



Поиск самой быстрой MRMC-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
изоляции транзакций

Горизонтальная
масштабируемость

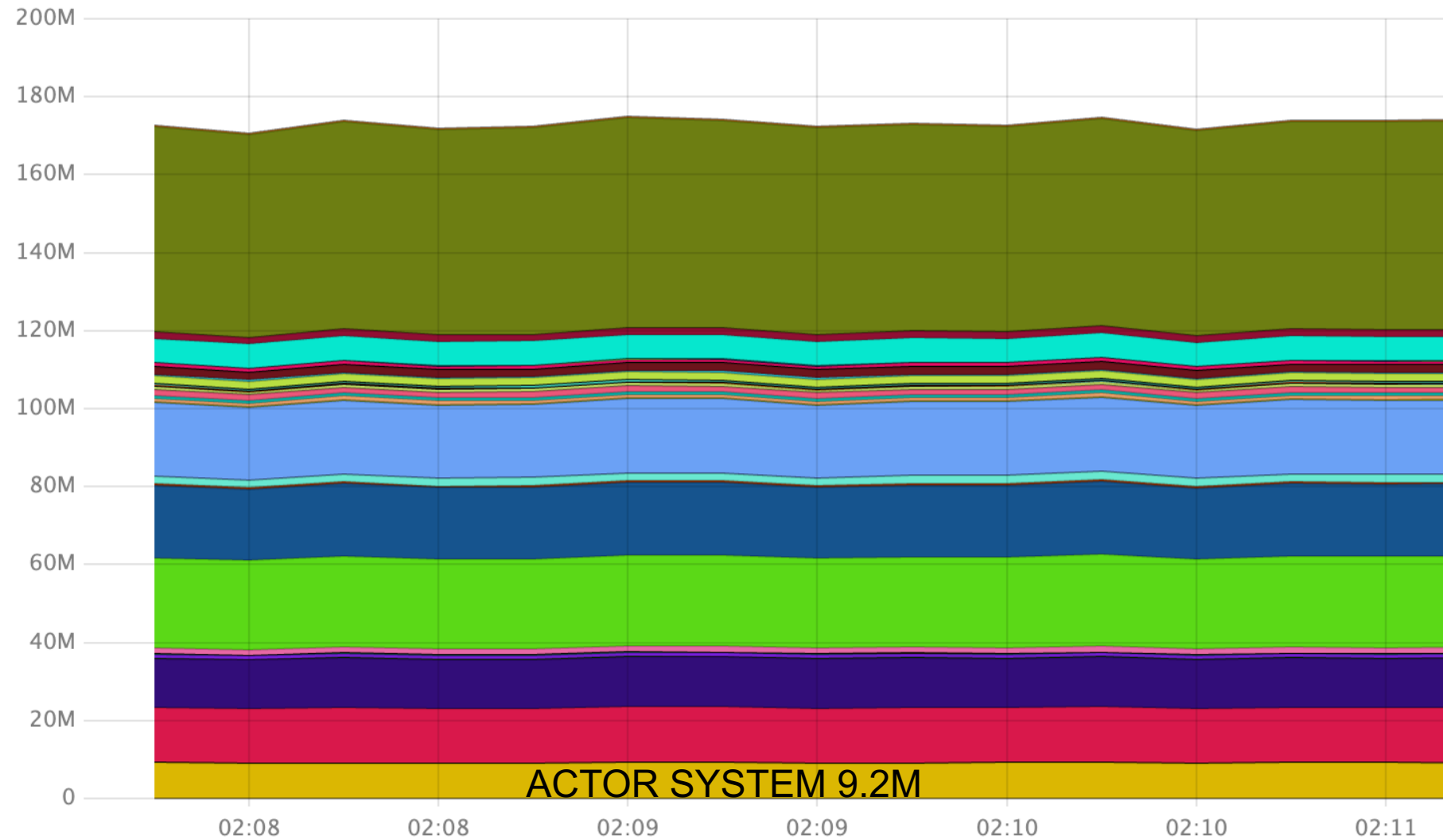
Кластеризуемая
Тысячи серверов

Проблема производительности

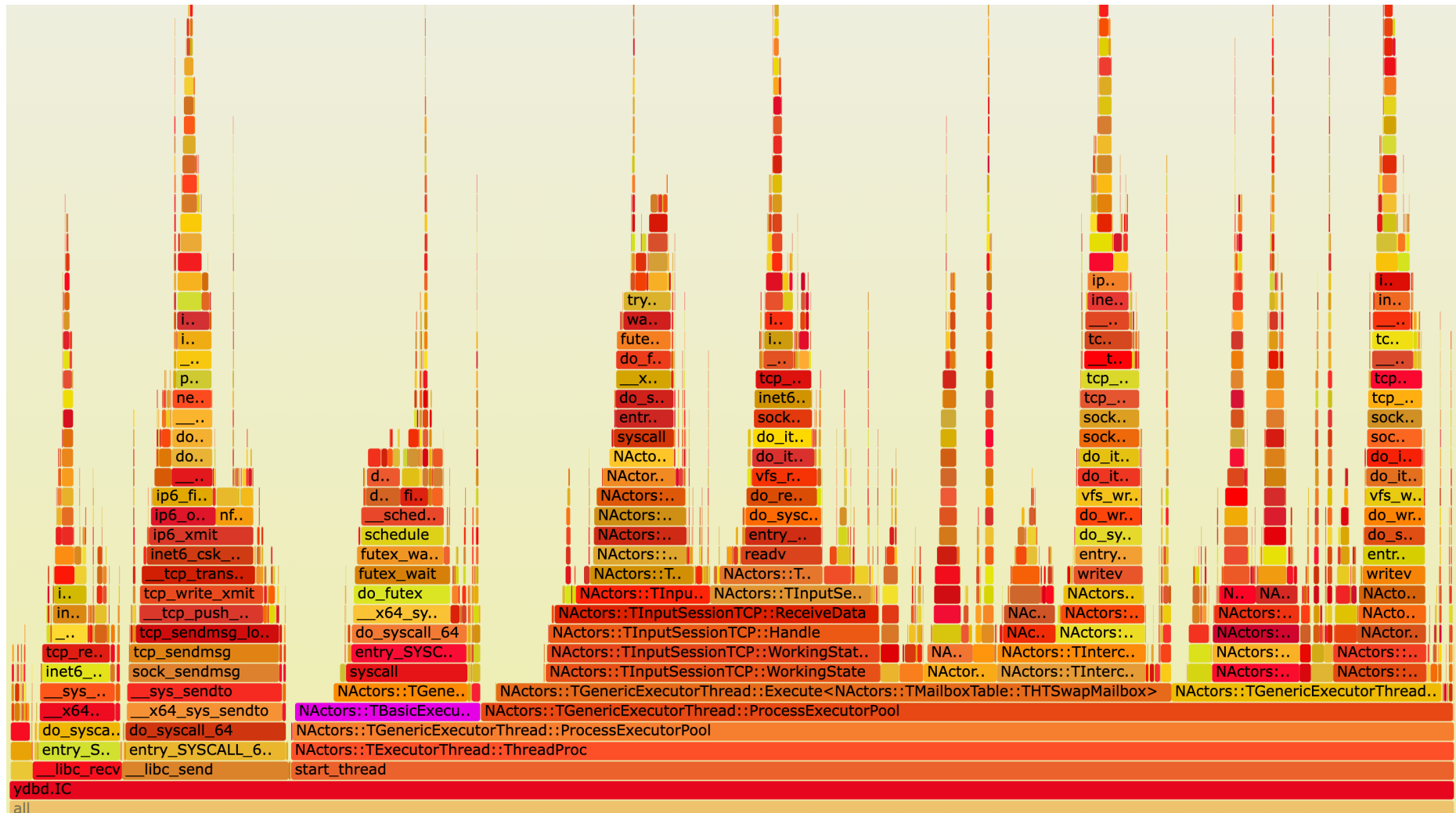
Высокая производительность YDB очень важна для нас.

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

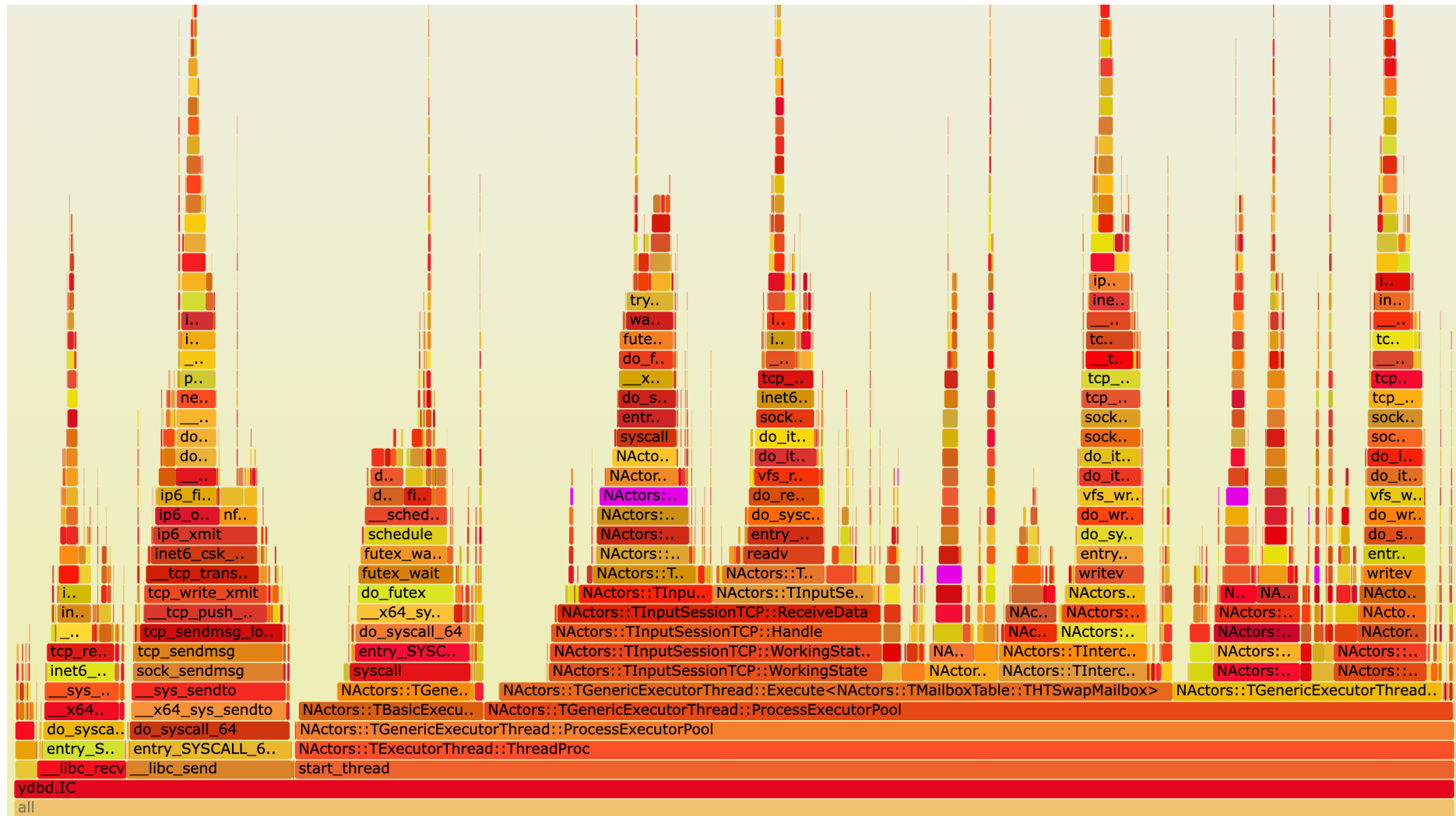
Где все тормозит



Где все тормозит



Где все тормозит

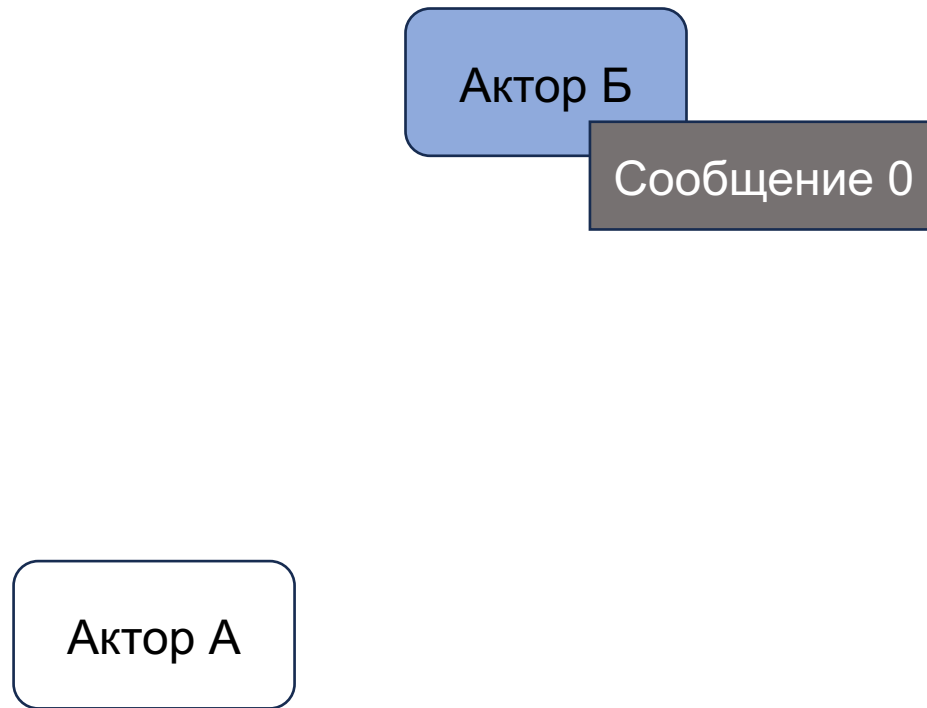


Matched: 15.5%

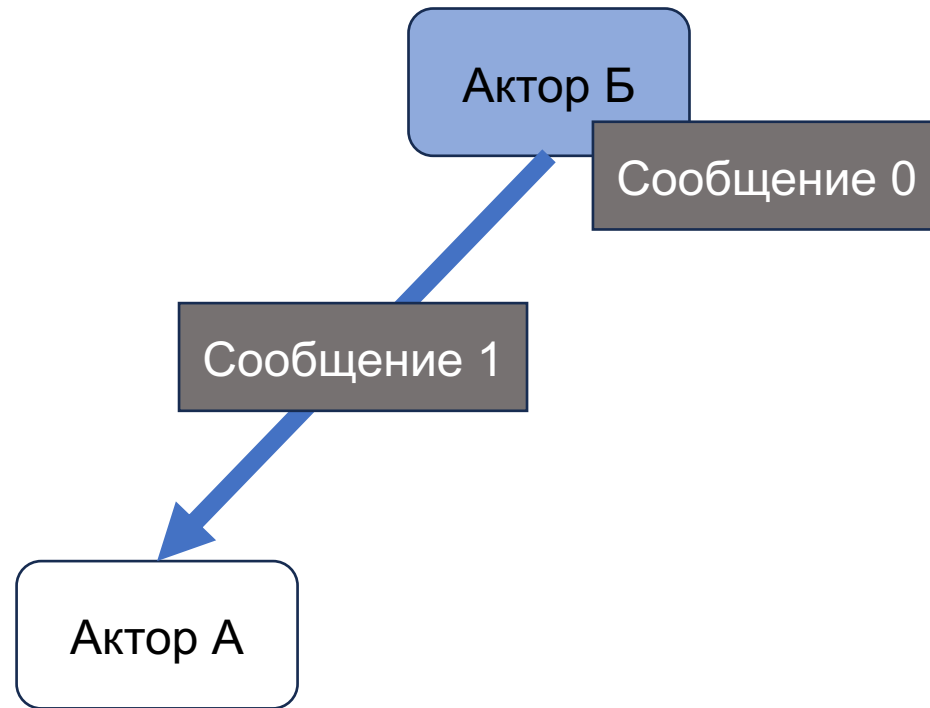
Что вообще такое акторная система?

Акторная система

Лежит в основе YDB



Акторная система



Акторная система

Актор Б

Сообщение 1

Актор А

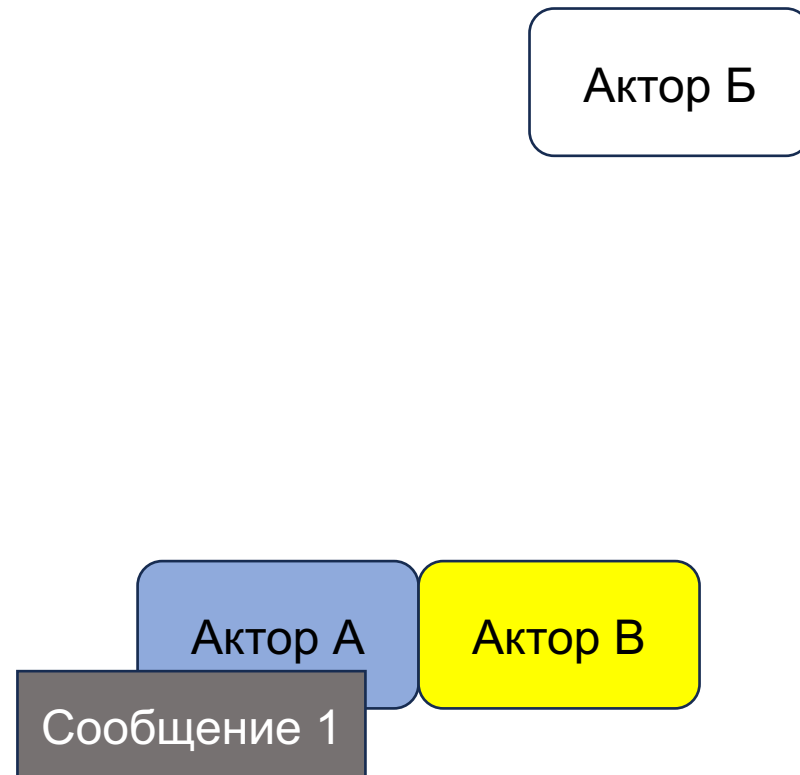
Акторная система

Актор Б

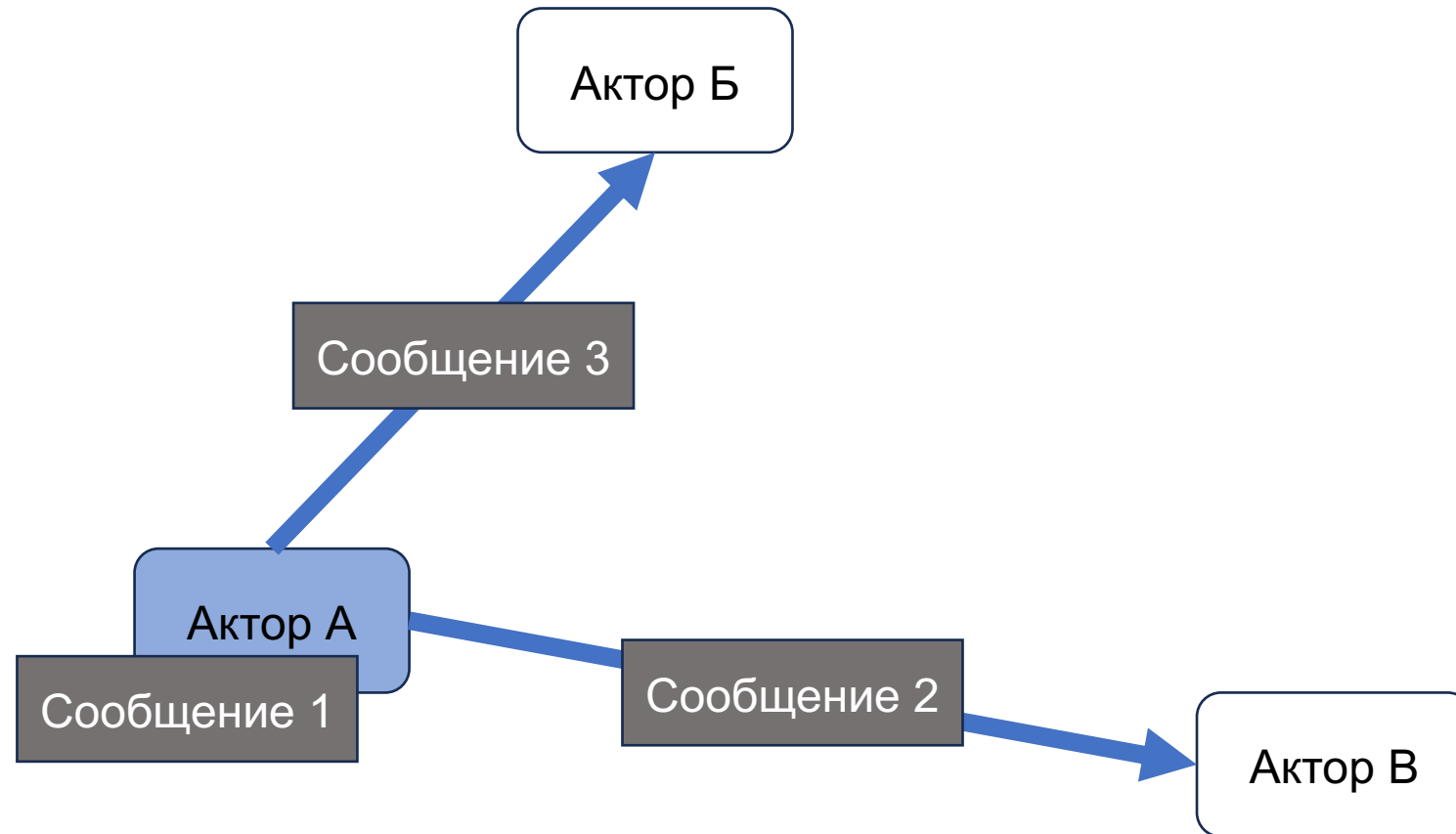
Актор А

Сообщение 1

Акторная система



Акторная система



Акторная система



Акторная система

Сообщение 3

Актор Б

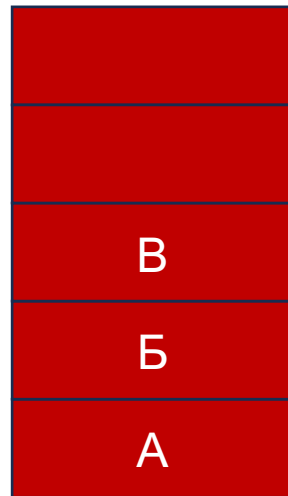
Сообщение 2

Актор В

Как происходит активация акторов?

Очередь активаций

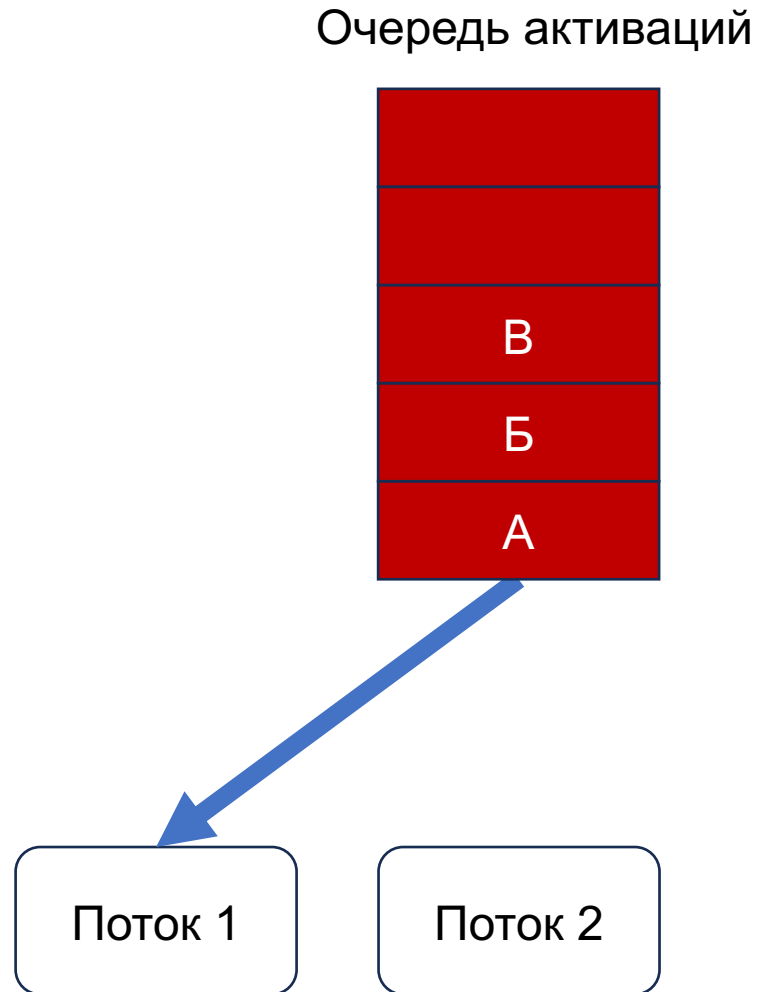
Очередь активаций



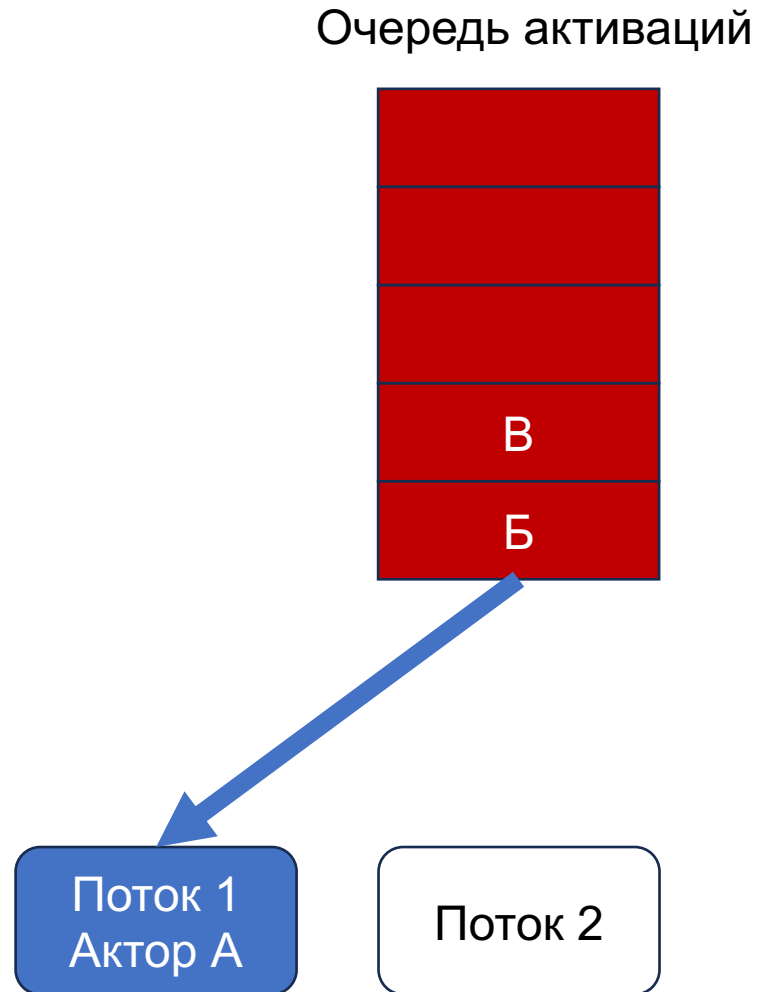
Поток 1

Поток 2

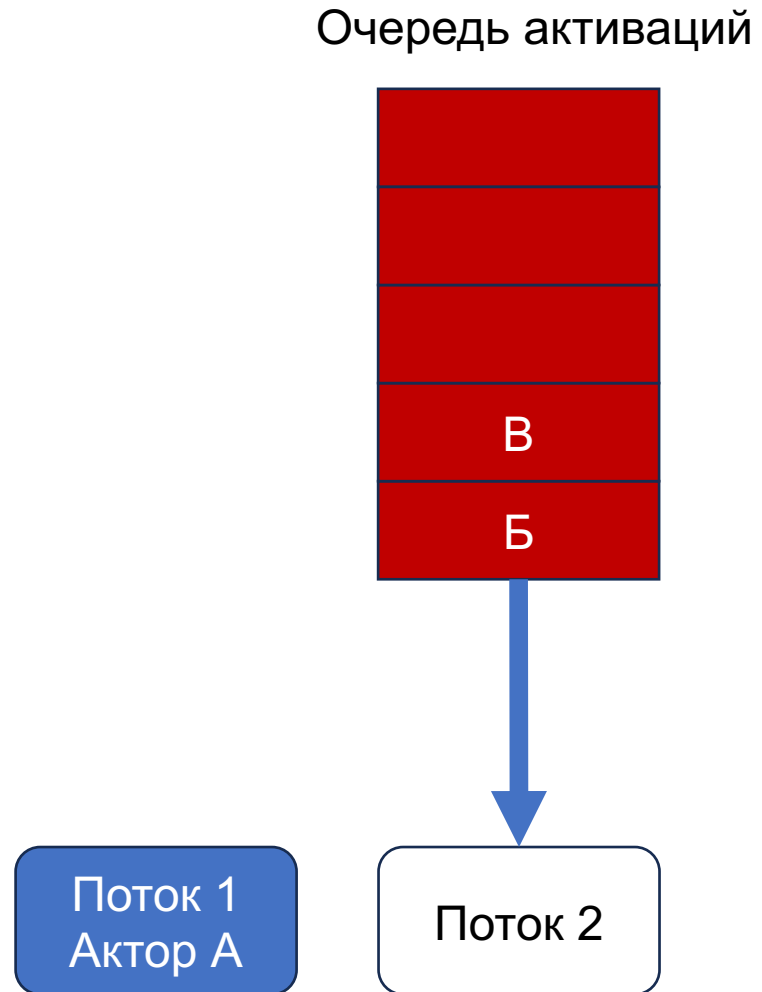
Очередь активаций



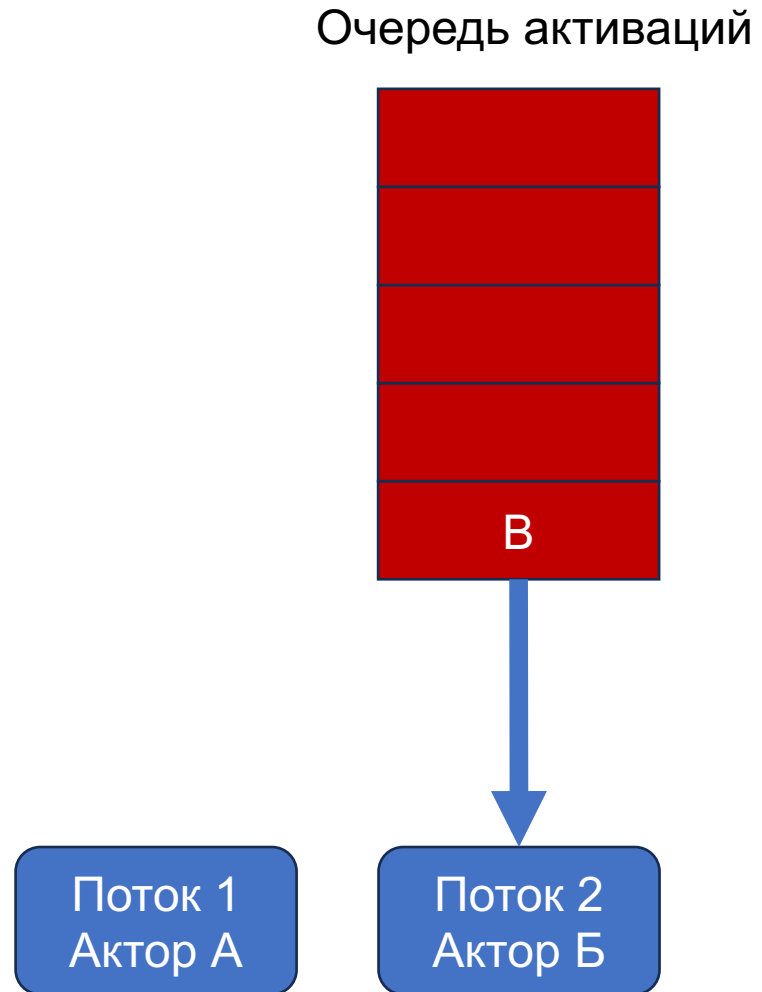
Очередь активаций



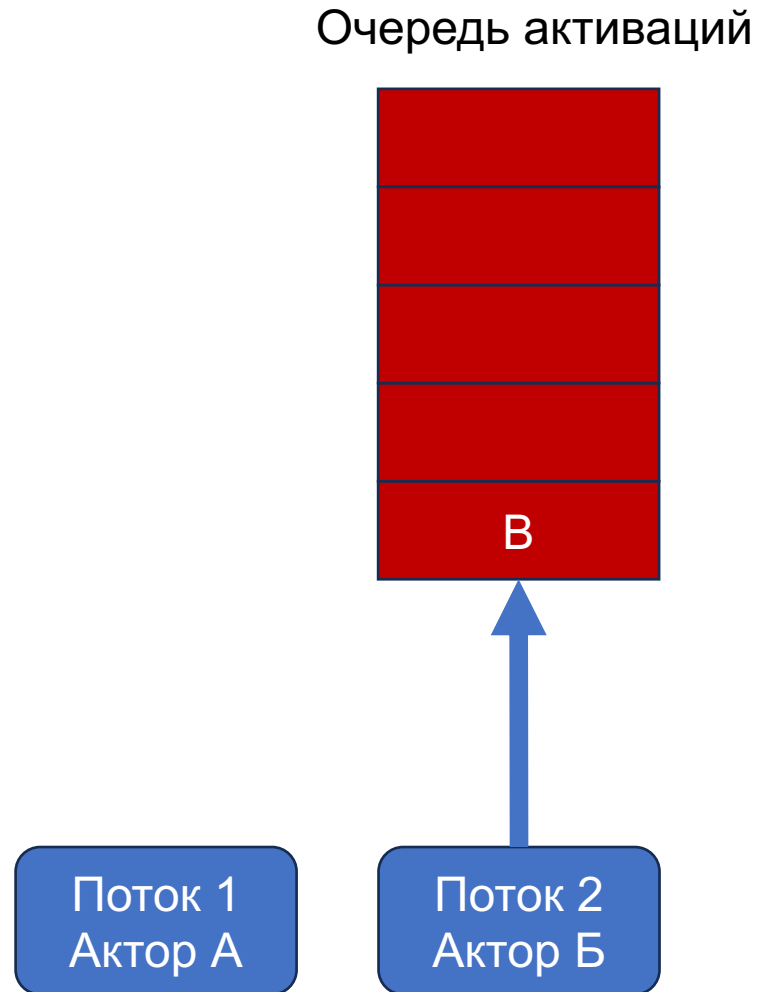
Очередь активаций



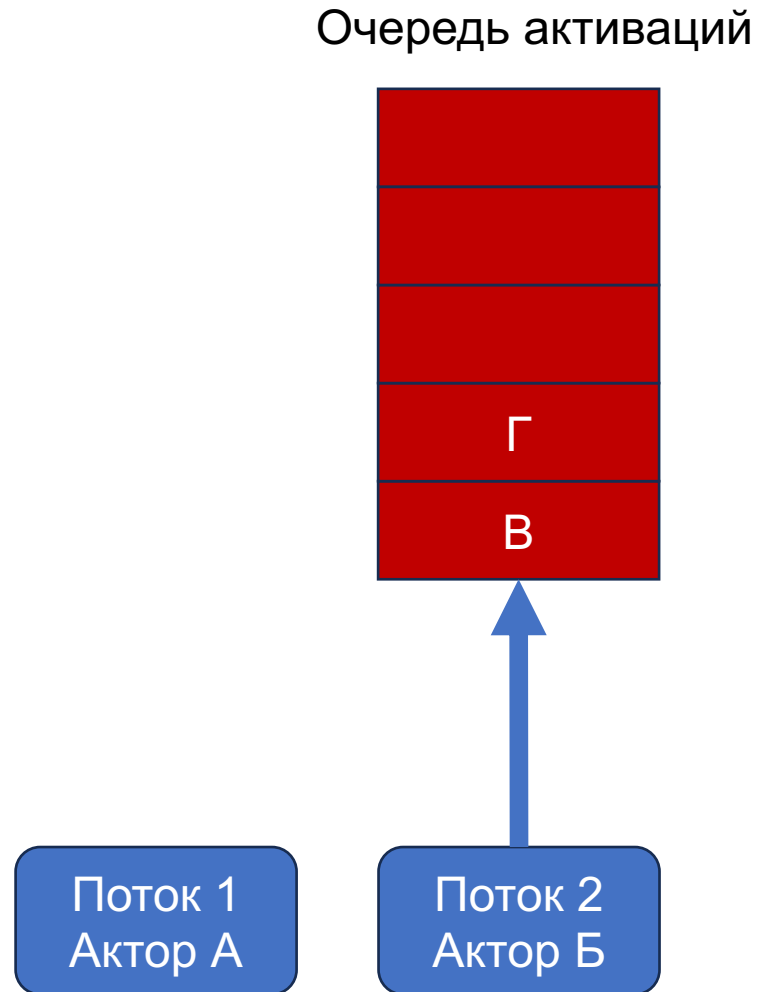
Очередь активаций



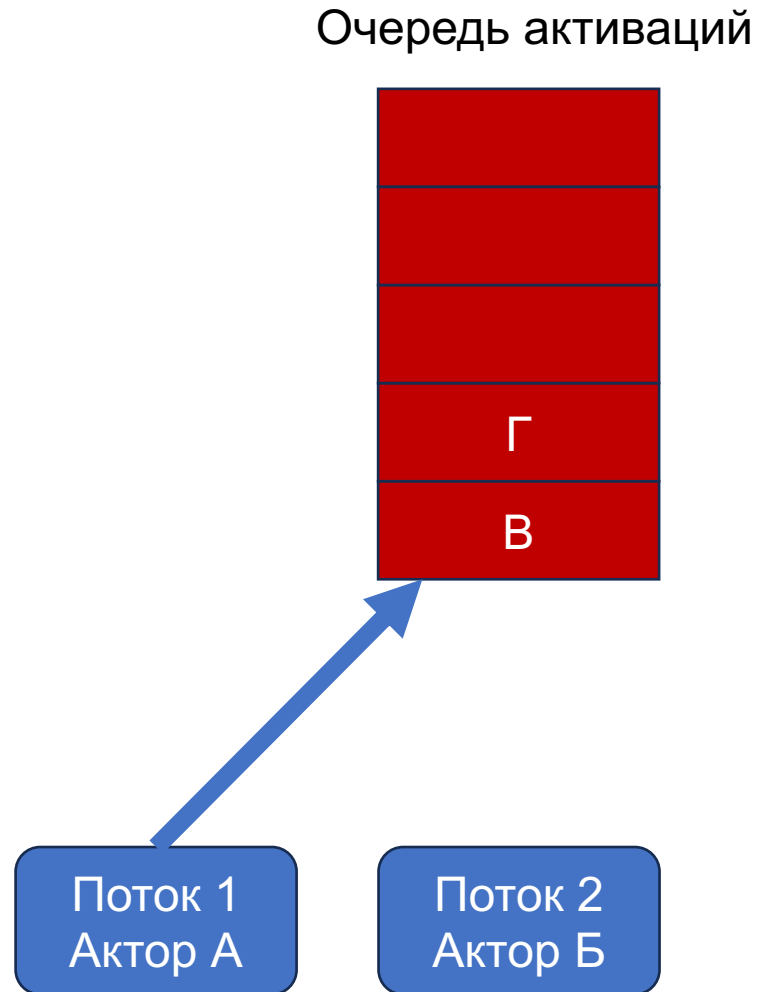
Очередь активаций



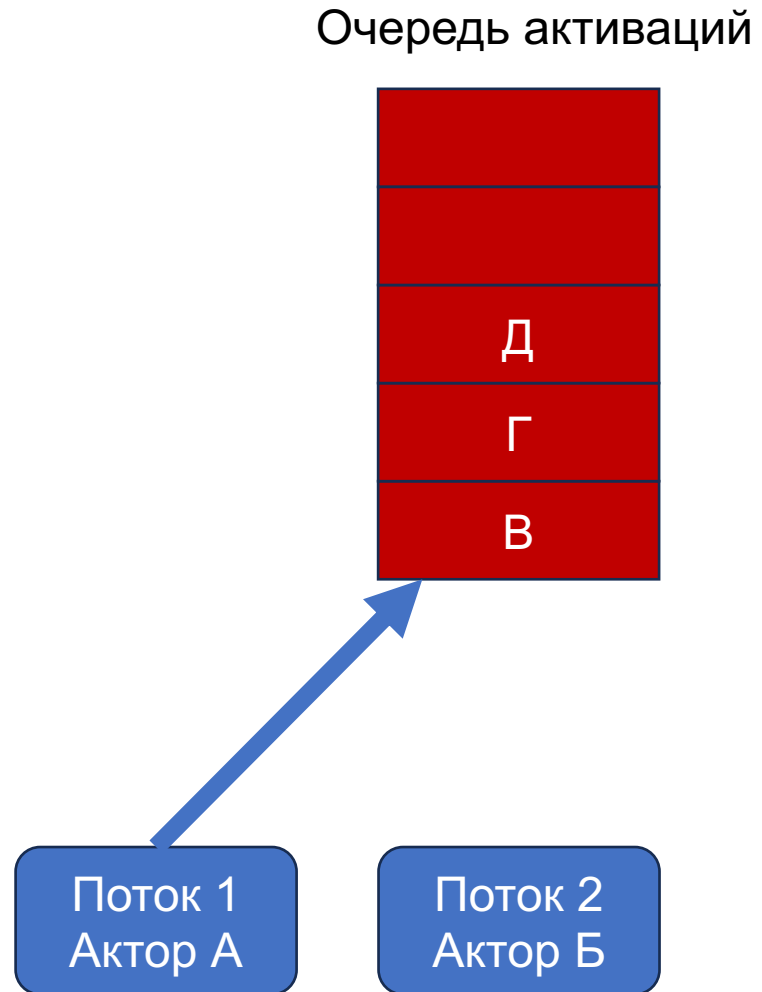
Очередь активаций



Очередь активаций



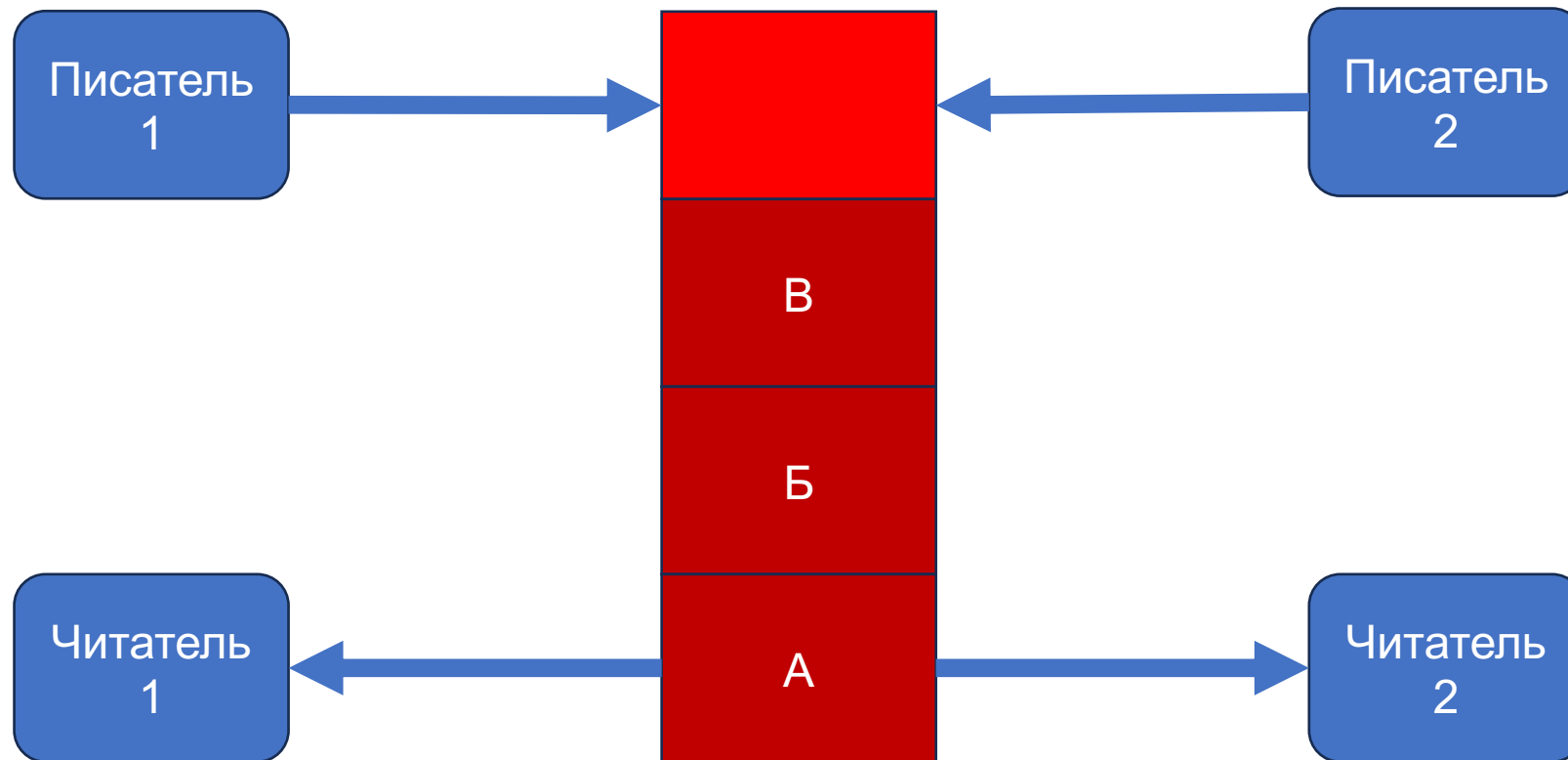
Очередь активаций



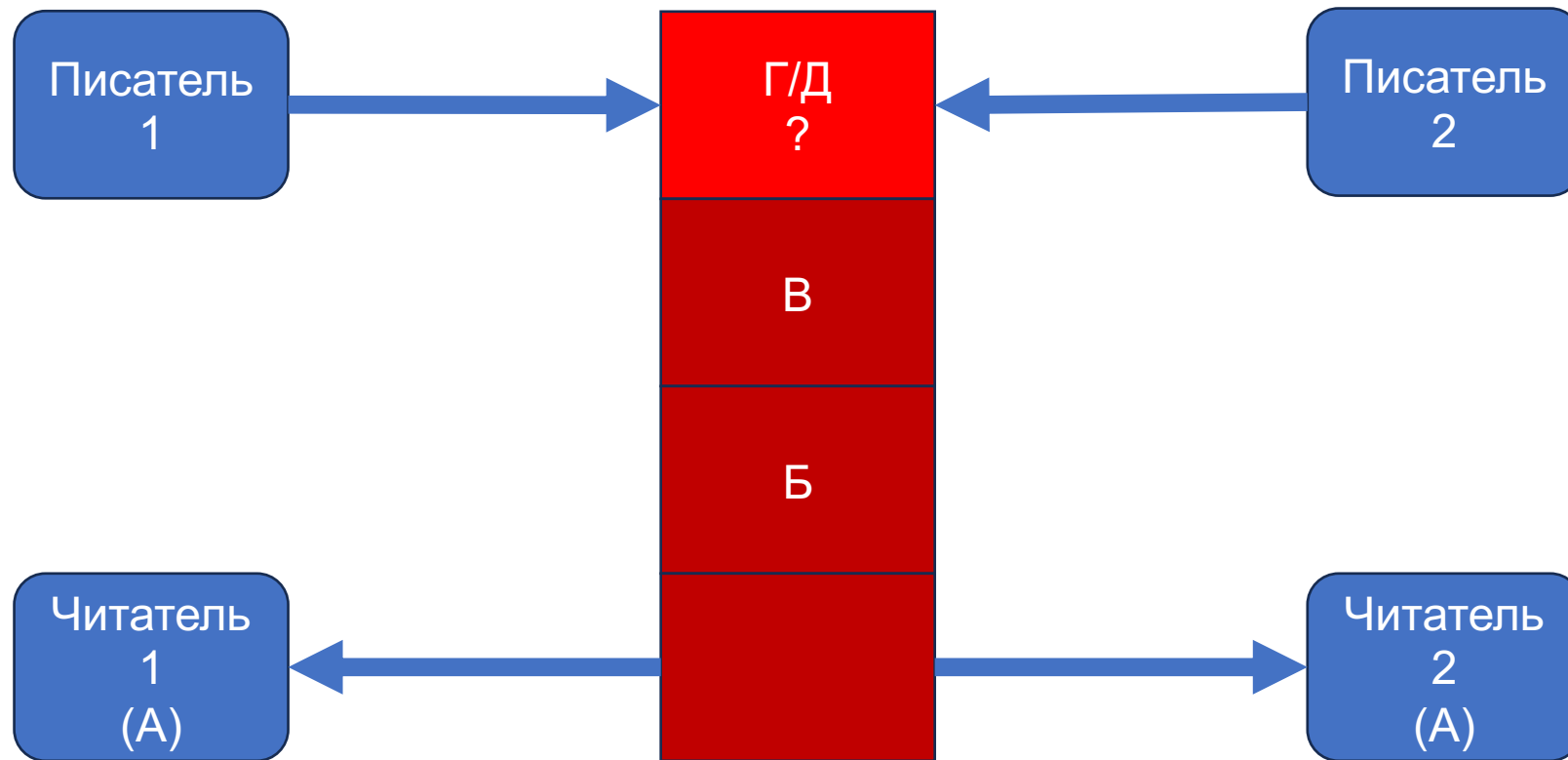
Как может быть устроена очередь активации?

