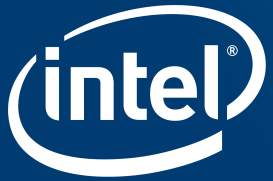


New Resource Agents for Pacemaker available to the Lustre community

Gabriele Paciucci, HPC Solution Architect

Copyright ® Intel Corporation 2016. All rights reserved. Intel and the Intel logo are trademarks of Intel Corporation in the U.S. and/or other countries.
* Other names and brands may be claimed as the property of others.

Legal Information

All information provided here is subject to change without notice. Contact your Intel representative to obtain the latest Intel product specifications and roadmaps

Tests document performance of components on a particular test, in specific systems. Differences in hardware, software, or configuration will affect actual performance. Consult other sources of information to evaluate performance as you consider your purchase. For more complete information about performance and benchmark results, visit http://www.intel.com/performance.

Intel technologies' features and benefits depend on system configuration and may require enabled hardware, software or service activation. Performance varies depending on system configuration. No computer system can be absolutely secure. Check with your system manufacturer or retailer or learn more at http://www.intel.com/content/www/us/en/software/intel-solutions-for-lustre-software.html.

Intel technologies may require enabled hardware, specific software, or services activation. Check with your system manufacturer or retailer.

You may not use or facilitate the use of this document in connection with any infringement or other legal analysis concerning Intel products described herein. You agree to grant Intel a non-exclusive, royalty-free license to any patent claim thereafter drafted which includes subject matter disclosed herein.

No license (express or implied, by estoppel or otherwise) to any intellectual property rights is granted by this document.

The products described may contain design defects or errors known as errata which may cause the product to deviate from published specifications. Current characterized errata are available on request.

Intel disclaims all express and implied warranties, including without limitation, the implied warranties of merchantability, fitness for a particular purpose, and non-infringement, as well as any warranty arising from course of performance, course of dealing, or usage in trade.

This document contains information on products, services and/or processes in development. All information provided here is subject to change without notice. Contact your Intel representative to obtain the latest forecast, schedule, specifications and roadmaps.

A "Mission Critical Application" is any application in which failure of the Intel Product could result, directly or indirectly, in personal injury or death. SHOULD YOU PURCHASE OR USE INTEL'S PRODUCTS FOR ANY SUCH MISSION CRITICAL APPLICATION, YOU SHALL INDEMNIFY AND HOLD INTEL AND ITS SUBSIDIARIES, SUBCONTRACTORS AND AFFILIATES, AND THE DIRECTORS, OFFICERS, AND EMPLOYEES OF EACH, HARMLESS AGAINST ALL CLAIMS COSTS, DAMAGES, AND EXPENSES AND REASONABLE ATTORNEYS' FEES ARISING OUT OF, DIRECTLY OR INDIRECTLY, ANY CLAIM OF PRODUCT LIABILITY, PERSONAL INJURY, OR DEATH ARISING IN ANY WAY OUT OF SUCH MISSION CRITICAL APPLICATION, WHETHER OR NOT INTEL OR ITS SUBCONTRACTOR WAS NEGLIGENT IN THE DESIGN, MANUFACTURE, OR WARNING OF THE INTEL PRODUCT OR ANY OF ITS PARTS.

Intel may make changes to specifications and product descriptions at any time, without notice. Designers must not rely on the absence or characteristics of any features or instructions marked "reserved" or "undefined". Intel reserves these for future definition and shall have no responsibility whatsoever for conflicts or incompatibilities arising from future changes to them. The information here is subject to change without notice. Do not finalize a design with this information.

Intel and the Intel logo are trademarks of Intel Corporation in the U.S. and/or other countries.

* Other names and brands may be claimed as the property of others.

© 2016 Intel Corporation

Copyright ® Intel Corporation 2016. All rights reserved. Intel and the Intel logo are trademarks of Intel Corporation in the U.S. and/or other countries.
* Other names and brands may be claimed as the property of others.

Agenda

- High availability and Lustre
- (Quick) Basic cluster setup
- Resource Agent for Lustre
- Constraints
- Improving monitor capabilities
- Conclusions

Copyright ® Intel Corporation 2016. All rights reserved. Intel and the Intel logo are trademarks of Intel Corporation in the U.S. and/or other countries.
* Other names and brands may be claimed as the property of others.

