Опыт миграции между дата-центрами

Сергей Бурладян
sburladyan@avito.ru

Михаил Тюрин
mtyurin@avito.ru

2016

Что нового вы узнаете?

- законодательные инициативы *точно* затрагивают непосредственно *вашу* техническую платформу

- законодательные инициативы затрагивают техническую платформу
- переезд большого нагруженного кластера между датацентрами по интернету возможен (*за конечное время*) !
- и с небольшим даунтаймом

- можно достаточно точно оценить необходимую пропускную полосу при переезде
- существуют ярко выраженные особенности шейпинга выделенного интернет маршрута

- при этом скорее всего вы будете вынуждены работать на новом оборудование / окружении
- существуют неожиданные особенности systemd

- результаты некоторых нагрузочных тестов
- закон Мура у кого-то всё еще может работать

Переезд «базы» — часть сложного процесса миграции всей площадки:

- приложение
- *картинки*
- индексы
- другие базы
- dwh
- crons
- …

30 SubTasks и 30 Dependent — 6 месяцев

Тестировали датацентр боем: отдаём картинки из Москвы

> 50% трафика площадки

Подлежит ли откату процесс переезда?

Нет, очень сложно

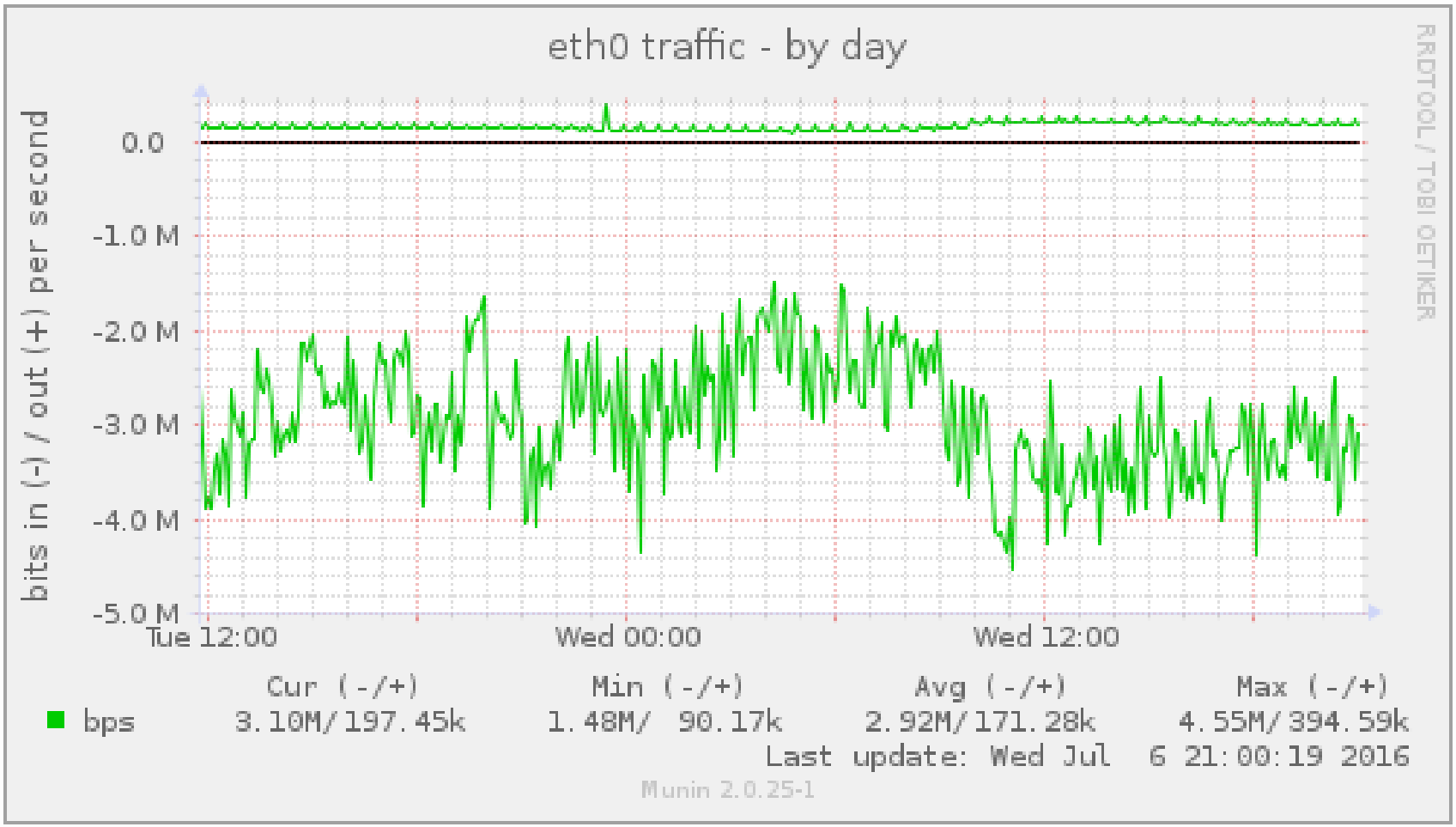
Интуитивное решение о применении репликации — правильно

