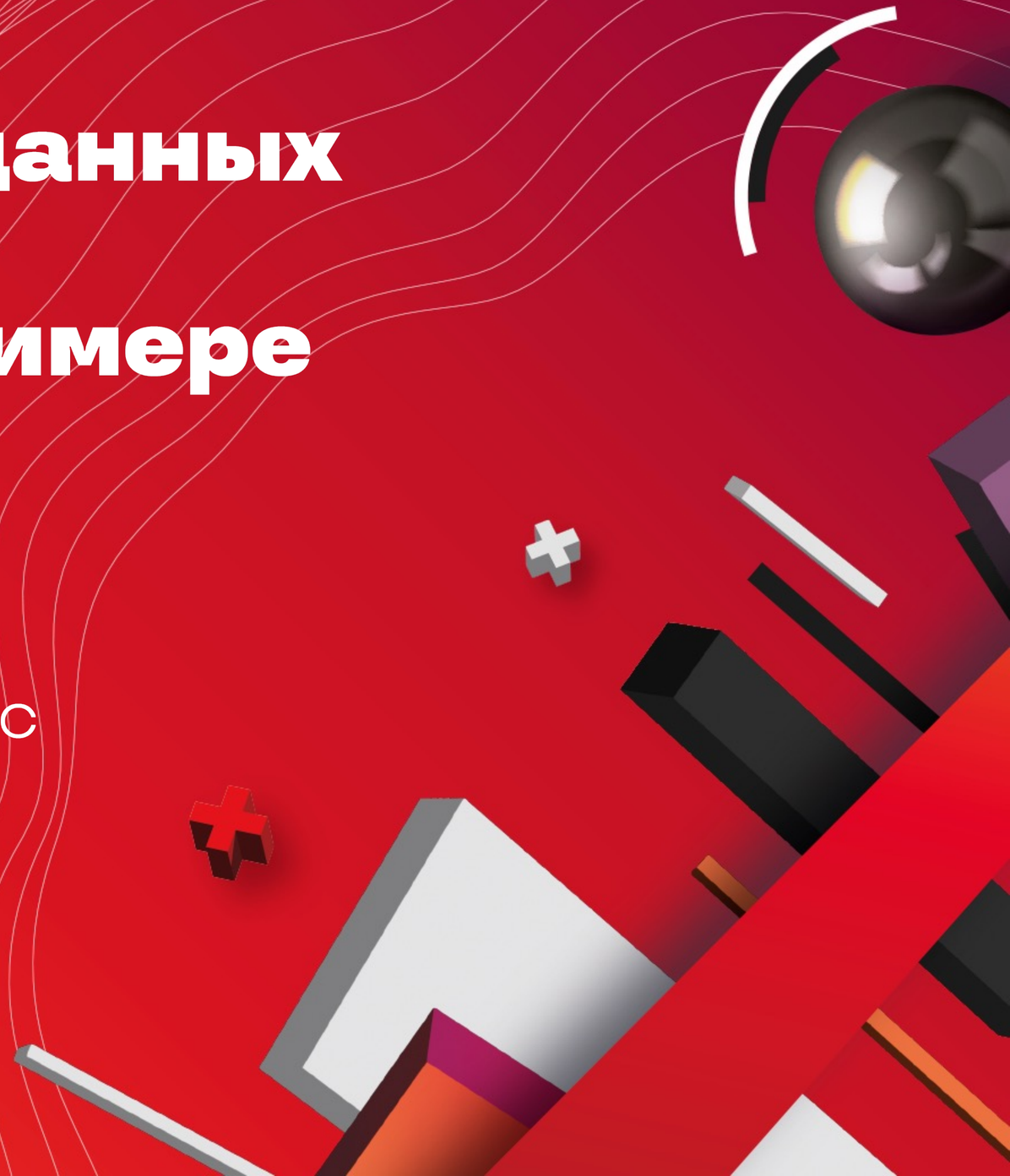
Особенности шин данных для очень больших инсталляций на примере YDB Topics

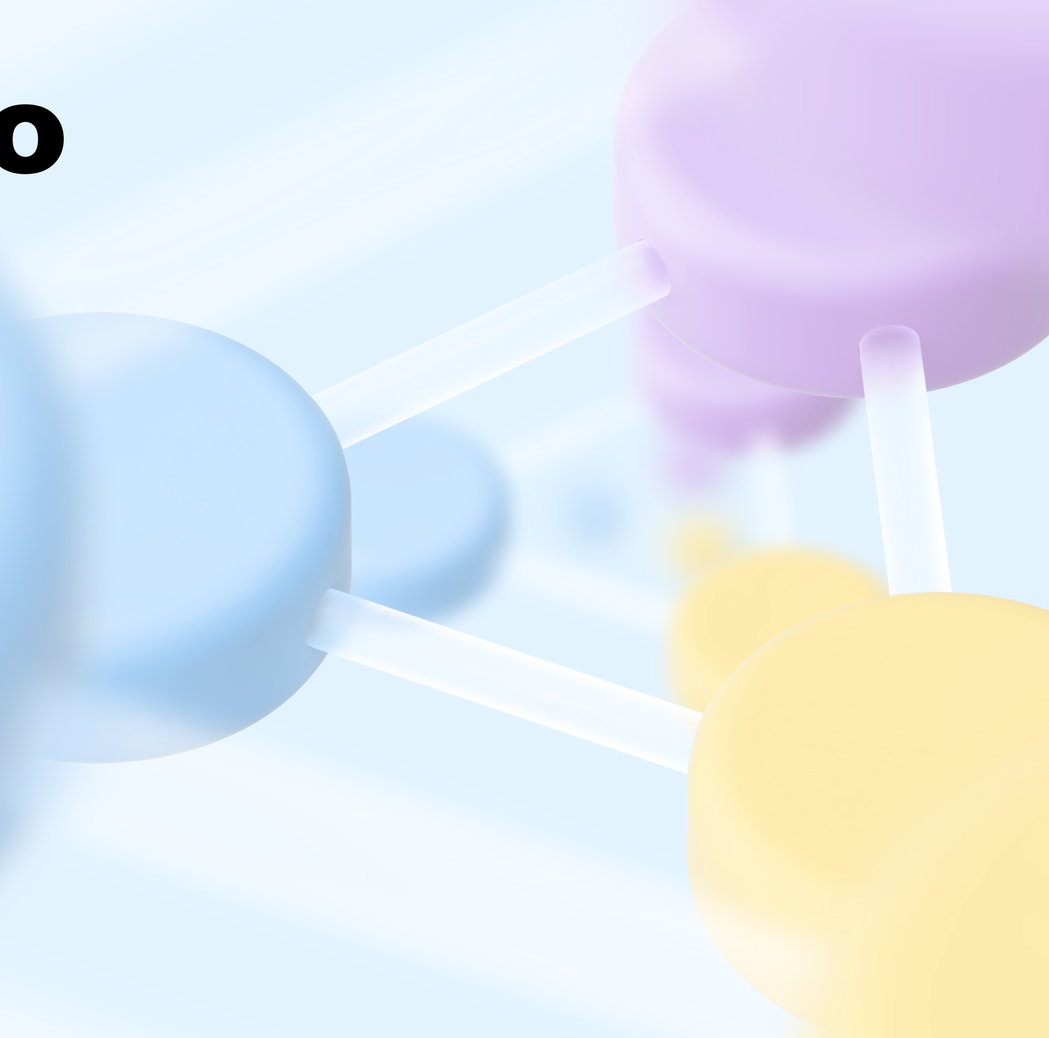
Алексей Дмитриев,
технический менеджер, Яндекс

HL 2023 HighLoad++



Почему большим компаниям сложно с Open Source

1. Что такое YDB Topics
2. Оборудование
3. Отказоустойчивость
4. Управление кластером
5. Self-service



Требования к шине передачи данных

- Гарантии доставки данных, работа в пределах SLA и модели отказа
- Минимальное количество оборудования
- Минимальные усилия по поддержке
- Интеграция с внутренней экосистемой

Что такое YDB Topics

- Масштабируемый сервис для надёжной передачи упорядоченных потоков сообщений
- Позволяет приложениям обмениваться сообщениями через очереди по модели pub/sub

