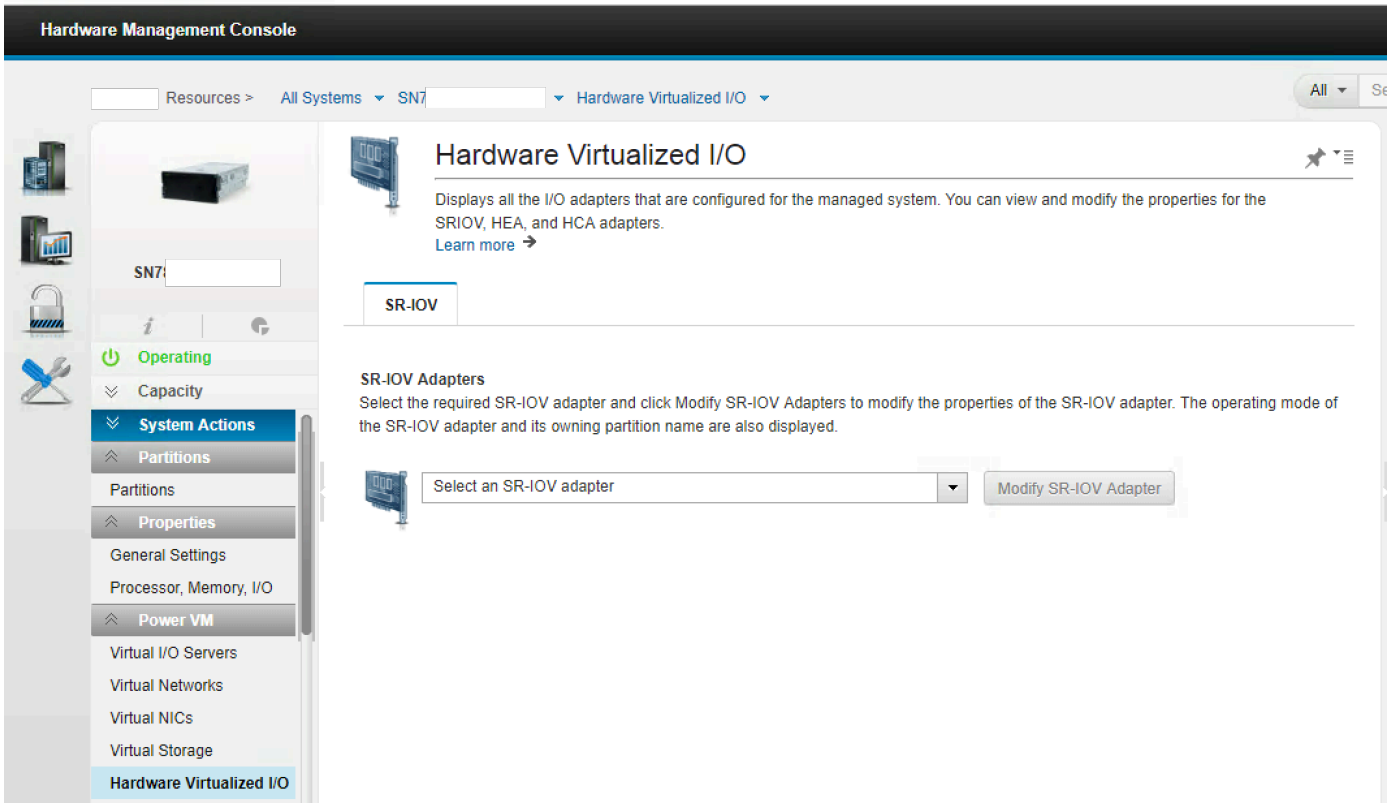
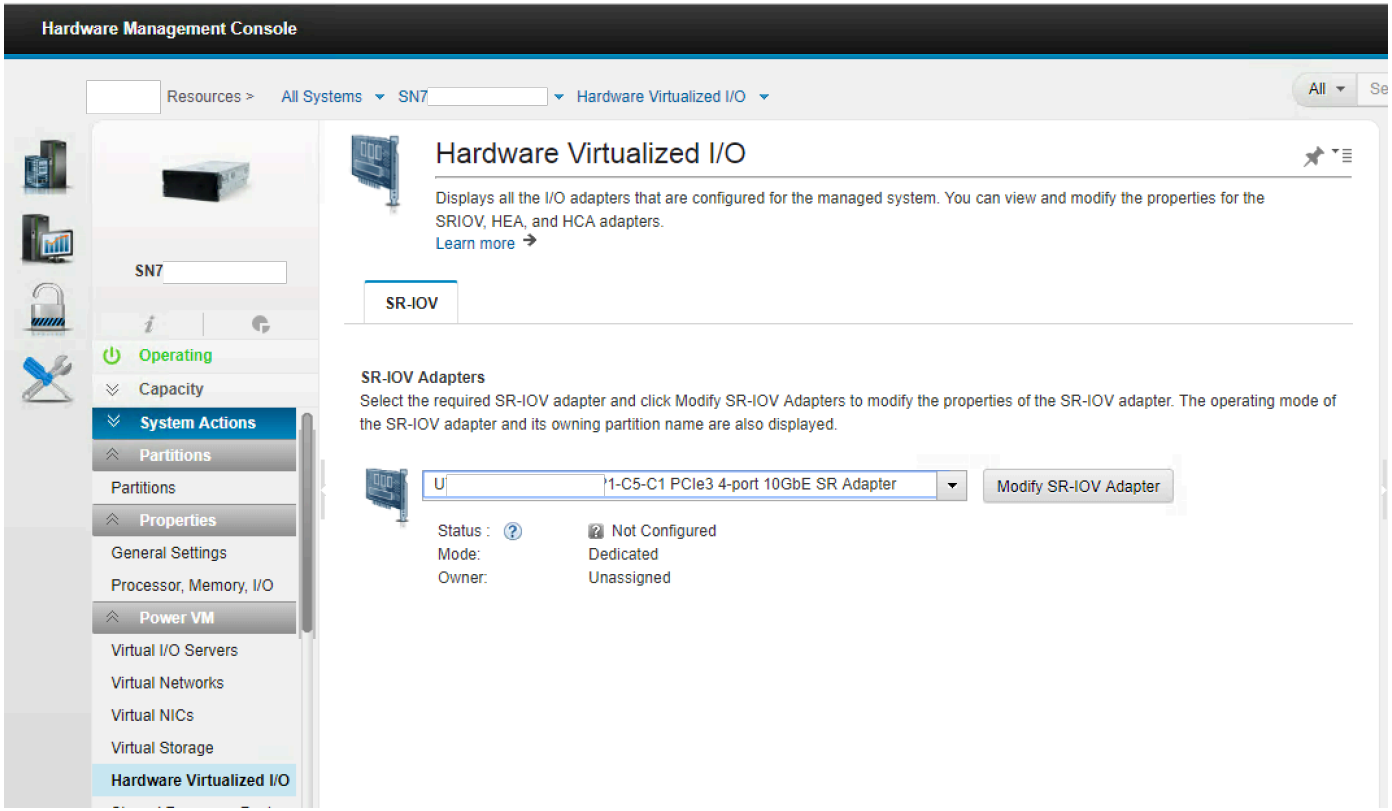
Configuring Native SR-IOV adapters on AIX with Link Aggregation

In this example we will configure Native SR-IOV adapters on AIX with Link Aggregation, on a POWER9 E980. This procedure assumes that the necessary switch configuration (for LACP/port channeling) has already been completed by a network specialist on each of the network switch ports.

1. **We start with two SR-IOV capable adapters (type EN16). We select the Power Server, where the adapters are located, and click on "Hardware Virtualized I/O". We are presented with the following screen:**



2. **From the drop down list of adapters, we select one of these for SR-IOV.**



3. **Click “Modify SR-IOV Adapter”. Select “Shared”. Click OK. It can take a while for the adapter to enter “shared” mode. So you may see a “spinning wheel” icon for several minutes. During this time, the adapter firmware is automatically being flashed by the system.**

