Внедрение YDB CDC на примере Yandex Monitoring

Егор Литвиненко

Yandex Infrastructure

Познакомимся

- Егор Литвиненко
- Работаю в Yandex Infrastructure
- Старший разработчик
  Yandex Observability Platform
- Помогаю людям мониторить их сервисы
- В коммерческой разработке 10 лет

  Lifestreet, Marketshare, Runawfe, Мультикарта,
  АйТи Консалтинг, Магнит…

Saint HighLoad++ 2023

Цели доклада

1

Разобраться, как обновлять данные в сервисе

Saint HighLoad++ 2023

Цели доклада

1

Разобраться, как обновлять данные в сервисе

2

Разобраться, как правильно приготовить YDB CDC

Saint HighLoad++ 2023

План

1. Теория
2. Практика использования CDC
3. Что пошло не так
4. Итоги

План

1. Теория
2. Практика использования CDC
3. Что пошло не так
4. Итоги

Yandex Monitoring в цифрах

Yandex Monitoring

700 млн

метрик в секунду записываем

Saint HighLoad++ 2023

Yandex Monitoring

700 млн

метрик в секунду записываем

это →

8,5 GB

в секунду

Saint HighLoad++ 2023

Yandex Monitoring

700 млн

метрик в секунду записываем

это →

8,5 GB

в секунду

Храним

6,4 PB

данных

Saint HighLoad++ 2023

Yandex Monitoring

700 млн

метрик в секунду записываем

это →

8,5 GB

в секунду

Храним

6,4 PB

данных

Храним

8 млн

алертов пользователей

Saint HighLoad++ 2023

Yandex Monitoring

700 млн
метрик в секунду записываем

это →

8,5 GB
в секунду

Храним
6,4 PB
данных

Храним
8 млн
алертов пользователей

Считаем
150 тысяч
алертов в секунду

Saint HighLoad++ 2023

Постановка задачи

Постановка задачи

- Сервис Yandex Monitoring запущен на тысячах нод

Saint HighLoad++ 2023

Постановка задачи

- Сервис Yandex Monitoring запущен на тысячах нод
- Сервису для работы нужны конфиги из таблиц базы

Saint HighLoad++ 2023

Постановка задачи

- Сервис Yandex Monitoring запущен на тысячах нод
- Сервису для работы нужны конфиги из таблиц базы
- Таблицы логически связаны и образуют иерархию

Saint HighLoad++ 2023

Постановка задачи

- Сервис Yandex Monitoring запущен на тысячах нод
- Сервису для работы нужны конфиги из таблиц базы
- Таблицы логически связаны и образуют иерархию
- Меняются редко

Saint HighLoad++ 2023

Постановка задачи

- Сервис Yandex Monitoring запущен на тысячах нод
- Сервису для работы нужны конфиги из таблиц базы
- Таблицы логически связаны и образуют иерархию
- Меняются редко
- Наша цель: чтобы ноды быстро забирали **только изменения**

Saint HighLoad++ 2023

Как мы решали задачу?

Как мы решали задачу?

- Таблицы небольшие

Saint HighLoad++ 2023

Как мы решали задачу?

- Таблицы небольшие
- Скачиваем таблицы целиком по таймеру

Saint HighLoad++ 2023

Как мы решали задачу?

- Таблицы небольшие
- Скачиваем таблицы целиком по таймеру
- Делаем это на каждой ноде

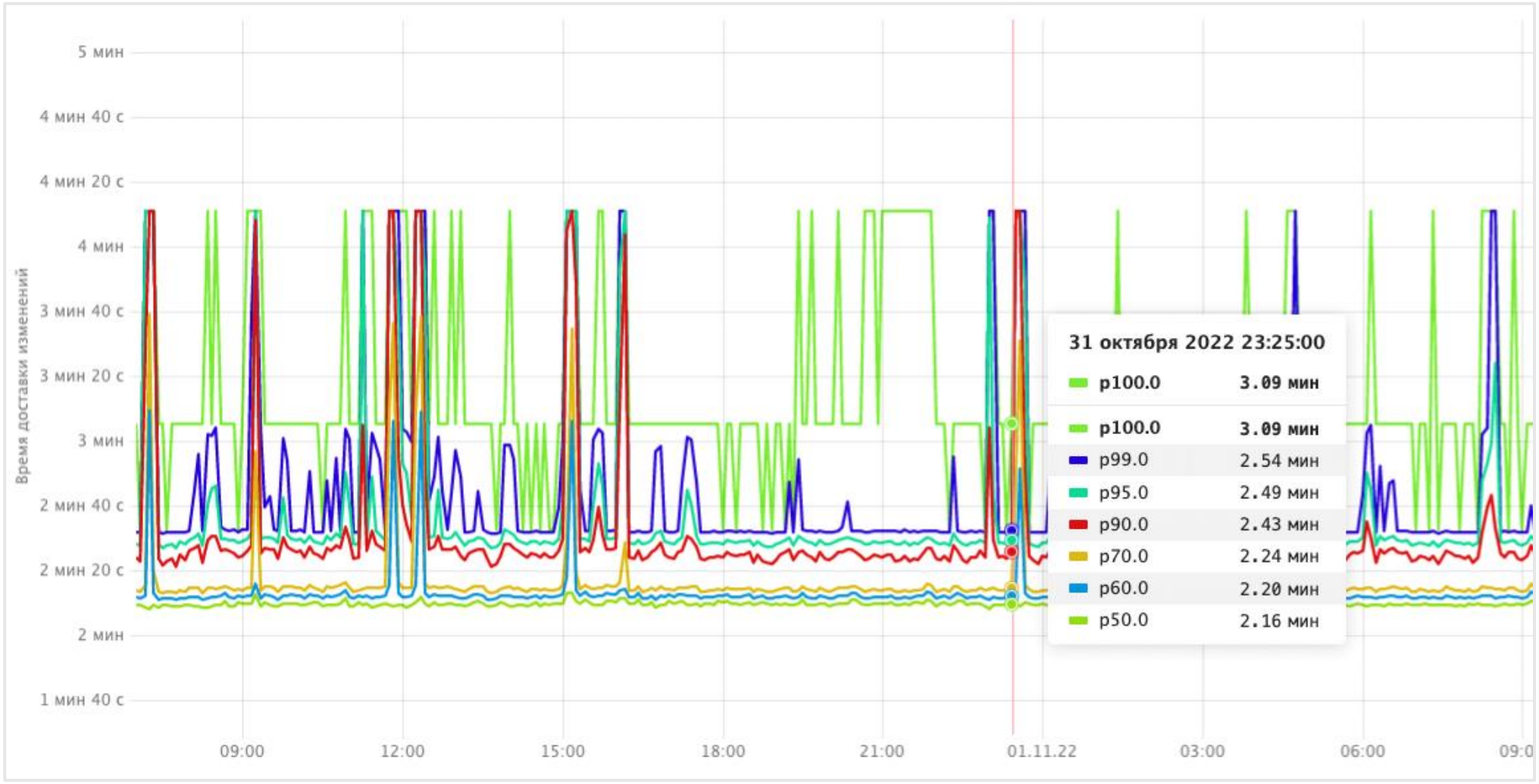
Saint HighLoad++ 2023

Как мы решали задачу?

- Таблицы небольшие
- Скачиваем таблицы целиком по таймеру
- Делаем это на каждой ноде
- Используем механизм **YDB ReadTable (gRPC stream)**

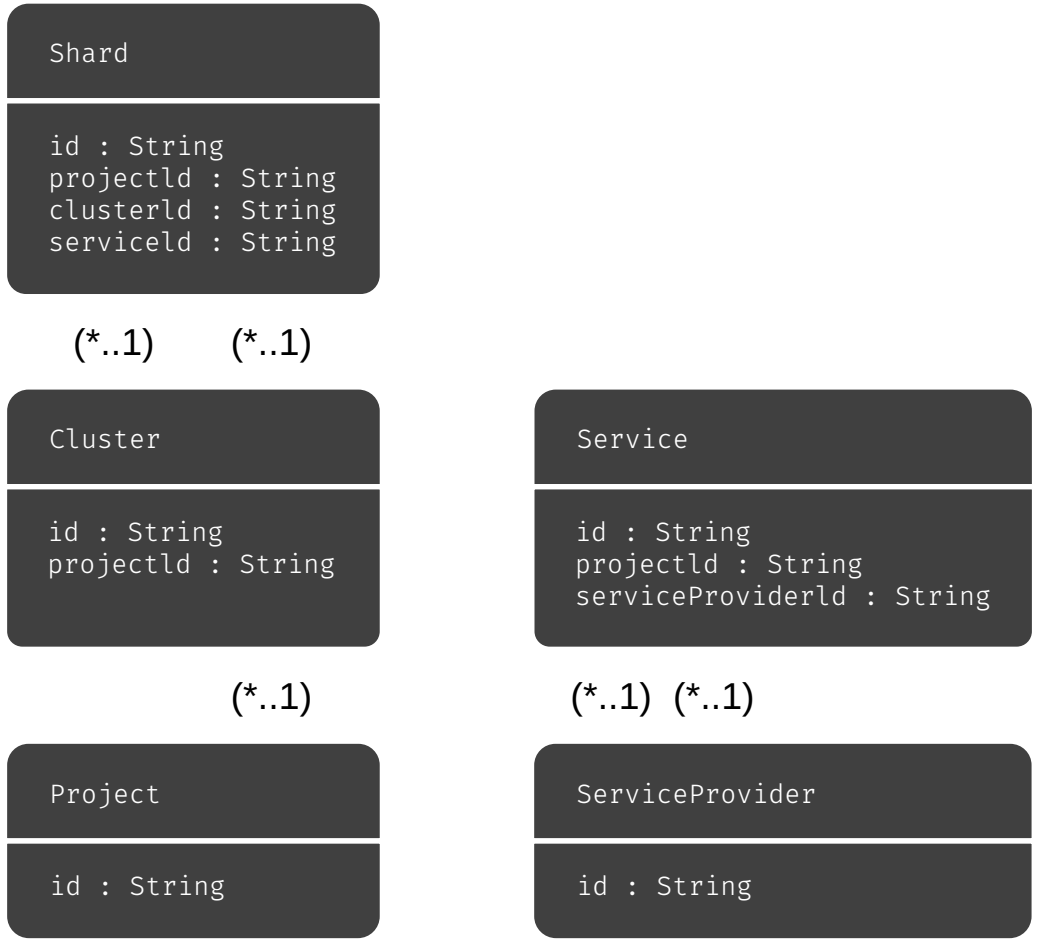
YDB ReadTable
clck.ru/34ccJd

Какой был лаг доставки?



Какие данные мы получаем?

