Kapitel 3

§3.3 Markov-Modelle mit endlicher Population

Das M/M/m/K//M Modell

In den bisher betrachteten Systemen wurde immer angenommen, dass die Kundenpopulation unendlich groß ist. Der Ankunftsprozess ist daher unabhängig von der Anzahl der Aufträge, die bereits im System sind. Diese Annahme wollen wir nun aufgeben.

Der Poisson-Prozess für die eintreffenden Kunden ist nur dann ein gutes Modell, wenn es sich sehr großes Resevoir von potentiellen Kunden verliegt. Liegt (etwa in einem kleinen Ort oder wenn die "Kunden" zu reparierende Machinen in einer Firma sind) nur eine beschränkte Menge M von Möglichen Kunden vor, so geht auch diese Größe in die Berechnung ein.

Nun sei angennomen, dass es genau M potentielle Kunden (Aufträge) für das System gibt. Jeder Kunde kann sich zu einem bestimmten Zeitpunkt in genau einem von drei möglichen Zuständen befinden:
1. außerhalb des Systems - er bereitet sich auf den Eintritt in das System vor;
2. innerhalb des Systems auf Bedienung wartend;
3. innerhalb des Systems in der Bedienung.

Es wird angenommen, dass die Verweilzeit eines Kunden außerhalb des Systems eine Exponentielverteilung mit dem Parameter λ besitzt. Die Verweilzeiten verschiedener Kunden seien stochastisch unabhängig.

M/M/1/K//M Modell (Maschine Repairman Model, K=M)

Die Rolle der Kunden in unserem Modell spielen dort Maschinen, die nach einer exponentialverteilten Operationszeit ausfallen und repariert werden müssen.
Der Bediener ist ein Techniker, der für eine Reparatur eine exponentialverteilte Bedienzeit benötigt. Da er immer nur eine Maschine gleichzeitig bedienen kann, müssen während einer Reparatur weitere ausfallende Maschienen warten. Die folgende Skizze veranschaulicht dieses System.

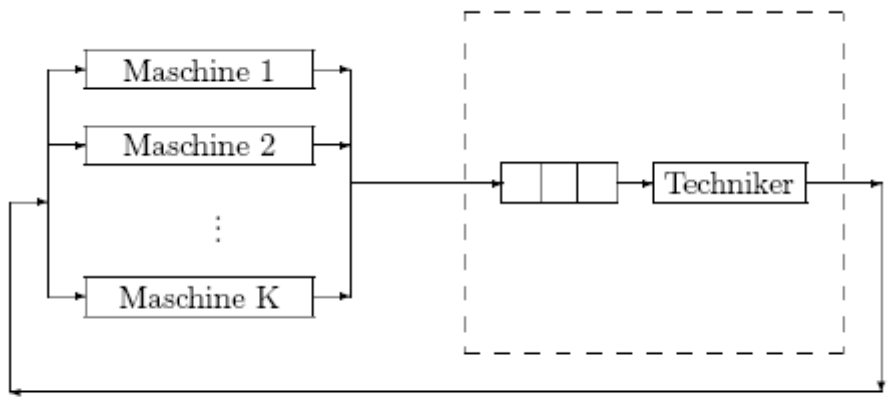

Abbildung Das M/M/1/M//M -System

Aus den Annahmen folgt: Wenn sich n Kunden im System befinden, sind $M - n$ Kunden außerhalb des Systems. Die Ankunftsrate beträgt jeden von ihnen λ. Die gesamte Ankunftsrate beim System ist in diesem Zustand also $(M - n)\,\lambda$.

Das System kann wieder als Geburts- und Todesprozess modelliert werden, indem man setzt

$$\lambda_{n\,n+1} = \lambda_n = \begin{cases} (M-n)\,\lambda & \text{falls } 0 \le n < M \\ 0 & \text{falls } n \ge M \end{cases}$$

$$\lambda_{n\,n-1} = \mu_n = \mu \quad \text{falls } 1 \le n \le M$$

Sei N –die Anzahl der Kunden im System. Wegen der Endlichkeit des Zustandsraumes folgt die Existenz der stationären Wahrscheinlichkeiten. Aus ihrer allgemeinen Darstellung folgt:

$$p_n = p_0 \prod_{i=0}^{n-1} \frac{\lambda_i}{\mu_{i+1}} = p_0 \prod_{i=0}^{n-1} \frac{\lambda(M-i)}{\mu} = p_0 \left(\frac{\lambda}{\mu}\right)^n \frac{M!}{(M-n)!}$$

p_0 erhält man aus der **Normierunsbedingung**:

