SRA OSS,INC.

PostgreSQLで動的にスケールアウト可能な負荷分散DBクラスタを作ろう!

db tech showcase Tokyo 2015
2015/6/10

SRA OSS, Inc. 日本支社
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

会社紹介

- 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

- 動的にスケールアウト可能な負荷分散DBクラスタ
  「サーバを複数台使って高い参照性能を得る」
  「サーバの追加はサービスを停止せずにできる」

PostgreSQL 組み込みのレプリケーション機能

＋

クラスタ管理ツール pgpool-II の
負荷分散 & 高可用化 機能

SRA OSS, INC.

アジェンダ

- データベースのクラスタリング
- PostgreSQL のレプリケーション機能
- pgpool-II のクラスタリング機能
- PostgreSQL と pgpool-II によるシステム構成
- スケールアウト性能
- デモ

SRA OSS, INC.

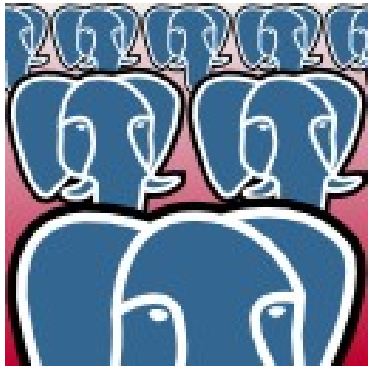
PostgreSQL とは

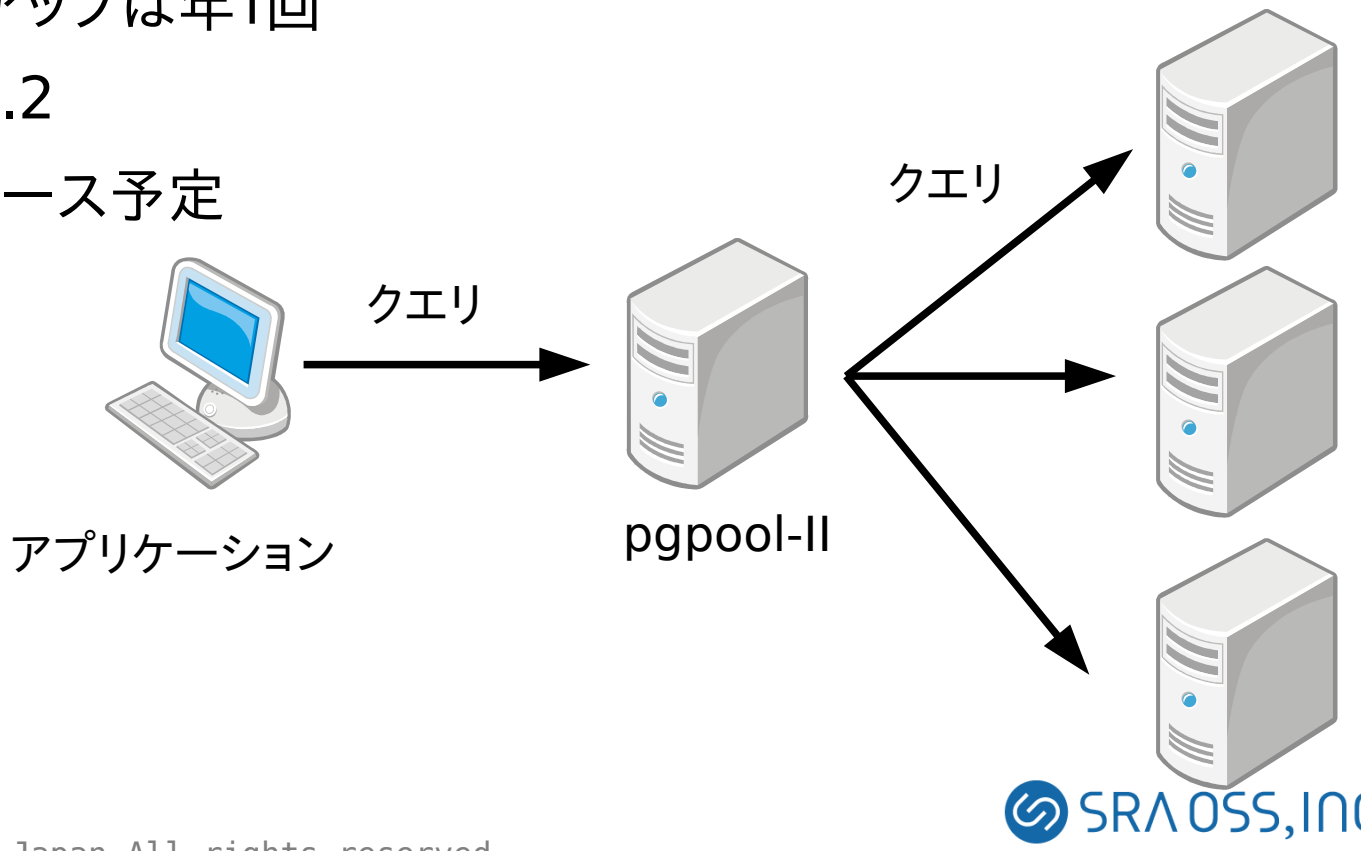
- 代表的なオープンソースのRDBMSの1つ
  - カリフォルニア大学で開発された研究用RDBMSのIngres(1970)を先祖に持つ
- オーナー企業を持たず、コミュニティによる開発が続けられている
  - 年1回のメジャーバージョンアップ
  - 最新リリースは9.4.3
  - 年内には 9.5 がリリース??
- PostgreSQLライセンスで配布
  - BSDタイプの緩いライセンス

SRA OSS, INC.

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
  - 最新リリースは 3.4.2
  - 秋ごろに 3.5 をリリース予定





SRA OSS, INC.

データベースクラスタリング

データベースクラスタリング

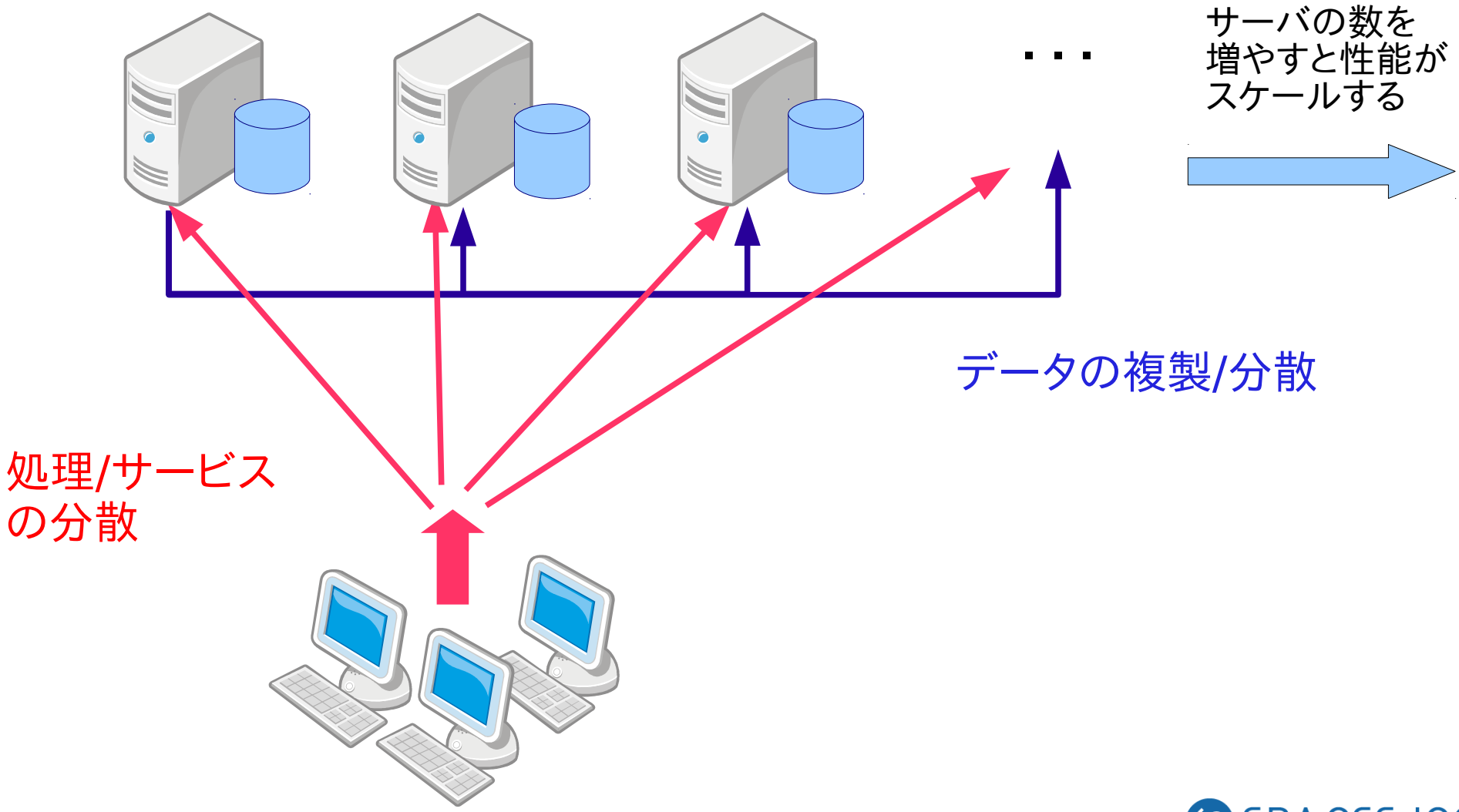
- 目的は？

  - 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
      → データの複製を複数もつことができる
  - 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
      → ノードを増やすことでスケールアウトさせたい
  - 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

SRA OSS, INC.

データベースのスケールアウト構成

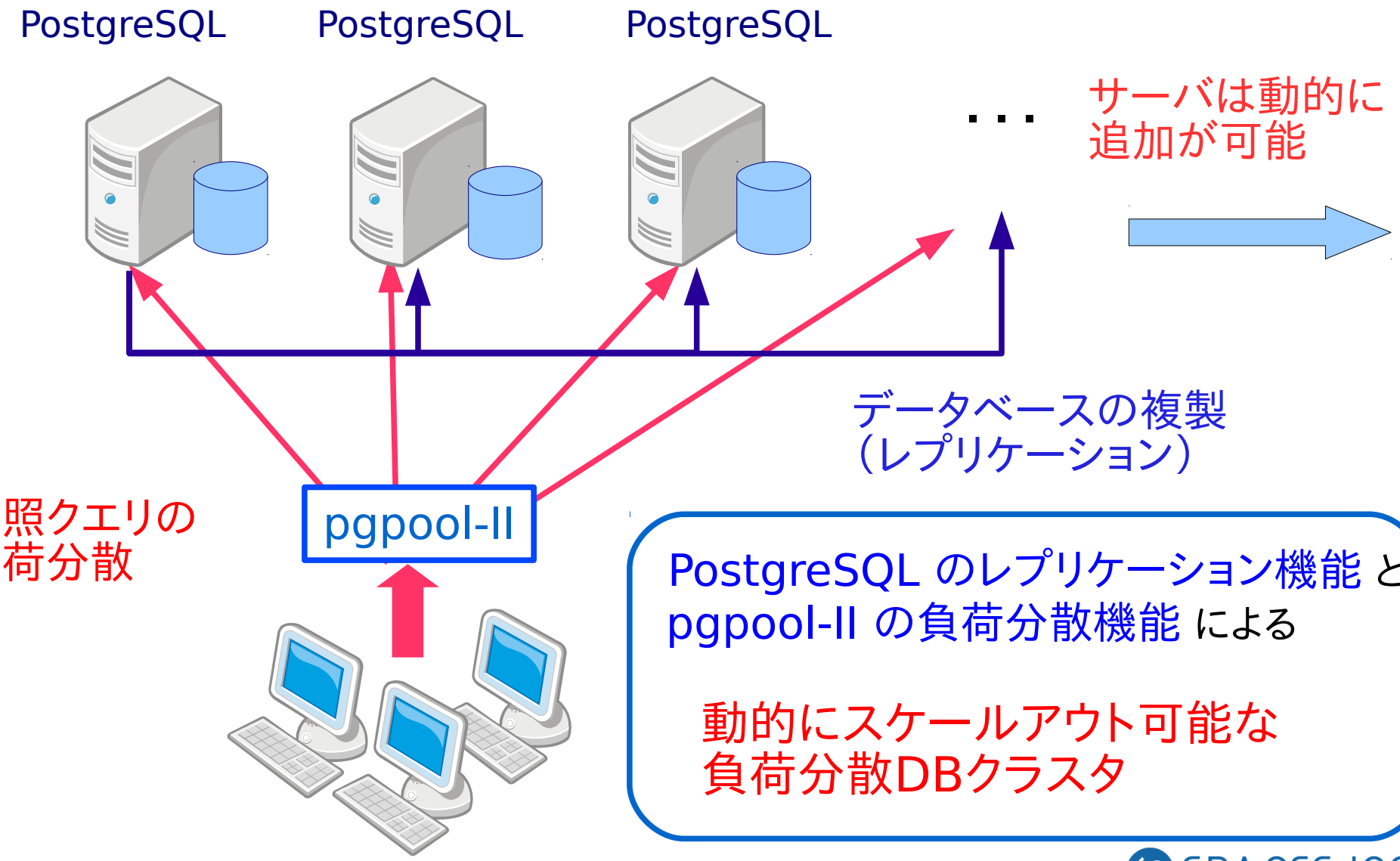
- 複数のデータベースサーバに処理を分散させる



SRA OSS, INC.

PostgerSQL + pgpool-II によるスケールアウト構成

- 複数の PostgreSQL サーバにクエリを分散させる



PostgreSQL のレプリケーション機能 と
pgpool-II の負荷分散機能 による

動的にスケールアウト可能な
負荷分散DBクラスタ

SRA OSS, INC.

