SES Tuning

Agenda

- Welcome and Introductions
- Why Tuning Is Important – Purpose-driven storage
- Architecture Best Practices
- Tuning Approach and Process – Follow my changes (*)
- Planning for Performance
- Tuning Specifics
- Case Study Examples
- Wrap Up – Feedback on what's working and what's needed

Why Tuning

Not just software configuration

Everything needs to work together to achieve high performance

Hardware, Software, Operating System, Patches, Network

Architecture Best Practices (5 min)

Node-specific Recommendations

- Gateway Nodes - #
- Monitor Nodes - #
- Storage Nodes –
  - Memory
  - Disk type ratio – SSD to Spinner

Networking Recommendations

- High-speed interconnect

Hardware, Software, Operating System, Patches, Network

Tuning General Steps

1. Rule out general issues.
2. Specify the problem that needs to be solved.
3. In case the degradation is new, identify any recent changes to the system.
4. Identify why the issue is considered a performance problem.
5. Specify a metric that can be used to analyze performance. For example, this metric could be latency, throughput, the maximum number of users that are simultaneously logged in, or the maximum number of active users.
6. Measure current performance using the metric from the previous step.
7. Identify the subsystem(s) where the application is spending the most time.
8. Monitor the system and/or the application and Analyze the data; categorize where time is being spent.
9. Tune the subsystem identified in the previous step.
10. Re-measure the current performance without monitoring, using the same metric as before.
11. If performance is still not acceptable, start over with Step 3.

