

PostgreSQLで高信頼性DBを構築しよう

OSC.Enterprise Osaka 2014

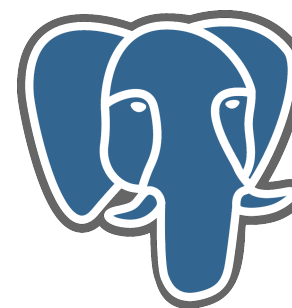
SRA OSS, Inc. 日本支社
取締役支社長
石井 達夫

自己紹介

- SRA OSS, Inc.で日本支社運営を担当
- PostgreSQLの開発者(コミッタ)
- PostgreSQL用のクラスタソフト “pgpool-II” のコミュニティリーダー

SRA OSS, Inc.のご紹介

- 1999年よりPostgreSQLサポートを中心にOSSビジネスを開始、2005年に現在の形に至る
- 主なビジネス
 - PostgreSQL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



PowerGres



ミッションクリティカル領域で 求められること

- 可用性(High Availability:HA)
- 性能(スケールアップ、スケールアウト)
- これらの要件を実現するために開発されたのがクラウドシステム

可用性(High Availability:HA)とは

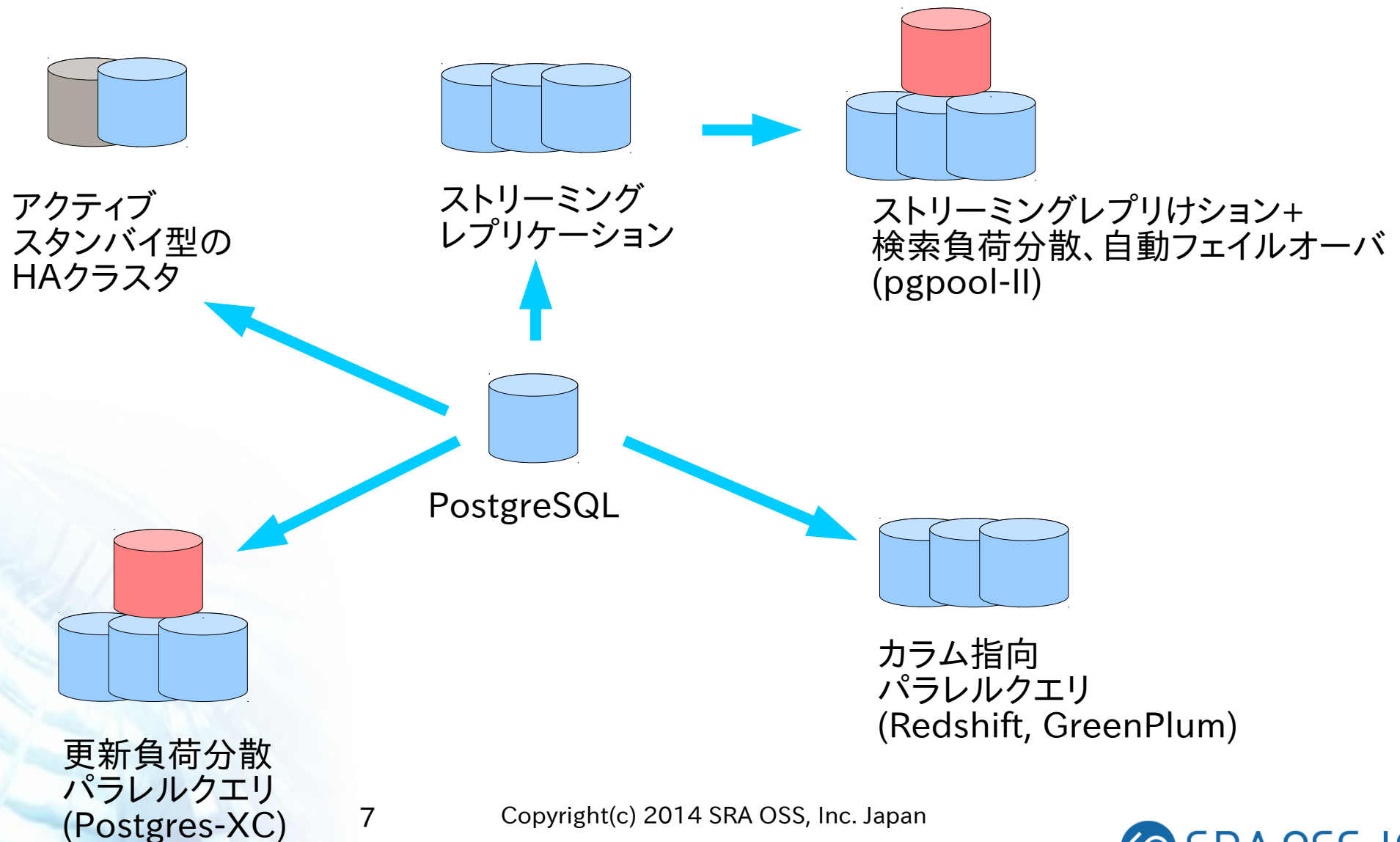
- システムが停止しないで使い続けられる能力のこと
 - 年間の稼働時間を目安にする(稼働率)
- 可用性を高めるには、以下のどちらかまたは両方を使う。今日は冗長化の方向でお話します(信頼性がさほど高くないコンポーネントを複数使って冗長化するのが最近のトレンド)
 - 構成要素の信頼性を上げる
 - 冗長化する

稼働率目標と実現手法

稼働率	年間動作不能時間	実現手法
99%	3日15時間36分	バックアップとリストア
99.9%	8時間46分	バックアップとリストアに加え、夜間待機が必要
99.99%	52分34秒	クラスタ+手動切り替え+夜間待機
99.999%	5分15秒	クラスタ+自動切り替え
99.9999%	32秒	特殊なハードウェアが必要