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

Bemerkung: p_0 steht in enger Beziehung zur Erlangschen B-Funktion (Erlang B-Formula):

$$p_0 = \frac{1}{\sum_{n=0}^{M} \frac{M!}{(M-n)!} \left(\frac{\lambda}{\mu}\right)^n} = \frac{\frac{(\mu/\lambda)^M}{M!}}{\sum_{n=0}^{M} \frac{1}{(M-n)!} \left(\frac{\mu}{\lambda}\right)^{M-n}} = \frac{\frac{(\mu/\lambda)^M}{M!}}{\sum_{n=0}^{M} \frac{1}{n!} \left(\frac{\mu}{\lambda}\right)^{M-n}} = B(M,\, \mu/\lambda)$$

Für die numerische Berechnung von p_0 lässt sich eine einfache Rekursionsformel entwickeln:

$$p_0 = :G(M)^{-1}$$

$$G(M+1) = \sum_{n=0}^{M+1} \frac{(M+1)!}{(M-n+1)!} (\lambda/\mu)^n =$$

$$= 1 + (M+1)\,\frac{\lambda}{\mu} \sum_{n=1}^{M+1} \frac{M!}{(M-(n-1))!} (\lambda/\mu)^{n-1} =$$

$$= 1 + (M+1)\,\frac{\lambda}{\mu}\, G(M)$$

Die Rekursion startet mit $G(0) = 1$.

Die Auslastung ρ des Bedieners ist gleich der Wahrscheinlichkeit, dass er tätig ist:

$$\rho = 1 - p_0$$

Im stationären Zustand muss die Abgangsrate beim System gleich der effektiven Zugangsrate α sein. Sofern der Bediener tätig ist, gehen pro Zeiteinheit μ Kunden ab. Die Wahrscheinlichkeit, dass der Bediener tätig ist, ist aber gleich ρ. Also

$$\alpha = \rho \cdot \mu$$

Die mittlere Wartezeit $W = \mathbb{E}[Q]$ erhält man durch folgende Überlegung:

Die Zeit Z für einen Zyklus setzt sich für jeden Kunden zusammen aus seiner Zeit außerhalb des Systems (seiner "Denkzeit" oder Operationszeit) O, seiner Wartezeit Q_q und seiner Bedienzeit B:

$$Z = O + Q_q + B$$

Für die Erwartungswerte folgt:

$$\mathbb{E}[Z] = \mathbb{E}[O] + \mathbb{E}\left[Q_q\right] + \mathbb{E}[B]$$

Der Kunde betritt also das System pro Zeiteinheit $1/\mathbb{E}[Z]$ mal. Die Zugangsrate α enthält die Zugänge aller Kunden; also

$$\alpha = \frac{M}{\mathbb{E}[O] + \mathbb{E}\left[Q_q\right] + \mathbb{E}[B]}$$

Daraus folgt

$$W_q = \mathbb{E}\left[Q_q\right] = \frac{M}{\alpha} - \mathbb{E}[O] - \mathbb{E}[B]$$

Für die mittlere Verweilzeit $W = \mathbb{E}[Q]$ (Antwortzeit) folgt daraus:

$$W = \mathbb{E}[Q] = \mathbb{E}\left[Q_q\right] + \mathbb{E}[B] = \frac{M}{\alpha} - \mathbb{E}[O] = \frac{M}{\rho\mu} - \frac{1}{\lambda} = \frac{M}{\mu(1-p_0)} - \frac{1}{\lambda}$$

Diese Formel ist auch als interaktives Antwortzeitgesetz bekannt; es spielt bei der Medellierung von Terminalsystemen eine wichtige Rolle.
Die mittlere Anzahl $L_q = \mathbb{E}\left[N_q\right]$ der Kunden in der Warteschlange und $L = \mathbb{E}[N]$ im System erhält man aus der Littleschen Formel.
Die Wahrscheinlichkeit, dass ein bestimmter Kunde sich im System befindet (dass eine bestimmte Maschine defekt) ist, errechnet man aus dem Anteil der Zeit eines Zyklus, in der sich der Kunde im System befindet.

Die wichtigsten Formeln sind in der folgenden Tabelle zusammengestellt.

