Linux Cluster next generation

Vladislav Bogdanov

Heartbeat

- Simple (mostly two-node) clusters
- IP (UDP: unicast, broadcast, multicast) or serial communication
- Limited functionality (esp. haresources mode), no clustered storage support
- Later: added crm
- Then split:
  - Heartbeat (messaging + membership)
  - Resource agents (OCF)
  - Cluster-glue (lrmd, stonith, “plumbing”, IPC, logging)
  - Cluster Resource Manager -> pacemaker
- Now: deprecated (except resource-agents and pacemaker)

RHCS

- CMAN – messaging + quorum + API (then plugin to OpenAIS, so only quorum + API)
- DLM – kernel + userspace
- GFS (then GFS2) – kernel + userspace
- CLVM
- Fencing framework
- Rgmanager - cluster resource manager
- Later: pacemaker support for CMAN
- Later: fencing via pacemaker
- Then: project split, DLM and GFS-utils are now separate packages
- Now: everything except them is deprecated after RHEL6

OpenAIS

- Totem protocol messaging – virtual synchrony
- SA Forum APIs – CPG, AMF, CKPT, EVT, LCK
- UDP multicast and broadcast communication
- CMAN became a plugin to OpenAIS
- Pacemaker supports OpenAIS via plugin
- DLM, GFS2 (and OCFS2) work with OpenAIS via CMAN only
- CLVM (mostly) works with OpenAIS
- DLM, GFS2 and OCFS2 do not work reliably without CMAN
- Later: project split, Totem and CPG go to corosync (1.x)
- Now: discontinued

Pacemaker (2-3 years ago)

- Runs on top of heartbeat, openais (corosync 1.x)
- Provides quorum
- Supports:
  - Resource
  - Clone
    - Anonymous
    - Globally-unique
  - Master-slave (with multi-master support)
  - Resource migration
  - Order and colocation constraints
  - Groups
  - Node attributes (volatile and non-volatile)
- Fencing (STONITH) – heartbeat agents
- Later: support for CMAN (rather ugly) and RHCS fence agents

Corosync 2.x

- Totem + CPG
- UDP unicast in addition to multicast and broadcast
- Multi-ring (active and passive)
- Libqb based
- Quorum (votequorum)
- Single-threaded
- No plugins anymore (OpenAIS, CMAN and pacemaker via plugin are not compatible)
- CMAP – dynamic cluster reconfiguration
  - Nodelist
- Dynamic nodelist (UDPU) members via CMAP
- More info – corosync_overview(8), cmap_keys(8), corosync.conf(5)

Votequorum (corosync 2.x)

- Configurable node votes
- Expected votes (cluster-wide)
- Special features
  - Two-node mode
  - WFA (wait-for-all) – no quorum until all configured nodes are seen simultaneously
  - LMS (last-man-standing) – dynamic expected_votes and quorum recalculation (down)
  - ATB (auto-tie-breaker) – partition with node with lowest known id remains quorate even with 50% of votes
  - AD (allow-downscale) – decrease expected_votes and quorum on clean shutdown, down to configured expected_votes
- More info - votequorum(5)

DLM

- Kernel component:
  - TCP/SCTP communications for locking itself
  - Now: in-kernel interface for fs-control
  - Now: Additional sysctls instead of user-space configuration
- User-space component (dlm_controld)
  - Before:
    - CMAN for both membership and quorum
    - RHCS (fenced) for fencing
  - Then:
    - CPG for membership (but CPG_NODE_DOWN event is not handled)
    - CMAN for quorum
    - RHCS (fenced) API for fencing
  - Now:
    - CPG for membership
    - Corosync quorum
    - Stonithd (pacemaker) for fencing
    - FS-control support is deprecated but still exists

GFS2

- Kernel component:
  - DLM for locking
  - Before: FS-control via user-space
  - Now (Fedora 17+ and RHEL7): In-kernel FS-control
- User-space component (gfs_controld)
  - Before:
    - DLM for locking control
    - CMAN for membership
    - RHCS (fenced) for fencing
  - Now (Fedora 17+ and RHEL7): obsolete

CLVM

- Locking support: DLM
- Membership support:
  - Before:
    - corosync (1.x)
    - OpenAIS (with buggy LCK service)
    - CMAN
  - Now:
    - CPG (corosync 2.x)
    - Legacy
- Quorum: missing (!!!), global stop on graceful shutdown of cluster node. Developers are insane. Patch exists.

Pacemaker (now)

- Support for corosync 2.x (messaging, quorum, parts of configuration via CMAP)
- CPG membership.
- CMAN, heartbeat, openais (with plugin) support is deprecated.
- Libqb IPC
- Order-sets (A then (B and C) then D)
- Colocation sets (A (B C) D)
  - Sequential yes/no
- Full stonithd (fencing daemon) rewrite
  - Client API
  - Fencing topology (A or (B and C))
  - Used by DLM 4.x
  - Can work without the rest of pacemaker

Pacemaker (cont)

- Full LRMD rewrite
- Systemd resources
- Nagios checks for container resources (VM)
- Tickets
  - GEO-clustering (with booth – PAXOS algorithm implementation)
  - Tickets are volatile cluster attributes
- Still missing:
  - Non-volatile cluster attributes (can be implemented with tickets and migratable 'Ticketer' resource-agent)
  - Migration of 'first' resource causes 'then' resources to restart
- Being developed
  - Remote LRMD execution

Management tools

- Console:
  - crmsh (from SUSE). Actively developed.
  - pcs (from RedHat). New development.
- Web-based:
  - hawk (from SUSE)
  - pcs (from RedHat)
- GUI:
  - Pacemaker-GUI (almost not supported)
  - LCMC (Java)

Linux cluster (Fedora17 and RHEL7)

- Corosync 2.x
  - Provides CPG and quorum
- Fence-agents from RHCS
- OCF resource-agents from heartbeat
- Pacemaker:
  - Uses:
    - CPG and quorum from corosync
    - Fence-agents for fencing
    - Resource-agents for resource management
  - Provides:
    - CIB (Cluster information Base)
    - CRMd (Cluster Resource Manager daemon)
    - Stonith (fencing) daemon and API
- DLM 4.x
  - Uses CPG and quorum from corosync
  - Uses fencing via stonith API (pacemaker)
- CLVM
  - Uses CPG (and quorum if patched) from corosync
- GFS2 (with in-kernel FS-control)
  - Uses DLM kernel component directly

Questions?

www.sam-solutions.com

Value of Talent. Delivered.