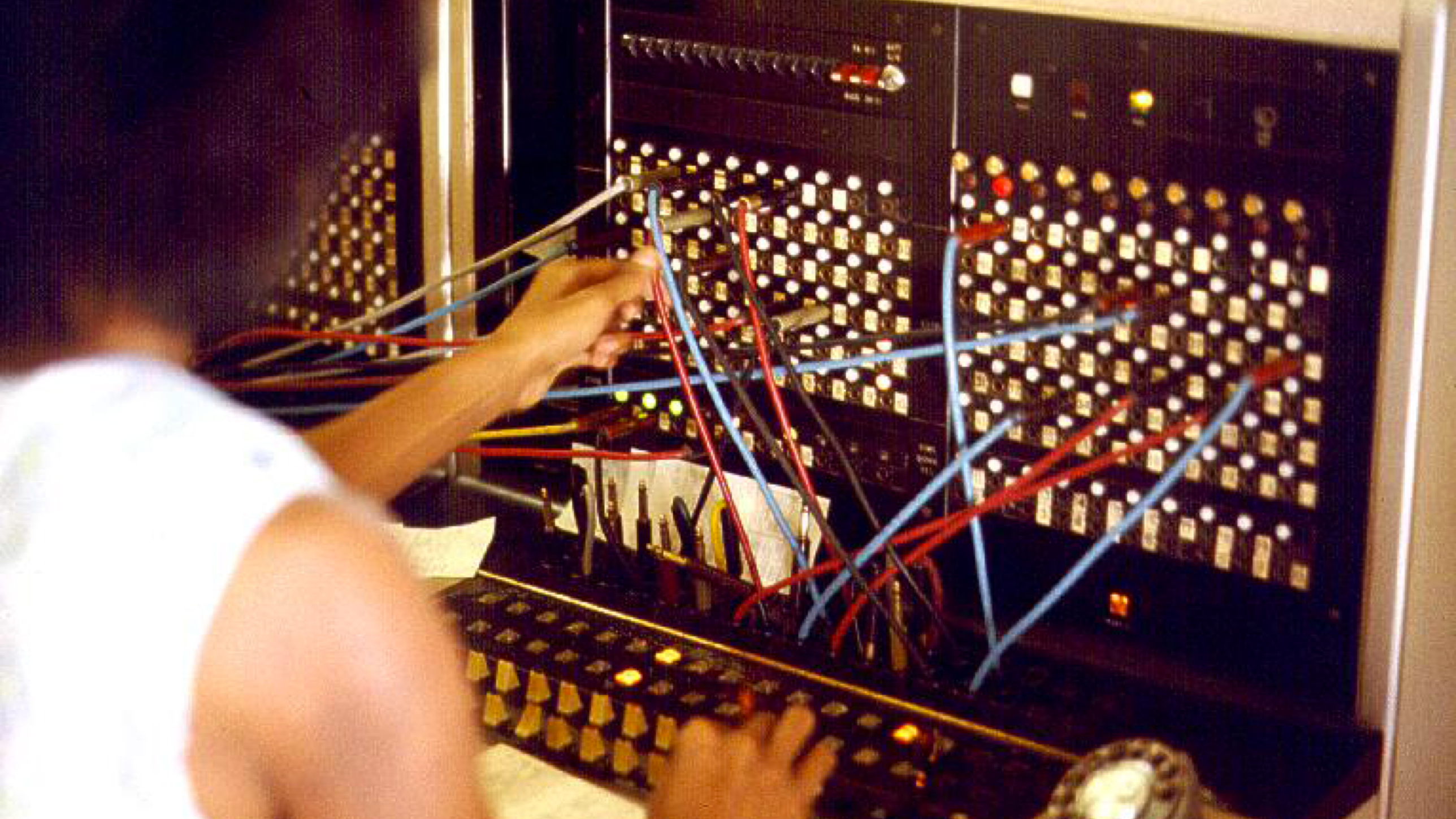

Яндекс

Яндекс Облако

Odyssey
архитектура, настройка, мониторинг

Андрей Бородин, руководитель подразделения разработки РСУБД с открытыми исходных кодом



PostgreSQL в Яндексе

Яндекс.Почта

- Сколько-то сотен миллионов пользователей
- 1+ триллион строк, 1+ миллион запросов в секунду

Яндекс.Облако

- Несколько петабайт Постгреса
- Много разных сервисов Яндекса живут в Облаке

Зачем экономить соединения с базой?

Зачем экономить соединения с базой?

1 backend == 1 process

Зачем экономить соединения с базой?

1 backend == 1 process
Много разных кэшей

- Relations cache
- Compiled PL\pgSQL
- Plans cache

OLTP-нагрузка

```
Snapshot
GetSnapshotData(Snapshot snapshot)
{
        ...
                /*
                 * Spin over procArray checking xid, xmin, and subxids.  The goal is
                 * to gather all active xids, find the lowest xmin, and try to record
                 * subxids.
                 */
                numProcs = arrayP->numProcs;
                for (index = 0; index < numProcs; index++)
                {
                        int                     pgprocno = pgprocnos[index];
                        PGXACT     *pgxact = &allPgXact[pgprocno];
                        TransactionId xid;
```

Где экономить соединения?

