

A016261

AIX Performance Tuning
Part 2 – I/O

Jaqui Lynch
jlynch@flagshipsg.com

IBM Power Systems
and IBM Storage
Technical University

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
- Flash Cache



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

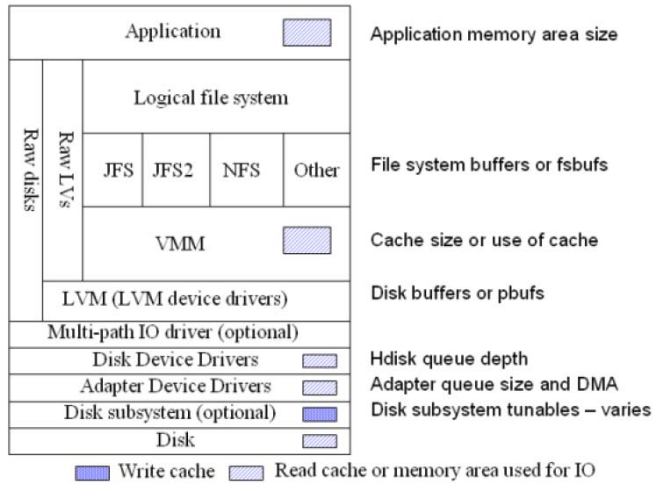


Figure 1 - AIX IO stack and basic tunables

5



IO Wait and why it is not necessarily useful

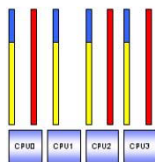
SMT2 example for simplicity

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=iss1IV63282>

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 -DI

	%tm	bps	tps	bread	bwrtn	rps	avg	min	max	wps	avg	min	max	avg	min	max	avg	avg	serv
	act					serv	serv	serv	serv	serv	serv	serv	serv	time	time	time	wqsz	sqsz	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

avgsserv 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

Rate of IOs submitted to a full queue per second

servqfull 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 -RDTI int count

iostat -RDTI 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         3            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=sg1iv63282>

This limitation got lifted for NPIV and the maximum is now 2048 provided you are at 6.1 tl9 (IV76258), 7.1 tl3 (IV76968) or 7.1 tl4 (IV76270).

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

Some lines removed to save space

Per Dan Braden:

Set num_cmd_elems to at least high active + high pending or 512+104=626

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 COMMANDS to queue to the adapter
 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 – see slide 13 for exceptions

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 xxxvg -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			NEW SET ALL TO
	AIXv5.3	AIXv6	AIXv7	
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 (loa)				
j2_maxPageReadAhead	128	128	128	as needed
j2_dynamicBufferPreallocation	16	16	16	as needed

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
- 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	#rbk	#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	8193.7	/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

2
3

23



Asynchronous I/O and Concurrent I/O



24



Async I/O - v5.3

Total number of AIOs in use

pstat -a | grep aio | wc -l
Maximum AIO servers started since boot
servers per cpu True
NB – maxservers is a per processor setting in AIX 5.3

AIO maxservers

lsattr -El aio0 -a maxservers
maxservers 320 MAXIMUM number of

Or new way for Posix AIOs is:

ps -k | grep aio | wc -l
4205

At AIX v5.3 tl05 this is controlled by aioo command

Also iostat -A

THIS ALL CHANGES IN AIX V6 – 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 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

##Restricted tunables

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

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

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

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 – 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

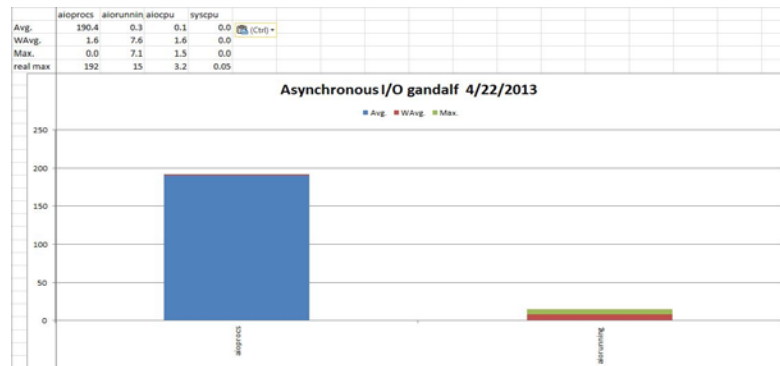
```
minservers = active number of CPUs or 10 whichever is the smaller number
maxservers = number of disks times 10 divided by the active number of CPUs
maxreqs    = 4 times the number of disks times the queue depth
```

