



Дешевле, Надёжнее, Проще
Хранение ПБ видео и фото в ОК

Развлекательная
социальная сеть
Одноклассники

2017

² До 2016 года



OBS

Фотографии

Музыка

Подарки

Видео

250 TB

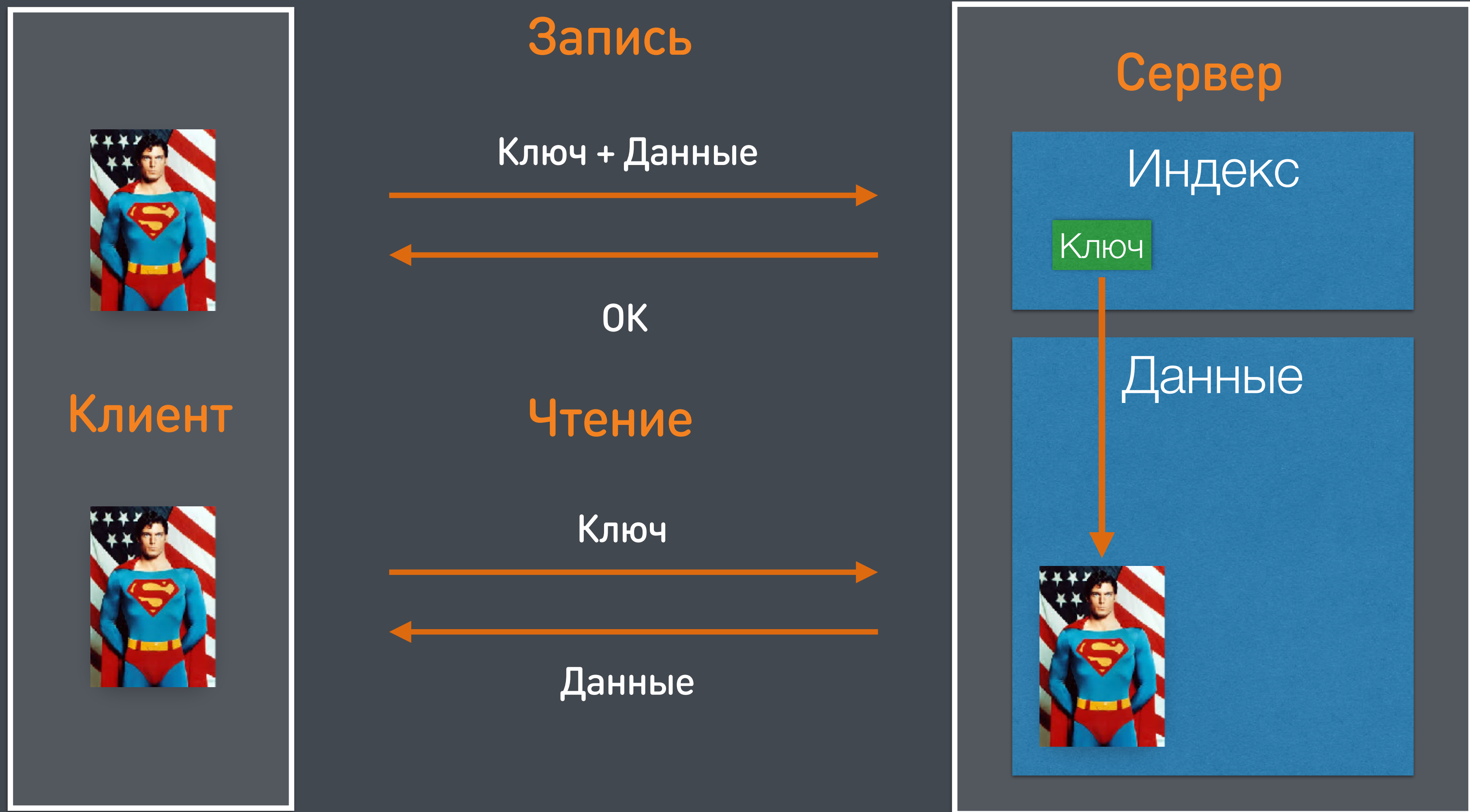
1 ПБ

2011

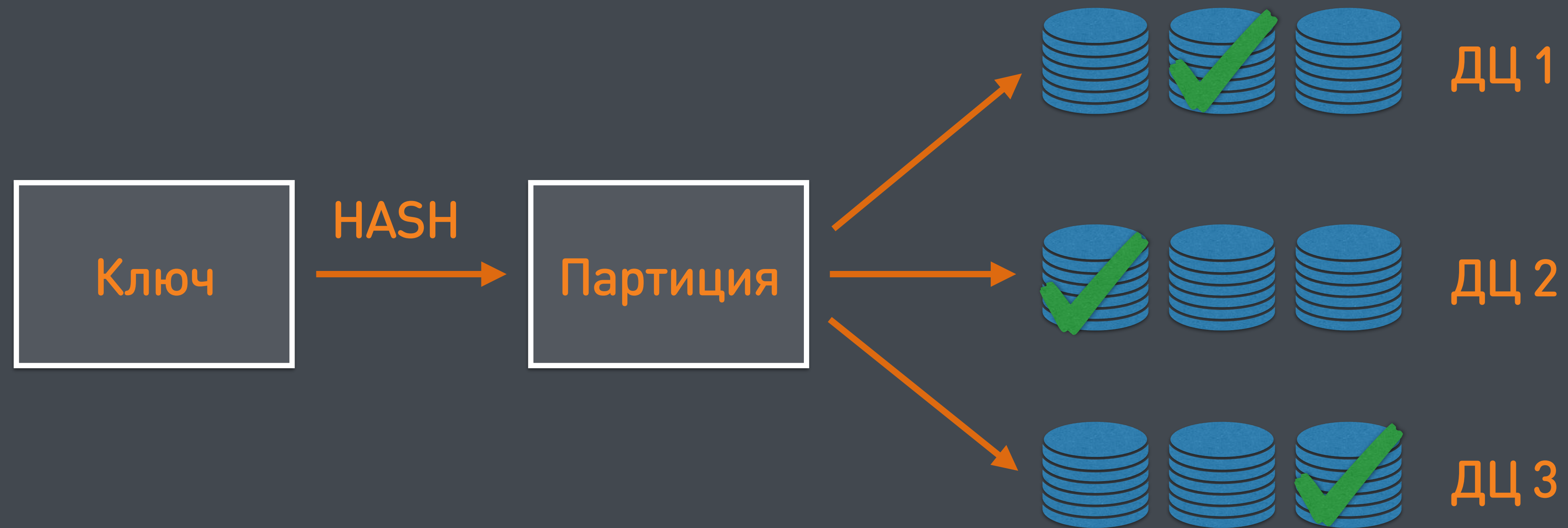
2013

https://www.slideshare.net/profyclub_ru/ss-9541712

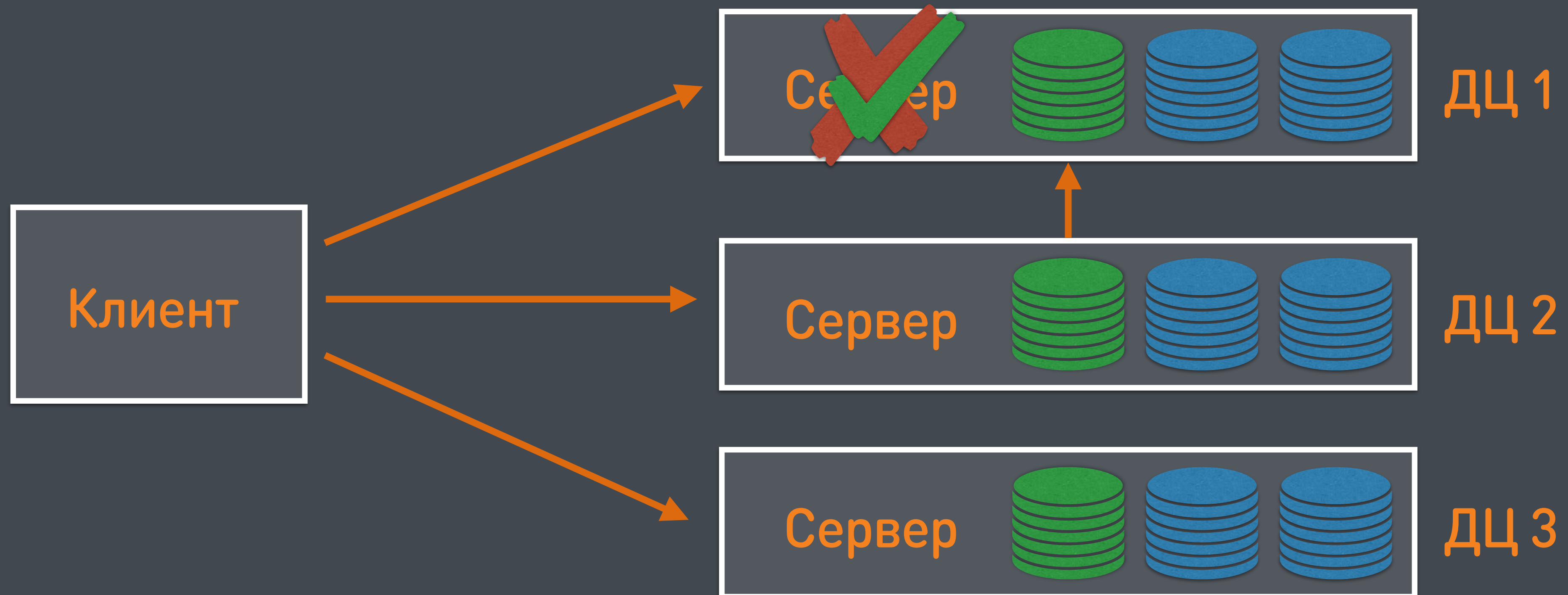
3 OBS



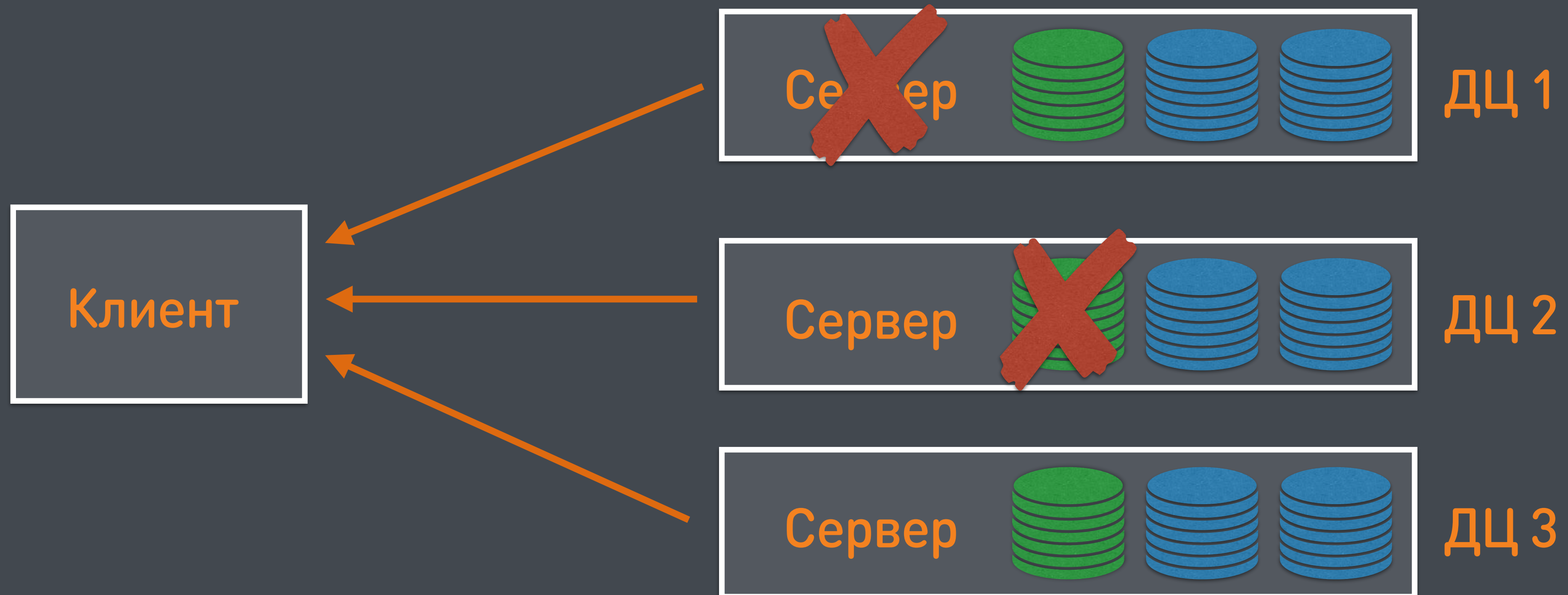
4 Распределение данных в OBS



5 Запись в OBS



6 Чтение в OBS



7 Нагрузка на OBS



250

К/с

Запросы

350

Гб/с

Трафик

8 Оборудование



600

Сервера 4U

18K

Диски 1 - 8 ТБ

9 А в чём проблема?



