より速く・より大きく・より強く
～企業ユーザから見たPostgreSQLの歴史と展望

PostgreSQL Conference Japan 2018

2018.11.22

NTT OSSセンタ 坂田 哲夫

Copyright©2018 NTT Corp. All Rights Reserved.

自己紹介

- 坂田 哲夫 (さかた･てつお)
- NTT OSSセンタ勤務
  - PostgreSQL担当 (2006～現在)
  - 自社開発ツールの維持管理、開発の支援など
    - 維持管理：PG-REX, pg_statsinfo etc.
    - 開発支援：暗号化関連
- コミュニティ活動
  - JPUG理事：勉強会担当(2005～現在)
  - PostgreSQL Enterprise Consortium にも参加

Copyright©2018 NTT corp. All Rights Reserved.

発表の背景

企業ユーザとしてのNTTグループ

Copyright©2018 NTT Corp. All Rights Reserved.

NTT OSSセンタとは

【OSSセンタのミッション】
NTTグループの収益性改善とビジネスモデル変革をOSSを活用して貢献する

OSSセンタの活動理念

私たちは、革新的で快適なICTサービスの安価な実現に貢献するため、全世界のOSSコミュニティと連携したオープンイノベーションやNTTグループのR&D機能を活かし、OSSを中核とする安定して利用できるソフトウェア開発とそのサポートサービスを提供します。

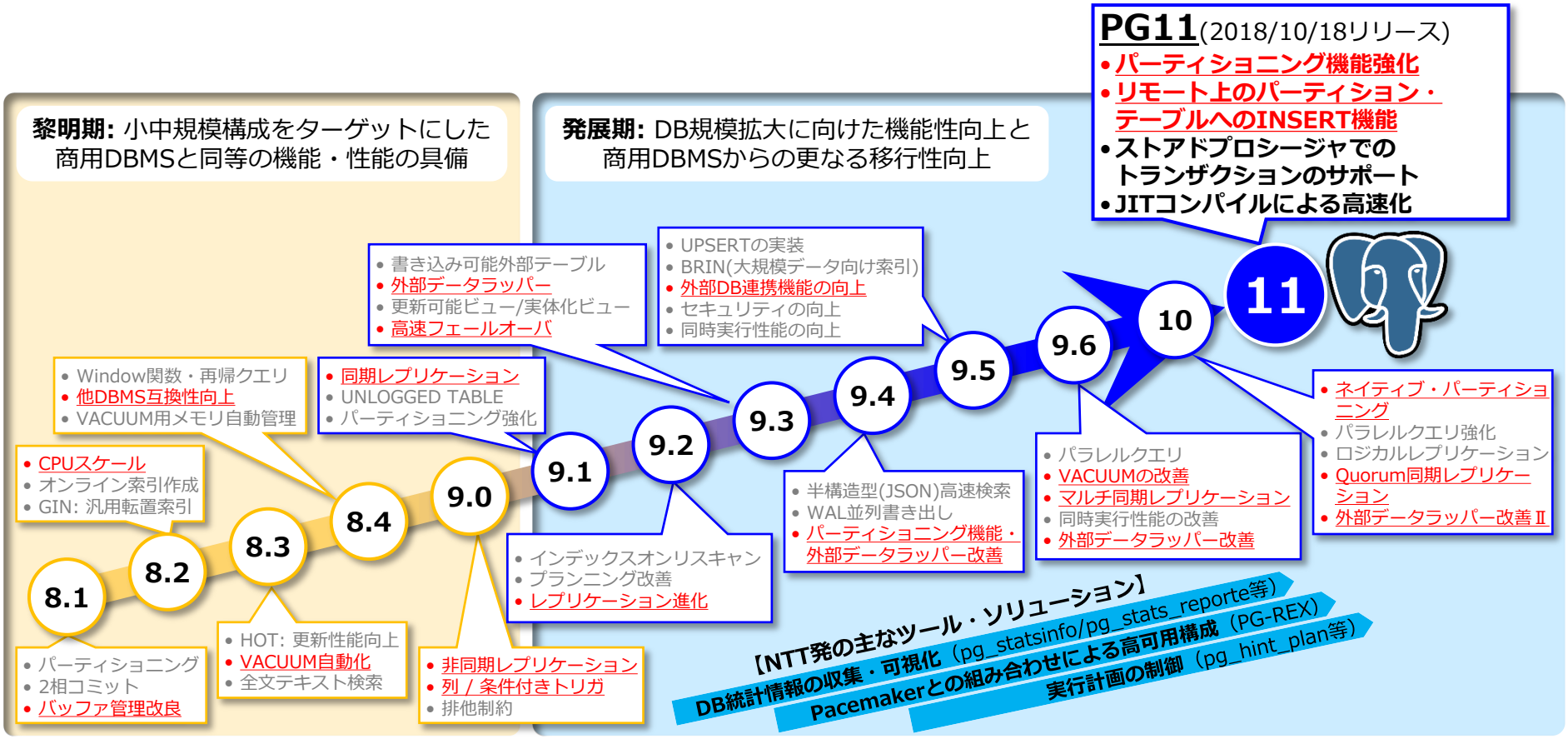
1)ICTシステムのライフサイクルにわたるTCO削減への貢献
2)革新的ICTサービス創生・ソリューション拡大への貢献
3)グローバルなオープンイノベーション市場で活躍できる人材の育成

*)**OSSVERT®**：OSs Suites VERified Technically（技術検証済みOSS組合せ）

Copyright©2018 NTT Corp. All Rights Reserved.

OSSセンタでのPostgreSQLへの開発貢献

PostgreSQLの機能・性能改善のための開発に貢献。国内外の会議に参加して活発に議論するなど、積極的なコミュニティ活動によってオープンイノベーションを推進しています。
また、高可用構成、実行計画制御といったエンタープライズ向け機能も開発し、NTTビジネスに貢献しています。



下線はNTTがコミュニティ提案し実現した機能

Copyright©2018 NTT Corp. All Rights Reserved.

より速く・より強く

POSTGRESQLの進化を振り返る

Copyright©2018 NTT Corp. All Rights Reserved.

何が重要か：DBMSの企業向けニーズ

- **より速く：性能**
  - 速度が速いことは、コストパフォーマンスが高いことに通じる
- **より大きく：スケール性**
  - ビジネスの成長と共に増加するデータを処理できる
    ⇒スケールするDB
- **より強く：可用性・セキュア性**
  - 故障に強い・災害に強いDB
  - セキュアなDB

Copyright©2018 NTT Corp. All Rights Reserved.

PostgreSQLの進化を振り返る

- **PostgreSQL 7.4(2003.11)の頃を振り返る**
  - NTTの取り組みはこのころ始まる
- **7.4の頃から「企業向け」の機能を強化し始める**
  - コミュニティ重鎮のB.Momjian氏の発言

今後は性能向上など、
エンタープライズ向けのニーズ
にも応えていきたい

※当時の東京での講演での発言

Copyright©2018 NTT Corp. All Rights Reserved.

PostgreSQL7.4の問題点

- **チェックポイントの実行中に性能が低下する**
  - その間PostgreSQLの応答が止まってしまう
- **マルチCPUで性能が出ない(scale upしない)**
  - 2CPUまでしかスケールしない
- **VACUUMに時間がかかる**
  - VACUUM時間がDBサイズを制約
- **DBの信頼性が低い**
  - ディスククラッシュ直前のDBが復元する手段がない
  - 故障時にフェイルオーバする手段がない

速くない

大きくならない

強くない

Copyright©2018 NTT Corp. All Rights Reserved.

