PGECons
PostgreSQL Enterprise Consortium

大規模DBを見据えた PostgreSQLの性能検証

2013年度活動成果報告

PostgreSQLエンタープライズ・コンソーシアム
WG1（性能WG）

アジェンダ

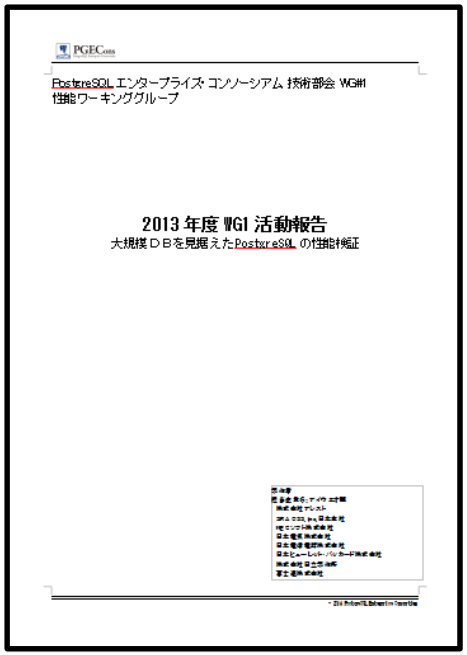
- WG1（性能ワーキンググループ）の今年度テーマ
- 今年度の成果物
- 実施体制
- 活動報告1: 定点観測（スケールアップ検証）
- 活動報告2:パーティショニング検証
- 活動報告3:ハードウェア活用（SSD）検証
- 活動報告4:スケールアウト検証（Postgres-XC）
- 2013年度活動をふりかえって
- 付録：検証環境について

WG1（性能ワーキンググループ）の今年度テーマ

- 定点観測（スケールアップ検証）
  - 多コアCPUでのPostgreSQL 9.3の到達点を検証
    - 9.2との比較、page checksumの性能影響
- パーティショニング検証
  - パーティションテーブルへの投入・検索・メンテナンス性能検証
- ハードウェア活用（SSD）検証
  - SSD採用時の性能向上を配置パターン別、アクセスプラン別に検証
- スケールアウト検証（Postgres-XC）
  - Postgres-XCのスケールアウト特性を検証

今年度の成果物

- 各テーマごとに実施した検証手順および結果を文書化
- 検証環境のハードソフト構成、検証結果データを公開
- 「2013年度WG1活動報告」として一冊にまとめた冊子を2014年4月に公開



実施体制

- 参加企業（企業名順）
  - 株式会社アイ・アイ・エム
  - 株式会社アシスト
  - SRA OSS, Inc.日本支社
  - NECソリューションイノベータ株式会社
  - 日本電気株式会社
  - 日本電信電話株式会社
  - 日本ヒューレット・パッカード株式会社
  - 株式会社日立製作所
  - 富士通株式会社（主査）

活動報告1

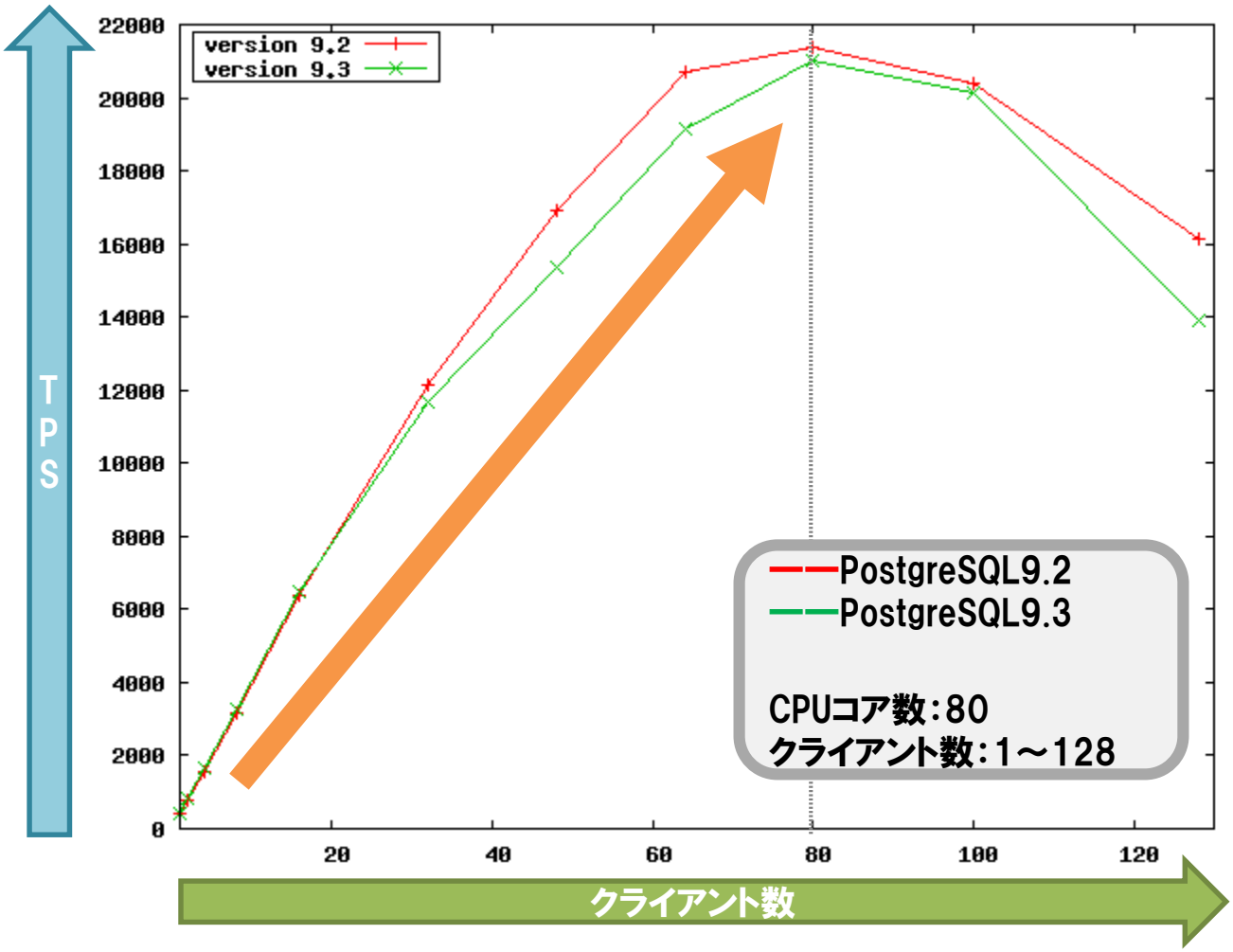
定点観測(スケールアップ検証)

定点観測(スケールアップ検証)概要

- pgbenchで計測
  - スケールファクタ1000で初期化(DBサイズは15GB)。
  - ランダムに10000行取得するカスタムスクリプトを300秒ずつ実行。pgbench-Sでは十分な負荷がかからないため。
  - 3回の計測結果の中央値を結果とした。
- 計測内容は以下3項目

