

EDB Postgres EFM3.9で実現する 同期レプリケーション Postgres HAクラスター

EnterpriseDB
高鶴 勝治

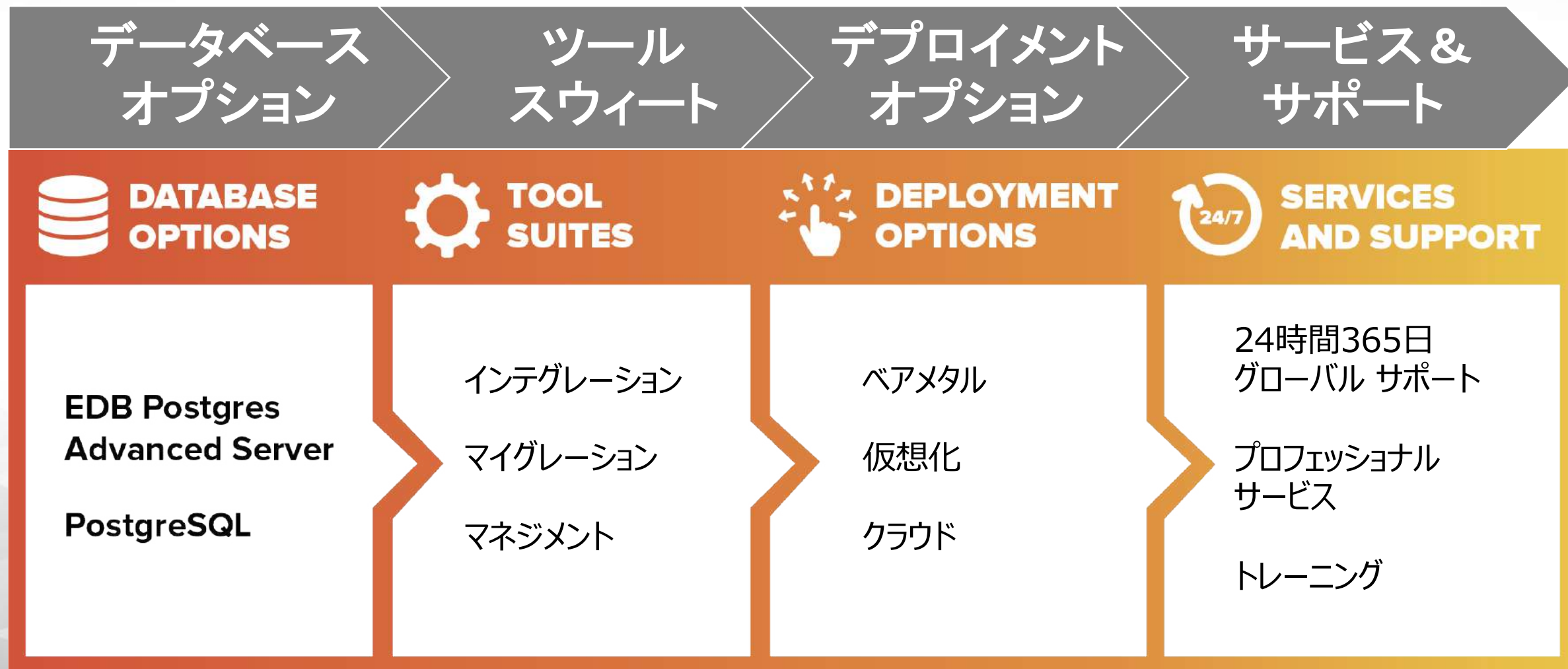


Agenda

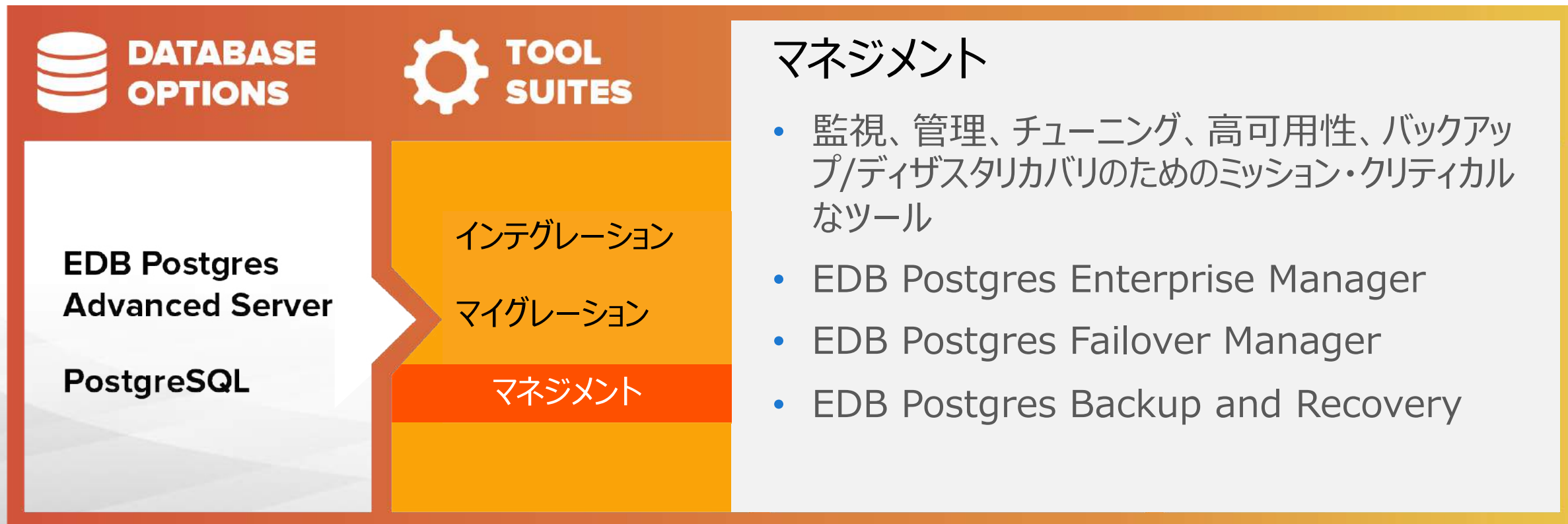
1. EDB Postgres Platform
2. EDB Postgres Failover Manager (EFM)
3. `reconfigure.sync.master` / `reconfigure.num.sync`
4. `use.replay.tiebreaker`

