RECHENZENTRUM TH MÜNCHEN
ARBEITSGRUPPE BETRIEBSSYSTEME

INTERNSCHRIFT Nr. 14

THEMA:

Theorie der Warteschlangen - kurze Einführung

VERFASSER:

wolf

DATUM:

21.10.1969

FORM DER ABFASSUNG

- ENTWURF
- AUSARBEITUNG
- [X] ENDFORM

SACHLICHE VERBINDLICHKEIT

- [X] ALLGEMEINE INFORMATION DISKUSSIONSGRUNDLAGE
- ERARBEITETER VORSCHLAG
- VERBINDLICHE MITTEILUNG
- VERALTET

ÄNDERUNGSZUSTAND

BEZUG AUF BISHERIGE INTERNSCHRIFTEN

Vorkenntnisse aus:

Erweiterung von:

Ersatz für:

BEZUG AUF KÜNFTIGE INTERNSCHRIFTEN

Vorkenntnisse zu:

Erweiterung in:

Ersetzt durch:

ANDERWEITIGE LITERATUR

siehe Literaturhinweise am Ende dieser Schrift

14/211069/Wo

5. Weitere Warteschlangenmodelle

5.1 Modelle mit mehreren (parallelen) Schaltern

5.1.1 Unendlich viele Schalter
5.1.2 Wartesystem mit m Schaltern
5.1.3 Verlustsystem mit m Schaltern

5.2 "Machine Minding" (endliches "Kundenreservoir")

6. Kopplung von Warteschlangenmodellen

6.1 Allgemeines

6.2 Warteschlangenmodelle mit Feedback

1. Überblick

Nachfolgend werden die Grundbegriffe und Fragestellungen der Theorie der Warteschlangen dargelegt, einige Modelle beschrieben und mögliche Anwendungen skizziert. Damit wird vor allem angestrebt, die Formulierung und Diskussion praktischer Probleme, die sich in dieses Gebiet einordnen lassen, zu erleichtern und zur Klärung dieser Probleme beizutragen. Da die Theorie nur für wenige, idealisierte Systeme brauchbare Lösungen liefert, sind oft nur grobe Abschätzungen möglich. Solche Schätzwerte, die im allgemeinen zu pessimistisch sind, können jedoch als obere Grenze bei der Planung und zur Kontrolle weiterreichender Methoden, vor allem zum Test von Simulationsprogrammen dienen. Wesentliche Quelle für die folgende Darstellung ist das IBM-Manual "Analysis of Some Queuing Models zu Real-Time Systems [1]. (Ziffern in eckigen Klammern verweisen auf das Literaturverzeichnis.)

2. Grundbegriffe der Warteschlangentheorie

Eine Warteschlange bildet sich, wenn "Kunden" vor einem oder mehreren "Schaltern" eintreffen, ehe die Abfertigung vorher angekommener Kunden abgeschlossen ist. Es hat sich eingebürgert, von "Kunden" vor einem "Schalter" oder einer "Bedienungsstation" (server) zu sprechen; dabei können tatsächlich etwa Kunden vor einem Bankschalter gemeint sein, aber auch z.B. rechenwillige Programme, die auf CPU-Zeit warten, Transportaufträge für den Trommelvermittler usw.

Ein Warteschlangenmodell läßt sich im wesentlichen durch den Ankunfts- und den Abfertigungsprozeß (und die Anzahl der Schalter) beschreiben. Daneben ist noch von Bedeutung, ob die Warteschlange Beschränkungen unterworfen ist und in welcher Reihenfolge sie abgearbeitet wird.

./. 2

2.1 Ankunftsprozeß

Im allgemeinen wird angenommen, daß ein ankommender Kunde zufällig und unabhängig von anderen Kunden aus einer Menge potentieller Kunden "ausgewählt wird". Wenn nicht ausdrücklich anders vermerkt, wird unterstellt, daß die Menge der möglichen Kunden hinreichend groß (theoretisch unendlich) ist, so daß die Anzahl der bereits eingetroffenen Kunden im Ankunftsprozeß nicht berücksichtigt werden muß. Ein einfaches Beispiel sind Ankünfte zu fest vorgegebenen Zeitpunkten, etwa in konstantem zeitlichem Abstand, jedoch ist ein solches Modell selten realistisch (z.B. bei manchen Systemen von Fließbändern) und außerdem relativ schlecht analytisch zu behandeln. Man wird statt dessen eine "zufällige Verteilung" der Ankünfte in der Zeit zugrundelegen. Es gibt zwei Möglichkeiten, den Ankunftsprozeß mittels einer Wahrscheinlichkeitsverteilung zu beschreiben.

2.1.1 Verteilung der Anzahl der Ankünfte

Man betrachtet die Anzahl der Ankünfte zu einem Zeitintervall, etwa t Sekunden. Es wird die Wahrscheinlichkeit $P_k(t)$ angegeben, daß genau k Kunden in t Sekunden eintreffen (diskrete Wahrscheinlichkeitsverteilung). Die meistgebrauchte Verteilung dieses Typs ist die Poissonverteilung (mit dem Parameter λt). Dabei werden die Wahrscheinlichkeiten $P_k(t)$ gegeben durch

$$P_k(t) = \frac{(\lambda t)^k}{k!} e^{-\lambda t}, \quad \lambda > 0, \; k = 0,1,2, \ldots$$

Der Erwartungswert (Mittelwert) der Anzahl der Ankünfte im Intervall der Länge t ist λt; λ ist also die mittlere Anzahl der Ankünfte pro Zeiteinheit.

./. 3

2.1.2 Verteilung der Zwischenankunftszeiten

Eine andere Möglichkeit, den Ankunftsprozeß zu beschreiben, besteht darin, die Zeiten zwischen dem Eintreffen aufeinanderfolgender Kunden zu betrachten. Diese Zwischenankunftszeit wird als Zufallsvariable T mit einer Verteilungsfunktion $F(t)$ = Wahrscheinlichkeit $(T<t) = W(T<t)$ vorgegeben. Meistgebrauchte Verteilung ist die (negative) Exponentialverteilung (stetige Verteilung) mit der Verteilungsfunktion

$$F(t) = w(T<t) = \begin{cases} 0 & \text{für } t < 0 \\ 1-e^{-\lambda t} & \text{für } t \geqq 0 \end{cases}$$

Der Erwartungswert $E(T)$ der Zwischenankunftszeit T ist $\frac{1}{\lambda}$. Durchschnittlich kommen also λ Kunden pro Zeiteinheit, weshalb λ oft als Ankunftsrate bezeichnet wird. (Für den Fall einer beliebigen Verteilungsfunktion $F(t)$ wird die Ankunftsrate durch $\frac{1}{E(T)}$ gegeben. Wie leicht nachzurechnen sind die Ankünfte genau dann Poisson-verteilt, wenn die Zwischenankunftszeiten exponentialverteilt sind. Beide Verteilungen beschreiben also denselben Ankunftsprozeß, der als Poissonankunft bzw. -prozeß, Zufalls- oder Random-Ankunft bezeichnet wird.

