Warteschlangentheorie in der Betriebswirtschaftslehre

Grundlagen und zentrale Zusammenhänge

Dipl.-Wi.-Ing., M.S. Justus Arne Schwarz und Prof. Dr. Raik Stolletz, Mannheim

Dipl.-Wi.-Ing., M.S. *Justus Arne Schwarz* ist Wissenschaftlicher Mitarbeiter am Lehrstuhl für ABWL und Produktion der Universität Mannheim. Bevorzugte Forschungsgebiete: Zeitabhängige Warteschlangennetze, Produktionshochlaufplanung.

Prof. Dr. *Raik Stolletz* ist Inhaber des Lehrstuhls für ABWL und Produktion der Universität Mannheim. Bevorzugte Forschungsgebiete: Gestaltung stochastischer Produktionssysteme, Ressourceneinsatz – und Ablaufplanung, Analyse von dynamischen Produktionssystemen.

Warteschlangensysteme werden in zahlreichen Bereichen der Betriebswirtschaftslehre zur Analyse und Planung genutzt. Die Warteschlangentheorie hat zum Ziel den Vorgang des Wartens mit mathematischen Methoden zu beschreiben und mittels Leistungsanalysen Gestaltungsrichtlinien abzuleiten. Dieser Beitrag gibt eine Übersicht zu Anwendungsgebieten, Modellierungsansätzen und Zusammenhängen auf dem Gebiet der Warteschlangentheorie.

Stichwörter: Warteschlangentheorie, Kendall Klassifikation, Littles Gesetz, Skaleneffekte

1. Warteschlangensysteme und die Ursachen des Wartens

Warteschlangensysteme werden in verschiedenen Bereichen der Betriebswirtschaftslehre zur Analyse und Planung genutzt, z. B. in der Dienstleistungsbranche, der Fertigungsindustrie und der Logistik. Ein klassisches einstufiges Wartesystem besteht aus zufällig eintreffenden Aufträgen (Jobs), einem Warteraum mit k Warteplätzen, in dem die Aufträge bis zum Beginn der Bearbeitung verweilen, und einem oder mehreren parallelen Bedienstationen (Servern), die die Aufträge mit stochastisch schwankenden Bedienzeiten abarbeiten (siehe *Abb. 1*). In *Tab. 1* sind Beispiele von Wartesystemen aufgeführt.

Zur Beschreibung des jeweils zugrunde liegenden Warteschlangenmodells wird folgende Charakterisierung verwendet: Das **Auftragsprofil** beschreibt den stationären oder zeitabhängigen Ankunftsprozess der Aufträge mittels der Wahrscheinlichkeitsverteilung der Zeit zwischen zwei aufeinanderfolgenden Aufträgen (Zwischenankunftszeit). Verlassen ungeduldige Kunden in einem Service-System die Warteschlange vor ihrer eigentlichen Bedienung, gehört auch die Ungeduldsverteilung zum Auftragsprofil.

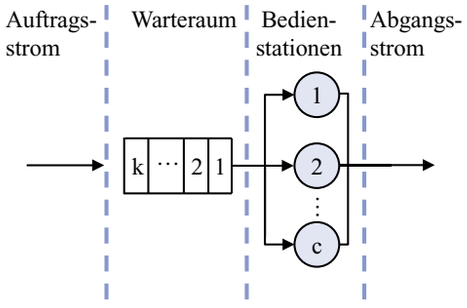


Abb. 1: Aufbau eines einstufigen Wartesystems

Anwendungsgebiet	Aufträge	Bedienstationen	Warteschlangendisziplin
Supermarkt	Kunden	Kassierer	First Come, First Served
Callcenter	Anrufe	Telefonist	Kundenklassen (Neu- u. Bestandskunden, Sprache)
Krankenhaus	Patienten	Arzt	Priorisierung von Notfällen
Hochregallager	Ein- und Auslageraufträge	Regalbediengerät	Priorisierung nach frühestem Liefertermin
Produktionssysteme	Fertigungsaufträge	Maschine	Priorisierung von Eilaufträgen
Informatik	Datenpakete	CPU	First Come, First Served

Tab. 1: Beispiele für Warteschlangen in der Praxis

https://doi.org/10.15358/0340-1650_2013_5_261

Das **Serverprofil** charakterisiert mittels der Verteilung der Bedienzeiten das Bedienverhalten. Die Wahrscheinlichkeitsverteilung kann ebenfalls stationär oder zeitabhängig sein. Können Bedienstationen ausfallen, wird auch dies zur Beschreibung des Servers angegeben. Mit der **Warteschlangendisziplin** wird festgelegt, nach welchem Prinzip wartende Kunden zur Bedienung herangezogen werden. Klassische Vorgehen sind die ankunftsorientierte First Come, First Served (FCFS)-Regel oder Prioritätsregeln.

