# PostgreSQL
# avito.ru

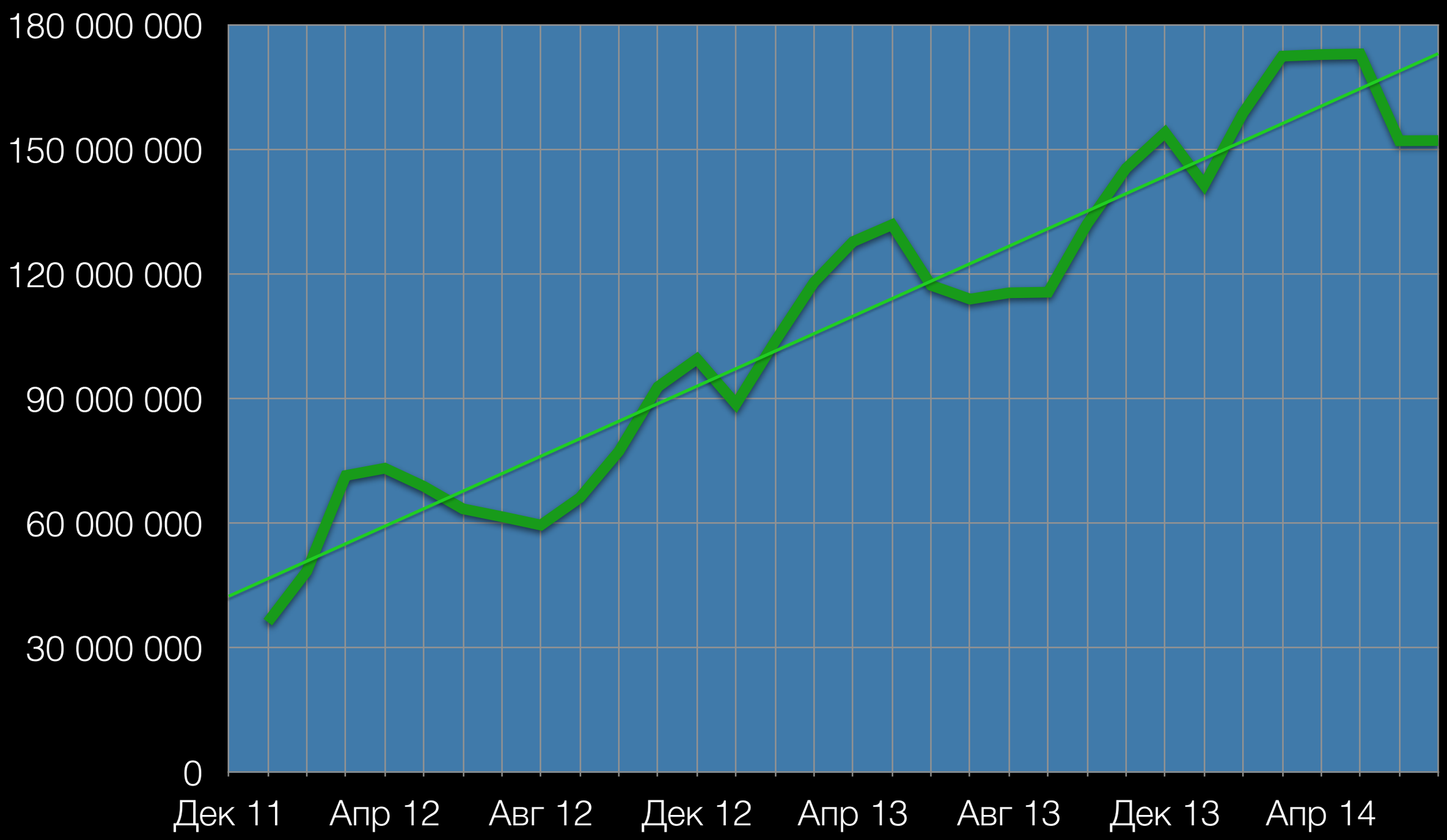
Михаил Тюрин
DBA, руководитель группы

pgday.ru
2014

# avito.ru

- самая большая доска объявлений в России (Европе)
- не только ru и не только сайт
- быстрый рост: до 250M hits/day за 5 лет
- до 1M новых объявлений в день
- ежемесячная аудитории 50M (5M/day)

# pageviews


180 000 000
150 000 000
120 000 000
90 000 000
60 000 000
30 000 000
0
Дек 11
Апр 12
Авг 12
Дек 12
Апр 13
Авг 13
Дек 13
Апр 14

# о чем доклад

- ретроспектива
- место в архитектуре
- параметры эксплуатации
- ключевые решения
- команда разработки
- текущие вызовы

