

Яндекс



История успеха Яндекс.Почты

Владимир Бородин

О Яндекс.Почте

- › Запущена в 2000
- › 10+ миллионов пользователей в сутки
- › 200.000 RPS в бэкенды web/mobile/imap
- › 150+ миллионов писем покладки в сутки
- › 20+ ПБ данных

Об этом рассказе

- › Миграция из Oracle в PostgreSQL
- › 300+ ТБ метаданных без избыточности
- › 250к запросов в секунду
- › OLTP: 80% чтений, 20% записи

Предыдущие попытки

- › MySQL
- › Самописное решение

О почтовых метаданных

The screenshot shows a Yandex Mail interface. The left sidebar contains a list of folders: **Inbox** (with a red arrow pointing to it), Hive, Notes, pgpool, PostgreSQL (3 / 54478), RabbitMQ, Sent Messages, Archive, Sent, Trash, Spam, and Drafts. Below these are filters: Flagged, Unread, Attachments (with a red arrow pointing to it), xakep.ru, and Important (with a red arrow pointing to it). The main inbox area shows a list of emails with columns for sender, subject, preview, and time. A red arrow points to the subject line of the second email: "Re: [HACKERS] Does Type Have = Operator?".

Sender	Subject	Preview	Time
Shawn	Re: [HACKERS] Re: Need help debugging why autovacuum seems "stuck" -- until I use superuser to vacuum freeze pg_database		20:24
David E. Wheeler	Re: [HACKERS] Does Type Have = Operator?	Oh, well crap. Maybe I'd be better off just comparing the plain text of the expressions as...	20:23
Josh berkus	Re: [HACKERS] Academic help for Postgres	Together with that, automated substitution of materialized views for query clauses. Also: opt...	19:58
Robert Haas	Re: [HACKERS] asynchronous and vectorized execution	On Wed, May 11, 2016 at 12:30 PM, Andres Freund <andres@anarazel.de> wr...	19:31
Alvaro Herrera	Re: [HACKERS] ALTER TABLE lock downgrades have broken pg_upgrade	Peter Eisentraut wrote: True. We have quite a few places in t...	19:09
Mike Broers	Re: [ADMIN] driving postgres to achieve benchmark results similar to bonnie++	Ok so I ran 6 parallel pgbench initializations at a relatively...	18:47
Andres Freund	Re: [HACKERS] HeapTupleSatisfiesToast() busted? (was atomic pin/unpin causing errors)	Same issue. If the dead tuple is noticed by he...	18:27
Jim Nasby	Re: [HACKERS] Add jsonb_compact(...) for whitespace-free jsonb to text	On 4/29/16 8:56 AM, Shulgin, Oleksandr wrote: +1. I've found t...	16:40
Ondřej Světlík	Re: [ADMIN] Autovacuum of pg_database	You are right, sorry I didn't mention it sooner. With regards Ondřej	15:07
Martín Marqués	[HACKERS] Minor documentation patch	Hi, Yesterday I was going over some consultancy and went to check some syntax for CREATE FU...	14:00
Ashutosh Bapat	Re: [HACKERS] Use %u to print user mapping's umid and userid	On Wed, May 11, 2016 at 1:34 PM, Etsuro Fujita <fujita.etsuro@lab.ntt...>	12:04
Etsuro Fujita	Re: [HACKERS] Odd oid-system-column handling in postgres_fdw	I'll add this to the next CF. Best regards, Etsuro Fujita	10:47
Etsuro Fujita	Re: [HACKERS] Optimization for updating foreign tables in Postgres FDW	Thanks for the review! I'll add this to the next CF. I think this sh...	10:31
Guillaume Lelarge	Re: [ADMIN] Major Version Upgradation in Replication Environment	Well, you still have to rsync the data directory (and all tablespaces' dir...	10:08
Marco Nietz	Re: [ADMIN] Memory and Swap	Linux tends to swap out to early with the default settings of swappiness, try to decrease it to 10 or 1 https:...	8:37
Noah Misch	Re: [HACKERS] what to revert	I discourage focusing on the statistical significance, because the hypothesis in question ("Applying revert...	8:37

- Compose
- Check mail
- Reply
- Reply all
- Forward
- Delete
- Spam!
- Unsubscribe
- Unread
- Label
- To folder
- Pin
- Add button
- More

[HACKERS] what to revert



Noah Misch noah@leadboat.com

To you and 4:  Kevin Grittner

Cc:  Tom Lane  Tomas Vondra  Andres Freund  postgresql-hackers@postgresql.org

Show conversation

I discourage focusing on the statistical significance, because the hypothesis in question ("Applying revert.patch to 4bbc1a7e decreases 'pgbench -S -M prepared -j N -c N' tps by 0.46%.") is already an unreliable proxy for anything we care about. PostgreSQL performance variation due to incidental, ephemeral binary layout motion is roughly +/-5%. Assuming perfect confidence that 4bbc1a7e+revert.patch is 0.46% slower than 4bbc1a7e, the long-term effect of revert.patch could be anywhere from -5% to +4%.