2013 — Production Яндекса поверх Apache Kafka

2017 — Production Яндекса поверх YDB

2022 — Выведен в Open Source

Немного цифр о YDB Topics в Яндексе

Скорость передачи: запись + чтение

×200 ГБ/с

×21 млн EPS

5 ДЦ

1000
команд

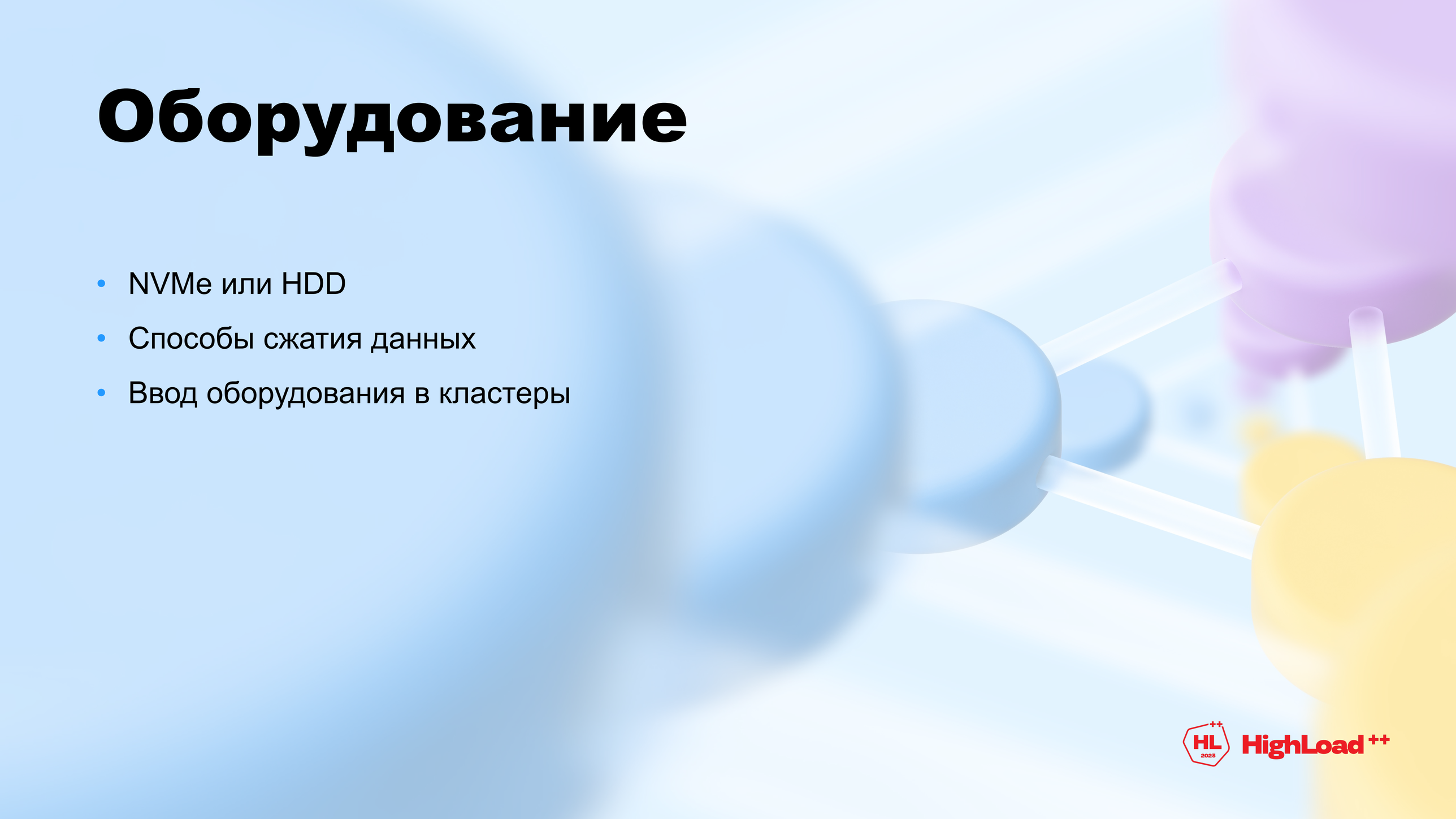
30k
топиков

1500
серверов

1
дежурный SRE

Оборудование

- NVMe или HDD
- Способы сжатия данных
- Ввод оборудования в кластеры



Общая информация

1500+

серверов в 5 ДЦ

18 ч.

минимальное время хранения данных

10 пб

общий объем хранения пользовательских данных

Диски

1

Использование NVMe-дисков дает экономию в серверах

2

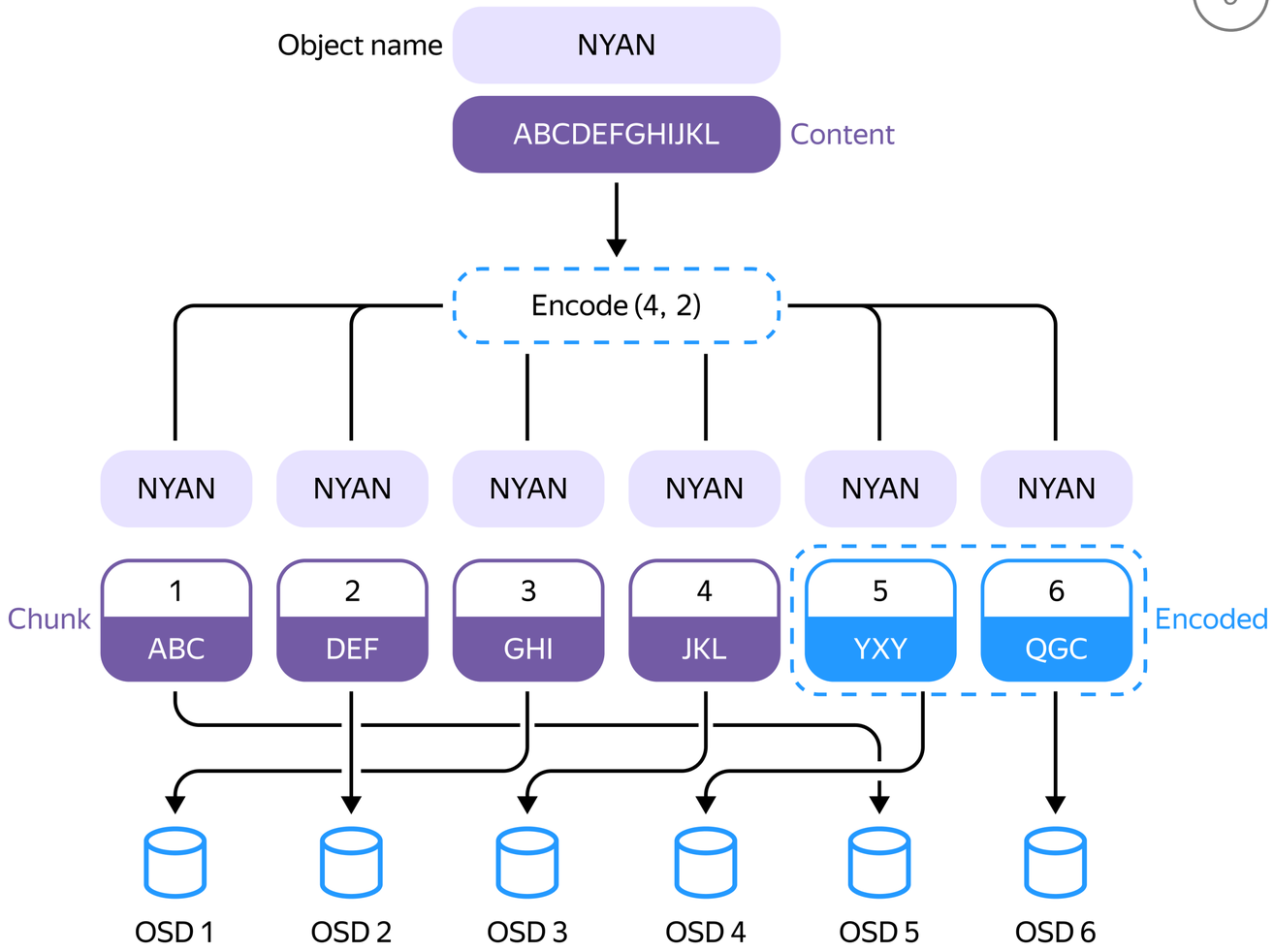
Узкое место — емкость NVMe-дисков

3

Число дисков определяет число серверов

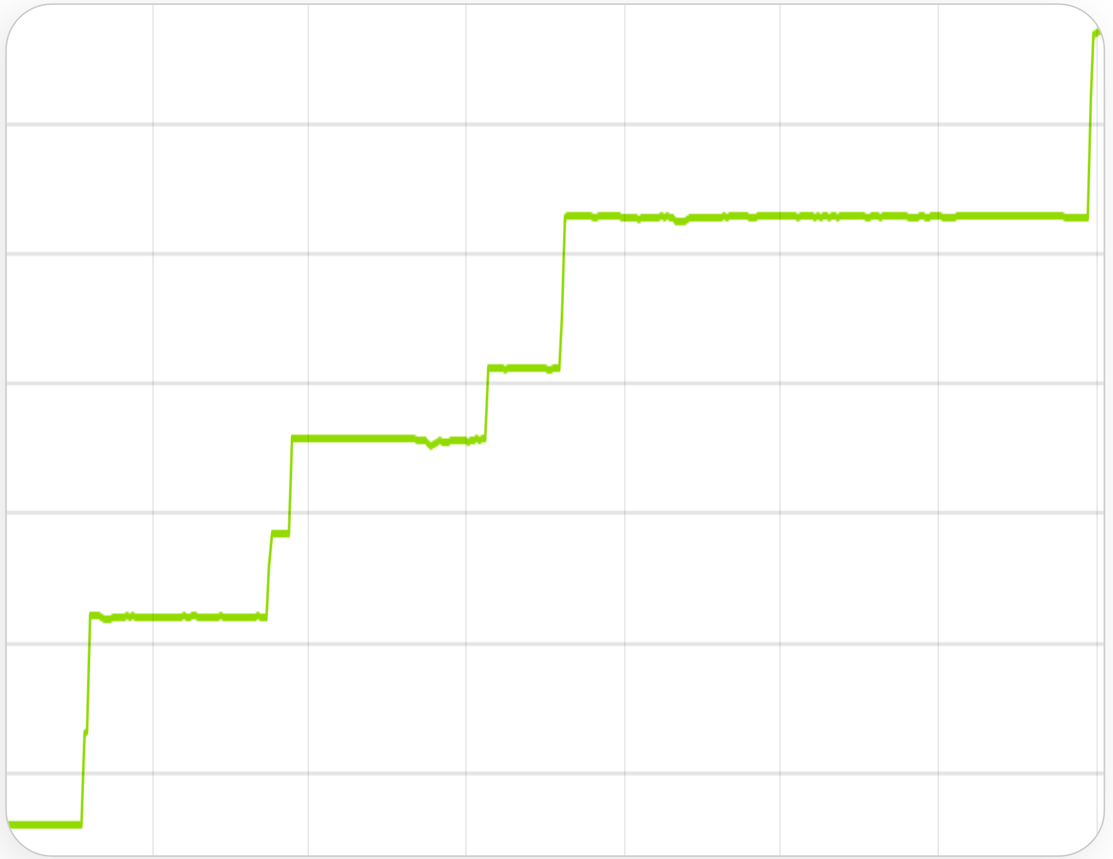
Erasure-кодирование

- Общий объем хранения пользовательских данных 10 ПБ
  - 15 ПБ при erasure-кодировании
  - 30 ПБ при репликации 3x
- При увеличении объема записи растет DWPD (Drive Writes Per Day)



