PostgreSQLの大規模運用を可能にするクラスタ技術のご紹介

SRA OSS, Inc. 日本支社
取締役支社長
石井　達夫

SRA OSS,INC.

自己紹介

- OSSの開発とビジネスに携わっています
  - PostgreSQLのコミッタ
  - PostgreSQL用のクラスタソフトpgpool-II の開発
- OSSの普及活動も行っています
  - オープンソースビジネス推進協議会
    - OSS全般の普及促進活動
  - PostgreSQLエンタープライズコンソシアム
    - PostgreSQLの技術的評価や普及促進活動

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, Hinemos, ZabbixなどのOSSのサポート、コンサルティング、導入構築
  - Postgres Plusの販売
  - PowerGresファミリーの開発、販売
  - PostgreSQL用の各種トレーニング

PowerGres

Hinemos

ZABBIX

CERTIFIED PARTNER

EDB ENTERPRISEDB

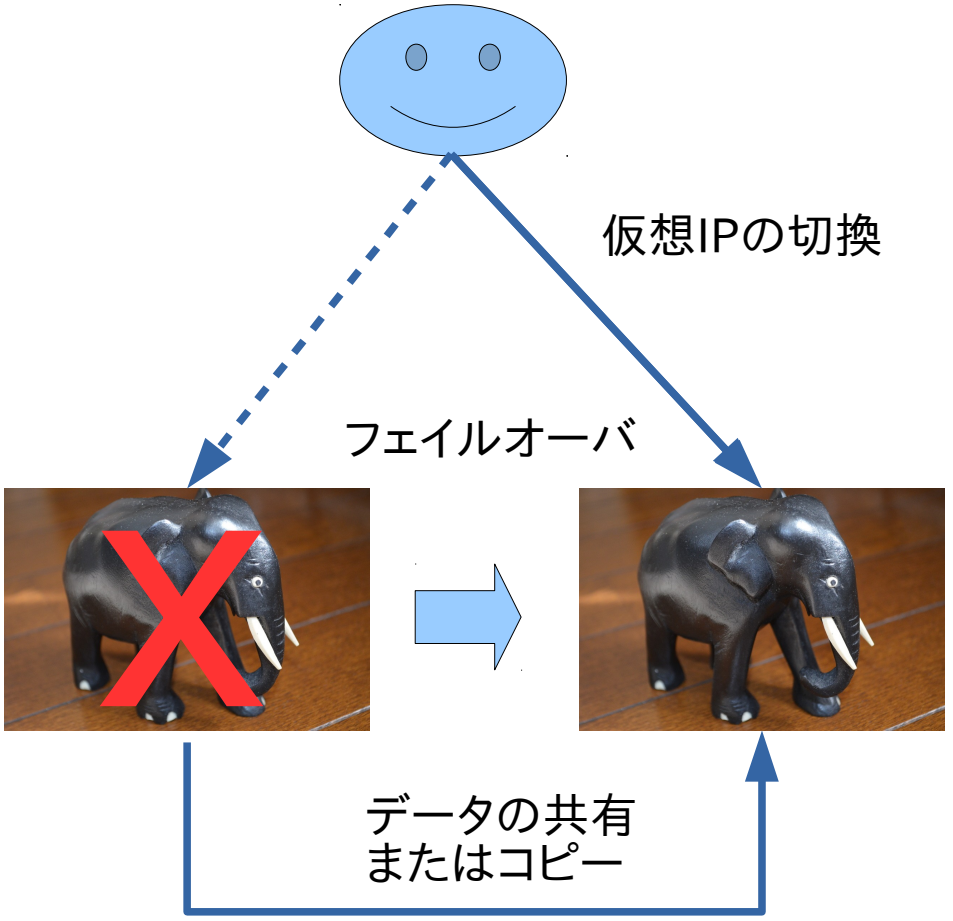
SRA OSS,INC.

PostgreSQLにおける クラスタリングの目的

- 可用性の向上(High Availability)
  - もっとも古典的かつ基本的なリクエスト
    - ダウンタイムの短縮
    - データ損失可能性の低減
- 性能の向上(Scale out)
  - 検索性能(read)の向上
    - ストリーミングレプリケーションによる検索負荷分散
    - postgres_fdwによる分散処理
    - Postgres-XC/XL/X2
  - 更新性能(write)の向上
    - Postgres-XC/XL/X2

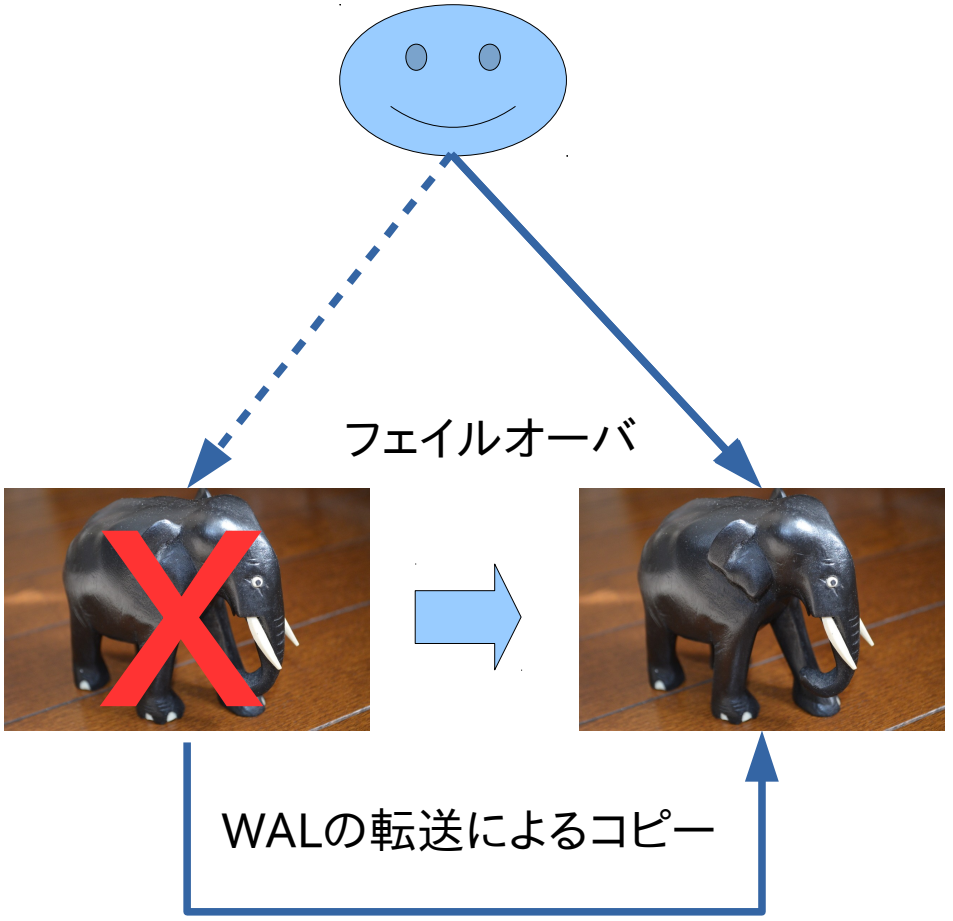
可用性向上のためのクラスタ技術：アクティブ・スタンバイ方式

- Pacemakerなどの汎用HAソフトを使って複数のPostgreSQLを管理
- 稼動系がダウンしたら待機系に切り替える
- DBの入ったディスク装置を共有する共有ディスク方式と、共有しない方式がある
- 性能は向上しない
- アプリケーションの修正は必要ない
  - 通常フェイルオーバ中はセッションが切断されるのでその対応は必要



