Session 1087

AIX Performance Tuning Part 3 - Network

Jaqui Lynch

Flagship Solutions Group

jlynch@flagshipsg.com

Edge 2016

The Premier IT Infrastructure Conference

Outthink status quo.

#ibmedge

IBM

Agenda

- Part 1
  - CPU
  - Memory tuning
  - Starter Set of Tunables
- **Part 3**
  - **Network**
  - **Performance Tools**
- Part 2
  - I/O
  - Volume Groups and File systems
  - AIO and CIO

2

FLAGSHIP SOLUTIONS GROUP

Tunables

- **The tcp_recvspace tunable**
  - The *tcp_recvspace* tunable specifies how many bytes of data the receiving system can buffer in the kernel on the receiving sockets queue.
- **The tcp_sendspace tunable**
  - The *tcp_sendspace* tunable specifies how much data the sending application can buffer in the kernel before the application is blocked on a send call.
- **The rfc1323 tunable**
  - The *rfc1323* tunable enables the TCP window scaling option.
  - By default TCP has a 16 bit limit to use for window size which limits it to 65536 bytes. Setting this to 1 allows for much larger sizes (max is 4GB)
- **The sb_max tunable**
  - The *sb_max* tunable sets an upper limit on the number of socket buffers queued to an individual socket, which controls how much buffer space is consumed by buffers that are queued to a sender's socket or to a receiver's socket. *The tcp_sendspace attribute must specify a socket buffer size less than or equal to the setting of the sb_max attribute*

3

UDP Send and Receive

udp_sendspace

Set this parameter to 65536, which is large enough to handle the largest possible
UDP packet. There is no advantage to setting this value larger

udp_recvspace

Controls the amount of space for incoming data that is queued on each UDP socket. Once the *udp_recvspace* limit is reached for a socket, incoming packets are discarded.

Set this value high as multiple UDP datagrams could arrive and have to wait on a socket for the application to read them. If too low packets are discarded and sender has to retransmit.

Suggested starting value for *udp_recvspace* is 10 times the value of *udp_sendspace*, because UDP may not be able to pass a packet to the application before another one arrives.

4

Some definitions

- TCP large send offload
  - Allows AIX TCP to build a TCP message up to 64KB long and send It in one call down the stack. The adapter resegments into multiple packets that are sent as either 1500 byte or 9000 byte (jumbo) frames.
  - Without this it takes 44 calls (if MTU 1500) to send 64KB data. With this set it takes 1 call. Reduces CPU. Can reduce network CPU up to 60-75%.
  - It is enabled by default on 10Gb adapters but not on VE or SEA.
- TCP large receive offload
  - Works by aggregating incoming packets from a single stream into a larger buffer before passing up the network stack. Can improve network performance and reduce CPU overhead.
- TCP Checksum Offload
  - Enables the adapter to compute the checksum for transmit and receive. Offloads CPU by between 5 and 15% depending on MTU size and adapter.

5

Large Receive

- Important note
  - Do not enable on the sea if used by Linux or IBM I client partitions (disabled by default)
  - Do not enable if used by AIX partitions set up for IP forwarding
  - Also called Receive TCP Segment Aggregation
  - If choose to enable this then make sure underlying adapter also has it enabled
  - *See* http://tinyurl.com/gpe5zgd for update on changes for Linux and Large receive
    - *Now supported if VIOS 2.2.4.10 with specific Linux levels*
      - *RHEL 7 rel 2 BE and LE*
      - *SLES 12 SP1*
      - *SLES 11 SP4*
      - *RHEL 6.8*
      - *Ubuntu 16.04 LTS*

6

Some more definitions

- MTU Size
  - The use of large MTU sizes allows the operating system to send fewer packets of a larger size to reach the same network throughput. The larger packets greatly reduce the processing required in the operating system, assuming the workload allows large messages to be sent. If the workload is only sending small messages, then the larger MTU size will not help. Choice is 1500 or 9000 (jumbo frames). Do not change this without talking to your network team.
- MSS – Maximum Segment Size
  - The largest amount of data, specified in bytes, that a computer or communications device can handle in a single, unfragmented piece.
  - The number of bytes in the data segment and the header must add up to less than the number of bytes in the maximum transmission unit (MTU).
- Computers negotiate MTU size
  - Typical MTU size in TCP for a home computer Internet connection is either 576 or 1500 bytes. Headers are 40 bytes long; the MSS is equal to the difference, either 536 or 1460 bytes.

7

More on MTU and MSS

Routed data must pass through multiple gateway routers.

We want each data segment to pass through every router without being fragmented.

If the data segment size is too large for any of the routers through which the data passes, the oversize segment(s) are fragmented.

This slows down the connection speed and the slowdown can be dramatic.

Fragmentation can be minimized by keeping the MSS as small as reasonably possible.

8

Starter set of tunables 3

Typically we set the following for both versions:

NETWORK
no -p -o rfc1323=1
no -p -o tcp_sendspace=262144
no -p -o tcp_recvspace=262144
no -p -o udp_sendspace=65536
no -p -o udp_recvspace=655360

Also check the actual NIC interfaces and make sure they are set to at least these values
You can't set udp_sendspace > 65536 as IP has an upper limit of 65536 bytes per packet

Check sb_max is at least 1040000 – increase as needed

9

ifconfig

ifconfig -a output

```
en0:
flags=1e080863,480<UP,BROADCAST,NOTRAILERS,RUNNING,SIMPLEX,MULTICAST,GROUPRT,
64BIT,CHECKSUM_OFFLOAD(ACTIVE),CHAIN>
        inet 10.2.0.37 netmask 0xfffffe00 broadcast 10.2.1.255
        tcp_sendspace 65536 tcp_recvspace 65536 rfc1323 0
lo0:
flags=e08084b<UP,BROADCAST,LOOPBACK,RUNNING,SIMPLEX,MULTICAST,GROUPRT,64BIT>
        inet 127.0.0.1 netmask 0xff000000 broadcast 127.255.255.255
        inet6 ::1/0
        tcp_sendspace 131072 tcp_recvspace 131072 rfc1323 1
```

These override no, so they will need to be set at the adapter.
Additionally you will want to ensure you set the adapter to the correct setting if it runs at less than GB, rather than allowing auto-negotiate

Stop inetd and use chdev to reset adapter (i.e. en0)
Or use chdev with the –P and the changes will come in at the next reboot
chdev -l en0 -a tcp_recvspace=262144 –a tcp_sendspace=262144 –a rfc1323=1 –P

On a VIO server I normally bump the transmit queues on the real (underlying adapters) for the aggregate/SEA
Example for a 1Gbe adapter:
chdev -l ent? -a txdesc_que_sz=1024 -a tx_que_sz=16384 -P

10

My VIO Server SEA

```
# ifconfig -a
en6:
flags=1e080863,580<UP,BROADCAST,NOTRAILERS,RUNNING,SIMPLEX,
MULTICAST,GROUPRT,64BIT,CHECKSUM_OFFLOAD(ACTIVE),CHAIN>

      inet  192.168.2.5   netmask 0xffffff00 broadcast 192.168.2.255
       tcp_sendspace 262144   tcp_recvspace 262144   rfc1323 1

lo0:
flags=e08084b,1c0<UP,BROADCAST,LOOPBACK,RUNNING,SIMPLEX,M
ULTICAST,GROUPRT,64BIT,LARGESEND,CHAIN>
      inet 127.0.0.1 netmask 0xff000000 broadcast 127.255.255.255
      inet6 ::1%1/0
       tcp_sendspace 131072   tcp_recvspace 131072   rfc1323 1
```

