groonga 開発予報

有限会社 未来検索ブラジル
矢田 晋

祝

本日リリースされた

groonga 1.2.8 には

grn_dat が含まれています

そこで今日は

皆さんにちょっと

ダブル配列の話を

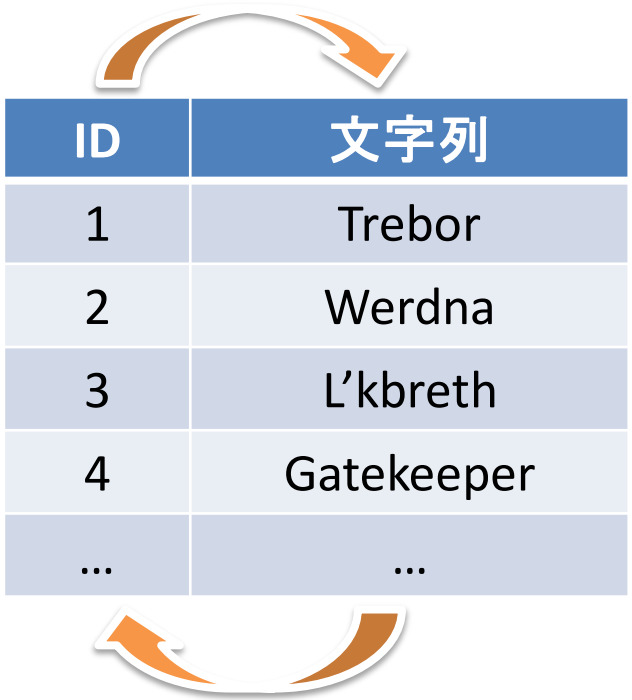
聞いてもらいます

とりあえず

grn_dat とは何か

grn_dat とは

- 文字列を ID と関連付けるモジュール
  - grn_pat, grn_hash の仲間



ID	文字列
1	Trebor
2	Werdna
3	L’kbreth
4	Gatekeeper
…	…

dat[1] == “Trebor”

dat[2] == “Werdna”

dat[“L’kbreth”] == 3

dat[“Gatekeeper”] == 4

grn_dat と仲間たち

- grn_pat – パトリシアトライ
  - 前方一致検索をサポートする
- grn_hash – ハッシュ表
  - 前方一致検索をサポートしない代わりに高速
- grn_dat – ダブル配列 

  - 前方一致検索をサポートする上に高速

いずれも

参照ロックフリー

grn_dat の特徴

- 前方一致検索と参照時間を重視
  - 文字列更新については後述

	grn_pat	grn_hash	grn_dat
前方一致検索	○	×	○
参照	○	◎	◎
更新	○	◎	△
サイズ	◎	○	△
文字列更新	×	×	○

前方一致検索とは

- Common Prefix Search
  - クエリの前半に一致する文字列を見つける
    “北海道” ⇒ “北”, “北海”, “北海道”
  - 用途: クエリから索引語への分割

- Predictive Search
  - クエリで始まる文字列を見つける
    “南斗” ⇒ “南斗孤鷲拳（シン）”, “南斗水鳥拳（レイ）”, etc.
  - 用途: クエリの補完・拡張

文字列更新とは

- ID を残して文字列のみを更新すること
  - 用途: テーブル情報の管理

ID	文字列
1	Trebor
2	Werdna
3	L’kbreth
4	Gatekeeper
…	

ID	文字列
1	Trebor
2	Werdna
3	L’kbreth
4	Sorn
	…

Update(“Gatekeeper”, “Sorn”)

grn_dat の役割

- grn_pat の代替として
  - 前方一致検索が必要なとき
  - 更新より参照の方が多いとき
  - メモリ使用量より参照時間を重視するとき

- テーブル情報の管理に使うと
  - テーブルやカラムの名前変更が可能になる
  - MySQL で ALTER TABLE RENAME が可能になる

技術的な情報がほしい方へ

- grn_dat 開発のポイント
  - ダブル配列の参照ロックフリー化
  - 更新の効率化
  - 前方一致検索の効率化

- 参考資料
  - 参照ロックフリーなダブル配列
    - http://groonga.org/ja/blog/2011/11/08/grn_dat.html