Zusammenfassung (Leistungsmerkmale bei M/M/1/M//M):

- mittlere Denkzeit, Operationszeit

$$\mathbb{E}[O] = \frac{1}{\lambda}$$

- mittlere Reparaturzeit, Bedienzeit

$$\mathbb{E}[B] = \frac{1}{\mu}$$

- $$p_n = \mathbb{P}[N = n] = \frac{M!}{(M-n)!}\left(\frac{\lambda}{\mu}\right)^n p_0 \text{ wobei}$$

$$\frac{1}{p_0} = \sum_{n=0}^{M} \frac{M!}{(M-n)!}\left(\frac{\lambda}{\mu}\right)^n = B\left(M, \frac{\mu}{\lambda}\right)$$

- Auslastung

$$\rho = 1 - p_0$$

- Durchsatz

$$\alpha = \rho\mu$$

- mittlere Anzahl von Kunden in der Warteschlange:

$$L_q = \mathbb{E}\left[N_q\right] = \alpha\,\mathbb{E}\left[Q_q\right]$$

- mittlere Wartezeit

$$W_q = \mathbb{E}\left[Q_q\right] = \frac{M}{\alpha} - \frac{1}{\lambda} - \frac{1}{\mu}$$

- mittlere Systemzeit

$$W = \mathbb{E}[Q] = W_q + \mathbb{E}[B] = \frac{M}{\alpha} - \frac{1}{\lambda}$$

- mittlere Anzahl von Kunden im System:

$$L = \mathbb{E}[N] = \alpha\,\mathbb{E}[Q]$$

-

$$\mathbb{P}[\text{Maschine } n \text{ defekt}] = \frac{\mathbb{E}[Q]}{\mathbb{E}[Q]+\mathbb{E}[O]}$$

Das M/M/1/M//M - System wurde in der Informatik zur Modellierung vieler Systeme benutzt. Wir wollen ein Modell für ein interaktives Rechensystem betrachten. Die Maschinen aus dem Repairman-Modell sind hier die M-Terminals, an denen Benutzer sitzen und nach einer exponentialverelten Denkzeit mit dem Paramete λ Aufträge in das Rechensystem eingeben. Jeder Benutzer ist entweder "denkend", hat einen Auftrag in der Warteschlange im Rechensystem oder hat einen Auftrag gerade in der Bedienung.
Das Rechensystem besitze nur einen Rechnerkern. Die Bedienzeiten seien exponentialverteilt mit der Rate μ.
Der Rechnerkern bearbeitet einen Auftrag nach dem anderen jeweils völlig bis zum Ende, er betreibt also kein Time Sharing. Den Informatiker interessieren besonders der Durchsatz α und die mittlere Antwortzeit W, und wie diese von der Anzahl M der Terminals abhängen.

Für dieses System ist das M/M/1/M//M-System offenbar ein geeignetes Modell. Aus der Tabelle oder unseren vorgehenden Überlegungen haben wir

$$\alpha(M) = \mu \cdot \rho = \mu \cdot (1 - p_0(M)) \text{ und}$$
$$W(M) = \frac{M}{\mu \cdot (1 - p_0(M))} - \frac{1}{\lambda}$$

wobei

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

Aus den Vorgaben lassen sich diese Funktionen berechnen. Über den Verlauf dieser Funktionen können wir folgende grobe Abschätzungen treffen:

Für den Fall, dass M=1 ist, hat der einzige vorhandene Auftrag im System keine Konkurrenz; seine Antwortzeit ist also seine Bedienzeit, d.h.

$$W(1) = \frac{M}{\mu(1-p_0(M))} - \frac{1}{\lambda} = \frac{\lambda+\mu}{\lambda\mu} - \frac{1}{\lambda} = \frac{1}{\mu} = \mathbb{E}[B]$$

Dieser Wert kann auf keinen Fall auch M>1 unterschritten werden. Die Funktion $W(M)$ muss also stets oberhalb der zur M-Achse parallelen Geraden mit dem W-Wert $\frac{1}{\mu}$ liegen.

Lassen wir nun $M \to \infty$ gehen, so wird die Zeit, die der Bediener untätig sein kann, gegen 0 gehen, d.h.

$p_0(M) \to 0$. Für diesen Fall nähert sich $W(M)$ also der folgenden Geraden:

$$\text{WG}(M) = \frac{M}{\mu \cdot (1 - p_0(M))} - \frac{1}{\lambda} \to \frac{M}{\mu} - \frac{1}{\lambda}$$

Da $(1 - p_0(M)) < 1$, muss $W(M)$ auch oberhalb dieser Geraden liegen. Es ergibt sich also das in der folgenden Skizze angedeutete Bild.