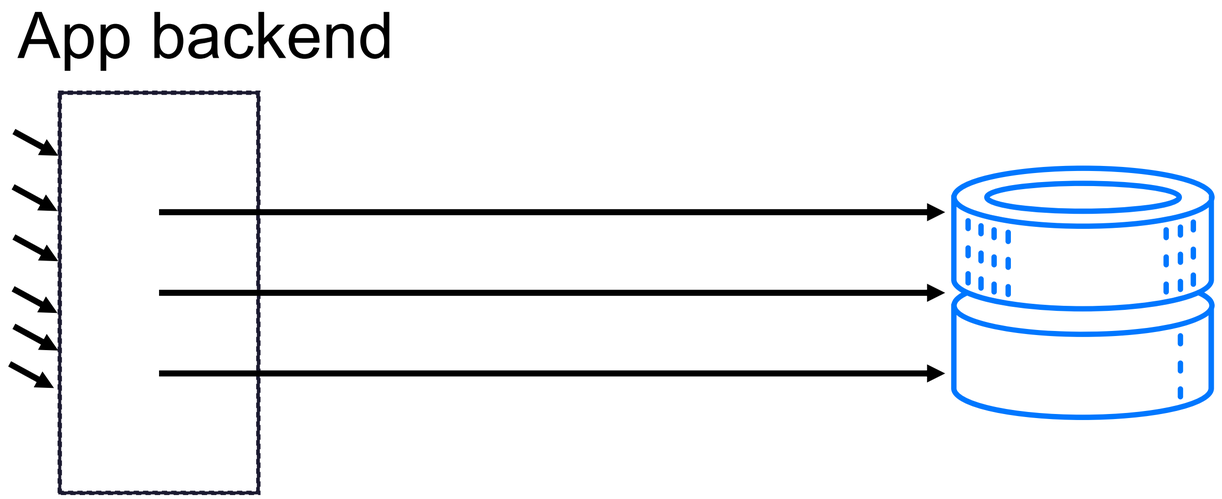
1 | Application-side pool

2 | Proxy pool

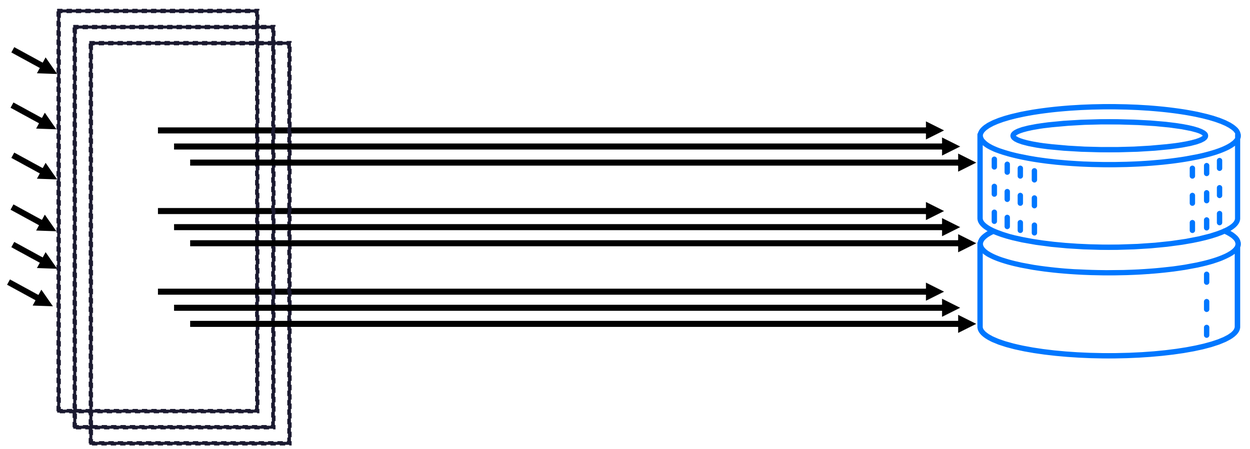
4 | DB built-in pooling

7 | Combinations

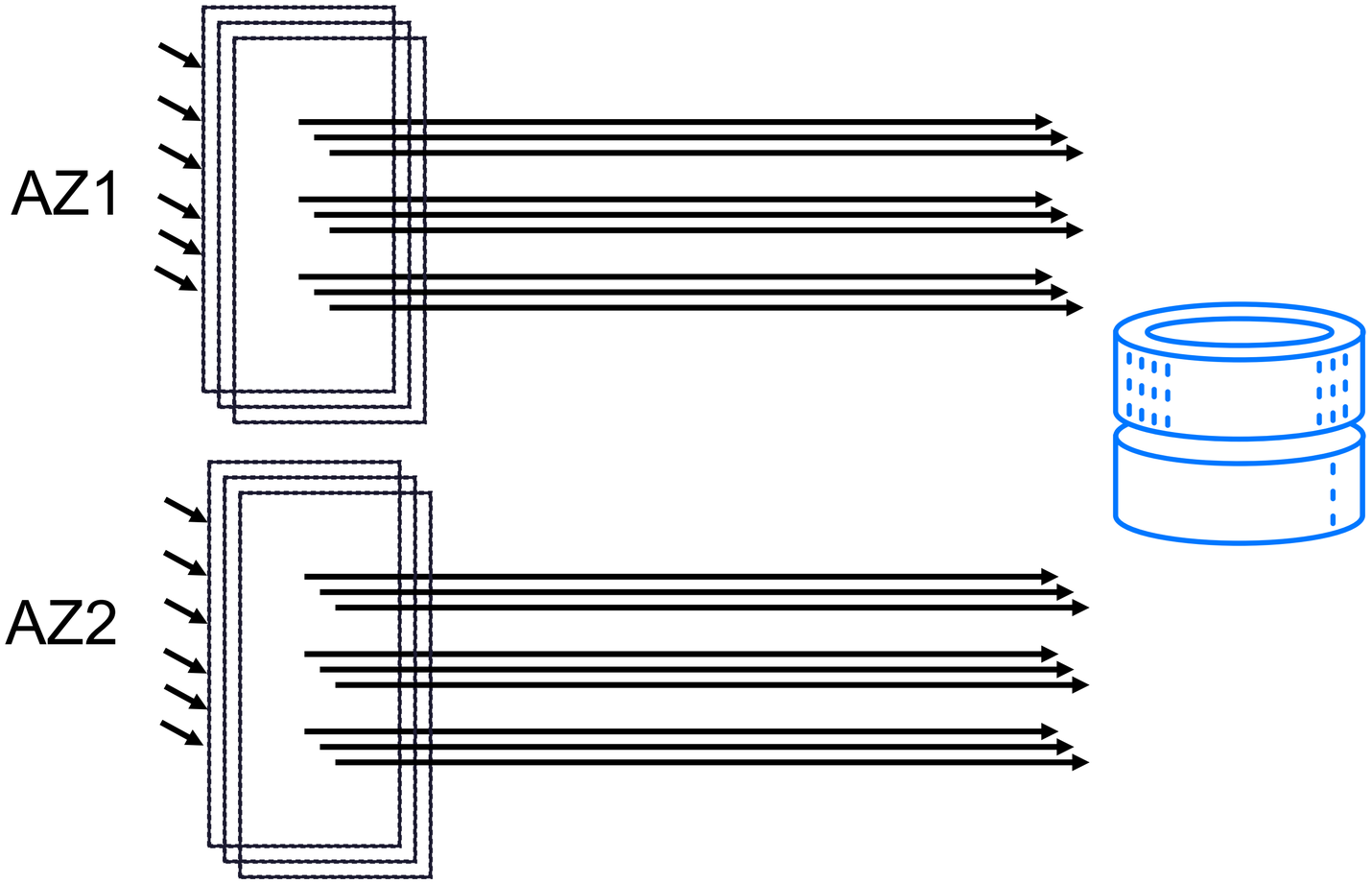
Application-side pool



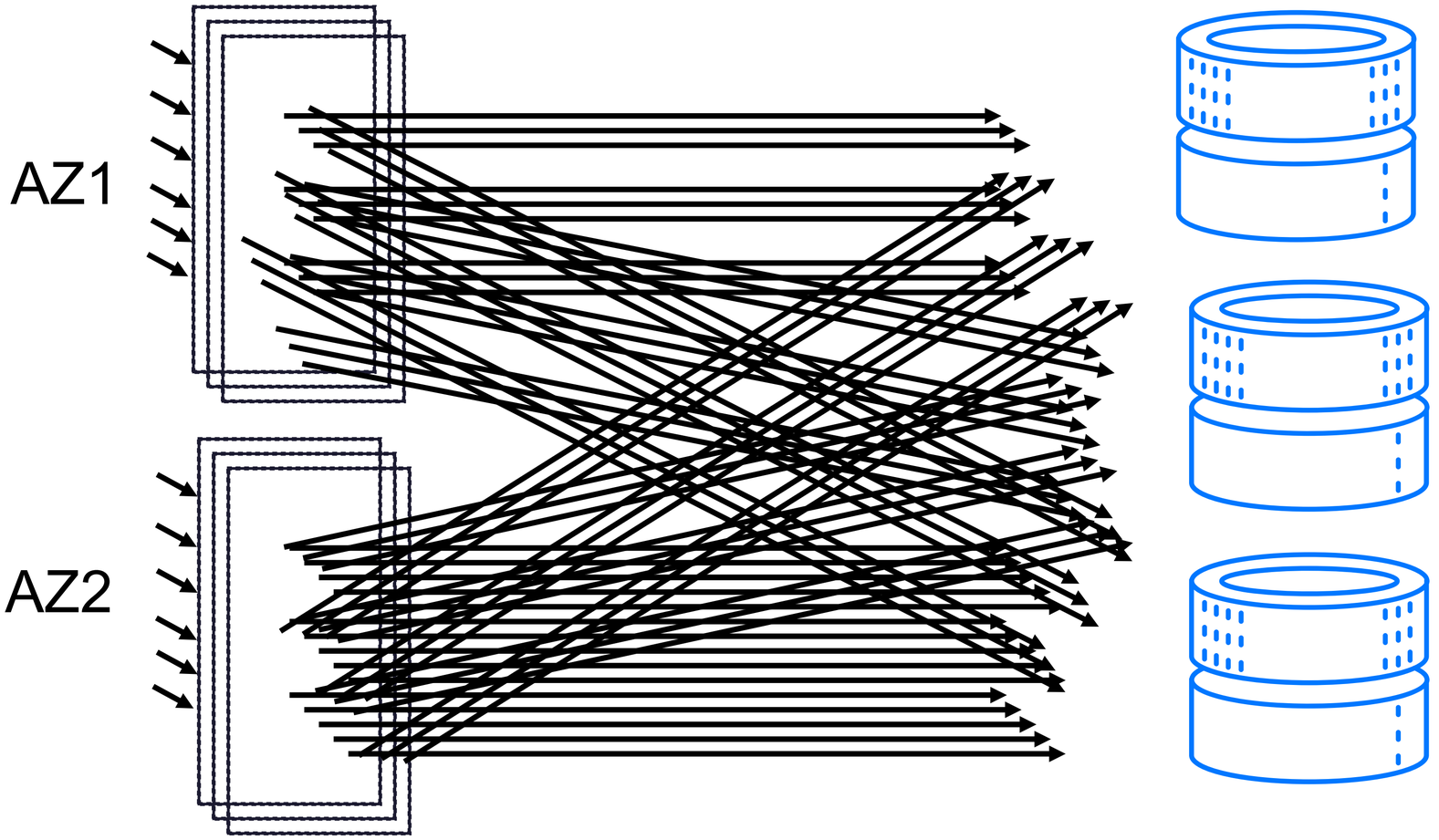
Распределение нагрузки между приложениями



В нескольких зонах доступности



С шардированием



Proxy poolers

Pgpool II
Crunchy-Proxy

- Много разной функциональности
- Только сессионный пулинг

PgBouncer

- Transaction pooling

Проблемы с PgBouncer

```
2017-03-13 10:48:23.995 24408 ERROR S: login failed: FATAL: too many connections for role "YYY"

psycopg2.OperationalError: ERROR:  pgbouncer cannot connect to server

>>> try:

...     conn = psycopg2.connect("port=6432 ...")

... except psycopg2.Error as e:

...     print(e.pgcode)

...

None

>>>
```

