SRA OSS, INC.

watchdog を構成する諸機能の内部動作

pgpool-II day 2015
2015/05/15

長田 悠吾
pgpool-II Global Development Group

自己紹介

- 長田 悠吾 (Yugo Nagata)

- pgpool-II 全般の対応
  - 開発、バグ修正、解析、ドキュメント、リリース、yum レポジトリサーバ、ビルドファーム、RPM、 pgpoolAdmin、・・・

- なんとなく watchdog 周りを担当することが多い
  - たまたま、最初に担当したのが watchdog 機能のテスト&デバッグだった

今日のお話

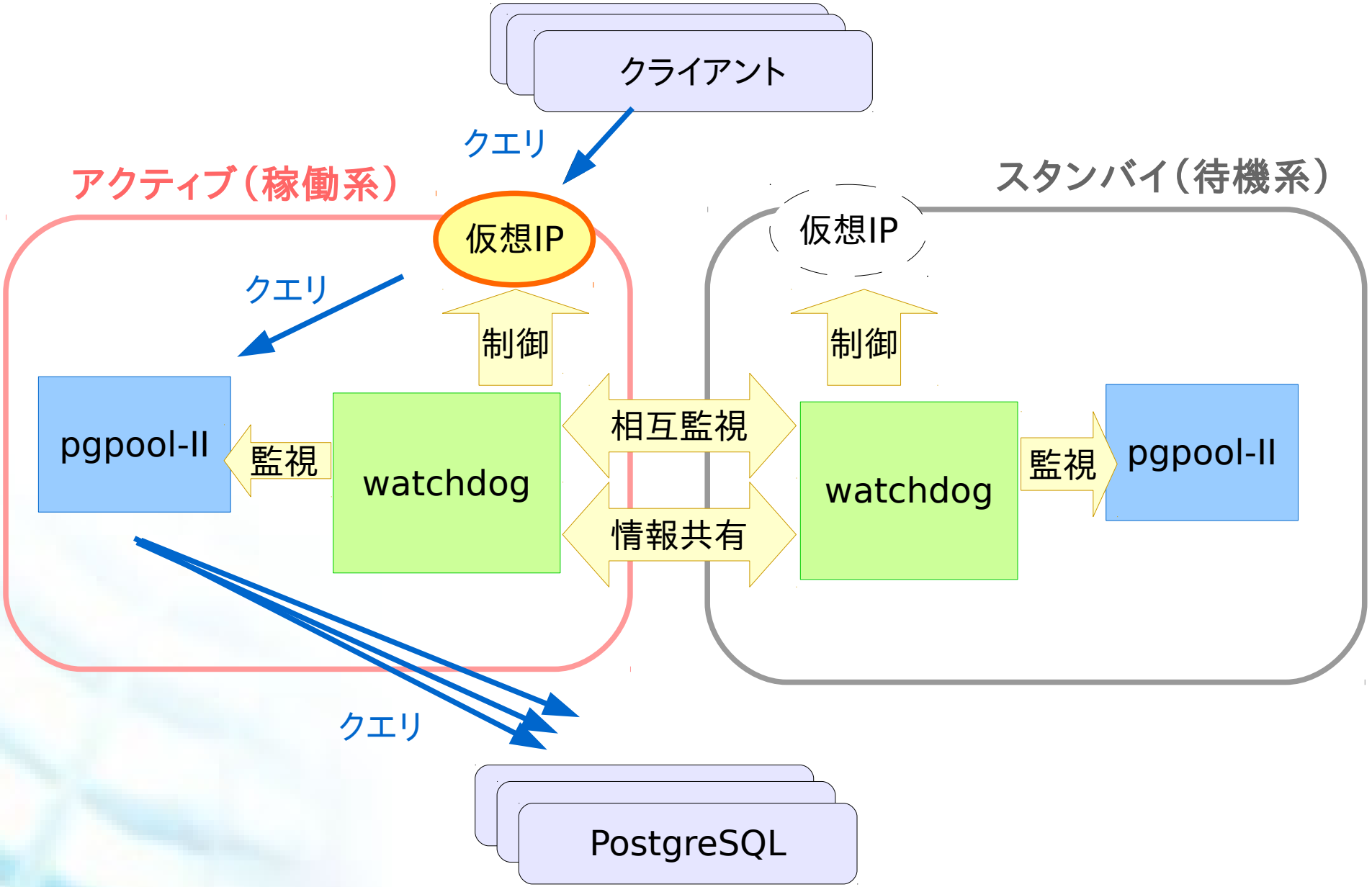
- Watchdog = pgpool-II 組み込みのHA機能
  - pgpool-II 自体が単一障害点（SPoF）になることを防ぐ

- 主な機能
  - pgpool-II の死活監視
  - 仮想 IP の制御
  - pgpool-II 間の情報共有

- その内部動作について、少しだけ詳しく話します
  - どんなプロセスで構成されるか
  - どんな情報をやりとりしているのか
  - どんな処理をしているのか

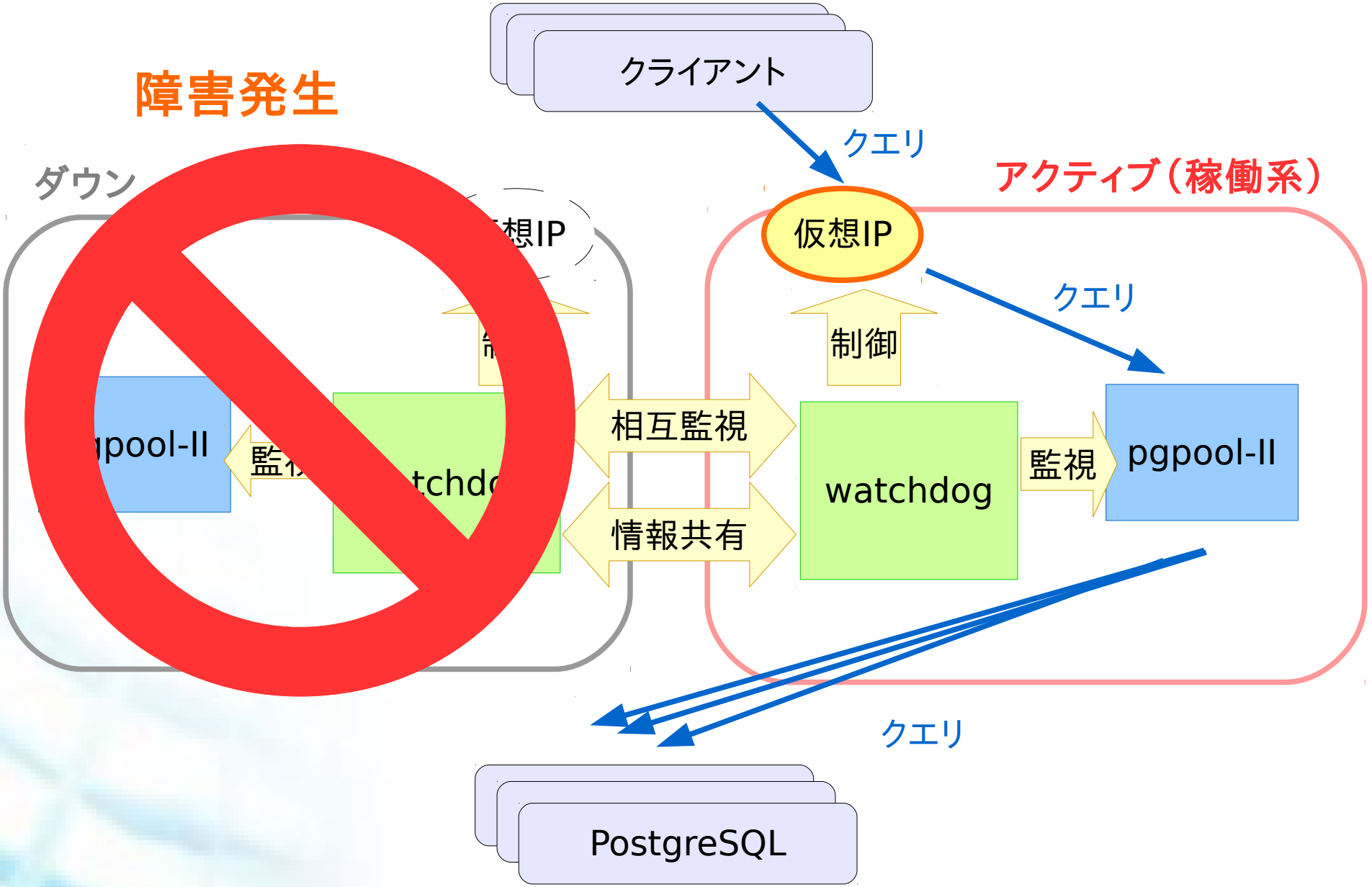
注意）pgpool-II 3.4 を前提にしています。

watchdog 機能概要



SRA OSS, INC.

