The Abdus Salam
International Centre
for Theoretical Physics

United Nations
Educational, Scientific and
Cultural Organization

IAEA
International Atomic Energy Agency

Measuring and Optimizing NFS/NAS performance

Dr. Clement Onime

onime@ictp.it

Overview

- NAS
- Measuring performance
- Optimizing/improving performance

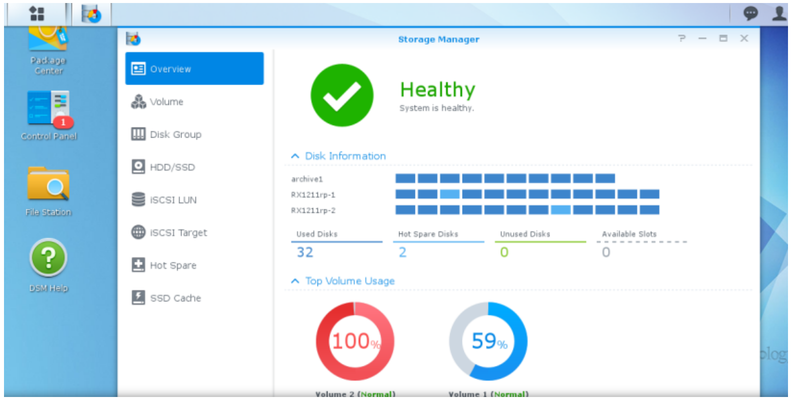
Network Attached Storage

- Generic term for storage devices that provide scalable storage and file-system for shared file level access.
- Wide range of commercial products and non-commercial

Images from https://en.wikipedia.org/wiki/Network-attached_storage and http://www.synology.com

NAS: Commercial products

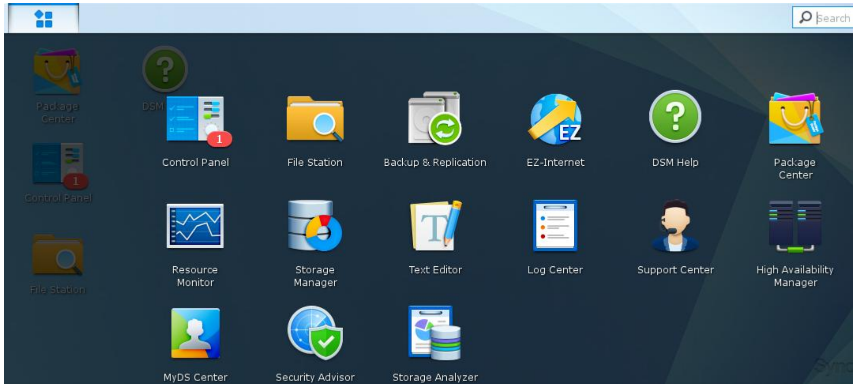
- Commercial NAS
  - Dedicated motherboards optimized for
    - I/O
    - Low power
  - Low to medium range
    - Custom Web User interface
    - Limited range of applications
  - High End
    - Performance and high availability
    - Data protection



Healthy
System is healthy.

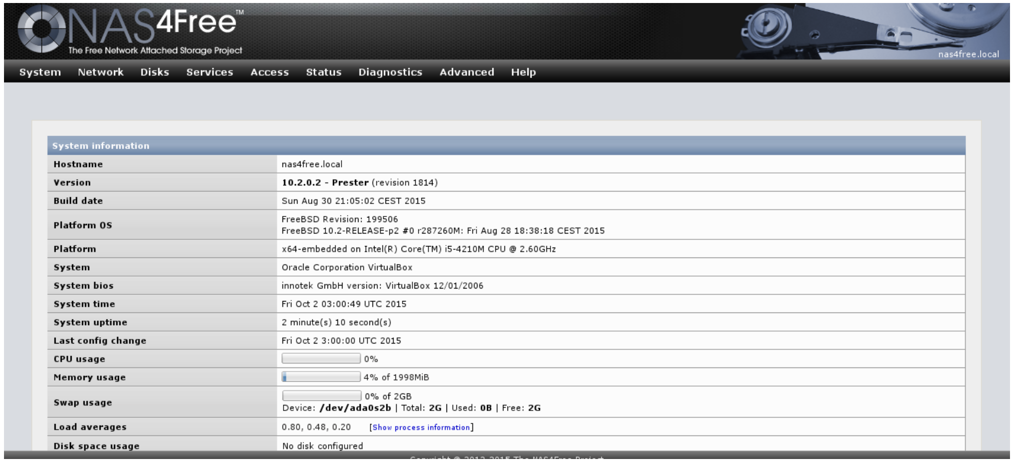
Used Disks 32
Hot Spare Disks 2
Unused Disks 0
Available Slots 0

