



Cluster Care: Reducing Downtime with Automated Node Failure Recovery

Robin Preble, HPC-OPS

Mentors: Conor Robinson and Graham Van Heule

August 8th, 2024

LA-UR-24-28425

The Problem

- Maintaining large clusters is labor-intensive
 - Thousands of nodes → frequent failures
 - Failures are often caused by well-documented issues with known solutions
- Manual intervention is time consuming and necessitates urgent responses
 - Employees may need to come in outside of regular working hours

Proposed Solution: Cluster Care

- Automate the detection and repair of failing nodes
- Benefits
 - Reduces workload for system administrators
 - Available 24/7
 - Improved availability for production workflows

Features: Modular Design

- Configurable
- Simple to maintain
- Adaptable

Example configuration file

```
src > etc > cluster_care > cluster_care.conf
1  # Individual module configs can be found in /etc/cluster_care/modules
2  [InputModules]
3  # Input modules are found at /var/cluster_care/modules/input
4  # Determine what input modules to use to determine node state, and by extension
5  # node reason (if the node is down) and whether a job is active.
6  state_modules=slurm,hcm_inf
7  # Determine what input modules to use to determine node type
8  type_modules=hcm_inf
9  # Determine what modules are used to check for messages with information about
10 # cluster state
11 message_modules=hcm_cluster
12
13 [ActionModules]
14 # Action Modules determine what action module to use for each action type.
15 # Action modules are located in /var/cluster_care/modules/action/.
16 # Each item here is in the format of <action type>=<action module>, and you
17 # can make any type of action type you want. So bob=bos_reboot, means that
18 # when trying to run the action bob on a group of nodes, it will use the
19 # bos_reboot action module.
20 reboot=shasta_bos_reboot
21 resume=slurm_resume
22 recheck=slurm_nhc_boot_recheck
23 update_firmware=slingshot_update_firmware
24
25 [Remediations]
26 profiles=RebootProfile,FirmwareProfile # Declare remediation profiles before defining them below
27 node_limit=50 # Maximum number of nodes you can perform an action on without manual override
28
29 [RebootProfile]
30 actions=reboot
31 types="compute"
32 states=1,2
33 reasons="slurm:Epilog error.*","slurm:Prolog error.*","slurm:NHC: check_ps_unauth_users.*"
34 job_active=0
35
36 [FirmwareProfile]
37 actions=update_firmware,reboot
38 types="compute"
39 states=1,2
40 reasons="slurm:NHC: hsn firmware version wrong*"
41 job_active=0
42
43
44
45
46
47
48
49
50
51
52
```

