

Найти за полсекунды

PHP/C

RabbitMQ

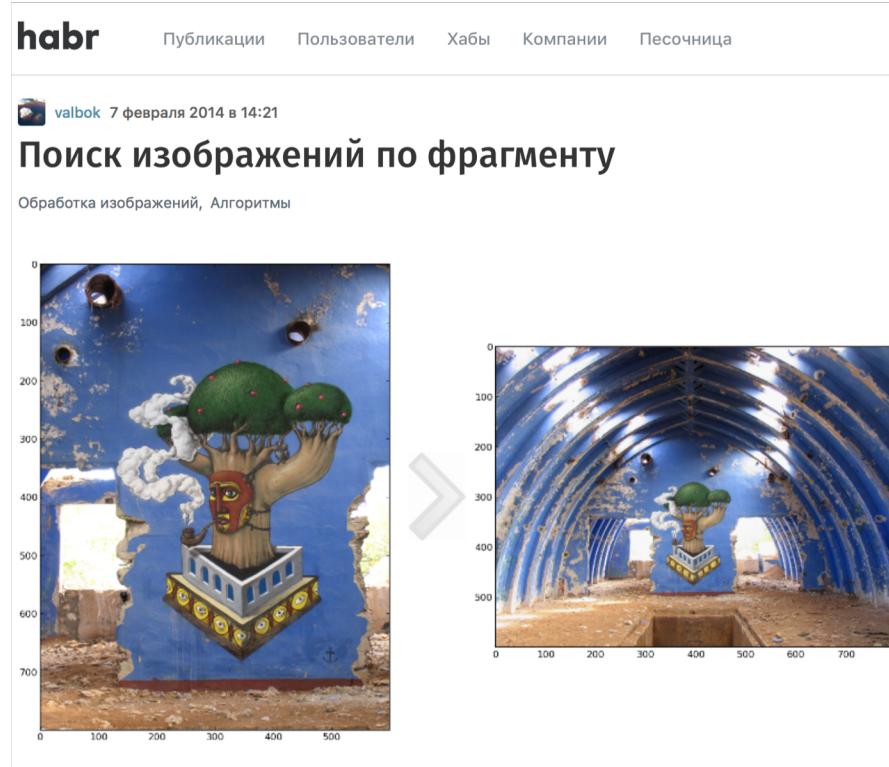
Backend

Алгоритмы

Петр Петренко

Senior software developer

skyeng



DevFest 2018, Novosibirsk, 23-25 November, 2018

Схожие изображения

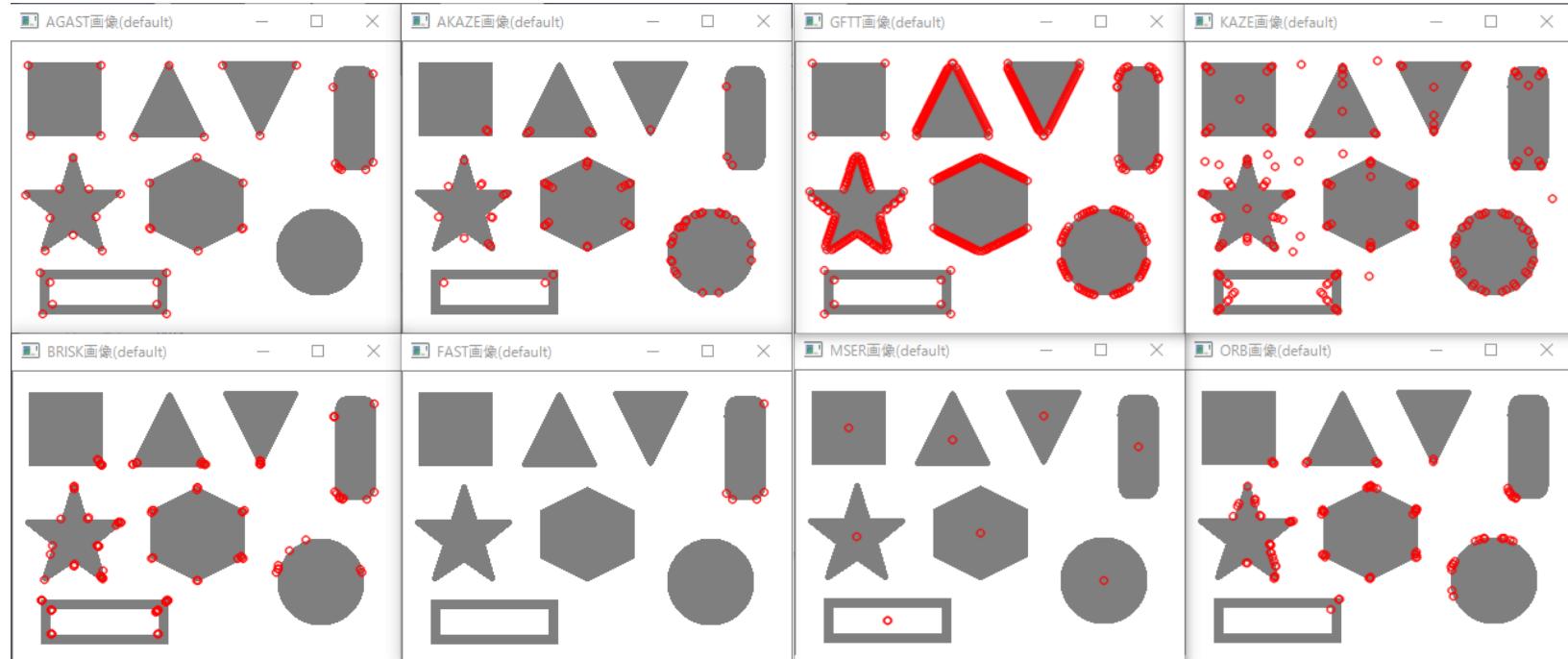
Миниатюры