EDB Postgres Platform

EDB POSTGRES PLATFORM



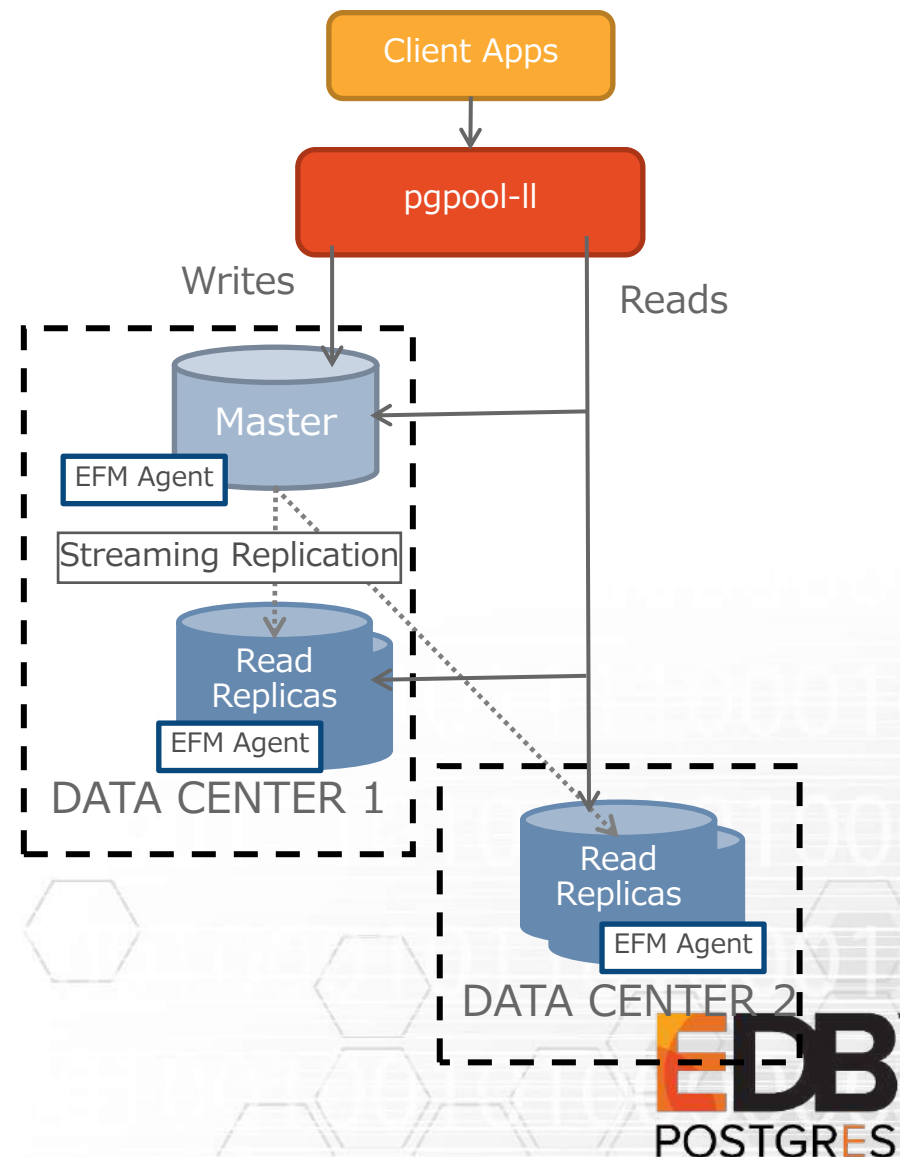
EDB POSTGRES PLATFORM MANAGEMENT SUITES



EDB Postgres Failover Manager (EFM)

EDB Postgres Failover Manager

- 簡単なセットアップ
- エージェント型 + Witnessにより、データベースのSplit-Brainの防止
- 仮想IP(VIP)を提供
- 複数のSlaveサーバの優先度を動的に指定可能
- pgpool-II との連携し、Read型のスケールアウトが可能
- データベースのヘルス・チェック & カスタム・モニタリング機能
- ストリーミング・レプリケーションの同期状態のチェック
- エージェント間のヘルス・チェック
- マスタ障害時自動フェイル・オーバーとクラスタ再構成
- 容易なスイッチ・オーバー/スイッチ・バック機能
- Fencingスクリプトによる外部リソースの制御
- クラスタ構成変更時のメール通知とHook処理の実行
- クラスタ構成変更時のメール送信機能



EFM 3.x の修正履歴

リリース	Ver.	修正内容 (抜粋)
2018/2	3.0	• EPAS10/PostgreSQL10対応 • フェイル・オーバ時の、昇格ノード以外の処理の実行 [script.remote.{pre, post}.promotion] • マスタ・ノードが孤立した場合にマスタDBの停止を制御 [stop.isolated.master] • カスタム・モニタリング機能を追加 [script.custom.monitor, custom.monitor.{interval, timeout, mode}]
2018/5	3.1	• PROMOTOコマンドのパラメタ追加 [-sourcename , -quiet] • 複数VIP対応 [virtualIp.single] • EFMエージェントがマスタDB接続不可の際のマスタDBの停止を制御 [stop.failed.master] • クラスタ構成変更時の efm.nodes の変更回避 [stable.nodes.file] • 通知レベルの制御 [notification.level]
2018/8	3.2	• ロードバランサーやpgpool-II 等との連携 [script.load balancer.{attach, detach}] • ホスト名のロギング
2018/10	3.3	• EPAS11/PostgreSQL11対応 • フェイル・オーバ前のVIP存在チェックの制御 [check.vip.before.promotion] • Syslogロギングのサポート [syslog.{host, port, protocol, facility, enabled}, file.log.enabled] • ロギング・レベルの変更 (TRACE/ DEBUG/INFO/WARN/ERROR) • VIPプロパティでのホスト名の使用
2019/1	3.4	• マスタ・ノード上のEFMエージェント停止時のフェイル・オーバ [master.shutdown.as.failure] • VIPのNICがbind.address のNICと異なる場合、VIP有無チェックをリトライ

EFM 3.x の修正履歴

リリース	Ver.	修正内容
2019/4	3.5	• script.load.balancer.{attach, detach} での新パラメタ [%t] • サーバ再起動後の接続タイムアウト [restart.connection.timeout] • 通知メール送信者の設定 [from.email] • recovery.conf 内の application name 対応 [application.name] • PROMOTO コマンドのパラメタ追加 [-noscripts]
2019/8	3.6	• 緩やかなメモリ・リークに対するFIX • より多くのスタンバイDB情報を用いた確実な、フェイル・オーバ • “cluster-status”レポートでの、Received/Replayed 情報出力 • Debianサポート [db.config.dir]
2019/10	3.7	• EPAS12/PostgreSQL12対応 [db.recovery.conf.dir -> db.recovery.dir]
2020/1	3.8	• restore.command (障害復旧時に、WALファイル取得) • reconfigure.sync.master (同期レプリケーション構成における障害時の挙動を制御)
2020/3	3.9	• サービス名変更 (efm-3.8 -> edb-efm-3.9) • reconfigure.num.sync (同期レプリケーション構成における障害時の挙動を制御) • use.replay.tiebreaker (Promotoされるスタンバイの優先順位に影響) • プロパティ名の変更

