PostgresPro

Системные баги под нагрузкой

Миша Жилин, Postgres Professional

C++ Russia, 17 мая 2026

Миша Жилин

- Физтех, 2008 год выпуска
- 18 лет занимаюсь производительностью программ
- Больше исследователь
- Меньше разработчик
- Обожаю Open Source, PostgreSQL и FreeBSD
- Работаю 5,5 лет в Postgres Professional

О чём доклад

PostgresPro

- Операционные системы работают как часы

О чём доклад

PostgresPro

- Операционные системы работают как часы
- Базы данных всегда возвращают корректный результат

О чём доклад

PostgresPro

- Операционные системы работают как часы
- Базы данных всегда возвращают корректный результат
- Системный софт - надёжный, качественный, проверенный

О чём доклад

PostgresPro

- Операционные системы работают как часы
- Базы данных всегда возвращают корректный результат
- Системный софт - надёжный, качественный, проверенный
- Или нет?!

О чём доклад

PostgresPro

- Операционные системы работают как часы
- Базы данных всегда возвращают корректный результат
- Системный софт - надёжный, качественный, проверенный
- Или нет?!
- Расскажу только то, что видел

О чём доклад

PostgresPro

- Операционные системы работают как часы
- Базы данных всегда возвращают корректный результат
- Системный софт - надёжный, качественный, проверенный
- Или нет?!
- Расскажу только то, что видел
- Что не видел, то не расскажу

О чём доклад

PostgresPro

- Поиск невозможного

О чём доклад

PostgresPro

- Поиск невозможного
- Несколько историй из личного опыта

О чём доклад

PostgresPro

- Поиск невозможного
- Несколько историй из личного опыта
- Включая много попыток поехать кукухой

О чём доклад

PostgresPro

- Поиск невозможного
- Несколько историй из личного опыта
- Включая много попыток поехать кукухой
- Мало кода

О чём доклад

PostgresPro

- Поиск невозможного
- Несколько историй из личного опыта
- Включая много попыток поехать кукухой
- Мало кода
- Много QR кодов (а может и немного)

Разгон

Intel SSD, 2019 год

Intel SSD, 2019 год

PostgresPro

- Пробный период запуска мобильного оператора

Intel SSD, 2019 год

PostgresPro

- Пробный период запуска мобильного оператора
- Отказ в распределенном дисковом пространстве CEPH

Intel SSD, 2019 год

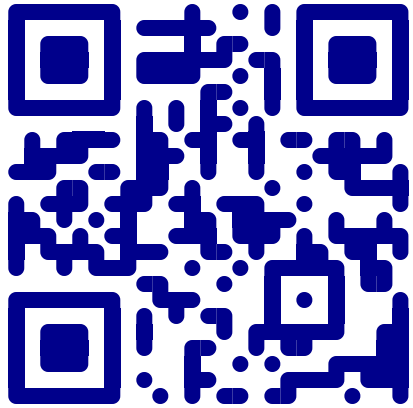
PostgresPro

- Пробный период запуска мобильного оператора
- Отказ в распределенном дисковом пространстве CEPH
- Зависание дисков Intel SSD S45210 и S4610

Intel SSD, 2019 год

PostgresPro

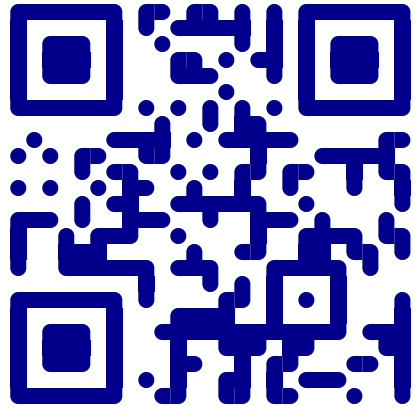
- Intel SSD D3-S4510 and Intel SSD D3-S4610 series 1.92 TB and 3.84 TB drives become unresponsive after 1,700 cumulative idle power-on hours
- Currently addressed in the Maintenance Release 1 (MR1) of Intel firmware (version XCV10110) as of March 2019



Intel SSD, 2019 год

PostgresPro

- Intel удалил все упоминания
- https://forums.servethehome.com/index.php?threads/psa-errors-with-intel-r-solid-state-drive-dc-s4510-and-s4610-series.26196/



WAL corruption, 2020 год

WAL corruption, 2020 год

PostgresPro

- Смена работы
  - Инженер по отказоустойчивости в кластерной группе
  - 4ая неделя испытательного срока

WAL corruption, 2020 год

PostgresPro

- Смена работы
  - Инженер по отказоустойчивости в кластерной группе
  - 4ая неделя испытательного срока
- Проблема от клиента

WAL corruption, 2020 год

PostgresPro

- Смена работы
  - Инженер по отказоустойчивости в кластерной группе
  - 4ая неделя испытательного срока
- Проблема от клиента
- База данных на 2 серверах: мастер и реплика

WAL corruption, 2020 год

PostgresPro

- Смена работы
  - Инженер по отказоустойчивости в кластерной группе
  - 4ая неделя испытательного срока
- Проблема от клиента
- База данных на 2 серверах: мастер и реплика
- Ошибка репликации
  - `[2020-10-27 12:03:02 MSK u= d= h= p=14570 l=99]`
  - `LOG: incorrect resource manager data checksum in record at 61E7/ED365B20]`

WAL corruption, 2020 год

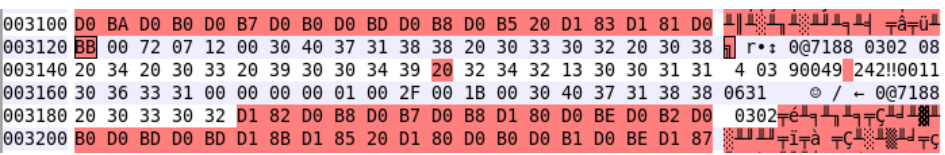
PostgresPro

- Смена работы
  - Инженер по отказоустойчивости в кластерной группе
  - 4ая неделя испытательного срока
- Проблема от клиента
- База данных на 2 серверах: мастер и реплика
- Ошибка репликации
  - `[2020-10-27 12:03:02 MSK u= d= h= p=14570 l=99]`
  - `LOG: incorrect resource manager data checksum in record at 61E7/ED365B20]`
- Неверная контрольная сумма - повреждение WAL журналов

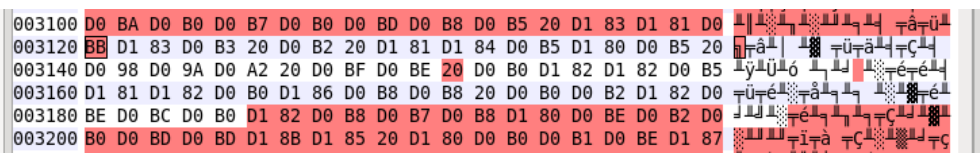
WAL corruption, 2020 год

PostgresPro

- Смена работы
  - Инженер по отказоустойчивости в кластерной группе
  - 4ая неделя испытательного срока
- Проблема от клиента
- База данных на 2 серверах: мастер и реплика
- Ошибка репликации
  - `[2020-10-27 12:03:02 MSK u= d= h= p=14570 l=99]`
  - `LOG: incorrect resource manager data checksum in record at 61E7/ED365B20]`
- Неверная контрольная сумма - повреждение WAL журналов
- Материал для анализа: `record at 61E7/ED365B20`



```
003100 D0 BA D0 B0 D0 B7 D0 B0 D0 BD D0 B8 D0 B5 20 D1 83 D1 81 D0
003120 BB 00 72 07 12 00 30 40 37 31 38 38 20 30 33 30 32 20 30 38
003140 20 34 20 30 33 20 39 30 30 34 39 20 32 34 32 13 30 30 31 31
003160 30 36 33 31 00 00 00 00 01 00 2F 00 1B 00 30 40 37 31 38 38
003180 20 30 33 30 32 D1 82 D0 B8 D0 B7 D0 B8 D1 80 D0 BE D0 B2 D0
003200 B0 D0 BD D0 BD D1 8B D1 85 20 D1 80 D0 B0 D0 B1 D0 BE D1 87
```



```
003100 D0 BA D0 B0 D0 B7 D0 B0 D0 BD D0 B8 D0 B5 20 D1 83 D1 81 D0
003120 BB D1 83 D0 B3 20 D0 B2 20 D1 81 D1 84 D0 B5 D1 80 D0 B5 20
003140 D0 98 D0 9A D0 A2 20 D0 BF D0 BE 20 D0 B0 D1 82 D1 82 D0 B5
003160 D1 81 D1 82 D0 B0 D1 86 D0 B8 D0 B8 20 D0 B0 D0 B2 D1 82 D0
003180 BE D0 BC D0 B0 D1 82 D0 B8 D0 B7 D0 B8 D1 80 D0 BE D0 B2 D0
003200 B0 D0 BD D0 BD D1 8B D1 85 20 D1 80 D0 B0 D0 B1 D0 BE D1 87
```

WAL corruption, 2020 год

PostgresPro

- Ошибка диска/OS: и всё что над ним RAID контролер, драйвера, блочной системы, файловой системы

WAL corruption, 2020 год

PostgresPro

- Ошибка диска/OS: и всё что над ним RAID контролер, драйвера, блочной системы, файловой системы
- Ошибка СУБД (не исключена вероятность “гуляющего” указателя)

WAL corruption, 2020 год

PostgresPro

- Ошибка диска/OS: и всё что над ним RAID контролер, драйвера, блочной системы, файловой системы
- Ошибка СУБД (не исключена вероятность “гуляющего” указателя)
- **Ошибка памяти**:
  - адрес повреждения в WAL файле – 0x366500 (выровнен на 256 байт)
  - размер повреждения - 64 байта (размер кэшлайна процессора)

WAL corruption, 2020 год

PostgresPro

- Через месяц ещё одно падение

WAL corruption, 2020 год

PostgresPro

- Через месяц ещё одно падение
- Потом ещё несколько

WAL corruption, 2020 год

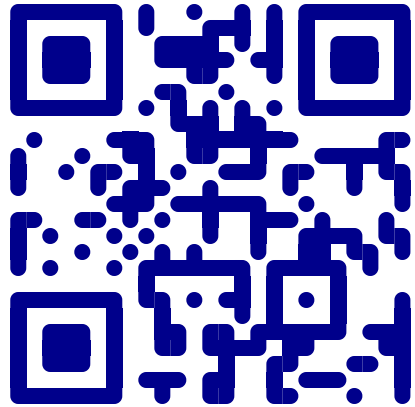
PostgresPro

- Через месяц ещё одно падение
- Потом ещё несколько
- Но всегда портился только WAL, данные проверяли - они все верные

WAL corruption, 2020 год

PostgresPro

- Двойной WAL буфер - защита от disk/memory corruption
- Настройка wal_sender_check_crc в PostgresPro Enterprise



Наступил 2021 год

WAL corruption, 2021 год

PostgresPro

- Изоляция проблемного сервера и доступ для экспериментов

WAL corruption, 2021 год

PostgresPro

- Изоляция проблемного сервера и доступ для экспериментов
- BullSequana S series
  - Несколько материнских плат
  - 4 терабайта памяти
  - 8 процессоров

WAL corruption, 2021 год

PostgresPro

- Изоляция проблемного сервера и доступ для экспериментов
- BullSequana S series
  - Несколько материнских плат
  - 4 терабайта памяти
  - 8 процессоров
- Недельные тесты
  - Ноль результатов

WAL corruption, 2021 год

PostgresPro

- Изоляция проблемного сервера и доступ для экспериментов
- BullSequana S series
  - Несколько материнских плат
  - 4 терабайта памяти
  - 8 процессоров
- Недельные тесты
  - Ноль результатов
- MemTest
  - Ноль результатов

WAL corruption, 2021 год

PostgresPro

- Полный ноль

WAL corruption, 2021 год

PostgresPro

- Полный ноль
- Фрустрации

WAL corruption, 2021 год

PostgresPro

- Полный ноль
- Фрустрации
  - А что ищем?

WAL corruption, 2021 год

PostgresPro

- Полный ноль
- Фрустрации
  - А что ищем?
  - Как задать вопрос?

WAL corruption, 2021 год

PostgresPro

- Полный ноль
- Фрустрации
  - А что ищем?
  - Как задать вопрос?
  - Что я не понимаю?

WAL corruption, 2021 год

PostgresPro

- Полный ноль
- Фрустрации
  - А что ищем?
  - Как задать вопрос?
  - Что я не понимаю?
- Google it, опытные коллеги

WAL corruption, 2021 год

PostgresPro

- Полный ноль
- Фрустрации
  - А что ищем?
  - Как задать вопрос?
  - Что я не понимаю?
- Google it, опытные коллеги
- Время менять подходы

WAL corruption, 2021 год

PostgresPro

- Основная гипотеза - это память

WAL corruption, 2021 год

PostgresPro

- Основная гипотеза - это память
- stress-ng - утилита нагрузки на железные ресурсы
  - `stress-ng --vm-rw 1000 --vm-rw-bytes 2500G --verify --metrics-brief -t 60m`

WAL corruption, 2021 год

PostgresPro

- Основная гипотеза - это память
- stress-ng - утилита нагрузки на железные ресурсы
  - `stress-ng --vm-rw 1000 --vm-rw-bytes 2500G --verify --metrics-brief -t 60m`
- Никаких ошибок программы

WAL corruption, 2021 год

PostgresPro

- Основная гипотеза - это память
- stress-ng - утилита нагрузки на железные ресурсы
  - `stress-ng --vm-rw 1000 --vm-rw-bytes 2500G --verify --metrics-brief -t 60m`
- Никаких ошибок программы
- Зато Fiber Channel начал мигать
  - `2021 Feb 15 23:43:57 dn-MDS-C9148S-1 %PORT-5-IF_DOWN_LINK_FAILURE: %$VSAN 2221%$ Interface fc1/15 is down (Link failure loss of signal)`

WAL corruption, 2021 год

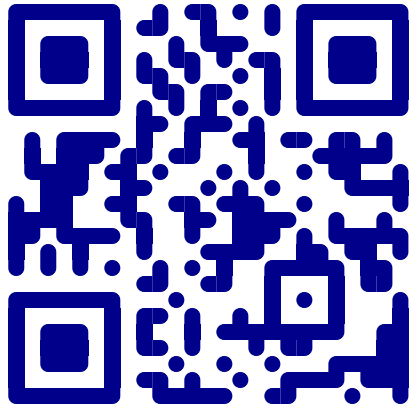
PostgresPro

- Основная гипотеза - это память
- stress-ng - утилита нагрузки на железные ресурсы
  - `stress-ng --vm-rw 1000 --vm-rw-bytes 2500G --verify --metrics-brief -t 60m`
- Никаких ошибок программы
- Зато Fiber Channel начал мигать
  - `2021 Feb 15 23:43:57 dn-MDS-C9148S-1 %PORT-5-IF_DOWN_LINK_FAILURE: %$VSAN 2221%$ Interface fc1/15 is down (Link failure loss of signal)`
- Время и идеи закончились

WAL corruption, 2021 год

PostgresPro

- Через полгода статья от техподдержки производителя железа
  - https://habr.com/ru/companies/atosservers/articles/562376/
- Обновление драйвера FC линков (драйвер устройства операционной системы Linux).