Volume 2 (Normal)
Volume 1 (Normal)



NAS: Commercial or non-commercial

- Non-commercial
  - Choose your own hardware
    - Doing it yourself (DIY)
      - Regular Unix O.S. server
        - » Linux/FreeBSD
    - Free implementations
      - Customized, stripped down Unix distributions
        - » NAS4Free, FreeNAS, OpenFiler, etc..



System information	
Hostname	nas4free.local
Version	**10.2.0.2 - Prester** (revision 1814)
Build date	Sun Aug 30 21:05:02 CEST 2015
Platform OS	FreeBSD Revision: 199506 FreeBSD 10.2-RELEASE-p2 #0 r287260M: Fri Aug 28 18:38:18 CEST 2015
Platform	x64-embedded on Intel(R) Core(TM) i5-4210M CPU @ 2.60GHz
System	Oracle Corporation VirtualBox
System bios	innotek GmbH version: VirtualBox 12/01/2006
System time	Fri Oct 2 03:00:49 UTC 2015
System uptime	2 minute(s) 10 second(s)
Last config change	Fri Oct 2 3:00:00 UTC 2015
CPU usage	0%
Memory usage	4% of 1998MiB
Swap usage	0% of 2GB Device: **/dev/ada0s2b** | Total: **2G** | Used: **0B** | Free: **2G**
Load averages	0.80, 0.48, 0.20 [Show process information]
Disk space usage	No disk configured

Notes on NAS products for HPC

- Scalable Capacity
  - Maximum capacity
  - Limits on FS
- Reliability
  - Periodic testing of individual hardware components
  - Immediate reporting of faults
  - Data-protection rebuild/recovery time
  - File system check (fsck) or recovery time
- Data protection
  - Block level protection
    - Redundant Array of Independent Disks (RAID)
  - Data replication to another device in near real-time
  - Ability to back-up data directly.
- Running Costs
  - Parts, licenses, expansion
- Interfaces
  - Throughput and concurrent access

Data protection of RAID levels

Level	Useable capacity	Data protection
RAID0	Size$_{min}$ * n	None
RAID1	Size$_{min}$	Failure of one single disk
RAID5	Size$_{min}$ * (n – 1)	Concurrent failure of one single disk
RAID6	Size$_{min}$ * (n – 2)	Concurrent failure of two disks
RAID1+0	Size$_{min}$ * (n/2)	Concurrent failure of more than two disks

Hardware RAID

Characteristics	Hardware RAID	BIOS RAID	Software RAID
Cache RAM	dedicated	shared	shared
Battery backup unit	Yes (48 hours)	No	No
Raw data disk Portability	Not recommended *(Works for same controller family)*	Not sure	Yes *(works for same O.S)*
Configuration tool	Dedicated firmware based	Firmware+Host O.S	Host O.S
Hot disk replacement	yes	No recommended	Not recommended
Performance enhancement	Yes (faster)	none	none

Notes on hardware RAID Volumes

- Typical unit presented to O.S
  - Provisioning is mostly about ability to expand
    - Reduction may require destroying and make a new one
- States
  - NORMAL
  - DIRTY
  - DEGRADED

Network File System

- Version 3 (NFS or NFSv3)
  - Most widely deployed implementation
  - Simple security system
    - IP address based
    - UID user authentication with POSIX/Unix permissions and ability to exclude UID=0 (root user).
- Version 4 (NFSv4)
  - Improved security thanks to kerberos 5 user authentication
- Version 4.1 (pNFS)
  - Improved performance: Separating metadata from data

