

pgpool-II で PostgreSQL のクラスタを 楽々運用しよう

OSC 2014 .Enterprise @ Tokyo
2014/12/12

SRA OSS, Inc. 日本支社
マーケティング部 OSS技術グループ

長田 悠吾

自己紹介

- 長田 悠吾 (ナガタ ユウゴ)
 - SRA OSS, Inc. 日本支社
 - マーケティング部 OSS技術グループ
 - pgpool-II 開発者
 - PostgreSQL 関連の技術調査
 - OSS の技術サポート
 - PostgreSQL の開発にも参加

SRA OSS, Inc. 日本支社

- 1999年よりPostgreSQLサポートを中心にOSSビジネスを開始
- PostgreSQL、Hinemos、Zabbix などのOSSサポート
- PowerGresファミリーの開発、販売
- トレーニング、導入、設計コンサルティングサービス



本日のお話

- PostgreSQLのクラスタリング機能
 - ストリーミングレプリケーション
- でも、それだけでは出来ないことも・・・
 - 負荷分散で検索性能をスケールアウトさせたい
 - サーバが落ちた場合には、人手を入れずに自動フェイルオーバーさせたい
- pgpool-II にお任せ！
 - 機能と使い方
 - 新バージョン 3.4 の新機能

PostgreSQLでクラスタリング

PostgreSQLとは

- 代表的なオープンソースのRDBMSの1つ
 - カリフォルニア大学で開発された研究用RDBMSのIngres(1970)を先祖に持つ
- オーナー企業を持たず、コミュニティによる開発が続けられている
 - 年1回のメジャーバージョンアップ
 - 9.4RC1が最近リリース
- PostgreSQLライセンスで配布
 - BSDタイプの緩いライセンス



クラスタリングしたい・・・

- データベースクラスタリングの目的は？

- 高可用性の確保

- サービスを停止させたくない
- 1つのデータベースが故障しても、別のデータベースが肩代わりする

- 参照負荷分散

- 大量のアクセスをさばきたい
- 負荷を分散して検索性能を向上

- 並列処理

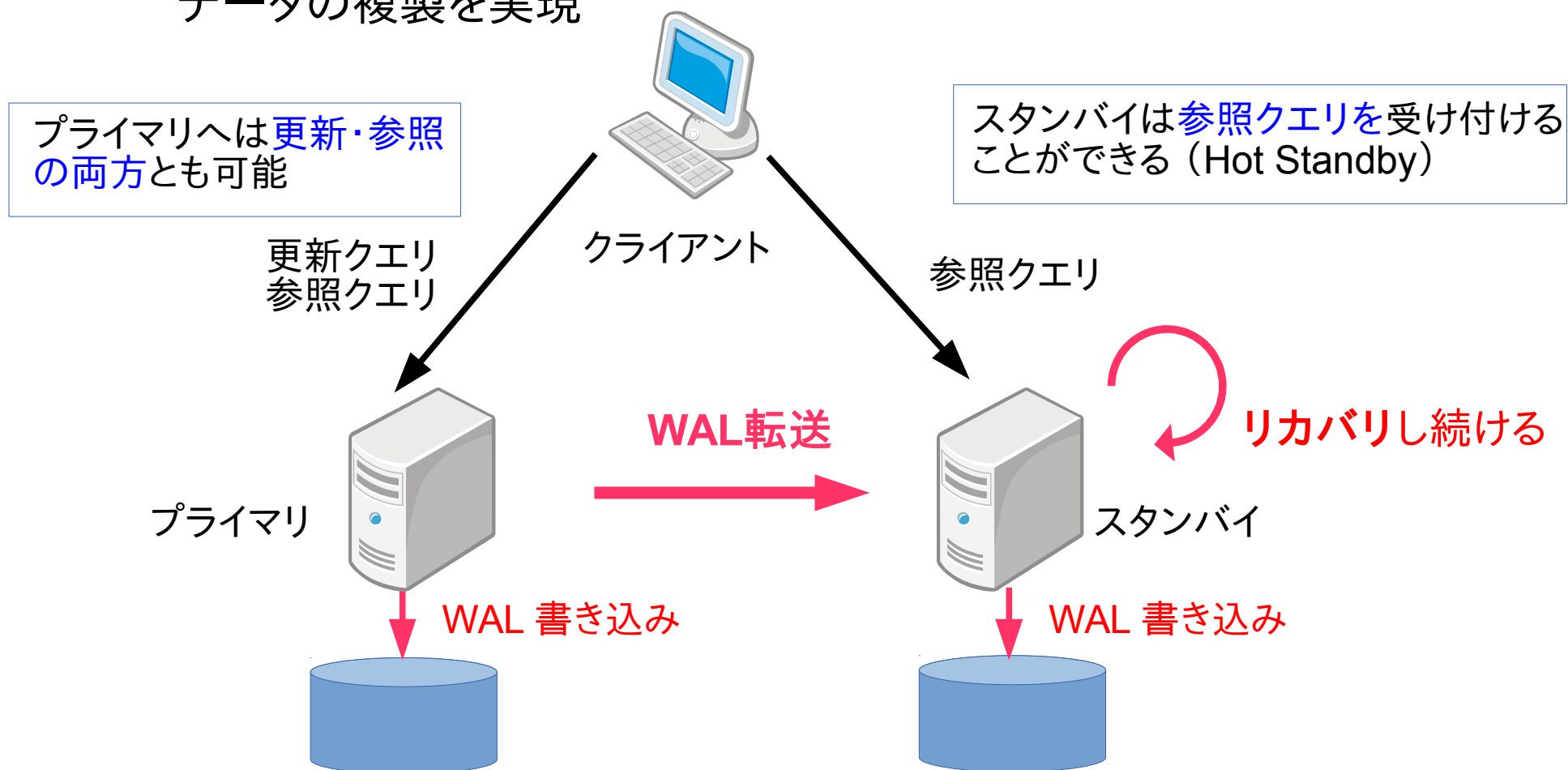
- 大量のデータを解析したい
- 複数のサーバで並列的に処理

PostgreSQLのクラスタ技術

- HAクラスタ
 - Pacemaker+DRBD、共有ストレージなどを利用
 - 待機側はサービス停止
- ストリーミングレプリケーション
 - PostgreSQL自体が持つ、非同期レプリケーション機能
 - プライマリ(更新可能) + 複数のスタンバイDB(検索のみ)
 - 簡単、確実、速い
- pgpool-II
 - クライアントとPostgreSQLの間に入ってレプリケーション機能を提供
 - コネクションプーリング、負荷分散、自動フェイルオーバーなど他の機能もある
- Postgres-XC
 - PostgreSQLを改造したクラスタシステム
 - 書き込み性能の負荷分散

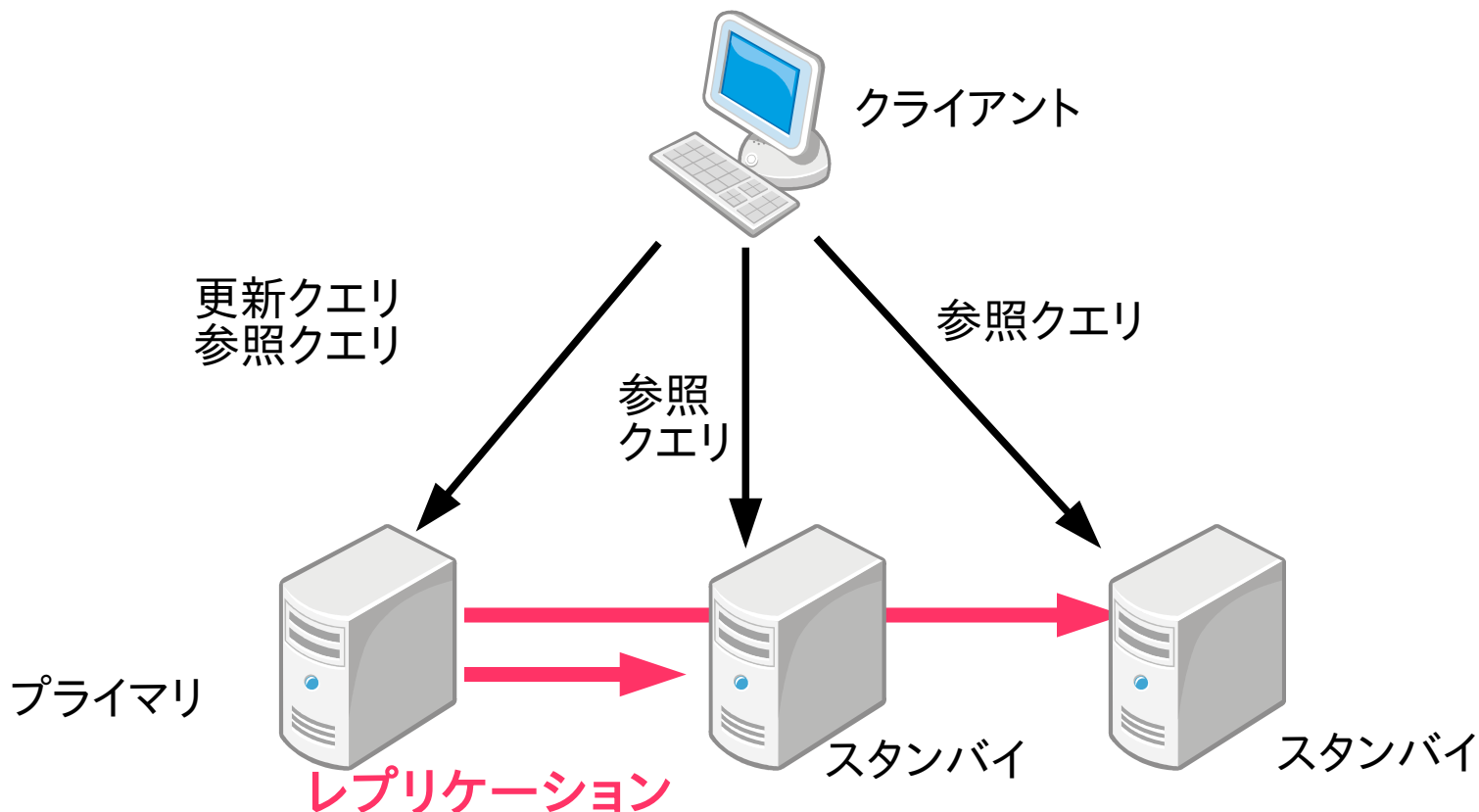
PostgreSQLのストリーミングレプリケーション

- ストリーミングレプリケーション (PostgreSQL 9.0 ~)
 - マスタからスレーブにトランザクションログ (WAL) を転送することによりデータの複製を実現