SR-IOV

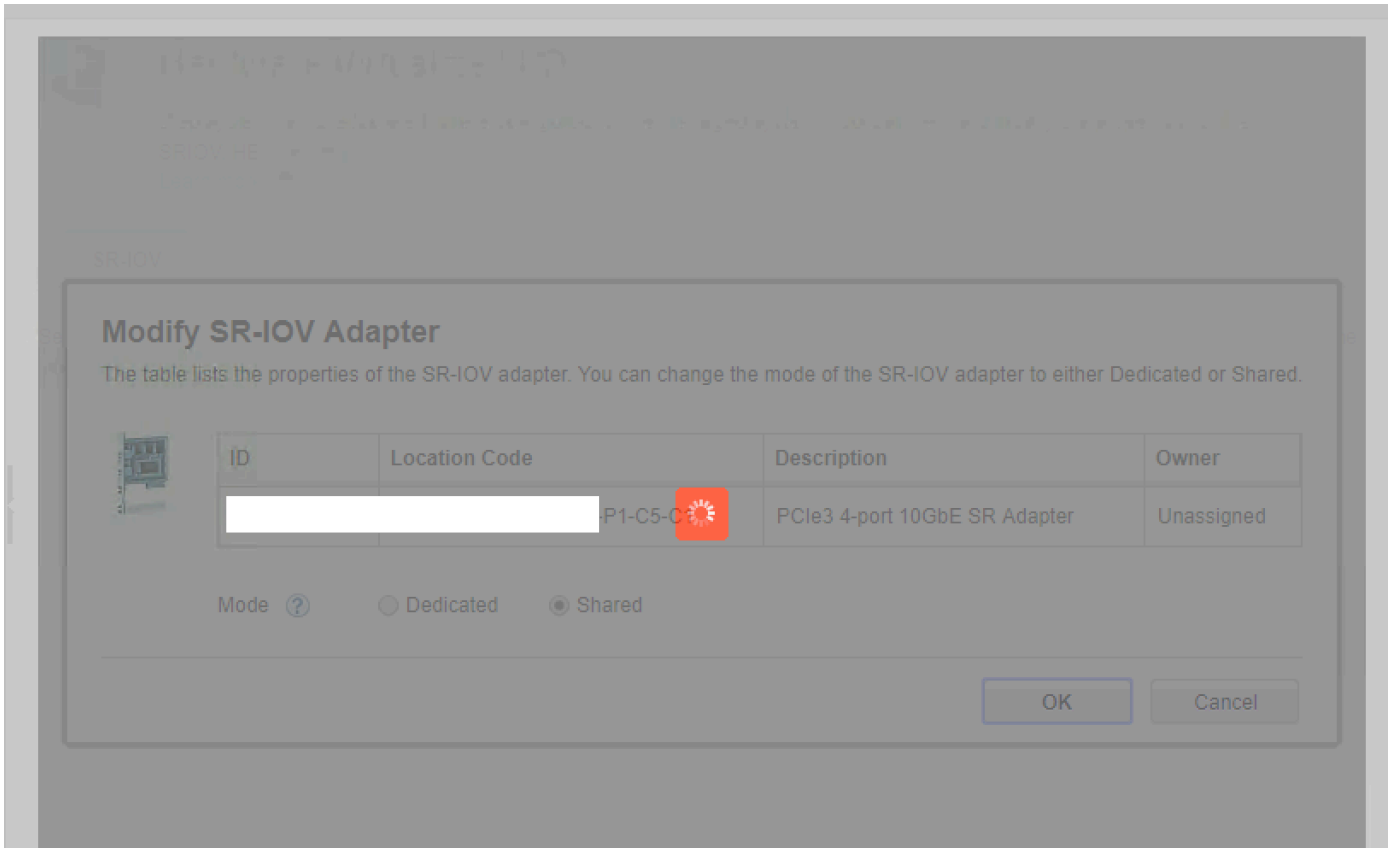
Modify SR-IOV Adapter

The table lists the properties of the SR-IOV adapter. You can change the mode of the SR-IOV adapter to either Dedicated or Shared.

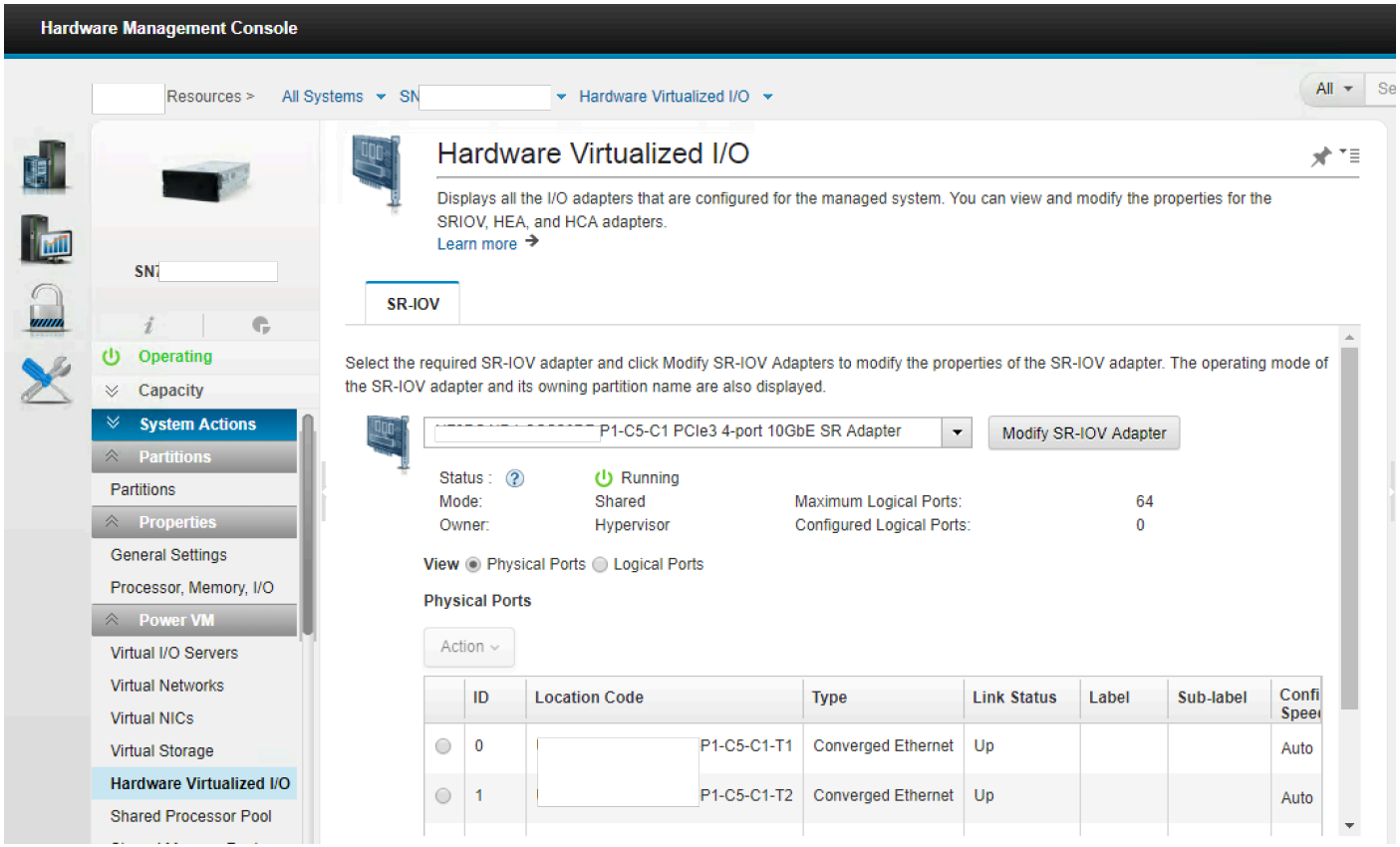
ID	Location Code	Description	Owner
	U P1-C5-C1	PCIe3 4-port 10GbE SR Adapter	Unassigned

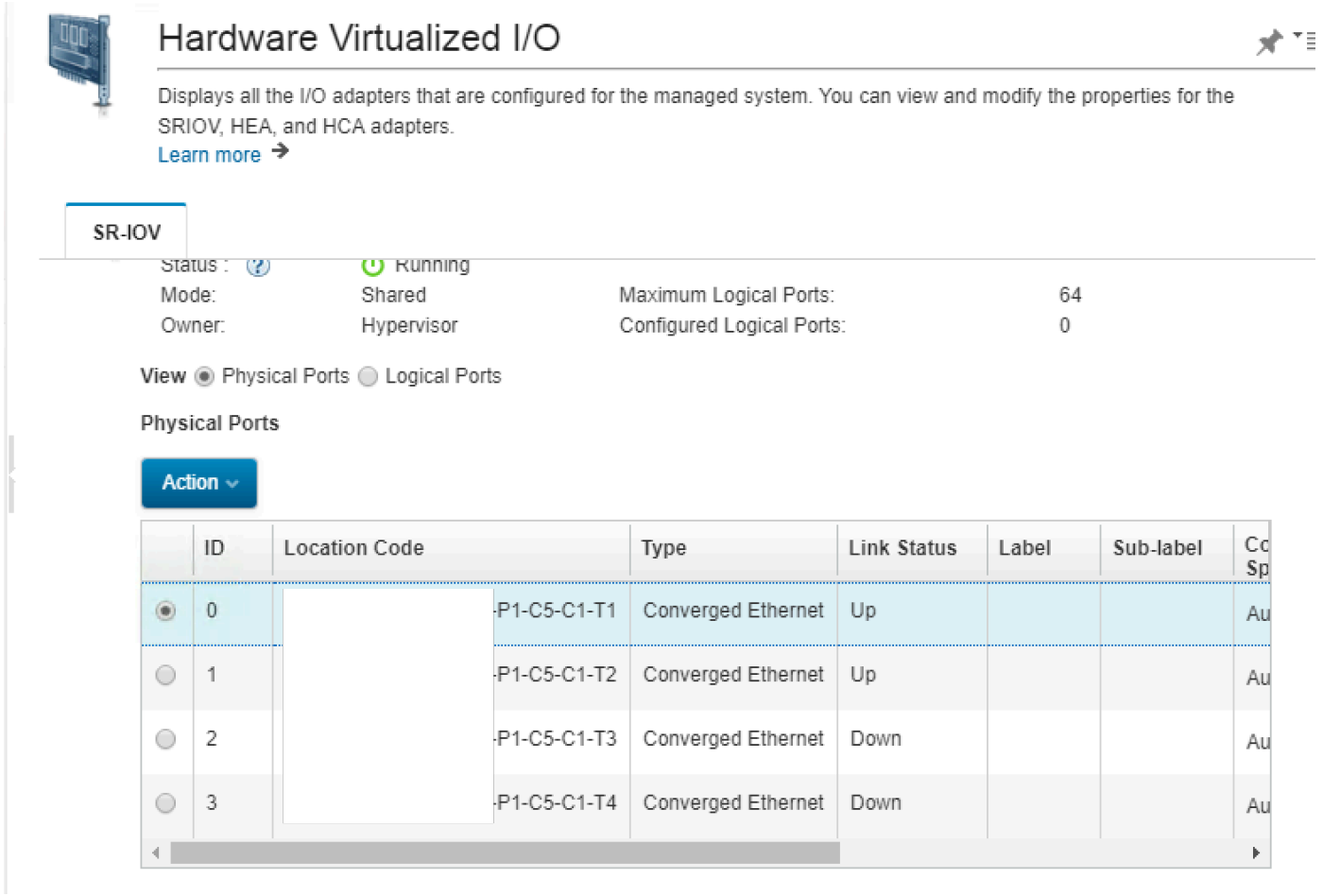
Mode (?) ○ Dedicated ◉ Shared

OK Cancel



4. **The adapter is now owned by the Hypervisor. A list of the physical ports and their status is displayed.**