High Availability and Failover in Lustre

- Service continuity is implemented in Lustre by a mechanism called “failover”
- Failover service runs in one location at a time, but can choose where
- Lustre services are tightly coupled to the data storage (targets)
- There are no user-space daemons: only *mount* or *u(n)mount*
- We need to guarantee no data corruption
- Lustre uses external HA framework

Copyright ® Intel Corporation 2016. All rights reserved. Intel and the Intel logo are trademarks of Intel Corporation in the U.S. and/or other countries.
* Other names and brands may be claimed as the property of others.

High Availability and Failover

- Storage targets must be configured with the NIDs of the Lustre servers hosting the Lustre service
- These are specified using *--failnode* or *--servicenode* options to mkfs.lustre:
  - failnode: older method primary/secondary
  - servicenode: **recommended** option, all hosts are equally able to run a given service, with no defined preferred primary

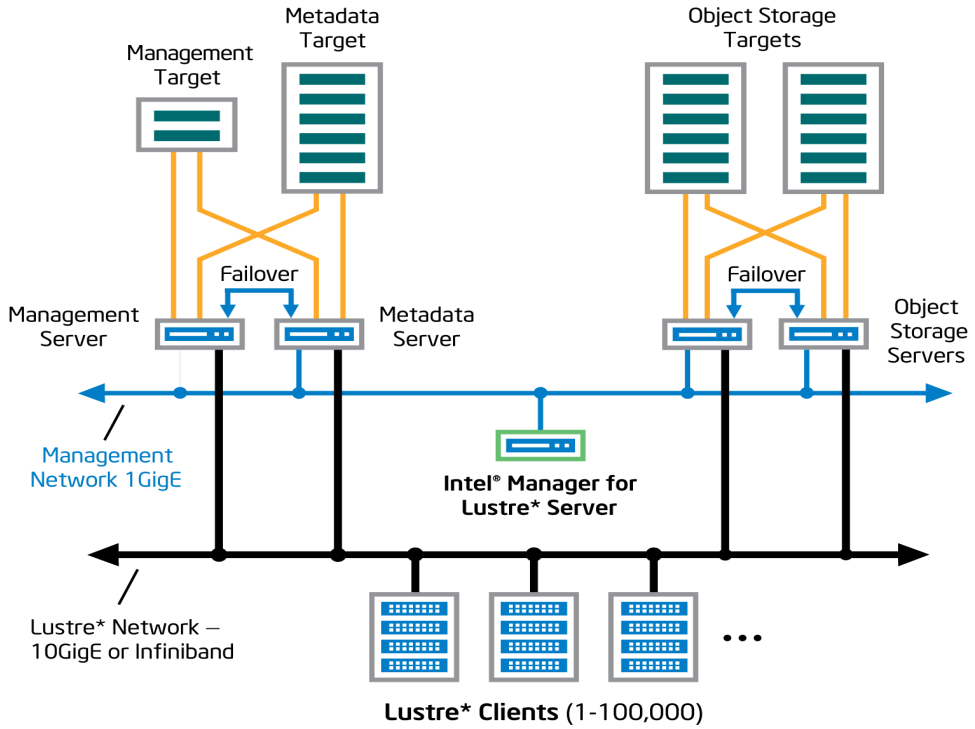
Copyright ® Intel Corporation 2016. All rights reserved. Intel and the Intel logo are trademarks of Intel Corporation in the U.S. and/or other countries.
* Other names and brands may be claimed as the property of others.

Pacemaker and Corosync

- The most commonly used HA cluster framework for Linux comprises two software applications used in combination:
  - Pacemaker – resource management
  - Corosync – cluster communications and low-level management
- Pacemaker is coming from Linux HA project (Heartbeat)
- Corosync is derived from the OpenAIS project
- In RHEL/CentOS 7, the high-availability framework has been rationalized around Pacemaker and Corosync version 2
- Red Hat also provides a command line tool called PCS (Pacemaker and Corosync Shell) that is available for both RHEL version 6 and version 7

Copyright ® Intel Corporation 2016. All rights reserved. Intel and the Intel logo are trademarks of Intel Corporation in the U.S. and/or other countries.
* Other names and brands may be claimed as the property of others.

Configuration for Two- Node Cluster



- Is the 2-way cluster the most common choice ?
- Each server requires several network interfaces:
  - A *management network* or public interface connection
  - A dedicated *cluster communication network* between paired servers
  - Routed network to fencing devices
  - *High performance data network*

Copyright ® Intel Corporation 2016. All rights reserved. Intel and the Intel logo are trademarks of Intel Corporation in the U.S. and/or other countries.
* Other names and brands may be claimed as the property of others.