Wikipediaより

PostgreSQLにおけるクラスタ技術のトレンド

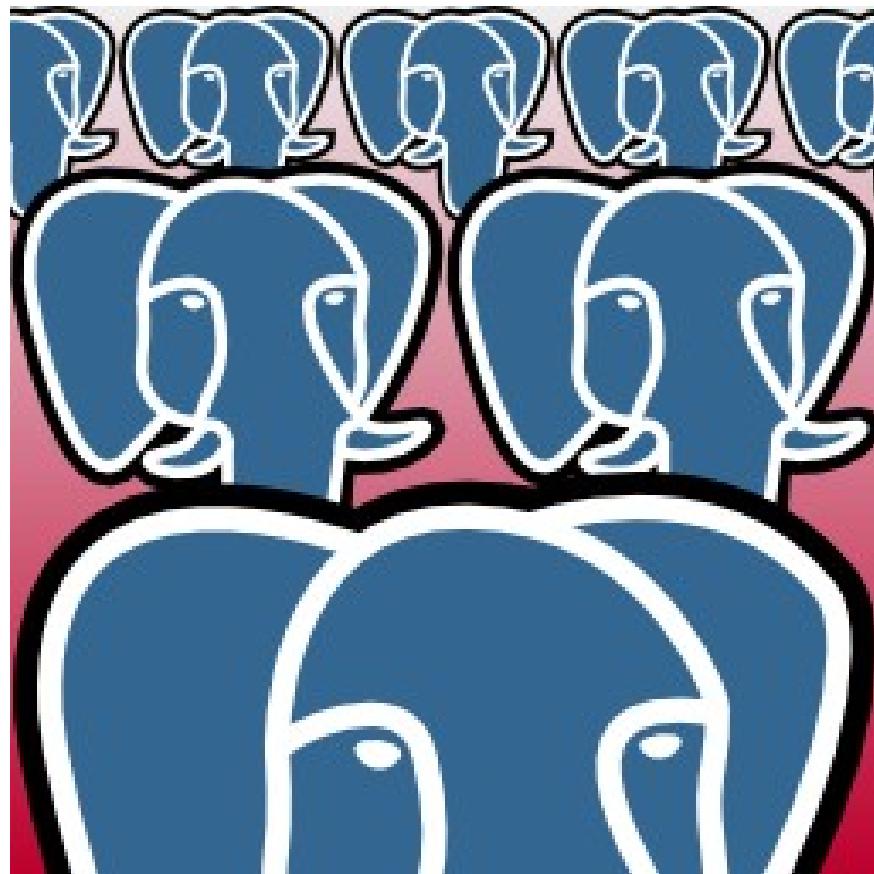


主なPostgreSQL用のクラスタ・レプリケーションツール

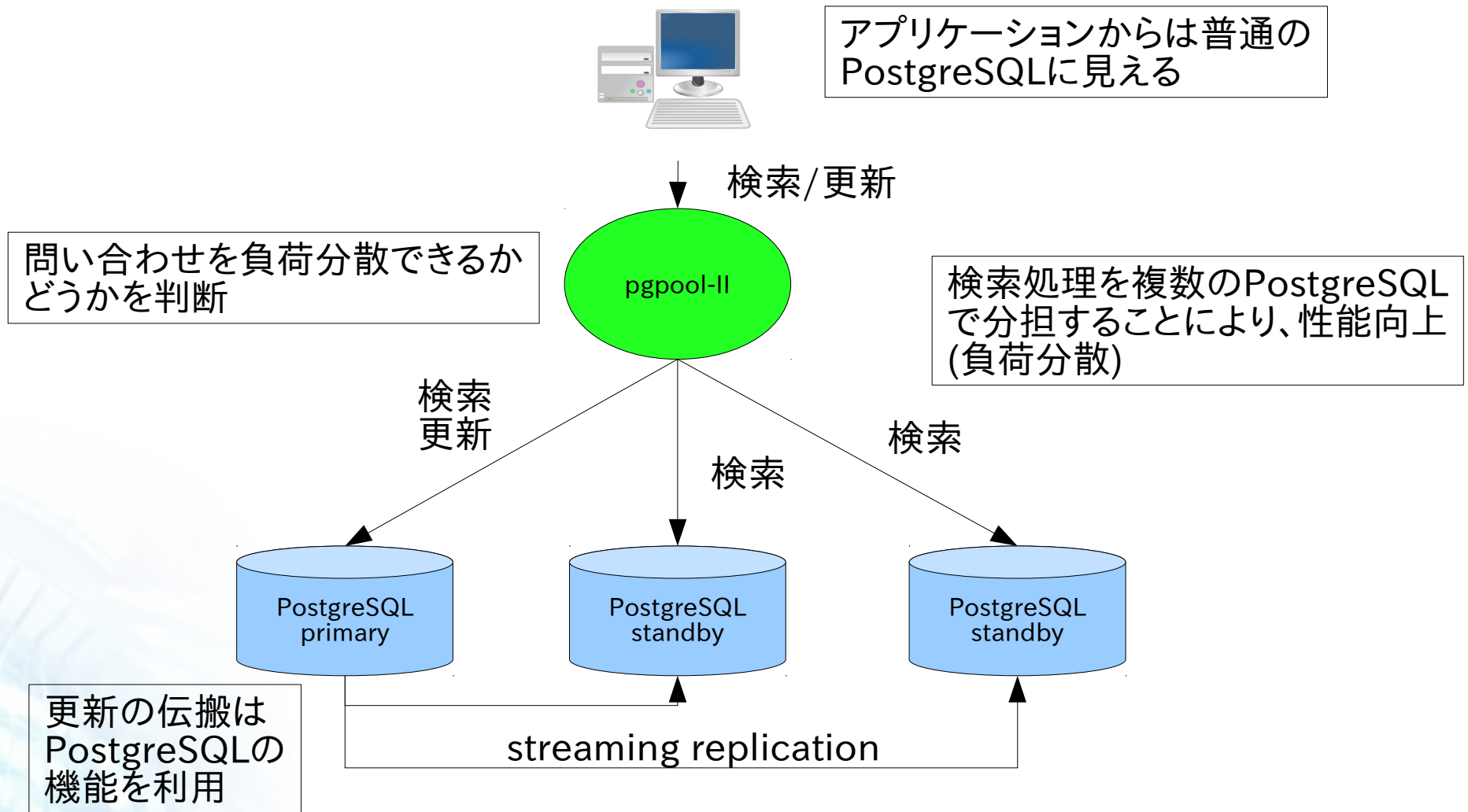
- ストリーミングレプリケーション
 - PostgreSQLネイティブの非同期レプリケーションシステム。master+複数のread only replica。簡単、確実、速い
 - でも自動フェイルオーバーや負荷分散はしてくれない
- pgpool-II
 - クライアントとPostgreSQLの間に入るproxyとして動作。pgpool-IIネイティブの同期レプリケーションと、ストリーミングレプリケーションの両方をサポート。コネクションプーリング、負荷分散、自動フェイルオーバー機能あり。
- Postgres-XC
 - PostgreSQLを改造した比較的新しいクラスタシステム。パラレルクエリと、書き込み負荷分散をサポート。まだ実績は少ない。