Hardware Management Console
Resources >
All Systems
SN
Hardware Virtualized I/O
All
Operating
Capacity
System Actions
Partitions
Partitions
Properties
General Settings
Processor, Memory, I/O
Power VM
Virtual I/O Servers
Virtual Networks
Virtual NICs
Virtual Storage
Hardware Virtualized I/O
Shared Processor Pool
Hardware Virtualized I/O
Displays all the I/O adapters that are configured for the managed system. You can view and modify the properties for the SRIOV, HEA, and HCA adapters.
Learn more
SR-IOV
Select the required SR-IOV adapter and click Modify SR-IOV Adapters to modify the properties of the SR-IOV adapter. The operating mode of the SR-IOV adapter and its owning partition name are also displayed.
P1-C5-C1 PCIe3 4-port 10GbE SR Adapter
Modify SR-IOV Adapter
Status :
Running
Mode:
Shared
Owner:
Hypervisor
Maximum Logical Ports:
64
Configured Logical Ports:
0
View
Physical Ports
Logical Ports
Physical Ports
Action
ID
Location Code
Type
Link Status
Label
Sub-label
Confi Spee
0
P1-C5-C1-T1
Converged Ethernet
Up
Auto
1
P1-C5-C1-T2
Converged Ethernet
Up
Auto



5. **Select one of the physical ports. Click on “Action” and select “Modify Physical Port”. Expand the “Advanced Section”. It is generally recommended to enabled flow control (ON) on 10GbE adapters. Please check with your network team to ensure that flow control is enabled on the switch port (and throughout the 10G network). The MTU size is 9000, by default on POWER9.**

Hardware Virtualized I/O

Displays all the I/O adapters that are configured for the managed system. You can view and modify the properties for the SRIOV, HEA, and HCA adapters.

Learn more

SR-IOV

Status : Running

Mode: Shared

Owner: Hypervisor

Maximum Logical Ports: 64

Configured Logical Ports: 0

View Physical Ports Logical Ports

Physical Ports

Action

View Physical Port

Modify Physical Port

		le	Type	Link Status	Label	Sub-label	Co Sp
		-P1-C5-C1-T1	Converged Ethernet	Up			Au
	1	P1-C5-C1-T2	Converged Ethernet	Up			Au
	2	P1-C5-C1-T3	Converged Ethernet	Down			Au
	3	P1-C5-C1-T4	Converged Ethernet	Down			Au

Modify SR-IOV Physical Port

You can modify the properties of a physical port that is on a Single Root I/O Virtualization (SR-IOV) capable adapter. The table lists the deta
of the physical port that you want to modify. You can also modify additional settings for this physical port.

ID	Location Code	Type	Link Status	Available Logical Ports	Configured Logical Ports
0	P1-C5-C1-T1	Converged Ethernet	Up	16 Ethernet	0 Ethernet

Label

Sub-label

Configured Capacity (%) 0.0

Available Capacity (%) 100.0

Configured Speed Auto

Negotiated Speed 10 Gbps

▸ **Advanced Section**

Learn more. →

OK Cancel

Modify SR-IOV Physical Port

You can modify the properties of a physical port that is on a Single Root I/O Virtualization (SR-IOV) capable adapter. The table lists the c
of the physical port that you want to modify. You can also modify additional settings for this physical port.

ID	Location Code	Type	Link Status	Available Logical Ports	Configured Logical Ports
0	-P1-C5-C1-T1	Converged Ethernet	Up	16 Ethernet	0 Ethernet

Label

Sub-label

Configured Capacity (%) 0.0

Available Capacity (%) 100.0

Configured Speed Auto

Negotiated Speed 10 Gbps

Advanced Section

MTU Size: 9000

Port Switch Mode: VEB

Flow Control: Off

Priority Flow Control: **Off**

Ethernet Logical Port Limits

Configured: 0

Maximum: 16

Total Supported: 16

Modify SR-IOV Physical Port

You can modify the properties of a physical port that is on a Single Root I/O Virtualization (SR-IOV) capable adapter. The table lists the c
of the physical port that you want to modify. You can also modify additional settings for this physical port.

ID	Location Code	Type	Link Status	Available Logical Ports	Configured Logical Ports
0	P1-C5-C1-T1	Converged Ethernet	Up	16 Ethernet	0 Ethernet