Примитивная очередь

Так никто не делает!



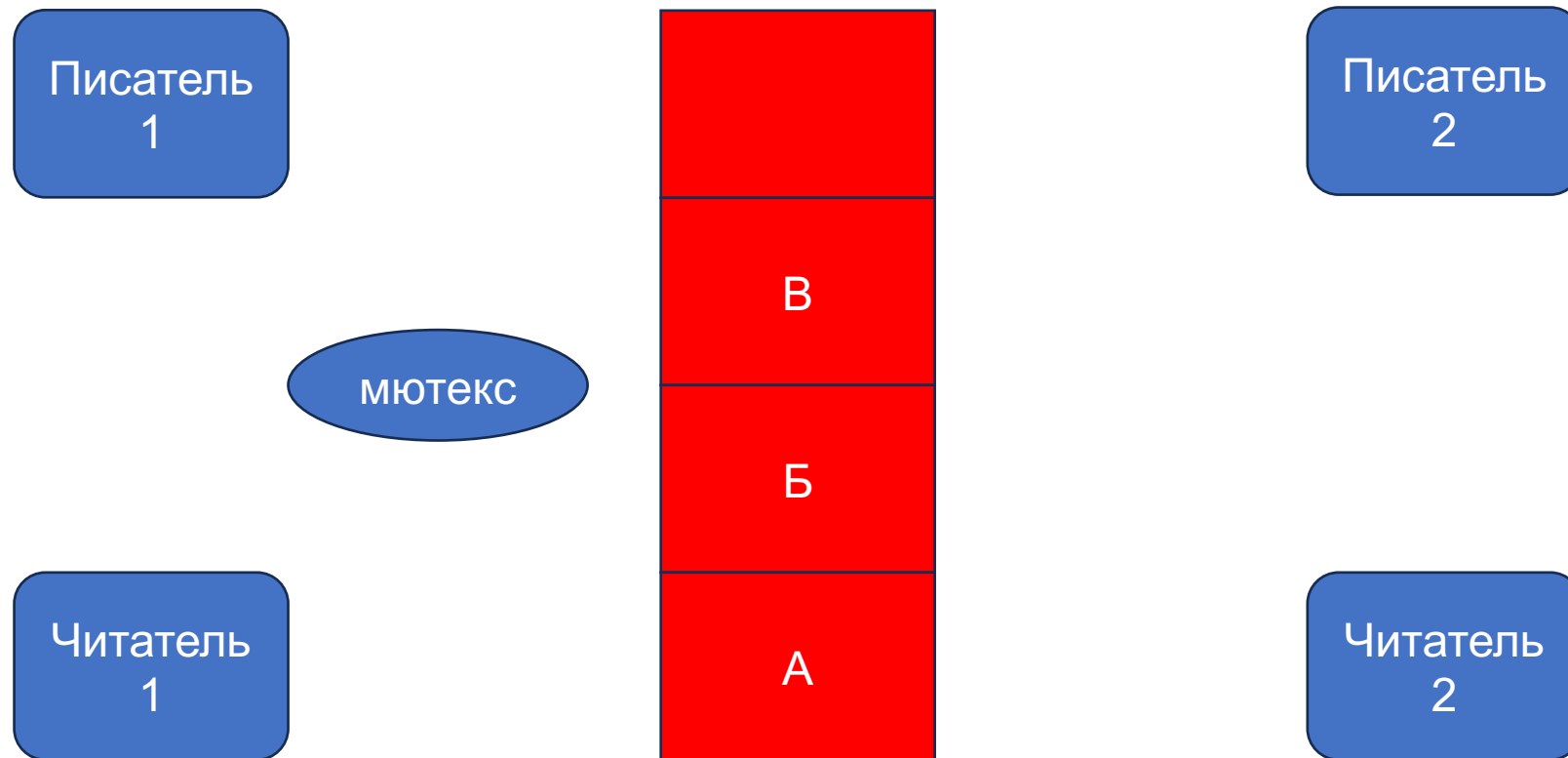
Примитивная очередь



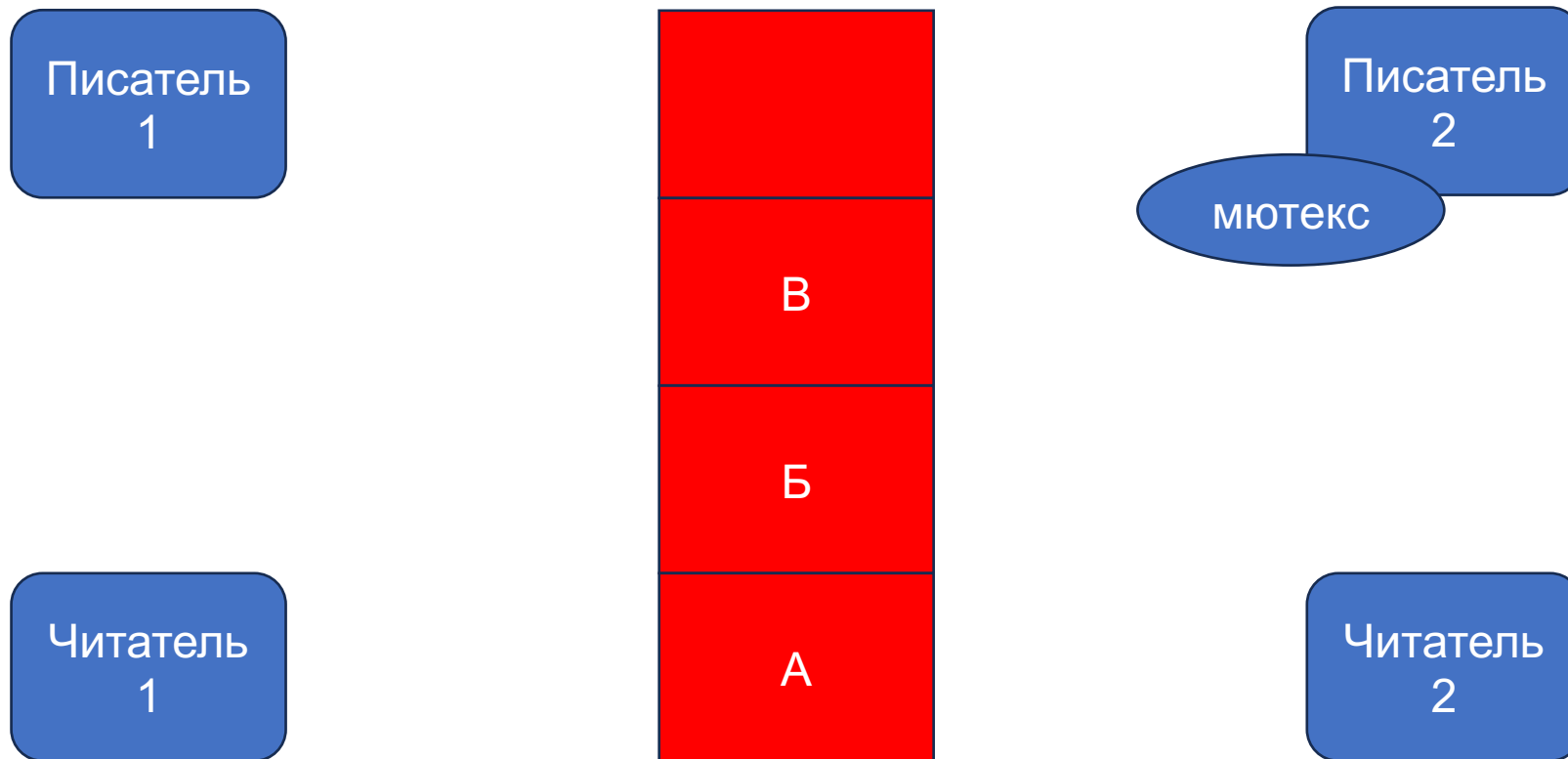
MPMC — Multiple Producers Multiple Consumers

Примитивная МРМС очередь с мютексом

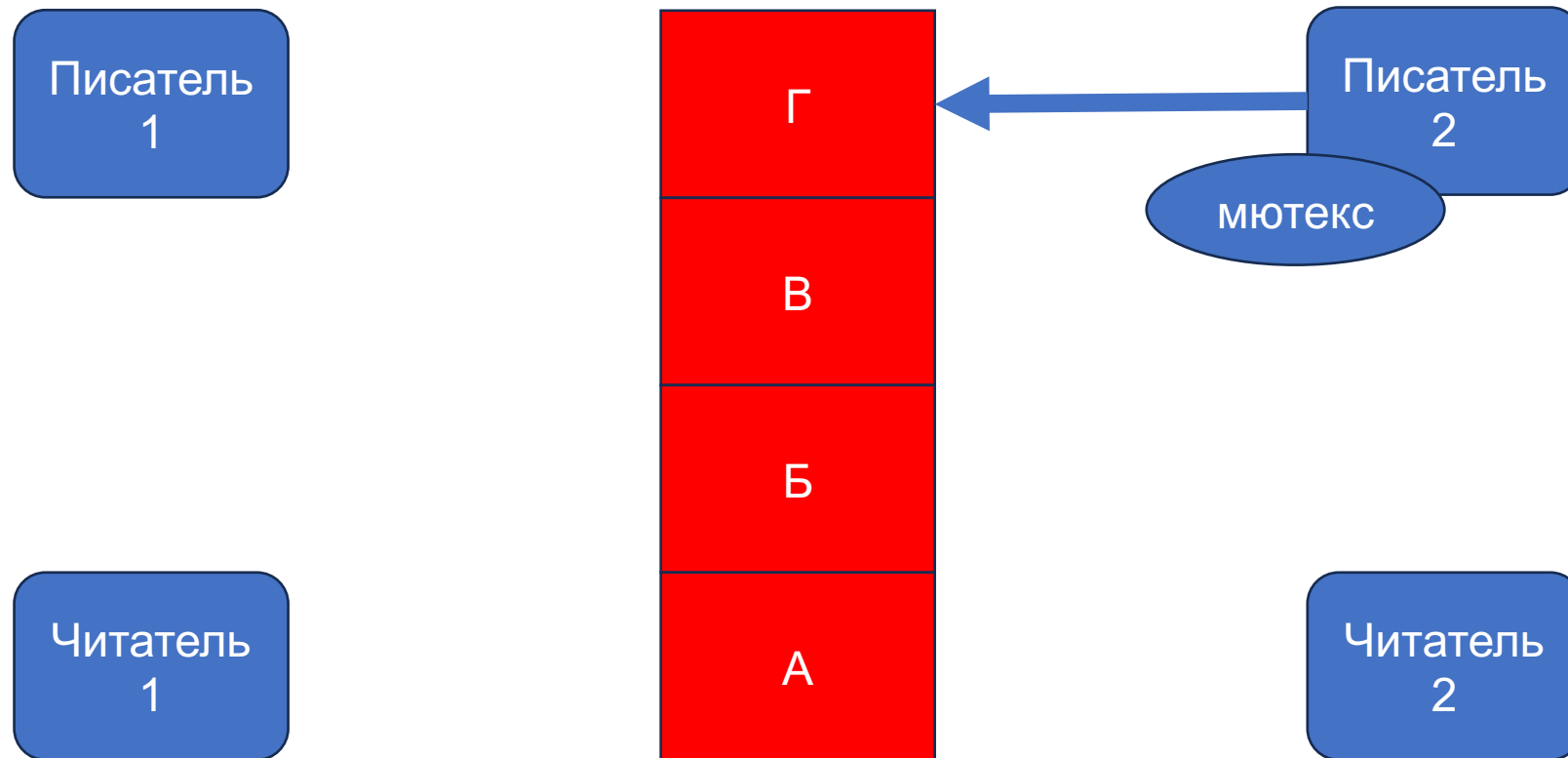
Все так делают.



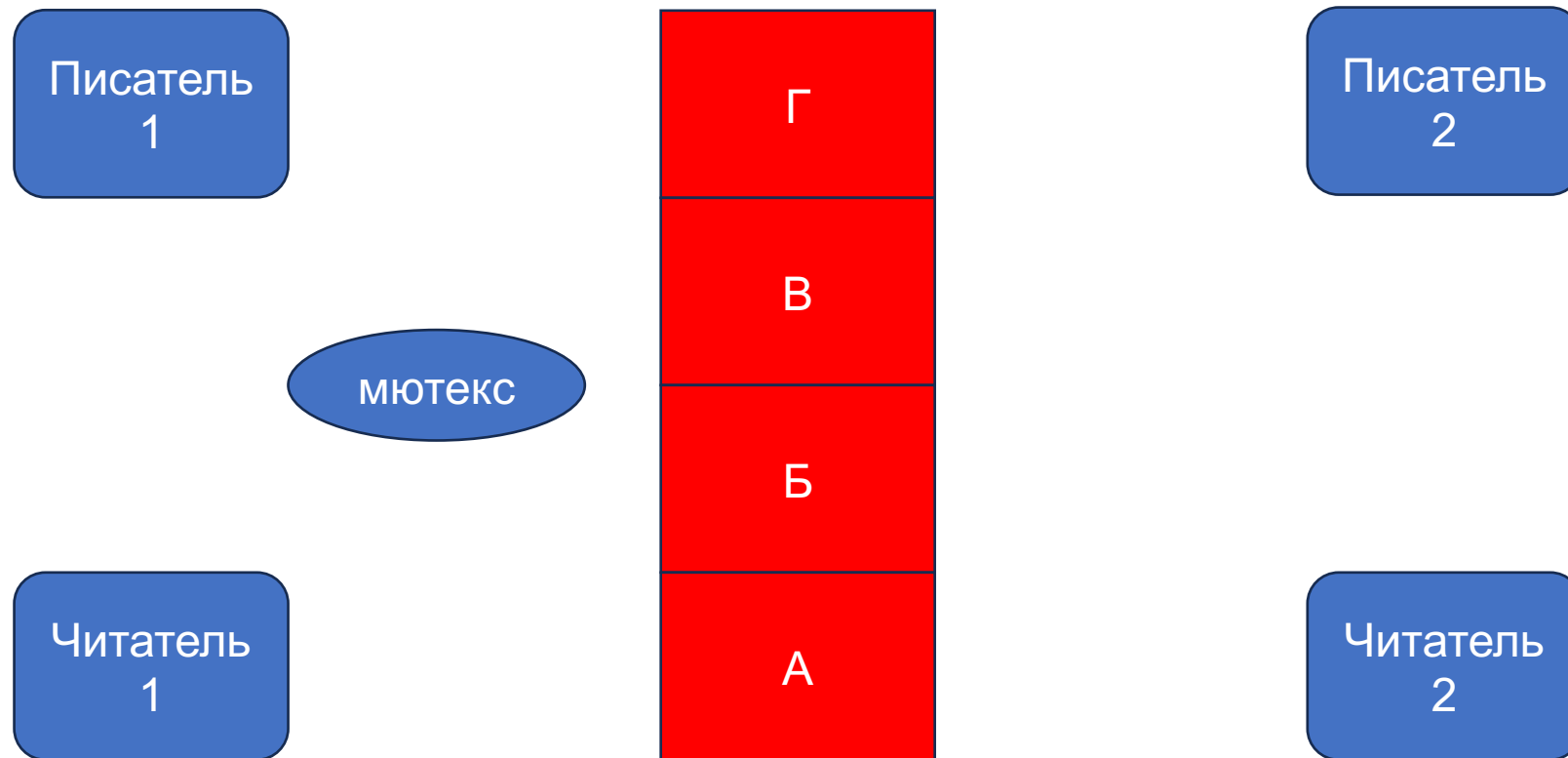
Примитивная МРМС очередь с мютексом



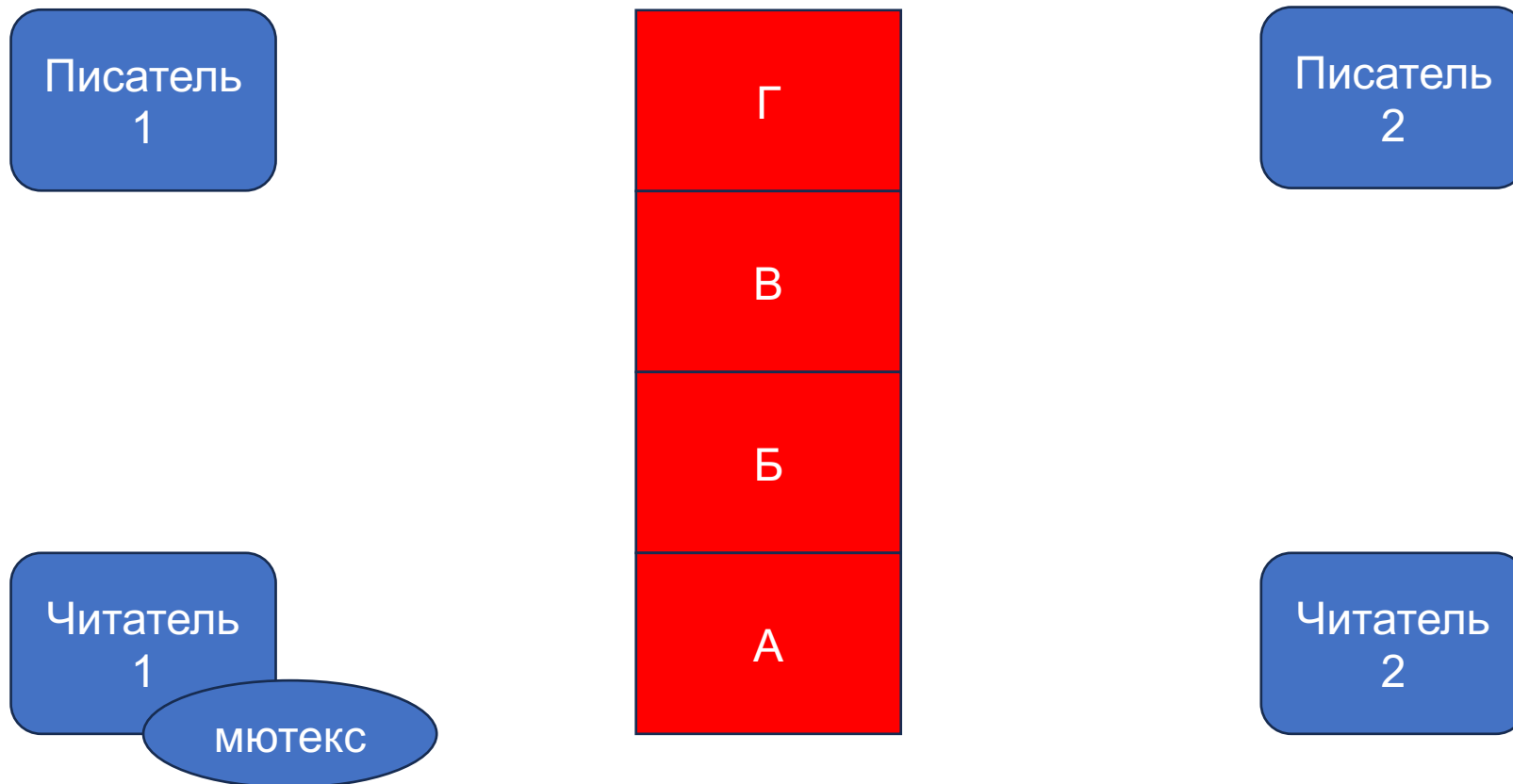
Примитивная МРМС очередь с мютексом



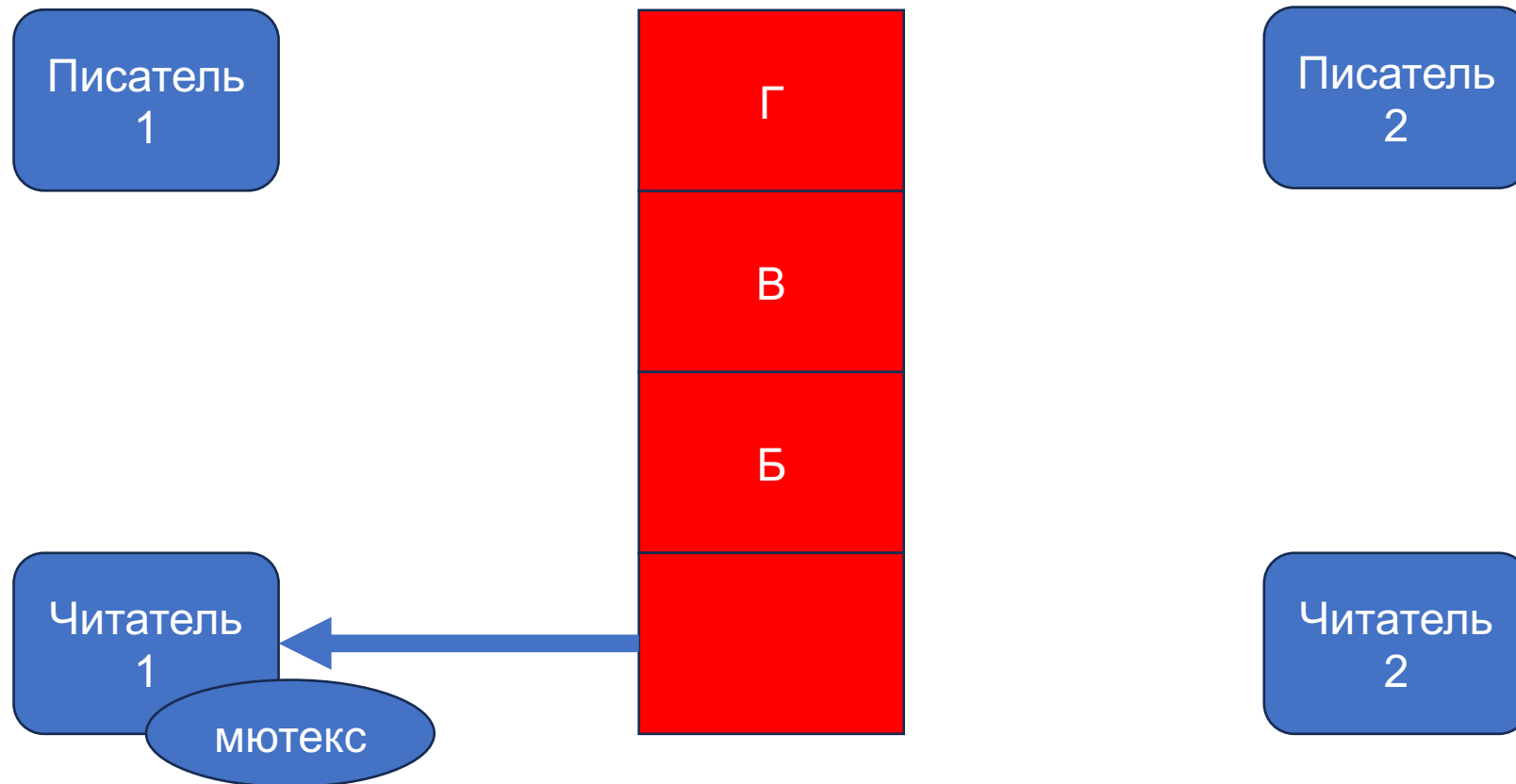
Примитивная МРМС очередь с мьютексом



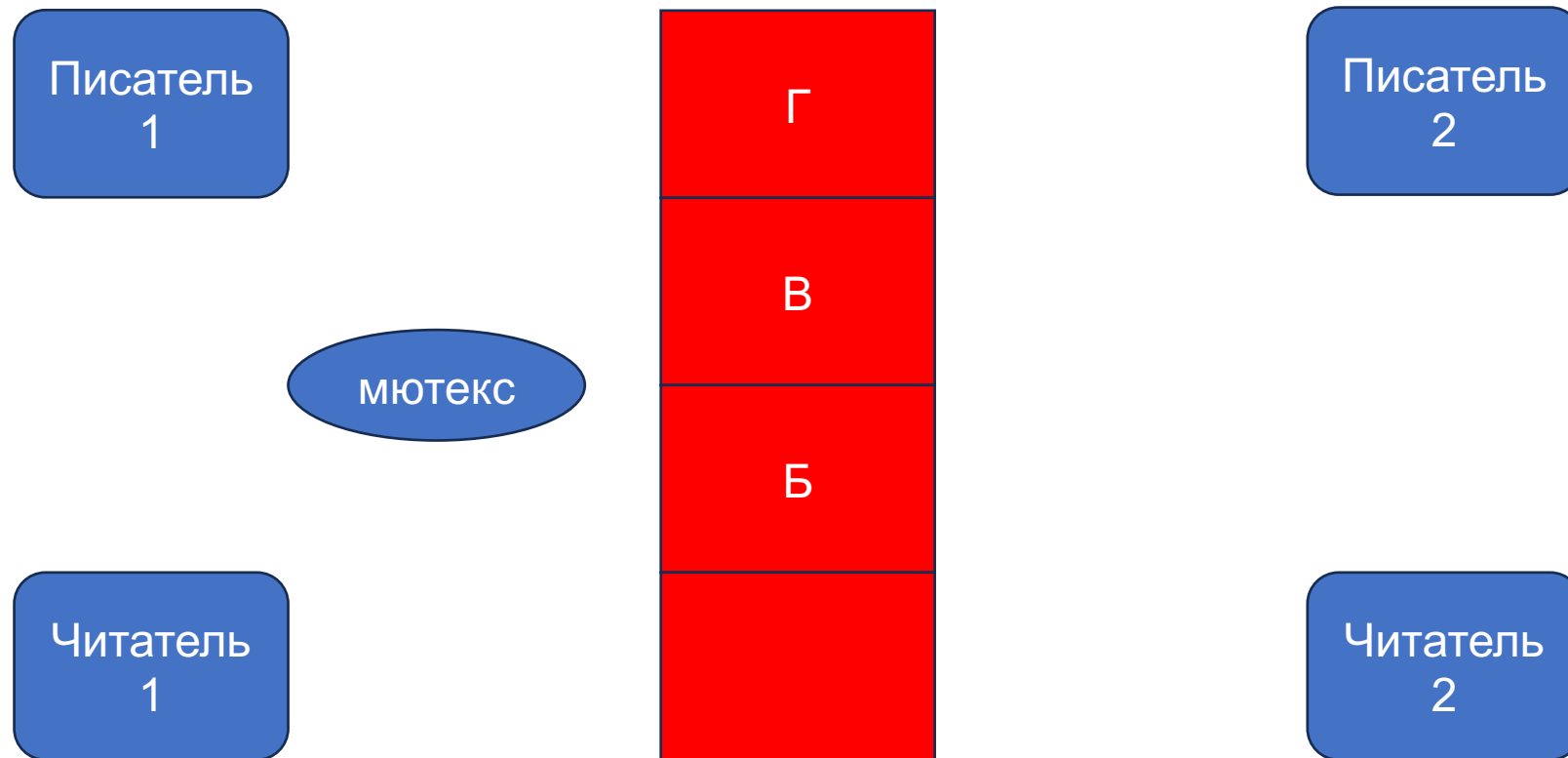
Примитивная МРМС очередь с мютексом



Примитивная МРМС очередь с мютексом

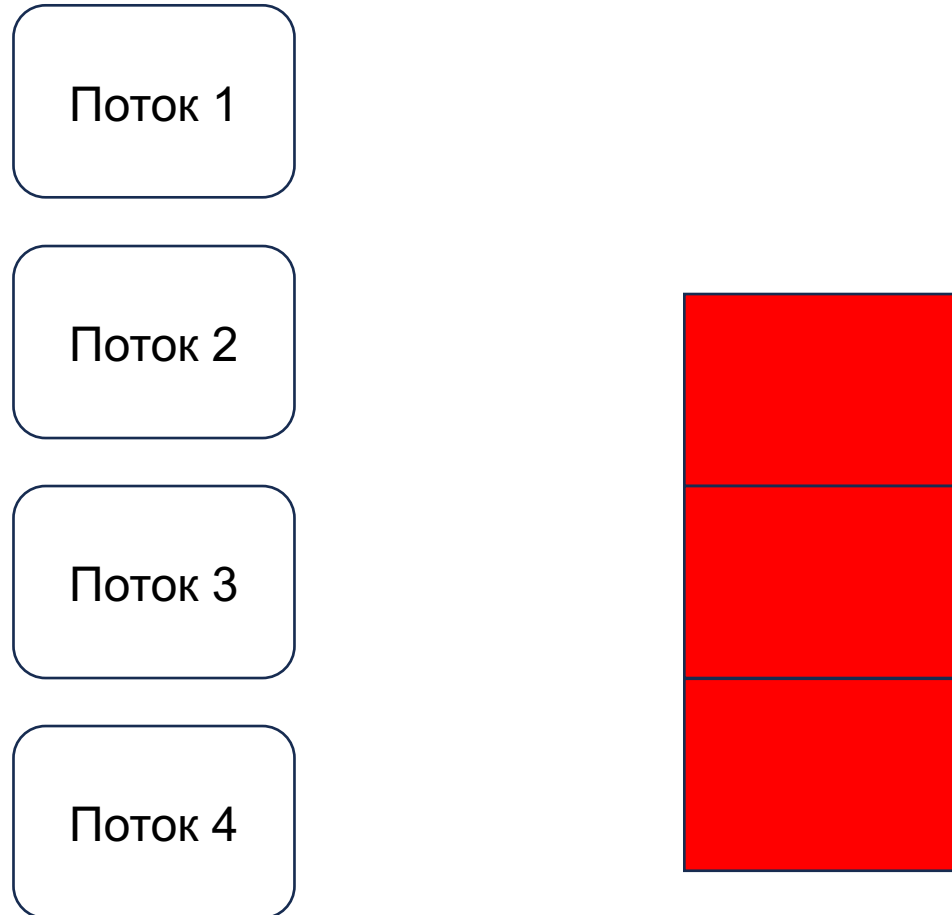


Примитивная МРМС очередь с мютексом

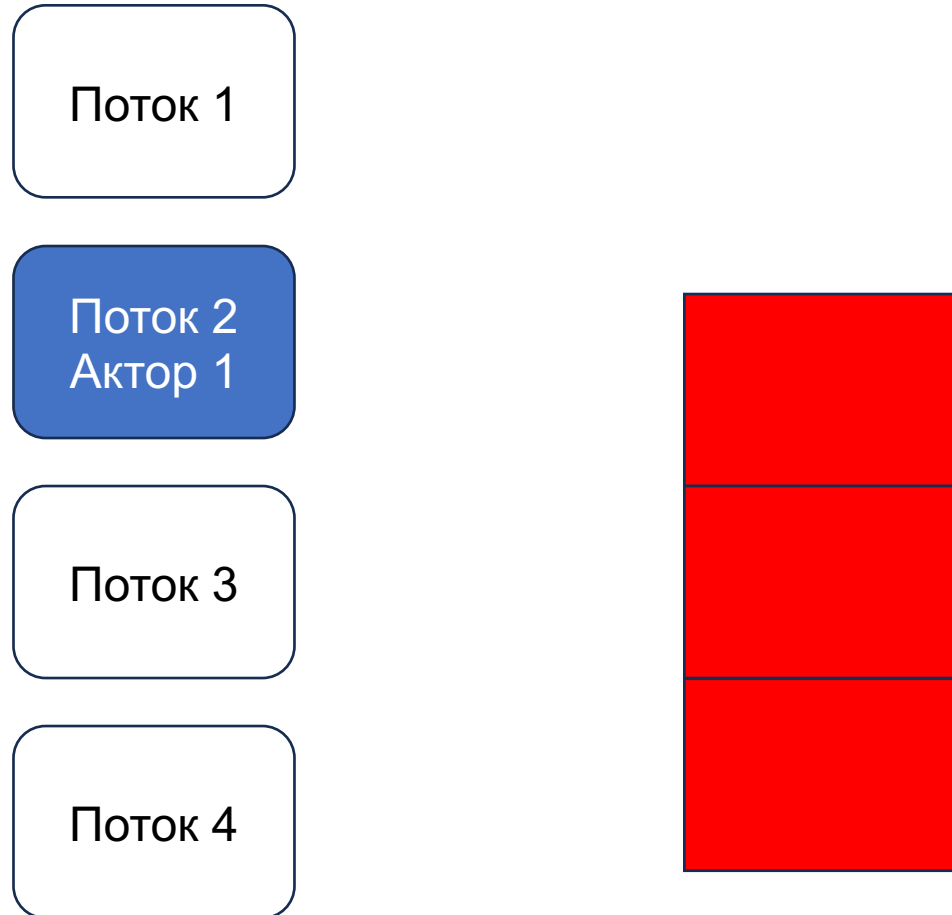


Как мы сравнивали производительность очередей?

