ADVISORY REGARDING SAS® GRID MANAGER WITH ISILON

ABSTRACT

This document outlines the best practices regarding SAS® Grid Manager with Isilon Storage

March 1, 2018

CONTENTS

OVERVIEW

VERSION SPECIFICITY

The information represented in this document pertains to the following product versions:

- SAS® 9 – all products including SAS® Grid Manager.
- DELL EMC Isilon Gen-5 Hardware: OneFS 8.1.1.0, 8.1.0.2, 8.0.0.6, 8.0.0.5, 8.0.0.4, 7.2.1.3, 7.2.1.1 or 7.2.0.3. Please do not use a version of OneFS prior to 7.2.0.3 with SAS Grid Manager.
- DELL EMC Isilon Gen-6 Hardware: OneFS 8.1.1.0, 8.1.0.2, 8.1.0.1. H500 is recommended Gen-6 model for SAS Grid Manager.

Note: Several RHEL versions can be used for the compute cluster. For OneFS 8.0.0.4 and above + OneFS 8.1.0.1 and above, RHEL 7.3 has delivered the best performance in conducted baseline performance testing.

SAS® GRID MANAGER

This document focuses on SAS Grid Manager, given its popularity for Big Data Analytics environment deployment, in which Isilon's feature set is of the greatest interest. Some references may be made to 'non-Grid' SAS environments to distinguish them from SAS Grid Manager environments, where the distinction is pertinent. SAS provides scalability through parallel processing and the ability to manage, access and process data in a distributed environment. With its ability to work with scalable procedures and I/O engines, it gives applications unmatched potential to scale up in SMP environments and scale out on network nodes simultaneously. Because SAS is analytically powerful, many SAS applications tend to be very data and/or compute intensive. As a result, the performance of these SAS applications may be improved by running in a Grid Manager environment.

DELL EMC ISILON

DELL EMC Isilon's network-attached storage (NAS) leverages scalable NAS performance using a clustered architecture. Powered by Isilon's OneFS™ Operating system, Isilon clusters provide multi-protocol access to large volumes of data over NFS, SMB, HTTP, FTP, and HDFS protocols. OneFS™ is designed to meet market demands for easy-to-deploy scale-out capacity, performance, and operational simplicity for large data enterprises.

SAS® GRID MANAGER ON ISILON

From a large-block IO performance perspective, the NFS protocol is not the highest-performing option for SAS applications. While many SAS and SAS Grid Manager applications can benefit enormously from Isilon's streaming sequential performance and scaling (core strengths of the Isilon architecture), the I/O demands and performance requirements of most SAS-based applications can face large-block WRITE performance issues on NFS based protocols, which is the underpinning of OneFS™.

Obtaining the best possible results with SAS products on DELL EMC Isilon NAS storage requires additional consideration of what SAS file stores to place on NFS-based, network attached Isilon, versus local storage. This document summarizes known SAS-on-Isilon Best Practices and Guidelines and provides links to other documents for details on many specific areas.

GENERAL CONSIDERATIONS

FITNESS AND SCOPE

SAS deploys to three major file system types – persistent, shared storage (SASDATA), non-persistent scratch working space (SASWORK/UTILLOC), and application storage space:

- **SASDATA** refers to persistent data stores shared between clients on a shared filesystem. SASDATA is typically a 70/30 to 80/20 R/W file system, utilizing large files with large block IO. SASWORK typically utilizes a 50/50 R/W IO predication. SASDATA suffices more successfully in shared file spaces, and under NFS protocols due to less large-block WRITE activity.

- **SASWORK/UTILLOC** refers to scratch working data areas exclusive to each SAS process. These spaces cannot be shared between individual SAS client processes. SASWORK is preferred to be on local, or non-NFS based storage, due to its high WRITE predications, which exacerbate known NFS cache coherency issues. For example, the high-speed block storage of DELL EMC's XtremIO, VMAX, VNX, and ScaleIO are better for the SASWORK/UTILLOC file systems.
- Application Storage Space: SAS Server Logs (metadata, object spawner, etc..) should be placed on local filesystems if the SAS install will use OneFS for shared data.

BASE CONFIGURATION

It is recommended to follow these guidelines for Isilon OneFS configuration for SAS Grid.

- OneFS 8.0.0.4 – No additional Isilon OneFS network tuning are required on OneFS
- If 7.2.x.x OneFS is used it is recommended to make the Isilon side network tuning listed in the Appendix
- SASWORK deployed on local SSD or flash-based DAS
- SASDATA deployed on Isilon

OPERATING SYSTEM

Operating Systems settings are distributed to meet a wide and varied range of environments. However, it is sometimes necessary to deviate from the standard OS settings to obtain the best performance possible under a given workload. Complete tuning guidelines can be found in the Additional Resources section of this advisory. As always it is a best practice to test these parameters and consult solution vendors on potential use before deploying into a production environment.

OS Recommendation:

- RHEL 7 – significant changes were made in the RHEL 7 TCP stack to optimize the stack, some tunings are suggested
- RHEL 6.x – May require more significant network tuning as listed in below to maximize performance with 8.0.0.x

NETWORK

When utilizing SAS Grid manager with NFS based shared data, it is required to operate and maintain a clean high performance non-lossy network to limit network interruption and provide as much available bandwidth as possible.