可用性向上のためのクラスタ技術：ストリーミングレプリケーション(1)

- PostgreSQL組み込みのレプリケーション技術である「ストリーミングレプリケーション」を利用
- プライマリからスタンバイへトランザクションログを転送してデータをコピーする
- 書き込み性能は向上しないが、複数のPostgreSQLで読み出し負荷を共有して性能を向上させることも可能
- アプリケーションの修正は必要
  - フェイルオーバ中はセッションが切断されるのでその対応は必要
  - 書き込み処理はプライマリにしか投げてはいけない（そのほか、ある種のロックはプライマリだけに投げるなどの考慮が必要）
- これらの対応が困難な場合には、すべてのDB処理をプライマリにのみ投げる（その場合でもフェイルオーバへの対応は必要）



可用性向上のためのクラスタ技術：ストリーミングレプリケーション(2)

- プライマリが故障した時に、どのスタンバイを昇格させるか決めておく必要がある(スタンバイの数が2以上の場合)
  - 同期レプリケーションを使っている場合は、同期スタンバイを昇格させる
  - 非同期レプリケーションを使っている場合は、固定ルールで昇格するスタンバイを決めておく方式と、動的に決める(最も遅延が少ないものを選ぶ、など)方式がある

プライマリ

スタンバイ1

スタンバイ2

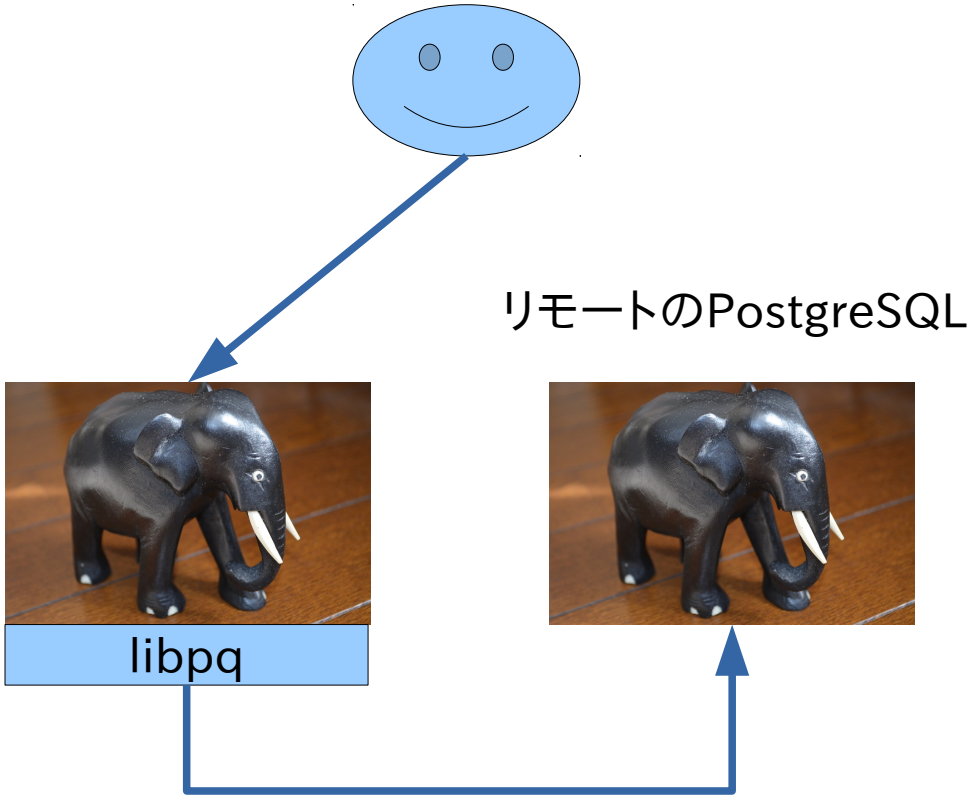
スタンバイ3

性能向上のためのクラスタ技術：ストリーミングレプリケーション(3)

- スタンバイサーバを複数設けて検索性能を向上させる（負荷分散）
  - 一つのSQLが分散処理されるわけではないので、多数のセッションが同時に実行されるような環境で効果が上がる
  - プライマリサーバが過負荷のときに、プライマリに検索処理をさせないようにするのも効果がある
  - どのセッションがどのスタンバイサーバに接続するかを決める必要がある
    - アプリケーションで行う（アプリケーションの改造が必要）
    - ミドルウェアで透過的に実施(Pgpool-IIなど)

可用性向上のためのクラスタ技術：postgres_fdw(1)

- あるPostgreSQLから別のPostgreSQLに接続、クエリを実行する拡張機能
- 外部データラッパのフレームワークを使っている
- “CREATE FOREIGN TABLE”文を使って、リモートのPostgreSQLにあるテーブルをあたかもローカルのテーブルであるかのように扱うことができる



性能向上のためのクラスタ技術：postgres_fdw(2)

- 外部テーブルへのSELET, UPDATE, DELETE, INSERTが可能
  - 外部テーブルへのアクセスは、常にREPEATABE READ分離レベルのトランザクションの内側で実行されるので、ローカルのトランザクションもREPEATABE READもしくはSERIALIZABLEで実行するのが良い
- 大きなテーブルを複数のサーバに分けて性能向上を狙う、などの用途が考えられる
  - PostgreSQL 9.5からは、継承を使ったパーティショニングに外部テーブルが利用可能になった
  - ただし、外部テーブルへのアクセスが並列に行われるわけではない

