

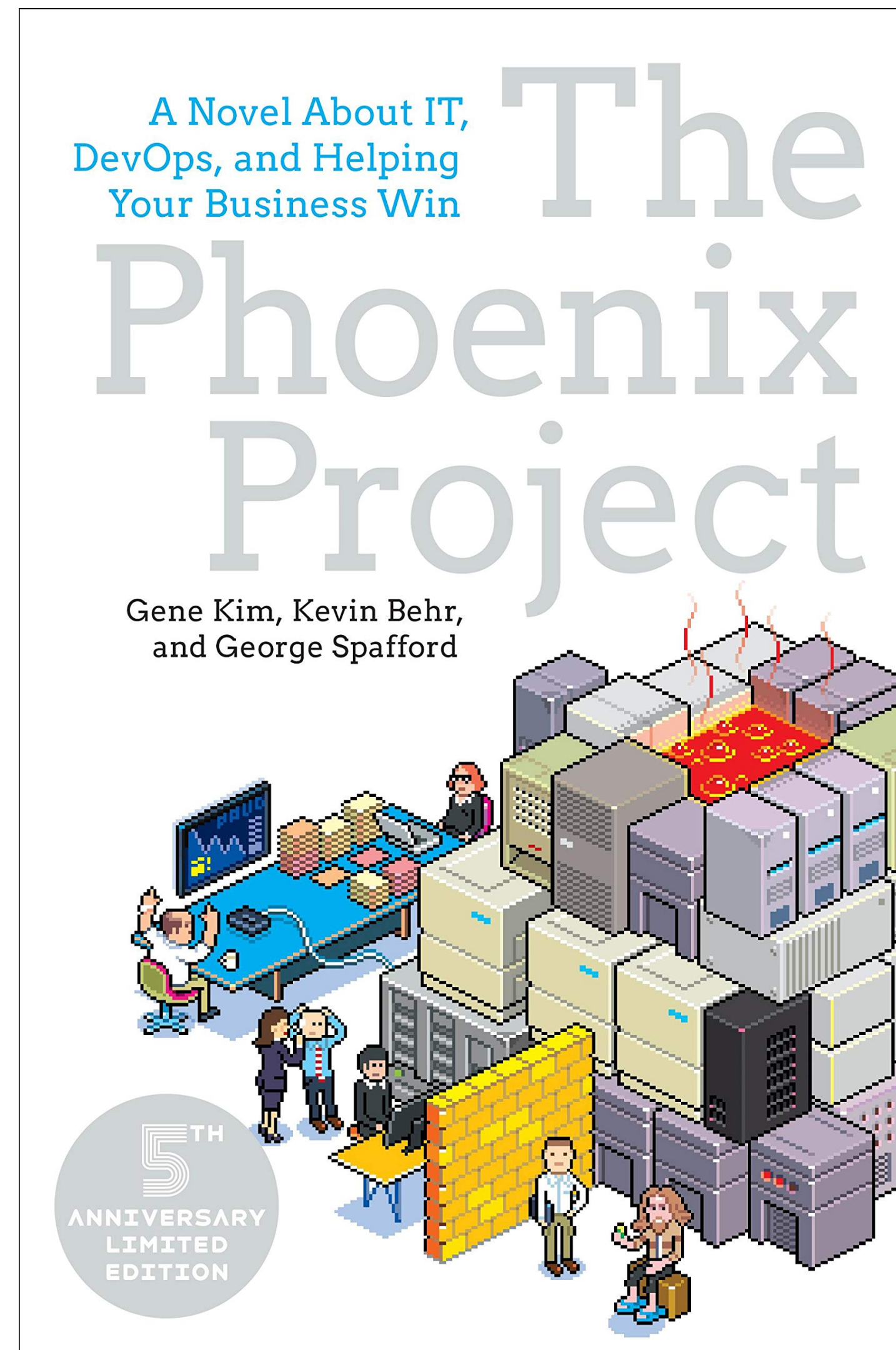
0 графике из книги «Проект Феникс»

