PostgreSQL 9.1 でつくる 高可用システムにまつわるエトセトラ

2012年11月30日

株式会社NTTデータ 基盤システム事業本部
海浦 隆一

NTT DATA

本日お話しすること

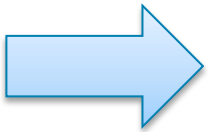
- NTTデータでのOSSの取り組み
- PostgreSQL 9．1 の商用システム適用
- 商用システム適用時の勘所
- 高可用性を実現したDBアプライアンス

NTTデータでのOSSの取り組み

NTTデータとOSS

NTT DATA

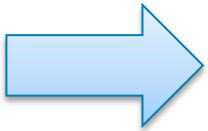
OSSは既に 当たり前に使われる ステージへ！



フルOSSモデルのスタック
－ PRORIZE フルOSSモデル －

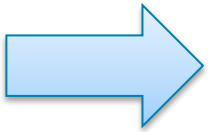
OSSの統合管理ツール
－ Hinemos －

OSSで新たな ビッグデータ の領域へ！



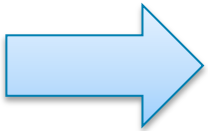
分散処理フレームワーク
－ Hadoop －

OSSを組み合わせて 仮想化基盤 を構築する！



フルオープン仮想化基盤構築ソリューション
－ OpenStack ＋ OpenFlow －

OSSで MC領域 をカバーする！



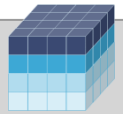
OSS-DBMS
－ PostgreSQL －

基盤構築ソリューション PRORIZE（プロライズ）

IAも。UNIXも。OSSも。
SOAも。仮想化も。

NTT DATA

検証済システム基盤構成 ＋ 設計済開発テンプレート



PRORIZE 検証済システム基盤構成

Web3層モデル	フルOSS Webモデル
	Webモデル（IAサーバ）
	Webモデル（UNIXサーバ）
SOAモデル	SOA基盤モデル（BPM）
	SOA基盤モデル（ESB）
	SOA基盤モデル（ハードウェアESB）

スタックモデルの中でも、OSSモデルは最も採用数が多い。
様々なサービス基盤でOSSモデルが稼動中！

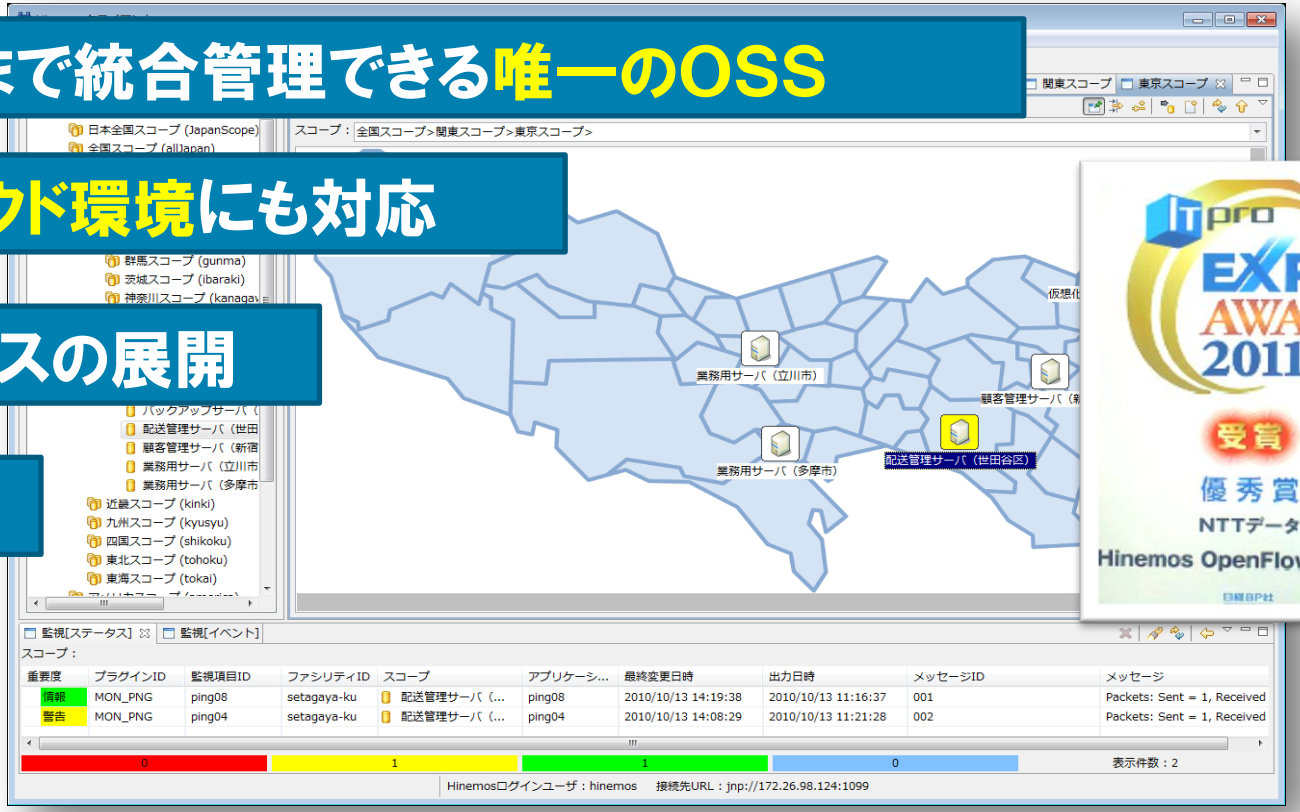
オープンソースの統合管理システム

監視からジョブまで統合管理できる唯一のOSS

仮想環境/クラウド環境にも対応

パートナービジネスの展開

先進的な機能



商用製品の代替ではない、Hinemos/OSSらしさで多くのパートナーと付加価値を伴いシステムの重要部を担う！

NTT DATA

加速するHadoop

急激に増加する様々な分野からのニーズ

cloudera 社と戦略的な協業

→ Hadoopディストリビューション「CDH」のサブスクリプション販売
より高度なサービスを提供できる体制を実現

お客様	規模など
経済産業省 様	100台
リクルート様	120台規模
国立図書館 様	40台規模
国立研究機関 様	70台規模

＋ 1000台超の規模

日本最大級のHadoopクラスタ（※当社調べ）の構築・運用
従来では予想もつかなかった領域のサービスをOSSで実現！

フルオープン仮想化基盤構築ソリューション

NTT DATA

課題	ソリューション
急速なビジネス変化 無駄のない投資	リーンスタートアップ
環境の配慮 電気代の高騰	アイドリングストップ
予期せぬ災害 大規模障害	ディザスタリカバリ
運用管理の問題	統合マネージメント

