Высокоскоростной алгоритм хэширования для системы хранения данных

Математико-механический факультет

Ершов Александр
группа 344
Научный руководитель:
Разработчик RAIDIX Маров А. В.

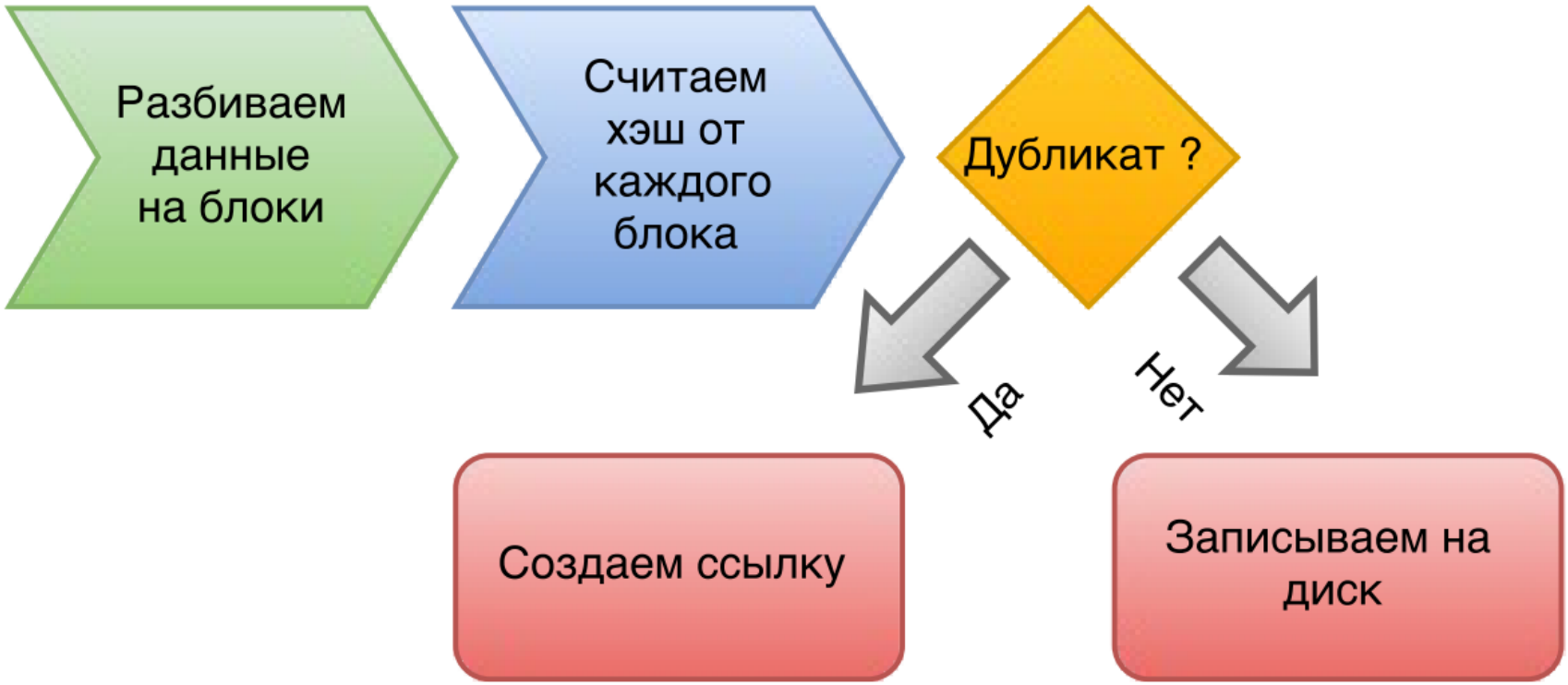
СПбГУ

29 сентября 2015 г.

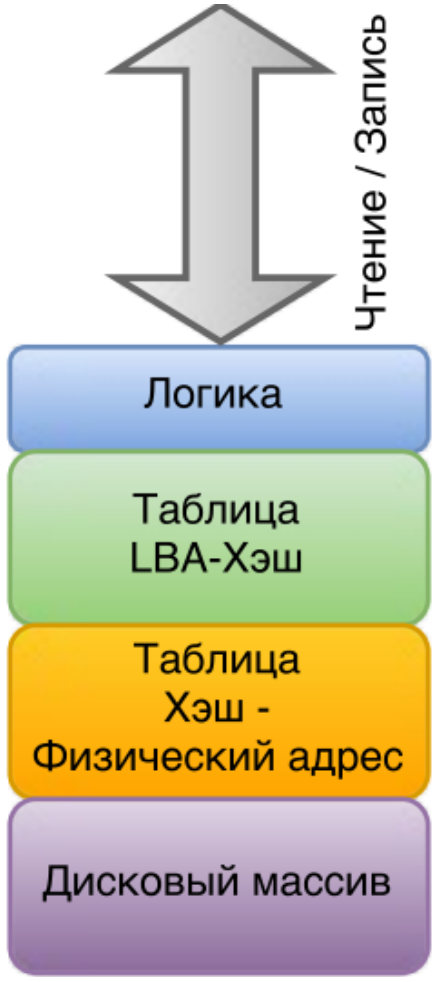
Дедупликация

- Процесс обработки данных, позволяющий избавится от хранения дубликатов
- File-level / Block-level
- Source / Target
- Inline / Post-process