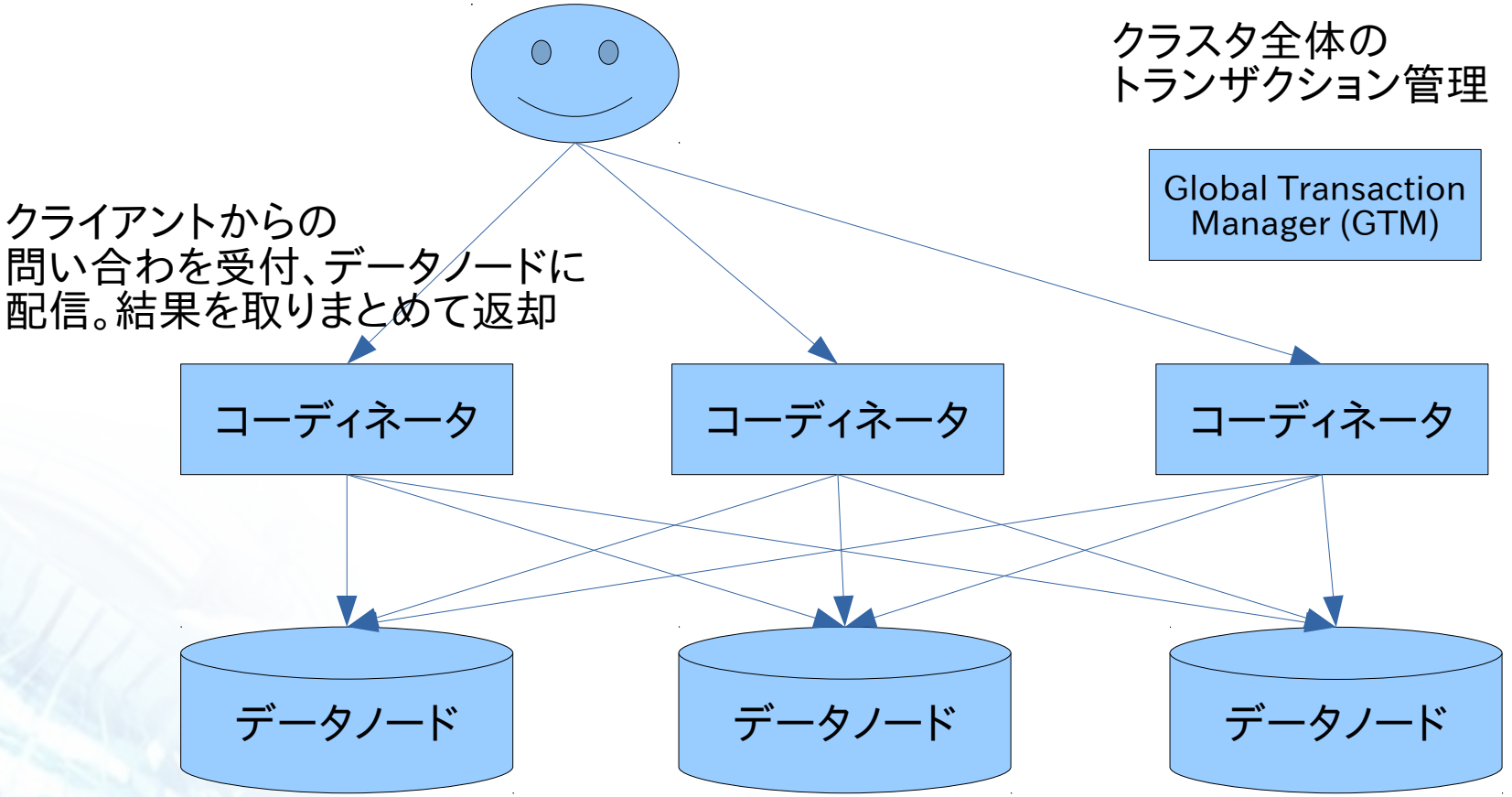
性能向上のためのクラスタ技術：postgres_fdw(3)

- 外部テーブルへのアクセスの最適化は、通常のオプティマイザのへpostgres_fdwが提供するフックが組み込まれることで行われる
  - WHERE/JOIN push down
    - (外部テーブルの結合の際に、)WHERE句を外部サーバに送って実行し、その結果だけを転送することによってデータ量を減らす
  - 処理に必要なカラムだけを転送することによってデータ量を減らす
  - 集約(countやsumなど)は最適化されない
- 外部テーブルの統計情報の取得方法を選択可能
  - 都度EXPLAINを外部テーブルで実施する
  - ANALYZEした結果をローカルに保存しておく

性能向上のためのクラスタ技術：postgres_fdw(4)

- 現状はまだ分散トランザクション機能を提供しているに過ぎず、性能向上のための工夫はユーザに任されている
- 外部テーブルの今後
  - 外部テーブルを使って性能向上のための本格的なクラスタを作ろうと考えている開発者がいる
  - 多くの機能拡張が必要になる
    - 集約関数のリモート実行
    - 並列処理
    - 複数サーバにまたがる読み取り一貫性の実現
    - 並列処理用のオプティマイザ
      - 現状は一部の最適化がpostgres_fdwの中で行われているに過ぎない

性能向上のためのクラスタ技術：Postgres-XC(1)



実際の問い合わせ処理を実行

性能向上のためのクラスタ技術：Postgres-XC(2)

- データノード間で並列に問い合わせが実行可能
- データをデータノードに分散（シャーディング）可能なので、検索性能のみならず、更新性能も向上
- 更新遅延がなく、ノード間の読み取り一貫性も保証されている
- 可用性は向上しない(むしろ低下する)
  - データノードのレプリケーションで対応

性能向上のためのクラスタ技術：Postgres-XC(3)

- PostgreSQLからのフォークであり、PostgreSQLへ追従するのが大変(2016/8現在、PostgreSQL 9.3相当)
- 一部の問い合わせは処理できない
- なかなかPostgres-XCが安定しなかったため、更にフォークを産んだ
  - Postgres-XL
    - 2016/8現在、PostgreSQL 9.5相当
- 最近Postgres-XCはPostgres-X2と名前を変えた
- Postgres-XLなどとの統合の動きもある

各種クラスタ技術のまとめ

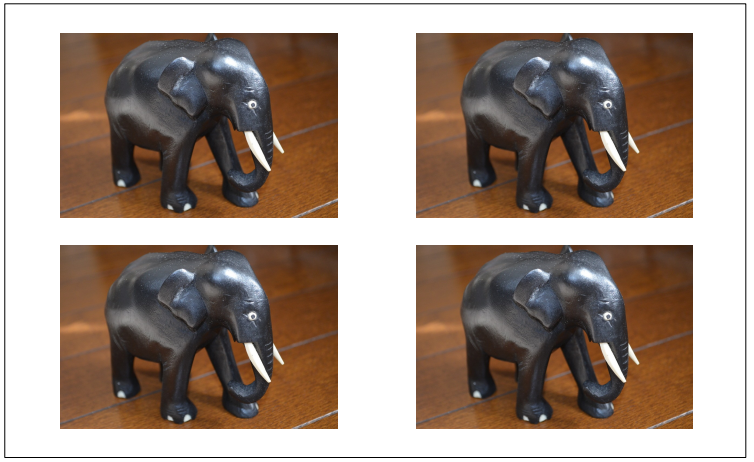
	可用性向上	検索性能向上	更新性能向上	アプリケーション変更度合い	PostgreSQL変更必要	実用性・実績
Pacemaker	○	X	X	○	○	○
Streaming replication	○	○	X	X	○	○
Postgres_FDW	X	▲	▲	○	○	○
Postgres-XC	X	▲	○	X	X	X

○:寄与する
X:寄与しない
▲:制限事項あり／工夫すれば寄与できる

Pgpool-IIのご紹介と使いどころ

PostgreSQLの群れ

- 象はパワフル
- 象が群れになればもっとパワフル!
- でも群れになれば管理が大変なのでは?