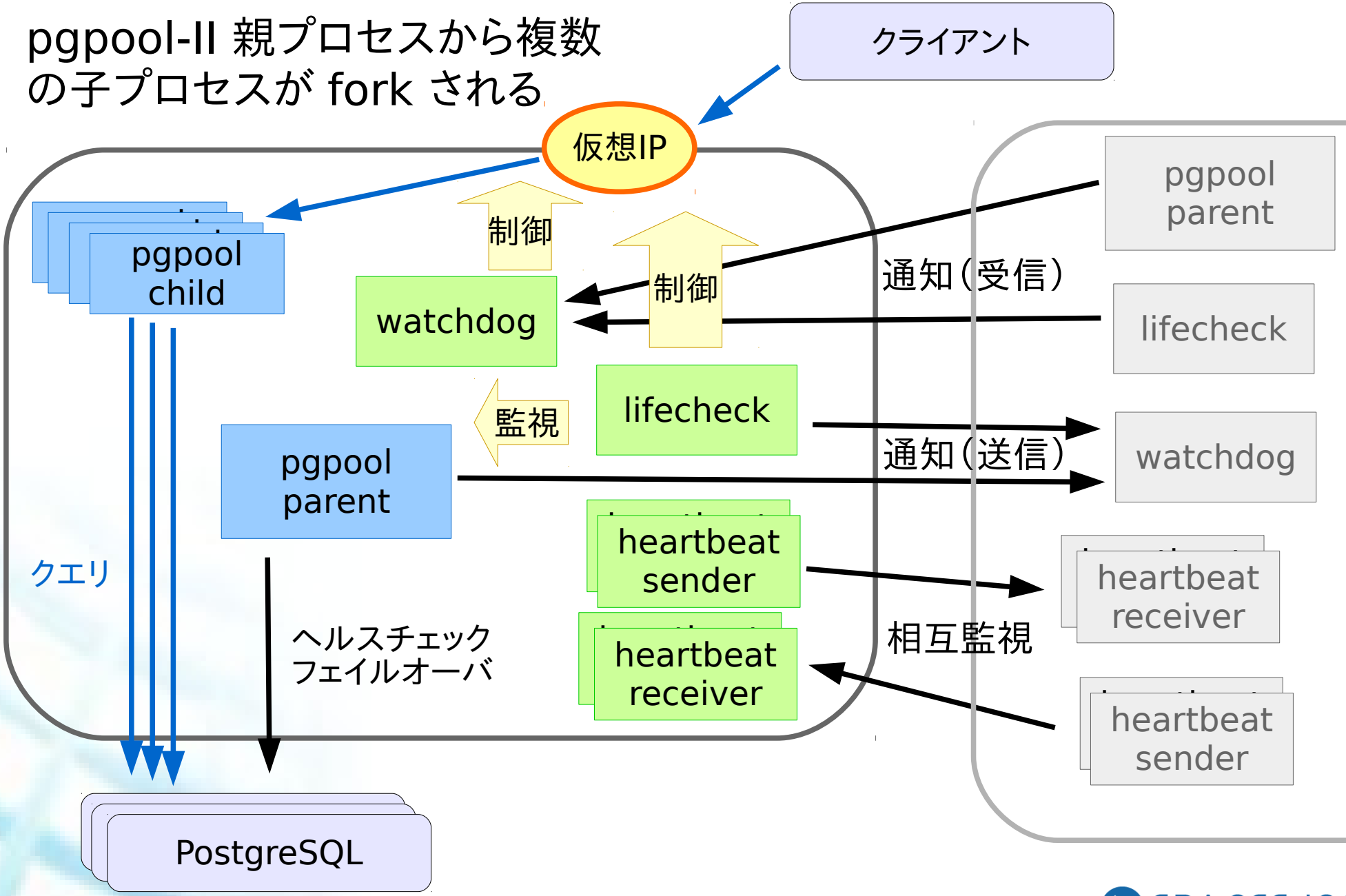
watchdog 機能概要



SRA OSS, INC.

