



IT NonStop
DataArt

Найти за полсекунды

Петр Петренко
Senior software developer

skyeng

Идея



habr

Публикации Пользователи Хабы Компании Песочница

 valbok 7 февраля 2014 в 14:21

Поиск изображений по фрагменту

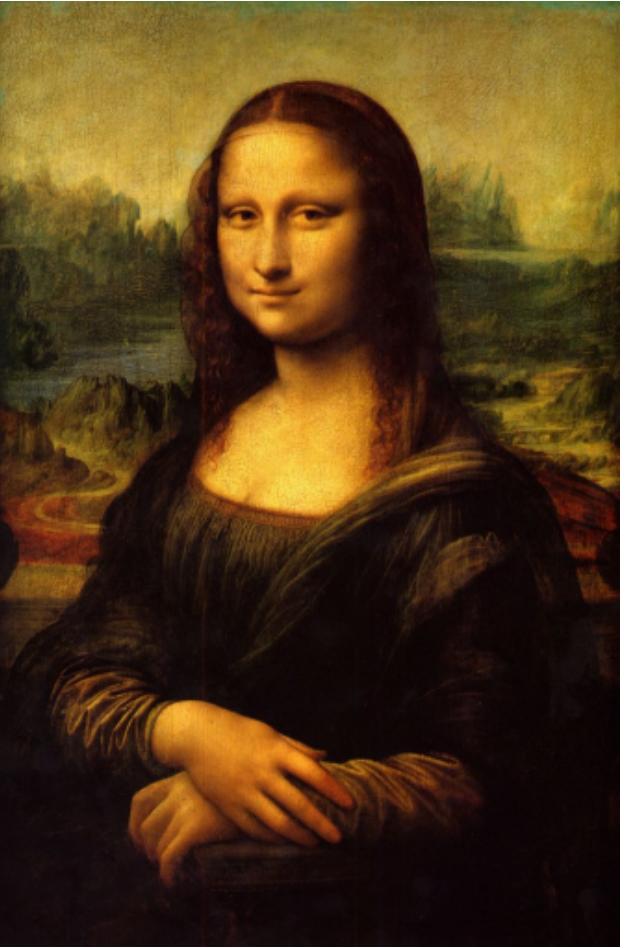
Обработка изображений, Алгоритмы



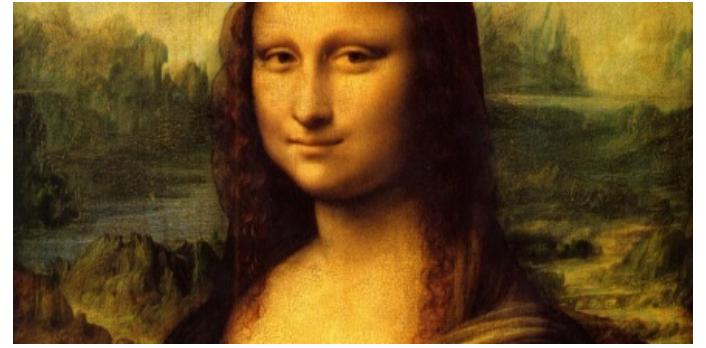
Схожие изображения



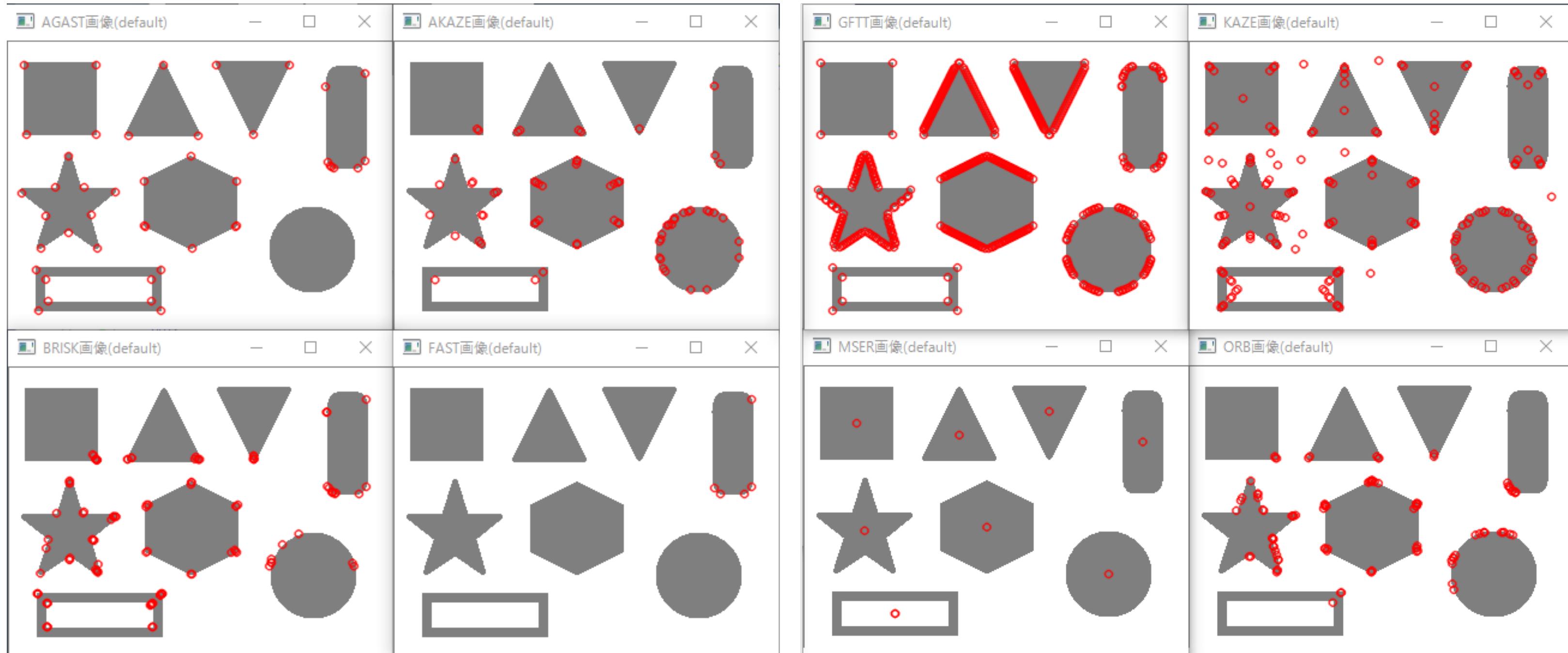
Миниатюры



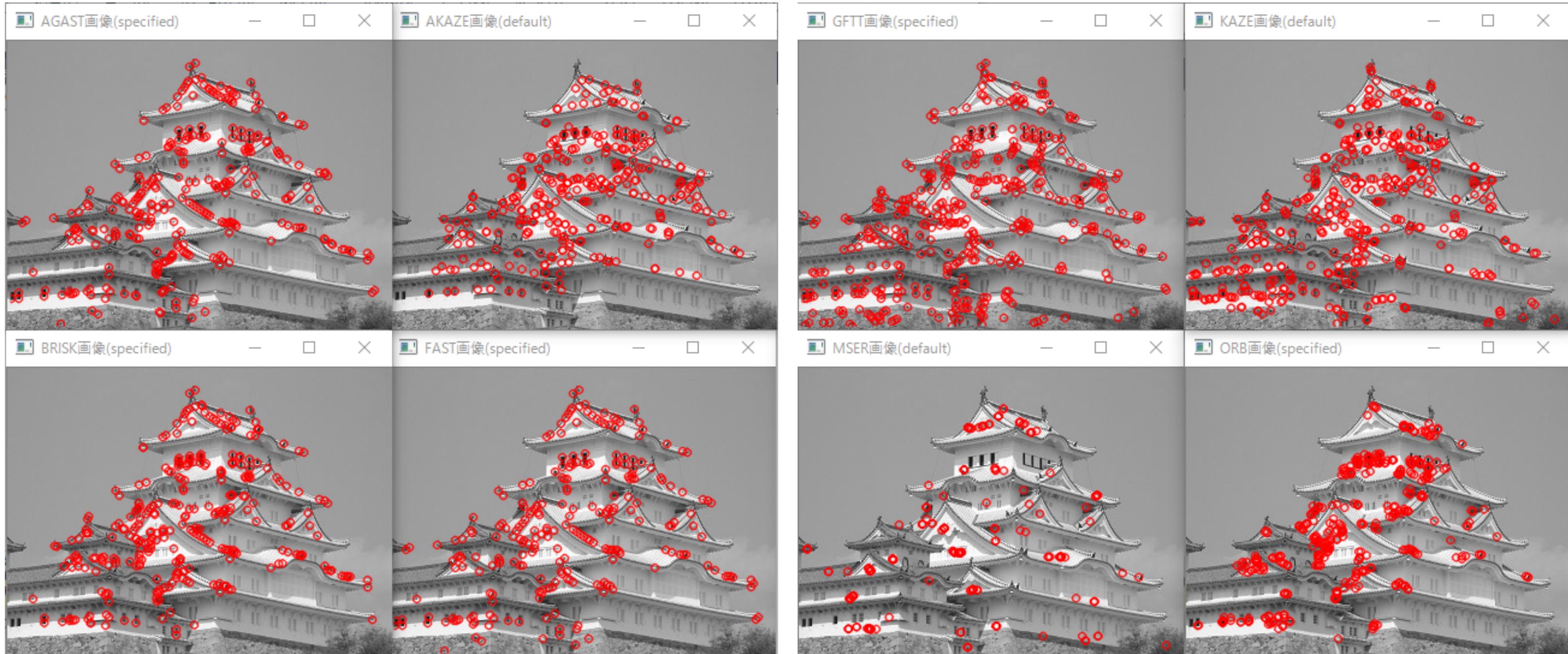
Полудубликаты



Ключевые точки