– Clients ←→ Server architecture

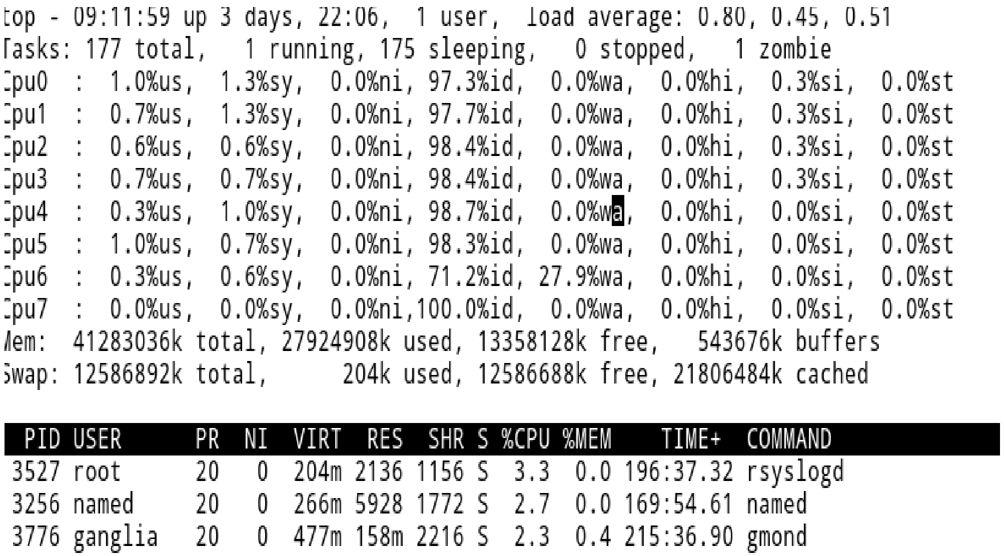
MEASURING NFS PERFORMANCE

Slowdown in performance

- Some general causes
  - Faulty connectivity (network)
  - Bad/faulty disk
  - Failing disk
  - Bad power supply unit
  - Server kernel panic or crash
    - Simply needs hard power cycle)
- Scalability issues
  - High load average on server
    - other processes/services on server
  - Overloaded, too many clients

Identifying bottlenecks

- Unix “top” command
  - Quickly determine if problem is CPU or I/O

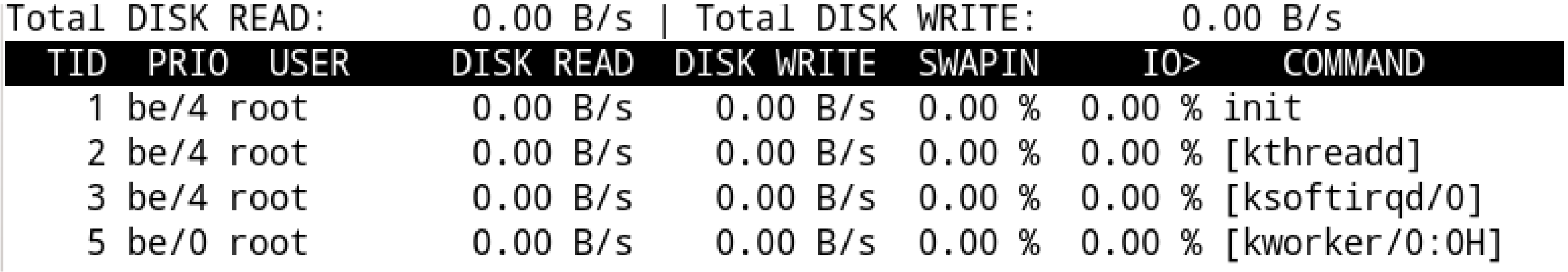


```
top - 09:11:59 up 3 days, 22:06,  1 user,  load average: 0.80, 0.45, 0.51
Tasks: 177 total,   1 running, 175 sleeping,   0 stopped,   1 zombie
Cpu0  :  1.0%us,  1.3%sy,  0.0%ni, 97.3%id,  0.0%wa,  0.0%hi,  0.3%si,  0.0%st
Cpu1  :  0.7%us,  1.3%sy,  0.0%ni, 97.7%id,  0.0%wa,  0.0%hi,  0.3%si,  0.0%st
Cpu2  :  0.6%us,  0.6%sy,  0.0%ni, 98.4%id,  0.0%wa,  0.0%hi,  0.3%si,  0.0%st
Cpu3  :  0.7%us,  0.7%sy,  0.0%ni, 98.4%id,  0.0%wa,  0.0%hi,  0.3%si,  0.0%st
Cpu4  :  0.3%us,  1.0%sy,  0.0%ni, 98.7%id,  0.0%wa,  0.0%hi,  0.0%si,  0.0%st
Cpu5  :  1.0%us,  0.7%sy,  0.0%ni, 98.3%id,  0.0%wa,  0.0%hi,  0.0%si,  0.0%st
Cpu6  :  0.3%us,  0.6%sy,  0.0%ni, 71.2%id, 27.9%wa,  0.0%hi,  0.0%si,  0.0%st
Cpu7  :  0.0%us,  0.0%sy,  0.0%ni,100.0%id,  0.0%wa,  0.0%hi,  0.0%si,  0.0%st
Mem:  41283036k total, 27924908k used, 13358128k free,   543676k buffers
Swap: 12586892k total,      204k used, 12586688k free, 21806484k cached

  PID USER      PR  NI  VIRT  RES  SHR S %CPU %MEM    TIME+  COMMAND
 3527 root      20   0  204m 2136 1156 S  3.3  0.0 196:37.32 rsyslogd
 3256 named     20   0  266m 5928 1772 S  2.7  0.0 169:54.61 named
 3776 ganglia   20   0  477m 158m 2216 S  2.3  0.4 215:36.90 gmond
```

