



Optimisation of Linux System Settings for Access to the JASMIN/CEMS Panasas Storage

C Del Cano Novales

September 2013

©2013 Science and Technology Facilities Council



This work is licensed under a [Creative Commons Attribution 3.0 Unported License](https://creativecommons.org/licenses/by/3.0/).

Enquiries concerning this report should be addressed to:


RAL Library
STFC Rutherford Appleton Laboratory
Harwell Oxford
Didcot
OX11 0QX

Tel: +44(0)1235 445384
Fax: +44(0)1235 446403
email: libraryral@stfc.ac.uk

Science and Technology Facilities Council reports are available online at: <http://epubs.stfc.ac.uk>

ISSN 1358-6254

Neither the Council nor the Laboratory accept any responsibility for loss or damage arising from the use of information contained in any of their reports or in any communication about their tests or investigations.



Optimisation of Linux System Settings for Access to the JASMIN/CEMS Panasas Storage

Del Cano Novales, Cristina (STFC,RAL,SC)
8/19/2013

Contents

Introduction and Motivation	3
Available parameters	5
Panasas mount options.....	5
Kernel settings	6
Ethernet flow control.....	7
Test environment.....	8
Panasas test environment	8
Test hosts	8
Current settings.....	8
Other settings	8
Test scripts	8
Tests	9
Other considerations	10
Virtual Machine.....	11
Reading	11
Test 1A:	12
Test 1B:.....	13
Test 1C:.....	14
Writing	15
Test 2A:	16
Test 2B:.....	17
Test 2C:.....	18
Test 2D:	19
Directory Listing	20
Virtual Machines – Optimal Settings	21
Comparison between current settings and settings obtained	22
Reading - Test Realm.....	22
Reading – Production Realm.....	22
Writing – Test Realm.....	23
Writing – Production Realm.....	23

Physical Host	24
Reading	24
Test 1A:	25
Test 1B:.....	26
Writing	27
Test 2A:	28
Test 2B:.....	29
Test 2C:.....	30
Directory Listing	31
Physical Hosts – Optimal Settings	32
Comparison between current settings and settings obtained	33
Reading - Test Realm.....	33
Reading – Production Realm.....	33
Writing – Test Realm.....	34
Writing – Production Realm.....	34
Conclusions and future work	35

Introduction and Motivation

The JASMIN/CEMS super-data-cluster provides fast, parallel storage and a computing environment for UK and European climate and earth observation communities.

Deployed at the Rutherford Appleton Laboratory as a joint project between RAL Space and the Scientific Computing Department, the system comprises a 4.6PB usable parallel storage, collocated with the Lotus HPC compute facility and a virtualisation infrastructure.

The JASMIN/CEMS cluster provides the infrastructure for data storage and services for the Centre for Environmental Data Archival, as well as data processing and scientific computation for the climate, earth system science and earth observation communities, including the ISIC/Satellite Applications Catapult funded by the UK Space Agency.

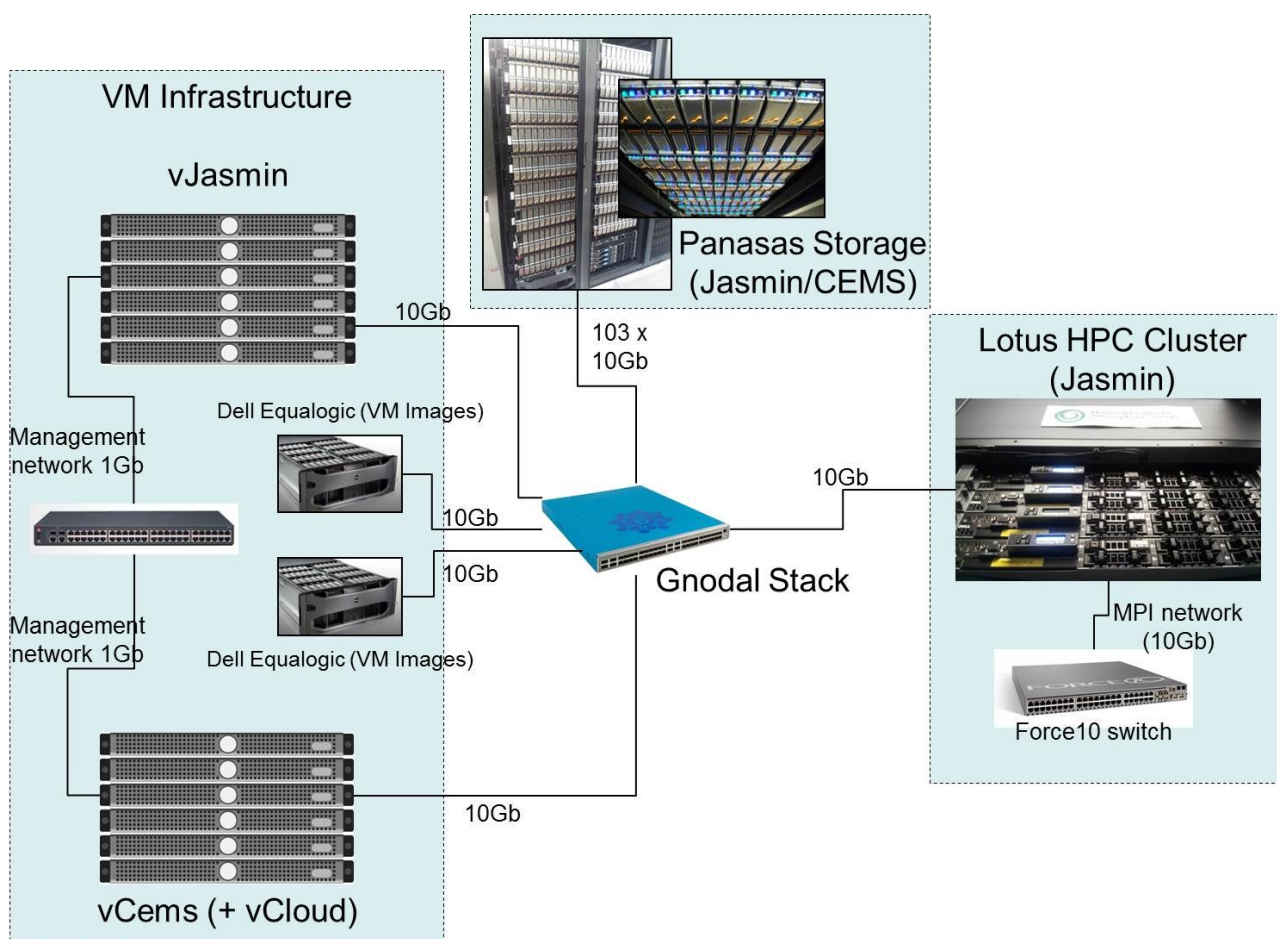


Figure 1: JASMIN/CEMS Architecture

Panasas is a very high-performance system which requires a high quality network for optimal performance. Network components and brands behave differently so Panasas recommends spending some time finding the optimal settings for best performance.

This report describes the investigation into the optimal system settings for increasing the storage access performance, both from physical hosts and from virtual machines.

Available parameters

This report focuses on three different types of parameters; Panasas mount options, kernel settings and Ethernet flow control.

Panasas mount options

- **iscsi-sock-rsize**: iSCSI receive socket buffer size (bytes). Panasas recommends a minimum value of 20480 for 10Gb networks.
- **iscsi-sock-wsize**: iSCSI send socket buffer size (bytes). This setting has an impact in write performance. It controls the amount of data that a StorageBlade can send to a client without an acknowledgement. Panasas recommends a value of 65536 for a 10Gb network.
- **max-async-writepages**: When the DirectFlow client is asked to sync out a file, the resulting writepages() can be processed asynchronously if less than n writepages() are already inflight. This option is used for tuning write bandwidth. The default value is 0 or disabled.
- **statahead-level**: The speed of directory lookups is improved by launching, in the background, requests for many directory entries at a time (a statahead). By default, a single page of directory entries is fetched. Typically one 4K page contains around 30 entries. This setting allows the number of stataheads issued in the background to be increased. Only DirectFlow clients that frequently walks a huge tree, for example rsync to backup files, are recommended to increase this number.

The available policies for statahead-level are as follows:

Policy	Statahead (assume page size is 4K)
0	OFF
1	1 directory page (default. around 30 entries)
2	2 directory page (around 60 entries)
3	3 directory page (around 90 entries)
4	4 directory page (around 120 entries)

Kernel settings

There are two groups of kernel settings. These are configured in `/etc/sysctl.conf` and will be enabled/disabled on the different tests.

The first set of settings has been recommended by Panasas.

The second set of settings is recommended for any hosts with 10Gb network interfaces and high latency networks, so the effect of these settings might be limited.¹

- Panasas driver recommended settings:

`vm.min_free_kbytes`: Amount of memory that the kernel should try to keep free at all times.

`kernel.shmmax`: Maximum size of shared memory segment (bytes).

`kernel.sem`: System semaphore parameters.

`net.ipv4.tcp_moderate_rcvbuf`: Autotuning parameter for the receiver buffer and TCP window size.

```
# Panasas driver recommended settings
vm.min_free_kbytes = 16384
kernel.shmmax = 2147483648
kernel.sem = 250 32000 100 128
net.ipv4.tcp_moderate_rcvbuf = 0
```

- 10Gb NIC recommended settings:

`net.core.rmem_max` and `net.core.wmem_max`: TCP max buffers size.

`net.ipv4.tcp_rmem` and `net.ipv4.tcp_wmem`: TCP buffer limits. The three values indicate the minimum, initial and maximum buffer size.

`net.core.netdev_max_backlog`: Length of the processor input queue.

```
# 10Gb nic settings
net.core.rmem_max = 16777216
net.core.wmem_max = 16777216
net.ipv4.tcp_wmem = 4096 65536 16777216
net.ipv4.tcp_rmem = 4096 87380 16777216
net.core.netdev_max_backlog = 30000
```

¹ <http://fasterdata.es.net/host-tuning/linux>

Ethernet flow control

Ethernet flow control allows a receiver to request the sender a temporary pause in the transmission to allow for processing of inbound data. A sender that receives a “pause” frame will stop transmitting for a short time.

Enabling flow control should result in a reduction in latency and an increase in transmission speed.