NFS MOUNT CONCEPTS

As discussed previously there are two sets of recommendations around NFS, with respect to NFS3, related to locking behaviors and SAS Grid Manager. As NFSv3 and NFSv4 have different locking behaviors it is important to use one version of the protocol for consistency.

Non-Shared data (SASWORK/UTILLOC) is only accessible to a single SAS session and cannot be shared between other SAS sessions. As such, coordinated file locking for SASWORK/UTILLOC between SAS client systems is not required. The NFS mount option to use ‘local lock=all’ or ‘nolock’ reduces unnecessary overhead for SASWORK/UTILLOC if NFS is used to hosts it (it is against our recommendations to host SASWORK/UTILLOC via NFS). While shared data (SASDATA), requires NFS-based locking in order for SAS to coordinate data access and prevent data corruption.

For non-SAS-Grid deployments, local locks will improve SAS performance. **One must assure that shared locks are not required** to coordinate file access with other applications which may be accessing the same data from other NFS clients. Failure to do so may result in data corruption.

The next section contains a short summary of the main NFS mount option considerations for SAS. For a definitive discussion of the options available on your specific NFS client platform, please consult your version-specific documentation such as the Unix/Linux nfs(5) manual pages.

MOUNT OPTIONS FOR NON-SHARED DATA: SASWORK

The recommendation is to deploy non-shared data (SASWORK, Logs and UTILLOC) on local storage devices on each host compute node. These storage devices ideally should be striped SSDs for adequate performance. These host-local areas are typically mounted to a single host via a dedicated directory and mount point.

It is **not recommended** to deploy non-shared data on Isilon. That said, there may be exceptions to this. For instance:

1. Single storage tier is required
2. SASWORK is already mapped to Isilon at existing customer and customer's SAS workload profile aligns with this configuration

In these exceptional situations where non-shared data is deployed on Isilon, the following NFS mount point configurations are recommended for SAS:

- nolock - Enables client-local lock management for greatly-improved performance. In as much as more modern versions of SAS do not perform file-locking operations in Non-Shared areas, one may see no benefit to this setting with specific workloads. However, since Non-Shared areas do not require inter-client coordination, it remains as a Best Practice to use this setting for Non-Shared storage areas.
- noatime
- nodiratime
- actimeo=86400 – That is one day in seconds, but any reasonably-large value should suffice
- rdirplus - Improves 'ls -l' performance. This is normally the default, but re-stating it should not be a problem.
- nocto - This option may provide some performance benefit, but may also result in cascaded performance issues with memory management. Its utility must be evaluated on a case-by-case basis by empirical testing.
- Make sure to explicitly define the host IP or Name that can mount, and have root access, for each export to reduce the risk another host can mount the Non-Shared data.
- vers=3

e.g.: vers=3,nolock,noatime,nodiratime,rdirplus, nocto, actimeo=86400

If Isilon is used for Non-Shared data, dedicated exports and directories should be created and used by the client mount point to isolate each grid node to dedicated space on Isilon to maximize NFS operation with remote Non-Shared data.

MOUNT OPTIONS FOR SHARED DATA: SASDATA

For Shared data areas located on Isilon:

- acdirmin=0,acdirmax=0 - Preferred options to accommodate the needs of SAS products.
- Alternatives such as noac and actimeo=0 have significantly-higher negative performance consequences and are therefore **not** recommended.
- noatime

- nodiratime

- rdirplus - improves 'ls -l' performance. This is normally the default, but re-stating it should not be a problem.

- vers=3

e.g.: vers=3, acdirmin=0, noatime, nodiratime, rdirplus

VALIDATE AND REVIEW ALL MOUNTS

Confirm NFS mount options under Linux by using the 'mountstats' command for each mount point. The mountstats command reveals the values that were actually chosen by each mount negotiation process; for various reasons, those values may not be the same as what is specified in /etc/fstab and later shown by the 'mount' command. Review and asses the actual mount configuration for deviation and address accordingly.

SETUID

Some SAS executables use 'setuid root'. Information on these files and what they do can be found in this linked document. If these files are accessed via NFS, the following constraints apply;

- The client's mount point **must not** use the 'noexec' or 'nosuid' mount options.

- The server's export **must not** suppress the root user's identity. (i.e, On Isilon do not map 'root' to 'nobody'.)

SAS® ON ISILON

Isilon offers SAS interesting feature sets uncommon in many other deployments. One main benefit of Isilon is ease and simplicity to scale the solution according to demand. It is generally much easier to increase the size of network attached storage than it is to manage LUNs on a typical array. Isilon also offers some very beneficial options around data protection and file layout documented further below. Several important settings for optimal configuration are listed below by category.

32BIT FILE ID'S FOR SAS 9.3

When using SAS 9.3 or earlier the filesystems needs to support 32-bit file ID's. Isilon supports 32bit file ID's but they are disabled by default. To enable 32bit File ID's:

```
isi nfs exports modify <file system ID number> –return-32bit-file-ids=yes
```

NFS EXPORTS

Since most NFS performance factors are associated with client mounts and client mount options, one could use a single NFS export to facilitate all of the mounts associated with a given SAS landscape. No specific export tuning are required, configuration of the exports should meet the requirements of your sas grid access requirements.