そろそろ説明は終わりにして

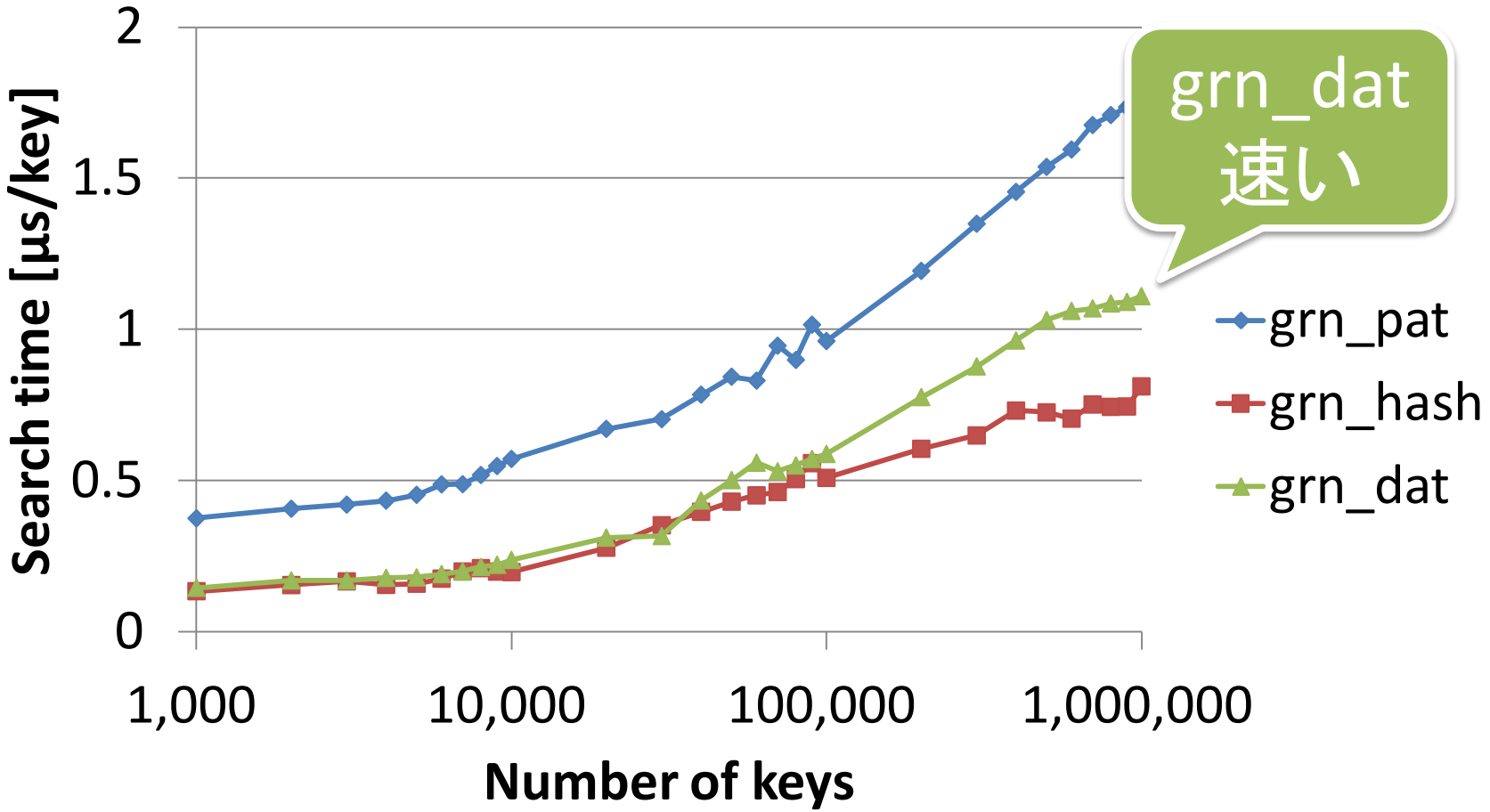
見せてもらおうか

新しいモジュールの性能とやらを

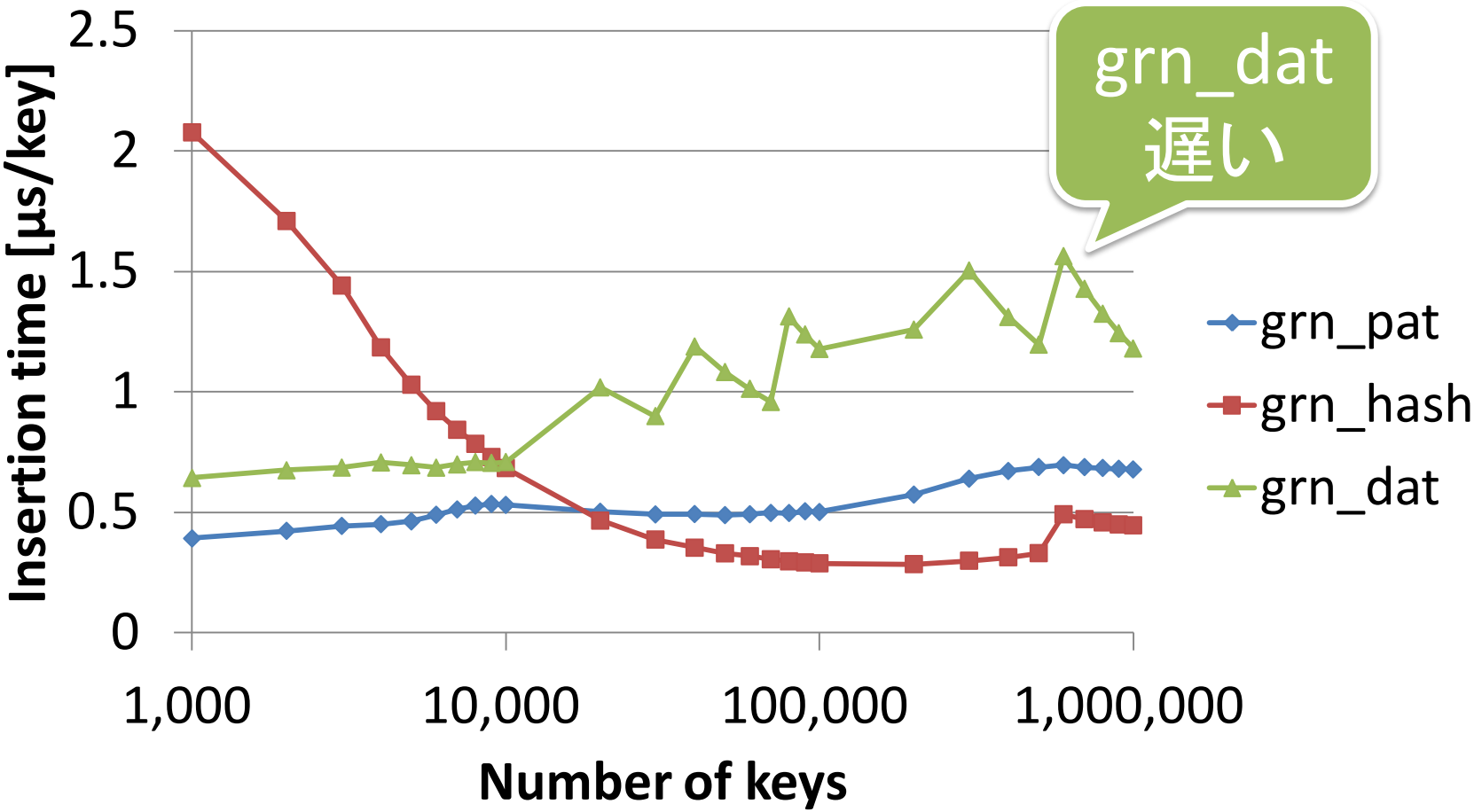
ベンチマーク(準備)

- データ
  - jawiki-20111111-all-titles-in-ns0
  - 先頭の 100 万件を使用
- 構築・参照方法
  - ランダム順に登録
  - ランダム順に登録文字列を参照
- 計測方法
  - 試行回数 11 で中央値を採用