2.2 Abfertigungsprozeß

Die wartenden Kunden benötigen "Dienstleistungen" irgendeiner Art und können an einem oder mehreren Schaltern abgefertigt werden. Wenn nicht anders vermerkt, wird angenommen, daß der Schalter während des ganzen betrachteten Zeitraums geöffnet ist, Service also jederzeit in Anspruch genommen werden kann. Der Angestellte am Schalter ist also frei nur während der Zeit, in der kein Kunde im System ist. Außerdem wird meist unterstellt, daß unmittelbar nach der Abfertigung eines Kunden mit der Bedienung des nächsten (sofern vorhanden) begonnen wird. Es existieren jedoch wichtige Modelle, in denen der "Schalter"

./. 4

nur zeitweise geöffnet ist (z.B. wegen Ausfall oder Wartung) oder in denen zwischen der Abfertigung von Kunden Zeit für "Verwaltungsaufgaben" (overhead) verbraucht wird.

Die Abfertigung am Schalter wird dadurch charakterisiert, wieviel Zeit pro Kunde benötigt wird. Falls diese von Kunde zu Kunde schwankt, liegt es nahe, sie durch eine Wahrscheinlichkeitsverteilung zu beschreiben. Die Abfertigungszeiten (Servicezeiten) werden als unabhängige, identisch verteilte Zufallsgrößen mit der Verteilungsfunktion H(t) betrachtet. Von großem theoretischem Interesse als Verteilungsfunktion für die Servicezeit S ist die (bereits beim Ankunftsprozeß betrachtete) negative Exponentialverteilung

$$H(t) = w(S\ t) = \begin{cases} 0 & \text{für } t<0 \\ 1-e^{-\mu t} & \text{für } t\geqq 0 \end{cases}$$

Hier ist $\frac{1}{\mu}$ der Erwartungswert der Servicezeit, μ die Abfertigungsrate. (Für beliebige Verteilungsfunktionen wird die Abfertigungsrate durch $\frac{1}{E(S)}$ gegeben.)

2.3 Warteschlange

Wenn Ankunft und Abfertigung zufälligen Schwankungen unterliegen, kann es zu Stauungen kommen. Man kann für eine auftretende Warteschlange beliebige Länge zulassen (unendlicher Warteraum) oder eine obere Grenze n für die Warteschlange vorgeben (endlicher Warteraum), wobei im letzteren Fall noch vereinbart werden muß, was mit den Kunden geschieht, die auf einen vollen Warteraum treffen. (Häufige Annahme: Kunde "geht verloren".)

Von Interesse ist weiterhin, in welcher Reihenfolge die Kunden bedient werden (queue discipline), ob in der Reihenfolge des Eintreffens, wie meist vorausgesetzt oder ob etwa der jeweils zuletzt gekommene Kunde zuerst bedient wird. Eine andere Möglichkeit besteht darin, den Kunden Prioritäten zuzuordnen und Kunden hoher Priorität vorrangig abzufertigen.

./. 5

3. Charakteristische Größen eines Warteschlangenmodells

Um ein Warteschlangenmodell zu beschreiben, braucht man ein Maß für die auftretende Belastung und für die Stauungen und Wartezeiten, die zu erwarten sind.

3.1 Verkehrsintensität

Das einfachste Maß für die Belastung eines Systems ist das

$$\text{Verhältnis} \quad \frac{\text{Ankunftsrate}}{\text{Abfertigungsrate}} = \frac{\text{mittlere Servicezeit}}{\text{mittlere Zwischenankunftszeit}}$$

Dieses Verhältnis wird als Verkehrsintensität ϱ bezeichnet (Einheit: Erlang):

$$\varrho = \frac{\frac{1}{E(T)}}{\frac{1}{E(S)}} = \frac{E(S)}{E(T)}$$

Für den Fall, daß T und S exponentialverteilt sind, ist $\varrho = \frac{\lambda}{\mu}$. Ein System mit 1 Schalter und unbegrenztem Warteraum arbeitet über längere Zeit nur dann zufriedenstellend, wenn die Verkehrsintensität weniger als 1 Erlang beträgt (der Fall $\varrho = 1$ bedarf besonderer Diskussion, ist jedoch nur von theoretischem Interesse); während für ein System mit m Schaltern eine Verkehrsintensität von weniger als m Erlang zugelassen werden kann. ϱ gibt für ein 1-Schalter-System außerdem den Bruchteil der Zeit an, während dem der Schalter belegt ist (Auslastung des Schalters), $1-\varrho$ dementsprechend die Leerzeit (idle time).

3.2 Länge der Warteschlange, Wartezeit

Von praktischem Interesse sind die folgenden Größen:

a) Erwartungswert und Verteilung der Anzahl der Kunden in der Warteschlange (ausschließlich eines evtl. gerade in Abfertigung begriffenen) bzw. im System (einschließlich des evtl. gerade in Abfertigung begriffenen), hier mit N_w bzw. N_q bezeichnet.

./. 6

b) Erwartungswert und Verteilung der Verweilzeit des Kunden in der Warteschlange (ohne Service) bzw. im System (einschließlich Service), hier mit T_w bzw. T_q bezeichnet.

Warnung: Bei Benützung von Formeln vergewissere man sich, ob der gerade in Abfertigung begriffene Kunde bzw. dessen Servicezeit mitgerechnet wird oder nicht. Es gilt im allgemeinen nicht $N_q = N_w + 1$, sondern $N_q = N_w + \varrho$.

3.3 Zeitabhängige Lösungen

Die Größen N_q, N_w, T_q und T_w hängen genau genommen nicht nur vom betrachteten Modell, sondern auch von den Bedingungen beim Start des Systems ab. In den meisten Untersuchungen wird jedoch angenommen, daß sich das System stabil verhält (bei 1-Schalter-Systemen mit unendlichem Warteraum also $\varrho < 1$) und daß bereits genügend Zeit vergangen ist, so daß der Einfluß der Anfangswerte vernachlässigt werden kann. Solche stationären Lösungen lassen sich für eine Reihe von Fällen relativ einfach ermitteln. In manchen Fällen interessiert man sich jedoch speziell für die Zeitabhängigkeit, z.B. für das Anwachsen der Warteschlange zu Beginn einer Periode hoher Verkehrsintensität (Stoßzeit) bzw. den Abbau eines Staues nach einer solchen Periode. Lösungen werden schon für einfache Modelle schwierig und unhandlich (vgl. Takács [4]), jedoch lassen sich in einigen Fällen Faustformeln finden, die grobe, aber leicht zu handhabende Abschätzungen liefern (vgl. Cox and Smith [2]).

