

Optimizing SAS® on Red Hat Enterprise Linux (RHEL) 6 & 7

Barry Marson, Red Hat, Inc.

ABSTRACT

Today, companies are increasingly utilizing analytics to discover new revenue and cost-saving opportunities. Many business professionals turn to SAS, a leader in business analytics software and service, to help them improve performance and make better decisions faster. Analytics are also being employed in risk management, fraud detection, life sciences, sports, and many more emerging markets. However, to maximize the value to the business, analytics solutions need to be deployed quickly and cost-effectively, while also providing the ability to readily scale without degrading performance. Of course, in today's demanding environments, where budgets are still shrinking and mandates to reduce carbon footprints are growing, the solution must deliver excellent hardware utilization, power efficiency, and return on investment.

To help solve some of these challenges, Red Hat and SAS have collaborated to recommend the best practices for configuring SAS 9 running on Red Hat Enterprise Linux.

INTRODUCTION

The scope of this document will cover best practices for configuring and optimizing SAS on Red Hat Enterprise Linux 6 and 7. Areas researched include file system selection, the I/O subsystem, and kernel tuning, both in bare metal and virtualized (KVM) environments. Additionally we now include Grid based configurations running with Red Hat Resilient Storage (GFS2 clusters).

This paper discusses:

- Setup & Tuning Requirements for All SAS Environments
- SAS with Red Hat Virtualization
- SAS Grid with Red Hat Shared File Systems
 - Red Hat Resilient Storage (formerly GFS2 clusters)
 - Red Hat Gluster Storage
 - Red Hat CEPH Storage

For information about Red Hat Enterprise Linux 5, see the Optimizing RHEL for SAS 9.2 document dated January 2011.

SETUP & TUNING REQUIREMENTS FOR ALL SAS ENVIRONMENTS

SAS environments have some very specific setup and tuning requirements to achieve optimal performance.

While there are several new SAS deployments in virtualized (KVM) and GRID based environments, many existing and new installations still require a standalone SAS installation. Its essential to read this chapter regardless of your type of installation.

One of the most important considerations when deploying SAS is choosing the right file system and configuring the I/O stack correctly. Even subtle mistakes can cause serious degradation in performance and scalability. While most performance configuration issues can be resolved, the challenge lies in finding these issues early enough to reduce system downtime. The most common mistakes usually require the recreation of the file system, which means data already on the file system needs to be backed up. To

further complicate matters, that archive process can be painfully slow depending on how things were configured. The information here is intended to help you make informed decisions before you deploy.

Areas covered here include:

- File System Selection
- File System Layout
- The I/O Stack
- Kernel Tuning for SAS

FILE SYSTEM SELECTION

Red Hat Enterprise Linux offers several choices of file systems. For stand alone systems, the choice has been pretty clear for several years that the best performing and scaling file systems in order of preference are XFS and ext4. Both of these file systems have extent based allocation schemes which help in their overall file system throughput

Ext3 simply cannot perform well with the large streaming I/O requirements of many SAS applications mostly due to its non extent based allocation scheme and its indirect block layout. Additionally, ext3 has file and file system size limitations noted in the table below.

GFS2 clustered file system can be used for SAS Grid Manager and is discussed later in this paper.

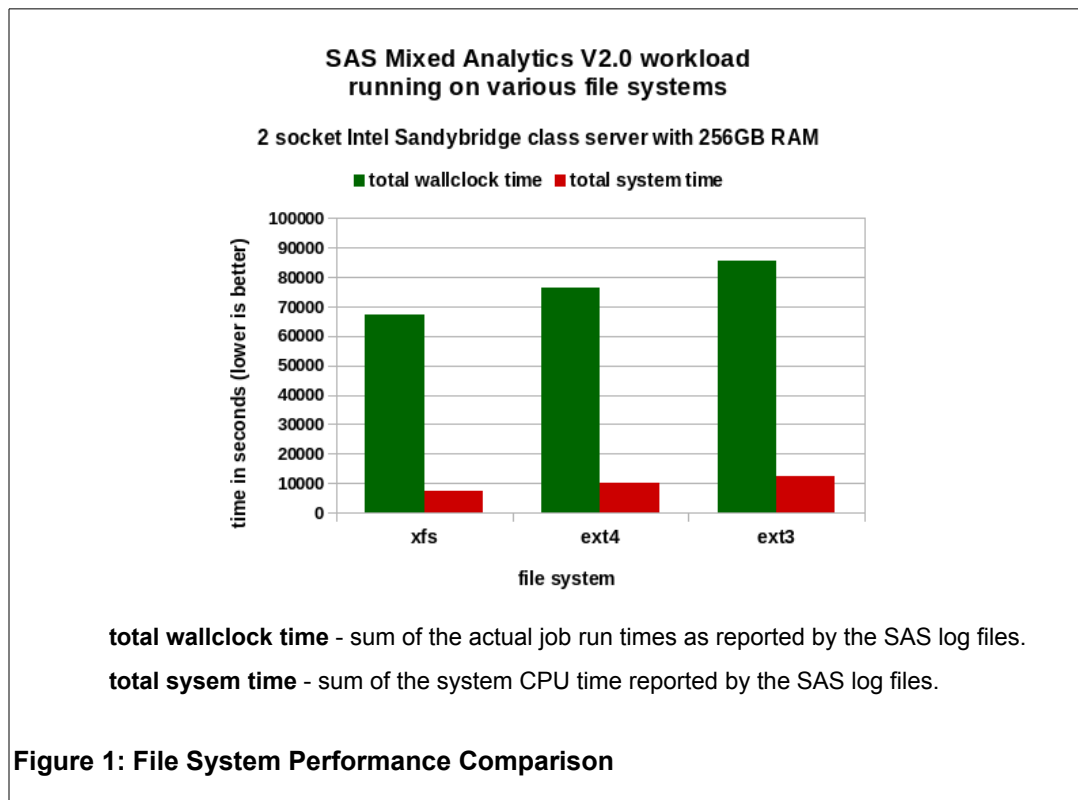
btrfs is still in technological preview and is not ready for production environments.