Видео

Популярность +

Загрузка 4K +

Стриминг =

1 ПБ в месяц

25 ПБ

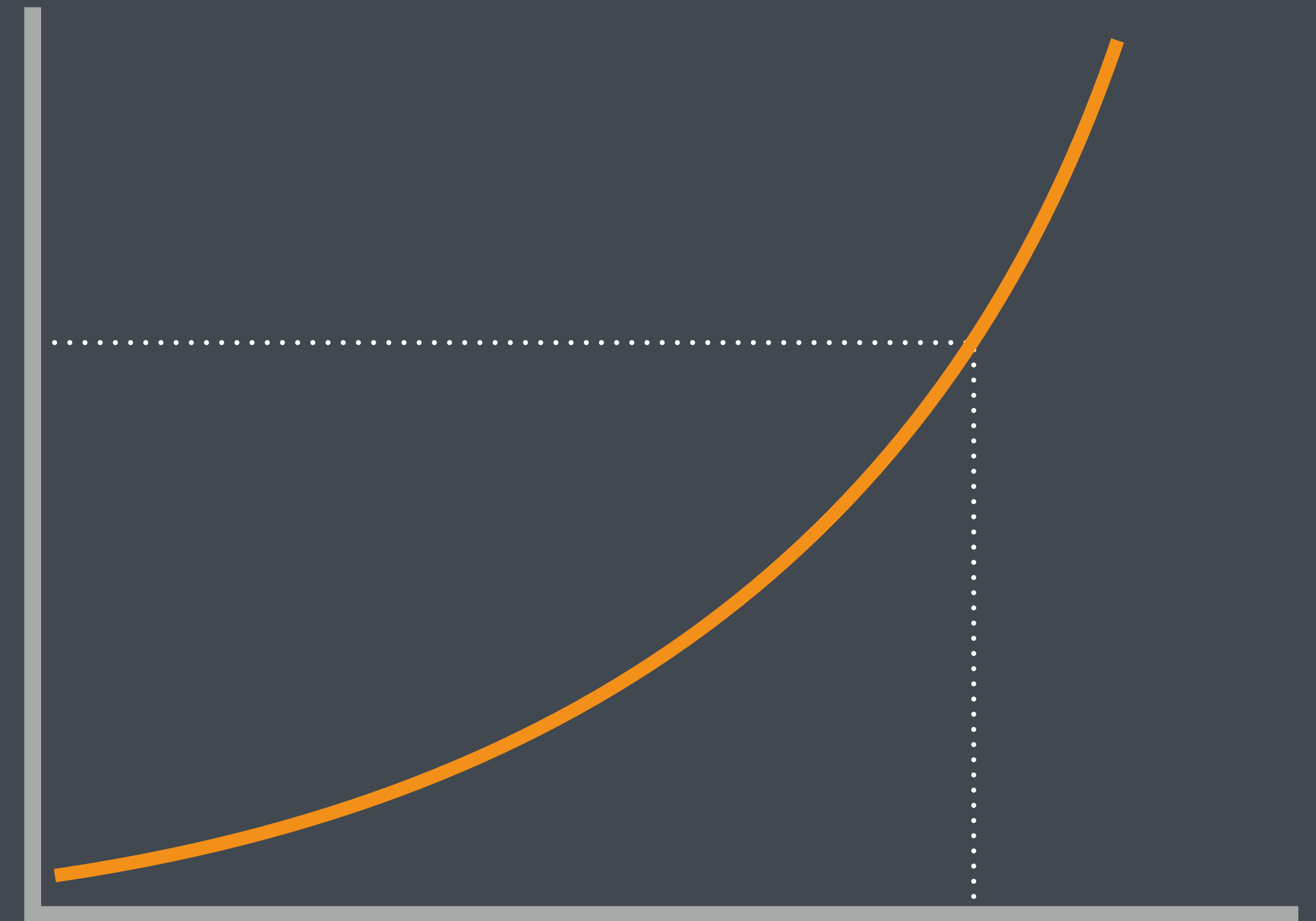
15 ПБ

1 ПБ

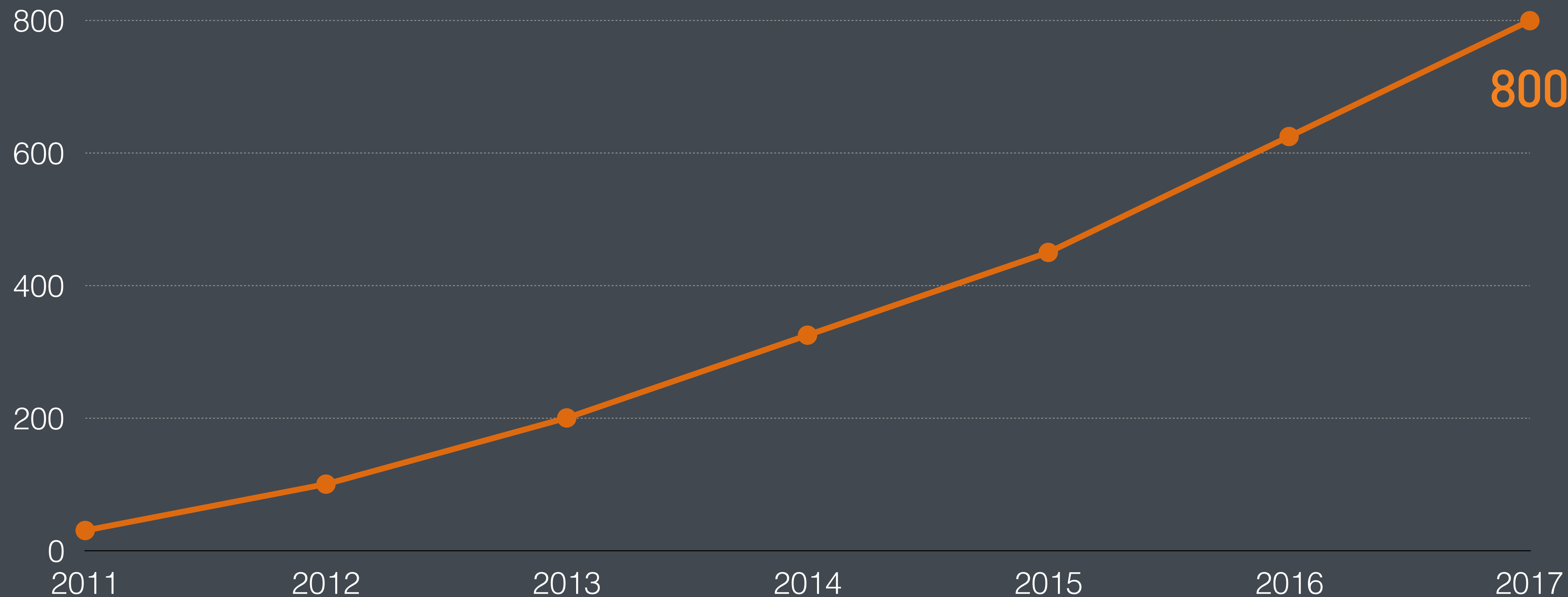
2013

2016

2017



10 Количество серверов в OBS + OCS (Фото + Видео)



11 Проблемы в 2016



Стоимость оборудования

Стоимость размещения

Нехватка места в ДЦ



OCS

Развлекательная
социальная сеть
Одноклассники



13 OCS — Цели



Хранить дешевле

Хранить ЕЩЁ надёжнее

Работа при отказе ДЦ + диска/сервера

Упростить эксплуатацию

14 Давай уменьшим избыточность



Схема репликации

Избыточность

Допустимая потеря дисков

Зеркало (1 + 1)

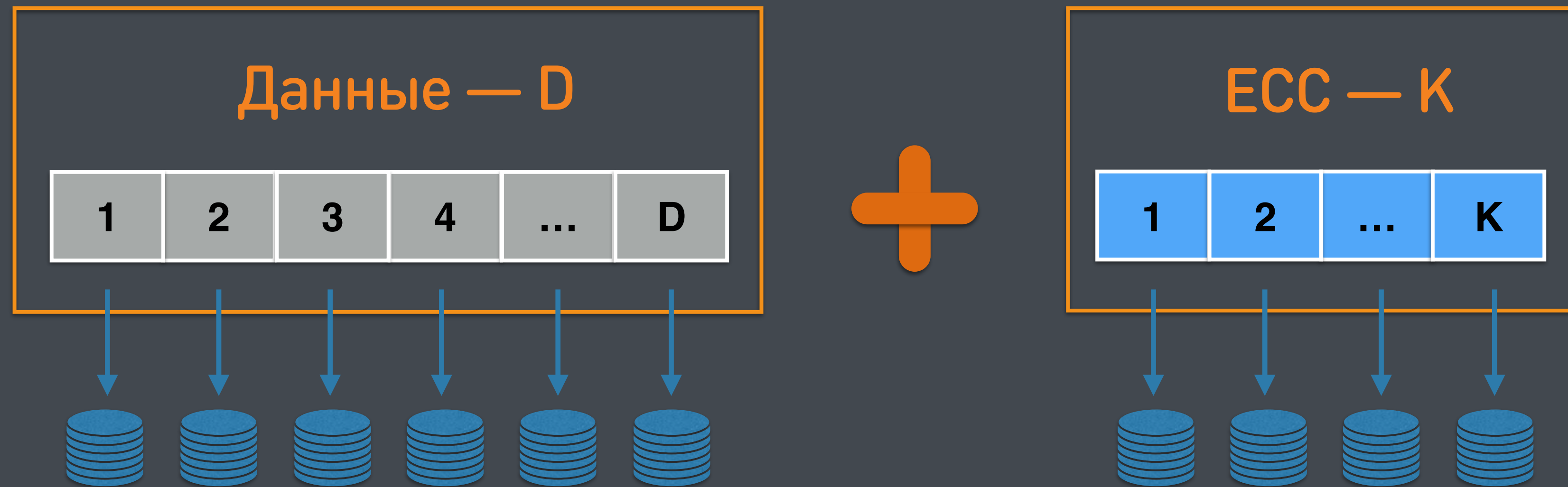
100%

XOR (2 + 1)

