
Яндекс

Яндекс Почта

PostgreSQL в Яндексе: история успеха №2

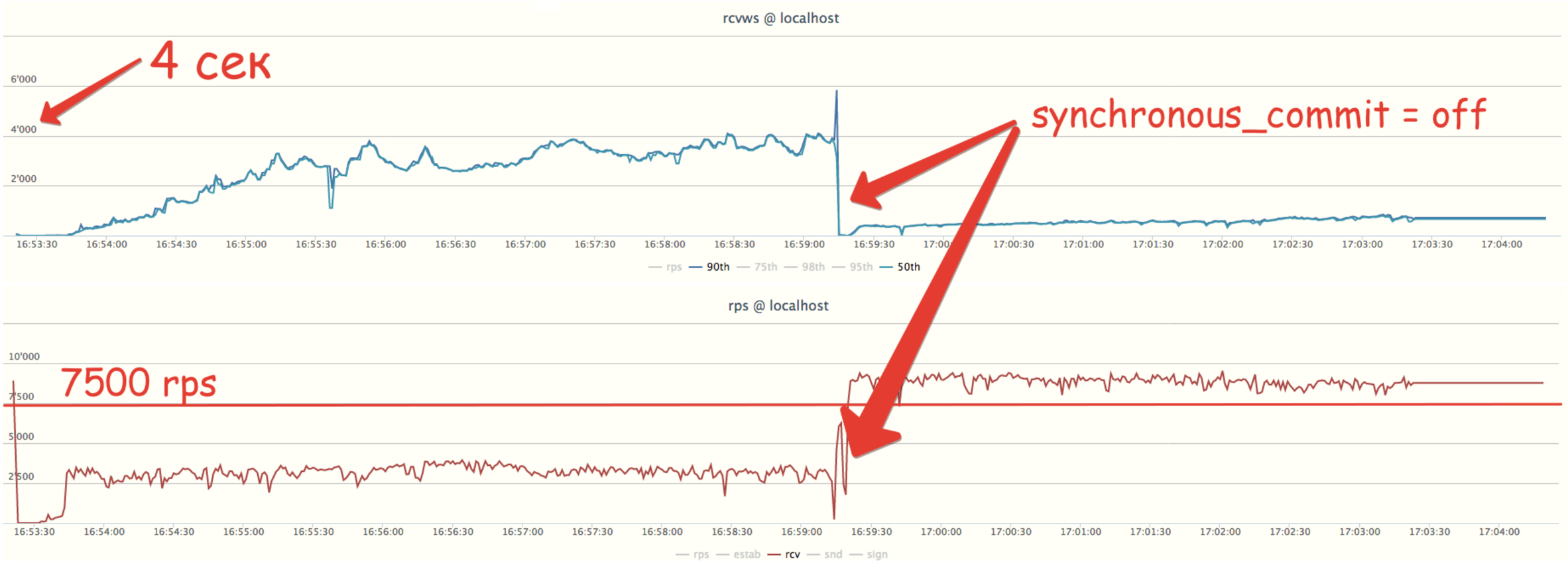
Владимир Бородин

events.yandex.ru/lib/talks/2382

Нужна небольшая табличка, с транзакциями

Разработчик

Первые проблемы



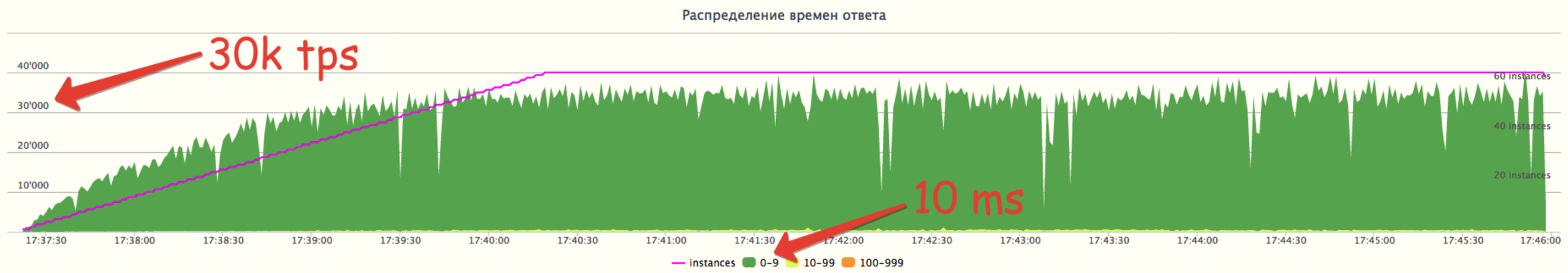
Первые проблемы

- Много мелких изменяющих запросов
- Очень медленно
- GDB показал, что виноват WALWriteLock
- synchronous_commit = off снимает симптомы
- commit_delay и commit_siblings не помогают