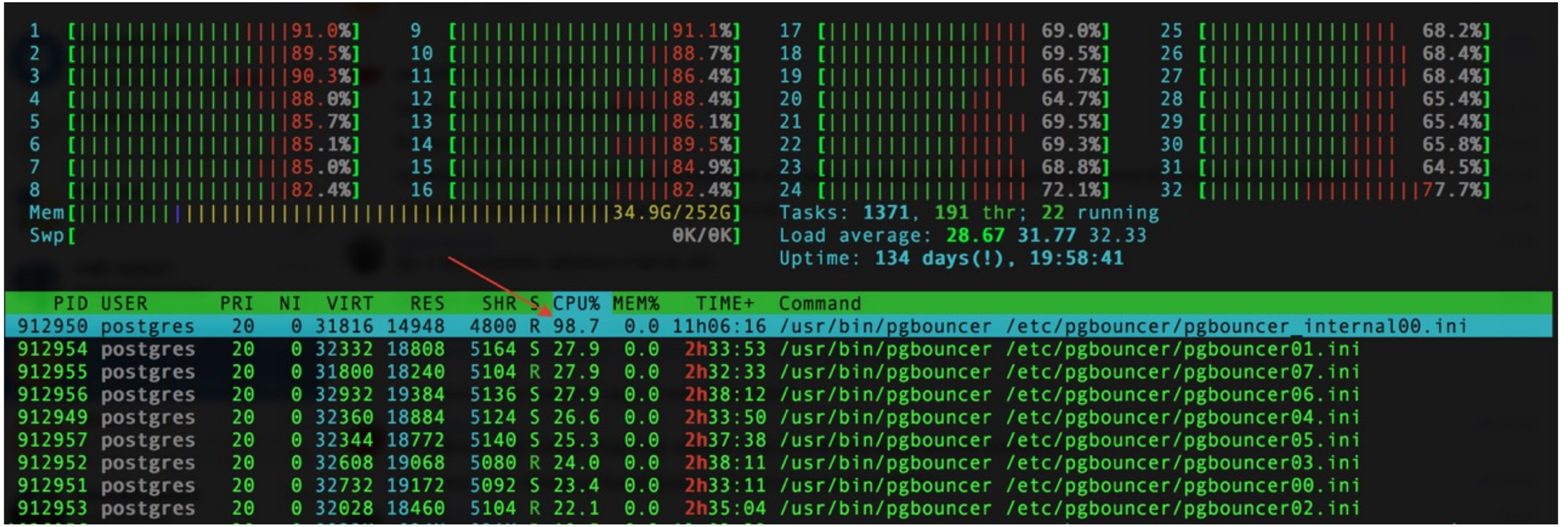
Проблемы с PgBouncer

```
  PID USER      PRI  NI  VIRT   RES   SHR S CPU% MEM%   TIME+  Command
 3548 pgbounce  10 -10 55344 16860   904 R 97.0  0.0 58h15:40 /usr/bin/pgbouncer -d -q /etc/pgbouncer/pgbouncer.ini
18561 postgres  20   0 66.1G  1804   712 S  7.0  0.0  5h51:48 postgres: wal writer process
21655 postgres  20   0 66.1G  3484  1876 S  5.0  0.0 50:00.07 postgres: wal sender process repl xivadb04d.mail.yandex.net(48473)
21688 postgres  20   0 66.1G  3484  1876 S  5.0  0.0 49:56.82 postgres: wal sender process repl xivadb04g.mail.yandex.net(36924)
26749 root      20   0 15968  1836  1048 R  3.0  0.0  0:02.06 htop
```

SO_REUSEPORT

https://lwn.net/Articles/542629/

```
+   if (af != AF_UNIX && cf_listen_reuseport == 1) {
+       int val = 1;
+       errpos = "setsockopt";
+       res = setsockopt(sock, SOL_SOCKET, SO_REUSEPORT, &val, sizeof(val));
+       if (res < 0)
+           goto failed;
+   }
```


1 [||||||||||||||||91.0%]
2 [||||||||||||||||89.5%]
3 [||||||||||||||||90.3%]
4 [||||||||||||||||88.0%]
5 [||||||||||||||||85.7%]
6 [||||||||||||||||85.1%]
7 [||||||||||||||||85.0%]
8 [||||||||||||||||82.4%]
9 [||||||||||||||||91.1%]
10 [||||||||||||||||88.7%]
11 [||||||||||||||||86.4%]
12 [||||||||||||||||88.4%]
13 [||||||||||||||||86.1%]
14 [||||||||||||||||89.5%]
15 [||||||||||||||||84.9%]
16 [||||||||||||||||82.4%]
17 [||||||||||||||| 69.0%]
18 [||||||||||||||| 69.5%]
19 [||||||||||||||| 66.7%]
20 [||||||||||||||| 64.7%]
21 [||||||||||||||| 69.5%]
22 [||||||||||||||| 69.3%]
23 [||||||||||||||| 68.8%]
24 [||||||||||||||| 72.1%]
25 [||||||||||||||| 68.2%]
26 [||||||||||||||| 68.4%]
27 [||||||||||||||| 68.4%]
28 [||||||||||||||| 65.4%]
29 [||||||||||||||| 65.4%]
30 [||||||||||||||| 65.8%]
31 [||||||||||||||| 64.5%]
32 [||||||||||||||| 77.7%]
Mem[||||||||||||||||||||||||||||||||34.9G/252G]
Swp[0K/0K]
Tasks: 1371, 191 thr; 22 running
Load average: 28.67 31.77 32.33
Uptime: 134 days(!), 19:58:41
PID USER PRI NI VIRT RES SHR S CPU% MEM% TIME+ Command
912950 postgres 20 0 31816 14948 4800 R 98.7 0.0 11h06:16 /usr/bin/pgbouncer /etc/pgbouncer/pgbouncer_internal00.ini
912954 postgres 20 0 32332 18808 5164 S 27.9 0.0 2h33:53 /usr/bin/pgbouncer /etc/pgbouncer/pgbouncer01.ini
912955 postgres 20 0 31800 18240 5104 R 27.9 0.0 2h32:33 /usr/bin/pgbouncer /etc/pgbouncer/pgbouncer07.ini
912956 postgres 20 0 32932 19384 5136 S 27.9 0.0 2h38:12 /usr/bin/pgbouncer /etc/pgbouncer/pgbouncer06.ini
912949 postgres 20 0 32360 18884 5124 S 26.6 0.0 2h33:50 /usr/bin/pgbouncer /etc/pgbouncer/pgbouncer04.ini
912957 postgres 20 0 32344 18772 5140 S 25.3 0.0 2h37:38 /usr/bin/pgbouncer /etc/pgbouncer/pgbouncer05.ini
912952 postgres 20 0 32608 19068 5080 R 24.0 0.0 2h38:11 /usr/bin/pgbouncer /etc/pgbouncer/pgbouncer03.ini
912951 postgres 20 0 32732 19172 5092 S 23.4 0.0 2h33:11 /usr/bin/pgbouncer /etc/pgbouncer/pgbouncer00.ini
912953 postgres 20 0 32028 18460 5104 R 22.1 0.0 2h35:04 /usr/bin/pgbouncer /etc/pgbouncer/pgbouncer02.ini