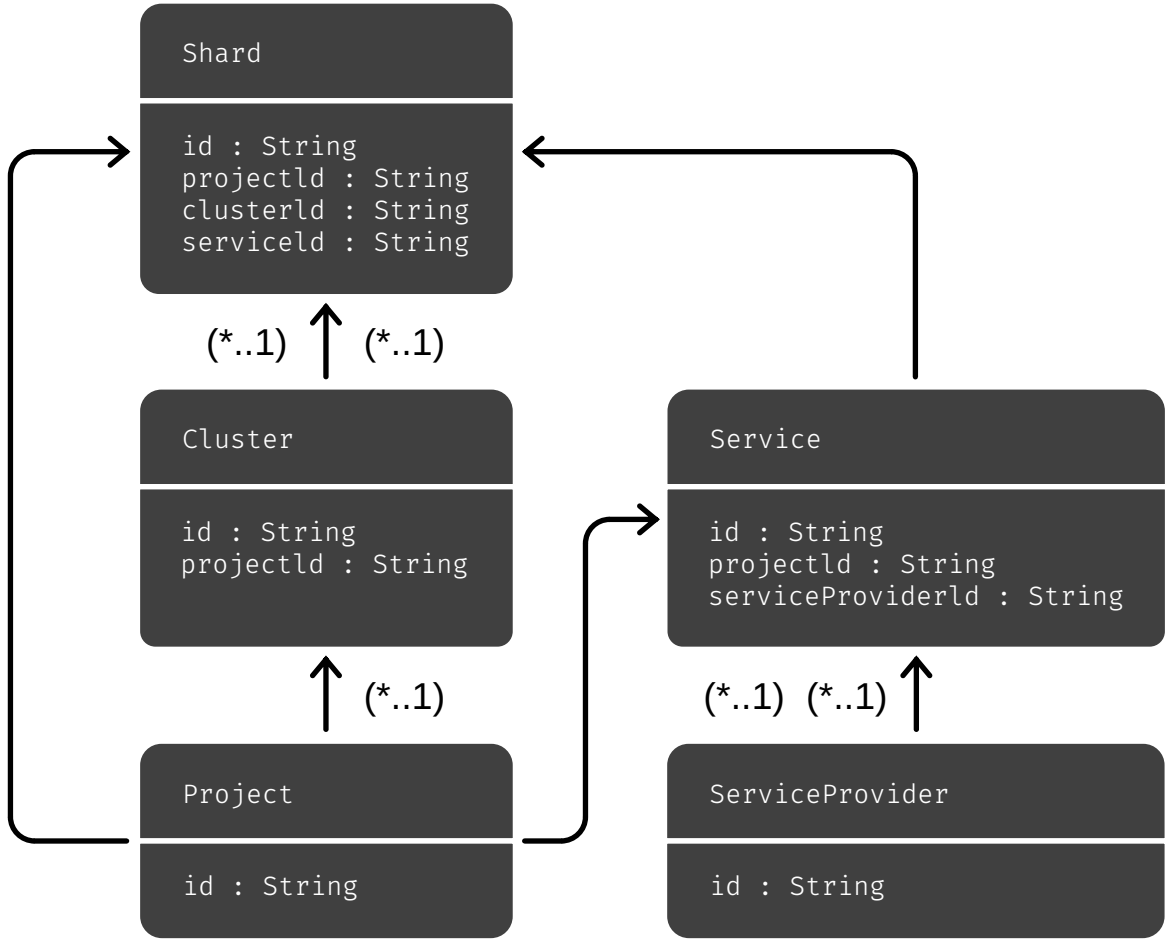
- Таблицы Project, Cluster, …



Saint HighLoad++ 2023

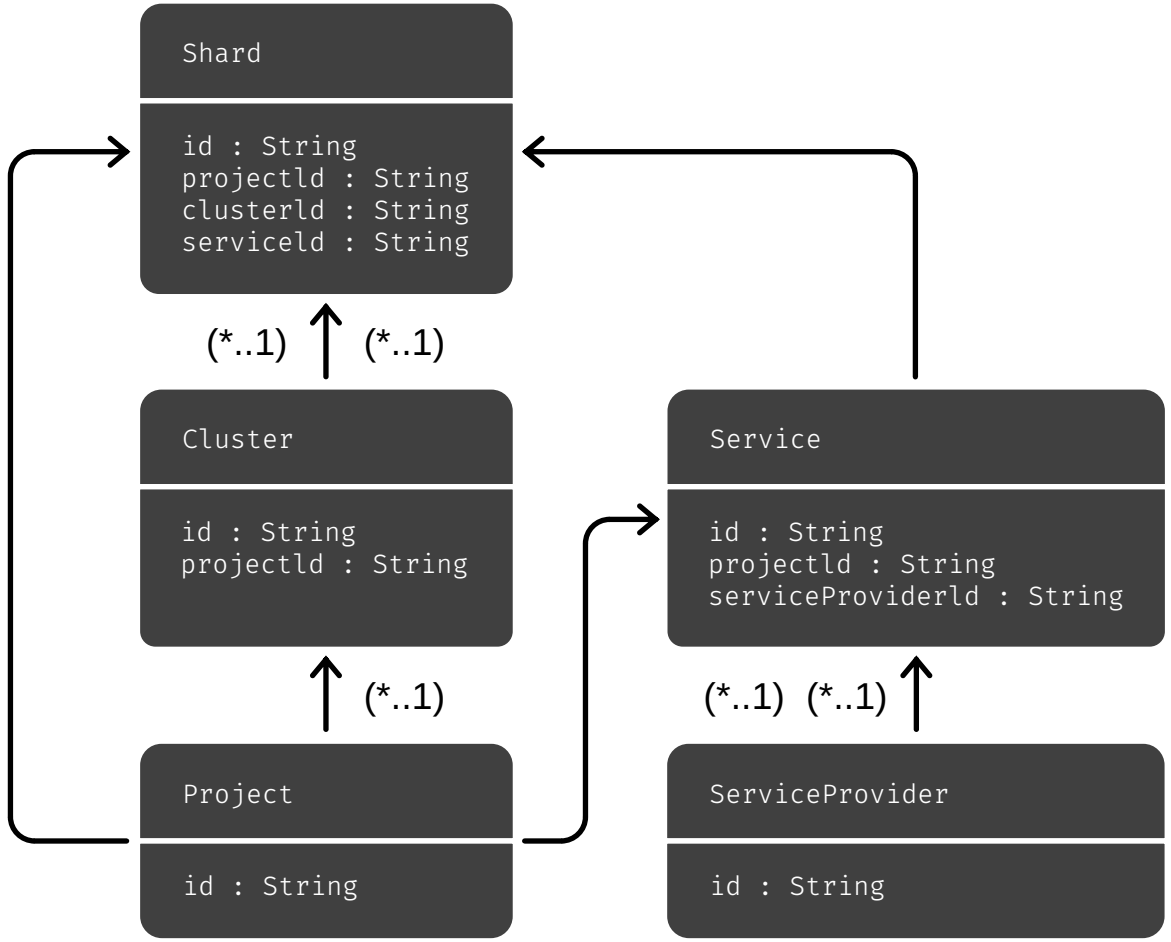
Какие данные мы получаем?

- Таблицы Project, Cluster, …
- Сущности связаны между собой



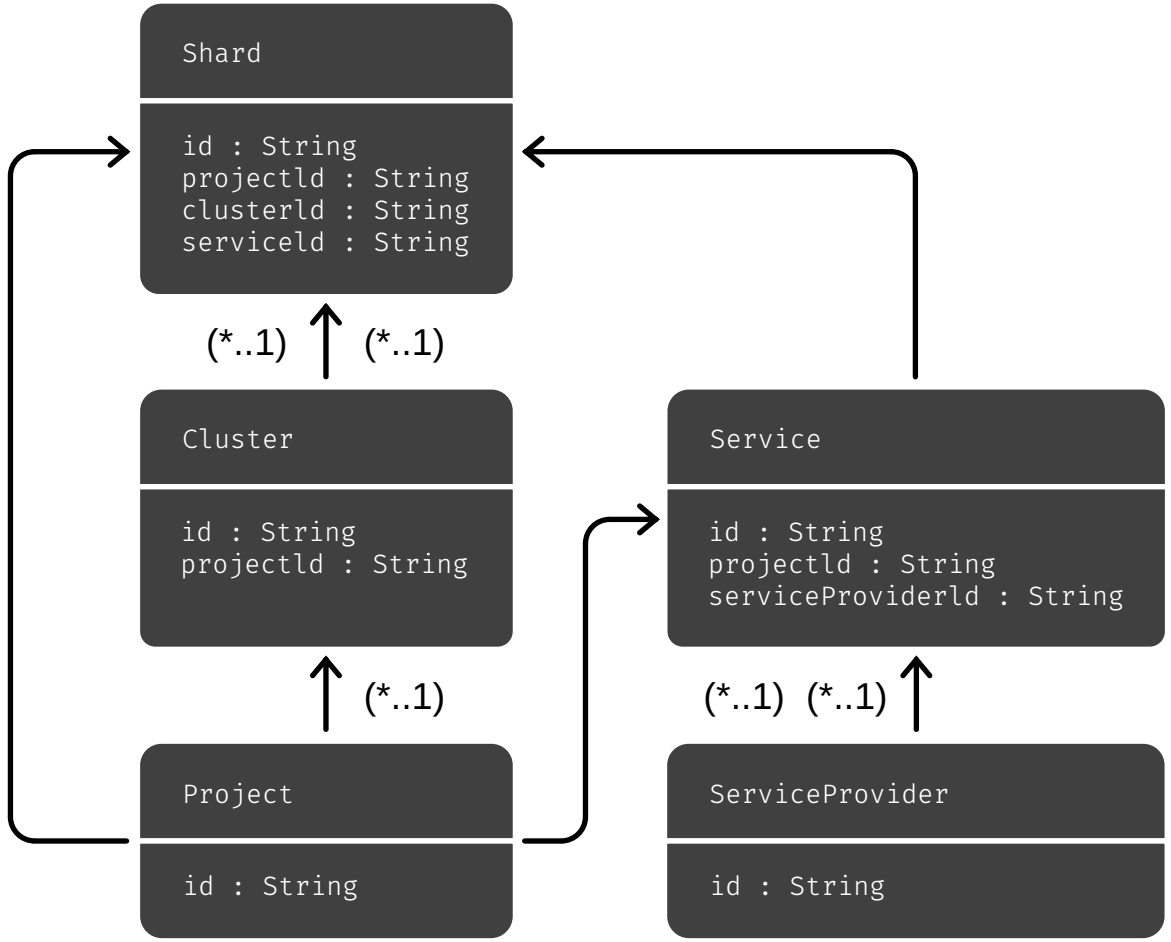
Какие данные мы получаем?

- Таблицы Project, Cluster, …
- Сущности связаны между собой
- В базе foreign keys нет



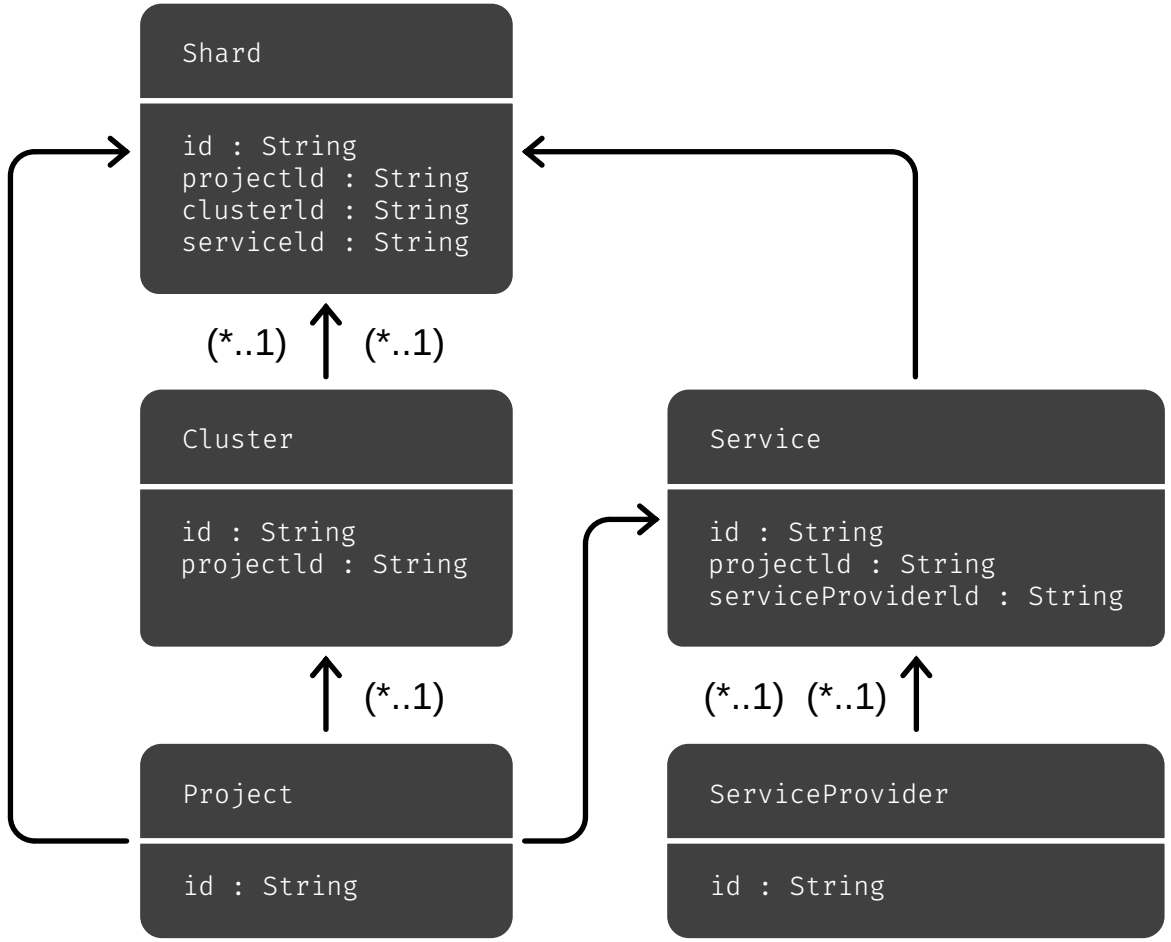
Какие данные мы получаем?

- Таблицы Project, Cluster, …
- Сущности связаны между собой
- В базе foreign keys нет
- Меняются в разных транзакциях



Какие данные мы получаем?

- Таблицы Project, Cluster, …
- Сущности связаны между собой
- В базе foreign keys нет
- Меняются в разных транзакциях
- Становятся только больше



Как брать из таблицы изменения

Как брать из таблицы изменения

- Slowly changing dimensions (SCD)

  Любая база данных

- Triggers on tables

  Любая база данных с триггерами

SCD

clck.ru/34cciJ

SCD Type4


```
SELECT * FROM SHARDS_HISTORY
WHERE SHARD_ID = “123”
AND UPDATED_AT > previous_read_time
ORDER BY UPDATED_AT DESC
LIMIT 1;

CREATE TRIGGER ON SHARDS
INSERT, UPDATE, DELETE …
```

Saint HighLoad++ 2023

SCD Type4


```
SELECT * FROM SHARDS_HISTORY
WHERE SHARD_ID = “123”
AND UPDATED_AT > previous_read_time
ORDER BY UPDATED_AT DESC
LIMIT 1;

// В YDB нет триггеров
CREATE TRIGGER ON SHARDS
INSERT, UPDATE, DELETE …
```

Saint HighLoad++ 2023

Как брать из таблицы изменения

- Slowly changing dimensions (SCD)

  Любая база данных

- Triggers on tables

  Любая база данных с триггерами

- Transaction log scanners

  Binlog в MySQL, Replication log в Postgres, …

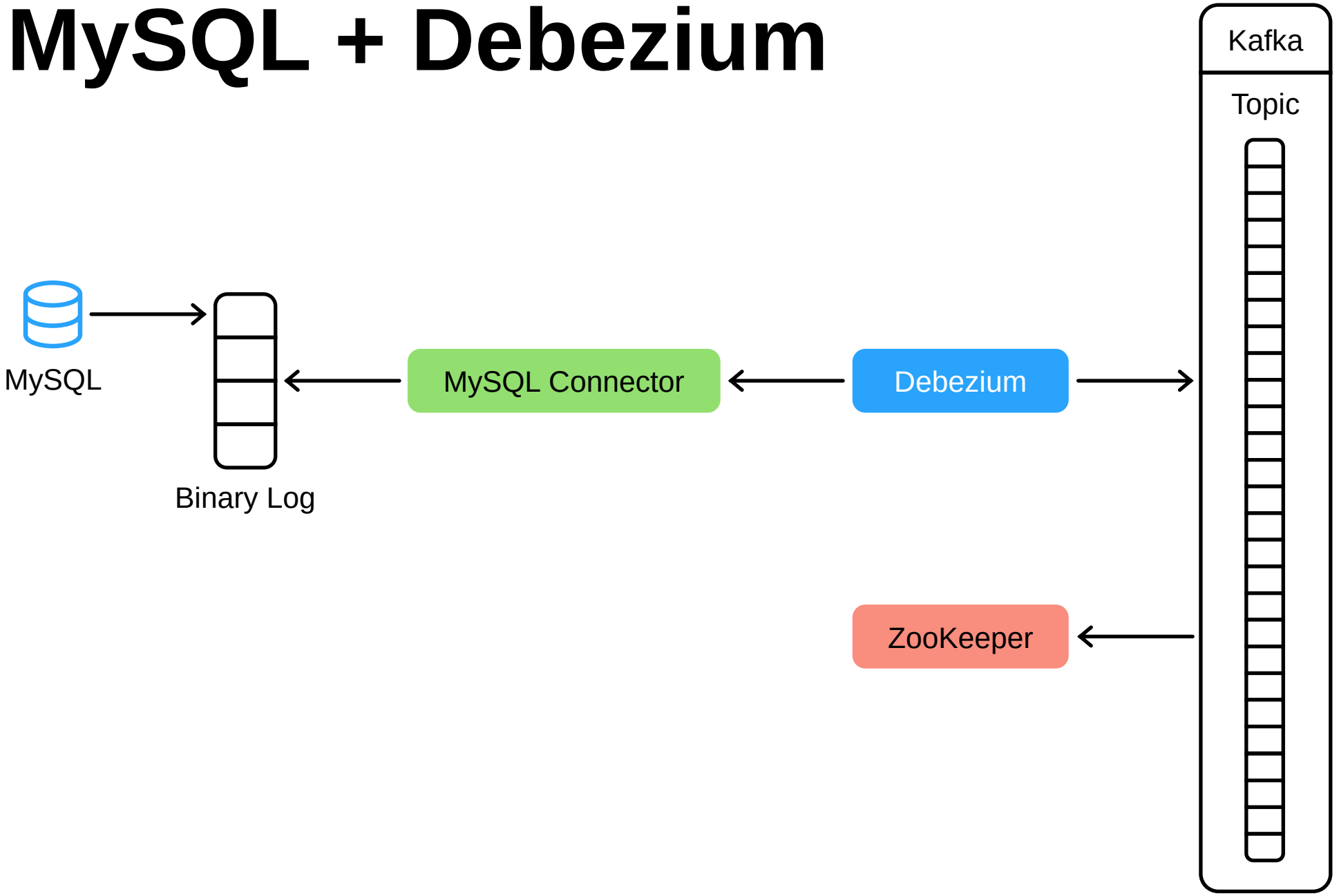
Saint HighLoad++ 2023

MySQL Binlog


```
mysql> SHOW BINARY LOGS;
+---------------+-----------+
| Log_name      | File_size |
+---------------+-----------+
| binlog.000130 |     27459 |
| binlog.000131 |     13719 |
| binlog.000132 |     43268 |
+---------------+-----------+
```