4. Warteschlangenmodelle mit 1 Schalter

Mathematisch am einfachsten zu behandeln sind Warteschlangenmodelle mit 1 Bedienungsstation, wobei die Ankünfte Poisson-verteilt, die Zwischenankunftszeiten also negativ exponentialverteilt und die Abfertigungszeiten ebenfalls exponentialverteilt sind; da dies außerdem eine brauchbare Näherung für eine Reihe von Problemen darstellt, wird es in 4.1 ausführlich dargestellt. In 4.2 wird die Annahme exponentialverteilter Servicezeit fallen-

gelassen, während in 4.3 und 4.4. auf den Einfluß der Abfertigungsreihenfolge und auf Prioritäten eingegangen wird.

Zur Kennzeichnung eines Warteschlangenmodells ist eine einfache Abkürzung üblich: Jedes Modell wird durch eine Angabe der Form A/B/c gekennzeichnet. Dabei ist c die Anzahl der Schalter, während A und B die Verteilung der Zwischenankunftszeiten bzw. der Servicezeiten angeben.
Es bedeutet

M: negative Exponentialverteilung
GI: beliebige Wahrscheinlichkeitsverteilung (general independent)
E_k: (spezielle) Erlangverteilung mit Parameter k (E_1=M)
D: starres Schema (deterministic)

Im ganzen Abschnitt 4 wird vorausgesetzt, daß die Menge der potentiellen Kunden unendlich, der Warteraum unbegrenzt und der Schalter ständig geöffnet ist. Weiterhin wird in 4.1 und 4.2 angenommen, daß die Kunden in der Reihenfolge des Eintreffens bedient werden.

4.1 M/M/1 (Zwischenankunfts- und Servicezeit exponentialverteilt, 1 Schalter)

4.1.1 Ankünfte

Bei diesem "klassischen" Modell ist die Anzahl der Ankünfte in einem Zeitintervall der Länge t Poisson-verteilt mit dem Parameter λt. Die Wahrscheinlichkeit $P_k(t)$ von genau k Ankünften im Intervall t ist also gegeben durch

$$P_k(t) = \frac{(\lambda t)^k}{k!} e^{-\lambda t}; \quad k = 0,1,2, \ldots$$

./. 8

Erwartungswert und Streuung sind λt, im Mittel kommen pro Zeiteinheit λ Kunden. Die Poissonverteilung gibt in vielen Fällen eine hinreichend genaue Annäherung an die realen Verhältnisse (z.B. Fernsprechverkehr). Falls das System nicht überlastet wird ($\varrho < 1$) und da keine Kunden verlorengehen, verlassen im Durchschnitt auch λ Kunden pro Zeiteinheit das System. λ gibt also nicht nur die Ankunftsrate, sondern auch den "Durchsatz" des Systems. Darüber hinaus gilt speziell bei dem Modell M/M/1, daß (für $\varrho < 1$) auch die Anzahl der in einem Intervall der Länge t abgefertigten Kunden Poisson-verteilt ist; anders formuliert: Ein solcher Schalter läßt die Verteilung der zeitlichen Abstände aufeinanderfolgender Kunden unverändert.

Die zur Poissonverteilung gehörige Verteilung der Zwischenankunftszeit T ist die negative Exponentialverteilung, wie man leicht verifiziert, denn (für $t \geq o$) gilt für die Verteilungsfunktion F(t):

$F(t)=w(T<t)=w$ (mindestens 1 Ankunft im Intervall der Länge t) =
$=1-w$ (keine Ankunft im Intervall der Länge t) =
$1-P_o(t)=1-e^{-\lambda t}$

Es gilt also

$$F(t) = \begin{cases} 0 & \text{für } t<0 \\ 1-e^{-\lambda t} & \text{für } t\geq 0 \end{cases}$$

Diese Verteilung hat den Erwartungswert $E(T)=\frac{1}{\lambda}$ und die Streuung $D^2(T)=\frac{1}{\lambda^2}$.

Um in praktischen Fällen zu entscheiden, ob die Poissonankunft eine brauchbare Annäherung darstellt, ermittelt man einen Schätzwert für den Variationskoeffizienten C der Verteilungsfunktion der Zwischenankunftszeiten ($C^2 = \frac{D^2(T)}{[E(T)]^2}$)
Für die Exponentialverteilung ist $C^2=1$. Es gilt nun folgende Faustregel: Falls $0,7 < C^2 < 1,3$, so kann die Exponentialverteilung benutzt werden. Kleinere Werte von C^2 weisen in

./. 9

Richtung auf einen Ankunftsprozeß mit konstanten Zeitabständen (exakt: $C^2=o$), während größere Werte von C^2 auf eine stärkere Zusammenballung der Ankünfte hindeuten. Die Annahme daß Poissonankunft vorliegt, wird für Werte von $C^2 < 1$ zu pessimistische Angaben liefern, kann also benutzt werden, um sich gegen den schlimmstmöglichen Fall abzusichern, während für $C^2 > 1$ die auftretenden Stauungen unterschätzt werden.

4.1.2 Abfertigung

Es wird angenommen, daß die Servicezeit S negativ exponentialverteilt ist, also folgende Verteilungsfunktion hat:

$$H(t) = w(S<t) = \begin{cases} 0 & \text{für } t<0 \\ 1-e^{-\mu t} & \text{für } t\geq 0 \end{cases}$$

Erwartungswert der Servicezeit (mittlere Servicezeit) ist $\frac{1}{\mu}$, Streuung $\frac{1}{\mu^2}$. Pro Zeiteinheit werden also durchschnittlich μ Kunden abgefertigt. Die Annahme, daß negative Exponentialverteilung der Servicezeit vorliegt, ist nur in wenigen praktischen Fällen realistisch (z.B. bei der Verteilung der Dauer von Telefongesprächen). Im allgemeinen streuen die Servicezeiten nicht so stark, so daß die Exponentialverteilung auf Werte führt, die als obere Grenze der Belastung angesehen werden können.

Daß die Exponentialverteilung mathematisch leicht zu handhaben ist, beruht auf der Tatsache, daß sie "nicht altert", d.h. die "Restservicezeit" eines bereits in Bedienung befindlichen Kunden die gleiche Verteilung wie die Servicezeit hat, exakt:

$$w(S>t+t' \mid S>t) = \frac{w(S>t+t' \wedge S>t)}{w(S>t)} = w(S>t')$$

Für das Beispiel des Fernsprechverkehrs würde das bedeuten: Wenn ein Gespräch bereits eine Stunde gedauert hat, so ist die Wahrscheinlichkeit, daß es in den nächsten 5 Minuten beendet wird, genauso groß, als hätte es gerade angefangen.