ISILON CONNECTIVITY

Isilon operates a DNS responder-based load balancer to distribute traffic across the Isilon cluster nodes. This feature is known as Smart Connect Advanced. It is recommended to use SmartConnect Advanced and Dynamic Pools for NFS. Dynamic Smart Connect Pools are required for nfs v3 IP failover functionality. Support for NFSv4 graceful failover is in OneFS 8.0. For additional information on implementing a Dynamic SmartConnect Zone, see the External Network Connectivity Guide.

In smaller OneFS clusters it is possible to manually define IP's for mount points from grid nodes to ensure an equal distribution of Isilon interfaces to grid nodes. As OneFS cluster get bigger it is suggested to create mount points using smartconnect zone names to automate connectivity and load balancing.

Note: When using the sizing calculation from the External Network Connectivity Guide for small grids (3/4/5 nodes) there may be too few IPs to spread mounts meaningfully (for instance for a five node grid if suggests allocation of 20 IP addresses – but with a requirement for 100 mount points that could be too few, putting 5 mounts on each IP on each client). In this case it would be better to allocate a greater number of addresses to the Smart Connect Pool.

In order to leverage the scale out capabilities of OneFS and optimize network connections from grid hosts to Isilon nodes it is suggested to use multiple mount points connected to OneFS so additional sessions from the client can leverage different nodes and interfaces to maximize the networking infrastructure utilized by OneFS.

An example would be to create multiple mount points to the shared data root on Isilon but connecting to different IP either specifically or using SmartConnect. In this example, each grid node can access the exported shared data via different mounts to increase the use of separate session and maximize all the interfaces and nodes available. So instead of all grid nodes used a single mount point /mnt/Isilon/data to access storage from the Isilon of /ifs/zone/sasdata, it is divided up and mounted by sub-areas, which are spread across IP addresses on the Isilon:

```
grid-node1 - /mnt/Isilon/data/area1 -- > IP1:/ifs/zone/sasdata/area1
grid-node1 - /mnt/Isilon/data/area2 -- > IP2:/ifs/zone/sasdata/area2
grid-node1 - /mnt/Isilon/data/area3 -- > IP3:/ifs/zone/sasdata/area3

grid-node2 - /mnt/Isilon/data/area1 -- > IP4:/ifs/zone/sasdata/area1
grid-node2 - /mnt/Isilon/data/area2 -- > IP5:/ifs/zone/sasdata/area2
grid-node2 - /mnt/Isilon/data/area3 -- > IP6:/ifs/zone/sasdata/area3
```

(Bigger Isilon clusters use SmartConnect FQDN as the target for the mount point):

```
grid-node1 - /mnt/Isilon/data/area1 -- > SCZ-FQDN:/ifs/zone/sasdata/area1
grid-node1 - /mnt/Isilon/data/area2 -- > SCZ-FQDN:/ifs/zone/sasdata/area2
grid-node1 - /mnt/Isilon/data/area3 -- > SCZ-FQDN:/ifs/zone/sasdata/area3

grid-node2 - /mnt/Isilon/data/area1 -- > SCZ-FQDN:/ifs/zone/sasdata/area1
grid-node2 - /mnt/Isilon/data/area2 -- > SCZ-FQDN:/ifs/zone/sasdata/area2
grid-node2 - /mnt/Isilon/data/area3 -- > SCZ-FQDN:/ifs/zone/sasdata/area3
```

By using multiple mounts to multiple IP we can increase the number of sessions and scale out more effectively, using OneFS SmartConnect allows the OneFS cluster to manage the distribution of connections and also the failover/failback capabilities of NFS v3.

The number of mounts created will depend on the IP allocation strategy and the number of assigned IP's and the client configuration

Recommendations:

- Assign the appropriate numbers of IP addresses to the SmartConnect Network Pool
- Use a Dynamic allocation methodology for the SmartConnect Network Pool
- Create multiple mount points from grid nodes to maximize use of the OneFS interfaces through multiple sessions

DATA ACCESS PATTERN

Isilon can optimize disk layout and cache utilization based on expected access patterns. Automation of these parameters is available through SmartPools. By Default, and without SmartPools, there is a pool default access pattern. Administrators manually denote the access pattern of specific files or directories. SmartPools enabled both file policies as well as a policy engine. These policies are defined at the file or folder level and determine the access pattern of files written to a location among various other options.

The behavior of the default "concurrency" setting has changed in OneFS 8.0 such that it implements an adaptive readhead algorithm. Because of this, the following recommendation should only be considered for OneFS 7.2.x versions and prior. For Shared data paths the recommended I/O optimization is "streaming" since the expected access pattern is large block contiguous I/O. The streaming I/O optimization increases the number of disks incorporated into file layout and enables a more aggressive cache pre-fetch algorithm. For Non-Shared data paths the recommended I/O optimization is "concurrency" as these regions exhibit both random and sequential I/O patterns.

Recommendation:

- One 8.0.x Clusters use the default Concurrency data access patterns
- One 7.2.x Clusters the data workflow can be evaluated against a streaming access pattern for Shared Data

SSD STRATEGY

Isilon can alter the behaviors around metadata handling to improve performance. By default, the Isilon uses an SSD Strategy known as L3 and this works well for most use cases. Alternatives are Metadata Read Acceleration and Metadata Read and Write acceleration. Whereas Metadata Read acceleration writes an extra mirror of the metadata to SSD, Metadata Read Write Acceleration makes every attempt to write all metadata onto SSDs. This accelerates metadata operations for both reads and writes but also uses much more SSD capacity. If Isilon is a dedicated cluster for use by SAS Grid Manager, then use Metadata Read and Write acceleration. Otherwise evaluate all workloads on the Isilon cluster before considering a change in order to strike the right balance across workloads. The Metadata Read strategy, which is implemented slightly differently by L3, is incorporated into the L3 SSD Strategy. Metadata Read and Write acceleration is not available with the L3 SSD Strategy.

