Your first date with Percona XtraDB Cluster

Krunal Bauskar
PXC Product Lead @ Percona
Agenda

- Why we need synchronous replication?
- How PXC helps achieve it
- Key features of PXC
- 5.7 exclusive features
- PXC as a complete HA solution
Traditional MySQL replication
Traditional MySQL replication
Traditional MySQL replication

- Replication delay (loss of transactions).
- Limited read scalability.
- No option for write-scalability.
- Switchover interval.
- No protection against network failure.
In search of better solution….

- What if all nodes can act as active master
- No single point failure
- Can protect against network outages
- Read/Write-scalability
- Easy to maintain/deploy
Percona XtraDB Cluster (PXC)

- PXC is multi-master solution implemented using
  - Percona-Server (from Percona (based on MySQL))
  - wsrep-replication plugin (from Codership)
  - Galera replication library (from Codership)

- PXC is independent product with complete life-cycle.

- PXC inherited from upstream products but has tons of local changes making it one of the best MySQL clustering solution.
PXC - Multi-Master

- Write to any node of the cluster and write will be replicated transparently to all other nodes of the cluster.
- Read from any node of the cluster and get consistent view of the data.
- Synchronous replication*
PXC - Multi-Master

- Write to any node of the cluster and write will be replicated transparently to all other nodes of the cluster
- Read from any node of the cluster and get consistent view of the data.
- Synchronous replication*
PXC - Multi-Master

- Write to any node of the cluster and write will be replicated transparently to all other nodes of the cluster.
- Read from any node of the cluster and get consistent view of the data.
- Synchronous replication*
But what makes PXC enterprise ready?

- Automatic Node Provisioning
- Ability to handle conflicting workload
- Flow-control
- Parallel processing
- Protection against network failure

- Cluster-safe-mode
- Security
- Performance
- Trackability
- Geo-distributed
Automatic Node Provisioning

- Bootstrapping cluster
- SST (State Snapshot Transfer)
  - rsync/mysqldump/xtrabackup*
- IST (Incremental Snapshot Transfer)
  - auto-catchup cluster state
Automatic Node Provisioning

- Bootstrapping cluster
- SST (State Snapshot Transfer)
  - rsync/mysqldump/xtrabackup*
- IST (Incremental Snapshot Transfer)
  - auto-catchup cluster state
Automatic Node Provisioning

- Bootstrapping cluster
- SST (State Snapshot Transfer)
  - rsync/mysqldump/xtrabackup*
- IST (Incremental Snapshot Transfer)
  - auto-catchup cluster state
Automatic Node Provisioning

- Bootstrapping cluster
- SST (State Snapshot Transfer)
  - rsync/mysqldump/xtrabackup*
- IST (Incremental Snapshot Transfer)
  - auto-catchup cluster state
Automatic Node Provisioning

- Bootstrapping cluster
- SST (State Snapshot Transfer)
  - rsync/mysqldump/xtrabackup*
- IST (Incremental Snapshot Transfer)
  - auto-catchup cluster state
Automatic Node Provisioning

- Bootstrapping cluster
- SST (State Snapshot Transfer)
  - rsync/mysqldump/xtrabackup*
- IST (Incremental Snapshot Transfer)
  - auto-catchup cluster state
Automatic Node Provisioning

- Bootstrapping cluster
- SST (State Snapshot Transfer)
  - rsync/mysqldump/xtrabackup*
- IST (Incremental Snapshot Transfer)
  - auto-catchup cluster state
Automatic Node Provisioning

- Bootstrapping cluster
- SST (State Snapshot Transfer)
  - rsync/mysqldump/xtrabackup*
- IST (Incremental Snapshot Transfer)
  - auto-catchup cluster state

SST OR IST
Automatic Node Provisioning

- Bootstrapping cluster
- SST (State Snapshot Transfer)
  - rsync/mysqldump/xtrabackup*
- IST (Incremental Snapshot Transfer)
  - auto-catchup cluster state
Automatic Node Provisioning

- Bootstrapping cluster
- SST (State Snapshot Transfer)
  - rsync/mysqldump/xtrabackup*
- IST (Incremental Snapshot Transfer)
  - auto-catchup cluster state
Automatic Node Provisioning