PostgreSQL のレプリケーション機能

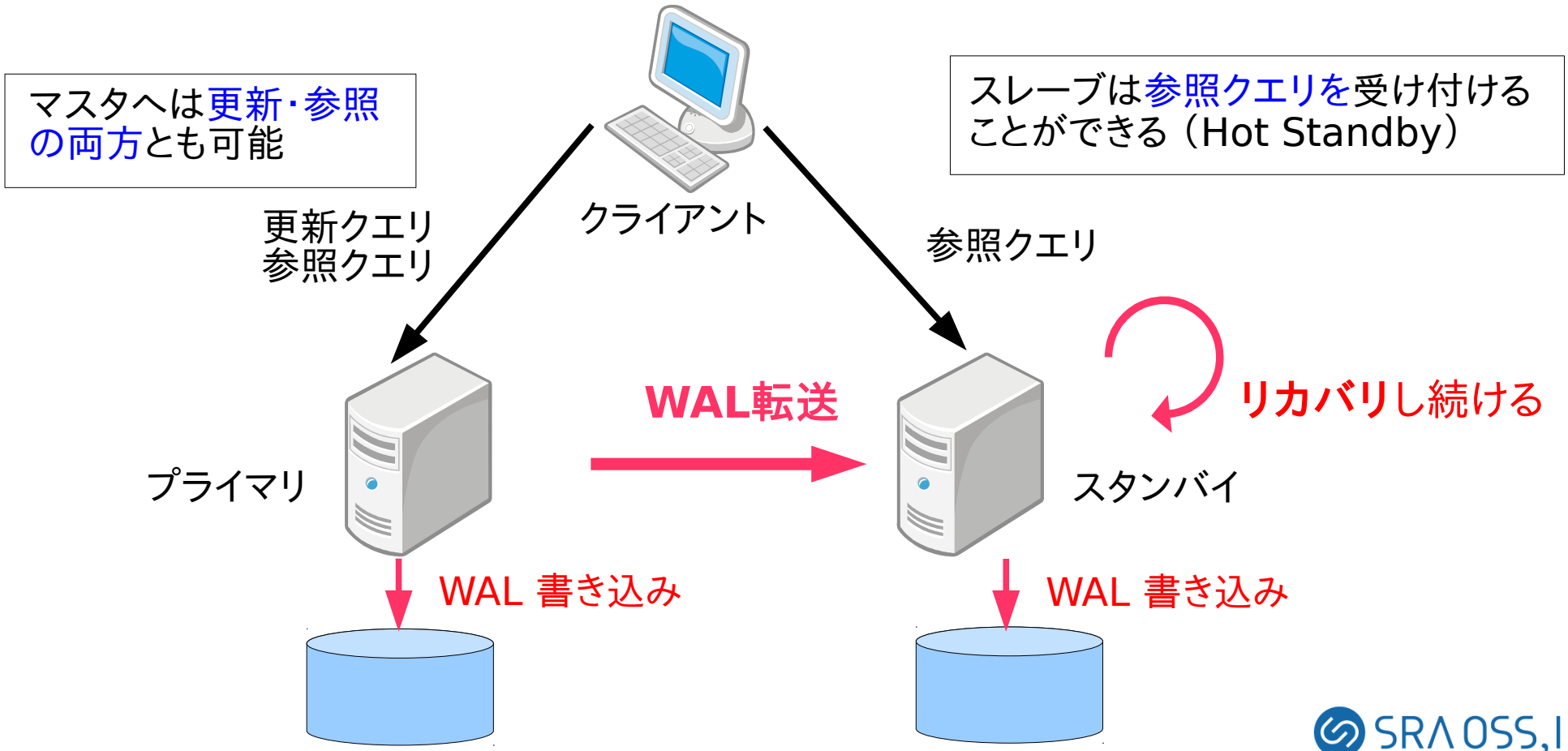
PostgreSQLのクラスタ技術

- 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の間に入って同期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

SRA OSS, INC.

PostgreSQL のレプリケーション機能

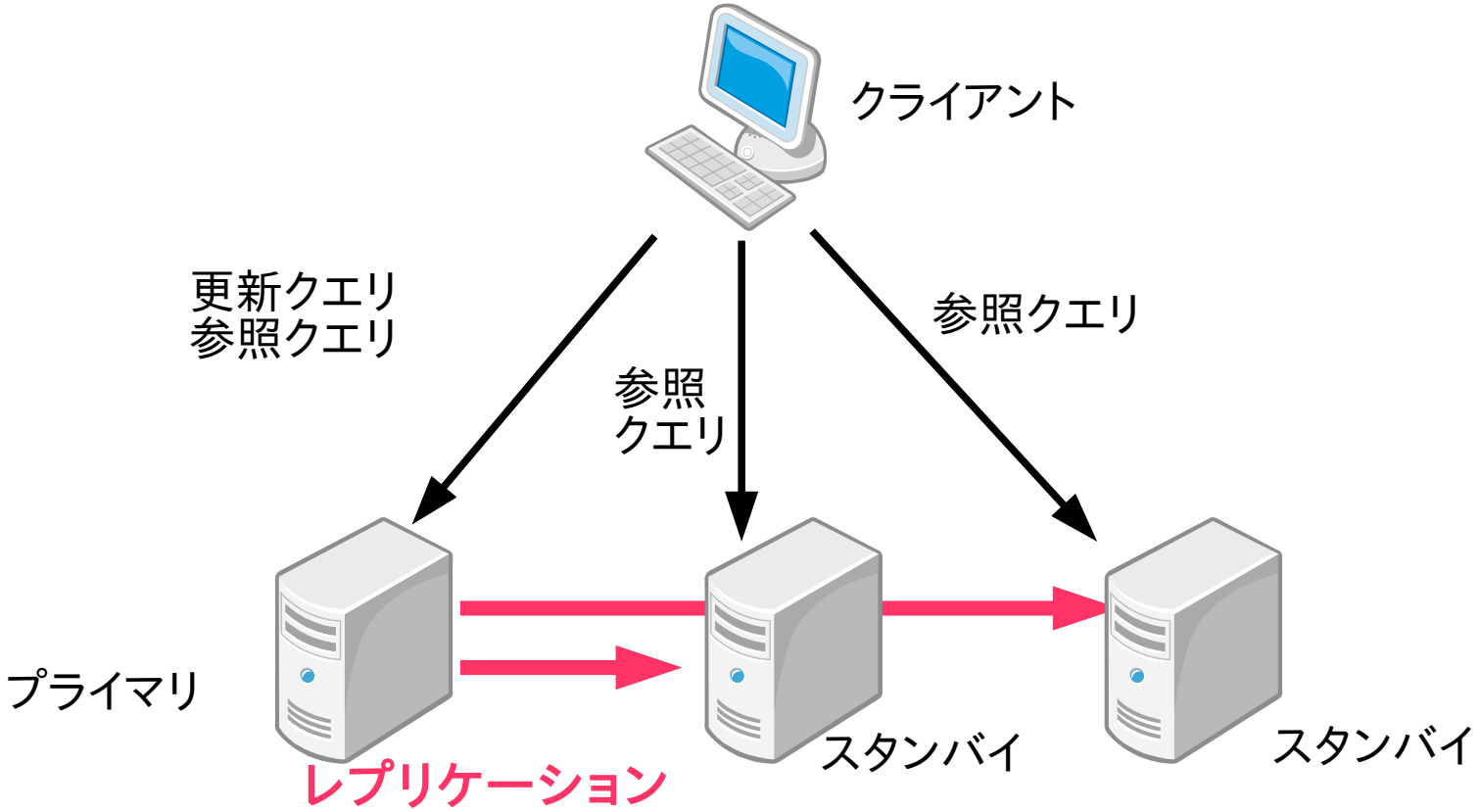
- ストリーミングレプリケーション（PostgreSQL 9.0 ～）
  - マスタからスレーブにトランザクションログ（WAL）を転送することによりデータの複製を実現
  - 対象はデータベース全体
  - 転送とリカバリの遅延のため、マスタとスレーブが常に同じ内容とは限らない



SRA OSS, INC.

複数スタンバイへのレプリケーション

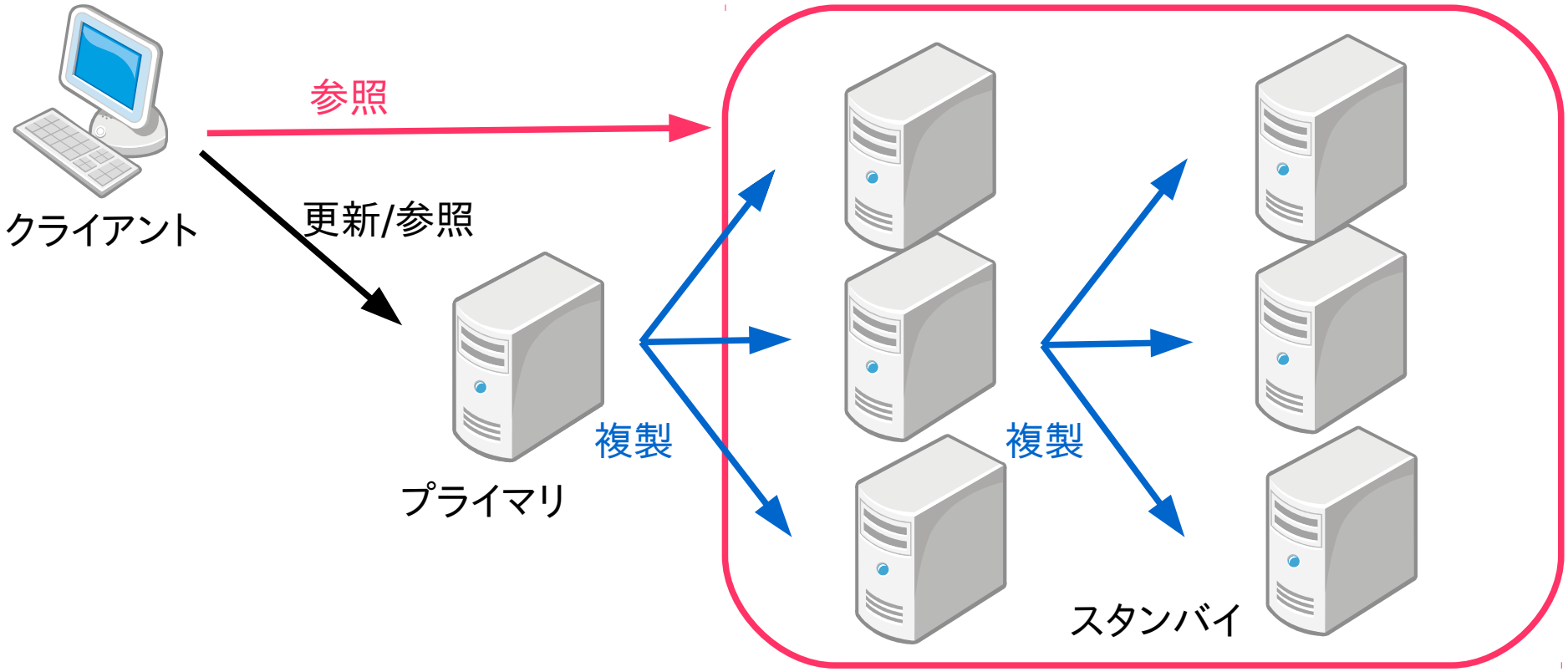
- 複数のスタンバイにレプリケーション可能
  - スレーブが参照クエリを受け付けることを利用して
    参照性能をスケールアウトさせることが可能



SRA OSS, INC.

カスケードレプリケーション

- カスケードレプリケーション（PostgreSQL 9.2 ～）
  - スレーブからさらに別のスレーブへのレプリケーションが可能
  - スタンバイ増加によるプライマリへの負荷の集中を回避



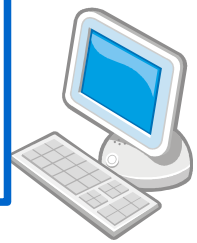
SRA OSS, INC.