ITpro EXPO2012 AWARD
優秀賞受賞！！

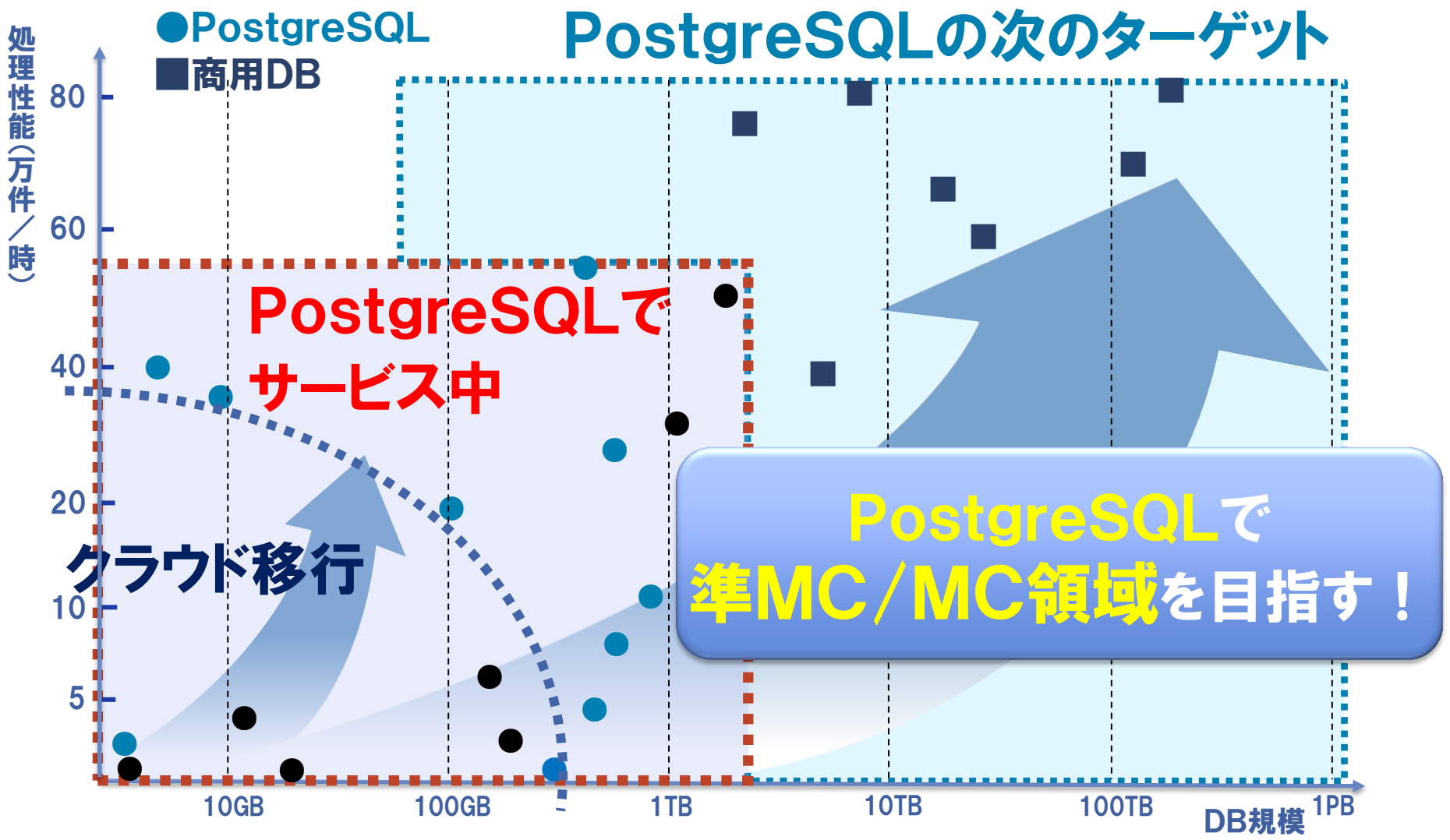
openstack™

＋

OpenFlow

OpenStackやOpenFlowなどのオープンな技術を活用し、
災害に強く、省電力で運用できる仮想化基盤を構築！

MC領域をカバーする PostgreSQL



出典:当社社内システム及び当社にてご提供中の商用システムをDB規模と処理性能で整理

NTT DATA

課題となる3つの柱

サポート体制

- 長期にわたるPostgreSQL等のOSSサポート
- さまざまな場面（検証・設計から運用トラブルまで）の手厚いサポート

技術開発／検証

- MC要件を満たす機能、ノウハウ蓄積
- 信頼性（ノンストップ・データロスなし）・運用性（導入・運用・解析の容易さ）

技術者育成

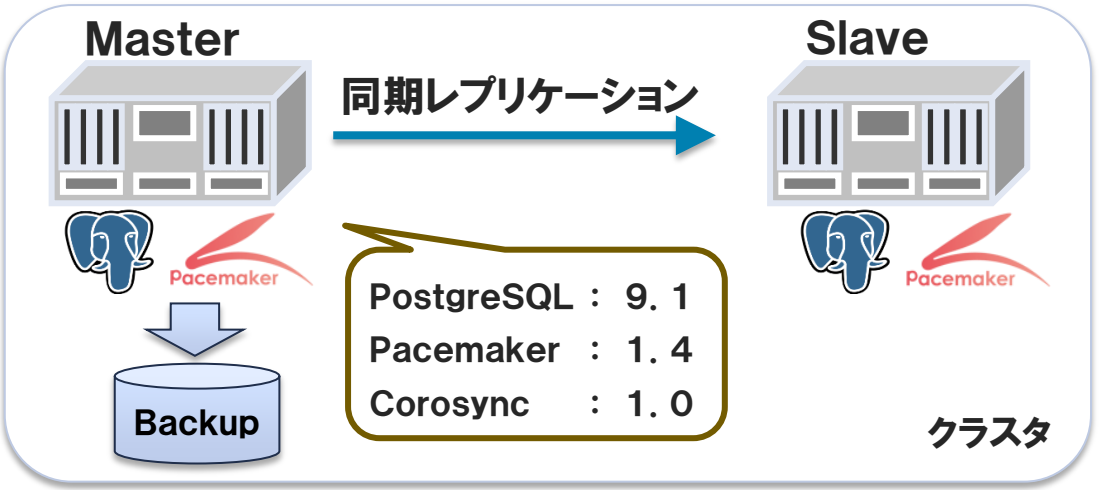
- OSSを扱えるだけでなく、中身を知ってサポート／コア開発ができる
- 緻密なサポートと自社ニーズを満たす機能の実現