Basic cluster setup

Copyright ® Intel Corporation 2016. All rights reserved. Intel and the Intel logo are trademarks of Intel Corporation in the U.S. and/or other countries.
* Other names and brands may be claimed as the property of others.

Configure the Basic HA Framework

- HA framework software to install (extra subscriptions for RHEL):

```
yum -y install pcs pacemaker corosync fence-agents
```

- To manage the cluster add an user account to the *haclient* group (*hacluster*)
- Set a password for the *hacluster* account
- Start the Pacemaker configuration daemon (pcsd) on all servers:

```
systemctl start pcsd.service
```

- Set up PCS authentication:

```
pcs cluster auth serverA serverB -u hacluster
```

Copyright ® Intel Corporation 2016. All rights reserved. Intel and the Intel logo are trademarks of Intel Corporation in the U.S. and/or other countries.
* Other names and brands may be claimed as the property of others.

Setup the cluster

- Configuration of the cluster:

```
pcs cluster setup --name Cluster01 serverA,serverA-HA
serverB,serverB-HA --transport udp --rrpmode passive --
token 17000
```

  - ServerA and ServerB (output from *uname -n)*
  - ServerA-HA and ServerB-HA (Ips on the cluster communication network)
  - Transport types: *udpu* (UDP unicast) and *udp* (used for multicast)
  - The redundant ring protocol (RRP) mode is specified by the *--rrpmode* flag
  - The *--token* timeout before a node is declared dead: **on a Lustre server we want to increase this value**
    - Pacemaker user space threads vs Lustre kernel space threads under heavy load

Copyright ® Intel Corporation 2016. All rights reserved. Intel and the Intel logo are trademarks of Intel Corporation in the U.S. and/or other countries.
* Other names and brands may be claimed as the property of others.

Status of the cluster

```
# pcs status
Cluster name: kapollo_oss
WARNING: no stonith devices and stonith-enabled is not false
Last updated: Thu Jun  2 05:19:48 2016          Last change: Thu Jun  2 05:20:20 2016 by hacluster via
crmd on kapollo02
Stack: corosync
Current DC: kapollo02 (version 1.1.13-10.el7_2.2-44eb2dd) - partition with quorum
2 nodes and 0 resources configured

Online: [ kapollo01 kapollo02 ]

Full list of resources:


PCSD Status:
  kapollo01: Online
  kapollo02: Online

Daemon Status:
  corosync: active/disabled
  pacemaker: active/disabled
  pcsd: active/disabled
```

Copyright ® Intel Corporation 2016. All rights reserved. Intel and the Intel logo are trademarks of Intel Corporation in the U.S. and/or other countries.
* Other names and brands may be claimed as the property of others.

Status of the corosync rings

```
# corosync-cfgtool -s
Printing ring status.
Local node ID 1
RING ID 0
  id = 10.70.227.11
  status = ring 0 active with no faults
RING ID 1
  id = 192.168.227.11
  status = ring 1 active with no faults
```

Copyright ® Intel Corporation 2016. All rights reserved. Intel and the Intel logo are trademarks of Intel Corporation in the U.S. and/or other countries.
* Other names and brands may be claimed as the property of others.