50%

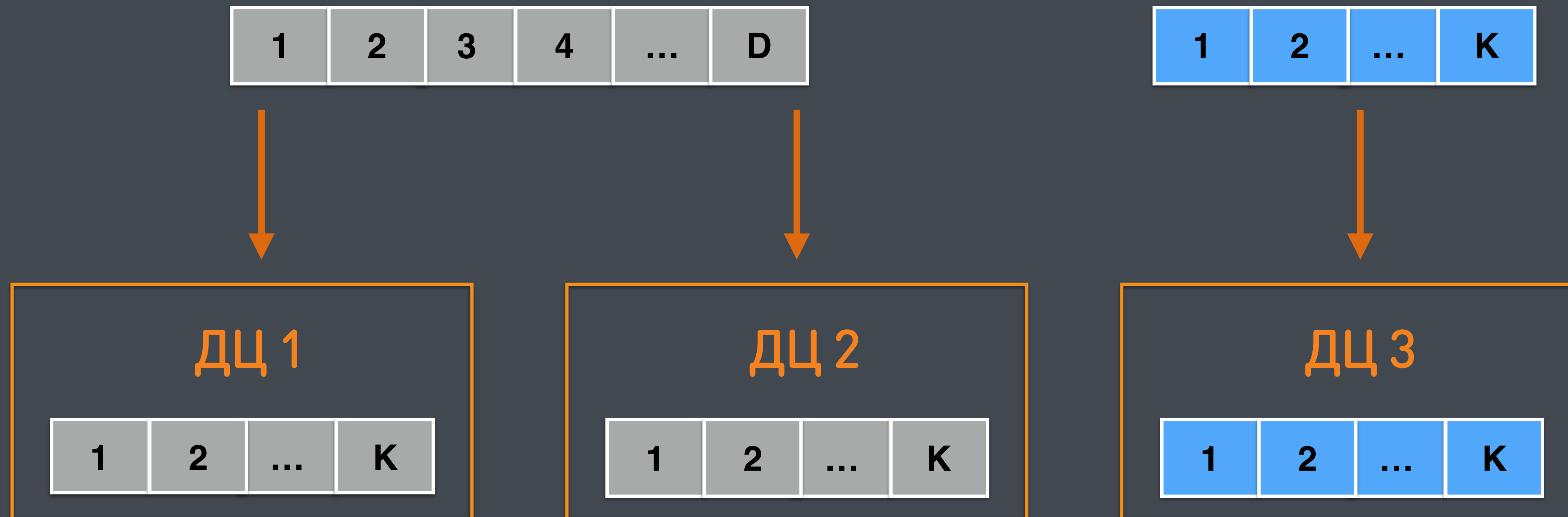


15 Коды Рида — Соломона



Можно потерять любые K дисков

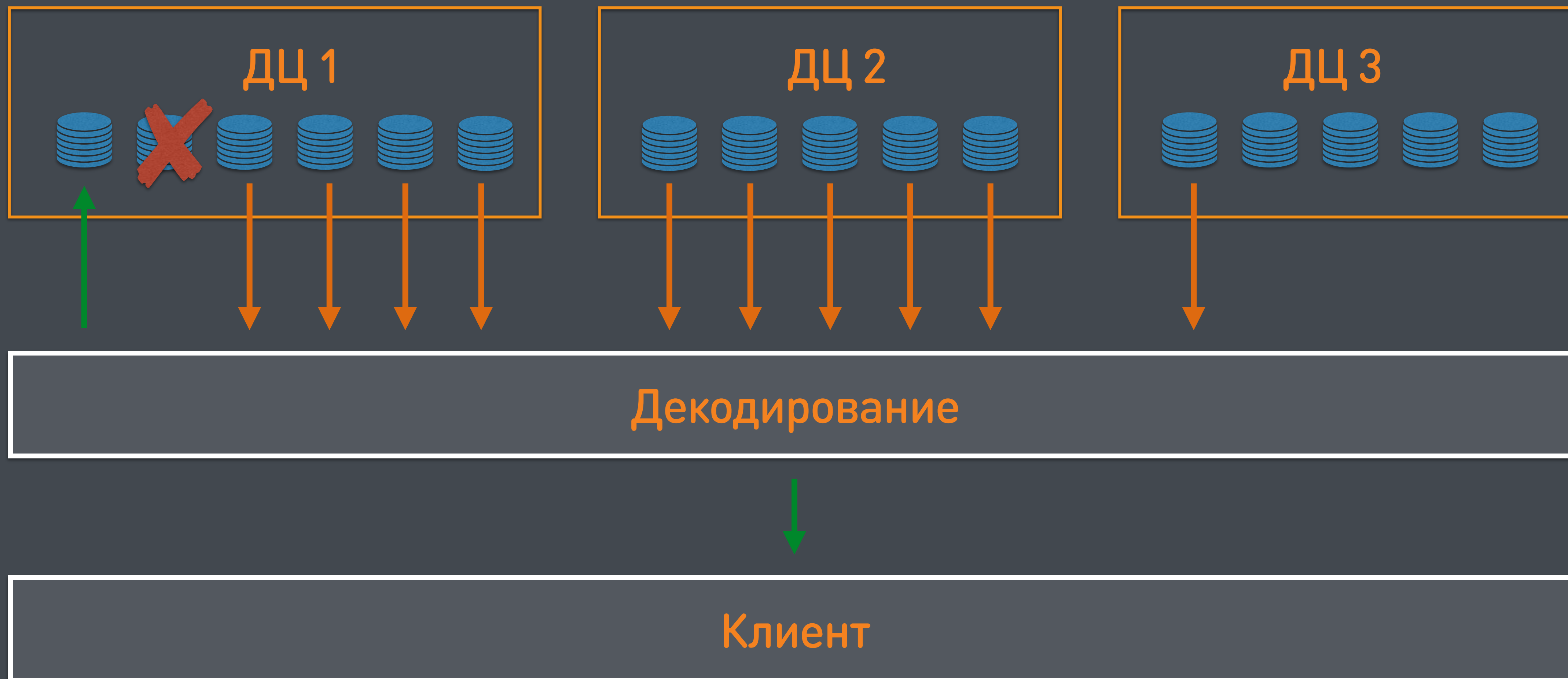
16 Распределение по 3 ДЦ



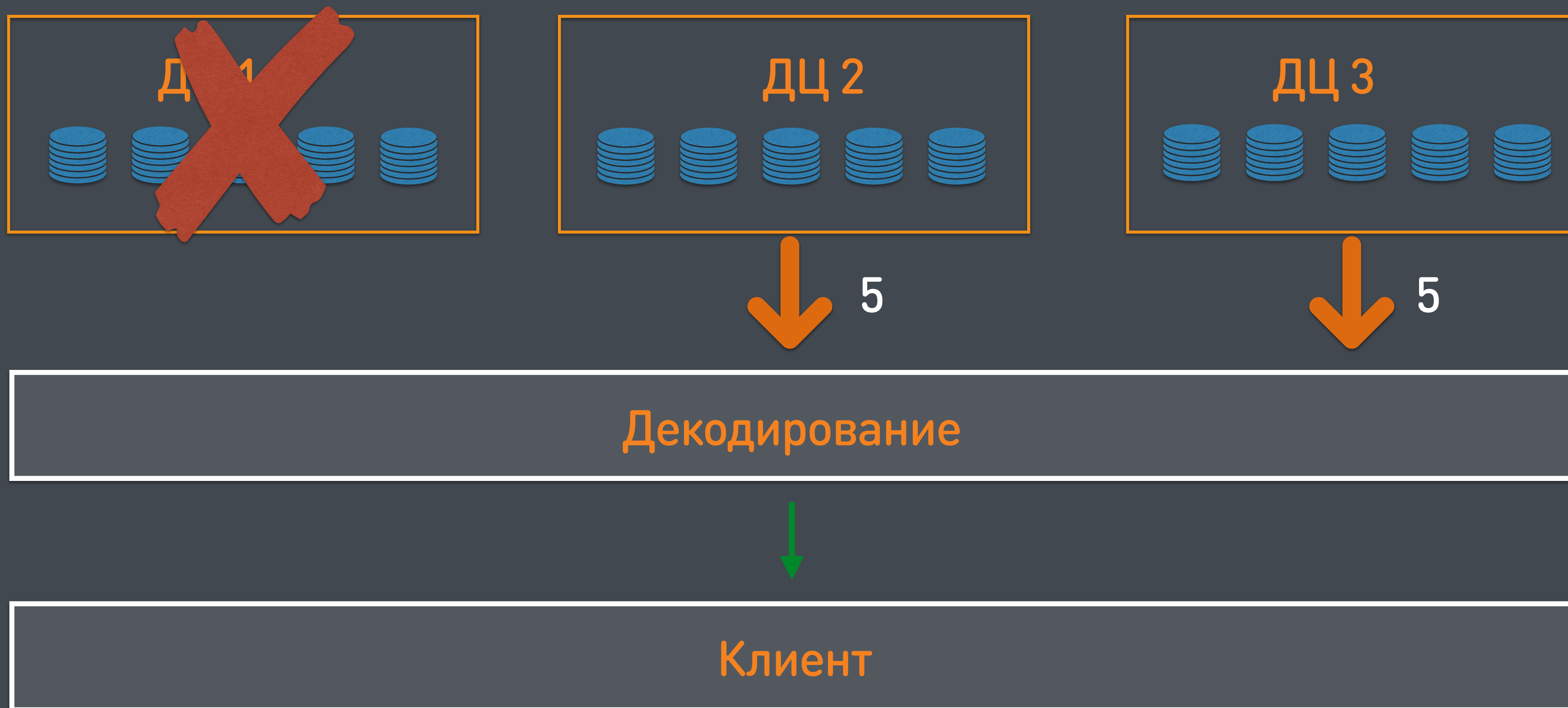
$$K \geq 1/3 (D + K)$$

Избыточность $\geq 50\%$

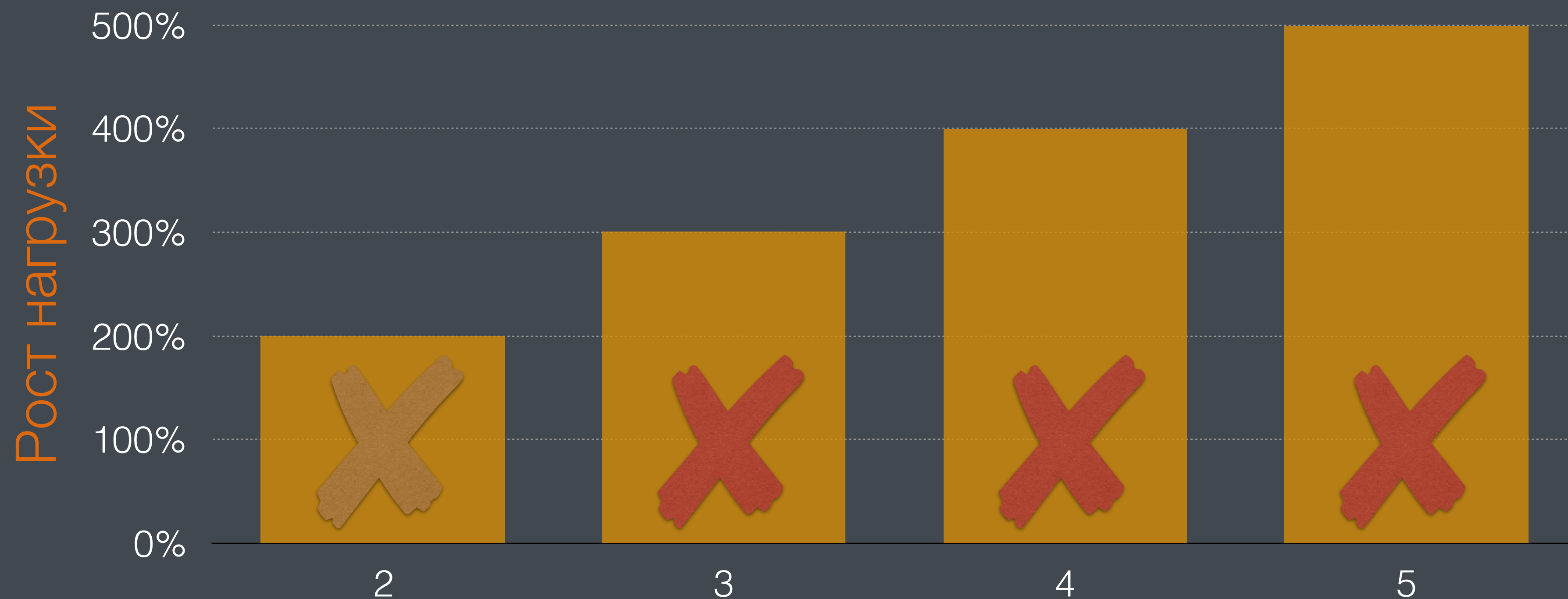
17 Восстановление данных 10 + 5



18 Отказ ДЦ

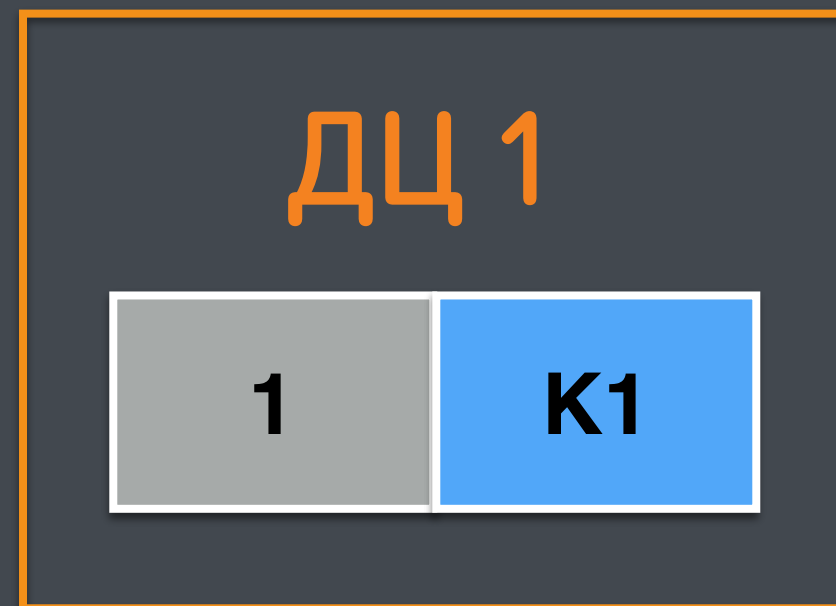


19 Трафик / IOPS при отказе ДЦ



Допустимая потеря дисков при избыточности 50%

20 Коды Рида-Соломона 3 + 3



Допустима потеря 3 дисков

Работа при отказе ДЦ + диска

Нагрузка при отказе ДЦ + 150%

Избыточность + 100%

21 Мы пошли другим путём



Данные

EvenOdd

1	2	3	4	5
---	---	---	---	---

ДЦ 1

E1	E2
----	----

6	7	8	9	10
---	---	---	---	----

ДЦ 2

E3	E4
----	----

XOR

XOR

X1	X2	X3	X4	X5
----	----	----	----	----

ДЦ 3

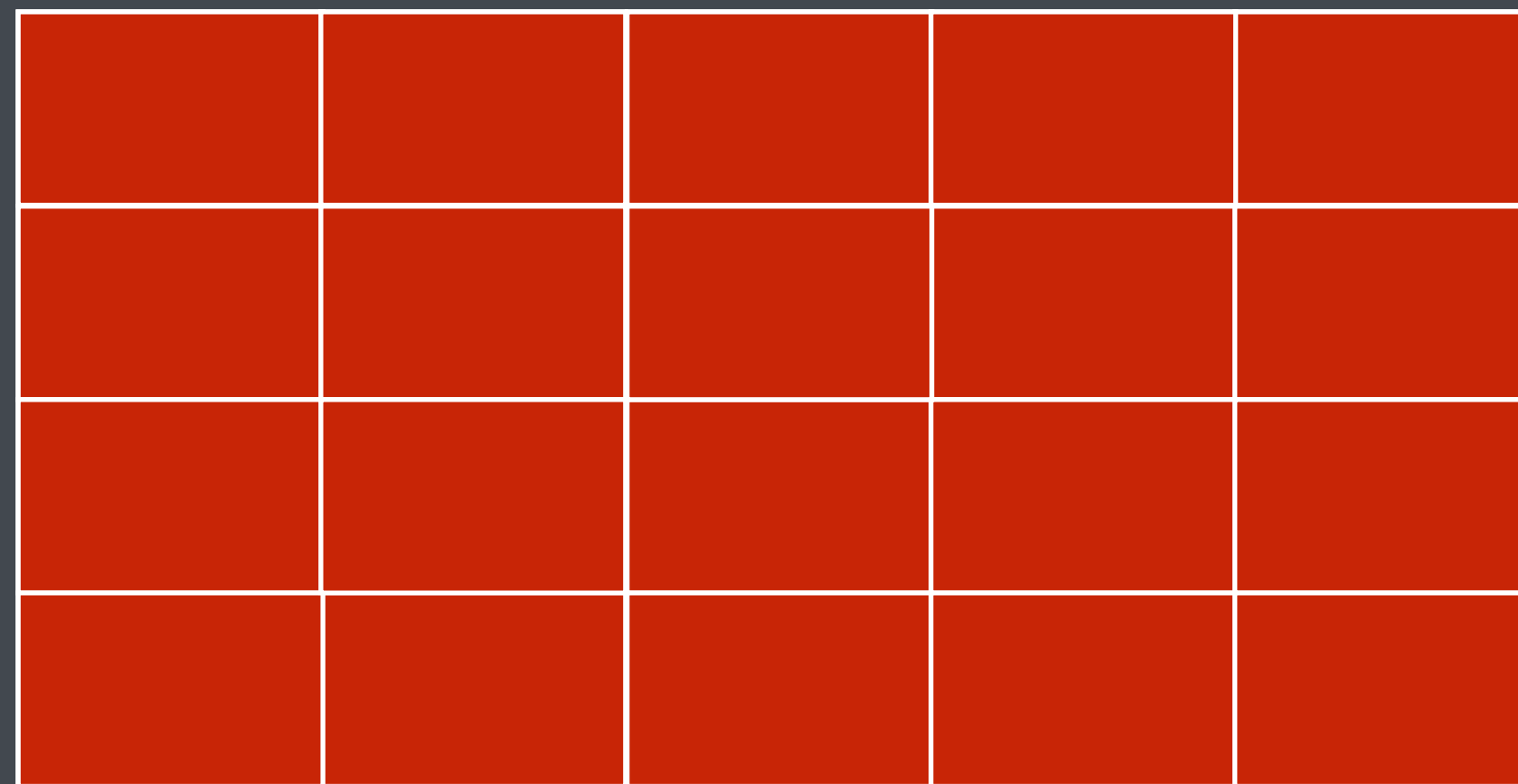
X6	X7
----	----

Данные — 10

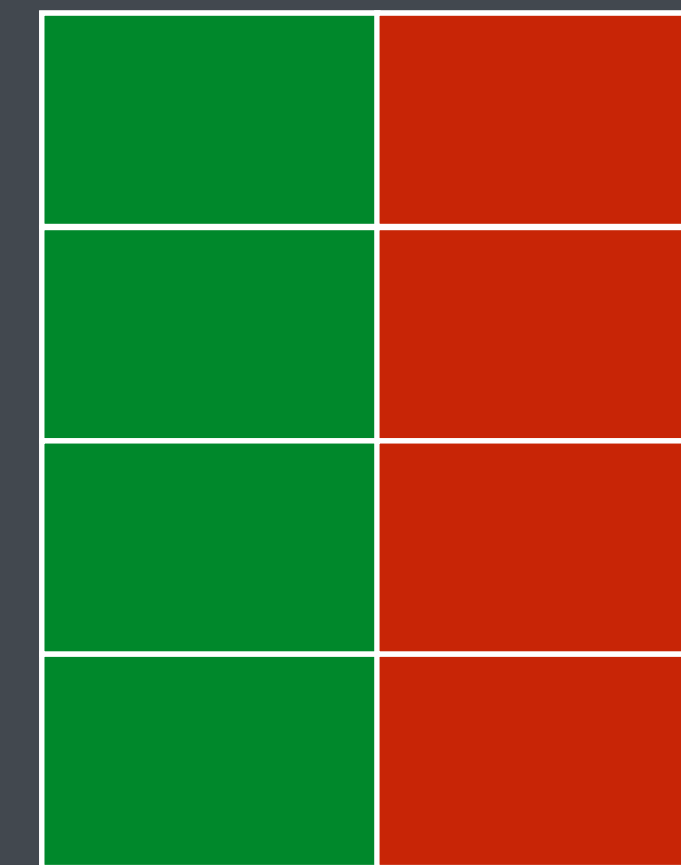
Коды — 11

Избыточность 110%

22 EvenOdd



XOR

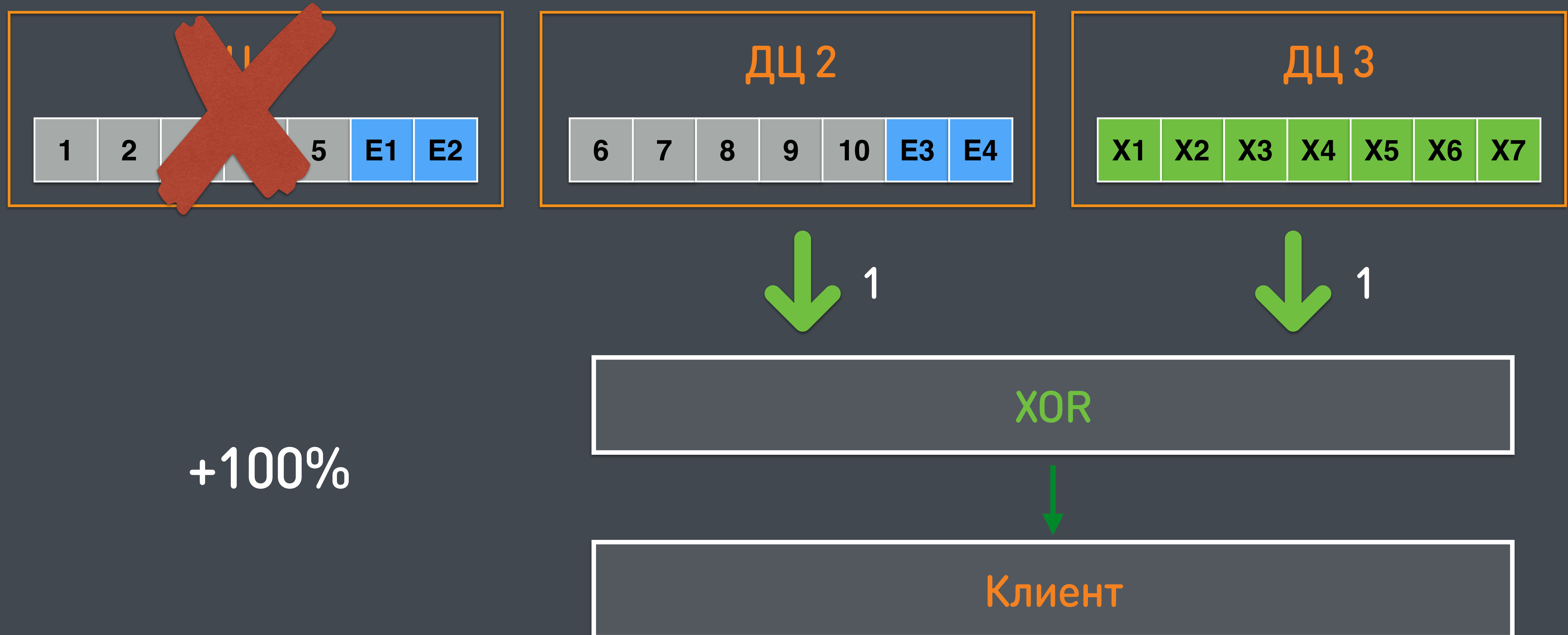


Можно потерять любые 2 диска

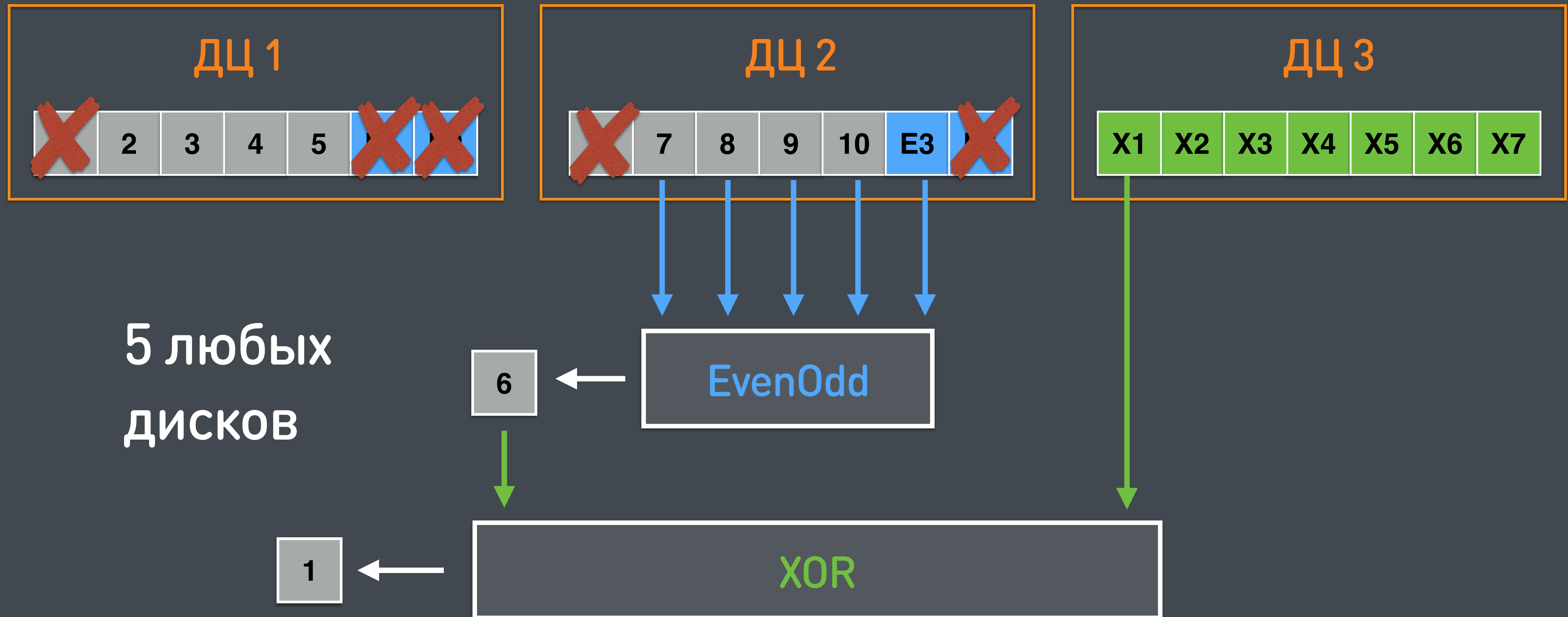
23 Восстановление данных



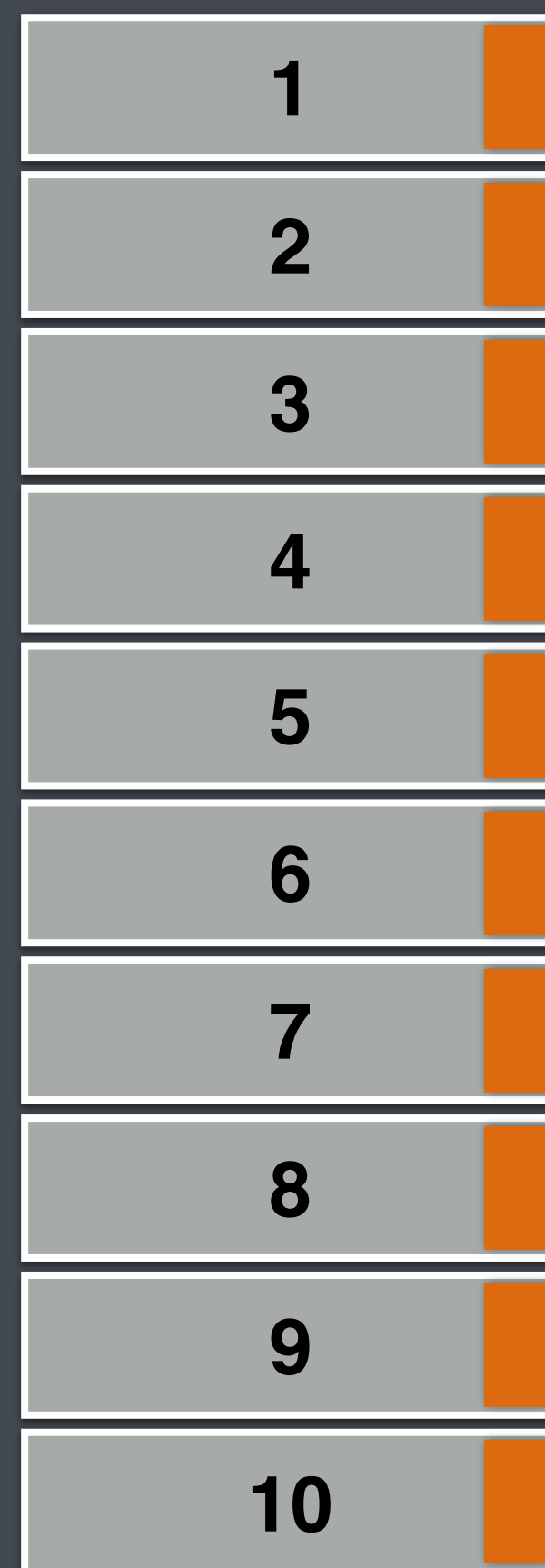
24 Отказ ДЦ



25 Много отказов



26 Как записать данные



10 X



Нагрузка
/IOPS



1 X