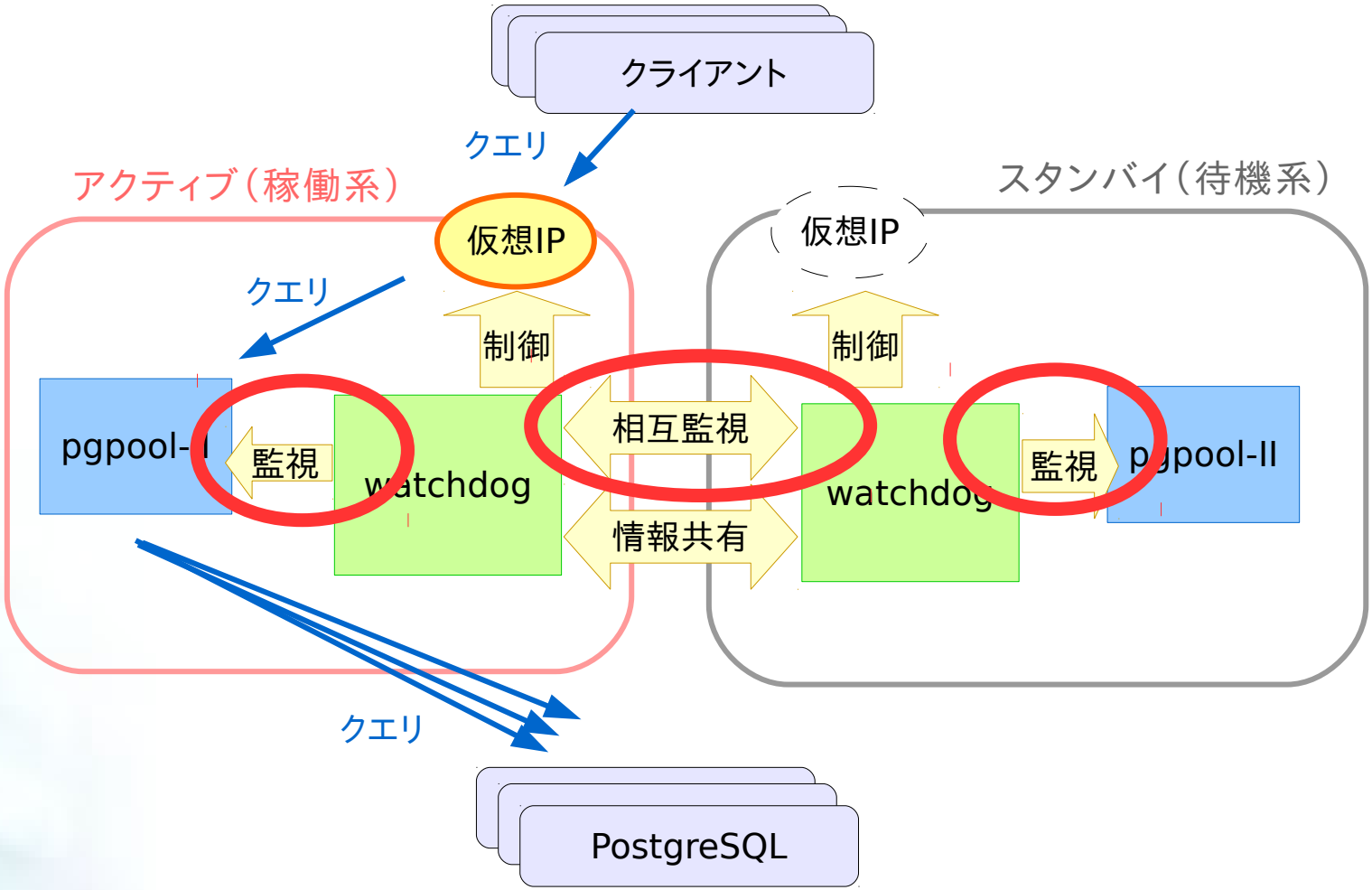
プロセス構成

- pgpool-II 親プロセスから複数の子プロセスが fork される



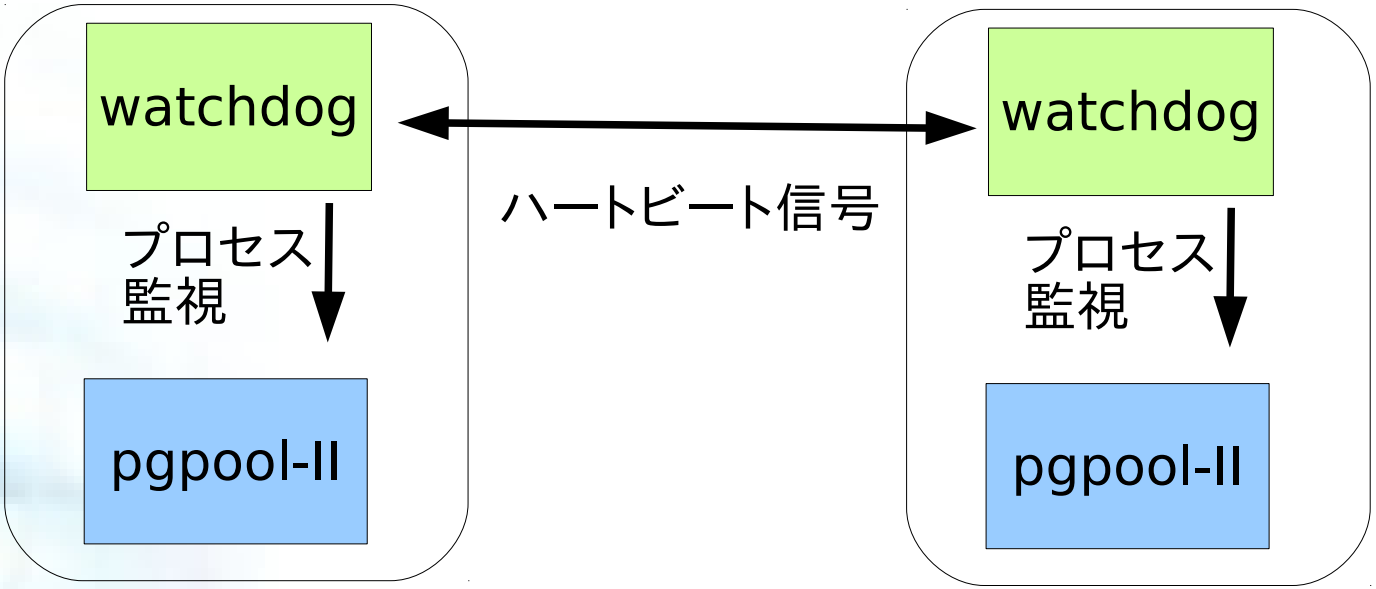
SRA OSS, INC.

死活監視



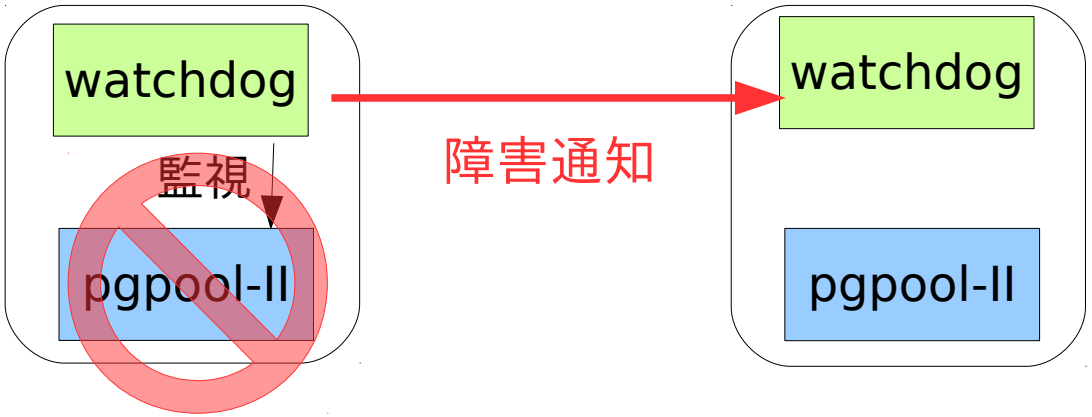
死活監視(life check):概要

- pgpool-II の障害発生の有無を監視する機能
  - 「ハートビート信号」を互いに交換することで、他の pgpool-II の障害を検知
    - 定期的にハートビートを送信
    - これが一定時間以上途切れたら、相手の pgpool-II は死んでいる
  - 自分自身の pgpool-II プロセスが生きているかも、定期的に確認

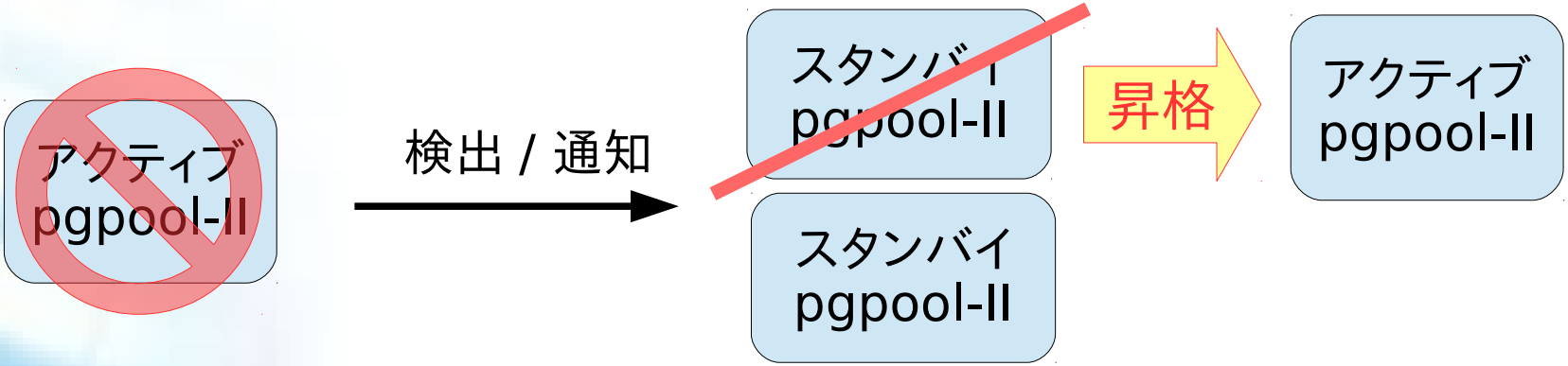


死活監視:障害が発生した場合の挙動

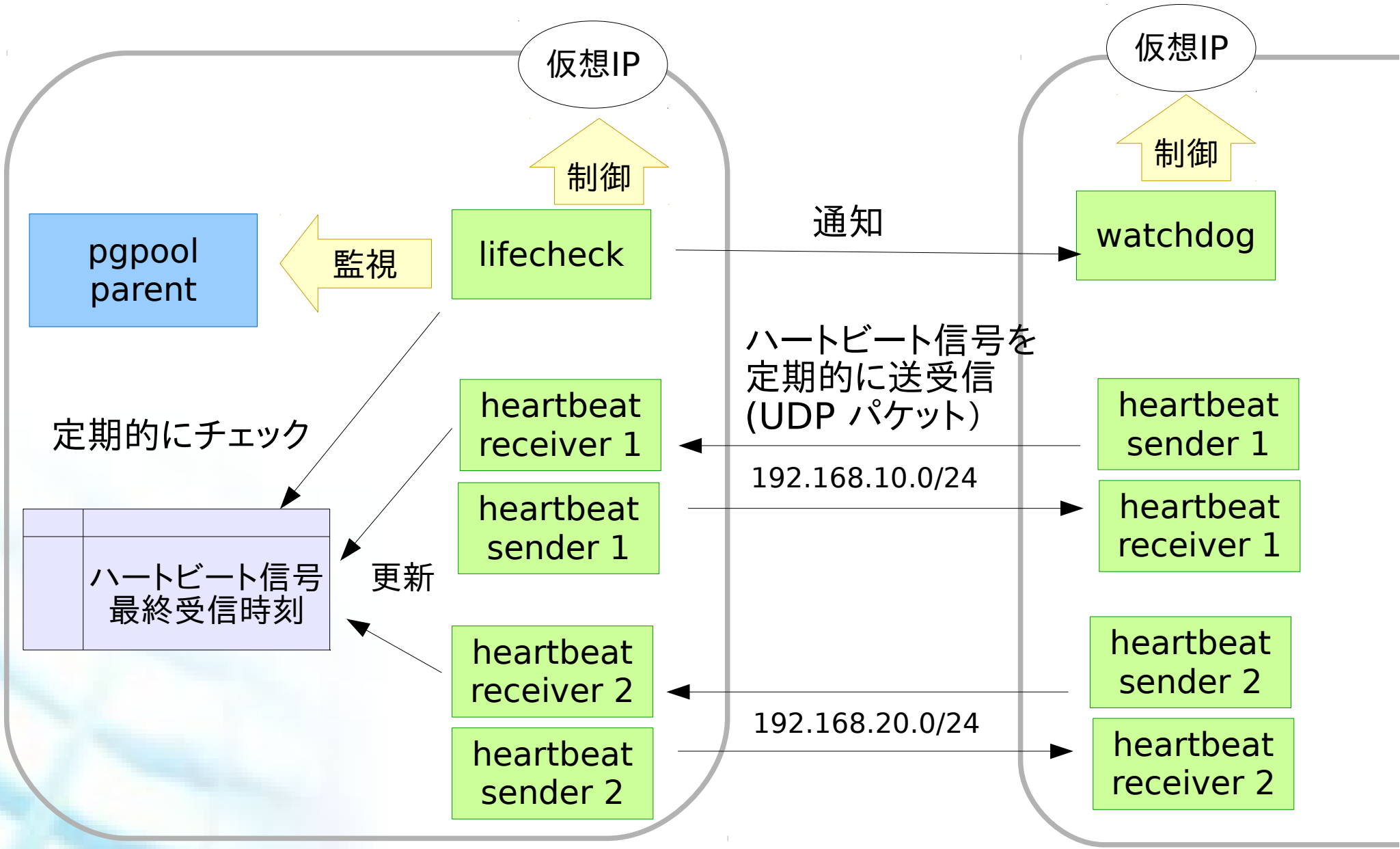
- 自分自身の障害を検出した場合
  - 自分に障害が発生したことを他の pgpool-II に通知