pgpool-IIとは

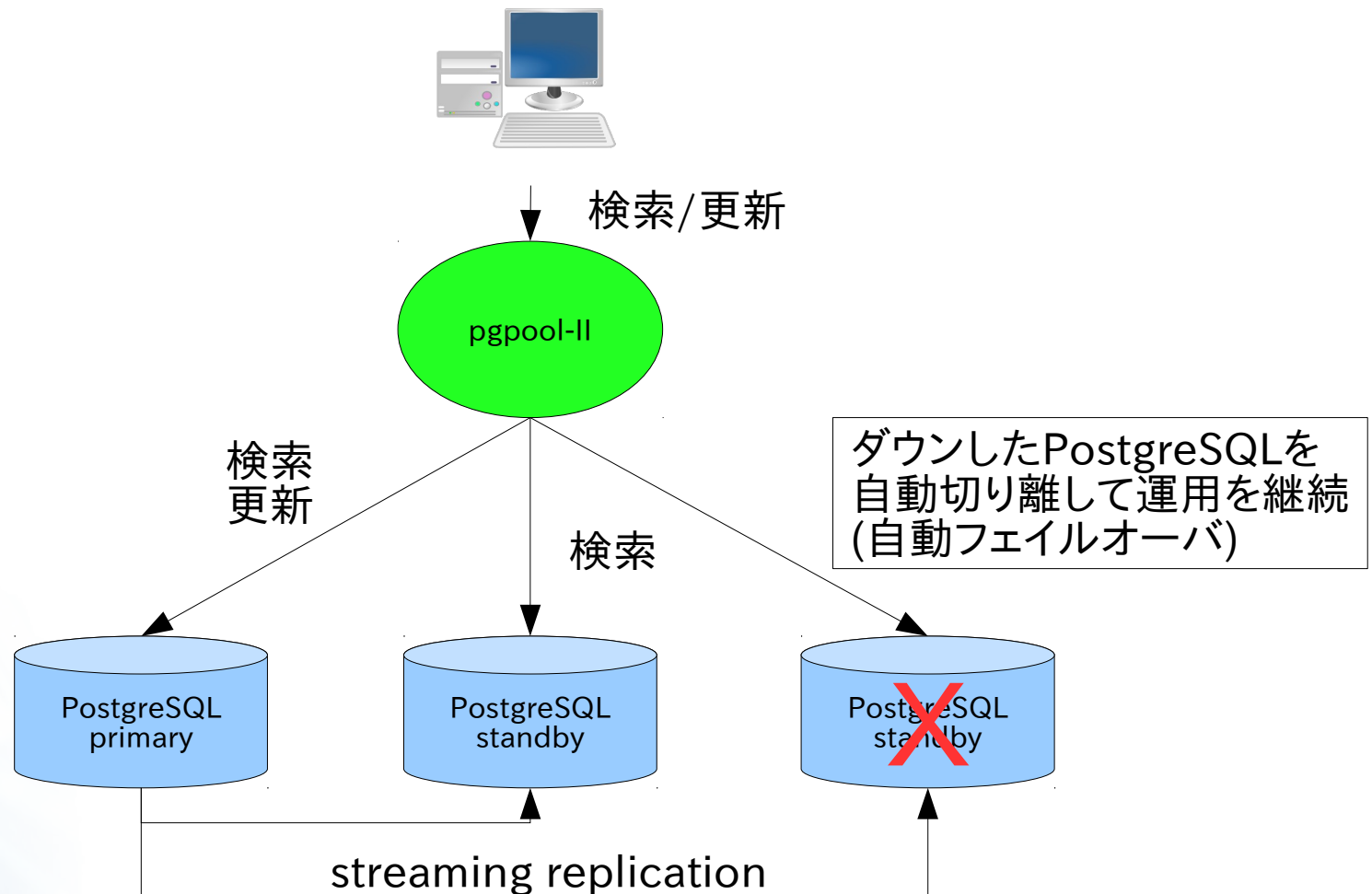
- PostgreSQLで使えるクラスタリングミドルウェア
- OSSとして公開 (BSDライセンス)
- PostgreSQLから独立したproxyとして動作、アプリケーションの改修は最低限
- 幅広いPostgreSQLのバージョンに対応
- 性能や可用性を高める多くの機能
- メジャーバージョンアップは年に1回、マイナーバージョンアップは2-3ヶ月に1回、最新バージョンは3.3系



