pPE52 - AIX Performance Tuning - Part 2 – I/O

Agenda

• Part 1
  • CPU
  • Memory tuning
  • Starter Set of Tunables

• Part 2
  • I/O
  • Volume Groups and File systems
  • AIO and CIO for Oracle

• Part 3
  • Network
  • Performance Tools
### Adapter Priorities affect Performance

<table>
<thead>
<tr>
<th>EET</th>
<th>Top</th>
<th>12Gbps or 8Gbps cables</th>
<th>Bottom</th>
<th>24Gbps</th>
<th>32Gbps only I/O Device 12Gbps7</th>
<th>Disc</th>
<th>Pri</th>
<th>Alloc</th>
</tr>
</thead>
<tbody>
<tr>
<td>Slot</td>
<td>Desc</td>
<td>Alloc</td>
<td>Slot</td>
<td>Desc</td>
<td>Pri</td>
<td>Alloc</td>
<td>Slot</td>
<td>Desc</td>
</tr>
<tr>
<td>C1</td>
<td>8GB DP fibre</td>
<td>1</td>
<td>lpar1</td>
<td>C1</td>
<td>8GB DP fibre</td>
<td>1</td>
<td>lpar1</td>
<td>C1</td>
</tr>
<tr>
<td>C2</td>
<td>4PT 1G/100/1000</td>
<td>3</td>
<td>lpar1</td>
<td>C2</td>
<td>4PT 1G/100/1000</td>
<td>3</td>
<td>lpar1</td>
<td>C2</td>
</tr>
<tr>
<td>C3</td>
<td>8GB DP fibre</td>
<td>5</td>
<td>vio2</td>
<td>C3</td>
<td>8GB DP fibre</td>
<td>5</td>
<td>vio2</td>
<td>C3</td>
</tr>
<tr>
<td>C4</td>
<td>4PT 1G/100/1000</td>
<td>6</td>
<td>vio2</td>
<td>C4</td>
<td>4PT 1G/100/1000</td>
<td>6</td>
<td>vio2</td>
<td>C4</td>
</tr>
<tr>
<td>C5</td>
<td>8GB DP fibre</td>
<td>2</td>
<td>vio1</td>
<td>C5</td>
<td>8GB DP fibre</td>
<td>2</td>
<td>vio1</td>
<td>C5</td>
</tr>
<tr>
<td>C6</td>
<td>4PT 1G/100/1000</td>
<td>4</td>
<td>vio1</td>
<td>C6</td>
<td>4PT 1G/100/1000</td>
<td>4</td>
<td>vio1</td>
<td>C6</td>
</tr>
<tr>
<td>D1</td>
<td>144GB disk</td>
<td>7</td>
<td>vio1</td>
<td>D1</td>
<td>144GB disk</td>
<td>7</td>
<td>vio1</td>
<td>D1</td>
</tr>
<tr>
<td>D2</td>
<td>144GB disk</td>
<td>8</td>
<td>1</td>
<td>D2</td>
<td>144GB disk</td>
<td>8</td>
<td>1</td>
<td>D2</td>
</tr>
<tr>
<td>D3</td>
<td>144GB disk</td>
<td>9</td>
<td>1</td>
<td>D3</td>
<td>144GB disk</td>
<td>9</td>
<td>1</td>
<td>D3</td>
</tr>
<tr>
<td>D4</td>
<td>144GB disk</td>
<td>10</td>
<td>1</td>
<td>D4</td>
<td>144GB disk</td>
<td>10</td>
<td>1</td>
<td>D4</td>
</tr>
</tbody>
</table>

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
  - 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
### Adapter bandwidth

