Автоматическая настройка параметров тиринга, зависящая от входящей нагрузки, в системе хранения данных

Смирнов Михаил Александрович

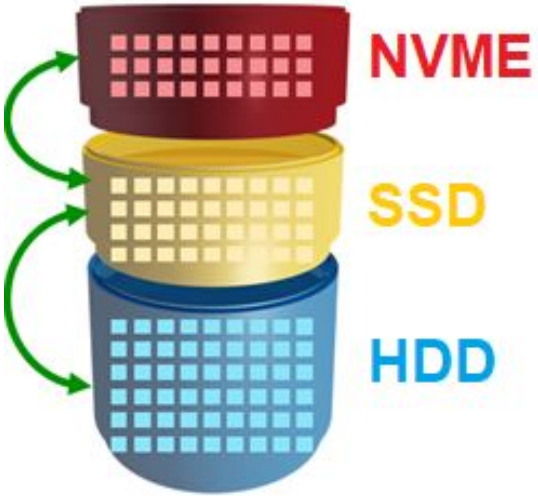
Научный руководитель:
д. ф.-м. н., проф. кафедры системного программирования СПбГУ Терехов А. Н.

Рецензент:
к. т. н., руководитель исследовательской лаборатории RAIDIX Лазарева С. В.

RAIDIX

- RAIDIX - это software Defined Storage (SDS) программно-определяемая система хранения данных
- Исследовательская лаборатория RAIDIX создаёт прототипы, которые передаются в отдел производства для внедрения в продукт

Организация многоуровневого хранения (Tiering)



1. В режиме реального времени производится сбор статистики операций чтения и записи
2. Периодически фоновый процесс анализирует статистику обращений и принимает решение о перераспределении данных
3. Данные перераспределяются по уровням в соответствии с заданной стратегией

Постановка задачи

Реализация модуля тира в системе RAIDIX, управляющего распределением данных на устройствах хранения.

1) Анализ существующих алгоритмов организации хранения, выбор оптимального и его улучшение

2) Реализация поддержки двух и трёх уровней тира, в зависимости от требований пользователя

3) Функциональное тестирование и измерение производительности

Проблема хранения метаданных в RAIDIX

- Метаданные
  1. Определяют распределение данных по уровням хранения
  2. Хранят статистику обращений
  3. Другая служебная информация
- Рост объёмов хранения в RAIDIX влечёт рост объёма метаданных. Выпадение метаданных из RAM наносит существенный урон производительности.
- Задача - исследовать возможности уменьшения объёма метаданных

Алгоритмы идентификации

	Фиксированный объём памяти для хранения	Экономие памяти	Ошибки распределения
Window-based Direct Address Counting	+	-	-
Multiple Hash Function	-	+	+
HotDataTrap	-	+	+

Существующие решения

IBM Easy Tier

- Период 24 часа
- Миграция в зависимости от нагрузки
- Размер блока 1 ГБ
- Поддерживает до трёх уровней

EMC FAST VP

- Периодичность и продолжительность миграции настраивается
- Размер блока 256 МБ
- Трёхуровневый тир

Реализация прототипа

Язык: C

Среда: Ядро Linux 3.10.0-327 x86_64

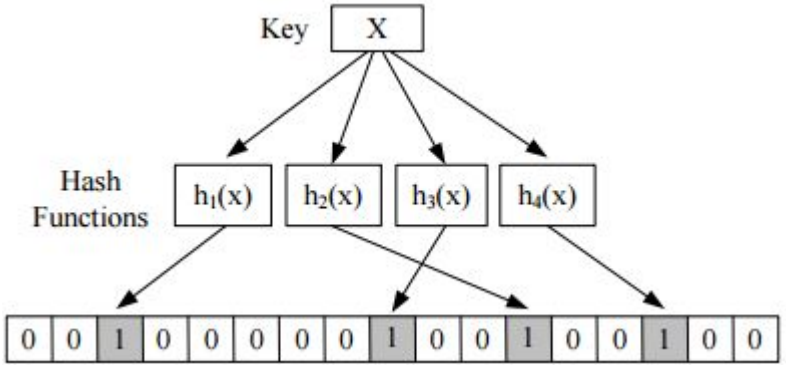
Функциональность:

1. Собственное логирование
2. Анализ статистики
3. Перенос данных по уровням хранения

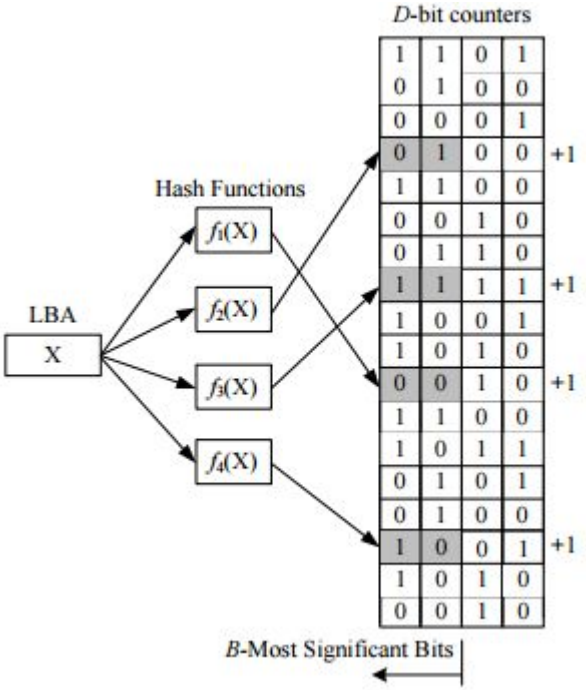
Особенности:

1. Возможен выбор двух или трёх уровней системы хранения
2. Возможно тестирование в промышленной системе RAIDIX

Реализация алгоритма идентификации



Bloom Filter



Multiple Hash Function

Необходимо подобрать:

1. Количество хэш-функций
2. Размер Bloom Filter
3. Размер элементов Bloom Filter
4. Размер ячейки памяти