***Reboot anytime the AIO Server parameters are changed

28



PROCAIO tab in nmon



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 lru 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 – 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 lru or sync 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 –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 -q to check blocksize
 - I give each instance its own filesystem and their redo logs are also separate
- Leave DISK_ASYNC_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 – 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, DMAP1: 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, DMAP1: no, VIX: no, EFS: no, ISNAPSHOT: no,
MAXEXT: 0, MountGuard: no)
```

It really helps if you give LVs meaningful names like /dev/lv_prodrredo 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 or in General

- 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:
1553 55 dio demoted:
```

35



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
- Note – you can't mix 512 and 4k disks in a VG

36



Flash Cache

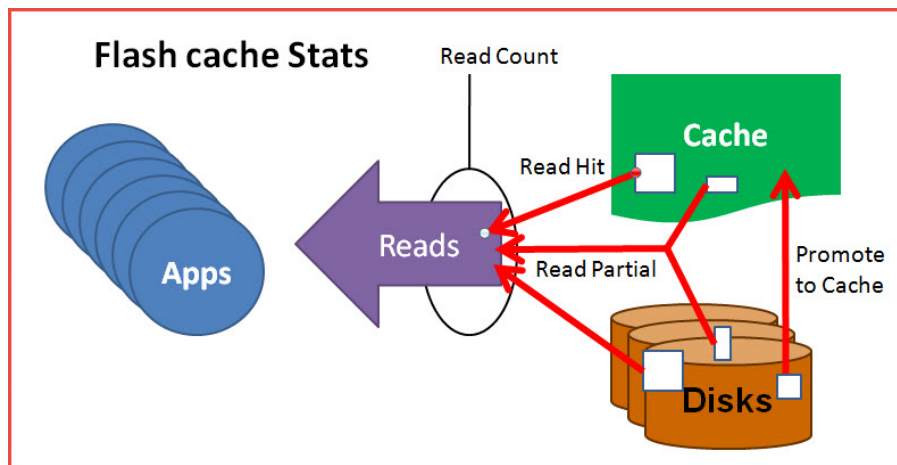
- Read only cache using SSDs. Reads will be processed from SSDs. Writes go direct to original storage device.
- Current limitations
 - Only 1 cache pool and 1 cache partition
- It takes time for the cache to warm up
- Blog
 - Nigel Griffiths Blog
 - <http://tinyurl.com/k7g5dr7>
 - Manoj Kumar Article
 - <http://tinyurl.com/mee4n3f>
- Prereqs
 - Server attached SSDs, flash that is attached using SAS or from the SAN
 - AIX 7.1 tl4 sp2 or 7.2 tl0 sp0 minimum
 - **Minimum** 4GB memory extra for every LPAR that has cache enabled
 - Cache devices are owned either by an LPAR or a VIO server
 - https://www.ibm.com/support/knowledgecenter/ssw_aix_72/com.ibm.aix.osdevice/caching_limitations.htm
 - Do not use this if your SAN disk is front-ended by flash already
 - Ensure following filesets are installed
 - `lspp -l | grep Cache` (on a 7.2 system)
 - `bos.pfcd.rte` 7.2.1.0 COMMITTED Power Flash Cache
 - `cache.mgt.rte` 7.2.1.0 COMMITTED AIX SSD Cache Device
 - `bos.pfcd.rte` 7.2.1.0 COMMITTED Power Flash Cache
 - `cache.mgt.rte` 7.2.1.0 COMMITTED AIX SSD Cache Device

10/15/2017

37



Flash Cache Diagram



Taken from Nigel Griffith's Blog at:

