AIX Performance Tuning I/O and Network

Jaqui Lynch
lynchj@forsythe.com

Handout will be at:
http://www.circle4.com/forsythe/aixperf-ionetwork.pdf

8/27/2015 1

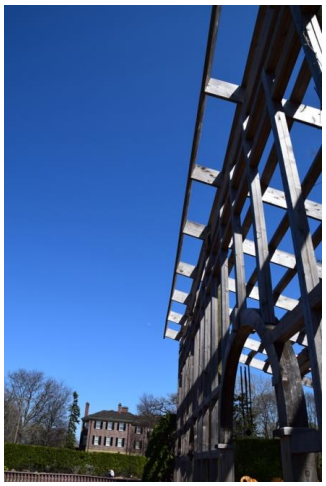
Agenda

- **I/O**
- **Volume Groups and File systems**
- **AIO and CIO**

- **Network**
- **nmon**

2

I/O



3

Rough Anatomy of an I/O

- LVM requests a PBUF
  - Pinned memory buffer to hold I/O request in LVM layer
- Then placed into an FSBUF
  - 3 types
  - These are also pinned
  - Filesystem — JFS
  - Client — NFS and VxFS
  - External Pager — JFS2
- If paging then need PSBUFs (also pinned)
  - Used for I/O requests to and from page space
- Then queue I/O to an hdisk (queue_depth)
- Then queue it to an adapter (num_cmd_elems)
- Adapter queues it to the disk subsystem
- Additionally, every 60 seconds the sync daemon (syncd) runs to flush dirty I/O out to filesystems or page space

4

From: AIX/VIOS Disk and Adapter IO Queue Tuning v1.2 – Dan Braden, July 2014

AIX IO Stack – Basic Tunables

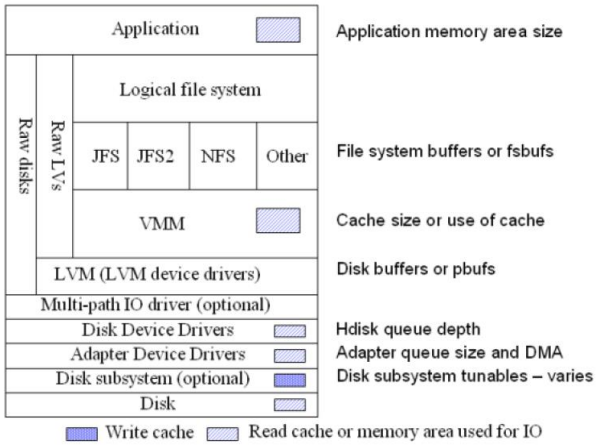
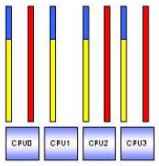


Figure 1 - AIX IO stack and basic tunables

5

IO Wait and why it is not necessarily useful



SMT2 example for simplicity

System has 7 threads with work, the 8th has nothing so is not shown
System has 3 threads blocked (red threads)
SMT is turned on
There are 4 threads ready to run so they get dispatched and each is using 80% user and 20% system

Metrics would show:
%user = .8 * 4 / 4 = 80%
%sys = .2 * 4 / 4 = 20%
Idle will be 0% as no core is waiting to run threads
IO Wait will be 0% as no core is idle waiting for IO to complete as something else got dispatched to that core
SO we have IO wait
BUT we don't see it
Also if all threads were blocked but nothing else to run then we would see IO wait that is very high

6

What is iowait? Lessons to learn

- iowait is a form of idle time
- It is simply the percentage of time the CPU is idle AND there is at least one I/O still in progress (started from that CPU)
- The iowait value seen in the output of commands like vmstat, iostat, and topas is the iowait percentages across all CPUs averaged together
  - This can be very misleading!
- High I/O wait does not mean that there is definitely an I/O bottleneck
- Zero I/O wait does not mean that there is not an I/O bottleneck
- A CPU in I/O wait state can still execute threads if there are any runnable threads

7

Basics

Data layout will have more impact than most tunables

Plan in advance

Large hdisks are evil

- I/O performance is about bandwidth and reduced queuing, not size
- 10 x 50gb or 5 x 100gb hdisk are better than 1 x 500gb
- Also larger LUN sizes may mean larger PP sizes which is not great for lots of little filesystems
- Need to separate different kinds of data i.e. logs versus data

The issue is queue_depth

- In process and wait queues for hdisks
- In process queue contains up to queue_depth I/Os
- hdisk driver submits I/Os to the adapter driver
- Adapter driver also has in process and wait queues
- SDD and some other multi-path drivers will not submit more than queue_depth IOs to an hdisk which can affect performance
- Adapter driver submits I/Os to disk subsystem
- Default client qdepth for vSCSI is 3
  - chdev –l hdisk? –a queue_depth=20 (or some good value)
- Default client qdepth for NPIV is set by the Multipath driver in the client

8

More on queue depth

- Disk and adapter drivers each have a queue to handle I/O
- Queues are split into in-service (aka in-flight) and wait queues
- IO requests in in-service queue are sent to storage and slot is freed when the IO is complete
- IO requests in the wait queue stay there till an in-service slot is free

- queue depth is the size of the in-service queue for the hdisk
  - Default for vSCSI hdisk is 3
  - Default for NPIV or direct attach depends on the HAK (host attach kit) or MPIO drivers used
- num_cmd_elems is the size of the in-service queue for the HBA

- Maximum in-flight IOs submitted to the SAN is the smallest of:
  - Sum of hdisk queue depths
  - Sum of the HBA num_cmd_elems
  - Maximum in-flight IOs submitted by the application
- For HBAs
  - num_cmd_elems defaults to 200 typically
  - Max range is 2048 to 4096 depending on storage vendor
  - As of AIX v7.1 tl2 (or 6.1 tl8) num_cmd_elems is limited to 256 for VFCs
    - See http://www-01.ibm.com/support/docview.wss?uid=isg1IV63282