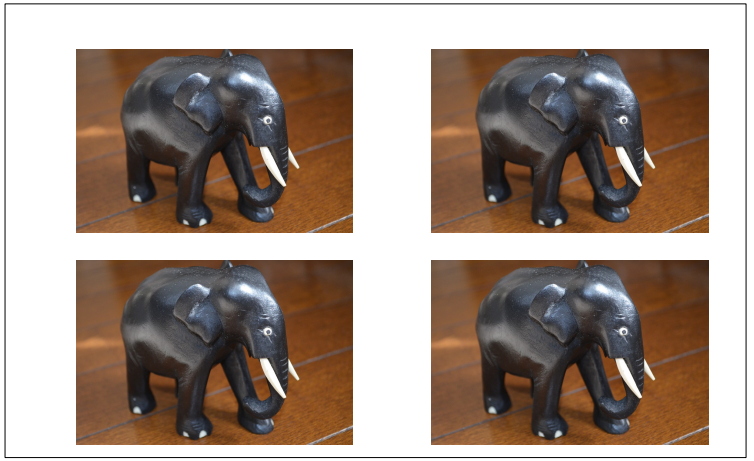
「象の群れ管理問題」の一例

- 群れにはリーダが必要です(プライマリサーバ)
- もしリーダが引退したら、新しいリーダを立てなければならない(フェイルオーバ、昇格)
- その際、他の象は新しいリーダに追従しなければならない
- リーダ以外の象は、働けなくなったら引退する(スタンバイのフェイルオーバ)
- 新しい象が群れに加わるときはスムーズに行われなければならない|
- 群れの象はお互いに助けあわなければならない (負荷分散)
- リーダにしかできない仕事がある (更新クエリ)

Pgpool-IIで「象の群れ管理問題」を解決

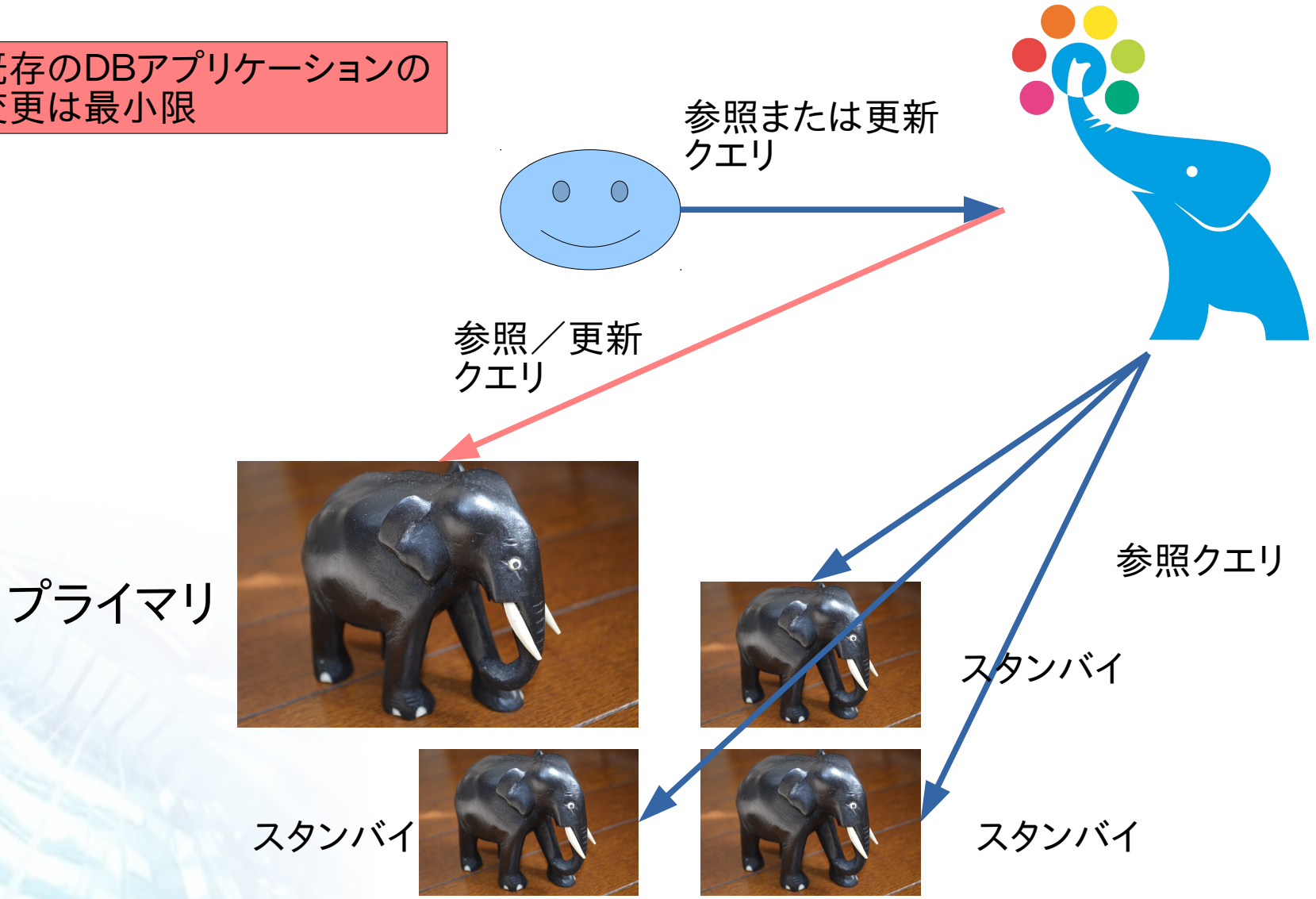
- Pgpool-IIを使うことにより、象の群れは単独の象のように見える
- ユーザ定義のフェイルオーバスクリプトにより、フェイルオーバ時にどのスタンバイが昇格するかのポリシーを決められる
- 「フォローマスターコマンド」で新しいプライマリへの自動追従も可能
- スタンバイがダウンしたら、自動的に群れからそのスタンバイは外されるので、クラスタとしての運用を継続できる
- クエリが検索クエリなら、負荷分散の対象となる
- クエリが更新クエリなら、プライマリサーバに送るI