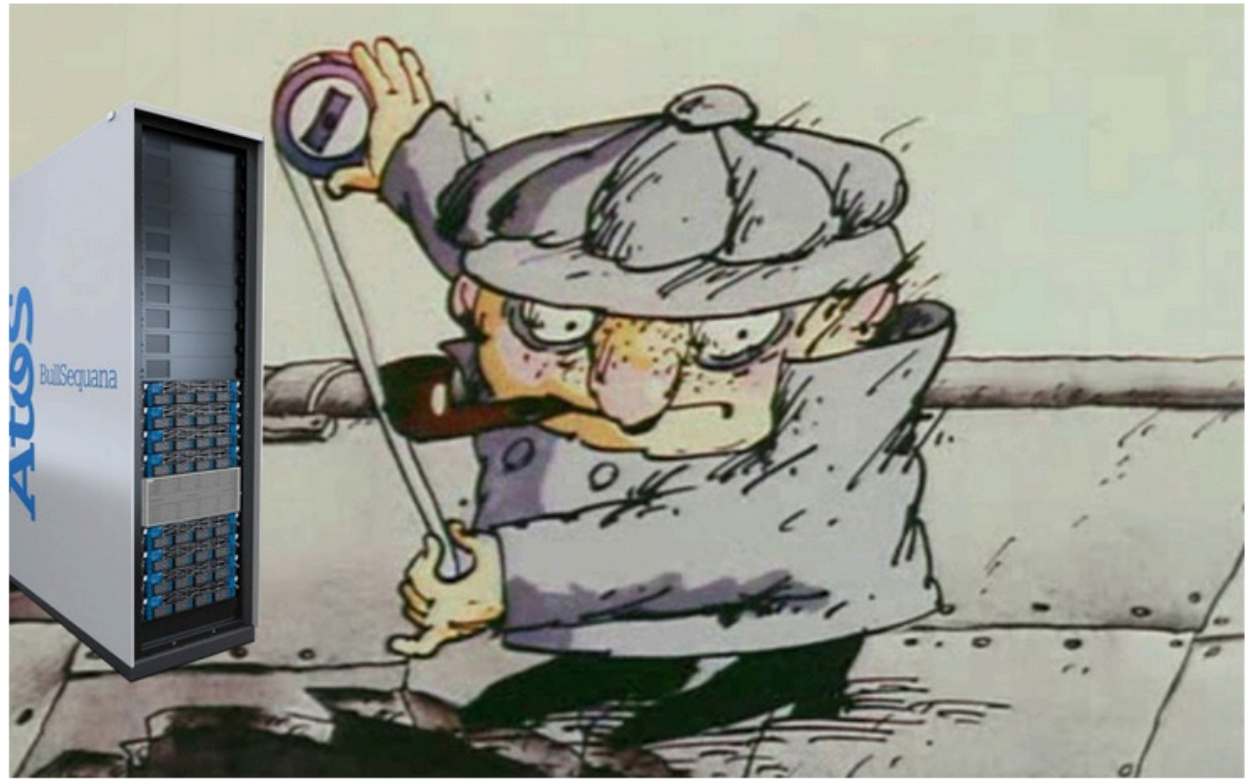


sharkattack 18 июн 2021 в 18:24

Следствие вели: пропажа FC-линков HBA Emulex на сервере Atos BullSequana S1600

6 мин 3.8K

Блог компании Atos, Высоконагруженные системы*, Тестирование IT-систем*, IT-компании



Цитатник

PostgresPro

- "За всё время этой проблемы пошатали и потрогали всё – саму ОС, FW, прошивку HW сервера, настройки HW сервера, параметры GRUB, настройки фабрики, свитчей и линков..."
- "Техподдержка врёт (ну или добросовестно заблуждается), поэтому приходится старательно все проверять самому"
- "Трогать и шатать при траблшутинге надо всё!"

Итоги

PostgresPro

- Обновления драйвера и реализации двойного буфера WAL журналов
- Больше повреждений WAL не было
- Кто виноват - до сих пор загадка
- Это успех или нет?
- Негативный опыт - тоже опыт

Полезные моменты

PostgresPro

- Нужен стенд. Люди, увы - не телепаты
- Если тест не выявляет дефект, то наверно не то нагружаем или не достаточно
- Полезно добавлять генераторы нагрузки на ресурсы ОС
- Пример - **stress-ng**

Прошли годы...

Прошли годы

PostgresPro

- Меня выбросили из кластерной группы как непрофпригодного
  - "Иди делай группу нагрузочных исследователей" 😜

Прошли годы

PostgresPro

- Меня выбросили из кластерной группы как непрофпригодного
  - "Иди делай группу нагрузочных исследователей" 😜
- Коллаборация с нашей техподдержкой
  - Они нам задачи
  - Мы им помощь в исследованиях

Прошли годы

PostgresPro

- Меня выбросили из кластерной группы как непрофпригодного
  - "Иди делай группу нагрузочных исследователей" 😜
- Коллаборация с нашей техподдержкой
  - Они нам задачи
  - Мы им помощь в исследованиях
- История с подвисанием сервера на Linux 5.15

Прошли годы

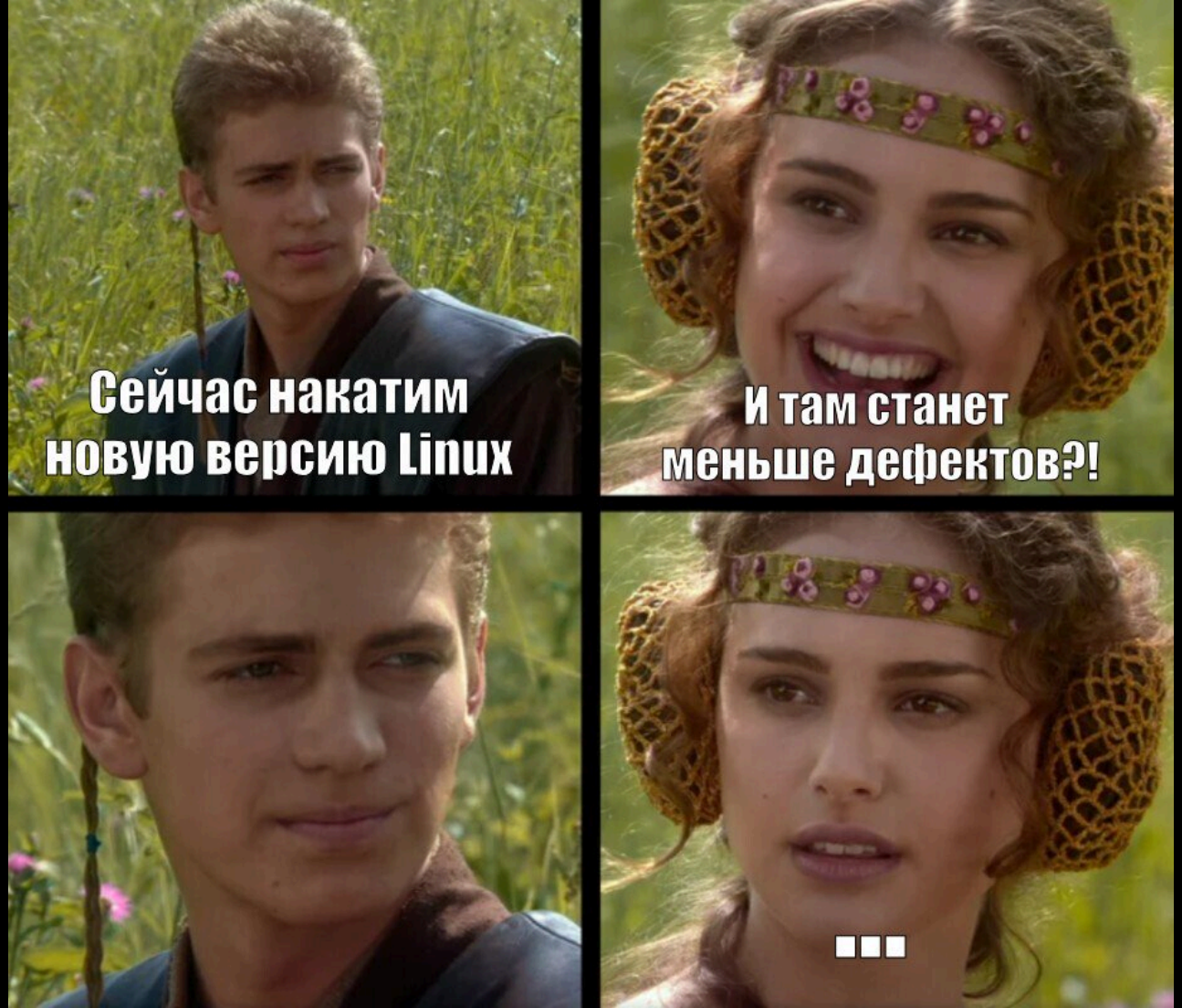
PostgresPro

- Меня выбросили из кластерной группы как непрофпригодного
  - "Иди делай группу нагрузочных исследователей" 😜
- Коллаборация с нашей техподдержкой
  - Они нам задачи
  - Мы им помощь в исследованиях
- История с подвисанием сервера на Linux 5.15
- Проблема масштабирования на lookup-а файлов в XFS

Прошли годы

PostgresPro

- Меня выбросили из кластерной группы как непрофпригодного
  - "Иди делай группу нагрузочных исследователей" 😜
- Коллаборация с нашей техподдержкой
  - Они нам задачи
  - Мы им помощь в исследованиях
- История с подвисанием сервера на Linux 5.15
- Проблема масштабирования на lookup-а файлов в XFS
- **Общая рекомендация**: обновитесь до новой версии операционной системы


Сейчас накатим
новую версию Linux
И там станет
меньше дефектов?!
...

segfault, 2024

segfault, 2024

PostgresPro

- 30 декабрь 2024 года

segfault, 2024

PostgresPro

- 30 декабрь 2024 года
- Миграция данных

segfault, 2024

PostgresPro

- 30 декабрь 2024 года
- Миграция данных
- Упала база данных по segfault

segfault, 2024

PostgresPro

- 30 декабрь 2024 года
- Миграция данных
- Упала база данных по segfault
- `[Mon Dec 30 17:49:50 2024] postgres[4092660]: segfault at 55edf40f20d0 ip 000055edf40f20d0 sp 00007ffc7a7773c8 error 15 in postgres[55edf3f67000+2a9000] likely on CPU 40 (core 16, socket 1)`

segfault, 2024

PostgresPro

- checkpointer (4092660)

```
#0  0×000055edf40f20d0 in ?? ()
#1  0×000055edf3f1a370 in hash_search (hashp=0×55edf5c1ada8, keyPtr=0×7ffc7a777480, action=HASH_ENTER, foundPtr=0×7ffc7a
#2  0×000055edf3dd2dd2 in RememberSyncRequest (ftag=0×7ffc7a777480, type=<optimized out>) at sync.c:565
#3  0×000055edf3dd2fb5 in RegisterSyncRequest (ftag=ftag@entry=0×7ffc7a777480, type=type@entry=SYNC_REQUEST, retryOnErro
```

- startup (4092662)

```
#0  0×000055edf40f20d0 in ?? ()
#1  0×000055edf3f1a370 in hash_search (hashp=0×55edf5c458f8, keyPtr=0×7ffc7a7781ec, action=HASH_FIND, foundPtr=0×0) at c
#2  0×000055edf3b40603 in XLogPrefetcherIsFiltered (blockno=1523809, rnode= ... , prefetcher=0×55edf5c375f8) at xlogprefet
#3  XLogPrefetcherNextBlock (pgsr_private=94480518837752, lsn=<optimized out>) at xlogprefetcher.c:693
#4  0×000055edf3b40f51 in lrq_prefetch (lrq=<optimized out>) at xlogprefetcher.c:256
#5  lrq_complete_lsn (lsn=67516543241680, lrq=0×55edf5c3aff8) at xlogprefetcher.c:294
```

segfault, 2024

PostgresPro

- checkpointer (4092660)

```
#0  0×000055edf40f20d0 in ?? ()
#1  0×000055edf3f1a370 in hash_search (hashp=0×55edf5c1ada8, keyPtr=0×7ffc7a777480, action=HASH_ENTER, foundPtr=0×7ffc7a
#2  0×000055edf3dd2dd2 in RememberSyncRequest (ftag=0×7ffc7a777480, type=<optimized out>) at sync.c:565
#3  0×000055edf3dd2fb5 in RegisterSyncRequest (ftag=ftag@entry=0×7ffc7a777480, type=type@entry=SYNC_REQUEST, retryOnErro
```

- startup (4092662)

```
#0  0×000055edf40f20d0 in ?? ()
#1  0×000055edf3f1a370 in hash_search (hashp=0×55edf5c458f8, keyPtr=0×7ffc7a7781ec, action=HASH_FIND, foundPtr=0×0) at c
#2  0×000055edf3b40603 in XLogPrefetcherIsFiltered (blockno=1523809, rnode= ... , prefetcher=0×55edf5c375f8) at xlogprefet
#3  XLogPrefetcherNextBlock (pgsr_private=94480518837752, lsn=<optimized out>) at xlogprefetcher.c:693
#4  0×000055edf3b40f51 in lrq_prefetch (lrq=<optimized out>) at xlogprefetcher.c:256
#5  lrq_complete_lsn (lsn=67516543241680, lrq=0×55edf5c3aff8) at xlogprefetcher.c:294
```

segfault, 2024

PostgresPro

- checkpointer (4092660)

```
#0  0×000055edf40f20d0 in ?? ()
#1  0×000055edf3f1a370 in hash_search (hashp=0×55edf5c1ada8, keyPtr=0×7ffc7a777480, action=HASH_ENTER, foundPtr=0×7ffc7a
#2  0×000055edf3dd2dd2 in RememberSyncRequest (ftag=0×7ffc7a777480, type=<optimized out>) at sync.c:565
#3  0×000055edf3dd2fb5 in RegisterSyncRequest (ftag=ftag@entry=0×7ffc7a777480, type=type@entry=SYNC_REQUEST, retryOnErro
```