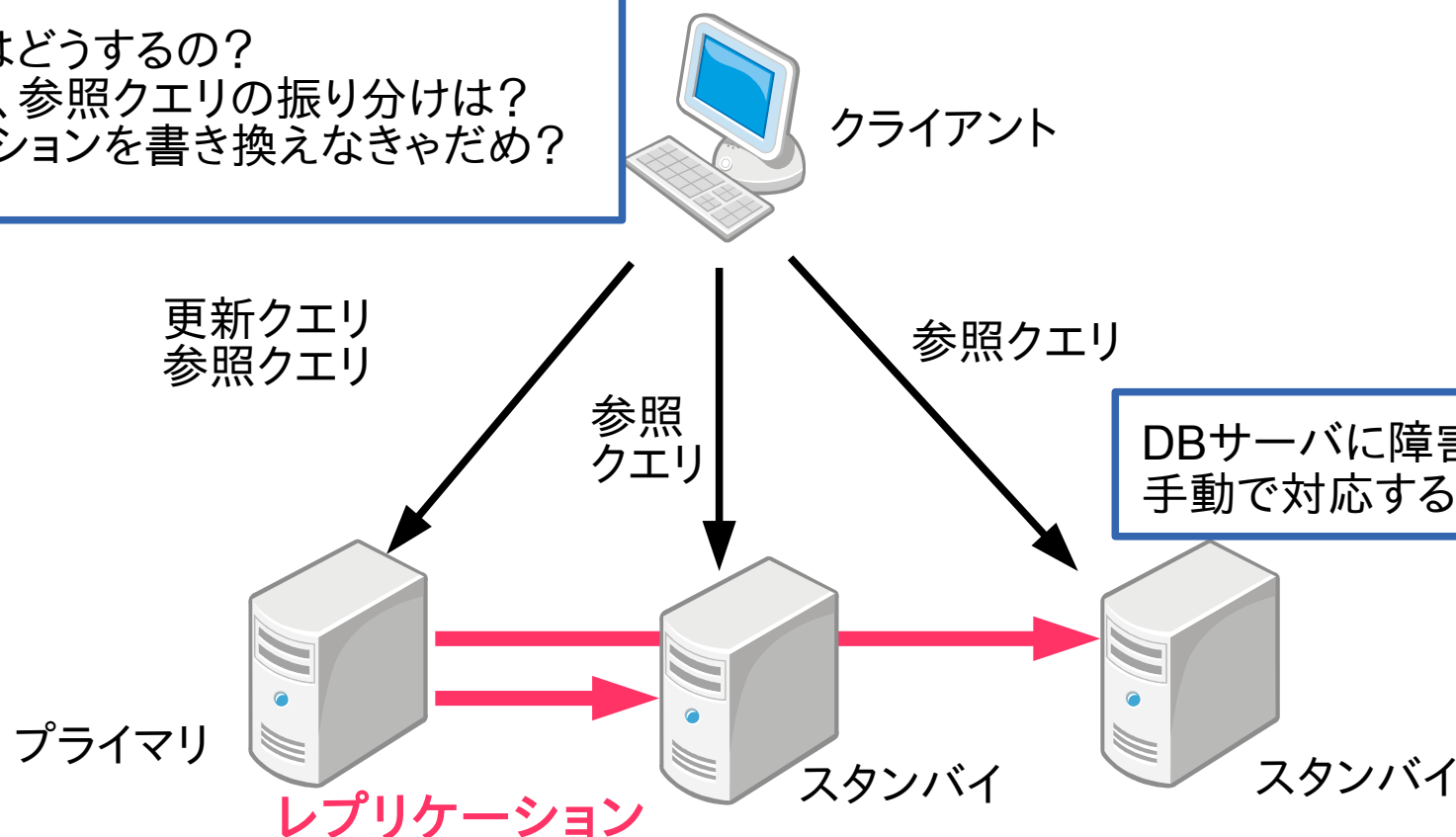
PostgreSQLのストリーミングレプリケーション

- 複数のスタンバイにレプリケーション可能
 - 参照性能のスケールアウトが可能
 - レプリケーションのカスケードリングも可能



しかし、PostgreSQLだけでは・・・

負荷分散はどうするの？
更新クエリ、参照クエリの振り分けは？
アプリケーションを書き換えなきゃだめ？

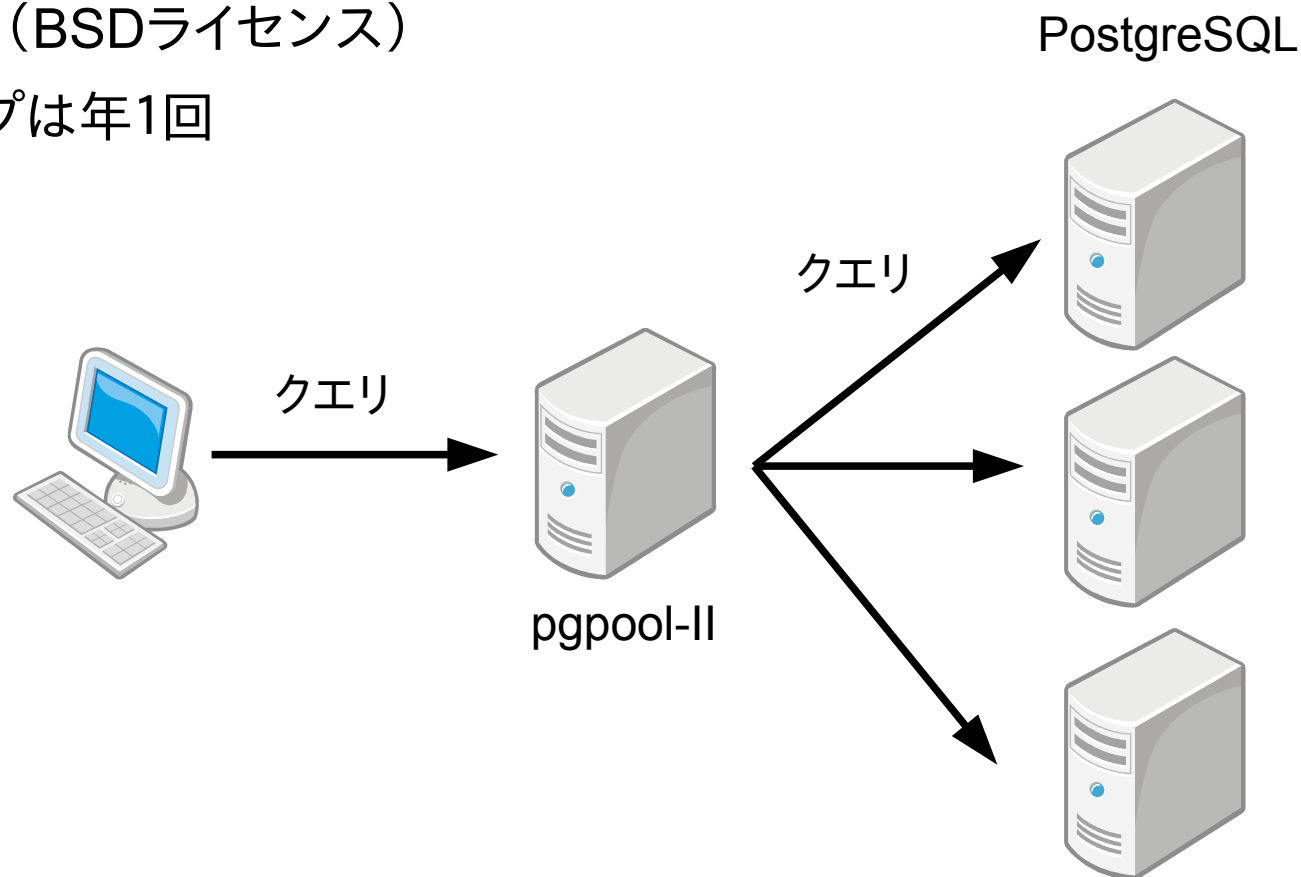
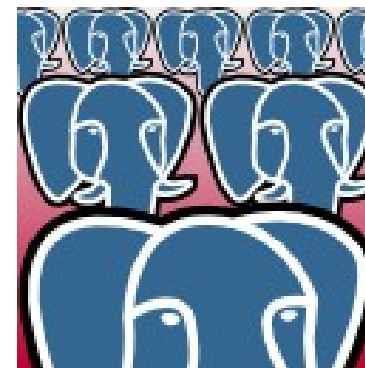


プライマリがダウンしたら更新ができなくなる?!
サービスが停止してしまう!?

そこで pgpool-II です

pgpool-II とは

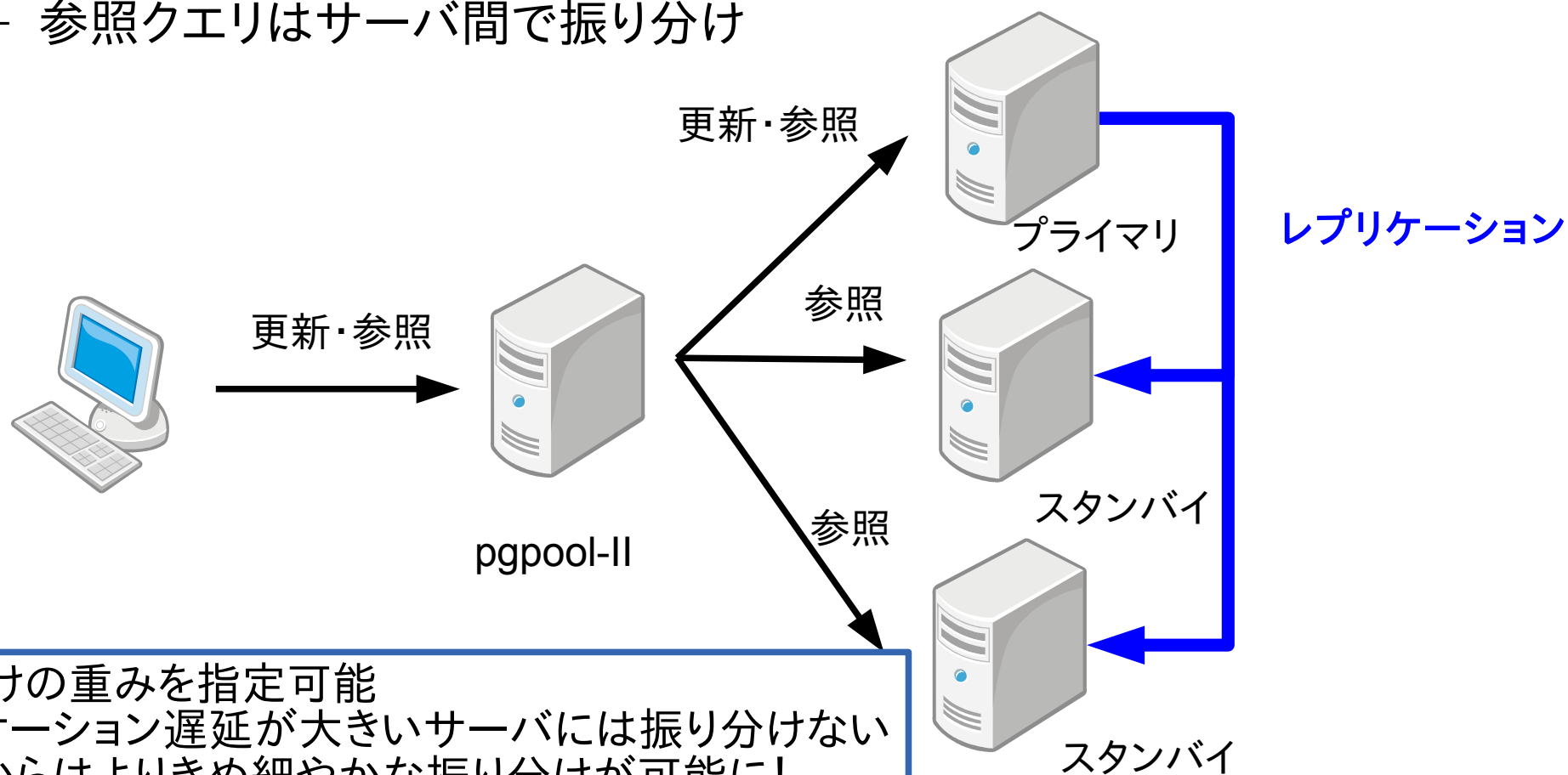
- アプリケーションとPostgreSQLの間に入って、クラスタリング機能を提供するミドルウェア
 - アプリケーションからは普通のPostgreSQLに見える
- オープンソースソフトウェア (BSDライセンス)
 - メジャーバージョンアップは年1回
 - 先月 3.4.0 がリリース
- 多彩な機能
 - コネクションプーリング
 - 参照負荷分散
 - クエリキャッシュ
 - ヘルスチェック
 - 自動フェイルオーバー
 - オンラインリカバリ



