SUSE Blog (/c/blog/)

Azure Shared Disks with “SLES for SAP / SLE HA 15 SP2”

By: Zhiqiang(Roger) Zhou (https://www.suse.com/c/author/zzhou/) | 546 views

Share

Microsoft Azure Shared Disks now supports SUSE Linux Enterprise Server for SAP Applications and SUSE Linux Enterprise High Availability Extension 15 SP1 and above, as announced at July 2020 by Microsoft (https://azure.microsoft.com/en-us/blog/announcing-the-general-availability-of-azure-shared-disks-and-new-azure-disk-storage-enhancements). With this new capability, it gives more flexibility to mission critical applications in the cloud environment, for example, SAP workload. Microsoft Azure Shared Disks provides high performance storage to the virtual machines running a SUSE Linux Enterprise Server operation system, and SUSE Linux Enterprise High Availability Extension adds the fault tolerance on top.

At the concept level, Microsoft Azure Shared Disks is not different from the other traditional shared disk technology on premises. This blog post mainly follows the latest SLE HA 15 SP2 Administration Guide (https://documentation.suse.com/sle-ha/15-SP2/html/SLE-HA-all/book-sleha-guide.html) to set up two use cases as outlined below (with the tuning of some parameters to accomodate the Azure environment).

- Active-Passive NFS server
- Active-Active OCFS2

NOTE: (In this blog, the following abbreviations are used.)
“SLES” stands for “SUSE Linux Enterprise Server”
“SLES for SAP” stands for “SUSE Linux Enterprise Server for SAP”
“SLE HA” stands for “SUSE Linux Enterprise High Availability Extension”
“SBD” stands for STONITH Block Device

Prerequisites – Azure Environment

To check the Azure Cloud Shell environment

At the local command line environment, the `azure-cli` version must be at 2.3.1 or higher.

```
suse@tumbleweed:~> az --version
```

Or, go directly to https://shell.azure.com , which is okay for the basic usage, but much less flexible for Linux admins.

To create two virtual machines and the shared disk

1. get SUSE image URN from the marketplace

```
URN=`az vm image list --publisher SUSE -f sles-sap-15-sp2-byos \
--sku gen2 --all --query "[-1].urn"|tr -d '"'`; echo $URN
```

2. create two virtual machines from scratch

NOTE: all regions with managed disks support Azure Shared Disks now, [click here] (https://docs.microsoft.com/en-us/azure/virtual-machines/linux/disks-shared).

```
LC=westus2
RG=asd_${LC}_001
URN="SUSE:sles-sap-15-sp2-byos:gen2:2020.09.21"

az group create --name $RG --location $LC
az ppg create -n ppg_$RG -g $RG -l $LC -t standard

az vm create --resource-group $RG --image $URN --ppg ppg_$RG \
  --admin-username $USER --ssh-key-values ~/.ssh/id_rsa.pub \
  --size Standard_D2s_v3 --name asd-sles15sp2-n1

az vm create --resource-group $RG --image $URN --ppg ppg_$RG \
  --admin-username $USER --ssh-key-values ~/.ssh/id_rsa.pub \
  --size Standard_D2s_v3 --name asd-sles15sp2-n2
```

3. To create the shared disk and attach it to virtual machines

```
DN=asd_shared_disk_152
az disk create -g $RG -n $DN -z 256 --sku Premium_LRS --max-shares 2

diskId=$(az disk show -g $RG -n $DN --query 'id' -o tsv); echo $diskId
az vm disk attach -g $RG --name $diskId --cachin None --vm-name asd-sles15sp2-n
az vm disk attach -g $RG --name $diskId --cachin None --vm-name asd-sles15sp2-n
```

General prerequisites for a basic HA cluster in Azure

1. Update SUSE Linux Enterprise Server

After logging in to the virtual machine, the first thing is to update SLES with the latest patches:

```
sudo SUSEConnect -r $REGCODE
sudo zypper up -y
sudo reboot
```

NOTE: The subscription registration code (REGCODE) can be retrieved from scc.suse.com

2. SBD must require watchdog

To enable and load softdog on all nodes:

```
sudo modprobe softdog; echo "softdog"|sudo tee /etc/modules-load.d/softdog.conf
```

NOTE: `softdog` is the only watchdog in the public cloud.

