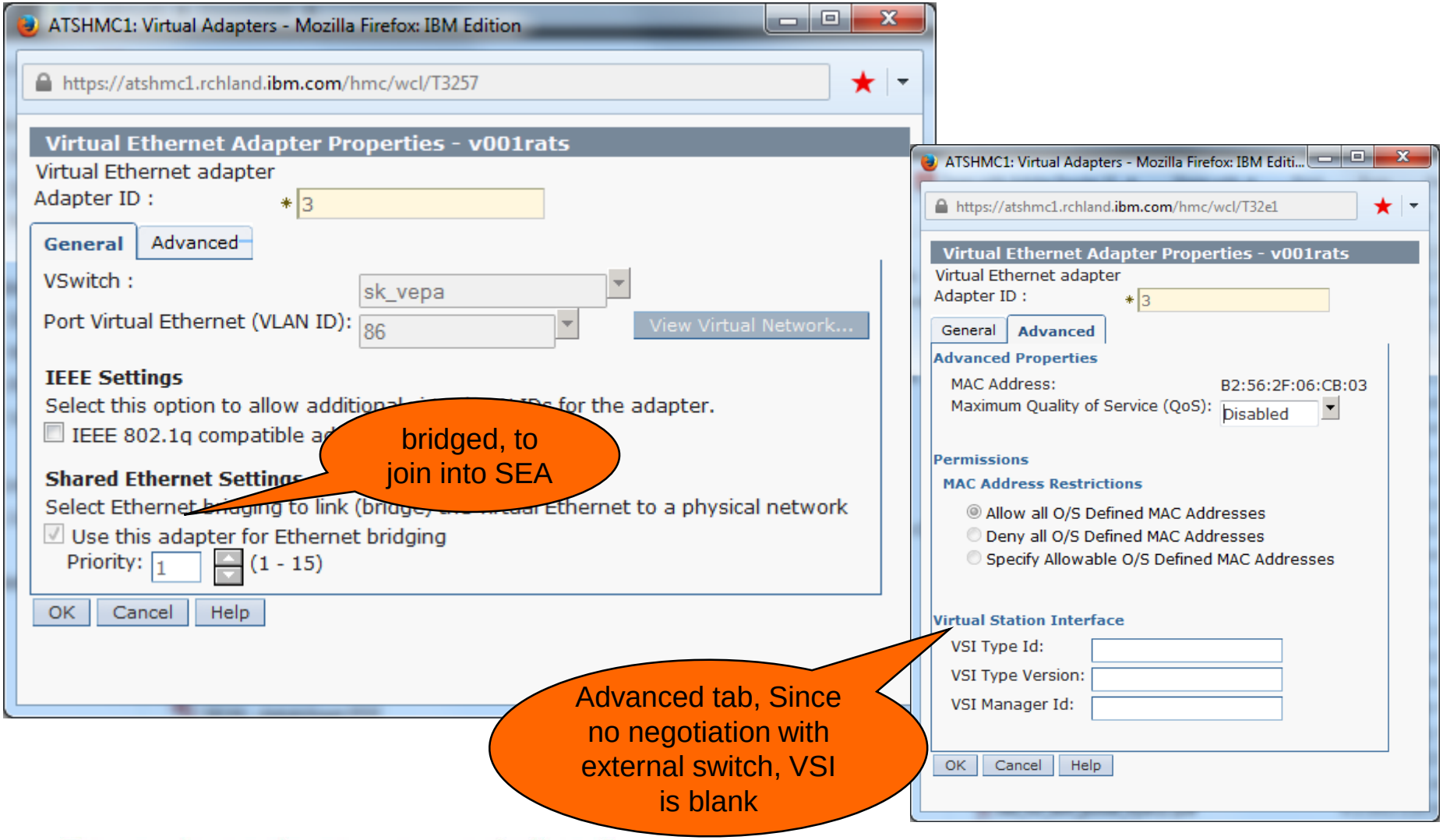


- If browser interface refuses, use command line toggle back and forth VEB - VEPA

chhwres -m <managedserver> -r virtualio --rsubtype vswitch -o s --vswitch sk_vepa -a switch_mode=VEPA --force

VEPA - Virtual Ethernet adapter, no VSI Profile data

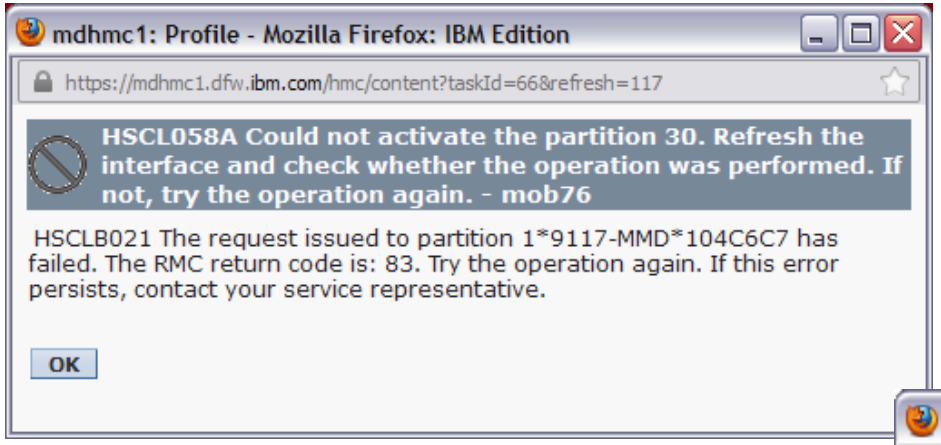
- **Can be configured at LPAR creation, or DLPAR modified**



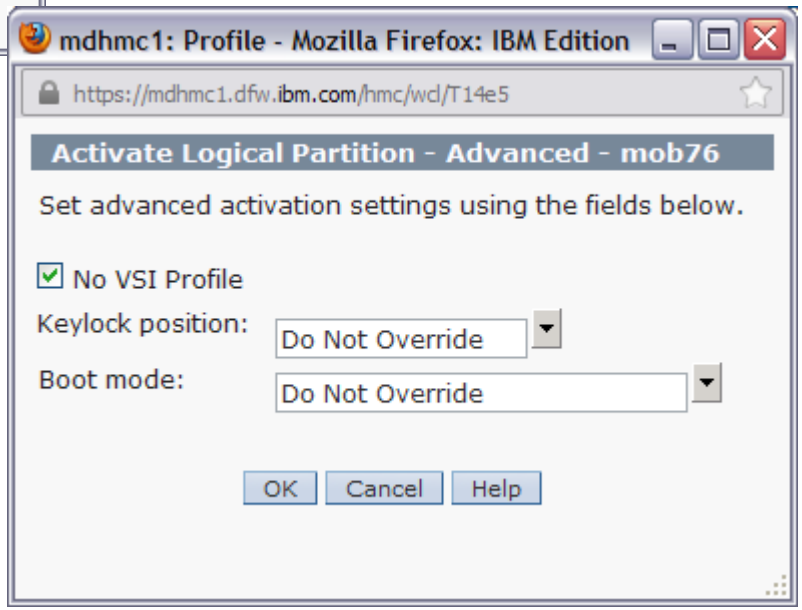
Power is performance redefined

VEPA – No VSI Profile checkbox

- **With Virtual Station Interface blank, you won't receive this error on activate**



- **Go back to activate, and checkbox "No VSI Profile" to bypass your config info**



VEPA – Other configuration effects

- lldpd was already running on the VIO server at 2.2.2.2

```
$ lssrc -s lldpd
Subsystem         Group            PID          Status
 lldpd            tcpip            6750426      active
```

- As root on VIO, you can check if any SEAs are already under lldpctl

```
# lldpctl show portlist
lldpctl: 0812-001 lldpd is currently not managing any ports
```

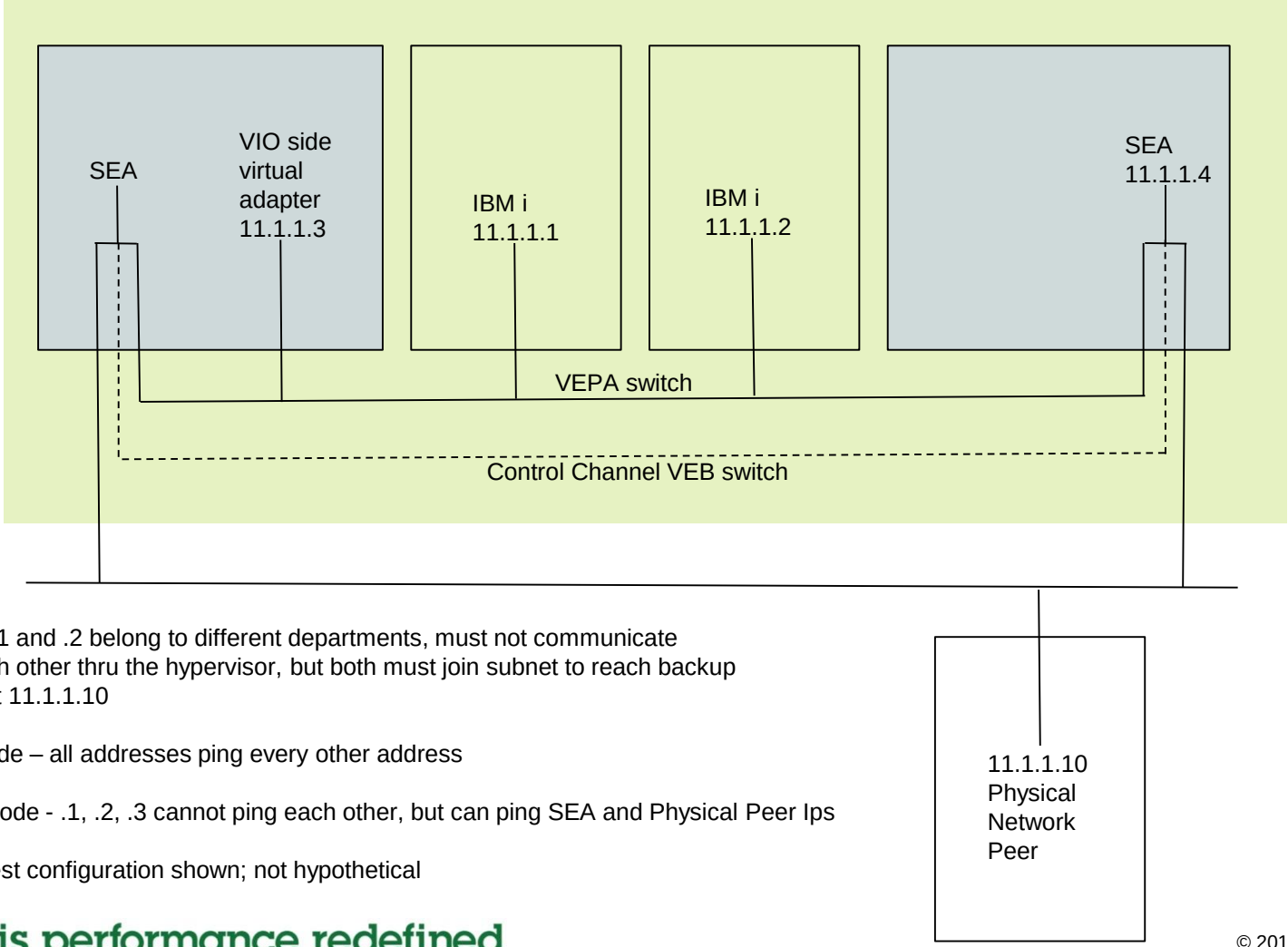
- There is an lldpsvc attribute on the SEA that you create. You will chdev it

```
$ lsdev -dev ent7 -attr | grep lldp
lldpsvc       no       Enable IEEE 802.1qbg services
$ chdev –dev ent7 –atttr lldpsvc=yes
```

