Warum steht man immer in der falschen Schlange?

Mathematischer Hintergrund zu einem alltäglichen Problem

Von Thomas Hanschke

Einen Großteil unserer Zeit verbringen wir in der Warteschlange: an der Kasse im Supermarkt, im Wartezimmer beim Arzt oder im Verkehrsstau auf dem Weg zur Arbeit. Jeder von uns kennt das Unbehagen, in einer langen Warteschlange zu stehen. Oftmals läßt sich nicht einmal abschätzen, wie lange man noch ausharren muß und ob es sich überhaupt lohnt abzuwarten.

Abgesehen von der psychischen Beeinträchtigung, die von einer langen Warteschlange ausgeht, sind Warteschlangen auch aus betriebswirtschaftlicher Sicht nicht wünschenswert. Denn die Zeit, die man in der Warteschlange zubringt, ist unproduktive Zeit, die weder den Kunden noch dem Betreiber des Systems zugute kommt. Den wirtschaftlichen Einfluß von Warteschlangen kann man am besten am Beispiel der Produktion verdeutlichen. Lange Durchlaufzeiten durch die Produktion haben zur Folge, daß man neue Produkte nicht schnell genug auf den Markt bringen kann und der Konkurrenz das Feld überlassen muß. Da lange Durchlaufzeiten mit hohen Beständen korreliert sind, entstehen durch die auf Bearbeitung wartenden Halbfertigfabrikate außerdem hohe Kapitalbindungskosten, die sich negativ auf das Betriebsergebnis auswirken.

Das Phänomen des Wartens wird seit fast einem Jahrhundert wissenschaftlich erforscht. Bereits 1917 publizierte der dänische Ingenieur A.K. Erlang, der bei einer Kopenhagener Telefongesellschaft beschäftigt war, eine mathematische Formel zur Dimensionierung von Fernsprechvermittlungsstellen. Die Fachleute erzählen sich, daß gerade diejenigen Länder, die Mitte des vergangenen Jahrhunderts über die schlechtesten Telefonsysteme verfügten, zumindest die besten Mathematiker auf dem Gebiet der Warteschlangentheorie hervorgebracht hätten. Mit dem Aufkommen der Datenverarbeitung werden diese Methoden auch zur Konzeption von Rechen-, Produktions- und Logistiksystemen verwendet. Ziel der Analysen ist es, bereits im Vorfeld der Planung Engpässe und Schwachstellen zu erkennen. Inzwischen sind mehrere tausend wissenschaftliche Publikationen über Warteschlangenprobleme erschienen, die sich auf die unterschiedlichsten Bereiche unseres täglichen Lebens beziehen. Angesichts einer so großen Zahl gesicherter Erkenntnisse läßt sich kaum erklären, warum wir heute noch so oft in der Warteschlange stehen ...

Das Grundmodell in einer Warteschlange

Die Warteschlangentheorie bedient sich zur Beschreibung von Bedienungssystemen eines einfachen Grundmodells. Es besteht aus dem sogenannten Bedienungsschalter, der über ein oder mehrere parallel arbeitende gleichartige Maschinen bzw. Arbeitsplätze verfügt, und aus einem Warteraum. Die Kunden treffen einzeln und zu zufälligen Zeitpunkten vor dem Bedienungsgerät ein. Ein neu ankommender Kunde wird bedient, sofern mindestens eines der Bedienungsgeräte frei ist, andernfalls muß er sich in die Warteschlange einreihen.

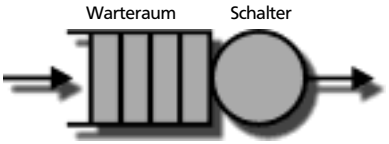

Bild 1: Das Grundmodell der Warteschlangentheorie

Die Begriffe Kunde und Schalter können in der Praxis unterschiedliche Bedeutung haben: Fahrzeuge, die an einer Ampel warten; Computerprogramme, die in einem Rechnerverbund zirkulieren; Telefonanrufe, die an einer Vermittlungsstelle einfallen; Werkstücke, die von einer Maschine bearbeitet werden; Patienten, die in einer Arztpraxis auf ihre Behandlung warten, usw.