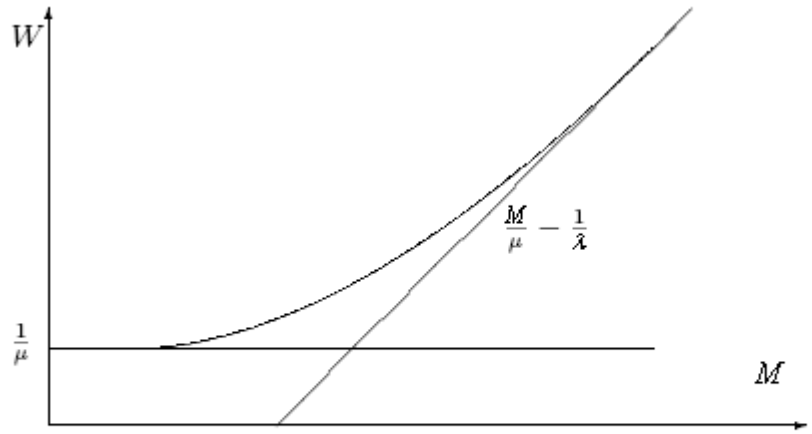

Abbildung Grenzgeraden der Antwortzeit im M/M/1/M//M -System

Analoge Überlegungen kann man nun bezüglich des Durchsatzes $\alpha(M)$ anstellen:
Bei M=1 berechnet man den Durchsatz zu

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

gäbe es keine gegenseitige Behinderung der Aufträge im System, wäre der Durchsatz bei M Terminals das M-fache. Da es aber mit wachsendem M zunehmend zu Behinderungen kommt, verläuft die Funktion $\alpha(M)$ unterhalb der Geraden

$$\mathrm{GL}(M) = M \cdot \frac{\lambda\mu}{\lambda+\mu}$$

Lassen wir M gegen unendlich wachsen, so wird der Bediener fast vollständig ausgelastet, d.h. es werden μ Aufträge pro Zeiteinheit durchgesetzt. Mehr kann der Bediener auf keinen Fall schaffen. Dieser Wert bildet also für den Durchsatz eine obere Schranke. Es ergibt sich also folgendes Bild.

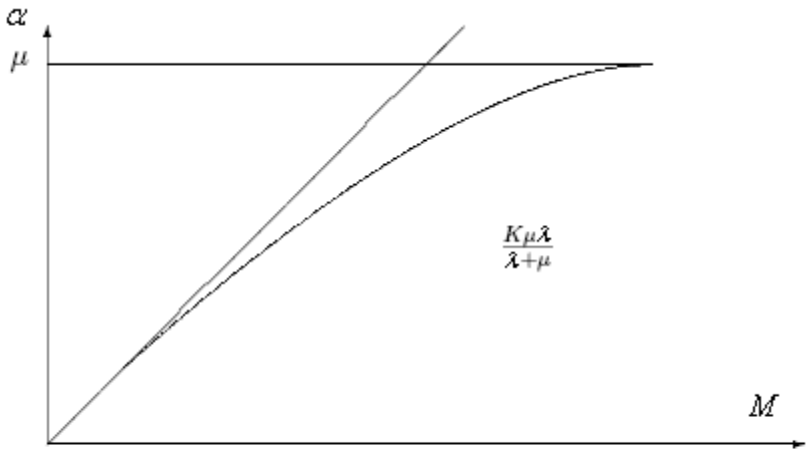

Abbildung Grenzgeraden des Durchsatzes im M/M/1/M//M -System

Die Skizzen zeigen, dass man anhand der Grenzgeraden bereits einige grobe Abschätzungen über das Verhalten des Systems bei Veränderungen der Anzahl der Terminals machen kann.

M/M/m/M//M Modell (Maschine Repairman Model)

Im vorhergehenden Abschnitt haben wir das Maschine Repairman Modell mit einem Techniker kennen gelernt, das zur Modellierung eines interaktiven Rechensystems mit einem Monoprozessorsysteme benutzt werden kann.Ein entsprechendes Modell kann man auch für Multiprozessorsysteme betrachten. Es seien also m Bediener im System vorhanden, wobei natürlich m≤M gelte. Die folgende Skizze beschreibt die Struktur des Systems.