If one wishes to make benchmark-driven decisions about single-digit performance changes, one must control for binary layout effects:
<http://www.postgresql.org/message-id/87vbitb2zp.fsf@news-spur.riddles.org.uk>
<http://www.postgresql.org/message-id/20160416204452.GA1910190@tornado.leadboat.com>

nm

--



today at 8:37

RELATED MESSAGES

- Noah Misch 8:37
I discourage focusing on the statisti...
- Andres Freund 0:06
Hm. Could you change max_connecti...
- Kevin Grittner 0:03
On Tue, May 10, 2016 at 2:41 PM, Ke...
- Kevin Grittner 10 may
On Tue, May 10, 2016 at 11:13 AM, T...
- Tomas Vondra 10 may
Hi /usr/src/postgresql/.../dist/...



ATTACHMENTS

LINKS

MESSAGES FROM NOAH MISCH

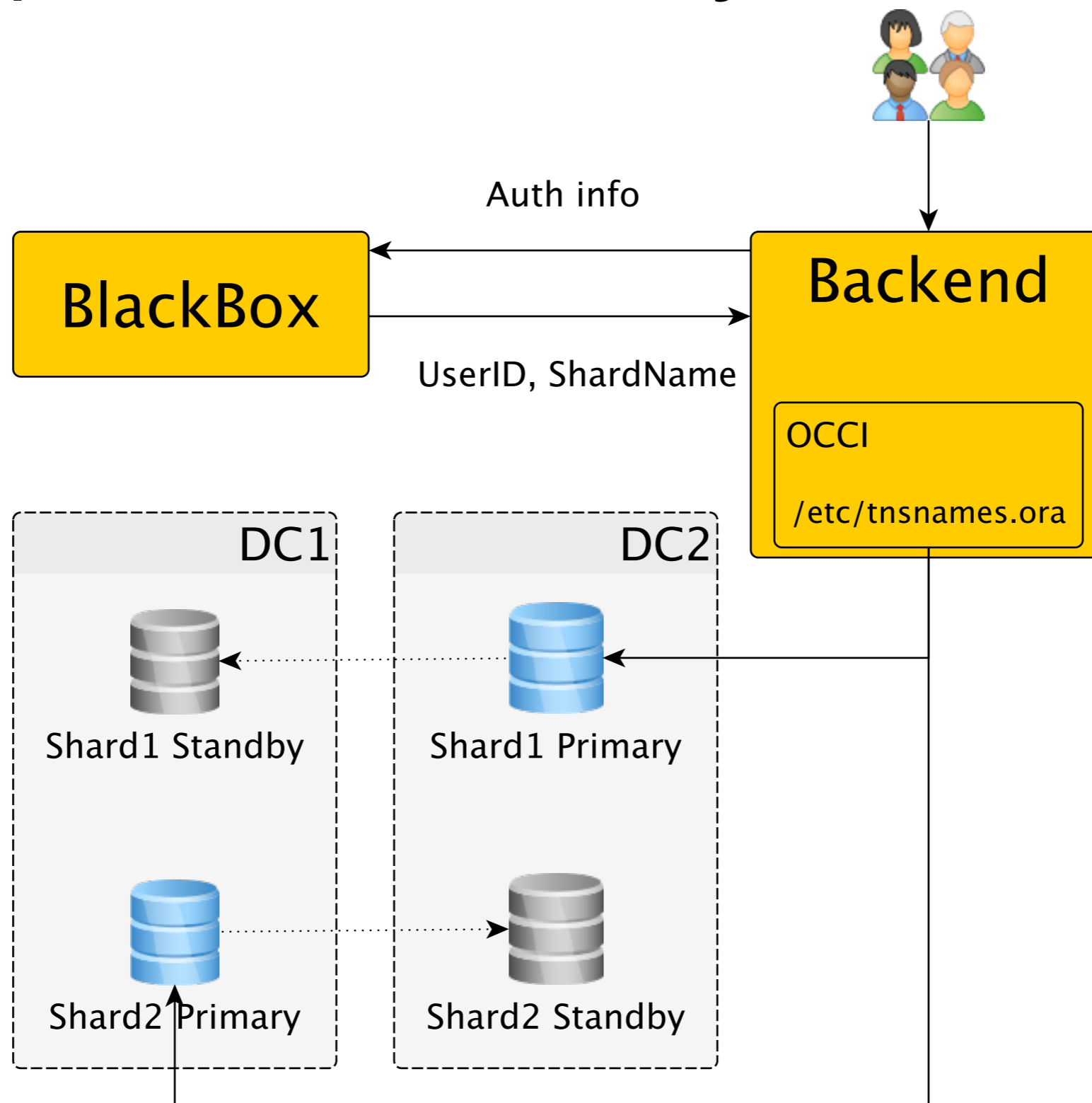
Назад в 2012



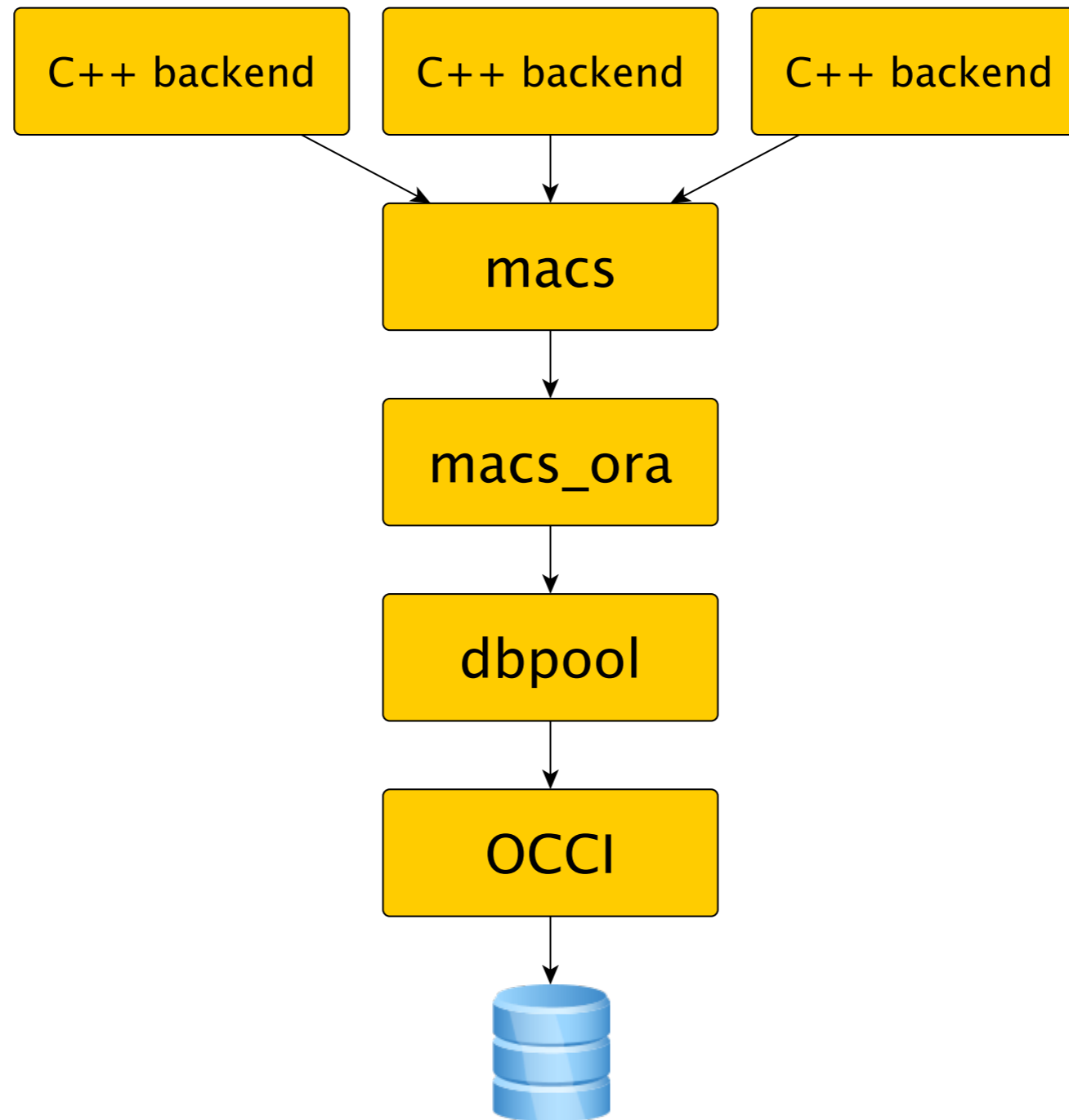
Метаданные Яндекс.Почты

- › Всё хранилось в Oracle
- › Очень много PL/SQL логики
- › Эффективная утилизация аппаратного обеспечения
 - 10+ ТБ данных на шард
 - Рабочий LA 100
- › Много ручных операций
- › Тёплые (SSD) и холодные (SATA) базы для разных пользователей
 - 75% SSD, 25% SATA

