

AWS上のPostgreSQL+pgpool-IIで 高可用性 / 負荷分散システムを目指す

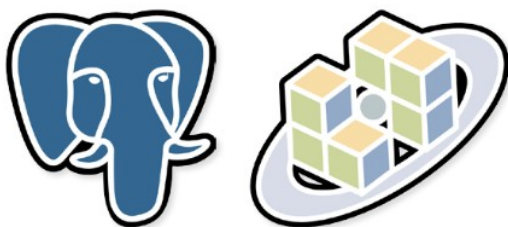
2014年11月8日(土)@新潟オープンソースセミナー2014

SRA OSS, Inc. 日本支社
OSS技術グループ

正野 裕大

自己紹介

- 正野 裕大 (マサノ ユウタ)
 - SRA OSS, Inc. 日本支社 OSS技術グループ
 - 主にAmazon Web Services (AWS) 関連のR&D活動に従事
- 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などのOSSサポート
 - PostgreSQLをベースとしたPowerGresファミリの開発・販売
 - トレーニング、導入・設計コンサルティング



本日お話する内容

- ① Amazon Web Services (AWS) 上でpgpool-IIを使って PostgreSQLのHA構成組みました
- ② 特に難しい話は出てこず、ベーシックな内容と導入事例のご紹介

1. HA構成とは
2. pgpool-IIとは
3. Amazon Web Servicesとは
4. AWS上でのpgpool-II+PostgreSQLの導入事例

本日本話する内容

- ① Amazon Web Services (AWS) 上でpgpool-IIを使って PostgreSQLのHA構成組みました
- ② 特に難しい話は出てこず、ベーシックな内容と導入事例のご紹介

1. HA構成とは

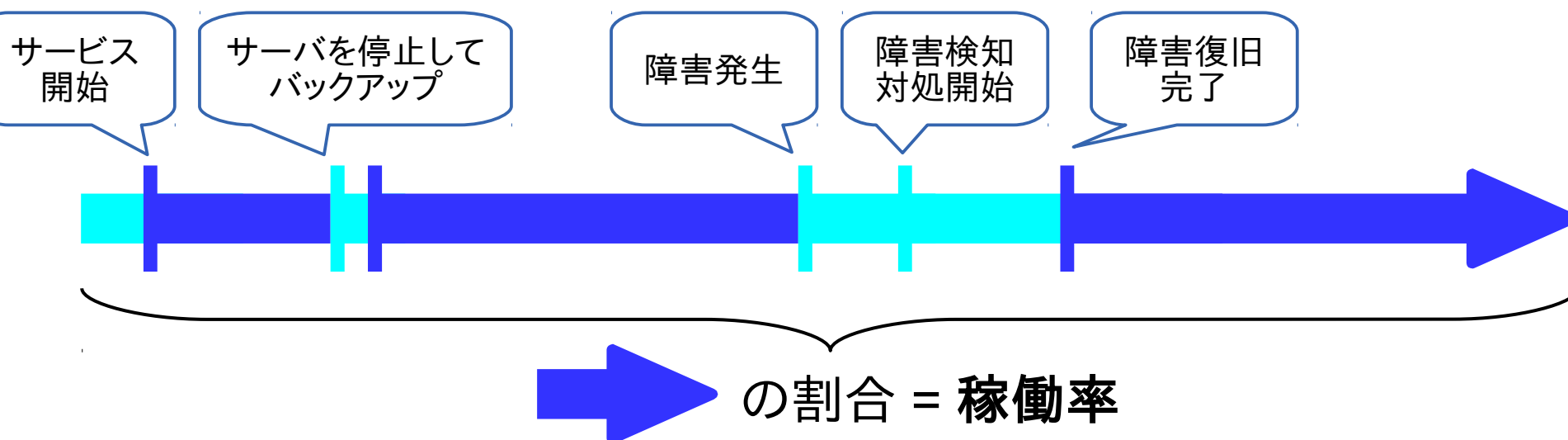
2. pgpool-IIとは

3. Amazon Web Servicesとは

4. AWS上でのpgpool-II+PostgreSQLの導入事例

高可用性 (High Availability) とは

サービスを停止しないで提供し続けられる能力のこと



ミッションクリティカル領域では
可能な限りサービスのダウンタイムを短くしたい

- 障害の発生確率を下げる
 - 構成要素の信頼性を上げるよりも冗長化
- メンテナンスの停止時間を短くする
- 障害検知から復旧するまでの時間を短くする

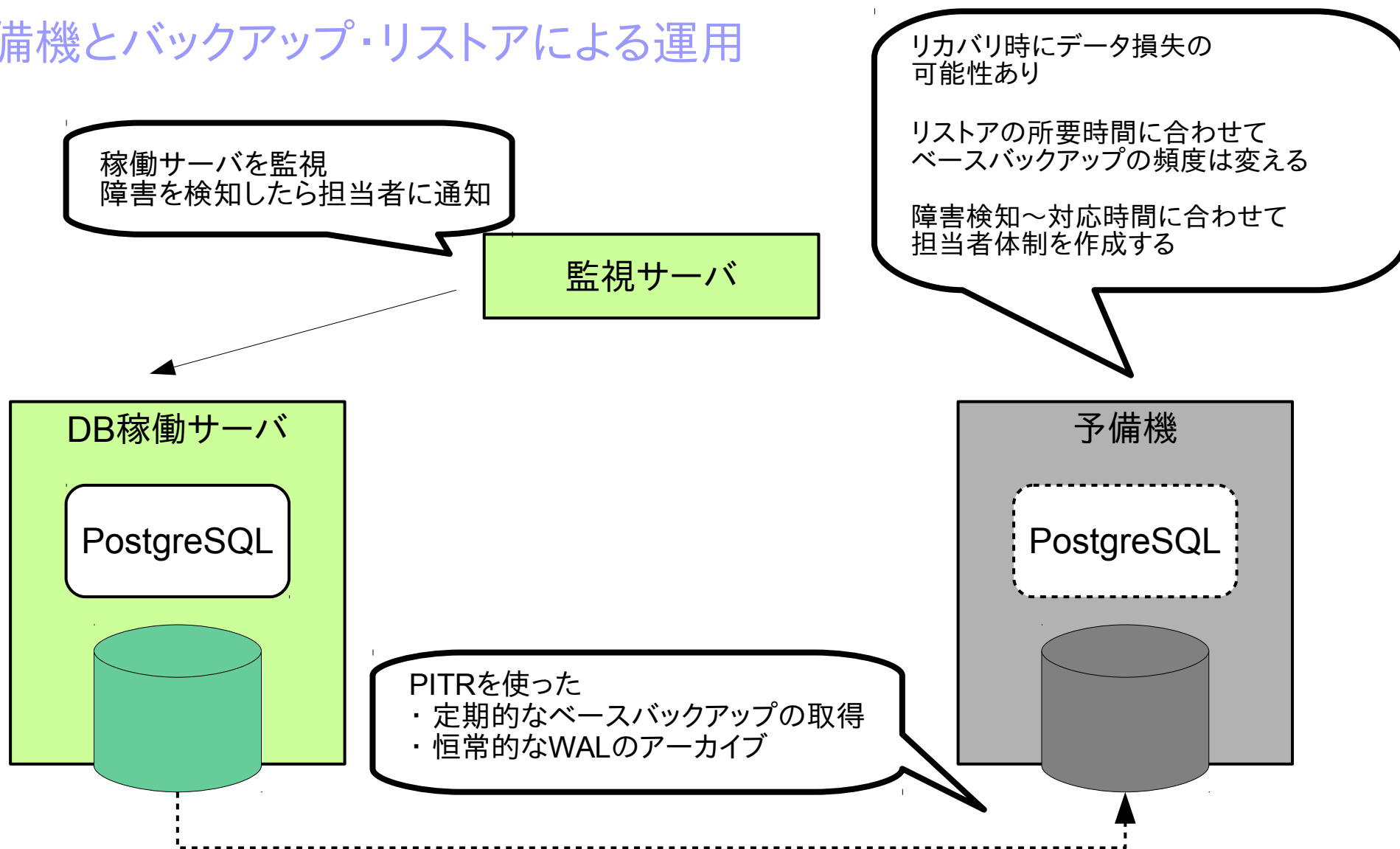
稼働率目標と実現方法

稼働率を高めたいならHA構成は必須

稼働率	年間停止時間	実現方法
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 + 夜間待機
99.99%	52分34秒	HA構成 + 手動切り替え + 夜間待機
99.999%	5分15秒	HA構成 + 自動切り替え
99.9999%	32秒	特殊なハードウェア

PostgreSQLのHA構成:稼働率 ~ 99%

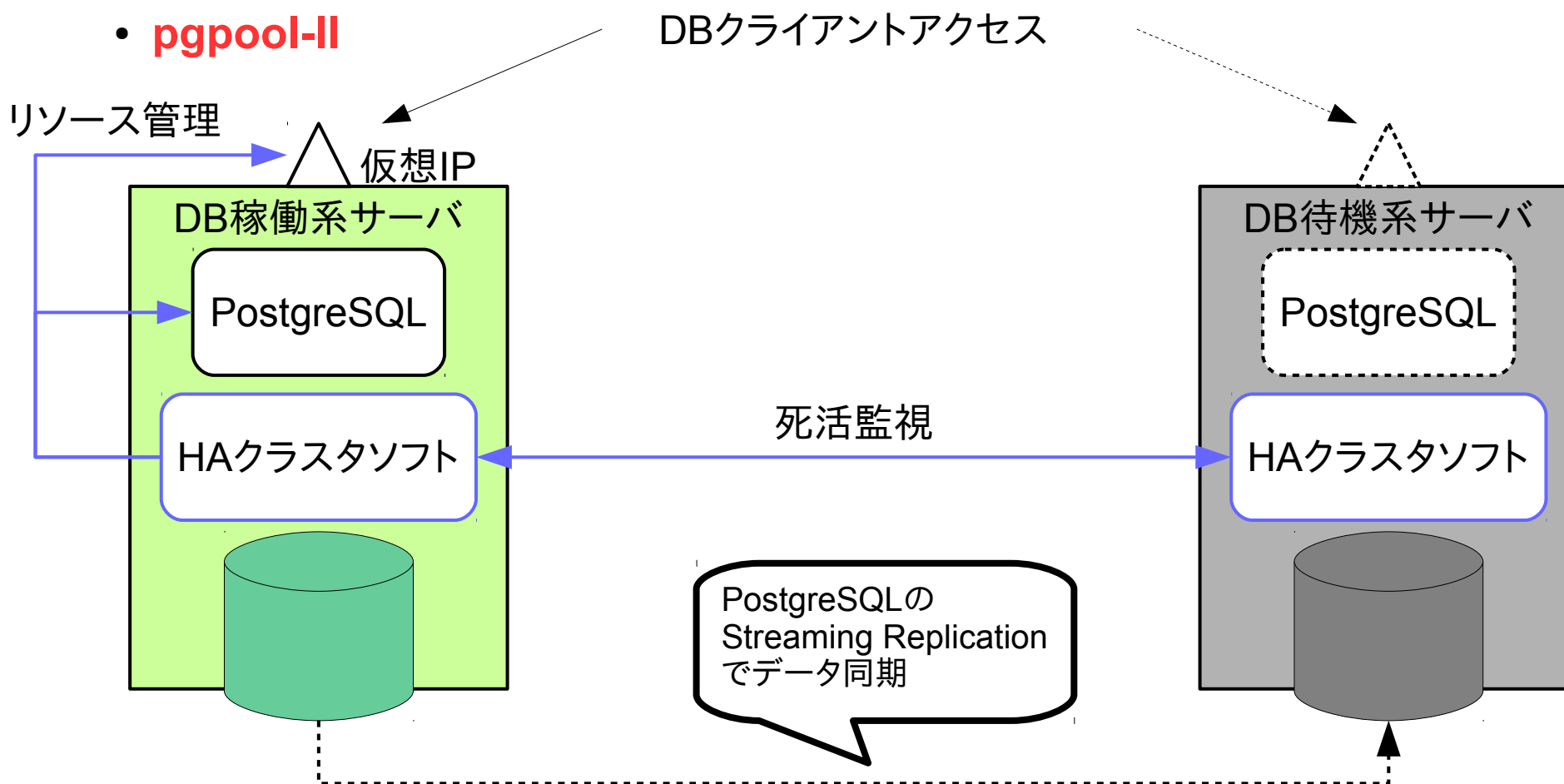
予備機とバックアップ・リストアによる運用



PostgreSQLのHA構成:稼働率 ~ 99.999%

Streaming ReplicationとHAクラスタリングソフトウェアの組み合わせ

- Pacemaker
- **pgpool-II**



本日お話する内容

- ① Amazon Web Services (AWS) 上でpgpool-IIを使ってPostgreSQLのHA構成組みました
- ② 特に難しい話は出てこず、ベーシックな内容と導入事例のご紹介

1. HA構成とは

2. pgpool-IIとは

3. Amazon Web Servicesとは

4. AWS上でのpgpool-II+PostgreSQLの導入事例

pgpool-IIとは

① PostgreSQL専用のクラスタ管理ツール

② BSDライセンスのオープンソースソフトウェア

- pgpool Global development Group が開発
[pgpool Wiki]