Как перенести базы с минимальным downtime?

Репликация!

- физическая — streaming, WAL архив // x2
- логическая — londiste // x2
- rpc — pgq + xrpcd.py // x2

Оценка необходимой пропускной способности



Оценка необходимой пропускной способности

- 12 серверов WAL ~ 100 — 150 Mbit
- 3 сервера londiste + rpc < 50 Mbit

- с учётом возможного роста до двух раз (переезд — 6 месяцев)

Первый вариант

NFS через интернет слишком круто

Сейчас просто включим стриминг в Москву и всё!

Оказывается

- Канал 2 Gbit
- Но на одну TCP сессию почему-то 2 Mbit

2 Gbit → 2 Mbit

Вариант два

- WAL архив (pbzip2):

  send-wals.sh + rsync

archive_command и т. п.: https://pgconf.ru/2016/89927

Вариант три

- send-wals.sh + parallel-rsync.sh

send-wals.sh:

```
get_new_files > "$FILES_LIST"
...
send_files < "$FILES_LIST" > "$SEND_OUTPUT"
...
mark_sended < "$FILES_LIST"
```

parallel-rsync.sh:

```
| xargs -0 -P "$PMAX" -n "$ARGS_NUM" \
    bash -e -o pipefail \
    -c 'dst_rsync "$@"' \
    -- "$DST_DIR" "${RSYNC_ARGS[@]}" --
```

Новая среда и оборудование

Проблема нового Debian

- pgbench Debian 6 vs Debian 7: в 2-3 раза!

- === 8.2 (jessie) | 3.16.7-ckt11-1+deb8u6 | AMD Opteron(tm) Processor 6136
- $ pgbench -SNn -r -f test.sql -p 6432 -U postgres -c 10 -T $((5 * 60)) avito_test
- tps = 7528.535949 (excluding connections establishing)

- === 6.0.3 (squeeze) | 2.6.32-5-amd64 | AMD Opteron(tm) Processor 6136
- $ pgbench -SNn -r -f test.sql -p 6432 -U postgres -c 10 -T $((5 * 60)) avito_test
- tps = 9284.494020 (excluding connections establishing)

- Debian 6 быстрее Debian 8 на ~ 20%

Новая среда и оборудование

- === 6.0.3 (squeeze) | 2.6.32-5-amd64 | AMD Opteron(tm) Processor 6136 2399.891 MHz | 9.2.13-1.pgdg60+1
- tps = 8212.192504 (excluding connections establishing)

- === 8.2 (jessie) | 3.16.7-ckt20-1+deb8u2 | Intel(R) Xeon(R) CPU E5-2697 v3 @ 2.60GHz | 9.2.14-1.pgdg80+1
- tps = 19067.844979 (excluding connections establishing)

- mbw

Особенности systemd

- /etc/systemd/logind.conf

  RemoveIPC=no

  https://www.postgresql.org/message-id/20151214224729.2624.99840@wrigleys.postgresql.org

