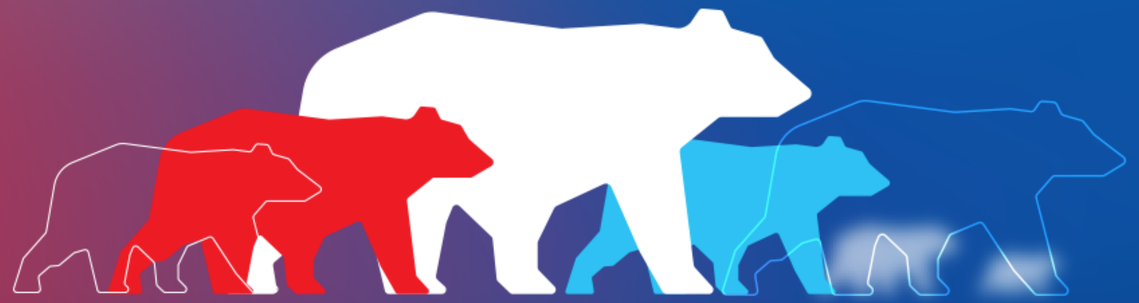
Яндекс Облако

Yandex Database — как мы обеспечиваем отказоустойчивость

Владислав Кузнецов, разработчик

HighLoad++ Siberia 2019

Профессиональная конференция для разработчиков высоконагруженных систем



План рассказа

01 | Общий обзор устройства Yandex Database

02 | Необходимость отказоустойчивости

03 | Tablet, единица отказоустойчивости

04 | Distributed storage, обеспечение отказоустойчивости

Yandex Database (YDB)

Yandex Database — это геораспределённая база данных, предоставляющая:

- Надёжное хранение данных с автоматической репликацией
- Механизм распределённых ACID-транзакций со строгой консистентностью
- Высокую пропускную способность при малом времени отклика
- Автоматическое восстановление после сбоев
- Декларативный язык запросов YQL (диалект SQL)
- Горизонтальную масштабируемость до тысяч нод

Наиболее частые отказы железа

Диски

- Время наработки на отказ (MTTF) одного диска — 1.4 млн часов или 160 лет
- Время наработки на отказ >= 1 диска из ~1000 серверов по 4 диска в каждом — 180 часов или 8 дней

Серверы

- Отказы железа
- Отключение серверов/стоек для обслуживания
- Выход из строя целых стоек (ToR switch, питание)
- Отключение дата-цетров (учения, питание, сеть)

Модель акторов

- Существует актор-система, внутри которой находятся акторы
- Каждый актор — легковесная State Machine
- Акторы общаются при помощи сообщений
- При обработке принятого сообщения актор может
  - Отправить новое сообщение
  - Создать новых акторов
  - Изменить свой State
  - Умереть

Модель акторов

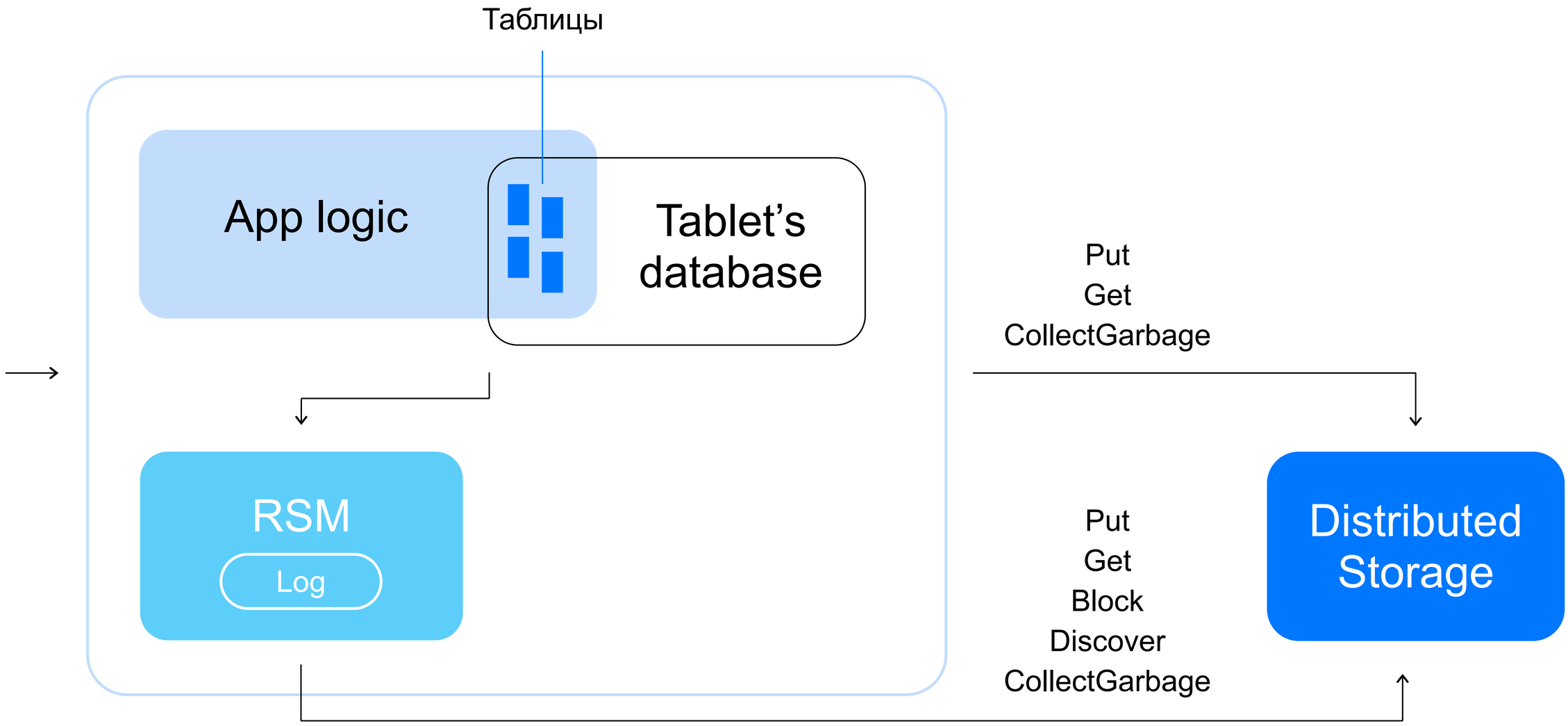
- YDB — распределённая актор-система
- Акторы могут быть запущены на любом сервере кластера
- Сообщения могут быть отправлены любому актору, в том числе на других серверах
- Актора можно быстро убить на одном сервере и поднять на другом (для балансировки или при сбоях сервера)

YDB Tablet

- Является актором, но имеет персистентное состояние
- Хранит все состояние и данные в Distributed Storage
- Единица отказоустойчивости системы
- Всегда работает только один экземпляр
- Поднимается на любой ноде кластера
- Набор небольших таблиц
- На кластере ~100к tablet

HighLoad++ Siberia 2019

YDB Tablet



HighLoad++ Siberia 2019

Tablet’s database

- У каждой tablet есть собственная база
- Все изменения в таблицах транзакционны и надёжно сохраняются в Distributed Storage
- Схема и все данные базы хранятся в Distributed Storage
- Работа с базой через код на C++

HighLoad++ Siberia 2019

Tablet — универсальный строительный блок системы

С использованием tablet построены разные логические сущности:

- Пользовательские таблицы
- Очереди
- Time-series хранилище данных мониторинга
- Системные tablet, необходимые для работы YDB

HighLoad++ Siberia 2019

Менеджмент tablet

- Существуют tablet двух категорий:
  - Системные (зарождаются самостоятельно)
  - Управляемые (контролируются системной tablet)
- Hive — системная tablet, управляющая другими tablet:
  - Знает обо всех управляемых tablet в системе
  - Запускает tablet, если необходимо (например, при отказе сервера)
  - Следит за статусом всех управляемых tablet
  - Балансирует tablet по серверам (по CPU, памяти)

HighLoad++ Siberia 2019

Архитектура
YDB Distributed Storage

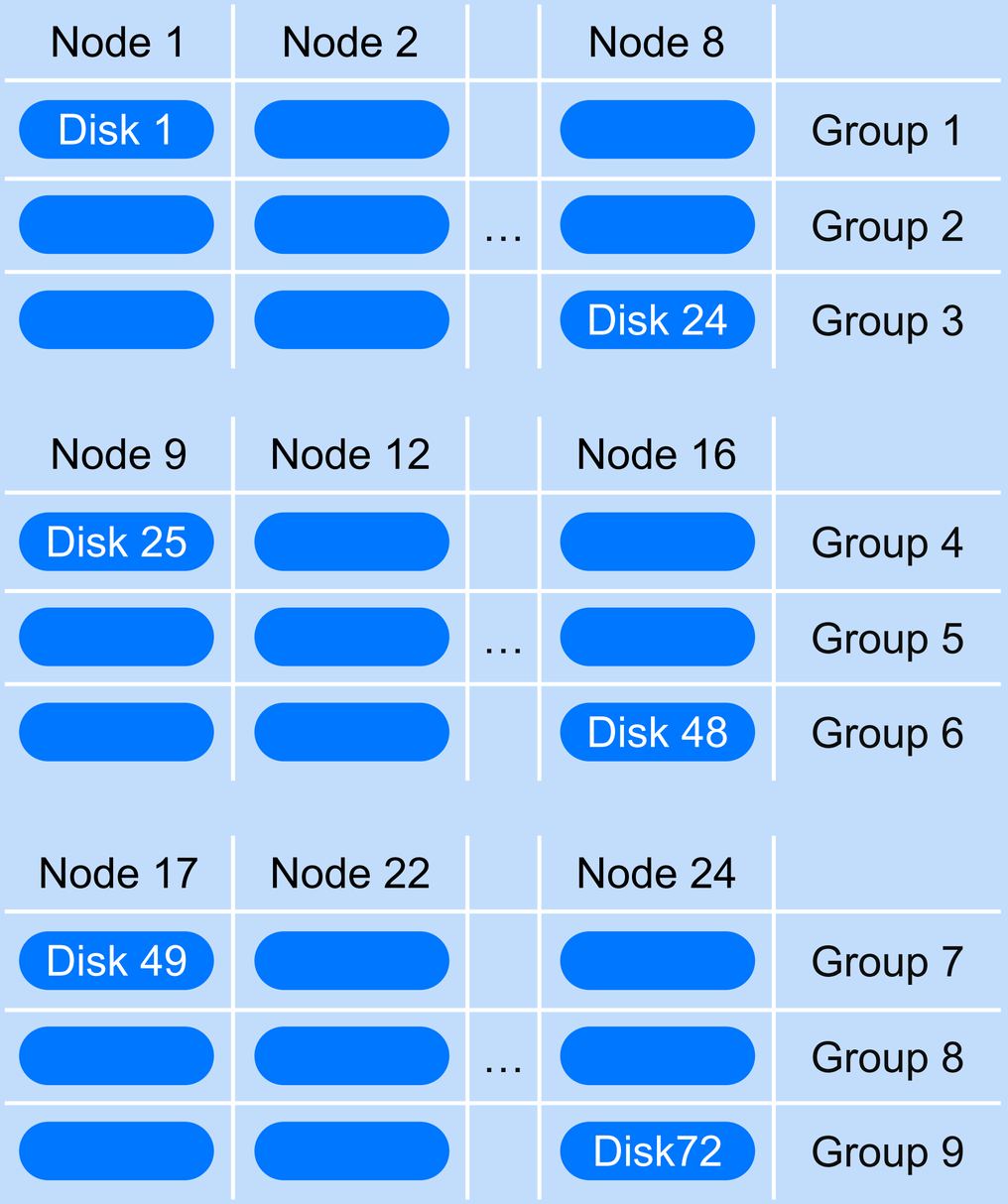
Distributed Storage

1. Создан для реализации отказоустойчивой работы Tablet
2. Хранит иммутабельные блобы произвольного размера (от 1B до 10MB)
3. Гарантии при записи блоба
   - Запись завершается успешно, только если блоб реплицирован и полностью записан
   - Гарантирует, что после успешной записи блоб не будет потерян
   - Если при записи невозможно гарантировать, что блоб был надёжно записан, возвращается ошибка
4. Обеспечивает настраиваемую избыточность хранения данных
   - 3dc mirror — георасределённая, реплики пишутся в 3 дата-центра
   - Erasure 4+2 — однодатацентровая, данные кодируются и пишутся в 6 частях. Для восстановления достаточно любых 4 частей.

HighLoad++ Siberia 2019

Distributed Storage

1. Распределённое хранилище
2. Состоит из набора групп
3. Каждая группа — непересекающийся набор дисков

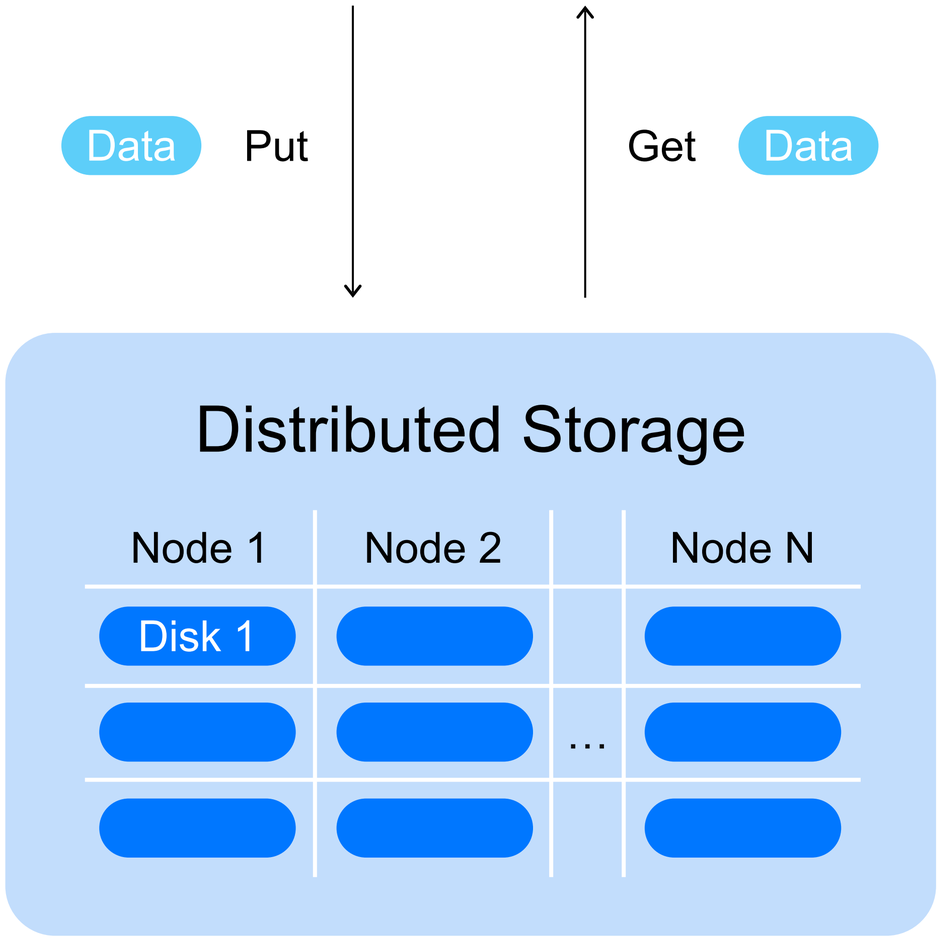


Distributed Storage

BlobID — идентификатор блоба, имеющий специальную структуру

> **Сообщения, обрабатываемые DSProxy**

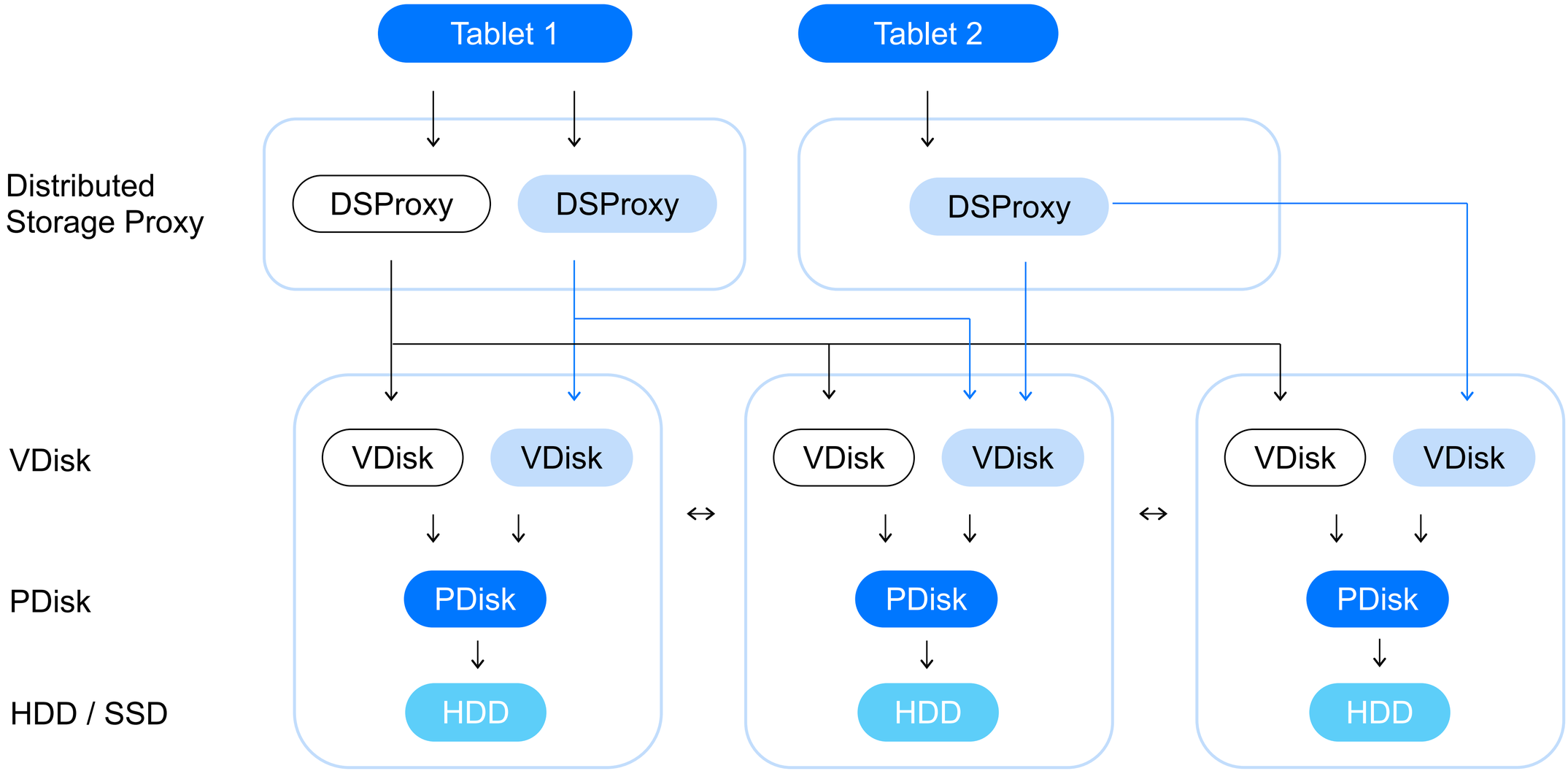
1. Put (blobID, data)
2. Get (blobID, offset, size)
3. Block
4. Discover
5. CollectGarbage



