# GR Failover Triggers and Scenarios



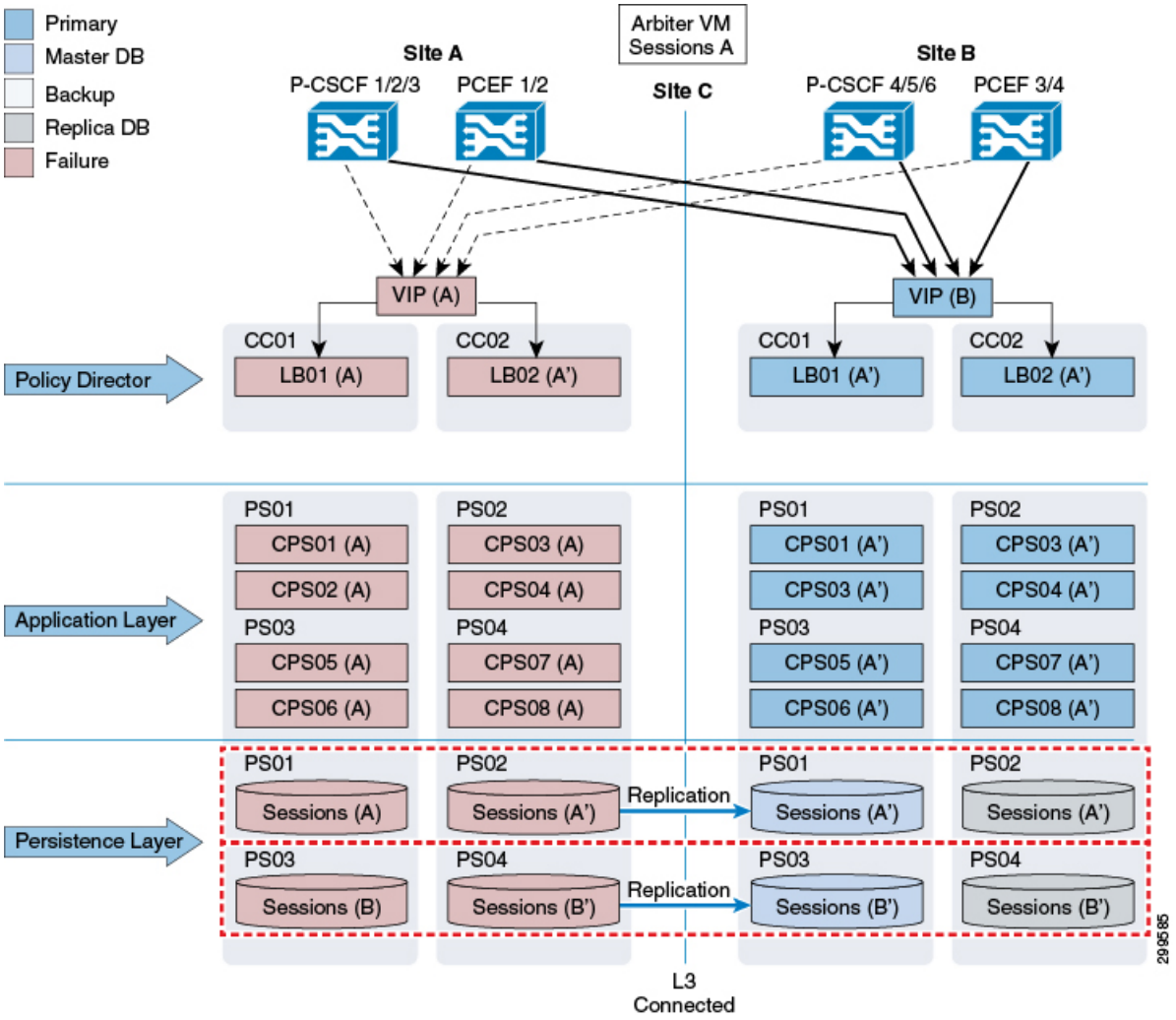
# Failover Triggers and Scenarios

In Geographic Redundancy, there are multiple scenarios which could trigger a failover to another site.

