YDB

Поиск самой быстрой MPMC-очереди

Алексей Станкевичус

ydb.tech

Для чего нам самая быстрая MPMC очередь

YDB: что это такое?

Собственная разработка

Начали в 2014

C++

Open Source — Apache 2.0

Реляционная СУБД со строгой консистентностью

ACID

YQL — SQL-диалект

CP с точки зрения CAP-теоремы

Serializable уровень изолированности транзакций

Горизонтальная масштабируемость

Кластеризуемая

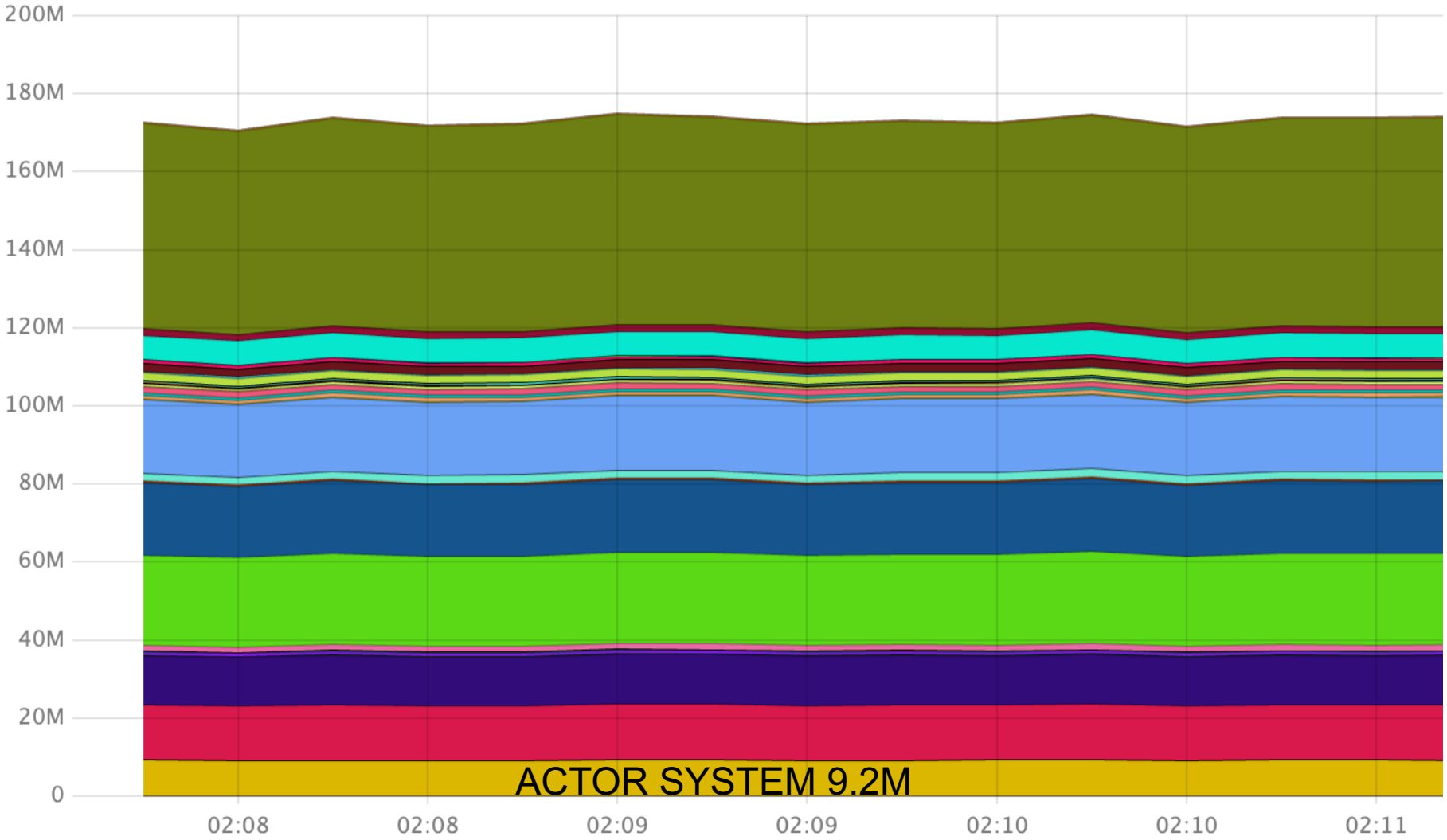
Тысячи серверов

Проблема производительности

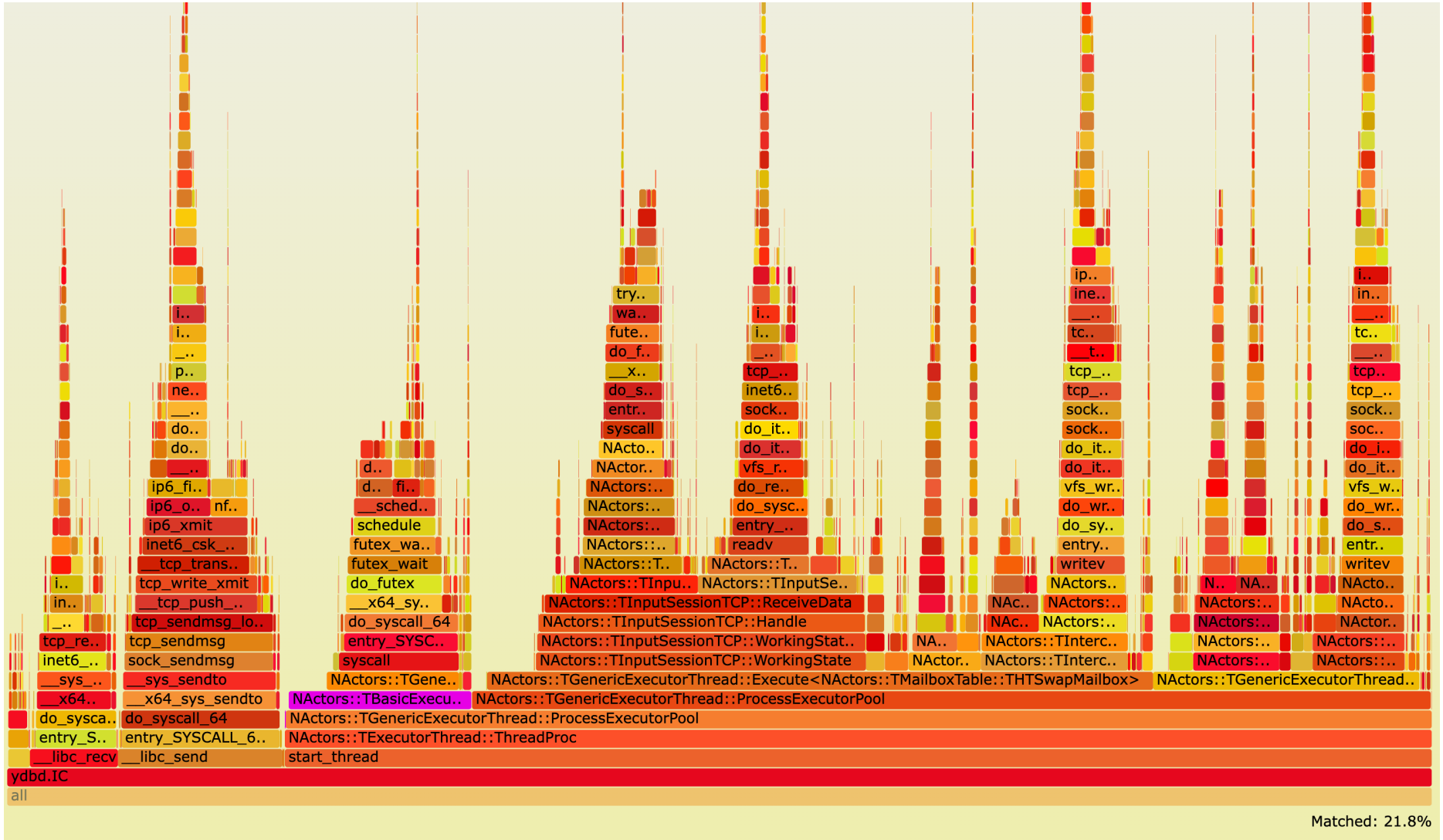
Высокая производительность YDB очень важна для нас.

YDB уже очень быстрая, но мы хотим, чтобы YDB стала еще быстрее.

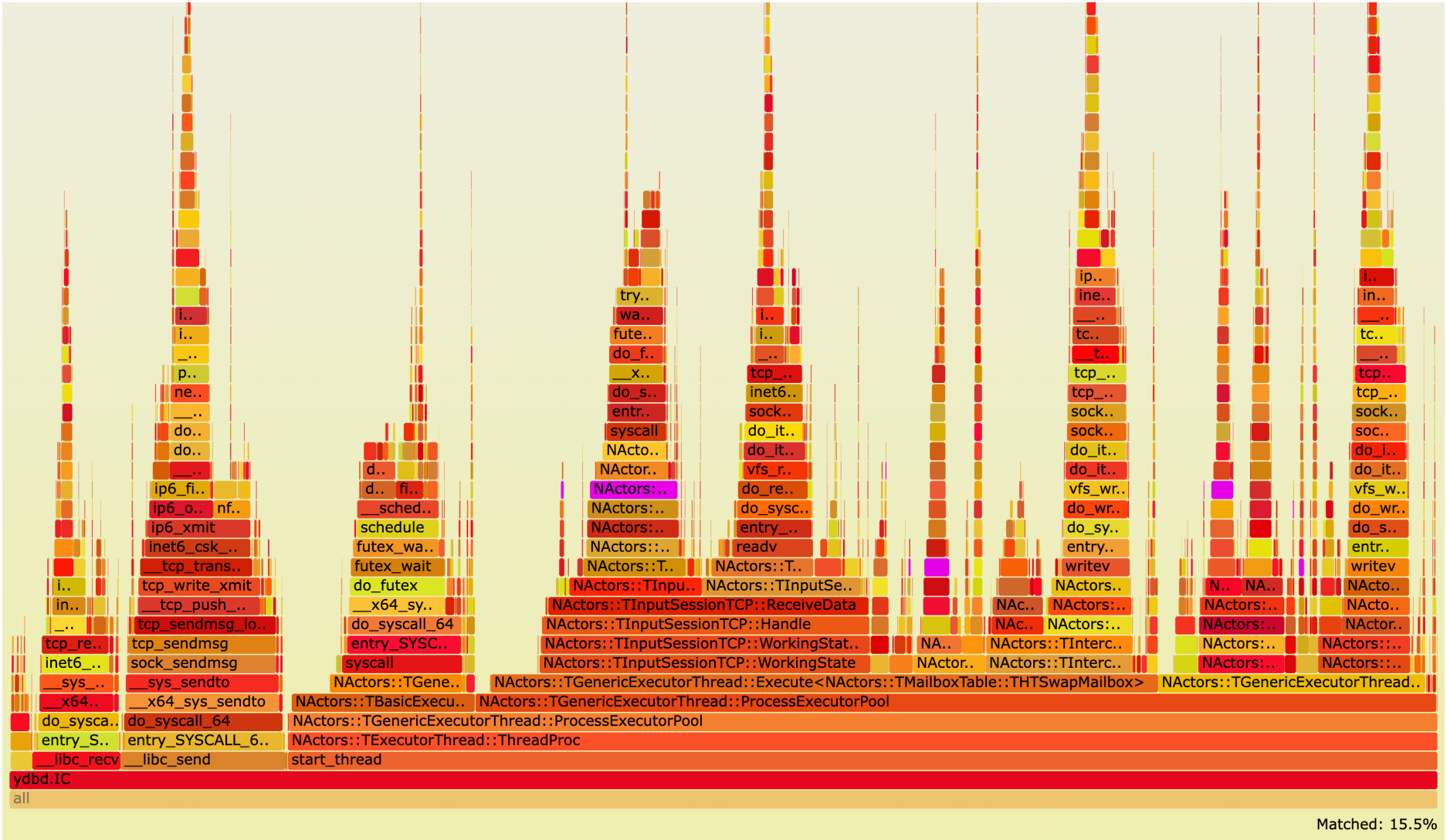
Где все тормозит



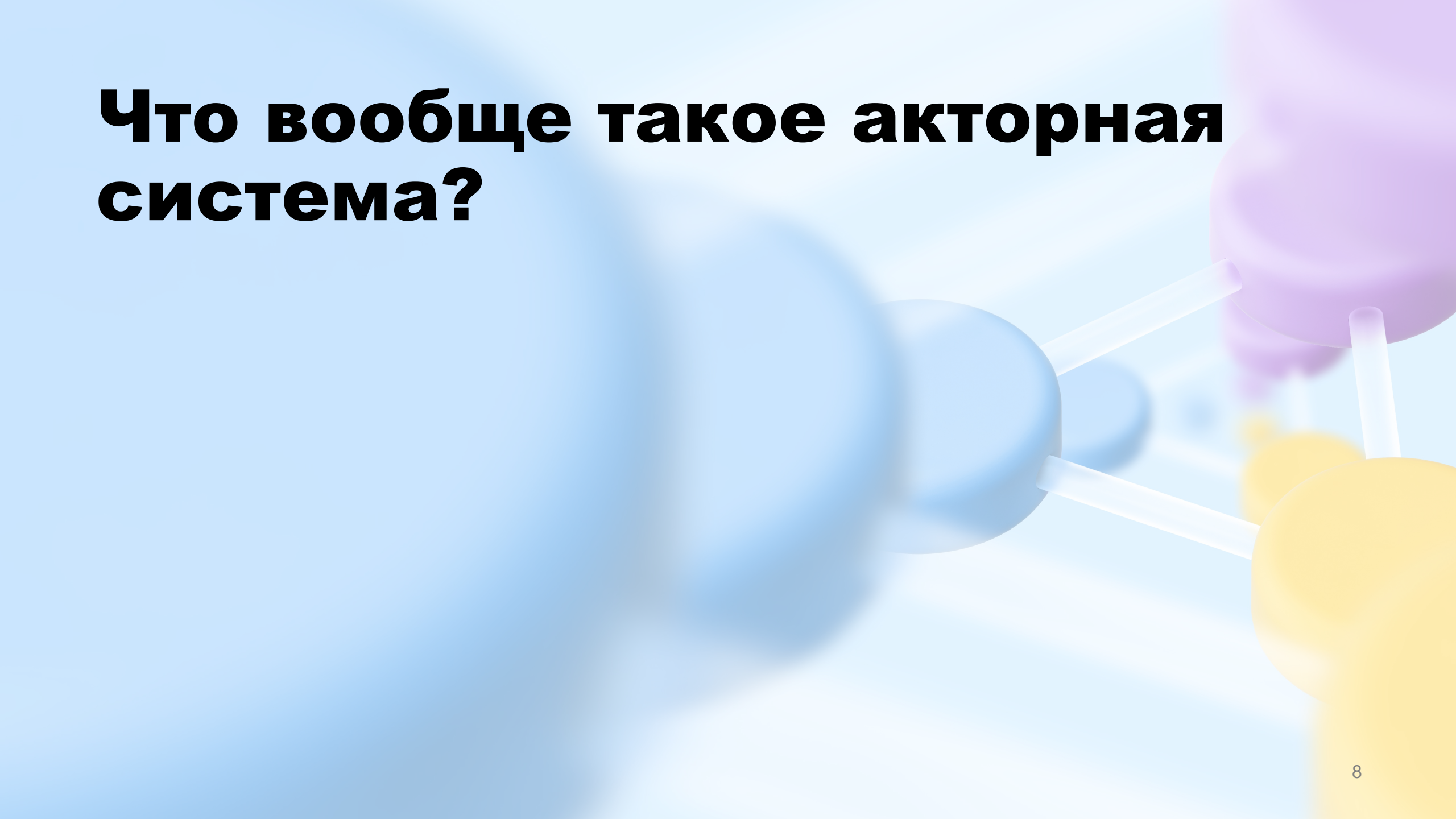
Где все тормозит



Где все тормозит

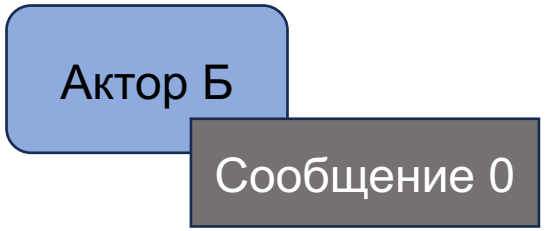


Что вообще такое акторная система?

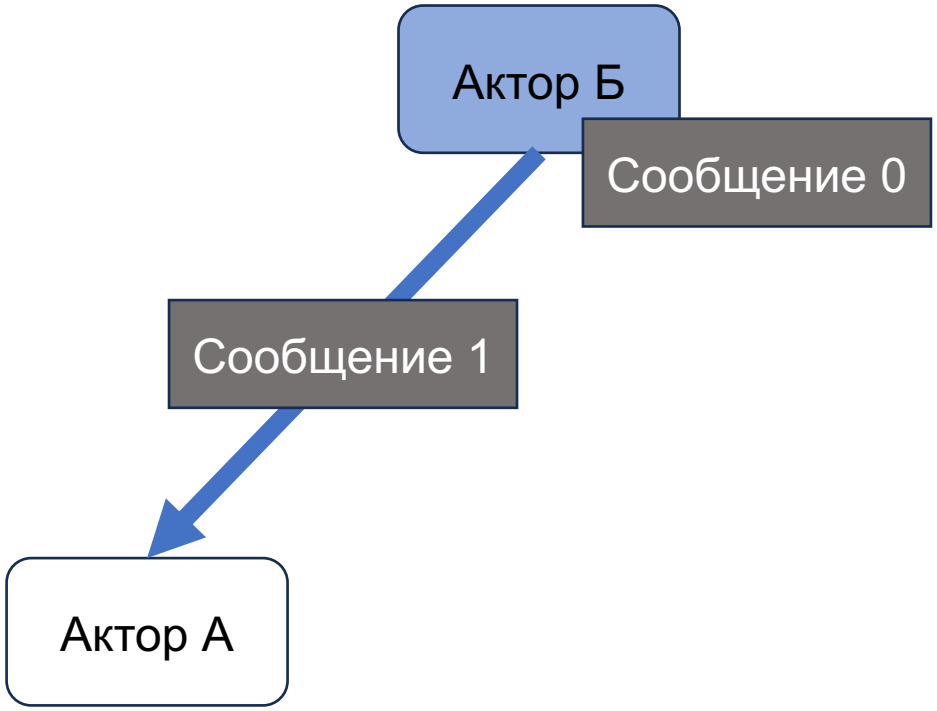


Акторная система

Лежит в основе YDB



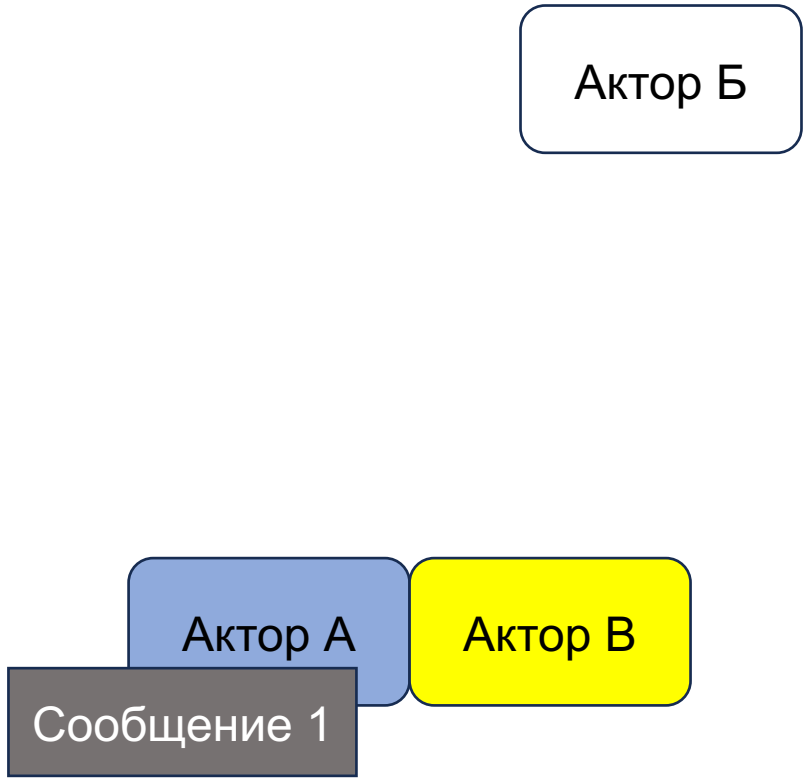
Акторная система



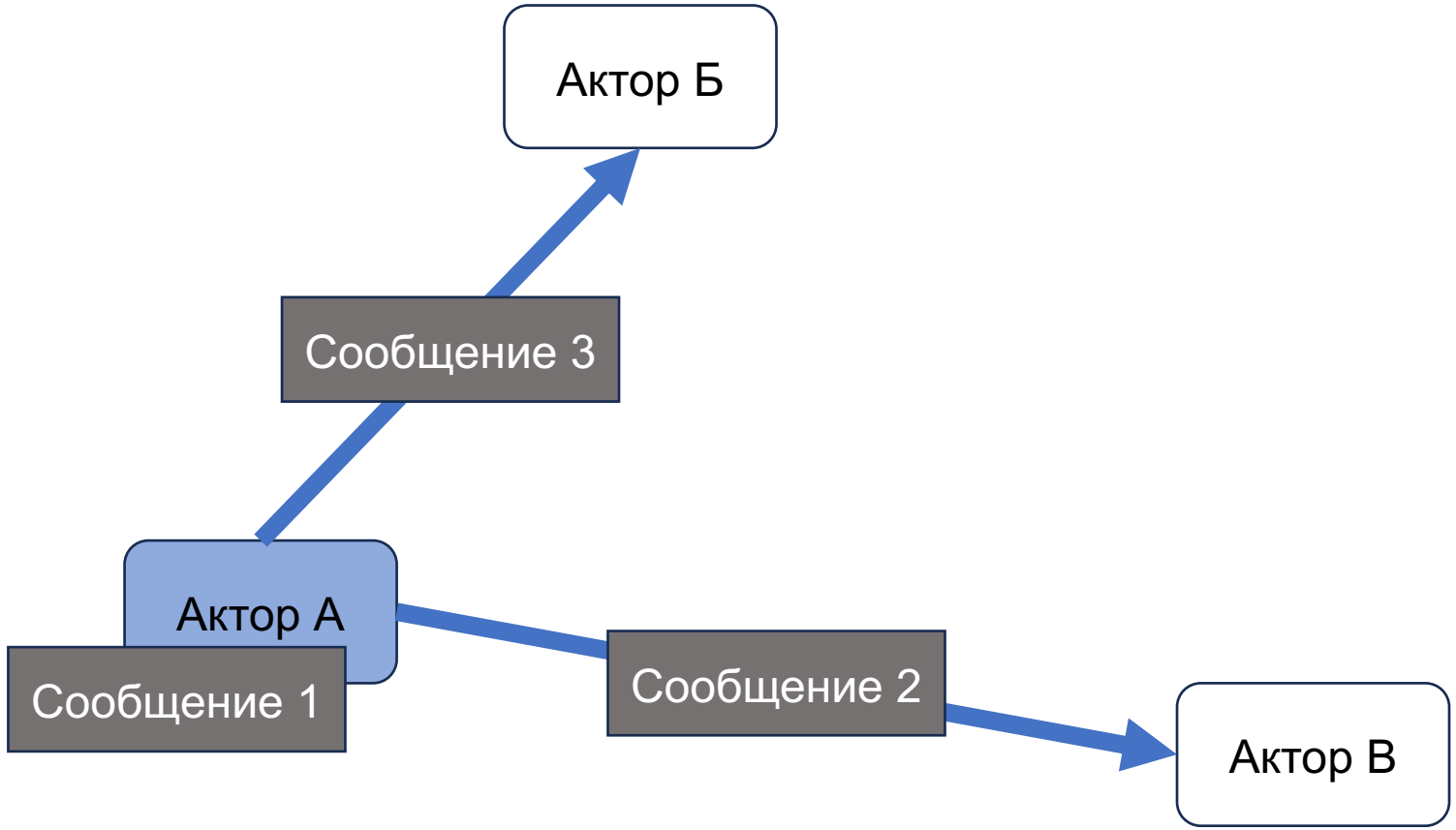
Акторная система

Акторная система

Акторная система



Акторная система

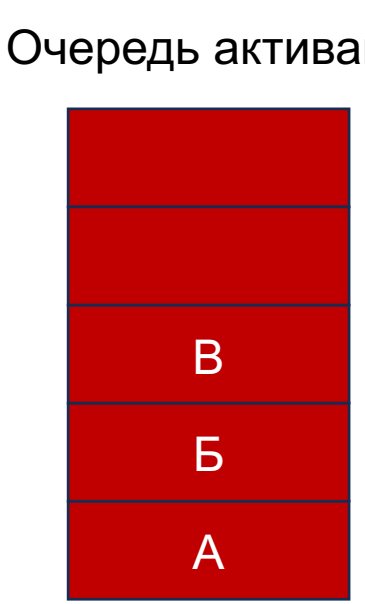


Акторная система

Акторная система

Как происходит активация акторов?

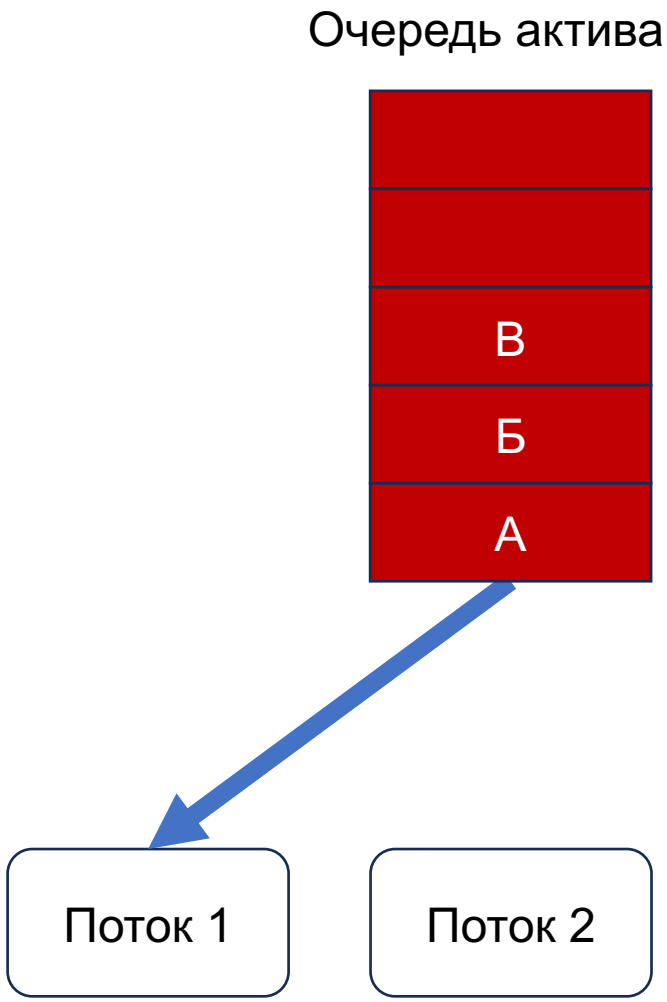
Очередь активаций



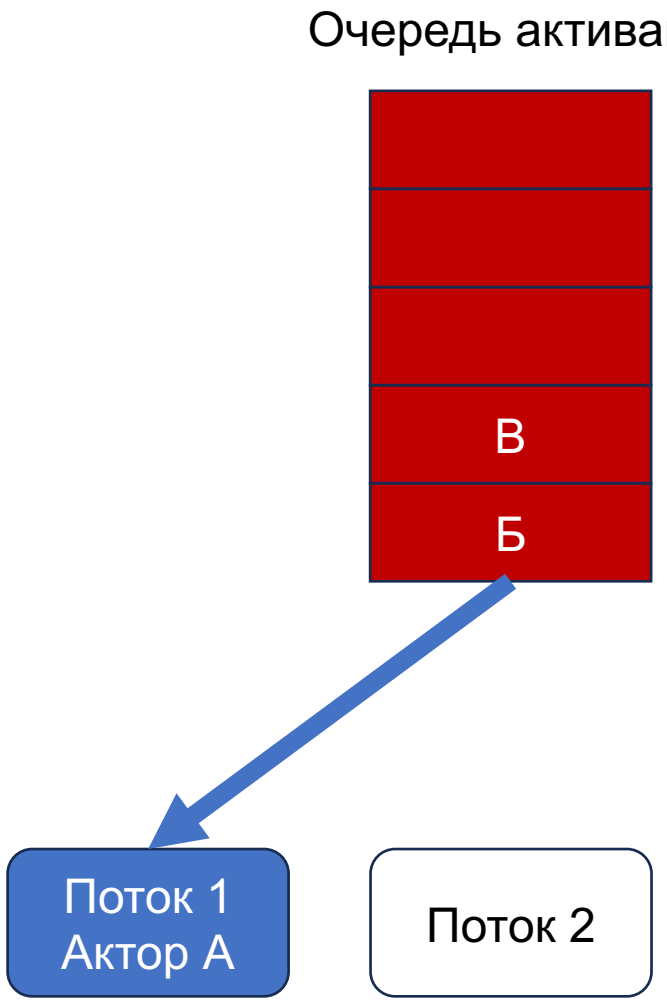
Поток 1

Поток 2

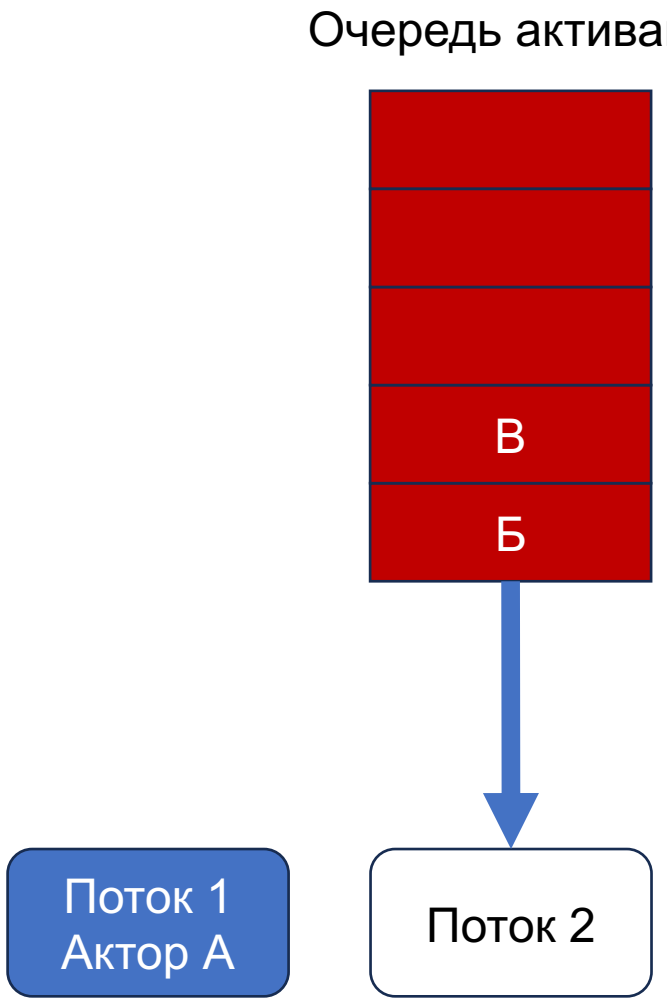
Очередь активаций



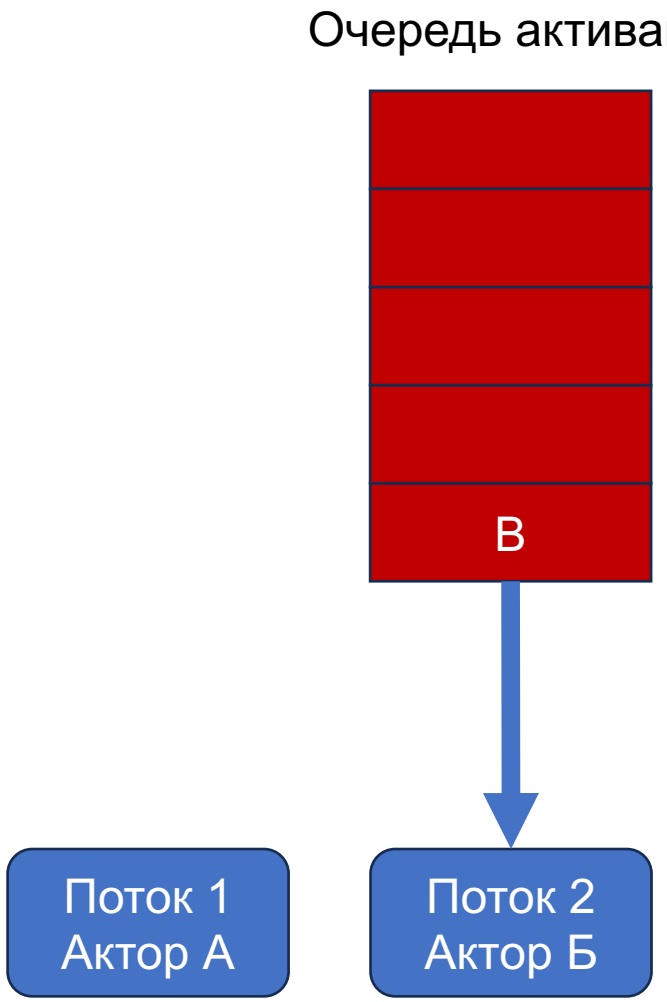
Очередь активаций



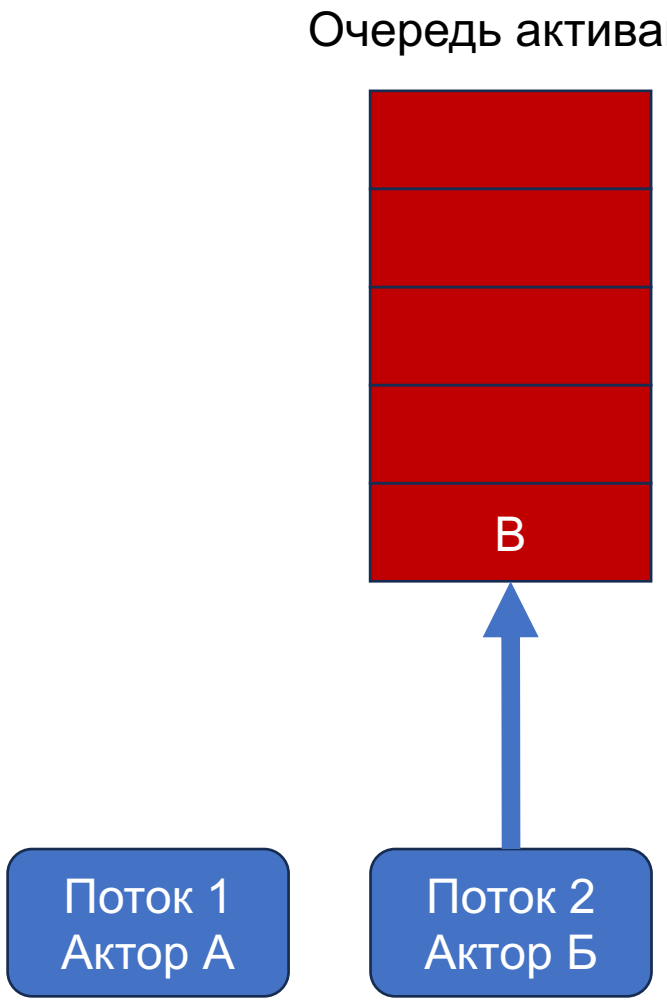
Очередь активаций



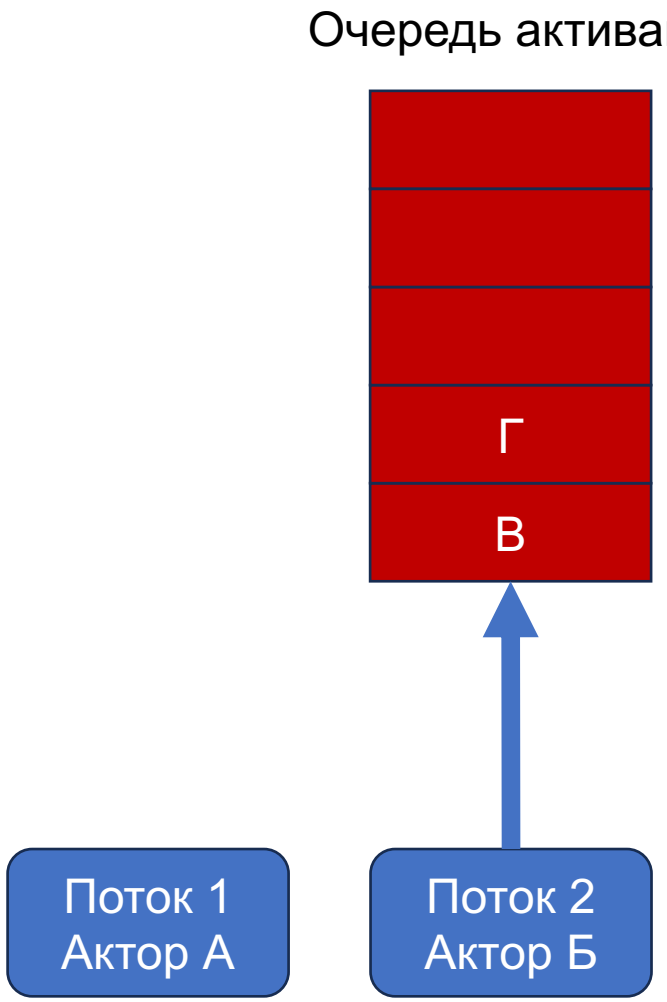
Очередь активаций



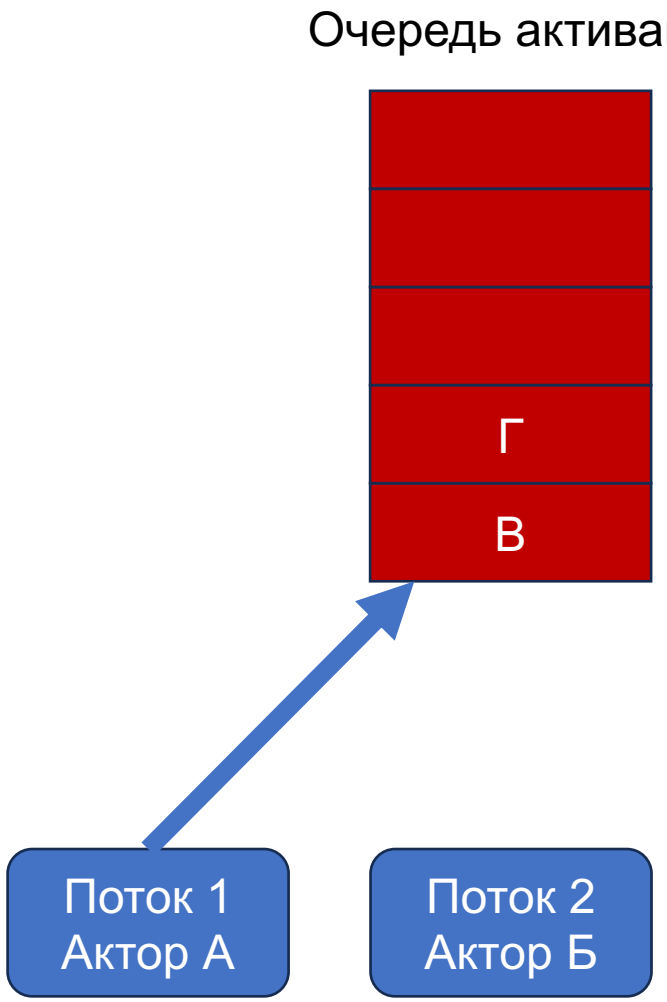
Очередь активаций



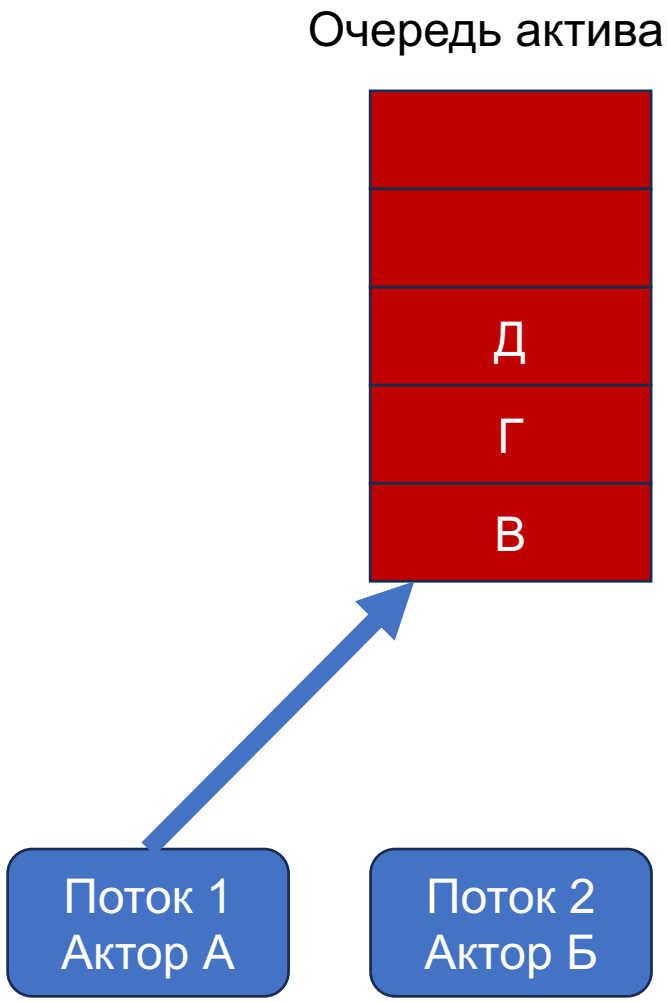
Очередь активаций



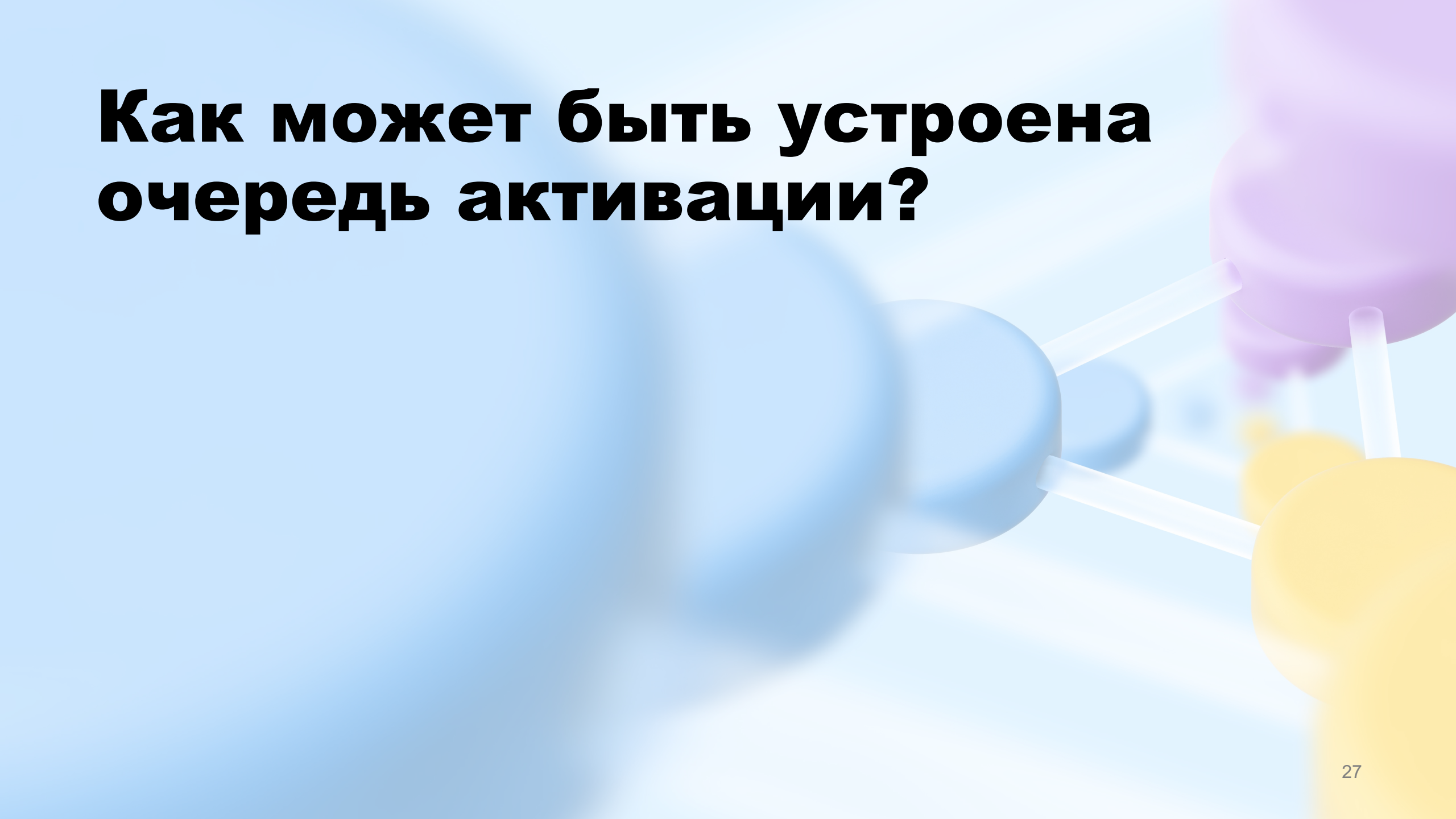
Очередь активаций



Очередь активаций

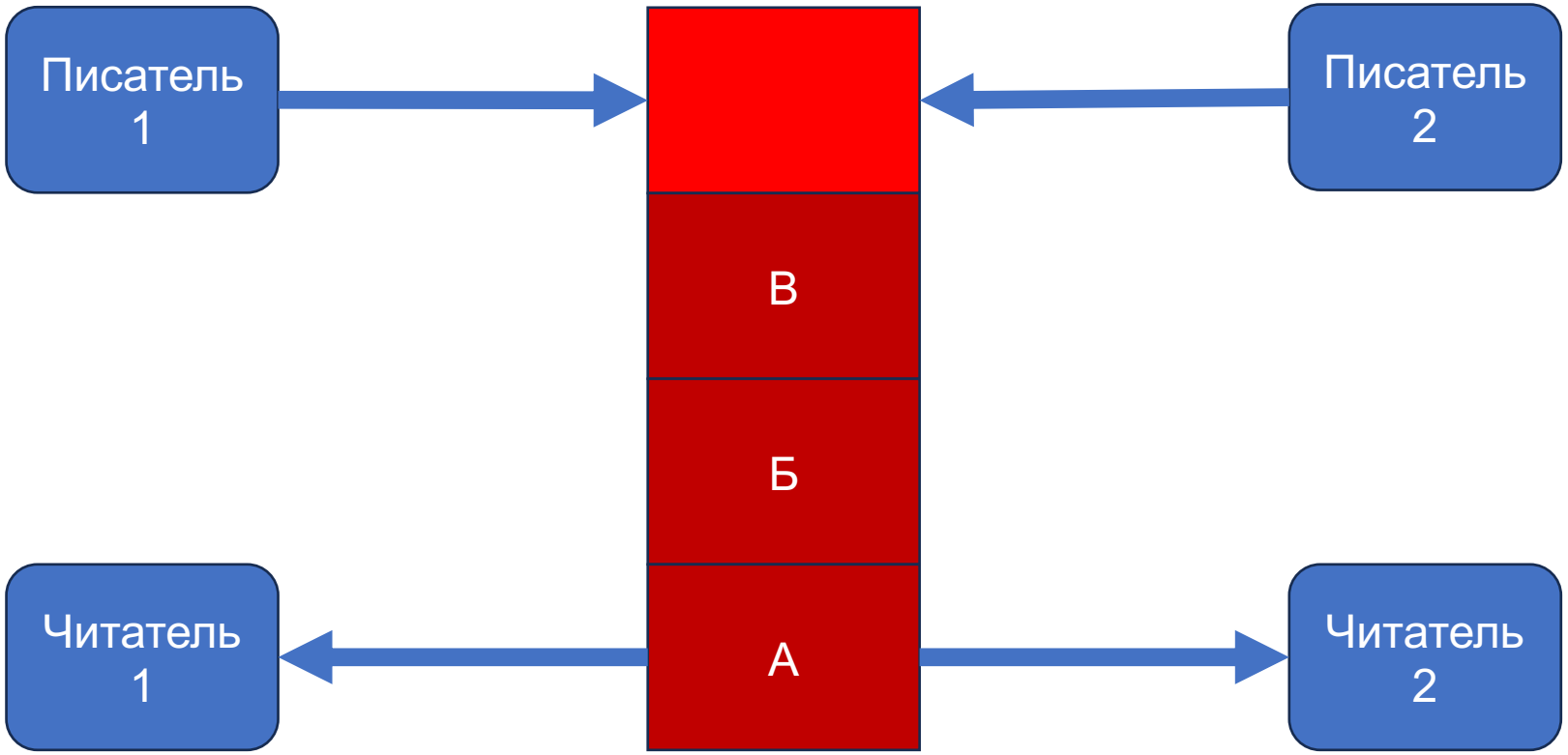


Как может быть устроена очередь активации?

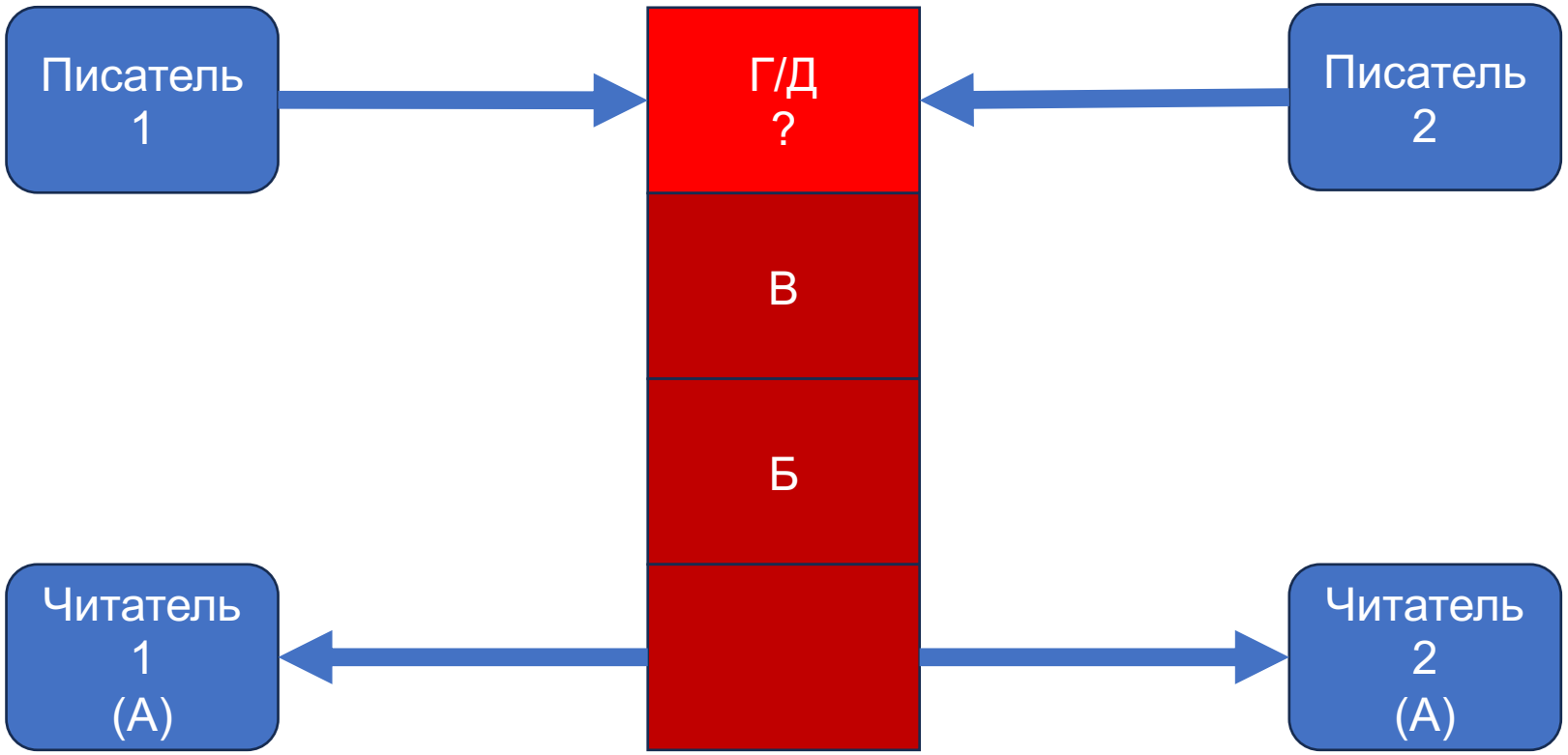


Примитивная очередь

Так никто не делает!



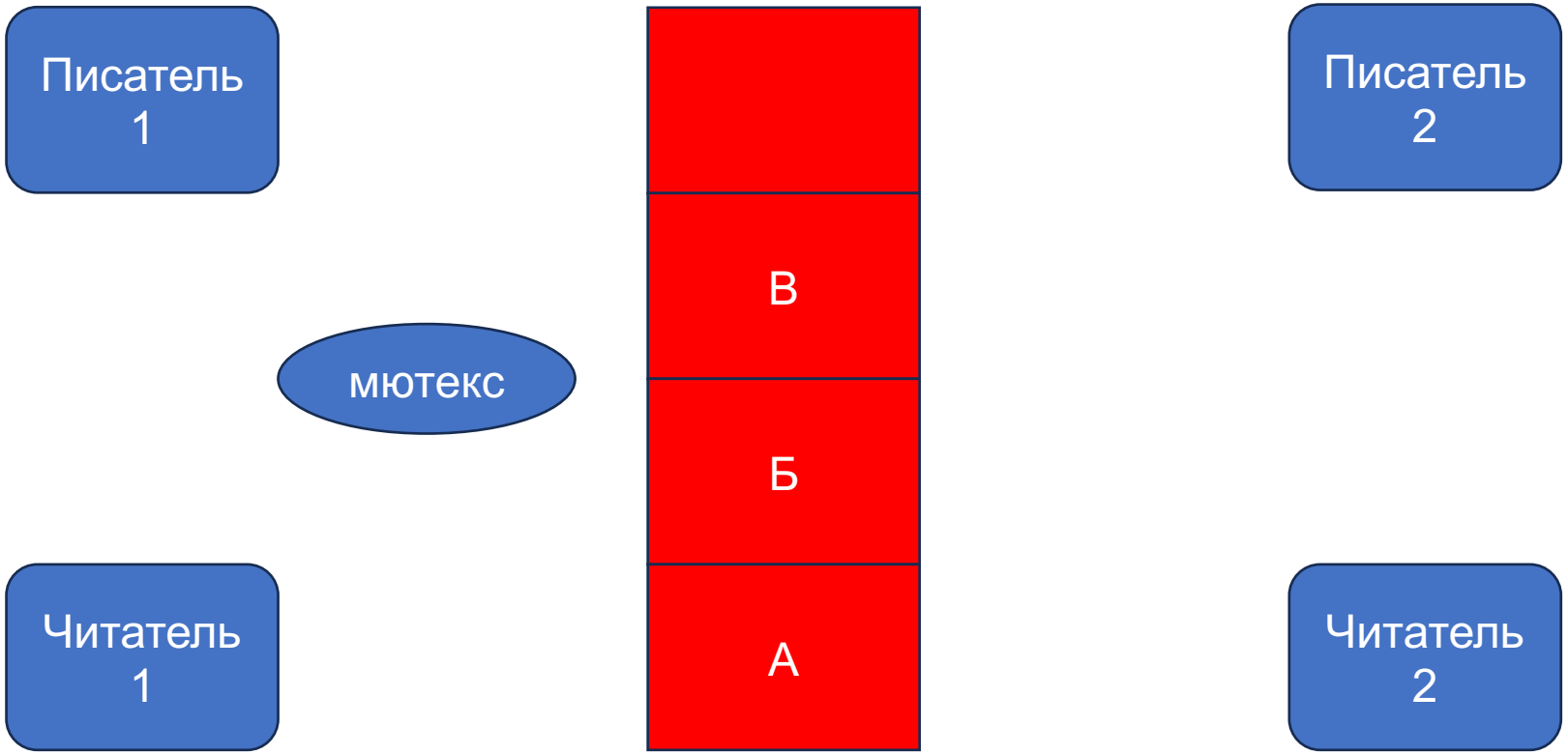
Примитивная очередь



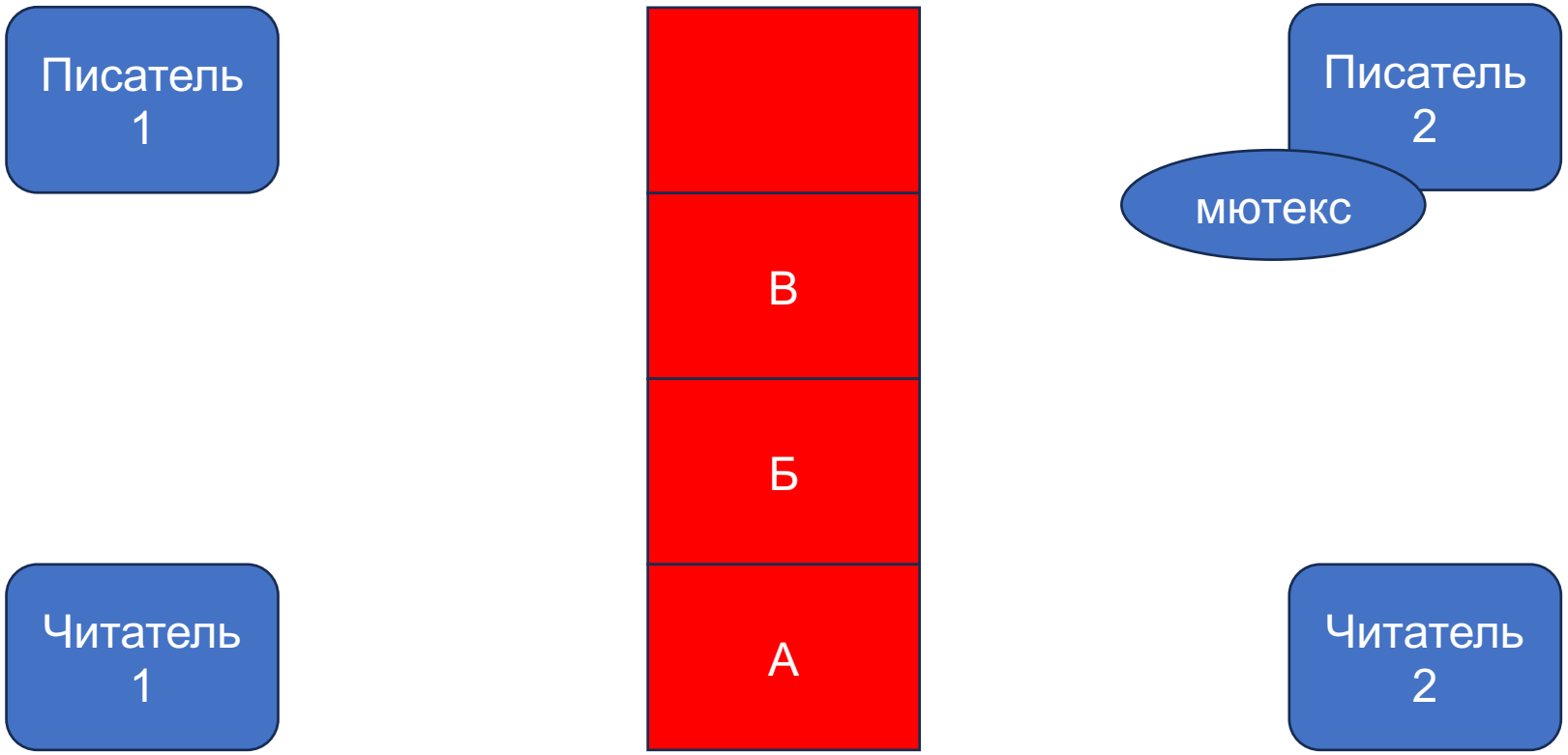
MPMC — Multiple Producers Multiple Consumers

Примитивная MPMC очередь с мютексом

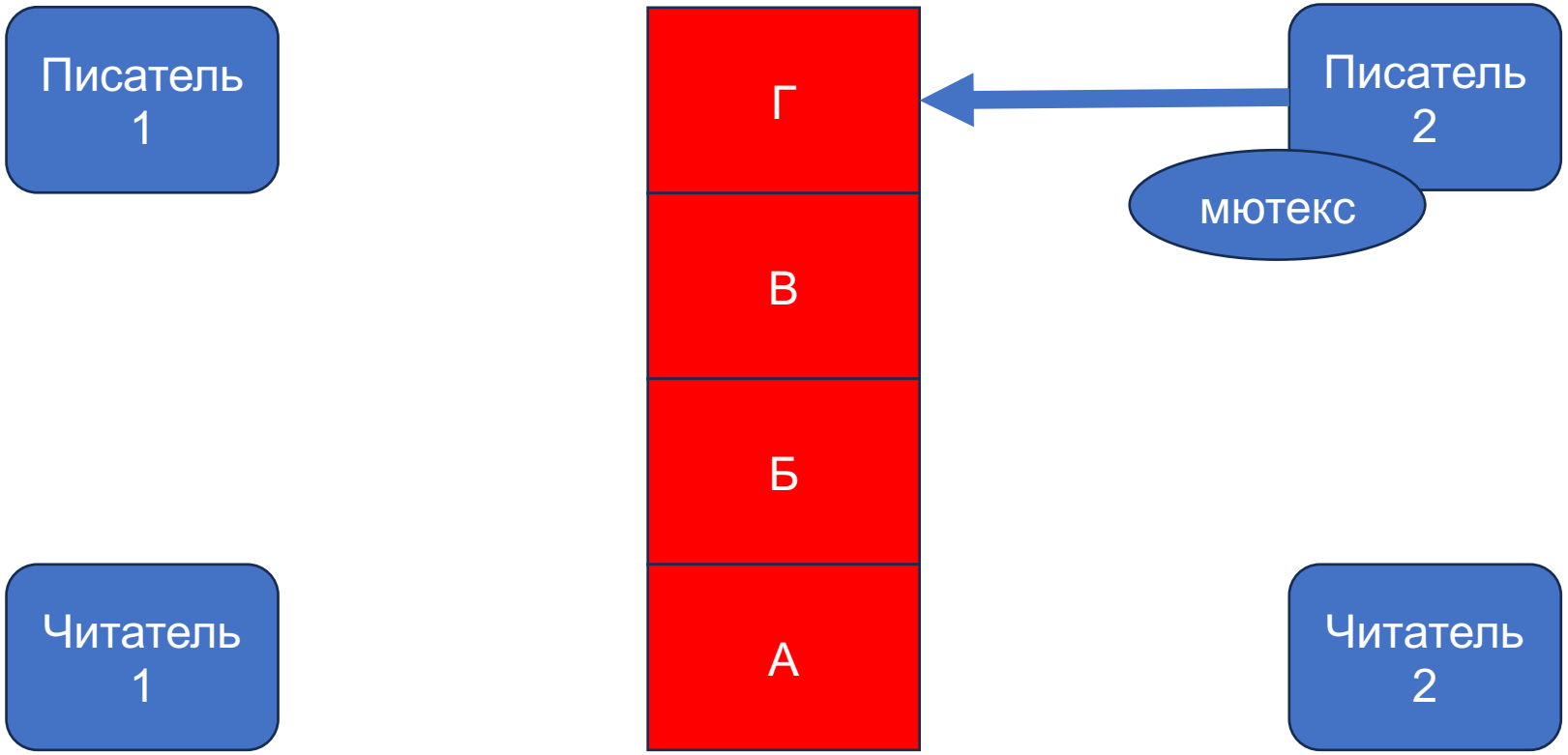
Все так делают.