Saint HighLoad++ 2023

MySQL + Debezium



Как брать из таблицы изменения

- Slowly changing dimensions (SCD)

  Любая база данных

- Triggers on tables

  Любая база данных с триггерами

- Transaction log scanners

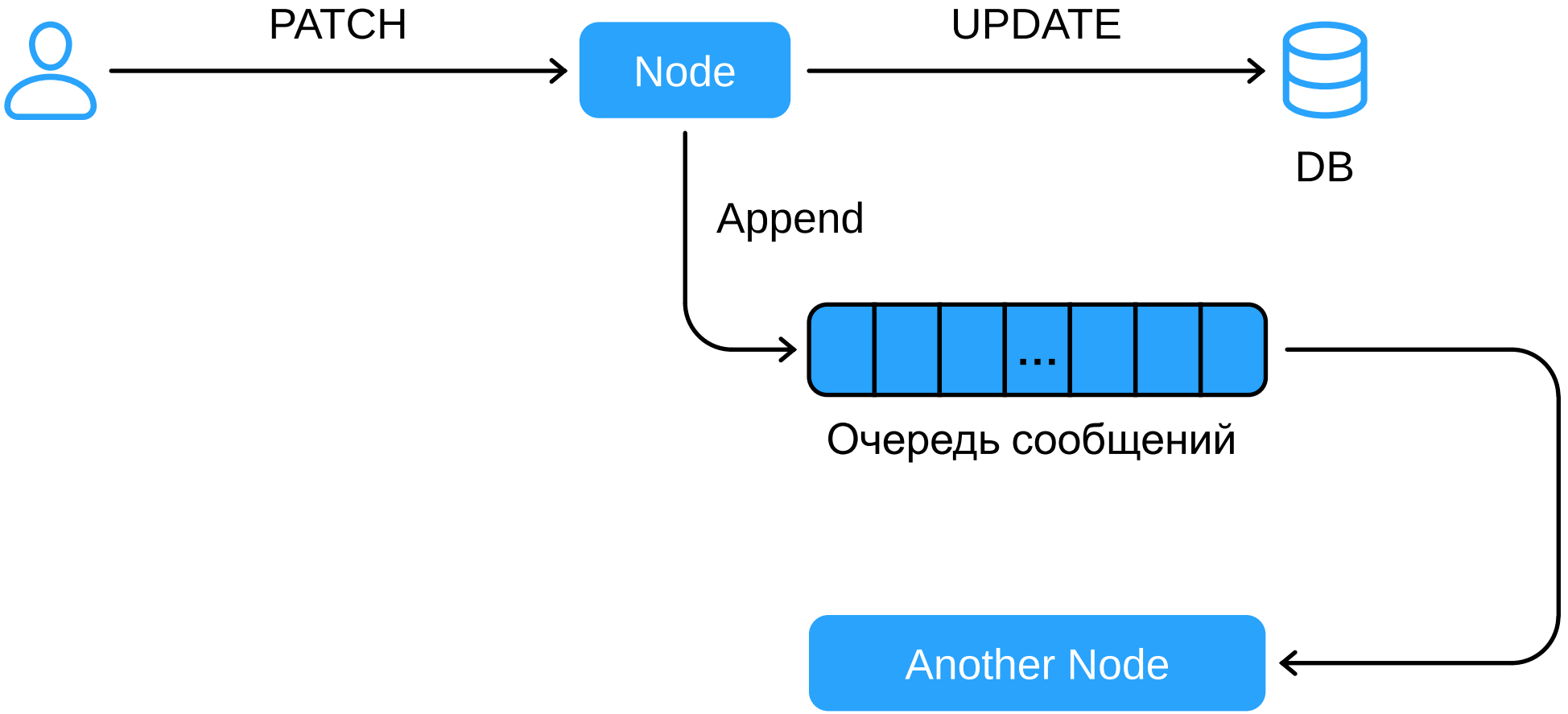
  Binlog в MySQL, Replication log в Postgres, …

- Event source

  Не зависит от базы данных

Saint HighLoad++ 2023

Event source



Как брать из таблицы изменения

- Slowly changing dimensions (SCD)

  Любая база данных

- Triggers on tables

  Любая база данных с триггерами

- Transaction log scanners

  Binlog в MySQL, Replication log в Postgres, …

- Event source

  Не зависит от базы данных

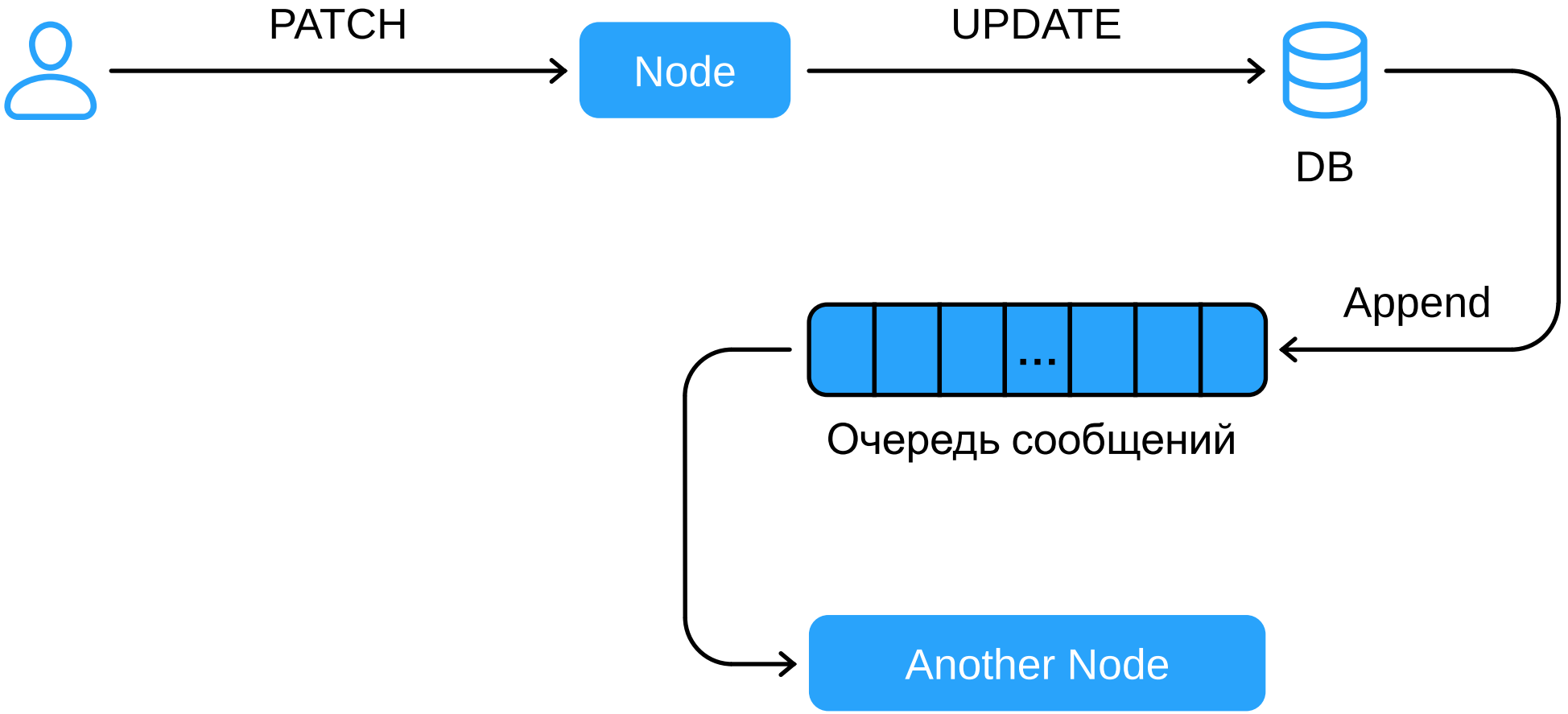
- Change Data Capture (CDC)

  YDB, CockroachDB, RethinkDB, Tarantool…

  Debezium, Databricks Live Table, Zendesk Daemon…

Saint HighLoad++ 2023

CDC


PATCH
Node
UPDATE
DB
Append
...
Очередь сообщений
Another Node

Сравнение подходов

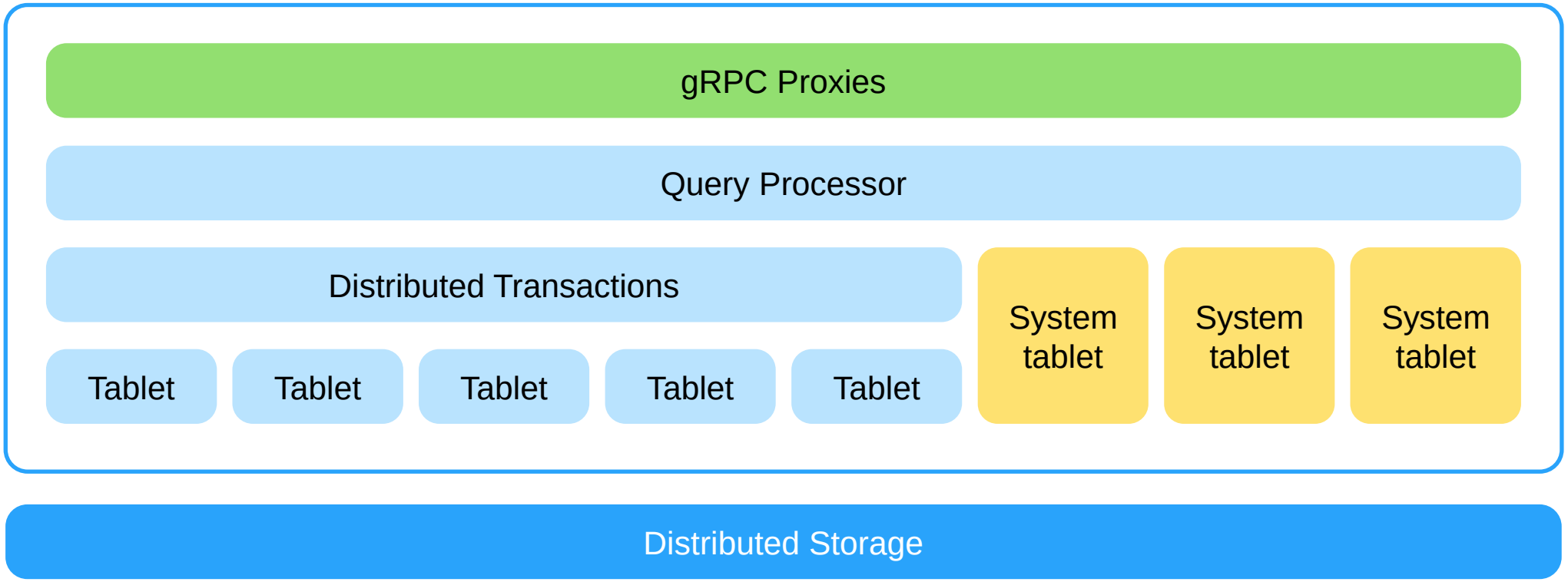
	SCD	Event sourcing	CDC
Писать код	Много	Много	Мало
Избыточность таблиц и индексов	Много	Нет	Нет

Saint HighLoad++ 2023

YDB

- Open-source распределённая база данных

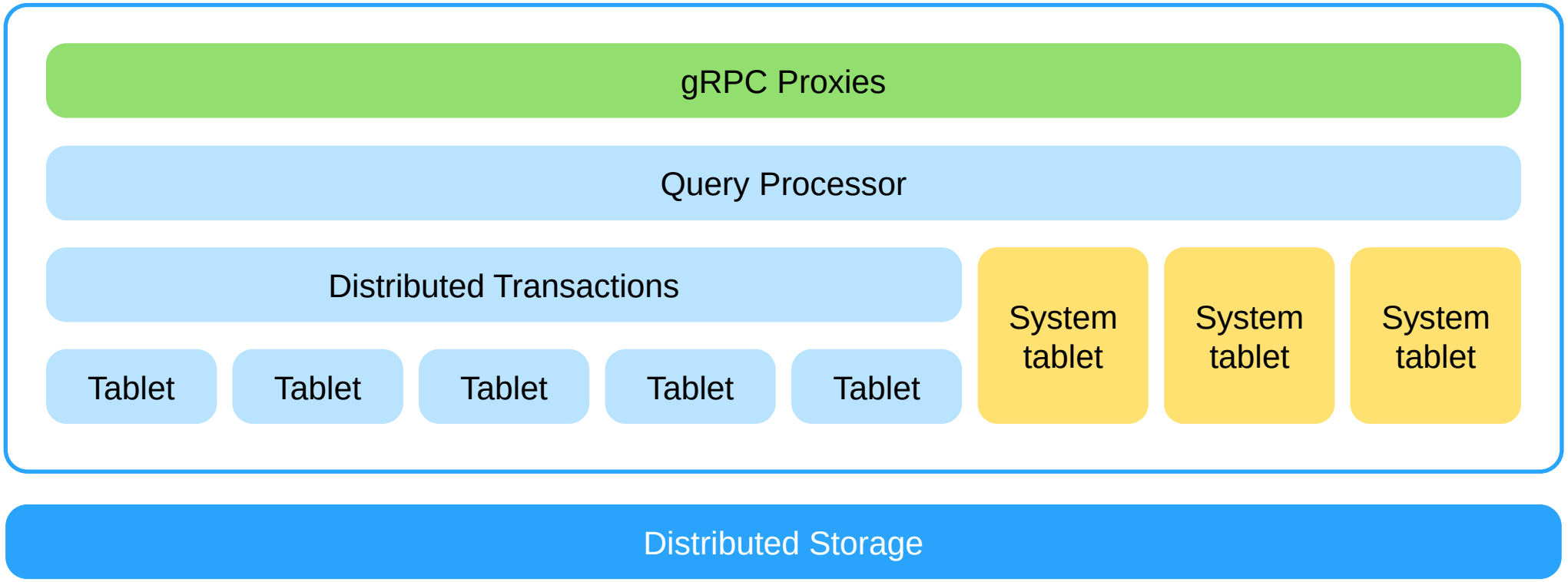
clck.ru/sG59F



YDB

- Open-source распределённая база данных
- Поддерживает очереди

clck.ru/sG59F



YDB CDC

YDB CDC

- Очереди партиционированы так же, как таблицы

Saint HighLoad++ 2023

YDB CDC

- Очереди партиционированы так же, как таблицы
- Все записи в очередь асинхронно

Saint HighLoad++ 2023

YDB CDC

- Очереди партиционированы так же, как таблицы
- Все записи в очередь асинхронно
- Упорядоченность изменений для primary key