11

Virtual Ethernet

Link aggregation
- Put vio1 aggregate on a different switch to vio2 aggregate
- Provides redundancy without having to use NIB
- Allows full bandwidth and less network traffic (NIB is pingy)
- Basically SEA failover with full redundancy and bandwidth

Pay attention to entitlement
- VE performance scales by entitlement not VPs

If VIOS only handling network then disable network threading on the virtual Ethernet
- chdev –dev ent? thread=0
- Non threaded improves LAN performance
- Threaded (default) is best for mixed vSCSI and LAN

http://www14.software.ibm.com/webapp/set2/sas/f/vios/documentation/perf.html

Turn on large send on VE adapters
- chdev –dev ent? –attr large_send=yes

Turn on large send on the SEA
- chdev –dev entx –attr largesend=1

NOTE do not do this if you are supporting Linux or IBM i LPARs with the VE/SEA

12

SEA Notes

Threaded versus Interrupt mode

Threading is the default and is designed for when both vSCSI and networking are on the same VIO server

- It improves shared performance
- Turning threading off improves network performance
- Only turn threading off if the VIO server only services network traffic

Failover Options

- NIB
  - Client side failover where there are extra unused adapters.
  - Very pingy and wasted bandwidth
  - Requires two virtual adapters and an additional NIB configuration per client
- SEA failover – server side failover.
  - Simpler plus you get to use the bandwidth on all the adapters
- SEA failover with loadsharing
  - Basically use two SEAs with different trunk priorities on the same VLANs

As of VIO 2.2.3 can get rid of control channel

- Requires VLAN 4095 to not be in use
- Requires HMC 7.7.8, VIOs 2.2.3 and firmware 780 minimum
- Not supported on MMB or MHB when announced
- mkvdev–sea ent0 –vadapter ent1 ent2 ent3 –default ent1 –defaulted 11 –attrha_mode=sharing

To find the control channel:

entstat–all ent? | grep–i"Control Channel PVID" where ent? Is the ent interface created above (probably ent4)

13

Network

Interface	Speed	MTU	tcp_sendspace	tcp_recvspace	rfc1323	tcp_nodelay	tcp_mssdflt
lo0 (loopback)	N/A	16896	131072	131072	1		
Ethernet	10 or 100 (Mbit)						
Ethernet	1000 (Gigabit)	1500	131072	65536	1		
Ethernet	1000 (Gigabit)	9000	262144	131072	1		
Ethernet	10 GigE	1500	262144	262144	1		
Ethernet	10 GigE	9000	262144	262144	1		
Ether Channel	Configures based on speed/MTU of the underlying interfaces.						
Virtual Ethernet	N/A	any	262144	262144	1		
InfiniBand	N/A	2044	131072	131072	1		

Above taken from AIX v7.1 Performance Tuning Guide

Check up to date information at:

Aix V5.3
http://www-01.ibm.com/support/knowledgecenter/api/content/ssw_aix_53/com.ibm.aix.prftungd/doc/prftungd/prftungd_pdf.pdf

AIX v6.1
http://www-01.ibm.com/support/knowledgecenter/ssw_aix_71/com.ibm.aix.performance/prftungd_pdf.pdf

AIX v7.1
http://www-01.ibm.com/support/knowledgecenter/ssw_aix_71/com.ibm.aix.performance/prftungd_pdf.pdf

14

10Gbe Ethernet Adapters

15

Network Performance and Throughput

- Depends on:
  - Available CPU power – entitlement at send/receive VIOs and client LPARs
    - **Scales by entitlement not by VPs**
  - MTU size
  - Distance between receiver and sender
  - Offloading features
  - Coalescing and aggregation features
  - TCP configuration
  - Firmware on adapters and server
  - Ensuring all known efixes are on for 10GbE issues

  - Pay attention to adapter type and placement
  - Use lsslot –c pci
    - This helps you figure out what kind of slots you have

16

Notes on 10GbE

- Using jumbo frames better allows you to use the full bandwidth – coordinate with network team first
  - Jumbo frames means an MTU size of 9000
  - Reduces CPU time needed to forward packets larger than 1500 bytes
  - Has no impact on packets smaller than 1500 bytes
  - Must be implemented end to end including virtual Ethernet, SEAs, etherchannels, physical adapters, switches, core switches and routers and even firewalls or you will find they fragment your packets
  - Throughput can improve by as much as 3X on a virtual ethernet
- Manage expectations
  - Going from 1GbE to 10GbE does not mean 10x performance
  - You will need new cables
  - You will use more CPU and memory
    - Network traffic gets buffered
    - This applies to the SEA in the VIOS
- Check that the switch can handle all the ports running at 10Gb
- Make sure the server actually has enough gas to deliver the data to the network at 10Gb

17

10GbE Tips

- Use flow control everywhere – this reduces the need for retransmissions
  - Need to turn it on at the network switch,
  - Turn it on for the adapter in the server
    - chdev –l ent? –a flow_cntrl=yes
- If you need significant bandwidth then dedicate the adapter to the LPAR
  - There are ways to still make LPM work using scripts to temporarily remove the adapter
- TCP Offload settings – largesend and large_receive
  - These improve throughput through the TCP stack
- Set largesend on (TCP segmentation offload) – should be enabled by default on a 10GbE SR adapter
  - AIX - chdev –l en? –a largesend=on
  - On vio – chdev –dev ent? –attr largesend=1
  - With AIX v7 tl1 or v6 tl7 – chdev –l en? –l mtu_bypass=on
- mtu_bypass
  - At 6.1 tl7 sp1 and 7.1 sp1
  - O/s now supports mtu_bypass as an option for the SEA to provide a persistent way to enable largesend
  - See section 9.11 of the AIX on POWER Performance Guide

18

10GbE Tips

- Try setting large_receive on as well (TCP segment aggregation)
  - AIX - chdev –l en? –a large_receive=on
  - VIO – chdev –dev ent? –attr large_receive=1
- If you set large_receive on the SEA the AIX LPARs will inherit the setting
- Consider increasing the MTU size (talk to the network team first) – this increases the size of the actual packets
  - chdev –l en? mtu=65535 (9000 is what we refer to as jumbo frames)
  - This reduces traffic and CPU overhead
- If you use ifconfig to make the changes it does not update ODM so the change does not survive a reboot - use chdev

19

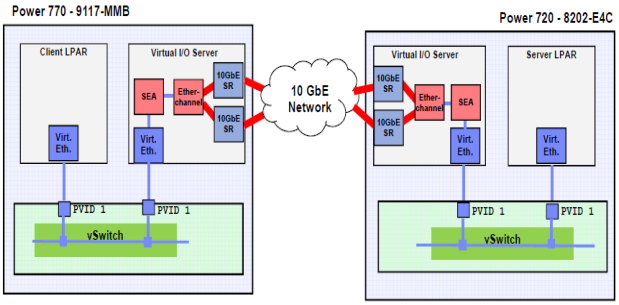
10GbE Tips

- Low CPU entitlement or too few VPs will impact network performance
  - It takes CPU to build those packets
- Consider using netperf to test
- Network speed between two LPARs on the same box is limited to the virtual Ethernet Speed which is about 0.5 to 1.5 Gb/s – much higher on up to date VIOS and HW
  - https://www.ibm.com/developerworks/community/blogs/aixpert/entry/powervm_virtual_ethernet_speed_is_often_confused_with_vios_sea_ive_hea_speed?lang=en
- The speed between two LPARs where one is on the SEA and the other is external is the lower of the virtual Ethernet speed above or the speed of the physical network
- But all VMs on a server can be sending and receiving at the virtual ethernet speed concurrently
- If 10Gb network check out Gareth's Webinar
  - http://public.dhe.ibm.com/systems/power/community/aix/PowerVM_webinars/7_10Gbit_Ethernet.wmv
  - Handout at: https://www.ibm.com/developerworks/wikis/download/attachments/153124943/7_PowerVM_10Gbit_Ethernet.pdf?version=1