参照負荷分散

参照負荷分散

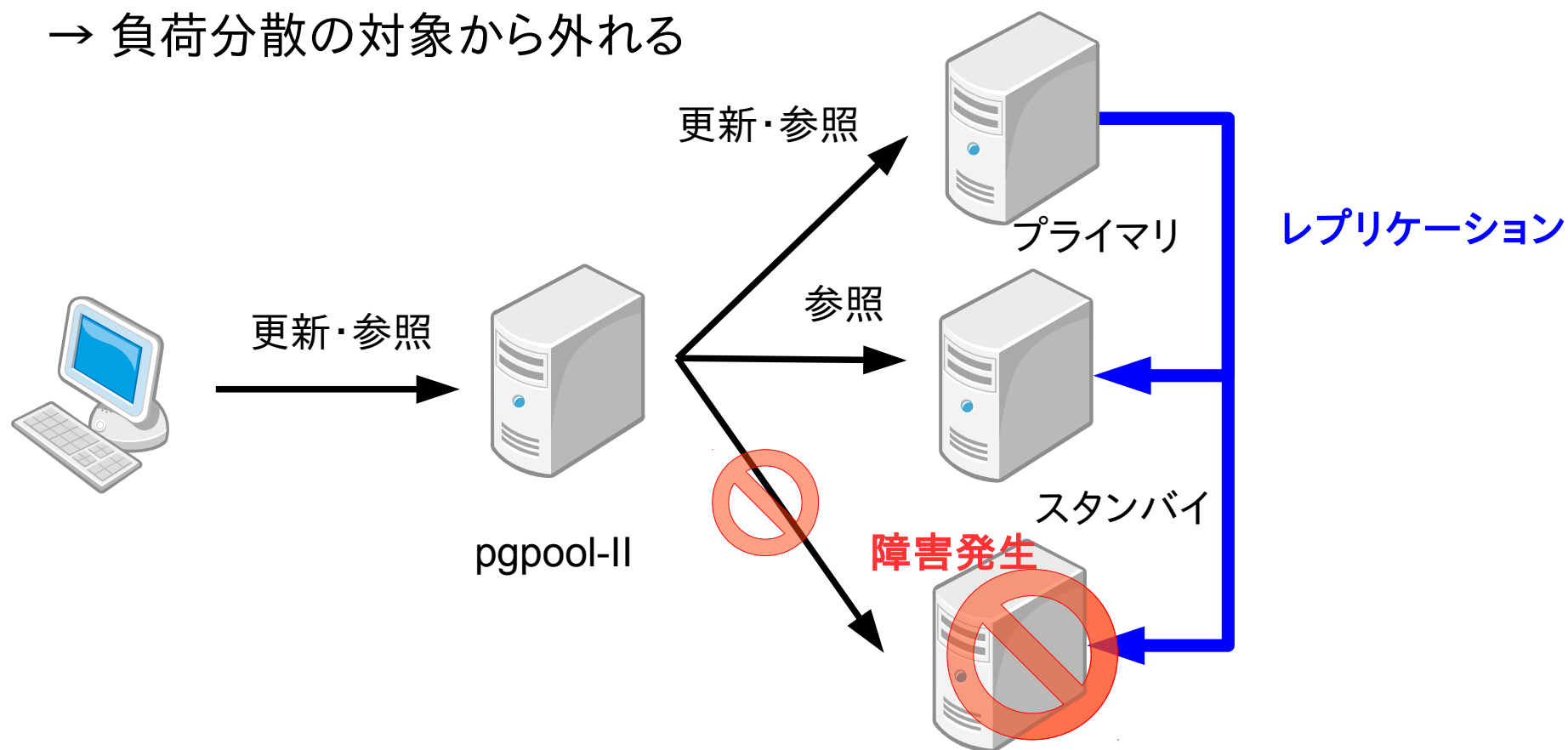
- クエリの自動振り分け
 - 更新クエリはプライマリサーバへ
 - 参照クエリはサーバ間で振り分け