9

Queue Depth

- Try sar –d, nmon –D, iostat -D
- sar –d 2 6 shows:

device	%busy	avque	r+w/s	Kbs/s	avwait	avserv
hdisk7	0	0.0	2	160	0.0	1.9
hdisk8	19	0.3	568	14337	23.5	2.3
hdisk9	2	0.0	31	149	0.0	0.9

- avque
  Average IOs in the wait queue
  Waiting to get sent to the disk (the disk's queue is full)
  Values > 0 indicate increasing queue_depth may help performance
  Used to mean number of IOs in the disk queue
- avwait
  Average time waiting in the wait queue (ms)
- avserv
  Average I/O service time when sent to disk (ms)
- See articles by Dan Braden:
  - http://www-03.ibm.com/support/techdocs/atsmastr.nsf/WebIndex/TD105745
  - http://www-03.ibm.com/support/techdocs/atsmastr.nsf/WebIndex/TD106122

10

iostat -Dl

	%tm act	bps	tps	bread	bwrtn	rps	avg serv	min serv	max serv	wps	avg serv	min serv	max serv	avg time	min time	max time	avg wqsz	avg sqsz	serv qfull
hdisk0	13.7	255.3K	33.5	682.7	254.6K	0.1	3	1.6	4	33.4	6.6	0.7	119.2	2.4	0	81.3	0	0	2.1
hdisk5	14.1	254.6K	33.4	0	254.6K	0	0	0	0	33.4	6.7	0.8	122.9	2.4	0	82.1	0	0	2.1
hdisk16	2.7	1.7M	3.9	1.7M	0	3.9	12.6	1.2	71.3	0	0	0	0	0	0	0	0	0	0
hdisk17	0.1	1.8K	0.3	1.8K	0	0.3	4.2	2.4	6.1	0	0	0	0	0	0	0	0	0	0
hdisk15	4.4	2.2M	4.9	2.2M	273.1	4.8	19.5	2.9	97.5	0.1	7.8	1.1	14.4	0	0	0	0	0	0
hdisk18	0.1	2.2K	0.5	2.2K	0	0.5	1.5	0.2	5.1	0	0	0	0	0	0	0	0	0	0
hdisk19	0.1	2.6K	0.6	2.6K	0	0.6	2.7	0.2	15.5	0	0	0	0	0	0	0	0	0	0
hdisk20	3.4	872.4K	2.4	872.4K	0	2.4	27.7	0.2	163.2	0	0	0	0	0	0	0	0	0	0
hdisk22	5	2.4M	29.8	2.4M	0	29.8	3.7	0.2	50.1	0	0	0	0	0	0	0.1	0	0	0
hdisk25	10.3	2.3M	12.2	2.3M	0	12.2	16.4	0.2	248.5	0	0	0	0	0	0	0	0	0	0
hdisk24	9.2	2.2M	5	2.2M	0	5	34.6	0.2	221.9	0	0	0	0	0	0	0	0	0	0
hdisk26	7.9	2.2M	4.5	2.2M	0	4.5	32	3.1	201	0	0	0	0	0	0	0	0	0	0
hdisk27	6.2	2.2M	4.4	2.2M	0	4.4	25.4	0.6	219.5	0	0	0	0	0	0	0.1	0	0	0
hdisk28	3	2.2M	4.5	2.2M	0	4.5	10.3	3	101.6	0	0	0	0	0	0	0	0	0	0
hdisk29	6.8	2.2M	4.5	2.2M	0	4.5	26.6	3.1	219.3	0	0	0	0	0	0	0	0	0	0
hdisk9	0.1	136.5	0	0	136.5	0	0	0	0	0	21.2	21.2	21.2	0	0	0	0	0	0

tps	Transactions per second – transfers per second to the adapter
avgserv	Average service time
Avgtime	Average time in the wait queue
avgwqsz	Average wait queue size If regularly >0 increase queue-depth
avgsqsz	Average service queue size (waiting to be sent to disk) Can't be larger than queue-depth for the disk
servqfull	Rate of IOs submitted to a full queue per second

Look at iostat –aD for adapter queues
If avgwqsz > 0 or sqfull high then increase queue_depth. Also look at avgsqsz.

Per IBM	Average IO sizes: read = bread/rps write = bwrtn/wps

Also try
iostat –RDTl int count
iostat –RDTl 30 5
Does 5 x 30 second snaps

11

Adapter Queue Problems

- Look at BBBF Tab in NMON Analyzer or run fcstat command
- fcstat –D provides better information including high water marks that can be used in calculations

- Adapter device drivers use DMA for IO
- From fcstat on each fcs
- NOTE these are since boot

FC SCSI Adapter Driver Information
No DMA Resource Count: 0
No Adapter Elements Count: 2567
No Command Resource Count: 34114051
Number of times since boot that IO was temporarily blocked waiting for resources such as num_cmd_elems too low

- No DMA resource – adjust max_xfer_size
- No adapter elements – adjust num_cmd_elems
- No command resource – adjust num_cmd_elems

- If using NPIV make changes to VIO and client, not just VIO
- Reboot VIO prior to changing client settings

12

Adapter Tuning

```
fcs0
bus_intr_lvl        115          Bus interrupt level                  False
bus_io_addr         0xdfc00      Bus I/O address                      False
bus_mem_addr        0xe8040000   Bus memory address                   False
init_link           al           INIT Link flags                      True
intr_priority       3            Interrupt priority                   False
lg_term_dma         0x800000     Long term DMA                        True
max_xfer_size       0x100000     Maximum Transfer Size                True      (16MB DMA)
num_cmd_elems       200          Maximum number of COMMANDS to queue to the adapter True
pref_alpa           0x1          Preferred AL_PA                      True
sw_fc_class         2            FC Class for Fabric                  True

Changes I often make (test first)
max_xfer_size       0x200000     Maximum Transfer Size    True    128MB DMA area for data I/O
num_cmd_elems       1024         Maximum number of COMMANDS to queue to the adapter True
```