20

10GbE Performance

- Diagram below shows all the places network traffic can be affected



Writeup by Nigel Griffiths on Virtual Ethernet Speeds:
https://www.ibm.com/developerworks/community/blogs/aixpert/entry/powervm_virtual_ethernet_speed_is_often_confused_with_vios_sea_ive_hea_speed?lang=en

Check out aPE3489 by Alexander Paul on 10Gb Ethernet Virtualization and Performance Update

21

Testing 10GbE Performance

- FTP is single threaded so not good for testing throughput
  - Unless you run lots of them concurrently
- Use iperf to test bandwidth
  - Useful for TCP and UDP benchmarks
  - Multithreaded
  - Can be run in client or server mode
  - On server run iperf –s
  - On client run something like iperf –c servername –t 60 –P 8
  - Has a GUI java frontend called jperf which allows you to change many settings
- Can also use netperf to test
  - Has TCP_STREAM and TCP_RR benchmarks built in
- jperf is also an option

22

Looking at Performance

23

Network Commands

- entstat –d or netstat –v (also –m and –I)
- netpmon
- iptrace (traces) and ipreport (formats trace)
- tcpdump
- traceroute
- chdev, lsattr
- no
- ifconfig
- ping and netperf or iperf
- ftp
  - Can use ftp to measure network throughput BUT is single threaded
  - ftp to target
  - ftp> put "| dd if=/dev/zero bs=32k count=100" /dev/null
  - Compare to bandwidth (For 1Gbit - 948 Mb/s if simplex and 1470 if duplex)
  - 1Gbit = 0.125 GB = 1000 Mb = 100 MB) but that is 100%

24

netstat -i

netstat -i shows the network interfaces along with input and output packets and errors. It also gives the number of collisions. The Mtu field shows the maximum ip packet size (transfer unit) and should be the same on all systems. In AIX it defaults to 1500.

Both Oerrs (number of output errors since boot) and Ierrs (Input errors since boot) should be < 0.025. If Oerrs>0.025 then it is worth increasing the send queue size. Ierrs includes checksum errors and can also be an indicator of a hardware error such as a bad connector or terminator.

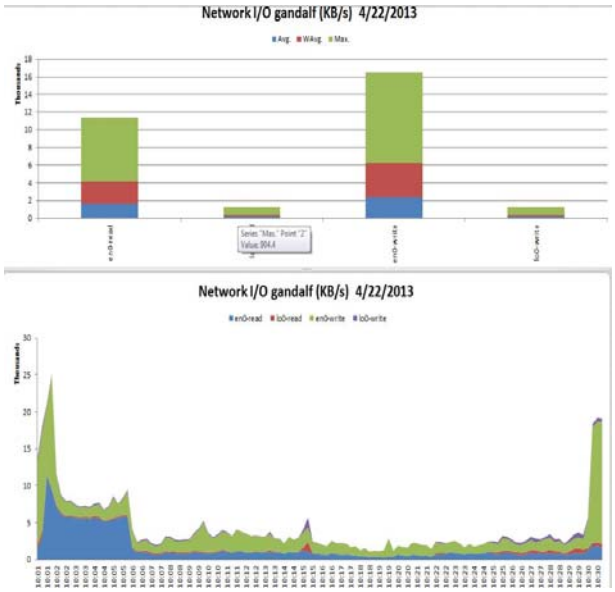
The Collis field shows the number of collisions since boot and can be as high as 10%. If it is greater then it is necessary to reorganize the network as the network is obviously overloaded on that segment.

```
# netstat -i
Name  Mtu   Network     Address                    Ipkts     Ierrs     Opkts     Oerrs     Coll
en6   1500  10.250.134  b740vio1                   4510939   0         535626    0         0

# netstat -i
Name  Mtu   Network     Address                    Ipkts     Ierrs     Opkts     Oerrs     Coll
en5   1500  link#2      a.aa.69.2b.91.c            6484659   0         3009061   0         0
en5   1500  10.250.134  b814vio1                   6484659   0         3009061   0         0
lo0   16896 link#1                                 1289244   0         1289232   0         0
lo0   16896 127         loopback                   1289244   0         1289232   0         0
lo0   16896 ::1%1                                  1289244   0         1289232   0         0
```

25

Net tab in nmon analyser



26

Other Network

- netstat –v
  - Look for overflows and memory allocation failures

    Max Packets on S/W Transmit Queue: 884

    S/W Transmit Queue Overflow: 9522
  - "Software Xmit Q overflows" or "packets dropped due to memory allocation failure"
    - Increase adapter xmit queue
    - Use lsattr –El ent? To see setting
  - Look for receive errors or transmit errors
  - dma underruns or overruns
  - mbuf errors

27

1Gb Adapter (4 port)

```
bnim: lsdev -C | grep ent0
ent0      Available 05-00      4-Port 10/100/1000 Base-TX PCI-Express Adapter (14106803)

bnim: lsattr -El ent0
chksum_offload yes          Enable hardware transmit and receive checksum  True
flow_ctrl      yes          Enable Transmit and Receive Flow Control       True
jumbo_frames   no           Transmit jumbo frames                          True
large_send     yes          Enable hardware TX TCP resegmentation          True
rxbuf_pool_sz  2048         Rcv buffer pool, make 2X rxdesc_que_sz         True
rxdesc_que_sz  1024         Rcv descriptor queue size                      True
tx_que_sz      8192         Software transmit queue size                   True
txdesc_que_sz  512          TX descriptor queue size                       True

bnim: lsattr -El en0
mtu          1500        Maximum IP Packet Size for This Device       True
mtu_bypass   off         Enable/Disable largesend for virtual Ethernet True
remmtu       576         Maximum IP Packet Size for REMOTE Networks   True
tcp_nodelay              Enable/Disable TCP_NODELAY Option            True
thread       off         Enable/Disable thread attribute              True
```

28

10Gb Adapter (SEA)

```
bnim: lsattr -El ent5
ha_mode       auto     High Availability Mode                                  True
jumbo_frames  no       Enable Gigabit Ethernet Jumbo Frames                    True
large_receive no       Enable receive TCP segment aggregation                  True
largesend     1        Enable Hardware Transmit TCP Resegmentation             True
nthreads      7        Number of SEA threads in Thread mode                    True
pvid          1        PVID to use for the SEA device                          True
pvid_adapter  ent4     Default virtual adapter to use for non-VLAN-tagged packets  True
queue_size    8192     Queue size for a SEA thread                             True
real_adapter  ent0     Physical adapter associated with the SEA                True
thread        1        Thread mode enabled (1) or disabled (0)                 True
virt_adapters ent4     List of virtual adapters associated with the SEA (comma separated) True

bnim: lsattr -El en7
mtu          1500       Maximum IP Packet Size for This Device      True
mtu_bypass   off        Enable/Disable largesend for virtual Ethernet True
remmtu       576        Maximum IP Packet Size for REMOTE Networks  True
tcp_nodelay             Enable/Disable TCP_NODELAY Option           True
thread       off        Enable/Disable thread attribute             True
```