- startup (4092662)

```
#0  0×000055edf40f20d0 in ?? ()
#1  0×000055edf3f1a370 in hash_search (hashp=0×55edf5c458f8, keyPtr=0×7ffc7a7781ec, action=HASH_FIND, foundPtr=0×0) at c
#2  0×000055edf3b40603 in XLogPrefetcherIsFiltered (blockno=1523809, rnode= ... , prefetcher=0×55edf5c375f8) at xlogprefet
#3  XLogPrefetcherNextBlock (pgsr_private=94480518837752, lsn=<optimized out>) at xlogprefetcher.c:693
#4  0×000055edf3b40f51 in lrq_prefetch (lrq=<optimized out>) at xlogprefetcher.c:256
#5  lrq_complete_lsn (lsn=67516543241680, lrq=0×55edf5c3aff8) at xlogprefetcher.c:294
```

segfault, 2024

PostgresPro

- Два процесса одновременно

segfault, 2024

PostgresPro

- Два процесса одновременно
- **hash_search**
  - Фундаментальный код
  - Горячий код (часто выполняемый)
  - Вероятность найти ошибку почти нулевая

segfault, 2024

PostgresPro

- Два процесса одновременно
- **hash_search**
  - Фундаментальный код
  - Горячий код (часто выполняемый)
  - Вероятность найти ошибку почти нулевая
- Трудность

segfault, 2024

PostgresPro

- Начинаем искать зацепки

segfault, 2024

PostgresPro

- Начинаем искать зацепки
- Вводные
  - Дамп памяти (coredump)
  - Конечное состояние регистров
  - Ассемблеровские инструкции

segfault, 2024

PostgresPro

- Начинаем искать зацепки
- Вводные
  - Дамп памяти (coredump)
  - Конечное состояние регистров
  - Ассемблеровские инструкции
- Поиск непротиворечивого сценария

segfault, 2024

PostgresPro

- Начинаем искать зацепки
- Вводные
  - Дамп памяти (coredump)
  - Конечное состояние регистров
  - Ассемблеровские инструкции
- Поиск непротиворечивого сценария
- Никаких вариантов, кроме одного, но нереального

segfault, 2024

PostgresPro

```
#0  0×000055edf40f20d0 in ?? ()
#1  0×000055edf3f1a370 in hash_search (hashp=0×55edf5c458f8, keyPtr=0×7ffc7a7781ec, action=HASH_FIND, foundPtr=0×0)
```

```
0×000055edf3f5870b <+123>:   cmp    $0×b,%r10d
0×000055edf3f5870f <+127>:   ja     0×55edf3f586b0 <hash_bytes+32>
0×000055edf3f58711 <+129>:   lea    0×1999b8(%rip),%rdi
0×000055edf3f58718 <+136>:   movslq (%rdi,%r10,4),%rsi
0×000055edf3f5871c <+140>:   add    %rdi,%rsi
0×000055edf3f5871f <+143>:   jmp    *%rsi
0×000055edf3f58721 <+145>:   nopl   0×0(%rax)
```

segfault, 2024

PostgresPro

```
#0  0×000055edf40f20d0 in ?? ()
#1  0×000055edf3f1a370 in hash_search (hashp=0×55edf5c458f8, keyPtr=0×7ffc7a7781ec, action=HASH_FIND, foundPtr=0×0)
```

```
0×000055edf3f5870b <+123>:   cmp    $0×b,%r10d
0×000055edf3f5870f <+127>:   ja     0×55edf3f586b0 <hash_bytes+32>
0×000055edf3f58711 <+129>:   lea    0×1999b8(%rip),%rdi
0×000055edf3f58718 <+136>:   movslq (%rdi,%r10,4),%rsi
0×000055edf3f5871c <+140>:   add    %rdi,%rsi
0×000055edf3f5871f <+143>:   jmp    *%rsi
0×000055edf3f58721 <+145>:   nopl   0×0(%rax)
```

segfault, 2024

PostgresPro

```
#0  0×000055edf40f20d0 in ?? ()
#1  0×000055edf3f1a370 in hash_search (hashp=0×55edf5c458f8, keyPtr=0×7ffc7a7781ec, action=HASH_FIND, foundPtr=0×0)
```

```
0×000055edf3f5870b <+123>:   cmp    $0×b,%r10d
0×000055edf3f5870f <+127>:   ja     0×55edf3f586b0 <hash_bytes+32>
0×000055edf3f58711 <+129>:   lea    0×1999b8(%rip),%rdi
0×000055edf3f58718 <+136>:   movslq (%rdi,%r10,4),%rsi
0×000055edf3f5871c <+140>:   add    %rdi,%rsi
0×000055edf3f5871f <+143>:   jmp    *%rsi
0×000055edf3f58721 <+145>:   nopl   0×0(%rax)
```

segfault, 2024

PostgresPro

```
#0  0×000055edf40f20d0 in ?? ()
#1  0×000055edf3f1a370 in hash_search (hashp=0×55edf5c458f8, keyPtr=0×7ffc7a7781ec, action=HASH_FIND, foundPtr=0×0)
```

```
0×000055edf3f5870b <+123>:   cmp    $0×b,%r10d
0×000055edf3f5870f <+127>:   ja     0×55edf3f586b0 <hash_bytes+32>
0×000055edf3f58711 <+129>:   lea    0×1999b8(%rip),%rdi
0×000055edf3f58718 <+136>:   movslq (%rdi,%r10,4),%rsi
0×000055edf3f5871c <+140>:   add    %rdi,%rsi
0×000055edf3f5871f <+143>:   jmp    *%rsi
0×000055edf3f58721 <+145>:   nopl   0×0(%rax)
```

segfault, 2024

PostgresPro

```
0×000055edf3f5870b <+123>:   cmp    $0×b,%r10d
0×000055edf3f5870f <+127>:   ja     0×55edf3f586b0 <hash_bytes+32>
0×000055edf3f58711 <+129>:   lea    0×1999b8(%rip),%rdi      # %rdi = (%rip+0×1999b8)      | %rdi = 0×55edf40f20d0
0×000055edf3f58718 <+136>:   movslq (%rdi,%r10,4),%rsi       # %rsi = mem(%rdi+(%r10*4))   | %r10 = 0
0×000055edf3f5871c <+140>:   add    %rdi,%rsi                # %rsi = %rsi + %rdi          | %rsi = 0×55edf40f20d0
0×000055edf3f5871f <+143>:   jmp    *%rsi                    # jump to %rsi                | segfault
0×000055edf3f58721 <+145>:   nopl   0×0(%rax)
```

segfault, 2024

PostgresPro

```
0x000055edf3f5870b <+123>:   cmp    $0xb,%r10d
0x000055edf3f5870f <+127>:   ja     0x55edf3f586b0 <hash_bytes+32>
0x000055edf3f58711 <+129>:   lea    0x1999b8(%rip),%rdi      # %rdi = (%rip+0x1999b8)      | %rdi = 0x55edf40f20d0
0x000055edf3f58718 <+136>:   movslq (%rdi,%r10,4),%rsi       # %rsi = mem(%rdi+(%r10*4))   | %r10 = 0
0x000055edf3f5871c <+140>:   add    %rdi,%rsi                # %rsi = %rsi + %rdi          | %rsi = 0x55edf40f20d0
0x000055edf3f5871f <+143>:   jmp    *%rsi                    # jump to %rsi                | segfault
0x000055edf3f58721 <+145>:   nopl   0x0(%rax)
```

segfault, 2024

PostgresPro

```
0×000055edf3f5870b <+123>:   cmp    $0×b,%r10d
0×000055edf3f5870f <+127>:   ja     0×55edf3f586b0 <hash_bytes+32>
0×000055edf3f58711 <+129>:   lea    0×1999b8(%rip),%rdi      # %rdi = (%rip+0×1999b8)      | %rdi = 0×55edf40f20d0
0×000055edf3f58718 <+136>:   movslq (%rdi,%r10,4),%rsi       # %rsi = mem(%rdi+(%r10*4))   | %r10 = 0
0×000055edf3f5871c <+140>:   add    %rdi,%rsi                # %rsi = %rsi + %rdi          | %rsi = 0×55edf40f20d0
0×000055edf3f5871f <+143>:   jmp    *%rsi                    # jump to %rsi                | segfault
0×000055edf3f58721 <+145>:   nopl   0×0(%rax)
```

segfault, 2024

PostgresPro

```
0×000055edf3f5870b <+123>:   cmp    $0×b,%r10d
0×000055edf3f5870f <+127>:   ja     0×55edf3f586b0 <hash_bytes+32>
0×000055edf3f58711 <+129>:   lea    0×1999b8(%rip),%rdi      # %rdi = (%rip+0×1999b8)      | %rdi = 0×55edf40f20d0
0×000055edf3f58718 <+136>:   movslq (%rdi,%r10,4),%rsi       # %rsi = mem(%rdi+(%r10*4))   | %r10 = 0
0×000055edf3f5871c <+140>:   add    %rdi,%rsi                # %rsi = %rsi + %rdi          | %rsi = 0×55edf40f20d0
0×000055edf3f5871f <+143>:   jmp    *%rsi                    # jump to %rsi                | segfault
0×000055edf3f58721 <+145>:   nopl   0×0(%rax)
```

segfault, 2024

PostgresPro

```
0×000055edf3f5870b <+123>:   cmp    $0×b,%r10d
0×000055edf3f5870f <+127>:   ja     0×55edf3f586b0 <hash_bytes+32>
0×000055edf3f58711 <+129>:   lea    0×1999b8(%rip),%rdi      # %rdi = (%rip+0×1999b8)      | %rdi = 0×55edf40f20d0
0×000055edf3f58718 <+136>:   movslq (%rdi,%r10,4),%rsi       # %rsi = mem(%rdi+(%r10*4))   | %r10 = 0
0×000055edf3f5871c <+140>:   add    %rdi,%rsi                # %rsi = %rsi + %rdi          | %rsi = 0×55edf40f20d0
0×000055edf3f5871f <+143>:   jmp    *%rsi                    # jump to %rsi                | segfault
0×000055edf3f58721 <+145>:   nopl   0×0(%rax)
```

segfault, 2024

PostgresPro

```
0×000055edf3f5870b <+123>:   cmp    $0×b,%r10d
0×000055edf3f5870f <+127>:   ja     0×55edf3f586b0 <hash_bytes+32>
0×000055edf3f58711 <+129>:   lea    0×1999b8(%rip),%rdi      # %rdi = (%rip+0×1999b8)      | %rdi = 0×55edf40f20d0
0×000055edf3f58718 <+136>:   movslq (%rdi,%r10,4),%rsi       # %rsi = mem(%rdi+(%r10*4))   | %r10 = 0
0×000055edf3f5871c <+140>:   add    %rdi,%rsi                # %rsi = %rsi + %rdi          | %rsi = 0×55edf40f20d0
0×000055edf3f5871f <+143>:   jmp    *%rsi                    # jump to %rsi                | segfault
0×000055edf3f58721 <+145>:   nopl   0×0(%rax)
```

segfault, 2024

PostgresPro

```
0×000055edf3f5870b <+123>:   cmp    $0×b,%r10d
0×000055edf3f5870f <+127>:   ja     0×55edf3f586b0 <hash_bytes+32>
0×000055edf3f58711 <+129>:   lea    0×1999b8(%rip),%rdi      # %rdi = (%rip+0×1999b8)      | %rdi = 0×55edf40f20d0
0×000055edf3f58718 <+136>:   movslq (%rdi,%r10,4),%rsi       # %rsi = mem(%rdi+(%r10*4))   | %r10 = 0
0×000055edf3f5871c <+140>:   add    %rdi,%rsi                # %rsi = %rsi + %rdi          | %rsi = 0×55edf40f20d0
0×000055edf3f5871f <+143>:   jmp    *%rsi                    # jump to %rsi                | segfault
0×000055edf3f58721 <+145>:   nopl   0×0(%rax)
```

segfault, 2024

PostgresPro

- Процессор зачитал значение равное **0** из адреса %rdi+%r10*4
- Должен был зачитать **0xffe667a7**

```
(gdb) p/x *(unsigned*)0×55edf40f20d0@12
$9 = {0×ffe667a7, ... }
```

segfault, 2024

PostgresPro

- Процессор зачитал значение равное **0** из адреса %rdi+%r10*4
- Должен был зачитать **0xffe667a7**

```
(gdb) p/x *(unsigned*)0×55edf40f20d0@12
$9 = {0×ffe667a7, ... }
```

segfault, 2024

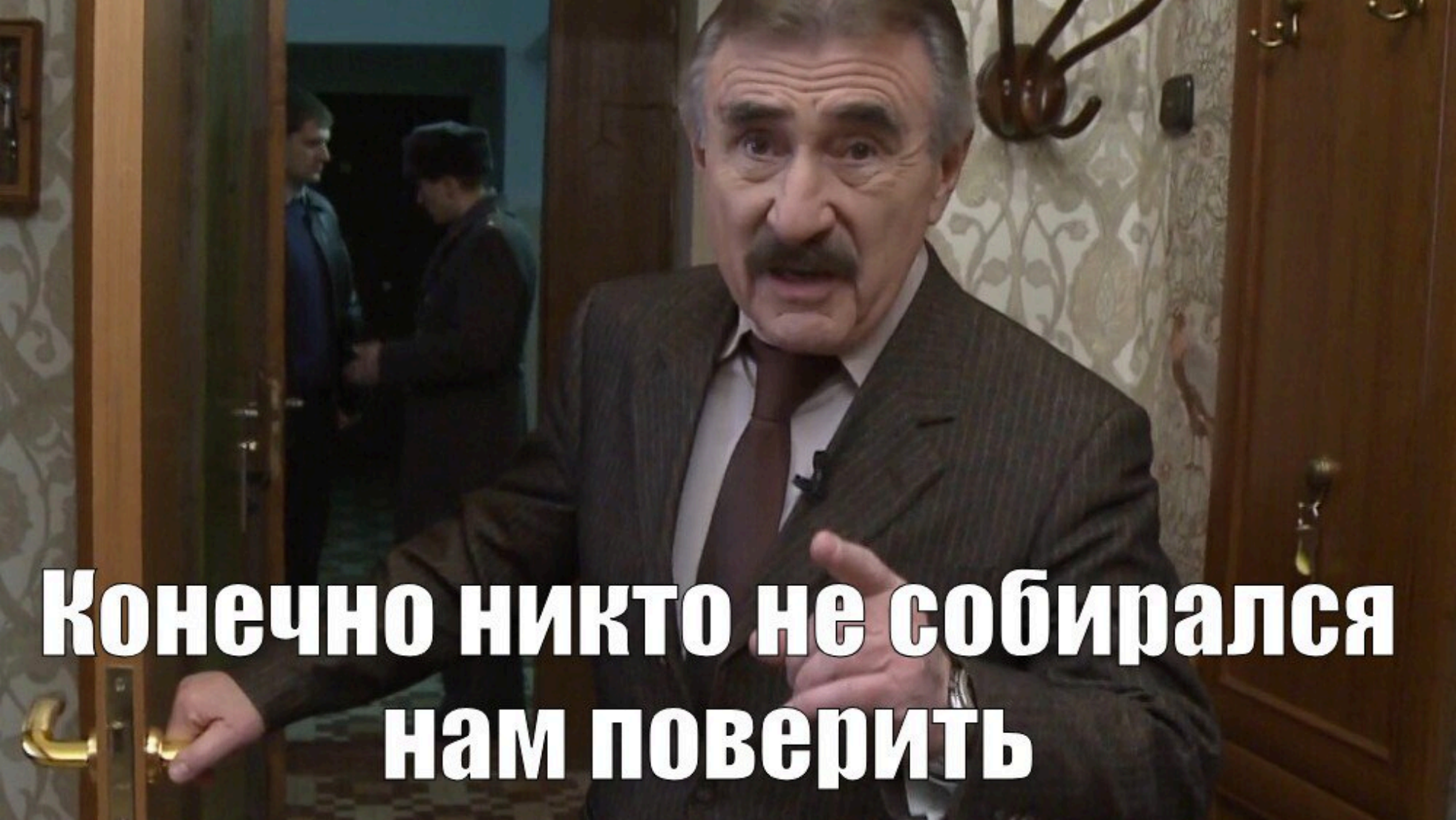
PostgresPro

- 31 декабря, 2:30 ночи: документ с анализом

segfault, 2024

PostgresPro

- 31 декабря, 2:30 ночи: документ с анализом
- "Главное - мы видим аномалию в работе системы, по какой-то причине процессор зачитал нулевое значения из ненулевой памяти (которая постоянная, никогда не меняется)"