Ключевые точки

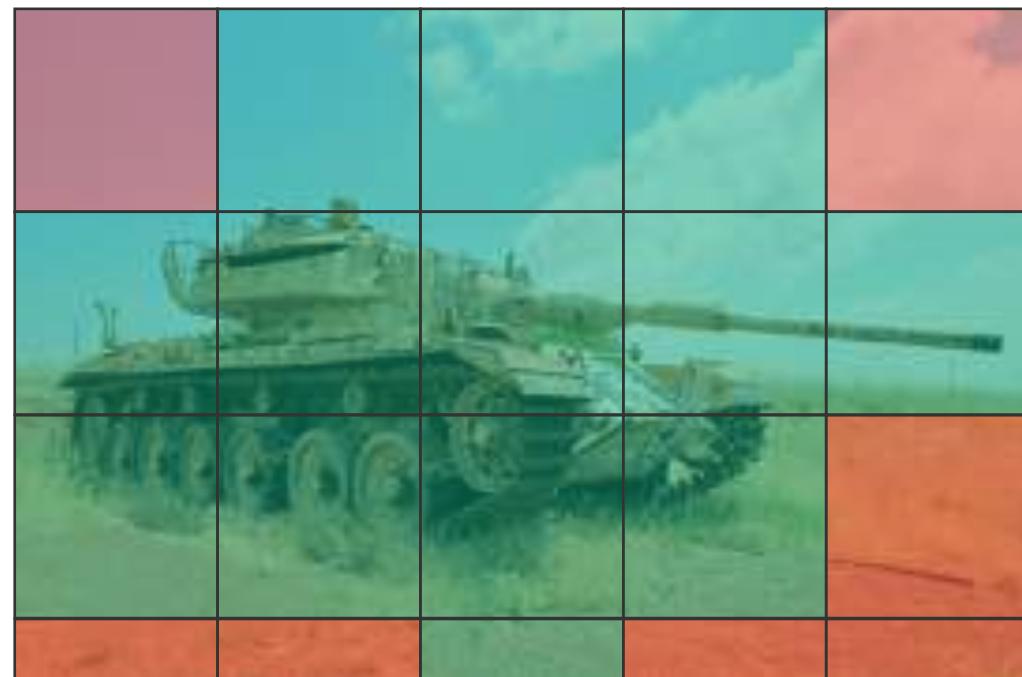


Алгоритмы поиска ключевых точек

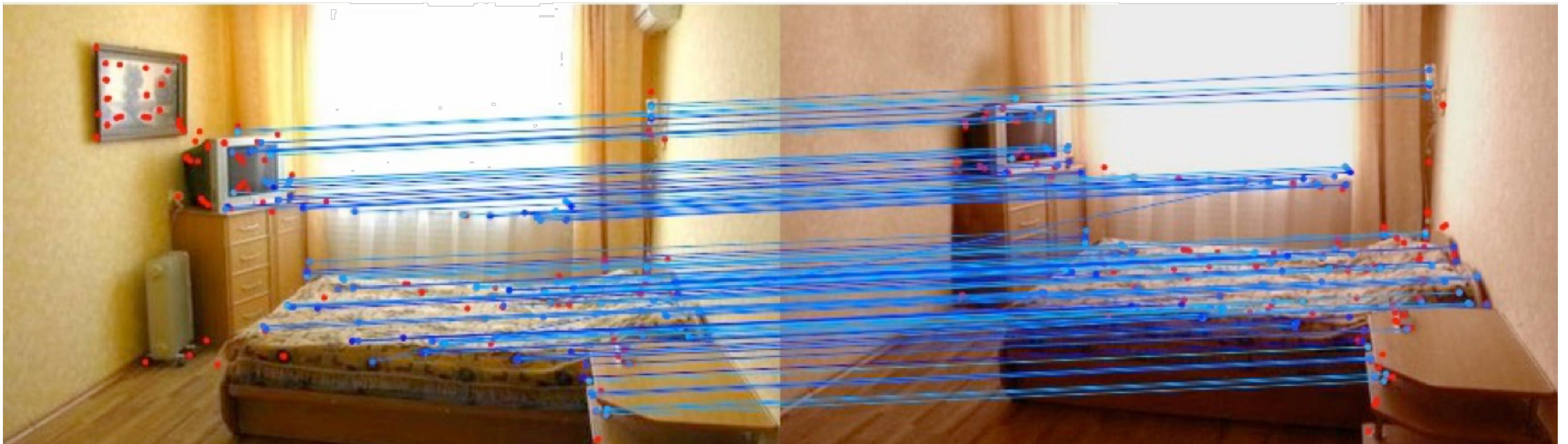


Алгоритм	Среднее покрытие	Максимальное покрытие	Минимальное покрытие	Среднее время выделения
FAST	81.60	237	4	<u>0.251 сек.</u>
BRISK	143.78	<u>367</u>	3	1.274 сек.
AKAZE	<u>155.57</u>	324	4	0.503 сек.

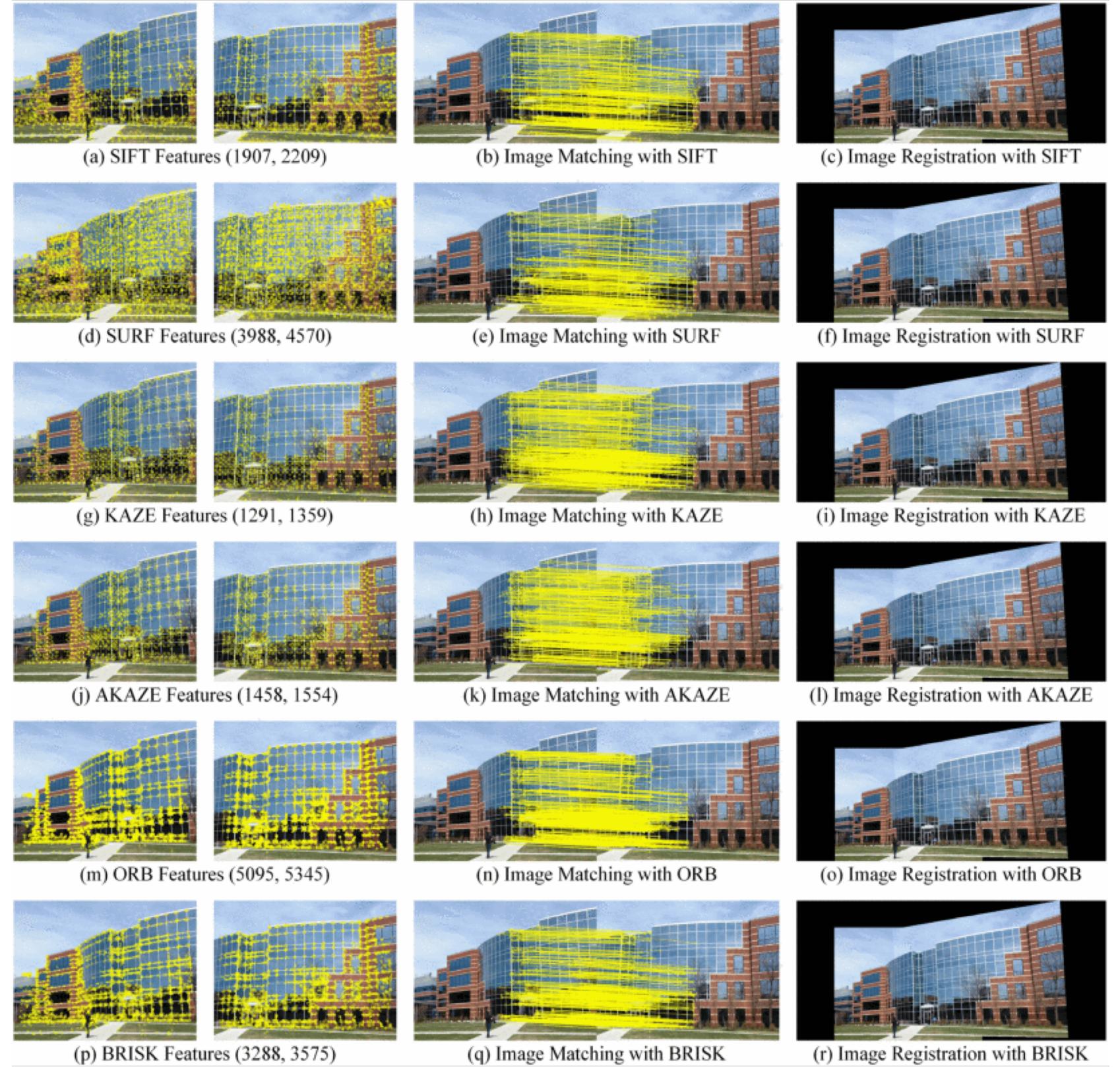
Покрытие



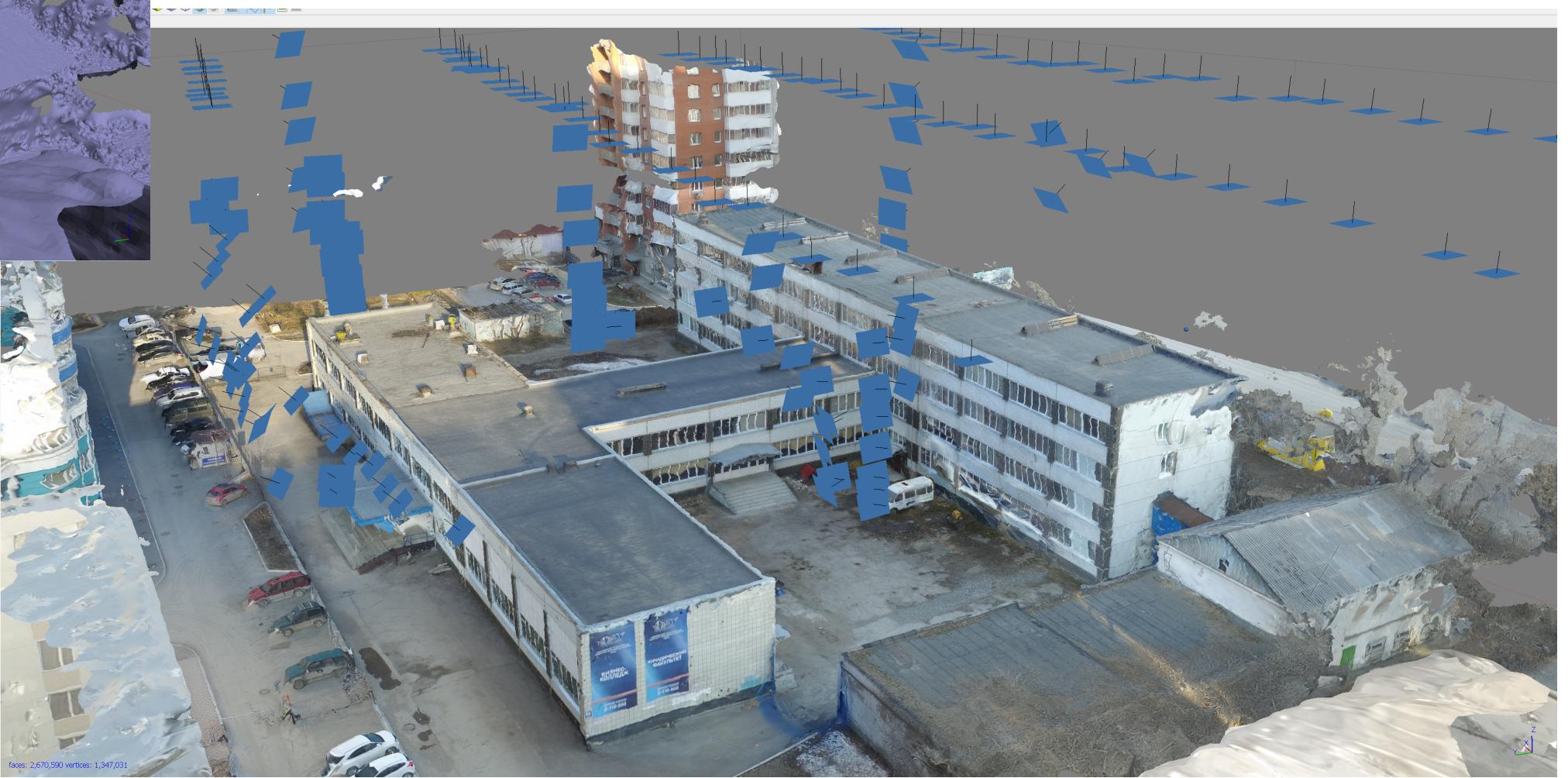
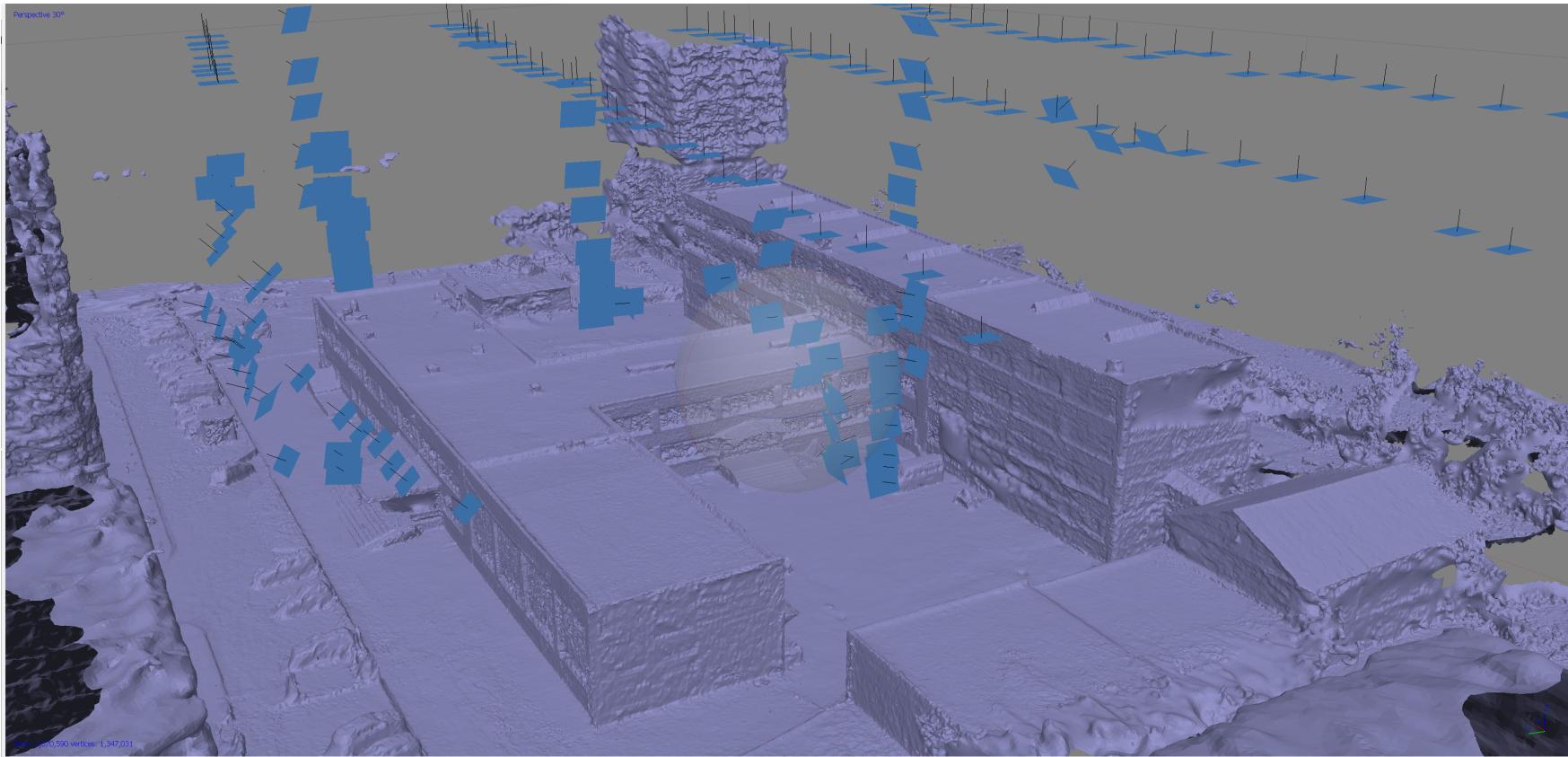
Сопоставление точек



