PostgreSQL, временные таблицы и фрагментация памяти

Смолкин Григорий
Петров Сергей

Участники

- Смолкин Григорий
- Петров Сергей
- Попов Николай
- Лубенникова Анастасия
- Федор Сигаев
- Пан Константин

- Жданюк Александр
- Елисеев Андрей

Стенд

Software

Centos 7.2

PostgreSQL 9.5.4

+1C patchset

1C Платформа 8.3.8

Cluster size: 350GB

Время: 10 часов

Hardware

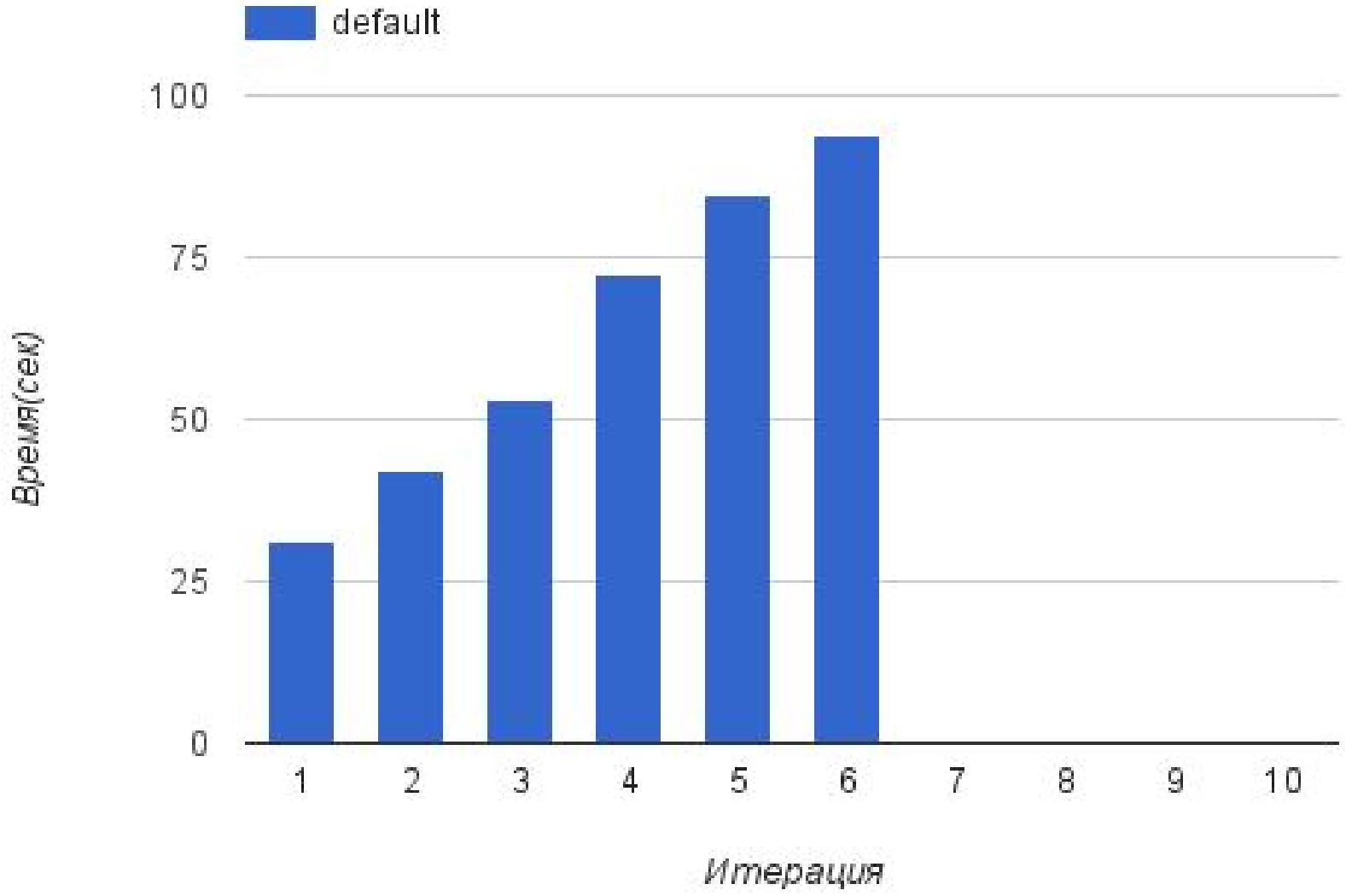
4 x 8core Intel(R)

E7-8837 2.67GHz

64GB RAM

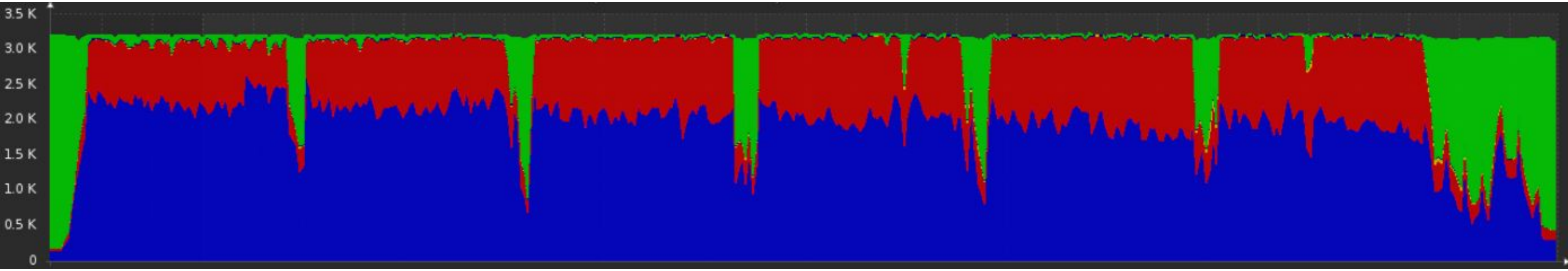
Fusion ioDrive

Первые результаты



Первые результаты

CPU LOAD

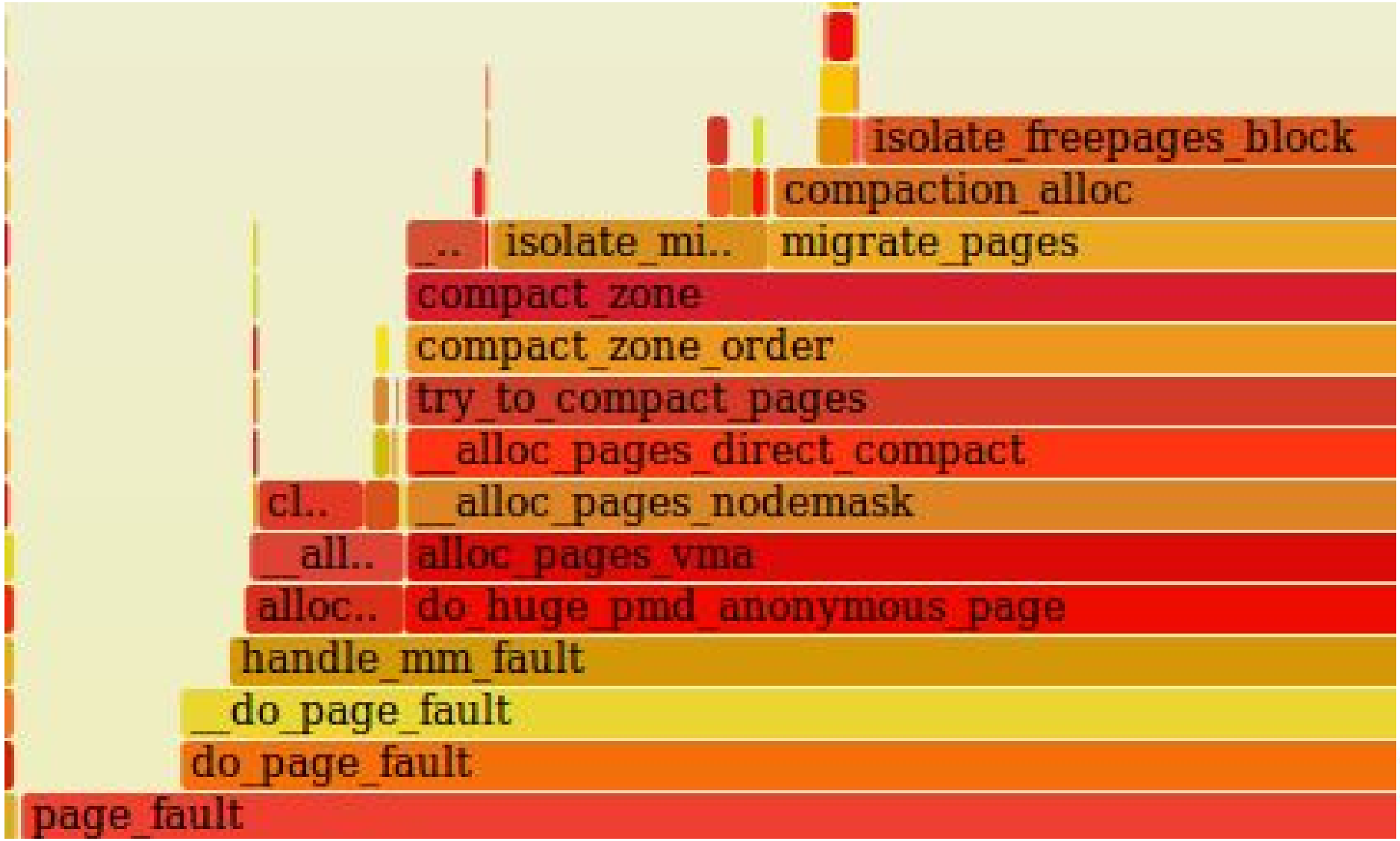


CPU: SYSTEM USER IDLE

Инструментарий

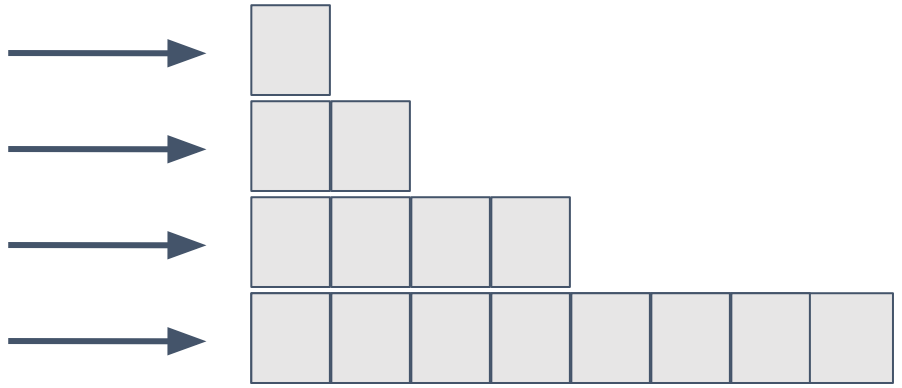
- mamonsu
- strace
- perf
- flamegraph
- atop+atopsar

Фрагментация памяти?



Memory Allocation

0	
1	
2	
3	
4	
5	
6	
7	
8	
9	
MAX ORDER	



2^0 page block(4KB)

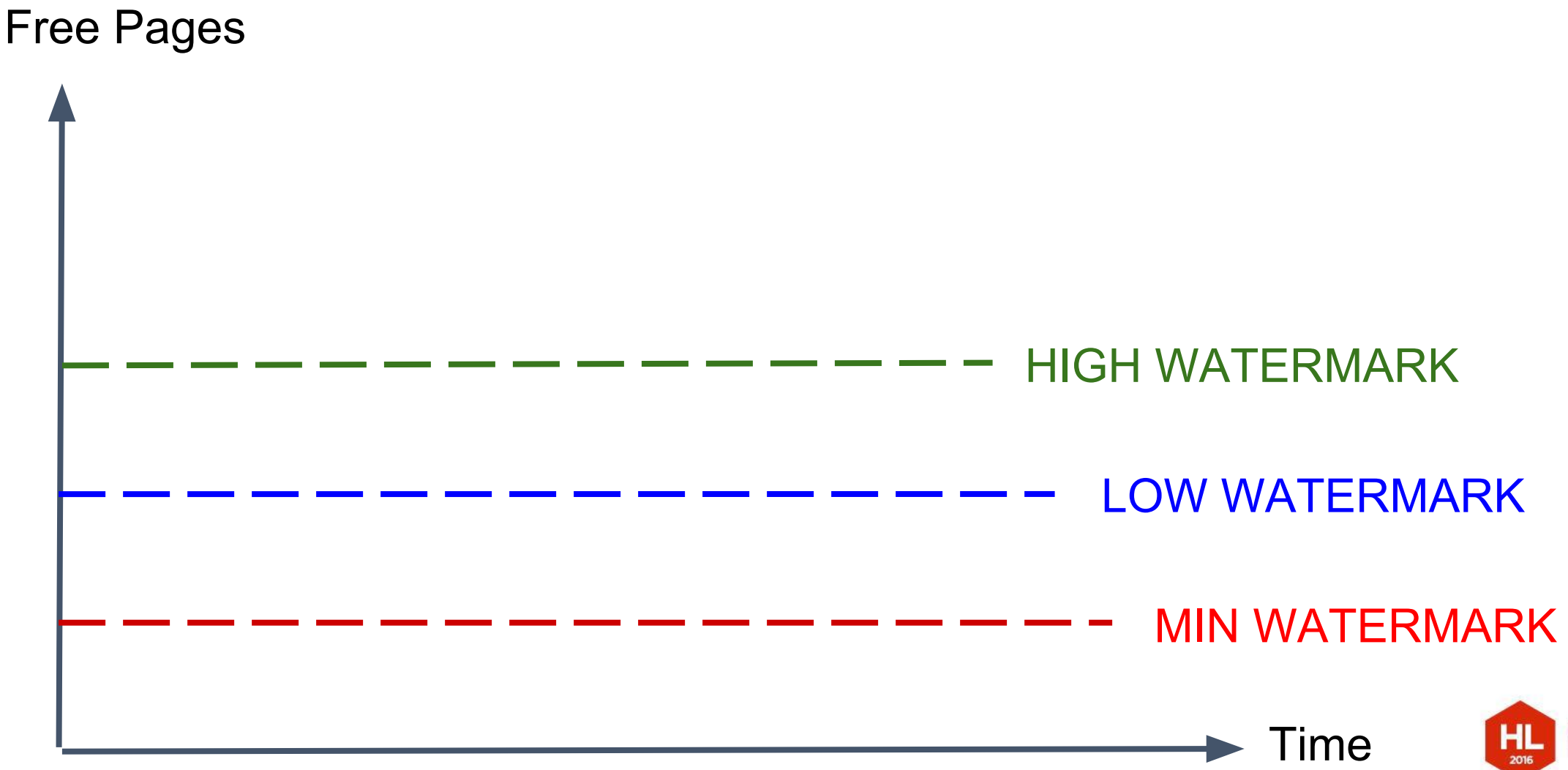
2^1 page block(8KB)

2^2 page block(16KB)