PostgreSQLの初期の進化

- **速く・大きく・強いDBに進化**
  - 8.0(2005)から9.1(2011)にかけて基本機能が充実
  - 関連企業の貢献(EnterpriseDB, 2nd Quadrant, Red Hat etc.)
  - 8.3以降、業務システムへの適用進む

課題	解決策	実現ver.	貢献企業
チェックポイントでの性能低下	負荷分散チェックポイント	8.3	NTT, EDB*1
マルチCPU	ロック方法改善	8.2, 8.3	Red Hat, NTT
VACUUM時間	マルチワーカ	8.3	NTT, 2Q*2
	HOT	8.3	EDB
DBの信頼性	物理バックアップ	8.0	2Q
	同期レプリケーション	9.0(非同期) 9.1(同期)	NTT, 2Q

*1 EDB⇒ EnterpriseDB
*2 2Q⇒ 2nd Quadrant

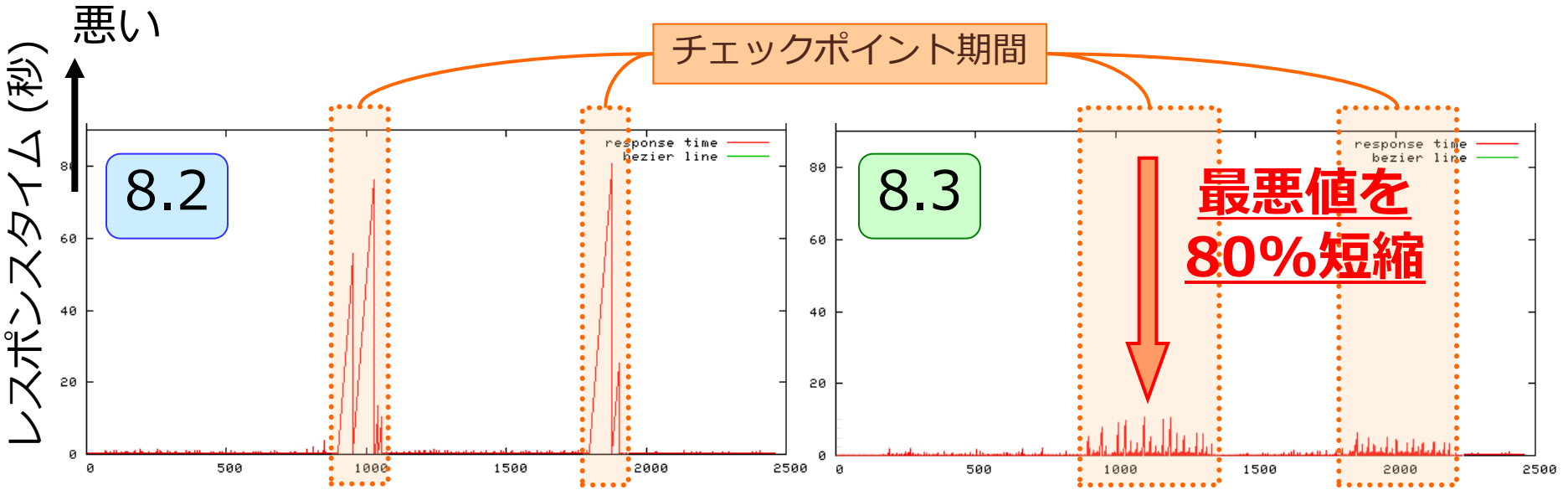
Copyright©2018 NTT Corp. All Rights Reserved.

10年前のデータに見る PostgreSQLの改善

Copyright©2018 NTT Corp. All Rights Reserved.

チェックポイント時の性能安定化

チェックポイント時に1分くらい「止まっていた」のが
10秒程度に。安定した性能でユーザの安心感高まる



チェックポイント(注)時の
レスポンスが悪い！

安定した性能を維持

負荷分散チェックポイント

（注）メモリとディスク上のデータとの同期処理

pgbench –c10 –s400 (10接続, 6GB)

Copyright©2018 NTT Corp. All Rights Reserved.

現在のPostgreSQLの到達点

Copyright©2018 NTT Corp. All Rights Reserved.

企業ニーズに応えるPostgreSQL

- **最近の大規模DB事例から**
  - DBサイズは数TB～数十TBを実現可能

利用分野	用途	PG ver.	DBやアプリケーションの特徴	資料★
通信	通信料金計算	9.2	• DBは数TB • read replicaによる負荷分散 • 約5000のバッチジョブ	1
電力	電力スマートメータ	9.4	• DBは11TB • 1日あたり2.4億件のデータ挿入	2
ネット	ソーシャルネット	9.3	• DBは全部で22TB • シャーディング(64サーバ) • システム全体で1.3M TPS	3
科学	天体DB	9.4	• DBは40TB • 1000万件/日の問合せ(≒100 TPS)	4

★末尾の参考文献を参照

Copyright©2018 NTT Corp. All Rights Reserved.

コミュニティ開発の実態

新機能への長い道のり

Copyright©2018 NTT Corp. All Rights Reserved.

コミュニティのメンバ企業による開発

- **NTT OSSセンタによる開発の事例を紹介**
- **ポイント**
  - コミュニティへのアピール
  - 長期にわたる開発と段階的な発展
  - コミュニティでの協力関係の”意味”

Copyright©2018 NTT Corp. All Rights Reserved.

レプリケーション開発前夜

- **当時のコミュニティの状況**
  - レプリケーション機能の提案がたくさんあった
    - Pgpool-II, Slony-I, Bucardo, Londiste, Continuent etc.
  - 一長一短があり、どれも決定版ではない
    - コミュニティの公式の立場は「本体でなく外部で実現」だった
- **レプリケーション機能の開発**
  - 8.1のころ、開発スタート(NTT内部)
    - DBの2重化による高信頼化を狙う
    - 8.3のころ、OSS版を公開
  - オタワのPGCon(2008)で発表
    - 信頼性向上に必須の機能であることをアピール

電話交換機用の
高信頼DBがルーツ

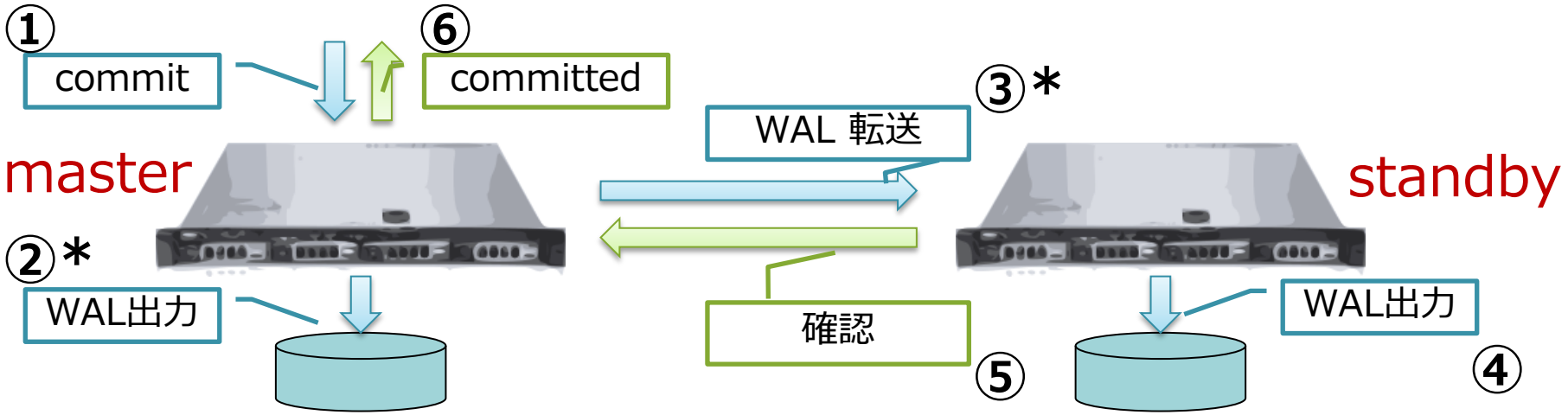
Copyright©2018 NTT Corp. All Rights Reserved.