BlobID

- Уникальный идентификатор данных, записанных в Distributed Storage
- В каждом blobID зафиксирован владелец данного блоба
- BlobID упорядочены, поэтому можно найти самые поздние
- Хранит «номер» изменения состояния таблетки

Устройство Distributed Storage



HighLoad++ Siberia 2019

PDisk

- Работает с физическими блочными устройствами напрямую
- Оптимизирован для быстрой записи множества мелких блобов
- Каждая запись пишется надёжно, минуя все кэши
- Имеет планировщик нагрузки, способный честно разделять полосу реального устройства между разными VDisk

VDisk — участник DS-группы

- Можно рассматривать как key-value-хранилище частей блобов
- Взаимодействие peer-to-peer всех VDisk’ов внутри группы
- Автоматически восстанавливается после потери данных (репликация)
- Синхронизируют метаинформацию о блокировках, барьерах сборки мусора

DSProxy

- Реализует логику Put, Get, Block, Discover, CollectGarbage
- Создается по требованию на любой ноде кластера
- Одна DSProxy пишет только на одну группу
- Знает, на каких серверах живут диски группы
- На все операции DSProxy никаких взаимодействий внутри группы

HighLoad++ Siberia 2019

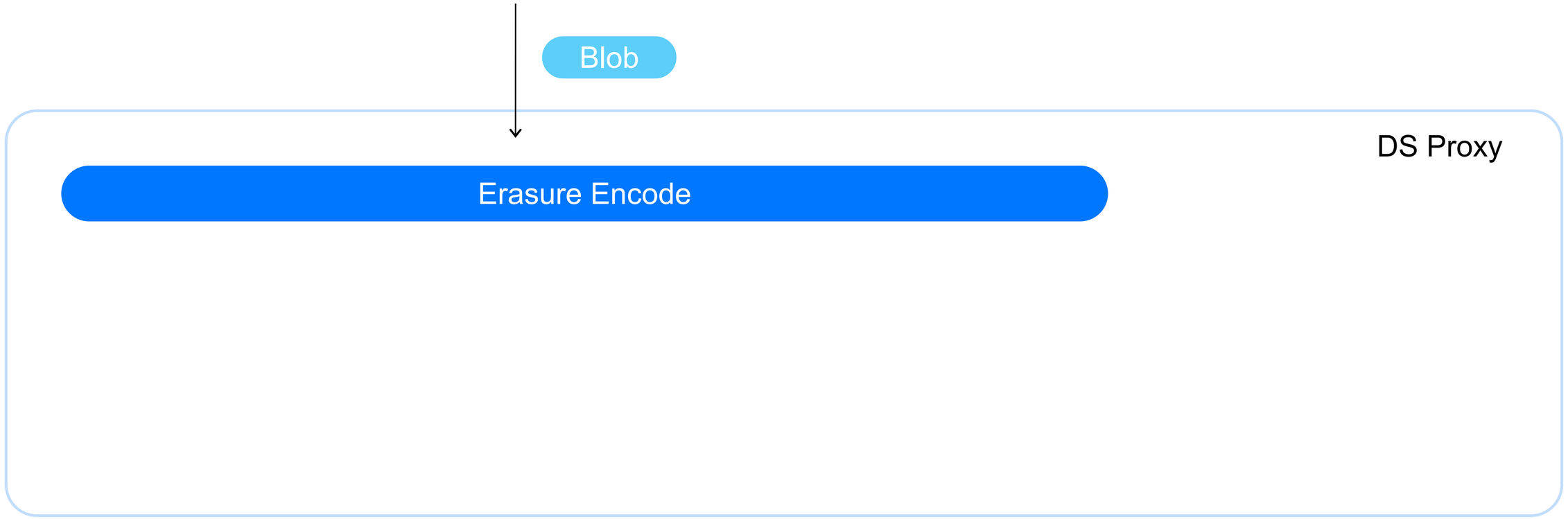
Устройство
YDB Distributed Storage Put

DSProxy Put

- Надёжно записывает блоб
- Для схемы erasure 4+2 нужно записать 6 частей
- Части должны быть записаны на разные диски
- Запись считается успешной, только когда записаны все 6 частей
- Допускается отказ двух дисков из группы
- 8 дисков — минимальный размер группы для erasure 4+2

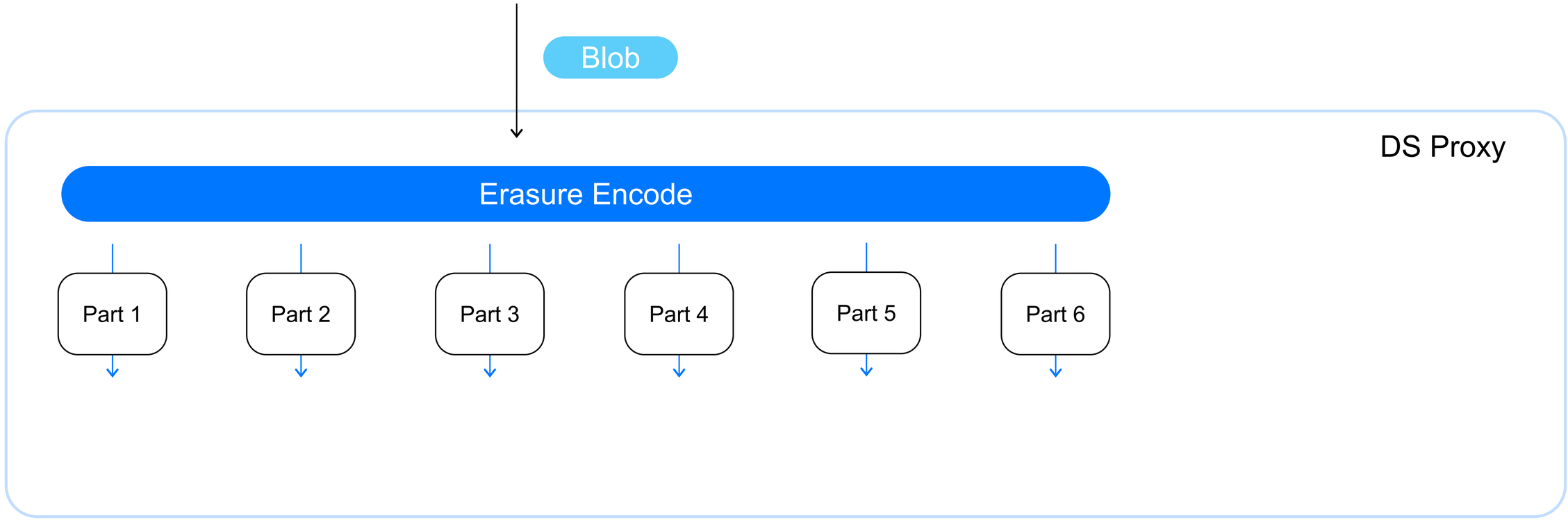
HighLoad++ Siberia 2019

DSProxy Put



HighLoad++ Siberia 2019

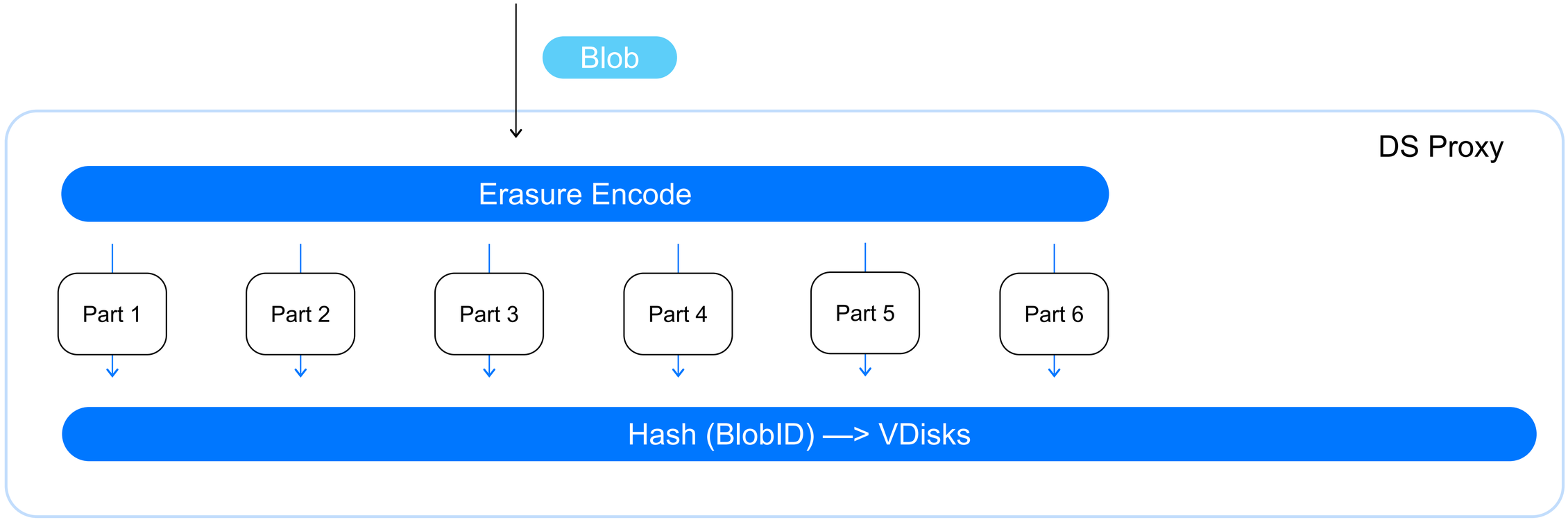
DSProxy Put



VDisk 1 | VDisk 2 | VDisk 3 | VDisk 4 | VDisk 5 | VDisk 6 | VDisk 7 | VDisk 8

HighLoad++ Siberia 2019

DSProxy Put



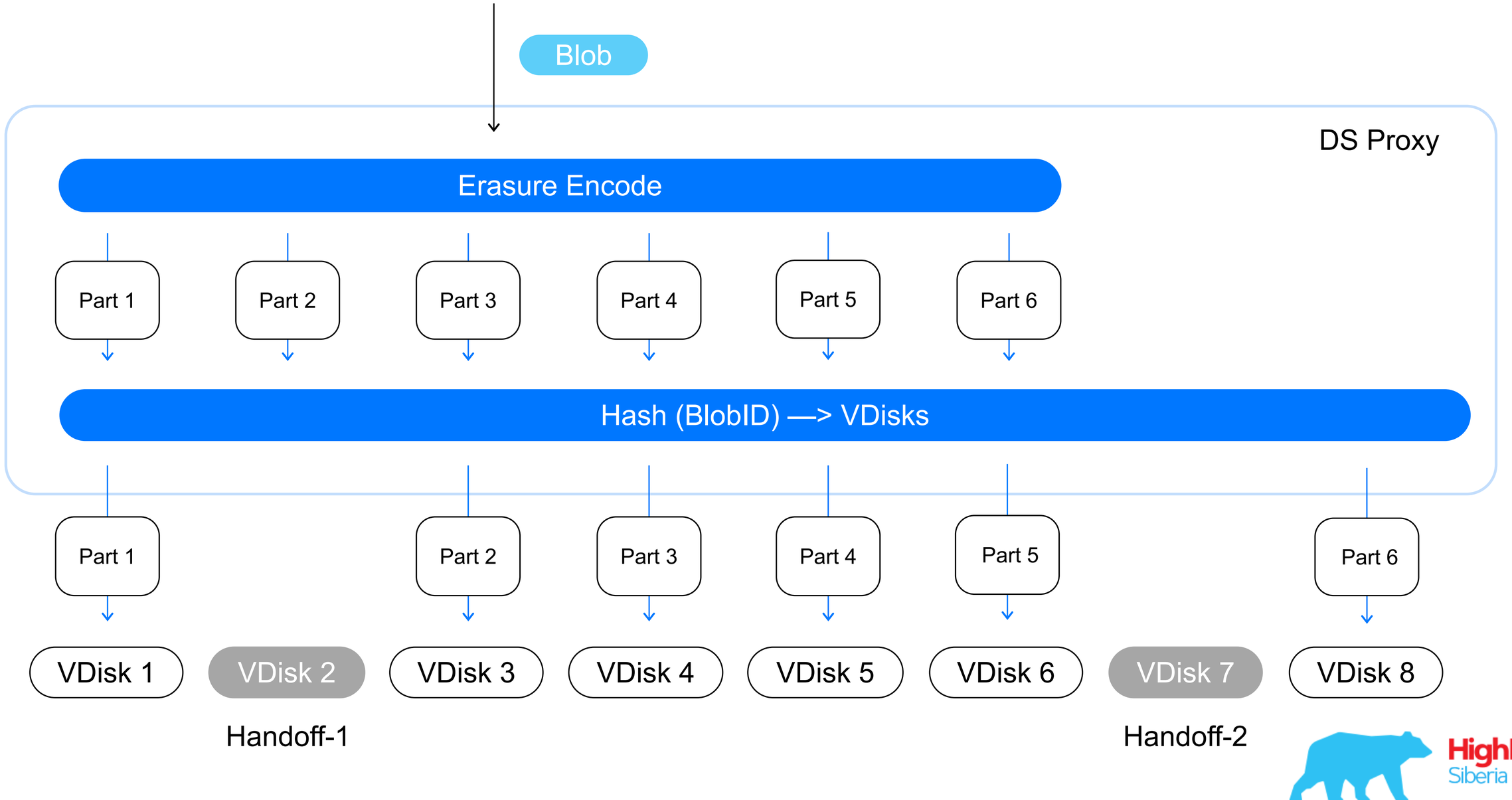
VDisk 1 | VDisk 2 | VDisk 3 | VDisk 4 | VDisk 5 | VDisk 6 | VDisk 7 | VDisk 8