Панорамы



3D модель по фото

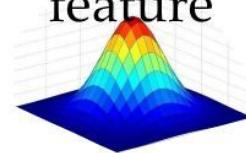




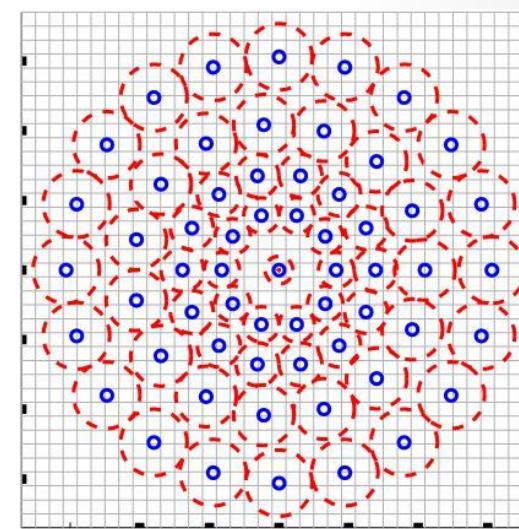
Дескрипторы

BRISK: Descriptor

2D Gaussian around each feature



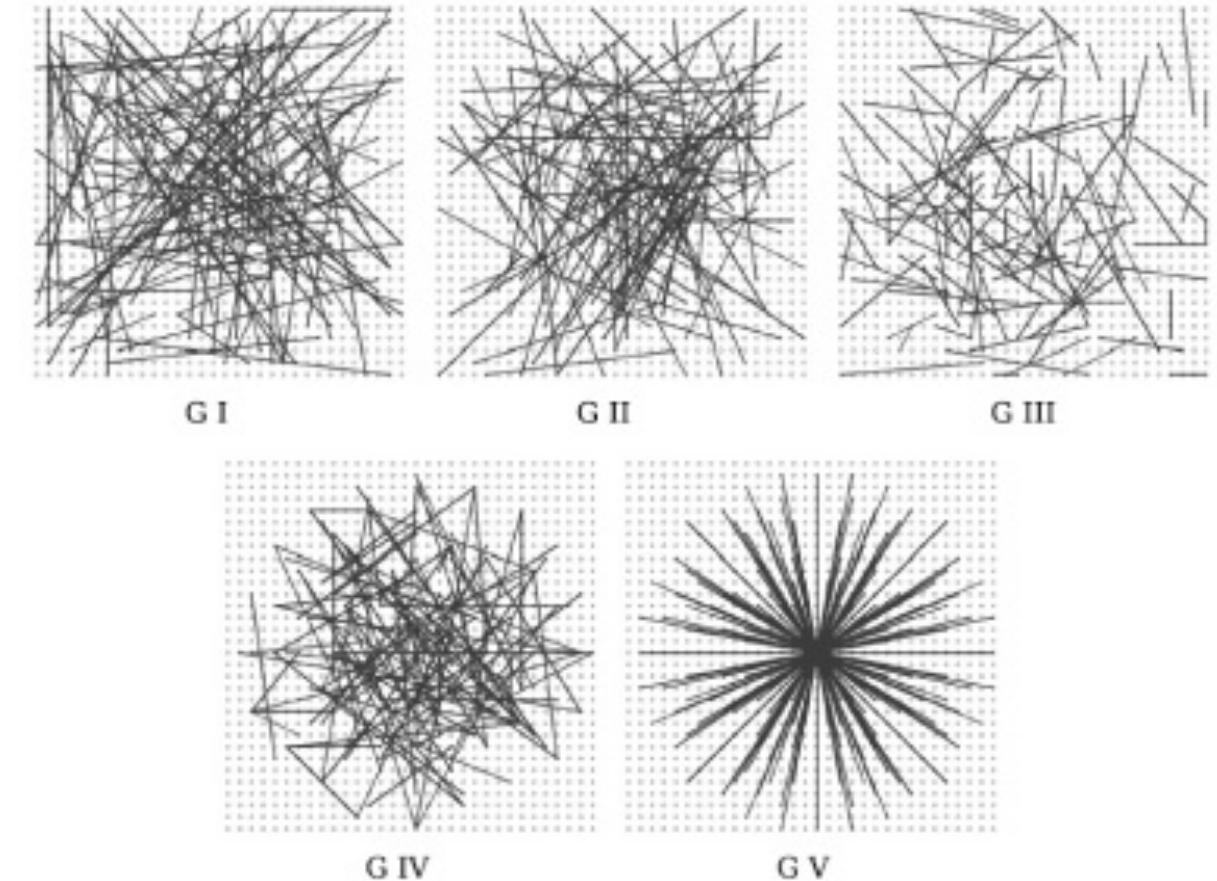
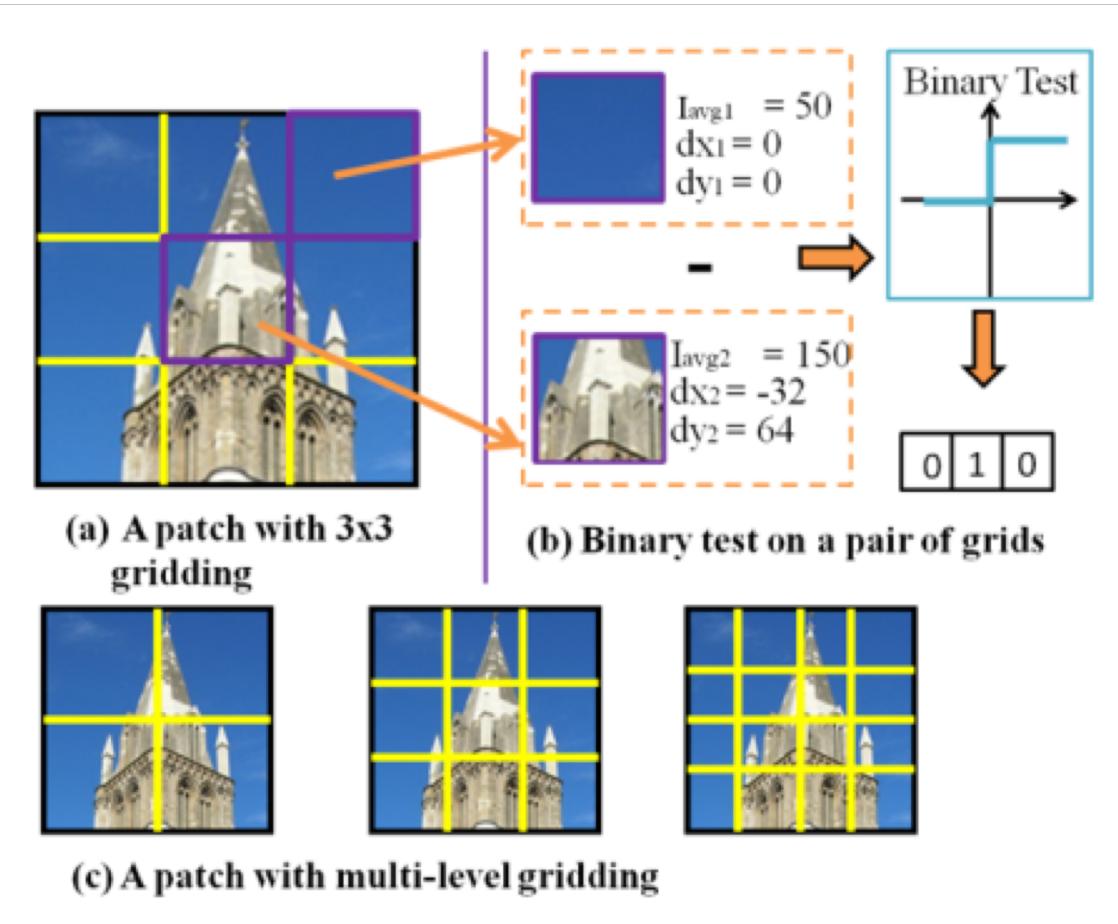
Robust to noise