- If you ever need to remove this SEA, you must first set lldpsvc back to no.
- The control channel between two VIOs, two SEAs, must NOT attach to the VEPA switch; it must attach to a VEB switch.
- When we were thinking full 802.1Qbg, physical adapter in a VEPA SEA may NOT be link aggregation or EtherChannel. I wonder if link aggregation SEA would in fact work, now that we do not negotiate VSI with physical switch.
- http://pic.dhe.ibm.com/infocenter/powersys/v3r1m5/advanced/content.jsp?topic=/p7hb1/iphb1_config_vsn.htm

VEPA – test scenario

```
chhwres -m <managedserver> -r virtualio --rsubtype vswitch -o s --vswitch sk_vepa -a switch_mode=VEPA --force
chhwres -m <managedserver> -r virtualio --rsubtype vswitch -o s --vswitch sk_vepa -a switch_mode=VEB --force
lshwres –m <managedserver> -r virtualio --rsubtype vswitch
```



LPARs .1 and .2 belong to different departments, must not communicate with each other thru the hypervisor, but both must join subnet to reach backup server at 11.1.1.10

VEB mode – all addresses ping every other address

VEPA mode - .1, .2, .3 cannot ping each other, but can ping SEA and Physical Peer Ips

Actual test configuration shown; not hypothetical

AIX Virtual Ethernet adapter

- **Virtual adapters in AIX**
  # chdev -l ent0 -a dcbflush_local=yes –P (in nim script, before first boot)
  ent0 changed

  You can apply this without -P before you have configured IP on the interface.

Hidden attribute. Not widely known.

- **At 7100-01-01-1141, (also 6100-04-05) we see the mtu_bypass ODM attribute**
  # chdev –l en0 –a mtu_bypass=on
  changes configured interface dynamically, and inserts ODM value; -P not required

- **You may set thread on IP interfaces, and likely leave ndogthreads at the default zero**
  # chdev –l en0 –a thread=on
  en0 changed

Thread - Physical or Virtual Ethernet Interfaces

- You may set thread on IP interfaces, and likely leave ndogthreads at the default zero
  # chdev –l en0 –a thread=on
  en0 changed

- With ndogthreads=0 (default), you get one thread per cpu for the interface

- thread on the interface, works in concert with the ndogthreads setting:

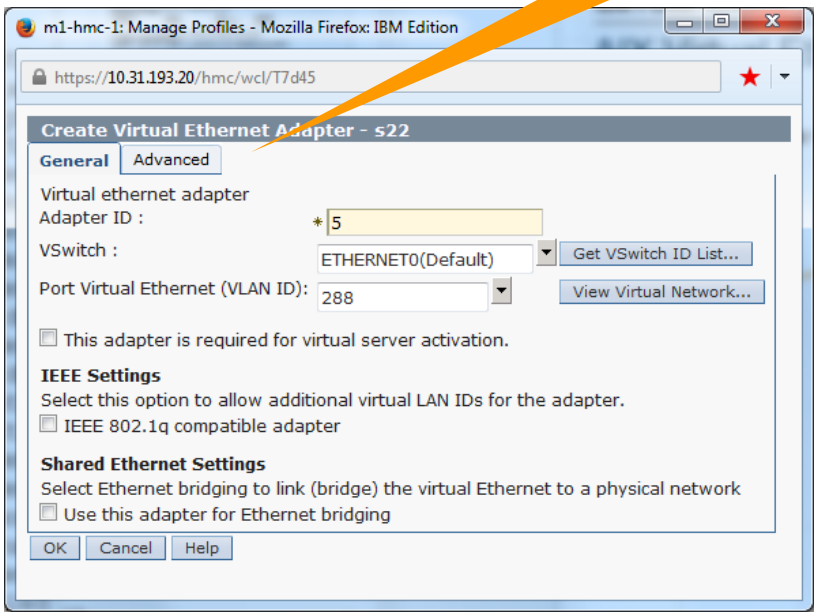
```
# no -h ndogthreads
Help for tunable ndogthreads:
Purpose:
Specifies the number of dog threads that are used during hashing.

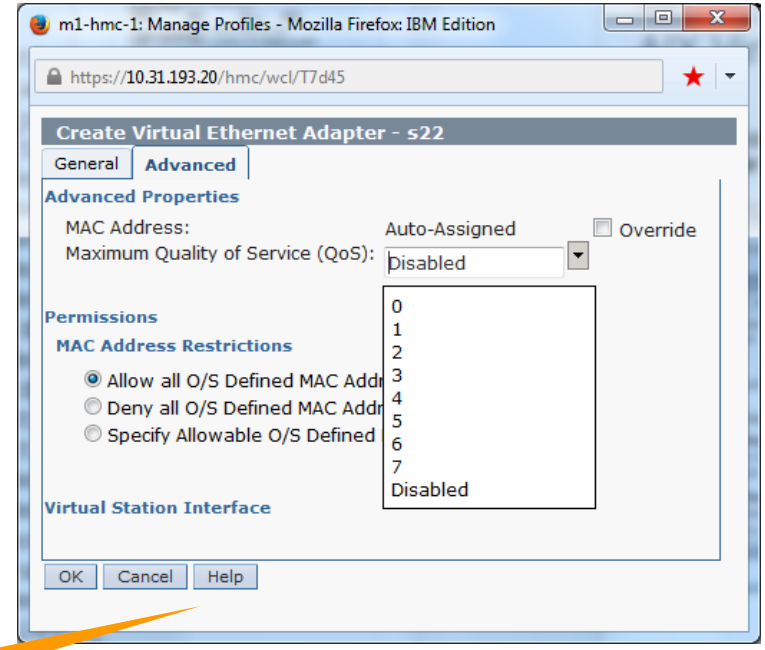
Values:
        Default: 0
        Range: 0 - 1024
        Type: Dynamic
        Unit: numeric
Tuning:
This option is valid only if dog threads are enabled for an interface. A value of 0 sets it to default ie dog
   threads equal to the number of CPUs. Max value is 1024. The minimum of tunable value and the number
   of cpus is taken as the number of dog threads during hashing.
```

AIX Virtual Ethernet adapter Quality of Service

- **When creating a virtual Ethernet adapter for a partition, you can specify Quality of Service on the adapter**

Advanced tab

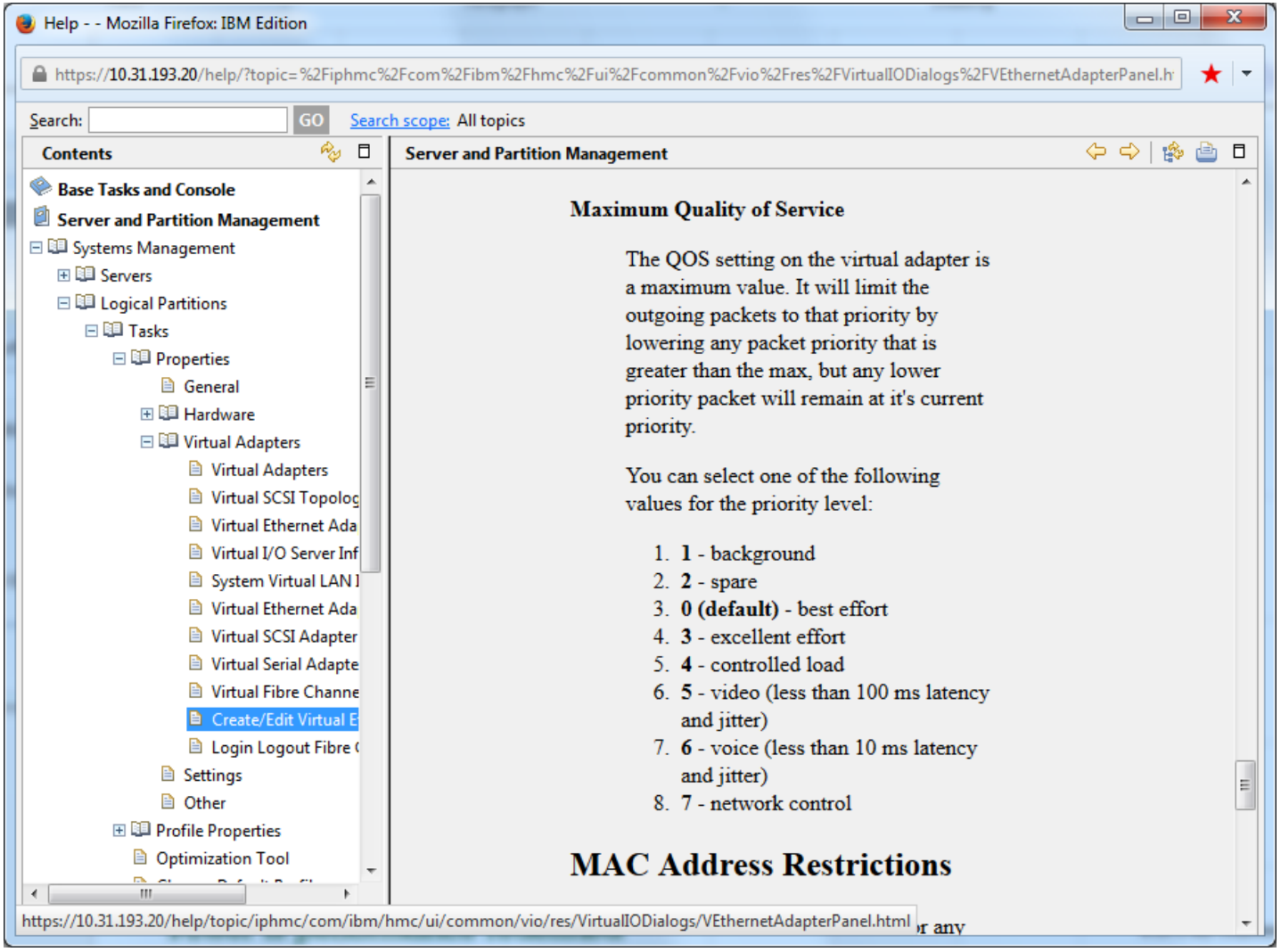




Help text next slide

AIX Virtual Ethernet adapter Quality of Service

- **QoS values**



AIX Virtual Ethernet adapter

If you happen to observe hypervisor send or receive failures…

```
# entstat -d ent0 | grep -i hypervisor
Hypervisor Send Failures: 0
Hypervisor Receive Failures: 4250
```

Only if Support makes you. They don't like to see Max Allocated above Min

- **You could review buffer allocation history on the virtual adapter**

```
# entstat -d ent0
…
…
Receive Information
  Receive Buffers
    Buffer Type              Tiny    Small   Medium    Large     Huge
    Min Buffers               512      512      128       24       24
    Max Buffers              2048     2048      256       64       64
    Allocated                 512      512      128       24       24
    Registered                512      511      128       24       24
    History
      Max Allocated           522     1349      133       29       47
      Lowest Registered       502      502      123       19       19
```

- **Consider increasing minimum tiny and minimum small to a level above Max Allocated**

```
# chdev -l ent0 -a min_buf_tiny=1024 -P
# chdev -l ent0 -a min_buf_small=2048 -P
```