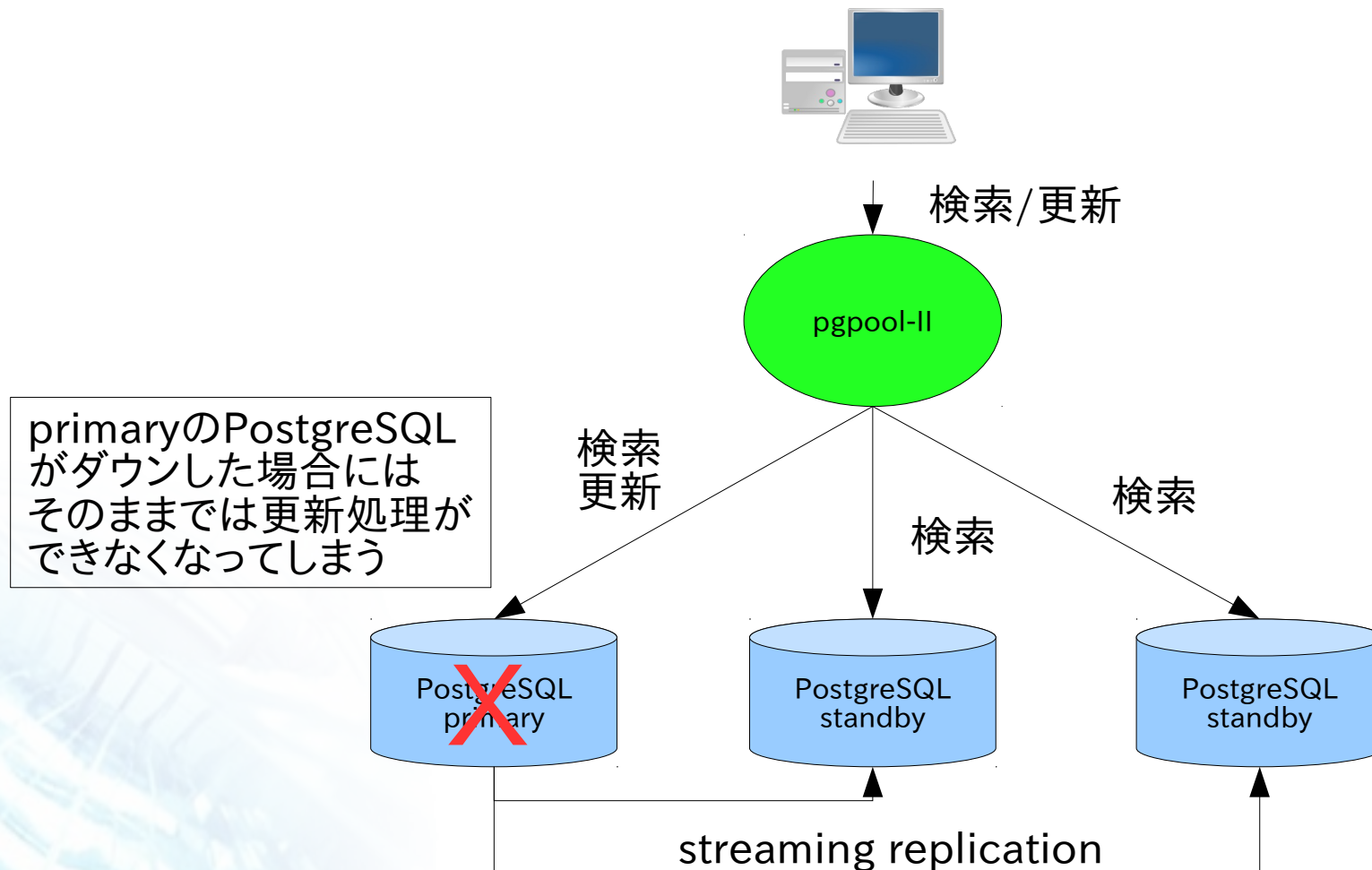
pgpool-IIによるクラスタの概念(1)



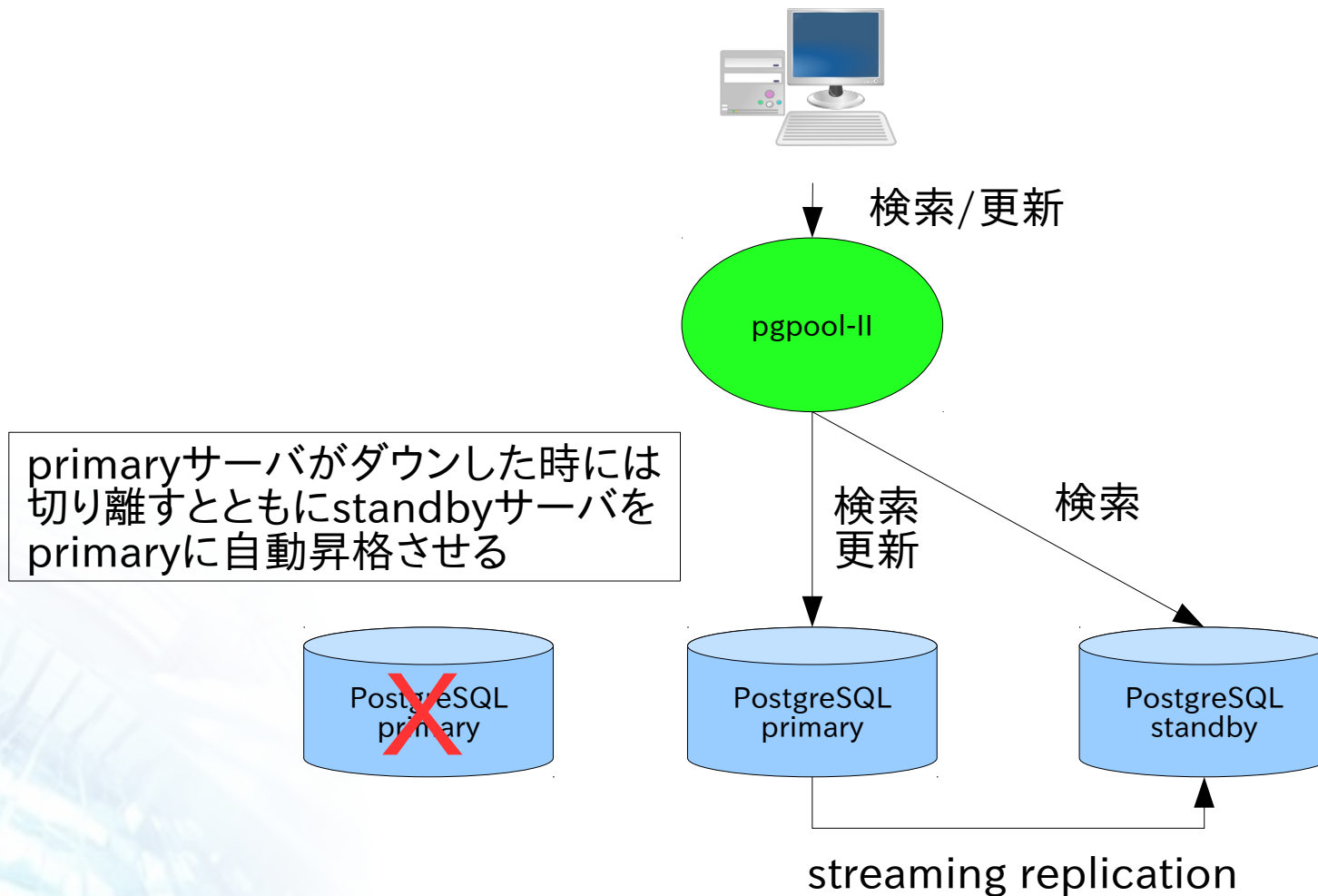
pgpool-IIによるクラスタの概念(2)