**Adapter Performance Chart**

<table>
<thead>
<tr>
<th>Adapter</th>
<th>FC</th>
<th>IOPS 4K</th>
<th>Sustained Sequential b/w</th>
</tr>
</thead>
<tbody>
<tr>
<td>2 Gbps FC adapter (single port)</td>
<td>5710</td>
<td>38,401</td>
<td>198 MB/s simplex, 385 MB/s duplex</td>
</tr>
<tr>
<td>4 Gbps FC adapter (single port)</td>
<td>5750</td>
<td>n/a</td>
<td>DDR slots: 400 MB/s simplex, ~750 MB/s duplex. SDR slots: 400 MB/s simplex, 500 MB/s duplex</td>
</tr>
<tr>
<td>4 Gbps FC adapter (dual)</td>
<td>5759</td>
<td>n/a</td>
<td>DDR slots: ~750 MB/s, SDR slots: ~500 MB/s</td>
</tr>
<tr>
<td>4 Gbps FC adapter PCIe</td>
<td>5773</td>
<td>n/a</td>
<td>400 MB/s simplex, ~750 MB/s duplex</td>
</tr>
<tr>
<td>4 Gbps FC adapter (dual) PCIe</td>
<td>5774</td>
<td>n/a</td>
<td>~750 MB/s</td>
</tr>
<tr>
<td>8 Gbps FC dual port PCIe</td>
<td>5735</td>
<td>142,000</td>
<td>750 MB/s per port simplex, 997 MB/s duplex per port</td>
</tr>
<tr>
<td>10 Gb FCoE PCIe Dual Port</td>
<td>5708</td>
<td>150,000</td>
<td>1475 MB/s simplex per adapter, 2000 MB/s duplex per</td>
</tr>
</tbody>
</table>

---

**disk_summ tab in nmon**

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

<table>
<thead>
<tr>
<th></th>
<th>Disk Read KB/s</th>
<th>Disk Write KB/s</th>
<th>IO/sec</th>
<th>Read-Write MB/s</th>
<th>R/W</th>
</tr>
</thead>
<tbody>
<tr>
<td>Avg.</td>
<td>23695.8</td>
<td>43912.8</td>
<td>393.4</td>
<td>65688.6</td>
<td>64.1</td>
</tr>
<tr>
<td>Real Max.</td>
<td>50481.1</td>
<td>92739.4</td>
<td>1140.8</td>
<td>118896.4</td>
<td>116.1</td>
</tr>
</tbody>
</table>
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
    - Client
    - External Pager
  - JFS
  - NFS and VxFS
  - 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
IO Wait and why it is not necessarily useful

**SMT2 example for simplicity**

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:
\[
\text{%user} = \frac{.8 \times 4}{4} = 80\%
\]
\[
\text{%sys} = \frac{.2 \times 4}{4} = 20\%
\]
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
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

Basics

• Data layout will have more impact than most tunables
  • Plan in advance

• Large hdisk 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 -1 hdisk -a queue_depth=20 (or some good value)
  • Default client qdepth for NPIV is set by the Multipath driver in the client
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 2 160.0 1.9
hdisk8 0.3 568 14337 23.5 2.3
hdisk9 0.0 31 149.0 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`
  Time waiting in the wait queue (ms)

- `avserv`
  I/O service time when sent to disk (ms)

- See articles by Dan Braden:
  - [http://www-03.ibm.com/support/techdocs/atsmastr.nsf/WebIndex/TD106122](http://www-03.ibm.com/support/techdocs/atsmastr.nsf/WebIndex/TD106122)