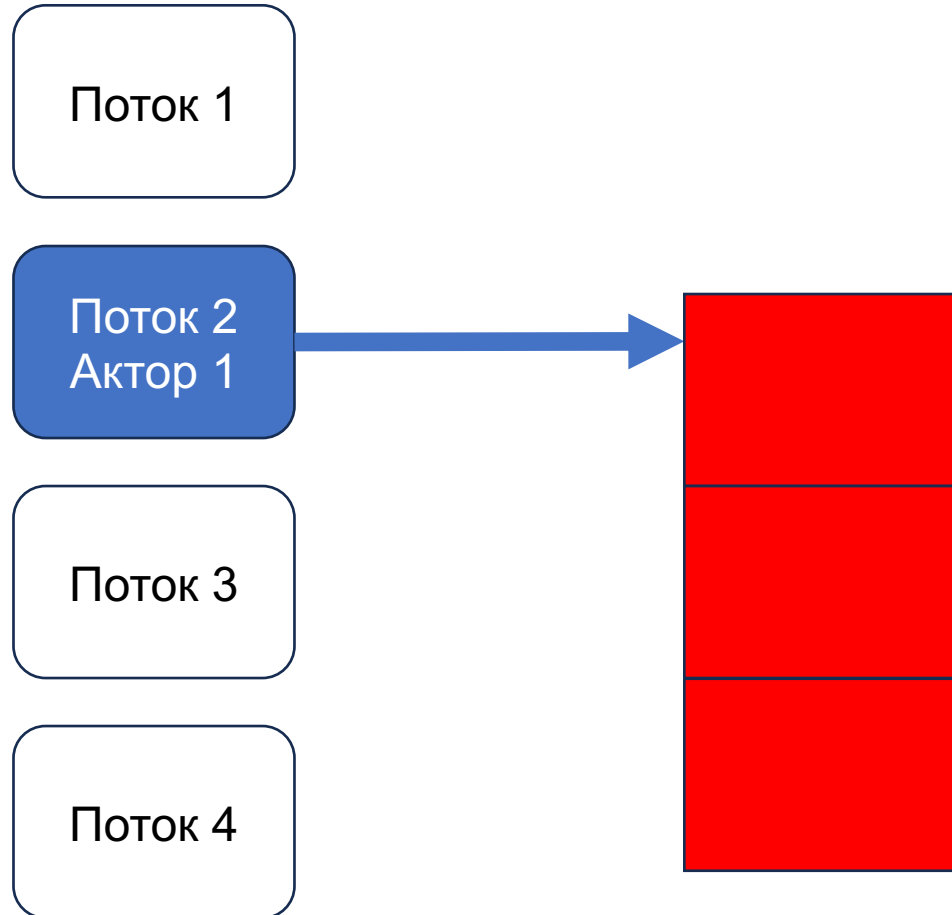
Описание бенчмарка



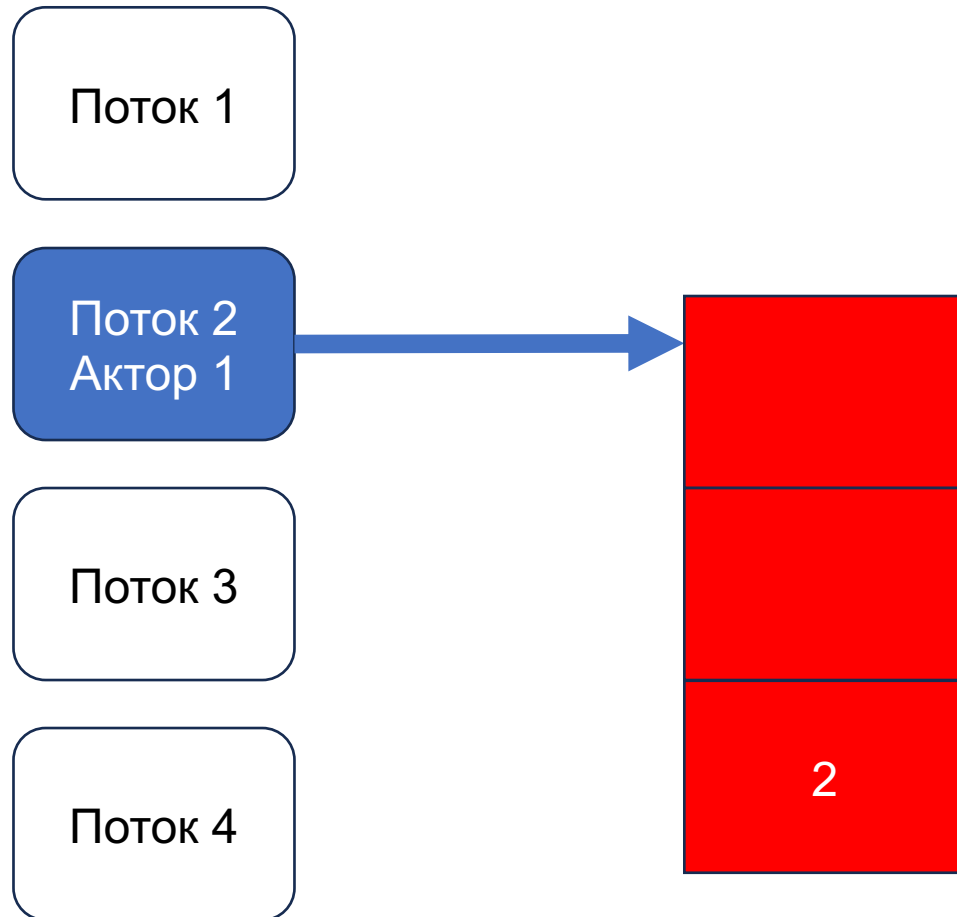
Описание бенчмарка



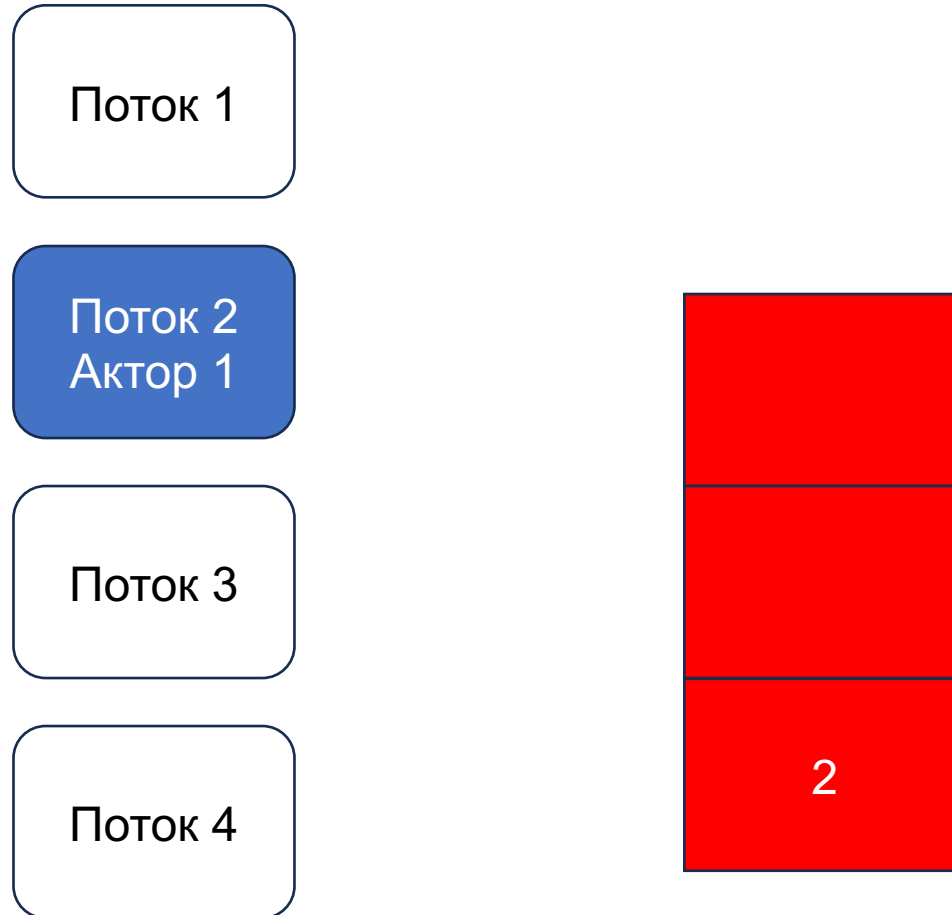
Описание бенчмарка



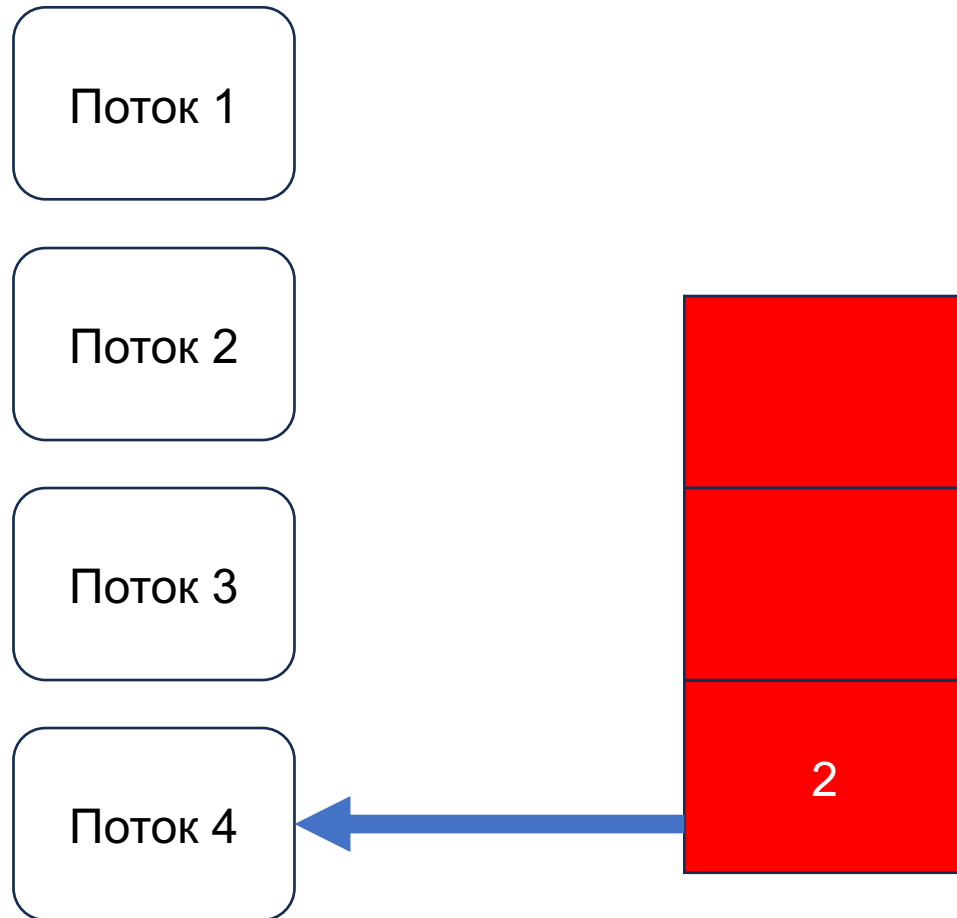
Описание бенчмарка



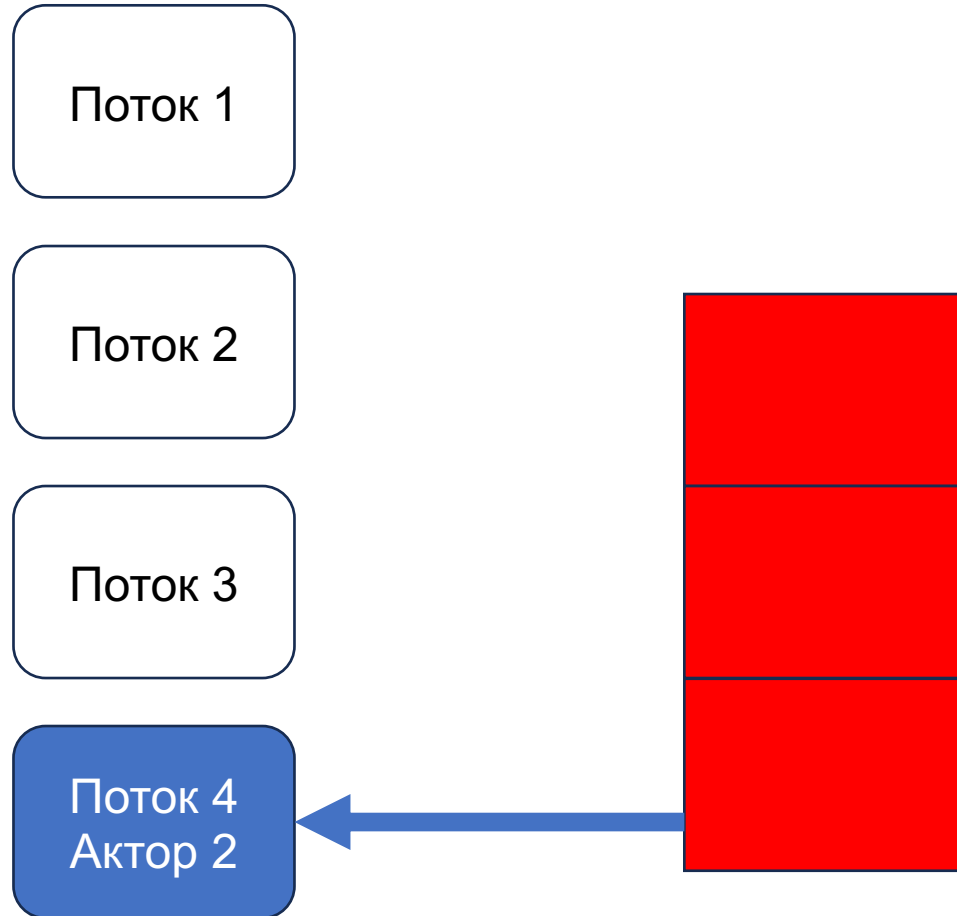
Описание бенчмарка



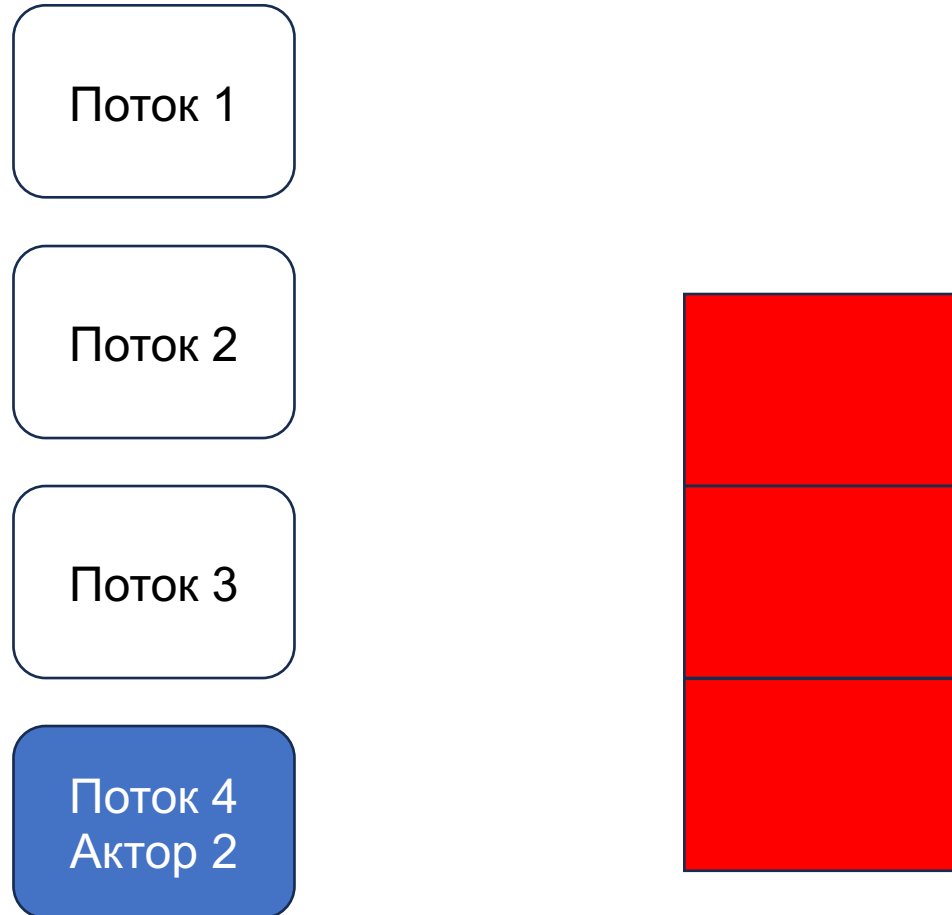
Описание бенчмарка



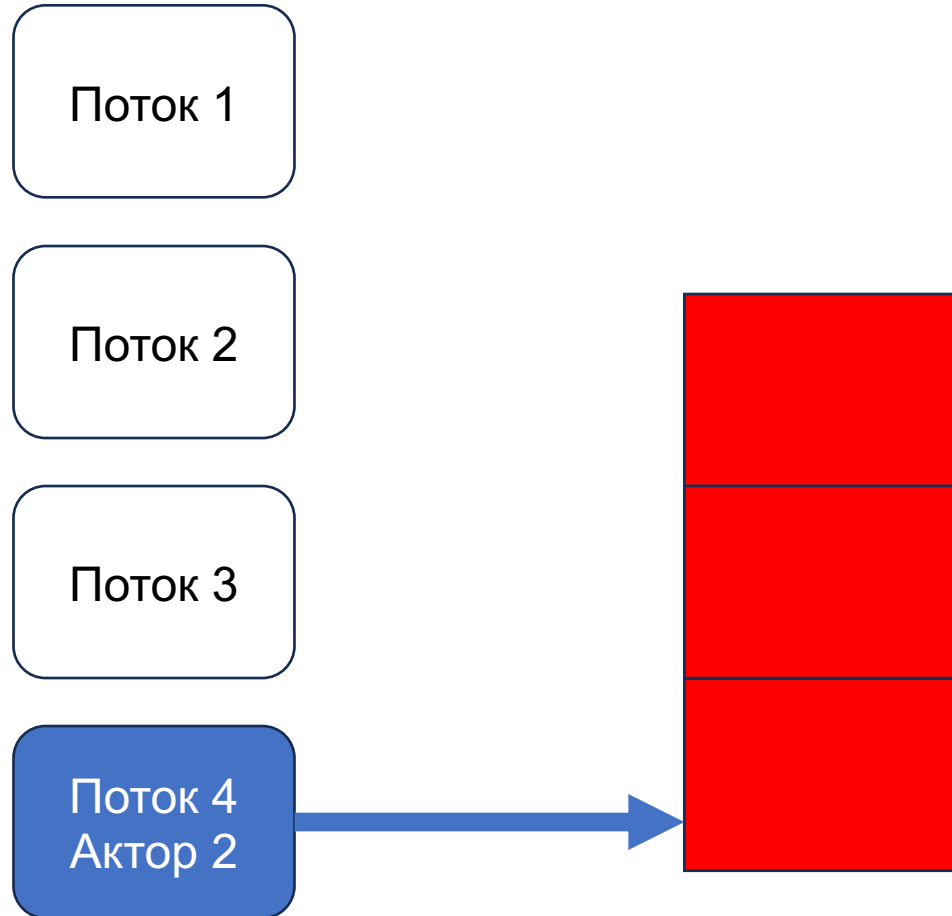
Описание бенчмарка



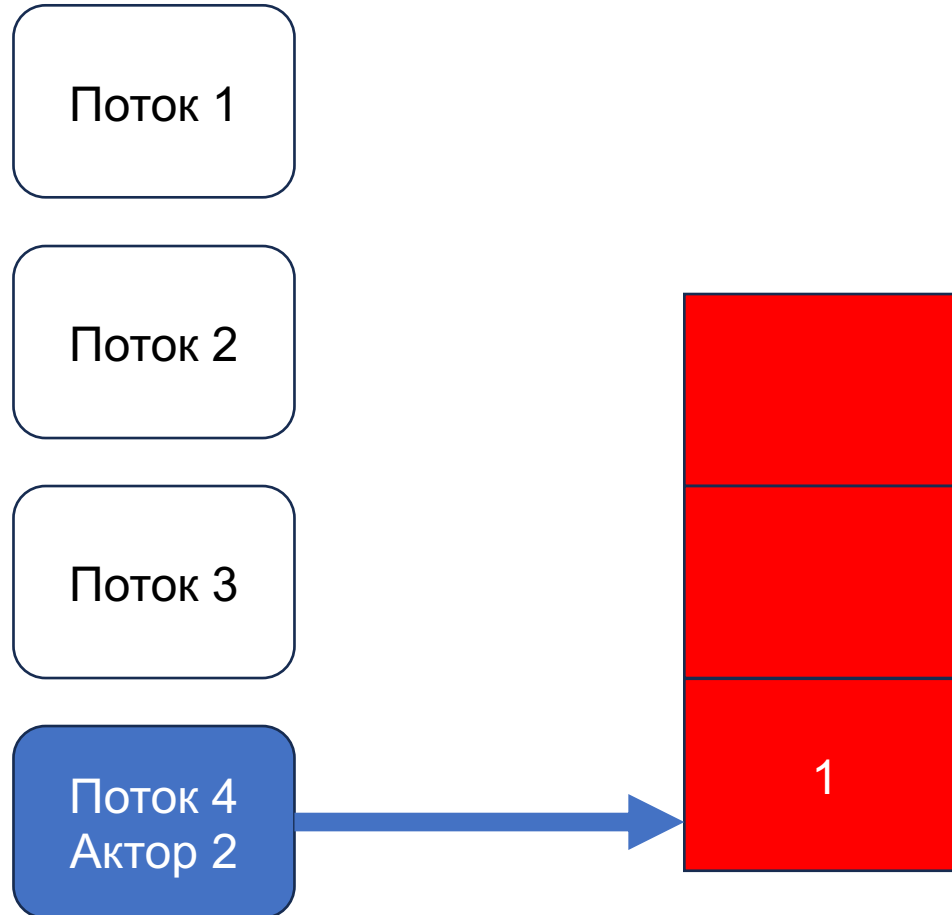
Описание бенчмарка



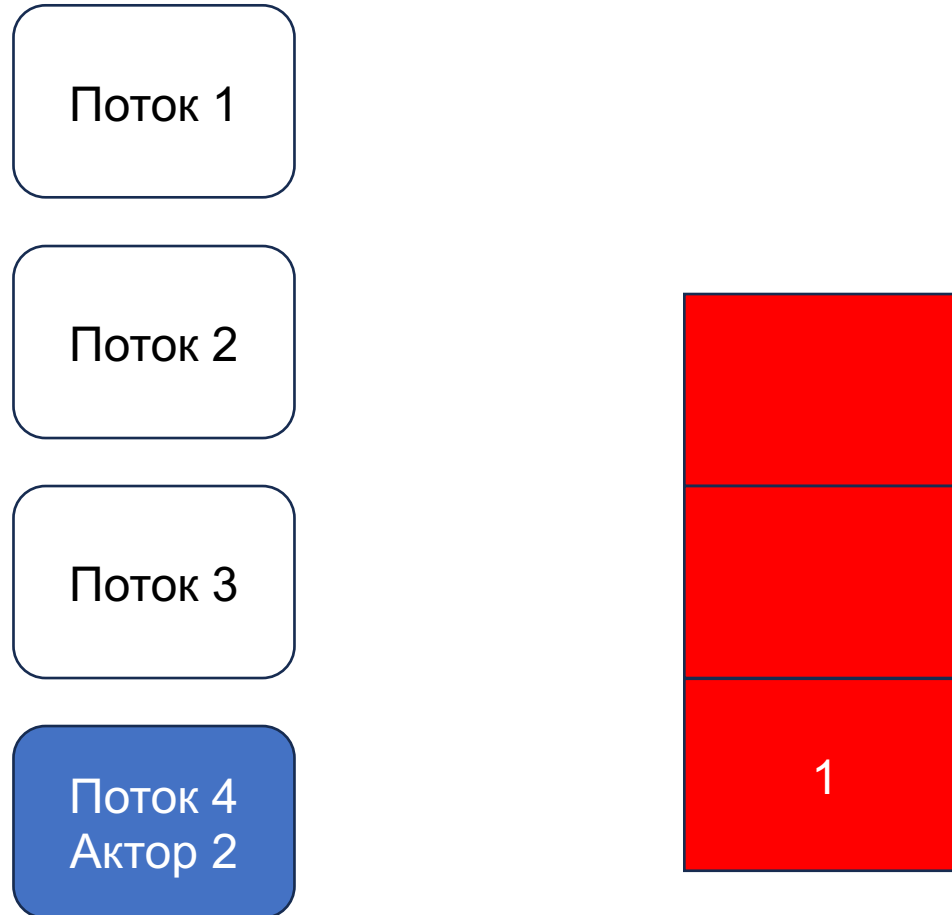
Описание бенчмарка



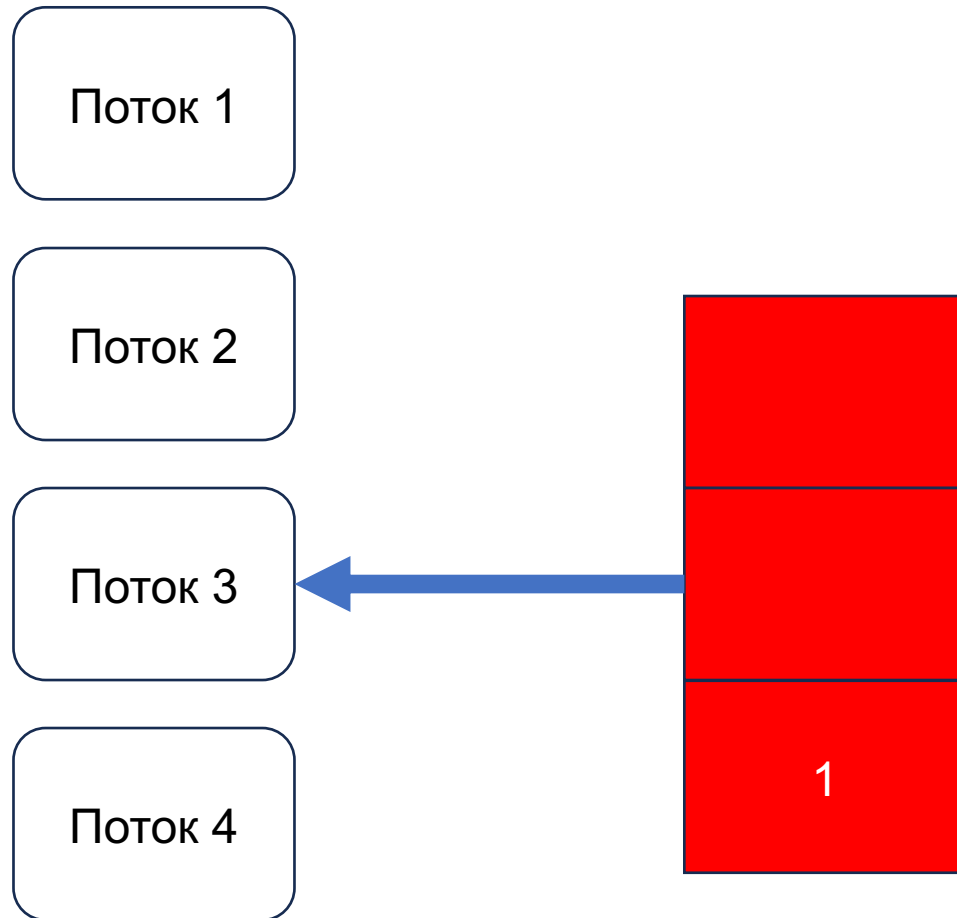
Описание бенчмарка



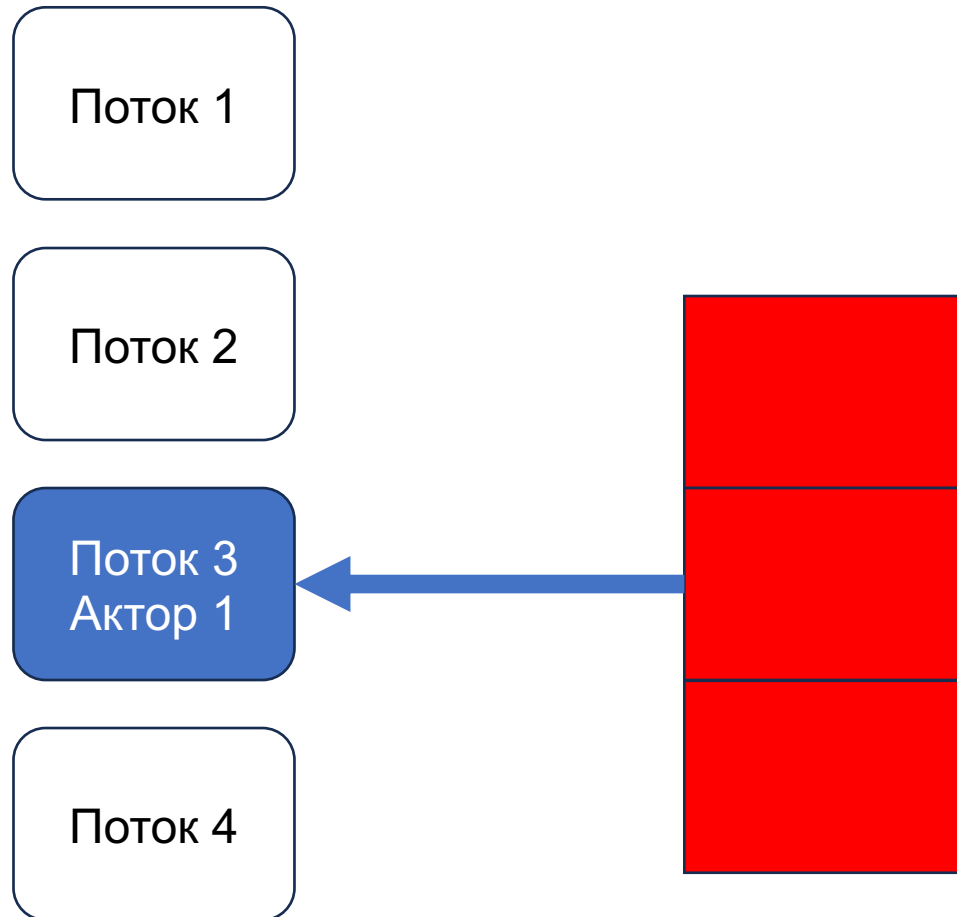
Описание бенчмарка



Описание бенчмарка



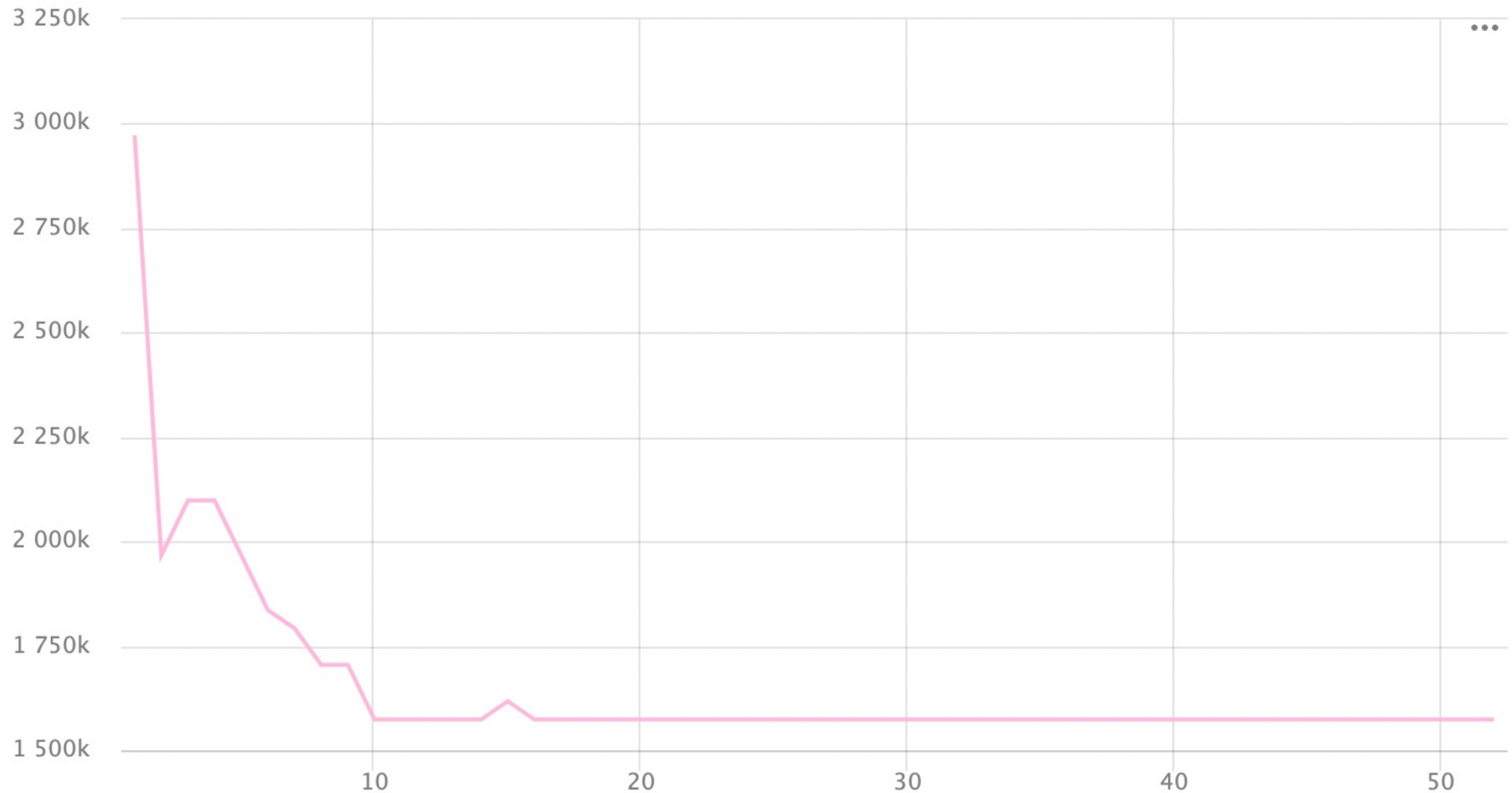
Описание бенчмарка



Описание бенчмарка

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

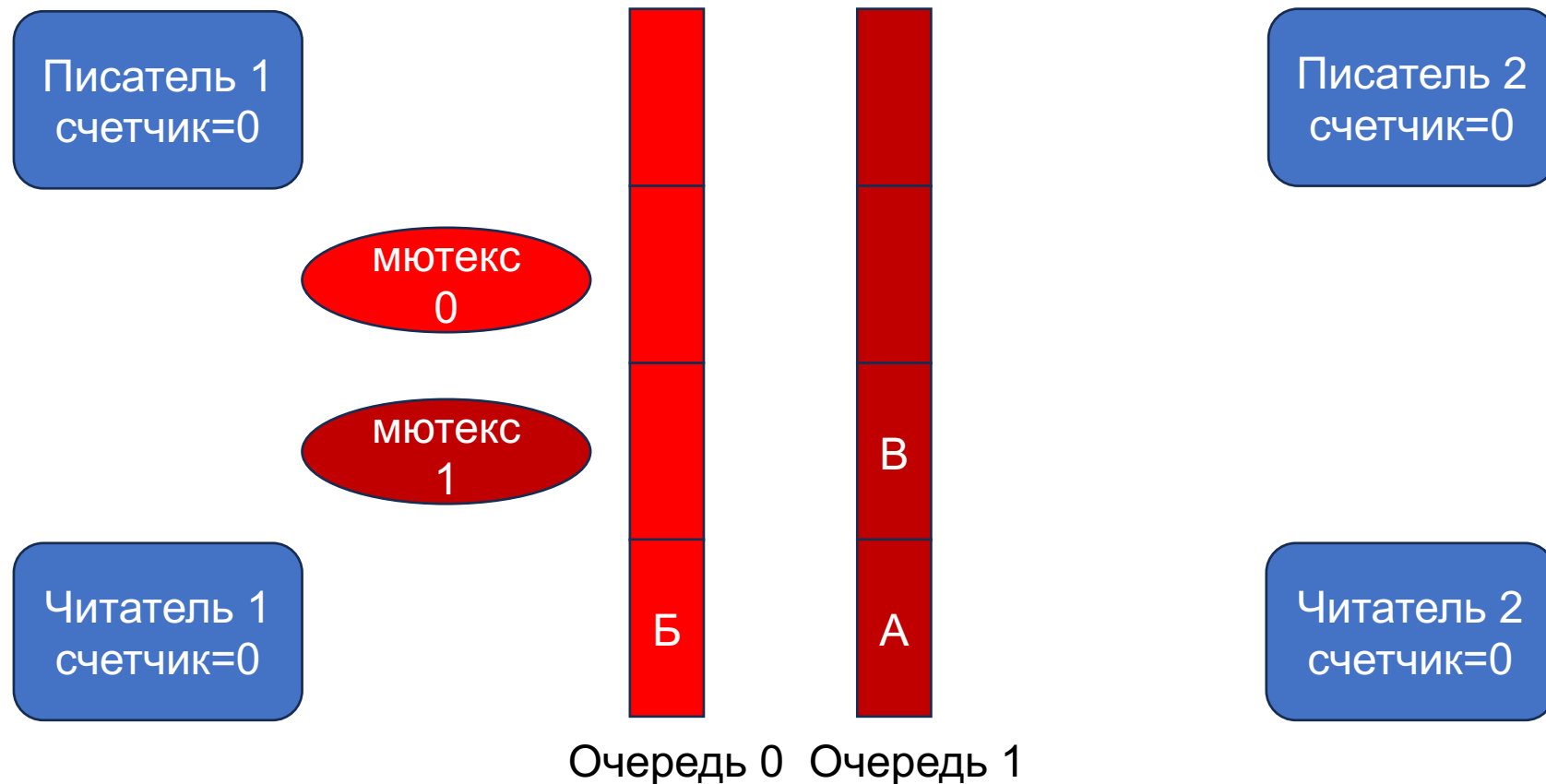
Примитивная МРМС очередь с мьютексом



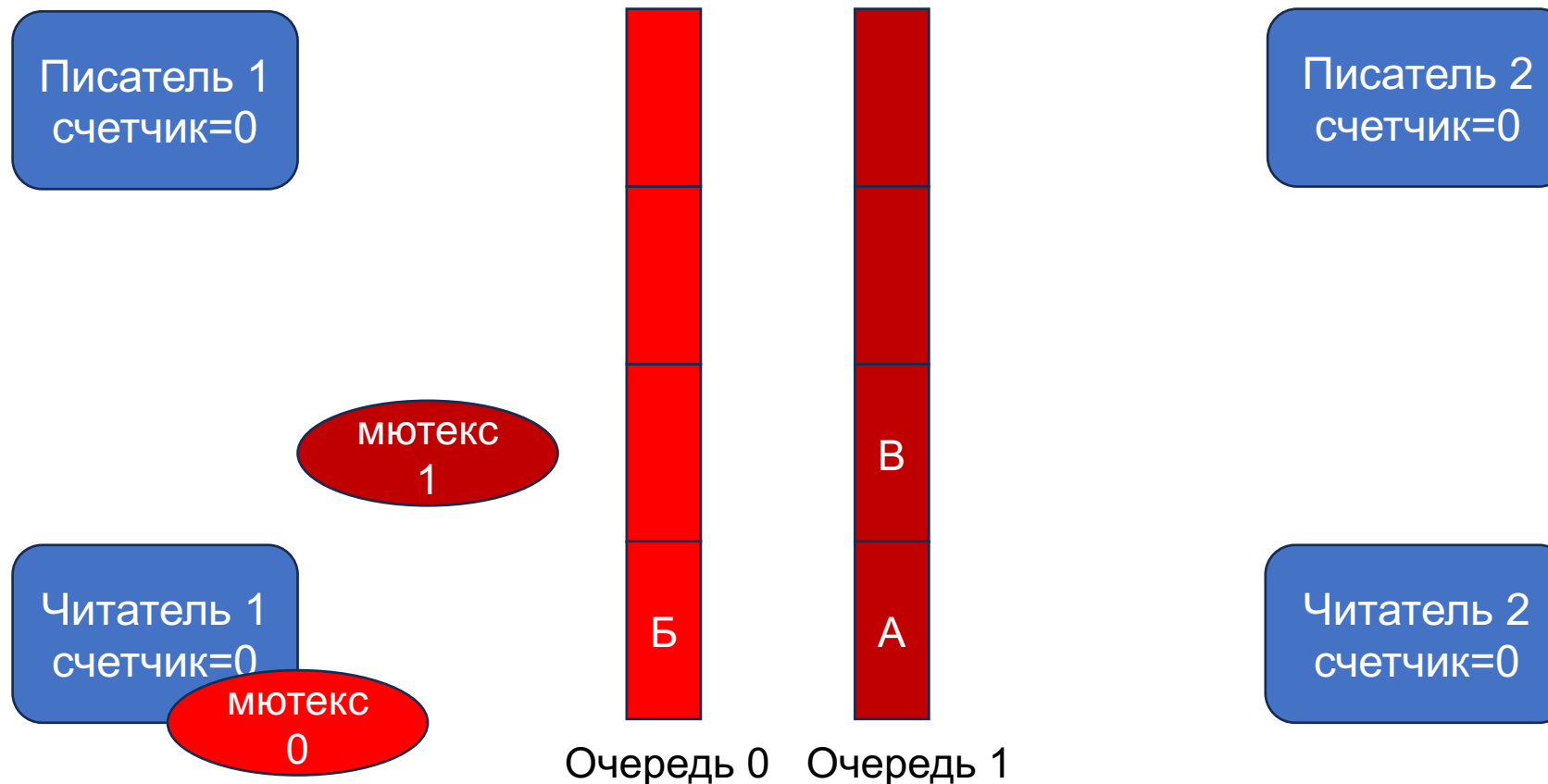
Исходное состояние до оптимизации

Revolving MPMC queue

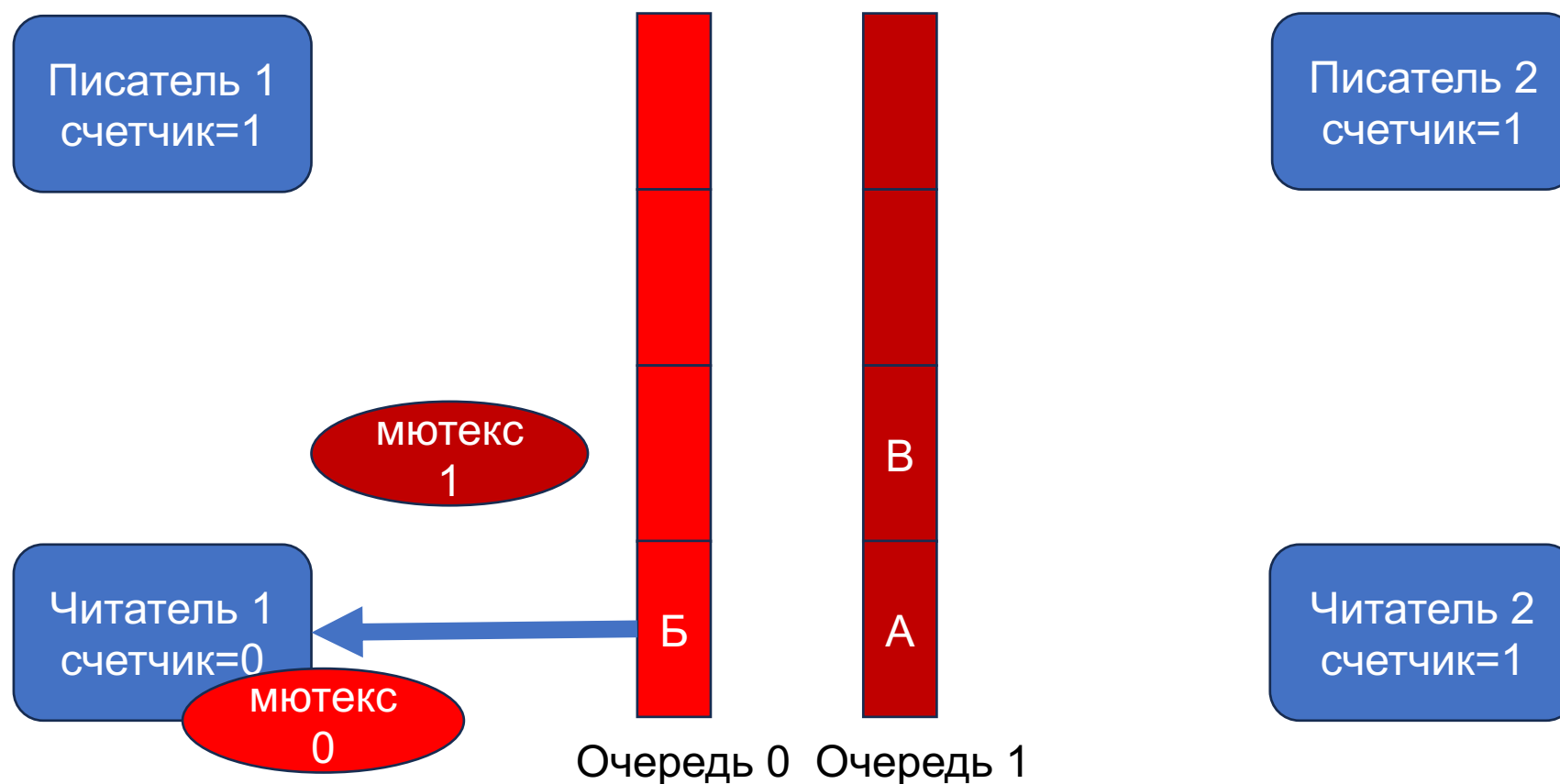
Исходное состояние до оптимизации: набор очередей с per-queue блокировками.