Иван Пономарёв, КУРС/МФТИ

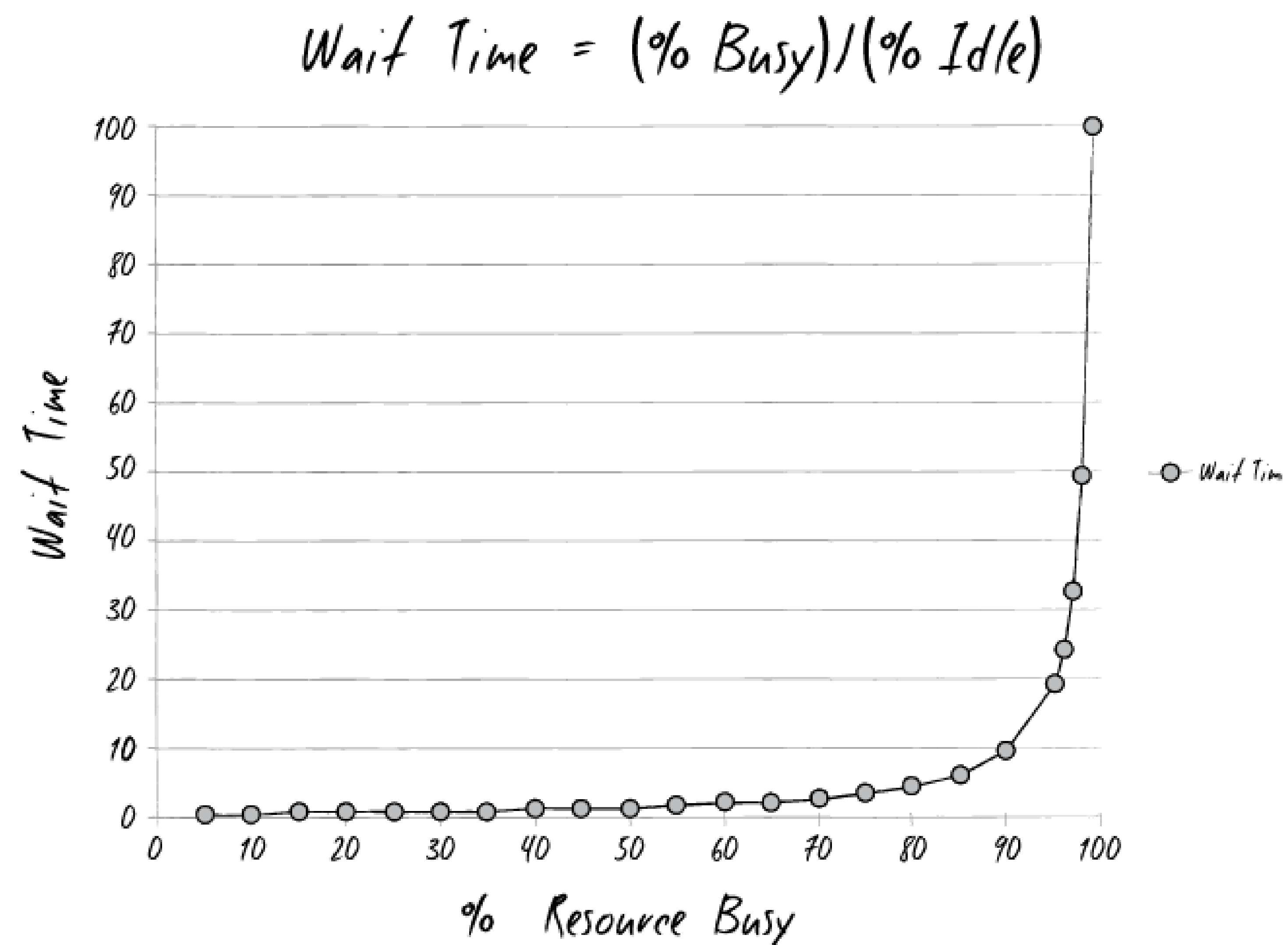
ponomarev@corchestra.ru

 [@inponomarev](https://twitter.com/inponomarev)

Кто читал эту книжку?



А кто знает, про что эта картинка?



WTF?

- Как так вышло, что чем «оптимальнее» загружен процессор, тем медленнее идёт процесс?
- И почему всё настолько плохо возле точки 100% загрузки процессора?

Любая очередь есть моделируемый процесс, подчиняющийся общим закономерностям

Любая очередь есть моделируемый процесс, подчиняющийся общим закономерностям

- Очередь из сообщений в топике Kafka,

Любая очередь есть моделируемый процесс, подчиняющийся общим закономерностям

- Очередь из сообщений в топике Kafka,
- очередь задач в вашей Жире,

Любая очередь есть моделируемый процесс, подчиняющийся общим закономерностям

- Очередь из сообщений в топике Kafka,
- очередь задач в вашей Жире,
- очередь на кассу в магазин,

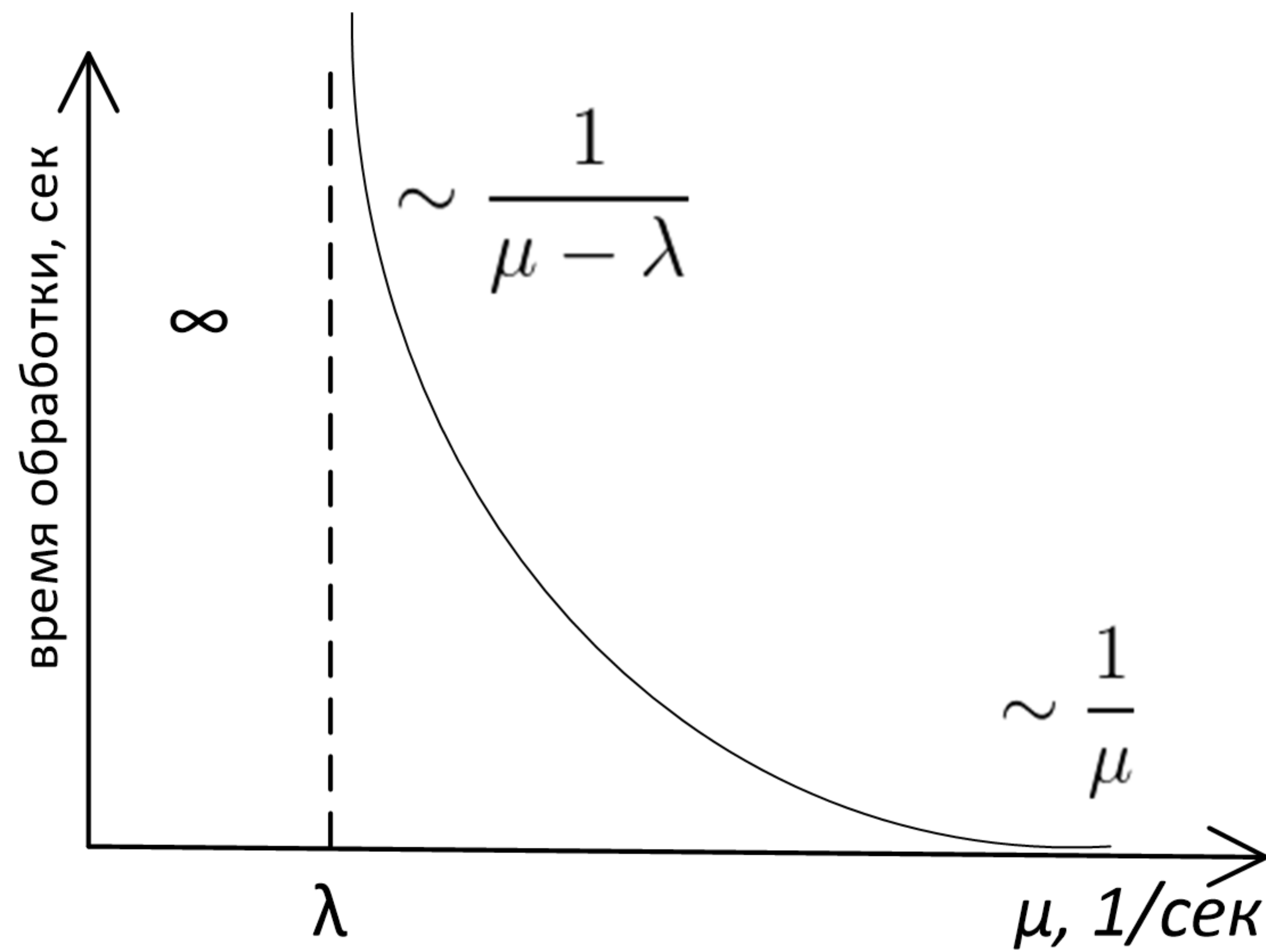
вообще любая очередь.