Отмена запроса

Клиент здорового человека

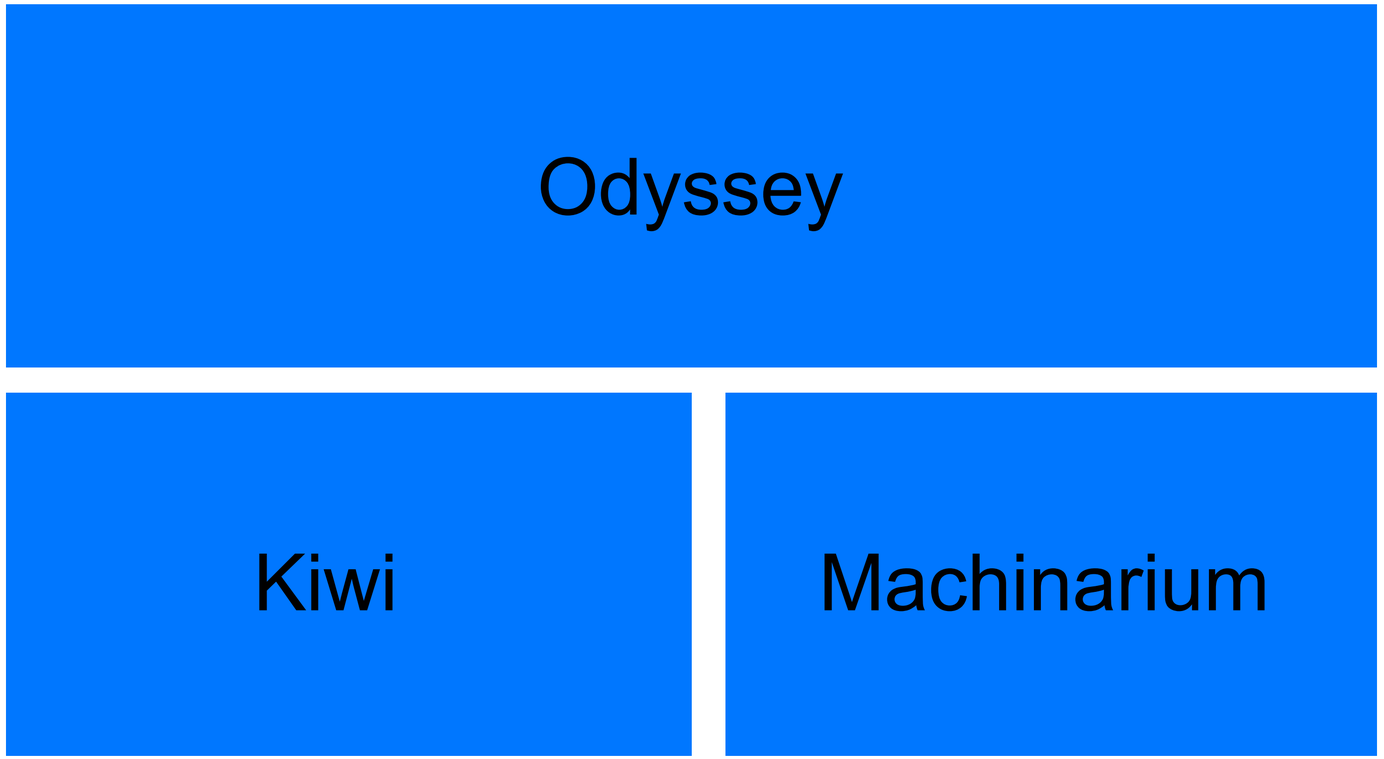
- Открывает новое соединение без аутентификации
- Отправляет токен отмены backend'a
- postgresql.org/docs/current/static/libpq-cancel.html

Клиент курильщика

- Отправит TCP reset

github.com/pgbouncer/pgbouncer/pull/79

Odyssey


Odyssey
Kiwi
Machinarium

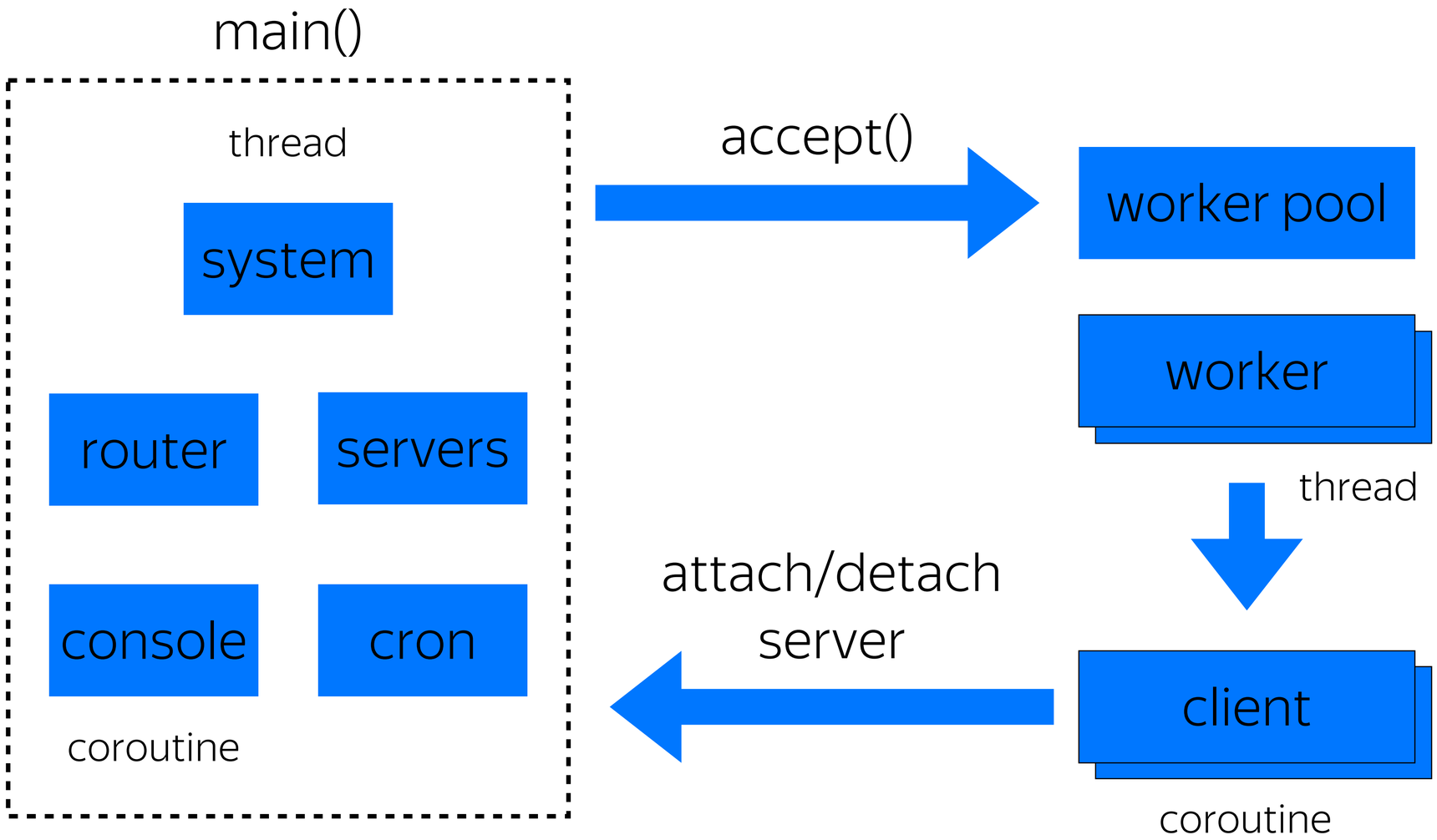
Kiwi

```
KIWI_API static inline machine_msg_t*
kiwi_fe_write_cancel(machine_msg_t *msg, uint32_t pid, uint32_t key)
{
    int size = sizeof(uint32_t) + /* len */
               sizeof(uint32_t) + /* special */
               sizeof(uint32_t) + /* pid */
               sizeof(uint32_t);  /* key */
    int offset = 0;
    if (msg)
        offset = machine_msg_size(msg);
    .....
```