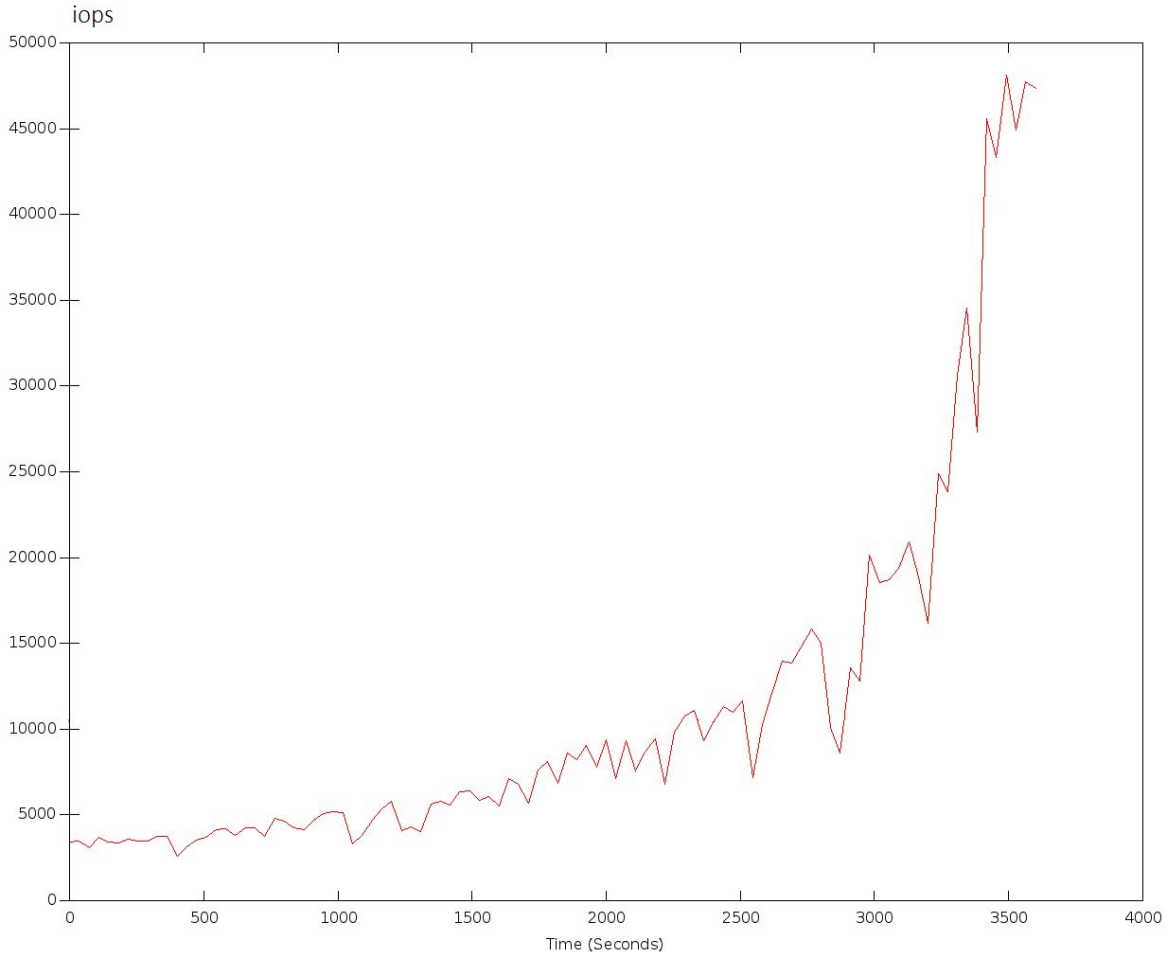
Схема экспериментов

- Моделируется нагрузка и запускается тестирование производительности RAIDIX
- Запускается механизм распределения данных
- Через 1 час или более тестирование приостанавливается и снимаются показатели производительности

Характеристики производительности:

1. Количество операций ввода/вывода в секунду (IOPS - input/output operations per second)
2. Объём перенесённых данных
3. Другие параметры

Тестирование корректности алгоритма



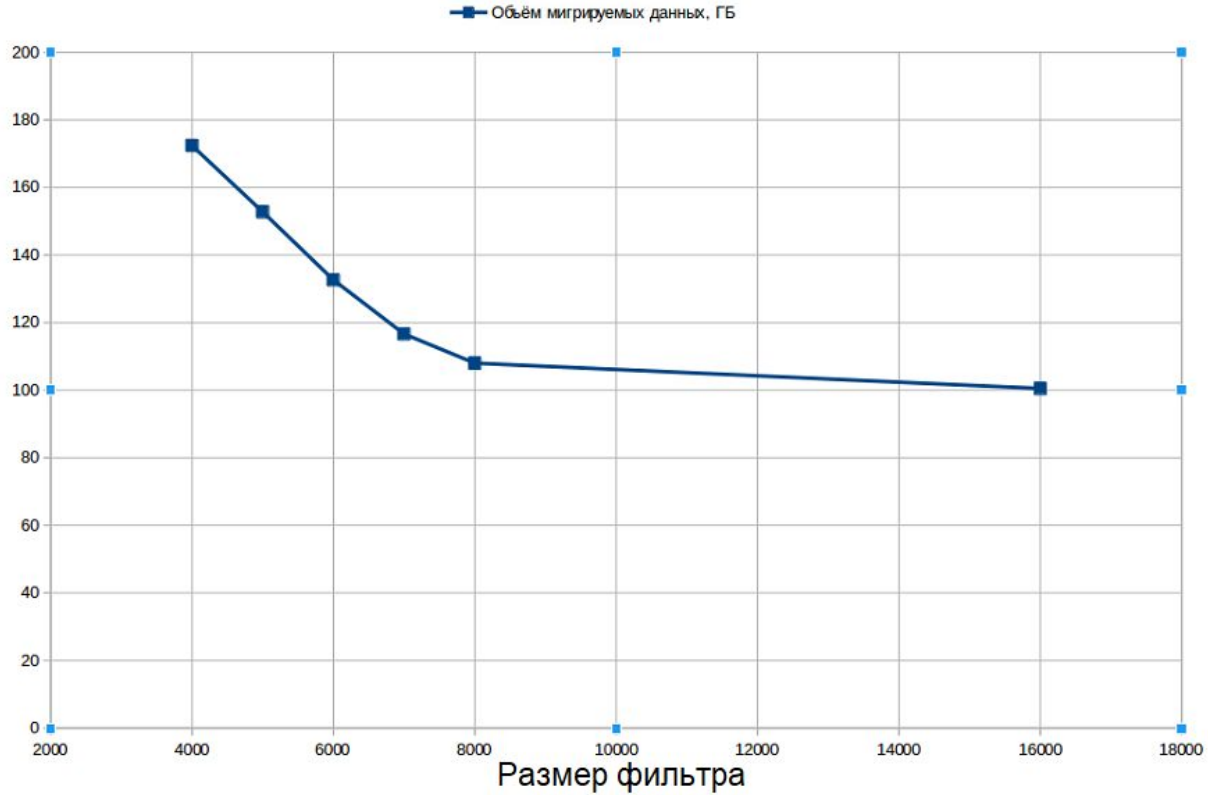
Параметры:

- Нагружается область 100 ГБ
- Миграция раз в минуту
- Длительность - 1 час

Вывод:

- Производительность выросла
- Все данные перенесены, так как под конец работы нагрузки на hdd-диски не было, все запросы отправлялись на ssd-диски.

Тестирование размера Bloom Filter

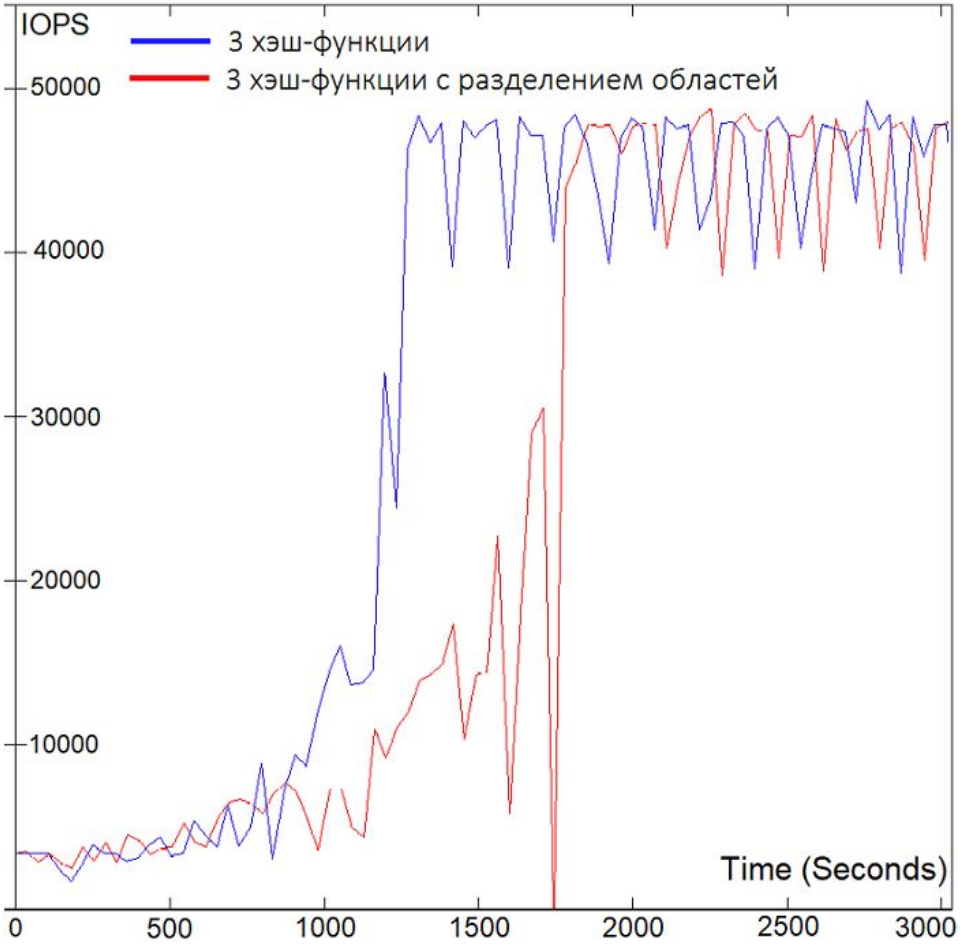


Параметры:

- Нагружается область 100 ГБ
- Миграция раз в минуту
- Длительность - 1 час
- Количество ячеек данных: 8136
- Меняется размер Bloom Filter от 4000 до 16000.

Вывод: оптимальным размером будем считать 7 тысяч.

Тестирование количества хэш-функций

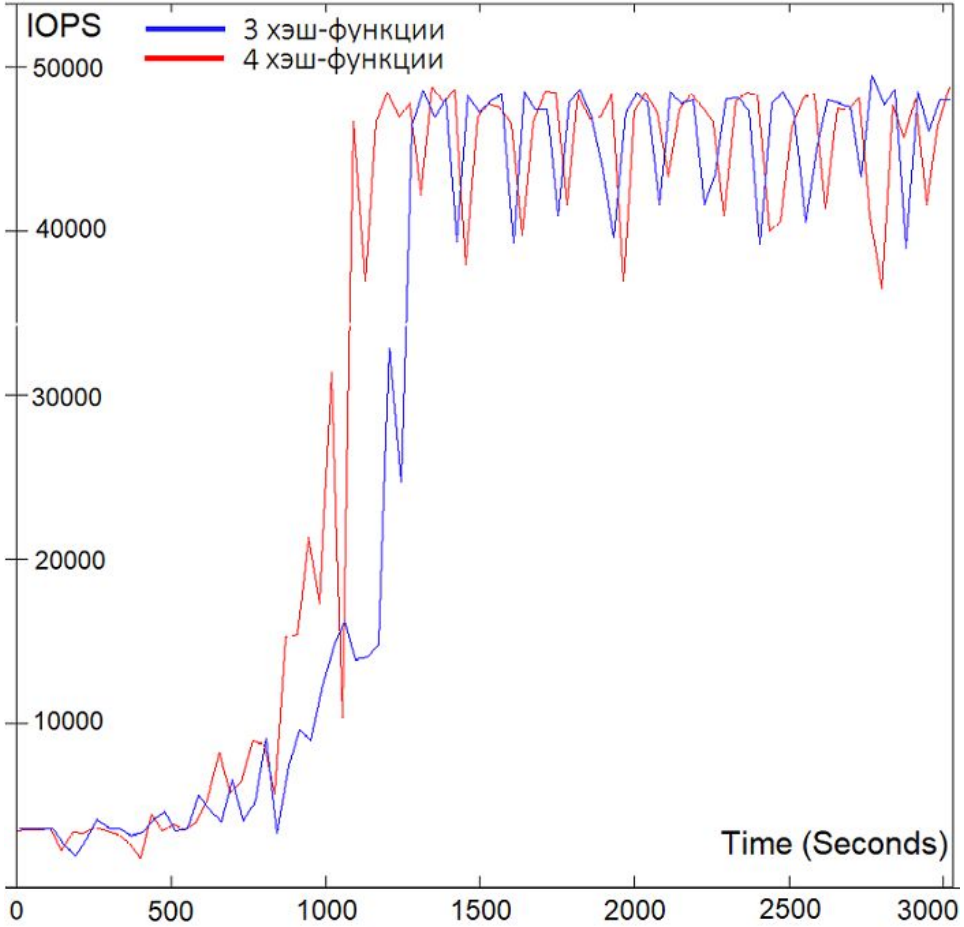


Параметры:

- Нагружается область 100 ГБ
- Миграция раз в минуту
- Длительность - 1 час

Вывод: разделение областей отрицательно сказалось на производительности.