Below is a File Size & File System Limits table comparing the file size and file system limits of these four file systems.

| | Red Hat Enterprise Linux 6 | | Red Hat Enterprise Linux 7 | |
|------------------|----------------------------|-------------------|----------------------------|-------------------|
| File System Type | Maximum File System Size | Maximum File Size | Maximum File System Size | Maximum File Size |
| XFS | 100 TB | 100 TB | 500 TB | 100 TB |
| ext4 | 16 TB | 16 TB | 50 TB | 16 TB |
| ext3 | 16 TB | 2 TB | 16 TB | 2 TB |
| gfs2 (shared) | 100 TB | 100 TB | 100 TB | 100 TB |

Table 1: File Size & File System Limits

Below is a File System Performance Comparison graph comparing file system performance when executing a SAS Mixed Analytic V2.0 workload on a simple 2 socket Intel server. The I/O stack for each file system was the same.



FILE SYSTEM LAYOUT

SASDATA

Whether you try to consolidate your data on a single large file system or break it up into several smaller ones, you need to weigh the need to handle tasks like backups and maintenance. Breaking file systems up may be the only option if your data sets are expected to exceed the file system limits noted below.

SAS SCRATCH SPACE

Due to the nature of how SAS temp file processing works, it is highly recommended that the SASWORK temp space always be a separate file system. This is regardless of whether you have a stand alone configuration or a shared file system for GRID. While all SAS file systems have a high I/O demands, this file system typically has the highest demand.

SAS UTILITY FILE LOCATION

The UTILLOC option specifies one or more locations for a type of utility file that is introduced as part of the multi-threaded procedures in SAS 9 architecture. By default, these files are put in the same directory as where SASWORK points. From a performance perspective, we strongly recommend that the location for these files are a separate file system from SASWORK as it also has high I/O demands.

THE I/O STACK

The I/O Stack effectively covers the block device underneath the file system all the way down to the presented LUN from the storage. Choosing and configuring your I/O stack is essential for optimum performance. If any one piece of the I/O stack is not configured properly, it can have a significant impact on your performance. Here we will discuss:

- Volume Management (LVM)
- Device Mapper Multipath I/O

VOLUME MANAGEMENT (LVM)

Logical Volume Manager (LVM) is used to create and manage LUNs made up of single and multiple block devices. These can be bottom level devices (*/dev/sdc*) or device mapper presented devices (discussed below). LVM is very useful when you need to construct a LUN bigger than any individual block device. This provides an easy mechanism for creating much larger file systems.

You may be wondering whether you should bother with LVM especially if you would prefer to present a single LUN from your storage array for each file system. The simple answer is both SAS and Red Hat **highly recommend** that you create **several** LUNs and present them to LVM to build your logical volumes. SAS typically recommends at least 8 LUNs since they deploy 1 writer thread per LUN. Some storage arrays perform better with several smaller presented LUNs vs. a single larger LUN. Its best to experiment with your resources first.

LVM also abstracts your devices with names so you don't have to worry about persistent device naming between reboots.

CONCATENATED VS. STRIPED LOGICAL VOLUMES

When you construct a logical volume from a volume group containing multiple LUNs, by **default**, LVM takes each successive LUN and appends it logically to the end of the prior one. This is known as a **concatenated** volume. This behavior makes it easier to remove a LUN from LVM but can be a serious performance bottleneck. The reason is if there isn't enough parallel activity across the file system then only certain LUNs will be performing I/O while others remain idle. If your volume has been created already, the easiest way to see if it's a concatenated volume is to run:

```
# lvs -o name,vg_name,size,attr,lv_size,stripes,stripesize
LV          VG          Attr      LSize  #Str  Stripe
lroot       vgRHEL6      -wi-ao--- 86.71g  1    0
lswap       vgRHEL6      -wi-ao--- 8.79g   1    0
```

In the above example the volume group shows a stripe count (**#Str**) of 1. That means even if you have multiple LUNs in your volume, they are not striped.

To get all your LUNs participating in parallel, tell LVM to create a **striped** volume. This stripes blocks across multiple LUNs (like hardware RAID0) to maximize and balance LUN utilization. This ensures, especially with streaming I/O that all LUNs will participate and dramatically improve performance.

```
# lvs -o name,vg_name,size,attr,lv_size,stripes,stripesize
LV          VG          Attr      LSize  #Str  Stripe
lASUITEinput  vg_sas      -wi-ao--- 150.00g  8    64.00k
lASUITEoutput  vg_sas      -wi-ao--- 150.00g  8    64.00k
lASUITEsaswork vg_sas      -wi-ao--- 150.00g  8    64.00k
```

The value in the **Stripe** column is the amount of data per LUN before I/O moves to the next LUN. The **#Str** tells us how many LUNs make up the logical volume. The effective stripe size is (**#Str x Stripe**). In the above case 8 x 64KB = 512MB I/O request would request I/O in parallel from all the LUNs. While this typically improves performance even when the LUNs are from the same storage array, the huge win is when LUNS from multiple arrays are configured.

By default, the order with which disks are built into LVM stripes is based on the order they were added into the LVM volume group. A recommendation is if you plan to stripe with multiple storage arrays, then add a LUN round robin from each of the arrays.

Below are some important **lvcreate(8)** options when creating striped logical volumes.

-i --stripes Stripes

Gives the number of stripes. This is equal to the number of physical volumes to scatter the logical volume.

-I --stripesize StripeSize

Gives the number of kilobytes for the granularity of the stripes.

The only drawback of using striped volumes is you cannot migrate data off a LUN to free it up. You would have to archive your data first and then tear down the logical volume. We think the performance benefits of LVM striping outweigh this loss of flexibility.

LVM READ AHEAD

SAS file systems need a significantly elevated read ahead value. This is due to the enormous amount of sequential data being read from files. When LVM creates a logical volume, it sets read ahead to automatic. For SAS environments, it's best to set read ahead to a static large value.

To tune read ahead, include this option with the **lvcreate(8)** command or if it was left out, you can change it later with **lvchange(8)**:

-r --readahead {ReadAheadValue|auto|none}

Sets the readahead mode for the given logical volume. The default mode is **auto**. When **ReadAheadValue** is specified the read ahead value becomes static. The default unit is in sectors (512 bytes) but other units like k (kilobytes), m (megabytes) can be used. This value is persistent between reboots.

Read ahead for SAS environments is discussed in depth in the **Additional Tuning Requirements** section.