<http://www.pgpool.net/mediawiki/jp/index.php/メインページ>

③ 多彩な機能

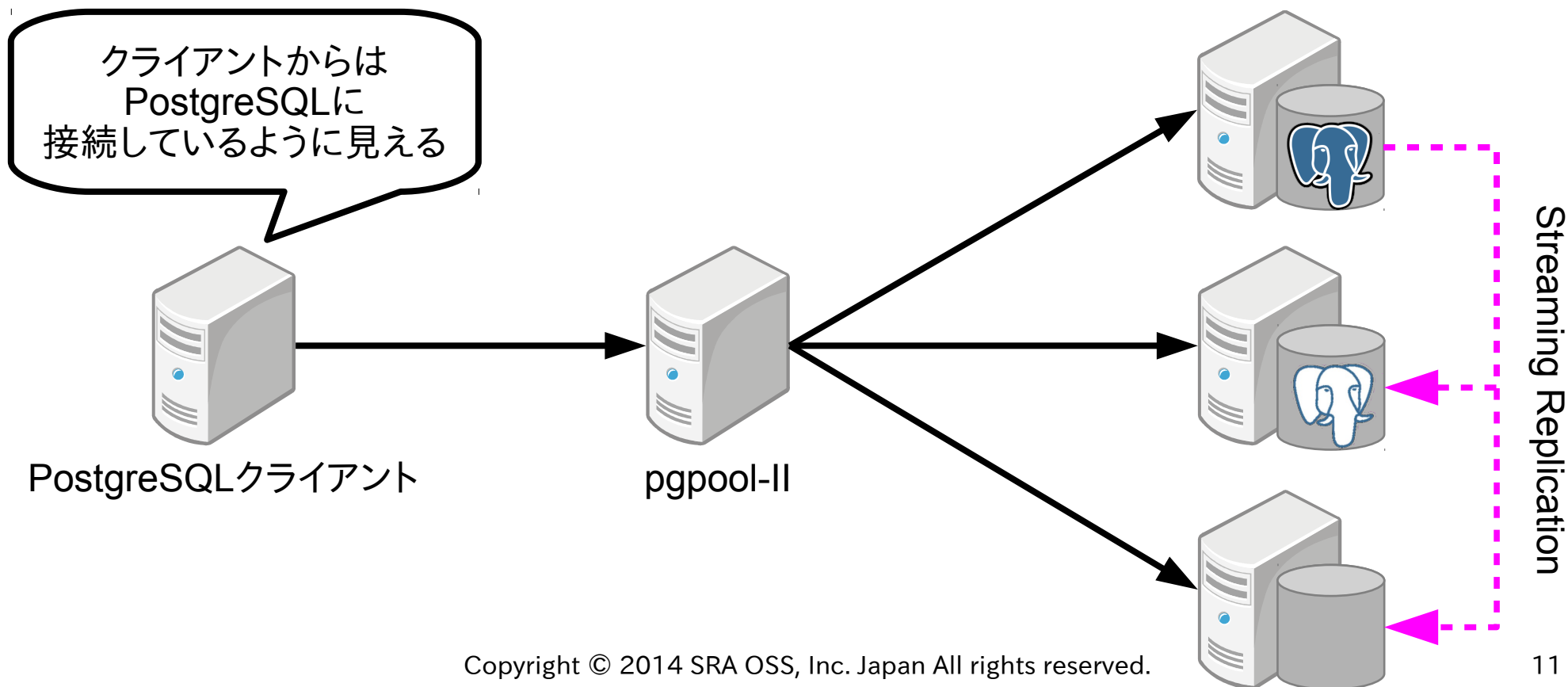
- コネクションプーリング
- 参照負荷分散
- オンメモリクエリキャッシュ
- 自動フェイルオーバー
- オンラインリカバリ
- レプリケーション
- watchdog



pgpool-IIの特徴

クライアントとPostgreSQLの間で動作して様々な機能を提供

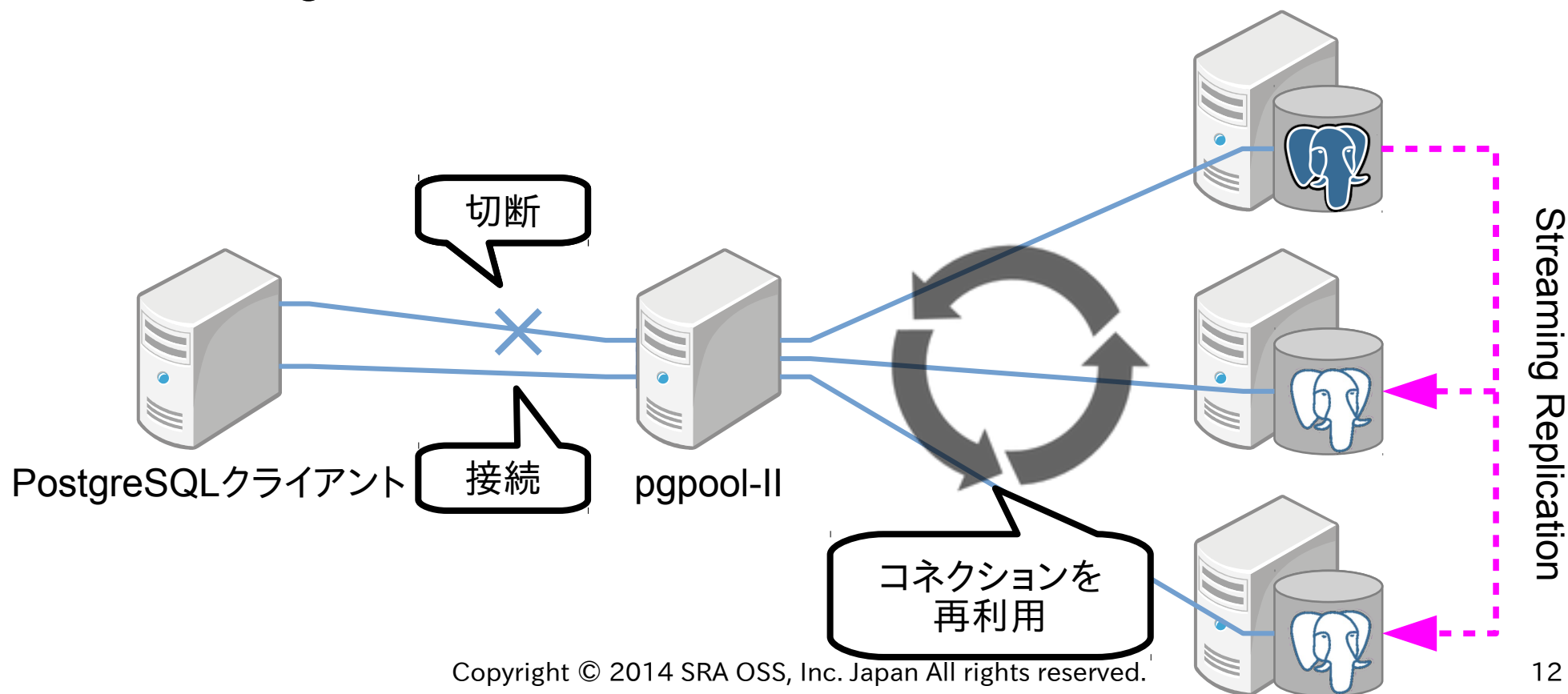
- ✓ Webサーバにおけるプロキシサーバのようなもの
- ✓ クライアントとPostgreSQLを透過的に繋いでクエリを中継
- ✓ 既存のシステムになるべくそのまま導入できるように設計されている



pgpool-IIのコネクションプーリング

クライアントのDB接続オーバーヘッドを抑制する

- ✓ pgpool-IIはpgpool-II～PostgreSQL間のDB接続を保持
- ✓ クライアントがpgpool-IIに接続時に上記の接続を再利用
- ✓ 直接PostgreSQLに接続するよりオーバーヘッドが小さくなる

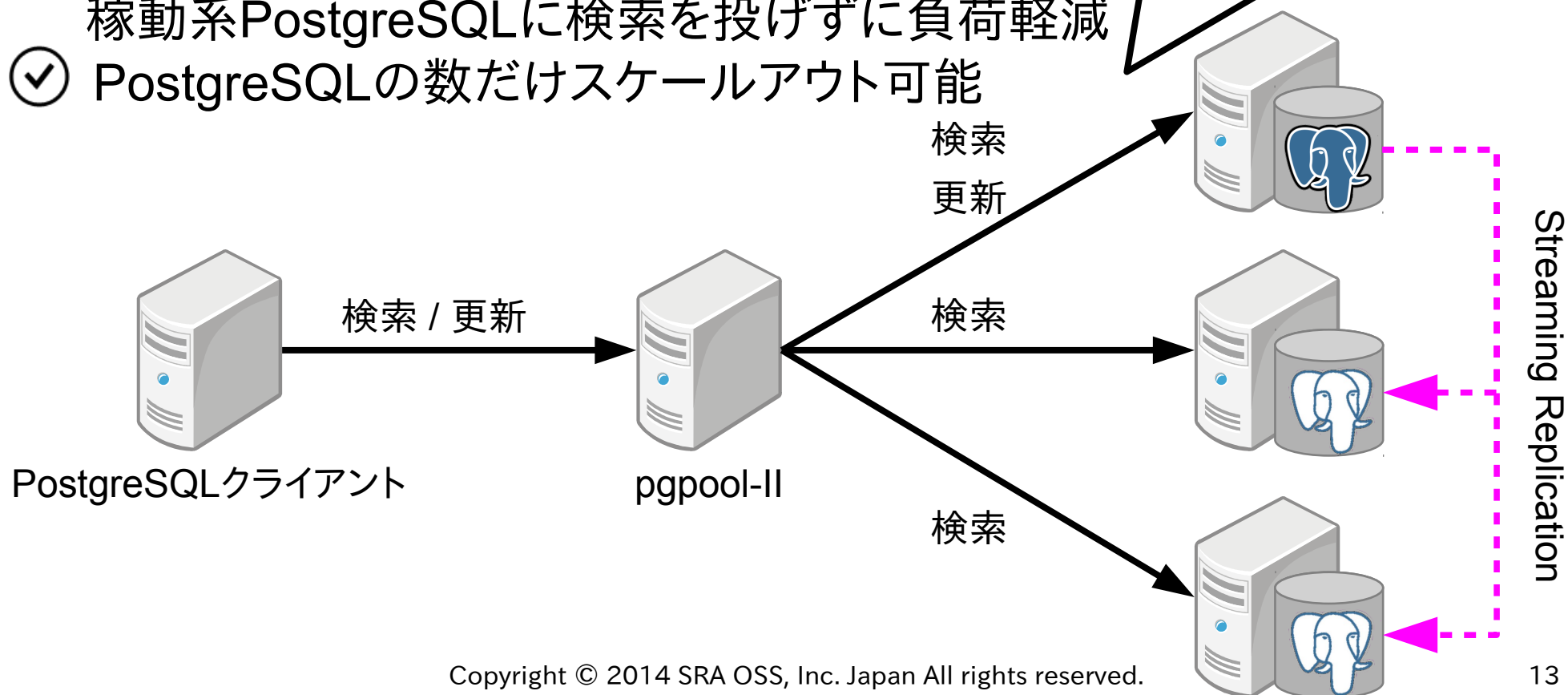


pgpool-IIの参照負荷分散

クライアントのクエリを自動的に振り分けて負荷を分散

- ✓ 更新系クエリは稼動系PostgreSQLに
- ✓ 参照系クエリはいずれかのPostgreSQLに
- ✓ 更新が重いシステムでは稼動系PostgreSQLに検索を投げずに負荷軽減
- ✓ PostgreSQLの数だけスケールアウト可能

参照系クエリの振り分け先は各DBサーバに対する重み付けで相対的に決められる



pgpool-IIの参照負荷分散

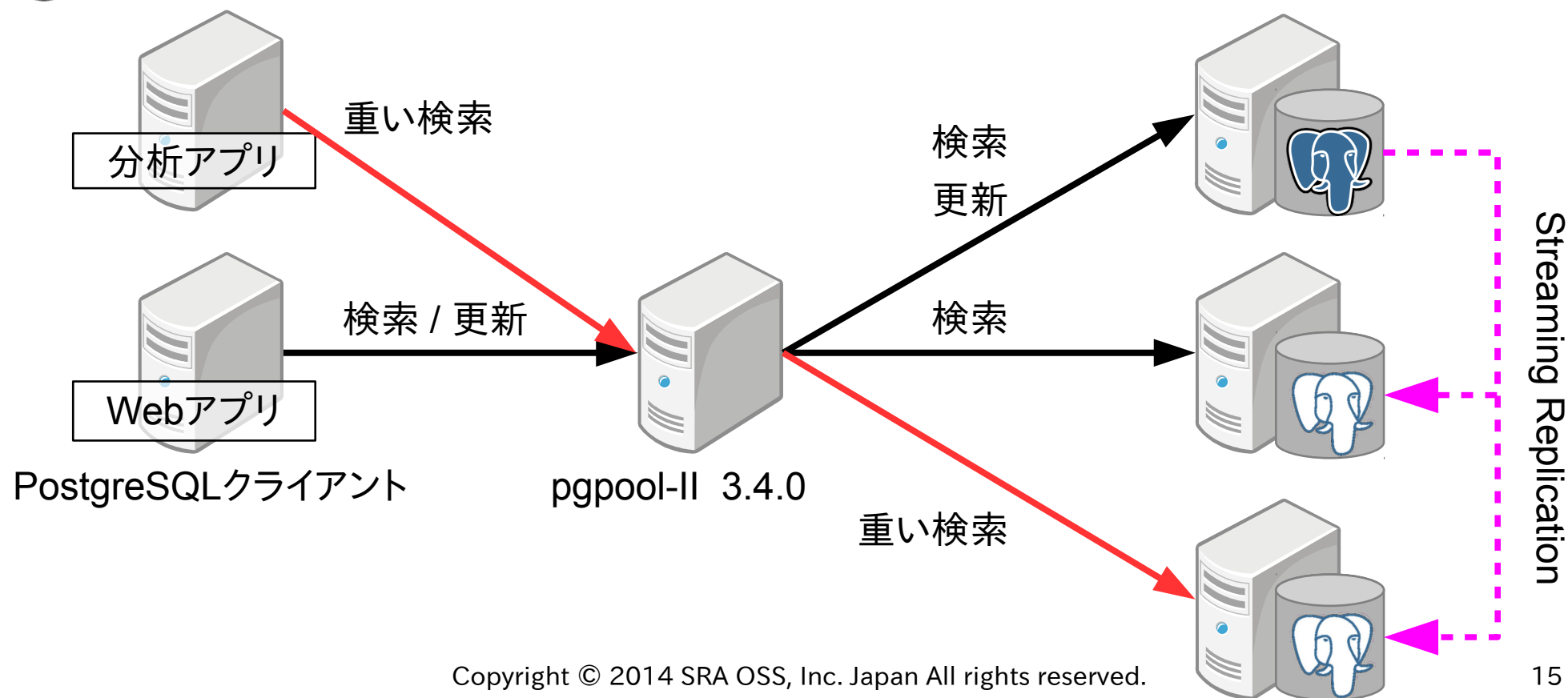
新バージョンのpgpool-II 3.4.0が昨日 (2014-11-07 Fri) リリースされました！



