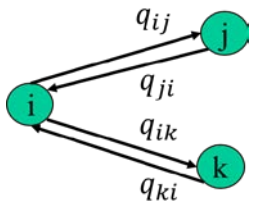
9 Continuous-Time Markov Chains (CTMCs)

See: P. Tran-Gia: "Einführung in die Leistungsbewertung und Verkehrstheorie", 2. Auflage, 2005, Oldenbourg

9.1 Basics about CTMCs

9.1.1 Definition

- Stochastic process with transitions from state i to another state j after exponentially distributed time $\left(A_{ij}(t) = 1 - e^{-q_{ij} \cdot t}\right)$
  - Process has memoryless property at any time instant.
  - "Transitions happen at rate q_{ij}"

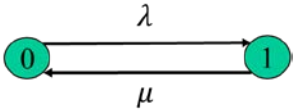


9.1.2 Simulation

Input: start state X, start time t
While(true)
$A_{min} = \infty; \quad next = -1$
For all $q_{XY} > 0$:
$A = \frac{-ln(U)}{q_{XY}}$
If $A < A_{min}$:
$A_{min} = A; next = Y$
$X = next; t += A_{min}$
TWH.count(X,t) // time-weighting histogram with 1 bin per state

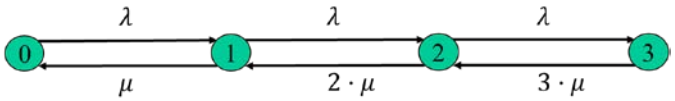
9.1.3 Application Example

- Single server
- Exponentially distributed inter-arrival time $A(t) = 1 - e^{-\lambda \cdot t}$
- Exponentially distributed service time $B(t) = 1 - e^{-\mu \cdot t}$
- Two states
  - 0: server idle
  - 1: server busy
- State transition diagram
  - Transitions marked with the corresponding exponential rates



9.1.4 Modelling „Aggregate“ Rates

- Consider M/M/3 loss model
  - 3 servers
  - State indicates number of occupied service units
  - State transition diagram:



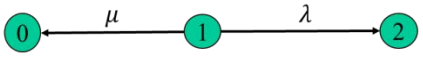
  - Transition with rate $k \cdot \mu$ from k to $k-1$ occupied service units – to be shown in the following
- k occupied service units, each with exponentially distributed service time $B^i_{single} = 1 - e^{-\mu \cdot t}$
  - What is the distribution of time B until a first service unit completes its service?
  - Answer:
    $1 - B(t) = P\left(\min_{0 \le i < k}\left(B^i_{single}\right) > t\right) = \prod_{0 \le i < k} P\left(B^i_{single} > t\right) =$
    $\prod_{0 \le i < k} e^{-\mu \cdot t} = e^{-k \cdot \mu \cdot t}$, i.e., $B(t) = 1 - e^{-k \cdot \mu \cdot t}$

9.1.5 Distribution of sojourn time S_i in State i

- $1 - S_i(t) = P\left(\min_{j \neq i}(A_{ij}) > t\right) = P(A_{ij} > t, j \neq i) = \prod_{j \neq i} P(A_{ij} > t) =$
  $\prod_{j \neq i} e^{-q_{ij} \cdot t} = e^{-(\sum_{j \neq i} q_{ij}) \cdot t}$ (13.2)
- Define rate to leave state i is $q_i = \sum_{j \neq i} q_{ij}$ (13.3)
  so that we get $1 - S_i(t) = e^{-q_i \cdot t}$

9.1.6 Transition Probabilities

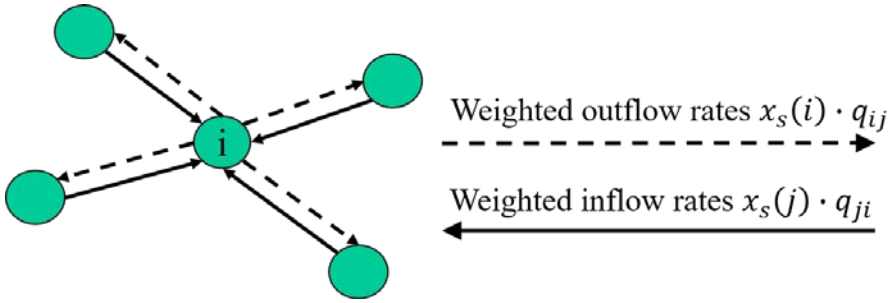
- Problem: process goes from state 1 to state 2 after A time (arrival event, $A(t) = 1 - e^{-\lambda \cdot t}$) or to state 0 after B time (service event, $B(t) = 1 - e^{-\mu \cdot t}$)



- What is the probability that the next state is 2 or 0, respectively?
- Solution
  - Next state is 2 with probability $\int_0^\infty a(t) \cdot P(B > t)dt = \int_0^\infty \lambda \cdot e^{-\lambda \cdot t} \cdot e^{-\mu \cdot t}dt = \frac{-\lambda}{\lambda+\mu} \cdot \int_0^\infty -(\lambda+\mu) \cdot e^{-(\lambda+\mu) \cdot t}dt = \frac{-\lambda}{\lambda+\mu} \cdot \left[e^{-(\lambda+\mu) \cdot t}\right]_0^\infty = \frac{\lambda}{\lambda+\mu}$
  - Next state is 0 with probability $\frac{\mu}{\lambda+\mu}$
- Generally
  - Process in state i
  - Next state is j with probability $\frac{q_{ij}}{\sum_{k \neq i} q_{ik}}$ (13.4)

9.2 Analysis of CTMCs

- Given
  - n+1 states, numbered $0 \leq i \leq n$
  - State transition rates q_{ij}
- Wanted: average state distribution
- Hint: stationary state distribution exists and equals average state distribution
- Property of the stationary state distribution $x_s = (x_s(0), \dots, x_s(n))$:
  - It's a distribution: $\sum_i x_s(i) = 1$ (13.5)
  - For any state i holds: $x_s(i) \cdot \sum_{j \neq i} q_{ij} = x_s(i) \cdot q_i = \sum_{j \neq i} x_s(j) \cdot q_{ji}$ (13.6)
    - Illustration



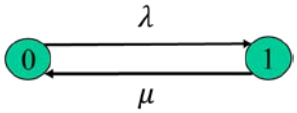
    - Background: **for any stationary state distribution x_s holds: sum of weighted outflow rates equals sum of weighted inflow rates**
    - $n + 2$ equations for $n + 1$ free variables $x_s(i), 0 \leq i \leq n$, but equations are linearly dependent and one of them can be eliminated

- Find the stationary state distribution x_s by solving the equation system above.
- How to express the Conditions (13.5) and (13.6) as a vector-matrix operation?
  - Reformulation of (13.6): $\sum_{j \neq i} q_{ji} \cdot x_s(j) - q_i \cdot x_s(i) = 0$
  - Define $q_{ii} = -q_i = -\sum_{j \neq i} q_{ij}$ (13.7)
    - q_{ii} is similar to rate q_i for leaving state i, but it is negative!
  - Define the **transition rate matrix** $Q = \begin{pmatrix} q_{00} & \cdots & q_{0,n} \\ \vdots & \ddots & \vdots \\ q_{n,0} & \cdots & q_{n,n} \end{pmatrix}$ (13.8)
    - Note: the sum of the entries in a line is 0.
    - The diagonal has negative entries.
    - Don't confuse with state transition probability matrix P where sum of entries in a line is 1!
  - Then the stationary state distribution x_s is defined by
    - $x_s \cdot \begin{pmatrix} 1 \\ \vdots \\ 1 \end{pmatrix} = 1$ according to (13.5) (13.9)
    - $x_s \cdot Q = (0 \quad \cdots \quad 0)$ according to (13.6) (13.10)

9.3 Comparative Calculation of Stationary State Distributions

9.3.1 CTMC

- State transition diagram

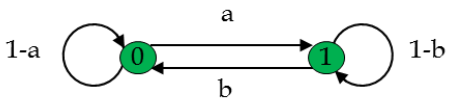


- Rate matrix $Q = \begin{pmatrix} -\lambda & \lambda \\ \mu & -\mu \end{pmatrix}$
- Conditions for stationary state distribution
  - (I) $x(0) + x(1) = 1 \rightarrow x(0) = 1 - x(1)$
  - (II) $x \cdot Q = (0, \cdots, 0)$
    - (IIa) $-\lambda \cdot x(0) + \mu \cdot x(1) = 0$
- Solve the equation system
  (I) in (IIa)
  $-\lambda \cdot [1 - x(1)] + \mu \cdot x(1) = 0$
  $-\lambda + \lambda \cdot x(1) + \mu \cdot x(1) = 0$
  $(\mu + \lambda) \cdot x(1) = \lambda$
  $x(1) = \frac{\lambda}{\mu+\lambda} \rightarrow x(0) = \frac{\mu}{\mu+\lambda}$
- Numerical example: $\lambda = \frac{2}{s}, \mu = \frac{3}{s} \rightarrow x(0) = 0.6, x(1) = 0.4$

Note: (II) does not need to be given in vector-matrix form, solving (13.5) and (13.6) is sufficient.

9.3.2 DTMC

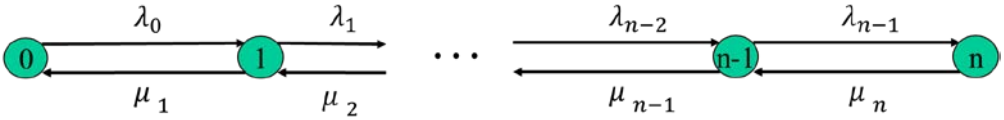
- State transition diagram



- State transition matrix $P = \begin{pmatrix} 1-a & a \\ b & 1-b \end{pmatrix}$
- Conditions for stationary state distribution
  - (I) $x(0) + x(1) = 1 \rightarrow x(0) = 1 - x(1)$
  - (II) $x \cdot P = x$
    - (IIb) $a \cdot x(0) + (1-b) \cdot x(1) = x(1)$
- Solve the equation system
  (I) in (IIb)
  $a \cdot [1 - x(1)] + (1-b) \cdot x(1) = x(1)$
  $a - a \cdot x(1) + x(1) - b \cdot x(1) = x(1)$
  $a = x(1) \cdot (1 + a - 1 + b)$
  $x(1) = \frac{a}{a+b} \rightarrow x(0) = \frac{b}{a+b}$
- Numerical example: $a = 0.2, b = 0.3 \rightarrow x(0) = 0.6, x(1) = 0.4$

