# Understanding Fault Tolerance

ClustrixDB provides fault tolerance by maintaining multiple copies of data throughout the cluster. This enables a cluster to experience the loss of node(s) or zone(s), without data loss and allowing the cluster to automatically resume operations.



## Built-in Fault Tolerance

By default, ClustrixDB is configured to accommodate a single node failure and automatically maintain 2 copies (replicas) of all data. As long as the cluster has sufficient replicas and a quorum of nodes is available, a cluster can lose a node without experiencing any data loss. Clusters with zones configured can lose a single zone.

### Configuring Fault Tolerance

The default settings for fault tolerance are generally acceptable for most clusters. However, as your cluster expands or contracts, or if your cluster is deployed in various Zones, you may want to adjust fault tolerance settings.

## Deploying Across Zones

ClustrixDB can be configured to be zone aware so that replicas (and acceptors) are place across different zones (AWS Availability Zones within the same Region, different server racks, different network switches, different power sources). When zones are configured, a cluster can lose an entire zone and automatically recover without loss of data.

## Using Replication

Setting up disaster recovery site for your ClustrixDB cluster will allow you to recover from catastrophic failures. Setting up a secondary ClustrixDB cluster for DR will also allow for easier transition to new releases. For information regarding the various replication configurations supported by ClustrixDB, please see Configuring Replication.

## MAX_FAILURES

ClustrixDB can be configured to survive more than one node (or zone) failure by changing the value of MAX_FAILURES, ensuring that all tables have additional replicas (and sufficient disk space), and that the cluster has a sufficient number of nodes (and zones). See ALTER CLUSTER SET MAX_FAILURES for more information.

> Maintaining additional replicas has a performance overhead, so increasing MAX_FAILURES increases latency for your cluster

## What Happens When a Node or Zone Fails?

When ClustrixDB experiences a node or zone failure:

- A node or zone fails a heartbeat check, meaning a node of zone can no longer communicate with other nodes of the cluster.
- A short group change occurs, after which the database becomes available (See Group Changes for more information).
- The timer associated with the global rebalancer_reprotect_queue_interval_s begins (default = 10 minutes).
- ClustrixDB establishes a queue of pending changes for the node's data and tracks all pending changes for that node or zone in that queue. This queue is necessary only if the failure is temporary.
- If the node returns within rebalancer_reprotect_queue_interval_s seconds, queued transactions are applied to the previously failed node(s) and processing resumes.
- If the node does not return within rebalancer_reprotect_queue_interval_s seconds, the queued transactions become unnecessary as the

Rebalancer will begin copying data of the failed node(s) or zone from replica versions in the cluster. To monitor this reprotect process, see Managing the Rebalancer.

- When the reprotect process completes, ClustrixDB will send a message indicating that full protection has been restored using Database Alerts. An entry to that effect also appears in the query.log.
- The failed/unavailable node(s) may then be removed from the cluster. See ALTER CLUSTER DROP.

## What Happens to Processes That Were Running?

Processes that were running when the heartbeat check failed will be impacted as shown:

| Process | Result |
|---|---|
| Queries (DML and DDL) | If the global autoretry is true and a transaction was submitted with autocommit enabled, the database will automatically retry any transactions that were in process following the creation of a new group. If the retried statements are unsuccessful, the application will receive an error.<br><br>ClustrixDB reads data exclusively from ranked replicas. If the failed node(s) contained any ranked replicas for a slice, ClustrixDB assigns that role to another replica of that slice elsewhere. |
| Replication | Replication processes will automatically restart at the proper binlog location following the group change. |
| Other Connections | Connections to nodes that are still in quorum will be reestablished and a new group will be formed with the available nodes.<br><br>Connections to non-communicative nodes will be lost. |