## Site Outage

As shown in the figure below, all P-GWs and P-CSCFs will direct traffic to the secondary site in the event of a complete outage of the primary site. Failover time will be dependent on failure detection timers on the P-GW and P-CSCF and the time it takes for the database replica set to elect a new Master database at the secondary site.

*Figure 1: Outage of Primary Site*




In order for Site A to be considered "ready for service" after an outage, all 3x tiers (Policy Director, Application Layer and Persistence Layer) must be operational.

At the Persistence (database replica set) level, MongoDB uses an operations log (oplog) to keep a rolling record of all operations that modify the data stored in the database. Any database operations applied on the Primary node are recorded on its oplog. Secondary members can then copy and apply those operations in an asynchronous process. All replica set members contain a copy of the oplog, which allows them to maintain the current state of the database. Any member can import oplog entries from any other member. Once the oplog is full, newer operations overwrite older ones.

When the replica members at Site A come back up after an outage and the connectivity between Sites A and B is restored, there are two possible recovery scenarios:

1. The oplog at Site B has enough history to fully resynchronize the whole replica set, for example the oplog did not get overwritten during the duration of the outage. In this scenario, the database instances at Site A will go into "Recovering" state once connectivity to Site B is restored. By default, when one of those instances catches up to within 10 seconds of the latest oplog entry of the current primary at Site B, the set will hold an election in order to allow the higher-priority node at Site A to become primary again.

2. The oplog at Site B does not have enough history to fully resynchronize the whole replica set (the duration of the outage was longer than what the system can support without overwriting data in the oplog). In this scenario, the database instances at Site A will go into "Startup2" state and stay in that state until we manually force a complete resynchronization (as they would be too stale to catch up with the current primary. A "too stale to catch up" message will appear in the mongodatabase.log or in the errmsg field when running rs.status(). For more information on manual resynchronization, Manual Recovery.

   During a complete resynchronization, all the data is removed from the database instances at Site A and restored from Site B by cloning the Site B session database. All Read and Write operations will continue to use Site B during this operation.

Recovery time, holding time for auto recovery and so on depends upon TPS, latency, oplog size. For optimum values, contact your Cisco Technical Representative.

In CPS Release 7.5.0 and higher releases, at the Policy Director level, there is an automated mechanism to check availability of the Master database within the local site. When the Master database is not available, the policy director processes will be stopped and will not process with any incoming messages (Gx/Rx).

- This check runs at Site A (primary site).
- This check runs every 5 seconds (currently not configurable) and will determine whether the Master Sessions database is at Site A.

  It is possible to configure which databases the script will monitor (Sessions, SPR, Balance). By default, only the Sessions database is monitored.
- If the Master database is not available at Site A, the two Policy Director Processes (Loadatabasealancers) of site A will be stopped or remain stopped if recovering from a complete outage (as described in this section).
- In case of two replica sets, if one of the two Replica sets Master database is not available at Site A, the two Policy Director Processes (Loadatabasealancers) of site A will be stopped or remain stopped if recovering from a complete outage and the second replica set Master database will failover from Site A to Site B.

These above mentioned checks will prevent cross site communication for read/write operations. Once the site is recovered, P-GWs and P-CSCFs will start directing new sessions to Site A again.

For existing sessions, P-GWs will continue to send traffic to Site B until a message for the session (RAR) is received from Site A. That will happen, for example, when a new call is made and the Rx AAR for the new session is sent by the P-CSCF to Site A. Also, for existing Rx sessions, the P-CSCF will continue to send the traffic to Site B.

# Gx Link Failure

As shown in the figure below, failure of the Gx link between a P-GW and the primary CPS node (Site A) results in the P-GW sending traffic to the secondary site (Site B). Failover time depends on failure detection timers on the P-GW.