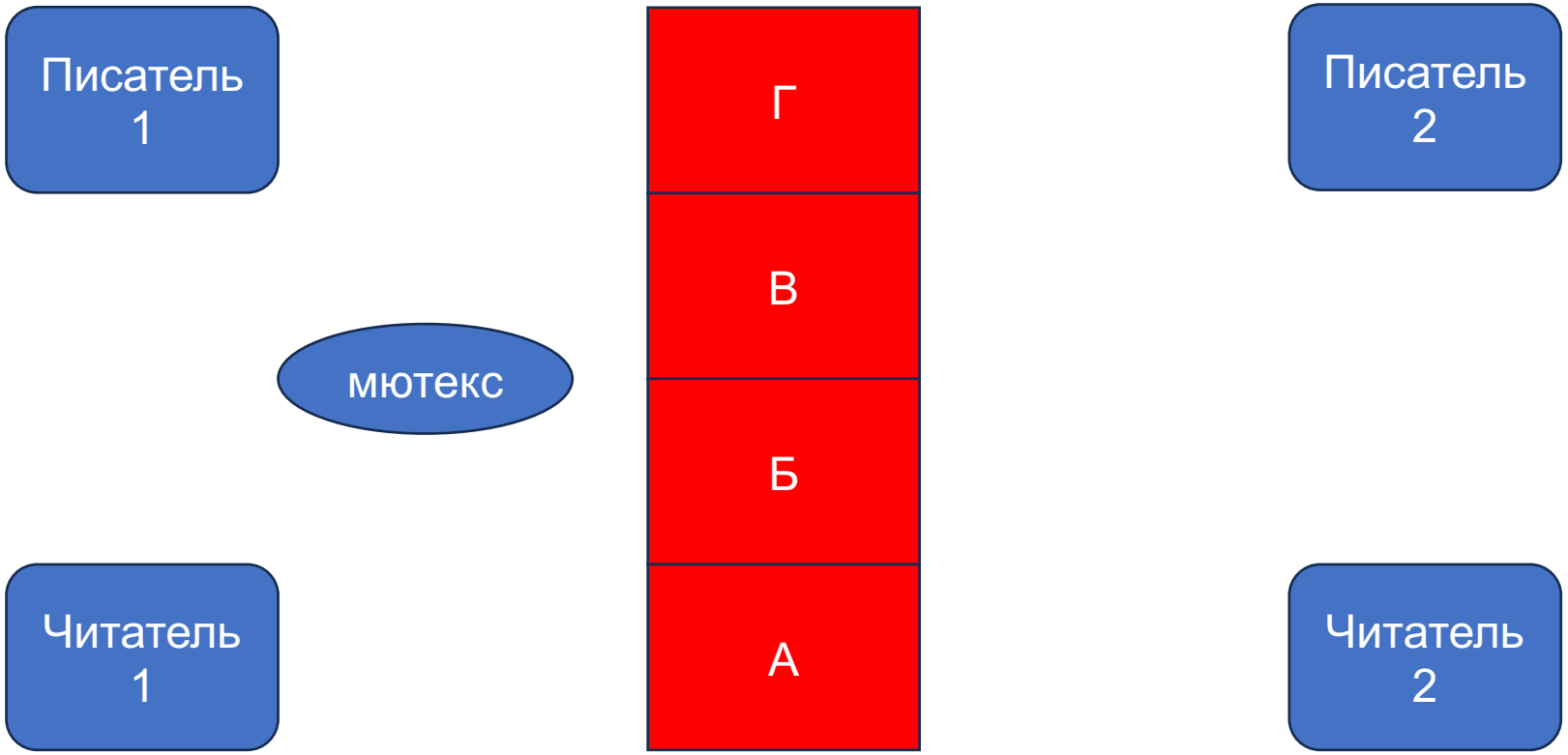
Примитивная MPMC очередь с мютексом



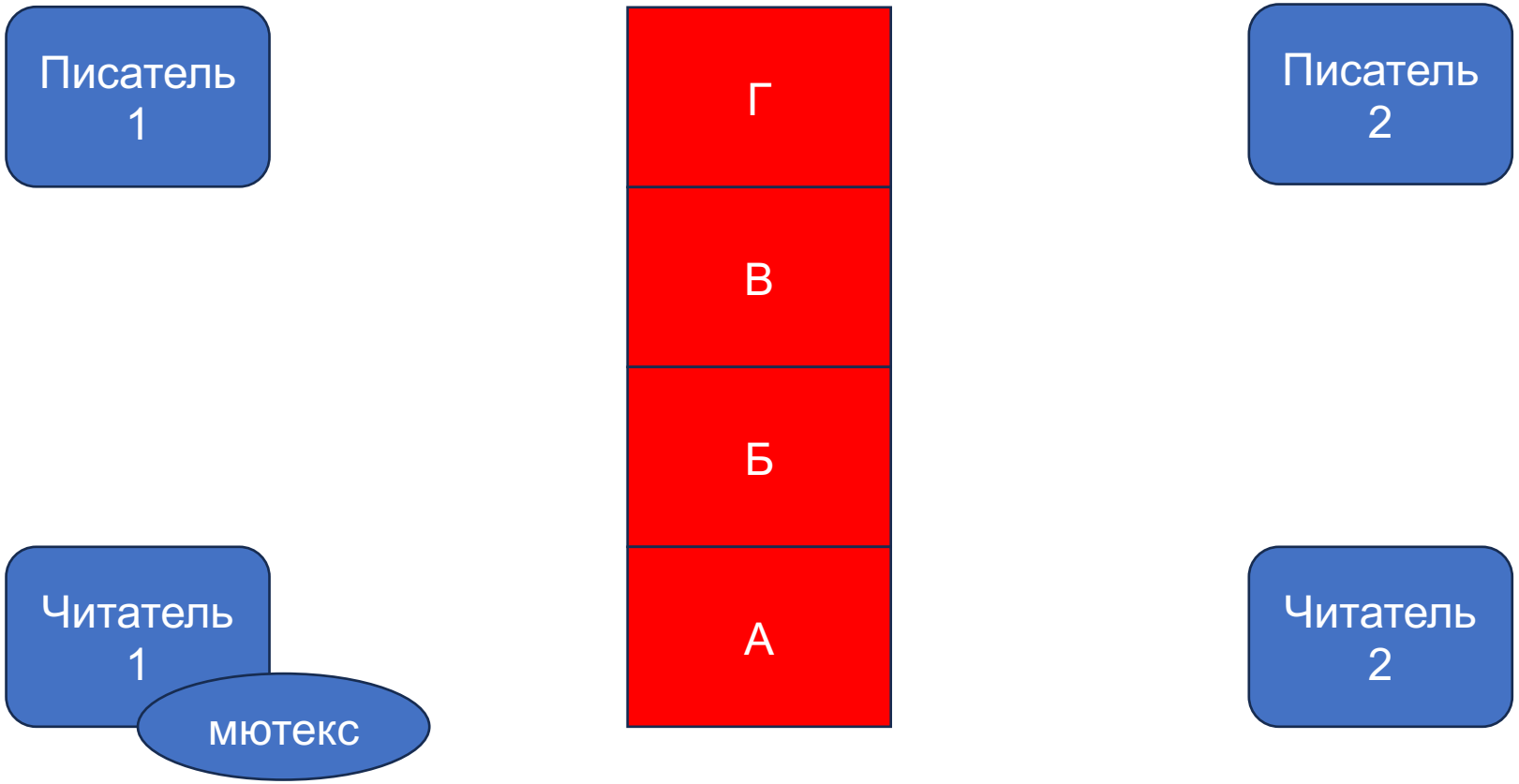
Примитивная MPMC очередь с мютексом



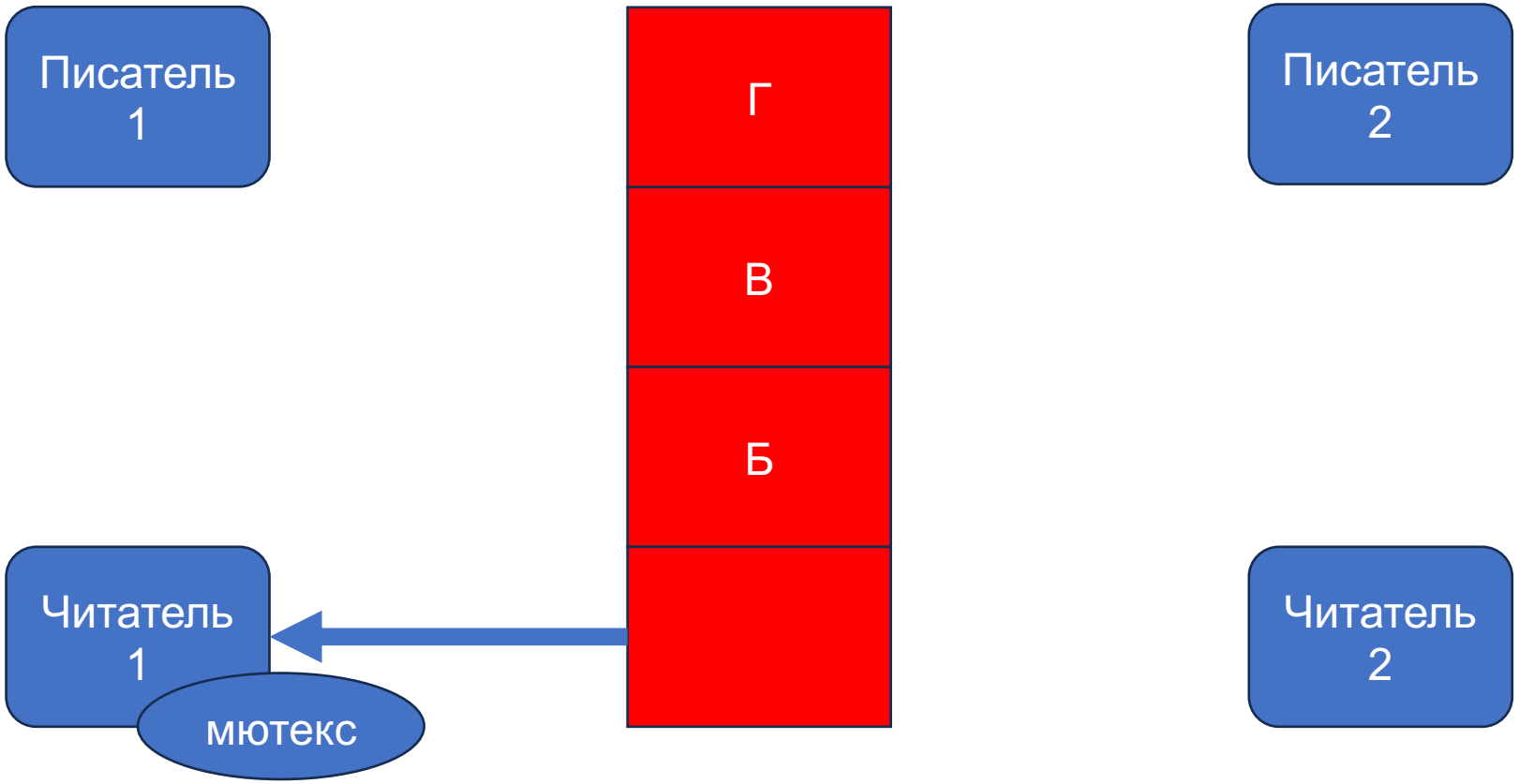
Примитивная MPMC очередь с мютексом



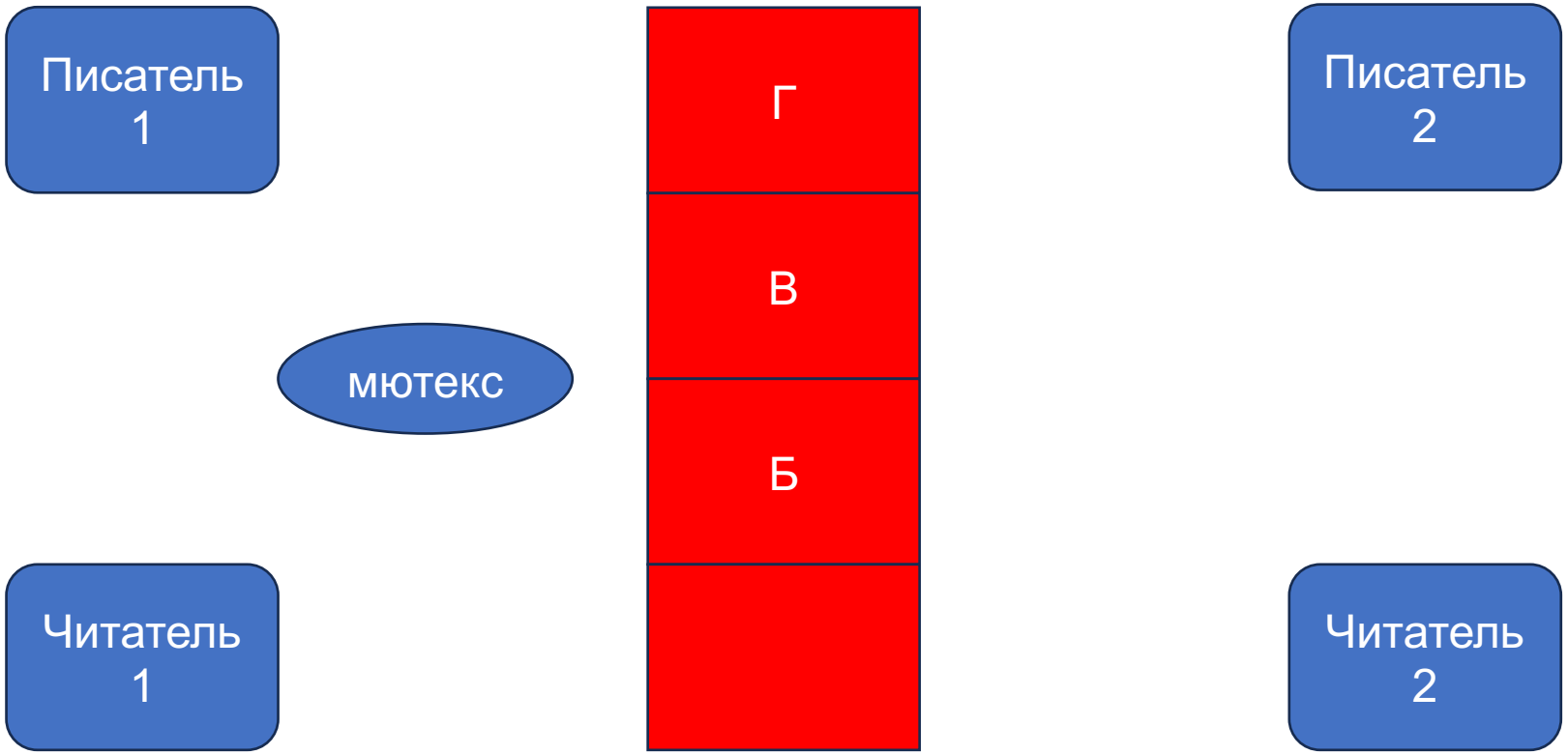
Примитивная MPMC очередь с мютексом



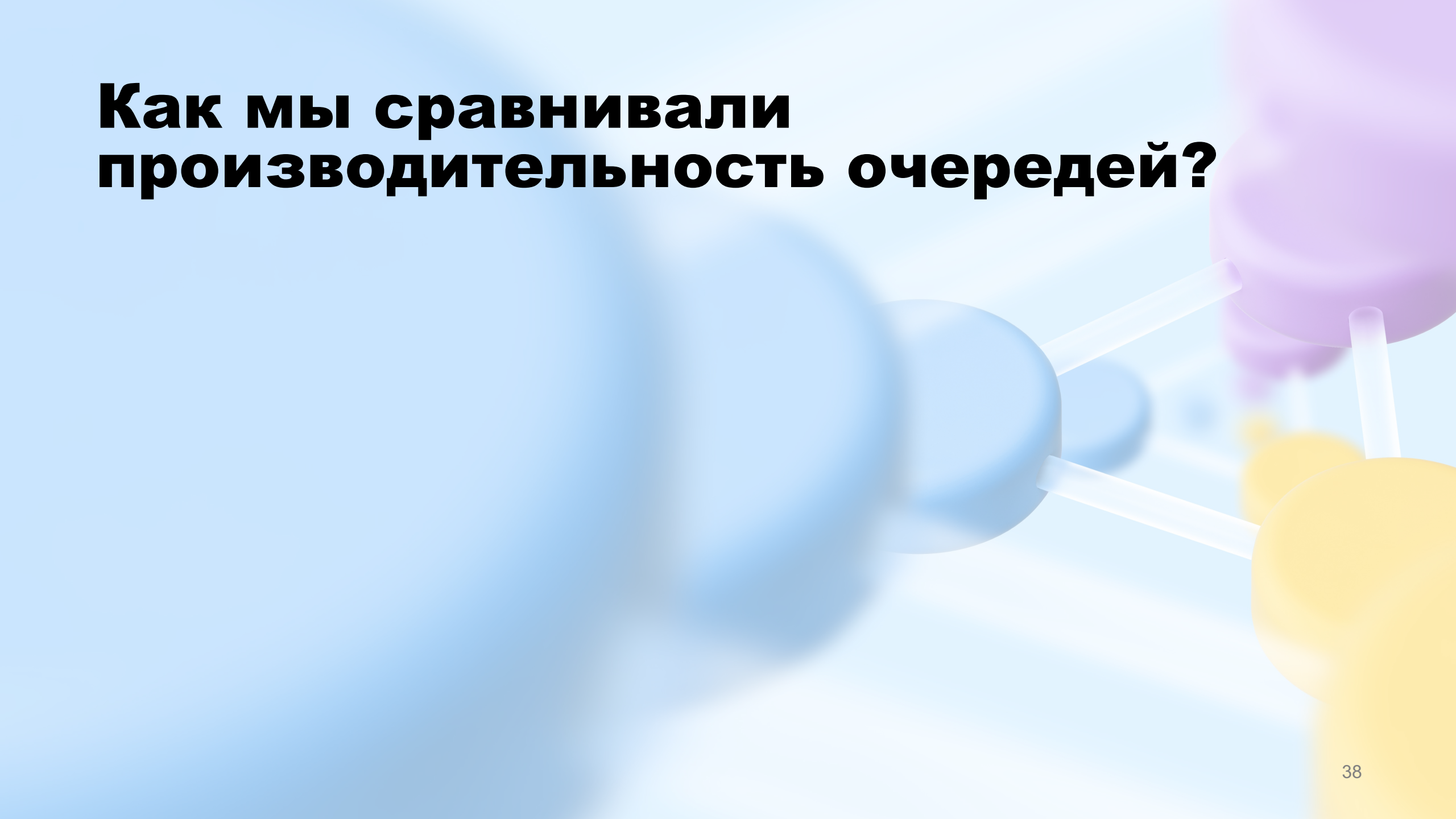
Примитивная MPMC очередь с мютексом



Примитивная MPMC очередь с мютексом



Как мы сравнивали производительность очередей?



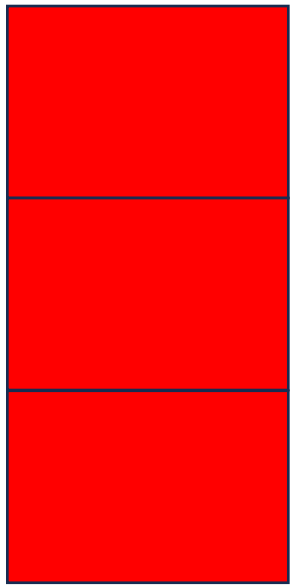
Описание бенчмарка

Поток 1

Поток 2

Поток 3

Поток 4



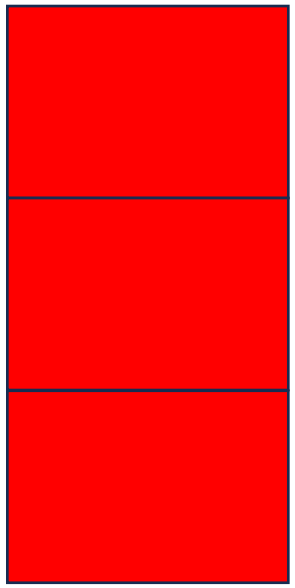
Описание бенчмарка

Поток 1

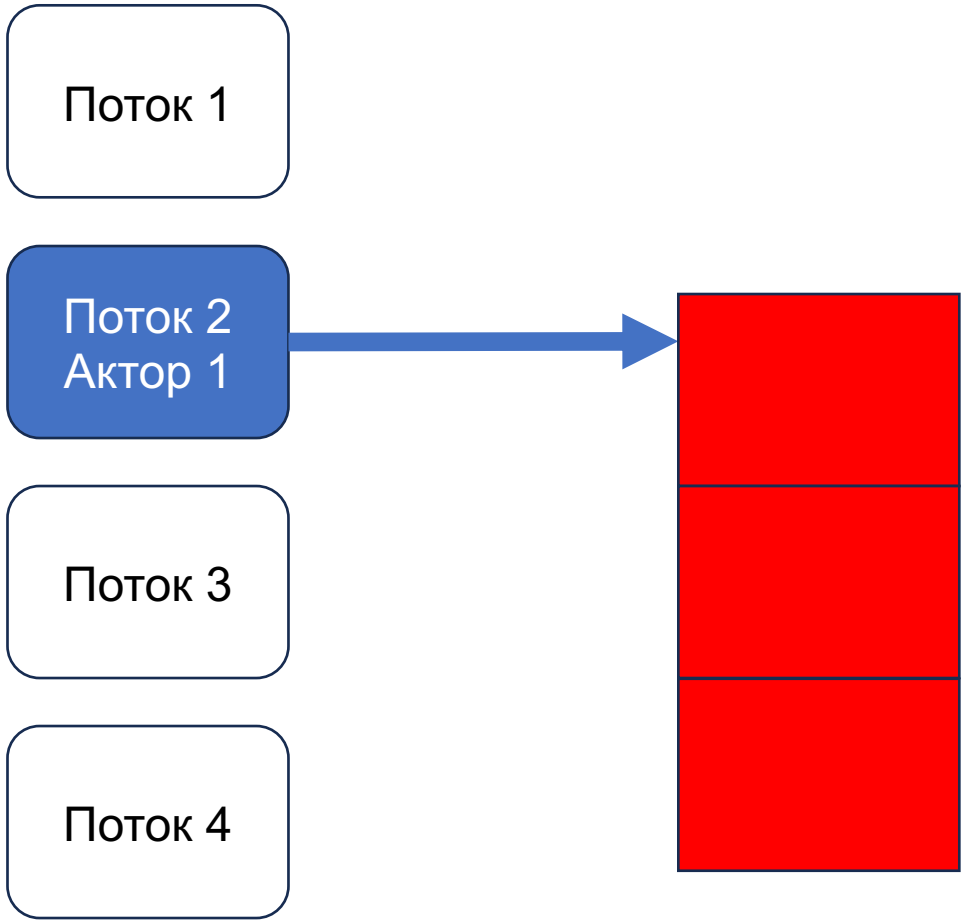
Поток 2
Актор 1

Поток 3

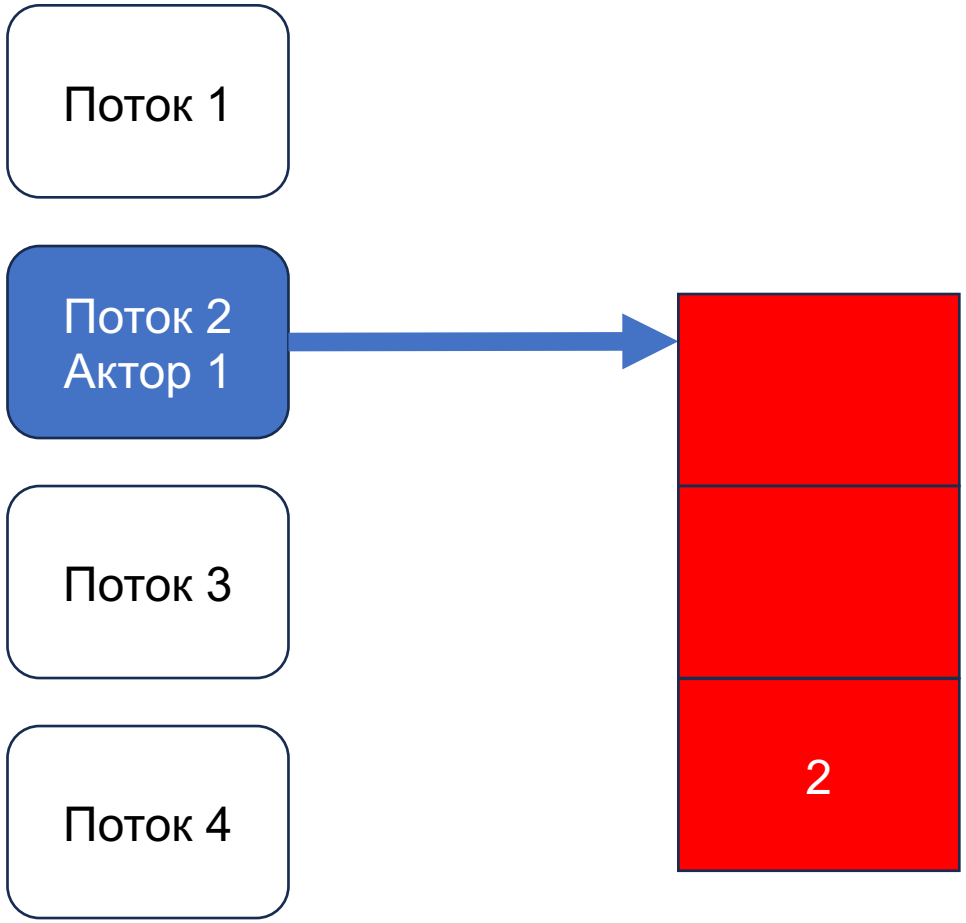
Поток 4



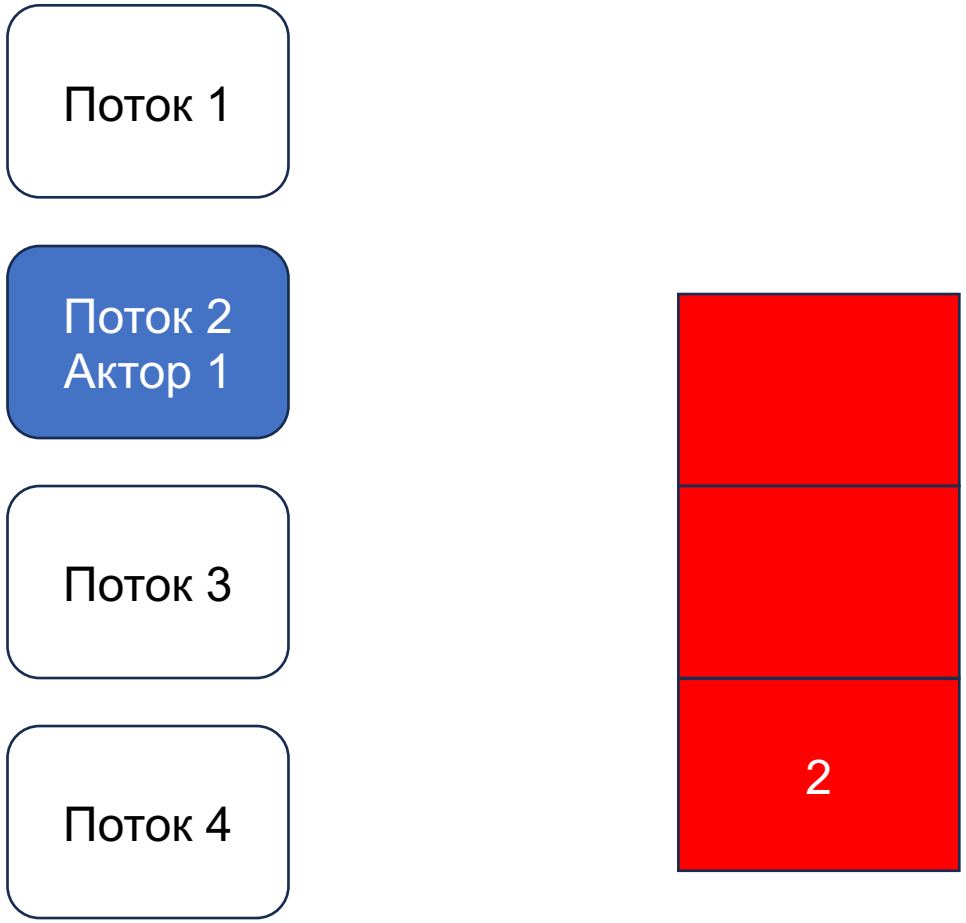
Описание бенчмарка



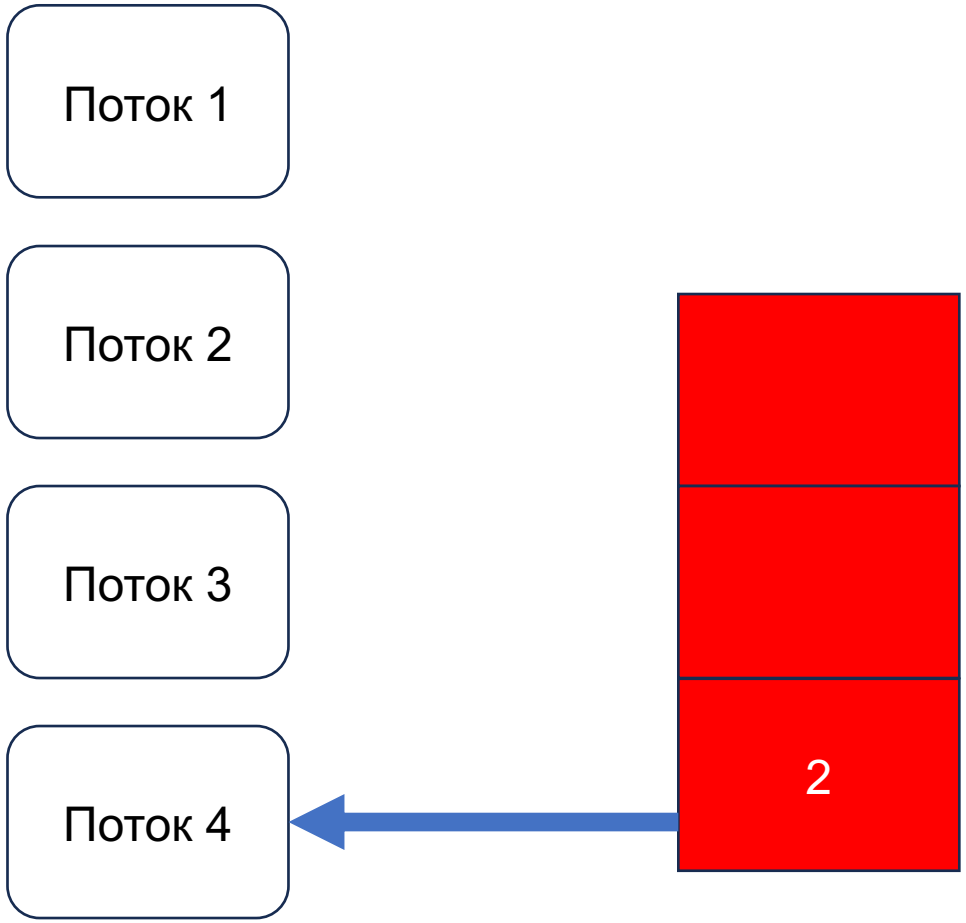
Описание бенчмарка



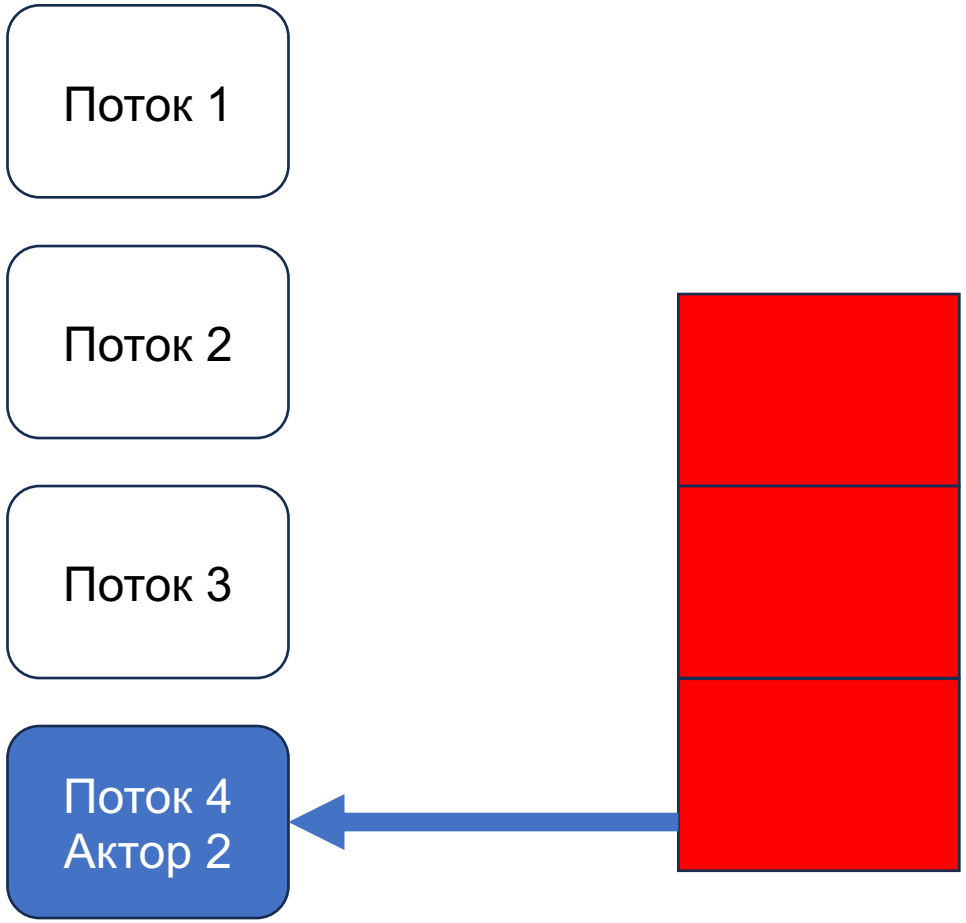
Описание бенчмарка



Описание бенчмарка



Описание бенчмарка



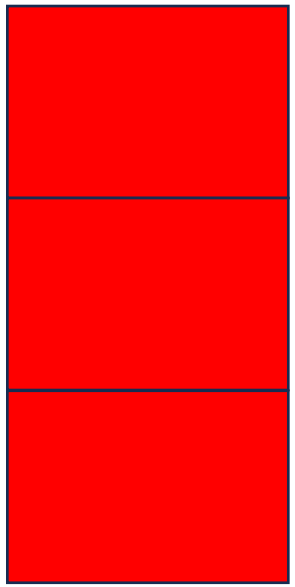
Описание бенчмарка

Поток 1

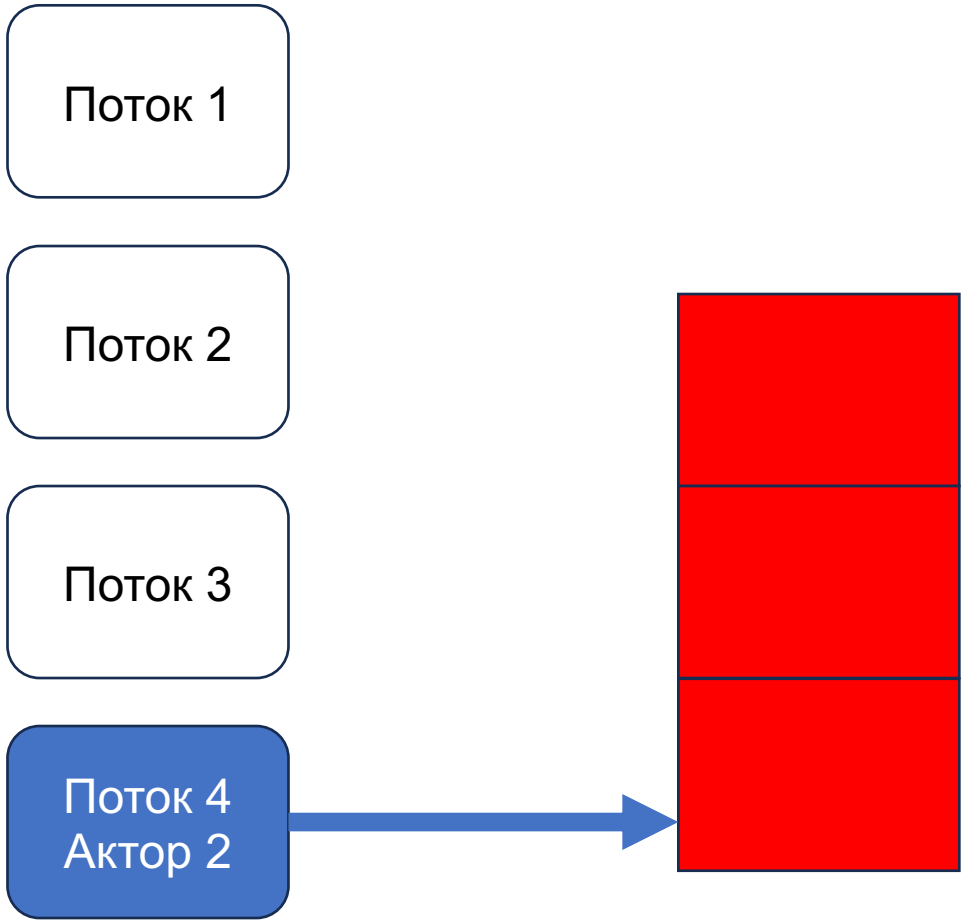
Поток 2

Поток 3

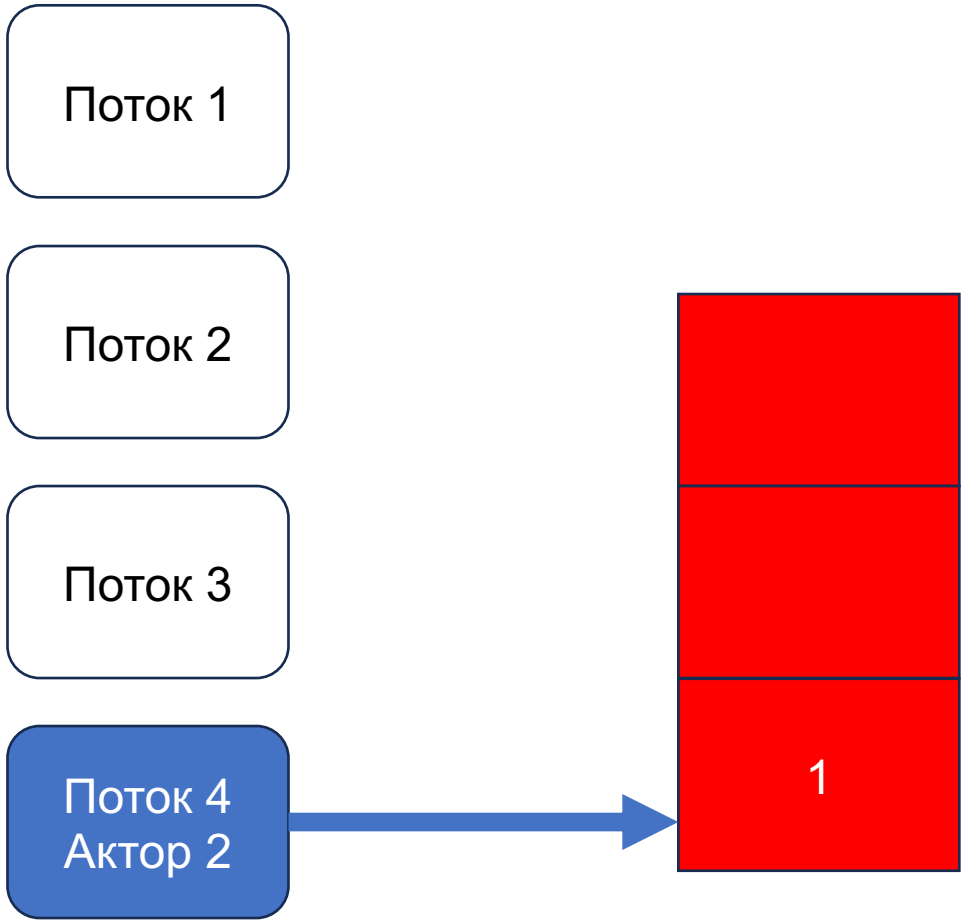
Поток 4
Актор 2



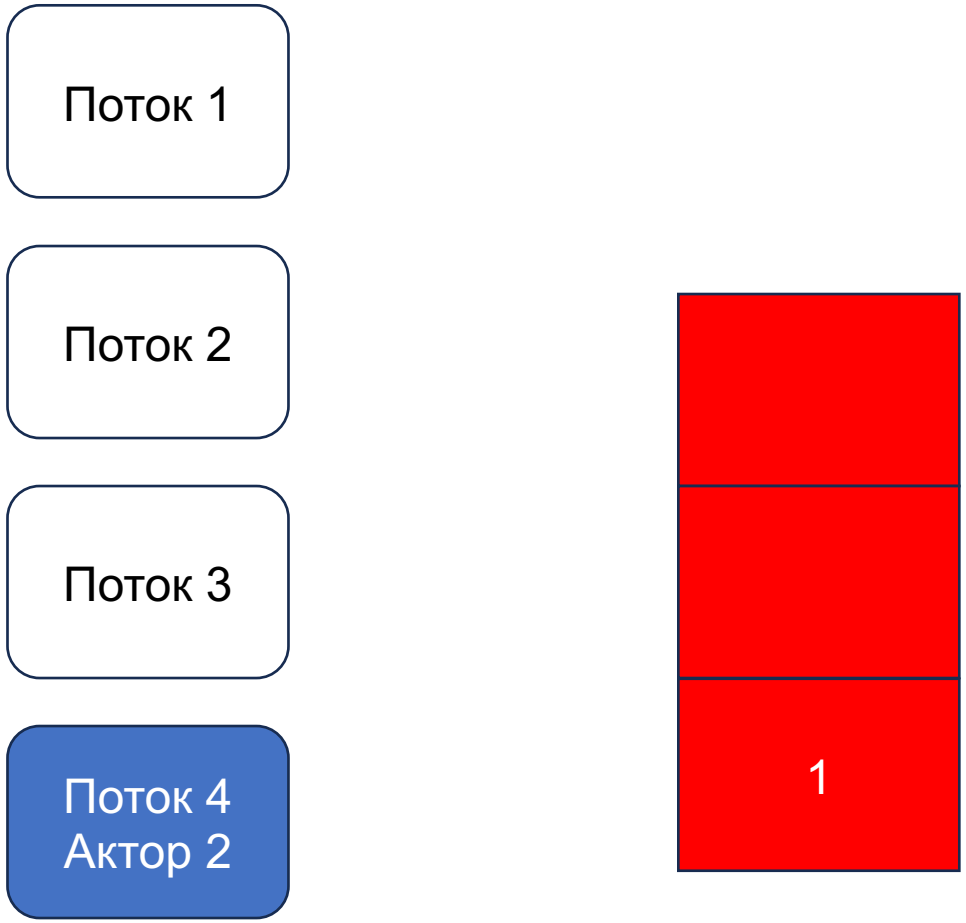
Описание бенчмарка



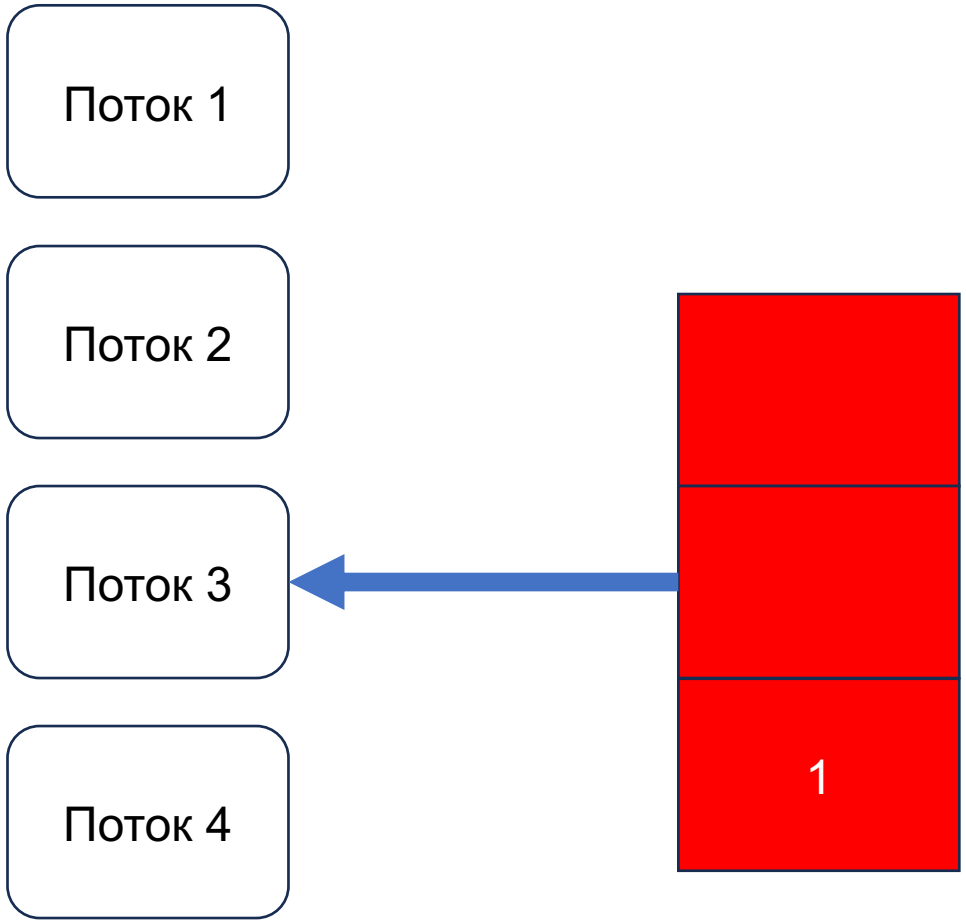
Описание бенчмарка



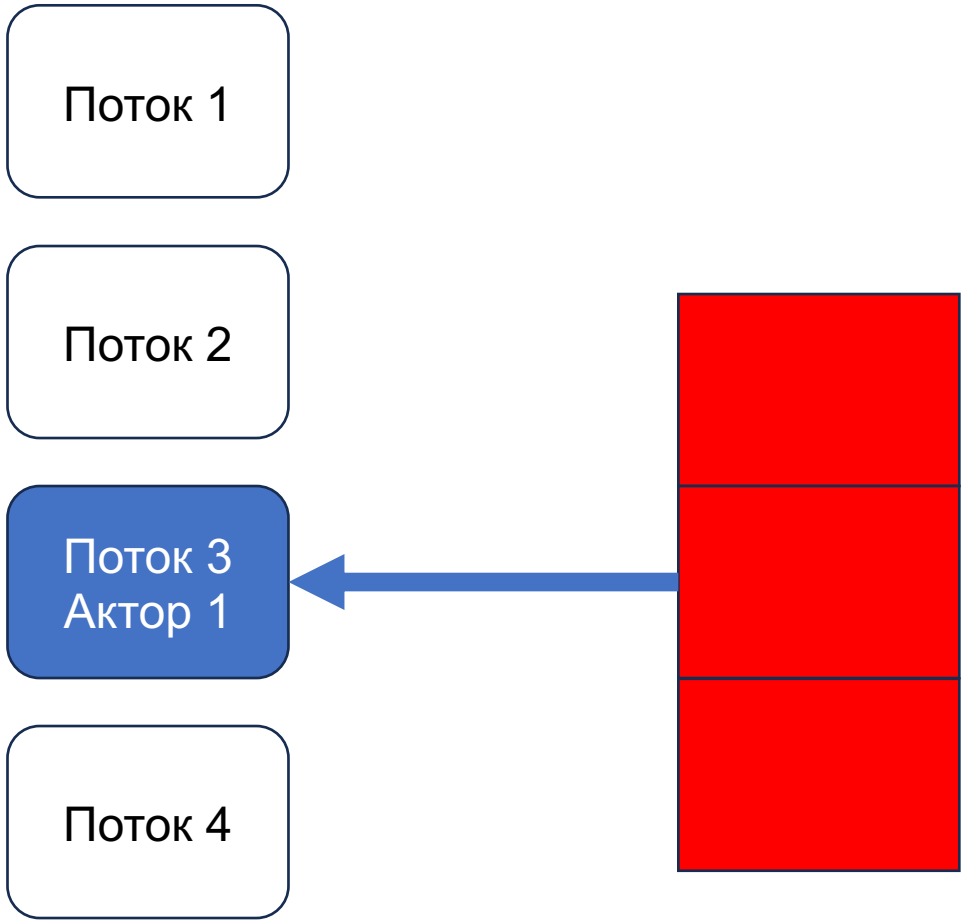
Описание бенчмарка



Описание бенчмарка



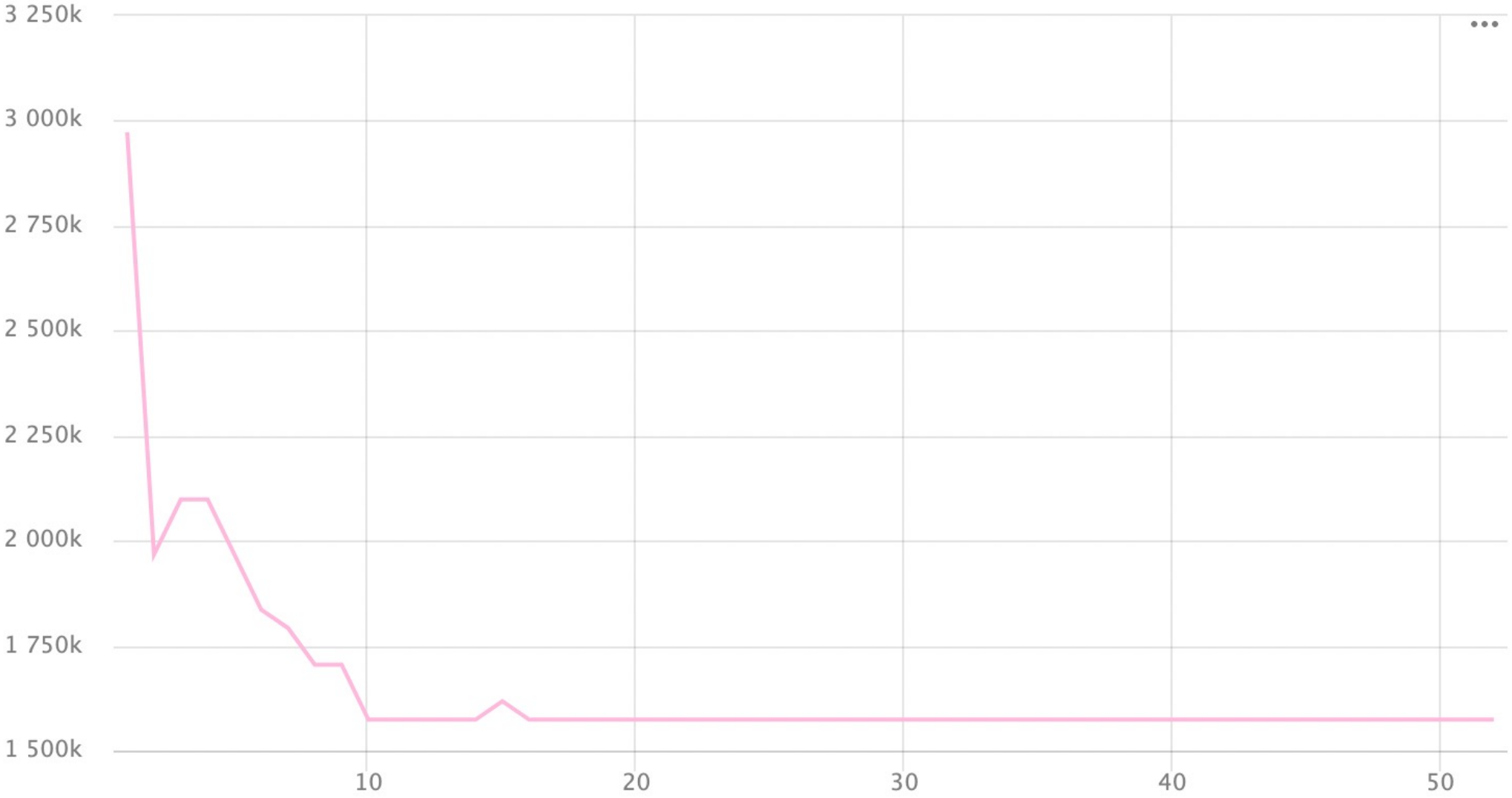
Описание бенчмарка



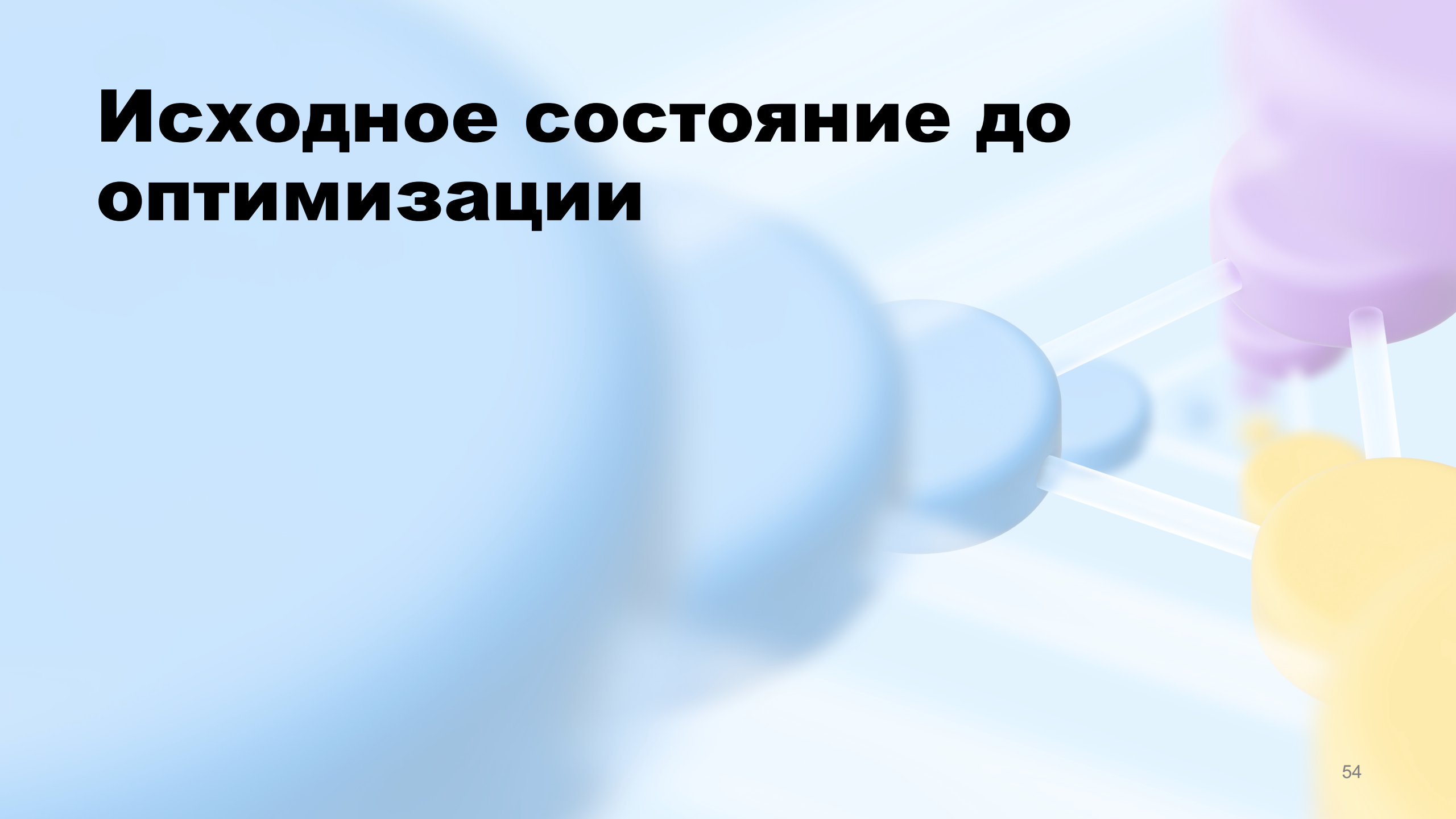
Описание бенчмарка

- 512 пар акторов
- От 1 до 56 потоков
- Измеряем количество активаций в секунду

Примитивная MPMC очередь с мютексом

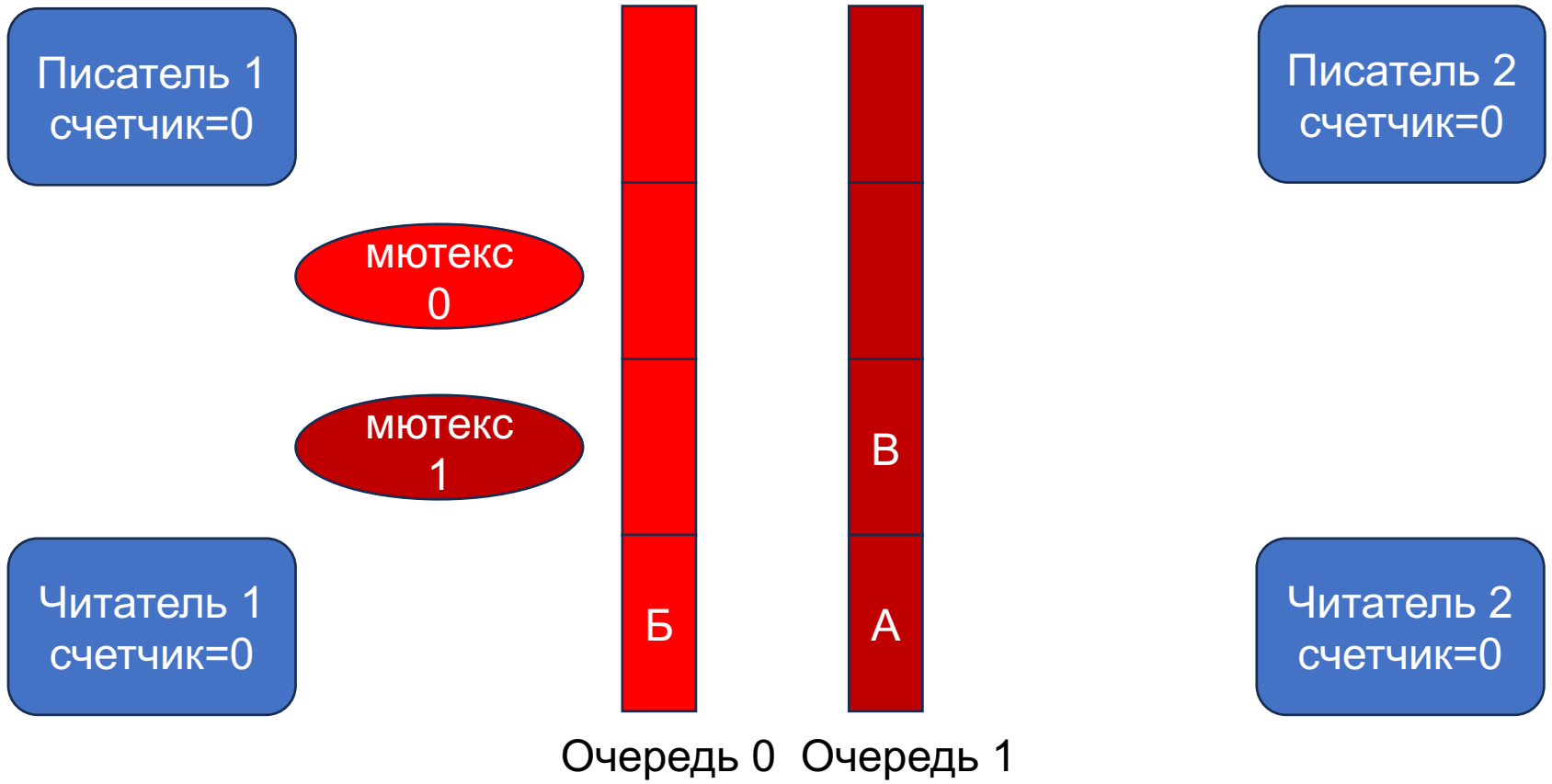


Исходное состояние до оптимизации

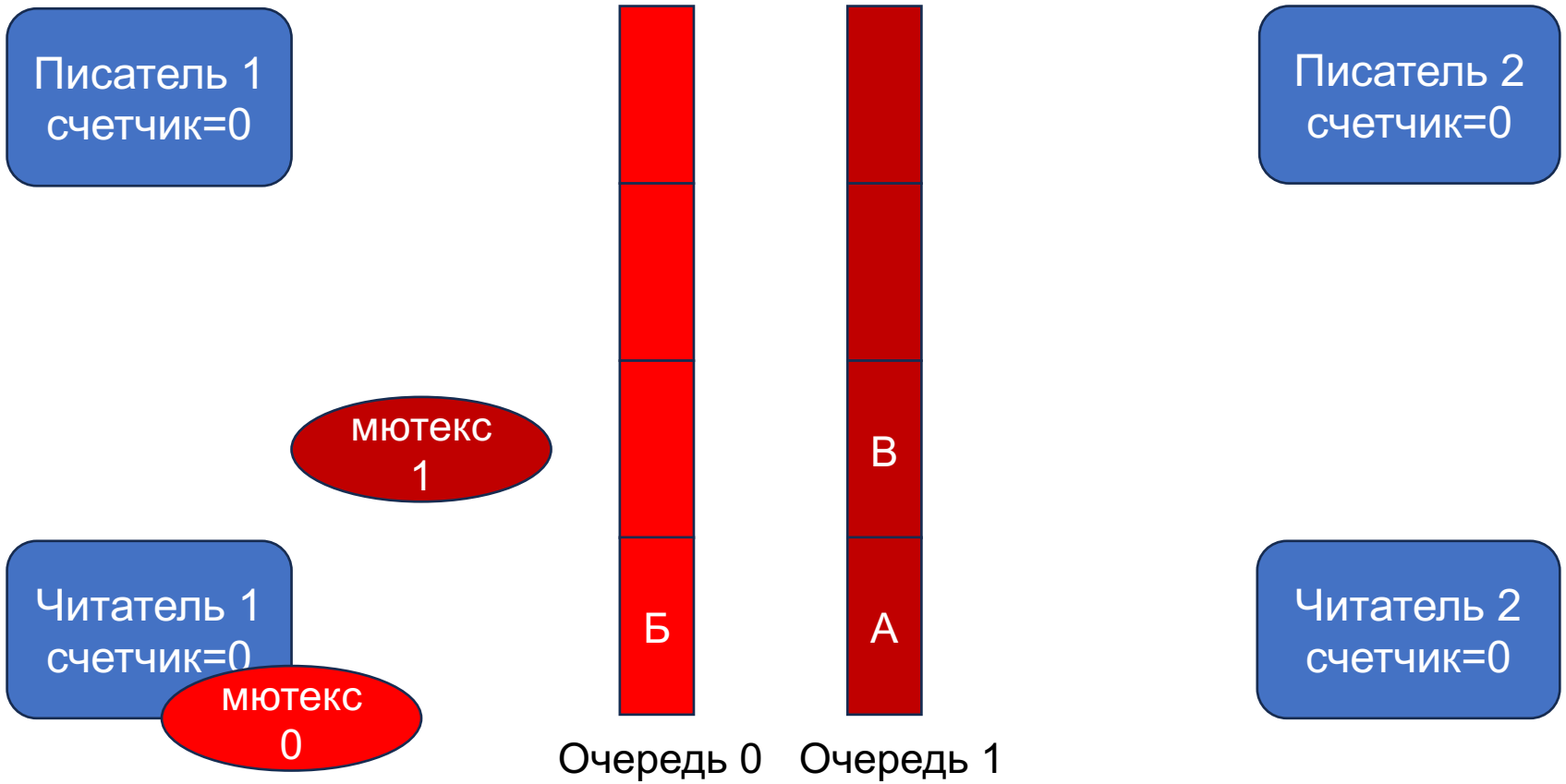


Revolving MPMC queue

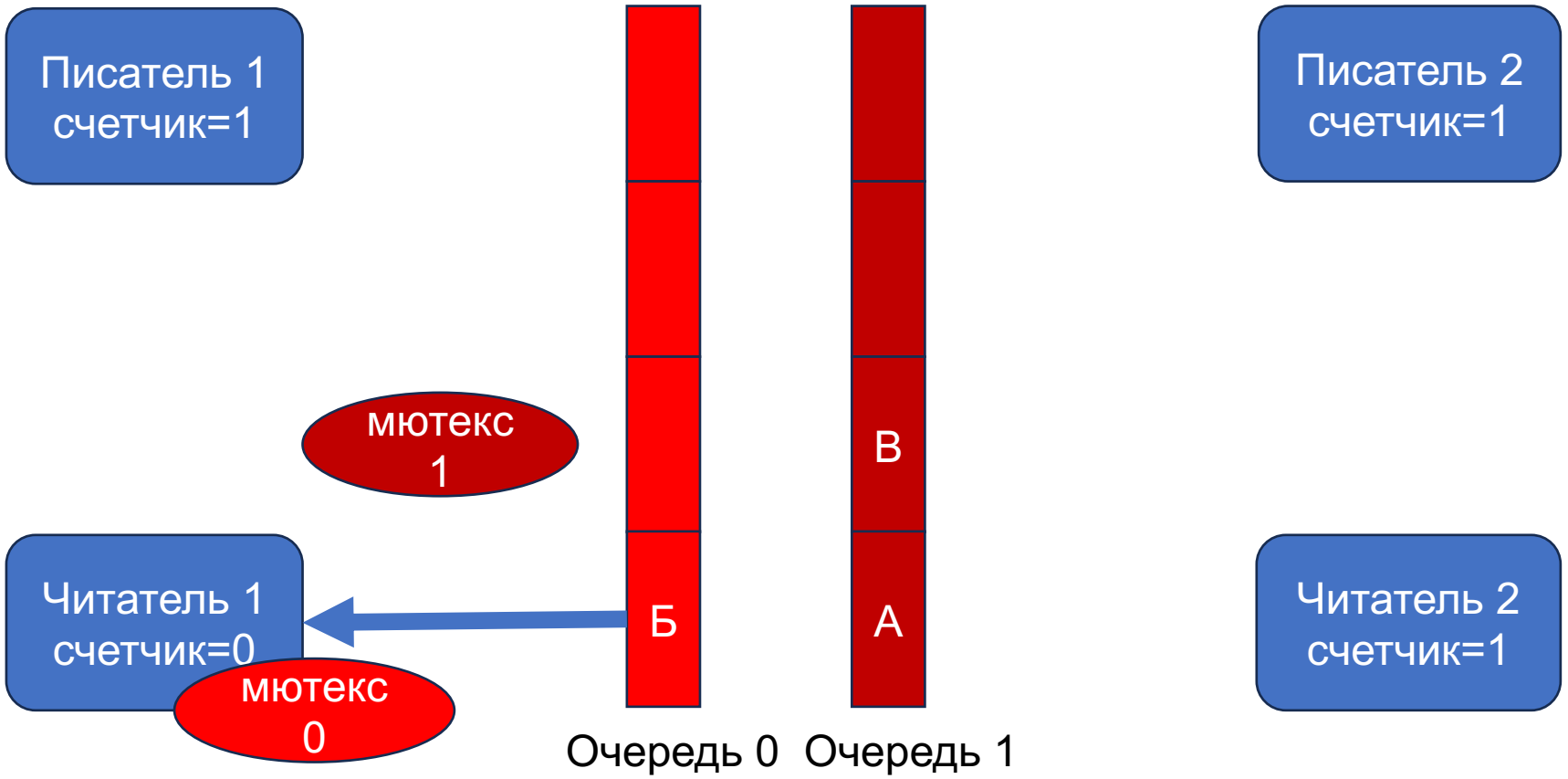
Исходное состояние до оптимизации: набор очередей с per-queue блокировками.