象の群れは単独の象のように見える

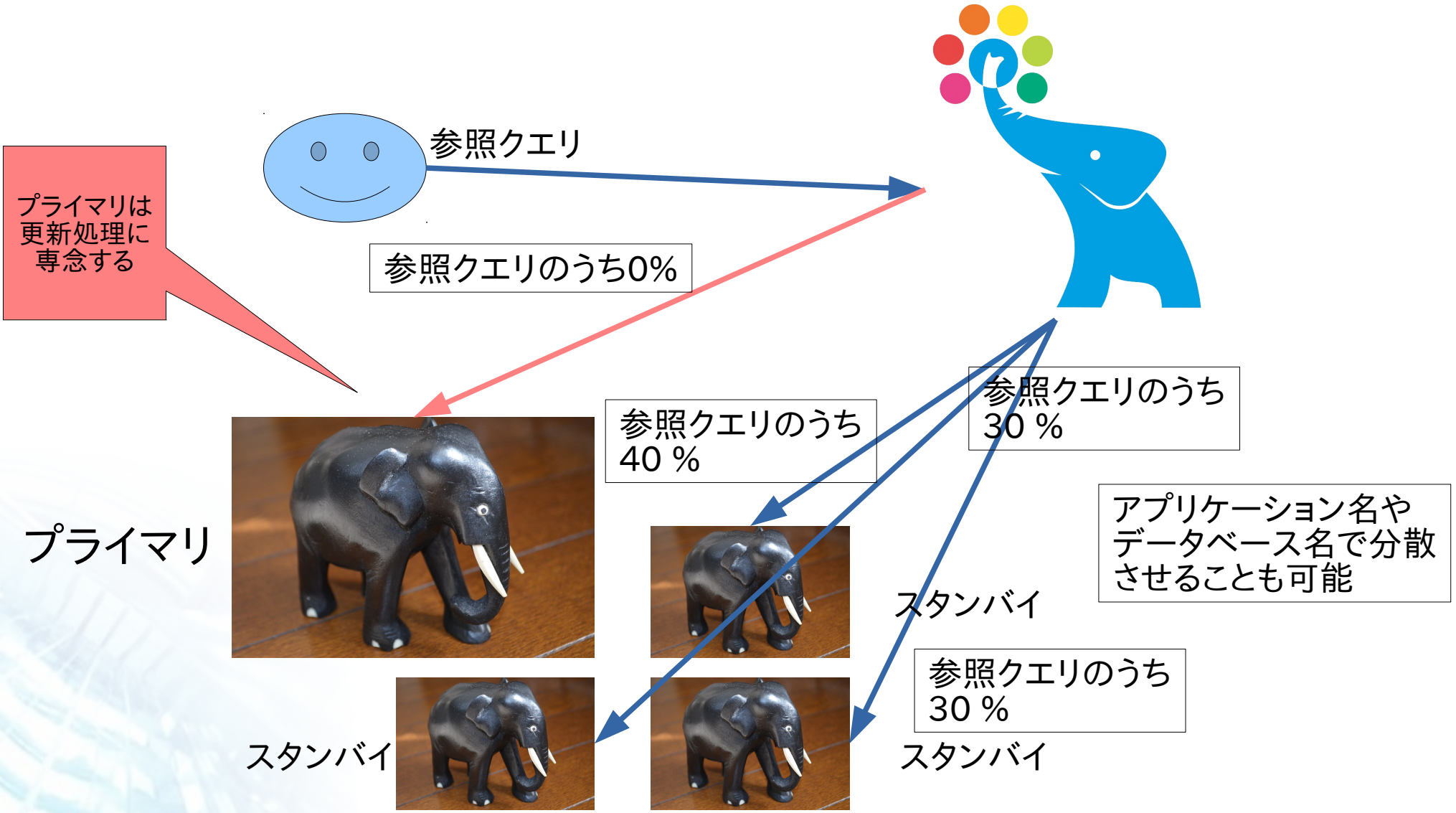


クエリのディスパッチ/ルーティング

既存のDBアプリケーションの
変更は最小限



負荷分散



スタンバイサーバがダウンした時

スタンバイサーバがダウンしたら、
Pgpool-IIがそのことを検知し、
クラスタリングの対象から取り除く

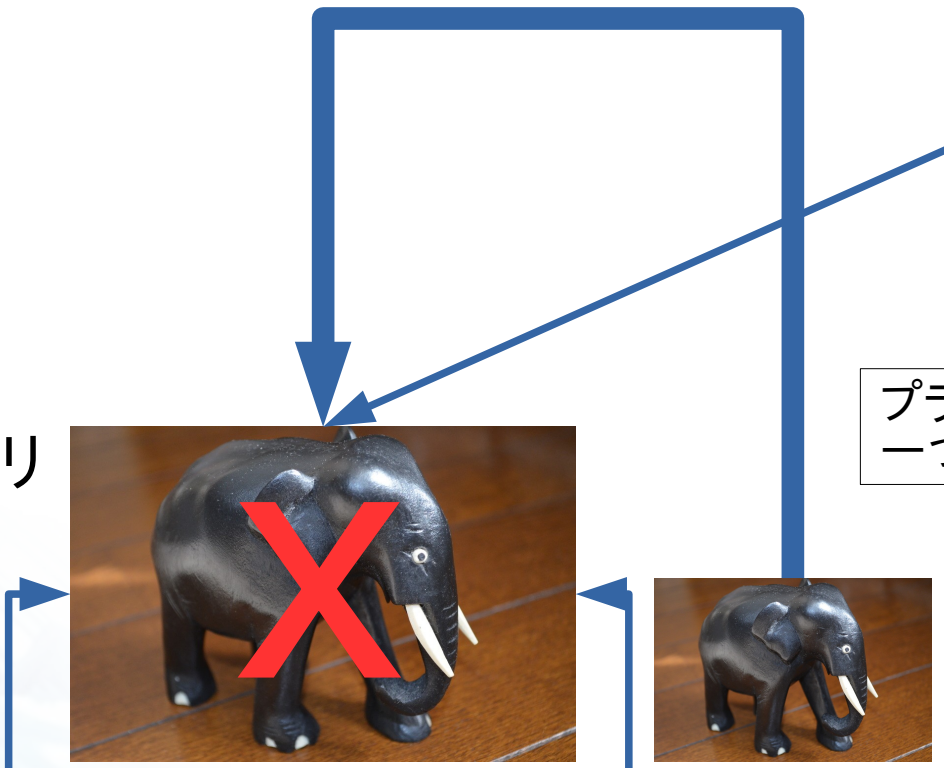
既存のセッションは再接続が
必要

プライマリ

スタンバイ

SRA OSS, INC.

プライマリサーバがダウンした時



プライマリがダウンすると、スタンバイの一つが昇格し、新しいプライマリになる

他のスタンバイは新しいスタンバイに追従するようになる

新しいスタンバイの追加

新しい象！

プライマリ

スタンバイ

新しいサーバは簡単に追加できる。
Pgpool-IIは新しいサーバにプライマリからデータをコピーし、他のサーバに影響を与えずにクラスタに新しいサーバを追加できる。
既存のセッションは切断されない。