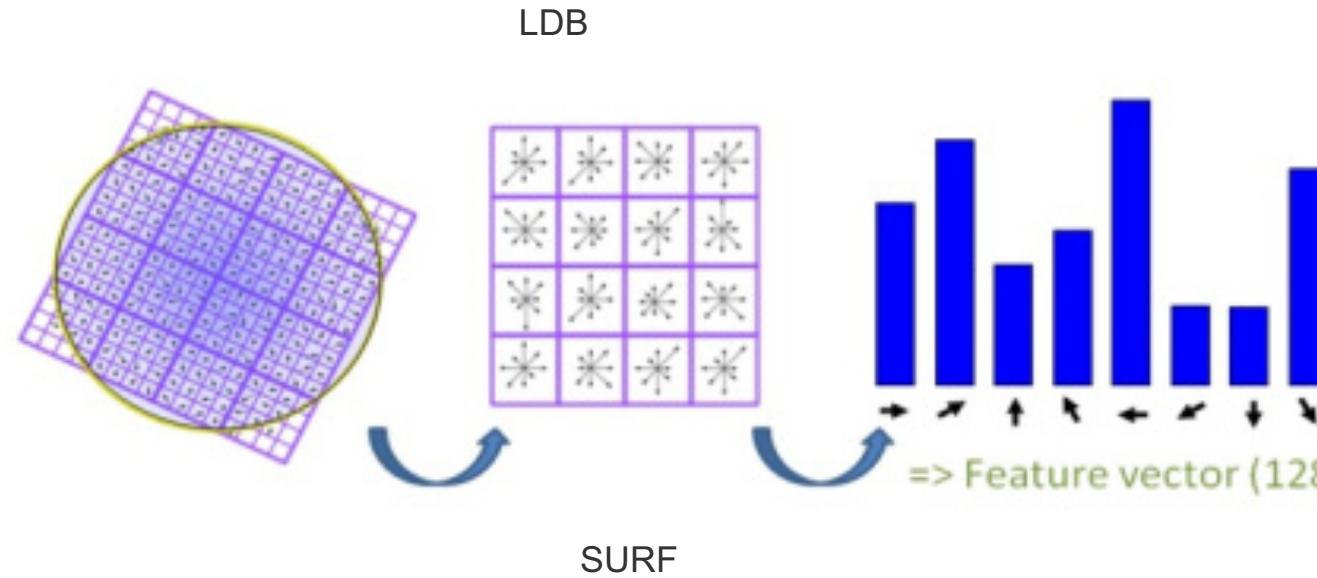
Centers: **BLUE** Gaussian: **RED**

32

slide: J. Heinly



BRISK



SURF

BRIEF

Сравнение дескрипторов

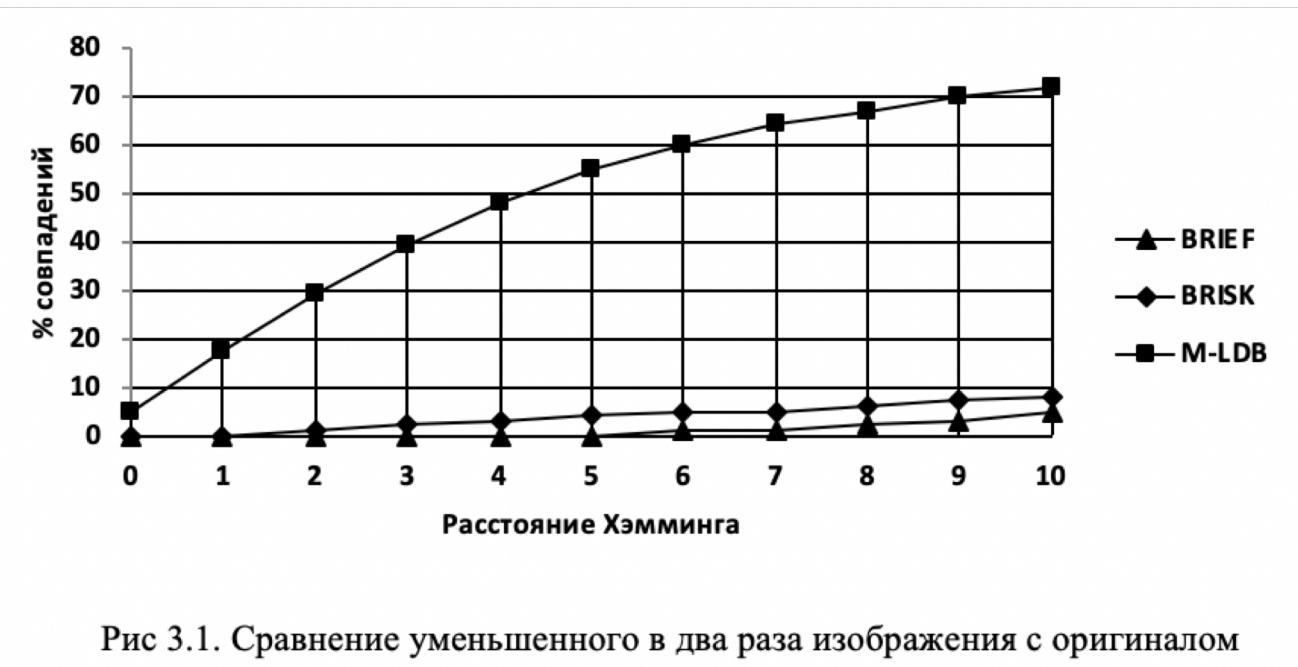


Рис 3.1. Сравнение уменьшенного в два раза изображения с оригиналом



Рис 3.3. Сравнение, обрезанного с разных сторон, изображения с оригиналом



Рис 3.4. Сравнение, обрезанного снизу, изображения с оригиналом

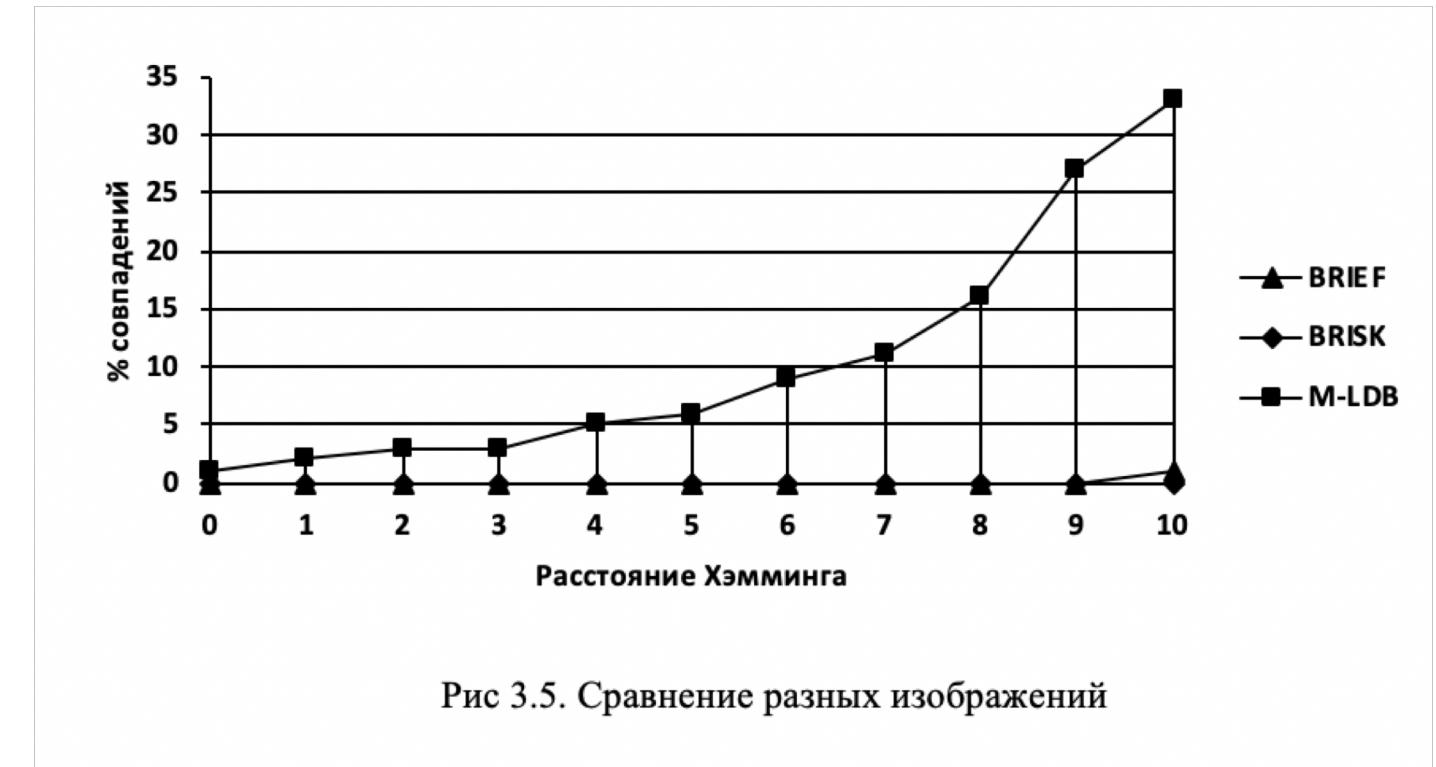
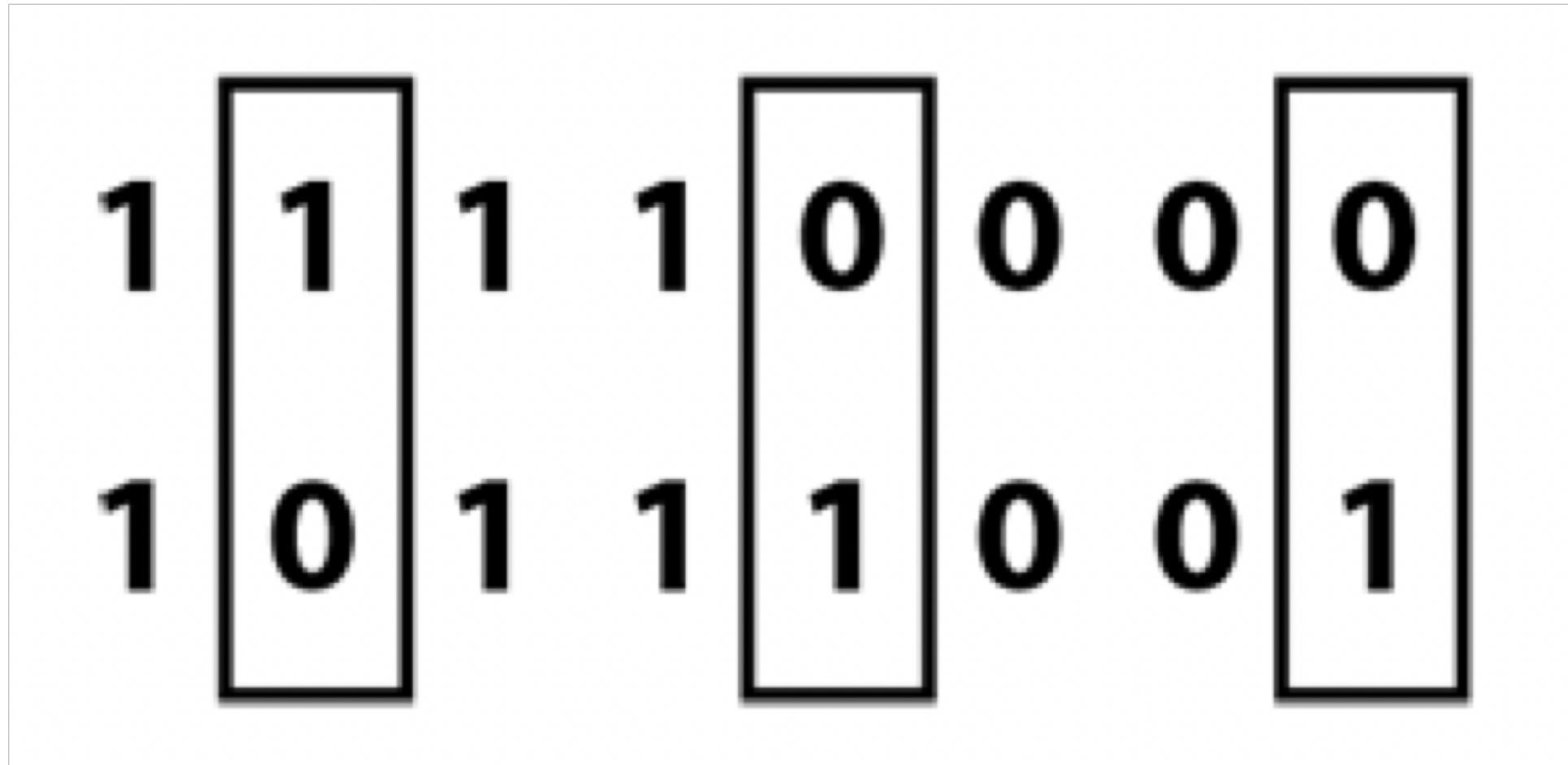
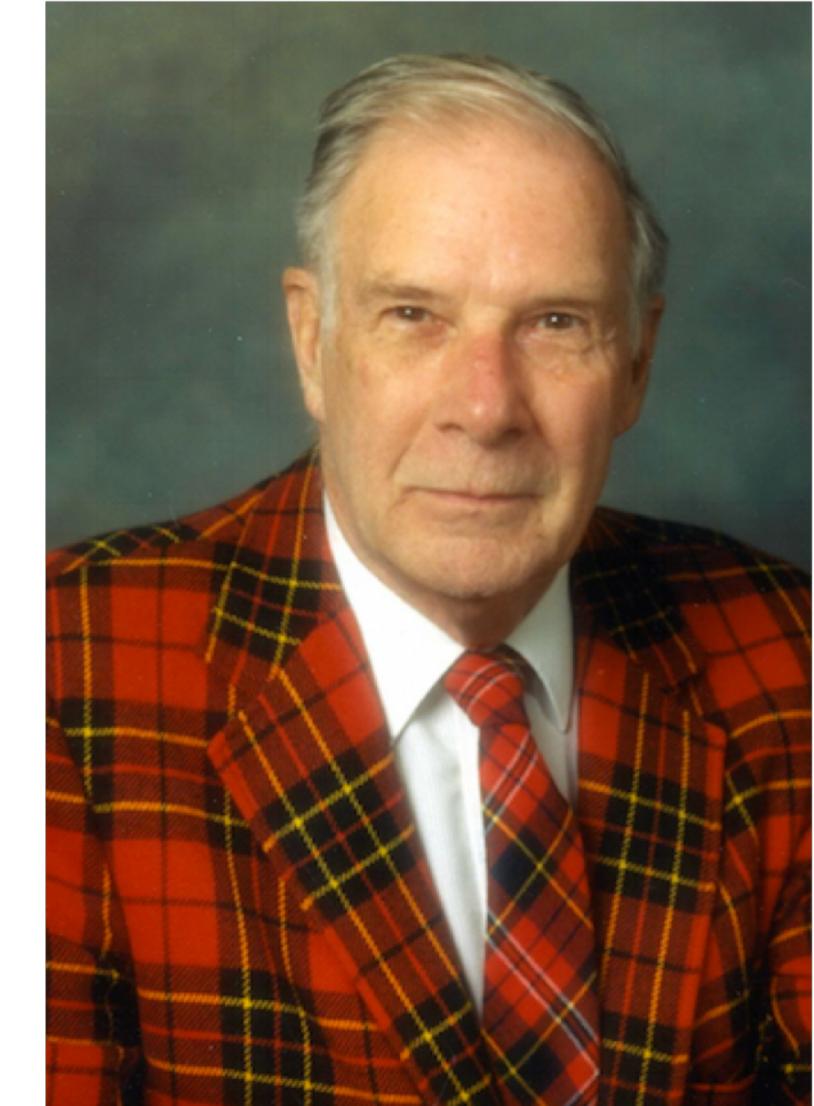