DEVICE MAPPER MULTIPATH I/O

If **either** your server or storage array has multiple connections/paths (Host Bus Adapters [HBA's] or network devices) to a switch, then your system is a candidate for configuring multipath I/O. We **highly** recommend you configure device mapper multipath I/O (MPIO) to provide fail-over and load balancing support. When properly configured, this provides:

- Continuous availability in case a path to a device fails
- Increased bandwidth support between the server and block device
- Easier manageability by providing a single abstracted device path to a presented LUN
- Having multiple and distributed paths to your storage, especially on larger socket systems, additionally helps reduce some latencies associated with Non Uniform Memory Architecture (NUMA) based systems

MPIO requires the **device-mapper-multipath** package which is typically not installed by default on your system. Additionally, MPIO will not start up unless the **/etc/multipath.conf** file exists at boot up time. Once both of these criteria are met, you will need to reboot your system.

The MPIO driver already has configuration information for several storage arrays. Sometimes configuration file additions and changes need to be made. The easiest way to see the present configuration is to run

```
# multipathd show config
```

This outputs everything the MPIO driver knows about. You can also run the **multipathd** command interactively as shown below.

```
# multipathd -k ""
multipathd> show config
```

To view your multipath devices, you can use the command:

```
# multipath -ll
...
msa13_6 (3600c0ff000db3c10c0f5d34d01000000) dm-7 HP,P2000 G3 FC
size=272G features='1 queue_if_no_path' hwhandler='0' wp=rw
|+- policy='round-robin 0' prio=130 status=active
| | 2:0:0:6 sdj 8:144 active ready running
| | 1:0:0:6 sddb 70:144 active ready running
| | 1:0:3:6 sdel 128:208 active ready running
| | 2:0:7:6 sdcg 69:208 active ready running
|+- policy='round-robin 0' prio=10 status=enabled
| 2:0:2:6 sdah 66:16 active ready running
| 2:0:3:6 sdat 66:208 active ready running
| 1:0:2:6 sddz 128:16 active ready running
| 1:0:7:6 sdgh 131:208 active ready running
...
```

The above snippet example shows a multipath device with 8 paths that are active, 4 of which will actively handle I/O with another 4 paths enabled and ready for fail-over.

DEVICE MAPPER MULTIPATH NAMING

If you look at the above multipath example, you will notice the device name **msa13_6** has been assigned as opposed to something like **mpath[a-z]**. This was accomplished by defining an alias name in the **/etc/multipath.conf** file. Alias definitions allow you to assign logical names to WWID for easier tracking. This is extremely useful for Resilient Storage configurations since you are not guaranteed the same logical name on each node. Below is the config file definition for the above alias:

```
multipaths {
    multipath {
        wwid 3600c0ff000db3c10c0f5d34d01000000
        alias msa13_6
    }
    ...
}
```

KERNEL TUNING FOR SAS

Having SAS run optimally out of the box is an ultimate desire from both SAS and Red Hat's perspective. We aren't quite there yet but as Red Hat Enterprise Linux has evolved and grown, better operating system defaults and tools help achieve optimum performance easier than ever. The primary areas of the kernel needing tuning for SAS environments include:

- CPU performance and power management
- I/O elevators for enterprise workloads
- Virtual Memory subsystem
- File system read ahead for improved sequential read performance
- File system I/O barriers for data integrity during a system crash

In this section, we will discuss how to tune these areas by discussing:

- Tuned with Red Hat Enterprise Linux 6 & 7
- Creating a Custom Tuned Profile
- Additional Tuning Requirements
- Tuning Beyond Tuned
- SAS Tuning File Examples

Work began in RHEL5 which provided the **ktune** service to provide a single means for adjusting a group of tunables. In RHEL6 we introduced **tuned** which provides even more tuning capability. Unfortunately **tuned** doesn't completely tune a RHEL system for SAS but is an excellent way to tune most of what SAS needs with a single command.

TUNED FOR RED HAT ENTERPRISE LINUX 6 (RHEL 6)

In RHEL 6, Red Hat extended the **ktune** utility to include the **tuned** tool which can tune various system parameters based off a profile definition.

For RHEL 6, the most recommended performance profile for file system I/O environments is **enterprise-storage**. This profile tunes:

- Adjusts the I/O elevator to **deadline** (versus CFQ default)
- Alters the power save mode from OnDemand to Performance
- Sets the VM reclaim parameters for **dirty_ratio** back to the RHEL 5 value of 40 (RHEL 6 adjusted default to 20)
- Scheduler tunable quantum back to RHEL 5 default of 10 milliseconds (RHEL 6)
- Default quantum is 4 milliseconds
- Enables Transparent Huge Pages (discussed below)
- Block device and LVM read ahead values increased by a factor of 4
- Remounts the file systems to disable I/O barriers using “**-o barrier=0**” (assumes enterprise class storage with battery backed up controllers). See **/proc/mount** to view the barrier settings on the server.

Early versions of RHEL 6, specifically versions earlier than RHEL 6.2 had improved performance when I/O barriers could be disabled. Significant improvements have been made to barrier code and the difference in performance is typically noise.

The **tuned** tool is **not** installed by default in RHEL6 so you will need to install the following rpms:

```
tuned-*.rpm  
tuned-utils-*.rpm
```

Useful **tuned** commands include:

- To **list** available profiles

```
# tuned-adm list  
Available profiles:  
- throughput-performance  
- desktop-powersave  
- sap  
- server-powersave  
- virtual-guest  
- default  
- virtual-host  
- laptop-battery-powersave  
- spindown-disk  
- latency-performance  
- laptop-ac-powersave  
- enterprise-storage  
Current active profile: default
```
- To **set** a specific profile

```
# tuned-adm profile enterprise-storage  
Stopping tuned: [ OK ]  
Switching to profile 'enterprise-storage'  
Applying deadline elevator: dm-0 dm-1 dm-2 sda sdb sdc sdd ... [ OK ]  
Applying ktune sysctl settings:
```

```

/etc/ktune.d/tunedadm.conf: [ OK ]
Calling '/etc/ktune.d/tunedadm.sh start': [ OK ]
Applying sysctl settings from /etc/sysctl.conf
Starting tuned: [ OK ]

```

- To **show** the active profile

```

# tuned-adm active
Current active profile:enterprise-storage
Service tuned: enabled, running
Service ktune: enabled, running

```

TUNED FOR RED HAT ENTERPRISE LINUX 7 (RHEL 7)

For RHEL 7, the tuned package is installed by default and for Server variant installs, is set to the **throughput-performance** profile by default. This profile effectively replaces **enterprise-storage** which was available in RHEL 6. The tune is effectively the same with two distinct differences.