9.2と9.3の参照性能の比較
9.3のpagechecksum機能有無での参照性能の比較
9.3のCPUコア数によるスケールアウト

9.2 と 9.3の参照性能の比較



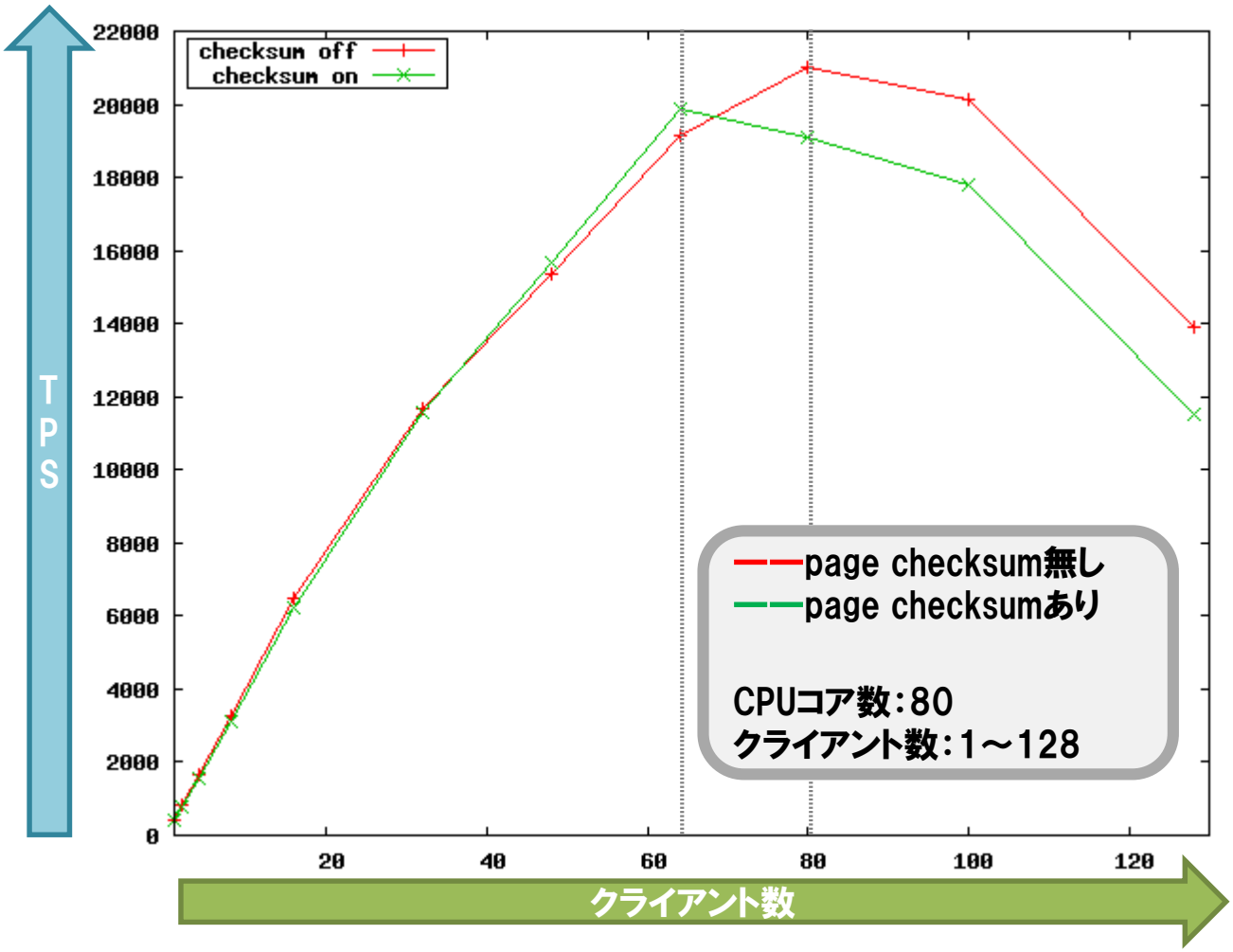
コア数と同じ80クライアントが最大TPSとなり、同傾向。

クライアント数32以降で、9.3は9.2より最大14%低い。

9.3の方が CPUidleが高い。

昨年度9.2 用にカスタマイズしたスクリプトだったので、9.3では十分な負荷がかからなかった？

9.3 page checksum有無での参照性能の比較(1)

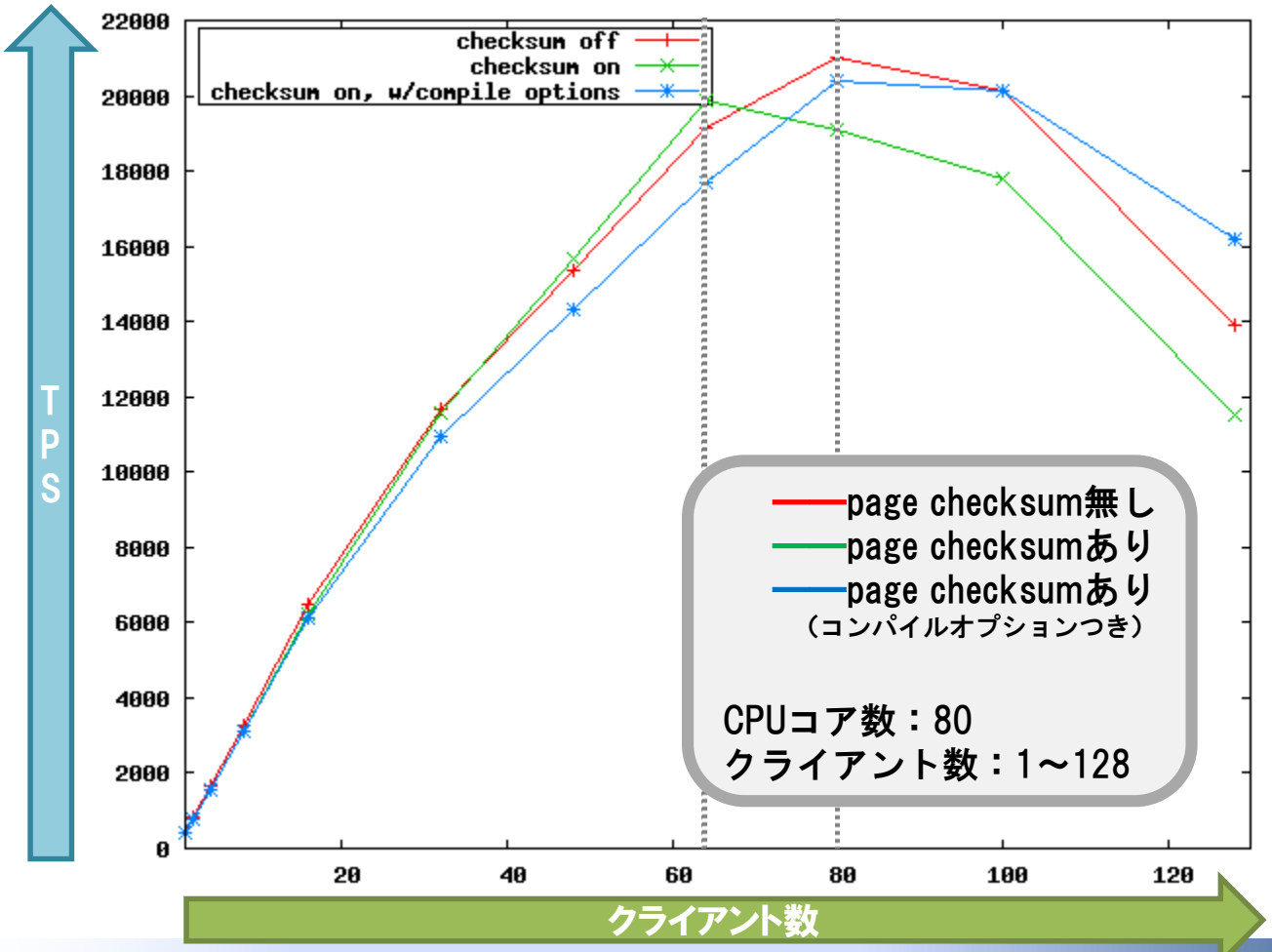


pagechecksumのオーバヘッドはおもに計算処理なので、クライアント数が多いと影響が大きい。

クライアント数が少なくCPU利用率が低い場合には、その影響はほとんど無いと言える。

9.3 page checksum有無での参照性能の比較(2)