Default TCP settings

Default AIX TCP settings are usually sufficient

Remember, Interface specific network options isno on by default. What you see with ifconfig is what is in force

```
# no -o use_isno
use_isno = 1

# ifconfig en0
en0:
flags=1e080863,4c0<UP,BROADCAST,NOTRAILERS,RUNNING,SIMPLEX,MULTICAST,GROU
PRT,64BIT,CHECKSUM_OFFLOAD(ACTIVE),LARGESEND,CHAIN>
        inet 9.19.51.148 netmask 0xffffff00 broadcast 9.19.51.255
         tcp_sendspace 262144 tcp_recvspace 262144 rfc1323 1
```

For physical adapters in AIX, tcp_sendspace, tcp_recvspace, rfc1323 may not be at the values shown on the above ifconfig

```
# chdev –l en0 –a tcp_sendspace=262144
# chdev –l en0 –a tcp_recvspace=262144
# chdev –l en0 –a rfc1323=1
```

TCP small packet, chatty conversations

- There are two ways that TCP slows down conversations that send small packets
- Nagle algorithm on sender prevents more than one small packet outstanding – you must wait for small segment to be acknowledged before you may transmit another
- Delayed Acknowledgement on receiver says it may wait up to 200 ms before sending acknowledgement, just In case data arrives on the socket to be transmitted
- TCP does a good job of aggregating small writes to the socket into full size segments, and then transmitting. But if you KNOW you have a small packet, time sensitive application, you can…
- # Ifconfig en0 tcp_nodelay 1 (a sender turning off nagle)
  # chdev –l en0 –a tcp_nodelay=1 (a sender setting nagle off for reboot)
- Do **NOT** set tcp_nodelayack, turning off delayed acknowledgements. Instead of sending 1 ACK for every 6-8 segments received, you will ACK **EVERY** segment, nearly doubling the packet rate on the connection, and using a lot more CPU.
- Optional – no –p –o tcp_nagle_limit=0 (or 1), no –p –o tcp_nagleoverride=1 (but remember, the isno settings should make these unnecessary)

TCP small packet, chatty conversations

- What if you make the changes on the previous slide, and see no difference? Your sockets based application may ALREADY be setting these options on the socket. Unless you are editing and compiling the source code, you don't control this

- int on=1;
  setsockopt(s, IPPROTO_TCP, TCP_NODELAY, &on, sizeof(on));
  http://publib.boulder.ibm.com/infocenter/pseries/v5r3/topic/com.ibm.aix.commtechref/doc/commtrf2/setsockopt.htm

Default NFS Settings

- Default AIX NFS settings are usually sufficient

```
# nfso -F -a | egrep "threads|socketsize“
          nfs_max_threads = 3891
           nfs_socketsize = 600000
       nfs_tcp_socketsize = 600000
        statd_max_threads = 50
```

- AIX NFS client mount options

  dio – direct io, bypass AIX caching of file pages written to NFS server (think Oracle rman backups to NAS). Reduces memory demand in AIX, reduces lrud running, reduces scans and frees, but it is **not** faster. Also be aware, this turns off readahead. If you ever had to restore from the same NAS, umount, and mount without dio

  biods=n AIX 53 defaulted to 4 biods per NFS mount, not sufficient. AIX 61, 71 default to 32 biods per NFS mount, usually sufficient.

- Do **not** expect NFS throughput to be close to what you measure at the TCP layer.

Recent NFS results – 1GB (8 Gb) file

3 simultaneous writes, AIX-SEA-SEA-AIX	**~120 sec**
3 simultaneous writes, AIX-SEA jumbo-SEA jumbo-AIX	~60 sec
3 simultaneous writes, AIX-AIX hypervisor	~20 sec
Single ftp write	~5.4 sec
single NFS sum (read) file	~5.8 sec
Single NFS cp (read) file	~2.9 sec
Single scp write (encryption)	~24 sec

Binary ftp with dd input, for network bandwidth

- The test is from AIX 5L Practical Performance Tools and Tuning Guide
  http://www.redbooks.ibm.com/abstracts/sg246478.html?Open

- To test ftp bandwidth between two peers, start with a .netrc file in one user's home directory like this:

```
 # cat ./.netrc
machine mob26.dfw.ibm.com login root password roots_password
macdef init
bin
put "|dd if=/dev/zero bs=8k count=2097152" /dev/null
quit
```

(note blank line in the file, after quit. chmod 700 .netrc)

- **One caveat: In the days of 10Gb Ethernet, a single FTP will NOT fill the pipe.**

- **You may real interest in what a single TCP connection can achieve.**

- **But it is not a viable test of total bandwidth available.**

Binary ftp with dd input for network bandwidth

- **Now, repeatedly send an 16GB file to the peer machine**

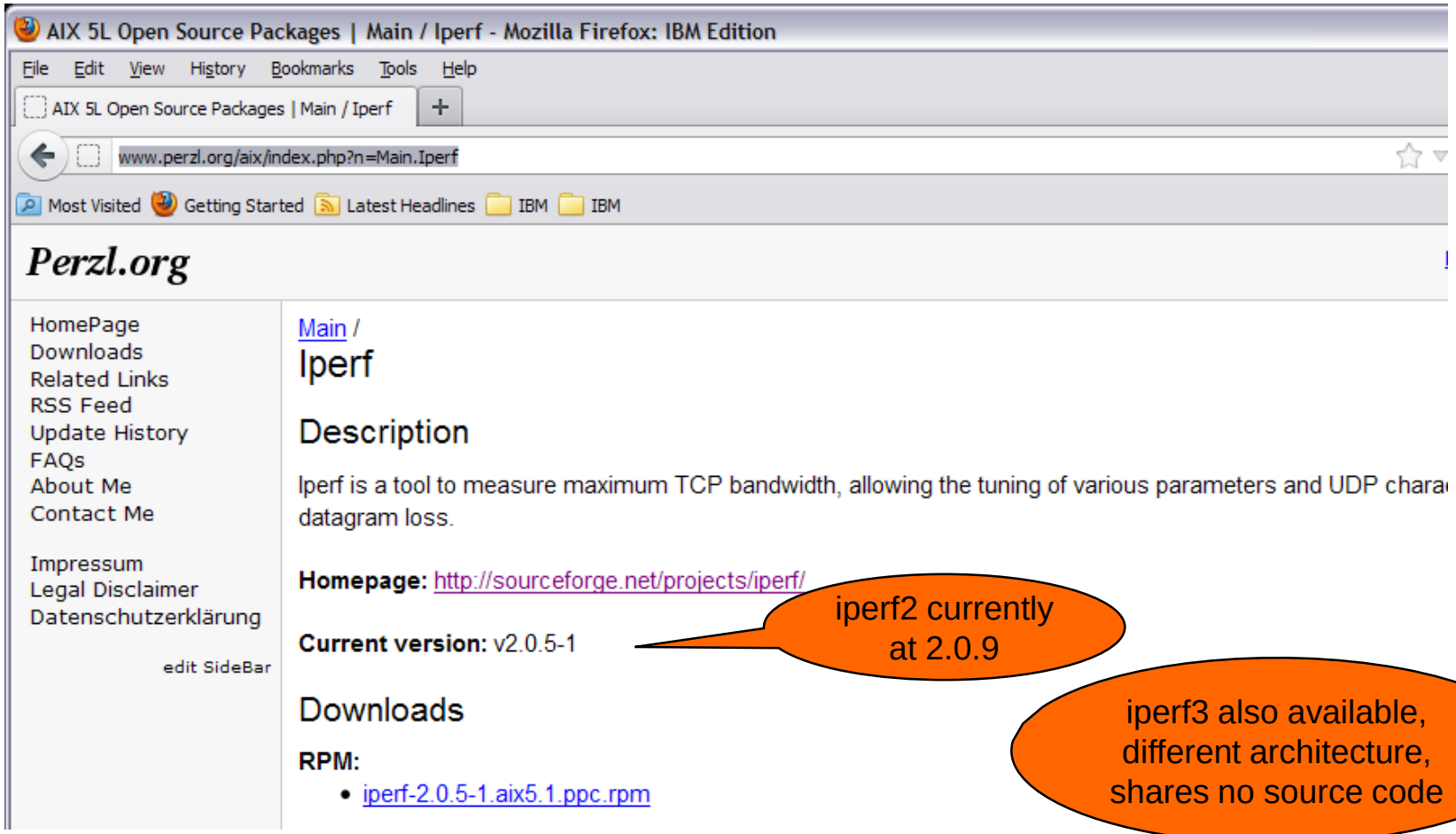
```
# while true
do
ftp mob26.dfw.ibm.com
done
```

- **Connected to mob26.dfw.ibm.com.**

```
220 mob26.dfw.ibm.com FTP server (Version 4.2 Wed Dec 23 11:06:15 CST 2009) read
y.
331 Password required for root.
230-Last unsuccessful login: Tue May  3 08:49:32 2011 on /dev/pts/0 from sig-9-6
5-204-36.mts.ibm.co
230-Last login: Thu May 26 17:17:15 2011 on ftp from ams28.dfw.ibm.com
230 User root logged in.
bin
200 Type set to I.
put "|dd if=/dev/zero bs=8k count=2097152" /dev/null
200 PORT command successful.
150 Opening data connection for /dev/null.
2097152+0 records in.
2097152+0 records out.
226 Transfer complete.
17179869184 bytes sent in 44.35 seconds (3.783e+05 Kbytes/s)
local: |dd if=/dev/zero bs=8k count=2097152 remote: /dev/null
quit
221 Goodbye.
ctl-c to quit.
```

iperf as alternative to ftp with dd

- **Google "iperf aix"**
- **http://www.perzl.org/aix/index.php?n=Main.Iperf**
- **(http://rpmfind.net/linux/rpm2html/search.php?query=iperf for linux)**



iperf server side

Actually, ifconfig shows what is truly in force

- **root@sq08.dfw.ibm.com / # iperf –s**

```
------------------------------------------------------------
Server listening on TCP port 5001
TCP window size: 16.0 KByte (default)
------------------------------------------------------------
[  4] local 9.19.51.90 port 5001 connected with 9.19.51.115 port 46393
[ ID] Interval       Transfer     Bandwidth
[  4]  0.0-10.0 sec  8.36 GBytes  7.17 Gbits/sec
[  4] local 9.19.51.90 port 5001 connected with 9.19.51.115 port 46396
[  5] local 9.19.51.90 port 5001 connected with 9.19.51.115 port 46397
[  4]  0.0-10.0 sec  6.01 GBytes  5.16 Gbits/sec
[  5]  0.0-10.0 sec  6.02 GBytes  5.17 Gbits/sec
[SUM]  0.0-10.0 sec  12.0 GBytes  10.3 Gbits/sec
[  4] local 9.19.51.90 port 5001 connected with 9.19.51.115 port 46399
[  5] local 9.19.51.90 port 5001 connected with 9.19.51.115 port 46400
[  6] local 9.19.51.90 port 5001 connected with 9.19.51.115 port 46401
[  4]  0.0-10.1 sec  4.78 GBytes  4.05 Gbits/sec
[  5]  0.0-10.1 sec  4.66 GBytes  3.95 Gbits/sec
[  6]  0.0-10.1 sec  4.88 GBytes  4.14 Gbits/sec
[SUM]  0.0-10.1 sec  14.3 GBytes  12.1 Gbits/sec

^Croot@sq08.dfw.ibm.com / # ifconfig en0
en0:
flags=1e080863,4c0<UP,BROADCAST,NOTRAILERS,RUNNING,SIMPLEX,MULTICAST,GRO
PRT,64BIT,CHECKSUM_OFFLOAD(ACTIVE),LARGESEND,CHAIN>
        inet 9.19.51.90 netmask 0xffffff00 broadcast 9.19.51.255
         tcp_sendspace 262144 tcp_recvspace 262144 rfc1323 1
```

