Яндекс

Архитектура бесконечного хранилища для пользовательского контента

Артём Соколов

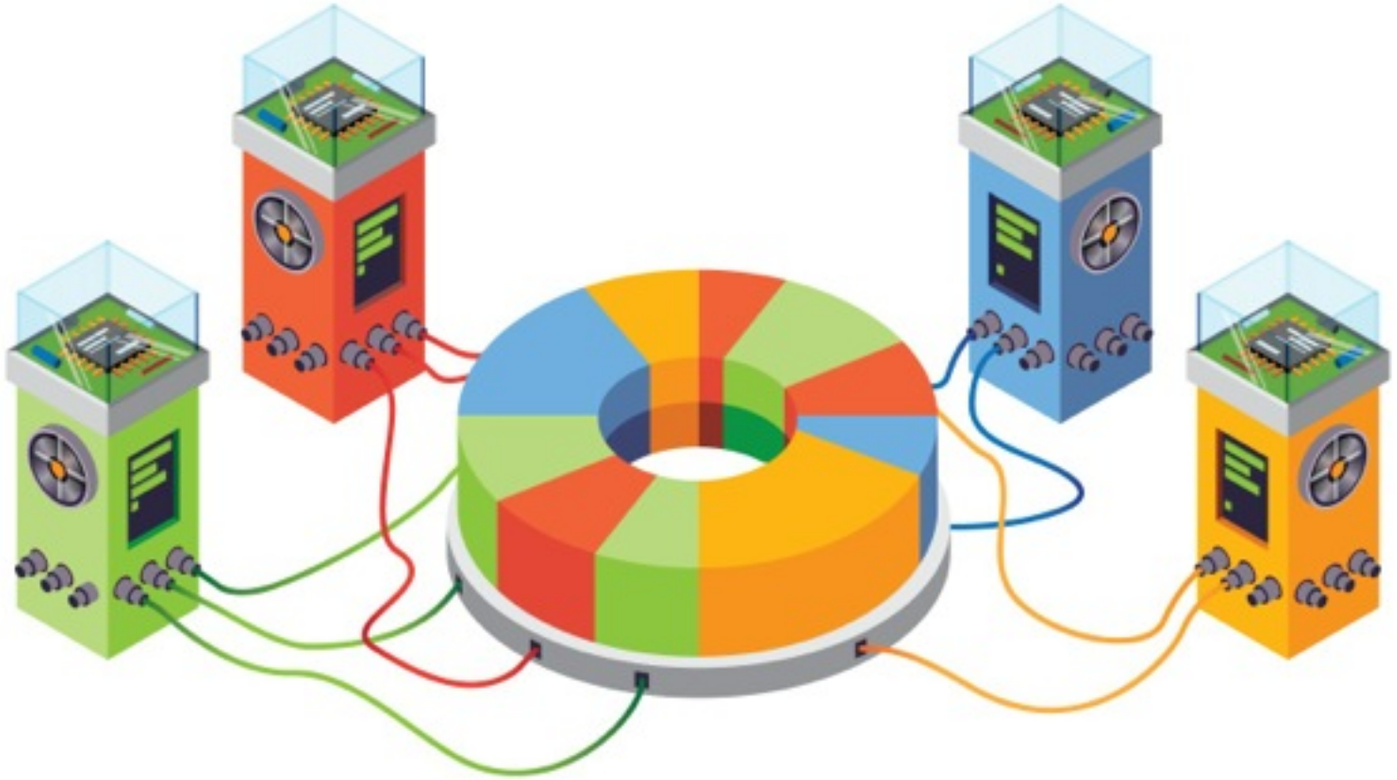
Я.Субботник в Минске, 30.08.2014

Немного цифр

- 4 Петабайта данных
- 10 миллиардов ключей
- 150 серверов
- 4500 дисков
- 5000 запросов в секунду

Elliptics

- key-value сторадж
- Междатацентровая репликация из коробки
- Динамическое определение машин в кластере
- Реплика — DHT



Elliptics

- key-value сторадж
- Междатацентровая репликация из коробки
- Динамическое определение машин в кластере
- Реплика — DHT

Когда DHT это хорошо

Подходит если нет роста объема данных

И имеет всякие плюшки

- key-value
- Естественная балансировка нагрузки
- Автоматическое адаптирование к изменению структуры кластера

Проблемы DHT

Тяжелая процедура масштабирования кластера

- грузим сеть и диски
- нельзя масштабировать больше одной реплики за раз

Все данные хранятся в одной куче

- трудно реализовать разные политики реплицируемости для разных типов данных
- трудно разделить данные по ДЦ
- трудно реализовать процедуру архивироания данных

В условиях постоянного роста объема данных скорость заливки данных может сравниться со скоростью добавления машин в кластер.