Конечно никто не собирался
нам поверить

segfault, 2024

PostgresPro

- Документ ушёл в ТП вендора ОС

segfault, 2024

PostgresPro

- Документ ушёл в ТП вендора ОС
- Ответ

segfault, 2024

PostgresPro

- Документ ушёл в ТП вендора ОС
- Ответ
  - **"Причины SIGSEGV во всех случаях (внезапно!) не внешние, а со стороны userspace, вызванные процессом postgres"**

segfault, 2024

PostgresPro

- Документ ушёл в ТП вендора ОС
- Ответ
  - **"Причины SIGSEGV во всех случаях (внезапно!) не внешние, а со стороны userspace, вызванные процессом postgres"**
- Так заканчивался 2024 год...

segfault, **2025**

segfault, 2025

PostgresPro

- Новые падения (6 января и т.д.) - случайный места, случайные программы (sshd)

segfault, 2025

PostgresPro

- Новые падения (6 января и т.д.) - случайный места, случайные программы (sshd)
- Больше доказательство о нулевой памяти

segfault, 2025

PostgresPro

- Новые падения (6 января и т.д.) - случайный места, случайные программы (sshd)
- Больше доказательство о нулевой памяти
- Гипотезы рождались как грибы

Гипотезы

PostgresPro

- Библиотеки сжатия

Гипотезы

PostgresPro

- Библиотеки сжатия
- Программные raid контроллеры

Гипотезы

PostgresPro

- Библиотеки сжатия
- Программные raid контроллеры
- Прошивки и микрокоды

Гипотезы

PostgresPro

- Библиотеки сжатия
- Программные raid контроллеры
- Прошивки и микрокоды
- FPU

Гипотезы

PostgresPro

- Библиотеки сжатия
- Программные raid контроллеры
- Прошивки и микрокоды
- FPU
- Гипотез много, фактов ноль

Гипотезы

PostgresPro

- Библиотеки сжатия
- Программные raid контроллеры
- Прошивки и микрокоды
- FPU
- Гипотез много, фактов ноль
- **"It’s a capital mistake to theorize without facts"** (c) Артур Конан Дойль

segfault, 2025

PostgresPro

- Падения продолжались (28 января, 7 февраля)

segfault, 2025

PostgresPro

- Падения продолжались (28 января, 7 февраля)
- **rasdaemon**: RAS (Reliability, Availability and Serviceability) logging tool

segfault, 2025

PostgresPro

- Падения продолжались (28 января, 7 февраля)
- **rasdaemon**: RAS (Reliability, Availability and Serviceability) logging tool
- Тук-тук, у вас битые планки памяти!

segfault, 2025

PostgresPro

- Падения продолжались (28 января, 7 февраля)
- **rasdaemon**: RAS (Reliability, Availability and Serviceability) logging tool
- Тук-тук, у вас битые планки памяти!
- Эх, опять мимо

segfault, 2025

PostgresPro

- Падения продолжались (28 января, 7 февраля)
- **rasdaemon**: RAS (Reliability, Availability and Serviceability) logging tool
- Тук-тук, у вас битые планки памяти!
- Эх, опять мимо
- Падали разные сервера

segfault, 2025

PostgresPro

- Падения продолжались (28 января, 7 февраля)
- **rasdaemon**: RAS (Reliability, Availability and Serviceability) logging tool
- Тук-тук, у вас битые планки памяти!
- Эх, опять мимо
- Падали разные сервера
- Делали анализ, писали отчёты, всегда одно и тоже - **обнуление памяти**!

Сложить лапки

Поиграться со стендом

segfault, 2025

PostgresPro

- Стенд от производителя железа

segfault, 2025

PostgresPro

- Стенд от производителя железа
- **Спасибо** за оперативность

segfault, 2025

PostgresPro

- Стенд от производителя железа
- **Спасибо** за оперативность
- Одно из падений: перестроение индексов в базе данных (rebuild index)

segfault, 2025

PostgresPro

- Стенд от производителя железа
- **Спасибо** за оперативность
- Одно из падений: перестроение индексов в базе данных (rebuild index)
- Делаем большую синтетическую базу

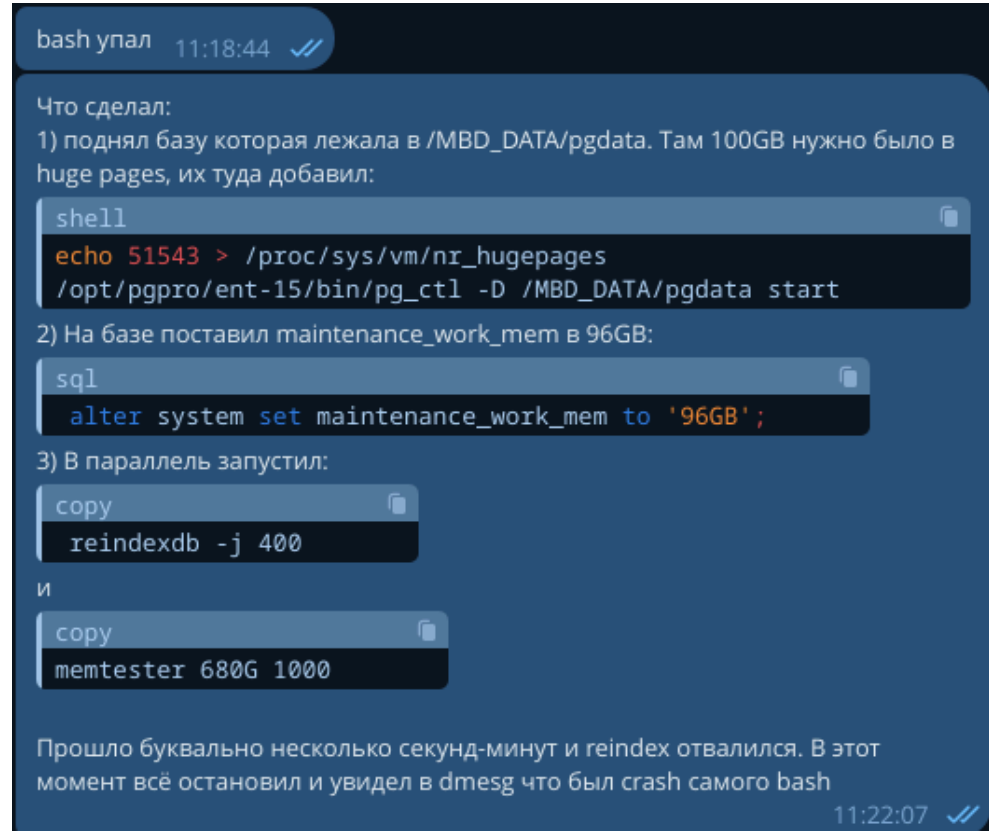
segfault, 2025

PostgresPro

- Стенд от производителя железа
- **Спасибо** за оперативность
- Одно из падений: перестроение индексов в базе данных (rebuild index)
- Делаем большую синтетическую базу
- Добавляем проверку памяти (memtester)

- 25 февраля

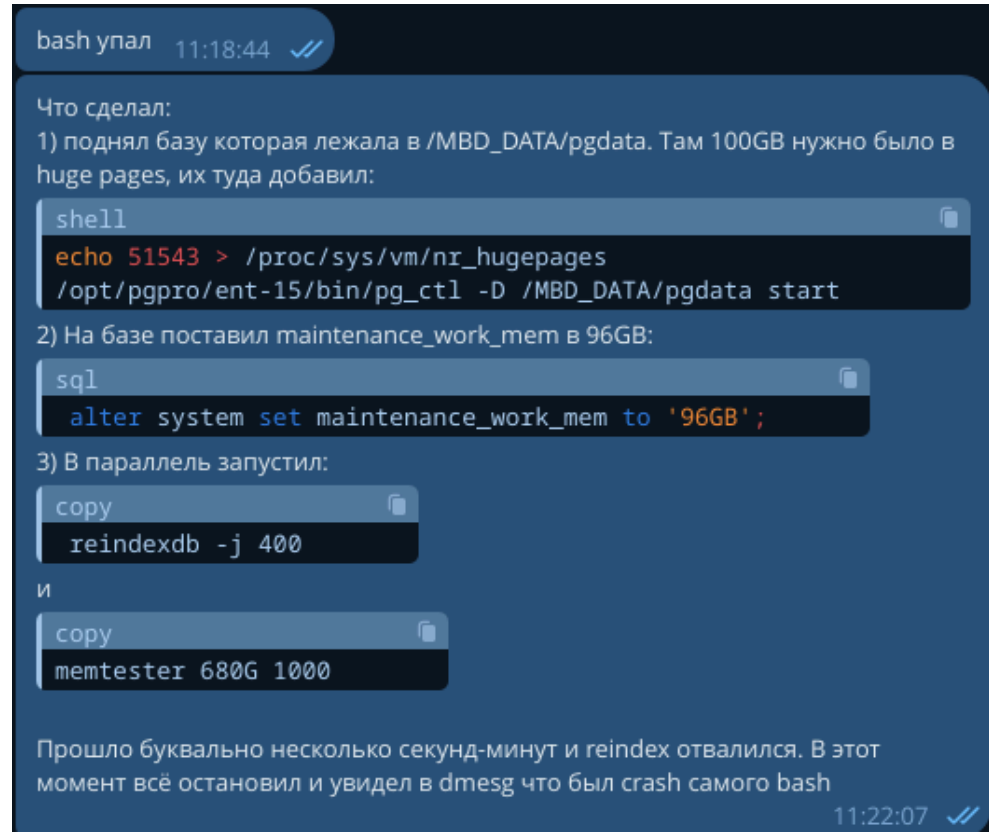
```
reindexdb -j 400
memtester 680G 1000
```



- 25 февраля

```
reindexdb -j 400
memtester 680G 1000
```

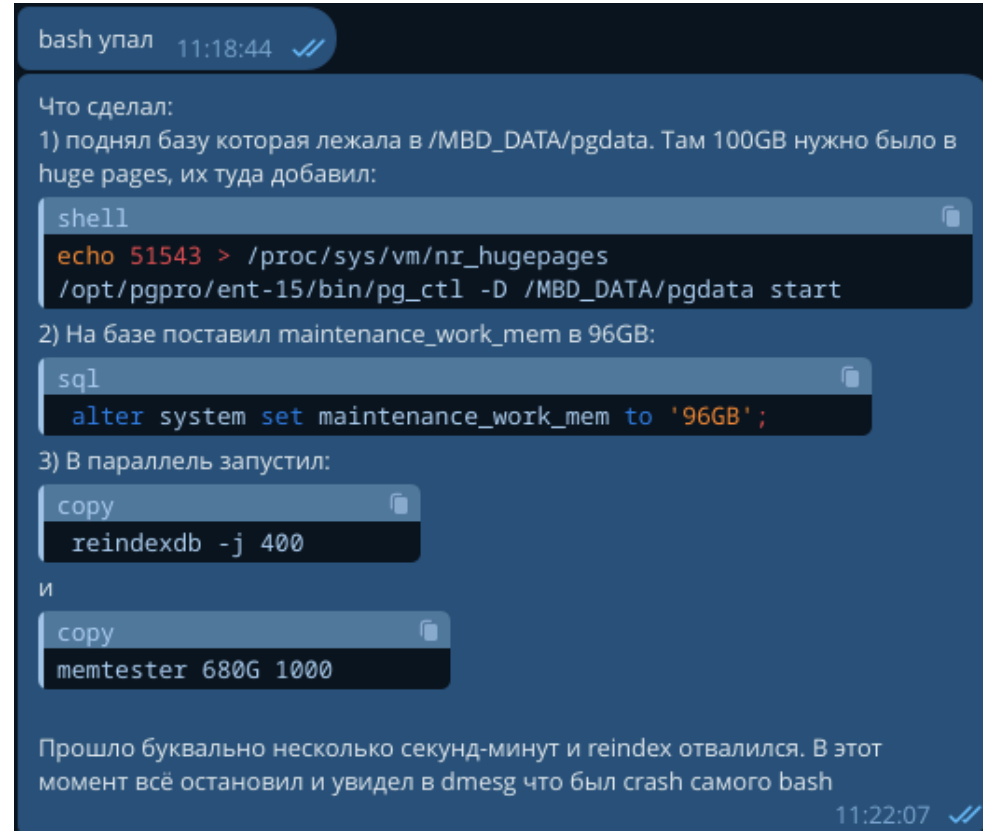
- 400 процессов перестроения индексов



- 25 февраля

```
reindexdb -j 400
memtester 680G 1000
```

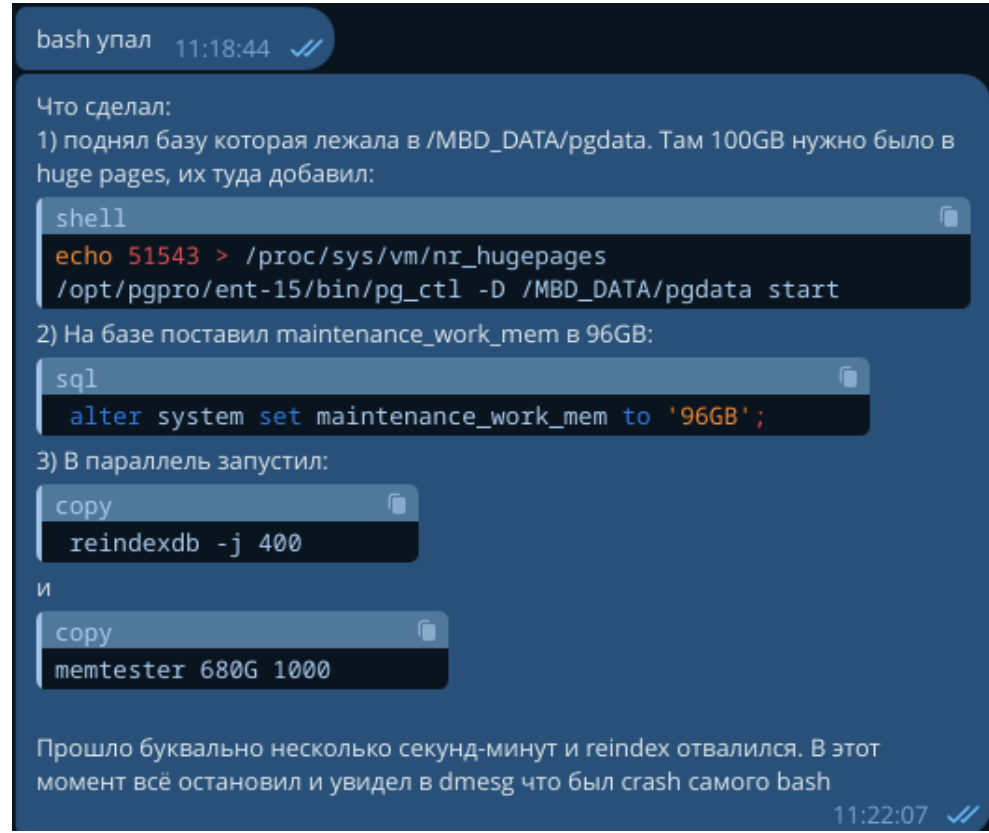
- 400 процессов перестроения индексов
- 1000 итераций проверки 680 гигабайт памяти



