NetBITS

AIX ◆ VIOS ◆ Network Protocols ◆ Network Drivers

Insightful tips about APARS, tuning, data collection and more

September 2019

Abstract:

This document has valuable information about known APARs, Network Tuning, Data Collection and useful tips about Network Protocols and Network Device Driver on AIX and VIOS. We want this NETbits to be valuable for you so please share your feedbacks and suggestions by sending email to aix_feedback@wwpdl.vnet.ibm.com.

Product data is subject to change without notice. This information may include technical inaccuracies or typographical errors. IBM may make improvements and/or changes in the product(s) and/or programs(s) at any time without notice. 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. While IBM has reviewed each item for accuracy in a specific situation, there is no guarantee that the same or similar results will be obtained elsewhere.

Contents

APARs

Note: If APAR is not available for download then SP level is not provided because SP level is not known at the time when this document is written.

Network Protocols:

(1) Vulnerability in FreeBSD affects AIX (CVE-2018-6922)

7.1 TL4 SP7 - IJ09620
7.1 TL5 SP4 - IJ09621
7.2 TL1 SP5 - IJ09623
7.2 TL2 SP3 - IJ09624
7.2 TL3 SP3 - IJ09625
2.2.5.50 - IJ09619
2.2.6.40 - IJ09619
3.1.0.20 - IJ09625

Status: Available for download

(2) netstat prints some messages incorrectly

7.1 TL04 - IJ16623
7.1 TL05 - IJ16622
7.2 TL03 - IJ16586
2.2.6 - IJ16624
3.1 - IJ16586

Status: Not available for download currently

(3) PLSO ENABLED SOCKETS DROP PLATFORM LARGE SEND PACKET AFTER LPM

7.1 TL04 - IJ14074
7.1 TL05 - IJ16334
7.2 TL03 - IJ14587

Status: Not available for download currently

Etherchannel:

(1) NIB Etherchannel with SR-IOV VF port problem

7.1 TL03 SP07 - IV77944
7.1 TL04 SP03 - IV82254
7.2 TL00 SP02 - IV80034
7.2 TL01 SP00 - IV82479
7.2 TL02 SP02 – IV80023
2.2.3.70 - IV80127
2.2.3.80 - IV80127
2.2.4.30 - IV80127
2.2.5.0 - IV80127

Status: Available for download

(2) Setting date can affect etherchannel

7.1 TL04 SP07 - IJ10325
7.1 TL05 SP03 - IJ06915
7.2 TL01 SP05 - IJ09841
7.2 TL02 SP03 - IJ09336
7.2 TL03 IJ06942
2.2.5.50 IJ05730
2.2.6.30 IJ05730
2.2.6.31 IJ05730
2.2.6.32 IJ05730

Status: Available for download

Network Device Driver:

Note: There are different network device drivers (e.g. vioentdd, lncentdd, shientdd, lnc2endd, mlxentdd, mlxcentdd etc.) for different ethernet adapter. To find name of the driver name used by ethernet adapter (e.g. ent0, ent1, ent2 etc.), please refer https://www-01.ibm.com/support/docview.wss?uid=ibm10883390.

(1) SMALL CHECKSUM OFFLOAD PACKETS SENT VIA SEA HAVE INVALID TCP SUM

6.1 TL09 SP12-1838 - IJ03186
7.1 TL05 SP02-1832 - IJ04407
7.2 TL01 SP05-1845 - IJ06063
7.2 TL02 SP02-1832 - IJ06064
7.2 TL03 SP00 - IJ04211
2.2.6.23 IJ03186
2.2.6.30 IJ03186

Driver affected: shientdd, lncentdd, lnc2entdd. mlxentdd, mlxcentdd
Status: Available for download

(2) FAILOVER FROM EEH ON SRIOV DEVICES IS NOT HAPPENING

6.1 TL09 SP12 IJ03337
7.1 TL05 SP03 IJ03333
7.2 TL03 IJ03281
2.2.6.31 IJ03337
2.2.6.30 IJ03337

Driver affected: shientdd, lnc2entdd. mlxentdd, mlxcentdd
Status: Available for download

(3) POSSIBLE ERRORS IN CERTAIN NETWORK ADAPTERS

6.1 TL09 SP12 - IJ10300
7.1 TL04 SP03 - IJ10369
7.1 TL05 SP03 - IJ10295
7.2 TL01 SP04 - IJ10370
7.2 TL02 SP01 - IJ10371
7.2 TL03 SP02 - IJ10293
7.2 TL04 - IJ10431
2.2.5.50 - IJ10300
2.2.6.32 - IJ10300