27 Проблема с записью



ДЦ 1

2
3
4
5
E3
E4

ДЦ 2

6
7
8
9
10
E3
E4

ДЦ 3

X1
X2
X3
X4
X5
X6
X7

28 Как писать данные



ДЦ 1

2
3
4
5
E3
E4

ДЦ 2

6
7
8
9
10
E3
E4

ДЦ 3

X1
X2
X3
X4
X5
X6
X7

29 У нас же есть OBS



2
3
4
5
E3
E4

6
7
8
9
10
E3
E4

X1
X2
X3
X4
X5
X6
X7

30 Чего мы добились



OBS

OCS

Избыточность

200 %

110 %

Допустимая потеря дисков

2

5



Архитектура

Развлекательная
социальная сеть
Одноклассники



Сегмент 84 ГБ

Данные 40 ГБ



33 Хранение блоков — сервер Данных

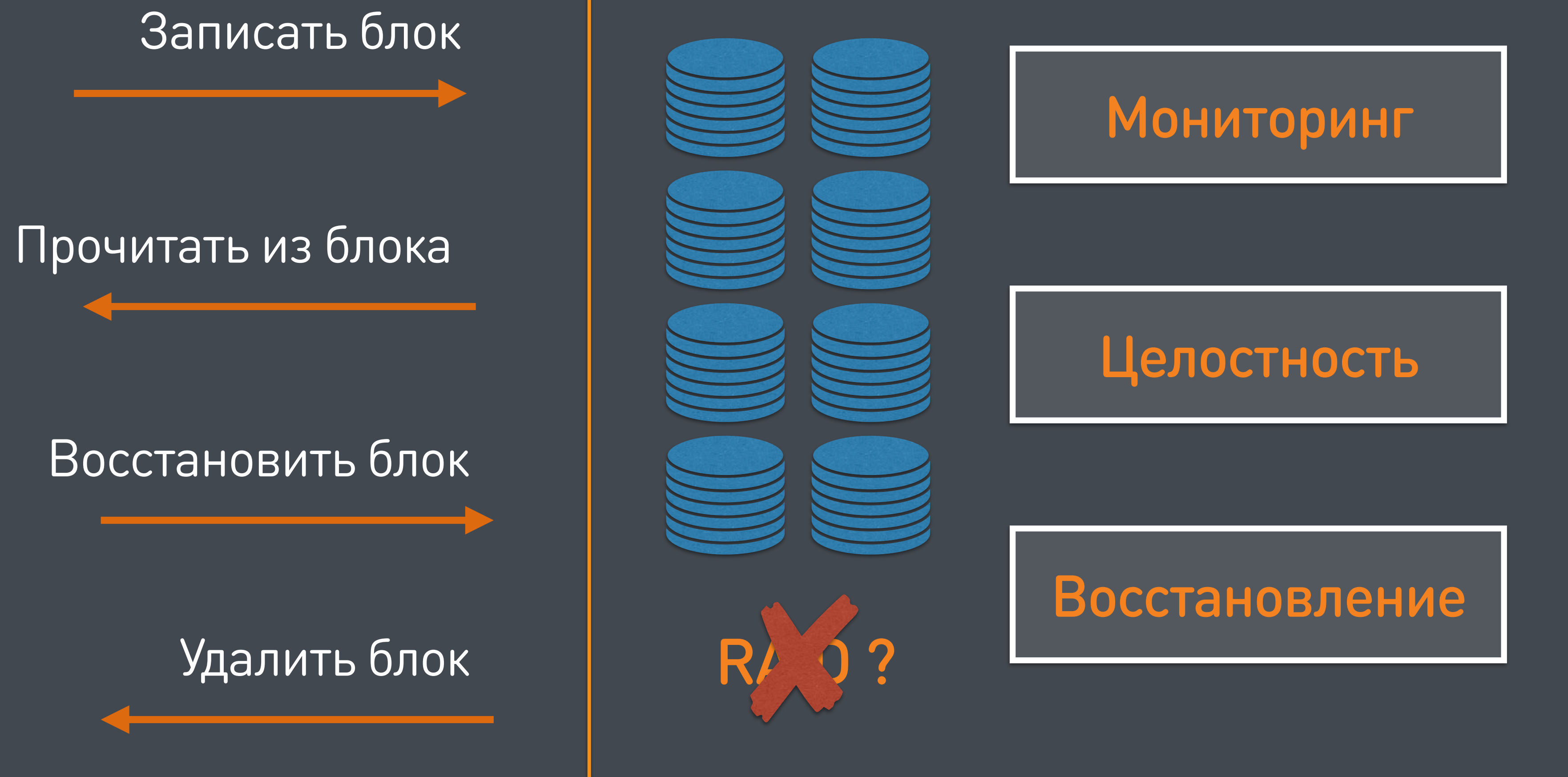
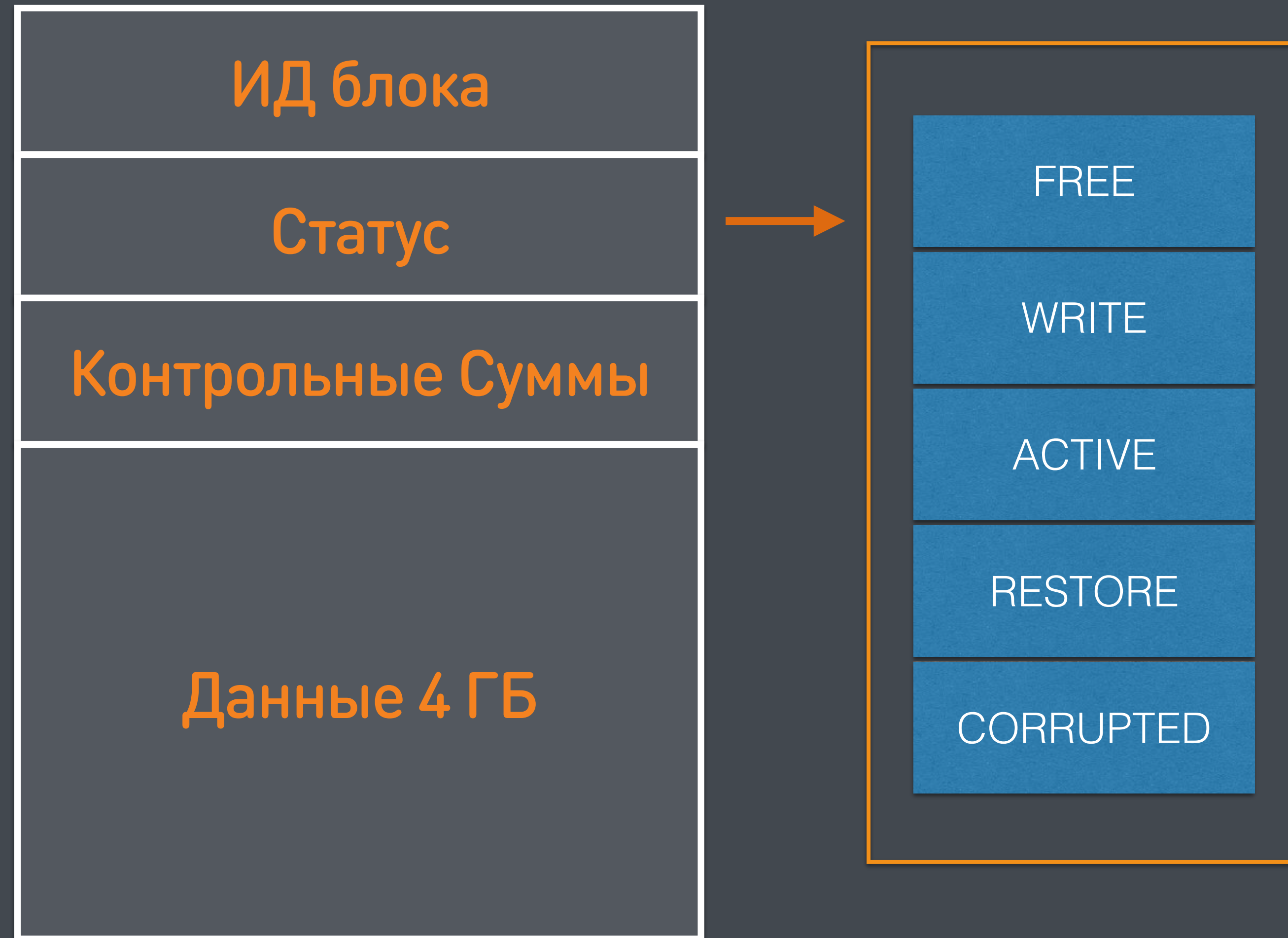




Таблица блоков
Блок 1
Блок 2
....
Блок 2



36 Работа с большими данными — one-nio



one.nio.net.Socket

one.nio.mem.DirectMemory

SendFile

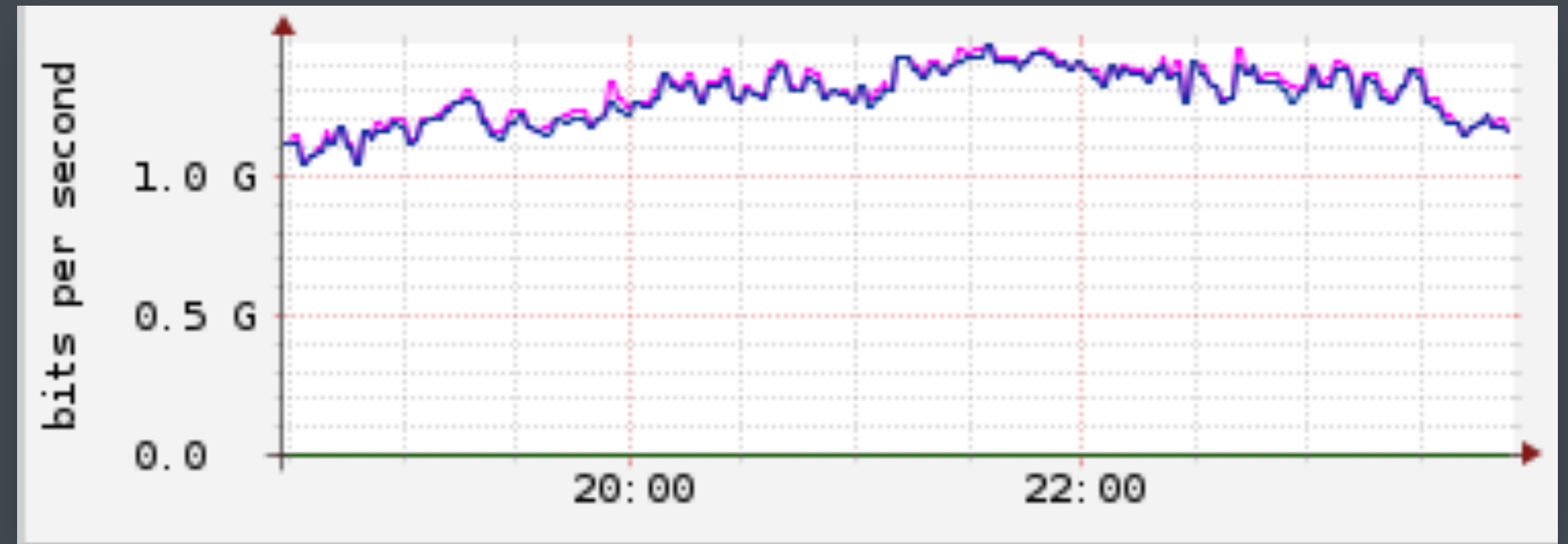
Offheap

github.com/odnoklassniki/one-nio

37 Производительность

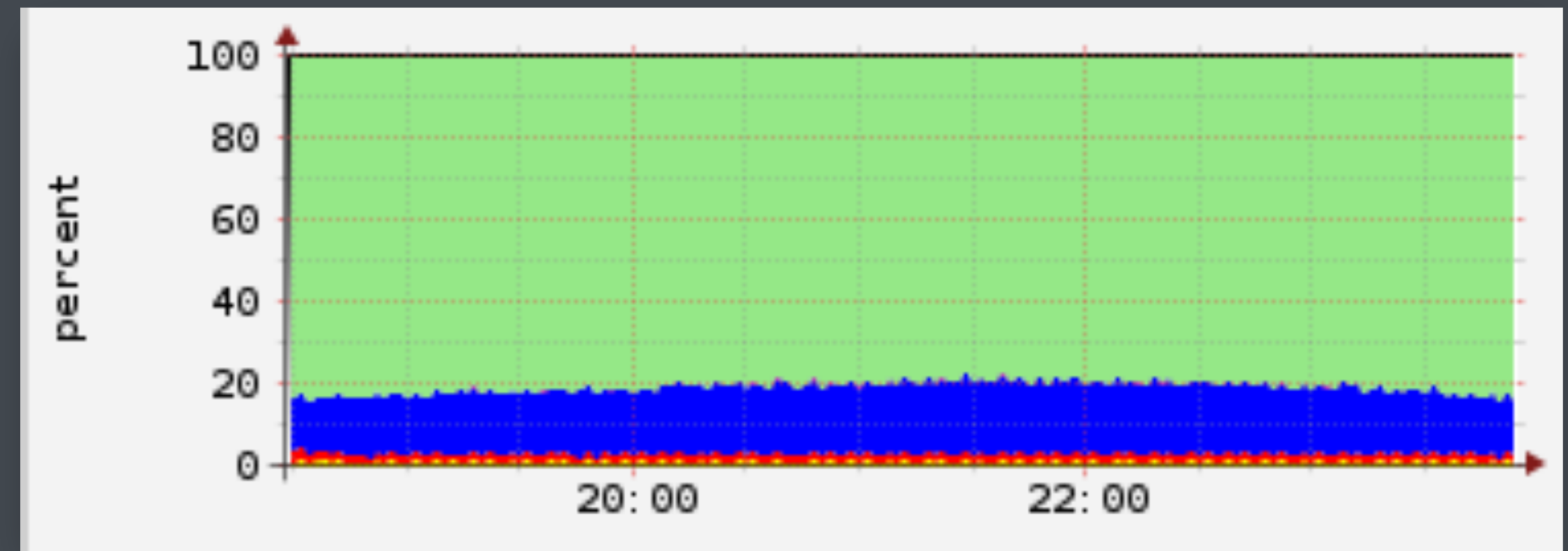


Номинальная нагрузка 1.5 Гб/с



Загрузка CPU User 1%

IO Wait 17%



38 Отказы дисков



Каждый диск изолирован:

Ошибки чтения/записи

Диски могут потеряться

Тормоза перед смертью

Единица кластера

Мониторинг

Пул обработчиков запросов

Шедулер задач

39 Замена диска в OBS



Отмонтировать



Заменить диск

Создать раздел



Замонтировать

Восстановить данные

40 А зачем нам файловая система?