Das Grundmodell kann auf vielfältige Weise variiert werden:

- Die Kunden werden nicht einzeln, sondern gruppenweise bedient (Wartesysteme mit Gruppenbedienung).
  Anwendung: Losfertigung in einem Produktionsbetrieb
- Einige Kunden verlassen das System, bevor sie bedient worden sind (Wartesysteme mit Zeitbeschränkungen).
  Anwendung: Lagerhaltung von verderblicher Ware
- Nicht alle Bedienungsgeräte stehen jedem Kunden zur Verfügung (Bedienungssysteme mit eingeschränkter Erreichbarkeit).

Anwendung: Fertigungsstraßen mit dedizierten Maschinen, Koppelanordnungen in einem Fernsprechnetz

- Einige Kunden scheuen sich, in das Bedienungssystem einzutreten, weil ihnen die Warteschlange zu lang erscheint (Wartesysteme mit ungeduldigen Kunden).
  Anwendung: übliches Kundenverhalten an einem Post-, Bank- oder Fahrkartenschalter
- Ein Kunde mit höherer Priorität verdrängt einen Kunden niedrigerer Priorität aus dem Bedienungsprozeß (Bedienungssysteme mit Prioritätssteuerung).
  Anwendung: Expreß-Los-Steuerung in einem Fertigungsprozeß
- Ein Kunde, der bei seiner Ankunft nicht sofort bedient werden kann, geht verloren (Verlustsysteme).
  Anwendung: Telefonate in einem Fernsprechnetz.

Der Strom der ankommenden Forderungen wird durch einen sogenannten Erneuerungsprozeß beschrieben. Dazu denken wir uns alle Forderungen in der Reihenfolge ihrer Ankünfte durchnumeriert. Die Zeitspanne I_n zwischen der Ankunft des $(n-1)$-ten und des n-ten Kunden wird als Zwischenankunftszeit bezeichnet. Von den Zufallsgrößen I_n $(n = 1,2,...)$ wird vorausgesetzt, daß sie stochastisch unabhängig und identisch verteilt sind mit dem Erwartungswert $\mathbf{E}[I]$ und dem Variationskoeffizienten c_I. Der Kehrwert $\lambda = 1/\mathbf{E}[I]$ heißt Ankunftsrate und gibt an, wieviele Kunden im Durchschnitt pro Zeiteinheit in das System einfallen.

Die Bedienungszeiten S_n $(n = 1,2,...)$ der aufeinanderfolgenden Kunden werden ebenfalls als stochastisch unabhängige und identisch verteilte Zufallsgrößen aufgefaßt. Für den zugehörigen Erwartungswert und den zugehörigen Variationskoeffizienten verwenden wir die Symbole $\mathbf{E}[S]$ und c_S. Der Kehrwert $\mu = 1/\mathbf{E}[S]$ heißt Bedienungsrate und gibt an, wieviele Kunden im Durchschnitt pro Zeiteinheit von dem Bedienungssystem abgefertigt werden können. Sind mehrere parallele und gleichartige Bedienungsgeräte vorhanden, erhöht sich die Bedienungsrate entsprechend der Anzahl der Geräte. Die Bedienungsregel legt fest, in welcher Reihenfolge die wartenden Kunden abgefertigt werden sollen. Folgende Regeln und Bezeichnungen sind gebräuchlich:

FIFO (FCFS)	First In, First Out (First Come, First Served). Die Bedienung erfolgt in der Reihenfolge der Ankünfte.
LIFO (LCFS)	Last In, First Out (Last Come, First Served). Die Bedienung erfolgt in umgekehrter Reihenfolge der Ankünfte.
SIRO	Selection In Random Order. Der nächste Kunde wird zufällig ausgewählt.
Non-preemptive priority	Relative Priorität. Manche Kunden werden gegenüber anderen Kunden vorrangig behandelt. Der laufende Bedienungsprozeß wird jedoch nicht unterbrochen.
Preemptive priority	Absolute Priorität. Besitzt der neu ankommende Kunde gegenüber den anderen Kunden im System eine höhere Priorität, so wird der laufende Bedienungsprozeß unterbrochen und mit der neuen Forderung fortgesetzt. Die alte Forderung wird zurückgestellt.
RR	Round Robin. Jeder Kunde kann den Bediener jeweils nur für ein bestimmtes Zeitintervall in Anspruch nehmen. Kunden, deren Abfertigung mehr Zeit benötigt, müssen sich deshalb mehrmals hintereinander in die Warteschlange einreihen.