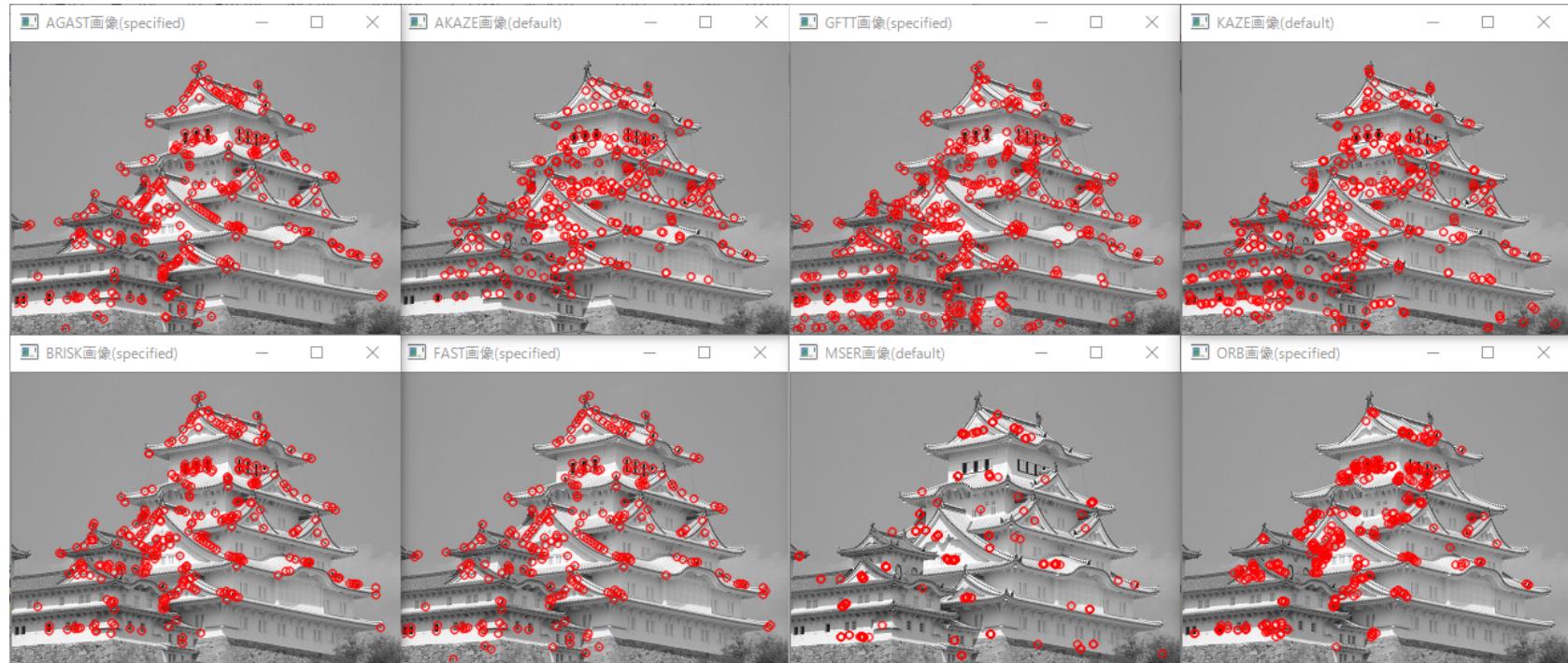
Полудубликаты



Ключевые точки



Ключевые точки



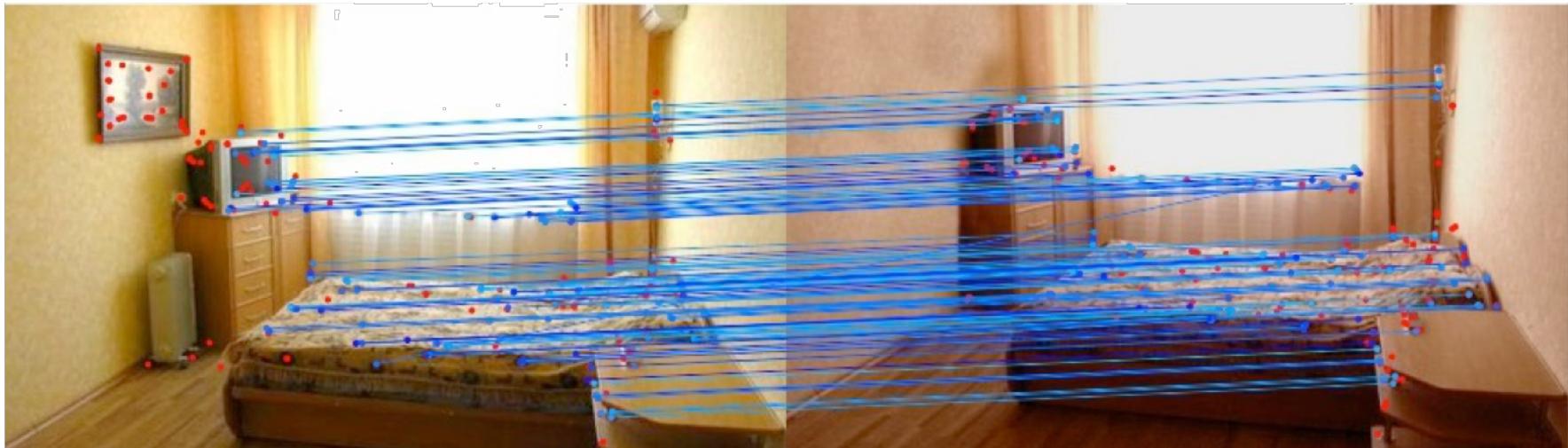
Алгоритмы поиска ключевых точек

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

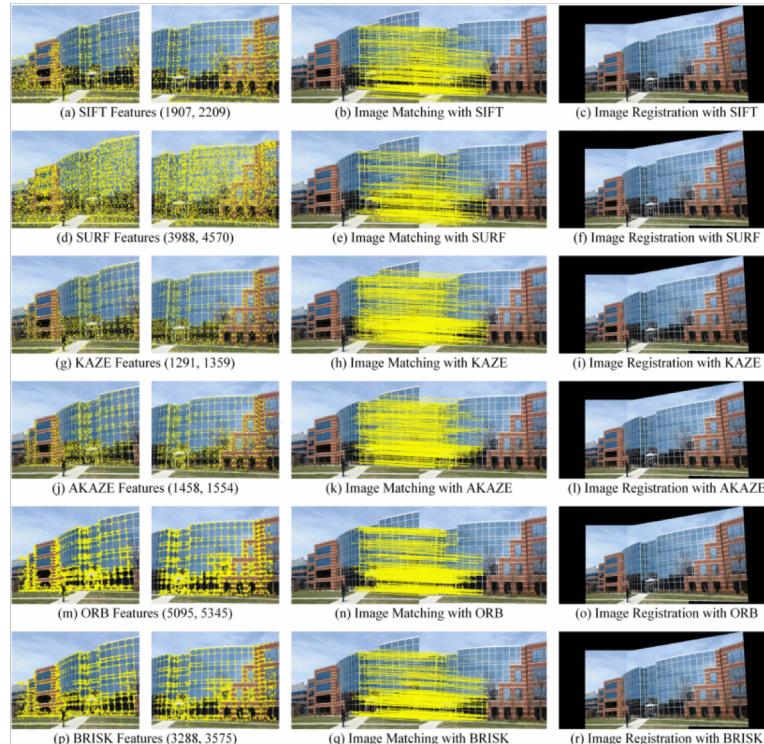
Покрытие



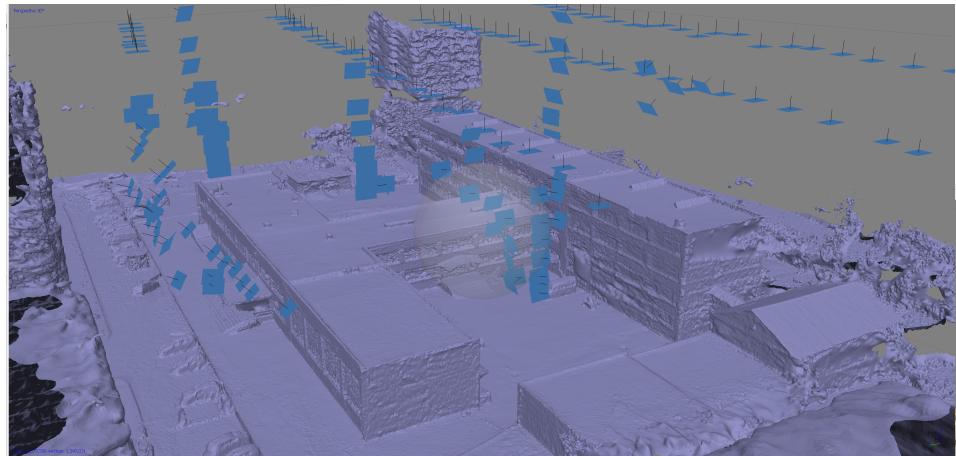