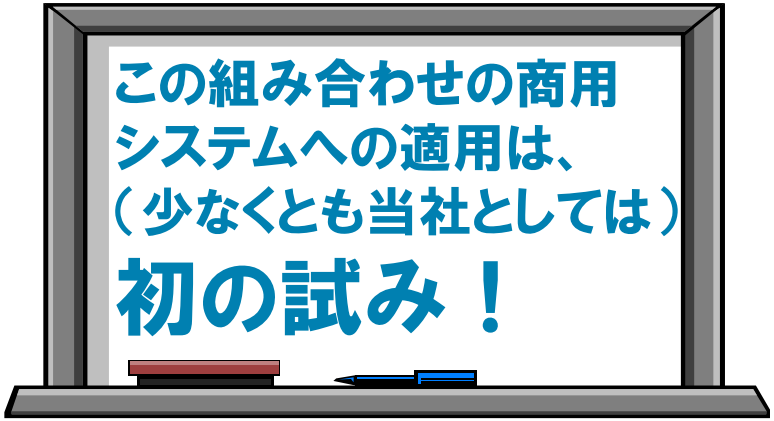
PostgreSQL 9．1 の商用システム適用

高可用DBサーバの商用システム適用

NTT DATA

□ NTTデータでは、**同期レプリケーション**を利用した高可用DBサーバを提供し、**商用システムに適用**しています。





適用までの取り組み

商用システム概要

各種コンテンツ・アプリケーション提供 SaaS 型プラットフォーム
BizXaaS® MaP

NTT DATA

システム概要と目的

- **地図上で物件管理情報の表示・編集**を行う
- 効率的なマーケティングを支援

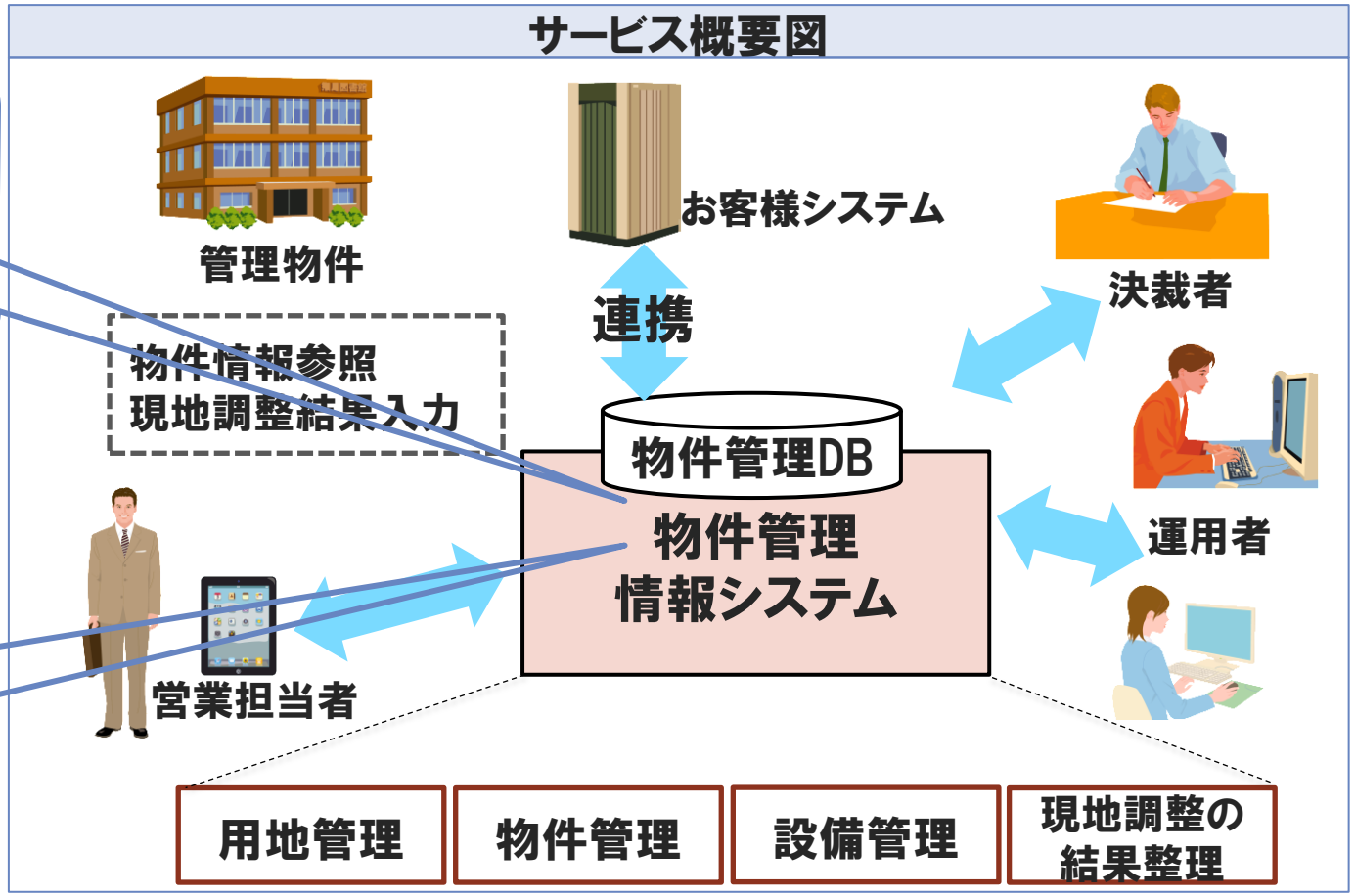
利用状況

- ユーザ数：**数万人**
- データ容量：**5～10TB**
- 運転時間帯：5時～24時

地図上に
物件管理情報を
プロットできる

訪問履歴や競合情報が
地図情報と併せて
管理できる

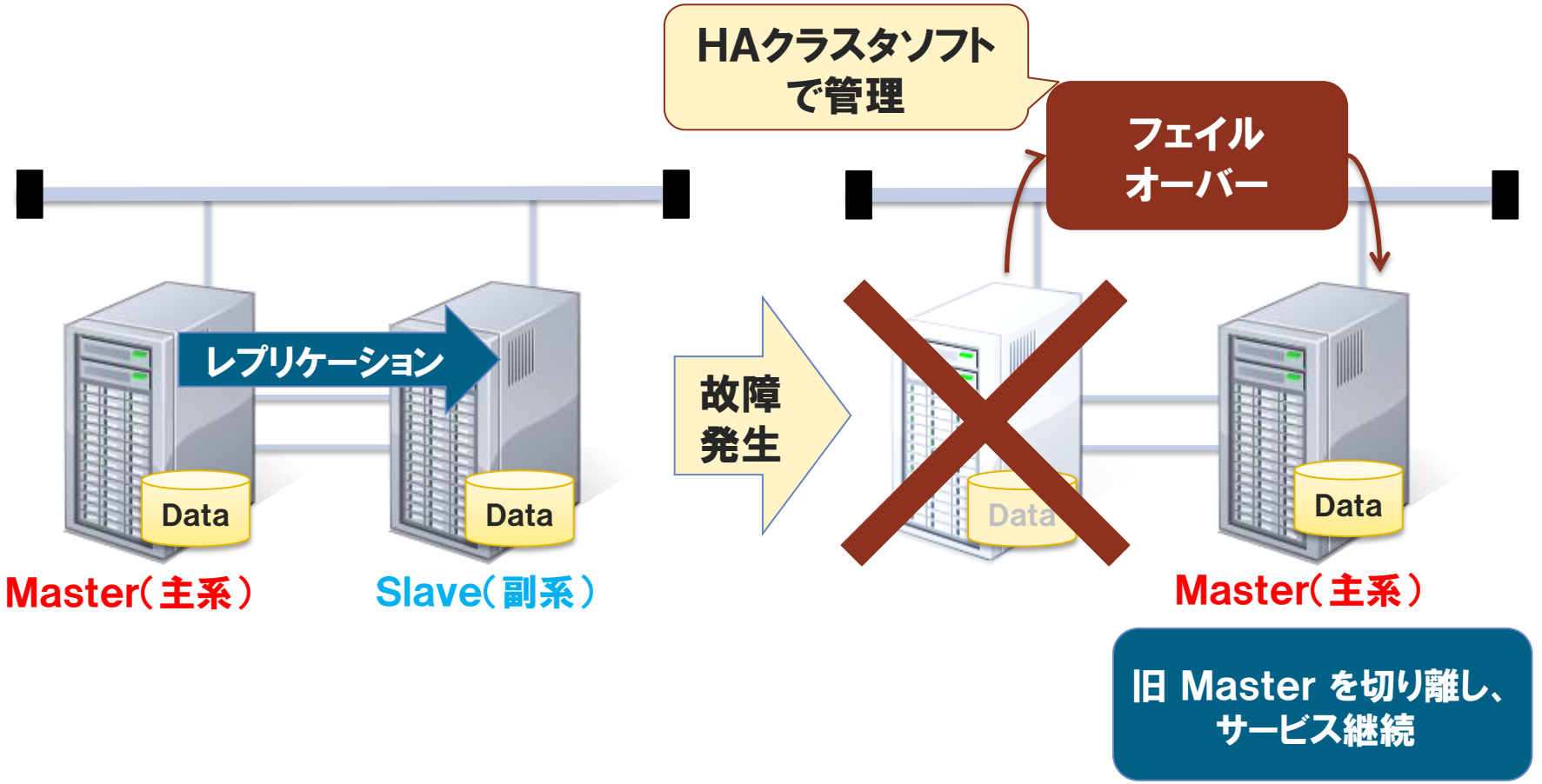
サービス概要図



DBサーバの構成

NTT DATA

- DBサーバは、物理マシン2台で共有ディスクレスの構成です。同期レプリケーションによりそれぞれのマシンのローカルディスクにデータを保持します。



HA クラスタとは？

NTT DATA

ロ HA(High Availability)クラスタとは、**コンピュータシステムの可用性(Availability)を高めることを目的とした、複数台のコンピュータによる連携構成(クラスタ)**のことを指します。