Получаем список блочных устройств

```
new RandomAccessFile('/dev/sdc');
```

MBR + GPT из Java

ИД диска в GPT

41 Что получилось?



Не надо монтировать, форматировать



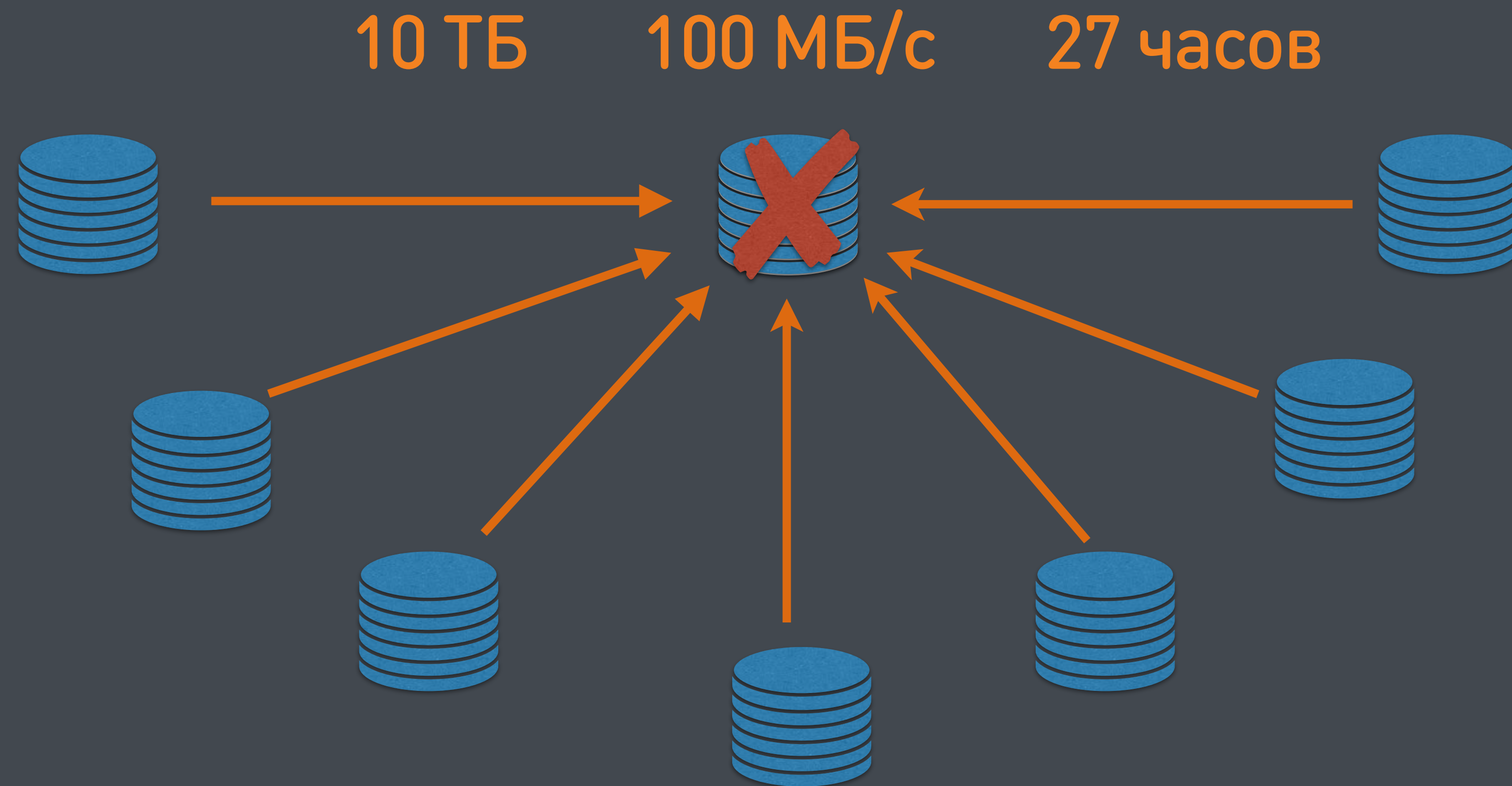
Автоматический ввод в эксплуатацию

Заменить диск

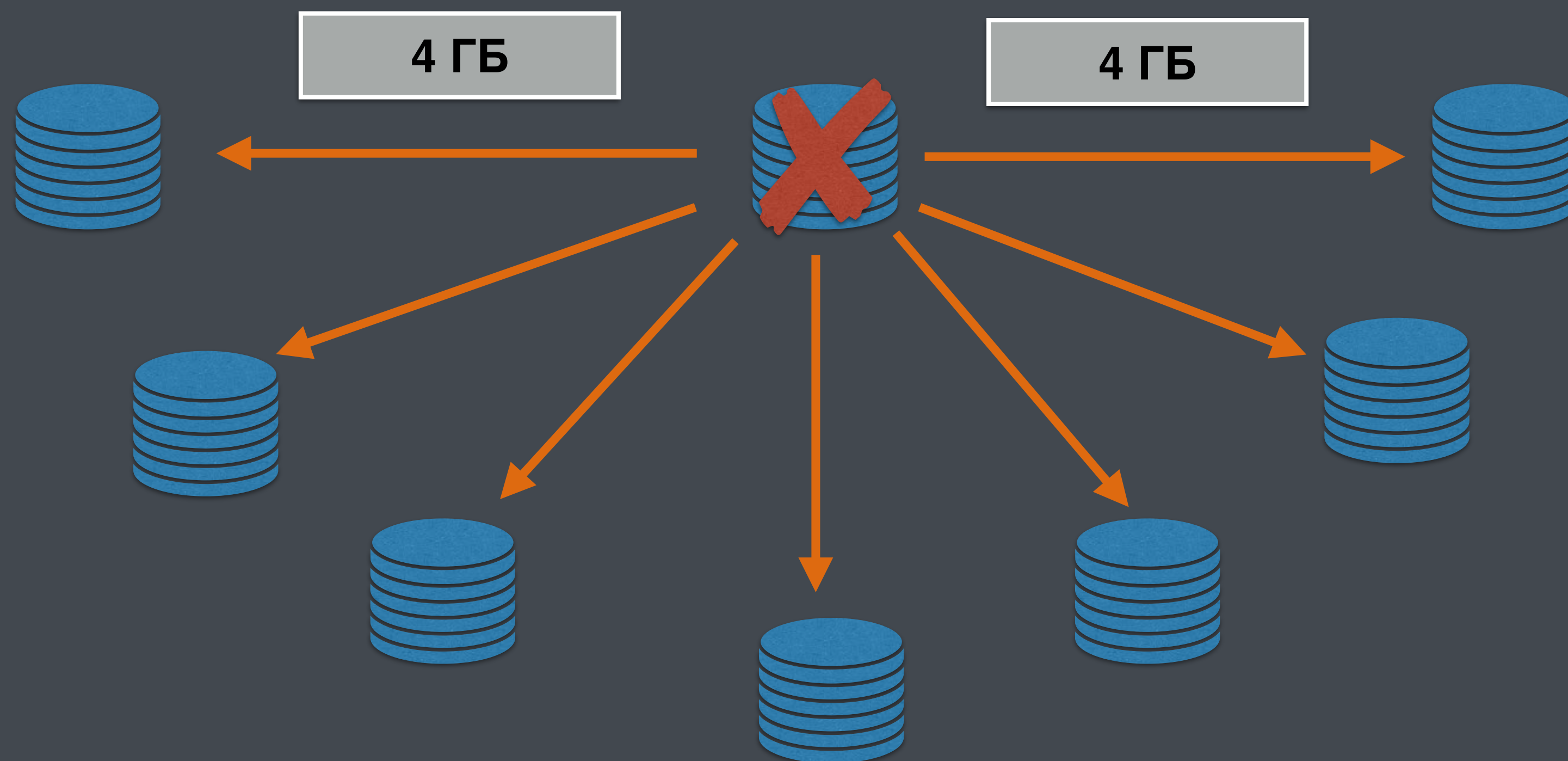
1 человек вместо 2-х



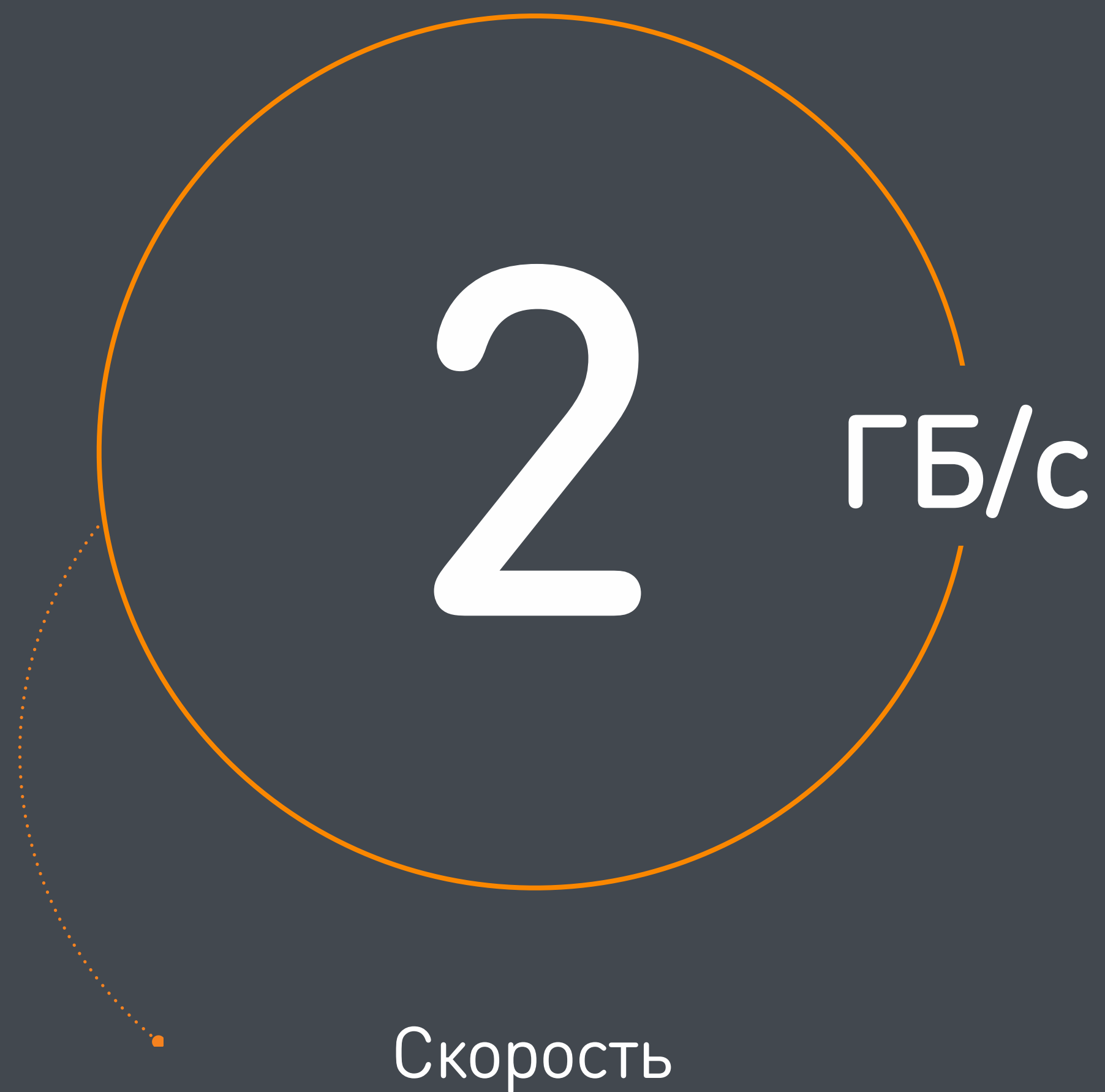
42 Восстановление данных в OBS



43 Восстановление данных в OCS



44 Скорость восстановления



45 Неравномерное заполнение дисков



Временная недоступность

Замена диска

Расширение кластера

46 Балансировка



Диск, заполненный > среднего



Диск, заполненный < среднего



Восстановить на B блок с A

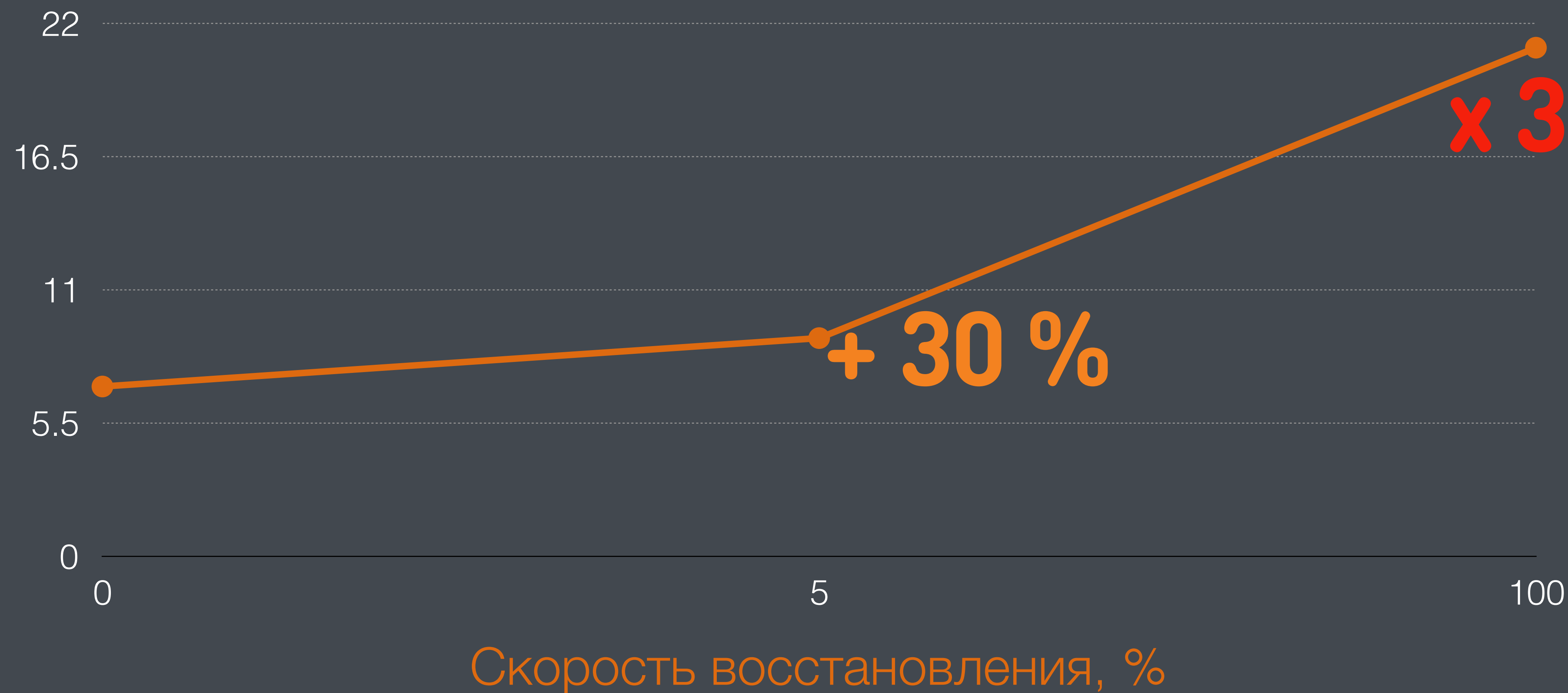


Одинаковый % заполнения

47 Время чтения при восстановлении данных



Время, мс



48 Что можно сделать?