その他のレプリケーション関連機能

- 同期レプリケーション (PostgreSQL 9.1～)
  - スタンバイ側のディスクに WAL が書き込まれることを保証するモード
    - 注意:データの同期を保証するものではない
- レプリケーションスロット（PostgreSQL 9.4～）
  - レプリケーション状態や付帯情報を保持する枠組み
    - スタンバイに必要な WAL が削除され、レプリケーション不能になるのを防止
    - スタンバイが参照しているデータが物理的に削除されてコンフリクトが起きるのを防止
- 論理デコーディング (PostgreSQL 9.4～)
  - テーブルへの更新内容をSQLレベルの更新情報として出力するモード
  - 将来的な機能拡張を実現するための基盤となる
    - 部分レプリケーション、マルチマスタレプリケーション、異種DBへのレプrケーション・・・・などの機能は今は実現できていない

PostgreSQL の機能だけではできないこと

肝心の負荷分散は？
更新クエリ、参照クエリの振り分けは？
　・・・アプリケーション側で対応？？



クライアント

更新クエリ
参照クエリ

参照
クエリ

参照クエリ

DBサーバを追加する場合は？
　・・・プライマリのバックアップ
　　＆同期を手動で？

プライマリ

レプリケーション

スタンバイ

スタンバイ

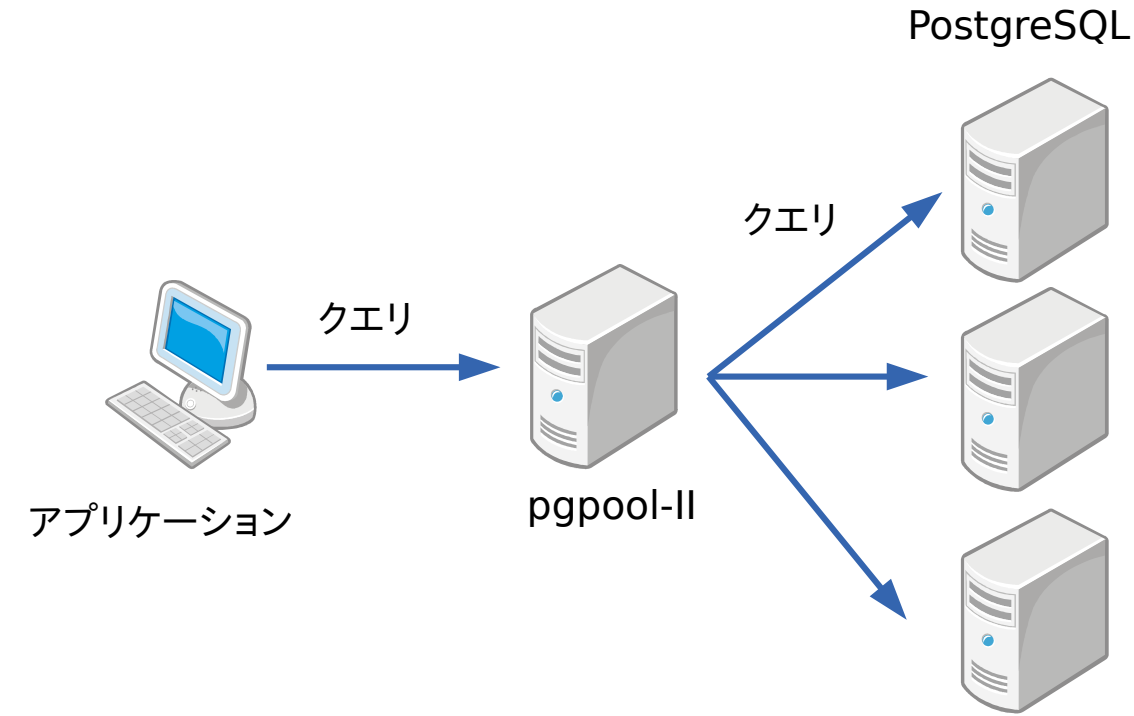
プライマリがダウンしたら更新クエリが処理できなくなる？!
　・・・スタンバイの昇格も手動で？

SRA OSS, INC.

pgpool-II のクラスタリング機能

pgpool-II の機能

- アプリケーションからは1台の PostgreSQL のように見える
  - PostgreSQL のSQLパーサが移植されており、同じように構文を理解可能
- 多彩な機能を持つ
  - 性能向上
    - 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
  - 高可用性
    - 自動フェイルオーバ
    - watchdog
  - クラスタ管理
    - オンラインリカバリ
  - クラスタとアプリケーションの親和性
    - クエリの自動振り分け



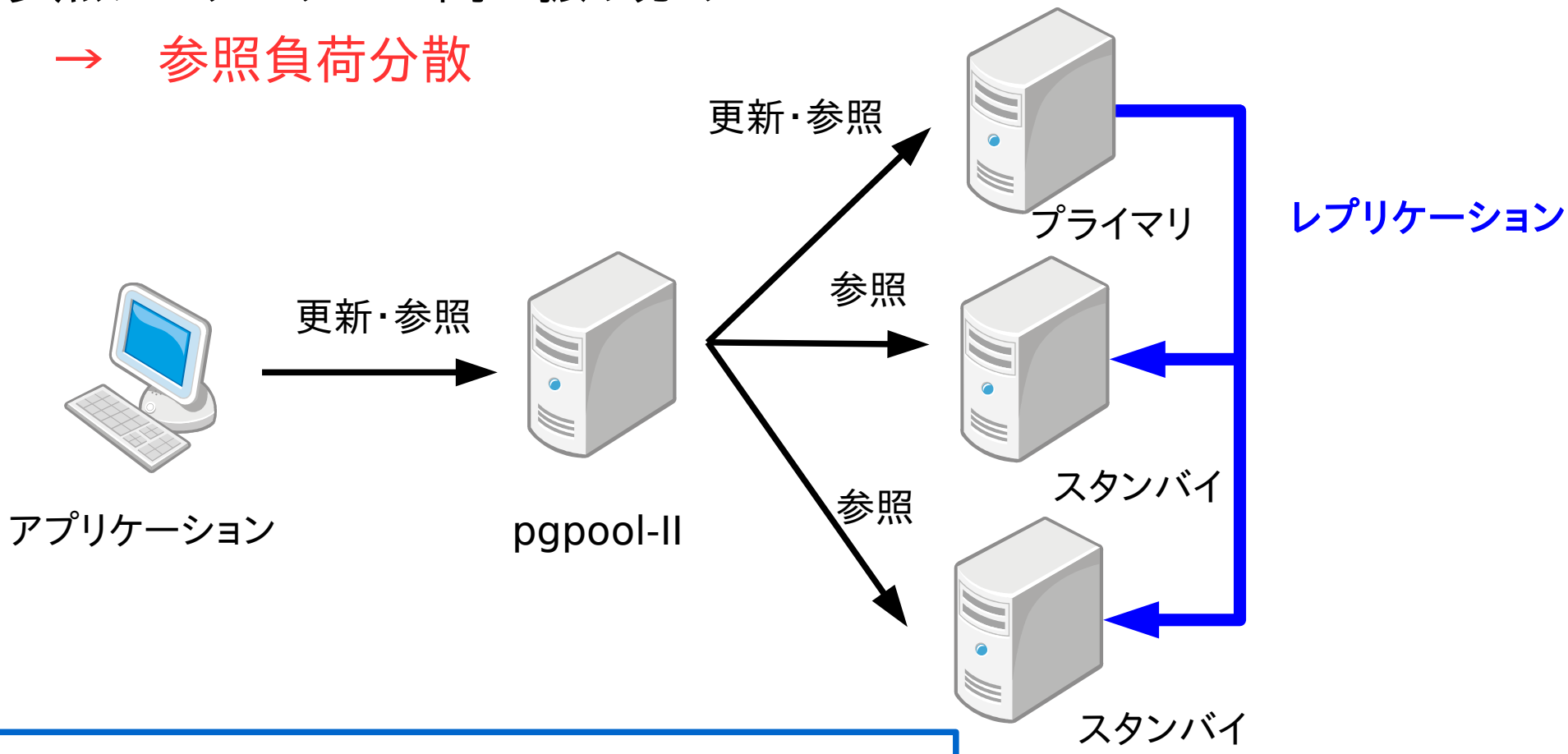
SRA OSS, INC.

クエリの振り分け

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

→　参照負荷分散

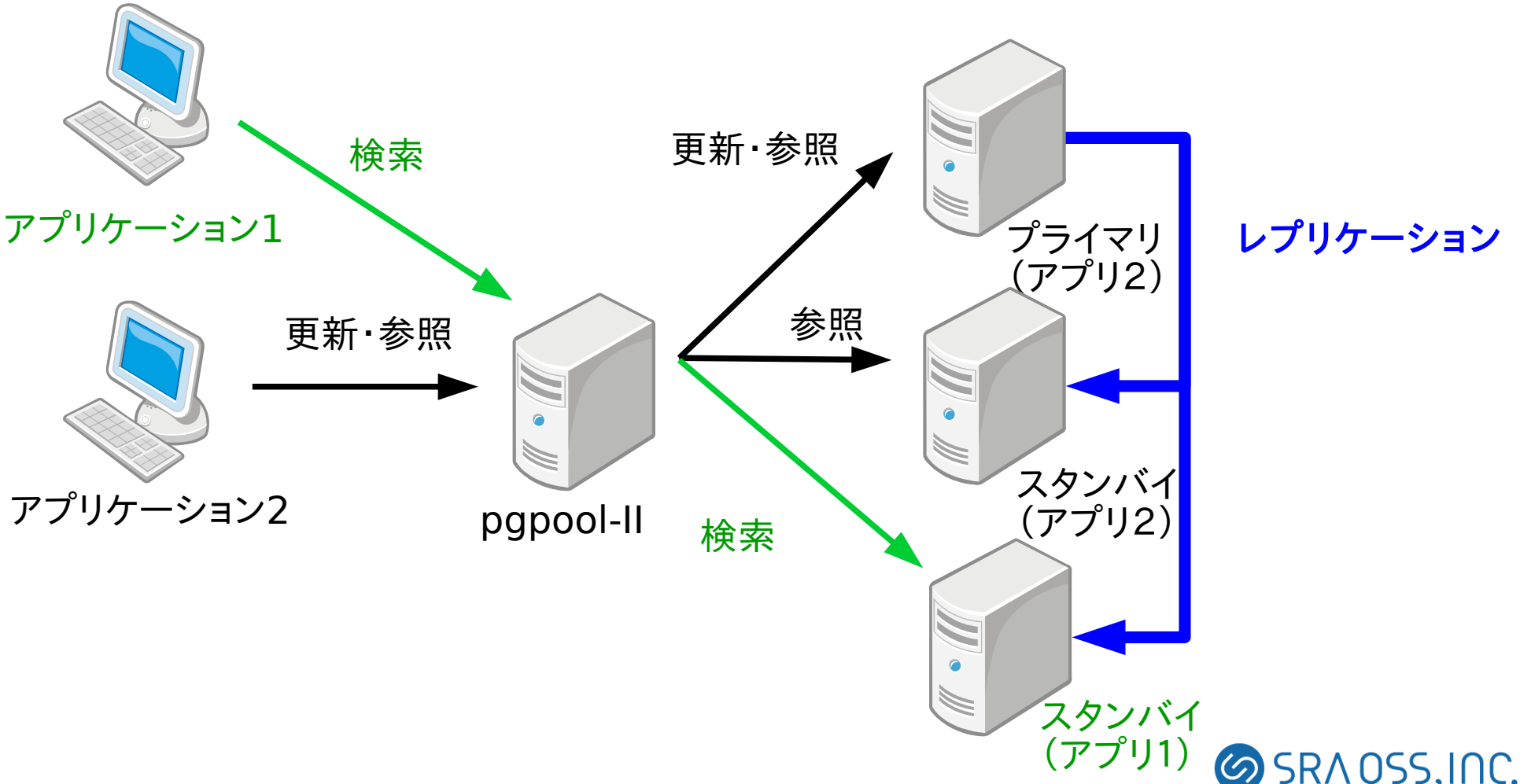


- 振り分けの重みを指定可能
- レプリケーション遅延が大きいサーバには振り分けない
- 3.4 からはよりきめ細やかな振り分けが可能に!

SRA OSS, INC.

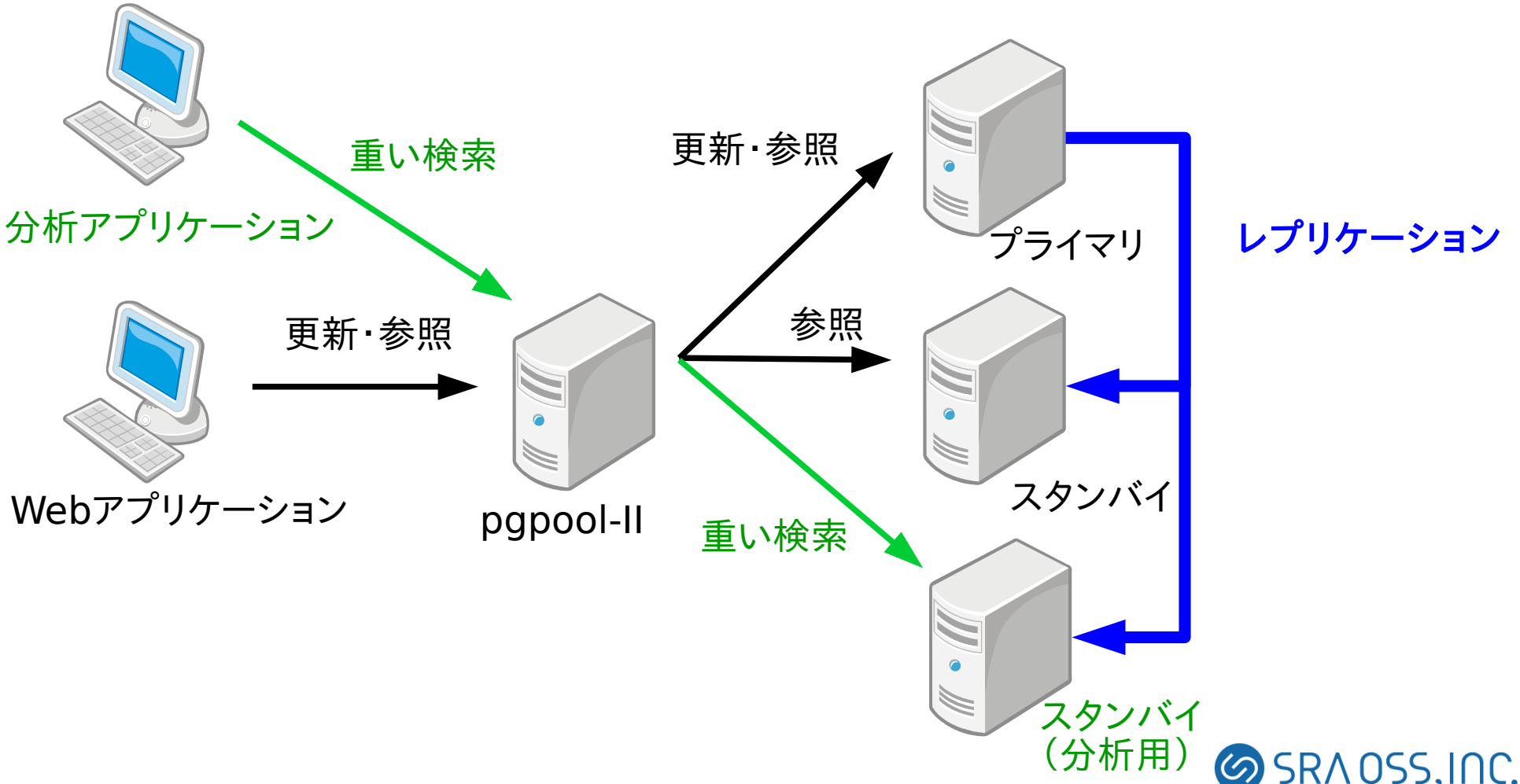
クエリの自動振り分け

- アプリケーション名、DB名にしたがって、クエリの振り分け先を指定できる (pgpool-II 3.4～)