データベースに障害が発生したら？

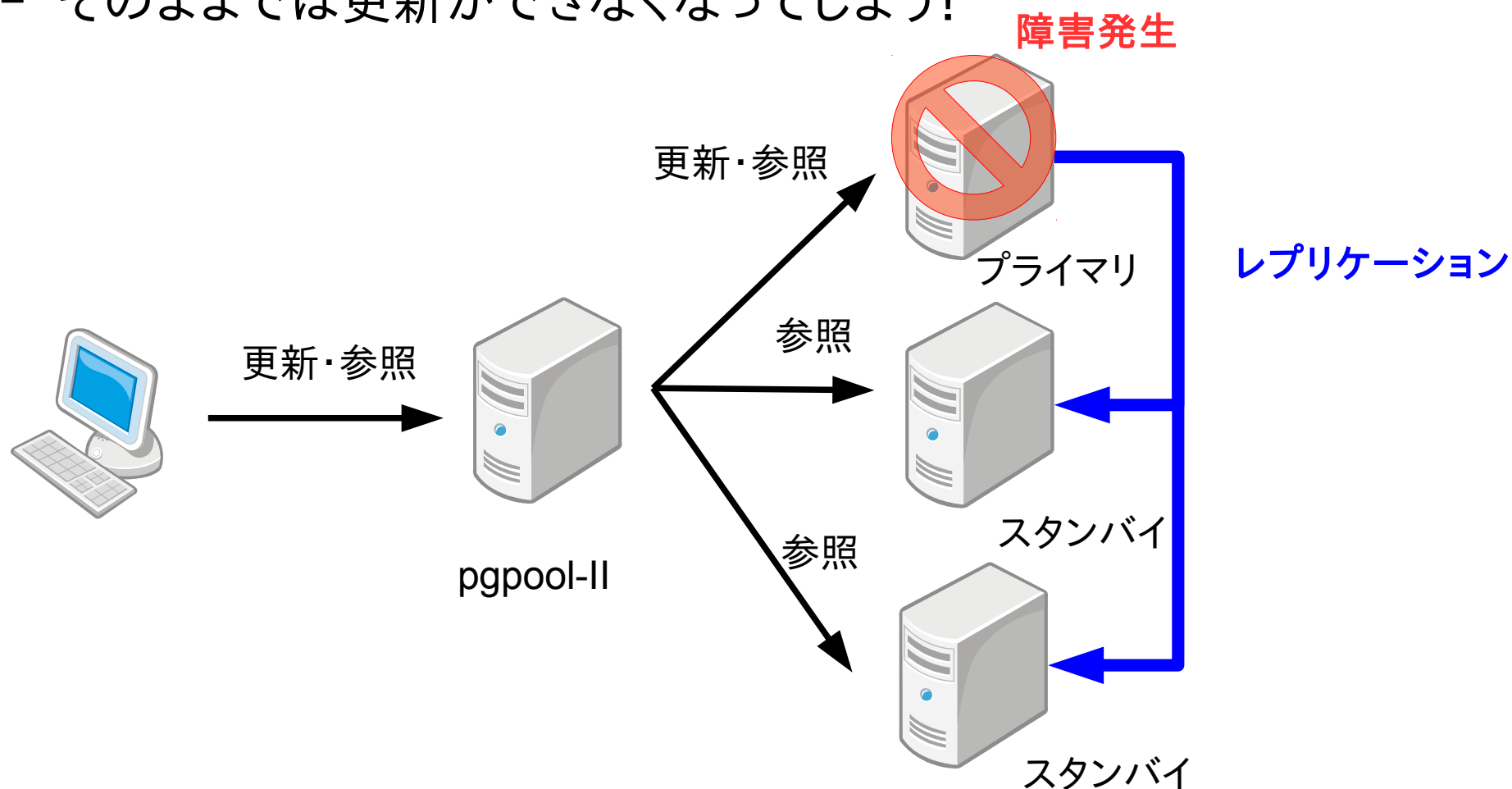
自動フェイルオーバー

- DBサーバの障害を自動検出（ヘルスチェック機能）
 - ダウンしたPostgreSQLを切り離す
→ 負荷分散の対象から外れる



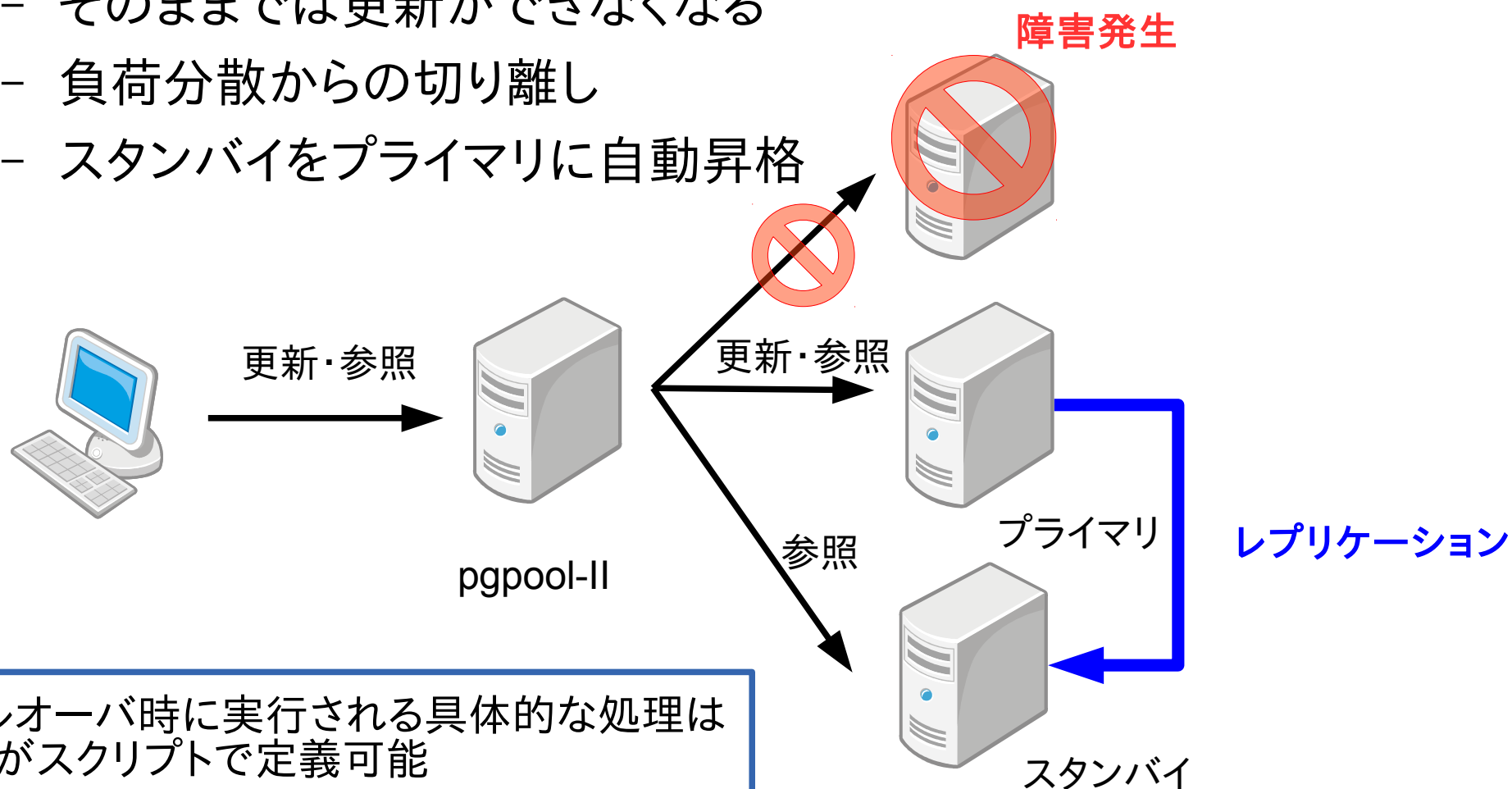
自動フェイルオーバー

- プライマリサーバに障害が発生した場合は？
 - そのままでは更新ができなくなってしまう！



自動フェイルオーバー

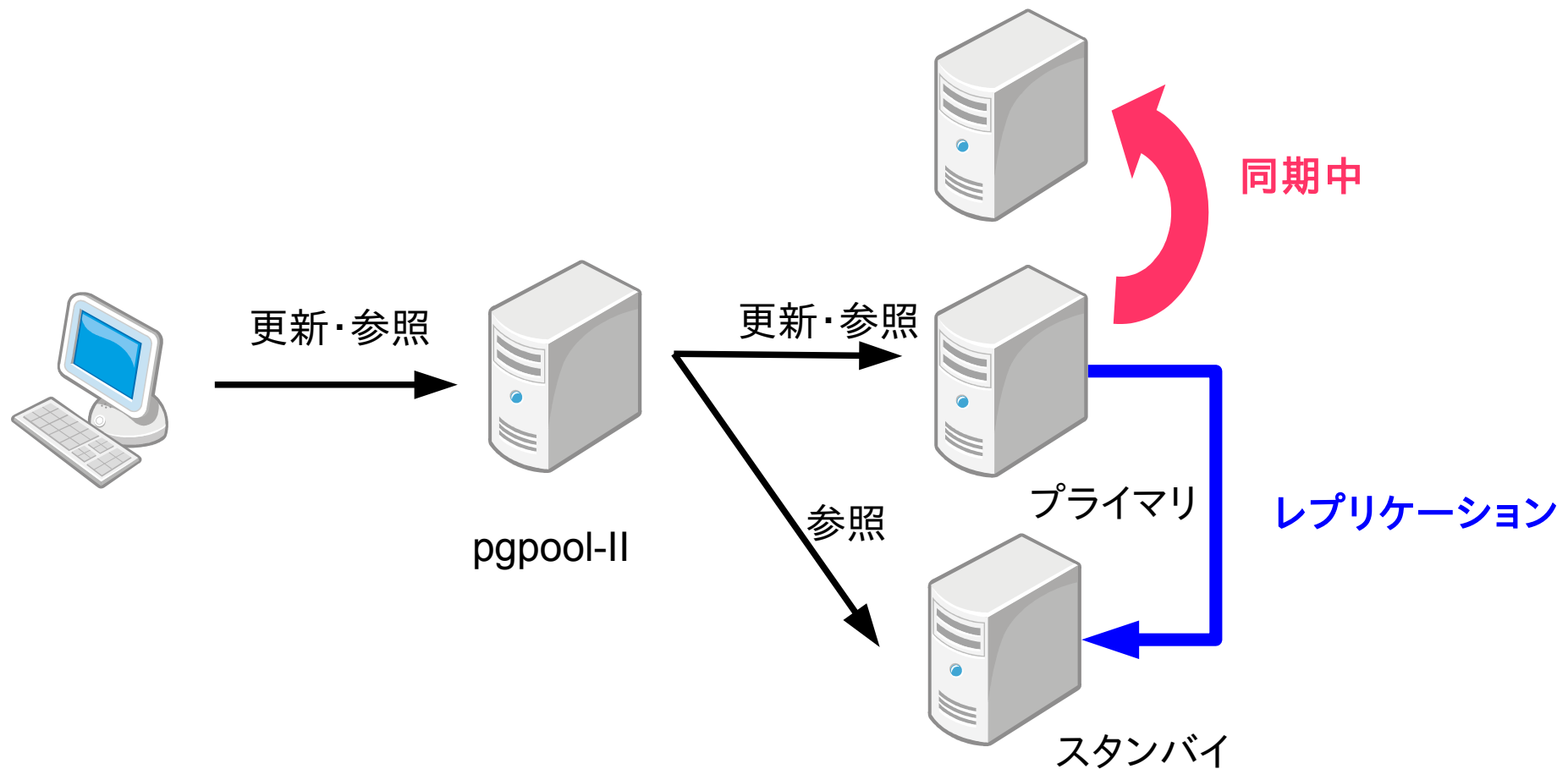
- プライマリサーバに障害が発生した場合
 - そのままでは更新ができなくなる
 - 負荷分散からの切り離し
 - スタンバイをプライマリに自動昇格



ダウンしたデータベースの復旧

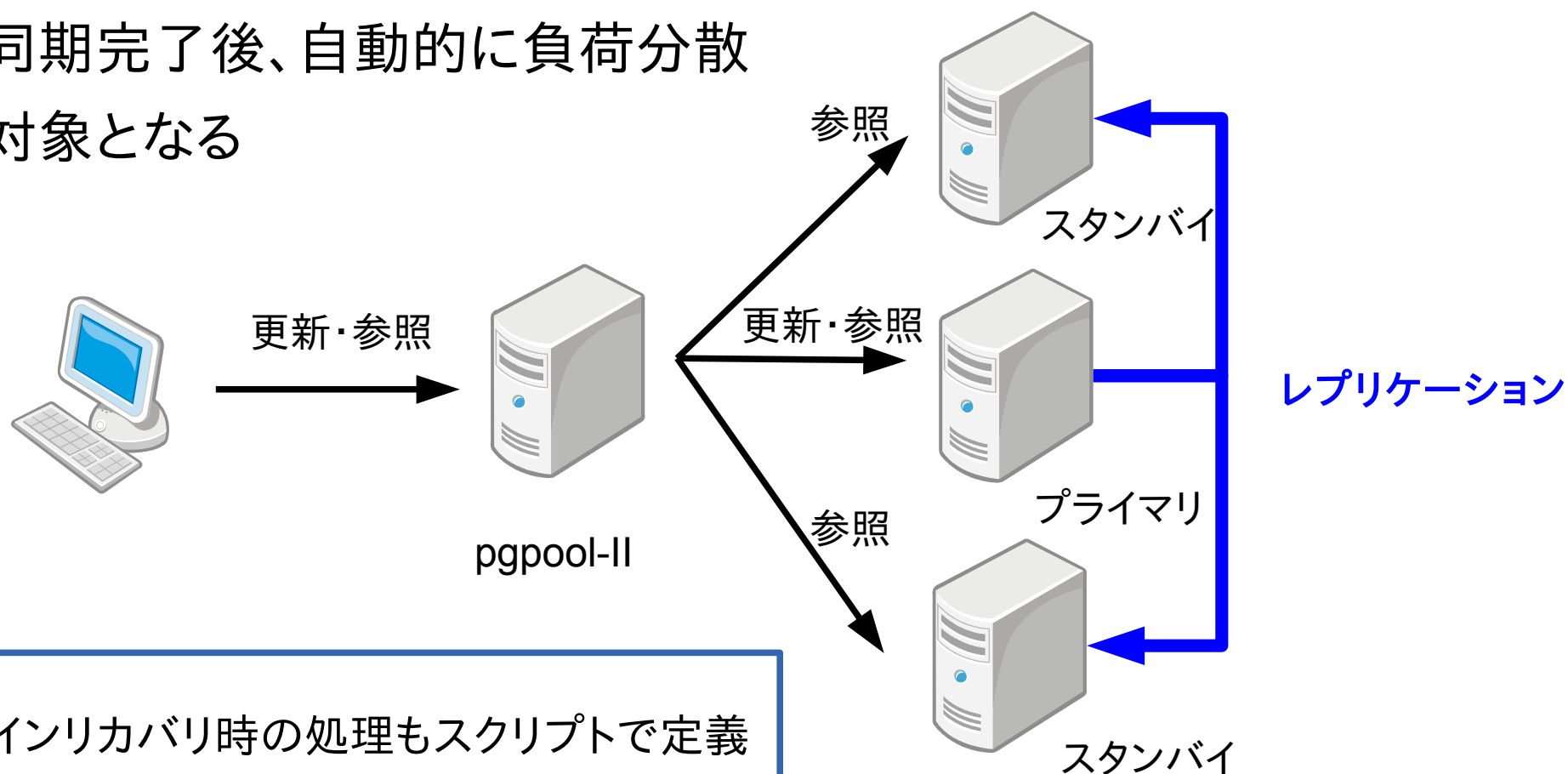
オンラインリカバリ

- ダウンしたスタンバイをプライマリに同期させる
- 同期中も更新が可能



オンラインリカバリ

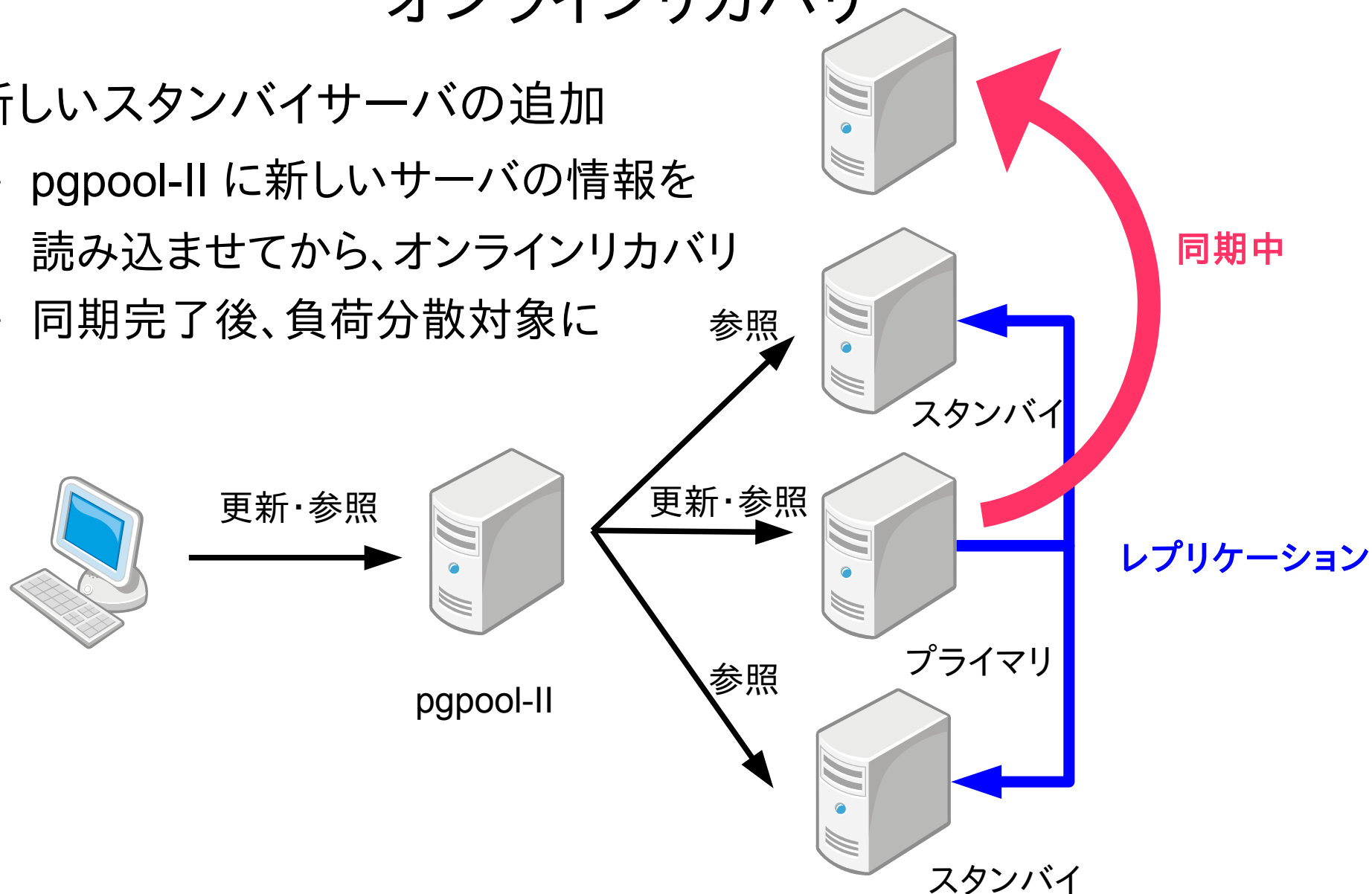
- ダウンしたスタンバイをプライマリに同期させる
- 同期中でも更新可能
- 同期完了後、自動的に負荷分散対象となる



