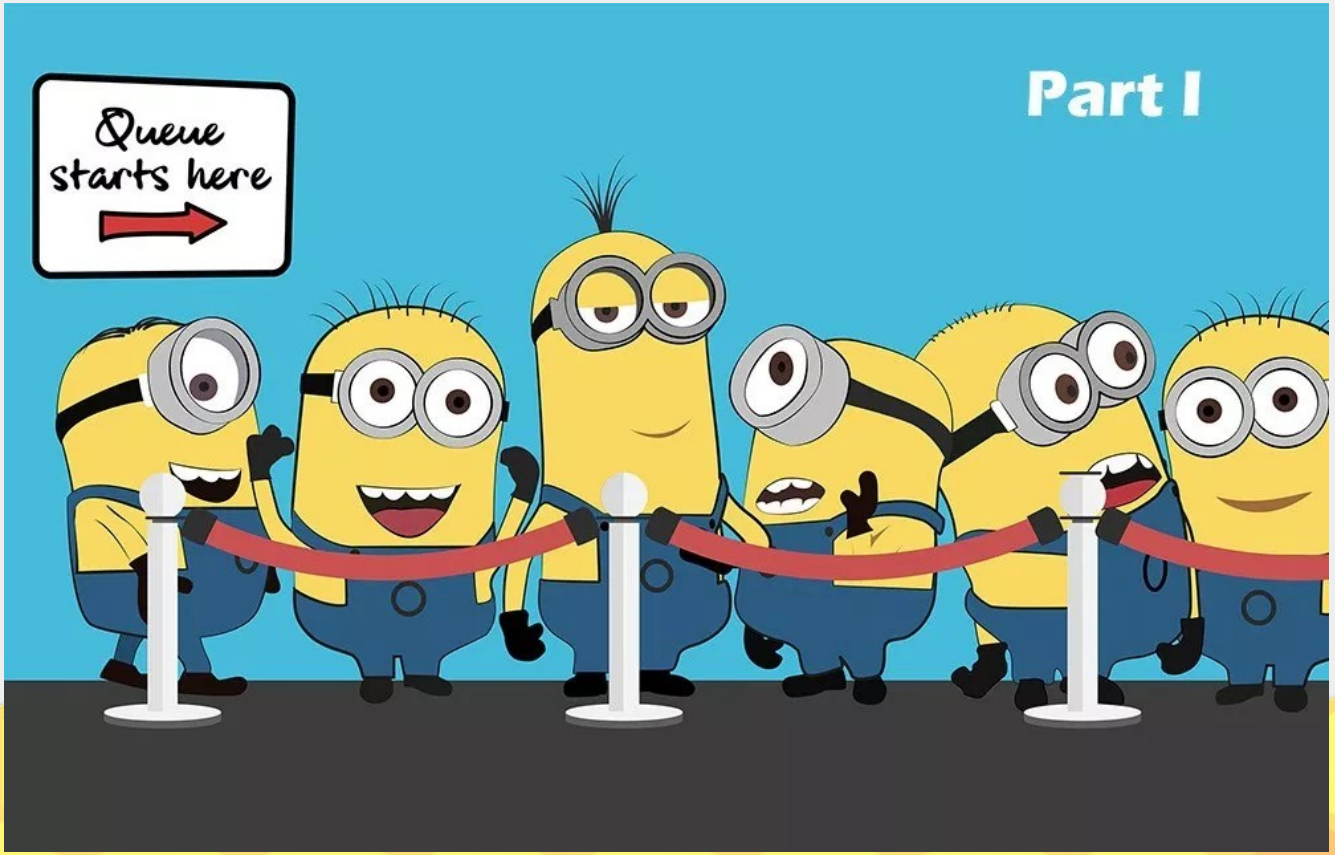
Skoltech

Skolkovo Institute of Science and Technology

Stochastic methods in Mathematical Modelling

Lecture 8. Queueing Systems.

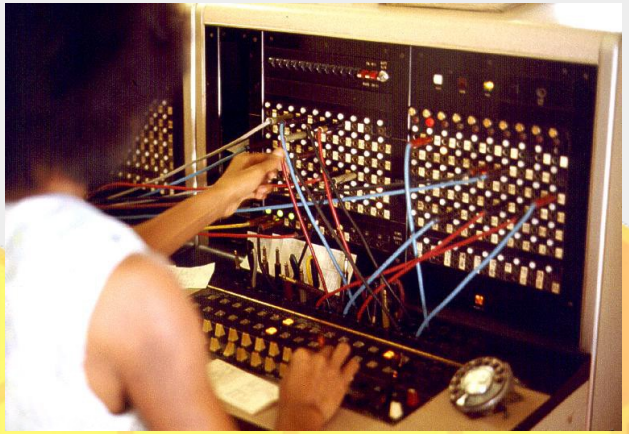


https://michaelscodingspot.com/c-job-queues/

Skoltech
Skolkovo Institute of Science and Technology

Queueing Systems

Agner Krarup Erlang published first paper in queueing theory in 1909. He modelled a number of phone calls arriving at an exchange by a Poisson process, thus, founding a queueing theory and telephone networks analysis.