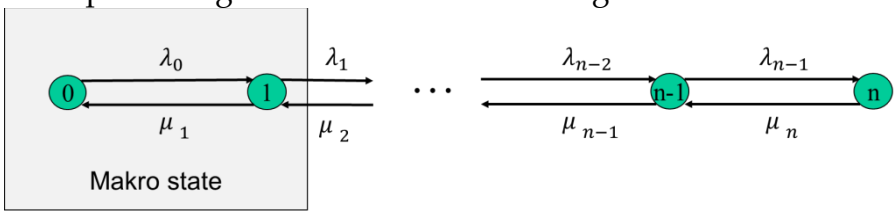
9.4 Birth-Death Processes

- Markov process
- Often (but not necessarily) one-dimensional state space, e.g., number of customers in the system
- State transitions only between "neighboring" states
- State transition diagram of a typical birth-death process with n+1 states



- This structure can be used to model M/M/s-K systems
  - M/M/n-0 (s=n, K=0)
    - Upward transition rates $\lambda_i = \lambda$
    - Downward transition rates $\mu_i = i \cdot \mu$
  - M/M/1-K (s=1, K=n-1)
    - Upward transition rates $\lambda_i = \lambda$
    - Downward transition rates $\mu_i = \mu$
  - M/M/s-K (K=n-s)
    - Upward transition rates $\lambda_i = \lambda$
    - Downward transition rates $\mu_i = \begin{cases} i \cdot \mu & \text{for } i \leq s \\ s \cdot \mu & \text{for } i > s \end{cases}$
- Equation system
  - $\sum_i x(i) = 1$
  - $\lambda_0 \cdot x(0) = \mu_1 \cdot x(1)$
  - $(\lambda_i + \mu_i) \cdot x(i) = \lambda_{i-1} \cdot x(i-1) + \mu_{i+1} \cdot x(i+1), i = 1, 2, \dots, n-1$
  - $\lambda_{n-1} \cdot x(n-1) = \mu_n \cdot x(n)$

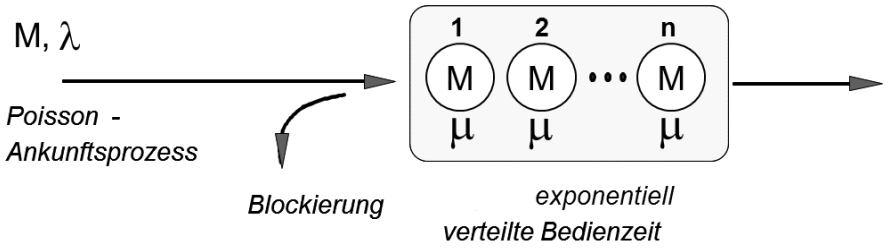
- Simplification of the calculation of the stationary state distribution
  - Macro state: set of neighboring states
  - Example: contiguous sets of nodes starting with state 0



  - Principle for new equations: sum of weighted inflow rates into macro state equals sum of weighted outflow rates out of macro state
  - Simplified but equivalent equation system
    - $\sum_i x(i) = 1$
    - $\lambda_{i-1} \cdot x(i-1) = \mu_i \cdot x(i), i = 1, 2, \dots, n$
  - Solution
    - $x(i) = x(0) \cdot \frac{\prod_{0<k\leq i} \lambda_{k-1}}{\prod_{0<k\leq i} \mu_k}, i = 1, 2, \dots, n$ and
    - $x(0) = \left(1 + \sum_{0<i\leq n} \frac{\prod_{0<k\leq i} \lambda_{k-1}}{\prod_{0<k\leq i} \mu_k}\right)^{-1}$

9.5 Loss System M/M/n-0

9.5.1 Model and Performance Metrics



State probabilities: $x(i) = \frac{\left(\frac{a^i}{i!}\right)}{\sum_{0 \le k \le n}\left(\frac{a^k}{k!}\right)}, i = 0, \ldots, n$ (Erlang formula)

with offered load $a = \frac{\lambda}{\mu}$ (pseudo unit: Erlang (Erl))

PASTA rule: Poisson Arrivals See Time Averages ⇒
Blocking probability $p_B = x(n)$ (Erlang-B formula)

Relative offered load $\rho = \frac{a}{n}$

Utilization: $\rho_u = \frac{a \cdot (1 - p_B)}{n}$

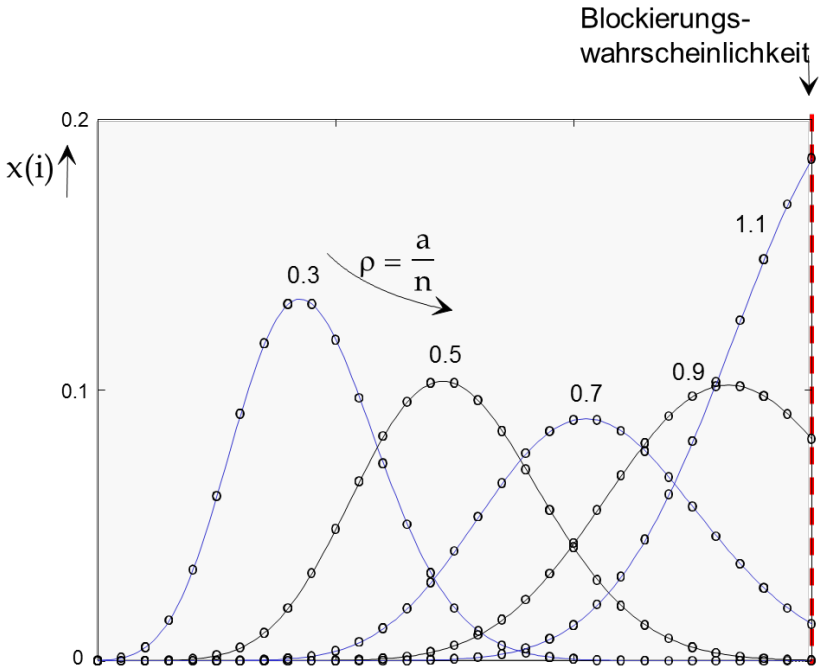


Abbildung 1: Stationary state distribution of a M/M/30-0 loss system for different relative offered load ρ.

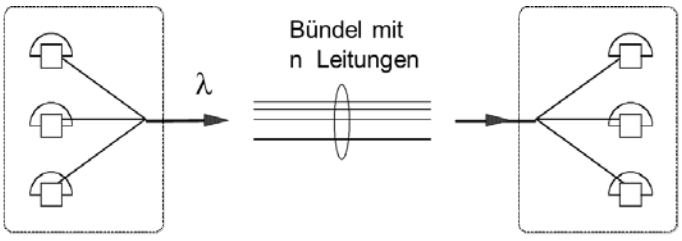
9.5.2 When is PASTA not applicable?

Assume a M/M/1 loss system *where requests arrive with rate λ in the presence of a free service unit and $\frac{\lambda}{2}$ in the presence of an occupied service unit.*
Thus, the arrival process depends on the system state. Therefore, it is not homogeneous and it is not a Poisson arrival process.
The system can be modelled with two states (0 free, 1 occupied).
The PASTA rule cannot be applied and the blocking probability is NOT $p_B = x(1)$.
The blocking probability is rather $p_B = \frac{\frac{\lambda}{2} \cdot x(1)}{\lambda \cdot x(0) + \frac{\lambda}{2} \cdot x(1)}$.

9.5.3 Application Example

Beispiel: Dimensionierung einer Ortsvermittlungsstelle (Telefon)

- Assumptions about
  - Arrival rates
    - One call per second arrives
    - Exponential inter-arrival time
    - $\Rightarrow \lambda = \frac{1}{s}$
  - Service rates
    - Calls take on average 90 s
    - Exponential call-holding time
    - $\Rightarrow \mu = \frac{1}{90s}$
  - $\Rightarrow$ Offered load $a = \frac{\lambda}{\mu} = 90$ Erl
- Needed: number of outgoing circuits so that the blocking probability of an incoming call is at most $p_B \leq 10^{-3}$
  - n=117 circuits are needed



n : Anzahl der Leitungen

B : Gesprächsdauer (E[B] = 90 sec)
λ : Anrufe/sec (Poisson Prozess) (λ=1 Anruf/sec)

⇨ Angebot
a=90

Dimensionierung mit
vorgegebener Dienstgüte $p_B \leq 10^{-3}$
(Quality of Service QoS)