オンラインリカバリ時の処理もスクリプトで定義

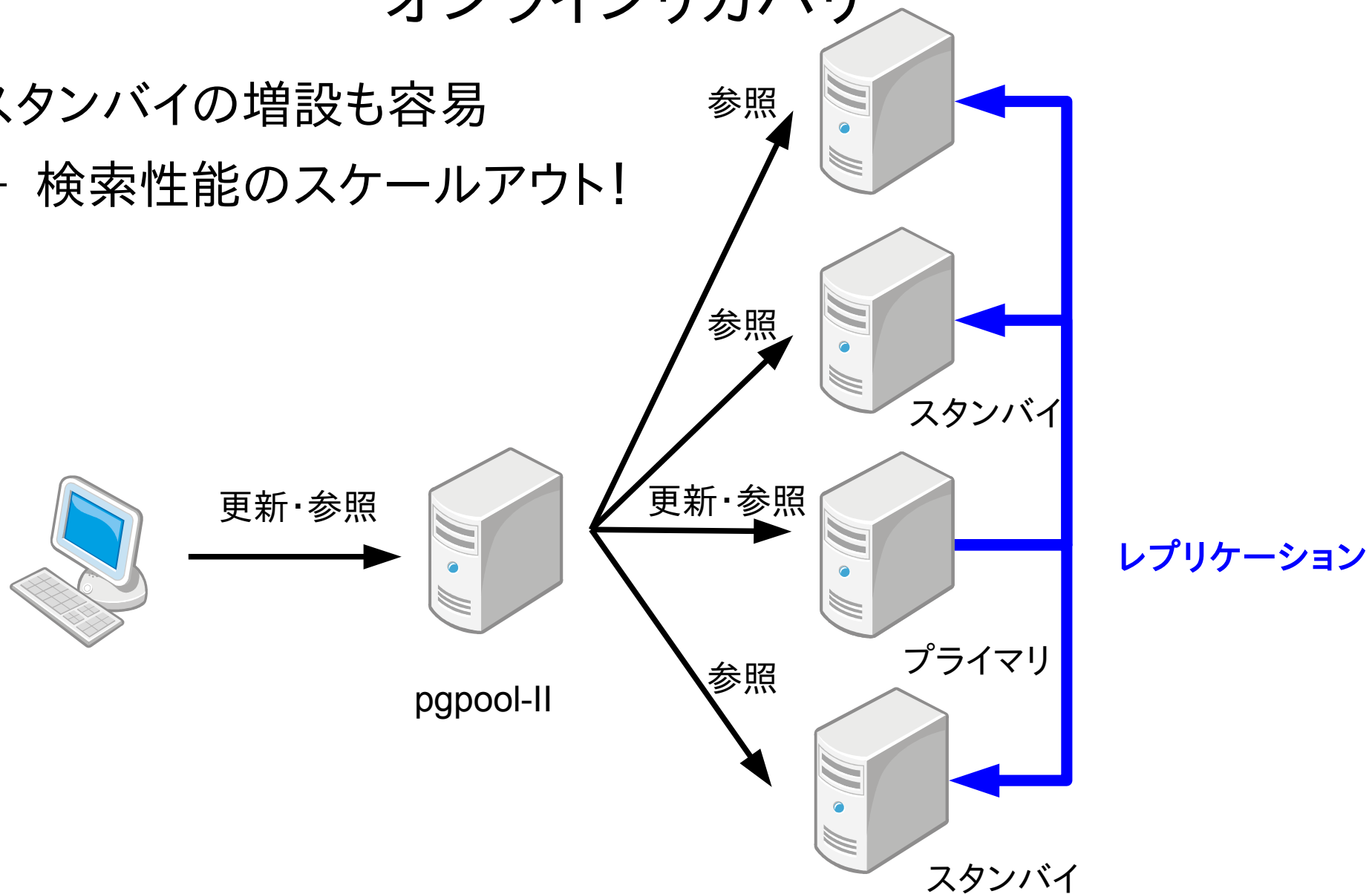
オンラインリカバリ

- 新しいスタンバイサーバの追加
 - pgpool-II に新しいサーバの情報を読み込ませてから、オンラインリカバリ
 - 同期完了後、負荷分散対象に



オンラインリカバリ

- スタンバイの増設も容易
 - 検索性能のスケールアウト!



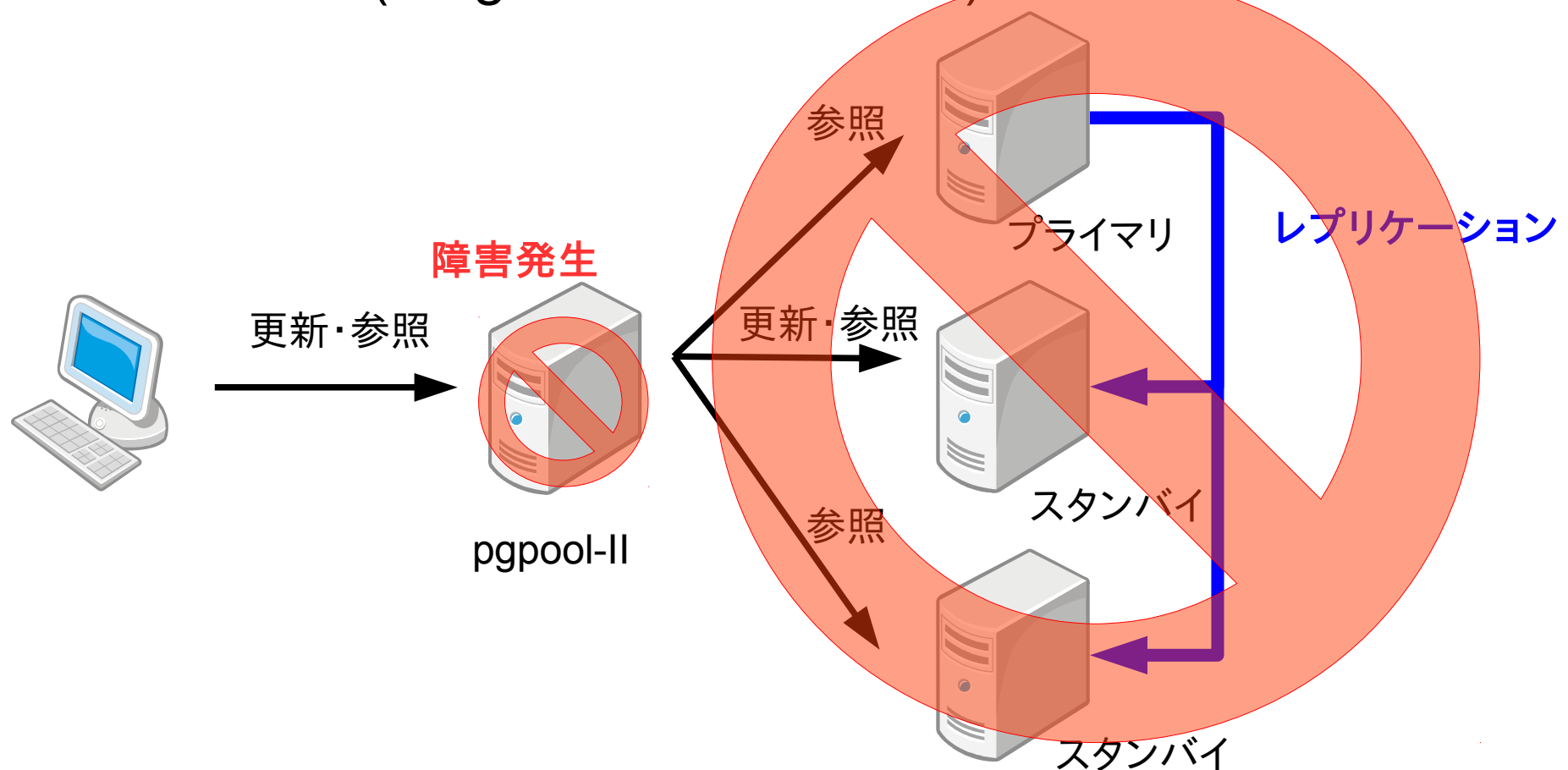
ここまでのまとめ

- 参照性能の負荷分散
 - 更新クエリと参照クエリの適切な振り分け
- 自動フェイルオーバー
 - データベース障害の自動検出&切り離し
 - プライマリがダウンしたら、スタンバイが新プライマリに昇格
- オンラインリカバリ
 - サービスを停めずにダウンしたサーバを復帰
 - 新しいスタンバイの追加も簡単
- 基本的にアプリケーションの書き換えは必要なし！

pgpool-IIに障害が発生したら？

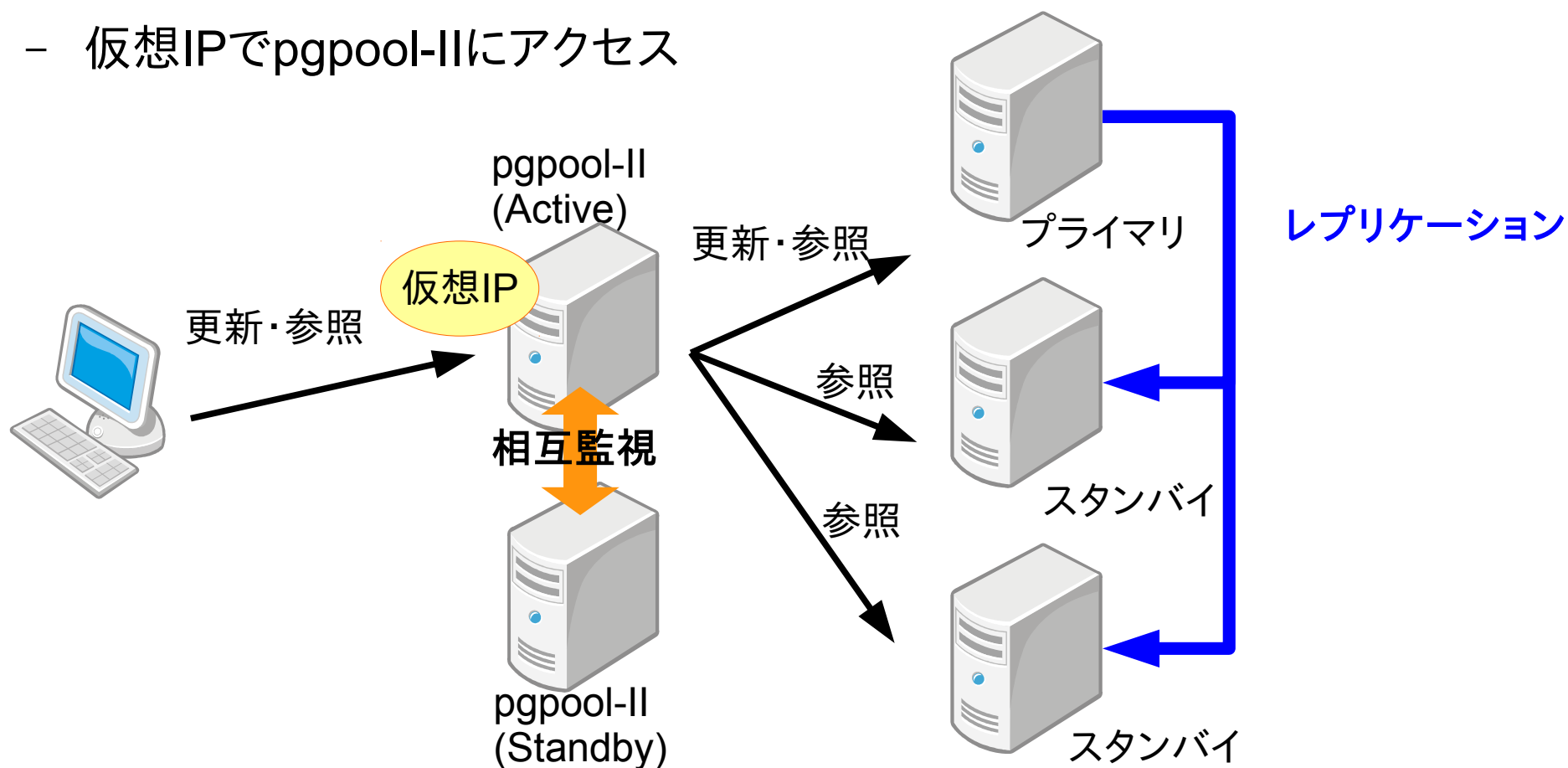
単一障害点？

- もし、pgpool-II に障害が発生したら？！
 - 単一障害点 (Single Point of Failure) じゃないの？



watchdog

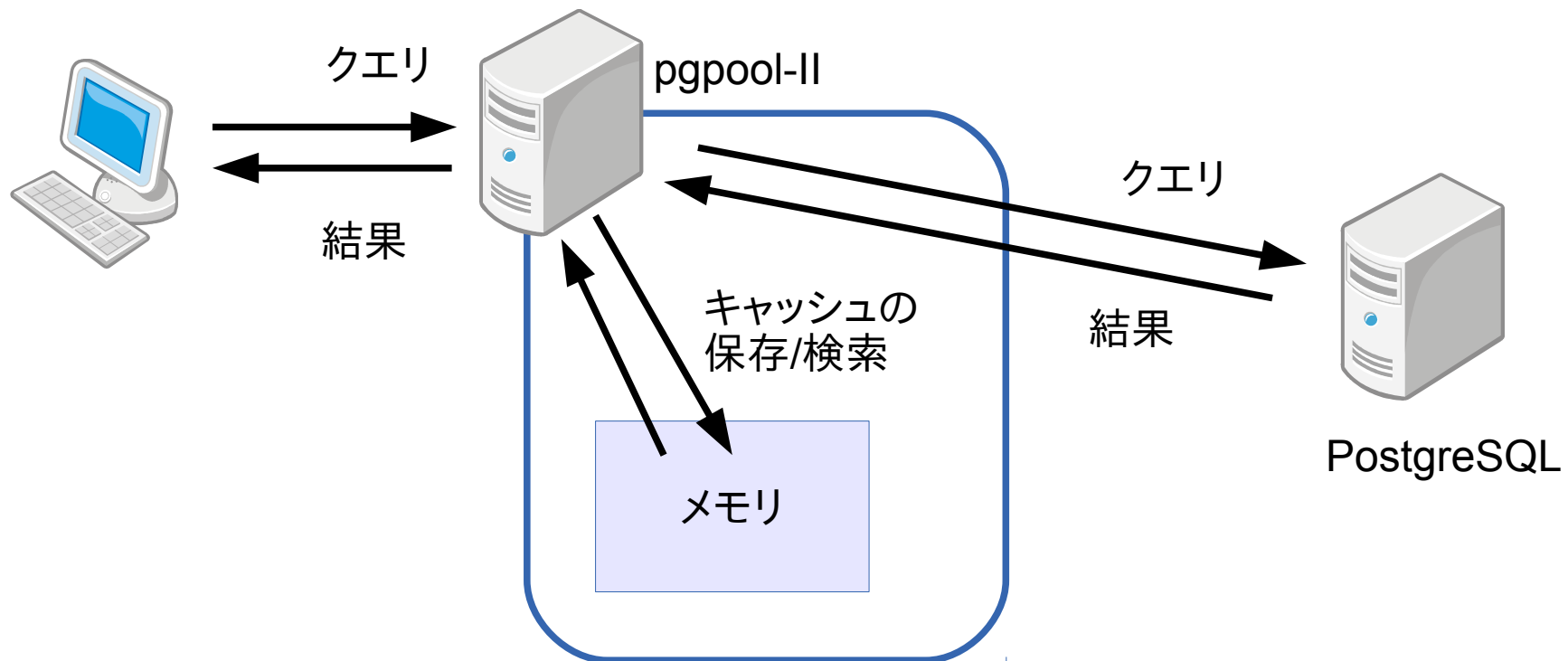
- pgpool-II 組み込みのHA機能
 - pgpool-II を Active/Standby 構成にする
 - 仮想IPでpgpool-IIにアクセス