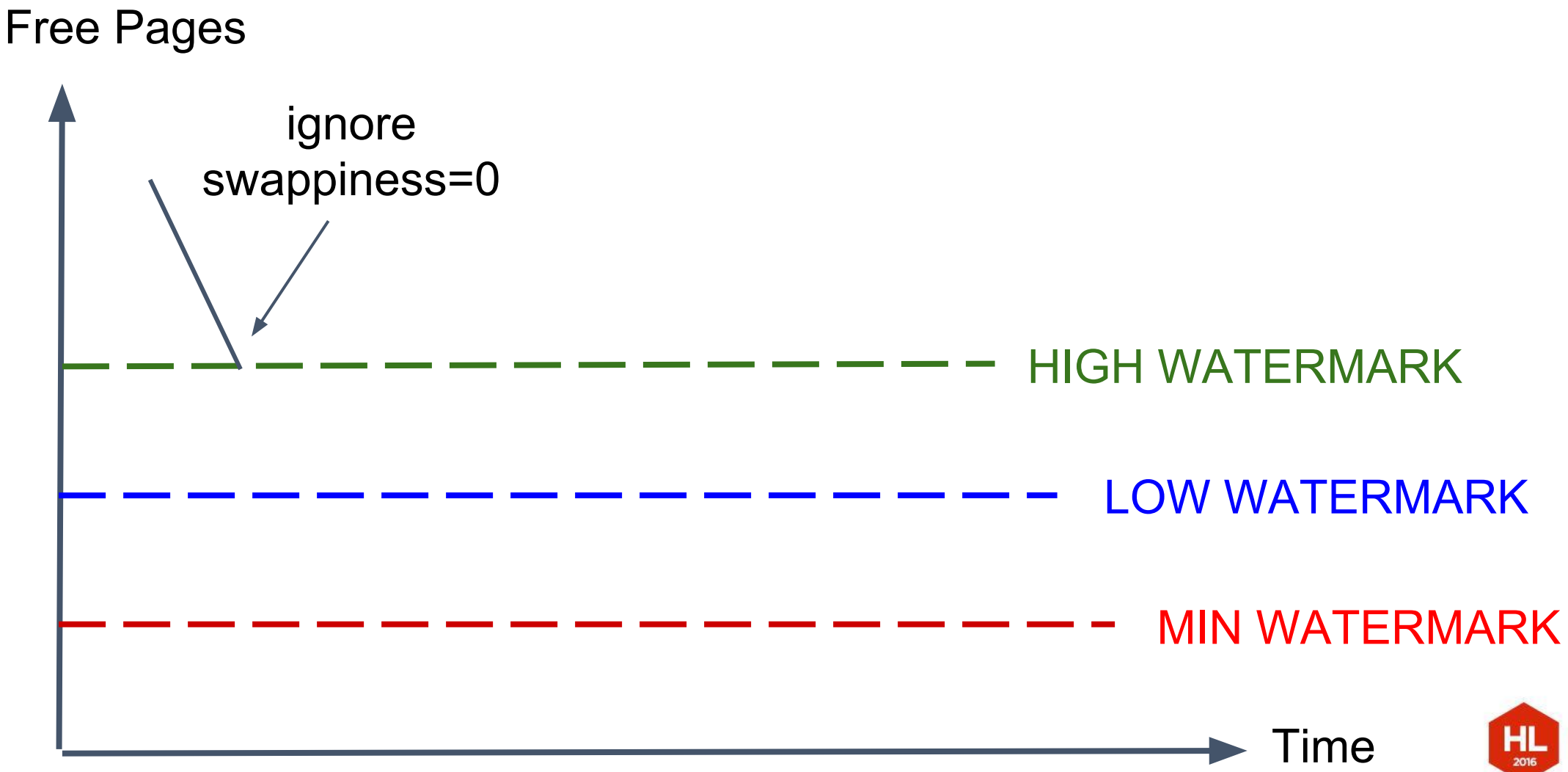
2^3 page block(32KB)