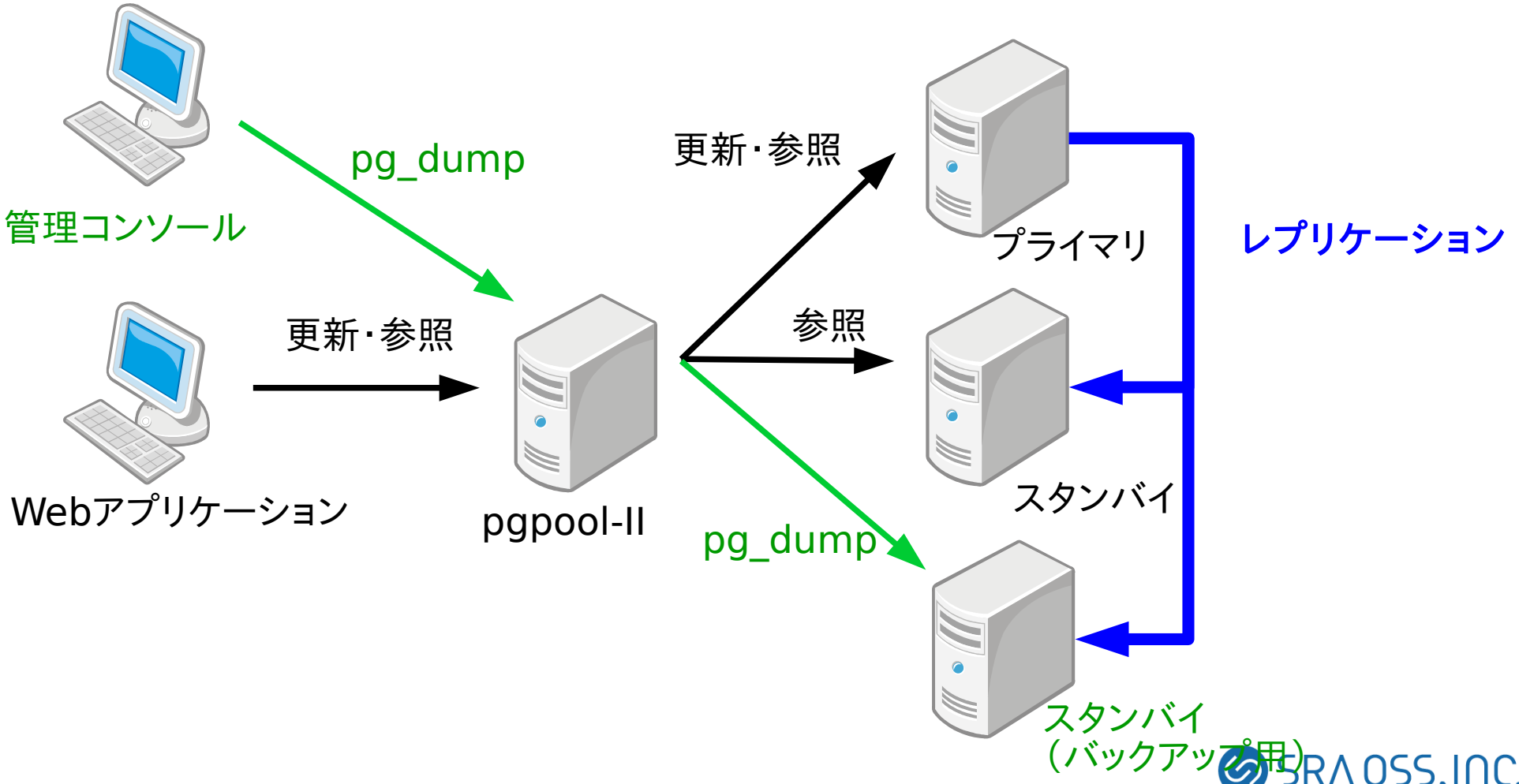
クエリの自動振り分け

- 例1）
  - 重い検索を行うデータ分析アプリケーションからのクエリは、分析専用のスタンバイサーバへ振り分ける
  - Web アプリケーションのクエリ処理を邪魔しない



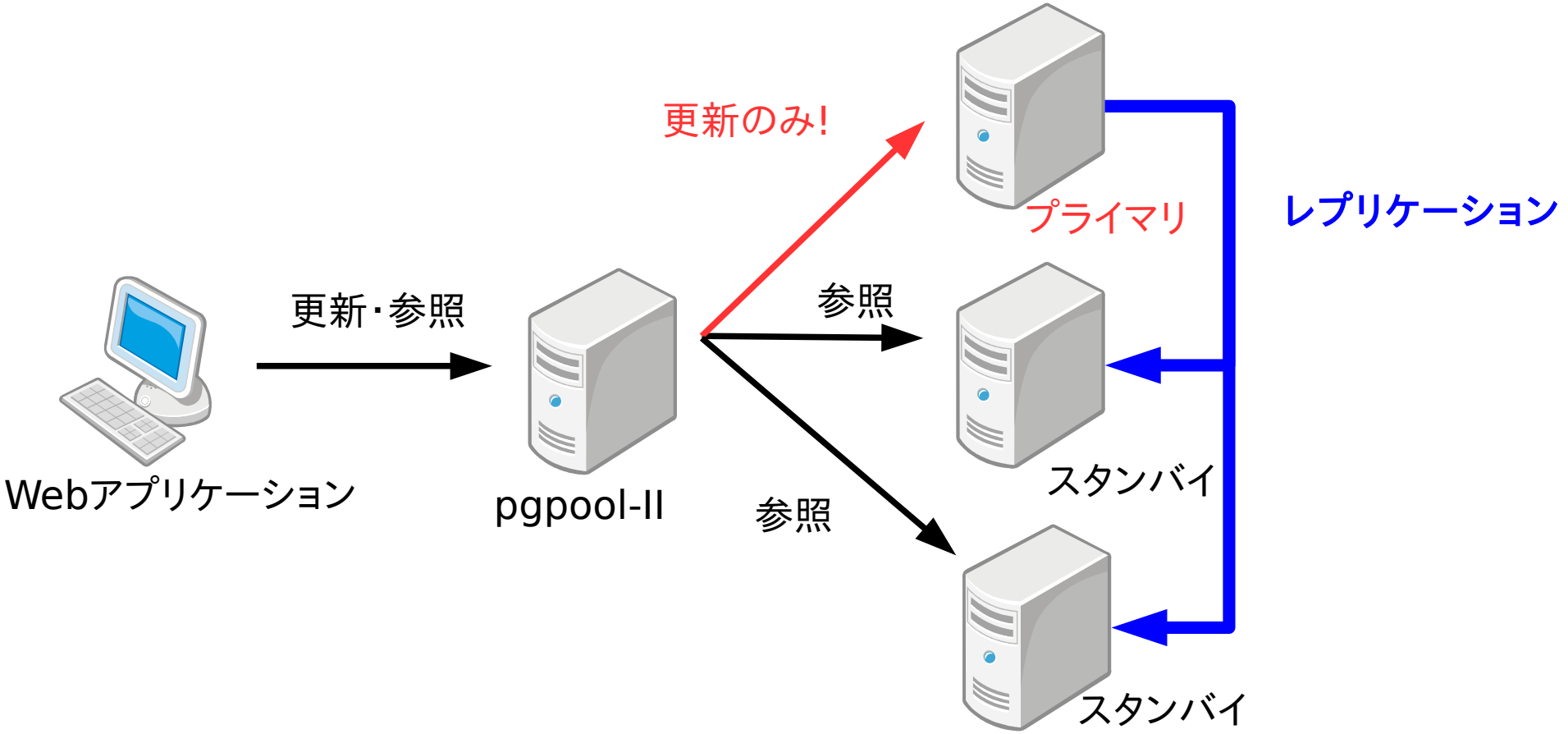
クエリの自動振り分け

- 例2）
  - バックアップツール（pg_dump)からのアクセスは、バックアップ専用のスタンバイサーバへ振り分ける
  - Web アプリケーションのクエリ処理を邪魔しない



クエリの自動振り分け

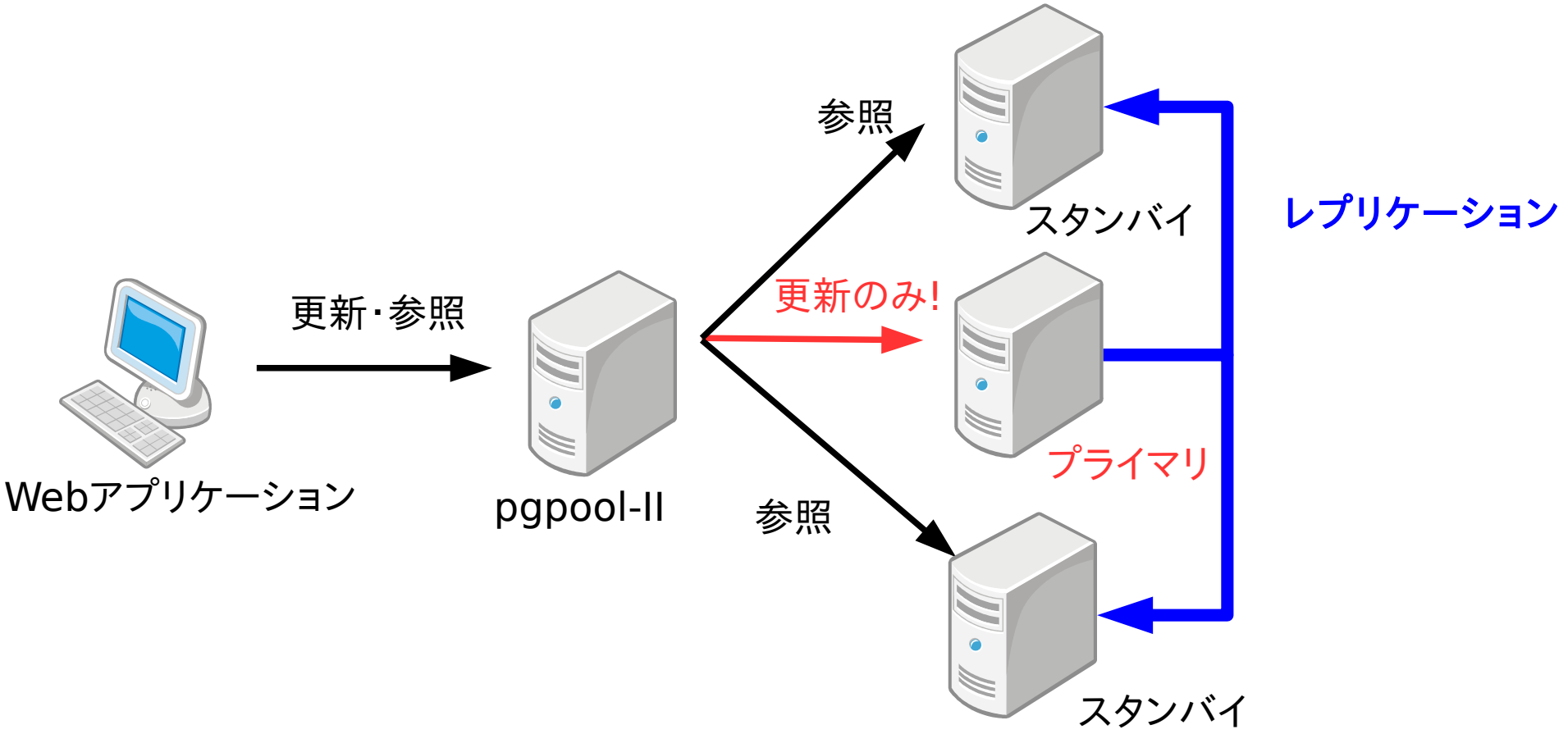
- 例3）
  - 参照クエリは全てスタンバイに送りたい



SRA OSS, INC.