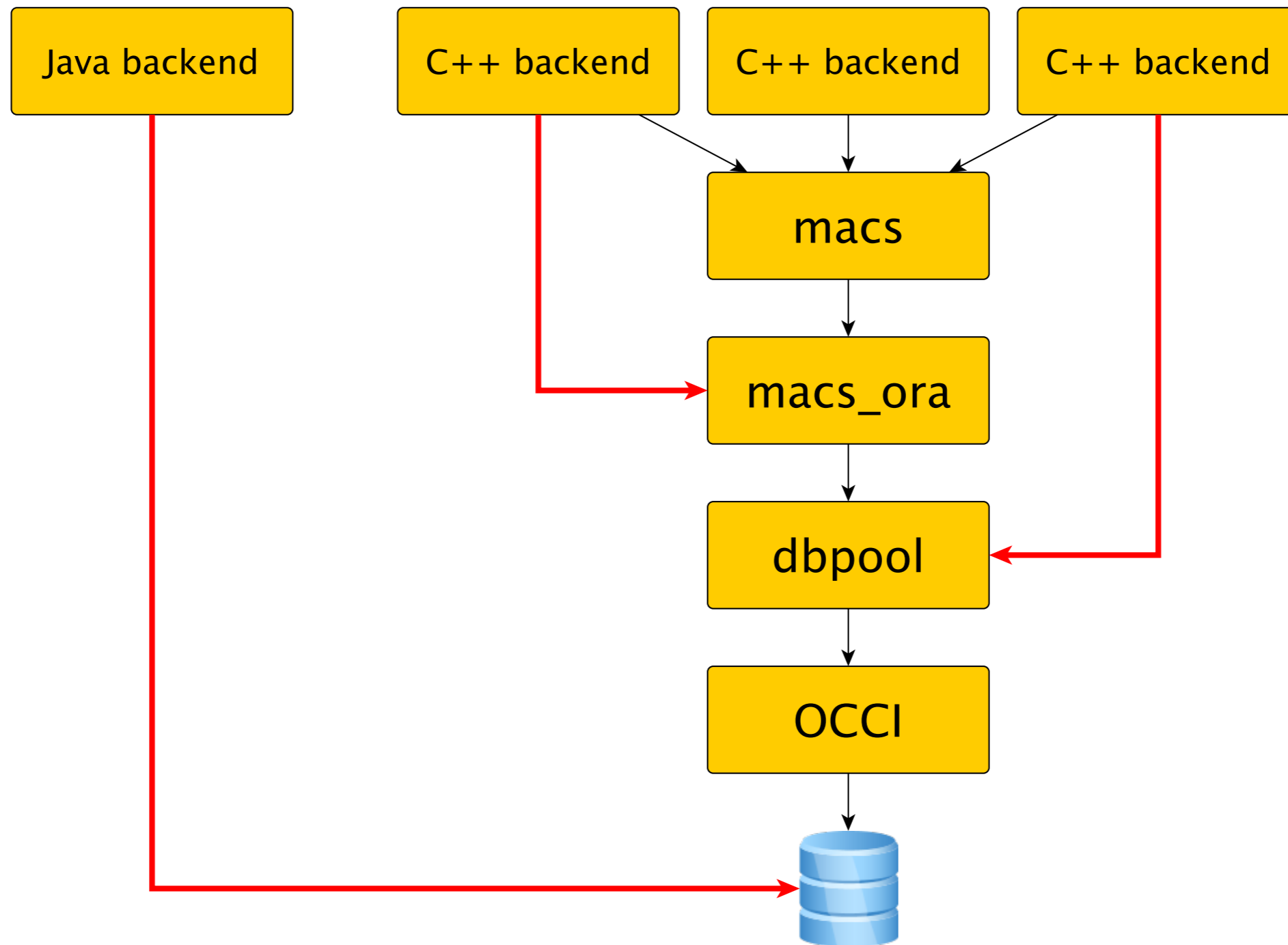
Шардирование и отказоустойчивость



Внутри приложения



Реальность



Наиболее популярные проблемы

- › PL/SQL deploy

 - Library cache

- › Множество ручных операций

 - Переключение мастера, наливка новой базы, переносы данных

- › Только синхронный интерфейс в OCCI

- › Проблемы с разработческими окружениями

- › Не очень отзывчивая поддержка



shop.oracle.com

Главная причина

Хронология



Эксперименты

- › Октябрь 2012 — политическое решение
 - Избавиться от Oracle за 3 года
- › Апрель 2013 — первые эксперименты с разными СУБД
 - PostgreSQL
 - Множество NoSQL решений
 - Самописное решение на основе поискового бэкенды
- › Июль 2013 — июнь 2014 — эксперимент со сборщиками
 - <https://simply.name/ru/video-pg-meetup-yandex.html>