Fencing mechanism

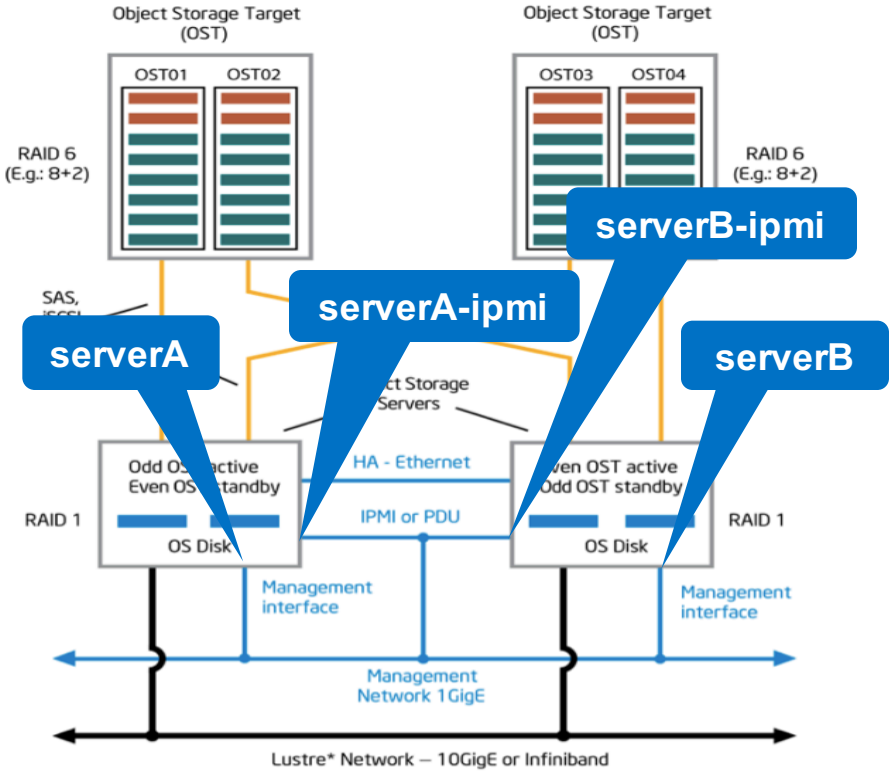
- In a high-availability environment, it is essential to isolate a failed component and remove it from production to avoid any corruption: fencing mechanism
- Pacemaker is using STONITH (Shoot the Other Node in the Head) as fencing mechanism
- Pacemaker has a set of software components called fencing agents that are used for this purpose

Copyright ® Intel Corporation 2016. All rights reserved. Intel and the Intel logo are trademarks of Intel Corporation in the U.S. and/or other countries.
* Other names and brands may be claimed as the property of others.

Fencing configuration

```
pcs  stonith create serverA-fence fence_ipmilan
ipaddr=serverA-ipmi passwd=XXX lanplus=yes login=XXX
pcmk_host_list=serverA

pcs stonith create serverB-fence fence_ipmilan
ipaddr=serverB-ipmi passwd=XXX lanplus=yes login=XXX
pcmk_host_list=serverB
```



Copyright ® Intel Corporation 2016. All rights reserved. Intel and the Intel logo are trademarks of Intel Corporation in the U.S. and/or other countries.
* Other names and brands may be claimed as the property of others.

Status of the cluster

```
# pcs status
Cluster name: kapollo_oss
Last updated: Thu Jun  2 05:22:57 2016          Last change: Thu Jun  2 05:21:03 2016 by root via
cibadmin on kapollo01
Stack: corosync
Current DC: kapollo02 (version 1.1.13-10.el7_2.2-44eb2dd) - partition with quorum
2 nodes and 2 resources configured

Online: [ kapollo01 kapollo02 ]

Full list of resources:

 kapollo01-ipmi (stonith:fence_ipmilan):        Started kapollo01
 kapollo02-ipmi (stonith:fence_ipmilan):        Started kapollo02

PCSD Status:
  kapollo01: Online
  kapollo02: Online

Daemon Status:
  corosync: active/disabled
  pacemaker: active/disabled
  pcsd: active/enabled
```

Copyright ® Intel Corporation 2016. All rights reserved. Intel and the Intel logo are trademarks of Intel Corporation in the U.S. and/or other countries.
* Other names and brands may be claimed as the property of others.

Resource Agent for Lustre

Copyright ® Intel Corporation 2016. All rights reserved. Intel and the Intel logo are trademarks of Intel Corporation in the U.S. and/or other countries.
* Other names and brands may be claimed as the property of others.

Pacemaker and Lustre

- The Lustre community provided some scripts for the (old) Heartbeat framework
- Pacemaker provides a generic Filesystem RA that give a basic support for Lustre on LDISKFS only
- Intel is providing 3 new RAs to:
  - support for ZFS* based Lustre file system: LU-8455
  - increase monitor capabilities for LNet: LU-8457
  - increase monitor capabilities for Lustre Services: LU-8458

Copyright ® Intel Corporation 2016. All rights reserved. Intel and the Intel logo are trademarks of Intel Corporation in the U.S. and/or other countries.
* Other names and brands may be claimed as the property of others.