クエリの自動振り分け

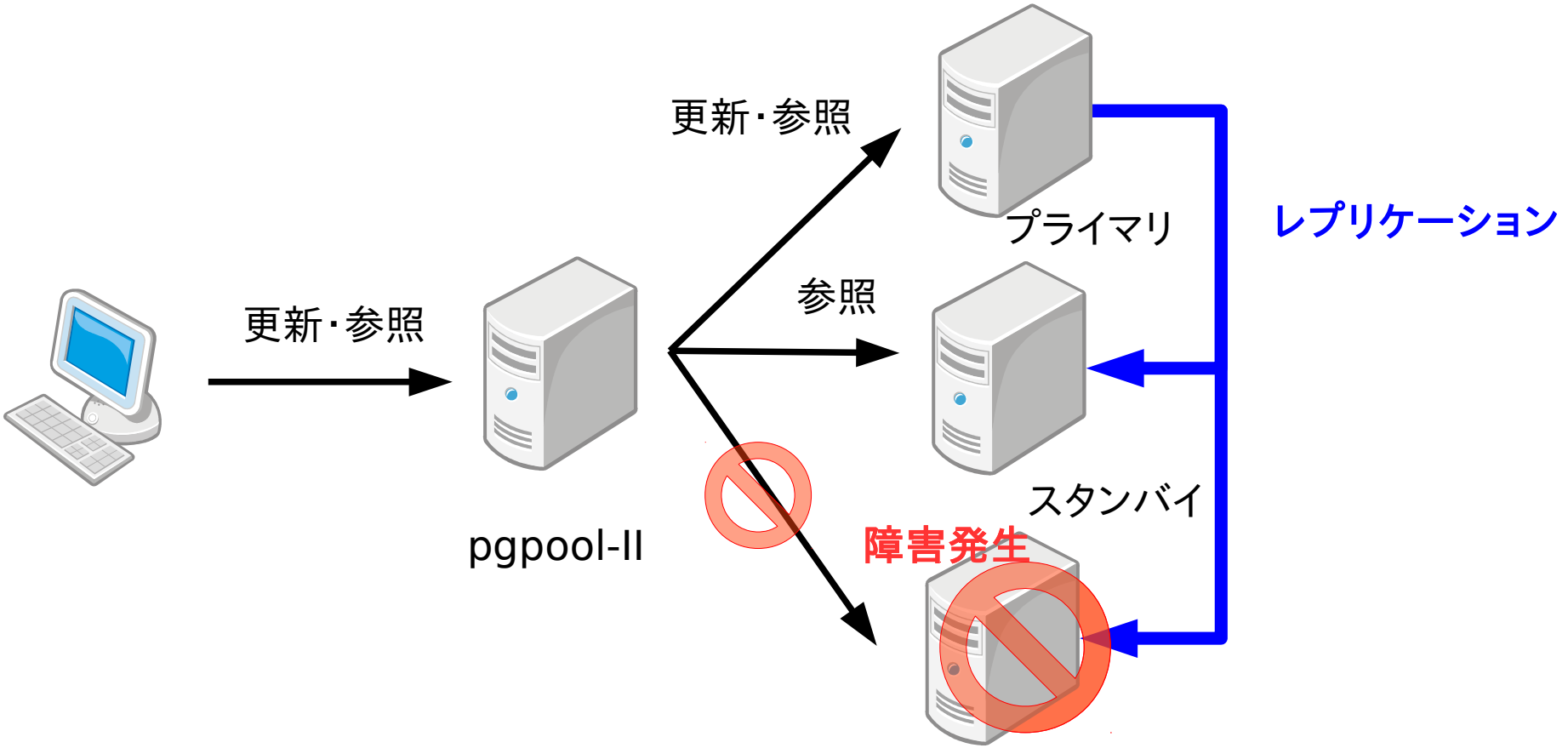
- 例3）
  - 参照クエリは全てスタンバイに送りたい
  - たとえ、プライマリノードが変わったとしても



ノードの障害

自動フェイルオーバ

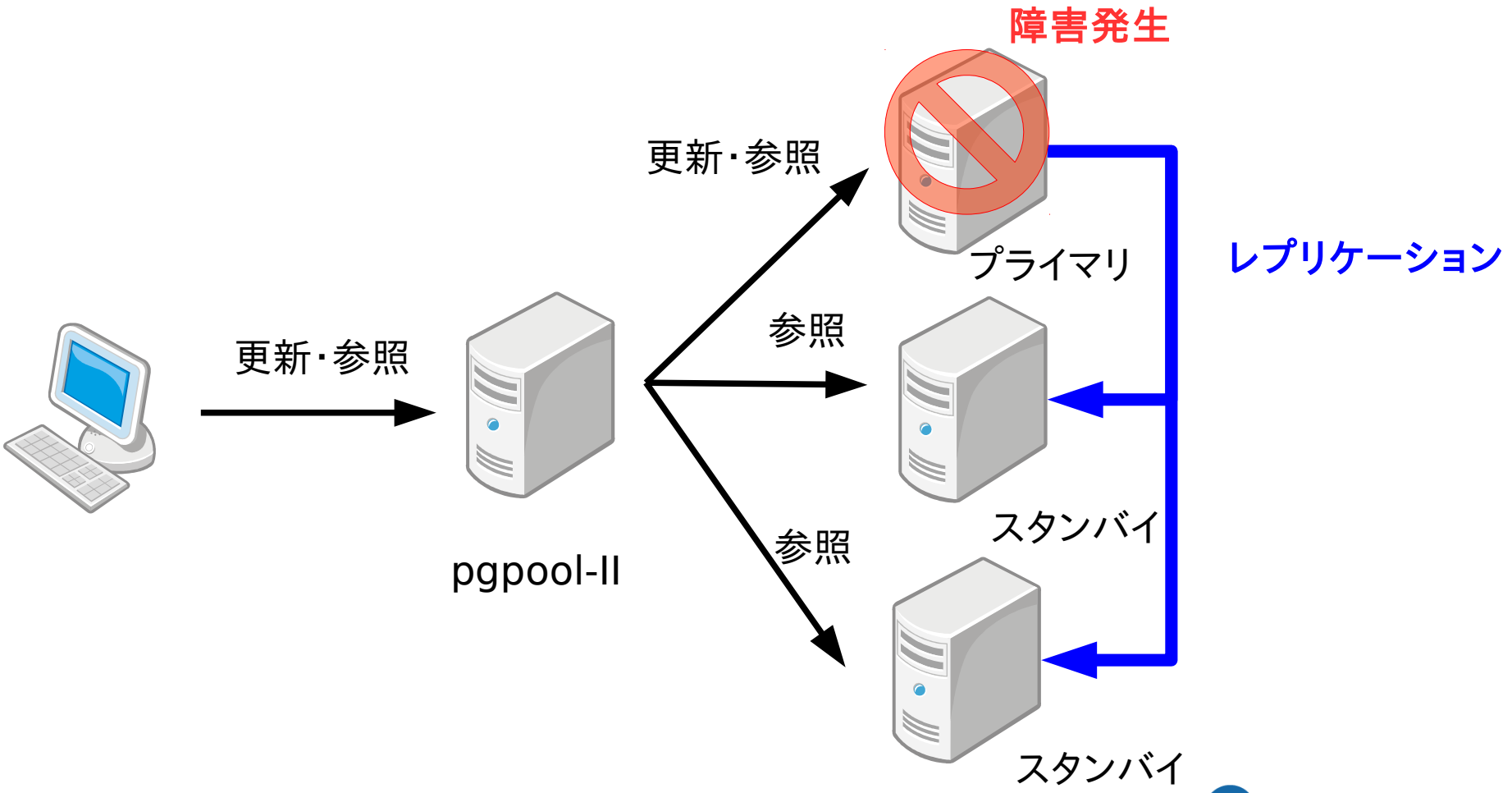
- DBサーバの障害を自動検出（ヘルスチェック機能）
  - ダウンしたPostgreSQLを切り離す
    → 負荷分散の対象から外れる



SRA OSS, INC.

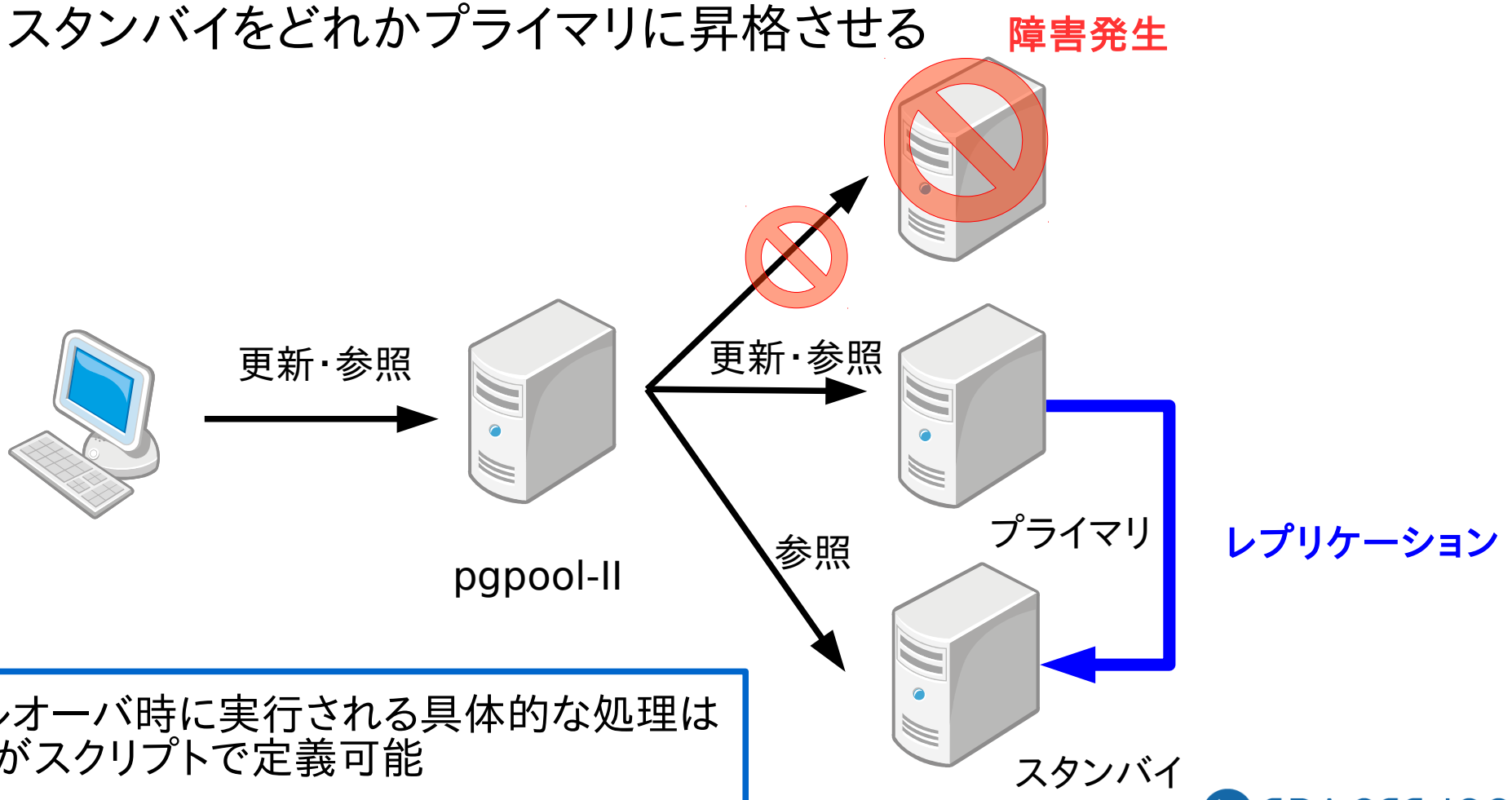
自動フェイルオーバ

- 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

- 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ってしまう
  - 負荷分散の対象から切り離す
  - スタンバイをどれかプライマリに昇格させる



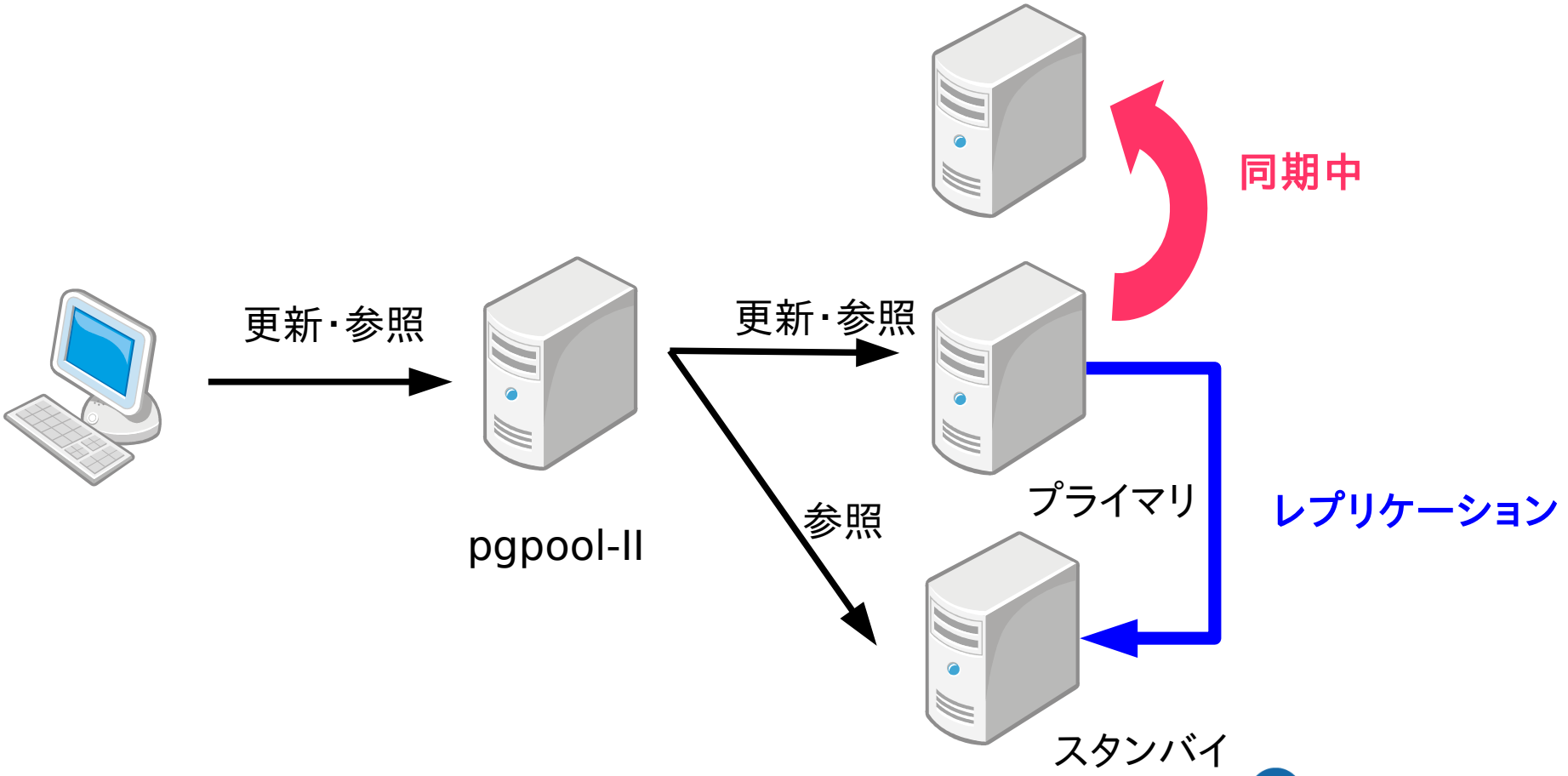
フェイルオーバ時に実行される具体的な処理は
ユーザがスクリプトで定義可能

SRA OSS, INC.