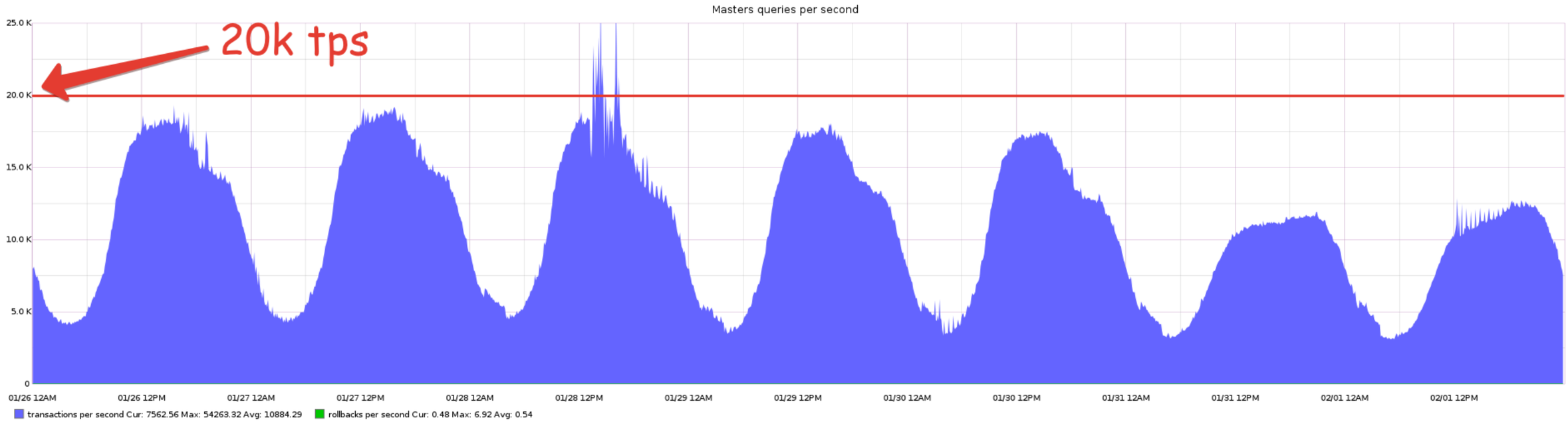
Первые решения

- pg_xlog на отдельном массиве,
- эта ФС смонтирована с nobarrier,
- wal_sync_method = fsync

Стрельба без приложения



Прямо сейчас



Нам надо ещё 10 ТБ положить

Менеджер

Детали

- Хранить надо 30 дней
- Чем старше данные, тем реже за ними приходят
- Во все читающие запросы можно добавить время

Partitioning

Варианты

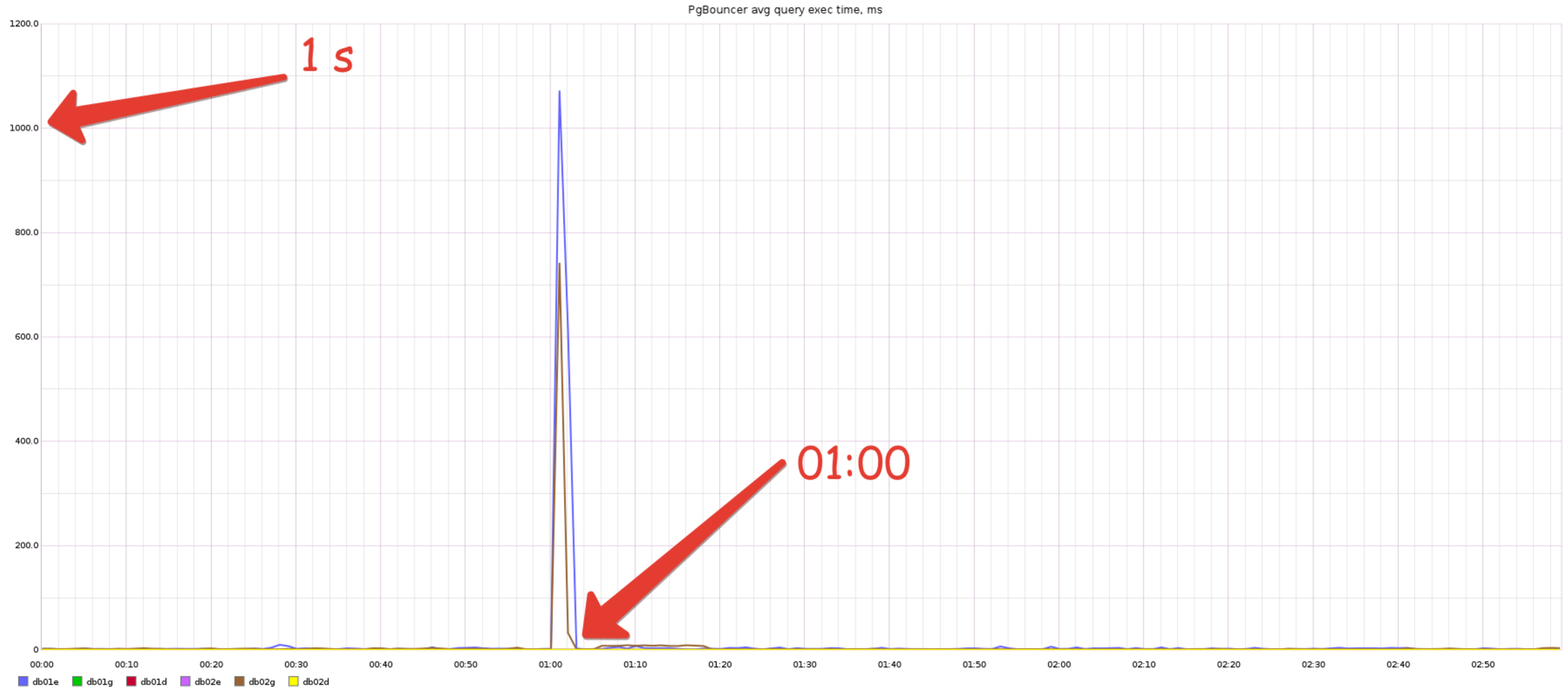
- pg_part
- pg_partman
- partition magic