Identifying top talkers

- Unix “iotop” command
  - Which user or processes are generating IO traffic



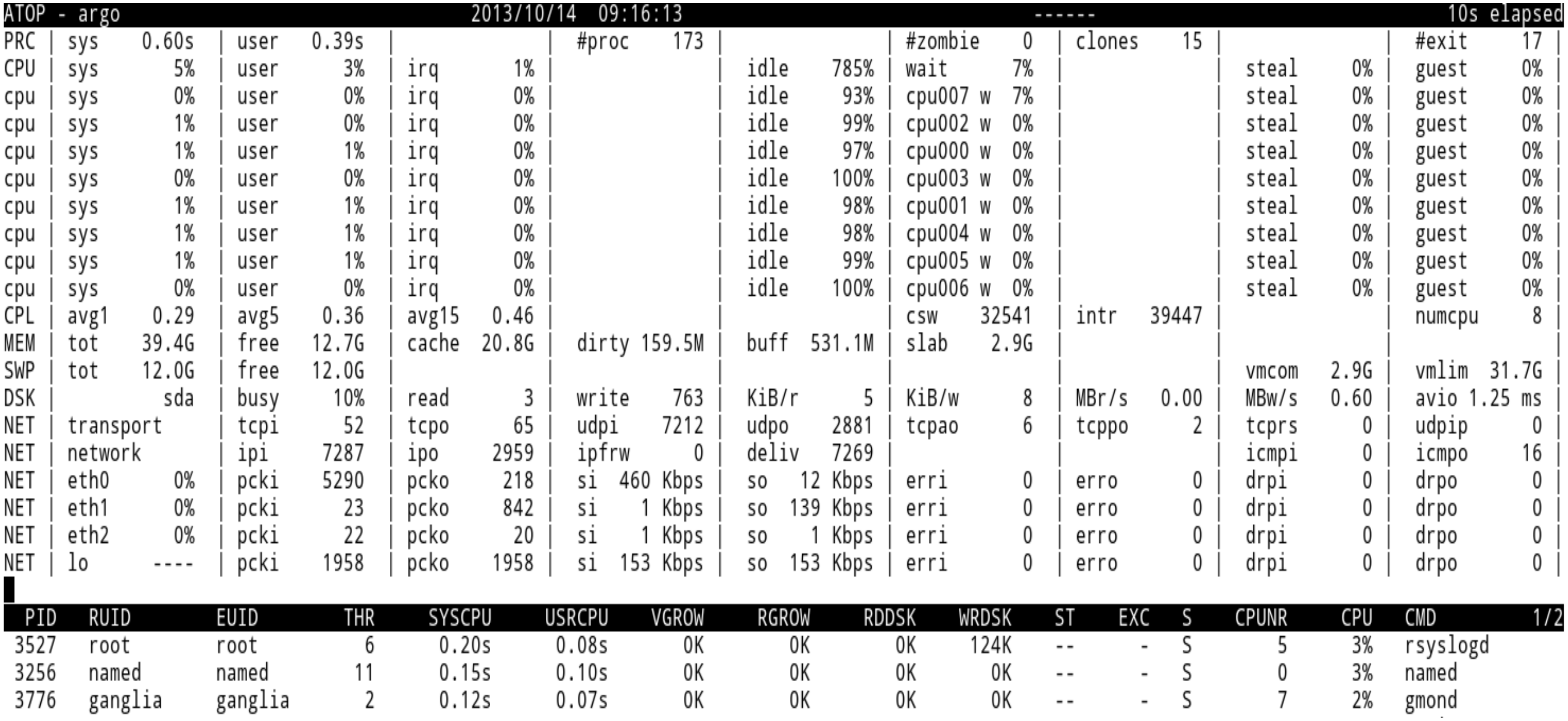
```
Total DISK READ:       0.00 B/s | Total DISK WRITE:       0.00 B/s
  TID  PRIO  USER     DISK READ  DISK WRITE  SWAPIN     IO>    COMMAND
    1 be/4 root       0.00 B/s    0.00 B/s  0.00 %  0.00 % init
    2 be/4 root       0.00 B/s    0.00 B/s  0.00 %  0.00 % [kthreadd]
    3 be/4 root       0.00 B/s    0.00 B/s  0.00 %  0.00 % [ksoftirqd/0]
    5 be/0 root       0.00 B/s    0.00 B/s  0.00 %  0.00 % [kworker/0:0H]
```