The Panasas shelves are flow control enabled, allowing them to both receive and transmit “pause” frames. Panasas recommends that flow control is enabled in all attached network equipment.

Ethernet flow control is not configurable on Virtual Machines when using the VMXNET3 VMware virtual network interface (default network interface for Red Hat Enterprise Linux 6).

Test environment

Panasas test environment

Testing is done in the Panasas test realm, currently comprising 9 shelves, three ActiveStor11 and six ActiveStor14.

The volume used for testing is /benchmarking/test, located on BladeSet 3 (Shelf 3). This is an ActiveStor11 shelf, which matches the current ActiveStor version in the JASMIN/CEMS production realm.

The Panasas test realm is currently running PanFS 5.0.0.c version.

Test hosts

Physical host: host004.jc.rl.ac.uk

96GB RAM

12 cores: Intel(R) Xeon(R) CPU X5675@3.07GHz

10Gb network connection to the Panasas storage through a Gnodal GS4008 10GbE Switch

Virtual machine: test01.jc.rl.ac.uk

2GB RAM

2 cores

10Gb network connection from the hosting Hypervisor. The connection is shared between virtual machines currently located in the physical host.

Current settings

- iscsi-sock-rsize: 16384
- iscsi-sock-wsize: 65536
- max-async-writepages: Not enabled
- statahead-level: Not enabled
- Kernel settings: Enabled
- Ethernet flow control: Enabled on physical hosts

Other settings

- Direct flow client: The client version used for testing is panfs-4.1.3-702504.9.x86_64.

Test scripts

- dd: The dd command is used by reading from the special device file /dev/zero and writing into the test volume. The size of the read/write block is 64k, and the file size is set on the different tests.
- ls: The ls command, combined with the time command is used to calculate the speed of directory listing.

Tests

All tests are run both in a physical host and in a virtual machine.

Each test is run with all the different combinations of the kernel settings and flow control, giving four test runs per test for virtual machines (flow control is not configurable) and eight test runs per test for the physical hosts.

- Read
 - iscsi-sock-rsize
- Write
 - iscsi-sock-wsize
 - max-async-writepages
- Directory listing
 - statahead-level

Other considerations

- The tests for this investigation have been run in the production network. The load of the network at the time of running the test cannot be controlled and has a large impact on the results, therefore this investigation is more interested in patterns, rather than specific numbers. Also for the same reason, tests are run several times and results averaged.
- Virtual machines reside on a physical host, in most cases sharing resources with other virtual machines. The test results therefore will depend on the load and traffic of the hosting physical machine at the time of the test.
- The test volume used is located on a bladeset comprising one shelf; production volumes and bladesets have different configurations. However, the results of this investigation should still be relevant for the production system.

Data is stored across a number of storage blades, depending on file size and configuration. Each shelf has 10 or 11 storage blades. The speed of accessing a file stored in less than 10/11 blades will be limited to 1Gb per blade, independently of the location of the data (Fig. 2 and 3). If the data is spread across more than 11 blades, the bandwidth will be limited to 10Gb by the single client.

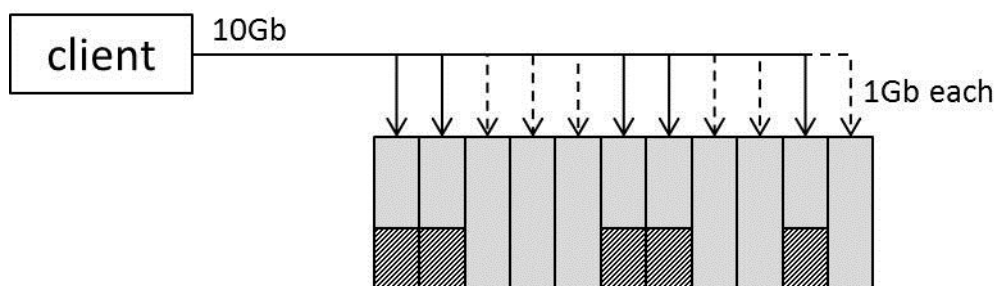


Figure 2: Volume on single shelf

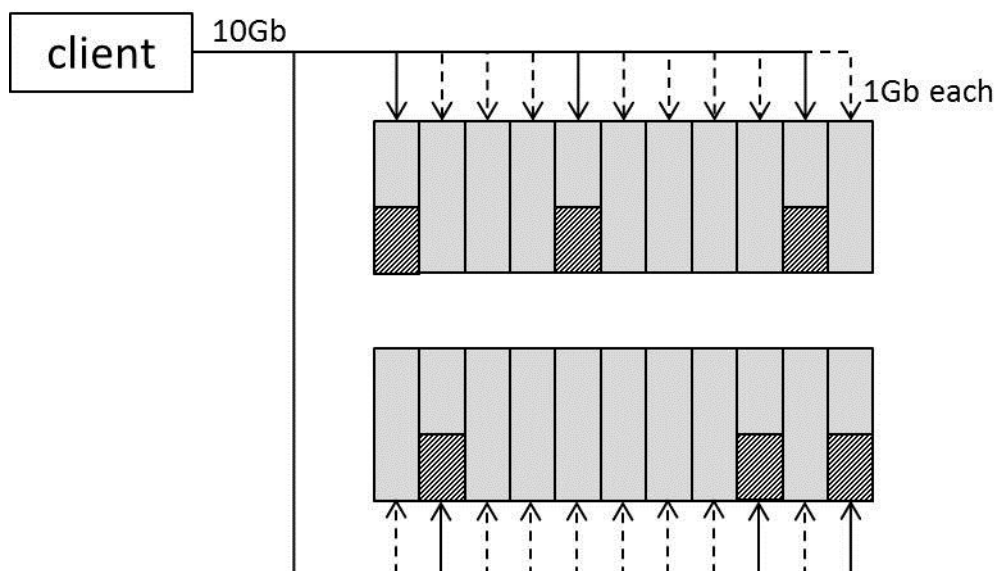


Figure 3: Volume on multiple shelves

Virtual Machine

Reading

The purpose of these tests is to calculate the optimal value for the `iscsi-sock-rsize` mount option and to compare the effects of the kernel settings on the average reading speed.

A `dd` read test is run over a range of `iscsi-sock-rsize` values, based on default values for the other settings. Every test will be run four times with the different combination of kernel settings.

Common Settings:

- `iscsi-sock-wsize` = 65536, recommended value by Panasas
- `max-async-writepages` not set
- `statahead-level` not set
- kernel settings: enabled or disabled
 - Run 1: Panasas settings enabled / 10Gb settings enabled
 - Run 2: Panasas settings disabled / 10Gb settings enabled
 - Run 3: Panasas settings enabled / 10Gb settings disabled
 - Run 4: Panasas settings disabled / 10Gb settings disabled
- flow control: Not configurable on a virtual machine

Test 1A:

Settings:

- iscsi-sock-rsize: from 20480 to 98304
- file size: 1.1 Gb => testfile1

Command run:

```
# dd if=/benchmarking/test/testfile1 of=/dev/null bs=64k  
# /usr/local/sbin/panfs_trace -j /benchmarking/test/testfile1
```

Results:

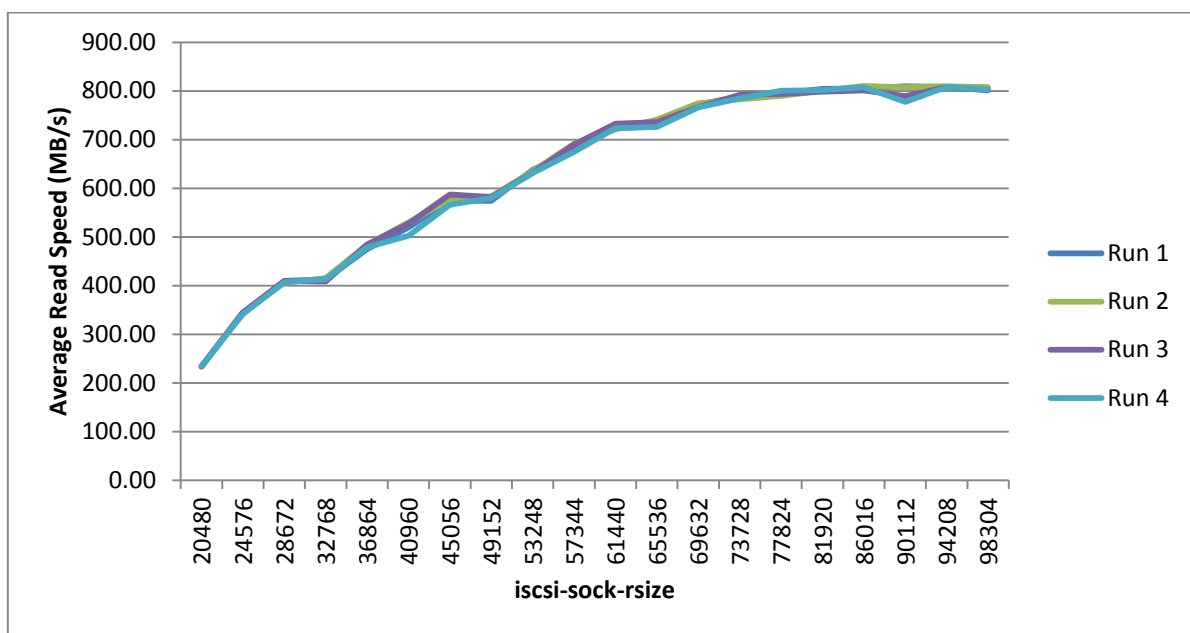


Figure 4: Test 1A

Conclusions:

- Average read speed clearly increases with the iscsi-sock-rsize, peaking at around 86016. Any setting over this value seems to give no benefit to the average reading speed.
- Kernel settings have no effect on the speed on this test.

Test 1B:

Based on the results of Test 1A, the range between the iscsi-sock-rsize values of 20480 and 69632 can be discarded as the average speed is much lower than for higher values.

A larger file size is chosen for this test to accentuate the differences on speed between the kernel settings combinations.