To verify if softdog is loaded:

```
sudo sbd query-watchdog
```

3. Prepare SBD Partition on one node

SBD needs a very small partition with 4MiB size [REF: SLE HA Guide – SBD Partition] (https://documentation.suse.com/sle-ha/15-SP2/html/SLE-HA-all/cha-ha-storage-protect.html#sec-ha-storage-protect-overview)

On one node, use `lsblk` to confirm the correct device name “/dev/sdX” refers to the shared disk, which might change upon OS reboot. SLES tries but doesn’t guarantee the persistent device name per reboot:

```
sudo lsblk
```

So, a partition label for partitions is intentionally assigned:

```
sudo parted /dev/sdc mklabel GPT
sudo parted /dev/sdc mkpart sbd-sles152 1MiB 5MiB
sudo parted /dev/sdc mkpart asd-data1 10GiB 20GiB
sudo parted /dev/sdc mkpart asd-data2 20GiB 30GiB
```

To verify the disk from another node:

```
sudo partprobe; sleep 5; sudo ls -l /dev/disk/by-partlabel/
```

1. Manually setup passwordless ssh for root login among the nodes

On both nodes, run:

```
sudo ssh-keygen
```

Then, append /root/.ssh/id_rsa.pub of both nodes to /root/.ssh/authorized_keys

Verify passwordless does work:

```
asd-sles15sp2-n1:~> sudo ssh asd-sles15sp2-n2
asd-sles15sp2-n2:~> sudo ssh asd-sles15sp2-n1
```

2. Now, let's bootstrap the basic cluster

```
asd-sles15sp2-n1:~> sudo crm cluster init -y -u -s /dev/disk/by-partlabel/sbd-s
```

NOTE: To focus on Azure shared disk topic, this blog will not cover how to make the VIP(10.0.0.9) works with Azure Load Balancer.

Wait until Node 1 finishes, then let Node 2 join:

```
asd-sles15sp2-n2:~> sudo crm cluster join -y -c asd-sles15sp2-n1
```

To monitor the cluster status:

```
asd-sles15sp2-n1:~> sudo crm_mon -rR
```

3. Basic cluster tuning in Azure

STEPS: Fine-tune SBD on-disk metadata

The sbd watchdog timeout in the on-disk metadata is used by the sbd daemon to initialize the watchdog driver. The default is 5 seconds. To add resilience against the foreseeable hiccup from the public cloud provider in case of planned maintenance activities, it might be good to enlarge the watchdog timeout to 60 seconds and the associated `msgwait` to 120 seconds. [REF: SLE HA Guide – Setting Up SBD with Devices] (https://documentation.suse.com/sle-ha/15-SP2/html/SLE-HA-all/cha-ha-storage-protect.html#sec-ha-storage-protect-fencing-setup)

Changing the SBD on-disk metadata requires to recreate the SBD disk on on node:

```
SBD_DEVICE=/dev/disk/by-partlabel/sbd-sles152
sudo sbd -d ${SBD_DEVICE} -1 60 -4 120 create
```

To verify the sbd device meta data:

```
sudo sbd -d ${SBD_DEVICE} dump
```

To re-initiate the watchdog driver you must restart the sbd daemon:

```
sudo crm cluster run "crm cluster restart"
```

STEPS: SBD stonith test

After the cluster is bootstrapped and running, we can play with SBD a bit. [REF: SLE HA Guide – Testing SBD and Fencing] (https://documentation.suse.com/sle-ha/15-SP2/html/SLE-HA-all/cha-ha-storage-protect.html#sec-ha-storage-protect-test):

```
SBD_DEVICE=/dev/disk/by-partlabel/sbd-sles152
sudo sbd -d ${SBD_DEVICE} message asd-sles15sp2-n2 reset
```