Driver affected: Incentdd, Inc2entdd
Status: Available for download

(4) MBUF MEMORY LEAK IN VIOENTDD

7.1 TL04 - IJ13158
7.2 TL03 - IJ14586

Driver affected: vioentdd
Status: Available for download

(5) DISABLE PLSO ON VEA FOR LOADBALANCER

7.1 TL05 - IJ18129
7.2 TL02 - IJ15766
7.2 TL03 - IJ18460

Driver affected: vioentdd
Status: Not available for download currently

(6) SHIENT_SW_ERR AND LDMP_COMPLETE ERROR IN ERRPT

7.1 TL04 – IJ18291
7.1 TL05 - IJ16499
7.2 TL02 - IJ17240
7.2 TL03 - IJ16925
2.2.6.51 - IJ17674

Driver affected: vioentdd
Status: Not available for download currently

Network Tuning:

(1) largesend/largereceive/mtu_bypass

For optimum throughput:

VIO Client:
mtu_bypass=on

VIO Server:
On SEA: largesend=1, large_receive=yes
On Real Adapter: large_send=on, large_receive=on

(2) jumbo frame (mtu 9000)

If jumbo_frame (i.e. mtu 9000) needs to used then make sure jumbo_frame is enabled on both source and destination hosts and all network devices (i.e. SEA, etherchannel, real adapter, switch, routers etc.).

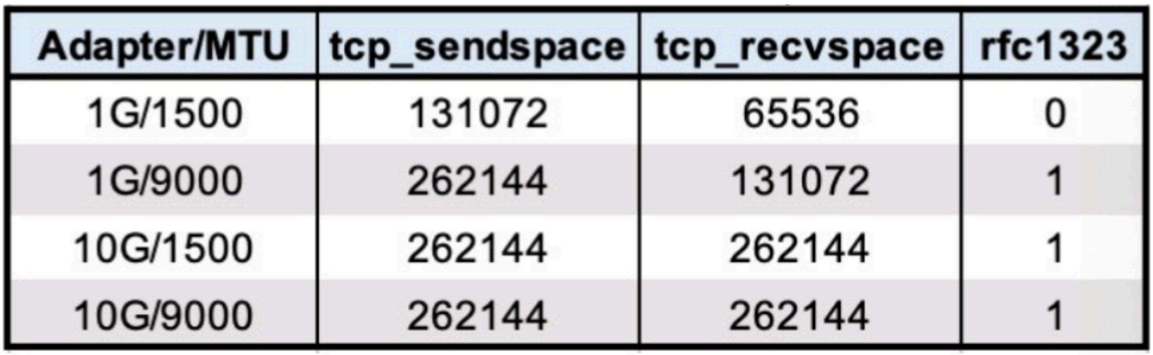
(3) tcp_sendspace/tcp_recvspace/rfc1323

tcp_sendspace specifies the TCP socket buffer size (in bytes) of for sending data.

tcp_recvspace specified the TCP socket buffer size (in bytes) for receiving data.

rfc1323 enables TCP enhancement specified by RFC1323. It allows tcp_sendspace/tcp_recvspace to exceed 64 Kbytes. To use rfc1323 for the connection it needs to be enabled at both source and destination hosts. i.e. If one host has rfc1323 disabled and 2nd host has it enabled then connection will not use rfc1323.

Default Values:



Adapter/MTU	tcp_sendspace	tcp_recvspace	rfc1323
1G/1500	131072	65536	0
1G/9000	262144	131072	1
10G/1500	262144	262144	1
10G/9000	262144	262144	1

Network Tuning:

How to check the setting?

Run “ifconfig -a” to check the settings on interface. If setting is done on interface then “ifconfig -a” will not show the setting. In that case, it uses system wide setting shown by “no -a”.

How to change setting?

To change setting on interface, use “ifconfig enX <option> <value>” or “chdev –l enX –a <option>=<value>”.

Note: change done using ifconfig does not survive the reboot but change done using chdev will survive the reboot.

(4) dog threads

By enabling the dog threads, the driver queues the incoming packet to the thread and the thread handles calling IP, TCP, and the socket code. Drivers, by default, call IP directly, which calls up the protocol stack to the socket level while running on the interrupt level. This minimizes instruction path length, but increases the interrupt hold time.