同期レプリケーションに関わる
`reconfigure.sync.master`／`reconfigure.num.sync`

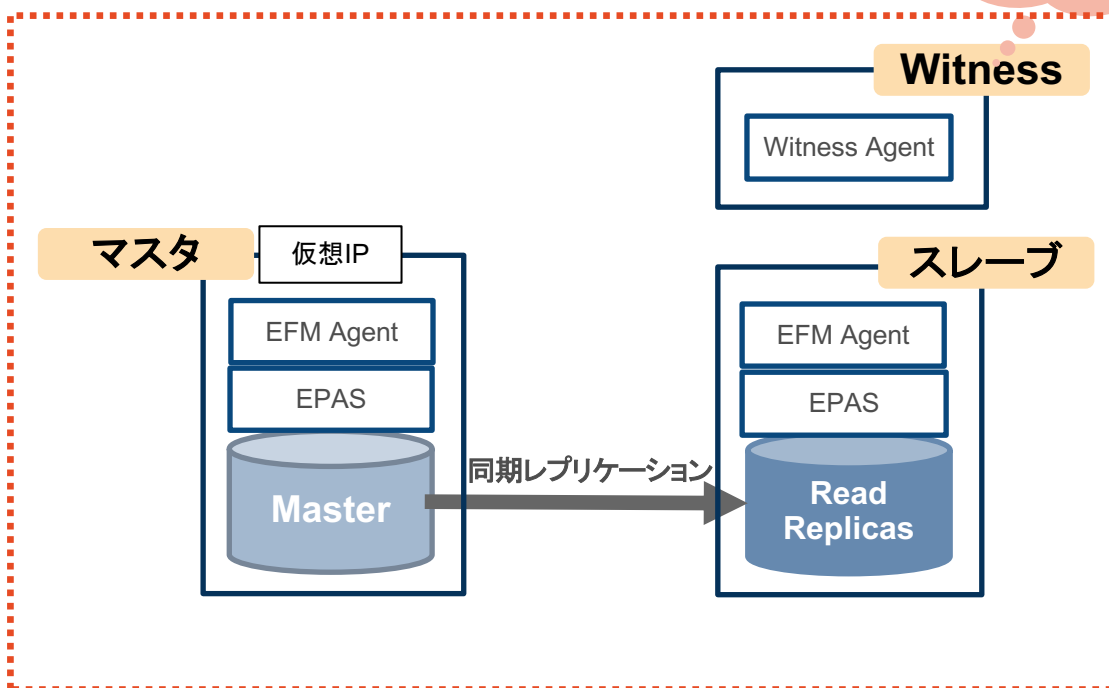
同期レプリケーションで重要な データベース・パラメタ SYNCHRONOUS_STANDBY_NAMES

- *The synchronous_standby_names parameter on the master node specifies the names and count of the synchronous standby servers that will confirm receipt of data, to ensure that the master nodes can accept write transactions.*
- synchronous_standby_names は、同期レプリケーションを行う、「対象となるSlaveサーバ」と「その数」を定義する重要なデータベース・パラメタ。同期レプリケーションが正常にできないと、マスターDBの更新トランザクションが停止してしまう。

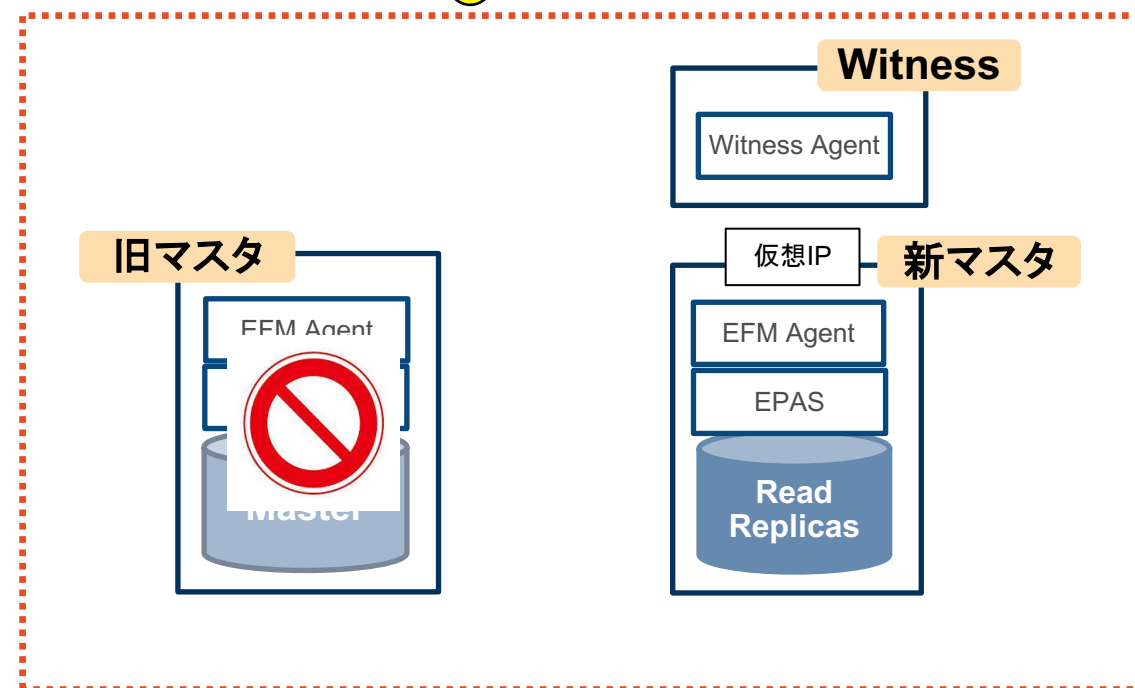
同期レプリケーションの停止(例)

- 2DB(マスタ+同期スレーブ1台)+Witness

2DB構成では
Witness必須



- マスタDB障害後、SlaveがEFMにより昇格 (promote) し、データ・ロスなしで、新マスターとしてサービスを継続
- 但し、スレーブが停止した場合は、マスター更新処理が停止するため、非同期レプリケーションに変更する処理が必要。



[EFM3.8] RECONFIGURE.SYNC.MASTER

- *Set the `reconfigure.sync.master` property to `true` to take the master database out of synchronous replication mode if the number of standby nodes drops below the level required.*
 - “TRUE” に設定すると、データベースで設定した、同期レプリケーションの数を下回る場合、同期レプリケーションを停止
- *Set `reconfigure.sync.master` to `false` to send a notification if the standby count drops, but not interrupt synchronous replication.*
 - “FALSE” に設定すると、データベースで設定した、同期レプリケーションの構成を変えることなく、通知メールのみ送信する

[EFM3.9] RECONFIGURE.NUM.SYNC

- *If set to true, Failover Manager will reduce the number of standbys needed in the master's # synchronous_standby_names property and reload the master configuration.*
 - "TRUE" に設定すると、データベースで設定した、同期レプリケーションの数を下回る場合、同期レプリケーションを停止
- *Failover Manager will not reduce the number below 1, taking the master # out of synchronous replication, unless the reconfigure.sync.master # property is also set to true.*
reconfigure.num.sync=false
 - "FALSE" に設定すると、データベースで設定した、同期レプリケーションの構成を変えることなく、通知メールのみ送信する