STEPS: Fine-tune the corosync timeout

Similarly to SBD, to add resilience against the foreseeable hiccup from the public cloud provider in case of planned maintenance activities, it might be good to enlarge corosync token timeout. However, be aware that this sacrifices the server recovery time for a real permanent failure, since corosync will take longer time to detect the failure in general. With that, it probably makes sense to change corosync `token` timeout to 30 seconds, and the associated `consensus` timeout to 36 seconds

Edit corosync.conf on all nodes of the cluster. REF: `man corosync.conf`:

```
sudo vi /etc/corosync/corosync.conf
```

```
token: 30000
consensus: 360000
```

Either propagate this change manually to all nodes, or let csync2 do it:

```
sudo csync2 -xv
```

On one node, let all nodes reload the corosync config:

```
sudo corosync-cfgtool -R
```

To verify the change:

```
sudo /usr/sbin/corosync-cmapctl |grep -w -e totem.token -e totem.consensus
```

Active-Passive NFS server

WARNING: This blog is only ... active/passive NFS cluster. If applications require to reclaim locks to work properly, a more advanced NFS solution should be implemented.

1. Prepare lvm and the filesystem

On one node, execute the following steps.

STEPS: Modify the lvm2 configuration

To edit:

```
sudo vi /etc/lvm/lvm.conf
```

> To set `system_id_source = "uname"`. It is "none" by default, which means lvm won't continue to operate any volume group which has been created with a systemid.
>
> To set `auto_activation_volume_list = []` to prevent any volume group get activated automatically at the OS level. Indeed, the volume group of HA LVM must be activated by Pacemaker at the cluster level. NOTE: you can ignore this configure for any `shared` volume group, since it requires the cluster components and can't be activated without the cluster stack.

To verify:

```
sudo lvmconfig global/system_id_source
sudo lvmconfig activation/auto_activation_volume_list
```

> system_id_source="uname"
> auto_activation_volume_list=[]

Either propagate this change manually to all nodes, or let csync2 do it:

```
sudo csync2 -xv
```

STEPS: create the logical volume

```
sudo pvcreate /dev/disk/by-partlabel/asd-data1
sudo vgcreate vg1 /dev/disk/by-partlabel/asd-data1
sudo lvcreate -l 50%VG -n lvl vg1
```

STEPS: Initialize the filesystem superblock

```
sudo mkfs.xfs /dev/vg1/lv1
```

2. Fine-tune the local NFS server configuration

To adjust NFSV4LEASETIME reasonably, change and verify /etc/sysconfig/nfs on all nodes:

```
sudo augtool -s set /files/etc/sysconfig/nfs/NFSV4LEASETIME 60
```

3. Bootstrap HA NFS server

On one node, run the following commands:

```
sudo crm configure \
primitive p_nfsserver systemd:nfs-server \
        op monitor interval=30s

sudo crm configure \
primitive p_vg1 LVM-activate \
        params vgname=vg1 vg_access_mode=system_id \
        op start timeout=90s interval=0 \
        op stop timeout=90s interval=0 \
        op monitor interval=30s timeout=90s

sudo crm configure \
primitive p_fs Filesystem \
        op monitor interval=30s \
        op_params OCF_CHECK_LEVEL=20 \
        params device="/dev/vg1/lv1" directory="/srv/nfs" fstype=xfs
```

NOTE: The `Filesystem` resource agent will create 'directory=' if it not exists yet

```
sudo crm configure \
primitive p_exportfs exportfs \
        op monitor interval=30s \
        params clientspec="*" directory="/srv/nfs" fsid=1 \
        options="rw,mp" wait_for_leasetime_on_stop=true
```

NOTE: By design, pacemaker might distribute `p_vg1 p_fs p_nfsserver p_exportfs admin-ip` on different nodes before creating the group resource as below. Resource failures might be reported accordingly, and they are false positives.

```
sudo crm configure group g_nfs p_vg1 p_fs p_nfsserver p_exportfs admin-ip
```