Label

Sub-label

Configured Capacity (%) 0.0

Available Capacity (%) 100.0

Configured Speed Auto

Negotiated Speed 10 Gbps

▼ **Advanced Section**

MTU Size: 9000

Port Switch Mode: VEB

Flow Control: Off

- On
- Off

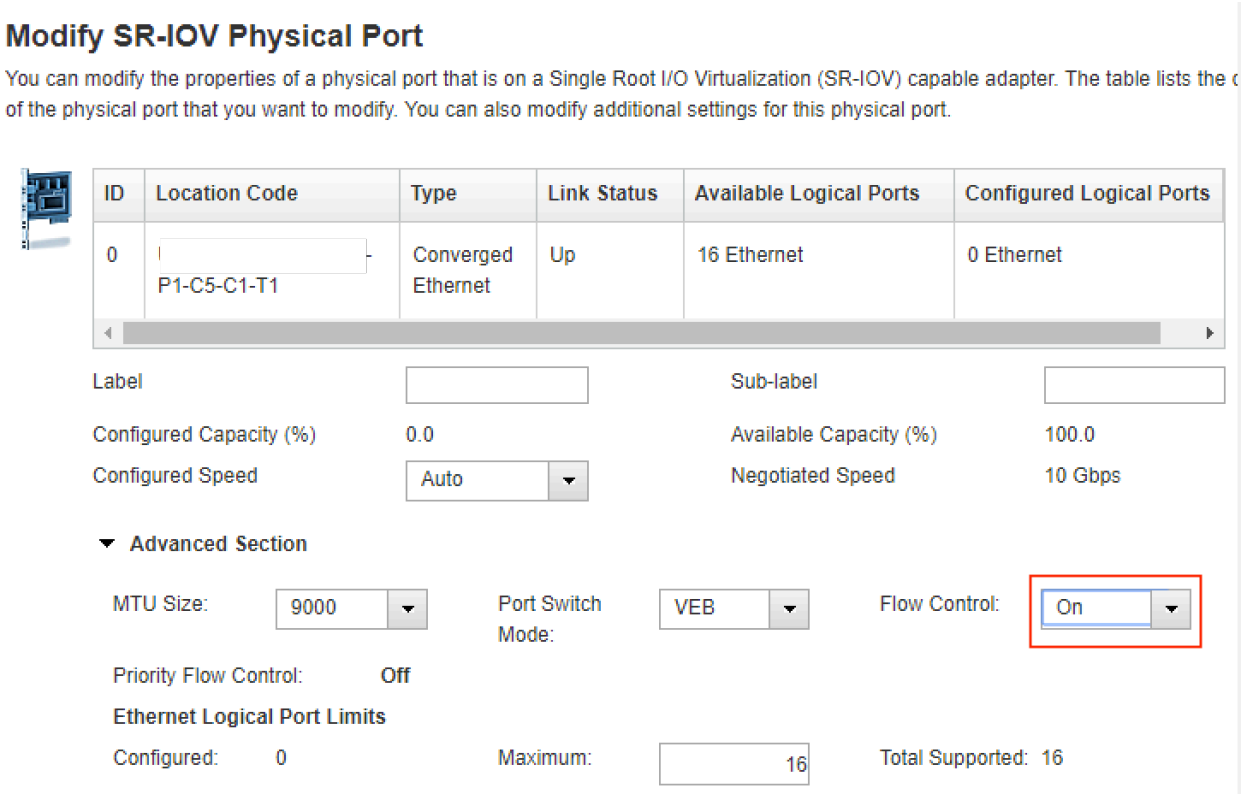
Priority Flow Control: **Off**

Ethernet Logical Port Limits

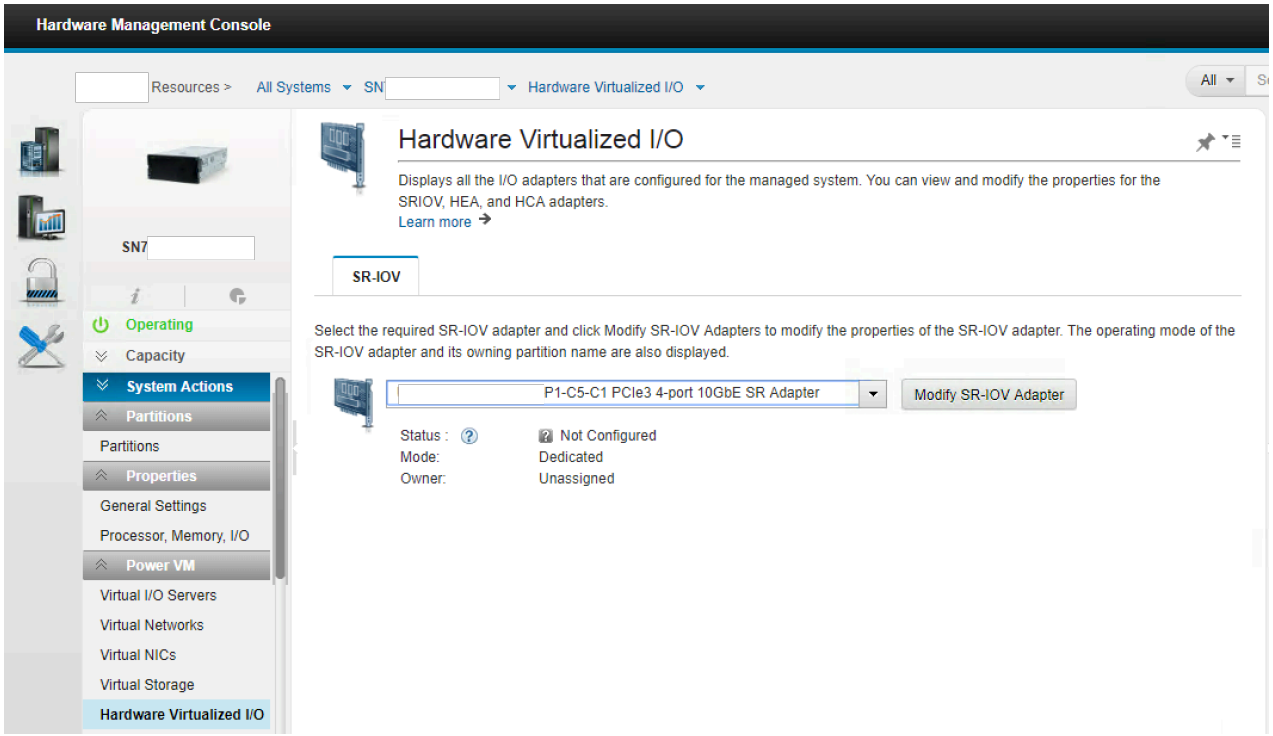
Configured: 0

Maximum: 16

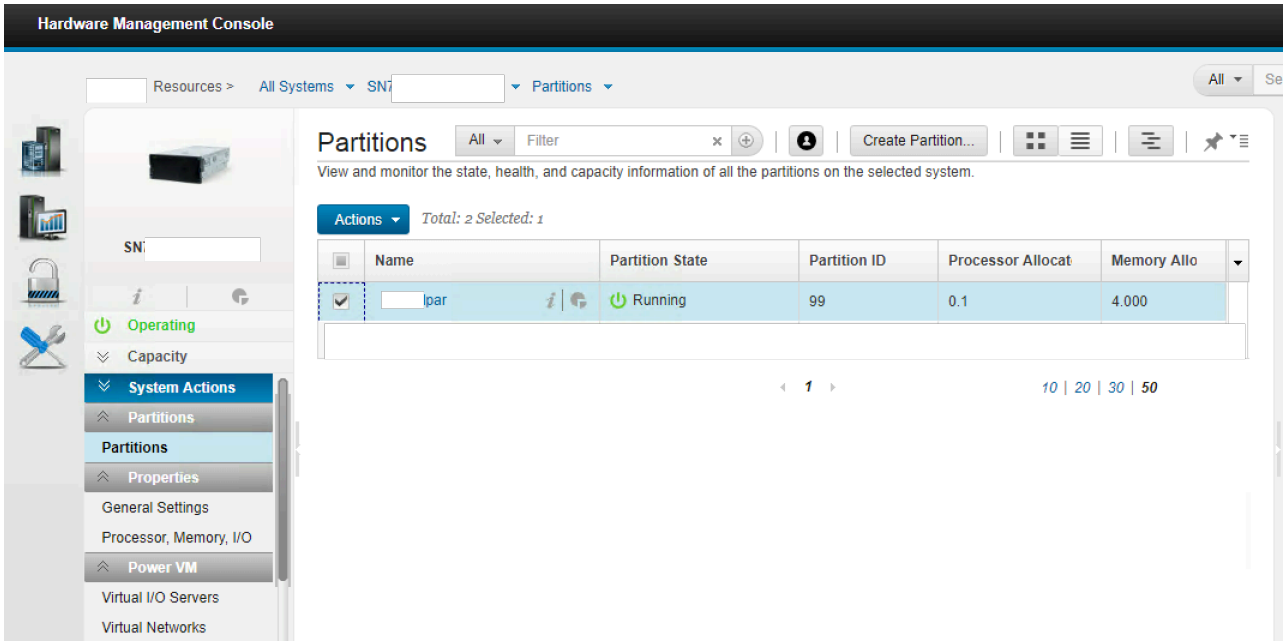
Total Supported: 16

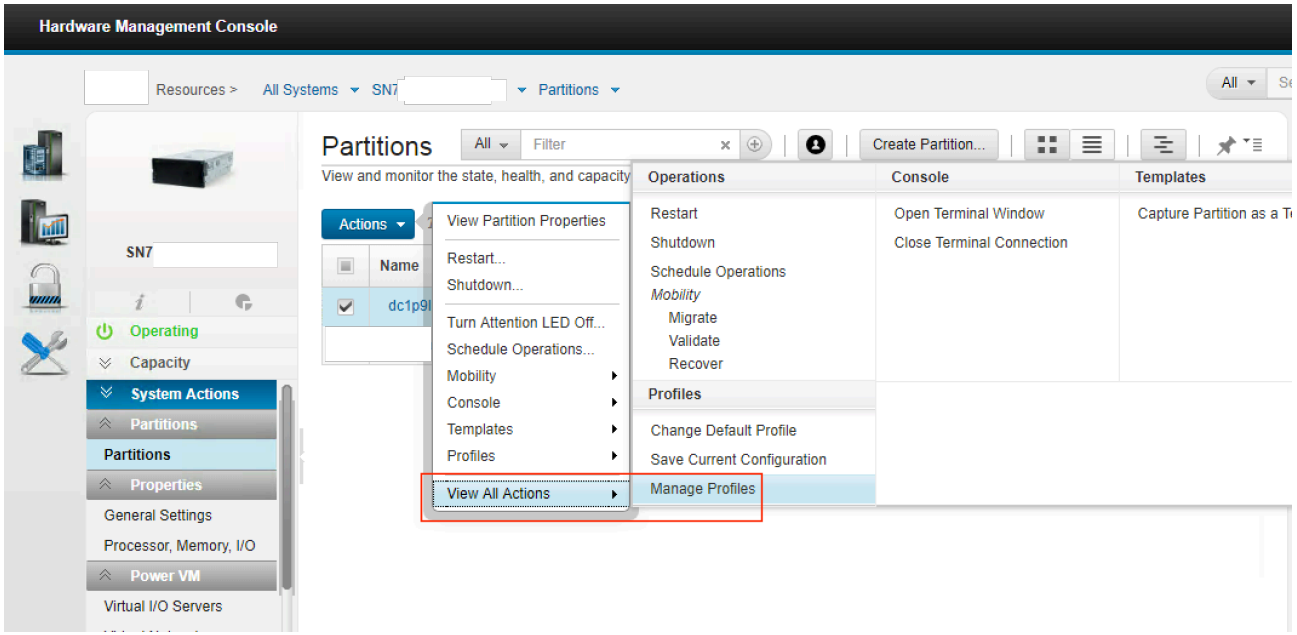


6. **Return to the drop down list adapters and select another adapter. Run through the same config step's as above.**