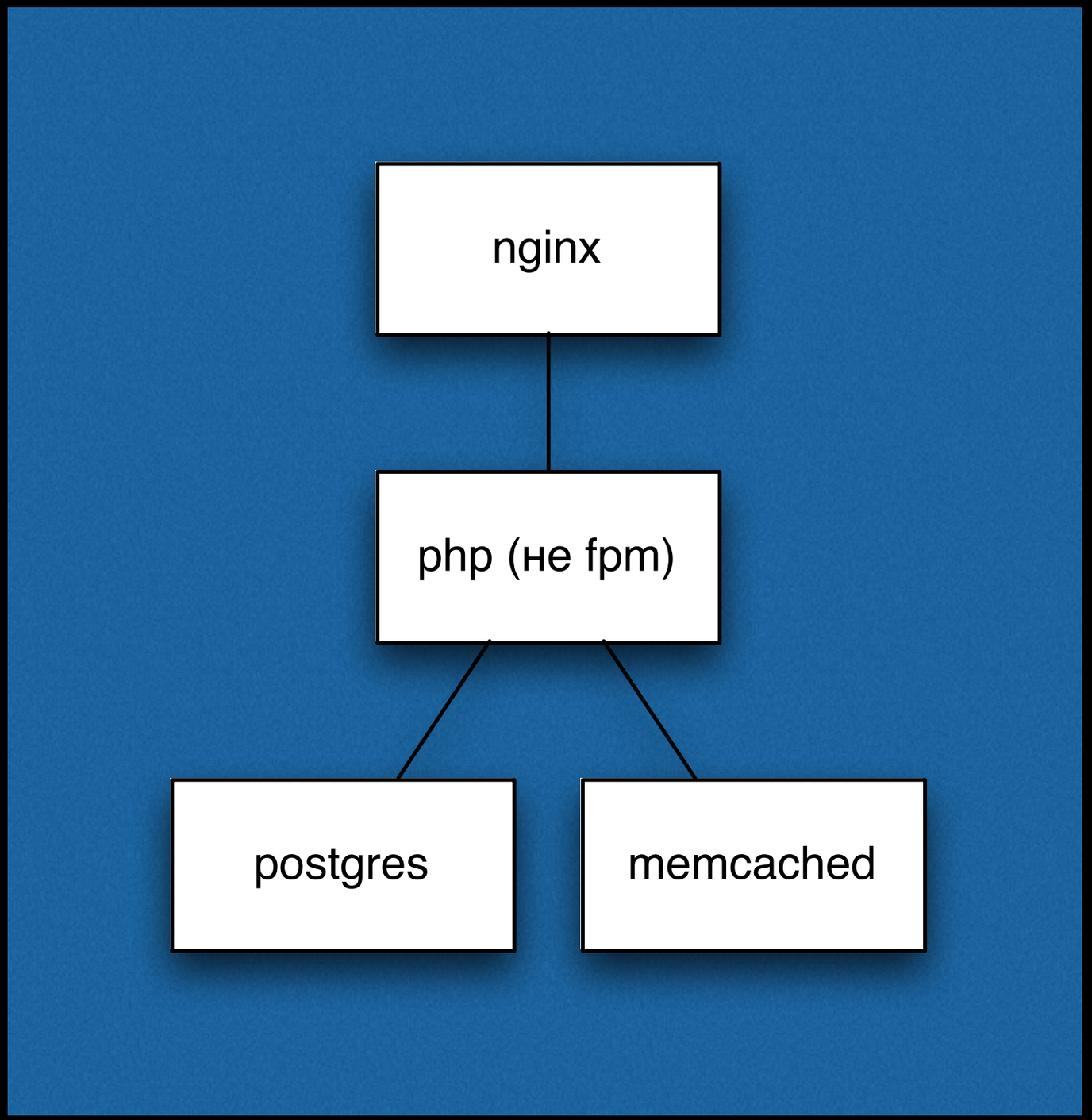
# начало

- сайт ru
- web backoffice
- ~10 человек вся компания // включая всех

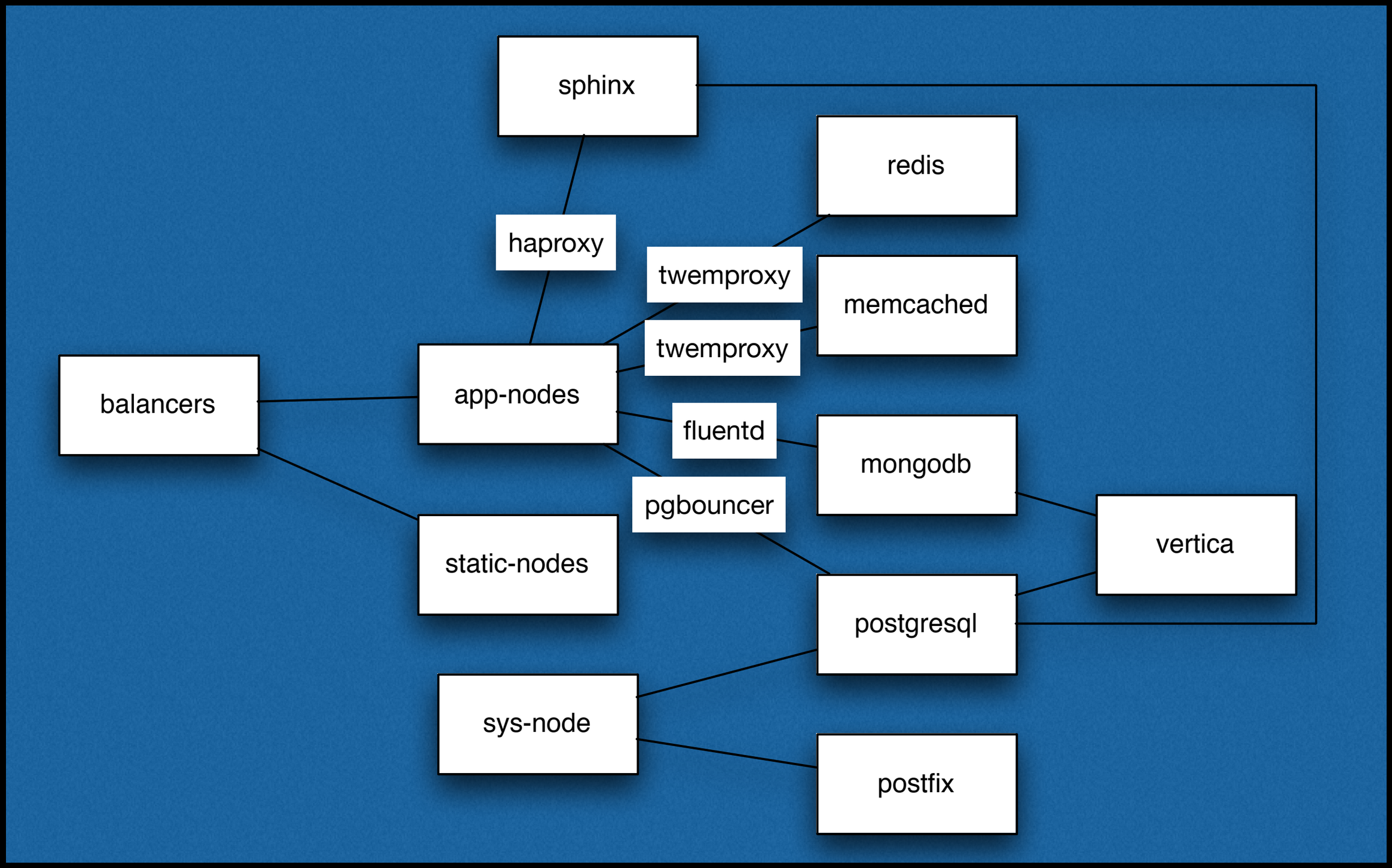
# а сейчас

- сайты: Россия, Украина, Марокко, Египет
- backoffice: support, moderation, bi, …
- ! mobile: web, iOS, android, WP
- api: mobile apps, partners, …
- 500+ человек // кто все эти люди?!