- HA クラスタの管理を行うソフトウェアを、HA クラスタソフトとします。
  HA クラスタソフトの代表例をいくつか列挙します。

HAクラスタソフト	分類
Cluster Pro	商用
Life Keeper	商用
PRIMECLUSTER	商用
HeartBeat	OSS
Pacemaker	OSS

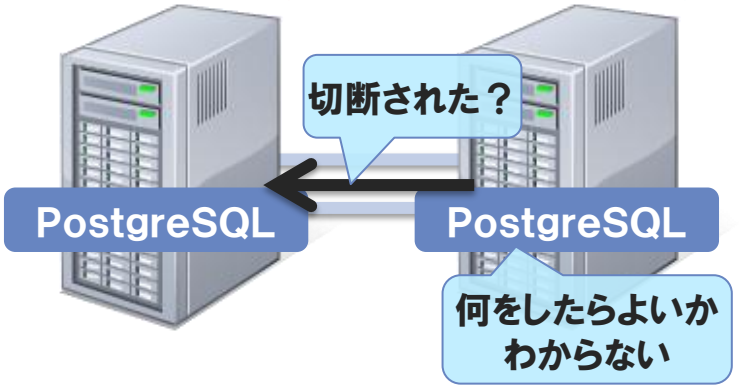
採用

なぜ HA クラスタソフトが必要？

NTT DATA

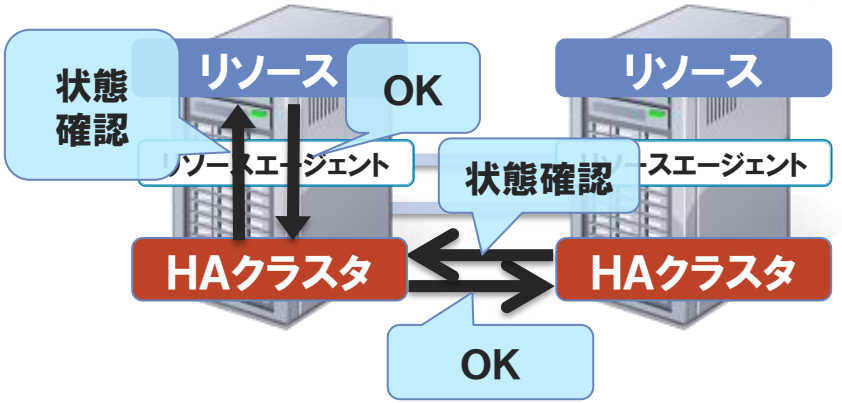
□ PostgreSQL には、フェイルオーバーを**自動的に行う機能がありません**。故障発生を検知して、各種処理を実施する HA クラスタソフトと組み合わせる必要があります。

PostgreSQL 単体だと・・・



Master から切断されたことは検知するが、そのまま待機する。

PostgreSQL と Pacemaker を組み合わせると・・・



互いに状態を確認し、問題が発生したら定義された**適切な処理**を行う。

商用システム適用時の勘所

商用システム適用にあたっての 5箇条

NTT DATA

□ PostgreSQL 9．1 の高可用商用システム適用にあたり、特に意識しておくとよい点は、下記の5つです。

検証は大事

サイジングは柔軟に
　－構成の決定は重要。柔軟な対応も必要－

商用システムは運用が要
　－切り替えは100％成功かつ即座の切り替え－

備えあれば憂いなし

監視で安心

5箇条 その1：検証は大事

NTT DATA

□ 検証および試験時に、全てのエラー項目、運用ミスなどはだしきり、運用を整理しましょう。

たとえば、今回実施した HAクラスタ（信頼性）の検証項目抽出の条件・・・

1	正常系試験	想定している運用操作を抽出し、**各操作が正常に実行できる**ことを確認する。
2	異常系試験	ハードウェア構成とリソース構成を考慮し、発生しうる故障種別を網羅した**単一故障試験**項目と、その組み合わせによる**多重故障試験**項目を抽出する。 例：現用系サービスLAN故障 × レプリケーションLAN現用・予備系が同時故障
3	長期安定試験	Pacemaker サービス起動後、DBに対し背景負荷をかけた状態で、一定時間連続走行させる。**走行中に各サービスが停止しない**こと、**メモリリークが発生していない**ことを確認する。
4	切替性能試験	ピーク負荷時だけでなく、DB（PostgreSQL）で実施されるVACUUMをはじめとする**メンテナンス**やバックアップ中に切り替えが発生した場合の時間も調査する。 また想定故障箇所については、その仕組み上、**最も切替が遅くなるもの**を重点的に調査し、最悪のシナリオでどれくらいの切替時間となるかを見定める。

検証中に発生した問題

NTT DATA

□ 案の定たくさんの課題・問題に直面しました。

- ９．１未対応のモジュール利用時に、コンパイルが通らない？
- RA どうしましょう？
- 切り戻し手順がよくわからない？
- クラスタ組んだけど切り替わらない？
- スプリットブレイン発生？
- . . . etc

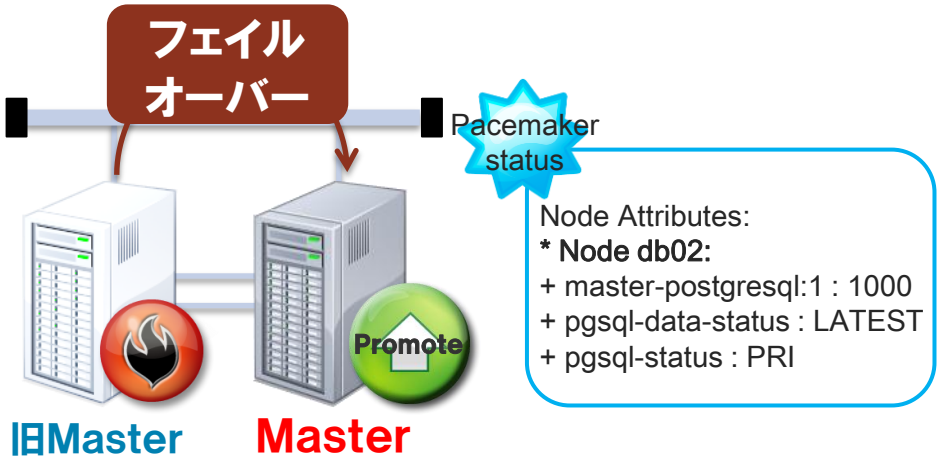
検証中に発生した問題：切り戻し失敗

NTT DATA

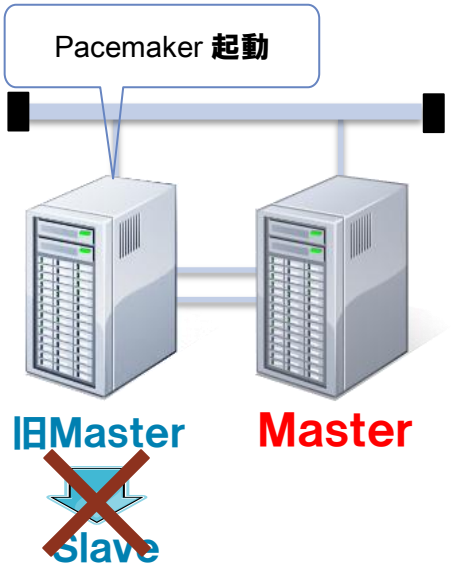
ロ フェイルオーバーが起こったあとクラスタの再構成（切り戻し）をしようと思ったけど、うまくできない。