Сопоставление точек



Панорамы



3D модель по фото

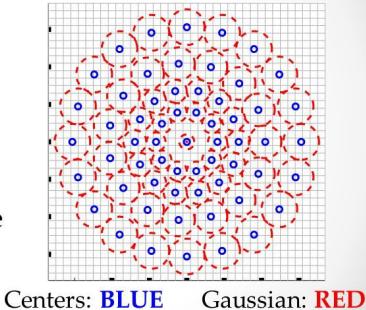


DevFest 2018, Novosibirsk, 23-25 November, 2018

Дескрипторы

BRISK: Descriptor

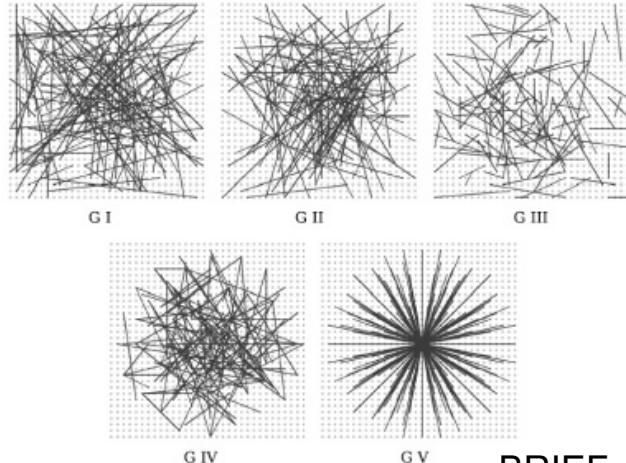
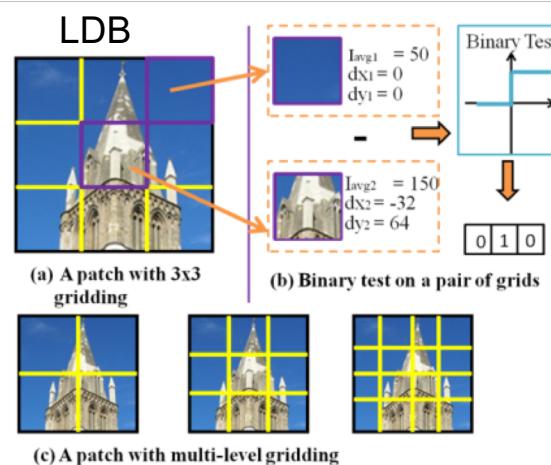
2D Gaussian around each feature