pg_partman

```
CREATE SCHEMA IF NOT EXISTS partman;
CREATE EXTENSION IF NOT EXISTS pg_partman SCHEMA partman;
SELECT partman.create_parent('xiva.notifications',
        'event_dt', 'time-static');
SELECT partman.apply_foreign_keys('xiva.notifications');
SELECT partman.apply_constraints('xiva.notifications');
UPDATE partman.part_config
        SET retention = '30 days'
        WHERE parent_table = 'xiva.notifications';
```

```
0 1 * * *   postgres psql xivadb -c "select partman.run_maintenance(false);"
 >/dev/null 2>&1
```

Обновление функции триггера на мастер-таблице

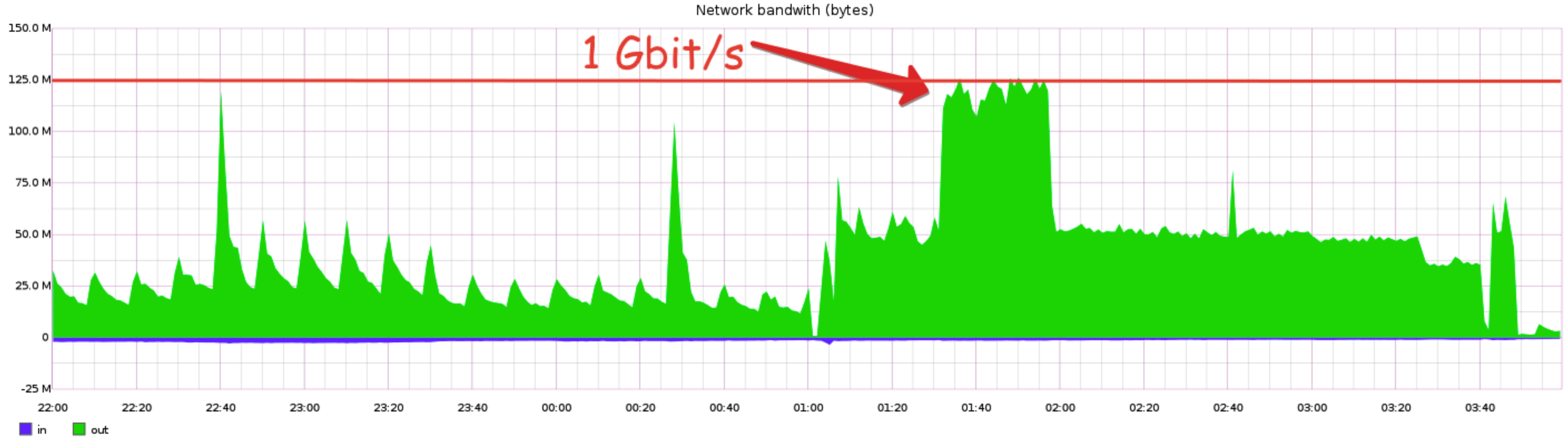


Перекладывание данных с SSD на SATA

Варианты

- ALTER TABLE foo SET TABLESPACE sata;
- http://reorg.github.io/pg_repack/

Большое количество WAL’ов



Решение

- Средствами БД победить не удалось
- На помощь пришли cgroups
- Ограничение на объём данных для группы процессов

```
group moving_to_sata {
    perm {
        admin {
            uid = postgres;
            gid = postgres;
        }
        task {
            uid = postgres;
            gid = postgres;
        }
    }
    blkio {
        blkio.throttle.write_iops_device="";
        blkio.throttle.read_iops_device="";
        blkio.throttle.write_bps_device="9:3 10485760";
        blkio.throttle.read_bps_device="9:2 10485760";
        blkio.weight="500";
        blkio.weight_device="";
    }
}
```

Решение

```
72      def get_list_of_local_pids(self):
73          try:
74              cur = self.conn.cursor()
75              cur.execute("SELECT pid FROM pg_stat_activity WHERE usename='postgres' AND \