./. 10

4.1.3 Anzahl der Kunden im System bzw. in der Warteschlange

Man interessiert sich für die Wahrscheinlichkeitsverteilung der Anzahl N_q der Kunden im System, also für die Größen $p_k=w(N_q=k)$ und für den Erwartungswert $E(N_q)$ und die Streuung $D^2(N_q)$. Die p_k sind davon abhängig, unter welchen Bedingungen das System gestartet wurde und wieviel Zeit seitdem vergangen ist, exakt ist also $p_k(t)$ zu schreiben. Man begnügt sich jedoch im allgemeinen mit den stationären Werten dieser Größen, nimmt also an, daß seit dem "Einschalten" soviel Zeit vergangen ist, daß der Einfluß der Anfangsbedingungen vernachlässigt werden kann.

Der stochastische Prozeß, der durch das Modell M/M/1 beschrieben wird, läßt sich unter anderem auch dadurch charakterisieren, daß für ein kleines Zeitintervall Δt gilt:

w (1 Ankunft in Δt) $= \lambda \Delta t + o(\Delta t)$;
w (1 Abfertigung in Δt) $= \mu \Delta t + o(\Delta t)$;
w (mehr als 1 Ankunft in Δt) $= o(\Delta t)$;
w (mehr als 1 Abfertigung in Δt) $= o(\Delta t)$.

Auf diesen Eigenschaften beruht nun das Verfahren zur analytischen Behandlung des Modells M/M/1 (das Verfahren bleibt auch bei einer leichten Verallgemeinerung der angeführten Eigenschaften anwendbar; man vgl. dazu Abschnitt 5): Man betrachtet Zustandsübergänge (Zustand gekennzeichnet durch die Anzahl der Kunden im System) in einem kleinen Zeitintervall Δt, wobei nur Zustandsübergänge in unmittelbar benachbarte Zustände betrachtet werden brauchen, drückt also die $p_k(t+\Delta t)$, $k=0,1,2, \ldots$ durch $p_{k-1}(t)$, $p_k(t)$, $p_{k+1}(t)$ aus. Man erhält so ein (im allgemeinen unendliches) System von Differenzengleichungen und daraus durch Grenzübergang ein System von Differentialgleichungen. Durch Nullsetzen der Differentialquotienten (stationäre Lösung!) erhält man ein (i.a. unendliches) lineares Gleichungssystem für die $p_k(k=0,1,2, \ldots)$.

./. 11

Zusammen mit der Forderung $\sum_{k=0}^{\infty} p_k=1$ (Wahrscheinlichkeitsverteilung!) erhält man (unter der Voraussetzung $\varrho<1$):

$$p_k=(1-\varrho)\varrho^k \quad , \quad k=0,1,2, \ldots$$

Insbesondere ist $p_0=1-\varrho$, also ϱ die Wahrscheinlichkeit, daß der Schalter belegt ist.

Aus den p_k erhält man den Erwartungswert für die Anzahl N_q der Kunden im System (einschließlich des Kunden in der Bedienungsstation)

$$E(N_q)=\frac{\varrho}{1-\varrho}$$

und die Streuung

$$D^2(N_q)=\frac{\varrho}{(1-\varrho)^2}$$

Für $\varrho\to 1$ wird nicht nur der Erwartungswert der Schlangenlänge beliebig groß, sondern auch die Streuung wächst unbegrenzt an. Das erklärt, warum Systeme unter hoher Belastung (ϱ nahe an 1) sich sehr instabil verhalten.

Analog erhält man für die Anzahl N_w der Kunden in der Warteschlange (ausschließlich des Kunden in der Bedienungsstation)

$$E(N_w)=\frac{\varrho^2}{1-\varrho}$$

Wie leicht nachzurechnen, gilt:

$$E(N_q)=E(N_w)+\varrho$$

4.1.4 Verweilzeit im System bzw. in der Warteschlange

Neben der Anzahl der Kunden im System bzw. in der Warteschlange interessiert man sich auch für die Zeit T_q, die ein Kunde von seiner Ankunft an im System (einschließlich Abfertigung) bzw. die Zeit T_w, die er in der Warteschlange (bis zum Beginn seiner Abfertigung) verbringt. Auf dem Weg über die Laplacetransformation erhält man für die Verteilungsfunktion $Q(t)$ der Verweilzeit im System T_q:

./. 12

$$Q(t) = w(T_q < t) = \begin{cases} 0 & \text{für } t<0 \\ 1-e^{-\mu(1-\varrho)t} & \text{für } t\geq 0 \end{cases}$$

(Exponentialverteilung mit Parameter $\mu(1-\varrho)$)

Analog ergibt sich für die Verteilungsfunktion W(t) der Verweilzeit in der Warteschlange (reine Wartezeit) T_W:

$$W(t) = w(T_W < t) = \begin{cases} 0 & \text{für } t<0 \\ 1-\varrho e^{-\mu(1-\varrho)t} & \text{für } t\geq 0 \end{cases}$$

(W(t) hat bei t=o einen Sprung der Höhe $1-\varrho$; das ist gerade die Wahrscheinlichkeit, daß die Wartezeit =o, der Schalter also frei ist.)

Weiterhin erhält man

$$E(T_q) = \frac{1}{\mu}\,\frac{1}{1-\varrho}$$

$$E(T_W) = \frac{1}{\mu}\,\frac{\varrho}{1-\varrho}$$

4.2 M/GI/1 (Zwischenankunftszeit exponentialverteilt, Servicezeit beliebig, 1 Schalter)

4.2.1 Ankunft und Abfertigung

Da in vielen Fällen zwar die Annahme von Poisson-Ankunft, nicht jedoch die Annahme exponentialverteilter Servicezeit gerechtfertigt ist, sei nun vorausgesetzt, daß die Servicezeiten unabhängige, identisch verteilte Zufallsgrößen mit der Verteilungsfunktion H(t) sind (M/GI/1).

Das Modell M/GI/1 wurde erstmals von Chintschin und Pollaczek behandelt, weshalb die in 4.2.2 angegebenen Formeln gelegentlich nach diesen Autoren benannt werden. Für die analytische Behandlung bedient man sich des folgenden Kunstgriffs: Man

./. 13

betrachtet die Anzahl der Kunden im System (den "Zustand" des Systems) genau zu den Zeitpunkten, in denen ein Kunde das System verläßt. Man erhält auf diese Weise eine in den Prozeß "eingebettete Markoffsche Kette". (Analog wären für das Modell GI/M/1 die Ankunftszeitpunkte zu betrachten. Für das Modell GI/GI/1 ist kein solches Vorgehen möglich; es kann lediglich die Bestimmung der Verteilungsfunktion der Wartezeit als Integralgleichung formuliert werden.)