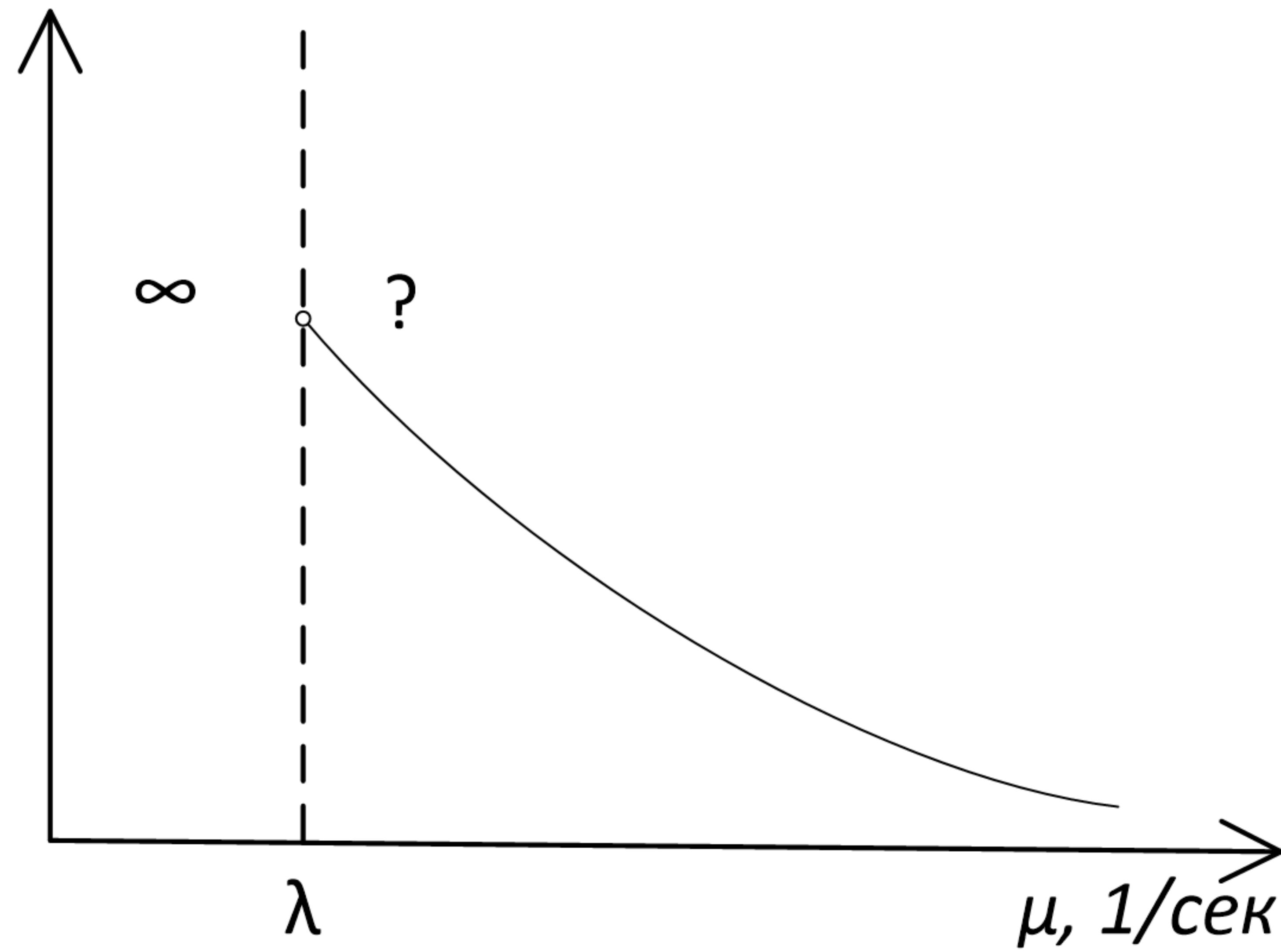
Параметры модели

- Средняя частота возникновения событий — λ ,
- Средняя пропускная способность обработчика — μ
- Загруженность обработчика — $\rho = \frac{\lambda}{\mu}$

Получаем такую картинку



Но почему не такую??

