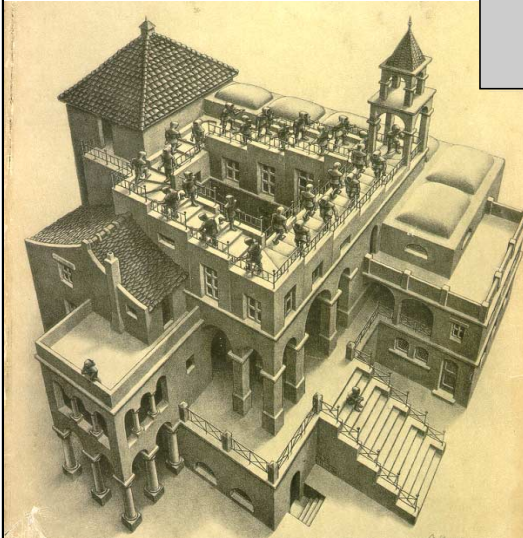


Nachrichtenverkehrstheorie Kap.2.3

- Weitere M/M/m/n - Systeme -



Universität Hannover
Institut für Allgemeine Nachrichtentechnik
Kommunikationsnetze
Prof. Dr.-Ing. K. Jobmann

Verkehrstheorie I 2.1.1

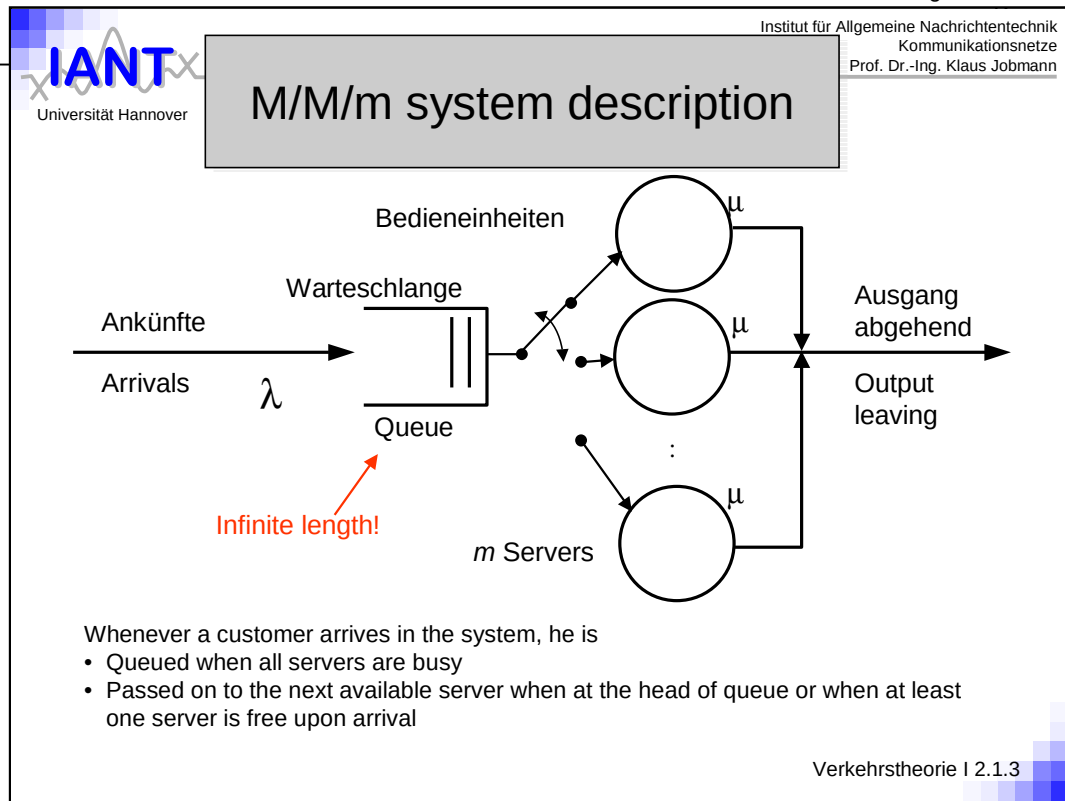
Literatur:

[schw88] Mischa Schwartz, Telecommunication Networks, Protocols, Modelling and Analysis, Addison Wesley Publishing Company, 1988, ISBN 0-201-16423-X

[Klei75] Leonard Kleinrock, Queuing Systems, Volume I, Volume II, John Wiley & Sons, New York...,1975, ISBN 0-471-49110-1

Gliederung

- M/M/1-System
- MM/1/N-System
 - ⇒ Verlust
 - ⇒ Durchsatz
- M/M/m-Systems (Waiting Systems)
- M/M/ ∞ -Systems (infinite server case)
- M/M/m/N-Systems (Loss Systems)



The system is such that a single queue with infinite length is connected to m servers. When a customer arrives in the system and at least one server is busy, he is forwarded to an arbitrary free server. If all servers are busy, the customer is queued and has to wait, until it is at the head of the queue and a server becomes free. The customer, then, is routed to this free server.