WAL shipping方式のレプリケーション

- **動作概要**
  - コミット時(②)にWALをスタンバイサーバに転送(③～④)
  - スタンバイへの書き込み後にコミットする(⑤,⑥)

特徴

- masterで「コミット済」のトランザクションのデータは、確実にstandbyに複製される
- オーバヘッドが小さく、高速に動作する

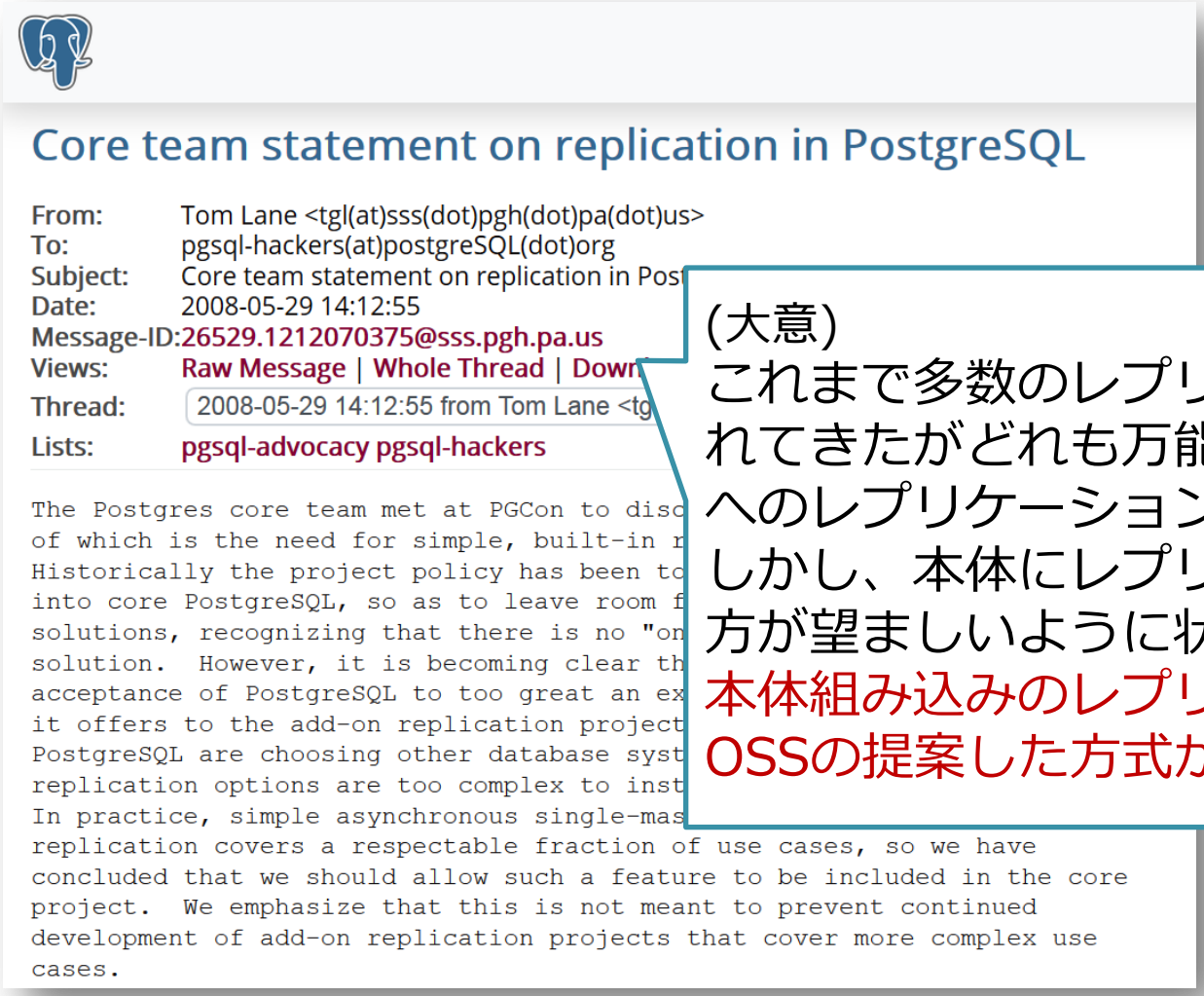


＊②と③～⑤は並行して実行される

Copyright©2018 NTT Corp. All Rights Reserved.

レプリケーション開発へ

- PGConで発表後のコアチームからのメッセージ



Core team statement on replication in PostgreSQL

From: Tom Lane <tgl(at)sss(dot)pgh(dot)pa(dot)us>
To: pgsql-hackers(at)postgreSQL(dot)org
Subject: Core team statement on replication in Post
Date: 2008-05-29 14:12:55
Message-ID: 26529.1212070375@sss.pgh.pa.us
Views: Raw Message | Whole Thread | Down
Thread: 2008-05-29 14:12:55 from Tom Lane <tg
Lists: pgsql-advocacy pgsql-hackers

The Postgres core team met at PGCon to disc
of which is the need for simple, built-in r
Historically the project policy has been to
into core PostgreSQL, so as to leave room f
solutions, recognizing that there is no "on
solution. However, it is becoming clear th
acceptance of PostgreSQL to too great an ex
it offers to the add-on replication project
PostgreSQL are choosing other database syst
replication options are too complex to inst
In practice, simple asynchronous single-mas
replication covers a respectable fraction of use cases, so we have
concluded that we should allow such a feature to be included in the core
project. We emphasize that this is not meant to prevent continued
development of add-on replication projects that cover more complex use
cases.

(大意)
これまで多数のレプリケーションが提案されてきたがどれも万能ではなく、本体機能へのレプリケーション実装は見送ってきた。
しかし、本体にレプリケーションがあった方が望ましいように状況が変わってきた…
本体組み込みのレプリケーションは、NTT OSSの提案した方式が良いと思われる。

https://www.postgresql.org/message-id/26529.1212070375@sss.pgh.pa.us

Copyright©2018 NTT Corp. All Rights Reserved.

レプリケーション開発の歴史

- **段階的な機能の発展**
  - 9.0で「非同期」レプリケーションが入る(NTT, 2Q, EDB)
  - 9.1で「同期」レプリケーションが入る(NTT)
    - コミット完了＝スタンバイに複製済み
  - 9.2でカスケードレプリケーション(NTT)
  - 9.3でスタンバイ→マスタの昇格を高速化(2Q, NTT)

バージョンが8.5でなく9に

提案機能を本体で実現

1分以内のfail over が可能に

ここまで提案から 約5年

Copyright©2018 NTT Corp. All Rights Reserved.