クエリキャッシュによる性能向上

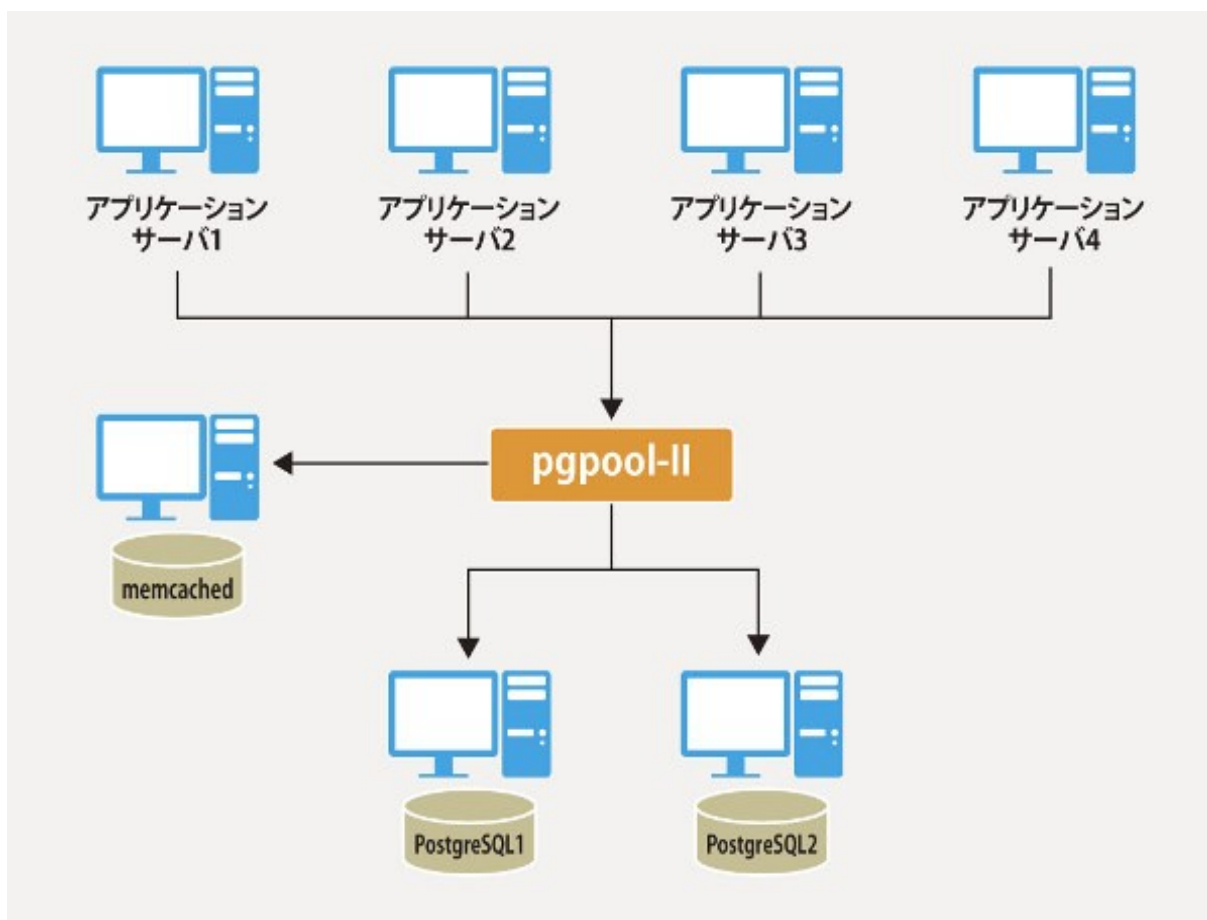
インメモリクエリキャッシュ

- SELECTクエリの結果をメモリ内にキャッシュする機能
 - 同じクエリが来たときに再利用する
 - DBへのアクセスが減り、応答速度が向上



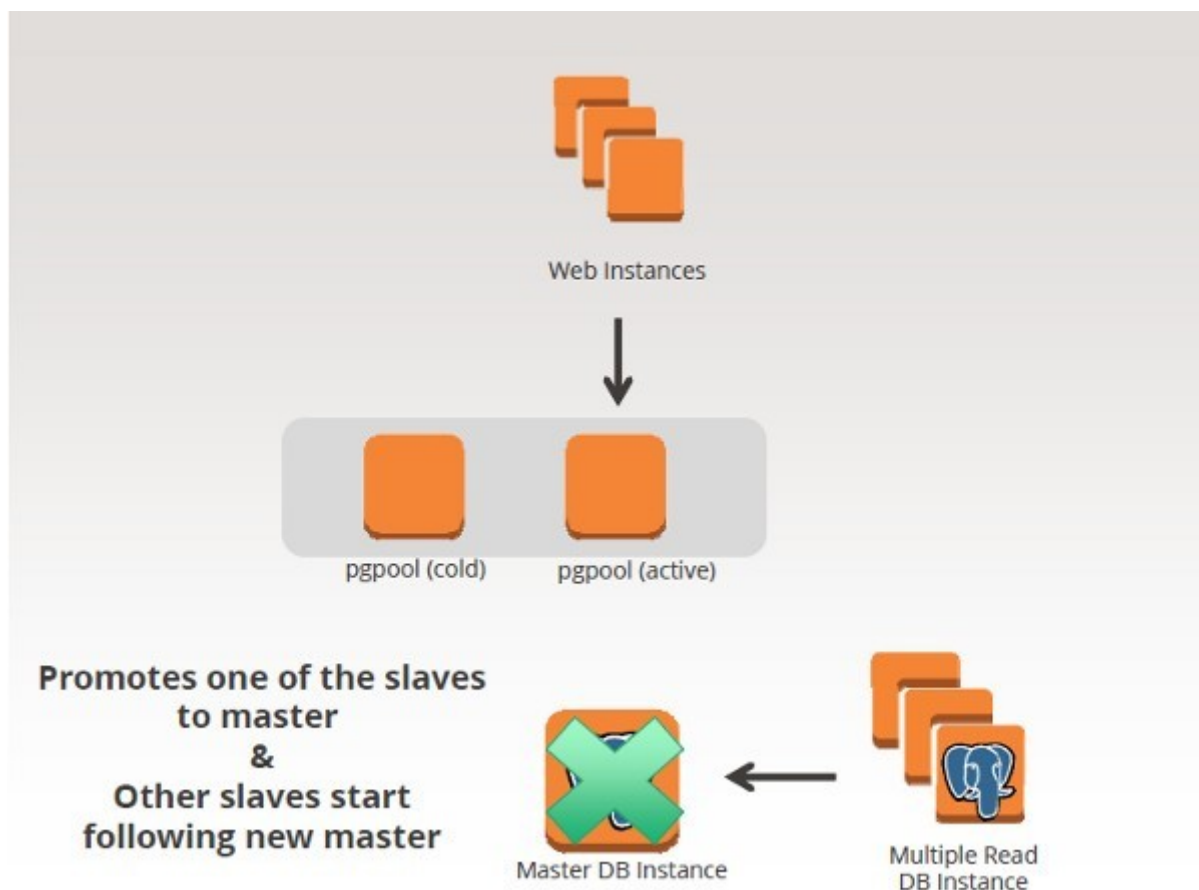
pgpool-II 活用事例

第一法規株式会社様 事例



- 大量の判例などを検索するシステム
- PostgreSQLのストリーミングレプリケーション + pgpool-II で、負荷分散による性能向上、可用性向上
- インメモリクエリキャッシュ機能を活用して検索性能を向上
- 一度発生したDB障害でも pgpool-II の自動フェイルオーバー機能により、サービスは停止することなく継続できた。

株式会社 Gengo様 事例



- 翻訳サービスのクラウドソーシング企業
- AWS上でシステム構築
- 3万トランザクション/日
- PostgreSQLのマスタの仮想イメージがある日突然壊れたトラブルをきっかけに、SPoF解消のため pgpool-II を導入
- 同時にPostgreSQLのバージョンアップ。オンラインリカバリ機能を活用しダウンタイムを最小限に。
- AWSによる強制インスタンス再起動メンテナンスも自動フェイルオーバー機能で乗り切った

pgpool-II の運用管理

pgpool-IIの運用管理コマンド

- pgpool コマンド
 - 起動
 - pgpool start
 - 停止
 - pgpool stop
 - 設定ファイルの読み込み
 - pgpool reload
 - 起動時オプション
 - デーモンモード、デバッグモード、設定ファイルの指定、バックエンドステータスの破棄、...

pgpool-IIの運用管理コマンド

- pcp コマンド
 - DBノードをpgpool-IIの管理下から外す
 - pcp_detach_node
 - DBノードをpgpool-IIの管理下に組み入れる
 - pcp_attach_node
 - オンラインリカバリを実行する
 - pcp_recovery_node
 - DBノードの状態を取得する
 - pcp_node_info
 - watchdog ステータスを取得する
 - pcp_watchdog_info

pgpool-IIの運用管理コマンド

- SHOW コマンド
 - pgpool-II の内部情報を取得できるクエリ
 - ノード情報、プロセス情報、コネクションプーリング、キャッシュヒット率、現在の設定、バージョン

```
postgres=# SHOW pool_nodes;
 node_id | hostname | port | status | lb_weight | role
-----+-----+-----+-----+-----+-----
 0      | /tmp    | 11002 | 2      | 0.333333 | primary
 1      | /tmp    | 11003 | 2      | 0.333333 | standby
 2      | /tmp    | 11004 | 2      | 0.333333 | standby
(3 rows)
```

```
postgres=# SHOW pool_version;
 pool_version
-----
 3.4.0 (tataraboshi)
(1 row)
```

GUI で管理したい

pgpoolAdmin

- pgpool-II のGUI管理ツール
 - PHPで書かれたWebアプリケーション
 - pgpool-II と DBノードの操作および状態確認が可能

pgpoolAdmin

pgpool Administration Tool

pgpoolステータス

ノードステータス

pgpool.conf設定

pgpoolAdmin設定

パスワード変更

ログアウト

pgpoolAdmin

- This is pgpoolAdmin on 6b3647dd1fb2
- For pgpool-II 3.4.
- Login
 - login user: postgres
 - superuser: yes

Backend info (PostgreSQL)