Automatic
SCALE IN - SCALE OUT
Handling workload conflicts

- **Optimistic Locking.** (No distributed locks)
- **Brute force abort:**
  - Forceful abort of conflicting transaction
- **Certification failure**
  - 2 conflicting transactions replicate at same time
Handling workload conflicts

- **Optimistic Locking.** (No distributed locks)

- **Brute force abort:**
  - Forceful abort of conflicting transaction

- **Certification failure**
  - 2 conflicting transactions replicate at same time

![Diagram showing conflict resolution](image)
Handling workload conflicts

- Optimistic Locking. (No distributed locks)
- Brute force abort:
  - Forceful abort of conflicting transaction
- Certification failure
  - 2 conflicting transactions replicate at same time

```
update t set i = 100
update t set i = 101
```
Handling workload conflicts

- Optimistic Locking. (No distributed locks)
- Brute force abort:
  - Forceful abort of conflicting transaction
- Certification failure
  - 2 conflicting transactions replicate at same time
Handling workload conflicts

- **Optimistic Locking.** (No distributed locks)
- **Brute force abort:**
  - Forceful abort of conflicting transaction
- **Certification failure**
  - 2 conflicting transactions replicate at same time
Handling workload conflicts

- Optimistic Locking. (No distributed locks)
- Brute force abort:
  - Forceful abort of conflicting transaction
- Certification failure
  - 2 conflicting transactions replicate at same time

```
update t set i = i + 1
update t set i = i + 2
```
Handling workload conflicts

▪ Optimistic Locking. (No distributed locks)
▪ Brute force abort:
  • Forceful abort of conflicting transaction
▪ Certification failure
  • 2 conflicting transactions replicate at same time
Handling workload conflicts

- Optimistic Locking. (No distributed locks)
- Brute force abort:
  - Forceful abort of conflicting transaction
- Certification failure
  - 2 conflicting transactions replicate at same time
Handling workload conflicts

Simply execute transaction.
Leave consistency to PXC.
Flow Control

- Dynamic control the workload processing to adjust nodes processing ability.

- Trx are queued. Queue full can cause flow-control.
Flow Control

- Dynamic control the workload processing to adjust nodes processing ability.

- Trx are queued. Queue full can cause flow-control.
Flow Control

- Dynamic control the workload processing to adjust nodes processing ability.

- Trx are queued. Queue full can cause flow-control.
Flow Control

- Dynamic control the workload processing to adjust nodes processing ability.

- Trx are queued. Queue full can cause flow-control.
Flow Control

- Dynamic control the workload processing to adjust nodes processing ability.

- Trx are queued. Queue full can cause flow-control.
Flow Control

- Dynamic control the workload processing to adjust nodes processing ability.

- Trx are queued. Queue full can cause flow-control.
Flow Control

- Dynamic control the workload processing to adjust nodes processing ability.

- Trx are queued. Queue full can cause flow-control.
Flow Control

Dynamic workload adaptation to maintain cluster state
Parallel Processing

- Multiple parallel worker threads to match-up with user-level workload threads

- They all beautifully coordinate to let the “FIRST COMMITTER WIN” (Commit/Local/Apply Monitor)
Parallel Processing

Seamless scalability through multi-threading
Network failure

- Quorum lost
  - Network outage, Machine failure
- Auto-recovery: Restart the cluster once all the nodes are back.
- Configurable timeout to adjust different network.
- Protection against Split-Brain
Network failure

- Quorum lost
  - Network outage, Machine failure
- Auto-recovery: Restart the cluster once all the nodes are back.
- Configurable timeout to adjust different network.
- Protection against Split-Brain
Network failure

- Quorum lost
  - Network outage, Machine failure
- Auto-recovery: Restart the cluster once all the nodes are back.
- Configurable timeout to adjust different network.
- Protection against Split-Brain
Network failure

- Quorum lost
  - Network outage, Machine failure
- Auto-recovery: Restart the cluster once all the nodes are back.
- Configurable timeout to adjust different network.
- Protection against Split-Brain
Network failure

- Quorum lost
  - Network outage, Machine failure
- Auto-recovery: Restart the cluster once all the nodes are back.
- Configurable timeout to adjust different network.
- Protection against Split-Brain
Network failure