⇨ n=117 Leitungen

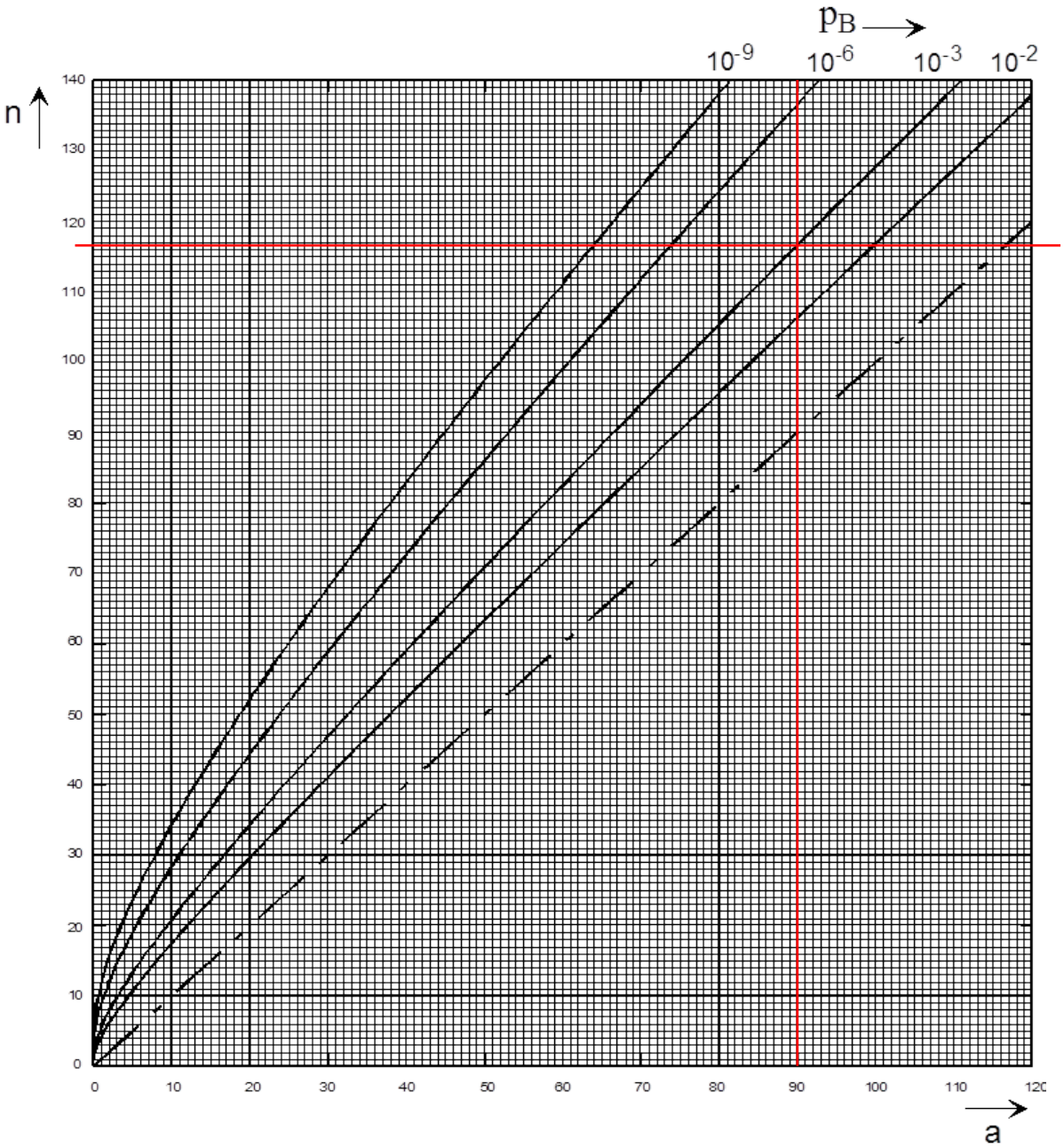


Abbildung 2: Benötigte Anzahl von Bedieneinheiten n bei einem M/M/n Verlustsystem um eine gewisse Blockierwahrscheinlichkeit p_B bei vorgegebener Last a nicht zu überschreiten. Für geringere Blockierwahrscheinlichkeiten werden mehr Bedieneinheiten benötigt.

9.5.4 Economy of Scale (Bündelungsgewinn) with M/M/n

- Bei größerem Angebot werden relativ weniger Leitungen benötigt um die gleiche Blockierwahrscheinlichkeit zu erreichen
- Bei größerem Angebot bessere Auslastung von entsprechend dimensionierten Systemen möglich

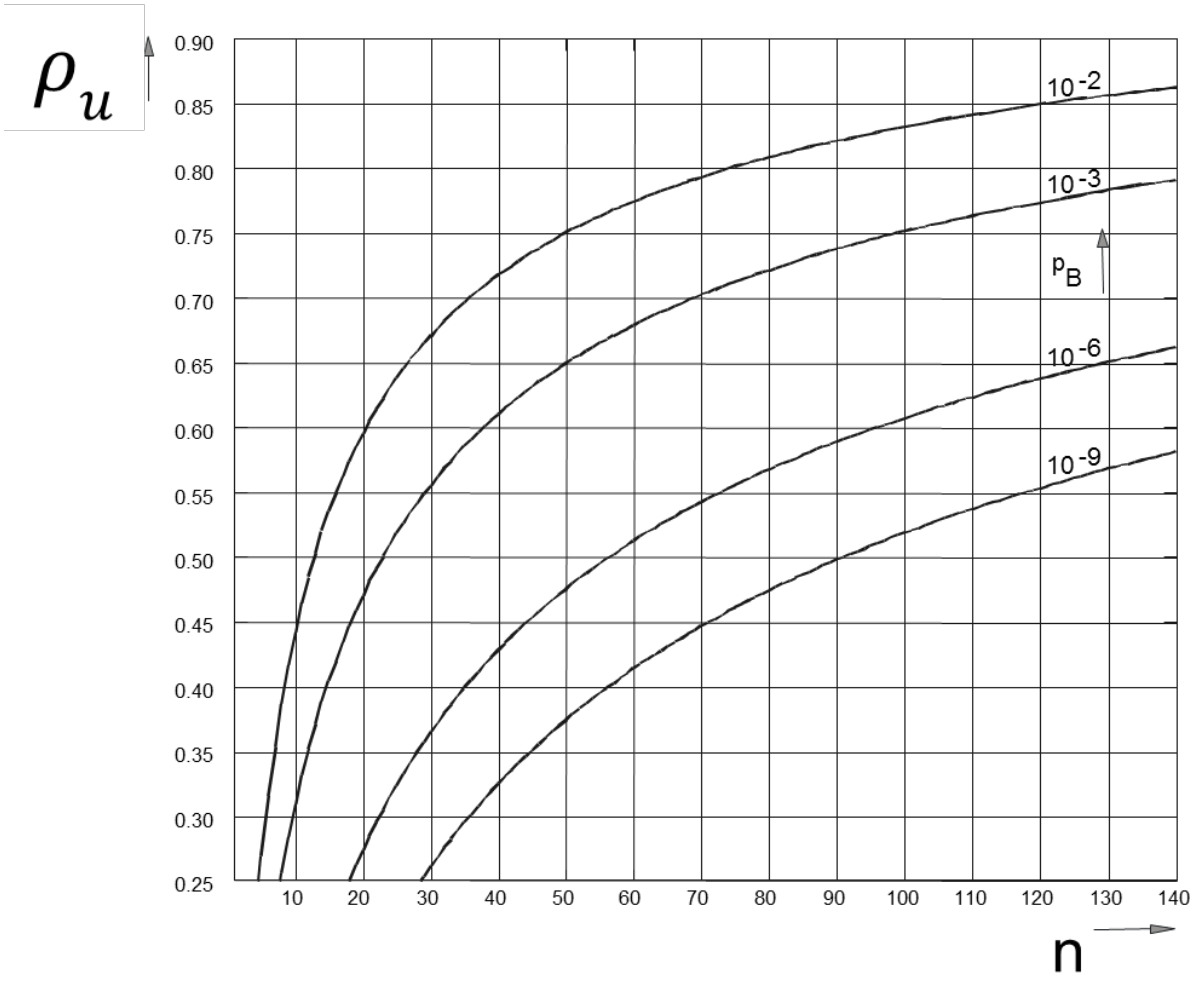


Abbildung 3: Auslastung von Bedieneinheiten, die so stark belastet sind, dass eine vorgegebene Blockierwahrscheinlichkeit eingehalten wird.

- Bei gleicher Blockierungswahrscheinlichkeit steigt die Auslastung der Leitungen ρ_u mit der Bündelgröße n
- Größere Leitungsbündel sind wirtschaftlicher
- Bündelungsgewinn lässt sich nicht beliebig steigern

- Andere Sichtweise:
  Wenn zwei kleine M/M/n Systeme zu einem großen M/M/2n System zusammengefasst werden, ist die Blockierwahrscheinlichkeit des großen Systems geringer als die der kleinen Systeme.

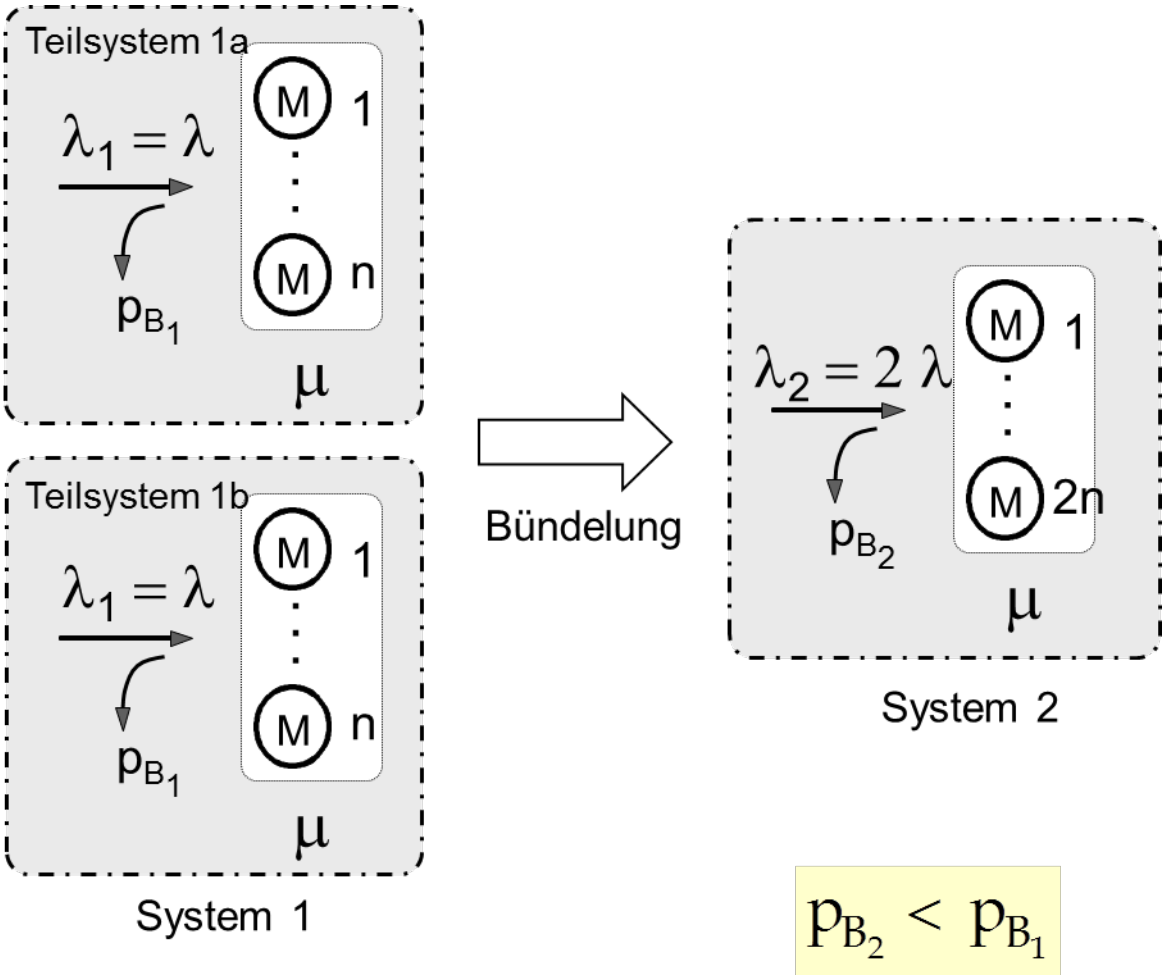
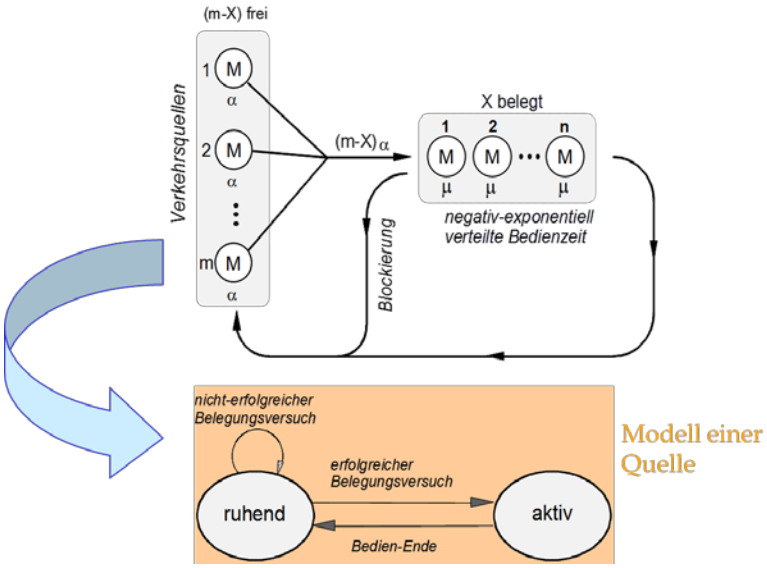


Abbildung 4: Combining customer arrivals and service unites of separate M/M/n-0 loss systems into a single one leads to lower blocking probabilities.