Settings:

- iscsi-sock-rsize: from 69632 to 98304
- file size: 11 Gb => testfile2

Command run:

```
# dd if=/benchmarking/test/testfile2 of=/dev/null bs=64k  
# /usr/local/sbin/panfs_trace -j /benchmarking/test/testfile2
```

Results:

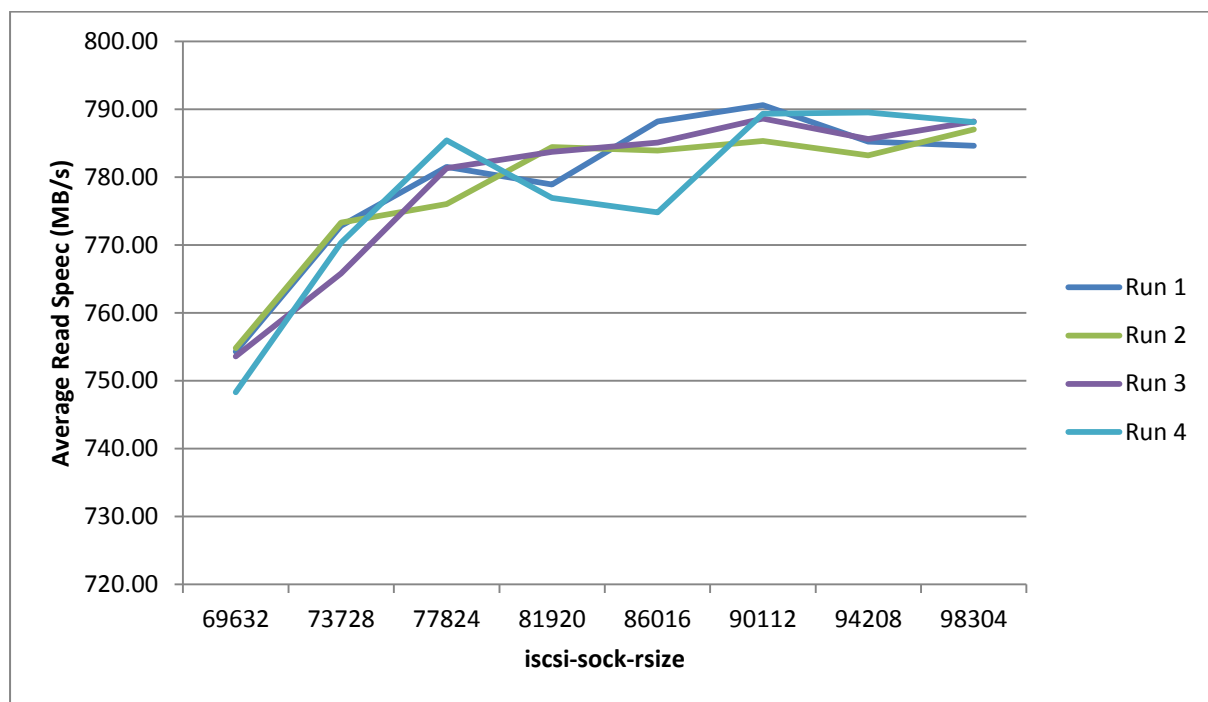


Figure 5: Test 1B

Conclusions:

- Optimal values for iscsi-sock-rsize on the 81920 – 98304 range.
- No obvious pattern on results for different kernel settings.

Test 1C:

Same test as 1B but with a very large file.

Settings:

- iscsi-sock-rsize: from 69632 to 98304
- file size: 111 Gb => testfile3

Command run:

```
# dd if=/benchmarking/test/testfile3 of=/dev/null bs=64k  
# /usr/local/sbin/panfs_trace -j /benchmarking/test/testfile3
```

Results:

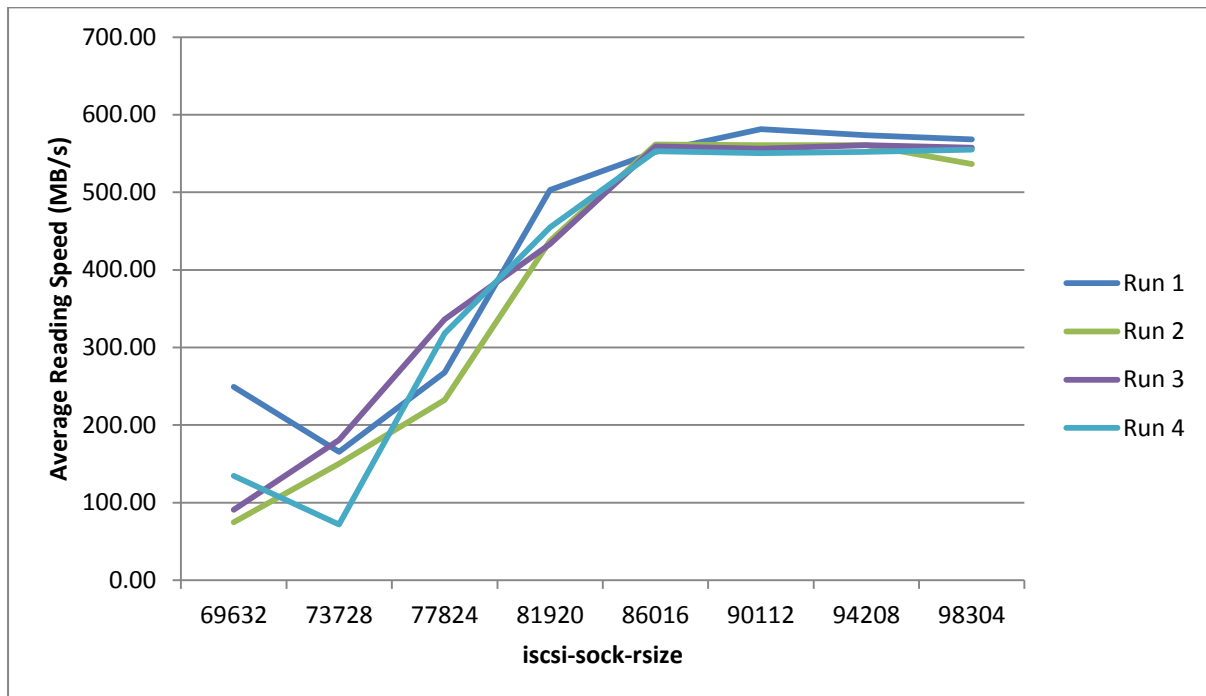


Figure 6: Test 1C

Conclusions:

- Optimal iscsi-sock-rsize = 90112
- Panasas and 10Gb kernel settings doesn't seem to affect average reading speed.

Writing

The following set of tests investigates the optimal value for the `iscsi-sock-wsize` mount option, the best value for the `max-async-writepages` setting, and compares the effects of the kernel settings on the average writing speed.

Common Settings:

- `iscsi-sock-rsize` = 90112, obtained from Read tests in previous section.
- `statahead-level` not set
- kernel settings: enabled or disabled
 - Run 1: Panasas settings enabled / 10Gb settings enabled
 - Run 2: Panasas settings disabled / 10Gb settings enabled
 - Run 3: Panasas settings enabled / 10Gb settings disabled
 - Run 4: Panasas settings disabled / 10Gb settings disabled
- flow control: Not configurable on a virtual machine

Test 2A:

Settings:

- iscsi-sock-wsize: from 20480 to 98304
- max-async-writepages not set
- filesize: 1.1GB => testfile1

Command run:

```
# dd if=/dev/zero of=/benchmarking/test/testfile1 bs=64k count=16384  
# /usr/local/sbin/panfs_trace -j /benchmarking/test/testfile1
```

Results:

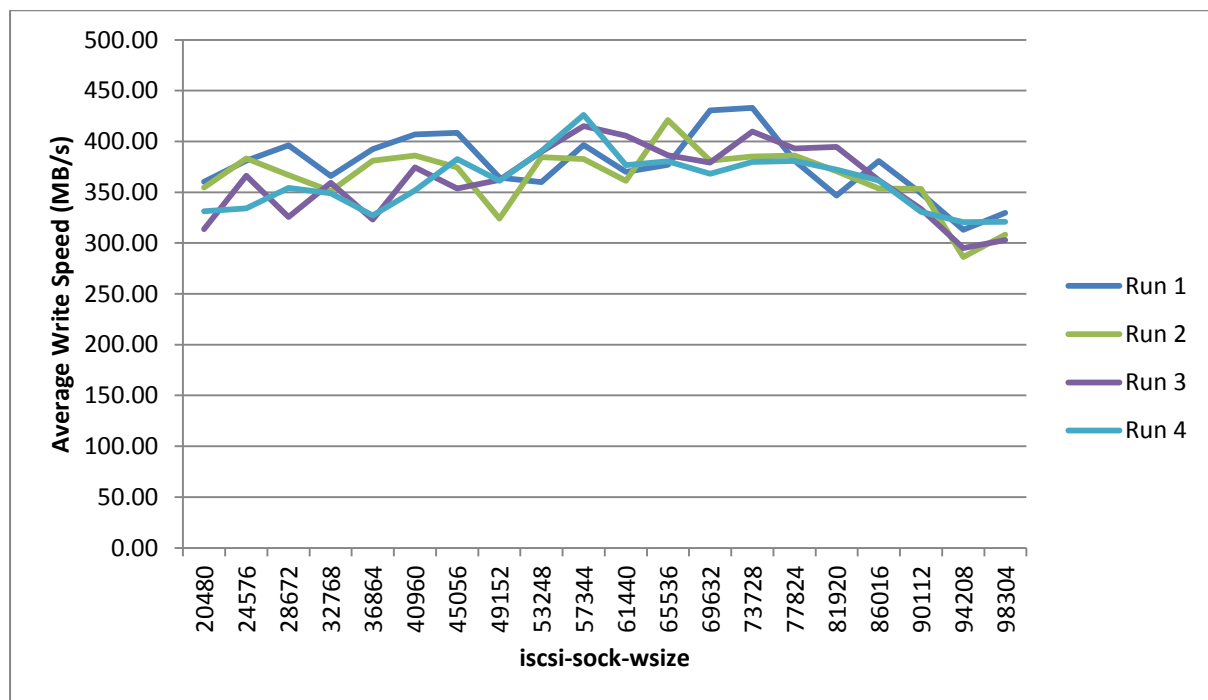


Figure 7: Test 2A

Conclusions:

- Although the graph is quite stable, there is slight increase in speed between iscsi-sock-wsize values of 53248 and 77824.
- It seems that kernel settings do now have an obvious impact on write speed.