- I/O barriers are not disabled with **throughput-performance**. The operating system queries the storage and disables barriers if the storage says it's safe to.
- The default I/O elevator is now **deadline** and is not modified by **tuned**.

Note, when you change **tuned** profiles in RHEL 7, the tool does not output anything unlike RHEL 6. Use:

```

# tuned-adm active
Current active profile: throughput-performance

```

to verify the tune has been enabled.

CREATING A CUSTOM TUNED PROFILE

One question which always comes up is what is the best way to handle the additional tuning requirements needed for SAS environments. The easiest solution is to create a custom **tuned** profile. We **highly recommend** this as opposed to modifying an existing profile. The reason is any file installed by RHEL has the potential to be changed from regular updates or errata. Creating your own profile protects you from such updates.

The tuning configuration files for RHEL 6 and RHEL 7 are different and are stored in different directory areas, but the principles for creating a custom tune and tweaking are the same.

RHEL6

1. Go to the tuning profiles directory at **/etc/tune-profiles**. There are sub-directories for each **tuned** profile name
2. Create a new directory that is the name of the profile you want to create. A typical name might be **sas-performance**.
3. Copy the files from the **enterprise-storage** profile directory into the new **sas-performance** directory.

```
# cp /etc/tune-profiles/enterprise-storage/* /etc/tune-profiles/sas-performance
```

4. Edit new file(s) as needed (discussed in the sections below)
5. Enable new profiles

```
# tuned-adm profile sas-performance
```

RHEL7

1. Go to the tuning profiles directory at **/usr/lib/tuned/profiles** . There are sub-directories for each **tuned** profile name.
2. Create a new directory that is the name of the profile you want to create. A typical name might be **sas-performance**.

3. Use **throughput-performance** as the basis for your custom profile
 - a) RHEL 7 supports class based profiles so you can inherit the characteristics of another profile (with an include statement) and then customization it further
 - b) Create a new file in the **sas-performance** directory called **tuned.conf**
 - c) The beginning of the file should include the **throughput-performance** profile with:

```
[main]
include=throughput-performance
```

4. Any changes made beyond the include will override the **throughput-performance** profile
5. Edit new config file as needed (discussed in the sections below)
6. Enable new profile

```
# tuned-adm profile sas-performance
```

ADDITIONAL TUNING REQUIREMENTS

Below are areas which need further tuning.

- Flash Storage (I/O elevators and more)
- I/O Barriers (RHEL 6 only)
- File system read ahead
- Transparent Huge Pages (THP)
- CPU performance and power management

Note: Before you make any changes to a tuned profile, make sure it is not enabled by setting to another profile.

FLASH STORAGE (I/O ELEVATORS AND MORE)

Typically we recommend the default I/O elevator (scheduler) be set to **deadline** for all storage devices making up SAS file systems. In RHEL 6, this tuning was handled directly by **tuned**. In RHEL 7, the default I/O elevator for server configurations is **deadline**. This was and **still** is the recommended setting for **spinning disk media**.

More and more SAS deployments are including **flash** based storage. This type of media typically performs best when their drivers/controllers take full control of I/O scheduling. Such devices should have their I/O elevators set to **noop**. The best way to tune I/O elevators in a mixed storage environment is to use a **udev** rule.

Below is an example of a **udev** rule file from a flash storage vendor.

```
# Use noop elevator for high-performance solid-state storage
ACTION=="add|change",SUBSYSTEM=="block",ENV{ID_VENDOR}=="PURE",ATTR{queue/scheduler}="noop"
# Reduce CPU overhead due to entropy collection
ACTION=="add|change", SUBSYSTEM=="block",ENV{ID_VENDOR}=="PURE",ATTR{queue/add_random}="0"
# Schedule I/O on the core that initiated the process
ACTION=="add|change",SUBSYSTEM=="block",ENV{ID_VENDOR}=="PURE",ATTR{queue/rq_affinity}="2"
```

udev rule files are stored in **/etc/udev/rules.d**. Naming conventions must be **NN-name.rules** where **NN** is a sorting number and name is a name associating what the rules in the file do. Rules are read in sorted order. The above example might be **01-purestorage.rules**.

RHEL6

Since RHEL 6 **tuned** modifies I/O elevators, You need to understand when resources actually get tuned. **udev** rules are executed before **tuned** starts up during boot-up. This means **tuned** will

wipe out some of the **udev** tuning. To solve this, it is recommended to rerun the **udev** rules after **tuned** has been brought up.

A way to get the **udev** rules rerun is to add the following calls in the **start()** function for your profile just before the return call.

```
udevadm control --reload-rules; udevadm trigger
```

A commented out example is shown in the RHEL 6 **SAS Tuning Example** further in this document.

RHEL7

Since **tuned** in RHEL 7 does not touch I/O elevators, simply define a **udev** rule.

*A word of **caution** when defining udev rules ...*

*Both Red Hat and SAS recommend **udev** rules only change characteristics about devices that are directly related to device behavior. That is I/O elevators, interrupt pinning, etc. Tuning such as **readahead** should be removed from **udev** rules since LVM and **tuned** handle them.*

I/O BARRIERS (RHEL 6 ONLY)

In RHEL 6, the **tuned** profile **enterprise-storage** specifically re-mounts all file systems to disable block layer write barriers. Write barriers ensure that certain I/O's make it through the device cache and are on persistent storage. If barriers are disabled on a device with a volatile (non battery backed) write back cache, file system corruption could occur during a crash or power loss.

If you are not sure whether you have battery backed protection then simply remove the calls in **/etc/tune-profiles/sas-performance/ktune.sh** which control it.

- In the **start()** function, remove
`disable_disk_barriers`
- In the **stop()** function, remove
`enable_disk_barriers`

FILE SYSTEM READ AHEAD

SAS file systems need a significantly elevated read ahead tuning value. This is due to the enormous amount of sequential data being read from files.

The block device that is actually mounted to present the file system is the **only** device which needs to be tuned. Tuning any block device underneath that mounted device will do nothing.

There are several ways to tune read ahead and it's important to set this up correctly or elevated read ahead may not occur or the read ahead rate could be altered automatically. Below are the most relevant block device types.