76                      (query LIKE 'CREATE%' OR query LIKE 'autovacuum%(to prevent wraparound)');")
77              res = cur.fetchall()
78              cur.close()
79              return [i[0] for i in res]
80          except Exception as err:
81              self.log.error(str(err))
82              sys.exit(1)
```

```
142         for pid in pids:
143             f = open('/cgroup/blkio/moving_to_sata/tasks', 'w')
144             f.write(str(pid))
145             f.close()
```

github.com/reorg/pg_repack/pull/34

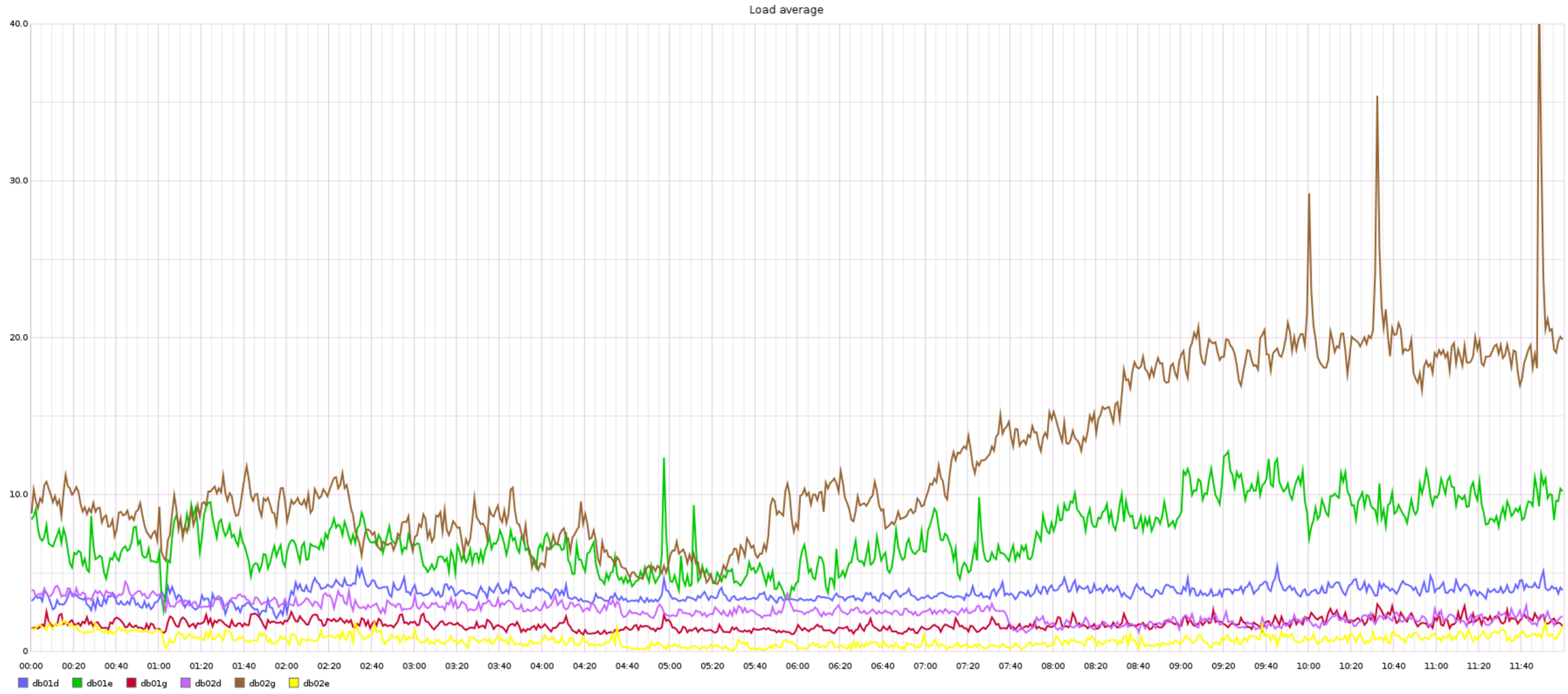
Autovacuum to prevent wraparound

```
xivastore01e/postgres # SELECT query FROM pg_stat_activity WHERE query ILIKE '%vacuum%';
                                query
-------------------------------------------------------------------------
 autovacuum: VACUUM xiva.notifications_p2014_12_26 (to prevent wraparound)
 autovacuum: VACUUM xiva.notifications_p2014_12_17 (to prevent wraparound)
 SELECT query FROM pg_stat_activity WHERE query ILIKE '%vacuum%';
 autovacuum: VACUUM xiva.notifications_p2014_12_29 (to prevent wraparound)
 autovacuum: VACUUM xiva.notifications_p2014_12_21 (to prevent wraparound)
 autovacuum: VACUUM xiva.notifications_p2014_12_22 (to prevent wraparound)
 autovacuum: VACUUM xiva.notifications_p2014_12_18 (to prevent wraparound)
 autovacuum: VACUUM xiva.notifications_p2014_12_19 (to prevent wraparound)
 autovacuum: VACUUM xiva.notifications_p2014_12_23 (to prevent wraparound)
 autovacuum: VACUUM xiva.notifications_p2014_12_20 (to prevent wraparound)
 autovacuum: VACUUM xiva.notifications_p2014_12_25 (to prevent wraparound)
(11 rows)

Time: 1.495 ms
xivastore01e/postgres #
```