Machinarium

```
void csw_worker(void *arg) {
    while (csw < 100000) {
        machine_sleep( time_ms: 0);
        csw++;
    }
}
void csw_runner(void *arg) {
    int rc = machine_coroutine_create(csw_worker,  arg: NULL);
    machine_join(rc);
    machine_stop();
}
void machinarium_test_context_switch(void) {
    machinarium_init();
    int id = machine_create( name: "test", csw_runner,  arg: NULL);
    machine_wait(id);
    machinarium_free();
}
```

Архитектура



yandex / odyssey

Unwatch 60 | Unstar 1.2k | Fork 41

Code | Issues 23 | Pull requests 5 | Projects 0 | Wiki | Security | Insights

Out of memory #52

Edit | New issue

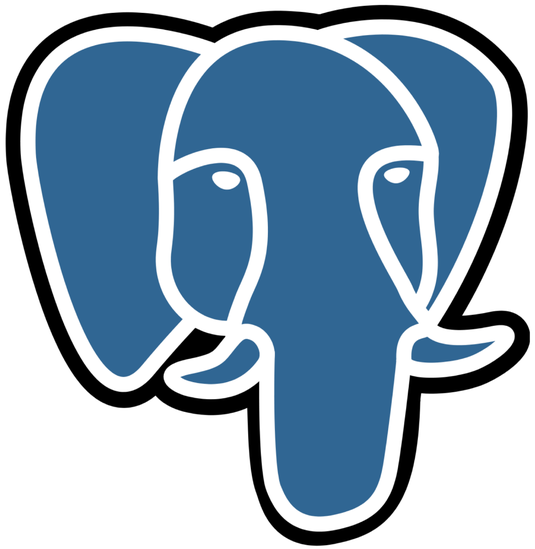
Closed **rzharkov** opened this issue on 19 Mar · 9 comments