Skoltech
Skolkovo Institute of Science and Technology

Queuing theory. Applications

Any kind of systems where arrivals (generally, coming units) and processing fits the framework

Call centres

Manufacturing

Supermarket check out

Airport border controls

Skoltech
Skolkovo Institute of Science and Technology

Poisson processes

Continuous version of Bernoulli process or, alternatively, random time intervals between successes

Assumptions

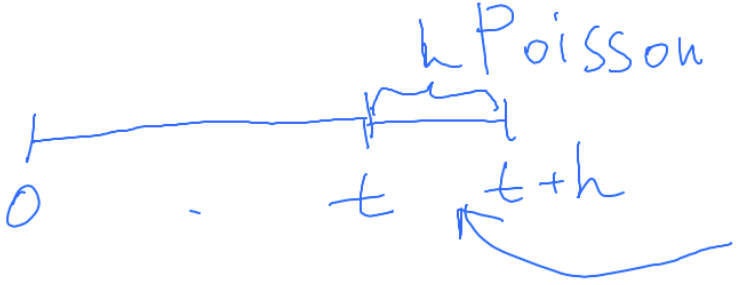
- Numbers of arrivals in disjoint intervals are independent
- For a small *h* the probability of *k* events P(*k*;*h*)

$$1-P_0(h)=\lambda h+o(h),$$
$$P_1(h)=\lambda h+o(h),$$
$$P_{k\geq 2}(h)=o(h).$$

- λ is the arrival rate of the process

For *k* successes within time interval *t*

$$P_k(t)=\frac{\lambda t}{k!}e^{-\lambda t}$$



$$\begin{cases} P_1(h) = \lambda h + o(h) \\ 1 - P_0(h) = \lambda h + o(h) \\ P_{h\geq 2}(h) = o(h) \end{cases}$$

4. $P_n(t+h)$

1. 0 successes $(t, t+h]$
   n events in $[0, t]$
2. 1 event in $[t, t+h]$
   $n-1$ events in $[0, t]$
3. $k \geq 2$ events in $[t, t+h]$
   $n-k$ events in $[0, t]$

$$P_n(t+h) = P_n(t)(1-\lambda h) + P_{n-1}(t)\lambda h + o(h) \rightarrow$$

$$\frac{P_n(t+h) - P_n(t)}{h} = -\lambda P_n + \lambda P_{n-1} + \frac{o(h)}{h}$$

$$\downarrow h \to 0$$

$$\frac{\partial P_n}{\partial t} = -\lambda P_n + \lambda P_{n-1}$$

$$\begin{cases} P_n'(t) = -\lambda P_n(t) + \lambda P_{n-1}(t), & n \geq 1 \\ P_0'(t) = -\lambda P_0(t), & n = 0 \end{cases}$$

initial condition

$$P_0(0) = 1$$

$$P_0(t) = e^{-\lambda t}$$

$$P_1(t) = \lambda t e^{-\lambda t}$$

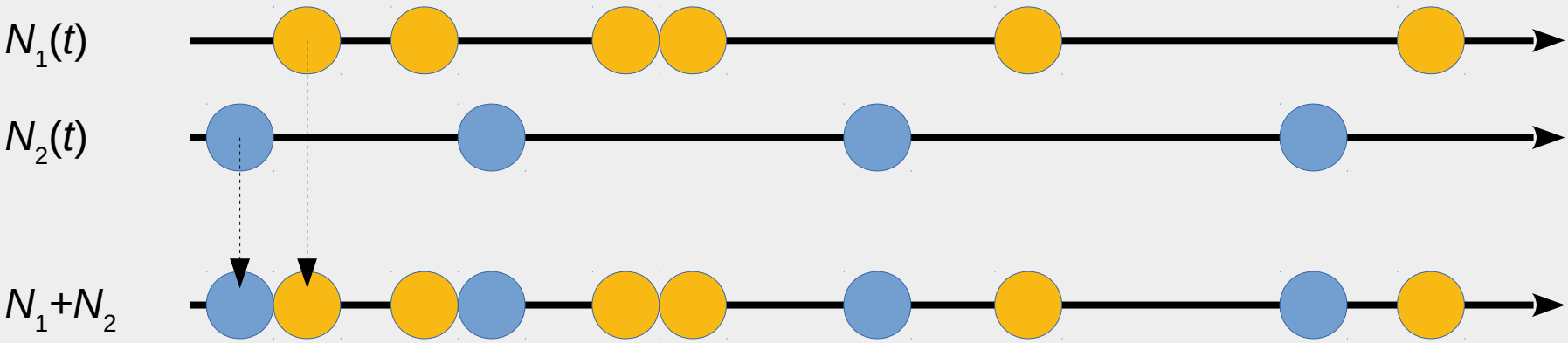
...