Autovacuum to prevent wraparound

```
xivastore01d/postgres # SELECT name, setting, boot_val
FROM pg_settings WHERE name IN
('autovacuum_freeze_max_age', 'vacuum_freeze_table_age');
-[ RECORD 1 ]------------------------
name     | autovacuum_freeze_max_age
setting  | 2000000000
boot_val | 200000000
-[ RECORD 2 ]------------------------
name     | vacuum_freeze_table_age
setting  | 1900000000
boot_val | 150000000

Time: 0.991 ms
xivastore01d/postgres #
```

Autovacuum to prevent wraparound



ERROR: database is not accepting commands to avoid wraparound data loss in database "xivadb"

PostgreSQL

Autovacuum to prevent wraparound

- Выключили throttling
- Последовательный VACUUM таблиц на SATA

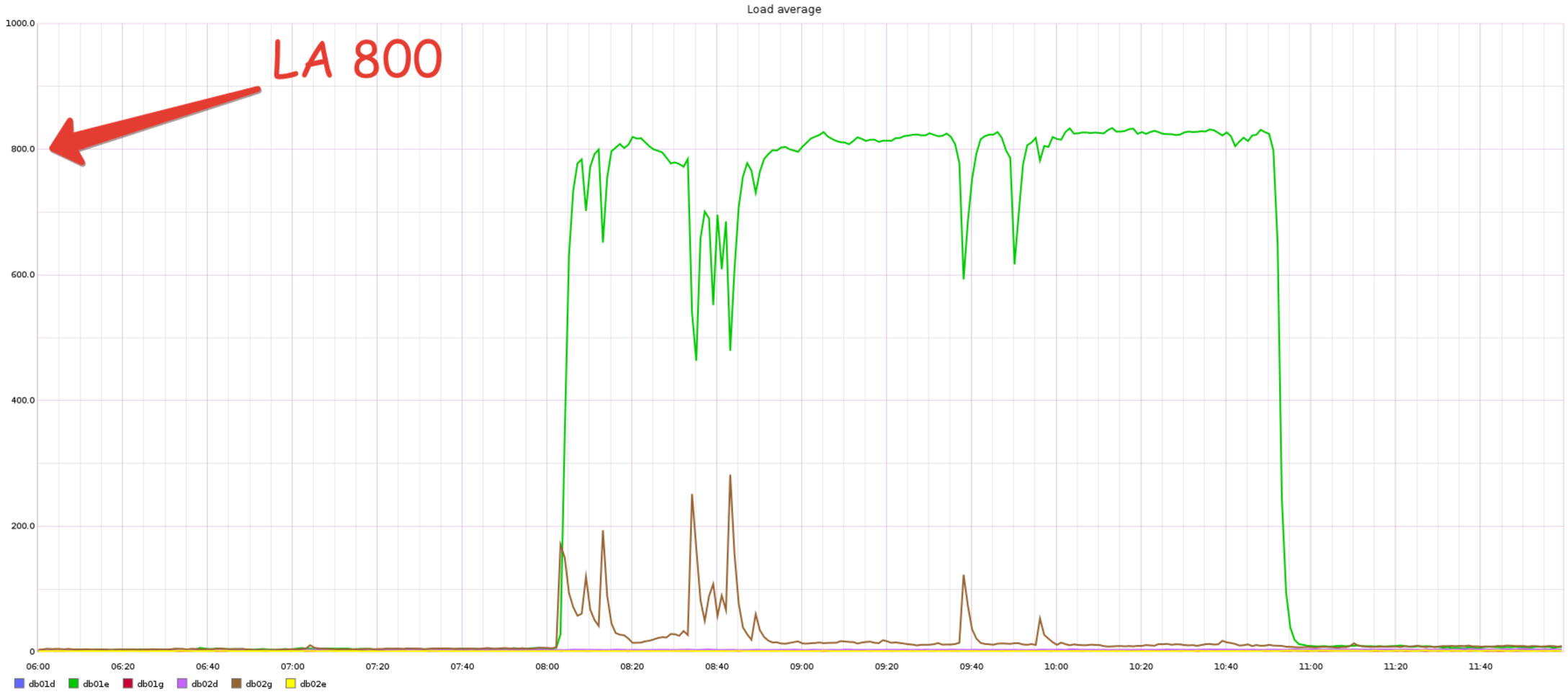
Бэкапы

```
root@xivastore01d ~ # fgrep 10404 /var/log/zkflock.log | fgrep 10609
2015-01-24 02:00:02 INFO     10404 Start subprocess: /etc/cron.yandex/pgb
ackup_database.sh (PID: 10609)
2015-01-25 11:53:43 INFO     10404 Child returned code: 0 (PID: 10609)
root@xivastore01d ~ # echo "SELECT '2015-01-25 11:53:43'::timestamp - '20
15-01-24 02:00:02'::timestamp;" | psql
    ?column?
---------------
 1 day 09:53:41
(1 row)

Time: 0.819 ms
root@xivastore01d ~ # grep '\--format=tar --gzip' /etc/cron.yandex/pgback
up_database.sh
        pg_basebackup --pgdata=${backup_path}/$date --format=tar --gzip -
-label=$date
root@xivastore01d ~ #
```

