Corosync and Qdevice news

Jan Friesse <jfriesse@redhat.com>

Clusterlabs Summit 2020
February 5, 2020

Part I

Corosync

Major changes in Corosync 3

- knet :)

Major changes in Corosync 3 (2)

- Enhanced statistics (`corosync-cmapctl -m stats`)
- Systemd startup notifications
- Environment variables removed (`COROSYNC_MAIN_CONFIG_FILE, ...`) and corosync arguments tied up
- Config file parser updated
- Reopening of log files supported (get rid of copytruncate logrotate method)
- Enable timestamps (and hi-res one)
- `ifdown` works (also for UDPU)
- Reduce totemsrp headers size (more space for application data)

Major changes in Corosync 3 (3)

- Remove CTS (not maintained and not very useful for Corosync), RDMA (unmaintained), Upstart, NSS, libcgroup
- totem is no longer shared library
- Blocking of unlisted IPs (knet ACL, UDPU)
- Consistent logging of node and ring ID
- Single CPG join list confchg event with all joined members is sent by node/group
- vqsim enhancements
- . . . and many many bug fixes

Future

- Improve config reload
- Synchronous logging of important events ("unexpected" fencing)
- Better statistics
- . . . and autotuning
- Integration of new knet features
- Corosync 4? Wire-compatibility?
- Move totemsrp into small, testable library (no network handling, no timers, . . .)?

Q&A

Part II

Qdevice

What is Qdevice?

- Independent arbiter for solving split-brain situations, stretch cluster
- Daemon running on every node of the cluster and using `Corosync votequorum` API and providing vote
- Currently only `net` model is implemented (`qdevice-net`)
- qdevice-net has support for multiple `algorithms` (LMS, FFSplit)
- Heuristics
  - Execute arbitrary number of commands
  - If all of them success whole heuristics sucess $\rightarrow$ no scoring

What is Qdevice? (cont Qnetd)

- 3rd side for Qdevice-net
- It is “clever” - responsible for decisions
- Supports TLS with both server and client (per cluster) certificates
- It’s able to handle multiple clusters
- No configuration file - all required information provided by cluster nodes
- No persistent state
- TCP based protocol designed with backwards/forwards compatibility in the mind since the very beginning

Major changes

- Not too much (Corosync 3 was a lot of work)
- Split from Corosync source tree (`https://github.com/corosync/corosync-qdevice`)
- Systemd startup notifications
- . . . and of course bugfixes

Future

- Clustered Qnetd (active/passive RA)
- Heuristics only model
  - Idea is to base vote only on heuristics result
  - Should be used in situations where 3rd side arbiter is already deployed
- Allow more than 1 vote for FFSplit
  - For situations when LMS is too strong and current FFSplit too weak
  - Be able to set arbitrary votes
- Redundant connections to Qnetd (important for LMS)
- Disk model as a closer replacement of `qdisk`

Q&A