Wartezeiten können selbst dann entstehen, wenn die Rate λ, mit der neue Aufträge das Wartesystem erreichen, kleiner ist als die Rate $c\mu$, wobei jeder der c Server Aufträge mit Rate μ bearbeiten kann. Dies ist durch die zufällige Variation der Zeit zwischen zwei Ankünften und der Bedienzeiten begründet. Treffen z. B. aufgrund unterdurchschnittlicher Zwischenankunftszeiten in einem Zeitintervall mehrere Aufträge am Wartesystem ein, während gerade ein Auftrag mit überdurchschnittlicher Bearbeitungsdauer abgearbeitet wird, müssen die ankommenden Aufträge warten.

Die Wahrnehmung der Wartezeit hängt oft stark von der Einstellung des Wartenden ab. Aus der **ökonomischen Perspektive** verursachen Wartezeiten z. B. unzufriedene Kunden oder Kosten für die Lagerung unfertiger Erzeugnisse. Der positive Effekt einer Warteschlange besteht in der Möglichkeit, schwankende Nachfragen auszugleichen. Dadurch kann eine hohe Auslastung einer knappen Ressource gewährleistet werden. Die **Warteschlangentheorie** hat zum Ziel diesen Trade-off mit mathematischen Methoden zu beschreiben und mittels Leistungsanalysen Gestaltungsrichtlinien für Wartesysteme abzuleiten.

2. Leistungsanalyse

Zur formalen Klassifikation einstufiger Systeme wird die *a/b/c/d* **Notation von Kendall** genutzt. Der Eintrag *a* beschreibt die Zwischenankunftszeitverteilung, *b* die Bedienzeitverteilung, *c* die Anzahl an parallelen Bedienstationen und *d* die maximale Anzahl an Aufträgen im System ($k + c$ in *Abb. 1*). Für die gängigsten Verteilungen werden folgende Abkürzungen verwendet:

- M – Exponentialverteilung
- D – Deterministische Zeiten
- E_k – Erlang-k-Verteilung
- G – Generelle Verteilung (gegeben durch Erwartungswert und Varianz)

Zur Messung des Systemverhaltens werden auftragsbezogene und systembezogene Leistungskenngrößen definiert. Klassische auftragsbezogene Kennwerte sind die Verteilungen der Wartezeit W_q und der Durchlaufzeit W bzw. deren Erwartungswerte $E[W_q]$ und $E[W]$. Die Durchlaufzeit gibt dabei die Summe aus Warte- und Bedienzeit an. Für Systeme mit beschränkter Systemkapazität ($k + c < \infty$) und solche, bei denen ungeduldige Kunden die Schlange verlassen, ist auch die Wahrscheinlichkeit der Bedienung *P(Service)* relevant.

Zur Bewertung des Systemzustandes werden die Verteilungen der Warteschlangenlänge L_q oder der Aufträge im System L herangezogen. Häufig wird der Vergleich verschiedener Systeme auf Basis der aus den Verteilungen berechneten Erwartungswerten $E[L_q]$ und $E[L]$ und deren Varianzen durchgeführt. Weiter wird die Auslastung ρ des Wartesystems genutzt, um Aussagen über die Effizienz des Ressourceneinsatzes zu erhalten.

3. Modellierungsansätze in der Warteschlangentheorie

Grundidee bei der Modellierung von Wartesystemen ist es, zunächst eine geeignete Beschreibung aller Zustände des Wartesystems zu finden. Für ein M/M/c/k System wäre dies z. B. die Anzahl der Kunden im System.

Unter bestimmten Verteilungsannahmen (Markovsysteme) lassen sich für Kennzahlen des Wartesystems im eingeschwungenen Zustand **analytische Lösungen** angeben (vgl. *Gross* et al., 2008). Der eingeschwungene Zustand ist dann erreicht, wenn sich die Eintrittswahrscheinlichkeiten der Systemzustände im Zeitverlauf nicht mehr verändern. Unter Ausnutzung der Markoveigenschaft werden die Übergangsraten von jedem Zustand zu allen anderen Zuständen bestimmt. Die auch als „Gedächtnislosigkeit" bekannte Markoveigenschaft impliziert, dass ein zukünftiges Ereignis nur vom aktuellen Zustand abhängt (gilt insbesondere für die Exponentialverteilung). Anschließend wird für jeden Zustand eine Gleichung aufgestellt. Dies erfolgt nach dem Prinzip „Summe der Zugangsraten in den Zustand = Summe der Abgangsraten aus dem Zustand heraus". Aus den einzelnen Gleichungen (Chapman-Kolmogorov-Gleichungen) und einer Normierungsbedingung ergibt sich ein lineares Gleichungssystem. Die Lösung dessen liefert die Zustandswahrscheinlichkeiten, aus denen die systembezogenen Kennzahlen abgeleitet werden.