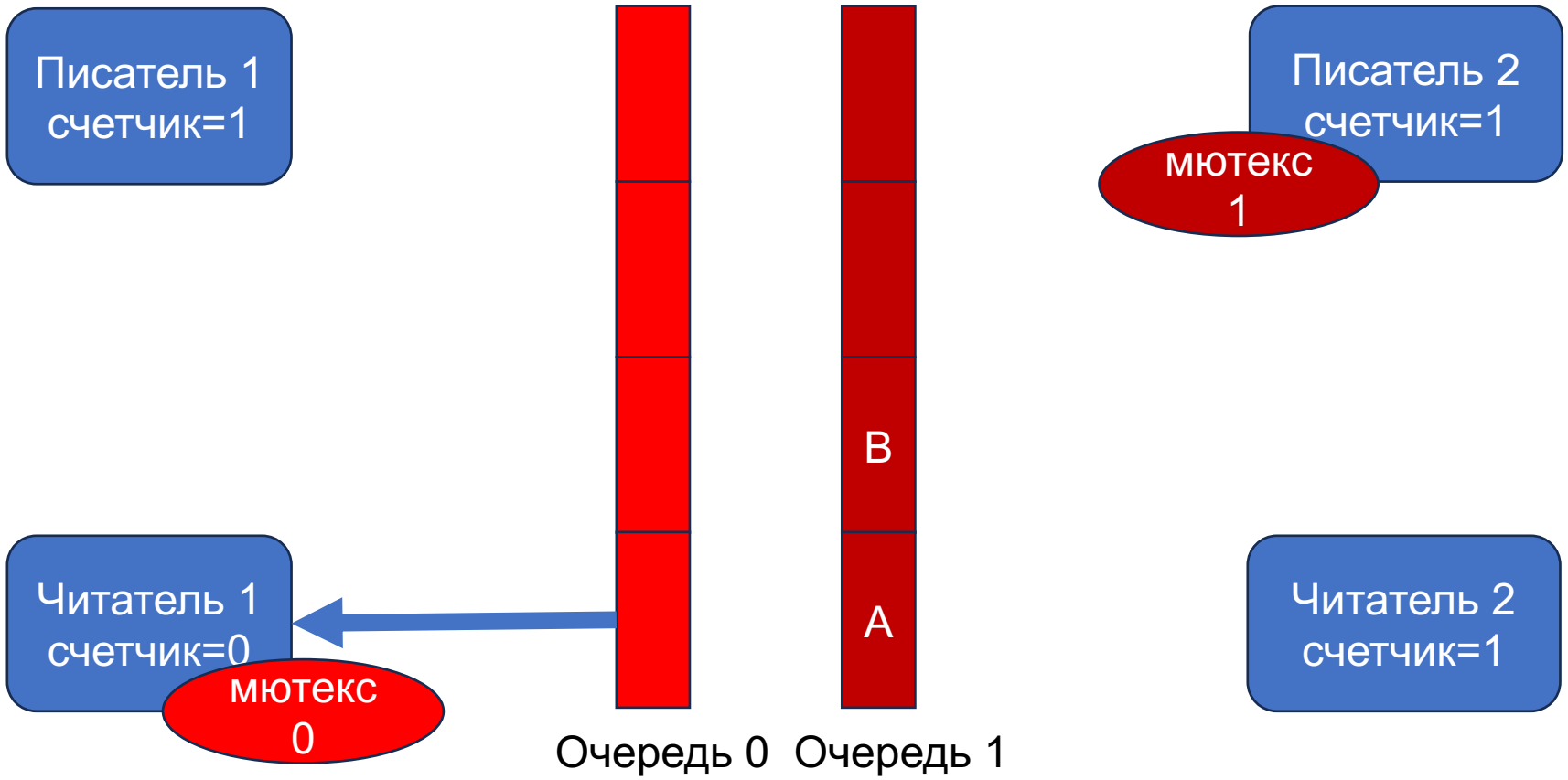
Revolving MPMC queue



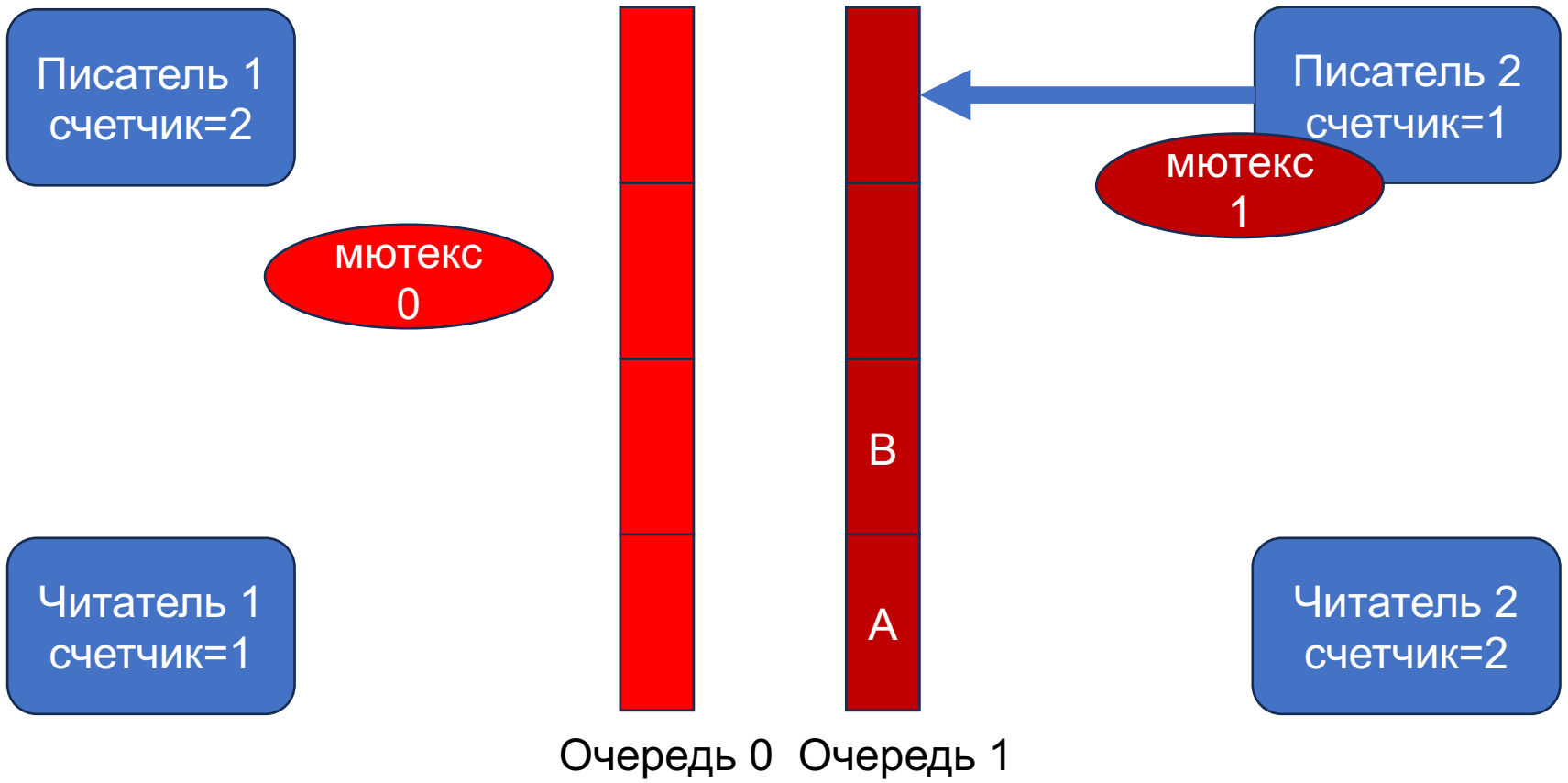
Revolving MPMC queue



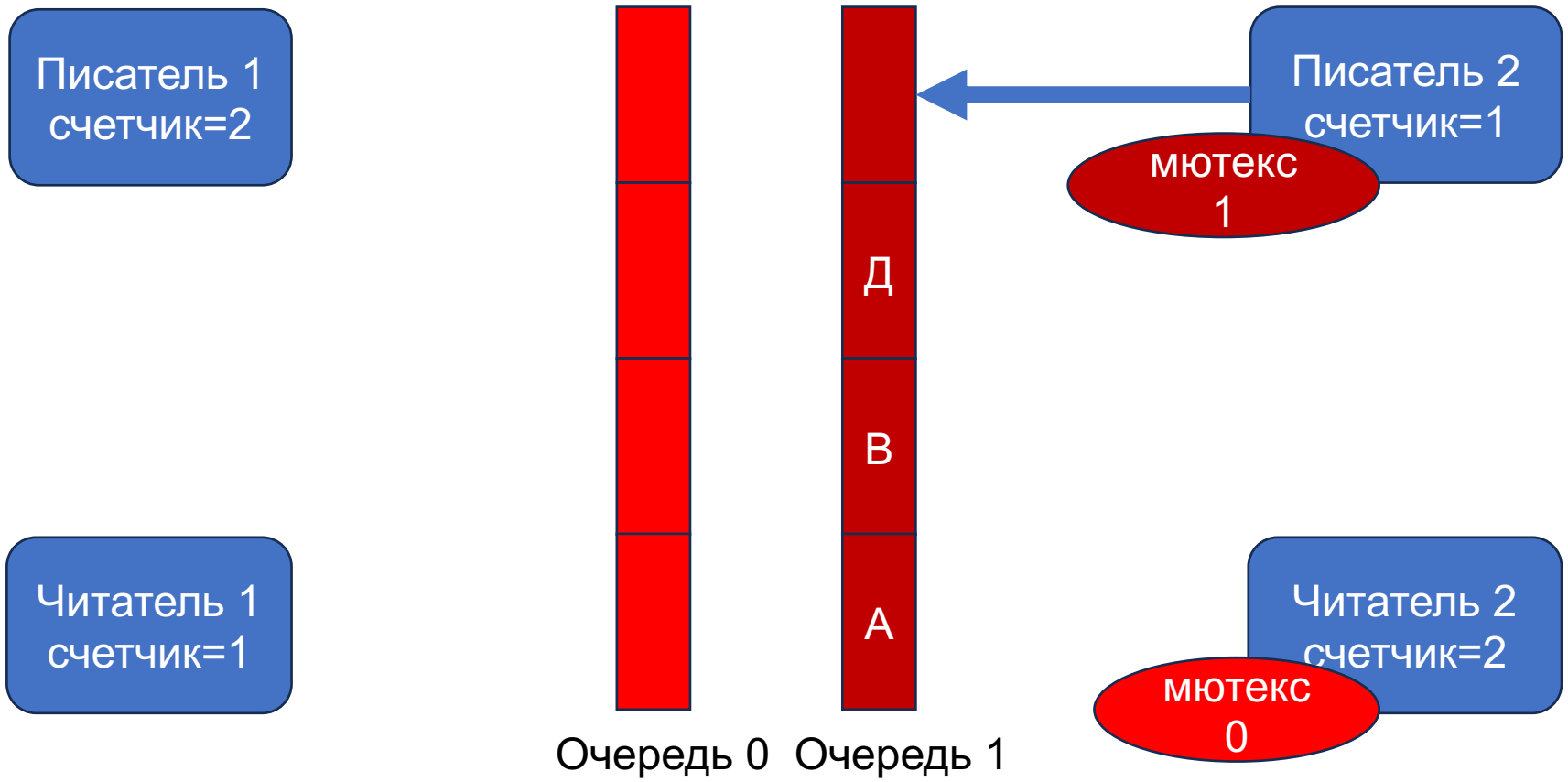
Revolving MPMC queue



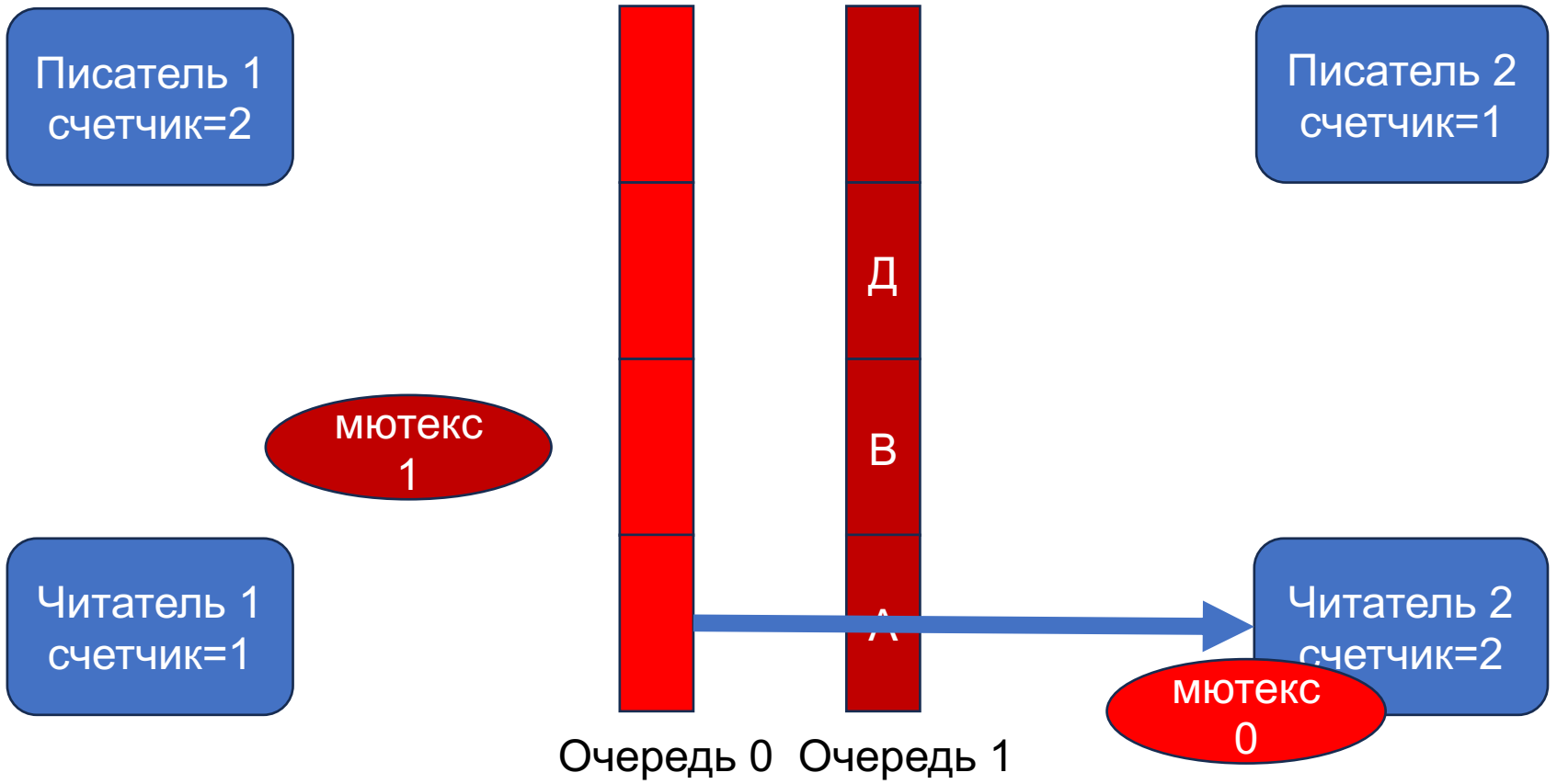
Revolving MPMC queue



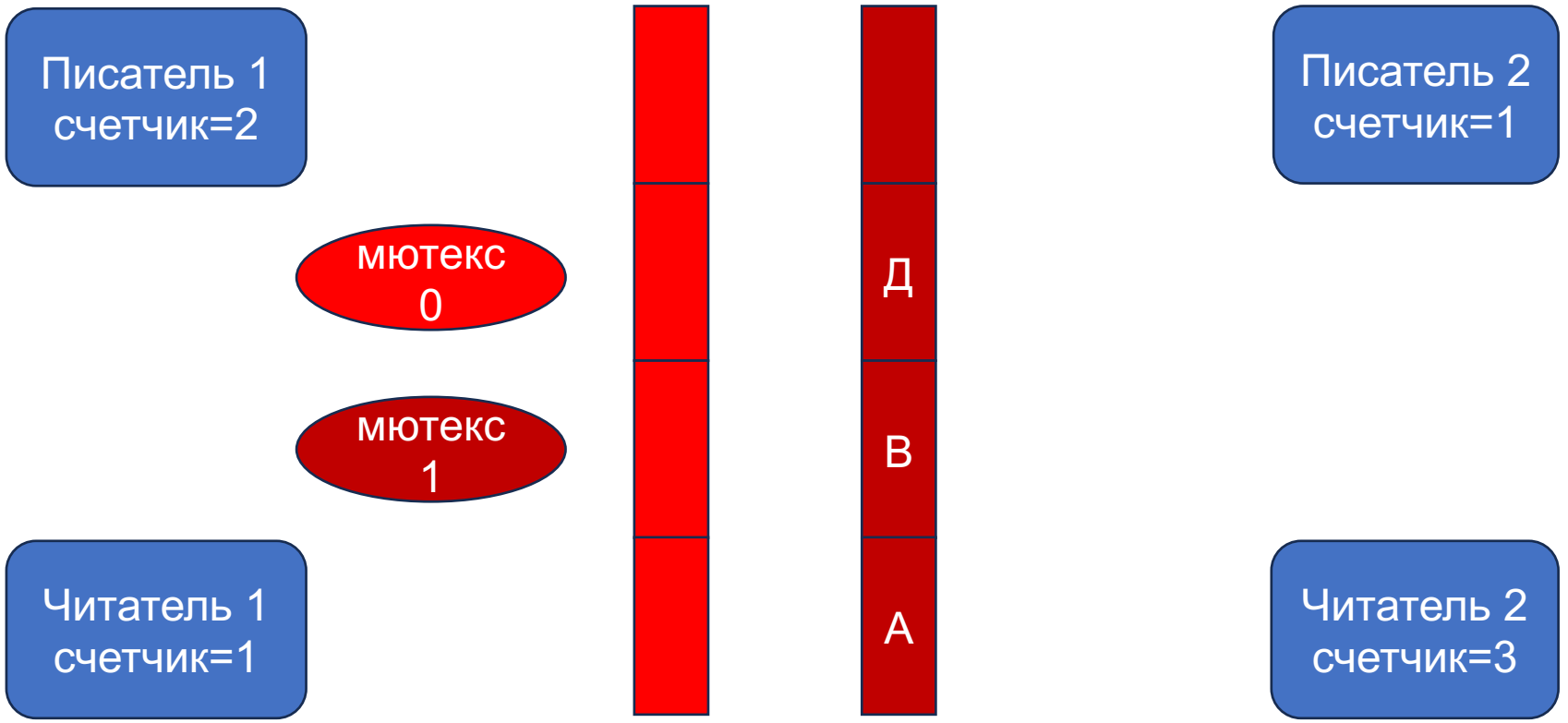
Revolving MPMC queue



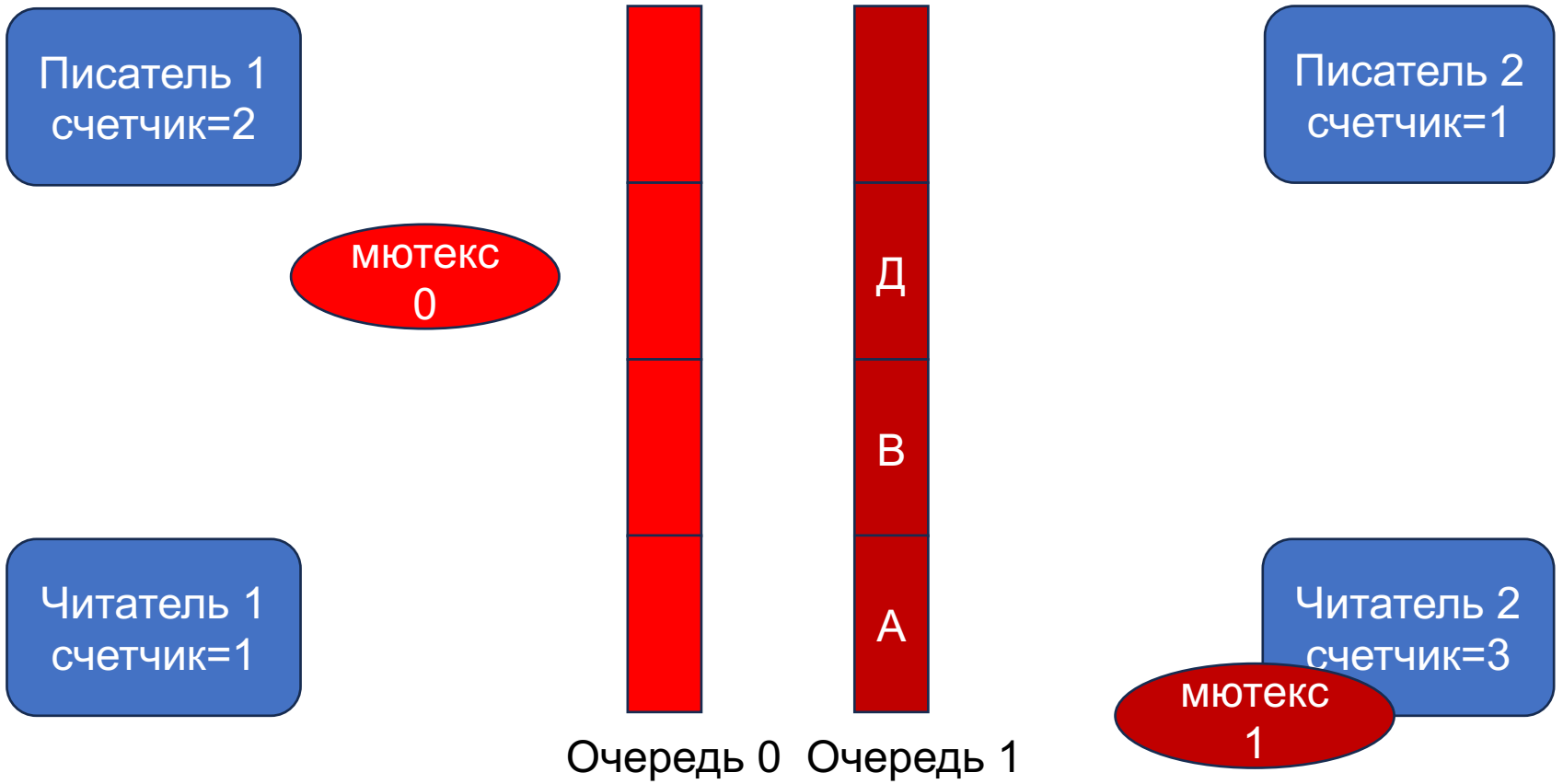
Revolving MPMC queue



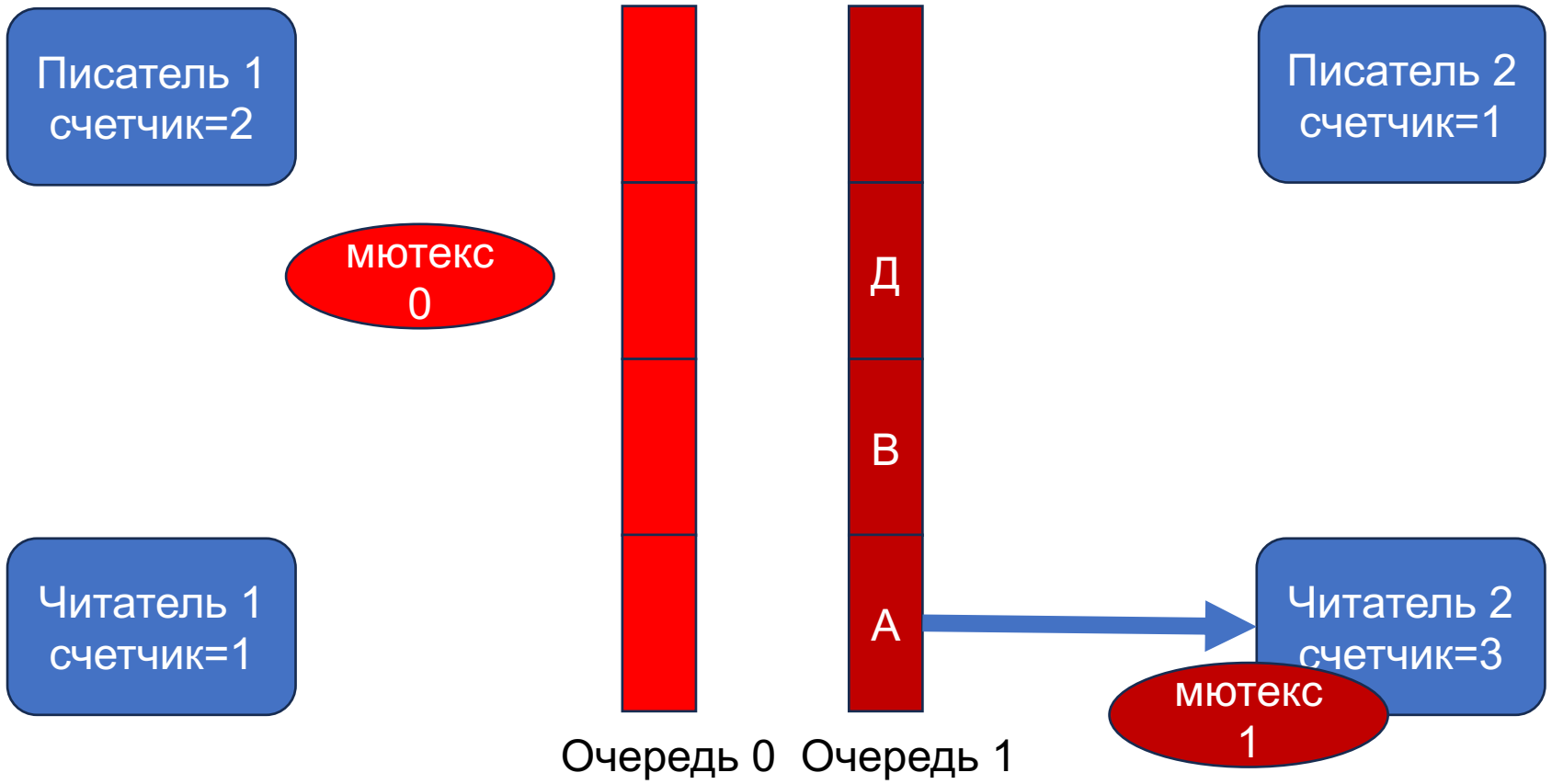
Revolving MPMC queue



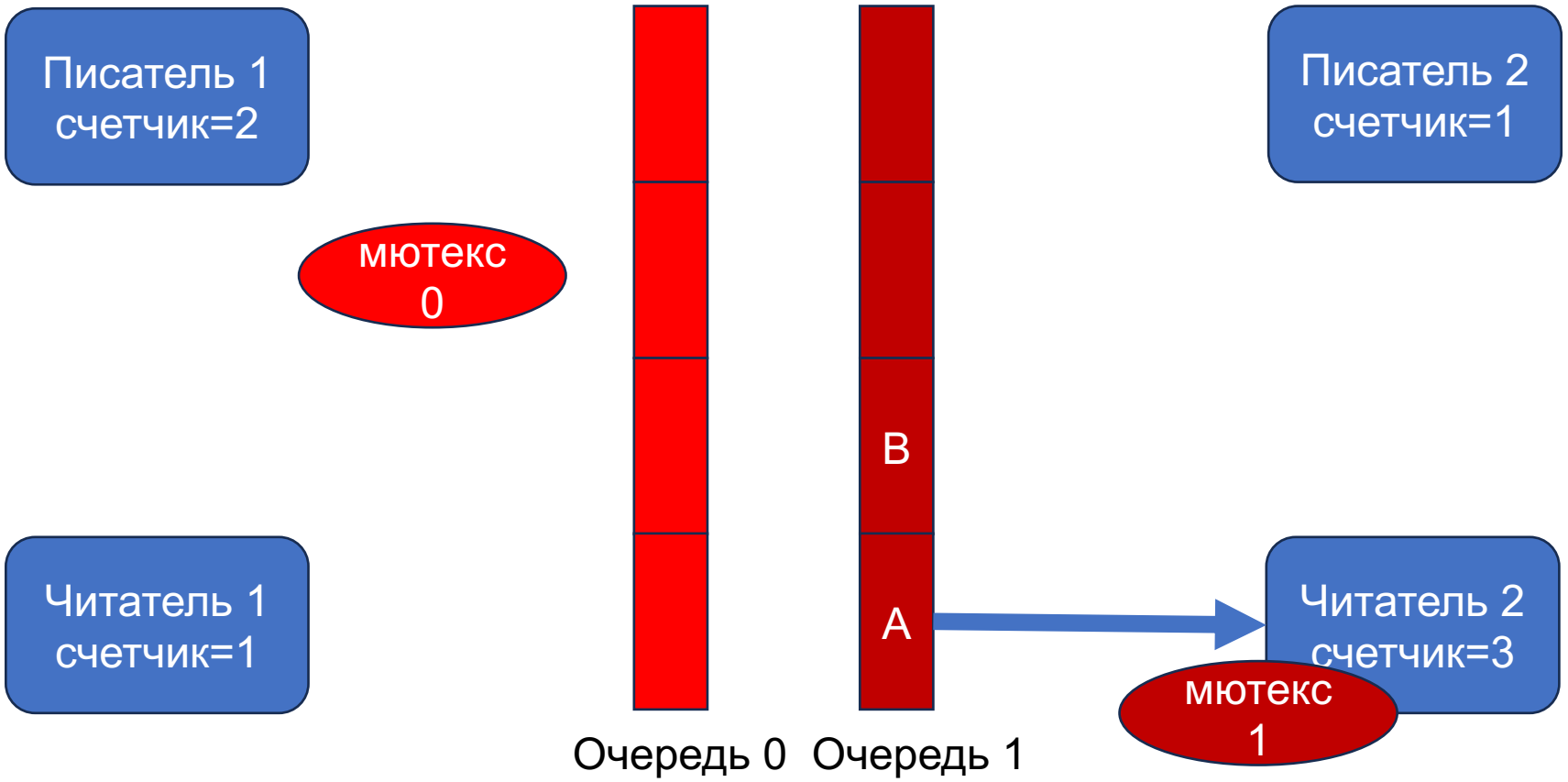
Revolving MPMC queue



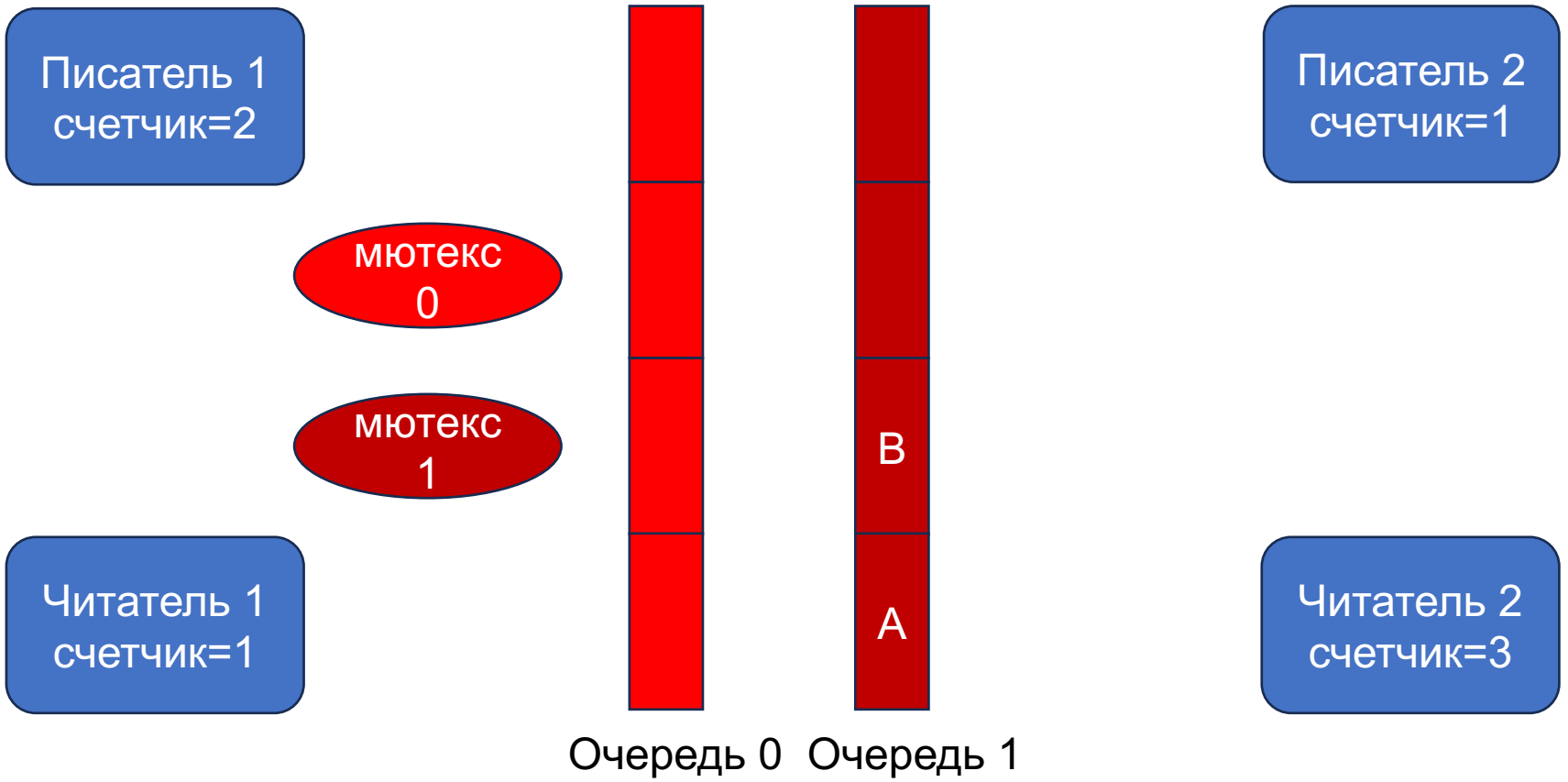
Revolving MPMC queue



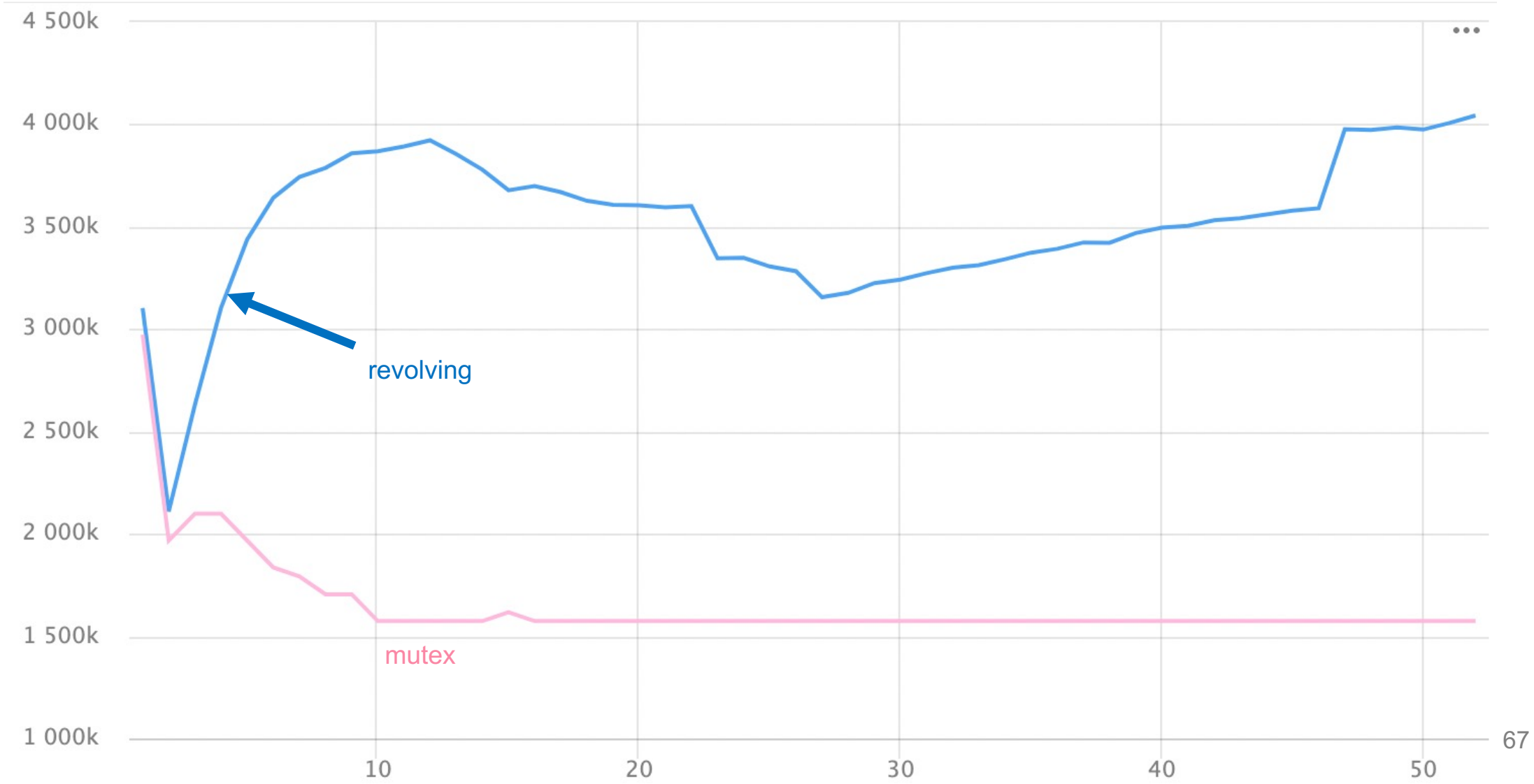
Revolving MPMC queue



Revolving MPMC queue

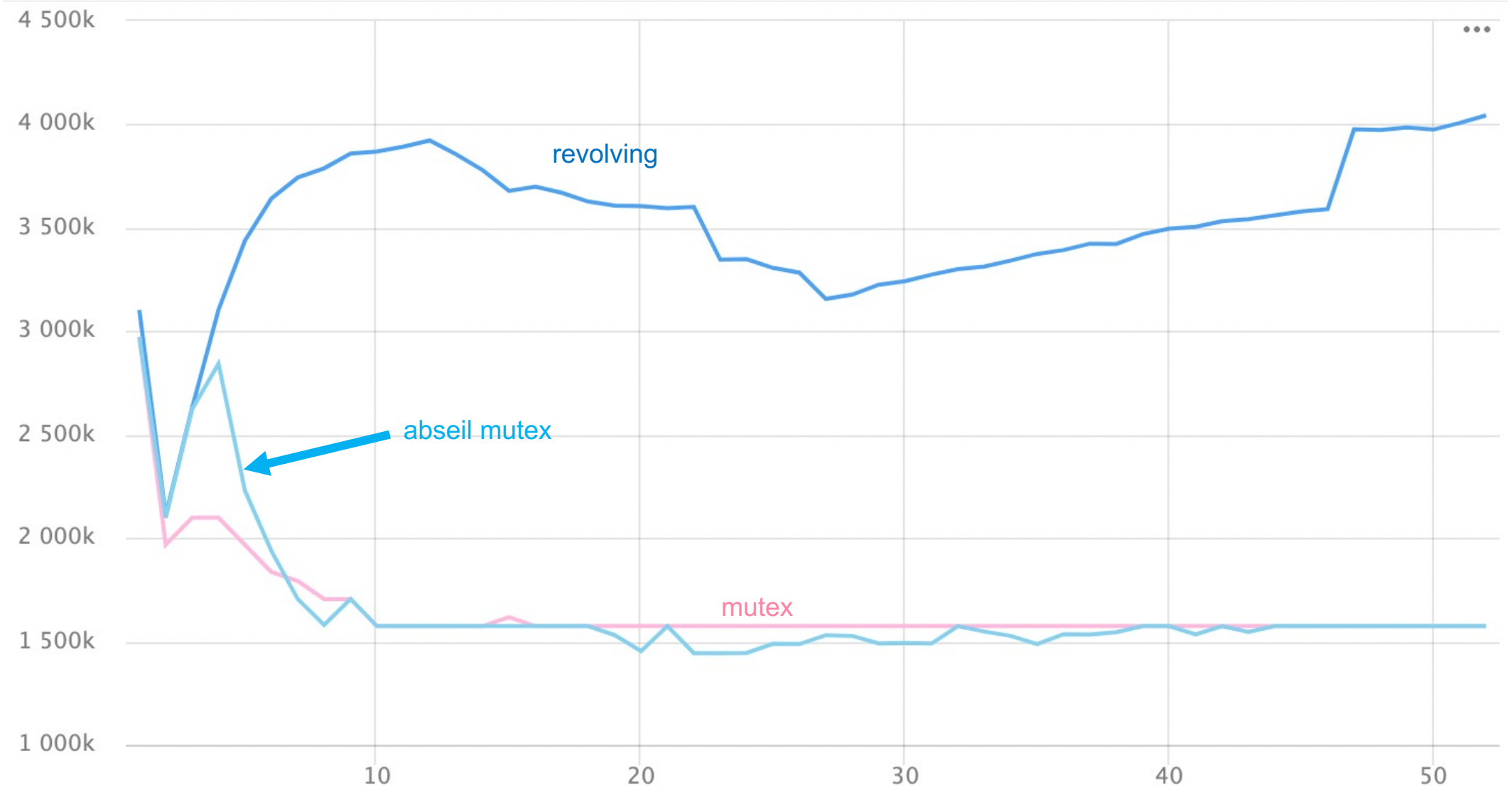


Revolving MPMC queue

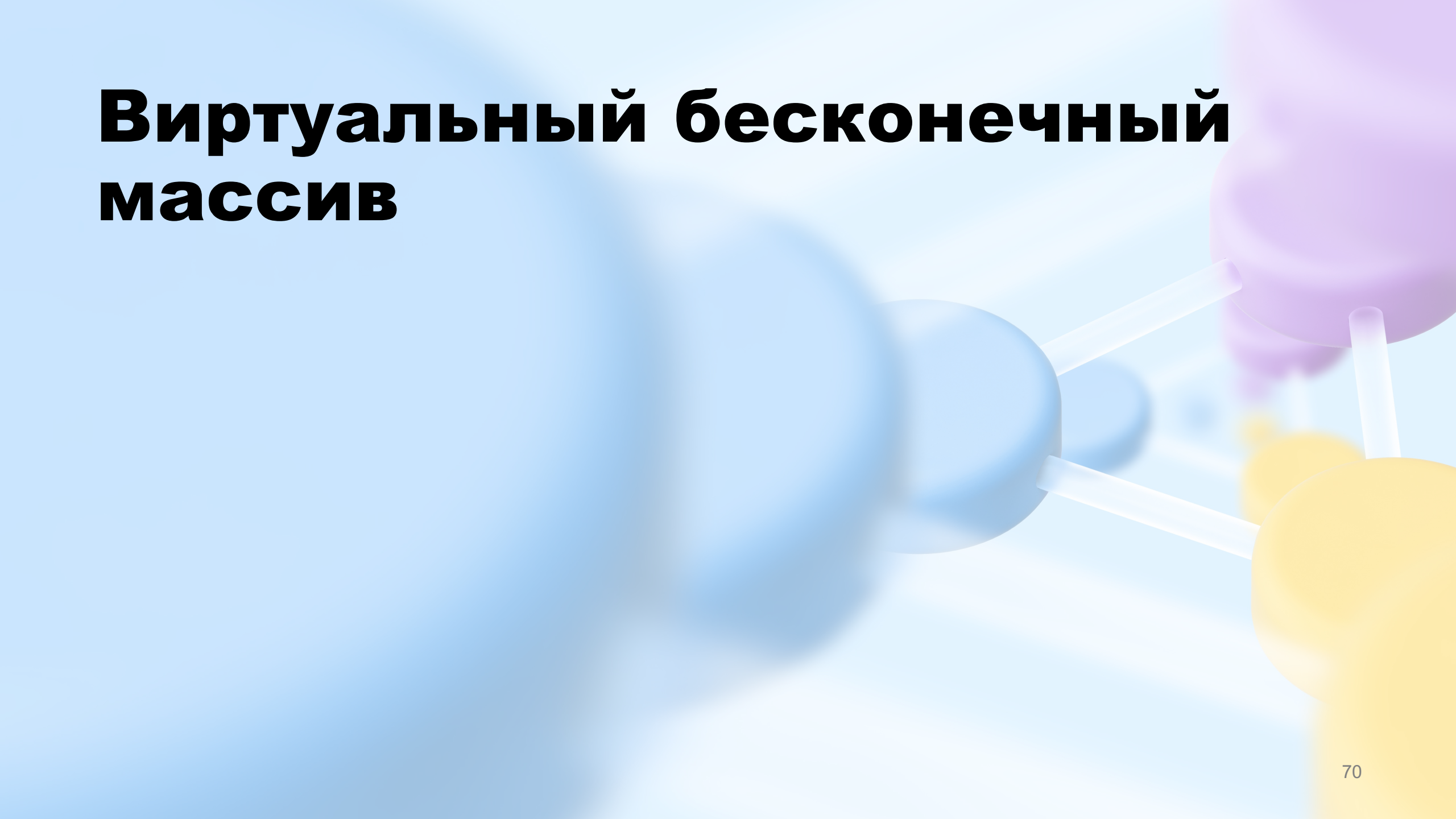


Быстрый mutex из abseil

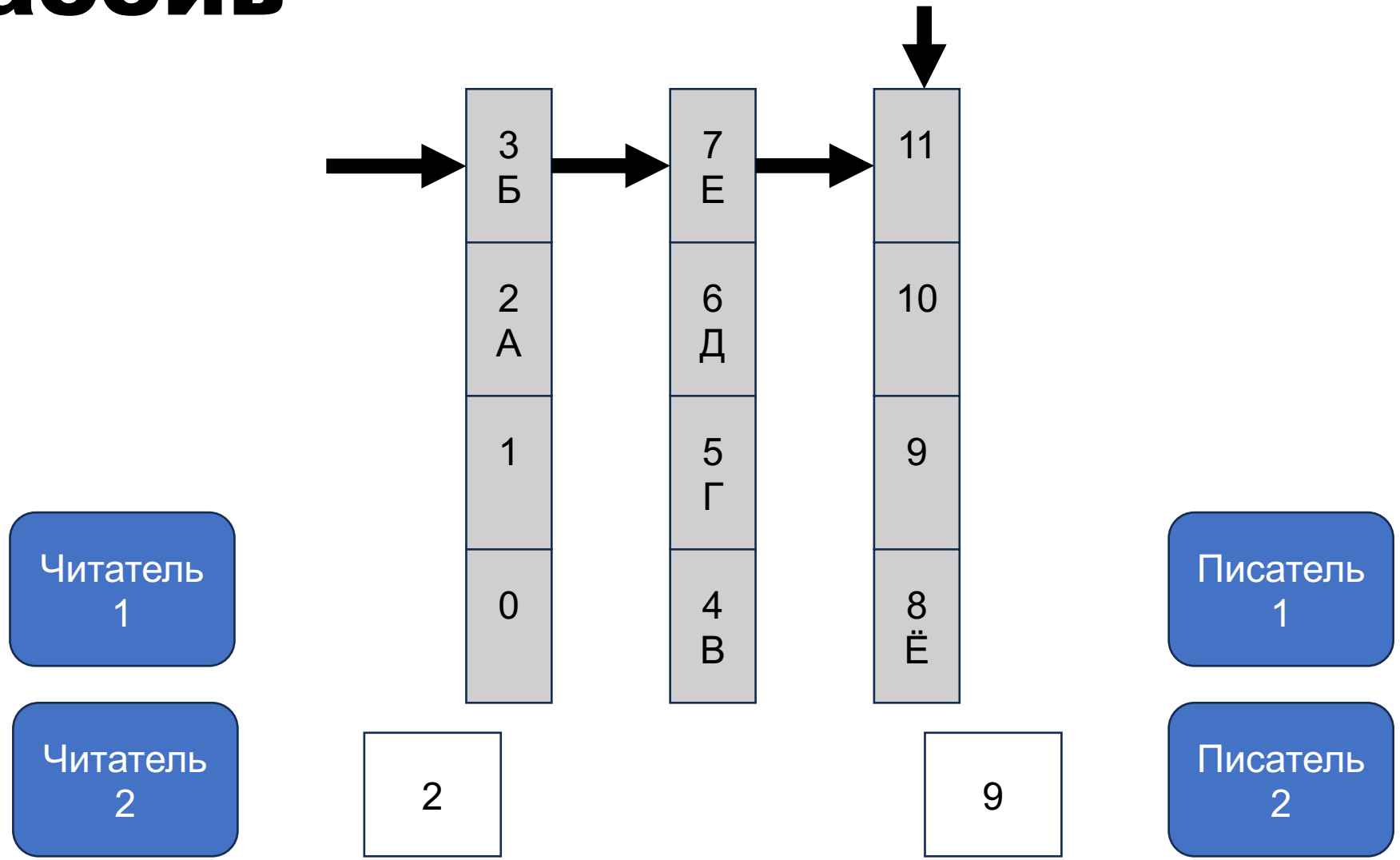
Быстрый mutex из abseil



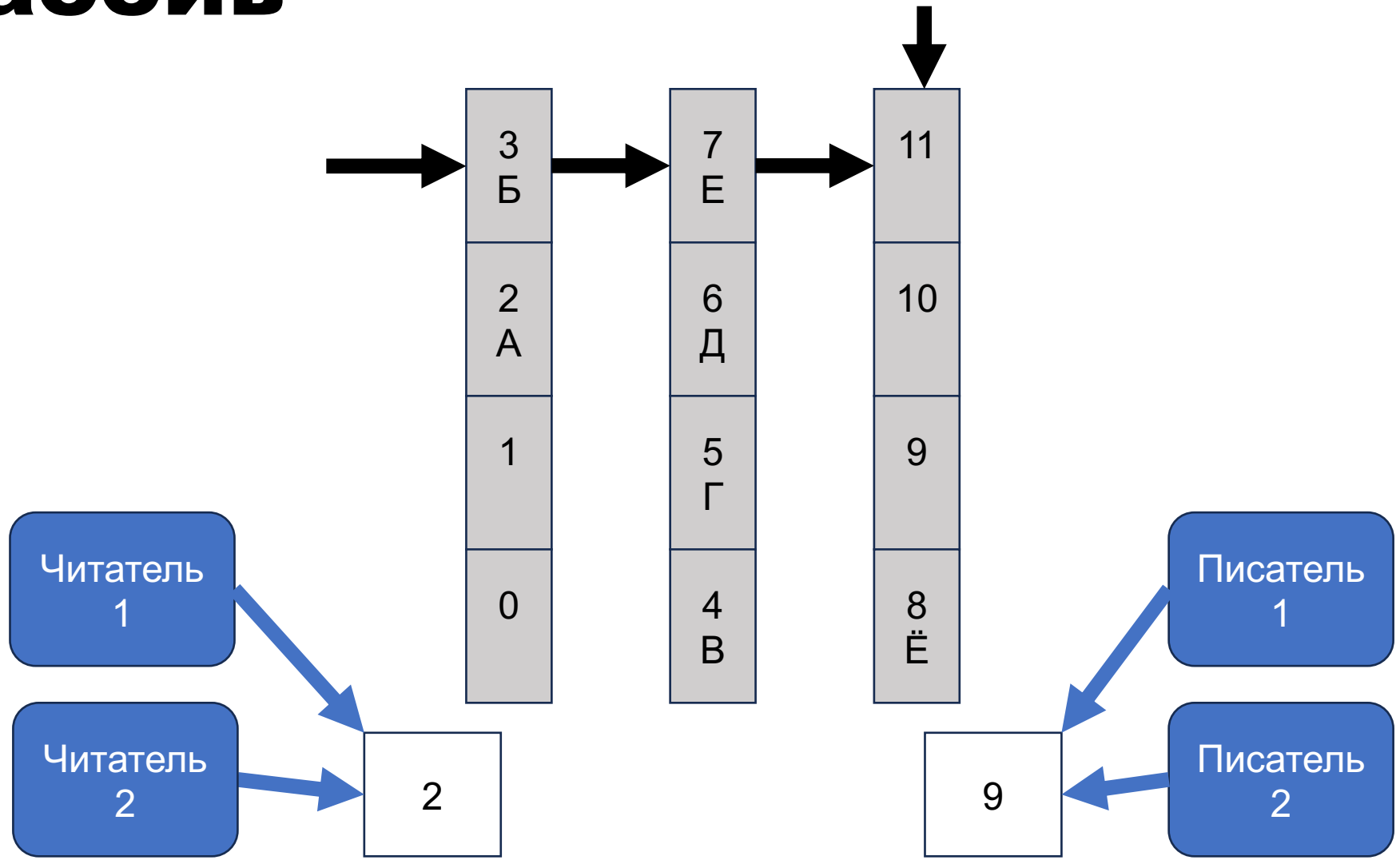
Виртуальный бесконечный массив



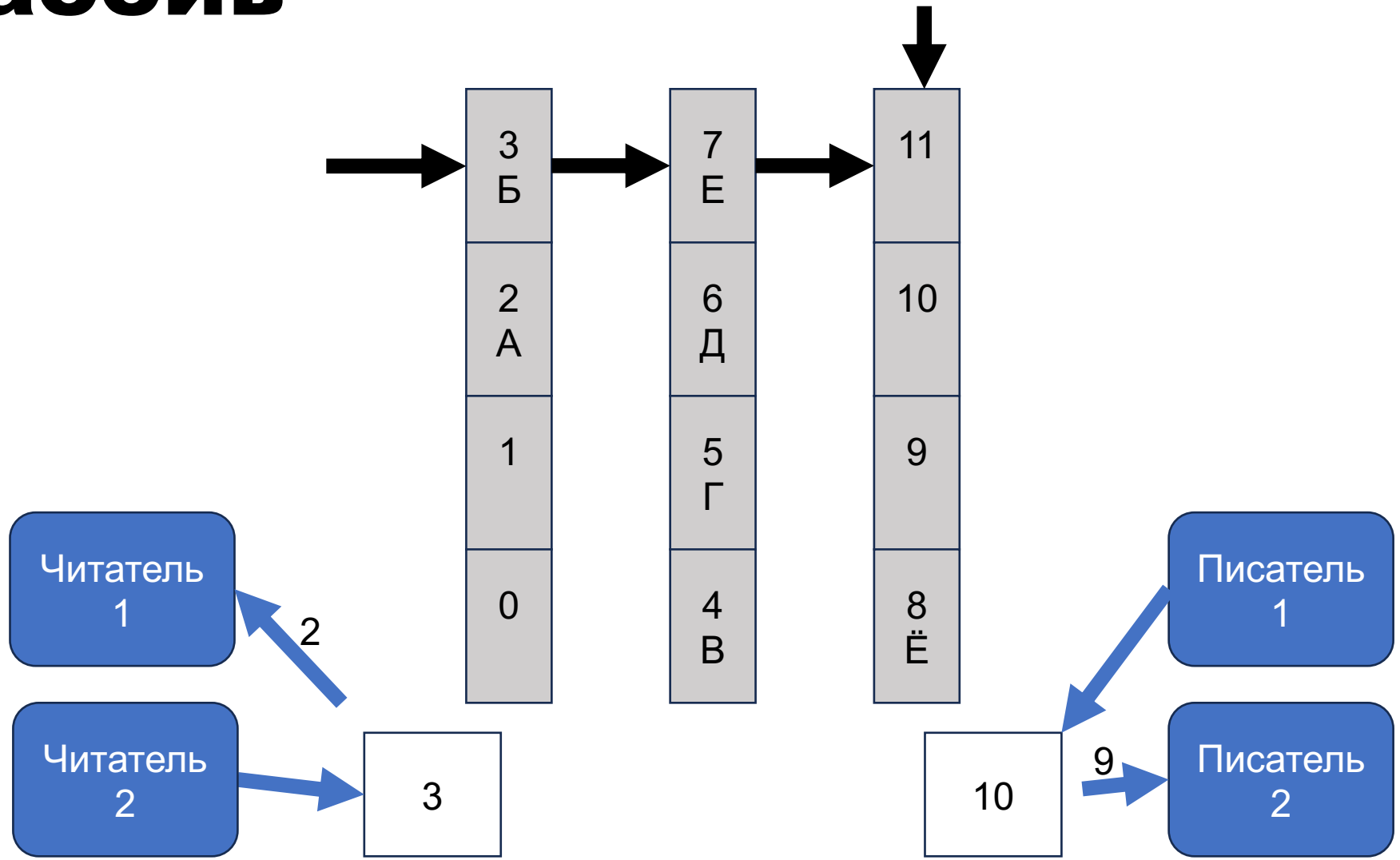
Виртуальный бесконечный массив



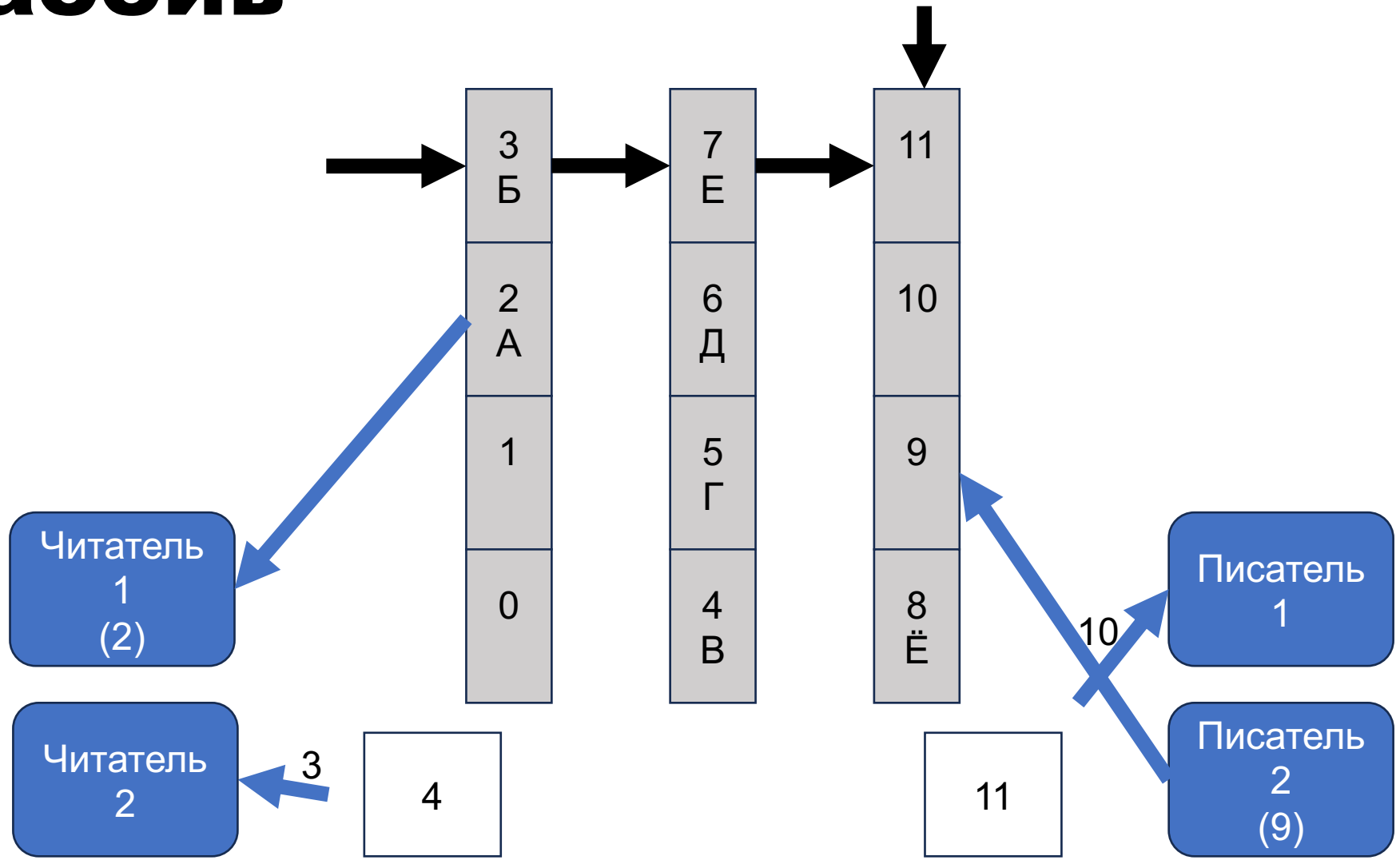
Виртуальный бесконечный массив



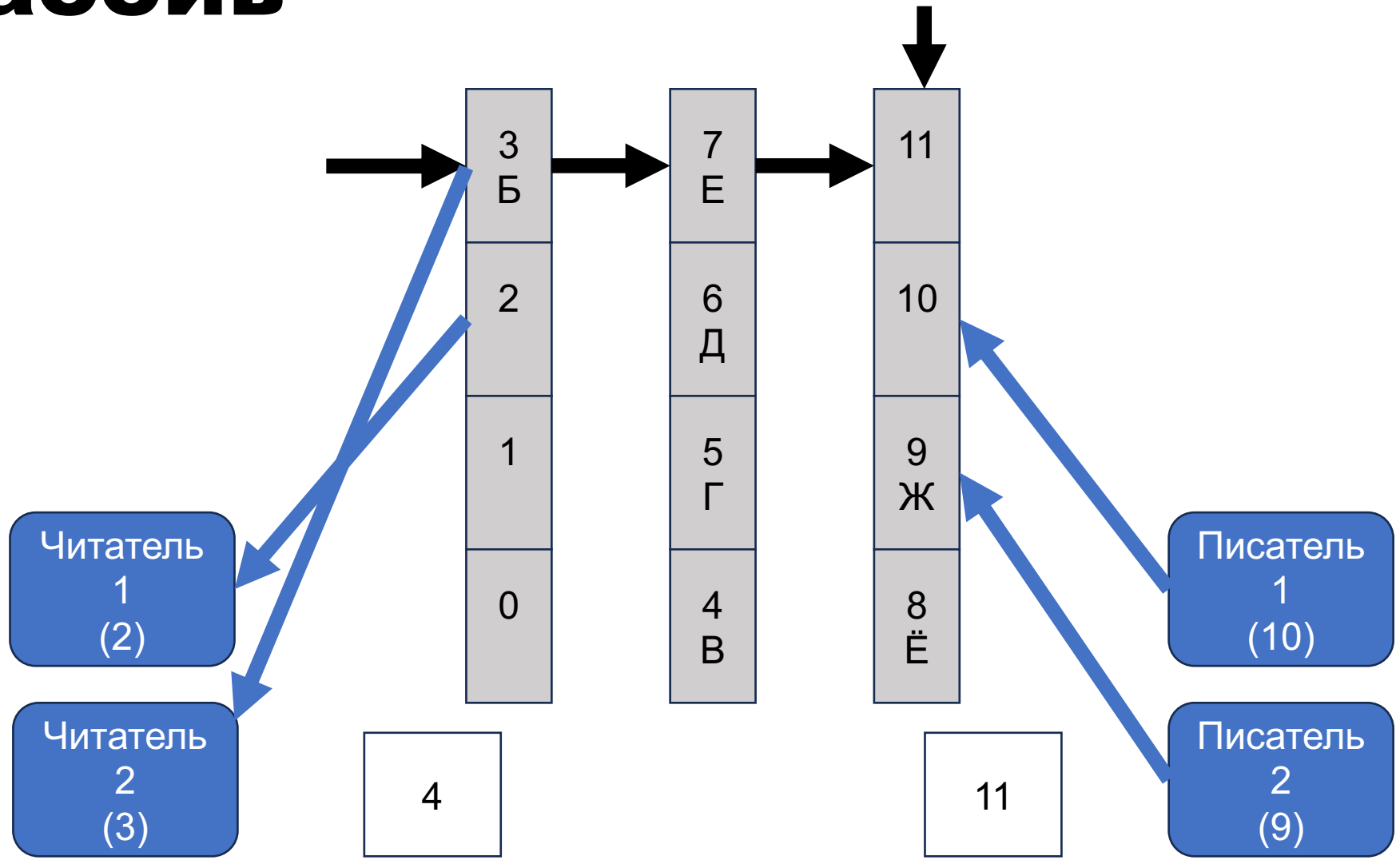
Виртуальный бесконечный массив



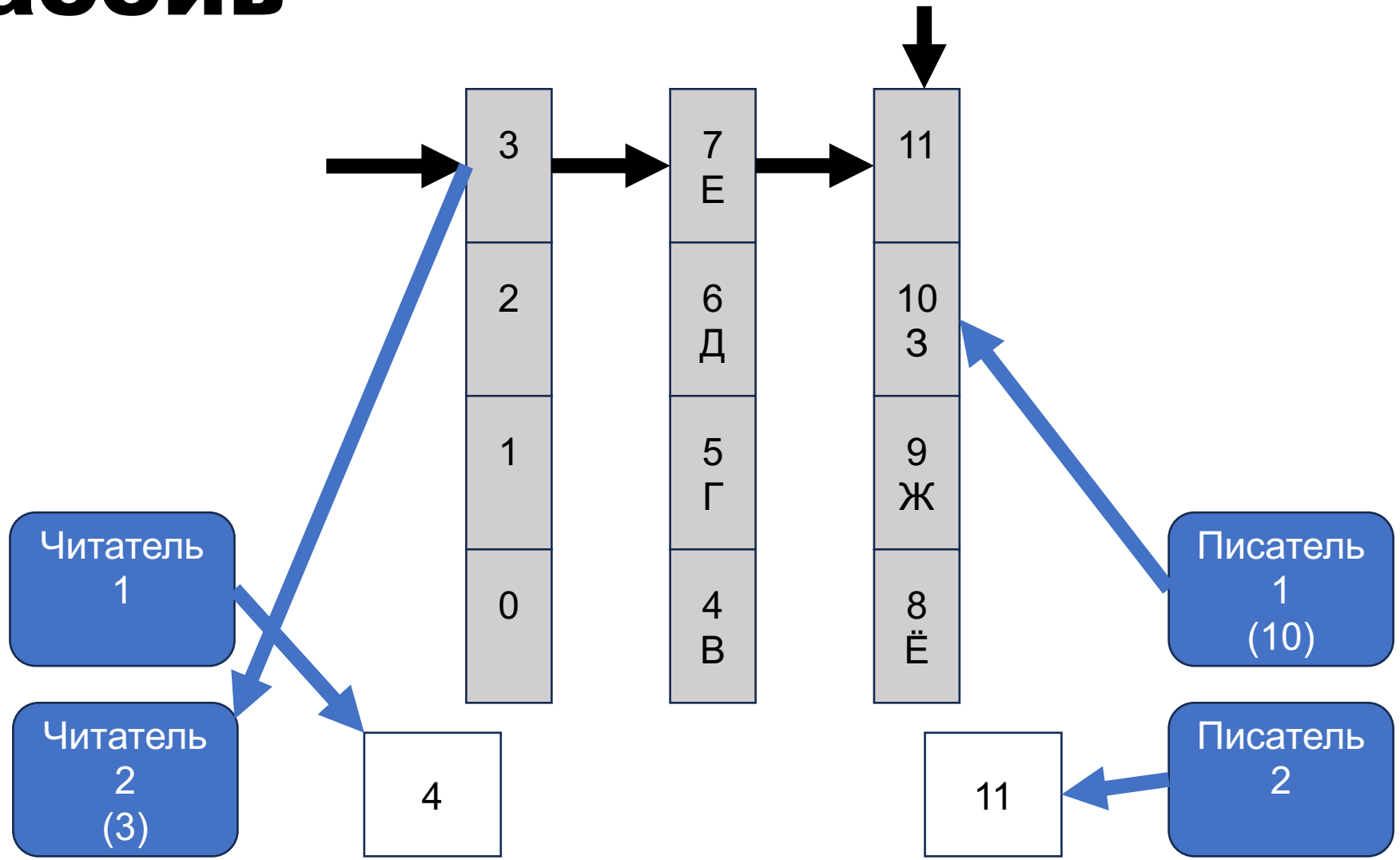
Виртуальный бесконечный массив



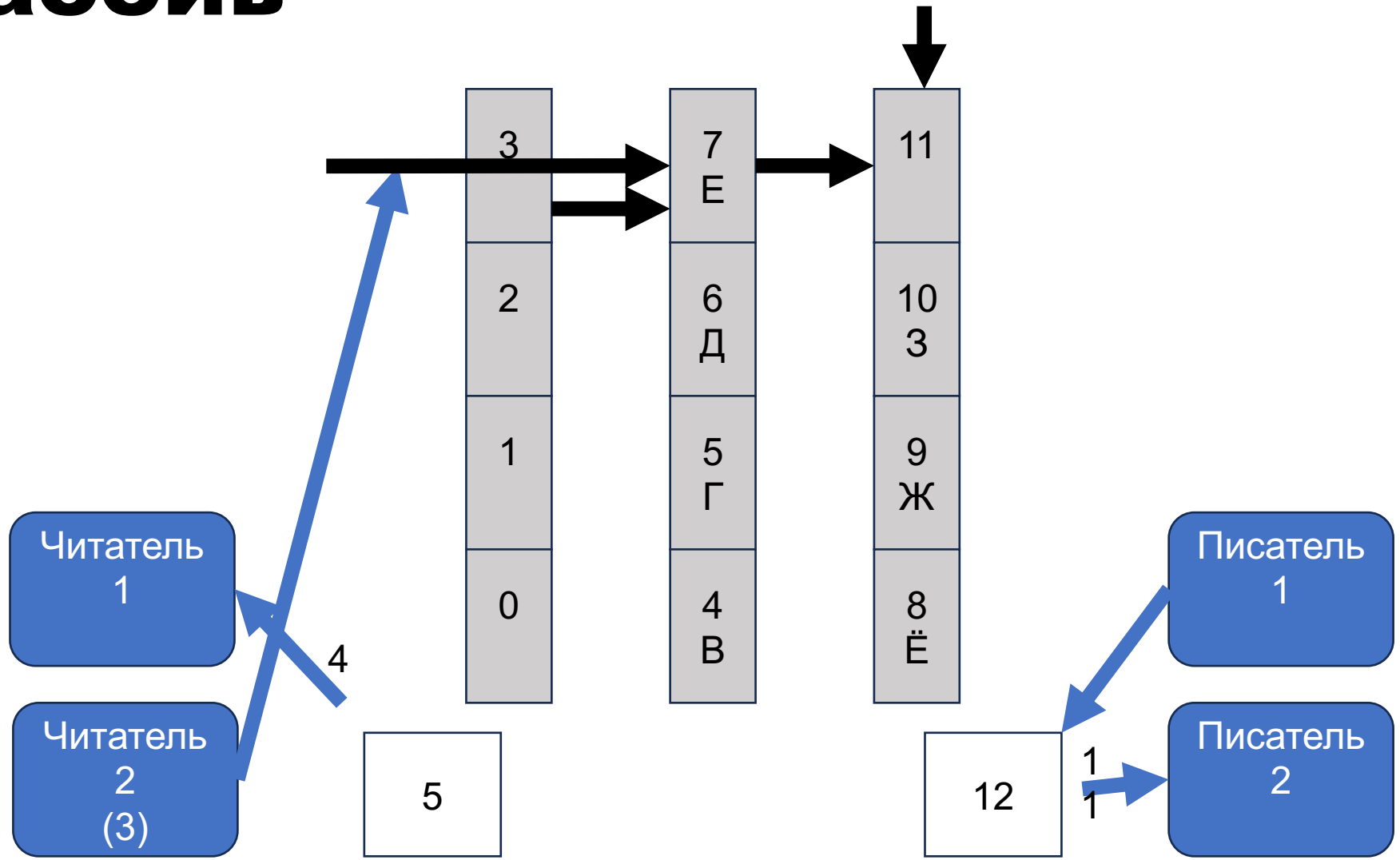
Виртуальный бесконечный массив



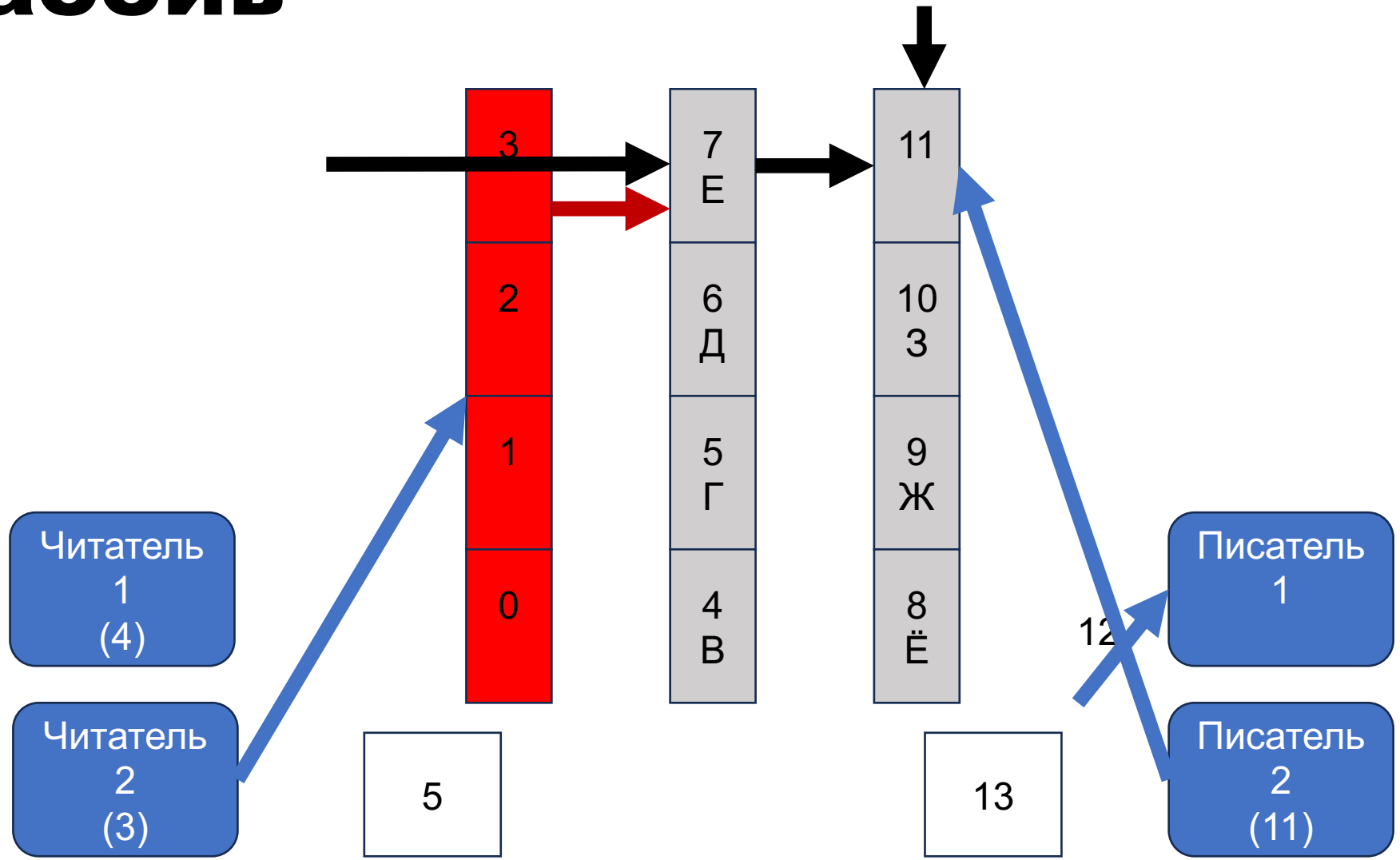
Виртуальный бесконечный массив



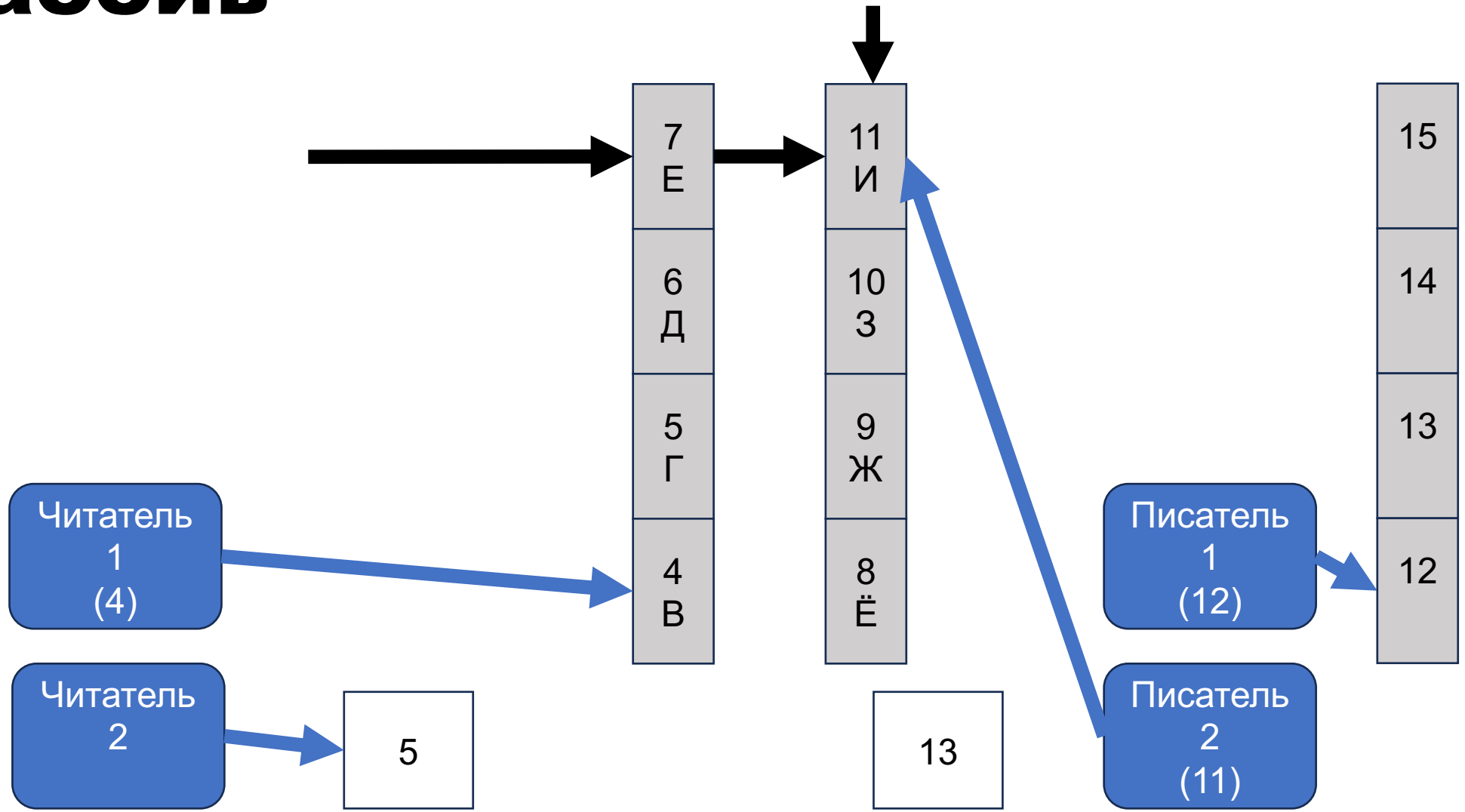