- 自分以外のアクティブ pgpool-II の障害を検出した場合
  - スタンバイのうち1つが、新しいアクティブに昇格する

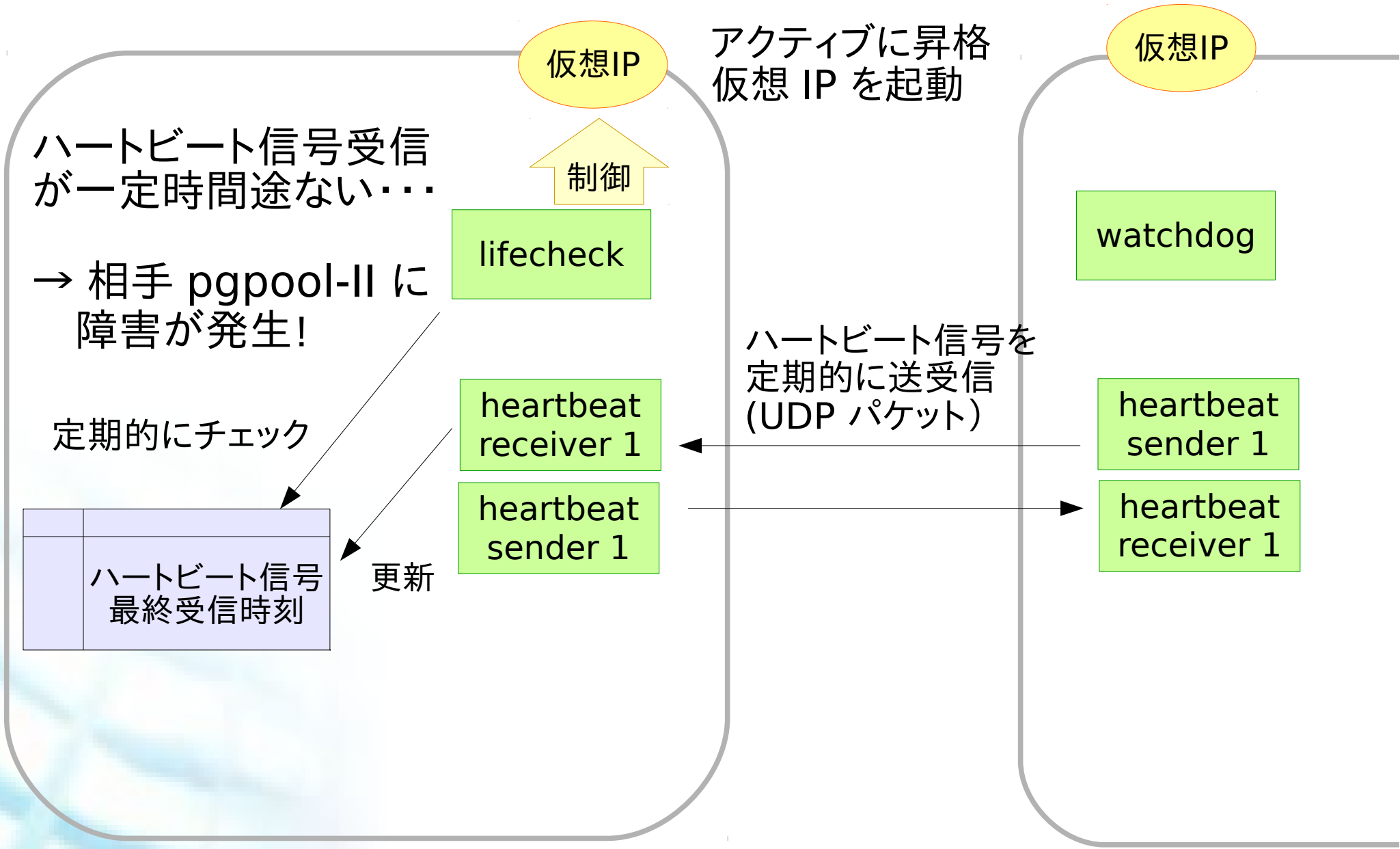


死活監視:全体像

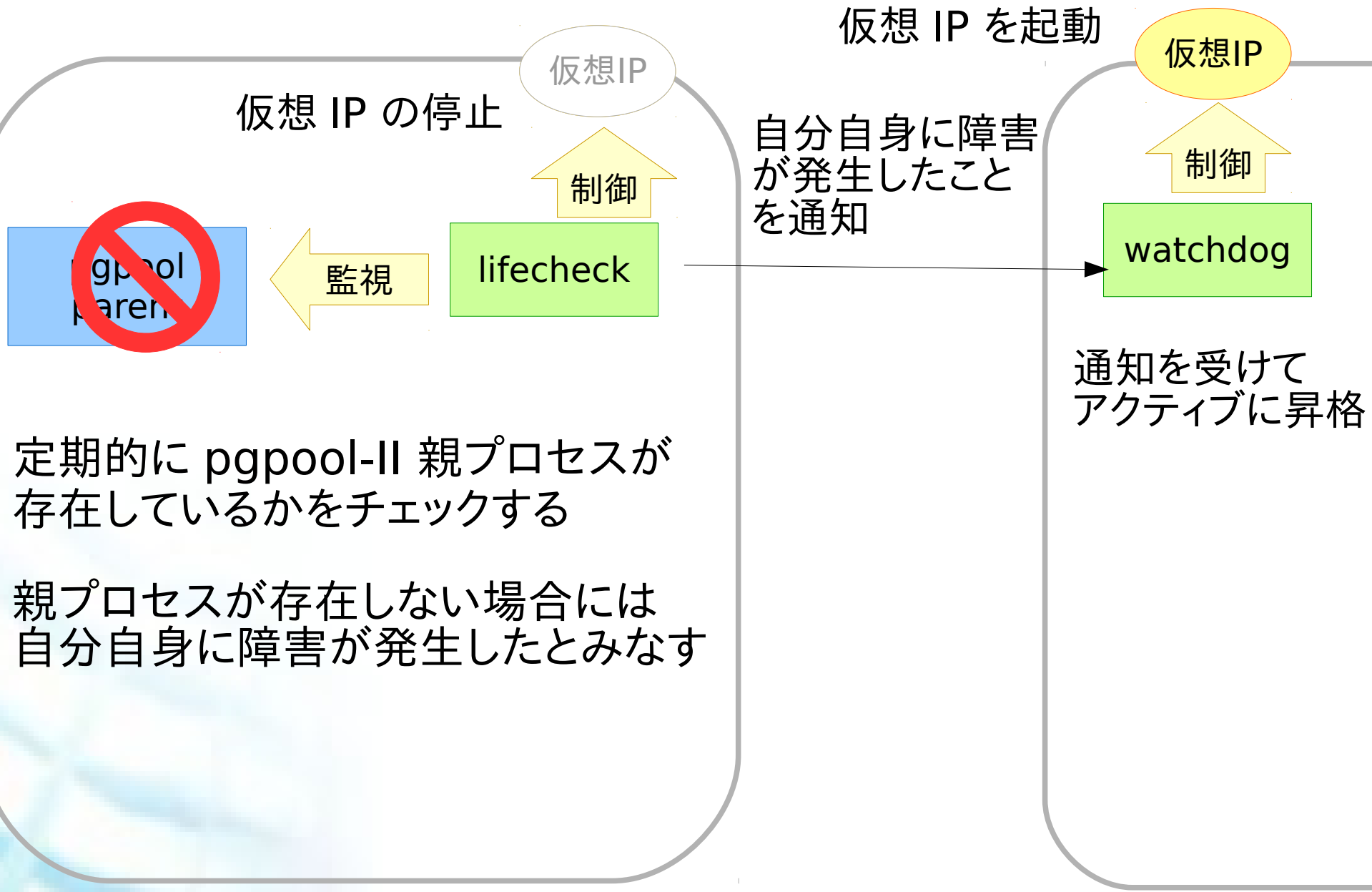


SRA OSS,INC.