Дедупликация : принцип работы



Архитектура системы



Постановка задачи

- Исследовать существующие хэш функции
- Реализовать свой вариант с использованием векторных вычислений
- Протестировать результаты

Обзор существующих хэш функций [1]

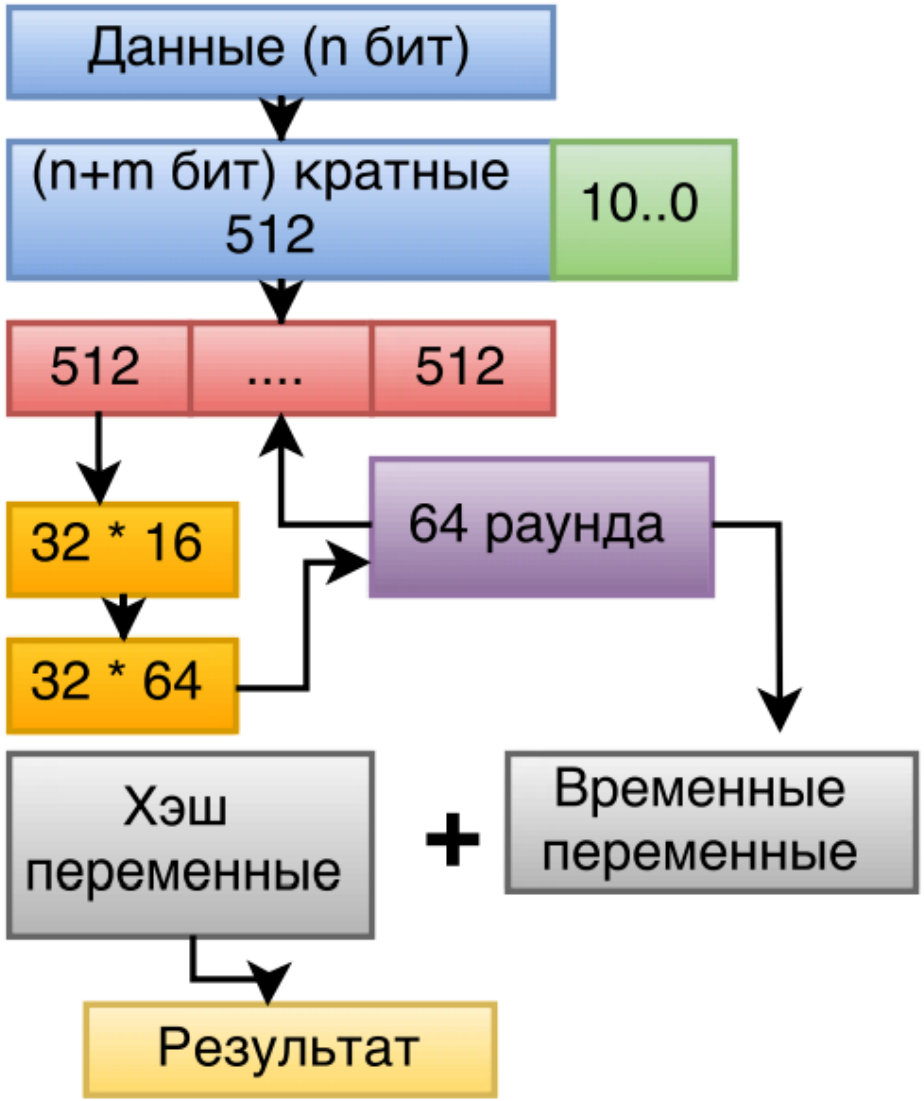
Функции	Размер блока	Размер слова	Кол-во раундов	Вывод	Год	Найдены коллизии
MD4	512	32	48	128	1990	+
MD5	512	32	64	128	1992	+
SHA-1	512	32	80	160	1995	+
SHA-256	512	32	64	256	2002	-
SHA-384	1024	64	80	384	2002	-
SHA-512	1024	64	80	512	2002	-