Barman 1.4.0

http://www.postgresql.org/about/news/1566/

Запрос во все партиции

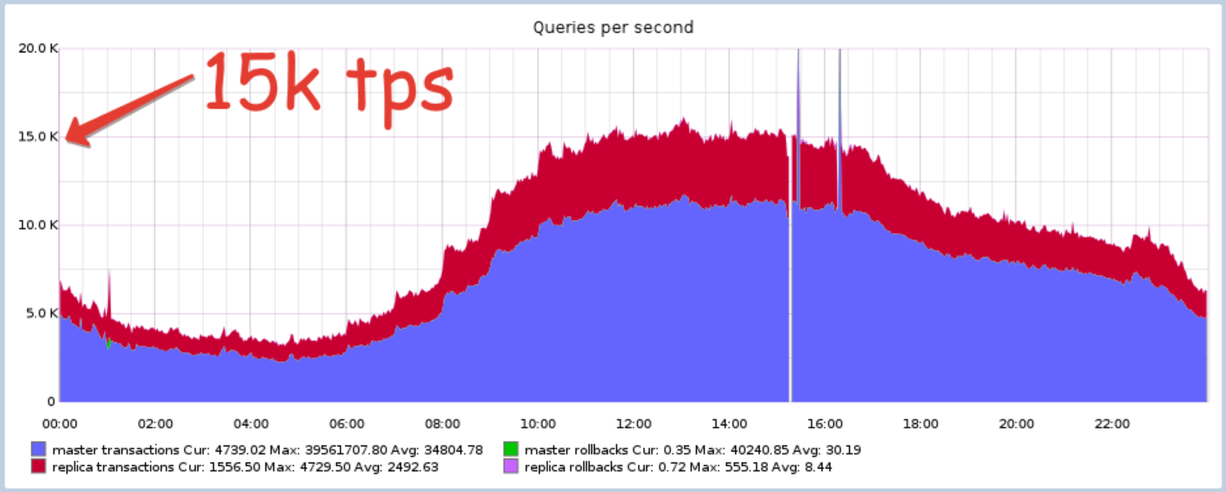


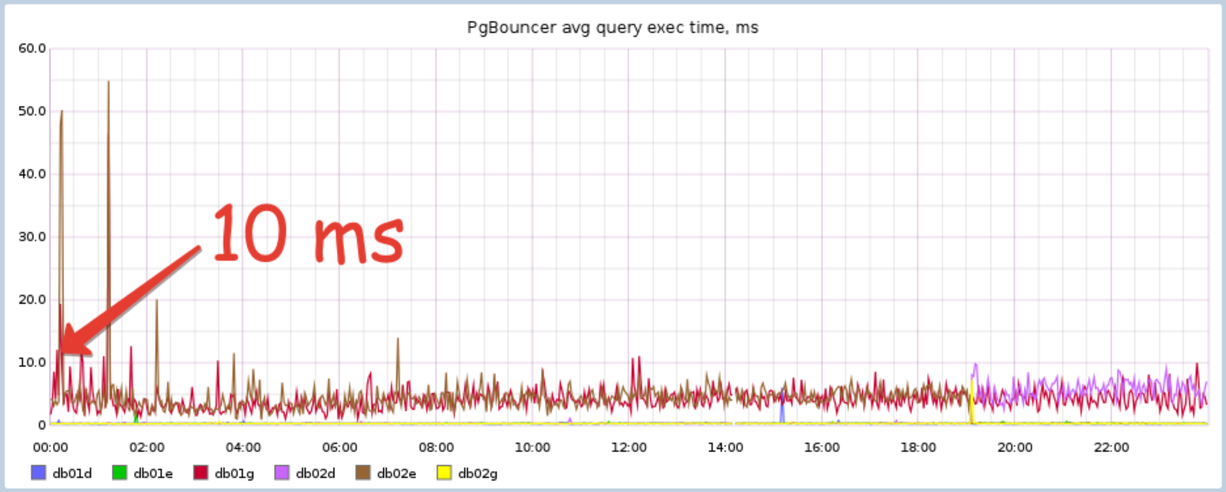
Запрос во все партиции

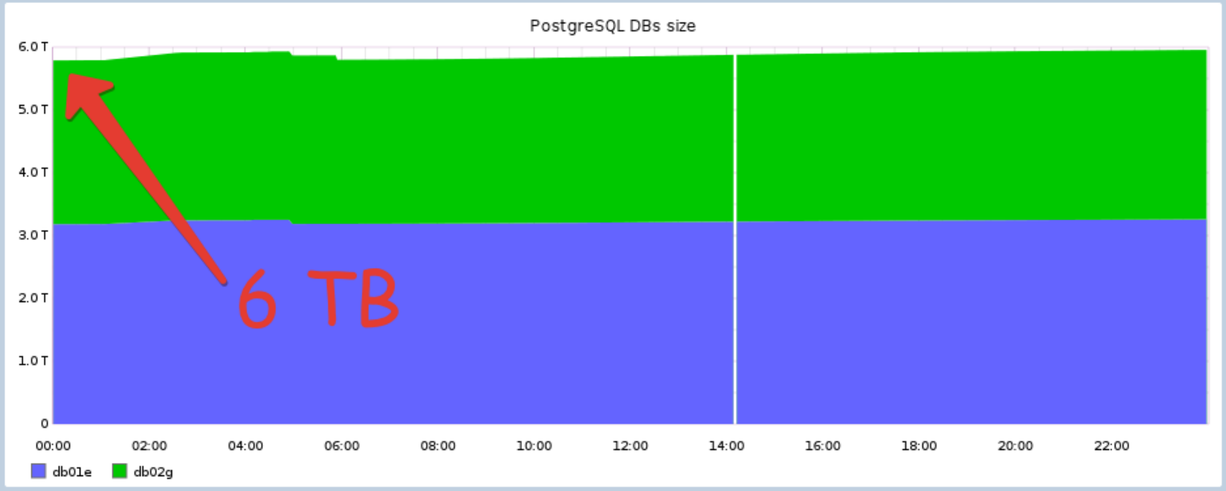
```
xivastore01e/xivadb # explain SELECT local_id FROM xiva.notifications
WHERE uid = 1 AND sid = 1 AND hash = 'blah'
AND event_dt = CURRENT_TIMESTAMP;
                                                                QUERY PLAN
-------------------------------------------------------------------------------------------
 Append  (cost=0.00..301.11 rows=36 width=8)
   ->  Seq Scan on notifications  (cost=0.00..0.00 rows=1 width=8)
         Filter: ((uid = 1) AND (sid = 1) AND (hash = 'blah'::text) AN
   ->  Index Scan using notifications_p2015_01_04_uid_sid_hash_key on
         Index Cond: ((uid = 1) AND (sid = 1) AND (hash = 'blah'::text
         Filter: (event_dt = now())
   ->  Index Scan using notifications_p2015_01_05_uid_sid_hash_key on
         Index Cond: ((uid = 1) AND (sid = 1) AND (hash = 'blah'::text
         Filter: (event_dt = now())
```