死活監視:ハートビート通信(相互監視)



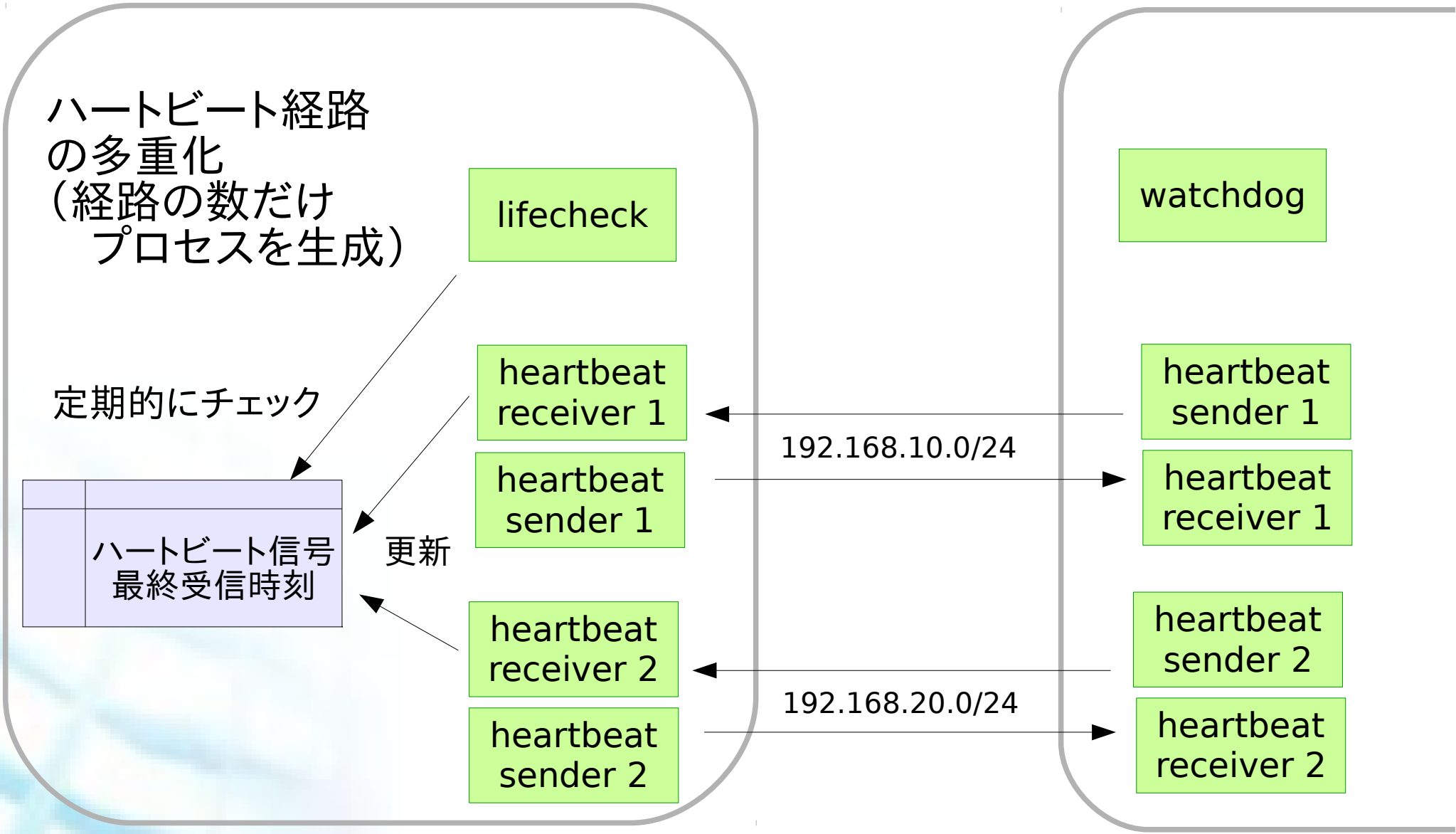
死活監視:プロセス監視(自己監視)