Identifying slow disks or network

- Unix “atop” command
  - Presents integrated view of CPU, RAM, network IO or individual disk IO

- Unix command “iptraf” presents view of network traffic for analysis

Output of “atop” command



```
ATOP - argo                       2013/10/14  09:16:13                     ------                   10s elapsed
PRC | sys    0.60s | user   0.39s |              | #proc    173 |              | #zombie    0 | clones    15 |              | #exit     17 |
CPU | sys       5% | user      3% | irq       1% |              | idle    785% | wait      7% |              | steal     0% | guest     0% |
cpu | sys       0% | user      0% | irq       0% |              | idle     93% | cpu007 w  7% |              | steal     0% | guest     0% |
cpu | sys       1% | user      0% | irq       0% |              | idle     99% | cpu002 w  0% |              | steal     0% | guest     0% |
cpu | sys       1% | user      1% | irq       0% |              | idle     97% | cpu000 w  0% |              | steal     0% | guest     0% |
cpu | sys       0% | user      0% | irq       0% |              | idle    100% | cpu003 w  0% |              | steal     0% | guest     0% |
cpu | sys       1% | user      1% | irq       0% |              | idle     98% | cpu001 w  0% |              | steal     0% | guest     0% |
cpu | sys       1% | user      1% | irq       0% |              | idle     98% | cpu004 w  0% |              | steal     0% | guest     0% |
cpu | sys       1% | user      1% | irq       0% |              | idle     99% | cpu005 w  0% |              | steal     0% | guest     0% |
cpu | sys       0% | user      0% | irq       0% |              | idle    100% | cpu006 w  0% |              | steal     0% | guest     0% |
CPL | avg1    0.29 | avg5    0.36 | avg15   0.46 |              |              | csw    32541 | intr   39447 |              | numcpu     8 |
MEM | tot    39.4G | free   12.7G | cache  20.8G | dirty 159.5M | buff  531.1M | slab    2.9G |              |              |              |
SWP | tot    12.0G | free   12.0G |              |              |              |              |              | vmcom   2.9G | vmlim  31.7G |
DSK |          sda | busy     10% | read       3 | write    763 | KiB/r      5 | KiB/w      8 | MBr/s   0.00 | MBw/s   0.60 | avio 1.25 ms |
NET | transport    | tcpi      52 | tcpo      65 | udpi    7212 | udpo    2881 | tcpao      6 | tcppo      2 | tcprs      0 | udpip      0 |
NET | network      | ipi     7287 | ipo     2959 | ipfrw      0 | deliv   7269 |              |              | icmpi      0 | icmpo     16 |
NET | eth0      0% | pcki    5290 | pcko     218 | si  460 Kbps | so   12 Kbps | erri       0 | erro       0 | drpi       0 | drpo       0 |
NET | eth1      0% | pcki      23 | pcko     842 | si    1 Kbps | so  139 Kbps | erri       0 | erro       0 | drpi       0 | drpo       0 |
NET | eth2      0% | pcki      22 | pcko      20 | si    1 Kbps | so    1 Kbps | erri       0 | erro       0 | drpi       0 | drpo       0 |
NET | lo      ---- | pcki    1958 | pcko    1958 | si  153 Kbps | so  153 Kbps | erri       0 | erro       0 | drpi       0 | drpo       0 |

  PID RUID       EUID        THR  SYSCPU  USRCPU  VGROW  RGROW  RDDSK  WRDSK ST EXC S  CPUNR  CPU CMD         1/2
 3527 root       root          6   0.20s   0.08s     0K     0K     0K   124K --   - S      5   3% rsyslogd
 3256 named      named        11   0.15s   0.10s     0K     0K     0K     0K --   - S      0   3% named
 3776 ganglia    ganglia       2   0.12s   0.07s     0K     0K     0K     0K --   - S      7   2% gmond
```