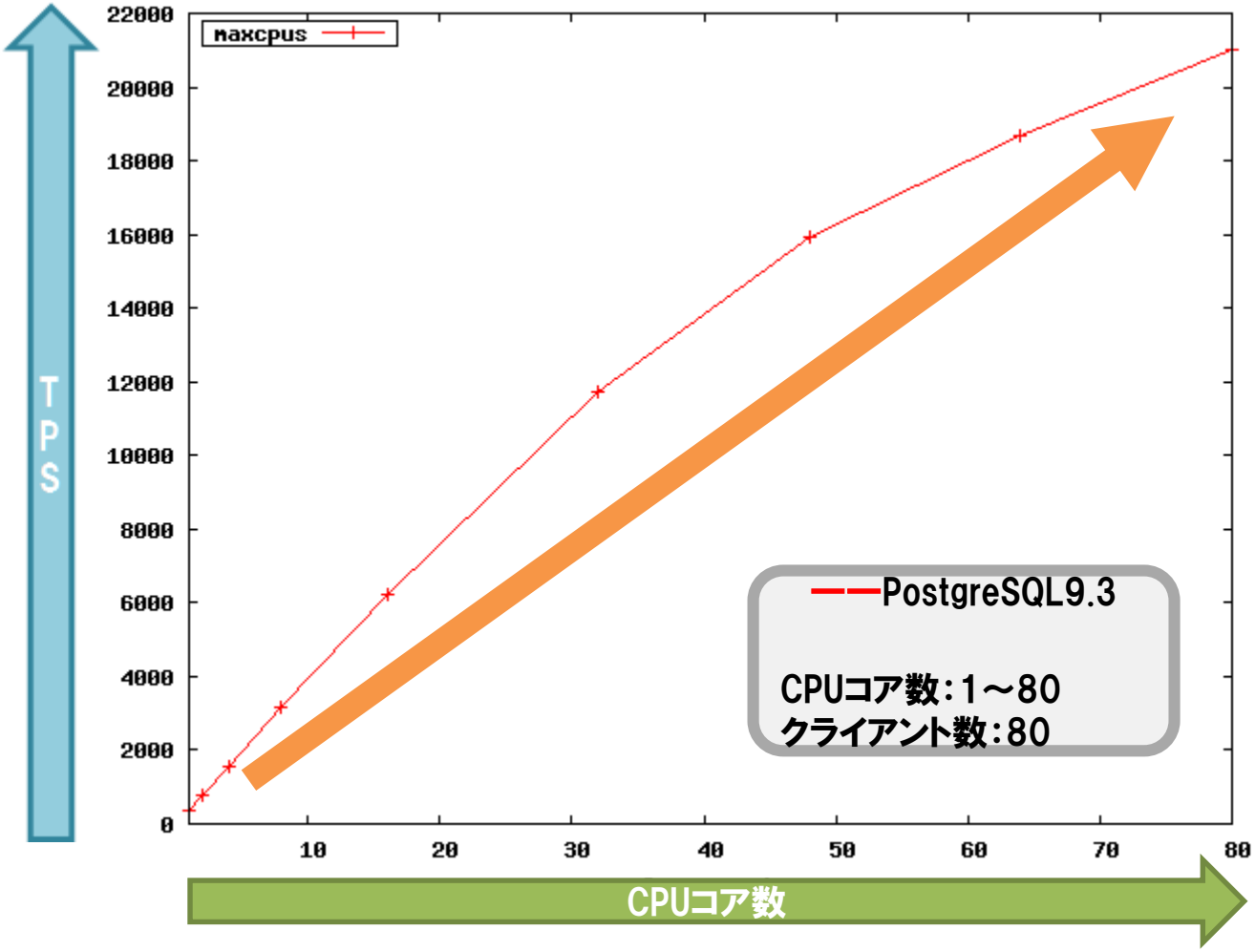
MakefileのCFLAGSに「-msse4.1 -funroll-loops -ftree-vectorize」をつけてインストールしたPostgreSQL9.3



同時接続クライアント数が多いときには、pagechecksum無しより、有りの方がよいTPSに。

コンパイルオプションの変更によって、pagechecksum処理以外のところにも良い影響が出た？

コア数変動での同時実行参照性能



同時接続クライアント数を80に固定し、コア数を変化させた。

TPSは、ほぼコア数に比例して向上し、PostgreSQLのスケールアウト性能が良好と言える。

活動報告2

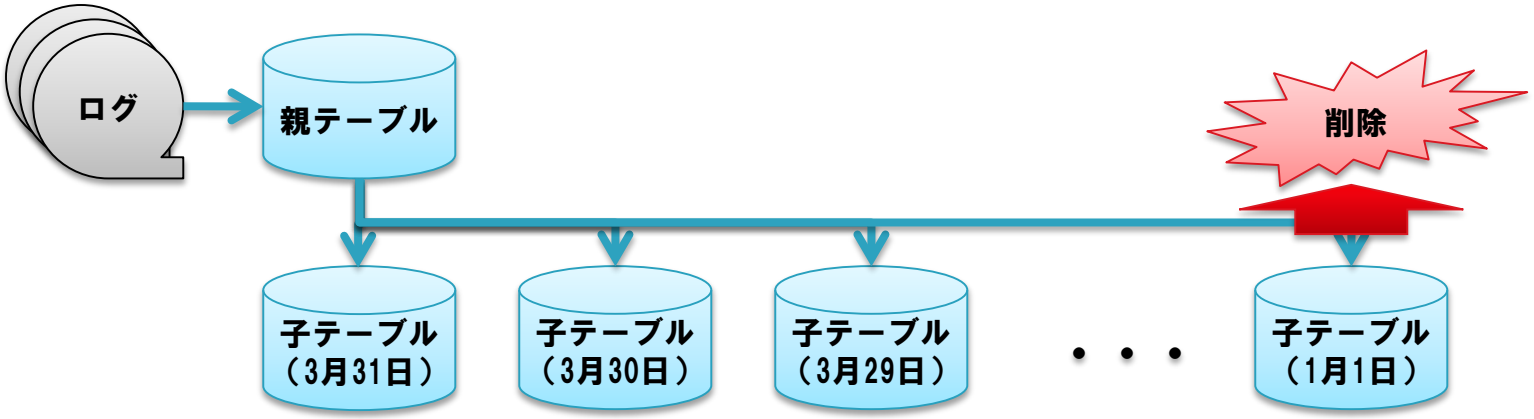
パーティショニング検証

検証概要

- 検証目的
  - 処理データ量の増大に対する対応手段としてのパーティショニングの効果
  - パーティショニングの更新処理方式による性能差検証
- 検証環境
  - DBサーバのサーバ環境（測定用クライアントはDBサーバ上で実行）

CPU	インテルXeon プロセッサー E5-2690(2.90GHz , 8C/16T, 20MB)*2
メモリ	128GB
OS	Redhat Enterprise Linux 6

- 検証シナリオ
  - アクセスログの格納先としてPostgreSQLのパーティションテーブルを使用
  - 1日分のデータに対する集計検索を実行
  - 一定期間（1,3,6ヶ月）データを保存し、日次で1日分のデータを削除

