

TUT-1071

Simplifying HA Cluster bootstrap and shutdown with SLE-HA 15 SP4



Roger Zhou

Sr. Engineering Manager

zzhou@suse.com

Agenda:

1. SLE-HA 15 SP4 highlights to improve UI/UX
2. Example scenario to setup a cluster
3. Things behind the scene

Preventing LVM2 autoactivation during OS boot
Cluster predefined configuration
Cluster with the disk-based SBD in VM
Cluster with the disk-less SBD
Cluster stack graceful shutdown

4. Future TODO list



SLE-HA 15 SP4 highlights to improve UI/UX

The continuous effort over years ...

- Easier initial setup with crmsh predefined bootstrap profiles
- Cluster graceful shutdown using crmsh "--all" option
It improves diskless sbd cluster and the dlm cluster
- ocfs2 'nocluster' mount option as the local mount without the cluster stack running
- pacemaker supports ocfl1
- corosync adds resilience to cope with the Public Cloud packet inspection hiccup

Example scenario to setup a cluster filesystem

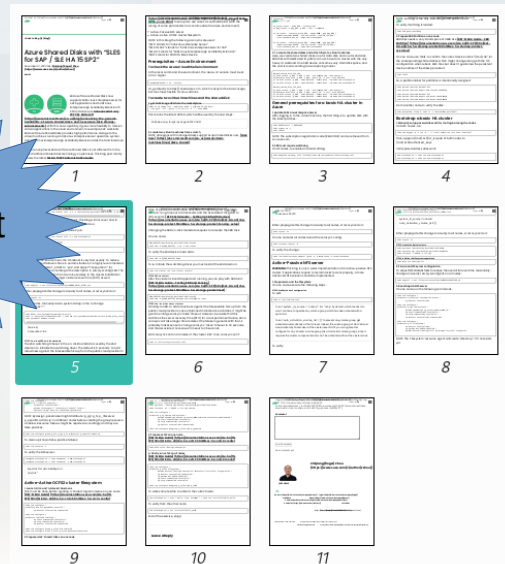
- Now, only need two commands, for example:

```
ssh -t 15sp4-1 crm cluster init -y \  
-s /dev/disk/by-partlabel/sbd-154 \  
-o /dev/disk/by-partlabel/ocfs2 \  
-m /srv/ocfs2 -C
```

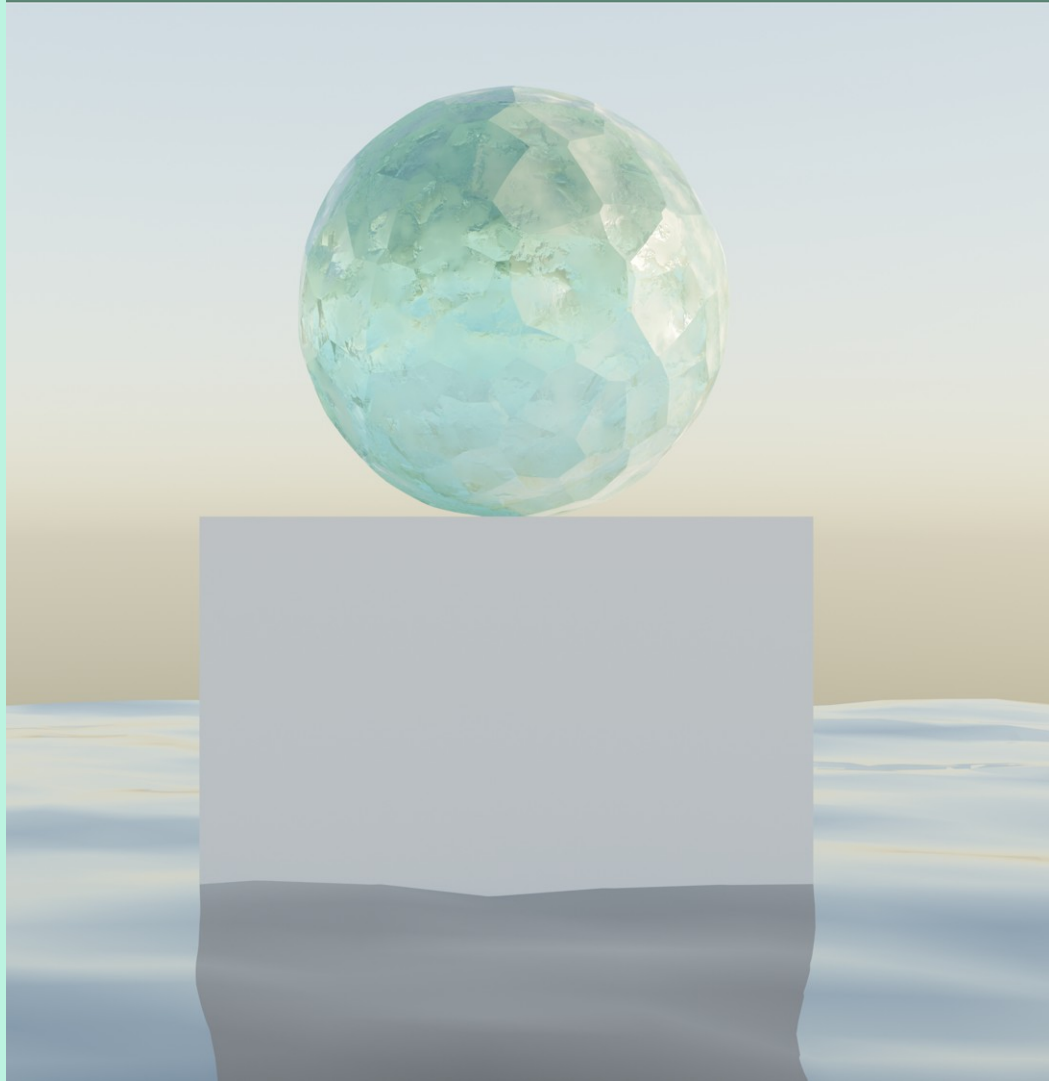
```
ssh -t 15sp4-2 crm cluster join -yc 15sp4-1
```