Прототип всей почты

› Август 2014 — декабрь 2014

› Прокладка всего боевого потока писем в PostgreSQL

Асинхронно

› Первоначальные решения со схемой данных

Важно для слоя абстракции

› Нагрузочное тестирование под живой нагрузкой

Выбор аппаратного обеспечения

› Множество другого опыта работы с PostgreSQL

<https://simply.name/ru/postgresql-and-systemtap.html>

Основная работа

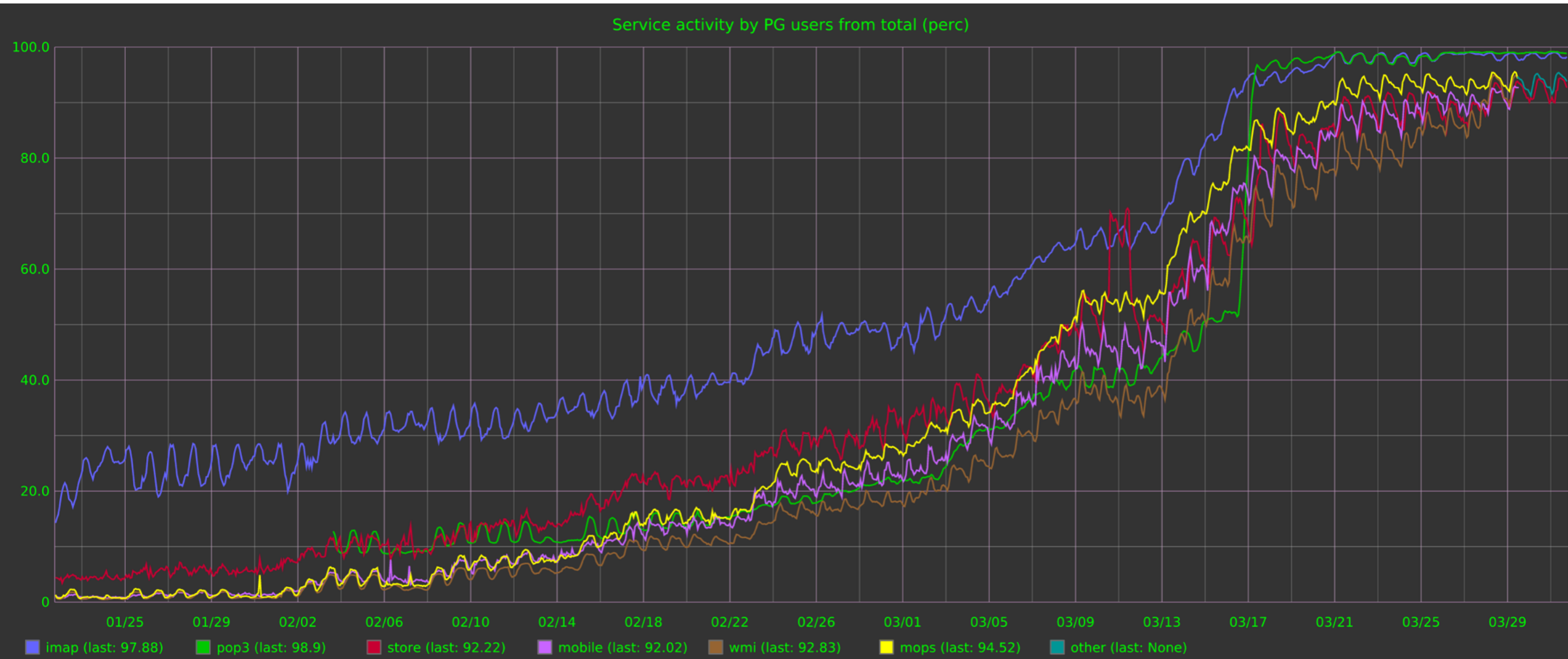
- › Январь 2015 — январь 2016 — разработка
- › Июнь 2015 — dog fooding
 - Ускорение разработки
- › Сентябрь 2015 — начало миграции неактивных пользователей
 - Исправление ошибок в коде переноса
 - Обратный перенос (план Б)
- › Январь 2016 — апрель 2016 — миграция



