Silke Meyer <s.meyer@heinlein-support.de>

Galera Cluster - Lessons learned

heinlein

Galera Cluster – Lessons learned
[SLAC 2017]
Silke Meyer <s.meyer@heinlein-support.de>

Ja, diese Folien stehen auch als PDF im Netz...

http://www.heinlein-support.de/vortrag

Überblick

- Kurze Einführung in Galera Cluster
- Wahl der passenden Replikationsmethode
- Datensicherung und -wiederherstellung
- Galera Arbitrator Daemon
- Monitoring von Galera Clustern
- Loadbalancing

Kurze Einführung in Galera Cluster

- MySQL-/MariaDB-Aufsatz für Cluster-Funktionalität
- synchrone Multi-Master Replikation
- entwickelt von Codership
- drei Distributionen (verändertes MySQL plus Galera):
  - Codership
  - MariaDB Cluster
  - Percona XtraDB Cluster
- ausgelegt für eine ungerade Anzahl von Knoten zur Vermeidung von Split-Brain beim Ausfall eines Knotens

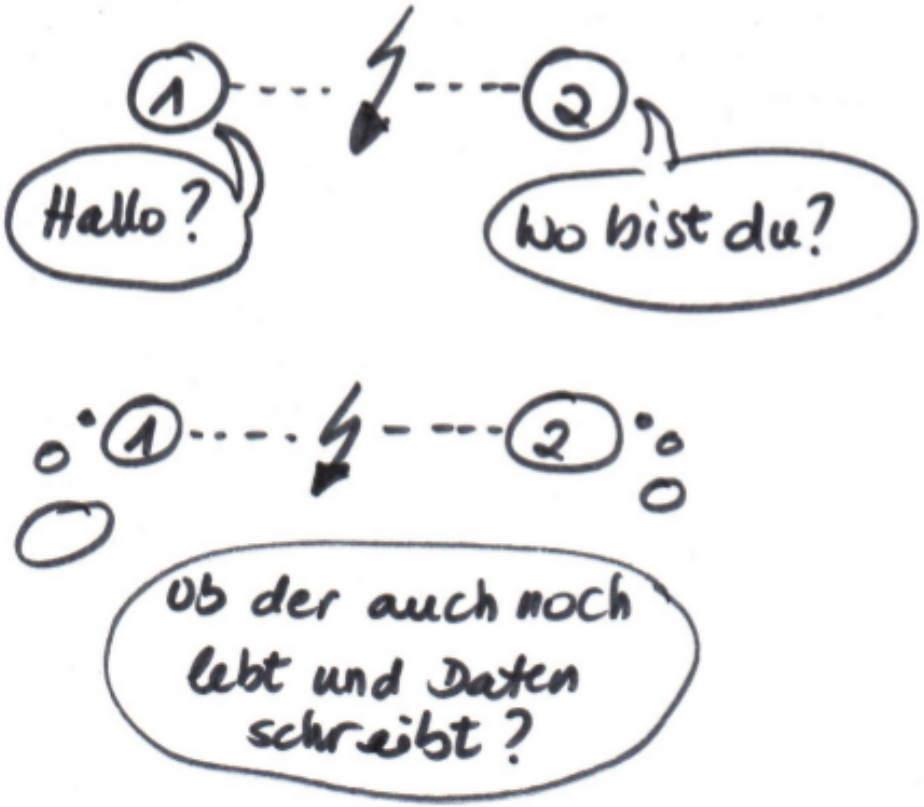
Features

- Paralleles Lesen und Schreiben auf alle Nodes, kein Slave Lag
- automatische „membership control“
- automatisches Failover / High Availability
- Wartung einzelner Nodes im laufenden Betrieb
- fühlt sich an wie MySQL / MariaDB

Die Sache mit dem Quorum

- Vermeidung von Split-Brain-Situationen
- Die „Primary Component“ ist befugt, Daten auszuliefern (wsrep_cluster_status primary).
- Beispiel für Node 1 von 3, der von der Mehrheit getrennt wird
  - wsrep_cluster_size 1
  - wsrep_local_state_comment Initialized
  - wsrep_cluster_status **non-primary**

Die Sache mit dem Quorum



Silke Meyer <s.meyer@heinlein-support.de>

Konfiguration I

```
bind-address = 0.0.0.0
binlog_format=ROW
default_storage_engine=InnoDB
innodb_autoinc_lock_mode = 2
innodb_flush_log_at_trx_commit = 0
query_cache_size = 0
query_cache_type = 0
```

Konfiguration II

```
wsrep_provider=/usr/lib/galera/libgalera_smm.so

wsrep_cluster_address="gcomm://10.192.7.11,10.192.7.
12,10.192.7.13"

wsrep_node_name = galeratest1

wsrep_sst_method = xtrabackup-v2

wsrep_sst_auth = sst:secret
```

Dateien in /var/lib/mysql

- galera.cache: write-set cache (gcache) für IST
- grastate.dat: Status des Knotens
- gvwstate.dat: Status der Primary Component aus Sicht des Nodes
- GRA_*.log: Logging von Replikationsfehlern

Wahl der passenden Methode für State Transfer

- SST (Snapshot State Transfer) versus IST (Incremental State Transfer)
- Binary Logs nur für Point in Time Recovery
- Methoden für SST:
  - logisch: mysqldump
  - physikalisch: rsync
  - physikalisch: xtrabackup / xtrabackup-v2 ← empfohlen
- Größe des gcache bestimmt, wie lange ein Knoten getrennt sein kann, ohne einen Full Sync (SST) zu brauchen.
- Beim State Transfer geht der Donor Node in den Status Donor/Desynced.