ノードの復旧

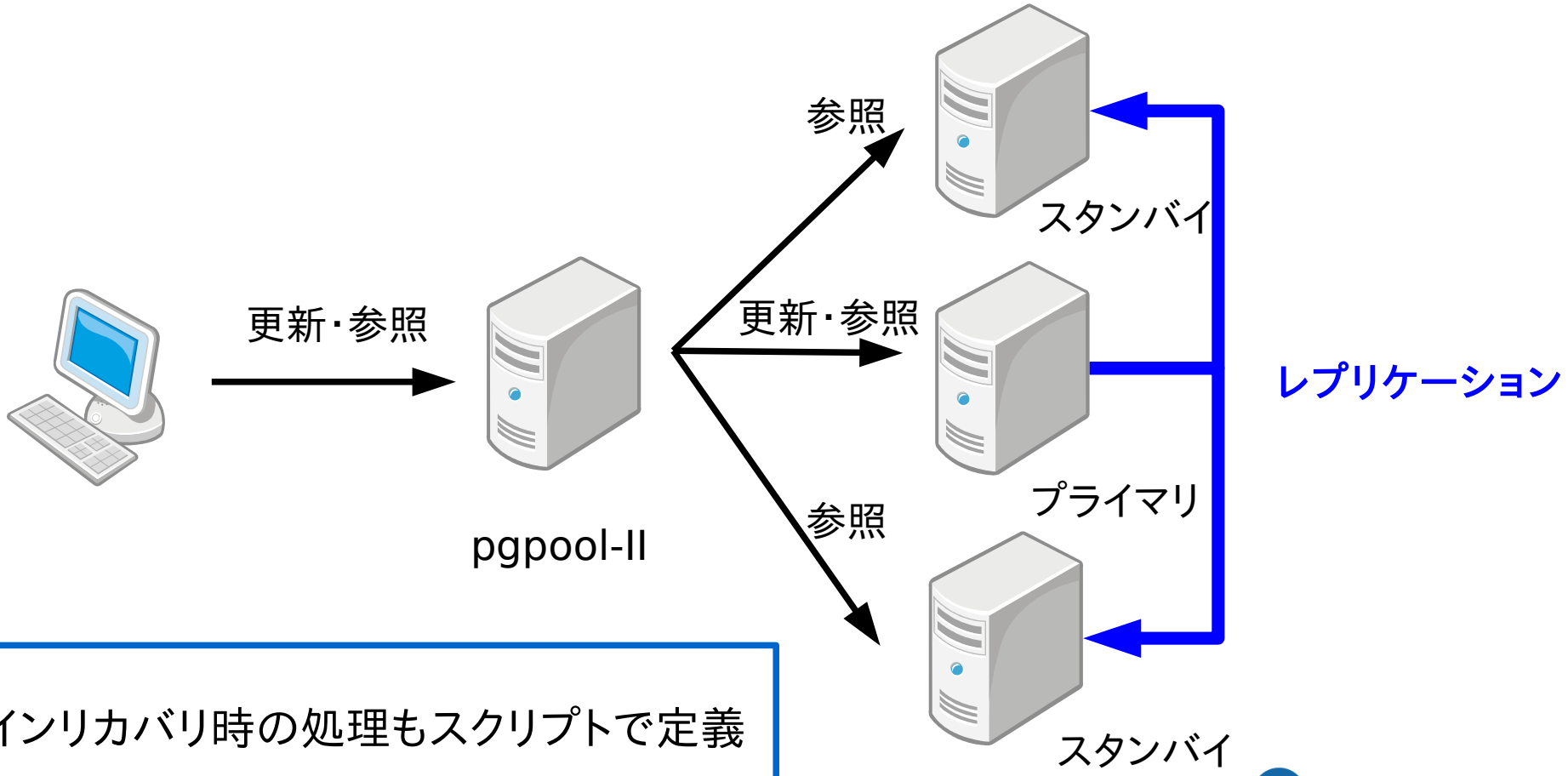
オンラインリカバリ

- ダウンしたスタンバイをプライマリに再同期させる
- 同期中もプライマリでは更新が可能



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再同期させる
- 同期中もプライマリでは更新が可能
- 同期完了後、マスタからのレプリケーションが再開され、自動的に負荷分散の対象となる



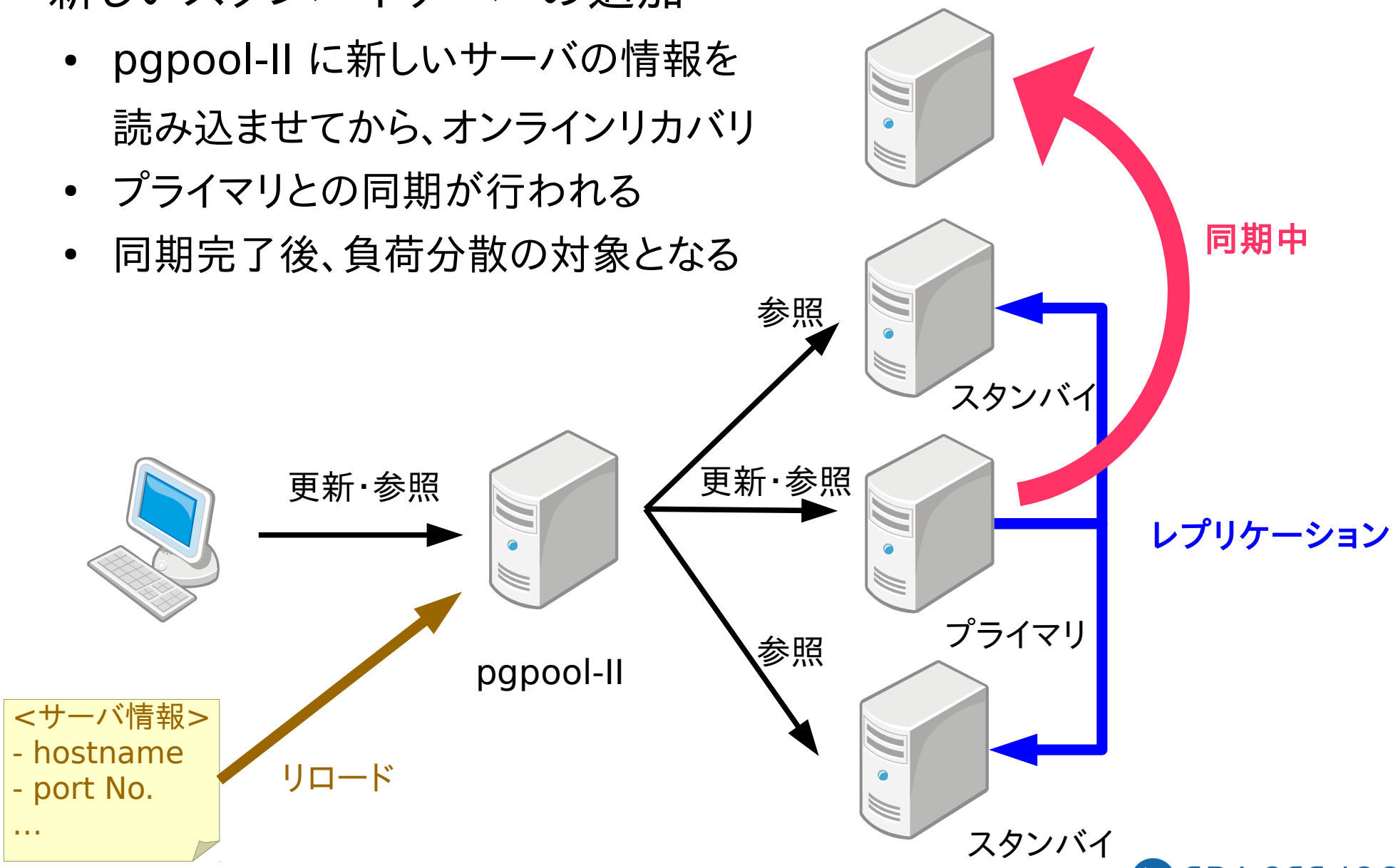
オンラインリカバリ時の処理もスクリプトで定義

SRA OSS, INC.

動的なノードの追加

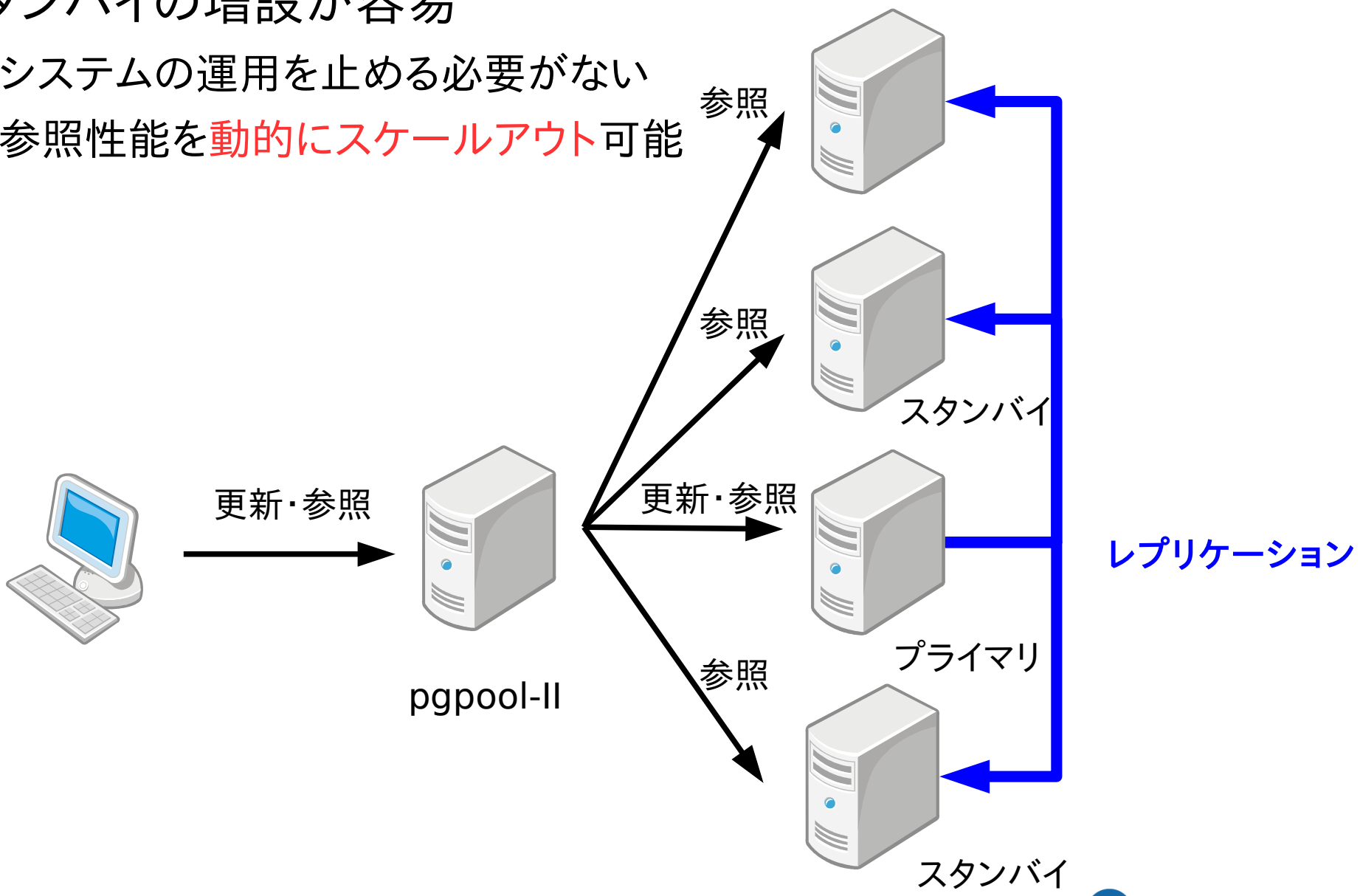
スタンバイサーバの追加

- 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
  - プライマリとの同期が行われる
  - 同期完了後、負荷分散の対象となる