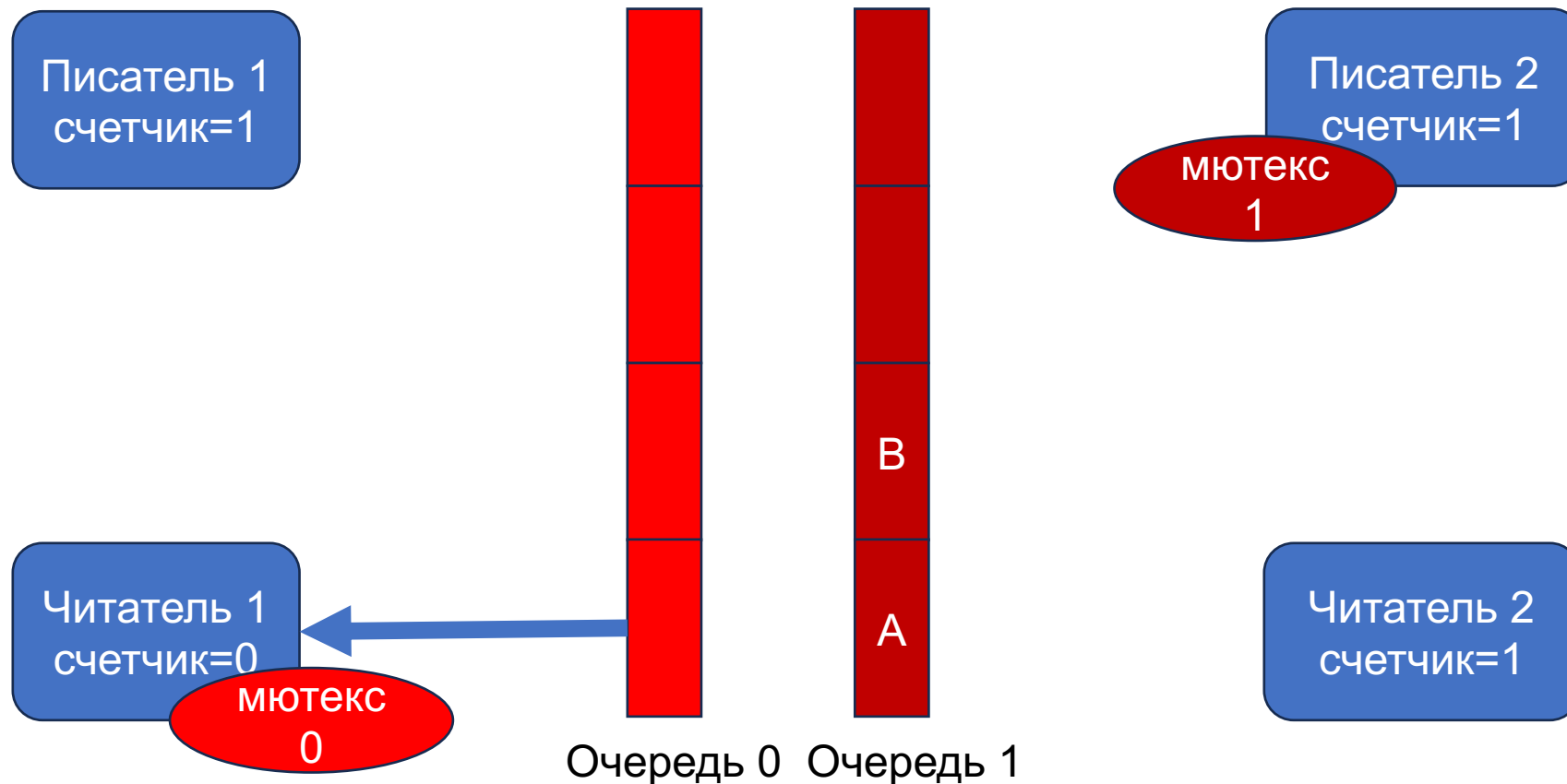
Revolving MPMC queue



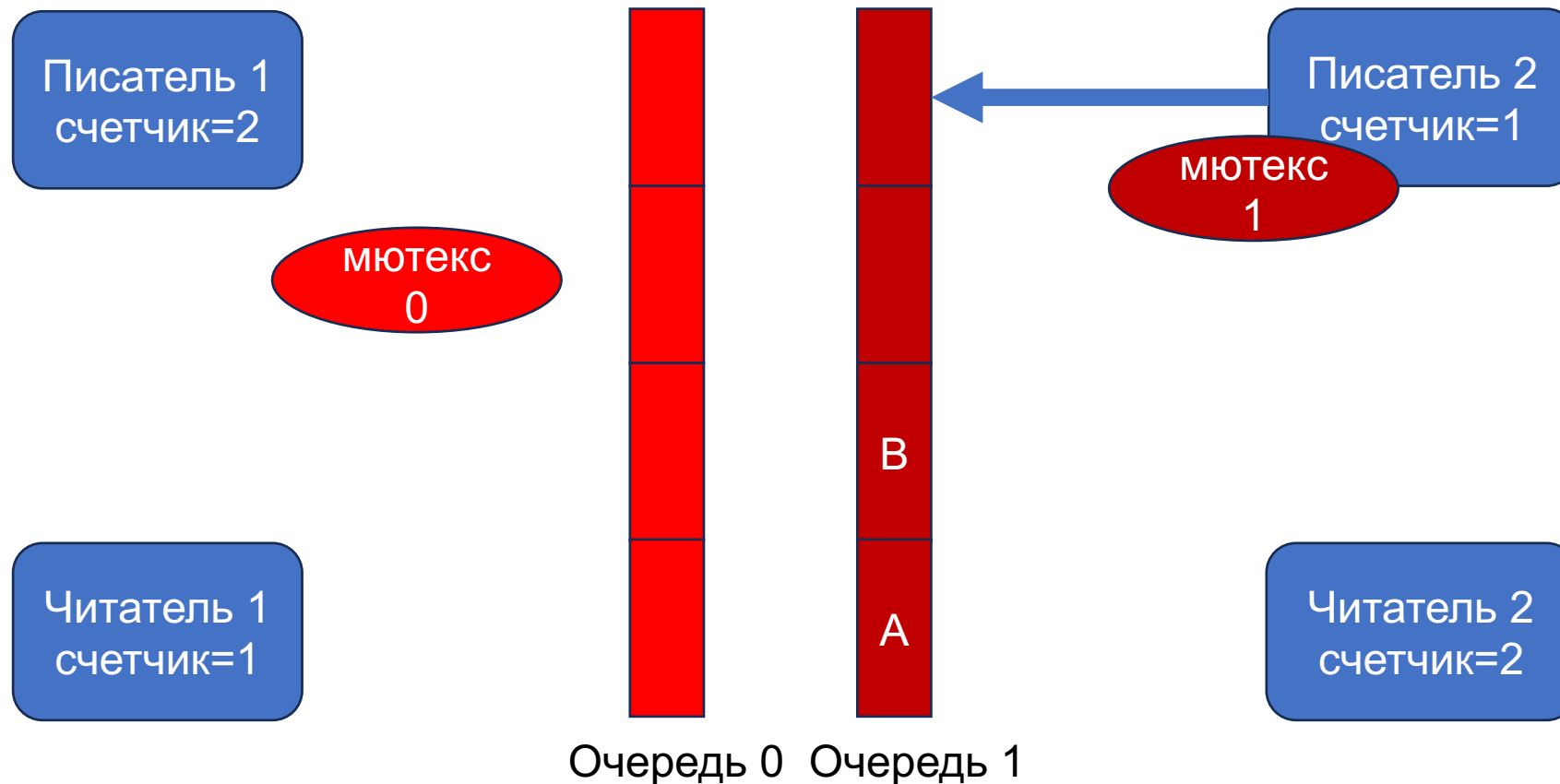
Revolving MPMC queue



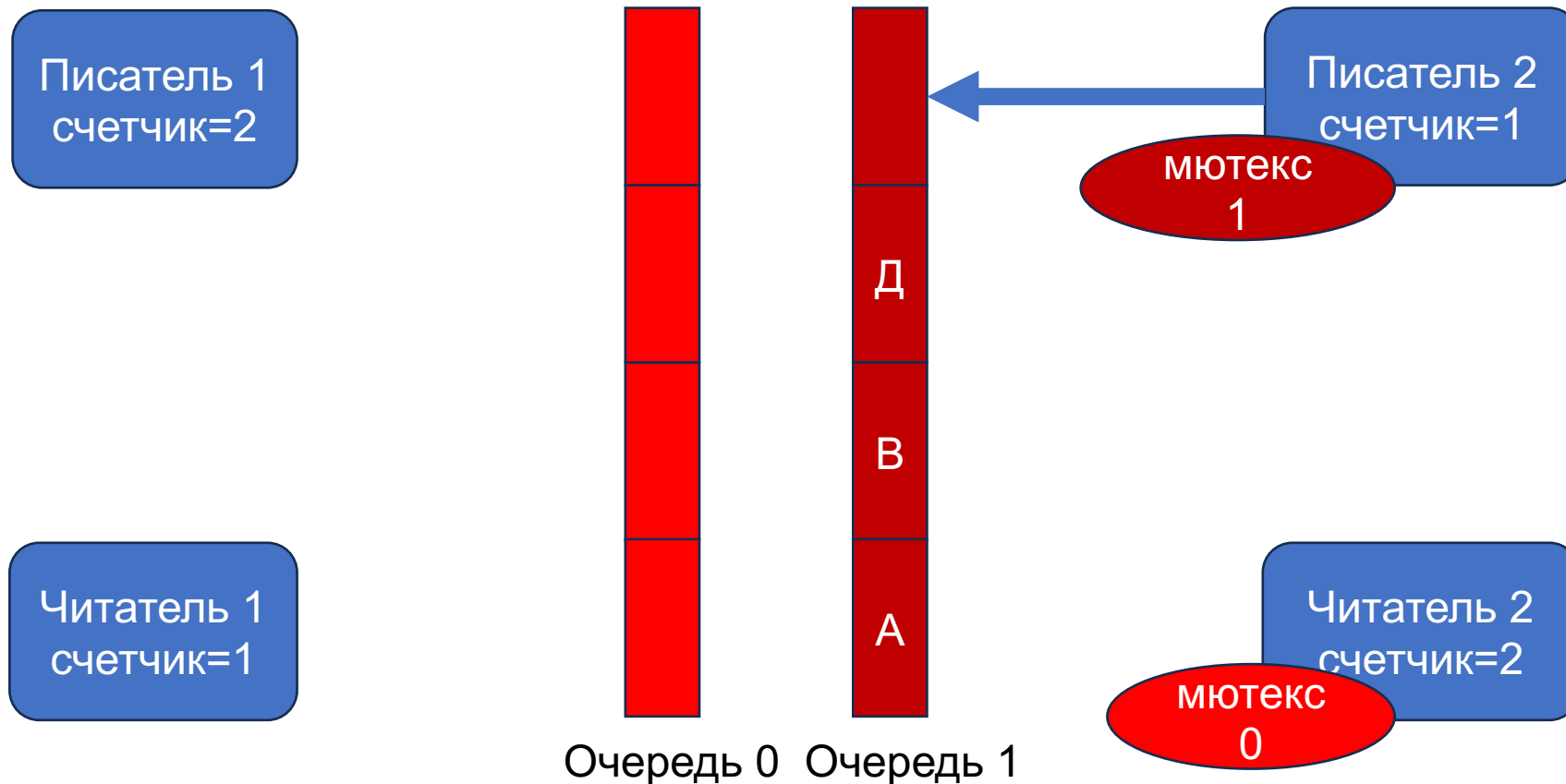
Revolving MPMC queue



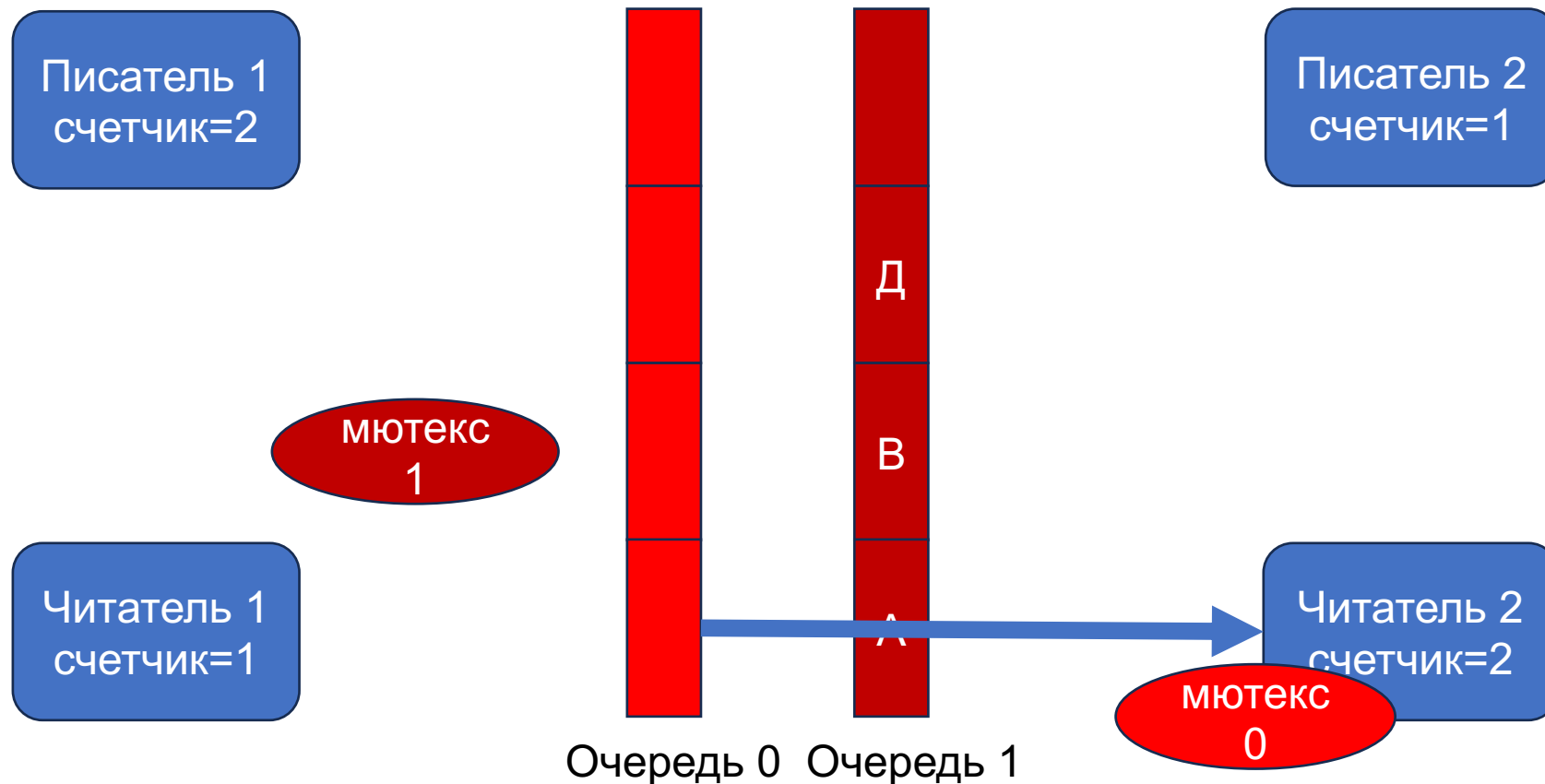
Revolving MPMC queue



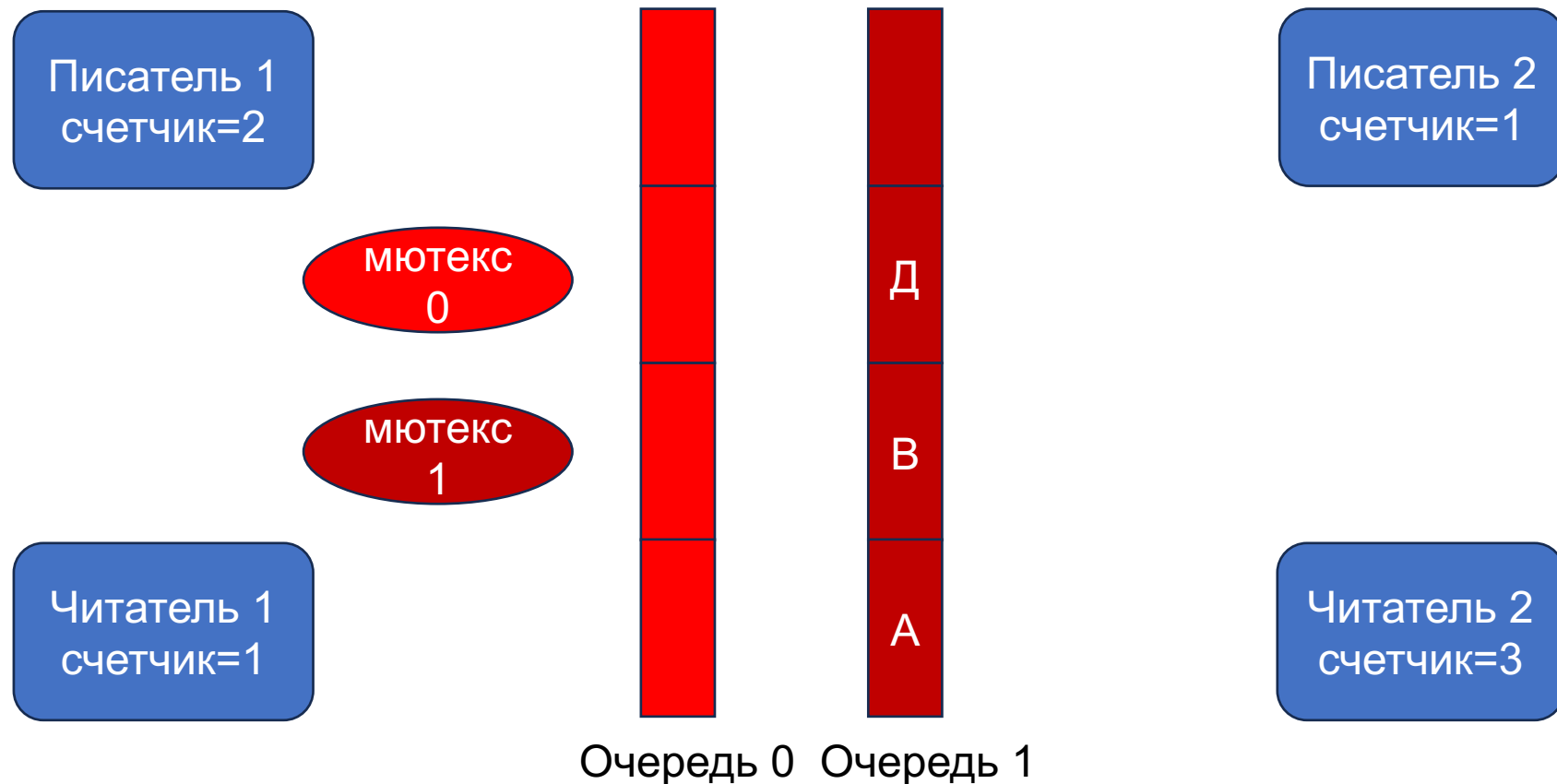
Revolving MPMC queue



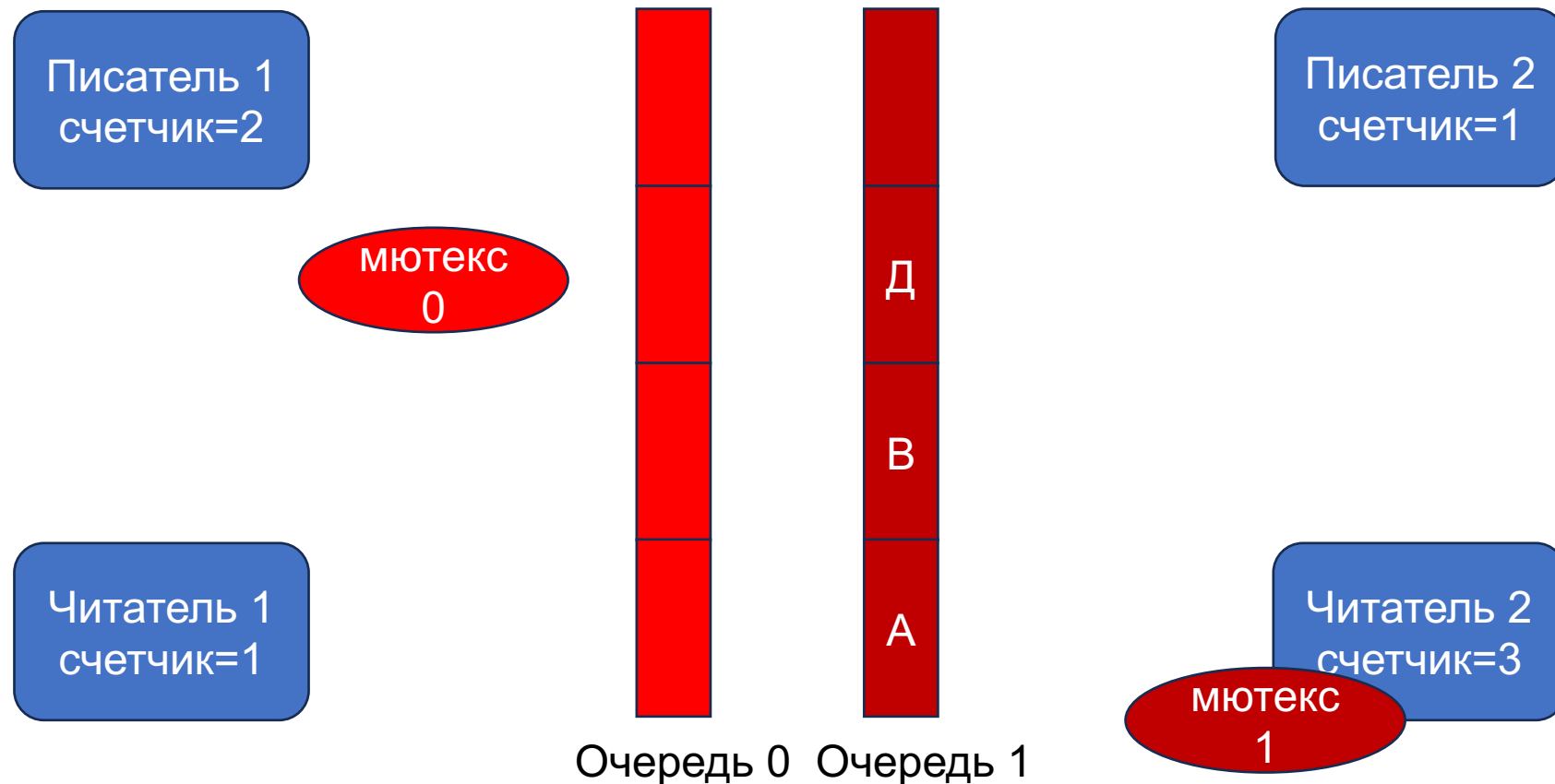
Revolving MPMC queue



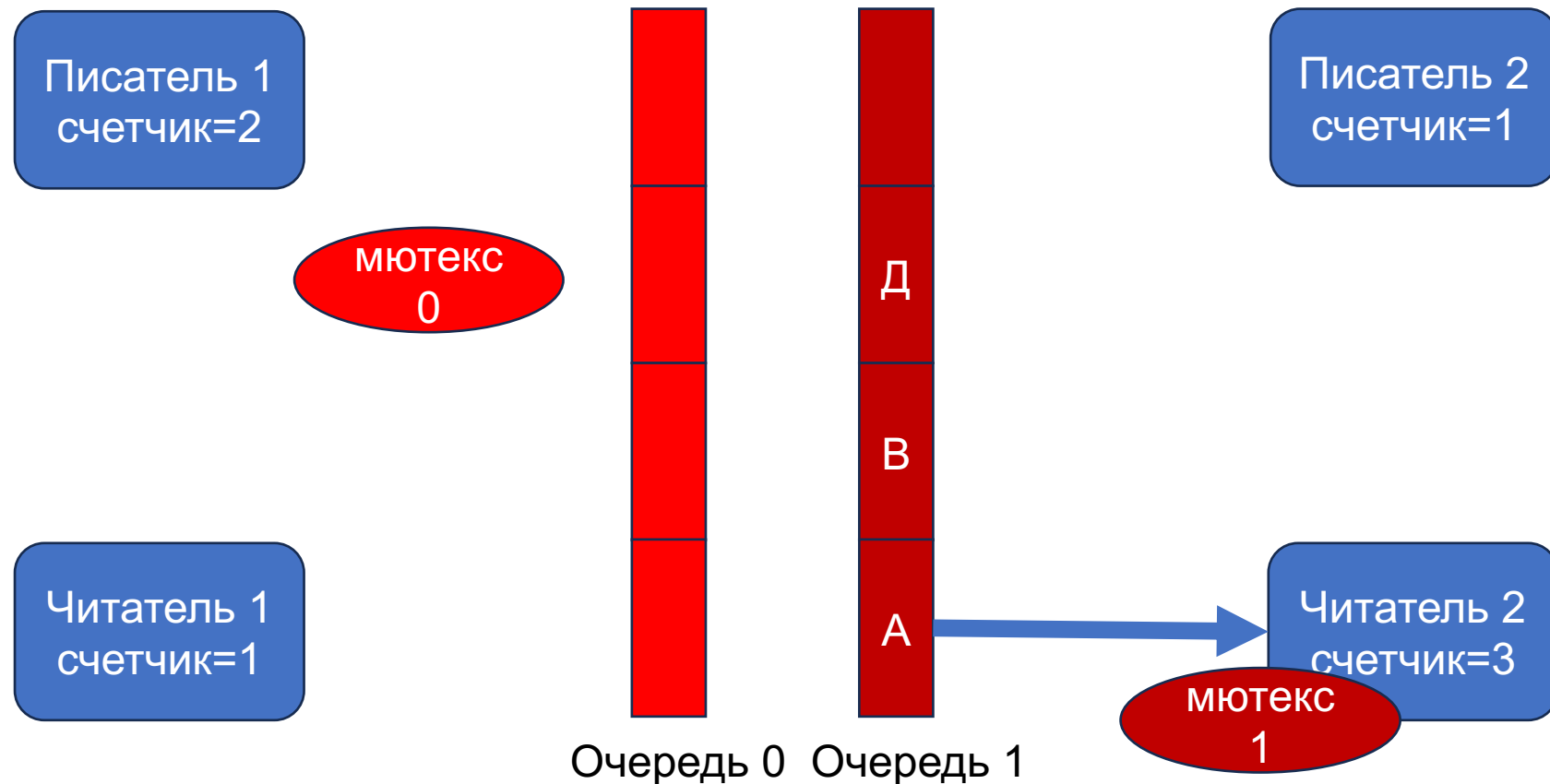
Revolving MPMC queue



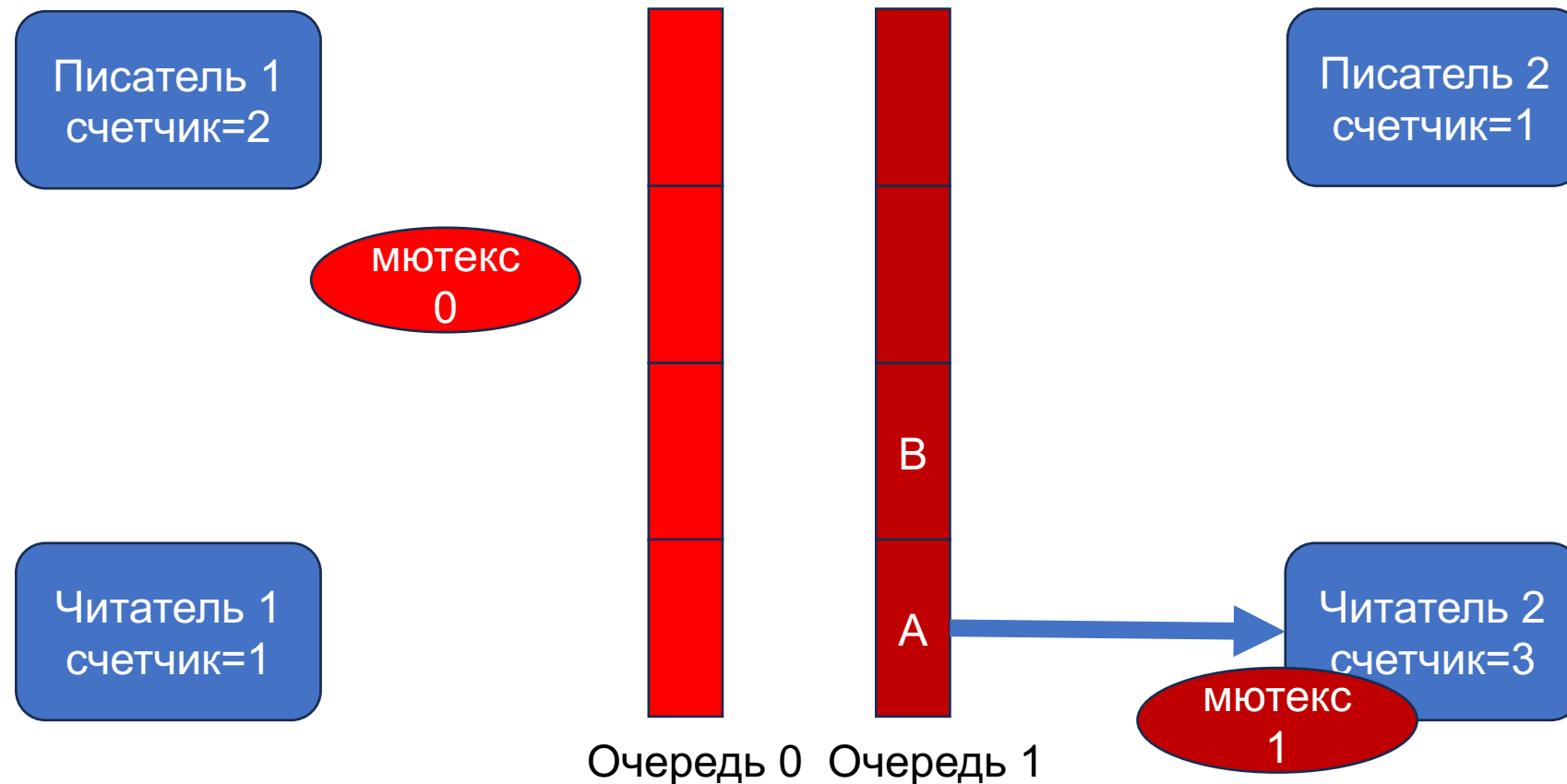
Revolving MPMC queue



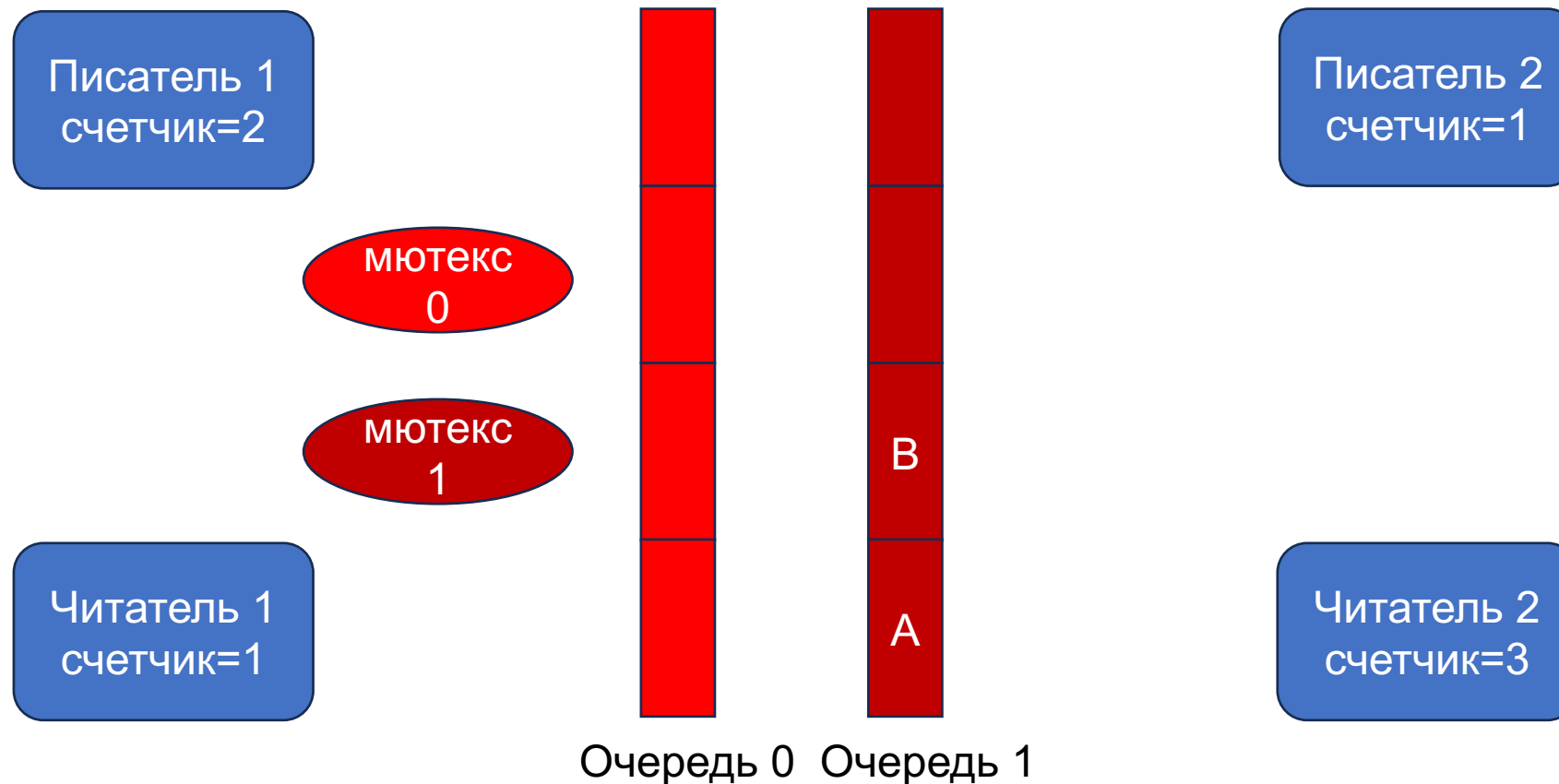
Revolving MPMC queue



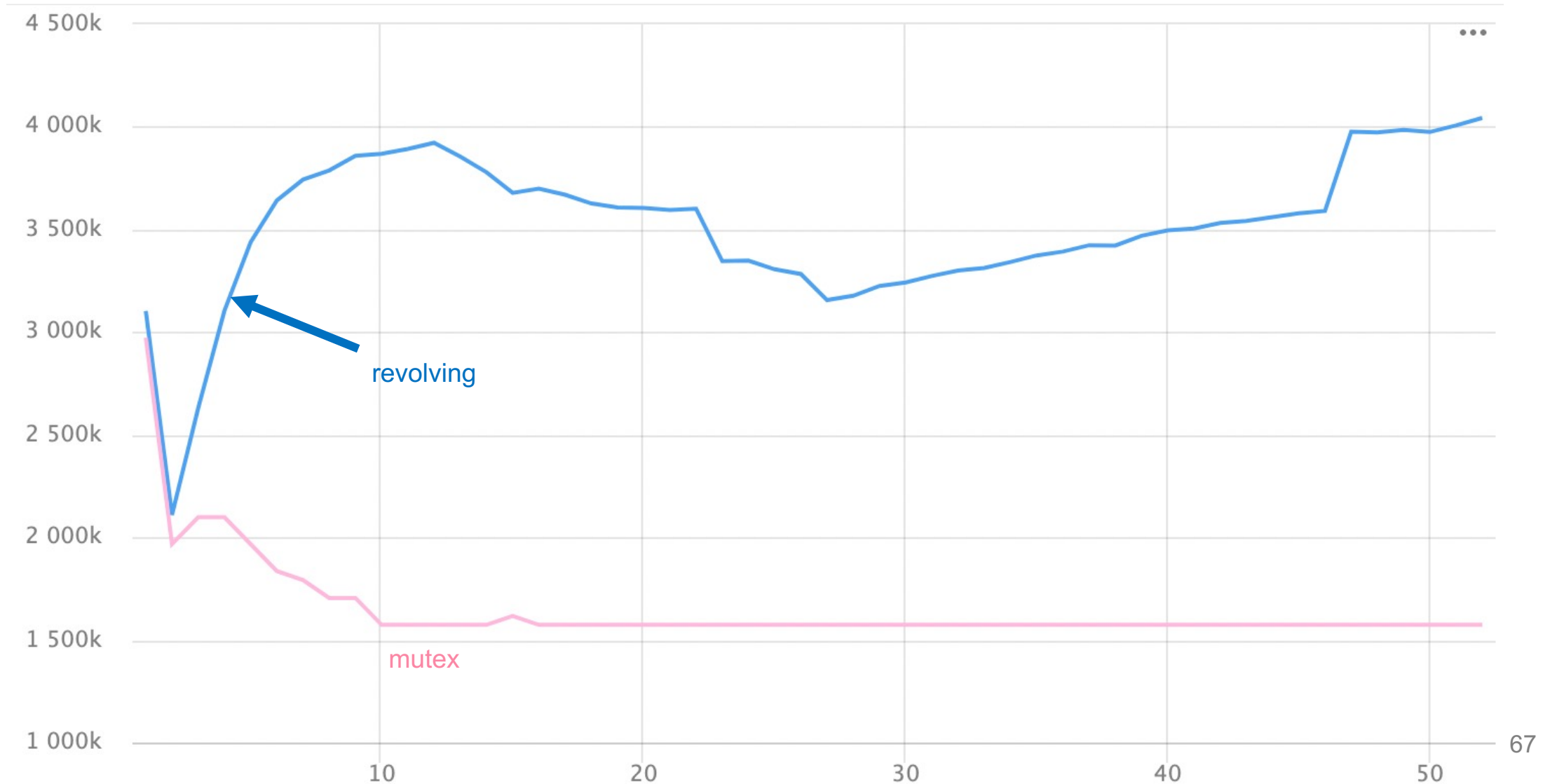
Revolving MPMC queue



Revolving MPMC queue

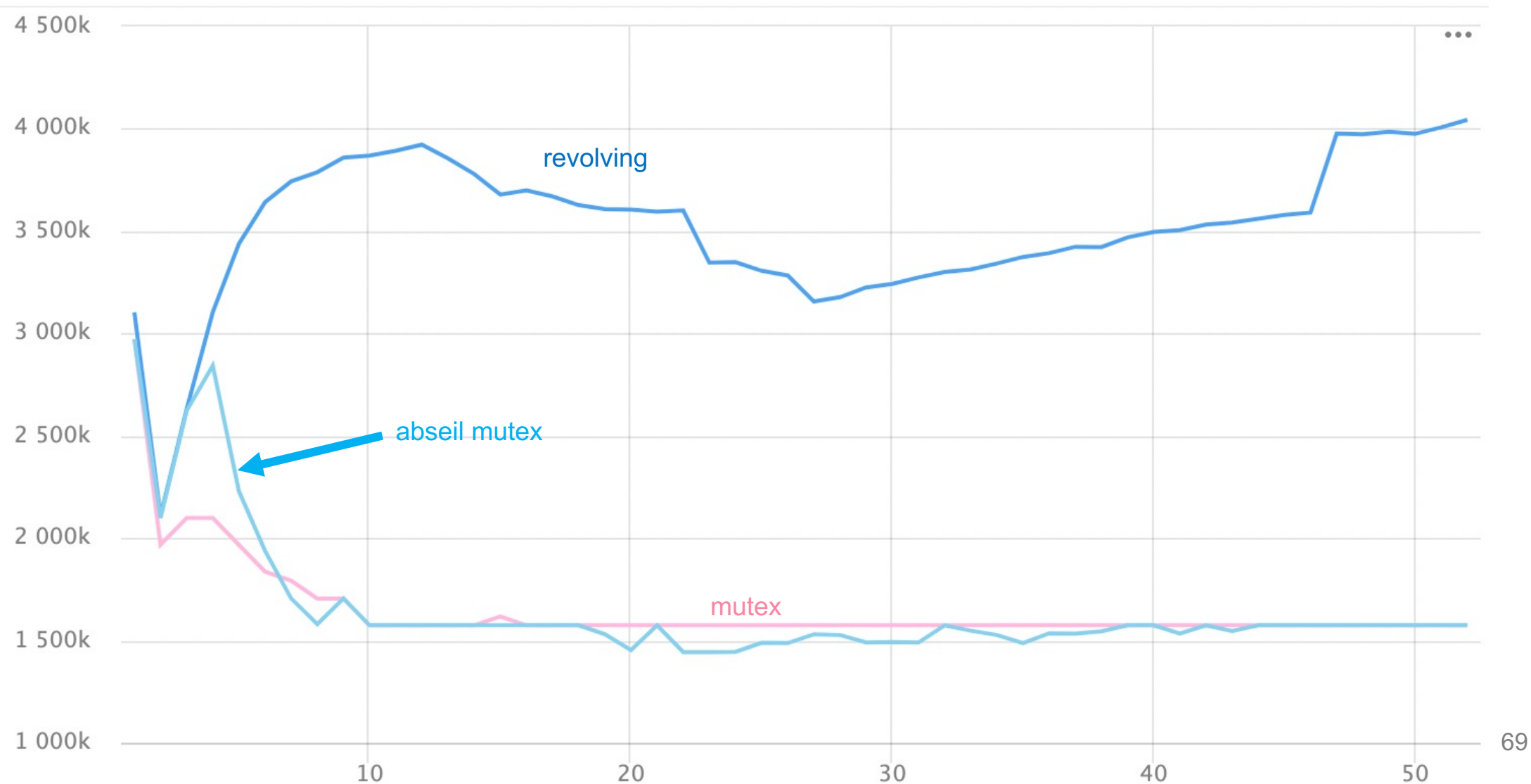


Revolving MPMC queue



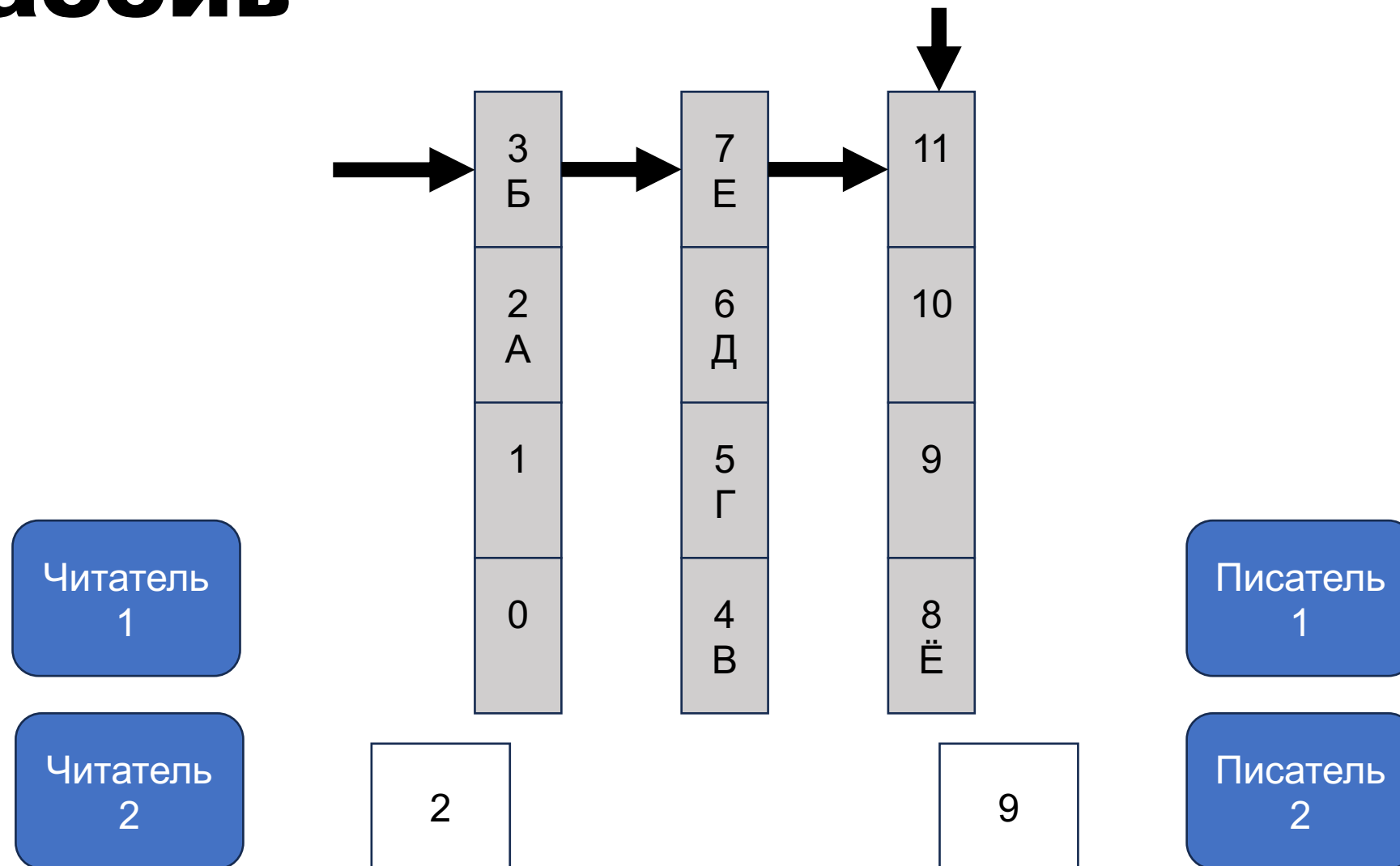
Быстрый mutex из abseil

Быстрый mutex из abseil

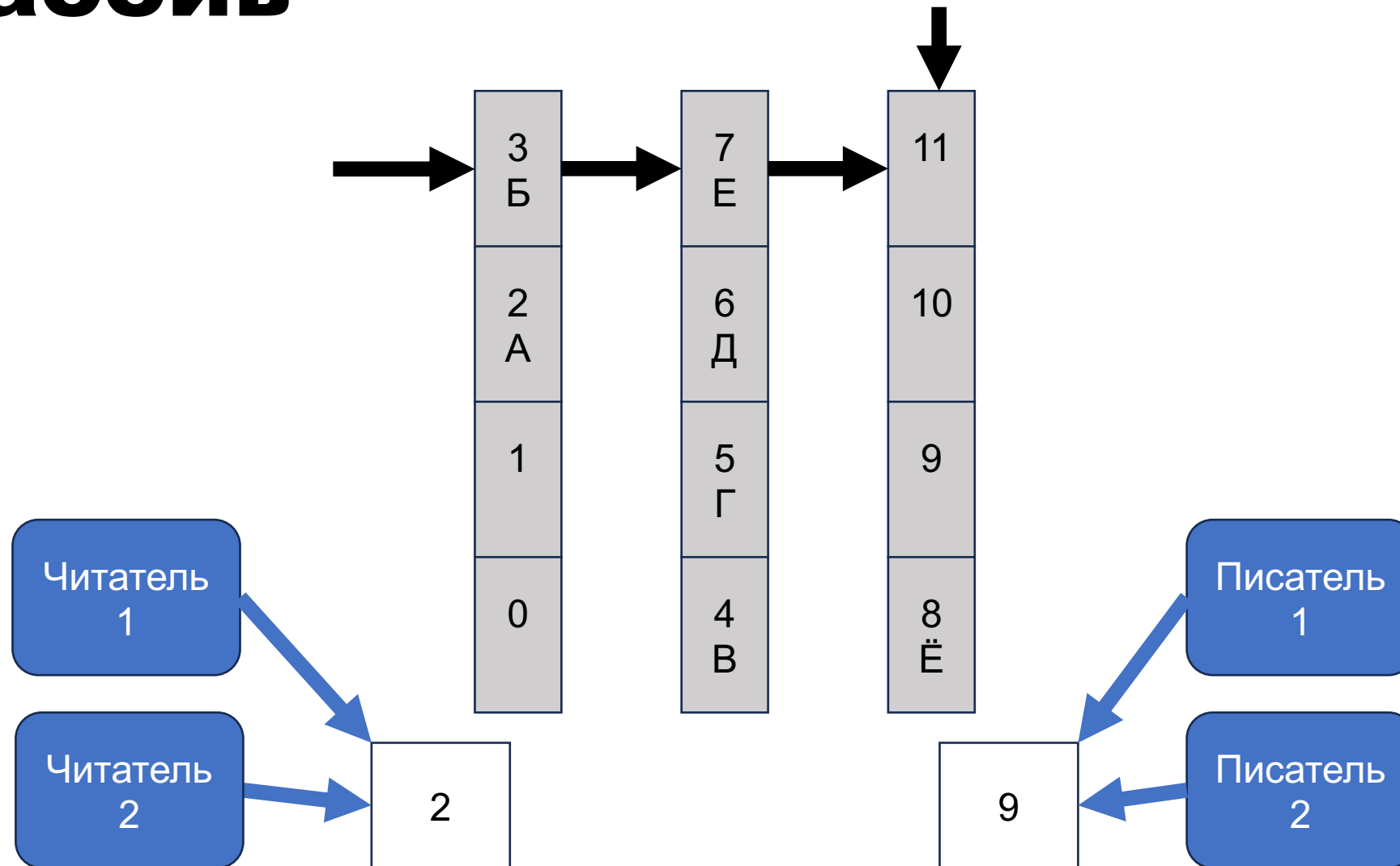


Виртуальный бесконечный массив

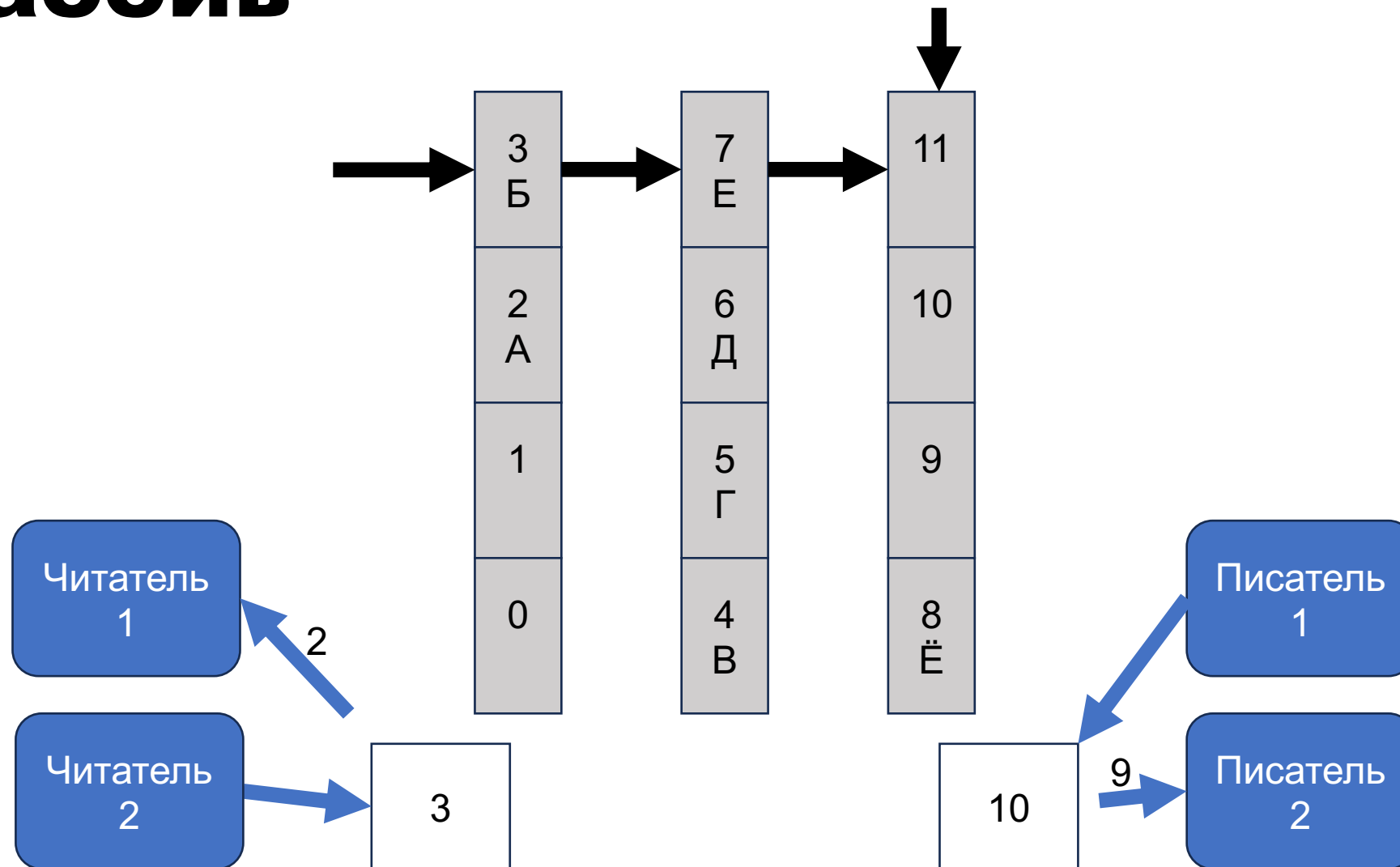
Виртуальный бесконечный массив



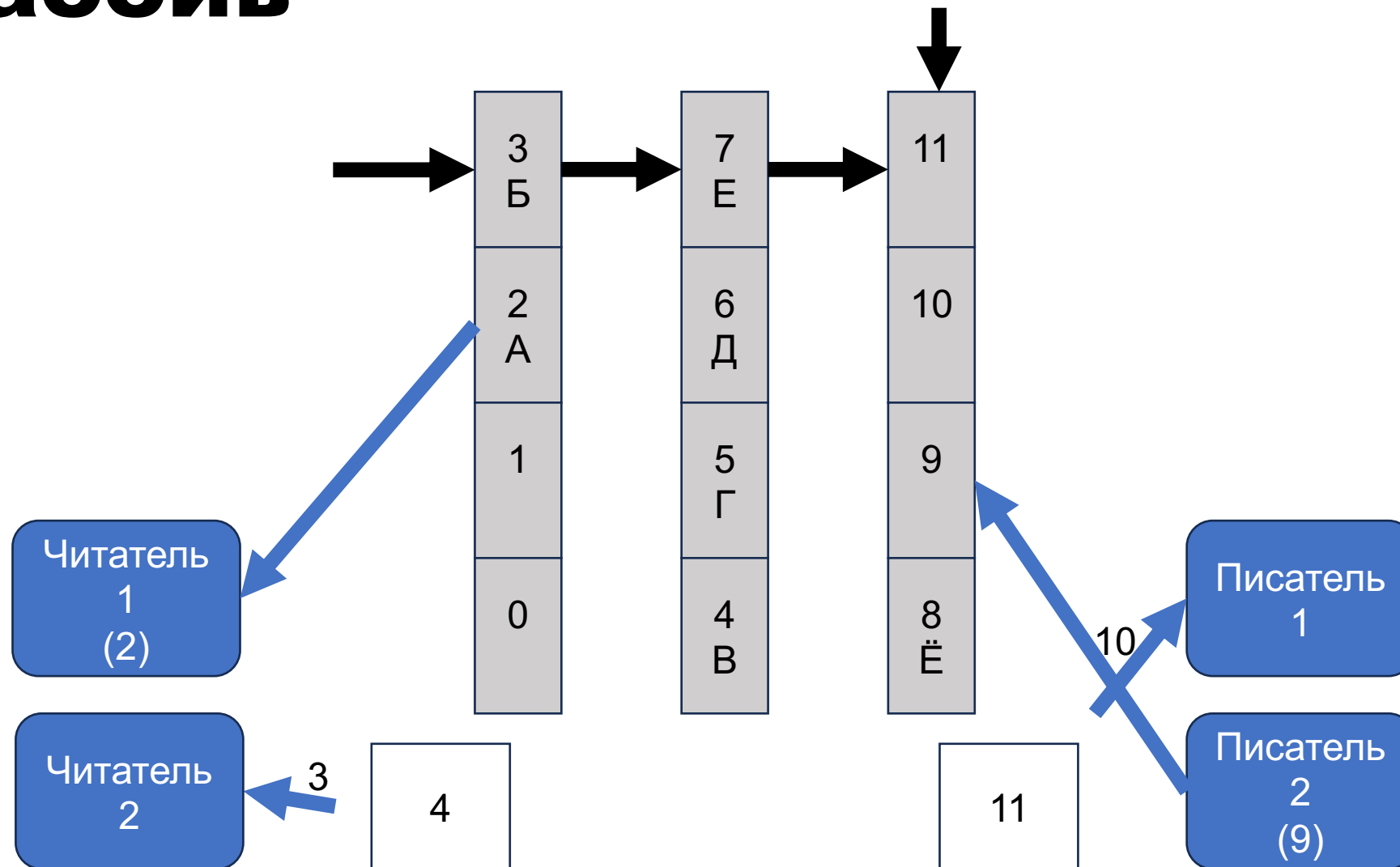
Виртуальный бесконечный массив



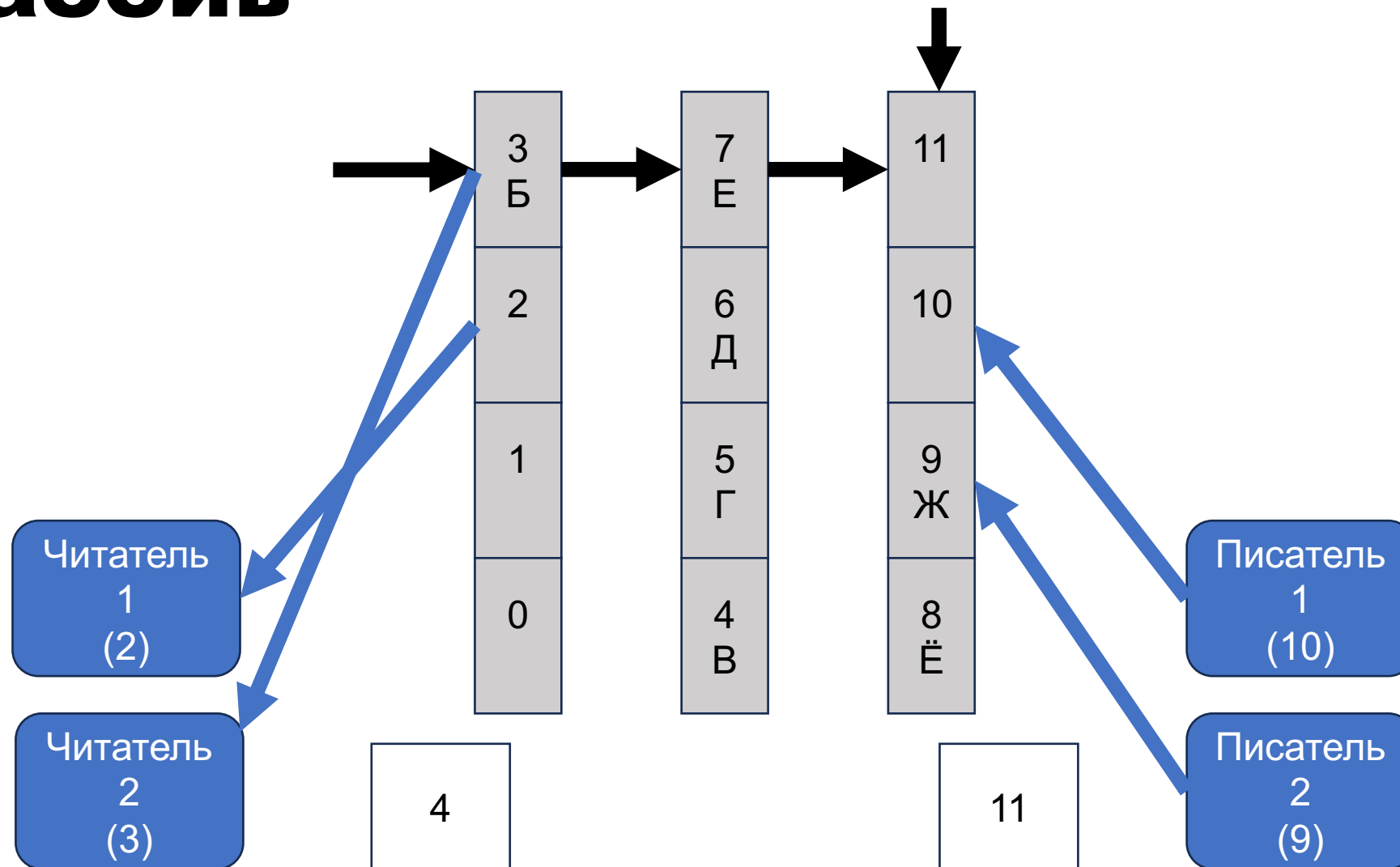
Виртуальный бесконечный массив



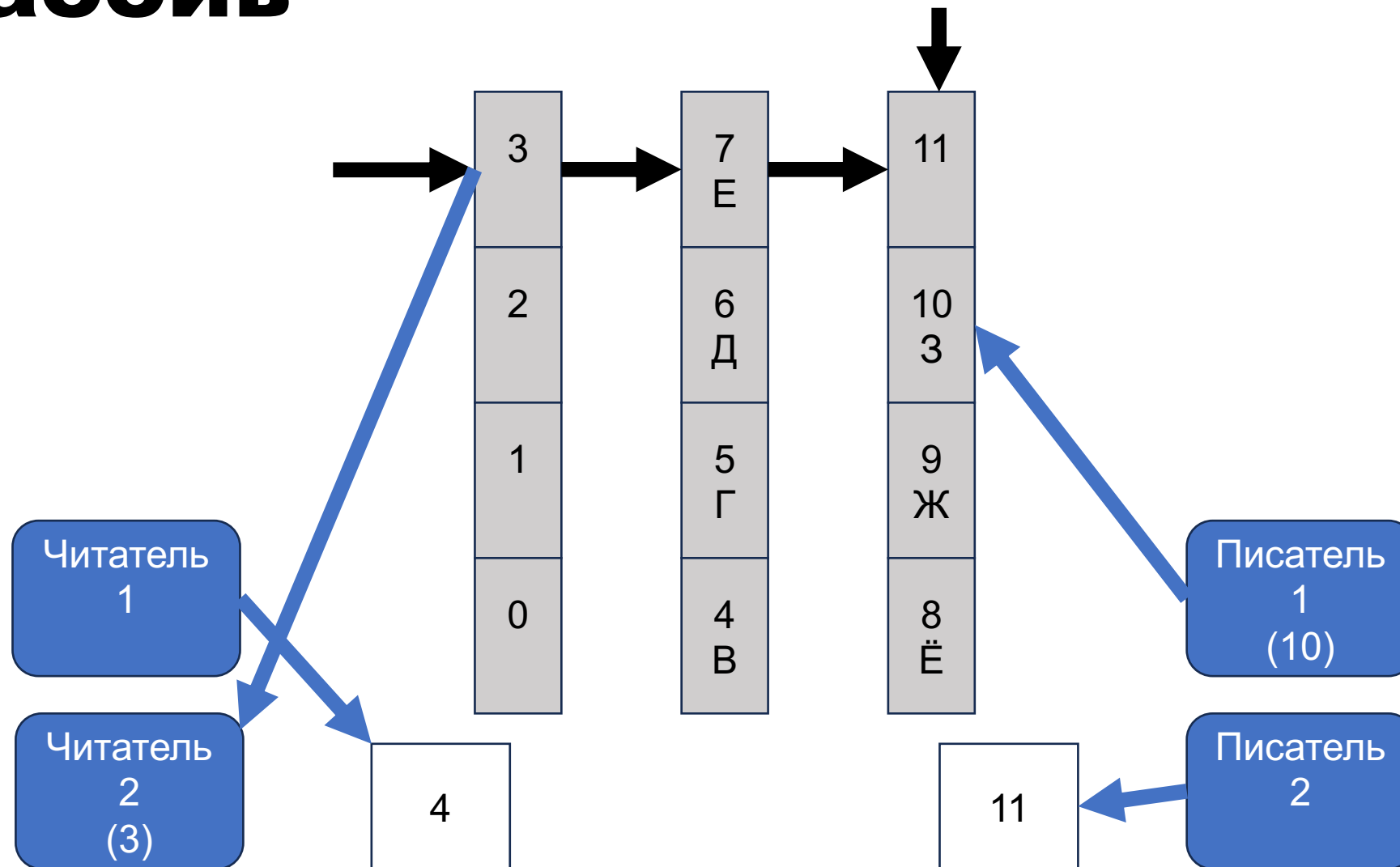
Виртуальный бесконечный массив



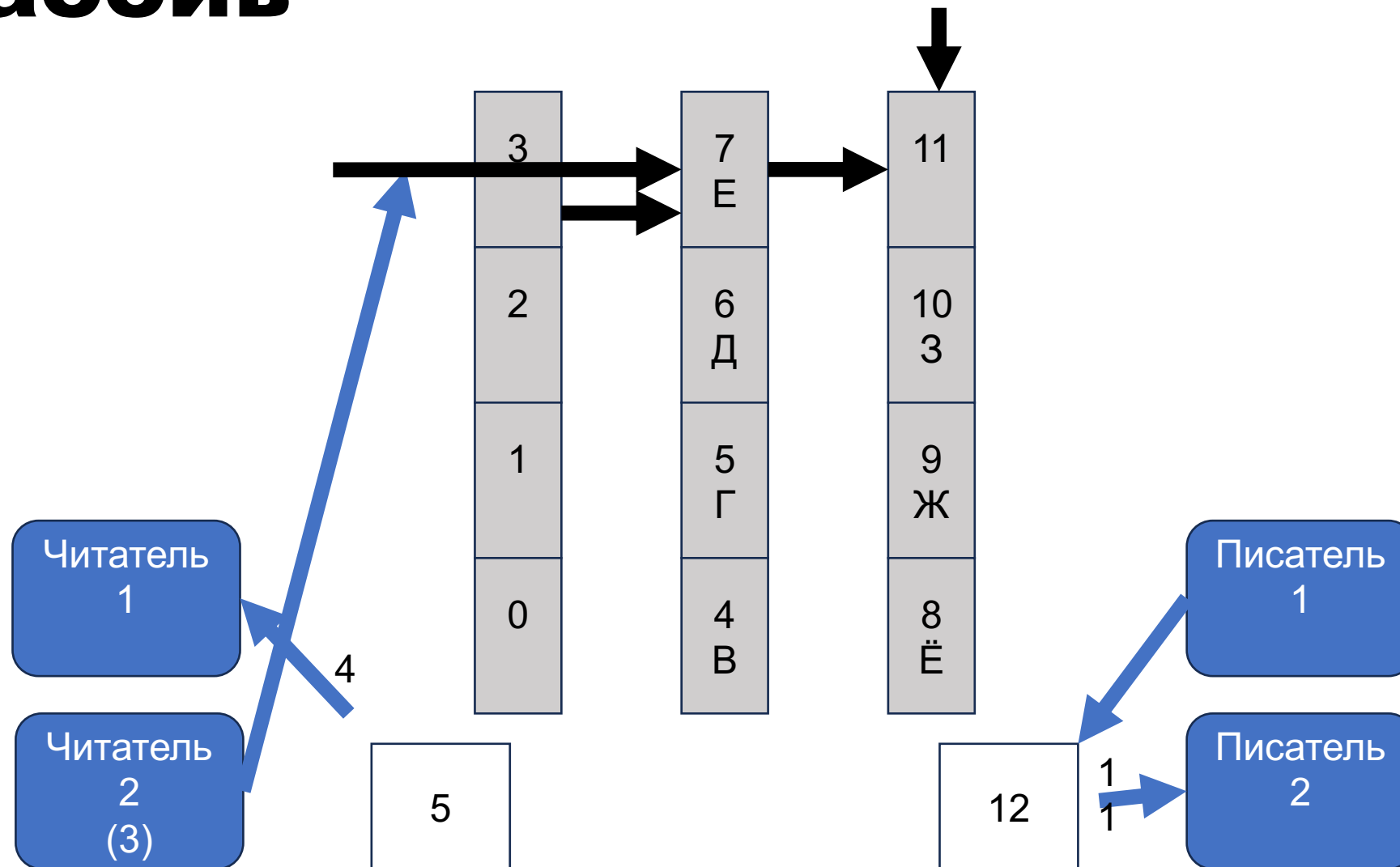
Виртуальный бесконечный массив



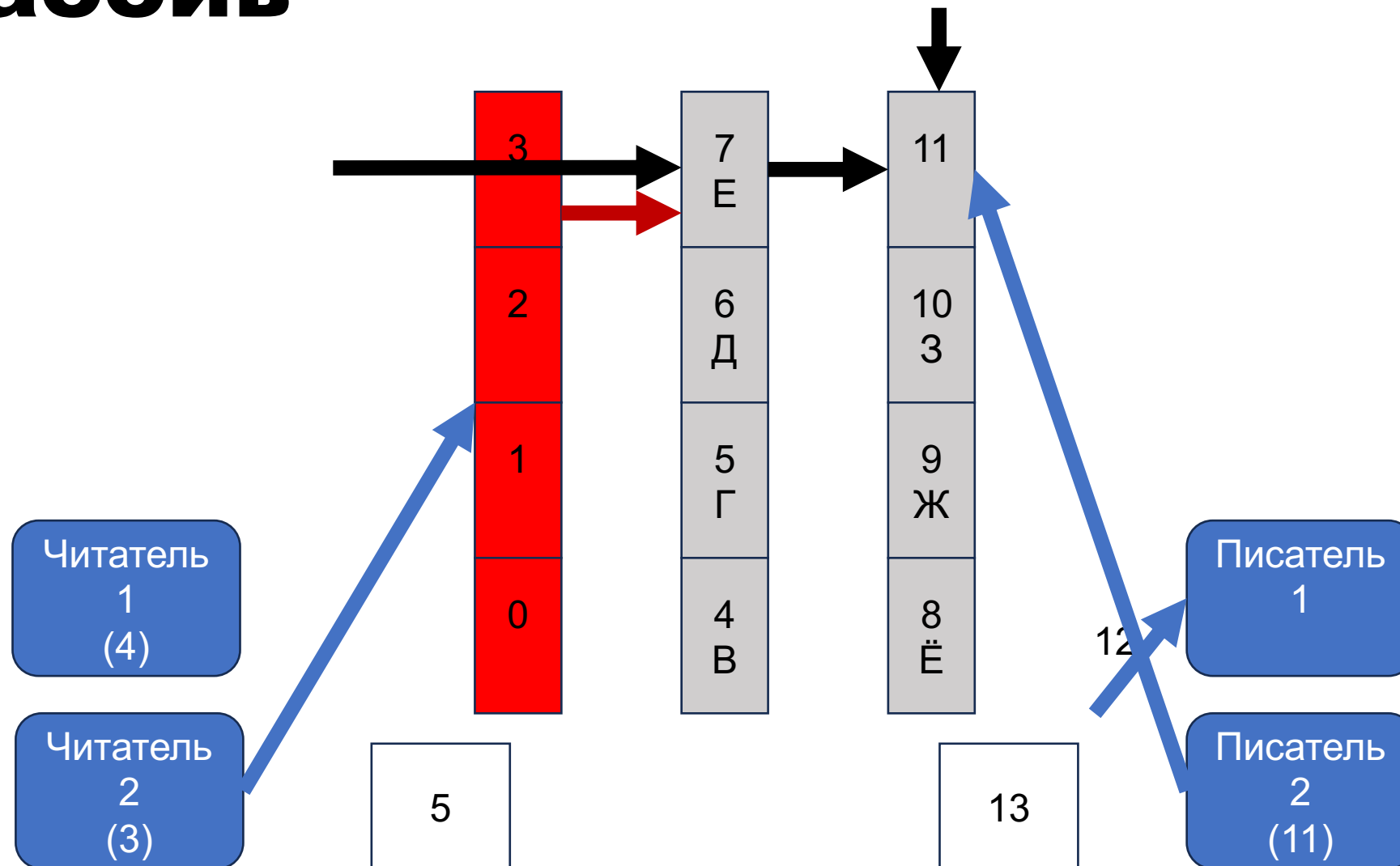
Виртуальный бесконечный массив



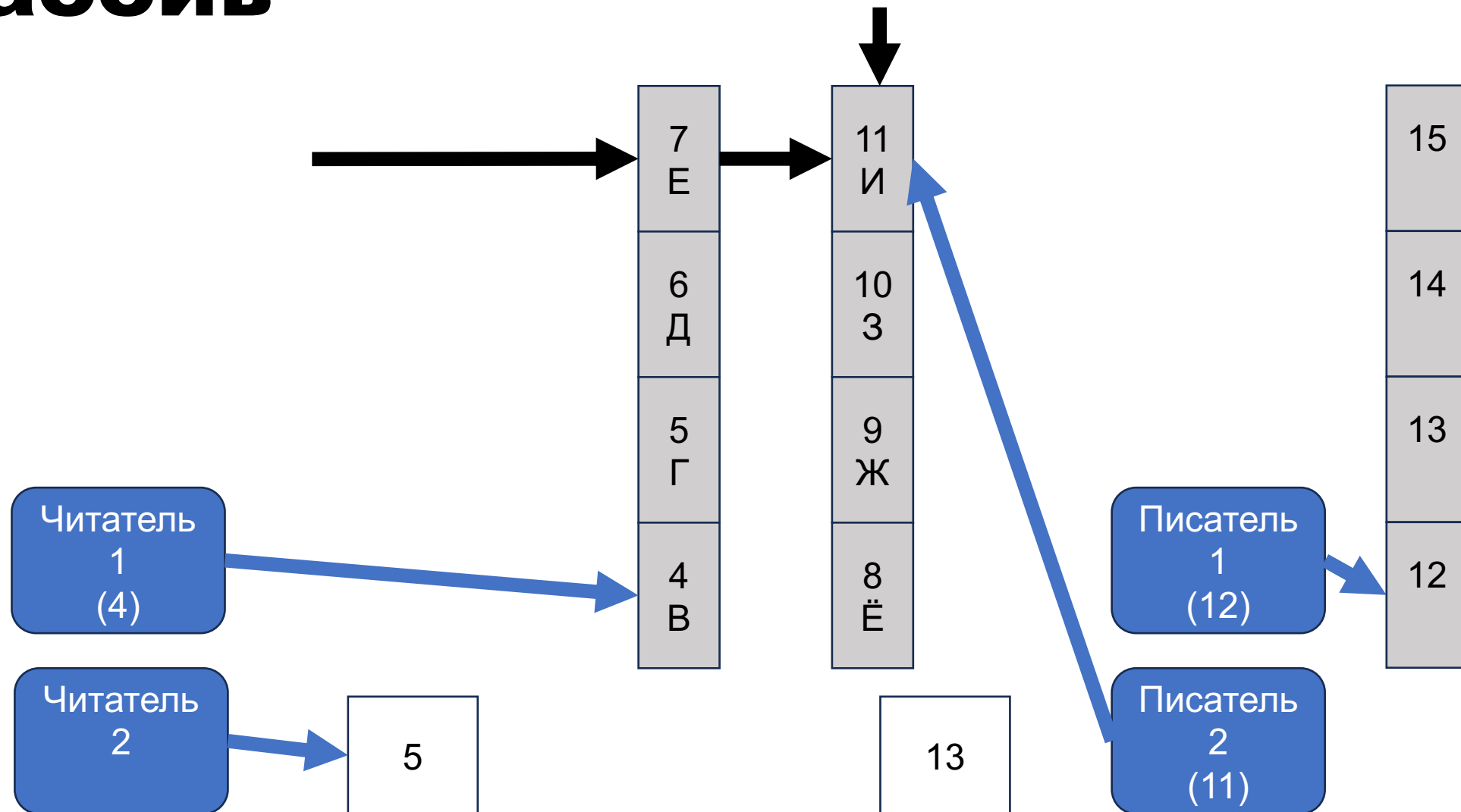
Виртуальный бесконечный массив



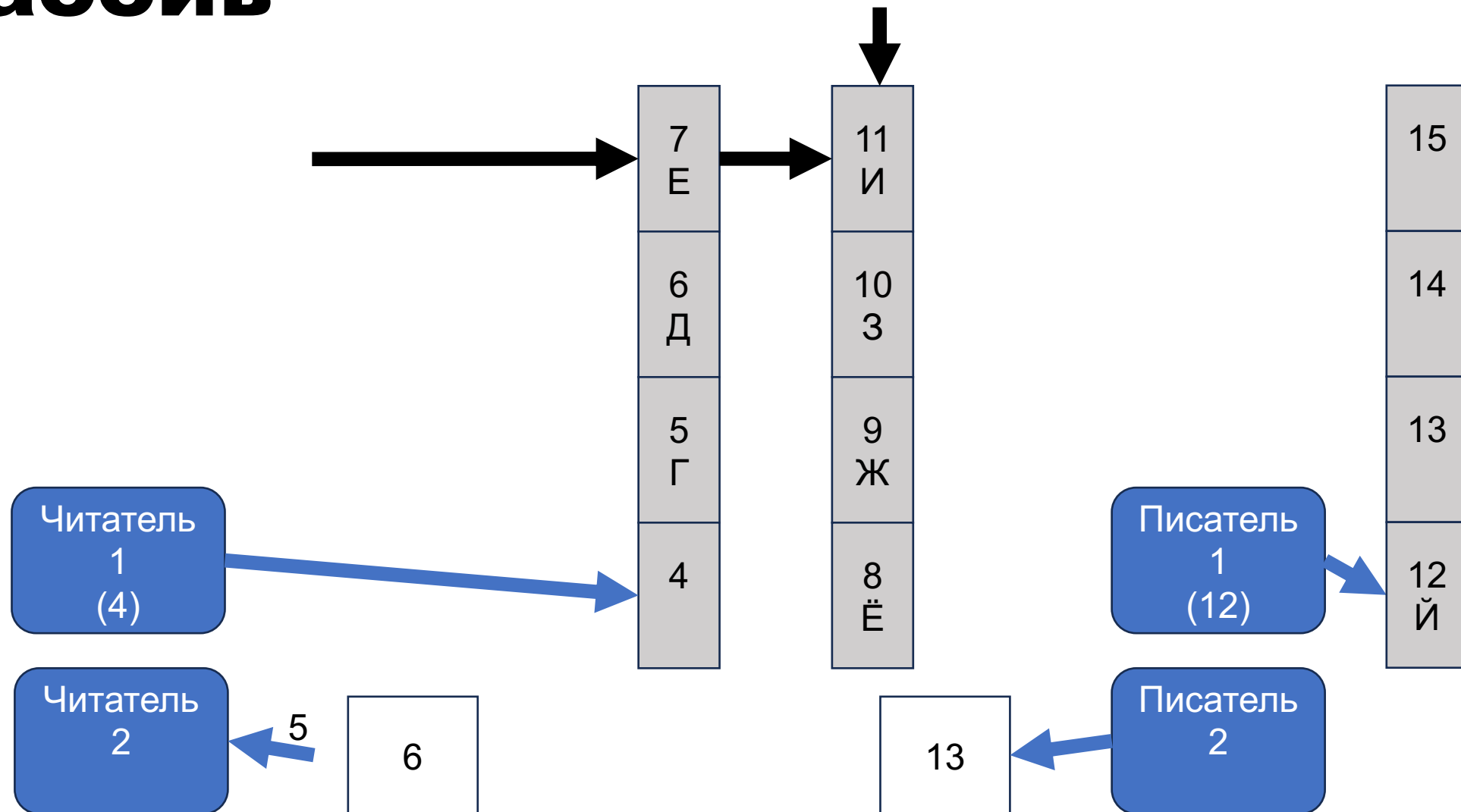
Виртуальный бесконечный массив



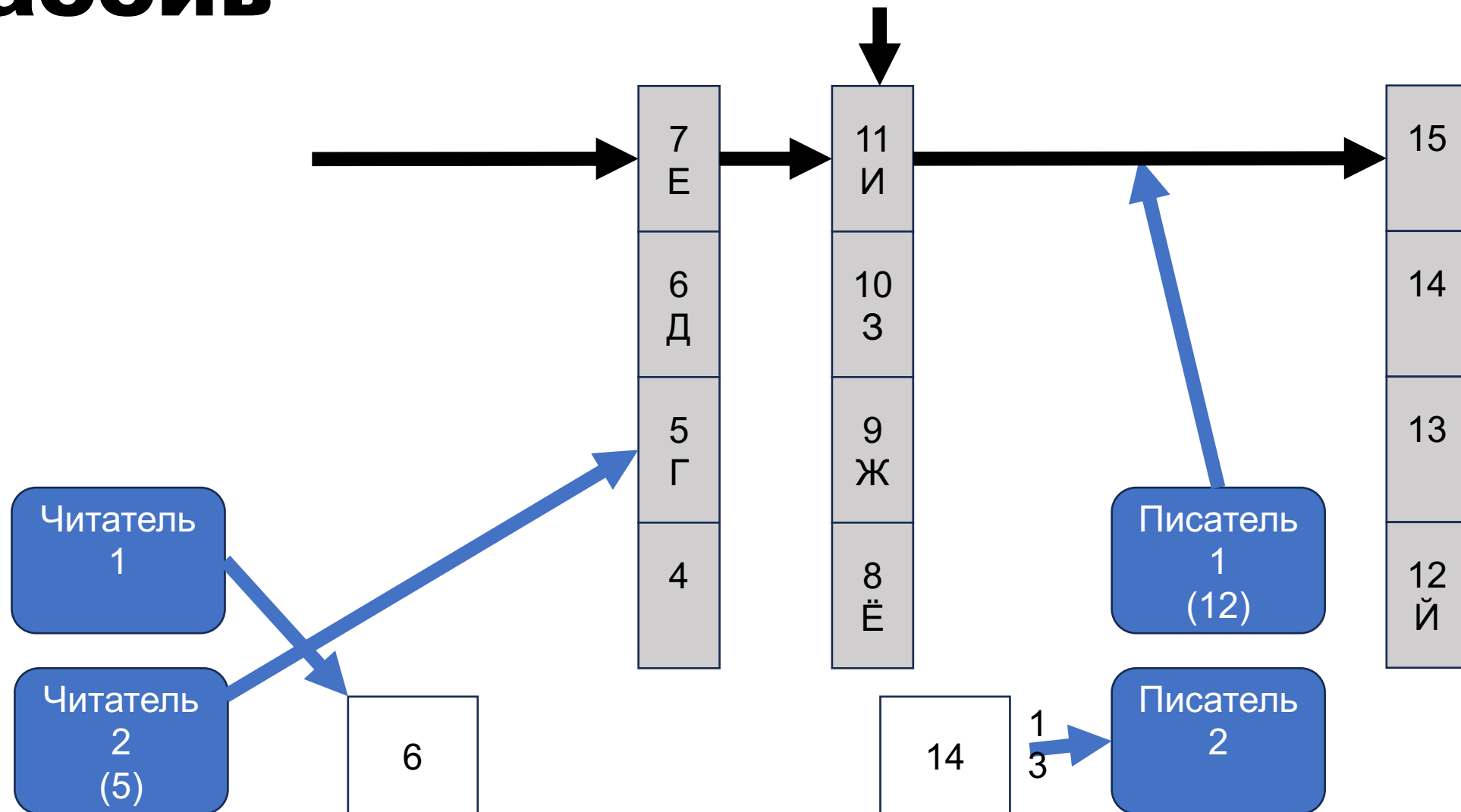