Die Bestimmung von auftragsbezogenen Kenngrößen erfolgt anschließend mittels ***Littles* Gesetz** (vgl. *Little*, 1961). Es stellt die Verbindung von system- und auftragsbezogenen Leistungskenngrößen her. Die erwartete Anzahl von Einheiten im System ergibt sich aus dem Produkt von effektiver Ankunftsrate λ_{eff} und erwarteter Wartezeit, also $E[L] = \lambda_{eff} E[W]$. Die effektive Ankunftsrate gibt dabei die Rate der tatsächlich bedienten Aufträge an. Es gilt also $\lambda_{eff} = \lambda P(Service)$. *Littles* Gesetz gilt sowohl für das gesamte Wartesystem als auch für den Warteraum mit $E[L_q] = \lambda_{eff} E[W_q]$ und ist darüber hinaus unabhängig von den Verteilungsannahmen. Aus dem Gesetz folgt weiter, dass sich der Bestand in einem System bei gleichem Durchsatz nicht reduzieren lässt, ohne die Durchlaufzeit der Aufträge im System zu reduzieren.

Wird die Markoveigenschaft der Exponentialverteilung aufgegeben, lassen sich oft nur **approximative Lösungen** angeben. Für das G/G/c/∞ ist eine der bekanntesten Formeln die Diffusionsapproximation von Kingman (vgl. *Curry/Feldman*, 2010):

$$E[W_q] \approx \frac{c_a^2 + c_b^2}{2} \times \frac{\rho^{\sqrt{2(c+1)}-1}}{c(1-\rho)} \times \frac{1}{\mu} \quad (1)$$

Hier gibt c_a^2 den quadrierten Variationskoeffizienten (Varianz/quadrierter Erwartungswert) des Ankunftsprozesses an und c_b^2 den entsprechenden Wert für den Bedienprozess. Die Auslastung ist gegeben mit $\rho = \lambda/(c\mu)$, wobei λ die Ankunftsrate und μ die Bedienrate ist.

Als weitere Methode zur Bestimmung des Verhaltens von Warteschlangensystemen kann auf computerbasierte **Simulation** zurückgegriffen werden. Vorteil der Simulation ist die Möglichkeit der detaillierten Abbildung komplexer Systeme. Ein entscheidender Nachteil dieser Herangehensweise ist der vergleichsweise hohe Zeitaufwand.

Die Modellierungsansätze können weiter nach der Betrachtungsweise der Zeit in zeitkontinuierliche und zeitdiskrete Modelle unterteilt werden. Während bei der kontinuierlichen Analyse zu jedem beliebigen Zeitpunkt zustandsverändernde Ereignisse auftreten können, kann dies bei der zeitdiskreten Analyse nur zu vorgegebenen Zeitpunkten geschehen.

4. Zentrale Zusammenhänge in Wartesystemen

4.1. Einflussfaktoren auf die Wartezeit

Anhand eines G/G/1/∞ Wartesystems werden im Folgenden die zentralen Treiber der Wartezeit gezeigt. In *Abb. 2* ist die mittels (1) bestimmte erwartete Wartezeit in Abhängigkeit der Gesamtvariabilität ($c_a^2 + c_b^2$) für verschiedene Auslastungen dargestellt. Die Wartezeit steigt proportional mit der Variabilität. Abhängig von der Gesamtvariabilität kann eine weitere Erhöhung einer ohnehin schon hohen Auslastung zu einem überproportionalen Anstieg der Wartezeit führen. Die unter Abschnitt 1 beschriebenen stochastischen Schwankungen führen besonders bei hohen Auslastungen zu übermäßigen Wartezeiten, weil ausgleichend wirkende Leerzeiten an der Bedienstation fehlen.

Für die Reduzierung der erwarteten Wartezeit ergeben sich zwei zentrale Stellhebel. Eine Möglichkeit stellt die Auslastungsreduzierung dar. Diese kann entweder durch Rückgang der ankommenden Aufträge (und damit eine Reduzierung des Durchsatzes) oder die Erhöhung der Bedienkapazität, z. B. durch eine Verkürzung der Bedienzeit oder dem Einsatz zusätzliche Server, erreicht werden. Eine Verringerung der Bedienzeit lässt sich durch Schulungen von Mitarbeitern oder Verbesserungen an Maschinen erreichen.