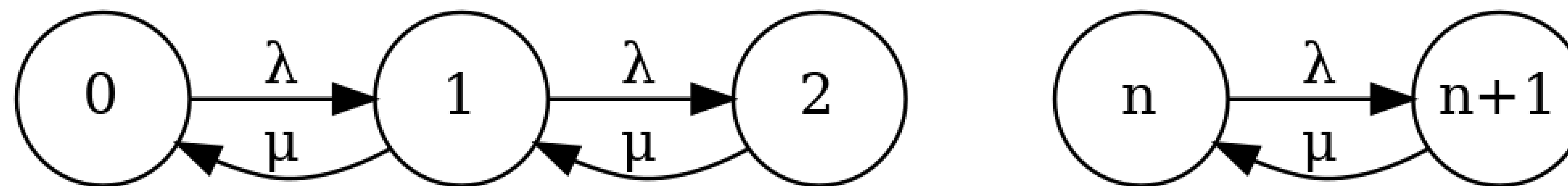


Наливаем себе пиво, начинается матан

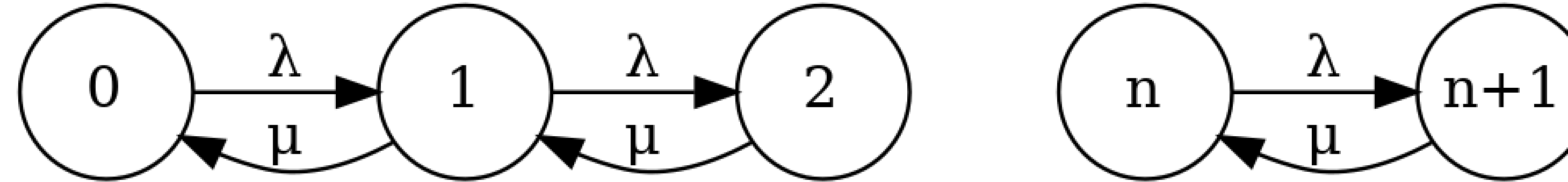


Модель (самая простая)

- События становятся в очередь в случайные моменты времени с распределением Пуассона,
- Обработчик событий затрачивает случайное время с экспоненциальным распределением,
- Система может находиться в счётном количестве состояний:



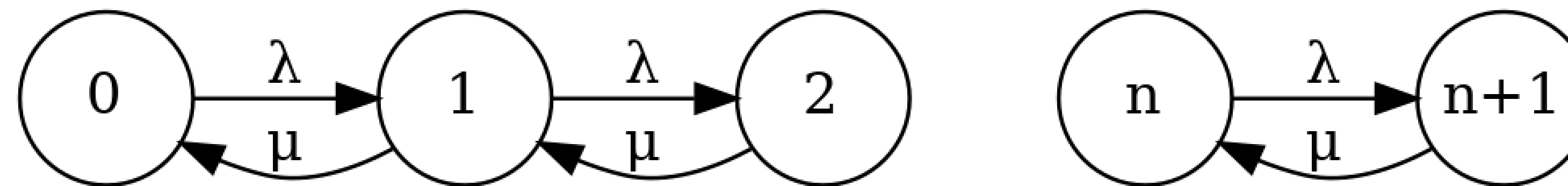
Вероятность нахождения в состоянии n в момент времени t



$$\begin{aligned} p_0(t + \Delta t) &= (1 - \lambda\Delta t)p_0(t) + \mu\Delta t p_1(t) + o(\Delta t), \\ p_n(t + \Delta t) &= \lambda\Delta t p_{n-1}(t) + (1 - (\lambda + \mu)\Delta t)p_n(t) \\ &\quad + \mu\Delta t p_{n+1}(t) + o(\Delta t). \end{aligned}$$