Виртуальный бесконечный массив



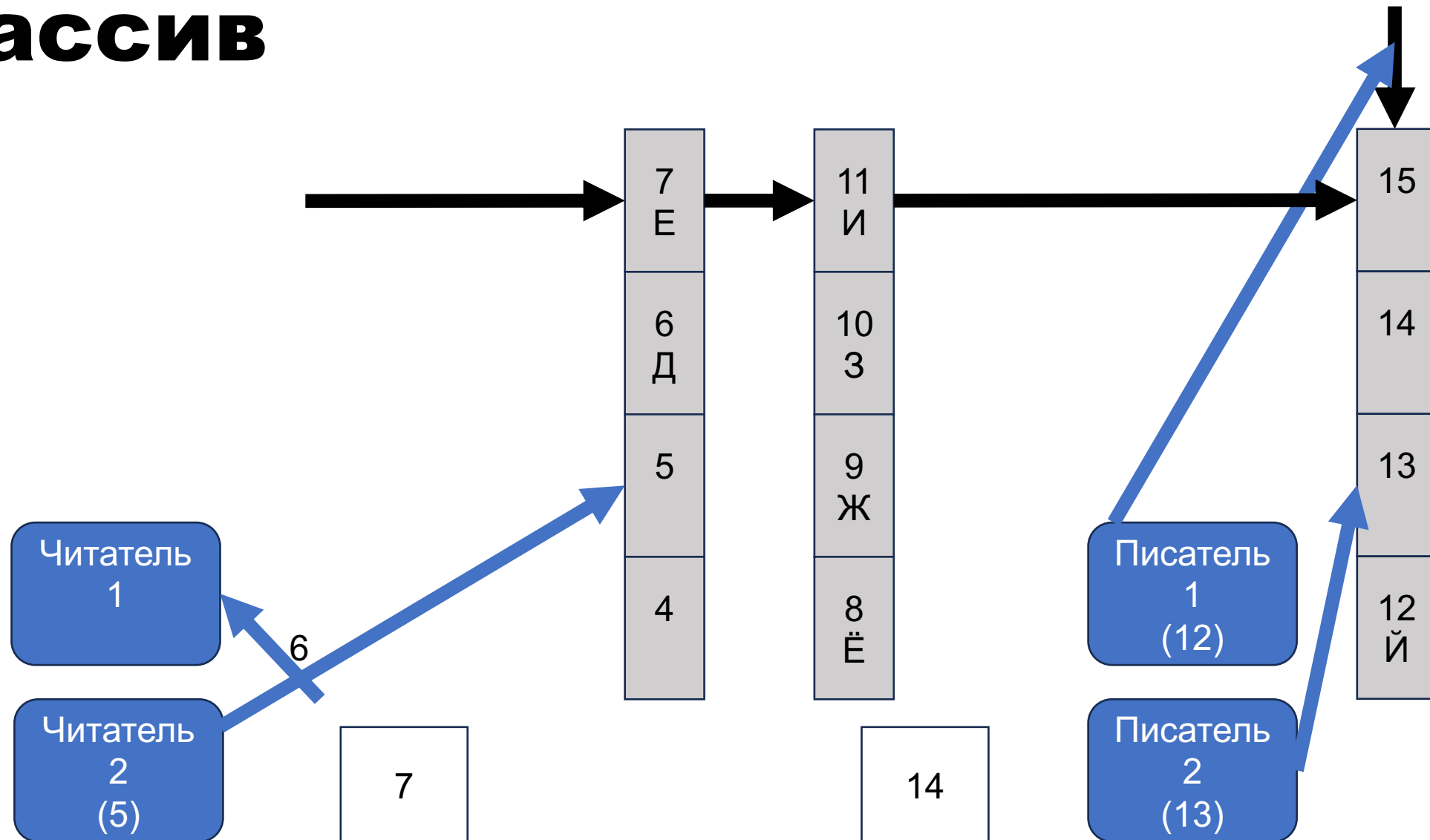
Виртуальный бесконечный массив



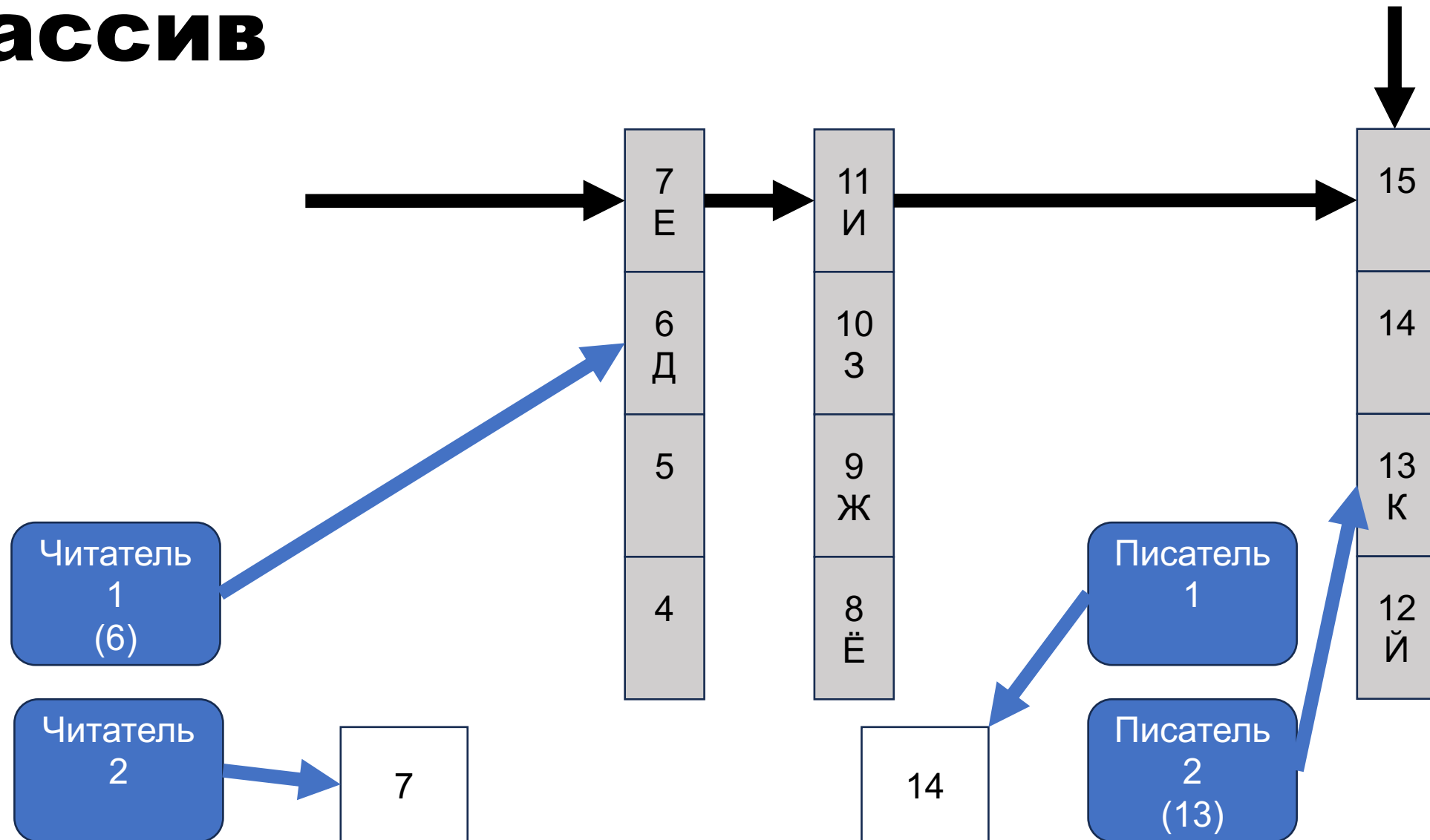
Виртуальный бесконечный массив



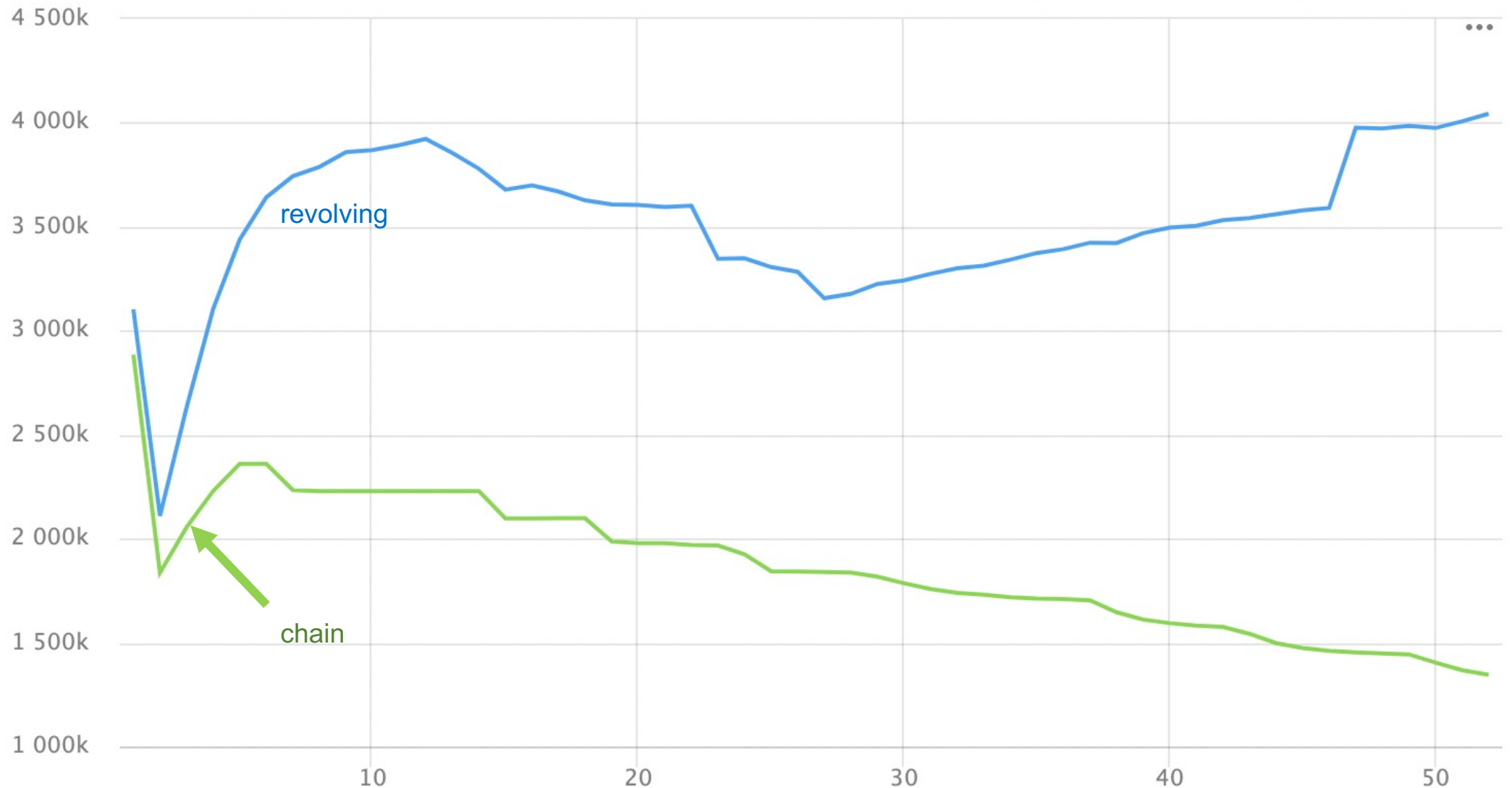
Виртуальный бесконечный массив



Виртуальный бесконечный массив

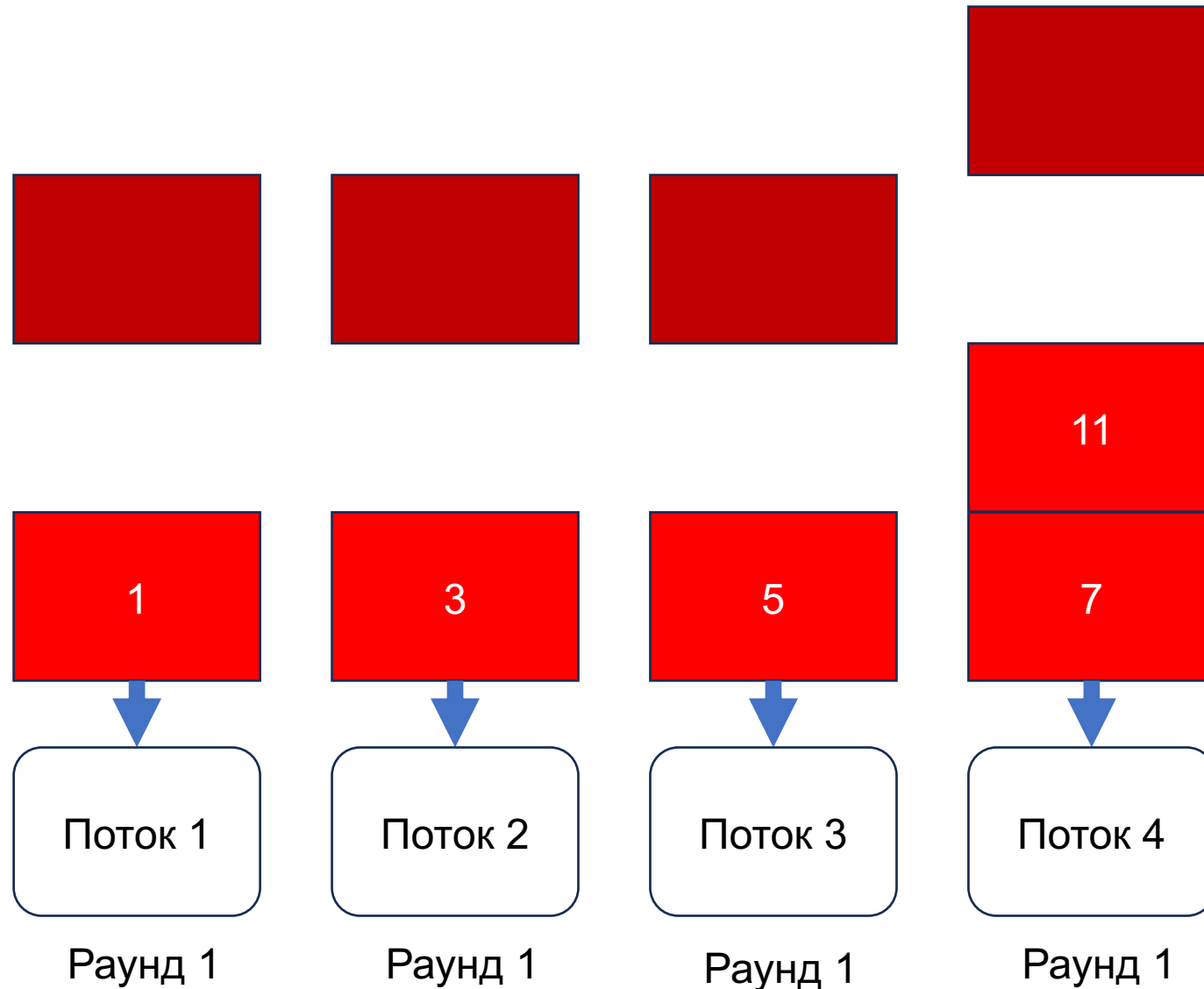


Виртуальный бесконечный массив

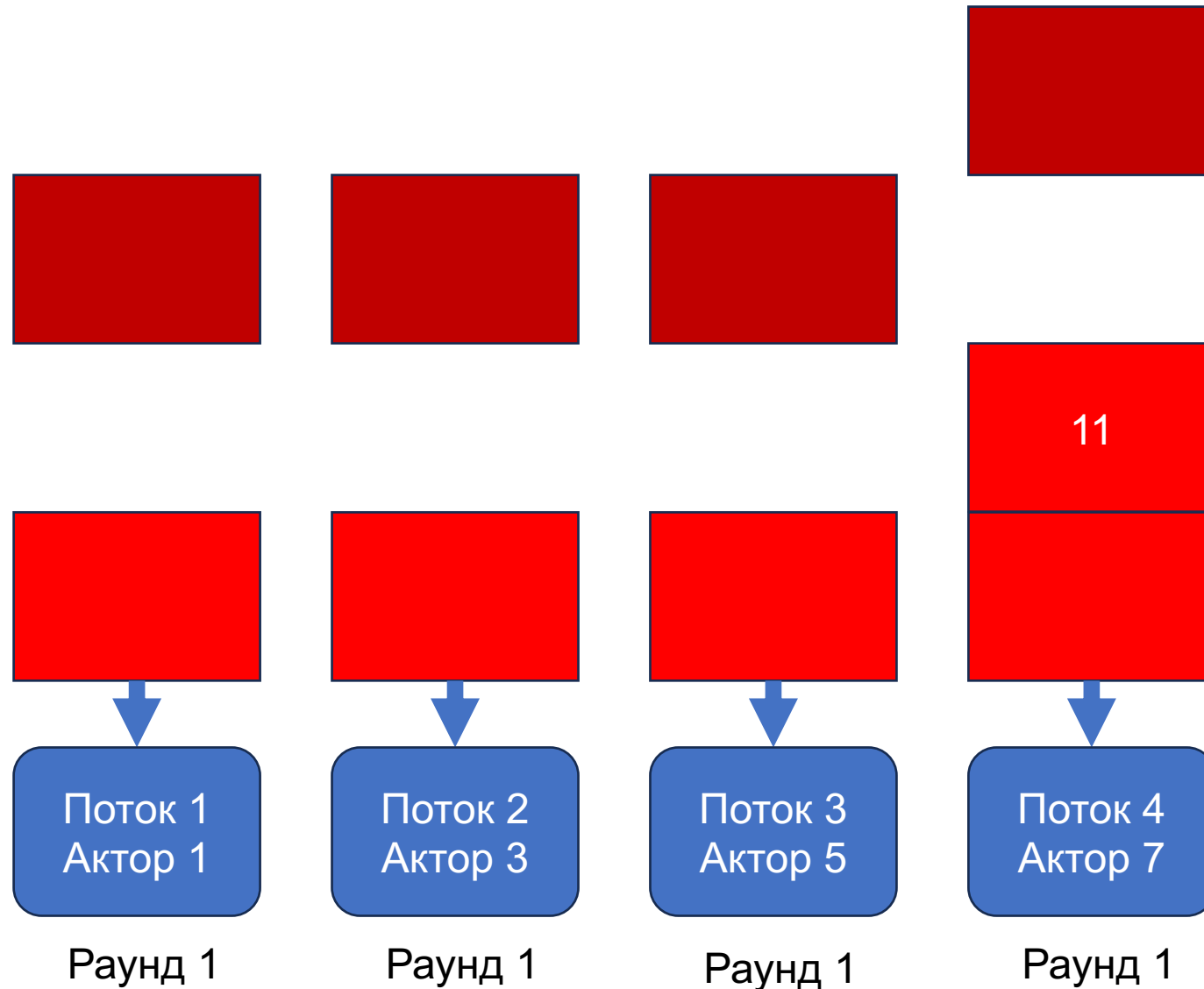


Пробуем применить ТРИЗ

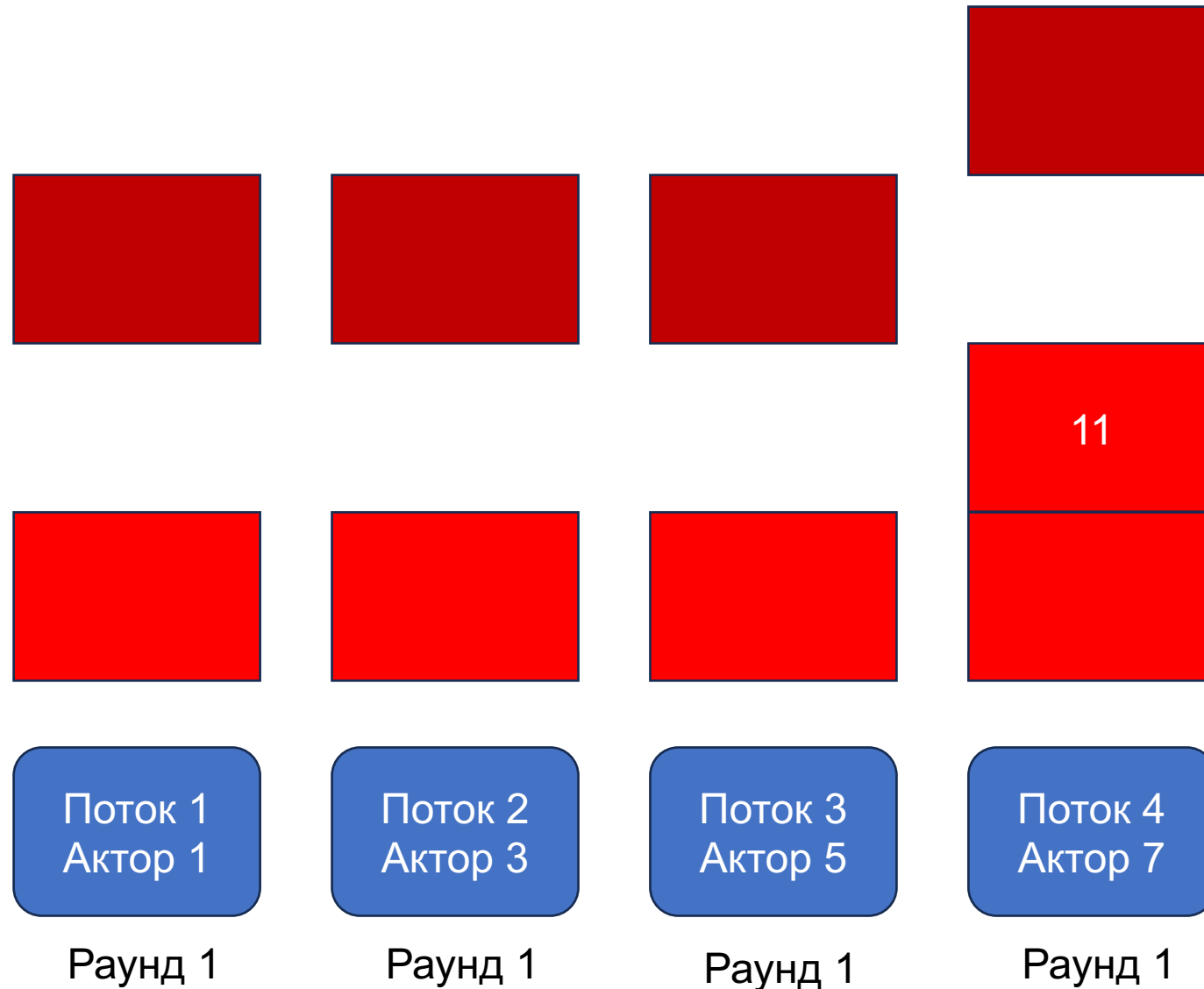
Отказ от единой очереди



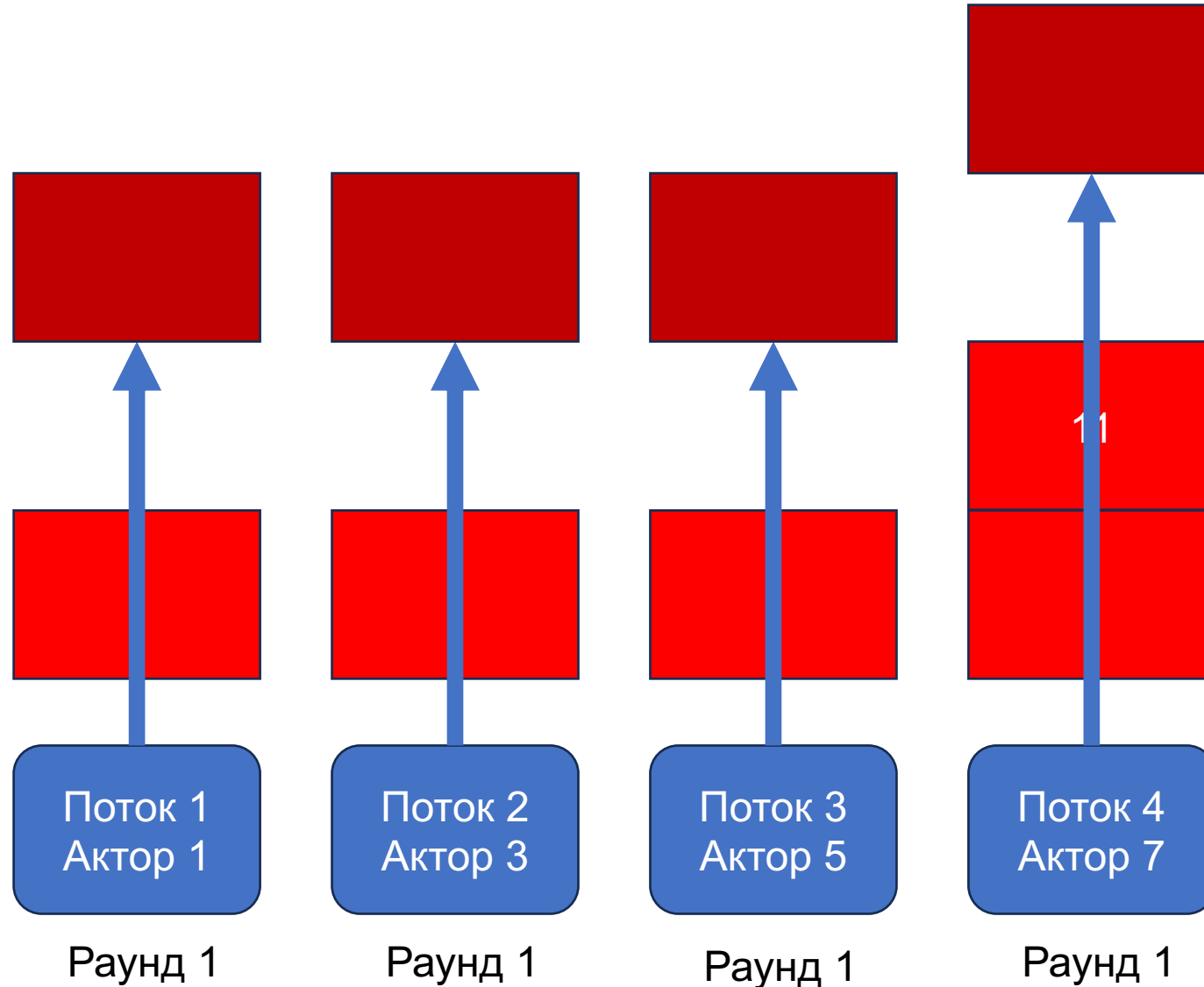
Отказ от единой очереди



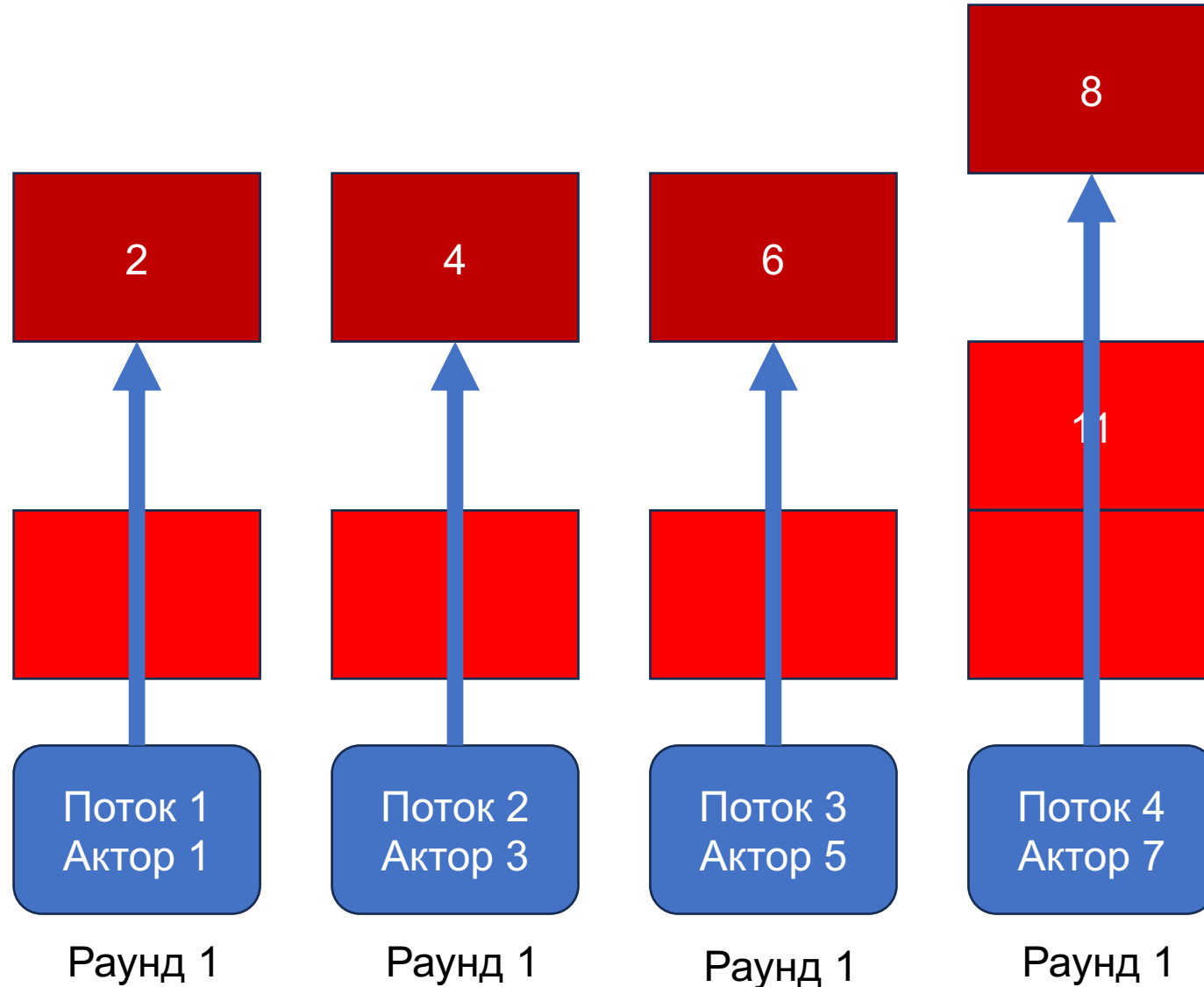
Отказ от единой очереди



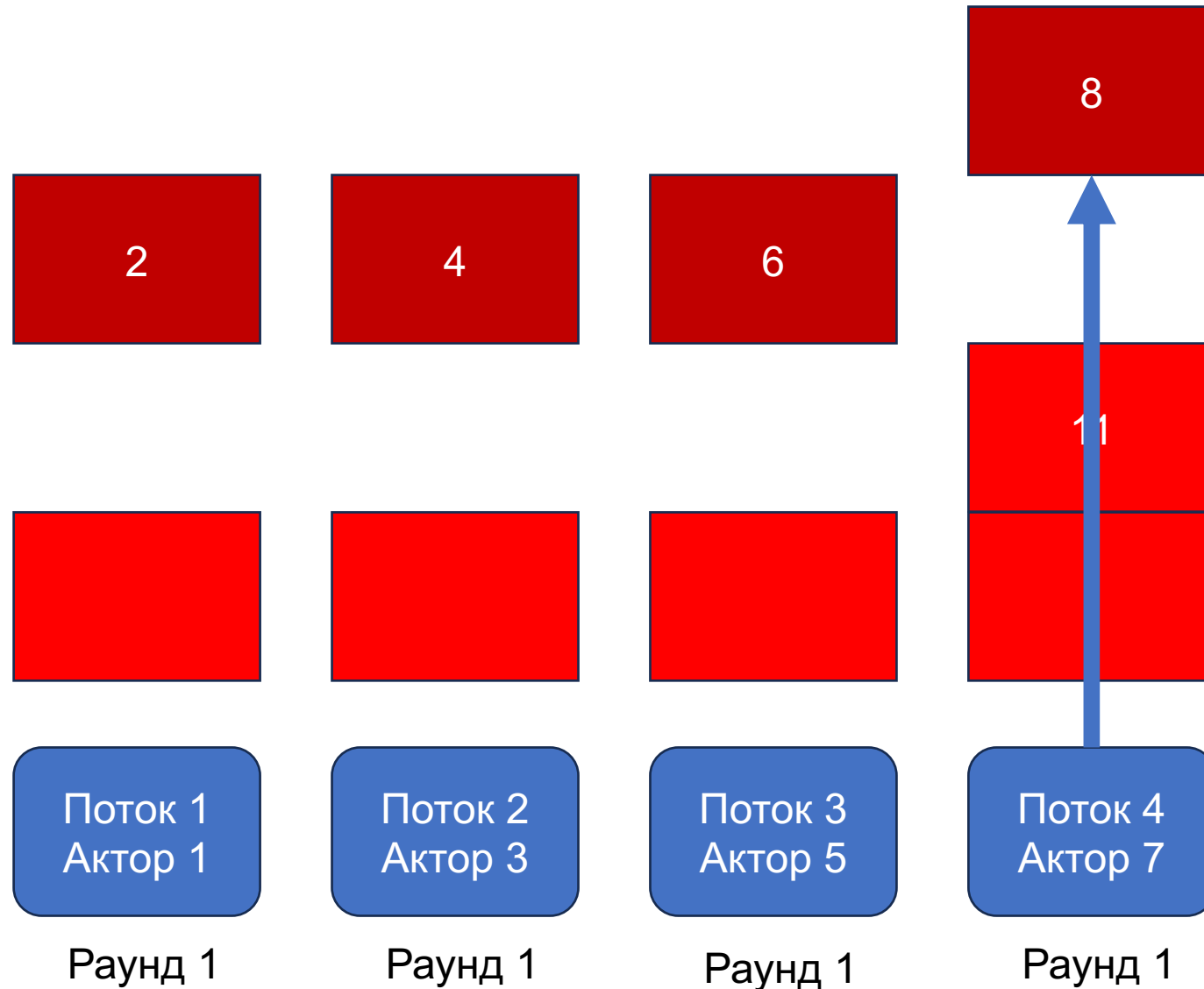
Отказ от единой очереди



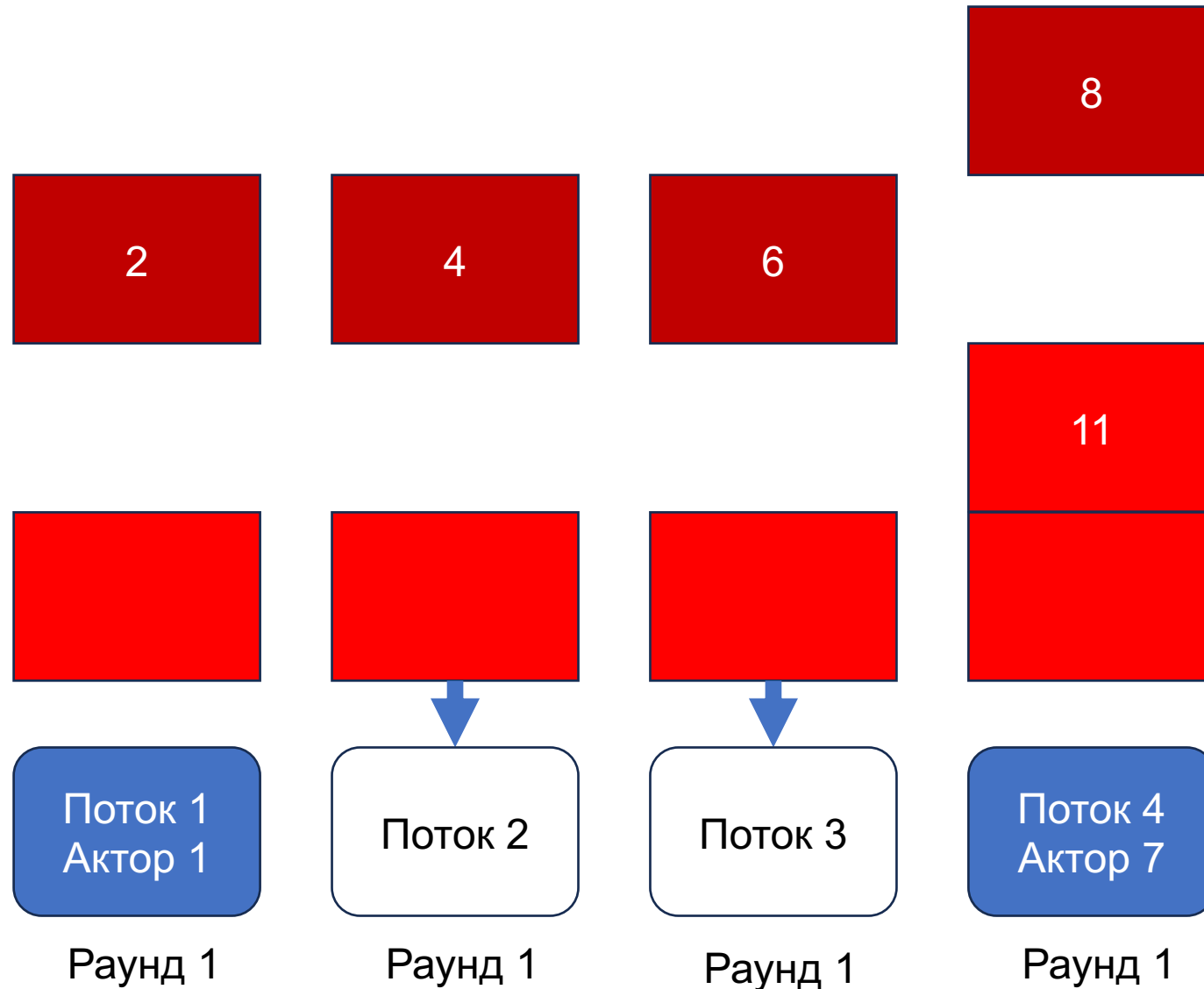
Отказ от единой очереди



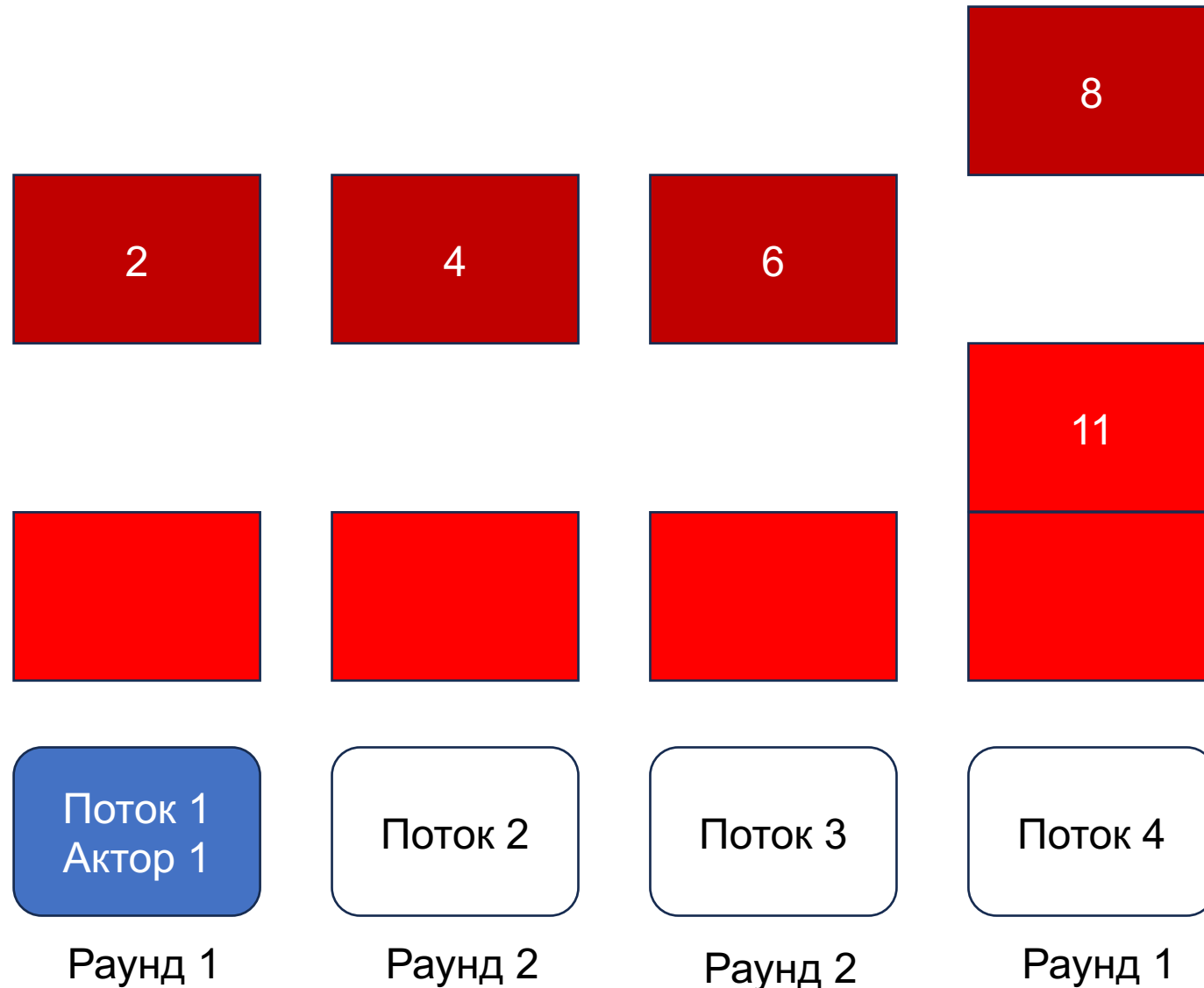
Отказ от единой очереди



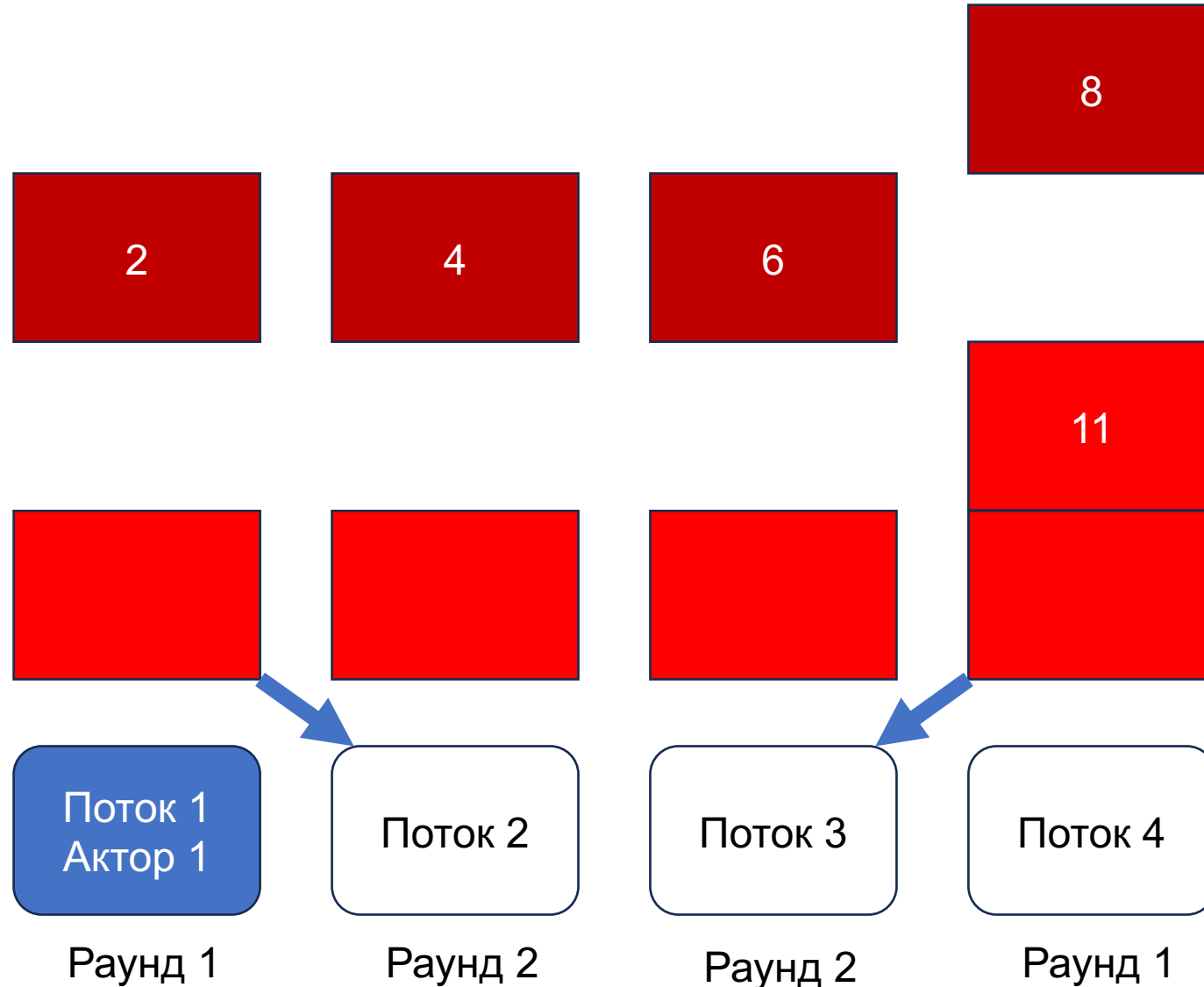
Отказ от единой очереди



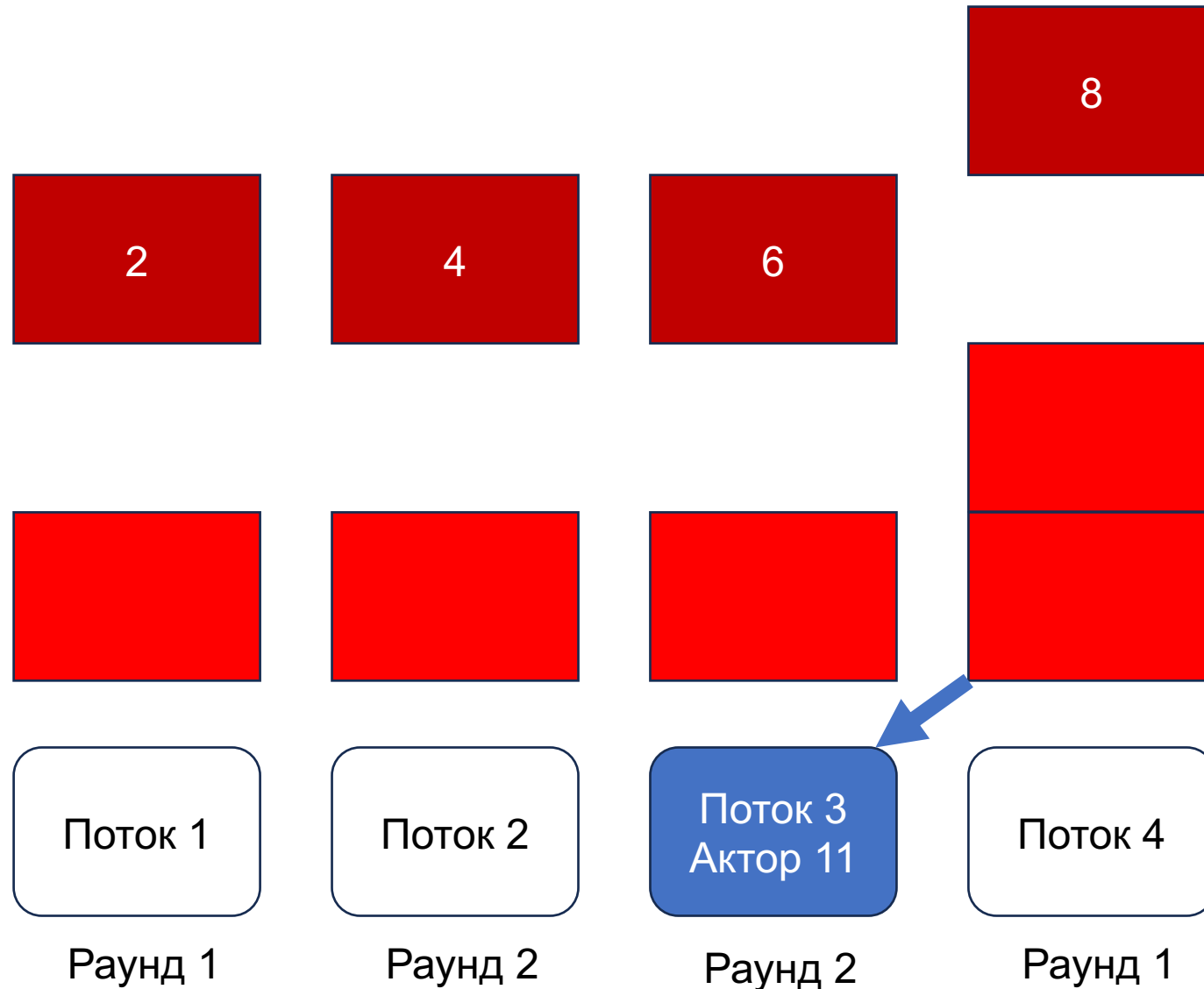
Отказ от единой очереди



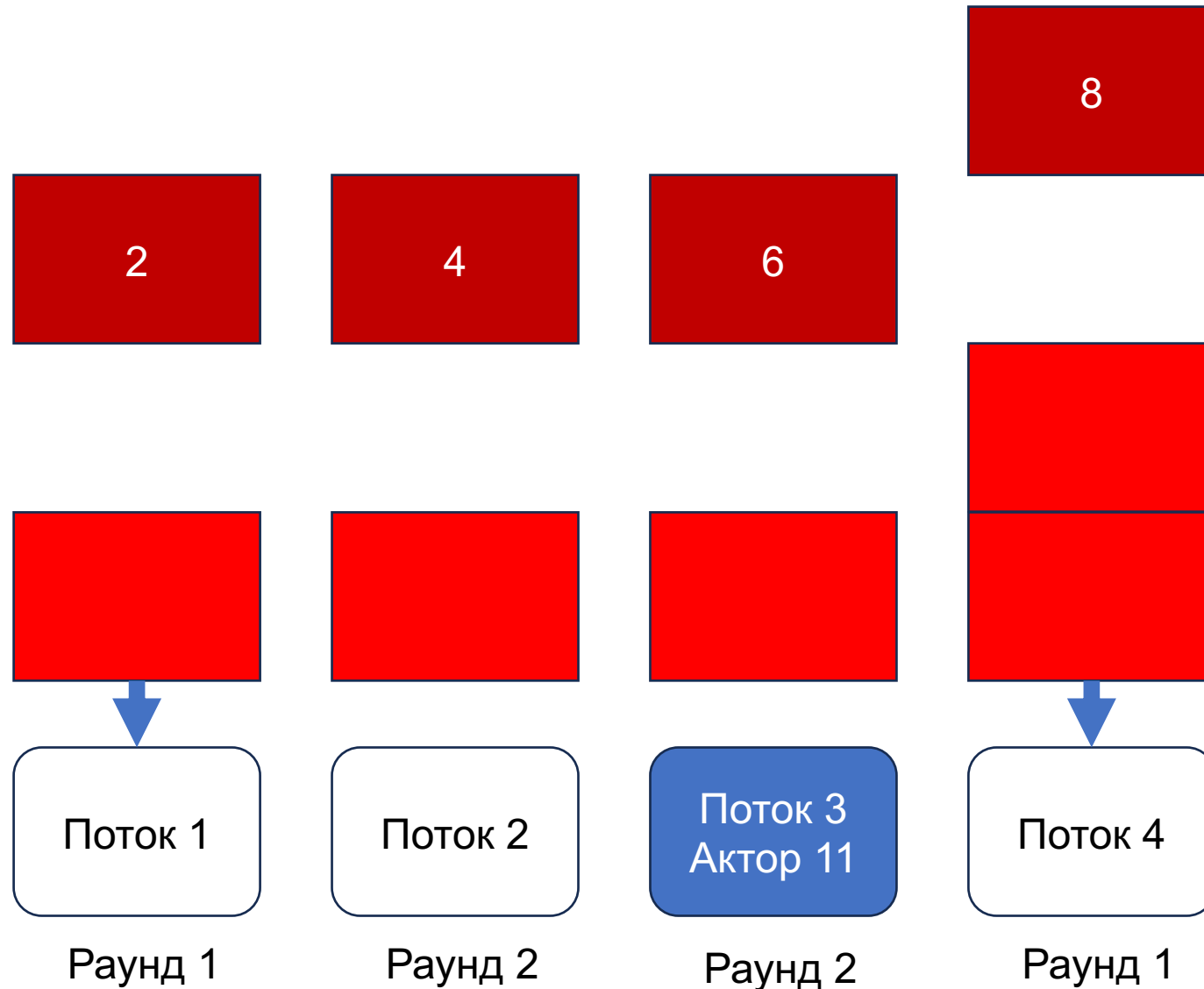
Отказ от единой очереди



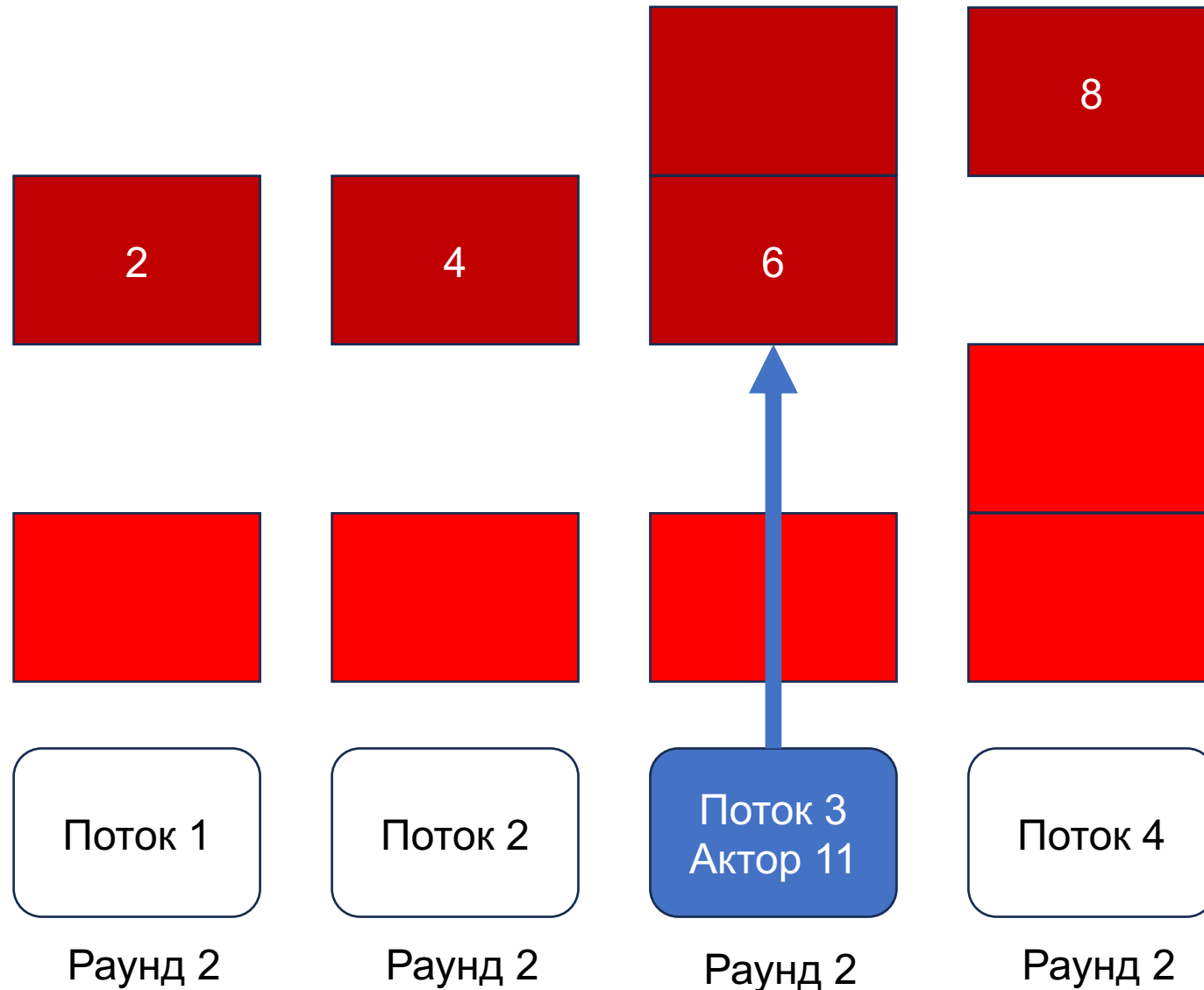
Отказ от единой очереди



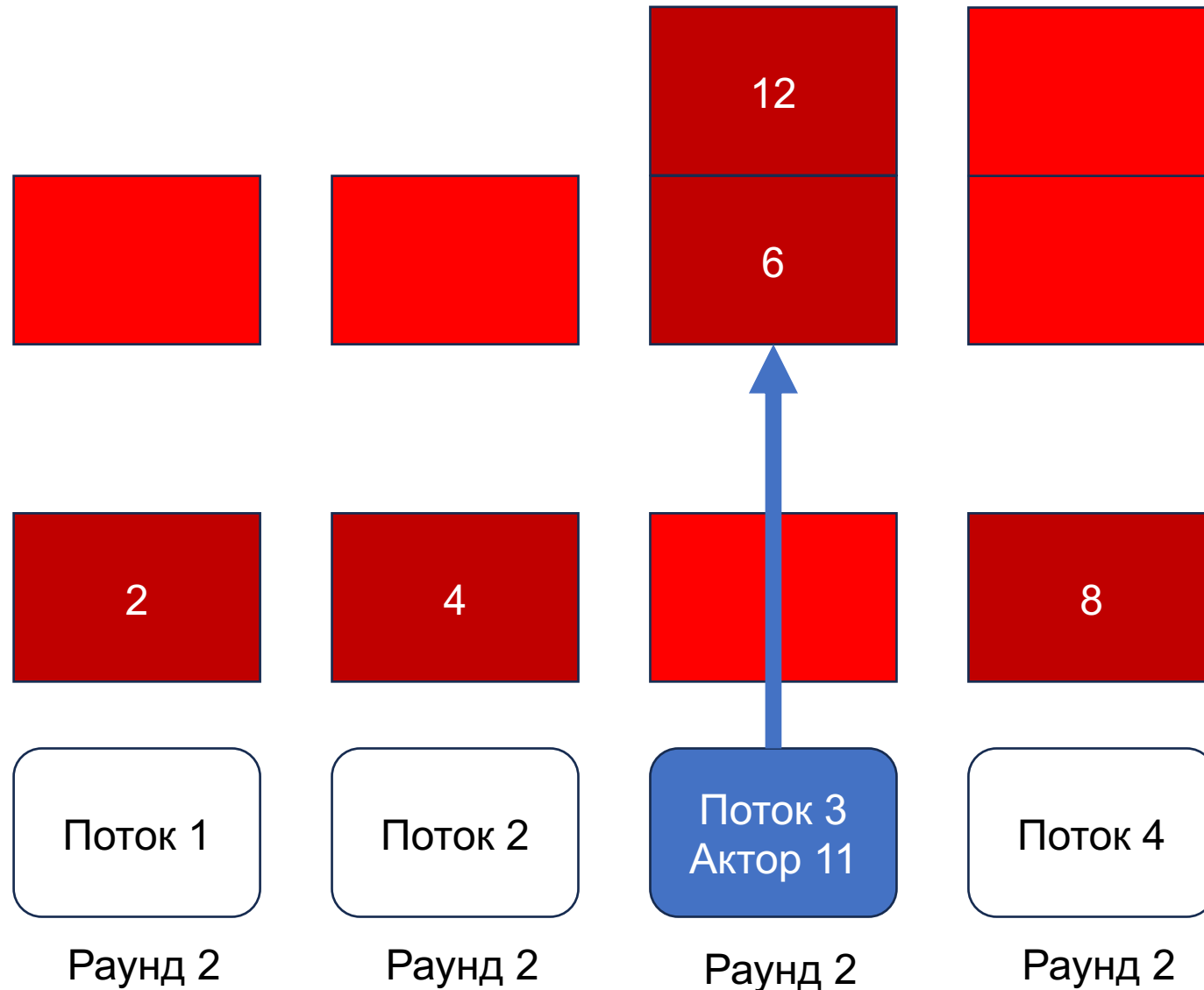
Отказ от единой очереди



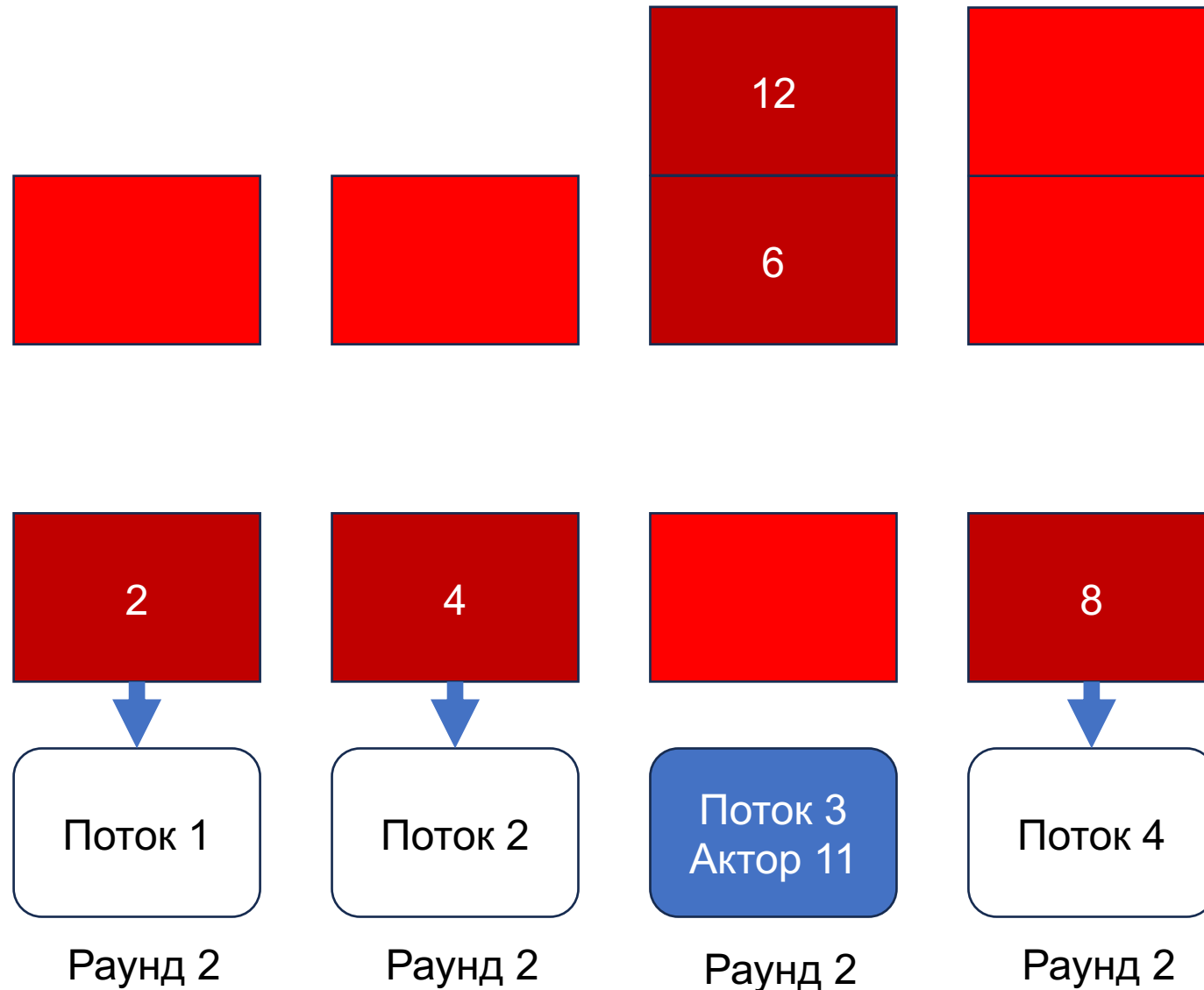
Отказ от единой очереди



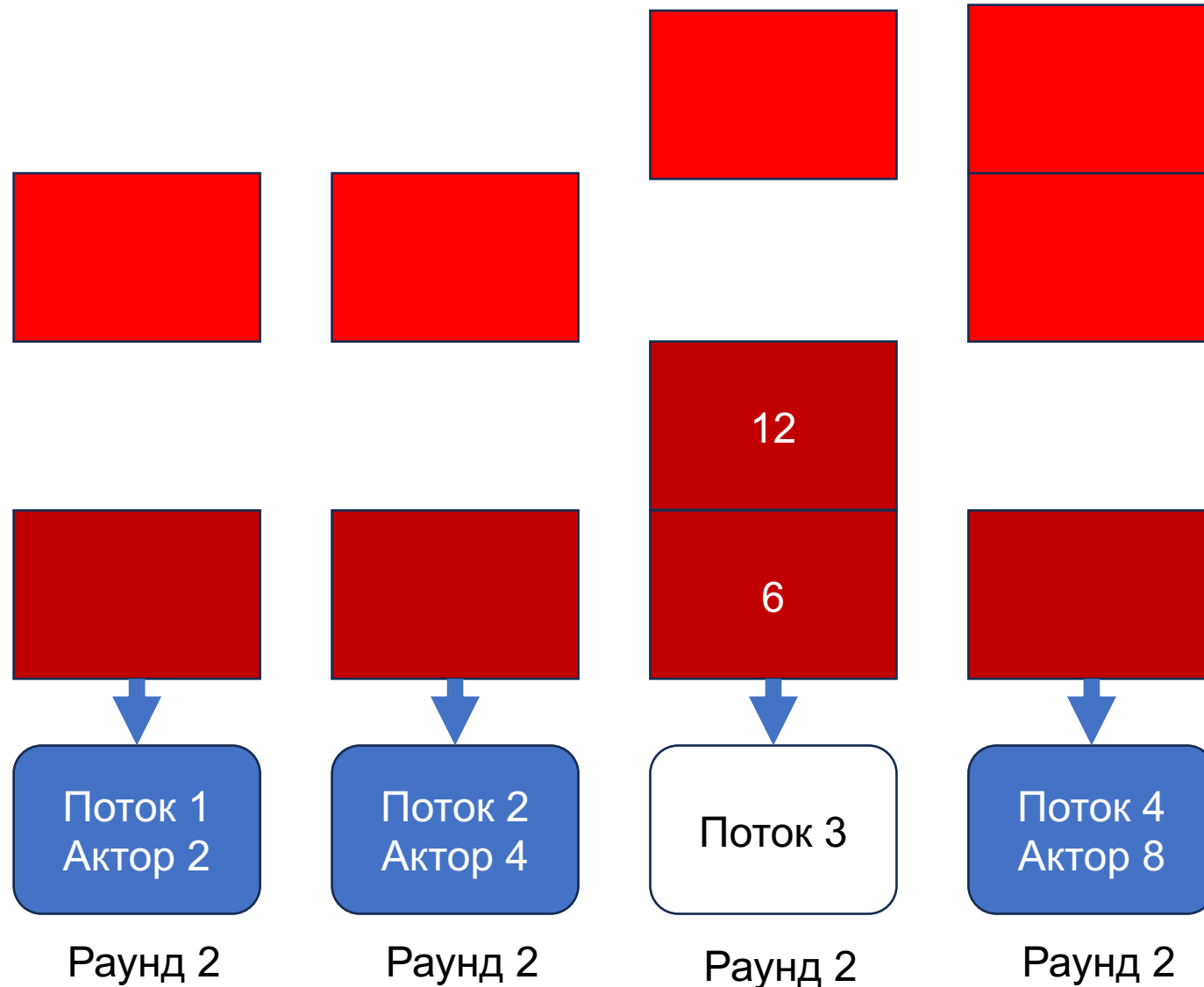
Отказ от единой очереди



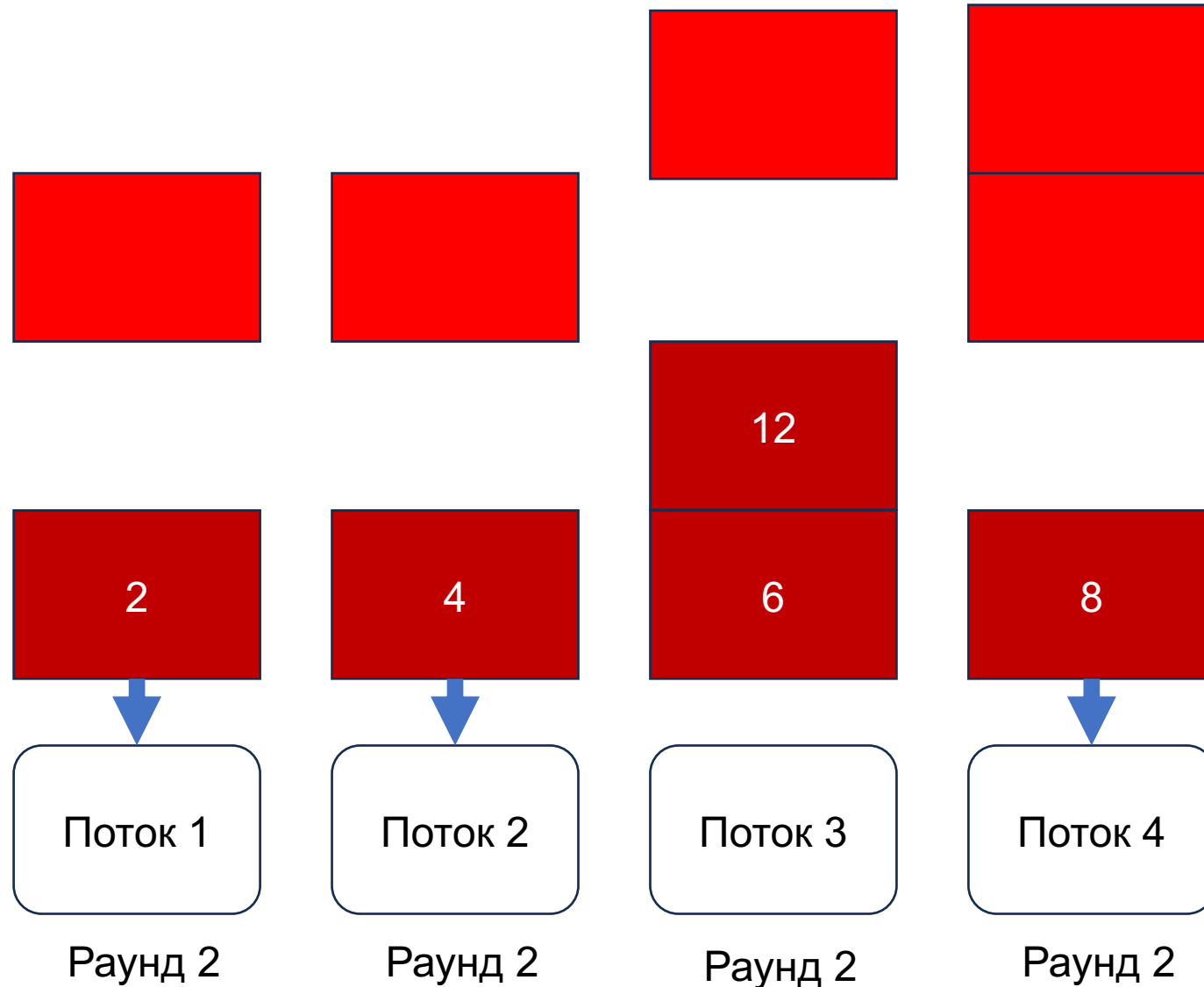
Отказ от единой очереди



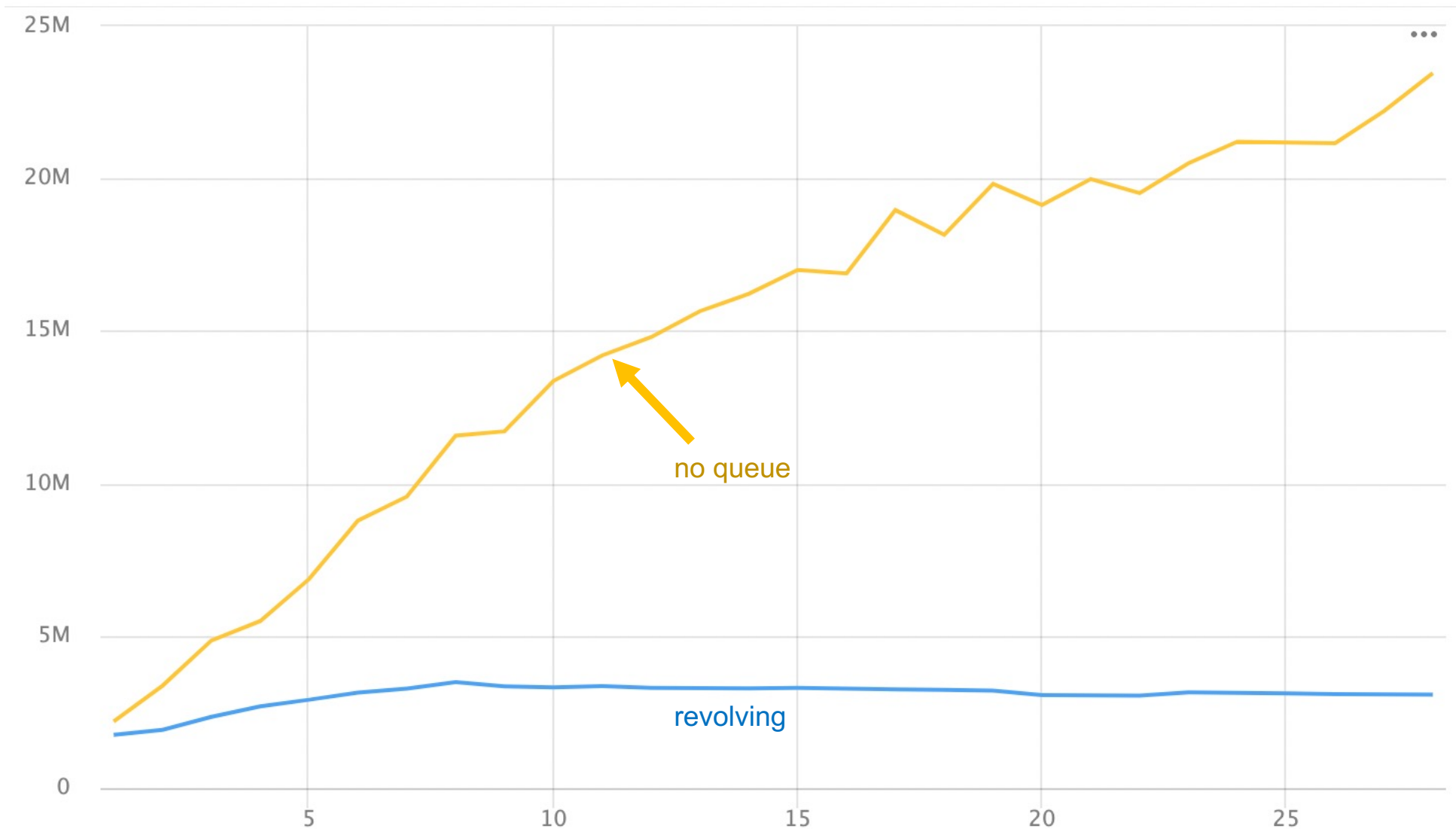
Отказ от единой очереди



Отказ от единой очереди



Отказ от единой очереди

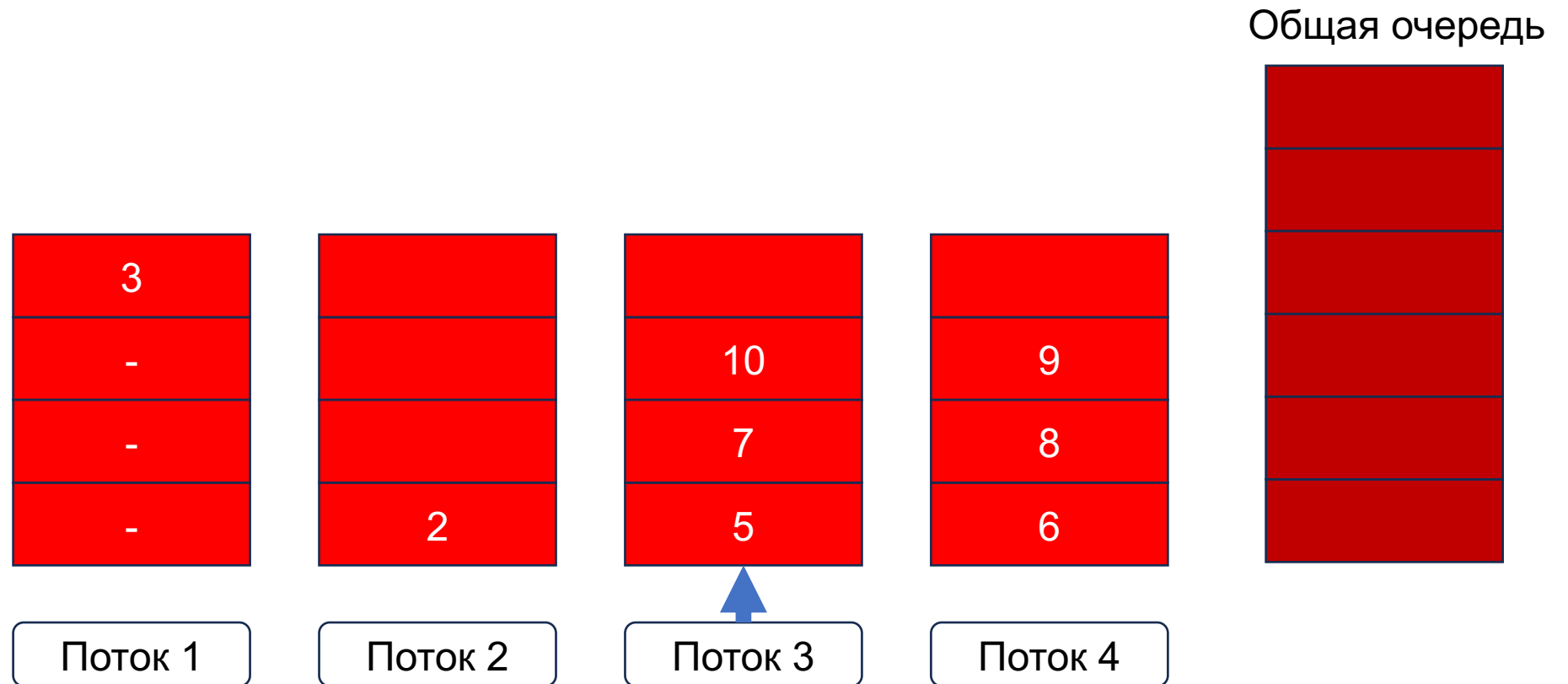