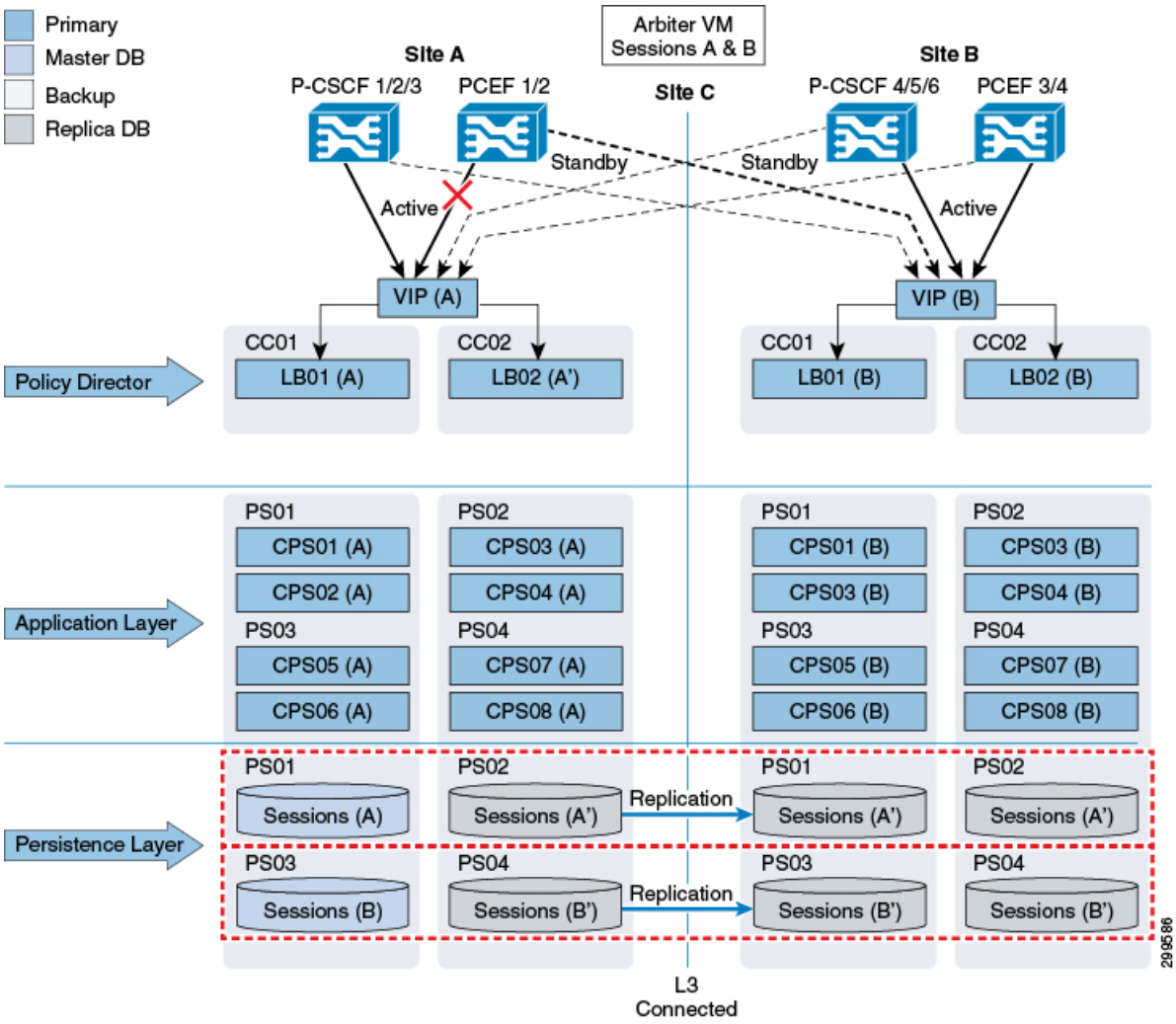
Gx transactions are processed at Site B.