Handoff-1

Handoff-2

HighLoad++ Siberia 2019

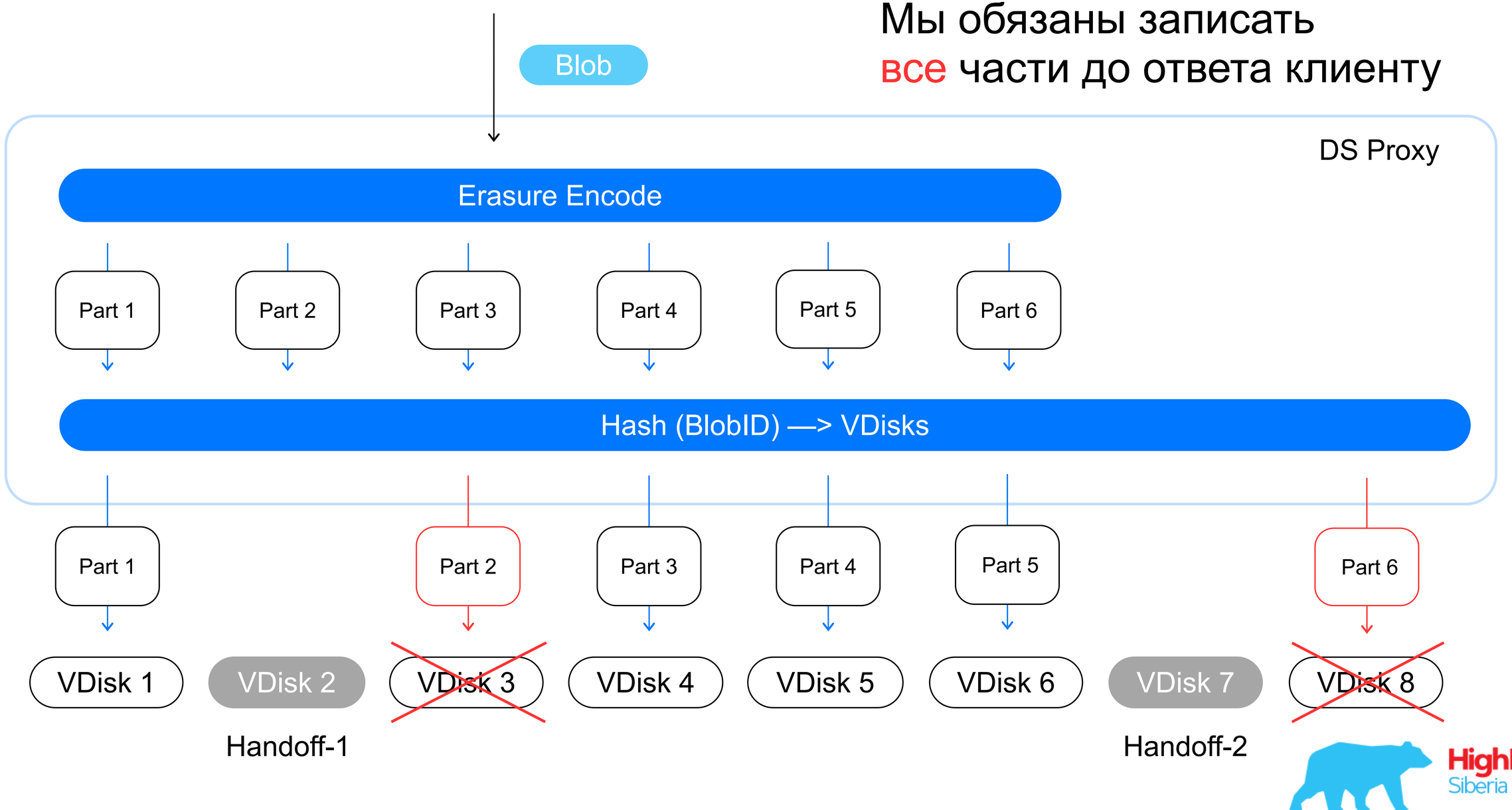
DSProxy Put



HighLoad++ Siberia 2019

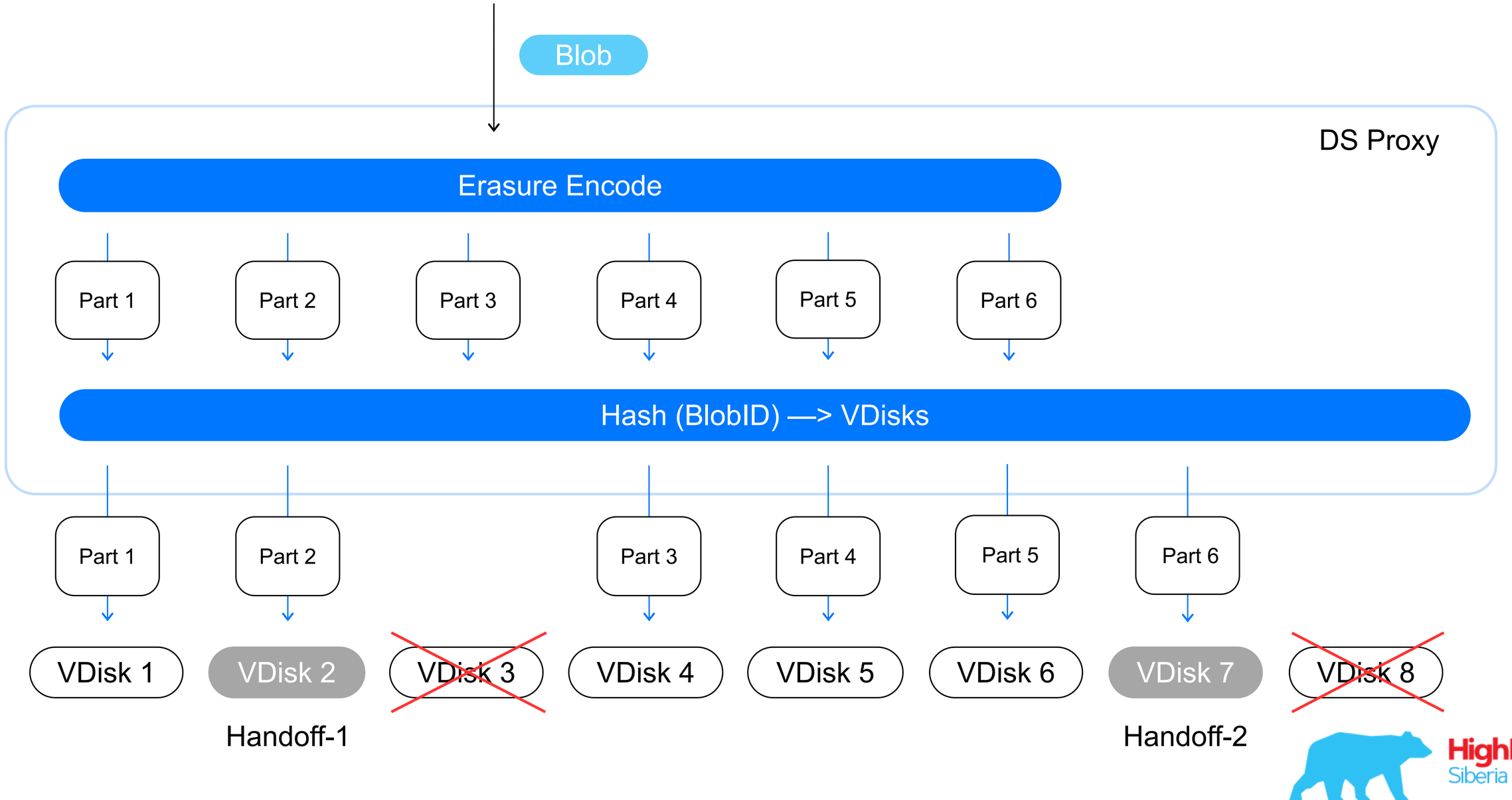
DSProxy Put

Мы обязаны записать
все части до ответа клиенту



HighLoad++ Siberia 2019

DSProxy Put



HighLoad++ Siberia 2019

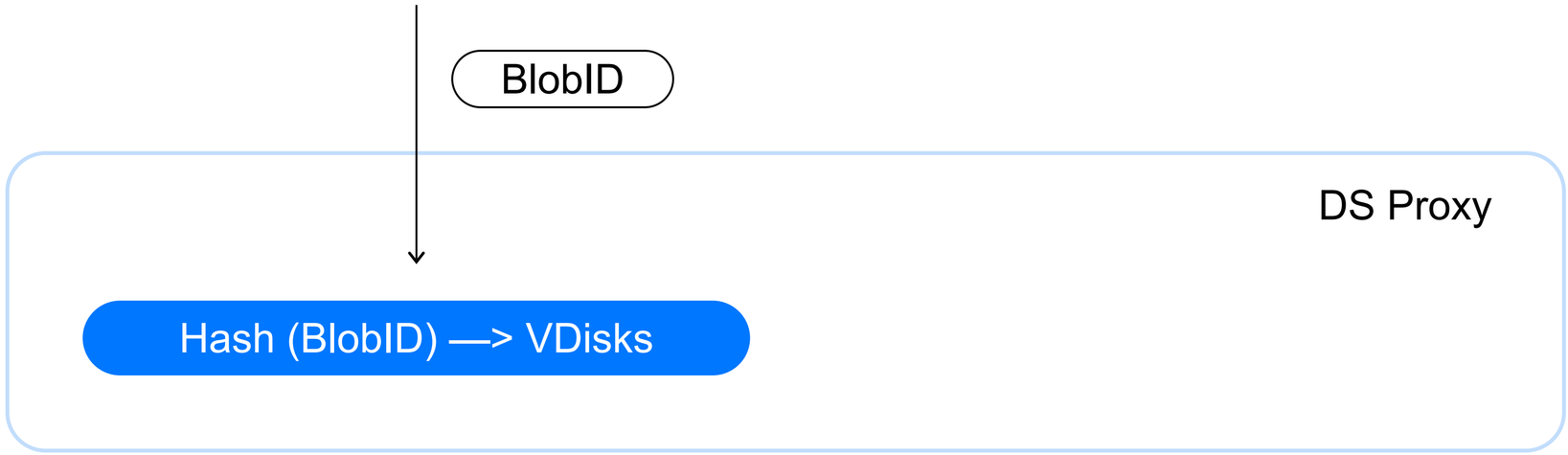
Устройство YDB Distributed Storage DSProxy Get

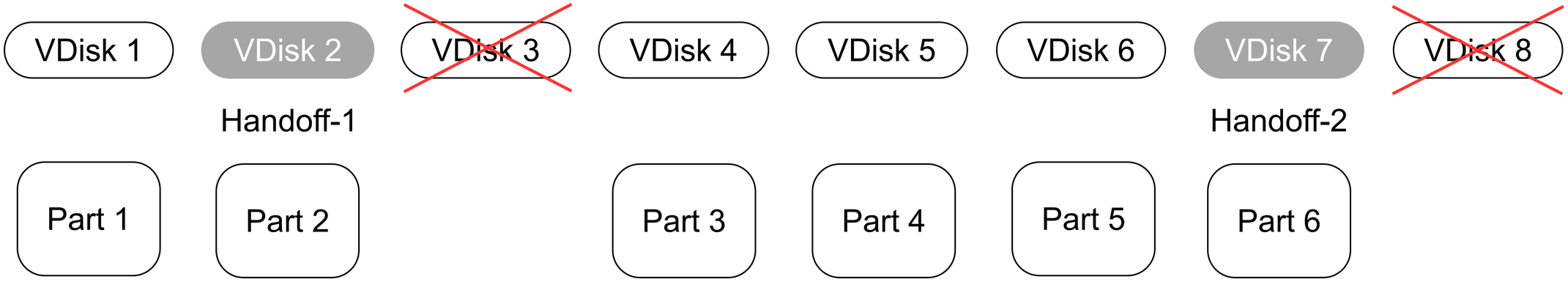
DSProxy Get

- Читает записанные блобы
- Некоторые части блоба могут быть потеряны
- Могут не работать 2 диска из группы
- Get знает, где искать все части, и читает их
- Может читать только 4 части с данными
- Может читать все 6 частей, ждать 4 ответа
  и после восстанавливать данные
  при помощи erasure