Чуть менее радикальный подход

Короткие per-thread очереди



Короткие per-thread очереди



Короткие per-thread очереди



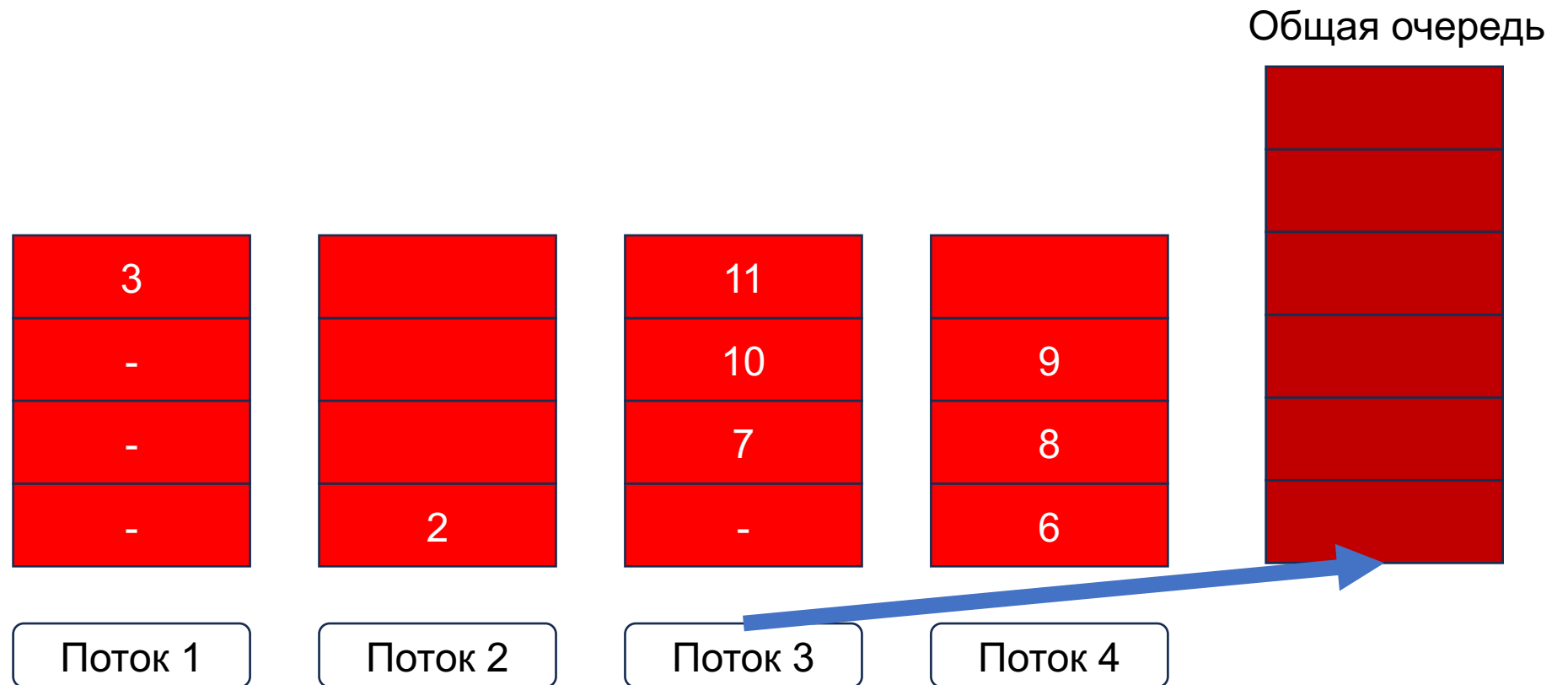
Короткие per-thread очереди



Короткие per-thread очереди



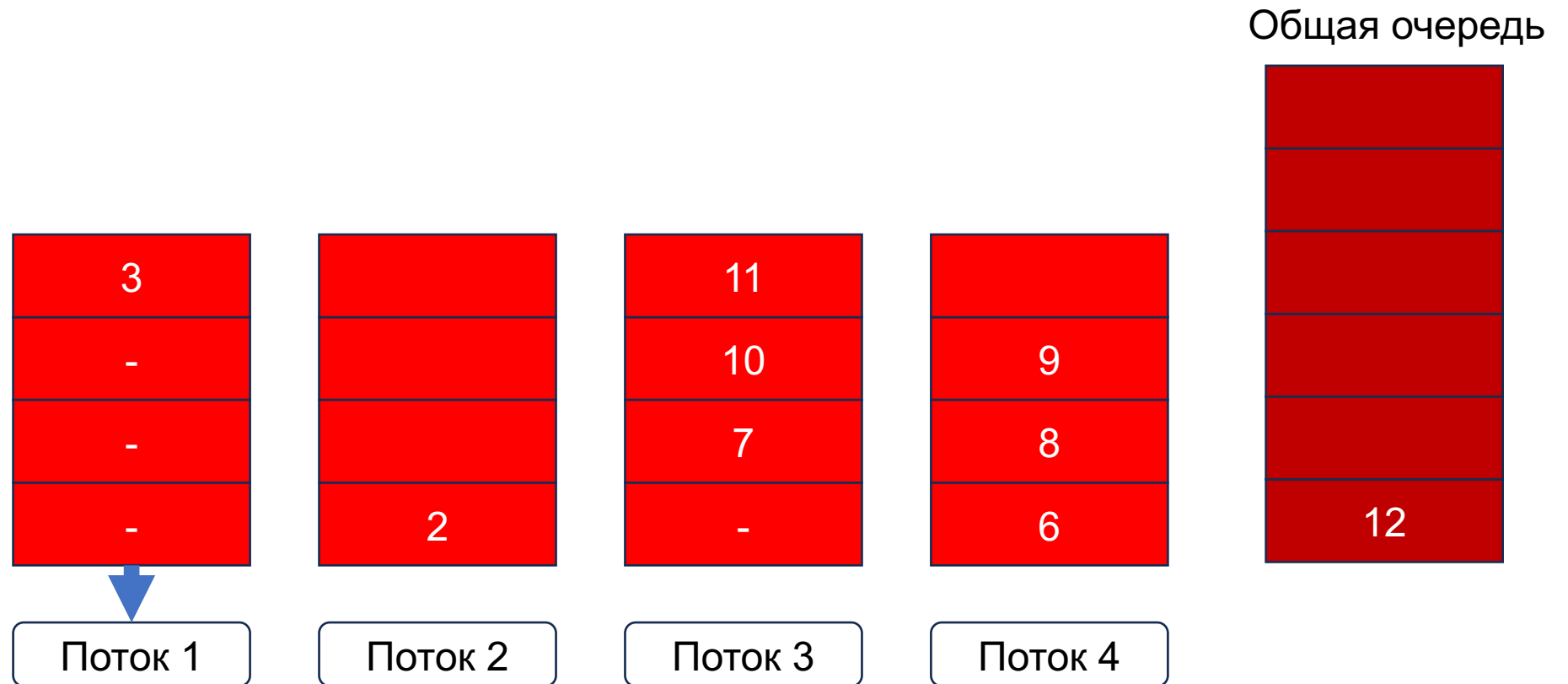
Короткие per-thread очереди



Короткие per-thread очереди



Короткие per-thread очереди



Короткие per-thread очереди



Короткие per-thread очереди



Короткие per-thread очереди



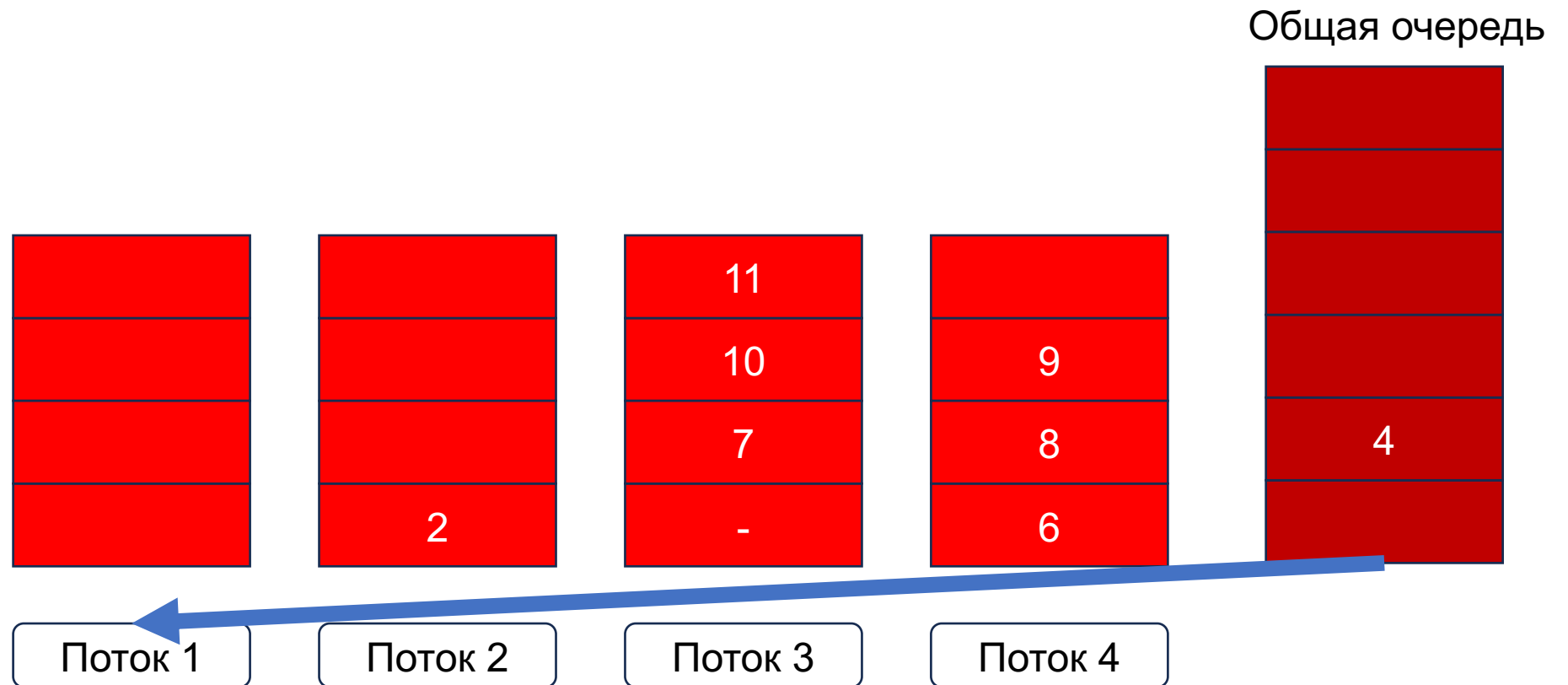
Короткие per-thread очереди



Короткие per-thread очереди



Короткие per-thread очереди



Короткие per-thread очереди



Короткие per-thread очереди



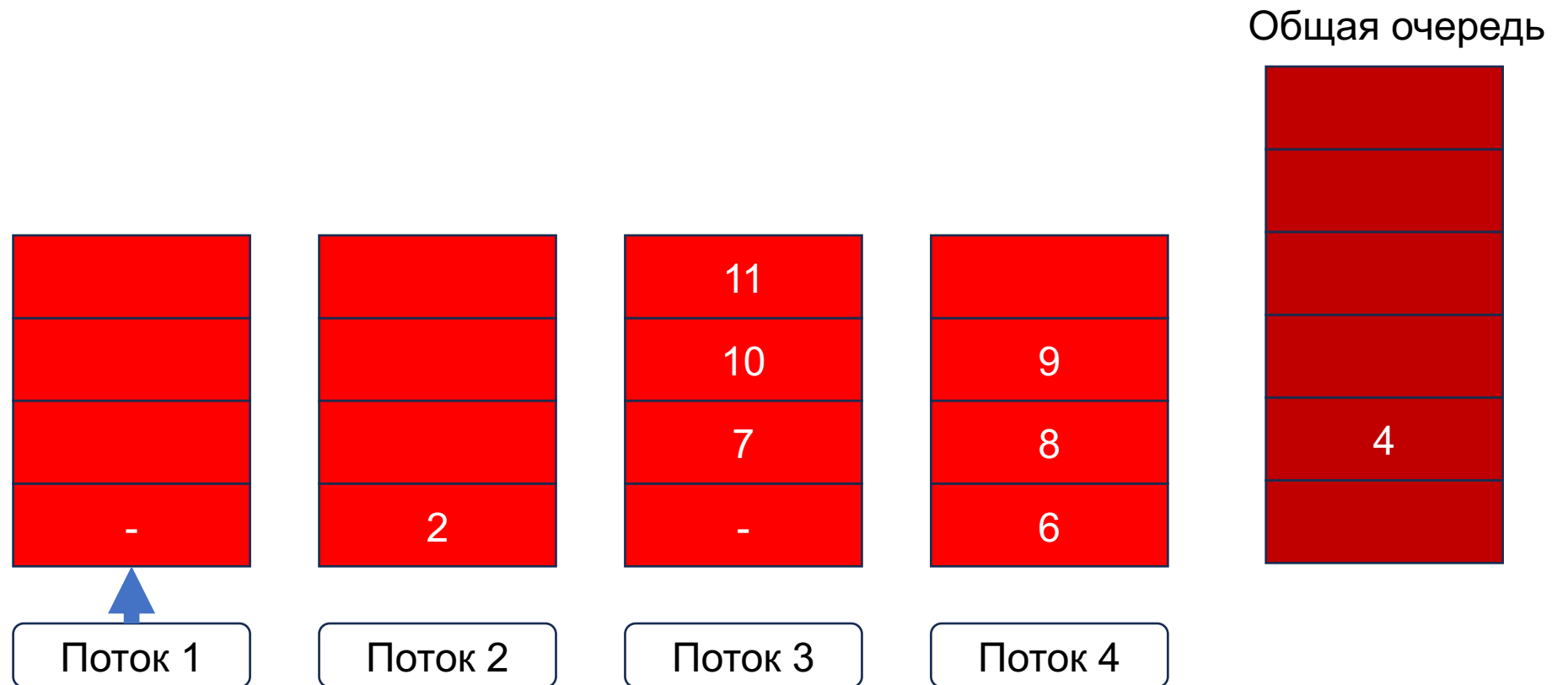
Короткие per-thread очереди



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



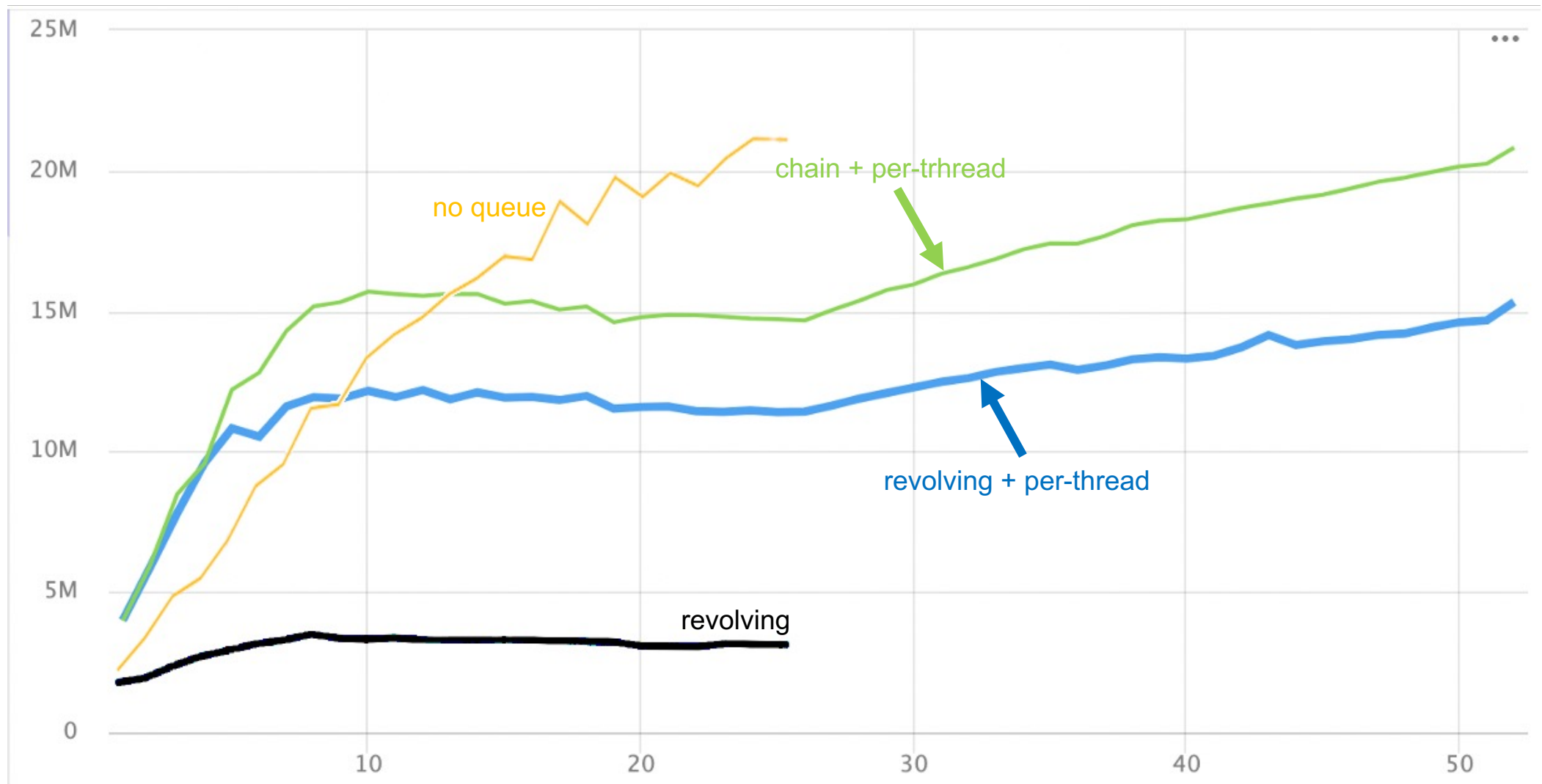
Короткие per-thread очереди



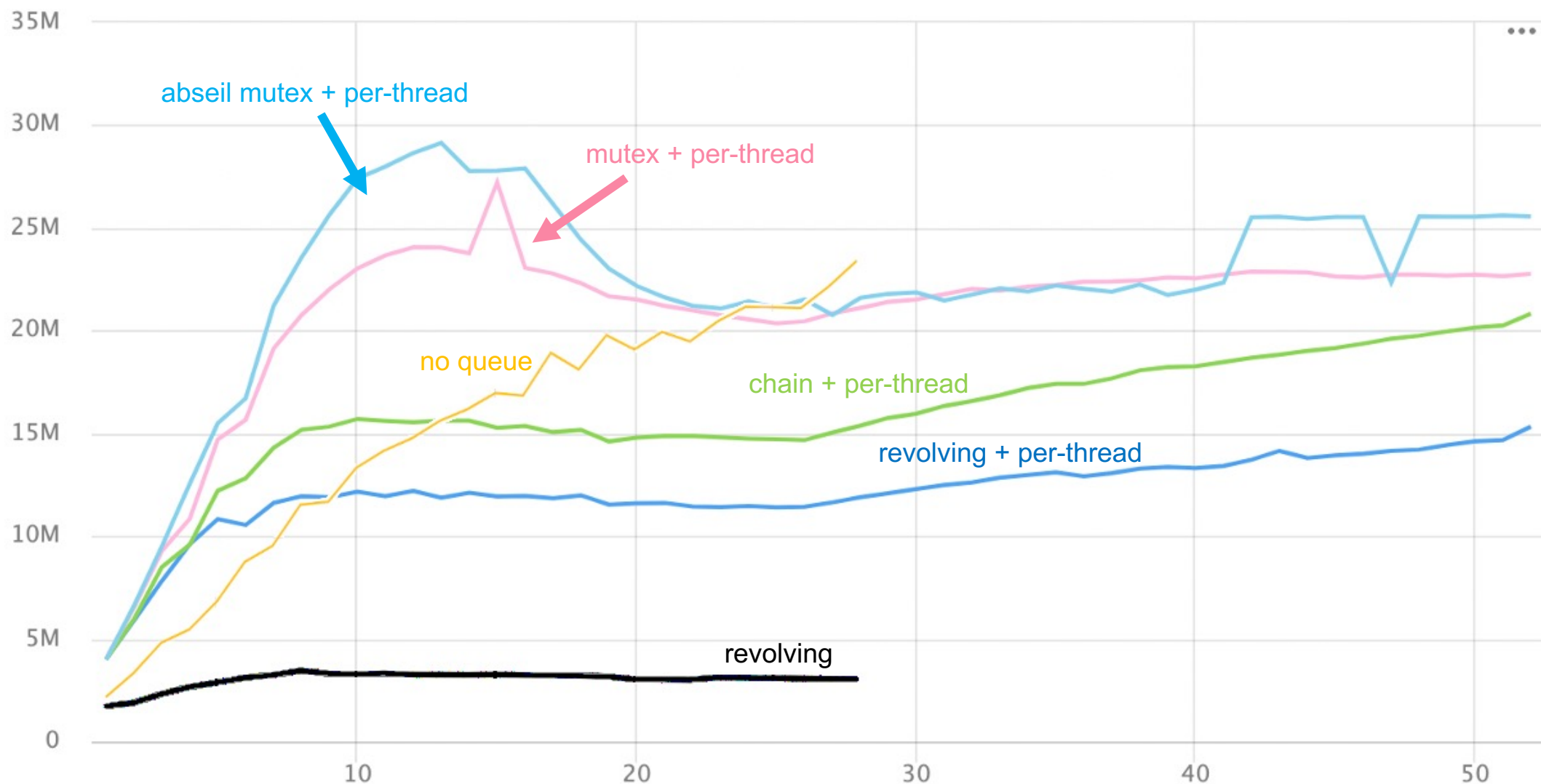
Короткие per-thread очереди



Короткие per-thread очереди

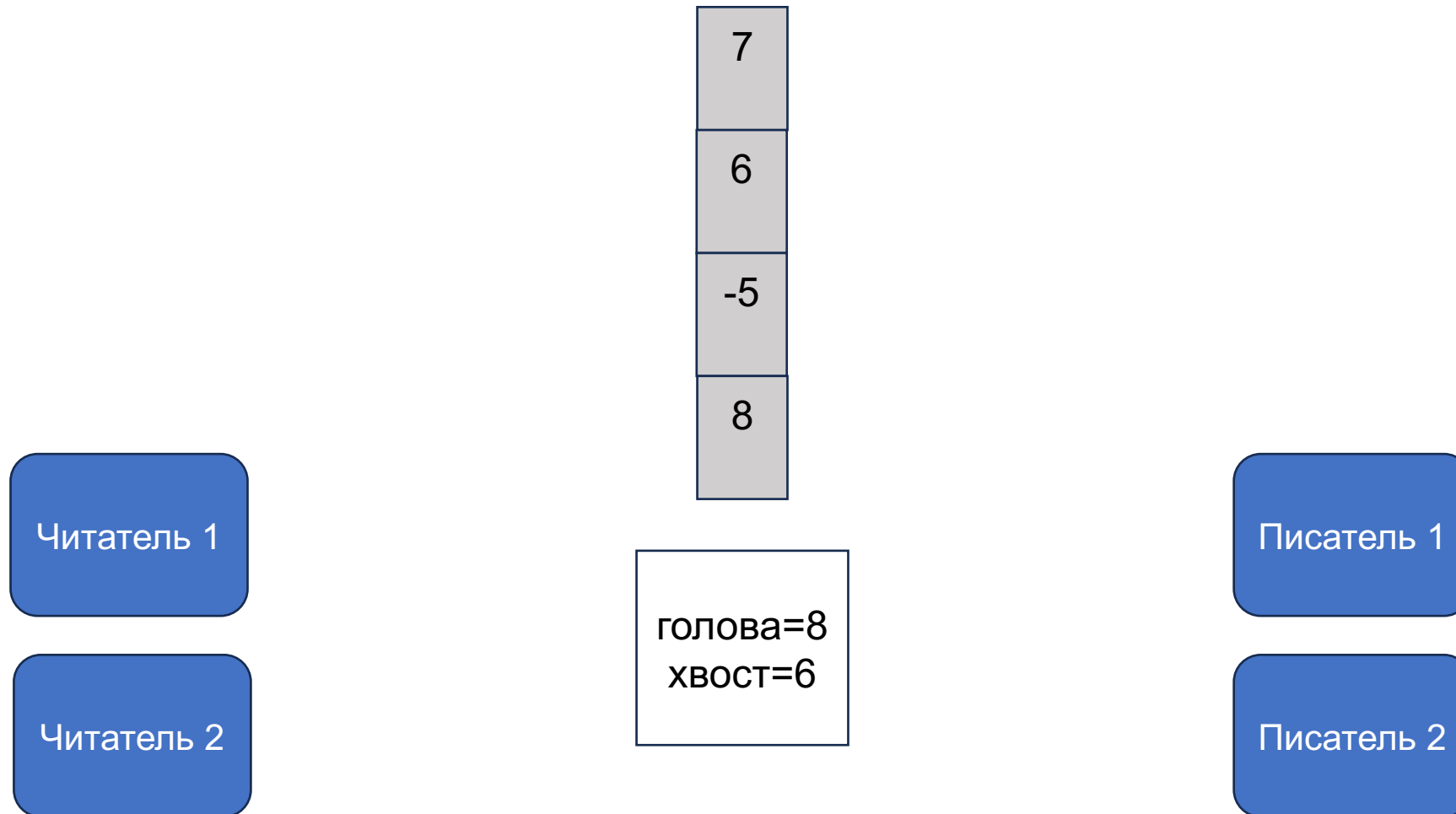


Короткие per-thread очереди

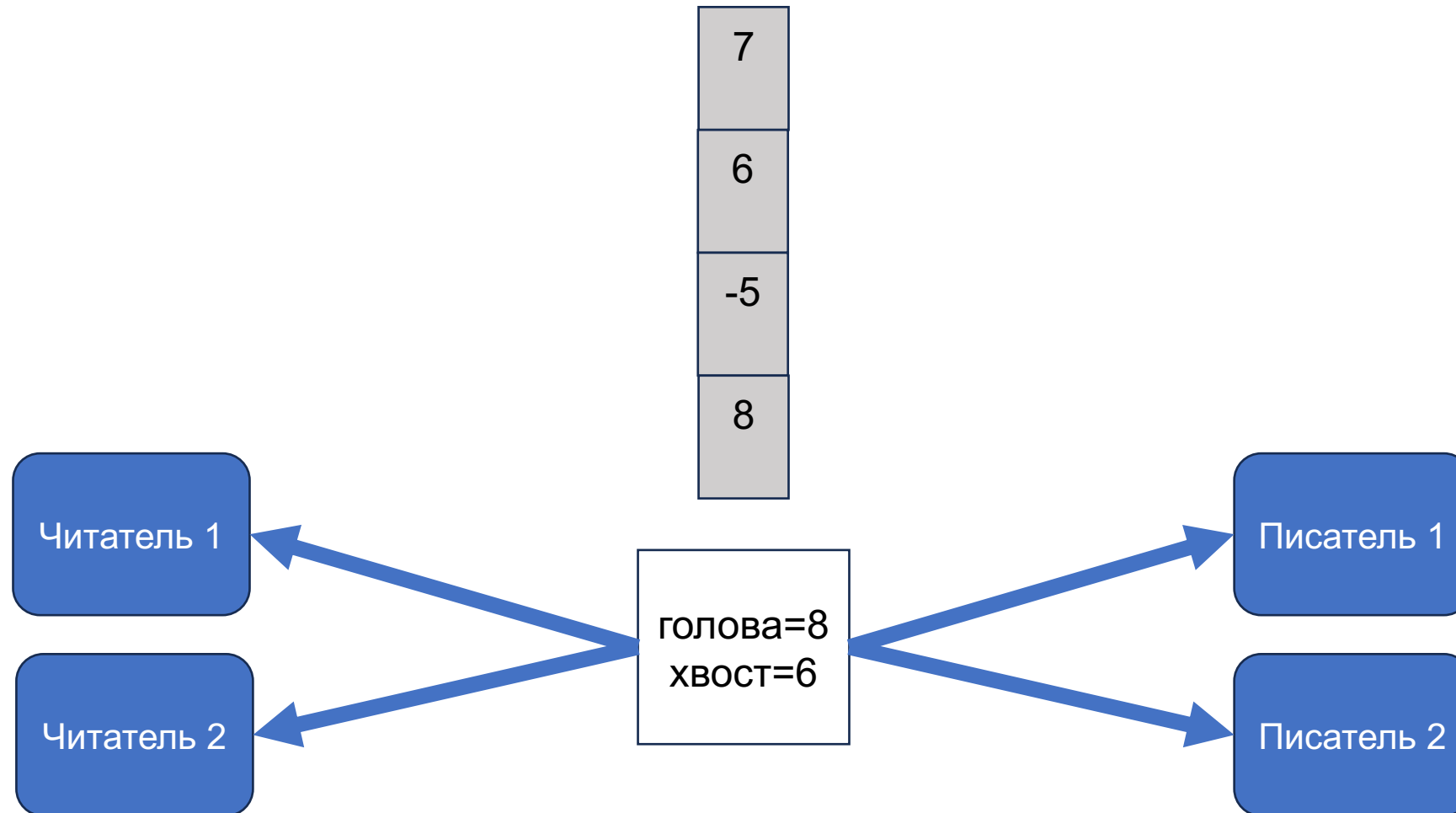


Оцениваем «скорость света» по циклическому буферу

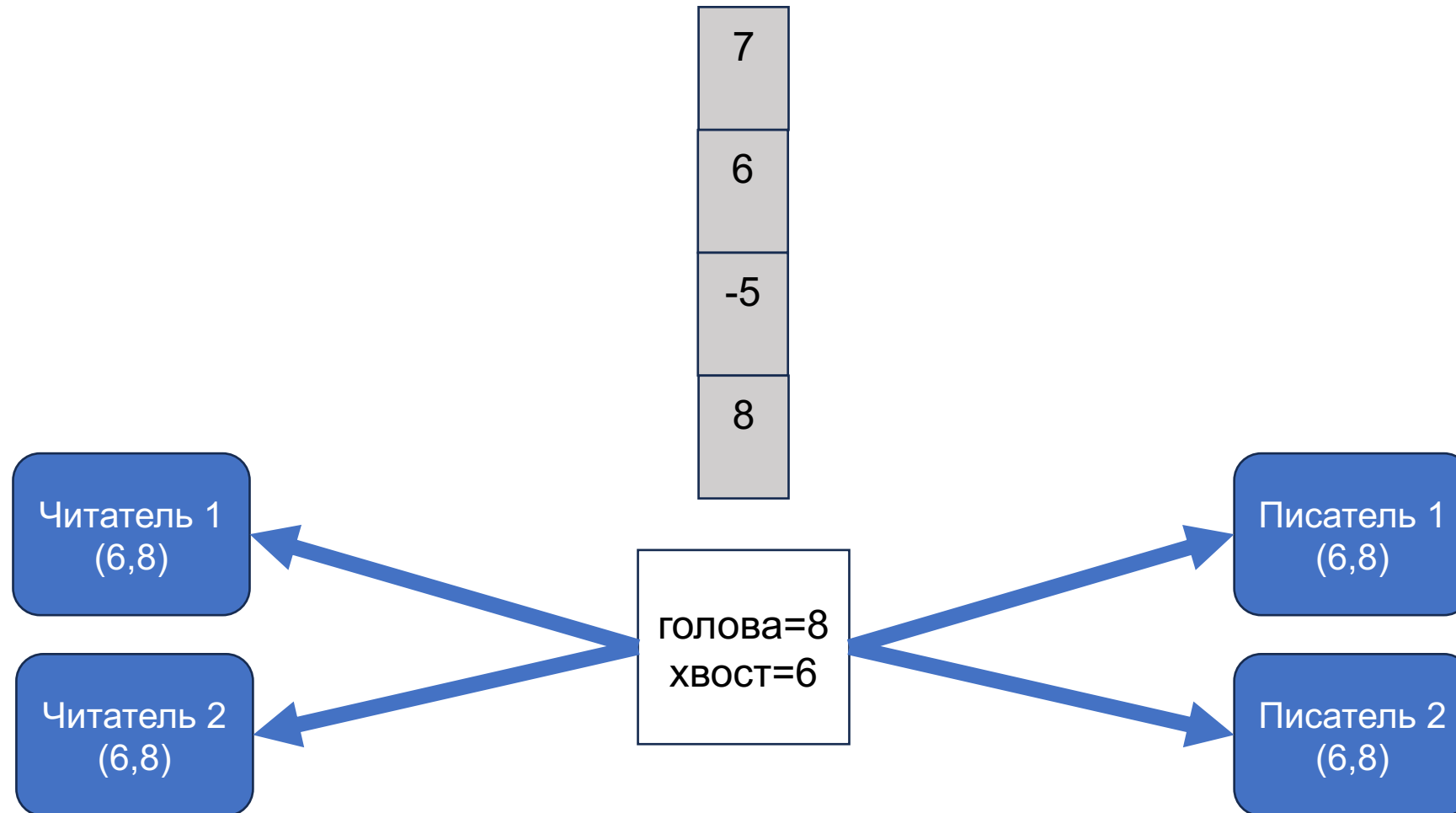
Очередь с циклическим буфером



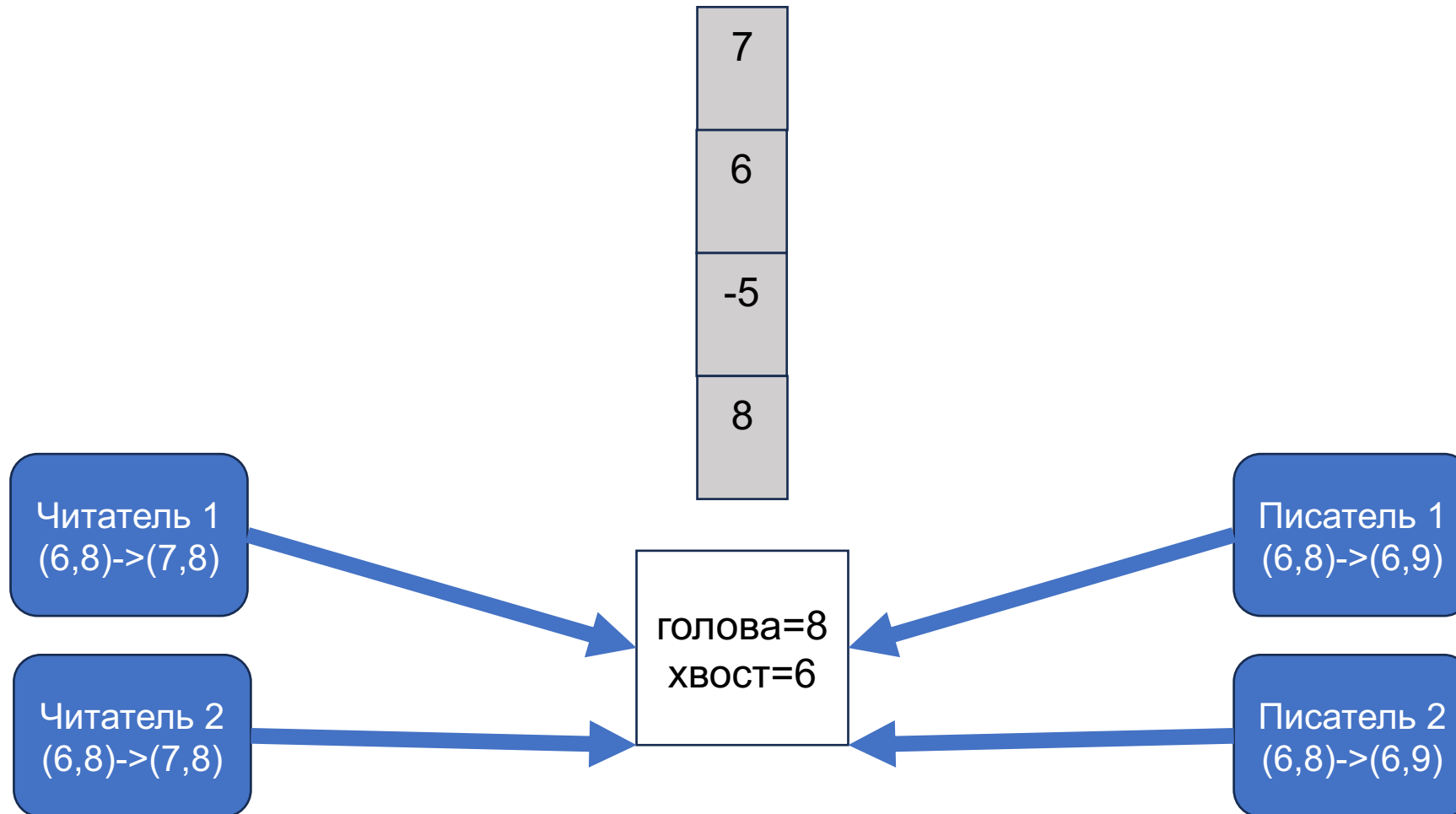
Очередь с циклическим буфером



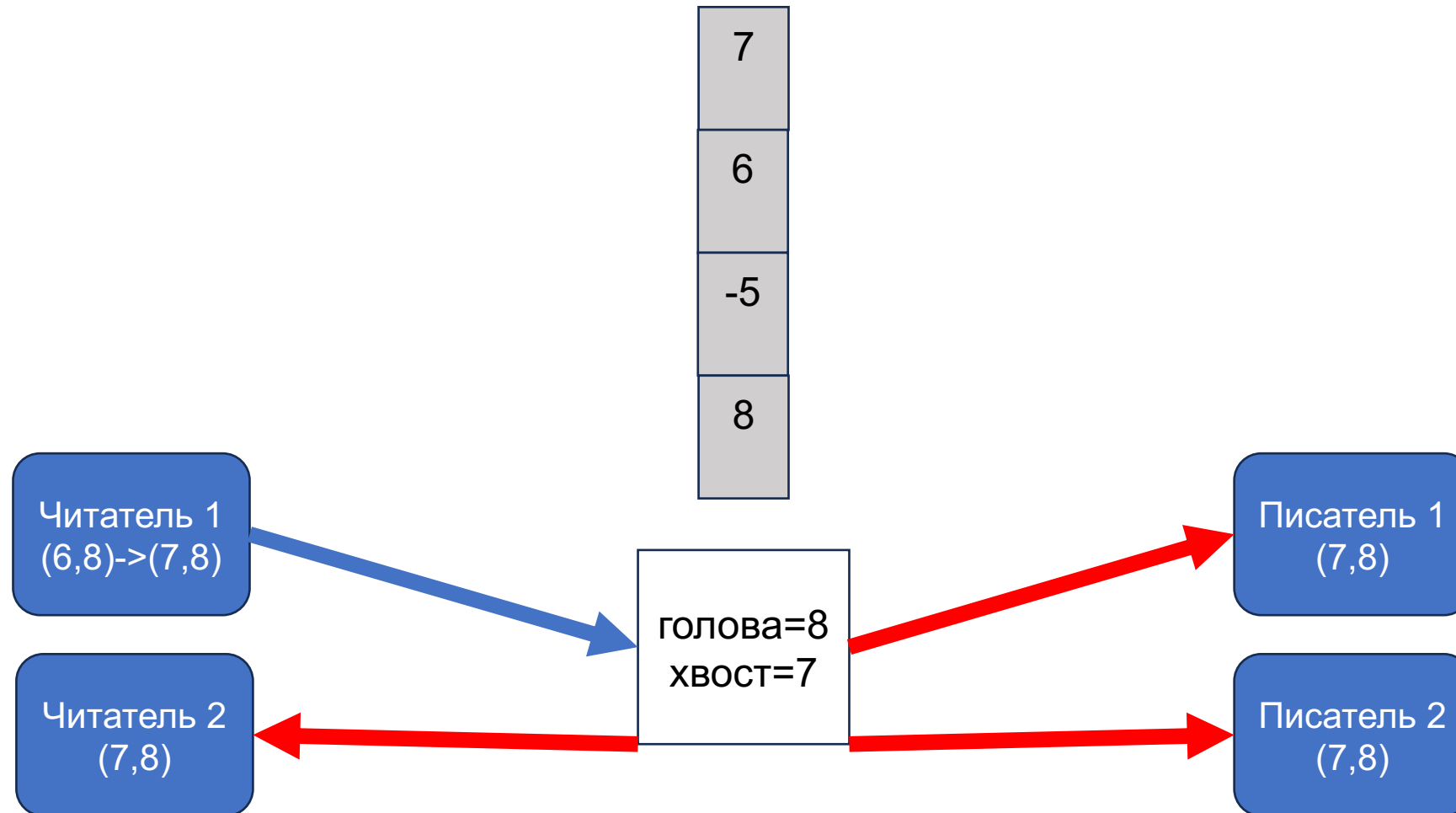
Очередь с циклическим буфером



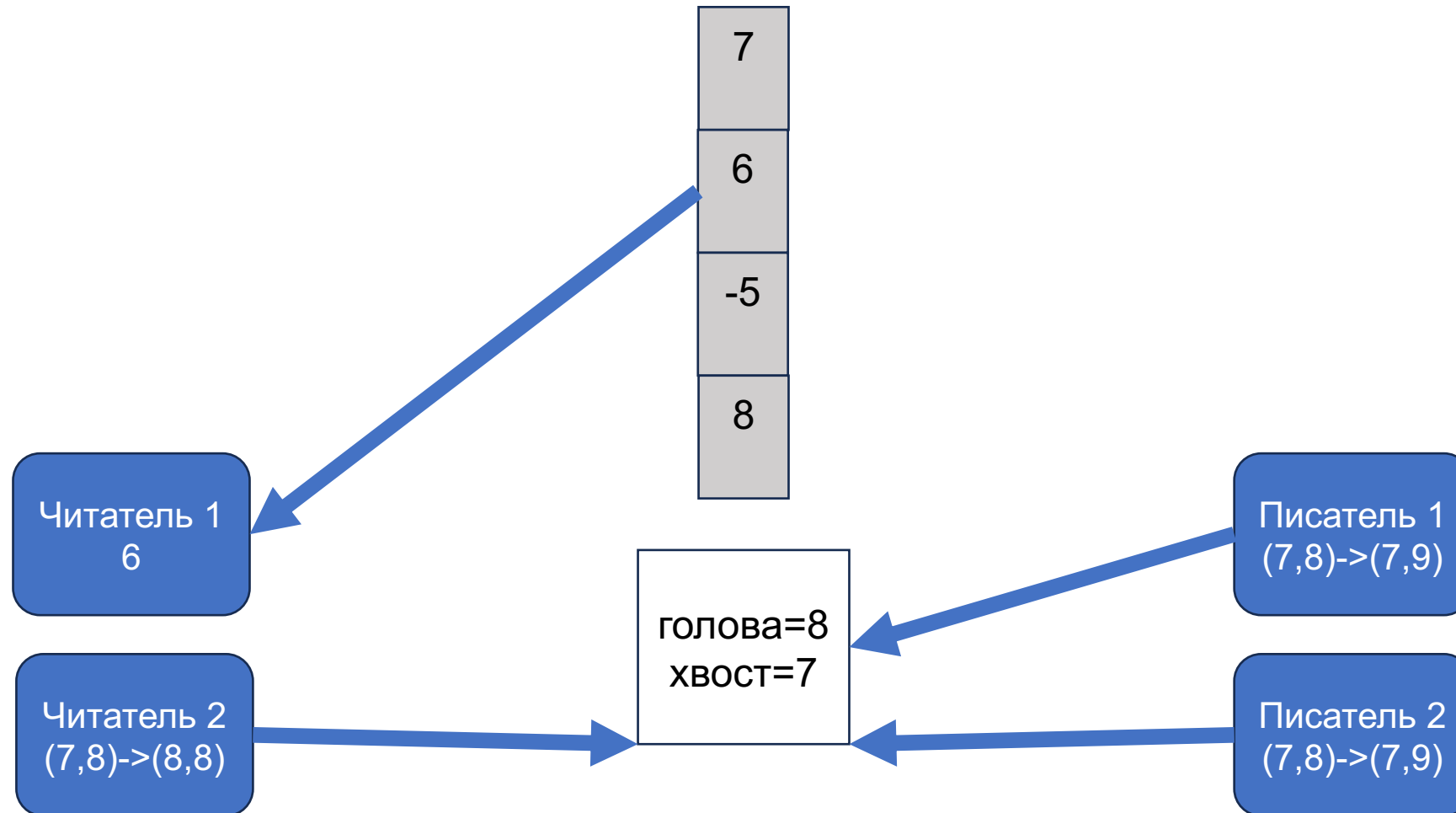
Очередь с циклическим буфером



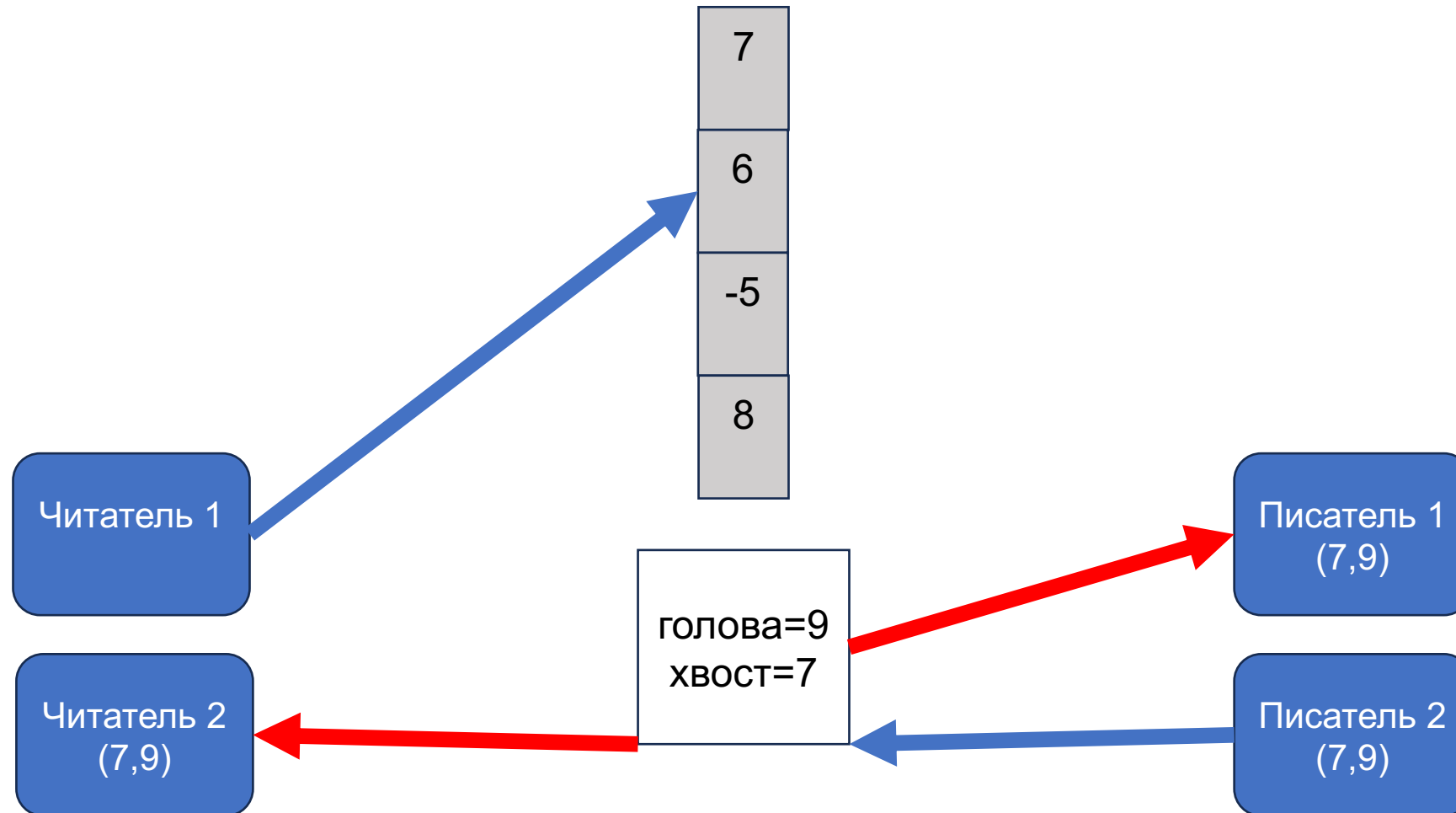
Очередь с циклическим буфером



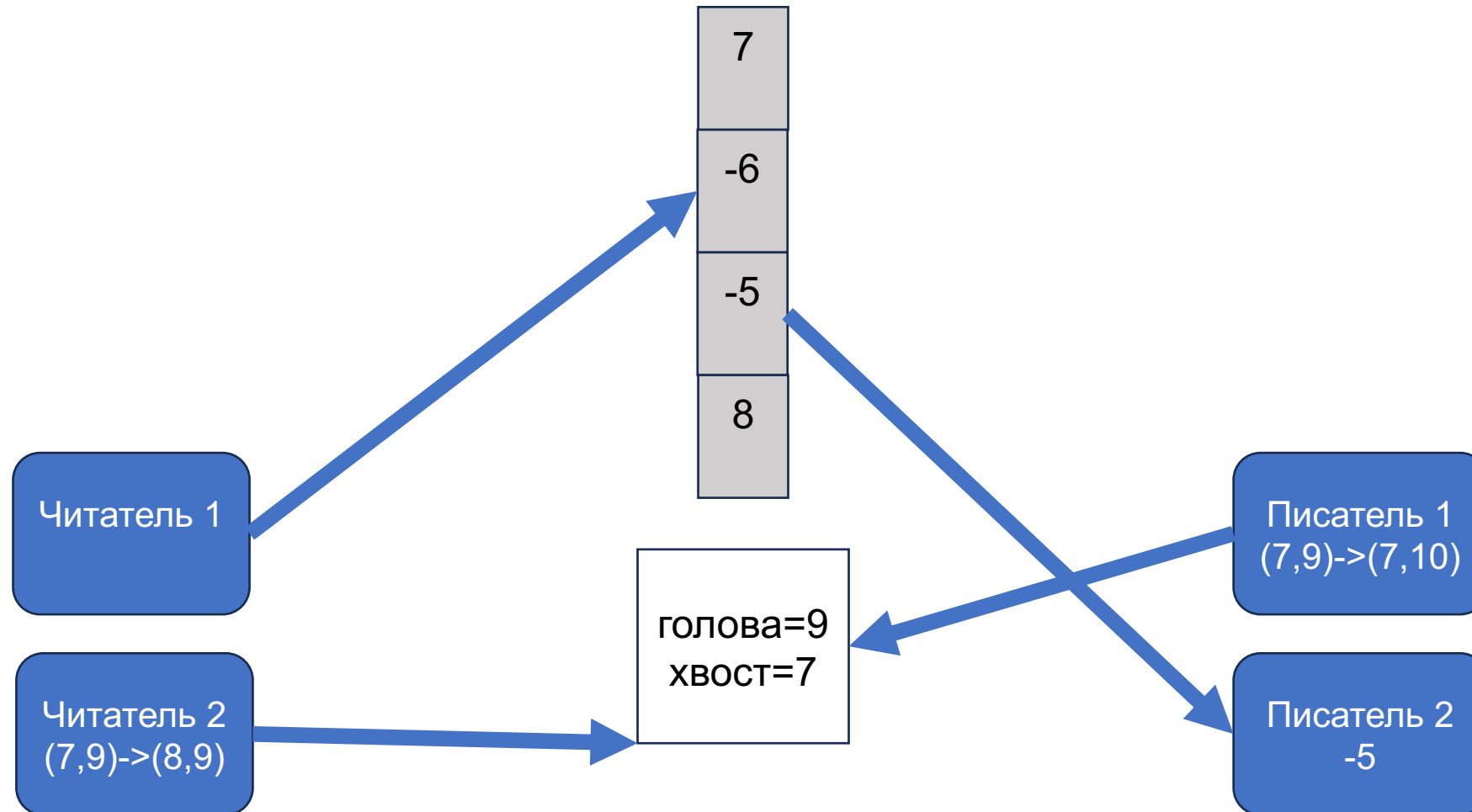
Очередь с циклическим буфером