pgpool-IIの参照負荷分散

新バージョンのpgpool-II 3.4.0ではきめ細かい参照負荷分散が可能に

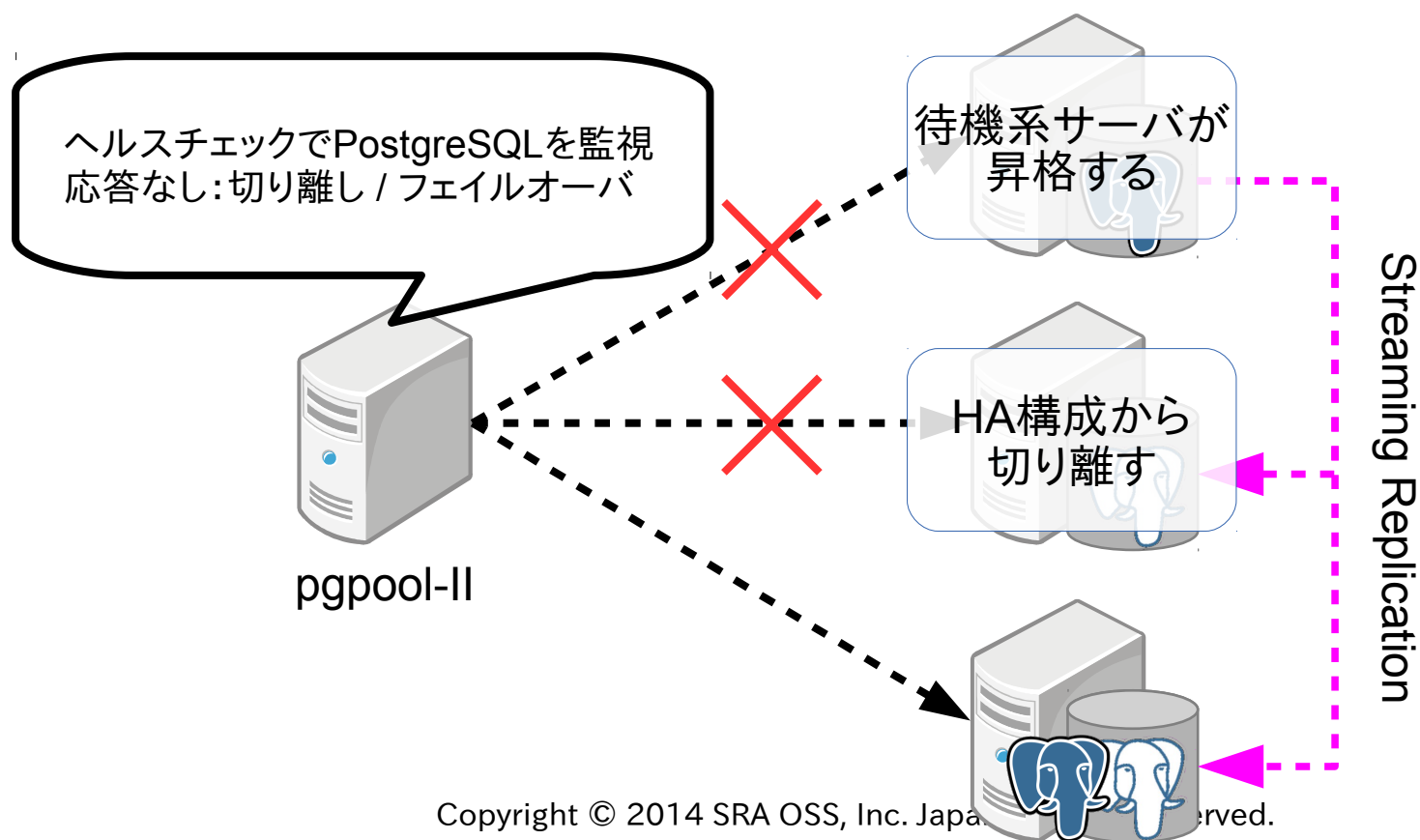
- ✓ アプリケーション名、DB名によって接続先を指定できる
- ✓ 分析用の重い参照系クエリを分析専用PostgreSQLに送信
- ✓ サービスに影響を与えなくする



pgpool-IIの自動フェイルオーバー

障害の発生したPostgreSQLを検知してHA構成から切り離し / 待機系昇格

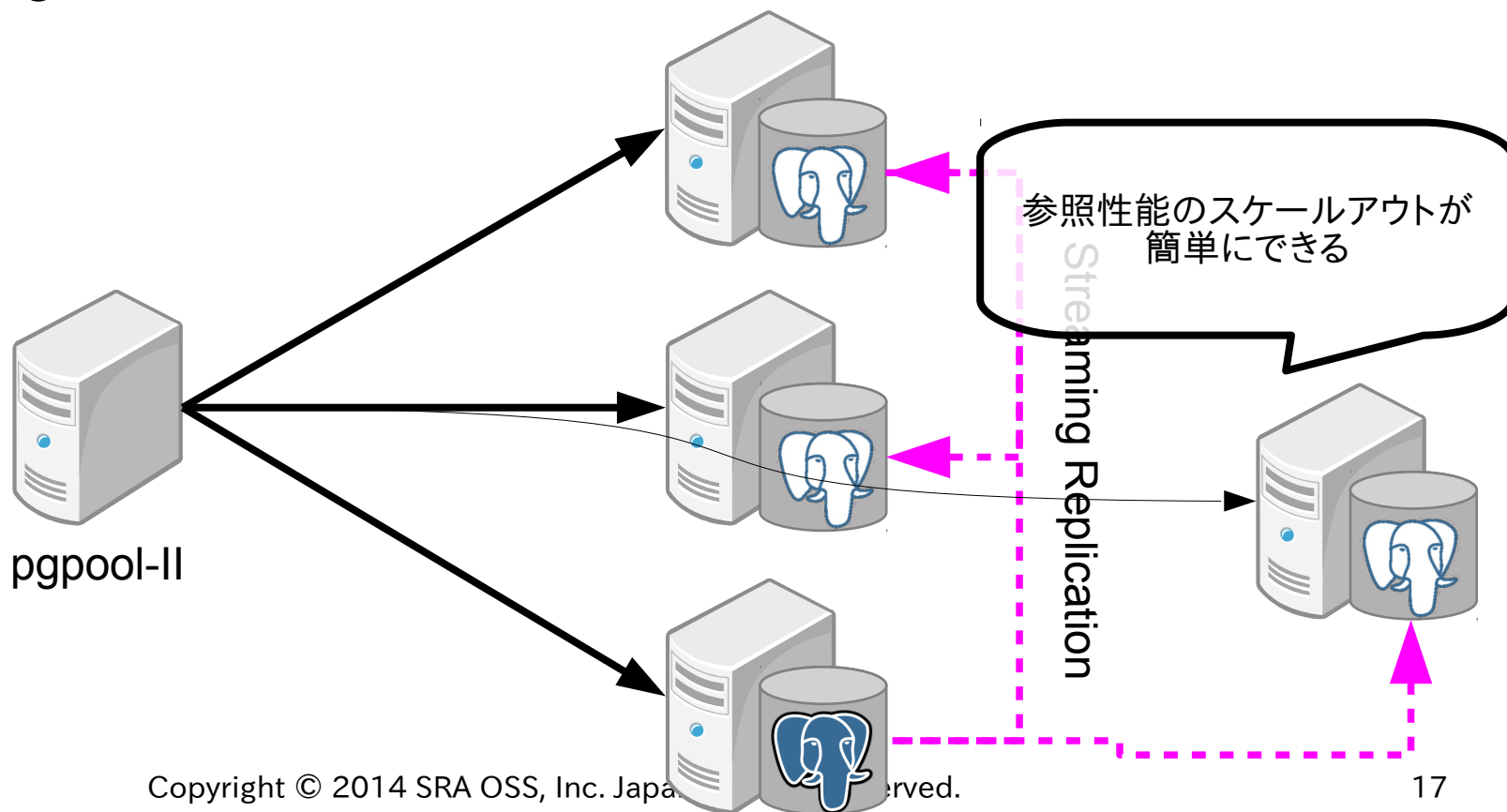
- ✓ 稼働系ダウン: 待機系PostgreSQLが自動的に稼働系に昇格
- ✓ 待機系ダウン: HA構成から切り離し



pgpool-IIのオンラインリカバリ

稼働系PostgreSQLを利用してダウンした待機系PostgreSQLを作成

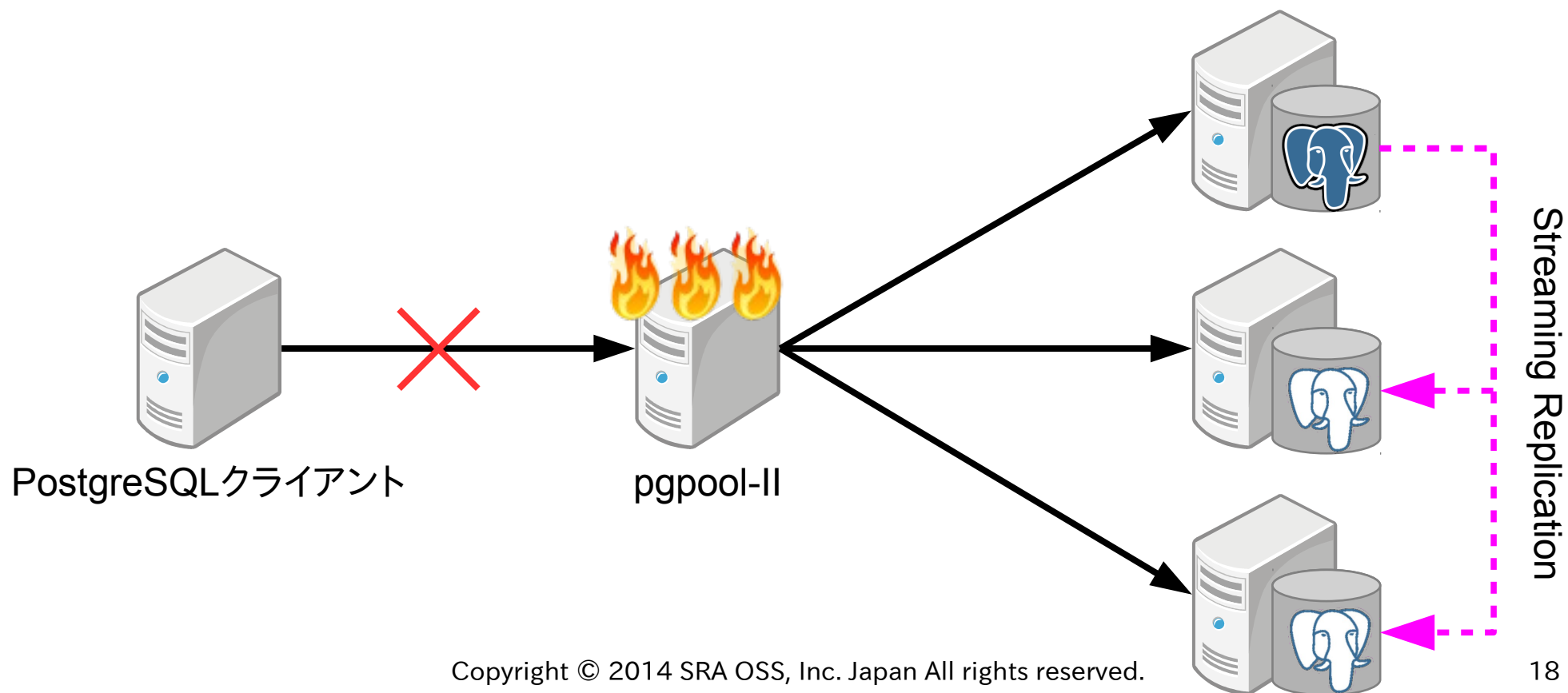
- ✓ ダウンした待機系を復旧させる
- ✓ 新しい待機系を作成する
- ✓ 稼働系PostgreSQLを停止させる必要はない



pgpool-IIのwatchdog

pgpool-II自体の可用性を高める機能

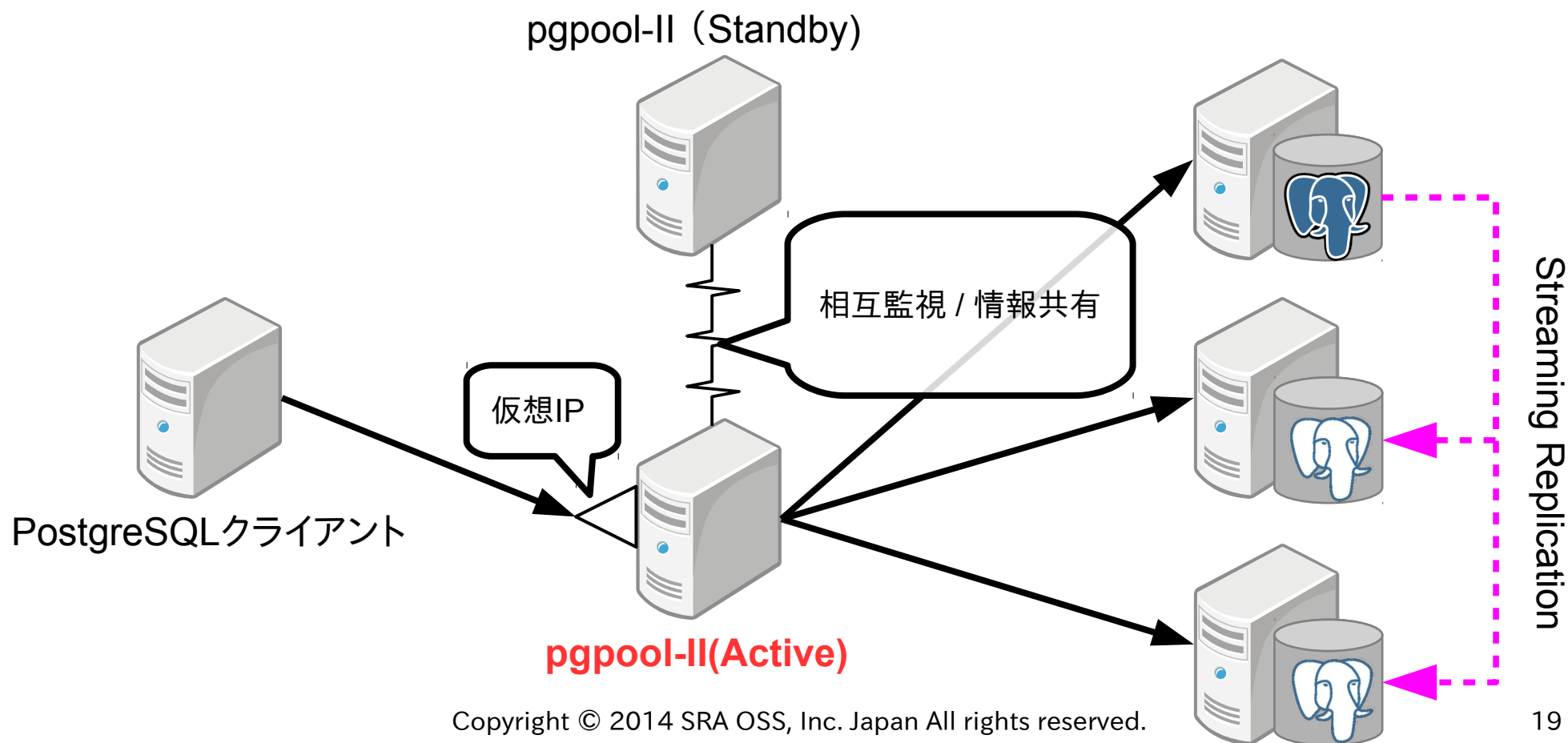
- ✓ pgpool-IIが単一障害点
- ✓ pgpool-IIがダウンしたらサービス停止



pgpool-IIのwatchdog

pgpool-II自体の可用性を高める機能

- ✓ pgpool-IIをActive/Standby構成にして単一障害点を回避
- ✓ クライアントはActive pgpool-IIに仮想IPでアクセスする