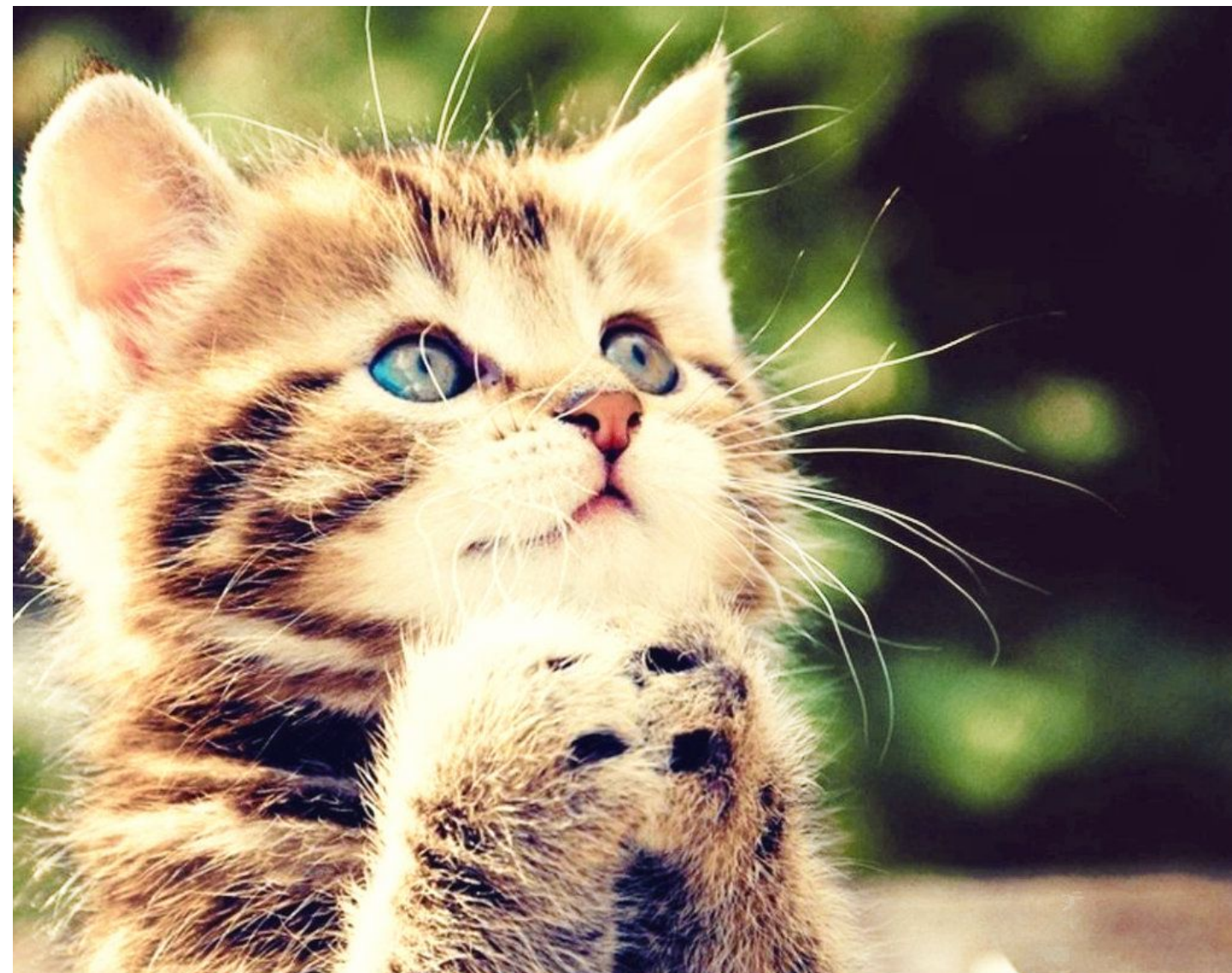


При $\Delta t \rightarrow 0$

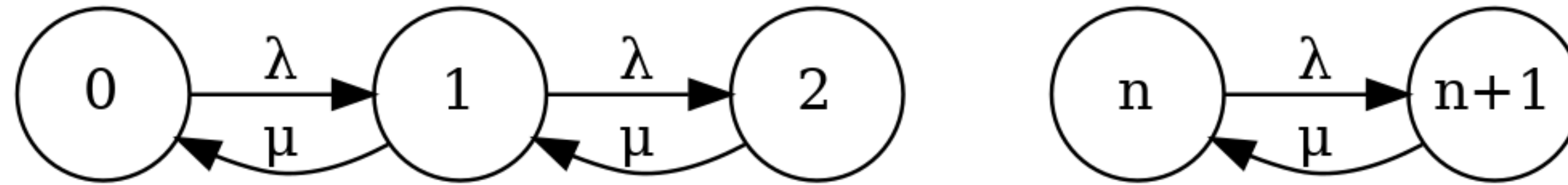


$$p'_0(t) = -\lambda p_0(t) + \mu p_1(t),$$

$$p'_n(t) = \lambda p_{n-1}(t) - (\lambda + \mu) p_n(t) + \mu p_{n+1}(t)$$



При стационарном поведении системы



$$0 = -\lambda p_0 + \mu p_1,$$

$$0 = \lambda p_{n-1} - (\lambda + \mu) p_n + \mu p_{n+1},$$

$$1 = \sum_{n=0}^{\infty} p_n$$



Решение всей системы

$$p_n = (1 - \rho)\rho^n,$$
$$n = 0, 1, 2 \dots, \quad \rho = \frac{\lambda}{\mu}$$



Средняя длина очереди?

$$\begin{aligned} E(L) &= \sum_{n=0}^{\infty} n p_n = \rho(1 - \rho) \sum_{n=0}^{\infty} n \rho^{n-1} = \\ &= \frac{\rho}{1 - \rho}. \end{aligned}$$