Что, если не DHT?

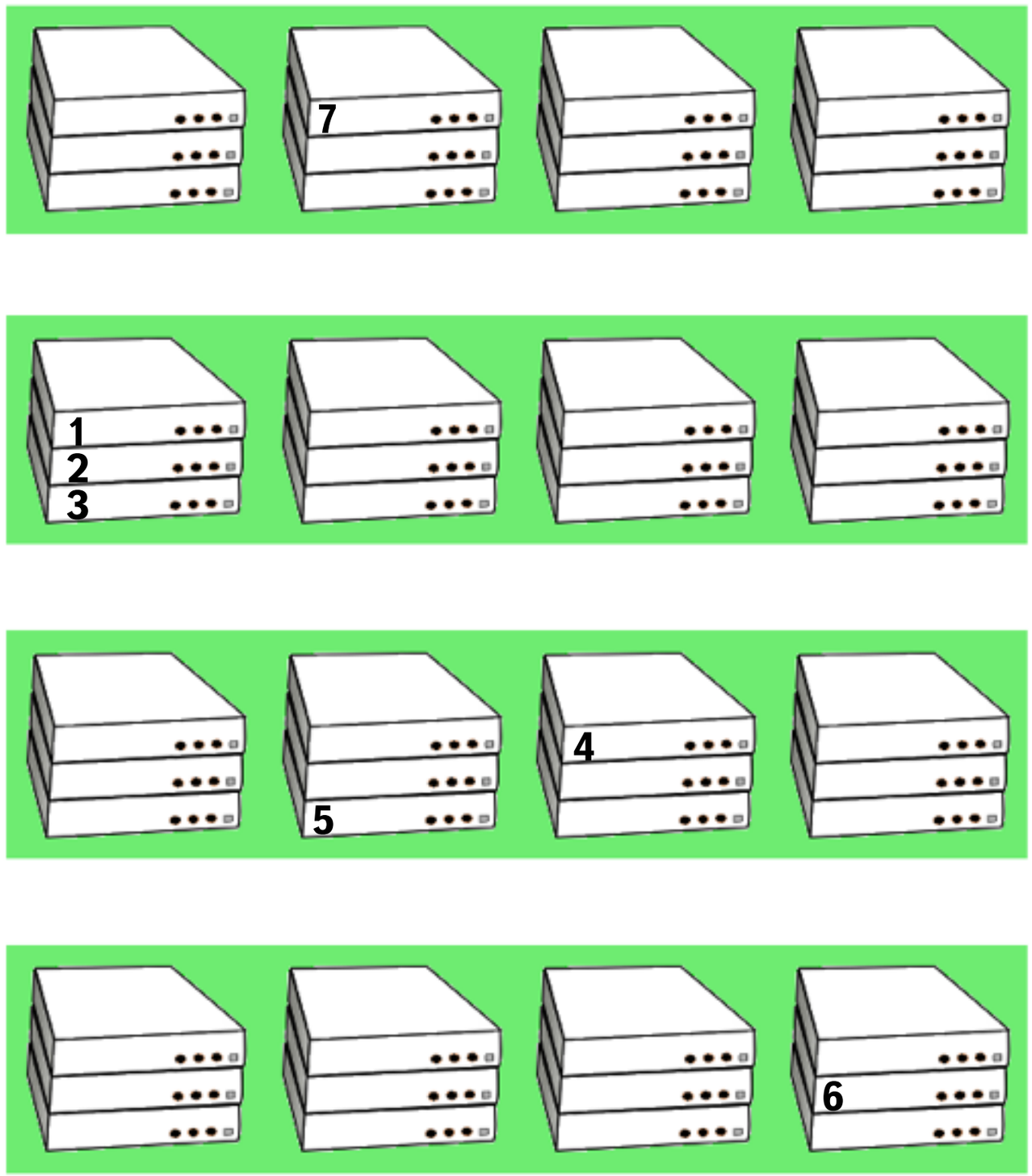
1

2

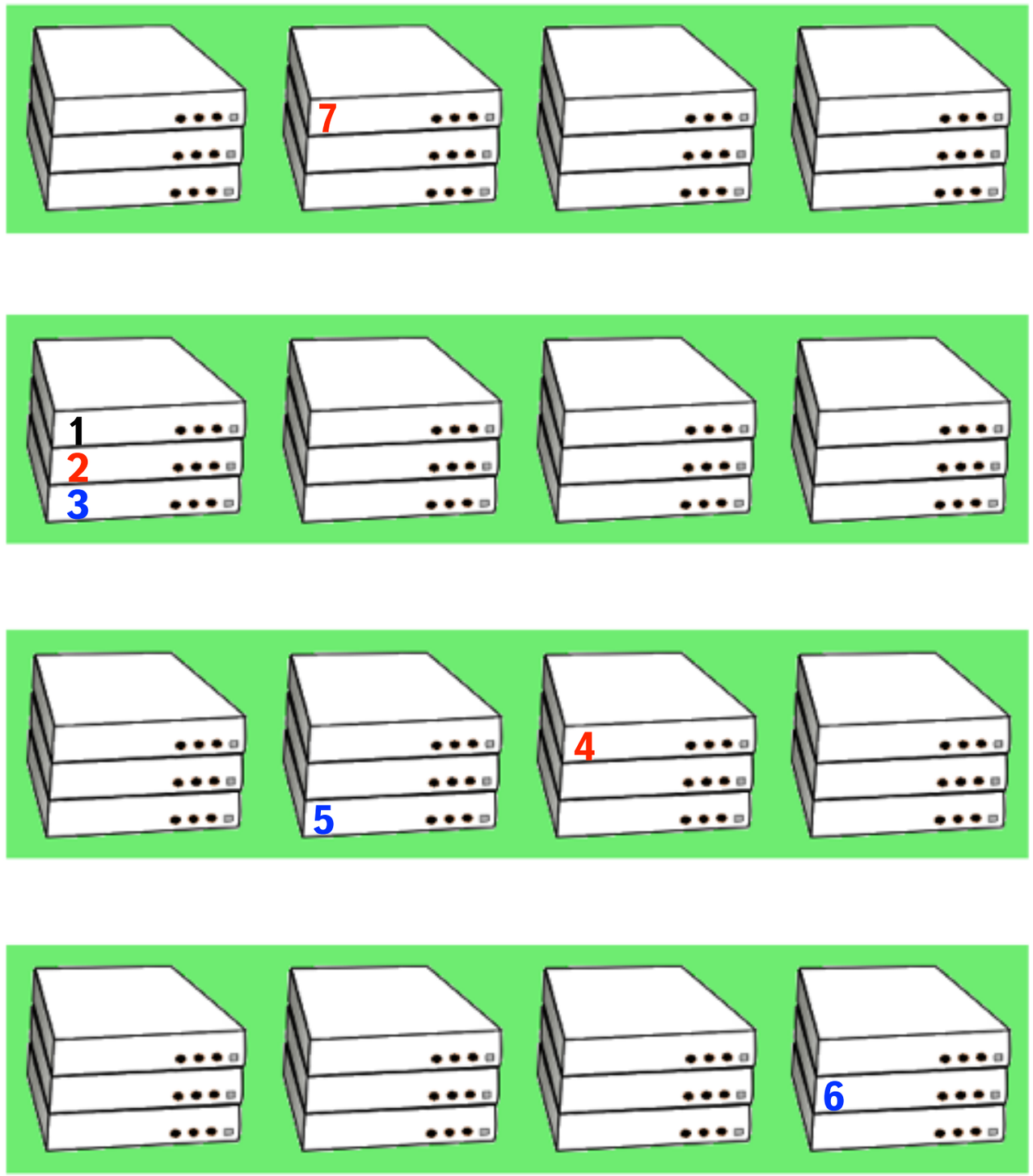
3

Mediastorage

Mediastorage



Mediastorage



Mastermind

- Знает обо всех группах в кластере
- Хранит информацию о каплах и неймспейсах
- Собирает статистику по кластеру
- Облегчает управление кластером

Mastermind для администратора

- Добавление новой группы
- Восстановление реплики
- Перенос группы на другую машину
- Объединение групп в каплы и настройка капла
- Объединение каплов в неймспейсы и настройка неймспейса

Mastermind для клиента

- Балансировка нагрузки
- Предоставление информации о структуре кластера

Mastermind

- Cocaine Worker
- Minion
- Console utility
- Flowmastermind
- libmastermind

Уходим от проблем DHT

- Легко масштабировать
- Гибкие настройки хранилища для каждого пользователя в отдельности
- Легко распределять данные по множеству ДЦ

Новые проблемы

- Пользователь должен запомнить, в какой капл записаны данные
- Большой кластер требует особого подхода

Над чем работаем прямо сейчас

- Кеширующие каплы
- Региональная программа
- Автоматическое обслуживание кластера

Спасибо за внимание!

https://github.com/reverbrain/elliptics

https://github.com/nobodyisme/mastermind

https://github.com/nobodyisme/mastermind-minion

https://github.com/nobodyisme/flowmastermind

https://github.com/yandex/libmastermind

https://github.com/yandex/mediastorage-proxy

derikon@yandex-team.ru