Often I raise this to 2048 – check with your disk vendor
lg_term_dma is the DMA area for control I/O

Check these are ok with your disk vendor!!!

```
chdev -l fcs0 -a max_xfer_size=0x200000 -a num_cmd_elems=1024 -P
chdev -l fcs1 -a max_xfer_size=0x200000 -a num_cmd_elems=1024 -P
```

At AIX 6.1 TL2 VFCs will always use a 128MB DMA memory area even with default max_xfer_size – I change it anyway for consistency
As of AIX v7.1 tl2 (or 6.1 tl8) num_cmd_elems there is an effective limit of 256 for VFCs
See http://www-01.ibm.com/support/docview.wss?uid=isg1IV63282

Remember make changes too both VIO servers and client LPARs if using NPIV
VIO server setting must be at least as large as the client setting

See Dan Braden Techdoc for more on tuning these:
http://www-03.ibm.com/support/techdocs/atsmastr.nsf/WebIndex/TD105745

13

fcstat –D - Output

```
lsattr -El fcs8
lg_term_dma   0x800000  Long term DMA                    True
max_xfer_size 0x200000  Maximum Transfer Size            True
num_cmd_elems 2048      Maximum number of COMMANDS to queue to the adapter True

fcstat -D fcs8
FIBRE CHANNEL STATISTICS REPORT: fcs8
........

FC SCSI Adapter Driver Queue Statistics
High water mark of active commands:  512
High water mark of pending commands:  104

FC SCSI Adapter Driver Information
 No DMA Resource Count: 0
 No Adapter Elements Count: 13300
 No Command Resource Count: 0
```

Adapter Effective max transfer value: 0x200000
The above tells you the max_xfer_size that is being used

Some lines removed to save space
Per Dan Braden:
Set num_cmd_elems to at least high active + high pending or 512+104=626

There is also an fcstat –e version as well - fcstat –e fcs0

14

My VIO Server and NPIV Client Adapter Settings

VIO SERVER

```
#lsattr -El fcs0
lg_term_dma       0x800000   Long term DMA                      True
max_xfer_size     0x200000   Maximum Transfer Size              True
num_cmd_elems     2048       Maximum number of COMMAND Elements True
```

NPIV Client (running at defaults before changes)

```
#lsattr -El fcs0
lg_term_dma       0x800000   Long term DMA                  True
max_xfer_size     0x200000   Maximum Transfer Size          True
num_cmd_elems     256        Maximum Number of COMMAND Elements True
```

NOTE NPIV client must be <= to settings on VIO
VFCs can't exceed 256 after 7.1 tl2 or 6.1 tl8

15

Tunables

16

vmstat –v Output – Not Healthy

3.0 minperm percentage
90.0 maxperm percentage
45.1 numperm percentage
45.1 numclient percentage
90.0 maxclient percentage

1468217 pending disk I/Os blocked with no pbuf	pbufs (LVM)
11173706 paging space I/Os blocked with no psbuf	pagespace (VMM)
2048 file system I/Os blocked with no fsbuf	JFS (FS layer)
238 client file system I/Os blocked with no fsbuf	NFS/VxFS (FS layer)
39943187 external pager file system I/Os blocked with no fsbuf	JFS2 (FS layer)

numclient=numperm so most likely the I/O being done is JFS2 or NFS or VxFS
Based on the blocked I/Os it is clearly a system using JFS2
It is also having paging problems
pbufs also need reviewing

17

lvmo –a Output

2725270 pending disk I/Os blocked with no pbuf
Sometimes the above line from vmstat –v only includes rootvg so use lvmo –a to double-check

vgname = rootvg
pv_pbuf_count = 512
total_vg_pbufs = 1024
max_vg_pbuf_count = 16384
pervg_blocked_io_count = 0 this is rootvg
pv_min_pbuf = 512
Max_vg_pbuf_count = 0
global_blocked_io_count = 2725270 this is the others

Use lvmo –v xxxxvg -a
For other VGs we see the following in pervg_blocked_io_count

	blocked	total_vg_bufs
nimvg	29	512
sasvg	2719199	1024
backupvg	6042	4608

lvmo –v sasvg –o pv_pbuf_count=2048 - **do this for each VG affected NOT GLOBALLY**

18

Parameter Settings - Summary

PARAMETER	DEFAULTS AIXv5.3	AIXv6	AIXv7	NEW SET ALL TO	
NETWORK (no)					
rfc1323	0	0	0	1	
tcp_sendspace	16384	16384	16384	262144 (1Gb)	
tcp_recvspace	16384	16384	16384	262144 (1Gb)	
udp_sendspace	9216	9216	9216	65536	
udp_recvspace	42080	42080	42080	655360	
MEMORY (vmo)					
minperm%	20	3	3	3	
maxperm%	80	90	90	90	JFS, NFS, VxFS, JFS2
maxclient%	80	90	90	90	JFS2, NFS
lru_file_repage	1	0	0	0	
lru_poll_interval	?	10	10	10	
Minfree	960	960	960	calculation	
Maxfree	1088	1088	1088	calculation	
page_steal_method	0	0 /1 (TL)	1	1	
JFS2 (ioo)					
j2_maxPageReadAhead	128	128	128	as needed – affects maxfree setting	
j2_dynamicBufferPreallocation	16	16	16	as needed – max is 256	

19

Other Interesting Tunables

- These are set as options in /etc/filesystems for the filesystem
- noatime
  - Why write a record every time you read or touch a file?
  - mount command option
  - Use for redo and archive logs
- Release behind (or throw data out of file system cache)
  - rbr – release behind on read
  - rbw – release behind on write
  - rbrw – both
- Use chfs to make the changes above
  - chfs -a options=rbrw,noatime /filesystemname
  - Needs to be remounted
- LOG=NULL
- Read the various AIX Difference Guides:
  - http://www.redbooks.ibm.com/cgi-bin/searchsite.cgi?query=aix+AND+differences+AND+guide
- When making changes to /etc/filesystems use chfs to make them stick