Test 2B:

Based on the results from Test 2A, this test focuses on the range between 53248 and 77824 for iscsi-sock-wsize, using a larger file.

Settings:

- iscsi-sock-wsize: from 53248 to 77824
- max-async-writepages not set
- filesize: 11GB => testfile2

Command run:

```
# dd if=/dev/zero of=/benchmarking/test/testfile2 bs=64k count=163840  
# /usr/local/sbin/panfs_trace -j /benchmarking/test/testfile2
```

Results:

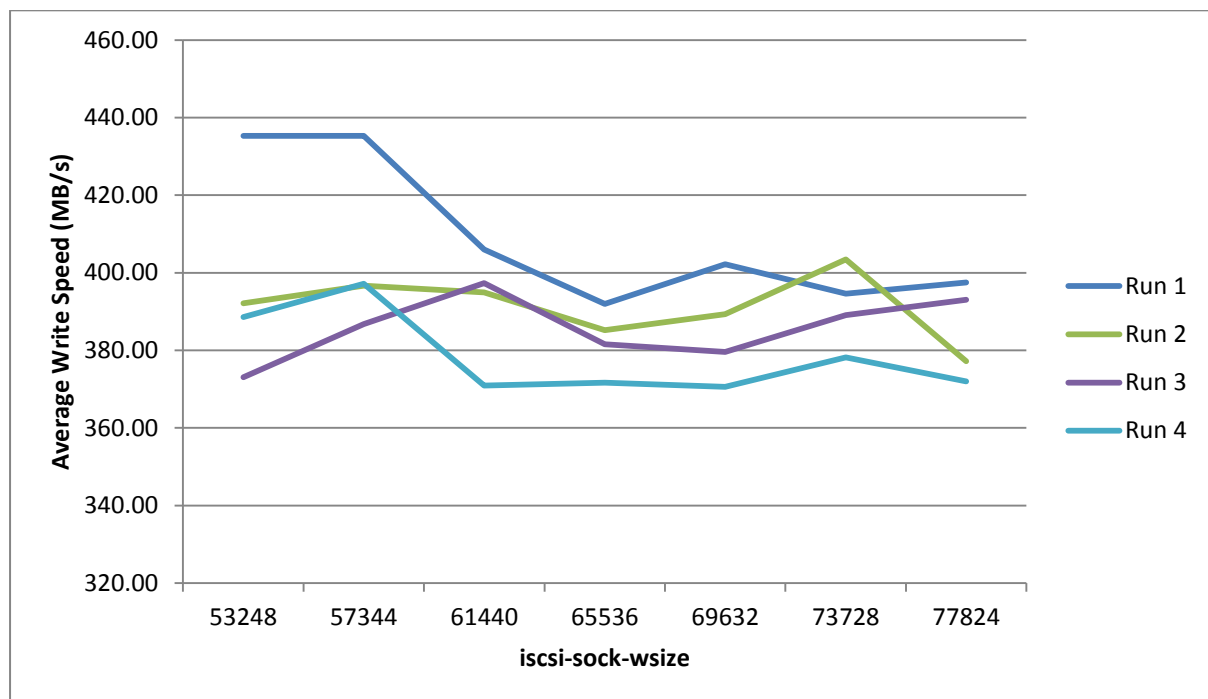


Figure 8: Test 2B

Conclusions:

- There is no obvious benefit in using any other value of iscsi-sock-wsize different to the recommended by Panasas (65536)

Test 2C:

Settings:

- iscsi-sock-wsize: from 20480 to 98304
- max-async-writepages not set
- filesize: 111GB => testfile3

Command run:

```
# dd if=/dev/zero of=/benchmarking/test/testfile3 bs=64k count=1638400  
# /usr/local/sbin/panfs_trace -j /benchmarking/test/testfile3
```

Results:

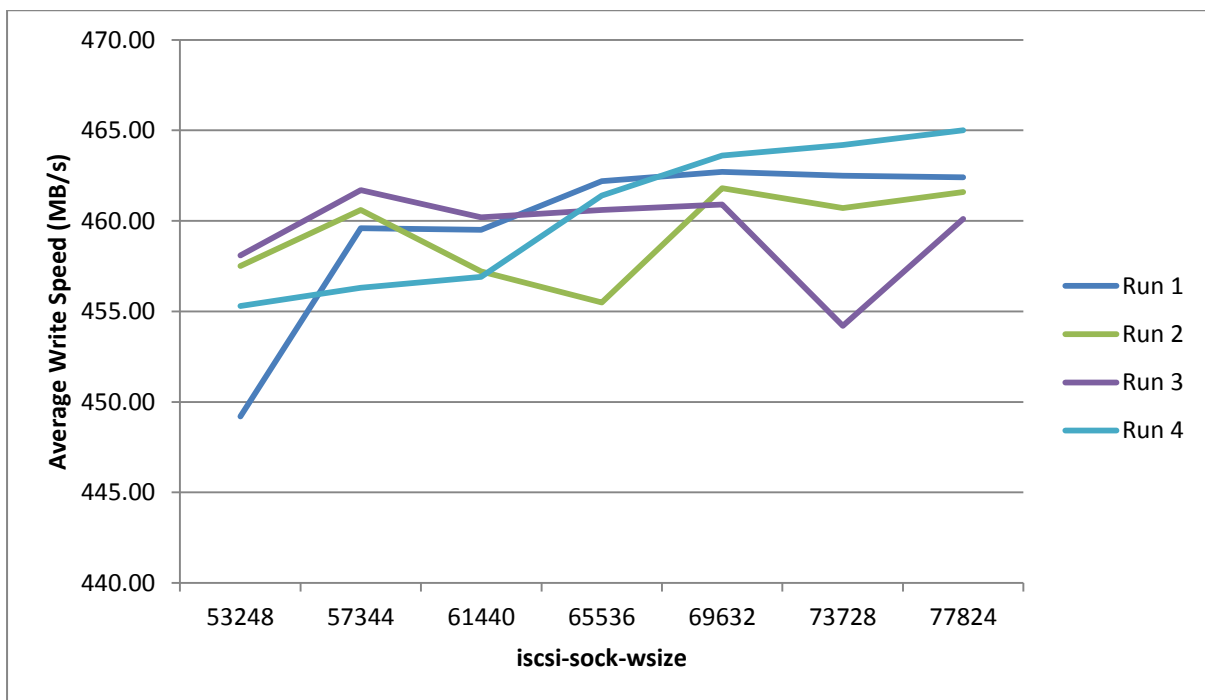


Figure 9: Test 2C

Conclusions:

- As seen in previous tests, the impact of increasing iscsi-sock-wsize is not relevant.
- The value chosen for iscsi-sock-wsize is then 65536.

Test 2D:

The purpose of this test is to find the optimal value for the max-async-writepages mount option based on the iscsi-sock-rsize and iscsi-sock-wsize obtained in previous tests.

Settings:

- iscsi-sock-wsize: 65536
- max-async-writepages: from 0 to 32
- filesize: 11GB => testfile2

Command run:

```
# dd if=/dev/zero of=/benchmarking/test/testfile2 bs=64k count=163840  
# /usr/local/sbin/panfs_trace -j /benchmarking/test/testfile2
```

Results:

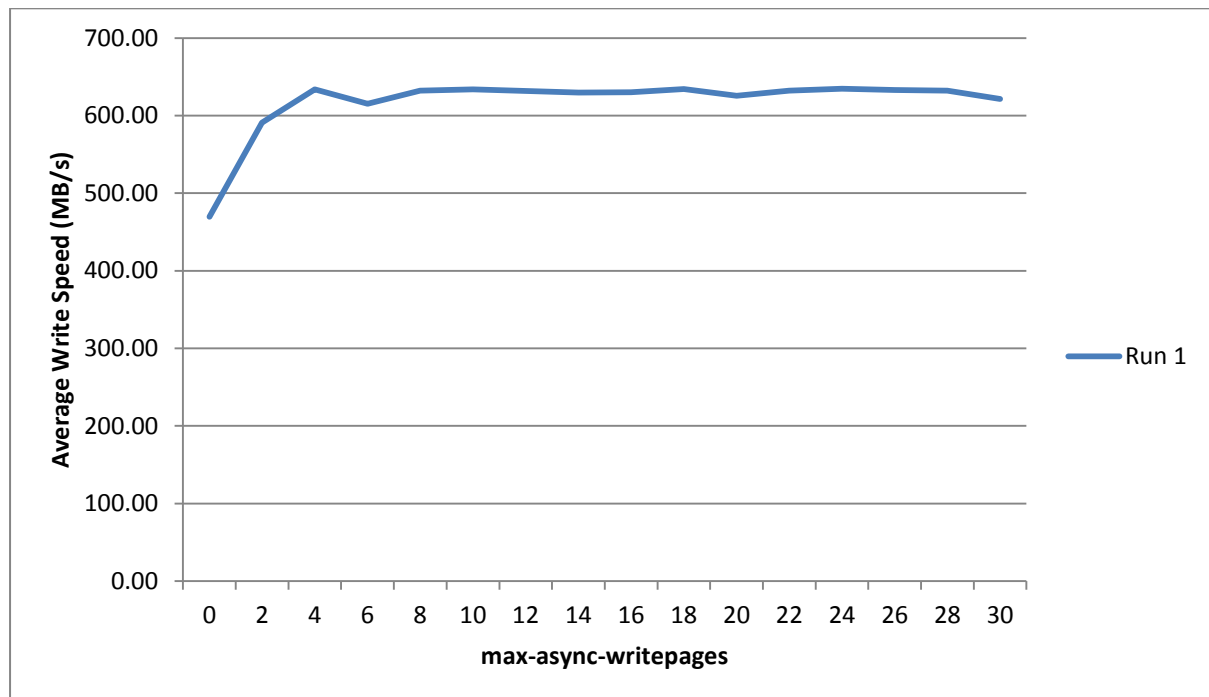


Figure 10: Test 2D

Conclusions:

- There is a very significant increase in average writing speed when enabling the max-async-writepages setting.
- Average speed flattens at around max-async-writepages=8. The choice of value is then 10.