The M/M/m-system, therefore, is identical to the M/M/1-system, except that m servers are available. In particular, the number of places in queue is unlimited.

The mean arrival rate of customers into the system is denoted by λ and each server has a mean service rate of μ . Both, arrival and service processes, are Poisson.

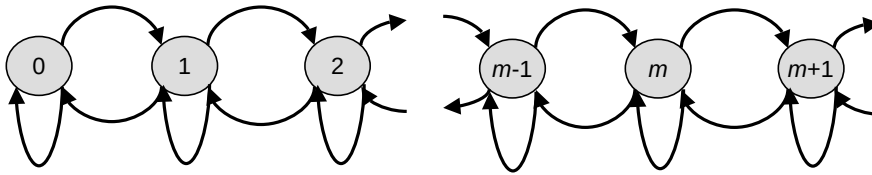
M/M/m Markov chain basic structure



Verkehrstheorie I 2.1.4

We consider again, as we did with the M/M/1-queue, a time-discrete Markov chain. The basic structure is what we search for first, i.e. which state is connected to which state. Assume again, that only a single arrival or exit from the system can happen per time slot.

M/M/m Markov chain transition probabilities

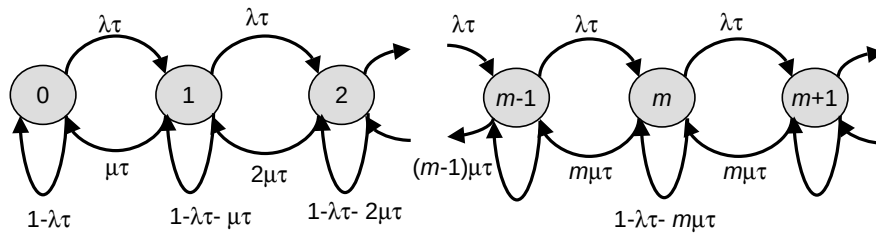


Verkehrstheorie I 2.1.5

The M/M/m Markov chain has the same structure as the M/M/1-System. However, the transition probabilities are not equal. We search for the transition probabilities now.

For this purpose, it is important to consider the number of busy servers in each state and the resulting probability that at least one of the servers finishes its work during the time slot period τ .

The M/M/m Markov chain



$$\begin{aligned} \lambda p_{n-1} &= n\mu p_n & n \leq m \\ \lambda p_{n-1} &= m\mu p_n & n > m \end{aligned} \Rightarrow p_n = \begin{cases} p_0 \frac{(m\rho)^n}{n!}; & n \leq m \\ p_0 \frac{m^m \rho^n}{m!}; & n > m \end{cases}$$

$$\text{where } \rho = \frac{\lambda}{m\mu} < 1$$

Verkehrstheorie I 2.1.6

Now that we have the Markov chain, we need to search for the detailed balance equations. The solution can be found by iteration, as in the case of the M/M/1-System:

for $n \leq m$:

$$\lambda p_{n-1} = n\mu p_n$$

$$p_n = \frac{\lambda}{n\mu} p_{n-1} = \frac{\lambda}{n\mu} \left(\frac{\lambda}{(n-1)\mu} p_{n-2} \right) = \dots = \frac{\lambda^n}{n! \mu^n} p_0$$

$$\text{with } \rho = \frac{\lambda}{m\mu} \Rightarrow \rho^n = \frac{\lambda^n}{m^n \mu^n} \Rightarrow \frac{\lambda^n}{\mu^n} = (m\rho)^n$$

$$\text{we get } p_n = \frac{(m\rho)^n}{n!} p_0$$

The part for $n > m$ is analogous

Solution for M/M/m system Determination of p_0

$$\begin{aligned}
 \sum_{n=0}^{\infty} p_n = 1 &= \sum_{n=0}^{m-1} \frac{(m\rho)^n}{n!} p_0 + \sum_{n=m}^{\infty} \frac{m^m \rho^n}{m!} p_0 = \\
 &\sum_{n=0}^{m-1} \frac{(m\rho)^n}{n!} p_0 + \sum_{n=m}^{\infty} \frac{(m\rho)^n}{m!} \frac{1}{m^{n-m}} p_0 \\
 \Leftrightarrow p_0 &= \left[\sum_{n=0}^{m-1} \frac{(m\rho)^n}{n!} + \sum_{n=m}^{\infty} \frac{(m\rho)^n}{m!} \frac{1}{m^{n-m}} \right]^{-1} \\
 &= \left[\sum_{n=0}^{m-1} \frac{(m\rho)^n}{n!} + \frac{(m\rho)^m}{m!(1-\rho)} \right]^{-1}
 \end{aligned}$$