- Logical Volume (LVM) Block Devices
- Simple or Multipath Block Devices
- Third Party Volume Manager Devices

READ AHEAD FOR LOGICAL VOLUME MANAGER (LVM) DEVICES

Earlier, we discussed the use of LVM to build and manage powerful striped block devices to house the various SAS file systems. When an LVM block device is configured, it is imperative to **only** use LVM to set read ahead on that block device. Using other mechanisms can conflict with LVM depending on who is setting the read ahead tune at a given time. The following mechanisms, which can tune read ahead, must **not** be used.

- **Tuned**

To disable read ahead in your **sas-performance** tuned profile:

RHEL 6

- Edit the file **/etc/tune-profiles/sas-performance/ktune.sh**
- The following line from the shell function **start()** should be removed
`multiply_disk_readahead 4`
- The following line from the shell function **stop()** should be removed
`restore_disk_readahead`

RHEL 7

- Edit the file **/usr/lib/tuned/sas-performance/tuned.conf**
- Add a **disk** subsystem entry with the following entry

```
[disk]
devices=!dm-*
```
- This will prevent **tuned** from tuning read ahead for any device mapper device.

- **udev** rules

Make sure no special **udev** rules define read ahead

- Tuning with **/sys/block/*/queue/read_ahead_kb**

Make sure you don't tune any device here either manually or with a script

By default, when LVM creates a logical volume, it sets the read ahead to automatic (**auto**). For SAS environments, we recommend you set read ahead **manually** to a large static value. The default unit value is in 512 byte sectors. You can specify a different unit as noted below. SAS typically sets read ahead in excess of 8 MB (8192).

To tune read ahead, include this option with your **lvcreate(8)** command or if it was left out, you can change it later with **lvchange(8)** :

```
-r --readahead {ReadAheadValue|auto|none}
```

Sets the readahead mode for the given logical volume. The default mode is **auto**. When **ReadAheadValue** is specified the read ahead value becomes static. The default unit is in sectors (512 bytes) but other units like **k** (kilobytes), **m** (megabytes) can be used. This value is persistent between reboots.

For example to change the read ahead value of an already existing LVM volume to 8MB

```
# lvchange --readahead 8m /dev/mapper/vg_sas-1SASDATA
```

To view the read ahead and other relevant configuration for LVM devices use:

```
# lvs -o name,vg_name,size,attr,lv_size,stripes,stripesize,lv_read_ahead
LV      VG      LSize  Attr    LSize  #Str  Stripe  Rahead
lroot   perf82_R7 272.66g -wi-ao--- 272.66g 1      0      auto
lswap   perf82_R7 4.00g   -wi-ao--- 4.00g   1      0      auto
1SASDATA vg_sas    3.00t   -wi-ao--- 3.00t   4      256.00k 8.00m
1SASWORK vg_sas    3.00t   -wi-ao--- 3.00t   4      256.00k 8.00m
1UTILLOC vg_sas    2.00t   -wi-ao--- 2.00t   4      256.00k 8.00m
```

Note that the **Rahead** column in **red** shows explicit read ahead tuning for the SAS logical volumes.

READ AHEAD FOR SIMPLE OR MULTIPATH BLOCK DEVICES

It is possible to build a SAS file system directly on top of a simple or multipath block device. This is typically **discouraged** for bare metal and most virtualized configurations for performance reasons.

There are situations though, when it can be useful in virtualized environments where the host has an already created striped LVM block device that is passed into the guest. This configuration is typically the exception rather than the norm. This tuning is discussed in the Virtualization section.

READ AHEAD FOR THIRD PARTY VOLUME MANAGER DEVICES

If you are using a third party volume manager, then consult their documentation regarding setting persistent read ahead. In such a case you follow all the tuning and setup instructions in the LVM section above that are not LVM specific. The volume manager must be the only tuner for read ahead.

TRANSPARENT HUGE PAGES (THP)

Transparent Huge Pages (THP) dynamically migrates standard x86_64 4KB pages to 2MB pages when available. This scanning and conversion is not typically needed in SAS environments when most of memory is devoted to the file system page cache. Both the **enterprise-storage** and **throughput-performance** profiles enable Transparent Huge Pages.

If you are running an older version of RHEL, specifically RHEL 6.0 or 6.1, then we **highly** recommend you **disable** Transparent Huge Pages. This feature creates much less overhead on SAS performance since RHEL 6.2 and RHEL7 and does not need to be disabled. If you are not sure, It's best to experiment with your workload.

To adjust tuning for a new profile **sas-performance**:

RHEL 6

- Edit the file `/etc/tune-profiles/sas-performance/ktune.sh`
- The shell function **start()** contains the following line:

```
set_transparent_hugepages always
```
- Change **always** to **never** to disable Transparent Huge Pages

RHEL 7

- Edit the file `/usr/lib/tuned/sas-performance/tuned.conf`
- Add a **vm** subsystem with the following entry

```
[vm]
transparent_huge_pages=never
```
- The **never** value will disable Transparent Huge Pages

When you enable this profile, Transparent huge pages will be disabled. You can verify by looking at:

```
# cat /sys/kernel/mm/transparent_hugepage/enabled
always madvise [never]
```

CPU POWER MANAGEMENT

Most of the CPU's of systems available for the past 5 years have the ability to reduce their power footprint by slowing down and even shutting down various subsystems when the system is under light load. The problem is that the performance of some workloads suffer when not a lot of compute cycles are needed. I/O and network loads fall into this category often. The workloads encounter delays due to the time it takes to bring the CPU's fully back online. To eliminate this latency and performance effect, **tuned** provides a mechanism to tell the operating system not to allow CPU's to migrate into deeper power management c states. To keep a CPU running at it's peak continuous rate, two things must be done.

First, the BIOS in your system needs to be set to **OS_CONTROL**. This will give the CPU drivers in the kernel the greatest ability to regulate CPU performance. Second, you need to modify your custom **tuned** configuration to tell the kernel you want the best peak continuous performance.