サマリー プロセス情報 **ノード情報** ログ

ノード情報

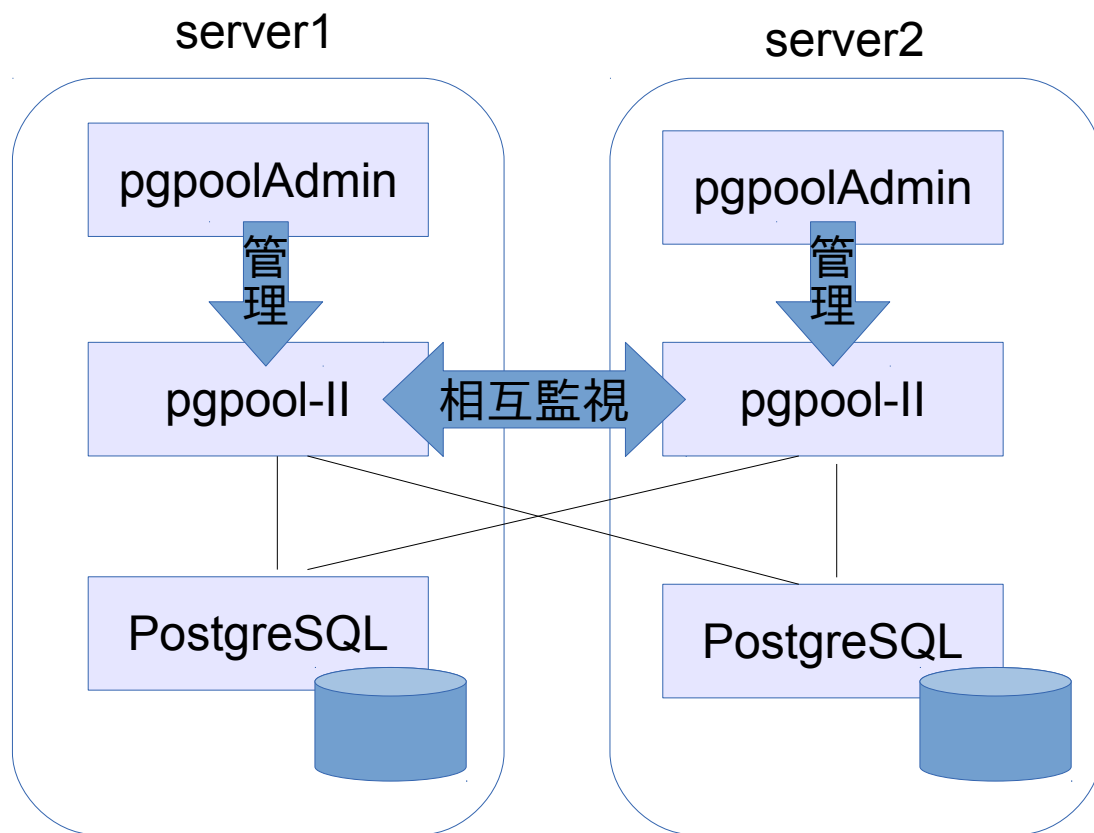
Refresh info: 10 seconds

IPアドレス	ポート	ステータス	ウェイト
node 0 /tmp	11002	ノード稼働中。接続有り。プライマリとして稼働中	postgres: アップ 0.333
node 1 /tmp	11003	ノード稼働中。接続有り。スタンバイとして稼働中	postgres: アップ 0.333
node 2 /tmp	11004	ノード稼働中。接続有り。スタンバイとして稼働中	postgres: アップ 0.333

導入が難しそう・・・？

pgpool-II インストーラ

- pgpool-II と pgpoolAdmin を2台のサーバにインストール
- 表示される質問に答えていくだけで設定が完了
 - レプリケーション
 - Watchdog
 - 負荷分散
 - ヘルスチェック
 - インメモリクエリチェック
 - フェイルオーバーバスキプト
- pgpool-II オフィシャルサイトから入手可能



テスト目的でとりあえず使ってみたい？

- pgpool-setup
 - pgpool-II の tar ball に同梱されているツール
 - コマンド1つで、ローカルマシン上にテスト環境（デモ環境）を構築
 - pgpool-II 1台 + 複数の PostgreSQL の構成
 - レグレッションテストでも利用されている
 - （余談ですが）
pgpool-II のビルドファームが運用開始
 - 毎日1回レグレッションテストを実行してメール通知

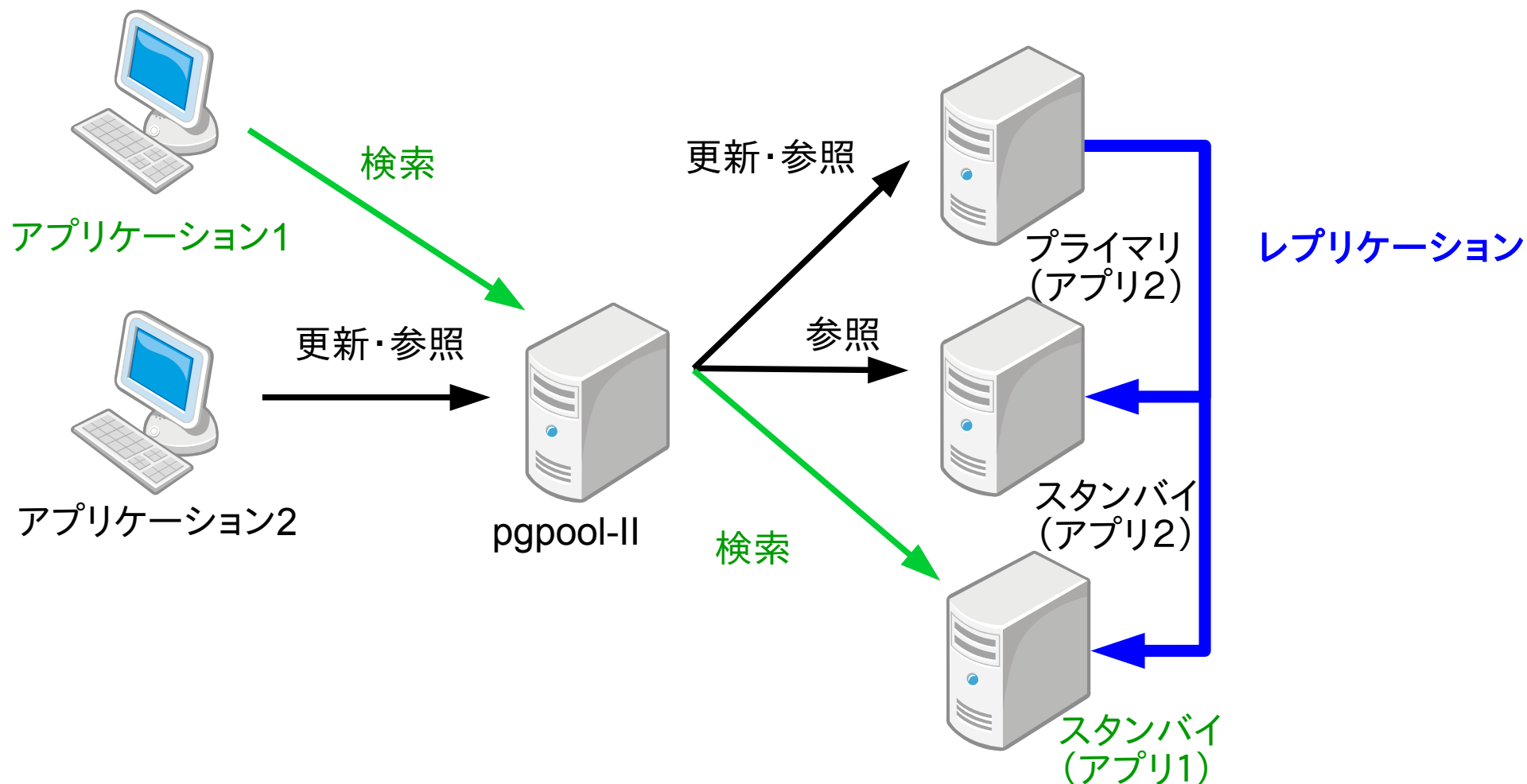
pgpool-II 3.4.0 の新機能

pgpool-II 3.4.0

- pgpool-II 3.4.0 が11月にリリース!
- 参照負荷分散のきめ細かい指定が可能に
 - アプリケーション名やデータベース名で、クエリの送り先を制御可能
- PostgreSQLへの接続待ちタイムアウトが指定可能に
 - AWSなどの仮想環境ではネットワーク遅延が大きく、これがフェイルオーバー発生の原因となっていた
- PostgreSQL 9.4 のパーサを移植
 - 新しい構文や機能に対応
- PostgreSQLのメモリマネージャ & 例外処理マネージャを移植
- IPv6対応
- etc....