Single thread, 2 threads, 3 threads. LPAR to LPAR, within machine

iperf client side

- **root@fahr / # iperf -c sq08**

```
------------------------------------------------------------
Client connecting to sq08, TCP port 5001
TCP window size:  256 KByte (default)
------------------------------------------------------------
[  3] local 9.19.51.115 port 46393 connected with 9.19.51.90 port 5001
[ ID] Interval       Transfer     Bandwidth
[  3]  0.0-10.0 sec  8.36 GBytes  7.18 Gbits/sec
root@fahr / # iperf -c sq08 -P 2
------------------------------------------------------------
Client connecting to sq08, TCP port 5001
TCP window size:  256 KByte (default)
------------------------------------------------------------
[  4] local 9.19.51.115 port 46397 connected with 9.19.51.90 port 5001
[  3] local 9.19.51.115 port 46396 connected with 9.19.51.90 port 5001
[ ID] Interval       Transfer     Bandwidth
[  4]  0.0-10.0 sec  6.02 GBytes  5.17 Gbits/sec
[  3]  0.0-10.0 sec  6.01 GBytes  5.16 Gbits/sec
[SUM]  0.0-10.0 sec  12.0 GBytes  10.3 Gbits/sec
root@fahr / # iperf -c sq08 -P 3
------------------------------------------------------------
Client connecting to sq08, TCP port 5001
TCP window size:  256 KByte (default)
------------------------------------------------------------
[  3] local 9.19.51.115 port 46401 connected with 9.19.51.90 port 5001
[  4] local 9.19.51.115 port 46399 connected with 9.19.51.90 port 5001
[  5] local 9.19.51.115 port 46400 connected with 9.19.51.90 port 5001
[ ID] Interval       Transfer     Bandwidth
[  3]  0.0-10.0 sec  4.88 GBytes  4.19 Gbits/sec
[  4]  0.0-10.0 sec  4.78 GBytes  4.10 Gbits/sec
[  5]  0.0-10.0 sec  4.66 GBytes  4.01 Gbits/sec
[SUM]  0.0-10.0 sec  14.3 GBytes  12.3 Gbits/sec
```

Hmm. Correct tcp_recvspace in this case

Single thread

2 threads

3 threads. LPAR to LPAR, within machine

iperf client side continued

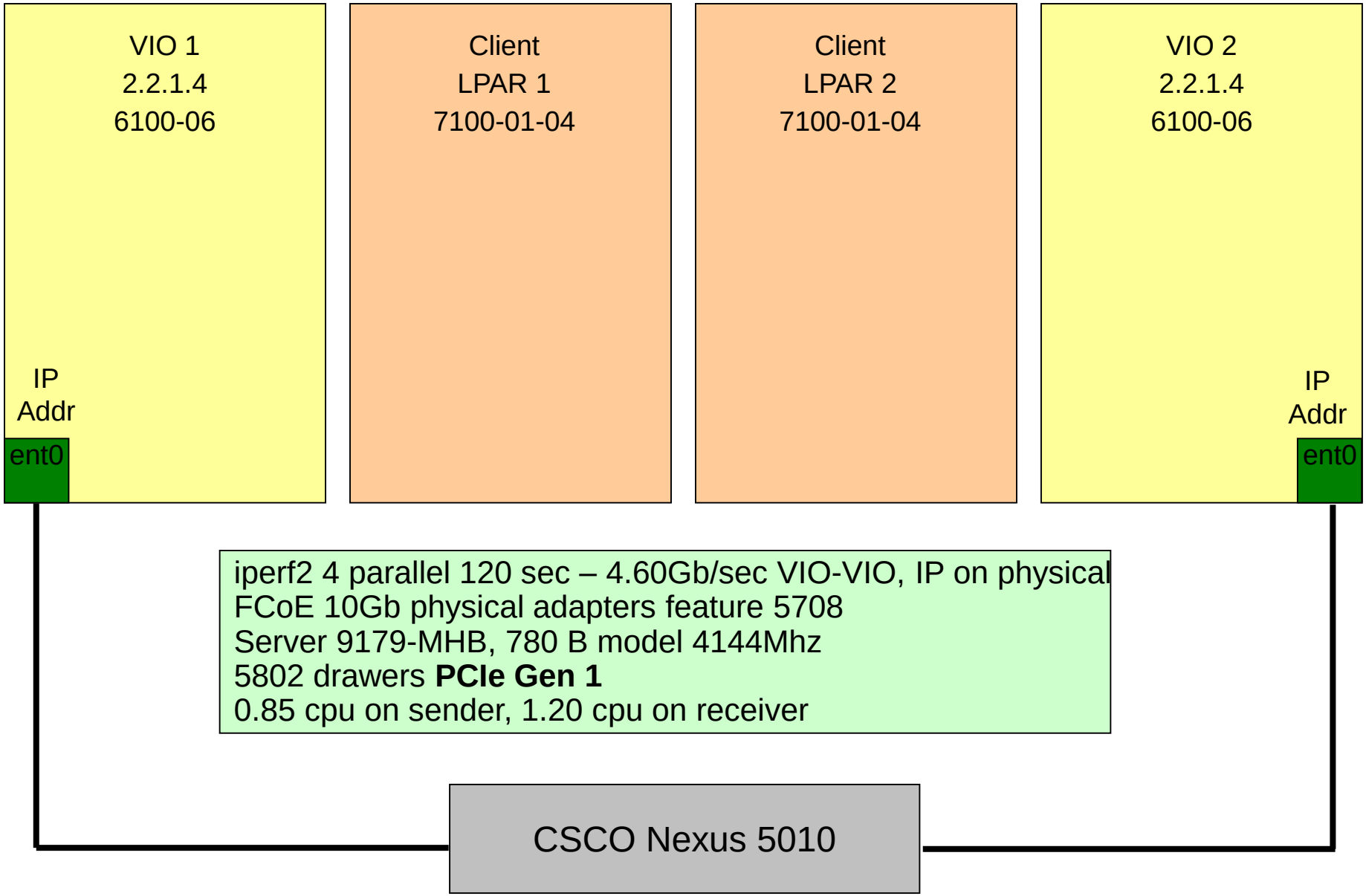
- **Miss any one setting, much lower thruput**

```
root@fahr /export/res # chdev -l en0 -a mtu_bypass=off
en0 changed
root@fahr /export/res # iperf -c sq08 -P 3
------------------------------------------------------------
Client connecting to sq08, TCP port 5001
TCP window size:  256 KByte (default)
------------------------------------------------------------
[  5] local 9.19.51.115 port 46634 connected with 9.19.51.90 port 5001
[  3] local 9.19.51.115 port 46632 connected with 9.19.51.90 port 5001
[  4] local 9.19.51.115 port 46633 connected with 9.19.51.90 port 5001
[ ID] Interval       Transfer     Bandwidth
[  5]  0.0-10.0 sec   455 MBytes   381 Mbits/sec
[  3]  0.0-10.0 sec   452 MBytes   379 Mbits/sec
[  4]  0.0-10.0 sec   482 MBytes   404 Mbits/sec
[SUM]  0.0-10.0 sec  1.36 GBytes  1.16 Gbits/sec
```