PAGE_SIZE = 4KB

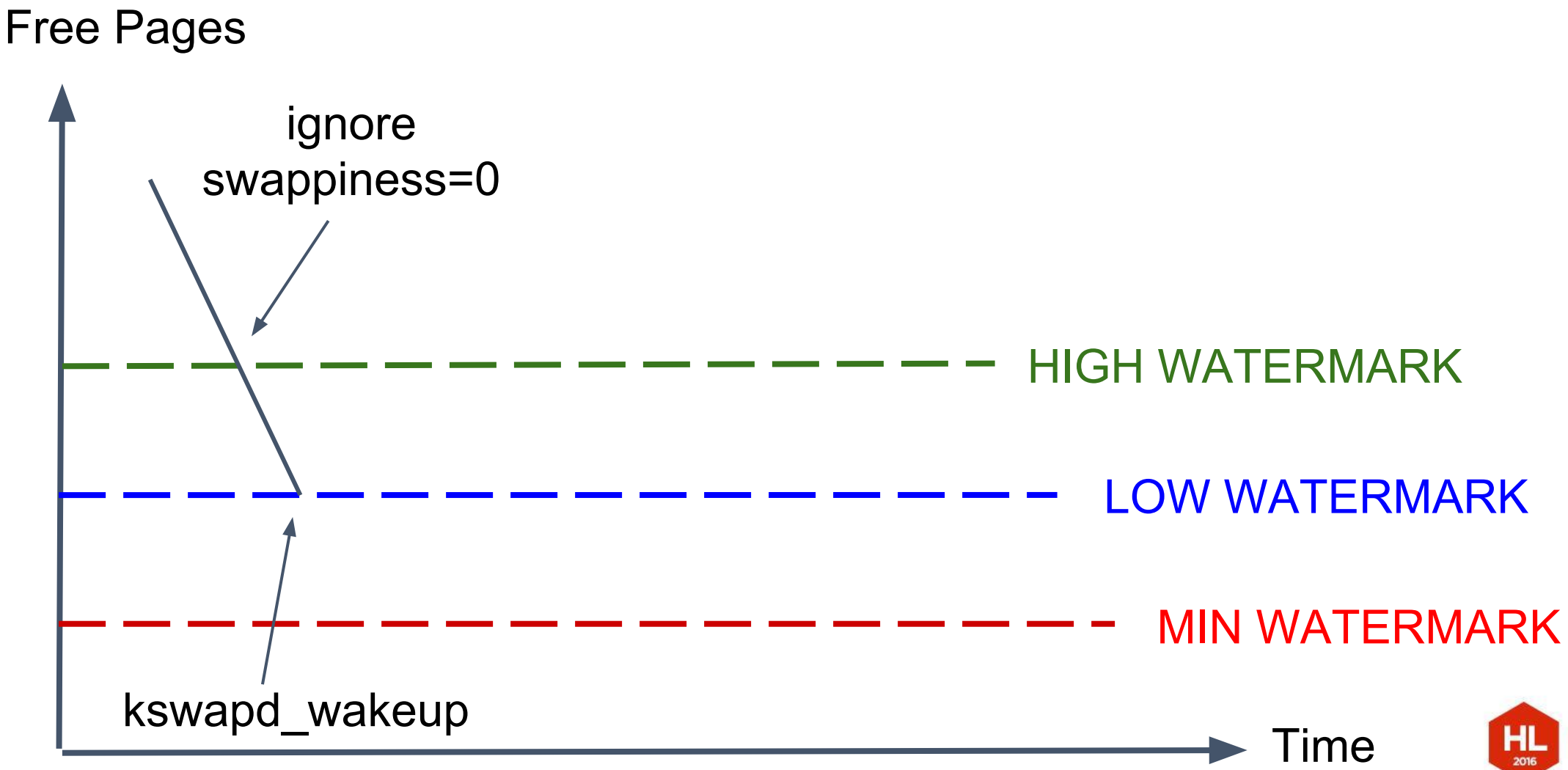
Memory Reclamation



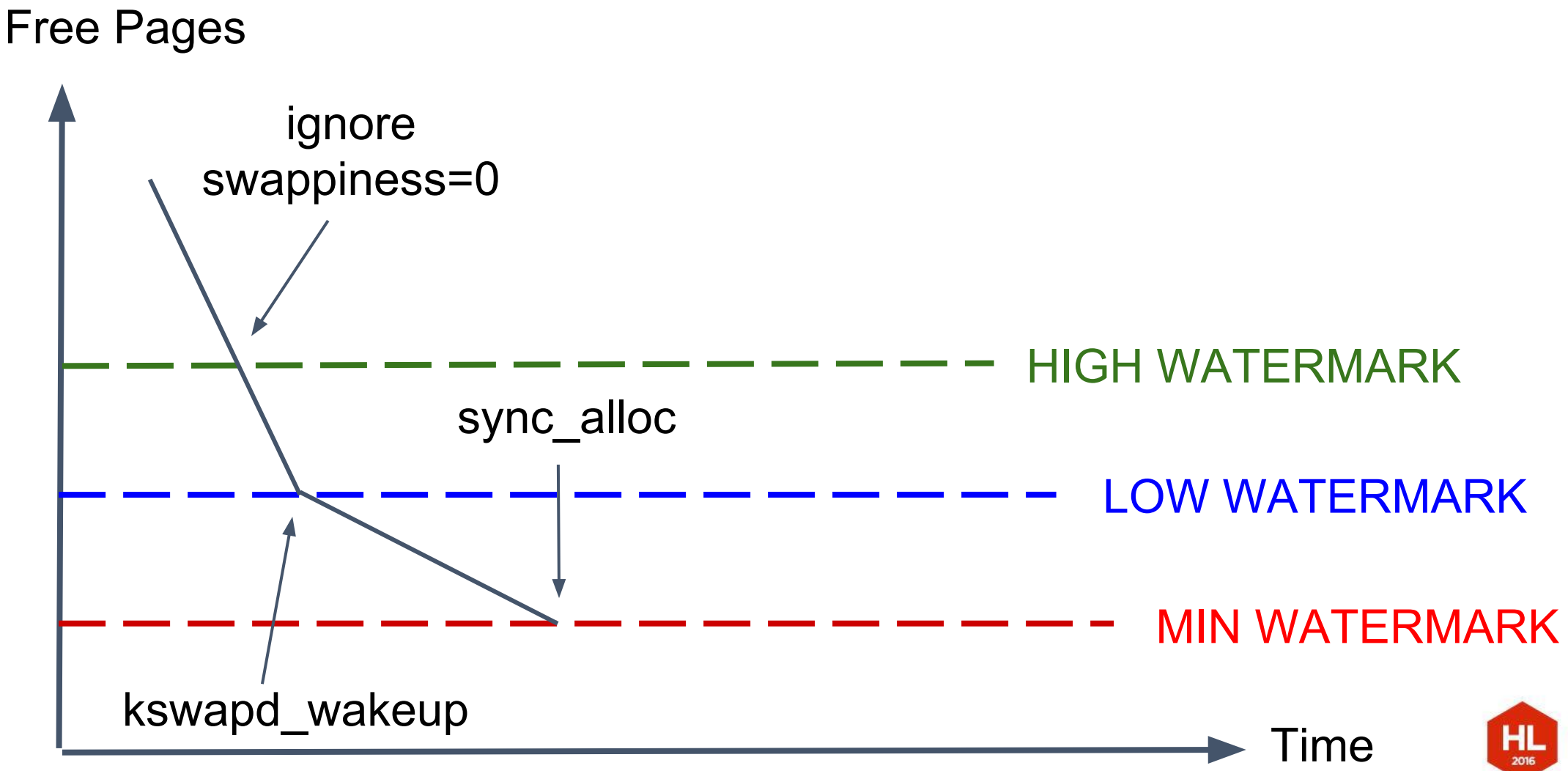
Memory Reclamation



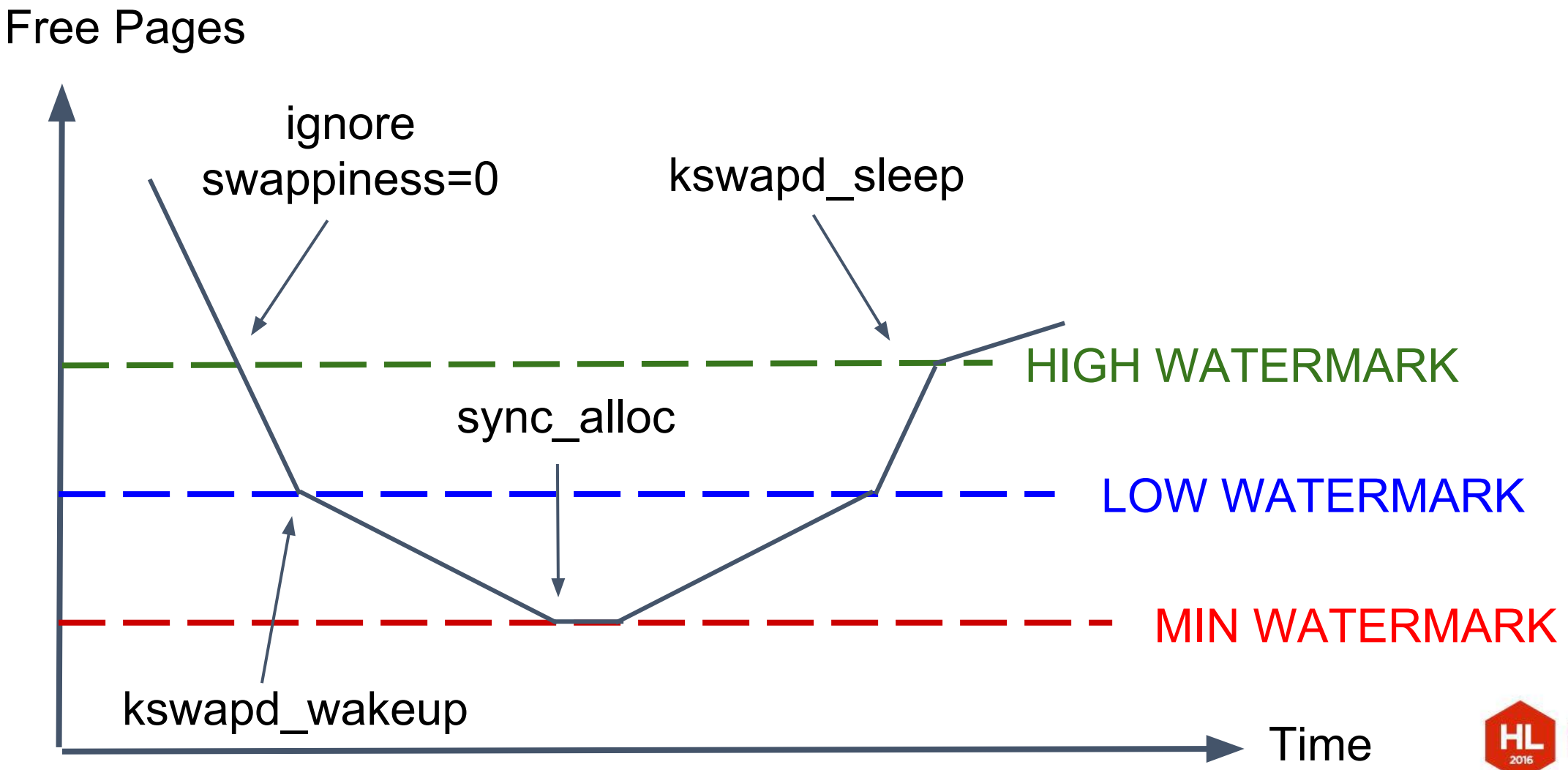
Memory Reclamation



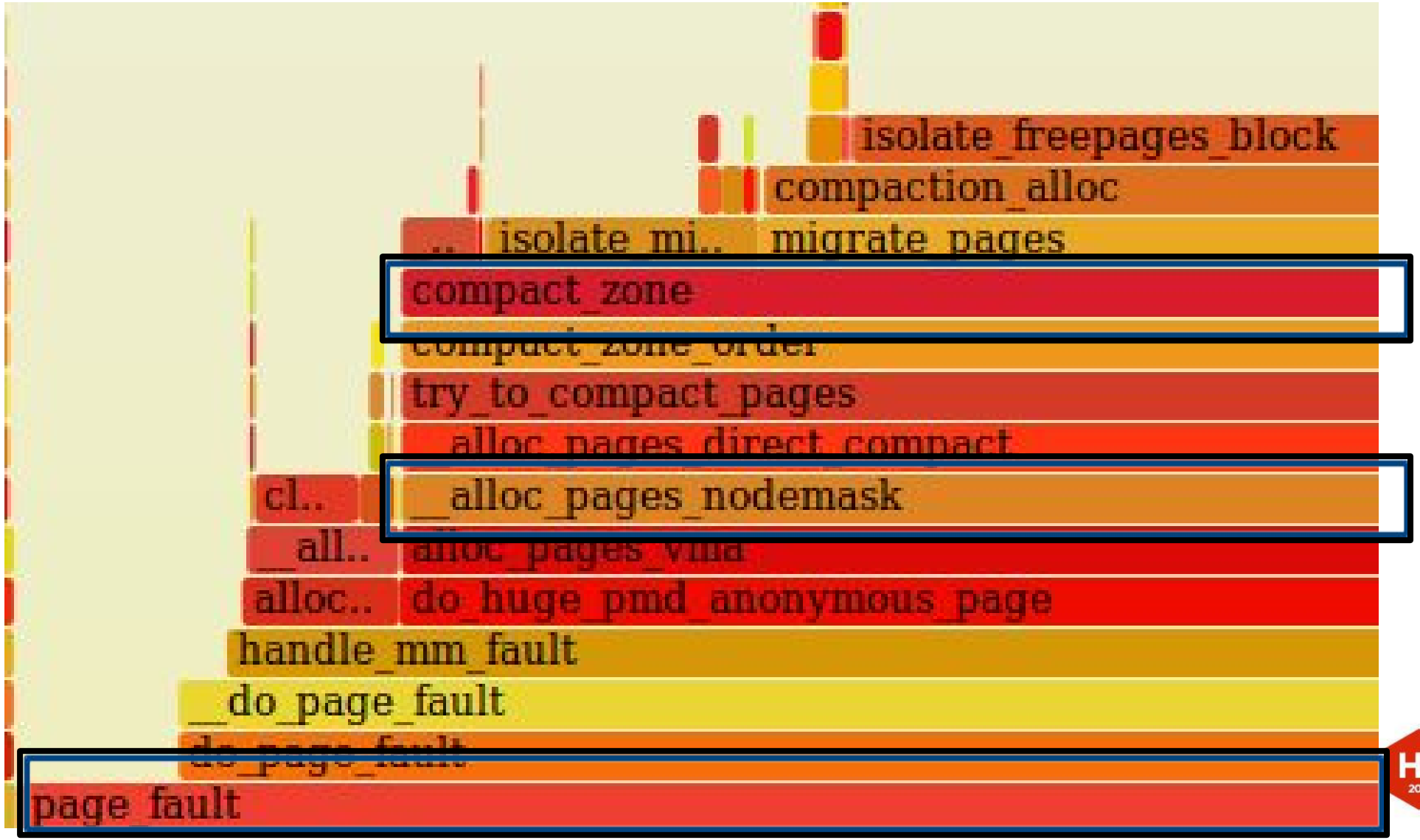
Memory Reclamation



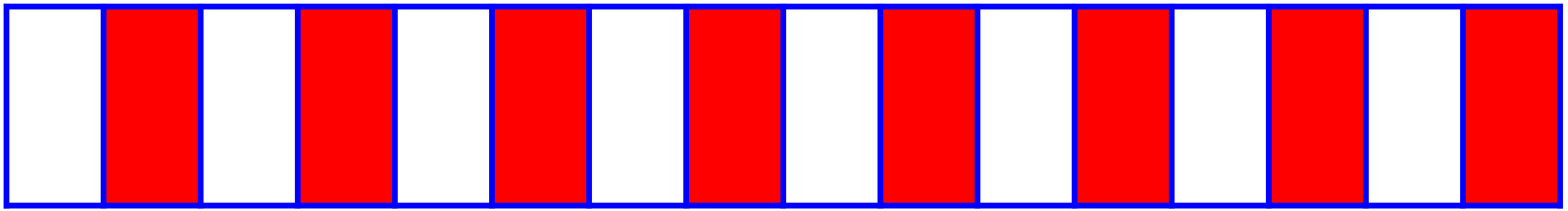
Memory Reclamation



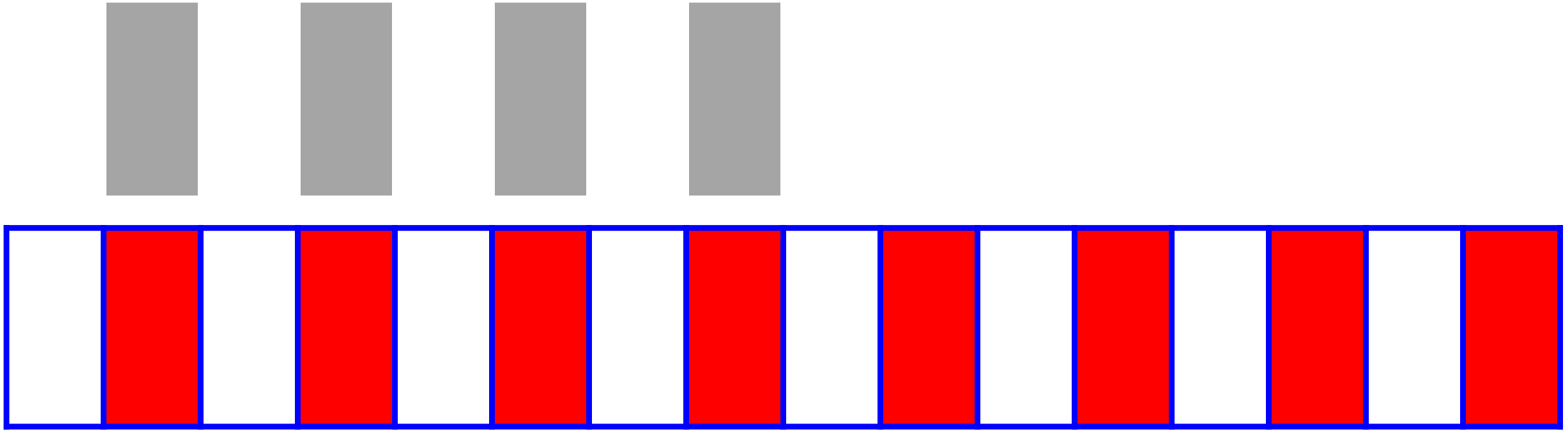
Дефрагментация памяти



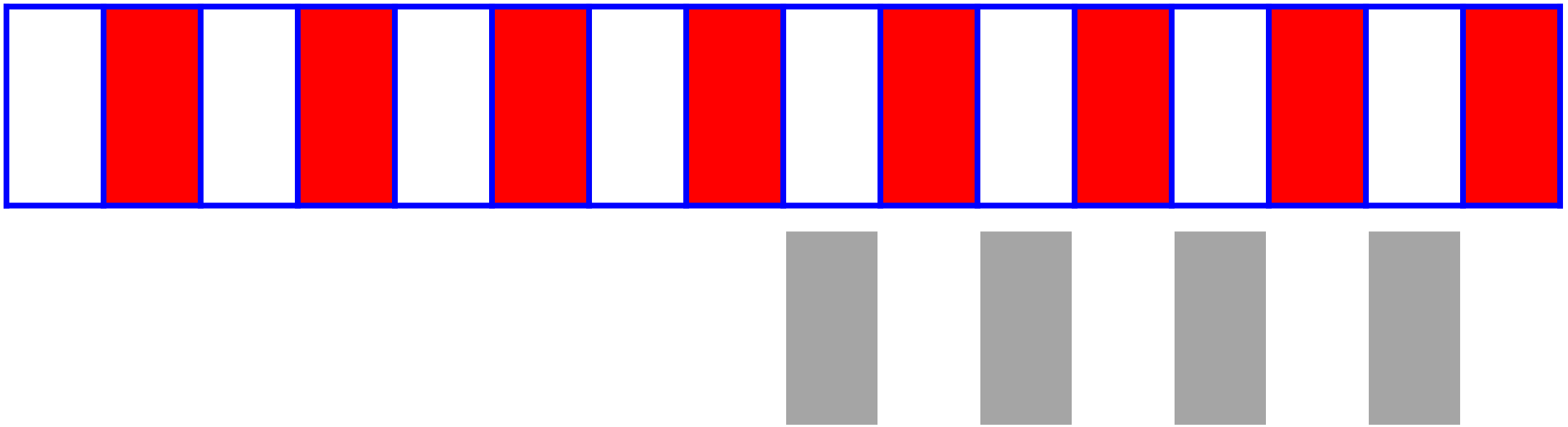
Дефрагментация памяти



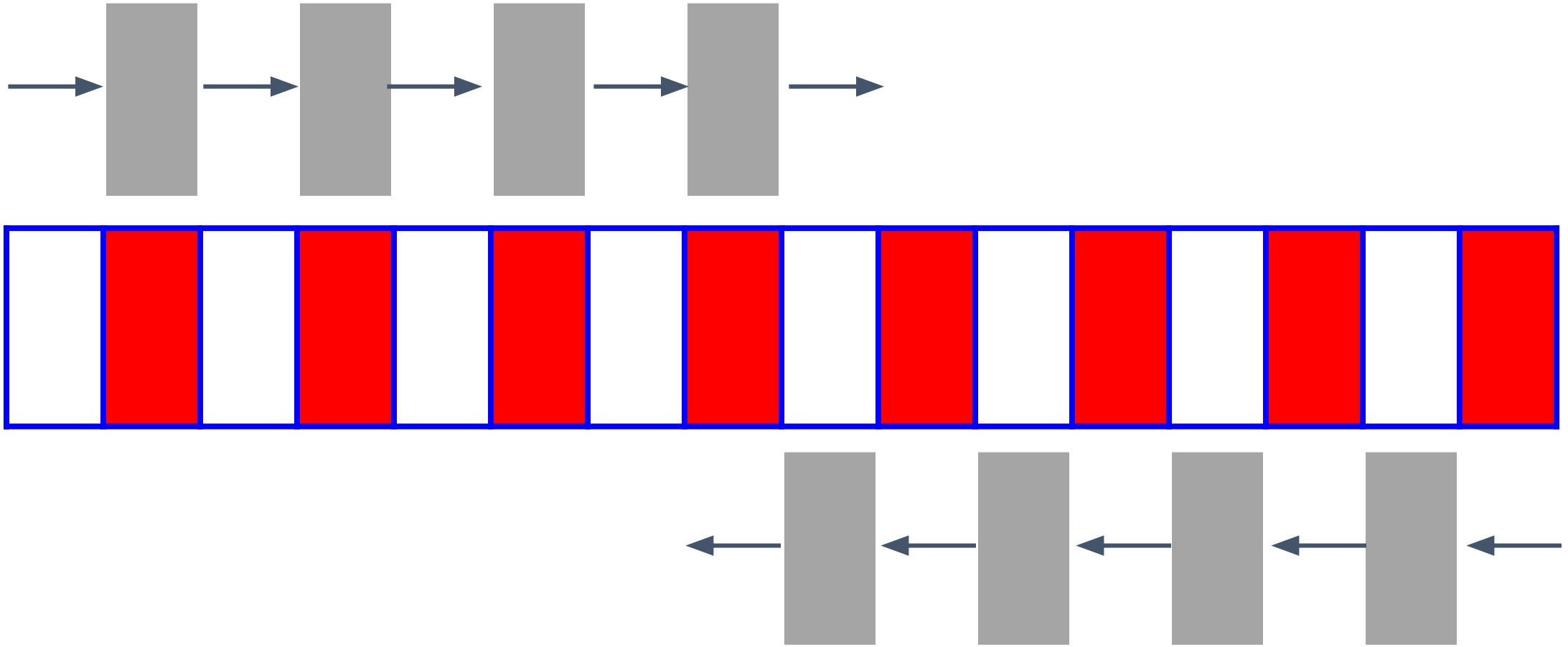
ISOLATE_MIGRATEPAGES



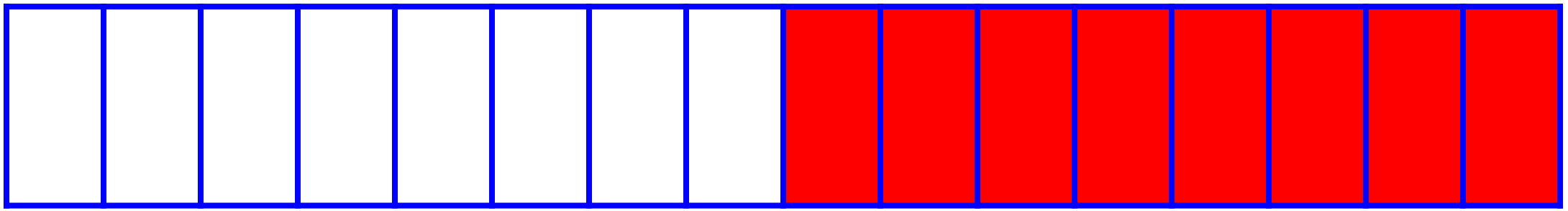
ISOLATE_FREEPAGES_BLOCK



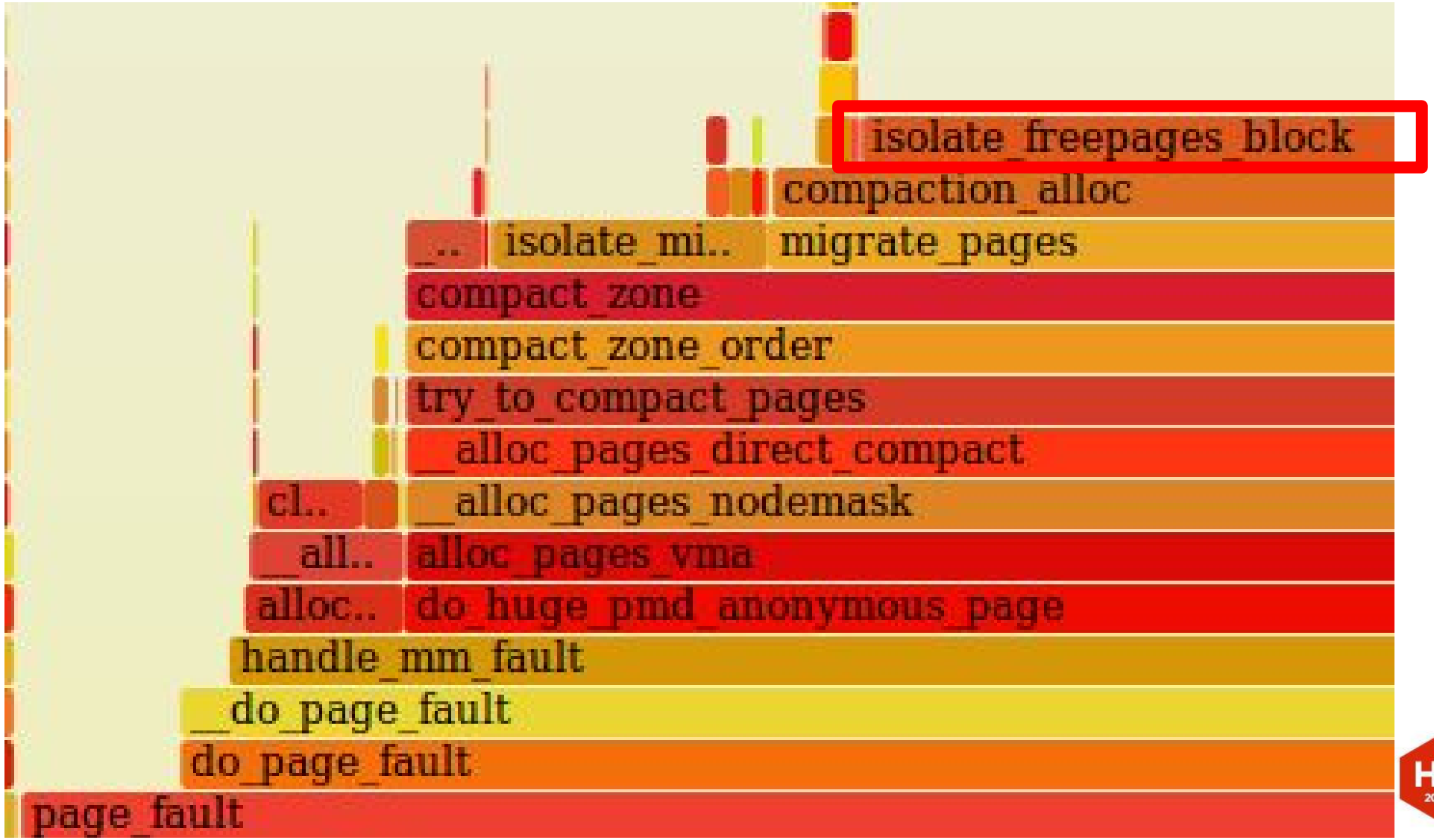
MIGRATE_PAGES



Дефрагментация памяти



ISOLATE_FREEPAGES_BLOCK



Что делать?

vm.min_free_kbytes:

+kswapd раньше начнет работу

+больше свободных страниц

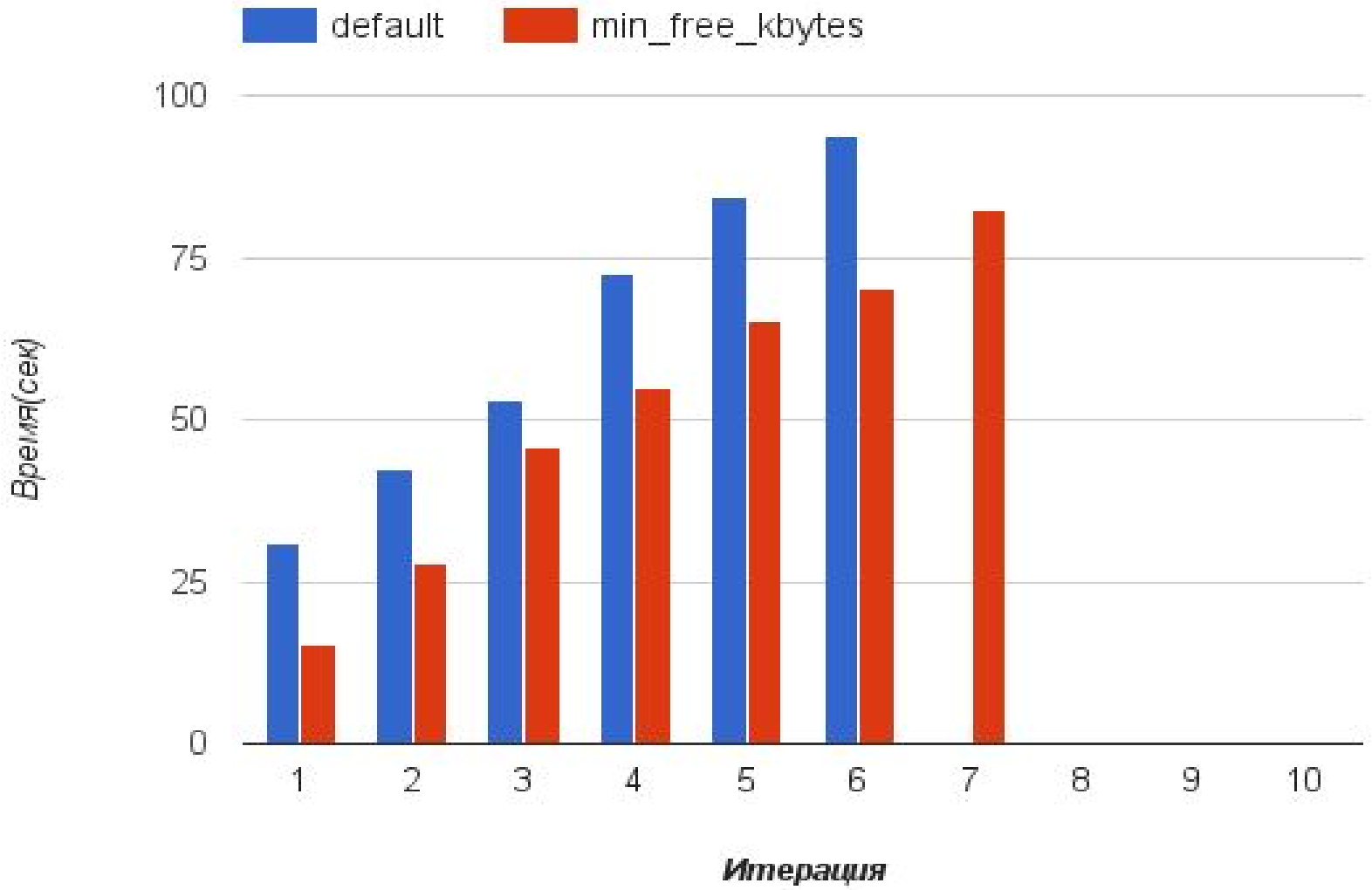
-меньше единовременно доступной памяти

vm.extfrag_threshold(500 by default):
/sys/kernel/debug/extfrag/extfrag_index (divided by 1000)

Node 0, zone Normal -1.000 -1.000 -1.000 -1.000 -1.000 -1.000 -1.000 0.983 0.992 0.996 0.998