$$P_n(t) = \frac{(\lambda t)^n}{n!} e^{-\lambda t}$$

Skoltech
Skolkovo Institute of Science and Technology

Poisson processes. Merging

Merging: Assume that we have two independent Poisson processes $N_1(t)$ and $N_2(t)$ with rates λ_1 and λ_2. Then $N(t)=N_1(t)+N_2(t)$ is also a Poisson process with the rate $\lambda = \lambda_1 + \lambda_2$



$$N(0)=N_1(0)+N_2(0)$$

$$1-P_0(h)=\lambda_1 h+\lambda_2 h+o(h),$$

$$P_1(h)=\lambda_1 h+\lambda_2 h,$$

$$P_2(h)=o(h).$$

For k successes within time interval t

Hence $$P_k(t)=\frac{(\lambda_1+\lambda_2)t}{k!}e^{-(\lambda_1+\lambda_2)t}$$

Skoltech

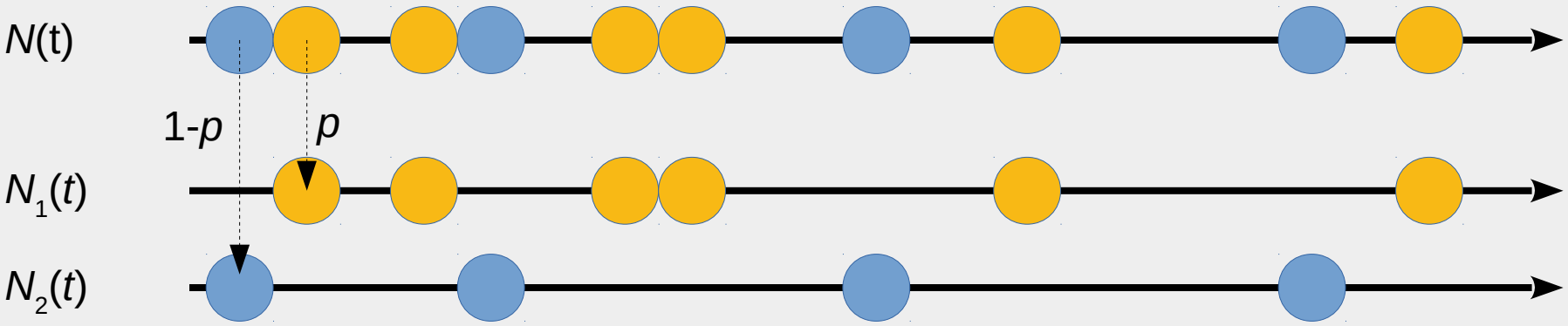
Skolkovo Institute of Science and Technology

Poisson processes. Splitting

Splitting: Let $N(t)$ will be a Poisson process with the rate λ. We split it into two Poisson processes $N_1(t)$ and $N_2(t)$ and the splitting is decided by the Bernoulli process with the probability p, i.e. with probabilities p and $1-p$ the arrival is classified as N_1 or N_2.

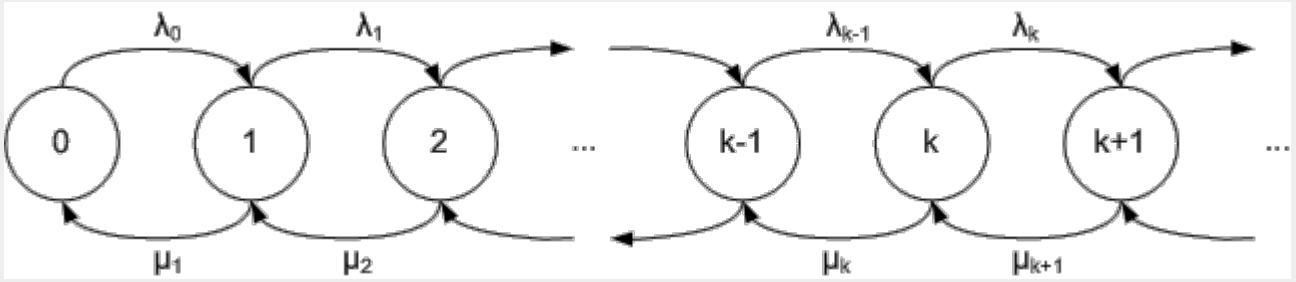
Then:

- $N_1(t)$ is a Poisson process with the rate $\lambda_1 = p\lambda$
- $N_2(t)$ is a Poisson process with the rate $\lambda_2 = (1-p)\lambda$
- $N_1(t)$ and $N_2(t)$ are independent