Tuning Approach and Process

Changes might require re-running the tuning steps – be proactive

Applying patches
Adding gateways
Adding storages nodes and/or disks

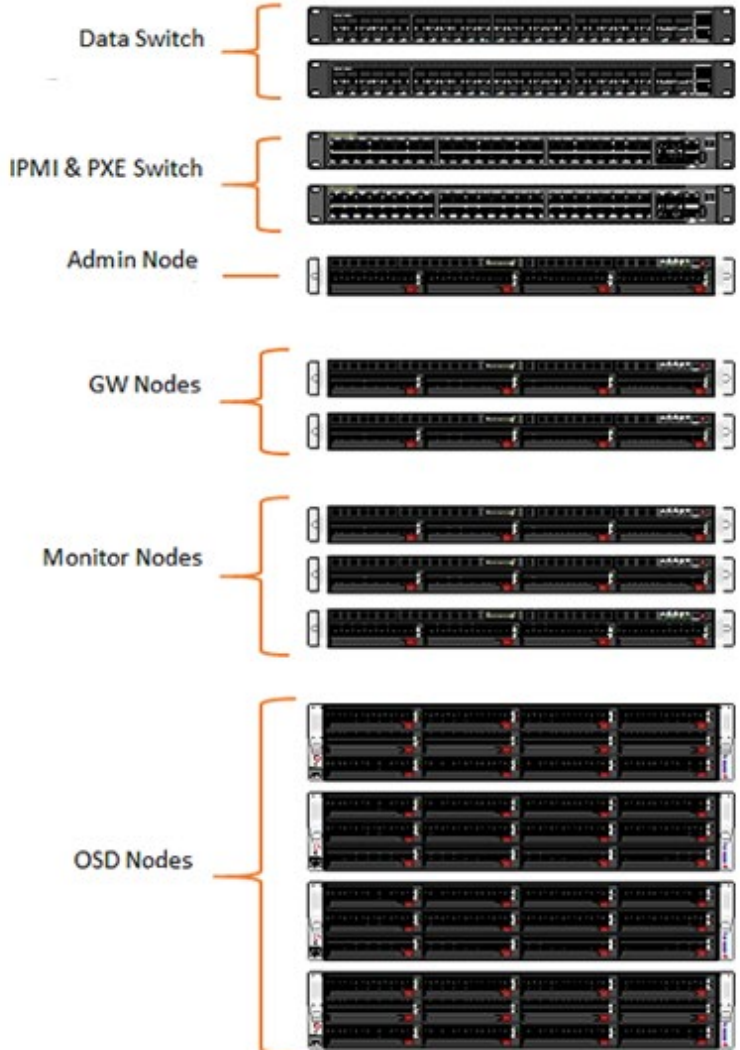
Tuning Approach and Process

Workload measurement

- Measurable and repeatable

Tools

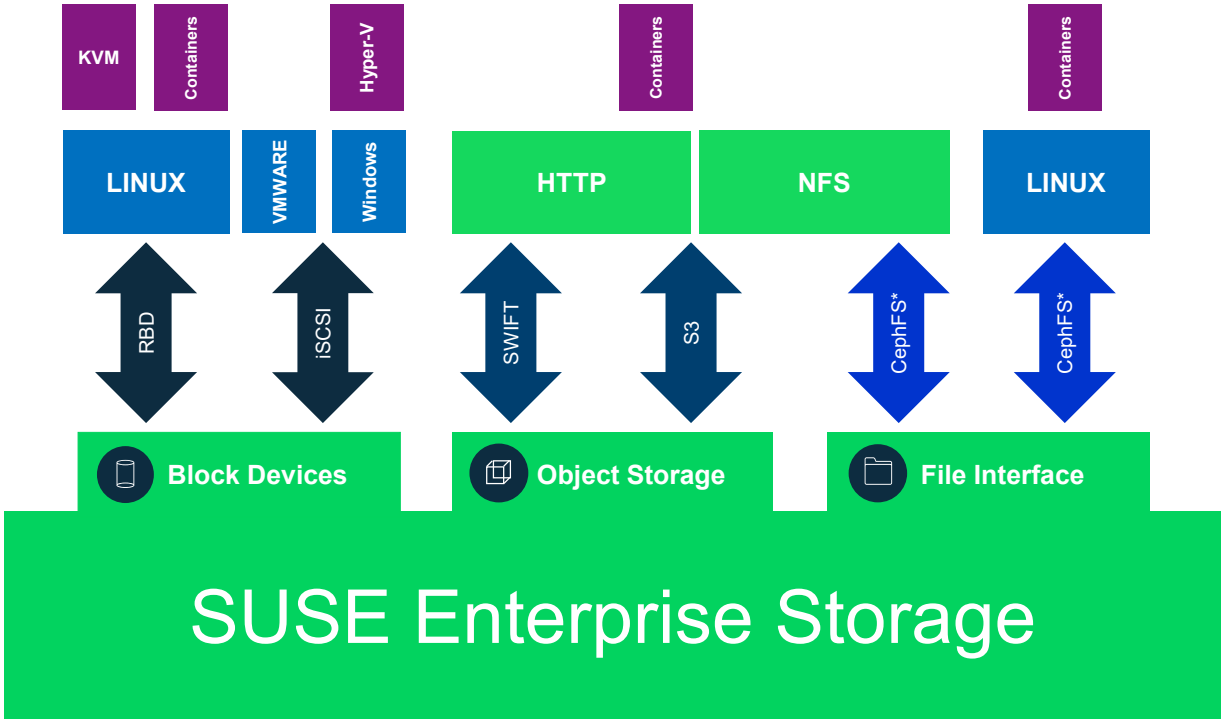
- OpenAttic – Ceph Manager
- Core Kernel/OS Tools


Data Switch
IPMI & PXE Switch
Admin Node
GW Nodes
Monitor Nodes
OSD Nodes

SUSE

ceph

Layering on the Environments



SES Cluster Settings

Jumbo Frames

Make sure jumbo frames are enabled on all devices in the communication chain. This includes Network Interface Card settings on the VMware server, all cluster nodes and any switches in-between.

To check / set this on SUSE Linux Enterprise Servers, verify the "MTU=' '" line for the relevant network interfaces in "`/etc/sysconfig/network/ifcfg-<dev>`".

To verify that Jumbo frames are properly configured, run a ping from the client to any one of the cluster nodes as follows:

```
:~ # ping -M do -s 8972 $insert_destination_IP
```

CPU C-States

In the BIOS of all the storage nodes, ensure that the servers are configured for maximum performance rather than efficiency.