Also need to look at Virtual Ethernet values as well as underlying real adapters

29

tcp_nodelayack

- tcp_nodelayack
  - Disabled by default
  - TCP delays sending Ack packets by up to 200ms, the Ack attaches to a response, and system overhead is minimized
  - Tradeoff if enable this is more traffic versus faster response
  - Reduces latency but increases network traffic
  - The *tcp_nodelayack* option prompts TCP to send an immediate acknowledgement, rather than the potential 200 ms delay. Sending an immediate acknowledgement might add a little more overhead, but in some cases, greatly improves performance.
  - Can help with Oracle performance and TSM restore performance
  - Can also flood the network
  - Dynamic change – recommend testing as a standalone change and monitoring network

  - To set – either: chdev -l en0 -a tcp_nodelay=1
  - OR: no -p -o tcp_nodelayack=1
  - See IBM articles at:
    - http://www-01.ibm.com/support/docview.wss?uid=swg21385899
    - http://www-01.ibm.com/support/docview.wss?uid=swg21449348

30

Other Network

- lparstat 2
  - High vcsw (virtual context switch) rates can indicate that your LPAR or VIO server does not have enough entitlement
- ipqmaxlen
  - netstat –s and look for ipintrq overflows
  - ipqmaxlen is the only tunable parameter for the IP layer
  - It controls the length of the IP input queue – default is 100
  - Tradeoff is reduced packet dropping versus CPU availability for other processing
- **Also check errpt – people often forget this**

31

TCP Analysis

```
netstat –p tcp
tcp:
        1629703864 packets sent
                684667762 data packets (1336132639 bytes)
                117291 data packets (274445260 bytes) retransmitted
        955002144 packets received
                1791682 completely duplicate packets (2687306247 bytes)
                0 discarded due to listener's queue full
        4650 retransmit timeouts
        0 packets dropped due to memory allocation failure
```

1. Compare packets sent to packets retransmitted – retransmits should be <5-10%
   1. Above is 0.168
2. Compare packets received with completely duplicate packets – duplicates should be <5-10%
   1. Above is 2.81
3. In both these cases the problem could be a bottleneck on the receiver or too much network traffic
4. Look for packets discarded because listeners queue is full – could be a buffering issue at the sender

32

IP Stack

ip:

955048238 total packets received
0 bad header checksums
0 fragments received
0 fragments dropped (dup or out of space)
0 fragments dropped after timeout

1. If bad header checksum or fragments dropped due to dup or out of space
   1. Network is corrupting packets or device driver receive queues are too small
2. If fragments dropped after timeout >0
   1. Look at ipfragttl as this means the time to life counter for the ip fragments expired before all the fragments of the datagram arrived. Could be due to busy network or lack of mbufs.
3. Review ratio of packets received to fragments received
   1. For small MTU if >5-10% packets getting fragmented then someone is passing packets greater than the MTU size

33

ipqmaxlen

Default is 100

Only tunable parameter for IP
Controls the length of the IP input queue
netstat –p ip
Look for ipintrq overflows

Default of 100 allows up to 100 packets to be queued up

If increase it there could be an increase in CPU used in the off-level interrupt handler
Tradeoff is reduced packet dropping versus CPU availability for other processing

34

netstat –v vio

SEA

Transmit Statistics:	Receive Statistics:
---	---
Packets: 83329901816	Packets: 83491933633
Bytes: 87482716994025	Bytes: 87620268594031
Interrupts: 0	Interrupts: 18848013287
Transmit Errors: 0	Receive Errors: 0
Packets Dropped: 0	**Packets Dropped: 67836309**
	Bad Packets: 0

Max Packets on S/W Transmit Queue: 374
S/W Transmit Queue Overflow: 0
Current S/W+H/W Transmit Queue Length: 0

Elapsed Time: 0 days 0 hours 0 minutes 0 seconds

Broadcast Packets: 1077222	Broadcast Packets: 1075746
Multicast Packets: 3194318	Multicast Packets: 3194313
No Carrier Sense: 0	CRC Errors: 0
DMA Underrun: 0	DMA Overrun: 0
Lost CTS Errors: 0	Alignment Errors: 0
Max Collision Errors: 0	**No Resource Errors: 67836309**

Virtual I/O Ethernet Adapter (l-lan) Specific Statistics:

Hypervisor Send Failures: 4043136
Receiver Failures: 4043136
Send Errors: 0
Hypervisor Receive Failures: 67836309

"No Resource Errors" can occur when the appropriate amount of memory can not be added quickly to vent buffer space for a workload situation.

You can also see this on LPARs that use virtual Ethernet without an SEA

35

Buffers as seen on VIO SEA or Virtual Ethernet

 # lsattr -El ent5

```
alt_addr        0x000000000000 Alternate Ethernet Address          True
chksum_offload  yes            Checksum Offload Enable             True
copy_buffs      32             Transmit Copy Buffers               True
copy_bytes      65536          Transmit Copy Buffer Size           True
desired_mapmem  0              I/O memory entitlement reserved for device False
max_buf_control 64             Maximum Control Buffers             True
max_buf_huge    64             Maximum Huge Buffers                True
max_buf_large   64             Maximum Large Buffers               True
max_buf_medium  256            Maximum Medium Buffers              True
max_buf_small   2048           Maximum Small Buffers               True
max_buf_tiny    2048           Maximum Tiny Buffers                True
min_buf_control 24             Minimum Control Buffers             True
min_buf_huge    24             Minimum Huge Buffers                True
min_buf_large   24             Minimum Large Buffers               True
min_buf_medium  128            Minimum Medium Buffers              True
min_buf_small   512            Minimum Small Buffers               True
min_buf_tiny    512            Minimum Tiny Buffers                True
poll_uplink     no             Enable Uplink Polling               True
poll_uplink_int 1000           Time interval for Uplink Polling    True
trace_debug     no             Trace Debug Enable                  True
use_alt_addr    no             Enable Alternate Ethernet Address   True
```

36

Buffers (VIO SEA or virtual ethernet)

Virtual Trunk Statistics
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	513	2042	128	24	24
Registered	511	506	128	24	24
History					
Max Allocated	532	**2048**	128	24	24
Lowest Registered	502	354	128	24	24

"Max Allocated" represents the maximum number of buffers ever allocated
"Min Buffers" is number of pre-allocated buffers
"Max Buffers" is an absolute threshhold for how many buffers can be allocated

chdev –l <veth> -a max_buf_small=4096 –P
chdev –l <veth> -a min_buf_small=2048 –P
Above increases min and max small buffers for the virtual ethernet adapter configured for the SEA above
Needs a reboot

Max buffers is an absolute threshold for how many buffers can be allocated
Use entstat –d (-all on vio) or netstat –v to get this information

37

UDP Analysis

netstat –p udp
udp:

42963 datagrams received
0 incomplete headers
0 bad data length fields
0 bad checksums
41 dropped due to no socket
9831 broadcast/multicast datagrams dropped due to no socket
0 socket buffer overflows
33091 delivered
27625 datagrams output

1. Look for bad checksums (hardware or cable issues)
2. Socket buffer overflows
   1. Could be out of CPU or I/O bandwidth
   2. Could be insufficient UDP transmit or receive sockets, too few nfsd daemons or too small nfs_socketsize or udp_recvspace

38

Detecting UDP Packet losses

- Run netstat –s or netstat –p udp
- Look under the ip: section for fragments dropped (dup or out of space)
    - Increase udp_sendspace

ip:

8937989 total packets received

……

0 fragments dropped (dup or out of space)