Виртуальный бесконечный массив



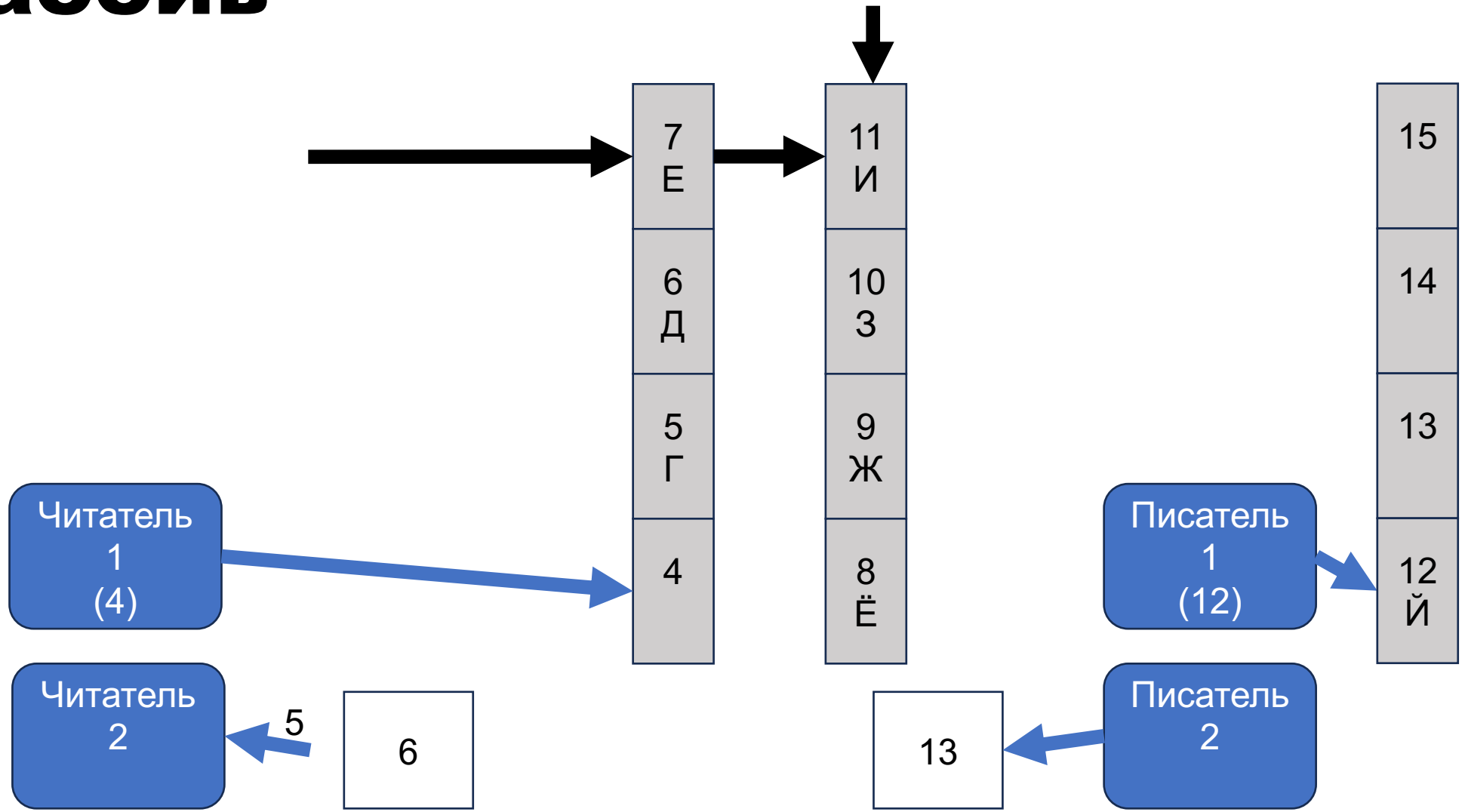
Виртуальный бесконечный массив



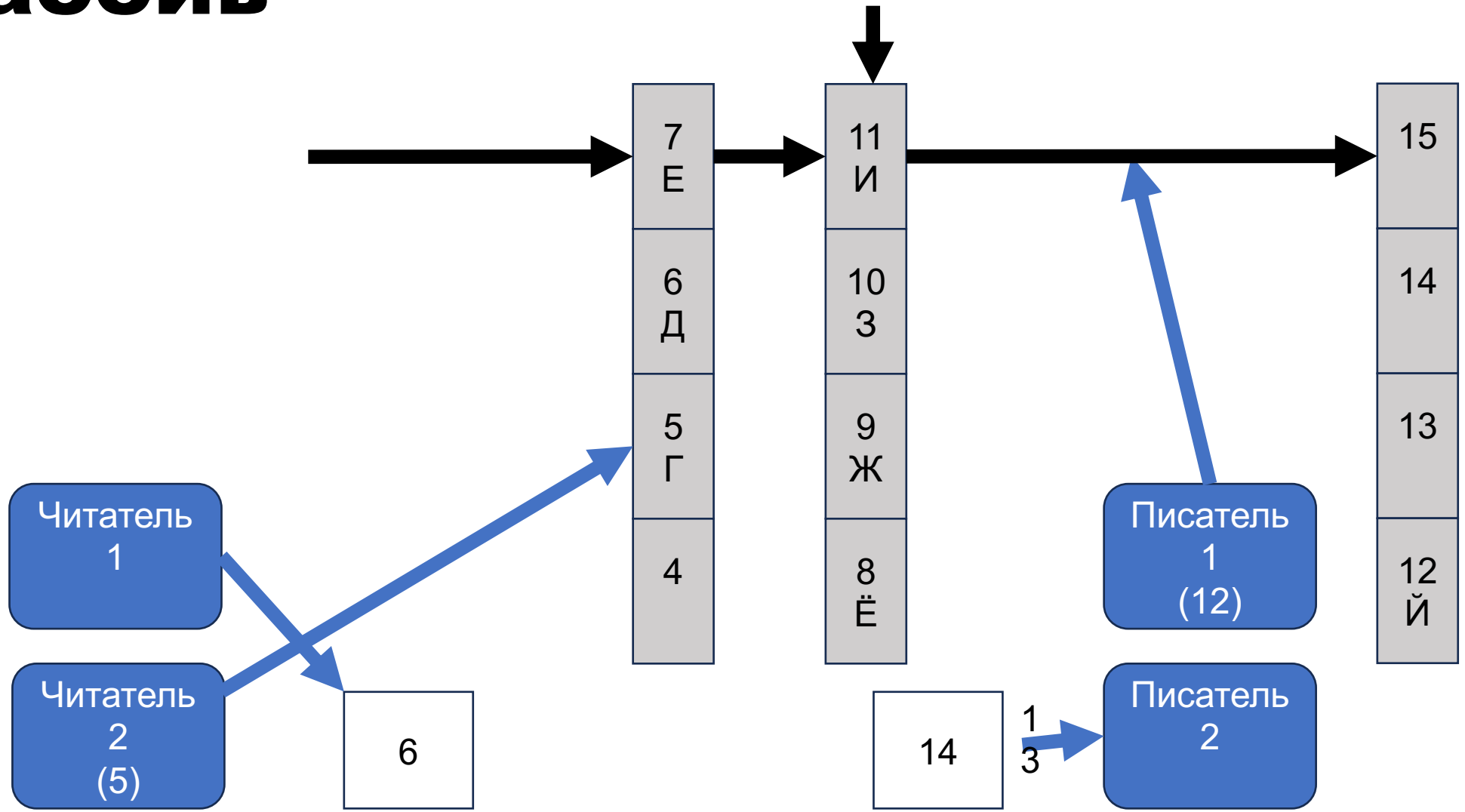
Виртуальный бесконечный массив



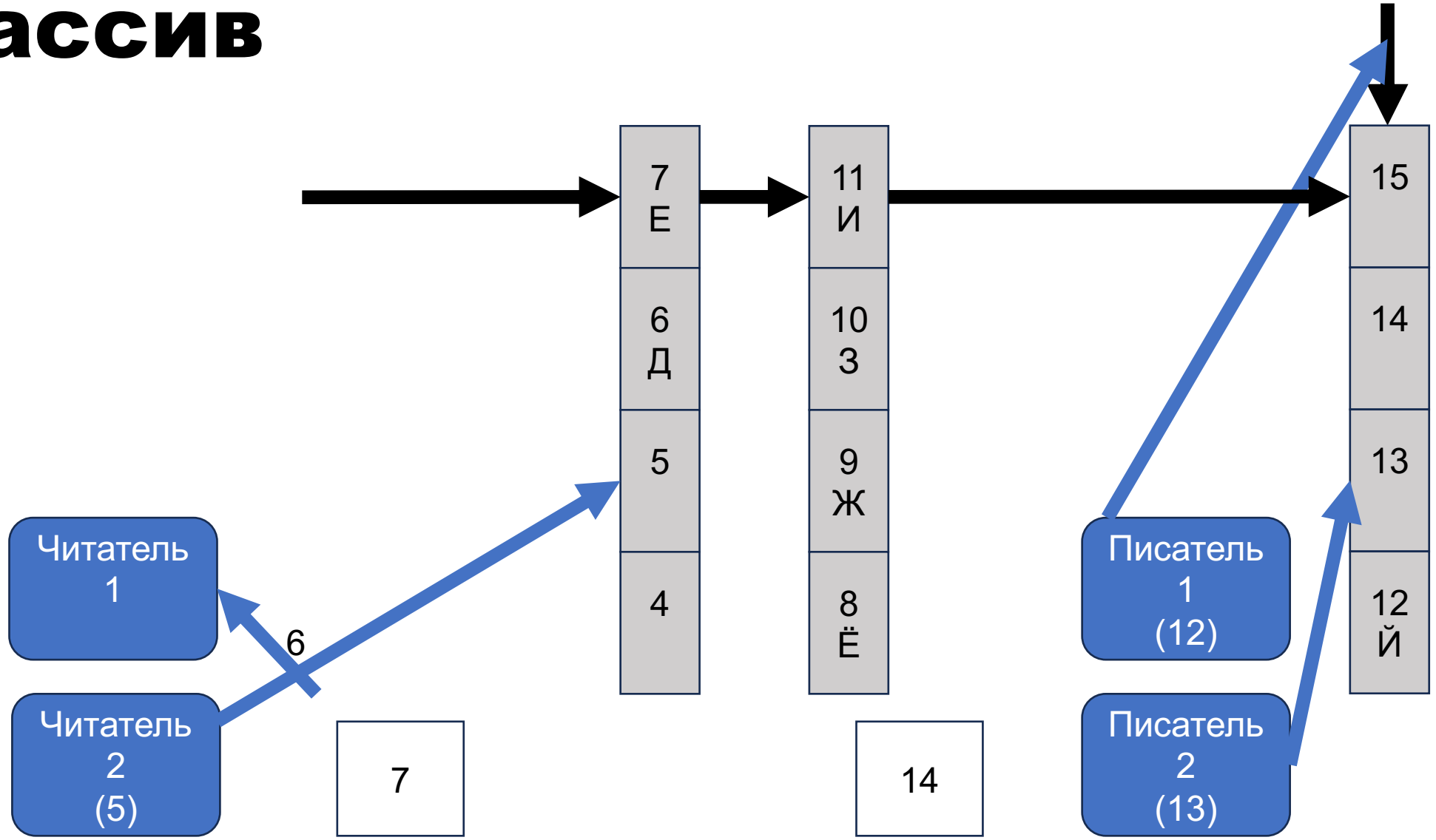
Виртуальный бесконечный массив



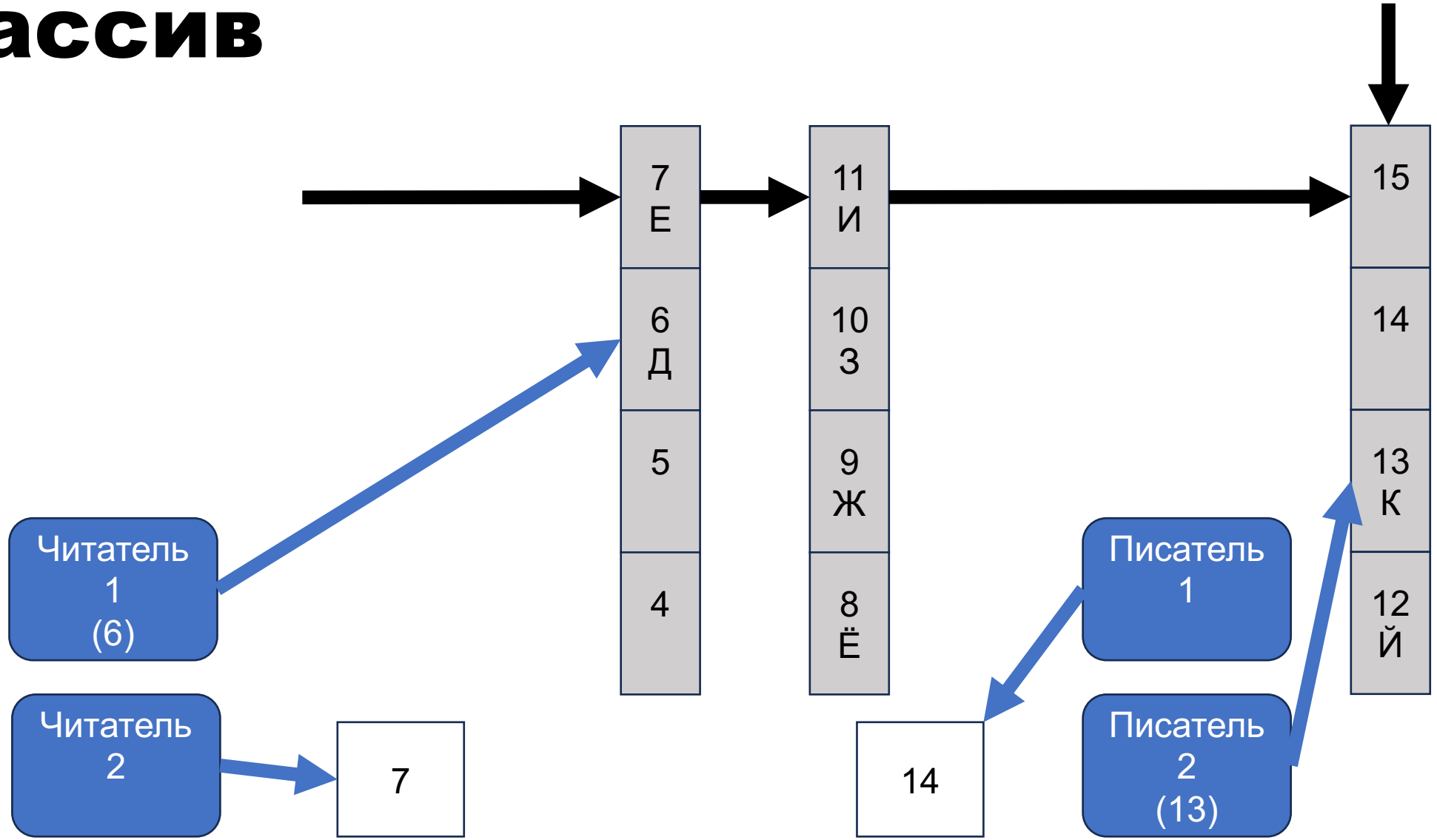
Виртуальный бесконечный массив



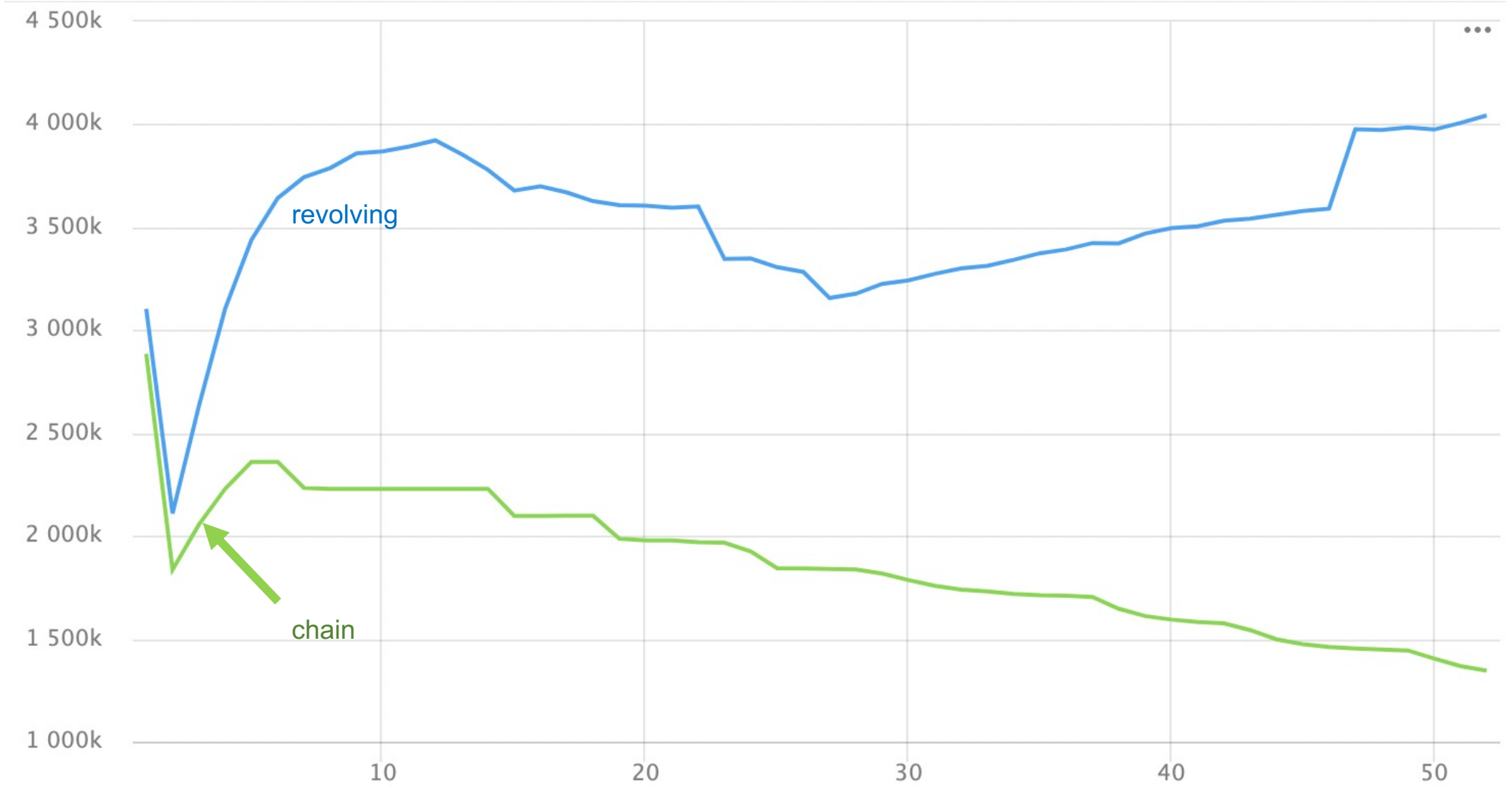
Виртуальный бесконечный массив



Виртуальный бесконечный массив

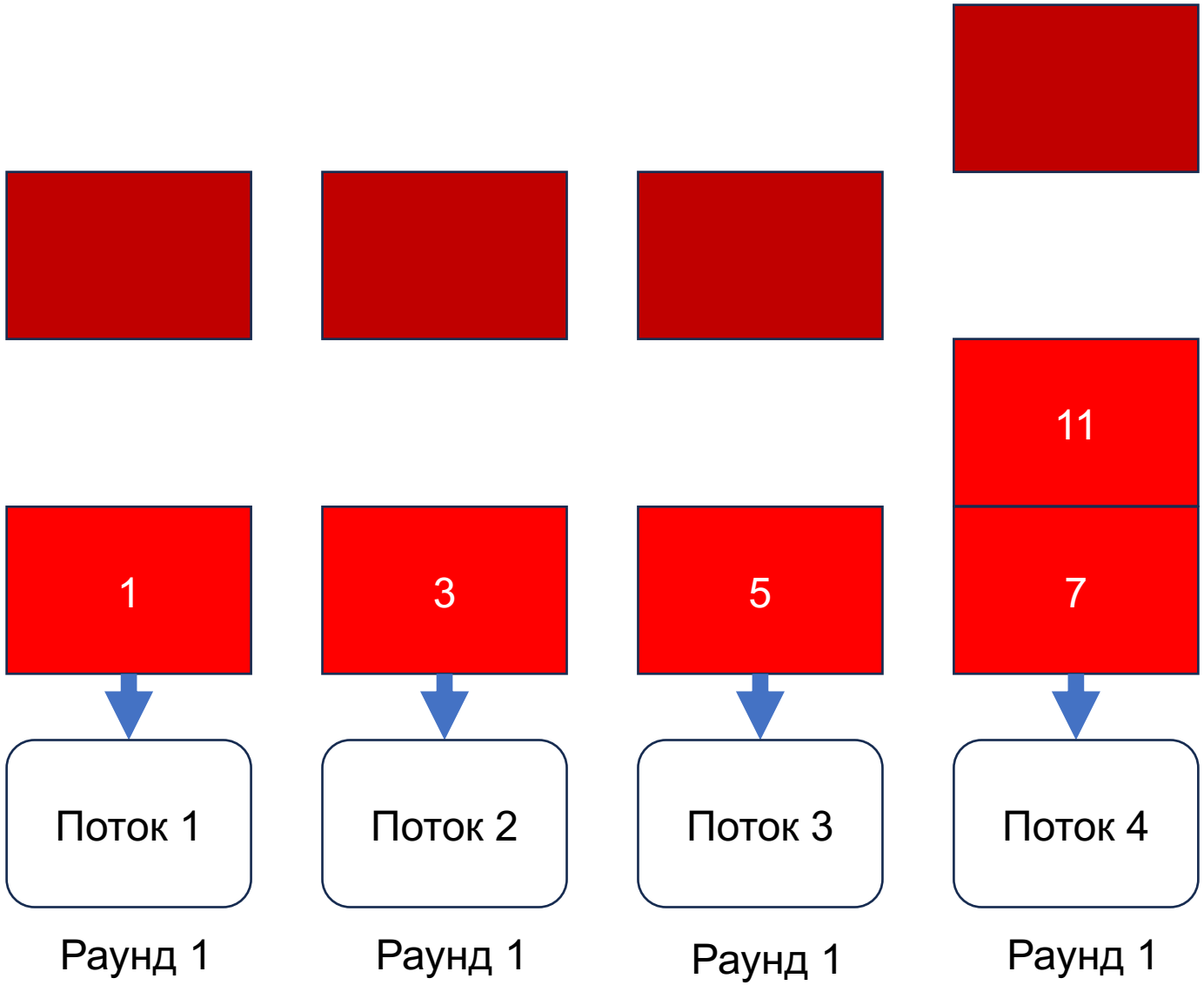


Виртуальный бесконечный массив

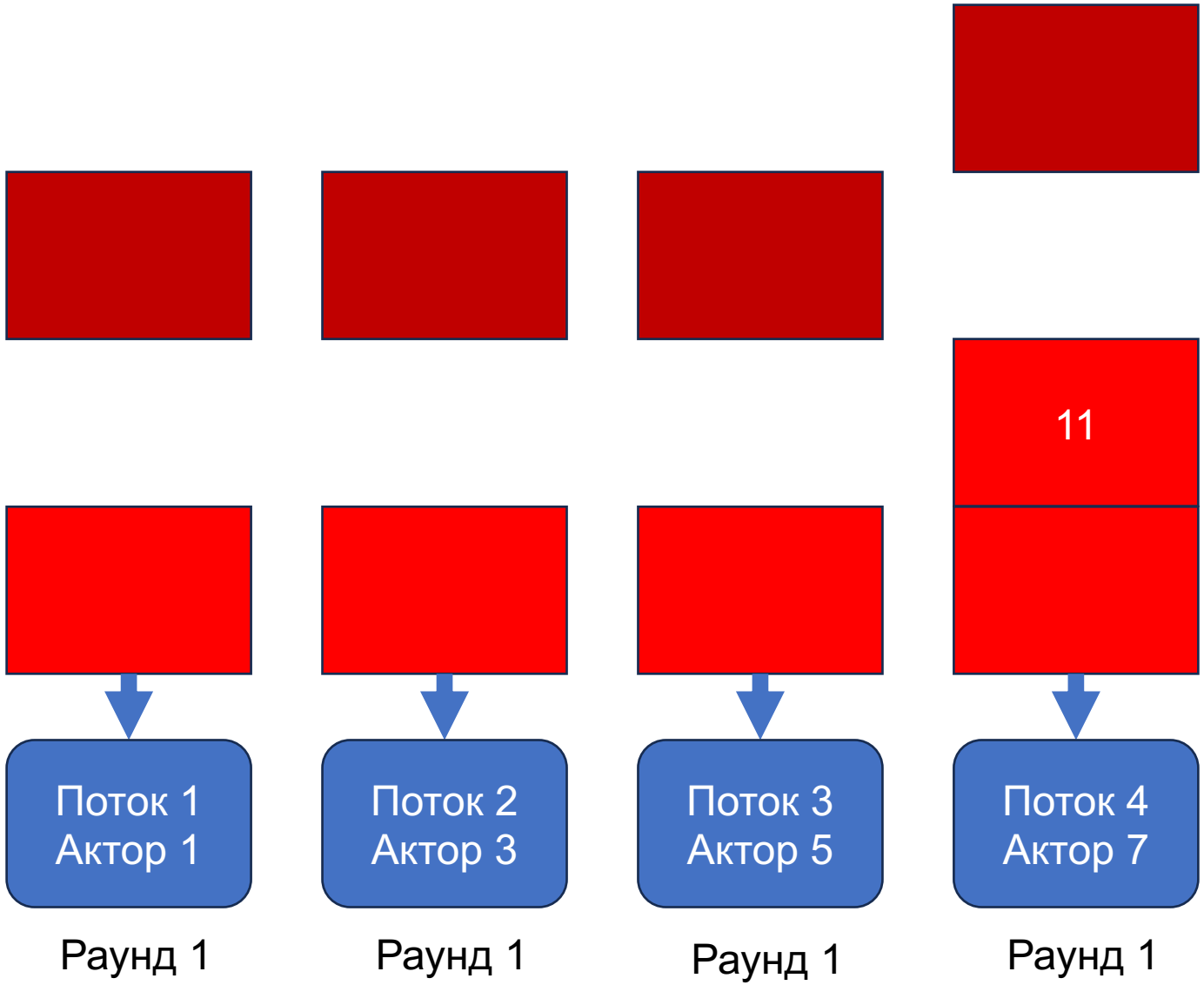


Пробуем применить ТРИЗ

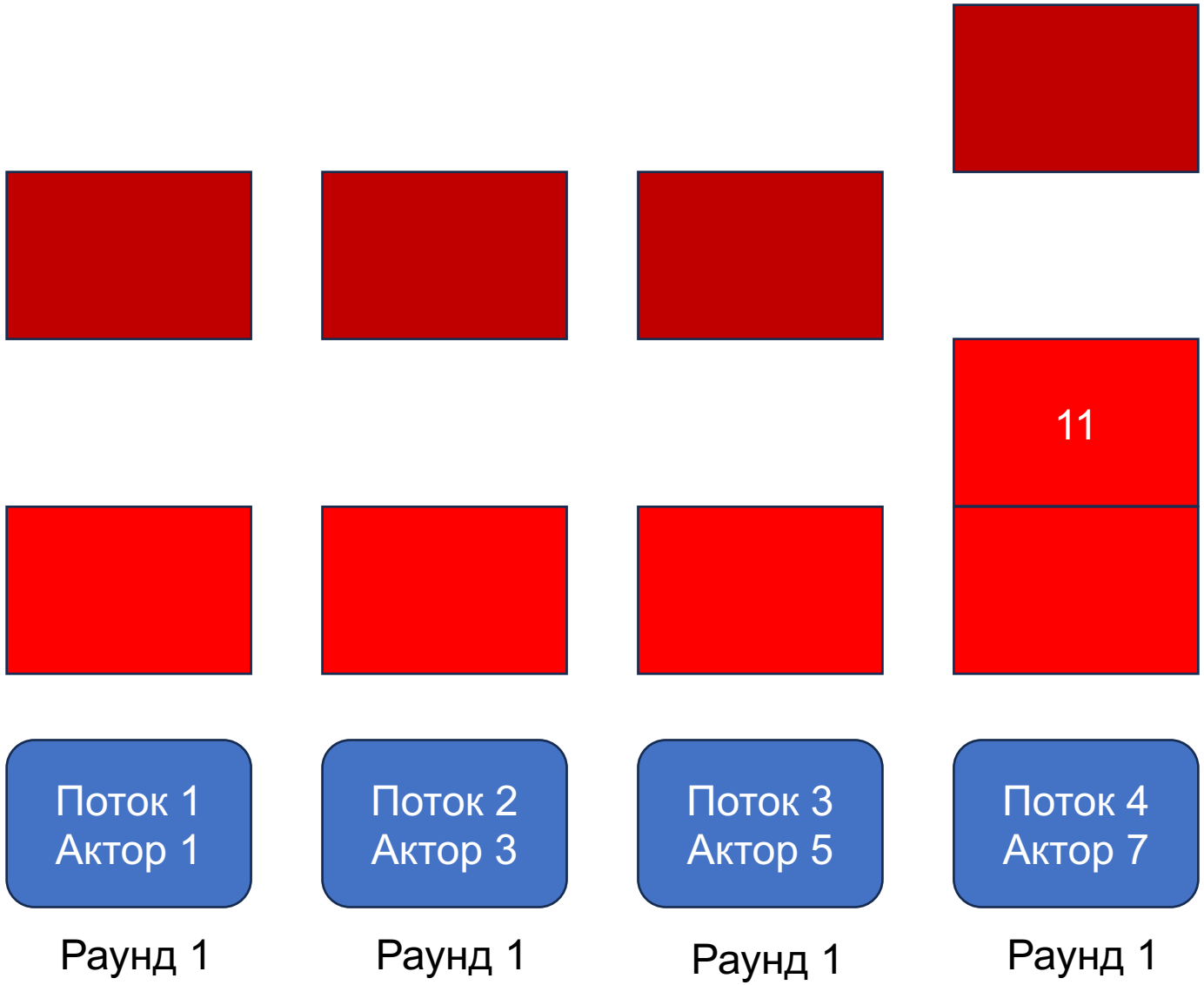
Отказ от единой очереди



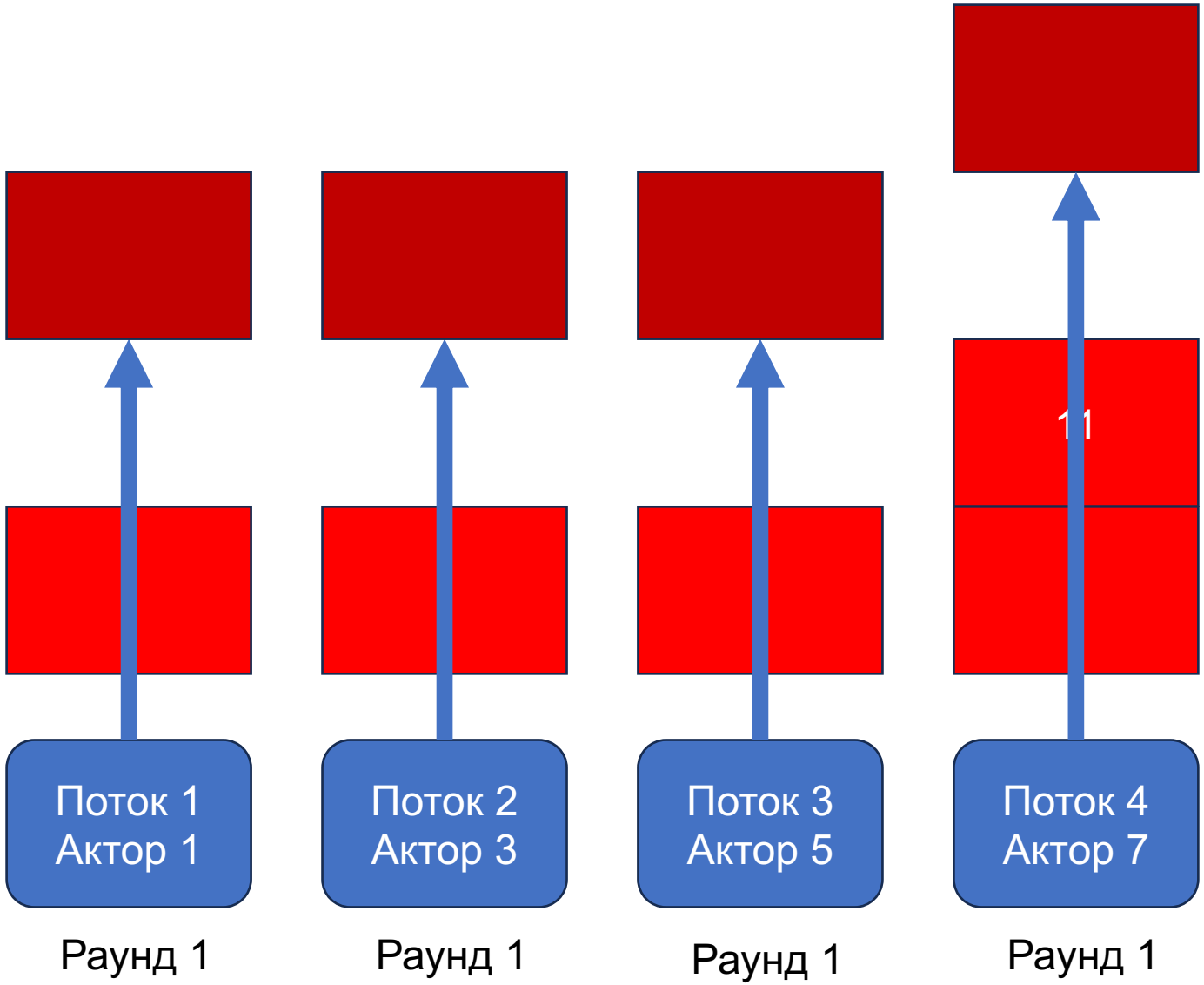
Отказ от единой очереди



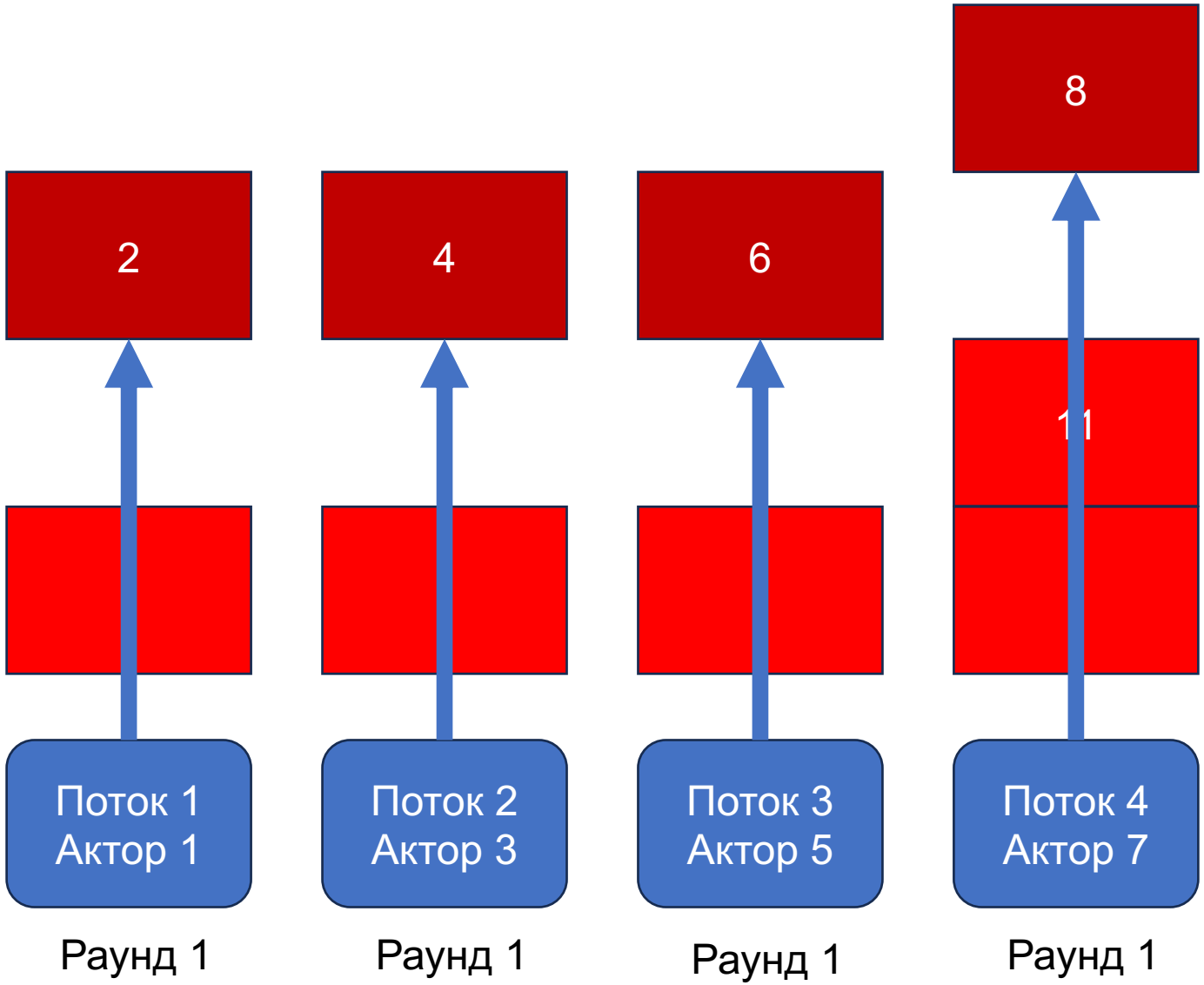
Отказ от единой очереди



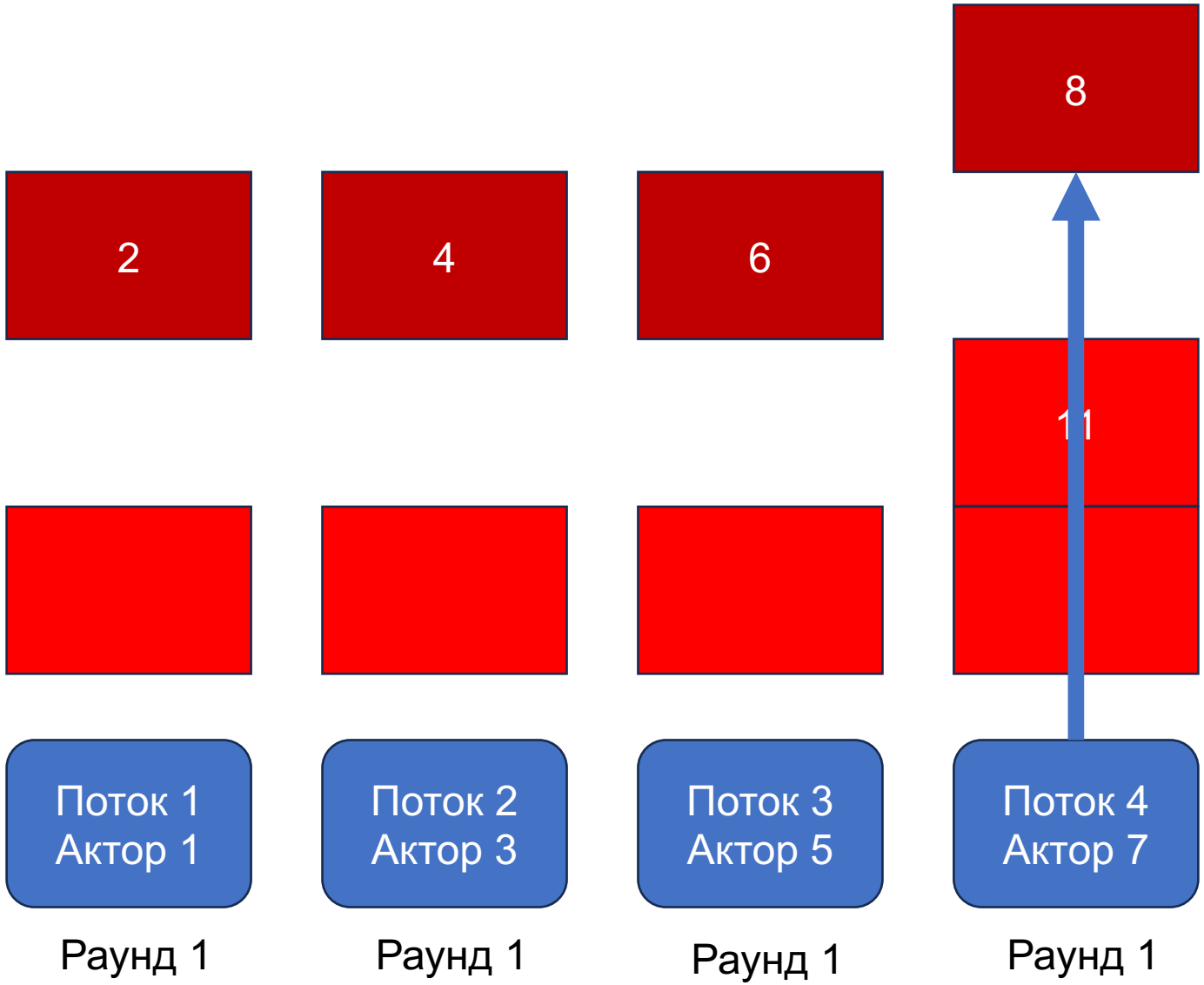
Отказ от единой очереди



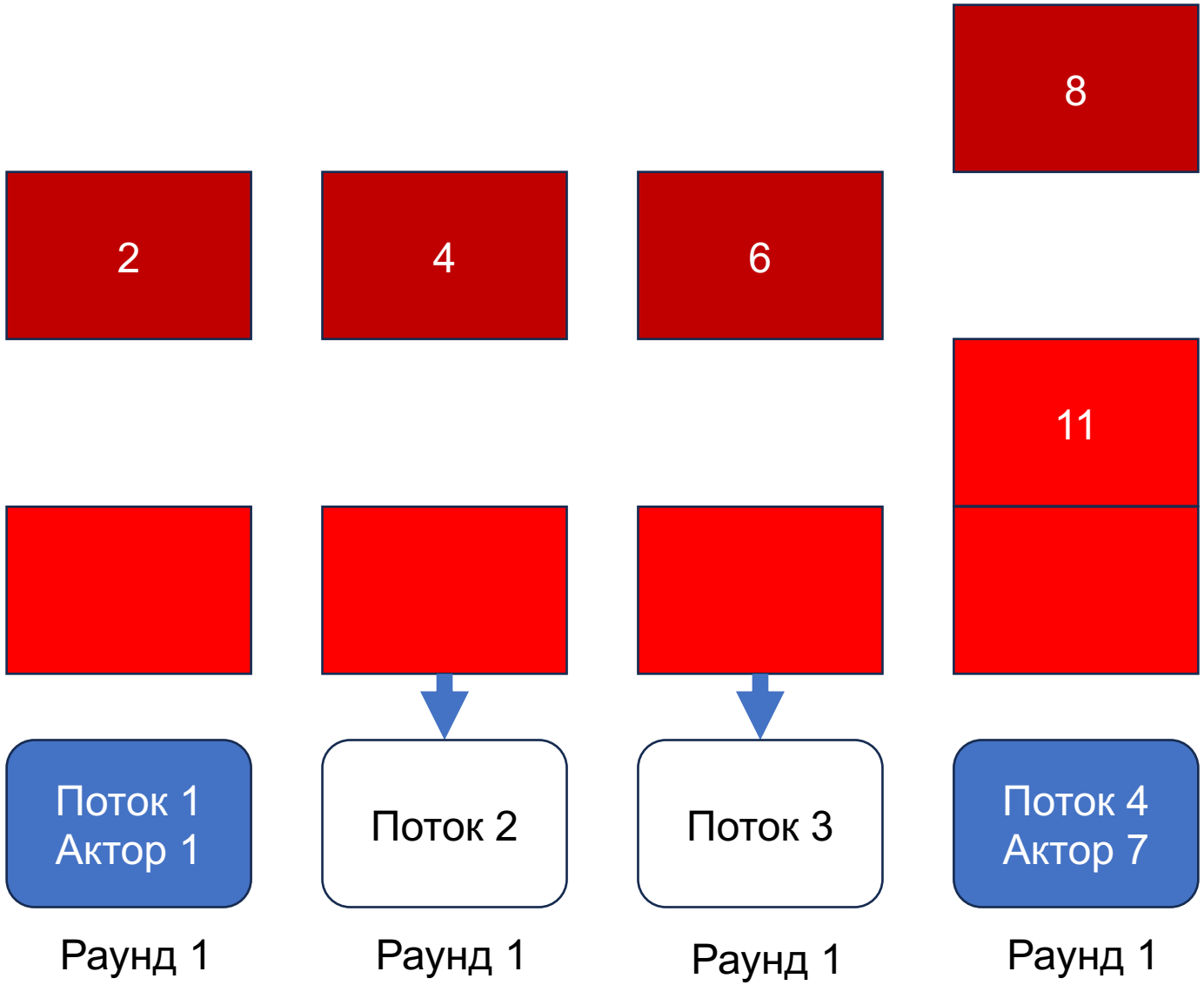
Отказ от единой очереди



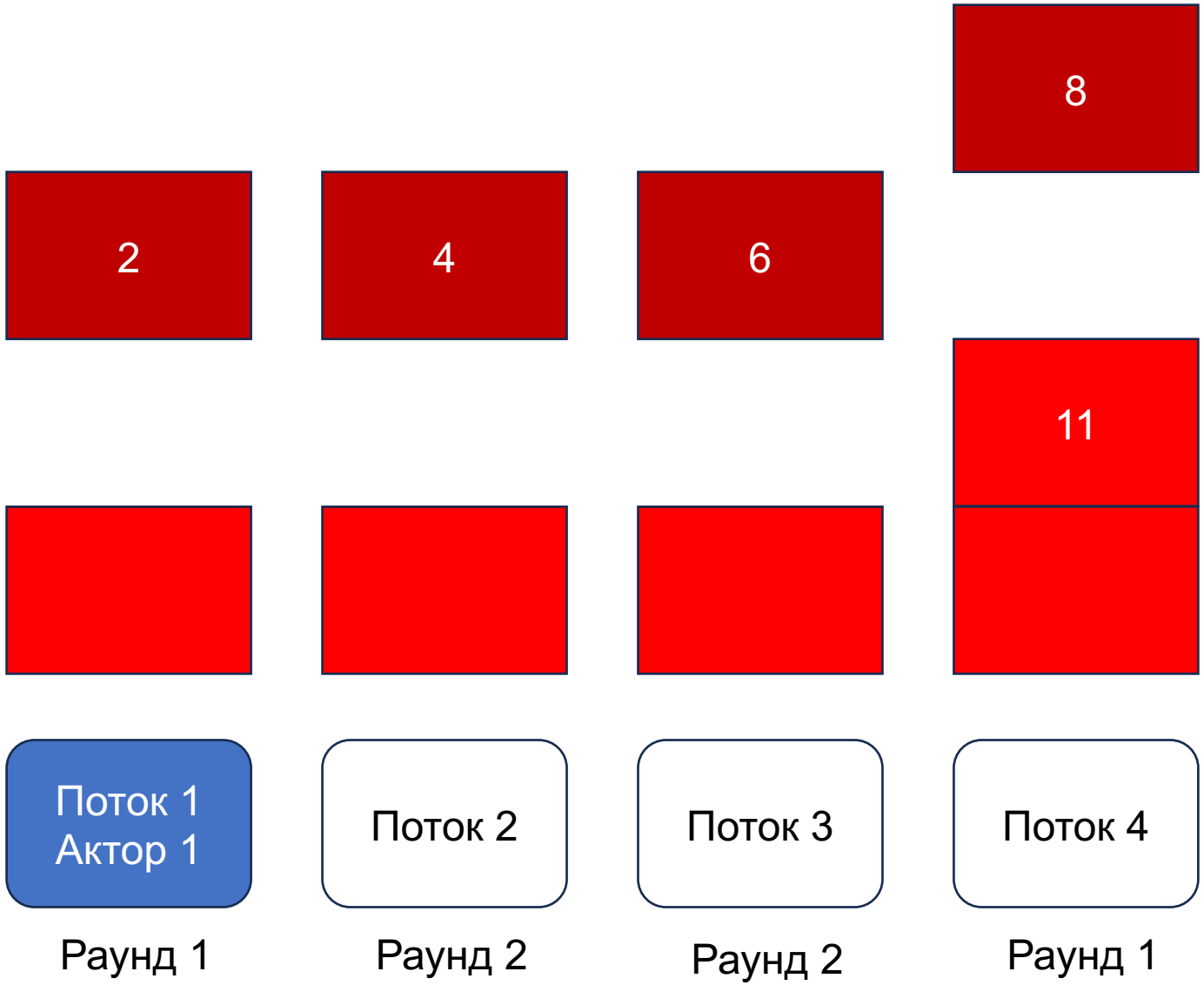
Отказ от единой очереди



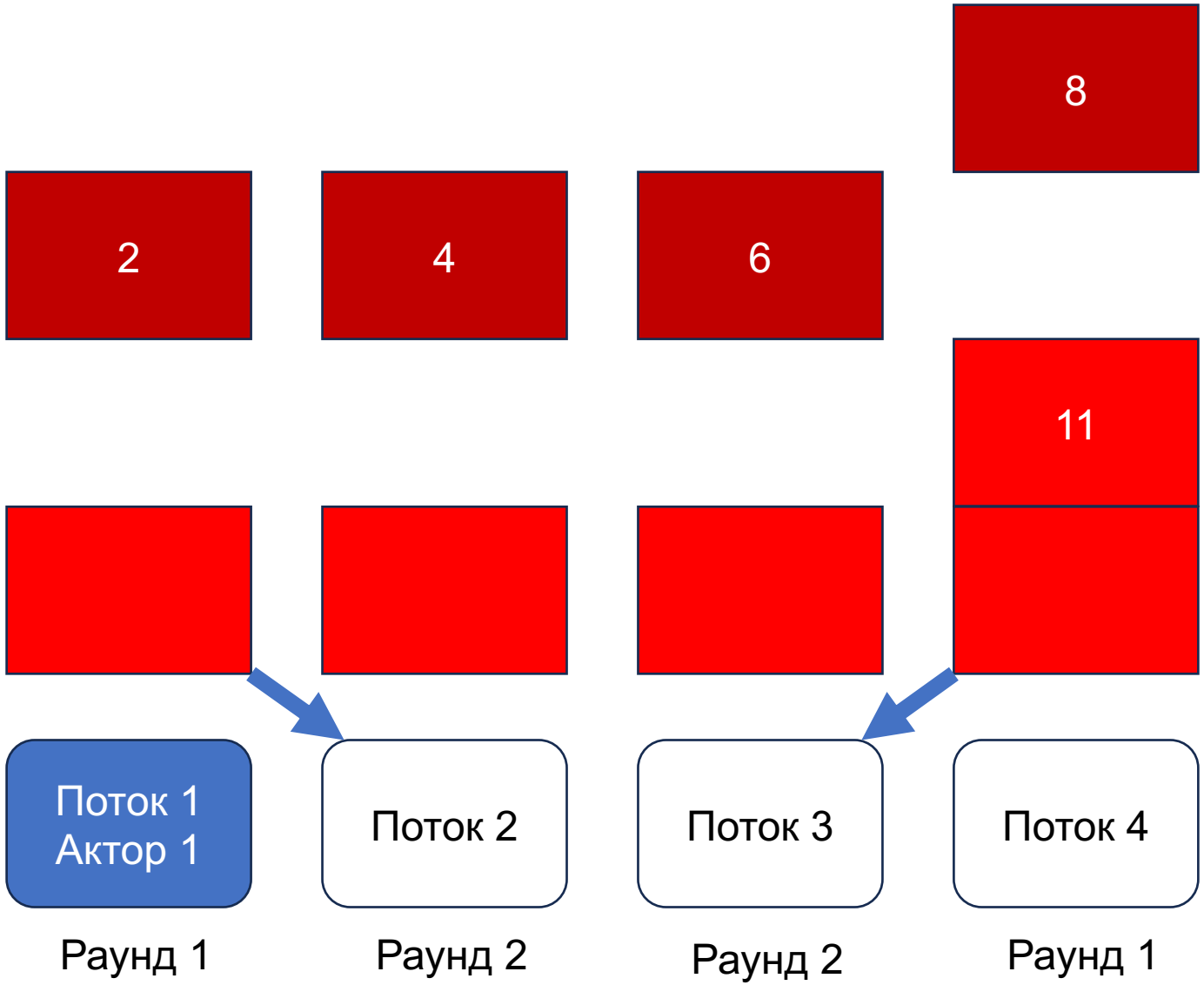
Отказ от единой очереди



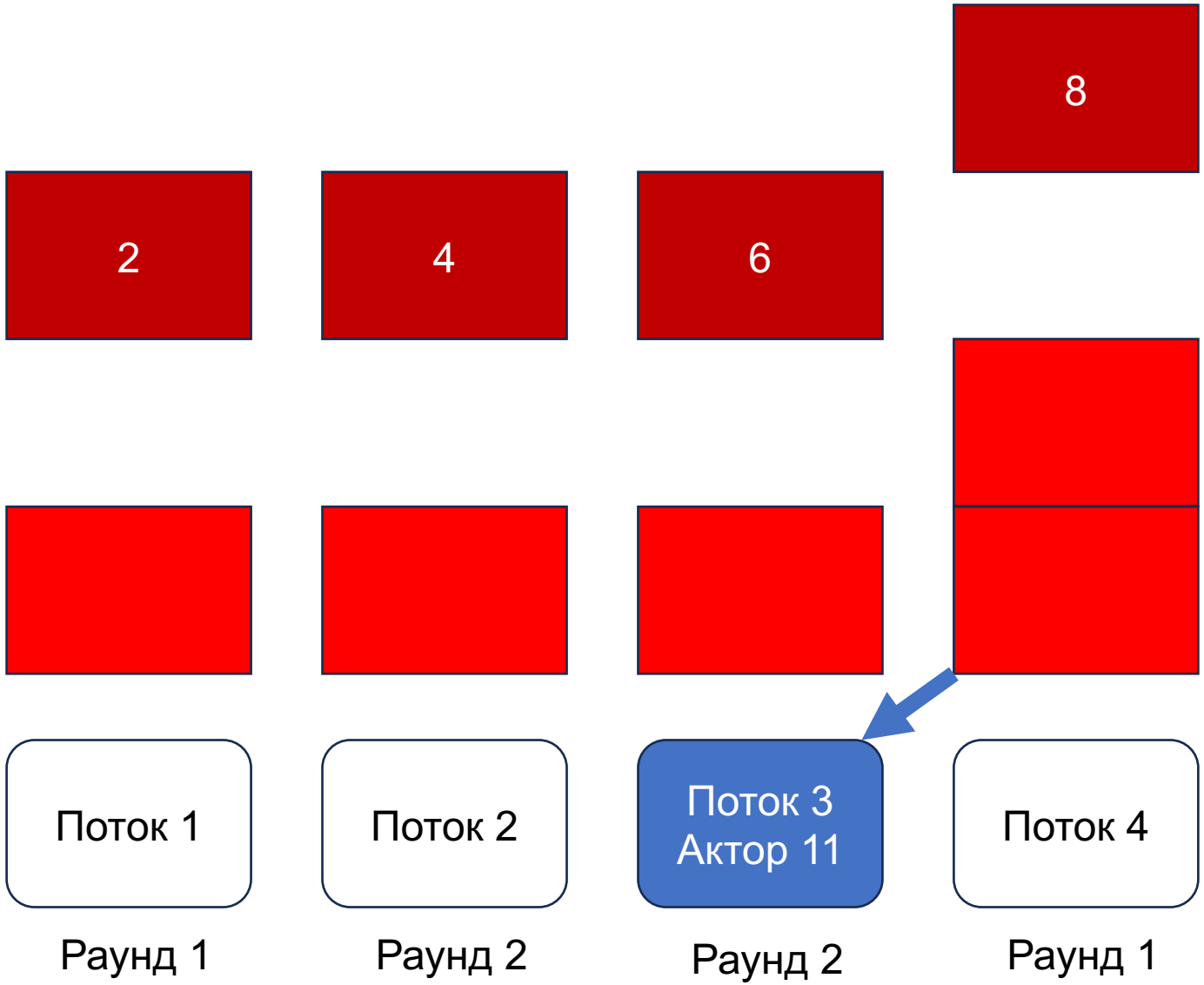
Отказ от единой очереди



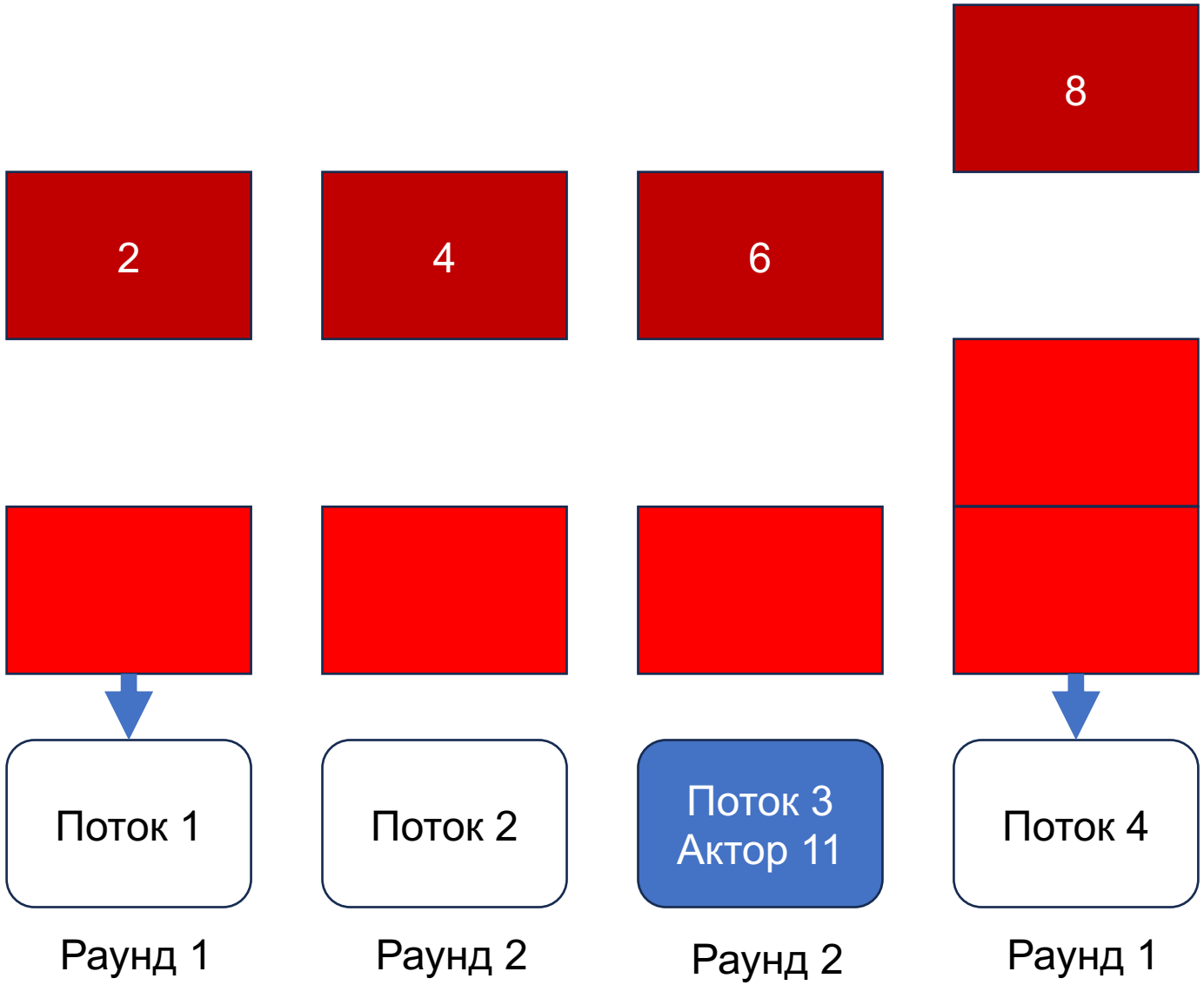
Отказ от единой очереди



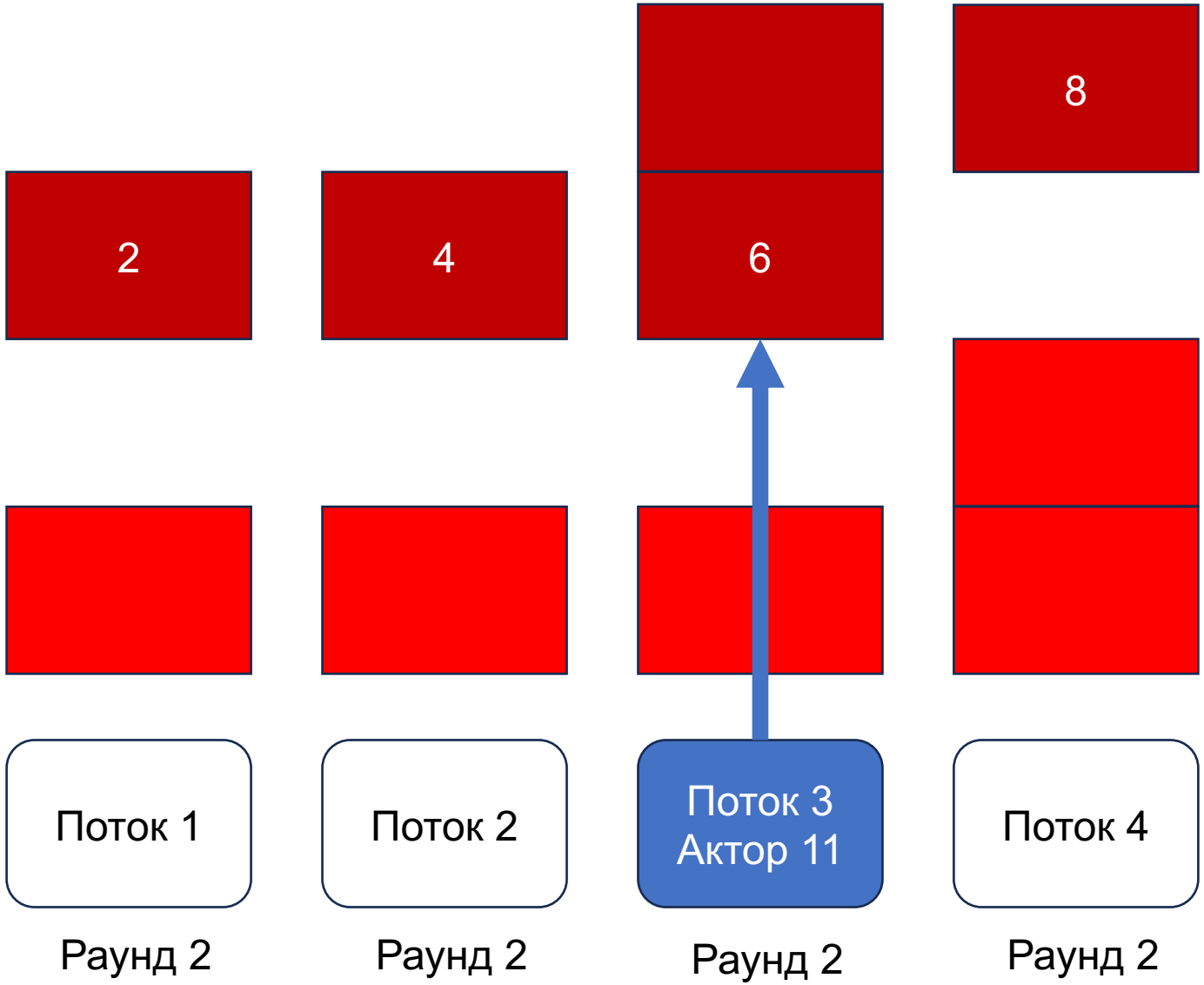
Отказ от единой очереди



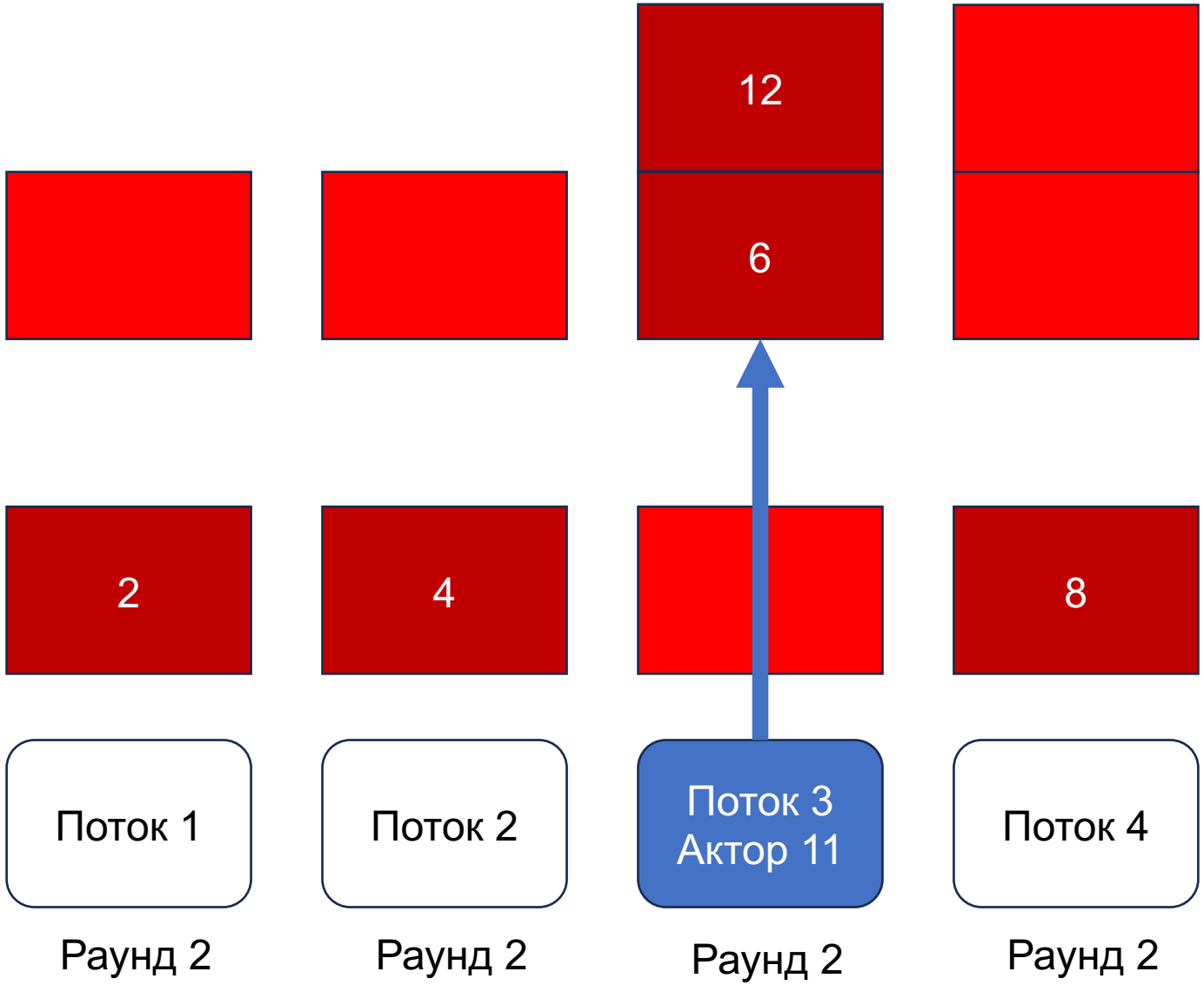
Отказ от единой очереди



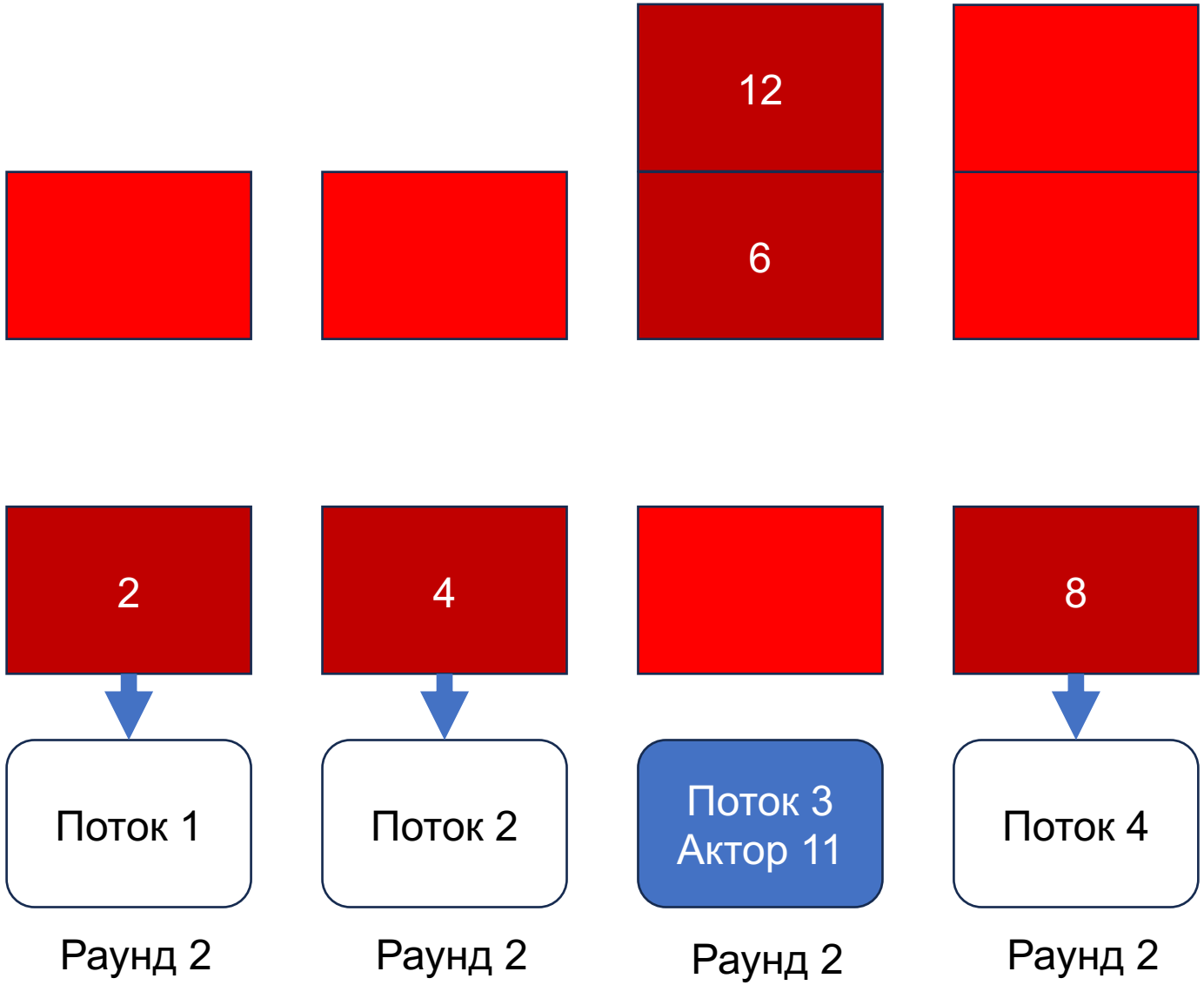
Отказ от единой очереди



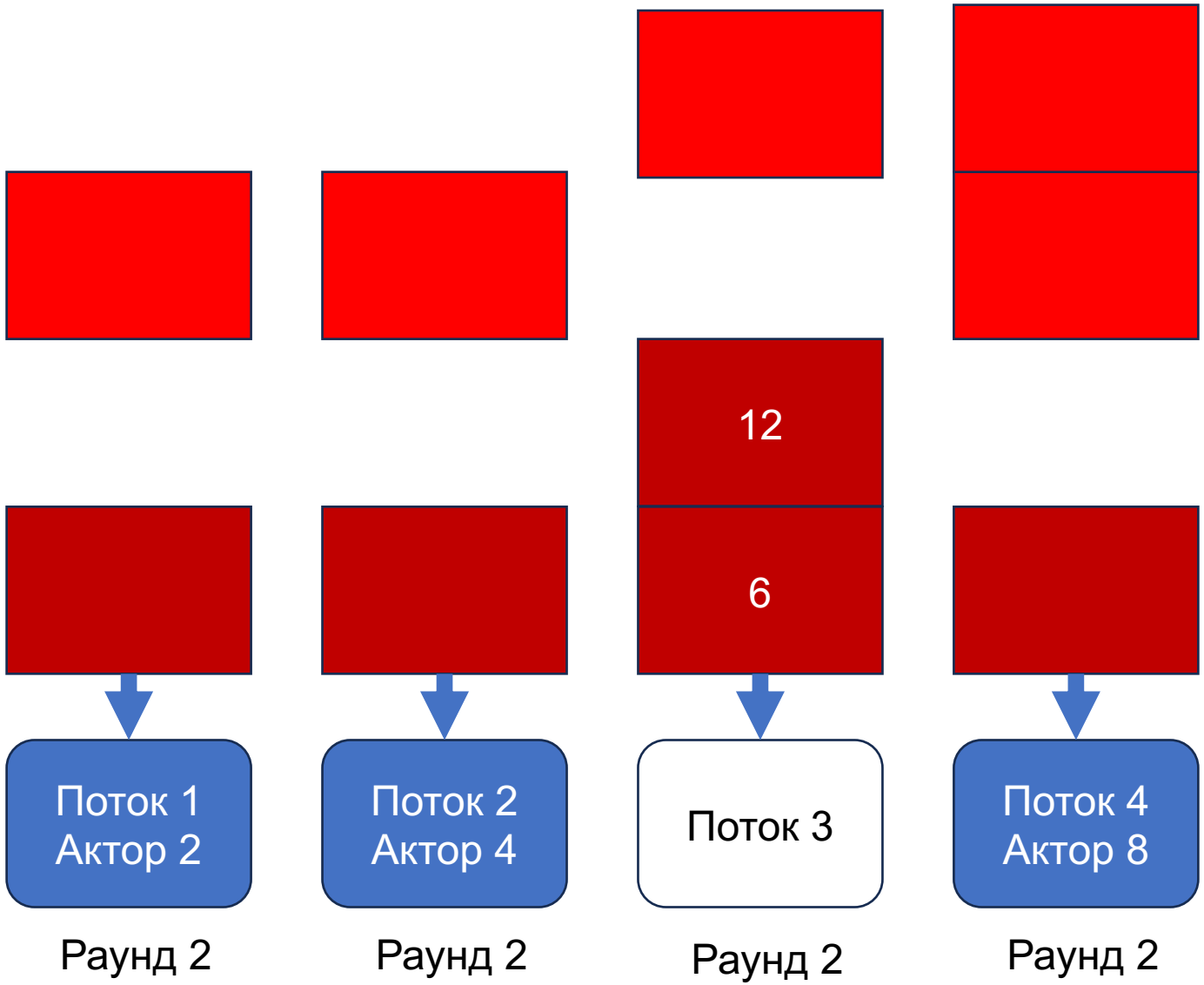
Отказ от единой очереди