On the SUSE Enterprise Storage servers, install the "`tuned`" package and configure it for "`throughput-performance`" by taking the following steps:

Install tuned with: `zypper in -y tuned`
Edit "`/etc/tuned/active_profile`" and add the line:
`throughput-performance`

Enable and start the tuned service: `systemctl enable tuned.service && systemctl start tuned.service`
Verify that the profile is properly in use by running: `tuned-adm active`

iSCSI GW Nodes

Values below 256 or 512

```
tpg_default_cmdsn_depth = 512 (64)

backstore_hw_queue_depth = 512 (Default 128)

backstore_queue_depth = 512 (Default 128)
```

RBD Image Object-Size

The default object-size for RBD images is 4MB. Because VMware limits the iSCSI I/O to a maximum of 512k, images to be used against VMware should be created with an object size of 1MB instead. Reducing the object size too much (e.g., matching the VMware maximum size of 512k) will result in a much higher number of RADOS objects per RBD.

This can be changed during the creation of the image

- **Via openATTIC**: Adjust the "`Object-size`" parameter
- **Via the command line**:

```
rbd create $pool_name/$image_name --size xxxxM/G/T --object-size 1M
```

VMWare Host Settings

Max Commands

On the initiators (VMware hosts):

```
MaxCommands = 512 (Default 128)
```

See the VMware documentation for details on how to change the setting.

ATS Locking

1. On the VMware server, make sure only ATS locking is enabled on the relevant Volume(s). To verify current settings on the VMware server, run:

```
:~] esxcli storage vmfs lockmode list
Volume Name     UUID                                 Type    Locking Mode  ATS
Compatible  ATS Upgrade Modes  ATS Incompatibility Reason
--------------  -----------------------------------  ------  ------------  --------------  -
---------------  --------------------------
datastore_ssd   5ae70fae-4f913adc-77aa-ac1f6b6442f0  VMFS-
6  ATS+SCSI              false  None               Device does not support ATS
datastore_spin  5ae723bd-a64fe2fc-bb36-ac1f6b6442f0  VMFS-
6  ATS+SCSI              false  None               Device does not support ATS
SESRBD        5afd58c3-13f582a6-596a-ac1f6b6442f0  VMFS-6  ATS                    true  No
upgrade needed
```

In the example above, the `SESRBD` volume shows as only having ATS locking enabled. For more information on how to upgrade volumes to ATS only, see the relevant VMware documentation.

A PTF (Program Temporary Fix) is needed on the iSCSI GW nodes to disable LIO SCSI Persistent Reservations. Please contact SUSE Technical Support to obtain this PTF, which will then need to be installed on all the iSCSI GW nodes after the step above. For details on how to properly apply a PTF, see TID 7016640 - Best practice for applying Program Temporary Fixes (PTFs).

MAX IO Size

By default, ESXi restricts iSCSI I/O to a maximum of 128k. This should be increased to the maximum of 512k. To list the current setting, run the following from the VMware server:

```
:~] esxcli system settings advanced list -o /ISCSI/MaxIoSizeKB
   Path: /ISCSI/MaxIoSizeKB
   Type: integer
   Int Value: 512
   Default Int Value: 128
   Min Value: 128
   Max Value: 512
   String Value:
   Default String Value:
   Valid Characters:
   Description: Maximum Software iSCSI I/O size (in KB) (REQUIRES REBOOT!)
```

In this example, the value is already increased to the maximum of 512k: "`Int Value: 512`"

To change the value from the default of 128k, run the following from the VMware console:

```
:~] esxcli system settings advanced set -o /ISCSI/MaxIoSizeKB -i 512
```

As stated, you must reboot the VMware host for the setting to take effect.