Robust to noise

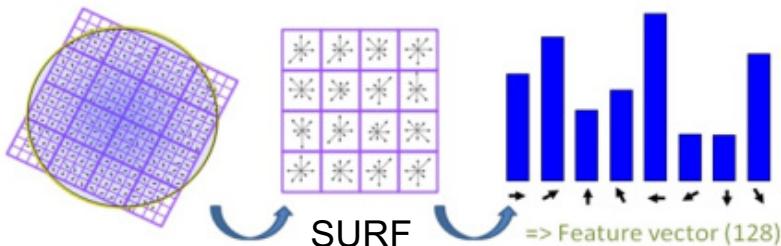
32

slide: J. Heinly

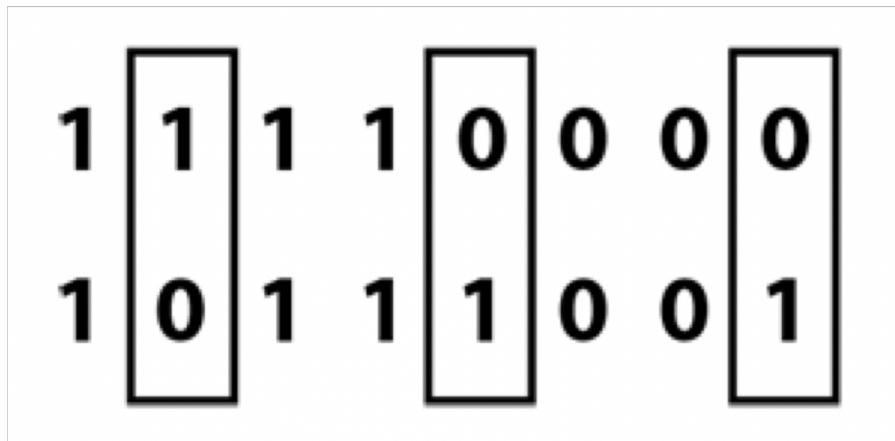


BRIEF

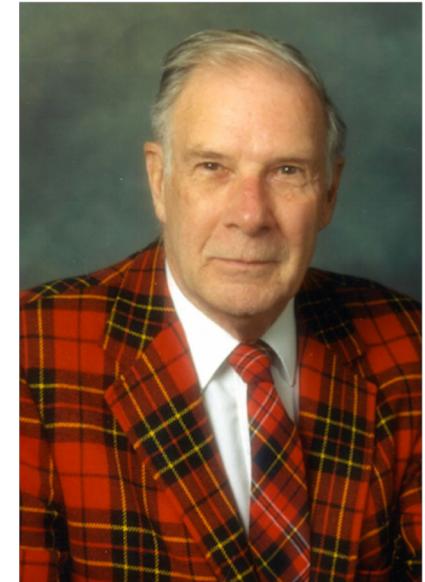
BRISK



Расстояние Хэмминга

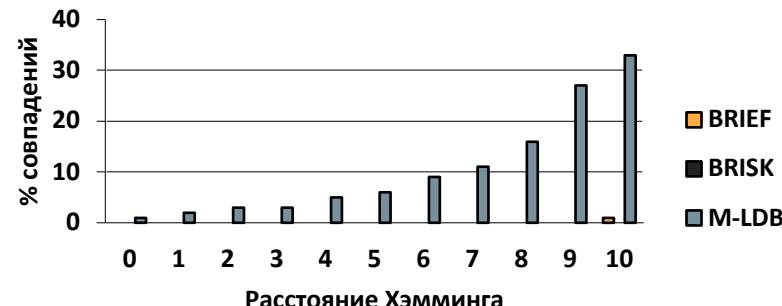
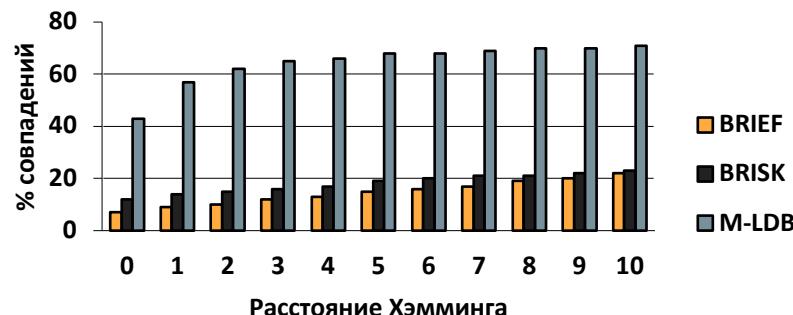
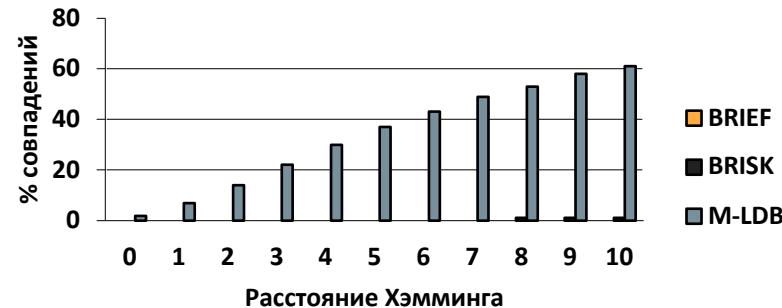
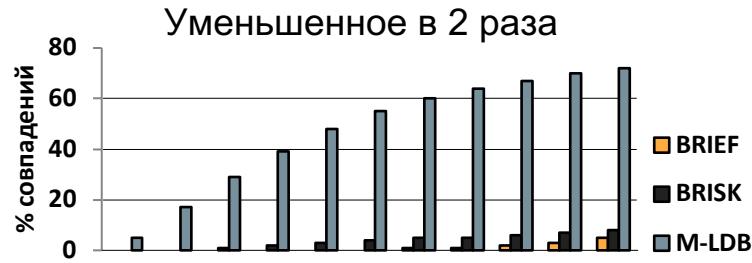


Число позиций, в которых соответствующие символы двух слов одинаковой длины различны