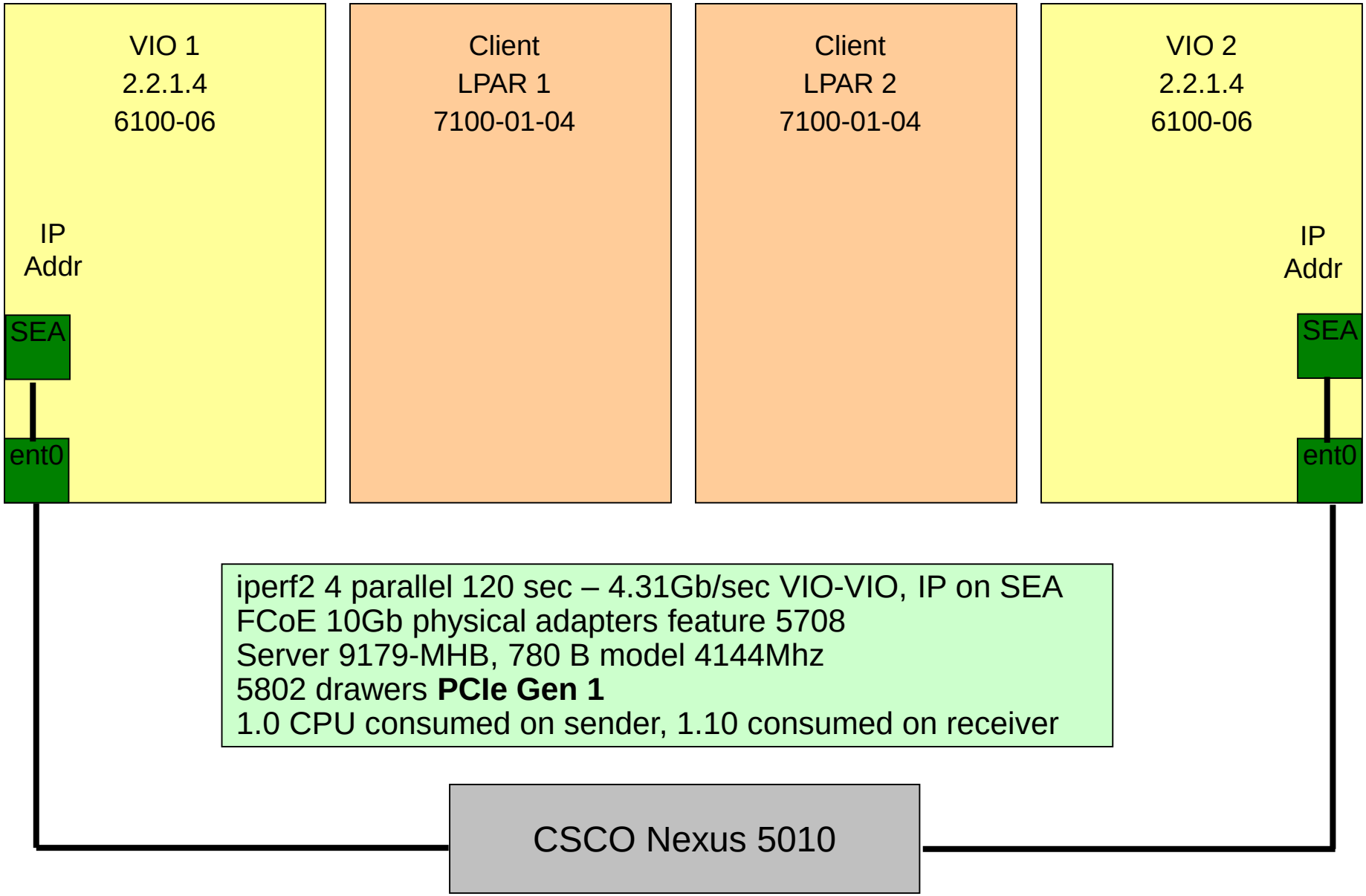
Turning off mtu_bypass

3 threads. LPAR to LPAR, within machine, MUCH LOWER THRUPUT

iperf2 thruput – FCoE adapter

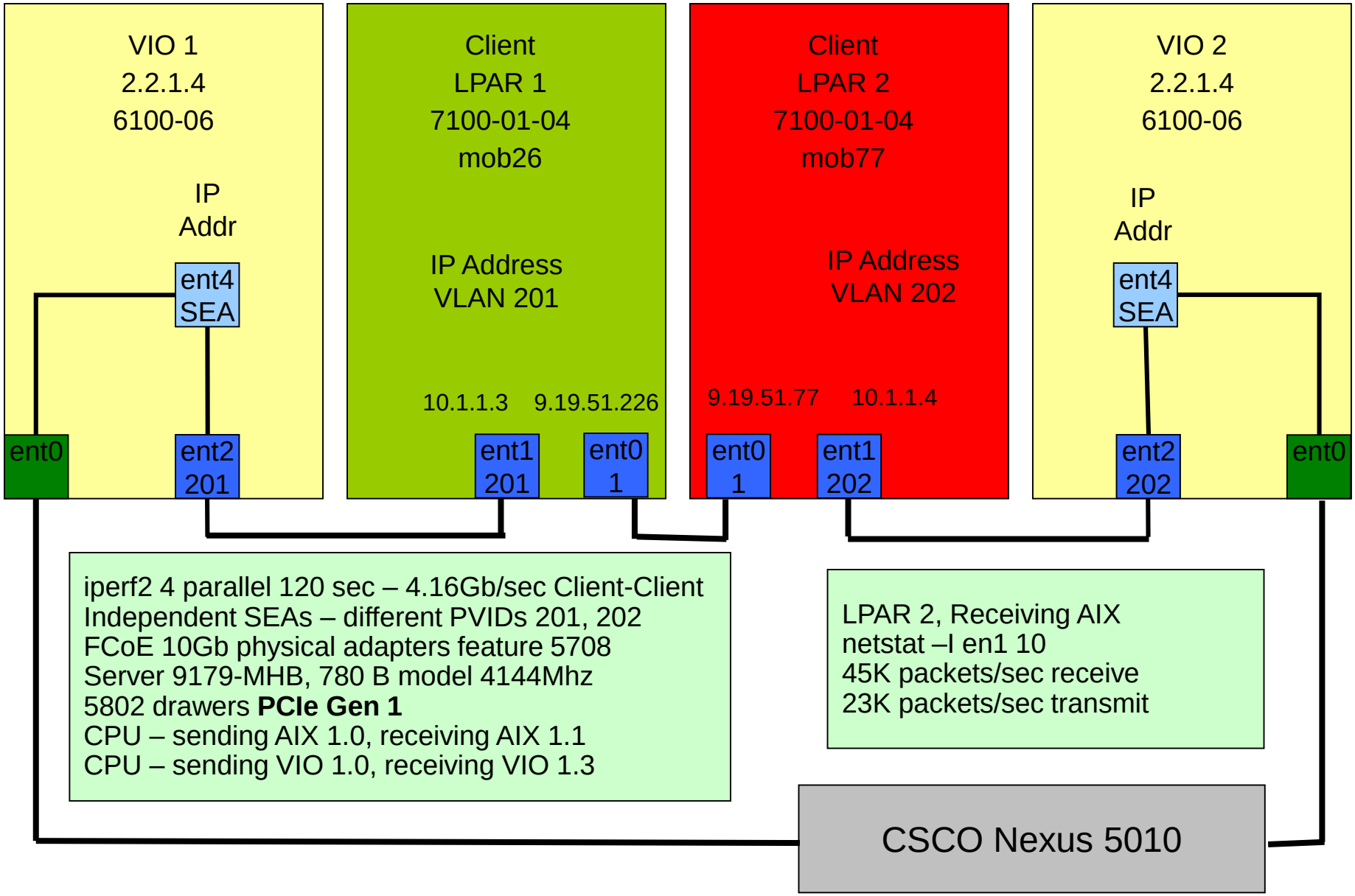


iperf2 thruput – FCoE adapter

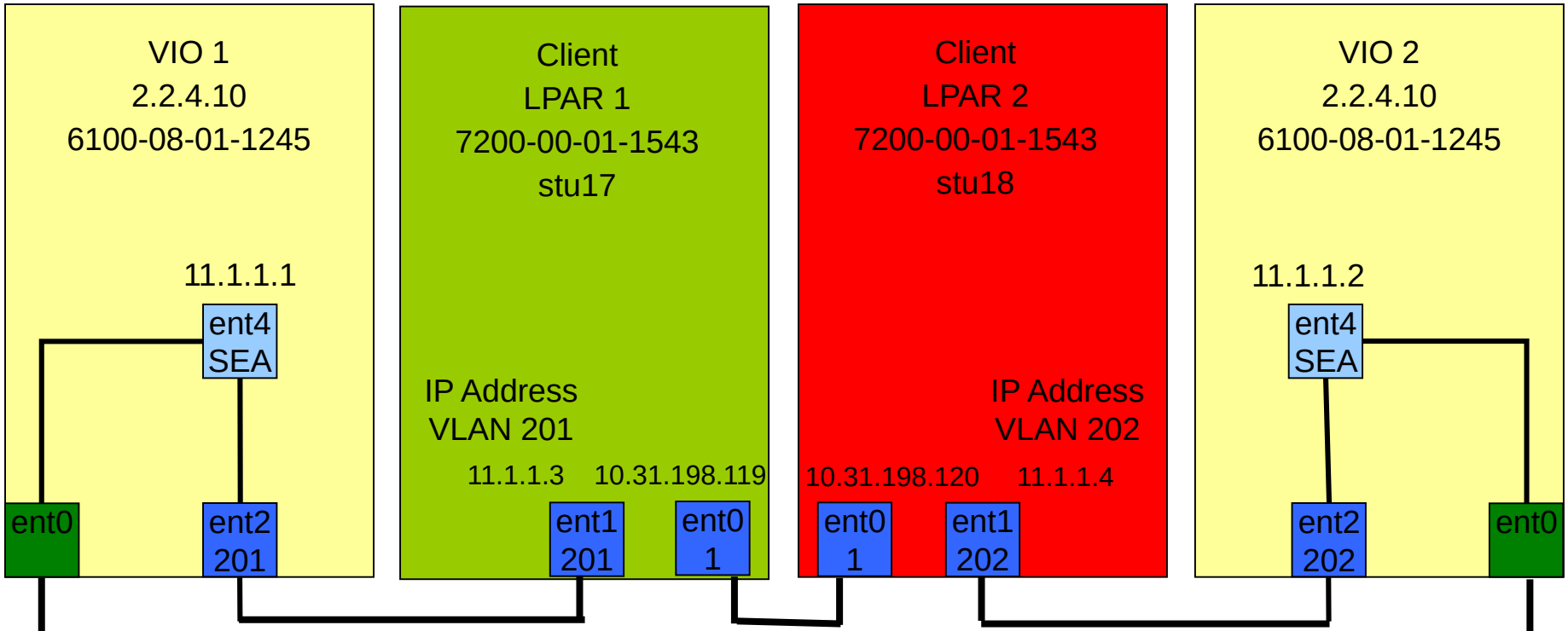


iperf2 4 parallel 120 sec – 4.31Gb/sec VIO-VIO, IP on SEA
FCoE 10Gb physical adapters feature 5708
Server 9179-MHB, 780 B model 4144Mhz
5802 drawers **PCIe Gen 1**
1.0 CPU consumed on sender, 1.10 consumed on receiver

iperf2 thruput – FCoE adapter, and SEA



iperf2 POWER8 – FCoE adapter, and SEA



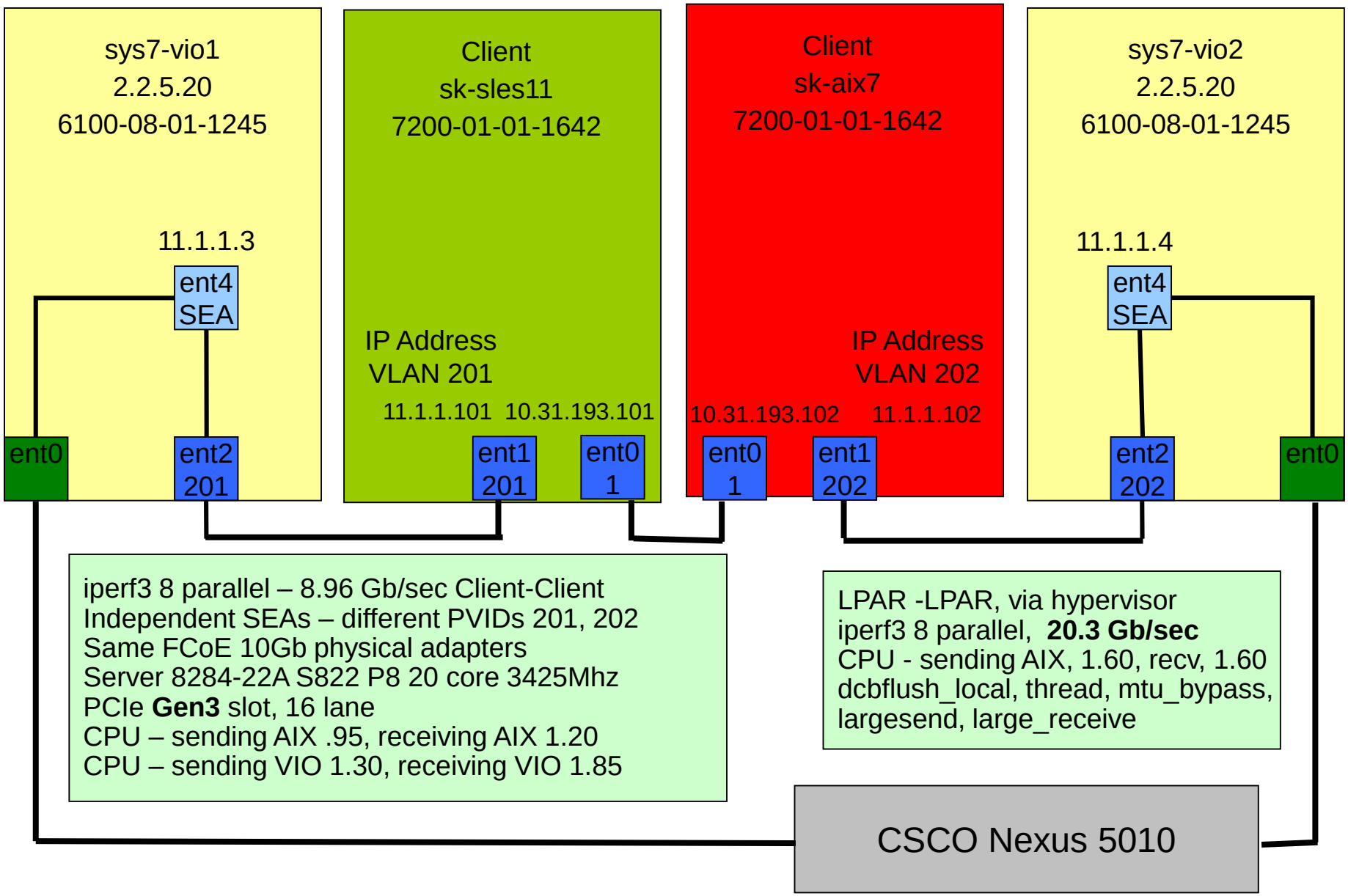
iperf2 8 parallel 30 sec – 8.95 Gb/sec Client-Client
Independent SEAs – different PVIDs 201, 202
Same FCoE 10Gb physical adapters
Server 8284-22A S822 P8 20 core 3425Mhz
PCIe **Gen3** slot, 16 lane
CPU – sending AIX .95, receiving AIX 1.2
CPU – sending VIO 1.3, receiving VIO 1.85

stu17 - stu18, via hypervisor
iperf2 8 parallel, 30 sec - **45Gb/sec**
CPU - sending AIX, 3.55, recv, 3.37
dcbflush_local, thread, mtu_bypass,
largesend, large_receive