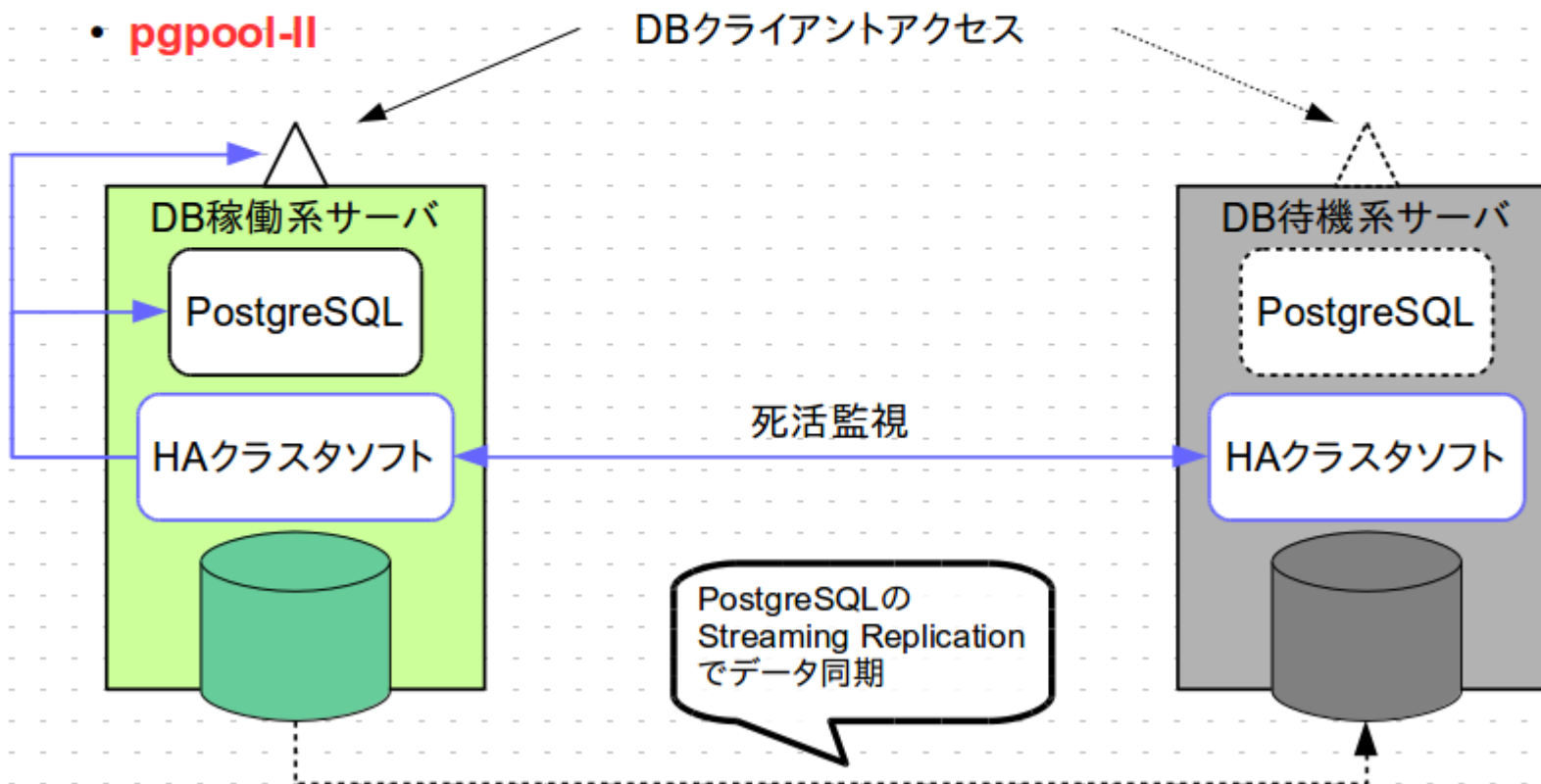


PostgreSQLのHA構成:稼働率 ~ 99.999%

Streaming ReplicationとHAクラスタリングソフトウェアの組み合わせ

- Pacemaker

- **pgpool-II**



本日お話する内容

- ① Amazon Web Services (AWS) 上でpgpool-IIを使ってPostgreSQLのHA構成組みました
- ② 特に難しい話は出てこず、ベーシックな内容と導入事例のご紹介

1. HA構成とは

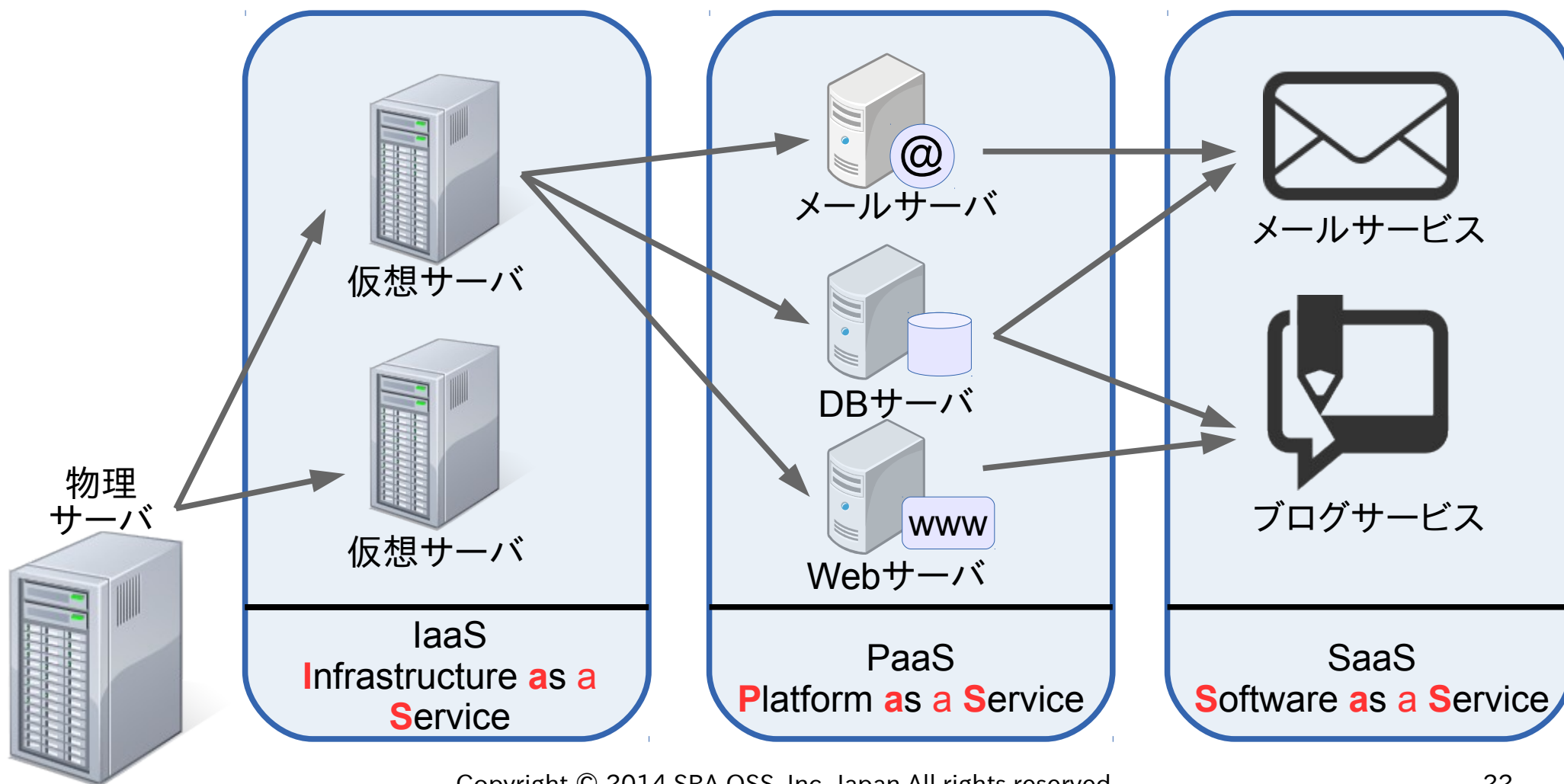
2. pgpool-IIとは

3. Amazon Web Servicesとは

4. AWS上でのpgpool-II+PostgreSQLの導入事例

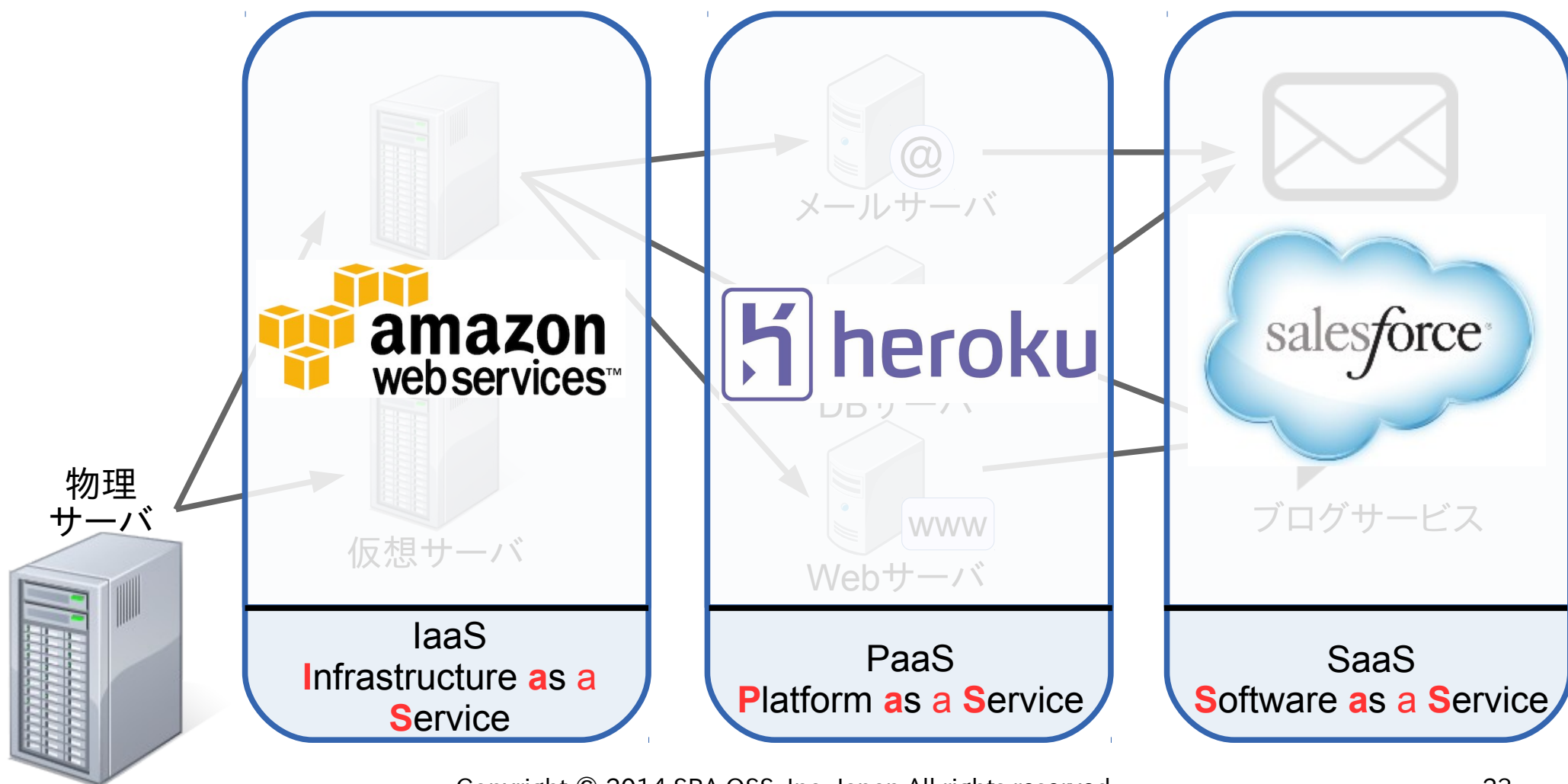
クラウドコンピューティングサービスの形態

- ① ネットワーク越しに仮想化されたリソースを提供
- ② 提供するレイヤ別に3つに分類できる



クラウドコンピューティングサービスの形態

- ✓ ネットワーク越しに仮想化されたリソースを提供
- ✓ 提供するレイヤ別に3つに分類できる



Amazon Web Services (AWS)

オンプレミスの仮想化と比較したAWSのメリット

	オンプレミス	AWS
初期コスト	必要	不要
利用コスト	余剰リソースにも発生	使った分だけ
インフラ調達	数週間～数カ月	数分
スケールアップ・ダウン	時間とコストがかかる	すぐに対応可能

[オンプレミスと AWS クラウドの違いとは? | アマゾン ウェブ サービス (AWS 日本語)]
<http://aws.amazon.com/jp/comparison-onpremise-aws/>

Amazon Web Services (AWS)

AWSサービス一覧



[AWS ベーシクトレーニング-トレーニング資料]

<http://www.slideshare.net/AmazonWebServicesJapan/aws-basic-20120629>

本日お話する内容

- ① Amazon Web Services (AWS) 上でpgpool-IIを使ってPostgreSQLのHA構成組みました
- ② 特に難しい話は出てこず、ベーシックな内容と導入事例のご紹介

1. HA構成とは

2. pgpool-IIとは

3. Amazon Web Servicesとは

4. AWS上でのpgpool-II+PostgreSQLの導入事例

株式会社Gengo様

gengo

翻訳サービスのクラウドソーシング企業

- 2009 年：東京でスタートアップ
- 2010 年：76 万ドルの資金調達に成功
- 2013 年：1200 万ドルの資金調達に成功

渋谷にオフィスに移転



サービス概要

- 翻訳して欲しい人と翻訳できる人(11000人以上が登録)をマッチング
- Web サイト上で or API を使って翻訳

[クオリティと生産性を両立するには | Gengo ゲンゴ]
<http://gengo.com/ja/how-it-works/>

翻訳可能な言語 (54 ペア)

From



英語



日本語



スペイン語 (ヨーロッパ)



中国語 (簡体)



ドイツ語



フランス語



ロシア語



イタリア語



ポルトガル語 (ブラジル)



スペイン語 (ラテンアメリカ)



フランス語 (カナダ)



韓国語



ポーランド語



ポルトガル語 (ヨーロッパ)



中国語 (繁体)

To



インドネシア語



韓国語



中国語 (簡体)



スペイン語 (ヨーロッパ)



中国語 (繁体)