<http://tinyurl.com/k7g5dr7>

10/15/2017

38



Flash Cache Setup 1

• Setup cache pool

- I have 4 x 387GB SSDs for my pool
- #Using hdisk2, 3, 8, 9 for the cache pool
- cache_mgt pool create -d hdisk2,hdisk3,hdisk8,hdisk9 -p cmpool0
- Above creates the pool cmpool0 and the VG cmpool0
- # TOTAL PPs: 2888 (2957312 megabytes)
- cache_mgt partition create -p cmpool0 -s 2957308M -P cm1part1

```
lsvg -l cmpool0
```

```
#cmpool0:
```

#LV NAME	TYPE	LPs	PPs	PVs	LV STATE	MOUNT POINT
#cm1part1	jfs	2888	2888	4	closed/syncd	N/A

```
lsvg -p cmpool0
```

```
cmpool0:
```

PV_NAME	PV STATE	TOTAL PPs	FREE PPs	FREE DISTRIBUTION
hdisk2	active	722	0	00..00..00..00..00
hdisk3	active	722	0	00..00..00..00..00
hdisk8	active	722	0	00..00..00..00..00
hdisk9	active	722	0	00..00..00..00..00

10/15/2017

39



Flash Cache Setup 2

• Assign source disks

Assign hdisk106-109 as the source disks to be cached from

```
cache_mgt partition assign -t hdisk106 -P cm1part1
```

```
cache_mgt partition assign -t hdisk107 -P cm1part1
```

```
cache_mgt partition assign -t hdisk108 -P cm1part1
```

```
cache_mgt partition assign -t hdisk109 -P cm1part1
```

These are the disks where my data is in filesystems that I will be working from

```
cache_mgt cache list
```

```
hdisk106,cm1part1,inactive
```

```
hdisk107,cm1part1,inactive
```

```
hdisk108,cm1part1,inactive
```

```
hdisk109,cm1part1,inactive
```

INACTIVE means cache is not started!

```
cache_mgt device list -l | grep pool
```

```
hdisk2,cmpool0
```

```
hdisk3,cmpool0
```

```
hdisk8,cmpool0
```

```
hdisk9,cmpool0
```

```
cache_mgt pool list -l
```

```
cmpool0,hdisk2,hdisk3,hdisk8,hdisk9
```

10/15/2017

40



Flash Cache Setup 3

```
cache_mgt partition list -l
cm1part1,2957312M,cmpool0,hdisk106,hdisk107,hdisk108,hdisk109
```

```
cache_mgt cache list
hdisk106,cm1part1,inactive
hdisk107,cm1part1,inactive
hdisk108,cm1part1,inactive
hdisk109,cm1part1,inactive
```

```
cache_mgt engine list -l
/usr/ccs/lib/libcehandler.a,max_pool=1,max_partition=1,tgt_per_cache=unlimited,cache_per_tgt=1
```

START THE CACHE SYSTEM

```
cache_mgt cache start -t all
-t all says start caching all the assigned hdisks
-t hdisk106 would just do hdisk106 if it has been assigned
```

TO UNASSIGN HDISKS

```
First stop caching then unassign
cache_mgt cache stop -t all
cache_mgt partition unassign -t hdisk106 -P cm1part1
cache_mgt partition unassign -t hdisk107 -P cm1part1
cache_mgt partition unassign -t hdisk108 -P cm1part1
cache_mgt partition unassign -t hdisk109 -P cm1part1
```

10/15/2017

41



Flash Cache Monitoring

```
cache_mgt monitor start
cache_mgt monitor stop
cache_mgt get -h -s
Above gets stats since caching started. But no average – it lists stats for every source disk so if you have 88 of
them it is a very long report
```

```
pfcrcs -a dump_stats
Undocumented but provides same statistics as above averaged for last 60 and 3600 seconds
Provides an overall average then the stats for each disk
```

Meaning of Statistics fields can be found at: <http://tinyurl.com/kesmvft>

Known Problems

```
cache_mgt list gets core dump if >70 disks in source – IV93772
http://www-01.ibm.com/support/docview.wss?crawler=1&uid=sg1IV93772
```

```
PFC_CAC_NOMEM error – IV91971
http://www-01.ibm.com/support/docview.wss?uid=sg1IV91971
D47E07BC 0413165117 P U ETCACHE NOT ENOUGH MEMORY TO ALLOCATE
Saw this when I had 88 source disks and went from 4 to 8 target SSDs
System had 512GB memory and only 256GB was in use
Waiting on an ifix when upgraded to AIX 7.2 tl01 sp2 – IJ00115s2a is new ifix
```

```
PFC_CAC_DASTOOSLOW
Claims DAS DEVICE IS SLOWER THAN SAN DEVICE
This is on hold till we install the apar for the PFC_CAC_NOMEM error
```