39

Detecting UDP Packet losses

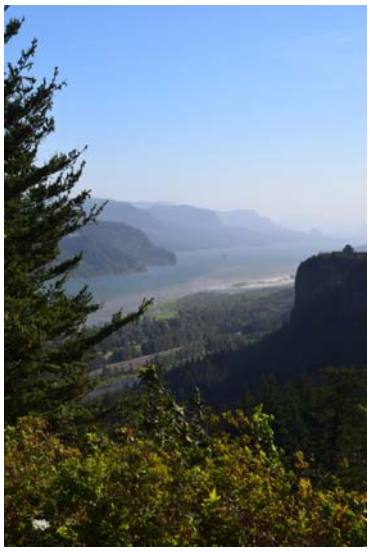
- Look under the udp: section for socket buffer overflows
    - These mean you need to increase udp_recvspace
- UDP packets tend to arrive in bursts so we typically set UDP receive to 10x UDP send. This provides staging to allow packets to be passed through.
- If a UDP packet arrives for a socket with a full buffer then it is discarded by the kernel
- Unlike TCP, UDP senders do not monitor the receiver to see if they have exhausted buffer space

udp:

1820316 datagrams received
0 incomplete headers
0 bad data length fields
0 bad checksums
324375 dropped due to no socket
28475 broadcast/multicast datagrams dropped due to no socket
0 socket buffer overflows
1467466 delivered
1438843 datagrams output

40

Performance Tools



41

DPO (Dynamic Platform Optimizer)

- PowerVM feature requiring firmware 760 and HMC 7.6.0 or greater
- DPO aware O/S
  - AIX v6.1 TL08, VIOS 2.2.2.0, AIX v7.1 TL02, IBM I 7.1 PTF MF56058, RHEL 7, SLES 12
  - Earlier systems will work but will need a reboot after running DPO
- No value on single-socket servers
- Used to improve system wide partition memory and processor placement (affinity)
- Tries to assign local memory to CPUs, reducing memory access time
- Launched via HMC using the optmem command
- lsoptmem command shows current and predicted memory affinity
- See chapter 15 of the PowerVM Managing and Monitoring Redbook SG24-7590
  - http://www.redbooks.ibm.com/redbooks/pdfs/sg247590.pdf

42

Tools

topas
- New –L flag for LPAR view

nmon

nmon analyzer
- Windows tool so need to copy the .nmon file over in ascii mode
- Opens as an excel spreadsheet and then analyses the data
- Also look at nmon consolidator

sar
- sar -A -o filename 2 30 >/dev/null
- Creates a snapshot to a file – in this case 30 snaps 2 seconds apart
- Must be post processed on same level of system

errpt

Check for changes from defaults

ioo, vmo, schedo, vmstat –v

lvmo

lparstat, mpstat

iostat

Check out Alphaworks for the Graphical LPAR tool

Ganglia - http://ganglia.info

Nmonrrd and nmon2web and pGraph

Commercial IBM
- PM for AIX
- Performance Toolbox
- Tivoli ITM

https://www.ibm.com/developerworks/wikis/display/WikiPtype/Other+Performance+Tools

43

Other tools

- filemon
  - filemon -v -o filename -O all
  - sleep 30
  - trcstop
- pstat to check async I/O in 5.3
  - pstat -a | grep aio | wc –l
- perfpmr to build performance info for IBM if reporting a PMR
  - /usr/bin/perfpmr.sh 300
- Network
- iPerf
  - http://www.oss4aix.org/download/RPMS/iperf/
- netperf
  - ftp://ftp.netperf.org/netperf

44

nmon and New Features for V12

- Must be running nmon12e or higher
- Nmon comes with AIX at 5.3 tl09 or 6.1 tl01 and higher BUT on 5.3 I download the latest version from the web so I get the latest v12 for sure
- Creates a file in the working directory that ends .nmon
- This file can be transferred to your PC and interpreted using nmon analyser or other tools

- Disk Service Times
- Selecting Particular Disks
- Time Drift
- Multiple Page Sizes
- Timestamps in UTC & no. of digits
- More Kernel & Hypervisor Stats *
- High Priority nmon
- Virtual I/O Server SEA
- Partition Mobility (POWER6)
- WPAR & Application Mobility (AIX6)
- Dedicated Donating (POWER6)
- Folded CPU count (SPLPAR)
- Multiple Shared Pools (POWER6)
- Fibre Channel stats via entstat

45

nmon Monitoring

- nmon -ft –AOPV^dMLW -s 15 -c 120
  - Grabs a 30 minute nmon snapshot
  - A is async IO
  - M is mempages
  - t is top processes
  - L is large pages
  - **O is SEA on the VIO**
  - P is paging space
  - V is disk volume group
  - d is disk service times
  - ^ is fibre adapter stats
  - W is workload manager statistics if you have WLM enabled

If you want a 24 hour nmon use:

nmon -ft –AOPV^dMLW -s 150 -c 576

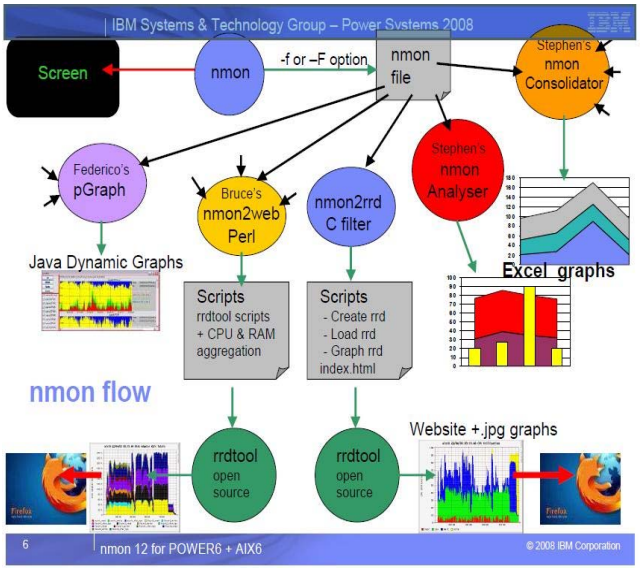
May need to enable accounting on the SEA first – this is done on the VIO
chdev –dev ent* -attr accounting=enabled

Can use entstat/seastat or topas/nmon to monitor – this is done on the vios
topas –E
nmon -O

VIOS performance advisor also reports on the SEAs

46

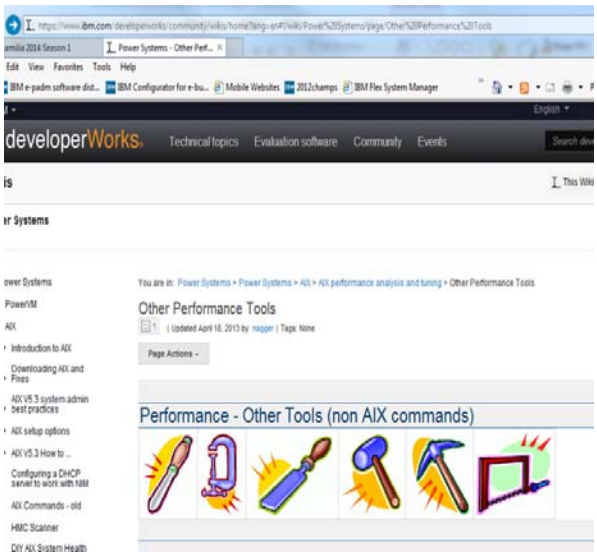
nmon



Courtesy Nigel Griffiths - IBM

47

Performance Wiki



Has links to:
HMC Scanner
AIX memory use analyzer
VIOS Performance Advisor
Java Performance Advisor
PowerVM Virtualization Perf. Advisor
Visual Performance Advisor
gmon
nmon
nmon analyser
nmon consolidator
nmon2web
nmon2rrd
pgraph
nstress – for stress testing