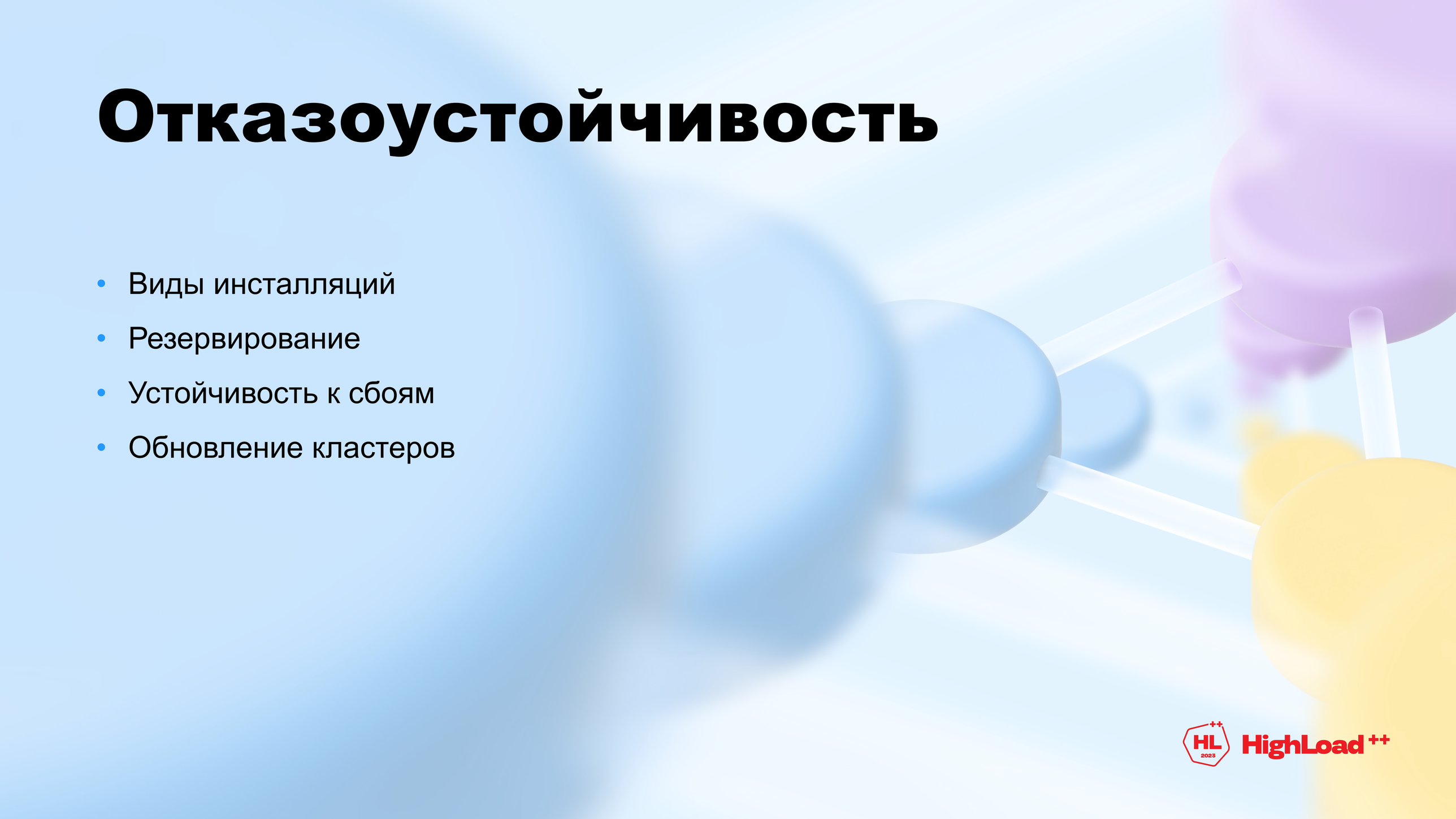
Добавление серверов

- Сервера вводятся партиями
- Обычно кластера расширяются
- Но и сжимаются тоже



Отказоустойчивость

- Виды инсталляций
- Резервирование
- Устойчивость к сбоям
- Обновление кластеров



Типы трафика

1

Передача биллинговых данных

2

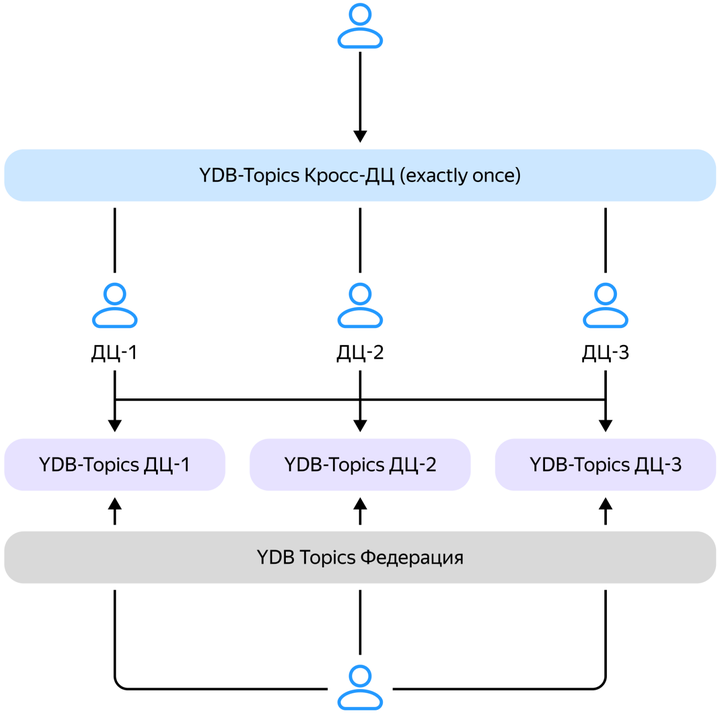
Передача журналов работы приложений

3

Передача данных реального времени

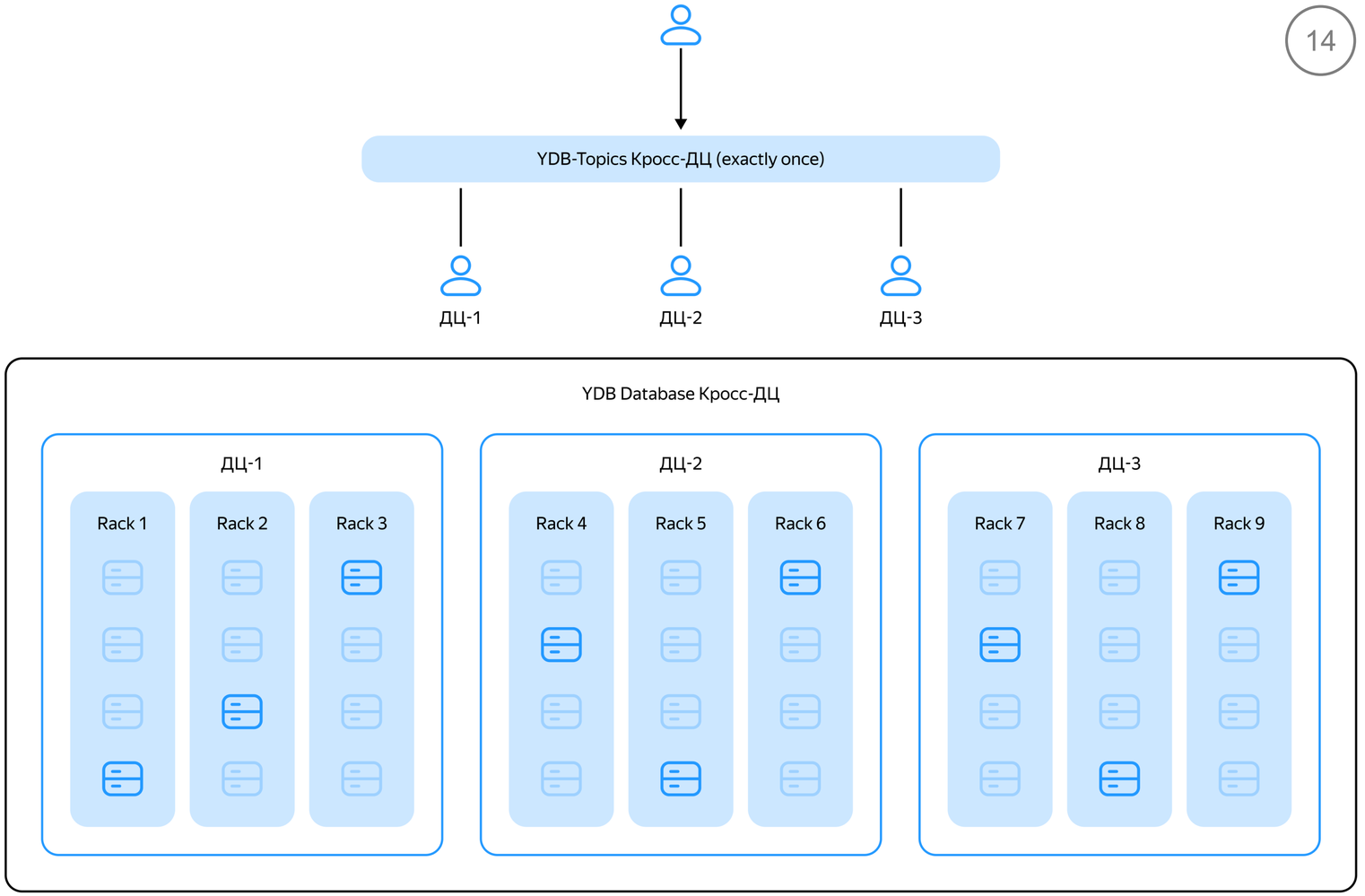
Инсталляции

- Виды инсталляций:
  Exactly-once и Федерация
- Exactly-once обеспечивает −1 ДЦ
  незаметно для пользователей
- Федерация управляет трафиком
  пользователей, переводя
  его в пропорциях в другие ДЦ



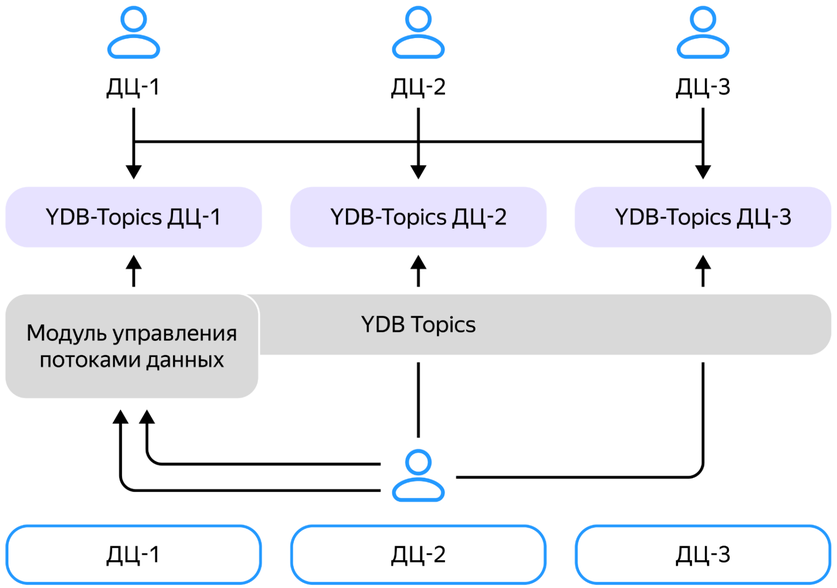
Кросс-ДЦ

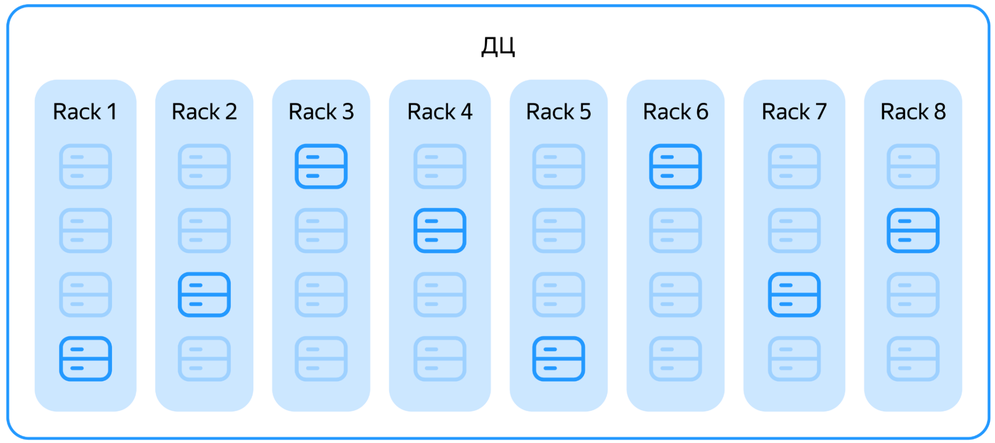
Основной сценарий — обработка exactly once (биллинговые данные), гарантия порядка и/или высокая доступность на чтение-запись