To adjust tuning for the new profile **sas-performance**:

RHEL 6

- Edit the file **/etc/tune-profiles/sas-performance**
- The shell function **start()** contains the following line:

```
set_cpu_governor performance
```
- **Insert** the following line after it:

```
/usr/libexec/tuned/pmqos-static.py cpu_dma_latency=1
```
- The shell function **stop()** contains the following line:

```
restore_cpu_governor
```
- **Add** the following line after it:

```
/usr/libexec/tuned/pmqos-static.py disable
```

RHEL 7

- Edit the file **/usr/lib/tuned/sas-performance/tuned.conf**
- Add a **cpu** subsystem entry with the following

```
[cpu]
force_latency=1
governor=performance
energy_perf_bias=performance
min_perf_pct=100
```

This will keep the CPU from going into deep power saving C states and in certain processors, keep the processor frequency at maximum. Once the profile is enabled, the **turbostat(8)** (Intel only) and **powertop(8)** tools can be used to monitor processor C state behavior. Note that more recent versions of **turbostat** require the **-debug** flag to view C state information. Check the manual page for further details.

Below is a RHEL 7 example of **turbostat** reporting summary information on an idle system before tuning:

```
# turbostat -S
Avg_MHz %Busy Bzy_MHz TSC_MHz SMI CPU%c1 CPU%c3 CPU%c6 CoreTmp Pkg%pc3 Pkg%pc6 time
5 0.25 1993 2933 0 0.42 4.15 95.18 44 0.00 0.00 5**
6 0.28 2031 2933 0 0.39 1.01 98.32 45 0.00 0.00 5**
5 0.27 1993 2933 0 0.73 1.97 97.03 45 0.00 0.00 5**
...
```

Note that the average processor speed **Bzy_MHz** on the idle system is a relatively low value frequency and **CPU%c6** tells us that the average state of the cores is over 95% in C state 6. Of note, on this particular hardware the time it takes to get from C state 6 to C state 0 is 200 usec.

Once the **sas-performance** profile is enabled, rerunning **turbostat** on an idle system will report something similar to this, which shows average processor speed **Bzy_MHz** at full speed and C states capped at 1 **CPU%c1**. Of note, on this particular hardware, migrating from C state 1 to C state 0 is only 3 usec.

```
# turbostat -S
Avg_MHz %Busy Bzy_MHz TSC_MHz SMI CPU%c1 CPU%c3 CPU%c6 CoreTmp Pkg%pc3 Pkg%pc6 time
6 0.18 3200 2933 0 99.82 0.00 0.00 48 0.00 0.00 5**
4 0.13 3200 2933 0 99.87 0.00 0.00 48 0.00 0.00 5**
4 0.13 3200 2933 0 99.87 0.00 0.00 48 0.00 0.00 5**
...
```

TUNING BEYOND TUNED

Tuned is a very useful tool for configuring your system for SAS specific needs. If your configuration needs additional tuning, it is best to contact SAS and Red Hat support to discuss your specific needs.

SAS TUNING FILE EXAMPLES

RHEL 6 /etc/tune-profiles/sas-performance/ktune.sh

```
#!/bin/sh

. /etc/tune-profiles/functions

start() {
    # Set CPU and power management to performance
    set_cpu_governor performance
    /usr/libexec/tuned/pmqos-static.py cpu_dma_latency=1

    # Disable Transparent Huge Pages
    set_transparent_hugepages never

    # Comment out to Not disable I/O Barriers
    disable_disk_barriers

    # Elevate readahead (disabled)
    #multiply_disk_readahead 4

    # EXAMPLE of rereading udev rules after tuned is finished
    # udevadm control --reload-rules ; udevadm trigger

    return 0
}

stop() {
    # Restore CPU/Power management settings
    restore_cpu_governor
    /usr/libexec/tuned/pmqos-static.py disable

    # Restore Transparent Huge Pages
    restore_transparent_hugepages

    # Comment out to Not disable I/O Barriers
    enable_disk_barriers

    # Reset readahead settings (disabled)
    #restore_disk_readahead

    return 0
}

process $@
```

RHEL 7 /usr/lib/tuned/sas-performance/tuned.conf

```
# Additional Tuning for SAS

# Profile inherits throughput-performance profile
[main]
include=throughput-performance

[cpu]
force_latency=1
governor=performance
energy_perf_bias=performance
min_perf_pct=100

[vm]
transparent_huge_pages=never

# Disable tuning read ahead on device mapper devices
```

```
[disk]
devices=!dm-*
```

SAS WITH RED HAT VIRTUALIZATION

With the ever increasing need for basic virtualization and cloud computing for provisioning system resources it became important early on to understand the best way to configure those environments for SAS. With respect to basic virtualization, we have characterized SAS performance in KVM environments. Since we had limited hardware resources and the SAS workload requires a large amount of compute, memory, and storage bandwidth, we only focused on a single large virtualized KVM guest.

Areas discussed include:

- Host Configuration Requirements
- SAS Guest Creation

Make sure you fully read **Setup and Tuning for All SAS Environments** as the information which follows builds on that information.

HOST CONFIGURATION REQUIREMENTS

Regardless of whether you deploy your guests on RHEL 6 or 7 hosts, the basic setup of the host is the same. The I/O and file system requirements are the same as with bare metal environments. The areas specific to virtualized environments which need discussion include:

- Tuned profile for KVM Host Environment
- Kernel Shared Memory (KSM)

TUNED PROFILE FOR KVM HOST ENVIRONMENT

The best **tuned** profile for the host environment is **virtual-host**. This profile assumes that SAS will never be running directly on the host. It focuses on allowing large running processes (KVM guests) and tunes the kernel to favor keeping those processes from paging out when memory resources are low. Additionally it makes sure Transparent Huge Pages (THP) are enabled. This is important because the merging of small **qemu-kvm** (KVM) process pages to 2MB pages shows a measurable performance improvement of **10%** when running SAS workloads.

virtual-host is a great starting point for tuning key components which are essential for good SAS performance. For peak performance, you will want to create a custom profile which tunes power management components. As noted in the basic setup discussion, you will want to create a custom **tuned** profile. We recommend calling it **sas-virtual-host**. Follow the instructions above to optimize **CPU power management**. Remember this change will keep your whole host platform running at peak performance for all guests.

KERNEL SHARED MEMORY (KSM)

This feature, which is available on RHEL KVM host environments has been shown to only benefit Microsoft KVM guest environments. It should be disabled for SAS on RHEL environments.

This can be disabled by running:

RHEL 6

```
# service ksm stop
# service ksmtuned stop
# chkconfig off ksm
# chkconfig off ksmtuned
```

RHEL 7

```
# systemctl stop ksm.service ksmtuned.service
# systemctl disable ksm.service ksmtuned.service
```

SAS GUEST CREATION

The creation of the KVM guests for SAS requires deciding how the resources of your host are going to be divided up. This section will discuss the best practices specific to SAS workloads.