パラメタ設定時の挙動 パターン①

両パラメタ: false (EFM3.7までの挙動と同じ)
⇒ 2台のスレーブ停止でトランザクション継続不可

	#1	#2	#3	reconfigure .num.sync	reconfigure .sync.master	synchronous_standby_names		Writable on MasterDB	efm39 cluster-status efm-aps12	select client_addr, application_name, state, sync_priority, sync_state from pg_stat_replication order by client_addr;
						value	comment			
Initial	○ (0)	○ (1)	○ (2)	FALSE	FALSE	1(efm01,efm02,efm03)	Initial Value@#node1	Yes	Agent Type Address Agent DB VIP Master 192.168.0.161 UP UP 192.168.100.198* Standby 192.168.0.162 UP UP 192.168.100.198 Standby 192.168.0.163 UP UP 192.168.100.198 Standby priority host list: 192.168.0.162 192.168.0.163	client_addr application_name state sync_priority sync_state -----+-----+-----+-----+----- 192.168.0.162 efm02 streaming 2 sync 192.168.0.163 efm03 streaming 3 potential
#1Down	✕ 0	○ (0)	○ (1)			1(efm01,efm02,efm03)	Value@node2	Yes	Agent Type Address Agent DB VIP Idle 192.168.0.161 UP UNKNOWN 192.168.100.198 Master 192.168.0.162 UP UP 192.168.100.198* Standby 192.168.0.163 UP UP 192.168.100.198 Standby priority host list: 192.168.0.163	client_addr application_name state sync_priority sync_state -----+-----+-----+-----+----- 192.168.0.163 efm03 streaming 3 sync
#2Down	○ 0	✕ 0	○ 0			1(efm01,efm02,efm03)	Value@node1	Yes	Agent Type Address Agent DB VIP Master 192.168.0.161 UP UP 192.168.100.198* Idle 192.168.0.162 UP UNKNOWN 192.168.100.198 Standby 192.168.0.163 UP UP 192.168.100.198 Standby priority host list: 192.168.0.163	client_addr application_name state sync_priority sync_state -----+-----+-----+-----+----- 192.168.0.163 efm03 streaming 3 sync
#3Down	○ 0	○ 0	✕ 0			1(efm01,efm02,efm03)	Value@node1	Yes	Agent Type Address Agent DB VIP Master 192.168.0.161 UP UP 192.168.100.198* Standby 192.168.0.162 UP UP 192.168.100.198 Idle 192.168.0.163 UP UNKNOWN 192.168.100.198 Standby priority host list: 192.168.0.162	client_addr application_name state sync_priority sync_state -----+-----+-----+-----+----- 192.168.0.162 efm02 streaming 2 sync
#3->#2Down	○ 0	✕ 0	✕ 0			1(efm01,efm02,efm03)	Value@node1	No (Waiting)	Agent Type Address Agent DB VIP Master 192.168.0.161 UP UP 192.168.100.198* Idle 192.168.0.162 UP UNKNOWN 192.168.100.198 Idle 192.168.0.163 UP UNKNOWN 192.168.100.198 Standby priority host list: (List is empty.)	client_addr application_name state sync_priority sync_state -----+-----+-----+-----+----- (0 行)

パラメタ設定時の挙動 パターン②

reconfig.sync.master: TRUE

⇒ 2台のスレーブ停止後も、トランザクション継続

	#1	#2	#3	reconfigure .num.sync	reconfigure .sync.master	synchronous_standby_names		Writable on MasterDB	efm39 cluster-status efm-aps12	select client_addr, application_name, state, sync_priority, sync_state from pg_stat_replication order by client_addr;
						value	comment			
Initial	○ (1)	○ (2)	○ (3)	FALSE	TRUE	1(efm01,efm02,efm03)	初期設定 Value@#node1	Yes	Agent Type Address Agent DB VIP Master 192.168.0.161 UP UP 192.168.100.198* Standby 192.168.0.162 UP UP 192.168.100.198 Standby 192.168.0.163 UP UP 192.168.100.198 Standby priority host list: 192.168.0.162 192.168.0.163	client_addr application_name state sync_priority sync_state -----+-----+-----+-----+----- 192.168.0.162 efm02 streaming 2 sync 192.168.0.163 efm03 streaming 3 potential
#1Down	✕ 0	○ 0	○ 0			1(efm01,efm02,efm03)	Value@node2	Yes	Agent Type Address Agent DB VIP Idle 192.168.0.161 UP UNKNOWN 192.168.100.198 Master 192.168.0.162 UP UP 192.168.100.198* Standby 192.168.0.163 UP UP 192.168.100.198 Standby priority host list: 192.168.0.163	client_addr application_name state sync_priority sync_state -----+-----+-----+-----+----- 192.168.0.163 efm03 streaming 3 sync
#2Down	○ 0	✕ 0	○ 0			1(efm01,efm02,efm03)	Value@node1	Yes	Agent Type Address Agent DB VIP Master 192.168.0.161 UP UP 192.168.100.198* Standby 192.168.0.162 UP UP 192.168.100.198 Standby 192.168.0.163 UP UP 192.168.100.198 Standby priority host list: 192.168.0.162 192.168.0.163	client_addr application_name state sync_priority sync_state -----+-----+-----+-----+----- 192.168.0.163 efm03 streaming 3 sync
#3Down	○ 0	○ 0	✕ 0			1(efm01,efm02,efm03)	Value@node1	Yes	Agent Type Address Agent DB VIP Master 192.168.0.161 UP UP 192.168.100.198* Standby 192.168.0.162 UP UP 192.168.100.198 Idle 192.168.0.163 UP UNKNOWN 192.168.100.198 Standby priority host list: 192.168.0.162	client_addr application_name state sync_priority sync_state -----+-----+-----+-----+----- 192.168.0.162 efm02 streaming 2 sync
#3->#2Down	○ 0	✕ 0	✕ 0			""(Empty Character String)	Value@node1	Yes	Agent Type Address Agent DB VIP Master 192.168.0.161 UP UP 192.168.100.198* Idle 192.168.0.162 UP UNKNOWN 192.168.100.198 Idle 192.168.0.163 UP UNKNOWN 192.168.100.198 Standby priority host list: (List is empty.)	client_addr application_name state sync_priority sync_state -----+-----+-----+-----+----- (0 行)

パラメタ設定時の挙動 パターン③

reconfig.num.master : TRUE

⇒ 2台のスレーブ停止でトランザクション継続不可

	#1	#2	#3	reconfigure .num.sync	reconfigure .sync.master	synchronous_standby_names		Writable on MasterDB	efm39 cluster-status efm-aps12	select client_addr , application_name , state, sync_priority , sync_state from pg_stat_replication order by client_addr;
						value	comment			
Initial	○ (1)	○ (2)	○ (3)	TRUE	FALSE	1(efm01,efm02,efm03)	Initial Value@#node1	Yes	Agent Type Address Agent DB VIP Master 192.168.0.161 UP UP 192.168.100.198* Standby 192.168.0.162 UP UP 192.168.100.198 Standby 192.168.0.163 UP UP 192.168.100.198 Standby priority host list: 192.168.0.162 192.168.0.163	client_addr application_name state sync_priority sync_state -----+-----+-----+-----+----- 192.168.0.162 efm02 streaming 2 sync 192.168.0.163 efm03 streaming 3 potential
#1Down	✕ 0	○ 0	○ 0			1(efm01,efm02,efm03)	Value@node2	Yes	Agent Type Address Agent DB VIP Idle 192.168.0.161 UP UNKNOWN 192.168.100.198 Master 192.168.0.162 UP UP 192.168.100.198* Standby 192.168.0.163 UP UP 192.168.100.198 Standby priority host list: 192.168.0.163	client_addr application_name state sync_priority sync_state -----+-----+-----+-----+----- 192.168.0.163 efm03 streaming 3 sync
#2Down	○ 0	✕ 0	○ 0			1(efm01,efm02,efm03)	Value@node1	Yes	Agent Type Address Agent DB VIP Master 192.168.0.161 UP UP 192.168.100.198* Idle 192.168.0.162 UP UNKNOWN 192.168.100.198 Standby 192.168.0.163 UP UP 192.168.100.198 Standby priority host list: 192.168.0.163	client_addr application_name state sync_priority sync_state -----+-----+-----+-----+----- 192.168.0.163 efm03 streaming 3 sync
#3Down	○ 0	○ 0	✕ 0			1(efm01,efm02,efm03)	Value@node1	Yes	Agent Type Address Agent DB VIP Master 192.168.0.161 UP UP 192.168.100.198* Standby 192.168.0.162 UP UP 192.168.100.198 Idle 192.168.0.163 UP UNKNOWN 192.168.100.198 Standby priority host list: 192.168.0.162	client_addr application_name state sync_priority sync_state -----+-----+-----+-----+----- 192.168.0.162 efm02 streaming 2 sync
#3->#2Down	○ 0	✕ 0	✕ 0			1(efm01,efm02,efm03)	Value@node1	No (Waiting)	Agent Type Address Agent DB VIP Master 192.168.0.161 UP UP 192.168.100.198* Idle 192.168.0.162 UP UNKNOWN 192.168.100.198 Idle 192.168.0.163 UP UNKNOWN 192.168.100.198 Standby priority host list: (List is empty.)	client_addr application_name state sync_priority sync_state -----+-----+-----+-----+----- (0 行)