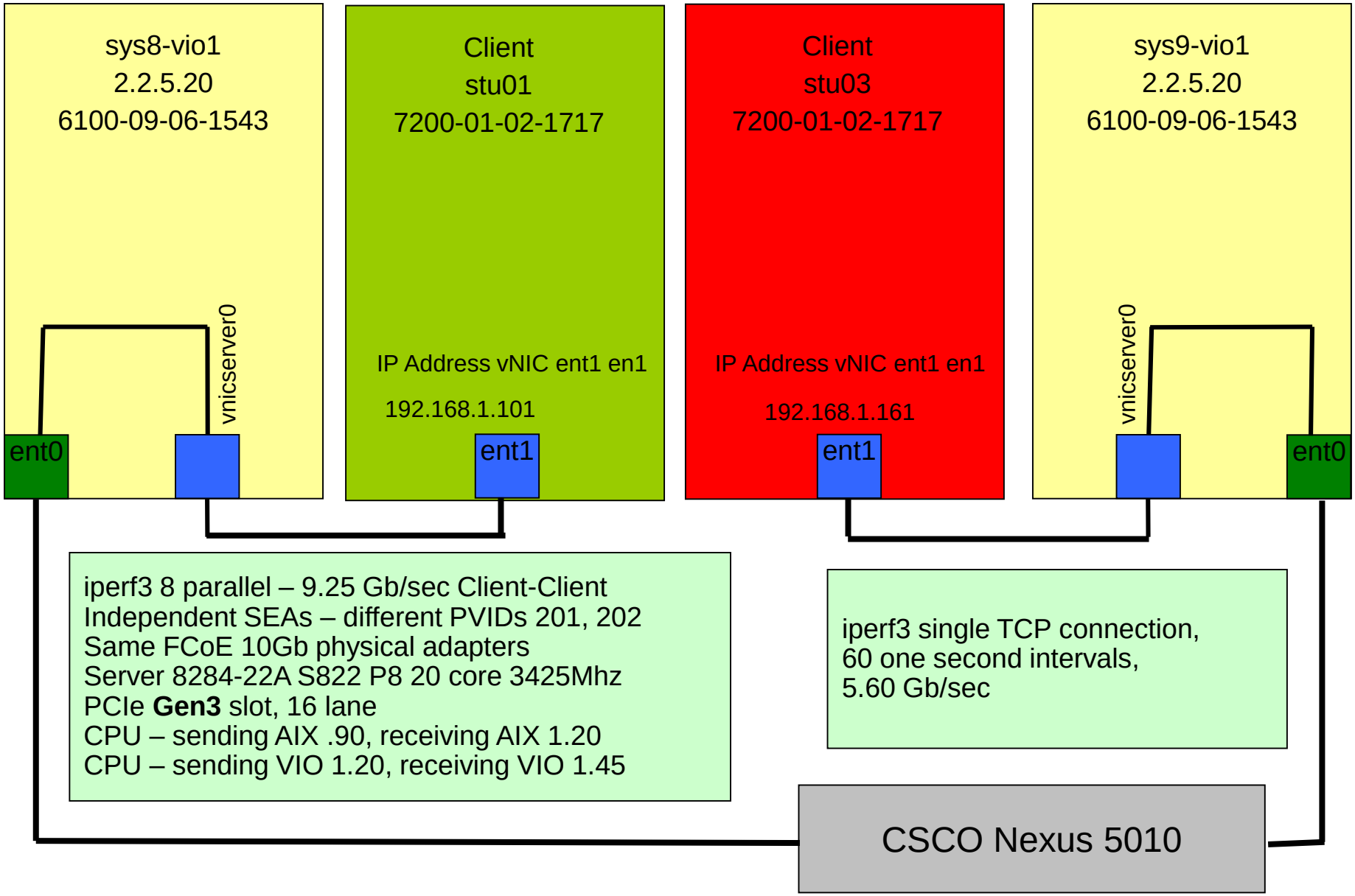
Caution:
20170714, iperf v2, still achieves this. iperf3 will **NOT**. See update next slide

xus 5010

iperf3 POWER8 – FCoE adapter, and SEA 20170714



iperf3 POWER8 – SR-IOV **VNIC** 20170825



Power is performance redefined

iperf2, iperf3, POWER8 – SR-IOV VF 20170905

sys8-vio1
2.2.5.20
6100-09-06-1543

Client
stu01
7200-01-02-1717

IP Address vNIC ent2 en2

172.16.1.101

ent1

Client
stu02
7200-01-02-1717

IP Address vNIC ent0 en0

172.16.1.131

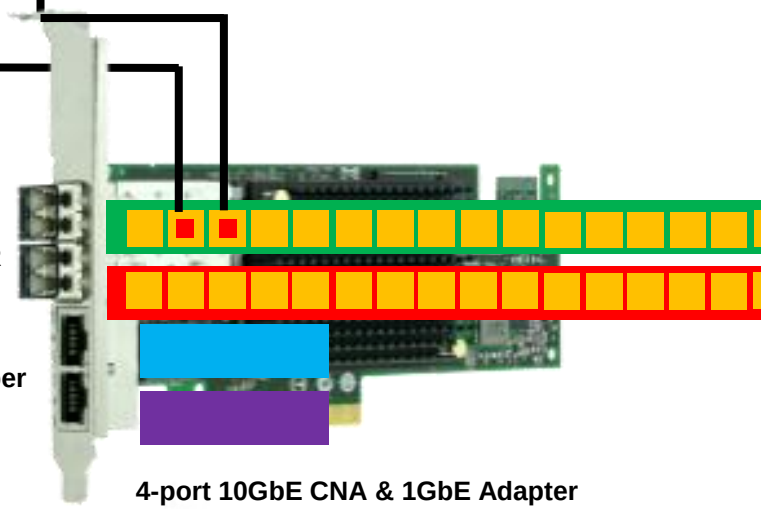
ent1

sys9-vio1
2.2.5.20
6100-09-06-1543

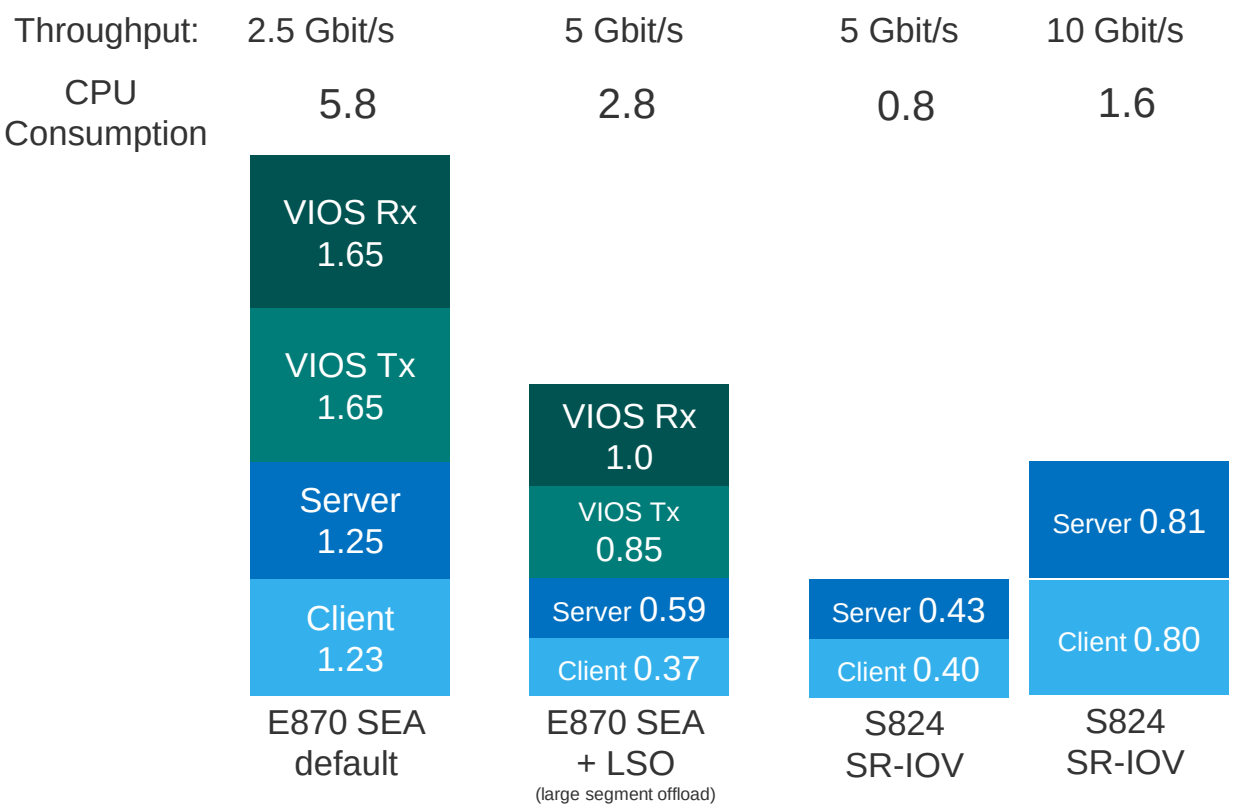
iperf3 8 parallel – 10.1Gb/sec Client-Client, VF
Sending AIX .55, receiving AIX .56

iperf2, 8 parallel - 10.1Gb/sec, Client-Client VF
Sending AIX, .47, receiving AIX .53

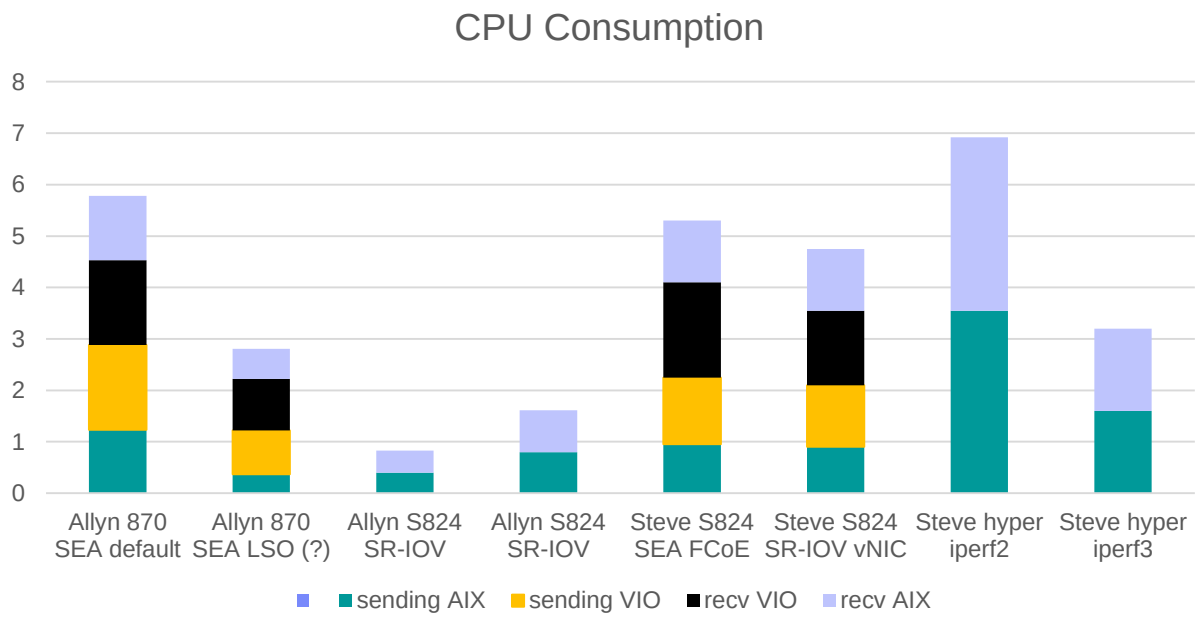
Both VF ports on same physical port of one SR-IOV adapter