extfrag_index >= vm.extfrag_threshold /1000 -> COMPACTION_CONTINUE

vm.min_free_kbytes=3%RAM



Временные таблицы

1. Привязаны к определенному серверному процессу
2. Не имеют механизма сбора статистики
3. Размещены в локальной памяти
4. Удаляются после отключения клиента
5. Постранично резервируют место на диске

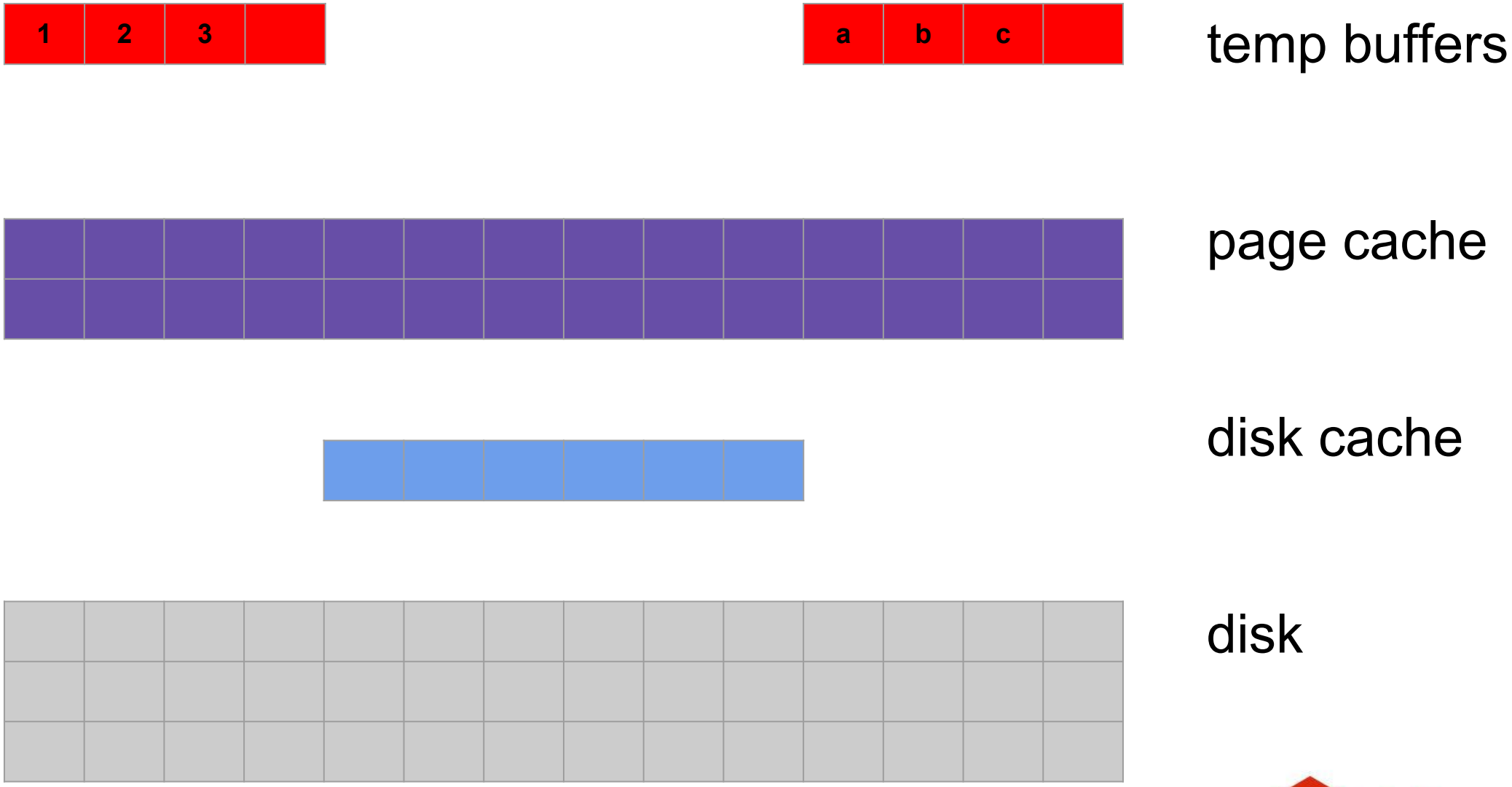
Резервирование места

```
open("base/13090/t2_1043562", O_RDWR)   = 40
write(40, "\0\0\0\0\0\0\0\0\0\0\0\0\0\0\0\0\0\0\0\0"..., 8192) = 8192
write(40, "\0\0\0\0\0\0\0\0\0\0\0\0\0\0\0\0\0\0\0\0"..., 8192) = 8192
write(40, "\0\0\0\0\0\0\0\0\0\0\0\0\0\0\0\0\0\0\0\0"..., 8192) = 8192
write(40, "\0\0\0\0\0\0\0\0\0\0\0\0\0\0\0\0\0\0\0\0"..., 8192) = 8192
write(40, "\0\0\0\0\0\0\0\0\0\0\0\0\0\0\0\0\0\0\0\0"..., 8192) = 8192
write(40, "\0\0\0\0\0\0\0\0\0\0\0\0\0\0\0\0\0\0\0\0"..., 8192) = 8192
```

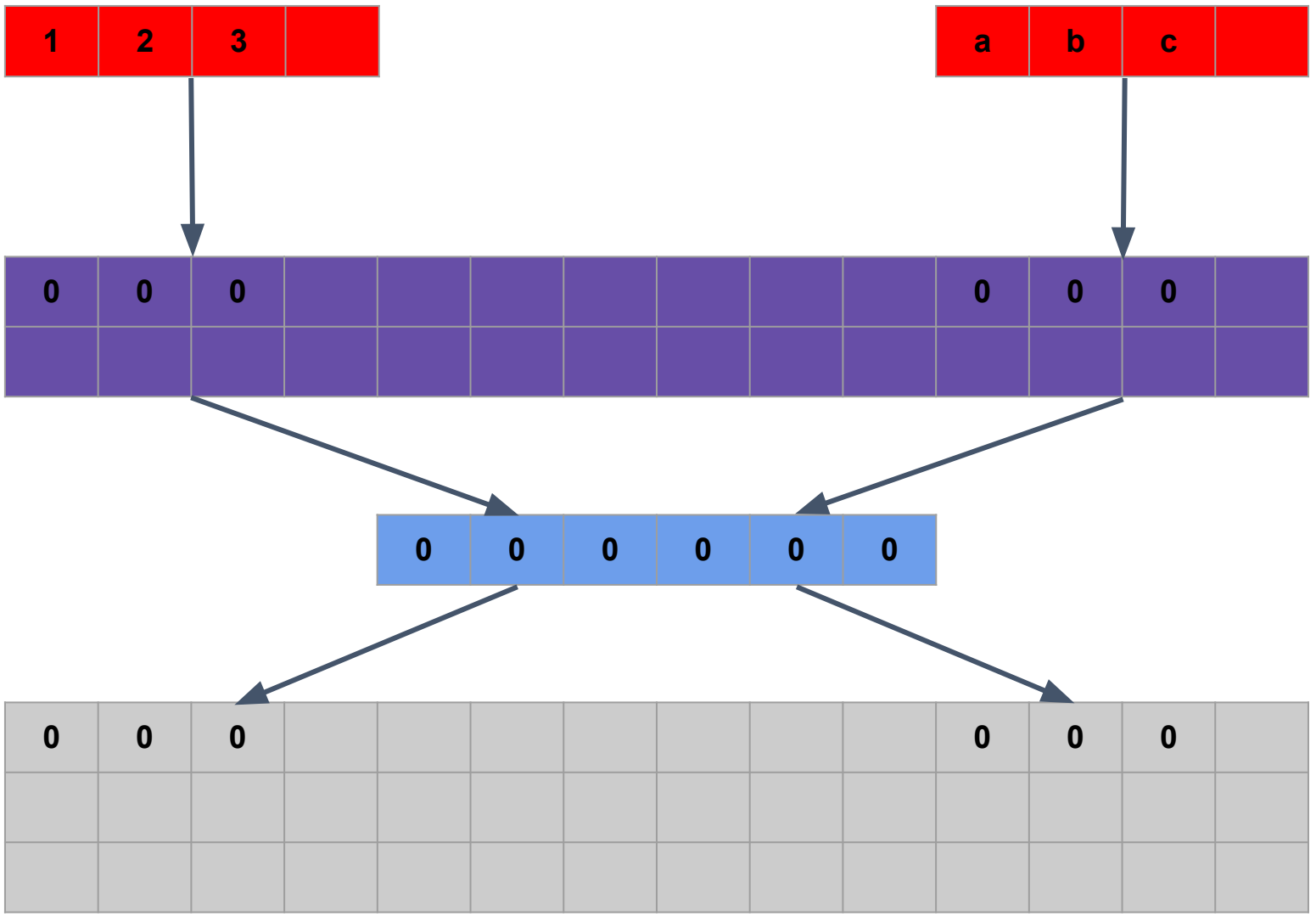
Почему это плохо

1. Фрагментация памяти(8KB = 2^1 * 4KB)
2. Тройная буферизация!!!