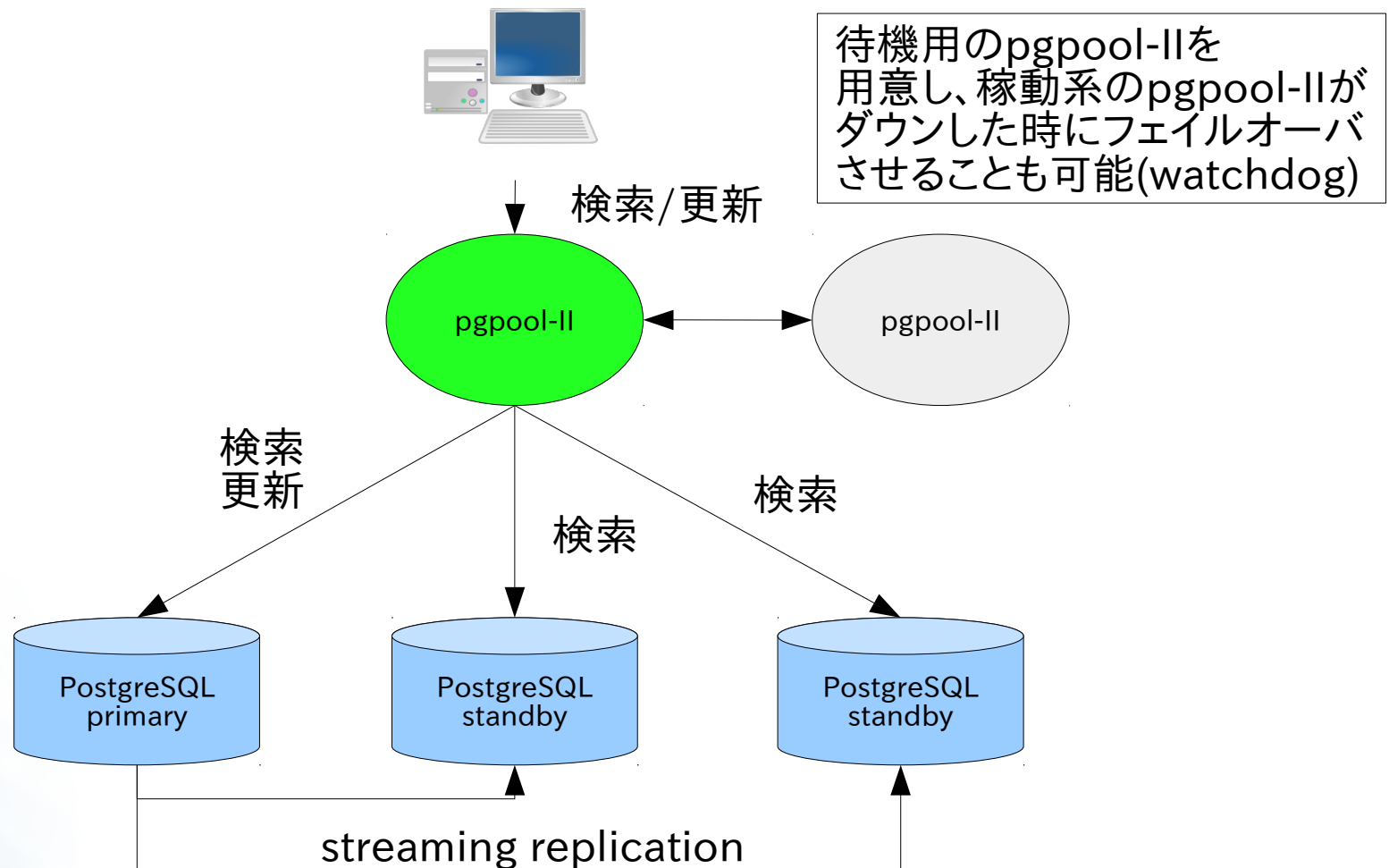
pgpool-IIによるクラスタの概念(3)