レプリケーション機能の開発を振り返る

- **機能の重要性をアピールすることは重要**
  - 大きな機能では、多数のメンバの協力が不可欠
- **機能の開発は段階的に進む**
  - 「非同期」→「同期」→「fail over高速化」
- **提案した企業以外も開発をサポート**
  - 2nd Quadrant, EnterpriseDBのメンバがレビューしてくれた

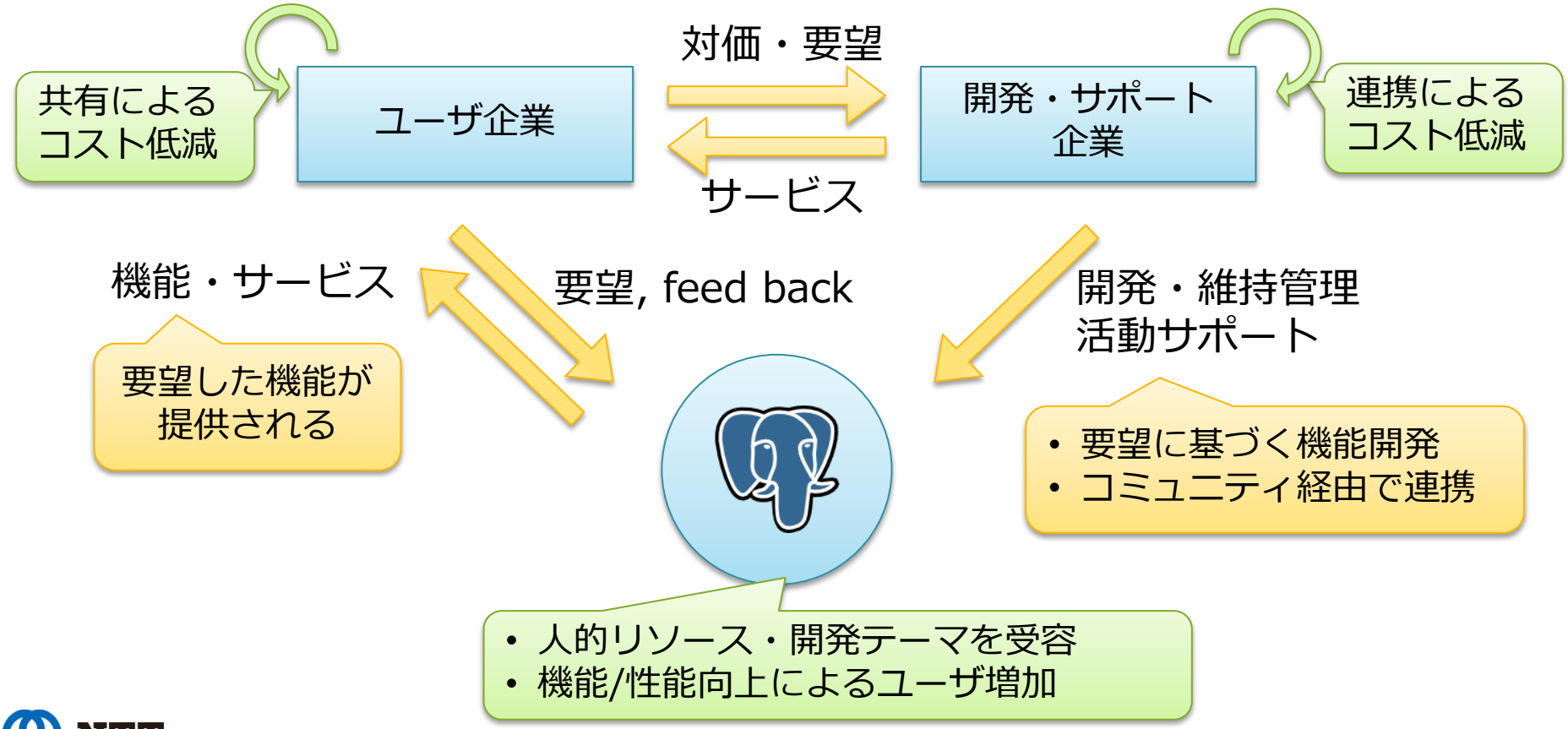
メンバの合意と自発的な協力による
ゆるやかな開発体制

その背景に**エコ・システム**

Copyright©2018 NTT Corp. All Rights Reserved.

開発と維持・利用のエコ・システム

- **PostgreSQL改善のメリットを共有**
  - 各プレイヤーはコミュニティを通じてメリットを共有
  - PostgreSQLの改善により更にメリットが増進



Copyright©2018 NTT Corp. All Rights Reserved.

より大きく・より強く

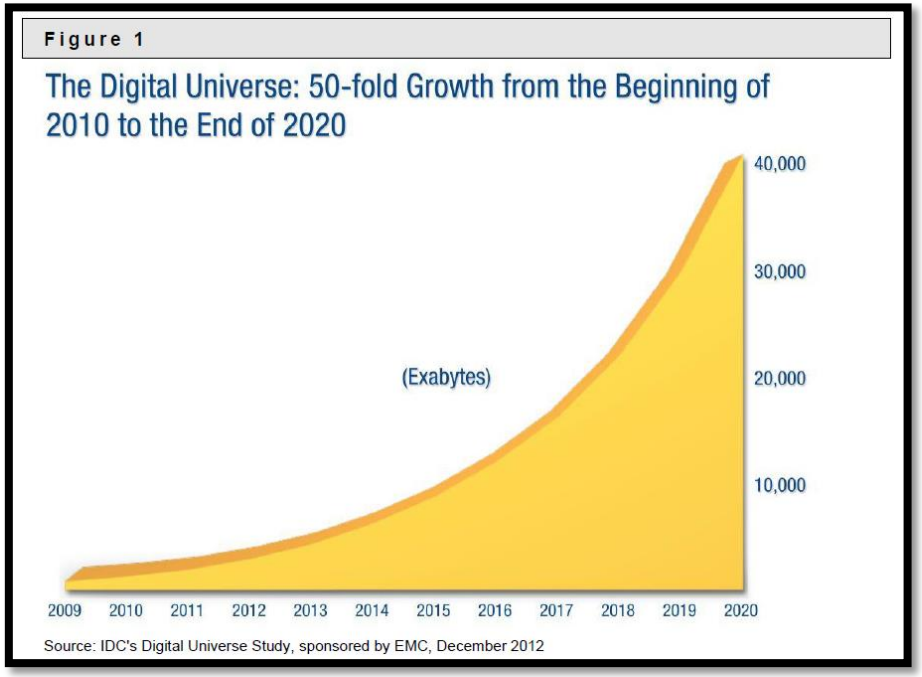
POSTGRESQLの将来

Copyright©2018 NTT Corp. All Rights Reserved.

PostgreSQLを取り巻く環境

- **インターネットの発達は大量データ処理をもたらした[1]**
  - データ量は2012年以降毎年2倍のペースで増える
  - 2020年には40ペタバイトに達する見込み
- **データの処理はクラウドに移行[1]**
  - 全データの約4割がクラウド上で処理

大規模データと
クラウド対応が
PostgreSQLの
発展の課題



[1] John Gantz and David Reinsel, "The Digital Universe in 2020", IDC report, 2012.