Watchdog: Pgpool-II組み込みの高可用性機能

アクティブPgpool

昇格

スタンバイPgpool

プライマリ

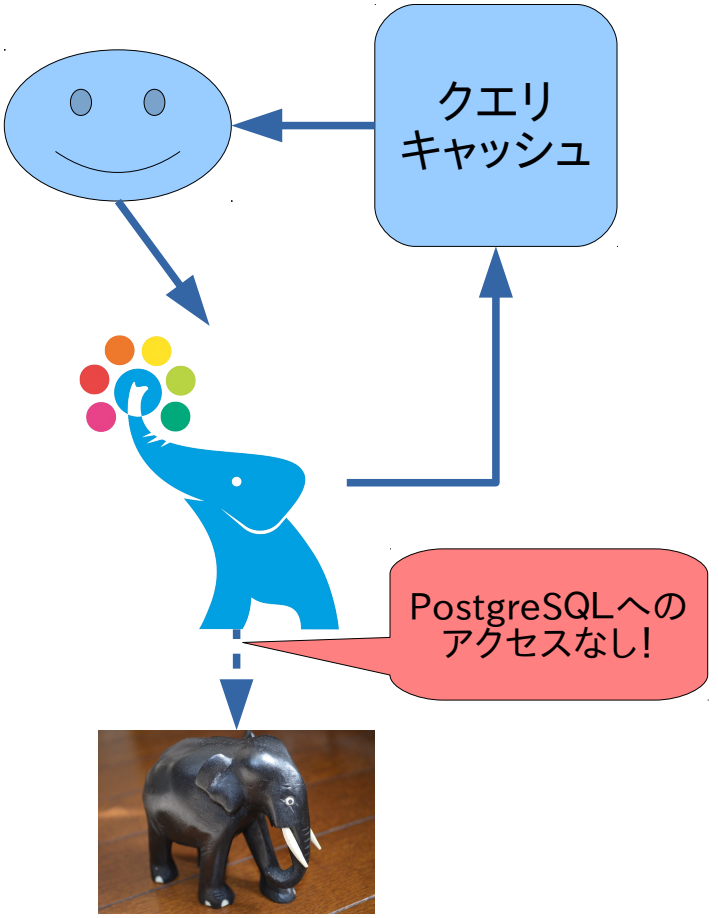
スタンバイ

アクティブPgpool-IIがダウンすると,
スタンバイPgpool-IIが昇格する

SRA OSS, INC.

インメモリクエリキャッシュ

- Pgpool-IIはクエリキャッシュを使ってクエリの結果を再利用する
- クエリキャッシュはメモリ上に置かれるので非常に高速
- その際にPostgreSQLアクセスは一切なし
- キャッシュ用のストレージは、共有メモリ火memcachedから選べる
- テーブルが更新されると、そのテーブルを参照したクエリキャッシュはすべて廃棄される
- タイムアウトベースのキャッシュ更新も可能



コミュニティサポートポリシー

- PostgreSQLと同様メジャーリリースとマイナーリリースがある
  - 3.x.y – x:vメジャーバージョン, y: マイナーバージョン
    - “3.5.3”: “3.5” がメジャーバージョンで, “3”がマイナーバージョン
  - 年に一度のメジャーバージョン
  - 年に3-4回のマイナーリリース
  - マイナーリリース間では互換性が保たれる
  - メジャーバージョン間では互換性は保証されない
- 最初のリリースから5年間バックパッチ(保守)を行う
  - つまり、常に5-6個のバージョンを保守している
- 5年以上のサポートが必要ならば、弊社にご相談ください