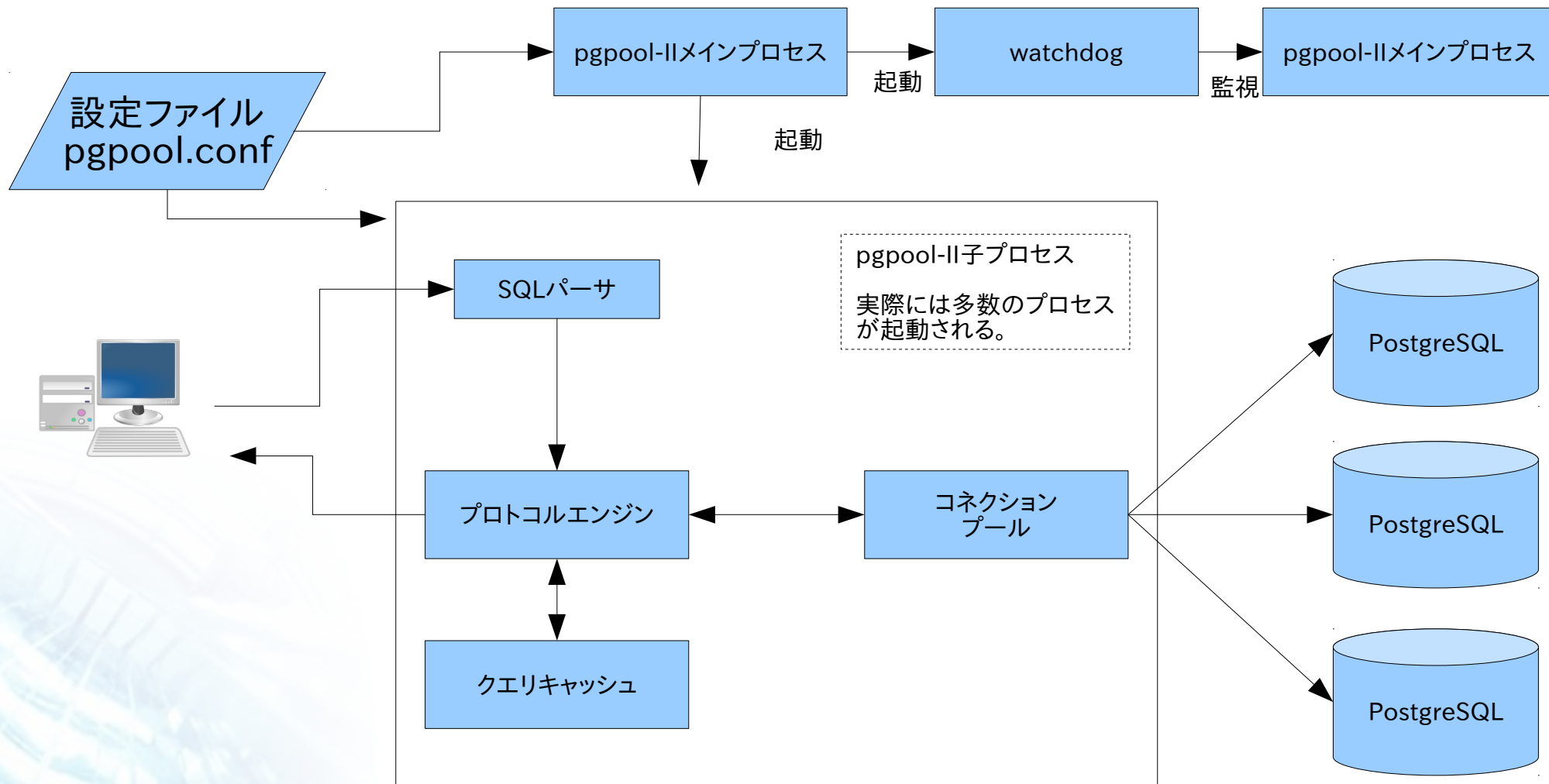
pgpool-IIによるクラスタの概念(4)



pgpool-IIによるクラスタの概念(5)



pgpool-IIの構造



pgpool-IIの必要とするリソース

- ディスクIOはほとんど発生しない
- ネットワーク性能が重要
- ユーザのセッションに対応して多数のプロセスが並行して動くので、CPU数が多い方が良いが、CPU負荷は軽いので、特別高速なCPUである必要はない
- 必要メモリ
 - 子プロセスあたり50MB位が目安

pgpool-IIのDB異常検技術

- 以下のどちらかまたは両方を使える
 - ヘルスチェック
 - 定期的にPostgreSQLにパケット(接続開始パケット)を投げて、応答が返るかどうかをチェック
 - 指定回数リトライしてそれでも駄目ならそのDBを切り離す(フェイルオーバー)
 - PostgreSQLとの通信の最中に通信エラーが起きたら、即座にDBを切り離すことも可能
 - 数秒から数十秒でフェイルオーバーできるが、敏感すぎるかも

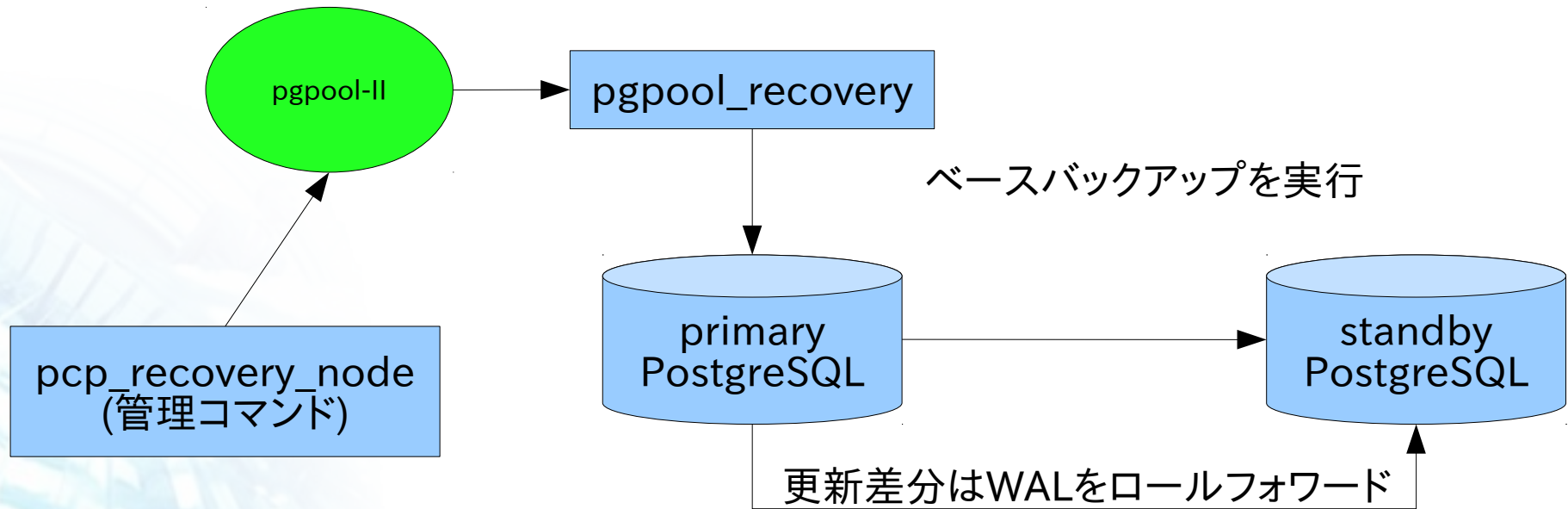


フェイルオーバー設定

- フェイルオーバーしたときに呼び出すスクリプトを設定可能
- 典型的な処理
 - primaryサーバがダウンした場合には、ユーザ好みのロジックで昇格候補のstandbyサーバを選び、昇格させる
 - フェイルオーバーした時に管理者にメールを送信することも可能