Федерация

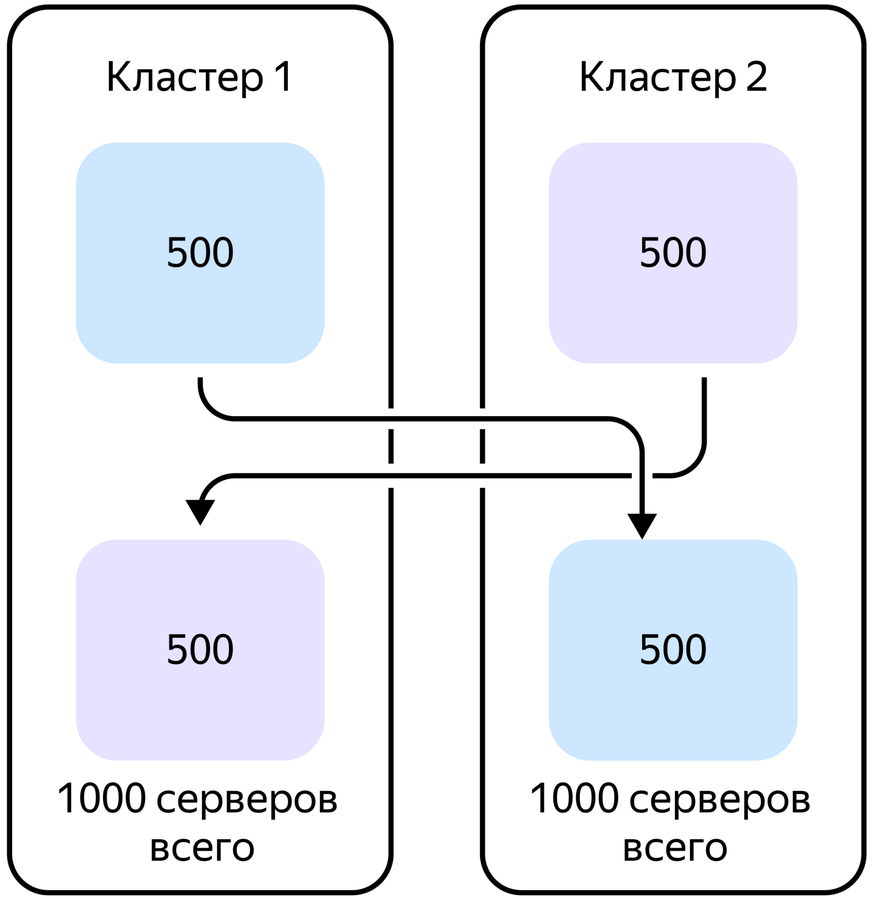
- Основной сценарий — поставка данных реального времени и журналов работы приложений
- Гарантии at least once
- Дешевле относительно кросс-ДЦ
- Модуль управления потоками данных координирует нагрузку на кластеры, управляя клиентом





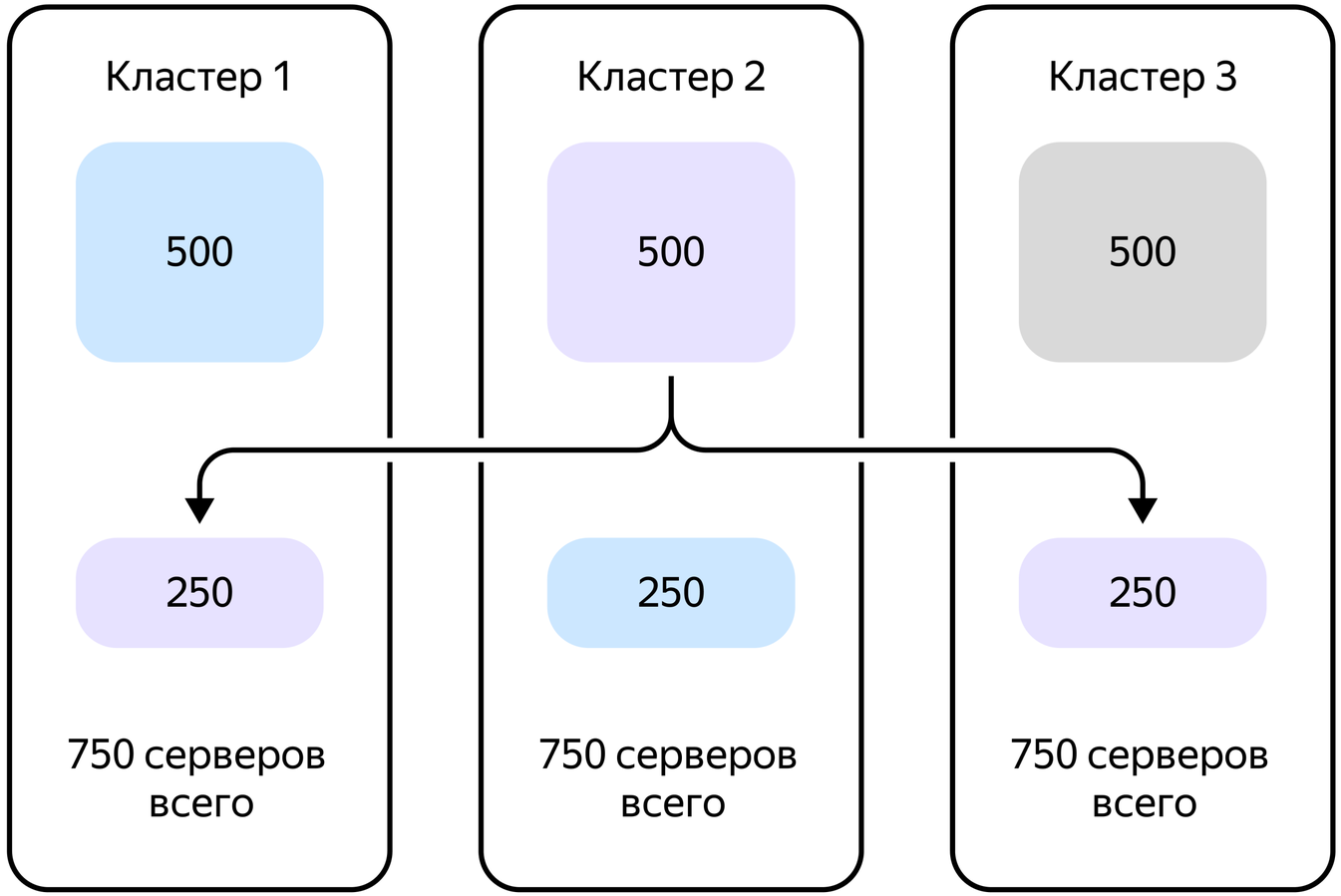
Резервирование

При двух доступных кластерах для работы при −1 ДЦ нужен 100% запас мощности в каждом кластере



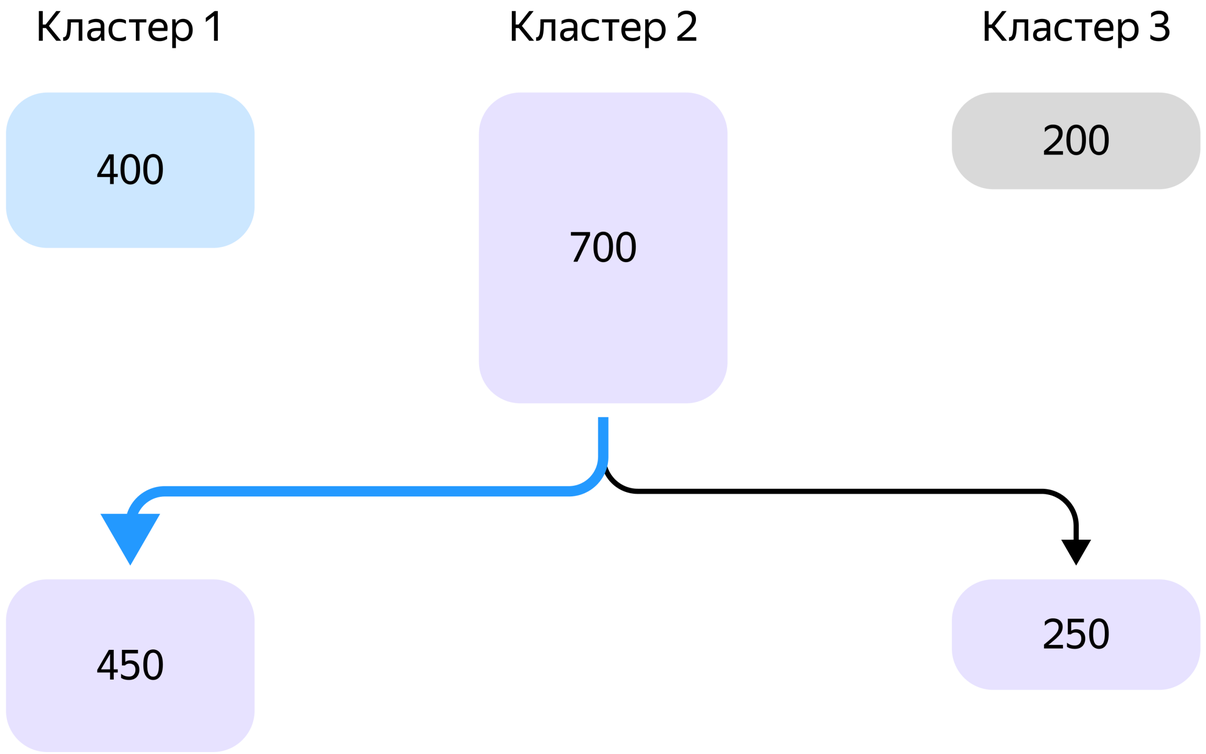
Резервирование

- При 3-х кластерах, запас в каждом кластере 50%
- При 5-и кластерах, запас в каждом кластере 25%