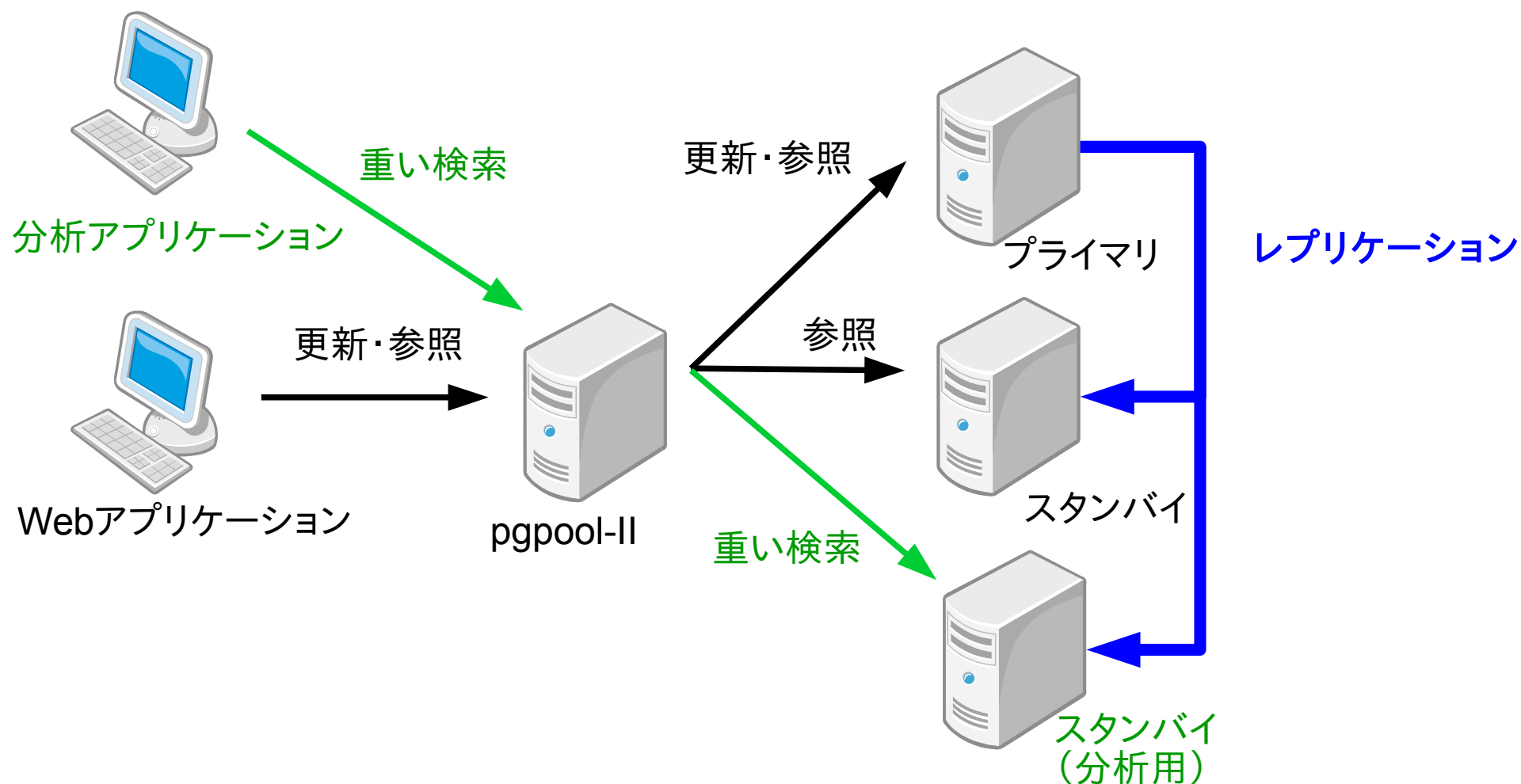
きめ細かな負荷分散

- アプリケーション名、DB名によって接続先が指定できる



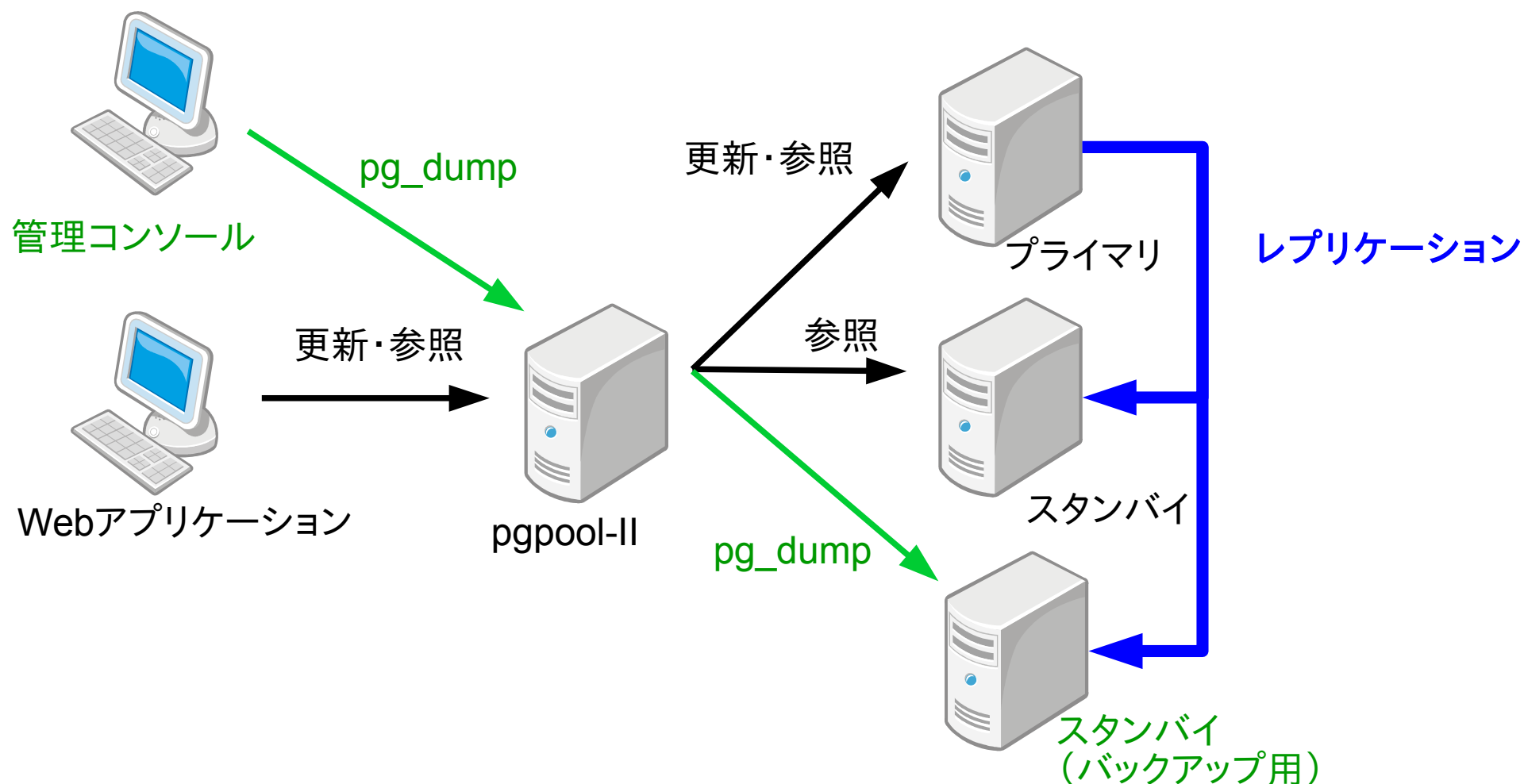
きめ細かな負荷分散

- アプリケーション名、DB名によって接続先が指定できる



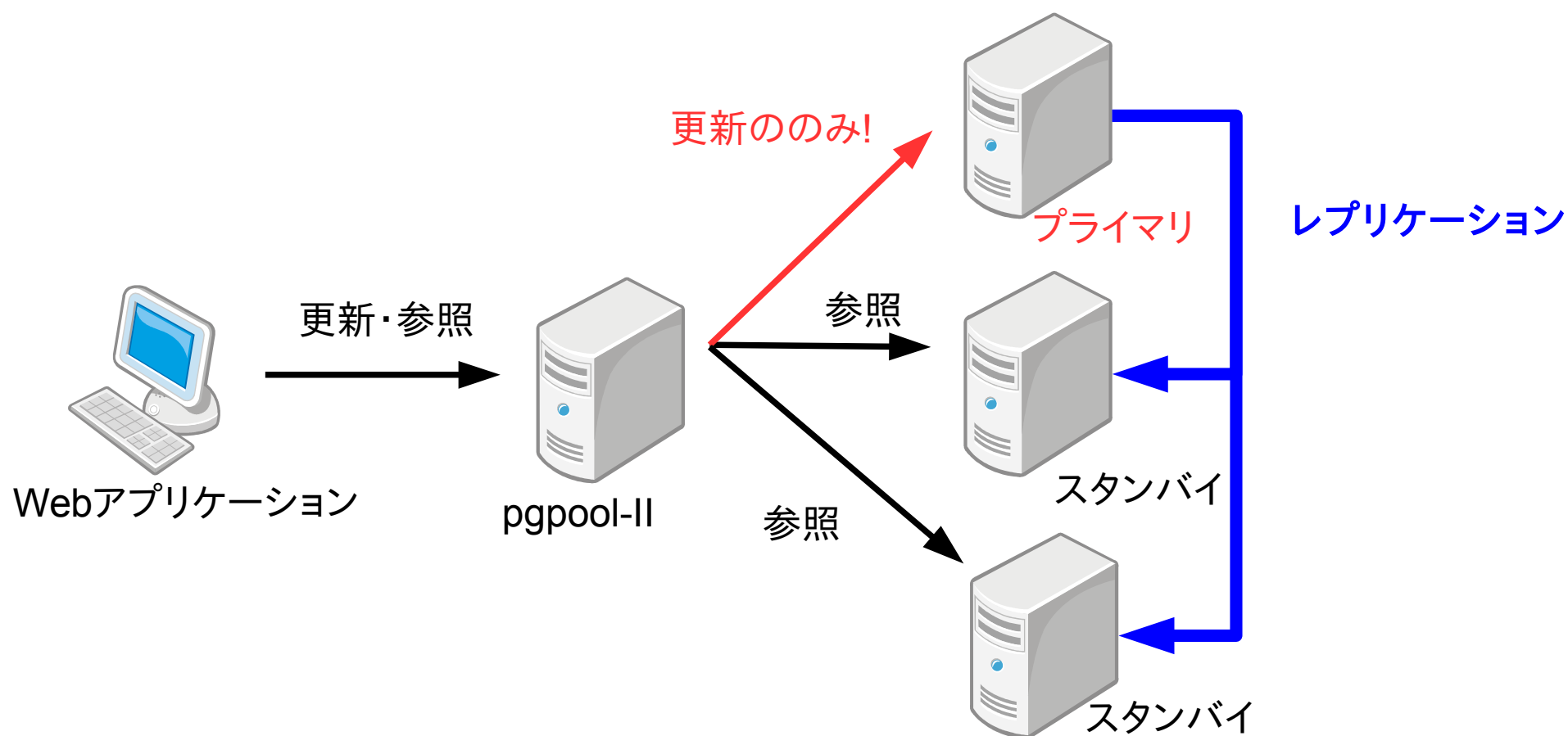
きめ細かな負荷分散

- アプリケーション名、DB名によって接続先が指定できる



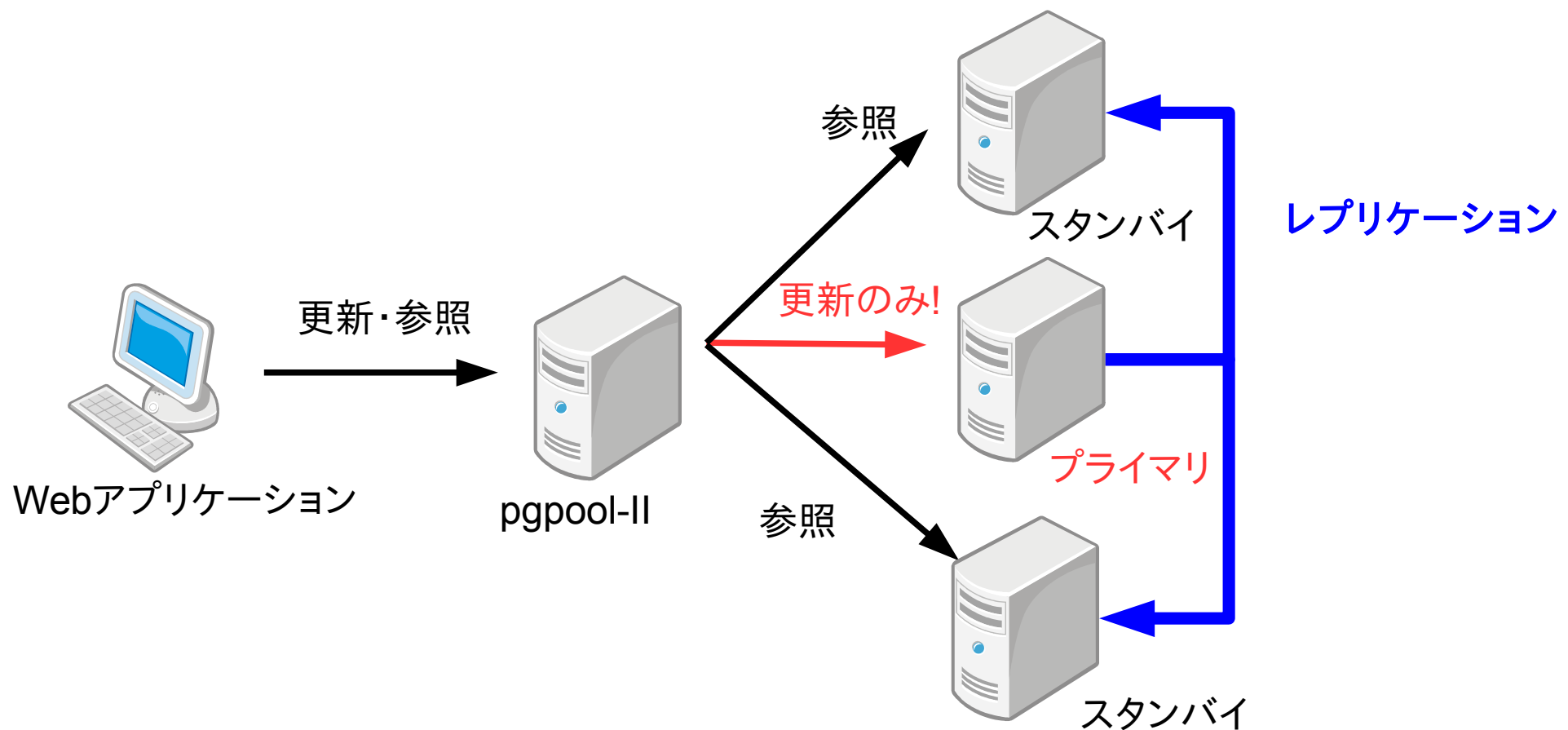
きめ細かな負荷分散

- アプリケーション名、DB名によって接続先が指定できる



きめ細かな負荷分散

- アプリケーション名、DB名によって接続先が指定できる



今後の予定

- pgpool-II 3.5 の開発
 - 性能改善
 - PostgreSQL 9.5 対応
 - pcp コマンドの改善
 - 取得できる内部情報の追加
 - 負荷分散の状況
- pgpool-II オフィシャルのyum レポジトリ
 - 近日公開予定

まとめ

- PostgreSQLのクラスタ技術
 - ストリーミングレプリケーション + pgpool-II がおすすめ!
- pgpool-II の機能
 - 負荷分散 & 自動クエリ振り分け
 - 自動フェイルオーバー
 - オンラインリカバリ & サーバの追加も容易
 - watchdog で pgpool-II 自体の高可用性化
 - インメモリクエリキャッシュ
 - PgpoolAdmin、インストーラ、テスト環境構築ツール
- pgpool-II 3.4.0 新機能の紹介
 - きめ細かな参照負荷分散の指定

参考URL

- pgpool-II オフィシャルサイト
 - <http://www.pgpool.net/>
 - <http://www.pgpool.net/jp/>
- SRA OSS, Inc. 日本支社
 - セミナー資料、事例情報、技術情報
 - <http://www.pgecons.org/>
- Let's Postgres
 - PostgreSQL 情報のポータルサイト
 - <http://lets.postgresql.jp/>

オープンソースとともに



URL: <http://www.sraoss.co.jp/>
E-mail: sales@sraoss.co.jp
Tel: 03-5979-2701