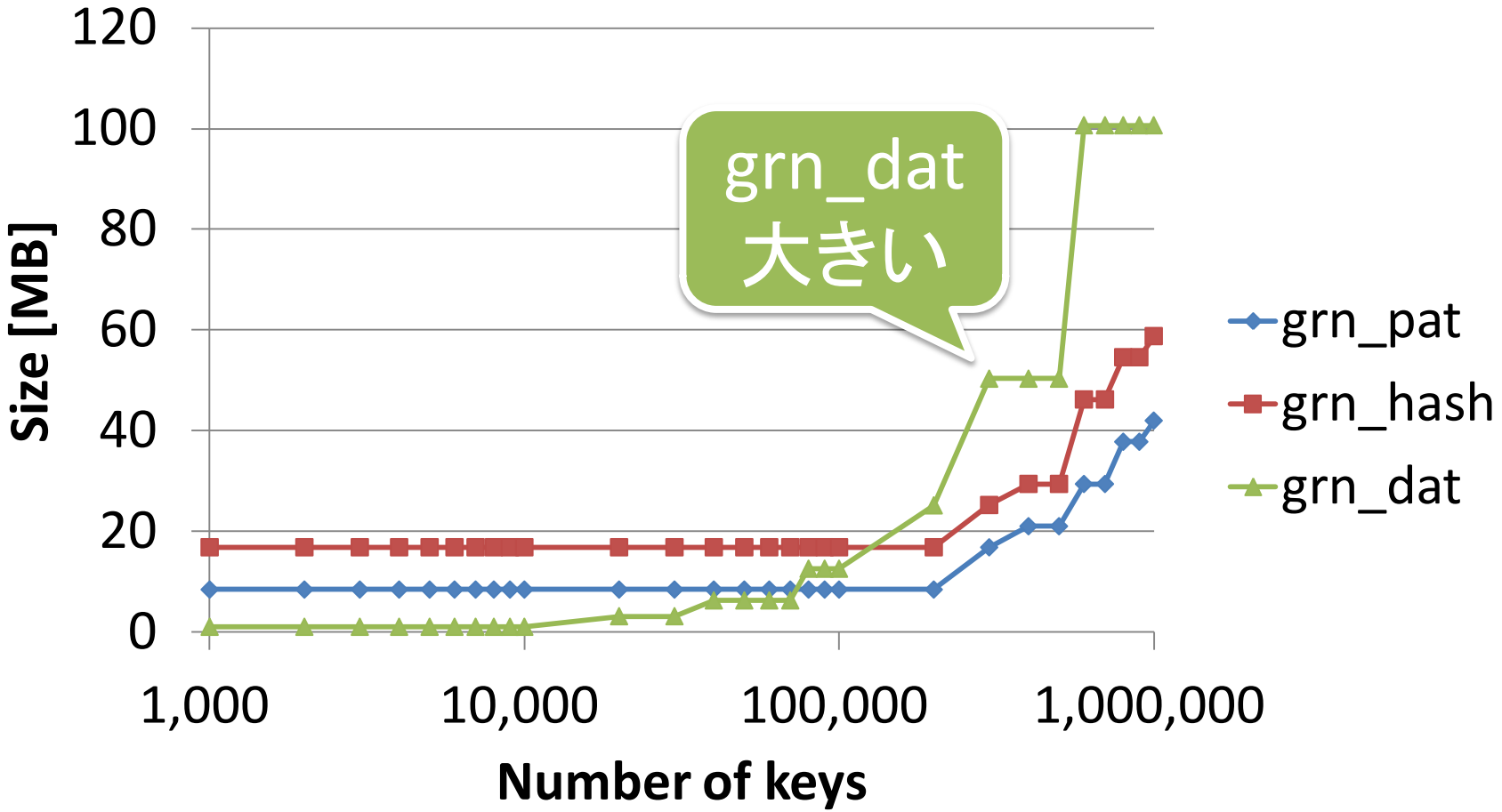
ベンチマーク（参照時間）


grn_dat
速い
Search time [µs/key]
2
1.5
1
0.5
0
1,000
10,000
100,000
1,000,000
Number of keys
grn_pat
grn_hash
grn_dat

ベンチマーク（構築時間）


grn_dat
遅い
Insertion time [µs/key]
2.5
2
1.5
1
0.5
0
1,000
10,000
100,000
1,000,000
Number of keys
grn_pat
grn_hash
grn_dat

ベンチマーク（サイズ）


120
100
80
60
40
20
0
Size [MB]
1,000
10,000
100,000
1,000,000
Number of keys
grn_dat
大きい
grn_pat
grn_hash
grn_dat