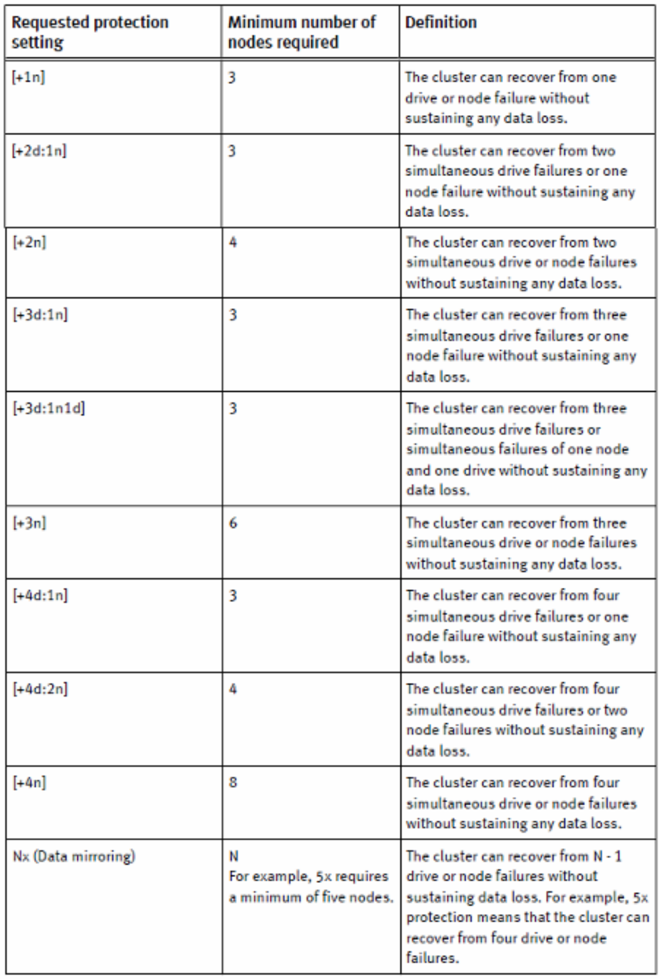
Recommendation:

- Use L3 caching SSD strategy, unless other workflows and performance have been evaluated for impact

DATA PROTECTION MODES

Isilon OneFS lets you specify the protection of a file or directory by setting its requested protection. This flexibility enables you to protect distinct sets of data at higher than default levels. Requested protection of data is calculated by OneFS and set automatically on storage pools within your cluster. The default setting is referred to as suggested protection, and provides the optimal balance between data protection and storage efficiency. For example, a suggested protection of N+2:1 means that two drives or one node can fail without causing any data loss. This would be the suggested default protection for a small cluster.

Requested protection settings determine the level of hardware failure that a cluster can recover from without suffering data loss.



Requested protection setting	Minimum number of nodes required	Definition
[+1n]	3	The cluster can recover from one drive or node failure without sustaining any data loss.
[+2d:1n]	3	The cluster can recover from two simultaneous drive failures or one node failure without sustaining any data loss.
[+2n]	4	The cluster can recover from two simultaneous drive or node failures without sustaining any data loss.
[+3d:1n]	3	The cluster can recover from three simultaneous drive failures or one node failure without sustaining any data loss.
[+3d:1n1d]	3	The cluster can recover from three simultaneous drive failures or simultaneous failures of one node and one drive without sustaining any data loss.
[+3n]	6	The cluster can recover from three simultaneous drive or node failures without sustaining any data loss.
[+4d:1n]	3	The cluster can recover from four simultaneous drive failures or one node failure without sustaining any data loss.
[+4d:2n]	4	The cluster can recover from four simultaneous drive failures or two node failures without sustaining any data loss.
[+4n]	8	The cluster can recover from four simultaneous drive or node failures without sustaining any data loss.
Nx (Data mirroring)	N For example, 5x requires a minimum of five nodes.	The cluster can recover from N - 1 drive or node failures without sustaining data loss. For example, 5x protection means that the cluster can recover from four drive or node failures.

For best results, we recommend that you accept at least the suggested protection for data on your cluster. You can always specify a higher protection level than suggested protection on critical files, directories, or node pools. As the cluster increases in node size the suggested protection will change to reflect the cluster size in order to effectively protect the data.

Increasing the requested protection of data also increases the amount of space consumed by the data on the cluster. The parity overhead for N + M protection depends on the file size and the number of nodes in the cluster. The percentage of parity overhead declines as the cluster gets larger. See the Appendix for the estimated percentage of overhead depending on the requested protection and the size of the cluster or node pool.

For Shared data the recommended protection policy, or any other hybrid protection policy, are applicable so that capacity is optimized. Employing this approach balances the needs for fault tolerance and performance with respect to capacity.