# как было



# как получилось



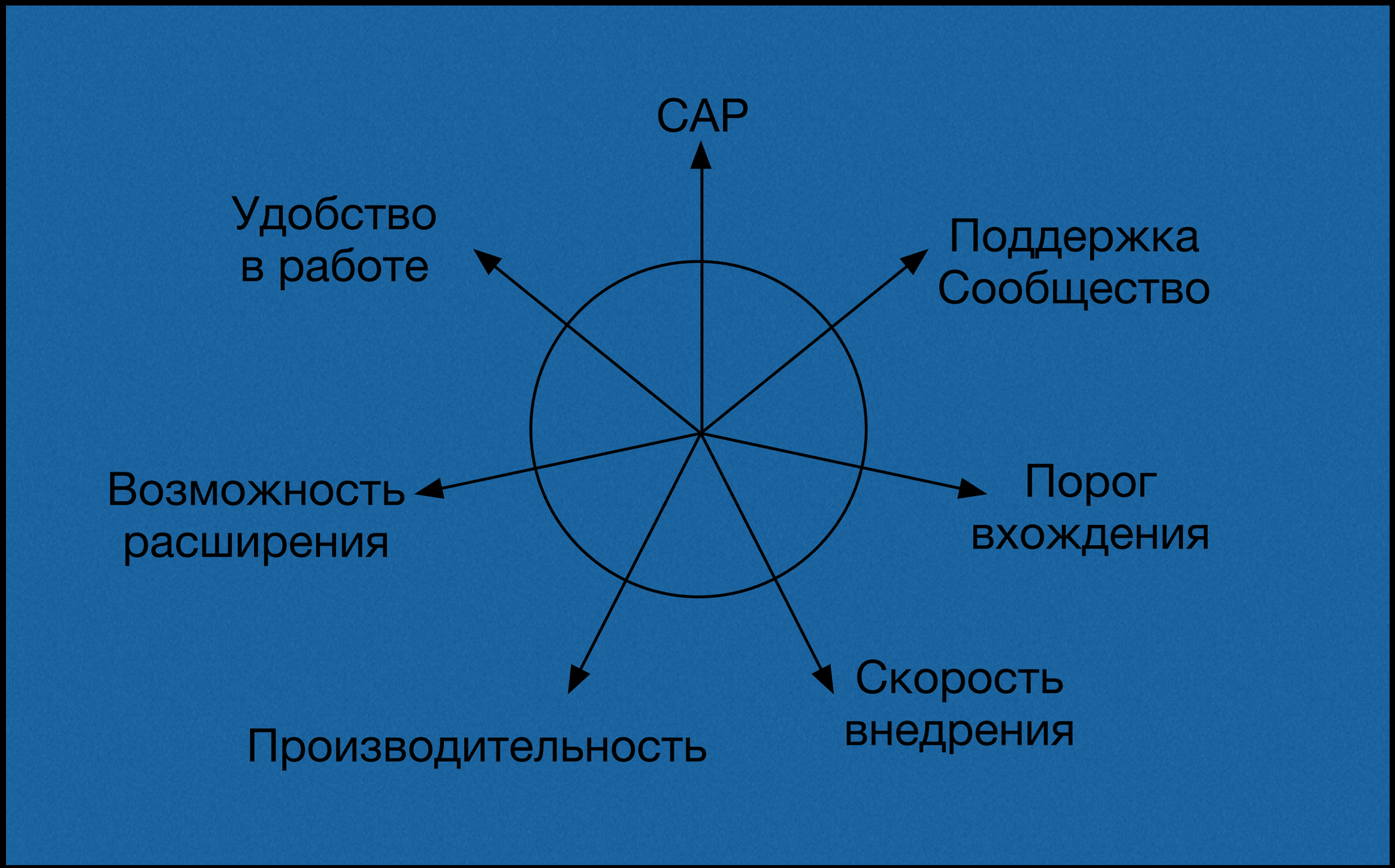
# еще раз

- app
- sys
- sphinx
- vertica
- postgres
- redis — дампится в базу

# php

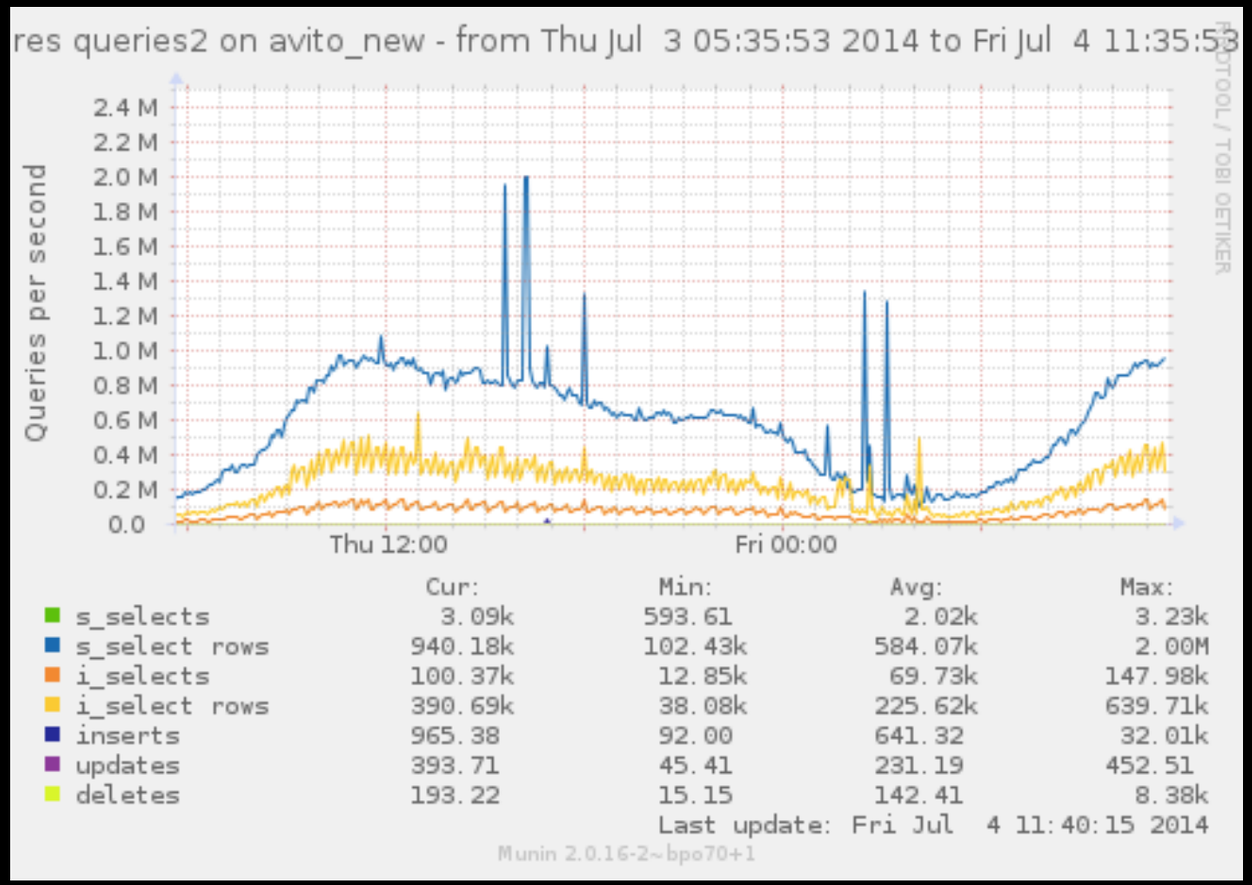
- pgbouncer в транзакшин режиме
- pdo с патчем
- свой класс
- ленивые коннекты и транзакция
- управление транзакциями
- OnCommit