スタンバイサーバの追加

- スタンバイの増設が容易
  - システムの運用を止める必要がない
  - 参照性能を動的にスケールアウト可能



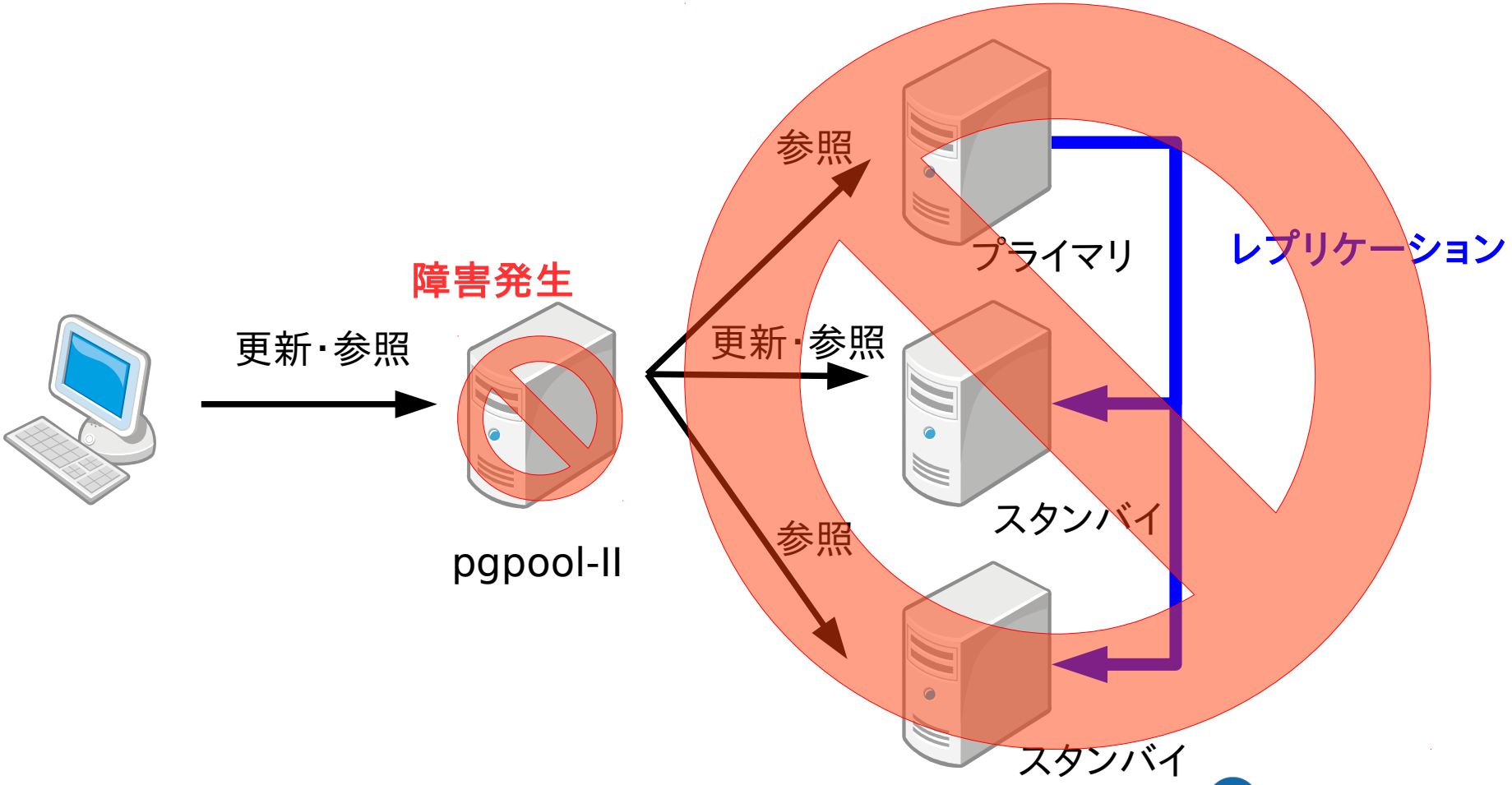
SRA OSS, INC.

PostgreSQL と pgpool-II のシステム構成

SRA OSS, INC.

単一障害点?

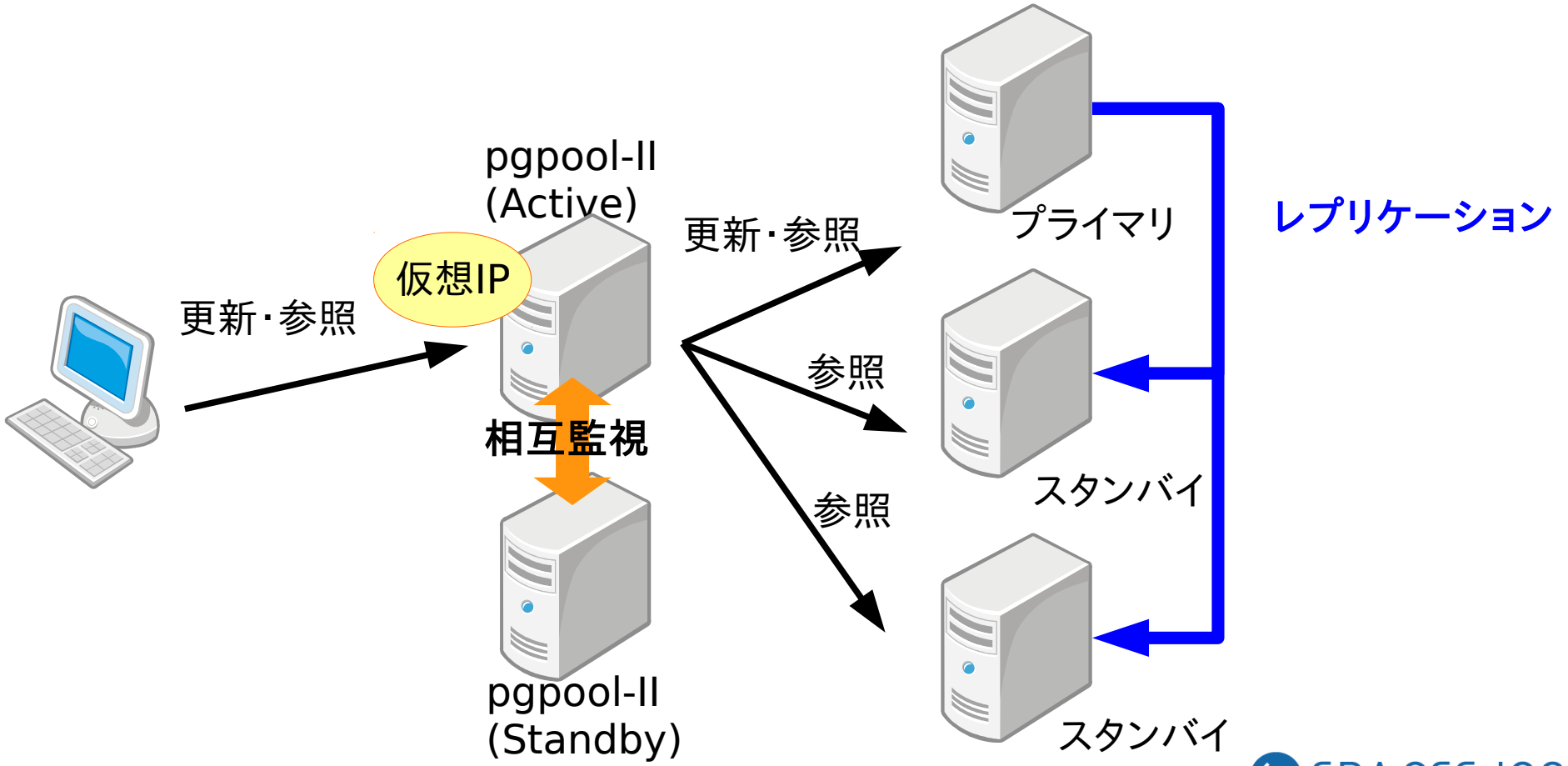
- もし、pgpool-II に障害が発生したら?!
  - 単一障害点 (Single Point of Failure) じゃないの?



SRA OSS, INC.

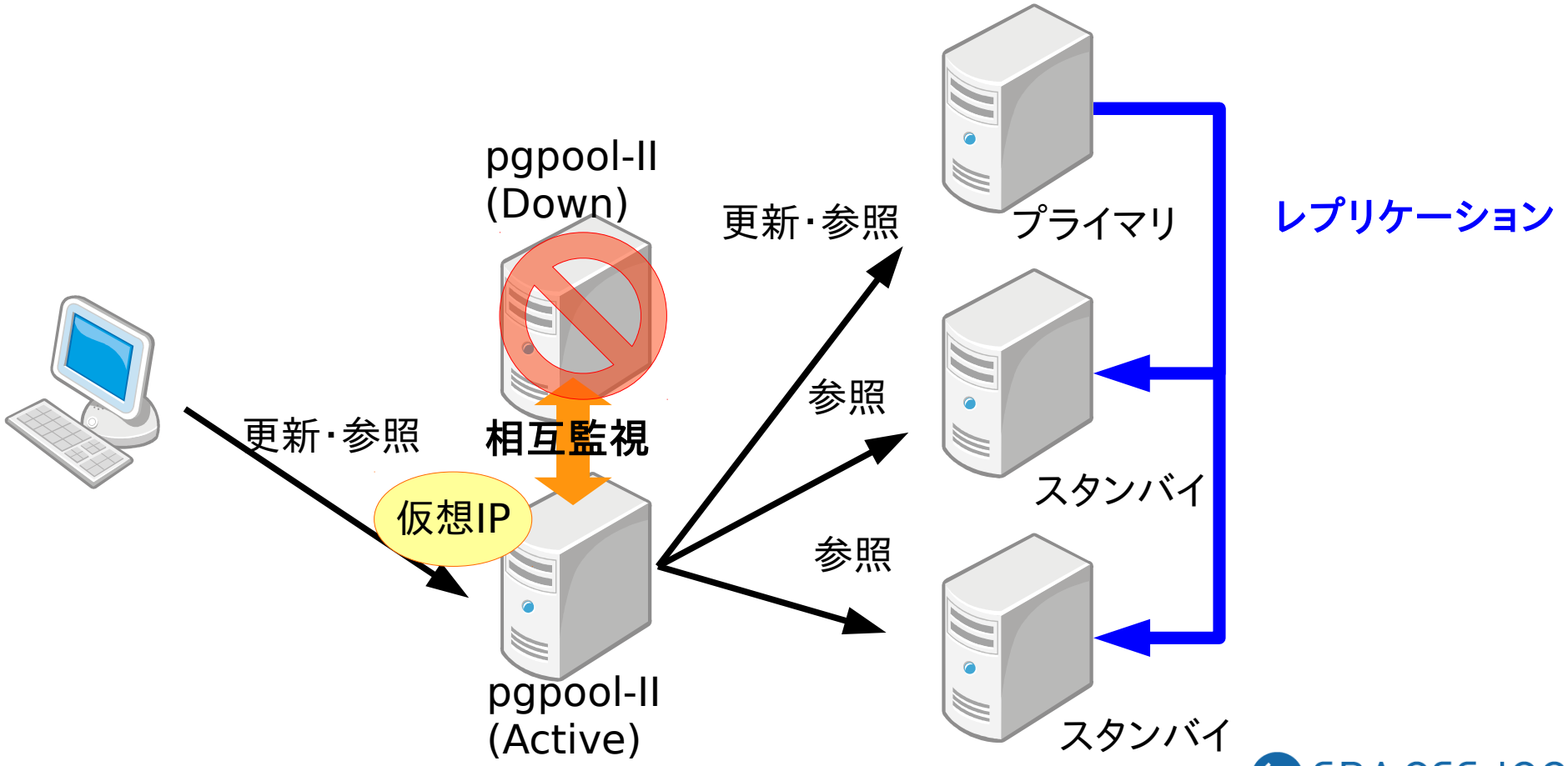
watchdog

- 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



マルチマスタ的構成

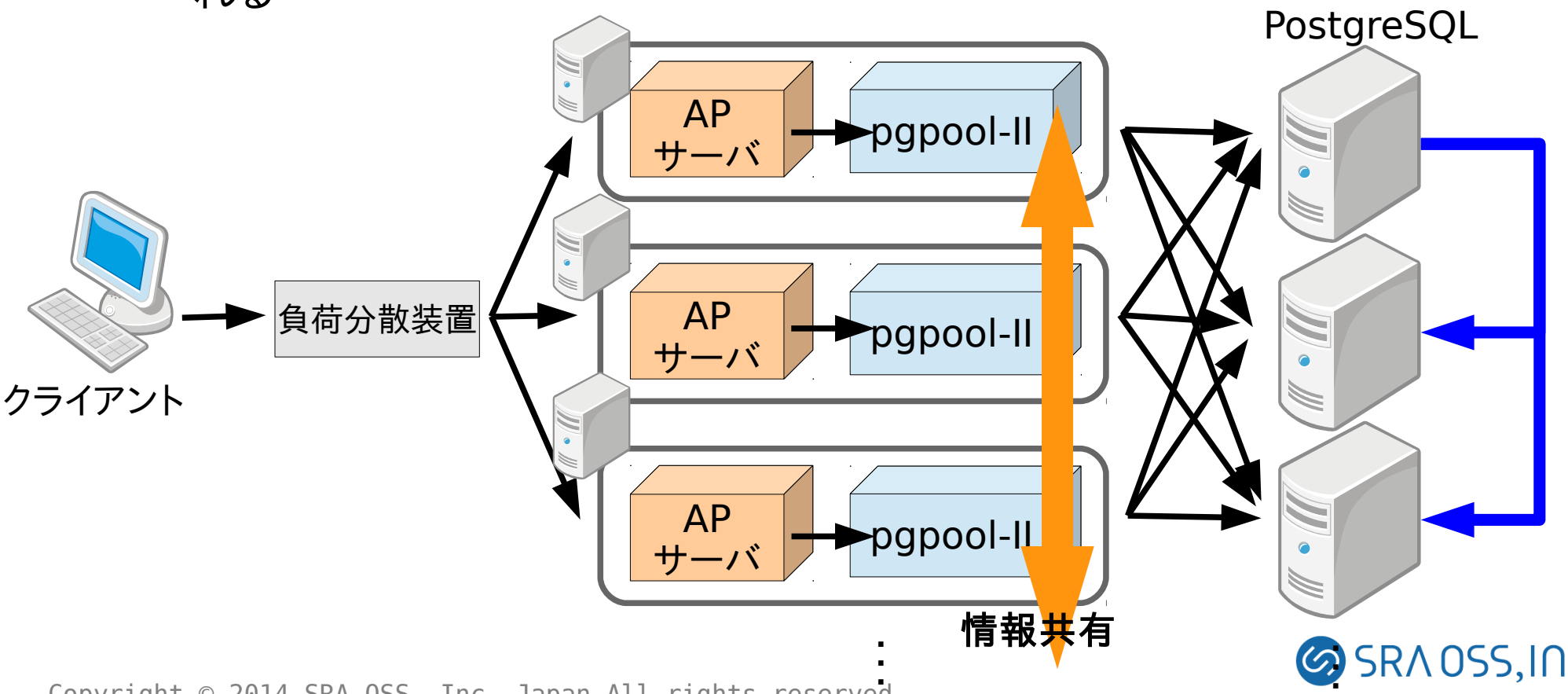
- Active pgpool-II に障害発生すると･･･
  - Standby pgpool-II が Active に昇格
  - 仮想IPでの付け替えが行われる



SRA OSS, INC.