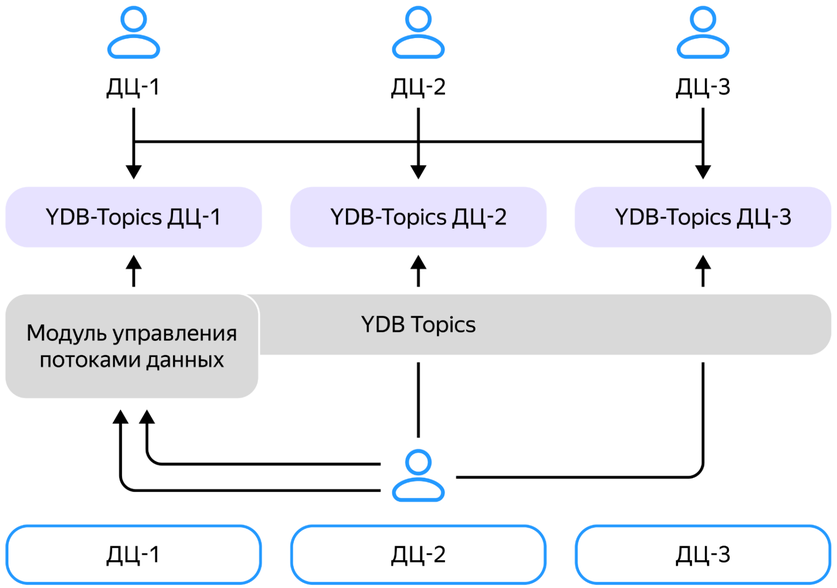
Резервирование

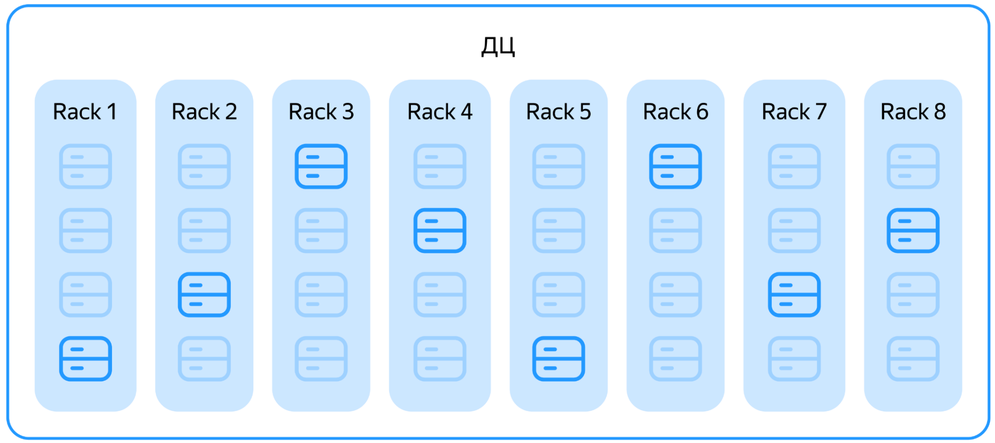
Кластеры могут быть разного размера, поэтому трафик делится в пропорциях



Федерация

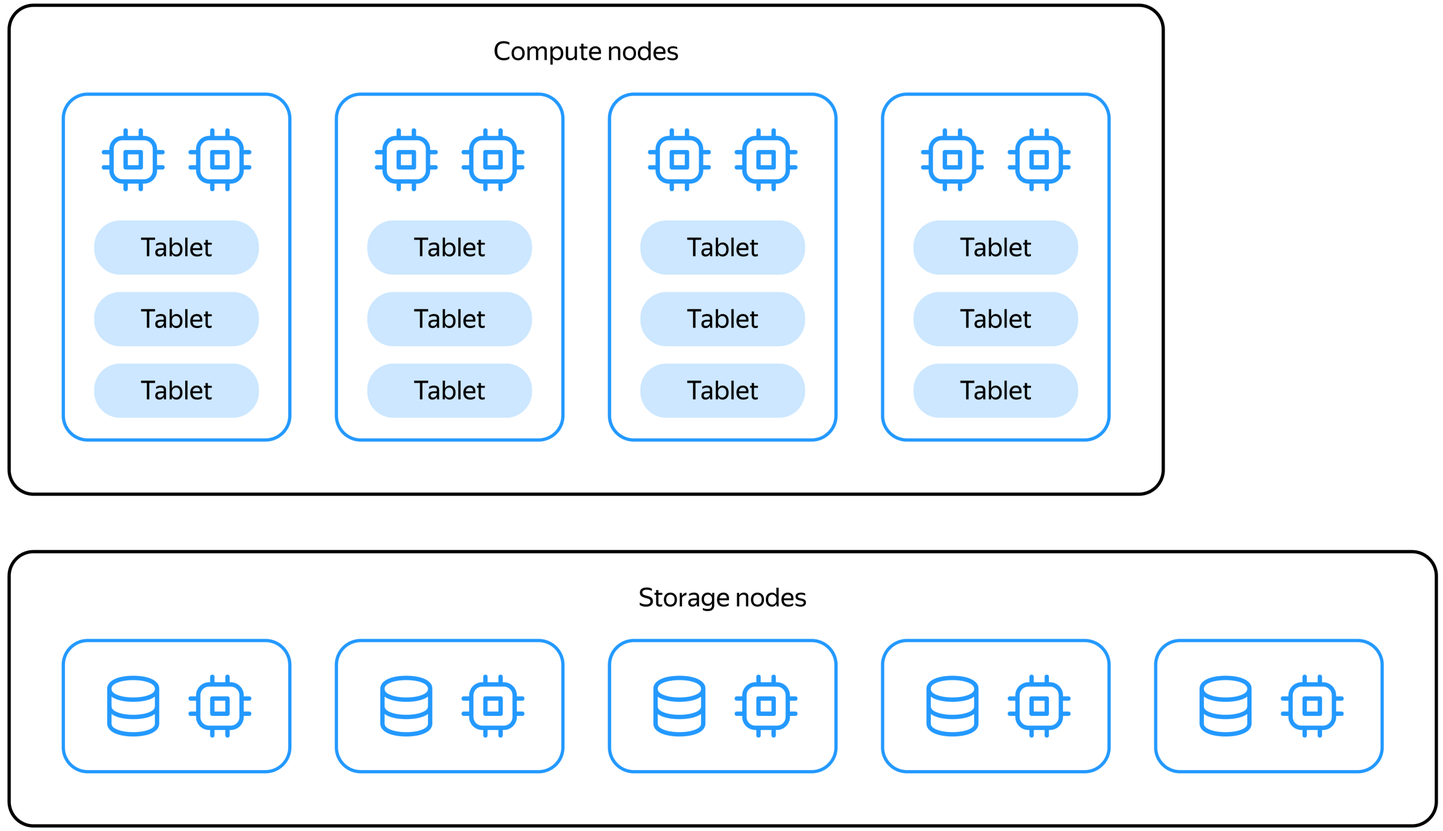
- Основной сценарий — поставка данных реального времени и журналов работы приложений
- Гарантии at least once
- Дешевле относительно кросс-ДЦ
- Модуль управления потоками данных координирует нагрузку на кластеры, управляя клиентом





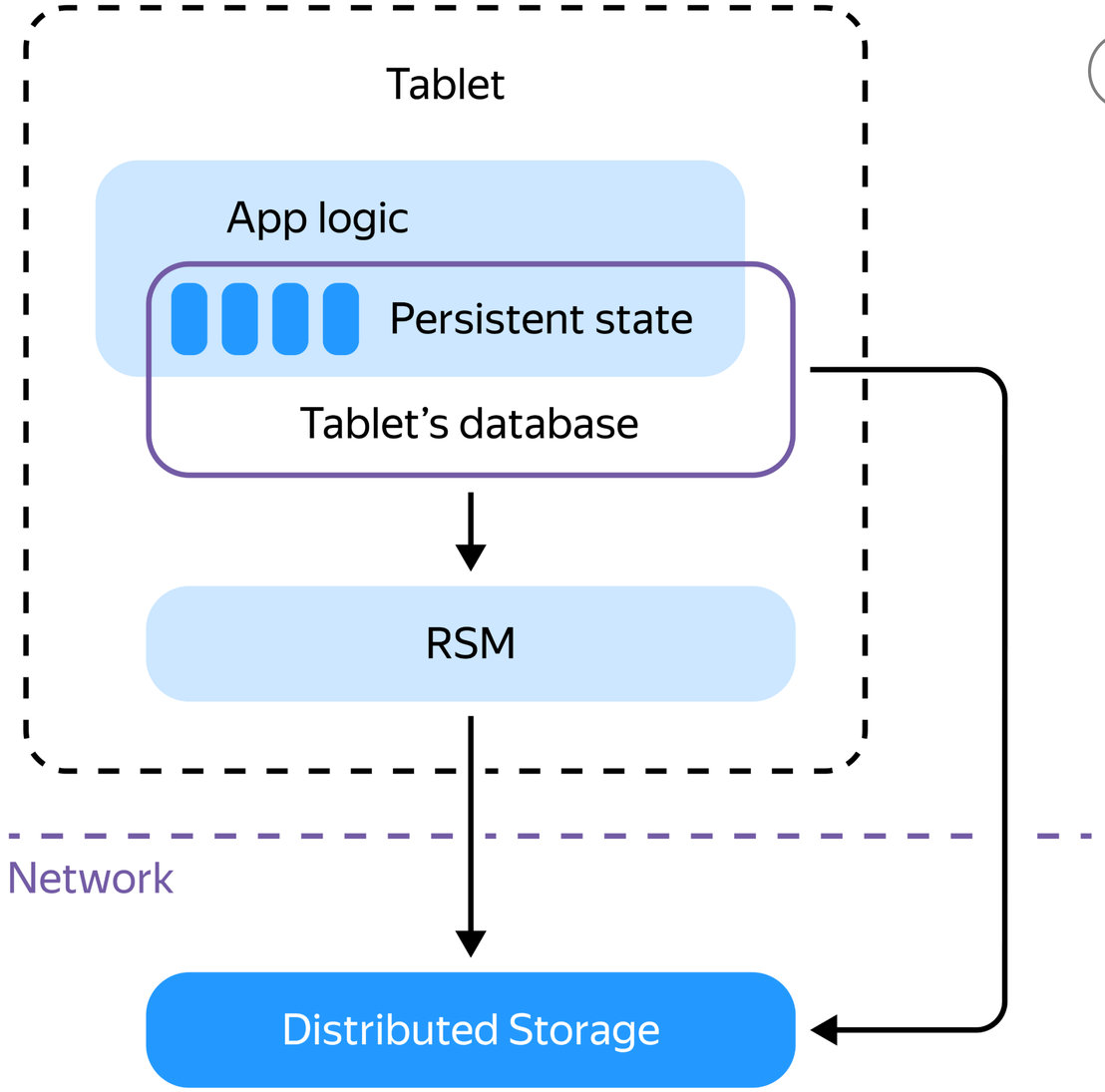
Разделение слоёв Compute и Storage

- Среды выполнения для таблеток и запросов запущены на вычислительных узлах
- Данные размещены на узлах хранения