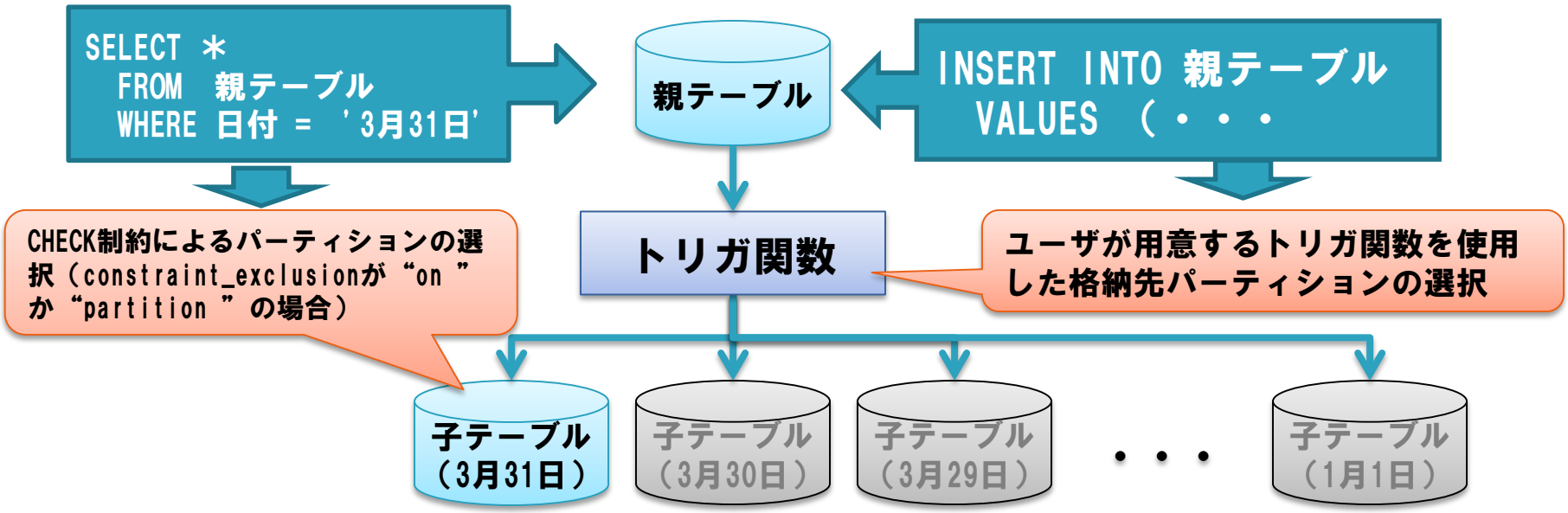


検証項目・方法・結果

検証項目	挿入性能	検索性能	運用性能
検証方法	トリガ関数の実装別性能 ■ 静的関数(PL/pgSQL) ■ 動的関数(PL/pgSQL) ■ C言語関数 ■ 各パーティションへの直接挿入(トリガ関数を使用しない)	1日分のデータを対象とした集計処理 ■ SeqScan ■ Index_Scan ■ Bitmap_Scan	1日分のデータ削除およびVACUUM ■ パーティションを指定したTRUNCATE ■ 条件指定のDELETE(非パーティション表) ■ VACUUM(非パーティション表)
比較対象	■ トリガ関数の実装別処理時間を比較 ■ トリガ関数／パーティション表への直接挿入比較	■ パーティション表／非パーティション表検索比較	■ パーティション表／非パーティション表性能比較
検証結果	■ 高速な順に、1.直接挿入、2.C言語関数、3.動的関数、4.静的関数。 ■ 静的関数は圧倒的に遅く、測定を一部断念	■ いずれの検索プランでも、パーティション表の応答時間が4～6倍高速 ■ 今回の検証シナリオでは、パーティション数が180以上の場合でもパーティショニングの利用を推奨	■ DELETEの応答時間300秒に対し、TRUNCATEは0.02秒と圧倒的にパーティション表が高速 ■ VACUUMの必要性も含め、検証シナリオの運用ではパーティションを推奨

PostgreSQLのパーティショニング

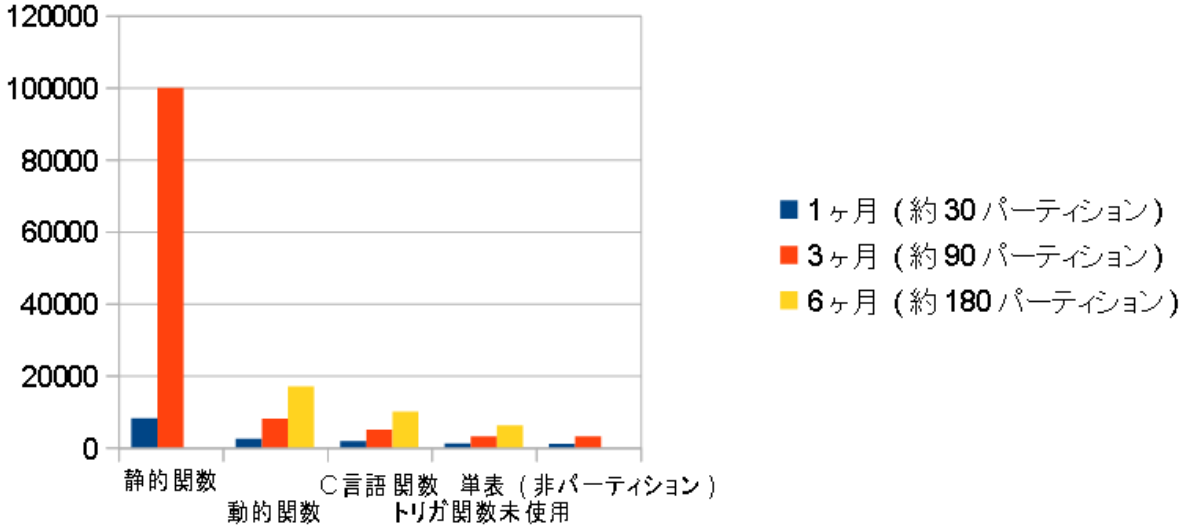
- テーブルの継承を利用したデータの分割格納
- トリガ関数による格納先決定
- CHECK制約を利用した検索先パーティションの選択



PostgreSQLでは一般的に100パーティション以上は非推奨

データ挿入性能測定結果

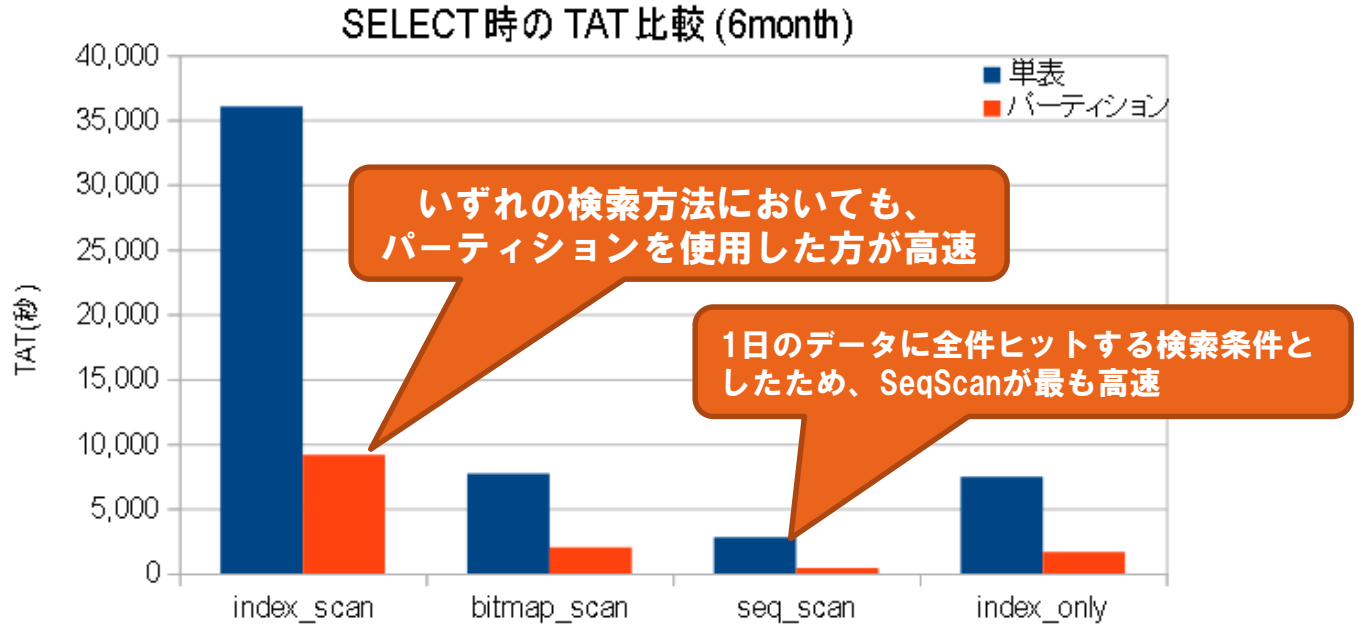
- 応答時間
  単表(※)＞C言語関数＞動的関数＞＞＞静的関数
  ※ トリガ関数を使用せず、パーティション表にレコードを直接挿入
- 静的関数のレコード挿入性能は非常に悪く、6ヶ月のデータ挿入は断念
- 動的関数はある程度良好な性能を記録し、実装が簡単でパーティション追加にも対応可能
- C言語関数はトリガ関数を使用した挿入の中では最も高速だが、実装は手間がかかり、実装ミスによるDBプロセス例外が発生する可能性もある



パーティションへの直接挿入やC言語関数は高速だが、実装は簡単ではない
データ量や要求性能に応じて方式を選択

検索性能測定結果

- 6ヶ月（約180パーティション）分のデータを格納したテーブルに対する検索性能
- 1日分のデータに対する集計では、いずれの検索プランでもパーティション表の方が高速
- 今回の検証シナリオでは1日分のデータサイズがメモリ容量よりはるかに大きく、必ずI/Oが発生する重い検索だったため、プラン生成によるオーバヘッドが問題にならなかった



検索内容次第で100パーティション以上でも検索性能面で有利な場合がある

運用性能測定(データ削除・VACUUM)

- パーティション表に対して削除処理によるデッドタプルVACUUMは不要
- TRUNCATEと検索条件付きのDELETEでは、処理時間および負荷ともに圧倒的にTRUNCATEの方が軽い

	TRUNCATE	VACUUM	ANALYZE
パーティション表 3ヶ月	0.02秒	--	--
パーティション表 6ヶ月	0.02秒	--	--
非パーティション表 6ヶ月	300秒 (DELETE)	3420秒	197秒