Recommendation:

- SASDATA – use at a minimum Suggested Protection Policy for the size of the cluster

NFS CLIENT TUNING SUGGESTIONS

It is recommended to validate your network and workflow for throughput and performance using a well understood test or workflow that is well understood for your environment; SAS-client-network-Isilon. By testing and understanding the behavior of a well-understood workflow we can look at making tuning recommendations to optimize the performance.

- Baseline the network with iperf (see appendix for additional details)
- Execute a parallel dd or use iotest.sh: available from http://support.sas.com/kb/59/680.html to baseline NFS performance
- Execute a SAS workflow

A SUGGESTED TESTING METHODOLOGY WOULD BE:

- Mount OneFS NFS filesystem
- cd to the SAS mount
- Create a “testdata” directory.
- Generate test data with the test or workflow selected.
- Monitor:
    - # nfsiostat 5 -adp /path/to/<sas-mount-directory>
    - # use Isilon IIQ
- Between every test run, clear the client cache:
    - # echo 3 > /proc/sys/vm/drop_caches

Many of the recommendation and tunings listed below come from testing the behavior of different setting with a standard test as described above. It is **highly** recommended to review the RedHat Network Performance Tuning Guide when evaluating and making network tunings. Specifically, that document led to the findings and changes to the clientside settings in internal testing as documented below.

Having baselined the NFS performance depending on the results obtained, additional network tunings can be made and evaluated. The following tuning guidance was tested and validate with the RHEL OS and OneFS version listed. The hardware configuration was a 4-node Isilon cluster consisting of X410 nodes with 256GB RAM, 10GbE bxe networking. The clients were multi-socket Intel servers with Intel 10GbE (ixgbe) network cards.

OBSERVED TUNINGS

RHEL 6.9 – OneFS 7.2.x

a. Modify the default OneFS 10GB sysctls (Appendix)

(No additional client side tunings were tested or validated at this time)

RHEL 6.9 – OneFS 8.0.0.x – These settings have been observed to provide significant performance improvements.

(Highly recommended to test and validate these tunings)

a. No modification to OneFS cluster sysctls alone
b. No modification to client TCP/net sysctls alone
c. If your client has >1 Intel CPU socket, fix NIC interrupt affinity so all IRQs land on one socket.
    - These recommendations apply to any NIC. The irqbalance policy documented here would need to be modified to suit.
    - For the Intel 10GbE card, there’s a script to do this supplied with the driver on sourceforge.net:

      https://sourceforge.net/p/e1000/bugs/383/#3f1d
      https://sourceforge.net/p/e1000/bugs/_discuss/thread/b5564986/3f1d/attachment/irqbalance-ixgbe-policy

d. Using ethtool to tune the card ring buffers to their maximum size (for Intel that’s 512 -> 4096 for both the RX and TX rings). For other cards, use “ethtool –g” to discover the available ring buffer settings and modify appropriately.
e. Turn **ON** Large Receive Offload on the card (using ethtool). This is a significant gain for 8.0.
f. Turn **OFF** GRO on the card (using ethtool).
g. Consider upgrading the NIC driver to the latest version. In the case of the ixgbe driver, this provides further performance gains as compared to the stock OS driver.

RHEL 7 – OneFS 7.2.x – These settings may yield some performance improvements.

(Test and evaluate if these tunings are required)

a. No modification to OneFS cluster sysctls alone
b. No modification to client TCP/net sysctls alone
c. If your client has >1 Intel CPU socket, fix NIC interrupt affinity so all IRQs land on one socket.
    - These recommendations apply to any NIC. The irqbalance policy documented here would need to be modified to suit.
    - For the Intel 10GbE card, there’s a script to do this supplied with the driver on sourceforge.net:

https://sourceforge.net/p/e1000/bugs/383/#3f1d
https://sourceforge.net/p/e1000/bugs/_discuss/thread/b5564986/3f1d/attachment/irqbalance-ixgbe-policy

d. Using ethtool to tune the card ring buffers to their maximum size (for Intel that's 512 -> 4096 for both the RX and TX rings). For other cards, use "ethtool –g" to discover the available ring buffer settings and modify appropriately.
e. For the Intel card, ensure Large Receive Offload is disabled (using ethtool). This interacts badly with the OneFS 7.2 bxe TSO setting.
f. Consider upgrading the NIC driver to the latest version. In the case of the ixgbe driver, this provides further performance gains as compared to the stock OS driver.

RHEL 7 – OneFS 8.0.0.x and OneFS 8.1.X – These settings may yield some performance improvements.

(Test and evaluate if these tunings are required)

a. No modification to OneFS cluster sysctls alone
b. No modification to client TCP/net sysctls alone
c. If your client has >1 Intel CPU socket, fix NIC interrupt affinity so all IRQs land on one socket.
   - These recommendations apply to any NIC. The irqbalance policy documented here would need to be modified to suit.
   - For the Intel 10GbE card, there's a script to do this supplied with the driver on sourceforge.net:

     https://sourceforge.net/p/e1000/bugs/383/#3f1d
     https://sourceforge.net/p/e1000/bugs/_discuss/thread/b5564986/3f1d/attachment/irqbalance-ixgbe-policy

d. Using ethtool to tune the card ring buffers to their maximum size (for Intel that's 512 -> 4096 for both the RX and TX rings). For other cards, use "ethtool –g" to discover the available ring buffer settings and modify appropriately.
e. Turn **ON** Large Receive Offload on the card (using ethtool). This is a significant gain for 8.0.
f. Turn **OFF** GRO on the card (using ethtool).
g. Consider upgrading the NIC driver to the latest version. In the case of the ixgbe driver, this provides further performance gains as compared to the stock OS driver.