HighLoad++
Siberia 2019

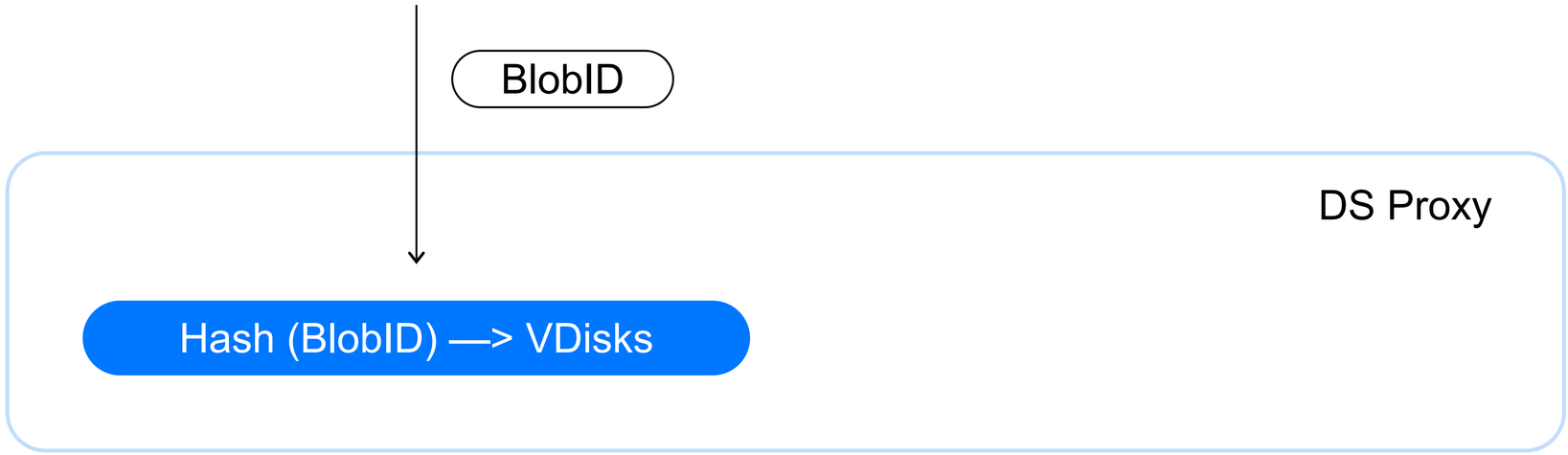
DSProxy Get

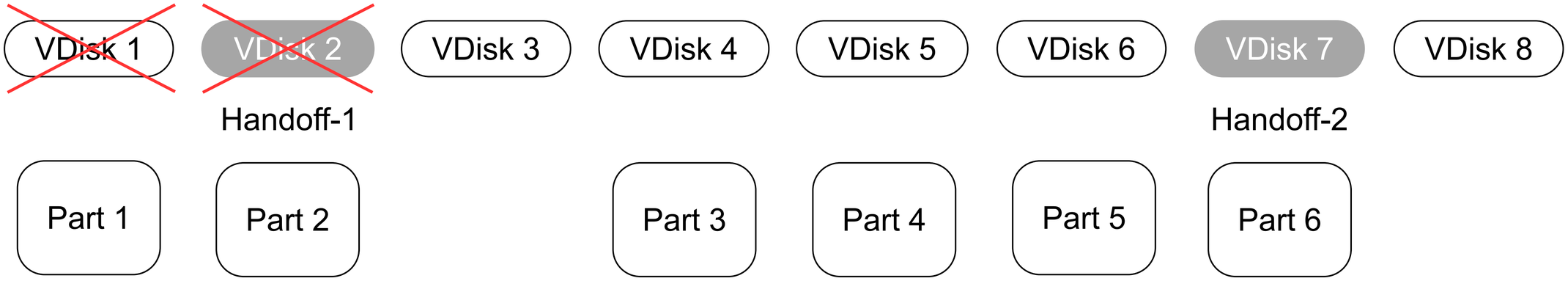




HighLoad++ Siberia 2019

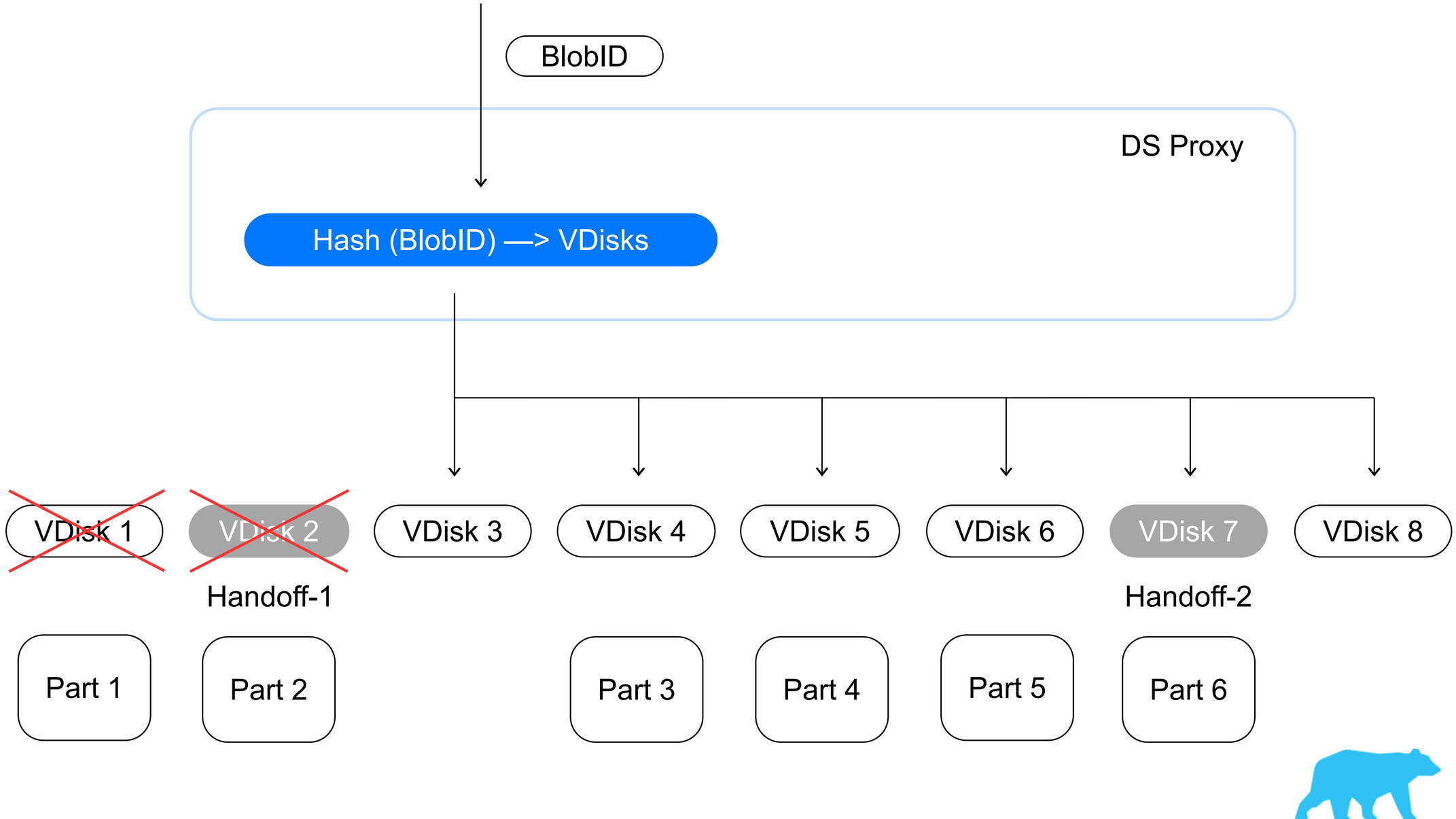
DSProxy Get





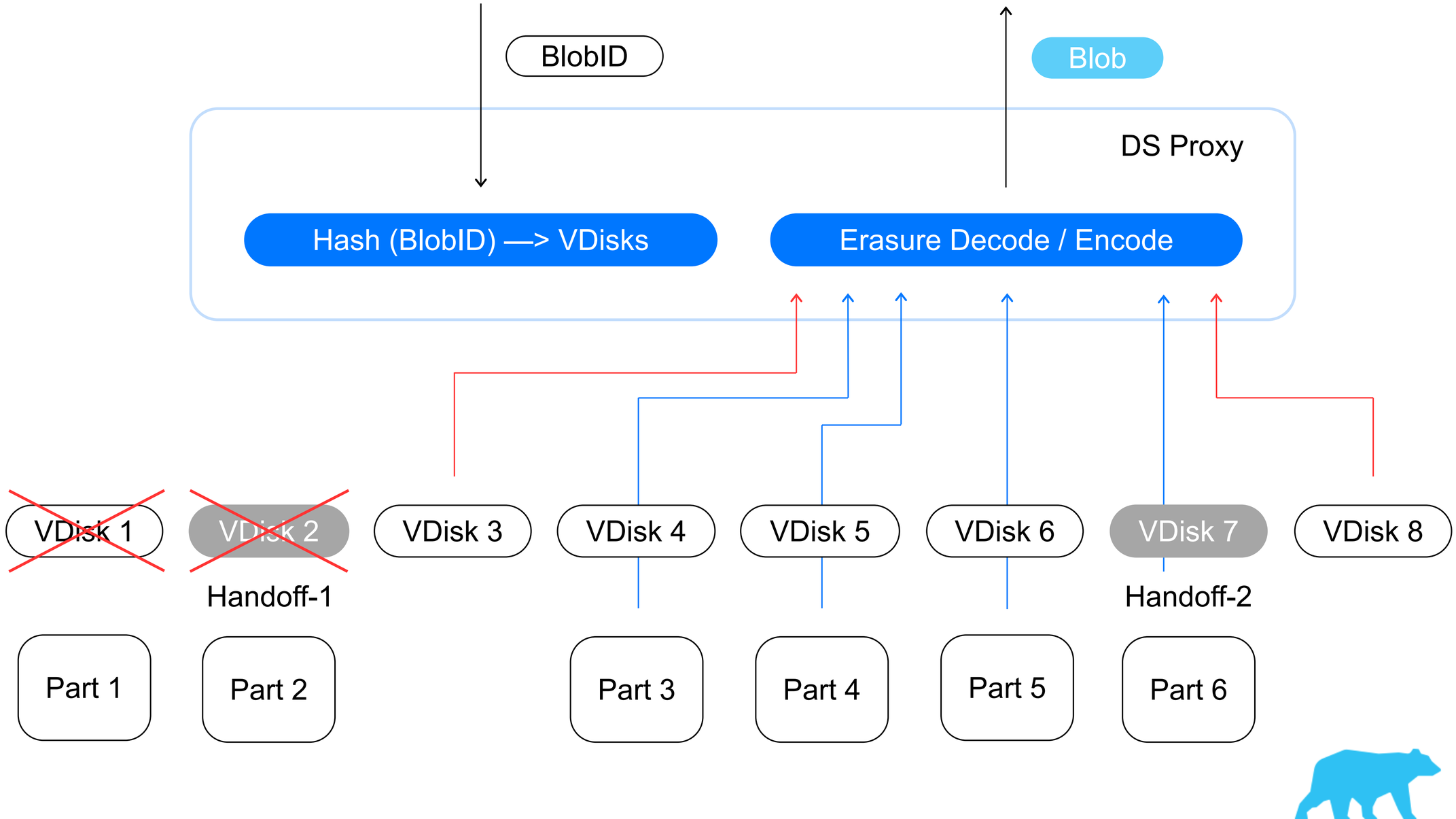
HighLoad++ Siberia 2019

DSProxy Get



HighLoad++ Siberia 2019

DSProxy Get



HighLoad++ Siberia 2019

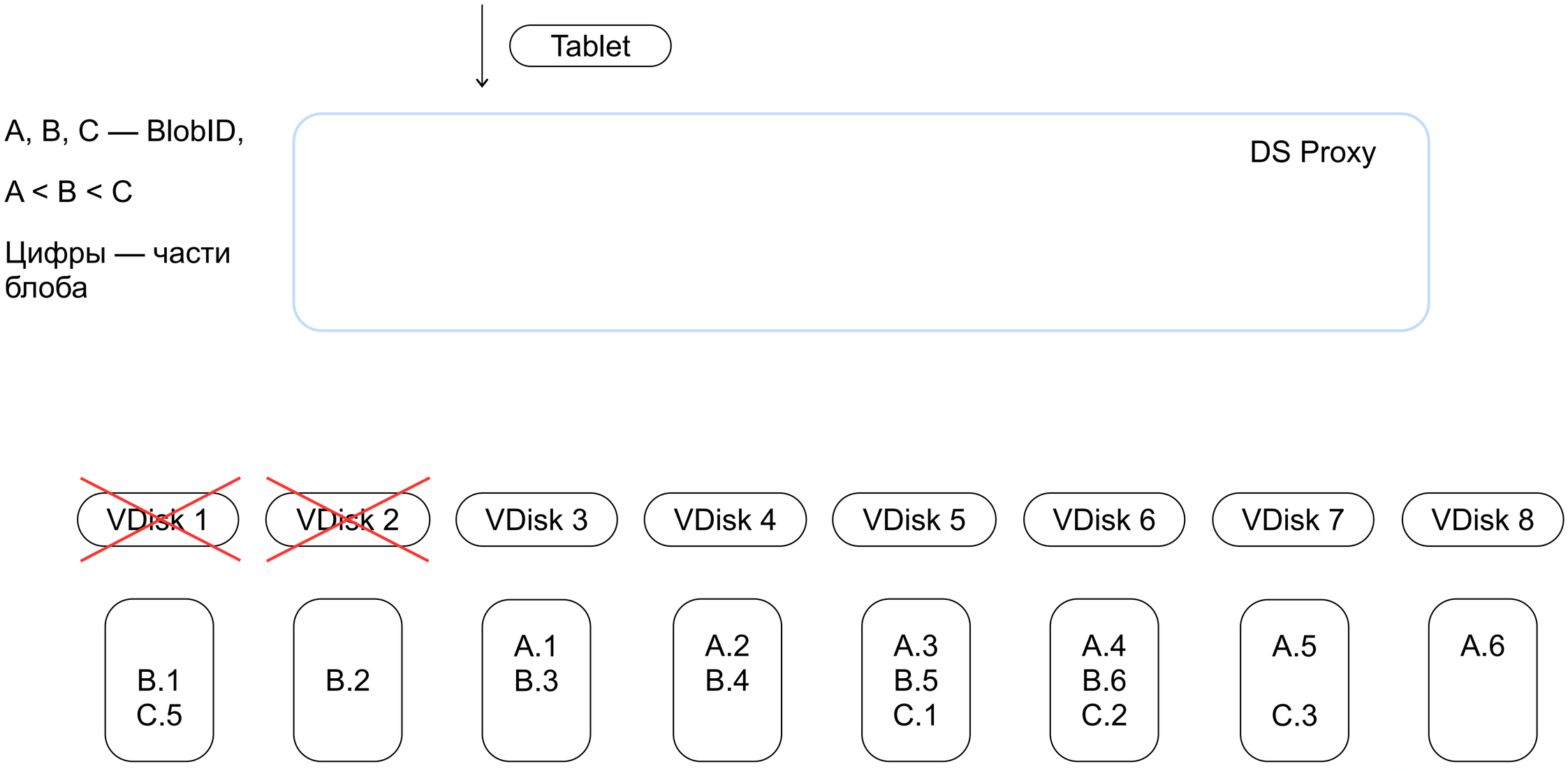
Устройство YDB Distributed Storage DSProxy Discover

DSProxy Discover

- Находит самый свежий записанный блоб
- Если блоб записан не во всех репликах, делает Get + Restore
- Для tablet информации в последнем записанном блобе достаточно для восстановления всего лога

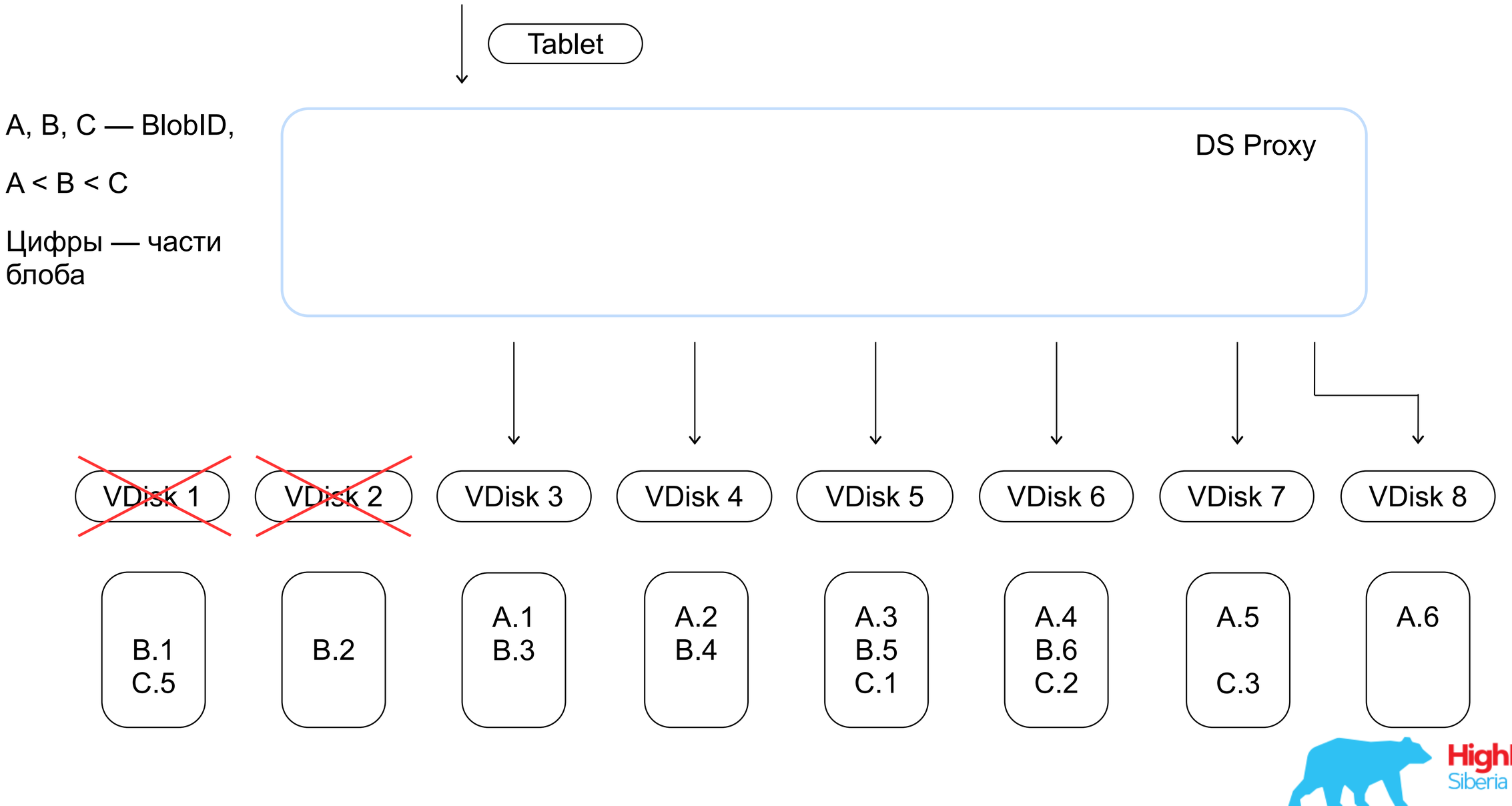
HighLoad++ Siberia 2019

DSProxy Discover



HighLoad++ Siberia 2019

DSProxy Discover



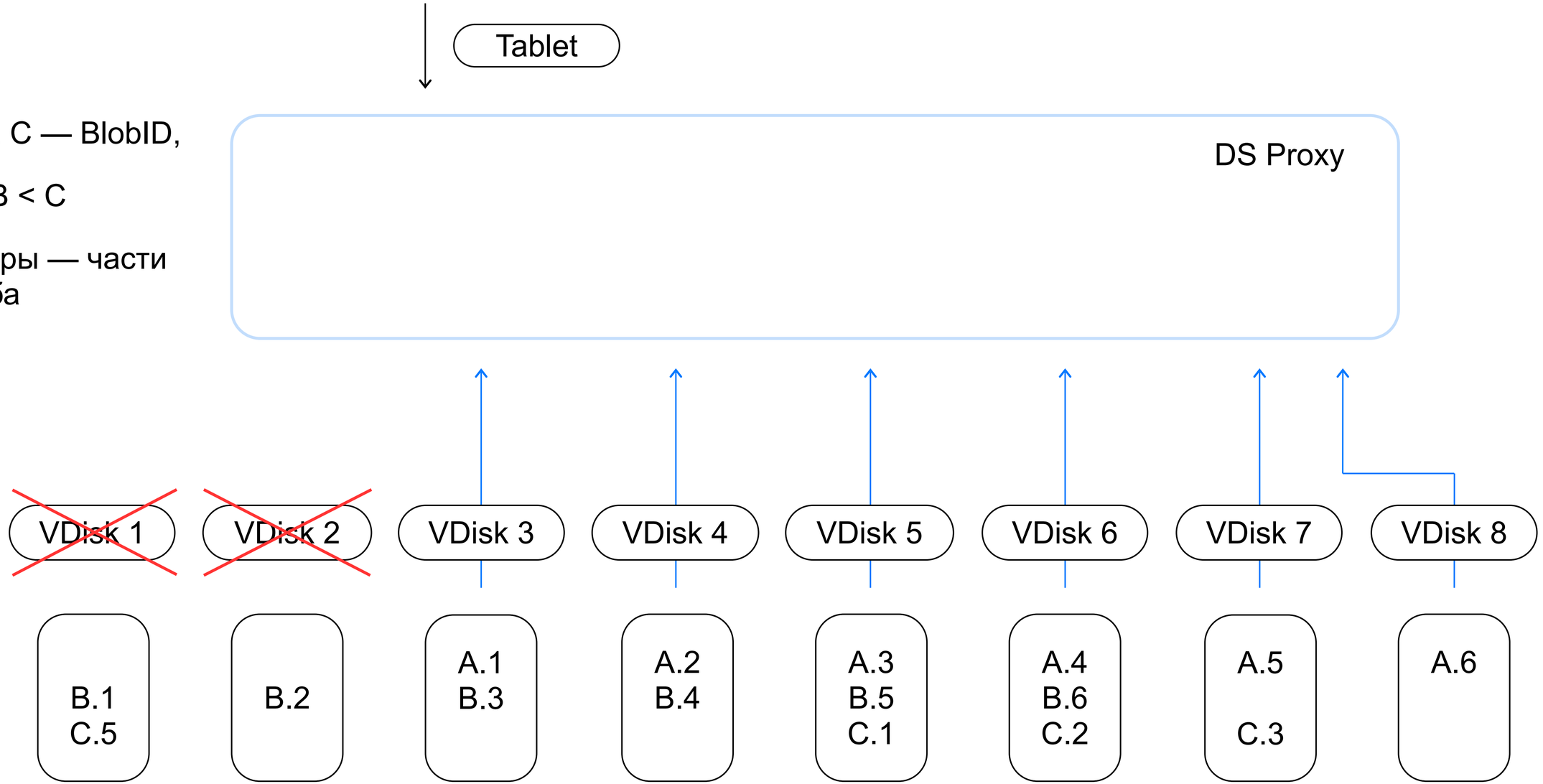
HighLoad++ Siberia 2019

DSProxy Discover

A, B, C — BlobID,

A < B < C

Цифры — части блоба

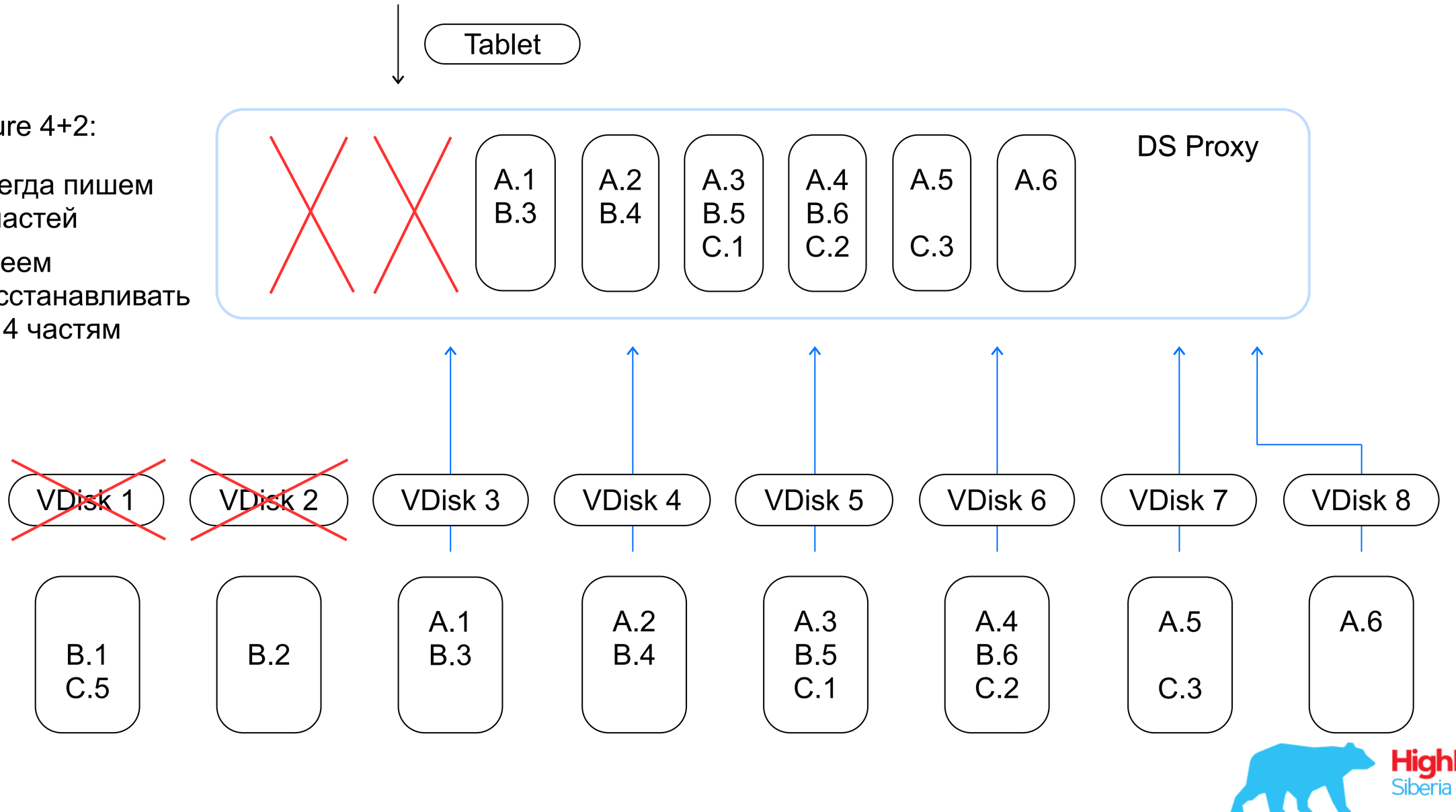


HighLoad++ Siberia 2019

DSProxy Discover

Erasure 4+2:

- Всегда пишем 6 частей
- Умеем восстанавливать по 4 частям

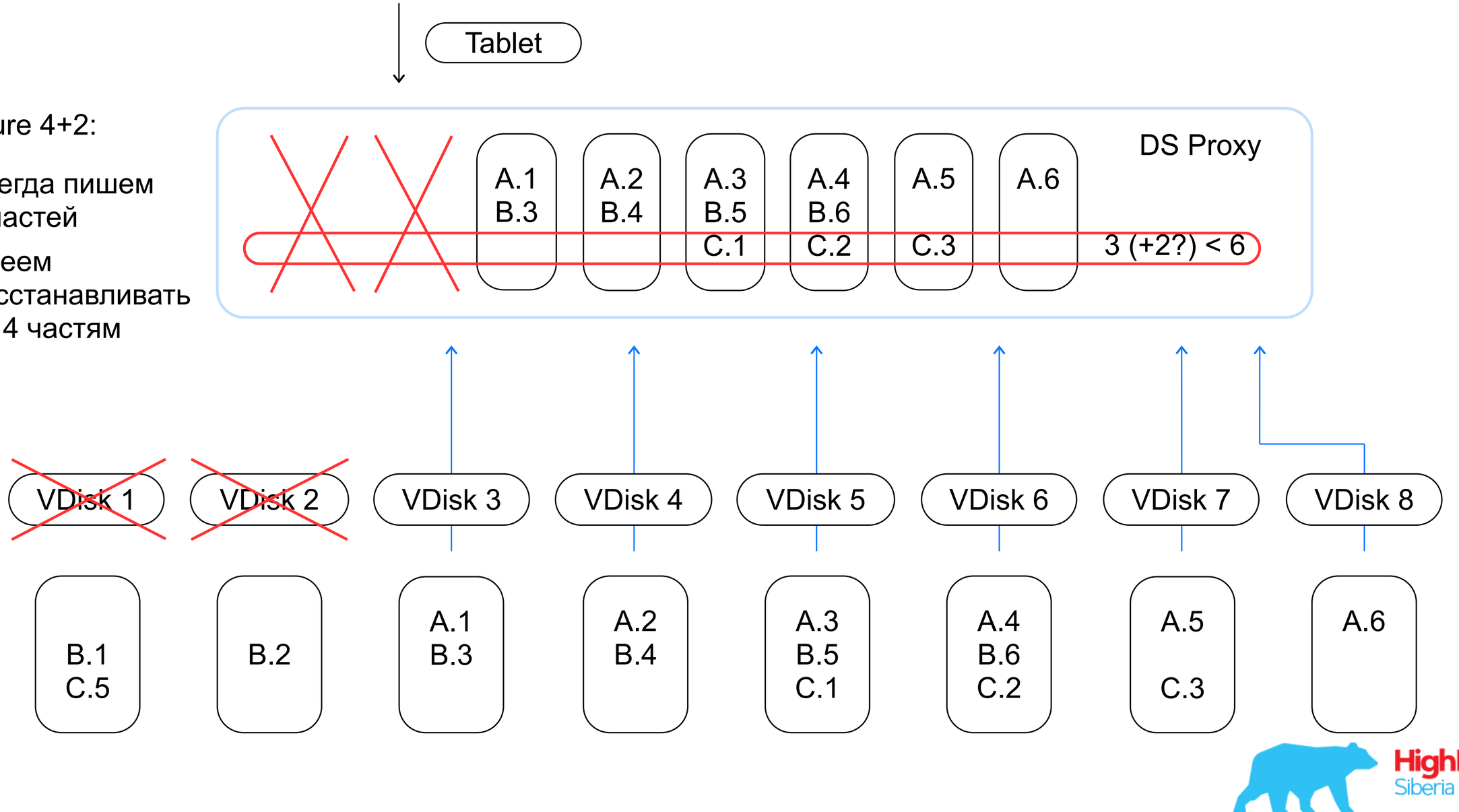


HighLoad++ Siberia 2019

DSProxy Discover

Erasure 4+2:

- Всегда пишем 6 частей
- Умеем восстанавливать по 4 частям

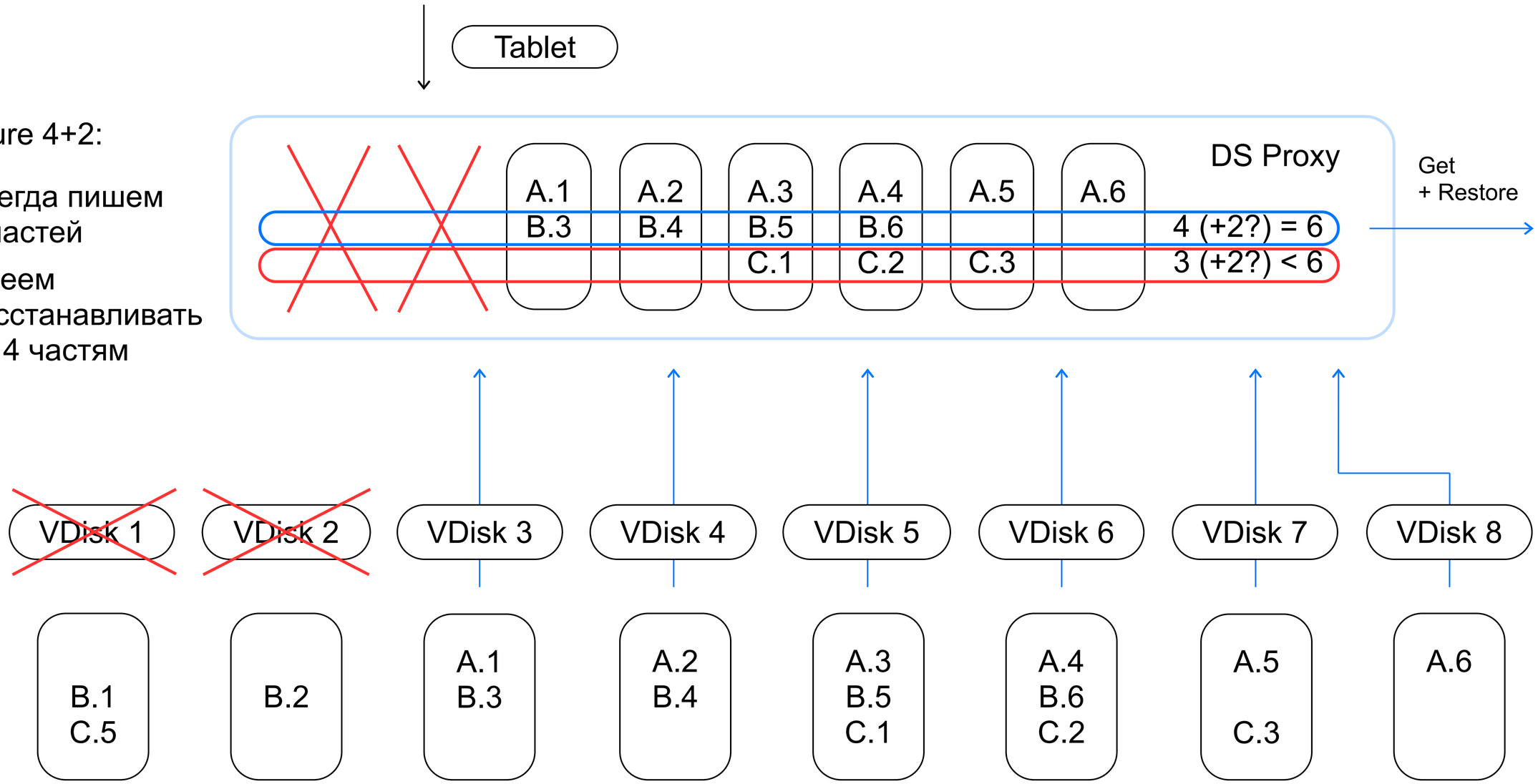


HighLoad++ Siberia 2019

DSProxy Discover

Erasure 4+2:

- Всегда пишем 6 частей
- Умеем восстанавливать по 4 частям

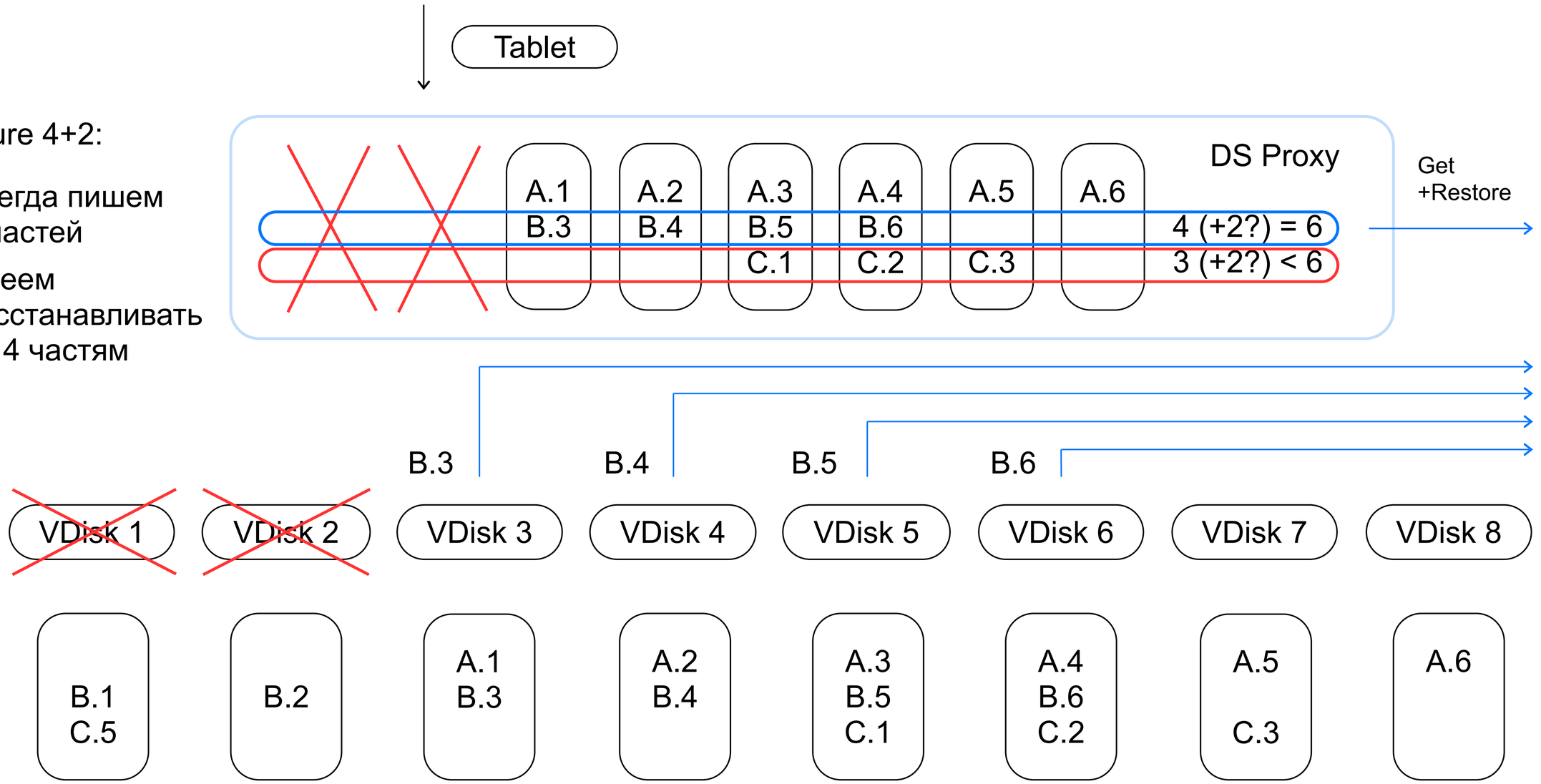


HighLoad++ Siberia 2019

DSProxy Discover

Erasure 4+2:

- Всегда пишем 6 частей
- Умеем восстанавливать по 4 частям

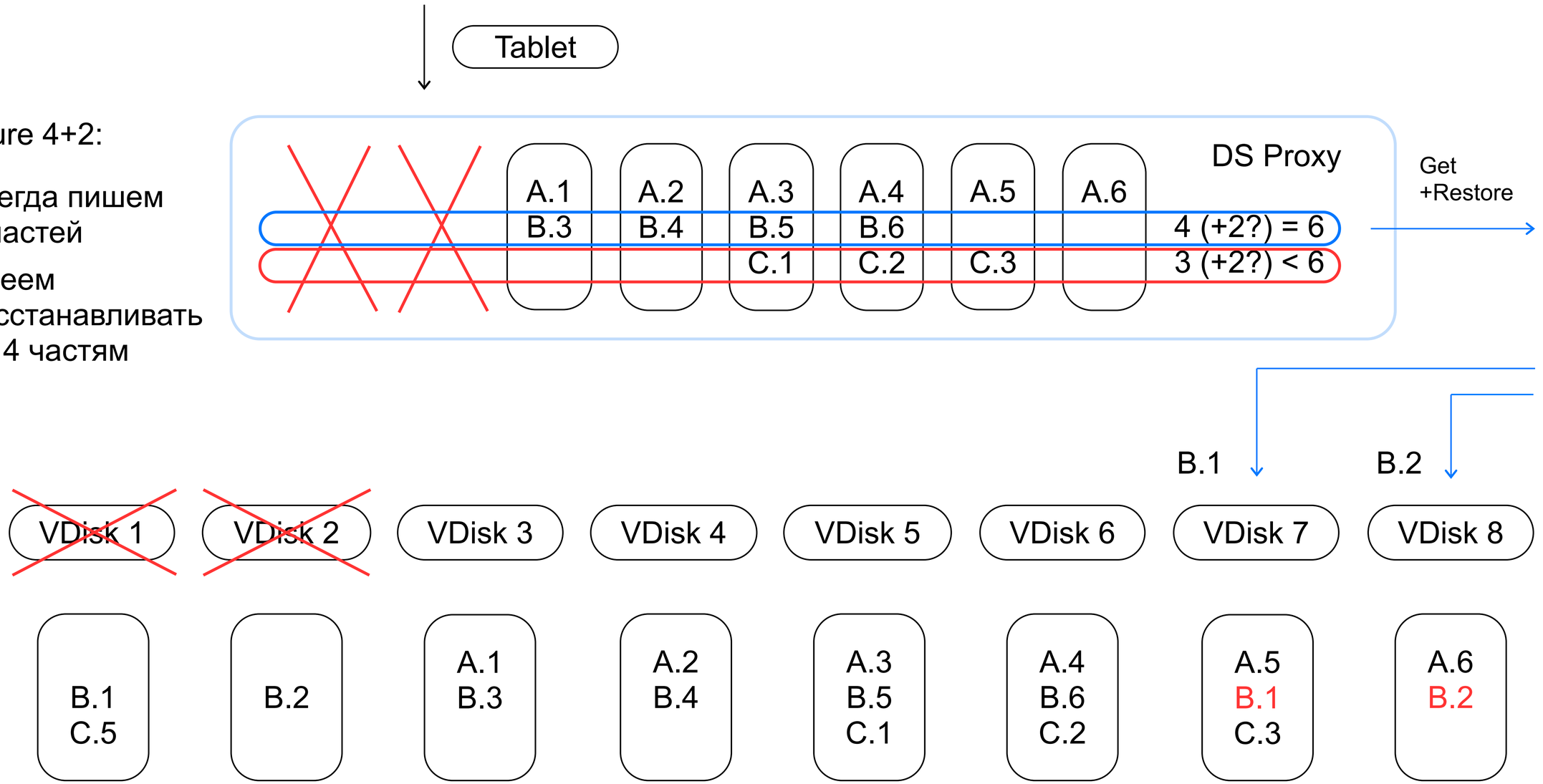


HighLoad++ Siberia 2019

DSProxy Discover

Erasure 4+2:

- Всегда пишем 6 частей
- Умеем восстанавливать по 4 частям

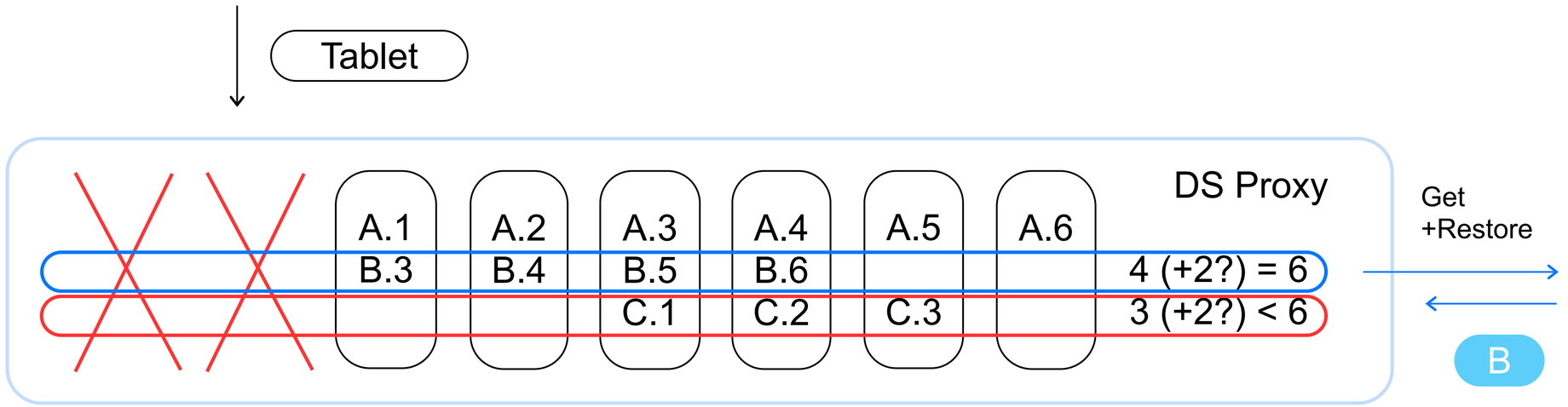


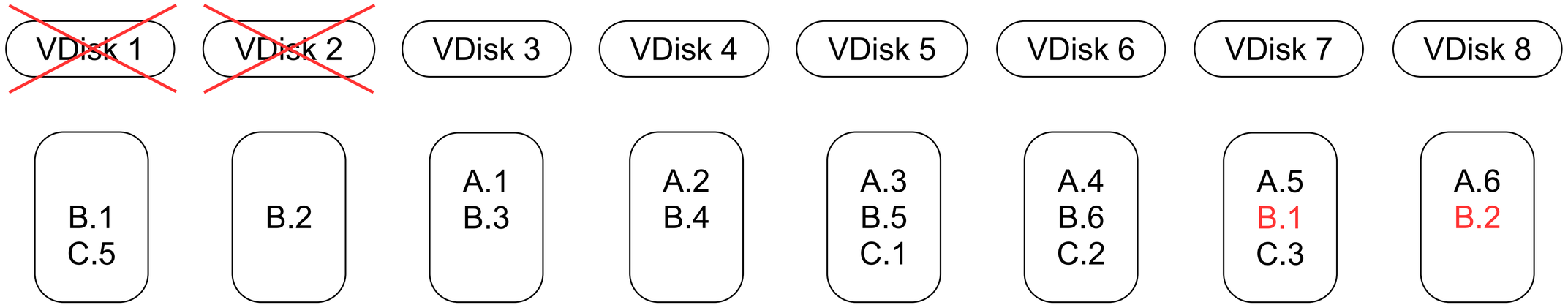
HighLoad++ Siberia 2019

DSProxy Discover

Erasure 4+2:

- Всегда пишем 6 частей
- Умеем восстанавливать по 4 частям



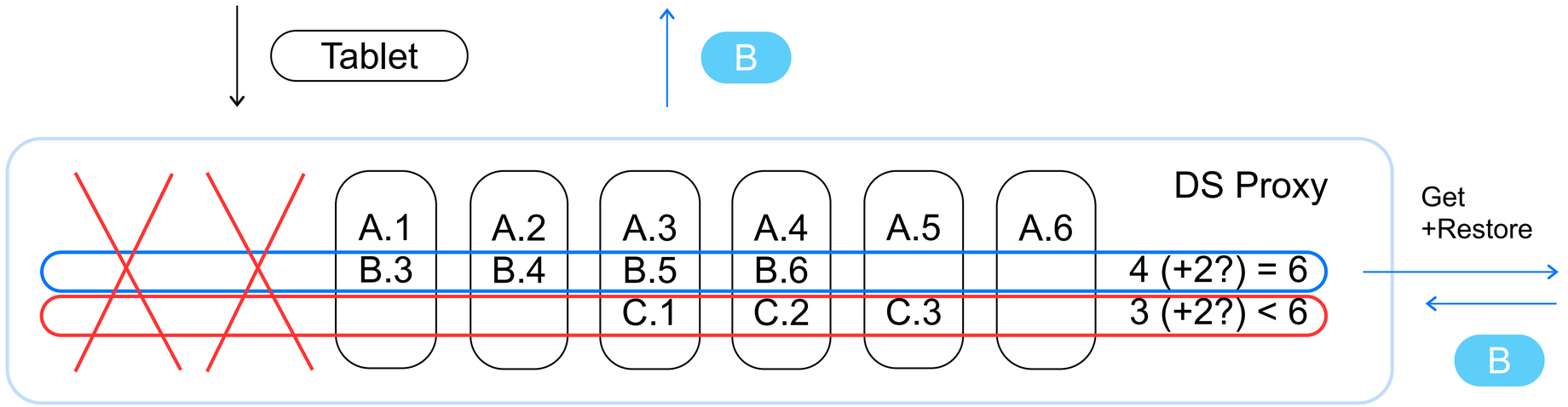


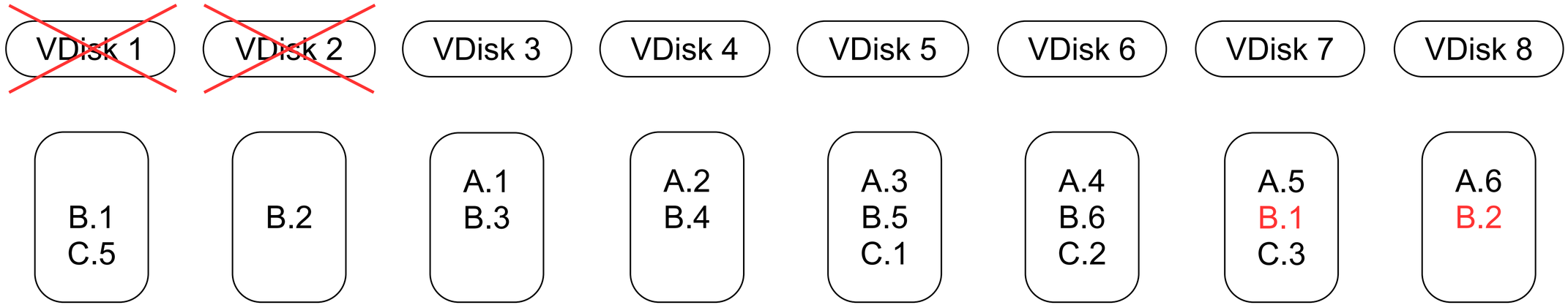
HighLoad++ Siberia 2019

DSProxy Discover

Erasure 4+2:

- Всегда пишем 6 частей
- Умеем восстанавливать по 4 частям





Устройство YDB Distributed Storage DSProxy Get + Restore

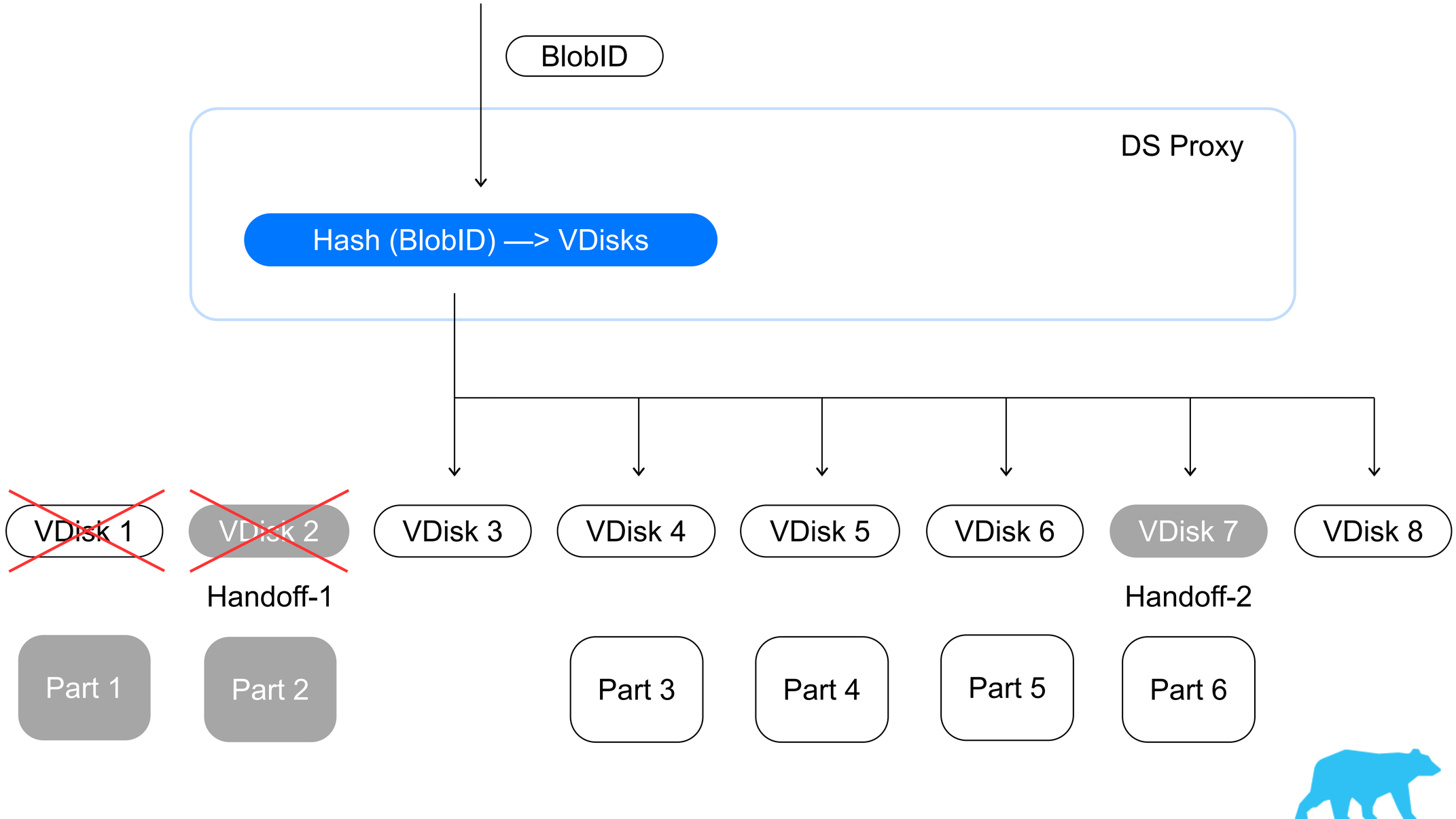
(Restore-on-read)

DSProxy Get + Restore

- Задача — прочитать блобы, возможно, записанные не до конца
- Восстановить их так, чтобы при новых ошибках их всё равно можно прочесть
- Нужно только для блобов, запись которых не была завершена