CFQ + `ioprio_set` (`one.nio.os.Proc`)

Клиентские запросы: высокий приоритет

Системные запросы: низкий приоритет

Фоновые процессы: `nice`, `idle`

Ограничение трафика

49 Расширение кластера

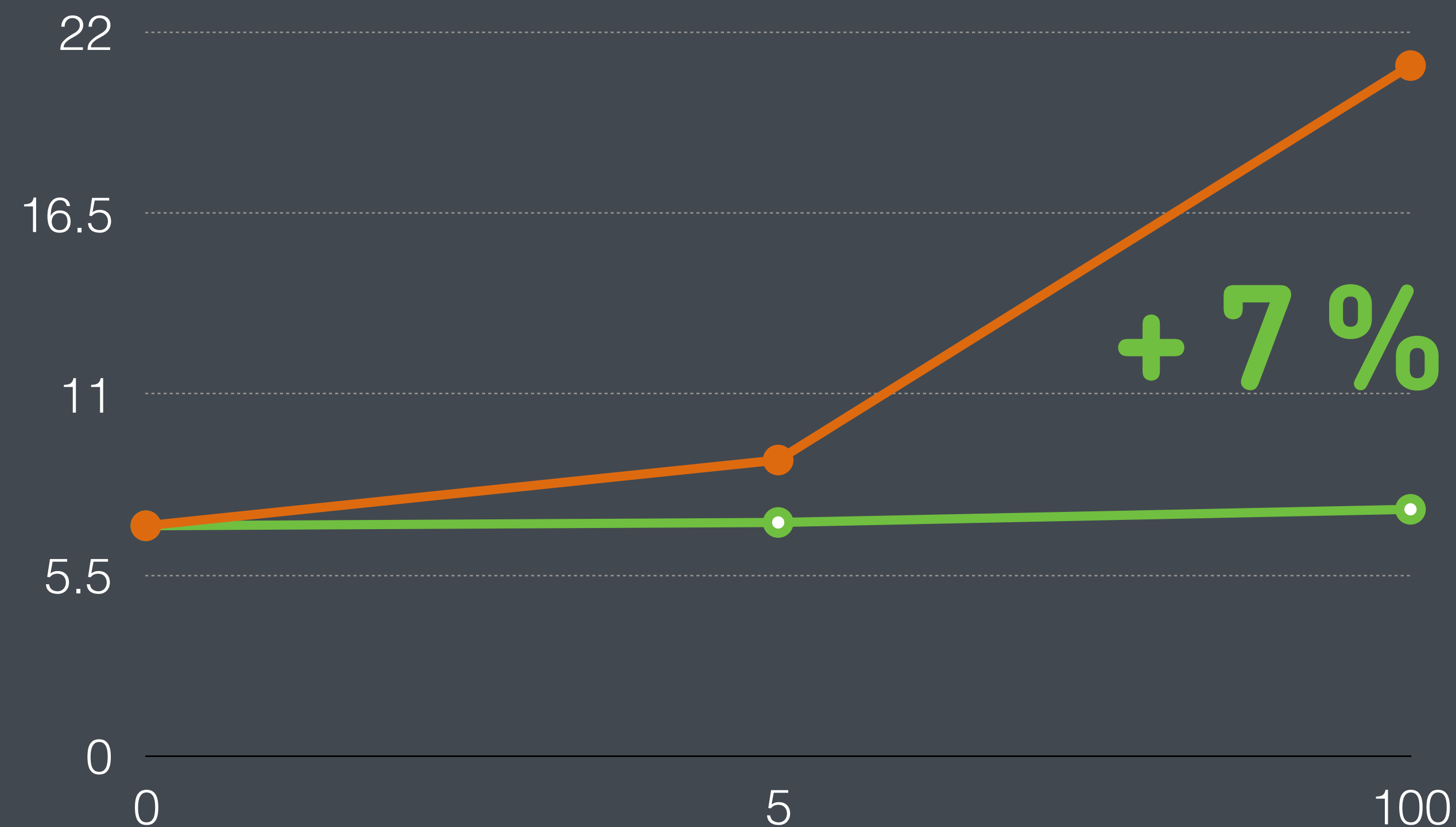


Добавлено 2000 дисков

Перемещено 6.75 ПБ

Заняло 14 дней

Время, мс



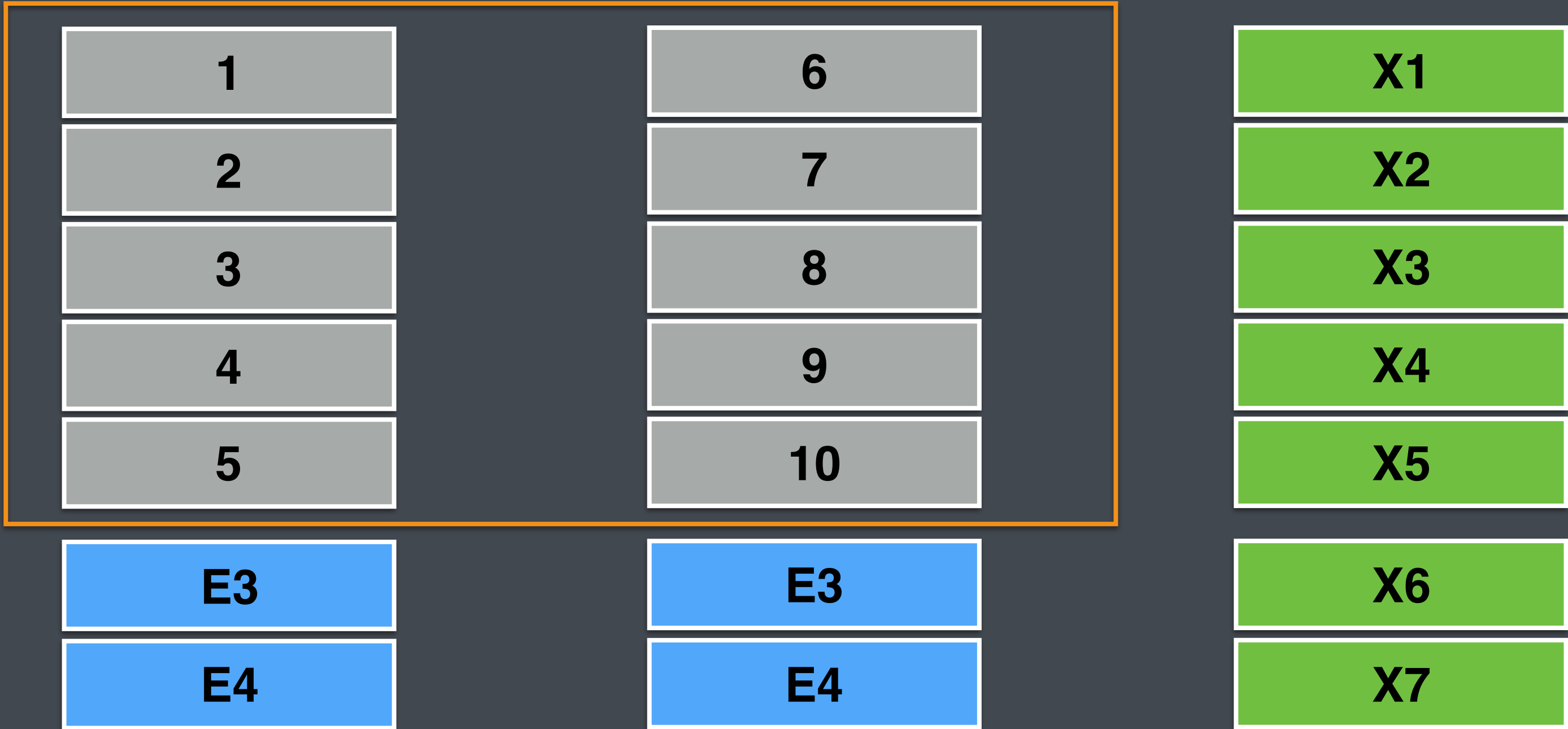
Скорость восстановления, %



Сегмент 84 ГБ

Блок 4 ГБ

Данные 40 ГБ



51 Как прочитать данные



Размер

ИД сегмента