To clean up known false-positive failures:

```
sudo crm_resource -C
```

To verify the NFS server:

```
suse@asd-sles15sp2-n2:~> sudo showmount -e asd-sles15sp2-n1
suse@asd-sles15sp2-n2:~> sudo showmount -e asd-sles15sp2-n2
```

```
Export list for asd-sles15sp2-n1:
/srv/nfs *
```

Active-Active OCFS2

1. Launch `dlm` and `lvmlockd`

This must be done before creating a `shared` logical volume on one node [REF: SLE HA Guide] (https://documentation.suse.com/sle-ha/15-SP2/html/SLE-HA-all/cha-ha-clvm.html#sec-ha-clvm-config):

```
sudo crm configure \
primitive dlm ocf:pacemaker:controld \
        op monitor interval=60 timeout=60

sudo crm configure \
primitive lvmlockd lvmlockd \
        op start timeout=90 interval=0 \
        op stop timeout=90 interval=0 \
        op monitor interval=30 timeout=90

sudo crm configure group g_ocfs2 dlm lvmlockd
sudo crm configure clone c_ocfs2 g_ocfs2 meta interleave=true
```

2. Prepare lvm2 `shared` disks on one node

```
sudo ls -l /dev/disk/by-partlabel/
sudo pvcreate /dev/disk/by-partlabel/asd-data2
sudo vgcreate --shared vg2-shared /dev/disk/by-partlabel/asd-data2
sudo lvcreate -an -l 50%VG -n lv1 vg2-shared

sudo crm configure \
primitive p_vg_shared LVM-activate \
        params vgname=vg2-shared vg_access_mode=lvmlockd activation_mode=shared
        op start timeout=90s interval=0 \
        op stop timeout=90s interval=0 \
        op monitor interval=30s timeout=90s

sudo crm configure modgroup g_ocfs2 add p_vg_shared
```

3. Prepare ocfs2 on one node

[REF: SLE HA Guide] (https://documentation.suse.com/sle-ha/15-SP2/html/SLE-HA-all/cha-ha-ocfs2.html#sec-ha-ocfs2-create)

```
sudo mkfs.ocfs2 /dev/vg2-shared/lv1
```

4. Finally run ocfs2 on all nodes

[REF: SLE HA Guide] (https://documentation.suse.com/sle-ha/15-SP2/html/SLE-HA-all/cha-ha-ocfs2.html#sec-ha-ocfs2-mount)

```
sudo crm configure \
primitive p_ocfs2 Filesystem \
        params device="/dev/vg2-shared/lv1" directory="/srv/ocfs2" fstype=ocfs2
        op monitor interval=20 timeout=40 \
        op_params OCF_CHECK_LEVEL=20 \
        op start timeout=60 interval=0 \
        op stop timeout=60 interval=0

sudo crm configure modgroup g_ocfs2 add p_ocfs2
```

To write some text file to ocfs2 at the current node:

```
asd-sles15sp2-n1:~> echo "'Hello' from `hostname`" | sudo tee /srv/ocfs2/hello_
```

To verify from the other node:

```
asd-sles15sp2-n2:~> cat /srv/ocfs2/hello_world
```

End of the exercise, enjoy!

Share

Tags: #SLES4SAP (https://www.suse.com/c/tag/sles4sap/), #SLES4SAP #Activation #Migratoin (https://www.suse.com/c/tag/sles4sap-activation-migratoin/), CSP (https://www.suse.com/c/tag/csp/), HA Storage (https://www.suse.com/c/tag/ha-storage/), High availability (https://www.suse.com/c/tag/high-availability/), Logical Volume Management (LVM) (https://www.suse.com/c/tag/logical-volume-management-lvm/), Microsoft Azure (https://www.suse.com/c/tag/microsoft-azure/), NFS (https://www.suse.com/c/tag/nfs/), OCFS2 (https://www.suse.com/c/tag/ocfs2/), Public Cloud (https://www.suse.com/c/tag/public-cloud/), SLE15 SP2 (https://www.suse.com/c/tag/sle15-sp2/), SLEHA (https://www.suse.com/c/tag/sleha/), SUSE (https://www.suse.com/c/tag/suse/) Category: Cloud Computing (https://www.suse.com/c/category/cloud-computing/), CSP (https://www.suse.com/c/category/partners/csp/), Digital Transformation (https://www.suse.com/c/category/digital-transformation/), Integrated Systems (https://www.suse.com/c/category/integrated-systems/), Microsoft (https://www.suse.com/c/category/partners/alliance-partners/microsoft/), Run SAP Solutions (https://www.suse.com/c/category/solutions/run-sap-solutions/), SAP (https://www.suse.com/c/category/partners/alliance-partners/sap/), SLES on Azure (https://www.suse.com/c/category/sles-on-azure/), SUSE Linux Enterprise High Availability Extension (https://www.suse.com/c/category/products/suse-linux-enterprise-high-availability-extension/), SUSE Linux Enterprise Server for SAP Applications (https://www.suse.com/c/category/products/suse-linux-enterprise-server-for-sap-applications/), Technical Solutions