Enabling the dog threads can increase capacity of the system in some cases, where the incoming packet rate is high, allowing incoming packets to be processed in parallel by multiple CPUs. The down side of the dog threads feature is that it increases latency under light loads and also increases host CPU utilization because a packet has to be queued to a thread and the thread has to be dispatched.

Guidelines when considering using dog threads are as follows:

- More CPUs than adapters need to be installed. Typically, at least two times more CPUs than adapters are recommended.
- Systems with faster CPUs benefit less. Machines with slower CPU speed may be helped the most.
- This feature is most likely to enhance performance when there is high input packet rate. It will enhance performance more on MTU 1500 compared to MTU 9000 (jumbo frames) on Gigabit as the packet rate will be higher on small MTU networks.
- The dog threads run best when they find more work on their queue and do not have to go back to sleep (waiting for input). This saves the overhead of the driver waking up the thread and the system dispatching the thread.

Network Tuning:

- The dog threads can also reduce the amount of time a specific CPU spends with interrupts masked. This can release a CPU to resume typical user-level work sooner.

- The dog threads can also reduce performance by about 10% if the packet rate is not fast enough to allow the thread to keep running. The 10% is an average amount of increased CPU overhead needed to schedule and dispatch the threads.

How to enable dog threads?

```
Ifconfig enX thread
```

How to disable dog threads?

```
Ifconfig enX -thread
```

Note: settings changed by ifconfig does not survive reboot. ifconfig is only command available to set dog threads.

Data Collection Steps:

Scenario 1: AIX VIO Client can’t communicate with remote system located outside of the managed system. e.g. can not ping remote system, can not establish a connection, connection hangs/disconnects etc.

vioc1 = VIO client can’t communicate with remote system
vios1 = primary VIO server of vioc1
remote1 = remote system that vioc1 can’t communicate

(1) On vioc1 and vios1

- Create a directory called /tmp/ibm (Make sure /tmp has enough free space to save data.)

- if SEA interface does not have IP address configured then run "ifconfig <SEA interface> up" to allow iptrace to snif the packets. e.g. if SEA interface is en4 and it does not have IP address configured then run "ifconfig en4 up".

```
ifconfig -a > /tmp/ibm/ifconfig.before
netstat -s > /tmp/ibm/netstat.s.before
netstat -v > /tmp/ibm/netstat.v.before
netstat -in > /tmp/ibm/netstat.in.before
netstat -rn > /tmp/ibm/netstat.rn.before
netstat -ano > /tmp/ibm/netstat.ano.before
arp -an > /tmp/ibm/arp.before
errpt -a > /tmp/ibm/errpt.before
```

(2) On vioc1

```
startsrc -s iptrace -a "-L 300000000 /tmp/ibm/iptrc.bin"
```

(3) On vios1

```
startsrc -s iptrace -a "-L 300000000 /tmp/ibm/iptrc.bin"
```

(4) On vioc1

Use step (a) or (b) or (c) depending upon the problem

(a) Can not ping remote system

```
ping -c 5 <IP address of remote1>
```

All 5 pings should fail.

(b) Can not establish a connection to remote system

Initiate the connection to remote system remote1 using IP address one time and let it fail.

(c) Connection hangs or disconnects

Recreate connection hangs or disconnects with remote system remote1

(5) On vioc1 and vios1

```
stopsrc -s iptrace
```

(6) On vioc1 and vios1

```
ifconfig -a > /tmp/ibm/ifconfig.after
netstat –s > /tmp/ibm/netstat.s.after
netstat -v > /tmp/ibm/netstat.v.after
netstat -in > /tmp/ibm/netstat.in.after
netstat -rn > /tmp/ibm/netstat.rn.after
netstat -ano > /tmp/ibm/netstat.ano.after
arp -an > /tmp/ibm/arp.after
errpt -a > /tmp/ibm/errpt.after
```

(7) Collect snap

on vioc1: run "snap -gtkc" as root
on vios1: run "snap" as padmin

(8) Data needs to be uploaded on IBM

(a) snap of vioc1
(b) snap of vios1
(c) All files in /tmp/ibm directory from vioc1 and vios1

(9) Depending upon the problem, provide following details.

(a) For connection failing to establish, connections hangs or connection disconnects problem, provide IP address of the remote system and tcp server port.

(b) For cannot ping remote system problem, provide IP address of the remote system.

Scenario 2: Application running on AIX VIO Client gives lower throughput when it communicates with AIX VIO Client on another managed system.

vioc1 = VIO client on 1st managed system
vios1 = primary VIOS server of vioc1

vioc2 = VIO client on 2nd managed system
vios2 = primary VIO server of vioc2

Application running on vioc1 gives lower throughput when it communicates with vioc2.