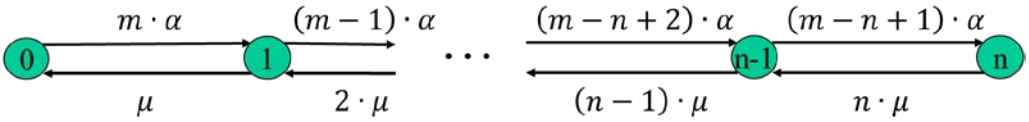
9.6 Loss System M/GI/n-0

Same formulae as for M/M/n-0

9.7 Loss System with a Finite Number of Sources



- Number of service units: n
- Number of sources: $m \geq n$
- Source can be inactive
  - Distribution function of inactive time is $I(t) = 1 - e^{-\alpha \cdot t}$
- At the end of an inactivity phase, the source makes a service request
  - Free service unit available
    - Source is serviced
    - Distribution function of service time is $B(t) = 1 - e^{-\mu \cdot t}$
    - Source becomes again inactive after service completion.
  - No free service unit available
    - Source is blocked
    - Source becomes inactive again
- State transition graph (state is number of busy service units)



- Distribution of busy service units
  - State probabilities: $x(i) = \frac{\binom{m}{i} \cdot (a^*)^i}{\sum_{0 \leq k \leq n} \binom{m}{k} \cdot (a^*)^k}, i = 0, \ldots, n$
  - With $a^* = \frac{\alpha}{\mu}$ (offered load of an inactive source)
- Blocking probability
  - $p_B = \frac{\binom{m-1}{n} \cdot (a^*)^n}{\sum_{0 \leq k \leq n} \binom{m-1}{k} \cdot (a^*)^k}$ (Engset formula)
  - Note: blocking probability is based on state probability for n busy service units in the presence of $m - 1$ customers, that's what the requesting customer sees.
- Formulae also hold for GI distributed service time.

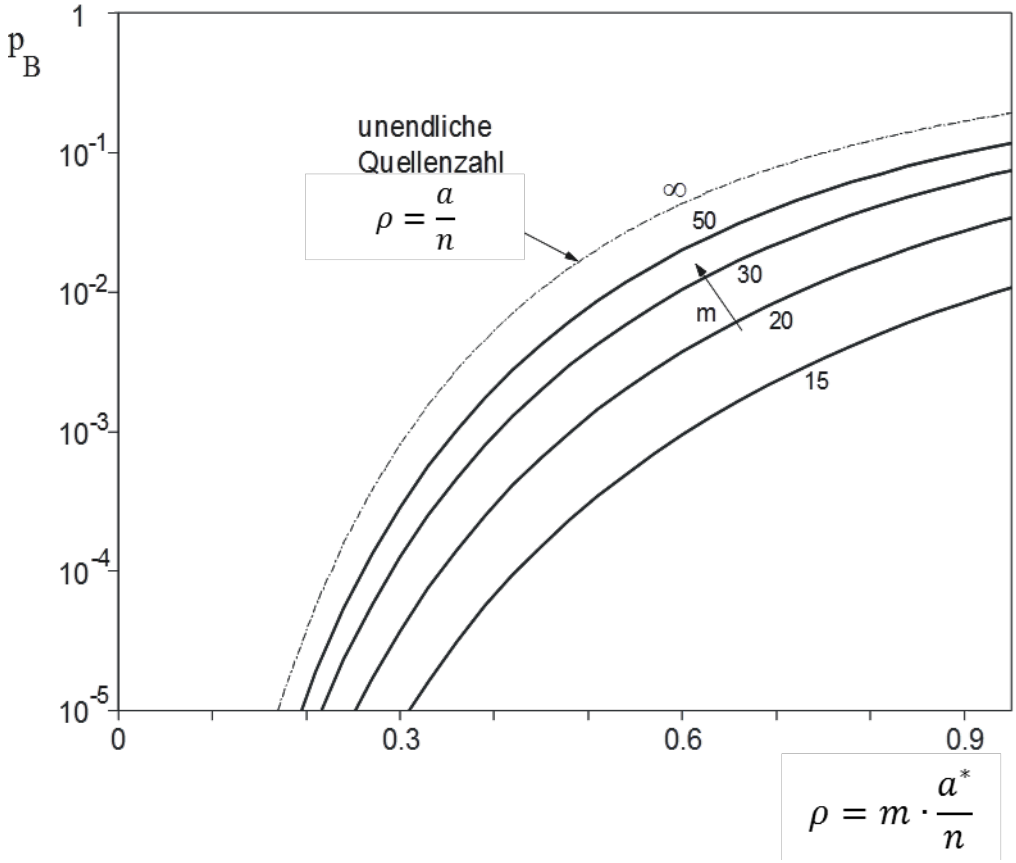
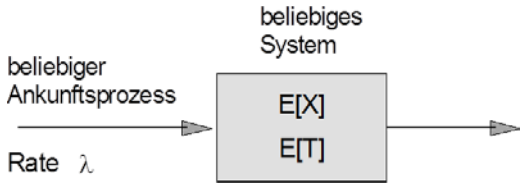


Abbildung 5: Blocking probabilities according to the Engset formula (The number of servers n is fixed but unknown, in any case $n < 15$). For an increasing number of customers m the blocking probability converges to the one of an M/M/n-0 loss system.

9.8 Theorem von Little

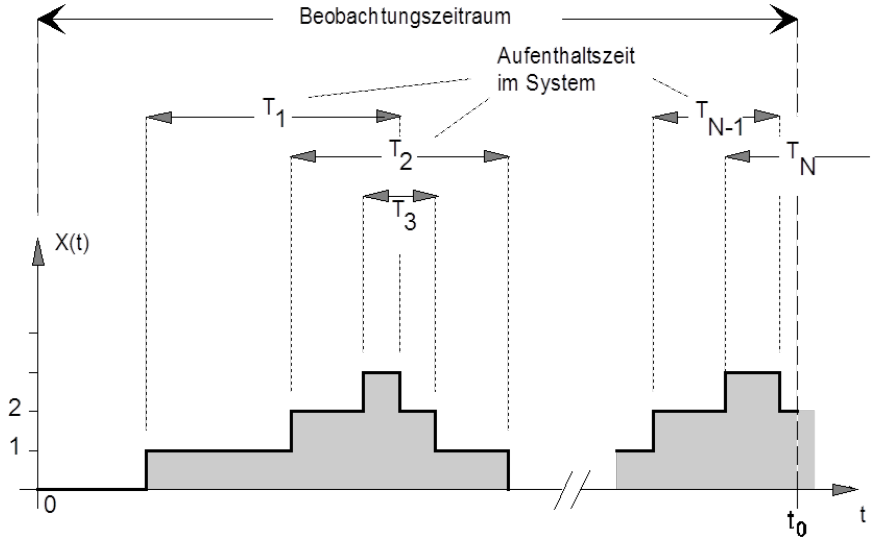


Nomenklatur (allgemein und im endlichen Beispiel für den Beweis)

- $\lambda, \bar{\lambda}$: mittlere Ankunftsrate des Ankunftsprozesses
- E[T], $\bar{T}$: mittlere Aufenthaltszeit im System
- E[X], $\bar{X}$: mittlere Anzahl von Anforderungen im System

Es gilt: $\lambda \cdot E[T] = E[X]$ (Theorem von Little)

Beweis



Wir betrachten endlichen Prozessausschnitt

- t_0: Dauer des betrachteten Prozesses
- N: Anzahl Ankünfte
- Es gilt
  - (1) $\bar{\lambda} = \frac{N}{t_0} \rightarrow t_0 = \frac{N}{\bar{\lambda}}$
  - (2) $\bar{X} = \frac{1}{t_0} \cdot \int_0^{t_0} X(t)dt$
  - (3) $\bar{T} \approx \frac{1}{N} \cdot \int_0^{t_0} X(t)dt \rightarrow \int_0^{t_0} X(t)dt \approx N \cdot \bar{T}$
- Nach Einsetzen von (1) und (3) in (2) folgt das Little-Theorem $\bar{X} \approx \bar{\lambda} \cdot \bar{T}$

9.9 Waiting System M/M/n-∞

- Can be modeled as birth-death process with infinite state space
- State indicates number of customers in the system
- State probabilities for $a < n$ (stability condition):

  $x(i) = \begin{cases} x(0) \cdot \frac{a^i}{i!} & i = 0, 1, \dots, n \\ x(n) \cdot \rho^{i-n} & i > n \end{cases}$ with relative offered load $\rho = \frac{a}{n}$ and

  $x(0) = \left(\sum_{0 \le k < n} \left(\frac{a^k}{k!} \right) + \frac{a^n}{n!} \cdot \frac{1}{1-\rho} \right)^{-1}$
- For $n = 1$
  - $x(0) = \left(1 + a \cdot \frac{1}{1-\rho}\right)^{-1} = \left(\frac{1-\rho}{1-\rho} + \frac{\rho}{1-\rho}\right)^{-1} = 1 - \rho$
  - $x(1) = x(0) \cdot a = (1-\rho) \cdot \rho;\ x(i) = (1-\rho) \cdot \rho^i$ sogar für alle $i \ge 0$
- Utilization ρ_u equals relative offered load ρ since $p_B = 0$.