20

filemon

Uses trace so don't forget to STOP the trace

Can provide the following information
- CPU Utilization during the trace
- Most active Files
- Most active Segments
- Most active Logical Volumes
- Most active Physical Volumes
- Most active Files Process-Wise
- Most active Files Thread-Wise

Sample script to run it:

```
filemon -v -o abc.filemon.txt  -O all -T 210000000
sleep 60
Trcstop
```

OR

```
filemon -v -o abc.filemon2.txt  -O pv,lv -T 210000000
sleep 60
trcstop
```

21

filemon –v –o pv,lv

```
Most Active Logical Volumes
------------------------------------------------------------------------
  util  #rblk          #wblk      KB/s        volume
description
------------------------------------------------------------------------
  0.66 4647264        834573    45668.9  /dev/gandalfp_ga71_lv     /ga71
  0.36       960      834565     6960.7  /dev/gandalfp_ga73_lv     /ga73
  0.13 2430816         13448    20363.1  /dev/misc_gm10_lv         /gm10
  0.11   53808         14800      571.6  /dev/gandalfp_ga15_lv     /ga15
  0.08   94416          7616      850.0  /dev/gandalfp_ga10_lv     /ga10
  0.07  787632          6296     6614.2  /dev/misc_gm15_lv         /gm15
  0.05    8256         24259      270.9  /dev/misc_gm73_lv         /gm73
  0.05   15936         67568      695.7  /dev/gandalfp_ga20_lv     /ga20
  0.05    8256         25521      281.4  /dev/misc_gm72_lv         /gm72
  0.04   58176         22088      668.7  /dev/misc_gm71_lv         /gm71
```

22

filemon –v –o pv,lv

Most Active Physical Volumes

util	#rblk	#wblk	KB/s	volume	description
0.38	4538432	46126	38193.7	/dev/hdisk20	MPIO FC 2145
0.27	12224	671683	5697.6	/dev/hdisk21	MPIO FC 2145
0.19	15696	1099234	9288.4	/dev/hdisk22	MPIO FC 2145
0.08	608	374402	3124.2	/dev/hdisk97	MPIO FC 2145
0.08	304	369260	3078.8	/dev/hdisk99	MPIO FC 2145
0.06	537136	22927	4665.9	/dev/hdisk12	MPIO FC 2145
0.06	6912	631857	5321.6	/dev/hdisk102	MPIO FC 2145

23

Asynchronous I/O and Concurrent I/O

24

Async I/O - v5.3

Total number of AIOs in use
pstat ï a | grep aios | wc ï l
Maximum AIOservers started since boot
of servers per cpu True
NB ï maxservers is a per processor setting in AIX 5.3

Or new way for Posix AIOs is:
ps ï k | grep aio | wc -l
4205

At AIX v5.3 tl05 this is controlled by aioo command
Also iostat ï A
THIS ALL CHANGES IN AIX V6 ï SETTINGS WILL BE UNDER IOO THERE
lsattr -El aio0
autoconfig defined STATE to be configured at system restart True
fastpath enable State of fast path True
kprocprio 39 Server PRIORITY True
maxreqs 4096 Maximum number of REQUESTS True
maxservers 10 MAXIMUM number of servers per cpu True
minservers 1 MINIMUM number of servers True

AIO maxservers
lsattr ï El aio0 ï a maxservers
maxservers 320 MAXIMUM number

AIO is used to improve performance for I/O to raw LVs as well as filesystems.

25

Async I/O – AIX v6 and v7

No more smit panels and no AIO servers start at boot
Kernel extensions loaded at boot
AIO servers go away if no activity for 300 seconds
Only need to tune maxreqs normally

ioo -a –F | more
aio_active = 0
aio_maxreqs = 65536
aio_maxservers = 30
aio_minservers = 3
aio_server_inactivity = 300
posix_aio_active = 0
posix_aio_maxreqs = 65536
posix_aio_maxservers = 30
posix_aio_minservers = 3
posix_aio_server_inactivity = 300

pstat -a | grep aio
22 a 1608e 1 1608e 0 0 1 aioPpool
24 a 1804a 1 1804a 0 0 1 aioLpool

You may see some aioservers on a busy system

##Restricted tunables
aio_fastpath = 1
aio_fsfastpath = 1
aio_kprocprio = 39
aio_multitidsusp = 1
aio_sample_rate = 5
aio_samples_per_cycle = 6
posix_aio_fastpath = 1
posix_aio_fsfastpath = 1
posix_aio_kprocprio = 39
posix_aio_sample_rate = 5
posix_aio_samples_per_cycle = 6

26

AIO Recommendations

Oracle now recommending the following as starting points

	5.3	**6.1 or 7 (non CIO)**
minservers	100	3 - default
maxservers	200	200
maxreqs	16384	65536 ï default

These are per LCPU
So for lcpu=10 and maxservers=100 you get 1000 aioservers

AIO applies to both raw I/O and file systems

Grow maxservers as you need to

27

iostat -A

iostat -A async IO

System configuration: lcpu=16 drives=15
aio: avgc avfc maxg maif maxr avg-cpu: % user % sys % idle % iowait

150 0 5652 0 12288 21.4 3.3 64.7 10.6

Disks:	% tm_act	Kbps	tps	Kb_read	Kb_wrtn
hdisk6	23.4	1846.1	195.2	381485298	61892856
hdisk5	15.2	1387.4	143.8	304880506	28324064
hdisk9	13.9	1695.9	163.3	373163558	34144512