Verkehrstheorie I 2.1.7

Some explanations for the last step in the slide above:

$$\sum_{n=m}^{\infty} \frac{(m\rho)^n}{m!} \frac{1}{m^{n-m}} = \sum_{n=m}^{\infty} \left(\frac{m\rho}{m} \right)^n \frac{1}{m! m^{-m}} = \frac{m^m}{m!} \sum_{n=m}^{\infty} \rho^n$$

The sum can be solved to:

$$\sum_{n=m}^{\infty} \rho^n = \sum_{n=0}^{\infty} \rho^n - \sum_{n=0}^{m-1} \rho^n = \frac{1}{1-\rho} - \frac{\rho^m - 1}{\rho - 1} = \frac{\rho^m}{1-\rho}$$

And the final result is:

$$\sum_{n=m}^{\infty} \frac{(m\rho)^n}{m!} \frac{1}{m^{n-m}} = \frac{(m\rho)^m}{m!(1-\rho)}$$

Note that the calculation of p_0 is necessary to compute the probabilities for all other states of the system. This computation requires somewhat more effort compared to an M/M/1-system. However, with given parameters, it can be done easily with a small computer program. All subsequent computations are straight forward and simple.

M/M/m delay situation

What is critical is the probability to find all servers busy, i.e. the case when an arriving customer has to wait in queue:

$$P\{\text{Queueing}\} = P_Q = \sum_{n=m}^{\infty} p_n = \sum_{n=m}^{\infty} \frac{p_0 m^m \rho^n}{m!} = \frac{p_0 (m\rho)^m}{m!} \sum_{n=m}^{\infty} \rho^{n-m}$$

$$P_Q = \frac{p_0 (m\rho)^m}{m!(1-\rho)}$$

This is known as the **Erlang-C formula (or Erlang's delay formula)**.

- Often used in telephone systems (call center example, manual exchange office)
- Used to estimate the probability to find e.g. all operators in a call center busy.
- The customer remains in queue, i.e. it continuously tries to get a free server.

Verkehrstheorie I 2.1.8

The danish mathematician A.K. Erlang was the foremost pioneer of queueing theory. The Erlang delay formula above is named after him to honour his merits in queueing theory.

Average numbers of customers in queue

The average number of customers in queue is given by

$$\begin{aligned} N_Q &= \sum_{n=m}^{\infty} (n-m) p_n = \sum_{i=0}^{\infty} i p_{m+i} = \\ &= \sum_{i=0}^{\infty} i p_0 \frac{m^m \rho^{m+i}}{m!} = p_0 \frac{(m\rho)^m}{m!} \sum_{i=0}^{\infty} i \rho^i = \\ &= p_0 \frac{(m\rho)^m}{m!} \frac{\rho}{(1-\rho)^2} \end{aligned}$$

With P_Q from the Erlang-C formula, we get

$$P_Q = \frac{p_0 (m\rho)^m}{m! (1-\rho)} \Rightarrow N_Q = P_Q \frac{\rho}{1-\rho}$$

Verkehrstheorie I 2.1.9

$i = n - m$

Comparison of M/M/m and M/M/1 systems

Note that

$$\frac{N_Q}{P_Q} = \frac{\rho}{1-\rho}$$

represents the expected number found in queue by an arriving customer conditioned on the fact that all servers are busy. It is independent of the number of servers for a given $\rho = \lambda/(m\mu)$!

Hence, as long as customers are waiting in queue, the queue size of the M/M/m systems behaves identically as in an M/M/1-system with service rate $m\mu$!

Average delay

The average waiting time in queue is given by Little's Law:

$$W = \frac{N_Q}{\lambda} = \frac{\rho \cdot P_Q}{\lambda(1-\rho)}$$

The average time in system per customer is, therefore:

$$T = \frac{1}{\mu} + W = \frac{1}{\mu} + \frac{\rho \cdot P_Q}{\lambda(1-\rho)} = \frac{1}{\mu} + \frac{P_Q}{m\mu - \lambda} \text{ with } \rho = \frac{\lambda}{m\mu}$$

Using Little's law again, the average number of customers in the system is:

$$N = \lambda T = \frac{\lambda}{\mu} + \frac{\lambda P_Q}{m\mu - \lambda} = m\rho + \frac{\rho \cdot P_Q}{1-\rho} \text{ with } \rho = \frac{\lambda}{m\mu}$$

In an M/M/1-system, we have:

$$N = \frac{\rho}{1-\rho}$$

Verkehrstheorie I 2.1.11

The comparison of N in the M/M/ m and the M/M/1 system shows again that the queue behaviour of the M/M/ m system is equal to that of the M/M/1 system, under the assumption that all servers are busy.