- KillMode=none

  SendSIGKILL=no

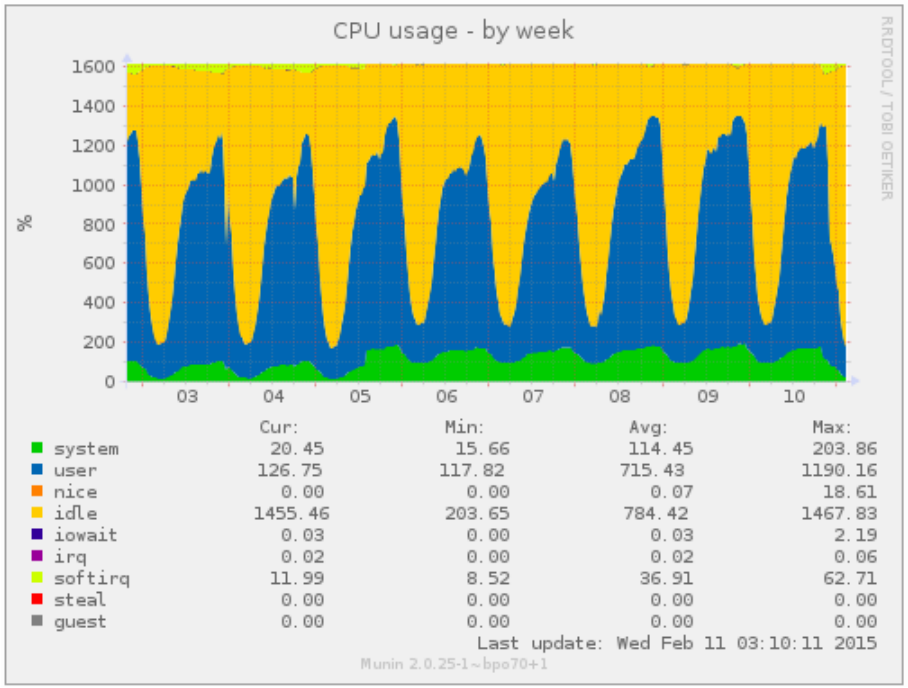
  https://bugzilla.suse.com/show_bug.cgi?id=906900

- sudo pg_ctlcluster 9.2 main stop -- -m f

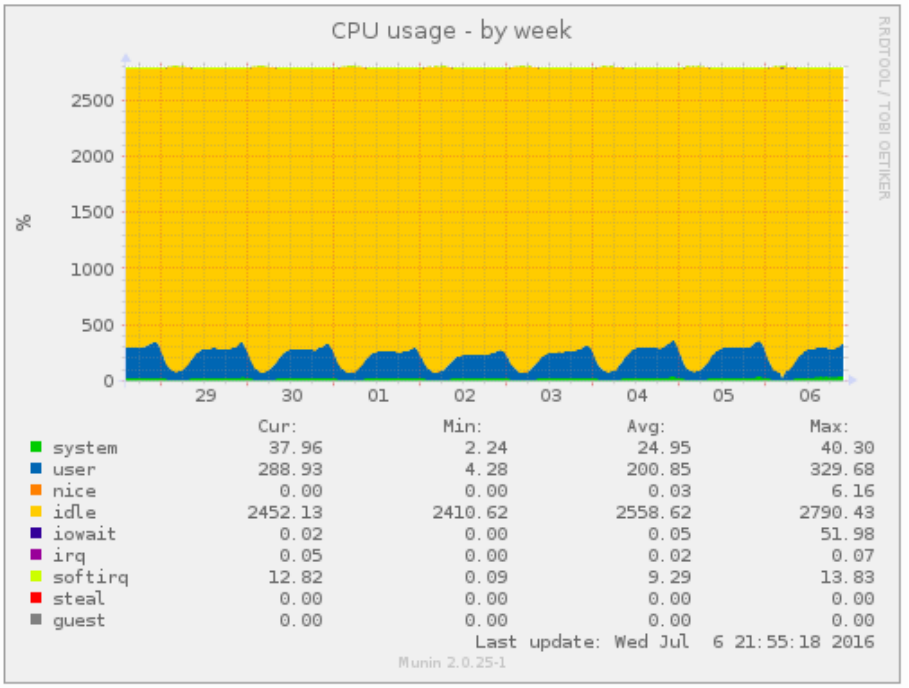
  sudo -u postgres pg_ctlcluster 9.2 main stop -m f

Закон Мура всё ещё работает

2015



2016



Спасибо за внимание!