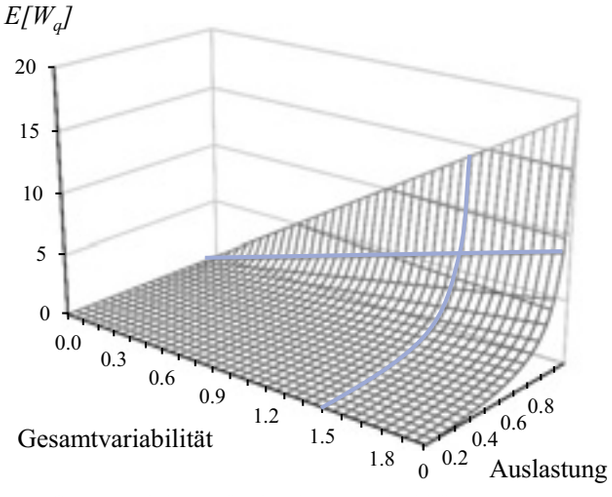


Abb. 2: Wartezeit in Abhängigkeit von Variabilität und Auslastung

Die zweite Möglichkeit stellt die **Variabilitätsreduktion** dar. Die Wartezeit lässt sich also auch verringern, wenn der Erwartungswert der Bedienzeit unverändert bleibt und nur die Variabilität der Bedienzeit reduziert wird. Eine Begrenzung der Variabilität im Auftragsstrom ist in Service-Systemen oft nur bedingt möglich. Bei der Produktionsplanung versucht man hingegen durch gezielte Nivellierung der Aufträge zu stabileren Produktionsbedingungen beizutragen.

4.2. Skaleneffekte in Warteschlangensystemen

Einen in der Betriebswirtschaft bekannten Zusammenhang beschreiben die Skaleneffekte (engl. **Economies of Scale**). Demzufolge können Systeme allein durch ihre Größe Kostenvorteile generieren. Dieser Effekt lässt sich mittels der Warteschlangentheorie für im gleichen Verhältnis steigende Ankunftsraten und Bedienstationen quantifizieren (vgl. *Helber/Stolletz*, 2004). Die Skaleneffekte zeigen sich in Form von reduzierten erwarteten Durchlauf- und Wartezeiten bzw. einer kürzeren erwarteten Warteschlangenlänge. Die Systemauslastung wird hingegen nicht beeinflusst. Die Größe der erzielbaren Einsparungen hängt dabei ab von: der Größe der verglichenen Systeme, der Variabilität in Ankunfts- und Bedienprozess sowie der Auslastung.

Die Begründung für diese Einsparungen liegt zum einen im Ausgleich von stochastischen Schwankungen durch die vorgeschaltete **gemeinsame Warteschlange** (vgl. M/M/2 System *Abb. 3*). Zum anderen werden im Vergleich zu zwei unabhängigen Wartesystemen Zustände vermieden, in denen sich vor einem Server eine Warteschlange bildet und der andere auf Aufträge wartet. Das numerische Beispiel in *Abb. 3* zeigt, dass sich sowohl die erwartete Warte-

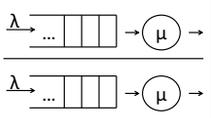
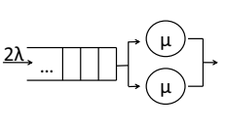
	2 unabhängige M/M/1 Systeme	**M/M/2 System mit gemeinsamem Warteraum**
		
Ankunftsrate	jeweils λ=0.9	ein Strom mit 2λ=1.8
E[W]	10.0	5.3
$E[W_q]$	9.0	4.3
E[L]	18.0	9.5
$E[L_q]$	16.2	7.7
ρ	0.9	0.9

Abb. 3: Skaleneffekte in Wartesystemen, μ=1

schlangenlänge als auch die erwartete Wartezeit bei einem zusammengeschlossenen System im Vergleich zu zwei unabhängigen Systemen mehr als halbiert.

5. Aktuelle Forschungsfragen in der Warteschlangentheorie

Ein Schwerpunkt der aktuellen Forschung ist die Herleitung analytischer Methoden für **Warteschlangennetzwerke**, z. B. für globale Liefernetzwerke oder Produktionslinien. Ein weiterer Fokus liegt auf der Planung **zeitabhängiger Warteschlangensysteme**. Gemeinsam haben beide Ströme das Ziel, durch ein verbessertes Verständnis der Wartesysteme einen wirtschaftlicheren Betrieb zu ermöglichen.

Literatur

Curry, G. L., R. M. Feldman, Manufacturing Systems Modeling and Analysis, 2. Aufl., Berlin/Heidelberg 2010.

Gross, D., J. F. Shortle, J. M. Thompson, C. M. Harris, Fundamentals of Queueing Theory, 4. Aufl., Hoboken 2008.

Helber, S., R. Stolletz, Call Center Management in der Praxis, Berlin/Heidelberg 2004.

Little, J. D. C., A Proof for the Queuing Formula: L = λ W, in: Operations Research, Vol. 9 (1961), S. 383–387.