Target's resource for ZFS (LU-8455)

```
pcs resource create ost00 ocf:heartbeat:LustreZFS pool=<pool name> volume=<volume name> mountpoint=<mount point> [OCF_CHEK_LEVEL=10]
```

- The script mount/umount MGT, MDT and OST targets and export/import of pools
- `pool` (pool name used during zpool create). The script is able to manage 1 target per pool
- `volume` (volume name defined during the Lustre formatting).
- `OCF_CHEK_LEVEL=10`. This option enable testing the health of the pool using the `zpool get health` command. DEGRADED mode is reported only. FAULTY pool is causing a failover
- **hostid protection must be configured**

Copyright ® Intel Corporation 2016. All rights reserved. Intel and the Intel logo are trademarks of Intel Corporation in the U.S. and/or other countries.
* Other names and brands may be claimed as the property of others.

ZFS RA assumptions and limitations

- ZFS RA script (LU-8455)
  - Pool import protection using the ZFS's hostid
    - ZFS pools cannot be imported on multiple nodes at one time
    - hostid blocks accidental pool import, but STONITH needed during failover
    - Multi-Mount Protection under development to prevent multiple imports
  - At the moment designed for 2 way cluster only
  - Single script for ZFS's pools import and Lustre mount
  - 1:1 pool/target

Copyright ® Intel Corporation 2016. All rights reserved. Intel and the Intel logo are trademarks of Intel Corporation in the U.S. and/or other countries.
* Other names and brands may be claimed as the property of others.

Status of the cluster

```
# pcs status

Online: [ kapollo01 kapollo02 ]

Full list of resources:

 kapollo01-ipmi (stonith:fence_ipmilan):        Started kapollo02
 kapollo02-ipmi (stonith:fence_ipmilan):        Started kapollo01
 ost00  (ocf::heartbeat:LustreZFS):     Started kapollo02
 ost01  (ocf::heartbeat:LustreZFS):     Started kapollo01
 ost02  (ocf::heartbeat:LustreZFS):     Started kapollo02
 ost03  (ocf::heartbeat:LustreZFS):     Started kapollo01
 ost04  (ocf::heartbeat:LustreZFS):     Started kapollo02
 ost05  (ocf::heartbeat:LustreZFS):     Started kapollo01
 ost06  (ocf::heartbeat:LustreZFS):     Started kapollo02
 ost07  (ocf::heartbeat:LustreZFS):     Started kapollo01
 ost08  (ocf::heartbeat:LustreZFS):     Started kapollo02
 ost09  (ocf::heartbeat:LustreZFS):     Started kapollo01
```

Copyright ® Intel Corporation 2016. All rights reserved. Intel and the Intel logo are trademarks of Intel Corporation in the U.S. and/or other countries.
* Other names and brands may be claimed as the property of others.

Constraints

Copyright ® Intel Corporation 2016. All rights reserved. Intel and the Intel logo are trademarks of Intel Corporation in the U.S. and/or other countries.
* Other names and brands may be claimed as the property of others.

Pacemaker's constraints and score

- The behavior of a resource in a cluster is determined by constraints:
  - location constraints
  - order constraints
  - colocation constraints
- When defining resource constraints, you specify a score for each constraint
- Scores are calculated on a per-resource basis and any node with a negative score for a resource cannot run that resource
- After calculating the scores for a resource, the cluster then chooses the node with the highest score

Copyright ® Intel Corporation 2016. All rights reserved. Intel and the Intel logo are trademarks of Intel Corporation in the U.S. and/or other countries.
* Other names and brands may be claimed as the property of others.

Suggested constraints for Lustre

- Preference constraint to enable manual balancing:
  - pcs constraint location ost00 prefers kapollo01=20
  - pcs constraint location ost00 prefers kapollo02=10
    - pcs resource relocate run
- Stickiness to prevent auto-failback:
  - pcs resource update ost00 meta resource-stickiness=1000

Copyright ® Intel Corporation 2016. All rights reserved. Intel and the Intel logo are trademarks of Intel Corporation in the U.S. and/or other countries.
* Other names and brands may be claimed as the property of others.

Status of the cluster

```
# pcs constraint show
Location Constraints:
  Resource: ost00
    Enabled on: kapollo01 (score:20) (id:location-ost00-kapollo01-20)
    Enabled on: kapollo02 (score:10) (id:location-ost00-kapollo02-10)

 # pcs resource  show ost00
  Resource: ost00 (class=ocf provider=heartbeat type=LustreZFS)
   Attributes: pool=zost00 volume=ost00 mountpoint=/mnt/ost00
   Meta Attrs: resource-stickiness=1000
   Operations: start interval=0s timeout=300s (ost00-start-interval-0s)
               stop interval=0s timeout=300s (ost00-stop-interval-0s)
               monitor interval=20s timeout=300s (ost00-monitor-interval-20s)
```

Copyright ® Intel Corporation 2016. All rights reserved. Intel and the Intel logo are trademarks of Intel Corporation in the U.S. and/or other countries.
* Other names and brands may be claimed as the property of others.