Tablet

- Хранит состояние в распределенном хранилище
- Пишет лог изменений в распределенное хранилище
- Может быть запущена на любой вычислительной ноде
- Существует в единственном экземпляре



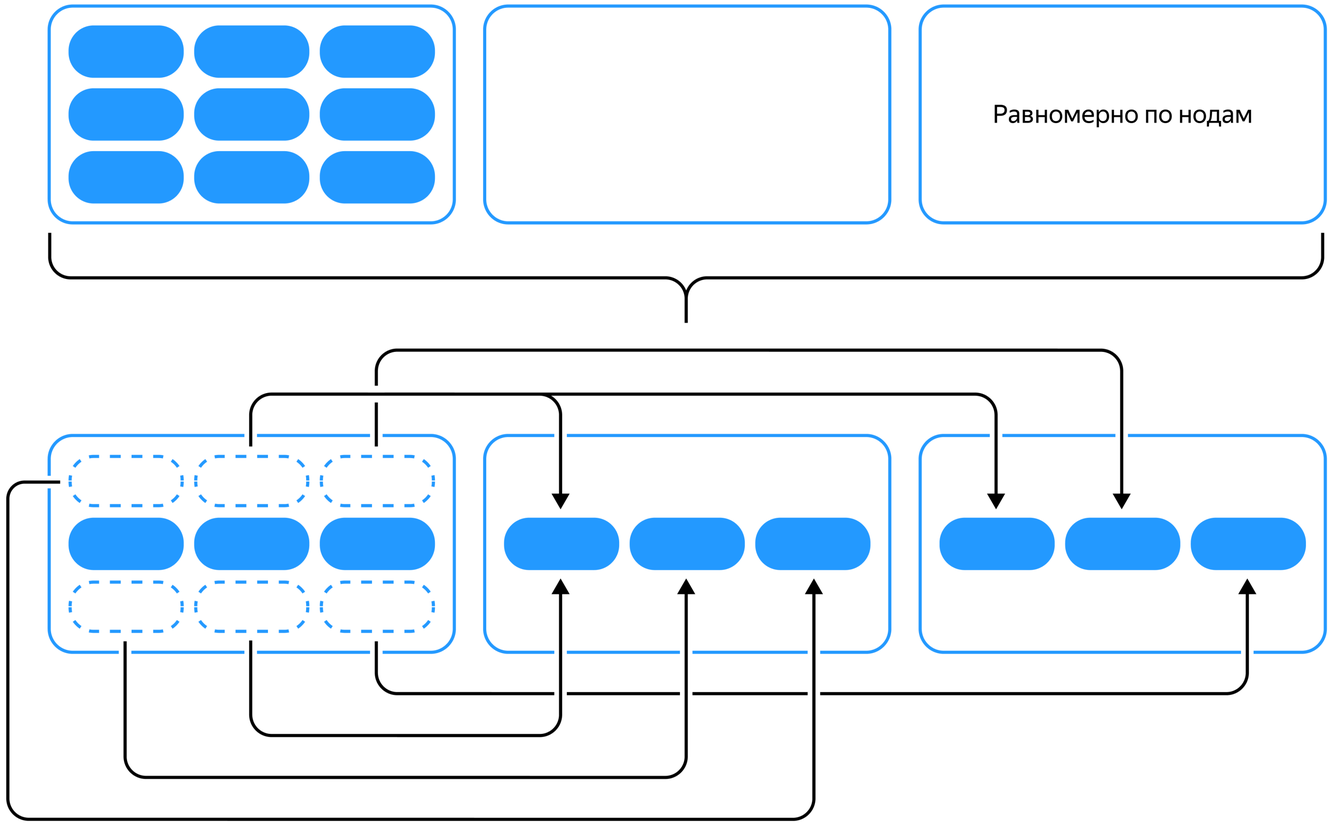
Балансировка таблеток

Балансировка для равномерной утилизации ресурсов кластера

Таблетки свободно перемещаются между узлами кластера

Клиентская балансировка сессий

- Смотрим, чтобы не было пустых нод
- Стараемся распределить нагрузку равномерно (таблетки одной таблицы будут по возможности распределены по разным нодам)



Устойчивость к сбоям

- Постоянная ребалансировка партиций для равномерной нагрузки
- **20+ ms** реакция на потерю сервера

Обновление кластера

- Нет возможности получить временный кластер в каждой зоне доступности на сотни серверов на время обновления
- Обновление производится под полной нагрузкой при регулярной активности пользователей в рабочее время
- Все версии совместимы между соседними версиями с возможностью наката-отката

Управление кластером

- Коммунальный или выделенный кластер
- Ресурсная модель
- Квотирование ресурсов
- Аудит