---

iostat -DI

| dev | act | bps | tps | btract | bwrite | avg | max | min | avg | wps | avg | max | avg | avg | serv | chg | sqsz | qfull |
|-----|-----|-----|-----|--------|--------|-----|-----|-----|-----|-----|-----|-----|-----|-----|-----|------|-----|------|------|
|     |     |     |     |        |        |     |     |     |     |     |     |     |     |     |     |     |     |     |
| hdisk0 | 19.7 | 255.3K | 34.4 | 682.4K | 0 | 0 | 0 | 0 | 0 | 0.1 | 3 | 1.6 | 4 | 10.4 | 10.8 | 139.2 | 2.4 | 0 | 81.3 | 0 | 0 | 2.1 |
| hdisk5 | 14.1 | 254.6K | 33.4 | 0 | 254.6K | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| hdisk16 | 17.7 | 254.6K | 33.4 | 0 | 254.6K | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| hdisk17 | 0.1 | 1.8K | 0.3 | 1.8K | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| hdisk15 | 6.4 | 2.3M | 4.9 | 2.3M | 279.1 | 4.8 | 19.5 | 2.9 | 97.5 | 0.1 | 7.8 | 1.1 | 14.4 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| hdisk18 | 0.1 | 2.1K | 0.5 | 2.1K | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| hdisk19 | 0.1 | 2.6K | 0.6 | 2.6K | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| hdisk20 | 1.4 | 257.4K | 2.4 | 257.4K | 0 | 2.4 | 27.7 | 0.2 | 163.2 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| hdisk21 | 5.1 | 29.8 | 2.4M | 0 | 29.8 | 1.7 | 0.2 | 50.1 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| hdisk22 | 13.3 | 12.2 | 2.3M | 0 | 12.2 | 16.4 | 0.2 | 248.5 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| hdisk24 | 12.2 | 5 | 2.3M | 0 | 5 | 34.6 | 0.2 | 221.9 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| hdisk26 | 7.9 | 2.1M | 4.5 | 2.1M | 0 | 4.5 | 32.1 | 0.1 | 201 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| hdisk27 | 6.2 | 2.1M | 4.5 | 2.1M | 0 | 4.5 | 25.4 | 0.6 | 219.5 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| hdisk28 | 1.2 | 2.1M | 4.5 | 2.1M | 0 | 4.5 | 30.3 | 3 | 101.6 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| hdisk29 | 6.8 | 1.2M | 4.5 | 2.2M | 0 | 4.5 | 26.6 | 3.1 | 210.2 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |

tps - Transactions per second – transfers per second to the adapter
avgtime - Average service time
avgtimeq - Average time in the wait queue
avgwqsz - Average wait queue size
avgwqsz - If regularly >0 increase queue_depth
avgqsz - Average service queue size (waiting to be sent to disk)
Can’t be larger than queue-depth for the disk
servfull - Number times the service queue was full

Look at iostat -aD for adapter queues
If avgwqsz > 0 or servfull high then increase queue_depth. Also look at avgqsz.
Per IBM
Average IO sizes:
read = bwrite/bps
write = bwrite/wps
nmon Disk Service Times

Adapter Queue Problems

- Look at BBBF Tab in NMON Analyzer or run fcstat command
- 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

- 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
Adapter Tuning

<table>
<thead>
<tr>
<th>fcs0</th>
<th>115</th>
<th>Bus interrupt level</th>
<th>False</th>
</tr>
</thead>
<tbody>
<tr>
<td>bus_intr_lv</td>
<td>0x00f0</td>
<td>Bus I/O address</td>
<td>False</td>
</tr>
<tr>
<td>bus_io_addr</td>
<td>0x8040000</td>
<td>Bus memory address</td>
<td>False</td>
</tr>
<tr>
<td>init_link</td>
<td>0x10</td>
<td>INIT Link flags</td>
<td>True</td>
</tr>
<tr>
<td>intr_priority</td>
<td>3</td>
<td>Interrupt priority</td>
<td>False</td>
</tr>
<tr>
<td>lg_term_dma</td>
<td>0x000000</td>
<td>Long term DMA</td>
<td>True</td>
</tr>
<tr>
<td>max_xfer_size</td>
<td>0x100000</td>
<td>Maximum Transfer Size</td>
<td>True</td>
</tr>
<tr>
<td>num_cmd_elems</td>
<td>200</td>
<td>Maximum number of COMMANDS to queue to the adapter</td>
<td>True</td>
</tr>
<tr>
<td>pref_alpa</td>
<td>0x1</td>
<td>Preferred AL_PA</td>
<td>True</td>
</tr>
<tr>
<td>sw_fc_class</td>
<td>2</td>
<td>FC Class for Fabric</td>
<td>True</td>
</tr>
</tbody>
</table>