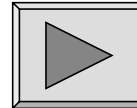
Example

Assume a communication link serving m Poisson streams with overall rate λ .

- In the first case, we have an M/M/1 system with a server with service rate $m\mu$
- In the second case, we have an M/M/m system with each server having a rate of μ

Compare the average time T in the system for both cases for

- a) Low load in the system
- b) High load in the system



Gliederung

- M/M/1-System
- MM/1/N-System
 - ⇒ Verlust
 - ⇒ Durchsatz
- M/M/m-Systems (Waiting Systems)
- **M/M/∞-Systems (infinite server case)**
- M/M/m/N-Systems (Loss Systems)

M/M/ ∞ systems The infinite server case

In the case where the number of servers is infinite, we obtain from the detailed balance equations:

$$\lambda \cdot p_{n-1} = \mu \cdot p_n, \quad n = 1, 2, \dots$$

$$\text{hence } p_n = p_0 \left(\frac{\lambda}{\mu} \right)^n \frac{1}{n!}$$

$$\text{with } \sum_{n=0}^{\infty} p_n = 1 \text{ we get } p_0 = \left[\sum_{n=0}^{\infty} \frac{\left(\frac{\lambda}{\mu} \right)^n}{n!} \right]^{-1} = e^{-\lambda/\mu} = e^{-\rho}$$

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

This means that in steady state, the number of customers in the system is Poisson distributed with parameter λ/μ . By Little's law, the average time in the system is

$$T = \frac{1}{\mu}$$

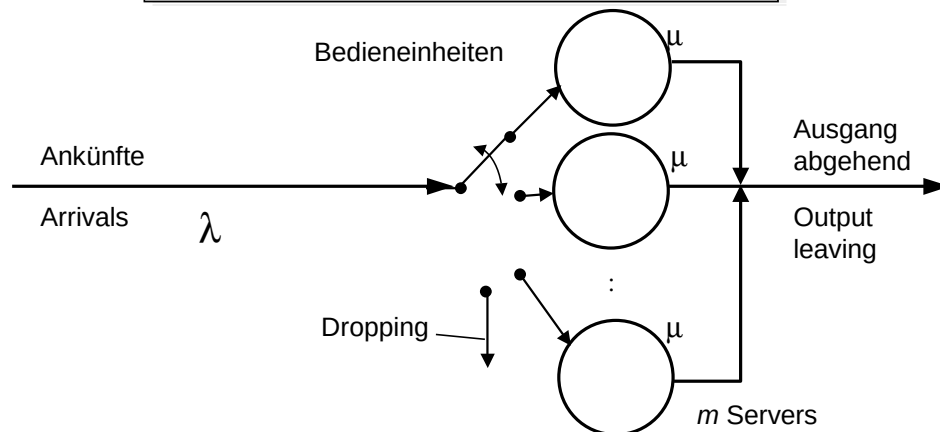
Verkehrstheorie I 2.1.14

The average time in the system is immediately obvious, since there is never waiting in queue.

Gliederung

- M/M/1-System
- MM/1/N-System
 - ⇒ Verlust
 - ⇒ Durchsatz
- M/M/m-Systems (Waiting Systems)
- M/M/ ∞ -Systems (infinite server case)
- M/M/m/m-Systems (Loss Systems)

Definition of the M/M/m/m system

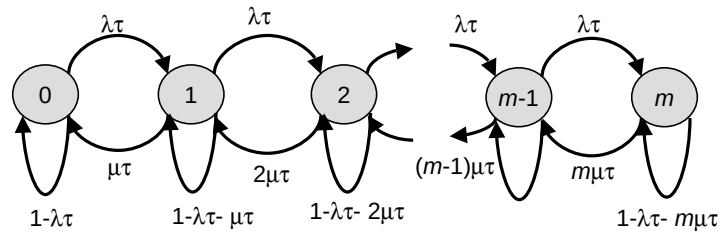


The last letter does not, as usual, denote the length of the queue but rather describes the maximum number of customers in the system. A customer is routed to a free server, if one is available. Otherwise, the customer is not allowed to enter the system but is simply dropped.