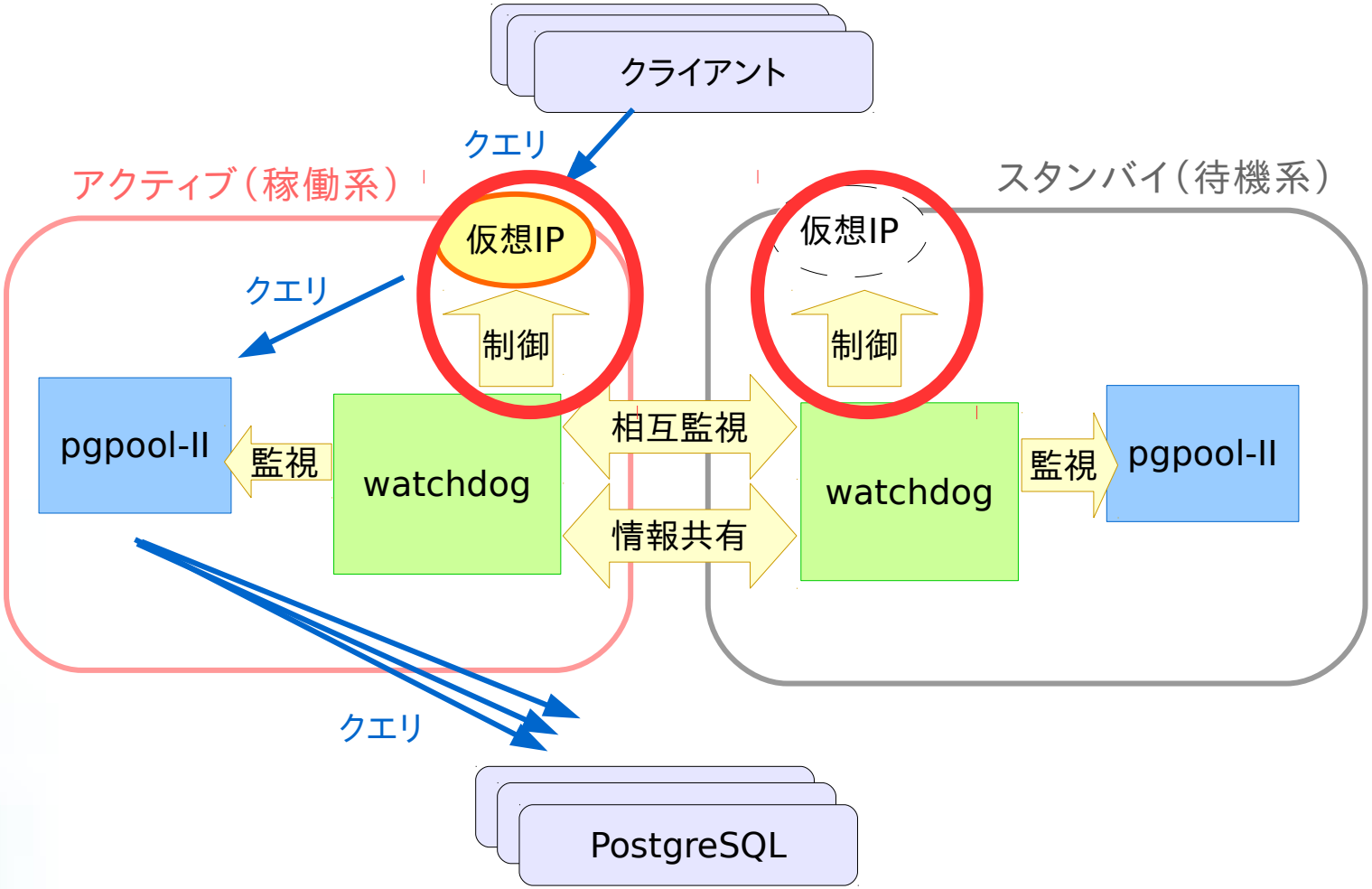
定期的に pgpool-II 親プロセスが
存在しているかをチェックする

親プロセスが存在しない場合には
自分自身に障害が発生したとみなす

死活監視:ハートビート通信の多重化



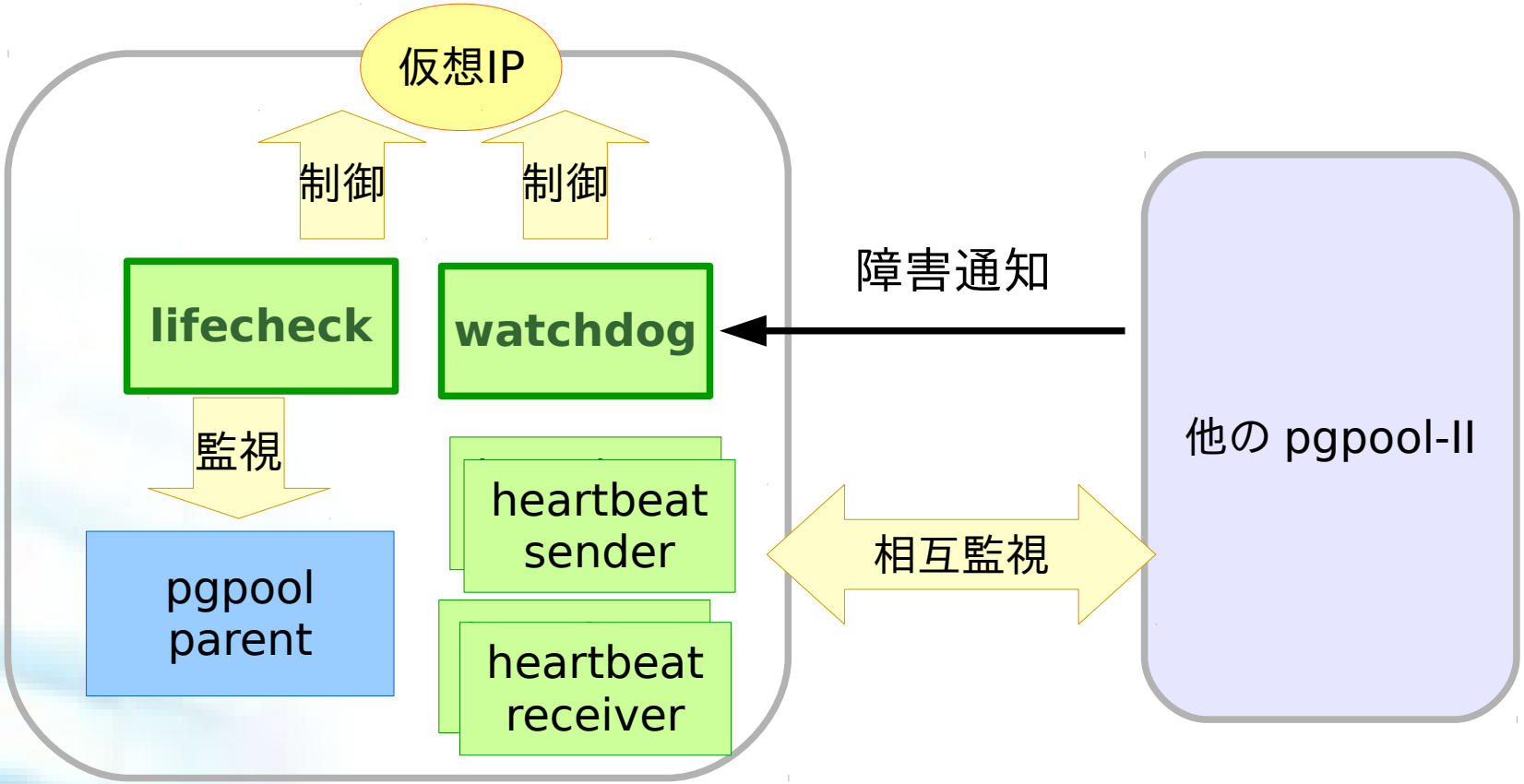
仮想 IP 制御



SRA OSS, INC.

仮想 IP の制御

- 仮想 IP 起動:
  - スタンバイ pgpool-II がアクティブに昇格するとき
- 仮想 IP 停止
  - アクティブ pgpool-II がダウンするとき



仮想 IP の制御:コマンド

- 仮想 IP の起動 / 停止の際に呼ばれるコマンド
  - デフォルトでは ifconfig コマンドを使用
    （次期バージョンからは ip コマンドを使用するよう変更されている）

- 仮想 IP 起動

```
if_up_cmd = 'ifconfig eth0:0 inet $_IP_$ netmask 255.255.255.0'
```

- 仮想 IP 停止

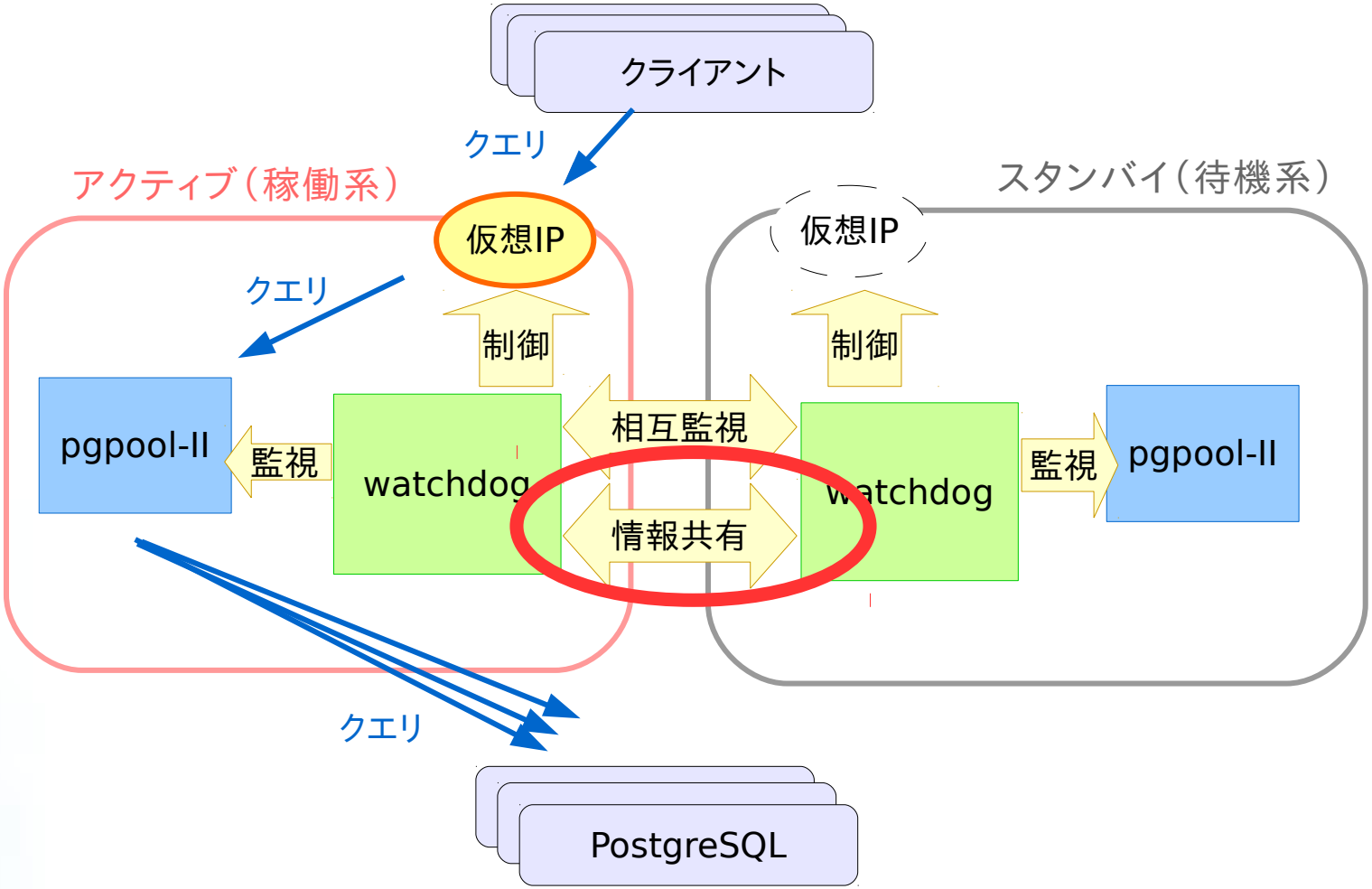
```
if_down_cmd = 'ifconfig eth0:0 down'
```