まとめると

前方一致検索ができて

参照時間に優れる

そういえば

「groonga 開発予報」

というタイトルでした

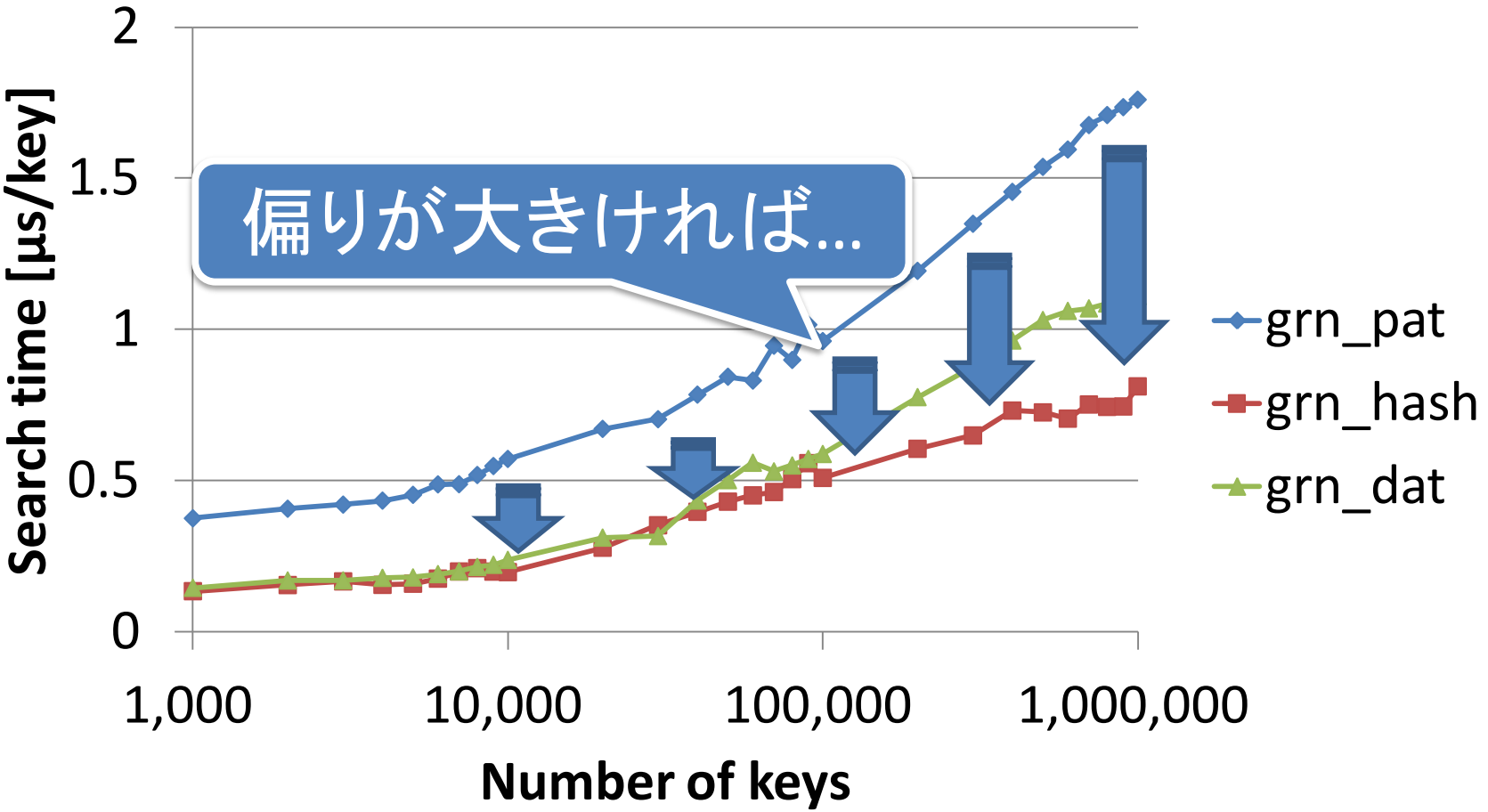
検討中の内容を紹介します

ひとつ

頻出する索引語をキャッシュ

http://groonga.org/ja/blog/2011/07/13/lexicon-cache.html