Zur symbolischen Kennzeichnung von Bedienungssystemen haben D.G. Kendall und B.W. Gnedenko die Notation

A/B/c/N

eingeführt. Die Buchstaben A und B markieren hierbei den Verteilungstyp der Zwischenankunftszeiten und der Bedienungszeiten. Der Buchstabe c steht für die Anzahl der parallelen Bediener und N bezeichnet die Kapazität des Warteraums. Für den Verteilungstyp sind folgende Abkürzungen gebräuchlich:

D	Deterministische Verteilung
M	Exponentialverteilung
E_k	Erlang-Verteilung mit Parameter k $(k = 1,2,...)$
H_k	Hyperexponentialverteilung mit Parameter k $(k = 1,2,...)$
PH	Phasentyp-Verteilung
G	Allgemeine Verteilung

Beispiel: Die Notation M/G/3/5 kennzeichnet ein Bedienungssystem mit exponentiell verteilten Zwischenankunftszeiten, beliebig verteilten Bedienungszeiten, drei parallelen Bedienern und einem Warteraum, in dem maximal 5 Kunden warten können.

Leistungsgrößen

Die Leistungsbewertung von Bedienungssystemen erfolgt auf der Basis folgender Prozesse:

- Die Länge der Warteschlange $(Q_t)_{t>0}$.
  Dieser Prozeß gibt an, wieviele Kunden sich zur Zeit t in der Warteschlange (vor dem Schalter) aufhalten.
- Der Prozeß der aufeinanderfolgenden Wartezeiten $(W_n)_{n>0}$.
  Die Zufallsvariable W_n bezeichnet die Zeit, die der n-te Kunde in der Warteschlange verweilt, ehe er bedient wird.

Zur Berechnung der Kenngrößen können verschiedene Methoden der Theorie der stochastischen Prozesse herangezogen werden. Die Eignung einer Methode hängt sehr stark davon ab, welche Verteilungstypen für die Zwischenankunfts- und Bedienungszeiten zugrundegelegt werden und ob zeitabhängige oder stationäre Größen berechnet werden sollen. Schon das Grundmodell der Warteschlangentheorie ist so kompliziert, daß es unter allgemeinen Verteilungsannahmen nicht exakt gelöst werden ▶

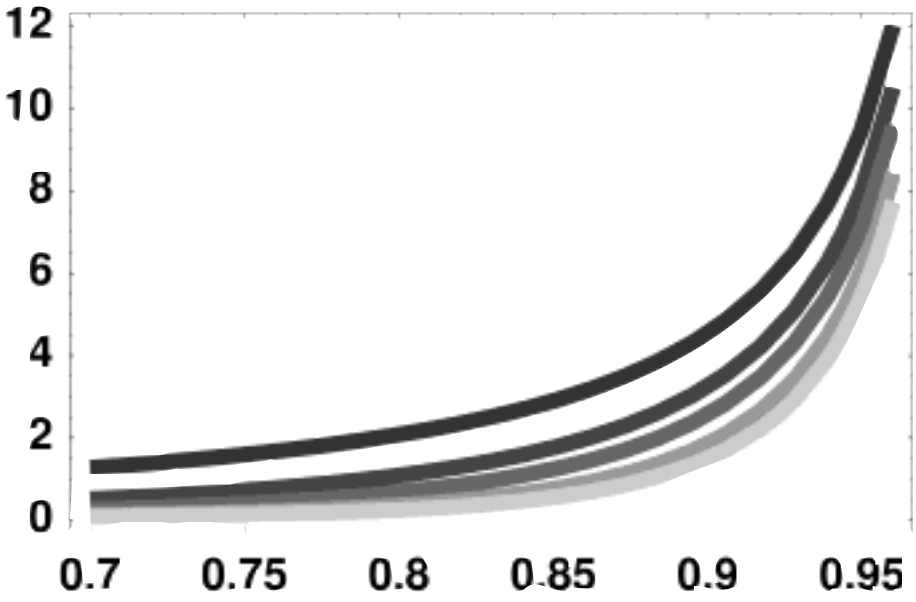

Bild 2: Die mittlere Warteschlangenlänge in Abhängigkeit von der Auslastung und der Variabilität des Systems.
Untere Kurve: geringe Variabilität. Obere Kurve: hohe Variabilität.