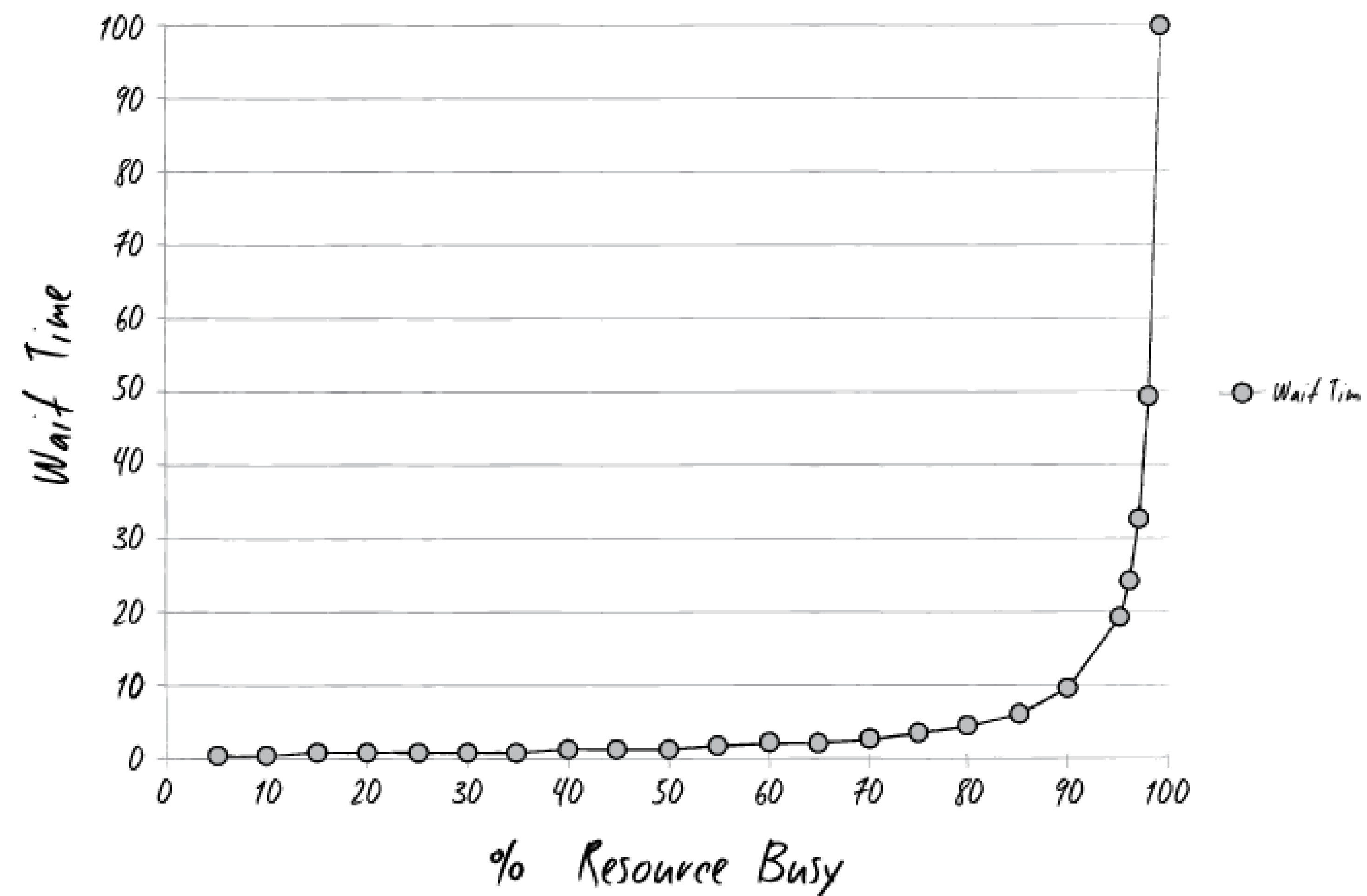


Итак...

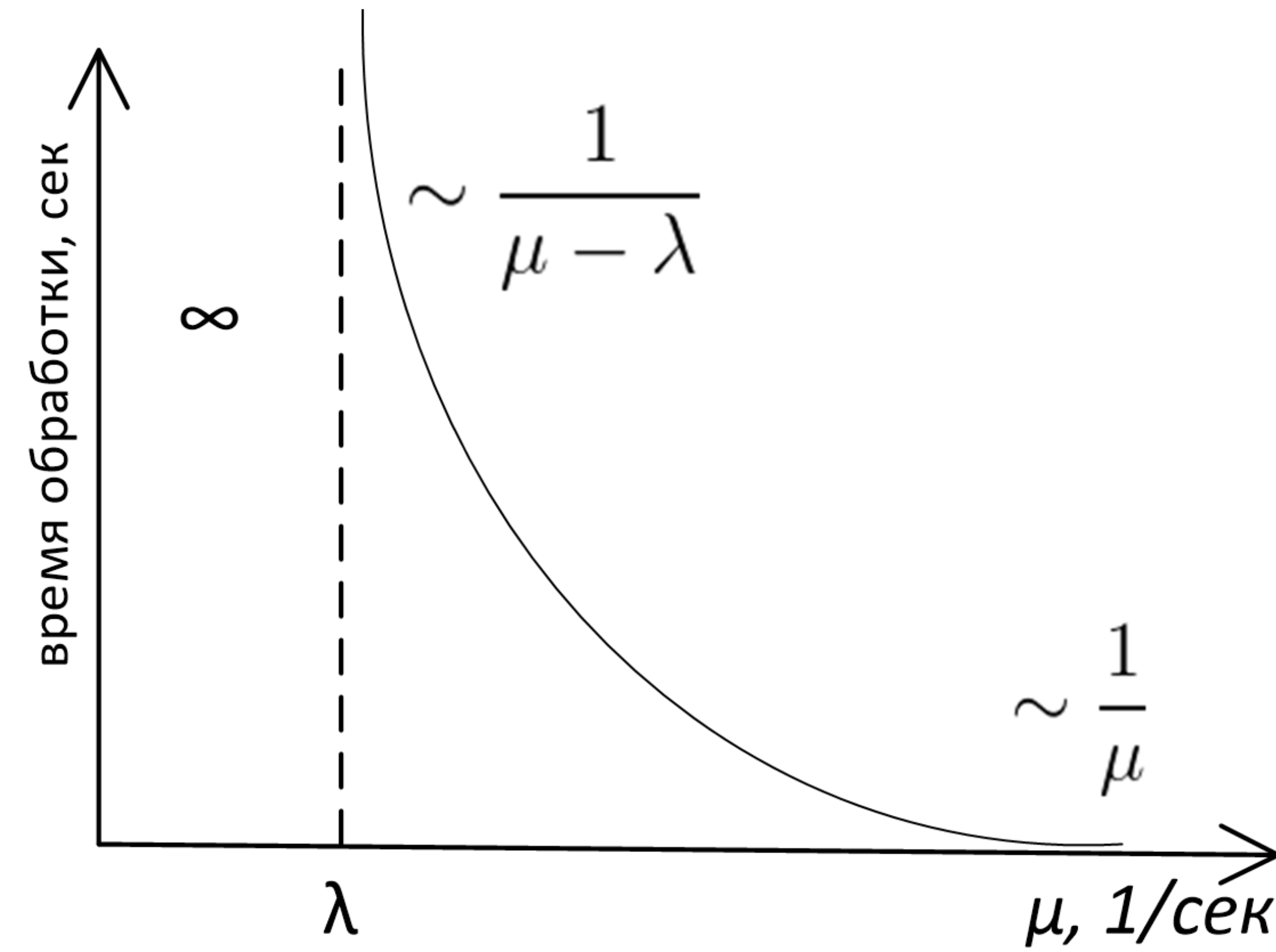
$$E(L) = \frac{\rho}{1 - \rho} .$$

Но ведь это...