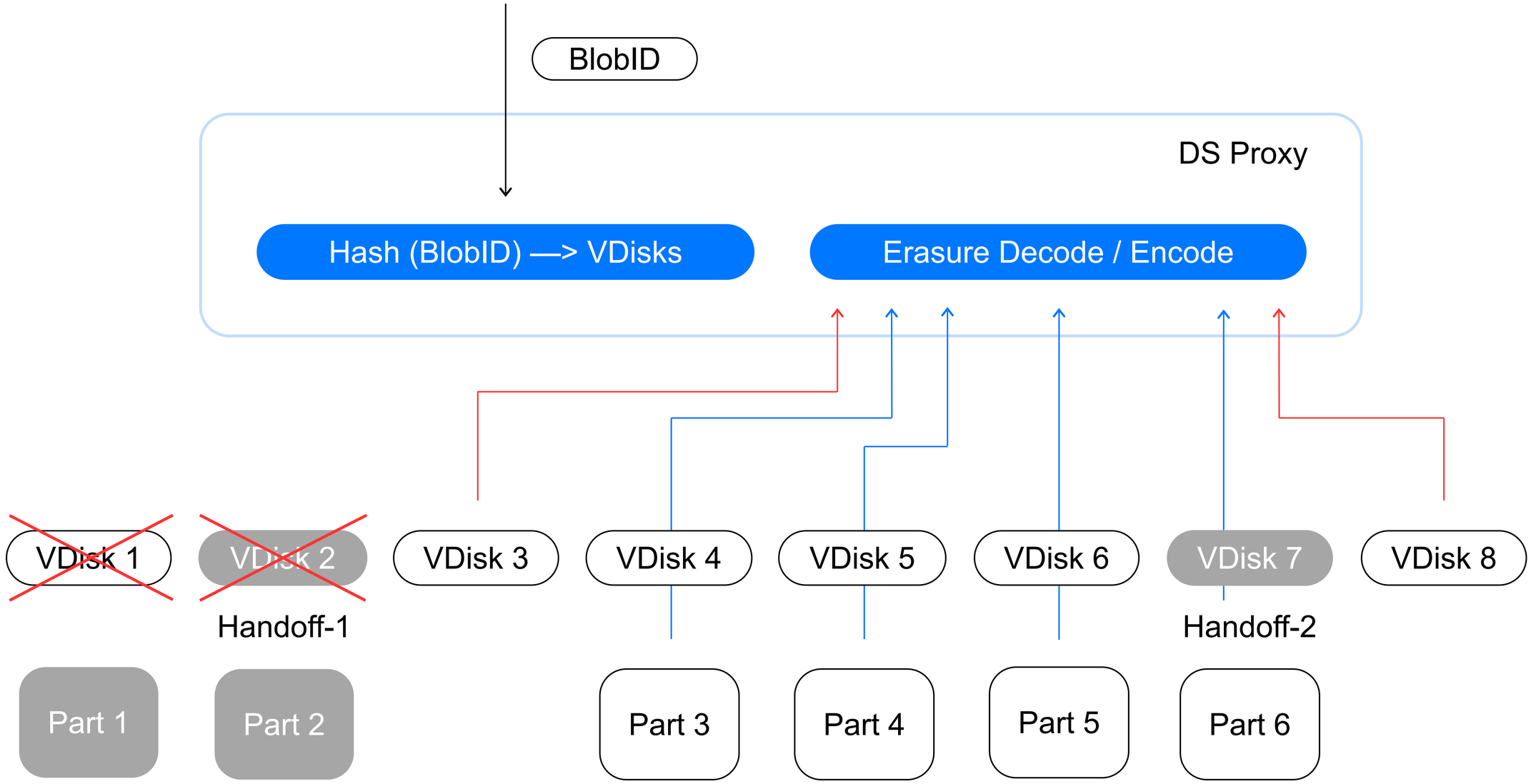
HighLoad++ Siberia 2019

DSProxy Get + Restore



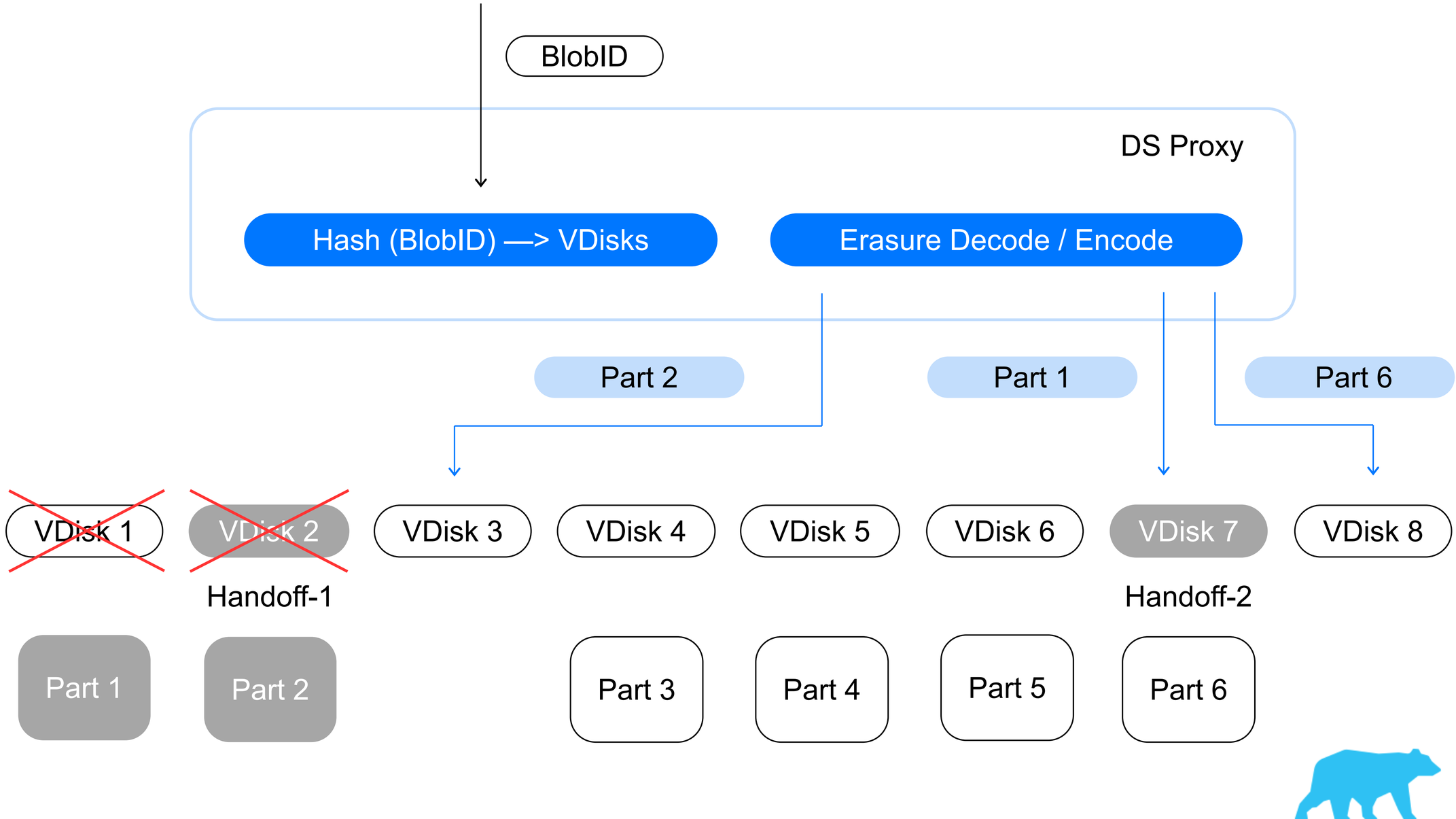
HighLoad++ Siberia 2019

DSProxy Get + Restore



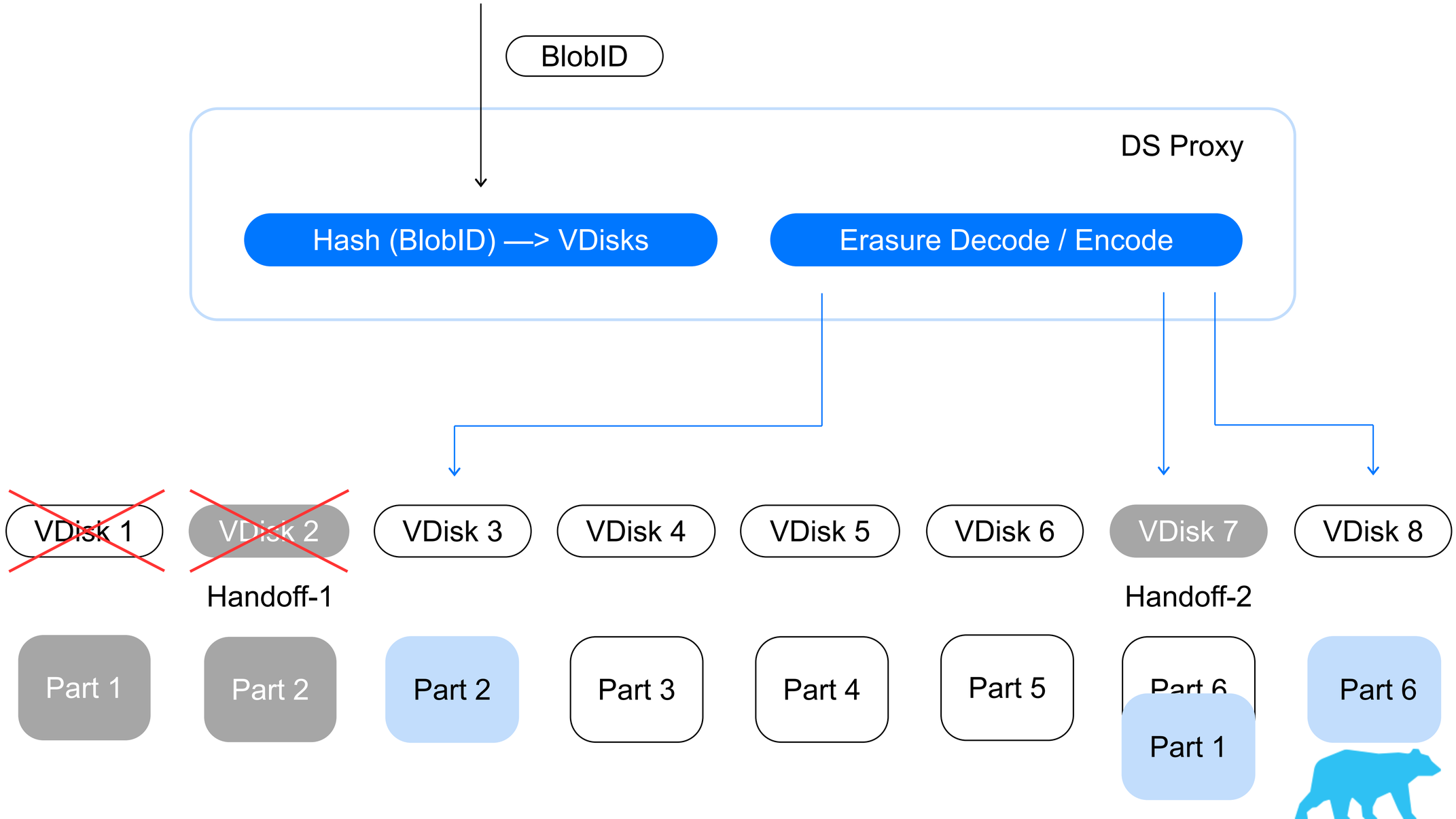
HighLoad++ Siberia 2019

DSProxy Get + Restore



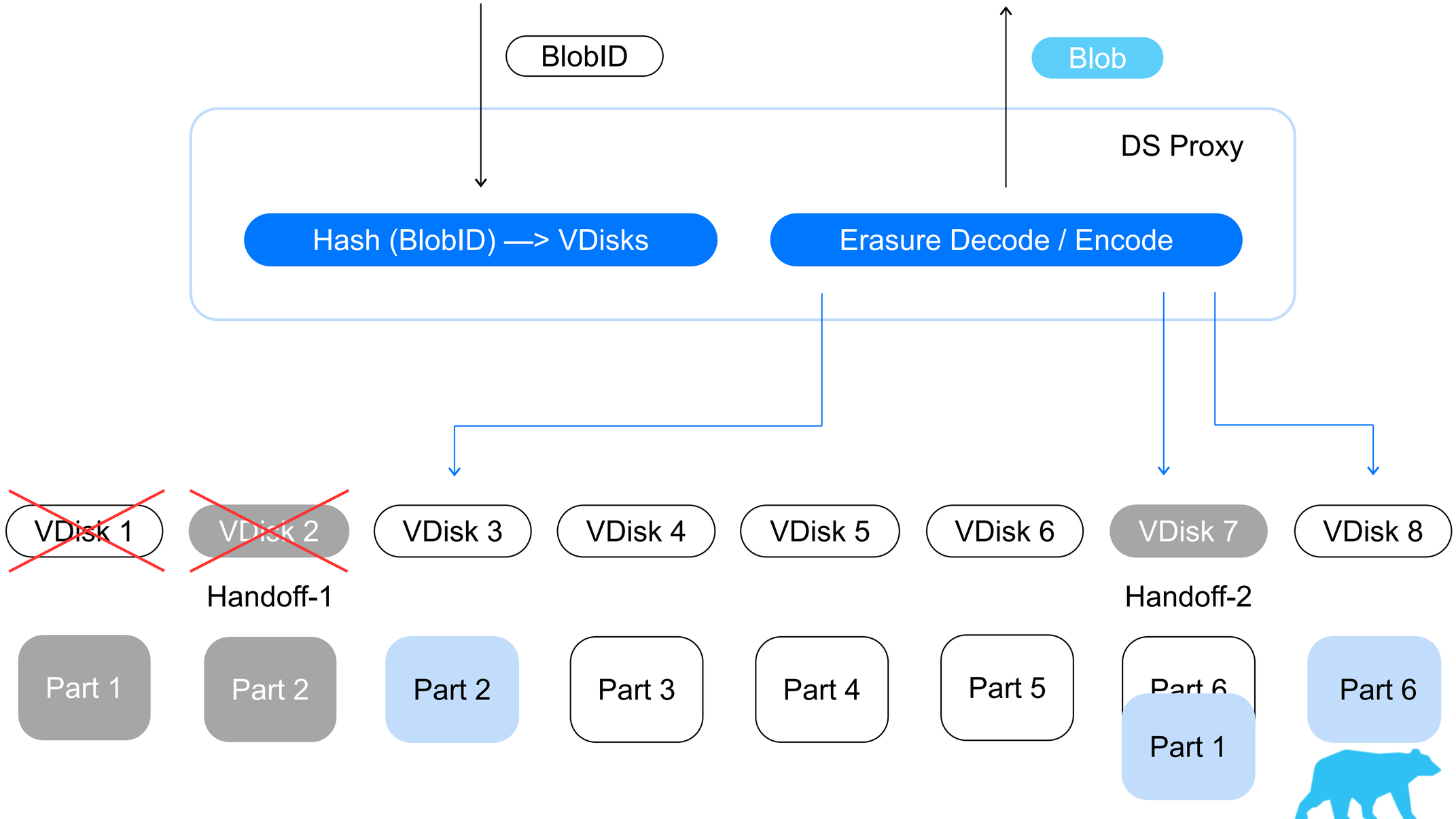
HighLoad++ Siberia 2019

DSProxy Get + Restore



HighLoad++ Siberia 2019

DSProxy Get + Restore



HighLoad++ Siberia 2019

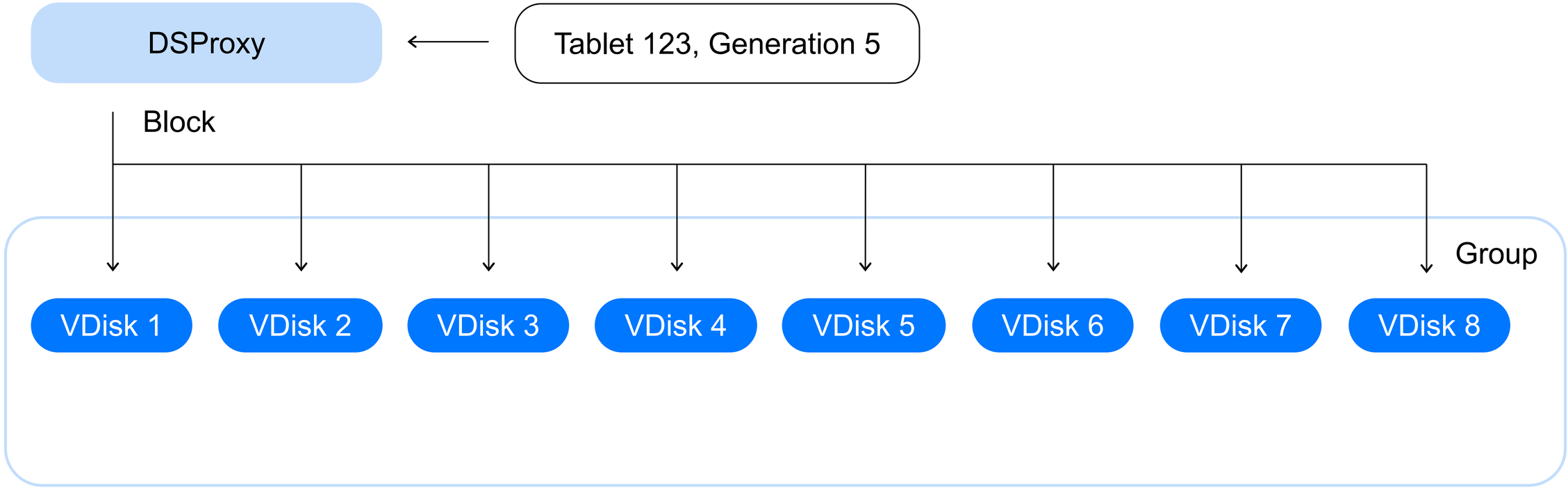
Устройство YDB Distributed Storage DSProxy Block

Блокировка группы, Block

- Решает проблему распределённого консенсуса
- Позволяет запускаться только единственному экземпляру tablet
- У каждой tablet имеется поколение
- При каждом рестарте tablet поколение увеличивается

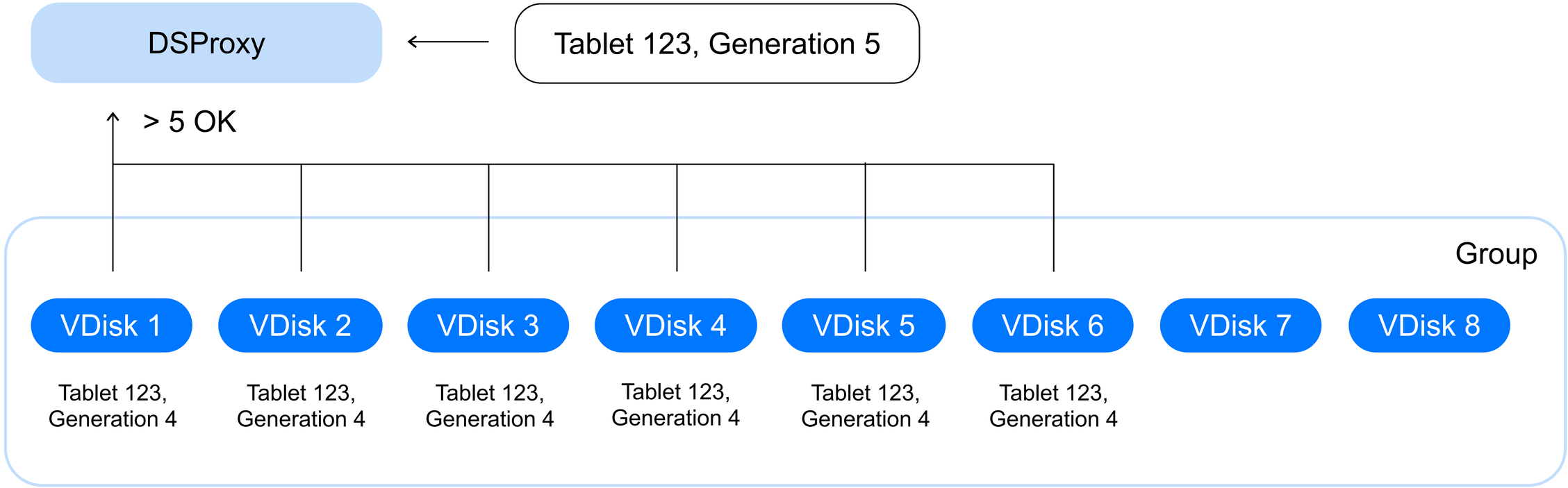
HighLoad++ Siberia 2019

4 + 2 Erasure, Block



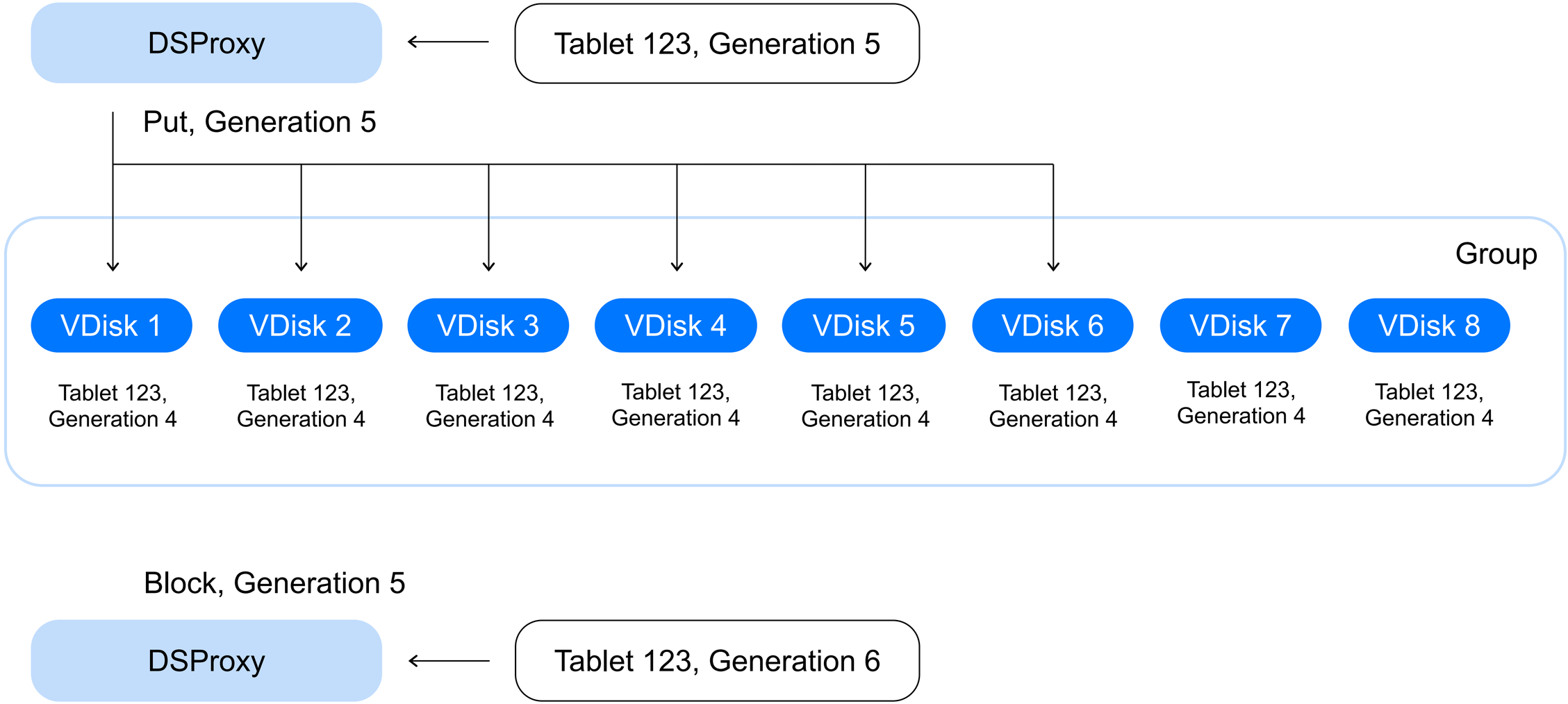
HighLoad++ Siberia 2019

4 + 2 Erasure, Block



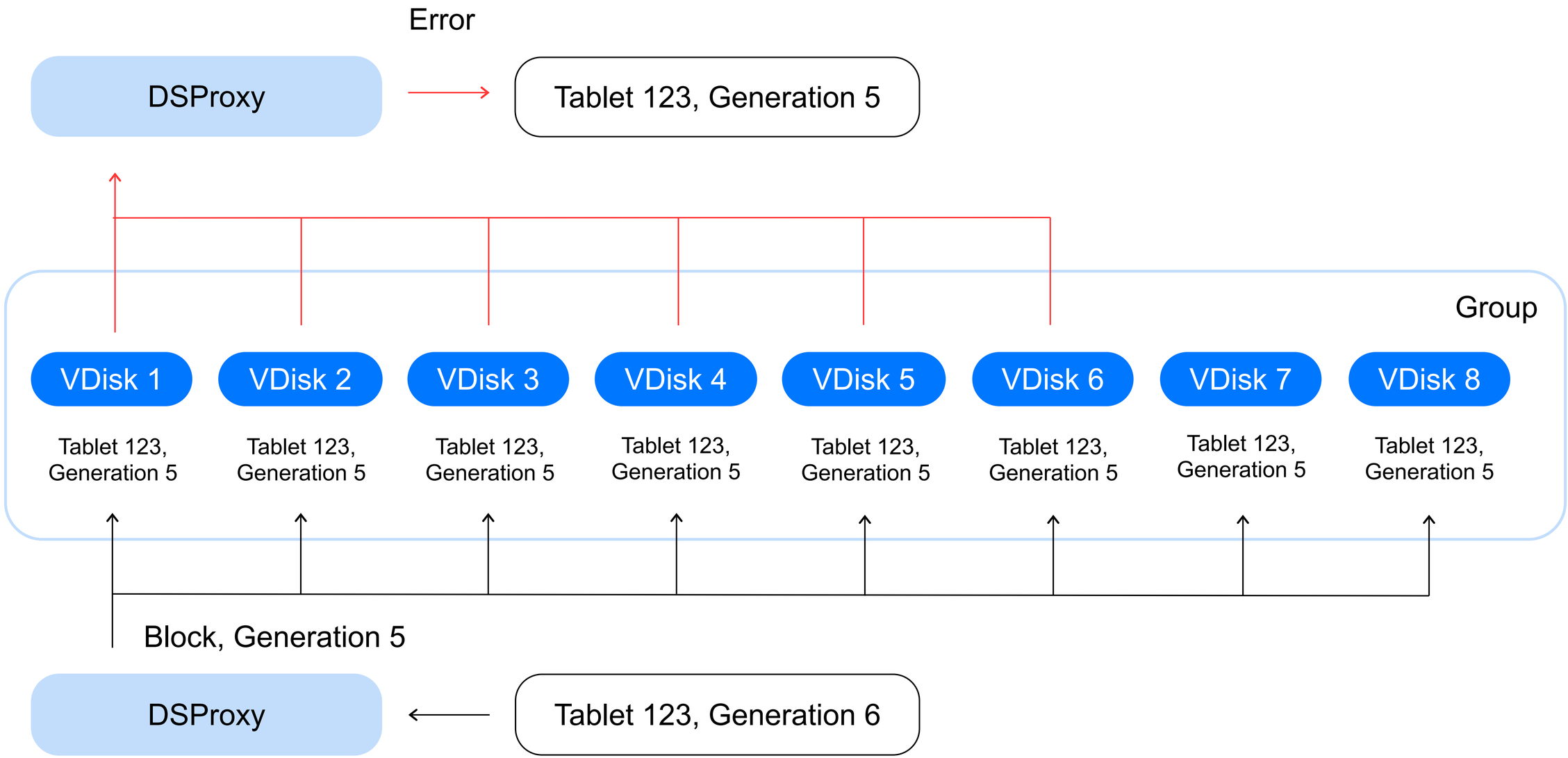
HighLoad++ Siberia 2019

4 + 2 Erasure, Block



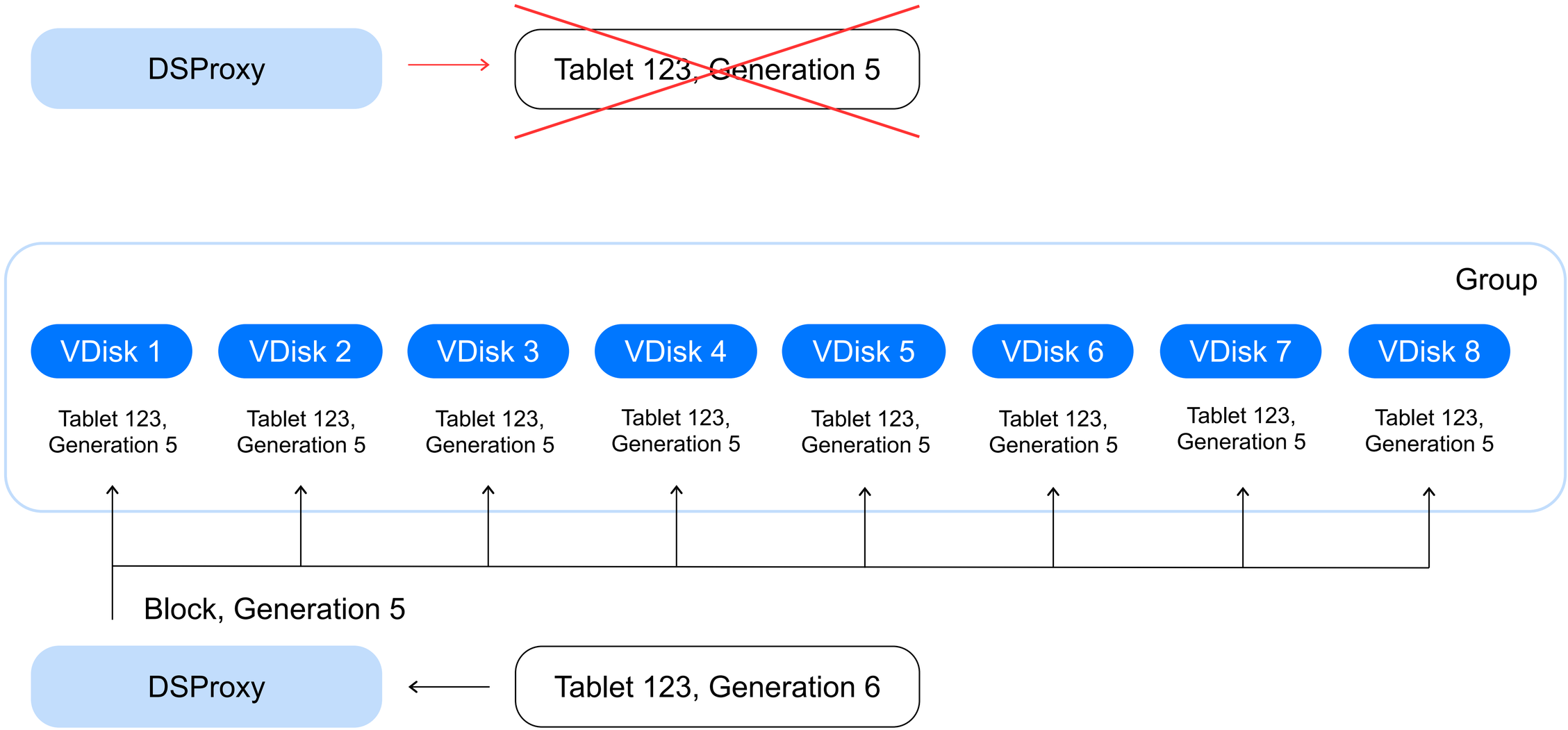
HighLoad++ Siberia 2019

4 + 2 Erasure, Block



HighLoad++ Siberia 2019

4 + 2 Erasure, Block



HighLoad++ Siberia 2019

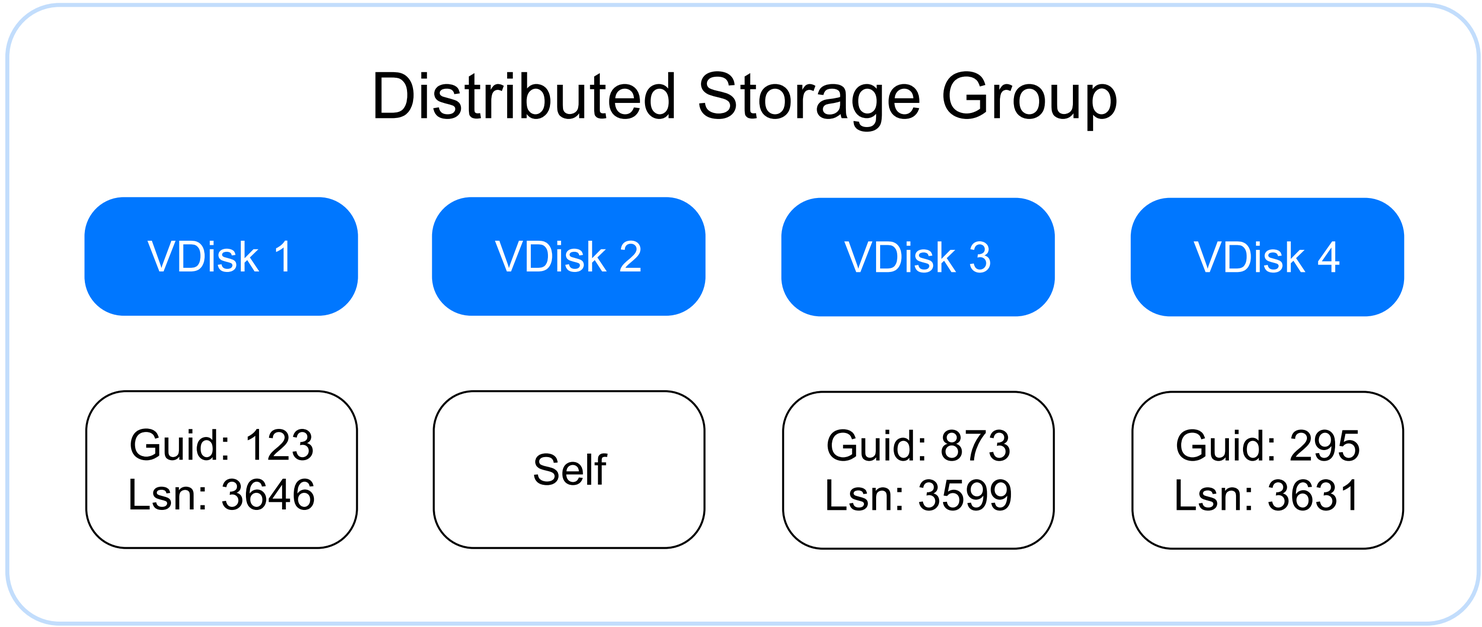
Устройство
YDB Distributed Storage VDisk