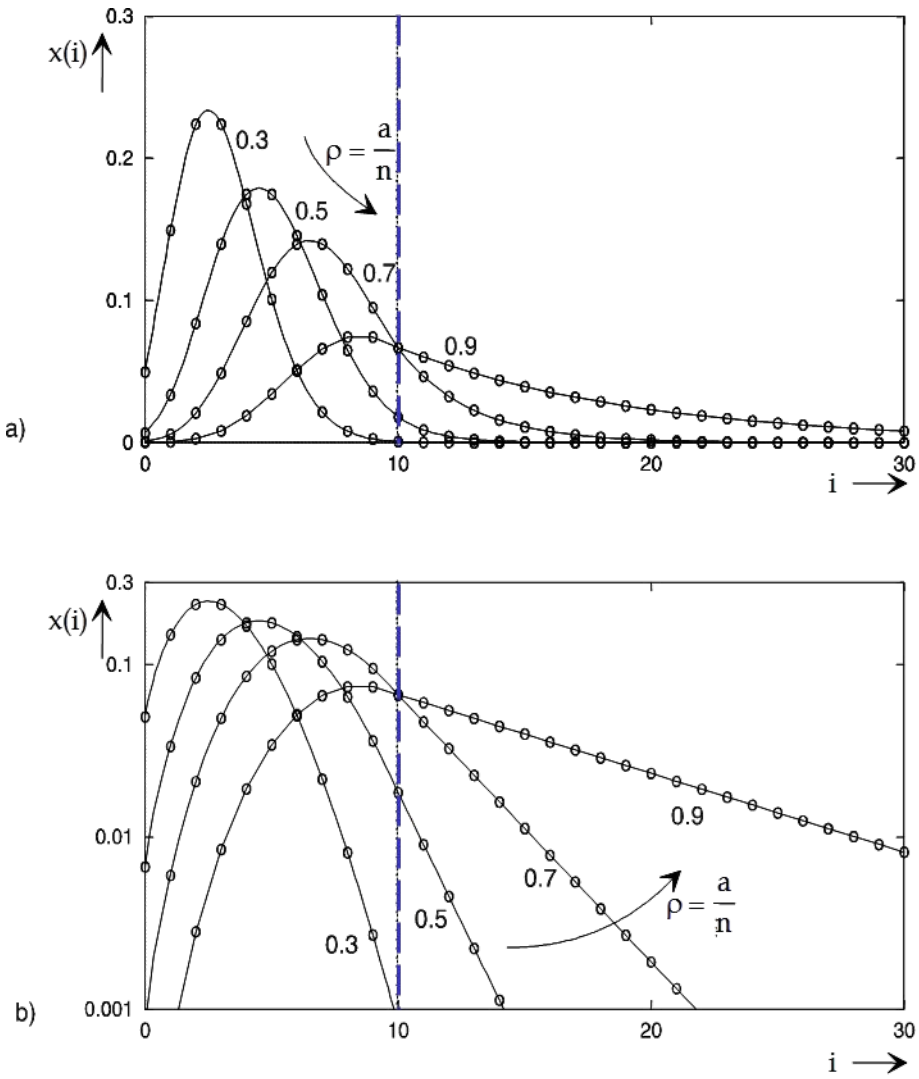


Abbildung 6: Stationary state distribution of an M/M/10-∞ waiting system for different relative offered load on a linearly and logarithmically scaled y-axis.

Waiting probability $p_W = \sum_{n \le i < \infty} x(i) = x(n) \cdot \frac{1}{1-\rho}$ (Erlang waiting formula)

- For $n = 1$: $p_W = 1 - x(0) = \rho$

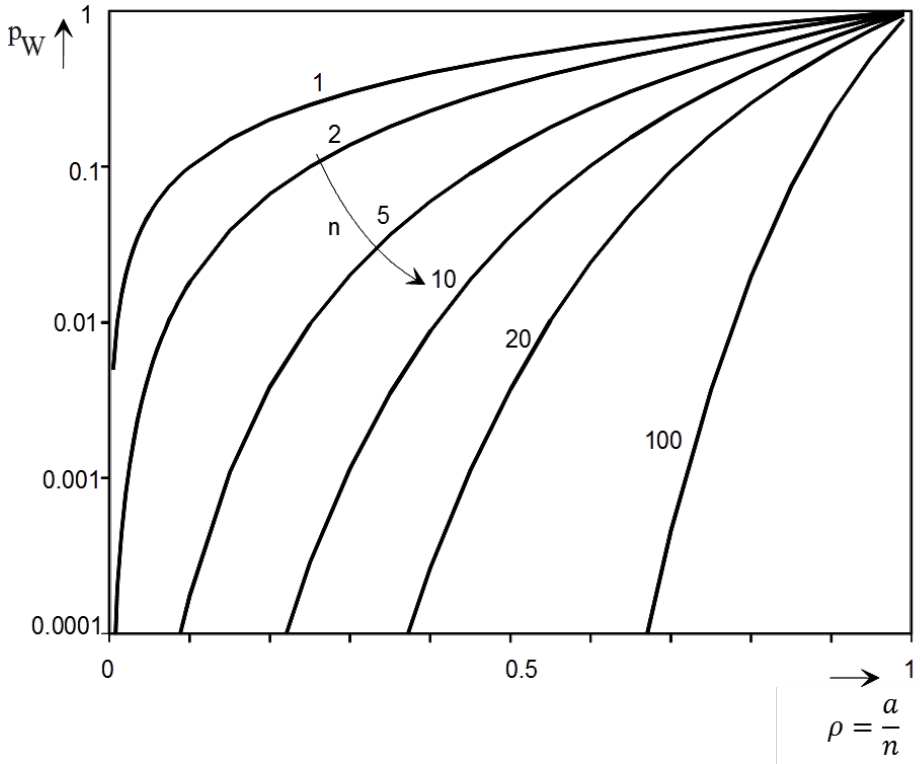


Abbildung 7: Waiting probabilities in a M/M/n-∞ waiting system. Economy of scale: systems with more service units reveal lower waiting probabilities for the same relative load.

X_w: number of waiting customers
Mean queue length:

$$E[X_w] = \sum_{n<i<\infty} (i-n) \cdot x(i) = x(n) \cdot \sum_{0<i<\infty} i \cdot \rho^i = x(n) \cdot \frac{\rho}{(1-\rho)^2} = p_W \cdot \frac{\rho}{1-\rho}$$

Mean waiting time of all customers $E[W]$
Consider system I: queue is traversed by all customers, some have waiting time $W = 0$

- $\lambda_I = \lambda$
- $E[X_I] = E[X_W]$

$\Rightarrow E[W] = E[T_I] = \frac{E[X_I]}{\lambda_I} = \frac{E[X_W]}{\lambda} = \frac{p_W}{\lambda} \cdot \frac{\rho}{1-\rho}$ (Little)

Mean waiting time of waiting customers $E[W_W]$
Consider system II: queue is traversed only by waiting customers

- $\lambda_{II} = \lambda \cdot p_W$
- $E[X_{II}] = E[X_W]$

$\Rightarrow E[W_W] = E[T_{II}] = \frac{E[X_{II}]}{\lambda_{II}} = \frac{p_W \cdot \frac{\rho}{1-\rho}}{\lambda \cdot p_W} = \frac{1}{\lambda} \cdot \frac{\rho}{1-\rho}$ (Little)

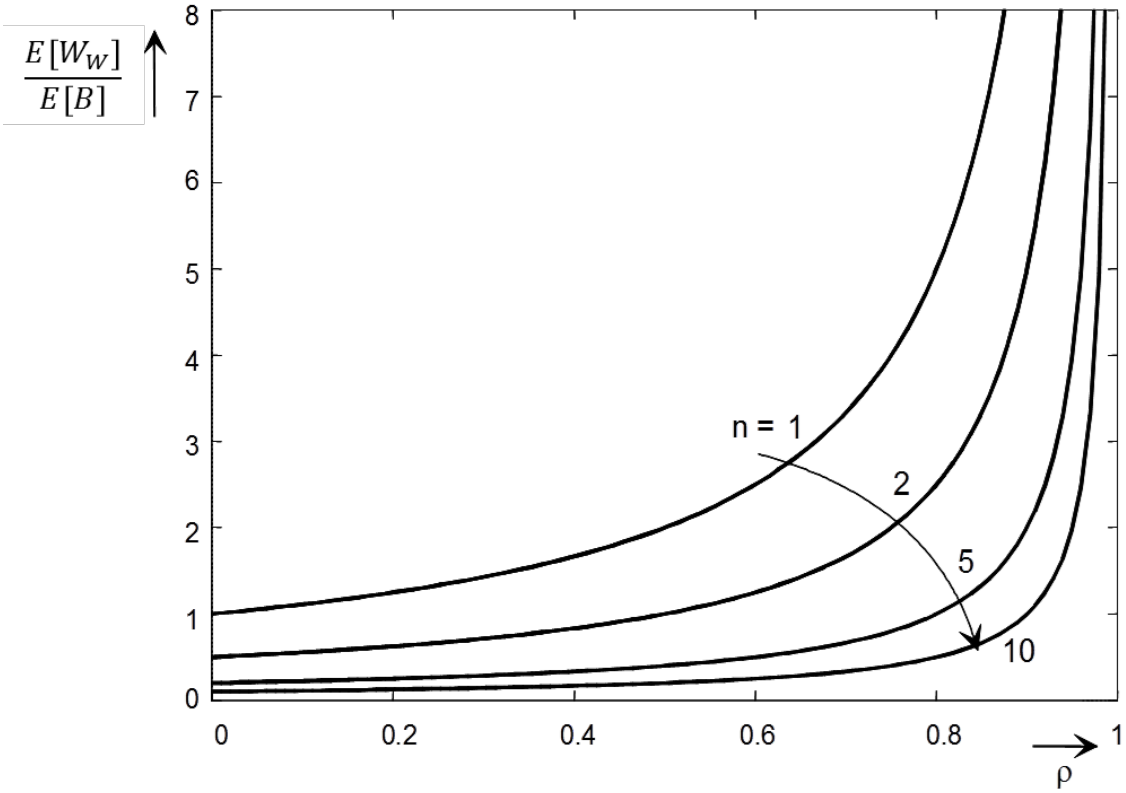


Abbildung 8: Mean waiting time of waiting customers in a M/M/n-∞ queuing system (normalized by $E[B] = 1/\mu$).

Why do we get a value of $1/n$ for $\rho \to 0$ in Abbildung 8?

- It's mean waiting time only for waiting customers ⇒ must be larger than zero, for $\rho \to 0$ we find waiting customers just very rarely
- Service unit in a blocking M/M/n-∞ system becomes available with rate $n \cdot \mu$ ⇒ mean time until then is $\frac{E[B]}{n}$.