$$\text{Wait Time} = (\% \text{ Busy}) / (\% \text{ Idle})$$



На пальцах это можно объяснить как-то так

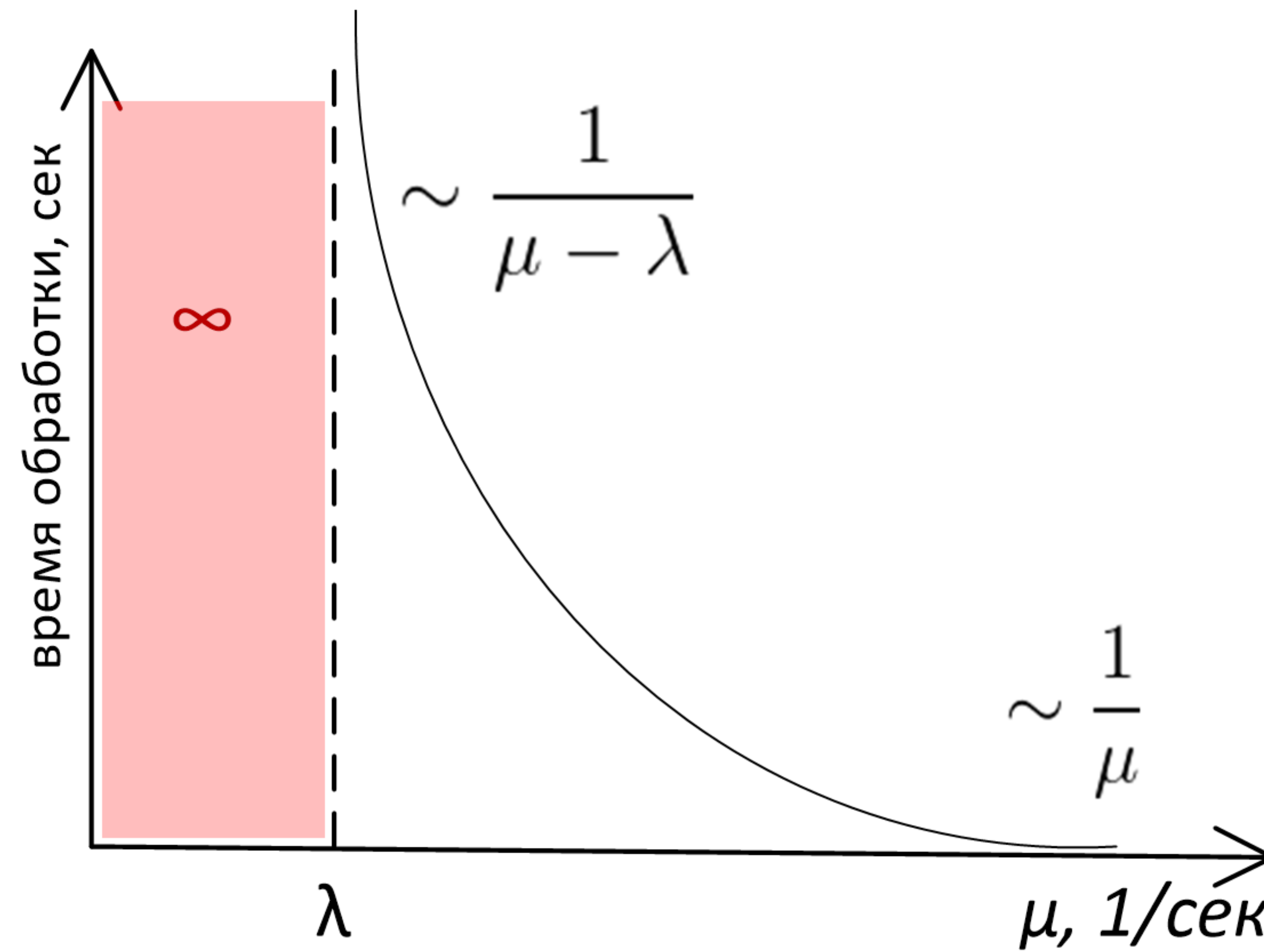


Выводы

Любая система с очередью может находиться в одном из трёх режимов

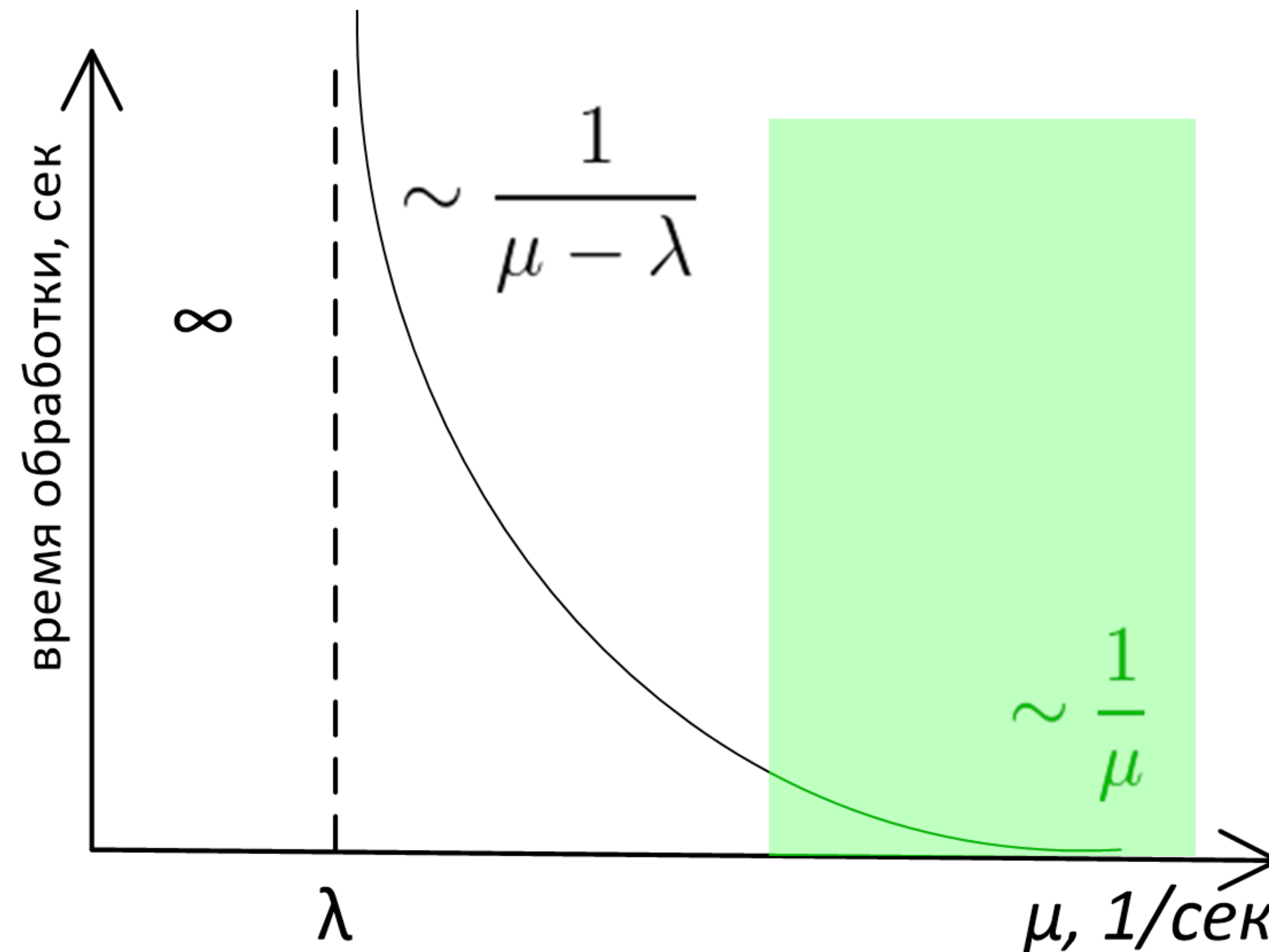
Нестабильный режим

$\mu < \lambda$ или $\mu = \lambda$ (нам конец)



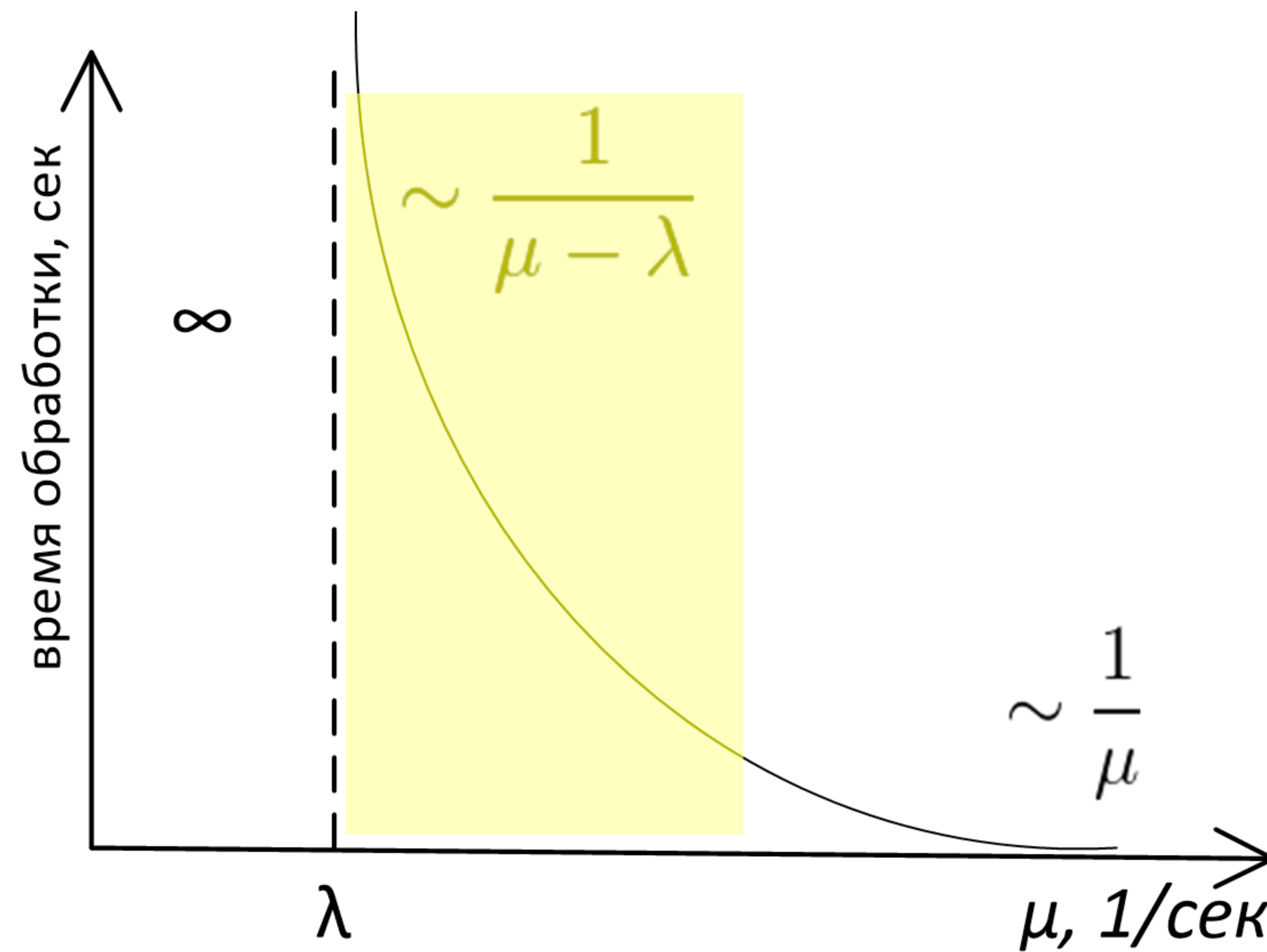
Режим с малым ожиданием в очереди (и малой загрузкой процессора)

$\mu \gg \lambda$ (очереди нет, но и процессор часто простаивает)



Возле точки насыщения

- $\mu > \lambda$, но ненамного!





Главный вывод

Оптимизация процессора и оптимизация потока — это «качели»: оптимизируя поток, мы должны лимитировать загрузку процессора.



У меня всё!

-  @inponomarev
-  ponomarev@corchestra.ru