STORAGE PRESENTATION

One of the most important aspects of configuring virtualized environments for SAS is to get the I/O storage for the SAS specific file systems correct.

There are two key ways to present storage to your virtualized guest.

- **PCI pass through devices**

Passing PCI devices through to the guest yields the least overhead compared to bare metal but provides the least flexibility. Expect overhead of around **2-5%** compared to bare metal. With these devices directly passed into the guest, it is possible to use the Logical Volume manager to build larger block devices out of smaller ones. As discussed in the Basic Setup section, the virtues of striped logical volumes holds true here as well.

- **virtIO**

If storage is going to be handled via virtIO, then it is extremely important to define that storage outside of the KVM guest. For example, if you want to take advantage of LVM and multi-LUN striping, you need to create that LVM block device first. RHEV-M management tools do not provide the functionality to create striped volumes.

Performance characterization has shown that SAS performs best with the very latest version of RHEL 7 host and KVM guest bits. For several years, virtIO performance running SAS workloads has had a **15%** overhead compared to bare metal. Recent enhancements in RHEL 7.1 have reduced that overhead to around **7%**.

If virtIO is to be used, the following is highly recommended. For each virtIO block device handling a SAS file system:

- Make sure the block device is a **raw** device not QCOW2. QCOW2 cannot handle the vast I/O requirements of SAS.
- Cache mode should be set to **none** which makes sure data written from the virtualized environment is committed to storage
- I/O mode should be set to **native** not threads.

KVM GUEST MEMORY ALLOCATION

Normally when KVM guests start, the KVM process allocates memory only on demand. SAS guests typically require significant amounts of memory because they like to make use of the page cache. We recommend forcing the guest to pre-allocated memory resources up front. This allows the host to effectively allocate Transparent Huge Pages quickly as well as reduce performance stutter early in the life of the guest.

PINNING KVM GUESTS

If your KVM guest is small enough to fit in a sub set of NUMA nodes on your host server, consider pinning the guest for better and more predictable performance.

TUNED PROFILE IN A VIRTUALIZED GUEST

For both RHEL 6 and 7, the **tuned** profile **virtual-guest** is the best tuning profile to start with. This profile is similar to **throughput-performance**. A custom profile is still needed for the environment running SAS. We recommend calling it **sas-virtual-guest**. The recommendations in **Additional Tuning Requirements** apply here as well however the following must be understood.

- I/O barriers should be left **enabled**.
- CPU performance and power management tuning is not needed and should not be added to the config file. That is handled by the host.
- If virtIO block devices are simply block devices to be managed by a volume manager (LVM or third party) inside the guest, then volume creation as well as read ahead tuning is setup identically to what is documented for **Setup & Tuning for All SAS Environments** as defined in **File System Read Ahead**.
- If virtIO block devices are **actually** LVM logical volumes defined from the host, then read ahead needs to be configured differently. This is because the guest has **no idea** LVM is even involved. In such a case **tuned will** be used to set readahead.

To adjust tuning for a new profile called **sas-virtual-guest**:

RHEL 6

- Edit the file **/etc/tune-profiles/sas-virtual-guest/ktune.sh**
- The shell function **start()** contains the following line:

```
multiply_disk_readahead 4
```
- Tuning readahead in RHEL 6 **tuned** is by a multiplier. By default, simple LUN readahead values are 128KB. The **tuned** multiplier elevates it 4 times to 512KB. SAS typically wants them to be between **8192 KB** and **16384 KB**
- To elevate them that much, **change** the multiplier from 4 to **64** or **128**.

RHEL 7

- Edit the file **/usr/lib/tuned/sas-virtual-guest/tuned.conf**
- Add a **disk** subsystem entry with the following entry

```
[disk]
readahead=>8192
```
- The unit is in KB by default.
- SAS typically wants them to be between **8192 KB** and **16384 KB**.
- The => instructs **tuned** to elevate a block device if the readahead value is not already above **8192**. If you have a LUN whose default readahead value is already elevated then tuned will not down tune it if its value is already higher than this value.

When you enable this profile, the read ahead values will be elevated. You can inspect readahead values with the **blockdev** command.

SAS GRID WITH RED HAT SHARED FILE SYSTEMS

Presently the best solution for entry level sized clusters for SAS Grid is to use Red Hat Resilient Storage (GFS2 clusters). This cluster specifically requires the GFS2 file system as the shared file system. This environment supports 2 to 16 nodes in a cluster. GFS2 can support a 100 TB file system. The maximum file size is 100 TB.

Presently SAS GRID is not supported with either Red Hat Gluster Storage or Red Hat CEPH Storage environment.

Other shared file systems are supported on Red Hat Enterprise Linux, but are not discussed here. More information can be found on SAS's website by reading "A Survey of Shared File Systems".

RED HAT RESILIENT STORAGE (GFS2 CLUSTERS)

As early as 2011, SAS began working with Red Hat to determine if a SAS Grid environment could be deployed with Red Hat Resilient Storage (formerly known as GFS2 clusters). There were already a few customers using this clustered environment with mixed success. SAS created a Grid workload called Calibration to validate whether the environment could handle the rigors of SAS Grid. This calibration workload is a meta-data intensive Grid workload broken up into 750 SAS jobs. Originally, SAS was unable to get reliable and repeatable performance results. Red Hat performance engineering and SAS worked closely with the Red Hat cluster and GFS2 file system team and after 18 months were able to validate and get SAS's acceptance that Red Hat Resilient Storage performed well.

SUPPORTED VERSIONS OF RED HAT ENTERPRISE LINUX

RHEL 6

Red Hat Resilient Storage 6.4 is the first update version of RHEL 6 which officially supports SAS Grid Manager. While there may have been attempts with customers using the environment in earlier versions of RHEL, neither Red Hat nor SAS recommend it. This is because the changes not only improved stability, but also performance and scalability. Due to the extensive changes in RHEL 6.4, engineering will not back port these fixes to earlier versions of RHEL 6. Additionally, earlier environments attempting to migrate forward to RHEL 6.4 or later should archive and rebuild their file systems. This is because one area heavily addressed was file system fragmentation.

RHEL 6.5 introduced additional GFS2 file system enhancements which include support for the Orlov allocator. Testing showed a **9%** improvement over RHEL 6.4 using the same 4 node cluster.