Es zeigt sich nun, daß die Erwartungswerte der charakteristischen Größen N_q, N_w, T_q, T_w nur von Momenten erster und zweiter Ordnung der Verteilungsfunktion der Servicezeit abhängen, d.h. es genügt, Erwartungswert $E(S)$ und Streuung $D^2(S)(=E(S^2)-[E(S)]^2)$ zu kennen. Bei praktischen Anwendungen wird man sich Schätzwerte für $E(S)$ und $D^2(S)$ verschaffen.

4.2.2 Anzahl der Kunden, Verweilzeit

Für den Fall einer allgemeinen Servicezeitverteilung lassen sich keine expliziten Formeln für die Verteilung der Anzahl der Kunden und der Verweilzeit angeben. Jedoch lassen sich die Erwartungswerte (und Streuungen) dieser Größen (auf dem Weg über die Laplace-Stieltjes-Transformation) gewinnen. Für die Anzahl N_q der Kunden im System erhält man:

$$E(N_q) = \varrho + \frac{\varrho^2}{2(1-\varrho)}\left[1+\frac{D^2(S)}{[E(S)]^2}\right]$$

Ähnlich ergibt sich für die Anzahl N_w der Kunden in der Warteschlange:

$$E(N_w) = \frac{\varrho^2}{2(1-\varrho)}\left[1+\frac{D^2(S)}{[E(S)]^2}\right]$$

Für die Verweilzeit T_q eines Kunden im System (einschließlich Service) erhält man:

$$E(T_q) = E(S) + \frac{\varrho E(S)}{2(1-\varrho)}\left[1+\frac{D^2(S)}{[E(S)]^2}\right]$$

Ähnlich ergibt sich für T_w (reine Wartezeit)

$$E(T_w) = \frac{\varrho E(S)}{2(1-\varrho)}\left[1+\frac{D^2(S)}{[E(S)]^2}\right]$$

./. 14

In diesen Formeln tritt die Größe $C_s^2 = \frac{D^2(s)}{[E(s)]^2}$, das Quadrat des Variationskoeffizienten der Servicezeitverteilung auf. Diese Größe charakterisiert also (bei gleichbleibender Verkehrsintensität) die zu erwartenden Stauungen bzw. Verzögerungen. Man sieht außerdem, daß bei gleicher mittlerer Servicezeit die Streuung der Servicezeiten das Verhalten des Systems bestimmt. Für den Fall konstanter Servicezeit ist $C_s^2=0$, für die Exponentialverteilung =1. Da in der Praxis C_s^2 meist einen Wert zwischen 0 und 1 besitzt, können die für die Grenzfälle $C_s^2=0$ bzw. =1 erhaltenen Werte für Anzahl der Kunden oder Verweilzeit als untere bzw. obere Schranke benutzt werden. Es existieren jedoch Ausnahmefälle, in denen für die Verteilung der Servicezeit $C_s^2>1$ gilt. Ein Beispiel für eine solche "hyperexponentielle" Verteilung ist eine Verteilung mit einer "zweigipfligen" Dichtefunktion mit großem Abstand zwischen den relativen Maxima; eine solche Verteilung liegt vor, wenn die Kunden in zwei Gruppen zerfallen, deren mittlere Servicezeiten sich erheblich mehr voneinander unterscheiden als die Streuung innerhalb der Gruppe. In diesem Fall würde die Näherung durch die Exponentialverteilung zu optimistische Aussagen liefern.

Es sei darauf hingewiesen, daß für $0<C_s^2<1$ die Möglichkeit besteht, als Approximation der Servicezeitverteilung eine (spezielle) Erlangverteilung mit Parameter k zu benutzen, wobei der Parameter k (k natürliche Zahl) gemäß dem Wert von C_s^2 bestimmt wird ($M/E_k/1$). Die Grenzfälle sind $E_1=M$ und E_∞: konstante Servicezeit.

4.3 Einfluß der Abfertigungsreihenfolge auf die Wartezeit

In 4.1 und 4.2 wurde angenommen, daß die Kunden in der Reihenfolge des Eintreffens bedient werden (first in, first out, kurz FIFO). Demgegenüber ist denkbar, daß jeweils der zuletzt angekommene Kunde zuerst bedient wird (last in, first out, kurz LIFO)

./. 15

oder daß der nächste zu bedienende Kunde "zufällig" aus den wartenden Kunden ausgewählt wird, wobei jeder die gleiche Wahrscheinlichkeit hat, gewählt zu werden (hier kurz mit RANDOM bezeichnet).

Es zeigt sich, daß in allen drei Fällen der Erwartungswert $E(T_W)$ der Wartezeit in der Schlange gleich ist, die Abfertigungsreihenfolge also keinen Einfluß auf die mittlere Wartezeit hat. Diese Aussage gilt darüberhinaus für jede Abfertigungsreihenfolge, die unabhängig von der Servicezeit ist. Dagegen unterscheiden sich die Streuungen $D^2(T_W)$ der Wartezeit; sie nehmen (für konstante und für exponentialverteilte Servicezeit) in der Reihenfolge FIFO, RANDOM, LIFO zu. LIFO führt also zu einem weniger stabilen Verhalten der Warteschlange als FIFO.

4.4 Warteschlangen mit Prioritäten

In vielen Fällen werden bestimmte Gruppen oder Typen von Kunden bevorzugt abgefertigt, sei es, daß es große "Kosten" verursacht, sie warten zu lassen, oder sei es, daß sie (vermutlich) nur eine kurze Servicezeit beanspruchen, so daß infolge ihrer vorrangigen Abfertigung die Warteschlange kürzer wird. Solche bevorzugten Kunden erhalten eine höhere Priorität, wobei im folgenden m Prioritätsklassen (m natürliche Zahl) vorausgesetzt werden, wobei 1 die höchste und m die niedrigste Priorität ist.

Man kann verschiedene Möglichkeiten diskutieren, wie sich die Ankunft eines Kunden der Priorität j auf im System befindliche Kunden mit Priorität $< j$ auswirkt. Zwei häufig betrachtete Fälle sind: Entweder wird die Bedienung eines Kunden unabhängig vom Eintreffen von Kunden höherer Priorität abgeschlossen und anschließend der Kunde mit der jeweils höchsten Priorität bedient (nicht unterbrechende Priorität, nonpreemptive priority), oder ein Kunde höherer Priorität verdrängt einen mit niedriger Priorität aus der Bedienungsstation und dessen Bedienung wird erst wieder aufgenommen, wenn kein Kunde höherer Priorität mehr im System ist (unterbrechende Priorität, preemptive priority, genauer: preemptive-resume priority).