Copyright©2018 NTT Corp. All Rights Reserved.

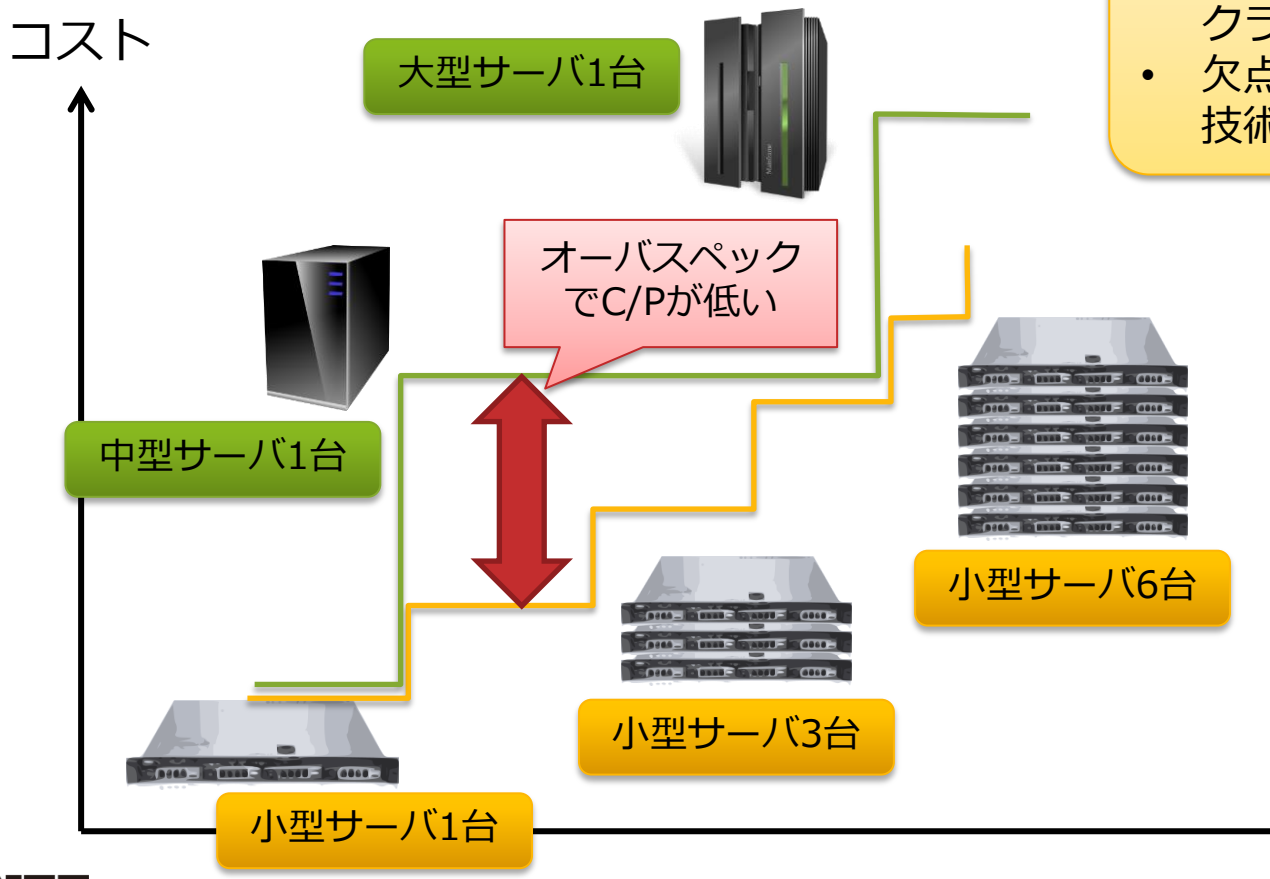
情報爆発時代のPostgreSQL

より大きなDBの実現

Copyright©2018 NTT Corp. All Rights Reserved.

大規模データへのアプローチ

- **スケールアップ**
  - より高性能なマシンに置き換える
- **スケールアウト**
  - マシンの数＝DBクラスタ



スケールアウトの特徴
- 利点
  スモールスタートできる
  C/Pが良い
  クラウドとも相性良い
- 欠点
  技術的な難易度が高い

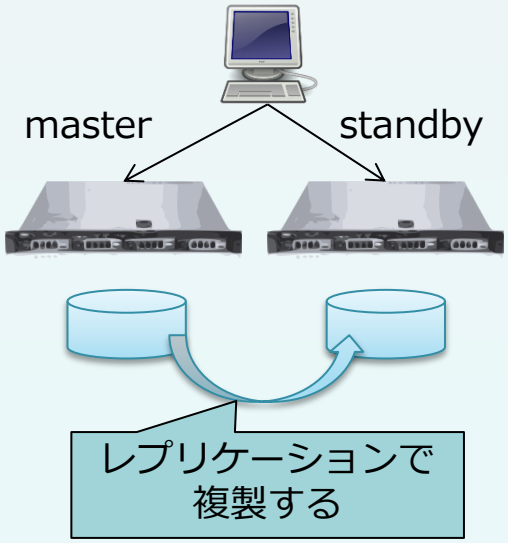
Copyright©2018 NTT Corp. All Rights Reserved.

代表的なScale Out向けDBクラスタ

レプリケーション

利点
- 実装が一番容易
- 参照負荷はスケール

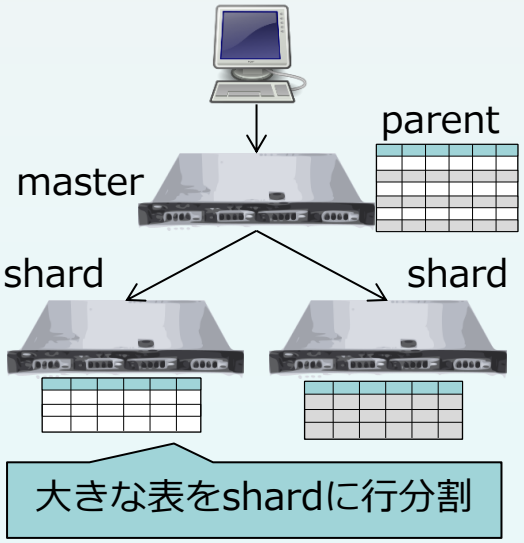
欠点
- 更新はマスタのみ (AP側で振り分ける)