Рис 3.5. Сравнение разных изображений

Расстояние Хэмминга



Число позиций, в которых соответствующие символы двух слов одинаковой длины различны



Ричард Уэсли Хэмминг

Линейный поиск



Ключевых точек на одно изображение : **~350**

Тестовая база данных : **~350 миллионов точек**

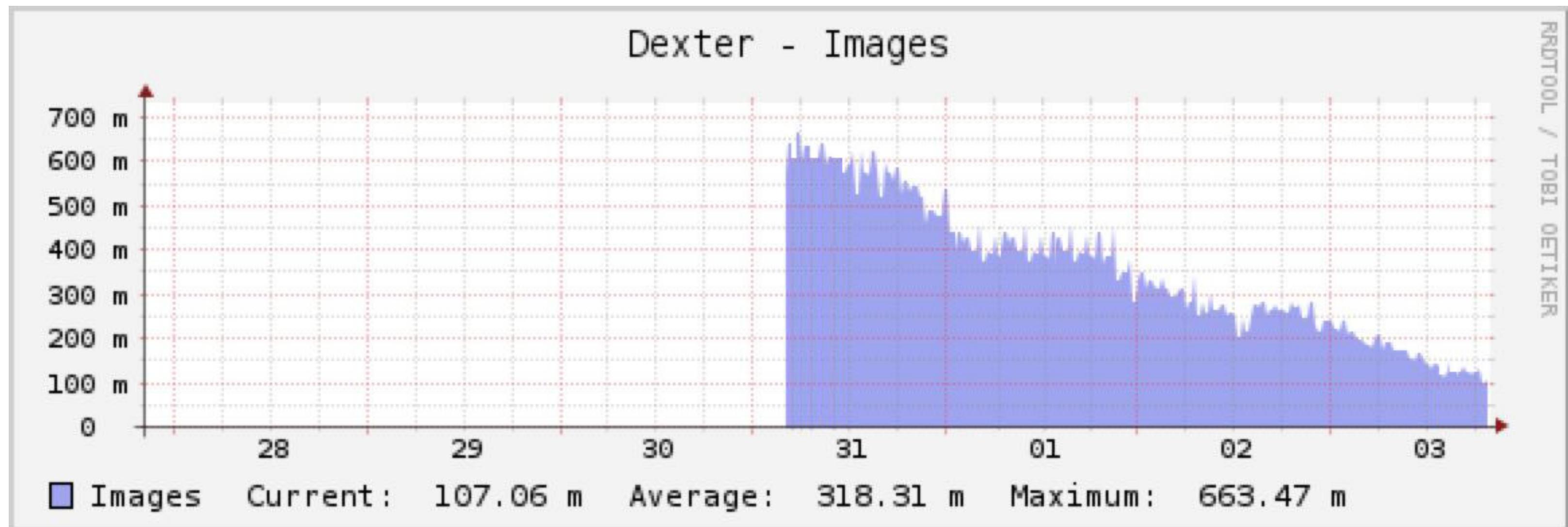
объем : **~ 10 гигабайт**

Линейный поиск одной ключевой точки в тестовой базе **~ 73.2 секунды** (без дисковых операций), всей картинки **более 7 часов.**