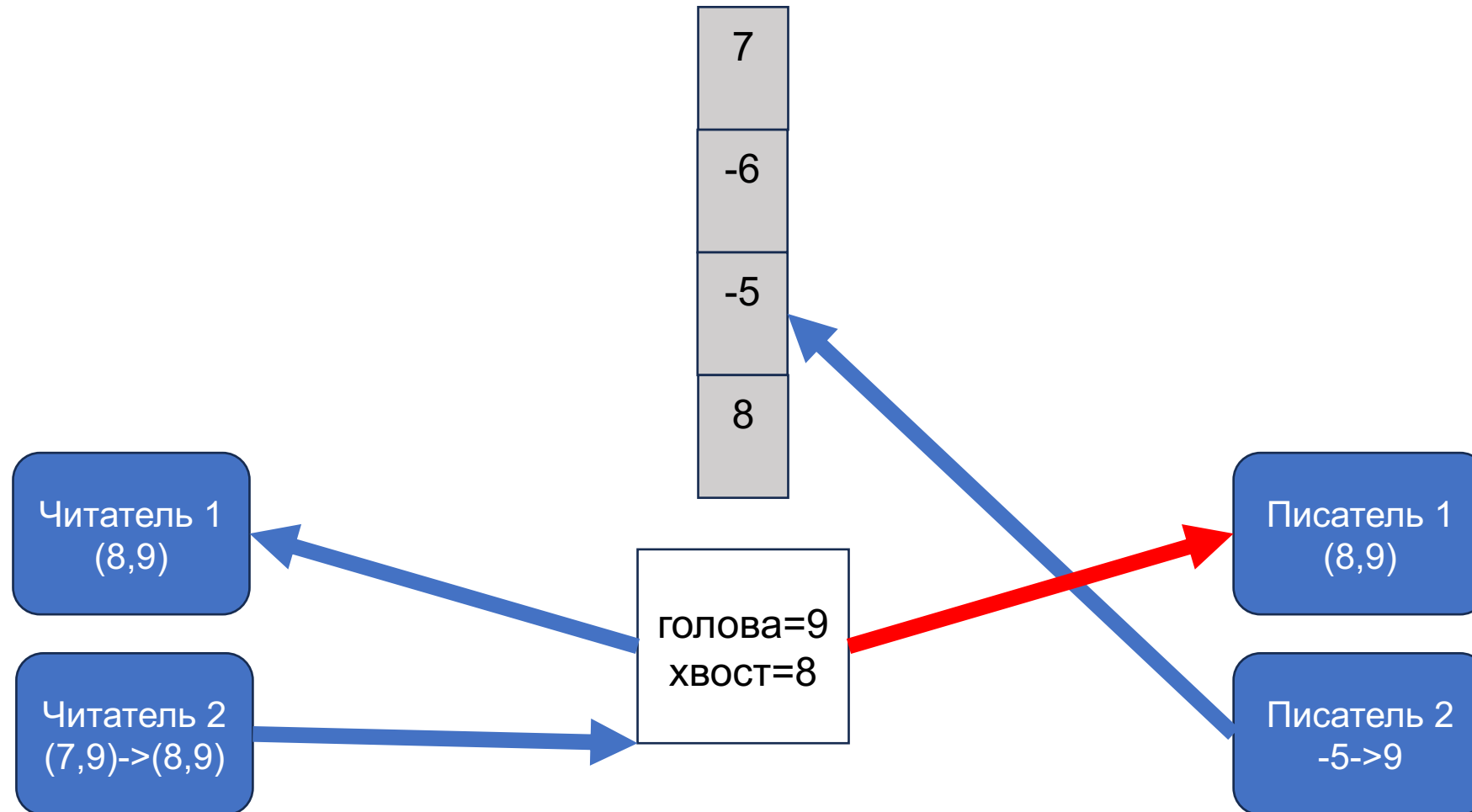
Очередь с циклическим буфером



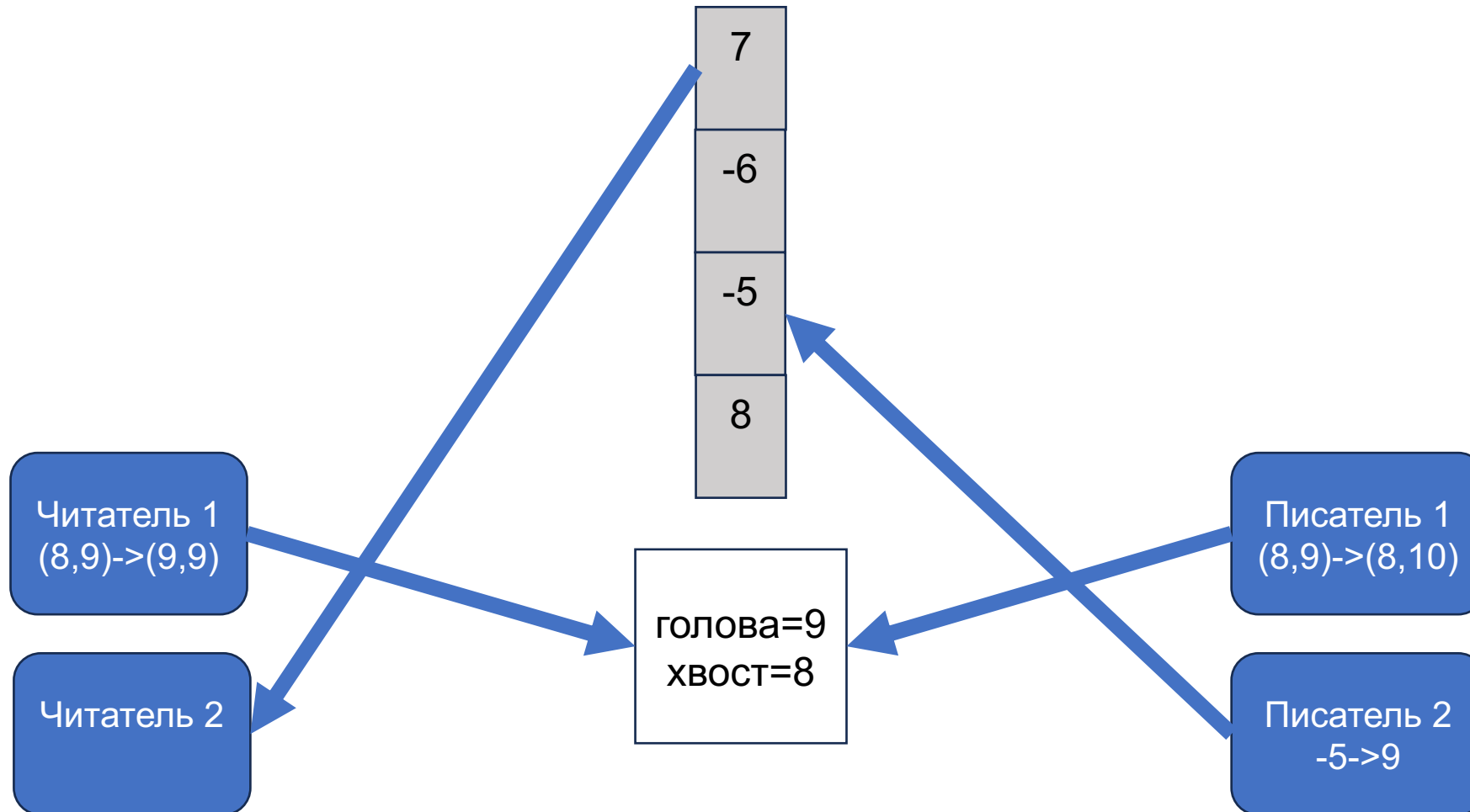
Очередь с циклическим буфером



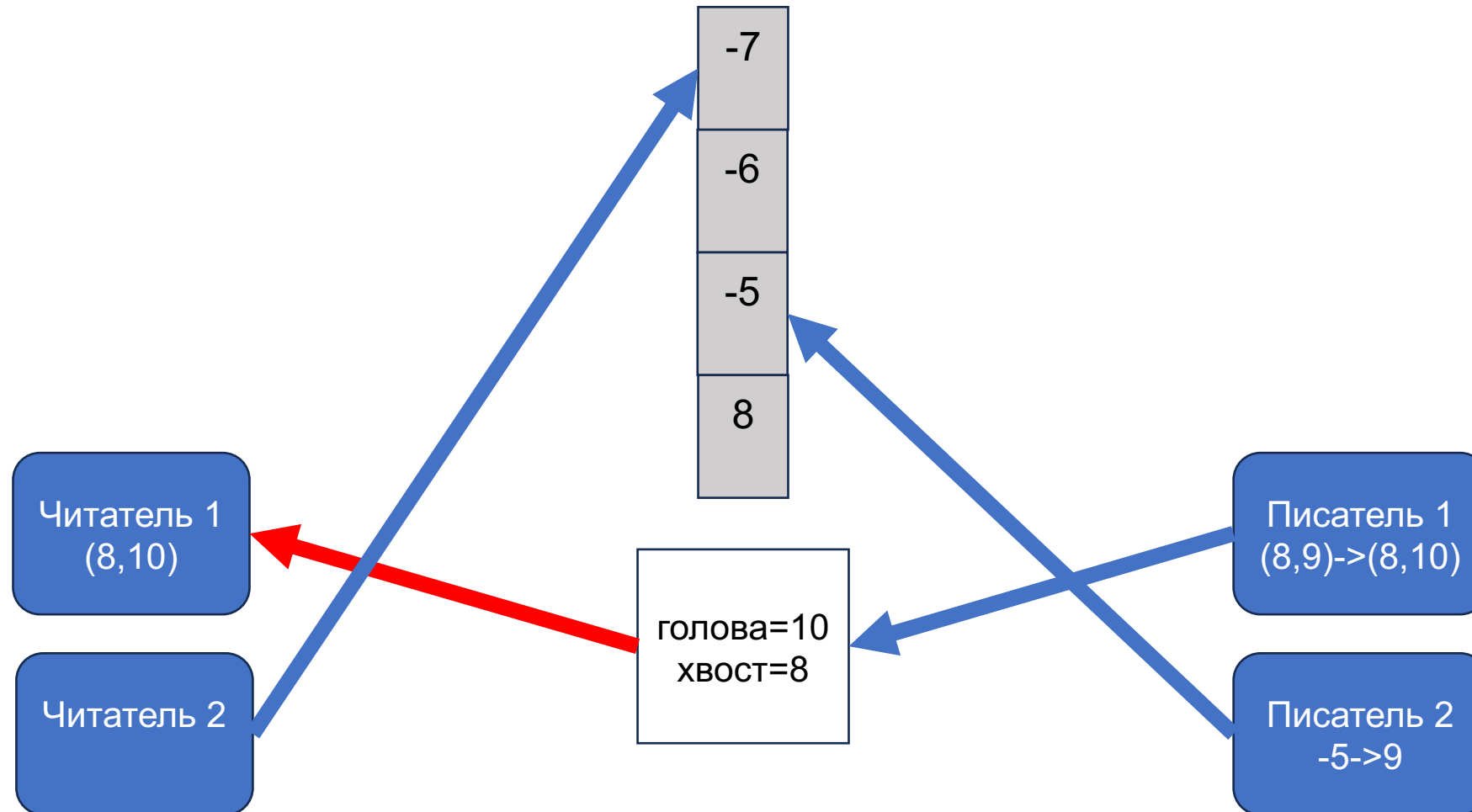
Очередь с циклическим буфером



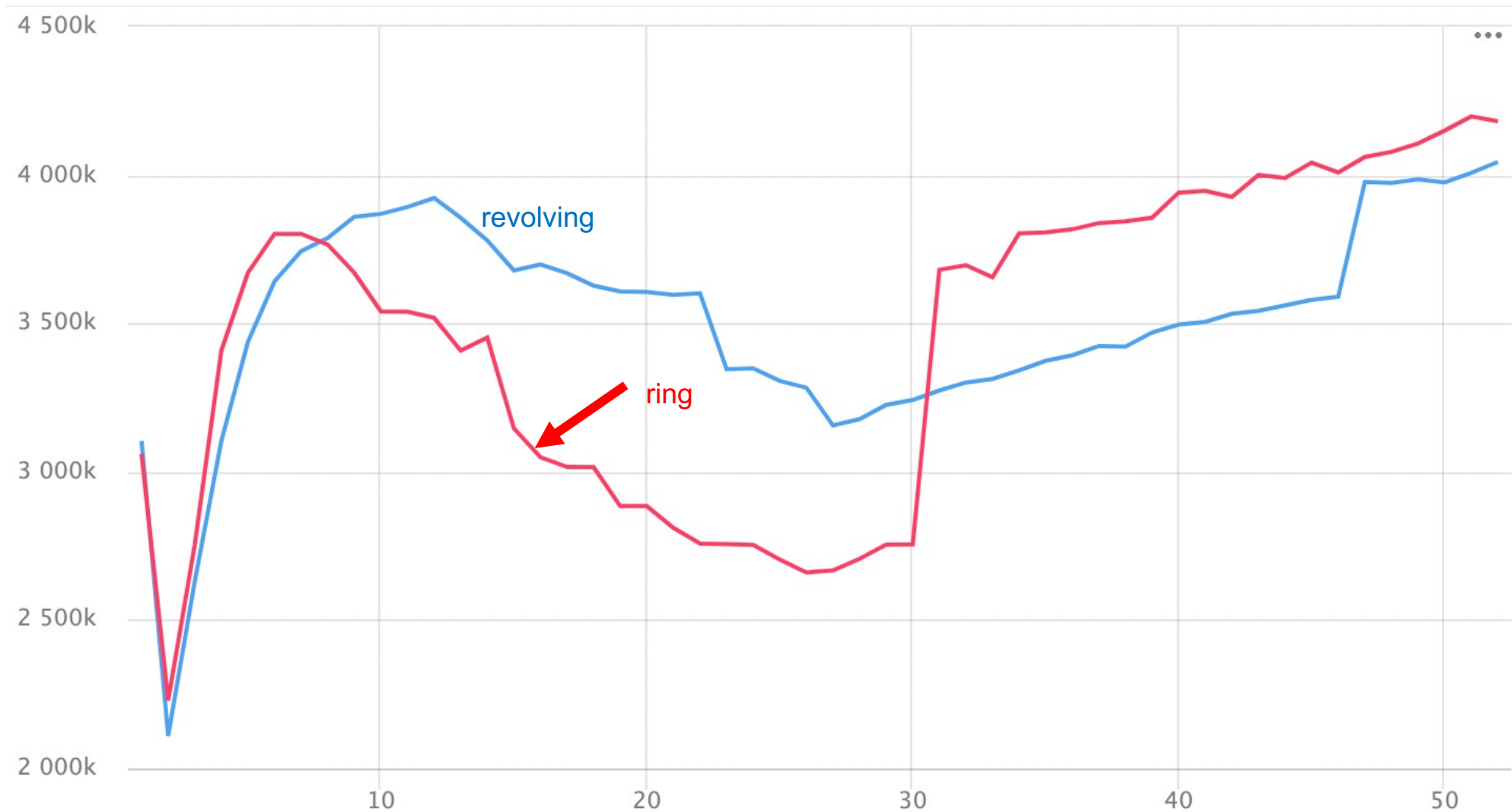
Очередь с циклическим буфером



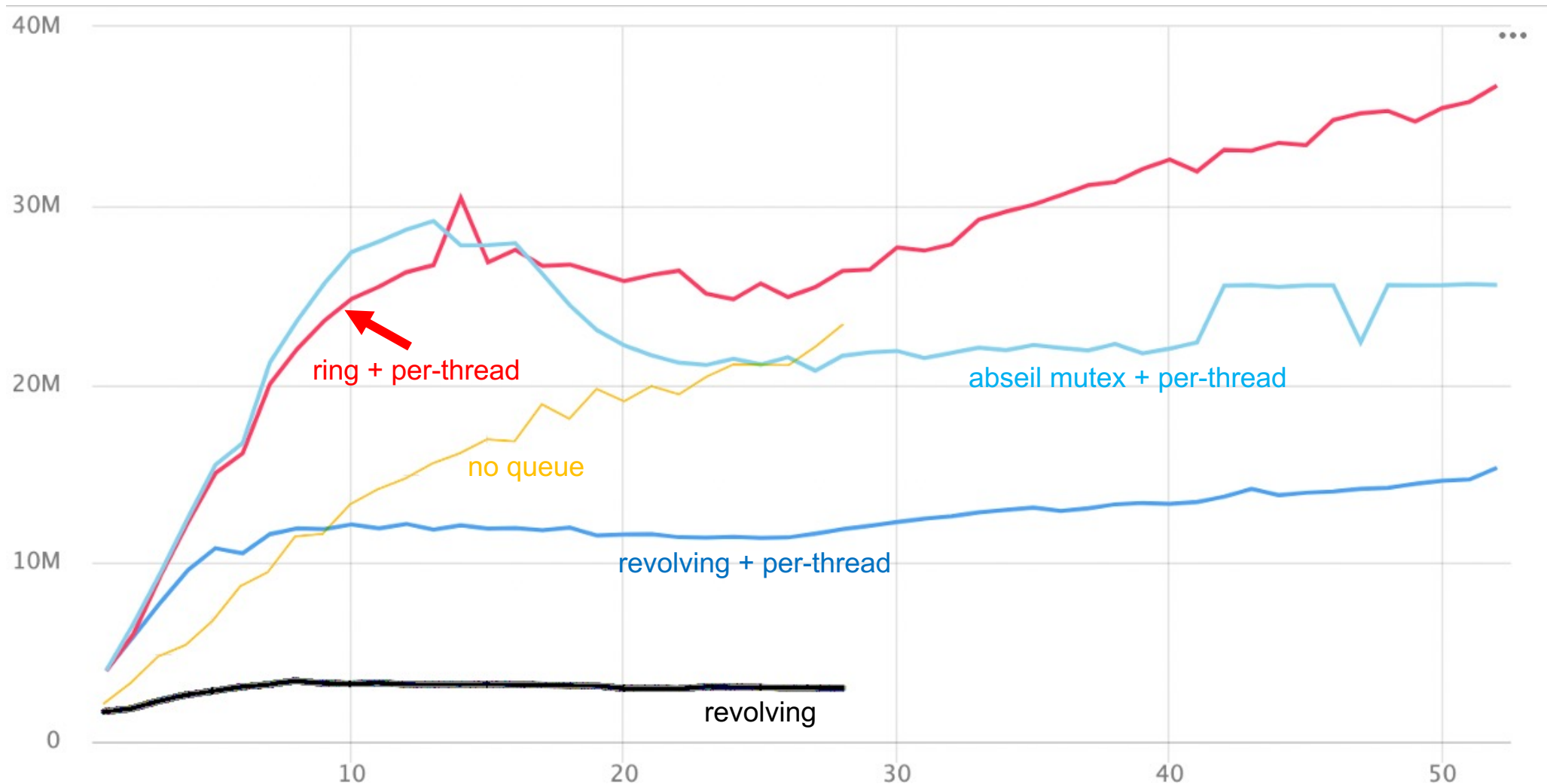
Очередь с циклическим буфером



Очередь с циклическим буфером



Очередь с циклическим буфером + короткие per-thread очереди



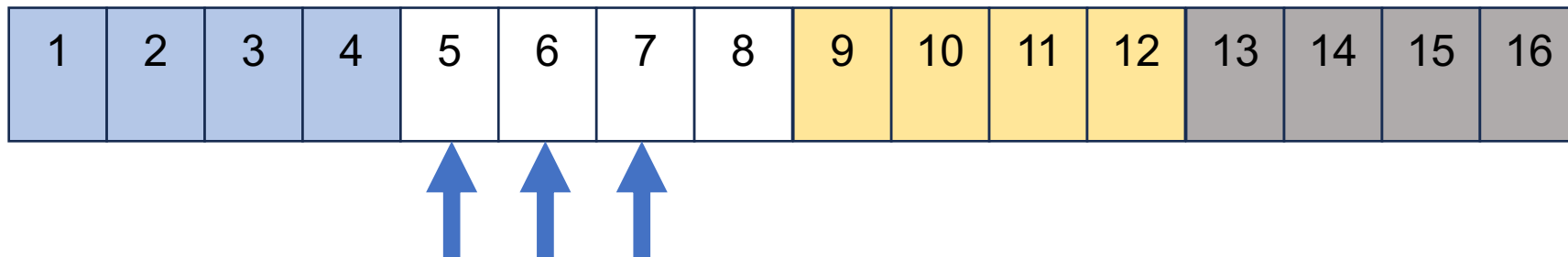
Важные доработки

Важные доработки

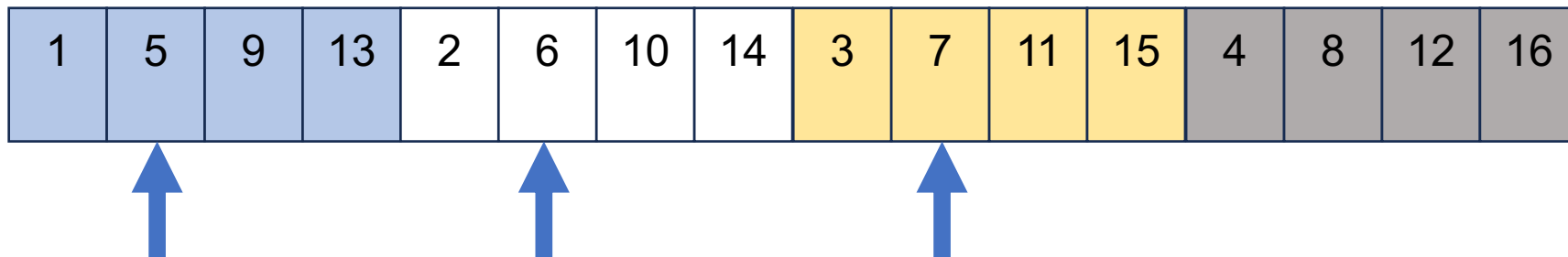
- Корректная обработка случаев, когда застрявший читатель/писатель отстал на круг и более
- Fallback на Revolving MPMC queue при переполнении

Раскладка по кэш-линиям

до:

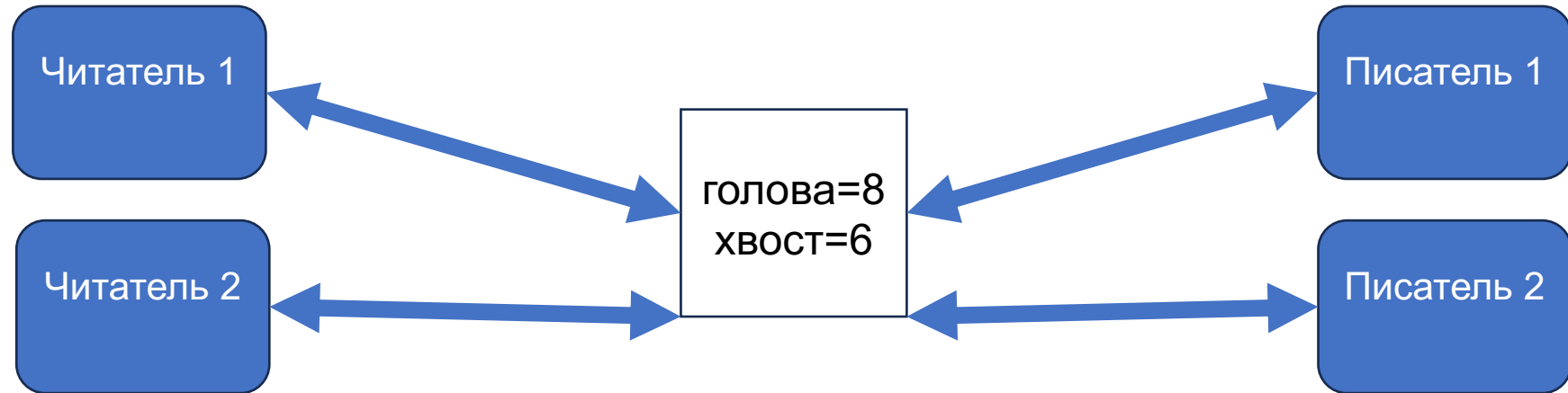


после:

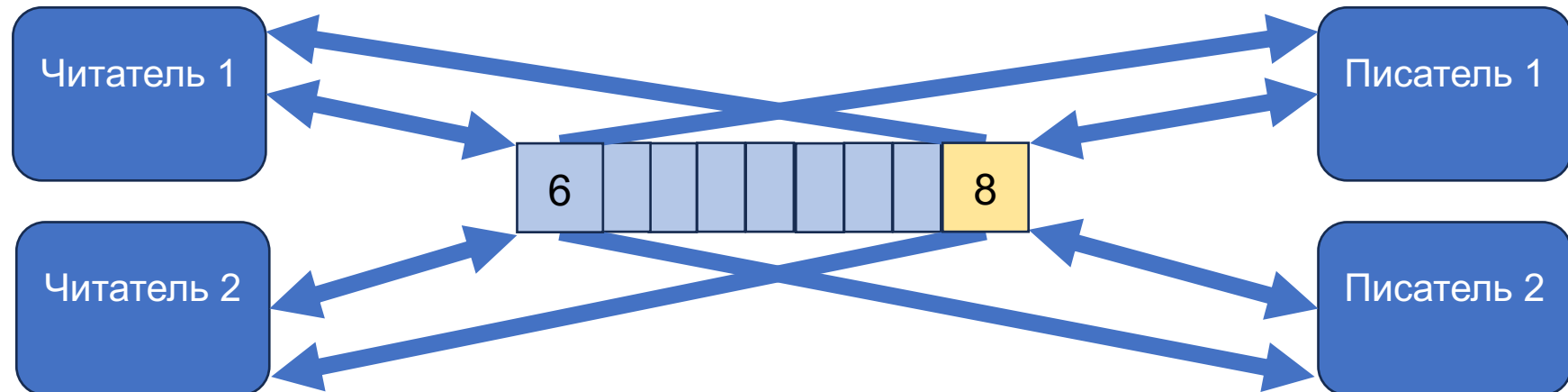


Горячие атомарные переменные

до:

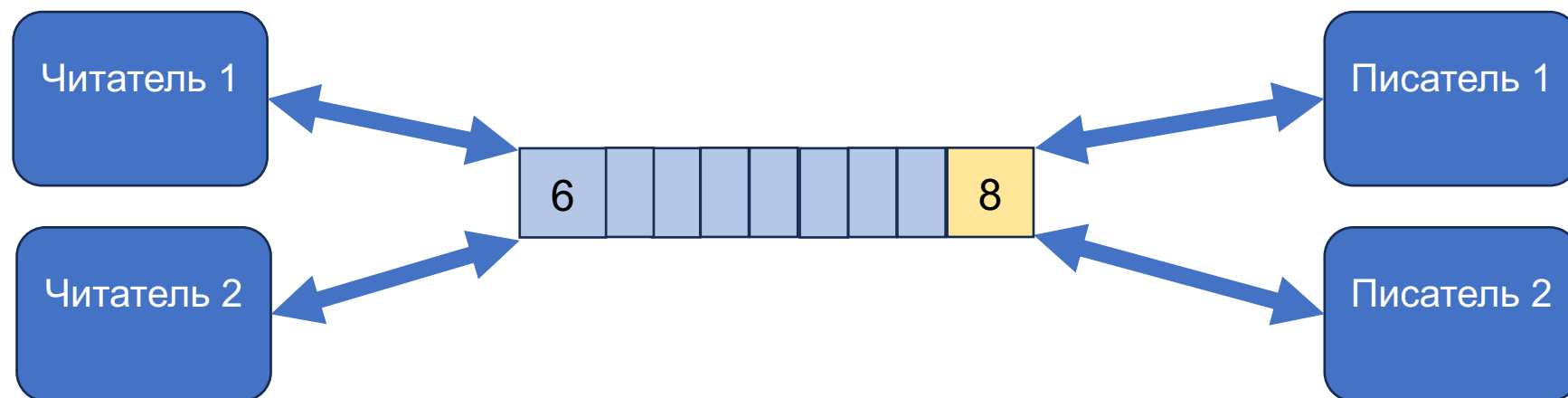


после:

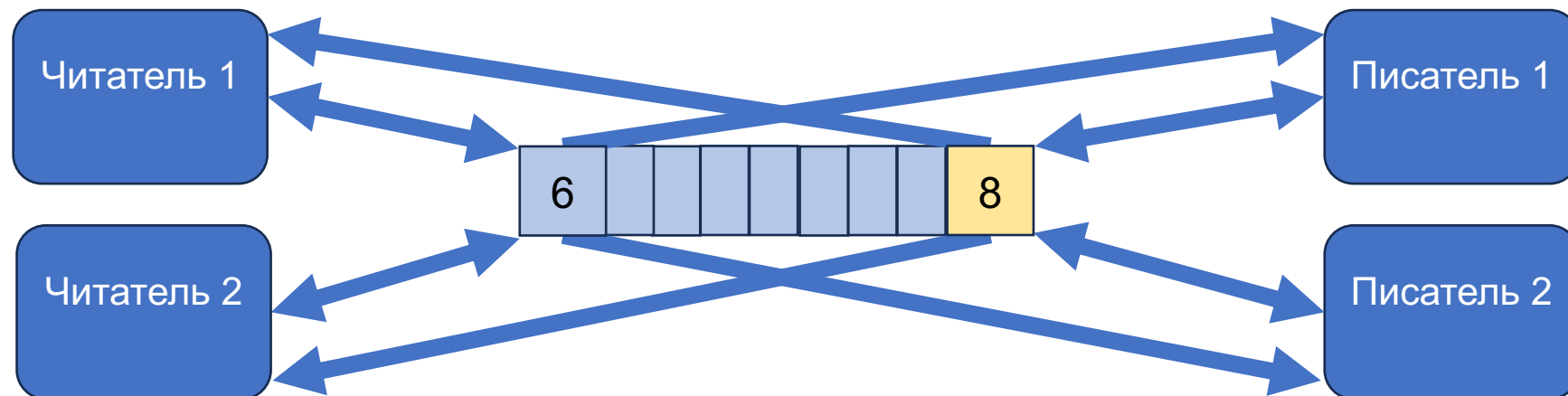


Быстро, но с гонкой, или медленно и без?

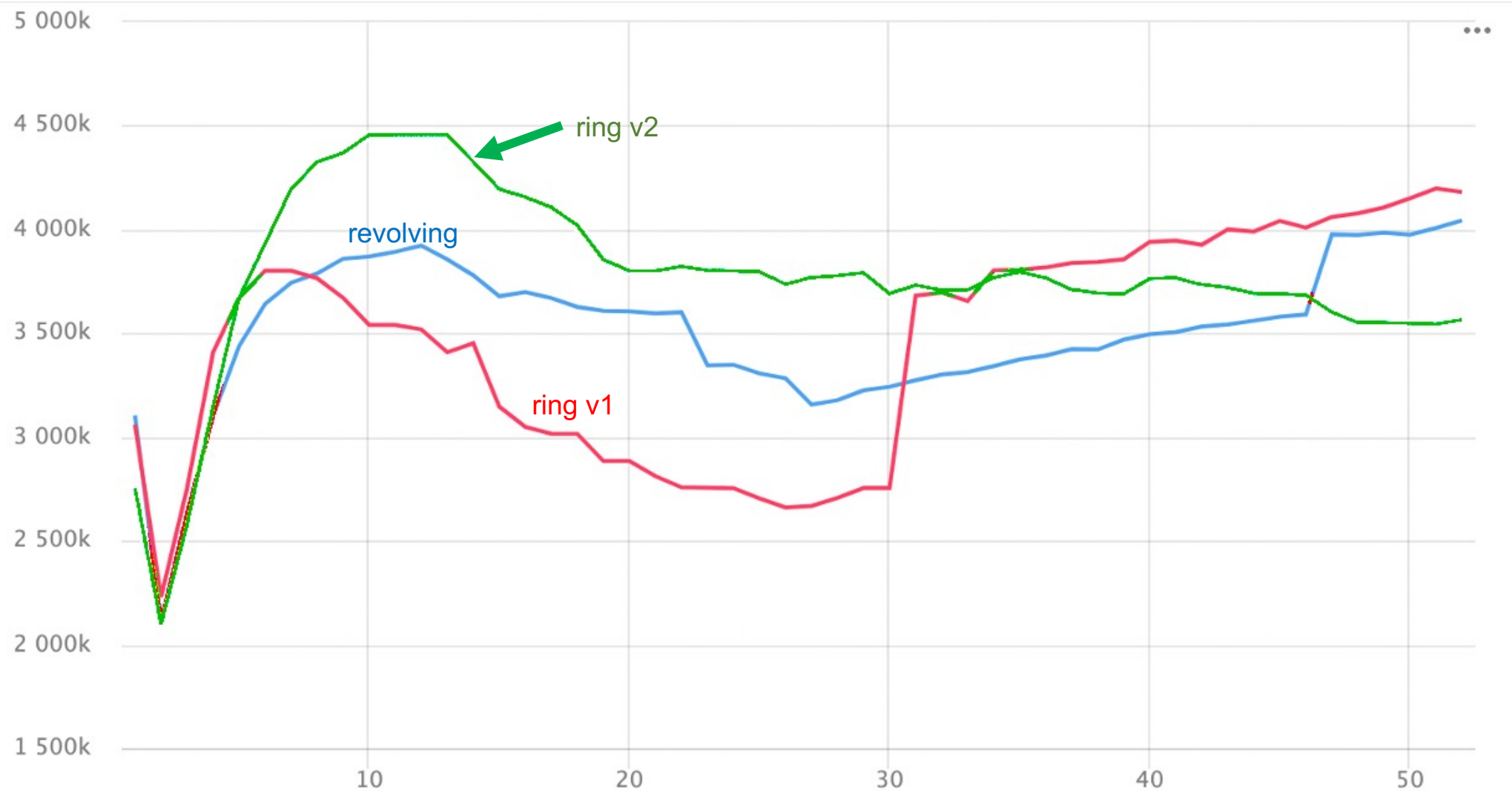
быстро:



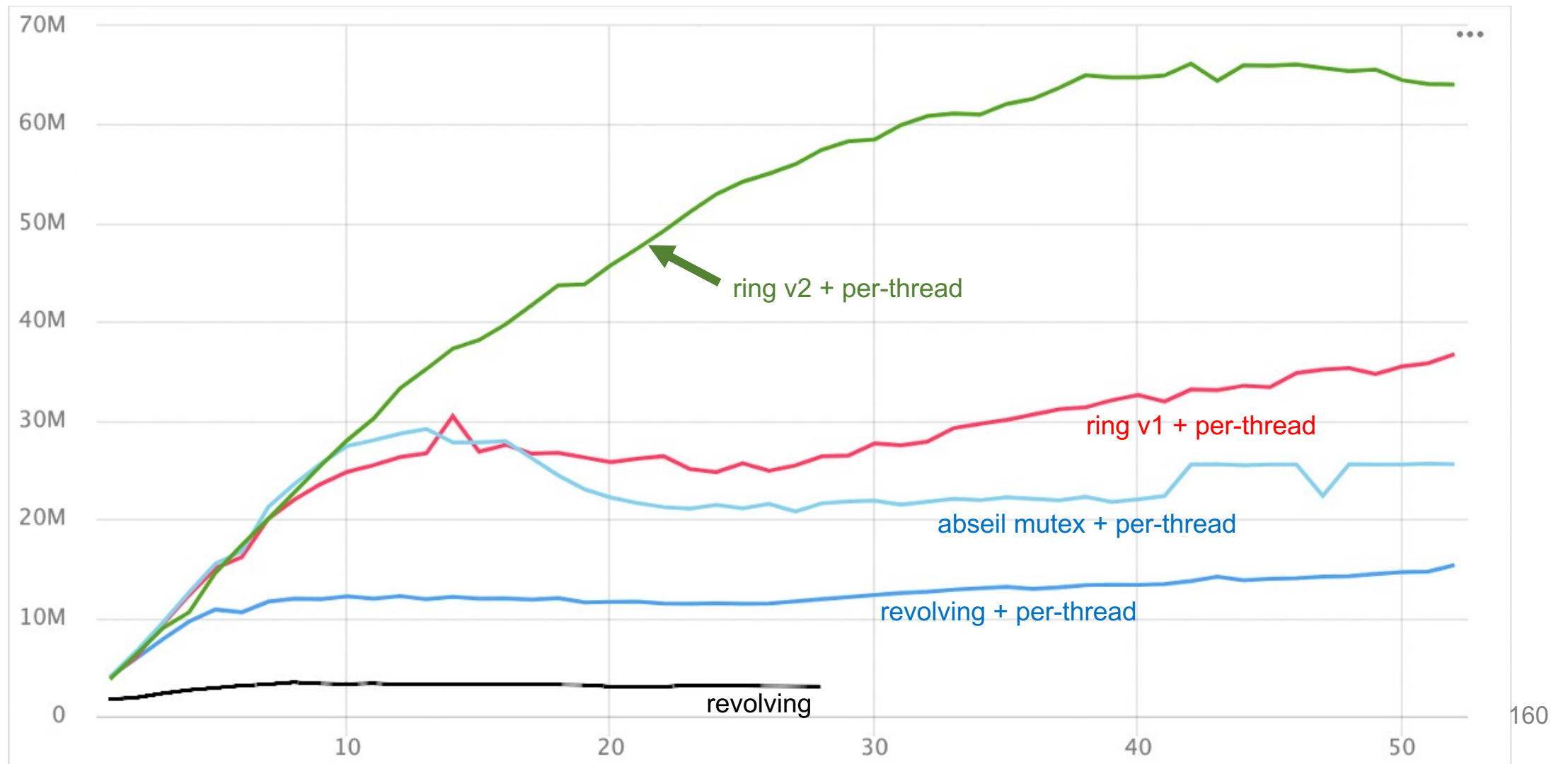
без гонки:



Очередь с циклическим буфером v2



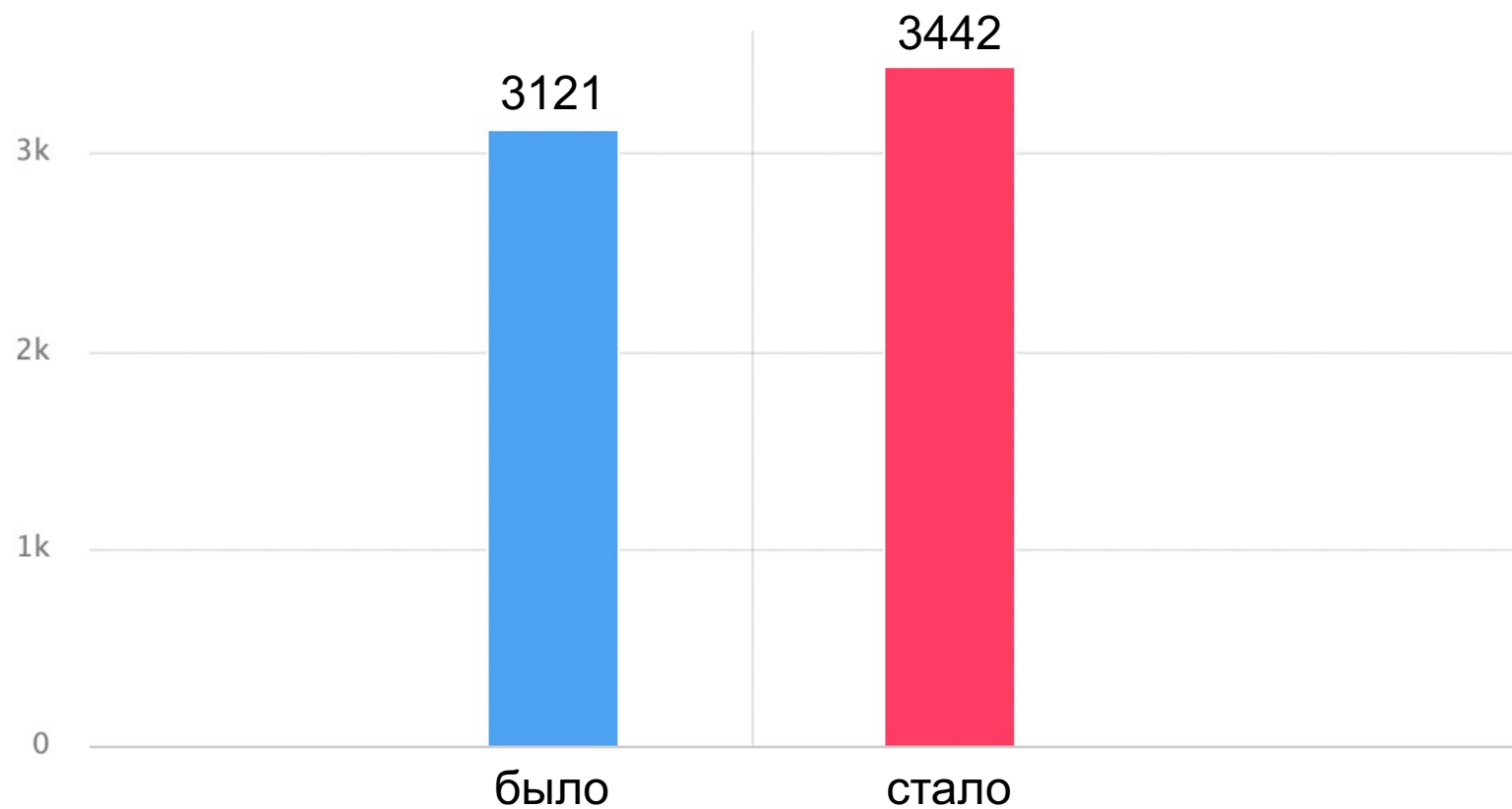
Очередь с циклическим буфером v2 + короткие per-thread очереди



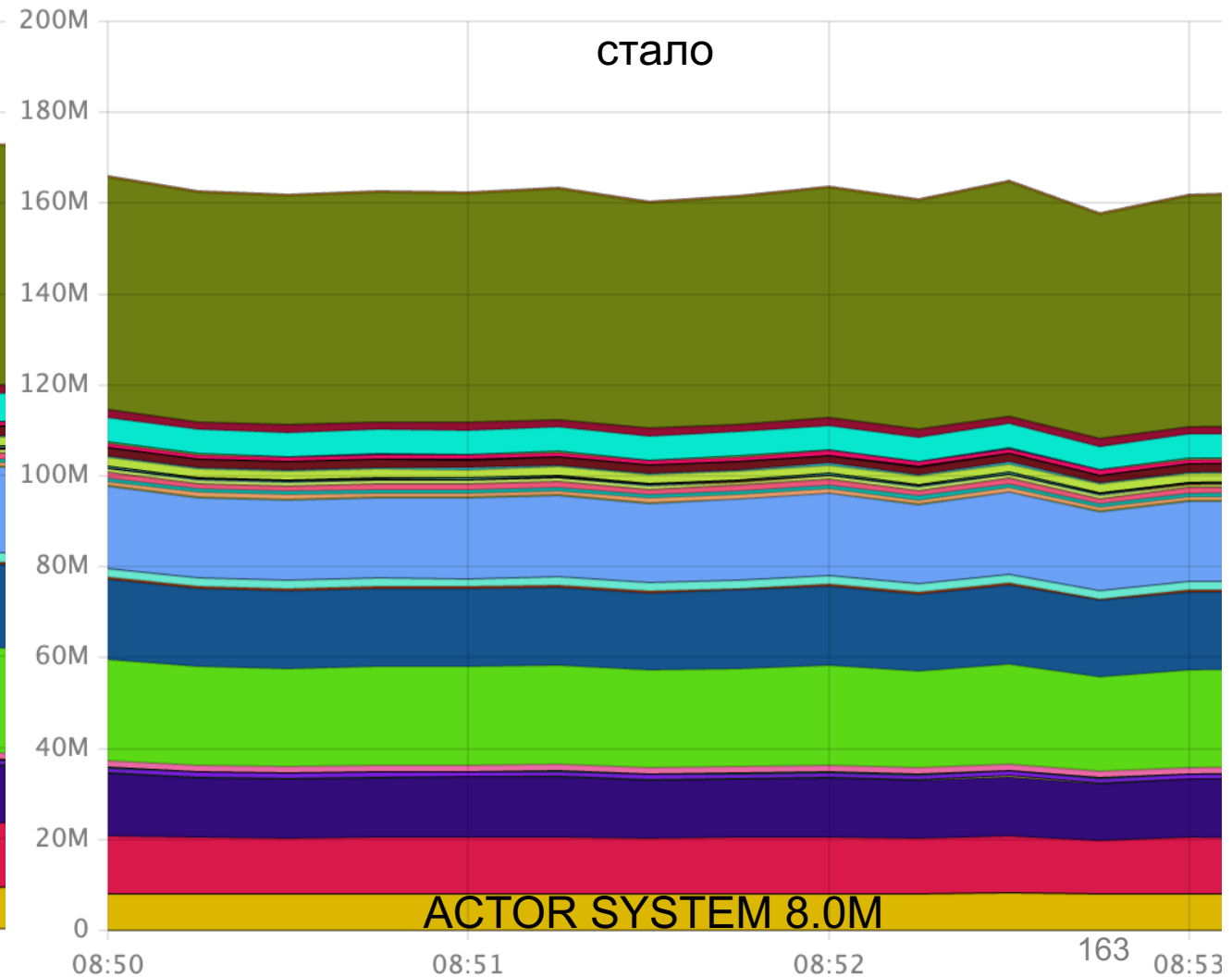
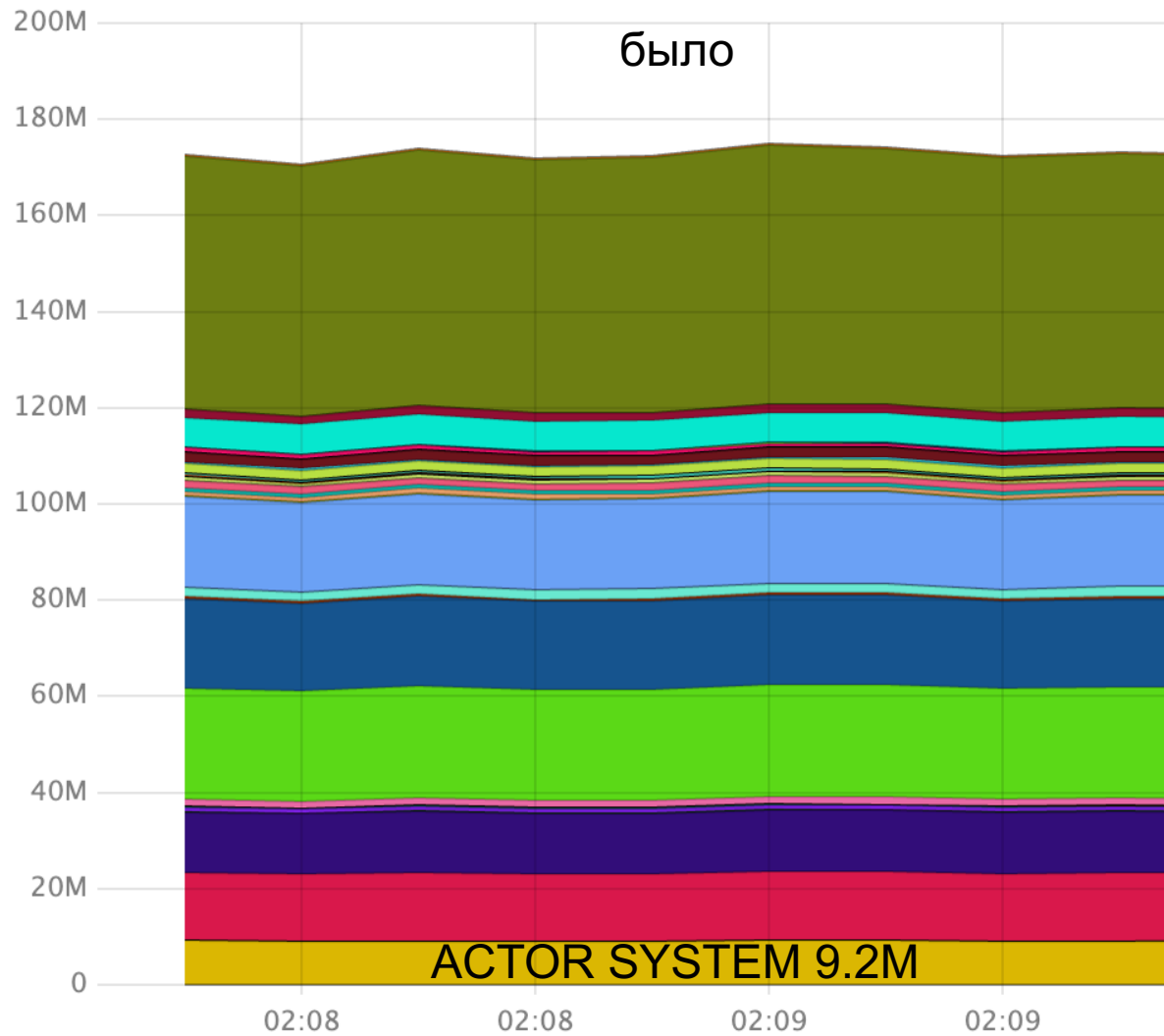
Влияние на производительность YDB

Результат

Нагрузочный тест YDB workload, транзакций в секунду



Результат



было:

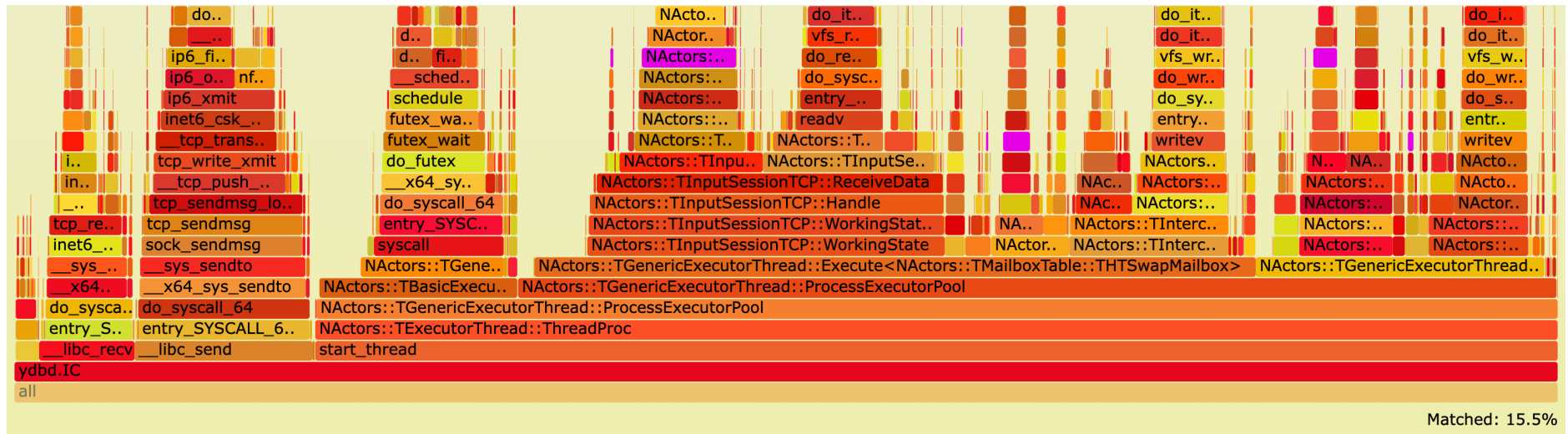