Directory Listing

The aim of this test is to investigate the effect of the statahead-level mount option in the time to run a directory listing (with ls) on a large directory.

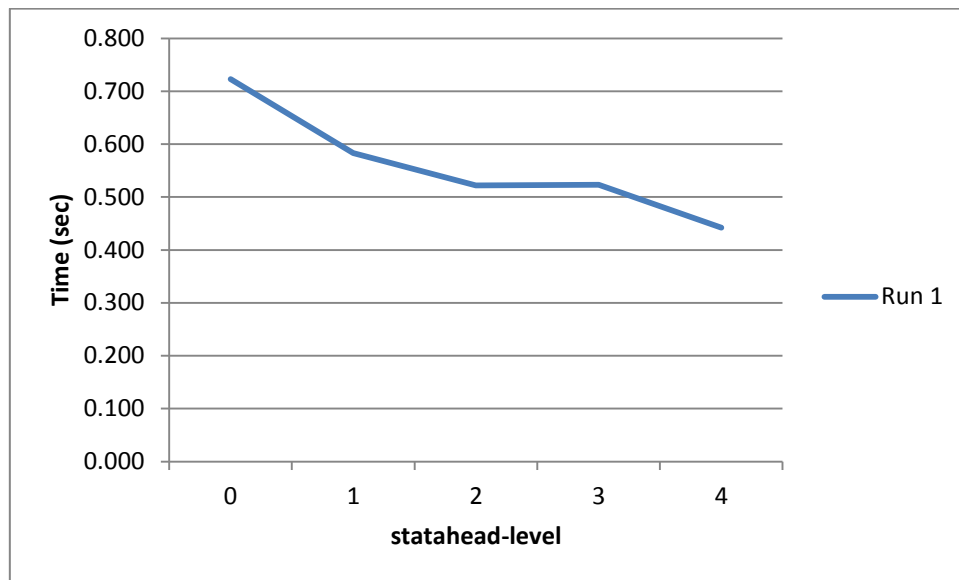


Figure 11: Directory Listing

Conclusions:

- There is a decrease in the time taken to list a directory when using the statahead-level mount option.
- The optimal value for statahead-level seems to be 4.

Virtual Machines – Optimal Settings

Mount Options	
iscsi-sock-rsize	90112
iscsi-sock-wsize	65536
statahead-level	4
max-async-writepages	10
Kernel Settings	No impact
Ethernet Flow Control	Not available

Comparison between current settings and settings obtained

Reading - Test Realm

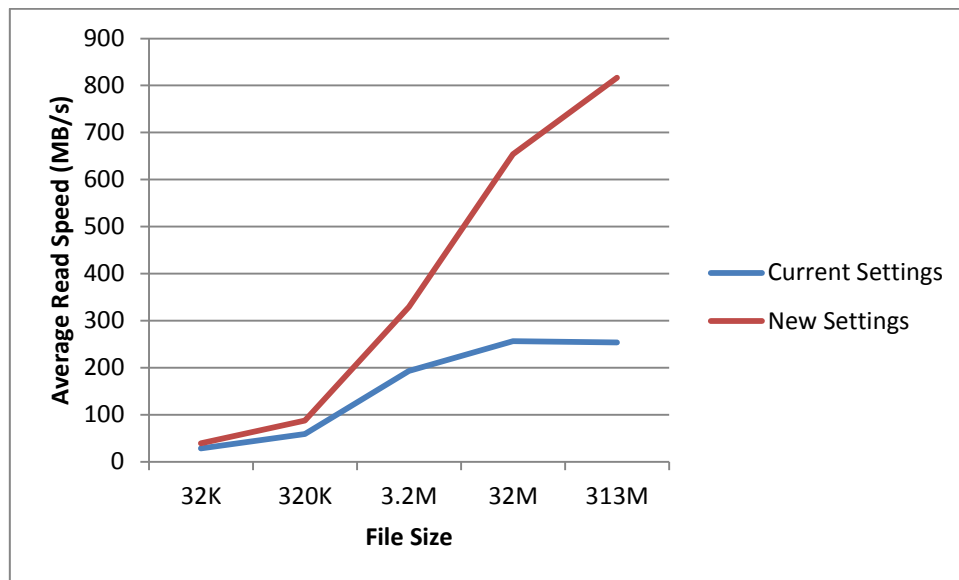


Figure 12: Reading - test realm

Reading - Production Realm

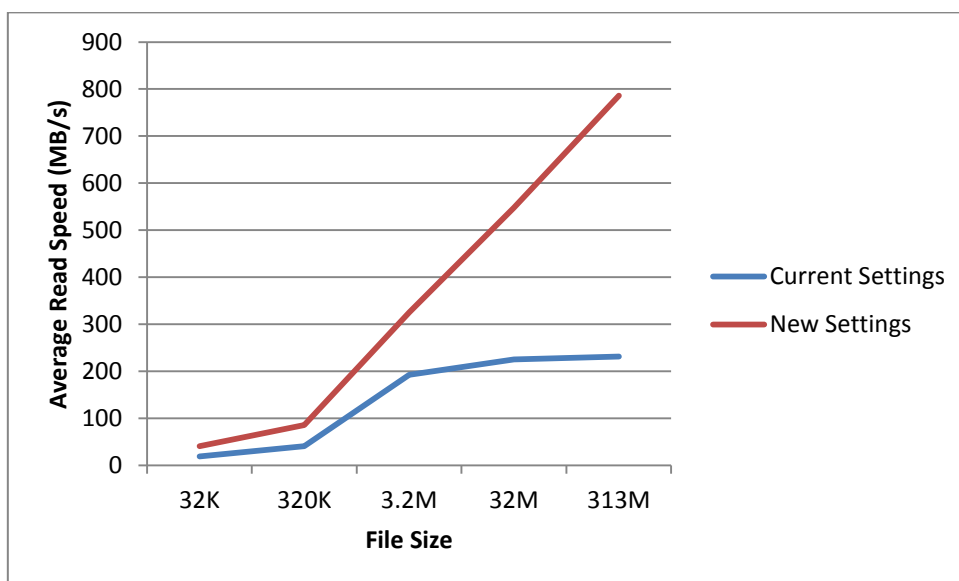


Figure 13: Reading - production realm

Writing - Test Realm

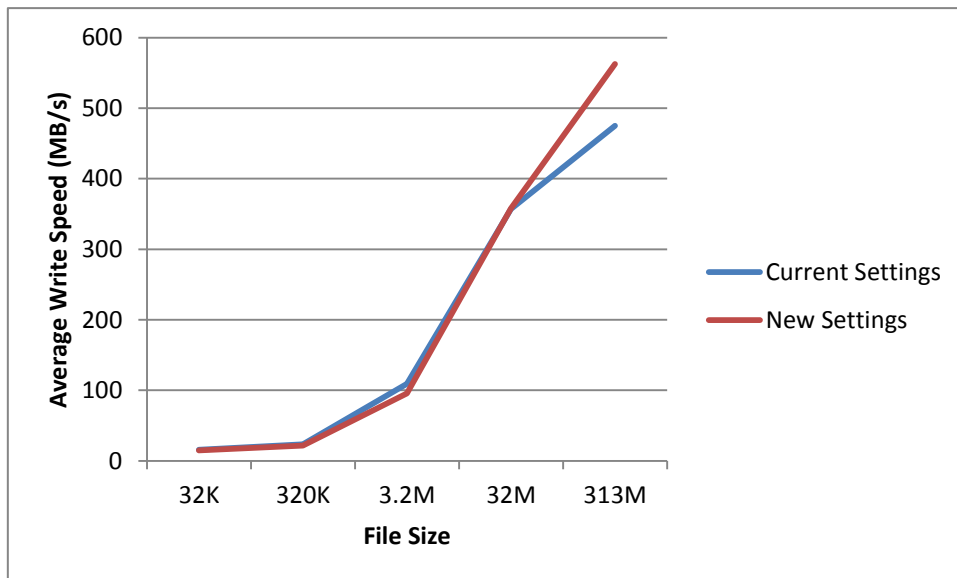


Figure 14: Writing - test realm

Writing - Production Realm

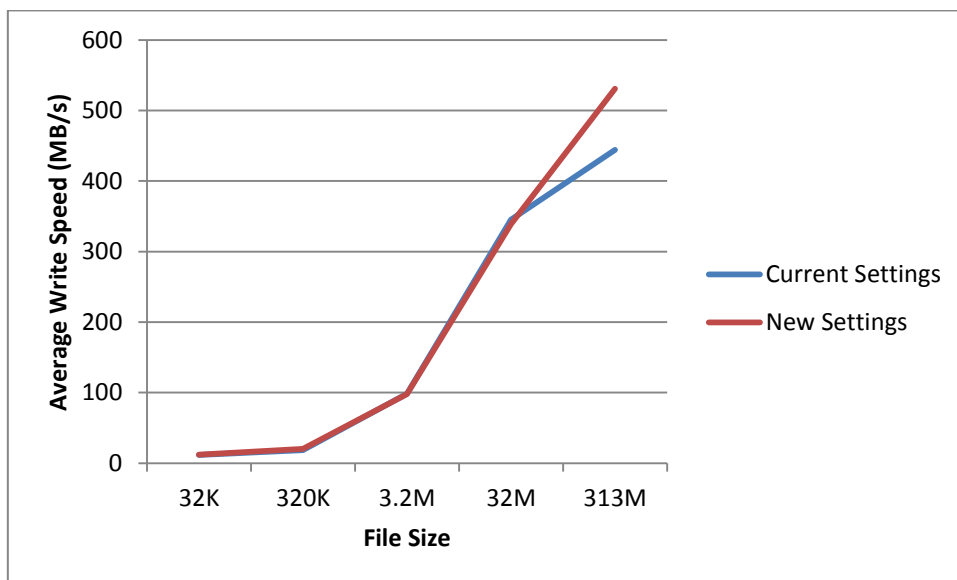


Figure 15: Writing - production realm

Physical Host

Reading

The purpose of these tests is to calculate the optimal value for the `iscsi-sock-rsize` mount option and to compare the effects of the kernel settings on the average reading speed.

A `dd` read test is run over a range of `iscsi-sock-rsize` values, based on default values for the other settings. Every test will be run eight times with the different combination of kernel settings and Ethernet flow control.