If maxg close to maxr or maxservers then increase maxreqs or maxservers

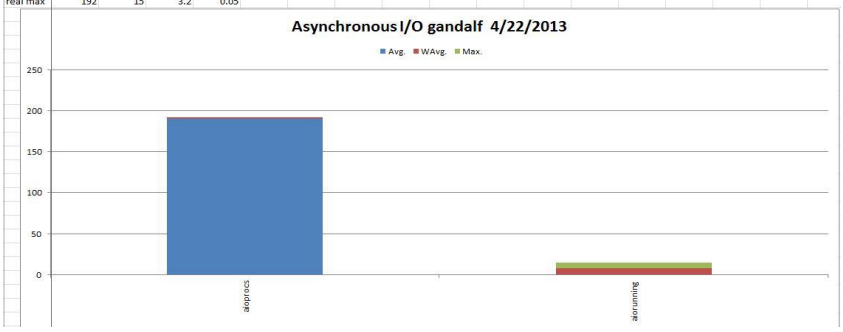
Old calculation – no longer recommended
minservers = active number of CPUs or 10 whichever is the smaller number
maxservers = number of disks *times 10 divided by the active number of CPUs*
maxreqs = 4 *times the number of disks times the queue depth*

***Reboot anytime the AIO Server parameters are changed

28

PROCAIO tab in nmon

	aioprocs	aiorunnin	aiocpu	syscpu
Avg.	190.4	0.3	0.1	0.0
WAvg.	1.6	7.6	1.6	0.0
Max.	0.0	7.1	1.5	0.0
real max	192	15	3.2	0.05



Maximum seen was 192 but average was much less

29

DIO and CIO

- **DIO**
  - Direct I/O
  - Around since AIX v5.1, also in Linux
  - Used with JFS
  - CIO is built on it
  - Effectively bypasses filesystem caching to bring data directly into application buffers
  - Does not like compressed JFS or BF (lfe) filesystems
    - Performance will suffer due to requirement for 128kb I/O (after 4MB)
  - Reduces CPU and eliminates overhead copying data twice
  - Reads are asynchronous
  - No filesystem readahead
  - No lrud or syncd overhead
  - No double buffering of data
  - Inode locks still used
  - Benefits heavily random access workloads

30

DIO and CIO

- **CIO**
  - Concurrent I/O ï AIX only, not in Linux
  - Only available in JFS2
  - Allows performance close to raw devices
  - **Designed for apps (such as RDBs) that enforce write serialization at the app**
  - Allows non-use of inode locks
  - Implies DIO as well
  - Benefits heavy update workloads
  - Speeds up writes significantly
  - Saves memory and CPU for double copies
  - No filesystem readahead
  - No lrud or syncd overhead
  - No double buffering of data
  - **Not all apps benefit from CIO and DIO – some are better with filesystem caching and some are safer that way**
- When to use it
  - Database DBF files, redo logs and control files and flashback log files.
  - Not for Oracle binaries or archive log files
- Can get stats using vmstat ï IW flags

31

DIO/CIO Oracle Specifics

- Use CIO where it will benefit you
  - Do not use for Oracle binaries
  - Ensure redo logs and control files are in their own filesystems with the correct (512) blocksize
    - Use lsfs ï q to check blocksizes
  - I give each instance its own filesystem and their redo logs are also separate
- Leave DISK_ASYNCH_IO=TRUE in Oracle
- Tweak the maxservers AIO settings
- Remember CIO uses DIO under the covers
- If using JFS
  - Do not allocate JFS with BF (LFE)
  - It increases DIO transfer size from 4k to 128k
  - 2gb is largest file size
  - Do not use compressed JFS ï defeats DIO

32

lsfs -q output

/dev/ga7_ga74_lv -- /ga74 jfs2 264241152 rw yes no
(lv size: 264241152, fs size: 264241152, block size: 4096, sparse files: yes, inline log: no, inline log size: 0, EAformat: v1, Quota: no, DMAPI: no, VIX: no, EFS: no, ISNAPSHOT: no, MAXEXT: 0, MountGuard: no)

/dev/ga7_ga71_lv -- /ga71 jfs2 68157440 rw yes no
(lv size: 68157440, fs size: 68157440, **block size: 512**, sparse files: yes, inline log: no, inline log size: 0, EAformat: v1, Quota: no, DMAPI: no, VIX: no, EFS: no, ISNAPSHOT: no, MAXEXT: 0, MountGuard: no)

It really helps if you give LVs meaningful names like /dev/lv_prodredo rather than /dev/u99

33

Telling Oracle to use CIO and AIO

If your Oracle version (10g/11g) supports it then configure it this way:
There is no default set in Oracle 10g do you need to set it
Configure Oracle Instance to use CIO and AIO in the init.ora (PFILE/SPFILE)

disk_async_io	=	true	(init.ora)
filesystemio_options	=	setall	(init.ora)

Note if you do backups using system commands while the database is up then you will need to use the 9i method below for v10 or v11

If not (i.e. 9i) then you will have to set the filesystem to use CIO in the /etc filesystems

options	=	cio	(/etc/filesystems)
disk_async_io	=	true	(init.ora)

Do not put anything in the filesystem that the Database does not manage
Remember there is no inode lock on writes

Or you can use ASM and let it manage all the disk automatically
Also read Metalink Notes #257338.1, #360287.1
See Metalink Note 960055.1 for recommendations

Do not set it in both places (config file and /etc/filesystems)

34

Demoted I/O in Oracle

- Check w column in vmstat -IW
- CIO write fails because IO is not aligned to FS blocksize
  - i.e app writing 512 byte blocks but FS has 4096
- Ends up getting redone
  - Demoted I/O consumes more kernel CPU
  - And more physical I/O
- To find demoted I/O (if JFS2)

trace –aj 59B,59C ; sleep 2 ; trcstop ; trcrpt –o directio.trcrpt

grep –i demoted directio.trcrpt

Look in the report for:

```
JFS2 IO dio demoted:
```

35

NETWORK

36

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*

37

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.

38

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.

39

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

40

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.

41

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.