RR Multipath Policy

By default, VMware uses the Most Recently Used (MRU) multipath policy. The RR policy, however, allows the initiator to fully utilize the maximum iSCSI session queue depth across all paths. To see the current MPIO settings on the VMware server, run:

```
:~] esxcli storage nmp device list

   Device Display Name: SUSE iSCSI Disk (naa.60014057ea3725b57a242008f1ee86c4)
   Storage Array Type: VMW_SATP_ALUA
   Storage Array Type Device Config: {implicit_support=on; explicit_support=on;
explicit_allow=on; alua_followover=on; action_OnRetryErrors=on; {TPG_id=0,TPG_state=AO}}
   Path Selection Policy: VMW_PSP_RR
   Path Selection Policy Device Config: {policy=rr,iops=100,bytes=10485760,useANO=0;
lastPathIndex=1: NumIOsPending=0,numBytesPending=0}
   Path Selection Policy Device Custom Config: iops=100
   Working Paths: vmhba64:C1:T0:L0, vmhba64:C0:T0:L0
   Is USB: false
```

In the above example above, the policy is already correctly set to use RR: "`Path Selection Policy: VMW_PSP_RR`"

For more information on how to see/modify VMware multipath policy settings, see VMware KB article 1017760.

Disable ATS for Heartbeat I/O

By default, VMware 5.5 Update 2 and later uses ATS for heartbeat I/O for VMFS 5 and higher datastores.

For details on how to view and configure VMware to use SCSI write operations instead for this, see the VMware KB article 2113956.

Veeam Backup

Pool Type

Consider using Replicated Pools only for the RBD (Rados Block Device) images, since the overhead involved with Erasure Coded Pools and the sequential type of backup / restore workload can result in a fairly large performance difference.

Increase the Windows iSCSI I/O Size

It is necessary to edit the registry of the Windows server. After this change, the server must be rebooted. To change this setting, edit the following key:

"[HKEY_LOCAL_MACHINE\SYSTEM\CurrentControlSet\Control\Class\{4d36e97b-e325-11ce-bfc1-08002be10318}\<MS iSCSI Initiator Instance Number>\Parameters]"

Note that the "...\<MS iSCSI Initiator Instance Number>\..." part of the registry link above will be something like "...\0001\..." or "...\0003\..." etc.

To ensure that the correct key is being changed, expand the "PersistentTargets" key below the "Parameters" entry and verify that the targets listed are the correct targets as exported via iSCSI.

Change the "MaxTransferLength" dword value to "0x00080000" (512K). It can also be tested with a value of "0x00100000" (1M).

Windows Multipath I/O (MPIO)

Using multiple iSCSI Gateways to export the target will allow using MPIO on the Windows client. Information on how to configure MPIO on Windows can be found in the online Microsoft Documentation, as described here.

To verify that MPIO is configured properly, run the example command below and then ensure that the target disk(s) are listed with multiple paths in the resulting "mpio.conf" file):

```
mpclaim.exe -v mpio.conf
```

Veeam Specific Settings

- "Enable parallel processing" should be enabled.
- "Enable storage latency control" should be disabled.
- For more information on these settings, see the Veeam online documentation.

- "Use per-VM backup files" should be enabled.
- For more information on this setting, see the Veeam user guide for VMware.

- The number of parallel tasks should be adjusted appropriately for the environment, to allow an optimal number of Virtual Machines to be backed up in parallel.
- For more information, see the Veeam User Guide for VMware vSphere.

Increase the Windows iSCSI I/O Size

It is necessary to edit the registry of the Windows server. After this change, the server must be rebooted. To change this setting, edit the following key:

"[HKEY_LOCAL_MACHINE\SYSTEM\CurrentControlSet\Control\Class\{4d36e97b-e325-11ce-bfc1-08002be10318}\<MS iSCSI Initiator Instance Number>\Parameters]"

Note that the "...\<MS iSCSI Initiator Instance Number>\..." part of the above registry link will be something like "...\0001\..." or "...\0003\..." etc.

To ensure that the correct key is being changed, expand the "PersistentTargets" key below the "Parameters" entry and verify that the targets listed are the correct targets as exported via iSCSI.

Change the "MaxTransferLength" dword value to "0x00080000" (512K). It can also be tested with a value of "0x00100000" (1M).

SUSE SES Tuning