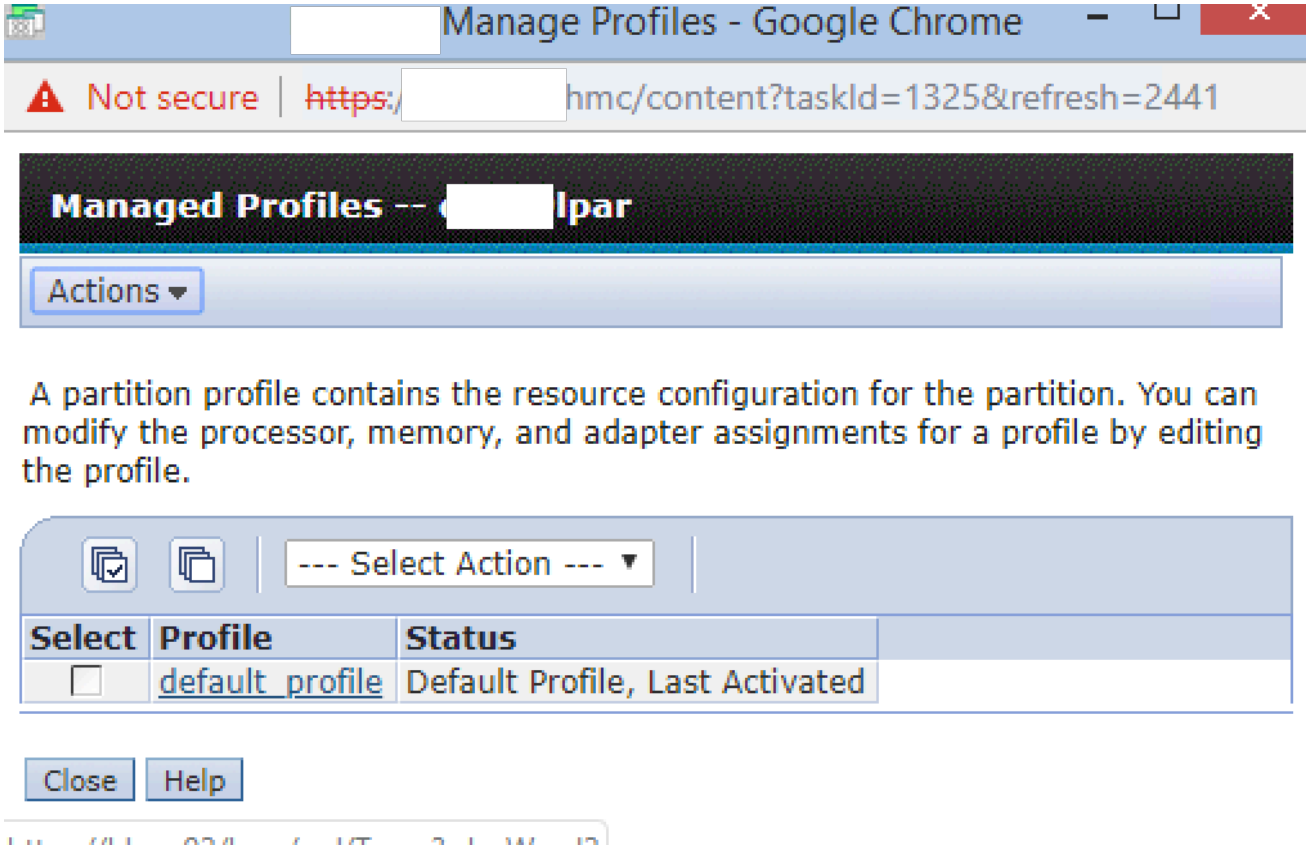


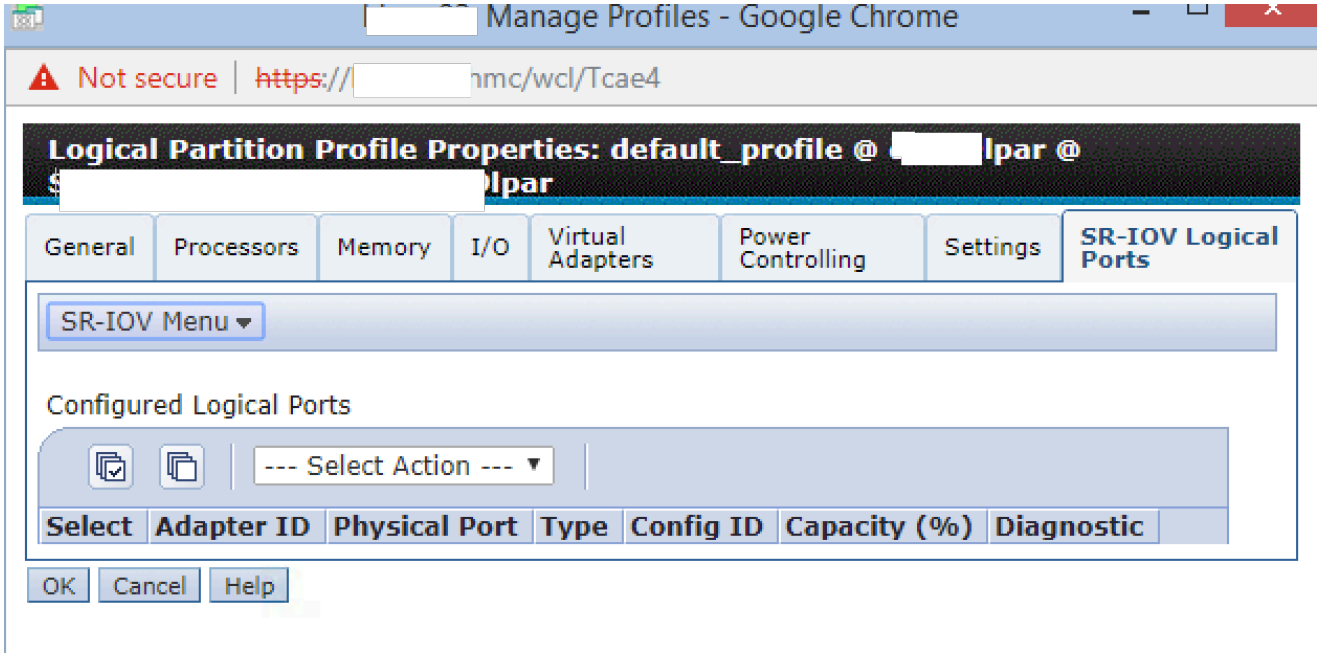
7. **Now we need to assign some logical ports from each of the physical ports to one of the LPARs on the system. Select the LPAR, “Actions”, “View All Actions” and “Managed Profiles”.**


Hardware Management Console
Resources >
All Systems
Partitions
All
Partitions
All
Filter
Create Partition...
View and monitor the state, health, and capacity information of all the partitions on the selected system.
Actions
Total: 2 Selected: 1
Name
Partition State
Partition ID
Processor Allocat
Memory Allo
lpar
Running
99
0.1
4.000
1
10 | 20 | 30 | 50
Operating
Capacity
System Actions
Partitions
Partitions
Properties
General Settings
Processor, Memory, I/O
Power VM
Virtual I/O Servers
Virtual Networks