kann. Es existieren allerdings Näherungsformeln, die sich in der Praxis recht gut bewährt haben und die die stochastische Funktionsweise von Bedienungssystemen transparent machen. Nach einer Formel von Allen-Cunnen gilt für die mittlere Warteschlangenlänge im stationären Fall:

$$E(Q) \approx \frac{\rho}{1-\rho} \cdot \sqrt{\rho^{c+1}} \cdot \frac{c_I^2 + c_S^2}{2}.$$

Hierbei bedeuten ρ die Auslastung des Systems und c^2_I sowie c^2_S die quadrierten Variationskoeffizienten der Zwischenankunfts- und Bedienungszeiten. Die Formel lehrt uns, daß die Warteschlange umso länger ist, je größer die Auslastung des Systems und die Variationskoeffizienten sind **(Bild 2).** Um eine kurze Warteschlange zu bekommen, muß man folglich genügend Kapazität bereitstellen oder die Variabilität des Systems gering halten.

Die mittlere Wartezeit im stationären Fall erhält man mit Hilfe der Formel von Little:

$$E(W) = \frac{E(Q)}{\lambda} \approx \frac{1}{\lambda} \cdot \frac{\rho}{1-\rho} \cdot \sqrt{\rho^{c+1}} \cdot \frac{c_I^2 + c_S^2}{2},$$

wobei λ die Inputrate des Systems bedeutet.

Die kurze oder die lange Warteschlange?

Nachdem wir die Grundbegriffe der Warteschlangentheorie erläutert haben, sind wir in der Lage, die in der Überschrift genannte Frage zu behandeln. Stehen wie im Supermarkt oder bei der Bundesbahn mehrere Kassen zur Auswahl, so neigt man dazu, diejenige mit der kürzesten Warteschlange auszuwählen. Doch offensichtlich garantiert diese Strategie nicht, daß man auch schneller abgefertigt wird. So entsteht der Eindruck, in der falschen Warteschlange zu stehen. Da die Arbeitsaufträge der einzelnen Kunden zufällig schwanken, könnte es sein, daß in der langen Warteschlange zufällig viele kleine Aufträge akkumuliert sind, während in der kurzen Warteschlange große Aufträge vorherrschen.

Als Beispiel betrachten wir zwei gleichartige M/M/1/∞-Bedienungssysteme. In dem einen warten zum Zeitpunkt unserer Ankunft m, in dem anderen n Kunden, wobei wir annehmen wollen, daß m kleiner oder gleich n ist. Die Wartezeit in der Schlange entspricht allgemein der Summe der Bedienungszeiten der Vorgänger, wobei mathematisch vereinfachend hinzukommt, daß die Restbedienungszeit des sich zum Zeitpunkt unserer Ankunft gerade im Bediener befindlichen Kunden aufgrund der sogenannten Gedächtnislosigkeit der Exponentialverteilung ebenfalls exponentiell verteilt ist mit dem Parameter μ. Aus der Wahrscheinlichkeitstheorie ist bekannt, daß die Summe von m unabhängigen mit dem Parameter μ exponentiell verteilten Zufallsgrößen Erlang-verteilt ist mit den Parametern μ und m:

$$P(W_{(m)} > t) = e^{-\mu t} \cdot \sum_{k=0}^{m-1} \frac{(\mu t)^k}{k!} \quad (t \geq 0).$$