Common Settings:

- `iscsi-sock-wsize` = 65536, recommended value by Panasas
- `max-async-writepages` not set
- `statahead-level` not set
- kernel settings and flow control: enabled or disabled
 - Run 1: Panasas settings enabled / 10Gb settings enabled/Flow Control enabled
 - Run 2: Panasas settings enabled / 10Gb settings enabled/Flow Control disabled
 - Run 3: Panasas settings disabled / 10Gb settings enabled/Flow Control enabled
 - Run 4: Panasas settings disabled / 10Gb settings enabled/Flow Control disabled
 - Run 5: Panasas settings enabled / 10Gb settings disabled/Flow Control enabled
 - Run 6: Panasas settings enabled / 10Gb settings disabled/Flow Control disabled
 - Run 7: Panasas settings disabled / 10Gb settings disabled/Flow Control enabled
 - Run 8: Panasas settings disabled / 10Gb settings disabled/Flow Control disabled

Test 1A:

Settings:

- iscsi-sock-rsize: from 20480 to 98304
- file size: 1.1 Gb => testfile1

Command run:

```
# dd if=/benchmarking/test/testfile1 of=/dev/null bs=64k  
# /usr/local/sbin/panfs_trace -j /benchmarking/test/testfile1
```

Results:

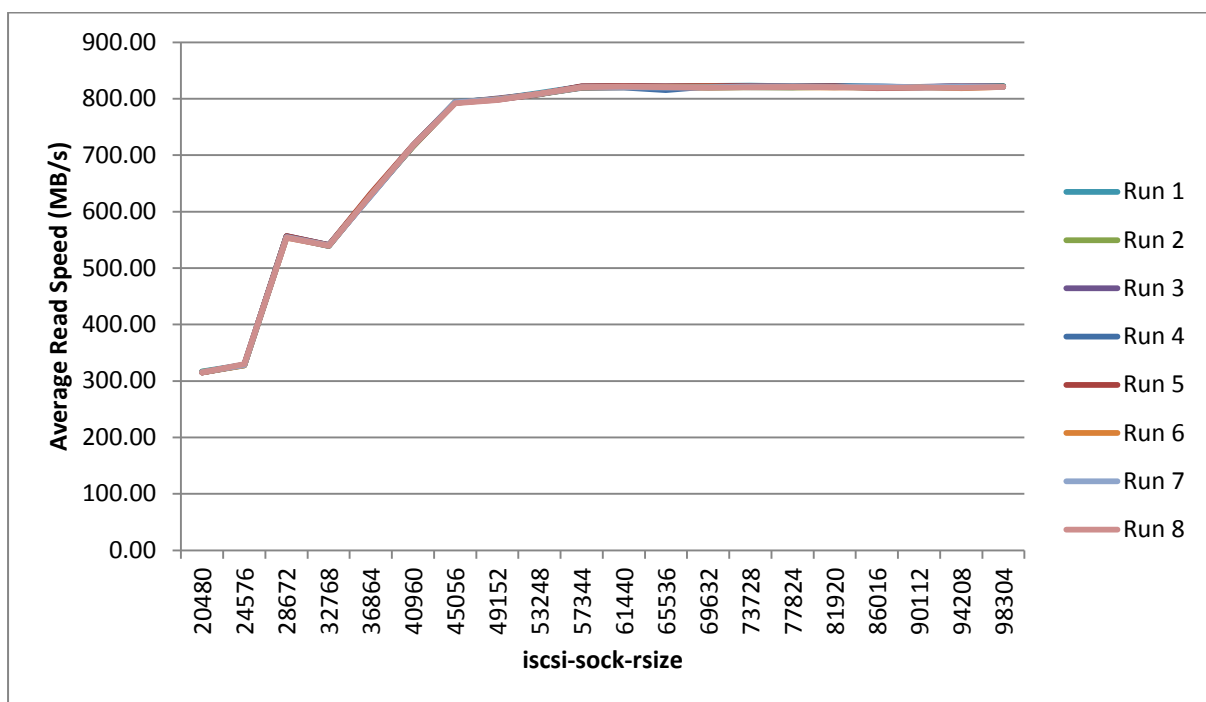


Figure 16: Test 1A

Conclusions:

- Average read speed increases with the iscsi-sock-rsize, peaking at around 57344.
- Kernel settings have no effect on the speed on this test.

Test 1B:

Based on the results of Test 1A, the range between the iscsi-sock-rsize values of 20480 and 49152 can be discarded as the average speed is lower than for higher values. In the same way, values higher than 81920 do not change with different iscsi-sock-rsize settings.

A larger file size is chosen for this test to accentuate the differences on speed between the kernel settings combinations.

Settings:

- iscsi-sock-rsize: from 53248 to 81920
- file size: 11 Gb => testfile2

Command run:

```
# dd if=/benchmarking/test/testfile2 of=/dev/null bs=64k
# /usr/local/sbin/panfs_trace -j /benchmarking/test/testfile2
```

Results:

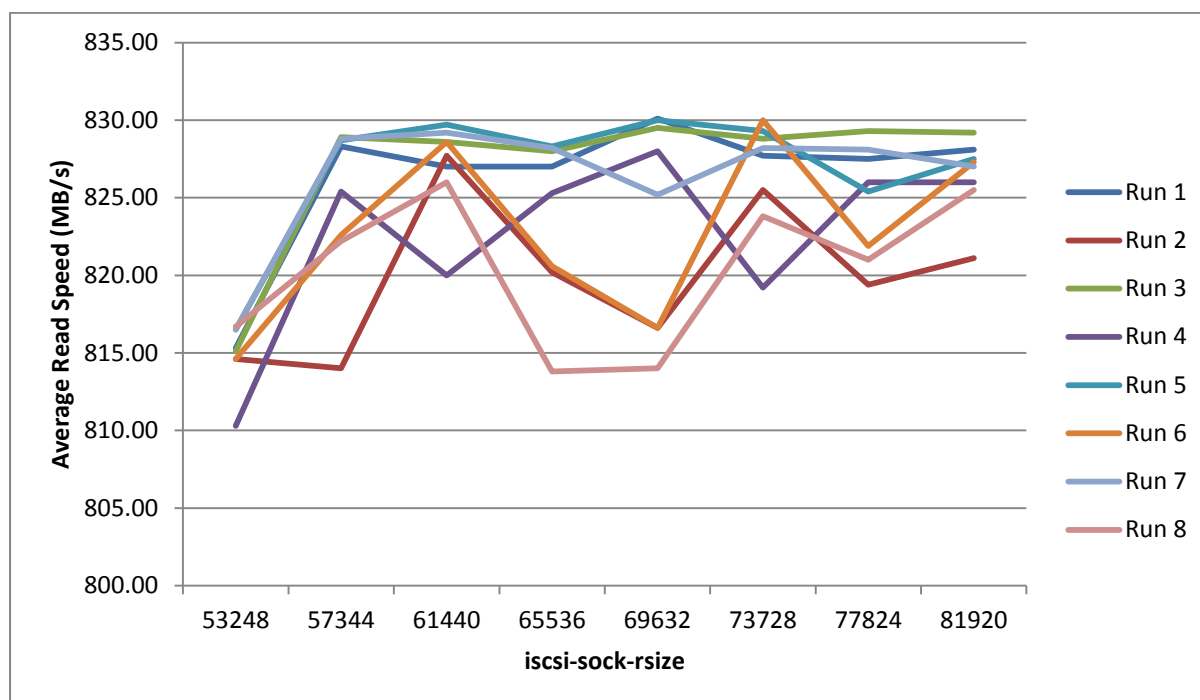


Figure 17: Test 1B

Conclusions:

- Optimal iscsi-sock-rsize = 57344
- Runs with Ethernet flow control disable seem to have a slight lower average speed than runs with flow control enabled.
- Panasas and 10Gb kernel settings doesn't seem to affect average reading speed.

Writing

The following set of tests investigates the optimal value for the `iscsi-sock-wsize` mount option, the best value for the `max-async-writepages` setting, and compares the effects of the kernel settings on the average writing speed.

Common Settings:

- `iscsi-sock-rsize` = 57344, obtained from Read tests in previous section.
- `statahead-level` not set
- kernel settings and flow control: enabled or disabled
 - Run 1: Panasas settings enabled / 10Gb settings enabled/Flow Control enabled
 - Run 2: Panasas settings enabled / 10Gb settings enabled/Flow Control disabled
 - Run 3: Panasas settings disabled / 10Gb settings enabled/Flow Control enabled
 - Run 4: Panasas settings disabled / 10Gb settings enabled/Flow Control disabled
 - Run 5: Panasas settings enabled / 10Gb settings disabled/Flow Control enabled
 - Run 6: Panasas settings enabled / 10Gb settings disabled/Flow Control disabled
 - Run 7: Panasas settings disabled / 10Gb settings disabled/Flow Control enabled
 - Run 8: Panasas settings disabled / 10Gb settings disabled/Flow Control disabled

Test 2A:

Settings:

- iscsi-sock-wsize: from 20480 to 98304
- max-async-writepages not set
- filesize: 1.1GB => testfile1

Command run:

```
# dd if=/dev/zero of=/benchmarking/test/testfile1 bs=64k count=16384  
# /usr/local/sbin/panfs_trace -j /benchmarking/test/testfile1
```

Results:

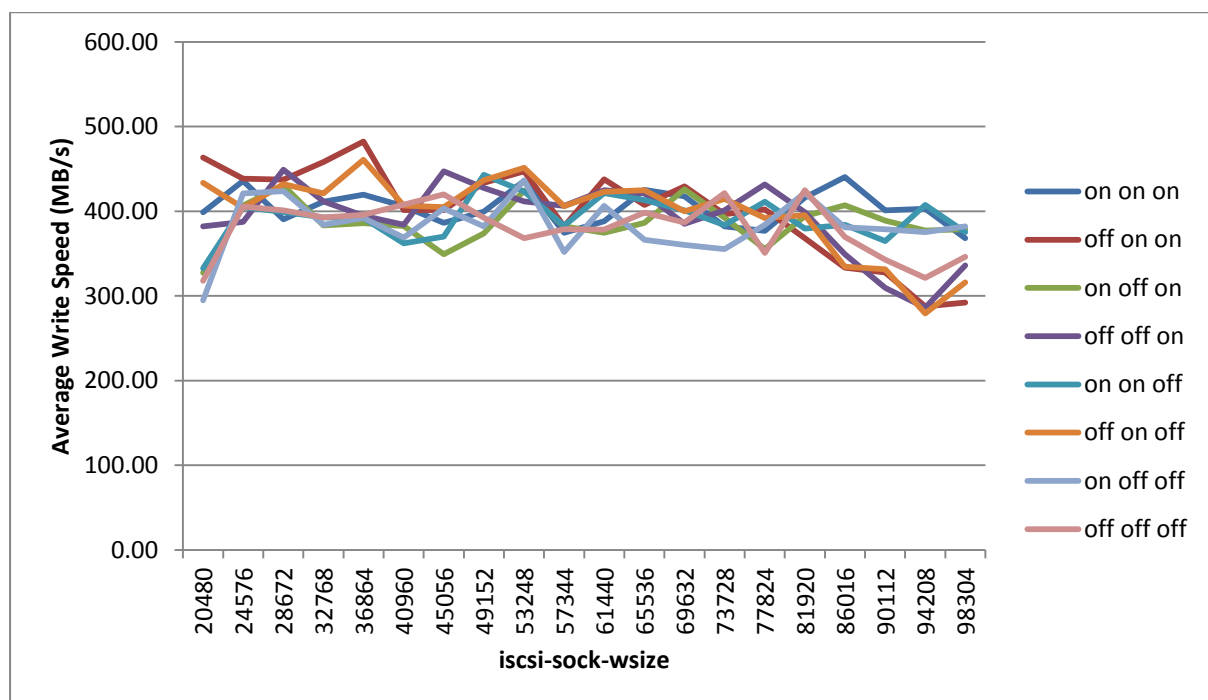


Figure 18: Test 2A

Conclusions:

- No increased speed with higher values of iscsi-sock-wsize.
- On higher values of iscsi-sock-wsize, Ethernet flow control has an impact on speed.

Test 2B:

Based on the results from Test 2A, this test focuses on the range between 53248 and 77824 for iscsi-sock-wsize, using a larger file.

Settings:

- iscsi-sock-wsize: from 53248 to 98304
- max-async-writepages not set
- filesize: 11GB => testfile2

Command run:

```
# dd if=/dev/zero of=/benchmarking/test/testfile2 bs=64k count=163840
# /usr/local/sbin/panfs_trace -j /benchmarking/test/testfile2
```

Results:

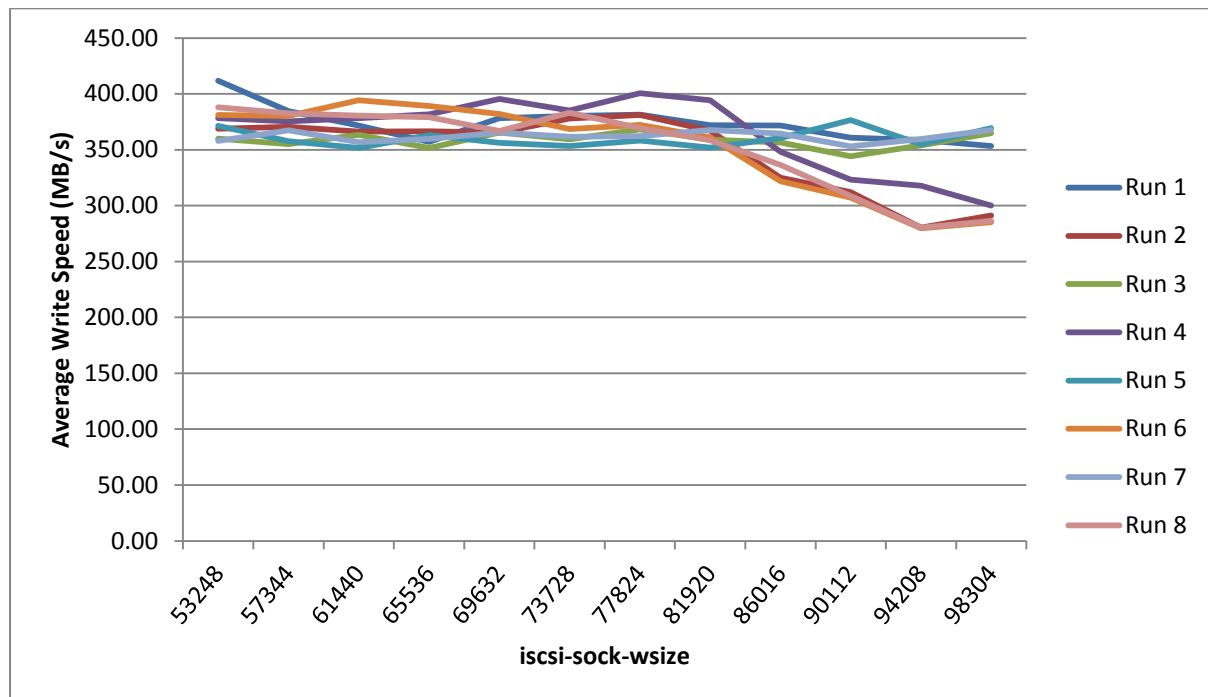


Figure 19: Test 2B

Conclusions:

- Average speed does not change up to iscsi-sock-wsize values of over 81920.
- At high values of iscsi-sock-wsize, the Ethernet flow control must be enabled, as the speed decreases when it's disabled.
- The optimal value for iscsi-sock-rsize is 65536, as there is no obvious improvement by changing it.

Test 2C:

The purpose of this test is to find the optimal value for the max-async-writepages mount option based on the iscsi-sock-rsize and iscsi-sock-wsize obtained in previous tests.

Settings:

- iscsi-sock-wsize: 65536
- max-async-writepages: from 0 to 32
- filesize: 11GB => testfile2

Command run:

```
# dd if=/dev/zero of=/benchmarking/test/testfile2 bs=64k count=163840  
# /usr/local/sbin/panfs_trace -j /benchmarking/test/testfile2
```

Results:

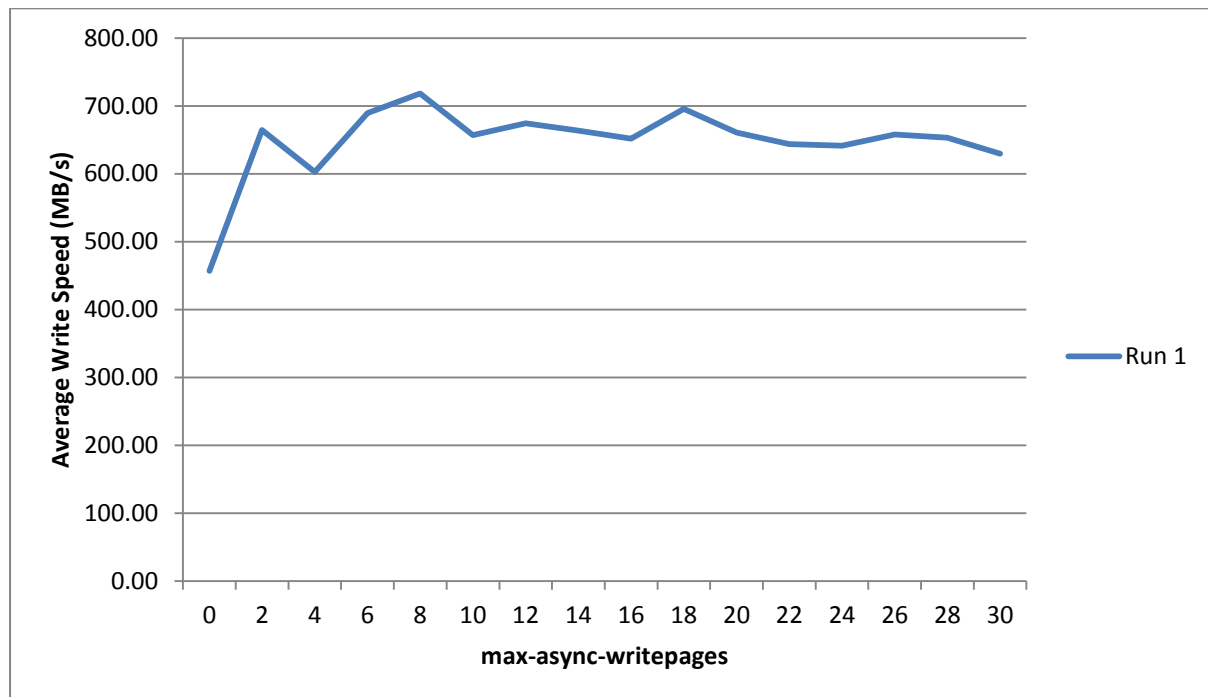


Figure 20: Test 2C

Conclusions:

- max-async-writepages should be enabled for a higher average write speed.
- Values over 8 don't improve the speed.
- A value of 10 is chosen as the optimal value to match the virtual machine setting.

Directory Listing

The aim of this test is to investigate the effect of the statahead-level mount option in the time to run a directory listing (with ls) on a large directory.

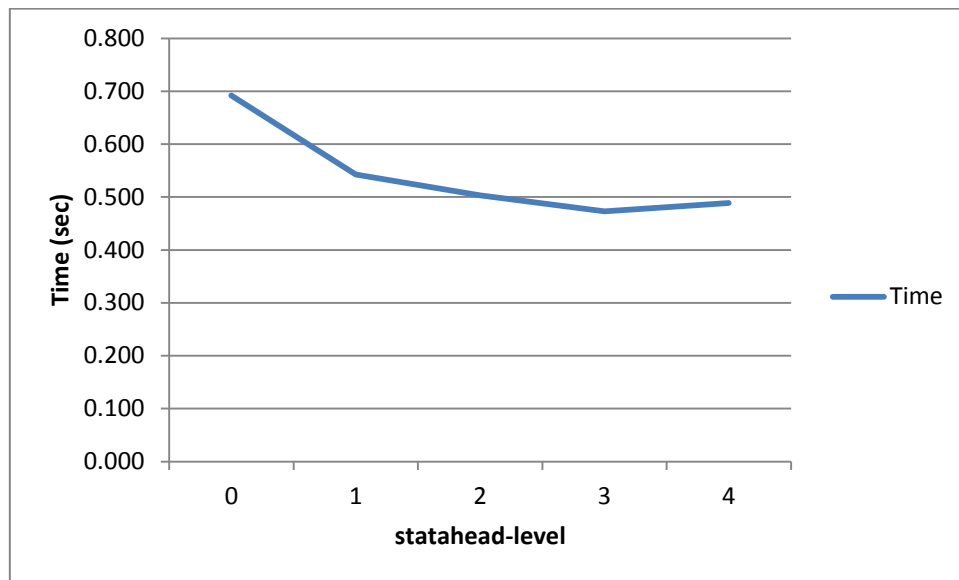


Figure 21: Directory Listing

Conclusions:

- Enabling the statahead-level mount option reduces the time it takes to run a directory listing.
- There isn't much difference between the values 3 and 4, so a value of 3 is chosen to match the virtual machine settings.