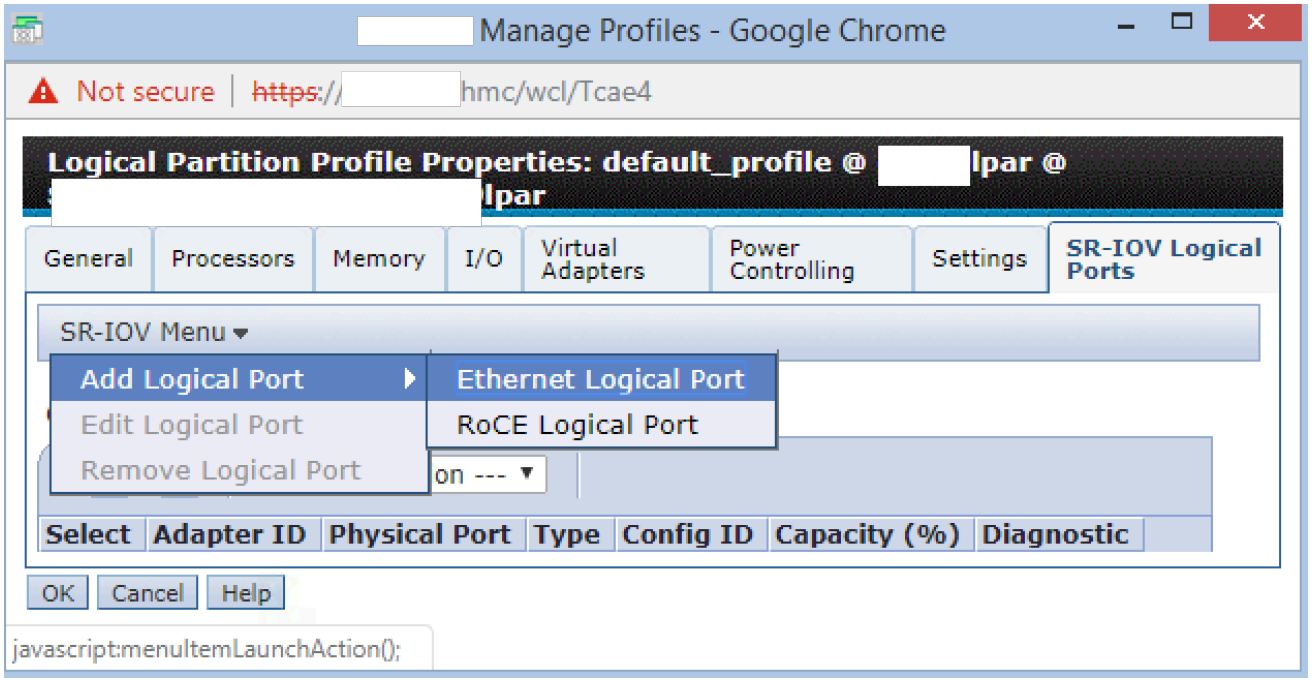


8. **Click on “default_profile”. Click on “SR-IOV Logical Ports”.**

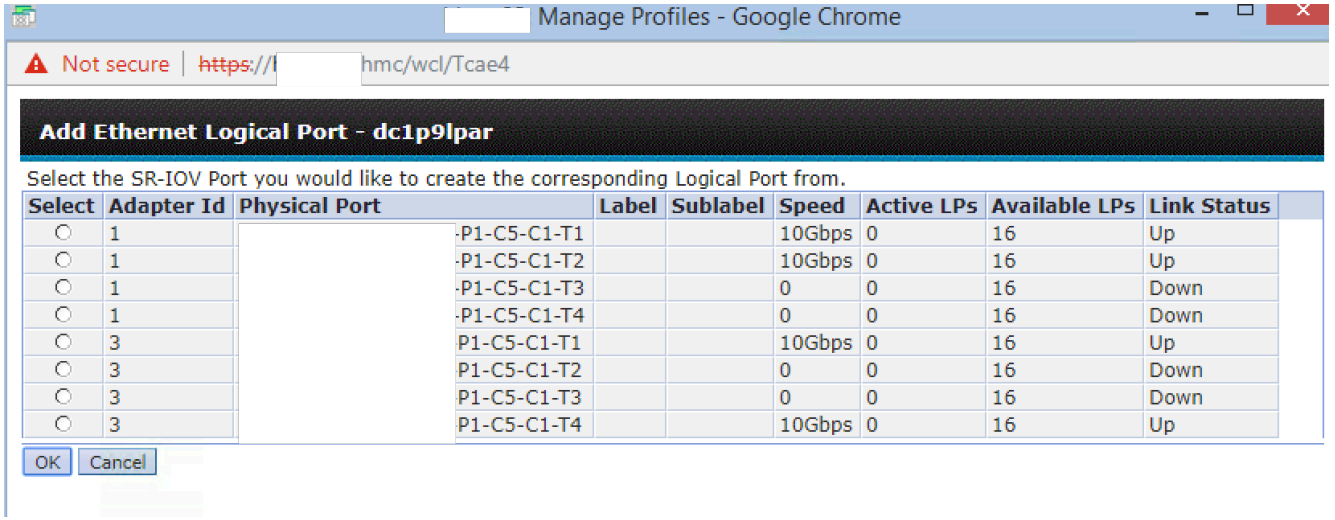

Manage Profiles - Google Chrome
Not secure
https:/
hmc/content?taskId=1325&refresh=2441
Managed Profiles --
lpar
Actions
A partition profile contains the resource configuration for the partition. You can modify the processor, memory, and adapter assignments for a profile by editing the profile.
--- Select Action ---
Select
Profile
Status
default_profile
Default Profile, Last Activated
Close
Help

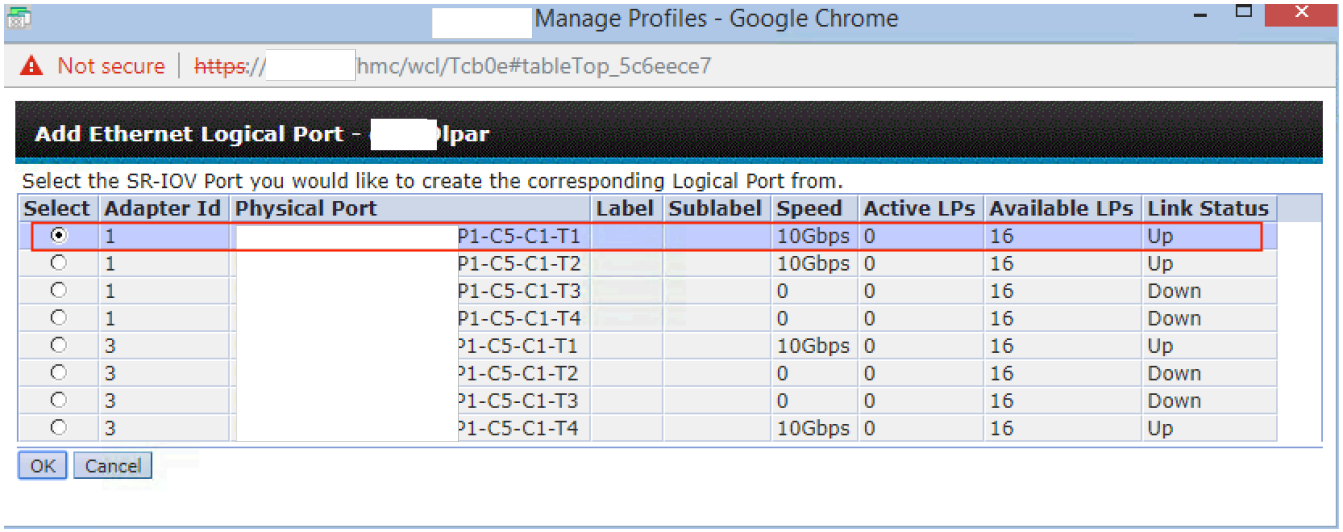


9. **Click on “Add Logical Port” and “Ethernet Logical Port”.**

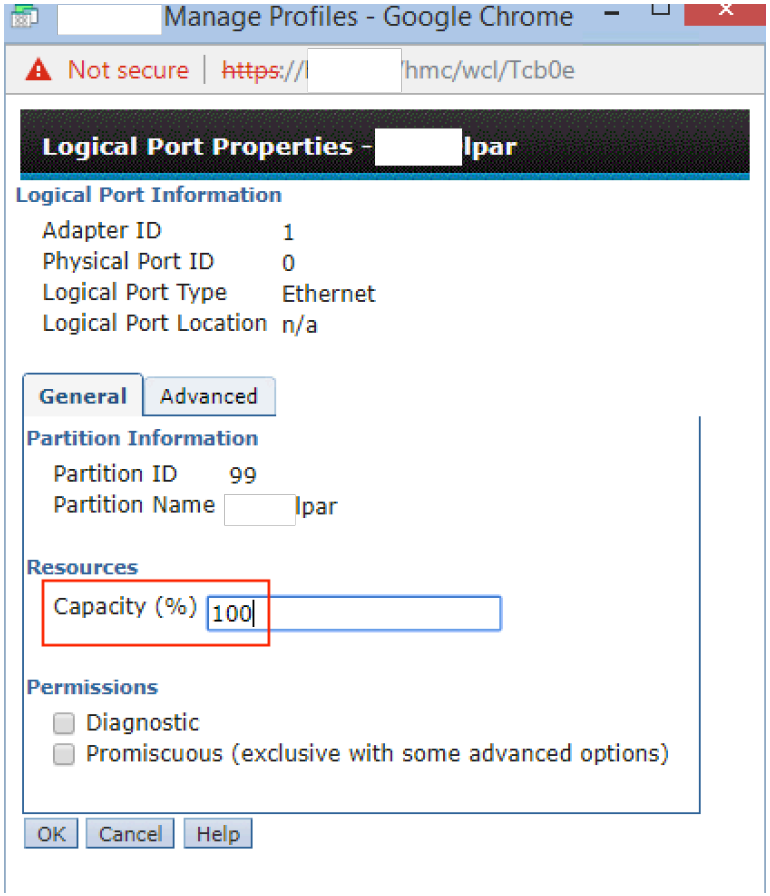


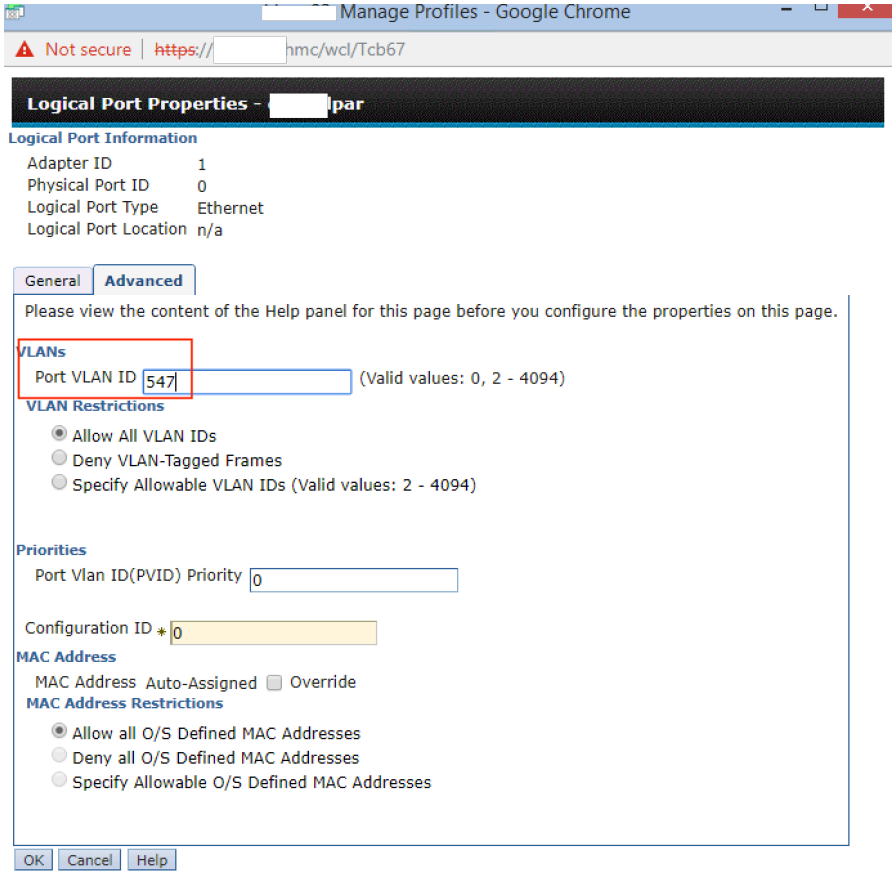
10. From the list of physical ports, select the first port from the first adapter i.e. UXXXX.XX1.XXXXXXX -P1-C5-C1-T1. Click OK.