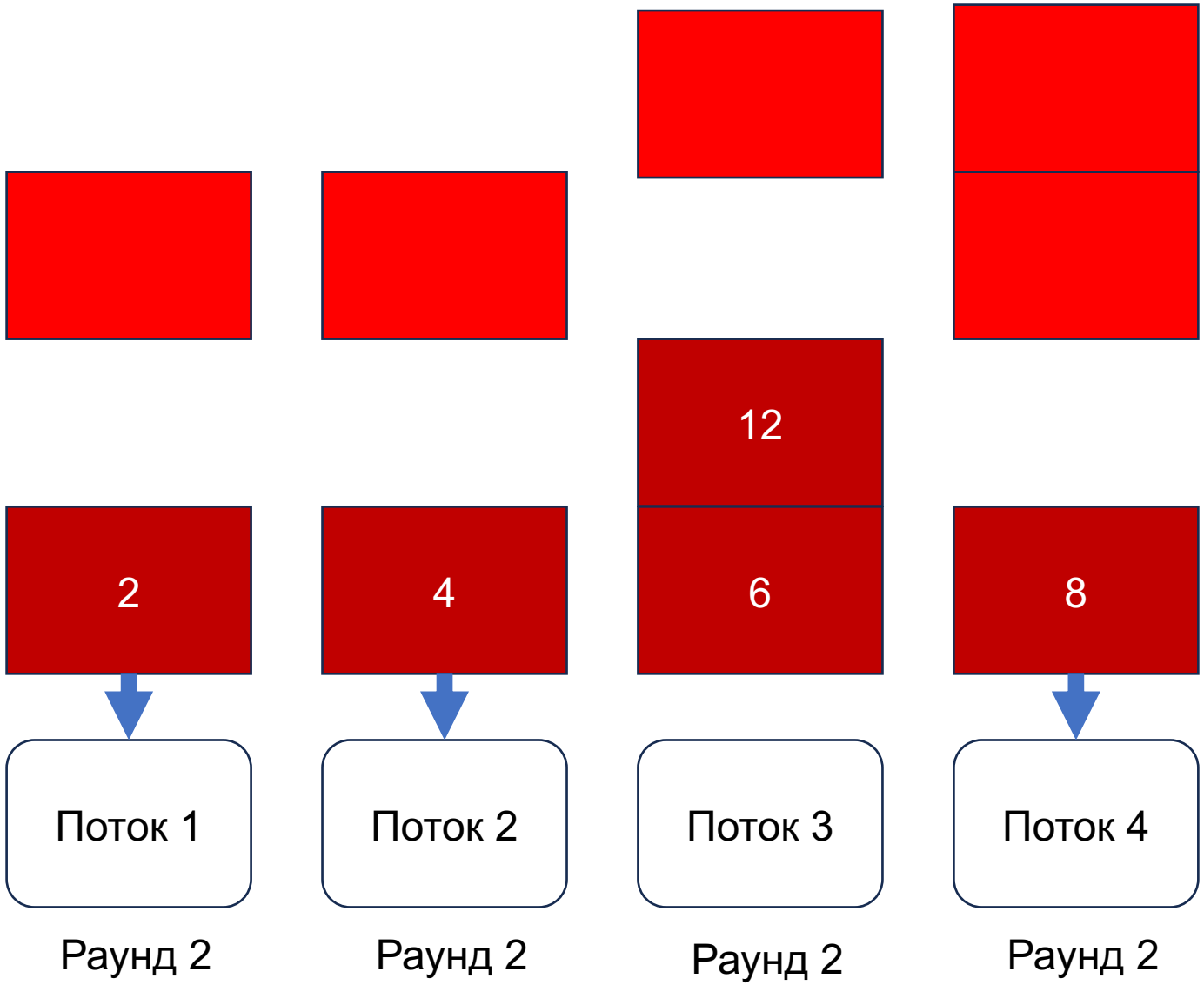
Отказ от единой очереди



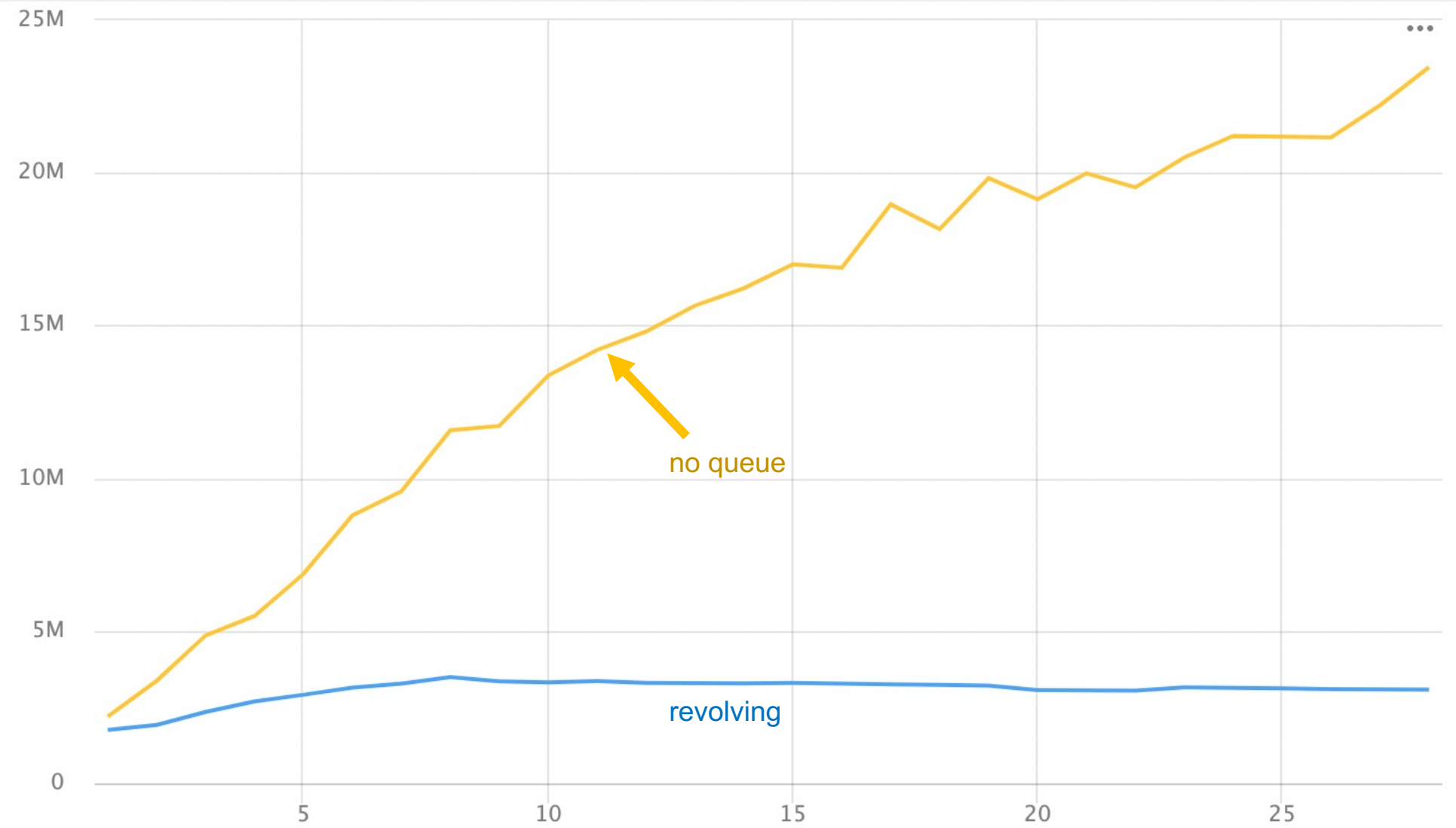
Отказ от единой очереди



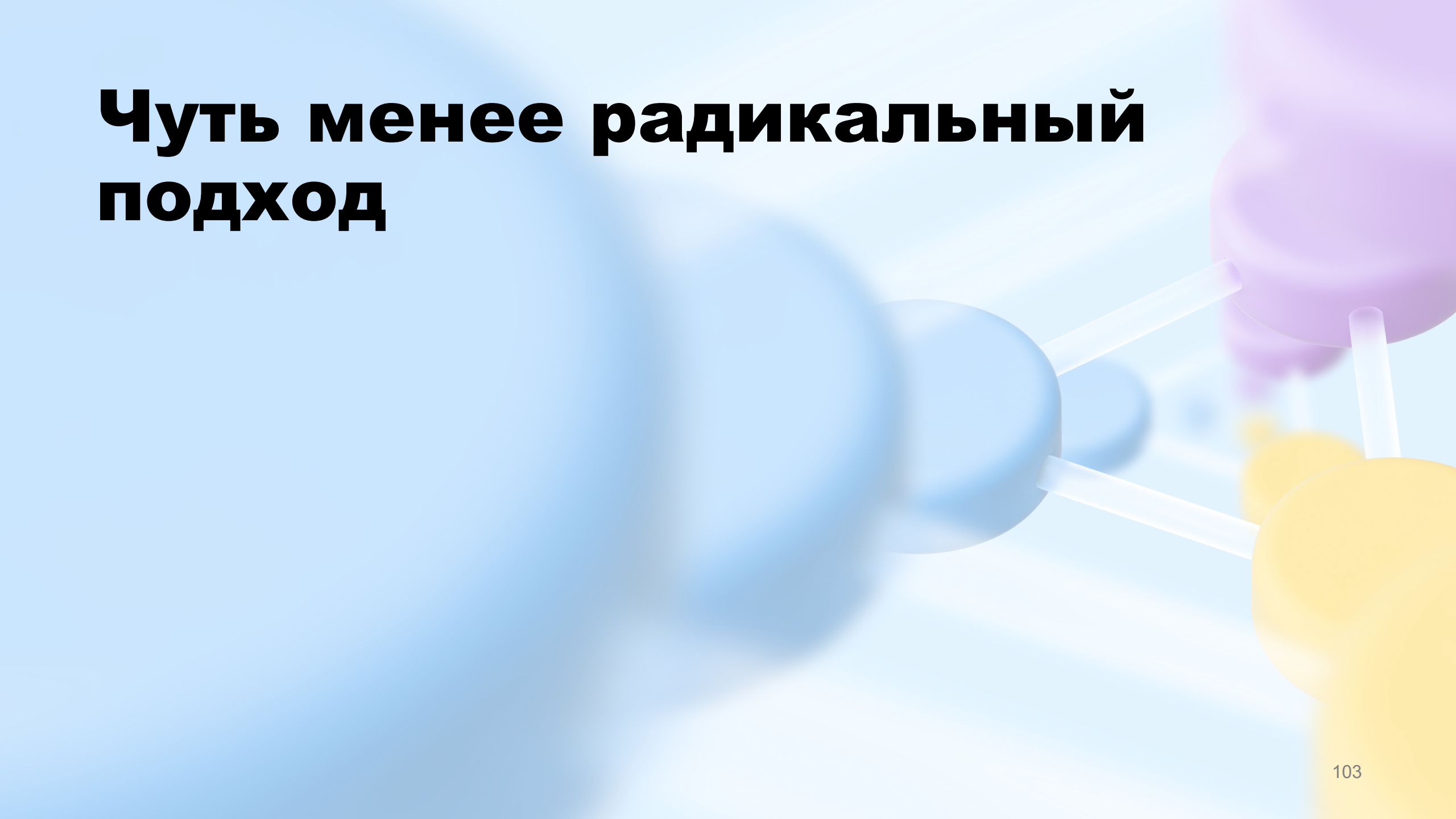
Отказ от единой очереди



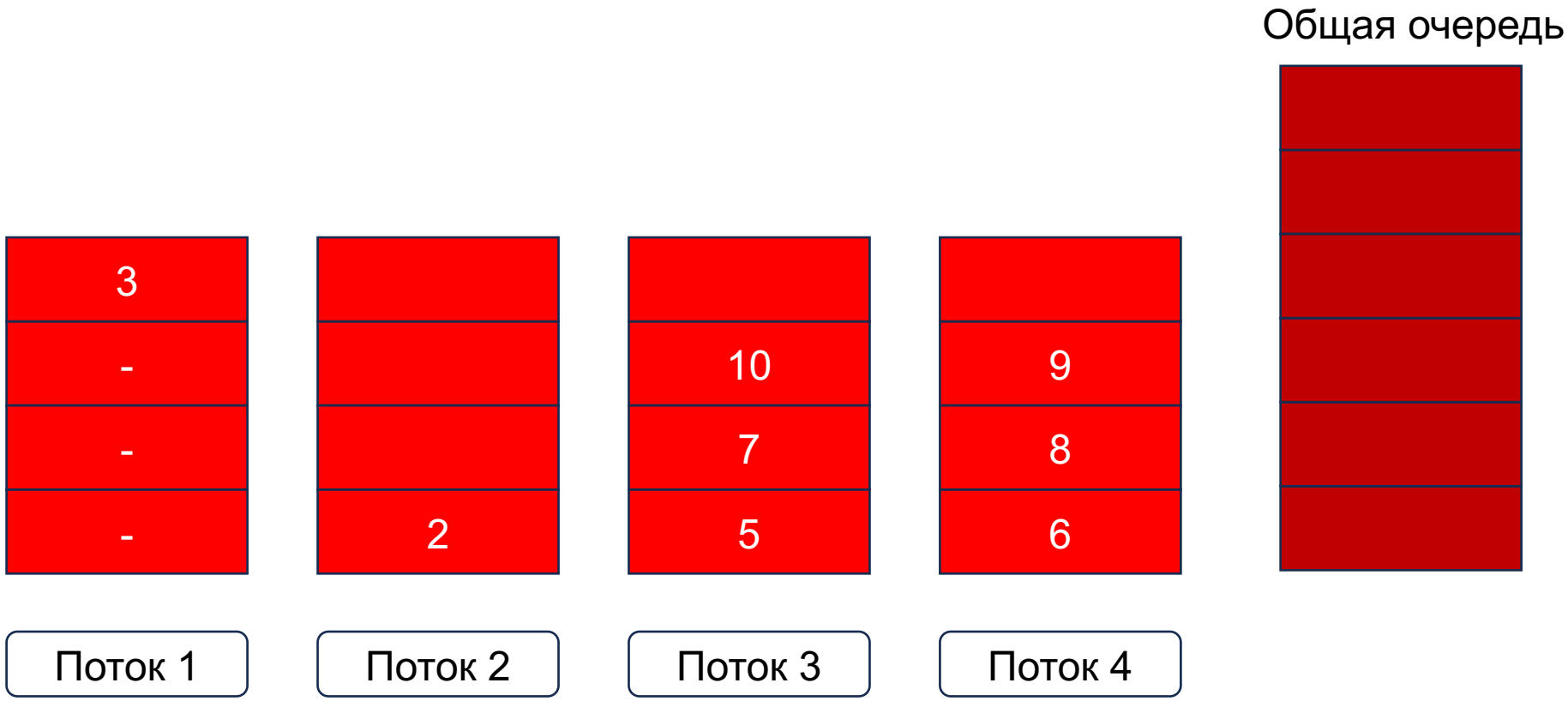
Отказ от единой очереди



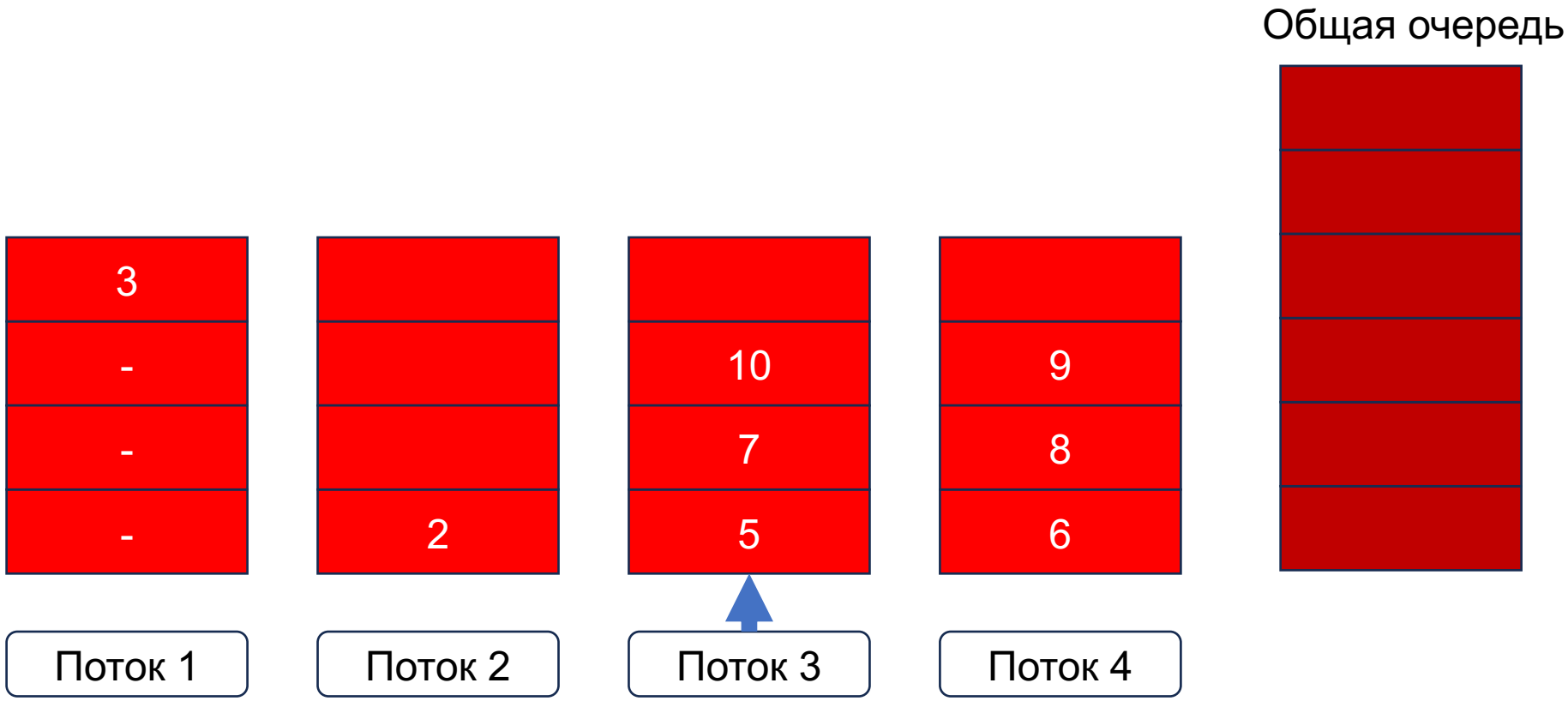
Чуть менее радикальный подход



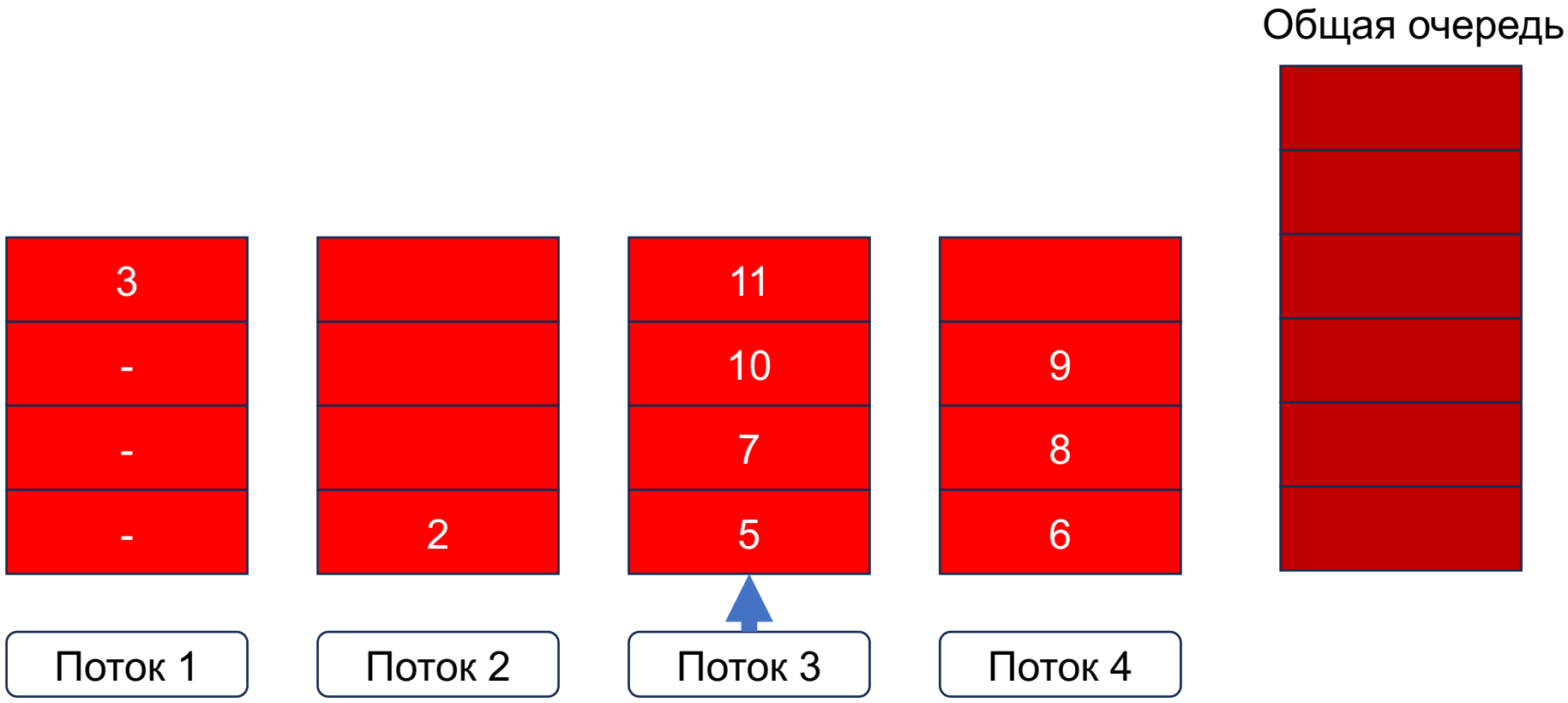
Короткие per-thread очереди



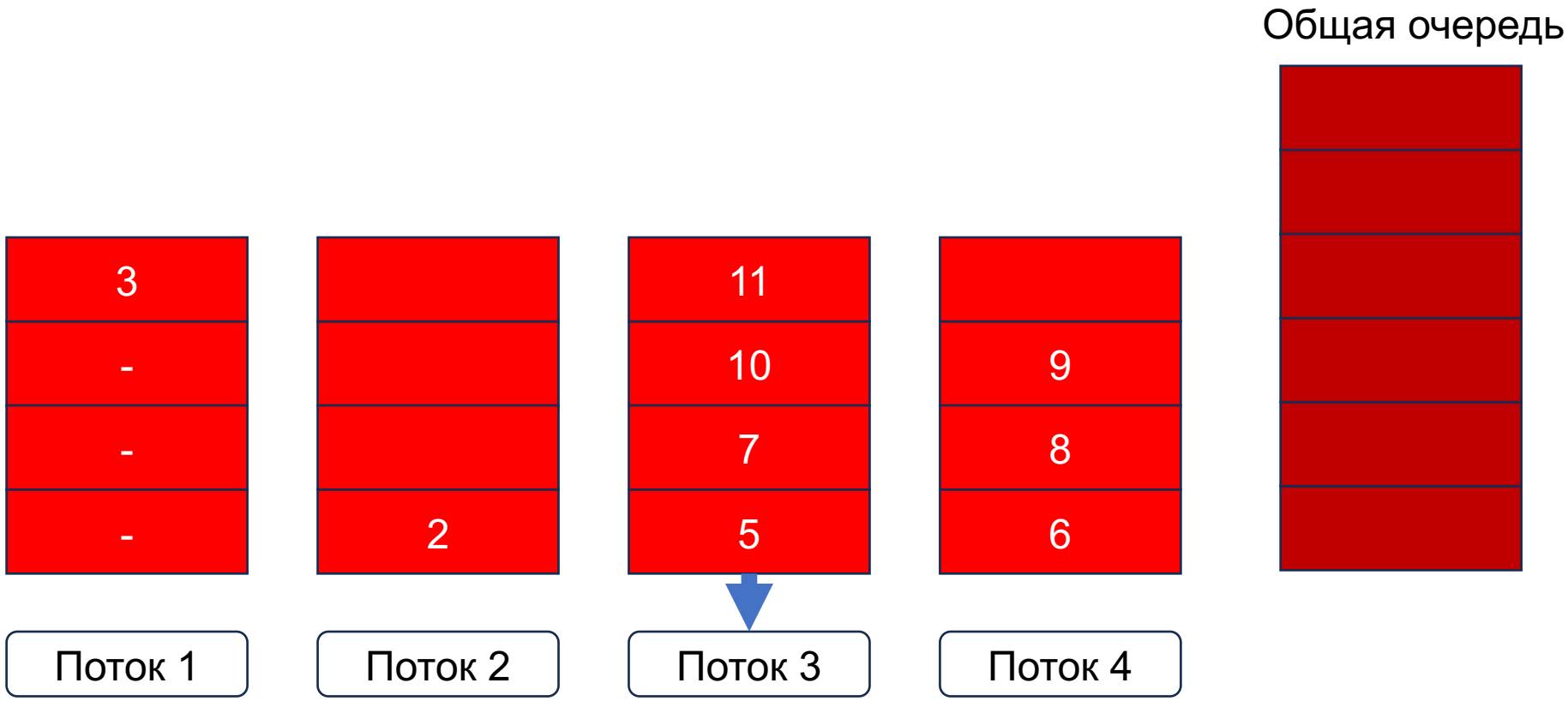
Короткие per-thread очереди



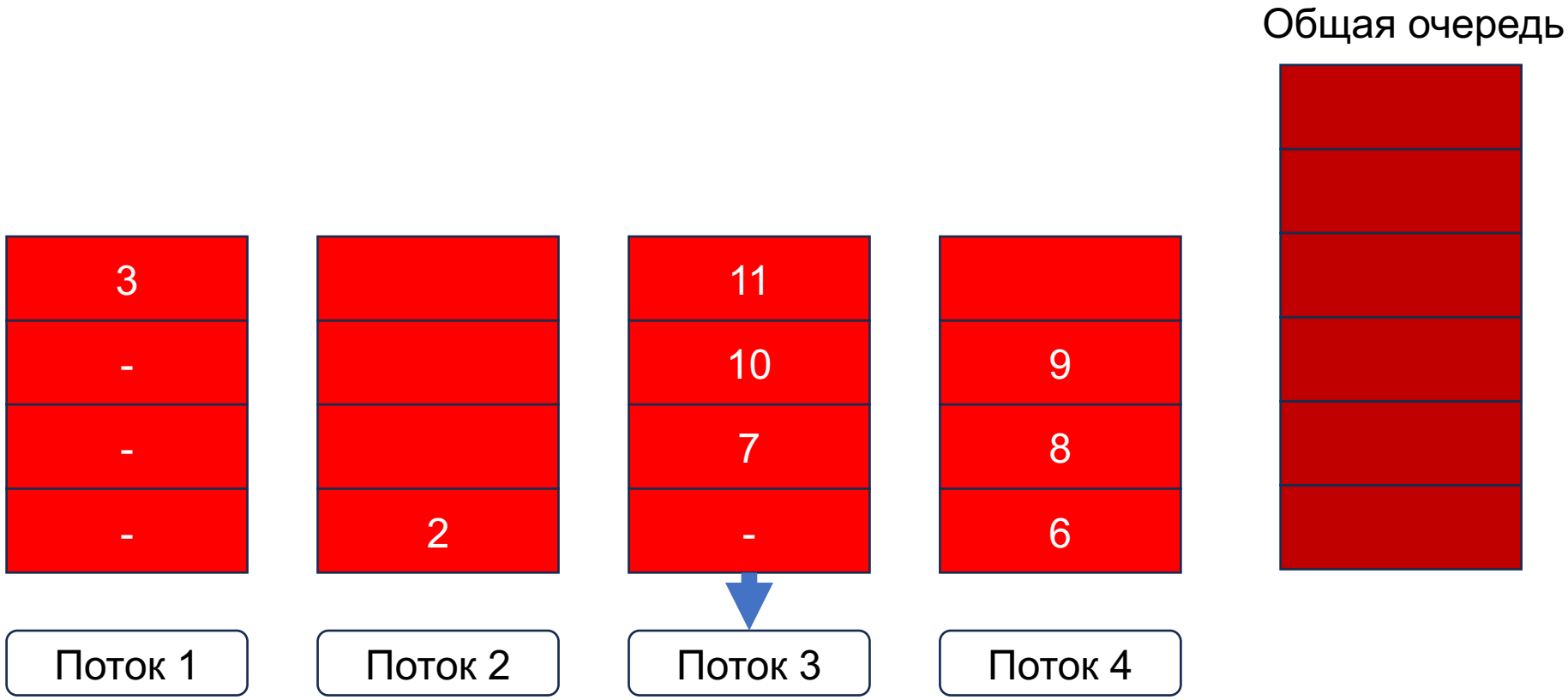
Короткие per-thread очереди



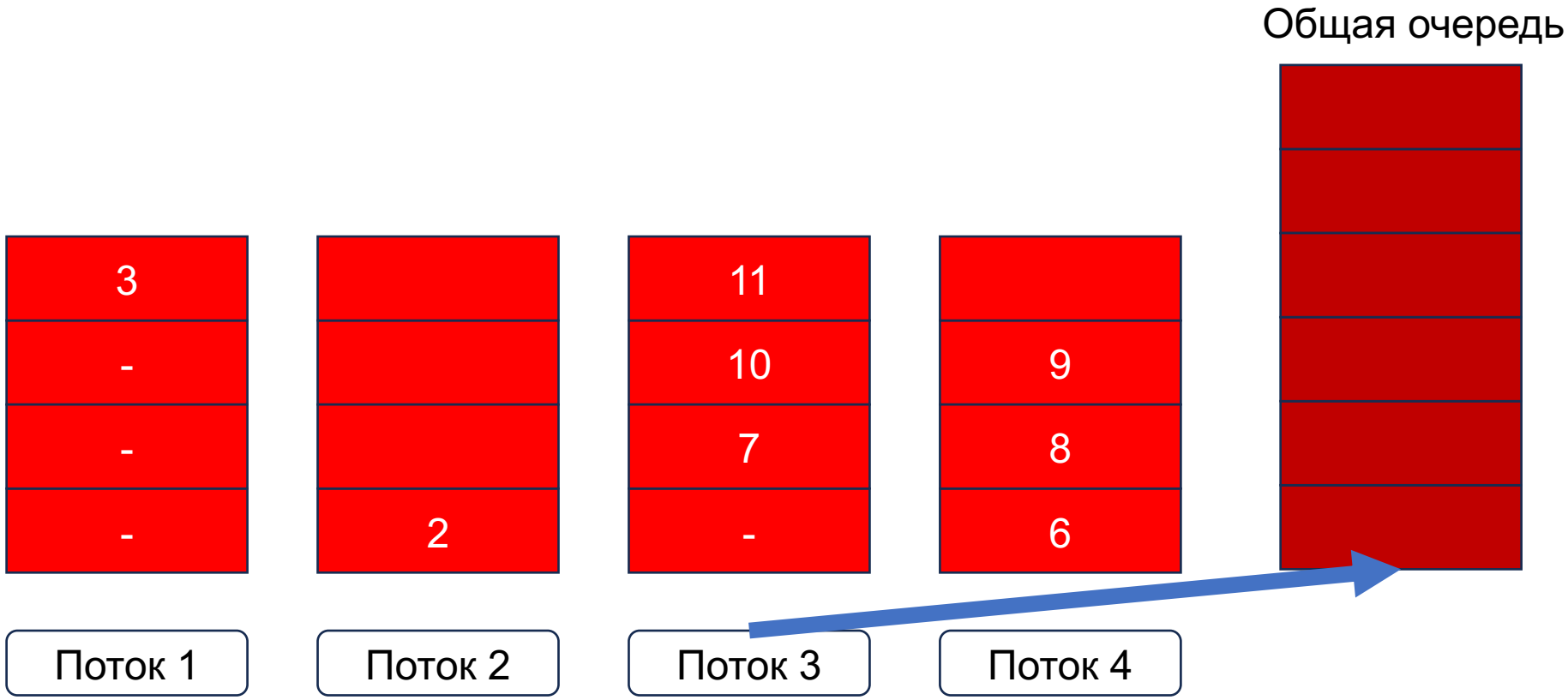
Короткие per-thread очереди



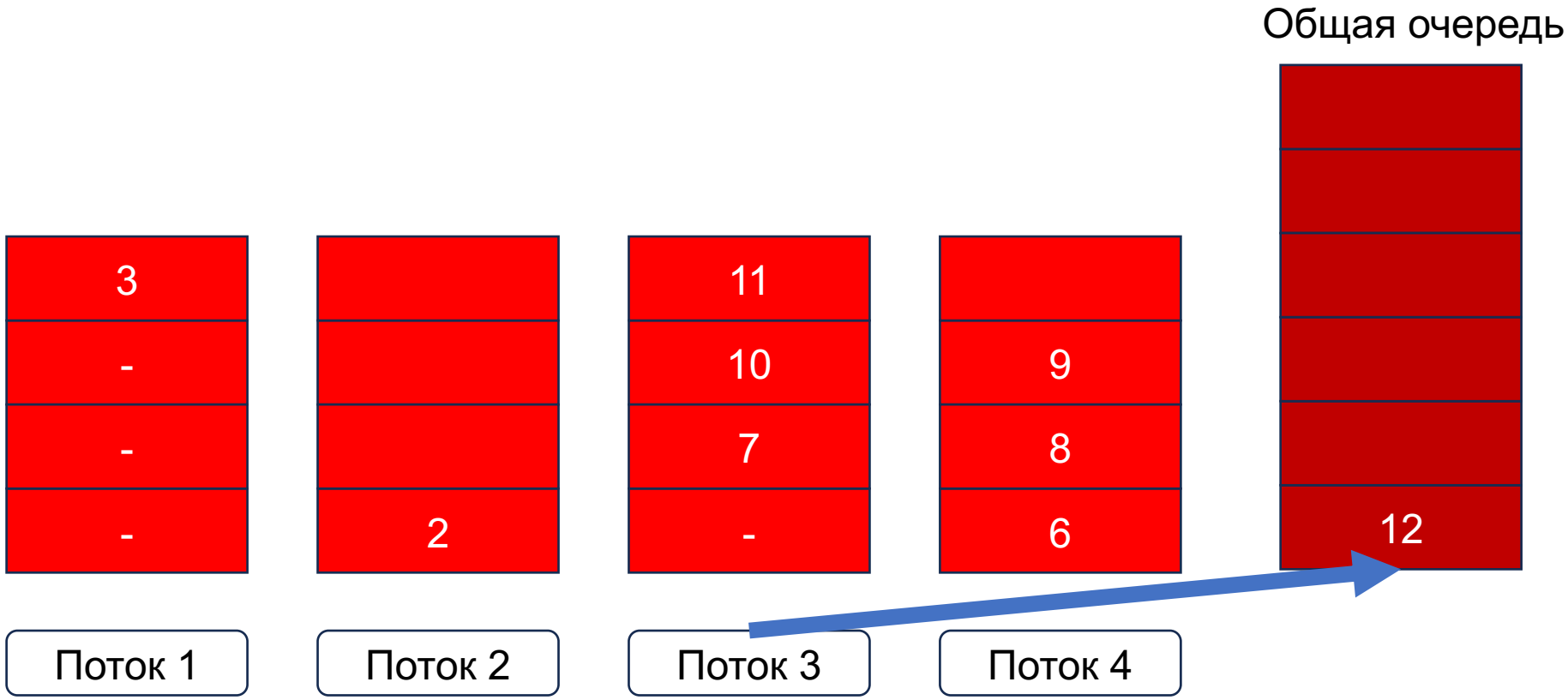
Короткие per-thread очереди



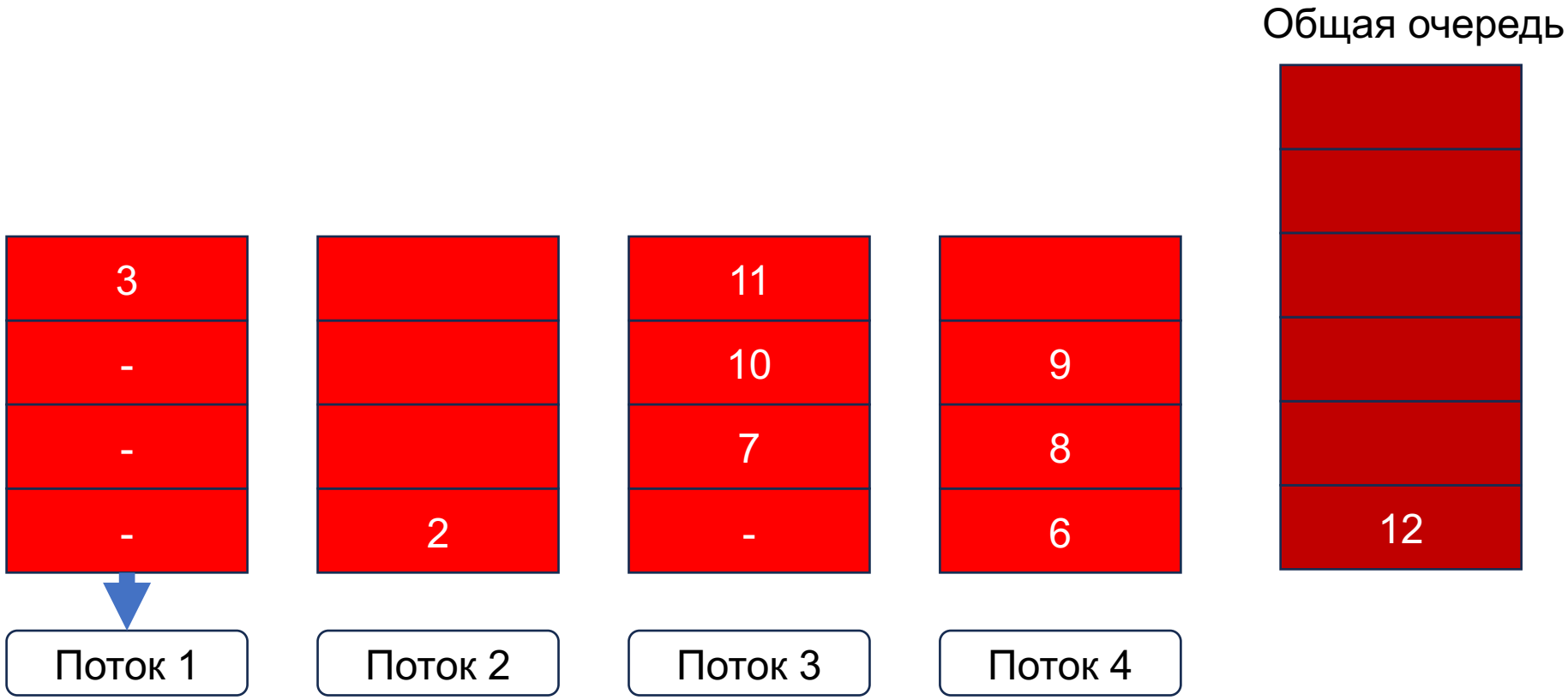
Короткие per-thread очереди



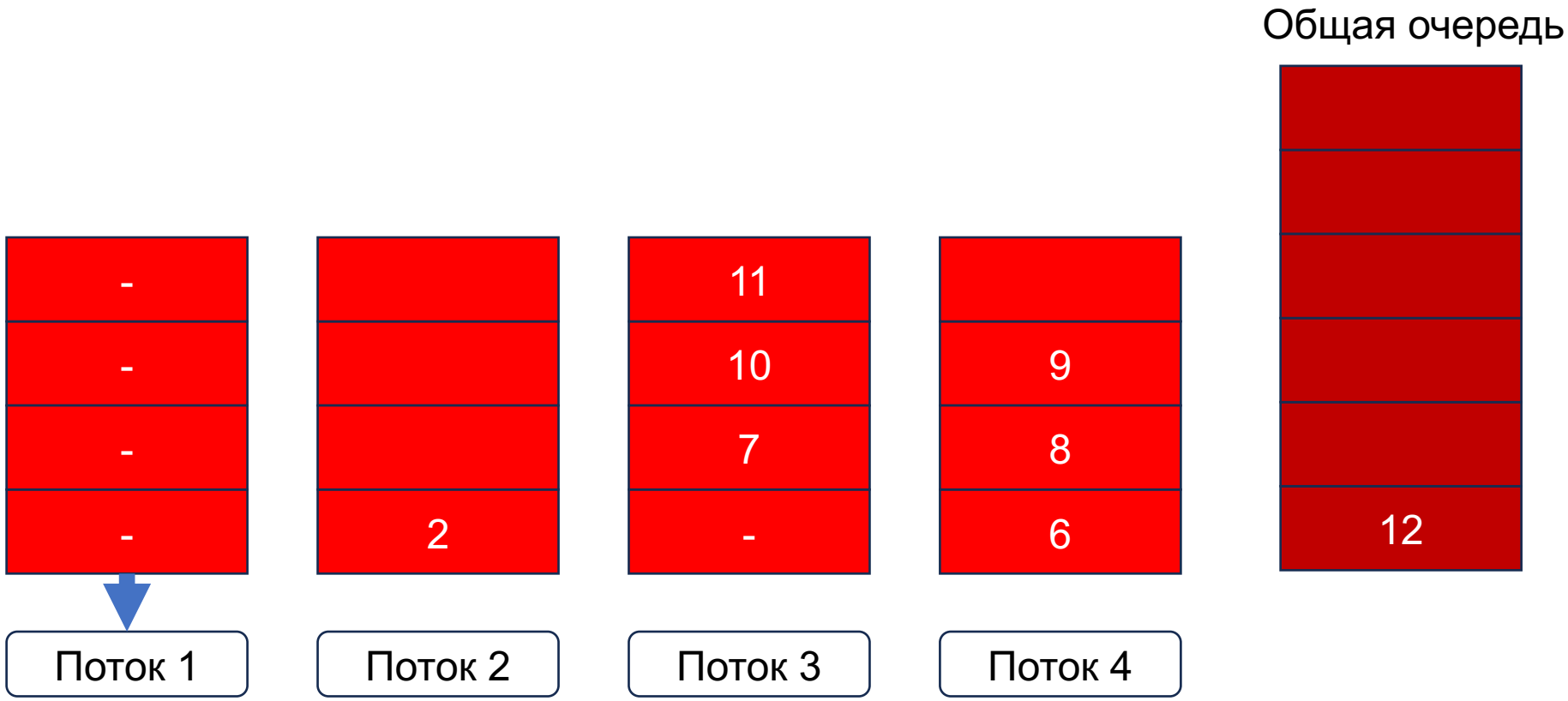
Короткие per-thread очереди



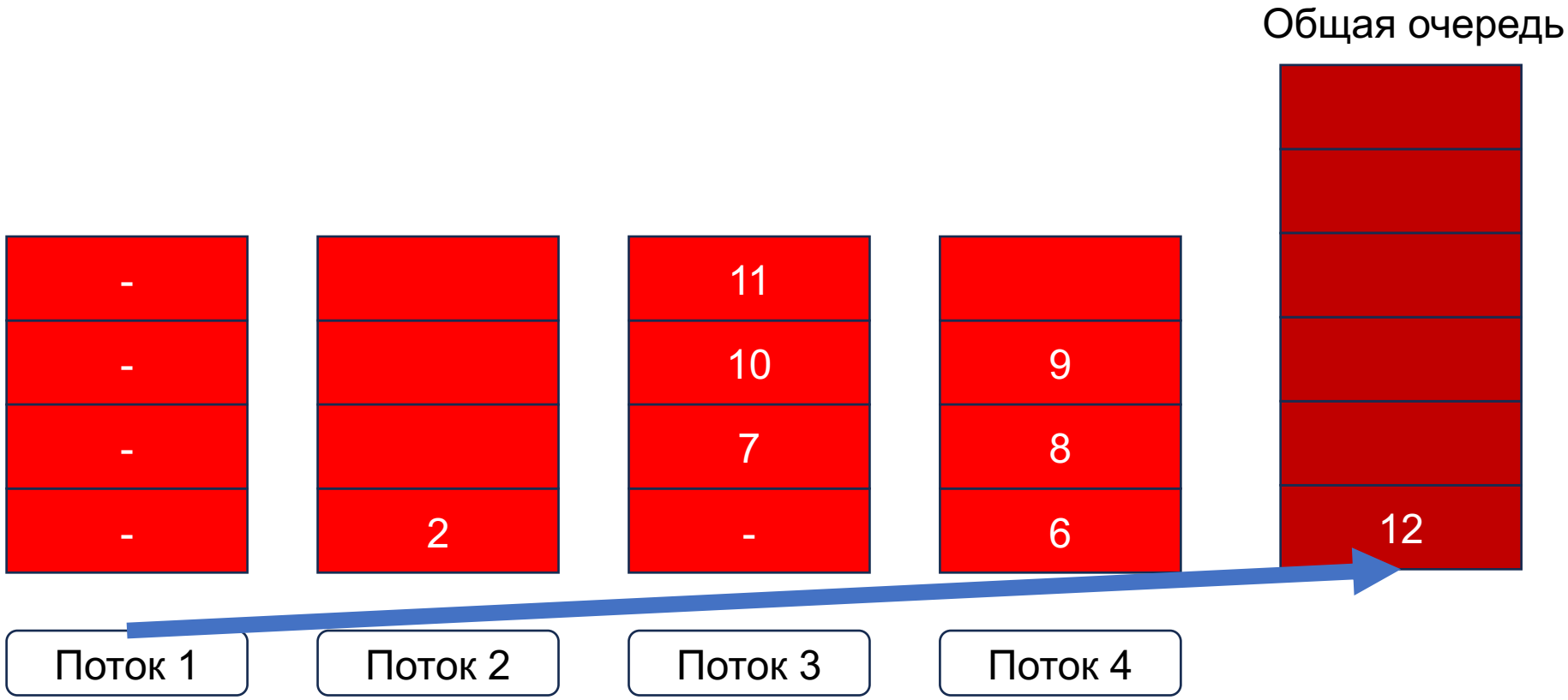
Короткие per-thread очереди



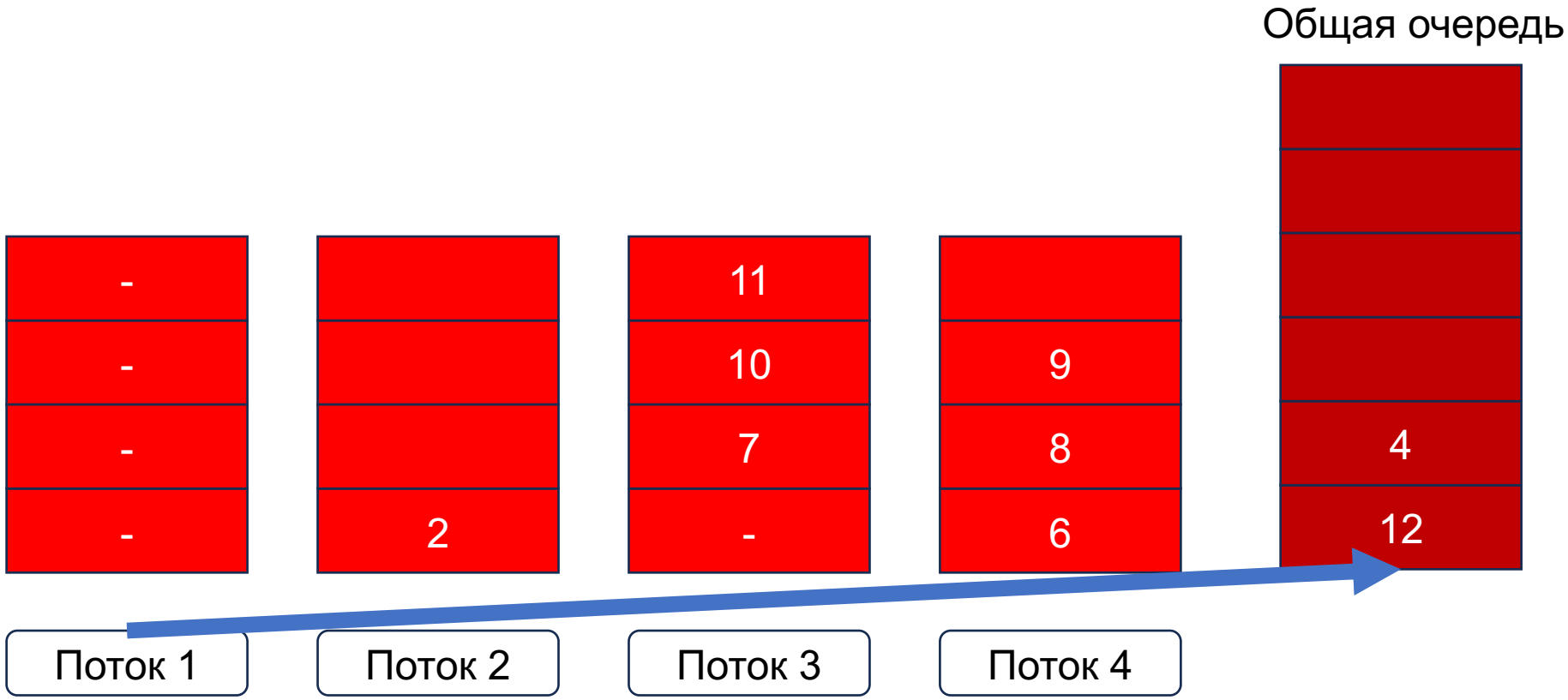
Короткие per-thread очереди



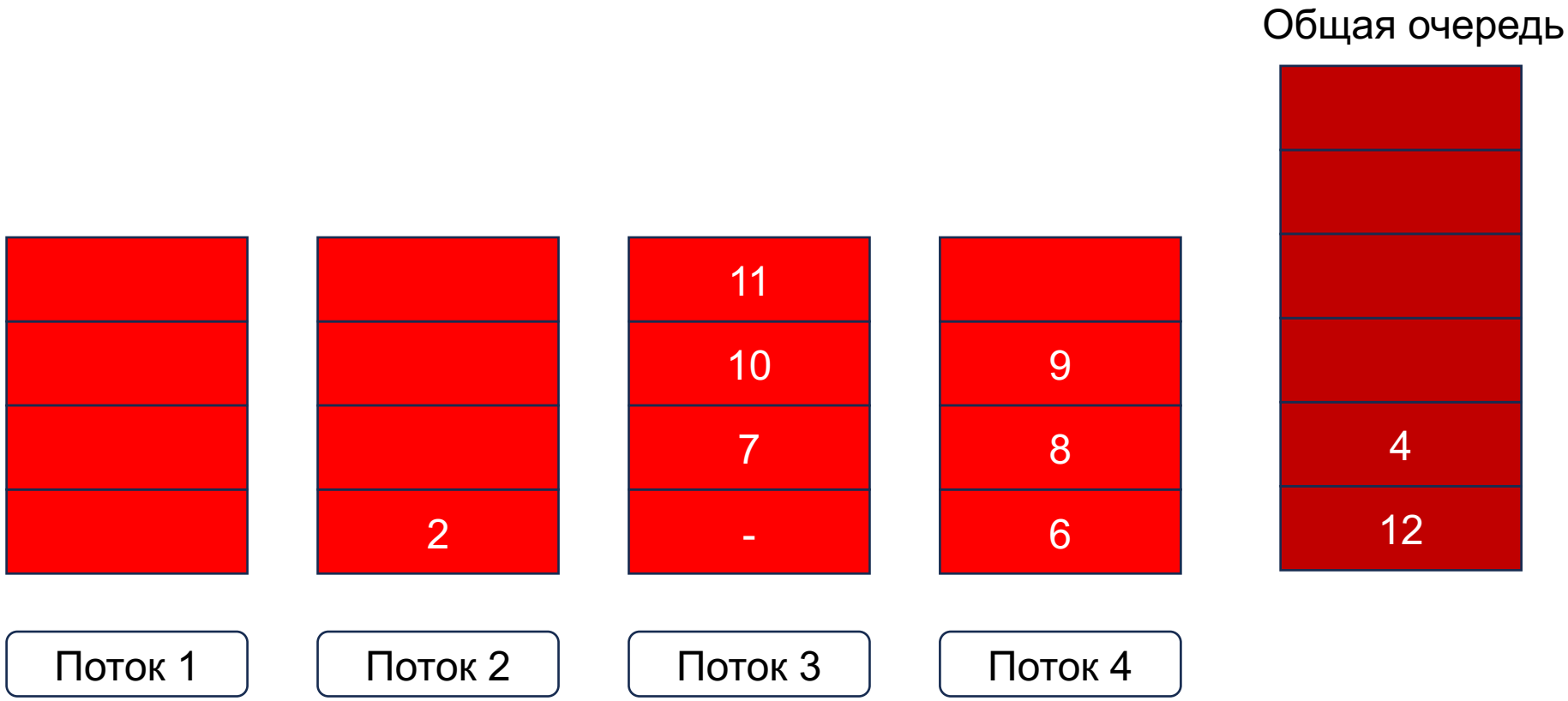
Короткие per-thread очереди



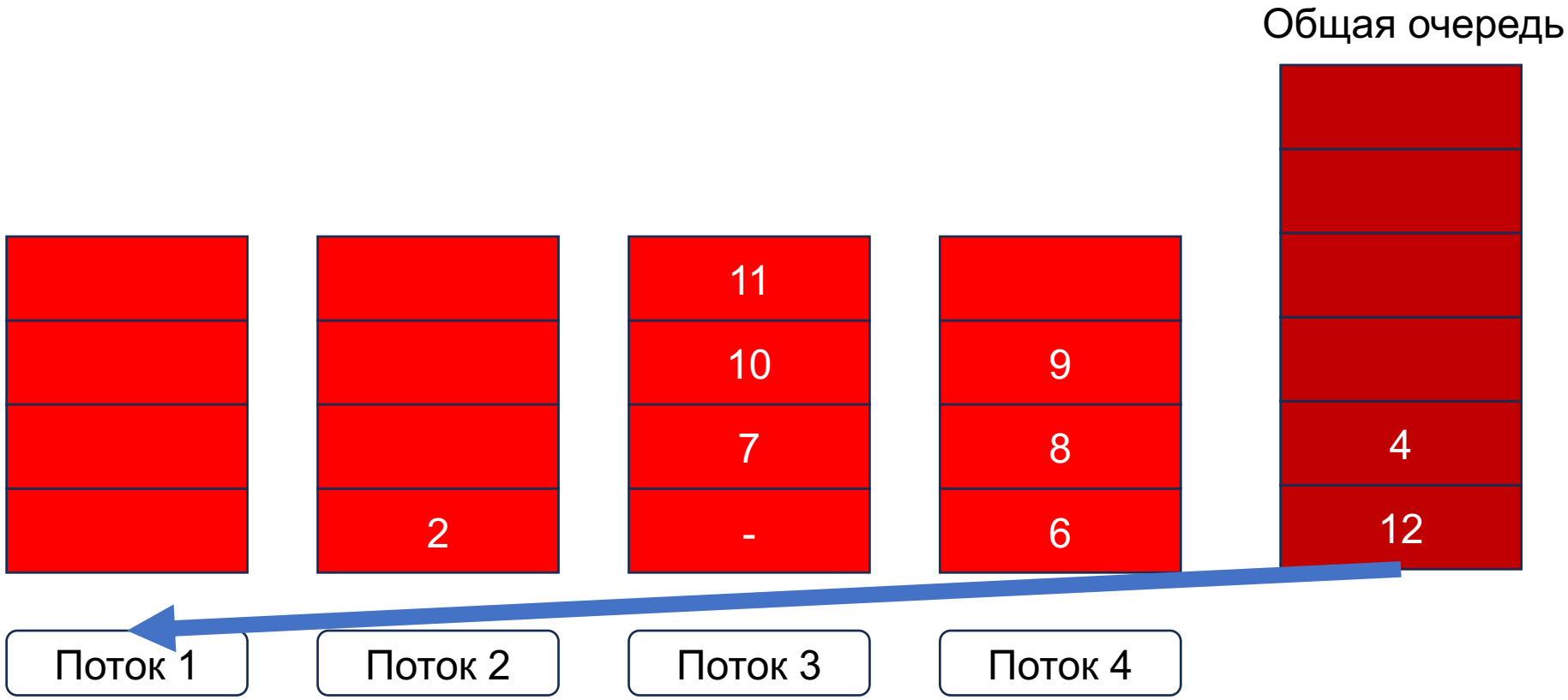
Короткие per-thread очереди



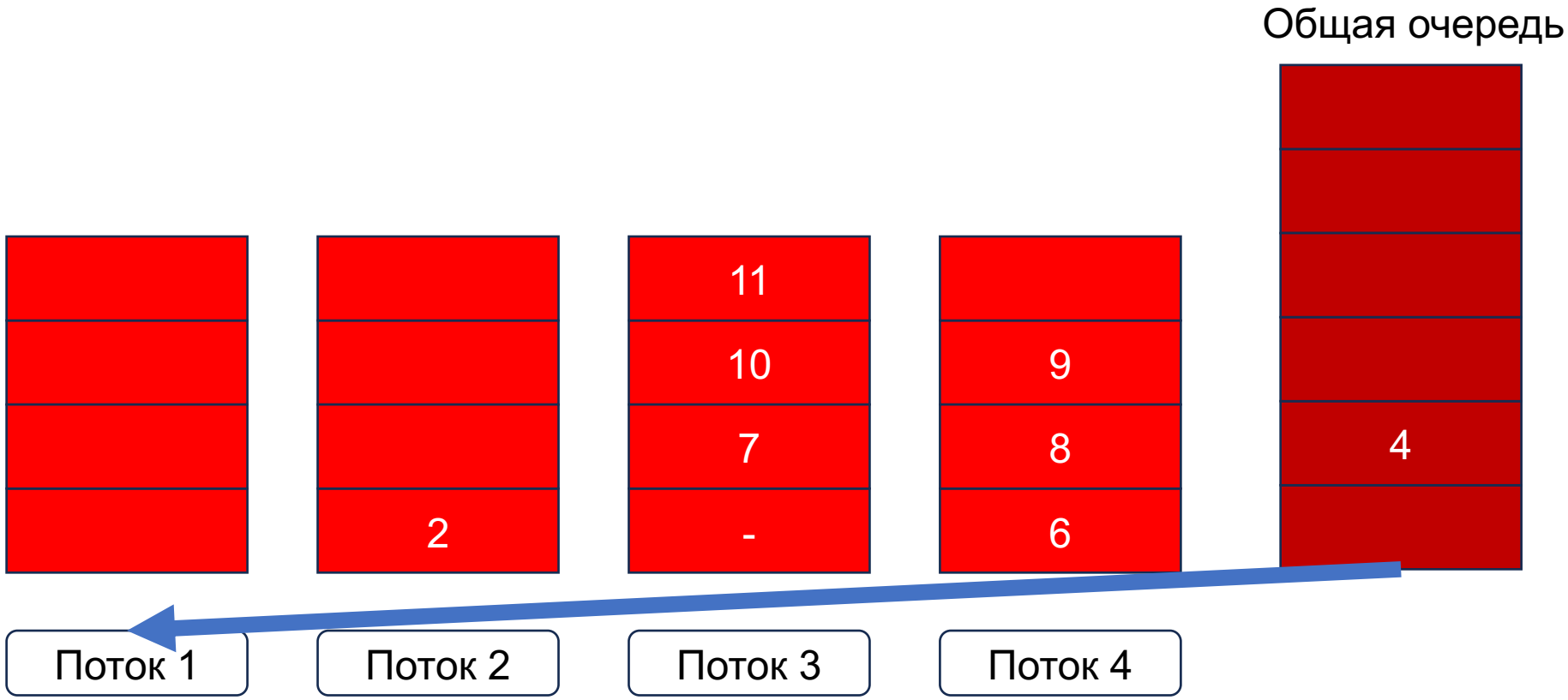
Короткие per-thread очереди



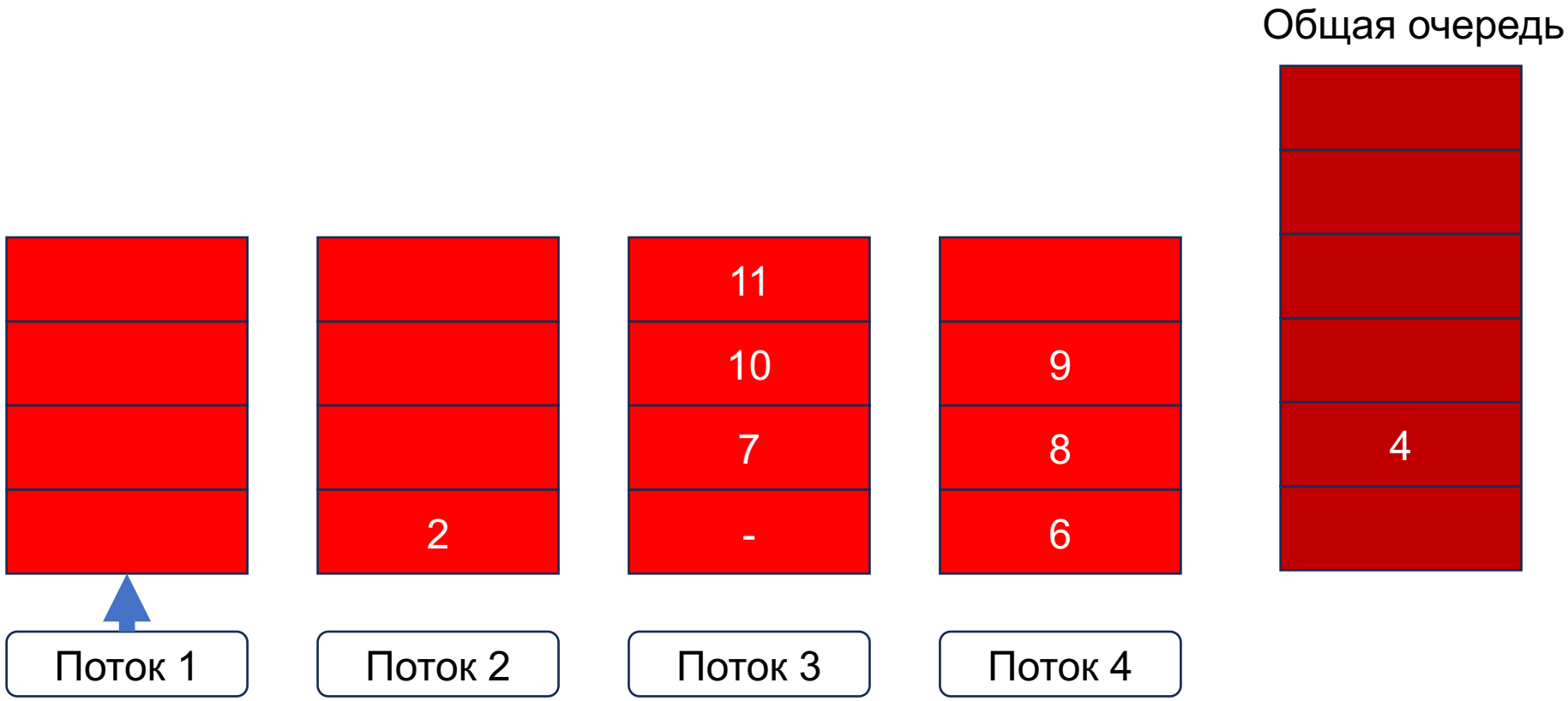
Короткие per-thread очереди



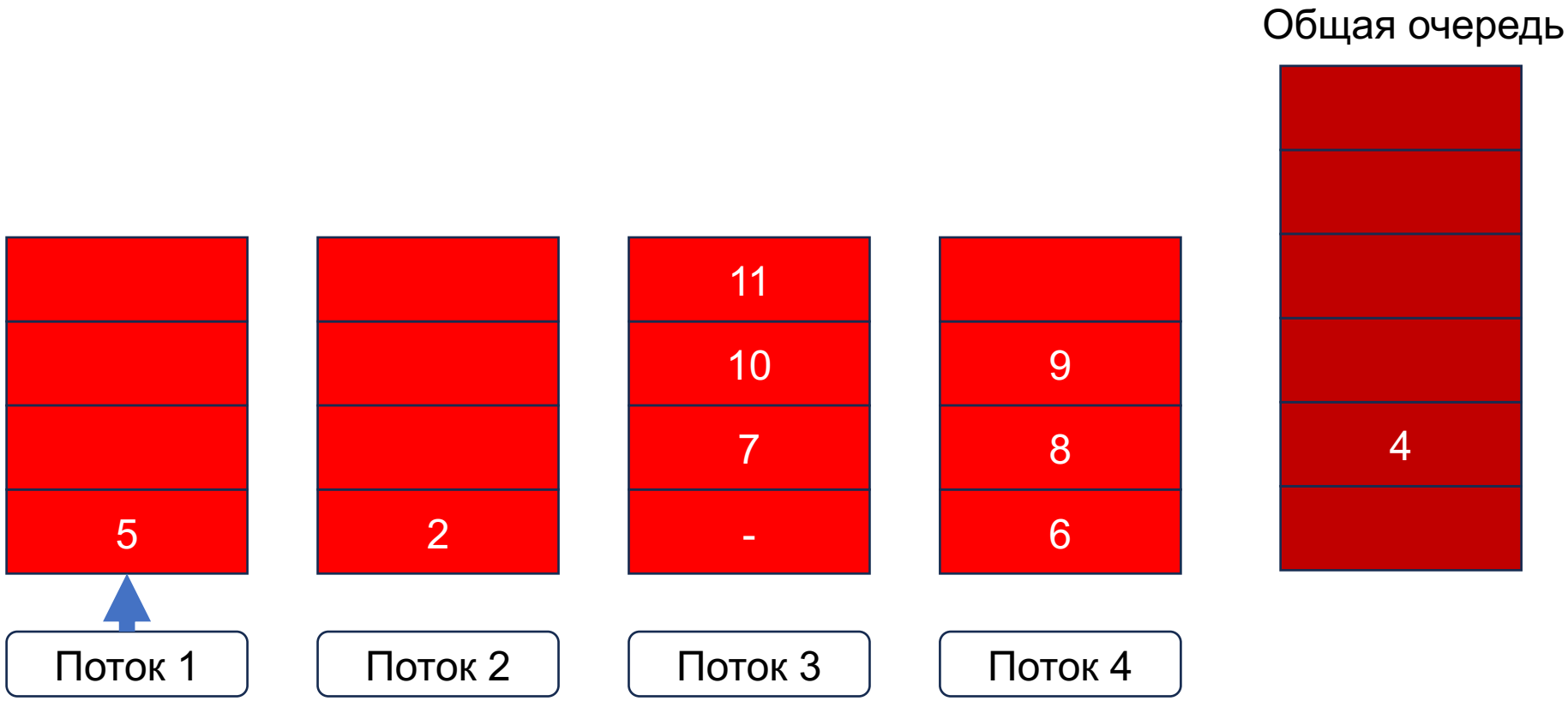
Короткие per-thread очереди



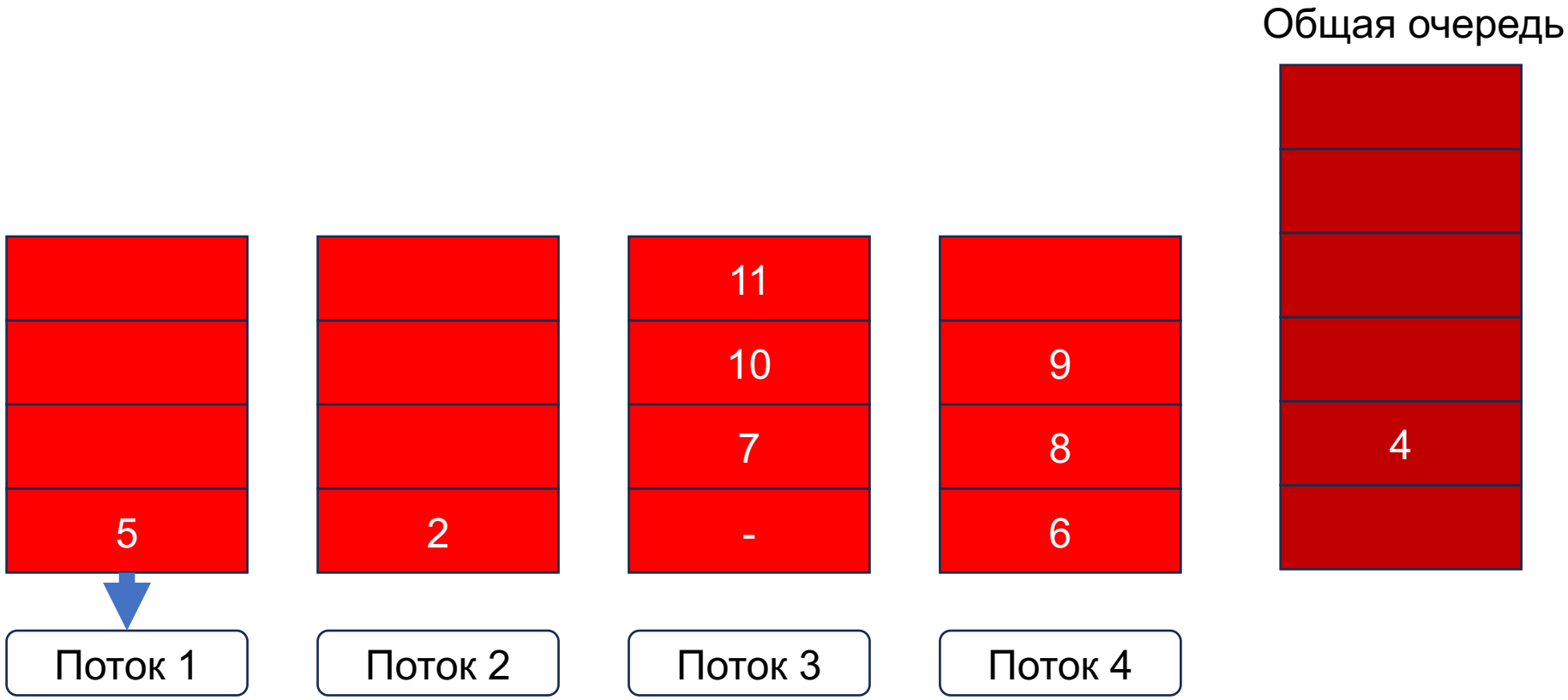
Короткие per-thread очереди