Skoltech
Skolkovo Institute of Science and Technology

Birth-death processes



$$\frac{dp_0(t)}{dt} = \mu_1 p_1(t) - \lambda_0 p_0(t)$$

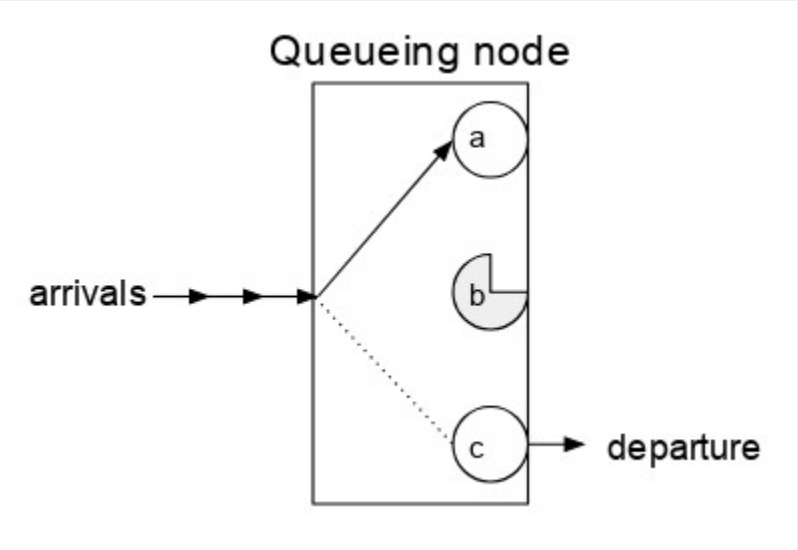
$$\frac{dp_k(t)}{dt} = \lambda_{k-1} p_{k-1}(t) + \mu_{k+1} p_{k+1}(t) - (\lambda_k + \mu_k) p_k(t), k = 1,2,\ldots$$

The graph is also a Markov chain!!

Skoltech
Skolkovo Institute of Science and Technology

Queueing Systems

Kendall's notation (extended) A/S/c/K/N/D



A denotes the time distribution between arrivals to the queue

S is the service time distribution

c is the number of service channels open at the node

K is the capacity of the queue

N is the size of the population of jobs to be served

D is the queueing discipline

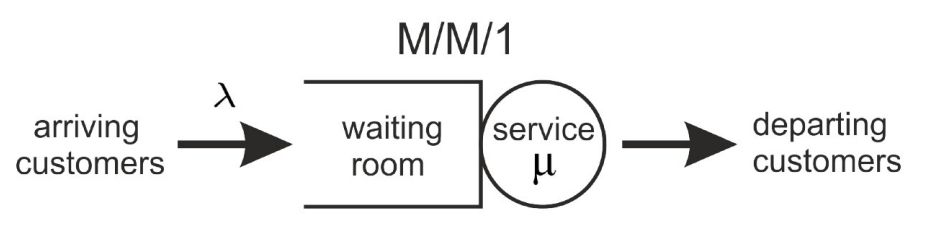
When the final three parameters are not specified (e.g. M/M/1 queue), it is assumed K = ∞, N = ∞ and D = “First in First out”

Skoltech
Skolkovo Institute of Science and Technology

Queueing Systems

Kendall's notation (examples)

M/M/1 queue: A queue with 1 server, arrival rate λ and departure rate μ.