[1] http://research.microsoft.com/pubs/64588/hash_survey.pdf

Инструменты

- Язык C
- SSE, AVX
- Intel intrinsic instructions

Реализация : Алгоритм SHA-256[2]

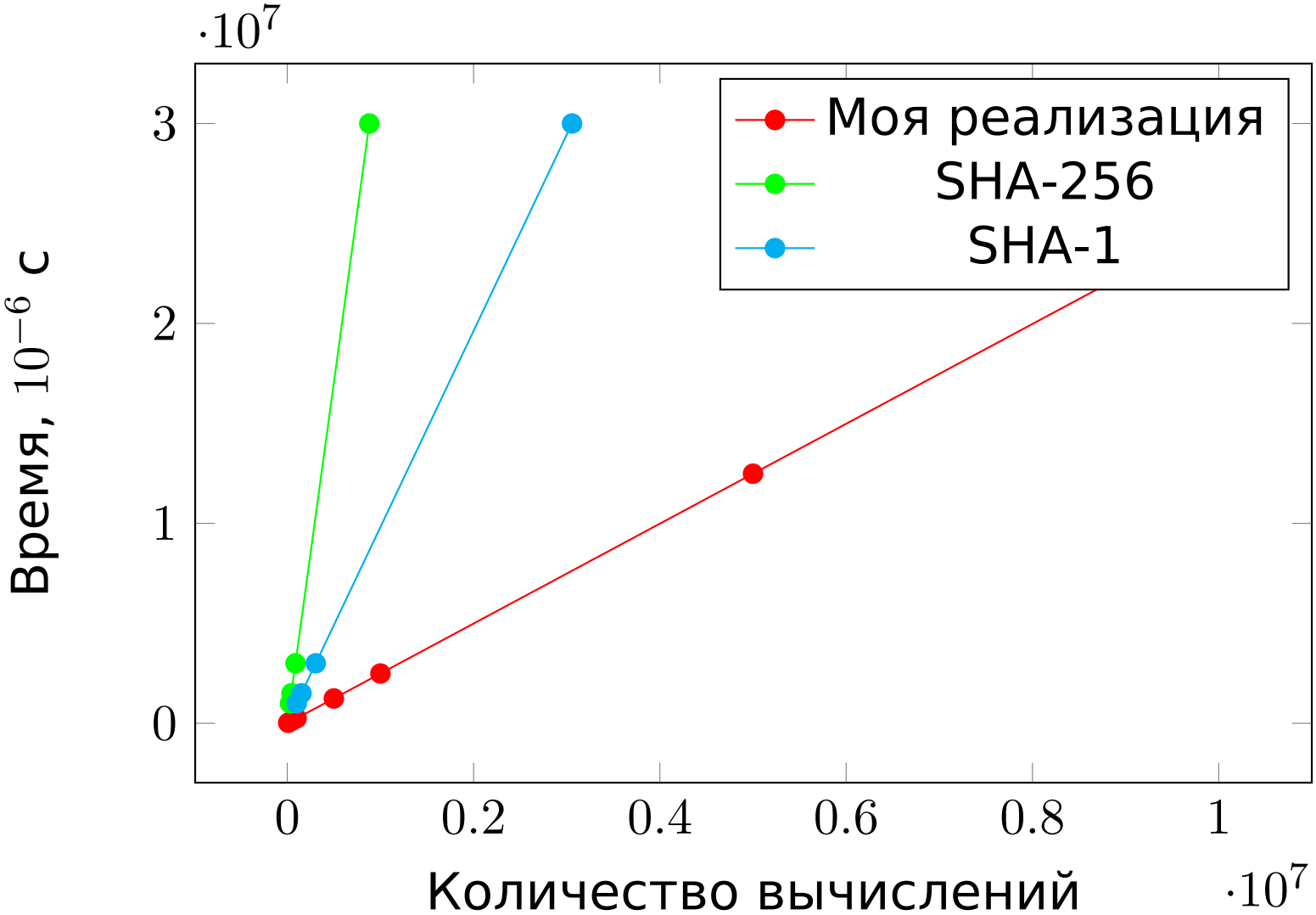


[2] https://tools.ietf.org/html/rfc4634

Реализация : Изменения

Hash	Размер блока	Размер слова (бит)	Количество раундов	Количество дополни - тельных блоков	Вывод (бит)
Sha-256	512 бит	32	64	48	256
Моя реализация	4 кб	128	256	0	256

Тестирование результатов



Коллизии найдены не были [3]

[3] Конфигурация : Ubuntu, AMD A10 2,3 Ггц, 6 Гб ОЗУ. Отклонения в измерениях 15%

Заключение

- Сделан обзор существующих хэш функций
- Реализован свой вариант с использованием векторных вычислений
- Протестированы результаты