Physical Hosts – Optimal Settings

Mount Options	
iscsi-sock-rsize	57344
iscsi-sock-wsize	65536
statahead-level	4
max-async-writepages	10
Kernel Settings	No impact
Ethernet Flow Control	Enabled

Comparison between current settings and settings obtained

Reading - Test Realm

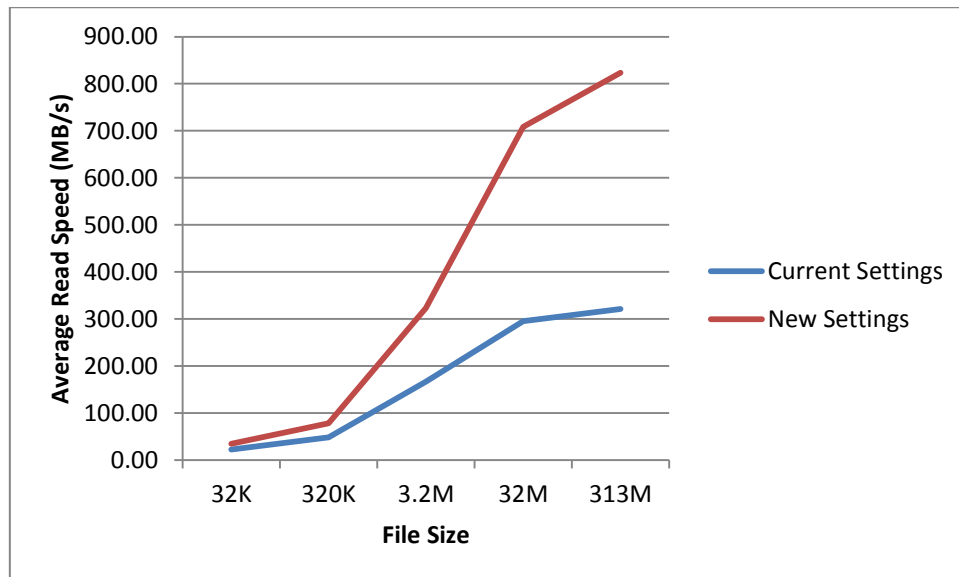


Figure 22: Reading - test realm

Reading - Production Realm

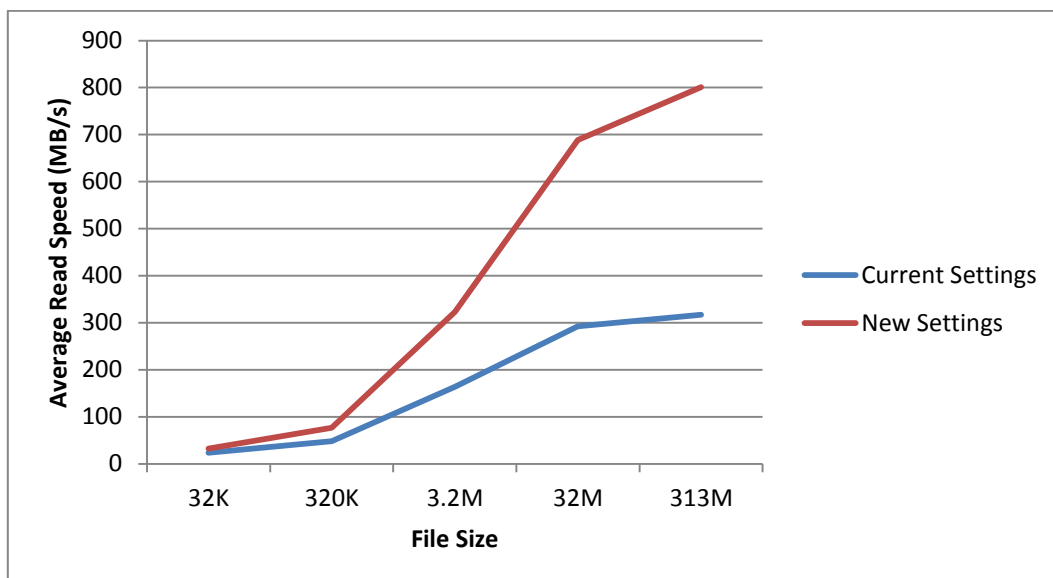


Figure 23: Reading - production realm

Writing - Test Realm

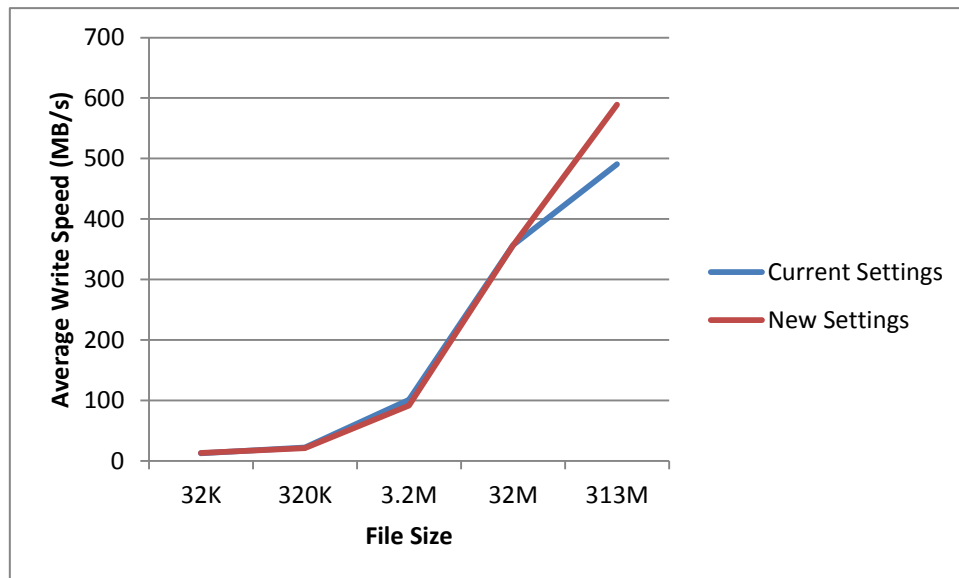


Figure 24: Writing - test realm

Writing - Production Realm

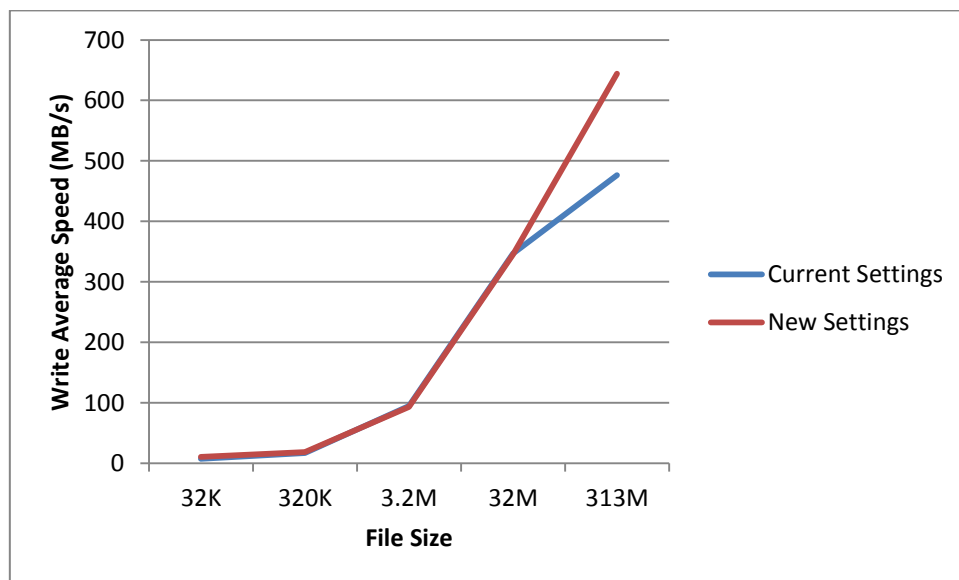


Figure 25: Writing - production realm

Conclusions and future work

- Kernel settings are used for optimising performance for wide area connections. All tests have been run on our local area network, so the different settings have no effect on any of the tests.
- Tests must be run with the client rpm that matches the server version to get all the benefits of the new Panasas features.
- There is no typical IO access pattern in JASMIN/CEMS that could be replicated with tools like IOZone or IOR, so a simple dd benchmark has been chosen for the tests. However, these tools might provide a different IO pattern it would be worth considering them for future investigation.
- New mount options are available at the time of finishing the report. New tests will be needed to investigate the impact on the storage access performance.
- The test environment in this report only covers ActiveStor11. ActiveStor14 shelves will need to be tested in the future.
- The results for the tests in this report are based on an environment where only the test client is configured with the new settings. Applying these new obtained settings to all the clients in the environment will obviously impact the performance of each client.
- An iperf network test run on the physical host used for the testing gives a maximum speed of 4.40Gb/sec (from test host to another host) or 7.22Gb/sec (from another host to test host), which might indicate a firmware issue on the network card. This is currently being investigated.
- Panasas and 10Gb kernel settings were originally recommended by Panasas. Looking closer at these settings after running the different tests, it was noticed that setting the TCP autotuning parameter to 0 (disabling autotuning) stops net.ipv4.tcp_wmem and rmem from having any effect, as these settings are used to set the bounds for autotuning. However, autotuning being turned off isn't a big concern as all the tests have been run in a consistently low latency network.