(https://www.suse.com/c/category/solutions/sector/federal/)
This entry was posted Tuesday, 10 November, 2020 at 10:58 am
You can follow any responses to this entry via RSS (https://www.suse.com/c/azure-shared-disks-excercise-w-sles-for-sap-or-sle-ha/feed/).

Leave a Reply

Logged in as Zhiqiang(Roger) Zhou (https://www.suse.com/c/wp-admin/profile.php). Log out? (https://www.suse.com/c/wp-login.php?action=logout&redirect_to=https%3A%2F%2Fwww.suse.com%2Fc%2Fazure-shared-disks-excercise-w-sles-for-sap-or-sle-ha%2F&_wpnonce=8113c22545)

Comment

Post Comment

No comments yet

SES 7 and SUSE Global Services: A Perfect Pairing (https://www.suse.com/c/ses-7-and-suse-global-services-a-perfect-pairing/)
SUSE releases fixes for new PLATYPUS attack (https://www.suse.com/c/suse-releases-fixes-for-new-platypus-attack/)
An Introduction to SUSE Manager for Retail (https://www.suse.com/c/an-introduction-to-suse-manager-for-retail/)
Azure Shared Disks with “SLES for SAP / SLE HA 15 SP2” (https://www.suse.com/c/azure-shared-disks-excercise-w-sles-for-sap-or-sle-ha/)
A New SUSE SAP Business One installer image is ready (https://www.suse.com/c/a-new-suse-sap-business-one-installer-image-is-ready/)

Popular Blog Posts

YaST Changes Arrive in Tumbleweed Snapshot – All You... (https://www.suse.com/c/yast-changes-arrive-in-tumbleweed-snapshot-all-you-need-to-know-about-the-new-storage-stack/)
Top 5 Reasons to Take the Cloud Foundry User Survey (https://www.suse.com/c/top-5-reasons-to-take-the-cloud-foundry-user-survey-this-week/)
Introducing Open Build Service, Version 2.10 (https://www.suse.com/c/introducing-open-build-service-version-2-10/)
Digital Transformation is all around us. Is your... (https://www.suse.com/c/digital-transformation-around-us-infrastructure-ready/)
Hola Barcelona! – SUSE @ TechEd – All... (https://www.suse.com/c/hola-barcelona-suse-teched-all-you-need-to-know/)

Categories

Select Category

Archives

Select Month

SUSE History

(/company/history/)

SUSE Management Team

(/company/executive-management/)

SUSE Logos & Imagery

(/brandcentral/suse/identity.php)

(https://www.suse.com) (https://www.suse.com/solutions/sector/federal/)

English

Careers (https://www.suse.com/company/careers/) Legal (https://www.suse.com/company/legal/)

Feedback About (https://www.suse.com/company/about/)

Communications Preferences (https://www.suse.com/company/subscribe/)

Contact Us (https://www.suse.com/contact/) Let's Chat

(https://www.suse.com/company/suse)

Sales Number: 49-911-740-53-0 Support: Open a Support Case (https://scc.suse.com/)

 Privacy and Cookie Policy (https://www.suse.com/company/policies/privacy/)