- NOTE

- C option indicates to use the lvmlockd based Cluster LVM for ocfs2.
- u option is no longer needed now. By default crmsh bootstraps corosync with udpu.



Behind the Scene during bootstrapping



Behind the scene

To use LVM2 in the cluster, we need prevent LVM2 autoactivation during os boot

- For the shared VG, it can not be activated without the cluster stack in the very early stage during the os boot.
- For the specified 'systemid' volumes, to prevent them getting activated during the os boot, LVM-Activate Resource Agent leverages a new `--setautoactivation lvm2` option. It defers lvm activation operations to the pacemaker stack.
- `crmsh` sets "no-quorum-policy=freeze" as the general advice for a dlm cluster. This avoids the "stop failure" to cause the node reset.

Behind the scene

crmsh has the predefined cluster configuration in `/etc/crm/profile.yml`

— For example, the typical cluster in “microsoft-azure”

```
corosync.totem.token 30000
```

```
sbd.watchdog_timeout 60
```



Behind the scene

Cluster with the disk-based SBD use case

- crmsh checks watchdog, and will configure and modprobe the softdog driver during the os boot, if need.
- to set pcmk_delay_max 30s if the cluster is 2-node without qdevice.
- to ensure the determinate state of the stonith operation
stonith-timeout > pcmk_delay_max + msgwait
- to prevent the fence failure to trigger double fencing
stonith-timeout > token + consensus

Behind the scene

Cluster with the disk-based SBD in VM environment (eg. kvm/xen, vmware, hyper-v)

- to avoid a node returns to the cluster too quickly and print the critical message "We were alleged just fenced by ..."

$\text{SBD_DELAY_START} > \text{token} + \text{consensus} + \text{pcmk_delay_max} + \text{msgwait}$

- to avoid systemd giveup starting sbd service
crmsh bootstrap adjusts TimeoutStartSec > SBD_DELAY_START