Please note that all validation on RHEL6 was done with the classic CMAN based cluster product

RHEL 7

Recently, Red Hat worked to get SAS Grid running on RHEL 7. We are pleased to announce that SAS Grid Manager is now supported on RHEL 7.1. At this time, the stability and performance enhancements required to support SAS in RHEL 7.1 have not been back ported to RHEL 7.0.

CLUSTER CONSIDERATIONS

ARCHITECTURE REVIEW

Due to the complexities of a cluster, both Red Hat and SAS require initiating a SAS Grid as well as RHEL Resilient Storage Architecture Review. This is to make sure that the customers functional configuration and performance expectations can be met. Additionally these reviews, along with this guide, help the customer make better decisions about deploying their configuration.

Additional information about the RHEL Resilient Storage Architectural Review can be found at "What information is required for an Architecture Review of Red Hat Enterprise Linux High Availability and/or Resilient Storage..."

STRETCH CLUSTERS

Both the Cluster Logical Volume daemon (**clvmd**) and GFS2 file system are not supported in a stretch cluster configuration and thus stretch clusters cannot be configured for SAS.

NUMBER OF NODES

Red Hat Resilient Storage supports configurations between 2 and 16 nodes. Red Hat has run with as many as 6 nodes, SAS has validated with 4, 8 and 12 nodes. Testing has been done with iSCSI as well as fiber channel attached storage.

Red Hat Resilient Storage is an excellent cluster file system for small configurations. It offers great performance and low end scalability when properly configured.

DEDICATED CLUSTER INTERCONNECT NIC

Heavy metadata traffic in a clustered environment generates significant cluster interconnect traffic. This traffic is typically in the form of several small network messages. While it is possible to share the network device, it is highly recommended to keep it isolated.

Additionally we highly recommend the cluster interconnect host names be defined in each `/etc/hosts` file to remove any chance of latencies with DNS look-ups.

EXCELLENT I/O CONNECTIVITY & PERFORMANCE

SAS processing puts a huge demand on the I/O subsystem. It is extremely important to make sure you have excellent connectivity from many / all sockets of your servers. When multipath I/O is available it should be utilized.

SAS SCRATCH SPACE CONSIDERATIONS

Some SAS Grid configurations may want to put their scratch space (SASWORK) scratch space on the same file system as their primary data. This is usually for convenience so the IT group only has to manage persistent data via the shared file system. SASWORK space **must** be on a **separate** file system from SASDATA. One of our validation tests, is to run multiple iterations of SAS Calibration. After each iteration we save the output files on the file system. This file system aging test validates how well the files are being allocated on SASDATA as well as how the file system performs as it gets filled up.

Testing with a separate SASDATA shared file system showed **predictable and repeatable** performance even when the file system was at 85% capacity.

Testing with the SASWORK directory on the SASDATA file system showed that even after the first iteration, high levels of file fragmentation had occurred. This continued to worsen with each iteration. File fragmentation hurts I/O performance because the storage needs to seek to different parts of the storage to find the next extent of data. This can be a huge performance loss especially with spinning disk drives. The kernel has to initiate many more smaller I/O requests regardless of disk type.

LOCAL FILE SYSTEMS

The only time SAS prefers SASWORK to be part of a shared file system is when a SAS application **requires** it for fail over, recovery or rescheduling. This is done by the SAS user adding checkpoint/restart SAS statements and logic to the SAS job. It is not based on a SAS product.

Customers who seek even more performance at the cost of losing re-animation ability and additional IT management can create local SASWORK file systems on each node. There is no question that a local SASWORK file system (when given the proper I/O resources and configuration) can outperform a shared file system. The local file system should be XFS. Again, you have to weigh the pros and cons.

DISTRIBUTED LOCK MANAGER (DLM) TUNING

RHEL 6

There are three tunables which improve DLM look-up performance. The three hash tables are:

```
DLM_LMBTBL_SIZE
DLM_RSBTL_SIZE
DLM_DIRTBL_SIZE
```

These tunables live in the config file `/etc/sysconfig/cman`. By default they are 1024 but should be elevated to their maximum (16384).

This tuning must be done **before** the shared file systems are mounted.

RHEL7

There is no DLM tuning for RHEL 7. The DLM hash table was rewritten for much better scalability.

RED HAT GLUSTER STORAGE (RHGS)

Red Hat's performance engineering team has made several attempts over the years to get the SAS Grid Calibration workload running in a Gluster environment. Each time we get closer, but the environment is still not suitable for the type of file system and I/O calls that SAS issues.

Today, Gluster storage is not an option today for SAS Grid.

RED HAT CEPH STORAGE (RHCS)

SAS specifically requires a POSIX compliant file system. Presently the CEPH engineering group is working on the cephfs interface which can provide that interface.

Today, CEPH storage is not an option for SAS Grid.

ACKNOWLEDGMENTS

Many thanks to individuals at both SAS and Red Hat for providing feedback and review of this document.

RECOMMENDED READING

- "Optimizing SAS for Red Hat Enterprise Linux (RHEL) 6 & 7"
This version of the paper will continue to evolve and be updated in the future. It is available on both the Red Hat and SAS web sites.
<https://access.redhat.com/articles/1593123>
http://support.sas.com/resources/papers/proceedings11/342794_OptimizingSASonRHEL6and7.pdf
- "What information is required for an Architecture Review of Red Hat Enterprise Linux High Availability and/or Resilient Storage..."
<https://access.redhat.com/solutions/125153>
- "A Survey of Shared File Systems"
http://support.sas.com/rnd/scalability/papers/SurveyofSharedFilepaper_20131010.pdf

CONTACT INFORMATION

Your comments and questions are valued and encouraged. Contact the author at:

Barry Marson
Red Hat, Inc.
bmarson@redhat.com

SAS and all other SAS Institute Inc. product or service names are registered trademarks or trademarks of SAS Institute Inc. in the USA and other countries. ® indicates USA registration.

Linux is a registered trademark of Linus Torvalds. Red Hat, Red Hat Enterprise Linux and the Red Hat "Shadowman" logo are registered trademarks of Red Hat, Inc. in the United States and other countries.

Intel, the Intel logo and Xeon are registered trademarks of Intel Corporation or its subsidiaries in the United States and other countries.

Other brand and product names are trademarks of their respective companies.