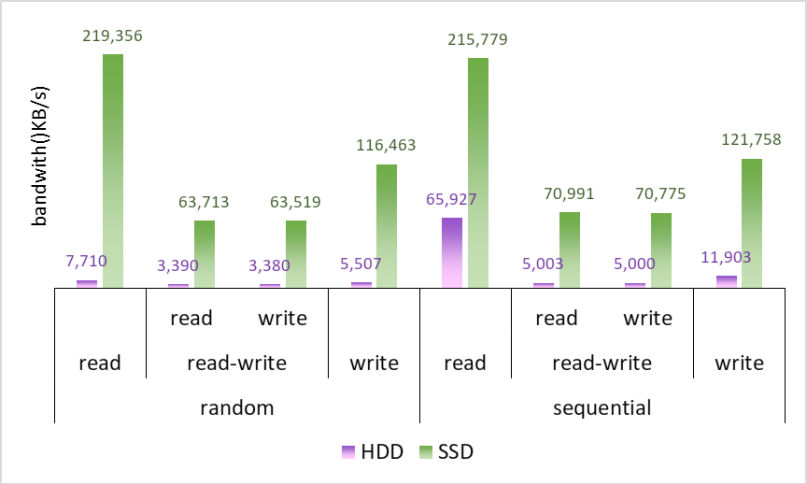
大容量データの運用(削除/VACUUM)にはパーティション表が有利

活動報告3

ハードウェア活用(SSD)検証

検証概要

- 検証目的
  - PostgreSQLのデータディレクトリ配下の資源をSSDに配置した場合の性能向上を検証
- 検証構成
  - PCIe SSD搭載マシン（HDDはディスクアレイを利用）
  - SATA SSD搭載マシン（HDDは内蔵HDDを利用）



ディスクI/O性能（PCIe SSD搭載マシン）

ディスクI/O性能（SATA SSD搭載マシン）

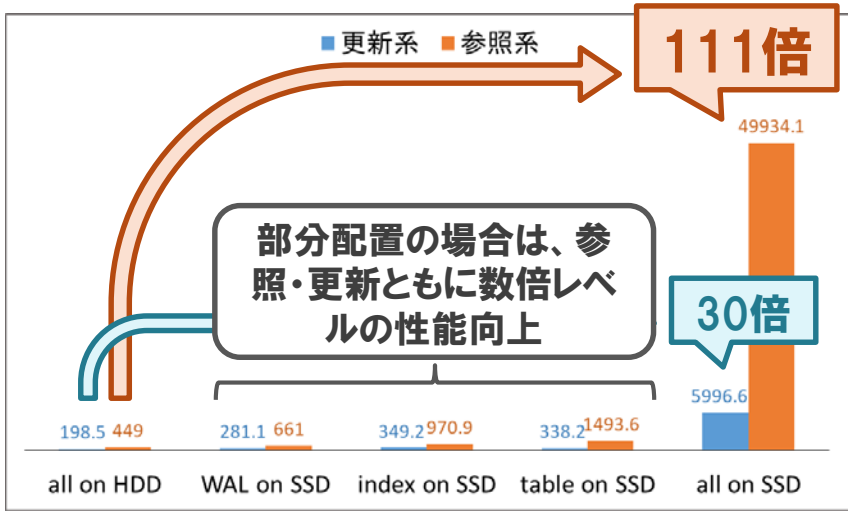
検証項目・方法・結果

検証項目	SSD配置パターン別インデックススキャン性能検証	インデックスオンリースキャン性能検証
検証方法	■ pgbenchのデフォルトスクリプト（TPC-Bライク）でTPSを測定 • 更新系と参照系 ■ PCIe SSD搭載マシン、SATA SSD搭載マシンそれぞれで測定	■ pgbenchのカスタムスクリプトでTPSを測定 • 完全一致の1件SELECT ■ UPDATE直後のインデックスオンリースキャンの性能も検証 • インデックスオンリーが失敗する場合 ■ PCIe SSD搭載マシン、SATA SSD搭載マシンそれぞれで測定
比較対象	■ データディレクトリ配下の資源の配置パターン毎に比較 • すべてHDD • データのみSSD • インデックスのみSSD • WALのみSSD • すべてSSD	■ データディレクトリ配下の資源の配置パターン毎に比較 • すべてHDD • すべてSSD
検証結果	■ データディレクトリ配下のすべての資源をSSDに配置した場合の性能が、他のパターンと比較して圧倒的に優位 ■ 最大で更新系30倍、参照系111倍	■ 資源をすべてSSDに配置することで最大119倍の性能向上を確認

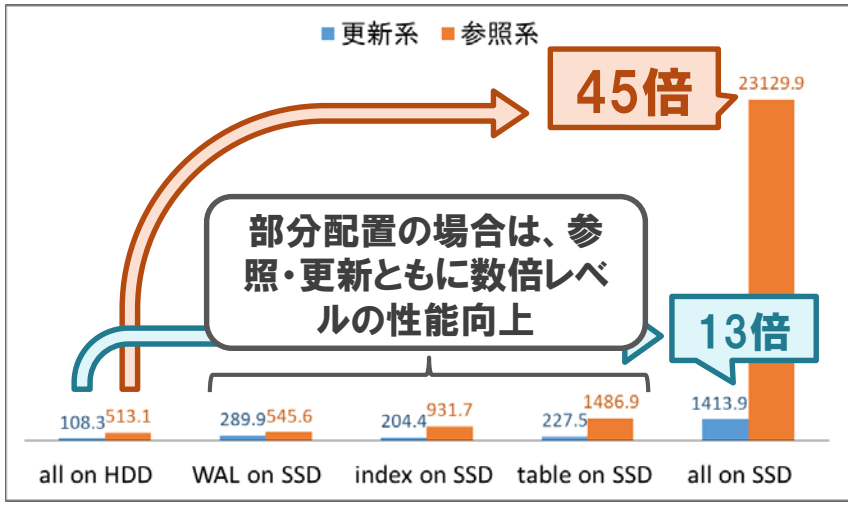
SSD配置パターン別インデックススキャン性能検証

■PCIe SSD搭載マシンの測定条件
- 共有メモリ:48GB
- pgbench設定:
  SF=26000, 実行時間=600秒

■SATA SSD搭載マシンの測定条件
- 共有メモリ:12GB
- pgbench設定:
  SF=6400, 実行時間=600秒



SSD配置パターン別インデックススキャン性能(PCIe SSD搭載マシン)

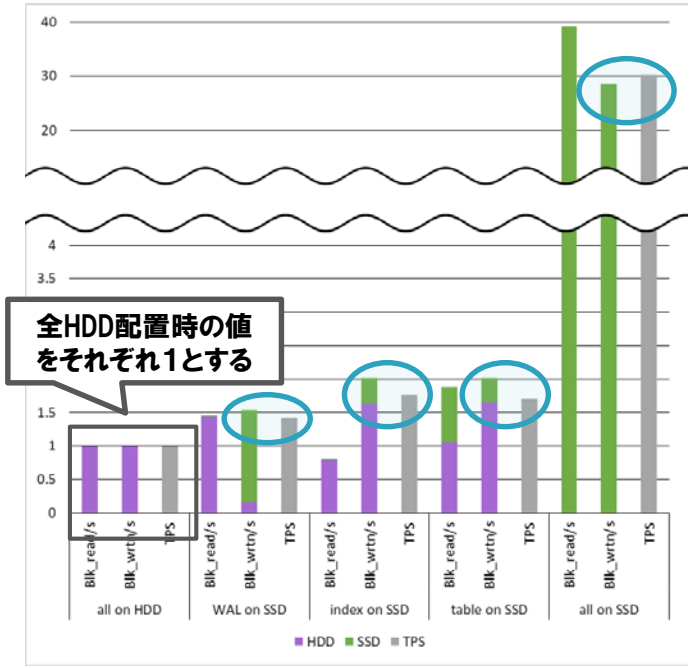


SSD配置パターン別インデックススキャン性能(SATA SSD搭載マシン)