Saint HighLoad++ 2023

YDB CDC

- Очереди партиционированы так же, как таблицы
- Все записи в очередь асинхронно
- Упорядоченность изменений для primary key
- Exactly Once

Saint HighLoad++ 2023

YDB CDC

- Очереди партиционированы так же, как таблицы
- Все записи в очередь асинхронно
- Упорядоченность изменений для primary key
- Exactly Once
- TTL по умолчанию 24 часа, максимум 30 дней

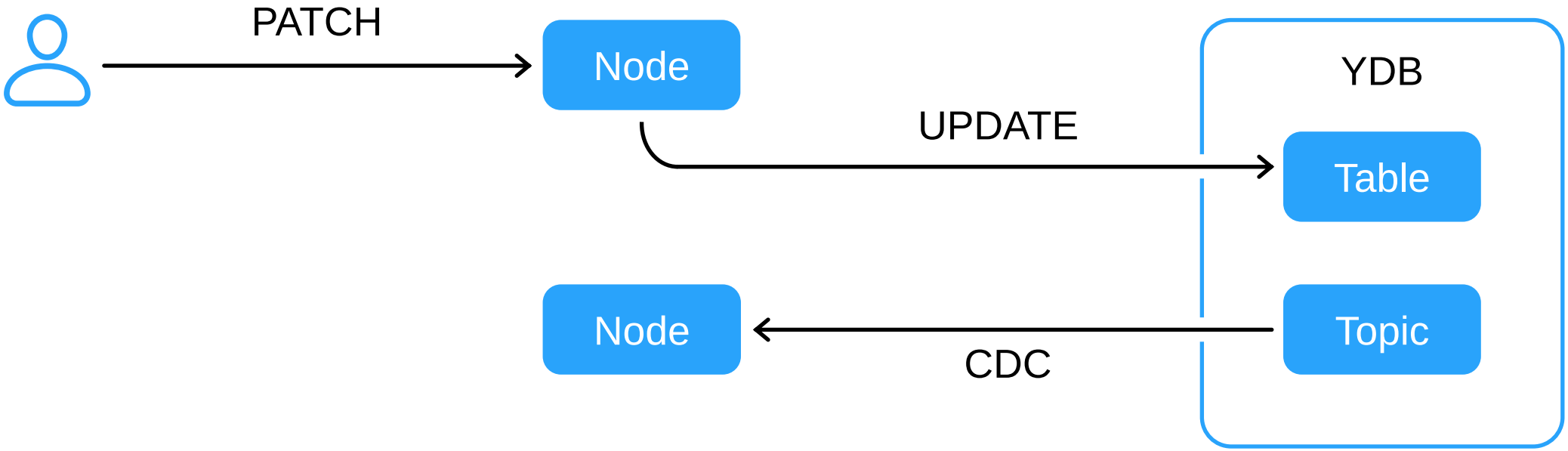
Saint HighLoad++ 2023

Пример сообщения YDB CDC


```
{
  "update": {},
  "newImage": {
    "createdAt": 1654947225226,
    "updatedAt": 1654947225226,
    "maxResponseSizeBytes": 10485760,
    "state": "RW",
    "maxFileSensors": 1000000,
    "protoQuotas": "KGQwsQI=",
       …,
    "maxMemSensors": 500000,
  },
  "key": [
    "composite_key_part1",
    "composite_key_part2"
  ]
}
```


Saint HighLoad++ 2023

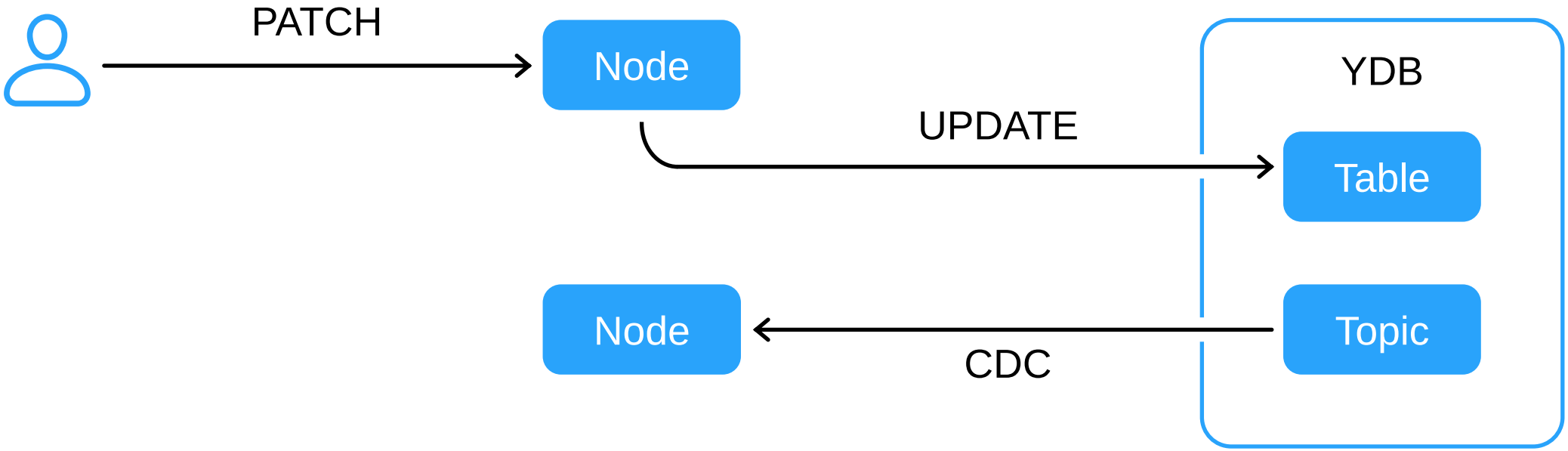
Схема YDB CDC



План

1. Теория
2. Практика использования CDC
3. Что пошло не так
4. Итоги

Схема YDB CDC



Первая загрузка таблицы

Первая загрузка таблицы

- Используем YDB ReadTable

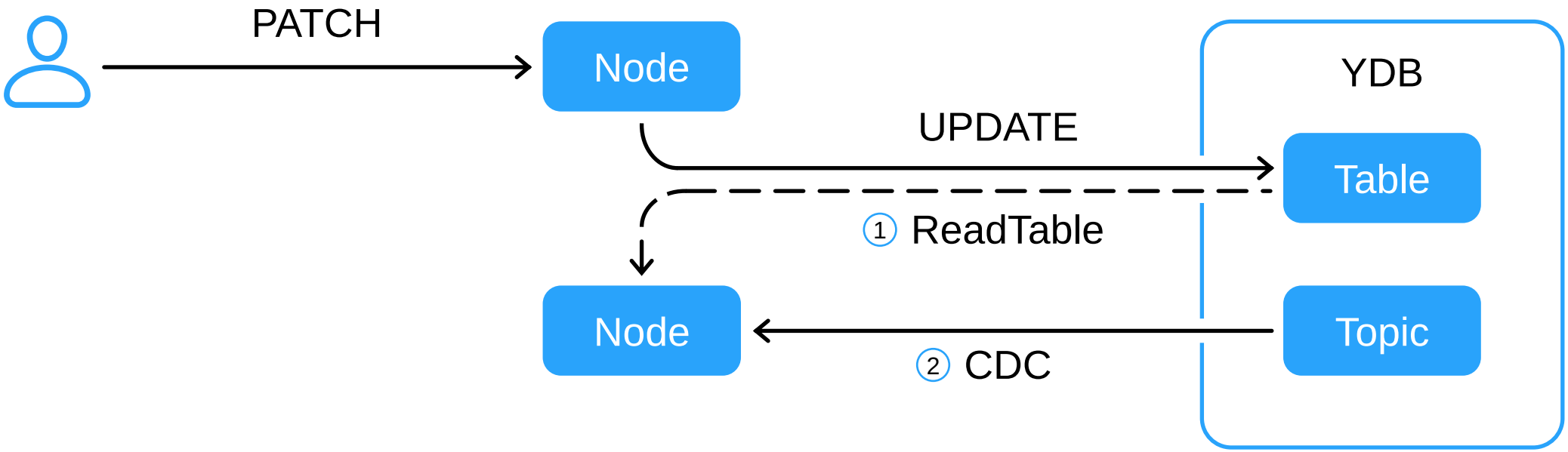
  gRPC stream

YDB ReadTable

clck.ru/34ccJd

Saint HighLoad++ 2023

Схема YDB CDC



Первая загрузка таблицы

- Используем YDB ReadTable

  gRPC stream

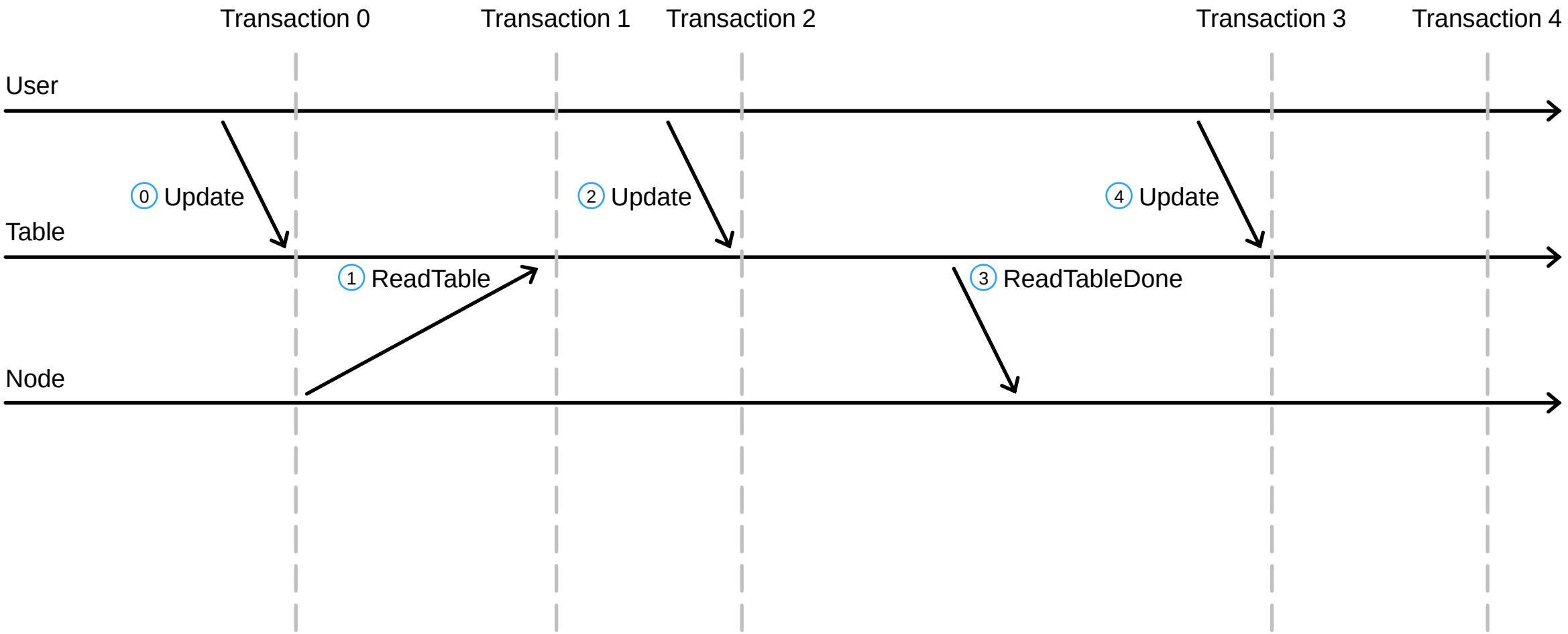
- Скачиваем только на старте

YDB ReadTable

clck.ru/34ccJd

Saint HighLoad++ 2023

Первая загрузка таблицы



Первая загрузка таблицы

- ReadTable читает из снэпшота,
  у которого есть virtual timestamp

virtual timestamp

clck.ru/34d94Q

Первая загрузка таблицы

- ReadTable читает из снэпшота, у которого есть virtual timestamp
- Сообщения тоже содержат virtual timestamp

virtual timestamp

clck.ru/34d94Q

Первая загрузка таблицы

- ReadTable читает из снэпшота, у которого есть virtual timestamp
- Сообщения тоже содержат virtual timestamp
- virtual timestamp:

  Coordinator time, transaction id

virtual timestamp

clck.ru/34d94Q

Первая загрузка таблицы

- ReadTable читает из снэпшота, у которого есть virtual timestamp
- Сообщения тоже содержат virtual timestamp
- virtual timestamp:

  Coordinator time, transaction id

- Можно мержить, фильтровать, сортировать строки

virtual timestamp

clck.ru/34d94Q

Пример virtual timestamp


```
{
  "update": {},
  "newImage": {
    "createdAt": 1654947225226,
    "updatedAt": 1654947225226,
    "maxResponseSizeBytes": 10485760,
    "state": "RW",
    "maxFileSensors": 1000000,
    "protoQuotas": "KGQwsQI=",
        …,
    "maxMemSensors": 500000,
  },
  "key": [
    "composite_key_part1",
    "composite_key_part2"
  ],
  "ts": [1670792400, 562949953607163]
}
```

Получение CDC после ReadTable

Получение CDC после ReadTable

- У очереди нельзя запросить сообщения по virtual timestamp 😢

Saint HighLoad++ 2023

Получение CDC после ReadTable

- У очереди нельзя запросить сообщения по virtual timestamp 😢
- Поэтому запрашиваем по времени снэпшота минус лаг