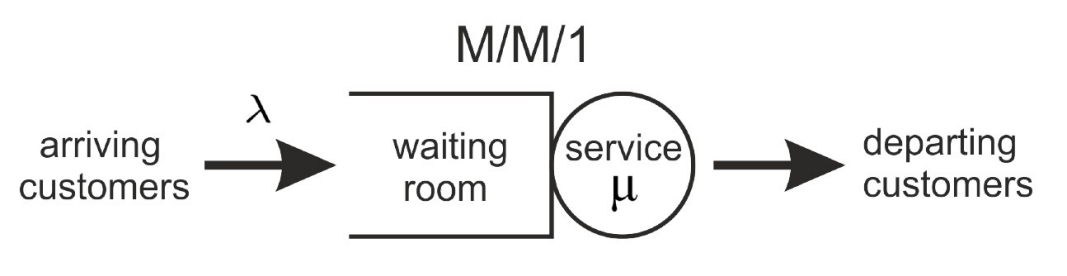


M/G/1 queue
M/D/1/∞ queue
M/D/k/∞ queue

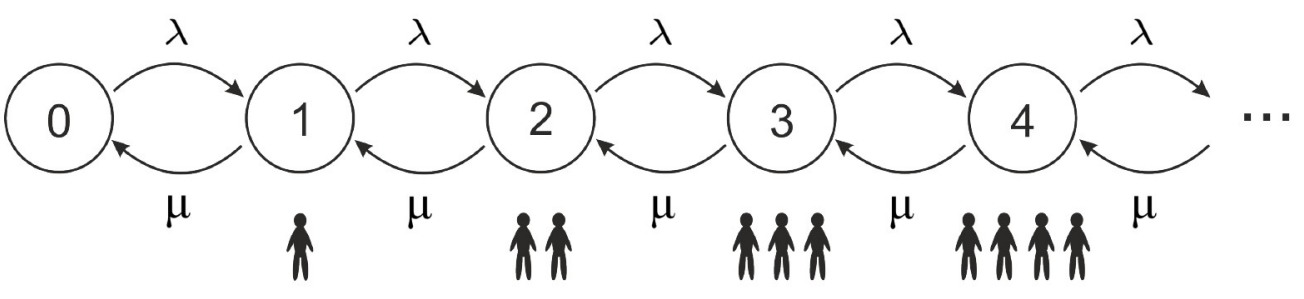
- **M** stands for *Markov* or *memoryless* and it means that arrivals occur according to a Poisson process. Arrivals may be also deterministic, **D**.
- **D** stands for *deterministic* and means that the jobs arriving at the queue require a fixed=deterministic amount of service/processing. Processing can also be stochastic, Markovian (or non-Markovian, in which case it is custom to denote it as **G** - generic service; arrival can also be **G**=generic).
- **k** describes the *number of servers* at the queueing node **k** = 1, 2, ... If there are more jobs at the node than the servers then jobs will queue and wait for service

Skoltech
Skolkovo Institute of Science and Technology

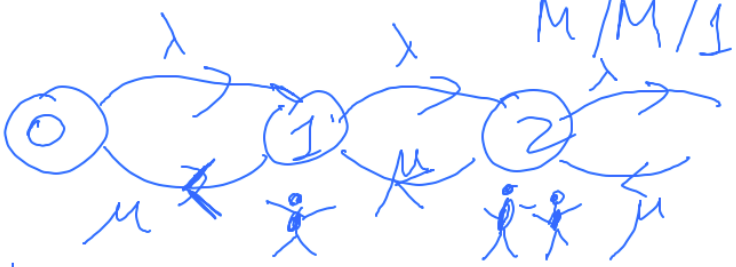
M/M/1 queue



Equivalent to the following birth-death process



M/M/1 Queue



$$\begin{cases} \dfrac{dp_0}{dt} = -\lambda p_0 + \mu p_1 \\ \dfrac{dp_n}{dt} = \lambda p_{n-1} - (\lambda+\mu) p_n + \mu p_{n+1}, \quad n \geq 1 \end{cases}$$

Steady-state analysis

$$\begin{cases} \mu p_1 = \lambda p_0 & 0 \text{ customers} \\ p_0 + \mu p_2 = (\lambda+\mu) p_1 & 1 \text{ customer} \\ \lambda p_{n-1} + \mu p_{n+1} = (\lambda+\mu) p_n & n \text{ customers} \end{cases} \Rightarrow$$

$$\frac{dp_i}{dt} = 0 \qquad \boxed{\rho \equiv \frac{\lambda}{\mu}}$$

$$p_1 = \frac{\lambda}{\mu} p_0$$

$$p_2 = \frac{1}{\mu}\left[(\lambda+\mu)\frac{\lambda}{\mu} p_0 - \lambda p_0\right] = \frac{\lambda^2}{\mu^2} p_0$$

$$\dots$$

$$p_n = \left(\frac{\lambda}{\mu}\right)^n p_0$$

$$1 = \sum_{n=0}^{\infty} p_n = p_0 \sum_{n=0}^{\infty} \rho^n = \frac{p_0}{1-\rho} \Leftarrow$$

$P_0 = 1-\rho$

$P_n = (1-\rho)\rho^n$

$\rho = \frac{\lambda}{\mu}$

$$\langle N \rangle = \sum_{n=0}^{\infty} n P_n = (1-\rho)\sum_{n=0}^{\infty} n\rho^n = \frac{\rho}{1-\rho} = \frac{\lambda}{\mu-\lambda}$$

$\lambda \to \mu$, $\langle N \rangle \to \infty$

if we start with m customers

$$P_n(t) = e^{-(\lambda+\mu)t}\left[\rho^{\frac{n-m}{2}} I_{n-m}(at) + \rho^{\frac{n-m-1}{2}} I_{n+1+m}(at) + \right.$$

$$\left. + (1-\rho)\rho^n \sum_{j=n+m+2} \rho^{-\frac{j}{2}} I_j(at)\right], \quad a = 2\sqrt{\lambda\mu}$$