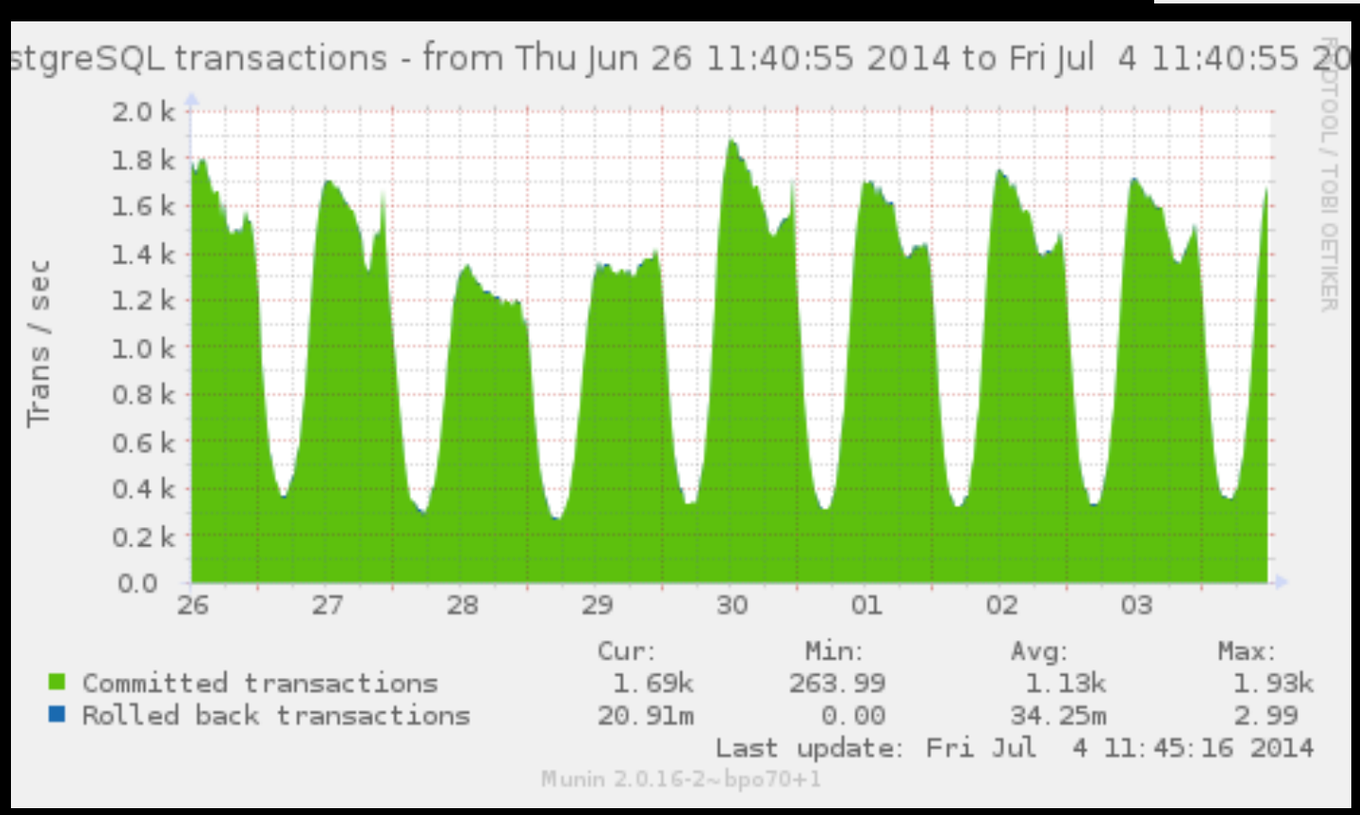
# как такое получилось



# хайлоад 1

- IUD мастера 1+K/s

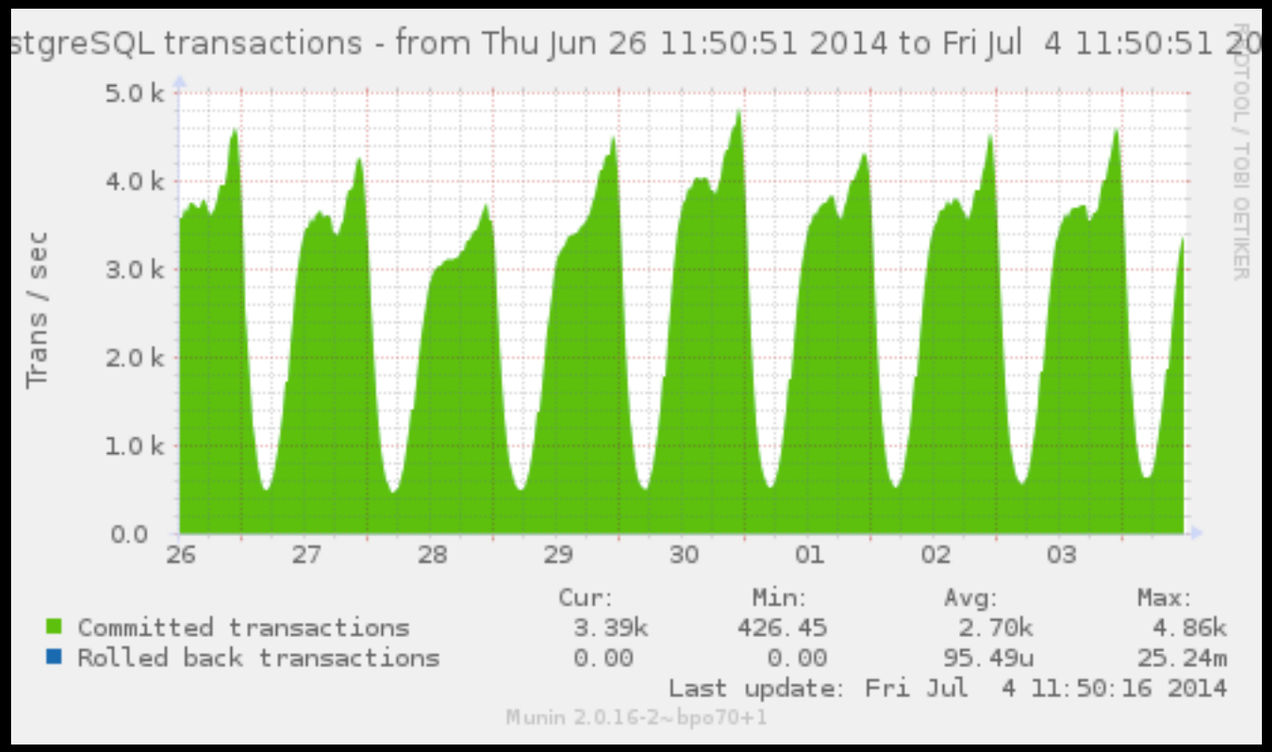


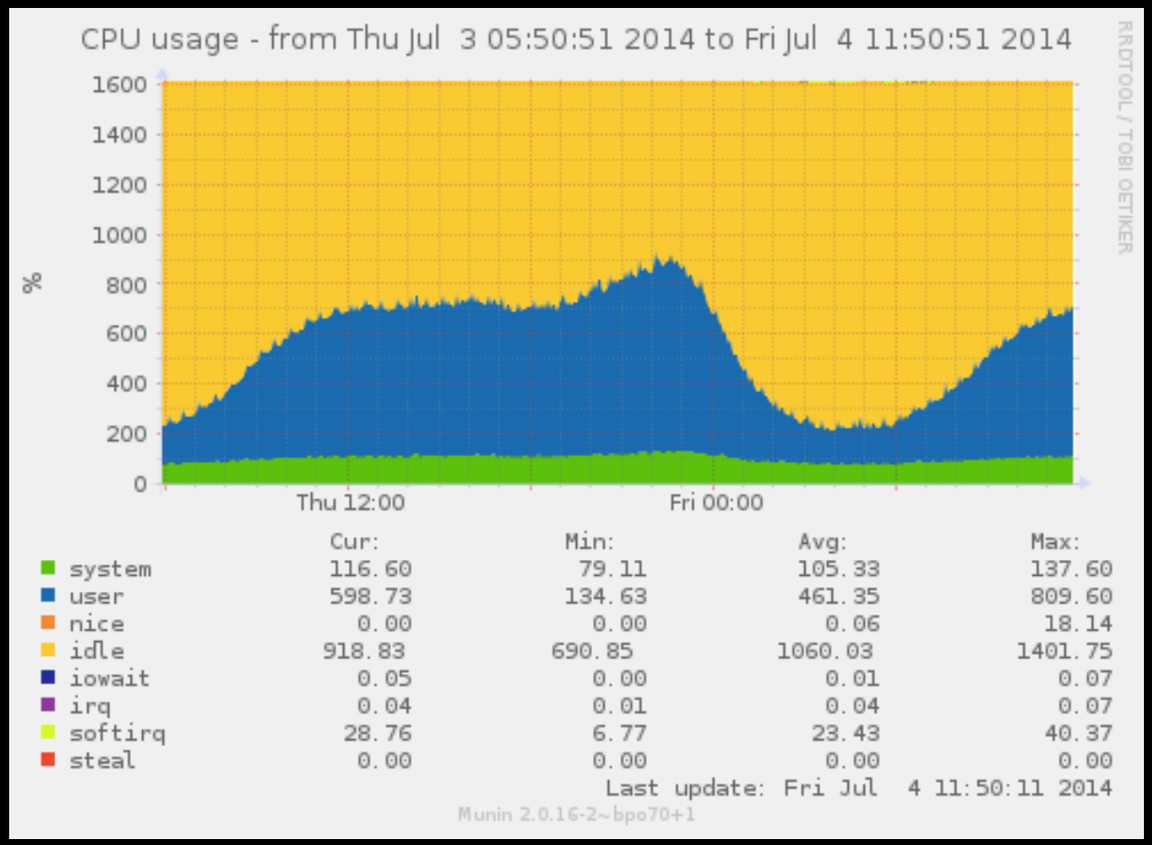



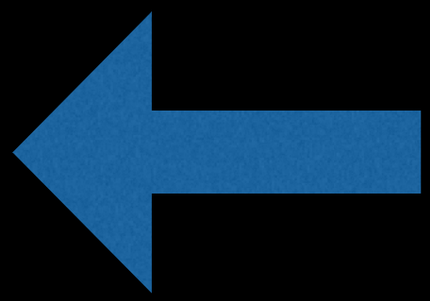

- transactions master 2K/s

# хайлоад 2

- transactions repca 5K/s
- cpu - 6-ядер

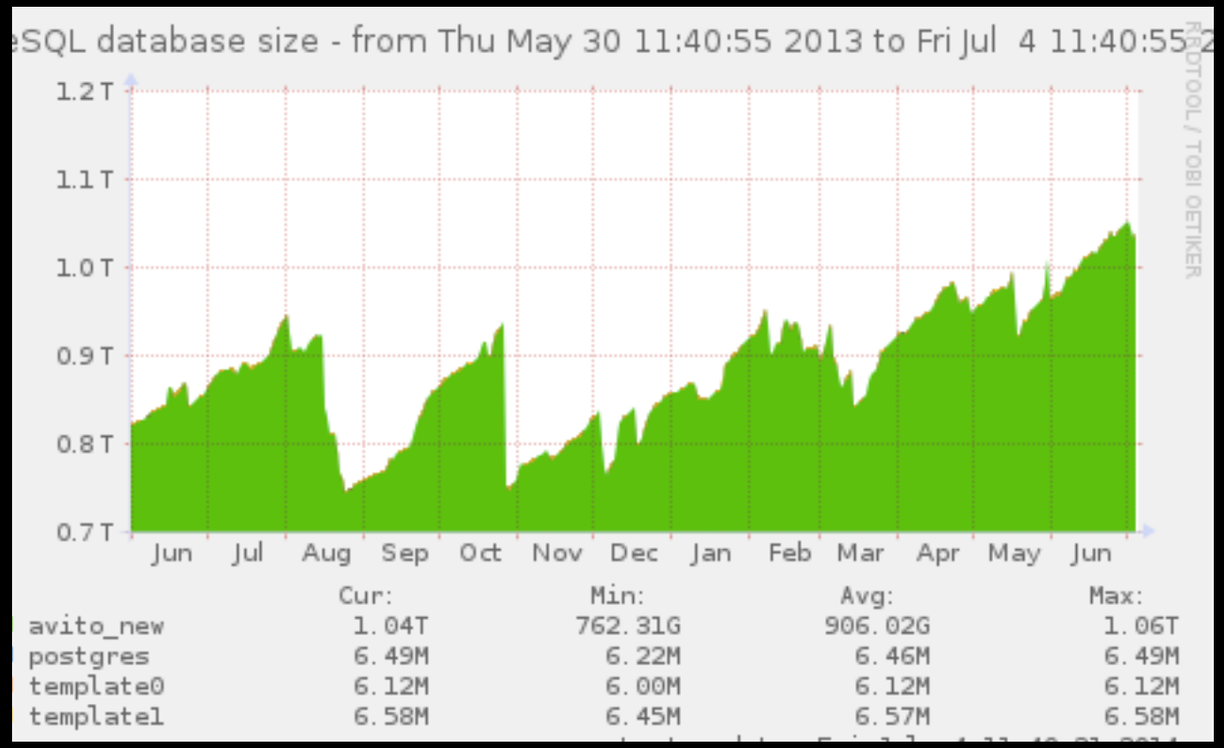


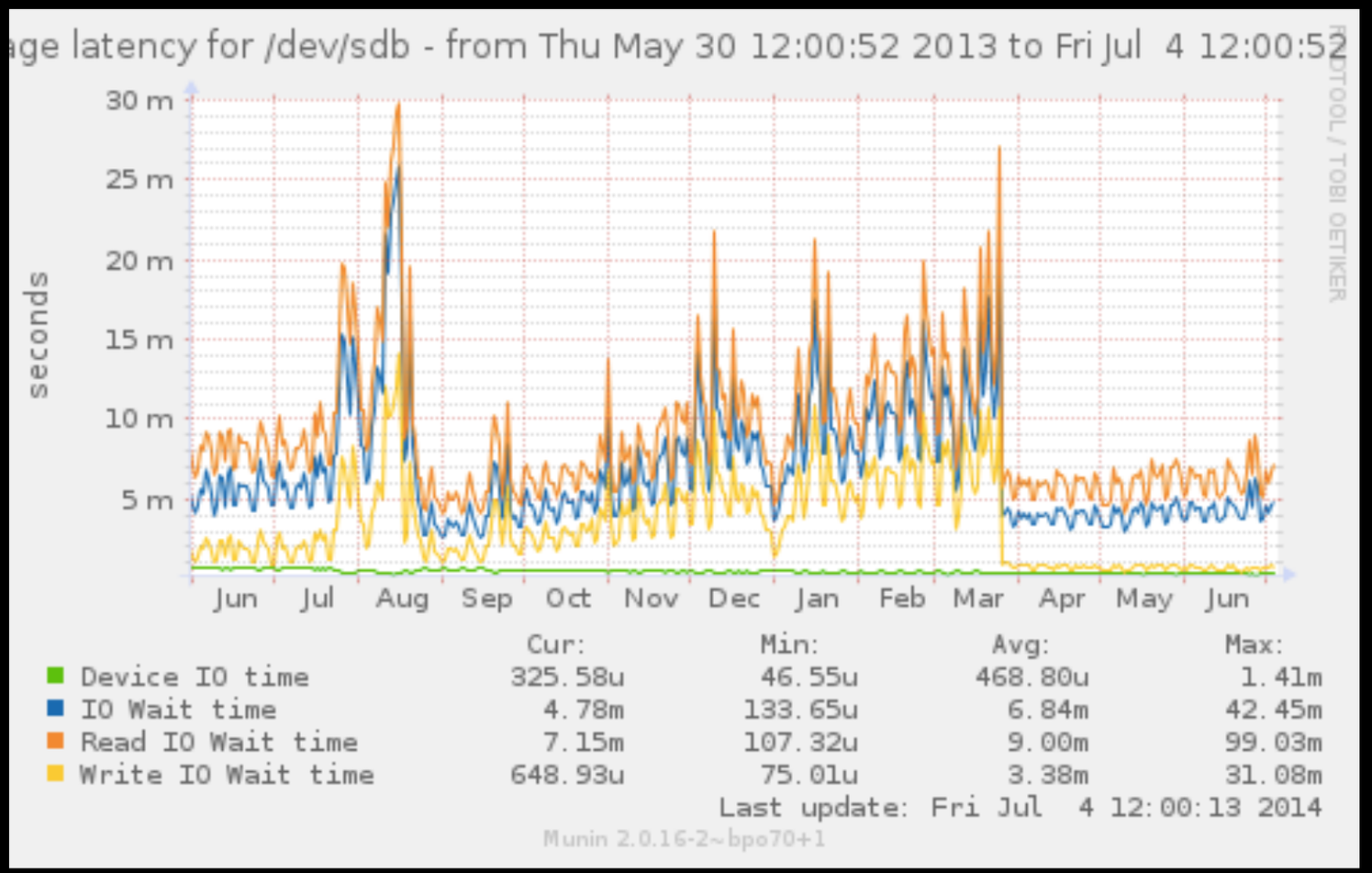



# хайлоад 3

- размер мастер 1+TB
- латенси мастера - OK
- запись на мастер - avg 40MB/s
- iops - 2-20K



# среда

- ( php + pdo => pgbouncer ) ==> ( pgbouncer => postgres )
- ! postgres **92** // pg_upgrade: 83 > 84 > 90 >> 92 >> ...
- >> ? 94/95: fdw/json