Changes I often make (test first)

max_xfer_size 0x200000 Maximum Transfer Size True
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 -f fcs0 -a max_xfer_size=0x200000 -a num_cmd_elems=1024 -P
chdev -f 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

Remember make changes to 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

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 2048  Maximum Number of COMMANDS Elements True

NOTE NPIV client must be <= to settings on VIO
vmstat –v Output TSM System – Fairly Healthy

Up 1 day 6 hours
3 memory pools
3.0 minperm percentage
90.0 maxperm percentage
12.1 numperm percentage
12.1 numclient percentage
90.0 maxclient percentage
76.8 percentage of memory used for computational pages

0 pending disk I/Os blocked with no pbuf
0 paging space I/Os blocked with no psbuf
1972 file system I/Os blocked with no fsbuf
318352 client file system I/Os blocked with no fsbuf
158410 external pager file system I/Os blocked with no fsbuf

Based on the blocked I/Os it is clearly a system using JFS2.
It is also experiencing some network problems – not necessarily NFS but network needs review.
Note – even with no JFS in the system you will see between 1700 and 2200 filesystem I./Os blocked with no fsbuf – no idea why but I see it all the time.

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
11173706 paging space I/Os blocked with no psbuf
2048 file system I/Os blocked with no fsbuf
238 client file system I/Os blocked with no fsbuf
39943187 external pager file system I/Os blocked with no fsbuf

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.
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

vname = 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 vgname-a
For other VGs we see the following in pervg_blocked_io_count:

<table>
<thead>
<tr>
<th>vgname</th>
<th>blocked</th>
<th>total_vg_pbufs</th>
</tr>
</thead>
<tbody>
<tr>
<td>nimvg</td>
<td>29</td>
<td>512</td>
</tr>
<tr>
<td>sasvg</td>
<td>2719199</td>
<td>1024</td>
</tr>
<tr>
<td>backupvg</td>
<td>6042</td>
<td>4608</td>
</tr>
</tbody>
</table>

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

Parameter Settings - Summary

<table>
<thead>
<tr>
<th>PARAMETER</th>
<th>DEFAULTS</th>
<th>AIXv5.3</th>
<th>AIXv6</th>
<th>AIXv7</th>
<th>NEW SET ALL TO</th>
</tr>
</thead>
<tbody>
<tr>
<td>NETWORK (no)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>rfc1323</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td></td>
</tr>
<tr>
<td>tcp_sendspace</td>
<td>16384</td>
<td>16384</td>
<td>16384</td>
<td>65536 (1Gb)</td>
<td></td>
</tr>
<tr>
<td>tcp_recvspace</td>
<td>16384</td>
<td>16384</td>
<td>16384</td>
<td>65536 (1Gb)</td>
<td></td>
</tr>
<tr>
<td>udp_sendspace</td>
<td>9216</td>
<td>9216</td>
<td>9216</td>
<td>65536</td>
<td></td>
</tr>
<tr>
<td>udp_recvspace</td>
<td>42080</td>
<td>42080</td>
<td>42080</td>
<td>655360</td>
<td></td>
</tr>
<tr>
<td>MEMORY (vmo)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>minperm%</td>
<td>20</td>
<td>3</td>
<td>3</td>
<td>3</td>
<td></td>
</tr>
<tr>
<td>maxperm%</td>
<td>80</td>
<td>90</td>
<td>90</td>
<td>90</td>
<td>JFS, NFS, VxFS, JFS2</td>
</tr>
<tr>
<td>maxclien%</td>
<td>80</td>
<td>90</td>
<td>90</td>
<td>90</td>
<td>JFS2, NFS</td>
</tr>
<tr>
<td>lru_file_repage</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td></td>
</tr>
<tr>
<td>lru_poll_interval</td>
<td>7</td>
<td>10</td>
<td>10</td>
<td>10</td>
<td></td>
</tr>
<tr>
<td>Minfree</td>
<td>960</td>
<td>960</td>
<td>960</td>
<td>calculation</td>
<td></td>
</tr>
<tr>
<td>Maxfree</td>
<td>1088</td>
<td>1088</td>
<td>1088</td>
<td>calculation</td>
<td></td>
</tr>
<tr>
<td>page_steal_method</td>
<td>0</td>
<td>0/1 (TL)</td>
<td>1</td>
<td>1</td>
<td></td>
</tr>
<tr>
<td>JFS2 (ioo)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>ji2_maxPageReadAhead</td>
<td>128</td>
<td>128</td>
<td>128</td>
<td>as needed</td>
<td></td>
</tr>
<tr>
<td>ji2_DynamicBufferPreallocation</td>
<td>16</td>
<td>16</td>
<td>16</td>
<td>as needed</td>
<td></td>
</tr>
</tbody>
</table>
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:

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
```
### Most Active Logical Volumes

<table>
<thead>
<tr>
<th>util</th>
<th>#rblk</th>
<th>#wblk</th>
<th>KB/s</th>
<th>volume</th>
<th>description</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.66</td>
<td>4647264</td>
<td>834573</td>
<td>45668.9</td>
<td>/dev/gandalfp_ga71_lv</td>
<td>/ga71</td>
</tr>
<tr>
<td>0.36</td>
<td>960</td>
<td>834565</td>
<td>6960.7</td>
<td>/dev/gandalfp_ga73_lv</td>
<td>/ga73</td>
</tr>
<tr>
<td>0.13</td>
<td>2430816</td>
<td>13448</td>
<td>20363.1</td>
<td>/dev/misc_gm10_lv</td>
<td>/gm10</td>
</tr>
<tr>
<td>0.11</td>
<td>53808</td>
<td>14800</td>
<td>571.6</td>
<td>/dev/gandalfp_ga15_lv</td>
<td>/ga15</td>
</tr>
<tr>
<td>0.08</td>
<td>94416</td>
<td>7616</td>
<td>850.0</td>
<td>/dev/gandalfp_ga10_lv</td>
<td>/ga10</td>
</tr>
<tr>
<td>0.07</td>
<td>787632</td>
<td>6296</td>
<td>6614.2</td>
<td>/dev/misc_gm73_lv</td>
<td>/gm73</td>
</tr>
<tr>
<td>0.05</td>
<td>8256</td>
<td>24259</td>
<td>270.9</td>
<td>/dev/misc_gm72_lv</td>
<td>/gm72</td>
</tr>
<tr>
<td>0.05</td>
<td>58176</td>
<td>22088</td>
<td>668.7</td>
<td>/dev/misc_gm71_lv</td>
<td>/gm71</td>
</tr>
</tbody>
</table>

### Most Active Physical Volumes

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

• Useful Commands
  • pcmpath query device
  • pcmpath query devstats
  • pcmpath query adapter
  • pcmpath query adaptstats
  • pcmpath query version
  • pcmpath query wwpn
  • pcmpath query port
  • pcmpath query portstats
  • pcmpath query essmap
  • sddpcm_get_config –Av
  • See example output in backup slides at end

ORACLE
Asynchronous I/O and Concurrent I/O
Async I/O - v5.3

Total number of AIOs in use
ps -al | grep aios | wc -l
Maximum AIOservers started since boot
NB - maxservers is a per processor setting in AIX 5.3

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 I0O THERE
Isattr -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.

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_wrt
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
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

```
ls /a -F | more
aio_active = 0
aio_maxreqs = 65536
aio_maxservers = 30
aio_server_inactivity = 300
posix_aio_active = 0
posix_aio_maxreqs = 65536
posix_aio_maxservers = 30
posix_aio_server_inactivity = 300
```

```
psstat -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

<table>
<thead>
<tr>
<th>#Restricted tunables</th>
</tr>
</thead>
<tbody>
<tr>
<td>aio_fastpath = 1</td>
</tr>
<tr>
<td>aio_fsfastpath = 1</td>
</tr>
<tr>
<td>aio_kprocprio = 39</td>
</tr>
<tr>
<td>aio_multitidsusp = 1</td>
</tr>
<tr>
<td>aio_sample_rate = 5</td>
</tr>
<tr>
<td>aio_samples_per_cycle = 6</td>
</tr>
<tr>
<td>posix_aio_fastpath = 1</td>
</tr>
<tr>
<td>posix_aio_fsfastpath = 1</td>
</tr>
<tr>
<td>posix_aio_kprocprio = 39</td>
</tr>
<tr>
<td>posix_aio_sample_rate = 5</td>
</tr>
<tr>
<td>posix_aio_samples_per_cycle = 6</td>
</tr>
</tbody>
</table>