42

Starter set of tunables 3

Typically we set the following for AIX v5.3 tl04 and higher:

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

43

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 0xffffe00 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

44

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
```

45

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

46

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)

47

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

48

10Gbe Ethernet Adapters

49

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

50

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

51

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

52

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**

53

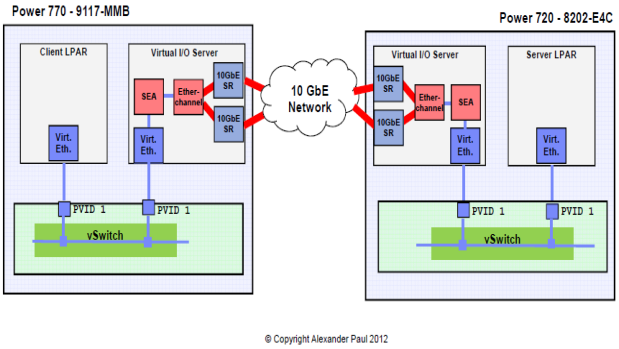
10GbE Tips

- 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
  - 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

54

10GbE Performance

Å 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 pPE27 by Alexander Paul on Network Performance Optimization for virtualized IBM POWER Systems

55

Testing 10GbE Performance

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

56

Looking at Performance

57

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%

58

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.

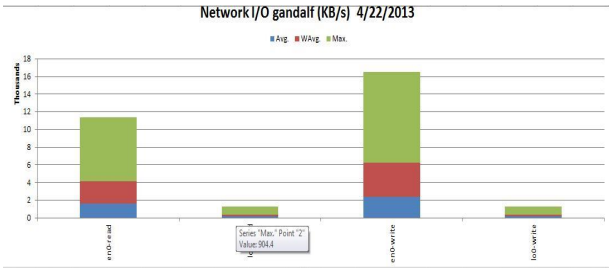
The Collis field shows the number of collisions since boot and can be as high as 10%. If it is greater then it is necessary to reorganize the network as the network is obviously overloaded on that segment.

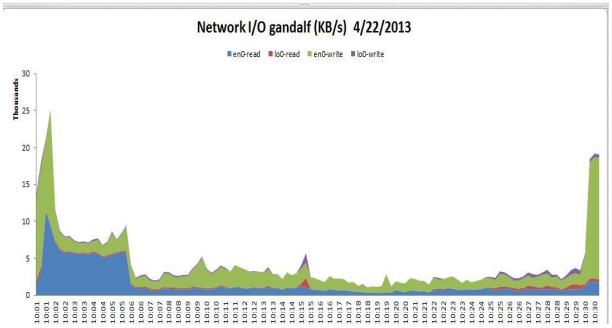
```
# netstat -i
Name  Mtu   Network    Address                    Ipkts      Ierrs      Opkts      Oerrs      Coll
en6   1500  10.250.134 b740vio1                   4510939    0          535626     0          0

# 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
```

59

Net tab in nmon analyser





60

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

61

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
```