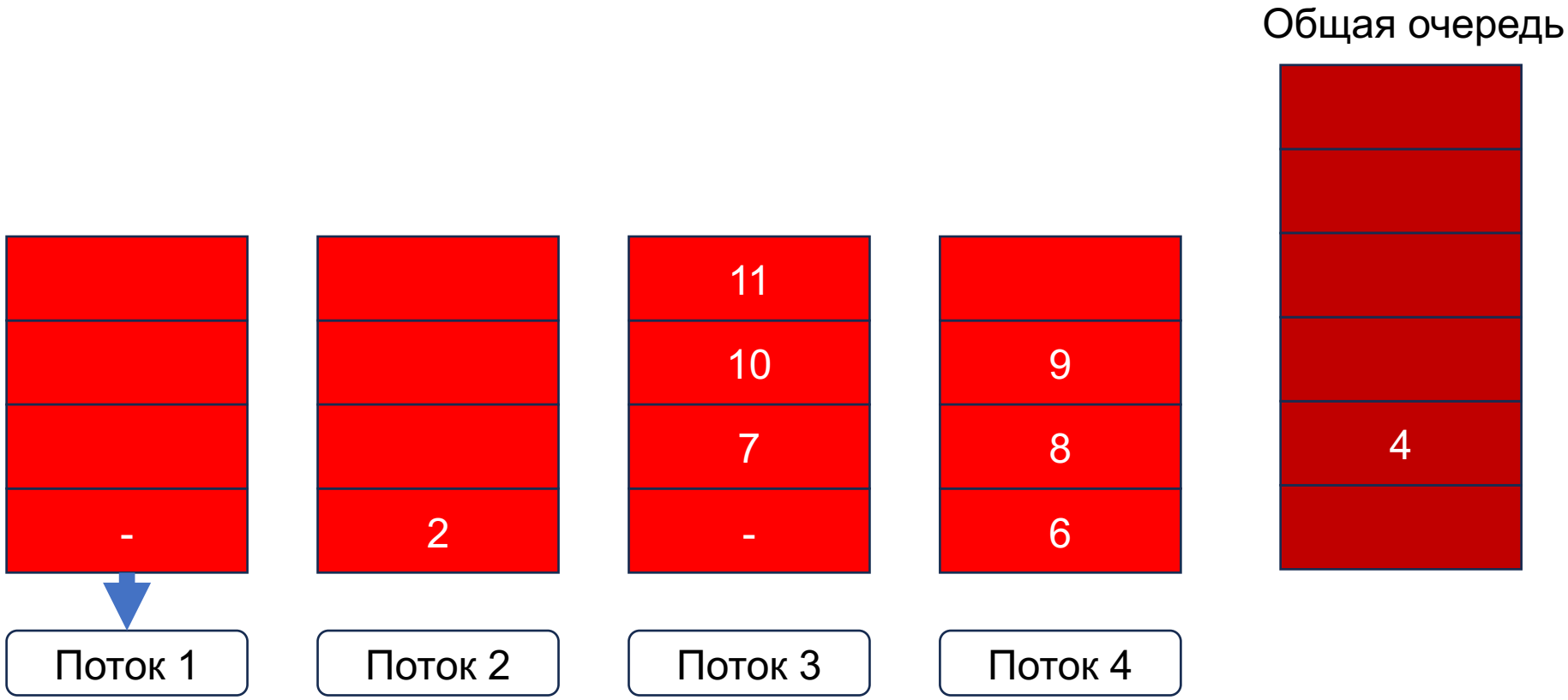
Короткие per-thread очереди



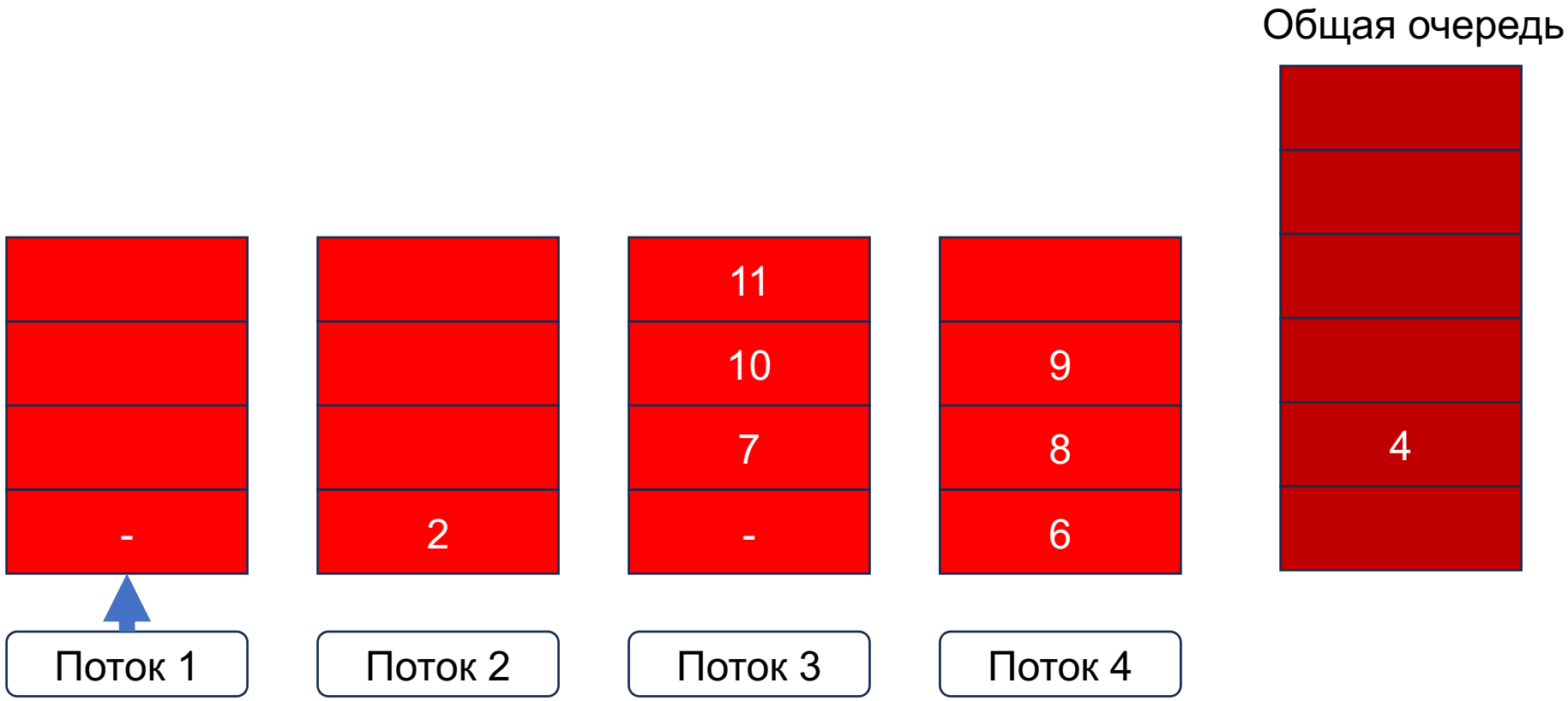
Короткие per-thread очереди



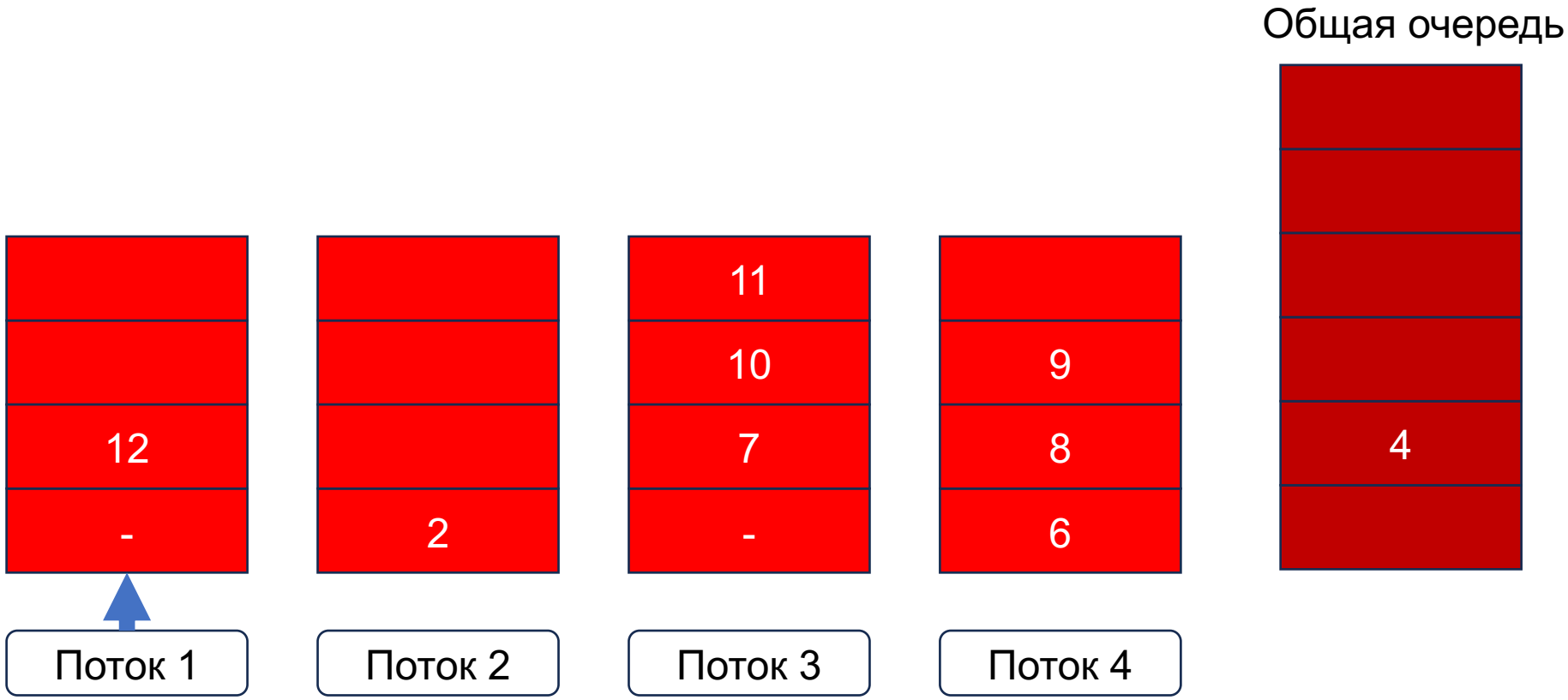
Короткие per-thread очереди



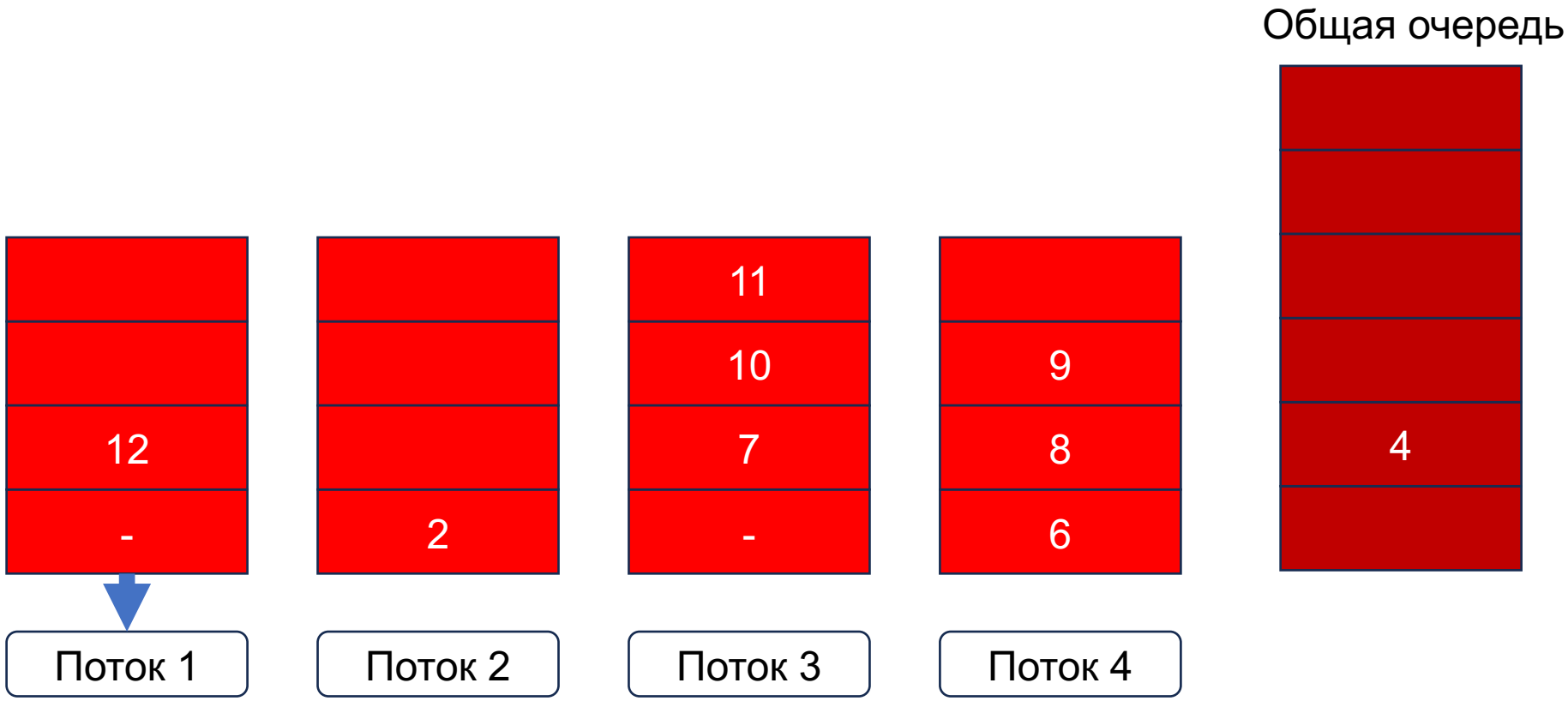
Короткие per-thread очереди



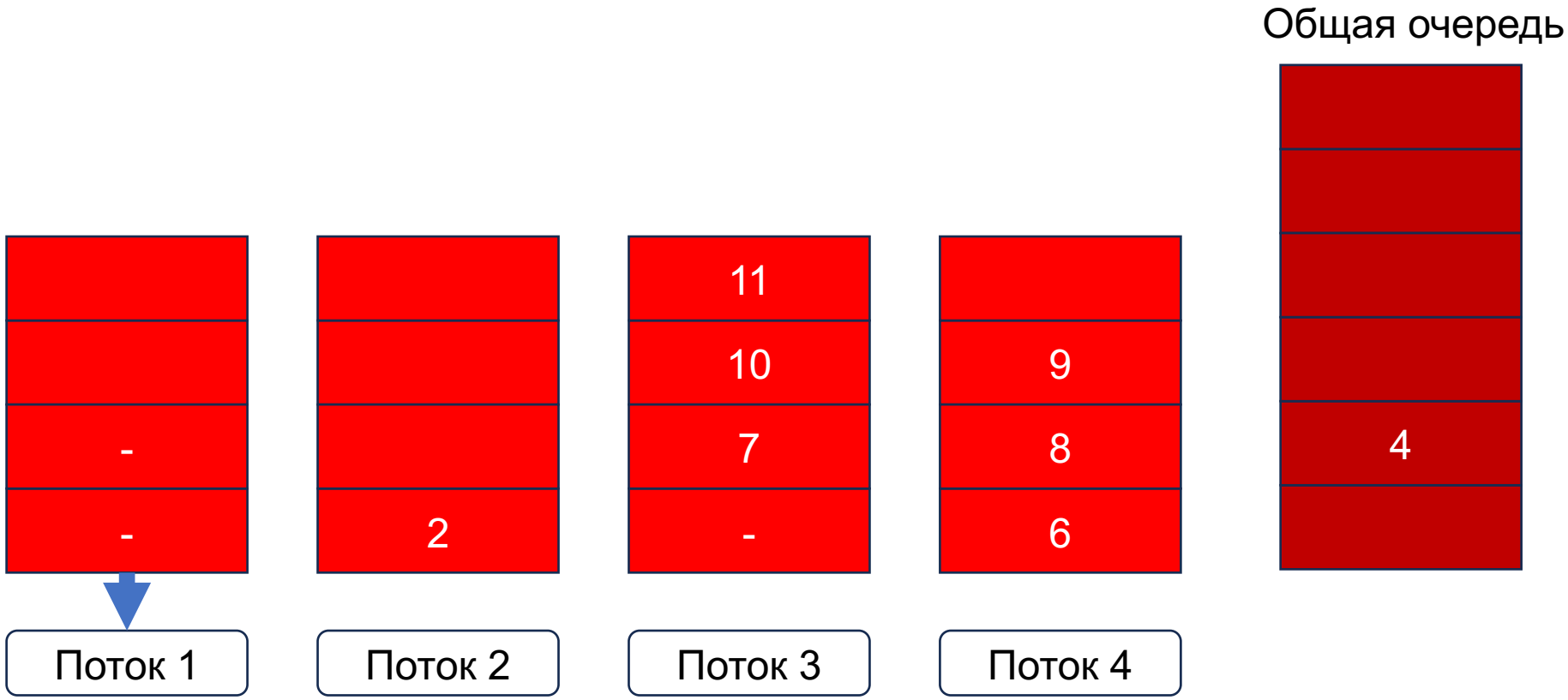
Короткие per-thread очереди



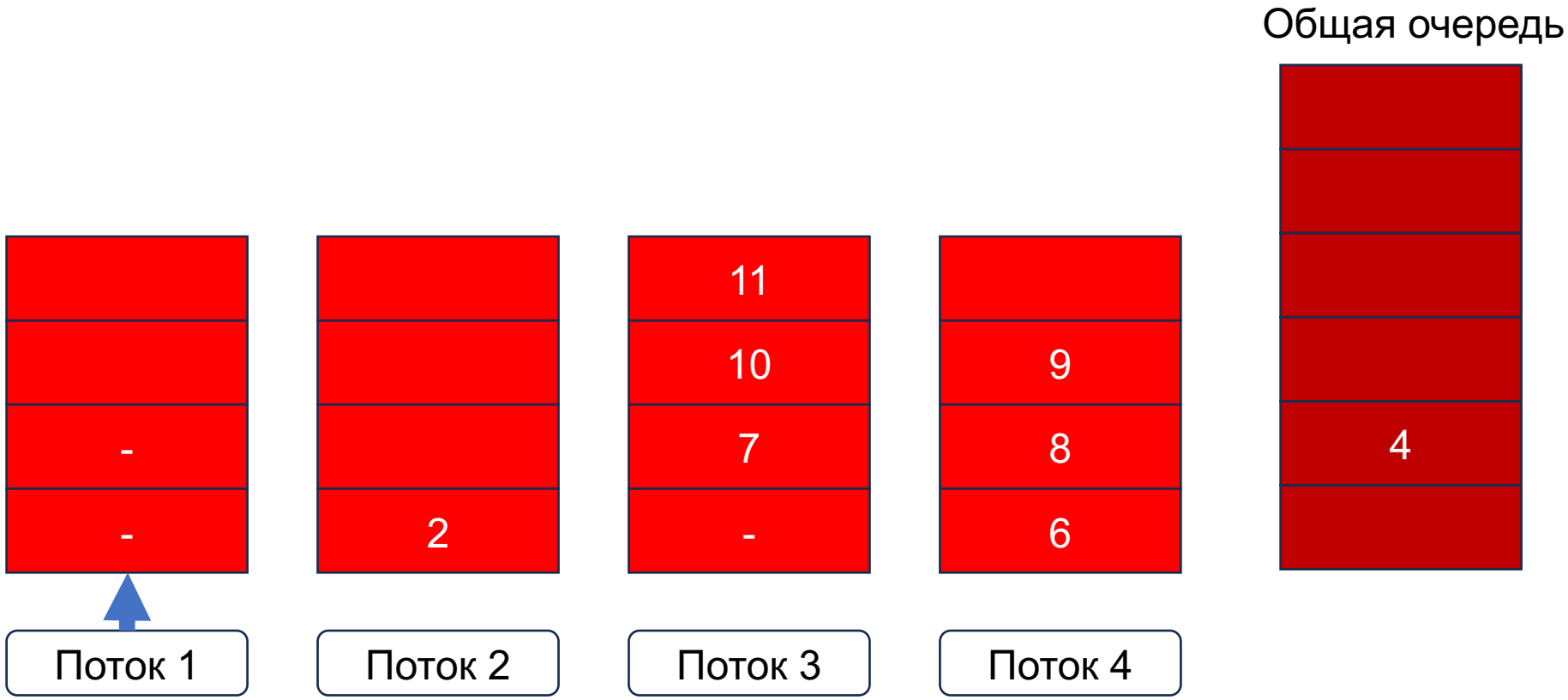
Короткие per-thread очереди



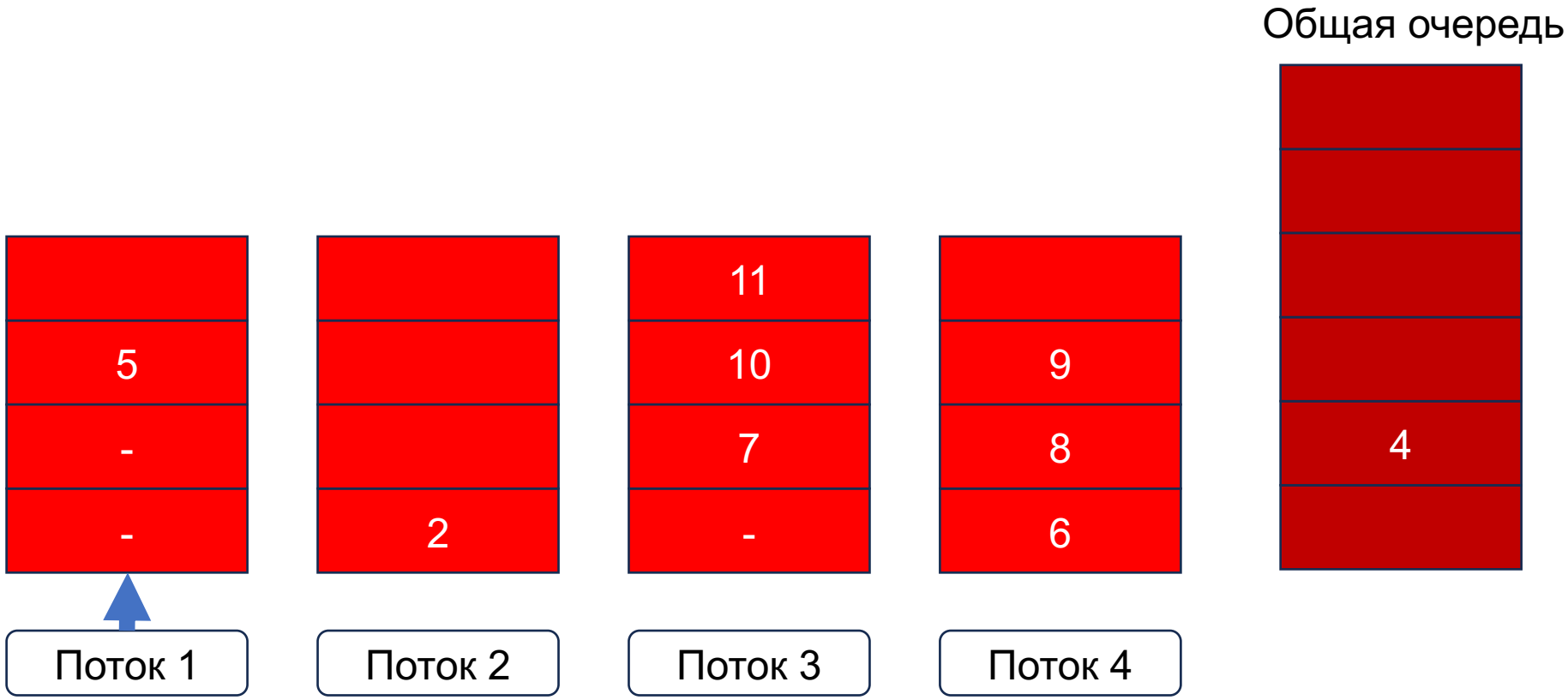
Короткие per-thread очереди



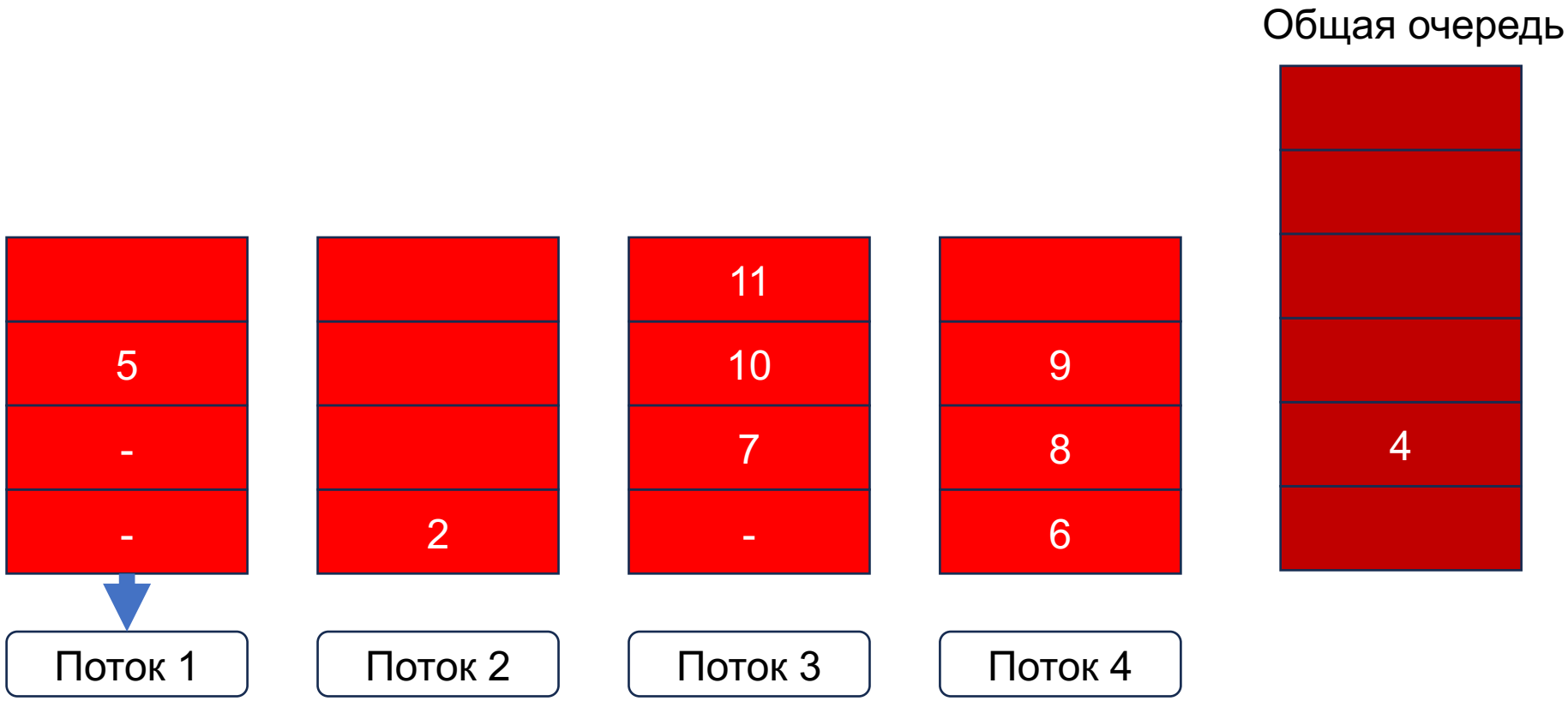
Короткие per-thread очереди



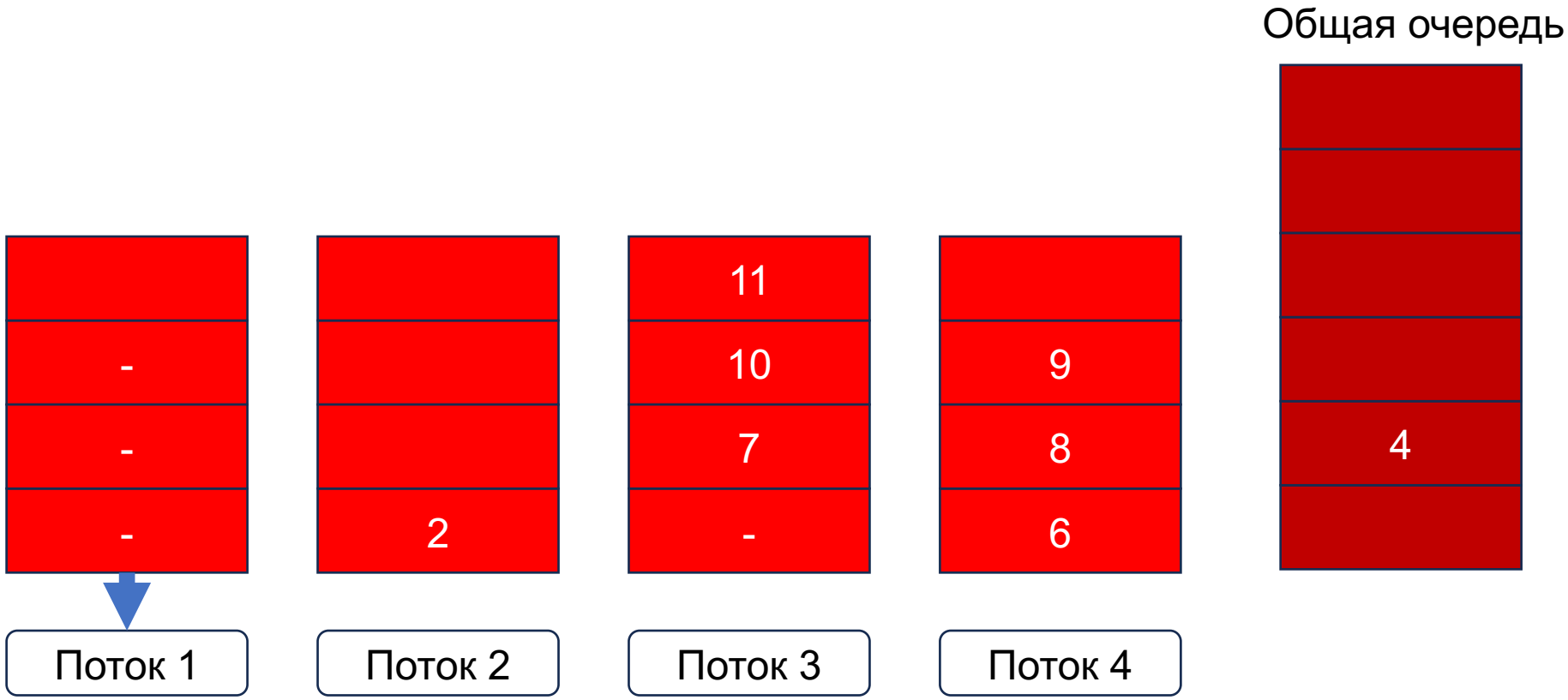
Короткие per-thread очереди



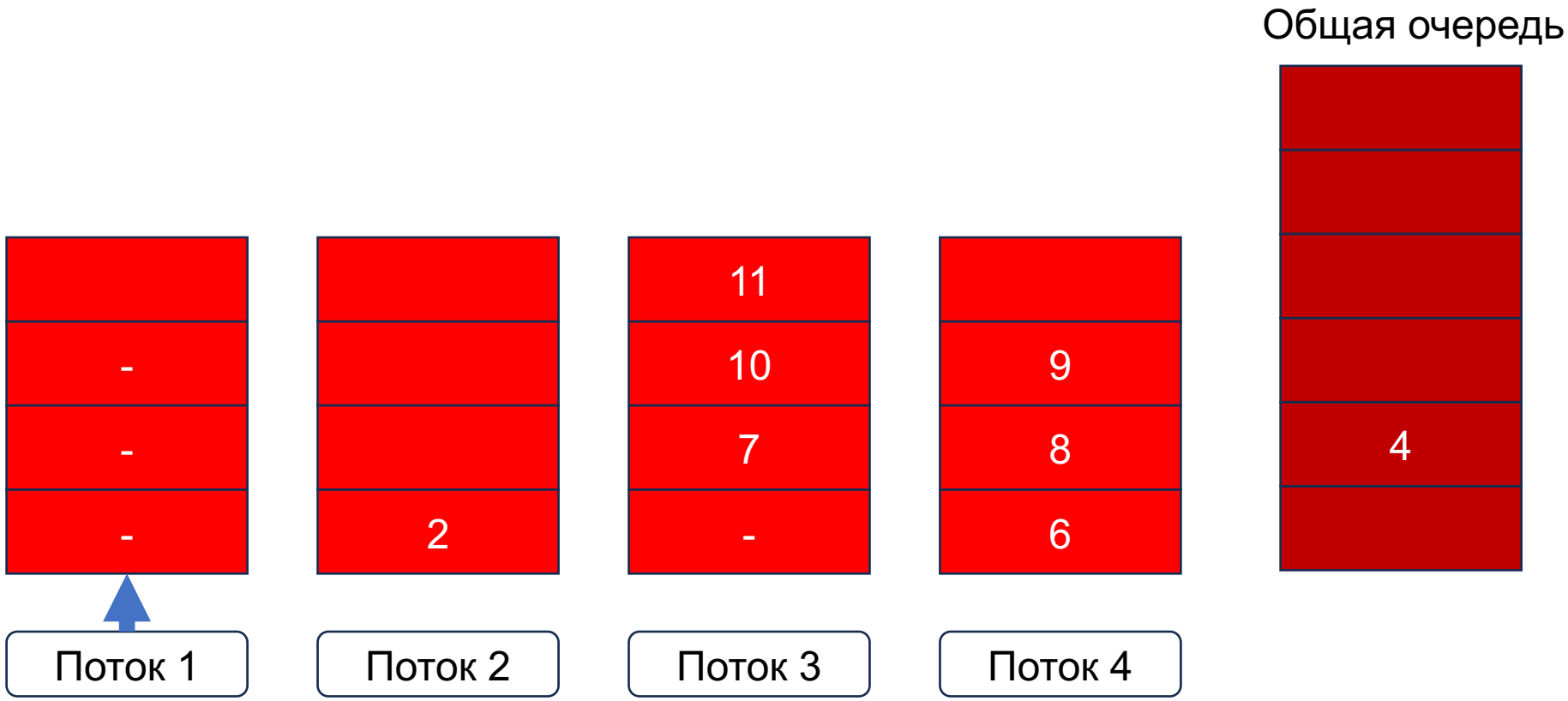
Короткие per-thread очереди



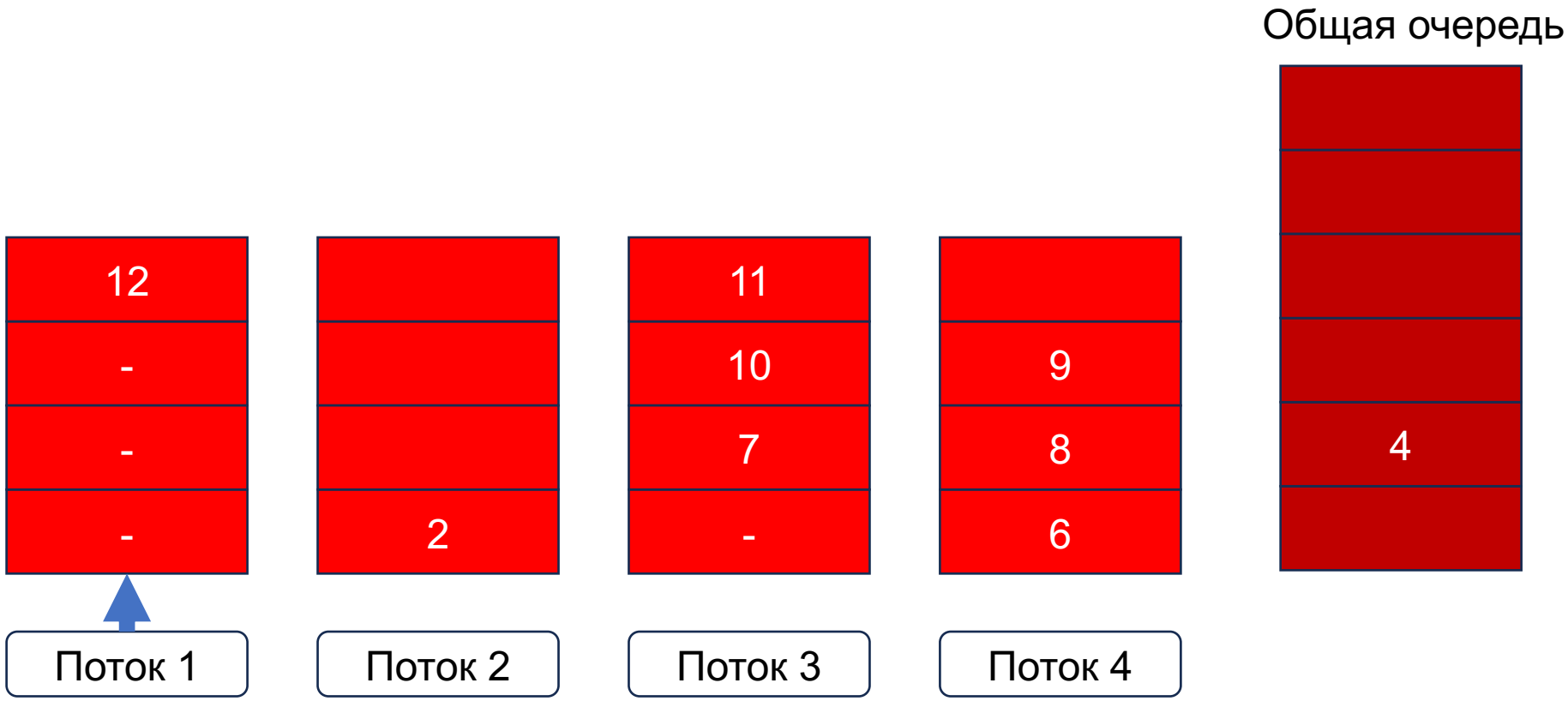
Короткие per-thread очереди



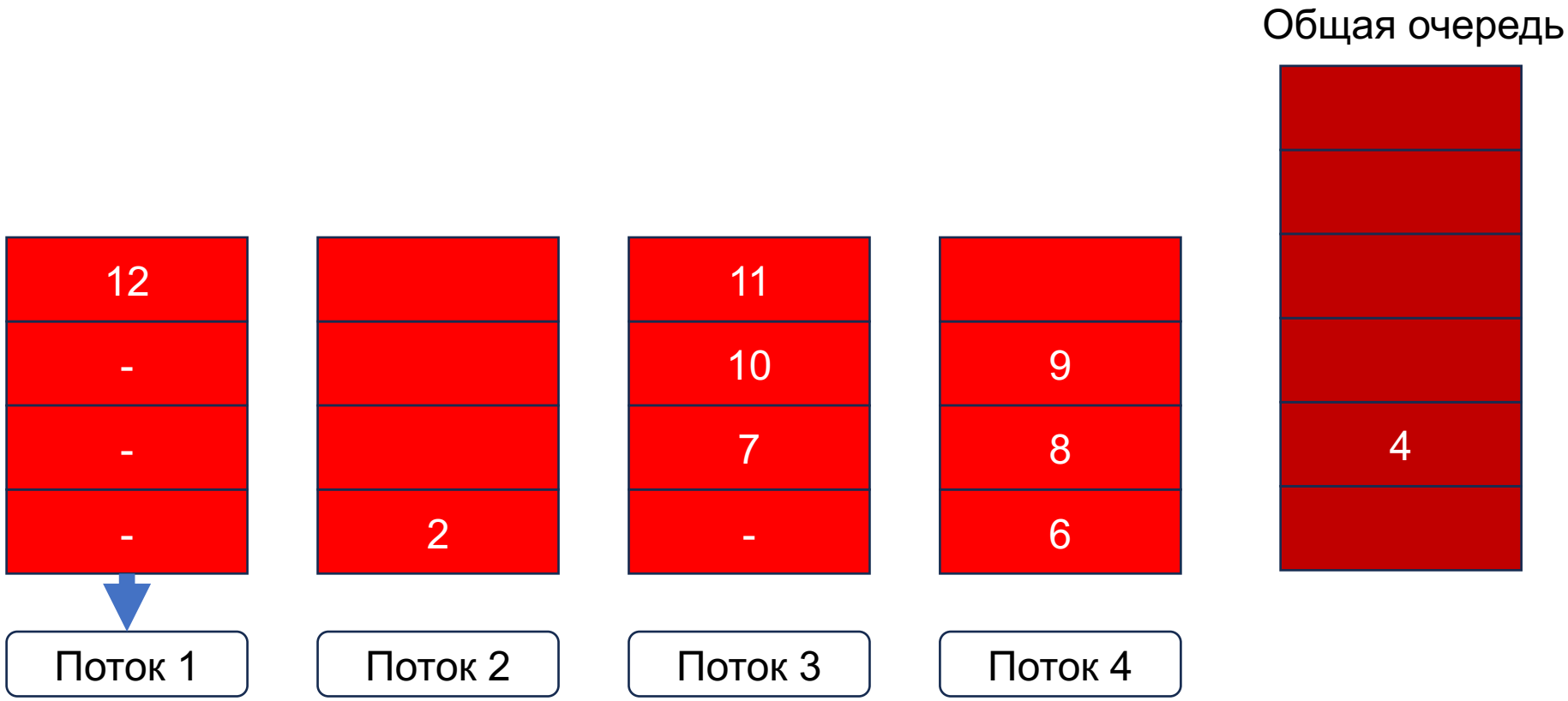
Короткие per-thread очереди



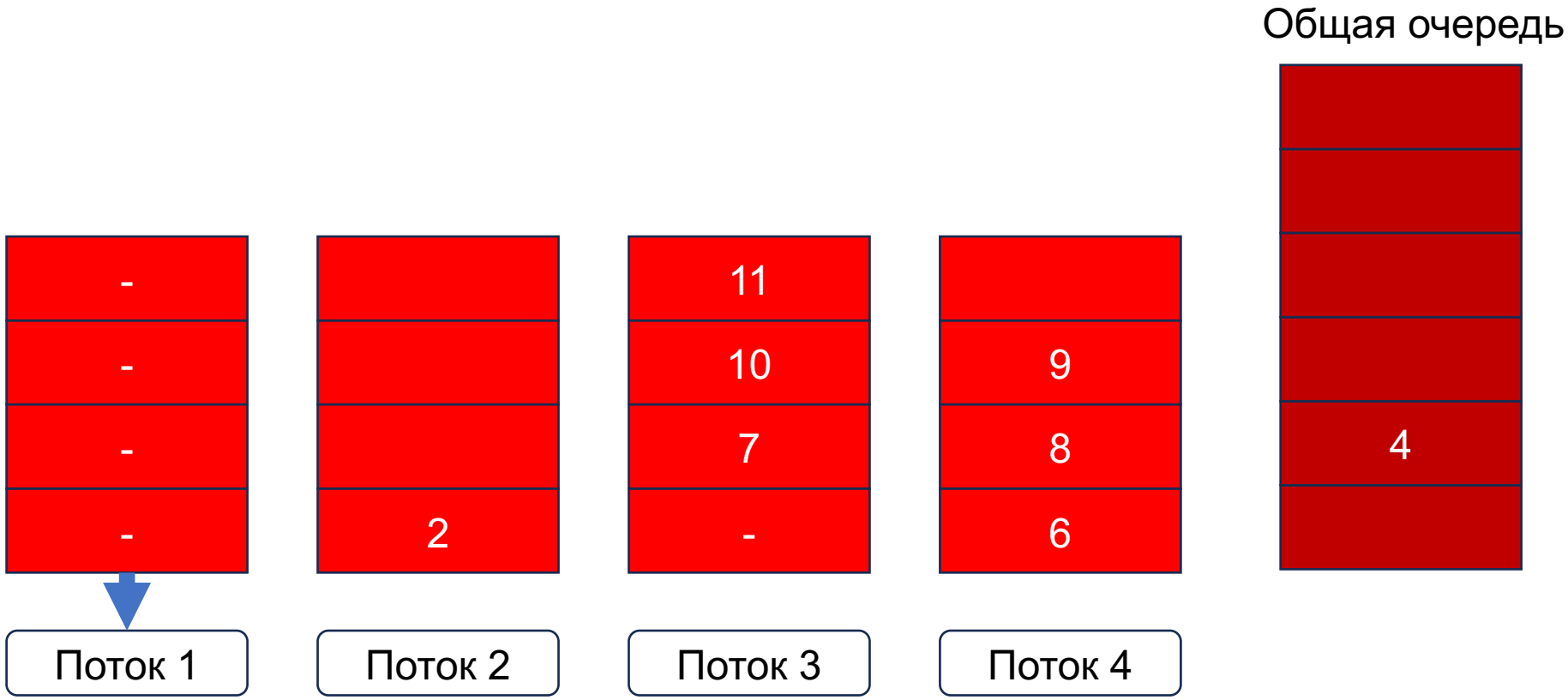
Короткие per-thread очереди



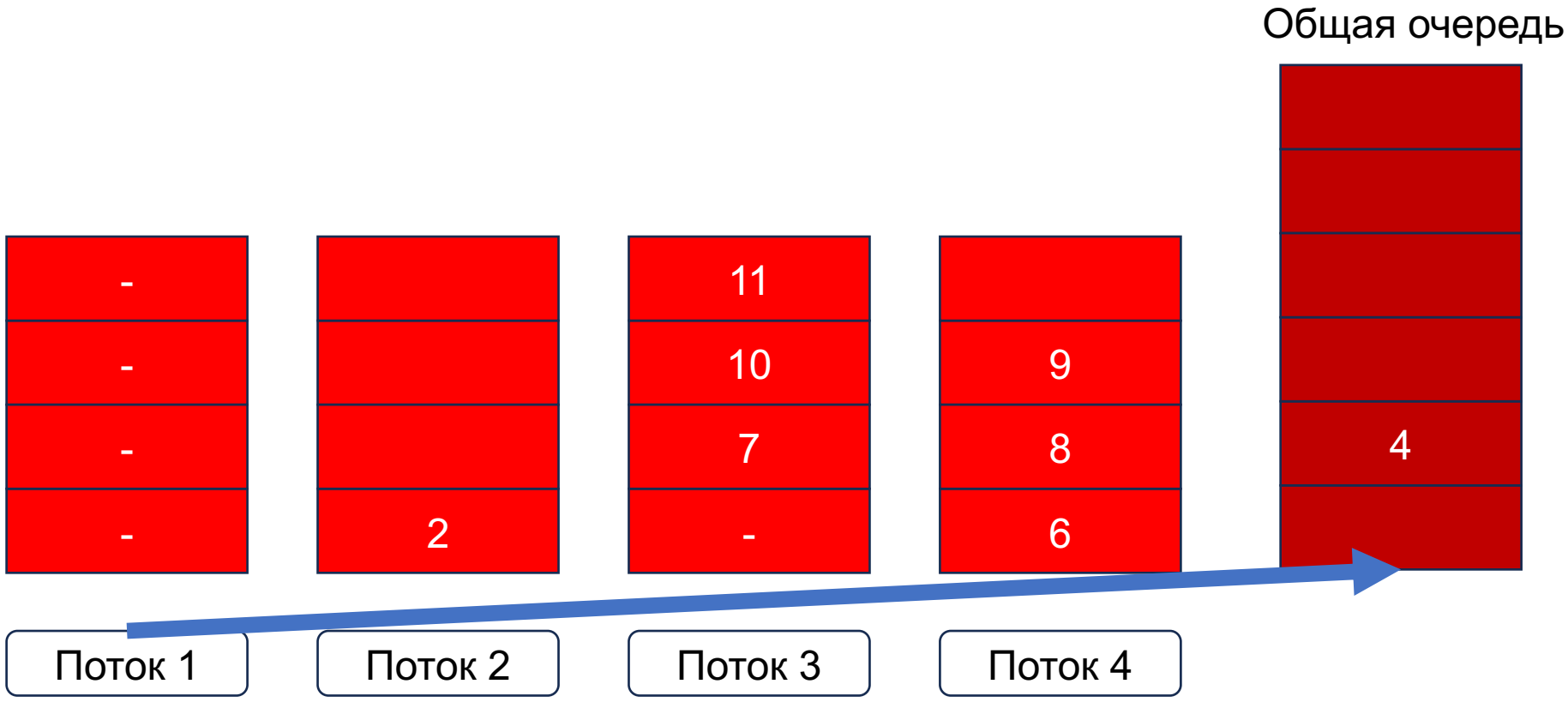
Короткие per-thread очереди



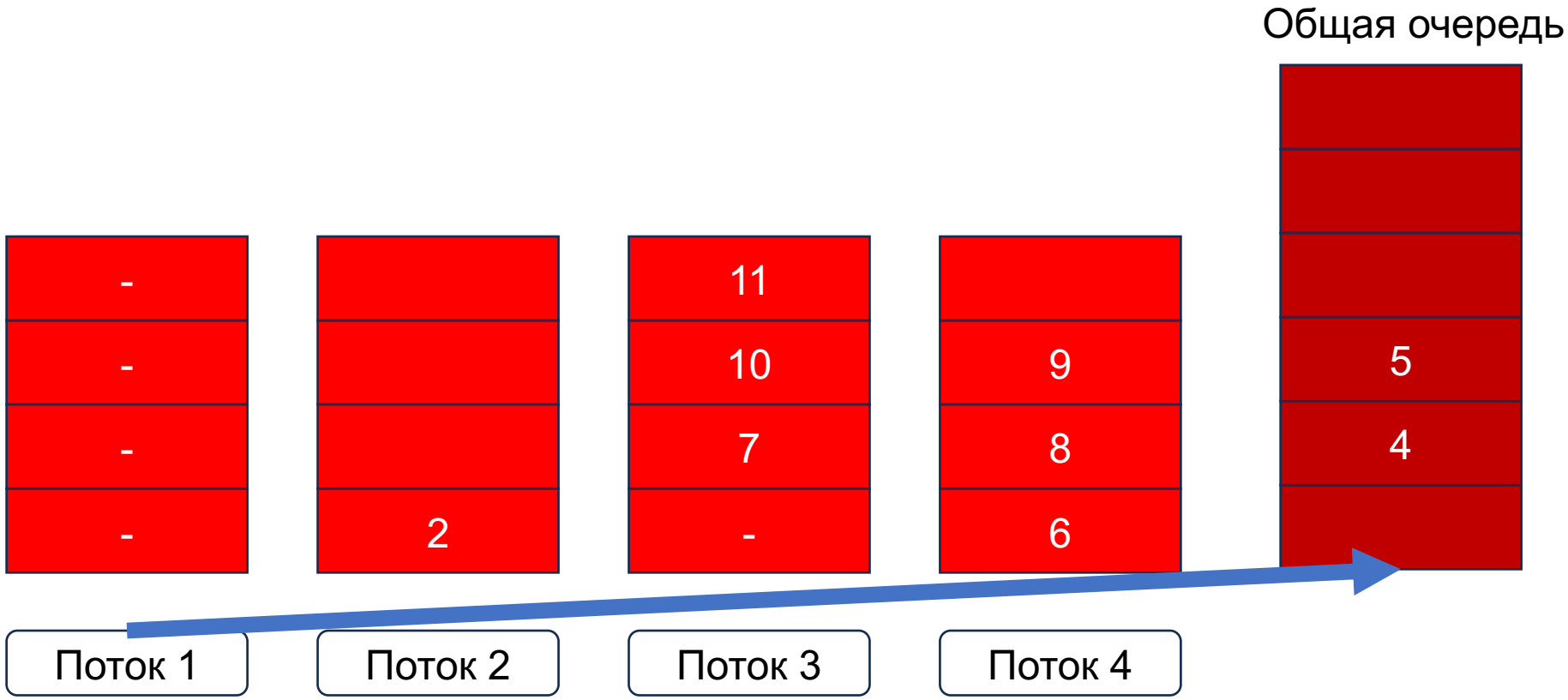
Короткие per-thread очереди



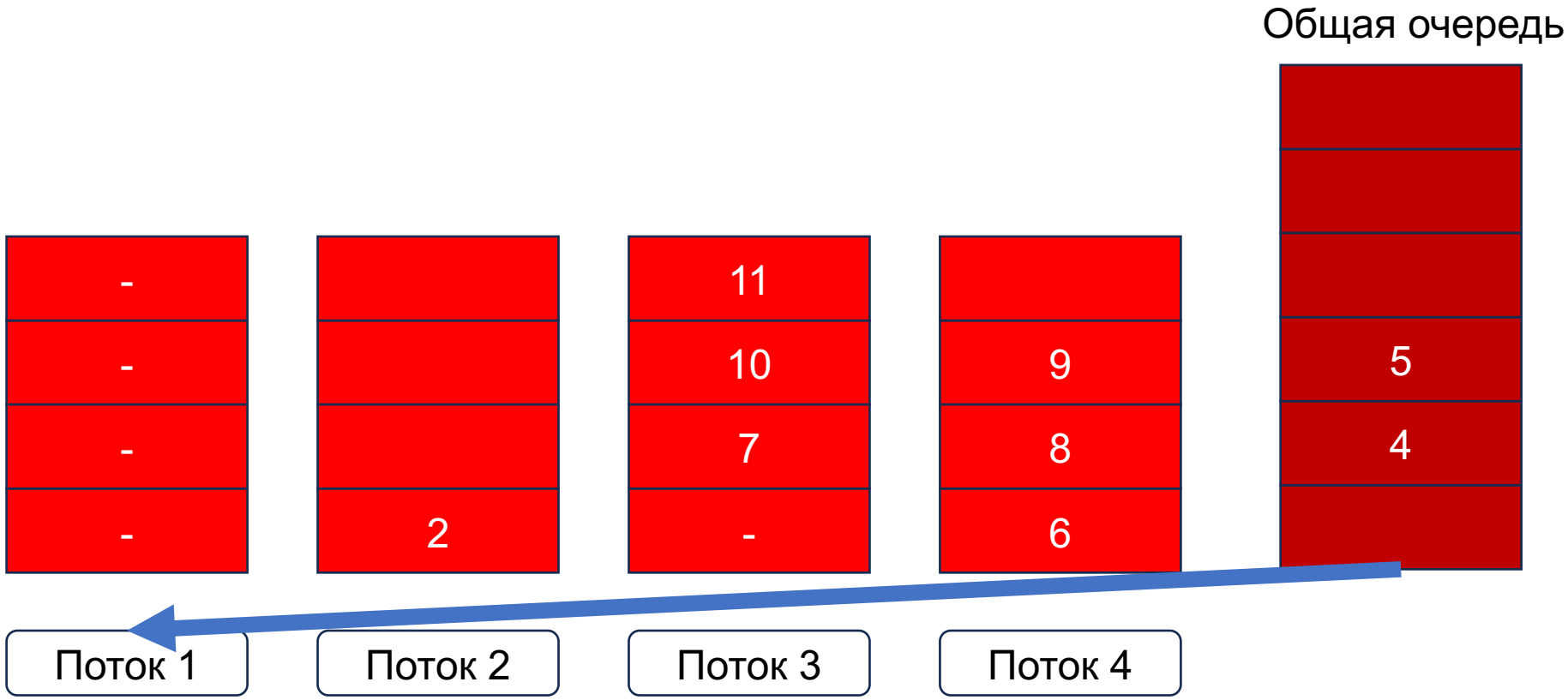
Короткие per-thread очереди



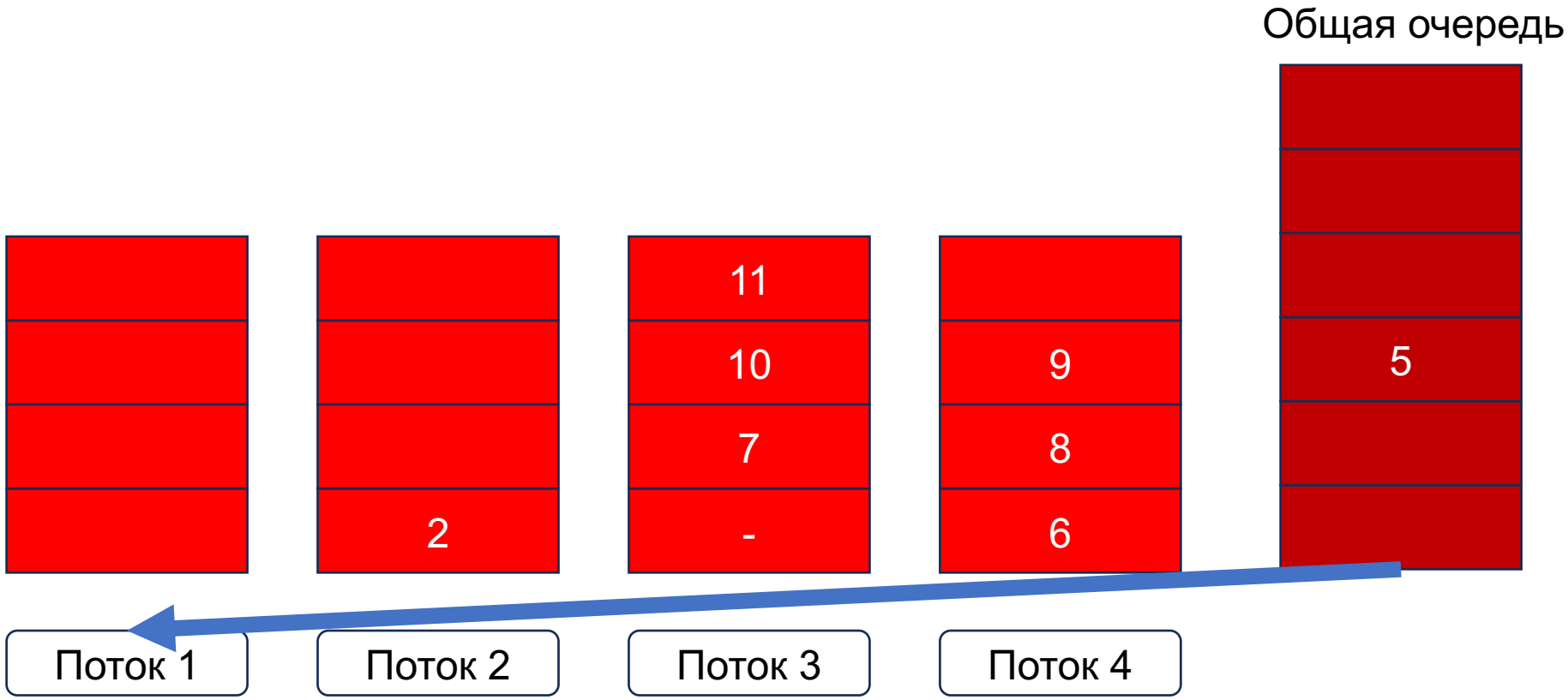
Короткие per-thread очереди



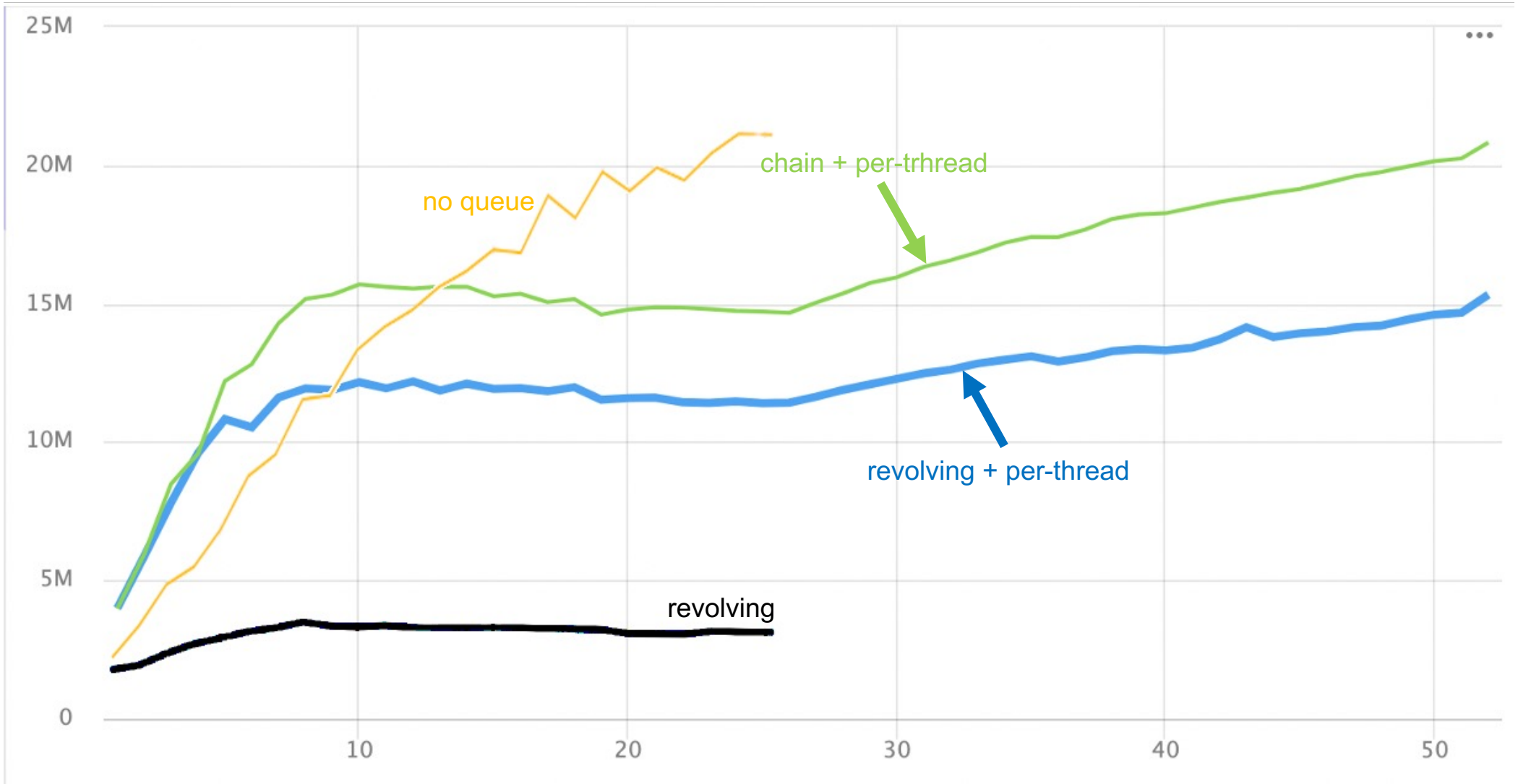
Короткие per-thread очереди



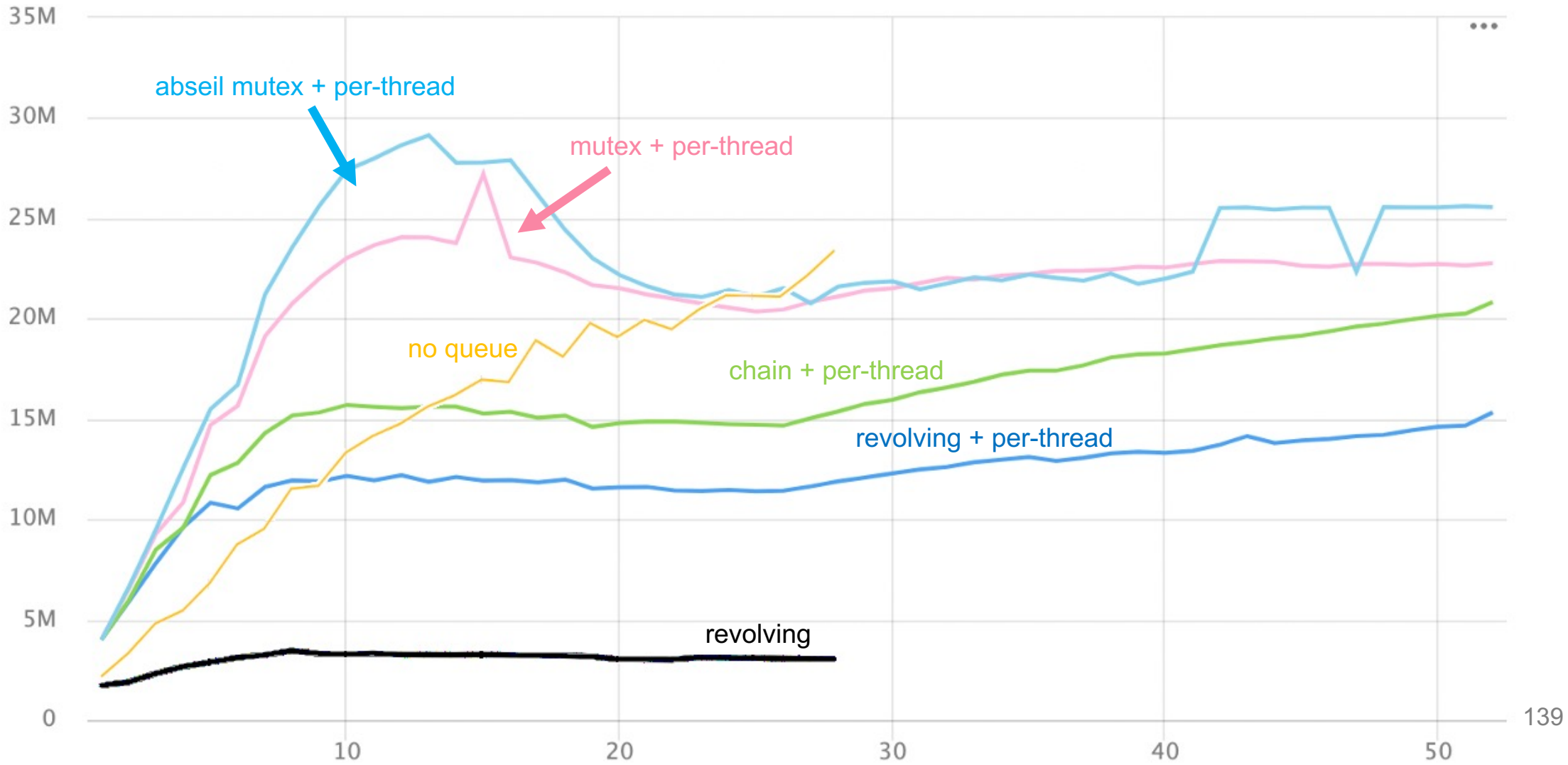
Короткие per-thread очереди



Короткие per-thread очереди

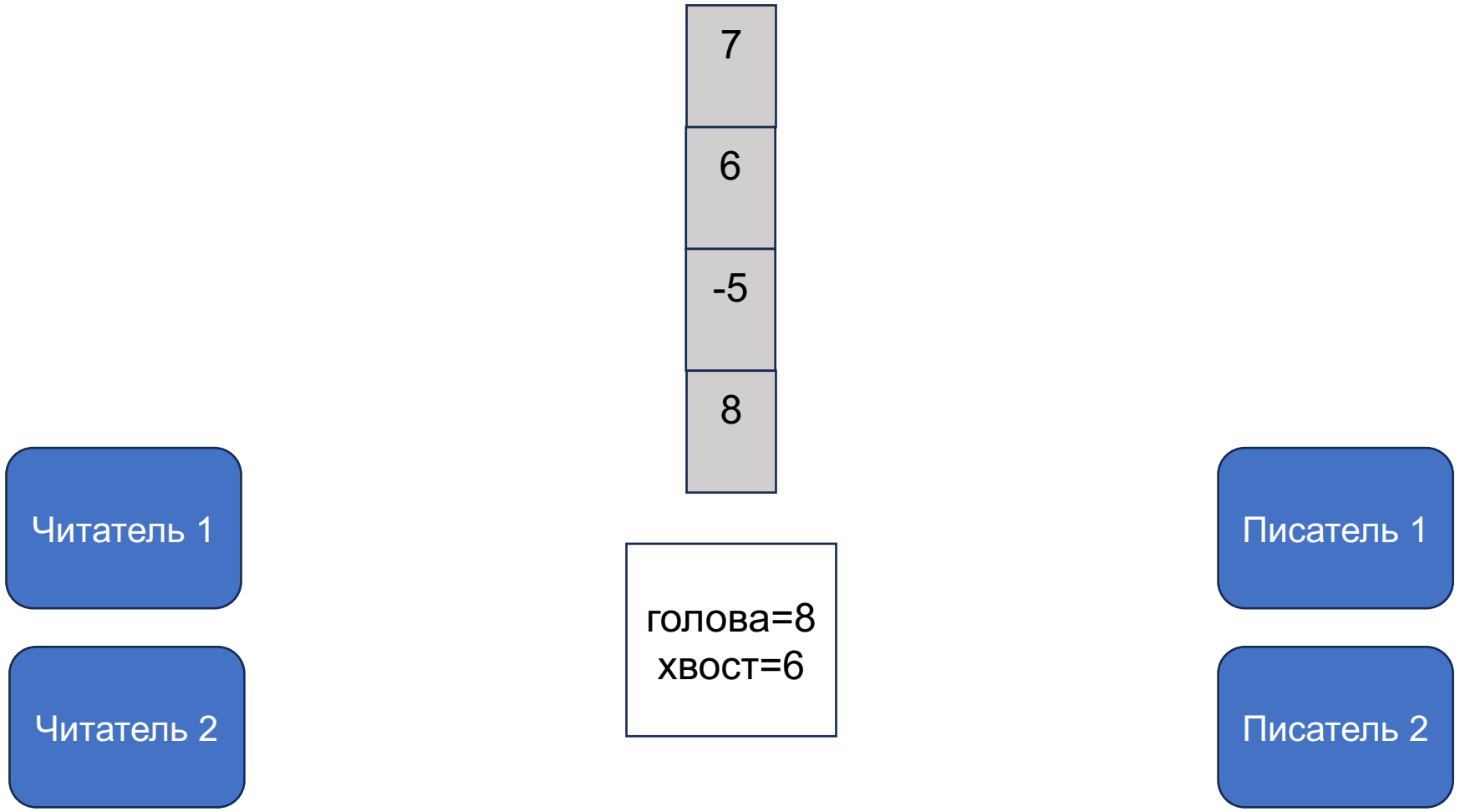


Короткие per-thread очереди