If a session already exists, the CPS0x(B) VM handling the transaction at Site B retrieves the subscriber's session from the Master Sessions (A) database at Site A. New sessions as well as session updates are written across to the Master database at Site A.

Gx responses towards the P-GW (for example CCA), as well as Rx messages such as ASR that may be generated as a result of Gx transaction processing, is sent from Site B.

After receiving an Rx AAR at Site A, the resulting Gx RAR is proxied from the lb at Site A to the lb at Site B (as the P-GW is not reachable from Site A).

***Figure 2: Gx Link Failure***




**Note** For SP Wi-Fi deployments, if a link fails between PCEF 1/2 and CPS Site A, all messages coming from PCEF 1/2 to Site B are processed but messages generated from Site A for PCEF 1/2 are not proxied from Site B. P-CSCF communication is not applicable for SP Wi-Fi deployments.

**Note** RADIUS-based policy control is no longer supported in CPS 14.0.0 and later releases as 3GPP Gx Diameter interface has become the industry-standard policy control interface.

## Failover Time Improvement

The subscriber impact for data calls (Gx only) has been reduced below 1 second for failures and recovery scenarios captured in following table:

***Table 1: Failure Triggers Supported for Data (Gx) Call***

| Failover Trigger Scenarios | Operational Impact | Timeout Durations (ms) |
|---|---|---|
| VM shutdown on both primary and secondary members on local site while other replica-sets are running | Shutdown during Maintenance Window of multiple VM's on local site | 800 ms or better |
| All replica-sets of local site are brought down | Power OFF of all session manager VM's | 800 ms or better |
| VM startup of both failed local members in a replica-set | Startup or multiple VM's during Maintenance Window | 800 ms or better |
| Replication or internal network link between two local members of a replica-set is down | None | 800 ms or better |