VDisk

- Можно рассматривать как key-value хранилище частей блобов
- Взаимодействие peer-to-peer всех VDisk’ов внутри группы
- Хранит данные блобов (в виде LSM-Tree)
- Хранит метаданные блобов
- Хранит метаданные блокировок и барьеров

HighLoad++ Siberia 2019

Синхронизация данных VDisk’a



- Синхронизация метаданных (BlobID, Keep / Don’t Keep Flags, Barriers) каждый с каждым в рамках группы
- VDiskGuid — уникальный идентификатор VDisk’a, генерируется при первом запуске группы
- Lsn — Log Sequence Number, каждая запись в логе идентифицируется Lsn
- Про каждый элемент хранимых данных VDisk знает, есть ли он на других VDisk’ах группы

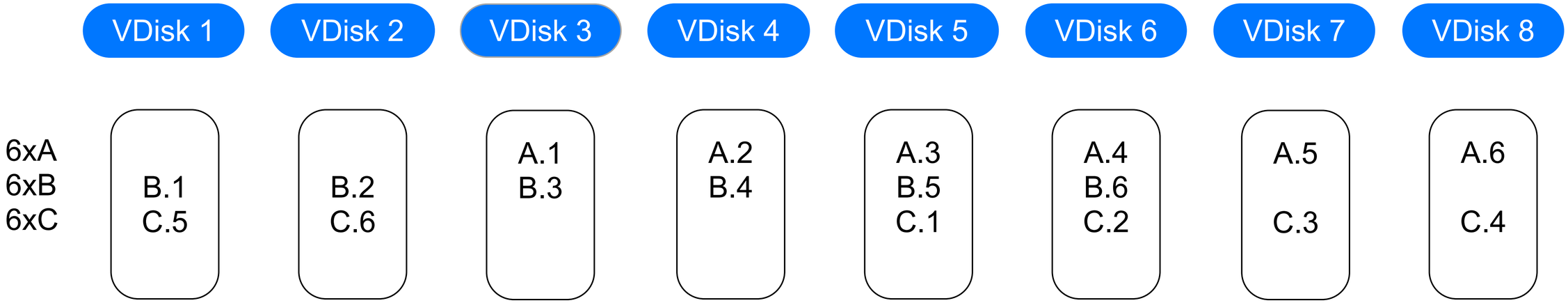
HighLoad++ Siberia 2019

Синхронизация VDisk

При выпадении / отставании диска на рестарте он синхронизирует свои метаданные со всеми другими VDisk в своей группе

Erasure 4+2:

- Всегда пишем 6 частей
- Умеем восстанавливать по 4 частям



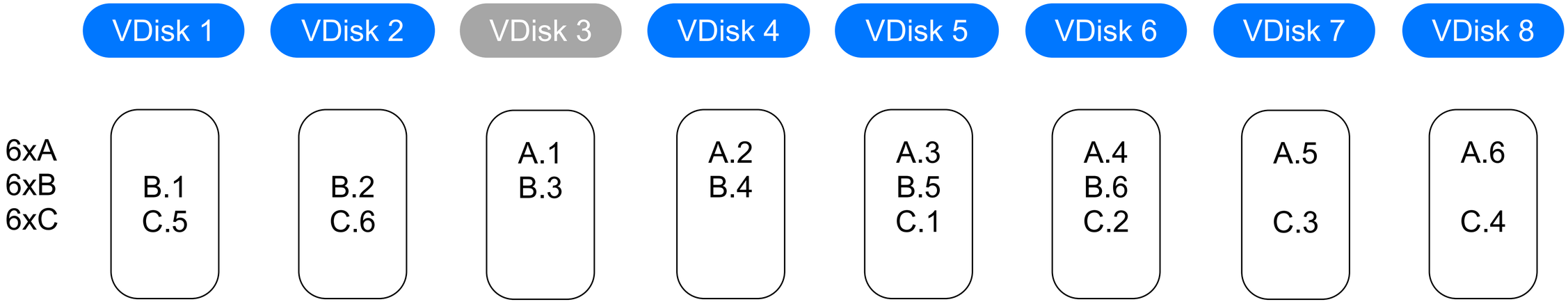
HighLoad++ Siberia 2019

Синхронизация VDisk

При выпадении / отставании диска на рестарте он синхронизирует свои метаданные со всеми другими VDisk в своей группе

Erasure 4+2:

- Всегда пишем 6 частей
- Умеем восстанавливать по 4 частям



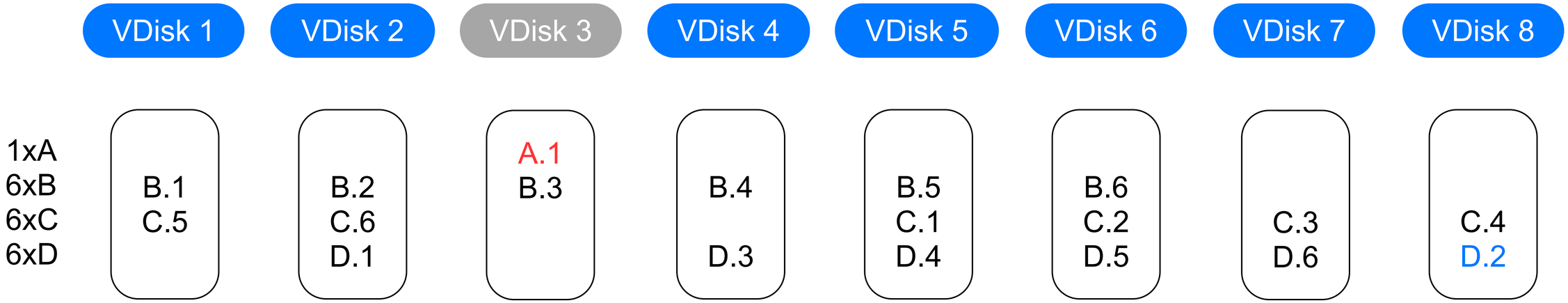
HighLoad++ Siberia 2019

Синхронизация VDisk

При выпадении / отставании диска на рестарте он синхронизирует свои метаданные со всеми другими VDisk в своей группе

Erasure 4+2:

- Всегда пишем 6 частей
- Умеем восстанавливать по 4 частям



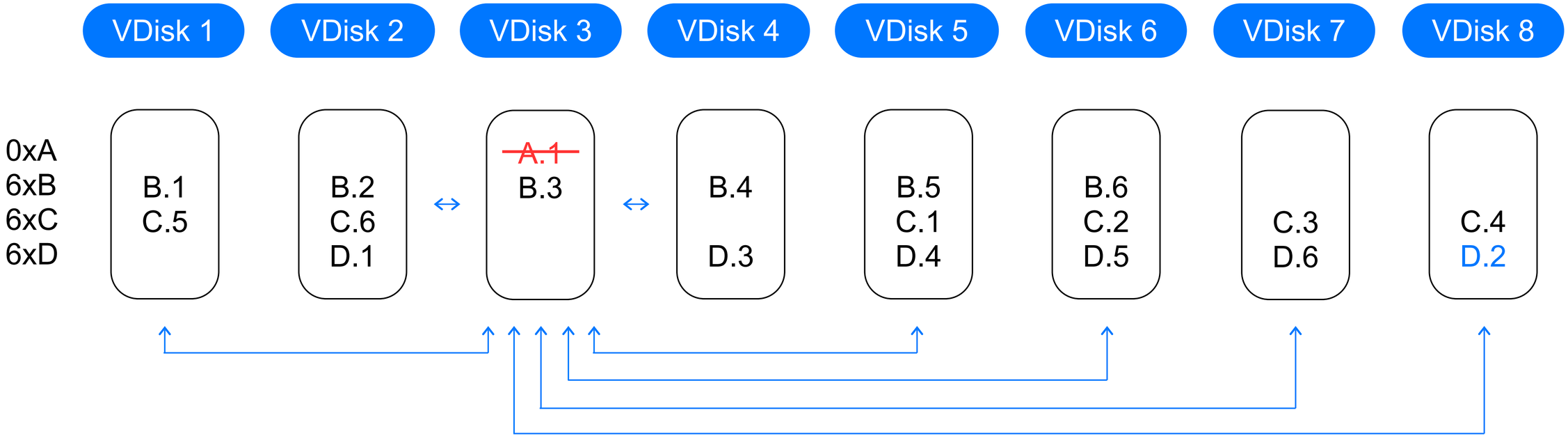
HighLoad++ Siberia 2019

Синхронизация VDisk

При выпадении / отставании диска на рестарте он синхронизирует свои метаданные со всеми другими VDisk в своей группе

Erasure 4+2:

- Всегда пишем 6 частей
- Умеем восстанавливать по 4 частям



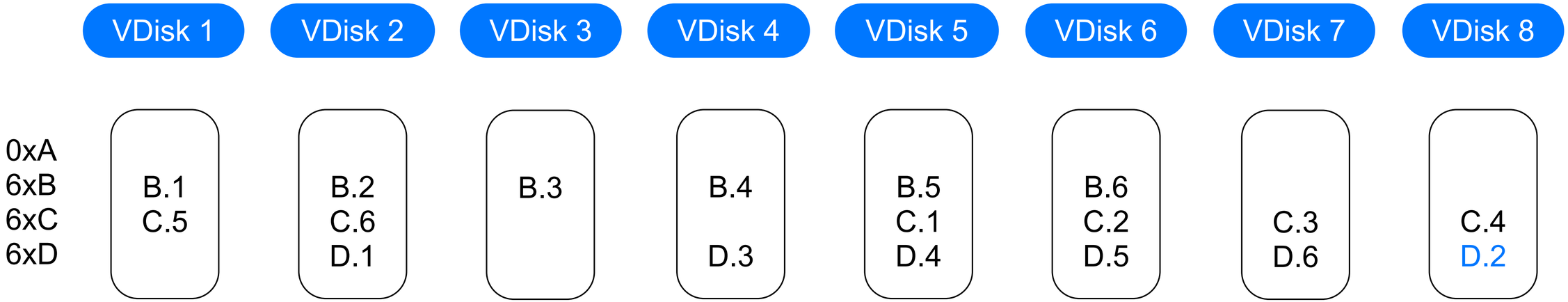
HighLoad++ Siberia 2019

Синхронизация VDisk

При выпадении / отставании диска на рестарте он синхронизирует свои метаданные со всеми другими VDisk в своей группе

Erasure 4+2:

- Всегда пишем 6 частей
- Умеем восстанавливать по 4 частям



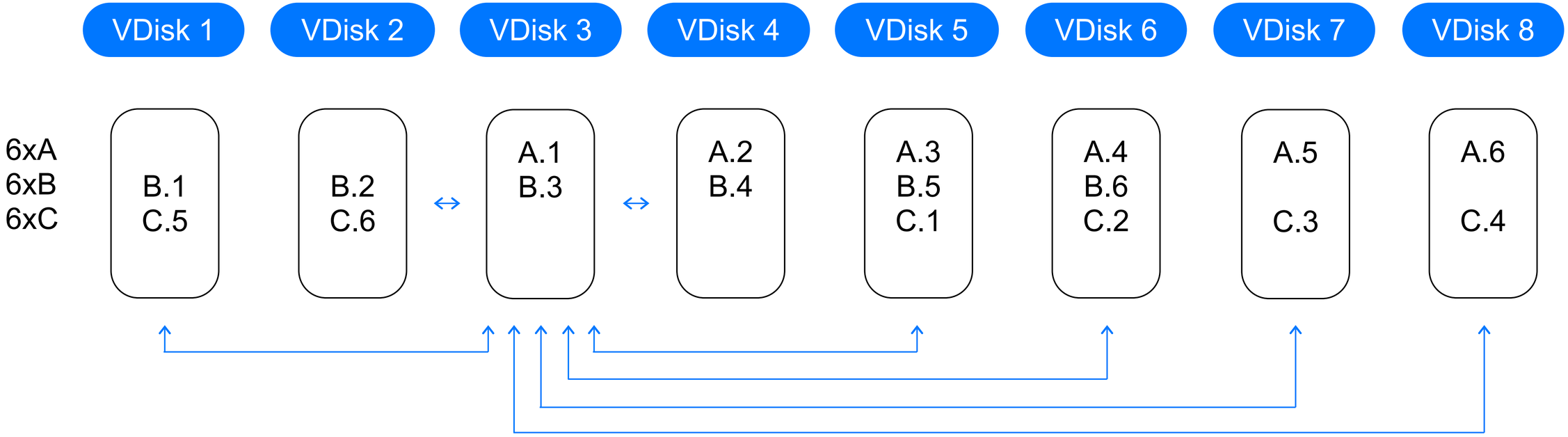
HighLoad++ Siberia 2019

Репликация VDisk

При удалении данных одного VDisk (замена диска) на рестарте VDisk закачивает все данные, которые были у него до замены, восстанавливая их при помощи данных с других дисков

Erasure 4+2:

- Всегда пишем 6 частей
- Умеем восстанавливать по 4 частям



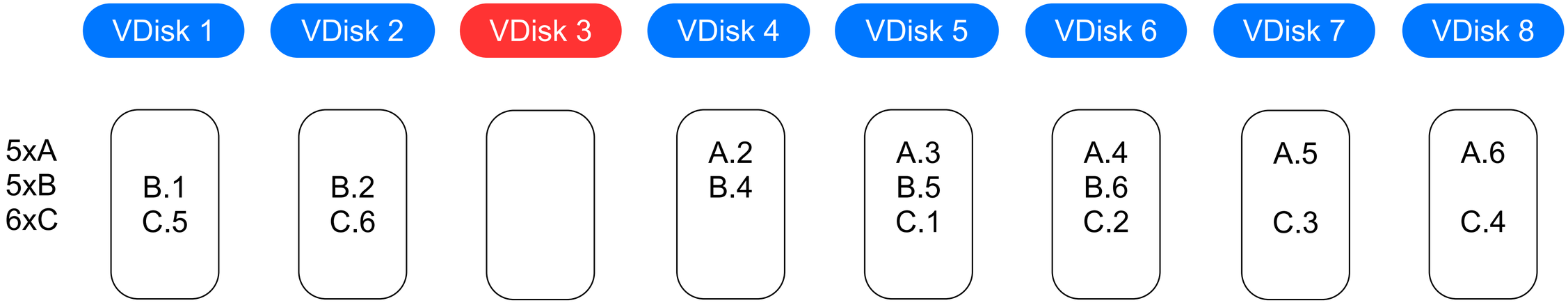
HighLoad++ Siberia 2019

Репликация VDisk

При удалении данных одного VDisk (замена диска) на рестарте VDisk закачивает все данные, которые были у него до замены, восстанавливая их при помощи данных с других дисков

Erasure 4+2:

- Всегда пишем 6 частей
- Умеем восстанавливать по 4 частям



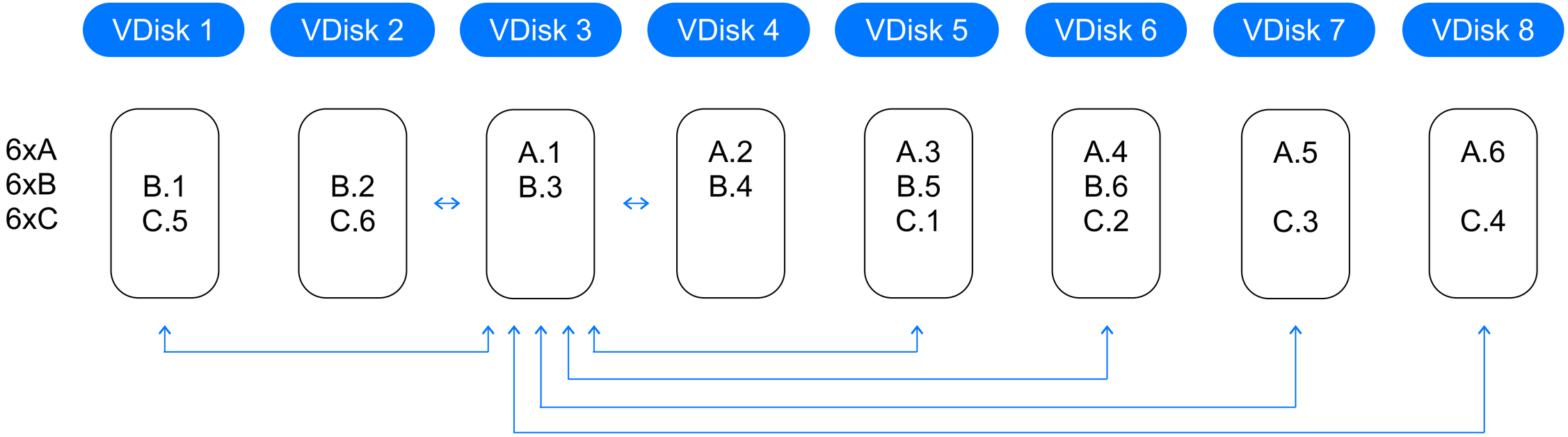
HighLoad++ Siberia 2019

Репликация VDisk

При удалении данных одного VDisk (замена диска) на рестарте VDisk закачивает все данные, которые были у него до замены, восстанавливая их при помощи данных с других дисков

Erasure 4+2:

- Всегда пишем 6 частей
- Умеем восстанавливать по 4 частям



HighLoad++ Siberia 2019

Ключевые показатели отказоустойчивости

- YDB выдерживает выпадение любых двух дисков из группы (группы из 8 и более дисков)
- Время наработки на отказ YDB-кластера из 1000 серверов — 100 лет:
  - Время наработки на отказ одного диска — 160 лет
  - Время наработки на отказ любого диска из 1000 нод — 8 дней
- Автоматически переживать отказ ДЦ без downtime
- Автоматически восстанавливаться после выпадении серверов / стоек за десятки миллисекунд

HighLoad++ Siberia 2019

Яндекс Облако

Спасибо за внимание!

Владислав Кузнецов

Разработчик

va-kuznecov@yandex-team.ru

HighLoad++ Siberia 2019