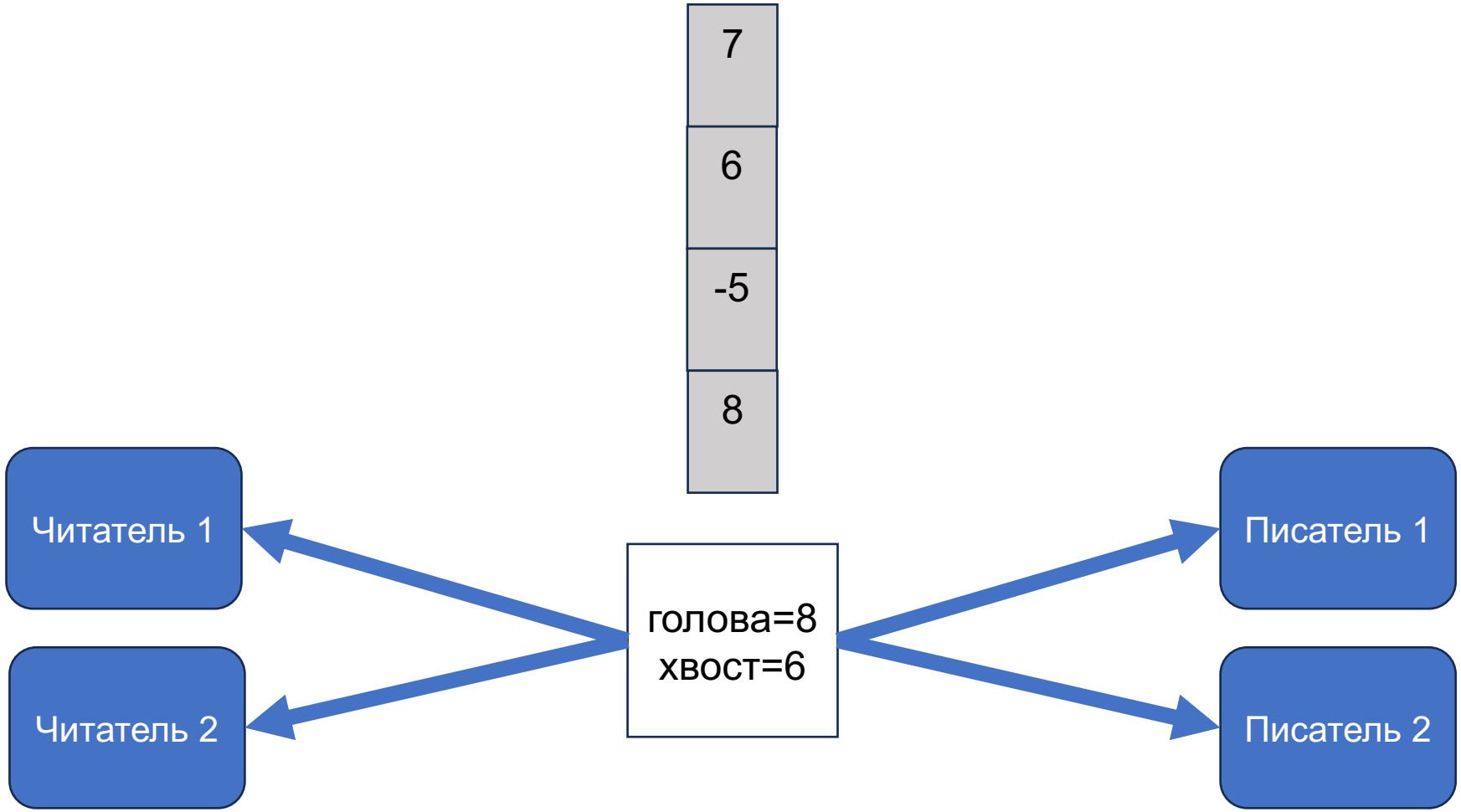


Оцениваем «скорость света» по циклическому буферу

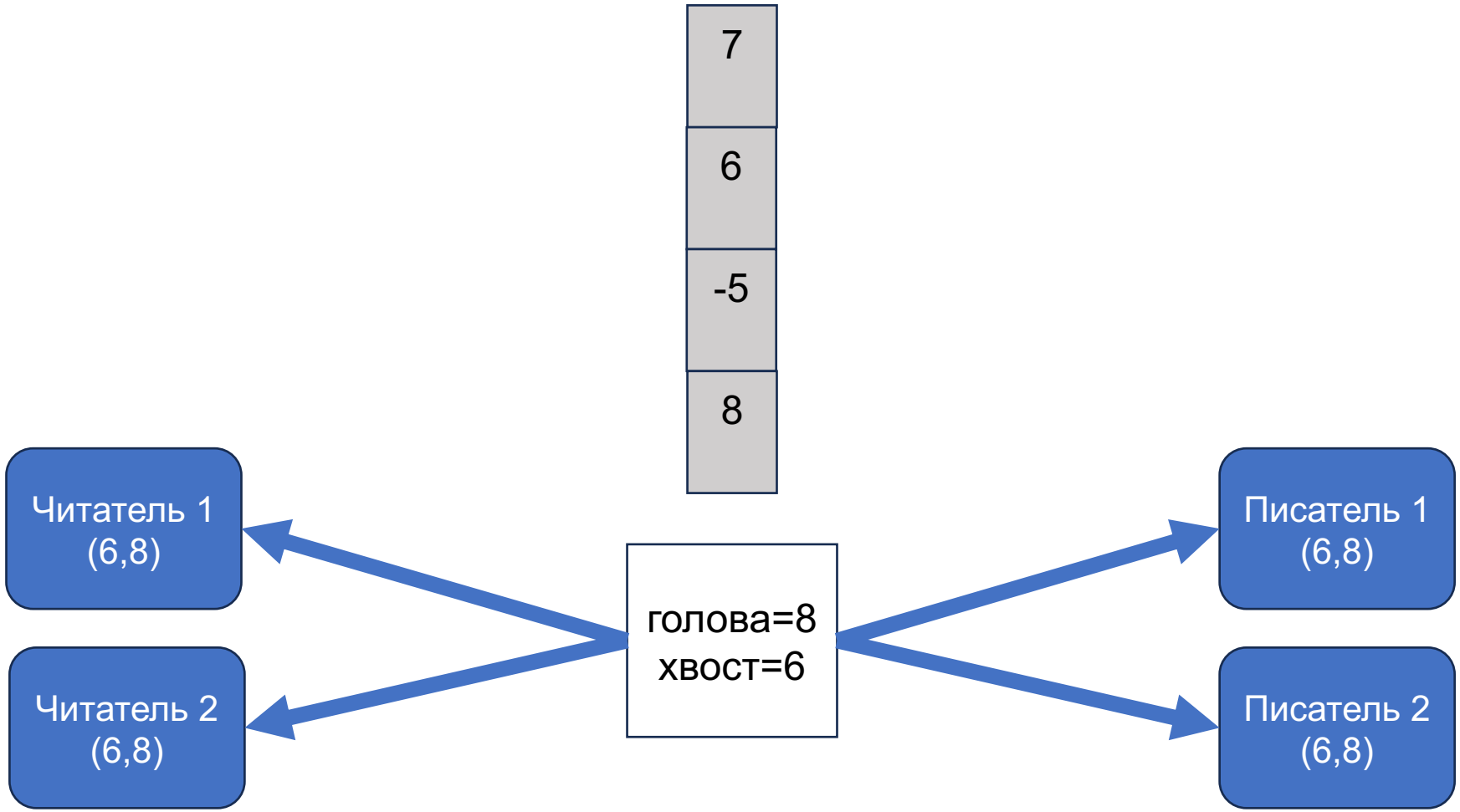
Очередь с циклическим буфером



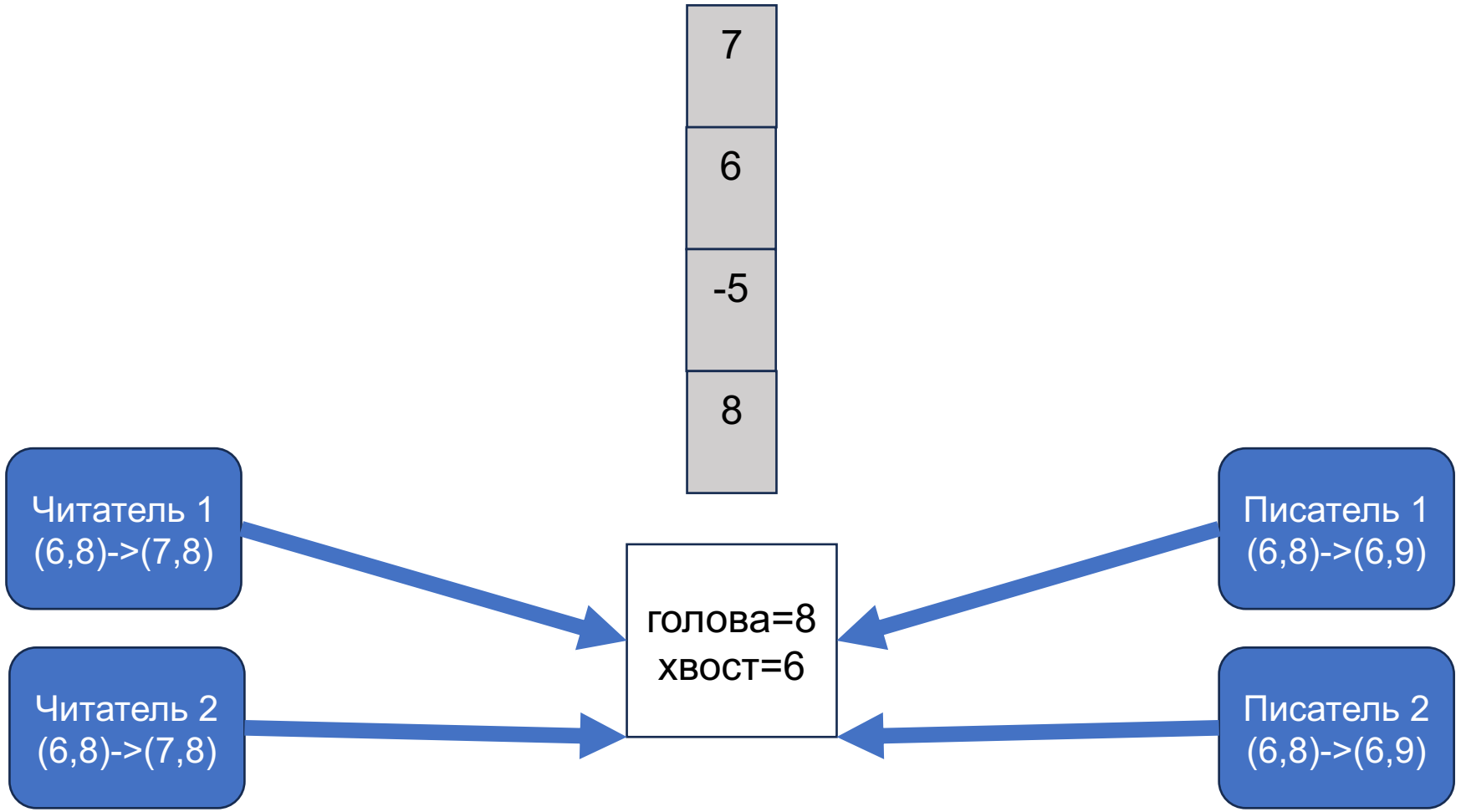
Очередь с циклическим буфером



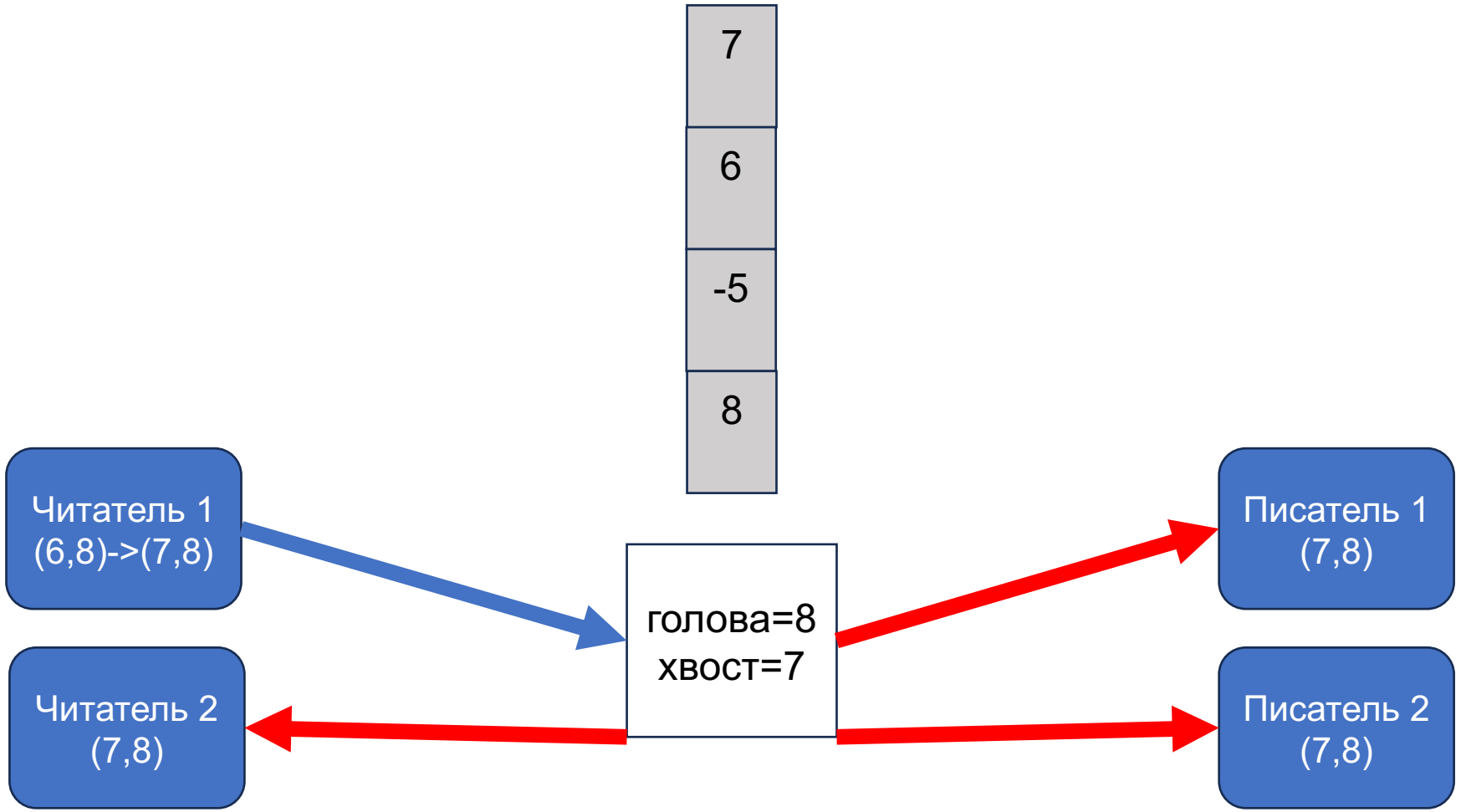
Очередь с циклическим буфером



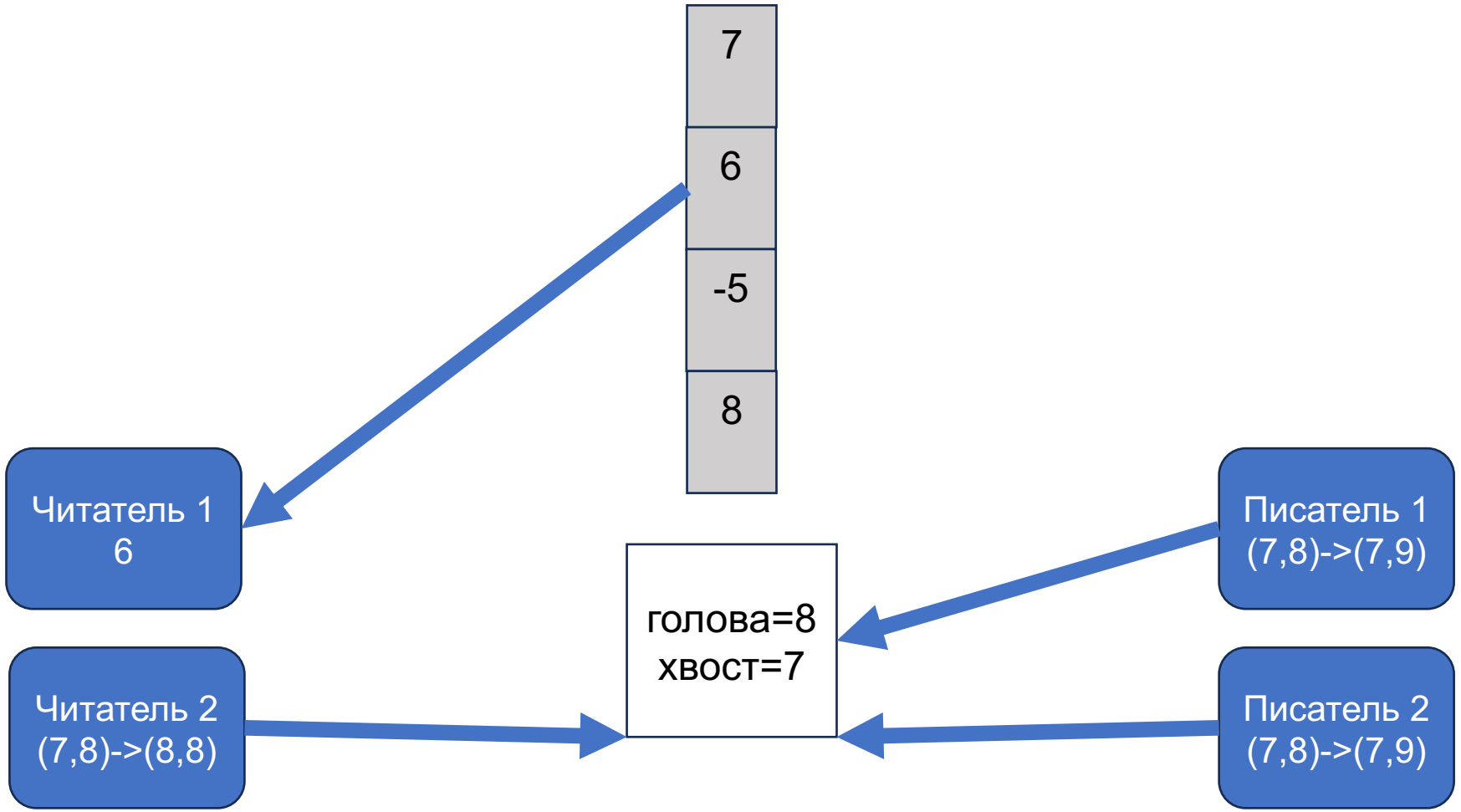
Очередь с циклическим буфером



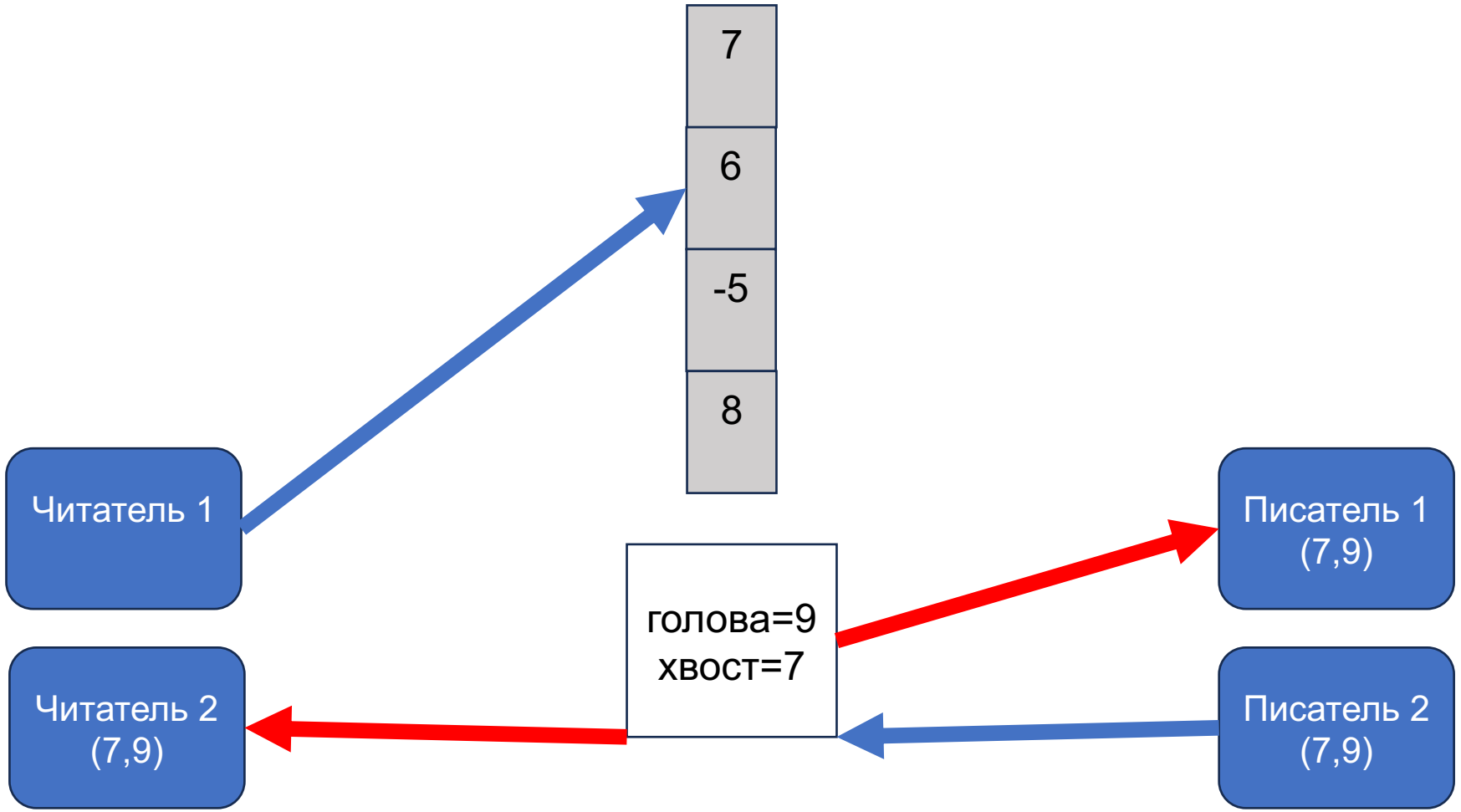
Очередь с циклическим буфером



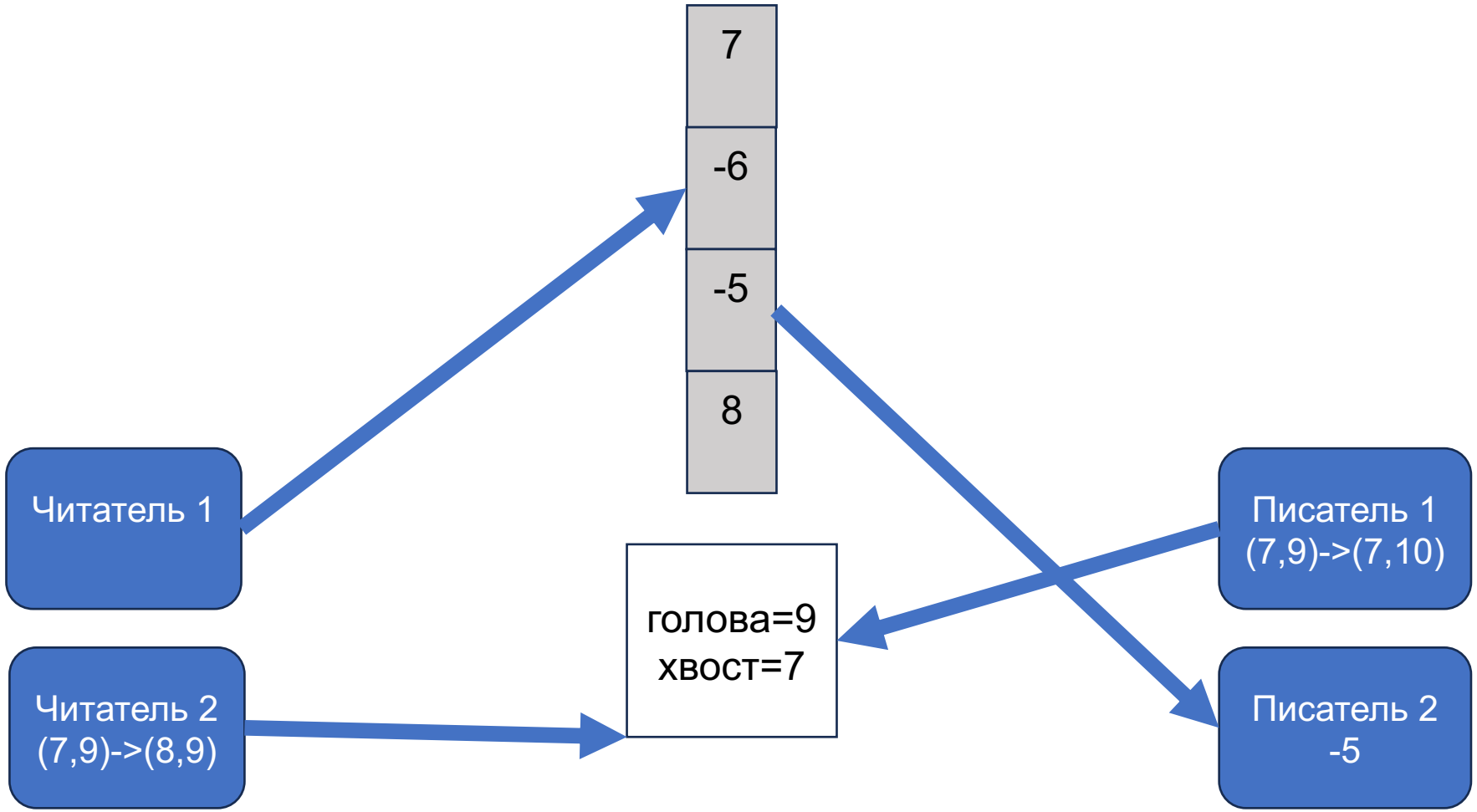
Очередь с циклическим буфером



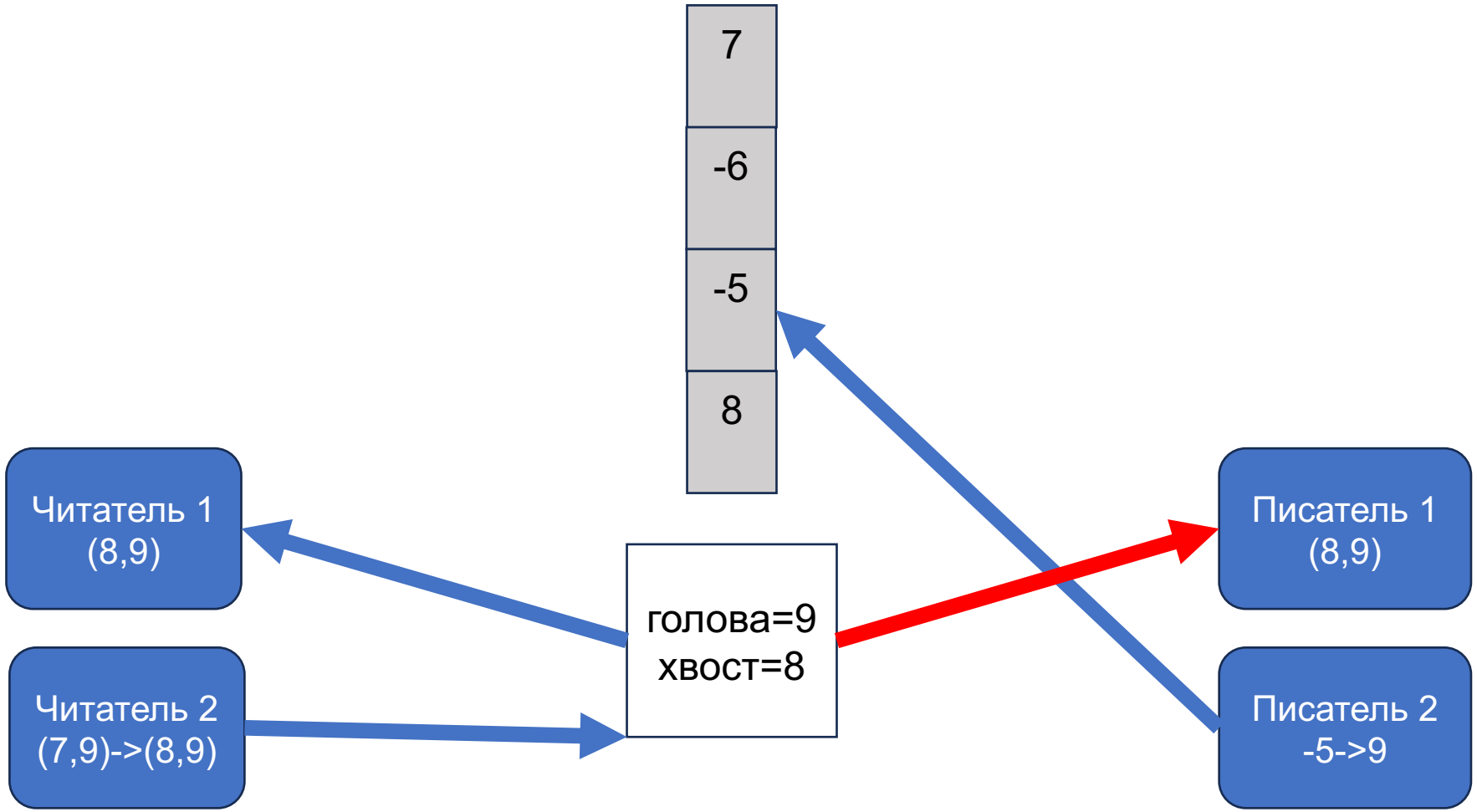
Очередь с циклическим буфером



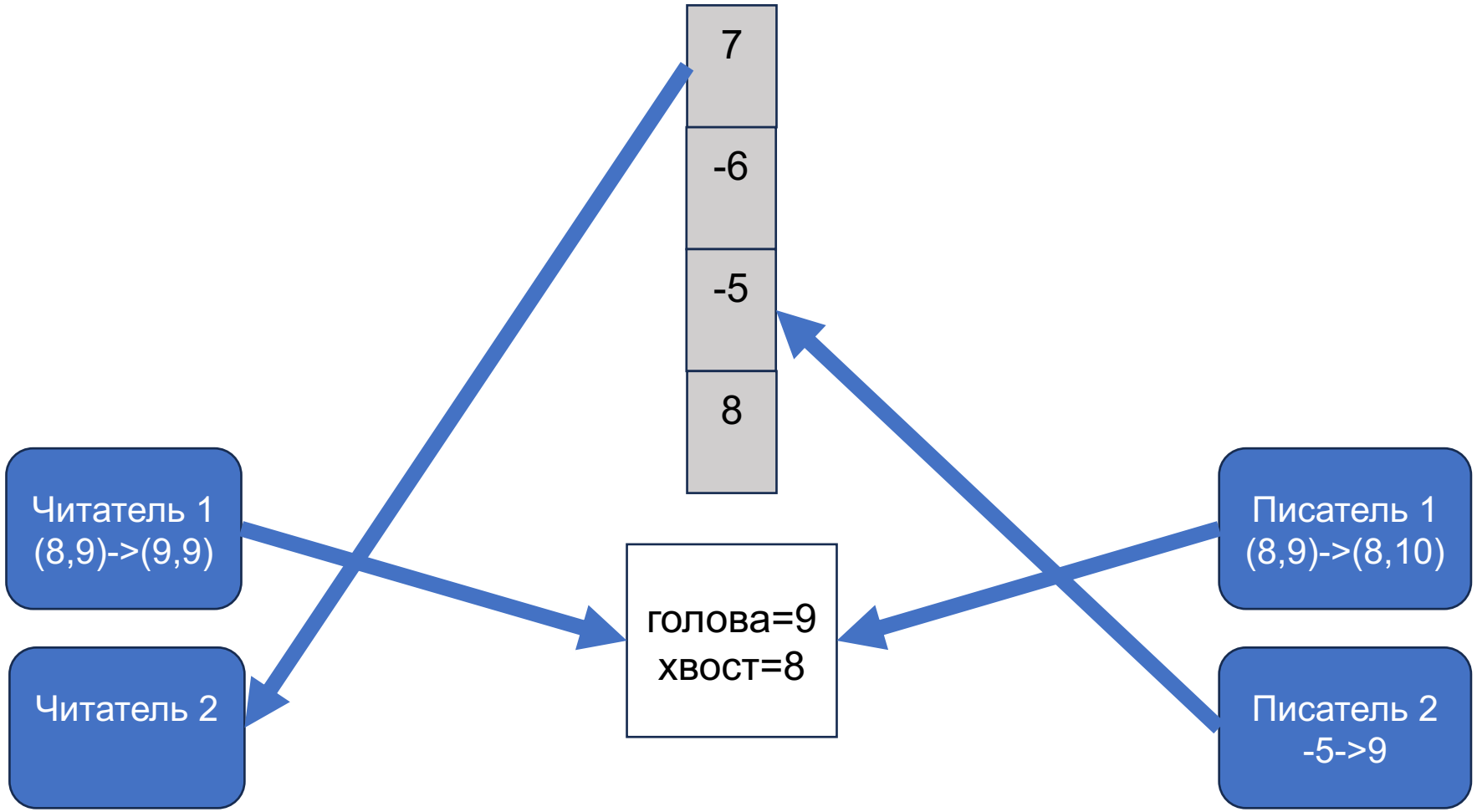
Очередь с циклическим буфером



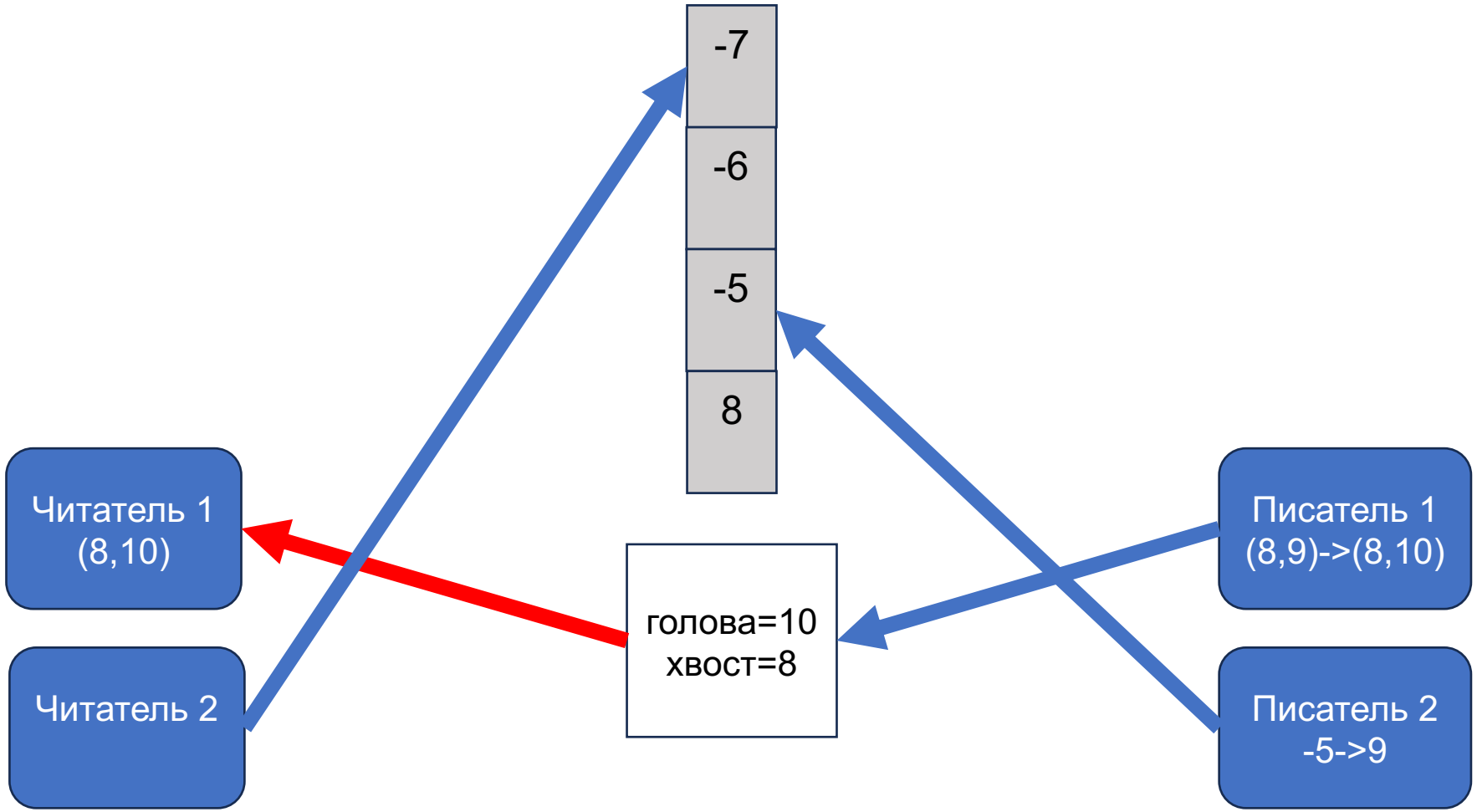
Очередь с циклическим буфером



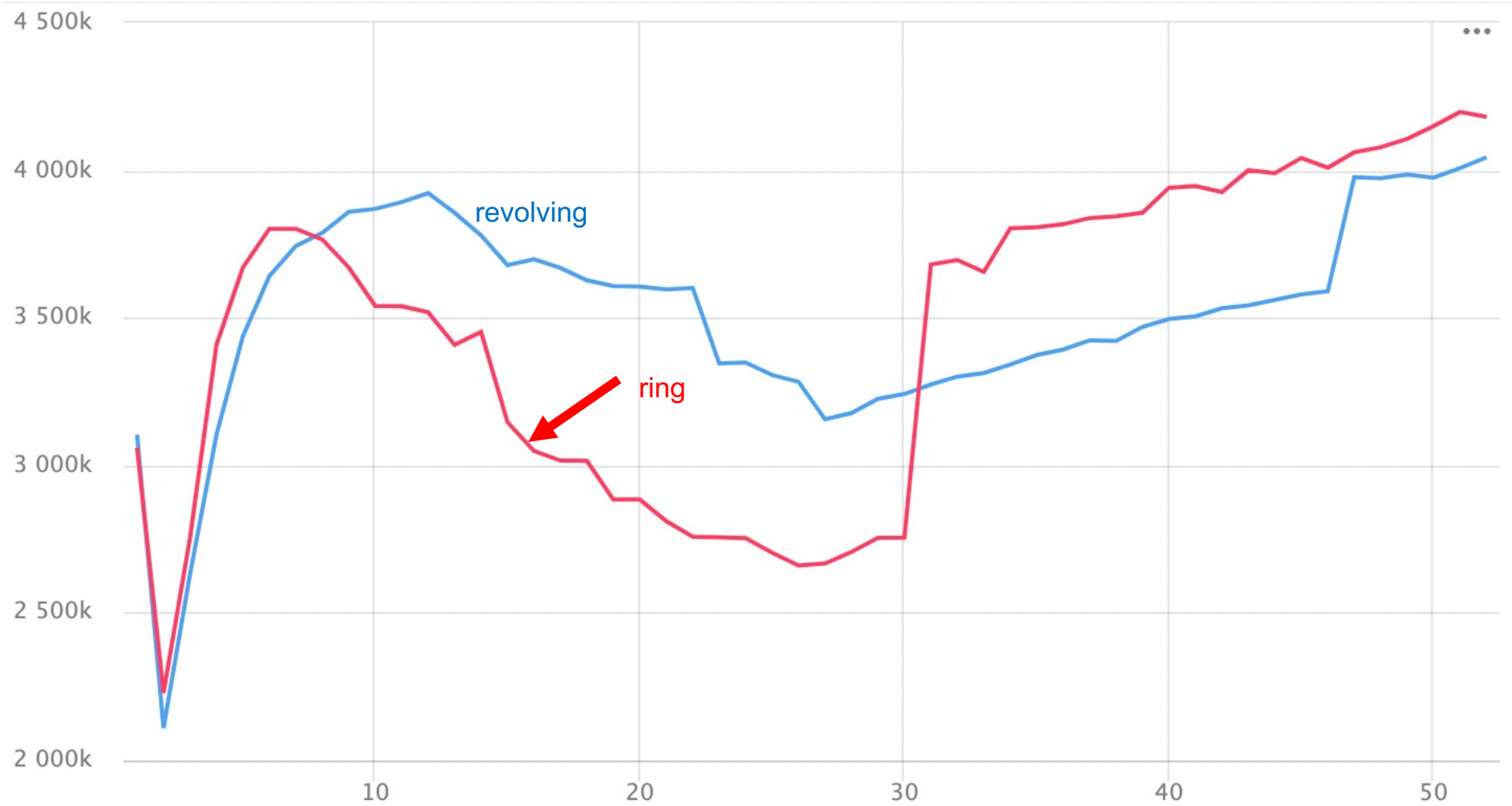
Очередь с циклическим буфером



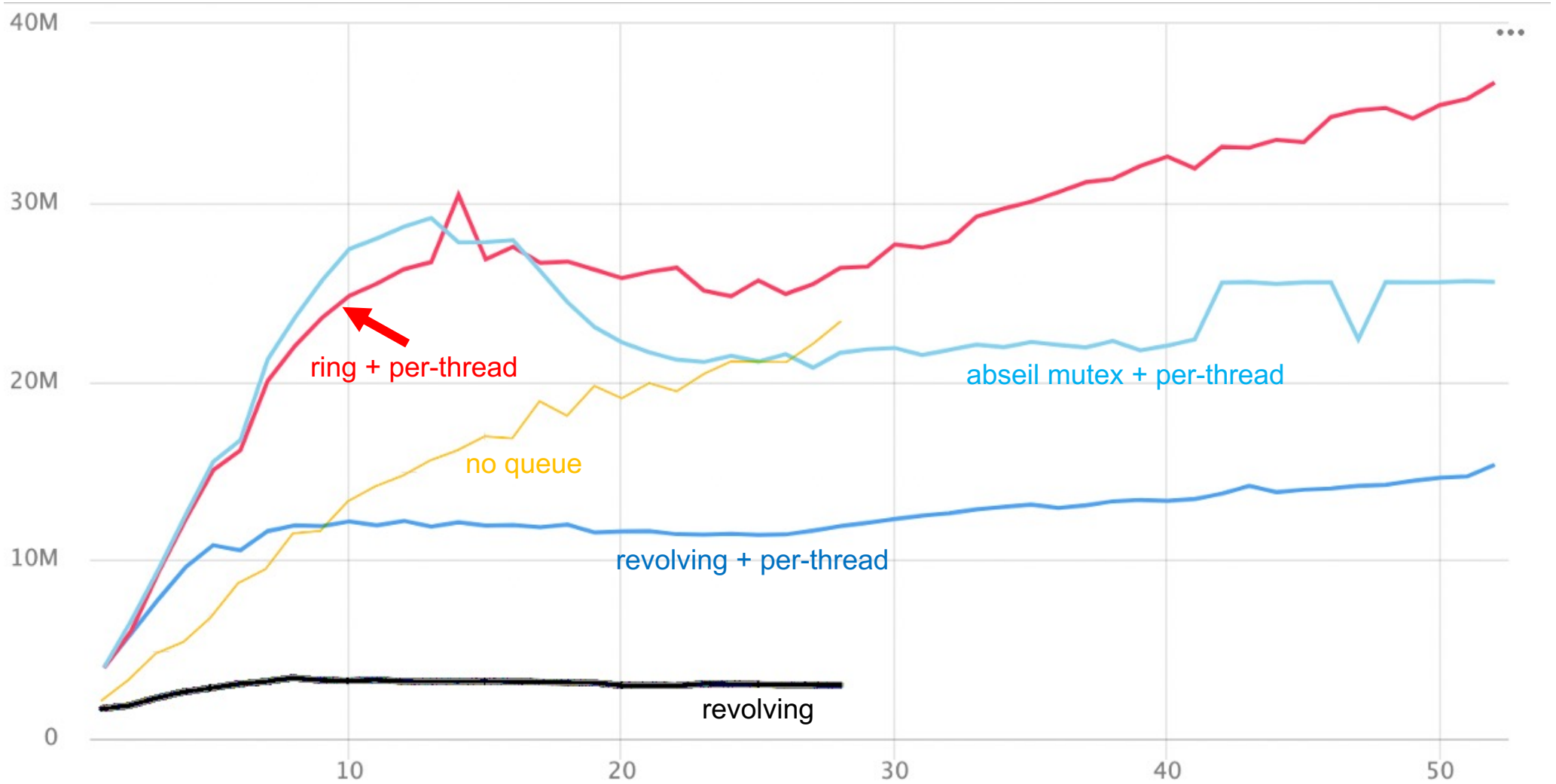
Очередь с циклическим буфером



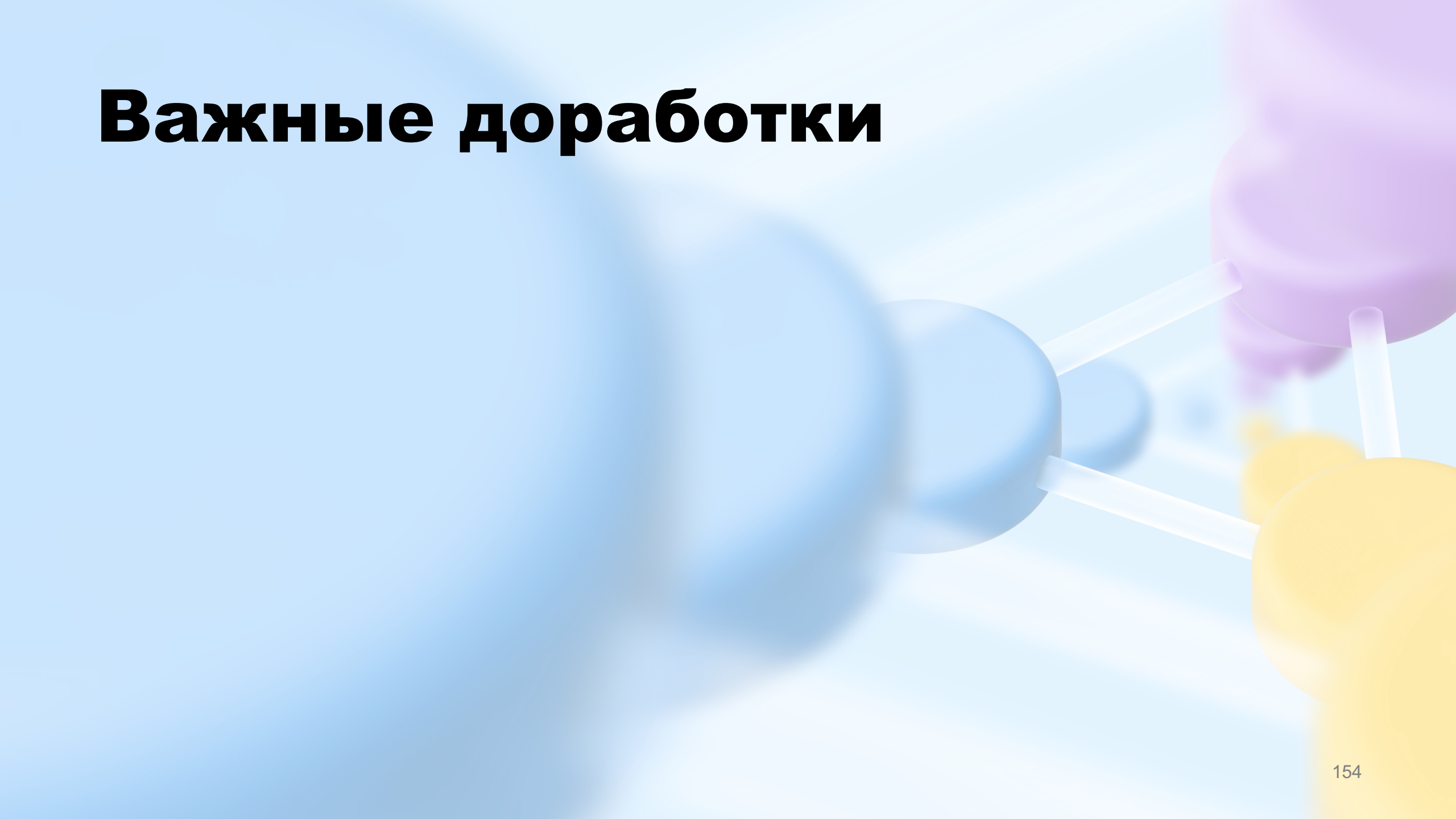
Очередь с циклическим буфером



Очередь с циклическим буфером + короткие per-thread очереди



Важные доработки

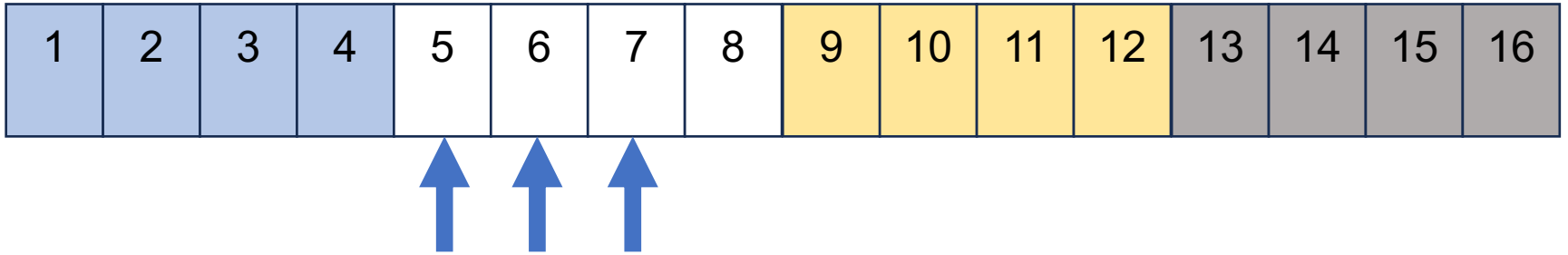


Важные доработки

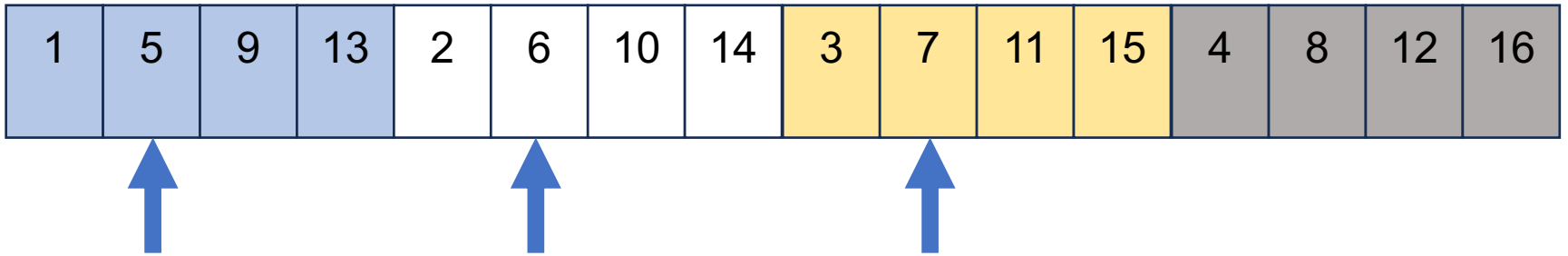
- Корректная обработка случаев, когда застрявший читатель/писатель отстал на круг и более
- Fallback на Revolving MPMC queue при переполнении

Раскладка по кэш-линиям

до:

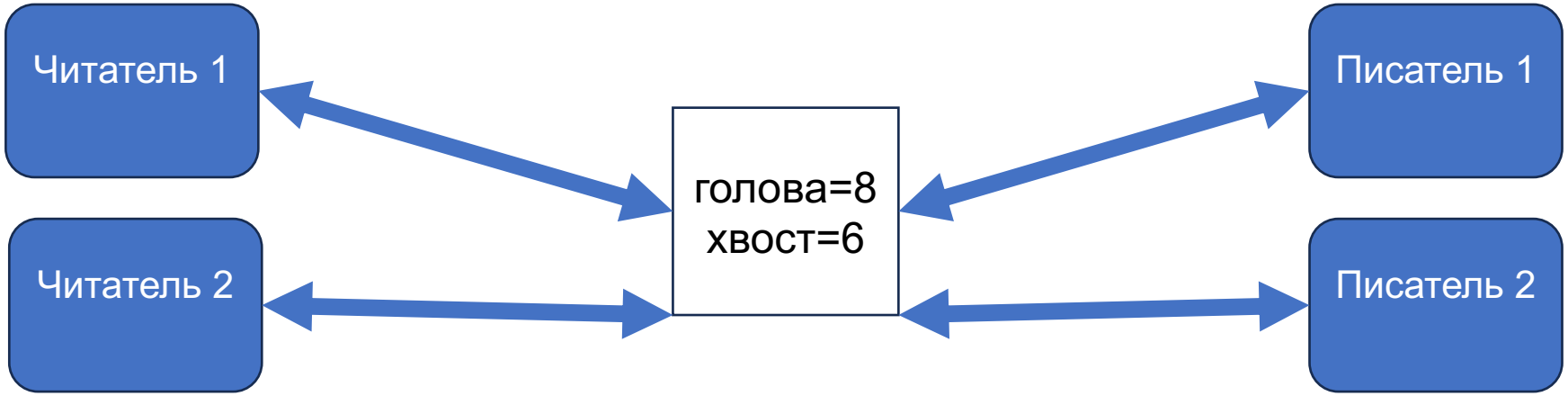


после:

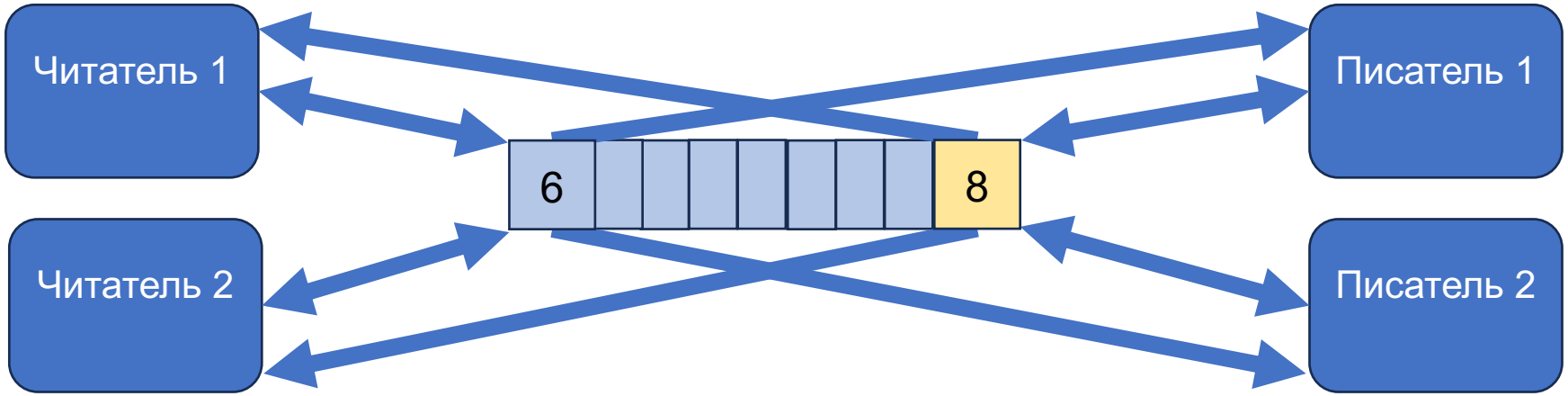