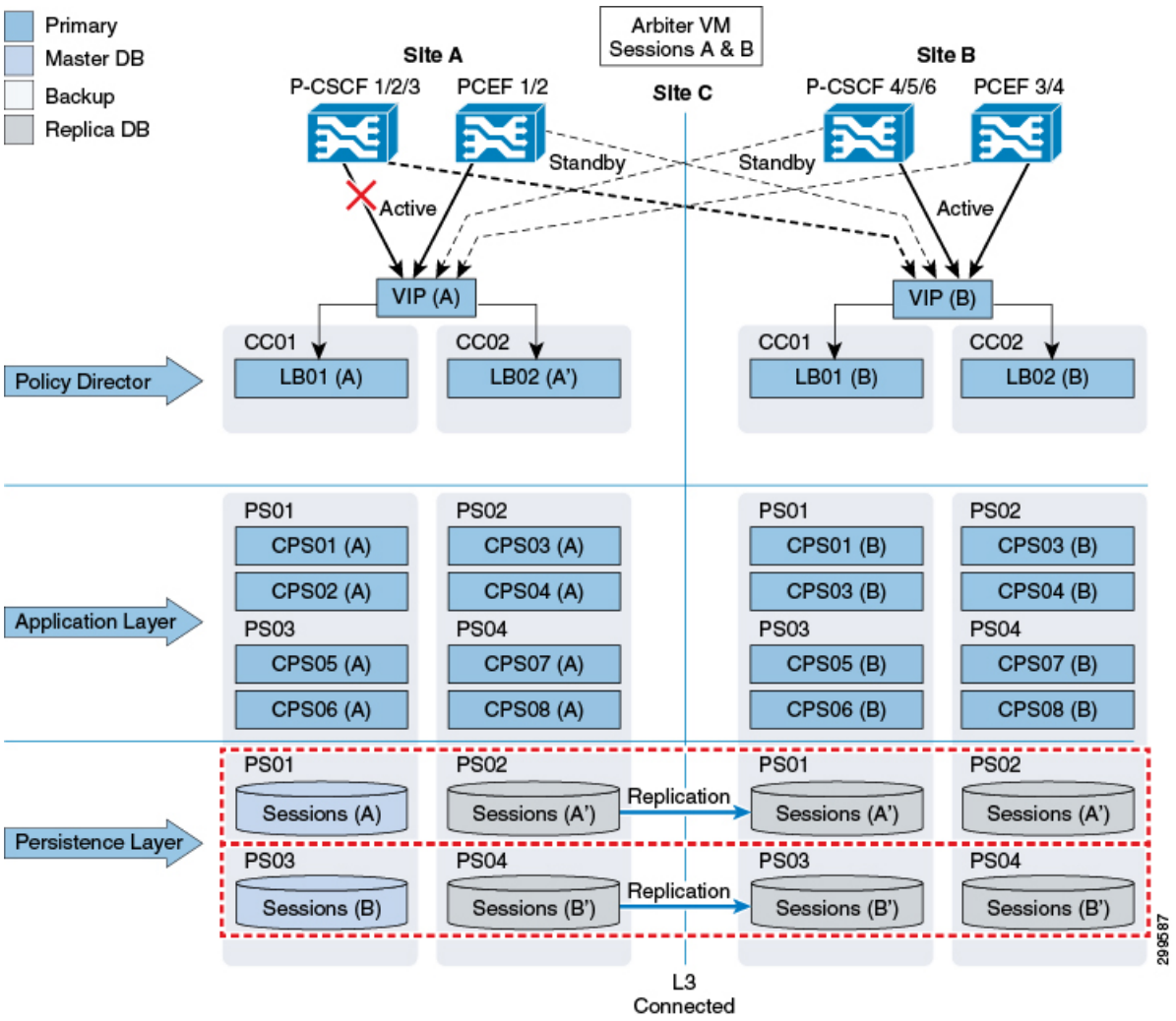
# Rx Link Failure

As shown in the figure below, failure of the Rx link between a P-CSCF and the primary CPS node (Site A) results in the P-CSCF sending traffic to the secondary site (Site B). Failover time dependents on failure detection timers on the P-CSCF.

Rx transactions is processed at Site B. The CPS0x(B) VM handling the transaction at Site B attempts to do the binding by retrieving the Gx session from the Master Sessions(A) database at Site A. Session information is also written across to the Master database at Site A.

The Rx AAA back to the P-CSCF as well as the corresponding Gx RAR to the P-GW is sent from Site B.