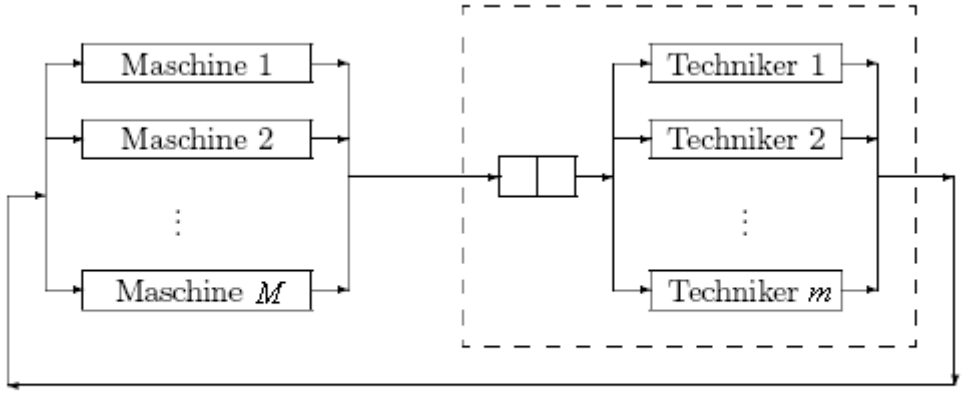

Abbildung Das *M/M/m/M//M* -System

Reparatur

Ein einfaches mathematisches Modell mit *m* Servern und Systemkapazität M (M>*m*) ist ein Geburts- und Todesprozess mit den Raten

$$\lambda_{n\,n+1} = \lambda_n = \begin{cases} (M-n)\,\lambda & \text{falls } 0 \le n < M \\ 0 & \text{falls } n \ge M \end{cases}$$

$$\lambda_{n\,n-1} = \mu_n = \begin{cases} n\,\mu & \text{falls } 1 \le n \le m \\ m\,\mu & \text{falls } n \ge m \end{cases}$$

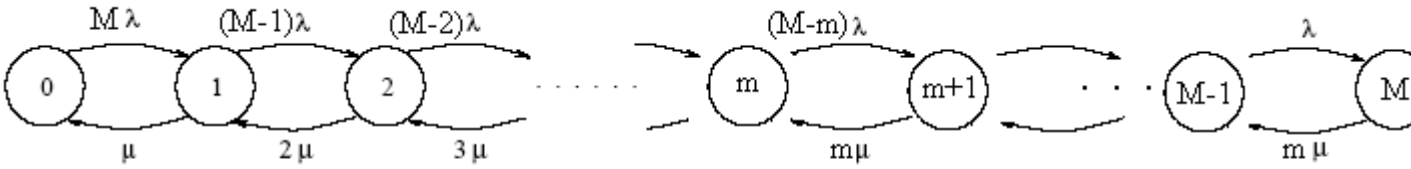

Figure M/M/m mit endlicher Population

Dieser Prozess hat den endlichen Zustandsraum I={0,1,...,M}. Die Gleichgewichtsverteilung erfüllt

$$p_n = \begin{cases} \binom{M}{n} \frac{\lambda^n}{\mu^n}\, p_0 & \text{falls } 1 \le n \le m \\ \binom{M}{n} \frac{n!}{m!\, m^{n-m}} \frac{\lambda^n}{\mu^n}\, p_0 & \text{falls } m \le n \le M \end{cases}$$

mit

$$p_0 = \sum_{n=0}^{m} \binom{M}{n} \frac{\lambda^n}{\mu^n} + \sum_{n=m+1}^{M} \binom{M}{n} \frac{n!}{m!\, m^{n-m}} \frac{\lambda^n}{\mu^n}.$$

Reparatur und Reserve

In einer Firma werden stets *M* Mashinen benötigt, die unabhängig voneinander arbeiten. Ist eine davon defekt, so wird dafür wenn möglich sofort eine Reservemashine eingesetzt, von denen es *Y* gibt. Jede Mashine hat auch nach einer Reparatur eine negativ exponential verteilte Lebensdauer mit Erwartungswert $\frac{1}{\lambda}$. Defekte Mashinen kommen in die Reparatur, wo an maximal $m \le Y$ Mashinen gleichzeitig Reparaturen durchgeführt werden können. Die Reparatur einer Mashine dauert eine negativ ($\mathbb{f}(t) = \mu e^{-\mu t}$, $t \ge 0$) exponential verteilte Zeitspanne mit Mittelwert $\frac{1}{\mu}$. Die Größen M, λ, μ sind vorgegeben. Ziel ist es, dass der Prozess nicht länger als 5% der Zeit mit weniger als M Mashinen auskommen muss.
Wir bezeichnen mit *n* den Zustand des Systems, bei dem *n* Mashinen defekt sind. Dann hat der Prozess den endlichen Zustandsraum I={0,1,...,M+Y} und es liegt ein Geburts- und Todesprozess vor mit den Raten