AIO Recommendations

Oracle now recommending the following as starting points

<table>
<thead>
<tr>
<th>5.3</th>
<th>6.1 or 7 (non ClO)</th>
</tr>
</thead>
<tbody>
<tr>
<td>minservers</td>
<td>100</td>
</tr>
<tr>
<td>maxservers</td>
<td>200</td>
</tr>
<tr>
<td>maxreqs</td>
<td>16384</td>
</tr>
</tbody>
</table>

These are per CPU
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
PROCAIO tab in nmon

Maximum seen was 192 but average was much less

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
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 irud 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 –IW flags

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 blocksizes**
    - I give each instance its own filesystem and their redo logs are also separate

- Leave DISKASYNCHIO=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
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

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)
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:

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:
  - Firmware history including release dates can be found at:
    - Power7 Midrange
    - Power7 High end
- 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
Useful Links

- Jaqui Lynch Articles
  - [http://www.ibmsystemsmag.com/authors/Jaqui-Lynch/](http://www.ibmsystemsmag.com/authors/Jaqui-Lynch/)

- Charlie Cler Articles
  - [http://www.ibmsystemsmag.com/authors/Charlie-Cler/](http://www.ibmsystemsmag.com/authors/Charlie-Cler/)

- Andrew Goade Articles
  - [http://www.ibmsystemsmag.com/authors/Andrew-Goade/](http://www.ibmsystemsmag.com/authors/Andrew-Goade/)

- Jaqui’s Upcoming Talks and Movies
  - Upcoming Talks
  - Movie replays
    - [http://www.circle4.com/movies](http://www.circle4.com/movies)

Useful Links

- AIX Virtual User Group site:
  - [http://www.tinyurl.com/ibmaixvug](http://www.tinyurl.com/ibmaixvug)

- UK PowerVM User group at:

- Nigel on Entitlements and VPs plus 7 most frequently asked questions

- AIXpert Blog

- 10 Golden rules for rPerf Sizing

- Youtube channel
  - [http://www.youtube.com/user/nigelargriffiths](http://www.youtube.com/user/nigelargriffiths)

- Jay Kruemke Twitter – chromeaix
  - [https://twitter.com/chromeaix](https://twitter.com/chromeaix)

- Nigel Griffiths Twitter – mr_nmon
  - [https://twitter.com/mr_nmon](https://twitter.com/mr_nmon)

- Gareth Coates Twitter – power_gaz
  - [https://twitter.com/power_gaz](https://twitter.com/power_gaz)
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#Other+PerformanceTools-VIOSPA

References

• Simultaneous Multi-Threading on POWER7 Processors by Mark Funk
• Processor Utilization in AIX by Saravanan Devendran
• SG24-7940 - PowerVM Virtualization - Introduction and Configuration
• SG24-7590 – PowerVM Virtualization – Managing and Monitoring
• SG24-8080 – Power Systems Performance Guide – Implementing and Optimizing
• SG24-8079 – Power 7 and 7+ Optimization and Tuning Guide
• Redbook Tip on Maximizing the Value of P7 and P7+ through Tuning and Optimization
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/