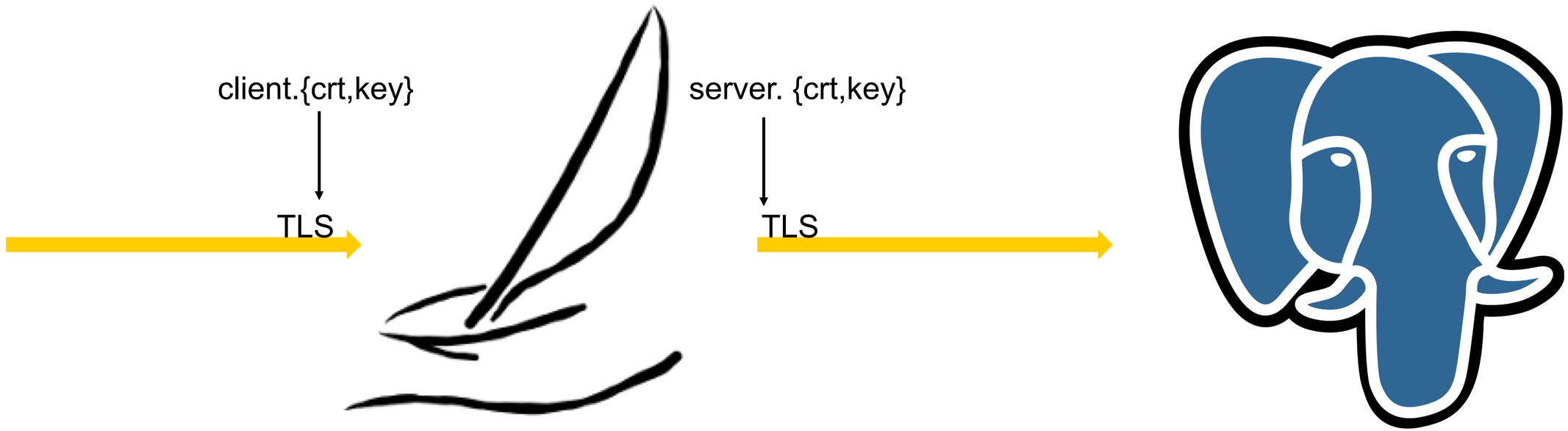
rzharkov commented on 19 Mar

Assignees

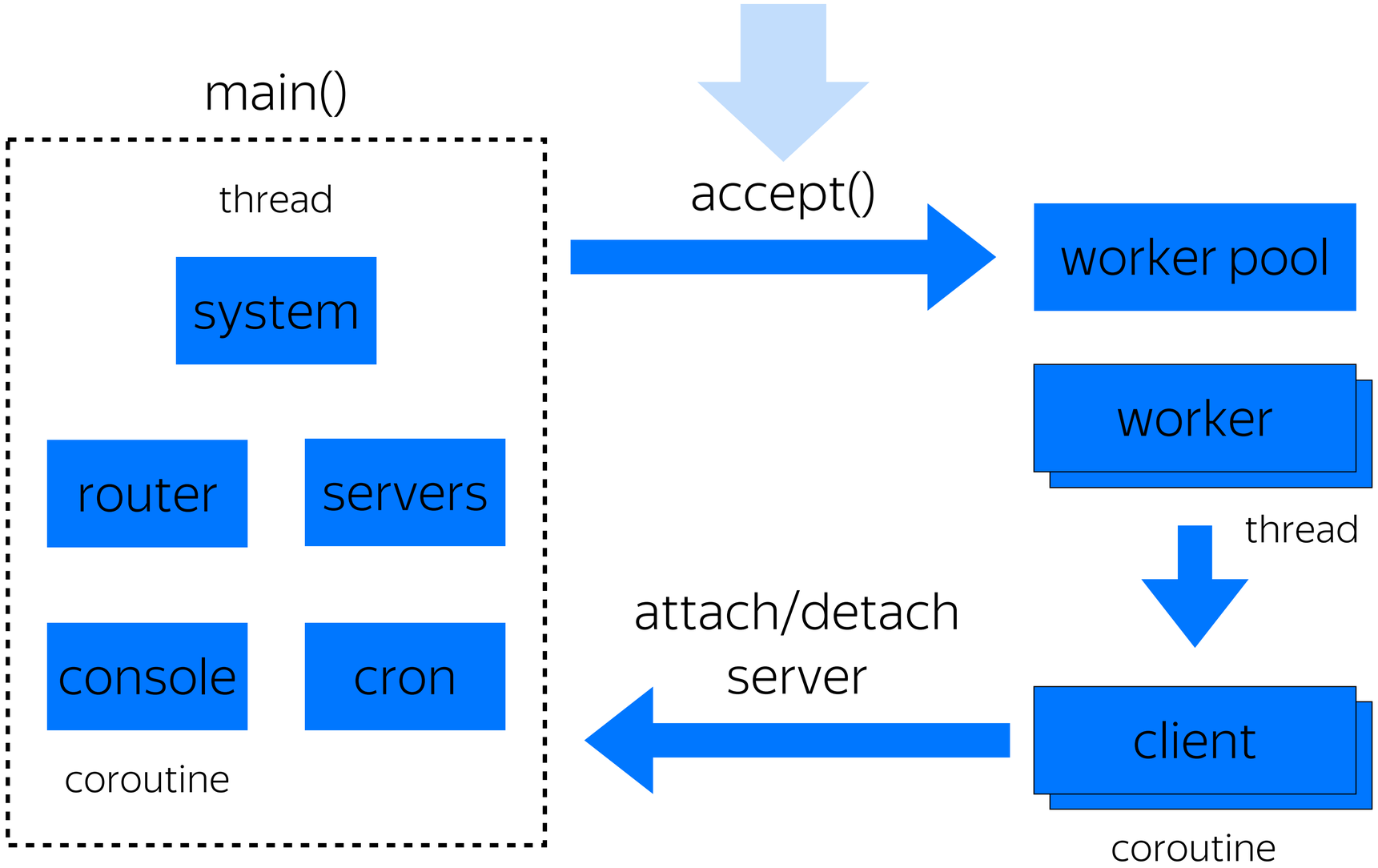
TLS

TLS




client.{crt,key}
TLS
server. {crt,key}
TLS

Торможение accept



GSSAPI

Add GSSAPI encryption support (Robbie Harwood, Stephen Frost)

This feature allows TCP/IP connections to be encrypted when using GSSAPI authentication, without having to set up a separate encryption facility such as SSL. In support of this, add hostgssenc and hostnogssenc record types in pg_hba.conf for selecting connections that do or do not use GSSAPI encryption, corresponding to the existing hostssl and hostnossl record types. There is also a new gssencmode libpq option, and a pg_stat_gssapi system view.

GSSAPI

Add GSSAPI encryption support (Robbie Harwood, Stephen Frost)

This feature allows TCP/IP connections to be encrypted when using GSSAPI authentication, without having to set up a separate encryption facility such as SSL. In support of this, add hostgssenc and hostnogssenc record types in pg_hba.conf for selecting connections that do or do not use GSSAPI encryption, corresponding to the existing hostssl and hostnossl record types. There is also a new gssencmode libpq option, and a pg_stat_gssapi system view.

Пулер должен понимать что за запрос к нему пришёл

Новое конфигурирование репликации

✓ **Enhance settings for replication support (#107)** — Browse files

master (#107)

efimkin authored and **x4m** committed 4 days ago — 1 parent f74abbf — commit 60d7229cdcbdf50968d1c3e5e83d0c8c047add5e

Showing **12 changed files** with **167 additions** and **32 deletions.** — Unified | Split

2 odyssey.conf

```
@@ -339,8 +339,6 @@ storage "postgres_server" {
339 339   #
340 340   #     "remote" - PostgreSQL server
341 341   #     "local"  - Odyssey (admin console)
342     - #     "replication" - physical replication, connects to walsender
343     - #     "replication_logical" - physical replication walsender and logical replication
344 342   #
345 343         type "remote"
346 344   #
```

Управление количеством серверов

Implement server_max_routing restriction

master (#104)

Browse files

x4m authored and **reshke** committed 8 days ago 1 parent 0bf792c commit 8496cde80aa292d0365c73315e2091469807c23e

Showing **7 changed files** with **51 additions** and **10 deletions.**

Unified | Split

9 ■■■■■ odyssey.conf

```
@@ -248,6 +248,15 @@ keepalive 7200
  #
  # client_max_routing 32
  
+ #
+ # Global limit of server connections concurrently being routed.
+ # We are opening no more than server_max_routing server connections concurrently.
+ # In future, this setting will be moved to storage section.
+ #
+ # Unset or zero 'server_max_routing' will set it's value equal to 2 * workers
+ #
+ # server_max_routing 4
+ 
  ###
  ### LISTEN
  ###
```

Управление количеством серверов

Это задача для ML!