I is a modified Bessel function

Busy period of the server

PDF: $f(t) = \frac{1}{t\sqrt{\rho}} e^{-(\lambda+\mu)t} I_1(2t\sqrt{\lambda\mu})$

$\langle t \rangle = \int_0^\infty t f(t)\,dt$

$\langle t \rangle$: $\mathcal{L}(f(t)) = \frac{1}{2\lambda}\left(\lambda+\mu+s - \sqrt{(\lambda+\mu+s)^2 - 4\lambda\mu}\right) = \hat{f}(s)$

Laplace transform

$\langle t \rangle, \langle t^2 \rangle, \langle t^3 \rangle$ are finite: $\mu > \lambda$

$$\hat{f}(s) = \int_0^\infty e^{-st} f(t)\,dt = \underbrace{\int_0^\infty f(t)\,dt}_{1} - s\underbrace{\int_0^\infty t f(t)\,dt}_{\langle t \rangle} + \ldots =$$

$$= 1 - \langle t \rangle s + \frac{\langle t^2 \rangle}{2} s^2 + \ldots$$

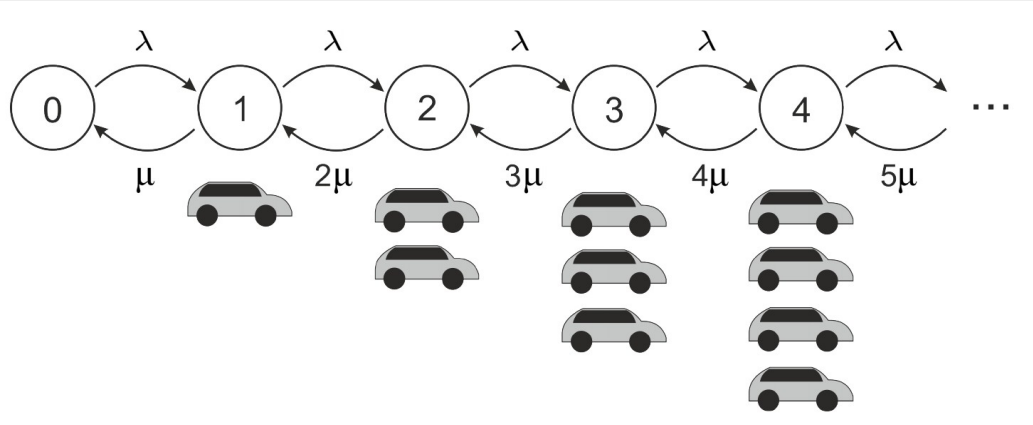
$$\hat{f}(s) = \frac{1}{2\lambda}\left(\lambda+\mu+s - \sqrt{\lambda^2+\mu^2+s^2+2\lambda\mu+2s(\lambda+\mu)-4\lambda\mu}\right) = \frac{1}{2\lambda}\left(\lambda+\mu+s-(\mu-\lambda)\sqrt{1+\frac{2s(\lambda+\mu)}{(\mu-\lambda)^2}}\right)$$

$$= 1 - \frac{s}{\mu-\lambda} \Rightarrow \boxed{\langle t \rangle = \frac{1}{\mu-\lambda}}$$

Skoltech
Skolkovo Institute of Science and Technology

M/M/∞ queue

Parking in the area of unlimited capacity



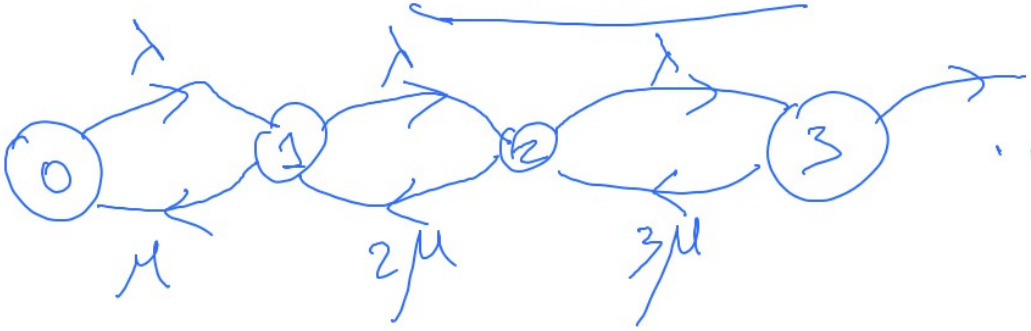
M/M/∞

Steady-state solution

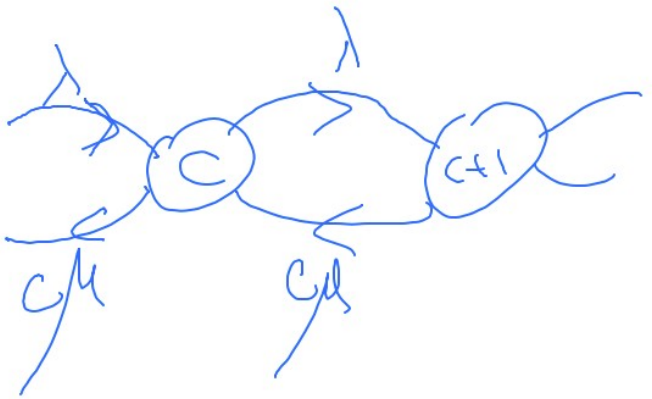
$$p_n = \frac{1}{n!}\left(\frac{\lambda}{\mu}\right)^n e^{-\lambda/\mu}$$