- 25 февраля

```
reindexdb -j 400
memtester 680G 1000
```

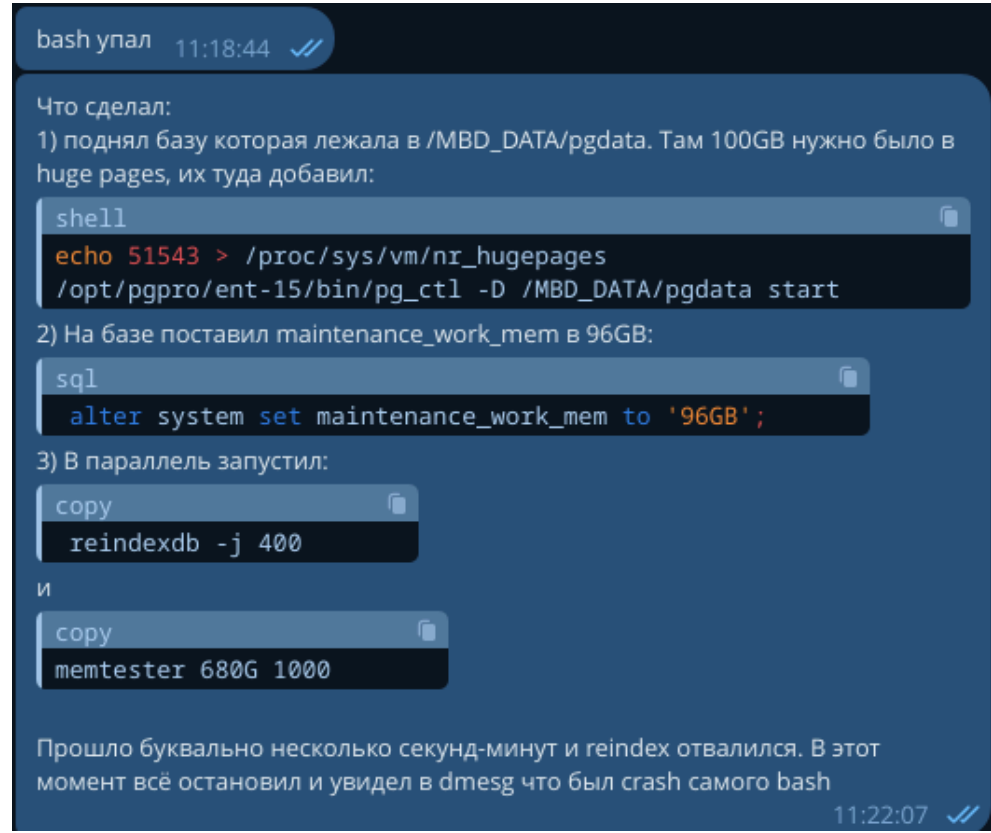
- 400 процессов перестроения индексов
- 1000 итераций проверки 680 гигабайт памяти
- Упал не postgres, а bash



- 25 февраля

```
reindexdb -j 400
memtester 680G 1000
```

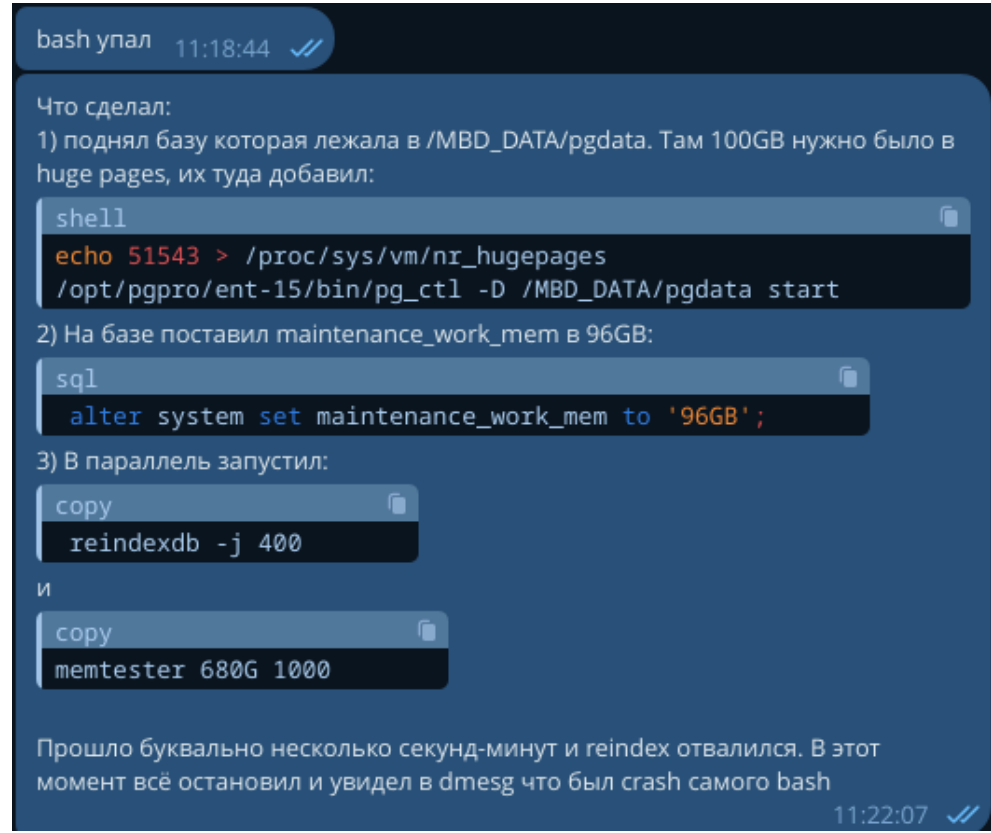
- 400 процессов перестроения индексов
- 1000 итераций проверки 680 гигабайт памяти
- Упал не postgres, а bash
- Воспроизводимый кейс



- 25 февраля

```
reindexdb -j 400
memtester 680G 1000
```

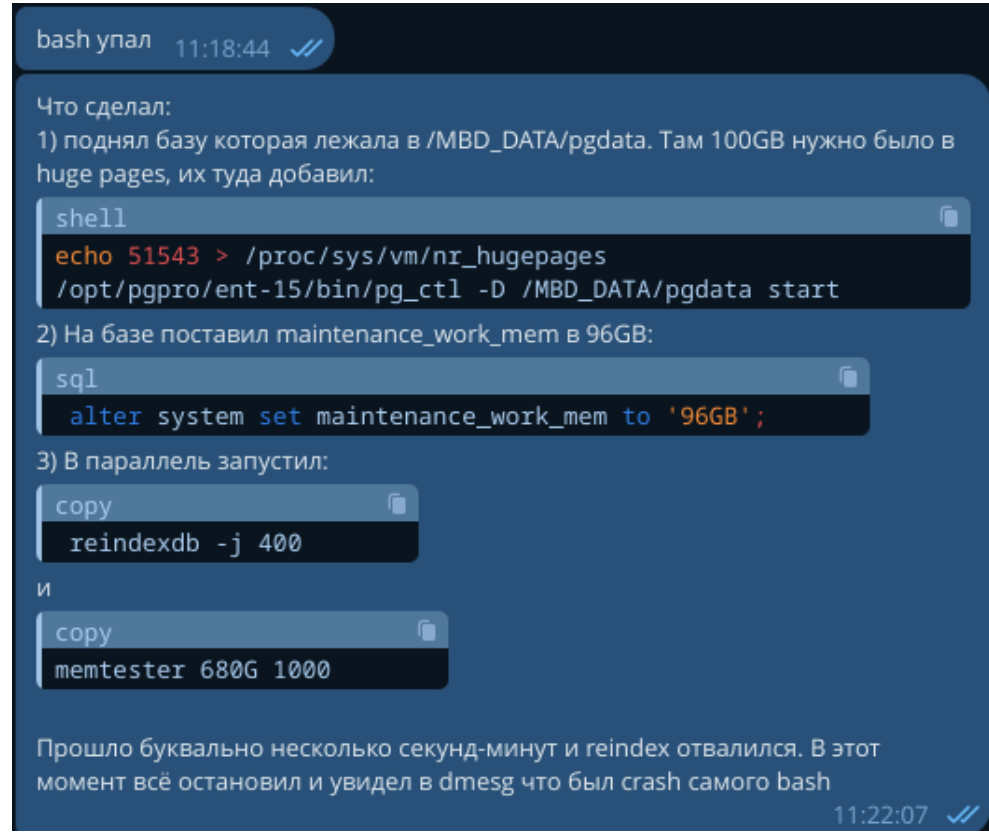
- 400 процессов перестроения индексов
- 1000 итераций проверки 680 гигабайт памяти
- Упал не postgres, а bash
- Воспроизводимый кейс
- Маленькая победа!



- 25 февраля

```
reindexdb -j 400
memtester 680G 1000
```

- 400 процессов перестроения индексов
- 1000 итераций проверки 680 гигабайт памяти
- Упал не postgres, а bash
- Воспроизводимый кейс
- Маленькая победа!
- Но что дальше?



segfault, 2025

PostgresPro

- Инструкция для проверки нового оборудования

segfault, 2025

PostgresPro

- Инструкция для проверки нового оборудования
- Глобальное обновление железа (Апрель-Май)

segfault, 2025

PostgresPro

- Инструкция для проверки нового оборудования
- Глобальное обновление железа (Апрель-Май)
  - Смена ОС

segfault, 2025

PostgresPro

- Инструкция для проверки нового оборудования
- Глобальное обновление железа (Апрель-Май)
  - Смена ОС
  - Увеличение памяти с 1TiB до 4TiB

segfault, 2025

PostgresPro

- Инструкция для проверки нового оборудования
- Глобальное обновление железа (Апрель-Май)
  - Смена ОС
  - Увеличение памяти с 1TiB до 4TiB
  - Замена software RAID на hardware RAID

segfault, 2025

PostgresPro

- Инструкция для проверки нового оборудования
- Глобальное обновление железа (Апрель-Май)
  - Смена ОС
  - Увеличение памяти с 1TiB до 4TiB
  - Замена software RAID на hardware RAID
- Ура! Никаких больше падений

segfault, 2025

PostgresPro

- Инструкция для проверки нового оборудования
- Глобальное обновление железа (Апрель-Май)
  - Смена ОС
  - Увеличение памяти с 1TiB до 4TiB
  - Замена software RAID на hardware RAID
- Ура! Никаких больше падений
- Успешный выход проекта в мае-июне

segfault, 2025

PostgresPro

- Инструкция для проверки нового оборудования
- Глобальное обновление железа (Апрель-Май)
  - Смена ОС
  - Увеличение памяти с 1TiB до 4TiB
  - Замена software RAID на hardware RAID
- Ура! Никаких больше падений
- Успешный выход проекта в мае-июне
- **Увы, никто тогда не знал что причина была не в железе...**

"Конец, конец… Концы в воду!"


"Иди, и без ёлки не возвращайся"

Осень 2025, снова жалобы

PostgresPro

- Различные клиенты

Осень 2025, снова жалобы

PostgresPro

- Различные клиенты
- Различное железо

Осень 2025, снова жалобы

PostgresPro

- Различные клиенты
- Различное железо
- Различные операционные системы

Осень 2025, снова жалобы

PostgresPro

- Различные клиенты
- Различное железо
- Различные операционные системы
- И не только PostgresPro, но PostgreSQL, Kafka и прочие

Осень 2025, снова жалобы

PostgresPro

- Различные клиенты
- Различное железо
- Различные операционные системы
- И не только PostgresPro, но PostgreSQL, Kafka и прочие
- **А точно ли дефект был в железе?** Ответ: **нет**

Осень 2025, воспроизведение

PostgresPro

- Linux kernel 6.1, XFS, 2 TiB RAM

Осень 2025, воспроизведение

PostgresPro

- Linux kernel 6.1, XFS, 2 TiB RAM
- Неожиданно воспроизвелось на наших серверах

Осень 2025, воспроизведение

PostgresPro

- Linux kernel 6.1, XFS, 2 TiB RAM
- Неожиданно воспроизвелось на наших серверах
- Подкрутили тест

Осень 2025, воспроизведение

PostgresPro

- Linux kernel 6.1, XFS, 2 TiB RAM
- Неожиданно воспроизвелось на наших серверах
- Подкрутили тест
- Воспроизвели на виртуалках с большим количеством vCPU и памяти

Осень 2025, воспроизведение

PostgresPro

- Linux kernel 6.1, XFS, 2 TiB RAM
- Неожиданно воспроизвелось на наших серверах
- Подкрутили тест
- Воспроизвели на виртуалках с большим количеством vCPU и памяти
- Ещё подкрутили тест

Осень 2025, воспроизведение

PostgresPro

- Linux kernel 6.1, XFS, 2 TiB RAM
- Неожиданно воспроизвелось на наших серверах
- Подкрутили тест
- Воспроизвели на виртуалках с большим количеством vCPU и памяти
- Ещё подкрутили тест
- Воспроизвели на маленьких виртуалках (4-8 vCPU)

Осень 2025, воспроизведение

PostgresPro

- Супер! Теперь можем что-нибудь поменять

Осень 2025, воспроизведение

PostgresPro

- Супер! Теперь можем что-нибудь поменять
- ext4 вместо XFS - **не** воспроизводится

Осень 2025, воспроизведение

PostgresPro

- Супер! Теперь можем что-нибудь поменять
- ext4 вместо XFS - **не** воспроизводится
- Debian вместо российской ОС - **не** воспроизводится

Осень 2025, воспроизведение

PostgresPro

- Супер! Теперь можем что-нибудь поменять
- ext4 вместо XFS - **не** воспроизводится
- Debian вместо российской ОС - **не** воспроизводится
- Магия!

Осень 2025, воспроизведение

PostgresPro

- Супер! Теперь можем что-нибудь поменять
- ext4 вместо XFS - **не** воспроизводится
- Debian вместо российской ОС - **не** воспроизводится
- Магия!
- Замена Linux kernel с вендорского на ванильное - **воспроизводится**

Осень 2025, воспроизведение

PostgresPro

- Давайте менять версии Linux kernel

Осень 2025, воспроизведение

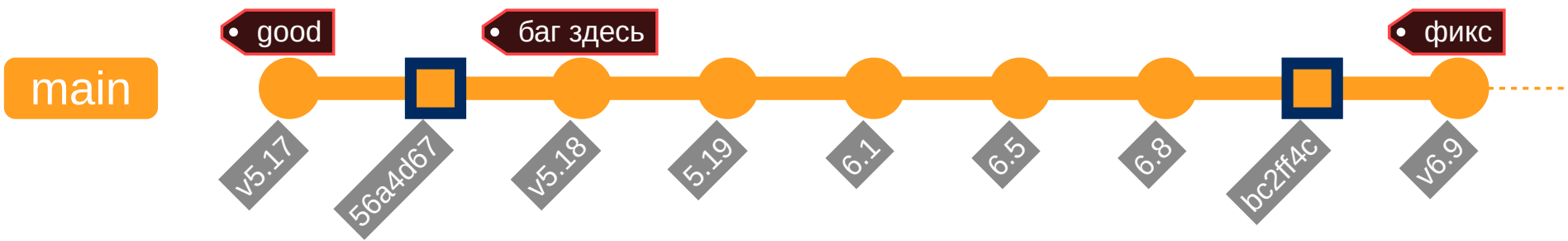
PostgresPro

- Давайте менять версии Linux kernel
- 5.18 — 6.8

Осень 2025, воспроизведение

PostgresPro

- Давайте менять версии Linux kernel
- 5.18 — 6.8
- Поиск git коммита, который добавил дефект


main
good
баг здесь
фикс
v5.17
56a4d67
v5.18
5.19
6.1
6.5
6.8
bc2ff4c
v6.9

Осень 2025, воспроизведение

PostgresPro

- Перебор коммитов

Осень 2025, воспроизведение

PostgresPro

- Перебор коммитов
- Проблемный коммит в драйвере клавиатуры!