Буферизация



Буферизация



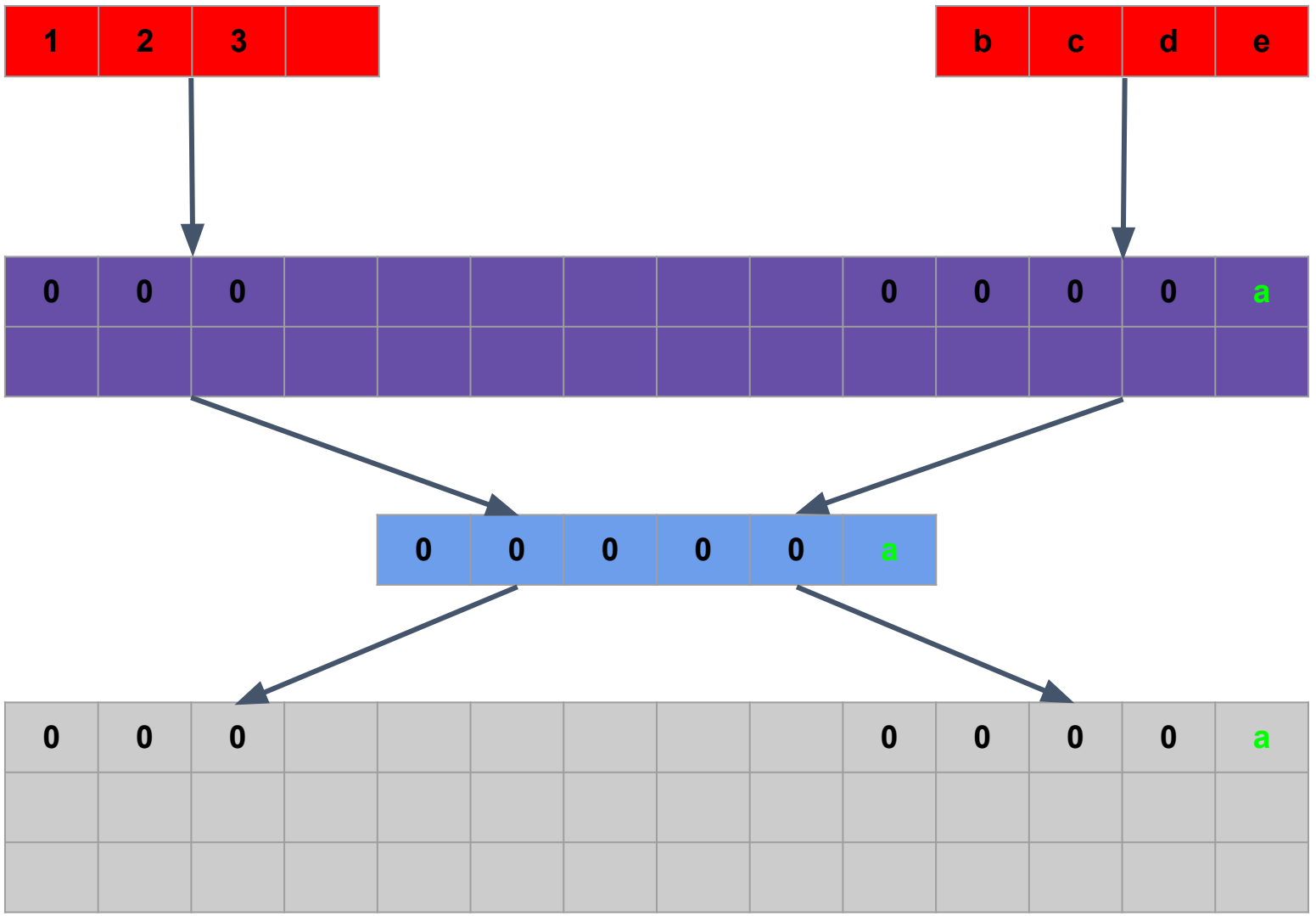
temp buffers

page cache

disk cache

disk

Буферизация



temp buffers

page cache

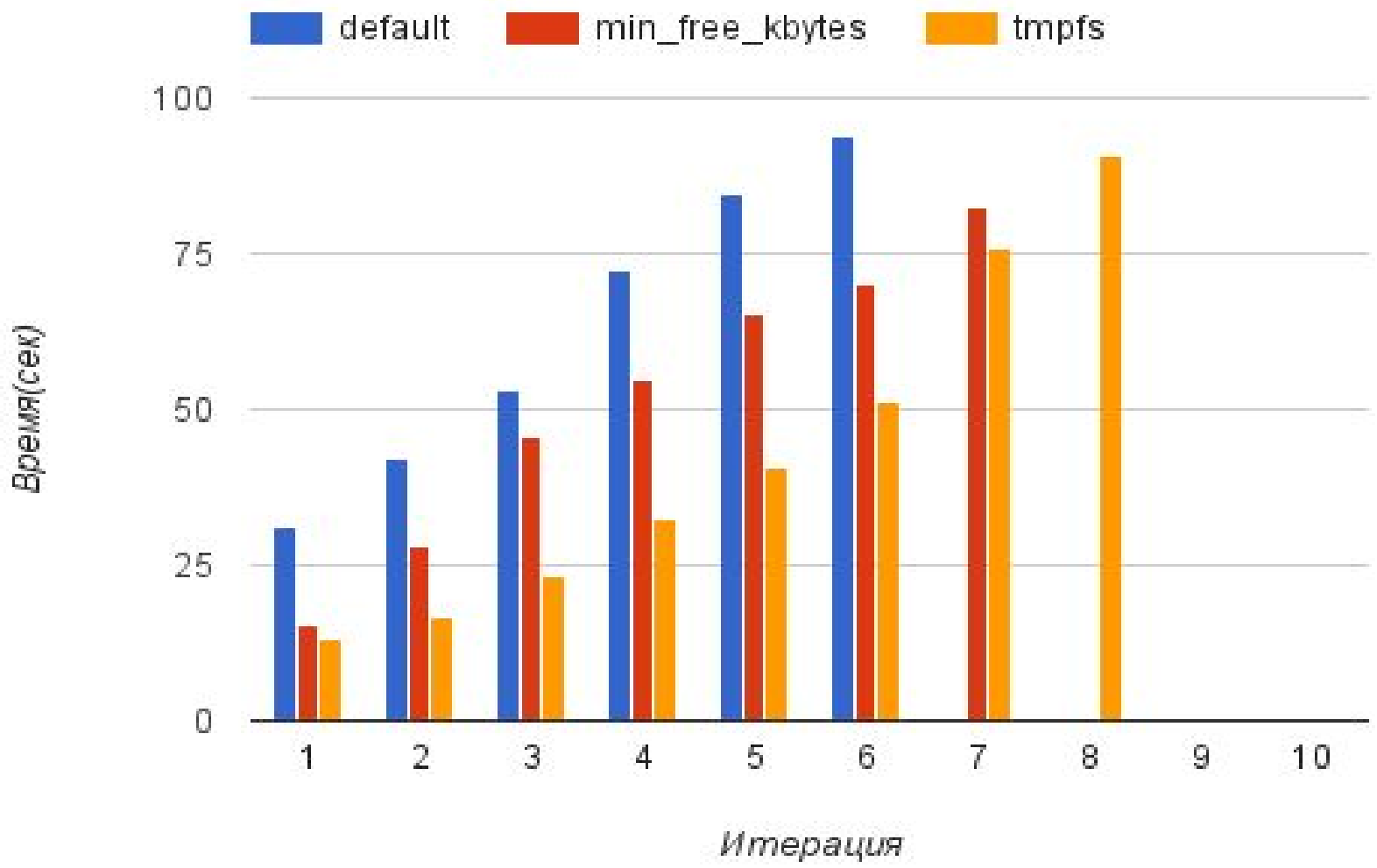
disk cache

disk

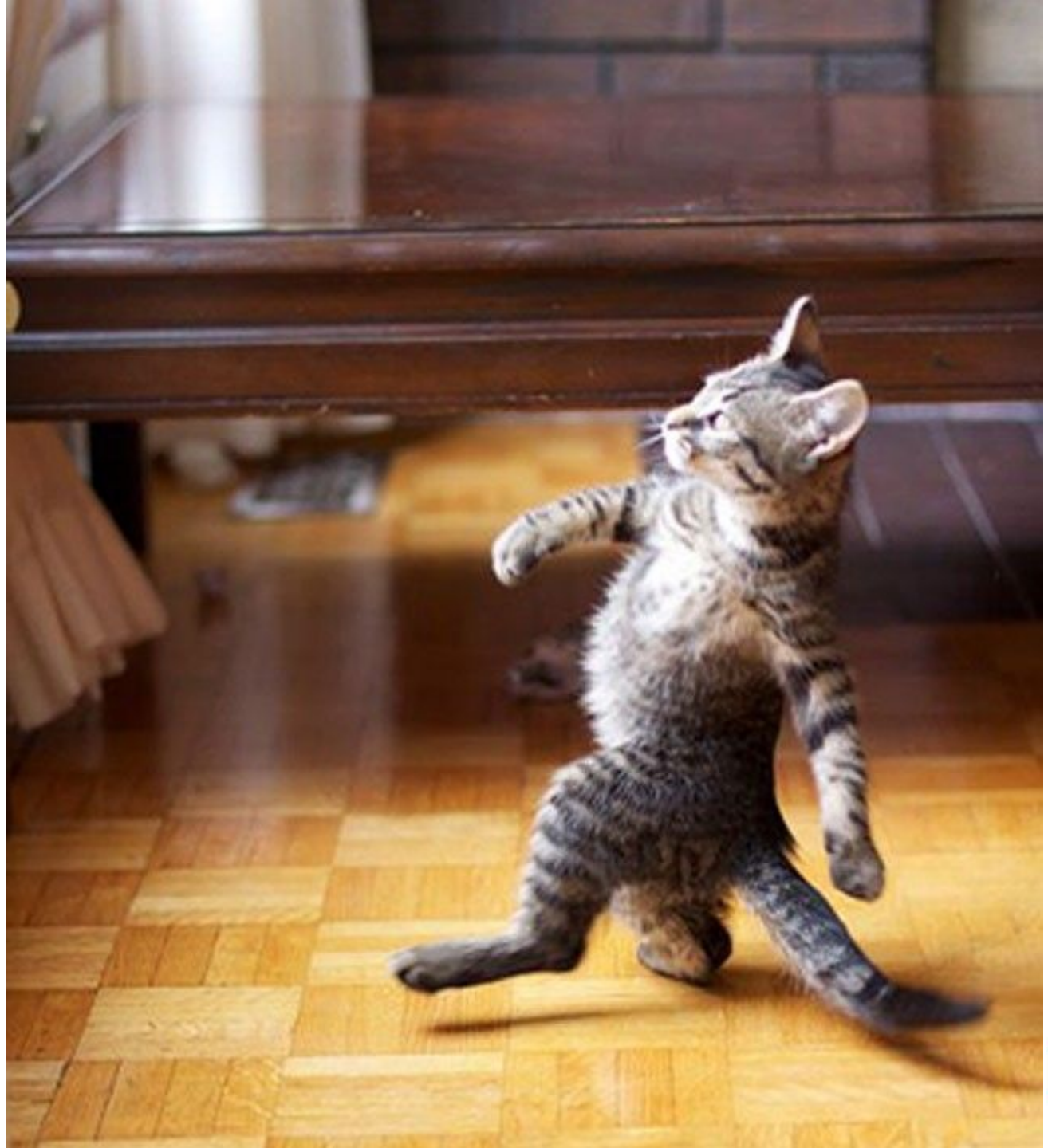
Что делать? TMPFS?

1. Фрагментация памяти(8KB = 2^1 * 4KB)
2. ~~Тройная буферизация!!!~~
2. Двойная буферизация :)

TMPFS



Мы отправились в отдел разработки

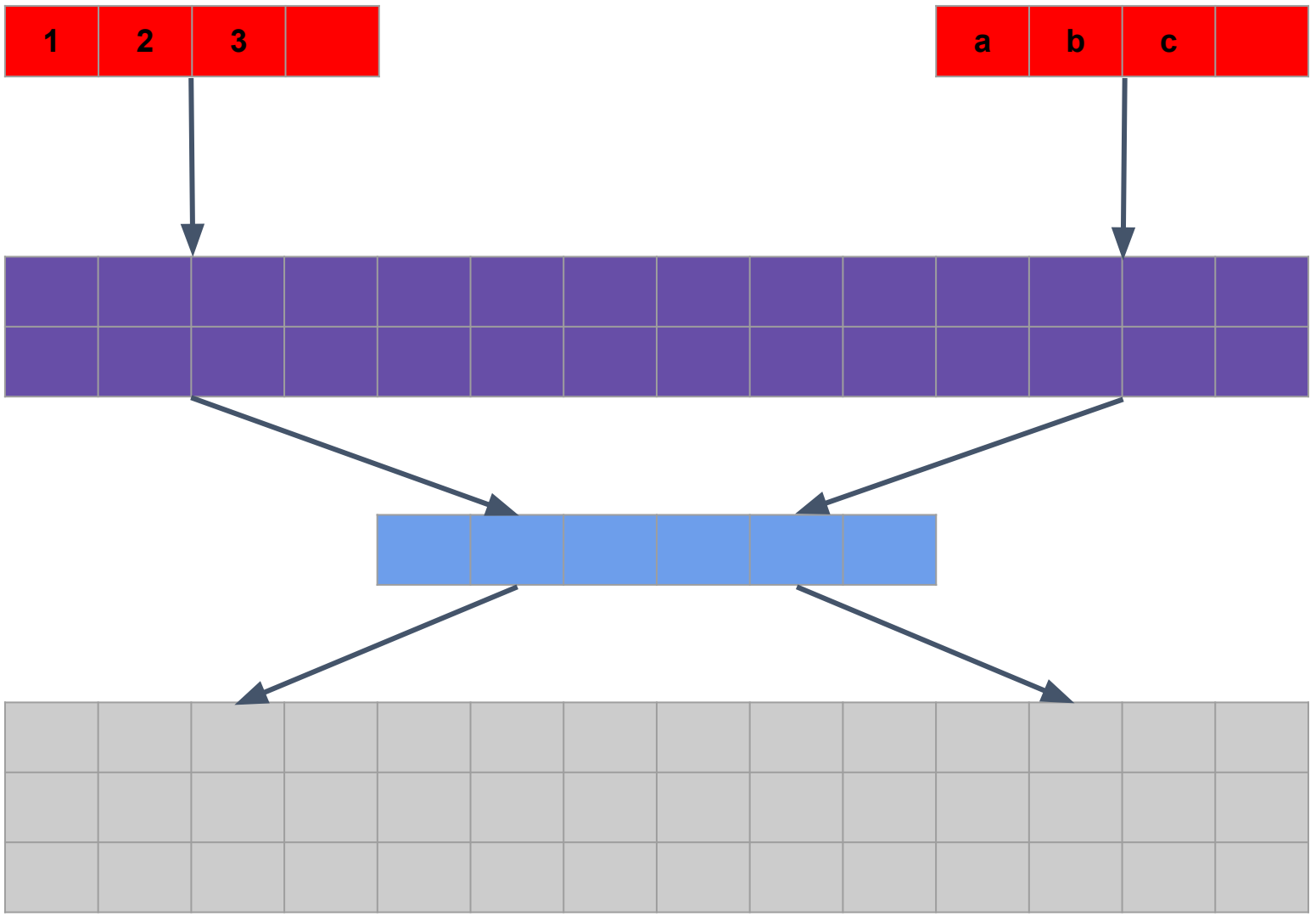


Патч №1: Отключение резервирования

Автор: Анастасия Лубенникова

- отключает резервирование для временных таблиц
- индексы, FSM, VM по-прежнему пишутся на диск
- PostgresPro Enterprise
- PostgresPro Enterprise 1C