./. 16

Die Ankünfte in den einzelnen Prioritätsklassen seien Poisson-verteilt mit den Ankunftsraten $\lambda_1, \lambda_2, \ldots \lambda_m$, damit sind auch die gesamten Ankünfte Poisson-verteilt mit der Ankunftsrate $\lambda = \sum_{i=1}^{m} \lambda_i$ Die Servicezeit S_j für Kunden der Prioritätsklasse j sei gemäß der Verteilungsfunktion $H_j(t)$ verteilt, die Servicezeit S für die Gesamtheit der Kunden also gemäß

$$H(t) = \frac{\lambda_1}{\lambda} H_1(t) + \frac{\lambda_2}{\lambda} H_2(t) + \ldots + \frac{\lambda_m}{\lambda} H_m(t)$$

Erwartungswert und 2. Moment von S_j sind gegeben durch

$$E(S_j) = \int_0^\infty t \, d H_j(t) \; ; \quad E(S_j^2) = \int_0^\infty t^2 d H_j(t);$$

Erwartungswert und 2. Moment von S ergeben sich zu

$$E(S) = \frac{\lambda_1}{\lambda} E(S_1) + \ldots + \frac{\lambda_m}{\lambda} E(S_m)$$

$$E(S^2) = \frac{\lambda_1}{\lambda} E(S_1^2) + \ldots + \frac{\lambda_m}{\lambda} E(S_m^2)$$

Der Anteil der Prioritätsklasse j an der Verkehrsintensität (und damit an der Belastung der Bedienungsstation) ist gegeben durch

$$\varrho_j = \lambda_j E(S_j),$$

der Anteil aller Prioritätsklassen $\leqq j$ also durch

$$u_j = \sum_{i=1}^{j} \varrho_i = \sum_{i=1}^{j} \lambda_i E(S_i) \qquad j = 1, \ldots m$$

Zusätzlich sei $u_0 = 0$ definiert.

.4.1 Nichtunterbrechende Priorität (nonpreemptive priority)

Für den Fall nichtunterbrechender Priorität ergibt sich (unter den obigen Annahmen) für die Wartezeit T_{wj} (bis zum Beginn der Abfertigung) eines Kunden der Prioritätsklasse j:

$$E(T_{wj}) = \frac{\lambda E(S^2)}{2(1-u_{j-1})(1-u_j)}, \qquad j = 1,2,\ldots m$$

./. 17

Hieraus erhält man für die Verweilzeit im System (einschließlich Service):

$$E(T_{qj}) = E(S_j) + \frac{\lambda E(S^2)}{2(1-u_{j-1})(1-u_j)}, \quad j = 1,2, \ldots m$$

Um einen einfachen Vergleich mit der Abfertigung ohne Prioritäten zu ermöglichen, sei speziell m=2, exponentialverteilte Servicezeit mit den Parametern μ_1 und μ_2, $\mu=\mu_1+\mu_2$, $\varrho = \frac{\lambda}{\mu}$ angenommen. Dann gilt:

$$E(T_{w1}) = \frac{1}{\mu} \frac{\varrho}{1-\varrho_1}$$

$$E(T_{w2}) = \frac{1}{\mu} \frac{\varrho}{(1-\varrho_1)(1-\varrho)}$$

Demgegenüber gilt ohne Prioritäten:

$$E(T_w) = \frac{1}{\mu} \frac{\varrho}{1-\varrho}$$

Man erhält daher folgende Beschleunigung bzw. Verzögerung

$$\frac{E(T_{w1})}{E(T_w)} = \frac{1-\varrho}{1-\varrho_1}; \quad \frac{E(T_{w2})}{E(T_w)} = \frac{1}{1-\varrho_1}; \quad \frac{E(T_{w1})}{E(T_{w2})} = 1-\varrho$$

4.2 Unterbrechende Priorität (preemptive-resume priortiy)

Im Fall unterbrechender Priorität kommen zur anfänglichen Wartezeit eines Kunden alle diejenigen Wartezeiten hinzu, die von Unterbrechungen durch Kunden höherer Priorität herrühren; es ist daher zweckmäßig, die Verweilzeit T_{qj} eines Kunden der Prioritätsklasse j im System (also einschließlich Service) zu betrachten:

$$E(T_{qj}) = \frac{1}{1-u_{j-1}} \left[E(S_j) + \frac{\sum_{i=1}^{j} \lambda_i E(S_i^2)}{2(1-u_j)} \right]$$

./. 18

Für Kunden der Priorität 1, also der höchsten Priorität, ergibt sich

$$E(T_{q1}) = E(S_1) + \frac{\lambda_1 E(S_1^2)}{2(1-\varrho_1)} = E(S_1) + \frac{\varrho_1 E(S_1)}{2(1-\varrho_1)} \left[1 + \frac{D^2(S_1)}{[E(S_1)]^2}\right]$$

Wie man durch Vergleich mit 4.2 feststellt und wie auf Grund des Modells zu erwarten war, bedeutet das, daß Kunden der höchsten Priorität so schnell abgefertigt werden, als wären sie allein im System.

Für weitere Untersuchungen über Warteschlangen und Prioritäten sei auf [8] und [9] verwiesen.

5. Weitere Warteschlangenmodelle

Wie bereits in 4.1.3 erwähnt, läßt sich das Verfahren zur Behandlung des Modells M/M/1 auch noch anwenden, wenn man die Eigenschaften des durch M/M/1 beschriebenen stochastischen Prozesses geeignet verallgemeinert. Es war bei M/M/1:

$$w\,(1 \text{ Ankunft in } \Delta t) = \lambda \Delta t + o(\Delta t).$$

$$w\,(1 \text{ Abfertigung in } \Delta t) = \mu \Delta t + o(\Delta t).$$

Nun läßt man zu, daß λ und μ von der Anzahl N_q der Kunden im System abhängen, also

$$w\,(1 \text{ Ankunft in } \Delta t \,|\, N_q=n) = \lambda_n \cdot \Delta t + o(\Delta t);$$

$$w\,(1 \text{ Abfertigung in } \Delta t \,|\, N_q=n) = \mu_n \cdot \Delta t + o(\Delta t).$$

Man leitet wiederum ein (i.a. unendliches) lineares Gleichungssystem für die (stationären) Wahrscheinlichkeiten $p_k = w\,(N_q=k)$ her. Damit eine stationäre Lösung existiert, muß die Reihe

$$S = 1 + \frac{\lambda_0}{\mu_1} + \frac{\lambda_0 \lambda_1}{\mu_1 \mu_2} + \ldots + \frac{\lambda_0 \cdots \lambda_{k-1}}{\mu_1 \cdots \mu_k} + \ldots$$

./. 19

konvergieren. (Der Fall $S = \infty$ wird im allgemeinen auf eine unbegrenzt wachsende Warteschlange hinweisen.) Damit gilt:

$$p_k = S^{-1} \frac{\lambda_0 \cdots \lambda_{k-1}}{\mu_1 \cdots \mu_k} \quad , \; k = 0,1,2, \ldots$$

(insbesondere $p_0 = S^{-1}$).