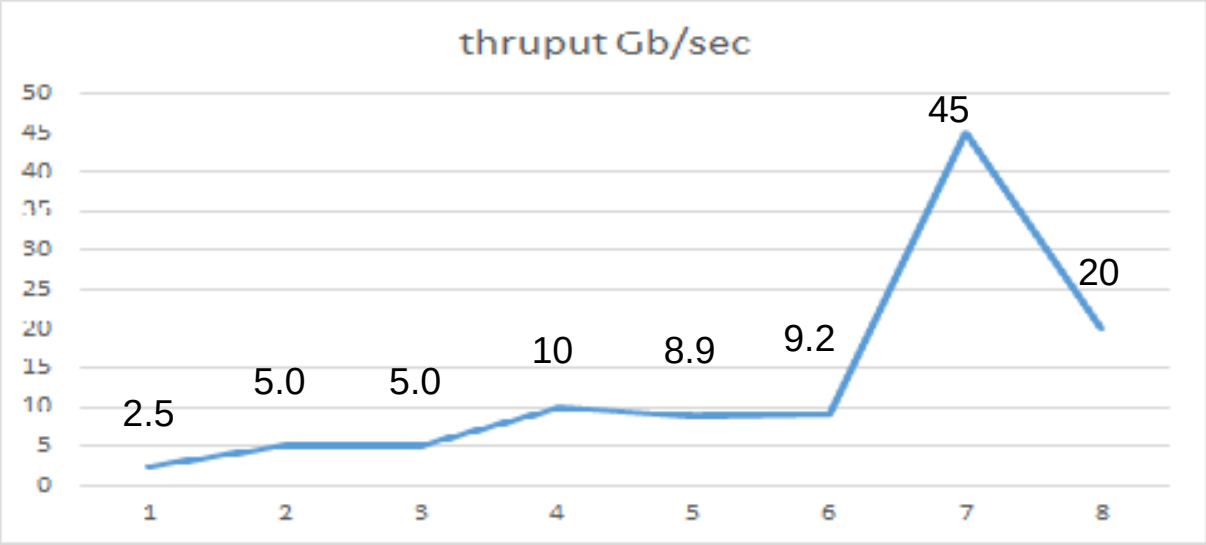


Total CPU Consumption: SEA / SR-IOV



SEA, SR-IOV (VF), and Power8 phyp virtual Ethernet





Power is performance redefined

SR-IOV Virtual Function (VF) configuration in NovaLink

- https://www.ibm.com/support/knowledgecenter/en/POWER8/p8eig/p8eig_cli.htm
- There is not yet a full command line reference for pvmctl
- Error / help text, not completely clear
- Here follows Glen Corneau's excellent configuration of VF in NovaLink

- In the NovaLink partition of a given machine, what SR-IOV adapters do you have?

```
   padmin@sys9-nova:~$ pvmctl  sriov list
SR-IOV Adapters
+----+---------+-------+---------------------------+-----------+--------------+--------+--------
| ID |  State  |  Mode |          Loc Code         | Port Stat |   Port Lbl   | Av Cap | Av LPs
+----+---------+-------+---------------------------+-----------+--------------+--------+--------
| 1  | Running | Sriov | U78CB.001.WZS0FG2-P1-C5-T1 |     Up    | sriov_nexus1 | 90.0%  |   19
|    |         |       | U78CB.001.WZS0FG2-P1-C5-T2 |     Up    | sriov_nexus1 | 90.0%  |   19
|    |         |       | U78CB.001.WZS0FG2-P1-C5-T3 |    Down   |     None     | 100.0% |   4
|    |         |       | U78CB.001.WZS0FG2-P1-C5-T4 |    Down   |     None     | 100.0% |   4
+----+---------+-------+---------------------------+-----------+--------------+--------+--------
```

SR-IOV Virtual Function (VF) configuration in NovaLink

What client LPAR would you like to assign SR-IOV VF to?

```
padmin@sys9-nova:~$ pvmctl vm list
Logical Partitions
+-----------+----+---------------+----------+-----------+---------------+------+-----+-----+
|    Name   | ID |     State     |   RMC    |    Env    |    Ref Code   | Mem  | CPU | Ent |
+-----------+----+---------------+----------+-----------+---------------+------+-----+-----+
| sys9-nova | 1  |    running    |   ----   | AIX/Linux | Linux ppc64le | 2560 |  2  | 0.5 |
|   stu19   | 4  |    running    |  active  | AIX/Linux |               | 2048 |  1  | 0.1 |
|   stu04   | 5  | not activated | inactive | AIX/Linux |    00000000   | 2048 |  1  | 0.1 |
|   stu11   | 12 |    running    |  active  | AIX/Linux |               | 2048 |  1  | 0.1 |
|   stu15   | 15 |    running    |  active  | AIX/Linux |               | 2048 |  1  | 0.1 |
|   stu07   | 16 |    running    |  active  | AIX/Linux |               | 2048 |  1  | 0.1 |
|   stu20   | 17 |    running    |  active  | AIX/Linux |               | 2048 |  1  | 0.1 |
|   stu16   | 18 |    running    |  active  | AIX/Linux |               | 2048 |  1  | 0.1 |
|   stu12   | 19 |    running    |  active  | AIX/Linux |               | 2048 |  1  | 0.1 |
|   stu08   | 20 |    running    |  active  | AIX/Linux |               | 2048 |  1  | 0.1 |
|   stu03   | 25 |    running    |  active  | AIX/Linux |               | 2048 |  2  | 0.4 |
+-----------+----+---------------+----------+-----------+---------------+------+-----+-----+
```

Create VF

```
padmin@sys9-nova:~$ pvmctl ethlp create --loc 'U78CB.001.WZS0FG2-P1-C5-T1'  -p id=25 --capacity=10
[PVME010501AD-0875] Cannot assign physical IO or logical port to partition with ID 25
while remote restart is enabled.

padmin@sys9-nova:~$ pvmctl vm update -s srr_enable=False -i id=25   (and try pvmctl ethlp create again)
```

SR-IOV Virtual Function (VF) configuration in NovaLink

What ethlp configuration do you have now? We have VNIC and VF in stu03

```
padmin@sys9-nova:~$ pvmctl ethlp list
SR-IOV Ethernet Logical Ports (VNIC)
+-------+-----------+-------------------------------+--------+------------+--------+
|  LPAR |    VIOS   |           Loc. Code           | Active | Fail. Pri. |  Cap.  |
+-------+-----------+-------------------------------+--------+------------+--------+
| stu03 | sys9-vio1 | U78CB.001.WZS0FG2-P1-C5-T1-S1 |   *    |     50     | 10.00% |
|       | sys9-vio2 | U78CB.001.WZS0FG2-P1-C5-T2-S2 |        |     50     | 10.00% |
+-------+-----------+-------------------------------+--------+------------+--------+
SR-IOV Ethernet Logical Ports (Non-VNIC)
+-------------------------------+------+-----------+--------+--------------+
|           Loc. Code           | PVID | Partition |  Cap.  |     MAC      |
+-------------------------------+------+-----------+--------+--------------+
| U78CB.001.WZS0FG2-P1-C5-T1-S3 |  0   |   stu03   | 10.00% | 2E7627AFC248 |
+-------------------------------+------+-----------+--------+--------------+


What if you are not NovaLink?  Configure SR-IOV VNIC, VF in HMC enhanced interface.
```

SR-IOV Virtual Function (VF) configuration in NovaLink

What does this look like in AIX? You may have to run cfgmgr for VF to come Available

```
root@stu03 / # lsdev -Cc adapter |grep ent[0-2]
ent0   Available       Virtual NIC Client Adapter (vnic)
ent1   Available       Virtual I/O Ethernet Adapter (l-lan)
ent2   Available 00-00 PCIe2 10GbE SFP+ SR 4-port Converged Network Adapter VF (df1028e214100f04)

root@stu03 / # chdev -l en2 -a mtu_bypass=on
en2 changed
root@stu03 / # chdev -l en2 -a thread=on
en2 changed
```

Shared Ethernet Adapter Linux effects

- For a long time, Linux on POWER did not handle large_receive on the SEA
- Recent success 3Q 2016 in SLES 11 SP4 (Think SAP HANA). The full "cheatsheet" for AIX and SLES 11 follows here

- 1) before SEA is configured, put dcbflush_local=yes on the trunked virtual adapters. If SEA is already configured, you may skip this
  $ chdev -dev entX -attr dcbflush_local=yes

- 2) configure SEA. largesend is on the SEA by default, put large_receive on also
  $ chdev -dev entY -attr large_receive=yes

- 3) In AIX client LPARs, before IP is configured, put dcbflush_local on virtual Ethernet adapters. If IP is already configured, you may skip this
  # chdev -l ent0 -a dcbflush_local=yes

- 4) Also in AIX, put thread and mtu_bypass on the interface en0
  # chdev -l en0 -a thread=on
  # chdev -l en0 -a mtu_bypass=on

Shared Ethernet Adapter Linux effects

- 5) For client partitions running SLES 11, start with SP4 - (uname -r 3.0.101-68-ppc64) then update to at least 77, and reboot. Current testing is at 3.0.101-80-ppc64

- 6) on the SLES partition console
  # rmmod ibmveth
  # modprobe ibmveth old_large_send=1
  # ethtool -K eth0 tso on (Do this for every virtual Ethernet adapter in the partition)

- 7) SLES - Verify tso is on
  # ethtool -k eth0
  Offload parameters for eth0:
  rx-checksumming: on
  tx-checksumming: on
  scatter-gather: on
  tcp-segmentation-offload: on
  udp-fragmentation-offload: off
  generic-segmentation-offload: on
  …
  …
- 8) Assure you have enough CPU in sending client partition, sending VIO, receiving VIO, and receiving partition

Shared Ethernet Adapter Linux effects

- 9) SLES 11 changes for network configuration to persist through a reboot

```
# echo "options ibmveth old_large_send=1"  >> /etc/modprobe.d/50-ibmveth.conf
# echo "ETHTOOL_OPTIONS_tso='-K iface tso on' " >> /etc/sysconfig/network/ifcfg-eth0.cfg
# echo "ETHTOOL_OPTIONS_tso='-K iface tso on' " >> /etc/sysconfig/network/ifcfg-eth1.cfg
```

Reading references for the SLES settings:
http://www-01.ibm.com/support/docview.wss?uid=isg3T1024094
https://www.ibm.com/developerworks/community/wikis/home?lang=en#!/wiki/W51a7ffcf4dfd_4b40_9d82_446ebc23c550/page/Taking%20advantage%20of%20networking%20large-send%20large-receive

- 10) What we observed:

SLES 11 ---> hypervisor ---> SLES 11	23.0 Gb/sec iperf, 8 TCP connect, 30 sec
SLES 11 <--- hypervisor <--- SLES 11	23.0 Gb/sec
SLES 11 ---> SEA -- 10Gb net -- SEA ---> SLES 11	8.95 Gb/sec
SLES 11 <--- SEA -- 10Gb net -- SEA <--- SLES 11	8.95 Gb/sec

Notices and Disclaimers

U.S. Government Users Restricted Rights - Use, duplication or disclosure restricted by GSA ADP Schedule Contract with IBM.