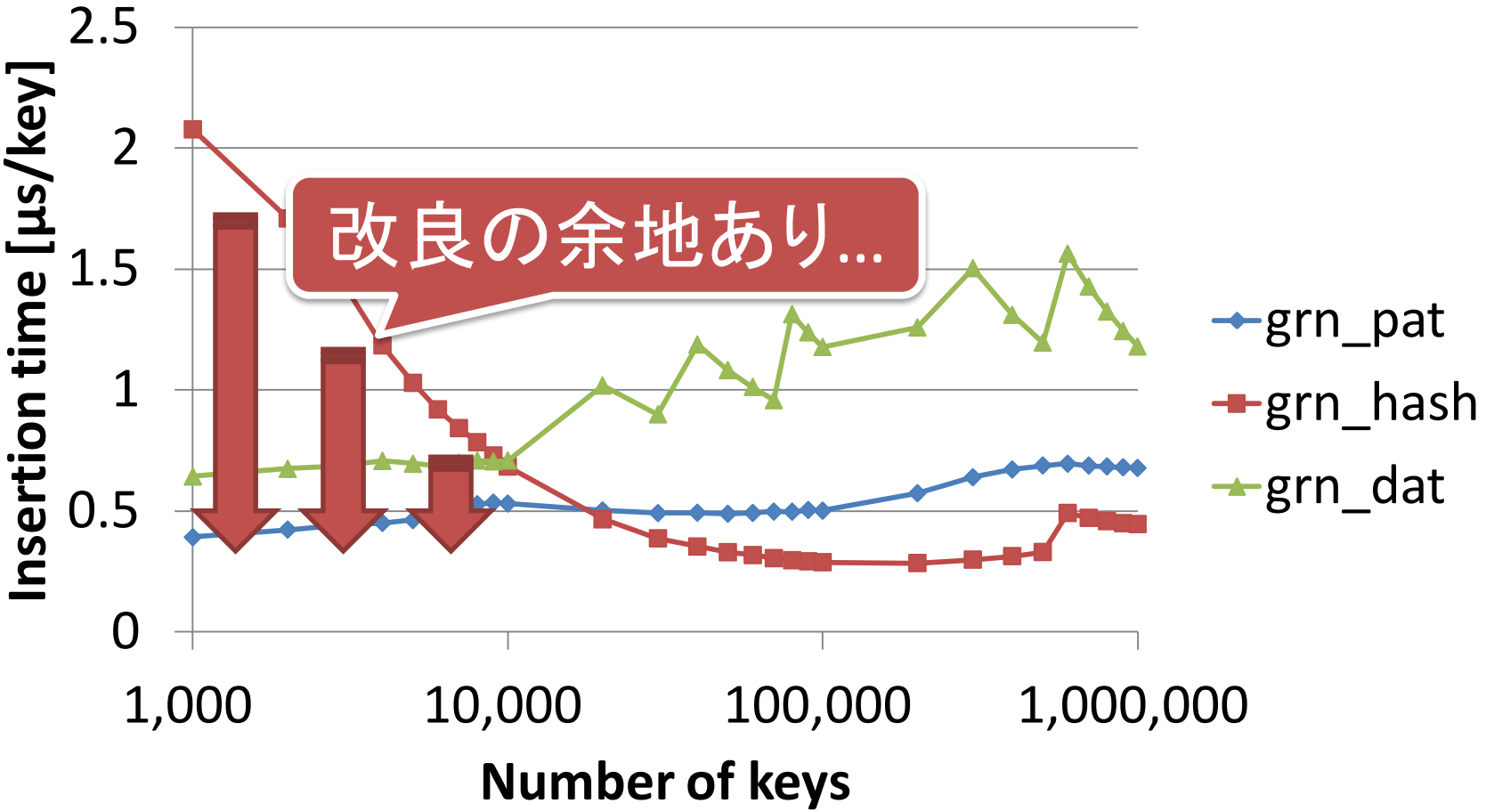
ベンチマーク（参照時間）


偏りが大きければ…
Search time [µs/key]
2
1.5
1
0.5
0
1,000
10,000
100,000
1,000,000
Number of keys
grn_pat
grn_hash
grn_dat