Features: Logging

- Activity is logged for tracking in Splunk
 - Verbose option available for additional details



```
2024-08-05T14:33:14.893768-06:00 rz-ncn-m003 cluster_care: {"category": "DEBUG", "type": "cluster_info", "sub_type": "node_data", "name": "ncn-w003", "node_type": "worker", "reason": [null], "job_active": null}
2024-08-05T14:33:14.899149-06:00 rz-ncn-m003 cluster_care: {"category": "DEBUG", "type": "cluster_info", "sub_type": "node_data", "name": "ncn-w004", "node_type": "worker", "reason": [null], "job_active": null}
2024-08-05T14:33:14.984398-06:00 rz-ncn-m003 cluster_care: {"category": "DEBUG", "type": "cluster_info", "sub_type": "node_data", "name": "ncn-w005", "node_type": "worker", "reason": [null], "job_active": null}
2024-08-05T14:33:14.989673-06:00 rz-ncn-m003 cluster_care: {"category": "DEBUG", "type": "cluster_info", "sub_type": "node_data", "name": "nid001000", "node_type": "compute", "reason": [null], "job_active": null}
2024-08-05T14:33:14.916073-06:00 rz-ncn-m003 cluster_care: {"category": "DEBUG", "type": "cluster_info", "sub_type": "node_data", "name": "nid001002", "node_type": "compute", "reason": [null], "job_active": null}
2024-08-05T14:33:14.922071-06:00 rz-ncn-m003 cluster_care: {"category": "DEBUG", "type": "cluster_info", "sub_type": "node_data", "name": "nid001061", "node_type": "compute", "reason": [null], "job_active": null}
2024-08-05T14:33:14.928414-06:00 rz-ncn-m003 cluster_care: {"category": "DEBUG", "type": "cluster_info", "sub_type": "node_data", "name": "nid001063", "node_type": "compute", "reason": [null], "job_active": null}
2024-08-05T14:33:14.938111-06:00 rz-ncn-m003 cluster_care: {"category": "INFO", "type": "remediation", "sub_type": "matches", "msg": "Found 5 nodes matching remediation profile 'RebootProfile'", "remediation_profile": "RebootProfile", "node_count": 5}
2024-08-05T14:33:23.561281-06:00 rz-ncn-m003 cluster_care: {"category": "INFO", "type": "action", "sub_type": "running_action", "msg": "Performing action 'reboot' on 5 nodes", "action_type": "reboot", "action_module": "shasta_bos_reboot", "nodes": ["nid001027", "nid001028", "nid001029", "nid001030", "nid001031"], "node_count": 5}
2024-08-05T14:33:29.156818-06:00 rz-ncn-m003 cluster_care: {"category": "INFO", "type": "action", "sub_type": "action_complete", "msg": "Finished 'reboot' action", "action_type": "reboot", "action_module": "shasta_bos_reboot", "nodes": ["nid001027", "nid001028", "nid001029", "nid001030", "nid001031"], "node_count": 5}
2024-08-05T14:33:29.164335-06:00 rz-ncn-m003 cluster_care: {"category": "INFO", "type": "remediation", "sub_type": "matches", "msg": "Found 3 nodes matching remediation profile 'FirmwareProfile'", "remediation_profile": "FirmwareProfile", "node_count": 3}
2024-08-05T14:33:32.418958-06:00 rz-ncn-m003 cluster_care: {"category": "INFO", "type": "action", "sub_type": "running_action", "msg": "Performing action 'update_firmware' on 3 nodes", "action_type": "update_firmware", "action_module": "slingshot_update_firmware", "nodes": ["nid001033", "nid001034", "nid001035"], "node_count": 3}
2024-08-05T14:35:01.146999-06:00 rz-ncn-m003 cluster_care: {"category": "INFO", "type": "action", "sub_type": "action_complete", "msg": "Finished 'update_firmware' action", "action_type": "update_firmware", "action_module": "slingshot_update_firmware", "nodes": ["nid001033", "nid001034", "nid001035"], "node_count": 3}
2024-08-05T14:35:01.152839-06:00 rz-ncn-m003 cluster_care: {"category": "INFO", "type": "action", "sub_type": "running_action", "msg": "Performing action 'reboot' on 3 nodes", "action_type": "reboot", "action_module": "shasta_bos_reboot", "nodes": ["nid001033", "nid001034", "nid001035"], "node_count": 3}
2024-08-05T14:35:06.762933-06:00 rz-ncn-m003 cluster_care: {"category": "INFO", "type": "action", "sub_type": "action_complete", "msg": "Finished 'reboot' action", "action_type": "reboot", "action_module": "shasta_bos_reboot", "nodes": ["nid001033", "nid001034", "nid001035"], "node_count": 3}
```

Example log file

Features: Interactive Mode

- Enables users to run the program manually

```
rz-ncn-m003:/users/rpreble/cluster_care # cluster_care -v
Malfunctioning and non-responsive nodes:
      NODES      STATE      REASON
      nid001005   drain    slurm:NHC: recheck
      nid001032   drain    slurm:NHC: cluster_care initiated reboot
      nid[001027-001031] drain    slurm:Epilog error.rpreble testing
      nid[001033-001035] drain    slurm:NHC: hsn firmware version wrong
      nid[001041,001047] drain    slurm:NHC: cluster_status initiated reboot
      nid[001048,001060,001062] down      slurm:Not responding

Found 5 nodes matching remediation profile 'RebootProfile'
      NODES      STATE      REASON
      nid[001027-001031] drain    slurm:Epilog error.rpreble testing

Prescribed remediation action(s): ['reboot']
Perform remediation on the above nodes? (y/n) y

Performing action 'reboot' on 5 nodes
Finished 'reboot' action
Found 3 nodes matching remediation profile 'FirmwareProfile'
      NODES      STATE      REASON
      nid[001033-001035] drain    slurm:NHC: hsn firmware version wrong

Prescribed remediation action(s): ['update_firmware', 'reboot']
Perform remediation on the above nodes? (y/n) y

Performing action 'update_firmware' on 3 nodes
Finished 'update_firmware' action
Performing action 'reboot' on 3 nodes
Finished 'reboot' action
rz-ncn-m003:/users/rpreble/cluster_care # █
```

Cluster Care running in
interactive mode

Questions?