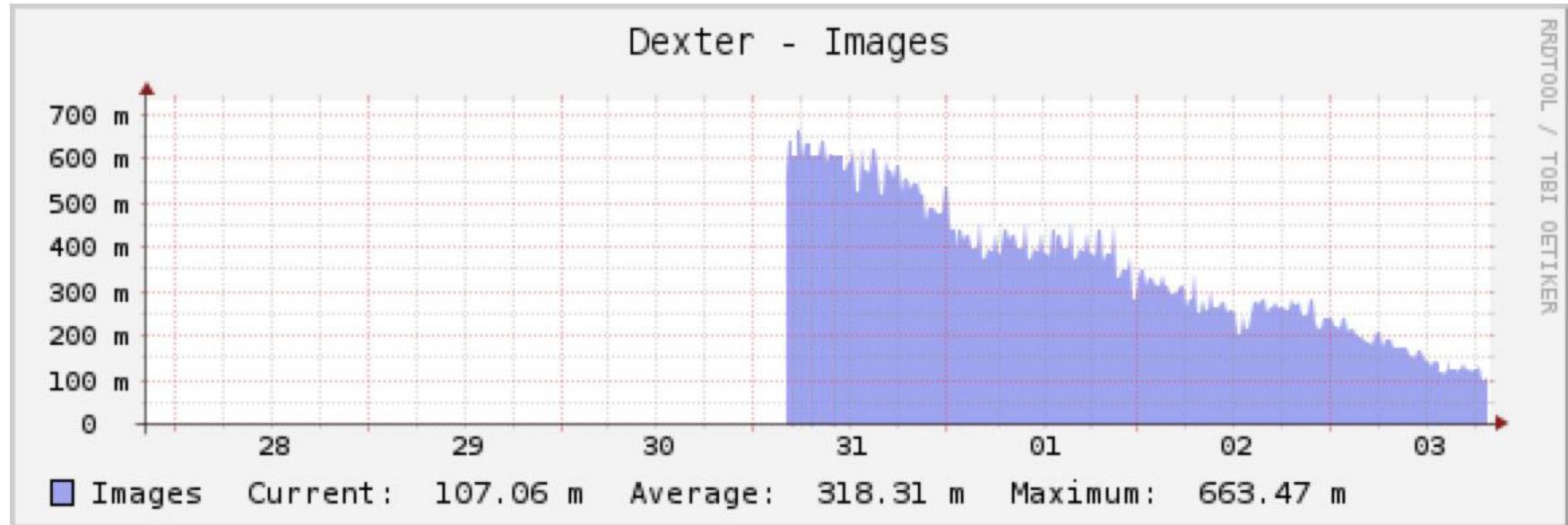
Richard
Уэсли
Хэмминг

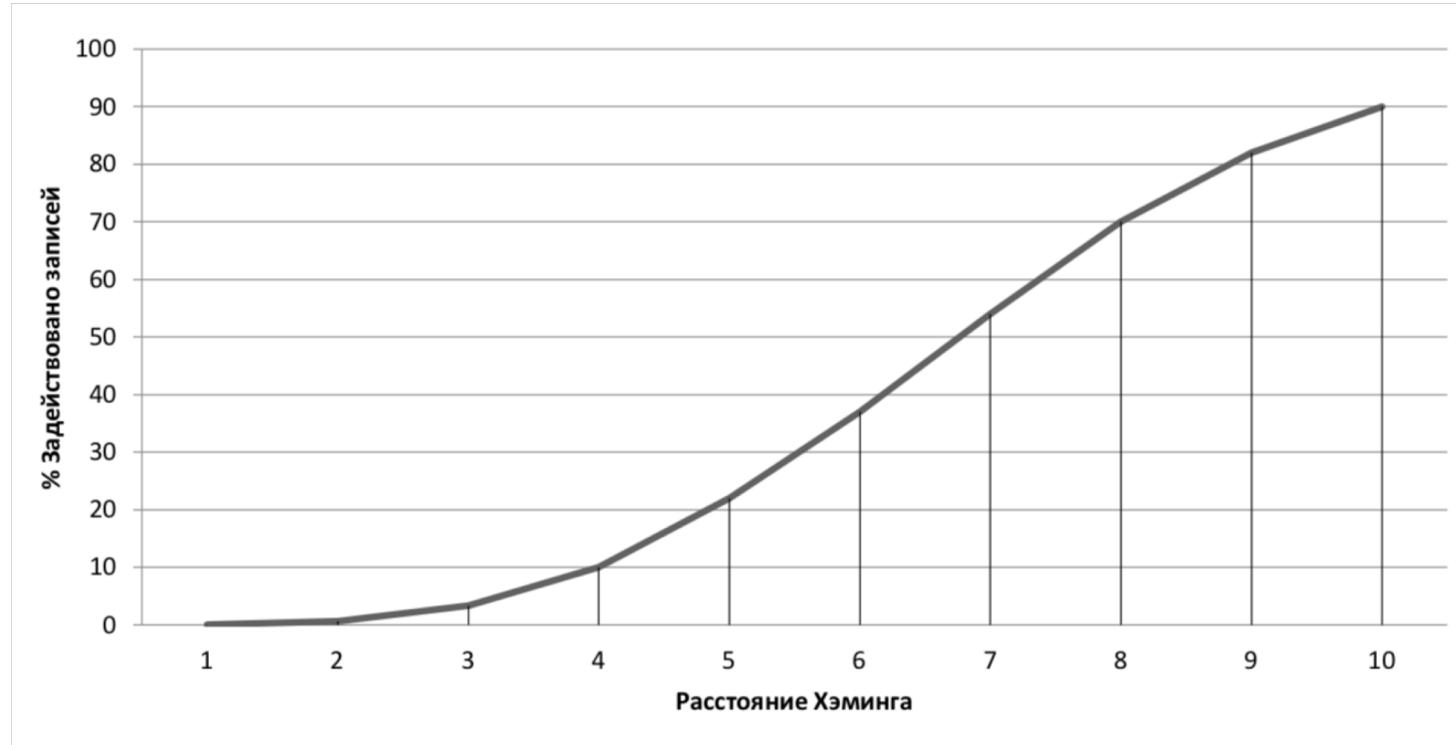
Сравнение дескрипторов



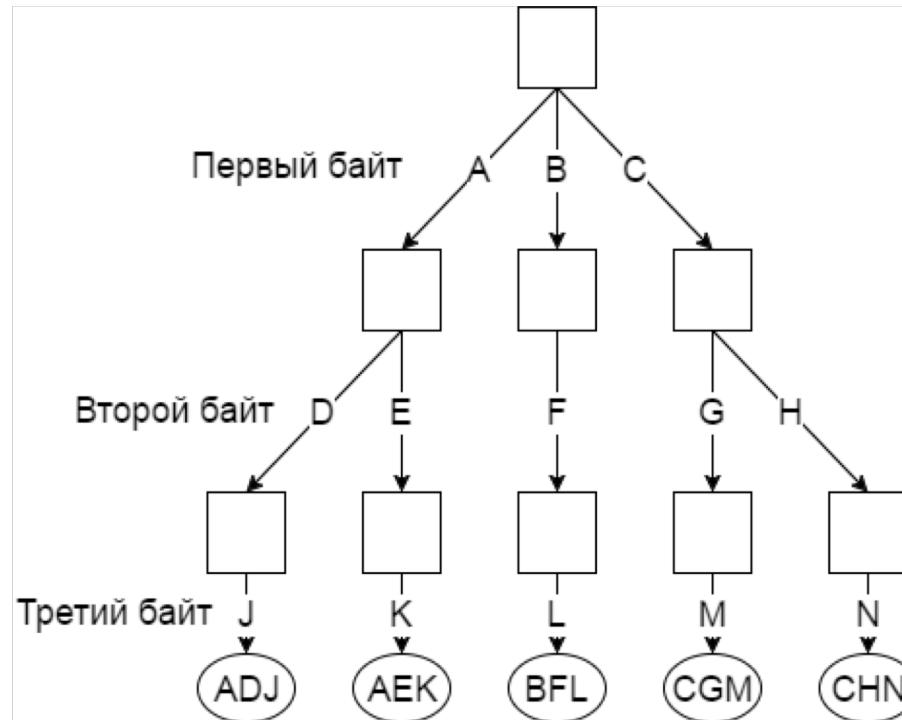
- Ключевых точек на одно изображение ~**350**
- Тестовая база данных: ~**350 миллионов точек**
- Объем база данных: ~**10GB**
- Линейный поиск одной ключевой точки в тестовой базе ~ **73.2 секунды** (без дисковых операций), всей картинки **более 7 часов.**