Distribution function of the waiting time of

- Non-waiting customers: $P(W_{\overline{W}} \leq t) = 1$
- Waiting customers: $P(W_W \leq t) = 1 - e^{-(1-\rho)\cdot n\cdot\mu\cdot t}$
- All customers: $P(W \leq t) =$
  $(1 - p_W) \cdot P(W_{\overline{W}} \leq t) + p_W \cdot P(W_W \leq t) =$
  $(1 - p_W) \cdot 1 + p_W \cdot \left(1 - e^{-(1-\rho)\cdot n\cdot\mu\cdot t}\right) =$
  $1 - p_W \cdot e^{-(1-\rho)\cdot n\cdot\mu\cdot t}$

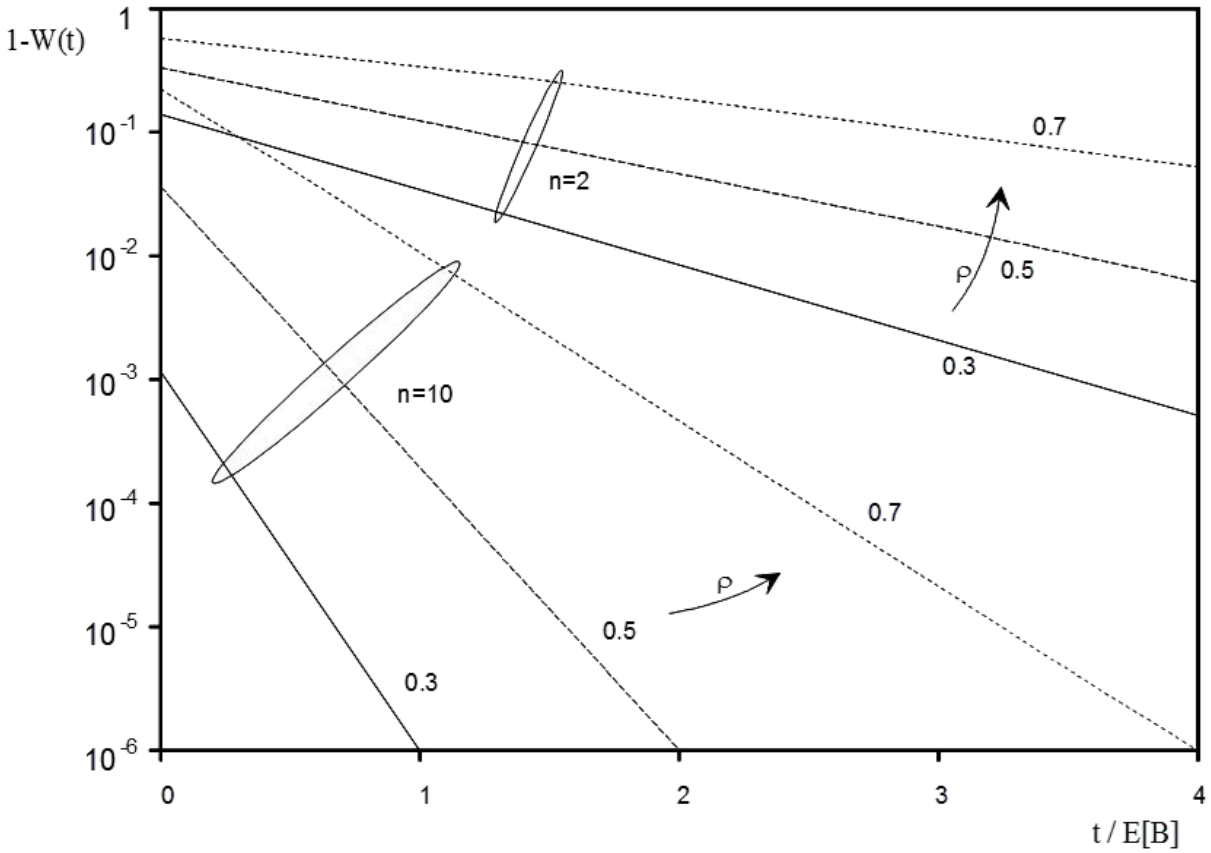


Abbildung 9: Complementary distribution function of the waiting time of all customers in an M/M/n-∞ waiting system.

Bündelungsgewinn in M/M/n-∞

- Systeme, die auf eine maximale Wartewahrscheinlichkeit bzw. Wartezeit dimensioniert sind, können für höhere angebotene Lasten wirtschaftlicher arbeiten (höhere Auslastung bei gleicher p_W bzw. E[W] möglich)
- Werden die Ressourcen zweier M/M/n-∞ Systeme zu einem großen M/M/2n-∞ zusammengefasst, dann weist dieses eine kleinere Wartewahrscheinlichkeit bzw. kleinere Wartezeiten auf.

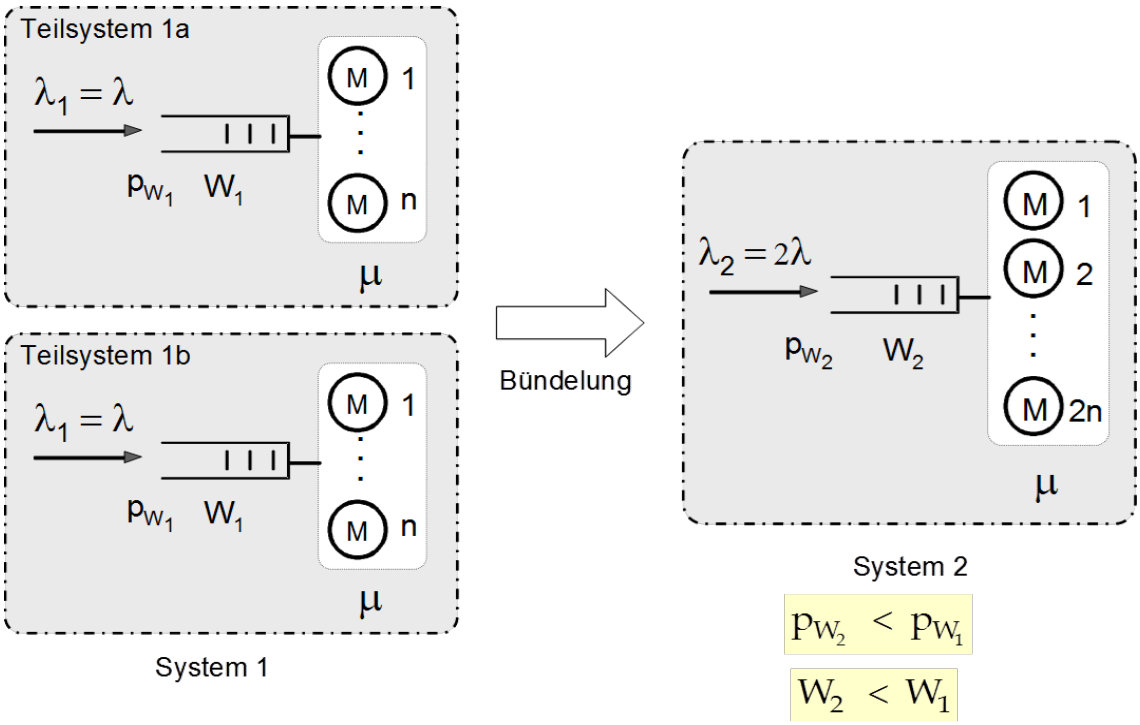


Abbildung 10: M/M/n-∞ waiting systems with more server units and the same relative load lead to lower waiting probabilities and shorter waiting times.

9.10 Waiting System M/GI/1-∞

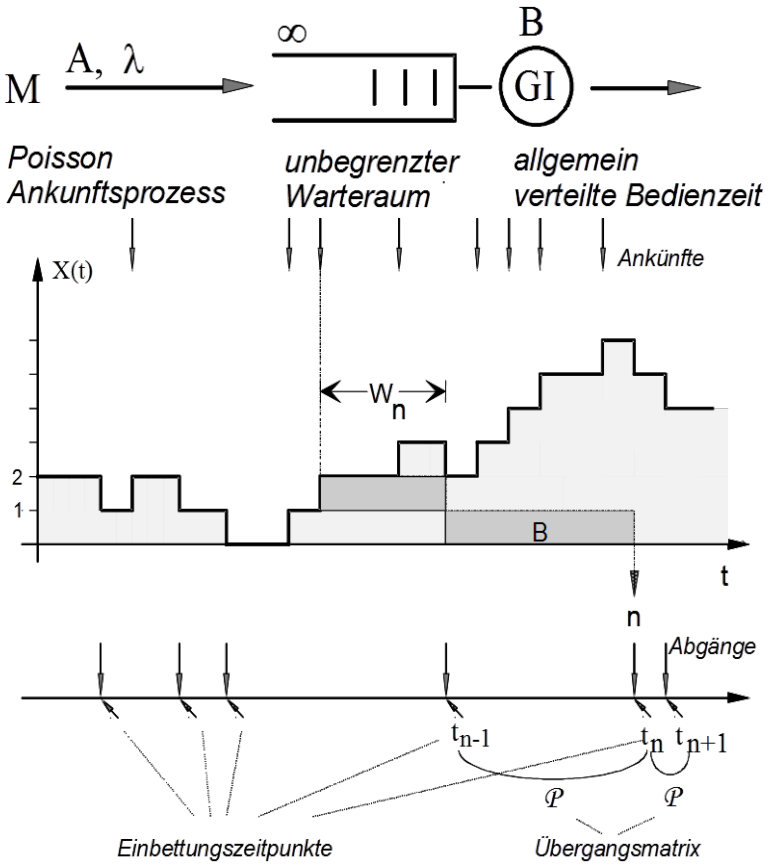


Abbildung 11: Embedded Markov chain with embedded points at service completion instants used for analysis.