パラメタ設定時の挙動 パターン④

両パラメタ: TRUE

⇒ 2台のスレーブ停止後も、トランザクション継続

	#1	#2	#3	reconfigure .num.sync	reconfigure .sync.master	synchronous_standby_names		Writable on MasterDB	efm39 cluster-status efm-epast2	select client_addr, application_name, state, sync_priority, sync_state from pg_stat_replication order by client_addr;
						value	comment			
Initial	○ (1)	○ (2)	○ (3)	TRUE	TRUE	1(efm01,efm02,efm03)	Initial Value@#node1	Yes	Agent Type Address Agent DB VIP Master 192.168.0.161 UP UP 192.168.100.198* Standby 192.168.0.162 UP UP 192.168.100.198 Standby 192.168.0.163 UP UP 192.168.100.198 Standby priority host list: 192.168.0.162 192.168.0.163	client_addr application_name state sync_priority sync_state -----+-----+-----+-----+----- 192.168.0.162 efm02 streaming 2 sync 192.168.0.163 efm03 streaming 3 potential
#1Down	✕ 0	○ 0	○ 0			1(efm01,efm02,efm03)	Value@node2	Yes	Agent Type Address Agent DB VIP Idle 192.168.0.161 UP UNKNOWN 192.168.100.198 Master 192.168.0.162 UP UP 192.168.100.198* Standby 192.168.0.163 UP UP 192.168.100.198 Standby priority host list: 192.168.0.163	client_addr application_name state sync_priority sync_state -----+-----+-----+-----+----- 192.168.0.163 efm03 streaming 3 sync
#2Down	○ 0	✕ 0	○ 0			1(efm01,efm02,efm03)	Value@node1	Yes	Agent Type Address Agent DB VIP Master 192.168.0.161 UP UP 192.168.100.198* Idle 192.168.0.162 UP UNKNOWN 192.168.100.198 Standby 192.168.0.163 UP UP 192.168.100.198 Standby priority host list: 192.168.0.163	client_addr application_name state sync_priority sync_state -----+-----+-----+-----+----- 192.168.0.163 efm03 streaming 3 sync
#3Down	○ 0	○ 0	✕ 0			1(efm01,efm02,efm03)	Value@node1	Yes	Agent Type Address Agent DB VIP Master 192.168.0.161 UP UP 192.168.100.198* Standby 192.168.0.162 UP UP 192.168.100.198 Idle 192.168.0.163 UP UNKNOWN 192.168.100.198 Standby priority host list: 192.168.0.162	client_addr application_name state sync_priority sync_state -----+-----+-----+-----+----- 192.168.0.162 efm02 streaming 2 sync
#3->#2Down	○ 0	✕ 0	✕ 0			""(Empty Character String)	Value@node1	Yes	Agent Type Address Agent DB VIP Master 192.168.0.161 UP UP 192.168.100.198* Idle 192.168.0.162 UP UNKNOWN 192.168.100.198 Idle 192.168.0.163 UP UNKNOWN 192.168.100.198 Standby priority host list: (List is empty.)	client_addr application_name state sync_priority sync_state -----+-----+-----+-----+----- (0 行)

パラメタ設定時の挙動 まとめ#1

- 同期スレーブ数=1台の時は、reconfigure.sync.master の設定が重要。TRUEに設定することで、同期スレーブがいなくなっても、マスタDBのトランクション継続が可能。
(例) synchronous_standby_name = '1(efm01,efm02,efm03)'
- reconfigure.num.sync の設定は影響しない

パラメタ設定時の挙動 パターン①

reconfig.sync.master: TRUE

⇒ 1台のスレーブ停止で、非同期モードに変更

	#1	#2	#3	reconfigure .num.sync	reconfigure .sync.master	synchronous_standby_names		マスター書込み可 否	Membership coordinator: 192.168.0.161	select client_addr, application_name, state, sync_priority, sync_state from pg_stat_replication order by client_addr;
						value	comment			
Initial	○ (1)	○ (2)	○ (3)	FALSE	TRUE	2(efm01,efm02,efm03)	Initial Value@#node1	Yes	Agent Type Address Agent DB VIP ----- Master 192.168.0.161 UP UP 192.168.100.198* Standby 192.168.0.162 UP UP 192.168.100.198 Standby 192.168.0.163 UP UP 192.168.100.198 Standby priority host list: 192.168.0.162 192.168.0.163	client_addr application_name state sync_priority sync_state ----- 192.168.0.162 efm02 streaming 2 sync 192.168.0.163 efm03 streaming 3 sync
#3Down	○ 0	○ 0	✕ 0			""(Empty Character String)	Value@node1	Yes	Agent Type Address Agent DB VIP ----- Master 192.168.0.161 UP UP 192.168.100.198* Standby 192.168.0.162 UP UP 192.168.100.198 Idle 192.168.0.163 UP UNKNOWN 192.168.100.198 Standby priority host list: 192.168.0.162	client_addr application_name state sync_priority sync_state ----- 192.168.0.162 efm02 streaming 0 async
#3->#2Down	○ 0	✕ 0	✕ 0			""(Empty Character String)	Value@node1	Yes	Agent Type Address Agent DB VIP ----- Master 192.168.0.161 UP UP 192.168.100.198* Idle 192.168.0.162 UP UNKNOWN 192.168.100.198 Idle 192.168.0.163 UP UNKNOWN 192.168.100.198 Standby priority host list: (List is empty.)	client_addr application_name state sync_priority sync_state ----- (0 行)