- Объем базы данных: ~85GB

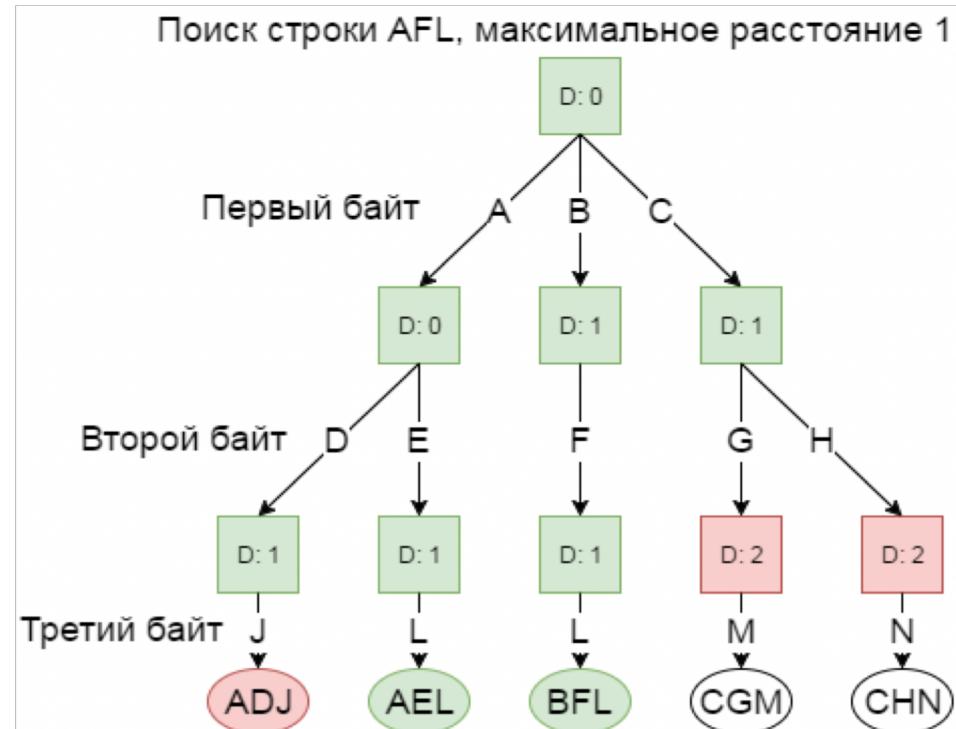




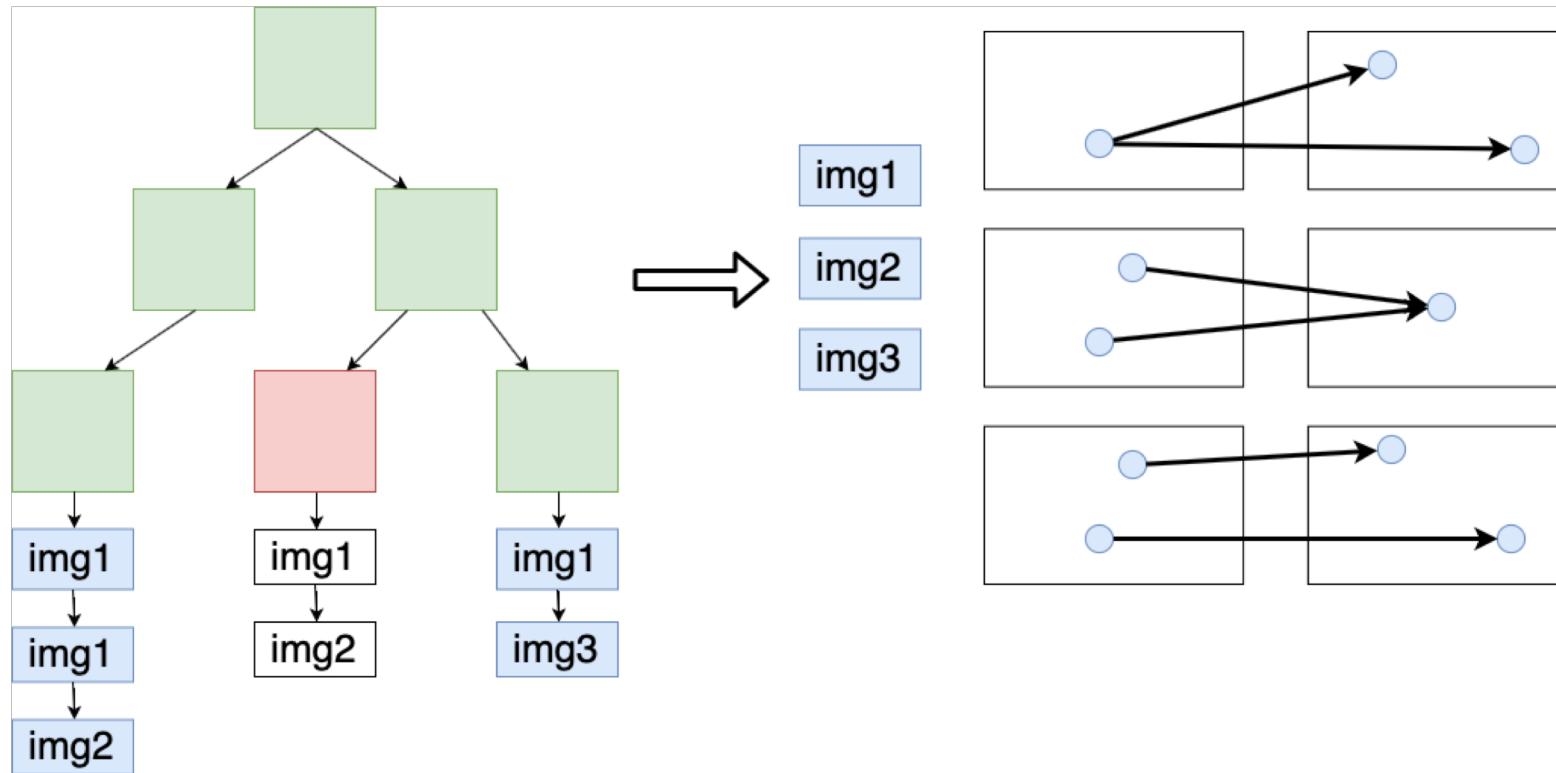
Предфиксное дерево



Предфиксное дерево



Поиск



Оптимизация

		Ключевых точек $10 * 10^9$ Узлов $3 * 10^9$			
1 байт Размер префикса	N байт (среднее 5.8) Префикс	1 байт Тип записи	4 байт Размер записи	8 байт Ссылка	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 $3 * 10^9$		32% 19.8 байт		Записей $10 * 10^9$ Узлов $3 * 10^9$		32% 11.8 байт	
1 байт (среднее 3 записи) Количество потомков	N байт (среднее 5.8) Префикс	1 байт Тип записи	4 байт Размер записи	8 байт Ссылка	3 бита Тип записи	5 бит Размер префикса	N байт (среднее 5.8) Префикс
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 байт) Префикс
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 байт) Префикс
19.8 * 3.3 + 1 = 66.34 байта База 185 GB		68% 19.8 байт		12.8 * 3.3 + 1 = 43.24 байта База 120 GB		3% 52.3 байт	

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				32% 11.8 байт				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				55% 9.8 байт				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	8% 14.8 байт				8% 12.8 байт				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				2% 15.8 байт				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% 39.9 байт				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													

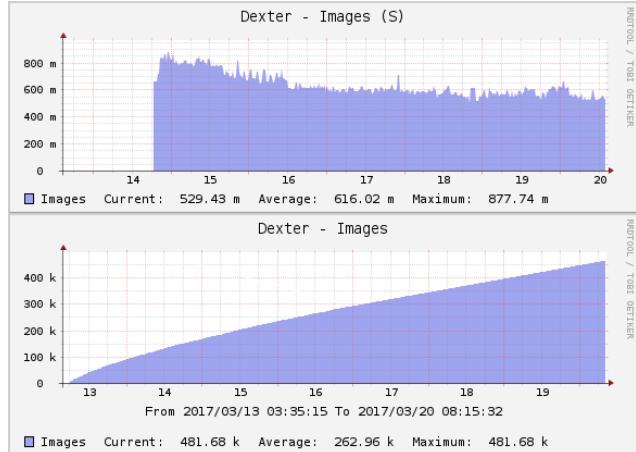
Префиксное дерево

- Объем база данных: ~7GB
- Поиск одной ключевой точки в тестовой базе примерно 3.5 миллисекунды, всей картинки примерно 1.3 секунды.