Saint HighLoad++ 2023

Получение CDC после ReadTable

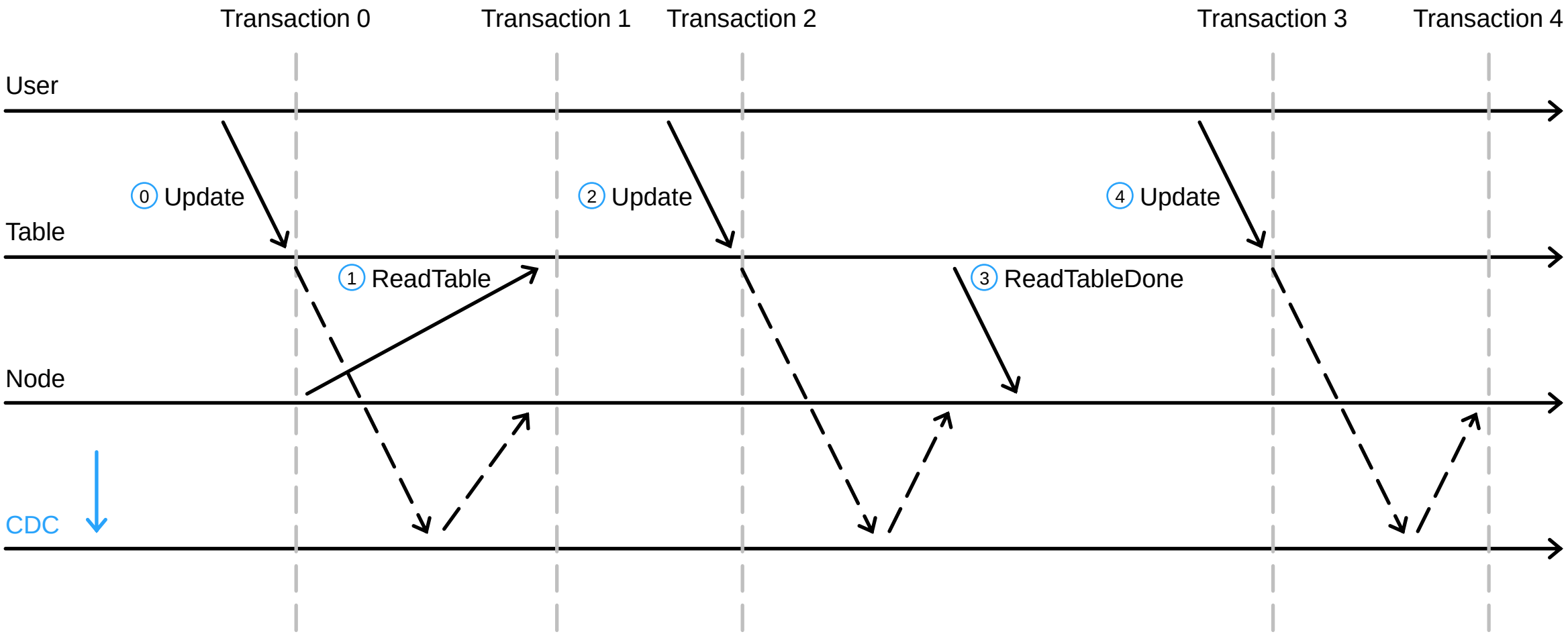


Схема таблиц

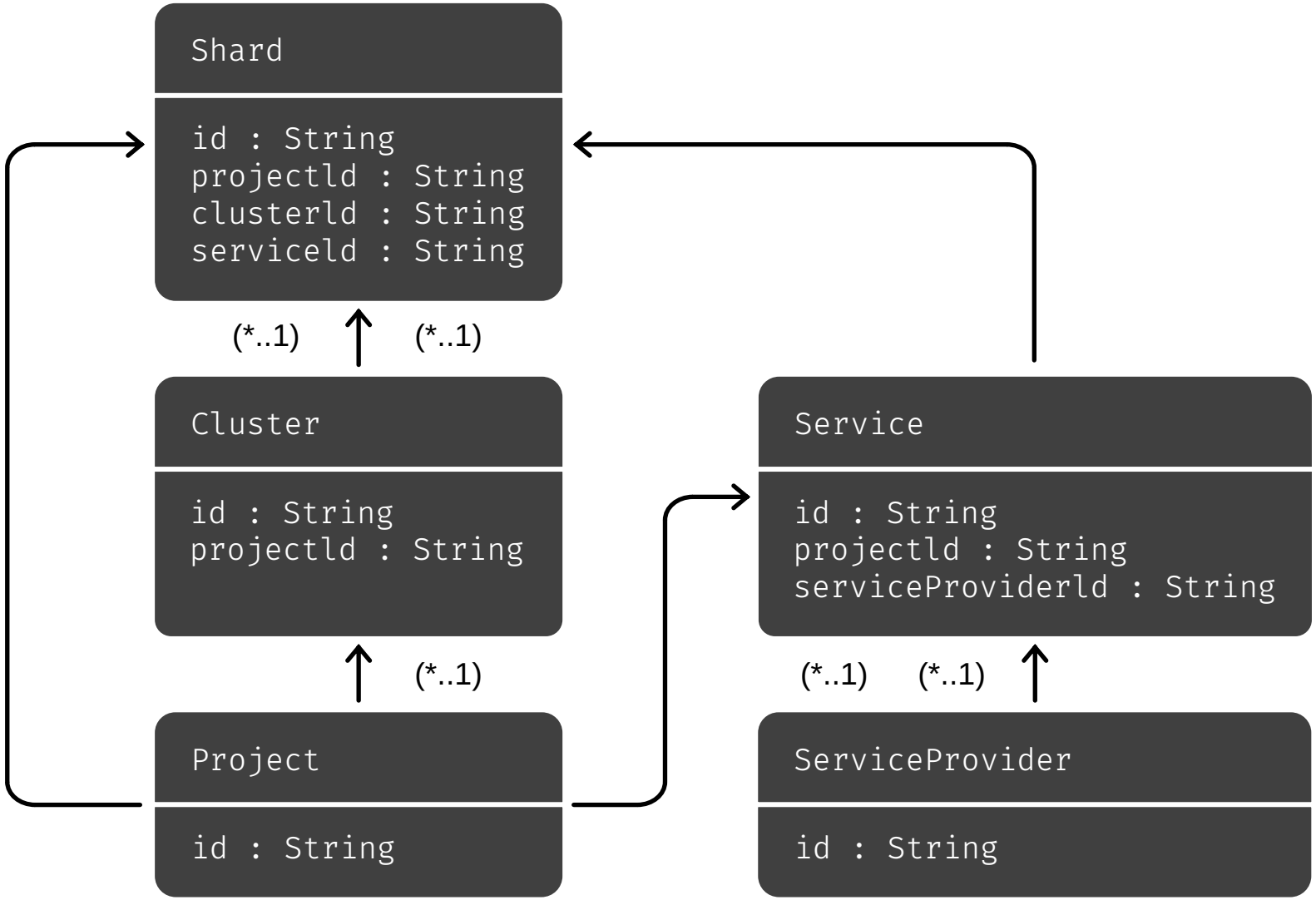
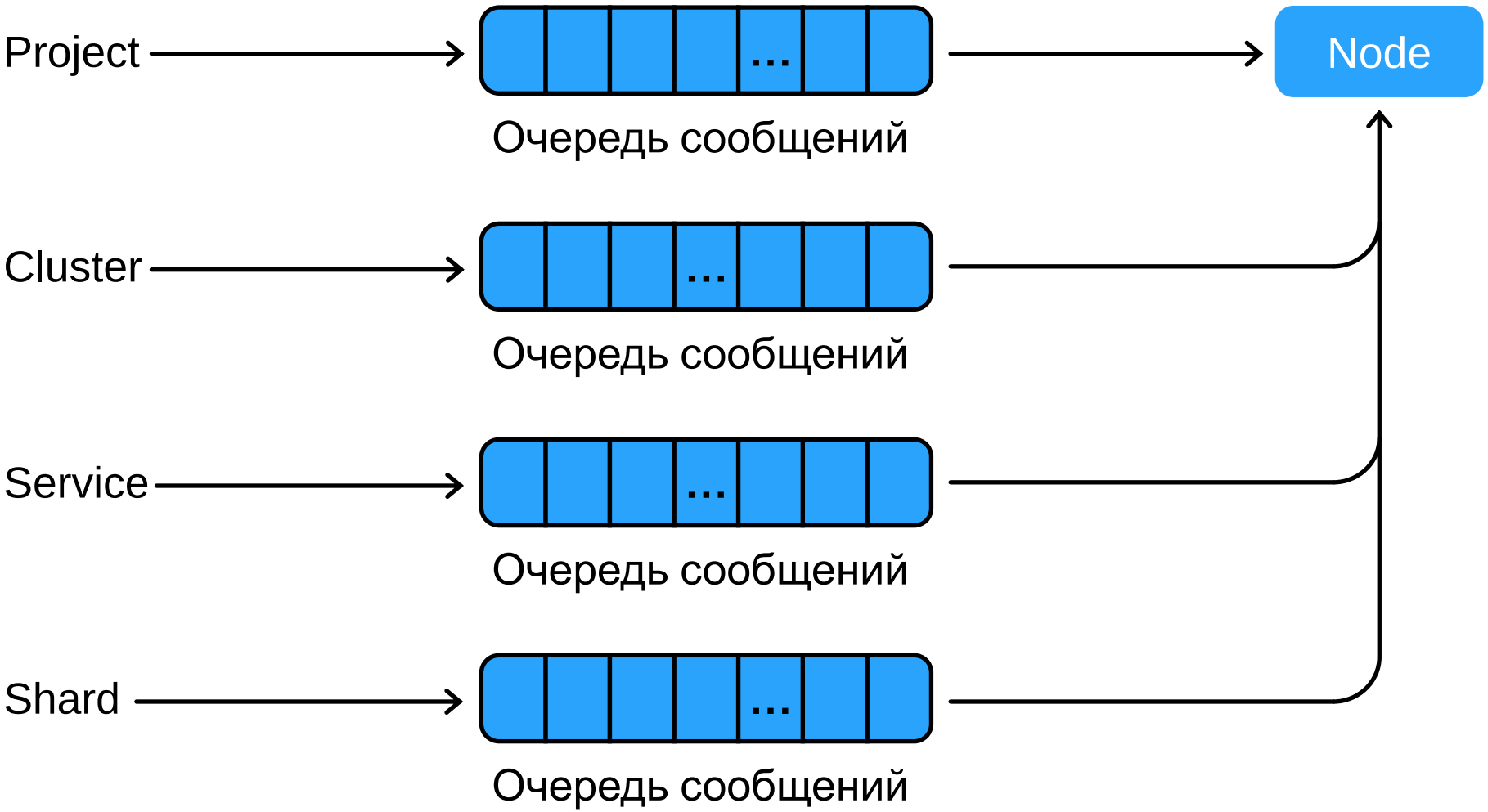


Схема очередей



Целостность без транзакций

Целостность без транзакций

- Считываем изменения из CDC

Saint HighLoad++ 2023

Целостность без транзакций

- Считываем изменения из CDC
- На лету валидируем все связи

Saint HighLoad++ 2023

Целостность без транзакций

- Считываем изменения из CDC
- На лету валидируем все связи
- Если состояние невалидно —
  не используем и ждём

Saint HighLoad++ 2023

Целостность без транзакций

- Считываем изменения из CDC
- На лету валидируем все связи
- Если состояние невалидно —
  не используем и ждём
- Если не дождались —
  перечитываем запросом

Saint HighLoad++ 2023

Схема YDB CDC

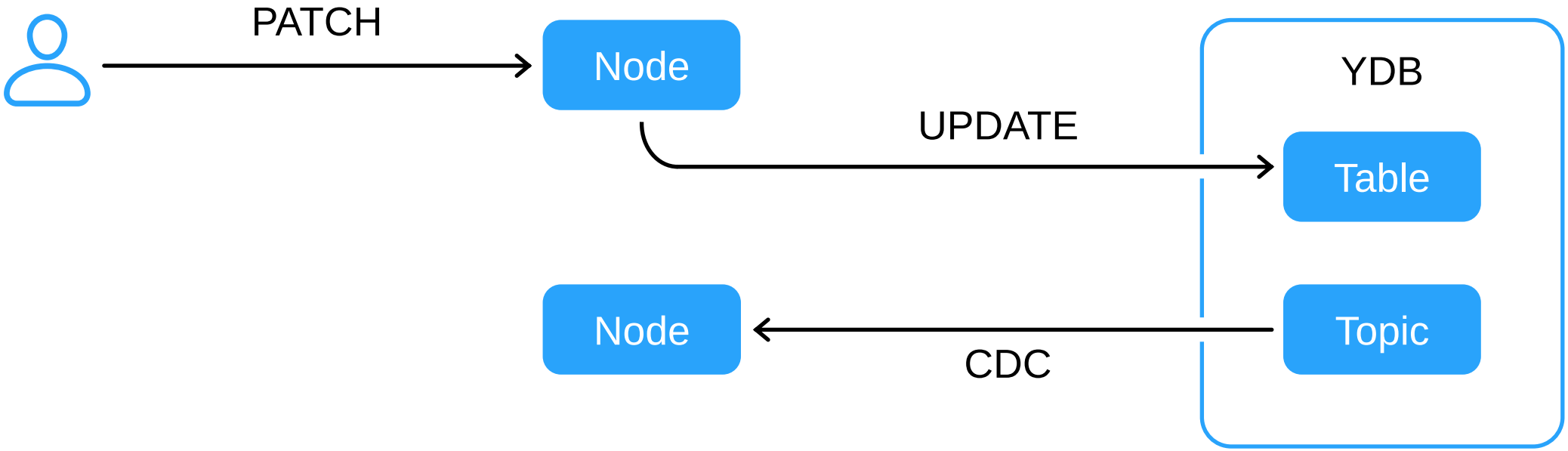
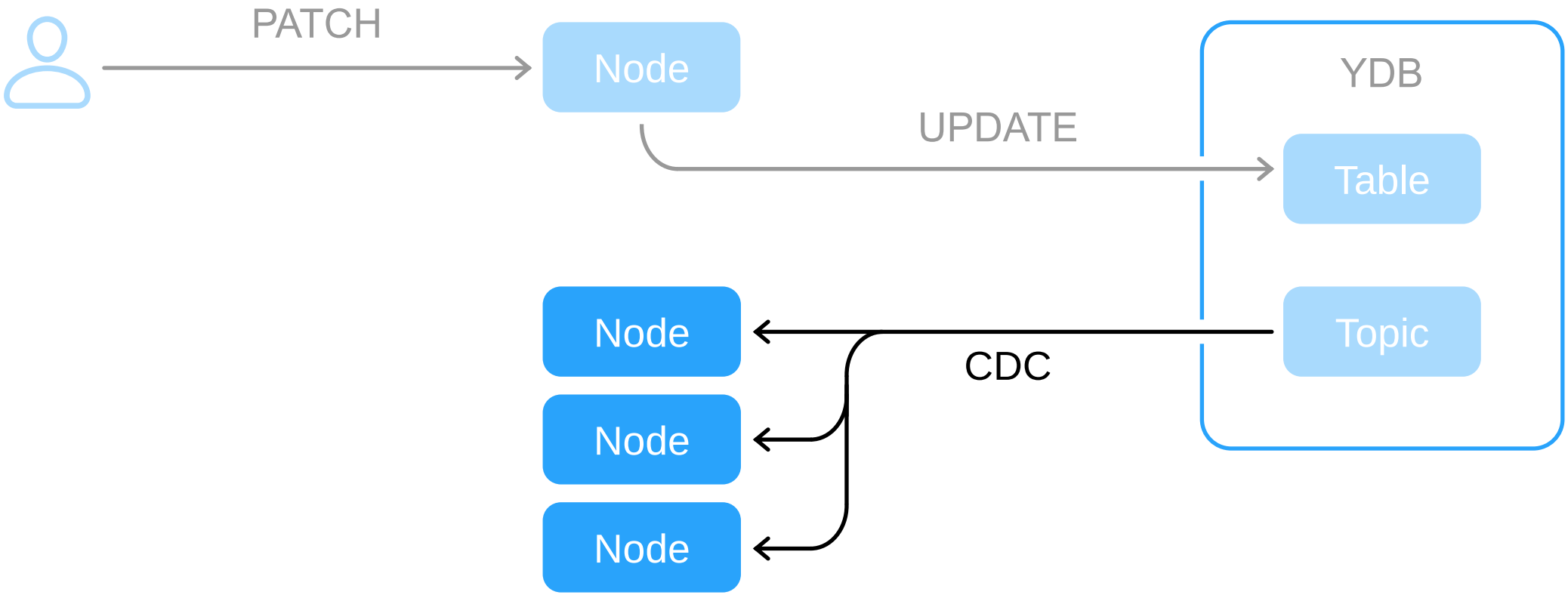