62

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

63

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

64

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**

65

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

66

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

67

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

68

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

69

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
```

70

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

71

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

72

Detecting UDP Packet losses

ÅRun netstat –s or netstat –p udp
ÅLook under the ip: section for fragments dropped (dup or out of space)
ÅIncrease udp_sendspace

ip:
8937989 total packets received
……
0 fragments dropped (dup or out of space)

73

Detecting UDP Packet losses

Å Look under the udp: section for socket buffer overflows
Å These mean you need to increase udp_recvspace
Å 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.
Å If a UDP packet arrives for a socket with a full buffer then it is discarded by the kernel
Å 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

74

Tips to keep out of trouble

- Monitor errpt
- Check the performance apars have all been installed
  - Yes this means you need to stay current
  - See Stephen Nasypany and Rosa Davidson Optimization Presentations
- Keep firmware up to date
  - In particular, look at the firmware history for your server to see if there are performance problems fixed
- Information on the firmware updates can be found at:
  - http://www-933.ibm.com/support/fixcentral/
- Firmware history including release dates can be found at:
  - Power7 Midrange
    - http://download.boulder.ibm.com/ibmdl/pub/software/server/firmware/AM-Firmware-Hist.html
  - Power7 High end
    - http://download.boulder.ibm.com/ibmdl/pub/software/server/firmware/AL-Firmware-Hist.html
  - Ensure software stack is current
  - Ensure compilers are current and that compiled code turns on optimization
  - To get true MPIO run the correct multipath software
  - Ensure system is properly architected (VPs, memory, entitlement, etc)
  - Take a baseline before and after any changes
- DOCUMENTATION

75

nmon

76

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

77

nmon Monitoring

Å nmon -ft –AOPV^dMLW -s 15 -c 120

- Å Grabs a 30 minute nmon snapshot
- Å A is async IO
- Å M is mempages
- Å t is top processes
- Å L is large pages
- Å **O is SEA on the VIO**
- Å P is paging space
- Å V is disk volume group
- Å d is disk service times
- Å ^ is fibre adapter stats
- Å 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

78

Thank you for your time

If you have questions please email me at:
lynchj@forsythe.com

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

Handout will be at:
http://www.circle4.com/forsythe/aixperf-ionetwork.pdf

79

Useful Links

- Charlie Cler Articles
  - http://www.ibmsystemsmag.com/authors/Charlie-Cler/
- 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
- IBM US Virtual User Group
    - http://www.tinyurl.com/ibmaixvug
- Power Systems UK User Group
    - http://tinyurl.com/PowerSystemsTechnicalWebinars

80

Useful Links

- AIX Wiki
  - https://www.ibm.com/developerworks/wikis/display/WikiPtype/AIX
- HMC Scanner
  - http://www.ibm.com/developerworks/wikis/display/WikiPtype/HMC+Scanner
- Workload Estimator
  - http://ibm.com/systems/support/tools/estimator
- Performance Tools Wiki
  - http://www.ibm.com/developerworks/wikis/display/WikiPtype/Performance+Monitoring+Tools
- Performance Monitoring
  - https://www.ibm.com/developerworks/wikis/display/WikiPtype/Performance+Monitoring+Documentation
- Other Performance Tools
  - https://www.ibm.com/developerworks/wikis/display/WikiPtype/Other+Performance+Tools
  - Includes new advisors for Java, VIOS, Virtualization
- VIOS Advisor
  - https://www.ibm.com/developerworks/wikis/display/WikiPtype/Other+Performance+Tools#OtherPerformanceTools-VIOSPA

81

References

- Simultaneous Multi-Threading on POWER7 Processors by Mark Funk
  - http://www.ibm.com/systems/resources/pwrsysperf_SMT4OnP7.pdf
- 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

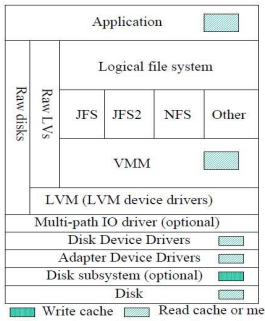
82

I/O Backup Slides

83

From: PE23 Disk I/O Tuning in AIX v6.1 – Dan Braden and Steven Nasypany, October 2010

The AIX IO stack



Application memory area caches data to avoid IO

NFS caches file attributes
NFS has a cached filesystem for NFS clients

JFS and JFS2 cache use extra system RAM
JFS uses persistent pages for cache
JFS2 uses client pages for cache

Queues exist for both adapters and disks
Adapter device drivers use DMA for IO
Disk subsystems have read and write cache
Disks have memory to store commands/data

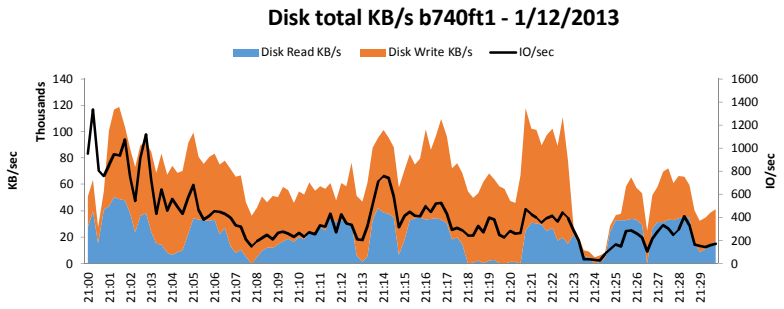
Write cache Read cache or memory area used for IO

IOs can be coalesced (good) or split up (bad) as they go thru the IO stack
IOs adjacent in a file/LV/disk can be coalesced
IOs greater than the maximum IO size supported will be split up