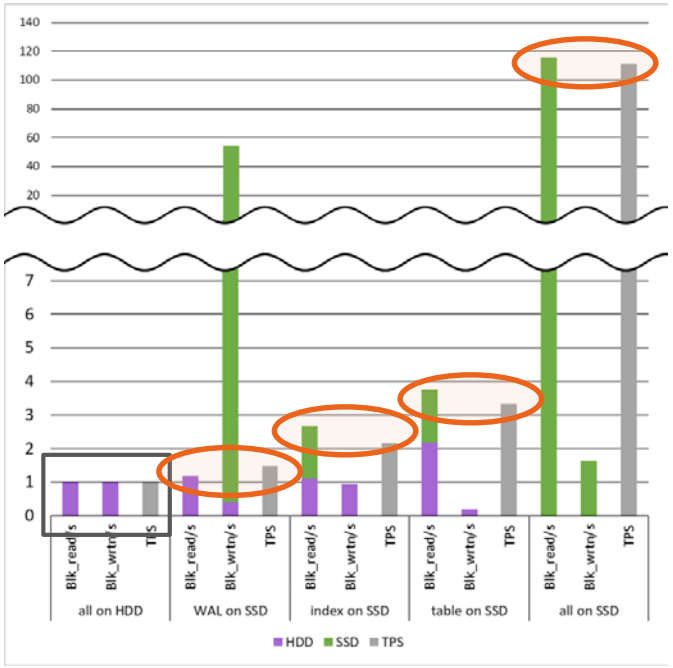
ディスクI/O性能がデータベース性能に強く影響していると推察されるため、すべてSSD配置の場合は大幅に性能改善。
ただし、部分配置での性能改善は限定的。
- データ、インデックス、WALそれぞれのディスクI/Oがデータベース性能に影響

SSD配置パターン別インデックススキャン性能検証

検証中の読み込み速度・書き込み速度・DB性能(TPS)を配置パターン間で比較



更新系のI/O性能とTPSの全HDD配置対比(PCIe SSD搭載マシン)



参照系のI/O性能とTPSの全HDD配置対比(PCIe SSD搭載マシン)

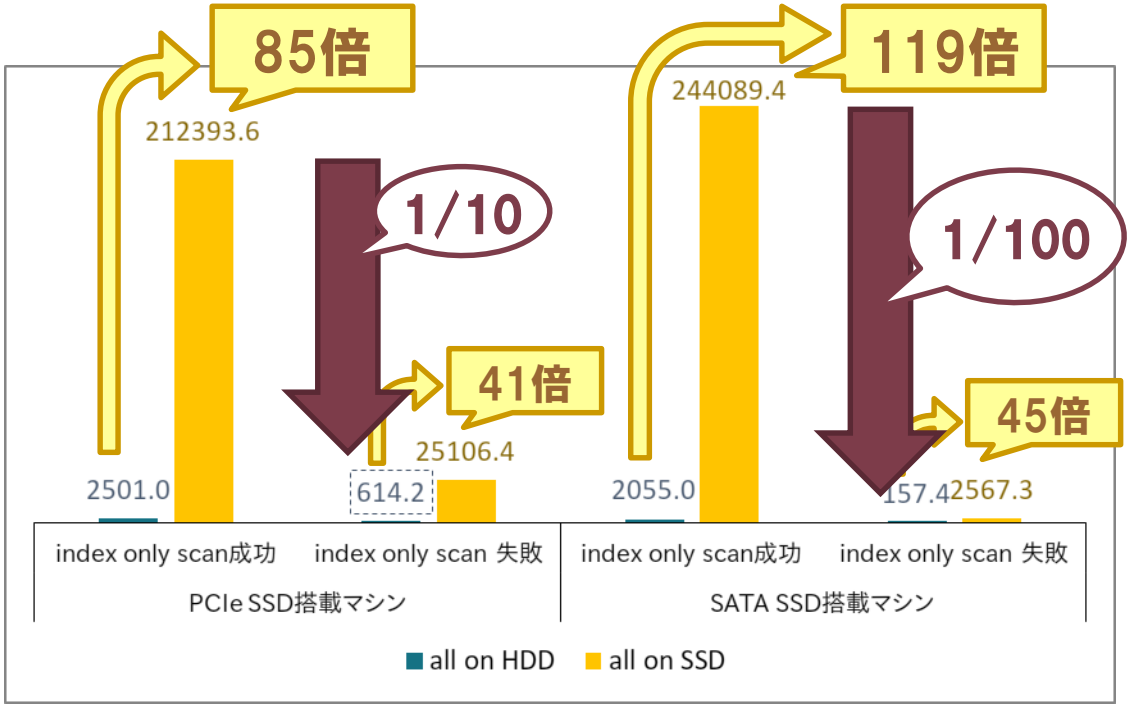
DB性能(TPS)が書き込み性能に依存
(更新系)

DB性能(TPS)が読み込み性能に依存
(参照系)

I/O性能がデータベース性能に強く影響していることが確認できた

- 本データ・トランザクションモデルでは、資源をすべてSSDに配置することで大きく性能改善

インデックスオンリースキャン性能検証



インデックスオンリースキャン成功／失敗時の性能検証

□測定方法・条件

- 共有メモリ、pgbench設定はインデックススキャンと同様
- スクリプトは独自

```
SELECT aid FROM pgbench_accounts
WHERE aid = :aid;
```

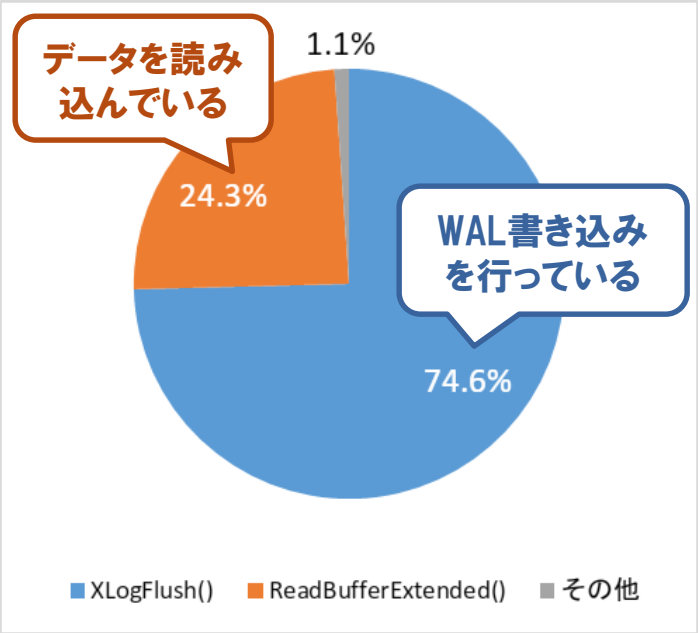
- 通常のインデックスオンリースキャン（成功）とUPDATE直後のインデックスオンリースキャン（失敗）の2パターンで計測