パラメタ設定時の挙動 パターン②

	#1	#2	#3	reconfigure . num. sync	reconfigure . sync. master	synchronous_standby_names		マスター書き込み可 否	Membership coordinator: 192.168.0.161	select client_addr, application_name, state, sync_priority, sync_state from pg_stat_replication order by client_addr;
						value	comment			
Initial	○ (1)	○ (2)	○ (3)	TRUE	FALSE	2(efm01,efm02,efm03)	Initial Value@#node1	Yes	Agent Type Address Agent DB VIP Master 192.168.0.161 UP UP 192.168.100.198* Standby 192.168.0.162 UP UP 192.168.100.198 Standby 192.168.0.163 UP UP 192.168.100.198 Standby priority host list: 192.168.0.162 192.168.0.163	client_addr application_name state sync_priority sync_state 192.168.0.162 efm02 streaming 2 sync 192.168.0.163 efm03 streaming 3 sync
#3Down	○ 0	○ 0	✕ 0			1(efm01,efm02,efm03)	Value@node1	Yes	Agent Type Address Agent DB VIP Master 192.168.0.161 UP UP 192.168.100.198* Standby 192.168.0.162 UP UP 192.168.100.198 Idle 192.168.0.163 UP UNKNOWN 192.168.100.198 Standby priority host list: 192.168.0.162	client_addr application_name state sync_priority sync_state 192.168.0.162 efm02 streaming 2 sync
#3->#2Down	○ 0	✕ 0	✕ 0			1(efm01,efm02,efm03)	Value@node1	No (Waiting)	Agent Type Address Agent DB VIP Master 192.168.0.161 UP UP 192.168.100.198* Idle 192.168.0.162 UP UNKNOWN 192.168.100.198 Idle 192.168.0.163 UP UNKNOWN 192.168.100.198 Standby priority host list: (List is empty.)	client_addr application_name state sync_priority sync_state (0 行)

パラメタ設定時の挙動 パターン③

	#1	#2	#3	reconfigure . num. sync	reconfigure . sync. master	synchronous_standby_names		マスター書き込み可 否	Membership coordinator: 192.168.0.161	select client_addr, application_name, state, sync_priority, sync_state from pg_stat_replication order by client_addr;
						value	comment			
Initial	○ (1)	○ (2)	○ (3)	TRUE	TRUE	2(efm01,efm02,efm03)	Initial Value@#node1	Yes	Agent Type Address Agent DB VIP Master 192.168.0.161 UP UP 192.168.100.198* Standby 192.168.0.162 UP UP 192.168.100.198 Standby 192.168.0.163 UP UP 192.168.100.198 Standby priority host list: 192.168.0.162 192.168.0.163	client_addr application_name state sync_priority sync_state 192.168.0.162 efm02 streaming 2 sync 192.168.0.163 efm03 streaming 3 sync
#3Down	○ 0	○ 0	✕ 0			1(efm01,efm02,efm03)	Value@node1	Yes	Agent Type Address Agent DB VIP Master 192.168.0.161 UP UP 192.168.100.198* Standby 192.168.0.162 UP UP 192.168.100.198 Idle 192.168.0.163 UP UNKNOWN 192.168.100.198 Standby priority host list: 192.168.0.162	client_addr application_name state sync_priority sync_state 192.168.0.162 efm02 streaming 2 sync
#3->#2Down	○ 0	✕ 0	✕ 0			""(Empty Character String)	Value@node1	Yes	Agent Type Address Agent DB VIP Master 192.168.0.161 UP UP 192.168.100.198* Idle 192.168.0.162 UP UNKNOWN 192.168.100.198 Idle 192.168.0.163 UP UNKNOWN 192.168.100.198 Standby priority host list: (List is empty.)	client_addr application_name state sync_priority sync_state (0 行)

パラメタ設定時の挙動 まとめ#2

- 同期スレーブ数=2台以上の時は、reconfigure.num.master の設定が重要。TRUEに設定するkとおで、同期スレーブが減少しても、マスタDBのトランクション継続が可能。
(例) Synchronous_standby_name = '2(efm01,efm02,efm03)'
Synchronous_standby_name = 'ANY 2(efm01,efm02,efm03)'
- 同期スレーブ数=2台以上でも、reconfigure.sync.master だけの設定では、縮退した同期レプリケーション構成にはならない
- 同期スレーブ数=1に至った時は、reconfigure.sync.master の設定が重要

まとめ

同期レプリケーションを行う場合は、以下のパラメタ設定を推奨

- `reconfigure.sync.master = TRUE`
- `reconfigure.num.master = TRUE`

留意点① `reconfigure.num.master=TRUE` AND/OR `reconfigure.sync.master=TRUE` の時、EFM は、マスターDBの “`synchronous_standby_names`” のみ変更する。

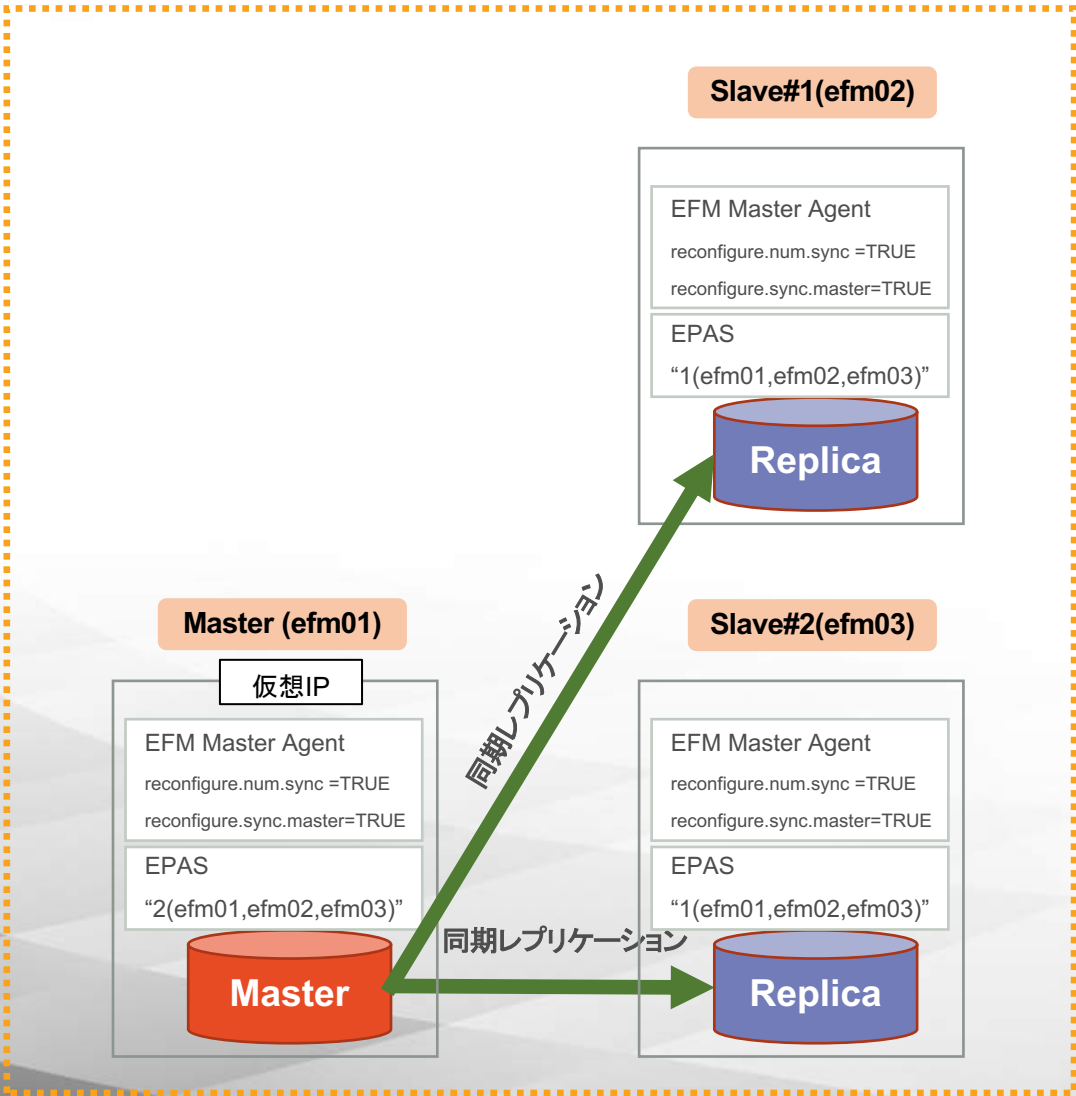
留意点② EFM3.9 では、`synchronous_standby_names` の調整が行われるのは、スレーブDB が、ダウンした時のみ。マスターDB停止後の、Promoteでは調整されない。