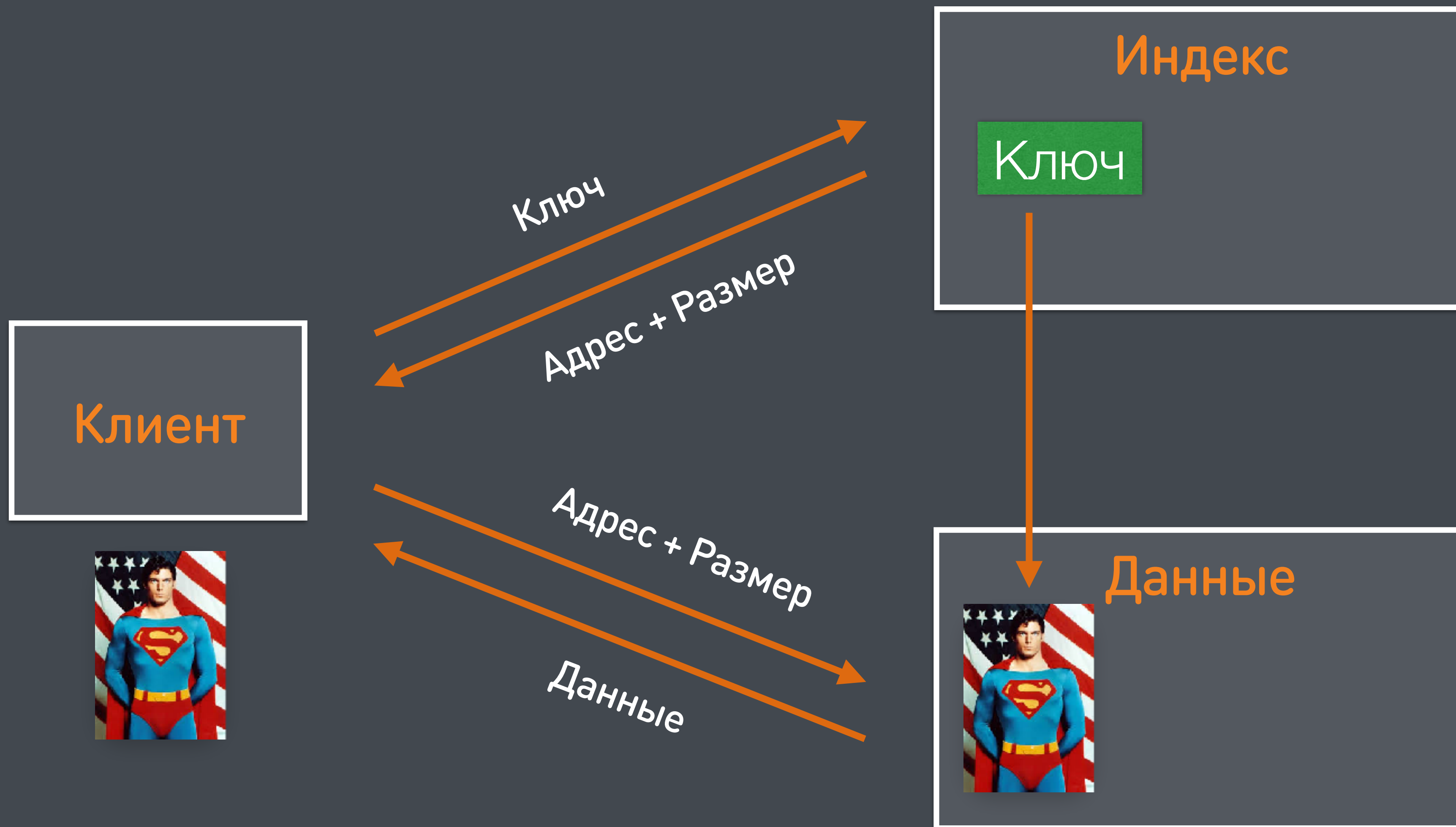
Диск

Смещение

Сервер

Адрес

52 Чтение данных



53 Индекс



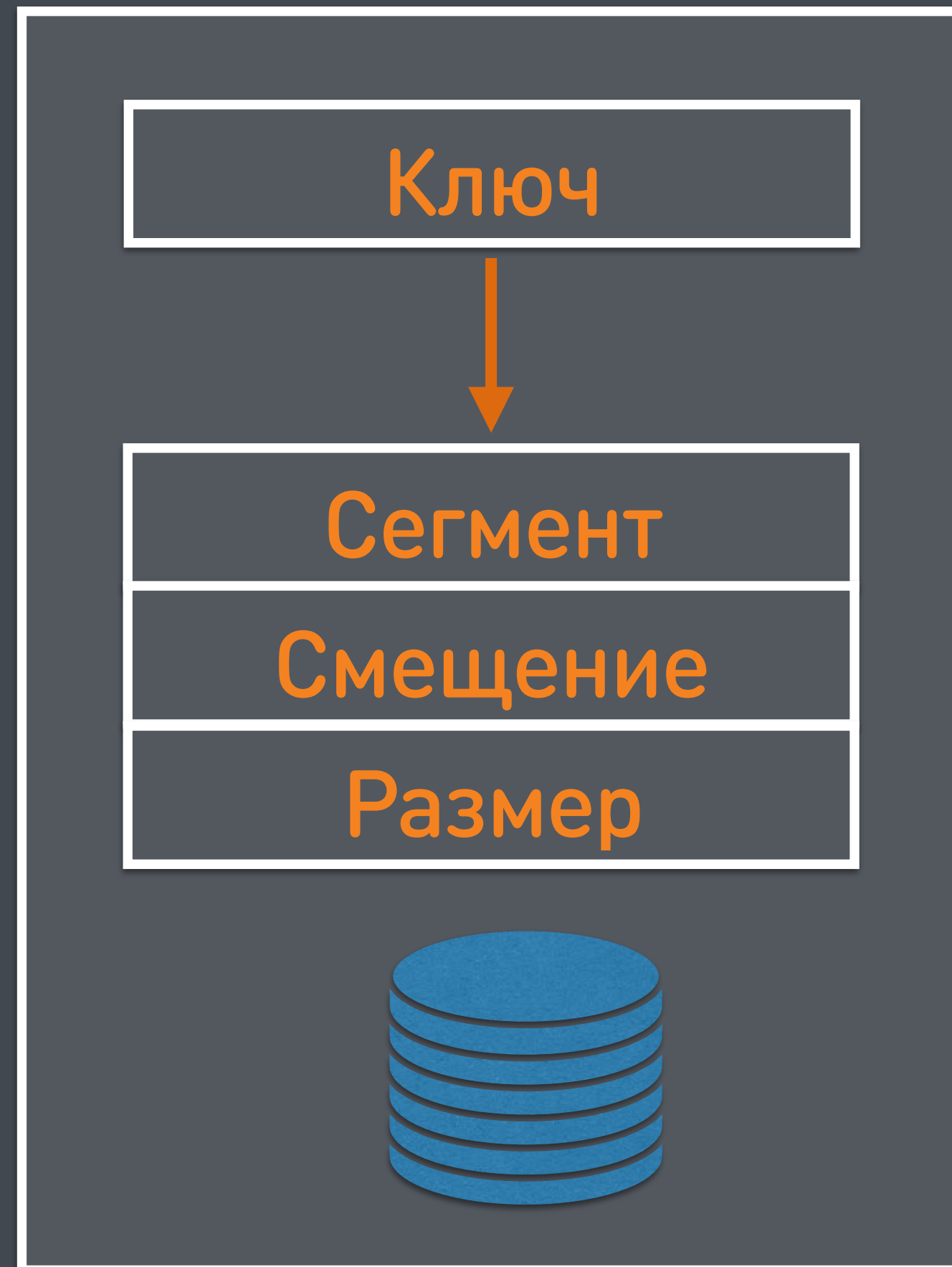
Cassandra 2.0 inside

RF = 3

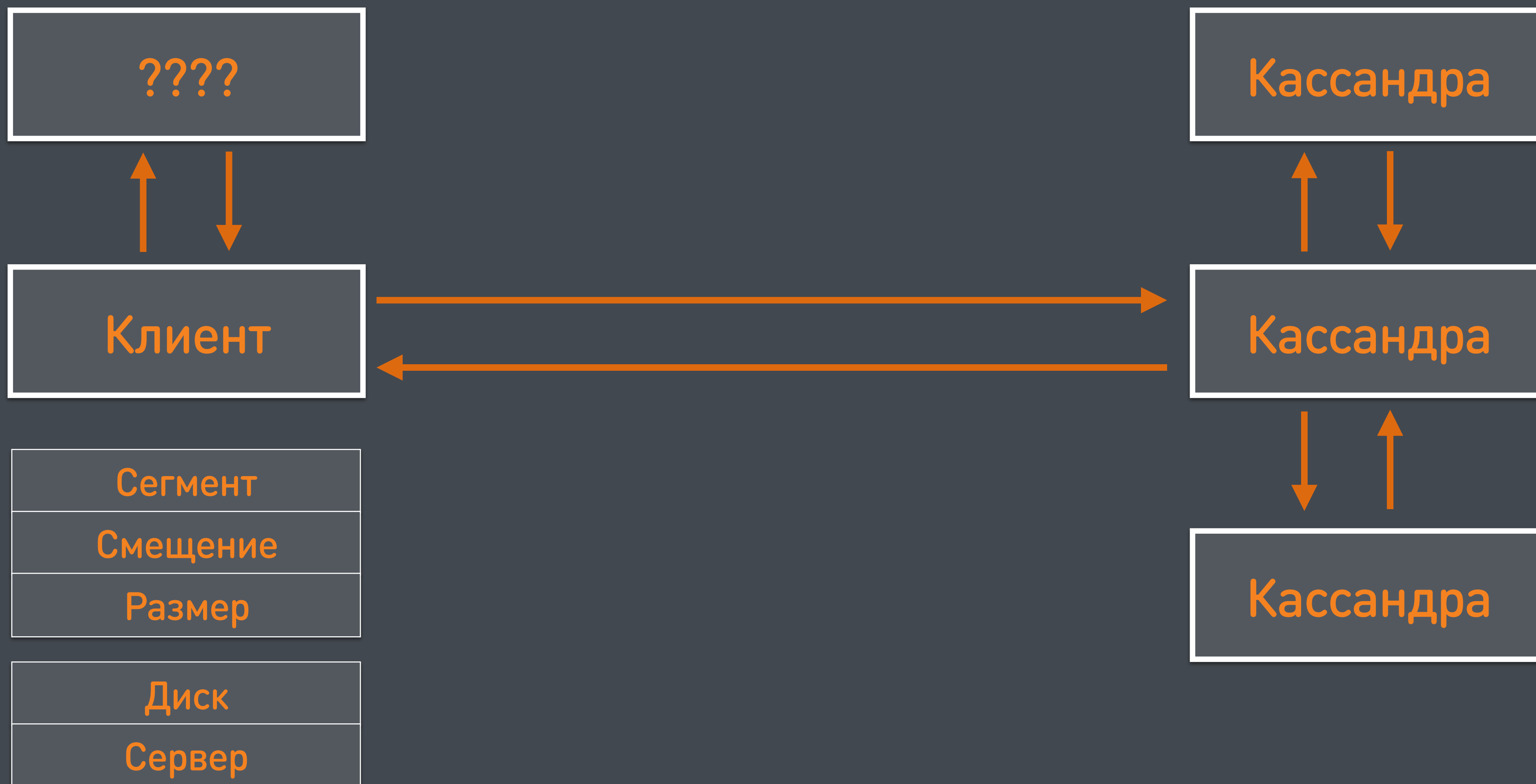
Распределение по ДЦ

Служебная информация

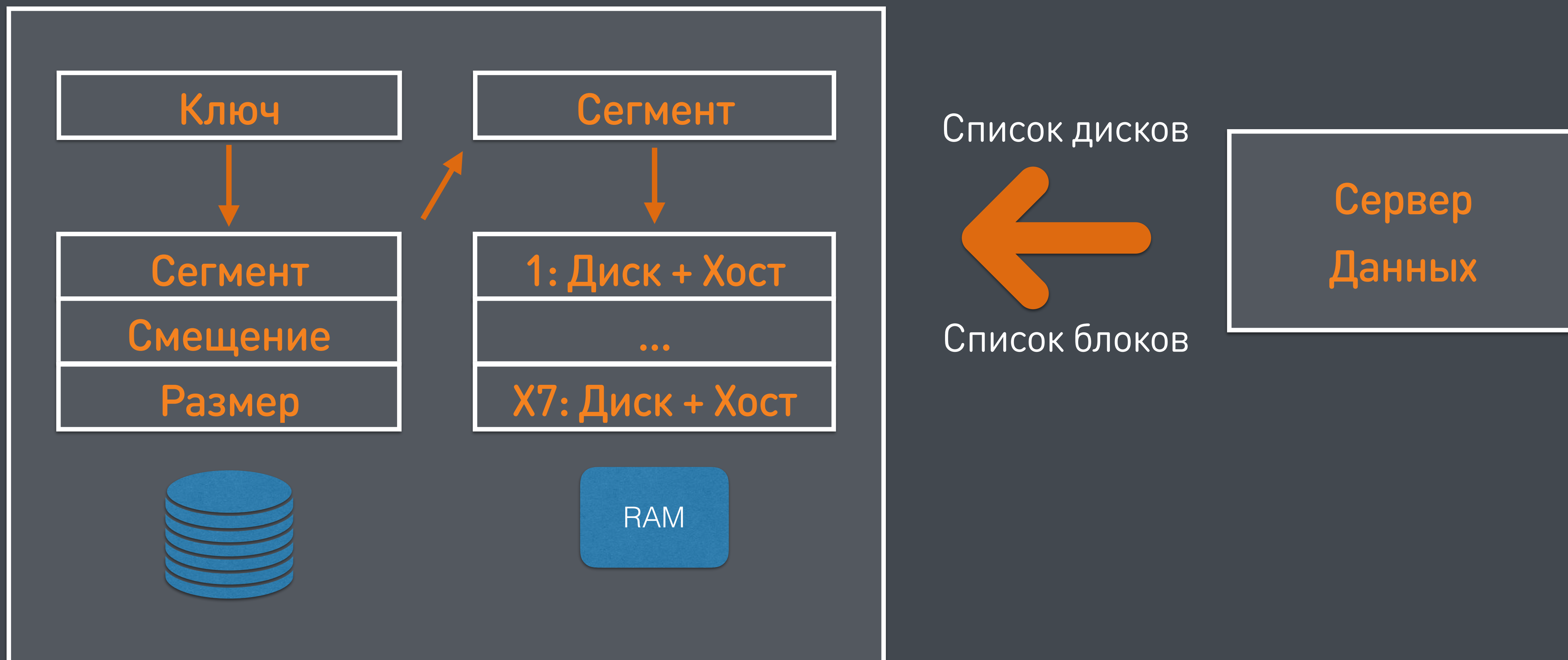
54 Данные в индексе



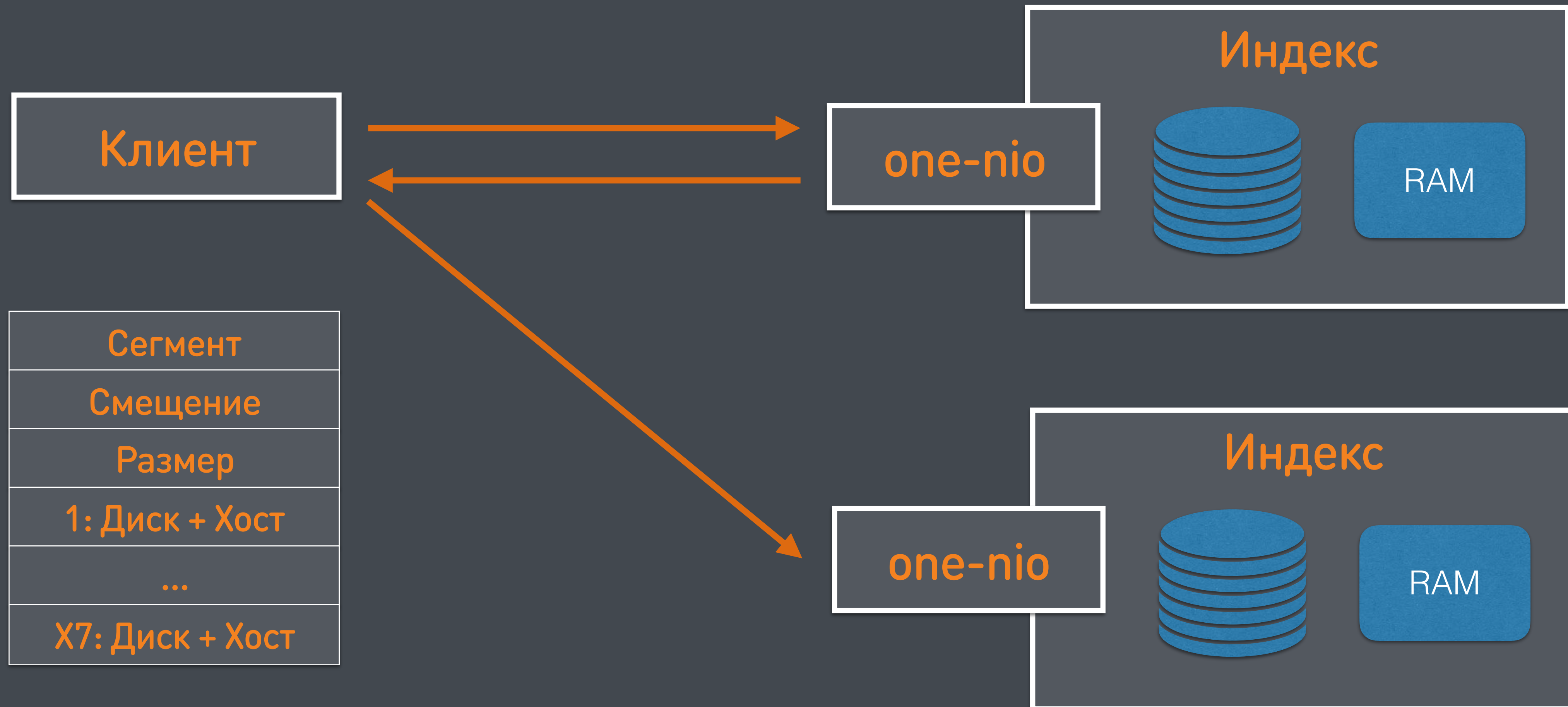
55 Чтение в Кассандре



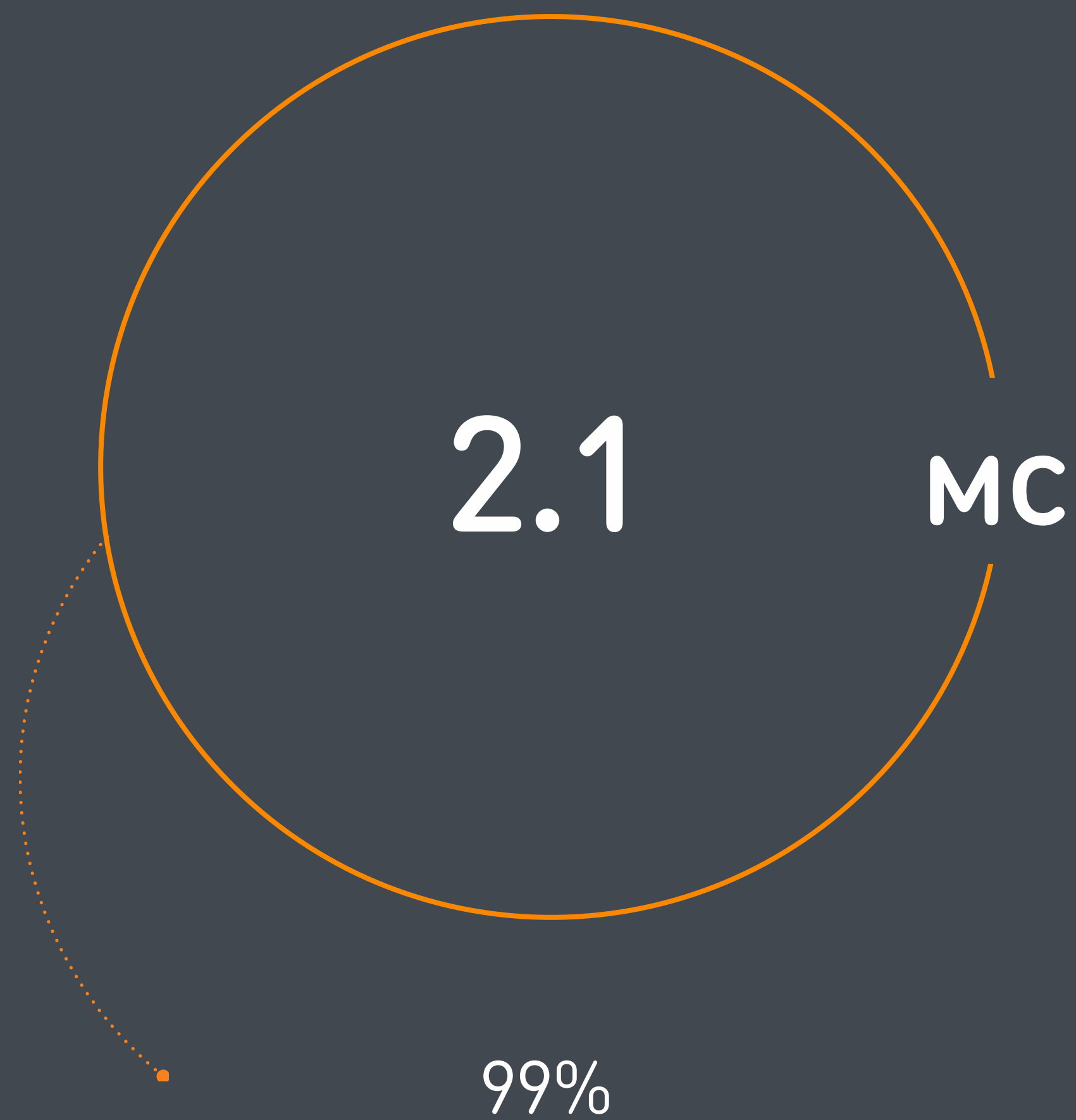
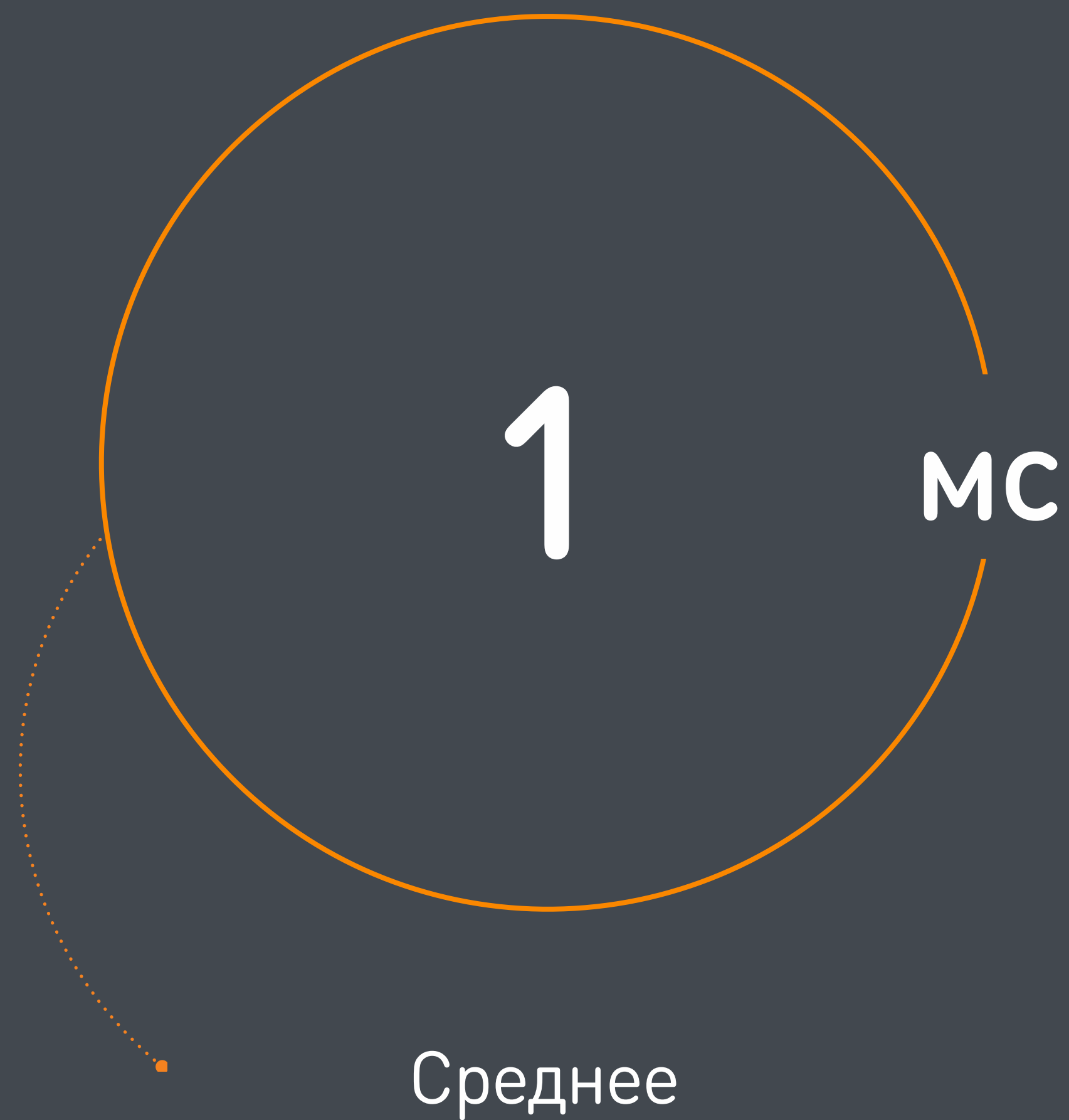
56 Устройство Индекса