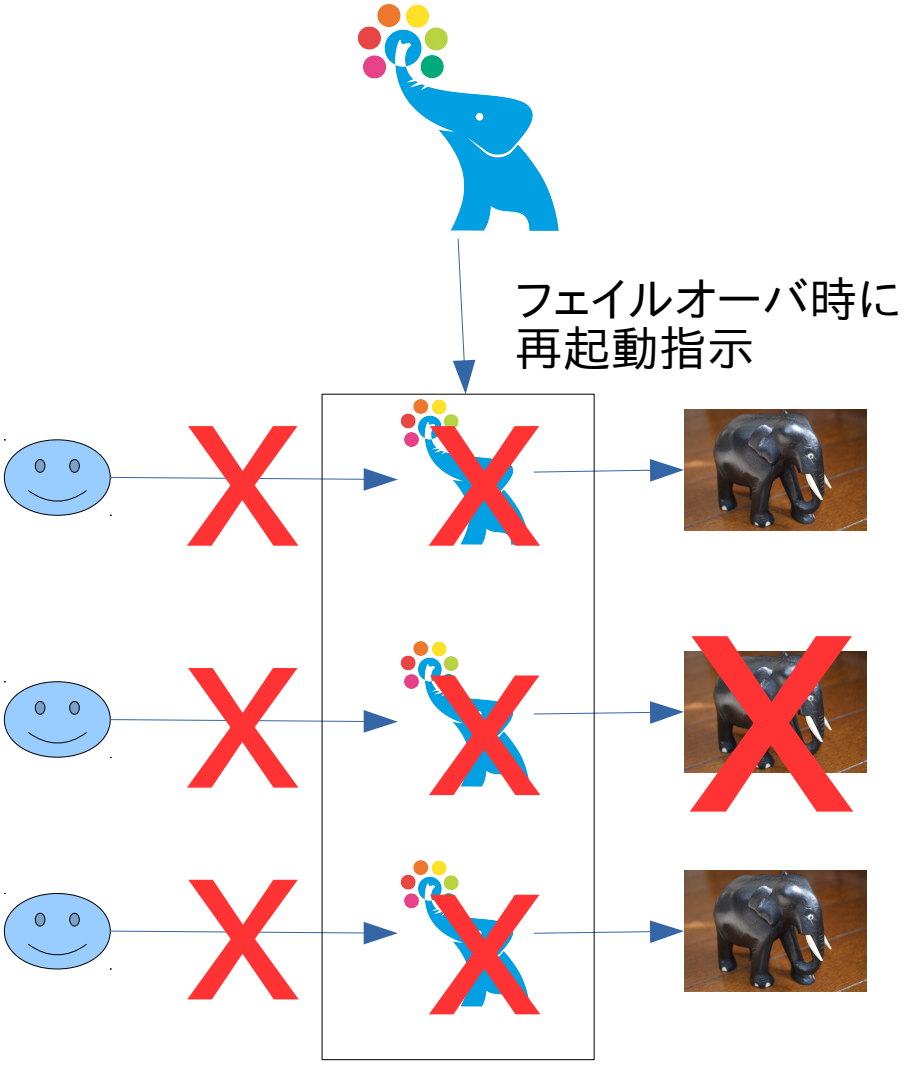
最新バージョン：Pgpool-II 3.6 が11月にリリースされます！

- フェイルオーバの改善
- PostgreSQL 9.6対応
- その他

フェイルオーバの改善

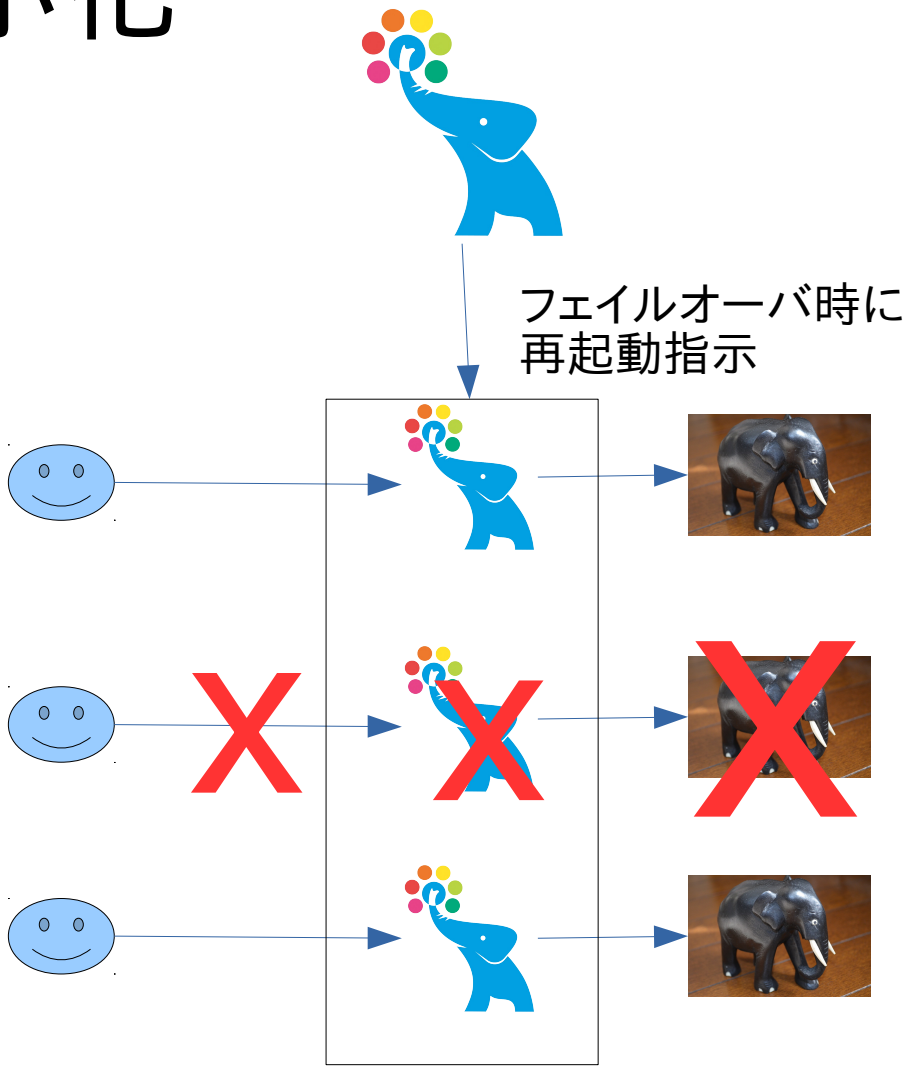
フェイルオーバにおける問題点

- 複数のPostgreSQLサーバを使ったクラスタ構成で、そのセッションが使用していないDBサーバがダウンした時にもセッションが一旦切断される



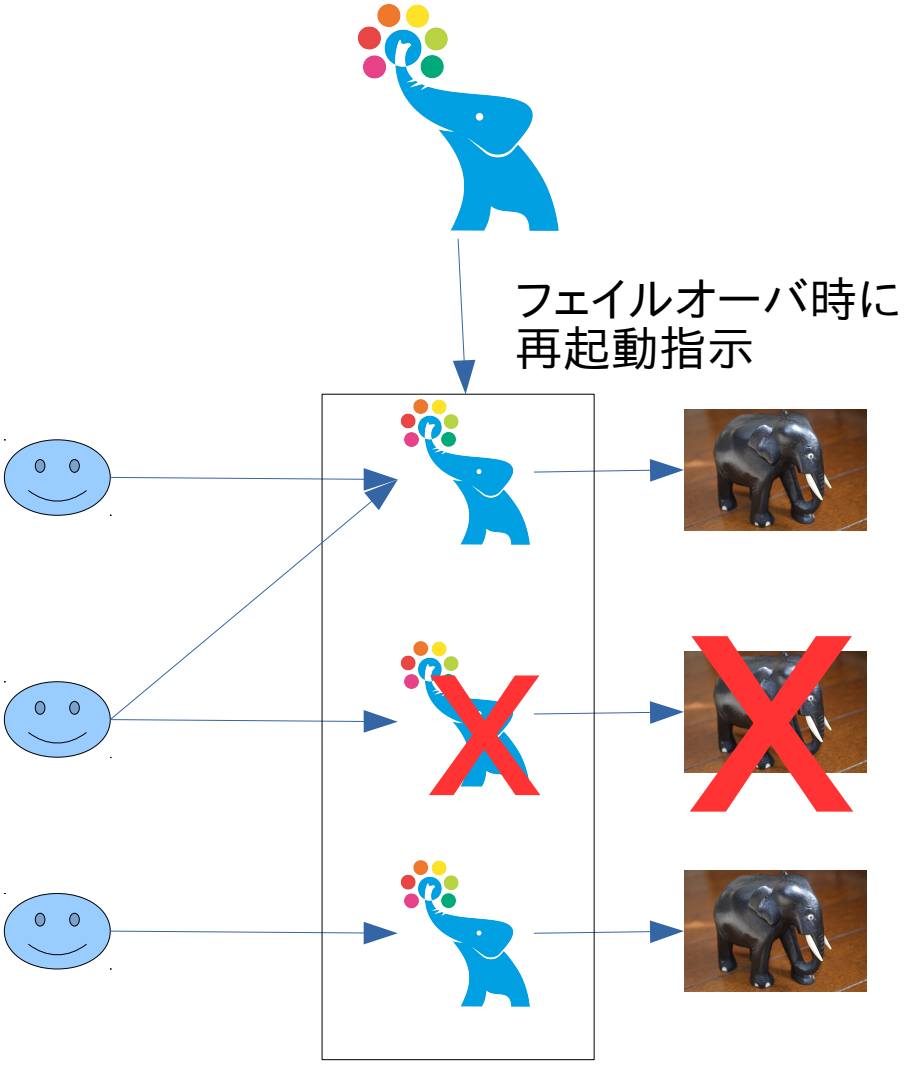
Pgpool-II 3.6では切断されるセッションを最小化

- 前提条件：ストリーミングレプリケーションモード、ダウンしたDBサーバはスタンバイ
- ダウンしたサーバを使っているセッションに関連したPgpool-II子プロセスのみ再起動指示
- それ以外のセッションに関連したPgpool-II子プロセスは、現在のセッションが終了後、自動的に再起動してダウン情報を反映する



PostgreSQLの計画停止が容易に

- あらかじめ、停止予定のDBサーバに対する負荷分散の重みを0に設定し、設定ファイルを再ロード
- 新しいセッションでは停止予定のDBを使わなくなる
- DBを停止しても影響を受けるセッションなし



使用しているDBサーバの確認がセッションごとに可能に

- 使用しているDB（負荷分散ノード）はセッションごとのプロパティなので、PCPコマンドなどでは確認ができない
- そこで、show pool_nodes で、現在のセッションの負荷分散ノードを確認可能に
- そのほか、レプリケーション遅延も確認可能に

```
psql -p 11000 -c "show pool_nodes" test
 node_id | hostname | port  | status | lb_weight |  role   | select_cnt | load_balance_node | replication_delay
---------+----------+-------+--------+-----------+---------+------------+-------------------+-------------------
 0       | /tmp     | 11002 | up     | 0.333333  | primary | 0          | false             | 0
 1       | /tmp     | 11003 | up     | 0.333333  | standby | 0          | true              | 0
 2       | /tmp     | 11004 | up     | 0.333333  | standby | 0          | false             | 0
(3 rows)
```

PostgreSQL 9.6対応

PostgreSQL 9.6 SQLパーサの移植

- Pgpool-IIでは、SQL文を正確に解析するためにSQLパーサを持っている
  - SQLパーサは、最新のPostgreSQLから移植
- 以下の新しい構文をサポート
  - COPY FROM INSERT … RETURNING TO …
  - ALTER FUNCTION … DEPENDS ON EXTENSION …
  - ALTER TABLE ADD COLUMN IF NOT EXISTS …
  - など