オンラインリカバリとは

- ダウンしたスタンバイDBあるいは新規に作成したスタンバイDBを、プライマリサーバに同期させる
- 同期中も更新可能
- 同期が完了したら自動的に管理対象に追加して負荷分散の対象となる

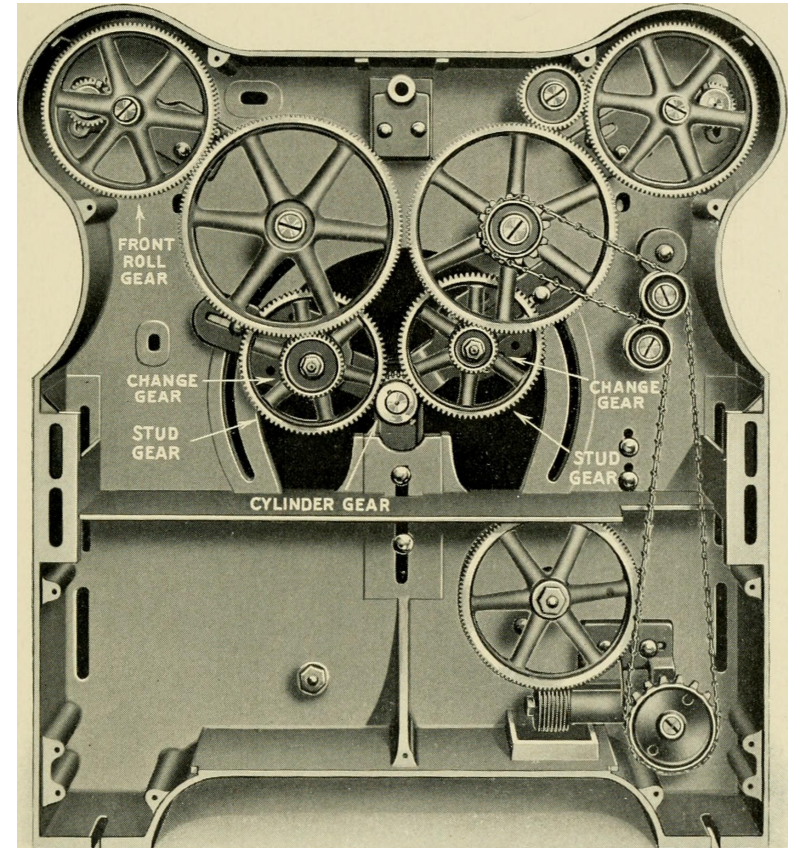


チューニング

- relcache_expire
 - pgpool-IIが持つテーブルのメタ情報のキャッシュ期限
 - 頻繁にテーブルを作成したりする場合は無期限ではなく短くする
- relcache_size
 - 上記キャッシュが保持できるテーブル数。テーブルが多い場合には増やす
- check_temp_table
 - テーブルが一時テーブルかどうかのチェックを行うかどうか。一時テーブルを使っていない場合はoffにすることにより、内部クエリの手数を減らせる
- memory_cache_enabled
 - 検索中心のシステムでは有効にすることにより性能向上
- 開発中の3.4では更にチューニング項目が増えます！

pgpool-IIの運用管理(1)

- pgpool-IIの起動
 - pgpool start
- pgpool-IIの停止
 - pgpool stop
- 設定ファイルの再読み込み
 - pgpool reload



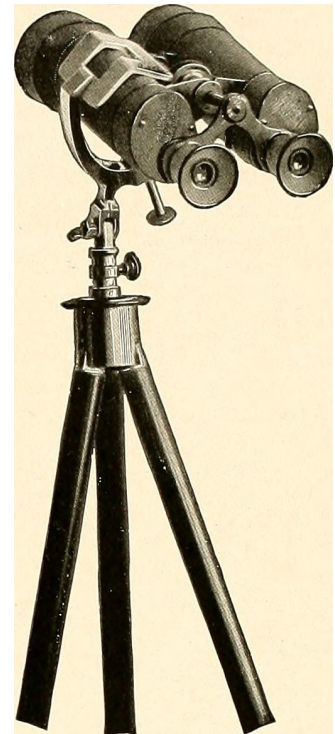
pgpool-IIの運用管理(2)

- DBノードの状態を表示

```
# show pool_nodes;
```

node_id	hostname	port	status	lb_weight	role
0	/tmp	11000	2	0.333333	primary
1	/tmp	11001	2	0.333333	standby
2	/tmp	11002	2	0.333333	standby

- pgpool-IIの管理下から外す
 - pcp_detach_node
- 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