タイ語

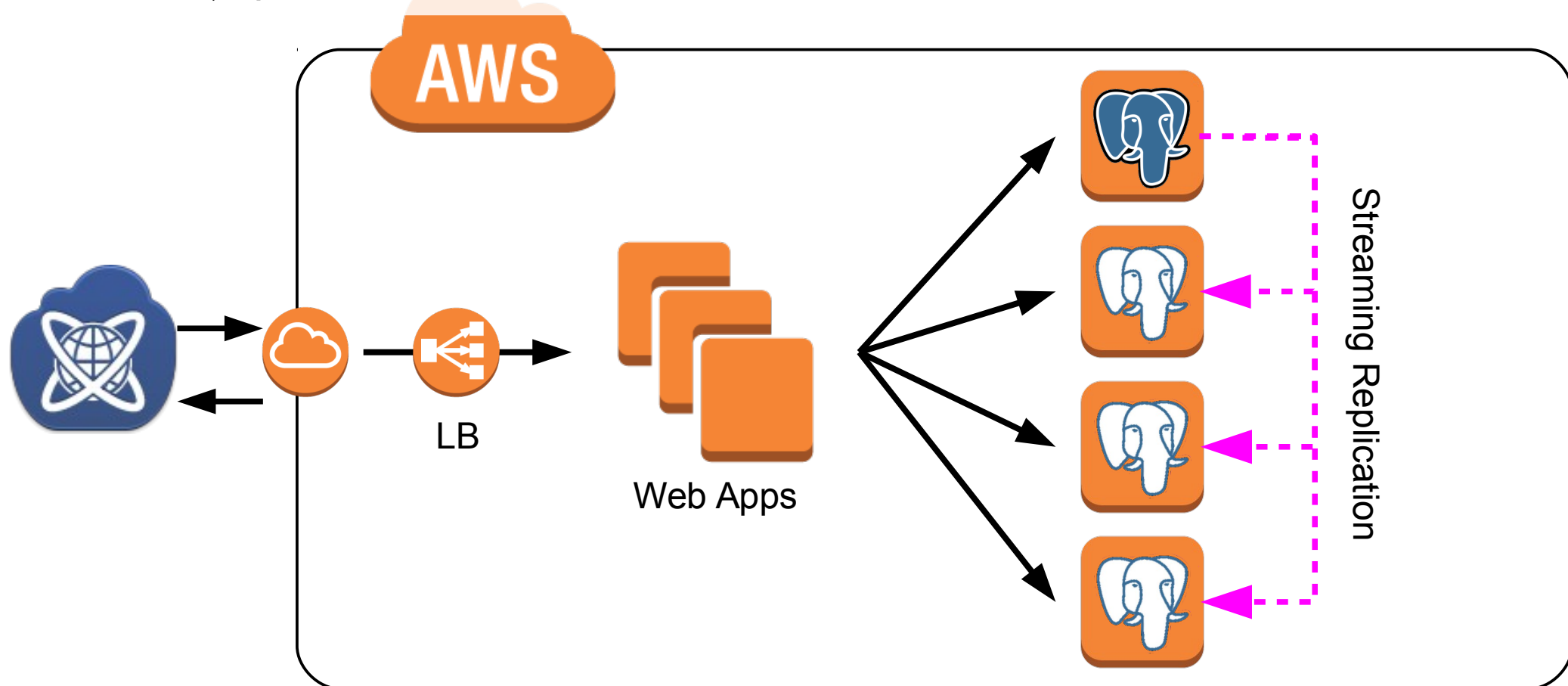


英語



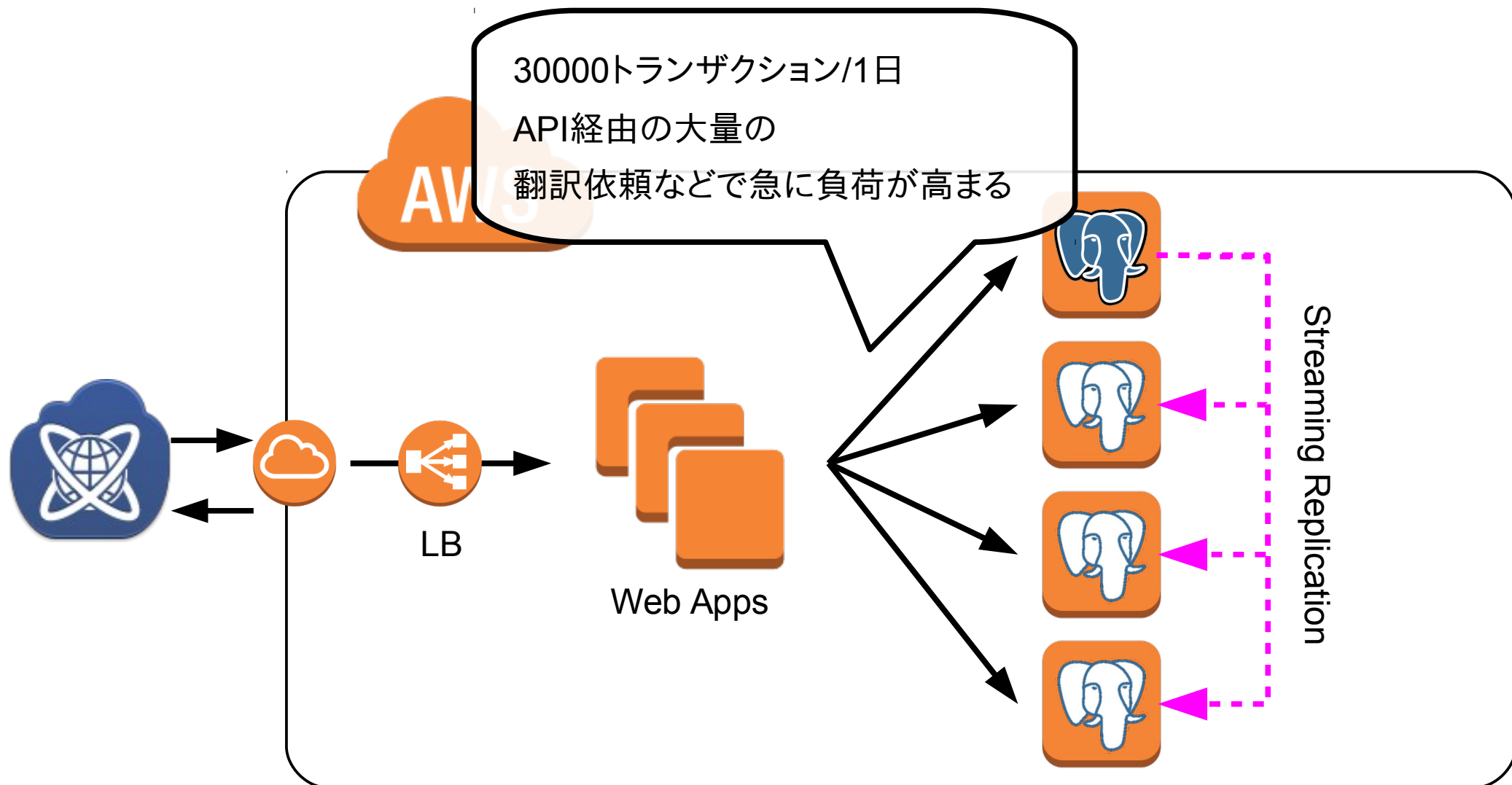
Gengo様の構成 (Before)

- PostgreSQL 9.1 on Ubuntu 11.10 - 14.10
- Primary * 1 + Standby * 8 の Streaming Replication 構成
- DBサイズは80GBほど



Gengo様の課題 1

- 「スパイク」なトランザクションで負荷が高まる



Many jobs 😊 ☹️

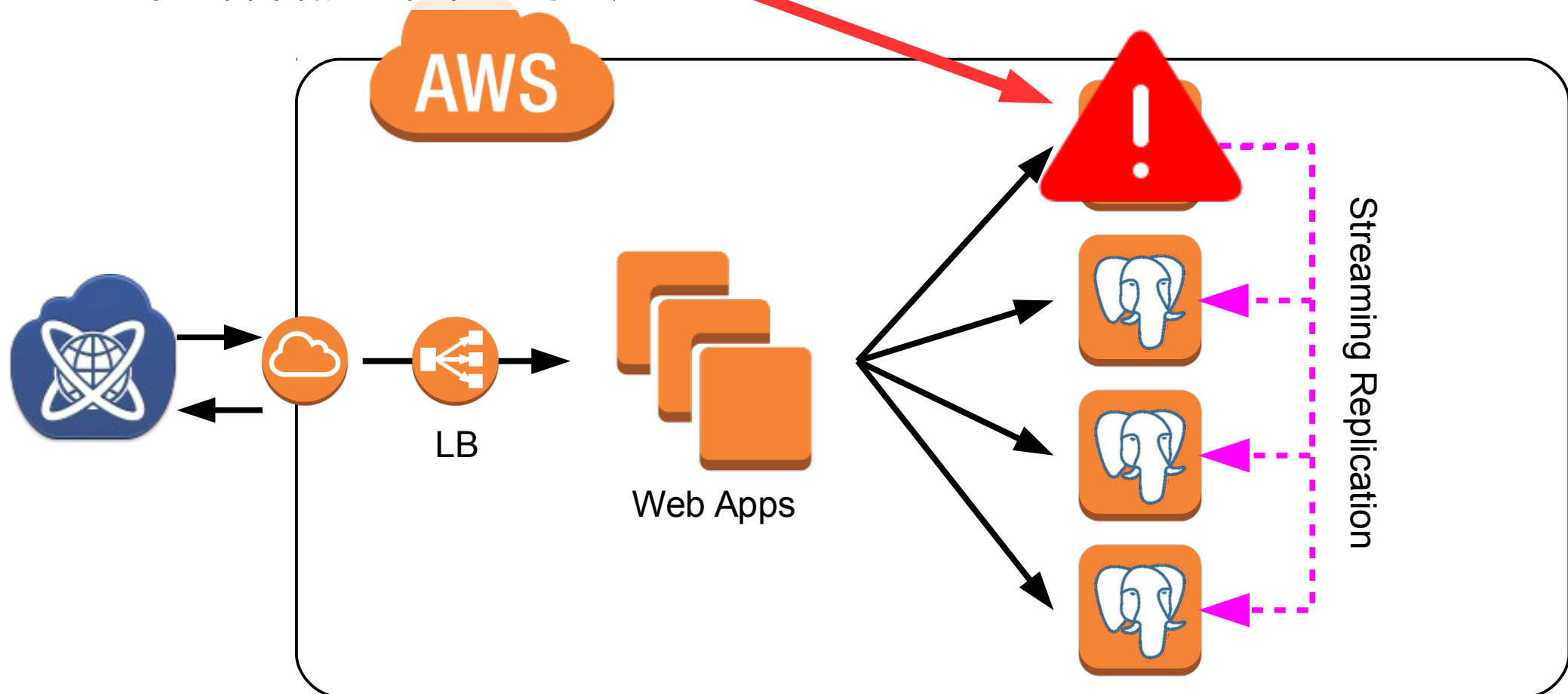


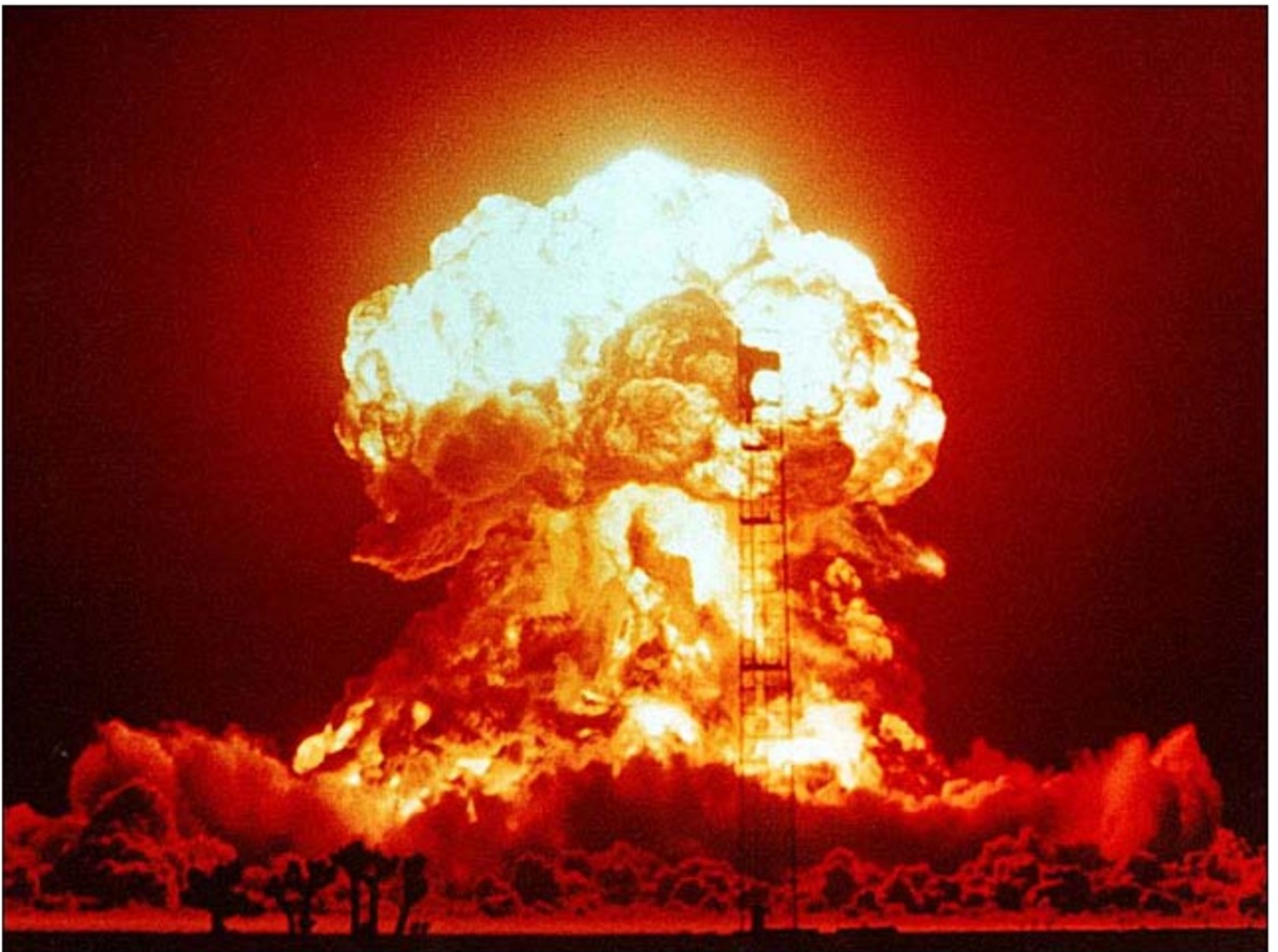
[2014年9月18日 第10回PostgreSQLエンタープライズ・コンソーシアムセミナー Gengo様スライドより]