$$\langle N \rangle = \frac{\lambda}{\mu}$$

M/M/C

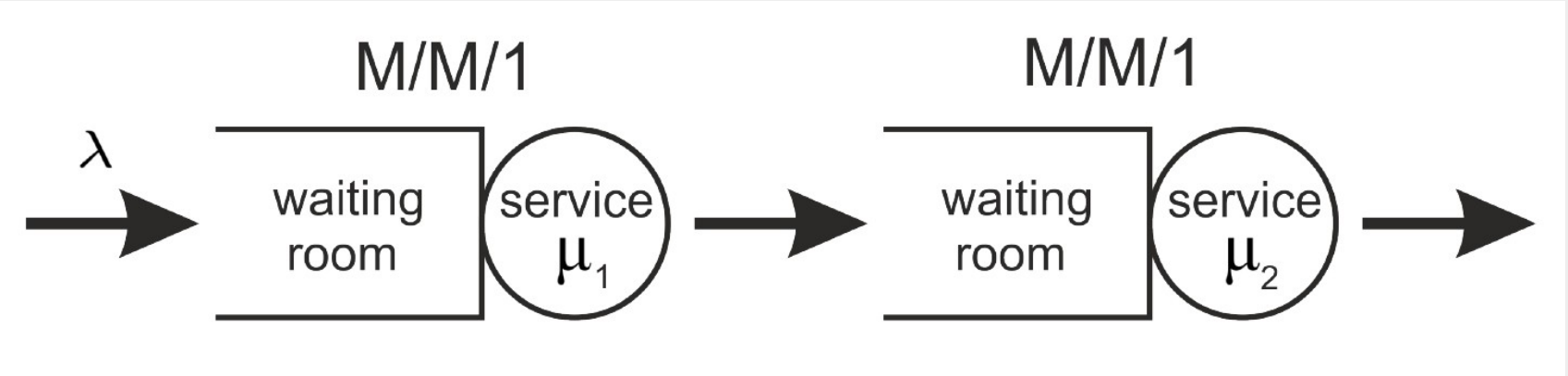


1 < C < ∞



Skoltech
Skolkovo Institute of Science and Technology

Tandem queues



Joint pdf for number of customers n_1, n_2 queueing in 1, 2

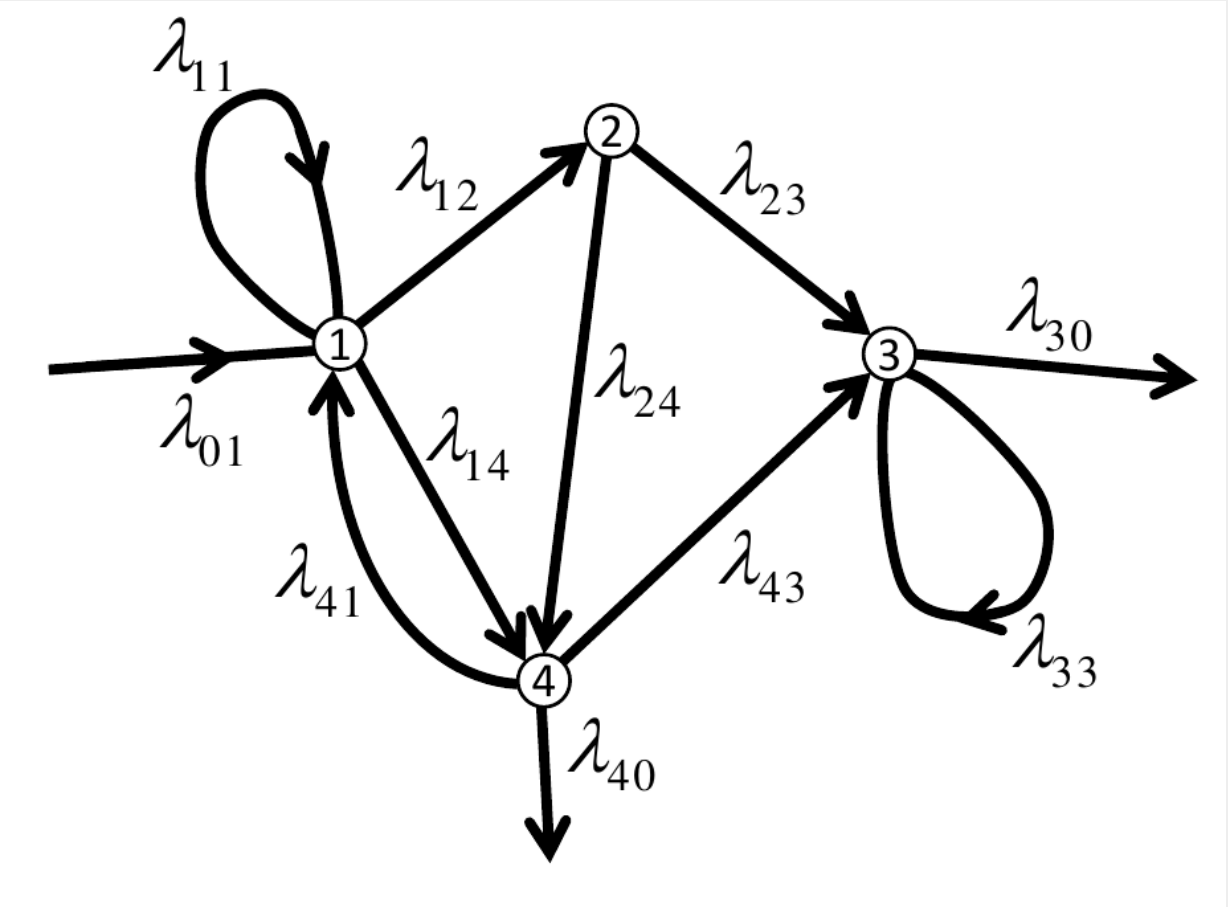
$$p(n_1, n_2) = (1-\rho_1)(1-\rho_2)\rho_1^{n_1}\rho_2^{n_2}$$

$$\rho_1 = \lambda/\mu_1, \rho_2 = \lambda/\mu_2$$

Skoltech

Skolkovo Institute of Science and Technology

Generalisation. Jackson Networks



Skoltech
Skolkovo Institute of Science and Technology

Heavy traffic limit

• The number of servers is fixed and the traffic intensity (utilization) λ/μ approaches unity (from below). The queue length approximation is the so-called “reflected Brownian motion”

• Traffic intensity is fixed and the number of servers and arrival rate are increased to infinity. Here the queue length limit converges to the normal distribution