10/15/2017

42



Flash Cache pfcra output

```

Sector size:          512 bytes
Fragment size:        1048576 bytes
Extent size:          1073741824 bytes
Cache size:           3100966387712 bytes

Cache state: Cache is warm

Reporting average statistics for the last 60 and 3600 seconds.

*****
* Global Cache Stats *
*****
Allocated Space:      1363706642432 (43.98% of total cache space)
Valid data:           1360417214464 (99.76% of allocated space)

Cache operations      |          60 sec |          3600 sec |
-----|-----|-----|
Hit Rate              |          100.00% |          98.53% |
Partial Hit Rate      |           0.00% |           0.00% |
Lookups               |           297 |         418824 |
Promotes              |           0 |           259 |
Partial Promotes      |           0 |           0 |
Server Promotes       |           0 |           0 |
Invalidates           |           0 |          1836 |
Purges                |           0 |           0 |
-----|-----|-----|
DAS I/O Stats
-----|-----|-----|
Avg. data read per second |         1419543048 |         2444089319 |
Avg. read request size   |           950784 |           846336 |
Avg. read latency (usec)  |           3616 |           3241 |
Avg. data written per second |           0 |           0 |
Avg. write request size  |           0 |           0 |
Avg. write latency (usec) |           0 |           0 |
-----|-----|-----|
SAN I/O Stats
-----|-----|-----|
Avg. data read per second |           0 |           56842 |
Avg. read request size   |           0 |           32768 |
Avg. read latency (usec)  |           0 |           3384 |
Avg. data written per second |           0 |           0 |
Avg. write request size  |           0 |           0 |
Avg. write latency (usec) |           0 |           0 |

```

10/15/2017

* Per-LUN Cache Stats *

43



Thank you for your time



If you have questions please email me at:

Jaqui@circle4.com or jlynch@flagshipsg.com

Also check out:

<http://www.circle4.com/movies/>

44



Useful Links

- Jaqui Lynch Articles
 - <http://www.circle4.com/jaqui/eserver.html>
- 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 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>