The primary goal of these tunings is to observe the impact the modification has on throughput and performance when observed and evaluated in your environment with your workflow. The complexity involved with today's modern infrastructure may require additional evaluation of these settings to determine the impact they have on NFS workflows with Isilon.

CONCLUSION

As a reminder, heavy WRITE intensive workloads, such as exist in Non-Shared file systems (SASWORK, UTILLOC, Metadata Logs), should be placed on non-NFS based, local storage. In addition, the SAS Metadata and SAS Object Spawner logs should be placed on local filesystems if the SAS install accesses OneFS. Shared data (SASDATA) should be placed on Isilon. It is highly recommended to validate the performance of the SAS Grid manager with Isilon OneFS and evaluate if additional tuning is required.

It is crucial to work with your Isilon engineer to plan, install, and tune your host utilizing the cluster to achieve maximum performance. The recommendations in advisory are applicable to majority of SAS-Grid and Isilon deployments. Exceptions may be needed for individual deployments driven by host system environment, unique workload requirements etc., (ex. deploying SASWork on Isilon). If guidance is needed for deviations from this advisory for your unique situation, please reach out to the contacts listed in the "CONTACT INFORMATION" section.

ADDITIONAL RESOURCES

These following references make occasional mention of NFS and NAS, and provide some background information about SAS I/O and performance in general. For operating SAS on Isilon NFS storage, this document is intended to supersede all others. Additional references will be added here as they become available. The following SAS-developed documents should be reviewed independently for additional background and insight regarding SAS I/O requirements;

Best Practices for Configuring IO for SAS

FAQ for Storage Configuration

http://support.sas.com/kb/59/680.html

https://access.redhat.com/sites/default/files/attachments/20150325_network_performance_tuning.pdf

https://sourceforge.net/p/e1000/bugs/_discuss/thread/b5564986/3f1d/attachment/irqbalance-ixgbe-policy

https://sourceforge.net/p/e1000/bugs/383/#3f1d

https://support.emc.com/docu58740_Isilon-External-Network-Connectivity-Guide---Routing,-Network-Topologies,-and-Best-Practices-for-SmartConnect.pdf

For Operating System tuning guidelines, please go to http://support.sas.com/kb/53/873.html

APPENDIX

EVALUATE A NETWORK USING IPERF

The iperf program tests the raw network throughput from the client to the server without the protocol layer. This enables you to establish a rough baseline of what raw traffic over the network looks like. Comparing the results you get from iperf with the approximate values listed in Table A - Average speed per interface type can help determine whether an underlying network issues exists. It is critical in establishing this baseline before continuing to test performance.

NOTE: Run iperf when there is minimal or no traffic on the network. Traffic on the network will skew the results. We are looking to evaluate core network capability with iperf
1. Open an SSH connection on any node in the cluster and log on using the "root" account.
2. From the command line, run: iperf -s

The output looks similar to the following:

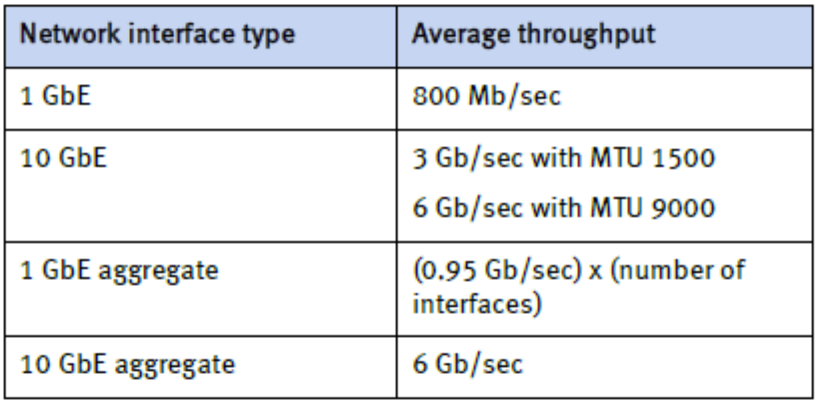
```
------------------------------------------------------------
Server listening on TCP port 5001
TCP window size: 128 KByte (default)
------------------------------------------------------------
```

3. From the client, run the following command:
iperf -c <IP address of the cluster>
The test runs. When the test completes, the output looks similar to the following:

```
------------------------------------------------------------
Client connecting to 192.168.167.11, TCP port 5001
TCP window size: 64.0 KByte (default)
------------------------------------------------------------
[ 3] local 192.168.167.2 port 60492 connected with 192.168.167.11 port 5001
[ ID] Interval Transfer Bandwidth
[ 3] 0.0-10.0 sec 6.71 GBytes 5.76 Gbits/sec
```

4. Find the throughput in the output. In the example above, the throughput is 5.76 Gbits/sec.

5. Repeat steps 3 and 4 two more times and consider the three throughput results.
• If the results are all fairly consistent, this provides a solid baseline of the network
• If the results are all very different, the problem is likely related to the physical network. (Investigate further)
• If two are consistent and one is very different, Rerun the tests, are the results consistant or highly variant still? (Variance may still indicate an underlying network issue)



Network interface type	Average throughput
1 GbE	800 Mb/sec
10 GbE	3 Gb/sec with MTU 1500 6 Gb/sec with MTU 9000
1 GbE aggregate	(0.95 Gb/sec) x (number of interfaces)
10 GbE aggregate	6 Gb/sec

Table A: Average speed per interface type

6. Compare the values you get by running iperf to the values in Table A - Average speed per interface type. The table indicates the average throughput you can expect to get on various interface types.