*Figure 3: Rx Link Failure*




**Note** This link failure model does not apply for SP Wi-Fi depolyments.

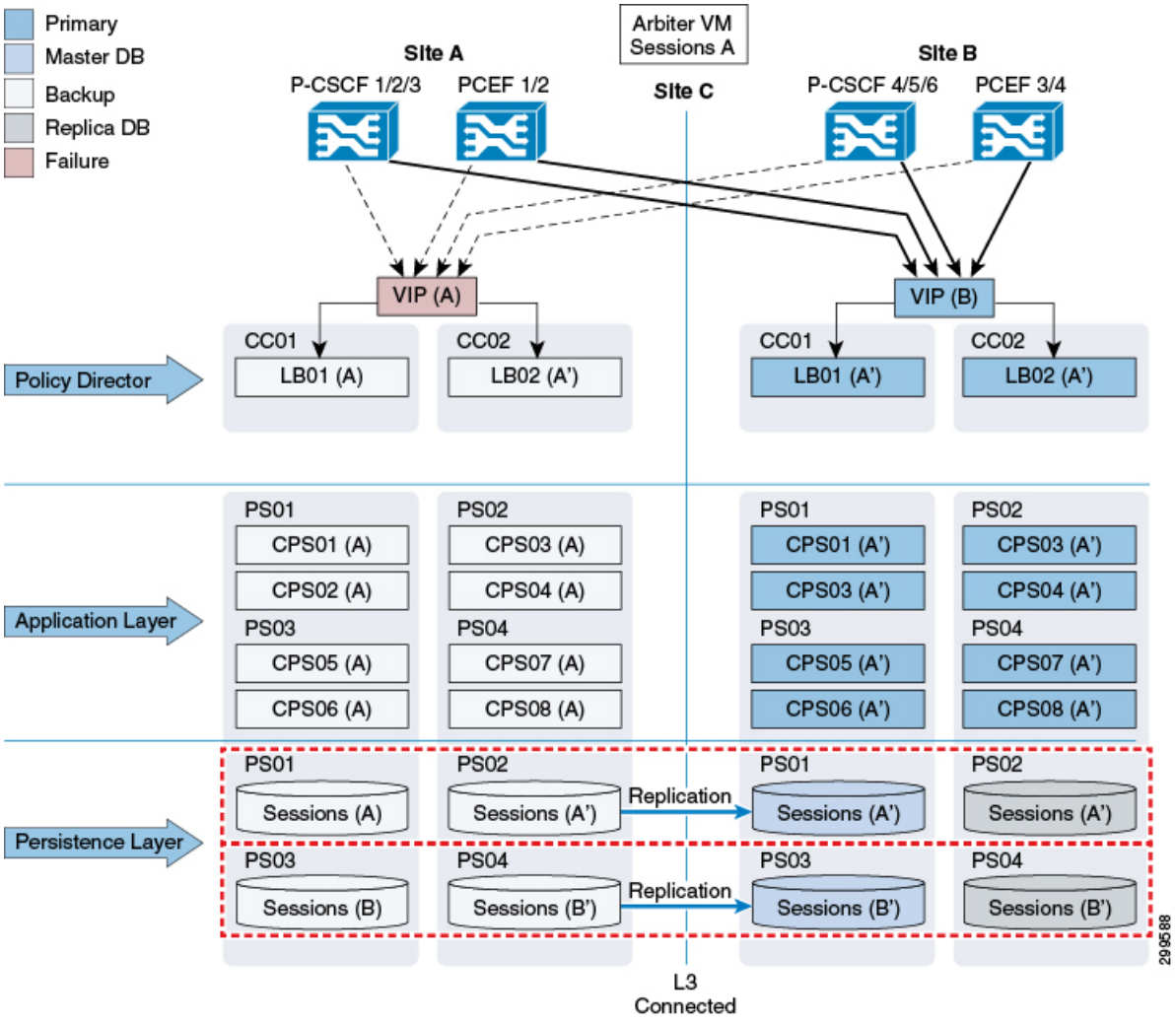
# Load Balancer VIP Outage

As shown in the figure below, all P-GWs and P-CSCFs will direct traffic to the secondary site if both Load Balancer at the primary site is not available (which leads the VIP to be not available). Failover time will be dependent on failure detection timers on the P-GW and P-CSCF.

In order to avoid database writes from Site B to Site A, the system can be configured to monitor VIP availability and, if VIP is not available, lower the priority of the database instances at Site A to force the election of a new Master database at Site B.

By default, VIP availability is monitored every 60 seconds.

*Figure 4: Load Balancer VIP Outage*




### Load Balancer/IP Outage

If the network between load balancers and their other communication end points, such as, PGW, fails, CPS will not detect this failure and will continue to operate as it is.

# Arbiter Failure

As shown in the figure below, the Arbiter is deployed in a non-redundant manner, as failure of the Arbiter alone does not have any impact on the operation of the Replica Set.

However, a subsequent failure, for example a complete outage of Site A while the Arbiter is down, would result in service interruption as the remaining database instances would not constitute a majority that would allow the election of a new Master database.

*Figure 5: Arbiter Failure*