<https://www.pgecons.org/2014/08/21/3157/>

Gengo様の課題 2

- ある日稼動系PostgreSQLサーバの**仮想イメージが壊れてしまう**（原因不明）
- 一時的にサービス停止に
- 単一障害点の存在が浮き彫りに





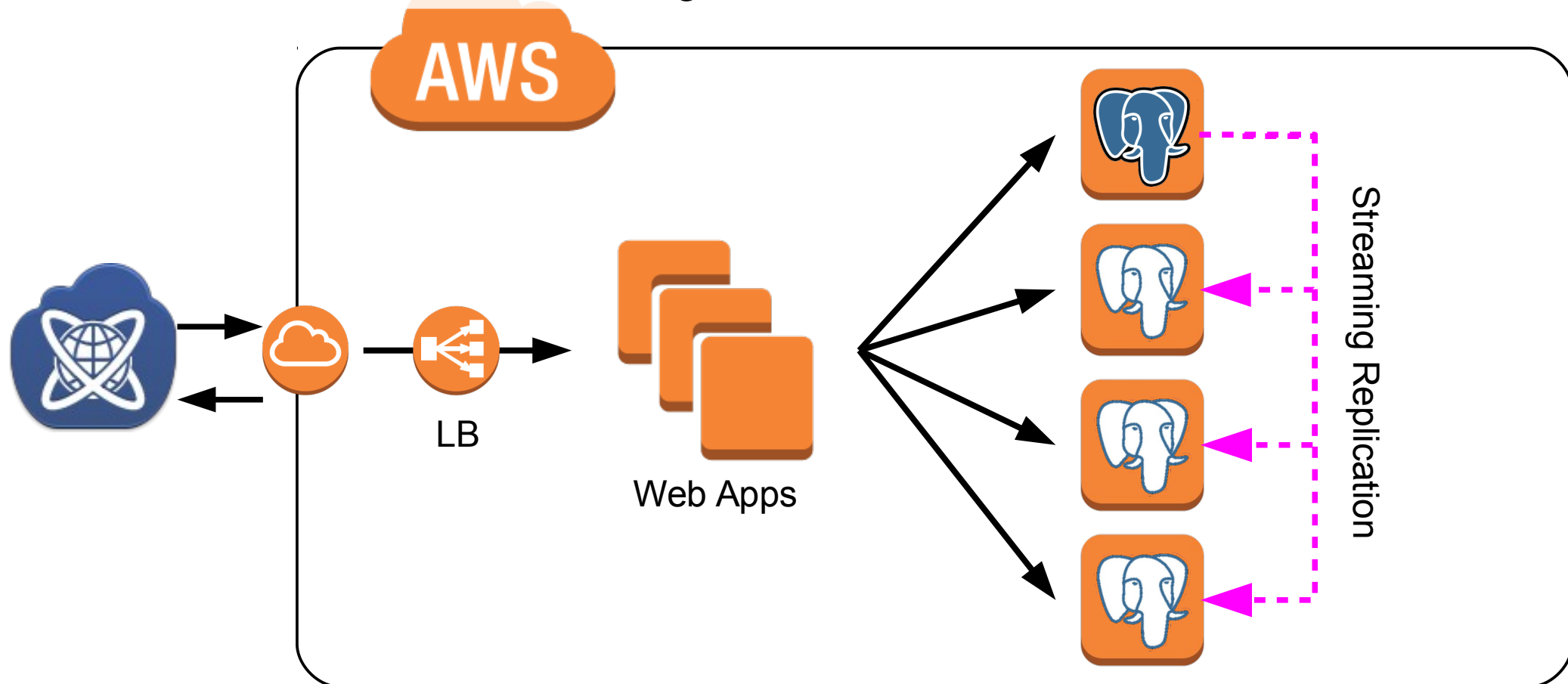
[2014年9月18日 第10回PostgreSQLエンタープライズ・コンソーシアムセミナー Gengo様スライドより]

<https://www.pgecons.org/2014/08/21/3157/>

Copyright © 2014 SRA OSS, Inc. Japan All rights reserved.

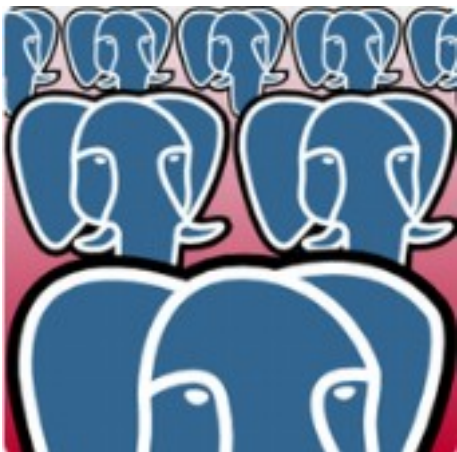
Gengo様の課題 3

- 稼働系サーバがダウンしたら自動フェイルオーバーしてほしい
- Web Apps 側の負荷分散以外に何かよい負荷分散方法はないか
- JSON型を使いたいのでPostgreSQLを9.3にアップグレードしたい



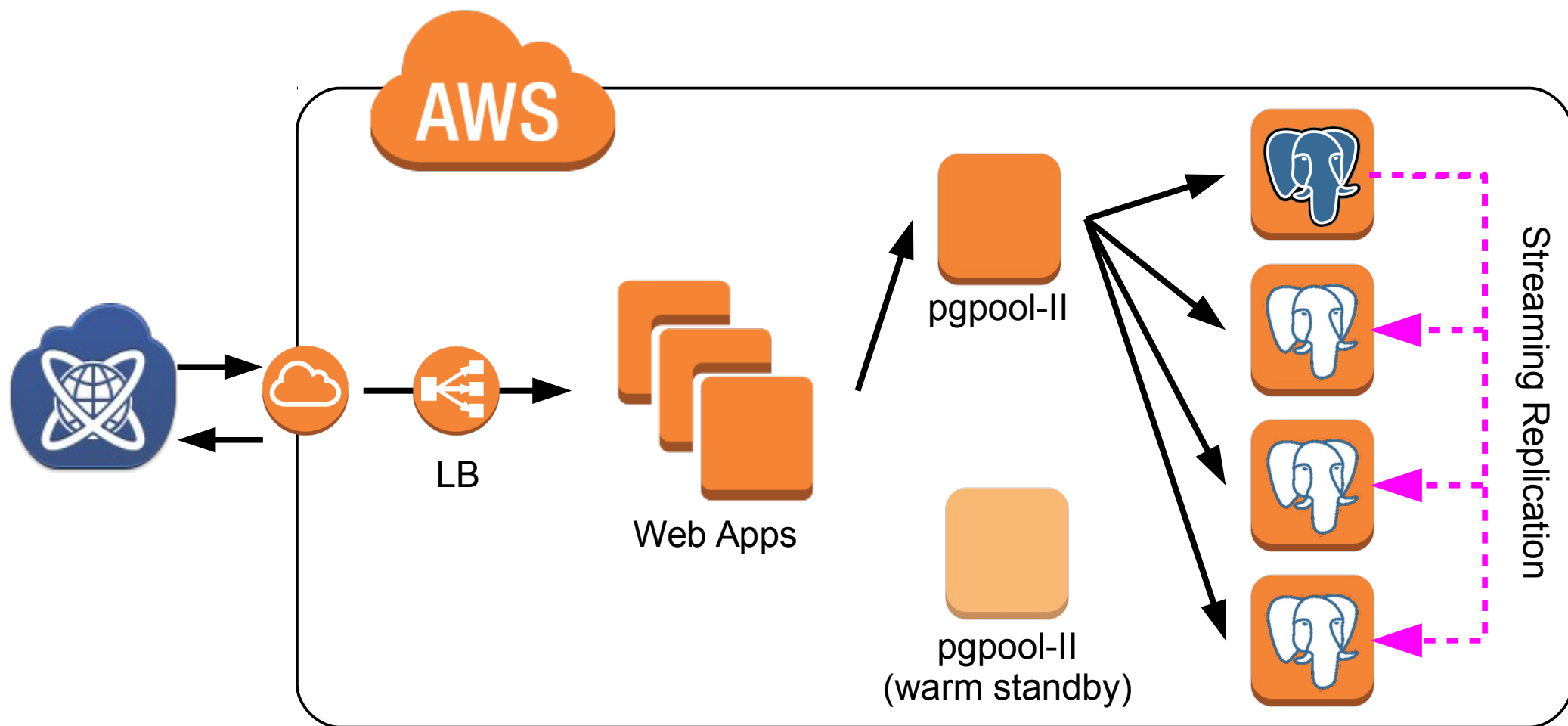
Gengo様の課題

- ✓ 多量のトランザクションを負荷分散
- ✓ 単一障害点を取り除く
- ✓ 自動フェイルオーバ
- ✓ PostgreSQLを9.1 -> 9.3にアップグレード
 - ダウンタイムは1時間でおねがいします

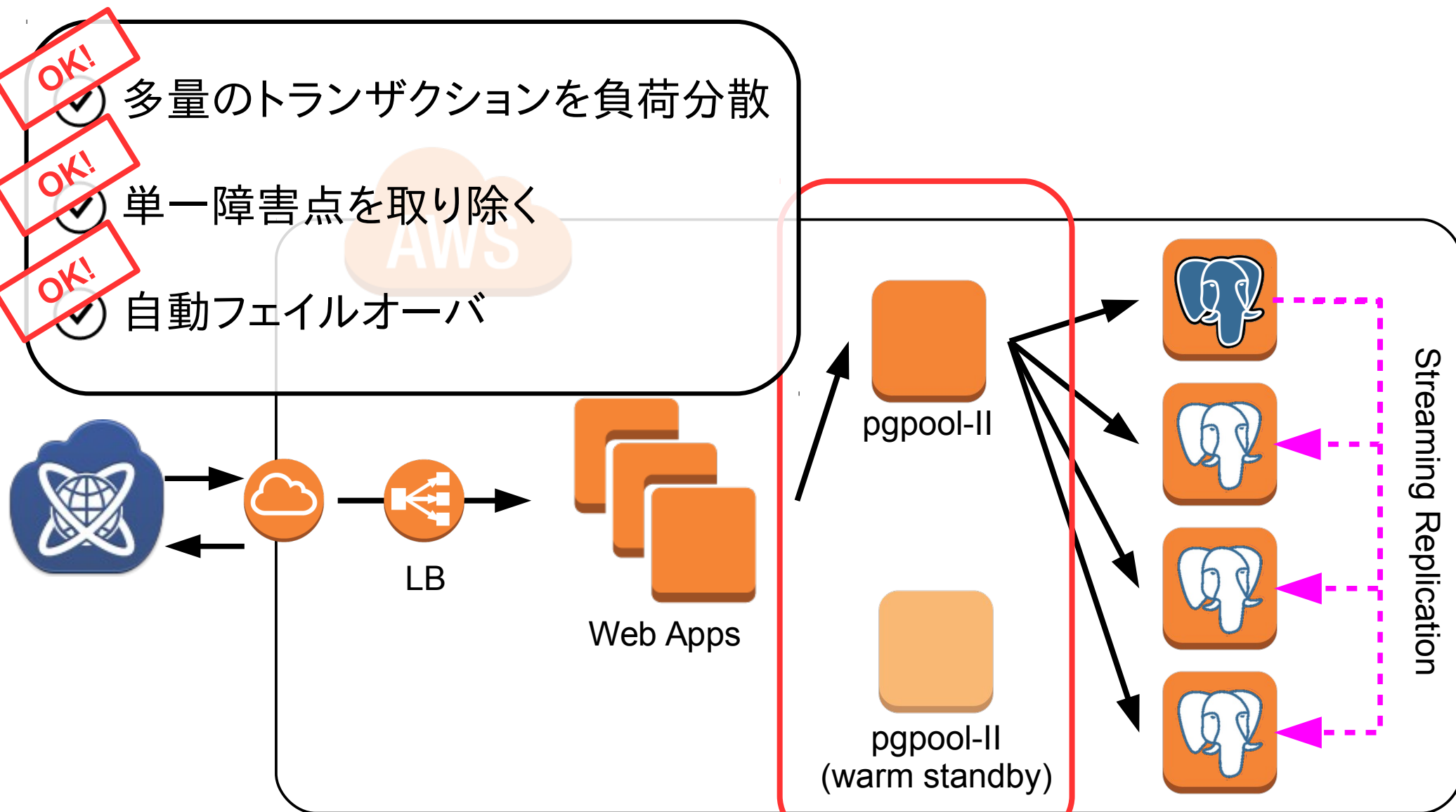


pgpool-IIの出番

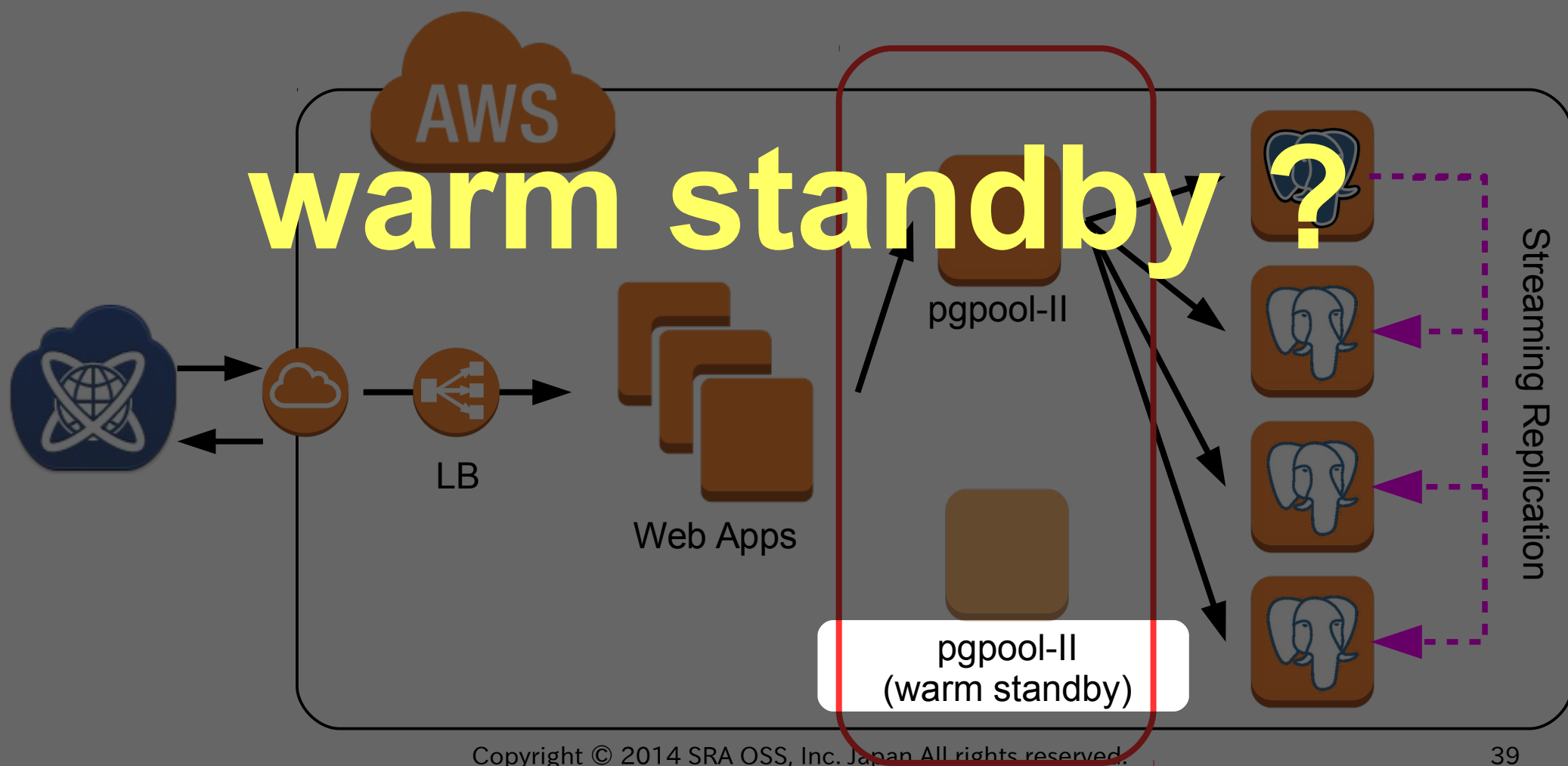
Gengo様の構成 (After)



Gengo様の構成 (After)



Gengo様の構成 (After)