発生事象

レプリケーション構成となっている状態で、
フェイルオーバーが発生。
正常に切り替わり、Master として起動



旧Master をクラスタに再度組み込み、レプリケーション構成を構築しようとしたが、Slaveとして動作しない！



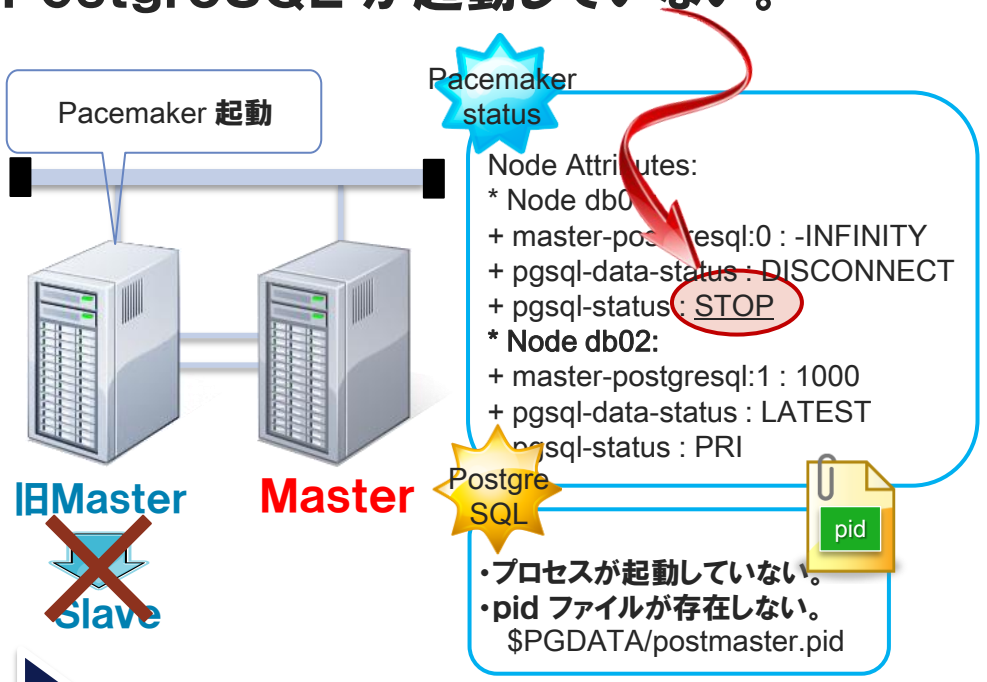
検証中に発生した問題：切り戻し失敗

NTT DATA

ロ フェイルオーバーが起こったあとクラスタの再構成（切り戻し）をしようと思ったけど、うまくできない。

原因①

PostgreSQL が起動していない。



PGSQL. lock ファイルが存在しているため、PostgreSQL は起動に失敗。

対処①

PGSQL. lock ファイルを削除する。

PGSQL.lock ファイル

・・・Master、Slaveのサーバ間で、データ不整合を避けるために作成。本ファイルが存在すると、PostgreSQL は起動できない。

作成者：Pacemaker の PostgreSQL RA
作成　：Master が起動したとき
削除　：正しい手順で Master が停止されたとき
※ Slave のマシンが存在しない状態で、Master が停止したとき

PostgreSQL 起動！

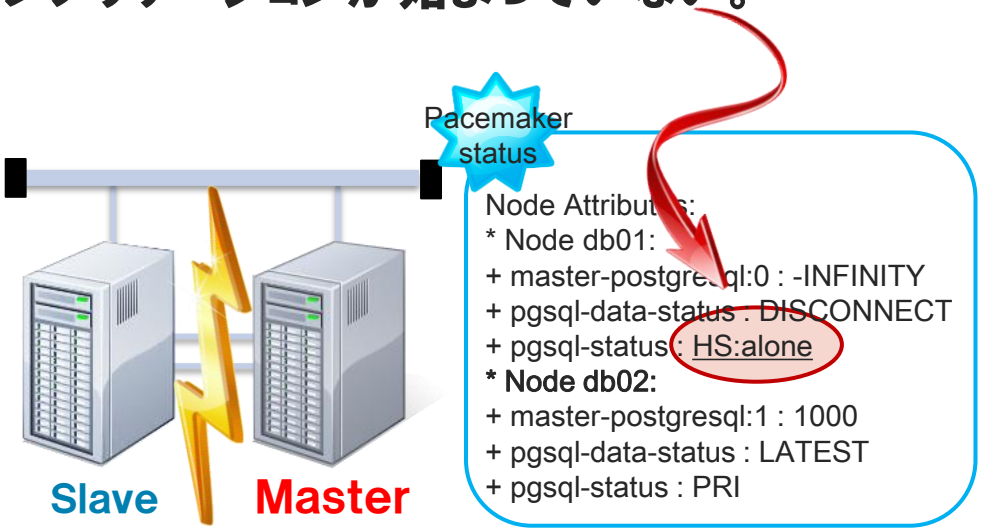
でもレプリケーションができていない。。。

検証中に発生した問題：切り戻し失敗

ロ フェイルオーバーが起こったあとクラスタの再構成（切り戻し）をしようと思ったけど、うまくできない。

原因②

レプリケーションが始まっていない。



タイムラインIDの不一致が発生している
ため、レプリケーションが開始できない。

対処②

タイムラインIDを合わせるために、再組込みの前にデータを手動で同期させる。

タイムライン ID

・・・Master／Slave のデータ整合性チェックに利用する。
Slave が Master と一致していなければ、データの整合性に問題がある可能性があるため、レプリケーションは開始されない。

変更者：PostgreSQL
変更　：Master が起動したとき。
※ PacemakerのRAは、マシンをいったん【Slave】として起動する。（※1）
その後、【Master】に昇格（promote）する。