Осень 2025, воспроизведение

PostgresPro

- Перебор коммитов
- Проблемный коммит в драйвере клавиатуры!
- Чё-то не то...

Осень 2025, воспроизведение

PostgresPro

- Перебор коммитов
- Проблемный коммит в драйвере клавиатуры!
- Чё-то не то...
- GitHub vs git-bisect

drivers: net: amd: lance: Remove this driver
lunn authored and kuba-moo committed 2 days ago
2fbd04d

drivers: net: 3com: 3c589: Remove this driver
lunn authored and kuba-moo committed 2 days ago
4ff8d06

drivers: net: 3com: 3c574: Remove this driver
lunn authored and kuba-moo committed 2 days ago
a7fbf27

drivers: net: 3com: 3c515: Remove this driver
lunn authored and kuba-moo committed 2 days ago
082b2e0

drivers: net: 3com: 3c509: Remove this driver
lunn authored and kuba-moo committed 2 days ago
91f3a27

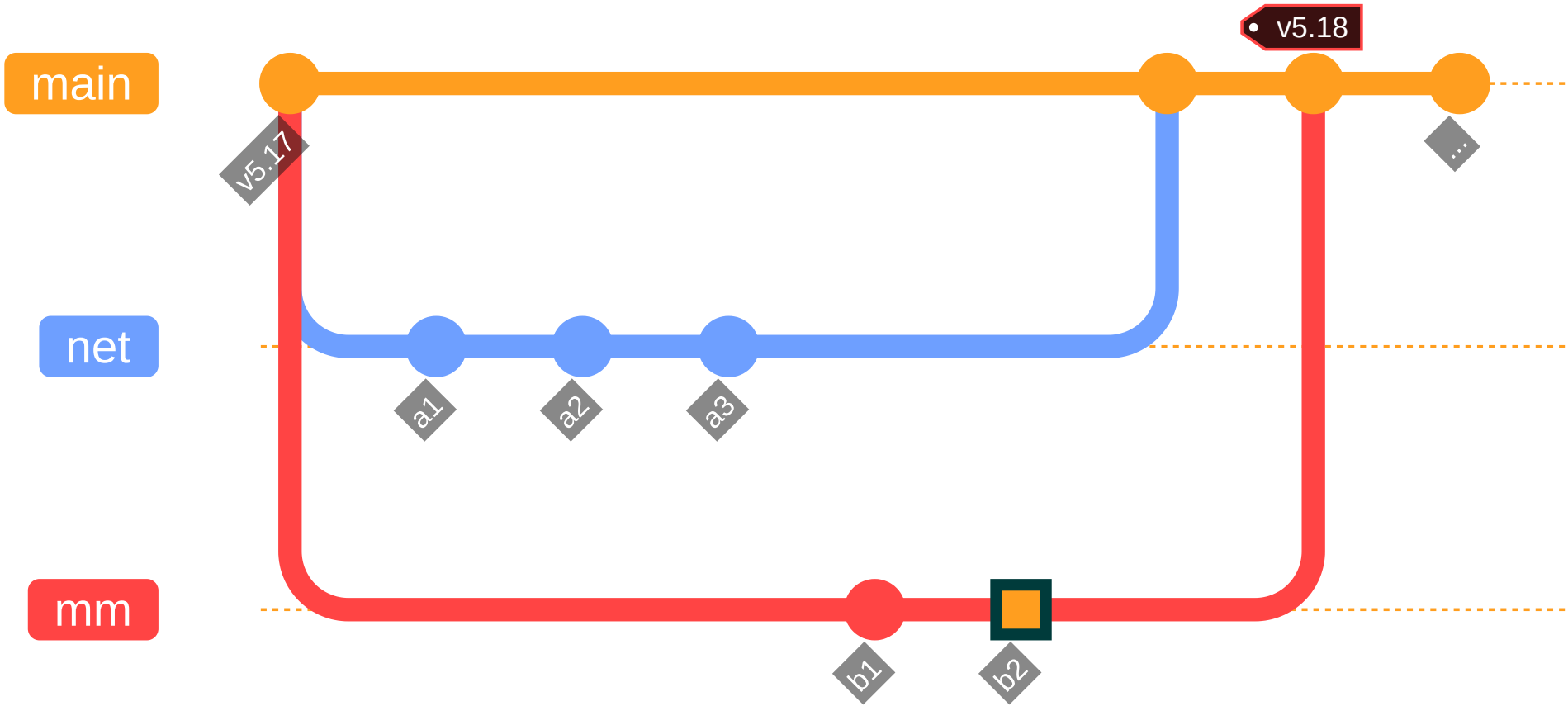
Merge patch series "eventpoll: fix ep_remove() UAF and follow-up cleanup"
brauner committed 2 days ago
Verified
ac8777c

eventpoll: drop vestigial epi->dying flag
brauner committed 2 days ago
Verified
07422c9

eventpoll: drop dead bool return from ep_remove_epi()
brauner committed 2 days ago
Verified
3a4551e

eventpoll: refresh eventpoll_release() fast-path comment
brauner committed 2 days ago
Verified
33e92e9

eventpoll: move f_lock acquisition into ep_remove_file()
brauner committed 2 days ago
Verified
d30deeb


main
net
mm
v5.17
v5.18
...
a1
a2
a3
b1
b2

ata: ahci: Drop low power policy board type	Mario Limonciello <mario.limonciell	2024-02-07 00:1
ata: ahci: do not enable LPM on external ports	Niklas Cassel <cassel@kernel.org>	2024-02-07 00:1
ata: ahci: drop hpriv param from ahci_update_initial_lpm_policy()	Niklas Cassel <cassel@kernel.org>	2024-02-07 00:1
ata: ahci: a hotplug capable port is an external port	Niklas Cassel <cassel@kernel.org>	2024-02-07 00:1
ata: ahci: move marking of external port earlier	Niklas Cassel <cassel@kernel.org>	2024-02-07 00:1
dt-bindings: ata: atmel: remove at91 compact flash documentation	Hari Prasath Gujulan Elango <Hari.P	2024-02-05 13:5
ata: pata_parport: make pata_parport_bus_type const	Ricardo B. Marliere <ricardo@marlie	2024-02-04 18:2
Merge tag 'dma-mapping-6.9-2024-03-11' of git://git.infradead.org/users/hch/dma-mapping	Linus Torvalds <torvalds@linux-fou	2024-03-13 19:2
dma-direct: Leak pages on dma_set_decrypted() failure	Rick Edgecombe <rick.p.edgecombe	2024-02-22 03:1
swiotlb: add debugfs to track swiotlb transient pool usage	ZhangPeng <zhangpeng362@huaw	2024-01-09 10:0
Merge tag 'lsm-pr-20240227' of git://git.kernel.org/pub/scm/linux/kernel/git/pcmoore/lsm	Linus Torvalds <torvalds@linux-fou	2024-02-28 04:0
Merge tag 'iommu-updates-v6.9' of git://git.kernel.org/pub/scm/linux/kernel/git/joro/iommu	Linus Torvalds <torvalds@linux-fou	2024-03-13 19:1
Merge branches 'arm/mediatek', 'arm/renesas', 'arm/smmu', 'x86/vt-d', 'x86/amd' and 'core' into next	Joerg Roedel <jroedel@suse.de>	2024-03-08 11:0
iommu/dma: Document min_align_mask assumption	Robin Murphy <robin.murphy@arm	2024-03-04 15:0
iommu: re-use local fwnode variable in iommu_ops_from_fwnode()	Krzysztof Kozlowski <krzysztof.kozl	2024-02-16 17:4
iommu: constify fwnode in iommu_ops_from_fwnode()	Krzysztof Kozlowski <krzysztof.kozl	2024-02-16 17:4
iommu: constify of_phandle_args in xlate	Krzysztof Kozlowski <krzysztof.kozl	2024-02-16 17:4
iommu: constify pointer to bus_type	Krzysztof Kozlowski <krzysztof.kozl	2024-02-16 17:4
iommu: Make iommu_report_device_fault() return void	Lu Baolu <baolu.lu@linux.intel.com>	2024-02-12 04:2
iommu: Make iopf_group_response() return void	Lu Baolu <baolu.lu@linux.intel.com>	2024-02-12 04:2
iommu: Track iopf group instead of last fault	Lu Baolu <baolu.lu@linux.intel.com>	2024-02-12 04:2
iommu: Improve iopf_queue_remove_device()	Lu Baolu <baolu.lu@linux.intel.com>	2024-02-12 04:2
iommu: Use refcount for fault data access	Lu Baolu <baolu.lu@linux.intel.com>	2024-02-12 04:2
iommu: Refine locking for per-device fault data management	Lu Baolu <baolu.lu@linux.intel.com>	2024-02-12 04:2
iommu: Separate SVA and IOPF	Lu Baolu <baolu.lu@linux.intel.com>	2024-02-12 04:2
iommu: Make iommu_queue_iopf() more generic	Lu Baolu <baolu.lu@linux.intel.com>	2024-02-12 04:2
iommu: Prepare for separating SVA and IOPF	Lu Baolu <baolu.lu@linux.intel.com>	2024-02-12 04:2
iommu: Merge iommu_fault_event and iopf_fault	Lu Baolu <baolu.lu@linux.intel.com>	2024-02-12 04:2
iommu: Remove iommu_[un]register_device_fault_handler()	Lu Baolu <baolu.lu@linux.intel.com>	2024-02-12 04:2
iommu: Merge iopf_device_param into iommu_fault_param	Lu Baolu <baolu.lu@linux.intel.com>	2024-02-12 04:2
iommu: Cleanup iopf data structure definitions	Lu Baolu <baolu.lu@linux.intel.com>	2024-02-12 04:2
iommu: Remove unrecoverable fault data	Lu Baolu <baolu.lu@linux.intel.com>	2024-02-12 04:2
iommu/arm-smmu-v3: Remove unrecoverable faults reporting	Lu Baolu <baolu.lu@linux.intel.com>	2024-02-12 04:2

selinux: reduce the object class calculations at inode init time	Paul Moore <paul@paul-moore.com	2024-01-19 04:4
Merge tag 'net-next-6.9' of git://git.kernel.org/pub/scm/linux/kernel/git/netdev/net-next	Linus Torvalds <torvalds@linux-fou	2024-03-13 03:4
Merge git://git.kernel.org/pub/scm/linux/kernel/git/netdev/net	Jakub Kicinski <kuba@kernel.org>	2024-03-12 06:3
nfp: flower: handle acti_netdevs allocation failure	Duoming Zhou <duoming@zju.edu.	2024-03-08 17:2
r8152: fix unknown device for choose_configuration	Hayes Wang <hayeswang@realtek.c	2024-03-08 10:5
devlink: Fix length of eswitch inline-mode	William Tu <witu@nvidia.com>	2024-03-10 19:4
net: phy: dp83822: Fix RGMII TX delay configuration	Tim Pambor <tp@osasysteme.de>	2024-03-05 14:0
Merge branch '1GbE' of git://git.kernel.org/pub/scm/linux/kernel/git/tnguy/net-queue	Jakub Kicinski <kuba@kernel.org>	2024-03-08 22:3
ice: fix stats being updated by way too large values	Przemek Kitszel <przemyslaw.kitsze	2024-02-27 17:3
igb: Fix missing time sync events	Vinicius Costa Gomes <vinicius.gom	2024-02-21 02:5
igc: Fix missing time sync events	Vinicius Costa Gomes <vinicius.gom	2024-02-21 02:5
net/rds: fix WARNING in rds_conn_connect_if_down	Edward Adam Davis <eadavis@qq.c	2024-03-05 03:1
dpll: fix dpll_xa_ref_*_del() for multiple registrations	Jiri Pirko <jiri@nvidia.com>	2024-03-06 18:1
net: phy: fix phy_get_internal_delay accessing an empty array	Kévin L'hôpital <kevin.lhopital@savc	2024-03-07 14:1
octeontx2-af: Fix devlink params	Sunil Goutham <sgoutham@marvel	2024-03-07 13:5
net: ip_tunnel: make sure to pull inner header in ip_tunnel_rcv()	Eric Dumazet <edumazet@google.c	2024-03-07 13:0
ipv6: fib6_rules: flush route cache when rule is changed	Shiming Cheng <shiming.cheng@m	2024-03-07 13:0
net: dsa: microchip: make sure drive strength configuration is not lost by soft reset	Oleksij Rempel <o.rempel@pengutr	2024-03-04 16:5
Merge branch 'nexthop-fix-two-nexthop-group-statistics-issues'	Jakub Kicinski <kuba@kernel.org>	2024-03-12 06:3
nexthop: Fix splat with CONFIG_DEBUG_PREEMPT=y	Ido Schimmel <idosch@nvidia.com>	2024-03-11 19:2
nexthop: Fix out-of-bounds access during attribute validation	Ido Schimmel <idosch@nvidia.com>	2024-03-11 19:2
nexthop: Only parse NHA_OP_FLAGS for dump messages that require it	Ido Schimmel <idosch@nvidia.com>	2024-03-11 19:2
nexthop: Only parse NHA_OP_FLAGS for get messages that require it	Ido Schimmel <idosch@nvidia.com>	2024-03-11 19:2
Merge tag 'for-netdev' of https://git.kernel.org/pub/scm/linux/kernel/git/bpf/bpf-next	Jakub Kicinski <kuba@kernel.org>	2024-03-12 04:0
bpf: move sleepable flag from bpf_prog_aux to bpf_prog	Andrii Nakryiko <andrii@kernel.org	2024-03-09 03:4
bpf: hardcode BPF_PROG_PACK_SIZE to 2MB * num_possible_nodes()	Puranjay Mohan <puranjay12@gma	2024-03-11 15:2
selftests/bpf: Add kprobe multi triggering benchmarks	Jiri Olsa <jolsa@kernel.org>	2024-03-12 00:1
Merge branch 'bpf-introduce-bpf-arena'	Andrii Nakryiko <andrii@kernel.org	2024-03-12 01:3
selftests/bpf: Add bpf_arena_htab test.	Alexei Starovoitov <ast@kernel.org>	2024-03-08 04:0
selftests/bpf: Add bpf_arena_list test.	Alexei Starovoitov <ast@kernel.org>	2024-03-08 04:0
selftests/bpf: Add unit tests for bpf_arena_alloc/free_pages	Alexei Starovoitov <ast@kernel.org>	2024-03-08 04:0
bpf: Add helper macro bpf_addr_space_cast()	Alexei Starovoitov <ast@kernel.org>	2024-03-08 04:0

segfault, 2025

PostgresPro

- git bisect: поиск бага по графу

```
git bisect start
git bisect good v5.17     # бага нет
git bisect bad v5.18      # баг есть
git bisect run ./test.sh  # → 56a4d67c264e
```

segfault, 2025

PostgresPro

- Коммит, который сломал (21 марта, 2022): mm/readahead: Switch to page_cache_ra_order
  - Масштаб: 3 файла, +5 / −5 строк

segfault, 2025

PostgresPro

- Коммит, который сломал (21 марта, 2022): mm/readahead: Switch to page_cache_ra_order
  - Масштаб: 3 файла, +5 / −5 строк
- Коммит, который исправил (5 марта, 2024): mm: free folios in a batch in shrink_folio_list()
  - Масштаб: 1 файла, +9 / −11 строк

segfault, 2025

PostgresPro

- Коммит, который сломал (21 марта, 2022): mm/readahead: Switch to page_cache_ra_order
  - Масштаб: 3 файла, +5 / −5 строк
- Коммит, который исправил (5 марта, 2024): mm: free folios in a batch in shrink_folio_list()
  - Масштаб: 1 файла, +9 / −11 строк
- Автор обоих коммитов: Matthew Wilcox, Oracle

segfault, 2025

PostgresPro

- Коммит, который сломал (21 марта, 2022): mm/readahead: Switch to page_cache_ra_order
  - Масштаб: 3 файла, +5 / −5 строк
- Коммит, который исправил (5 марта, 2024): mm: free folios in a batch in shrink_folio_list()
  - Масштаб: 1 файла, +9 / −11 строк
- Автор обоих коммитов: Matthew Wilcox, Oracle
- Знаем коммиты, хотим ли мы знать причину? Ответ: **не хотим**

segfault, 2025

PostgresPro

- Коммит, который сломал (21 марта, 2022): mm/readahead: Switch to page_cache_ra_order
  - Масштаб: 3 файла, +5 / −5 строк
- Коммит, который исправил (5 марта, 2024): mm: free folios in a batch in shrink_folio_list()
  - Масштаб: 1 файла, +9 / −11 строк
- Автор обоих коммитов: Matthew Wilcox, Oracle
- Знаем коммиты, хотим ли мы знать причину? Ответ: **не хотим**
- Знание по ядру есть на уровне исследователя, не программиста

segfault, 2025

PostgresPro

- Но хотим понять, почему на Debian не воспроизводится!