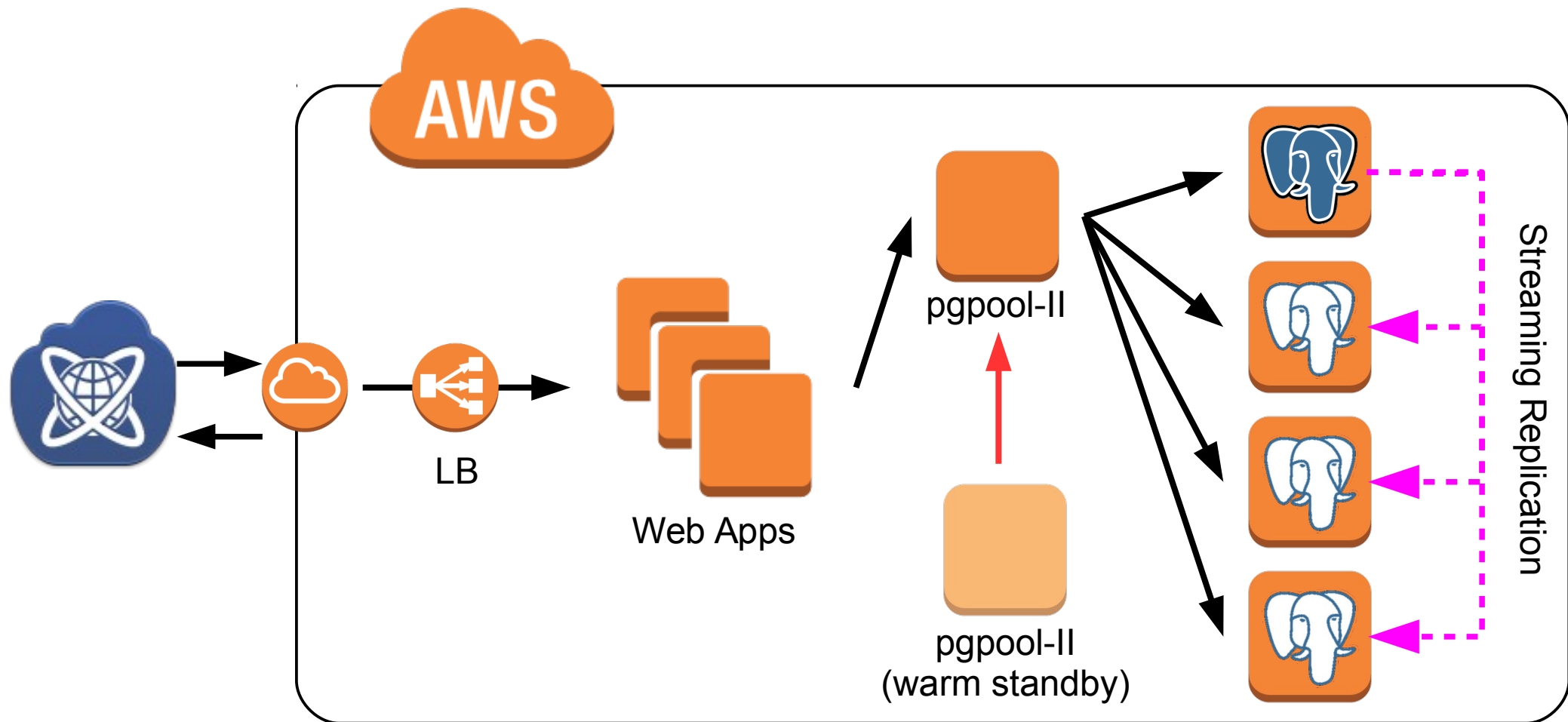
| AWS のハマりどころ



AWSでは仮想IPが使えません

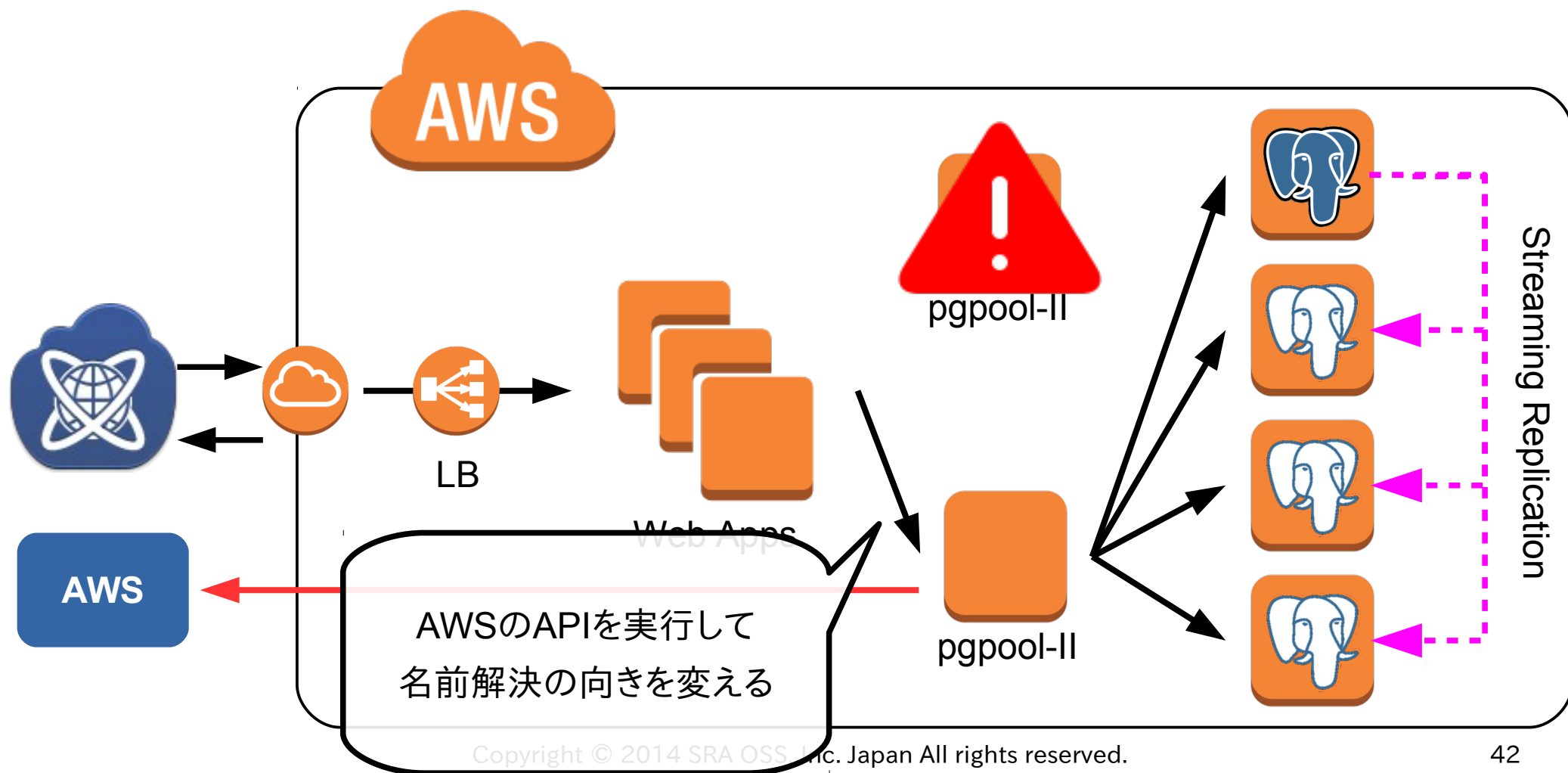
AWS的対策

☑ “SELECT 1”でモニタリング



AWS的対策

- ✓ “SELECT 1”でモニタリング
- ✓ AWSのRoute53(DNSサーバ)を使ってアドレス変更



AWS的対策

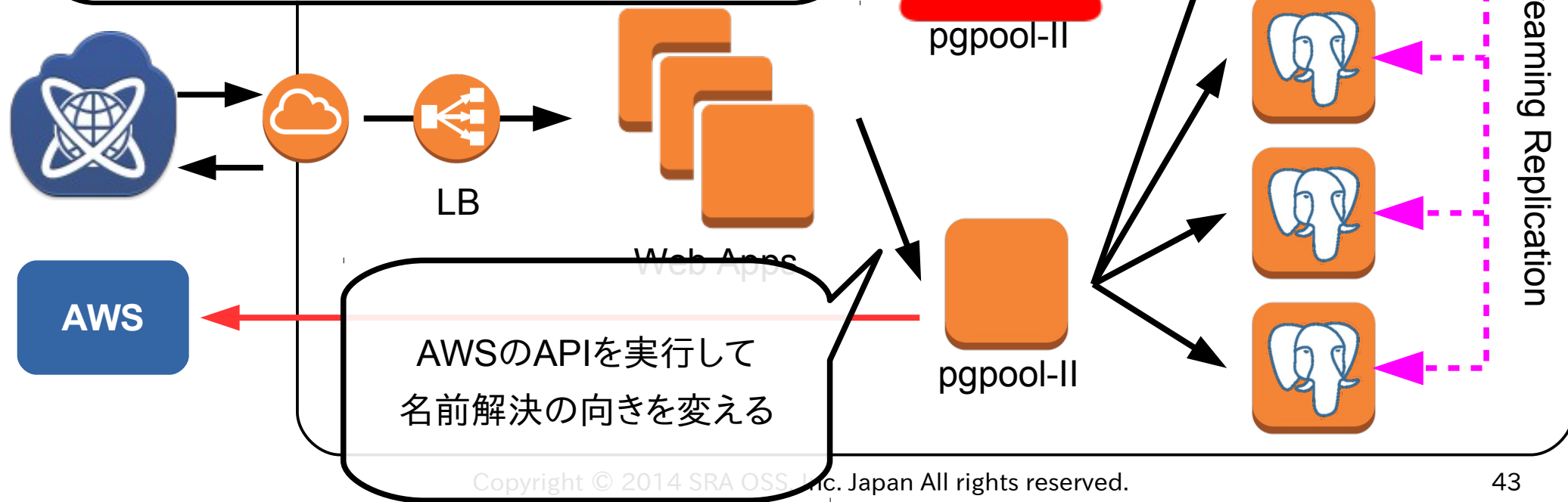
“SELECT 1”でモニタリング

OK! 多量のトランザクションを負荷分散を使ってアドレス変更

OK! 単一障害点を取り除く

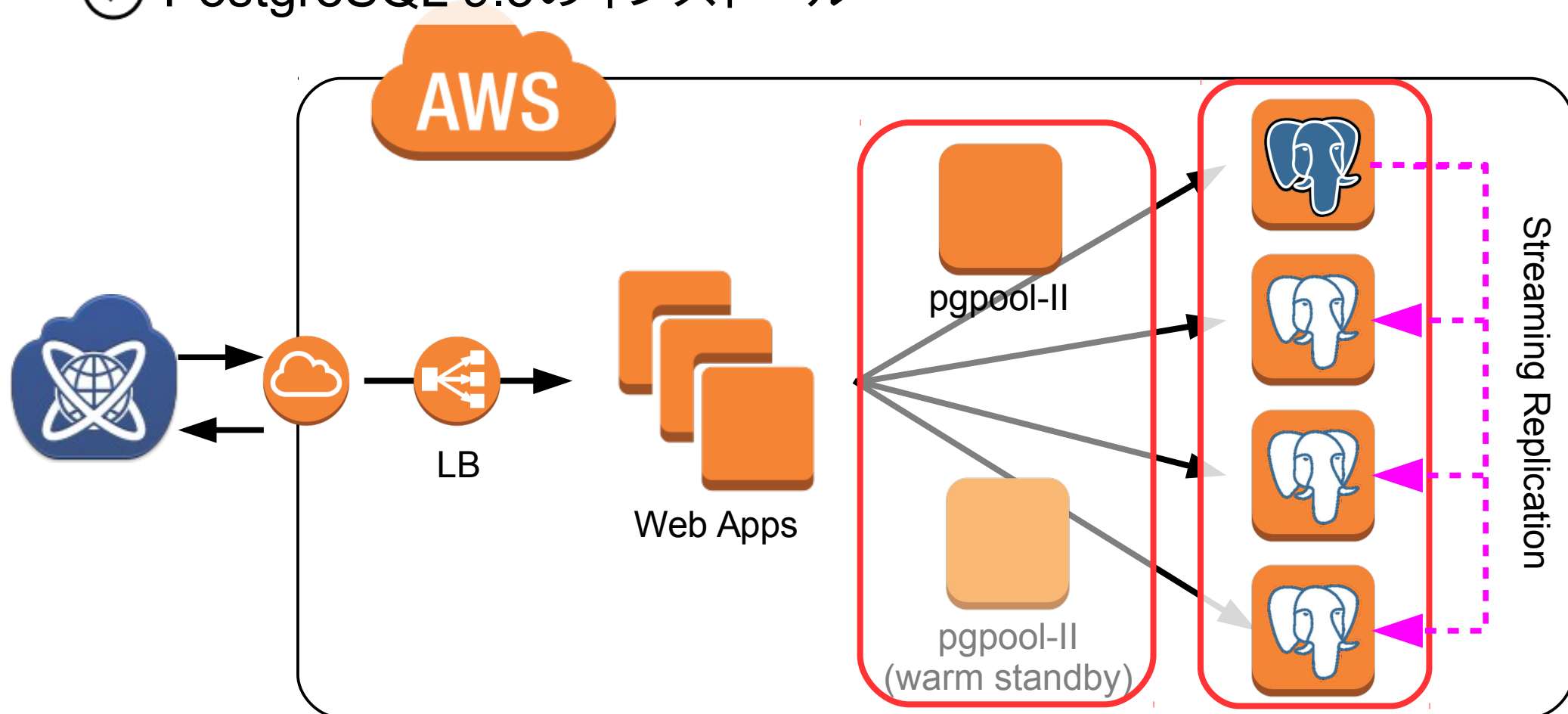
OK! 自動フェイルオーバー

pgpool-IIも



導入作業について ～ダウンタイムは1時間でおねがいします～
サービス稼働中に影響のない作業は済ませておく

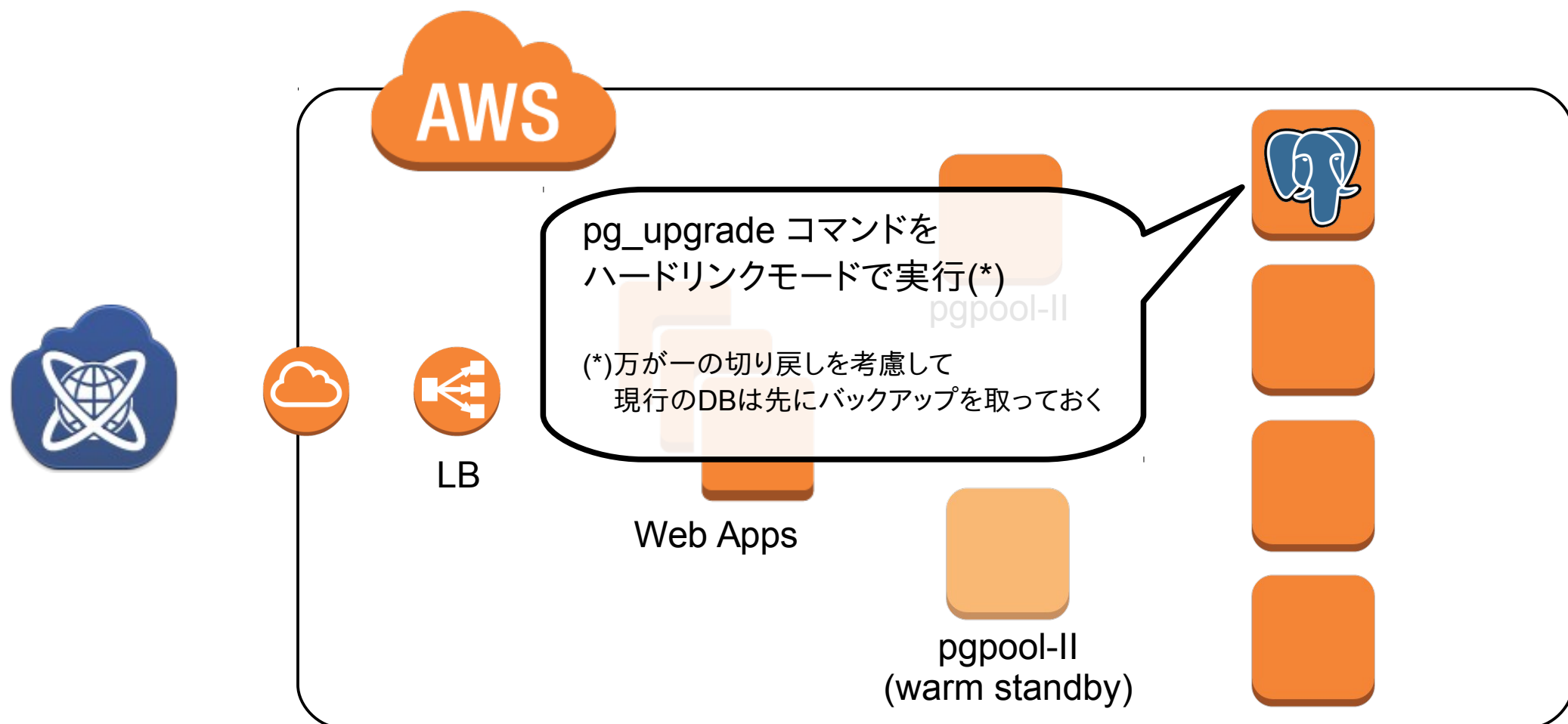
- ✓ pgpool-IIのセットアップ
- ✓ PostgreSQL 9.3のインストール



導入作業について ～ダウンタイムは1時間でおねがいします～

サービス停止

☑ データベースクラスタのアップデート

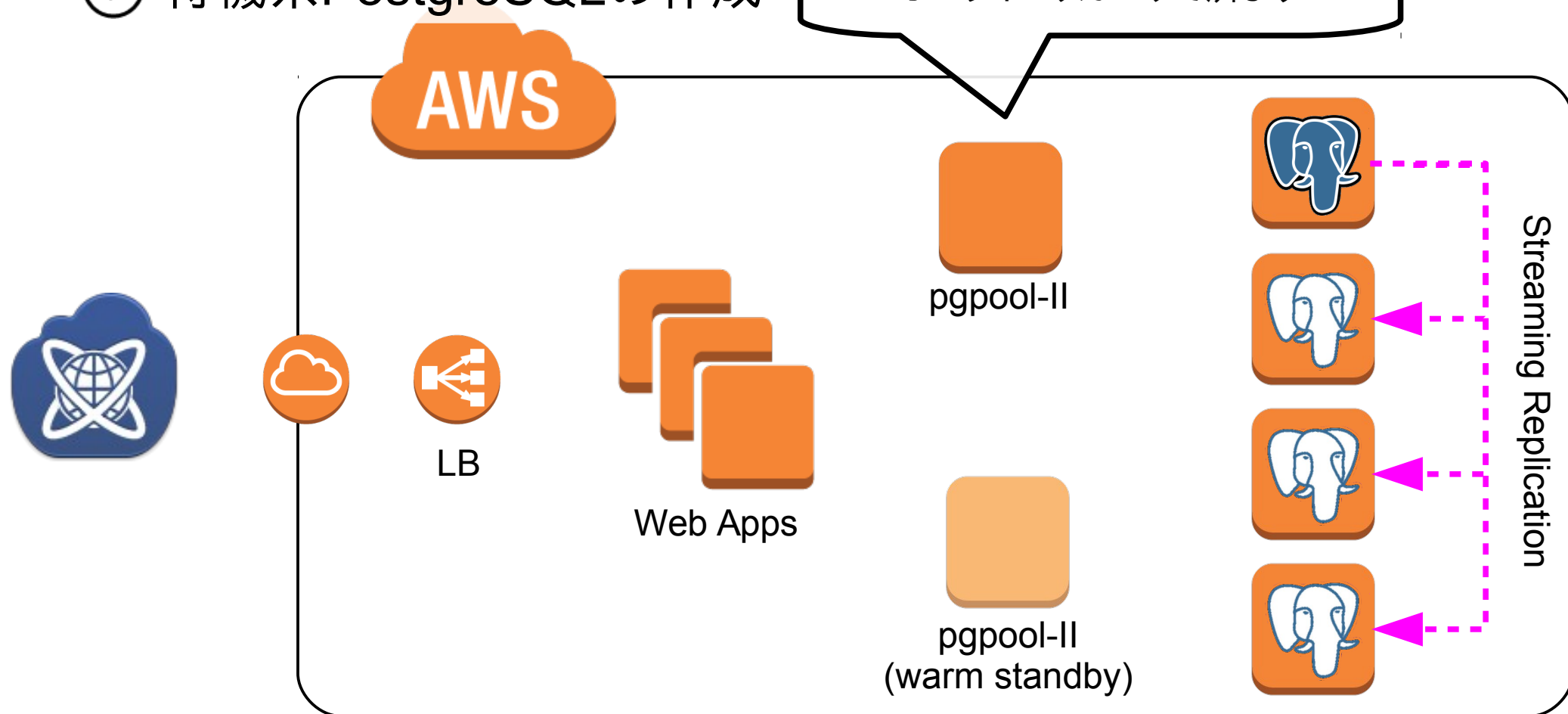


導入作業について ～ダウンタイムは1時間でおねがいします～

サービス停止

- ① データベースクラスタのアップデート
- ② 待機系PostgreSQLの作成

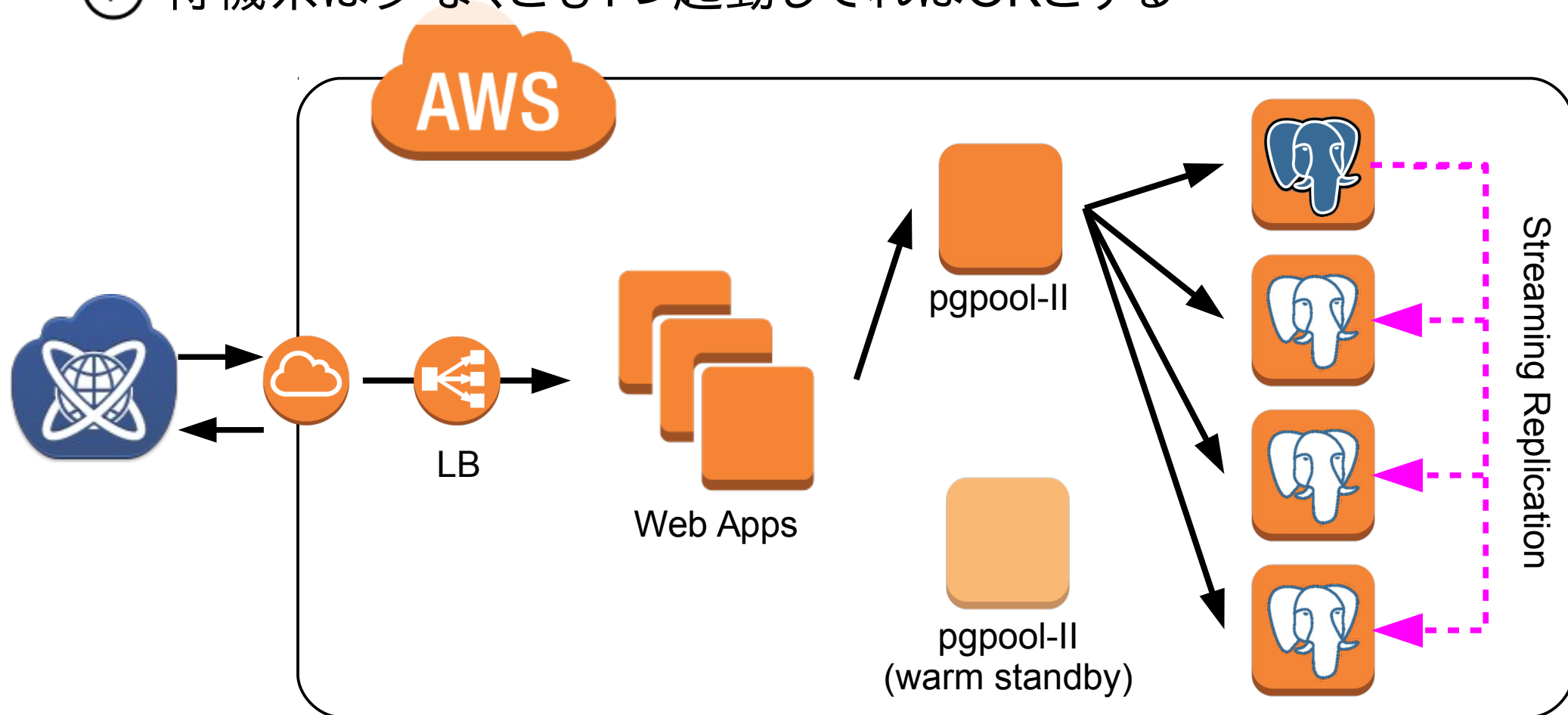
オンラインリカバリで済ます



導入作業について ～ダウンタイムは1時間でおねがいします～

1時間以内で作業フィニッシュ!

- ✓ ただし作業は全てスクリプトで自動化