同期レプリケーション再開！

（※1）挙動の詳細については本講演では割愛します。
本日後半の講演や Linux-HA の公開資料などを参照してください。

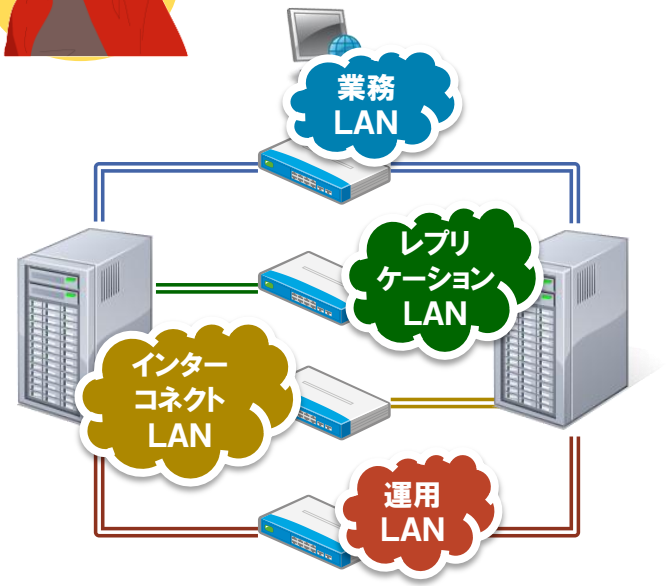
5箇条 その2：サイジングは柔軟に

NTT DATA

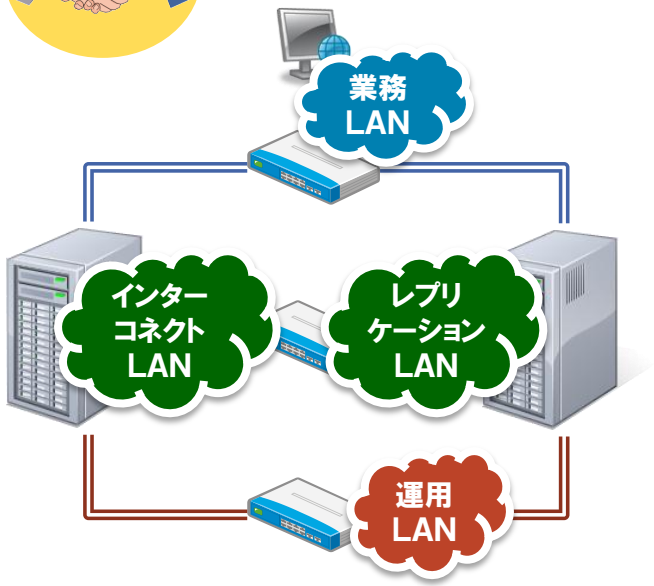
□ 適切なサイジングを行うことは、システムの安定運用に繋がります。技術的に理想である構成はありますが、実現可能な案に**柔軟に対応する**ことも必要です。

たとえば、スイッチ設置数、ポート数を考慮したネットワーク設計・・・

理想　それぞれ bonding して
4つのネットワークを構築



現実　それぞれ bondingはするけれど
環境都合でネットワークは3つに

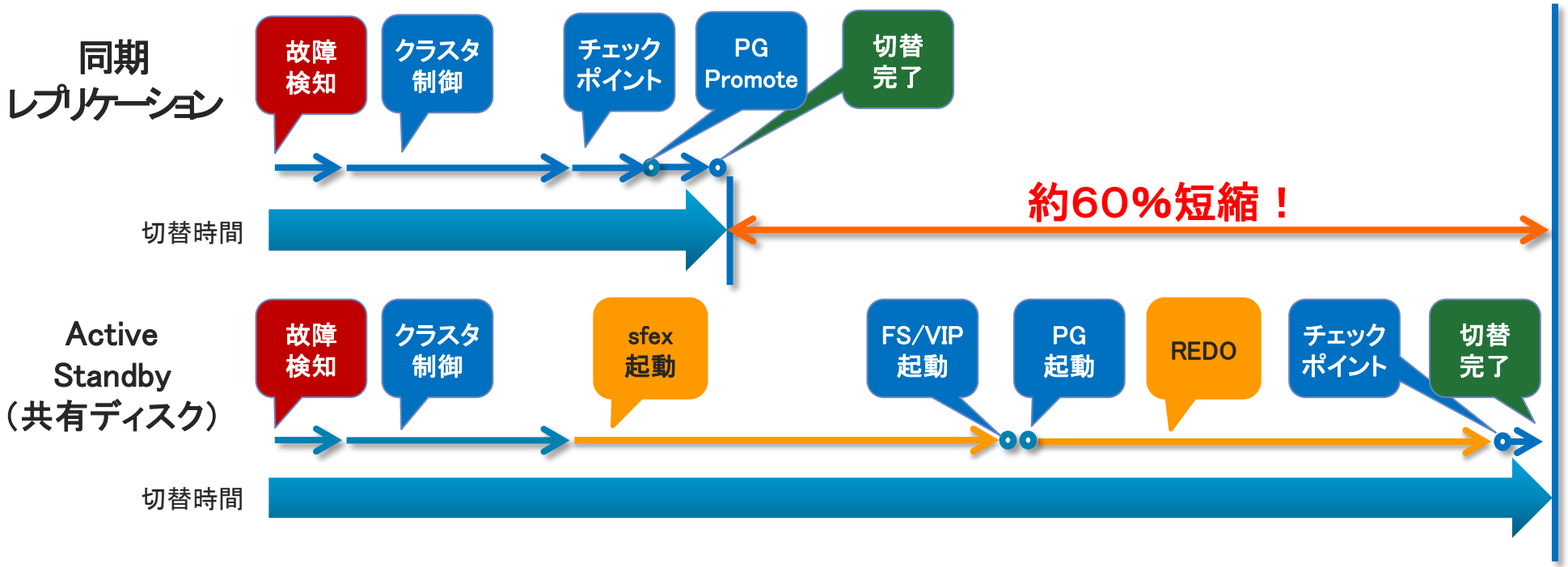


5箇条 その3：商用システムは運用が要

NTT DATA

ロ 商用システムでは運用が重要です。
いかに効率良く運用し、問題を発生させない、または発生した問題を早急に収束させることがポイントです。

たとえば、切替時間短縮を目的とした同期レプリケーション構成・・・



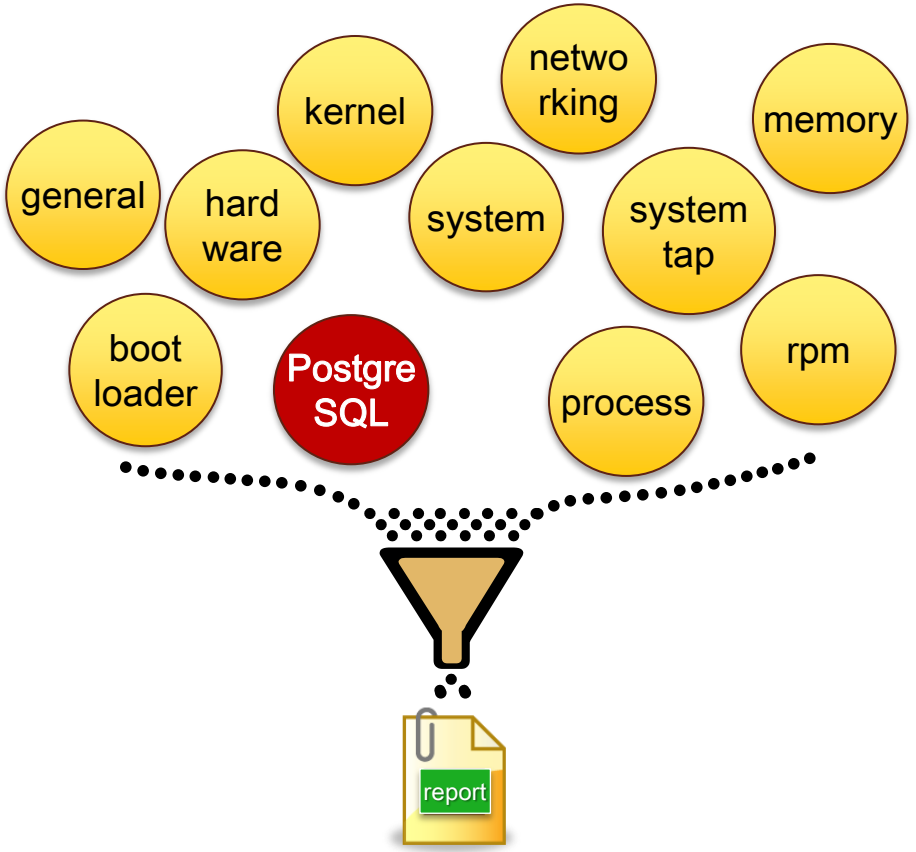
※当社検証環境において、もっとも時間がかかると想定される電源故障時の切替時間を計測