Note that these values are not absolute; they are meant to be used as a guide.

- If your throughput results are substantially slower than the throughput listed in the table, the problem might be related to the physical network. It is highly recommended to investigate and troubleshoot further.

ETHTOOL

Some basic ethtool commands are listed below:

Review features:

```
ethtool --show-features where "<ethX>" is your NIC device
```

Review Ring Buffer parameters:

```
ethtool  -g "<ethX>" is your NIC device
```

Modify LRO or GRO on the Interface card:

```
ethtook -K <ethX> lro on gro off" where "<ethX>" is your NIC device
```

Consult the man ethtool pages for additional information.

ISILON 7.X NETWORK INTERFACE TUNINGS

When using OneFS 7.x code the following network sysctl should be modified from the default options:

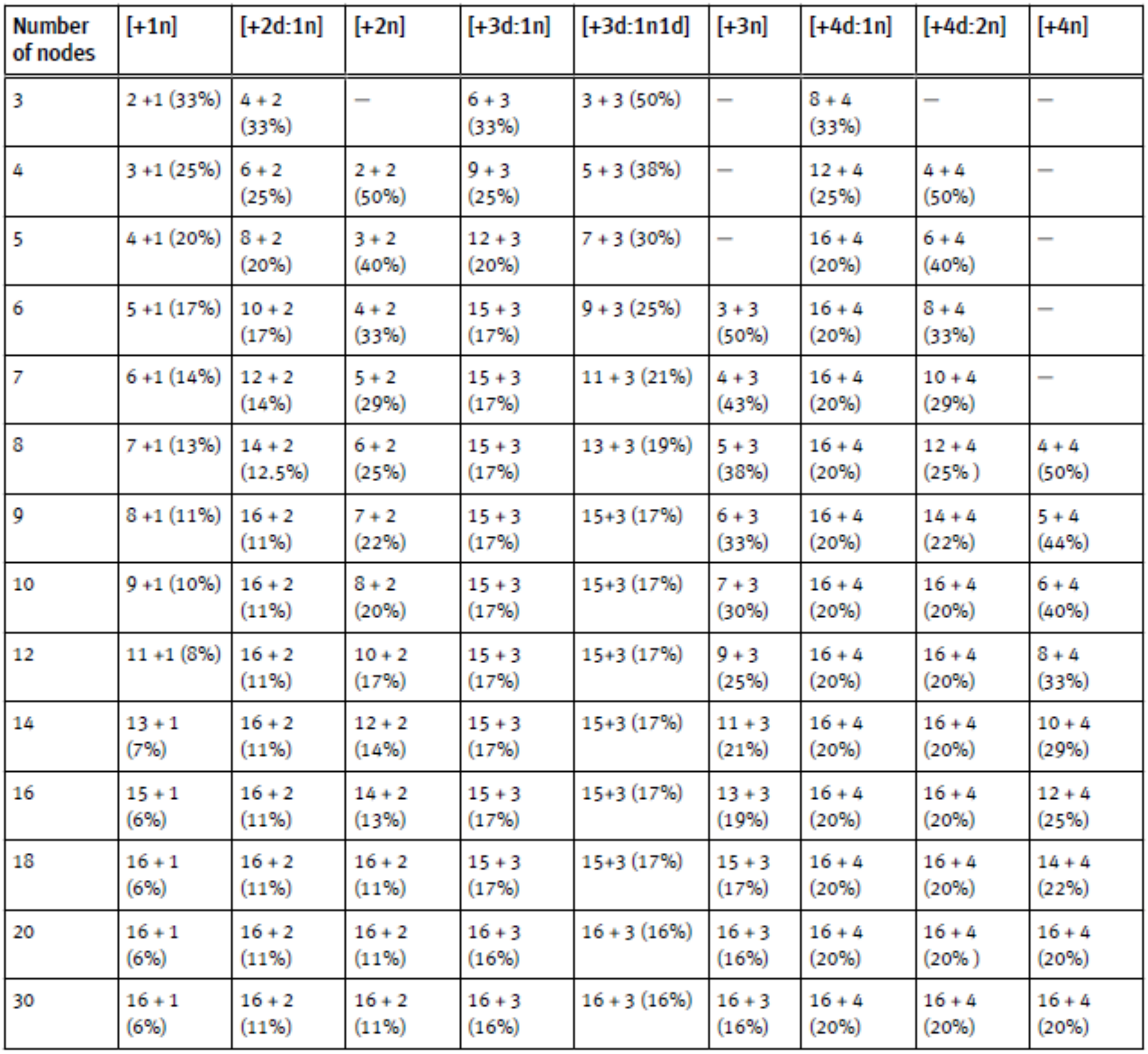
sysctl OID	Recommended	Default
kern.ipc.maxsockbuf	*16777216*	*2097152*
net.inet.tcp.recvspace	*524288*	*131072*
net.inet.tcp.recvbuf_inc	*32768*	*16384*
net.inet.tcp.recvbuf_max	*2097152*	*262144*
net.inet.tcp.sendspace	*524288*	*131072*
net.inet.tcp.sendbuf_inc	*16384*	*8192*
net.inet.tcp.sendbuf_max	*2097152*	*262144*

It is recommended to validate any sysctl's after any oneFS upgrades to ensure they persisted through the upgrade process depending on how they were set.