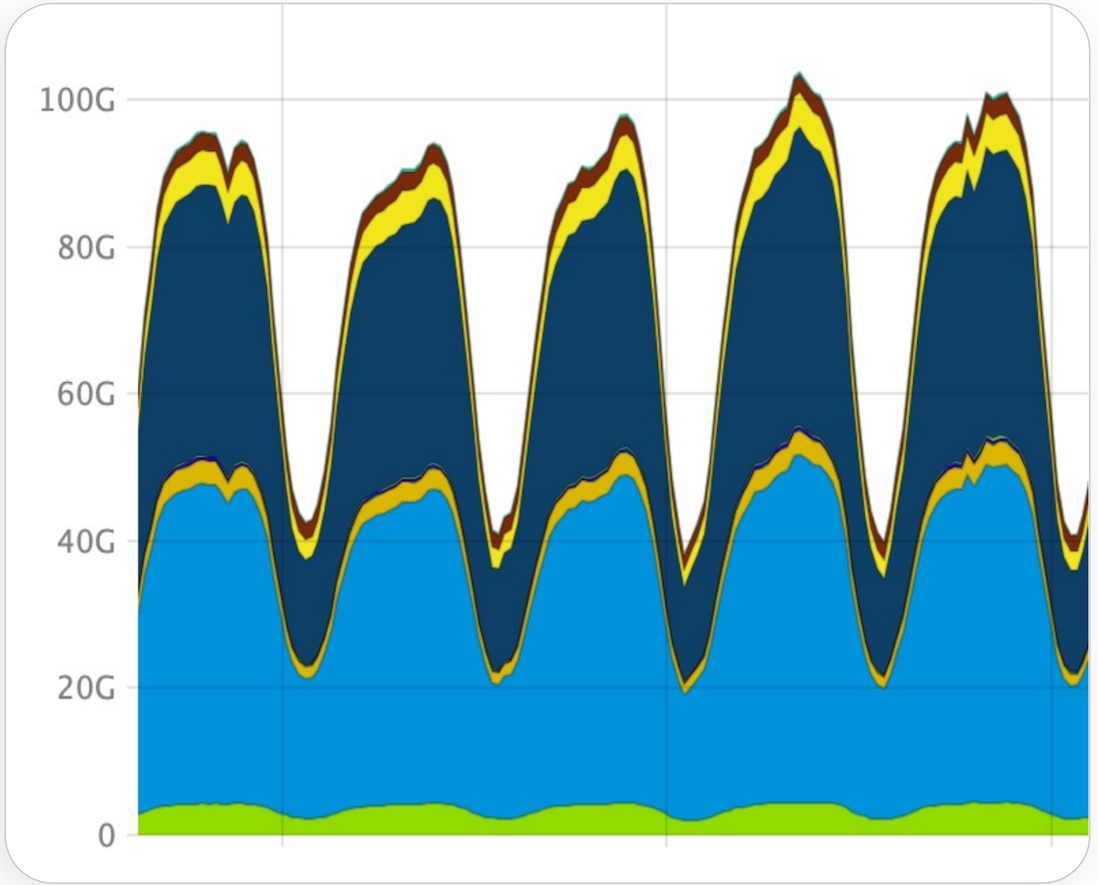
Коммунальный кластер

~100 ГБ/с

используется в ДЦ-1 при 359 GB/s выданных квот

70%

оборудования экономит коммунальный кластер



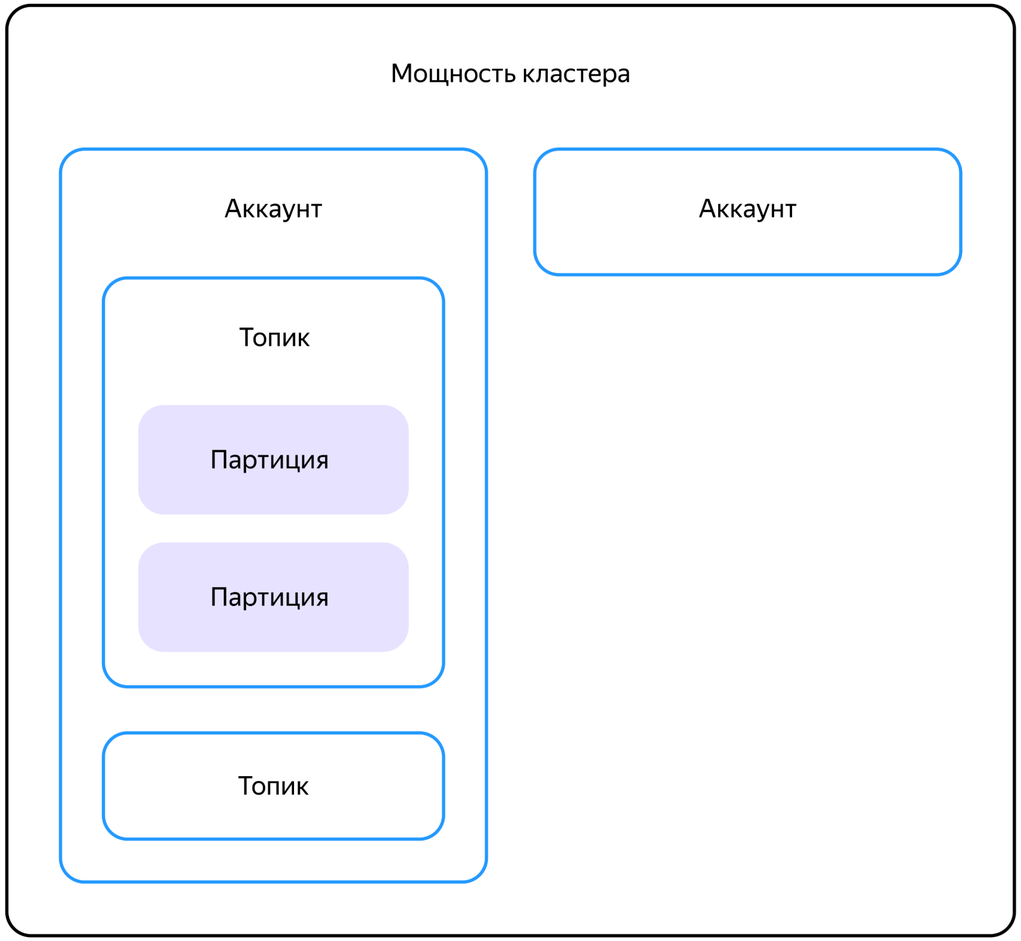
Один кластер или много

Критические системы создают выделенные кластера YDB Topics для снижения blast radius

	Один общий кластер	Выделенные кластеры
Эффективность оборудования	Высокая	Средняя
Надежность	Средняя	Выше средней
Расширение кластера	Легкое	Сложное
Стоимость единицы мощности	Низкая	Высокая

Ресурсная модель

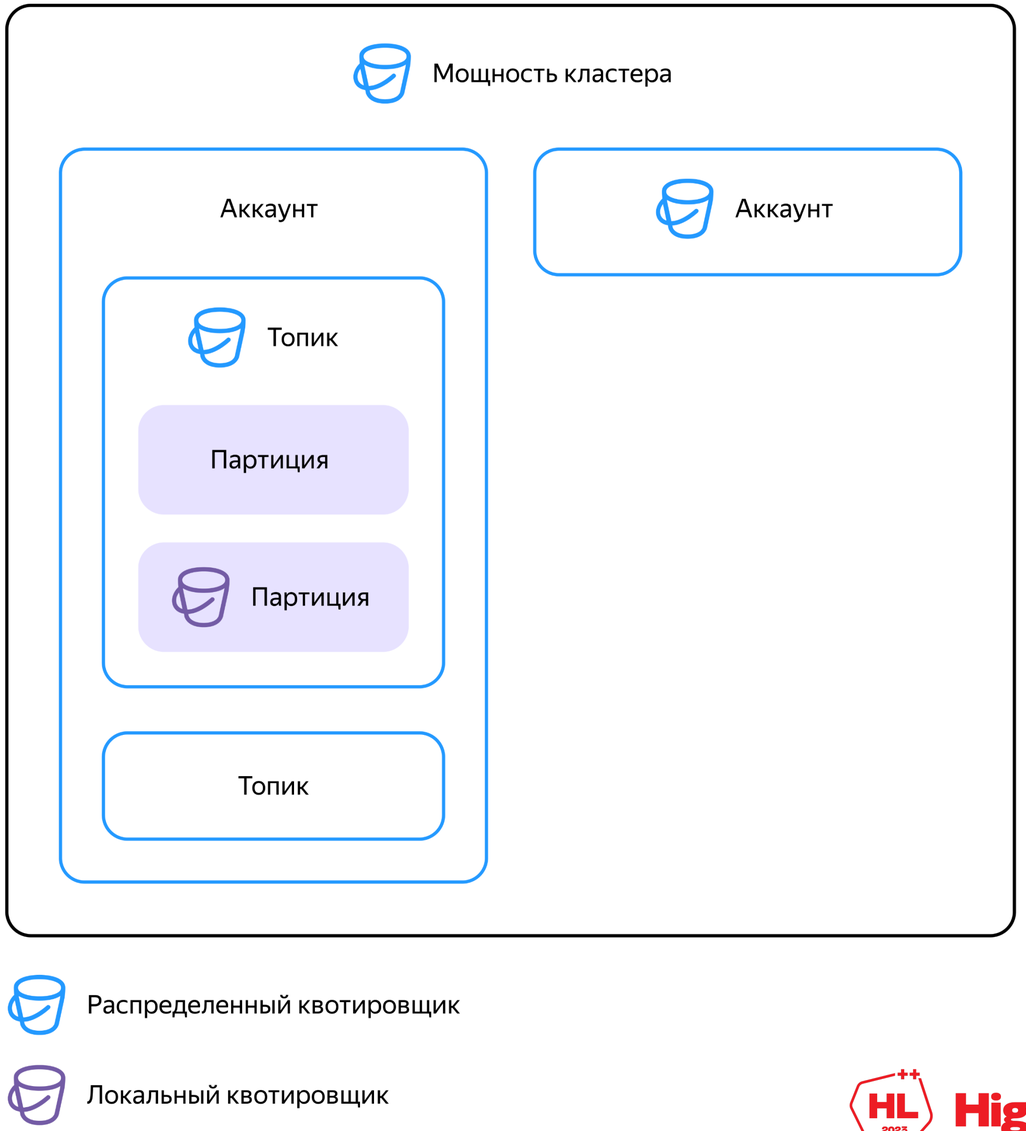
- Аккаунт. Квота пропускной способности и объема хранимых данных в инсталляции (федерация, кросс-ДЦ). Интегрированы с ИС Яндекса
- Топик. Выполняет передачу данных в пределах лимита пропускной способности и объема хранимых данных
- Партиция. Единица масштабируемости топиков



Квотирование

Rate limiting на

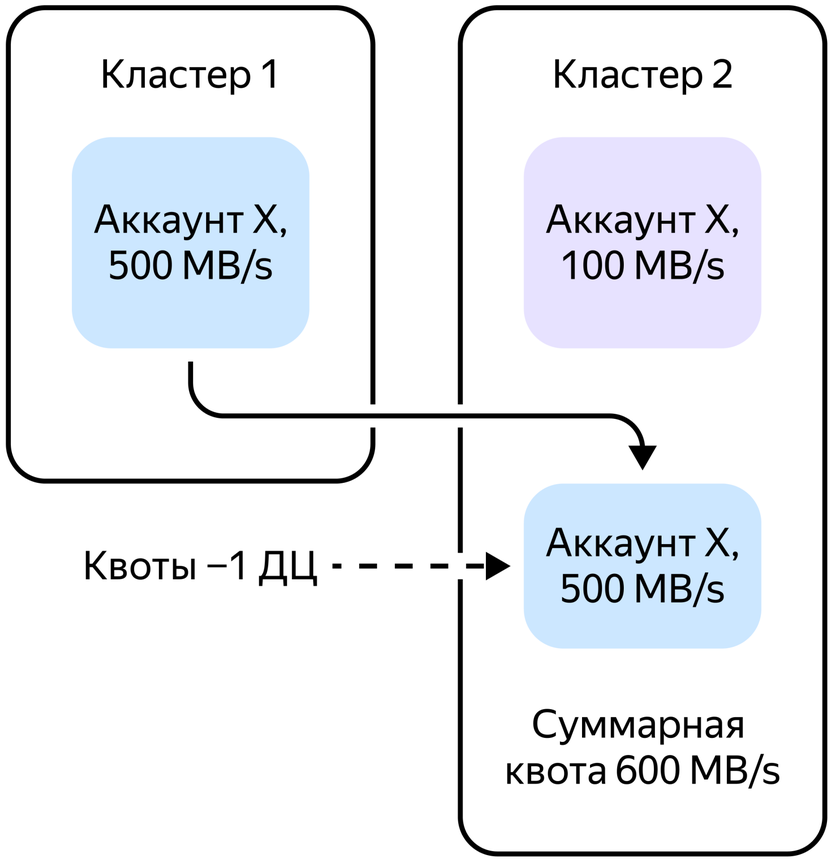
- Запись во всех аккаунты
- Скорость записи и объем хранения в аккаунт в пределах кластера
- Запись в топик в пределах кластера
- Запись в партицию
- Чтение из партиции



Квоты −1 ДЦ

Отдельный вид квот

- Аналогичны стандартным квотам
- Выдаются в случае потери ДЦ
- Автоматически прибавляются к квоте в Аккаунте в зависимости от конфигурации