Waiting probability: $p_W = \rho$

Mean waiting time of all customers: $E[W] = E[B] \cdot \frac{\rho \cdot (1+(c_B)^2)}{2 \cdot (1-\rho)} = \frac{\lambda \cdot E[B^2]}{2 \cdot (1-\rho)}$

Mean waiting time of waiting customers: $E[W_W] = \frac{E[W]}{p_W} = E[B] \cdot \frac{(1+(c_B)^2)}{2 \cdot (1-\rho)}$ with coefficient of variation c_B of service time B

Higher moments of waiting time of all customers (Takács recursion formula):

$$E[W^0] = 1$$

$$E[W^k] = \frac{\lambda}{(1-\rho)} \cdot \sum_{0<i\leq k} \binom{k}{i} \cdot \frac{E[B^{i+1}]}{i+1} \cdot E[W^{k-i}]$$

$$k = 1: \qquad E[W] = \frac{\lambda \cdot E[B^2]}{2 \cdot (1-\rho)}$$

$$k = 2: \quad E[W^2] = 2 \cdot E[W]^2 + \frac{\lambda \cdot E[B^3]}{3 \cdot (1-\rho)}$$

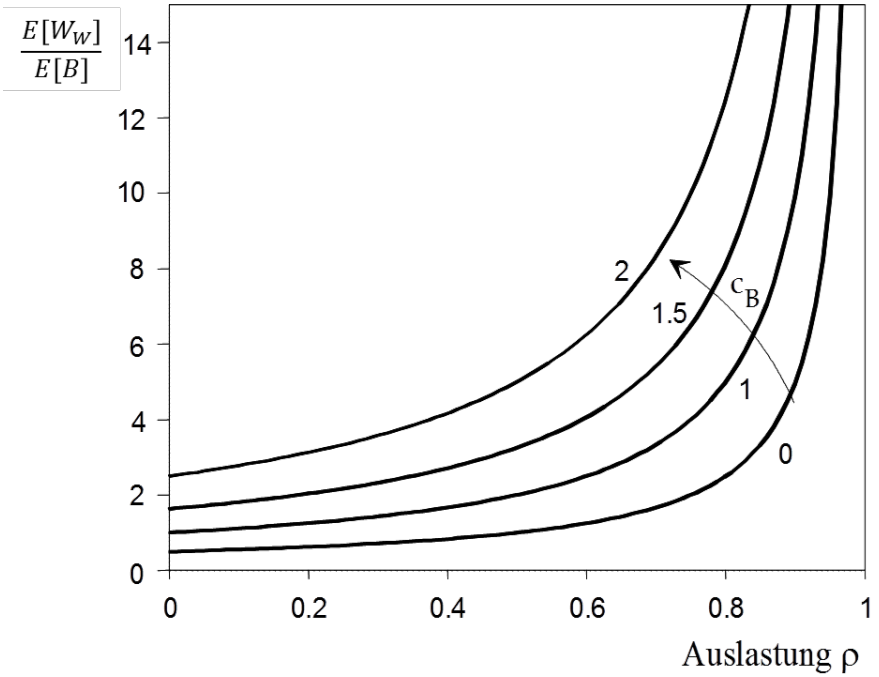


Abbildung 12: Mean waiting time of waiting customers for a M/G/1-∞ waiting system normalized by E[B].

The distribution function of the waiting time may be approximated by a Gamma distribution (M. Menth, R. Henjes, C. Zepfel, P. Tran-Gia: "Gamma-Approximation for the Waiting Time Distribution Function of the M/GI/1-∞ Queue", in 2nd Conference on Next Generation Internet Networks (NGI), April 2006, Valencia, Spain).

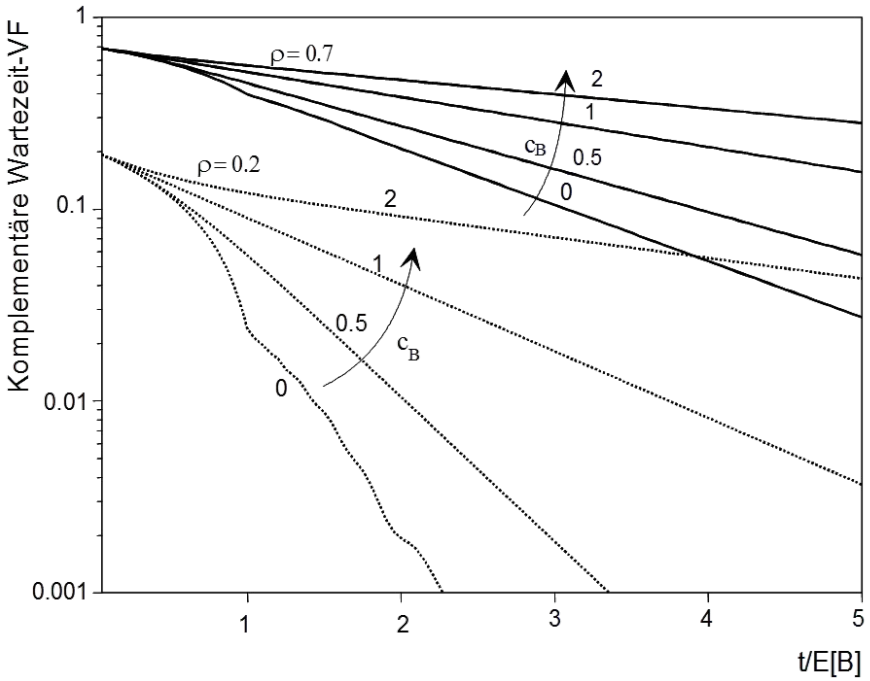


Abbildung 13: Complementary distribution function of the waiting time of all customers for a M/G/1-∞ waiting system normalized by E[B] (approximated by a discrete-time analysis method).

9.11 Waiting System M/D/1-∞

See: [VBI15] V.B. Iversen, "TELETRAFFIC ENGINEERING and NETWORK PLANNING", 2015,
http://orbit.dtu.dk/files/118473571/Teletraffic_34342_V_B_Iversen_2015.pdf

9.12 Waiting System n·D/D/1-∞

Approximationsformel für die Wartezeitverteilung von n·D/D/1-∞

- n: Anzahl der periodischen Flüsse
- A: konstante Zwischenankunftszeit einer Quelle
- B: konstante Kundenbedienzeit
- Daraus ergibt sich Systemauslastung $\rho = \frac{n \cdot B}{A}$
- $P(W > t) \approx \exp\left(\left(\frac{-2 \cdot t}{B}\right) \cdot \left(\frac{t}{n \cdot B} + 1 - \rho\right)\right)$ (aus COST 242)
- Randbedingung für die Gültigkeit der Formel
  - $\rho < 1$
  - Die Wartezeit kann nicht länger dauern als A

    $\Rightarrow$ t>A zu verwenden ist nicht sinnvoll
  - o $\Rightarrow$ relativ geringe mittlere Wartezeit sogar für sehr hohe Systemauslastungen von knapp unter 100%

- Illustration des Wartezeitverhaltens von n*D/D/1-∞ Wartesystemen in Abhängigkeit von Periode und Anzahl gemultiplexter Flüsse
  (Quelle: Michael Menth und Stefan Mühleck: „Packet Waiting Time for Multiplexed Periodic On/Off Streams in the Presence of Overbooking", International Journal of Communication Networks and Distributed Systems (IJCNDS), vol. 4, no. 2, pages 207 – 229, 2010)

  Das nD/D/1-∞ System ist periodisch, so dass sich die Wartezeiten innerhalb einer Prozessrealisierung nach jeder Periode wiederholen. Allerdings unterscheiden sich die Wartezeiten in unterschiedlichen Prozessrealisierungen, weil sich die Phasen der Sendezeitpunkte innerhalb einer Periode (Sendemuster) unterscheiden.

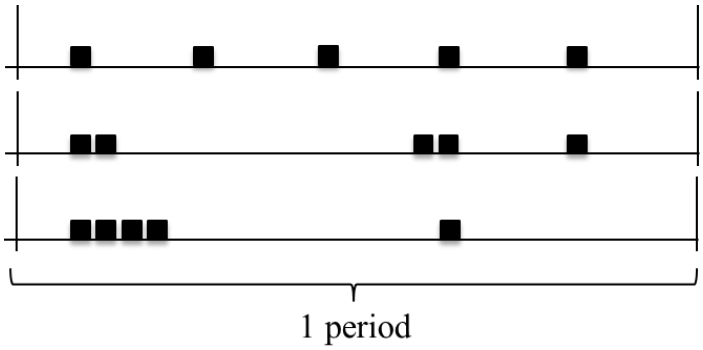


Abbildung 14: Unterschiedliche Sendemuster (Phasen der Sendezeitpunkte innerhalb einer Periode).

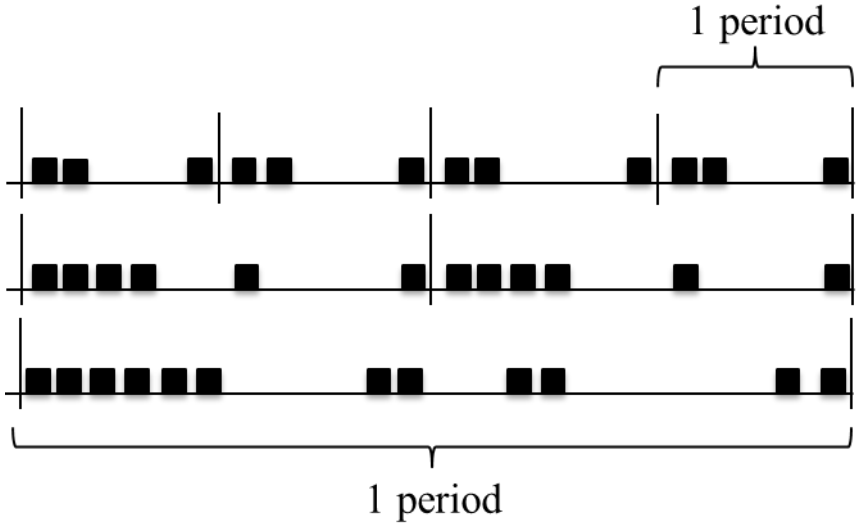


Abbildung 15: Wenn im nD/D/1-∞ System die Perioden bei gleicher Last länger werden, dann senden mehr Quellen pro Periode. Es können sich größere Bursts bilden und somit längere Wartezeiten auftreten.

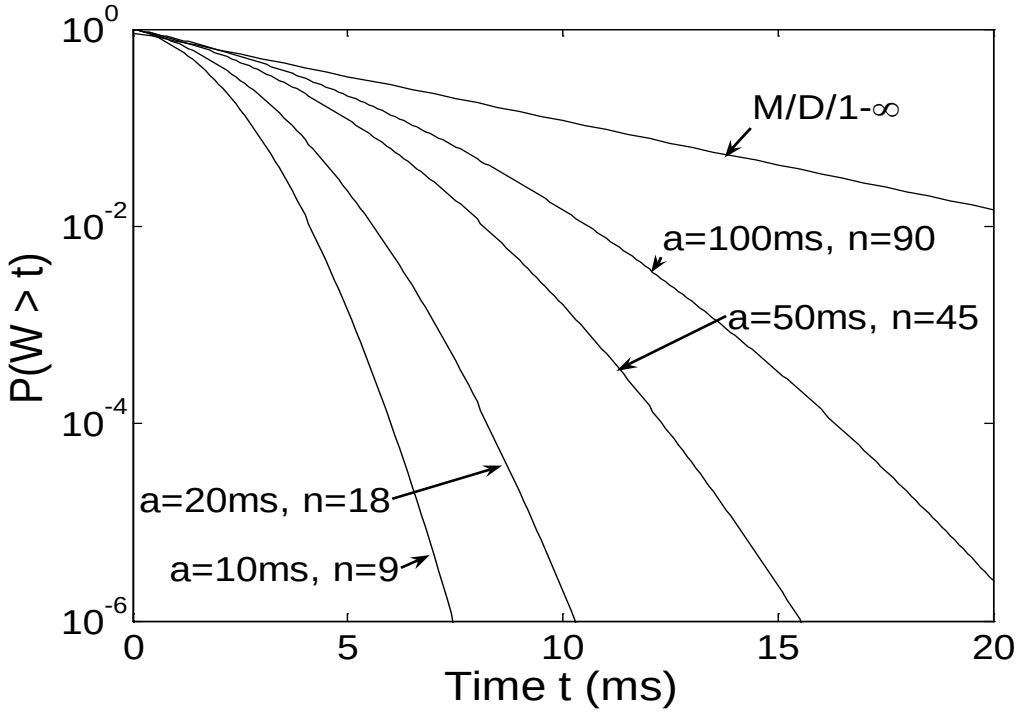


Abbildung 16: Complementary cumulative distribution function (ccdf) of the waiting time of an n*D/D/1-∞ waiting system. The number of periodic sources is n. Utilization $\rho = 0.9$ and packet size $B \sim 1\ ms$ are fixed. The period a and the number of multiplexed flows vary. With increasing period, the waiting time converges to the one of an M/D/1-∞ waiting system with utilization $\rho = 0.9$.