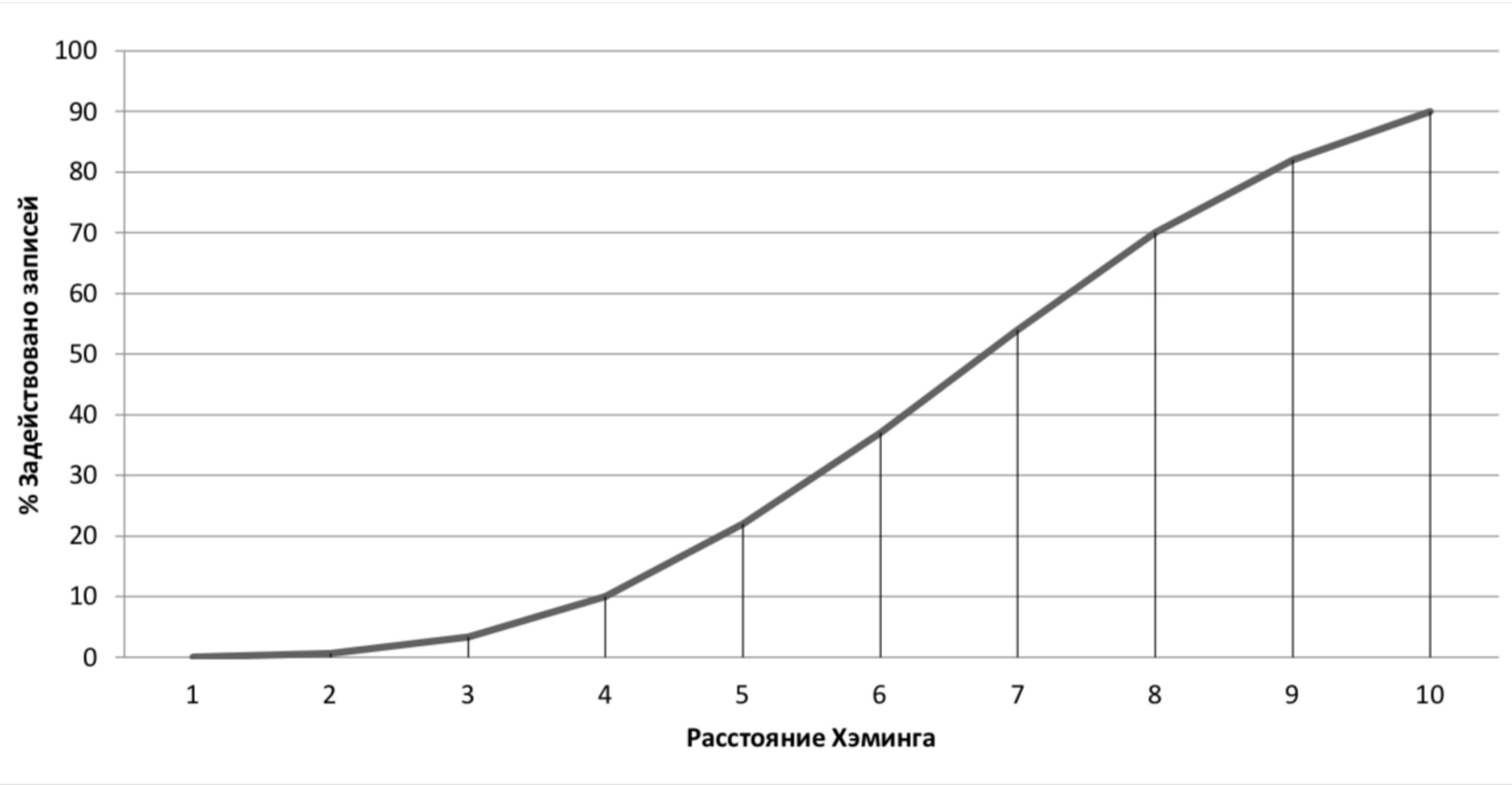
HEngine



Объем базы данных: ~ 85 гигабайт



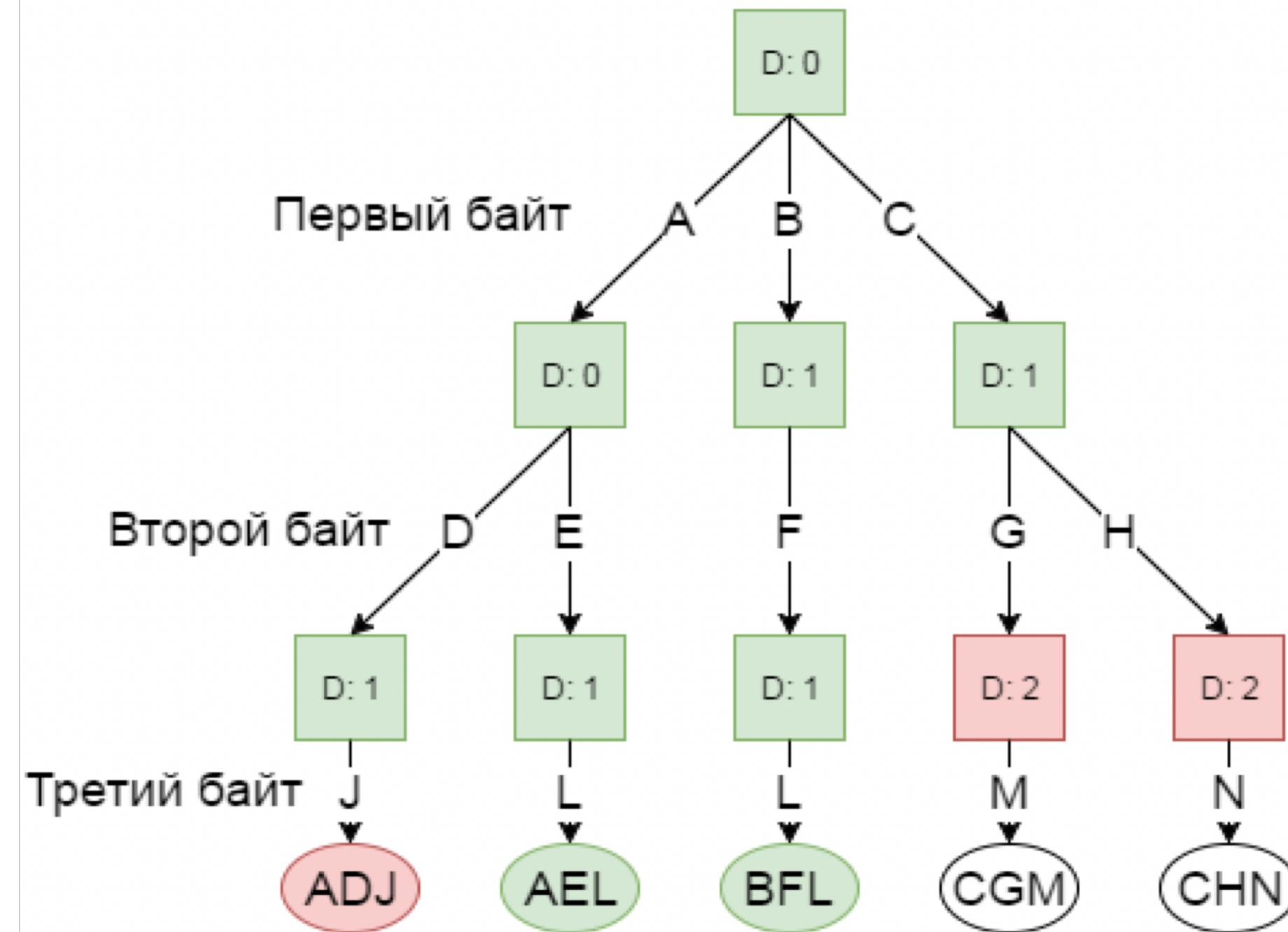
VP-дерево



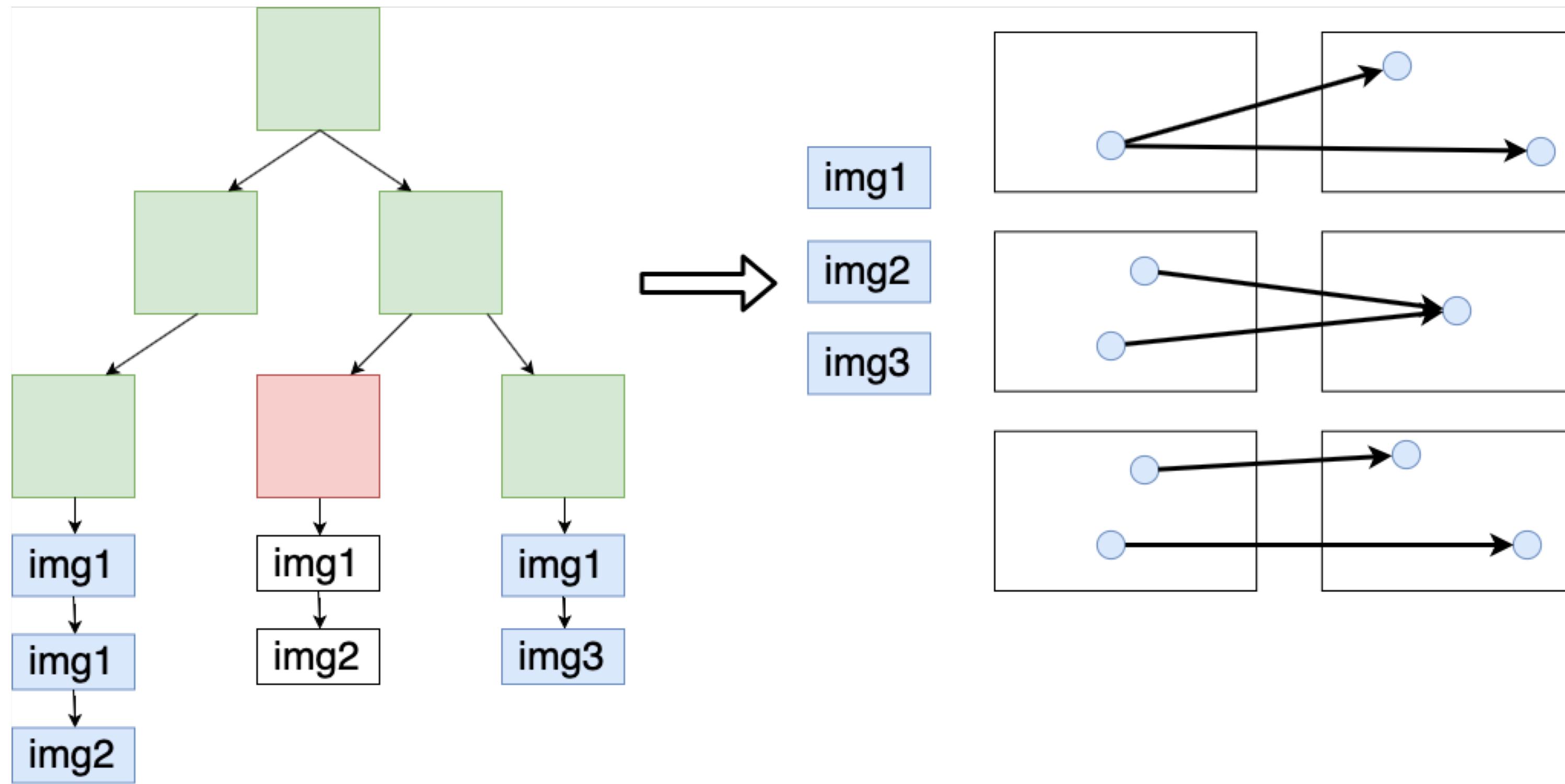


Префиксное дерево

Поиск строки AFL, максимальное расстояние 1



Поиск



Оптимизация



		Ключевых точек $10 * 10^9$			Узлов $3 * 10^9$	
		1 байт (среднее 3 записи) Количество потомков	1 байт Тип записи	4 байт Размер записи	8 байт Ссылка	
1 байт Размер префикса	N байт (среднее 5.8) Префикс					32% 19.8 байт
1 байт Размер префикса	N байт (среднее 5.8 байт) Префикс	1 байт Тип записи	4 байт Размер массива	N * 4 байт (среднее 8 байт) Ссылка		68% 19.8 байт
$19.8 * 3.3 + 1 = 66.34$ байта			База 185 GB			

Оптимизация



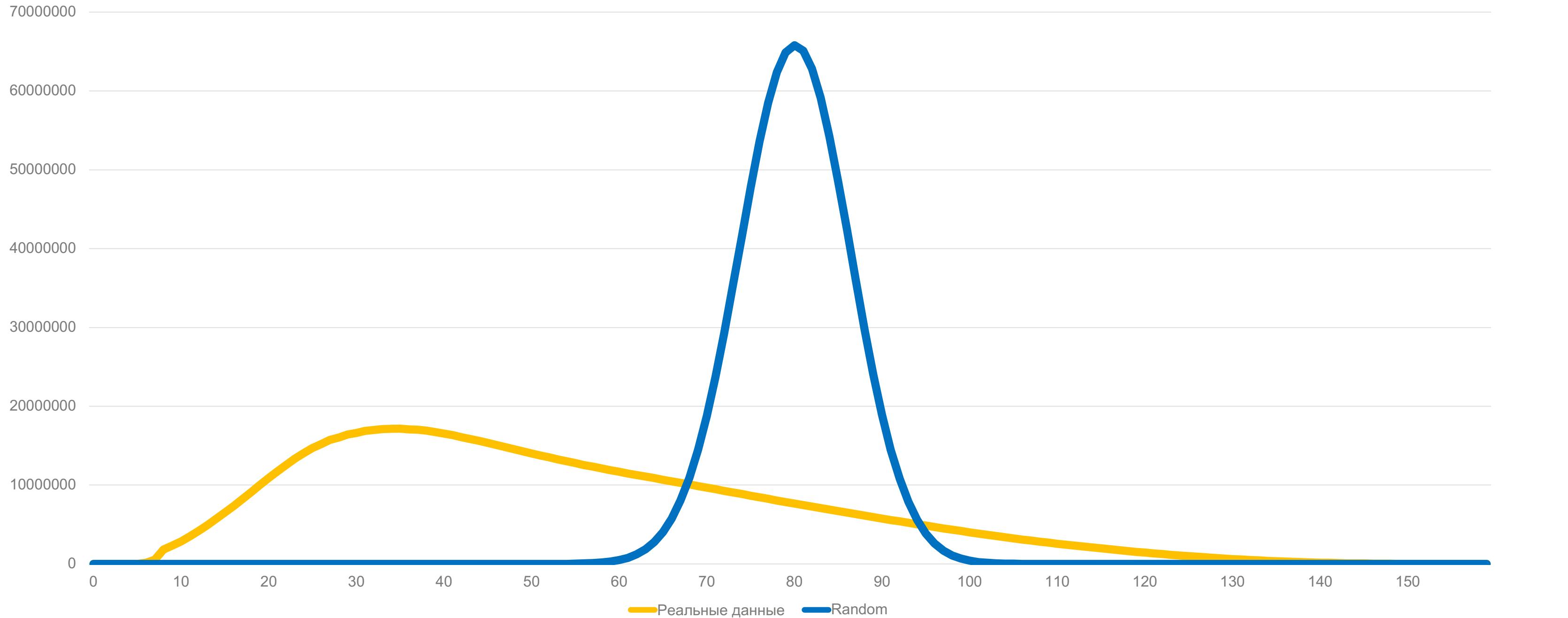
Ключевых точек $10 * 10^9$ Узлов $3 * 10^9$					Записей $10 * 10^9$ Узлов $3 * 10^9$																																												
<table border="1"> <tr> <td>1 байт (среднее 3 записи) Количество потомков</td> <td></td> <td></td> <td></td> <td></td> </tr> <tr> <td>1 байт Размер префикса</td> <td>N байт (среднее 5.8) Префикс</td> <td>1 байт Тип записи</td> <td>4 байт Размер записи</td> <td>8 байт Ссылка</td> </tr> <tr> <td>1 байт Размер префикса</td> <td>N байт (среднее 5.8 байт) Префикс</td> <td>1 байт Тип записи</td> <td>4 байт Размер массива</td> <td>N * 4 байт (среднее 8 байт) Ссылка</td> </tr> </table>					1 байт (среднее 3 записи) Количество потомков					1 байт Размер префикса	N байт (среднее 5.8) Префикс	1 байт Тип записи	4 байт Размер записи	8 байт Ссылка	1 байт Размер префикса	N байт (среднее 5.8 байт) Префикс	1 байт Тип записи	4 байт Размер массива	N * 4 байт (среднее 8 байт) Ссылка	<table border="1"> <tr> <td>3 бита Тип записи</td> <td>5 бит Размер префикса</td> <td>N байт (среднее 5.8) Префикс</td> <td>5 байт Ссылка</td> <td></td> </tr> <tr> <td>3 бита Тип записи</td> <td>5 бит Размер префикса</td> <td>N байт (среднее 5.8 байт) Префикс</td> <td>4 байта Id картинки</td> <td></td> </tr> <tr> <td>3 бита Тип записи</td> <td>5 бит Размер префикса</td> <td>N байт (среднее 5.8 байт) Префикс</td> <td>4 байта Id картинки</td> <td>4 байта Id картинки</td> </tr> <tr> <td>3 бита Тип записи</td> <td>5 бит Размер префикса</td> <td>N байт (среднее 5.8 байт) Префикс</td> <td>4 байта Id картинки</td> <td>4 байта Id картинки</td> </tr> <tr> <td>3 бита Тип записи</td> <td>5 бит Размер префикса</td> <td>N байт (среднее 5.8 байт) Префикс</td> <td>4 байт Размер массива</td> <td>N * 4 байт (среднее 41.5 байт) Ссылка</td> </tr> </table>					3 бита Тип записи	5 бит Размер префикса	N байт (среднее 5.8) Префикс	5 байт Ссылка		3 бита Тип записи	5 бит Размер префикса	N байт (среднее 5.8 байт) Префикс	4 байта Id картинки		3 бита Тип записи	5 бит Размер префикса	N байт (среднее 5.8 байт) Префикс	4 байта Id картинки	4 байта Id картинки	3 бита Тип записи	5 бит Размер префикса	N байт (среднее 5.8 байт) Префикс	4 байта Id картинки	4 байта Id картинки	3 бита Тип записи	5 бит Размер префикса	N байт (среднее 5.8 байт) Префикс	4 байт Размер массива	N * 4 байт (среднее 41.5 байт) Ссылка
1 байт (среднее 3 записи) Количество потомков																																																	
1 байт Размер префикса	N байт (среднее 5.8) Префикс	1 байт Тип записи	4 байт Размер записи	8 байт Ссылка																																													
1 байт Размер префикса	N байт (среднее 5.8 байт) Префикс	1 байт Тип записи	4 байт Размер массива	N * 4 байт (среднее 8 байт) Ссылка																																													
3 бита Тип записи	5 бит Размер префикса	N байт (среднее 5.8) Префикс	5 байт Ссылка																																														
3 бита Тип записи	5 бит Размер префикса	N байт (среднее 5.8 байт) Префикс	4 байта Id картинки																																														
3 бита Тип записи	5 бит Размер префикса	N байт (среднее 5.8 байт) Префикс	4 байта Id картинки	4 байта Id картинки																																													
3 бита Тип записи	5 бит Размер префикса	N байт (среднее 5.8 байт) Префикс	4 байта Id картинки	4 байта Id картинки																																													
3 бита Тип записи	5 бит Размер префикса	N байт (среднее 5.8 байт) Префикс	4 байт Размер массива	N * 4 байт (среднее 41.5 байт) Ссылка																																													
32% 19.8 байт					32% 11.8 байт																																												
68% 19.8 байт					55% 10.8 байт																																												
19.8 * 3.3 + 1 = 66.34 байта					8% 14.8 байт																																												
База 185 GB					2% 18.8 байт																																												
12.8 * 3.3 + 1 = 43.24 байта					3% 52.3 байт																																												
База 120 GB																																																	