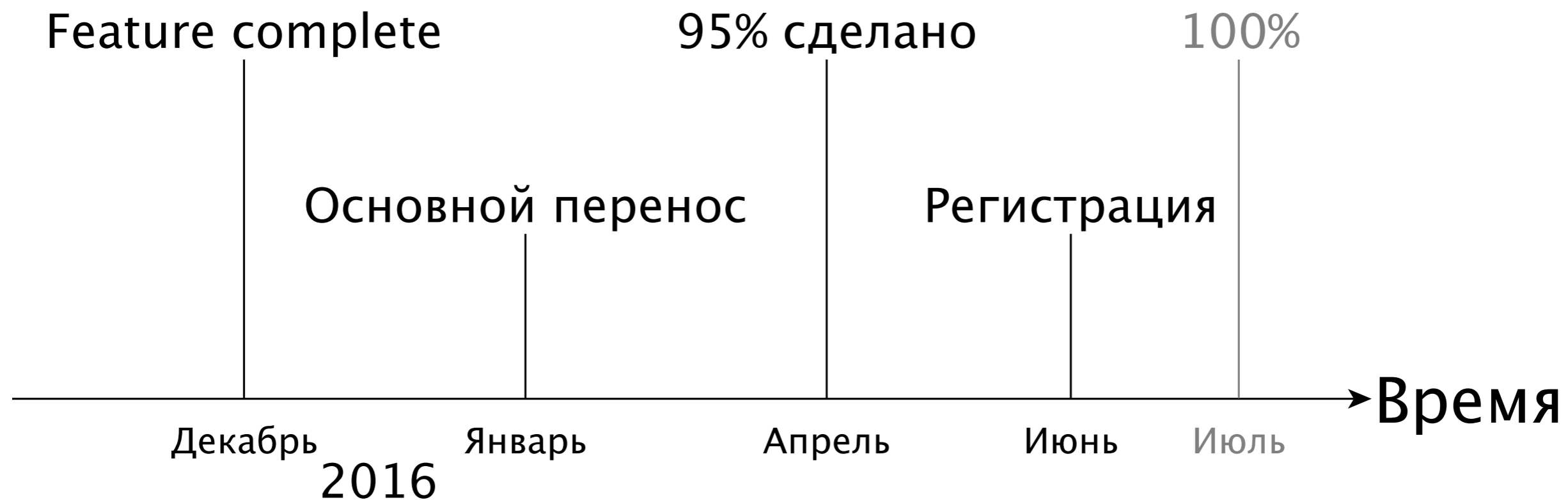
10 человеколет

Переписывание всего кода для работы с Oracle и PostgreSQL

Миграция



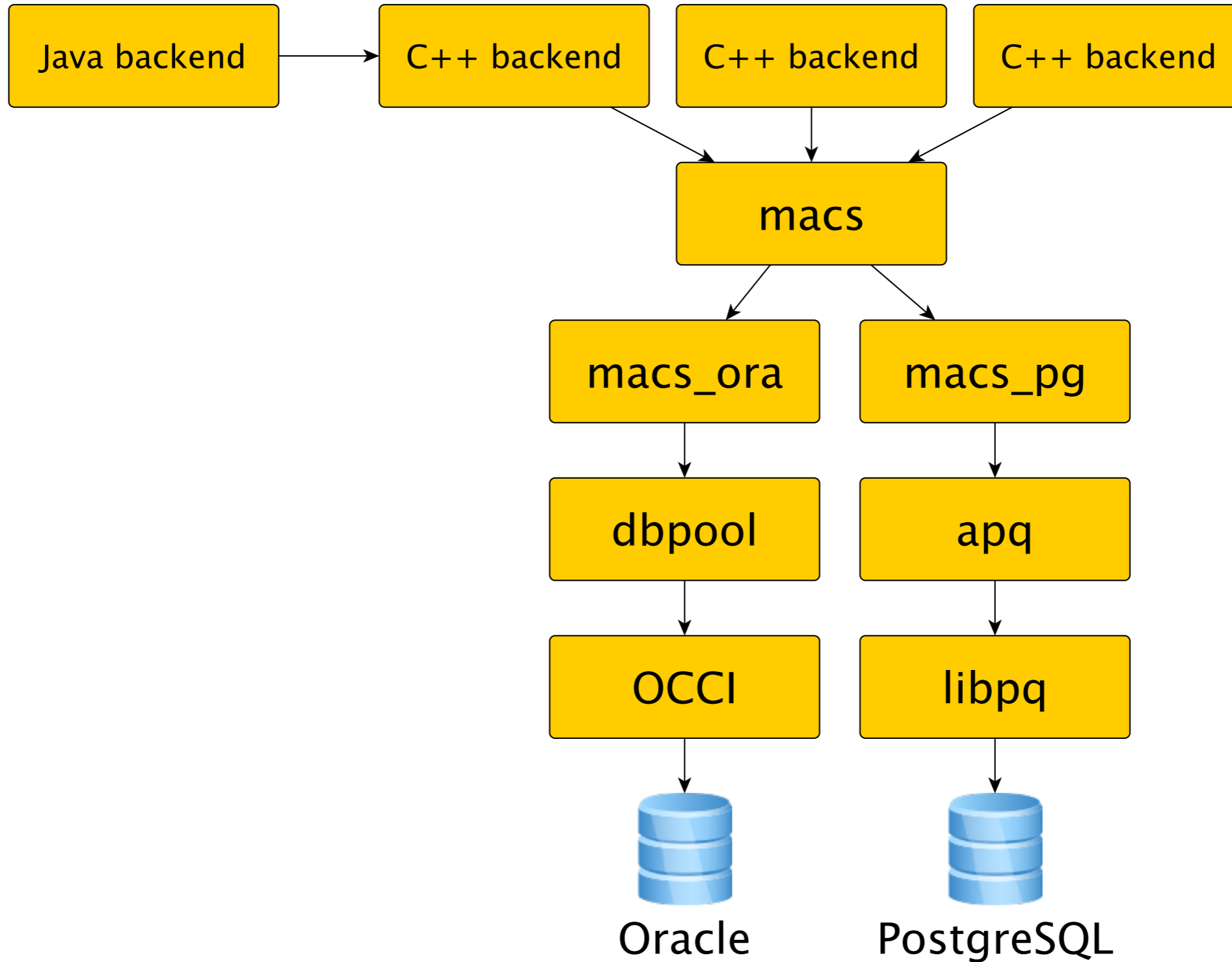
Завершение



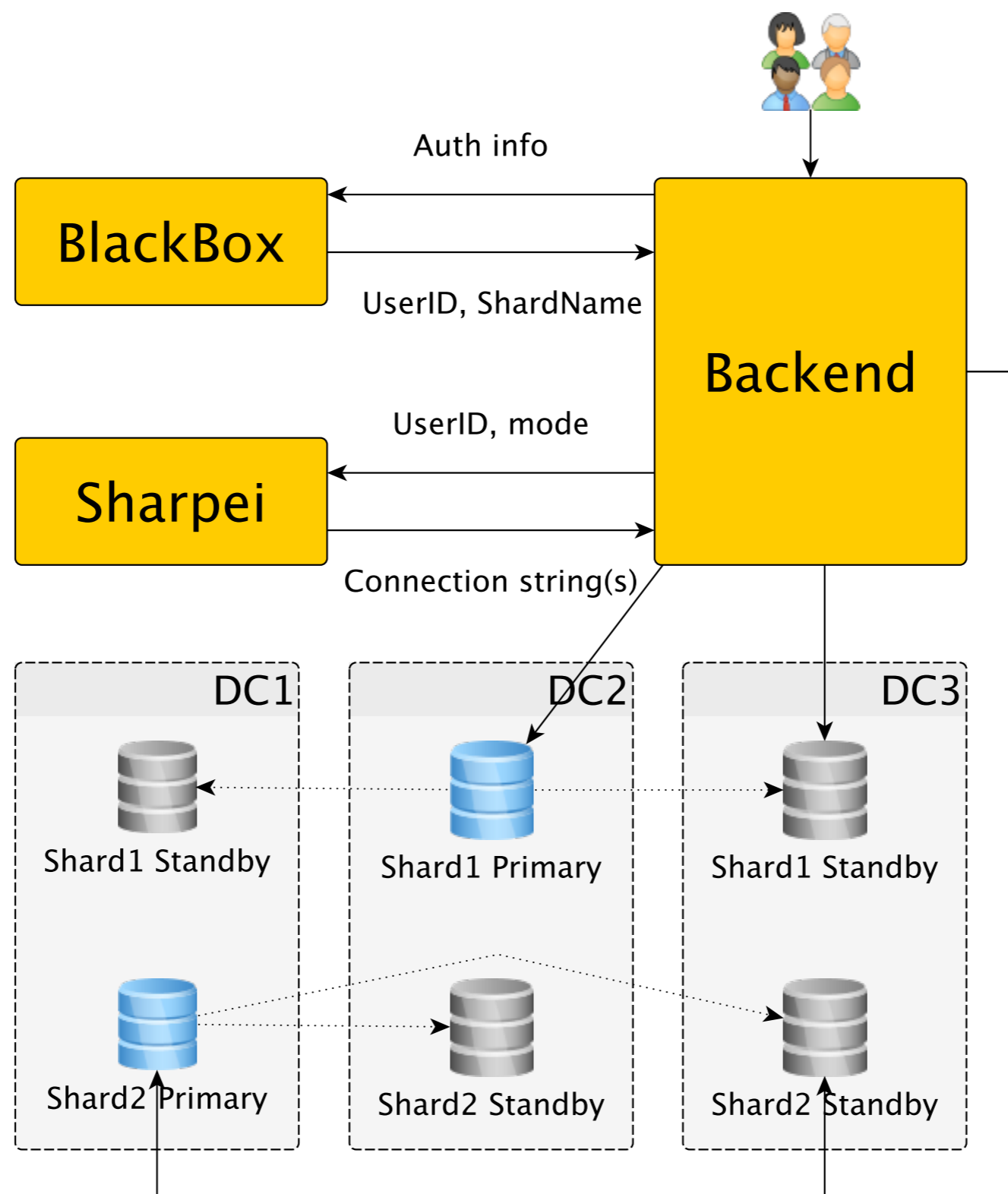
Основные изменения



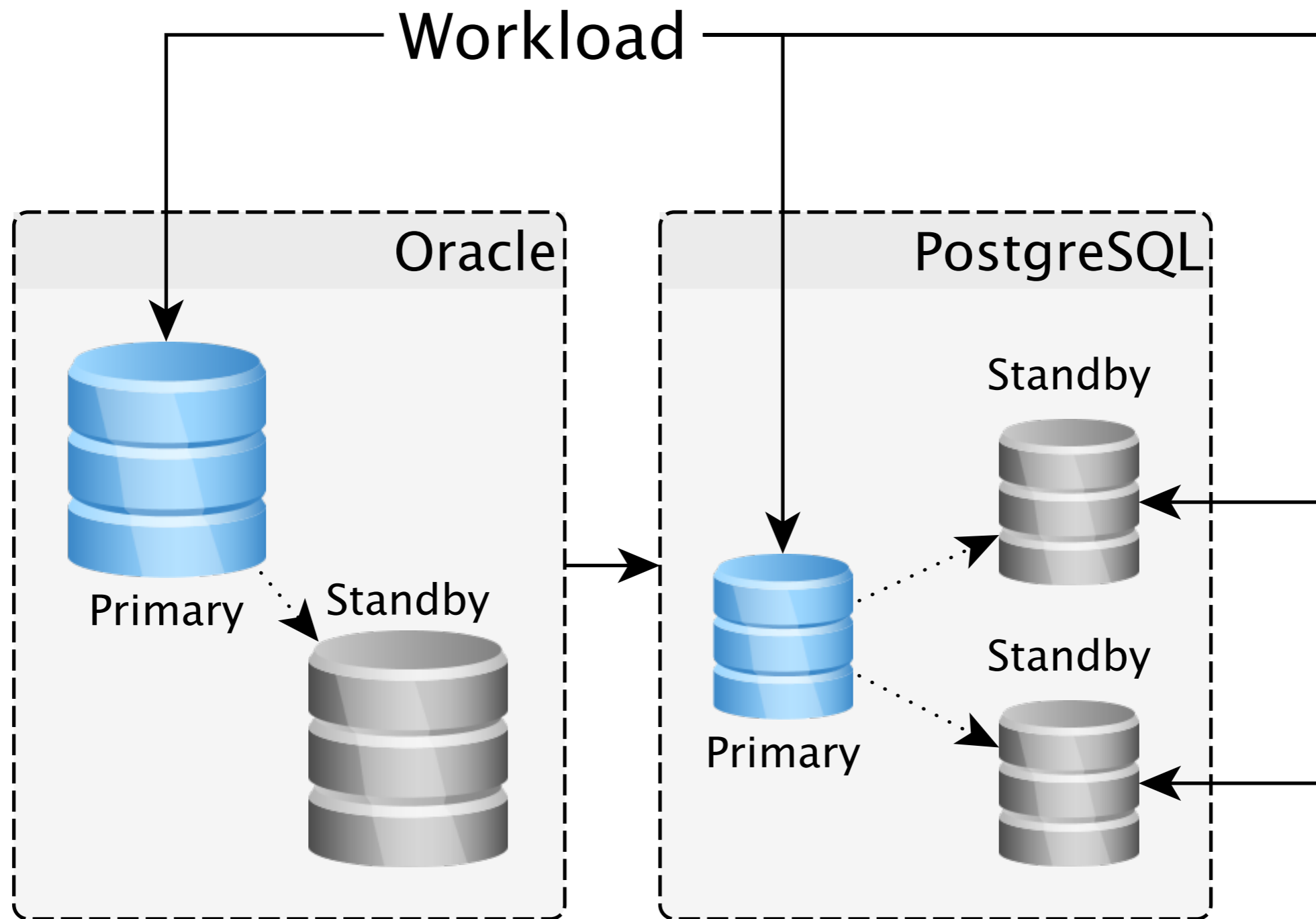
macs



Шардирование и отказоустойчивость



Аппаратное обеспечение



Аппаратное обеспечение

- › Тёплые базы (SSD) для большинства активных пользователей
- › Холодные базы (SATA) для всех неактивных пользователей
- › Горячие базы для очень активных пользователей
 - 2% пользователей создают 50% нагрузки
- › Автоматизация переноса пользователей между типами шардов
- › TBD: перемещение старых писем активных пользователей с SSD на SATA

Идентификаторы

В Oracle все идентификаторы (mid, fid, lid, tid) глобально уникальны

- › Диапазоны для sequence'ов каждого шарда в отдельной БД
- › NUMBER(20, 0) — 20 байт

В PostgreSQL идентификаторы уникальны в пределах одного логина

- › Вместо уникального mid уникальная пара (uid, mid)
- › Bigint + bigint — 16 байт

Изменения схемы

- › Больше конкуренция за одну страницу индекса
 - Обычный B-Tree вместо реверсивных индексов
- › Ревизии для всех объектов
 - Возможность читать с реплик только актуальные данные
 - Инкрементальные обновления для IMAP и мобильных
- › Денормализация части данных
 - Массивы и GIN
 - Композитные типы

Пример

```