- 仮想 IP 起動後はネットワーク内のARP キャッシュを更新する

```
arping_cmd = 'arping -U $_IP_$ -w 1'
```

- AWS の CLI などを実行するカスタムコマンドを設定することも可能
  - クラウドでの利用のためには、さらなる検証 & 改良が必要

情報共有



情報共有：pgpool-II サーバ情報

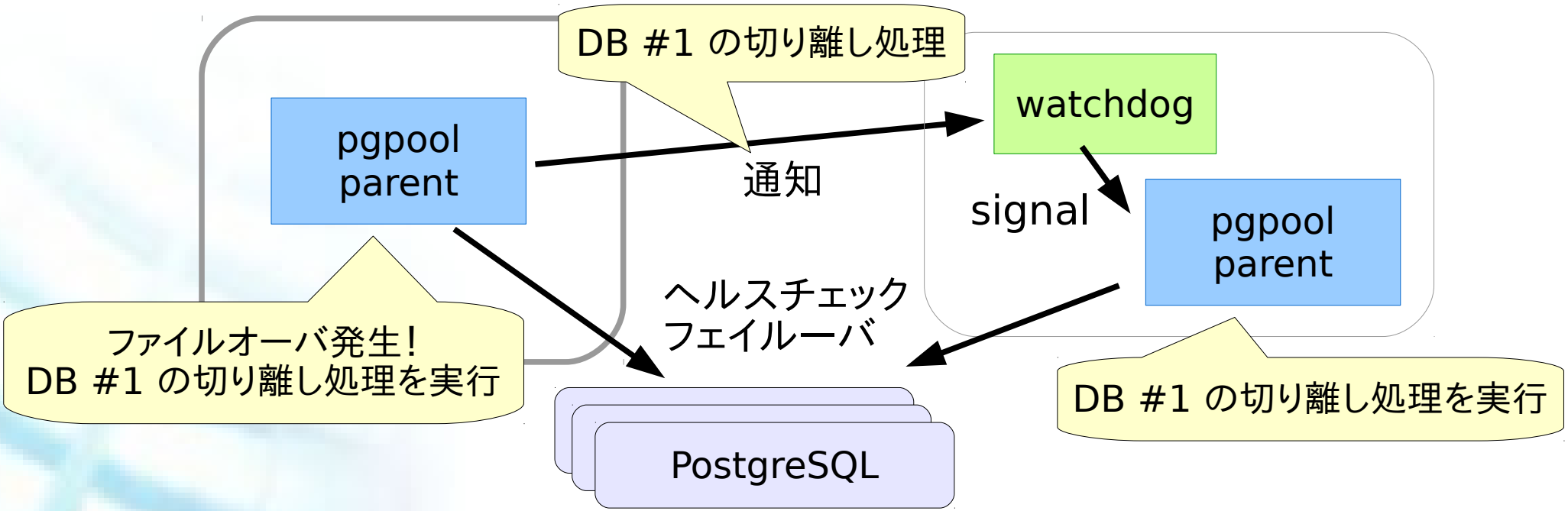
- pgpool-II 間で互いの情報を共有している

  - ホスト名、ポート番号
    - 通信の他、各 pgpool-II の ID 情報として使われる
  - ステータス
    - アクティブ / スタンバイ / ダウン など
  - 仮想 IP 設定
    - 全ての pgpool-II で同じである必要がある
  - 起動時刻
    - 新アクティブを決める際に利用される

- 起動時に、他の pgpool-II に対して自分のサーバ情報を送信する
  - 再起動時には、そのサーバに関する情報が更新される

情報共有:DBノード情報の共有

- DB ノード情報
  - 各バックエンド DB が pgpool-II の管理下にあるか or ないか、
    プライマリ DB か or スタンバイ DB か、などといった情報
- DB のフェイルオーバやフェイルバックなどが発生した時
  - 他の pgpool-II にイベントの内容を通知する
  - 通知の種類:「切り離し」「復帰」「昇格」「オンラインリカバリの開始/終了」

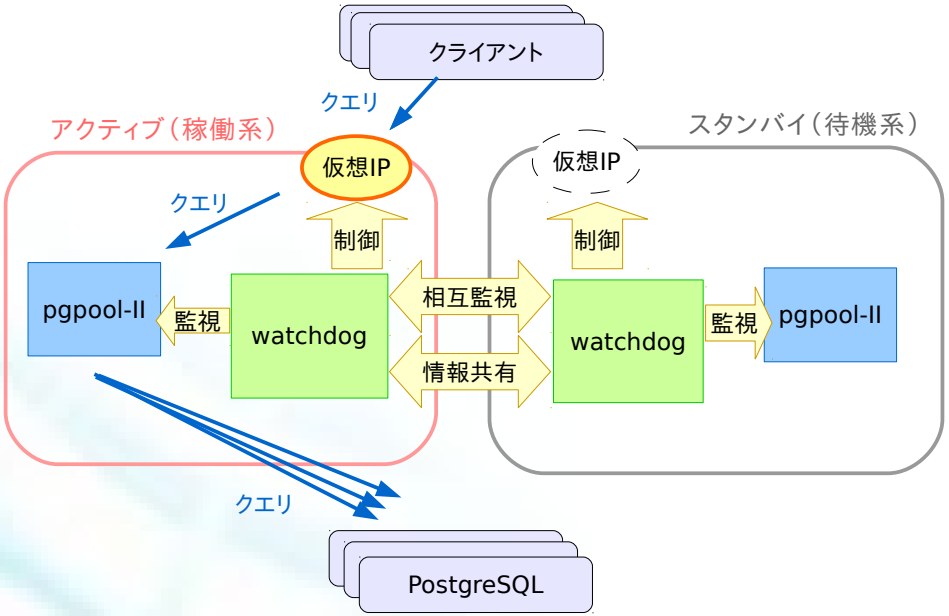


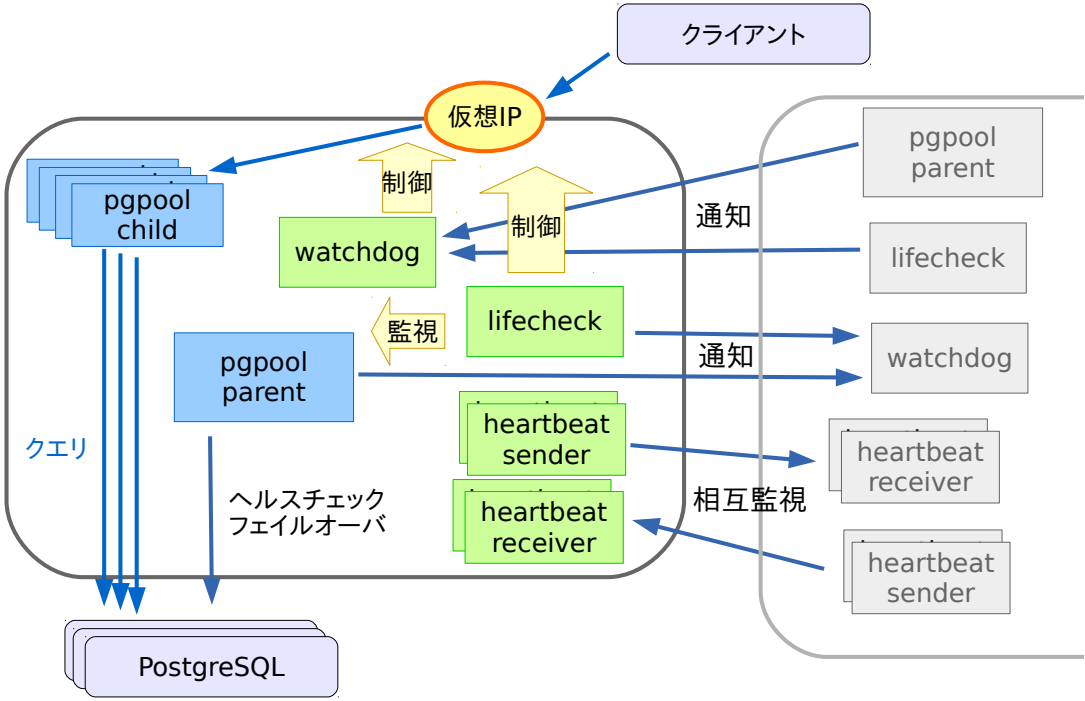
参考情報

- 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.sraoss.co.jp/
- Let's Postgres
  - PostgreSQL 情報のポータルサイト
  - http://lets.postgresql.jp/
- メーリングリスト
  - pgpool-general-jp@sraoss.jp
  - pgpool-general@pgpool.net

最後に

- watchdog の動きが何となくイメージできたでしょうか？
  - 死活監視
  - 仮想 IP 制御
  - 情報共有





- 分からないことがあれば、後で直接でもメールでもご質問ください（可能な範囲で）お答えいたします！

SRA OSS, INC.