マルチマスタ的構成

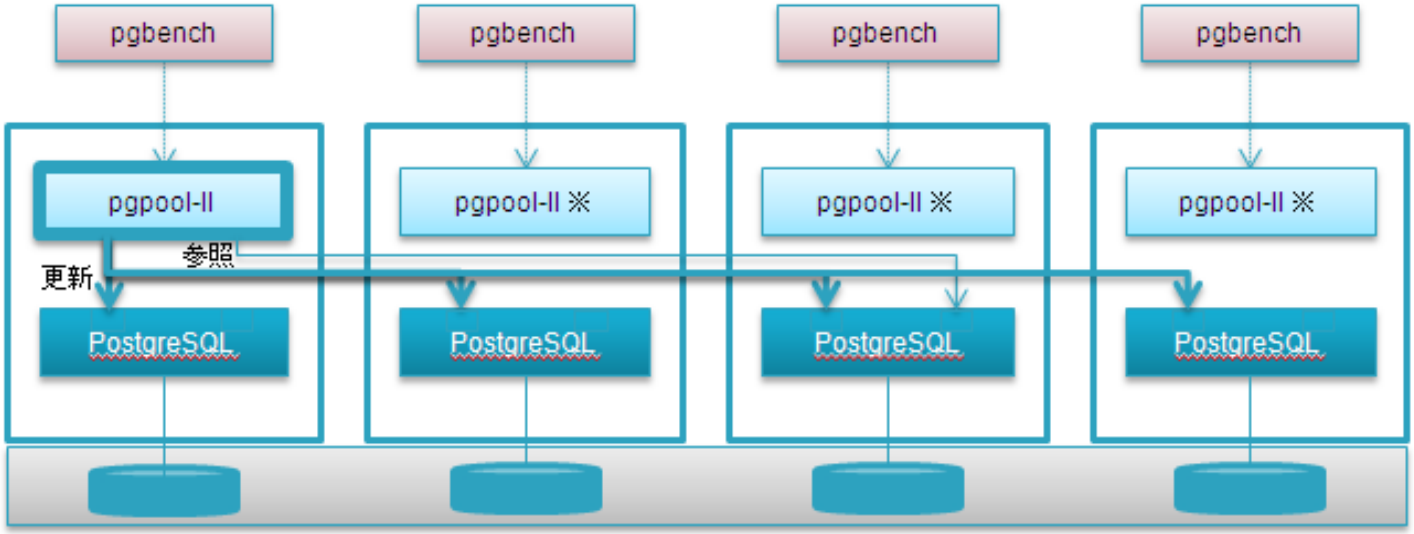
- APサーバと pgpool-II を1台のサーバに同居させた構成
  - pgpoo-II が冗長化されている
  - APサーバ/pgpool-II のペアを増やすことで、APサーバの性能をスケールアウト可能
  - Watchdog の機能により pgpool-II 間でバックエンド情報が共有される



スケールアウト性能

スケールアウト性能

- 参照性能は本当にスケールアウトするか？
- pgpool-II(3.2.1) & PostgreSQL(9.2.1)で、ノード数を増やすと全体の処理能力が向上するかを確認
  - 1～4台の PostgreSQL で検証
  - マルチマスタ的構成と似た構成
  - APサーバに相当する位置に、ベンチマークツール（pgbench）が配置されている。

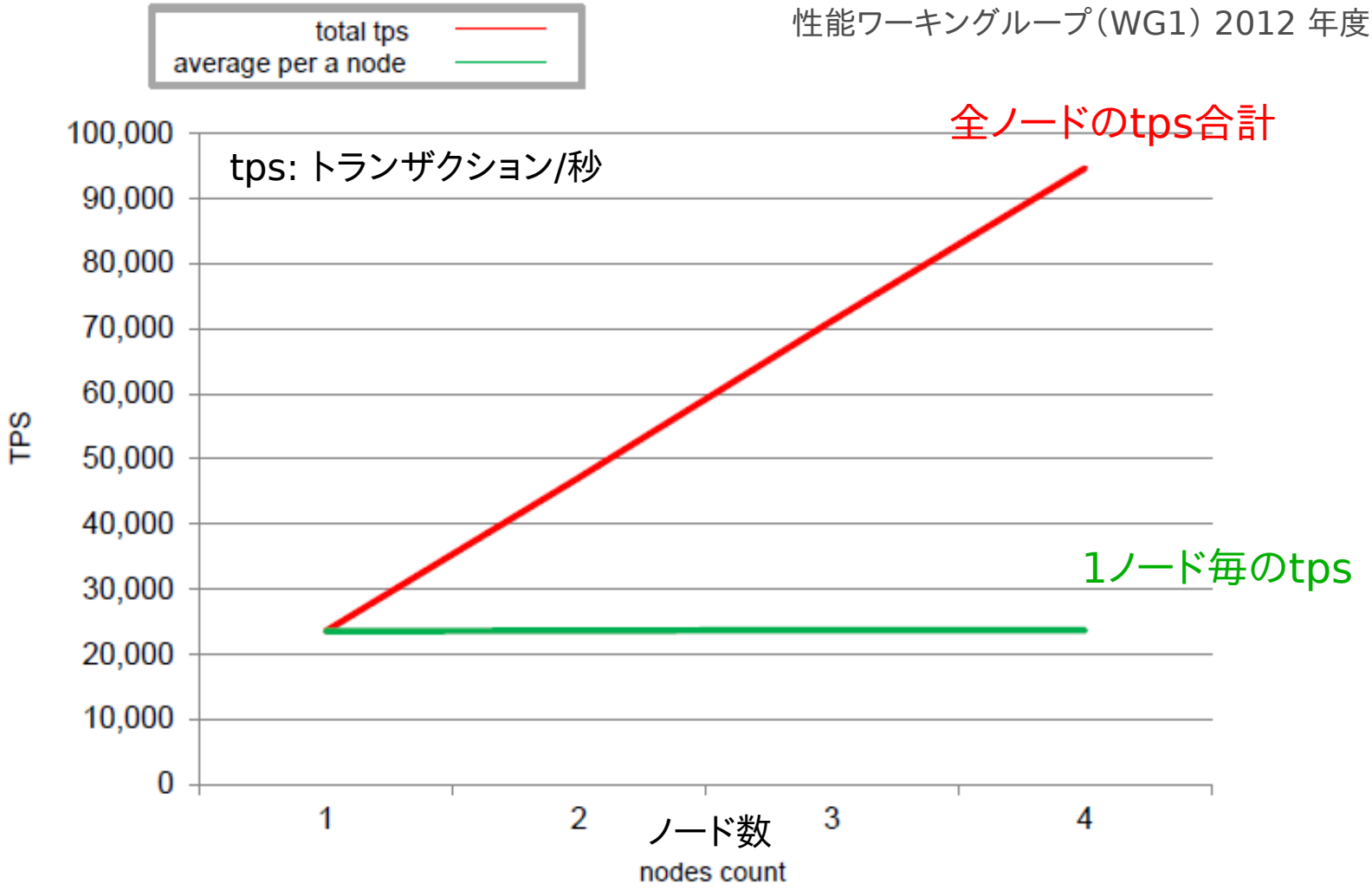


※PostgreSQL エンタープライズ・コンソーシアム
性能ワーキンググループ（WG1）2012 年度成果物より引用

SRA OSS, INC.

スケールアウト性能（結果）

※PostgreSQL エンタープライズ・コンソーシアム
性能ワーキンググループ（WG1）2012 年度成果物より引用



ノード数が増えるほど、合計の tps が増える（スケールメリットあり）

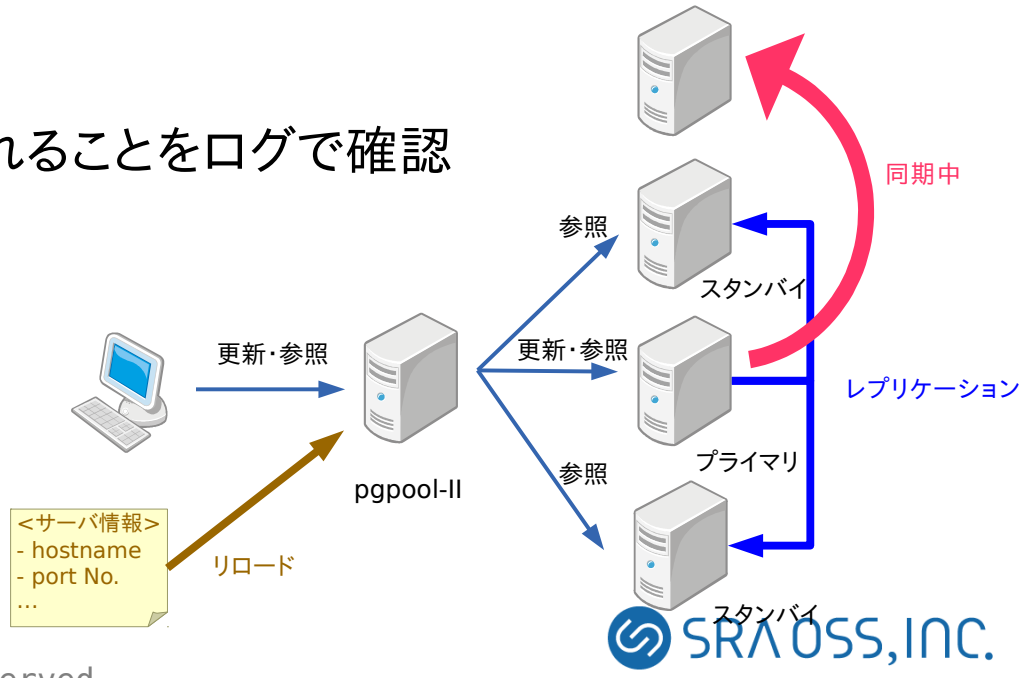
SRA OSS, INC.

デモ

デモの内容

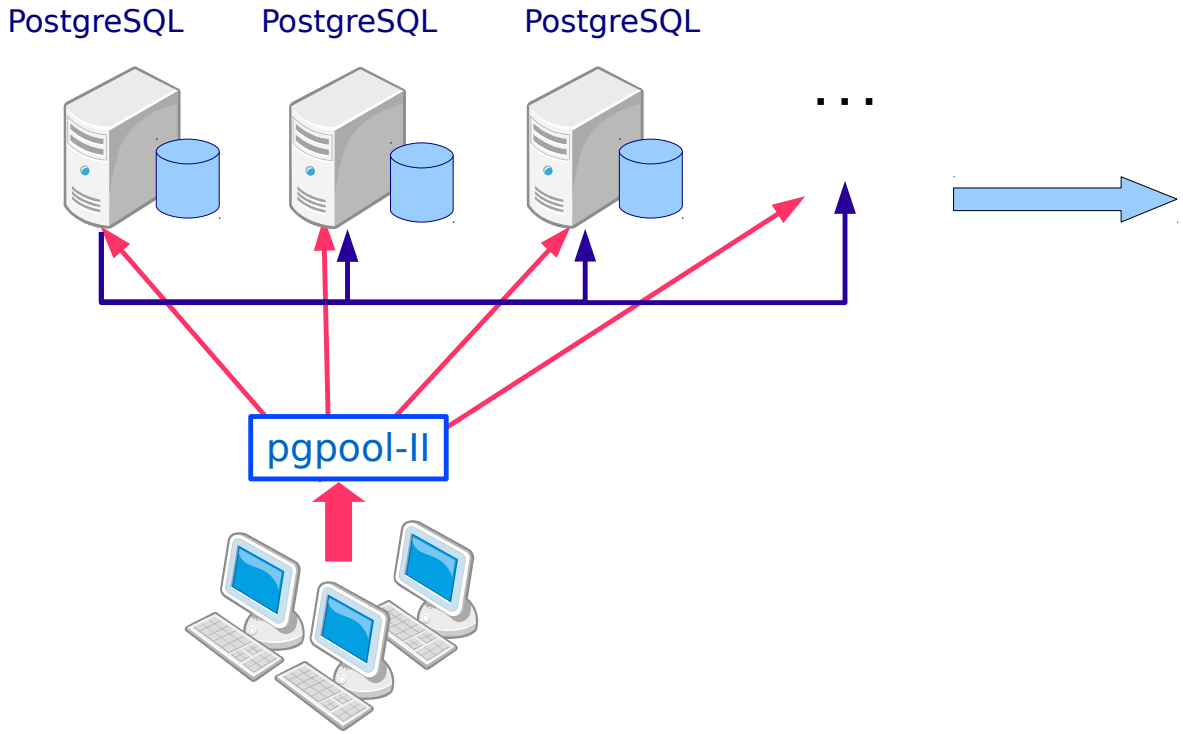
- 運用中のシステムに新しいスレーブを追加
  - ベンチマーク処理を中断せずに新しいノードが追加可能
  - 自動的に新しいスレーブに参照クエリが振り分けられるようになる

- デモ手順
  1. PostgreSQL サーバ2台 + pgpool-II のクラスタを作成
  2. ログを確認しながら、ベンチマーク (pgbench) を実行
  3. 3台目のサーバ情報を pgpoo-II に登録
  4. オンラインリカバリを実行
  5. 新しいサーバにクエリが振り分けられることをログで確認



まとめ

- PostgreSQL 組み込みのレプリケーション機能
  - ストリーミングレプリケーション
    ＝ 信頼性の高い非同期レプリケーション
- pgpool-II のクラスタリング機能
  - 参照負荷分散
  - クエリの自動振り分け
  - 自動フェイルオーバ
  - オンラインリカバリ
  - 新規ノードの動的な追加



- これらの組み合わせによる

動的にスケールアウト可能な負荷分散データベースクラスタ

SRA OSS, INC.

参考URL

- PostgreSQL ドキュメント
  - http://www.postgresql.jp/document/9.4/html/
- 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/
- PostgreSQL エンタープライズコンソーシアム (PGECons)
  - PostgreSQL の検証報告書
  - http://www.pgecons.org/

SRA OSS, INC.

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

SRA OSS, INC.