- modules: **hstore**, intarray, pg_buffercahce, dblink, auto_explaine, plproxy, pg_stat_statements, pgstattuple
- **plpgsql**/sql: бизнес логика
- plperlu: csv, files; plpythonu: json, zip

# далее

- python + psycopg2; perl + DBI; bash + psql
- **skytools2**: londiste, pgq, php/python consumer, php deamon
- munin + собственные плагины («системные» / «продуктовые»)
- собственные скрипты
    - архив: pitr + pbzip2 (4 потока)
    - myvac, truncator, optimazer, rotator

# подсистемы

- сайт
    - мастер базы — синхронная запись (+очереди)
    - wal реплика основного мастера — чтение
        - pgq реплика «активных айтимов» — чтение + индексация
        - pgq платежная реплика — чтение
- хранилище (архив внутри) — 8 нод по 2 базы

# далее

- индексация сайта / индексация админки
- експорт в dwh
- xrpcd: денормализация, геокодер, …
- ! архив и резервирование ВСЕГО

# repca

- реплика активных айтимов
- матвью — активные айтимы
- londiste с мастера на реплику // PGQ!
- реплика в шаред буферах — OK
- резервирование — ! два коснумера на одну очредь (knowhow)

# repca — indexer

- sphinx
- ! pause replication
- 10 indexers
- uftp + rsync
- monitoring
- тоже два

# archive + (!hot)standby

- pitr
- в nfs по ssh через standby
- сжатие валов
- pg_basebackup
- 4 дня глубина
- pg_archivecleanup