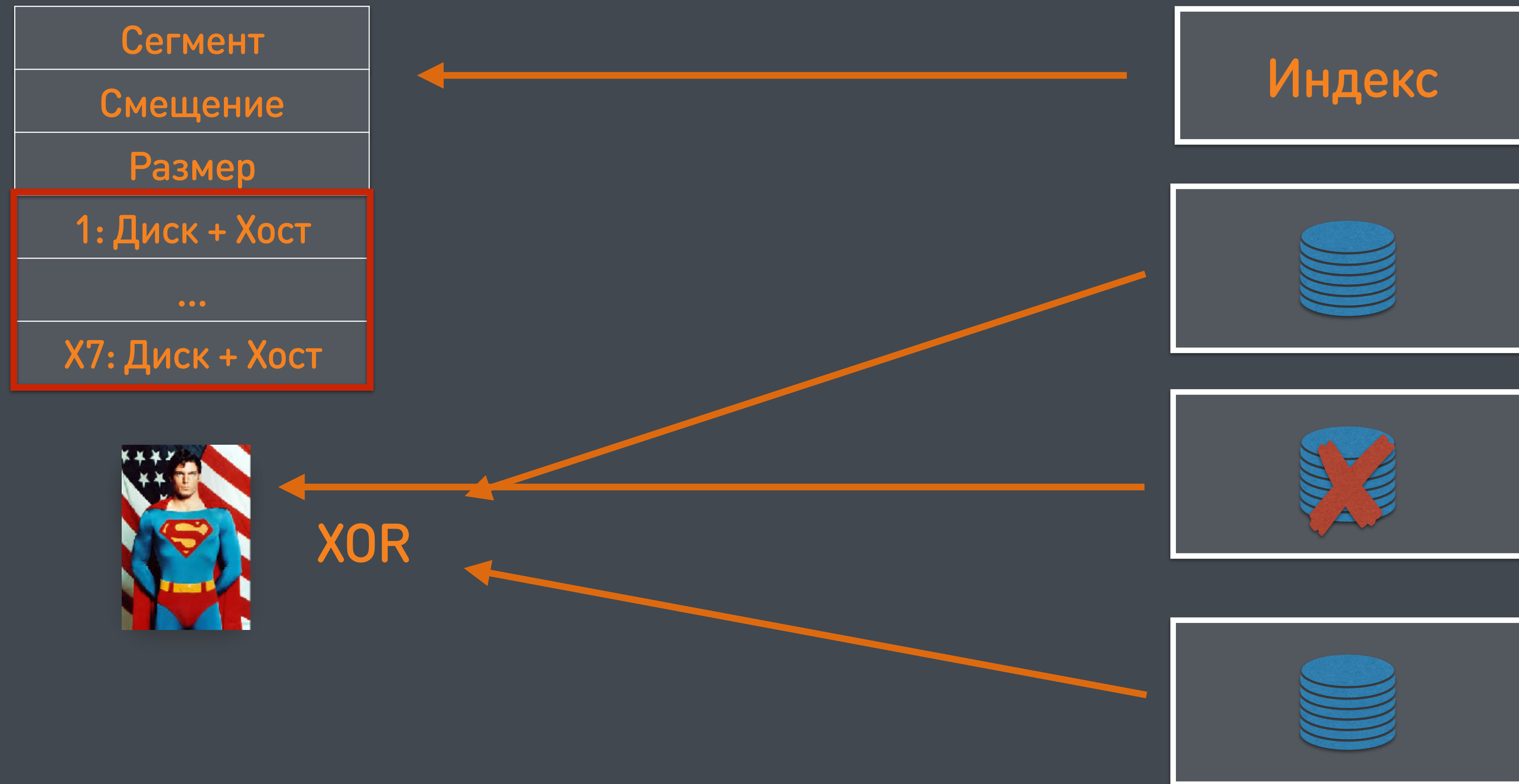
57 Спекулятивное чтение из индекса



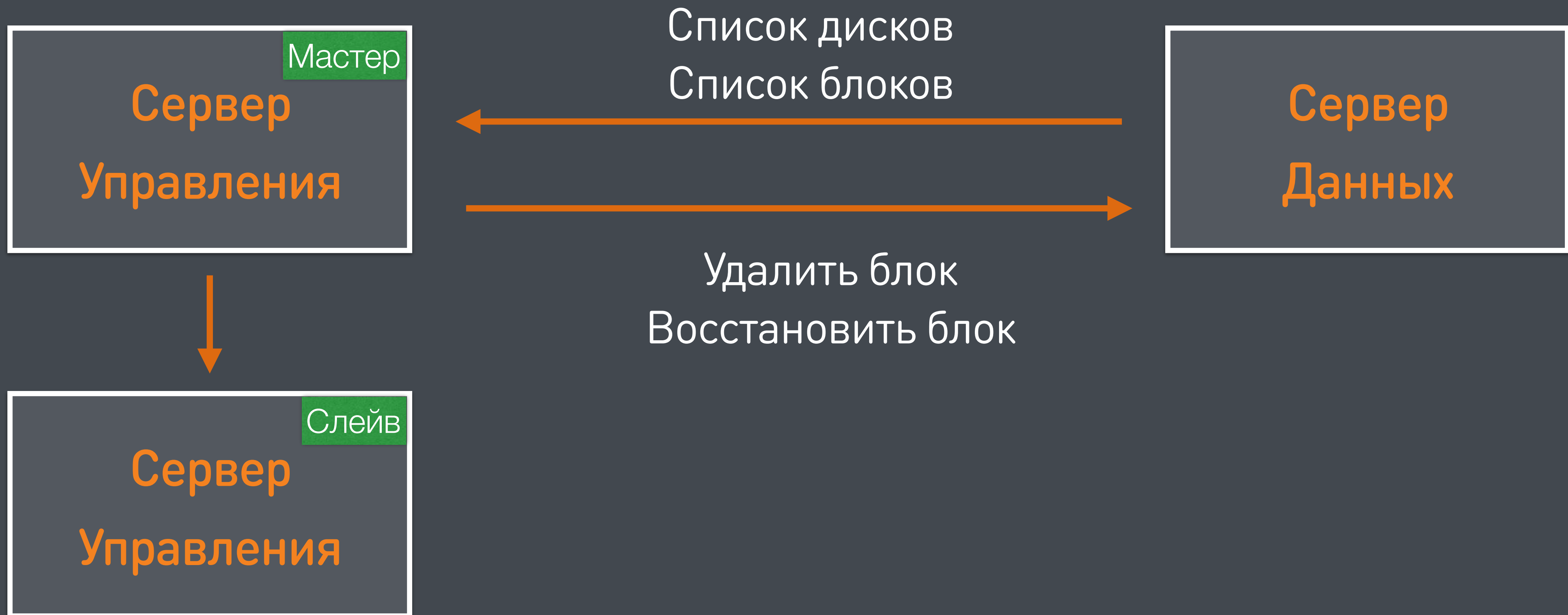
58 Время чтения из индекса



59 Восстановление данных на лету



60 Управление кластером



61 Мастер / слейв

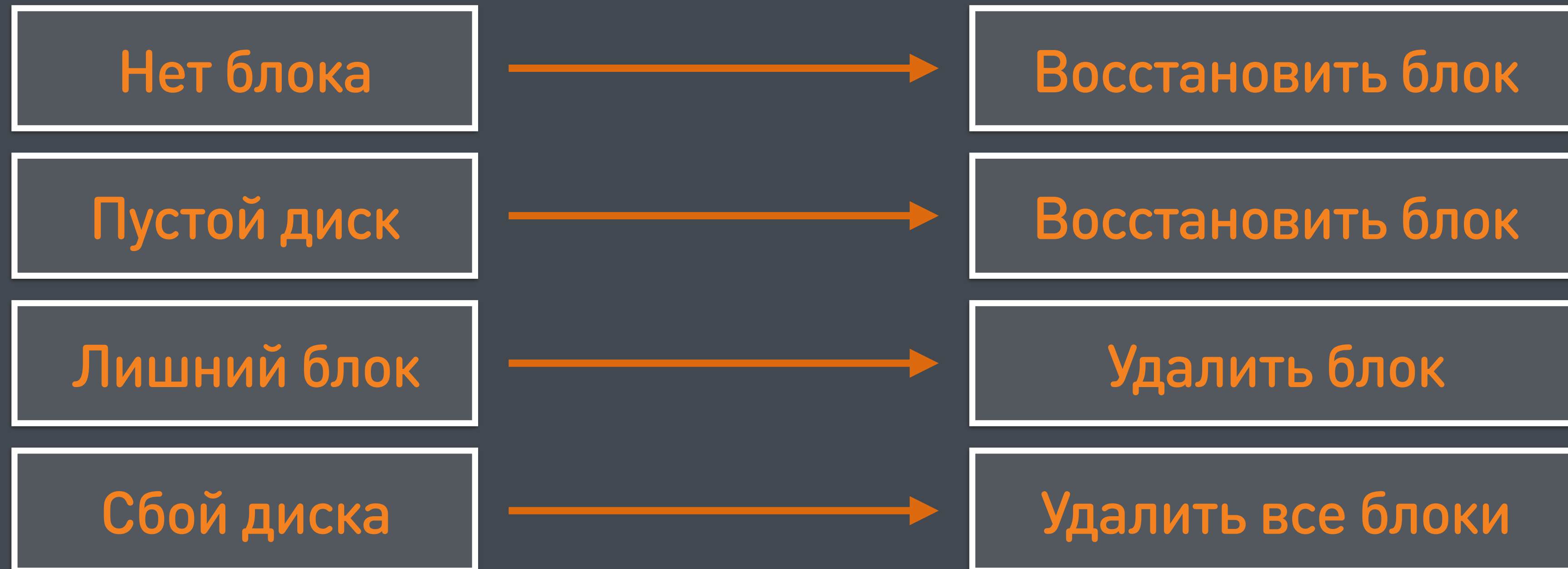


Проще принимать решения

Автоматический Leader Election

Потеря мастера не критична

62 Принятие решений





Результаты

Развлекательная
социальная сеть
Одноклассники



64 Что в OCS сегодня



Видео

Фото

Старше 140 дней

Старше 1.5 лет

80% данных

40% данных

60% запросов

40% запросов

65 Оборудование



240 серверов данных

8400 дисков

12 индексов

150 заменённых дисков

4 расширения кластера

66 Результаты



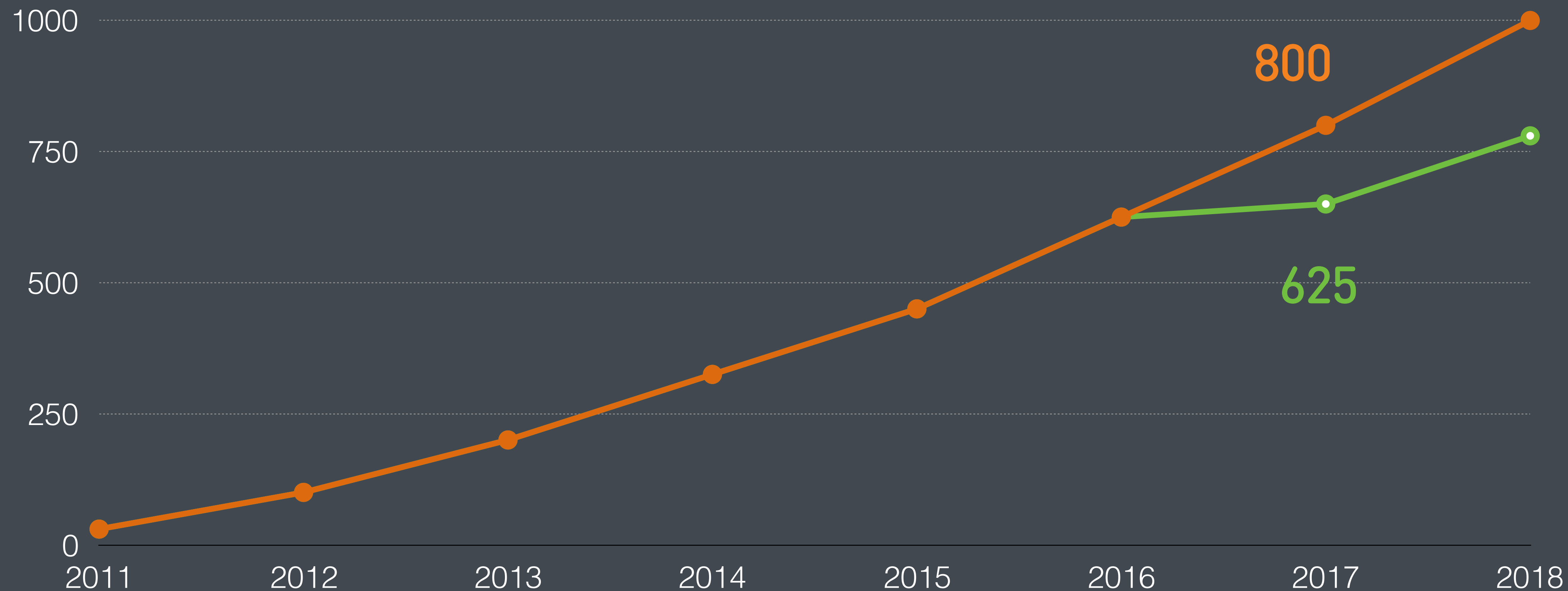
-19 ПБ

Место

99.9997 %

Доступность

67 Количество серверов в OBS + OCS (Фото + Видео)





ОДНОКЛАССНИКИ