And many other performance tools

48

VIOS Advisor

- https://www.ibm.com/developerworks/community/wikis/home?lang=en#!/wiki/Power%20Systems/page/VIOS%20Advisor
- Application that collects performance metrics and does a health check
- Productized in VIOS 2.2.2
  - part command
- Current downloadable version is 030813A
- Also new HMC performance & capacity monitor
  - HMC v8.8
  - http://www-01.ibm.com/support/docview.wss?uid=nas8N1020114

49

PowerVP

•Included with PowerVM Enterprise, optional add-on to Standard

•Real time graphical monitor
•Customizable thresholds and alerts
•Replay abilities
•System and drill down performance views
•AIX, Linux and IBM i VMs
•Background data collection
•**PowerVP v1.1.3 Enhancements**
- –Export of PowerVP performance data to external repository
- –Integration with VIOS performance advisor
- –New thresholds and alerts
- –LE PowerVM guest support
- –E850 hardware support
- –PowerVP performance monitor browser support

50

Thank you for your time

If you have questions please email me at:
jaquilynch@gmail.com

Also check out:
http://www.circle4.com/movies/

52

Useful Links

- Jaqui Lynch Articles
  - http://www.ibmsystemsmag.com/authors/Jaqui-Lynch/
  - https://enterprisesystemsmedia.com/author/jaqui-lynch
- Jay Kruemke Twitter – chromeaix
  - https://twitter.com/chromeaix
- Nigel Griffiths Twitter – mr_nmon
  - https://twitter.com/mr_nmon
- Gareth Coates Twitter – power_gaz
  - https://twitter.com/power_gaz
- Jaqui's Upcoming Talks and Movies
  - Upcoming Talks
    - http://www.circle4.com/forsythetalks.html
  - Movie replays
    - http://www.circle4.com/movies

53

Useful Links

- HMC Scanner
  - https://www.ibm.com/developerworks/community/wikis/home?lang=en#!/wiki/Power%20Systems/page/HMC%20Scanner
- Workload Estimator
  - http://ibm.com/systems/support/tools/estimator
- Performance Tools Wiki
  - https://www.ibm.com/developerworks/community/wikis/home?lang=en#!/wiki/Power%20Systems/page/AIX%20Performance%20Commands
  - Performance Monitoring
  - https://www.ibm.com/developerworks/community/wikis/home?lang=en#!/wiki/Power%20Systems/page/Performance%20Monitoring%20Tips%20and%20Techniques
  - Other Performance Tools
  - https://www.ibm.com/developerworks/community/wikis/home?lang=en#!/wiki/Power+Systems/page/Other+Performance+Tools
  - Includes new advisors for Java, VIOS, Virtualization
- VIOS Advisor
  - https://www.ibm.com/developerworks/community/wikis/home?lang=en#/wiki/Power%20Systems/page/VIOS%20Advisor

54

References

- Processor Utilization in AIX by Saravanan Devendran
  - https://www.ibm.com/developerworks/mydeveloperworks/wikis/home?lang=en#/wiki/Power%20Systems/page/Understanding%20CPU%20utilization%20on%20AIX
- Rosa Davidson Back to Basics Part 1 and 2 –Jan 24 and 31, 2013
  - https://www.ibm.com/developerworks/mydeveloperworks/wikis/home?lang=en#/wiki/Power%20Systems/page/AIX%20Virtual%20User%20Group%20-%20USA
- SG24-7940 - PowerVM Virtualization - Introduction and Configuration
  - http://www.redbooks.ibm.com/redbooks/pdfs/sg247940.pdf
- SG24-7590 – PowerVM Virtualization – Managing and Monitoring
  - http://www.redbooks.ibm.com/redbooks/pdfs/sg247590.pdf
- SG24-8171 – Power Systems Performance Optimization
  - http://www.redbooks.ibm.com/redbooks/pdfs/sg248171.pdf
- Redbook Tip on Maximizing the Value of P7 and P7+ through Tuning and Optimization
  - http://www.redbooks.ibm.com/technotes/tips0956.pdf

55

Backup Slides

56

NETWORK TUNING in AIX

See article at:
http://www.ibmsystemsmag.com/aix/administrator/networks/network_tuning/

Replay at:
http://www.circle4.com/movies/

57

Network Performance and Throughput

Table 6. Maximum network payload speeds versus duplex TCP streaming rates

Network type	Raw bit rate (Mbits)	Payload rate (Mb)	Payload rate (MB)
10 Mb Ethernet, Half Duplex	10	5.8	0.7
10 Mb Ethernet, Full Duplex	10 (20 Mb full duplex)	18	2.2
100 Mb Ethernet, Half Duplex	100	58	7.0
100 Mb Ethernet, Full Duplex	100 (200 Mb full duplex)	177	21.1
1000 Mb Ethernet, Full Duplex, MTU 1500	1000 (2000 Mb full duplex)	1811 (1667 peak) [1]	215 (222 peak) [1]
1000 Mb Ethernet, Full Duplex, MTU 9000	1000 (2000 Mb full duplex)	1936 (1938 peak) [1]	231 (231 peak) [1]
10 Gb Ethernet, Full Duplex, MTU 1500	10000 (20000 Mb full duplex)	14400 (18448 peak) [1]	1716 (2200 peak) [1]
10 Gb Ethernet, Full Duplex, MTU 9000	10000 (20000 Mb full duplex)	18000 (19555 peak) [1]	2162 (2331 peak) [1]
FDDI, MTU 4352 (default)	100	97	11.6
ATM 155, MTU 1500	155 (310 Mb full duplex)	180	21.5
ATM 155, MTU 9180 (default)	155 (310 Mb full duplex)	236	28.2
ATM 622, MTU 1500	622 (1244 Mb full duplex)	476	56.7
ATM 622, MTU 9180 (default)	622 (1244 Mb full duplex)	884	105

[1] The values in the table indicate rates for dedicated adapters on dedicated partitions. Performance for 10 Gigabit Ethernet adapters in virtual Ethernet Adapter (in VIOS) or Shared Ethernet Adapters (SEA) or for shared partitions (shared LPAR) is not represented in the table because performance is impacted by other variables and tuning that is outside the scope of this table.

58 AIX v7.1 - http://www-01.ibm.com/support/knowledgecenter/ssw_aix_71/com.ibm.aix.performance/prftungd_prftuse

Valid Adapters for P7 and P7+

- 770
  - Multifunction Cards – up to one per CEC
    - 1768 Integrated Multifunction Card with Copper SFP+ - Dual 10Gb copper and dual 10/100/1000MB copper ethernet
    - 1769 Integrated Multifunction Card with SR Optical - Dual 10Gb optical and dual 10/100/1000MB copper ethernet
- PCIE Adapters
  - 5284/5287 PCIE2 – 2 port 10GbE SR (5284 is low profile)
  - 5286/5288 PCIE2 – 2 port 10GbE SFP+ Copper (5286 is low profile)
  - 5769 PCIE1.1 – 1 port 10GbE SR
  - 5772 PCIE1.1 – 1 port 10GbE LR
  - EC27/EC28 PCIE2 – 2 port 10GbE RoCE SFP+ (EC27 is low profile)
  - EC29/EC30 PCIE2 – 2 port 10Gbe RoCE SR (EC29 is low profile)
  - 5708 PCIE – 2 port 10Gb FCoE converged network adapter
- Basically SR is fibre and SFP+ is copper twinax
- **If using SFP+ IBM only supports their own cables** – they come in 1m, 3m and 5m and are 10GbE SFP+ active twinax cables
- Use the PCIE2 cards wherever possible
- RoCE – Supports the InfiniBand trade association (IBTA) standard for remote direct memory access (RDMA) over converged Ethernet (RoCE)
- More information on adapters at:

http://pic.dhe.ibm.com/infocenter/powersys/v3r1m5/topic/p7hcd/pcibyfeature_77x_78x.htm