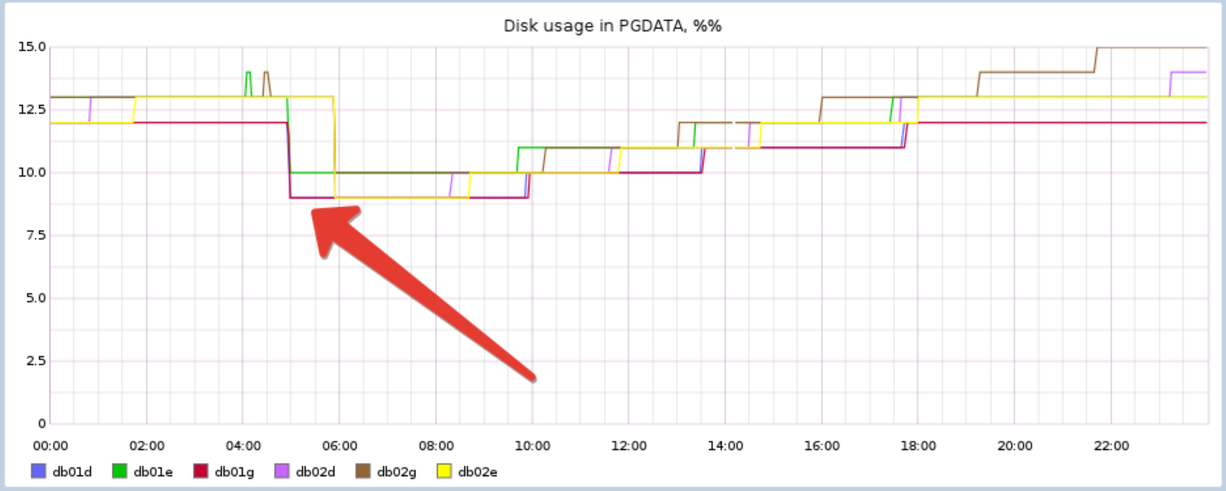
Решение

```
      18  +    start_of_day timestamp with time zone;
      19  +    end_of_day timestamp with time zone;
18    20   BEGIN
19    21       IF (i_event_dt IS NULL) THEN
20    22           SELECT current_timestamp INTO i_event_dt;
21    23       END IF;
      24  +    SELECT date_trunc('day', i_event_dt),
      25  +        date_trunc('day', i_event_dt + '1 day'::interval)
      26  +        INTO start_of_day, end_of_day;
22    27
23    28       SELECT sid INTO service_id FROM xiva.services WHERE service_name = i_service;
24    29       IF NOT FOUND THEN
          @@ -60,8 +65,8 @@ BEGIN
60    65                   WHERE n.uid = i_uid
61    66                       AND n.sid = service_id
62    67                       AND n.hash = i_hash
63        -                    AND n.event_dt >= date_trunc('day', i_event_dt)
64        -                    AND n.event_dt < date_trunc('day', i_event_dt + '1 day'::interval);
      68  +                    AND n.event_dt >= start_of_day
      69  +                    AND n.event_dt < end_of_day;
```


Queries per second
15k tps
20.0 K
15.0 K
10.0 K
5.0 K
0
00:00 02:00 04:00 06:00 08:00 10:00 12:00 14:00 16:00 18:00 20:00 22:00
master transactions Cur: 4739.02 Max: 39561707.80 Avg: 34804.78
replica transactions Cur: 1556.50 Max: 4729.50 Avg: 2492.63
master rollbacks Cur: 0.35 Max: 40240.85 Avg: 30.19
replica rollbacks Cur: 0.72 Max: 555.18 Avg: 8.44


PgBouncer avg query exec time, ms
10 ms
60.0
50.0
40.0
30.0
20.0
10.0
0
00:00 02:00 04:00 06:00 08:00 10:00 12:00 14:00 16:00 18:00 20:00 22:00
db01d db01e db01g db02d db02e db02g


PostgreSQL DBs size
6 TB
6.0 T
5.0 T
4.0 T
3.0 T
2.0 T
1.0 T
0
00:00 02:00 04:00 06:00 08:00 10:00 12:00 14:00 16:00 18:00 20:00 22:00
db01e db02g

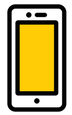

Disk usage in PGDATA, %%
15.0
12.5
10.0
7.5
5.0
2.5
0
00:00 02:00 04:00 06:00 08:00 10:00 12:00 14:00 16:00 18:00 20:00 22:00
db01d db01e db01g db02d db02g db02e

Вопросы?

Контакты

Владимир Бородин

Системный администратор

 +7 (495) 739 70 00, доб. 7255

 @man_brain

 d0uble@yandex-team.ru

 http://simply.name