M/M/∞ queue. Heavy traffic limit

If N_t is the number of customers at time t and $\lambda/\mu \to \infty$ then the variable

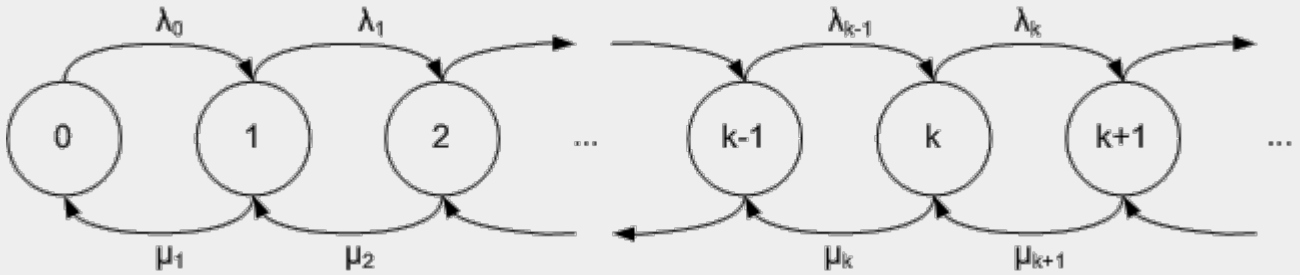
$$X_t = \frac{N_t - \lambda/\mu}{\sqrt{\lambda/\mu}}$$

Behaves according to Ornstein-Uhlenbeck process

$$dX_t = -X dt + \sqrt{2}\, dW_t \longrightarrow \text{Brownian motion}$$

Skoltech
Skolkovo Institute of Science and Technology

Exercise. M/M/c queue



1. Program a Markov chain for M/M/c queue for a general c (parameter)
2. Then study the case c=1 and approach and plot the steady state distribution $p_n=(1-\lambda/\mu)(\lambda/\mu)^n$ for $\lambda=0.2$, $\mu=0.6$
3. Perform the same for $c \to \infty$ to check $p_n=(1/n!)(\lambda/\mu)^n \exp(-\lambda/\mu)$

Skoltech

Skolkovo Institute of Science and Technology

Literature

1. F. P. Kelly. Reversibility and Stochastic Networks (Wiley, Chichester, 1979, reprinted 1987, 1994)