RECENT WORK AND FUTURE PLANS FOR SBD

A very personal View

Klaus Wenninger
kwenning@redhat.com
September 6, 2017
SBD
Storage Based Death

- Fencing
- Watchdog Observation & Heartbeats
- ‘Poison Pill’ Messaging
- Suicide based on Quorum & Health
Is node X capable of causing corruption?

NO
POISON-PILL MESSAGING
Node-A fencing Node-B via shared Disk

Node-A puts Poison-Pill into Messaging-Slot of Node-B

<table>
<thead>
<tr>
<th></th>
<th>clear</th>
</tr>
</thead>
<tbody>
<tr>
<td>Node-A</td>
<td></td>
</tr>
<tr>
<td>Node-B</td>
<td>reset</td>
</tr>
<tr>
<td>Node-C</td>
<td></td>
</tr>
</tbody>
</table>

Node-B periodically checks Messaging-Slot
WATCHDOG LOOP
Basic Principle

➔ Simple Loop monitors Complex Software Component(s)
➔ (Hardware)-Watchdog triggered in this Loop
➔ Stuck Observation-Code >>> no Triggering
➔ (Hardware)-Watchdog >>> defined Reboot-Timeout
SBD DEPLOYMENT SCENARIOS
By Number of Shared Disks

NO SHARED STORAGE

- Solely based on Watchdog, Quorum & Node-Health Status from Pacemaker
- &ge; 3 Nodes required under all circumstances
- Quorum-Based-Fencing after stonith-watchdog-timeout

SINGLE SHARED DISK

- Quorate Partition survives without Disk thus Single Disk is no SPOF

3 SHARED DISKS

- Disks are redundant by themselves - Content if quorate Number of Disks (2) visible
- No Quorum Info from Pacemaker needed
SBD DEPLOYMENT SCENARIOS
Details and their Degree of Support

SUPPORTED  RECENTLY SUPPORTED  CANDIDATES
SBD DEPLOYMENT SCENARIOS
Support for 2-Node-Clusters with a single shared Disk

SINGLE SHARED DISK
➔ Quorate Partition survives without Disk thus Single Disk is not SPOF
➔ For getting usable Quorum Info from Pacemaker >=3 Nodes
➔ 2-Node-Clusters are always quorate when seen the Partner once

RHEL 7.4 approach for 2-Node-Clusters
➔ Quorum Info from Pacemaker useless
➔ dynamically read 2-Node-config from Corosync
➔ count Members in pacemaker-cpg-protocol
➔ Survive if either
  ◆ Both Nodes are in cpg-protocol
  ◆ Or Disk is available
SBD FUTURE FEATURES
Wish-List as of my very personal Perception

➔ Sharing of watchdog-device with other Consumers
  ◆ via systemd
  ◆ Replicated as device using cuse
  ◆ ...
➔ Heartbeat to Hypervisors without virtual /dev/watchdog
➔ Optional periodic Write-Access-Test to Storage
➔ (better) Support for pacemaker-remote Scenarios
➔ Arbitrary Mix of Nodes fenced via arbitrary Fencing-Methods in a single Cluster
  ◆ Quorum-based-watchdog-fencing
  ◆ Poison-Pill-Fencing
  ◆ Other pacemaker-supported Fencing Methods
➔ Easier & more foolproof Handling of Timeouts and their Dependencies
SYSTEMD AS WATCHDOG-PROVIDER
For SBD

➔ Have systemd handle /dev/watchdog
➔ Hardware-Watchdog observes the systemd ‘mainloop’
➔ Systemd provides Heartbeat-Observation for SBD
➔ Systemd has to be inspected carefully
  ◆ E.g. watchdog-triggered up to 50x in shutdown-retry-loop is a no-go for sbd-purposes - violates shutdown within defined timeout requirement
➔ Possible Solution for multiple Watchdog-Consumers on one System
➔ Systemd can start/trigger a Heartbeat-Daemon for Hypervisors like vmware & virtualbox
SBD & PACEMAKER-REMOTE

REDUNDANT DISKS
➔ Assure that Remote-Node-Name is used by SBD or map in Pacemaker

SETUPS IN NEED OF QUORUM-INFO FROM PACEMAKER
➔ Sit happily without Connection and no Resources running
➔ Don’t trigger Watchdog while disconnected with running Resources
➔ Some Experiments and Thoughts: https://github.com/ClusterLabs/sbd/pull/14
➔ Alternative: Disable sbd-fencing for Remote-Nodes
SINGLE SHARED DISK - FROM EACH SBD INSTANCE POV

➔ In a Split Situation the invisible internal Arbiter of the Disk-Replication-Solution switches one Side to fail on all Disk Accesses

➔ Especially interesting in 2-Node-Clusters

➔ Disk-Replication-Solution switches one side to read-only
  ◆ Prevents Data-Corruption
  ◆ basically desirable as outdated but consistent data is provided
  ◆ SBD doesn’t detect the read-only access
  ◆ Write & Readback would have to be implemented
THANK YOU

Klaus Wenninger
kwenning@redhat.com
September 6, 2017