Схема YDB CDC



Лимиты YDB CDC по умолчанию

Лимиты YDB CDC по умолчанию

2 MB max per sec

на подписчика

Saint HighLoad++ 2023

Лимиты YDB CDC по умолчанию

2 MB max per sec

на подписчика

10 max

подписчиков на очередь

Saint HighLoad++ 2023

Лимиты YDB CDC по умолчанию

2 MB max per sec
на подписчика

10 max
подписчиков на очередь

5 max
очередей на таблицу

Saint HighLoad++ 2023

Лимиты YDB CDC по умолчанию

2 MB max per sec
на подписчика

10 max
подписчиков на очередь

5 max
очередей на таблицу

100 MB max per sec
итого на таблицу

Saint HighLoad++ 2023

Как не надо было делать

Как не надо было делать

- Обновления редкие

Saint HighLoad++ 2023

Как не надо было делать

- Обновления редкие
- Хотим подписать на очередь 1500 нод

Saint HighLoad++ 2023

Как не надо было делать

- Обновления редкие
- Хотим подписать на очередь 1500 нод
- Готово 

Saint HighLoad++ 2023

Как не надо было делать

- Обновления редкие
- Хотим подписать на очередь 1500 нод
- Готово
- Надо было: зеркалировать
  или повышать лимиты

Повышать лимиты
clck.ru/34d9y6

Как не надо было делать

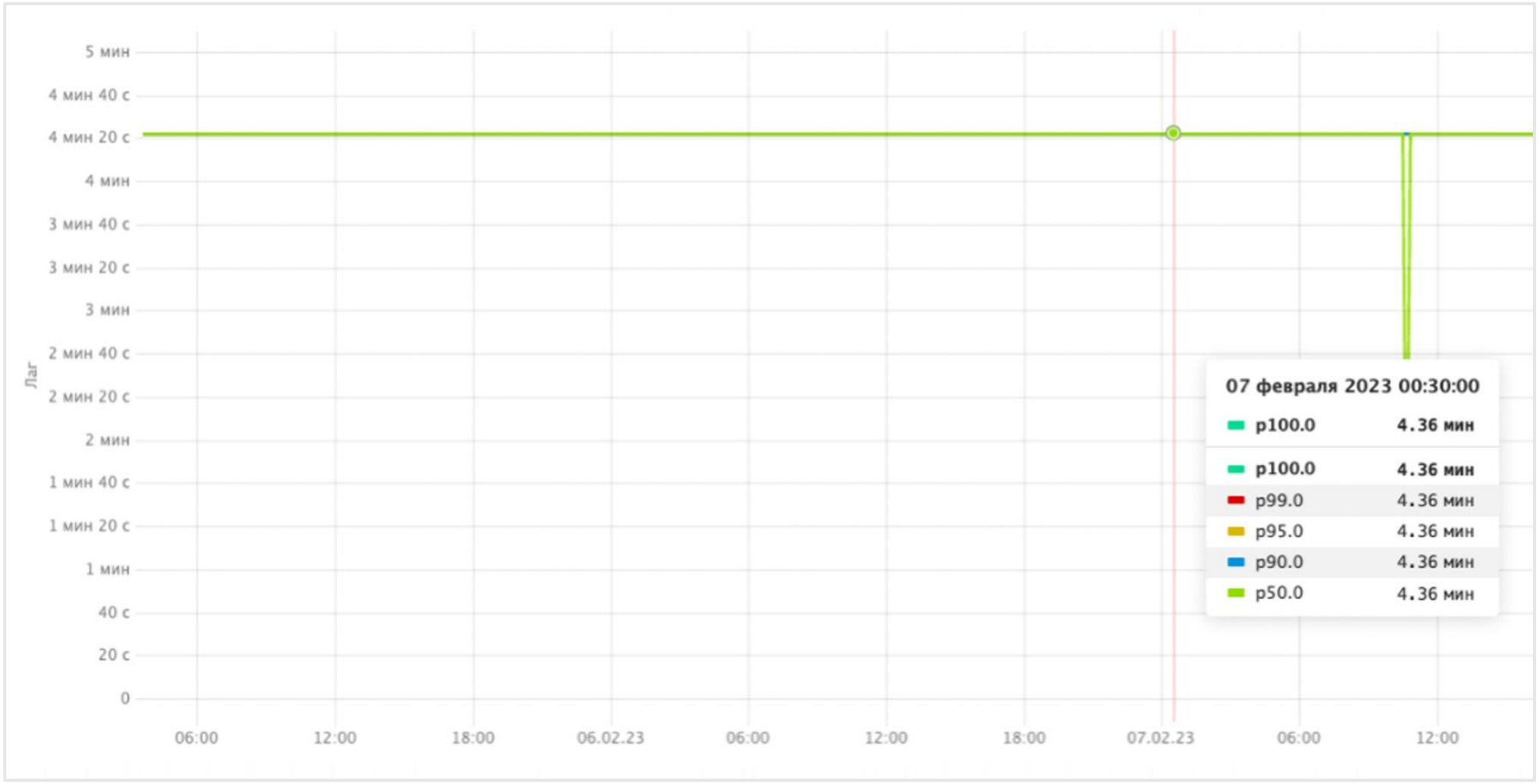
- Обновления редкие
- Хотим подписать на очередь 1500 нод
- Готово
- Надо было: зеркалировать
  или повышать лимиты
- На самом деле наш сценарий
  уже поддерживается

Повышать лимиты
clck.ru/34d9y6

План

1. Теория
2. Практика использования CDC
3. Что пошло не так
4. Итоги

Проблема: лаг доставки 4 минуты



Saint HighLoad++ 2023

Решение: лаг доставки 4 минуты

- Перезагружаем ноды

Saint HighLoad++ 2023

Решение: лаг доставки 4 минуты

- Перезагружаем ноды
- Они загружают первый раз таблицы

Saint HighLoad++ 2023

Решение: лаг доставки 4 минуты

- Перезагружаем ноды
- Они загружают первый раз таблицы
- Это долго, отсюда большие персентили вверху