すべてHDDに配置した場合：ディスクI/Oがボトルネック
すべてSSDに配置した場合：I/Oネックが解消、CPU性能が左右

なぜインデックスオンリースキャン失敗時にデータベース性能が大幅に落ち込むのか？

インデックスオンリースキャン性能検証

なぜインデックスオンリースキャン失敗時にデータベース性能が大幅に落ち込むのか？



インデックスオンリースキャン性能検証時のスタックトレース
（SATA SSD搭載マシン）

当初の予想

データをヒープから取得する性能がネック

⇒ 性能の落ち込みは限定的

実際の結果

WALへの書き込みが発生
データ取得＋WAL書き込み性能がネック

⇒ 大幅に性能ダウン

活動報告4

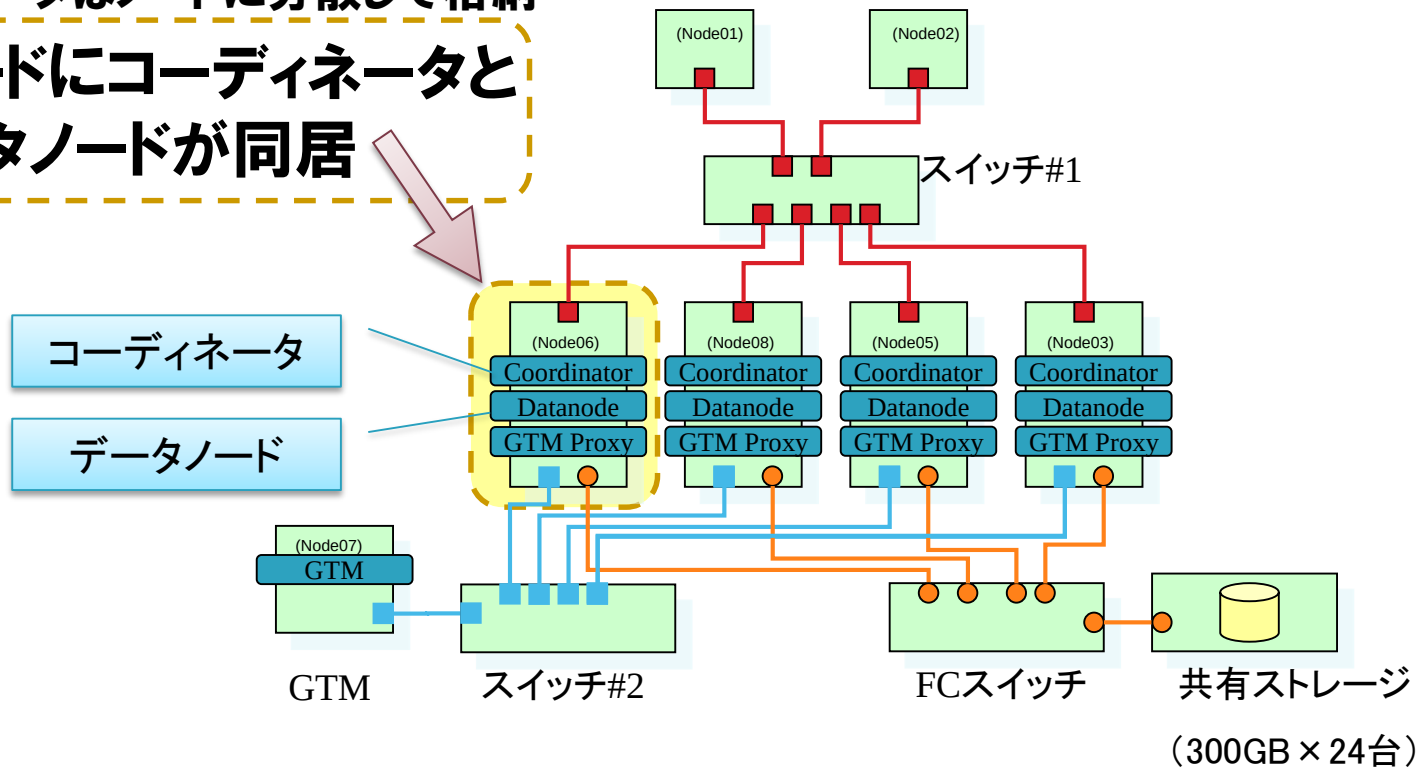
スケールアウト検証（Postgres-XC）

検証目的

- Postgres-XCのスケールアウト性を検証する
  - 異なる条件でのスケールアウト性を明らかにする
  - 実システムの要望に沿った2つの検証シナリオを設定
    - スループット要求が増大するシステムへの適用を想定
    - 取り扱うデータ量が増大するシステムへの適用を想定
  - スループット向上シナリオ
    - DBサイズを固定したまま、クラスタを構成するノード数を増やす
  - DBサイズ拡張シナリオ
    - DBサイズと比例してクラスタ数を増やす

検証構成

- Postgres-XCのコミュニティの推奨構成で測定
  - 2012年度の検証と基本的に同じ
  - 1ノードから最大4ノード構成まで測定
    - データはノードに分散して格納
  - 1ノードにコーディネータとデータノードが同居

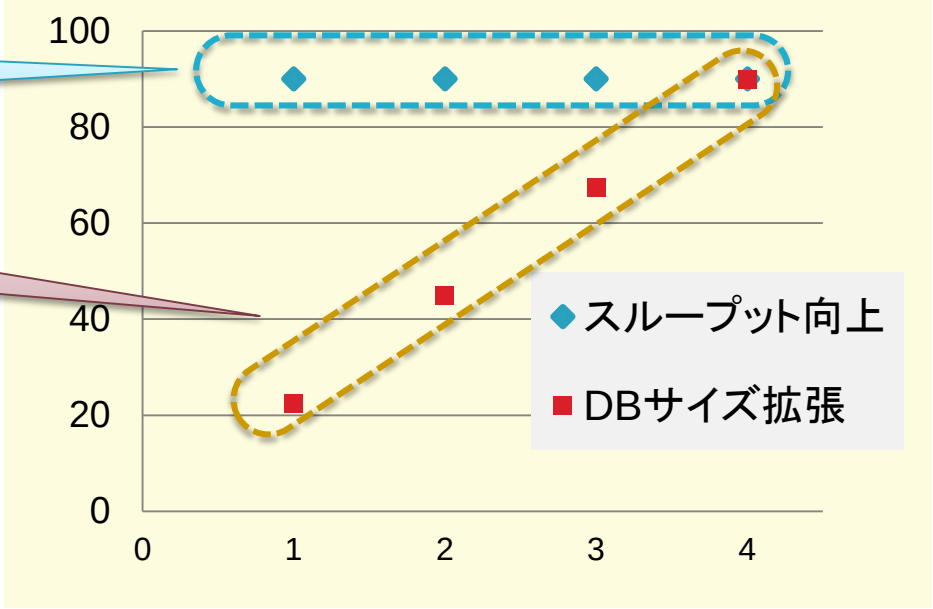


検証項目・方法・結果

		性能向上シナリオ	DBサイズ拡張シナリオ
検証方法（共通項目）		■ pgbenchを使用し、10分間の平均スループットを測定 ■ pgbenchの4つの表の全てを、各ノードに均等に分散して格納 ■ PostgreSQL 9.2系と性能を比較（相対値の基準）	
検証結果	参照	■ ノード数を増やすと性能が大きく向上 ■ キャッシュの効果が顕著	■ ノード数に比例して性能が向上
	更新	■ ノード数に比例して性能が向上	■ ノード数を増やしても性能は一定

スループット向上シナリオ
- DBサイズを固定して、ノード数を増やす

DBサイズ拡張シナリオ
- ノード数と比例してDBサイズを増やす



ノード数とDBサイズ[GB]

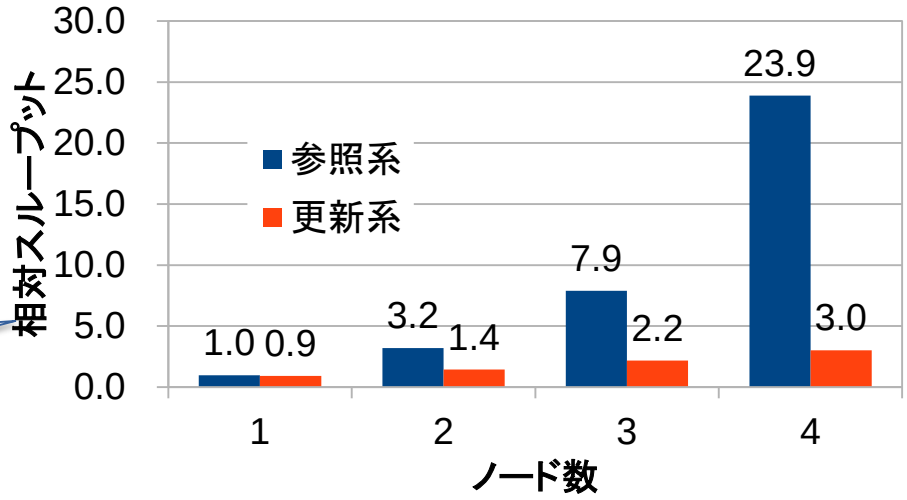
シナリオ	1	2	3	4
スループット向上	90	90	90	90
DBサイズ拡張	22.5	45	67.5	90

スループット向上シナリオ

- 参照系でノード数を1から4に増やしたとき
  - スループットが約24倍に向上
    - ストレージ上のデータがメモリに載るため、台数倍以上のスループット向上が見られた（キャッシュ効果）
- 更新系でノード数を1から4に増やしたとき
  - ノード数に比例して、スループットは約3倍に向上
    - 更新ではキャッシュ効果が見られない
    - ストレージがボトルネック（コミット時のディスクのフラッシュに起因）

相対スループット1の大きさ [TPS]