Буферизация



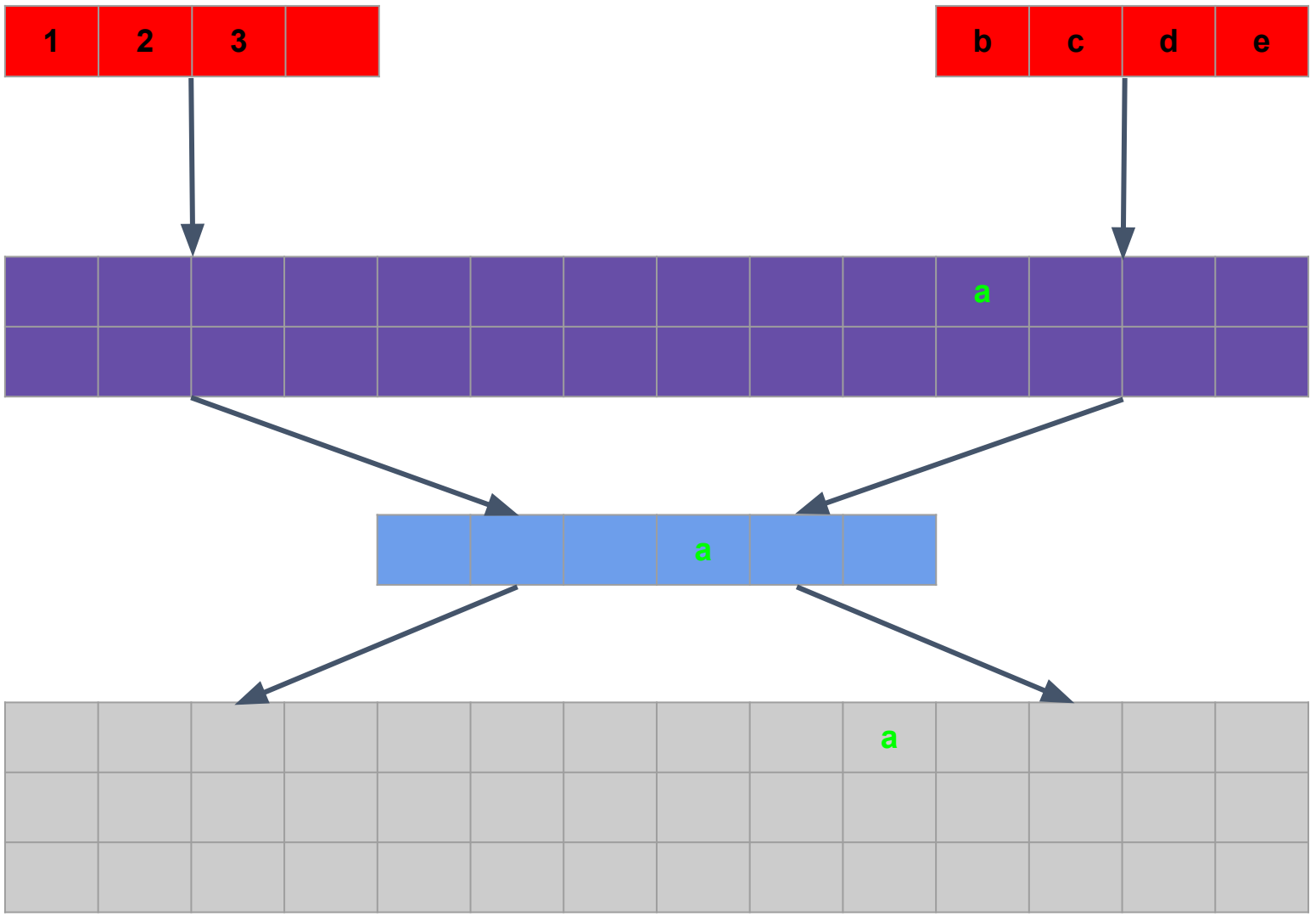
temp buffers

page cache

disk cache

disk

Буферизация



temp buffers

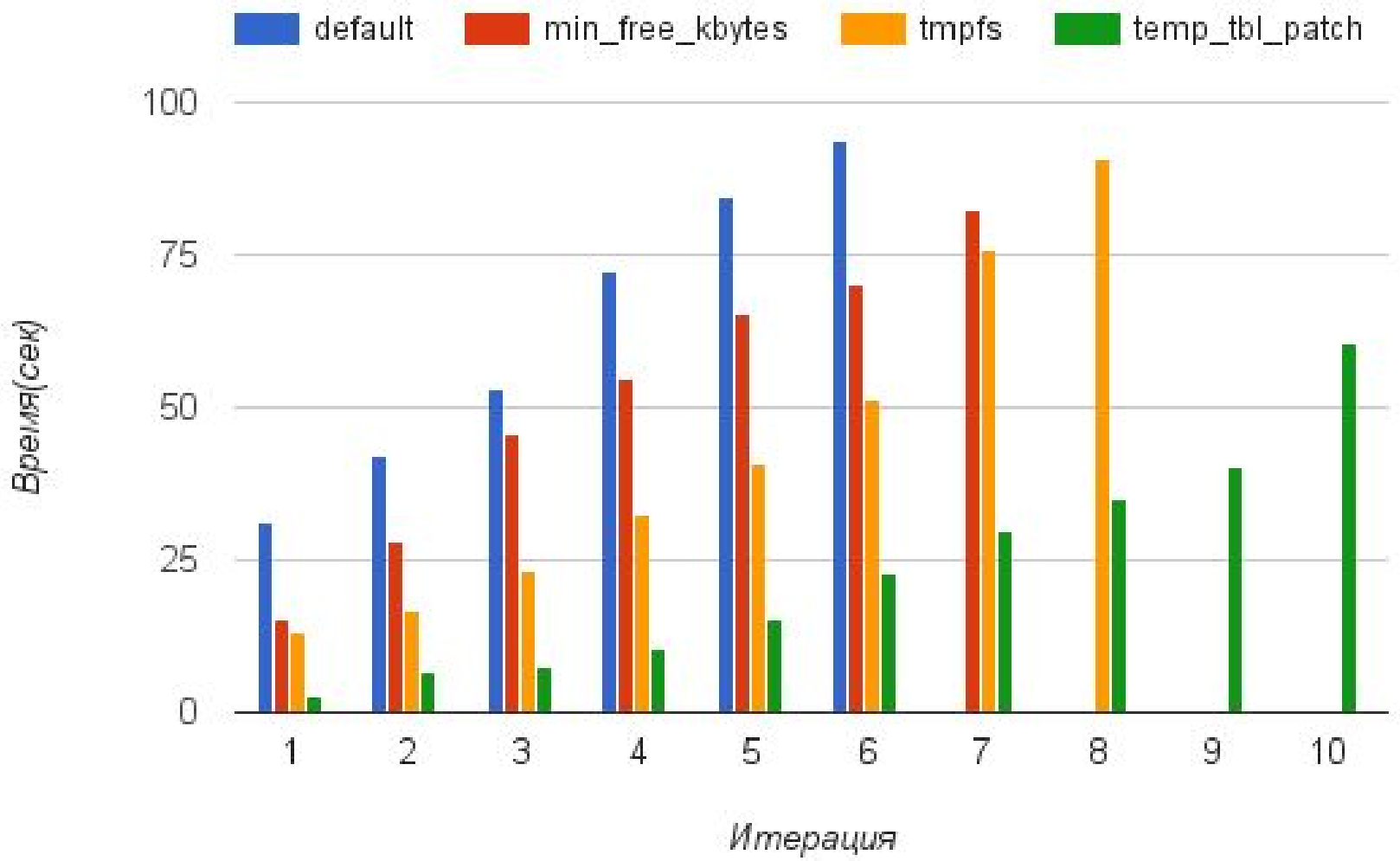
page cache

disk cache

disk

I/O

Патч №1: Отключение резервирования



Осиротевшие временные таблицы

Orphan temp tables

Условия появления:

- Крэш постгреса(питание, oom killer, etc.)
- Не хватило памяти на локи

Симптомы:

- out of shared memory
- autovacuum «found orphan table»

Orphan temp tables

Почему это плохо?

- Распухание каталога
- Вакуум не удаляет осиротевшие таблицы
- Автовакуум спамит в лог

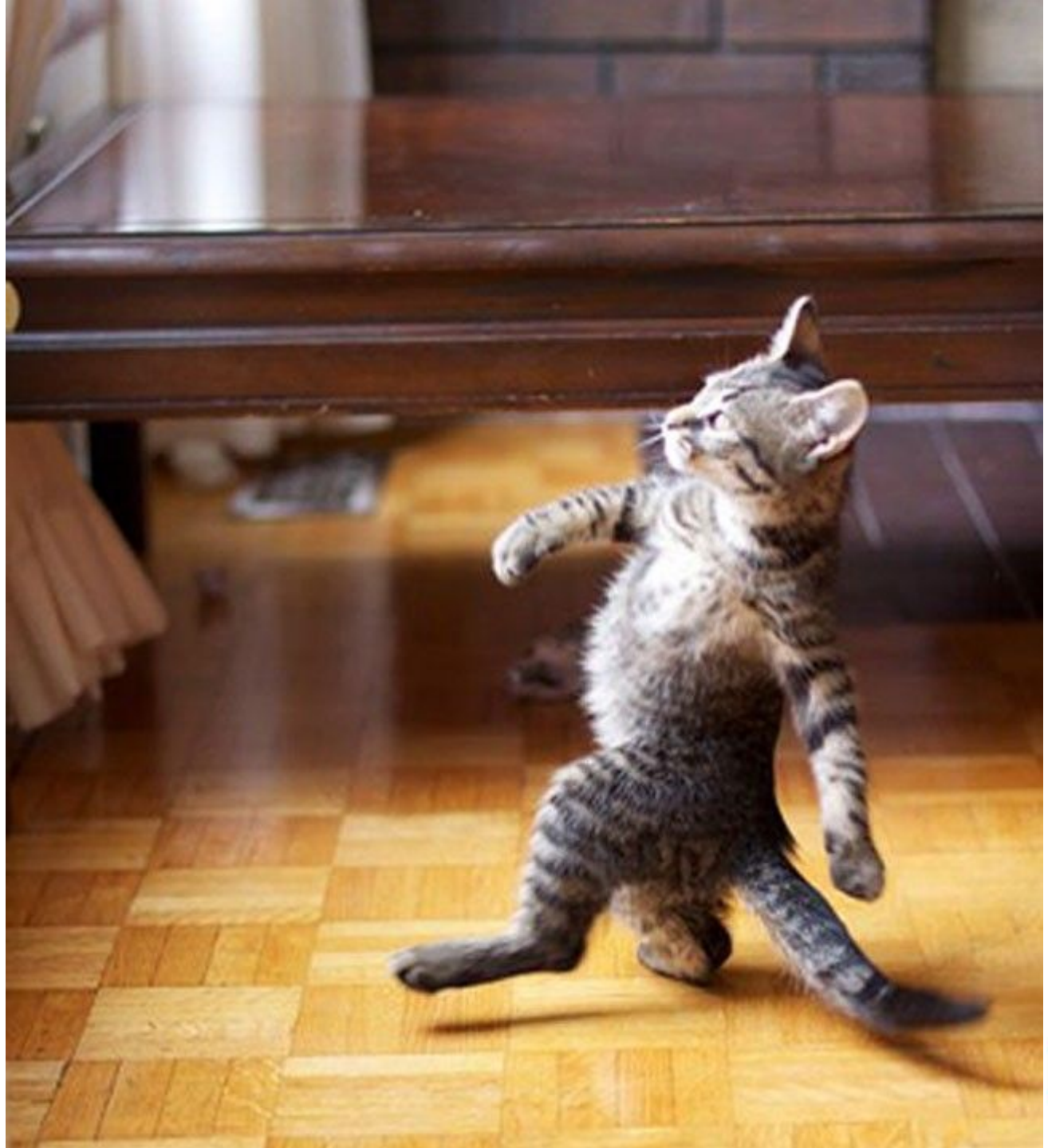
Orphan temp tables

Что делать?

- Увеличить размер lock_table:
  lock_table = max_locks_per_transaction * (max_conn + max_pred_locks_per_transaction)

- Удалить схемы с временными таблицами:
  DROP SCHEMA pg_temp_n CASCADE;
  DROP SCHEMA pg_toast_temp_n CASCADE;

Мы отправились в отдел разработки



Патч №2: Orphan tables clean up

Автор: Константин Пан

- корректная работа с локами
- keep_orphan_tables = boolean
- https://commitfest.postgresql.org/11/831/
- PostgresPro Enterprise
- PostgresPro Enterprise 1C

В советском PostgreSQL

Статистика

Определяет

План

online_analyze

Расширение, призванное решить проблему сбора статистики для временных таблиц:

online_analyze

Расширение, призванное решить проблему сбора статистики для временных таблиц:

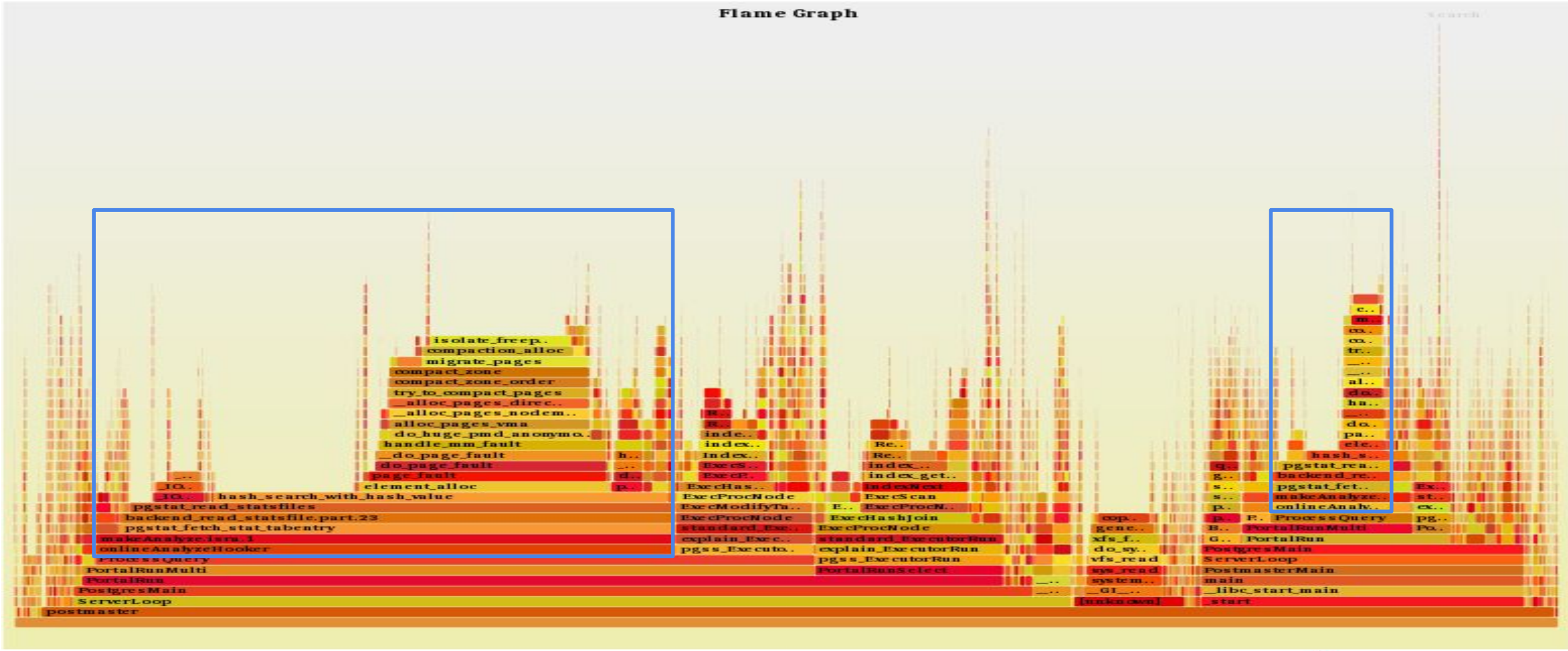
- Принудительно выполняет ANALYZE

online_analyze

Расширение, призванное решить проблему сбора статистики для временных таблиц:

- Принудительно выполняет ANALYZE
- Полагается на статистику stats collector`a

online_analyze

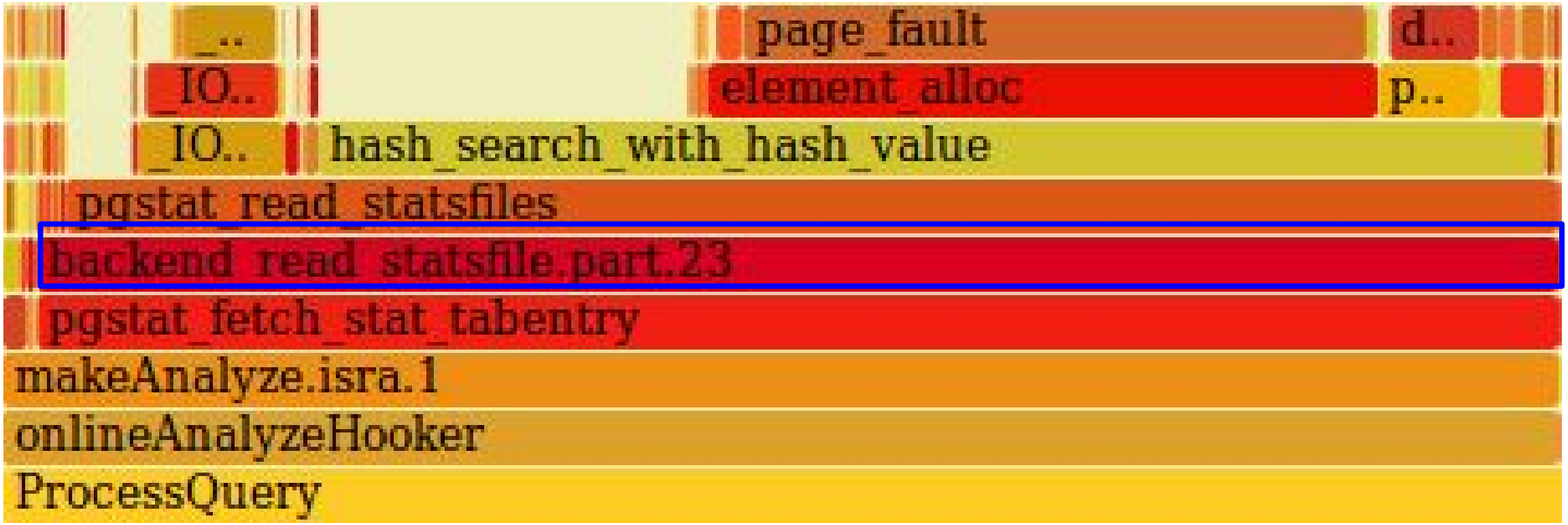


online_analyze

При большом кол-ве таблиц:

- Чтение файла со статистикой становится дорогим

online_analyze

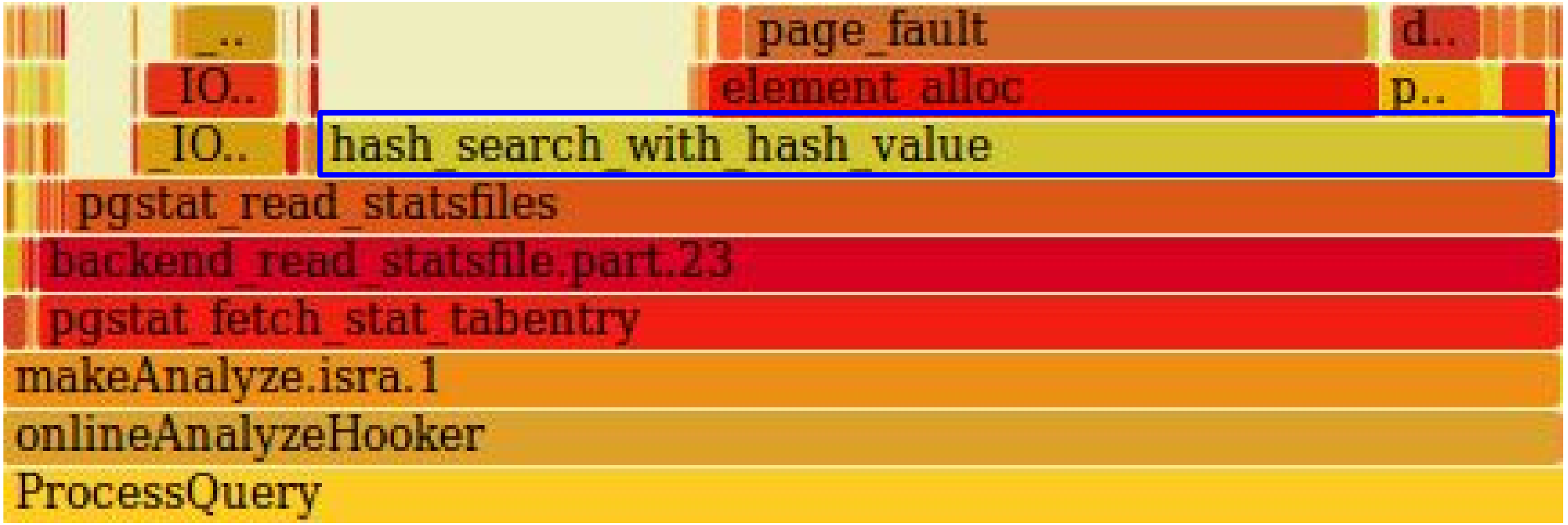


online_analyze

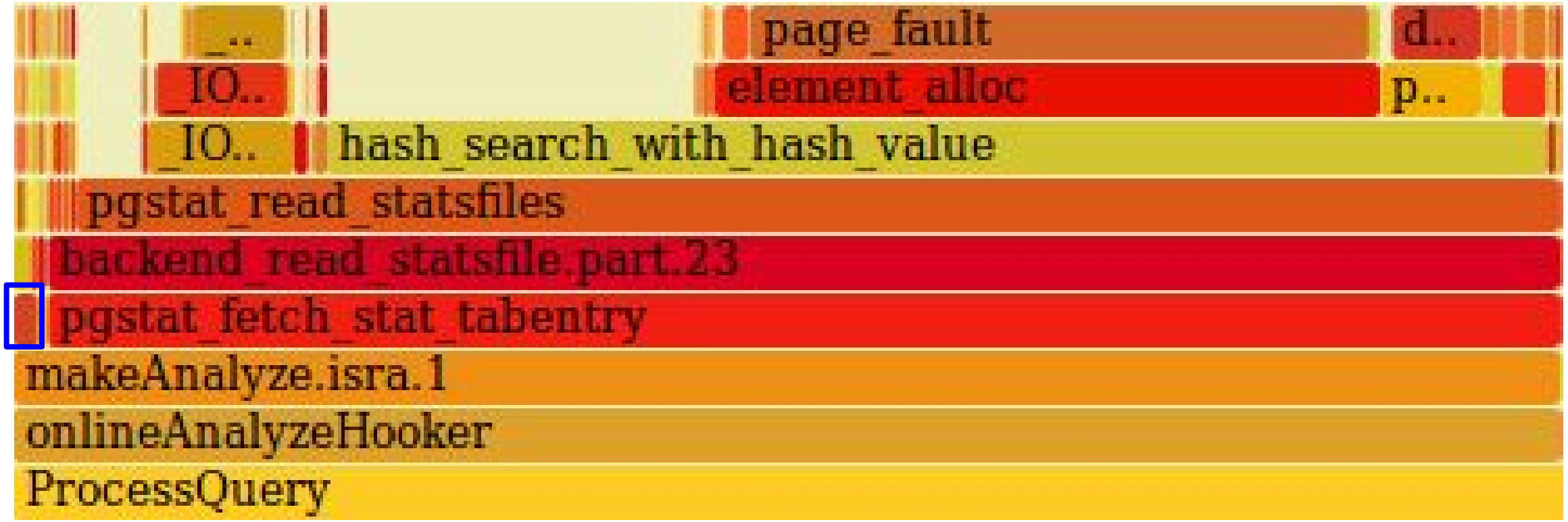
При большом кол-ве таблиц:

- Чтение файла со статистикой становится дорогим
- Поиск по прочитанному хэшу становится дорогим

online_analyze

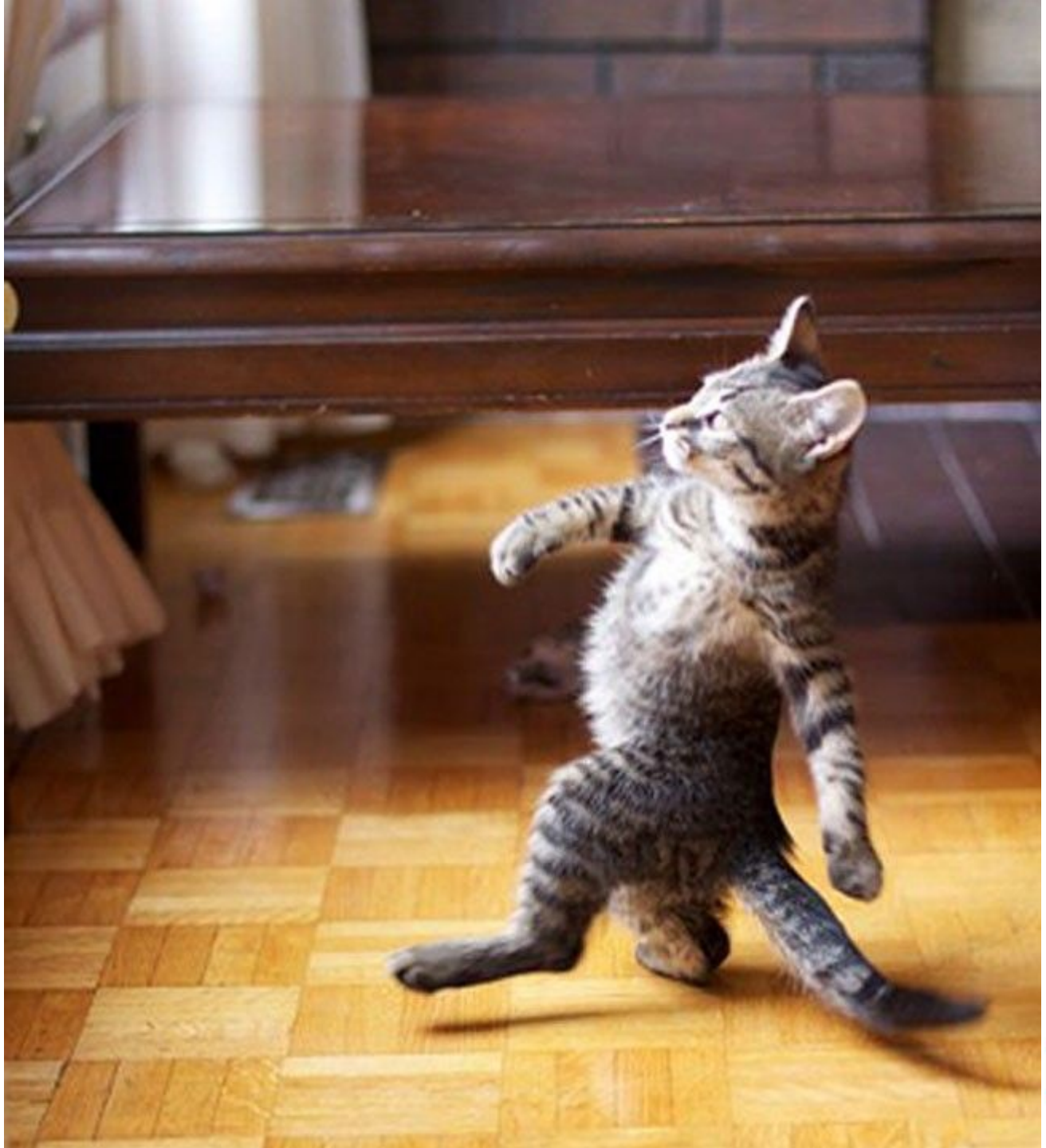


online_analyze



online_analyze

Отключить?