NOTE SFP+ adapters are not available for B model 770s

59

Adapter Options and Defaults

Table 7. Adapters and their available options, and system default settings

Adapter type	Feature code	TCP checksum offload	Default setting	TCP large send	Default setting
GigE, PCI, SX & TX	2969, 2975	Yes	OFF	Yes	OFF
GigE, PCI-X, SX and TX	5700, 5701	Yes	ON	Yes	ON
GigE dual port PCI-X, TX and SX	5706, 5707	Yes	ON	Yes	ON
10 GigE PCI-X LR and SR	5718, 5719	Yes	ON	Yes	ON
10/100 Ethernet	4962	Yes	ON	Yes	OFF
ATM 155, UTP & MMF	4953, 4957	Yes (transmit only)	ON	No	N/A
ATM 622, MMF	2946	Yes	ON	No	N/A

60

PCI Adapter transmit Queue Sizes

Table 10. Examples of PCI adapter transmit queue sizes

Adapter Type	Feature Code	ODM attribute	Default value	Range
IBM 10/100 Mbps Ethernet PCI Adapter	2968	tx_que_size	8192	16-16384
10/100 Mbps Ethernet Adapter II	4962	tx_que_sz	8192	512-16384
Gigabit Ethernet PCI (SX or TX)	2969, 2975	tx_que_size	8192	512-16384
Gigabit Ethernet PCI (SX or TX)	5700, 5701, 5706, 5707	tx_que_sz	8192	512-16384
10 Gigabit Ethernet PCI-X (LR or SR)	5718, 5719	tx_que_sz	8192	512-16384
ATM 155 (MMF or UTP)	4953, 4957	sw_txq_size	2048	50-16384
ATM 622 (MMF)	2946	sw_txq_size	2048	128-32768
FDDI	2741, 2742, 2743	tx_queue_size	256	3-2048

For adapters that provide hardware queue limits, changing these values will cause more real memory to be consumed on receives because of the control blocks and buffers associated with them. Therefore, raise these limits only if needed or for larger systems where the increase in memory use is negligible. For the software transmit queue limits, increasing these limits does not increase memory usage. It only allows packets to be queued that were already allocated by the higher layer protocols.

61

PCI Adapter Receive Queue Sizes

Table 11. Examples of PCI adapter receive queue sizes

Adapter Type	Feature Code	ODM attribute	Default value	Range
IBM 10/100 Mbps Ethernet PCI Adapter	2968	rx_que_size rx_buf_pool_size	256 384	16, 32 ,64, 128, 26 16-2048
10/100 Mbps Ethernet PCI Adapter II	4962	rx_desc_que_sz rxbuf_pool_sz	512 1024	100-1024 512-2048
Gigabit Ethernet PCI (SX or TX)	2969, 2975	rx_queue_size	512	512 (fixed)
Gigabit Ethernet PCI-X (SX or TX)	5700, 5701, 5706, 5707, 5717, 5768, 5271, 5274, 5767, and 5281	rxbuf_pool_sz rxdesc_que_sz	2048 1024	512-16384,1 128-3840,128
10 Gigabit PCI-X (SR or LR)	5718, 5719	rxdesc_que_sz rxbuf_pool_sz	1024 2048	128-1024, by 128 512-2048
ATM 155 (MMF or UTP)	4953, 4957	rx_buf4k_min	x60	x60-x200 (96-512)
ATM 622 (MMF)	2946	rx_buf4k_min rx_buf4k_max	256 [2] 0 [1]	0-4096 0-14000
FDDI	2741, 2742, 2743	RX_buffer_cnt	42	1-512

62

txdesc_que_sz

Some drivers allow you to tune the size of the transmit ring or the number of transmit descriptors.

The hardware transmit queue controls the maximum number of buffers that can be queued to the adapter for concurrent transmission. One descriptor typically only points to one buffer and a message might be sent in multiple buffers. Many drivers do not allow you to change the parameters.

Adapter type	Feature code	ODM attribute	Default value	Range
Gigabit Ethernet PCI-X, SX or TX	5700, 5701, 5706, 507	txdesc_que_sz	512	128-1024, multiple of 128

63

Definitions – tcp_recvspace

tcp_recvspace specifies the system default socket buffer size for receiving data. This affects the window size used by TCP. Setting the socket buffer size to 16KB (16,384) improves performance over Standard Ethernet and token-ring networks. The default is a value of 4096; however, a value of 16,384 is set automatically by the rc.net file or the rc.bsdnet file (if Berkeley-style configuration is issued).

Lower bandwidth networks, such as Serial Line Internet Protocol (SLIP), or higher bandwidth networks, such as Serial Optical Link, should have different optimum buffer sizes. The optimum buffer size is the product of the media bandwidth and the average round-trip time of a packet. tcp_recvspace network option can also be set on a per interface basis via the chdev command.

Optimum_window = bandwidth * average_round_trip_time

The tcp_recvspace attribute must specify a socket buffer size less than or equal to the setting of the sb_max attribute

Settings above 65536 require that rfc1323=1 (default is 0)

64

Definitions – tcp_sendspace

tcp_sendspace Specifies the system default socket buffer size for sending data. This affects the window size used by TCP. Setting the socket buffer size to 16KB (16,384) improves performance over Standard Ethernet and Token-Ring networks. The default is a value of 4096; however, a value of 16,384 is set automatically by the rc.net file or the rc.bsdnet file (if Berkeley-style configuration is issued).

Lower bandwidth networks, such as Serial Line Internet Protocol (SLIP), or higher bandwidth networks, such as Serial Optical Link, should have different optimum buffer sizes. The optimum buffer size is the product of the media bandwidth and the average round-trip time of a packet. tcp_sendspace network option can also be set on a per interface basis via the chdev command.

Optimum_window = bandwidth * average_round_trip_time

The tcp_sendspace attribute must specify a socket buffer size less than or equal to the setting of the sb_max attribute

Settings above 65536 require that rfc1323=1 (default is 0)

65

Definitions – netstat -m

netstat -m s used to analyze the use of mbufs in order to determine whether these are the bottleneck. The no -a command is used to see what the current values are. Values of interest are thewall, lowclust, lowmbuf and dogticks.

An mbuf is a kernel buffer that uses pinned memory and is used to service network communications. Mbufs come in two sizes - 256 bytes and 4096 bytes (clusters of 256 bytes).

Thewall is the maximum memory that can be taken up for mbufs. Lowmbuf is the minimum number of mbufs to be kept free while lowclust is the minimum number of clusters to be kept free. Mb_cl_hiwat is the maximum number of free buffers to be kept in the free buffer pool and should be set to at least twice the value of lowclust to avoid thrashing.

NB by default AIX sets thewall to half of memory which should be plenty. It is now a restricted tunable.

```
# no -a -F | grep thewall
        thewall = 1572864
# vmstat 1 1

System configuration: lcpu=4 mem=3072MB ent=0.50
```

66

netstat –m – Field meanings

You can use the **netstat -Zm** command to clear (or zero) the mbuf statistics. This is helpful when running tests to start with a clean set of statistics. The following fields are provided with the **netstat -m** command:

Field name	Definition
By size	Shows the size of the buffer.
inuse	Shows the number of buffers of that particular size in use.
calls	Shows the number of calls, or allocation requests, for each sized buffer.
failed	Shows how many allocation requests failed because no buffers were available.
delayed	Shows how many calls were delayed if that size of buffer was empty and theM_WAIT flag was set by the caller.
free	Shows the number of each size buffer that is on the free list, ready to be allocated.
hiwat	Shows the maximum number of buffers, determined by the system, that can remain on the free list. Any free buffers above this limit are slowly freed back to the system.
freed	Shows the number of buffers that were freed back to the system when the free count when above the hiwat limit.