ふたつ

不安定なハッシュ表を調整

ベンチマーク（構築時間）

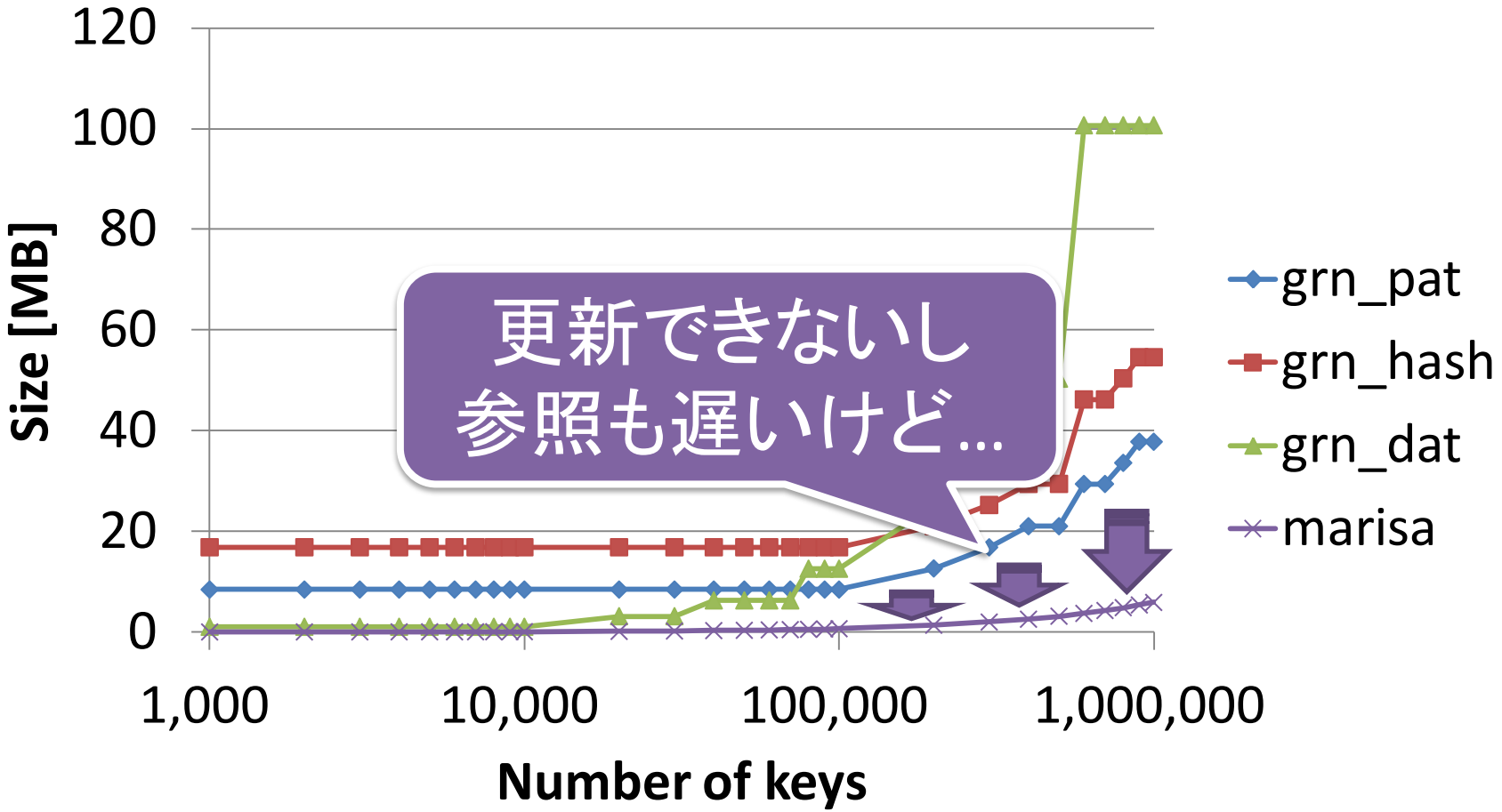

Insertion time [µs/key]
2.5
2
1.5
1
0.5
0
改良の余地あり…
1,000
10,000
100,000
1,000,000
Number of keys
grn_pat
grn_hash
grn_dat

みっつ

見る機会の少ないデータを圧縮

http://code.google.com/p/marisa-trie/

ベンチマーク（サイズ）


120
100
80
60
40
20
0
Size [MB]
更新できないし
参照も遅いけど…
grn_pat
grn_hash
grn_dat
marisa
1,000
10,000
100,000
1,000,000
Number of keys

他にもありますが

このくらいで勘弁してください

grn_dat と愉快な仲間たちに

絞って紹介しました

次回

「groonga の野望」

お楽しみに