Manage Profiles - Google Chrome
Not secure
https://
hmc/wcl/Tcae4
Add Ethernet Logical Port - dc1p9lpar
Select the SR-IOV Port you would like to create the corresponding Logical Port from.
Select
Adapter Id
Physical Port
Label
Sublabel
Speed
Active LPs
Available LPs
Link Status
1 -P1-C5-C1-T1 10Gbps 0 16 Up
1 -P1-C5-C1-T2 10Gbps 0 16 Up
1 -P1-C5-C1-T3 0 0 16 Down
1 -P1-C5-C1-T4 0 0 16 Down
3 P1-C5-C1-T1 10Gbps 0 16 Up
3 P1-C5-C1-T2 0 0 16 Down
3 P1-C5-C1-T3 0 0 16 Down
3 P1-C5-C1-T4 10Gbps 0 16 Up
OK
Cancel

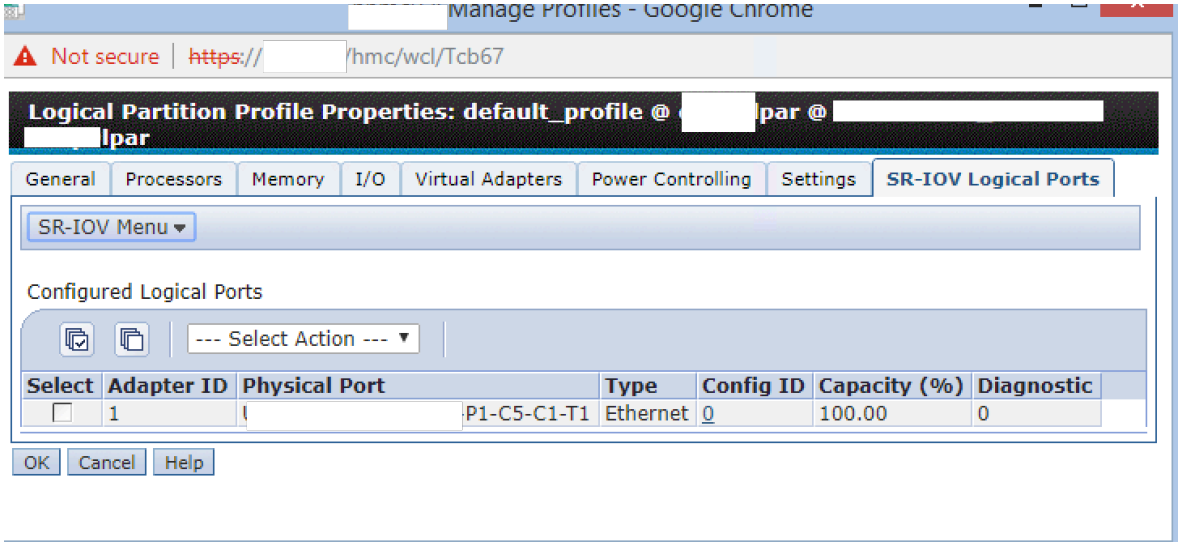

Manage Profiles - Google Chrome
Not secure
https://
hmc/wcl/Tcb0e#tableTop_5c6eece7
Add Ethernet Logical Port -
lpar
Select the SR-IOV Port you would like to create the corresponding Logical Port from.
Select
Adapter Id
Physical Port
Label
Sublabel
Speed
Active LPs
Available LPs
Link Status
1 P1-C5-C1-T1 10Gbps 0 16 Up
1 P1-C5-C1-T2 10Gbps 0 16 Up
1 P1-C5-C1-T3 0 0 16 Down
1 P1-C5-C1-T4 0 0 16 Down
3 P1-C5-C1-T1 10Gbps 0 16 Up
3 P1-C5-C1-T2 0 0 16 Down
3 P1-C5-C1-T3 0 0 16 Down
3 P1-C5-C1-T4 10Gbps 0 16 Up
OK
Cancel

11. Change “Capacity (%)” to 100. Click on advanced and enter the required VLAN ID in “Port VLAN ID”. Click OK. For link aggregation devices, the adapter capacity must be set to 100%. Please refer the links in the references section for details on why this is a recommended and mandatory requirement.

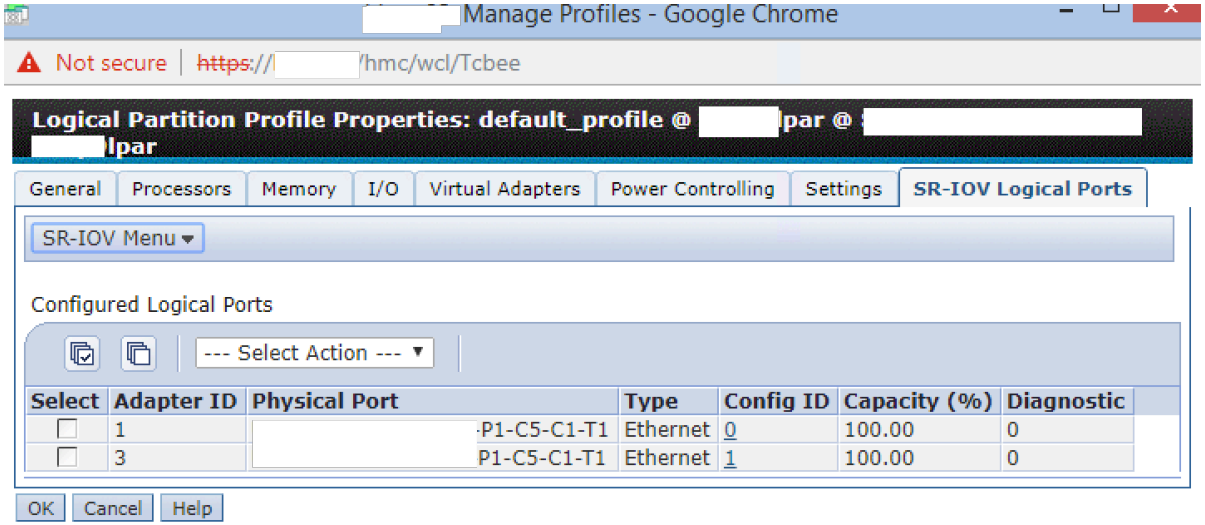




12. A new logical port will appear in the LPAR profile.



13. Perform the same task for the other adapter i.e. UXXXX.XX1.XXXXXXX -P1-C5-C1-T1. Change “Capacity (%)” to 100. Click on advanced and enter the same VLAN ID (547) in “Port VLAN ID”. You will end up with two logical ports in the LPAR profile.



14. Shutdown the LPAR and reactivate for the updated profile to take effect. After the LPAR has finished booting, login into the system and check that two new VF adapters are available in AIX.

```
# lsdev -Cc adapter
ent0   Available 05-00 PCIe3 4-Port 10GbE SR Adapter VF(df1028e21410e304)
ent1   Available 06-00 PCIe3 4-Port 10GbE SR Adapter VF(df1028e21410e304)
pkcs11 Available       ACF/PKCS#11 Device
vsa0   Available       LPAR Virtual Serial Adapter
vscsi0 Available       Virtual SCSI Client Adapter
vscsi1 Available       Virtual SCSI Client Adapter

# lscfg -vpl ent0 | grep Phys
    Physical Location: UXXXX.XX1.XXXXXXX-P1-C5-C1-T1-S1

# lscfg -vpl ent1 | grep Phys
    Physical Location: UXXXX.XX2.XXXXXXX-P1-C5-C1-T1-S1
```

15. Check that both interfaces are up and on the correct VLAN.

```
# entstat -d ent0 | grep -i link
Physical Port Link Status: Up
Logical Port Link Status: Up
        PCIe Link Speed: Unknown
# entstat -d ent1 | grep -i link
Physical Port Link Status: Up
Logical Port Link Status: Up
        PCIe Link Speed: Unknown

# entstat -d ent0 | grep -i vlan
Port VLAN (Priority:ID): 0:0547
VLAN ACL Status: DisabledEnabled VLAN IDs: None
# entstat -d ent1 | grep -i vlan
Port VLAN (Priority:ID): 0:0547
VLAN ACL Status: DisabledEnabled VLAN IDs: None
```

16. Configure the Link Aggregation device on AIX now.

```
# smit etherchannel
Add 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                                                                   |
```

```
                                  Add An EtherChannel / Link Aggregation

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

                                                        [Entry Fields]
  EtherChannel / Link Aggregation Adapters             ent0,ent1+
  Enable Alternate Address                             no+
  Alternate Address                                   []+
  Enable Gigabit Ethernet Jumbo Frames                 yes+
```

```
  Mode                                                  8023ad+
  IEEE 802.3ad Interval                                 long+
  Hash Mode                                             src_dst_port+


                                                              COMMAND STATUS

Command: OK            stdout: yes           stderr: no

Before command completion, additional instructions may appear below.

ent2 Available
```