Behind the scene

Cluster with the disk-less sbd use case

- `stonith-timeout > stonith-watchdog-timeout`

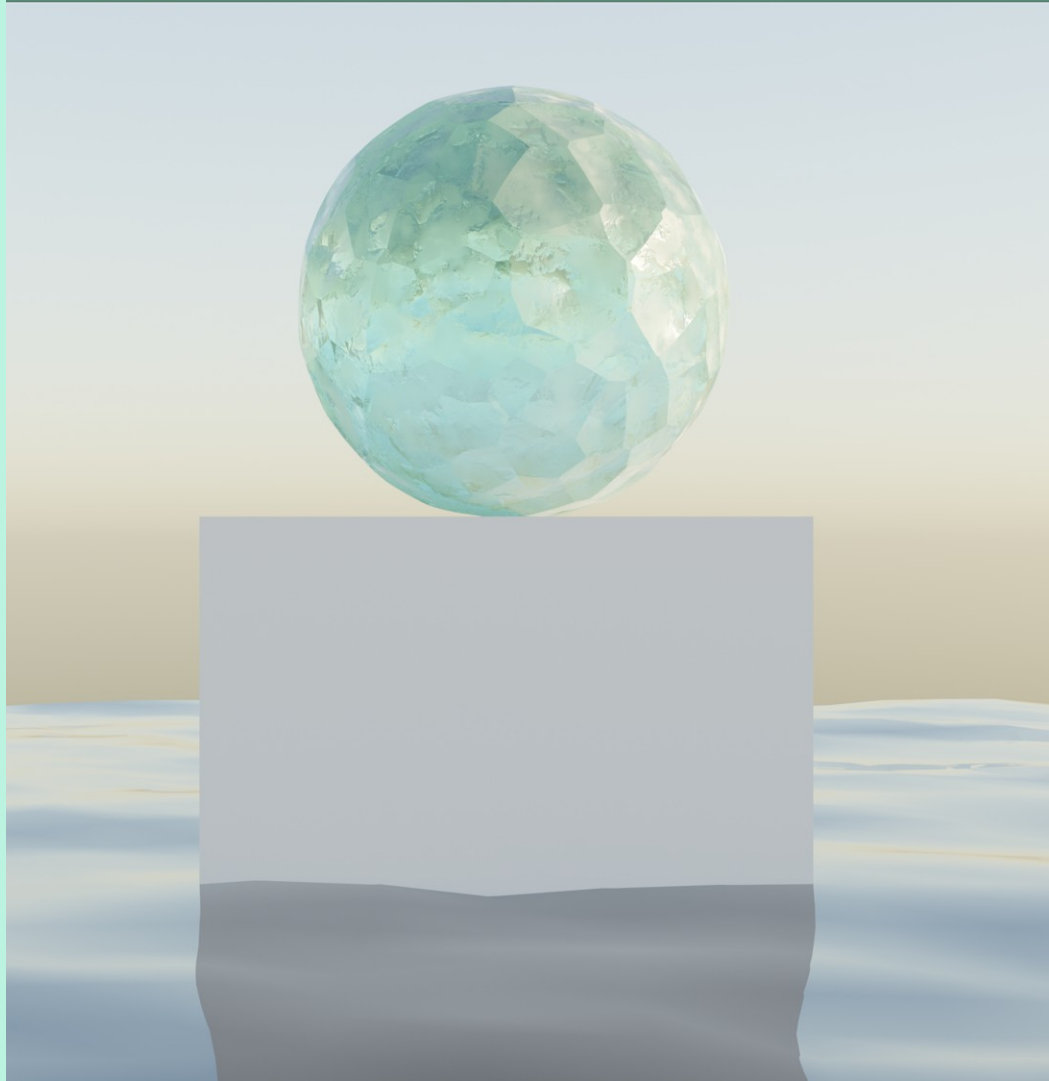
`stonith-watchdog-timeout = 2*SBD_WATCHDOG_TIMEOUT`

- For the VM environment

`SBD_DELAY_START > token + consensus + 2*SBD_WATCHDOG_TIMEOUT`

Behind the Scene

cluster graceful shutdown



Behind the scene: graceful shutdown of the whole cluster

- ``crm cluster stop --all`` stops pacemaker at all nodes first, then stops corosync afterwards at all nodes.
- This intends to solve two issues of the cluster shutdown in the node-by-node approach. In that case, once the node loses quorum:
 1. the diskless sbd cluster will trigger self reset
 2. the dlm will hang and it prevents shutdown
- Furthermore, `dlm_tool` implements a brand new ``set_config`` option. `crmsh` leverages it to shutdown the hanging dlm above.

Future TODO

The continuous effort to improve UI/UX ...

- bootstrap the priority fence delay feature for the 2-node cluster
- gradually start the large cluster, eg. tens of nodes
- speedup time by isolating csync2 for the normal bootstrapping
- ...
- Looking forward to your feedback. They count!

Thank you



For more information, contact SUSE at:

+1 800 796 3700 (U.S./Canada)

+49 (0)911-740 53-0 (Worldwide)

Maxfeldstrasse 5

90409 Nuremberg

www.suse.com

© SUSE LLC. All Rights Reserved. SUSE and the SUSE logo are registered trademarks of SUSE LLC in the United States and other countries. All third-party trademarks are the property of their respective owners.

General Disclaimer: This document is not to be construed as a promise by any participating company to develop, deliver, or market a product. It is not a commitment to deliver any material, code, or functionality, and should not be relied upon in making purchasing decisions. SUSE makes no representations or warranties with respect to the contents of this document, and specifically disclaims any express or implied warranties of merchantability or fitness for any particular purpose. The development, release, and timing of features or functionality described for SUSE products remains at the sole discretion of SUSE. Further, SUSE reserves the right to **revise** this document and to make changes to its content, at any time, without obligation to notify any person or entity of such revisions or changes.

All SUSE marks referenced in this presentation are trademarks or registered trademarks of SUSE, LLC, Inc. in the United States and other countries. All third-party trademarks are the property of their respective owners.