Управление клиентскими соединениями

✓ **Implement login timeout**

master (#81) 1.0

Browse files

x4m authored and **reshke** committed on 1 Dec 2019 — 1 parent f43ac71 — commit eb7299748d55b28703a3f53753b3064789d656da

Showing 17 changed files with **37 additions** and **24 deletions**.

Unified | Split

4 odyssey.conf

```
@@ -292,6 +292,10 @@ listen {
292 292   #     tls_key_file ""
293 293   #     tls_cert_file ""
294 294   #     tls_protocols ""
    295 +
    296 + #   client_login_timeout
    297 + #   Prevent client stall during routing for more that client_login_timeout milliseconds.
    298 + #   Defaults to 15000.
295 299   }
296 300
297 301   ###
```

Управление клиентскими соединениями

✓ Implement client_max_routing (#72)

master (#72) 1.0

Browse files

x4m committed on 15 Oct 2019 Verified 1 parent d5ba841 commit 298cdca895e7e69d3640f242d0ed08e3a220d261

Showing 8 changed files with 50 additions and 0 deletions. Unified Split

10 odyssey.conf

```
@@ -238,6 +238,16 @@ keepalive 7200
 #
 # client_max 100
 
+ #
+ # Global limit of client connections concurrently being routed.
+ # Client connection is being routed after it is accepted and until it's startup
+ # message is read and connection is assigned route to the database. Most of the
+ # routing time is occupied with TLS handshake.
+ #
+ # Unset or zero 'client_max_routing' will set it's value equal to 4 * workers
+ #
+ # client_max_routing 32
+
 ###
 ### LISTEN
 ###
```

Настройка и мониторинг

Listen

```
listen {
    host "*"
    port 6432
    backlog 128
    tls "require"
    client_login_timeout
    server_login_retry
}
```

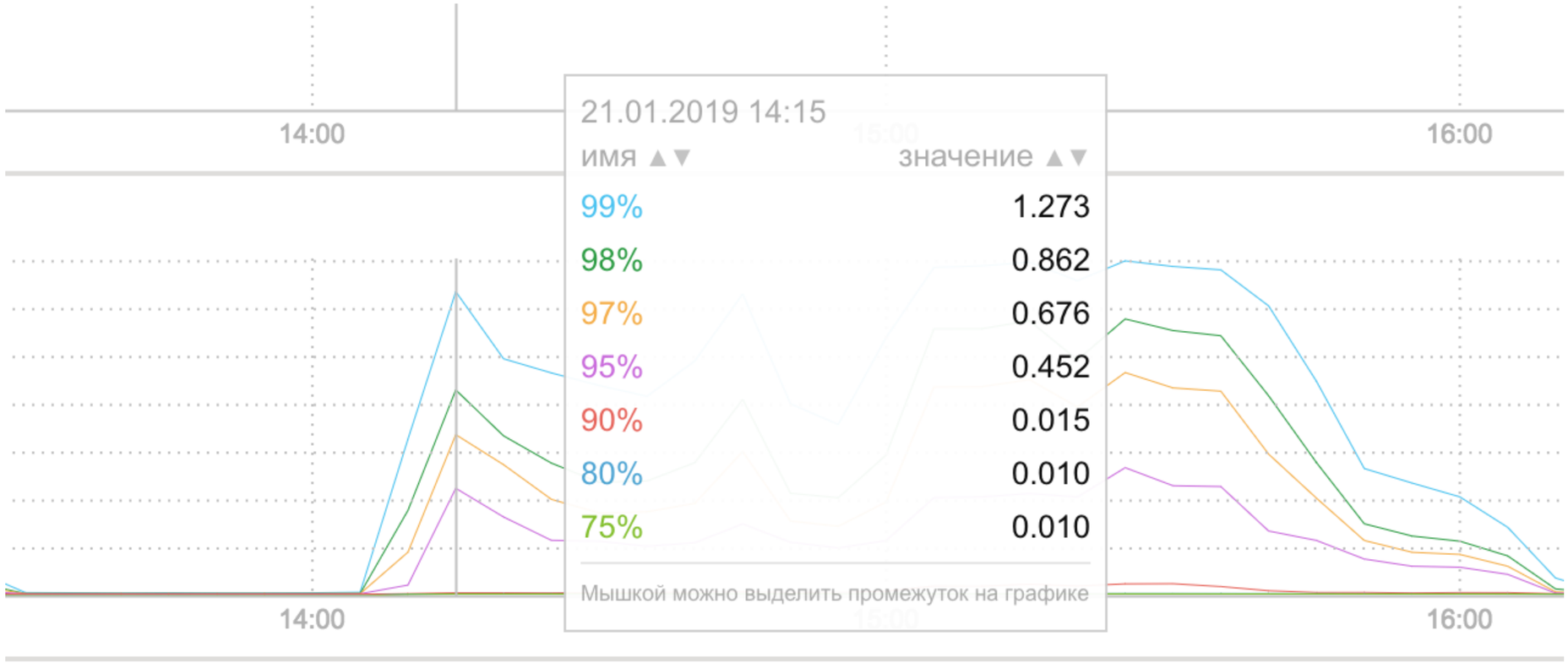
Storage

```
storage "postgres_server" {
#   "remote" - PostgreSQL server
#   "local"  - Odyssey (admin console)
    type "remote"
    host "localhost"
    port 5432
    tls "disable"
    server_max_routing 4
}
```

Database + User

```
database default {
    user default {
        authentication "scram-sha-256"
        password ""
        client_max 100
        storage "postgres_server"
#       storage_db "database"
#       storage_user "test"
#       storage_password "test"
        pool "transaction"
        pool_size 0
        pool_ttl 60
        pool_discard no
        pool_cancel yes
        pool_rollback yes
        client_fwd_error yes
        application_name_add_host yes
        log_debug no
        quantiles "0.99,0.95,0.5"
    }
}
```

Квантили запросов



Мониторинг

- Utilization
- Saturation
- Errors

Автоматическое обновление

- В Одиссее нет Online restart
- И даже reload вызывает всплески под нагрузкой
- Обязательно мониторить все открытые интерфейсы

Сообщество

SCRAM аутентификация

Команды PAUSE\RESUME

Вопросы?

Яндекс Облако

Спасибо!

Андрей Бородин

Руководитель подразделения разработки РСУБД с открытым исходным кодом

x4mmm@yandex-team.ru

@x4mmm