Example: Telephone system.

Verkehrstheorie I 2.1.16

M/M/m/m system Markov chain formulation



Detailed balance equation:

$$\lambda p_{n-1} = n\mu p_n \quad n \leq m \quad \Rightarrow \quad p_n = p_0 \frac{\left(\frac{\lambda}{\mu}\right)^n}{n!}$$

Solving for $\sum_{n=0}^m p_n = 1$, we get:

$$p_0 = \left[\sum_{n=0}^m \left(\frac{\lambda}{\mu}\right)^n \frac{1}{n!} \right]^{-1}$$

Verkehrstheorie I 2.1.17

Blocking probability

The blocking probability B is given by the case that all m servers are busy, i.e. by p_m :

$$p_m = \frac{\frac{(\lambda / \mu)^m}{m!}}{\sum_{n=0}^m \frac{(\lambda / \mu)^n}{n!}} = B$$

This formula is also known as Erlang-B formula. It finds wide use in evaluating the blocking probability in voice telephony systems.

Note that the loss formula can be shown to hold even if the service time has mean $1/\mu$ but arbitrary probability distribution.

Offered load in Erlangs [Erl]

The offered load A is the expected value of the number of busy lines b :

$$A = \sum_{n=0}^{\infty} n \cdot P\{b = n\} = \sum_{n=0}^{\infty} n \cdot p_n \quad \begin{array}{l} b - \text{number of busy lines} \\ p_n - \text{probability that } n \text{ lines are busy} \end{array}$$

With a stationary stochastic process, A can be calculated as a time average value.
The model is:

- Observation of the system at equidistant times with interval Δt
- Summing up all numbers of busy lines at all observation instants

$$A = \frac{1}{N \cdot \Delta t} \cdot \sum_{i=0}^N \Delta t \cdot n(t_0 + i \cdot \Delta t) = \frac{1}{N} \cdot \sum_{i=0}^N n(t_0 + i \cdot \Delta t)$$

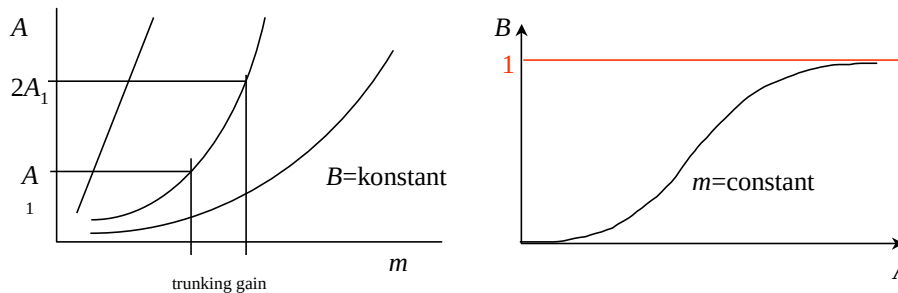
$n(t)$ - number of busy lines at time t

N - number of observation instants

Note that $A = \lambda / \mu$

Trunking gain

$$B = \frac{A^m}{m!} \sum_{n=0}^m \frac{A^n}{n!}$$



Verkehrstheorie I 2.1.20

The trunking gain describes the fact that, as the number m of servers grows, the offered load that can be carried at constant loss probability B increases more than proportionally.

Example (from table):

With 2 servers, i.e. $m = 2$, and $B = 1\%$, the offered load that can be carried is 0.153, i.e. $\rho = 0.153$ (7% of the available server capacity)

With 10 servers, i.e. $m = 10$, and $B = 1\%$, the offered load that can be carried is 4.46, i.e. $\rho = 4.46$ (on average, 4.46 customers can wish to enter the system at the same time. This corresponds to almost 50% of the available server capacity)

Example

- a) Low load in the system
- b) High load in the system

M/M/1 system:
$$T_1 = \frac{1}{m\mu} + \frac{P_{Q,1}}{m\mu - \lambda}$$

M/M/m system:
$$T_m = \frac{1}{\mu} + \frac{P_{Q,m}}{m\mu - \lambda}$$

a) P_Q very small in both cases, hence $\frac{T_m}{T_1} \cong m$, i.e. m times larger av. delay in M/M/m system

b) P_Q and ρ close to 1 in both cases, hence $\frac{1}{m\mu} \ll \frac{1}{m\mu - \lambda}$ and $\frac{1}{\mu} \ll \frac{1}{m\mu - \lambda}$
and therefore $\frac{T_m}{T_1} \cong 1$, i.e. the delays are almost equal in heavy load situations

(why is this so?)