17. Configure an IP address on the Link Agg interface (en2).

```
# smit tcpip

 Minimum Configuration & Startup
                              +-------------------------------------------------------------------------+
                              |                        Available Network Interfaces                     |
                              |                                                                         |
                              | Move cursor to desired item and press Enter.                            |
                              |                                                                         |
                              |   en0   0M-00   Standard Ethernet Network Interface                     |
                              |   en1   0P-00   Standard Ethernet Network Interface                     |
                              |   en2           Standard Ethernet Network Interface                     |

                                                  Minimum Configuration & Startup

                             To Delete existing configuration data, please use Further Configuration menus

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

                                                     [Entry Fields]
* HOSTNAME                                          [p9lpar]
* Internet ADDRESS (dotted decimal)                 [10.1.1.39]
  Network MASK (dotted decimal)                     [255.255.255.0]
* Network INTERFACE                                  en2
  NAMESERVER
           Internet ADDRESS (dotted decimal)        []
```

```
          DOMAIN Name                                [] 
  Default Gateway
       Address (dotted decimal or symbolic name)     [10.1.1.1]


                                                                COMMAND STATUS

Command: OK            stdout: yes           stderr: no

Before command completion, additional instructions may appear below.

en2
p9lpar
inet0 changed
en2 changed
inet0 changed
```

18. Review the IP configuration.

```
# ifconfig -a
en2:
flags=1e080863,18c0<UP,BROADCAST,NOTRAILERS,RUNNING,SIMPLEX,MULTICAST,GROUPRT,64BIT,CHECKSUM_OFFLOAD(ACTIVE),LARGESEND,CHAIN>
        inet 10.1.1.39 netmask 0xffffff00 broadcast 10.1.1.255
         tcp_sendspace 262144 tcp_recvspace 262144 rfc1323 1
lo0: flags=e08084b,c0<UP,BROADCAST,LOOPBACK,RUNNING,SIMPLEX,MULTICAST,GROUPRT,64BIT,LARGESEND,CHAIN>
        inet 127.0.0.1 netmask 0xff000000 broadcast 127.255.255.255
        inet6 ::1%1/64
         tcp_sendspace 131072 tcp_recvspace 131072 rfc1323 1
```

19. Change the MTU size to 9000.

```
# chdev -l en2 -a mtu=9000
en2 changed
```

20. Review the routing table and ensure the correct default gateway is configured. Try pinging the gateway address.

```
# netstat -nr
Routing tables
Destination        Gateway           Flags    Refs     Use  If    Exp  Groups
```

```
Route Tree for Protocol Family 2 (Internet):
default            10.1.1.1          UG        1        14 en2      -      -
10.1.1.0          10.1.1.39        UHSb      0         0 en2      -      -   =>
10.1.1/24         10.1.1.39        U         0         0 en2      -      -
10.1.1.39         127.0.0.1         UGHS      0         0 lo0      -      -
10.1.1.255        10.1.1.39        UHSb      0         0 en2      -      -
127/8              127.0.0.1         U         3    103012 lo0      -      -

Route Tree for Protocol Family 24 (Internet v6):
::1%1              ::1%1             UH        0        54 lo0      -      -

# ping 10.1.1.1
PING 10.1.1.1: (10.1.1.1): 56 data bytes
64 bytes from 10.1.1.1: icmp_seq=0 ttl=255 time=0 ms
64 bytes from 10.1.1.1: icmp_seq=1 ttl=255 time=0 ms
64 bytes from 10.1.1.1: icmp_seq=2 ttl=255 time=0 ms
64 bytes from 10.1.1.1: icmp_seq=3 ttl=255 time=0 ms
```

21. If the ping test fails, check that the Link Agg devices are “in sync” across both partners. They should all be “IN_SYNC”.

```
# entstat -d ent2 | grep -i sync
                Synchronization: IN_SYNC
                Synchronization: IN_SYNC
                Synchronization: IN_SYNC
                Synchronization: IN_SYNC
```

The configuration is now complete.

NOTE: If you are attempting to configure native SR-IOV on AIX, without Link Aggregation and you wish to use jumbo frames (MTU=9000), you need to change the jumbo_frames attribute on the entX device before you attempt to change the MTU to 9000. Failure to do so will result in the following error (and failure of the interface as a result).

```
# chdev -l en0 -a mtu=9000
Method error (/usr/lib/methods/chgif):
        0514-068 Cause not known.
0821-532 ioctl (SIOCSIFMTU): A return value of a math subroutine is not within machine precision.
0821-103 : The command /usr/sbin/ifconfig en0 inet  10.1.1.7 arp netmask 255.255.255.0 mtu 9000 up largesend -monitor -thread
failed.
```

```
0821-007 cfgif: ifconfig command failed.
        The status of"en0" Interface in the current running system is uncertain.
0821-103 : The command /usr/lib/methods/cfgif -len0 failed.
chgif: cfgif failed to configure ISNO attributes
0821-532 ioctl (SIOCSIFMTU): A return value of a math subroutine is not within machine precision.
0821-103 : The command /usr/sbin/ifconfig en0 inet 10.1.1.7 arp netmask 255.255.255.0 mtu 9000 up largesend -monitor -thread
failed.
0821-229 chgif: ifconfig command failed.
The status of"en0" Interface in the current running system is uncertain.
```

Perform the following steps to enable jumbo frames and set the MTU correctly.

```
# ifconfig en0 down detach
# lsattr -El ent0 | grep -i jumbo_f
jumbo_frames    no              Request jumbo frames                    True
# chdev -l ent0 -a jumbo_frames=yes
ent0 changed
# lsattr -El ent0 | grep -i jumbo_f
jumbo_frames    yes             Request jumbo frames                    True
# chdev -l en0 -a mtu=9000
en0 changed
# chdev -l en0 -a state=up
# mkdev -l inet0
```

NOTE: SR-IOV adapters and firmware updates:

When you switch the adapter into "shared" mode (for SR-IOV), it is automatically flashed to the firmware level included with the system firmware (e.g. 11.2.211.37). This level is different from the level available in dedicated mode.

"This fix updates the adapter firmware to 11.2.211.37 for the following Feature Codes: EN15, EN17, EN0H, EN0J, EN0M, EN0N, EN0K, EN0L, EL38, EL3C, EL56, and EL57.
The SR-IOV adapter firmware level update for the shared-mode adapters happens under user control to prevent unexpected temporary outages on the adapters. A system reboot will update all SR-IOV shared-mode adapters with the new firmware level. In addition, when an adapter is first set to SR-IOV shared mode, the adapter firmware is updated to the latest level available with the system firmware (and it is also updated automatically during maintenance operations, such as when the adapter is stopped or replaced). And lastly, selective manual updates of the SR-IOV adapters can be performed using the Hardware Management Console (HMC). To selectively update the adapter firmware, follow the

steps given at the IBM Knowledge Center for using HMC to make the updates:
https://www.ibm.com/support/knowledgecenter/en/POWER9/p9efd/p9efd_updating_sriov_firmware.htm.
Note: Adapters that are capable of running in SR-IOV mode, but are currently running in dedicated mode and assigned to a partition, can be updated concurrently either by the OS that owns the adapter or the managing HMC (if OS is AIX or VIOS and RMC is running). "

Full description here: ftp://ftp.software.ibm.com/software/server/firmware/VM-Firmware-Hist.html

So, if you flash the firmware on the adapter, in dedicated mode and then switch the adapter back to shared mode, the firmware level is returned to the level that is included with the system firmware.

Useful HMC commands for displaying SR-IOV adapters:

```
lshwres -m FRAME1 -r sriov --rsubtype adapter
lshwres -m FRAME1 -r sriov --rsubtype physport --level ethc
lshwres -m FRAME1 -r sriov --rsubtype logport --level eth
lshwres -m FRAME1 -r sriov --rsubtype physport --level ethc -F ,phys_port_loc,priority_flow_control_active,trans_flow_control
lshwres -m FRAME1 -r sriov --rsubtype physport --level ethc -F
,phys_port_loc,priority_flow_control_active,trans_flow_control,,recv_flow_control,config_recv_flow_control,,config_trans_flow_con
trol
lshwres -m FRAME1 -r sriov --rsubtype logport --level eth -F lpar_name,location_code
```

Recommended SR-IOV References:

http://aix4admins.blogspot.com/2016/01/sr-iov-vnic.html
https://www.ibm.com/developerworks/community/wikis/home?lang=en#!/wiki/Power%20Systems/page/SR-IOV%20Frequently%20Asked%20Questions
http://aix4admins.blogspot.com/2011/08/hmc-command-line-commands-have-help.html

Special Considerations (a work-around):

In a couple of cases, we encountered an issue, on just a couple of network ports, where LACP would not come up. The issue was isolated to a Nexus 7K switch. LACP configured on these ports, connected to the 7k, would not come up. After much troubleshooting and tracing (on the

switch), it was discovered that the switch was receiving the packets but as untagged instead of tagged. The only difference being the switch & port type and f/w level. Anything connected to other switch types, such as a Nexus 5000 (or FEX ports) worked fine.

In an effort to prevent any further delays to the project, we implemented a work-around (just for those ports only). Essentially we started tagging at the AIX level and not at the logical port layer. We removed the VLAN ID from the logical port configuration and then configured a VLAN tagged device in AIX (and configured IP on this new VLAN interface). This allowed the ports (and LACP) to come up and for the network to function.

The remainder of the LPARs in the environment were left as-is, with the VLAN ID set at the PVID level on the logical port. They continued to work as expected. It appears be something specific to the switch (firmware perhaps?), but we did not ascertain the root cause of the issue.

For information on configurating a VLAN device on AIX, please refer to the AIX Knowledge Center:
https://www.ibm.com/support/knowledgecenter/en/ssw_aix_72/com.ibm.aix.networkcomm/adapters_vlan.htm