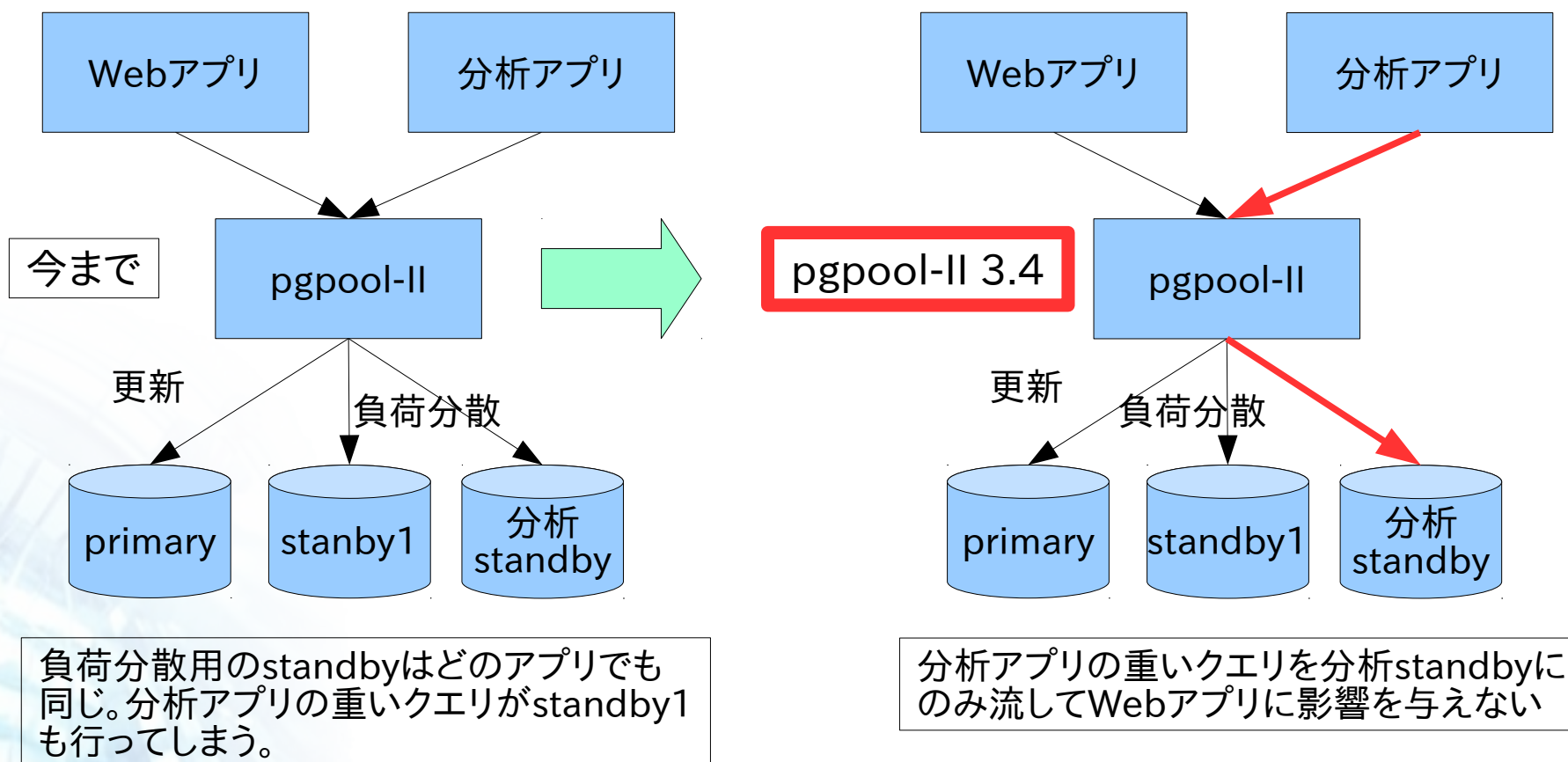


pgpool-IIの運用管理(3)

- 新しいstandbyノードを追加する
 - PostgreSQLのインストールと、initdbだけやっておく(起動はしない)
 - pgpool.confに新しいDBの情報を記述
 - “pgpool reload”でpgpool.confの再読み込み
 - DBノードがダウン状態でpgpool-IIに認識される
 - pcp_recovery_nodeを実行、DBの同期を取り、オンライン状態にする
- 新しくpgpool-IIに接続したセッションから新しいDBノードを利用できるようになる

開発中のpgpool-II 3.4.0の新機能!(1)

- 検索負荷分散のきめ細かな指定が可能になりました
 - アプリケーション名による指定、DB名による指定が可能



開発中のpgpool-II 3.4.0の新機能!(2)

- 具体的な指定の仕方
 - app_name_redirect_preference_list =
'psql:primary,myapp[0-4]:1,myapp[5-9]:standby'
 - database_redirect_preference_list =
'postgres:primary,mydb[0-4]:1,mydb[5-9]:2'
- アプリケーション名とデータベース名に正規表現が利用可能

開発中のpgpool-II 3.4.0の新機能!(3)

- 性能・信頼性に関する新しい設定項目
 - check_unlogged_table
 - unloggedテーブルを使用していない場合にチェックをスキップしてマスタサーバの負荷を軽減できる
 - listen_backlog_multiplier
 - 同時接続数が非常に多い場合に接続待ち行列を増やして性能と安定性を向上可能
 - connect_timeout
 - コネクション時のタイムアウトを調整して不安定、高負荷環境でネットワーク接続エラーによるフェイルオーバを低減することが可能
 - allow_sql_comments
 - 負荷分散に影響を与えずにSQLコメント(`/* comment */`)が使用可能

開発中のpgpool-II 3.4.0の新機能!(4)

- ログに関する機能向上
 - print_user
 - ログに接続ユーザ名を印字
 - ログレベルの導入
 - PostgreSQL同様debug[1-5],log,notice, errorなどのログレベルが設定できる
- その他
 - PostgreSQL 9.4対応
 - たとえば“REFRESH MATERIALIZED VIEW CONCURRENTLY”を正しく認識できる
 - IPv6対応
 - PostgreSQLサーバのアドレス、pgpool-IIのバインドアドレスにIPv6を使用可能
 - PostgreSQLのメモリーマネージャ、例外処理マネージャを移植
 - 信頼性、性能の向上
 - ソースツリーの改良
 - 可読性のアップ
 - PostgreSQLの状態を記録するステータスファイルがテキストファイルになり、可読性が向上

まとめ

- 可用性とは
- PostgreSQLで使えるクラスタ技術
- pgpool-IIの紹介



参考URL

- SRA OSSのサイト(多数のスライドや技術情報あり)
 - <http://www.sraoss.co.jp>
- pgpool-IIオフィシャルサイト
 - <http://www.pgpool.net>
 - http://www.pgpool.net/pgpool-web/contrib_docs/simple_sr_setting2_3.3/index-ja.html
- PostgreSQLに関する各種技術情報
 - <http://lets.postgresql.jp>

オープンソースとともに



SRA OSS, INC.