Тестирование количества хэш-функций



Параметры:

- Нагружается область 100 ГБ
- Миграция раз в минуту
- Длительность - 1 час

Вывод: у алгоритма с четырьмя хэш-функциями рост производительности происходит быстрее.

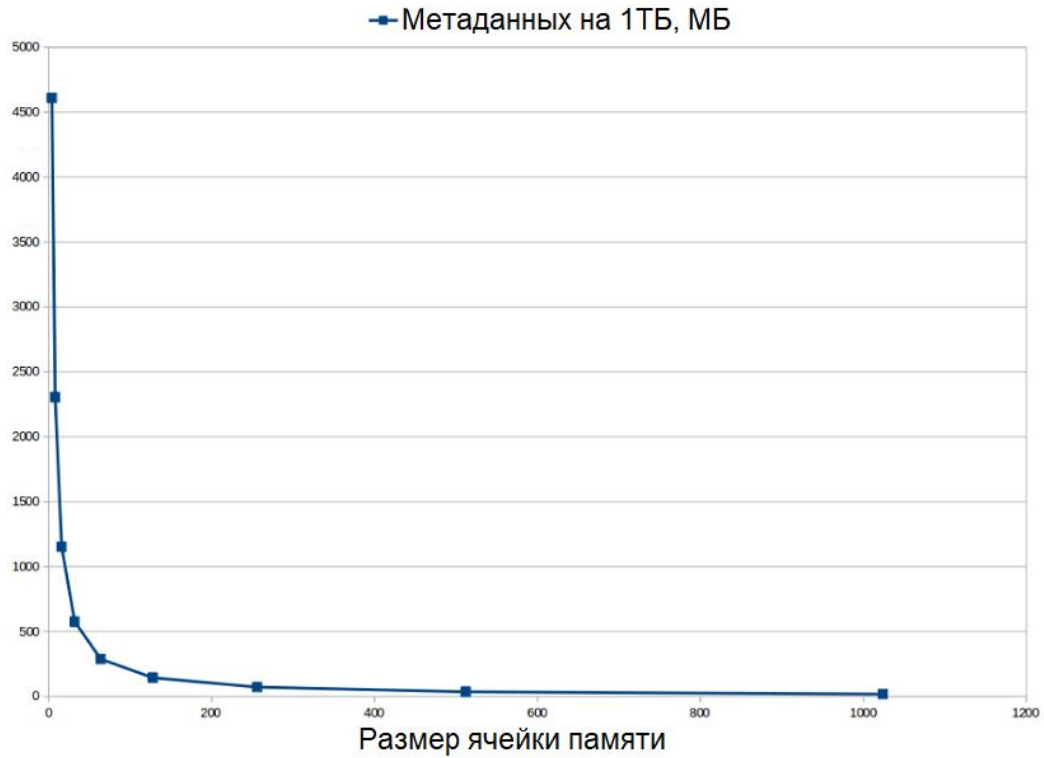
Сравнение с тиром без миграции

Параметр сравнения	С миграцией	Без миграции	С миграцией лучше?
Средняя скорость потока по всему тиру, iops	2300	1800	+
Средняя скорость первого потока по области, iops	34000	1800	+
Средняя скорость второго и последующих потоков по области, iops	15000	1800	+
Объём мигрируемых данных, ГБ	360	0	-

Параметры:

- Нагружаемые области - 100 ГБ
- Миграция раз в час
- Длительность - 36 часов

Метаданные



Для ячейки размером 4 МБ необходимо 4,5 ГБ метаданных на 1 ТБ данных.

Для ячейки размером 1 ГБ необходимо 18 МБ метаданных на 1 ТБ данных.

Основные результаты работы

1) Проведён анализ алгоритмов распределения данных и выбран алгоритм наилучшим образом удовлетворяющий потребностям СХД RAIDIX
2) Реализован прототип тира, организующий миграцию данных между уровнями хранения
3) Проведено тестирование производительности с использованием синтетической нагрузки
4) На основании тестирования произведён выбор ключевых параметров алгоритма
5) Прототип и результаты исследования переданы в отдел производства
6) Результаты представлены на конференции “СПИСОК”