segfault, 2025

PostgresPro

- Но хотим понять, почему на Debian не воспроизводится!
- Ядра одинаковые, но есть и параметры загрузки ядра. Grub

segfault, 2025

PostgresPro

- Но хотим понять, почему на Debian не воспроизводится!
- Ядра одинаковые, но есть и параметры загрузки ядра. Grub
- init_on_free = 0 во всех зарубежных

segfault, 2025

PostgresPro

- Но хотим понять, почему на Debian не воспроизводится!
- Ядра одинаковые, но есть и параметры загрузки ядра. Grub
- init_on_free = 0 во всех зарубежных
- init_on_free = 1 в российских

Проверьте свои серверы

PostgresPro

- Ядро 5.18 — 6.8 + XFS + init_on_free = обратитесь к вендору ОС
- Фикс в 6.9, но не помечен как исправление
- Бэкпорта нет: на LTS-ядрах фикс сам не придёт
- Нужен bug report в bugzilla
- А теперь прикол...

segfault или не segfault?

segfault или не segfault?

PostgresPro

- SIGSEGV - это сигнал
- Сигнал можно обработать
- Пишем свой обработчик

segfault или не segfault?

PostgresPro

```
sigaction_t sigact;

sigact.sa_sigaction = my_segfault_handler;
sigact.sa_flags = SA_RESETHAND | ...;

...

sigaction(SIGSEGV, &sigact, NULL);
```

segfault или не segfault?

PostgresPro

```
sigaction_t sigact;

sigact.sa_sigaction = my_segfault_handler;
sigact.sa_flags = SA_RESETHAND | ... ;

...

sigaction(SIGSEGV, &sigact, NULL);
```

segfault или не segfault?

PostgresPro

```
sigaction_t sigact;

sigact.sa_sigaction = my_segfault_handler;
sigact.sa_flags = SA_RESETHAND | ...;

...

sigaction(SIGSEGV, &sigact, NULL);
```

segfault или не segfault?

PostgresPro

```
sigaction_t sigact;

sigact.sa_sigaction = my_segfault_handler;
sigact.sa_flags = SA_RESETHAND | ...;

...

sigaction(SIGSEGV, &sigact, NULL);
```

segfault или не segfault?

PostgresPro

- Ночью создавался индекс

segfault или не segfault?

PostgresPro

- Ночью создавался индекс
- С утра DBA находит 3 файла от нашего обработчика
  - crash_1764605313879686_pid2130208.log
  - crash_1764605308648131_pid1867399.log
  - crash_1764605308641242_pid2130207.log

segfault или не segfault?

PostgresPro

- Ночью создавался индекс
- С утра DBA находит 3 файла от нашего обработчика
  - crash_1764605313879686_pid2130208.log
  - crash_1764605308648131_pid1867399.log
  - crash_1764605308641242_pid2130207.log
- Упало создание индексов
  - program_invocation_name: postgres: mydb: postgres myuser local CREATE INDEX

segfault или не segfault?

PostgresPro

- ВНИМАНИЕ! Никто не пострадал

segfault или не segfault?

PostgresPro

- ВНИМАНИЕ! Никто не пострадал
- Индекс создался

segfault или не segfault?

PostgresPro

- ВНИМАНИЕ! Никто не пострадал
- Индекс создался
- В логах нет ошибок

segfault или не segfault?

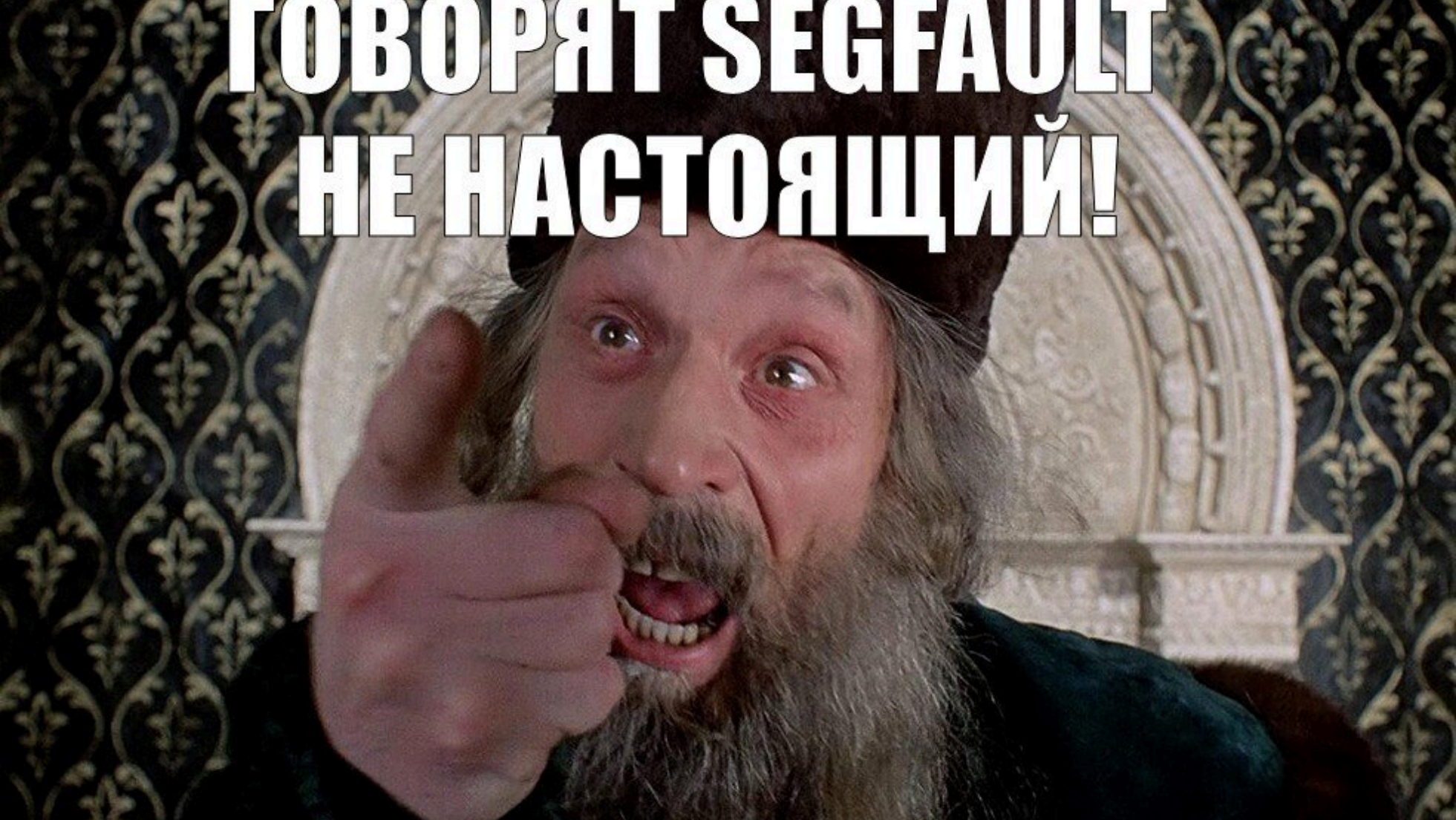
PostgresPro

- ВНИМАНИЕ! Никто не пострадал
- Индекс создался
- В логах нет ошибок
- Процессы не падали, база не перезагружалась!

segfault или не segfault?

PostgresPro

- ВНИМАНИЕ! Никто не пострадал
- Индекс создался
- В логах нет ошибок
- Процессы не падали, база не перезагружалась!
- Индекс валидный!


ГОВОРЯТ SEGFAULT
НЕ НАСТОЯЩИЙ!

Без обработчика

PostgresPro

- Случается SEGFAULT

Без обработчика

PostgresPro

- Случается SEGFAULT
- Инструкция процессора по адресу записанному в регистр RIP

Без обработчика

PostgresPro

- Случается SEGFAULT
- Инструкция процессора по адресу записанному в регистр RIP
- Ядерный обработчик логирует, делает coredump и убивается процесс

С обработчиком

PostgresPro

- Случается SEGFAULT

С обработчиком

PostgresPro

- Случается SEGFAULT
- Вызывается наш обработчик

С обработчиком

PostgresPro

- Случается SEGFAULT
- Вызывается наш обработчик
- Делаем полезное действие

С обработчиком

PostgresPro

- Случается SEGFAULT
- Вызывается наш обработчик
- Делаем полезное действие
- Потом возвращается обработчик по умолчанию (благодаря SA_RESETHAND)

С обработчиком

PostgresPro

- Случается SEGFAULT
- Вызывается наш обработчик
- Делаем полезное действие
- Потом возвращается обработчик по умолчанию (благодаря SA_RESETHAND)
- Операционная система заканчиваем обработку сигнала


ГОРОД ЗАСЫПАЕТ,
ПРОСЫПАЕТСЯ
МАФИЯ...

С обработчиком

PostgresPro

- Повторяется инструкция процессора в тех же самых условиях

С обработчиком

PostgresPro

- Повторяется инструкция процессора в тех же самых условиях
- Значит, повторяется и SIGSEGV / SEGFAULT

С обработчиком

PostgresPro

- Повторяется инструкция процессора в тех же самых условиях
- Значит, повторяется и SIGSEGV / SEGFAULT
- А иногда и не повторяется

С обработчиком

PostgresPro

- Повторяется инструкция процессора в тех же самых условиях
- Значит, повторяется и SIGSEGV / SEGFAULT
- А иногда и не повторяется
- Вторая попытка позже первой (разница до 1 секунды)

С обработчиком

PostgresPro

- Страница пустая только некоторый короткий интервал времени

С обработчиком

PostgresPro

- Страница пустая только некоторый короткий интервал времени
- Структуры памяти восстанавливаются

С обработчиком

PostgresPro

- Страница пустая только некоторый короткий интервал времени
- Структуры памяти восстанавливаются
- Ядерный обработчик не вызывается

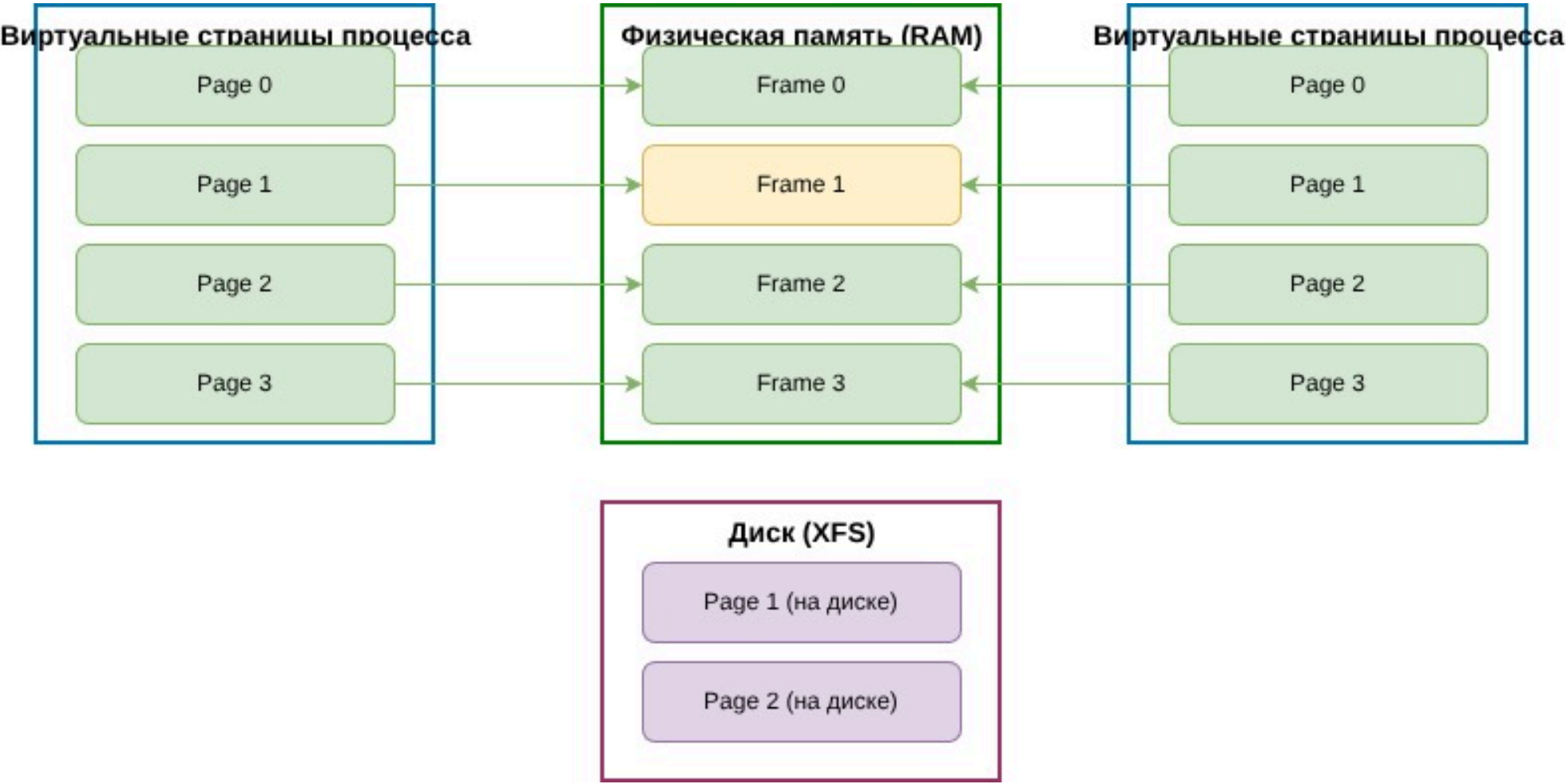
А почему раньше не помогало?