- ✓ 待機系は少なくとも1つ起動してればOKとする





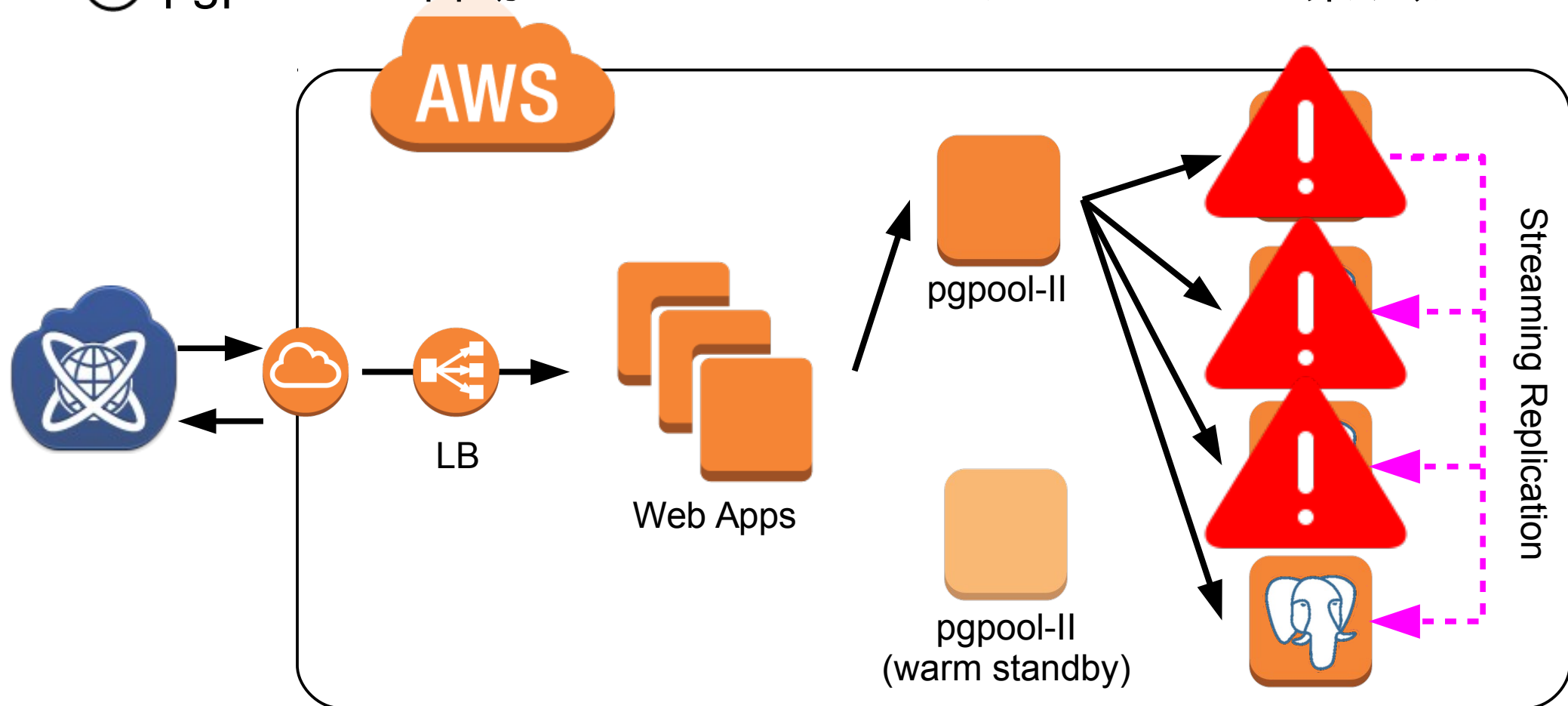
Wow
Such thanks.

[2014年9月18日 第10回PostgreSQLエンタープライズ・コンソーシアムセミナー Gengo様スライドより]
<https://www.pgecons.org/2014/08/21/3157/>

Gengo様その後

Xenの脆弱性対策に伴う強制インスタンス再起動メンテナンスに巻き込まれる

- ✓ 待機系1台を除きPostgreSQLサーバが再起動
- ✓ pgpool-IIの自動フェイルオーバーでメンテナンスなしで乗り切った



その他のTIPS

pgpool-IIのヘルスチェックがタイムアウトしてしまう

- ① 仮想環境特有のネットワークがプツプツ切れる問題
- ① pgpool-IIはヘルスチェック時にselect()コールを投げている
ソースコードにパッチを当ててタイムアウト値を延長した
(将来的にはユーザが指定できるようになる予定)

商用クラスタリングソフトウェアはAWSの仮想IP問題にどう対応している？

- ① NATインスタンスと仮想ネットワークインタフェースとルーティングルールをAWS APIで操作してなんとか仮想IP的な方法を実現
- ① 試みにpgpool-IIでも同様の手法を実験してみたら
一応動作した
が、あまり現実的ではない

まとめ

- ① PostgreSQLの高可用性+負荷分散 on AWS
- ① pgpool-IIで構築できました

ただし

- ① AWSでは仮想IPが使えないので代替手段が必要
- ① 死活監視のタイムアウト値は長めに
- ① AWSのメンテナンス再起動は不可避なので対策は考えておく

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