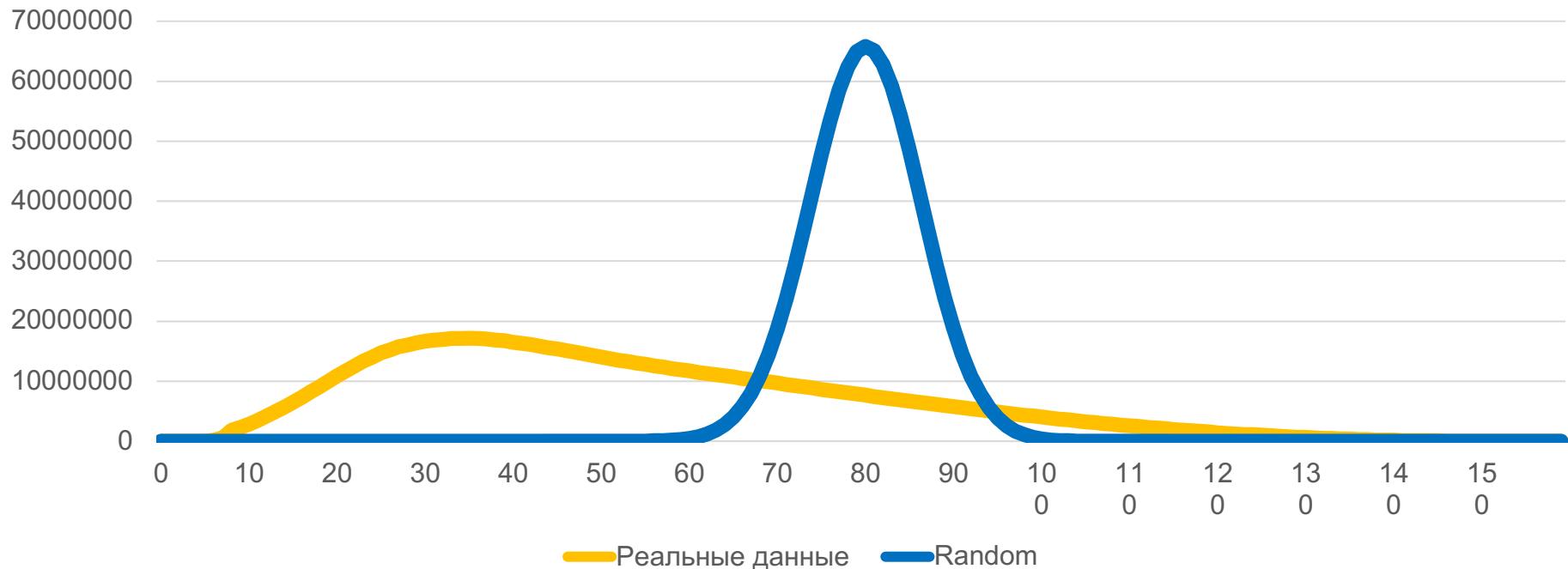
$1 * 10^6 - 7 \text{ GB/M}$

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

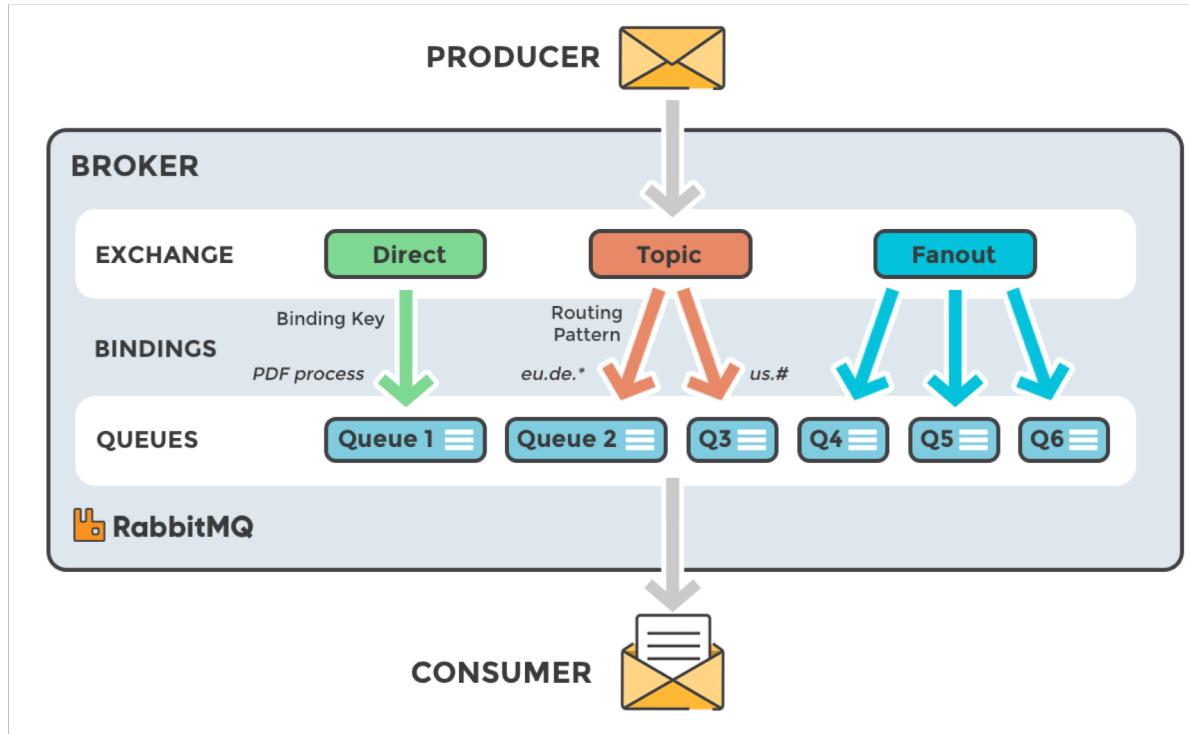
$30 * 10^6 - 4 \text{ GB/M}$



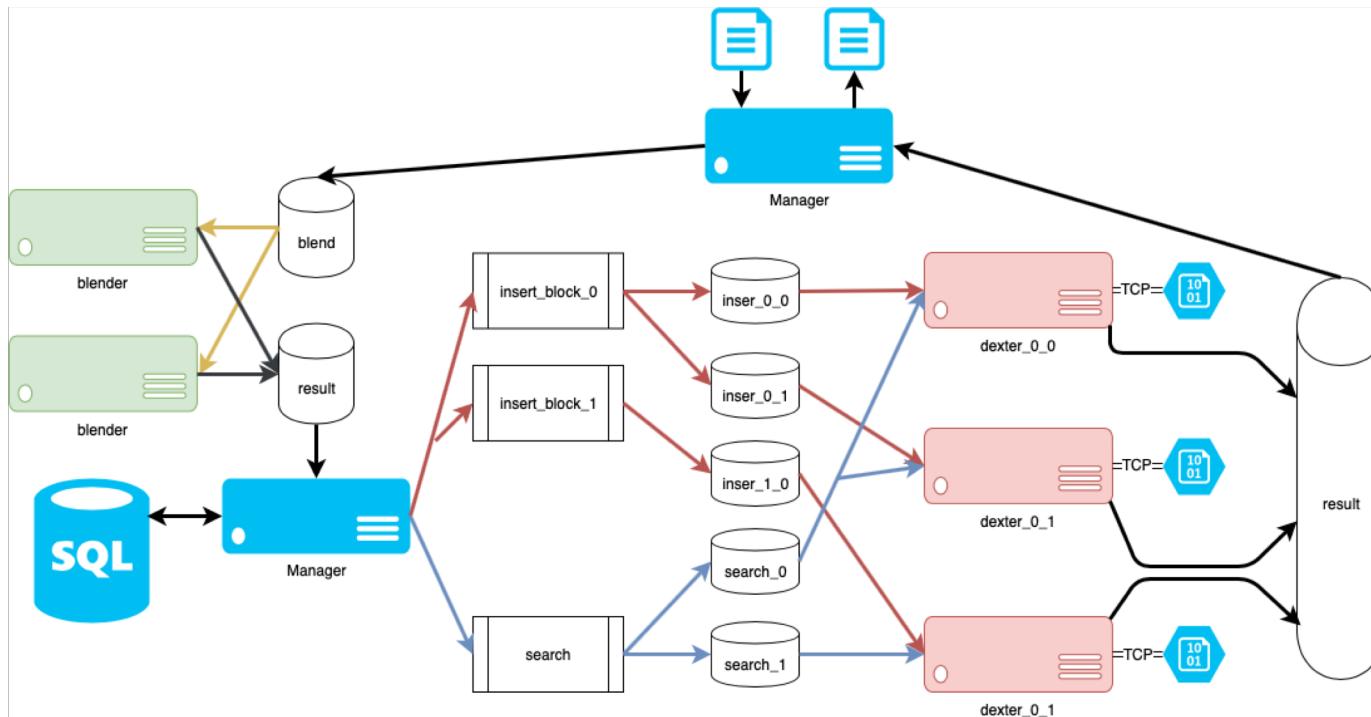
Неравномерное распределение



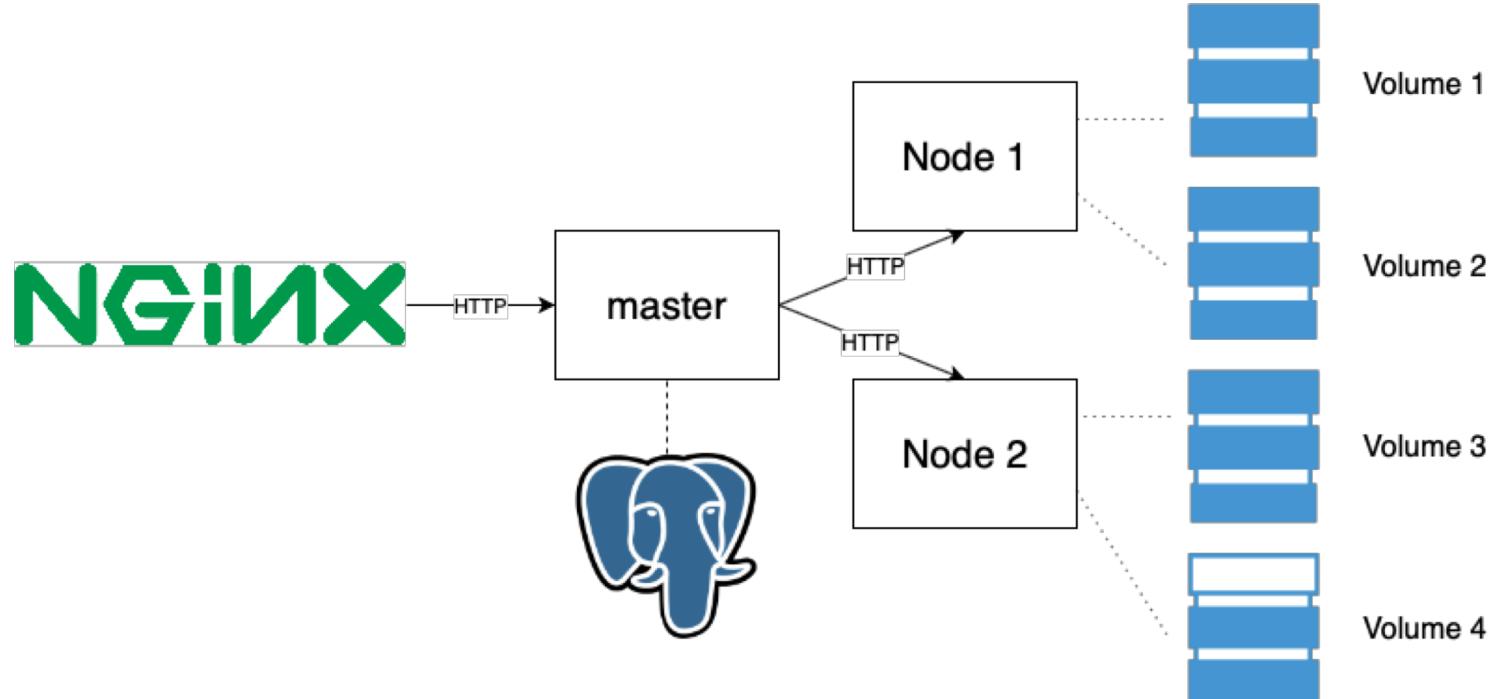
RabbitMQ



Конвейер поиска

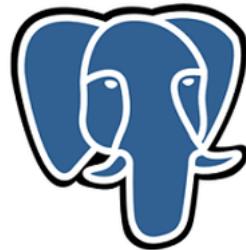


Хранение картинок

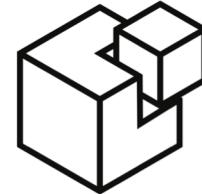