Benchmarking NFS performance

- Use the Unix application "iozone" to measure
  - Sequential read/writes, rewrites and rereads
  - Random read/writes, rewrites and rereads
  - Others
    - Backwards read, etc.
  - Direct output to native spreadsheet format for plotting

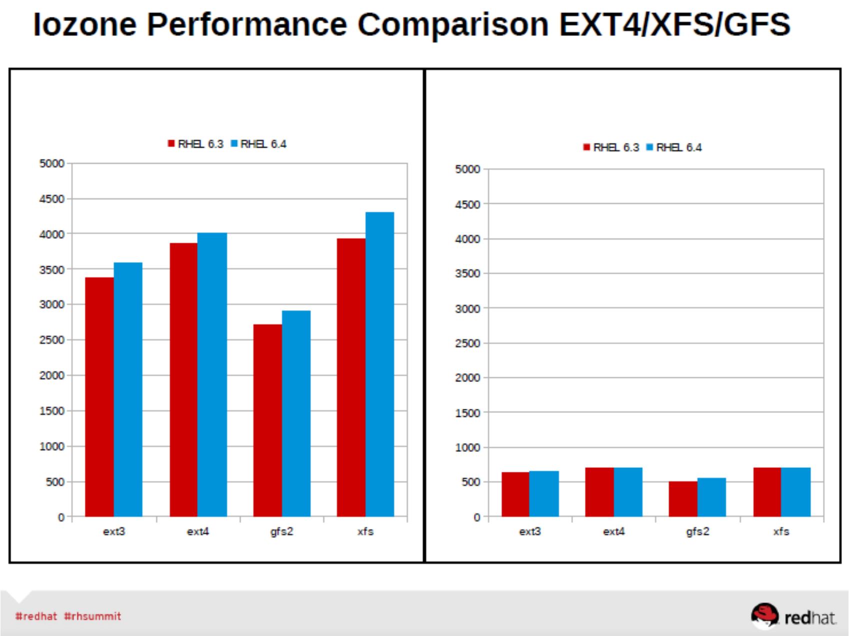


Image from https://rhsummit.files.wordpress.com/2013/06/shak-jeder-summit-perf-analysis-and-tuning-part-2-2013.pdf

Iozone basic operation

- Basic syntax for NFS mount-point (/mnt/foo)
  - iozone –azcR -U /mnt/foo –f /mnt/foo/testfile –b exceloutput.xls > logfile

- *Important Note: Understanding benchmarking data is better when comparing with prior baseline measurements.*

Mitigating CPU effects in iozone

- Options
  - Processor affinity of all threads/processes (“-P # ”)
  - Limit on the least number of processes (“-l n ”)
  - Processor cache purges (“-p ”)
  - CPU cache size (“-S size ”)
  - CPU cache line size (“-L size ”)
  - Thread synchronization (“-x ”)

File-system access modes

- Options
  - Forcing synchronous file access or enforcing the O_SYNC option when creating/opening files (“-o ”)
  - Files locking used for all IO requests (“-W ”)
  - Memory mapped file access (either synchronous or asynchronous mode) used in the client (“-B ”)
  - POSIX Asynchronous IO used in the client, which should not be confused with Asynchronous IO implemented on some Servers, such as Linux and HP-UX.)) (“-H n ” or “-k n ”)
  - Include fsync and fflush calls in the timings (“-e ”)

Concurrent access by many nodes

- iozone -t 2 -r 64 -s 1024 -+m filename
- Where the file "filename" lists the node-hostname, work-directory and iozone-binary:

```
Compute-0-0 /home/foo /usr/bin/iozone
Compute-0-1 /home/foo /usr/bin/iozone
```