Горячие атомарные переменные

до:

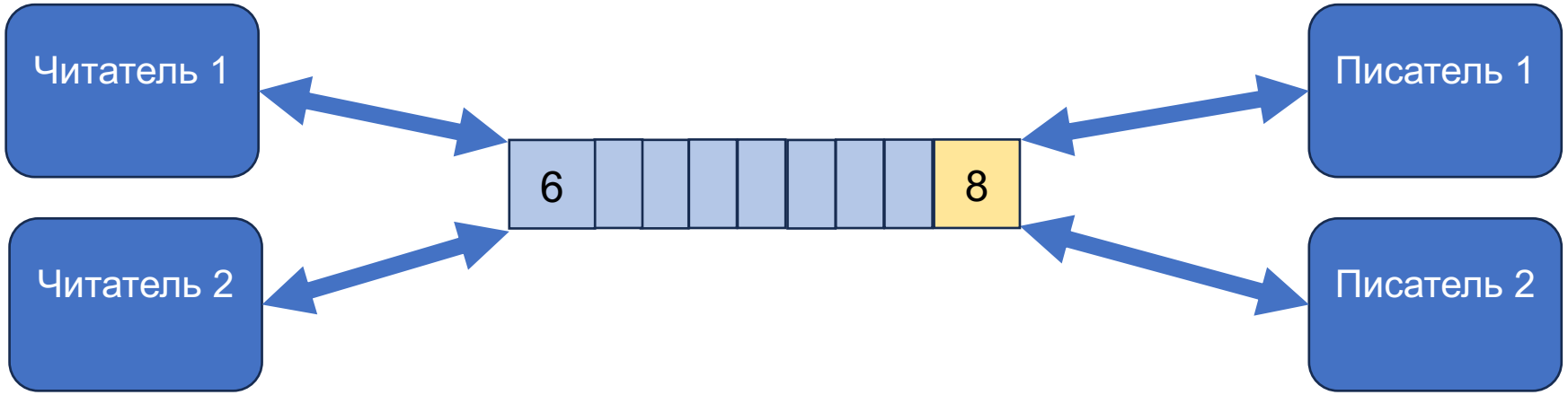


после:

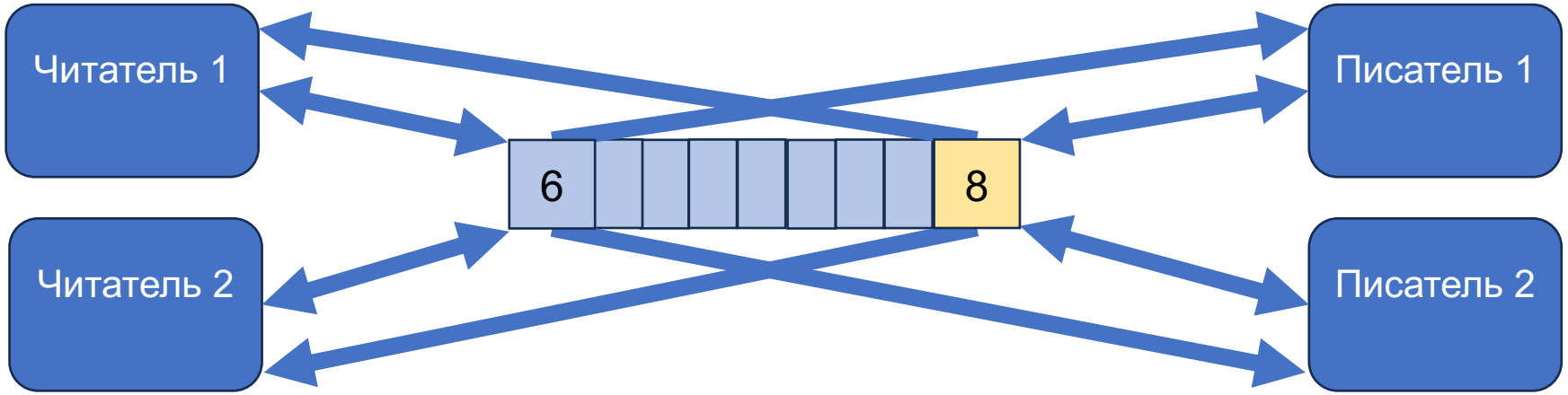


Быстро, но с гонкой, или медленно и без?

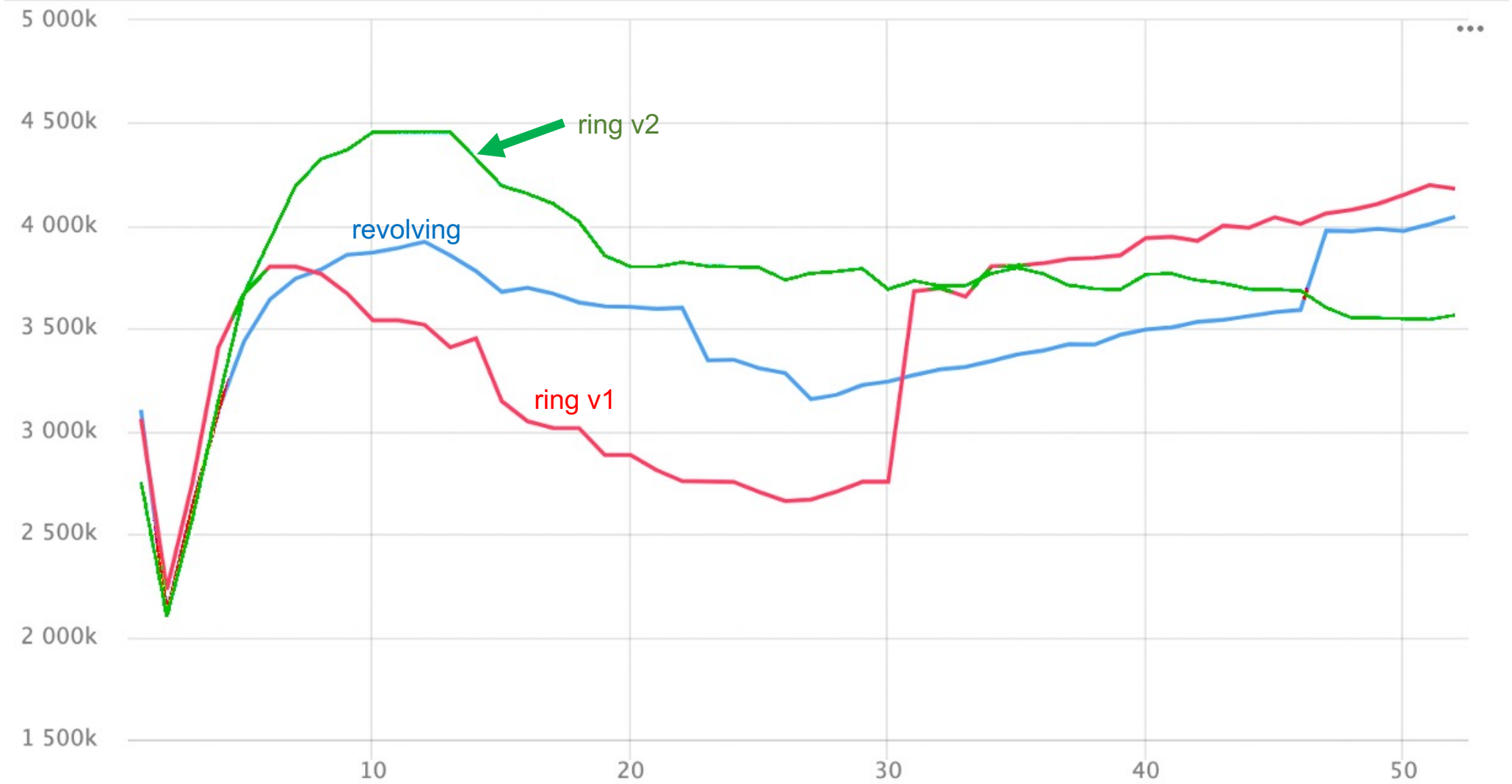
быстро:



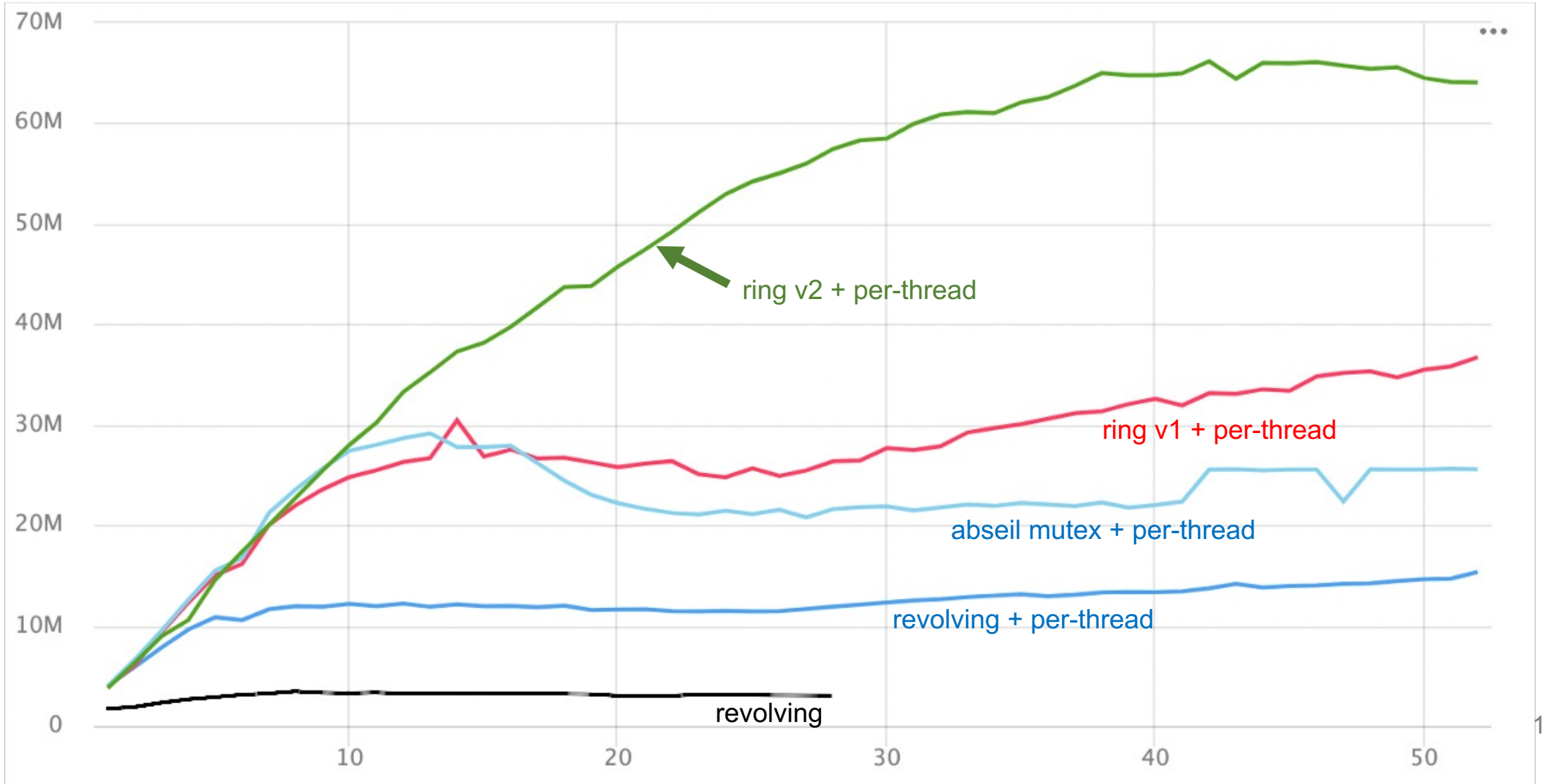
без гонки:



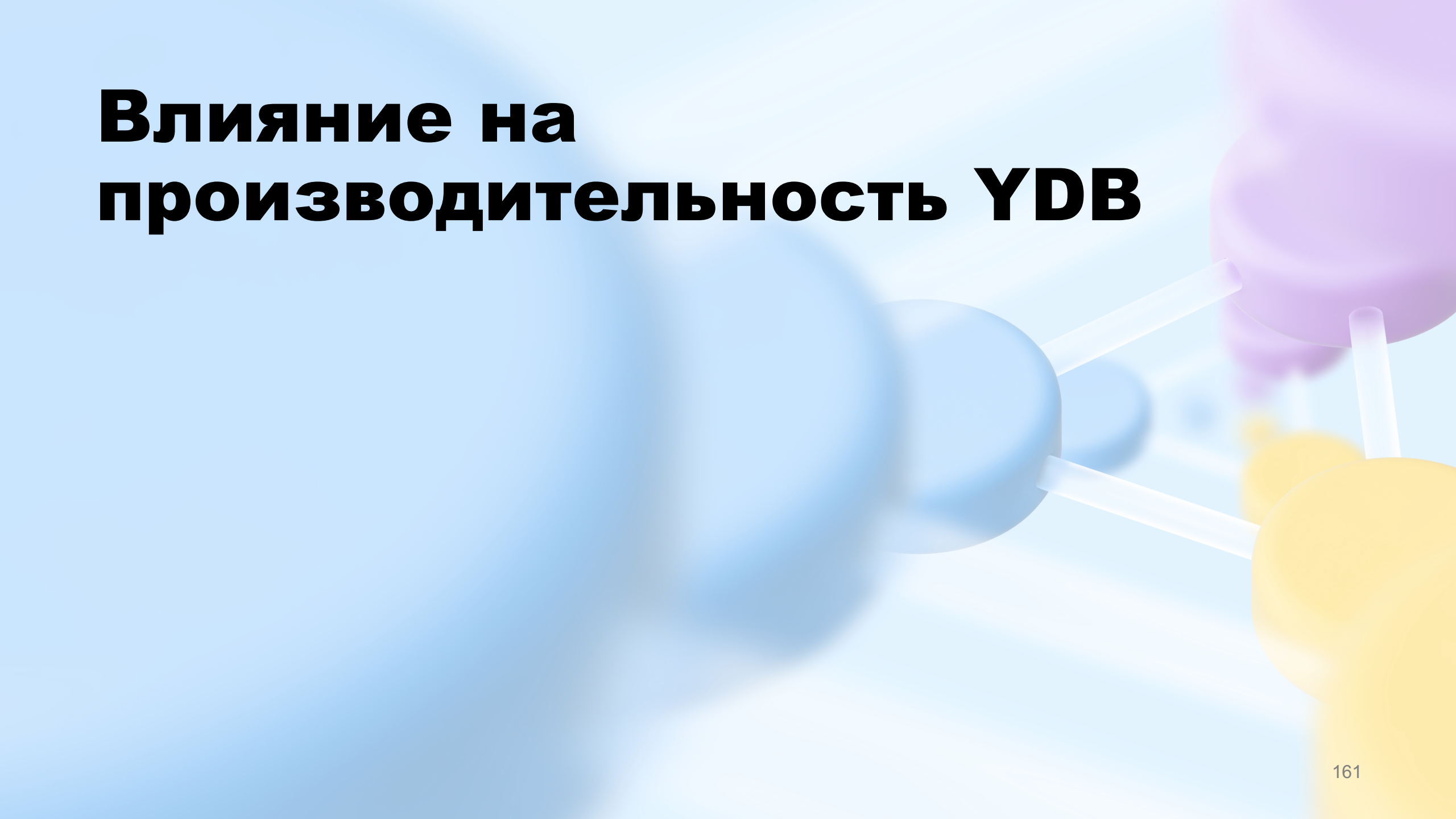
Очередь с циклическим буфером v2



Очередь с циклическим буфером v2 + короткие per-thread очереди

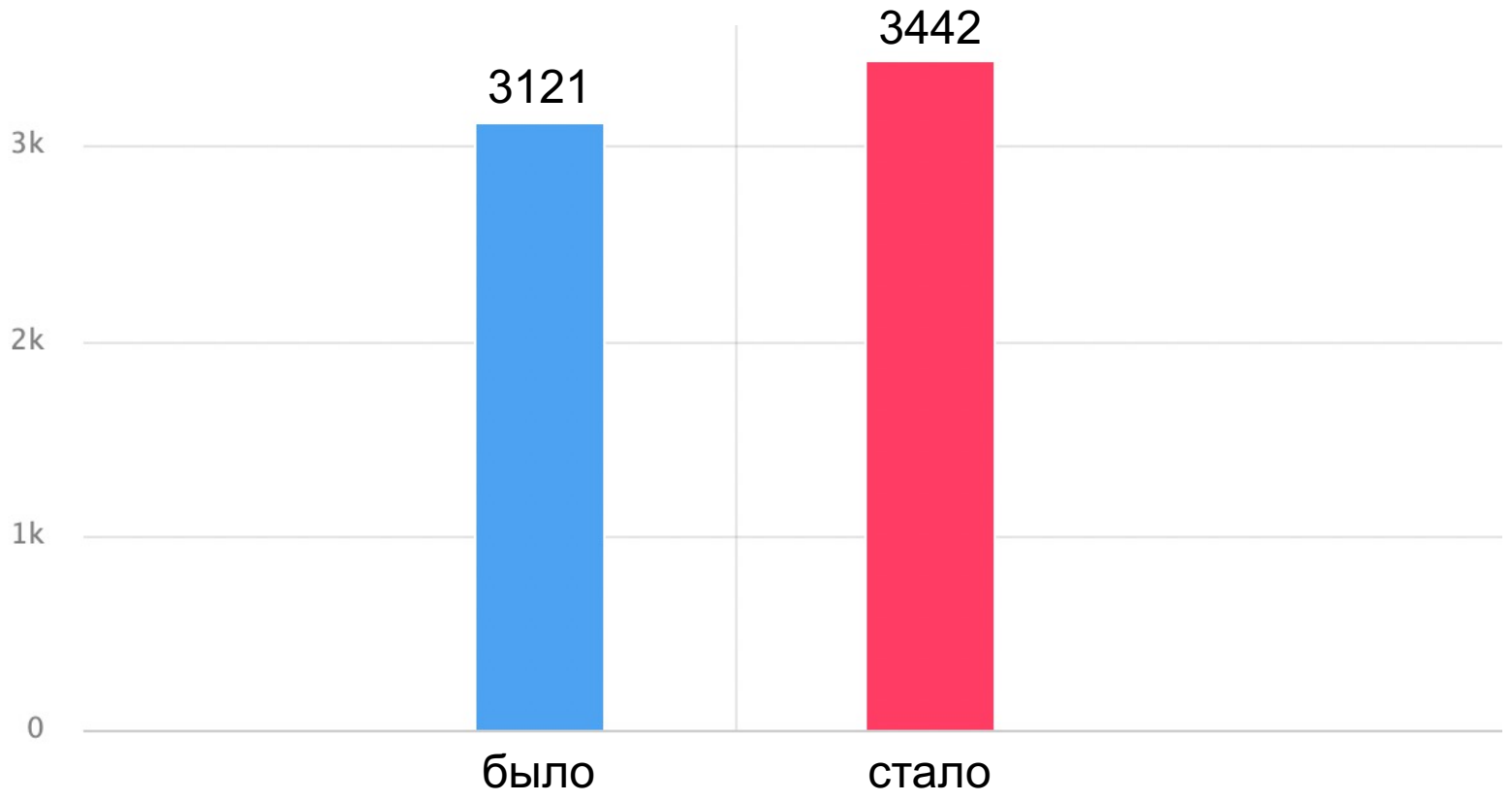


Влияние на производительность YDB

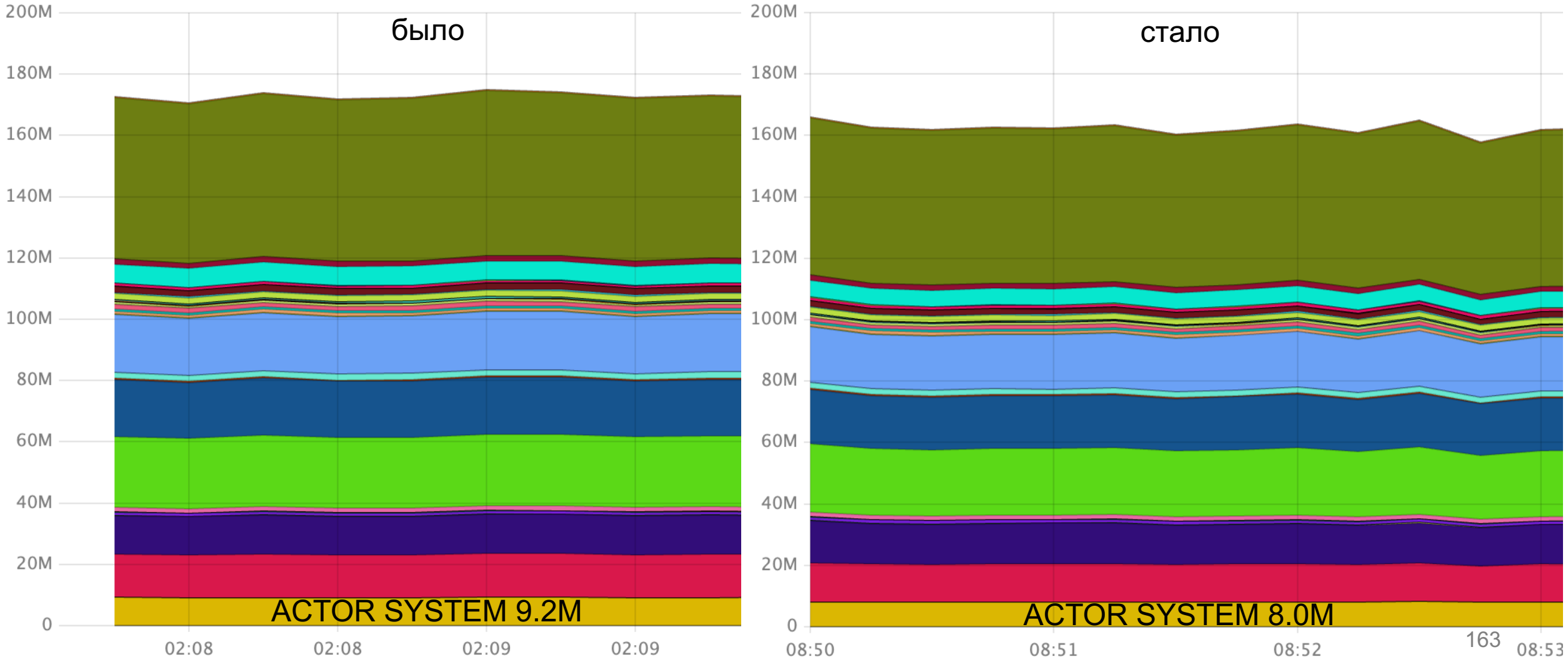


Результат

Нагрузочный тест YDB workload, транзакций в секунду

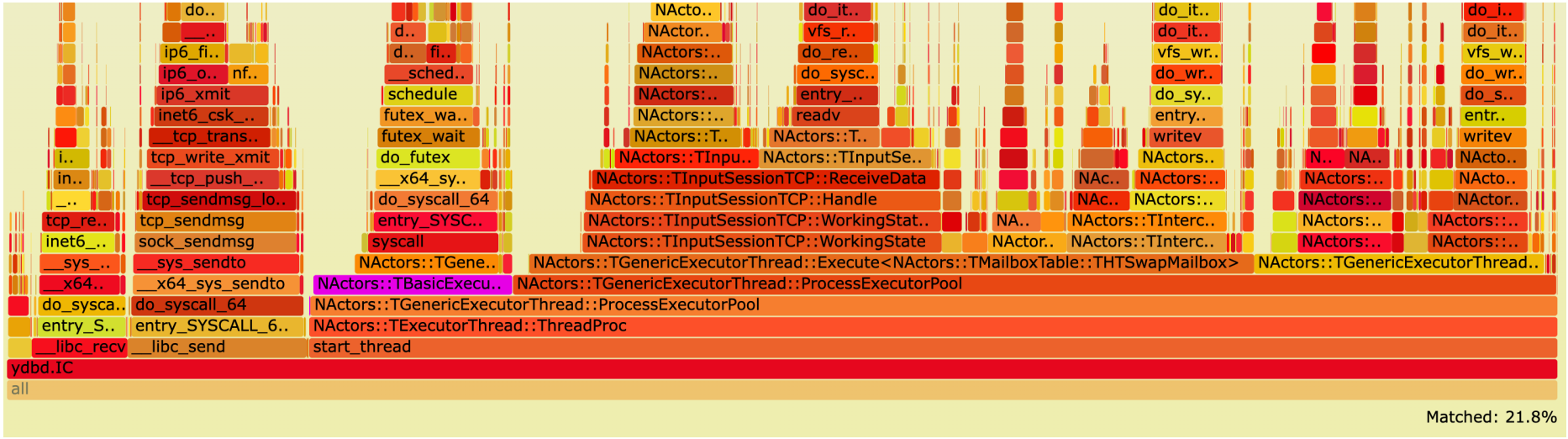


Результат

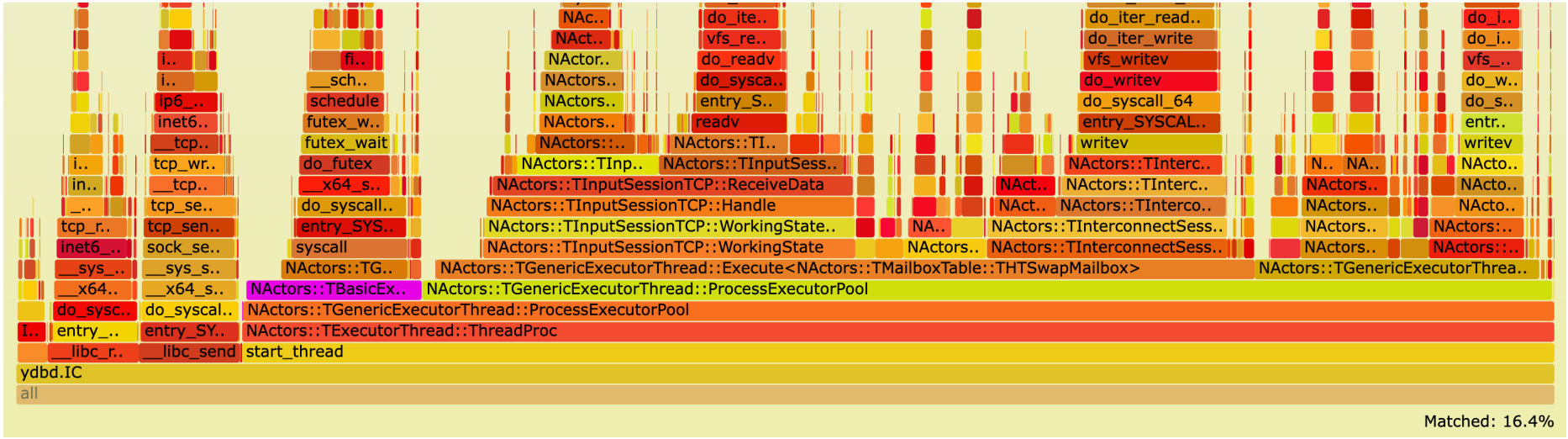

было
200M
180M
160M
140M
120M
100M
80M
60M
40M
20M
0
02:08
02:08
02:09
02:09
ACTOR SYSTEM 9.2M
стало
200M
180M
160M
140M
120M
100M
80M
60M
40M
20M
0
08:50
08:51
08:52
08:53
ACTOR SYSTEM 8.0M

Результат

было:

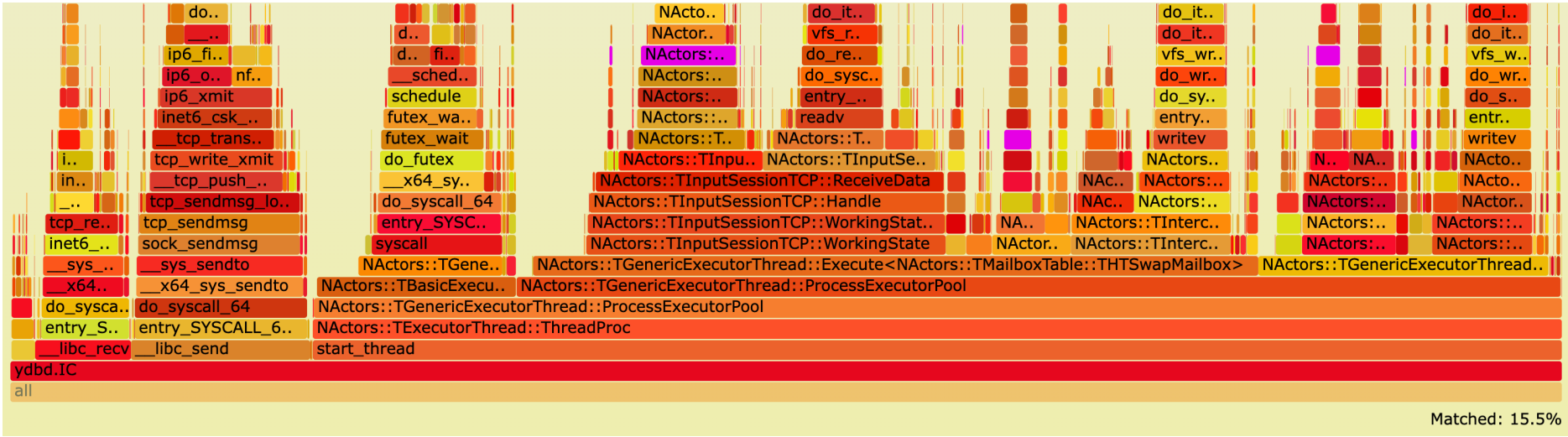


стало:

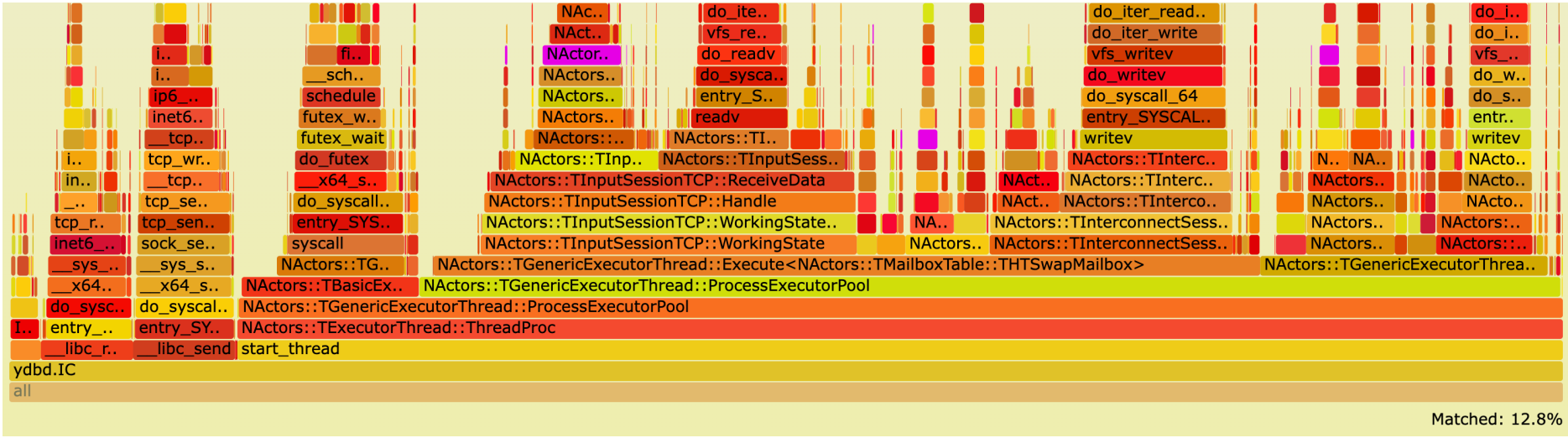


Результат

было:



стало:



Где наши очереди

- Revolving
  https://github.com/ydb-platform/ydb/blob/53259c90946aa5172013b3adb11f6037c3d288c4/ydb/library/actors/util/unordered_cache.h
- Chain
  https://gitlab.com/agrianius/mt_queue/-/blob/master/mpmc_buffer_chain.hh
- Ring v2
  https://github.com/ydb-platform/ydb/blob/53259c90946aa5172013b3adb11f6037c3d288c4/ydb/library/actors/util/mpmc_ring_queue.h

YDB

Спасибо!

Алексей Станкевичус,
ydb.tech