IBM

84

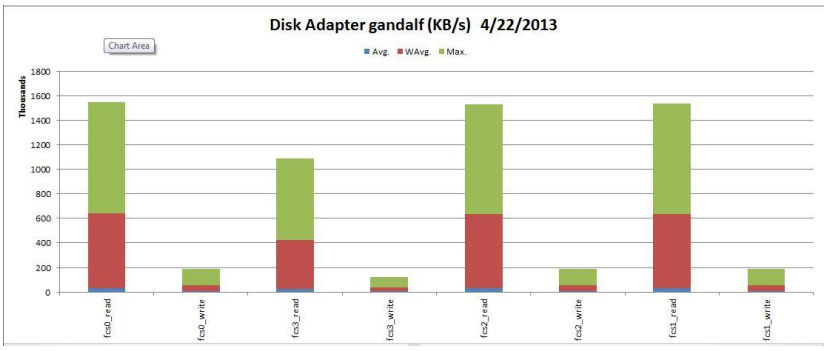
disk_summ tab in nmon



	Disk Read KB/s	Disk Write KB/s	IO/sec	Read+Write	R/W MB/Sec
Avg.	21695.8	43912.8	393.4	65608.6	64.1
Real Max	50481.1	92739.4	1340.8	118896.4	116.1

85

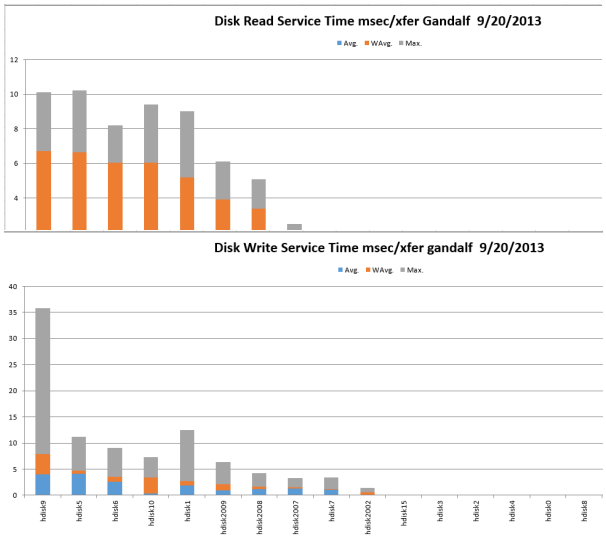
IOadapt tab in nmon



Are we balanced?

86

nmon Disk Service Times



87

Adapters

88

Adapter Priorities affect Performance

Power 770 Layout			9117-MMC											
CEC		Top	123456 has GX cables			Bottom	2468ab				5877 pcie only I/O Drawer 123487			
	Slot	Desc	Pri	Alloc		Slot	Desc	Pri	Alloc		Slot	Desc	Pri	Alloc	IOC
	C1	8GB DP fibre	1	lpar1		C1	8GB DP fibre	1	lpar1		C1	8GB DP fibre	1	vio1	1
	C2	4PT 10/100/1000	3	lpar1		C2	4PT 10/100/1000	3	lpar1		C2	4PT 10/100/1000	3		1
	C3	8GB DP fibre	5	vio2		C3	8GB DP fibre	5	vio1		C3		5		1
	C4	4PT 10/100/1000	6	vio2		C4	4PT 10/100/1000	6	vio1		C4	8GB DP fibre	2	vio2	2
	C5	8GB DP fibre	2	vio1		C5	8GB DP fibre	2	vio2		C5	4PT 10/100/1000	4		2
	C6	4PT 10/100/1000	4	vio1		C6	4PT 10/100/1000	4	vio2		C6	4GB DP fibre	6	lpar1	2
											C7	4GB DP fibre	7		3
	D1	146GB disk		vio1		D1	146GB disk		vio1		C8		8		3
	D4	146GB disk		vio2		D4	146GB disk		vio2		C9		9		3
											C10		10		3

Check the various Technical Overview Redbooks at http://www.redbooks.ibm.com/

89

Power8 – S814 and S824 Adapter Slot Priority

S814 S824 Adapter Slots

			S814 / S824 (1 socket populated)			S824 (2 sockets populated)		
ID	Slot	Type	Feature	Description	Use	Feature	Description	Use
P1-C2	1	PCIe3 x8	Not available with 1-socket populated					
P1-C3	2	PCIe3 x16				EN0H	4-port FCoE (2x 10GbE + 2x 1Gb	VIO-1
P1-C4	3	PCIe3 x8						
P1-C5	4	PCIe3 x16				EN0H	4-port FCoE (2x 10GbE + 2x 1Gb	VIO-1
P1-C6	5	PCIe3 x16	EN0A	2-port 16Gb FC	VIO-1	EN0A	2-port 16Gb FC	VIO-1
P1-C7	6	PCIe3 x16	EN0A	2-port 16Gb FC	VIO-1	EN0A	2-port 16Gb FC	VIO-1
P1-C8	7	PCIe3 x8	EN0A	2-port 16Gb FC	VIO-2	EN0A	2-port 16Gb FC	VIO-2
P1-C9	8	PCIe3 x8	EN0A	2-port 16Gb FC	VIO-2	EN0A	2-port 16Gb FC	VIO-2
P1-C10	9	PCIe3 x8	EN0W	4-port 1GbE (required)		5899	4-port 1GbE (required)	
P1-C11	10	PCIe3 x8	EN0H	4-port FCoE (2x 10GbE + 2x 1Gb	VIO-1	EN0H	4-port FCoE (2x 10GbE + 2x 1Gb	VIO-2
P1-C12	11	PCIe3 x8	EN0H	4-port FCoE (2x 10GbE + 2x 1Gb	VIO-2	EN0H	4-port FCoE (2x 10GbE + 2x 1Gb	VIO-2
			Available Slot Priority: 6, 5, 7, 8, 10, 11			Available Slot Priority: 6, 5, 4, 2, 1, 3, 7, 8, 10, 11		

90

I/O Bandwidth – understand adapter differences

- PCIe2 LP 8Gb 4 port Fibre HBA
  - Data throughput 3200 MB/ps FDX per port
  - IOPS 200,000 per port
  - http://www.redbooks.ibm.com/technotes/tips0883.pdf
  - Can run at 2Gb, 4Gb or 8Gb
- PCIe2 8Gb 1 or 2 port Fibre HBA
  - Data throughput 1600 MB/s FDX per port
  - IOPS Up to 142,000 per card

Above are approximate taken from card specifications
Look at DISK_SUMM tab in nmon analyzer
Sum reads and writes, figure out the average and max
Then divide by 1024 to get MB/s

91

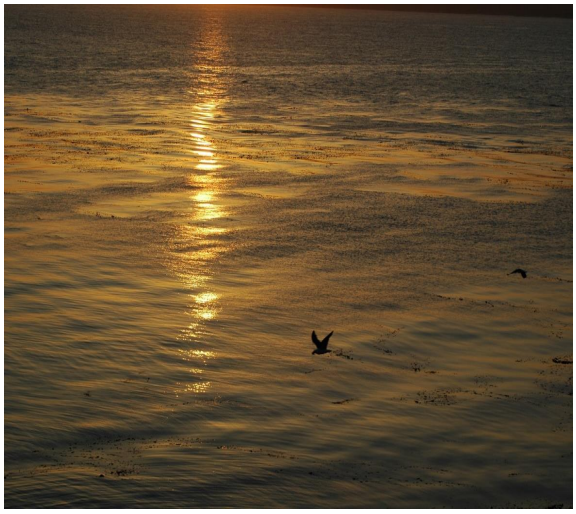
Adapter bandwidth

Adapter Performance Chart

Adapter	FC	IOPS 4K	Sustained Sequential b/w
2 Gbps FC adapter (single port)	5716	38,461	198 MB/s simplex, 385 MB/s duplex
4 Gbps FC adapter (single port)	5758	n/a	DDR slots: 400 MB/s simplex, ~750 MB/s duplex, SDR slots: 400 MB/s simplex, 500 MB/s duplex
4 Gbps FC adapter (dual)	5759	n/a	DDR slots: ~750 MB/s, SDR slots: ~500 MB/s
4 Gbps FC adapter PCI-e	5773	n/a	400 MB/s simplex, ~750 MB/s duplex
4 Gbps FC adapter (dual) PCI-e	5774	n/a	~750 MB/s
8 Gbps FC dual port PCI-e	5735	142,000	750 MB/s per port simplex, 997 MB/s duplex per port 1475 MB/s simplex per adapter, 2000 MB/s duplex per
10 Gb FCoE PCIe Dual Port	5708	150,000	930 MB/s per port simplex, 1900 MB/s per port duplex 1630 MB/s simplex per adapter, 2290 MB/s duplex per adapter

92

Network
Backup Slides



93

NETWORK TUNING in AIX

See article at:
http://www.ibmsystemsmag.com/aix/administrator/networks/network_tuning/
Replay at:
http://www.circle4.com/movies/

94

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.

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

95

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

96

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

97

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.

98

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

99

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

100

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)

101

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)

102

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
```

103

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 the**M_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.

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

104

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

105

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

106

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

107

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
```

108

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

109

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

110