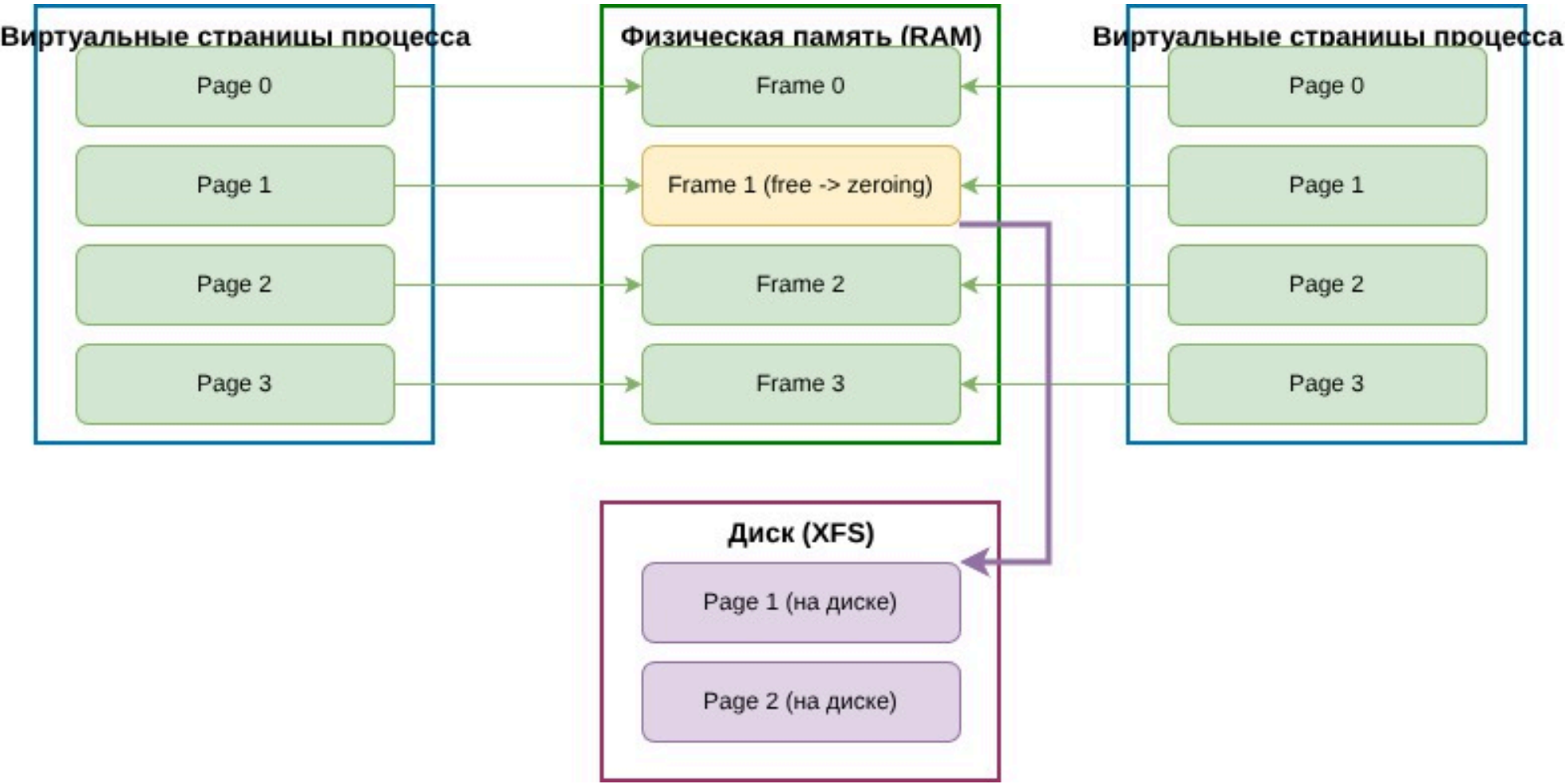
PostgresPro

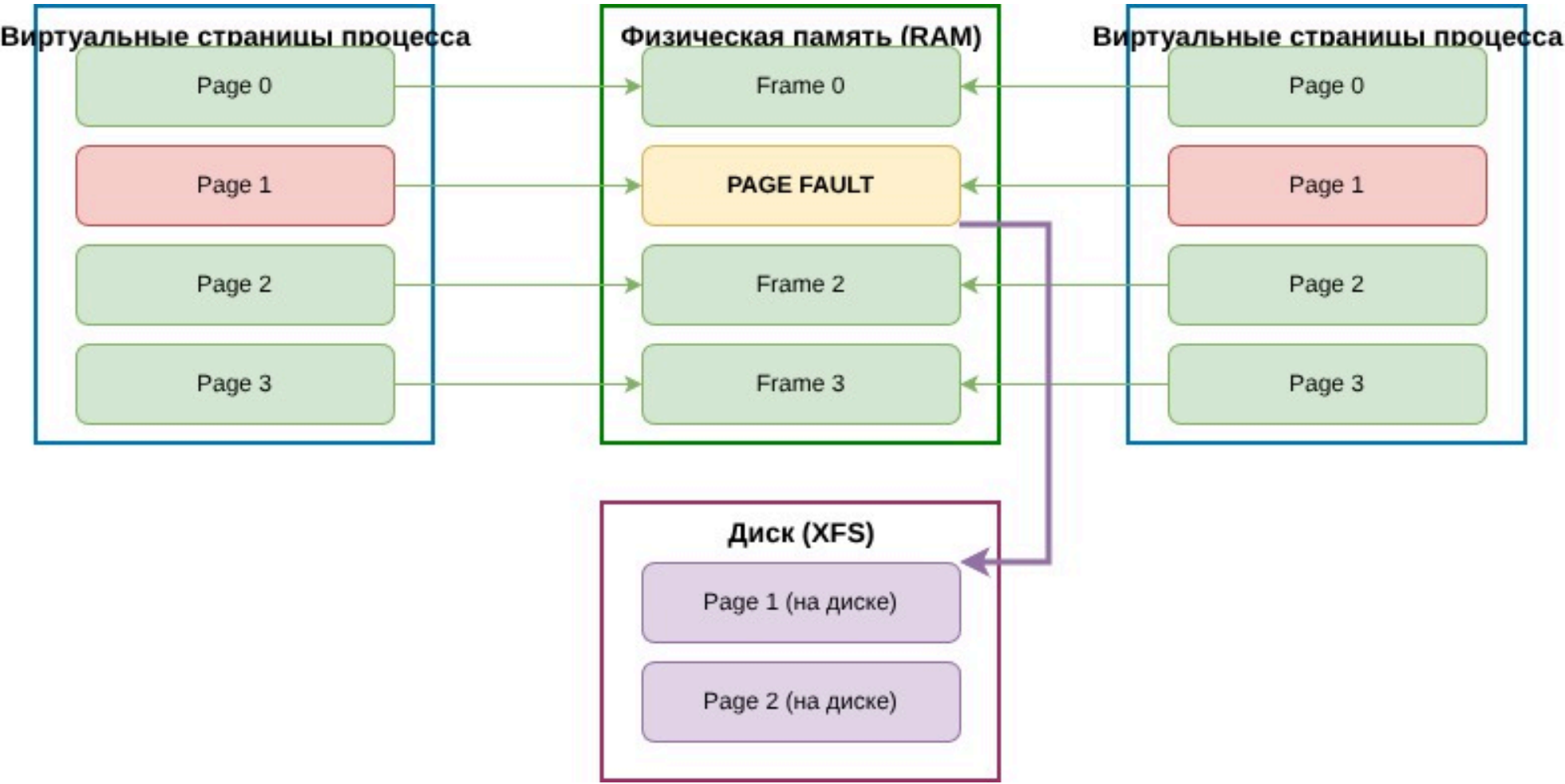
- Обнулялись страницы из сегментов отличных от .text
- "0x00 0x00" валидная инструкция

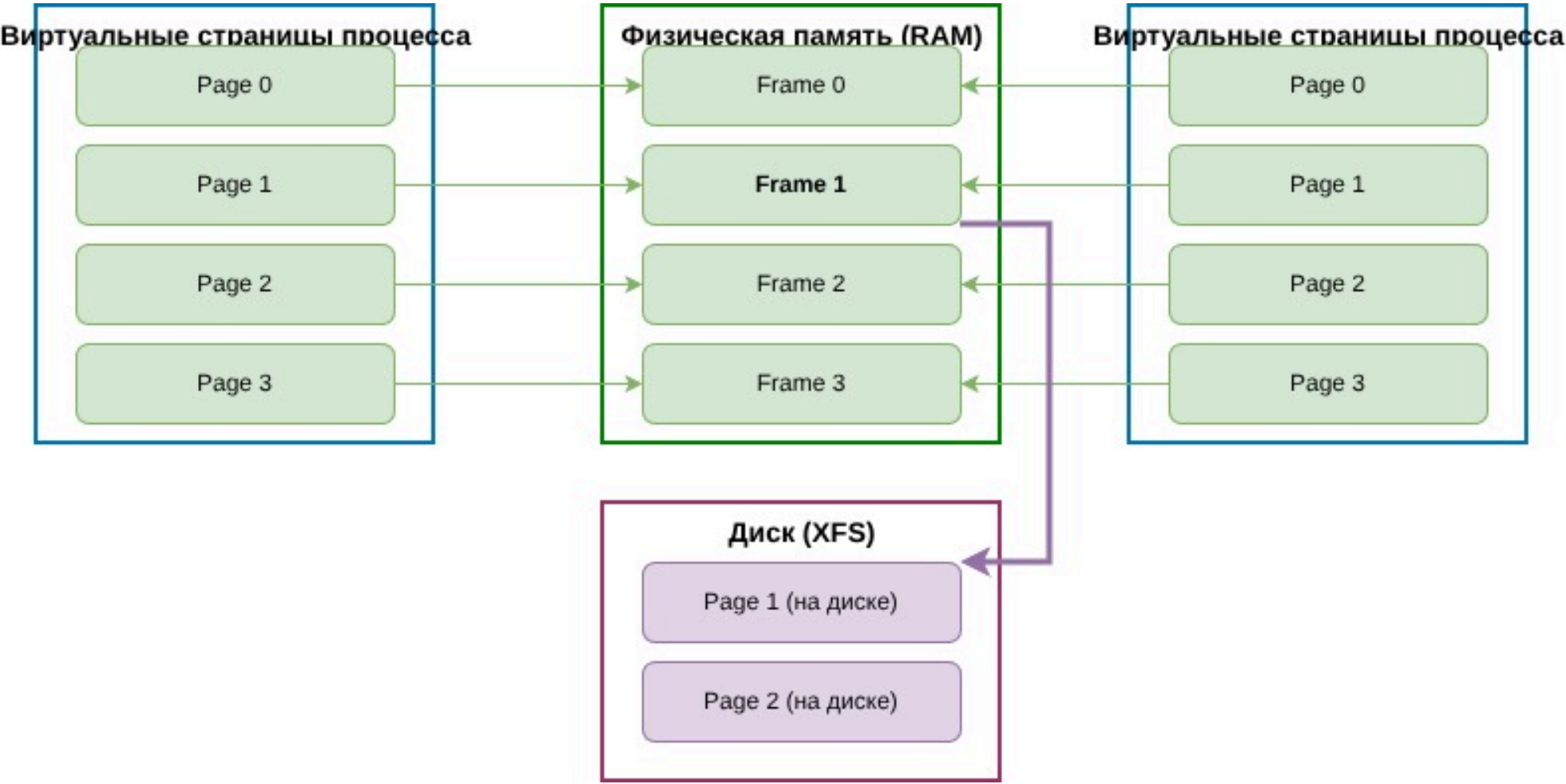
```
00 00 = add [rax], al
```

- Если RAX содержит невалидный адрес, то падает сразу
- Если RAX - валидный адрес, то... падает, но потом


Виртуальные страницы процесса
Page 0
Page 1
Page 2
Page 3
Физическая память (RAM)
Frame 0
Frame 1
Frame 2
Frame 3
Виртуальные страницы процесса
Page 0
Page 1
Page 2
Page 3
Диск (XFS)
Page 1 (на диске)
Page 2 (на диске)


Виртуальные страницы процесса
Page 0
Page 1
Page 2
Page 3
Физическая память (RAM)
Frame 0
Frame 1 (free -> zeroing)
Frame 2
Frame 3
Виртуальные страницы процесса
Page 0
Page 1
Page 2
Page 3
Диск (XFS)
Page 1 (на диске)
Page 2 (на диске)


Виртуальные страницы процесса
Page 0
Page 1
Page 2
Page 3
Физическая память (RAM)
Frame 0
PAGE FAULT
Frame 2
Frame 3
Виртуальные страницы процесса
Page 0
Page 1
Page 2
Page 3
Диск (XFS)
Page 1 (на диске)
Page 2 (на диске)


Виртуальные страницы процесса
Page 0
Page 1
Page 2
Page 3
Физическая память (RAM)
Frame 0
Frame 1
Frame 2
Frame 3
Виртуальные страницы процесса
Page 0
Page 1
Page 2
Page 3
Диск (XFS)
Page 1 (на диске)
Page 2 (на диске)

Открытые вопросы

PostgresPro

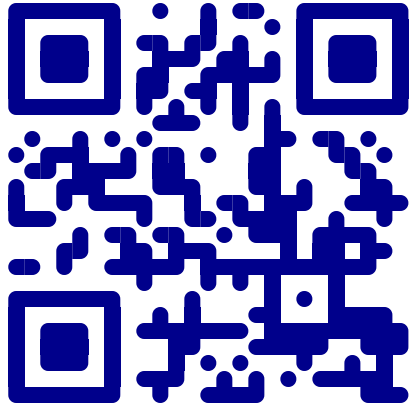
- Падаем до Page Fault или после?
- Дефект или недоделанная фича?
- Сможет ли AI агент помочь в поиске похожих проблем?
- Какие инструменты использовать для диагностики?

Заключение

Заключение

PostgresPro

- Декабрь 2025 года
- Технический бюллетень о проблеме
- Передано российским вендорам и ИСП РАН
- Статья на Хабр, где куда больше деталей



Guideline

PostgresPro

- Нужен стенд. Люди, увы - не телепаты
- Детальный анализ условий падений
  - что за процессы работали в этот момент, какие запросы/действия выполняли
- Если тест не выявляет дефект –> не то нагружается или не достаточно
- Гипер-усиливайте тест, чтобы сократить время воспроизведения
- Уменьшайте ресурсы постепенно, переводите на виртуалки
- Меняйте версии софта и параметры
- Полезно добавлять генераторы нагрузки на ресурсы ОС
- git-bisect / memtester / stress-ng / rasdaemon

Ну а это конец...?

PostgresPro

- Ребята из одного банка воспроизвели SEGFAULT и на **init_on_free=0**
- Но ушло не 1 минута, а 50 часов!
- Тикет в Linux bugzilla - я так и не завёл

Благодарности

PostgresPro

- Своим коллегам - Алексею Махмутову, Юре Соколову и Тимофею Мелко
- Коллегам из других компаний - Александру Бурцеву и Ко (product owner МБД Скала^р)
- И многих других компаниям и сочувствовавшим!
- Славе Смирнову из ПК Heisenbug за **ОГРОМНУЮ ПОМОЩЬ** в подготовке доклада!

Спасибо, что выслушали!

А данный доклад закончился

Вопросы?

Миша Жилин
Telegram: @mizhka
E-mail: mizhka@{gmail.com, freebsd.org}