$$\lambda_{n\,n+1} = \lambda_n = \begin{cases} M\,\lambda & \text{falls } 0 \le n \le Y \\ (M + Y - n)\,\lambda & \text{falls } Y \le n \le M + Y \\ 0 & \text{falls } n \ge M + Y \end{cases}$$

$$\lambda_{n\,n-1} = \mu_n = \begin{cases} n\,\mu & \text{falls } 1 \le n \le m \\ m\,\mu & \text{falls } n \ge m \end{cases}$$

Die Gleichgewichtsverteilung erfüllt

$$p_n = \begin{cases} \frac{M^n}{n!}\,\frac{\lambda^n}{\mu^n}\,p_0 & \text{falls } 0 \le n \le m \\ \frac{M^n}{m^{n-m}\,m!}\,\frac{\lambda^n}{\mu^n}\,p_0 & \text{falls } m \le n \le Y \\ \frac{M^Y\,M!}{(M-n-Y)!\,m!\,m^{n-m}}\,\frac{\lambda^n}{\mu^n}\,p_0 & \text{falls } Y \le n \le M + Y \end{cases}$$

wobei man $p_0$durch Normierung $\sum_{n=0}^{M+Y} p_n = 1$erhält.
Will man erreichen, daß höchstens 5% der Zeit weniger als *M* Mashinen intakt sind, so muss

$$\sum_{n=Y+1}^{Y+M} p_n < 0.05$$

gelten.
Ob man mehr Reservemashinen oder mehr Reparaturkapazität einsetzt, hängt von konkreten Kostenüberlegung ab.
In einem anderen Fall können aber Reparatur und Reservemaschinen so teuer sein, dass man besser mehr Systemausfälle einkalkuliert.

Die wichtigsten Formeln für dieses Modell sind in der folgenden Tabelle zusammengestellt.

Zusammenfassung (Leistungsmerkmale bei M/M/m/M//M):

- mittlere Denkzeit, Operationszeit

$$\mathbb{E}[O] = \frac{1}{\lambda}$$

- mittlere Reparaturzeit, Bedienzeit

$$\mathbb{E}[B] = \frac{1}{\mu}$$

- $$p_n = \mathbb{P}[N = n] = \binom{M}{n}\left(\frac{\lambda}{\mu}\right)^n p_0 \qquad \text{für } 1 \le n \le m$$

$$p_n = \mathbb{P}[N = n] = \frac{n!}{m!\,m^{n-m}}\binom{M}{n}\left(\frac{\lambda}{\mu}\right)^n p_0 \text{ für } m \le n \le M$$

wobei

$$\frac{1}{p_0} = 1 + \sum_{n=1}^{M} \frac{p_n}{p_0}$$

- Auslastung

$$\rho = \frac{\alpha}{m\,\mu}$$

- Durchsatz

$$\alpha = \frac{M}{\mathbb{E}[O] + \mathbb{E}\left[Q_q\right] + \mathbb{E}[B]}$$

- mittlere Anzahl von Kunden in der Warteschlange:

$$L_q = \mathbb{E}\left[N_q\right] = \sum_{n=m+1}^{M} (n - m)\; p_n$$

- mittlere Wartezeit

$$W_q = \mathbb{E}\left[Q_q\right] = \mathbb{E}\left[N_q\right] \frac{\mathbb{E}[O] + \mathbb{E}[B]}{M - \mathbb{E}\left[N_q\right]}$$

- mittlere Systemzeit

$$W = \mathbb{E}[Q] = W_q + \mathbb{E}[B]$$

- mittlere Anzahl von Kunden im System:

$$L = \mathbb{E}[N] = \alpha\, \mathbb{E}[Q]$$

-

$$\mathbb{P}[\text{Maschine } n \text{ defekt}] = \frac{\mathbb{E}[Q]}{\mathbb{E}[Q] + \mathbb{E}[O]}$$