Оптимизация



		1 байт (среднее 3 записи) Количество потомков		Записей $10 * 10^9$ Узлов $3 * 10^9$				1 байт (среднее 3 записи) Количество потомков		Записей $10 * 10^9$			
3 бита Тип записи	5 бит Размер префикса	N байт (среднее 5.8) Префикс	5 байт Ссылка	32% 11.8 байт		3 бита Тип записи	5 бит Размер префикса	N байт (среднее 5.8) Префикс	5 байт Ссылка	32% 11.8 байт			
3 бита Тип записи	5 бит Размер префикса	N байт (среднее 5.8 байт) Префикс	4 байта Id картинки	55% 10.8 байт		3 бита Тип записи	5 бит Размер префикса	N байт (среднее 5.8 байт) Префикс	3 байта Id картинки	55% 9.8 байт			
3 бита Тип записи	5 бит Размер префикса	N байт (среднее 5.8 байт) Префикс	4 байта Id картинки	4 байта Id картинки	4 байта Id картинки	8% 14.8 байт		3 бита Тип записи	5 бит Размер префикса	N байт (среднее 5.8 байт) Префикс	3 байта Id картинки	3 байта Id картинки	8% 12.8 байт
3 бита Тип записи	5 бит Размер префикса	N байт (среднее 5.8 байт) Префикс	4 байта Id картинки	4 байта Id картинки	4 байта Id картинки	2% 18.8 байт		3 бита Тип записи	5 бит Размер префикса	N байт (среднее 5.8 байт) Префикс	3 байта Id картинки	3 байта Id картинки	2% 15.8 байт
3 бита Тип записи	5 бит Размер префикса	N байт (среднее 5.8 байт) Префикс	4 байт Размер массива	N * 4 байт (среднее 41.5 байт) Ссылка		3% 52.3 байт		3 бита Тип записи	5 бит Размер префикса	N байт (среднее 5.8 байт) Префикс	2 байт Размер массива	N * 3 байт (среднее 31.1 байт) Ссылка	3% 39.9 байт
12.8 * 3.3 + 1 = 43.24 байта		База 120 GB				11.7 * 3.3 + 1 = 39.61 байта		База 110 GB		-75 GB			

Неравномерное распределение





Префиксное дерево

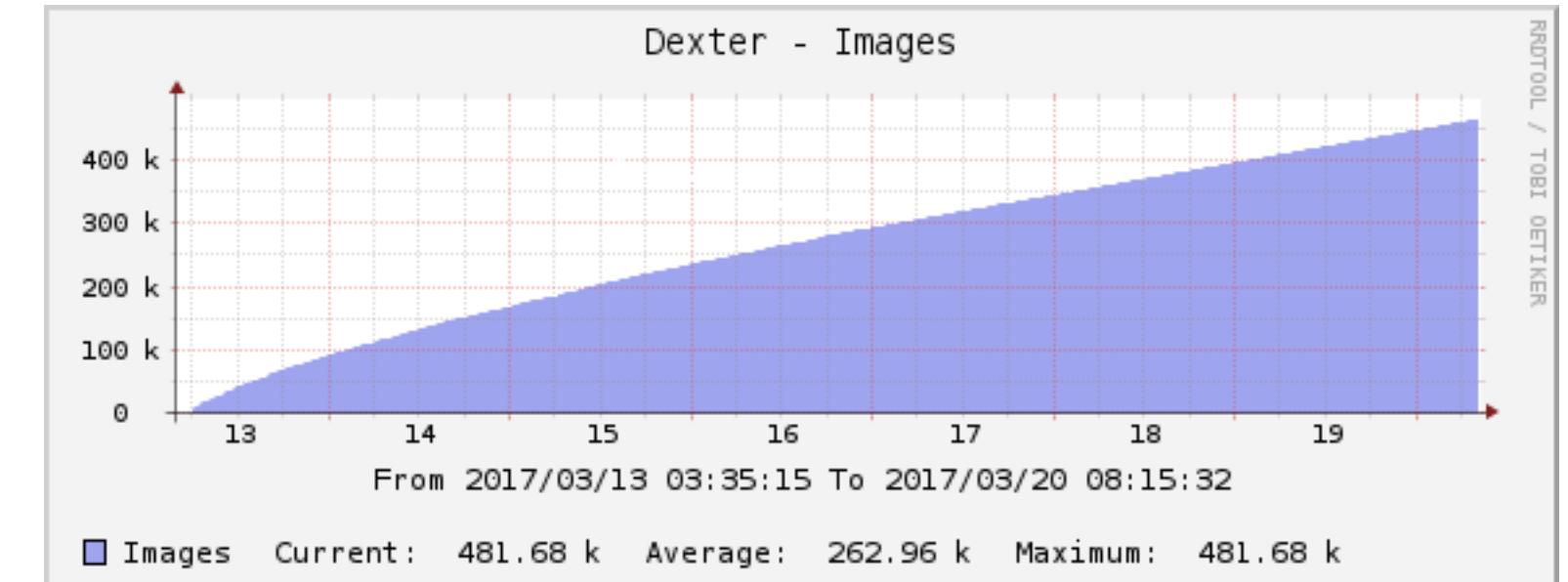
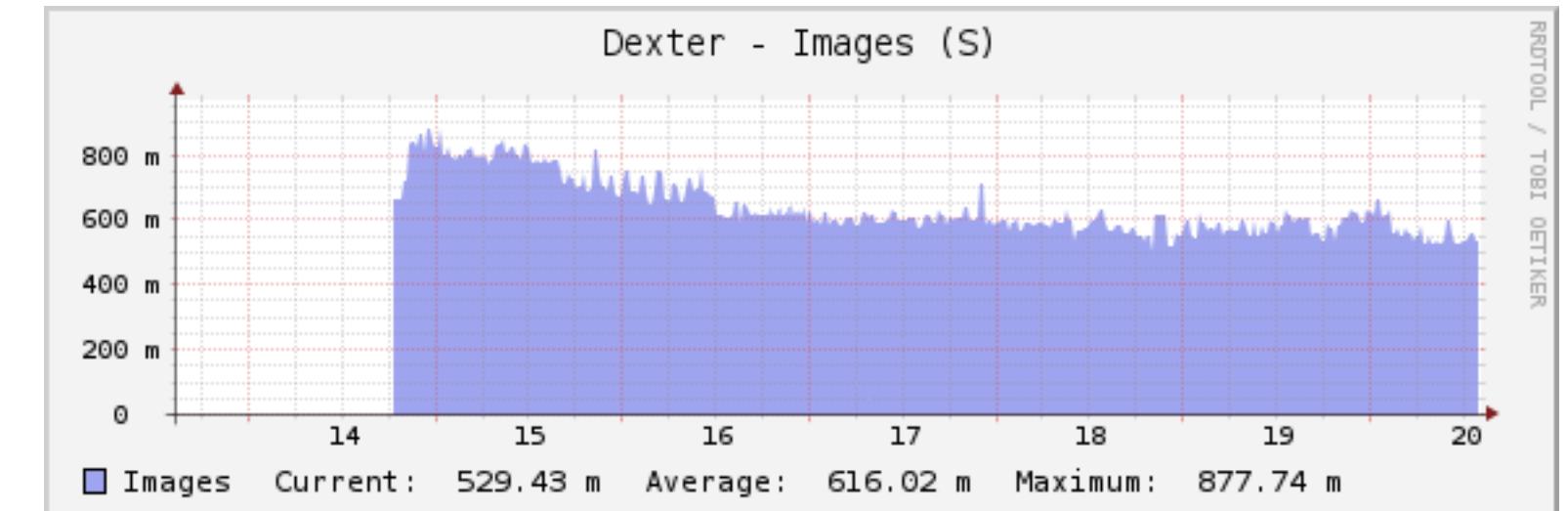
Объем файла с деревом: ~ 7 гигабайт

Поиск одной ключевой точки в тестовой базе примерно **3.5 миллисекунды**, всей картинки примерно **1.3 секунда**.