xdb01g/maildb M # \dS mail.box
```

```
Table "mail.box"
```

Column	Type	Modifiers
uid	bigint	not null
mid	bigint	not null
lids	integer[]	not null

```
<...>
```

```
Indexes:
```

```
"pk_box" PRIMARY KEY, btree (uid, mid)
```

```
"i_box_uid_lids" gin (mail.ulids(uid, lids)) WITH (fastupdate=off)
```

```
<...>
```

```
xdb01g/maildb M #
```

Хранимая логика

- › PL/pgSQL прекрасен
- › Сильно сократили количество логики
 - Только для гарантии логической целостности данных
- › Сильно увеличили покрытие тестами
 - Цена ошибки очень высока
- › Простой deploy

Подход к обслуживанию

- › SaltStack

 - Детальный diff между текущим и ожидаемым состоянием

- › Все изменения схемы и кода через миграции

- › Все частые операции автоматизированы


- › Репрезентативные тестовые окружения

Проблемы



До начала основного переноса

- › Problem with ExclusiveLock on inserts
- › Checkpoint distribution
- › ExclusiveLock on extension of relation with huge shared_buffers
- › Hanging startup process on the replica after vacuuming on master
- › Replication slots and isolation levels
- › Segfault in BackendIdGetTransactionIds
- › Значительно больше решено без помощи сообщества



В любой непонятной
ситуации виноват
autovacuum

Oracle DBA

Диагностика

- › <https://simply.name/ru/pg-stat-wait.html>
- › [Wait_event in pg_stat_activity \(9.6\)](#)
- › <https://simply.name/ru/slides-pgday2015.html>