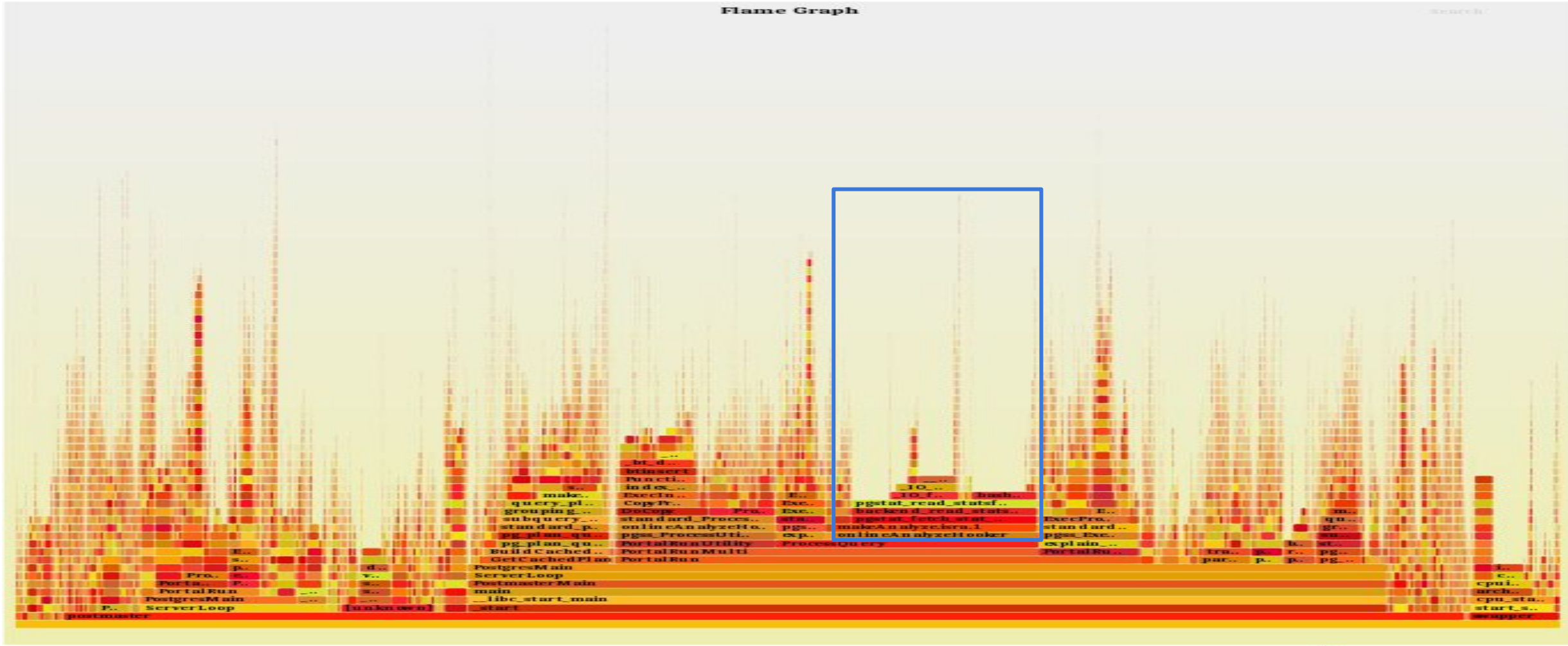


Патч №3: online_analyze

Автор: Фёдор Сигаев

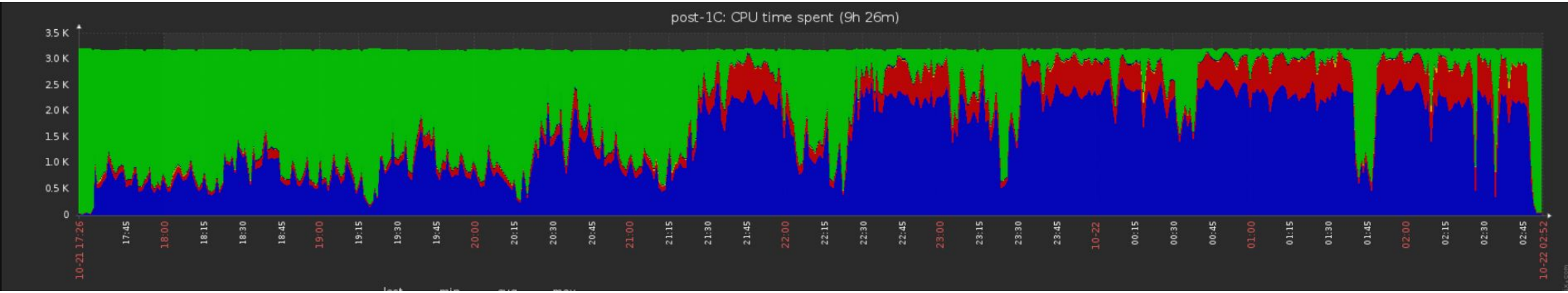
1. Реже обращаемся за статистикой
2. WIP - хранить статистику в локальной памяти

Патч для online_analyze



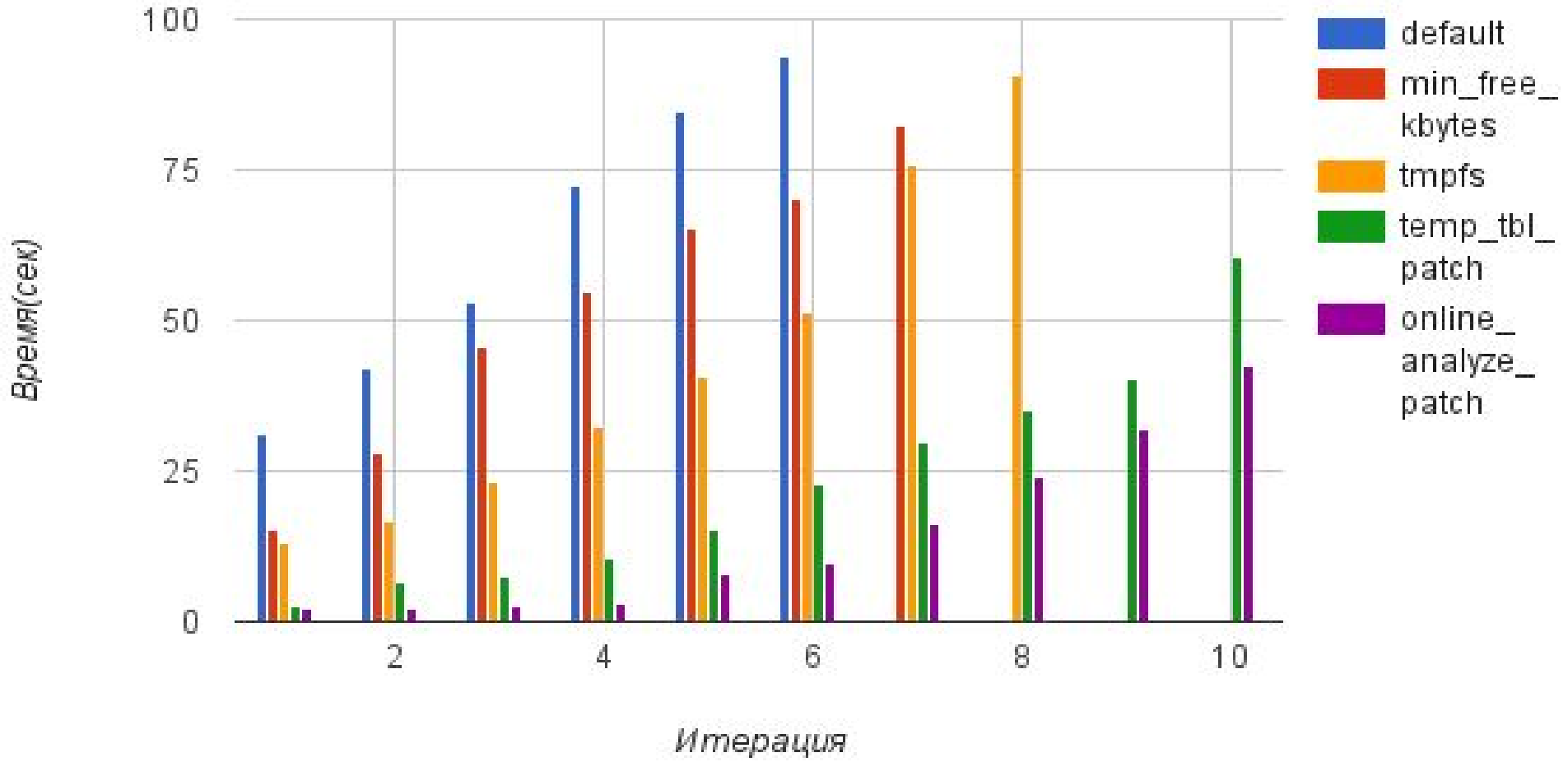
online_analyze patch

CPU LOAD



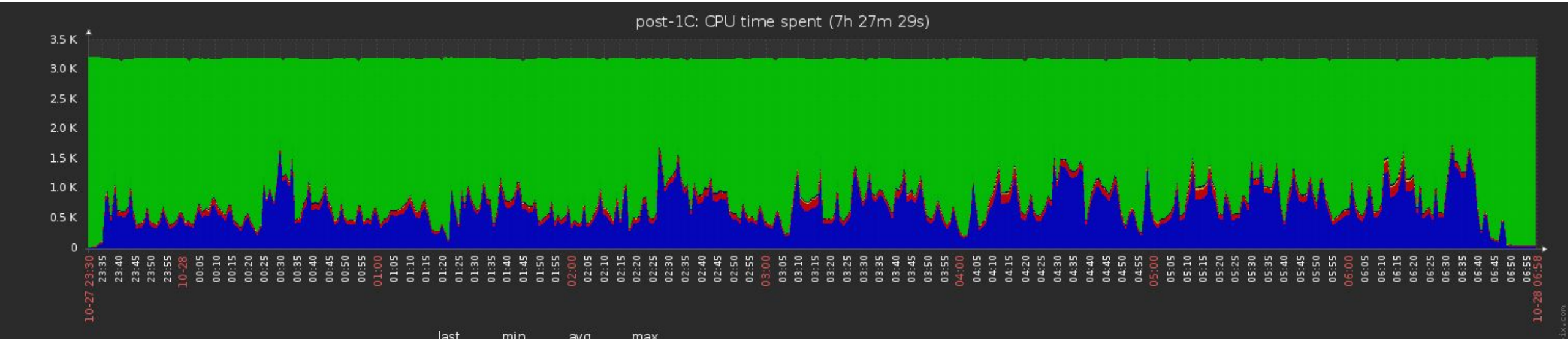
CPU: IDLE SYSTEM USER

online_analyze patch



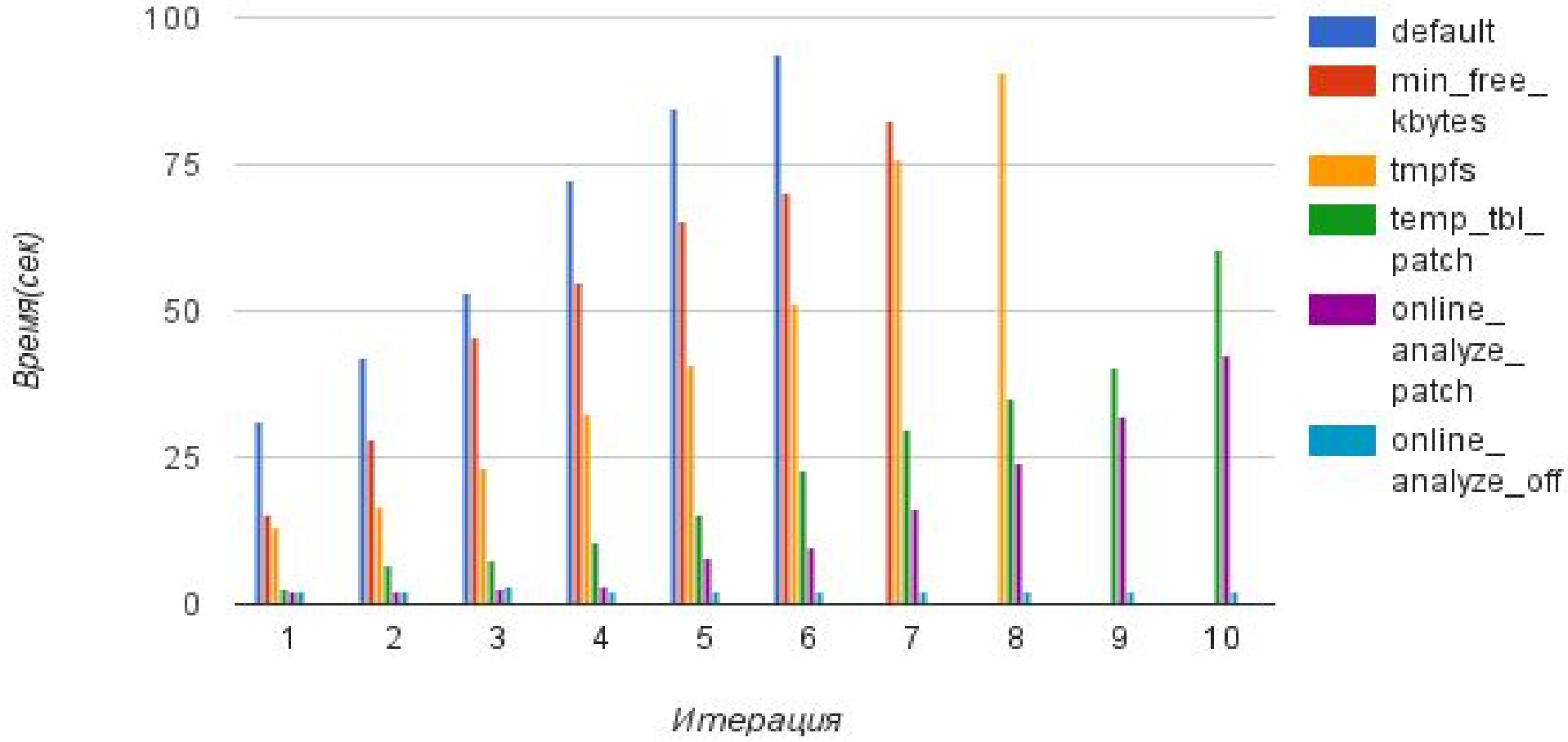
online_analyze off

CPU LOAD



CPU: IDLE SYSTEM USER

online analyze off



Вопросы ?

Спасибо за внимание!

Контакты:

g.smolkin@postgrespro.ru

s.petrov@postgrespro.ru

Итерация

Итерация Ключевая операция	Целевое время	Количество замеров	APDEX	Продолжительность Средняя	 Минима	 Макси
Итерация №1. BPM: 52		5 522	0,983	1,626	0,095	39,316
Авансовый отчет проведение	2,00	220	0,993	0,879	0,694	3,754
Груп. банковских выписок с 100 до 200 док	180,00	1	1,000	17,197	17,197	17,197
Документ список. авансовый отчет	1,00	220	1,000	0,175	0,097	0,203
Документ список. корректировка долга	1,00	600	1,000	0,171	0,095	0,194
Документ список. поступление товаров услуг	1,00	300	1,000	0,126	0,097	0,206
Документ список. приходный кассовый ордер	1,00	220	1,000	0,182	0,157	0,441
Документ список. расходный кассовый ордер	1,00	220	1,000	0,160	0,138	0,275
Документ список. реализация товаров услуг	1,00	300	1,000	0,128	0,104	0,377
Книга покупок по постановлению № 1137 за	30,00	100	1,000	12,413	11,864	14,205
Книга продаж по постановлению № 1137 за месяц	30,00	100	0,500	37,520	35,876	39,316
Корректировка долга проведение	2,00	600	0,949	1,479	0,893	7,383
Передача материалов в эксплуатацию проведение	2,00	360	0,997	0,613	0,410	2,513
Поступление на расчетный счет проведение	2,00	440	0,994	0,747	0,551	6,338
Поступление товаров услуг проведение	2,00	300	0,990	1,193	0,714	2,514
Приходный кассовый ордер проведение	2,00	220	0,998	0,761	0,500	5,344
Расходный кассовый ордер проведение	2,00	220	0,989	0,996	0,712	4,556
Реализация товаров услуг проведение	2,00	300	0,995	1,162	0,972	2,235
Регистрация с- ф на аванс заполнение т ч	150,00	1	1,000	21,437	21,437	21,437
Списание материалов из эксплуатации проведение	3,00	360	0,996	1,439	0,886	4,376
Списание с расчетного счета проведение	2,00	440	1,000	0,946	0,684	1,847