Unter gewissen einschränkenden Voraussetzungen lassen sich eine Reihe von Modellen, wie anschließend gezeigt wird, auf einen Ankunfts- und Abfertigungsprozeß des oben beschriebenen Typs zurückführen.

5.1 Modelle mit mehreren (parallelen) Schaltern

5.1.1 Unendlich viele Schalter

Dies wäre das Modell eines idealisierten Telefonvermittlungssystems, bei dem jeder eintreffende Anruf sofort eine freie Leitung erhält. Jeder der parallelen Schalter hat dieselbe Funktion und ankommende Kunden können zufällig auf freie Schalter verteilt werden. (Es bildet sich also keine eigentliche Warteschlange.) Zwischenankunftszeiten und Abfertigungszeiten seien exponentialverteilt mit den Parametern λ bzw. μ ($M/M/\infty$). Man denkt sich nun die parallelen Schalter in 1 Schalter zusammengefaßt, dessen Abfertigungsrate mit der Anzahl der Kunden im System linear zunimmt, setzt also

$$\mu_n = n\mu, \quad \lambda_n = \lambda \; , \quad n = 0,1,2, \ldots$$

Das führt auf $S = e^{\frac{\lambda}{\mu}}$ und damit auf

$$p_k = \frac{(\frac{\lambda}{\mu})^k}{k!} e^{-\frac{\lambda}{\mu}} \quad , \; k = 0,1,2, \ldots$$

Die Anzahl der Kunden im System ist also Poissonverteilt.

Man erhält daher für die mittlere Anzahl von Kunden im System = mittlere Anzahl der belegten Schalter:

$$E(N_q) = \frac{\lambda}{\mu}; \quad D^2(N_q) = \frac{\lambda}{\mu}$$

./. 20

Weiterhin ergibt sich für die Wahrscheinlichkeit, daß höchstens m Schalter belegt sind

$$w(N_q \leqq m) = e^{-\frac{\lambda}{\mu}} \sum_{k=1}^{m} \frac{\left(\frac{\lambda}{\mu}\right)^k}{k!}$$

(tabelliert als "cumulative Poisson function""oder "Erlang function").

Da man in der Praxis nicht unendlich viele Schalter vorsehen kann, wird man zwar zunächst das idealisierte Modell betrachten, sich aber danach fragen, wieviel Schalter z.B. mit einer Wahrscheinlichkeit von 99 % maximal benötigt werden. Entsprechend viele Schalter wird man dann vorsehen und in Kauf nehmen, daß das System in 1 % der Zeit nicht wunschgemäß arbeitet.

Die Ergebnisse von 5.1.1 bleiben gültig auch für den Fall beliebiger Serviceverteilung (M/GI/∞), wenn man $\frac{1}{\mu}$ durch E(S) ersetzt.

5.1.2 Wartesystem mit m Schaltern

Es werden dieselben Annahmen wie in 5.1.1 gemacht, lediglich wird die Maximalzahl der verfügbaren Schalter auf m begrenzt (M/M/m), außerdem wird unendlicher Warteraum vorausgesetzt. Es ist also

$$\lambda_n = \lambda \text{ für } n = 0,1,2, \ldots$$

$$\mu_n = \begin{cases} n\mu \text{ für } 0 \leqq n < m \\ m\mu \text{ für } n \geqq m \end{cases}$$

Das liefert

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

und damit

$$p_k = \begin{cases} \dfrac{(\frac{\lambda}{\mu})^k}{k!} \; S^{-1} & \text{für } k < m \\ \dfrac{(\frac{\lambda}{\mu})^k}{m!\, m^{k-m}} \; S^{-1} & \text{für } k \geqq m \end{cases}$$

5.1.3 Verlustsystem mit m Schaltern

Bei diesem klassischen, bereits von Erlang behandelten Modell können maximal m Kunden an parallelen Schaltern bedient werden; außerhalb der Schalter ist kein zusätzlicher Warteraum vorhanden. Sind alle m Schalter belegt, so gehen eintreffende Kunden verloren, d.h. ihre Ankunft wird ignoriert (Telefonvermittlung mit m parallelen Leitungen).

Die sonstigen Voraussetzungen bleiben wie in 5.1.2 (M/M/m).

Hier ist

$$\lambda_n = \begin{cases} \lambda \text{ für } n < m \\ 0 \text{ für } n \geqq m \end{cases}$$

$$\mu_n = n\mu \text{ für } n \leqq m$$

(für $n > m$ ist μ_n nicht definiert!)

Die Wahrscheinlichkeit, daß k Kunden im System, also k Schalter belegt sind, ergibt sich zu

$$p_k = \frac{\frac{1}{k!}\left(\frac{\lambda}{\mu}\right)^k}{\sum_{n=0}^{m} \frac{1}{n!}\left(\frac{\lambda}{\mu}\right)^n} \qquad k = 0,1,2, \ldots m$$

Interessant ist vor allem die Wahrscheinlichkeit p_m, daß ein eintreffender Kunde "verlorengeht" (Verlustwahrscheinlichkeit).

Auch die Ergebnisse von 5.1.3 bleiben gültig für beliebige Servicezeitverteilung, wenn man $\frac{1}{\mu}$ durch E(S) ersetzt.