(1) do following on HMC for on vioc1 and vioc2, vios1 and vios2 if they have shared processor.

on HMC 8 (Enhanced view)

- select the lpar
- select General Properties
- click on "Advanced" button on top right corner
- Make sure "Enable Performance Information Collection" option is checked.

(2) On vios1 and pvios2

If SEA interface does not have IP address configured then run "ifconfig <SEA interface> up" to allow iptrace to snif the packets. e.g. if SEA interface is en4 and it does not have IP address configured then run "ifconfig en4 up".

(3) on vioc1, vioc2, vios1 and vios2

- create a directory called /tmp/ibm

```
ifconfig -a > /tmp/ibm/ifconfig.before
netstat -s > /tmp/ibm/netstat.s.before
netstat -v > /tmp/ibm/netstat.v.before
netstat -in > /tmp/ibm/netstat.in.before
netstat -rn > /tmp/ibm/netstat.rn.before
netstat -ano > /tmp/ibm/netstat.ano.before
arp -an > /tmp/ibm/arp.before
errpt -a > /tmp/ibm/errpt.before
```

(4) on vioc1, vioc2, vios1 and vios2

```
lparstat -h -t 3 > /tmp/ibm/lparstat.out 2> /dev/null &
```

(5) On vioc1, vioc2, vios1 and vios2

```
startsrc -s iptrace -a "-L 300000000 /tmp/ibm/iptrc.bin"
```

(6) Recreate lower throughput problem for approximately 10 to 15 seconds.

(7) On vioc1, vioc2, vios1 and vios2

```
stopsrc -s iptrace
```

(8) On vioc1, vioc2, vios1 and vios2

```
ifconfig -a > /tmp/ibm/ifconfig.after
netstat -s > /tmp/ibm/netstat.s.after
netstat -v > /tmp/ibm/netstat.v.after
netstat -in > /tmp/ibm/netstat.in.after
netstat -rn > /tmp/ibm/netstat.rn.after
netstat -ano > /tmp/ibm/netstat.ano.after
arp -an > /tmp/ibm/arp.after
errpt -a > /tmp/ibm/errpt.after
```

(9) On vioc1, vioc2, vios1 and vios2

kill the lparstat

kill the lparstat

(10) Data needs to be uploaded on IBM

- snap of vioc1, vioc2, vios1 and vios2.
- All files in /tmp/ibm directory from vioc1, vioc2, vios1 and vios2.

(11) Provide the tcp server port of the application giving lower throughput

Scenario 3: AIX host with physical ethernet adapter can’t communicate with remote AIX host with physical ethernet adapter.

hostA = AIX host with physical adapter that can't communicate with hostB
hostB = AIX host with physical adapter that can't communicate with hostA

(1) on hostA and hostB

- create a directory called /tmp/ibm

```
ifconfig -a > /tmp/ibm/ifconfig.before
netstat -s > /tmp/ibm/netstat.s.before
netstat -v > /tmp/ibm/netstat.v.before
netstat -in > /tmp/ibm/netstat.in.before
netstat -rn > /tmp/ibm/netstat.rn.before
arp -an > /tmp/ibm/arp.before
errpt -a > /tmp/ibm/errpt.before
```

(2) on hostA and hostB

```
startsrc -s iptrace -a "-L 300000000 /tmp/ibm/iptrc.bin"
```

(3) on hostA

```
ping -c 5 <IP address of hostB>
```

I assume all 5 pings will fail. Put ping output in /tmp/ibm/ping.out.
Note: Please remember to save output in /tmp/ibm/ping.out.

(4) on hostA and and hostB

```
stopsrc -s iptrace
```

(5) on hostA and hostB

```
ifconfig -a > /tmp/ibm/ifconfig.after
netstat -s > /tmp/ibm/netstat.s.after
netstat -v > /tmp/ibm/netstat.v.after
netstat -in > /tmp/ibm/netstat.in.after
netstat -rn > /tmp/ibm/netstat.rn.after
arp -an > /tmp/ibm/arp.after
errpt -a > /tmp/ibm/errpt.after
```

(6) on hostA and hostB

run "snap -gtkc" to collect snap.

(7) Data needs to be uploaded on IBM

- snap from hostA and hostB
- All files in /tmp/ibm directory on hostA and hostB

Useful Tips:

Question: I have following setup to use jumbo frame (i.e. MTU 9000) but communication is not working. What else I can check to communicate using jumbo frame?