シャーディング

利点
- 実装が比較的容易
- 参照負荷はスケール

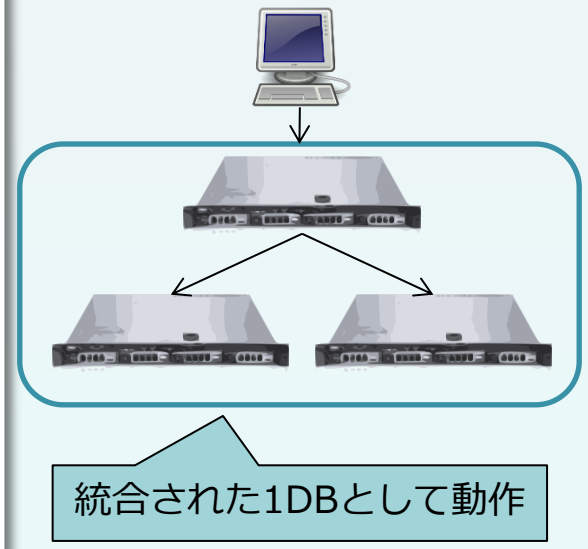
欠点
- 更新の一貫性に制約



分散DBMS

利点
- 参照負荷・更新負荷ともスケール
- 一貫性が保証される

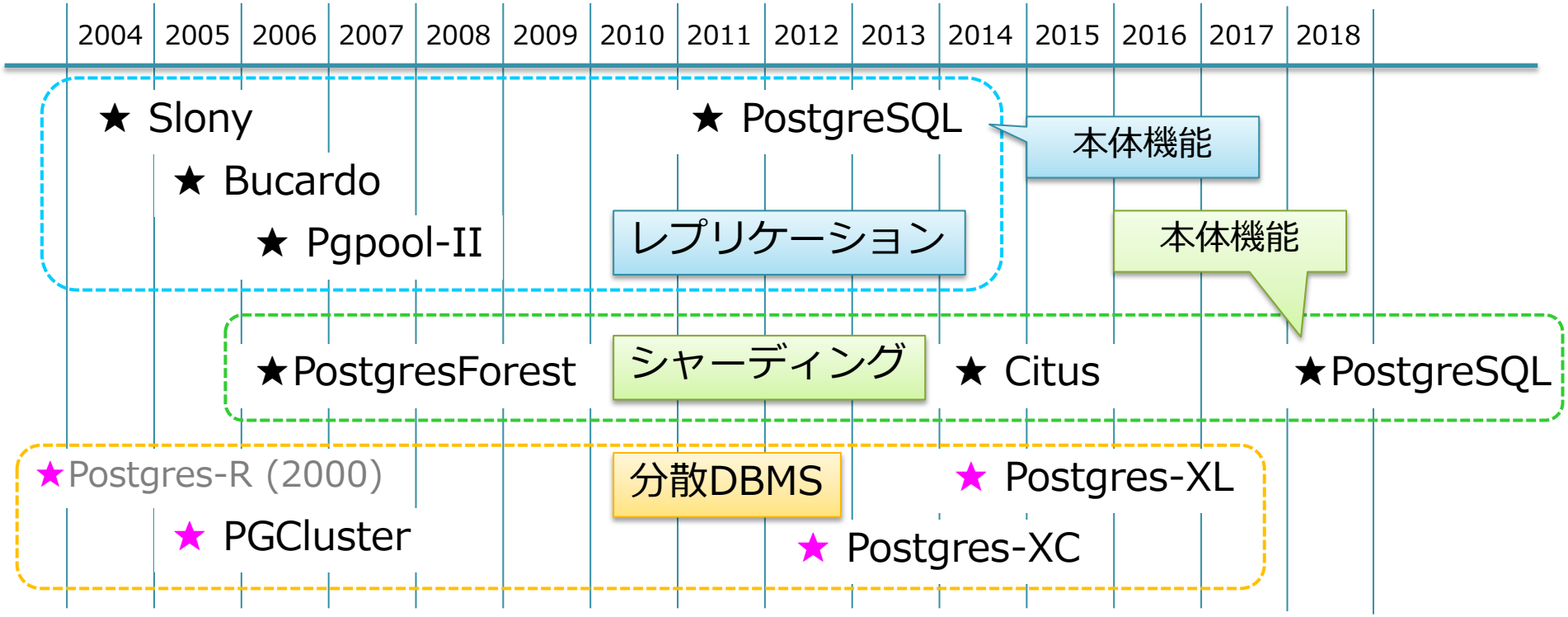
欠点
- 実装の難易度が高い



Copyright©2018 NTT Corp. All Rights Reserved.

DBクラスタの開発年表

- **レプリケーション, シャーディング, 分散DBMSの各方式が提案**
  - 実現方法は★外部拡張と★本体改造
- **要望が強い機能は本体に取り込まれる**
  - レプリケーション(9.0～9.1)、シャーディング(9.3～11)
  - 次の目標は本体機能の分散DBMS化



★ 本体改造

Copyright©2018 NTT Corp. All Rights Reserved.

PostgreSQLの分散DBMS化

- **分散DBMSはPostgreSQL本体で実現するのが理想**
  - 幅広い用途に対応可能
  - 開発も維持も高コスト

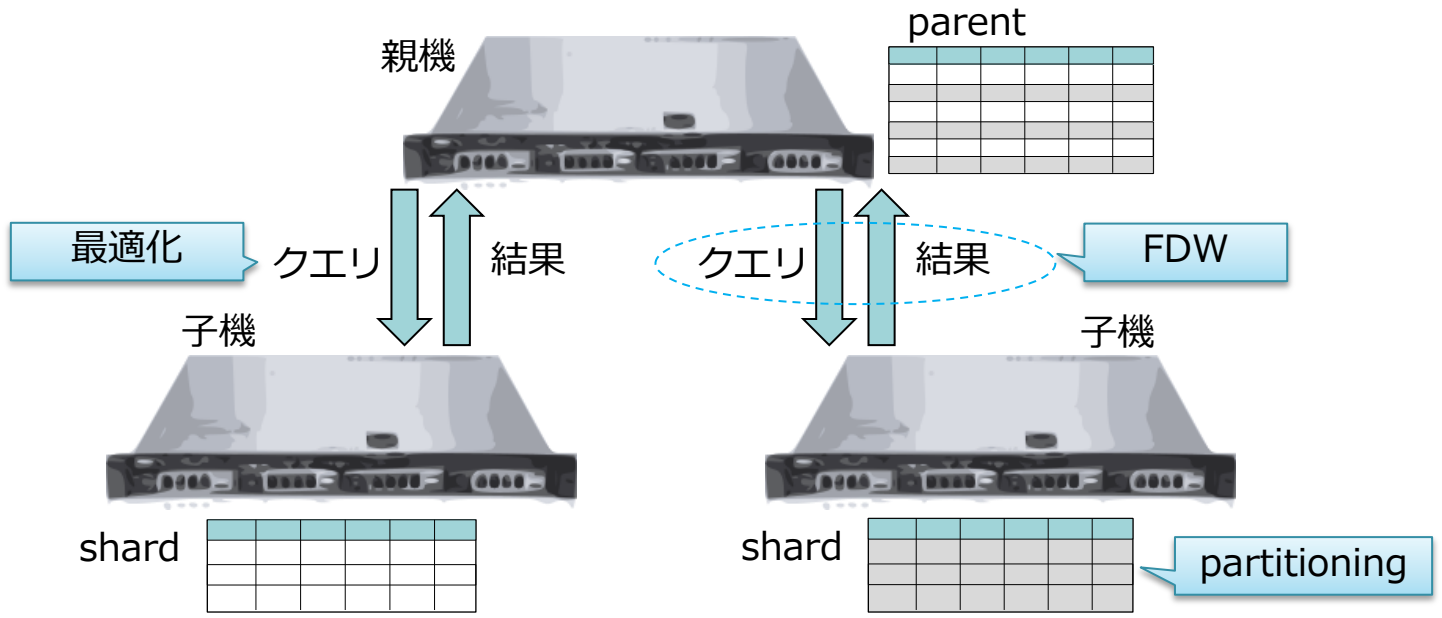
  } コミュニティ全体で取り組むことで メリット最大化・リスク低減

- **シャーディングからのアプローチ**
  - シャーディング＋トランザクション管理で分散DBMS化
  - 段階的アプローチが可能
    - コミュニティ開発では重要
    - シャーディング自体が段階的に開発された

Copyright©2018 NTT Corp. All Rights Reserved.

