OSSDEV-2015: **ReOpenLDAP**

Кластерный LDAP для «Больших телекомов»

Леонид Юрьев
Петер-Сервис РнД, Сколково

Кластерный LDAP для Больших телекомов

PETER-SERVICE

- Решения для крупных операторов связи
- Полный цикл: проектирование, разработка, внеднение, поддержка
- более 20 лет
- > 100 миллионов абонентов
- http://www.billing.ru

Кластерный LDAP для Больших телекомов

Зачем нам это?

- LDAP «прописан» как протокол взаимодействия Telco-подсистем
- Сервер LDAP является обязательным компонентом
- Решили обзавестить своим «кубиком»

Кластерный LDAP для Больших телекомов

Требования, однако:

- от 10К обновлений в секунду
- от 25К запросов в секунду, с перспективой роста
- 10^7-10^8 записей
- Multi-master репликация
- Гео-распределенный кластер 2x2
- 24x7

Кластерный LDAP для Больших телекомов

Кандидаты:

- 389DS, OpenDJ, ApacheDS, OpenLDAP...
- Желаемую нагрузку по-записи не выдержал никто
- OpenLDAP:
  - LMDB внутри, MVCC
  - Линейное масштабироание по-чтению
  - Нет взаимодеградации чтение/запись
  - Производительность по-записи упирается в диск
  - Есть mmapped-режим, ядро сохрянит базу при аварии

Кластерный LDAP для Больших телекомов

OpenLDAP, проблемы:

- Утечки памяти
- Падения в коде БД-движка
- Падения при проблемах в сети
- Странное поведение репликации…
- Регулярные переполнения базы

Кластерный LDAP для Больших телекомов

Facepalm:

- Symas Corp принимали репорты и всё…
- Начали разбираться сами:
  - Распухание FreeDB → пробный LIFO патч
  - Ошибки double-free и ±1
  - Valgrind
  - Race conditions и поехали…
- Адъ и содомия в исходном коде

Кластерный LDAP для Больших телекомов

Почему форк:

- **5000** предупреждений, 99% безобидные, 1% баги
- Качество кода «спорное», технический долг захлестывает
- ***Нам нужно ехать, а не шашечьки***
- А также:
  - Патч завернули из-за __VA_ARGS__
  - Нет актуальных компиляторов без поддержки __VA_ARGS__
  - __VA_ARGS__ уже были в исходниках
  - Никто не проверяет собираемость на старых платформах
  - Устранение других предупреждений не заинтересовало меинтейноров OpenLDAP

Кластерный LDAP для Больших телекомов

OpenLDAP = facepalm!

1) Проект не ориентирован на развитие усилиями сообщества

2) … 5) … 10)

Так НЕ надо делать ПО!

Please, don't create a software anymore, especially multi-threaded, and a protocols related to replication too.

This will be enough to make the World better, and save my time ;)

Кластерный LDAP для Больших телекомов

Наш ReOpenLDAP:

- Забыли про windows и устаревшие системы
- На FreeBSD меинтейнера не нашлось
- Сконцентрировались на затребованном:
  - только **Linux и x86_64**(backtrace feature)
  - не старше RHEL6
  - GCC либо Clang
  - Стабильность, Репликация, Производительность.

Кластерный LDAP для Больших телекомов

Основные доработки:

- Инфраструктура: чекер памяти, IDKFA...
- Тесты: теперь они могут не падать
- LMDB: собственная версия, отдельная тема (aka MDBX)
- SLAPD: выявленно и устраненно «100500»
- Репликация: наша боль...

Кластерный LDAP для Больших телекомов

MDBX (наша версия движка базы данных):

- Будет доклад на HighLoad++2015:
  - http://www.highload.ru/2015/abstracts/1831.html
  - 2 – 3 ноября, Москва, Крокус-Экспо
  - С результатами бенчмарков и сравнениями с Sophia, WiredTiger, LevelDB…
- Гарантия консистентности для WRITEMAP
- LIFO
- OOM-Handler
- mdb_chk

Кластерный LDAP для Больших телекомов

SLAPD (1):

- Больше нет проблемы «отставших читателей»
  - Dreamcatcher минимизирует вероятность
  - Oom-handler гарантирует
- BigLock
  - Устраняет волатильность: UUID → DN

Кластерный LDAP для Больших телекомов

SLAPD (2):

- «Кворум» при репликации
  - Переводит узел в readonly при нарушении связанности
- Limit-concurrent-refresh
  - В разы уменьшает пиковое потребление RAM и CPU
  - Внезапно кластер синхронизируется в разы быстрее
- Checkpoints
  - Периодически с посекундной точностью
  - По объему изменений

Кластерный LDAP для Больших телекомов

SLAPD (3)

- Репликация, много переделок, но еще болеет:
  - Доработали тесты и стали замечать
  - https://github.com/ReOpen/ReOpenLDAP/issues/43
  - Только в августе наши способ/сценарий воспроизведения
- ≈ 50% кода переписано
- До 2-3 раз снижено кол-во транзакций
  - Пропуск лишних обновлений contextCSN
- ...

Кластерный LDAP для Больших телекомов

Что дальше?

- Получили приемлемую версию:
  - Наиболее стабильная из известных нам
  - Наиболее проиводительная
  - Много мозолей и пота
- Были планы зарелизиться летом, но репликация...
- Петер-Сервис с вероятность 50%:
  - Стратегически выберет другую платформу
  - Будет поддерживать релиз до вывода из эксплуатации
- Будем пробовать передать в сообщество, но не в родительский проект.