AIX VIO Client:

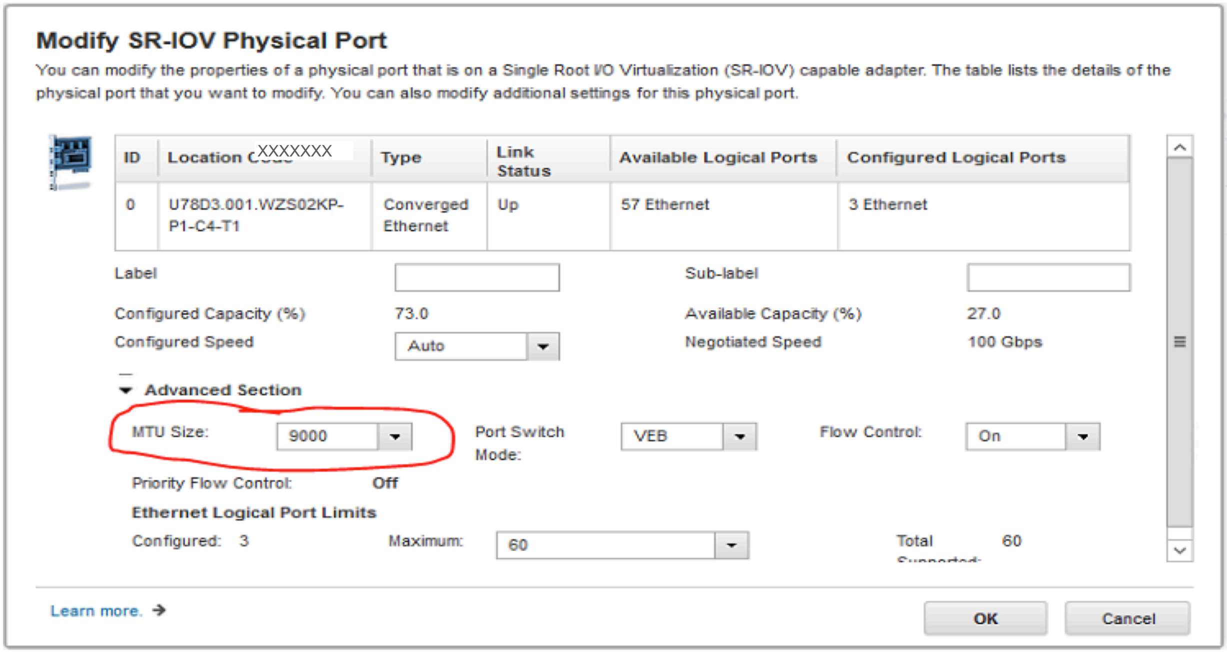
ent0 - Virtual Ethernet Adapter   en0 - mtu: 9000

VIOS:

SEA - ent9 - jumbo_frames: yes
    Real Adapter - ent1 – 2-port 100Gbit RoCE PCIe3 Adapter VF (b31514101410f704)
        Feature Code: EC3M
        Physical location: U78D3.001.XXXXXXX-P1-C4-T1-S1
        jumbo_frames: yes
    pvid adapter - ent2

Switches/Routers/Remote host: jumbo frame is enabled

Answer: ent3 is a VF meaning it is a virtual port created from physical port U78D3.001.XXXXXXX-P1-C4-T1. There is a "MTU Size" attribute for the physical port that can be changed using HMC. The default value is 1500. If it is set to 1500 then communication using jumbo frame packet does not work. It needs to be set to 9000. See HMC screen capture below. If attributes are not set as per above then driver will log MLXCENT_HW_TMP_ERR error in errpt.



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 details of the physical port that you want to modify. You can also modify additional settings for this physical port.

ID	Location CXXXXXXX	Type	Link Status	Available Logical Ports	Configured Logical Ports
0	U78D3.001.WZS02KP-P1-C4-T1	Converged Ethernet	Up	57 Ethernet	3 Ethernet

Label:

Sub-label:

Configured Capacity (%): 73.0

Available Capacity (%): 27.0

Configured Speed: Auto

Negotiated Speed: 100 Gbps

Advanced Section

MTU Size: 9000

Port Switch Mode: VEB

Flow Control: On

Priority Flow Control: Off

Ethernet Logical Port Limits

Configured: 3

Maximum: 60

Total Supported: 60