- Use the -+d option to testing for data corruption

Iozone example output

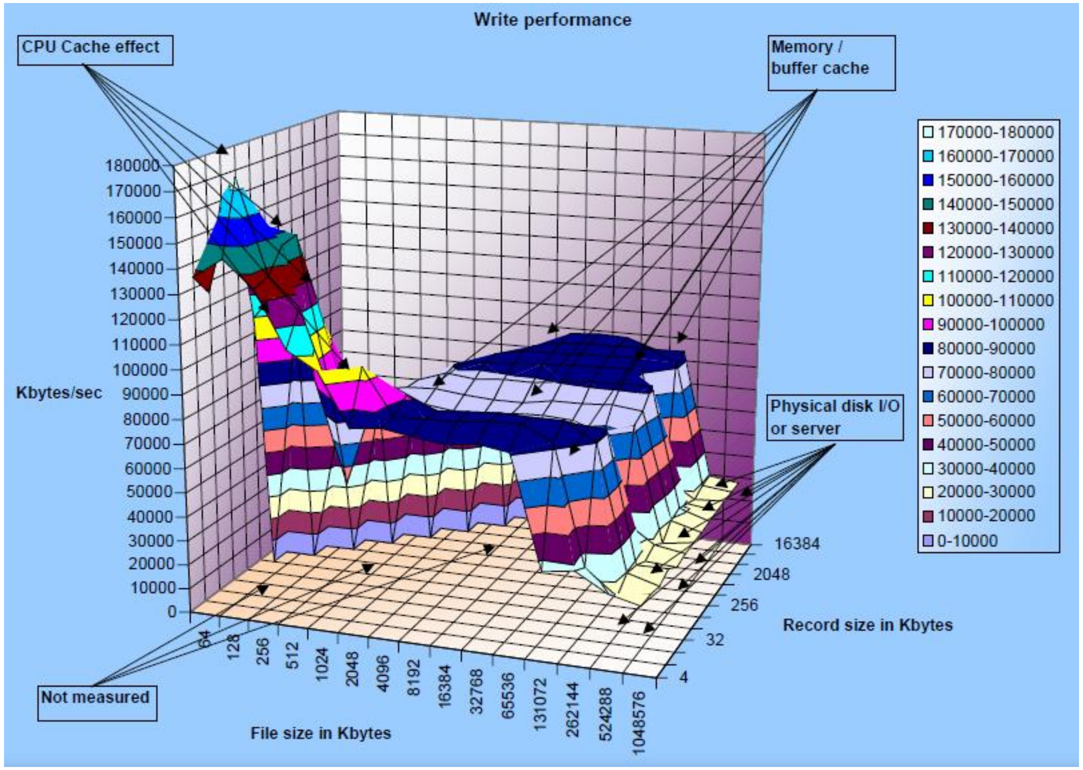


Image from http://www.iozone.org/docs/NFSClientPerf_revised.pdf

OPTIMIZING NFS

Visual inspection of NAS devices

- LED /Lights
  - On disks
  - Network ports (both computer & network device)
  - Power supply
- Damaged/broken cables
  - Broken heads, old cables
- High temperatures can also degrade the MTBF

Periodic on-line monitoring

- Monitoring
  - Use smartd for SMART monitoring of disks
    - periodic self testing of disks, predict disk faults and sends e-mail notifications
  - Use Ganglia or NAGIOS
    - Monitor hardware, status and occupancy/capacity
- Periodic benchmarking (iozone)
  - File-system on both server and client side.
    - Can show trends

Server side configuration

- Exports
  - Restrict writes only to certain clients can improve performance
    - /etc/exports
      - None, read-only, read-write (with wildcards)
- Use NFS v3 when enhanced security checks are not needed
  - NFS v4 requires Keberos.
- When using quota control
  - Use server side tools to set and manage quota

Client side configuration

- Switching to on-demand mount can help performance
  - Automounter
    - auto.master
    - auto.home
- Additional performance tuning
  - mount options
    - rsize and wsize

Server performance tuning

- Vertical scaling *(bigger single server)*
  - More or faster RAM (and/or CPU)
  - More network connections
  - More NFS daemons
- Horizontal scaling *(more physical servers)*
  - Requires partitioning of data
  - Works best with automounter based client mounting
- Linux Kernel parameters (only if power is protected)
  - vm.dirty_background_ratio
  - vm.dirty_ratio

Summary

- Optimizing NFS storage solution for your HPC clusters is important for performance.
- Both preventive maintenance and periodic benchmarking can help to detect and cure NFS related performance problems.

Thank you