5箇条 その4：備えあれば憂いなし

NTT DATA

ロ トラブルが発生しないように準備しておくことは当然です。
トラブルが発生した場合の対処も検討しておくことが重要です。

たとえば、故障発生時のログ収集を効率よく行うためのしくみ・・・



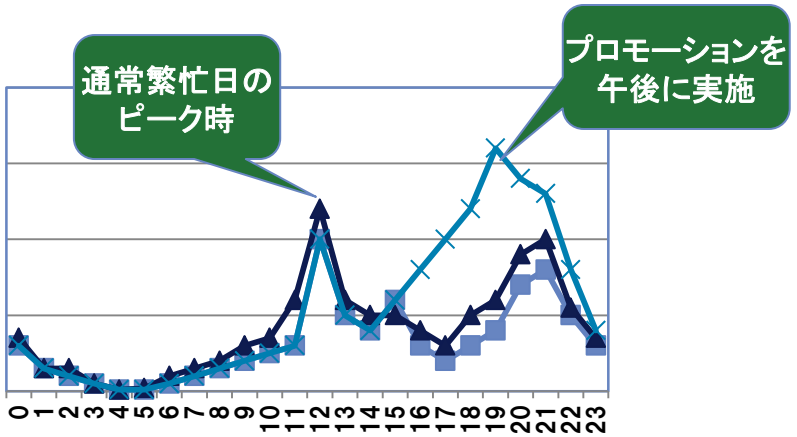
- コマンドで一つで必要なログを全て取得できる sosreport を利用します。
- スクリプト化など一括で取得するしくみを準備しておくことで、利便性の向上と取得時間の短縮を図ります。
- デフォルトでは取得できない PostgreSQL の稼働情報、ログ等々の取得もできるよう、機能追加しました。

5箇条 その5：監視で安心

NTT DATA

ロ ハードウェアの状態やプロセスのリアルタイム監視を行います。加えて、過去の性能情報等も記録し、プロアクティブな対策の検討ができるようにしておくとさらに安心です。

たとえば、アクセス情報を記録してピーク時の対応を検討・・・



図：例（とある1日のアクセス数）

- プロモーションなど、特定のイベント後の局所的なアクセス増加と成長率を考慮して、スケールアウトの検討を行います。
- システム要件（最大値）に到達する時期の再計算ができ、システムライフサイクルの見直し等を行います。

おわりに

商用システムへの適用にあたって

NTT DATA

□ PostgreSQL 9．1 を商用システムに適用し、**安定運用の実績**を作りました。

- 異常時の運用フローの整理・検討は徹底的に
  ⇒ 商用システムの肝はやはり運用！
- Pacemaker の商用システム適用は積極的に
  ⇒ サービス開始後、異常な挙動は 0 ！
- 同期レプリケーションは切替時間短縮の選択肢として有効
  ⇒ 性能への影響もありますので、考慮が必要です
- タイムラインIDの扱いについては、今後に期待
  ⇒ 柔軟な運用、クラスタ再組込み時間の短縮できませんか
- JPUG、コミュニティからの情報が頼り（何か見つけたらすぐ連絡します）
  ⇒ PostgreSQL マニュアルの誤りを修正し、投稿しました

高可用性を実現したDBアプライアンス

高可用性システム構築の敷居：高

NTT DATA

- 高可用な商用システム構築の5箇条は大事です。しかし、全てを充分に実施する場合は、システム構築に対する敷居が高くなってしまいます。

検証は大事

→ 検証マシンの調達、検証項目作成、検証実施．．．

サイジングは柔軟に

→ 案件適性やSLAの確認、諸元からのサイジング．．．

商用システムは運用が要

→ 運用手順の整理、運用手順書の作成．．．

備えあれば憂いなし

→ 都度スクリプトの作りこみ、情報取得範囲の検討．．．

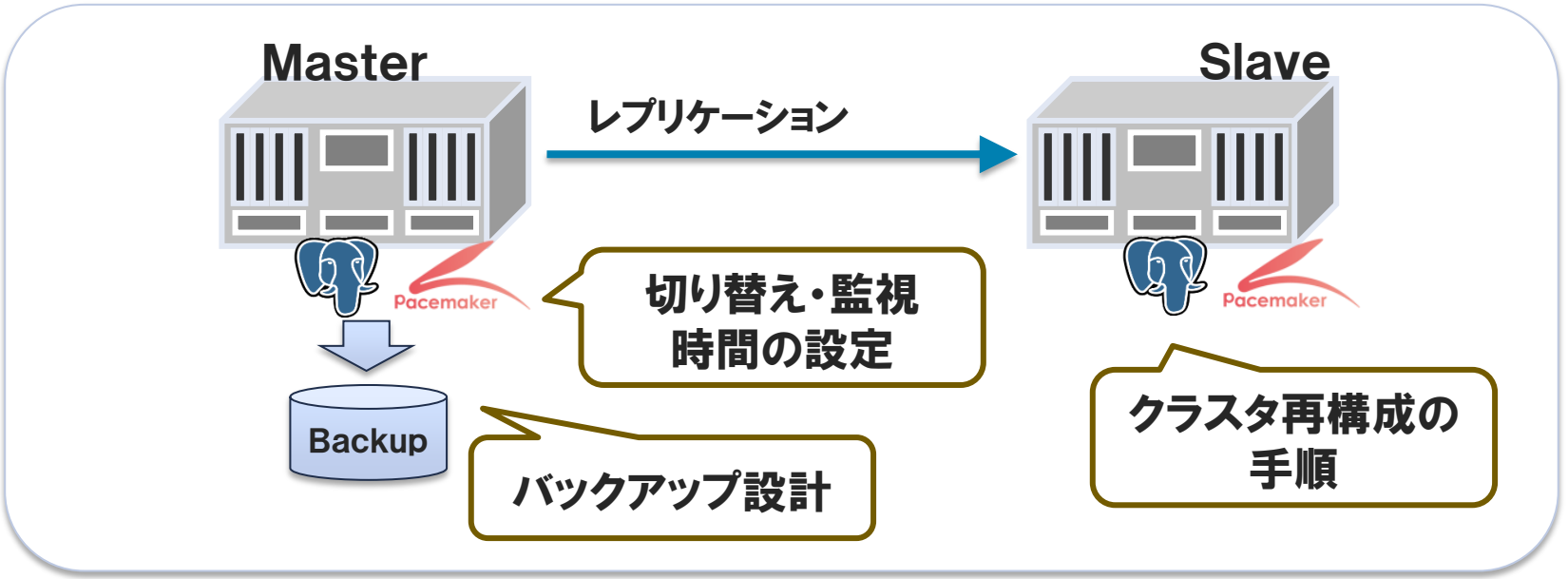
監視で安心

→ 監視パラメータの調査、設定、検証．．．

NTT DATA

- 有識者がしっかり手順を踏むことで、HA構成の高可用DBサーバ（PostgreSQL）の構築・運用はできます。
  その分、多くの**コストと時間**が必要です。