./. 22

5.2 "Machine Minding" (endliches "Kundenreservoir")

Diesem als "Machine Minding" oder "Machine Interference" bezeichneten Modell liegt folgende Vorstellung zugrunde: m gleichartige Maschinen fallen in gewissen Abständen aus und brauchen "Wartung". Sie stehen deshalb unter der Kontrolle von einem oder mehreren "Wartungsleuten". Es wird angenommen, daß jeweils zum frühest möglichen Zeitpunkt mit der Wartung einer ausgefallenen Maschine begonnen wird. Vorzugeben ist die Verteilung der Zeit für die Wartung (Servicezeit) und die Verteilung der Zeit vom Abschluß der Instandsetzung einer Maschine bis zum nächsten Ausfall. Hier sei angenommen, daß beide Zeiten exponentialverteilt mit den Parametern λ und μ sind und daß 1 Wartungsmann für m Maschinen vorhanden ist. Der wesentliche Unterschied gegenüber früher behandelten Modellen besteht darin, daß das Reservoir der möglich "Kunden" endlich ist. Um die am Anfang von 5 skizzierte Methode anwenden zu können, ist zu setzen:

$$\lambda_n = \begin{cases} (m-n)\,\lambda & \text{für } n \leqq m \\ 0 & \text{für } n > m \end{cases}$$

$$\mu_n = \mu$$

Damit ergibt sich für die Wahrscheinlichkeit, daß k Maschinen ausgefallen sind:

$$p_k = \frac{\frac{1}{(m-k)!}\left(\frac{\mu}{\lambda}\right)^{m-k}}{\sum_{n=0}^{m} \frac{\left(\frac{\mu}{\lambda}\right)^n}{n!}}, \qquad k = 0,1,2, \ldots m$$

Diesem Modell läßt sich auch folgende Interpretation geben: Von insgesamt m Konsolen kommen Anfragen an den zentralen Rechner. Von einer bestimmten Konsole kann eine weitere Anfrage erst nach Bearbeitung der vorhergehenden Anfrage von dieser Konsole kommen.

./. 23

6. Kopplung von Warteschlangenmodellen

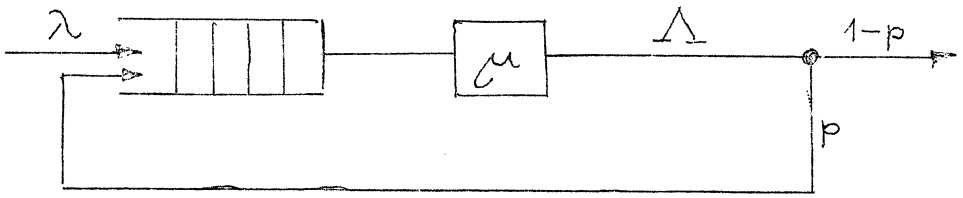
6.1 Allgemeines

Z.B. bei der Planung eines Betriebssystems hat man in vielen Fällen mehrere gekoppelte Warteschlangenmodelle zu betrachten, also ein "Netzwerk" von Bedienungssationen. Die wesentlichen neuen Züge gegenüber den früher betrachteten Modellen sind die Verzweigung bzw. Zusammenführung von Kundenströmen und die Hintereinanderschaltung von Bedienungsstationen. Es ist keine exakte Methode bekannt, Modelle dieser Art im allgemeinen Fall analytisch zu behandeln. Lediglich für den Spezialfall, daß alle Zwischenankunfts- und Servicezeiten exponentialverteilt sind und überall beliebig viel Warteraum zur Verfügung steht, gibt es eine einfache Lösungsmethode (Satz von J.R. Jackson, vgl.[1]). Selbst das Problem der Hintereinanderschaltung zweier Schalter A und B bereitet Schwierigkeiten, sobald die Servicezeit in A eine beliebige Verteilung hat; die Ankünfte der Kunden in B sind dann im allgemeinen nicht mehr unabhängig und das Problem, ein Modell mit allgemeinen, nicht unabhängigen Zwischenankunftszeiten zu behandeln, ist ungelöst.

6.2 Warteschlangenmodelle mit Feedback

Der einfachste Fall einer Kopplung liegt vor, wenn das Warteschlangenmodell nur aus 1 Schalter mit Rückkopplung (Feedback) besteht, wenn also ein Teil der bearbeiteten Kunden nach dem Verlassen dieses Schalters wieder in die Warteschlange zurückgestellt wird.

Ein einfaches Modell dieses Typs ist das folgende:



./. 24

Die Ankünfte der "neuen Kunden" seien Poisson-verteilt (Ankunftsrate λ), die Abfertigungszeiten exponentialverteilt (Abfertigungsrate μ). Die Kunden, die den Schalter verlassen, werden mit der Wahrscheinlichkeit p wieder in die Warteschlange zurückgestellt. Für den Durchsatz Λ des Systems gilt dann

$$\Lambda = \lambda + p\Lambda$$

also

$$\Lambda = \frac{\lambda}{1-p}$$

Daraus ergibt sich für die Verkehrsintensität

$$\varrho = \frac{\lambda}{\mu(1-p)}$$

Für die mittlere Verweilzeit im System bei 1 Durchgang erhält man

$$E(T_q^{(1)}) = \mu \frac{1}{1-\varrho}$$

für die mittlere Anzahl der Durchgänge

$$E(N) = \frac{1}{1-p}$$

woraus sich für die gesamte mittlere Verweilzeit im System ergibt:

$$E(T_q) = E(N) \cdot E(T_q^{(1)}) = \frac{1}{\mu(1-\varrho)(1-p)}$$

Die in der Praxis interessierenden Modelle weisen jedoch gegenüber dem hier behandelten u.a. folgende Komplikationen auf (vgl. [6], [7]):

a) gestutzte Servicezeitverteilung: Die Bearbeitung eines Kunden darf (in 1 Durchlauf) maximal 9 Zeiteinheiten dauern; Kunden, die innerhalb dieses Zeitquantums nicht abgefertigt werden können, werden in die Warteschlange zurückgestellt und ihre Bedienung wird im nächsten Durchgang weitergeführt (Round Robin);

./. 25

b) Verwaltungsaufwand (overhead) beim Wechsel von einem Kunden zum nächsten;

c) endlicher Warteraum;

d) Prioritäten, die entweder von außen vorgegeben oder dynamisch ermittelt werden (foreground-background queues).

Literaturhinweise:

[1] Analysis of Some Queuing Models in Real-Time Systems, IBM Manual Nr. F20-0007-0, New York 1965

[2] D.R. Cox and W.L. Smith - Queues, New York 1961

[3] J. Martin - Design of Real-Time Computer Systems, New Jersey 1967

[4] L. Takács - Introduction to the Theory of Queues, New York 1962

[5] S. Karlin - A First Course in Stochastic Processes, New York, London 1966

[6] E.G. Coffmann and L. Kleinrock - Feedback Queuing Models for Time-Shared Systems, JACM, Oktober 1968

[7] J.M. McKinney - A Survey of Analytical Time-Sharing Models, Computing Surveys, Juni 1969

[8] W. Wagner - Wartezeiten un Prioritäten, aus W. Händler (Herausgeber) - Teilnehmer-Rechensysteme, München, Wien 1968

[9] W. Wagner - Über ein kombiniertes Warte-Verlust-System mit Prioritäten, Stuttgart 1968