45



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>

46



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>

47



Backup Slides

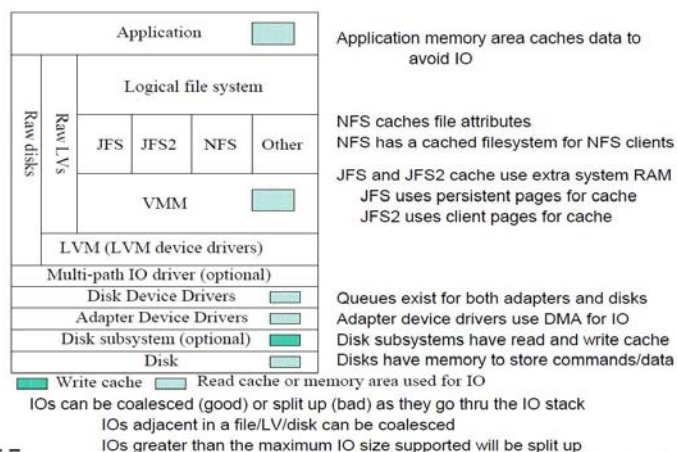


48



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

The AIX IO stack



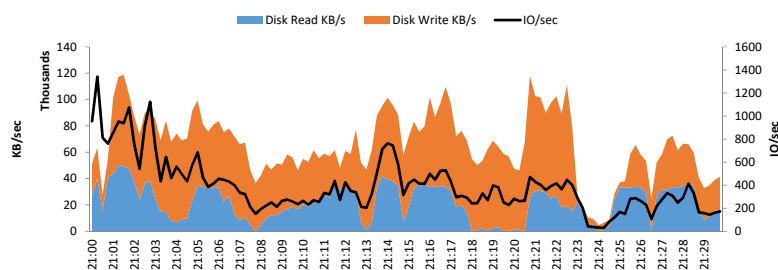
© 2010 IBM Corporation

49



disk_summ tab in nmon

Disk total KB/s b740ft1 - 1/12/2013

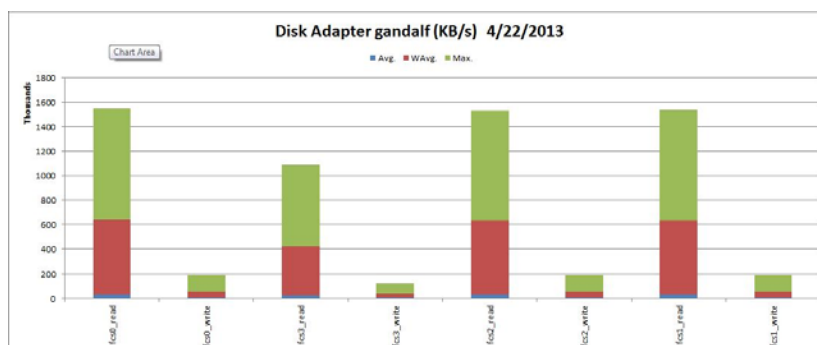


	Disk Read KB/s	Disk Write KB/s	IO/sec	Read+Write	R/W
Avg.	21695.8	43912.8	393.4	65608.6	64.1
Real Max	50481.1	92739.4	1340.8	118896.4	116.1

50



IOadapt tab in nmon

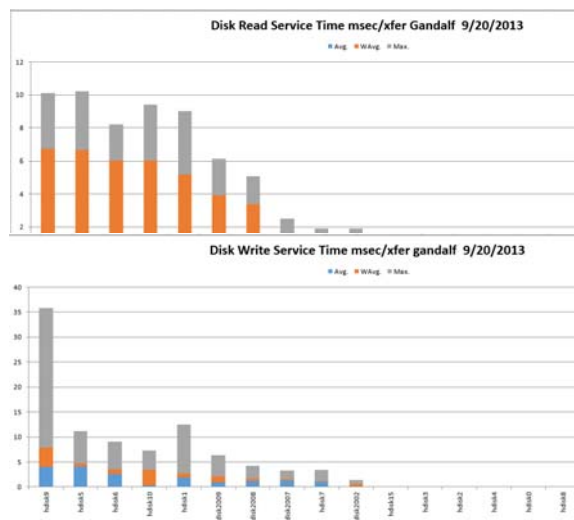


Are we balanced?

51



nmon Disk Service Times



52



Adapters



53



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
				C1	8GB DP fibre	1	lpar1	C1	8GB DP fibre	1	lpar1	C1	8GB DP fibre	1	vio1
				C2	4PT 10/100/1000	3	lpar1	C2	4PT 10/100/1000	3	lpar1	C2	4PT 10/100/1000	3	
				C3	8GB DP fibre	5	vio2	C3	8GB DP fibre	5	vio1	C3		5	
				C4	4PT 10/100/1000	6	vio2	C4	4PT 10/100/1000	6	vio1	C4	8GB DP fibre	2	vio2
				C5	8GB DP fibre	2	vio1	C5	8GB DP fibre	2	vio2	C5	4PT 10/100/1000	4	
				C6	4PT 10/100/1000	4	vio1	C6	4PT 10/100/1000	4	vio2	C6	4GB DP fibre	6	lpar1
												C7	4GB DP fibre	7	
				D1	146GB disk		vio1	D1	146GB disk		vio1	C8		8	
				D4	146GB disk		vio2	D4	146GB disk		vio2	C9		9	
												C10		10	

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

54



Power8 – S814 and S824 Adapter Slot Priority

S814 S824 Adapter Slots

ID	Slot	Type	S814 / S824 (1 socket populated)			S824 (2 sockets populated)		
			Feature	Description	Use	Feature	Description	Use
P1-C2	1	PCIe3 x8	Not available with 1-socket populated					
P1-C3	2	PCIe3 x16						
P1-C4	3	PCIe3 x8						
P1-C5	4	PCIe3 x16						
P1-C6	5	PCIe3 x16						
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		

55



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

56



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

© 2012 IBM Corporation

57