	do..	d..	NActo..	do_it..	do_it..	do_i..
	__..	d..	NActor..	vfs_r..	do_it..	do_it..
	ip6_fi..	d.. fi..	NActors::	do_re..	vfs_wr..	vfs_w..
	ip6_o.. nf..	_sched..	NActors::	do_sysc..	do_wr..	do_wr..
	ip6_xmit	schedule	NActors::	entry_..	do_sy..	do_s..
	inet6_csk..	futex_wa..	NActors:::	readv	entry..	entr..
	_tcp_trans..	futex_wait	NActors::T..	NActors::T..	writew	writew
i..	tcp_write_xmit	do_futex	NActors::TInpu..	NActors::TInputSe..	NActors..	N.. NA..
in..	_tcp_push..	_x64_sy..	NActors::TInputSessionTCP::ReceiveData		NAC.. NActors::	NActors::
__.	tcp_sendmsg_lo..	do_syscall_64	NActors::TInputSessionTCP::Handle		NAC.. NActors::	NActors::
tcp_re..	tcp_sendmsg	entry_SYSC..	NActors::TInputSessionTCP::WorkingStat..	NA..	NActors::TInterc..	NActors::
inet6_..	sock_sendmsg	syscall	NActors::TInputSessionTCP::WorkingState	NActor..	NActors::TInterc..	NActors::
sys..	_sys_sendto	NActors::TGene..	NActors::TGenericExecutorThread::Execute<NActors::TMailboxTable::THTSwapMailbox>			NActors::TGenericExecutorThread..
x64..	__x64_sys_sendto	NActors::TBasicExecu..	NActors::TGenericExecutorThread::ProcessExecutorPool			
do_sysca..	do_syscall_64	NActors::TGenericExecutorThread::ProcessExecutorPool				
entry_S..	entry_SYSCALL_6..	NActors::TExeutorThread::ThreadProc				
__libc_rcv	__libc_send	start_thread				
ydbd.IC						
all						

Matched: 21.8%

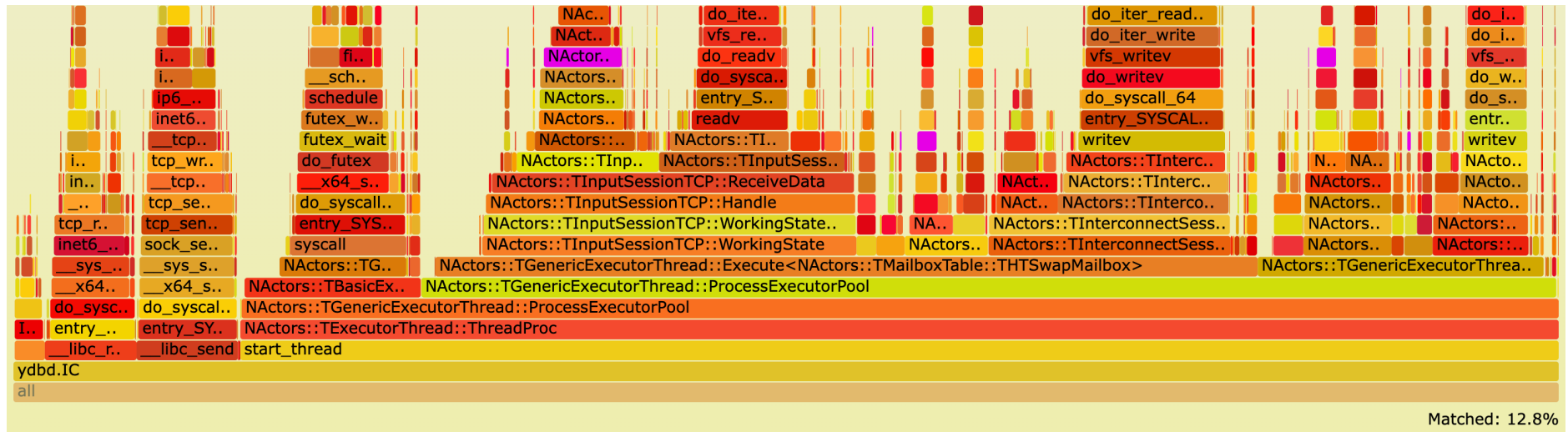
[illegible]

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/agrianus/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



Спасибо!

**Алексей Станкевичус,
ydb.tech**