67 http://www-01.ibm.com/support/knowledgecenter/ssw_aix_71/com.ibm.aix.performance/prftungd_pdf.pdf

netstat –v – Field meanings

Transmit and Receive Errors
Number of output/input errors encountered on this device. This field counts unsuccessful transmissions due to hardware/network errors. These unsuccessful transmissions could also slow down the performance of the system.

Max Packets on S/W Transmit Queue
Maximum number of outgoing packets ever queued to the software transmit queue. An indication of an inadequate queue size is if the maximal transmits queued equals the current queue size (xmt_que_size). This indicates that the queue was full at some point. To check the current size of the queue, use the lsattr -El adapter command (where adapter is, for example, ent0). Because the queue is associated with the device driver and adapter for the interface, use the adapter name, not the interface name. Use the SMIT or the chdev command to change the queue size.

S/W Transmit Queue Overflow
Number of outgoing packets that have overflowed the software transmit queue. A value other than zero requires the same actions as would be needed if the Max Packets on S/W Transmit Queue reaches the xmt_que_size. The transmit queue size must be increased.

http://www-01.ibm.com/support/knowledgecenter/ssw_aix_71/com.ibm.aix.performance/prftungd_pdf.pdf

68

netstat –v – Field meanings

Broadcast Packets
Number of broadcast packets received without any error. If the value for broadcast packets is high, compare it with the total received packets. The received broadcast packets should be less than 20 percent of the total received packets. If it is high, this could be an indication of a high network load; use multicasting. The use of IP multicasting enables a message to be transmitted to a group of hosts, instead of having to address and send the message to each group member individually.

DMA Overrun
The DMA Overrun statistic is incremented when the adapter is using DMA to put a packet into system memory and the transfer is not completed. There are system buffers available for the packet to be placed into, but the DMA operation failed to complete. This occurs when the MCA bus is too busy for the adapter to be able to use DMA for the packets. The location of the adapter on the bus is crucial in a heavily loaded system. Typically an adapter in a lower slot number on the bus, by having the higher bus priority, is using so much of the bus that adapters in higher slot numbers are not being served. This is particularly true if the adapters in a lower slot number are ATM adapters.

Max Collision Errors
Number of unsuccessful transmissions due to too many collisions. The number of collisions encountered exceeded the number of retries on the adapter.

http://www-01.ibm.com/support/knowledgecenter/ssw_aix_71/com.ibm.aix.performance/prftungd_pdf.pdf

69

netstat –v – Field meanings

Late Collision Errors
Number of unsuccessful transmissions due to the late collision error.

Timeout Errors
Number of unsuccessful transmissions due to adapter reported timeout errors.

Single Collision Count
Number of outgoing packets with single (only one) collision encountered during transmission.

Multiple Collision Count
Number of outgoing packets with multiple (2 - 15) collisions encountered during transmission.

Receive Collision Errors
Number of incoming packets with collision errors during reception.

No mbuf Errors
Number of times that mbufs were not available to the device driver. This usually occurs during receive operations when the driver must obtain memory buffers to process inbound packets. If the mbuf pool for the requested size is empty, the packet will be discarded. Use the netstat -m command to confirm this, and increase the parameter thewall.

http://www-01.ibm.com/support/knowledgecenter/ssw_aix_71/com.ibm.aix.performance/prftungd_pdf.pdf

70

Definitions – netstat -v

netstat -v is used to look at queues and other information. If Max packets on S/W transmit queue is >0 and is equal to current HW transmit queue length then the send queue size should be increased. If the No mbuf errors is large then the receive queue size needs to be increased.

```
# netstat -v | grep Queue
Max Packets on S/W Transmit Queue: 0
S/W Transmit Queue Overflow: 0
Current S/W+H/W Transmit Queue Length: 0
Current HW Transmit Queue Length: 0

# netstat -v | grep mbuf
No mbuf Errors: 0
```

71

Network Speed Conversion

	power of 2	bits		= 1
	2^10	1024	=	kilobyte
	2^20	1048576	=	megabyte
	2^30	1073741824	=	gigabyte
	2^40	1.09951E+12	=	terabyte
	2^50	1.1259E+15	=	petabyte
	2^60	1.15292E+18	=	exabyte
	2^70	1.18059E+21	=	zettabyte
	2^80	1.20893E+24	=	yottabyte
	2^90	1.23794E+27	=	lottabyte
To Convert:	See Tab			
bits or Bytes	B			
Kbits or KBytes	K			
Mbits or Mbytes	M			
Gbits or Gbytes	G			

Try converter at: http://www.speedguide.net/conversion.php

72

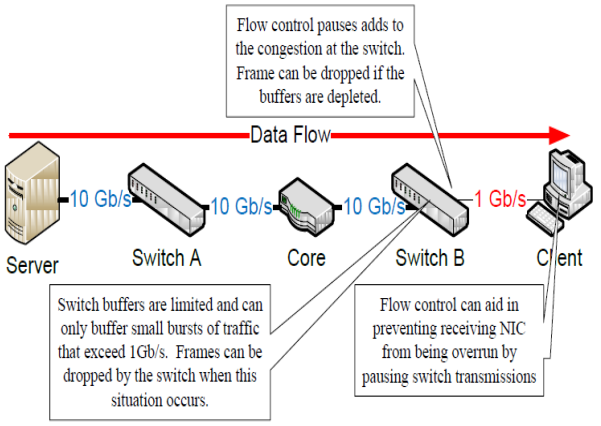
Network Speed Conversion

Converts Gigabits or Gigabytes								
1 Kilobyte =	1024	bytes	1 Megabyte =	1048576	bytes	1 gigabyte =	1073741824	bytes
Enter number Gbps:	bytes/sec (Bps)	bytes/min (Bpm)	Kbytes/sec (KBps)	Kbytes/min (KBpm)	Mbytes/sec (MBps)	Mbytes/min (MBpm)	Gbytes/sec (GBps)	Gbytes/min (GBpm)
1	134217728	8053063680	131072	7864320	128	7680	0.125	7.5
	bits/sec (bps)	bits/min (bpm)	Kbits/sec (Kbps)	Kbits/min (Kbpm)	Mbits/sec (Mbps)	Mbits/min (Mbpm)	Gbits/sec (Gbps)	Gbits/min (Gbpm)
	1073741824	64424509440	1048576	62914560	1024	61440	1	60
Enter number GBps:	bytes/sec (Bps)	bytes/min (Bpm)	Kbytes/sec (KBps)	Kbytes/min (KBpm)	Mbytes/sec (MBps)	Mbytes/min (MBpm)	Gbytes/sec (GBps)	Gbytes/min (GBpm)
0.125	134217728	8053063680	131072	7864320	128	7680	0.125	7.5
	bits/sec (bps)	bits/min (bpm)	Kbits/sec (Kbps)	Kbits/min (Kbpm)	Mbits/sec (Mbps)	Mbits/min (Mbpm)	Gbits/sec (Gbps)	Gbits/min (Gbpm)
	1073741824	64424509440	1048576	62914560	1024	61440	1	60

73

Speed Bottlenecks

Although flow control can prevent buffers from being depleted in one area, it may shift the congestion to the next device that is throttling the traffic in response to received pause frames.



From Nathan Flowers – Performance Issues with 10Gb Ethernet v1.0.0. 2012

74