# linux + hardware

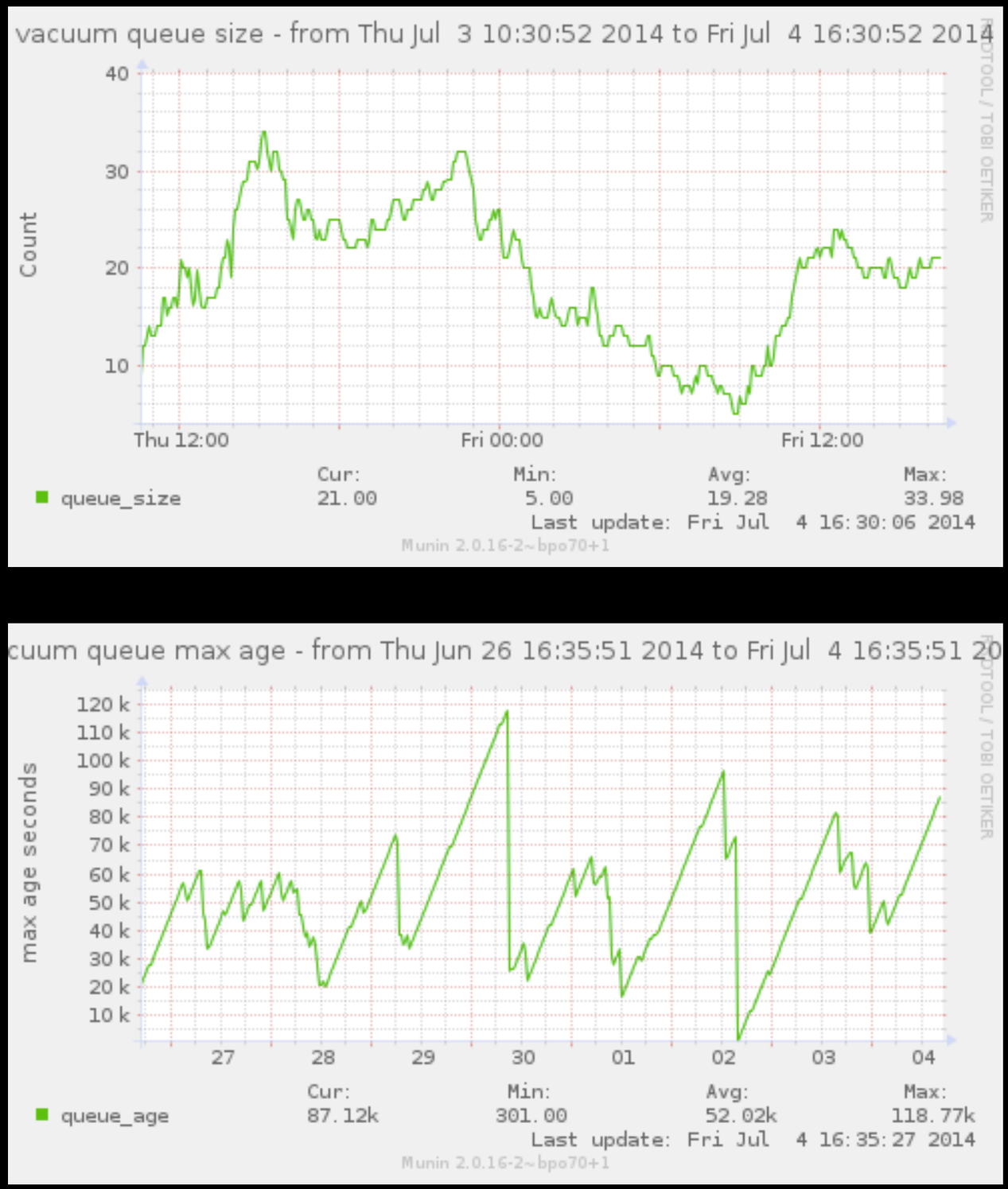
- linux debian 2.6.32 - 3.2. …
- raid10 + bbu
- numa off + interleaving on
- swap off
- ext4: noatime, wo barriers
- small read ahead
- noop планировщик io

# postgresql.conf

- shared buffers / work men / effective cache size
- max connections
- checkpoint / bgwriter
- autovacuum !!! on
- sync commit off
- log collector to tmpfs / what to log
- stat direcory to tmpfs
- ! fsync / tablespace => tmpfs — во дают ребята

# VACUUM

- ON
- наша новая мерилка



# команда

- 7+ разработчиков  + 0,5 админа
- тимлид
- dba + sql
- system adm + monitoring
- php
- python
- bash + perl
- pappet

# вызовы: архив

- база растет просто от времени
- это поражает воображение
- надо иметь архив
- есть варианты
- вопрос согласованности

# вызовы: вызов приложения из базы

- движок «асинхронной» обработки
- примеры: инвалидация кеша / апдейт счетчиков / продуктовые логи
- очереди-транспорт
- pgq php consumer
- можно «оживить web»

# вызовы: dev db

- текущая схема
    - бекап на дев сервере
    - ежедневные дампы схем и справочников
        - выкатка на живой базе — опасно
        - но очень просто
- растет и команда
- контроль изменений
- цель: выкатка базы тоже по кнопке
- сложно: нужен второй авито

# SSD

- уже не дорого
- много места и
- x20 к read iops (вау!) // запись точно не хуже

*Sergey Burladyan added a comment - 10/июн/14 20:59*

*сравнение с HDD, random read:*

***HDD: 54 000 блоков: 76 секунд 0.00141***

***SDD: 100 000 блоков: 8 секунд 0.00008***

- внедряем — уже пол года проигрывает валы и гоняет запросы
- trim — не трогать дефолты

# вместо заключения

- сделали на postgres топ5 сайт
- после яндекс-вк-ок-мейл
- суп из топора

> *«Набор фич и расширений, легендарная надёжность PostgreSQL, наличие встроенной репликации, средств резервирования и архивирования — весь потенциал нашел своё воплощение, а наличие открытого профессионального комьюнити не оставляет шансов к неэффективной реализации.»*

Спасибо!

**Вопросы**

**mtyurin@avito.ru**