俺のキュー
(っぽいもの)

青柳公右平

動機

- クローラーとかボットを作ってる
- マルチプロセス、マルチスレッドで大量の処理をしたい
- 対象が被らないようにして、バンバン処理してほしい
- PostgreSQL以外のミドルウェアを準備するのがめんどくさい
- テーブルを使って、キューにしてみる
- 実はあんまり順番が重要じゃないのでキューではない・・・

テーブル

```
CREATE TABLE queues (
  id BIGSERIAL NOT NULL PRIMARY KEY
  ,created_at TIMESTAMPTZ NOT NULL
  ,started_at TIMESTAMPTZ
  ,data JSONB NOT NULL
)
```

データ投入

```
INSERT INTO queues (created_at, data)
SELECT
  NOW()
  ,'{}'
FROM
  generate_series(1, 10000)
```

方針

- started_atがnullのレコードを1件取ってくる
- そのレコードを更新する
- 更新する際にはstarted_atがnullか再度検証する
- 更新できたらOK

```
SELECT t1.id, t1.data

INTO w_record

FROM queues AS t1

WHERE t1.started_at IS NULL
LIMIT 1;


UPDATE queues SET
started_at = NOW()
id = w_record.id AND started_at IS NULL;
```

問題点

- 大量にスレッドがあると同じレコードに集中して、1つのスレッド以外は全員失敗する。
- まだレコードはたくさんあるのに(><;
- ORDER BY RANDOM()してもいいが、レコード数が多くなると重い。
- なのでOFFSETをRANDOMにする。

```
SELECT t1.id, t1.data

INTO w_record

FROM queues AS t1

WHERE t1.started_at IS NULL
OFFSET (random() * 5)::INT + 1
LIMIT 1;


UPDATE queues SET
started_at = NOW()
id = w_record.id AND started_at IS NULL;
```

だいたいうまくいく

- しかし、OFFSET以上にレコードが無いとSELECTできない。
- なのでOFFSET 0で再検索

以上をまとめてストアードプロシージャにした

https://github.com/aoyagikouhei/pesudo-queue

実測環境

- MacBook Air (13-inch, Late 2010)
- DBサーバ さくらのVPS 2GB 3Core
- 4プロセス25スレッド
- 10000レコード

実測

- 1プロセス25スレッド→31.516023
  多重度による変化なし
- 4プロセス25スレッド多重度0→20.821634
- 4プロセス25スレッド多重度25→15.810679
- 4プロセス25スレッド多重度100→15.969905
- 4プロセス25スレッド多重度10000→19.655184

以上

- こんなんでいいんですかね？
- なんか他にいい方法ありますかね？
- ちゃんとミドルウェア用意した方がいいですかね？

ご清聴
ありがとうございました