Pgpool-II 3.6その他の改良

PGPOOL SET

- PostgreSQLの“SET”コマンドに相当するもの
- Pgpool-IIの一部の設定変数をセッション内で変更可能
  - セッション終了後に自動リセットされる
  - 構文は、“PGPOOL SET 変数名 TO 値”
- SETできる変数
  - client_idle_limit
  - client_idle_limit_in_recovery
  - log_statement
  - log_per_node_statement
  - log_min_messages
  - client_min_messages
  - log_error_verbosity
  - allow_sql_comments
  - check_temp_table
  - check_unlogged_table

PGPOOL SHOW

- Pgpool-IIの設定変数を個別に表示
  - PGPOOL SHOW 変数名
- 「変数グループ」(“logical group”)別の表示も可能
  - “backend”, “other_pgpool”(他のwatchdogノード)、“heartbeat”

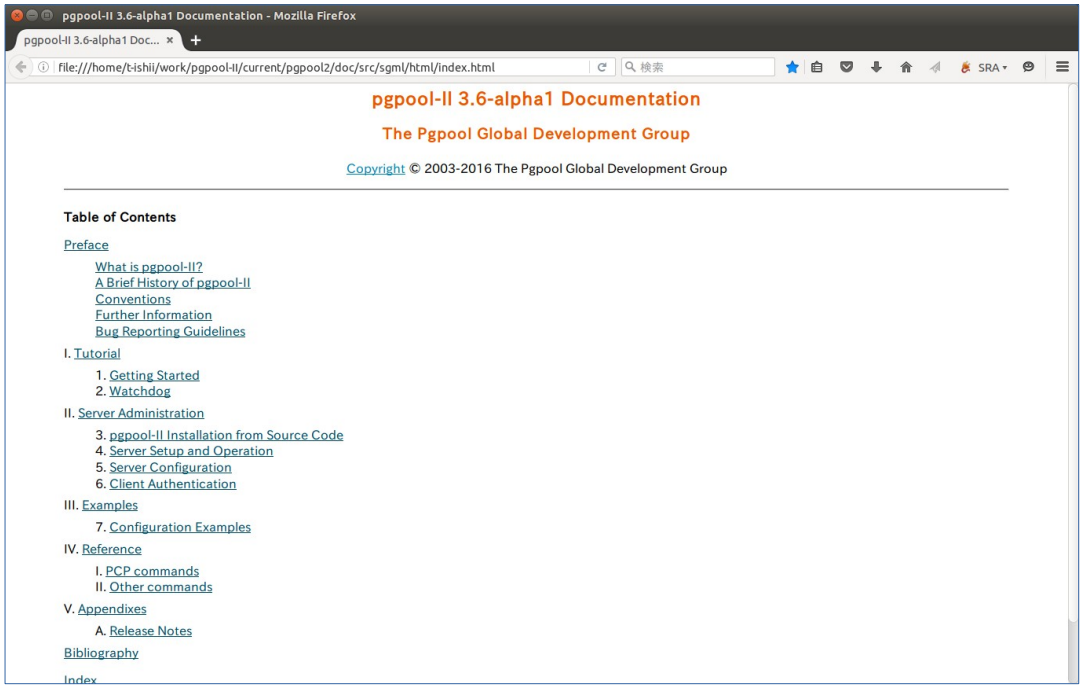
```
postgres=# pgpool show backend;
          item           |              value              |                   description
-------------------------+---------------------------------+-----------------------------------------------
 backend_hostname0       | 127.0.0.1                       | hostname or IP address of PostgreSQL backend.
 backend_port0           | 5434                            | port number of PostgreSQL backend.
 backend_weight0         | 0                               | load balance weight of backend.
 backend_data_directory0 | /var/lib/pgsql/data             | data directory of the backend.
 backend_flag0           | ALLOW_TO_FAILOVER               | Controls various backend behavior.
 backend_hostname1       | 192.168.0.1                     | hostname or IP address of PostgreSQL backend.
 backend_port1           | 5432                            | port number of PostgreSQL backend.
 backend_weight1         | 1                               | load balance weight of backend.
 backend_data_directory1 | /home/usama/work/installed/pg   | data directory of the backend.
 backend_flag1           | ALLOW_TO_FAILOVER               | Controls various backend behavior.
(10 rows)
```

pg_terminate_backend対応

- pg_terminate_backendとは
  - PostgreSQLの組み込み関数の一つ。プロセスIDを指定して、特定のバックエンドプロセスを終了することができる。主に、無限ループに入ってしまった、ロックを掴みっぱなしになっている、などの状態になったバックエンド終了させるための機能
- Pgpool-IIにとっての問題点
  - DBのシャットダウンと同じエラーコードがpg_terminate_backendの実行でもPostgreSQLから返るため、Pgpool-IIはDBがシャットダウンされたと解釈してフェイルオーバしてしまう
- 解決策
  - pg_terminate_backendの引数を調べ、指定プロセスIDがどれかのセッションで使われているバックエンドである場合には、該当エラーコードがPostgreSQLから返ってきたとしても、フェイルオーバを起こさず、該当セッションを切断するだけにする
- 制限事項
  - pg_terminate_backendの引数は単純整数でなければならない
    - SELECT pg_terminate_backend(pid) from strange_table; とかは駄目

ドキュメントフォーマットの変更

- 今までは、手打ちのHTMLを使用
  - メインテナンスが大変
  - 索引や目次などを作れない
- Pgpool-II 3.6では、PostgreSQLと同様、SGML → HTMLという仕掛けを採用
- 最初はSGMLファイルを作るのが大変だが、保守は楽になるはず（と信じてます）
- ボランティア募集中！



今後の予定

- Pgpool-II 3.6を今年のリリースは11月頃の予定
  - http://pgpool.net/mediawiki/index.php/pgpool-II_3.6_development
    にて逐次開発状況を公開中
  - 9月終わりくらいにベータリリースの予定なので、是非ベータテストにご協力を!

各種クラスタ技術のまとめ

	可用性向上	検索性能向上	更新性能向上	アプリケーション変更度合い	PostgreSQL変更必要	実用性・実績
Pacemaker	○	X	X	○	○	○
Streaming replication	○	○	X	X	○	○
Postgres_FDW	X	▲	▲	○	○	○
Postgres-XC	X	▲	○	X	X	X
Pgpool-II	○	○	X	○	○	○

○:寄与する
X:寄与しない
▲:制限事項あり／工夫すれば寄与できる

URLなど

- pgpool-II公式サイト
  - http://www.pgpool.net
- Postgres-X2
  - https://github.com/postgres-x2
- SRA OSS
  - http://www.sraoss.co.jp

宣伝

- 第1回 Pgpool-II 内部情報勉強会のお知らせ
  - 2016年9月7日(水) 池袋SRAグループビル9F会議室にて、15:00-17:00
  - 参加費は無料
- 以下のような方におすすめ
  - Pgpool-IIをすでに使っており、更なる使いこなしのため内部構造に関心がある
  - Pgpool-IIの開発に関心があり、内部構造を知っておきたい
  - 申し込みはこちらを参照
    - http://www.pgpool.net/pipermail/pgpool-general-jp/2016-August/001412.html

Thank you!