留意点③ スレーブが、再び、EFMクラスタに参加しても、`synchronous_standby_names` の調整は行われない。

構成例 # 1: 2 DB(マスタ+同期スレーブ1台)+WITNESS



構成例#2: 3DB(マスタ+同期スレーブ2台)



Slave-DB
障害



Promoto するスレーブ決めに関わる use.replay.tiebreaker

[EFM3.9] USE.REPLAY.TIEBREAKER

- *Use replay LSN value for tiebreaker when choosing a standby to promote # before using failover priority. Set this property to true to consider replay location as more important than failover priority (as seen in cluster-status command) when choosing the “most ahead” standby to promote. use.replay.tiebreaker=true*

→ PromotoするスタンバイDBの決め方に影響

“TRUE”に設定すると、“WAL Replayed LSN”が最新のスタンバイDBを採用する

“FALSE”に設定すると、EFM上の“Standby priority host list”のTop PriorityのスタンバイDBを採用する

このパラメタの利点

- EFMは、当初、非同期レプリケーションを基本に開発されたため、EFMは、データベース・レイヤにおいて、スタンバイDBのプライオリティは、フラットと考えている
- 実際は、同期レプリケーションを行う場合、データベース・レイヤでは、`synchronous_standby_names`の値で、プライオリティ付け(同期 or 非同期)されているため、フェイル・オーバする場合は、データベース上で、プライオリティの高い(=最新の状態の)スタンバイDBを、昇格(Promote)しなければならない。
- そのため、従来のバージョンでは、EFMのHook機能を使って、EFM上のプライオリティを、データベース上のものに合わせる必要があった。
- 新パラメタにより、EFMは、最新のスタンバイにDBを昇格することができるので、Quorum Commit の利用が可能。