今回の事例システムの構築（DB）に必要なタスクの一部



NTT DATA

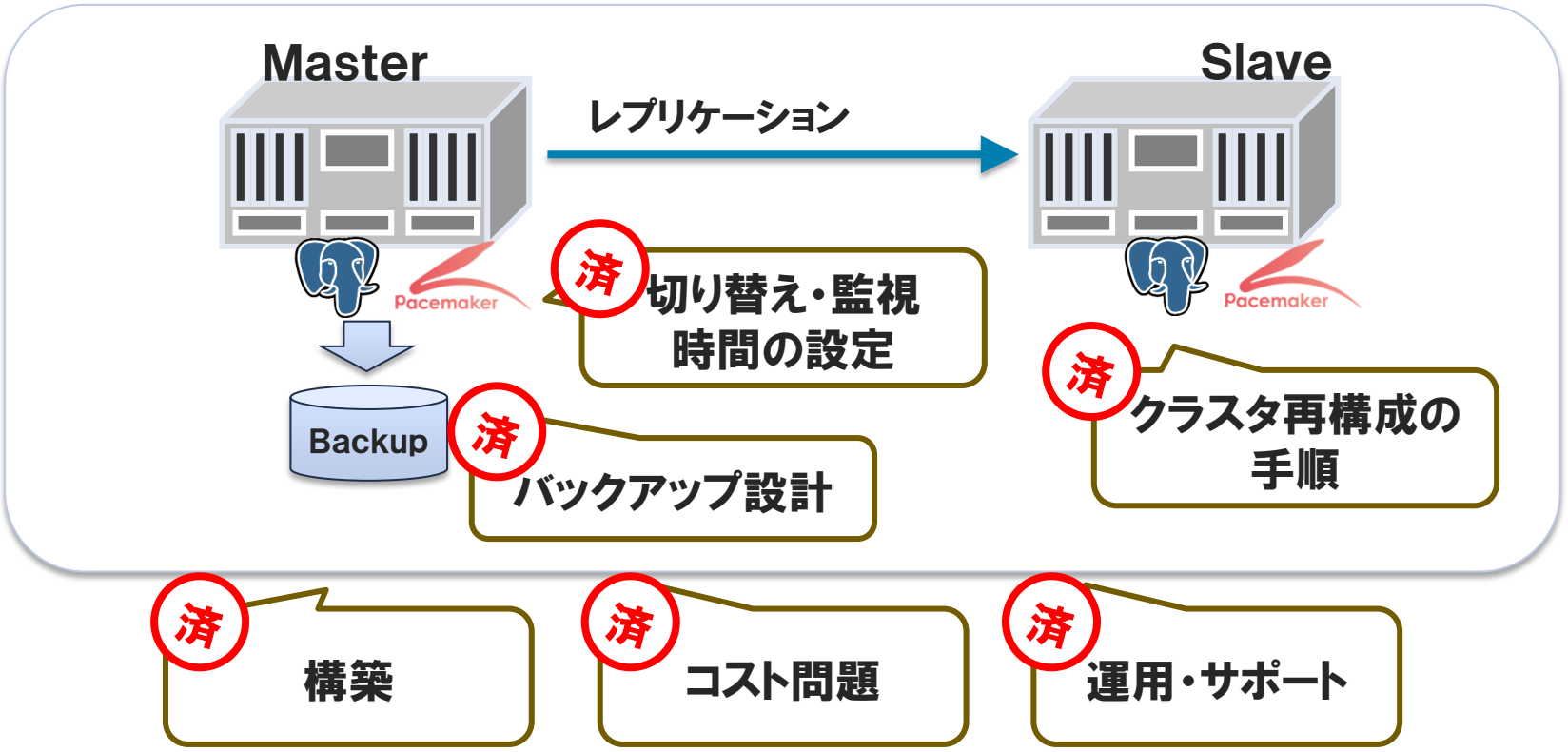
高可用性を実現したDBアプライアンス

はじめました

NTT DATA

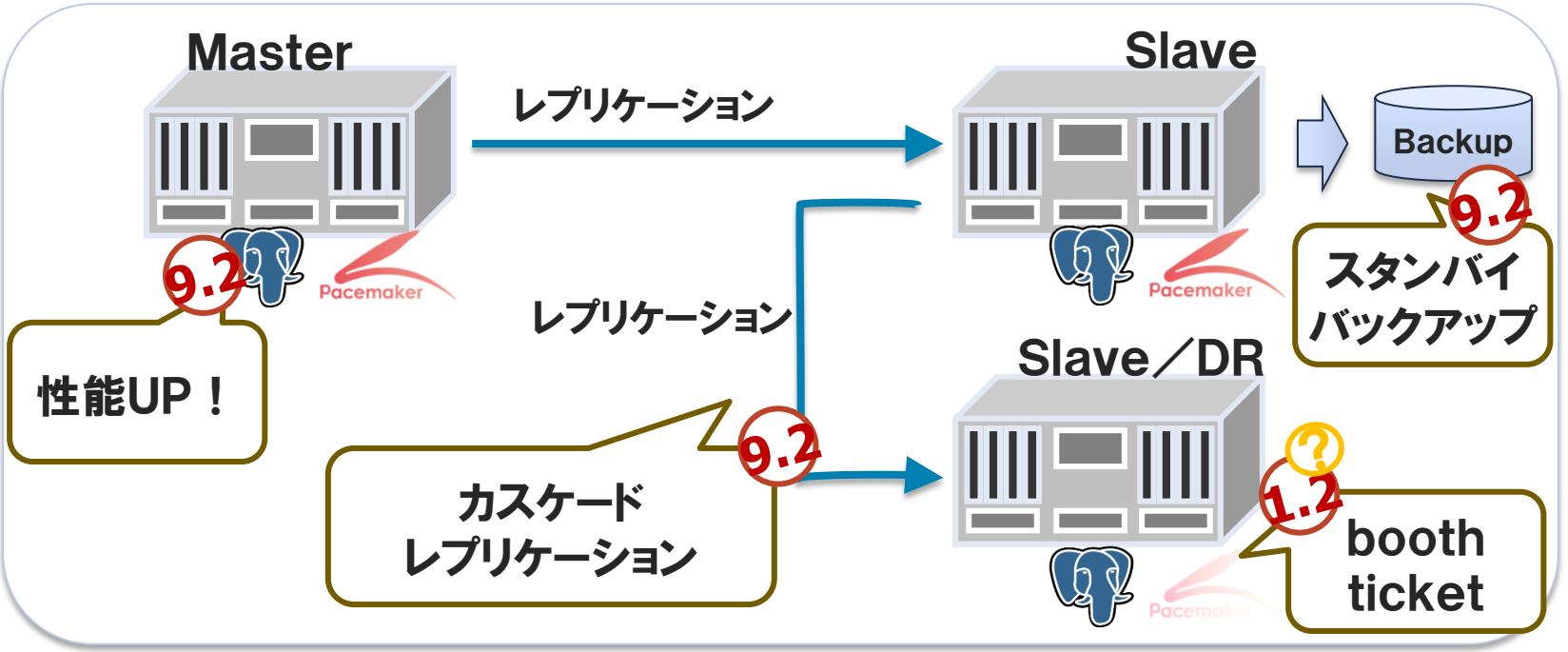
□ 高可用なシステムを、すぐにご利用でき、簡単な操作での運用が可能でかつ安心のサポート込み。さらに低価格でご提供します。

今回の事例システムの構築(DB)に必要なタスクは当然、実施**済**でご提供



NTT DATA

□ 最新バージョンのPostgreSQLを利用することにより、高性能化、高機能化を検討しています。最近ご要望を多くいただいている、ディザスタリカバリ機能もご提供いたします。

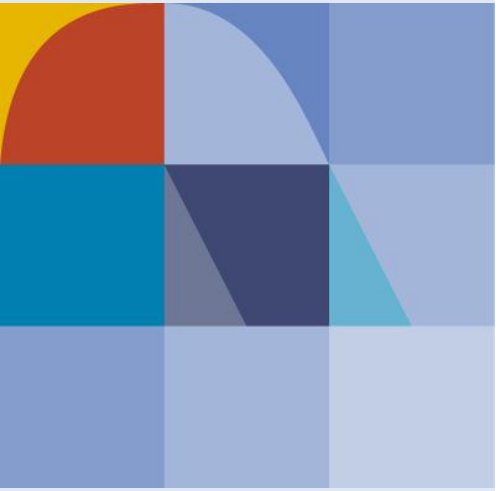


最新をｗａｔｃｈして、最適な機能を採用！

□ GresCubeとは・・・

新機能開発を継続して実施します

次の領域へどんどん拡大していく
ＧｒｅｓＣｕｂｅ の今後にご注目ください



NTT DATA

変える力を、ともに生み出す。