予備スライド

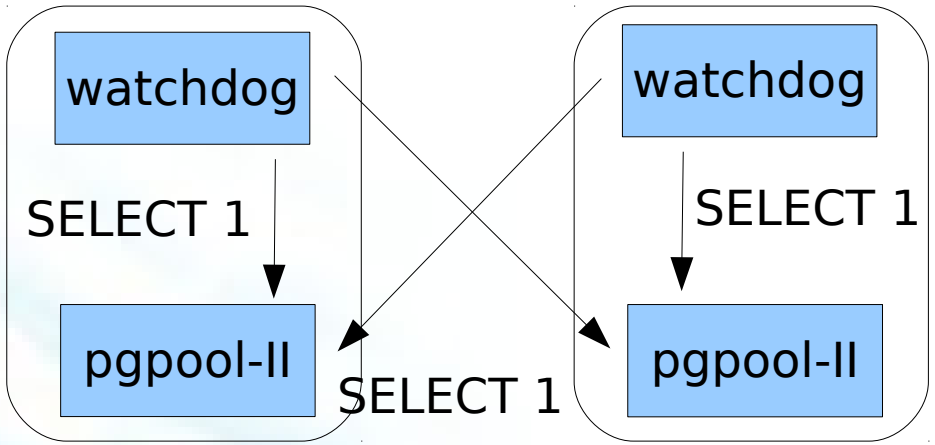
watchdog ステータス

- ステータス
  - アクティブ（稼働系）
    - 仮想 IP を保持している pgpool-II　（全体で1つのみ存在）
  - スタンバイ（待機系）
    - 仮想 IP を保持していない pgpool-II
    - アクティブに昇格することができる
  - ダウン
    - 障害が発生したとみなされている pgpool-II
- pgpool-II 起動時の動作
  - 最初に起動した pgpool-II がアクティブとして起動する
  - 2番目以降に起動した pgpool-II はアクティブに対してクラスタ参加申請
  - これが受理されるとスタンバイとして起動できる

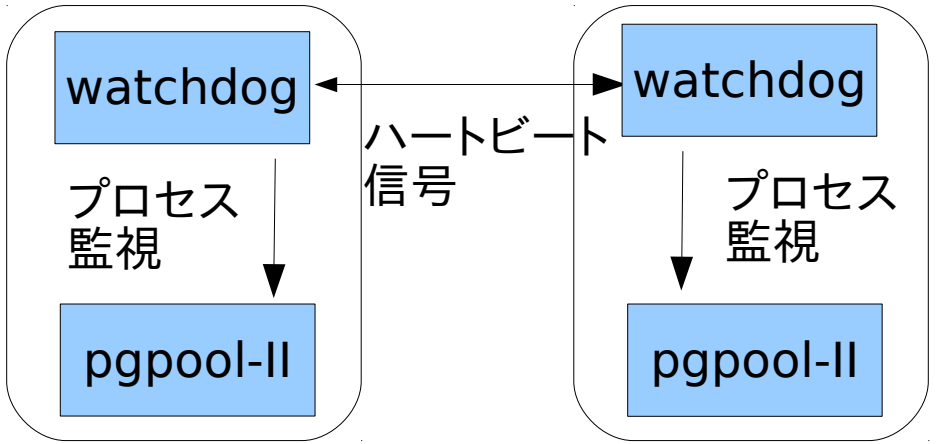
死活監視(life check):概要

- pgpool-II の障害発生の有無を監視する機能
  - 死活監視には2種類のモードがある

- クエリモード(非推奨)
  - 「SELECT 1」などのクエリを発行して pgpool-II の応答をチェック



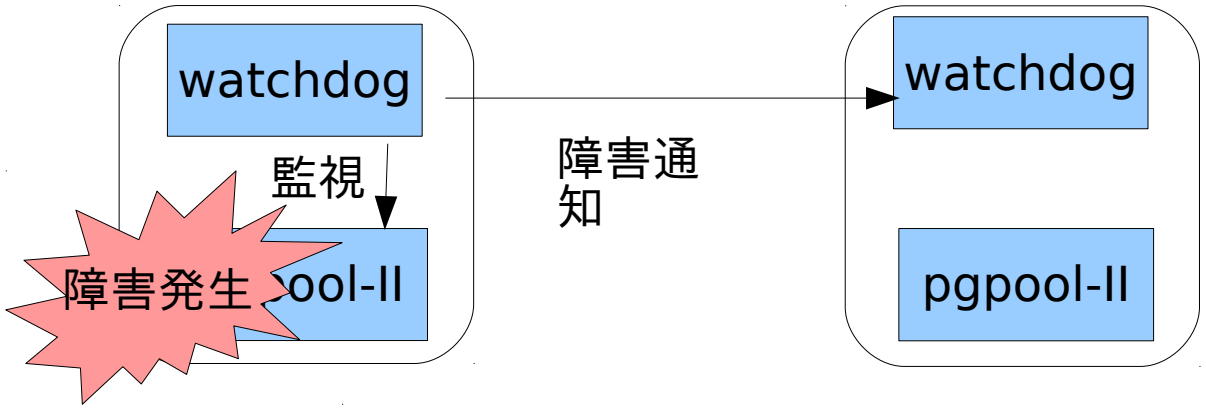
- ハートビートモード(標準)
  - 「ハートビート信号」の交換によって、他の pgpool-II の障害を検知



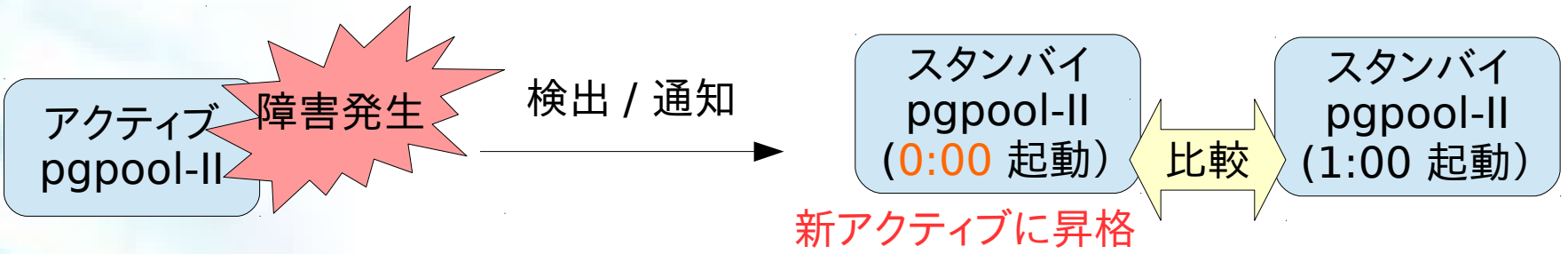
以降は、ハートビートモードの使用を前提に進めます。

死活監視:障害が発生した場合の挙動

- 自分自身の障害を検出した場合
  - 自分に障害が発生したことを他の pgpool-II に通知



- 自分以外のアクティブ pgpool-II の障害を検出した場合、あるいは、アクティブ pgpool-II から障害通知(上述)を受け取った場合
  - スタンバイのうち1つが、新しいアクティブに昇格する
    - 最も起動時間の早い pgpool-II



仮想 IP の制御: root 権限

- 「仮想 IP の制御には root 権限が必要」

1. root ユーザで pgpool-II を起動する
2. sudo 権限のあるユーザで pgpool-II を起動する
   - 仮想 IP 制御コマンドを “sudo ifconfig ...” などに設定
3. ifconfig コマンド等に setuid を設定する

```
# chmod 4755 /usr/sbin/ifconfig
```

   - 一般ユーザが root 権限でコマンドを実行可能になる
   - 実際には「pgpool-II 実行ユーザ専用の ifconfig コマンド」を用意するのがよい

情報共有:セキュアな通信

- セキュリティの問題
  - watchdog 通信のプロトコルを知っていれば、pgpool-IIになりすまして他の pgpool-II に影響を与えることができる
- 対応策
  - 全ての pgpool-II で共通の「認証キー」を設定
    - 認証キーの異なる pgpool-II からの通信は無視する