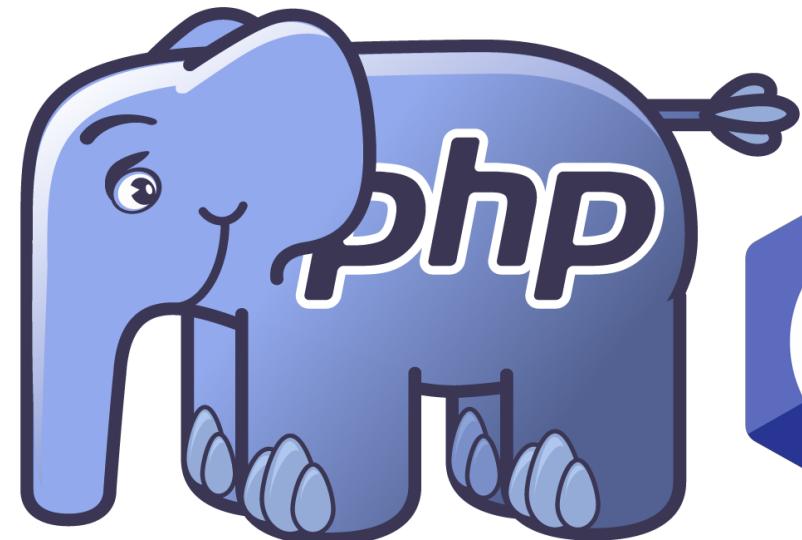


Разработка и администрирование

DevFest
Siberia 2018

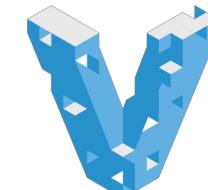


PostgreSQL



SALTSTACK

Nagios®



VAGRANT



CACTI

DevFest 2018, Novosibirsk, 23-25 November, 2018

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



Спасибо!

Петр Петренко

skyeng

Senior software developer

 @AVALLAC.RU

 @AVALLAC_NSK

 @AVALLAC

 mail@avallac.ru