⇒ 非同期でも、up-to-dateノードにPromote

POSTGRES

パラメタ設定時の挙動 パターン②

パラメタ: TRUE

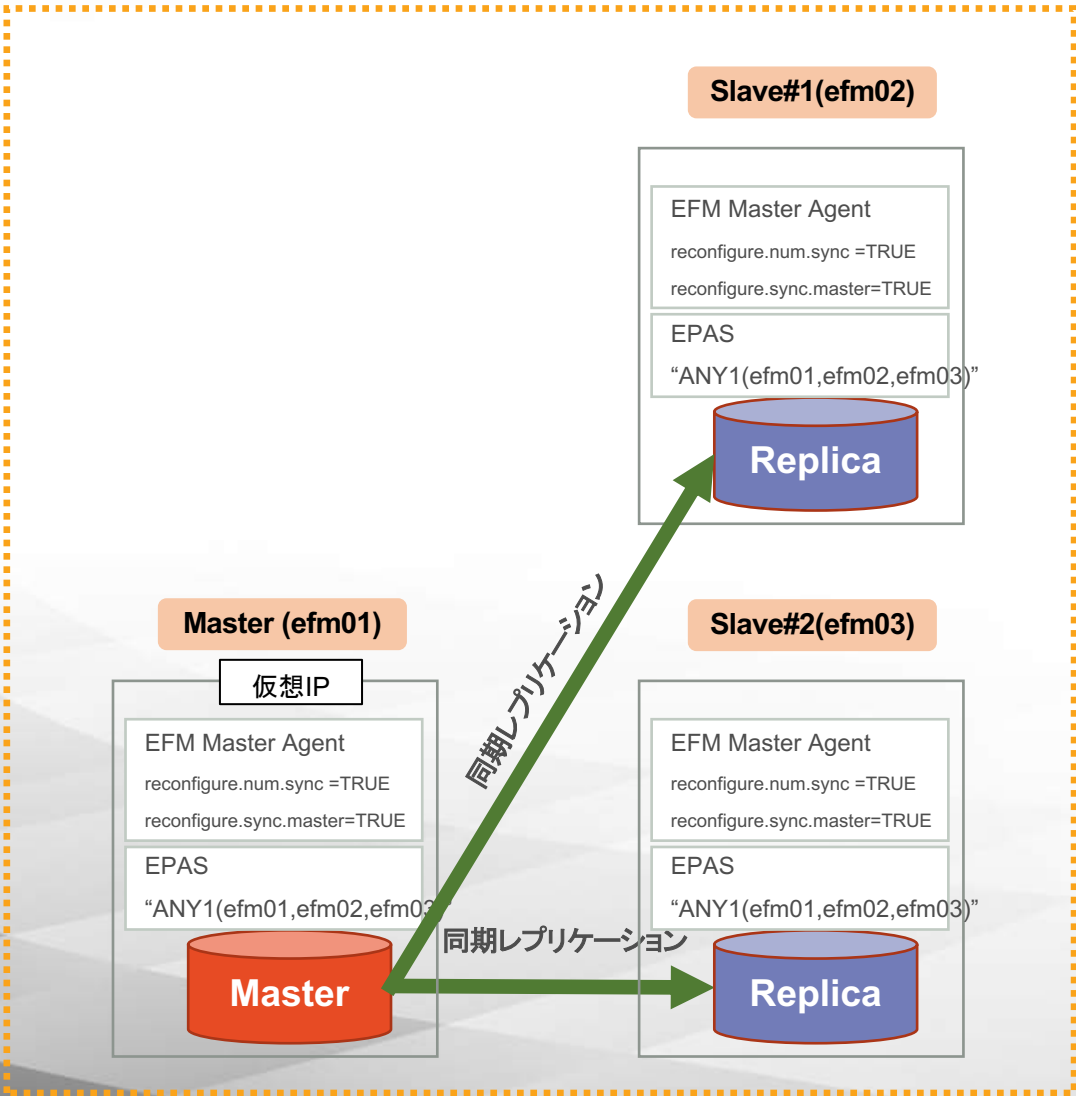
⇒ 同期でも、up-to-dateノードにPromote

	synchronous_standby_names	use.replay.tiebreaker	just before Failover	after Failover	Promoted Node
			client_addr application_name state sync_priority sync_state sent_lsn write_lsn flush_lsn replay_lsn write_lag flush_lag replay_lag	client_addr application_name state sync_priority sync_state sent_lsn write_lsn flush_lsn replay_lsn write_lag flush_lag replay_lag	
Sync	"2(efm01,efm02,efm03)"	FALSE	<pre> 192.168.0.162 efm02 streaming 2 sync 4/7448B940 4/7448B940 4/7448B940 4/7448B940 00:00:00.000929 00:00:00.015735 00:00:00.015744 192.168.0.163 efm03 streaming 3 sync 4/7448B940 4/7448B940 4/7448B940 4/7448B940 00:00:00.000873 00:00:00.003697 00:00:00.003705 (2 rows) Promote Status: DB Type Address WAL Received LSN WAL Replayed LSN Info ----- Master 192.168.0.161 4/7448B940 Standby 192.168.0.162 4/7448B940 4/7448B940 Standby 192.168.0.163 4/7448B940 4/7448B940 </pre>	<pre> 192.168.0.163 efm03 streaming 3 sync 4/750089F0 4/750089F0 4/750089F0 4/750089F0 Cluster Status: efm-epas12 Agent Type Address Agent DB VIP ----- Idle 192.168.0.161 UP UNKNOWN 192.168.100.198 Master 192.168.0.162 UP UP 192.168.100.198* Standby 192.168.0.163 UP UP 192.168.100.198 </pre>	Node#2
		TRUE	<pre> 192.168.0.162 efm02 streaming 2 sync 4/7A188710 4/7A188710 4/7A188710 4/7A00CBD0 00:00:00.002163 00:00:00.013425 00:00:10.215976 192.168.0.163 efm03 streaming 3 sync 4/7A188710 4/7A188710 4/7A188710 4/7A188710 00:00:00.001037 00:00:00.004101 00:00:00.006736 Standby priority host list: 192.168.0.162 192.168.0.163 Promote Status: DB Type Address WAL Received LSN WAL Replayed LSN Info ----- Master 192.168.0.161 4/7A188710 Standby 192.168.0.162 4/7A188710 4/7A00CBD0 Standby 192.168.0.163 4/7A188710 4/7A188710 </pre>	<pre> 192.168.0.162 efm02 startup 2 potential 4/7B0000A0 4/7B0000A0 4/7B0000A0 00:00:00.307959 00:00:00.310952 00:00:00.310952 Cluster Status: efm-epas12 Agent Type Address Agent DB VIP ----- Idle 192.168.0.161 UP UNKNOWN 192.168.100.198 Standby 192.168.0.162 UP UP 192.168.100.198 Master 192.168.0.163 UP UP 192.168.100.198* </pre>	Node#3

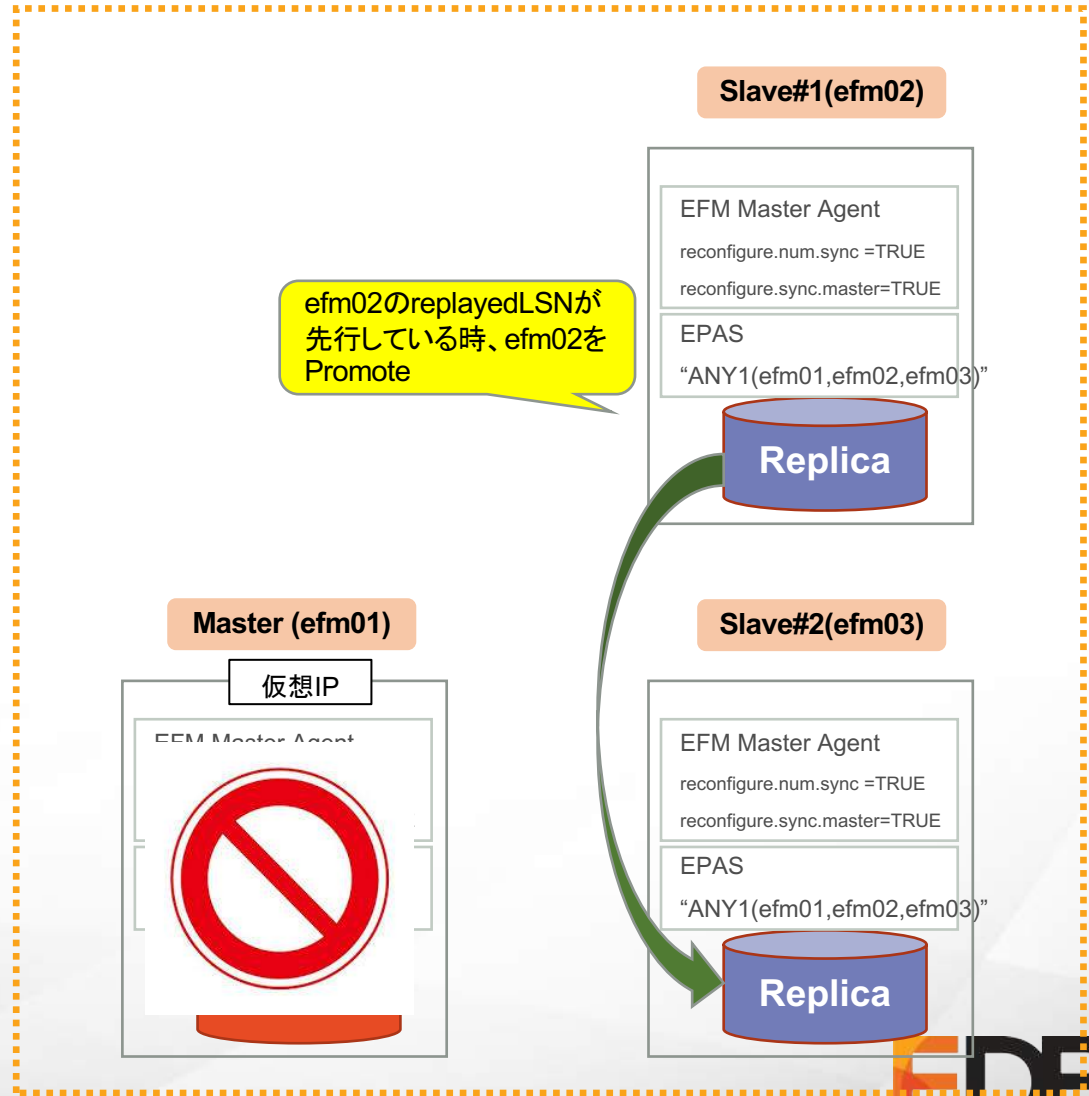
まとめ

`use.replay.tiebreaker = TRUE` の場合、EFM 上のスタンバイ・プライオリティではなく、最もWALが適用されているDBにPROMOTEを行う

構成例#3: 3DB(マスタ+同期スレーブ2台(クォーラム))



Master-DB
障害



THANK YOU

merci
grazie
spasiba
kam ouen
gratizias
tak
manana
mahalo
hvala
cheers
toda
gracias
grassie
thank you
danki
kitos
welalin
mahalo
danki
thanks
takk
gracias
domo arrigato
danke
gratitude
merci
na gode
mesi
modupe
talofa
miigwetch
kitos
takk
dziekuje