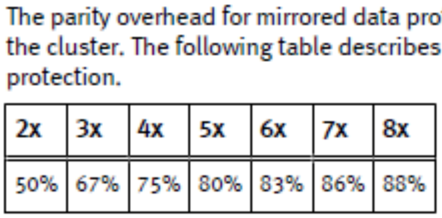
REQUESTED PROTECTION DISK SPACE USAGE

Increasing the requested protection of data also increases the amount of space consumed by the data on the cluster. The parity overhead for N + M protection depends on the file size and the number of nodes in the cluster. The percentage of parity overhead declines as the cluster get larger.

The following table describes the estimated percentage of overhead depending on the requested protection and the size of the cluster or node pool.



Number of nodes	[+1n]	[+2d:1n]	[+2n]	[+3d:1n]	[+3d:1n1d]	[+3n]	[+4d:1n]	[+4d:2n]	[+4n]
3	2 +1 (33%)	4 + 2 (33%)	—	6 + 3 (33%)	3 + 3 (50%)	—	8 + 4 (33%)	—	—
4	3 +1 (25%)	6 + 2 (25%)	2 + 2 (50%)	9 + 3 (25%)	5 + 3 (38%)	—	12 + 4 (25%)	4 + 4 (50%)	—
5	4 +1 (20%)	8 + 2 (20%)	3 + 2 (40%)	12 + 3 (20%)	7 + 3 (30%)	—	16 + 4 (20%)	6 + 4 (40%)	—
6	5 +1 (17%)	10 + 2 (17%)	4 + 2 (33%)	15 + 3 (17%)	9 + 3 (25%)	3 + 3 (50%)	16 + 4 (20%)	8 + 4 (33%)	—
7	6 +1 (14%)	12 + 2 (14%)	5 + 2 (29%)	15 + 3 (17%)	11 + 3 (21%)	4 + 3 (43%)	16 + 4 (20%)	10 + 4 (29%)	—
8	7 +1 (13%)	14 + 2 (12.5%)	6 + 2 (25%)	15 + 3 (17%)	13 + 3 (19%)	5 + 3 (38%)	16 + 4 (20%)	12 + 4 (25%)	4 + 4 (50%)
9	8 +1 (11%)	16 + 2 (11%)	7 + 2 (22%)	15 + 3 (17%)	15+3 (17%)	6 + 3 (33%)	16 + 4 (20%)	14 + 4 (22%)	5 + 4 (44%)
10	9 +1 (10%)	16 + 2 (11%)	8 + 2 (20%)	15 + 3 (17%)	15+3 (17%)	7 + 3 (30%)	16 + 4 (20%)	16 + 4 (20%)	6 + 4 (40%)
12	11 +1 (8%)	16 + 2 (11%)	10 + 2 (17%)	15 + 3 (17%)	15+3 (17%)	9 + 3 (25%)	16 + 4 (20%)	16 + 4 (20%)	8 + 4 (33%)
14	13 + 1 (7%)	16 + 2 (11%)	12 + 2 (14%)	15 + 3 (17%)	15+3 (17%)	11 + 3 (21%)	16 + 4 (20%)	16 + 4 (20%)	10 + 4 (29%)
16	15 + 1 (6%)	16 + 2 (11%)	14 + 2 (13%)	15 + 3 (17%)	15+3 (17%)	13 + 3 (19%)	16 + 4 (20%)	16 + 4 (20%)	12 + 4 (25%)
18	16 + 1 (6%)	16 + 2 (11%)	16 + 2 (11%)	15 + 3 (17%)	15+3 (17%)	15 + 3 (17%)	16 + 4 (20%)	16 + 4 (20%)	14 + 4 (22%)
20	16 + 1 (6%)	16 + 2 (11%)	16 + 2 (11%)	16 + 3 (16%)	16 + 3 (16%)	16 + 3 (16%)	16 + 4 (20%)	16 + 4 (20%)	16 + 4 (20%)
30	16 + 1 (6%)	16 + 2 (11%)	16 + 2 (11%)	16 + 3 (16%)	16 + 3 (16%)	16 + 3 (16%)	16 + 4 (20%)	16 + 4 (20%)	16 + 4 (20%)



The parity overhead for mirrored data protection is not affected by the number of nodes in the cluster. The following table describes the parity overhead for requested mirrored protection.

2x	3x	4x	5x	6x	7x	8x
50%	67%	75%	80%	83%	86%	88%

CONTACT INFORMATION

Mike Kirkpatrick
DELL EMC Corporation
176 South Street
Hopkinton, MA 01748
508-435-1000
mike.kirkpatrick@isilon.com

Margaret Crevar
SAS Institute Inc.
100 SAS Campus Dr
Cary, NC 27513-8617 United States
919-531-7095
margaret.crevar@sas.com

Special Thanks to Tony Brown from SAS, Tim Wright from DELL EMC, and Bob Sneed from DELL EMC for their review and input.