Saint HighLoad++ 2023

Решение: лаг доставки 4 минуты

Решение: лаг доставки 4 минуты

Смотрим
на лаги доставки
в очереди YDB
и загрузку

Загрузка
clck.ru/34dAM7

Saint HighLoad++ 2023

Решение: лаг доставки 4 минуты

Смотрим
на лаги доставки
в очереди YDB
и загрузку

Загрузка
clck.ru/34dAM7

Тяжёлые метрики,
когда много
подписчиков

Исправлено в Upstream

Upstream
clck.ru/34dANP

Решение: лаг доставки 4 минуты

Смотрим
на лаги доставки
в очереди YDB
и загрузку

Загрузка
clck.ru/34dAM7

Тяжёлые метрики,
когда много
подписчиков

Исправлено в Upstream

Upstream
clck.ru/34dANP

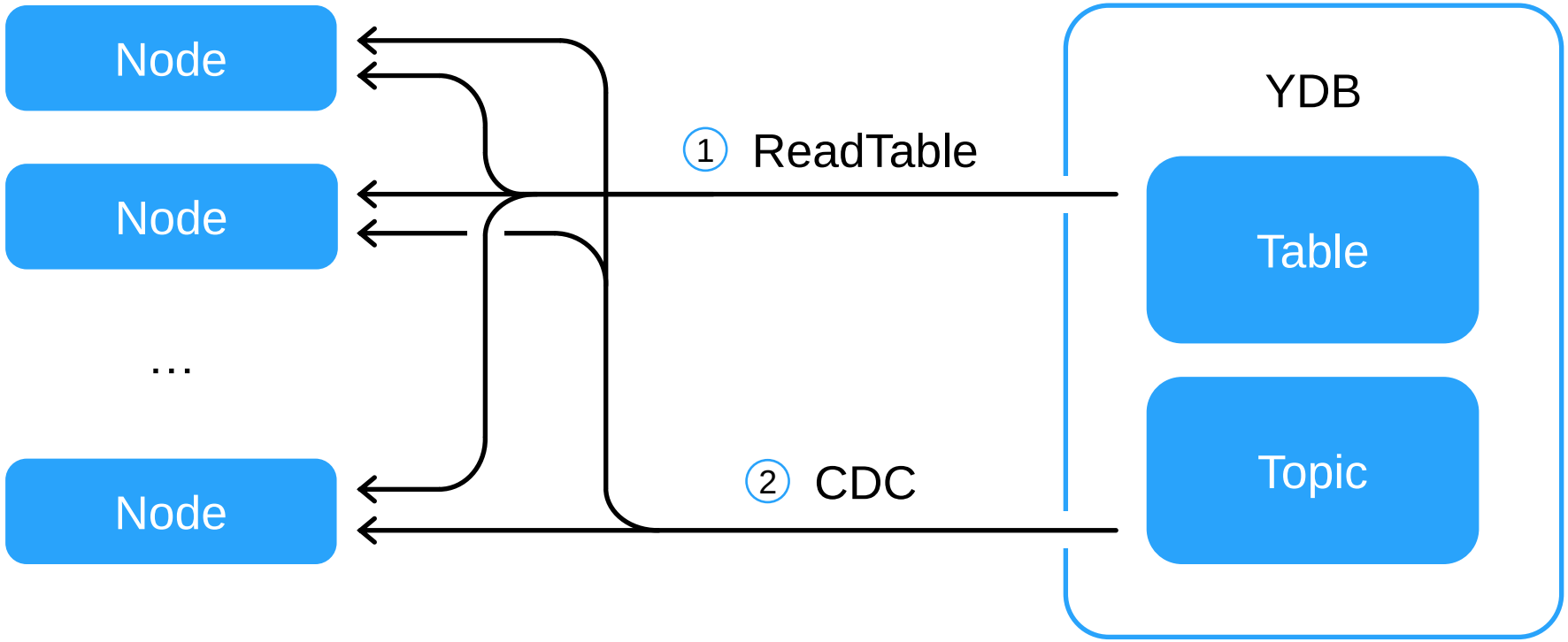
Тяжёлые
SQL-запросы

Оптимизировать

Добавлять ресурсы

Performance
clck.ru/34hXJg

Проблема: массовый рестарт



Решение: массовый рестарт

1

Повторяем упавшие
ReadTable на клиентах

Saint HighLoad++ 2023

Решение: массовый рестарт

1

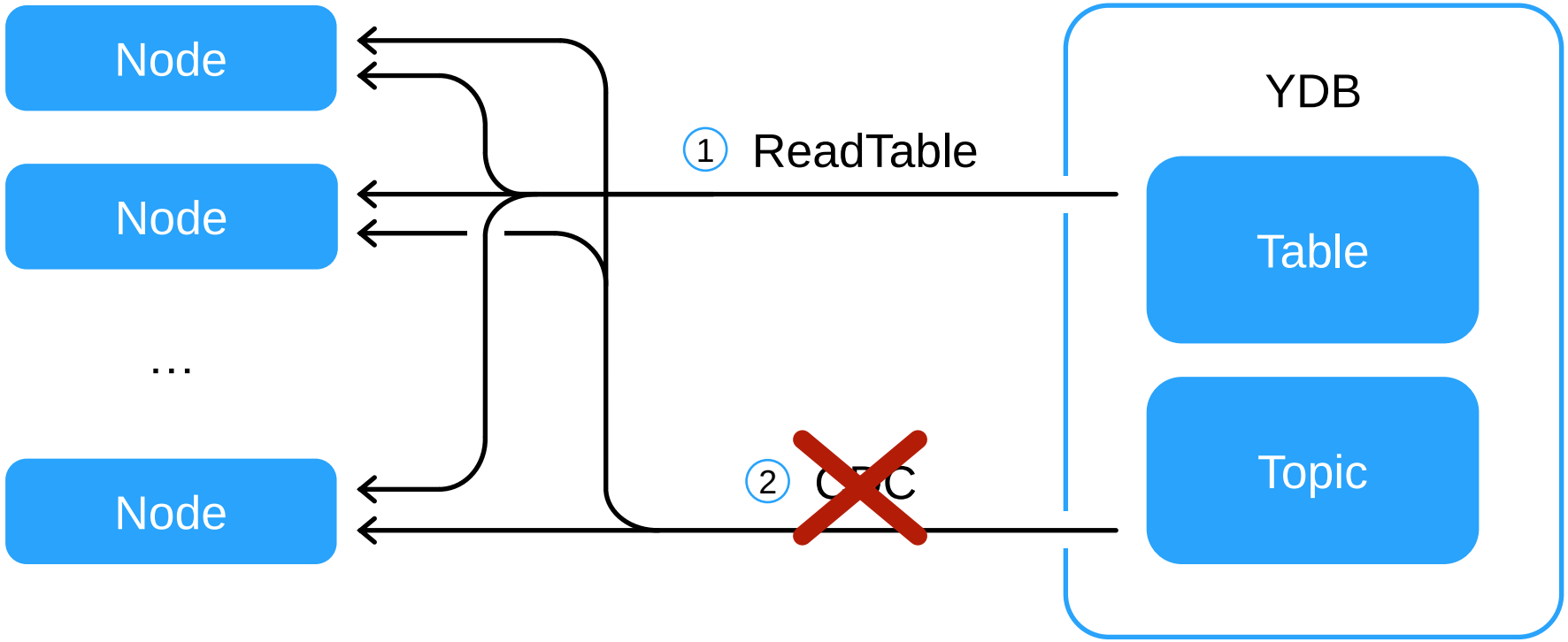
Повторяем упавшие ReadTable на клиентах

2

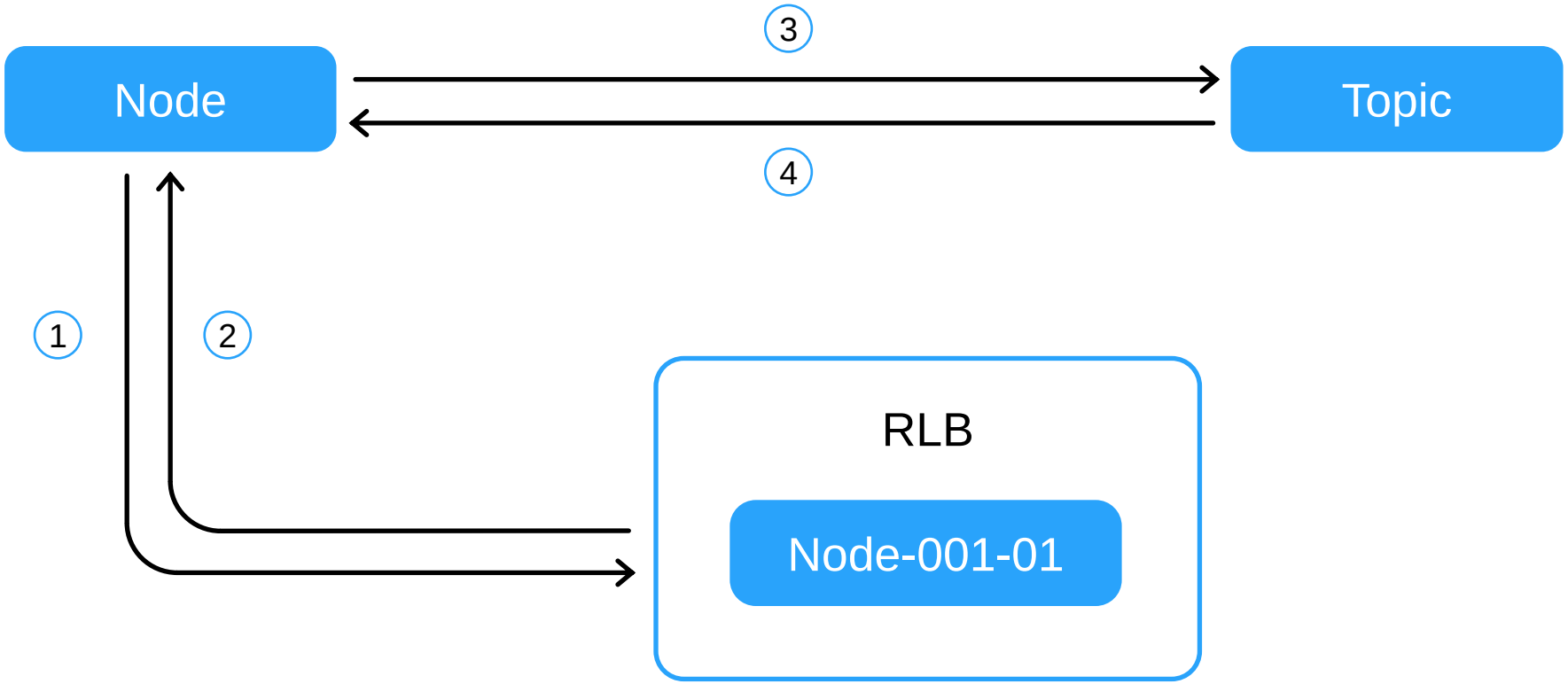
Distributed semaphore на клиентах

Saint HighLoad++ 2023

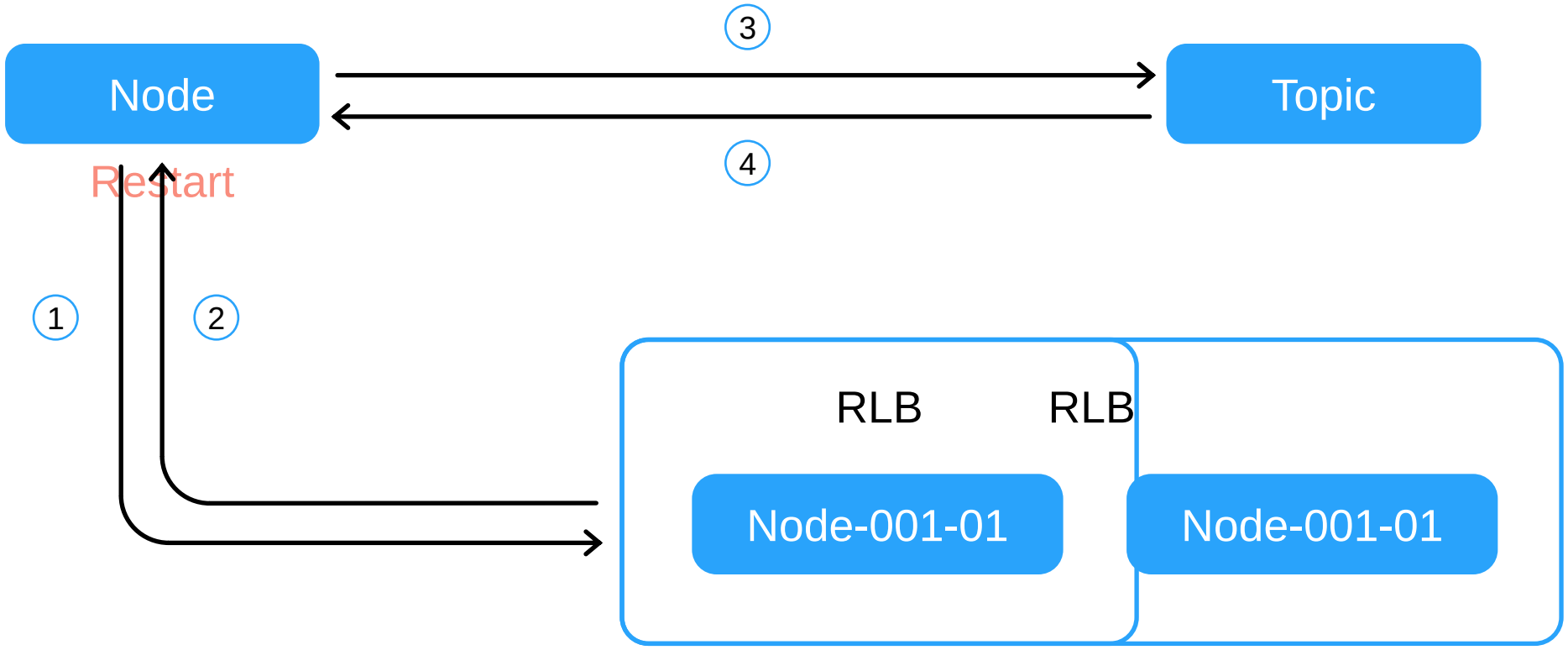
Проблема: рестарт сервиса



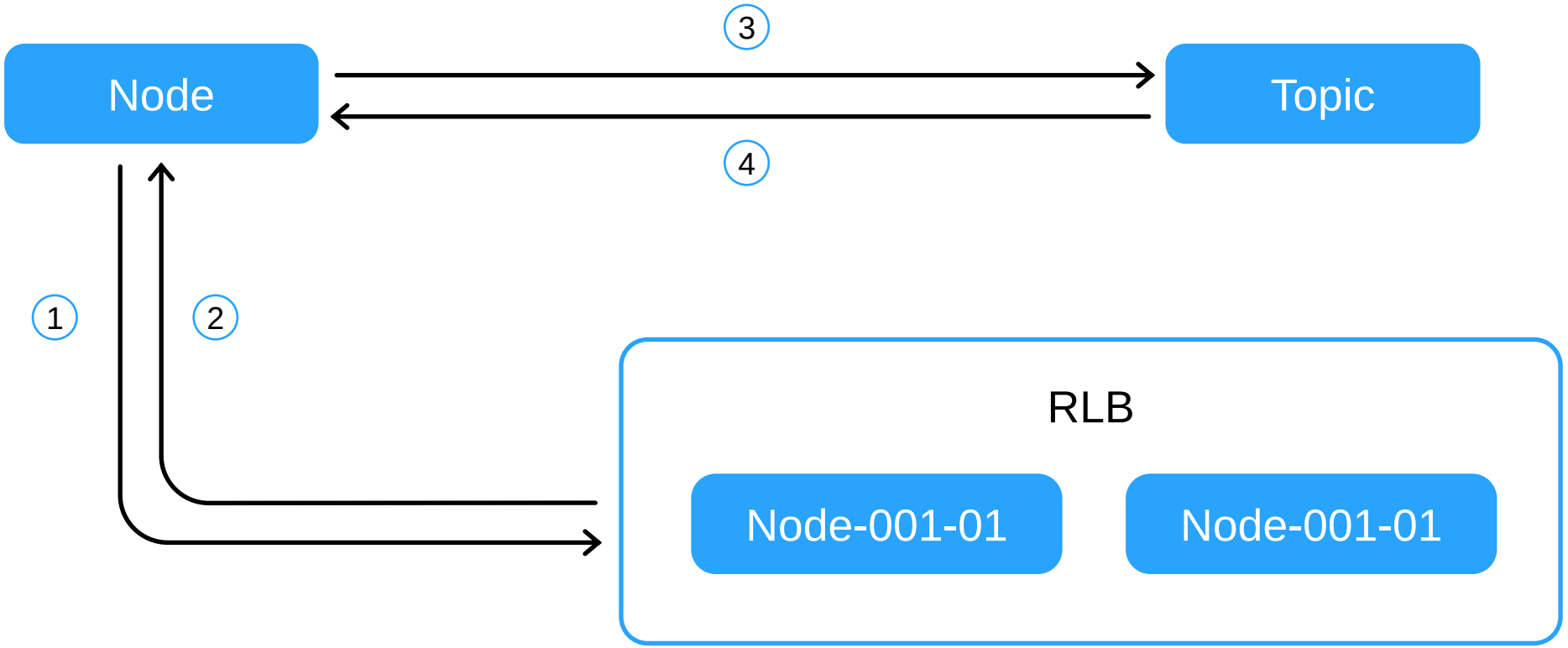
Проблема: рестарт сервиса



Проблема: рестарт сервиса



Проблема: рестарт сервиса

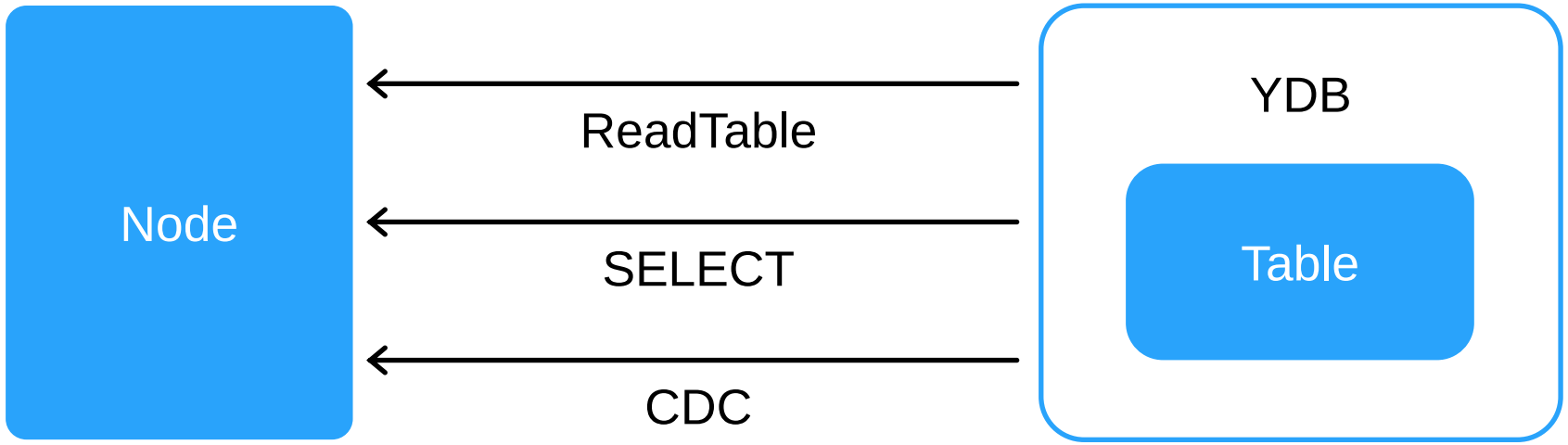


Решение: рестарт сервиса

Client GRPC keep alive/idle timeout — помогают умирать старой сессии

Saint HighLoad++ 2023

Проблема: методы чтения



Решение: методы чтения

Решение: методы чтения

- Round-trip tests

  Подняли локально YDB

  Записали сущность

  Прочитали через CDC

  Проверили, что все поля заполнились корректно

Saint HighLoad++ 2023

Решение: методы чтения

- Round-trip tests

  Подняли локально YDB

  Записали сущность

  Прочитали через CDC

  Проверили, что все поля заполнились корректно

- Унификация кода, читающего данные

Saint HighLoad++ 2023

Решение: методы чтения

- Round-trip tests

  Подняли локально YDB

  Записали сущность

  Прочитали через CDC

  Проверили, что все поля заполнились корректно

- Унификация кода, читающего данные
- Флаг, включающий один метод

  У нас ReadTable

Saint HighLoad++ 2023

Флаг на ReadTable

Флаг на ReadTable

- Когда всё хорошо, конфиги совпадают с небольшим лагом

Saint HighLoad++ 2023

Флаг на ReadTable

- Когда всё хорошо, конфиги совпадают с небольшим лагом
- Если найдено расхождение, то ставится флаг

Saint HighLoad++ 2023

Флаг на ReadTable

- Когда всё хорошо, конфиги совпадают с небольшим лагом
- Если найдено расхождение, то ставится флаг
- Расхождение может возникнуть

  Баги в коде при выкатке новых версий

Saint HighLoad++ 2023

Флаг на ReadTable

- Когда всё хорошо, конфиги совпадают с небольшим лагом
- Если найдено расхождение, то ставится флаг
- Расхождение может возникнуть

  Баги в коде при выкатке новых версий

  Потеря сетевого соединения и т. п.