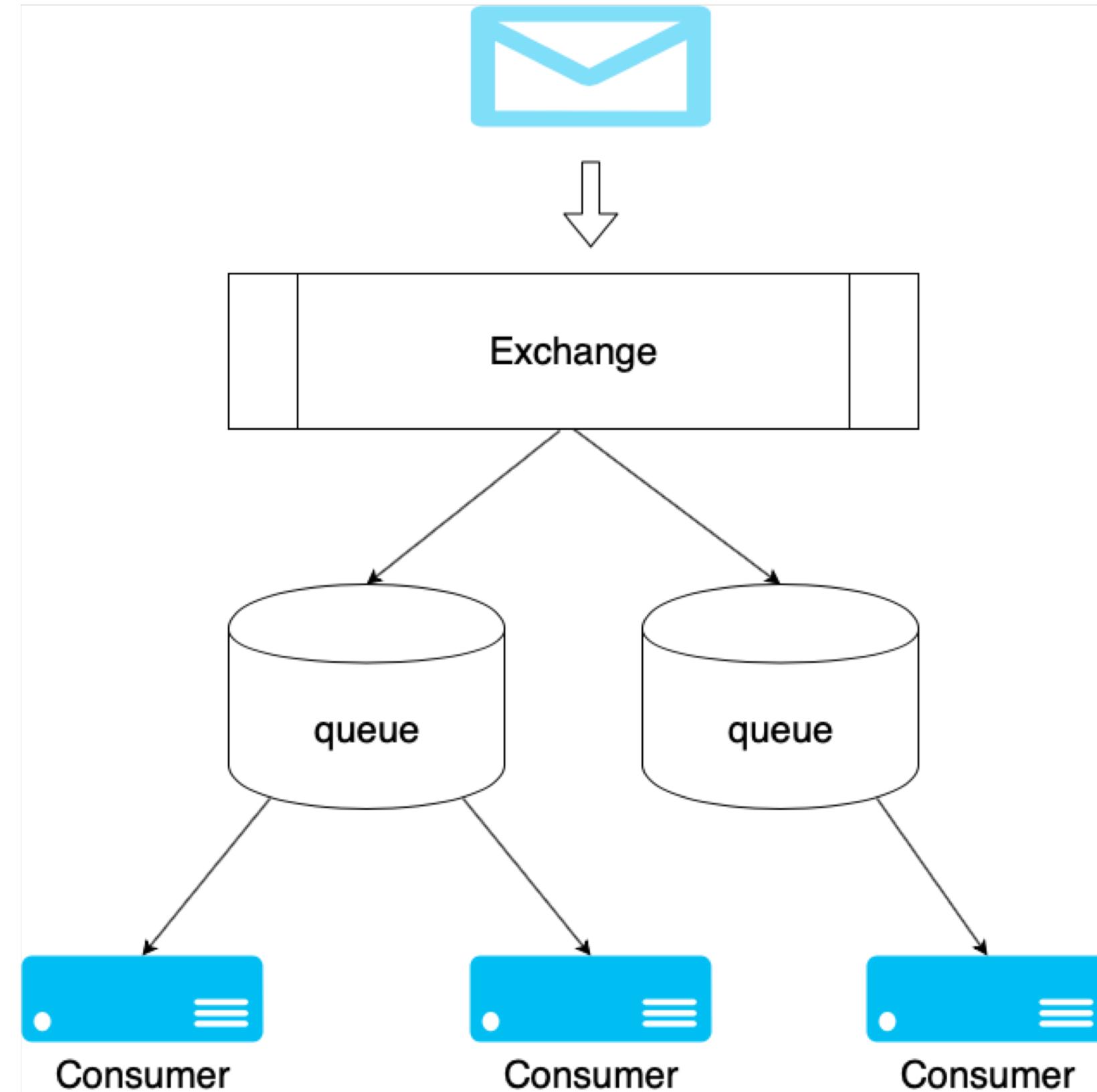
$1 * 10^6 - 7 \text{ GB/M}$

$3 * 10^6 - 5 \text{ GB/M}$

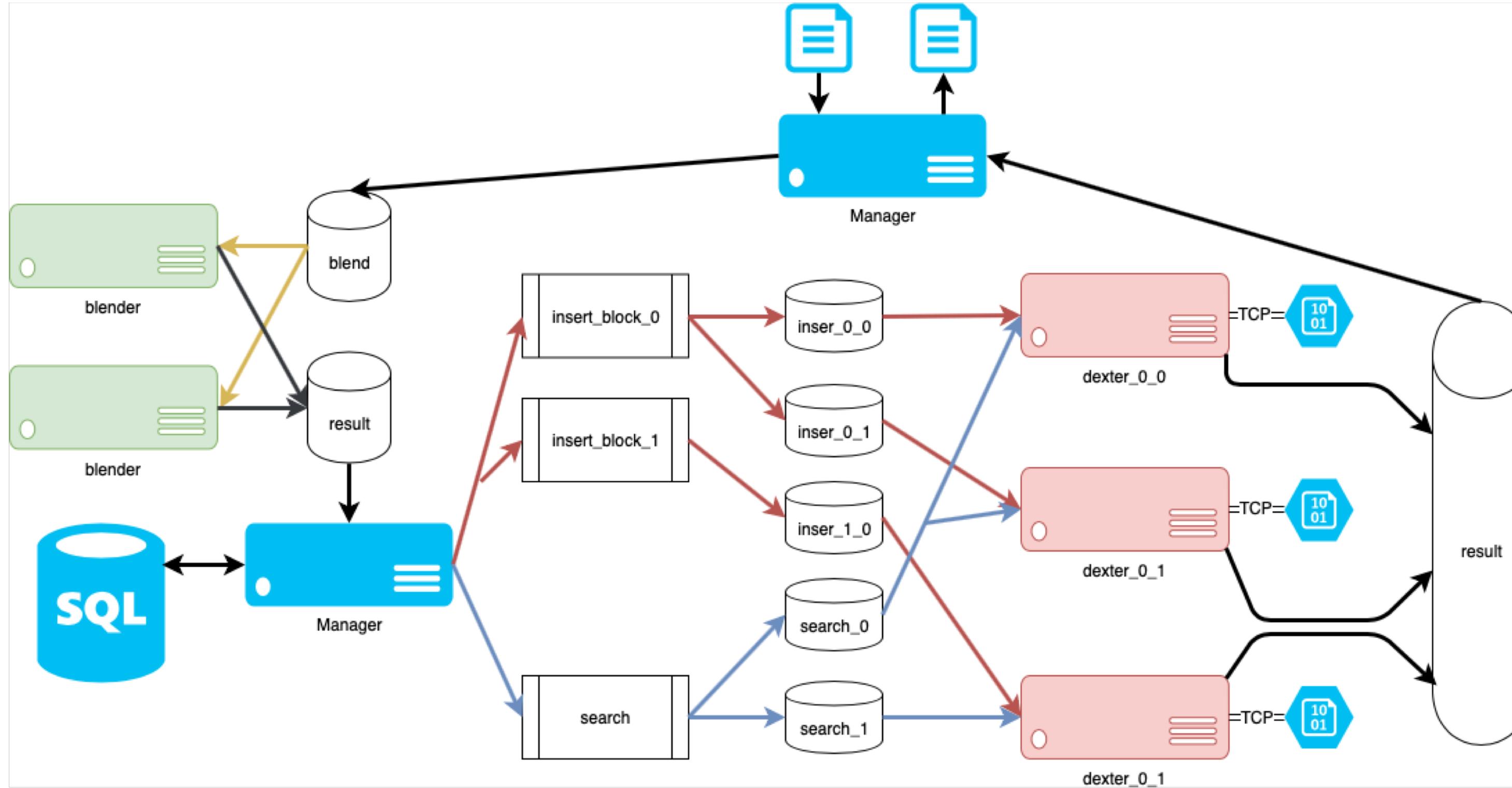
$30 * 10^6 - 4 \text{ GB/M}$



RabbitMQ



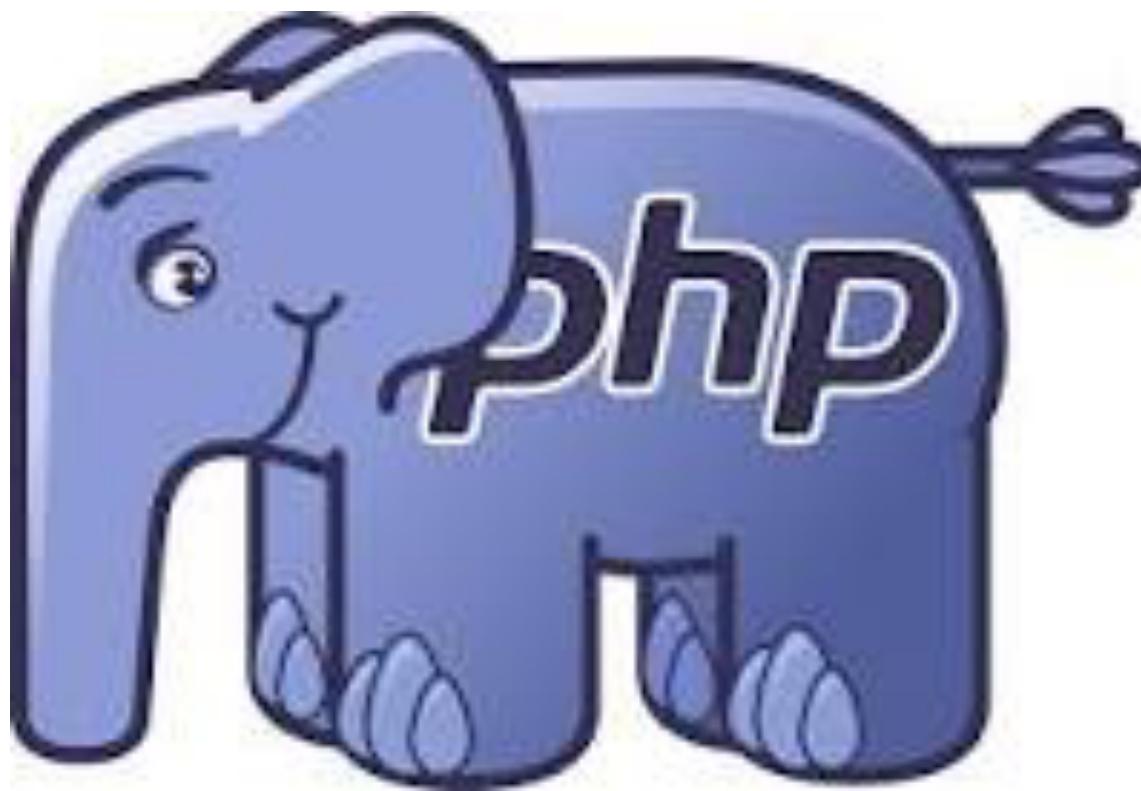
Конвейер поиска



Разработка и администрирование



PostgreSQL



SALTSTACK

Nagios®



Заключение



- Для rocket science нужно только желание
- Знания из ВУЗа лишними не бывают
- Pet проекты развивают
- Выступления помогают копнуть вглубь



Вопросы?

MAIL@AVALLAC.RU
<https://t.me/AVAIAC>

CTrie



Прототип:

- PHP
- Структуры а памяти
- Хранение в файле
- Один поток
- TCP

CTrie



Этап 1:

- C++
- Объекты кешируются в памяти
- Поиск по файлу
- Один поток
- TCP

CTrie



Этап 2:

- C++
- Объекты в памяти, ленивая загрузка
- Хранение в файле
- Многопоточность pthread, эксплюзивная блокировка для операции вставки
- TCP

CTrie



Этап 3:

- С
- Единый блок памяти
- Проблема фрагментации
- Хранение в файле
- Многопоточность pthread, вставка без блокировки
- TCP

Заголовок

Размер базы (64 бита)	Размер корня (64 бита)	Смещение корня (64 бита)
--------------------------	---------------------------	-----------------------------

Нода

Тип 1	Количество потомков (8 бит)				
	Размер префикса (8 бит)	Префикс (8*N бит)	Тип записи (8 бит)	Размер узла (14 бит)	Смещение узла (42 бита)
Тип 2	Размер префикса (8 бит)	Префикс (8*N бит)	Тип записи (8 бит)	Размер массива в байтах (4 байта)	Массив из uint32_t

CTrie



Этап 4:

- С
- Единый блок памяти, **низкоуровневые операции**
- Проблема фрагментации
- Хранение в файле
- Многопоточность pthread, вставка без блокировки
- TCP

Заголовок					
		Версия (64 бита)	Размер базы (64 бита)	Размер корня (64 бита)	Смещение корня (64 бита)
Нода					
Тип 0	Количество потомков (8 бит)				
Тип 1	Тип записи (3 бита)	Размер префикса (5 бит)	Префикс (8*N бит)	Смещение узла (40 бит)	0 - 1 Тб
Тип 2	Тип записи (3 бита)	Размер префикса (5 бит)	Префикс (8*N бит)	Смещение узла (40 бит)	1 - 2 Тб
Тип 3	Тип записи (3 бита)	Размер префикса (5 бит)	Префикс (8*N бит)	Размер массива (32 бита)	Массив из uint32_t (128+ бит)
Тип 4	Тип записи (3 бита)	Размер префикса (5 бит)	Префикс (8*N бит)	ID картинки uint32_t (32 бита)	
Тип 5	Тип записи (3 бита)	Размер префикса (5 бит)	Префикс (8*N бит)	ID картинки uint32_t (32 бита)	ID картинки uint32_t (32 бита)
			Префикс (8*N бит)	ID картинки uint32_t (32 бита)	ID картинки uint32_t (32 бита)



Этап 5:

- С
- Единый блок памяти + **мтар**, низкоуровневые операции
- Проблема фрагментации
- Хранение в файле
- Многопоточность pthread, вставка без блокировки
- TCP
- **AVL дерево в заголовке**