Аудит

1

Полная детализация сессий чтений и записи

2

Запись: Topic, партиция, IP, producer, метаданные

3

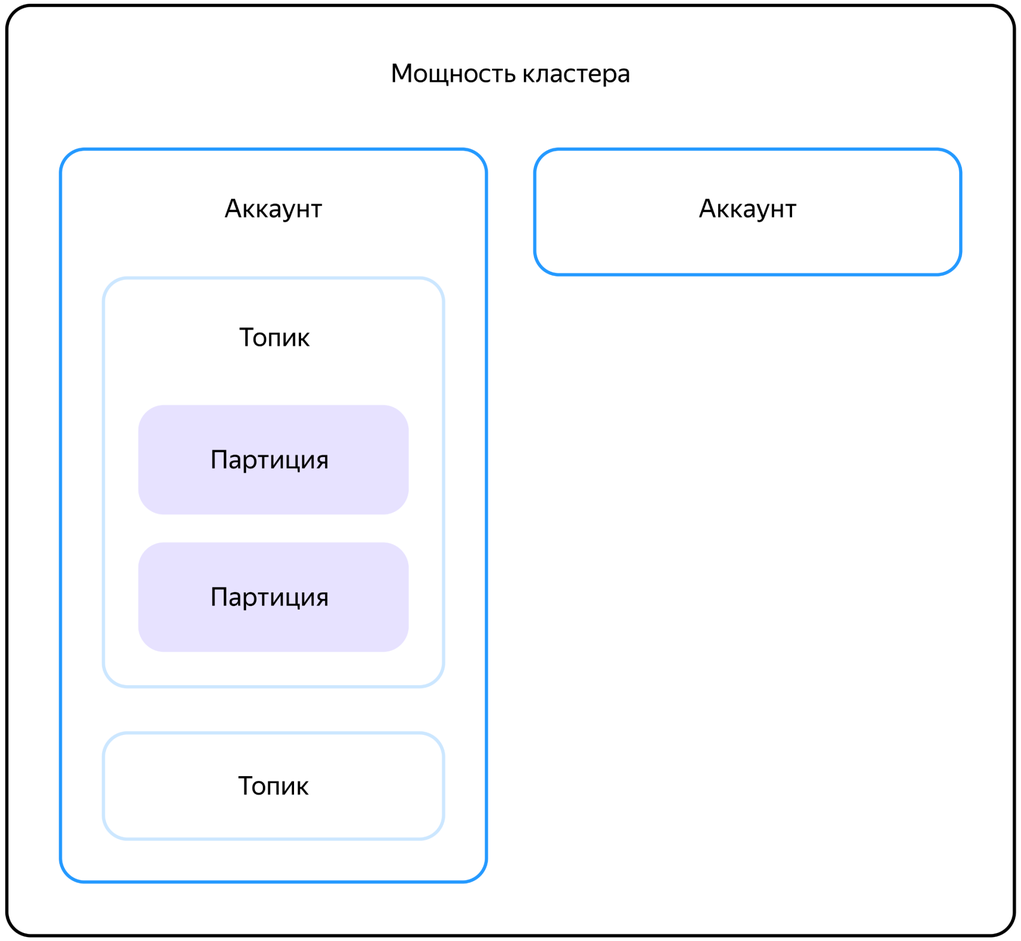
Чтение: Topic, IP, consumer, метаданные

Self-service

- Виды действий, выполняемых пользователями
- Диагностика и мониторинг

Self-service

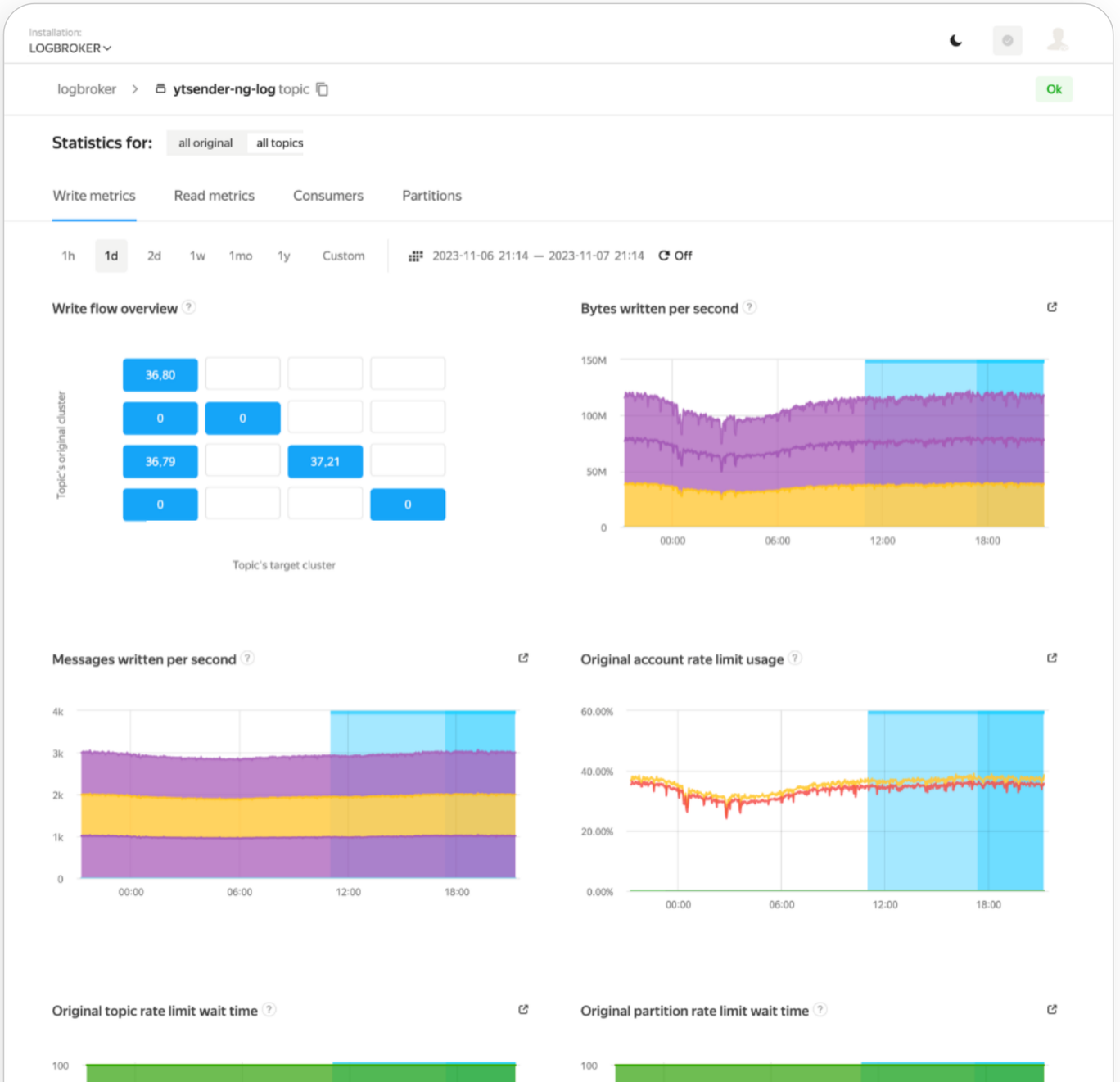
- Аккаунты заводятся автоматически за счет интеграции с ИС Яндекса
- Топиками, партициями, consumer’ами пользователи управляют самостоятельно



Диагностика и мониторинг

Графики

- Скорости записи-чтения с распределением по ДЦ
- Распределения задержек по ДЦ, по consumer’ам, по партициям и т. д.
- Числа сессий чтения-записи



Диагностика и мониторинг

Пользователи настраивают
алерты поверх графиков
и самостоятельно реагируют

Большинство решений Open Source изначально не создаются для больших компаний

1

Только self service спасает от перегрузки поддержкой оборудования и пользователей

2

Нужно квотировать все, на масштабе все становится хуже и заметнее

3

Нужно автоматизировать все, на масштабе все становится хуже и заметнее

YDB Topics

1

Распределенная шина данных для данных любого объема

2

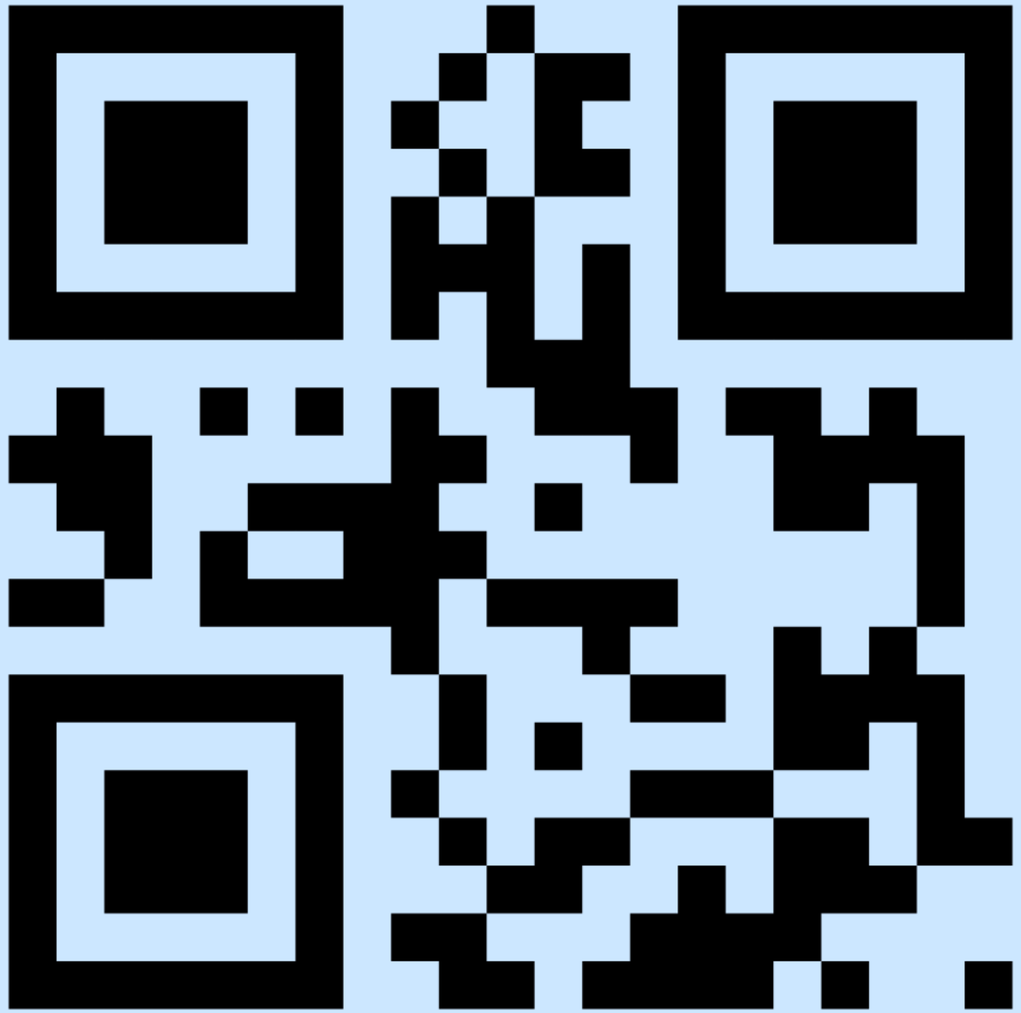
Высоконадежная и высокодоступная

3

С поддержкой протокола Apache Kafka

Платформа YDB

- OLTP, OLAP, Топики; Потоковая, Федеративная обработка данных
- Open Source с лицензией Apache 2.0
- Горизонтальное масштабирование на тысячи узлов
- Автоматические отказоустойчивость и катастрофоустойчивость
- Доступно в Yandex Cloud в виде Managed/Serverless YDB, Yandex Data Streams, Yandex Query



ydb.tech

Голосуйте за мой доклад

Алексей Дмитриев,
технический менеджер,
Яндекс

HL 2023 HighLoad ++