Saint HighLoad++ 2023

Флаг на ReadTable

- Когда всё хорошо, конфиги совпадают с небольшим лагом
- Если найдено расхождение, то ставится флаг
- Расхождение может возникнуть

  Баги в коде при выкатке новых версий

  Потеря сетевого соединения и т. п.

  Космическое излучение меняет битик

Saint HighLoad++ 2023

Как всё это мониторить

- Метрики сервиса

Saint HighLoad++ 2023

Как всё это мониторить

- Метрики сервиса

  Задержка доставки

Saint HighLoad++ 2023

Как всё это мониторить

- Метрики сервиса

  Задержка доставки

  Время последнего события
  (cross dc, cross nodes)

Saint HighLoad++ 2023

Как всё это мониторить

- Метрики сервиса

  Задержка доставки

  Время последнего события
  (cross dc, cross nodes)

  Объёмы чтения и спайки

Как всё это мониторить

- Метрики сервиса
  - Задержка доставки
  - Время последнего события (cross dc, cross nodes)
  - Объёмы чтения и спайки
  - Ошибки обработки

Saint HighLoad++ 2023

Как всё это мониторить

- Метрики сервиса
  - Задержка доставки
  - Время последнего события (cross dc, cross nodes)
  - Объёмы чтения и спайки
  - Ошибки обработки
- Метрики YDB
  - Метрики топиков
  - Задержка чтения
  - Сообщения в очереди
  - Загрузку CPU

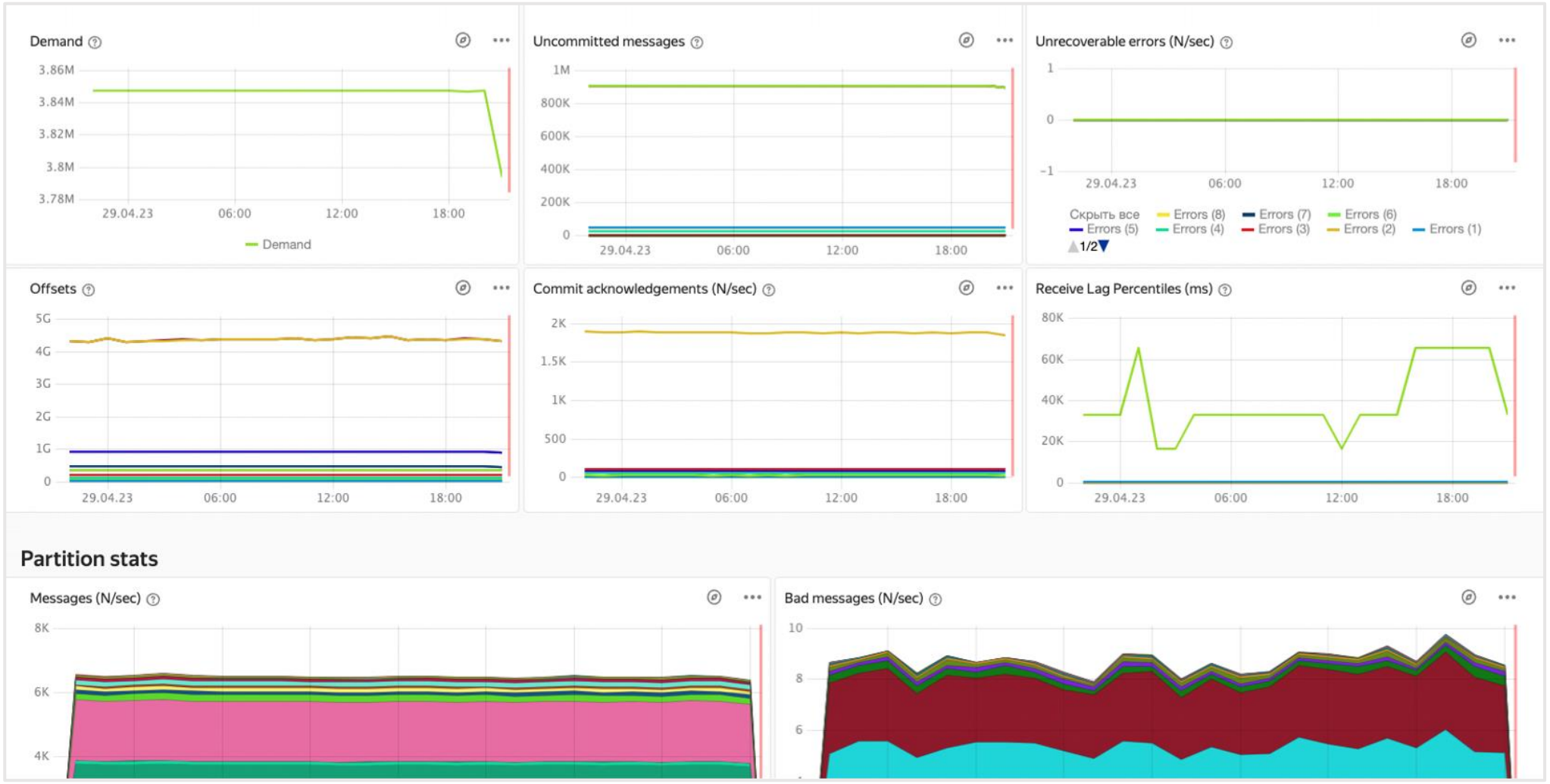
Метрики топиков

https://clck.ru/34mn8M

CPU

clck.ru/34dAM7

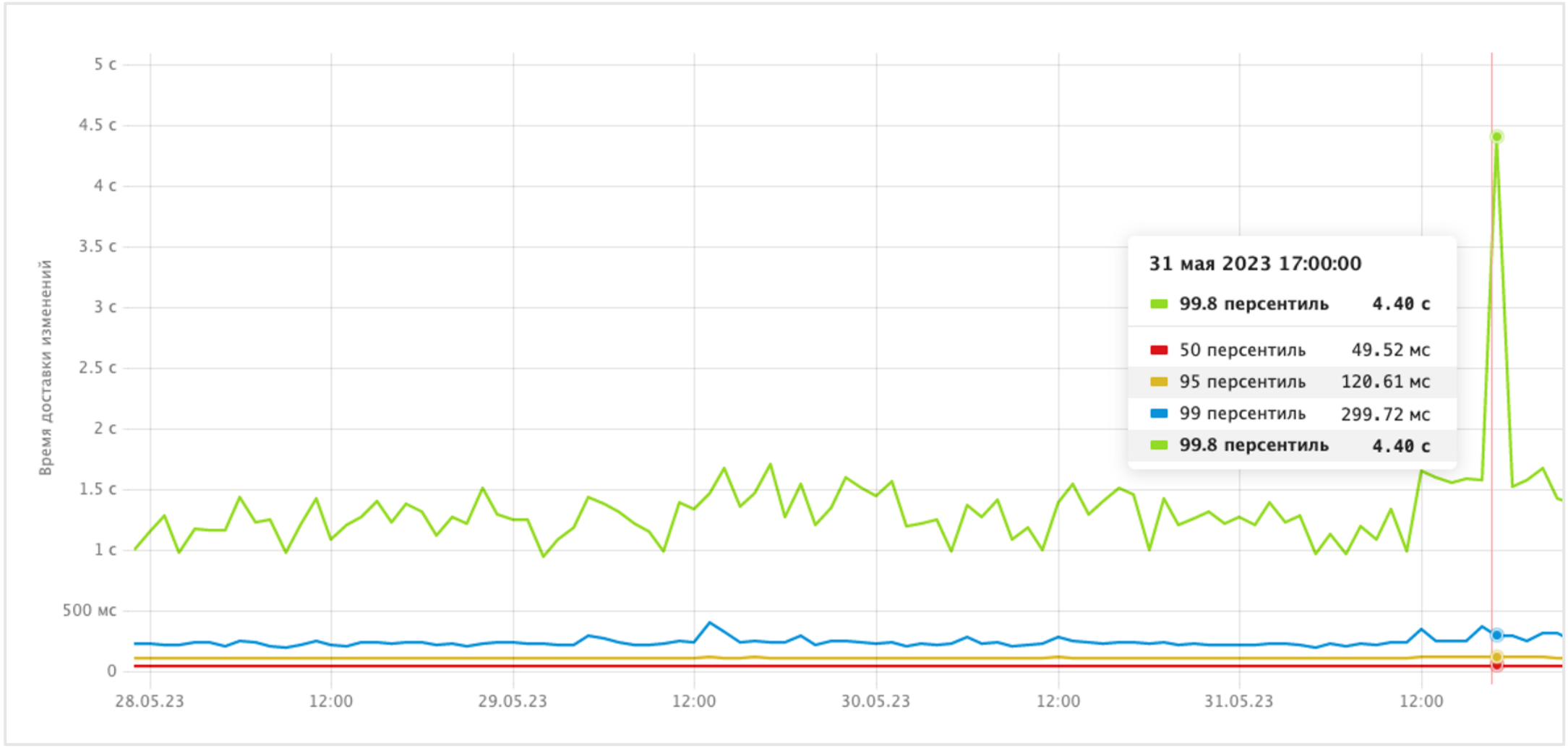
Как всё это мониторить



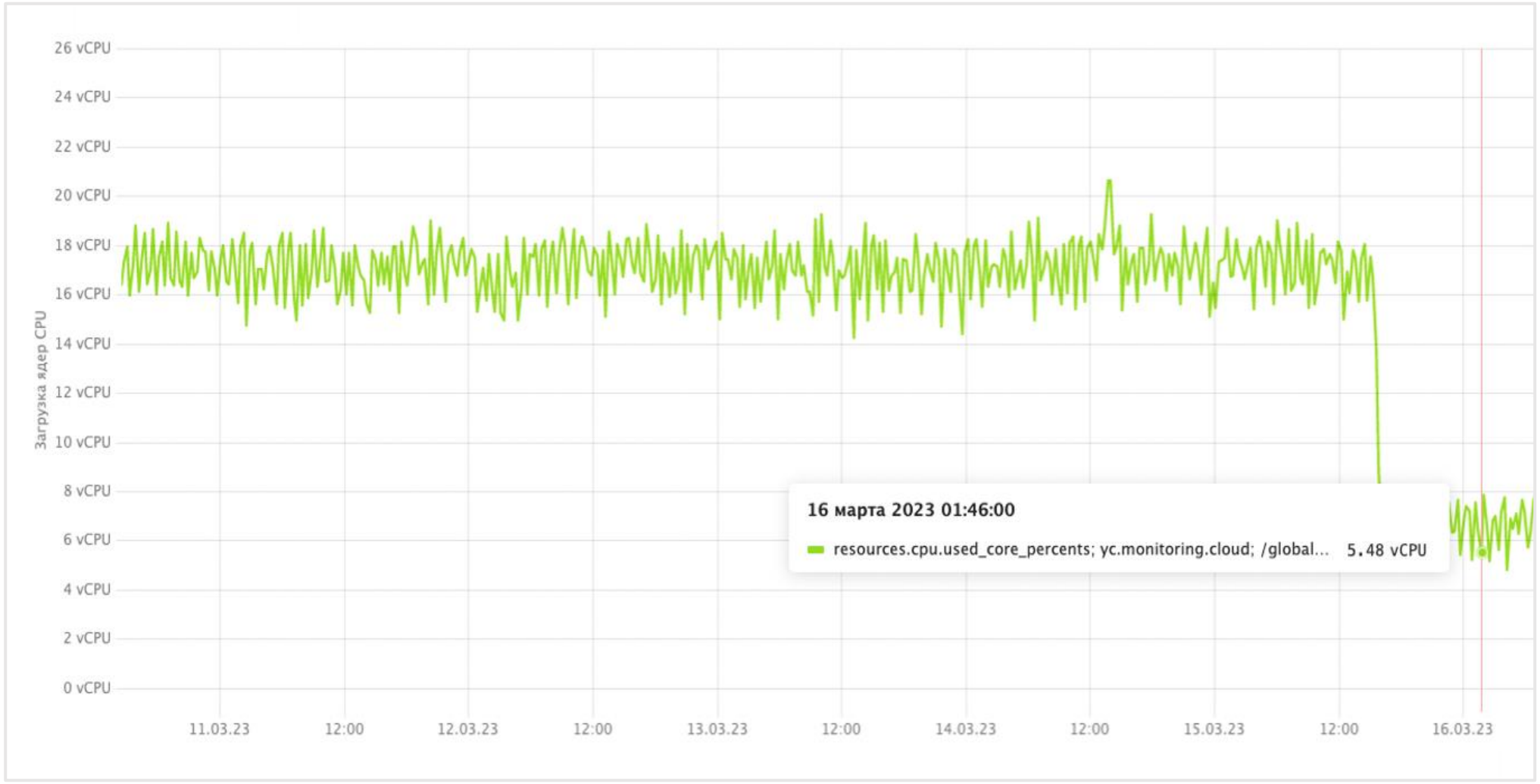
План

1. Теория
2. Практика использования CDC
3. Что пошло не так
4. Итоги

Какой стал лаг доставки?



Как упал CPU?



Saint HighLoad++ 2023

Наш CDC в цифрах

1530
подписчиков на очередь

~6 тыс.
сообщений в секунду

80 MB
в секунду в среднем

180 MB
в секунду в пике

<800
миллисекунд —
99 персентиль
лага доставки

В 2 раза
упала загрузка
ядер CPU

Saint HighLoad++ 2023

Дальнейшие планы

- Ещё больше таблиц переводить на CDC
- Протестировать CDC initial table scan
- Проверить возможности YDB SDK v2
- Планируется в CDC

  Консистентность между очередями

  Эфемерные подписчики

initial table scan
clck.ru/34dCo2

YDB SDK v2
clck.ru/34dCp7

Выводы

- Получили реализацию механизма Change Data Capture на YDB CDC, которую мы научились готовить
- Получили быструю доставку конфигураций на 1500 подписчиков
- Улучшили доступность мониторинга для наших пользователей

Saint HighLoad++ 2023

Голосуйте за мой доклад

Егор Литвиненко
Yandex Infrastructure

Infrastructure

Saint HighLoad++ 2023