シャーディングへのアプローチ

- **シャーディングの要素技術：**
  パーティショニング＋FDW＋最適化
  - パーティショニング：表を論理的にシャードに分割する
    - V.10で”宣言的パーティショニング”を実現
  - FDW：親機から子機にクエリを送って処理させる
    - V.9.1でコア機能を実現。9.3で外部DBにPostgreSQLが接続可能
  - 最適化：子機で出来る処理は子機で実行 (push down)
    - 9.5(結合), 10(集約)



Copyright©2018 NTT Corp. All Rights Reserved.

分散DBMSに向けての課題

- **分散したサーバ上でのトランザクション管理機能が必要**
  - Atomic Commit 機能がVer.12に向けて提案中
  - 将来的には更に効率的な方式が望ましい
- **SQL実行系の強化**
  - 非同期実行(asynchronous exec.)による効率化
  - パーティション化された表の効率的な処理
    - 複数パーティションにまたがる表への統合的なインデックス
- **運用系の充実**
  - バックアップやレプリケーションの実現、クラスタの高可用化
  - 業務システムへの適用に不可欠な機能

Copyright©2018 NTT Corp. All Rights Reserved.

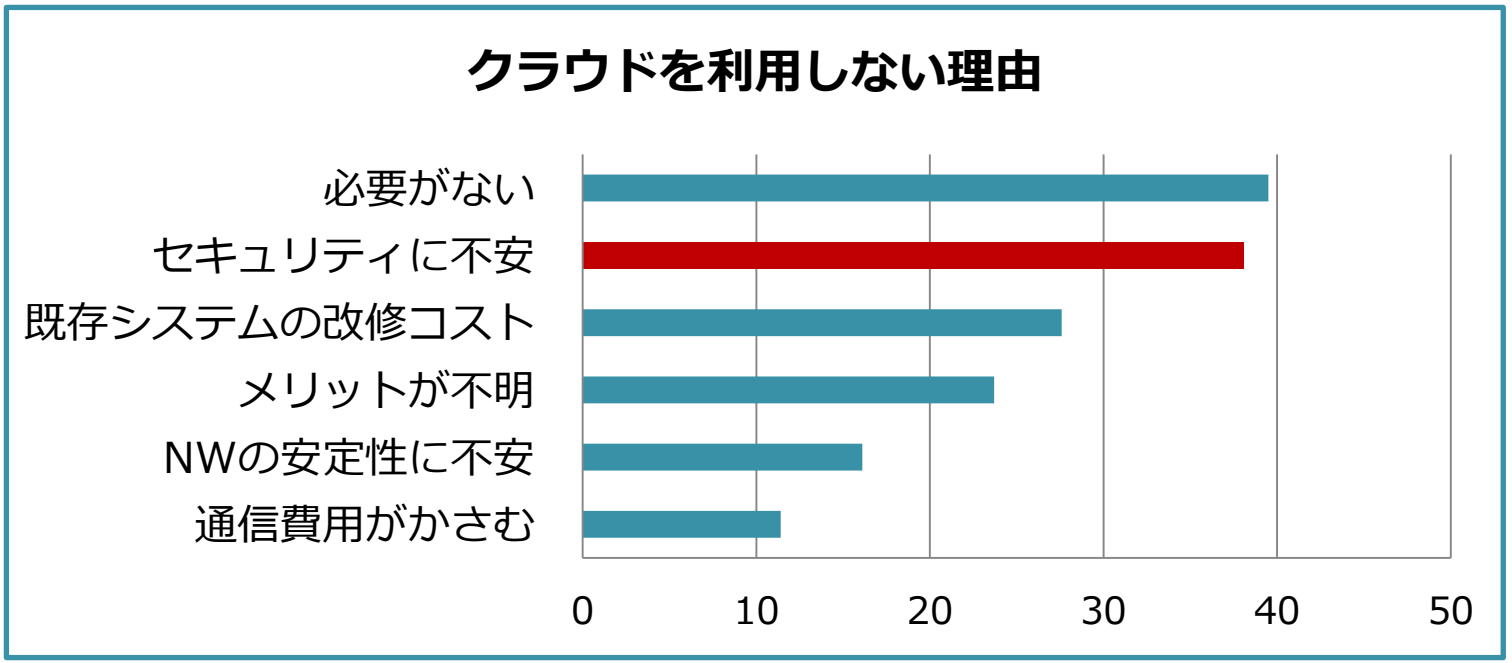
クラウド時代のPostgreSQL

より強いDBの実現

Copyright©2018 NTT Corp. All Rights Reserved.

クラウド化はより強いDBを求める

- **全世界のデータの約4割はクラウドで処理される＊1**
- **国内企業の約6割はクラウドを利用＊2**
  - 利用しない理由の第2位はセキュリティ不安



※クラウドを利用しないと回答した企業(約43%)にその理由を問うたもの

＊1：John Gantz and David Reinsel, "The Digital Universe in 2020", IDC report, 2012.
＊2：総務省, “平成29年通信利用動向調査”, 2018.

Copyright©2018 NTT Corp. All Rights Reserved.

DBMSのセキュリティ強化とは？

- **DBMSのセキュア化の3分野***
  - アクセス・利用の制限
    - 認証機能：正しいユーザを識別
    - 権限機能：ユーザの適切な権限
  - 不正追跡・監視
    - 監査ログ：データ操作の記録
  - データの秘匿
    - データの暗号化
    - 通信の暗号化

セキュリティの強化は
これらすべてが必要

＊ IPA “非機能要求グレード2018”を基に作成.

Copyright©2018 NTT Corp. All Rights Reserved.

PostgreSQLのセキュア化の歴史

- **セキュリティ機能3分野の強化**
  - 認証と権限の開発は活発だがそれ以外は低調

ver.	認証・権限	監査	暗号化
9.0以前	• LDAP認証(8.2)	• 基本のログ監査	• pg_crypto(8.3)
9.1～9.3	• SE Linux連携(9.1)		
9.4～9.5	• 行単位セキュリティ(9.5)	• (pgaudit) ★	• (TED for PG)★
9.6	• デフォルトロール		
10	• scram認証		
11	• デフォルトロール強化		

★ カッコつきは3rd party OSS。特記なき項目は本体機能

Copyright©2018 NTT Corp. All Rights Reserved.

セキュリティ機能の現状と課題

- **認証と権限** 
  - 現状：基本機能はOK
    - ユーザを一意に識別して認証できる
    - 一般ユーザのために適切な権限が設定できる
  - 課題
    - 特権ユーザの権限が強すぎる⇒補助的な制限手段が必要
- **監査** 
  - 現状：基本機能はOK
    - 本体のログ監査やpgauditを用いて多くのニーズに対応できる
  - 課題
    - 特権ユーザが監査を無効化できてしまう⇒補助的な制限手段が必要
- **暗号化** 
  - 現状：最低限の機能はある
    - pgcryptoでデータを安全に暗号化できる
  - 課題
    - AP側の暗号化の対応作業が多い
    - 鍵管理のための作り込みが必要

NTT OSSセンタの取り組みは、
14:10からBトラックにて

Copyright©2018 NTT Corp. All Rights Reserved.