Резервные копии

› В Oracle бэкапы ($inc0 + 6 * inc1$) и archive логи \approx размер БД

› В PostgreSQL с barman'ом $\approx N * \text{размер БД}$, где $N > 5$

WAL'ы сжимаются, а бэкапы нет

File-level increment'ы толком не работают

Все операции однопоточные и очень медленные

› Для 300 ТБ данных понадобилось бы ≈ 2 ПБ под бэкапы

› <https://github.com/2ndquadrant-it/barman/issues/21>

› <http://pgday.ru/ru/2016/papers/80>

Во время основного переноса

› Проблемы не с PostgreSQL

› Проблемы с данными

Очень много legacy за 10+ лет

Ошибки в коде переноса

Завершение



Нам не хватает в PostgreSQL

- › Declarative partitioning
- › Хороший recovery manager
 - Параллелизм/сжатие/page-level increment'ы
 - Частичное восстановление (например, одной таблицы) в online
- › Дальнейшее развитие интерфейса ожиданий
- › Большой разделяемый кэш, O_DIRECT и асинхронное I/O
- › Quorum commit

Итоги

- › 1 РВ с отказоустойчивостью (100+ миллиардов строк)
- › 250к TPS
- › Три календарных года / 10+ человеколет
- › Быстрее deploy / эффективнее использование времени DBA
- › Рефакторинг кода всего бэкенда
- › В 3 раза больше аппаратного обеспечения
- › Ни одной крупной аварии пока :)
- › Linux, nginx, postfix, PostgreSQL

Вопросы?

Владимир Бородин

Системный администратор



+7 (495) 739 70 00, доб. 7255



@man_brain



d0uble@yandex-team.ru



<https://simply.name>