- Quorum lost
  - Network outage, Machine failure
- Auto-recovery: Restart the cluster once all the nodes are back.
- Configurable timeout to adjust different network.
- Protection against Split-Brain
Network failure

- Quorum lost
  - Network outage, Machine failure
- Auto-recovery: Restart the cluster once all the nodes are back.
- Configurable timeout to adjust different network.
- Protection against Split-Brain
Network failure

Auto rejoin and protection against incomplete view
Geo-distributed

- **Advance Availability** with Geo-distribution of cluster nodes

Data-Center-1

Data-Center-2
Geo-distributed

- **Advance Availability** with Geo-distribution of cluster nodes

```
Data-Center-1
  NU1 ┌─┐  NU2 ─NU3
  |   └─┘
  NA1 └─┐
        │
        │
        │
        NA2 ─NA3
Data-Center-2
```
Lot more flexible topology

- PXC cluster can act ASYNC SLAVE, ASYNC MASTER.
- 2 PXC cluster could be connected through a ASYNC replication link.
Flexible topology

Flexible Topology
Cluster-Safe-Mode [5.7]

- Blocks workload that is not cluster safe.
- `pxc_strict_mode`
  - **ENFORCING** (default)
  - MASTER
  - PERMISSIVE
  - DISABLED
Security

- **Data secure:**
  - In-flight/during transit (SST, IST, replication traffic through SSL sockets)
  - at-rest (through encrypted tablespace) [5.7]

[Single option configuration supported `pxc-encrypt-cluster-traffic=ON/OFF`][5.7]
Performance [5.7]

- Industry Leading Performance Solution (5.7.17+ onwards)
- For all workload: OLTP/POINT UPDATE/etc…
- Quicker node rejoins
Performance [5.7]

- Industry Leading Performance Solution (5.7.17+ onwards)
- For all workload: OLTP/POINT UPDATE/etc…
- Quicker node rejoins
Trackability

- Can track lot of important aspects including
  - show status (Flow-control, IST progress, queue size, replication latency, applied/committed upto, cache size)
  - performance_schema.pxc_cluster_view
  - Various other internal objects exposed through PFS
  - Error/information logs.
  - Improved debugging messages for SST
  - Improved visibility of thread state through show processlist
PXC as a complete HA solution

- Complete HA solution should have:
  - Stable cluster
  - Load Balancer
  - Simplified and Unified view of the system
Load Balancer

- PXC can operate with multiple load balancers like HAProxy, ProxySQL, etc...
- PXC suggest use of ProxySQL
  - Integrated and closed development.
  - Feature rich load balancer (lot more features getting added)
  - Custom PXC script aids simplified PXC configuration (auto-discovery of PXC nodes).
Load Balancer

- **PXC maintenance mode**
  - Abrupt graceful shutdown can cause disruption in workload till Load-Balancer readjust the load.
  - Maintenance mode is like advance warning helping load balancer (ProxySQL only) to make a note of this and pre-adjust workload.
  - Also applicable if node needs to stop active traffic for maintenance purpose.

```
pxc_maint_mode
DISABLED
SHUTDOWN
MAINTENANCE
pxc_maint_transition_period
```
Simplified and Unified view

- PMM (Percona Monitoring and Management)
  - Complete integrated with PXC
  - Simple way to take control of complete HA system

https://pmmddemo.percona.com/
PXC as a complete HA solution

- Complete HA solution should have:
  - Stable cluster
  - Load Balancer
  - Simplified and unified view of the system
Connect back.....

- mail me: krunal.bauskar@percona.com
- PXC forum https://www.percona.com/forums/questions-discussions/percona-xtradb-cluster
- PXC @ JIRA https://jira.percona.com/projects/PXC/issues

Q&A

RATE

THE TALK
Thank You Sponsors!!

Alibaba Cloud
Microsoft
VividCortex

DATACORE
facebook
Grafana
Shannon Systems
timescale
yelp
box
solarwinds

Dynimize
TwinDB
PingCAP
Altinity
ObjectRocket
MySQL
Navicat
GitHub
cPanel
BlazingDB
Google Cloud

datanami
HPCwire
database
Packt
ODBMS.org
EnterpriseTech