Closing

コミュニティの課題と展望

Copyright©2018 NTT Corp. All Rights Reserved.

PostgreSQLの課題と展望

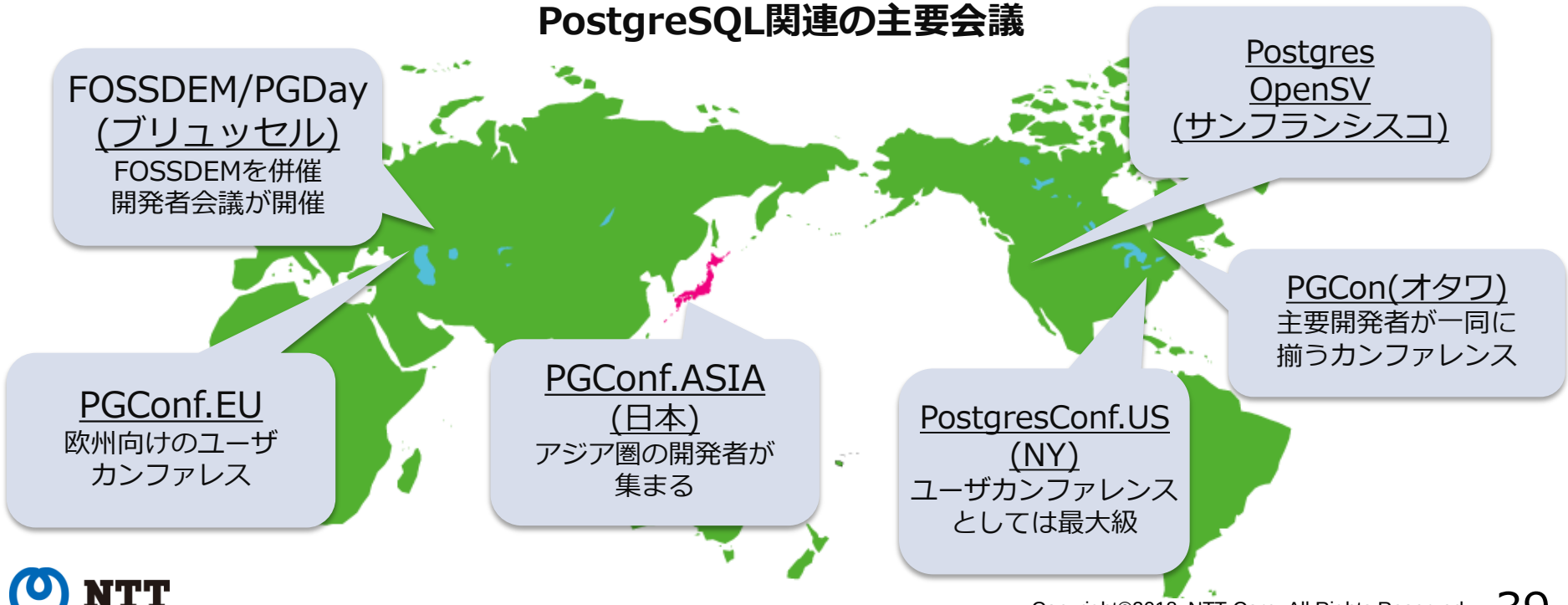
- **課題**
  - 高度な要望に応えられる開発体制が必要
    - ユーザニーズの分析から設計・実装まで、衆知を集める必要あり
    - 開発スピードの加速
- **展望**
  - 高度な機能の追加によってユーザの増加が期待できる
    - 大規模DB：IoT, Analitics, Bio-informatics etc.
    - セキュア化：金融、ヘルスケア・医療、行政 etc.
  - コミュニティの活性化
    - 大きなメリットと大きな投資/リスク：関係者の合意・納得が必要
      ⇒コミュニティ活動(イベント・ML)の重要度が増す

Copyright©2018 NTT Corp. All Rights Reserved.

活発化する国際カンファレンス

- **2008年からは”PGCon”がオタワで継続開催**
  - スタート当初は唯一の国際カンファレンスだった
- **現在は欧米亜で複数の定期カンファレンスが開催**
  - 開発者向け・ユーザ向けともに盛況

PostgreSQL関連の主要会議



Copyright©2018 NTT Corp. All Rights Reserved.

おわりに

- **PostgreSQLの発展は企業システムのニーズにマッチする方向で進んできた**
  - より速く、より大きく、より強く
  - 現在ではPostgreSQLは企業システム向けDBの有力な選択肢
- **PostgreSQLの開発は”エコシステム”が担う**
  - 大きな機能は複数企業がゆるやかに連携して手掛けている
  - 開発はもちろん、利用することもエコシステムへの参加
- **インターネットの発展はPostgreSQLの進化を導く**
  - データの大規模化対応とDBのセキュリティ強化が必要
  - 難易度の高い開発に向けてコミュニティでの議論が活性化

Copyright©2018 NTT Corp. All Rights Reserved.

ご清聴ありがとうございました

Copyright©2018 NTT Corp. All Rights Reserved.

謝辞・参考文献

- **謝辞**
  - 本稿に対してコメントを頂き、PostgreSQLの検証データ等を作成してくれたNTT OSSセンタのメンバ各位に謝意を表する。
  - P.18, 26, 27, 30のサーバのイラストはhttps://openclipart.org/ から借用した。ライセンスは Creative Commons CC0 1.0 である。
- **P.11の参考資料**

1. 朝倉, 山田, “料金系基幹システムへのPostgreSQL導入事例”, PGECons 2015年事例セミナー資料
   https://www.pgecons.org/wp-content/uploads/2015/09/PGECons3_NTT5_20150911.pdf
2. 石井, “電力自由化を陰で支えるPostgreSQL”, PGConf ASIA 2016資料
   http://www.pgconf.asia/JP/wp-content/uploads/2016/12/1023b0605223d2bdc2ba777d346fd504.pdf
3. Blomqvist, V., “Using PostgreSQL in Tantan”, PGConf ASIA 2016資料
   http://www.pgconf.asia/JP/wp-content/uploads/2016/12/From-0-to-350bn-rows-in-2-years-PgConfAsia2016-v1.pdf
4. Bartunov, O., “PostgreSQL Universal Database”, PGConf ASIA 2017資料
   http://www.pgconf.asia/JA/2017/wp-content/uploads/sites/2/2017/12/D1-K1.pdf

Copyright©2018 NTT Corp. All Rights Reserved.

注釈

- P.4 NTT OSSセンタは2006年4月発足
- P.8 Bruce Momjian氏はPostgreSQLコアチームの一人。この発言は坂田の記憶
- P.10 この表はP.5の「黎明期」に対応する
- P.12 PostgreSQL 8.2, 8.3の性能比較
- P.19 メール発信者のTom Lane氏はコアチームの一人
- P.21 OSSのエコシステムというときは、ここで説明する「関連する企業や人の集まり」という意味と、「連携して動作するソフトウェアの集まり」という意味とがある
- P.27 通常「シャーディング」という時は複数サーバに表データを水平分割して分散処理することを言う。ここではトランザクション管理をしない単なるシャーディングを指して用いた
- P.28 各OSS製品の提案時期は最初のOSS公開時を各製品のWebサイト等で確認した。ただしPostgres-RはOSS公開していない
- P.39 2018年のPGConf.EUは400名超が参加し、世界最大の規模だったとのこと

Copyright©2018 NTT Corp. All Rights Reserved.