参照	910	更新	760

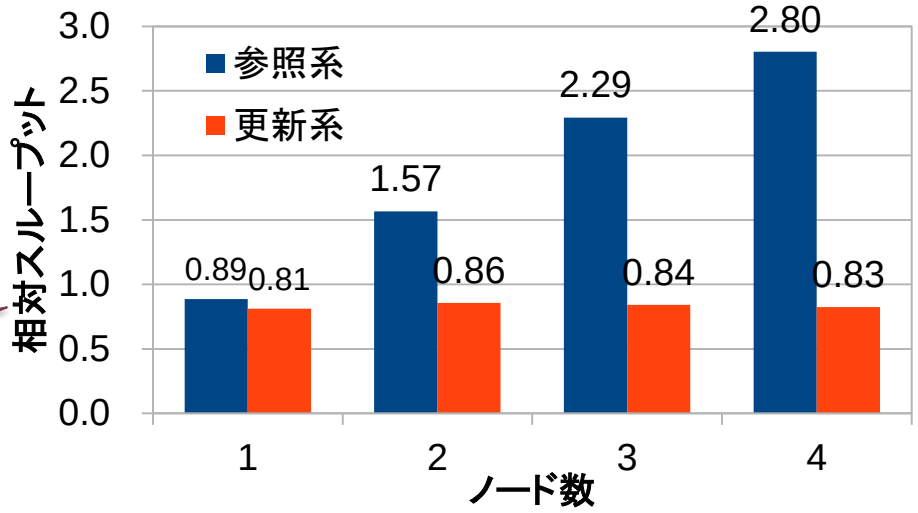


DBサイズ拡張シナリオ

- 参照系でノード数を1から4に増やしたとき
  - ノード数に比例して、スループットが約2.8倍に向上
    - メモリ量とDBサイズの比が一定のため、キャッシュ効果はない
- 更新系でノード数を1から4に増やしたとき
  - スループットはほぼ横ばい
    - 1トランザクションで複数ノードのデータを更新（pgbenchの特性）
    - 1回のコミット要求に参加するノード数は4ノード時に平均1.75
    - コミット処理はディスクのフラッシュがあり、重い

相対スループット1の大きさ [TPS]

参照	7753	更新	2777



活動報告のまとめ

2013年度の活動を振り返って

2013年度活動をふりかえって

- 昨年できなかったパーティショニング機能の検証をはじめ、今年度も多くの有意義な検証を行うことができました。
- さまざまなメディアで紹介されたとおり、競合することもある各企業のメンバーが、共通の目的のもとに活動し、交流を深めるということはとても意義のあることです。

2013年度活動をふりかえって

- WG検討会では報告書に載せきれない貴重な情報を数多く聞くことができました。来年度はぜひ一緒に活動しましょう！

PGECons

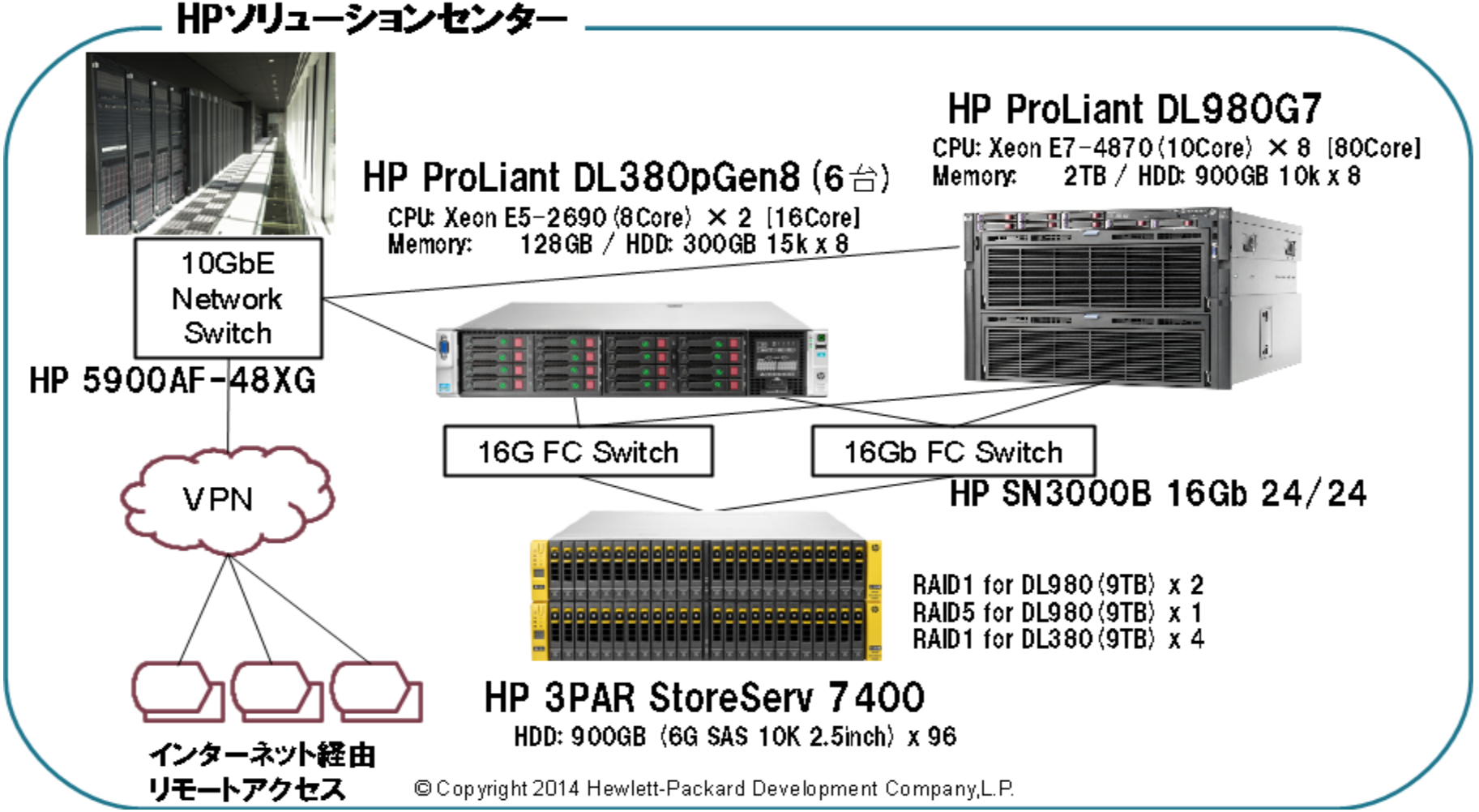
PostgreSQL Enterprise Consortium

付録：今回使用した検証環境について

- 定点観測（スケールアップ）およびパーティショニング検証では日本ヒューレット・パッカード株式会社様から検証環境をご提供いただきました
- ハードウェア活用（SSD）検証では富士通株式会社様から検証環境をご提供いただきました
- スケールアウト検証では日本電気株式会社様から検証環境をご提供いただきました
- 株式会社 アイ・アイ・エム様に「性能評価サービス」をご提供いただきました。
- PostgreSQLエンタープライズ・コンソシアムとして御礼を申し上げます

付録：検証環境１ （提供：日本ヒューレット・パッカード株式会社）

- 定点観測（スケールアップ）およびパーティショニング検証での使用機器／設備



付録：検証環境2 （提供：富士通株式会社）

■ ハードウェア活用（SSD）検証での使用機器／設備

富士通トラステッド・クラウド・スクエア

★PRIMERGY RX300 S7
PCIe SSD搭載マシンとして使用
CPU:インテルXeon E5-2690（2.90GHz）8コア×2
メモリ:192GB
内蔵HDD:900GB×2 RAID1
内蔵SSD:1.2TB×1（PCIe SSD）

★PRIMERGY RX300 S8
SATA SSD搭載マシンとして使用
CPU:インテルXeon E5-2697v2 （2.70GHz）12コア×2
メモリ:48GB
内蔵HDD:600GB×2 RAID1
内蔵SSD:200GB×2（SATA SSD） RAID1

★ETERNUS DX80 S2
ディスクアレイ装置として使用
ディスク容量:600GB×5 RAID5

検証ルーム

サーバルーム

最新の当社サーバ、ストレージ機器を約300台完備し、事前導入検証やベンチマーク、ICTシステム検証が可能です。

詳細情報は以下をご確認ください。
http://jp.fujitsu.com/facilities/tcs/

付録：検証環境3 （提供：日本電気株式会社）

■ スケールアウト検証での使用機器／設備

NECプラットフォームイノベーションセンター

- Server
  - Express5800/B120d （Blade）6台
    CPU：Xeonプロセッサー E5-2470 (2.30GHz 8Core 20M) ×2
    メモリ：32GB
    HDD：300GB (2.5型SAS 15,000rpm) ×2 RAID1
  - Express5800/R120d-2M （ラックサーバ）1台
    CPU：Xeonプロセッサー E5-2690 (2.90GHz, 8Core, 20MB) ×2
    メモリ：32GB
    HDD：2.5 型 146.5GB(6Gb/s SAS, 15,000 rpm)×6 RAID1

- Storage （データベース格納領域）
  - iStorage/M300 1台
    - コントローラ数： 2台 （FCコントローラ）
    - キャッシュメモリ： 16GB
    - FCポート数： 8個 （8G対応 4ポート／1コントローラ）
    - ディスク：SAS 3.5型 15krpm 600GB ×24玉
      - RAID10 (6台) × 4