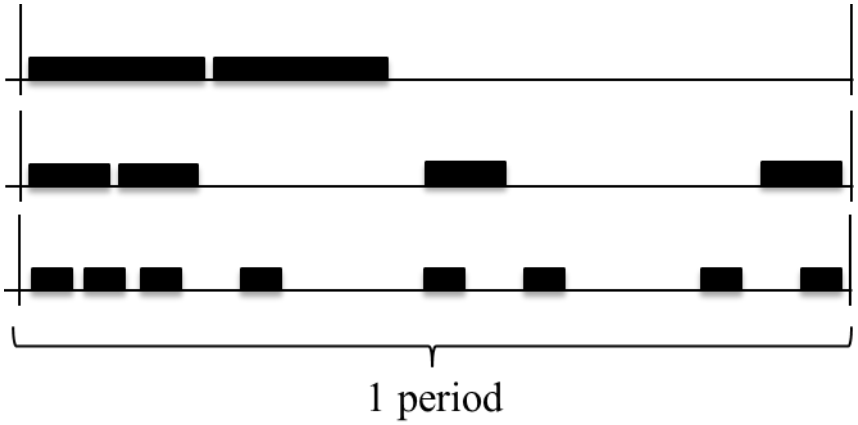


Abbildung 17: Wenn im nD/D/1-∞ System die Anzahl der Sender bei gleicher Last und Periodenlänge zunimmt, werden die Pakete kürzer. Die Last verteilt sich besser über die Periode und Wartezeiten werden kürzer.

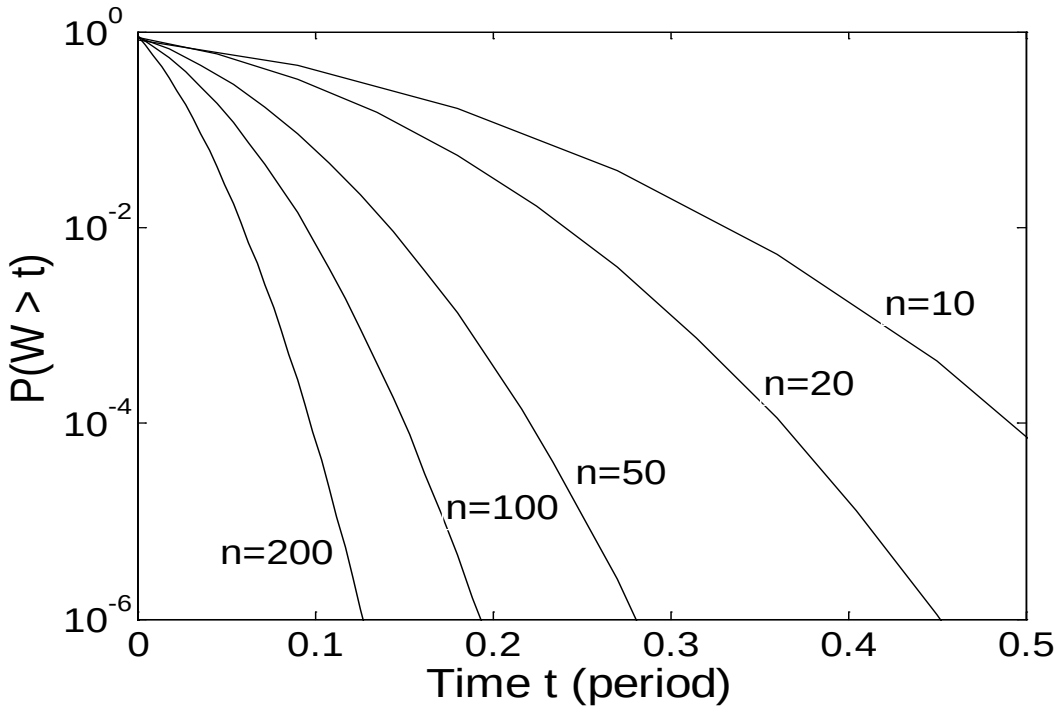


Abbildung 18: Complementary cumulative distribution function (ccdf) of the waiting time of an n*D/D/1-∞ waiting system. Utilization $\rho = 0.9$ and period a are fixed. The number of periodic sources n and the packet size B vary. With increasing number of sources, the waiting time decreases.

9.13 Multi-Rate M/GI/∞ Queue

- K. W. Ross and D. H. K. Tsang. The Stochastic Knapsack Problem. IEEE/ACM Transactions on Networking, 37(7):740–747, 1989
- Service class $s, 1 \leq s \leq S$ has
  - Markovian arrival rate λ_s
  - Markovian service rate μ_s
  - Offered load $a_s = \frac{\lambda_s}{\mu_s}$
  - Requires c_s service units
- S-dimensional state $x = (x_1, \ldots, x_S)$ with x_s indicating the number of requests in service of type s
- Link capacity required for state x

$$c(x) = \sum_{1 \leq s \leq S} c_s \cdot x_s$$

- Computation of state probabilities

$$p(x) = \prod_{1 \leq s \leq S} \frac{(a_s)^{x_s}}{x_s!} \cdot e^{-a_s}$$

- Formulae exist for probability that less than c capacity is required P(C≤c)
- Application
  - Traffic mix of small and large flows
    - 64 kb/s voice, 256 kb/s low-bitrate video, 2048 kb/s high-bitrate video
  - What's the probability for certain traffic rates on a link without admission control (blocking)?
  - M. Menth, R. Martin, J. Charzinski: "Capacity Overprovisioning for Networks with Resilience Requirements", ACM SIGCOMM, Sept. 2006, Pisa, Italy

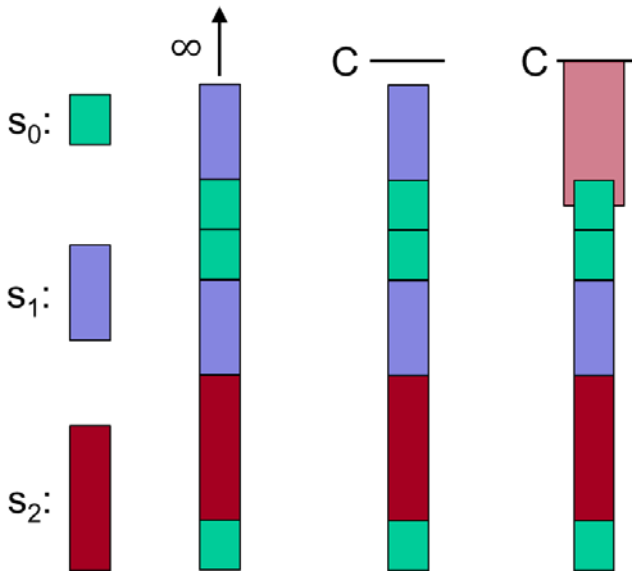


Abbildung 19: Multirate M/M/∞, M/M/C, and M/M/C with trunk reservation. The rectangles represent jobs/customers and their sizes are proportional to their required number of servers.

9.14 Kaufman & Roberts Formula for Blocking Probabilities in a Multi-Dimensional Loss System ("Multirate M/M/C-0")

Problem

- Server with C service units
- S different service classes
- Service class $s, 1 \leq s \leq S$ has
  - Markovian arrival rate λ_s
  - Markovian service rate μ_s
  - Offered load $a_s = \frac{\lambda_s}{\mu_s}$
  - Requires c_s service units
- New request for service s blocked if less than c_s free service units available
- What are the probabilities $x(c)$ for c busy service units?

Solution

- $x(c) = \frac{\tilde{x}(c)}{\sum_{0 \leq c \leq C} \tilde{x}(c)}$
- $\tilde{x}(c) = \begin{cases} 0 & \text{for} \quad c < 0 \\ 1 & \text{for} \quad c = 0 \\ \sum_{1 \leq s \leq S} a_s \cdot \frac{c_s}{c} \cdot \tilde{x}(c - c_s) & \text{for} \quad 0 < c \leq C \end{cases}$

Application

- Flows of different rates demand for admission to a link and are blocked if remaining capacity does not suffice
- Computation of service-specific blocking probabilities $p_{B,s} = \sum_{C - c_s < c \leq C} x(c)$
- May be used for link dimensioning: how much capacity C is needed to keep blocking probabilities below a certain value?
- M. Menth: "Efficient Admission Control and Routing in Resilient Communication Networks", Doktorarbeit, Universität Würzburg, Am Hubland, 97074 Würzburg, 2004, http://www.opus-bayern.de/uni-wuerzburg/volltexte/2004/994/pdf/Menth04.pdf

Observation: services with larger c_s experience larger $p_{B,s}$

9.15 Multirate M/M/C-0Twith Trunk Reservation

- Improvement for more fairness: equal $p_{B,s}$ for all services
- Block new request if less than $c_{max} = \max_{1 \leq x \leq S}(c_s)$ free service units available
- Modifications of equations above lead to sufficiently accurate approximation
  - $\tilde{x}(c) = \begin{cases} 0 & \text{for} \quad c < 0 \\ 1 & \text{for} \quad c = 0 \\ \sum_{1 \leq s \leq S} a_s \cdot \frac{c_{TR}(C,c,c_s)}{c} \cdot \tilde{x}(c - c_s) & \text{for} \quad 0 < c \leq C \end{cases}$
    with $c_{TR}(C, c, c_s) = \begin{cases} c_s & \text{for} \quad c - c_s \leq C - c_{max} \\ 0 & \text{for} \quad c - c_s > C - c_{max} \end{cases}$
  - $p_{B,s} = \sum_{C - c_{max} < c \leq C} x(c)$