Datensicherung und -wiederherstellung

- Verbreitet: Asynchroner Slave an einem Galera-Node (klass. Master-Slave)
- Laut Doku: Backup-Skript triggern über Galera Arbitrator Daemon
- Wiederherstellung
  - Backup holen
  - 1. Node: `mysqld_safe --wsrep-new-cluster`
  - Weitere Nodes normal starten
  - „`service mysql start … failed!`“ kann gelogen sein! Log lesen!
  - systemd unterstützt kein Bootstrappen mit `--wsrep-new-cluster`

Galera Arbitrator Daemon (garbd) I

- Quorumsbeschaffung bei Cluster mit gerader Anzahl von Nodes
- Beispiel: Kunde mit 2 Web- und 2 Datenbank-Servern, garbd auf einem der Webserver mitlaufen lassen

```
root@arbi:~# cat /etc/default/garb
GALERA_NODES="10.192.7.11:4567, 10.192.7.12:4567"
GALERA_GROUP="Galera Testcluster"
LOG_FILE="/var/log/garbd.log"
```

Demo: Quorumsbeschaffung mit garbd

- Wir haben 2 Datenbank-Server und garbd auf einem Webserver.
- Szenario I:
  - 1. DB-Server und garbd laufen durch
  - 2. DB-Server fällt aus
  - Cluster läuft weiter.
- Szenario II:
  - Beide Datenbank-Server und garbd laufen.
  - Das Netzwerk zwischen den DB-Servern ist kaputt.
  - Cluster läuft weiter.

Galera Arbitrator Daemon (garbd) II

- Triggern eines Backup-Skriptes im Cluster bei erfolgreicher Verbindung mit der primären Komponente
- Keine Sicherung von inkonsistentem Zustand
- Eigenes Backup-Skript: /usr/bin/wsrep_sst_**backup**

```
root@arbi:~# cat /etc/cron.d/galera-backup-trigger

30 2 * * * root /usr/bin/garbd --address
gcomm://10.192.7.13:4567?
gmcast.listen_addr=tcp://0.0.0.0:4444 --group
"Galera Testcluster" --donor galeratest3 --sst
backup --log /var/log/garbd/backup-galera-
testcluster.log
```

Monitoring I

- das gewohnte DB-Monitoring weiternutzen ;)
- jeden Galera-Knoten monitoren
- Monitoring des Netzwerks dazwischen
- Galera-Status (Client-Abfragen):

```
MariaDB> show global status like „wsrep%“;
wsrep_local_state_comment Synced
wsrep_cluster_size 3
wsrep_cluster_status Primary
wsrep_ready ON
```

Monitoring II: Bottleneck Detection

- `wsrep_flow_control_paused:` Wie viel Prozent der Zeit hat das Cluster auf einen langsamen Node gewartet?
- `wsrep_flow_control_sent:` Wie oft hat diese Node darum gebeten, dass die anderen auf ihn warten?
- `wsrep_local_bf_aborts` und `wsrep_local_cert_failures:` Aufzeichnung nicht erfolgreich abgeschlossener Transaktionen
- `State „query end“ in SHOW PROCESSLIST?`

Loadbalancing

- Gute Erfahrungen mit HAProxy
- weg von der Appliance
- nicht alle Datenbankanfragen über ein und dieselbe Service-IP schicken auf der Appliance schicken
- Jeder Client kann feiner eingestellt werden und selbst die Verfügbarkeit der Datenbank prüfen.
- (Noch?) keine Erfahrungen mit Galera Loadbalancer… anyone?

Fragen?
Diskussionsbedarf?

- Danke für das Interesse!
- Natürlich und gerne stehe ich Ihnen jederzeit mit Rat und Tat zur Verfügung und freue mich auf neue Kontakte.
  - Silke Meyer
  - Mail: s.meyer@heinlein-support.de
  - Telefon: 030/40 50 51 – 51
- Wenn's brennt:
  - Heinlein Support 24/7 Notfall-Hotline: 030/40 505 - 110

Wir suchen:
Admins, Consultants, Trainer!

Wir bieten:
Spannende Projekte, Kundenlob, eigenständige Arbeit, keine Überstunden, Teamarbeit

...und natürlich: Linux, Linux, Linux...

http://www.heinlein-support.de/jobs

Silke Meyer <s.meyer@heinlein-support.de>

heinlein

Heinlein Support hilft bei allen Fragen rund um Linux-Server

HEINLEIN AKADEMIE

Von Profis für Profis: Wir vermitteln die oberen 10% Wissen: geballtes Wissen und umfangreiche Praxiserfahrung.

HEINLEIN HOSTING

Individuelles Business-Hosting mit perfekter Maintenance durch unsere Profis. Sicherheit und Verfügbarkeit stehen an erster Stelle.

HEINLEIN CONSULTING

Das Backup für Ihre Linux-Administration: LPIC-2-Profis lösen im CompetenceCall Notfälle, auch in SLAs mit 24/7-Verfügbarkeit.

HEINLEIN ELEMENTS

Hard- und Software-Appliances und speziell für den Serverbetrieb konzipierte Software rund ums Thema eMail.