Information in these presentations (including information relating to products that have not yet been announced by IBM) has been reviewed for accuracy as of the date of initial publication and could include unintentional technical or typographical errors. IBM shall have no responsibility to update this information. THIS DOCUMENT IS DISTRIBUTED "AS IS" WITHOUT ANY WARRANTY, EITHER EXPRESS OR IMPLIED. IN NO EVENT SHALL IBM BE LIABLE FOR ANY DAMAGE ARISING FROM THE USE OF THIS INFORMATION, INCLUDING BUT NOT LIMITED TO, LOSS OF DATA, BUSINESS INTERRUPTION, LOSS OF PROFIT OR LOSS OF OPPORTUNITY. IBM products and services are warranted according to the terms and conditions of the agreements under which they are provided.

IBM products are manufactured from new parts or new and used parts. In some cases, a product may not be new and may have been previously installed. Regardless, our warranty terms apply."

Any statements regarding IBM's future direction, intent or product plans are subject to change or withdrawal without notice.

Performance data contained herein was generally obtained in a controlled, isolated environments. Customer examples are presented as illustrations of how those customers have used IBM products and the results they may have achieved. Actual performance, cost, savings or other results in other operating environments may vary.

References in this document to IBM products, programs, or services does not imply that IBM intends to make such products, programs or services available in all countries in which IBM operates or does business.

Workshops, sessions and associated materials may have been prepared by independent session speakers, and do not necessarily reflect the views of IBM. All materials and discussions are provided for informational purposes only, and are neither intended to, nor shall constitute legal or other guidance or advice to any individual participant or their specific situation.

It is the customer's responsibility to insure its own compliance with legal requirements and to obtain advice of competent legal counsel as to the identification and interpretation of any relevant laws and regulatory requirements that may affect the customer's business and any actions the customer may need to take to comply with such laws. IBM does not provide legal advice or represent or warrant that its services or products will ensure that the customer is in compliance with any law

Notices and Disclaimers Con't.

Information concerning non-IBM products was obtained from the suppliers of those products, their published announcements or other publicly available sources. IBM has not tested those products in connection with this publication and cannot confirm the accuracy of 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. IBM does not warrant the quality of any third-party products, or the ability of any such third-party products to interoperate with IBM's products. IBM EXPRESSLY DISCLAIMS ALL WARRANTIES, EXPRESSED OR IMPLIED, INCLUDING BUT NOT LIMITED TO, THE IMPLIED WARRANTIES OF MERCHANTABILITY AND FITNESS FOR A PARTICULAR PURPOSE.

The provision of the information contained herein is not intended to, and does not, grant any right or license under any IBM patents, copyrights, trademarks or other intellectual property right.

IBM, the IBM logo, ibm.com, Aspera®, Bluemix, Blueworks Live, CICS, Clearcase, Cognos®, DOORS®, Emptoris®, Enterprise Document Management System™, FASP®, FileNet®, Global Business Services ®, Global Technology Services ®, IBM ExperienceOne™, IBM SmartCloud®, IBM Social Business®, Information on Demand, ILOG, Maximo®, MQIntegrator®, MQSeries®, Netcool®, OMEGAMON, OpenPower, PureAnalytics™, PureApplication®, pureCluster™, PureCoverage®, PureData®, PureExperience®, PureFlex®, pureQuery®, pureScale®, PureSystems®, QRadar®, Rational®, Rhapsody®, Smarter Commerce®, SoDA, SPSS, Sterling Commerce®, StoredIQ, Tealeaf®, Tivoli®, Trusteer®, Unica®, urban{code}®, Watson, WebSphere®, Worklight®, X-Force® and System z® Z/OS, are trademarks of International Business Machines Corporation, registered in many jurisdictions worldwide. Other product and service names might be trademarks of IBM or other companies. A current list of IBM trademarks is available on the Web at "Copyright and trademark information" at: www.ibm.com/legal/copytrade.shtml.

Special notices (cont.)

Revised December 2, 2010

Notes on benchmarks and values

The IBM benchmarks results shown herein were derived using particular, well configured, development-level and generally-available computer systems. Buyers should consult other sources of information to evaluate the performance of systems they are considering buying and should consider conducting application oriented testing. For additional information about the benchmarks, values and systems tested, contact your local IBM office or IBM authorized reseller or access the Web site of the benchmark consortium or benchmark vendor.

IBM benchmark results can be found in the IBM Power Systems Performance Report at http://www.ibm.com/systems/p/hardware/system_perf.html .

All performance measurements were made with AIX or AIX 5L operating systems unless otherwise indicated to have used Linux. For new and upgraded systems, the latest versions of AIX were used. All other systems used previous versions of AIX. The SPEC CPU2006, LINPACK, and Technical Computing benchmarks were compiled using IBM's high performance C, C++, and FORTRAN compilers for AIX 5L and Linux. For new and upgraded systems, the latest versions of these compilers were used: XL C for AIX v11.1, XL C/C++ for AIX v11.1, XL FORTRAN for AIX v13.1, XL C/C++ for Linux v11.1, and XL FORTRAN for Linux v13.1.

For a definition/explanation of each benchmark and the full list of detailed results, visit the Web site of the benchmark consortium or benchmark vendor.

TPC	http://www.tpc.org
SPEC	http://www.spec.org
LINPACK	http://www.netlib.org/benchmark/performance.pdf
Pro/E	http://www.proe.com
GPC	http://www.spec.org/gpc
VolanoMark	http://www.volano.com
STREAM	http://www.cs.virginia.edu/stream/
SAP	http://www.sap.com/benchmark/
Oracle, Siebel, PeopleSoft	http://www.oracle.com/apps_benchmark/
Baan	http://www.ssaglobal.com
Fluent	http://www.fluent.com/software/fluent/index.htm
TOP500 Supercomputers	http://www.top500.org/
Ideas International	http://www.ideasinternational.com/benchmark/bench.html
Storage Performance Council	http://www.storageperformance.org/results

Revised December 2, 2010

Notes on HPC benchmarks and values

The IBM benchmarks results shown herein were derived using particular, well configured, development-level and generally-available computer systems. Buyers should consult other sources of information to evaluate the performance of systems they are considering buying and should consider conducting application oriented testing. For additional information about the benchmarks, values and systems tested, contact your local IBM office or IBM authorized reseller or access the Web site of the benchmark consortium or benchmark vendor.

IBM benchmark results can be found in the IBM Power Systems Performance Report at http://www.ibm.com/systems/p/hardware/system_perf.html .

All performance measurements were made with AIX or AIX 5L operating systems unless otherwise indicated to have used Linux. For new and upgraded systems, the latest versions of AIX were used. All other systems used previous versions of AIX. The SPEC CPU2006, LINPACK, and Technical Computing benchmarks were compiled using IBM's high performance C, C++, and FORTRAN compilers for AIX 5L and Linux. For new and upgraded systems, the latest versions of these compilers were used: XL C for AIX v11.1, XL C/C++ for AIX v11.1, XL FORTRAN for AIX v13.1, XL C/C++ for Linux v11.1, and XL FORTRAN for Linux v13.1. Linpack HPC (Highly Parallel Computing) used the current versions of the IBM Engineering and Scientific Subroutine Library (ESSL). For Power7 systems, IBM Engineering and Scientific Subroutine Library (ESSL) for AIX Version 5.1 and IBM Engineering and Scientific Subroutine Library (ESSL) for Linux Version 5.1 were used.

For a definition/explanation of each benchmark and the full list of detailed results, visit the Web site of the benchmark consortium or benchmark vendor.

SPEC	http://www.spec.org
LINPACK	http://www.netlib.org/benchmark/performance.pdf
Pro/E	http://www.proe.com
GPC	http://www.spec.org/gpc
STREAM	http://www.cs.virginia.edu/stream/
Fluent	http://www.fluent.com/software/fluent/index.htm
TOP500 Supercomputers	http://www.top500.org/
AMBER	http://amber.scripps.edu/
FLUENT	http://www.fluent.com/software/fluent/fl5bench/index.htm
GAMESS	http://www.msg.chem.iastate.edu/gamess
GAUSSIAN	http://www.gaussian.com
ANSYS	http://www.ansys.com/services/hardware-support-db.htm
	Click on the "Benchmarks" icon on the left hand side frame to expand. Click on "Benchmark Results in a Table" icon for benchmark results.
ABAQUS	http://www.simulia.com/support/v68/v68_performance.php
ECLIPSE	http://www.sis.slb.com/content/software/simulation/index.asp?seg=geoquest&
MM5	http://www.mmm.ucar.edu/mm5/
MSC.NASTRAN	http://www.mscsoftware.com/support/prod%5Fsupport/nastran/performance/v04_sngl.cfm
STAR-CD	www.cd-adapco.com/products/STAR-CD/performance/320/index/html
NAMD	http://www.ks.uiuc.edu/Research/namd
HMMER	**http://hmmer.janelia.org/**
	http://powerdev.osuosl.org/project/hmmerAltivecGen2mod

Revised December 2, 2010

Notes on performance estimates

rPerf for AIX

rPerf (Relative Performance) is an estimate of commercial processing performance relative to other IBM UNIX systems. It is derived from an IBM analytical model which uses characteristics from IBM internal workloads, TPC and SPEC benchmarks. The rPerf model is not intended to represent any specific public benchmark results and should not be reasonably used in that way. The model simulates some of the system operations such as CPU, cache and memory. However, the model does not simulate disk or network I/O operations.

- **rPerf estimates are calculated based on systems with the latest levels of AIX and other pertinent software at the time of system announcement. Actual performance will vary based on application and configuration specifics. The IBM eServer pSeries 640 is the baseline reference system and has a value of 1.0. Although rPerf may be used to approximate relative IBM UNIX commercial processing performance, actual system performance may vary and is dependent upon many factors including system hardware configuration and software design and configuration. Note that the rPerf methodology used for the POWER6 systems is identical to that used for the POWER5 systems. Variations in incremental system performance may be observed in commercial workloads due to changes in the underlying system architecture.**

All performance estimates are provided "AS IS" and no warranties or guarantees are expressed or implied by IBM. Buyers should consult other sources of information, including system benchmarks, and application sizing guides to evaluate the performance of a system they are considering buying. For additional information about rPerf, contact your local IBM office or IBM authorized reseller.

==========

CPW for IBM i

Commercial Processing Workload (CPW) is a relative measure of performance of processors running the IBM i operating system. Performance in customer environments may vary. The value is based on maximum configurations. More performance information is available in the Performance Capabilities Reference at: www.ibm.com/systems/i/solutions/perfmgmt/resource.html

Revised April 2, 2007