Improving monitor capabilities

Copyright ® Intel Corporation 2016. All rights reserved. Intel and the Intel logo are trademarks of Intel Corporation in the U.S. and/or other countries.
* Other names and brands may be claimed as the property of others.

Increasing the monitor capabilities

- Monitor capabilities are limited to:
  - Operating System crashes
  - Hardware failures
  - Split brain
  - Targets not mounted
  - Ring 0 / Ring 1 failures
- We developed additional scripts to monitor:
  - LNet outages - healthLNET
  - Lustre servers status – healthLUSTRE
  - Lustre targets status - healthLUSTRE

Copyright ® Intel Corporation 2016. All rights reserved. Intel and the Intel logo are trademarks of Intel Corporation in the U.S. and/or other countries.
* Other names and brands may be claimed as the property of others.

Clone RA

- These agents are **not** ZFS specific
- Pacemaker's resource **clone** technique provides active/active monitoring
- The cloned monitor resource is actively monitoring LNet outages or Lustre services outages and update a specific variables in the pacemaker CIB database
- If the variables is below a specific threshold, a pacemakers' constraint is triggered and all the resources available on the faulted node are moved on the healthy node

Copyright ® Intel Corporation 2016. All rights reserved. Intel and the Intel logo are trademarks of Intel Corporation in the U.S. and/or other countries.
* Other names and brands may be claimed as the property of others.

LNet monitor RA (LU-8457)

- Creation and options:

```
pcs resource create healthLNET ocf:pacemaker:healthLNET
dampen=5s multiplier=1000 lctl=true device=ib0
host_list="192.168.211.215@o2ib0 192.168.211.216@o2ib0" –
clone
```

  - `dampen` The time to wait (dampening) further changes occur OCF_RESKEY_dampen=5s
  - `lctl` Option to enable lctl ping. The default is true OCF_RESKEY_lctl=true
  - `device` Device used for the LNET network. We assume the same device accross the cluster
  - `host_list` NIDs (if lctl is selected) or IPs associated to the device selected

Copyright ® Intel Corporation 2016. All rights reserved. Intel and the Intel logo are trademarks of Intel Corporation in the U.S. and/or other countries.
* Other names and brands may be claimed as the property of others.

LNet monitor RA

- Constraints:

  ```
  pcs constraint location <Resource Name> rule score=-
  INFINITY pingd lt 1 or not_defined pingd
  ```

  – `pingd` This is the variable the clone RA is updating in the CIB database

Copyright ® Intel Corporation 2016. All rights reserved. Intel and the Intel logo are trademarks of Intel Corporation in the U.S. and/or other countries.
* Other names and brands may be claimed as the property of others.

Lustre monitor RA (LU-8458)

- Creation and options:

```
pcs resource create healthLUSTRE
ocf:pacemaker:healthLUSTRE dampen=5s  --clone
```

  - `dampen` The time to wait (dampening) further changes occur OCF_RESKEY_dampen=5s
  - Testing: `lctl get_param health_check`
    - LBUG are reported
    - I/O errors are also triggered
  - By default Lustre enable `panic_on_lbug`

Copyright ® Intel Corporation 2016. All rights reserved. Intel and the Intel logo are trademarks of Intel Corporation in the U.S. and/or other countries.
* Other names and brands may be claimed as the property of others.

Lustre monitor RA

- Constraints:

  `pcs constraint location <Resource Name> rule score=-INFINITY lustred lt 1 or not_defined lustred`

  - `lustred` This is the variable the clone RA is updating in the CIB database

Copyright ® Intel Corporation 2016. All rights reserved. Intel and the Intel logo are trademarks of Intel Corporation in the U.S. and/or other countries.
* Other names and brands may be claimed as the property of others.