SDDPCM
Examples
sddpcm

pcmpath query device

DEV#: 26  DEVICE NAME: hdisk26  TYPE: 2145  ALGORITHM: Load Balance

===========================================================================
Path#  Adapter/PathName     State  Mode         Select   Errors
0*     fscsi1/path0         OPEN   NORMAL      0         0
1       fscsi2/path12       OPEN   NORMAL      0         0
2*     fscsi2/path12       OPEN   NORMAL      0         0
3       fscsi1/path15       OPEN   NORMAL      0         0
4*     fscsi0/path8         OPEN   NORMAL      0         0
5       fscsi0/path10       OPEN   NORMAL      0         0
6*     fscsi3/path4         OPEN   NORMAL      0         0
7       fscsi3/path6         OPEN   NORMAL 22500688  0

sddpcm

pcmpath query devstats

DEV#: 26  DEVICE NAME: hdisk26

===========================================================================
I/O:  Total Read     Total Write     Active Read    Active Write     Maximum
      20060369    96466878            0            1              20
SECTOR: 1545992323  923040348         0            8           10240

Transfer Size:  <= 512  <= 4k  <= 16K  <= 64K  > 64K
          4736800  102527863  4163516  1834356  3264712
### sddpcm

**pcmpath query adapter**

Total Dual Active and Active/Asymmetric Adapters : 4

<table>
<thead>
<tr>
<th>Adpt#</th>
<th>Name</th>
<th>State</th>
<th>Mode</th>
<th>Select</th>
<th>Errors</th>
<th>Paths</th>
<th>Active</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>fscsi1</td>
<td>NORMAL</td>
<td>ACTIVE</td>
<td>2939082738</td>
<td>0</td>
<td>296</td>
<td>294</td>
</tr>
<tr>
<td>1</td>
<td>fscsi3</td>
<td>NORMAL</td>
<td>ACTIVE</td>
<td>2976510807</td>
<td>0</td>
<td>296</td>
<td>294</td>
</tr>
<tr>
<td>2</td>
<td>fscsi0</td>
<td>NORMAL</td>
<td>ACTIVE</td>
<td>2986133005</td>
<td>0</td>
<td>296</td>
<td>294</td>
</tr>
<tr>
<td>3</td>
<td>fscsi2</td>
<td>NORMAL</td>
<td>ACTIVE</td>
<td>2944614956</td>
<td>0</td>
<td>296</td>
<td>294</td>
</tr>
</tbody>
</table>

### sddpcm

**pcmpath query adaptstats**

Total Dual Active and Active/Asymmetric Adapters : 4

**Adapter # : 0**

<table>
<thead>
<tr>
<th>I/O:</th>
<th>Total Read</th>
<th>Total Write</th>
<th>Active Read</th>
<th>Active Write</th>
<th>Maximum</th>
</tr>
</thead>
<tbody>
<tr>
<td>SECTOR</td>
<td>155083544632</td>
<td>7126792194</td>
<td>0</td>
<td>0</td>
<td>42321</td>
</tr>
</tbody>
</table>

**Adapter # : 1**

<table>
<thead>
<tr>
<th>I/O:</th>
<th>Total Read</th>
<th>Total Write</th>
<th>Active Read</th>
<th>Active Write</th>
<th>Maximum</th>
</tr>
</thead>
<tbody>
<tr>
<td>SECTOR</td>
<td>160450579441</td>
<td>71807565761</td>
<td>0</td>
<td>0</td>
<td>42692</td>
</tr>
</tbody>
</table>

**Adapter # : 2**

<table>
<thead>
<tr>
<th>I/O:</th>
<th>Total Read</th>
<th>Total Write</th>
<th>Active Read</th>
<th>Active Write</th>
<th>Maximum</th>
</tr>
</thead>
<tbody>
<tr>
<td>SECTOR</td>
<td>161163171012</td>
<td>71921771564</td>
<td>0</td>
<td>0</td>
<td>42375</td>
</tr>
</tbody>
</table>

**Adapter # : 3**

<table>
<thead>
<tr>
<th>I/O:</th>
<th>Total Read</th>
<th>Total Write</th>
<th>Active Read</th>
<th>Active Write</th>
<th>Maximum</th>
</tr>
</thead>
<tbody>
<tr>
<td>SECTOR</td>
<td>155636065200</td>
<td>71309018625</td>
<td>0</td>
<td>0</td>
<td>42246</td>
</tr>
</tbody>
</table>