Status of the cluster

```
# pcs status


Online: [ kapollo01 kapollo02 ]

Full list of resources:

 kapollo01-ipmi (stonith:fence_ipmilan):        Started kapollo02
 kapollo02-ipmi (stonith:fence_ipmilan):        Started kapollo01
 ost00  (ocf::heartbeat:LustreZFS):     Started kapollo02
 ost01  (ocf::heartbeat:LustreZFS):     Started kapollo01
 ost02  (ocf::heartbeat:LustreZFS):     Started kapollo02
 ost03  (ocf::heartbeat:LustreZFS):     Started kapollo01
 ost04  (ocf::heartbeat:LustreZFS):     Started kapollo02
 ost05  (ocf::heartbeat:LustreZFS):     Started kapollo01
 ost06  (ocf::heartbeat:LustreZFS):     Started kapollo02
 ost07  (ocf::heartbeat:LustreZFS):     Started kapollo01
 ost08  (ocf::heartbeat:LustreZFS):     Started kapollo02
 ost09  (ocf::heartbeat:LustreZFS):     Started kapollo01
 Clone Set: healthLNET-clone [healthLNET]
     Started: [ kapollo01 kapollo02 ]
 Clone Set: healthLUSTRE-clone [healthLUSTRE]
     Started: [ kapollo01 kapollo02 ]
```

Copyright ® Intel Corporation 2016. All rights reserved. Intel and the Intel logo are trademarks of Intel Corporation in the U.S. and/or other countries.
* Other names and brands may be claimed as the property of others.

Status of the cluster

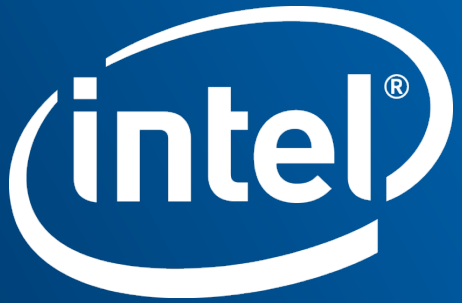
```
# pcs constraint show --full
Location Constraints:
  Resource: ost00
    Enabled on: kapollo01 (score:20) (id:location-ost00-kapollo01-20)
    Enabled on: kapollo02 (score:10) (id:location-ost00-kapollo02-10)
    Constraint: location-ost00
      Rule: score=-INFINITY boolean-op=or  (id:location-ost00-rule)
        Expression: pingd lt 1  (id:location-ost00-rule-expr)
        Expression: not_defined pingd  (id:location-ost00-rule-expr-1)
    Constraint: location-ost00-1
      Rule: score=-INFINITY boolean-op=or  (id:location-ost00-1-rule)
        Expression: lustred lt 1  (id:location-ost00-1-rule-expr)
        Expression: not_defined lustred  (id:location-ost00-1-rule-expr-1)
```

Copyright ® Intel Corporation 2016. All rights reserved. Intel and the Intel logo are trademarks of Intel Corporation in the U.S. and/or other countries.
* Other names and brands may be claimed as the property of others.

Conclusion

- HA implementation and tuning in Lustre is a complex topic
- RAs scripts are in production and heavily tested, but a wider adoption can improve robustness and add new functionalities like:
  - Support cluster with >2 nodes
  - Split ZFS import and Lustre mount scripts
  - Improve low level monitor capabilities (integration with ZED ?)

Copyright ® Intel Corporation 2016. All rights reserved. Intel and the Intel logo are trademarks of Intel Corporation in the U.S. and/or other countries.
* Other names and brands may be claimed as the property of others.


intel®

LDISKFS configuration for HA

Copyright ® Intel Corporation 2016. All rights reserved. Intel and the Intel logo are trademarks of Intel Corporation in the U.S. and/or other countries.
* Other names and brands may be claimed as the property of others.

Creation of the target's resource for LDISKFS

- Filesystem default script:

```
pcs resource create <Reource Name> ocf:heartbeat:Filesystem device="/dev/mapper/mpatha"
fstype="lustre" directory="/lustrefs/mgs"
```

  - `ocf:heartbeat:Filesystem.` The script must be located in /usr/lib/ocf/resource.d/heartbeat with permission 755 on all the lustre server nodes.
  - `device` (required): The name of block device for the filesystem, or -U, -L options for mount, or NFS mount specification.
  - `directory` (required): The mount point for the filesystem.
  - `fstype` (required): The type of filesystem to be mounted.

Copyright ® Intel Corporation 2016. All rights reserved. Intel and the Intel logo are trademarks of Intel Corporation in the U.S. and/or other countries.
* Other names and brands may be claimed as the property of others.