Damit läuft unsere Fragestellung auf den Vergleich zweier Erlang-verteilter Zufallsgrößen hinaus, wobei die eine Erlang-verteilt ist mit den Parametern μ und *m* und die andere Erlang-verteilt ist mit den Parametern μ und *n*. Gefragt ist also nach der Wahrscheinlichkeit, daß $W_{(m)}$ größer ist als $W_{(n)}$, in Zeichen: $P(W_{(m)} > W_{(n)})$. Diese Wahrscheinlichkeit kann mit dem Instrumentarium der bedingten Wahrscheinlichkeit ausgerechnet werden. Indem man die Gesetzmäßigkeiten der Erlang-Verteilung ausnutzt, bekommt man:

$$P(W_{(m)} > W_{(n)}) = \int_0^\infty e^{-\mu} \cdot \sum_{k=0}^{m-1} \frac{(\mu t)^k}{k!} \cdot \frac{\mu^n \cdot t^{n-1}}{(n-1)!} e^{-\mu} dt = \sum_{k=0}^{m-1} \binom{n+k-1}{k} \left(\frac{1}{2}\right)^{n+k}.$$

Diese Formel haben wir für verschiedene Parameter *m* und *n* ausgewertet **(Bild 3)**. Man kommt zu dem Ergebnis, daß die Wahrscheinlichkeit, in der kürzeren Schlange länger ausharren zu müssen als in der längeren, sich deutlich von Null unterscheidet.

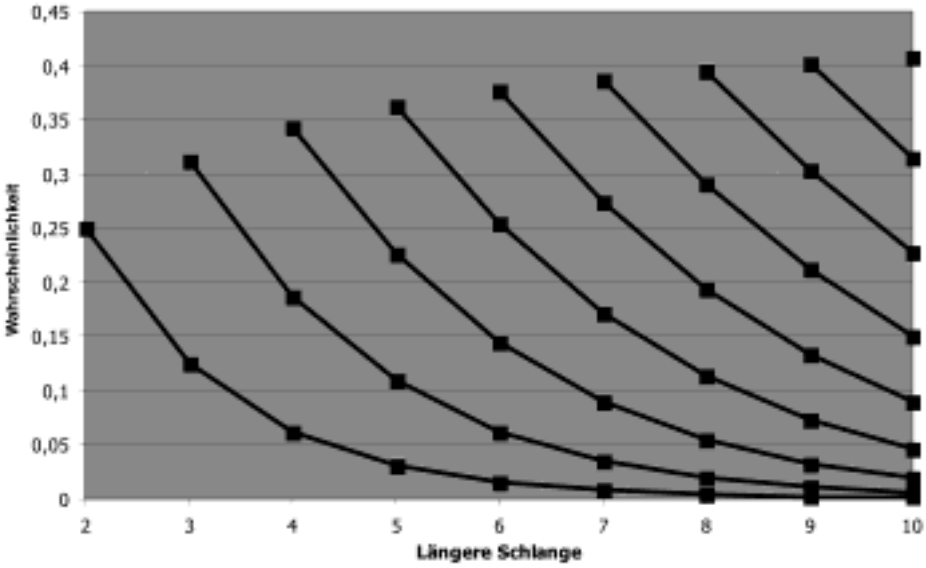

Bild 3: Die Wahrscheinlichkeit $P(W_{(m)} > W_{(n)})$ für verschiedene Werte *m* und *n*.
Die Kurven beziehen sich von links nach rechts auf die Werte m = 1,2,3,...,9.

Optimierung

Für alle günstiger ist es, in einer gemeinsamen Warteschlange zu warten. Nicht nur, daß sich jetzt die Frage nach der falschen Schlange gar nicht erst stellt, es ergibt sich auch eine deutlich kürzere Wartezeit. Um diesen Sachverhalt zu bestätigen, muß man nur die mittlere Wartezeit des Mehrbediener-Systems mit der mittleren Wartezeit des äquivalenten Systems von Einbediener-Systemen (c = 1) vergleichen. Dabei ist lediglich zu beachten, daß jeder Strang nur mit dem c-ten Teil des Inputstroms belastet wird. Als Verhältniszahl bekommt man:

$$F = \frac{\sqrt{\rho^{c+1}}}{\rho} \cdot \frac{1}{c}.$$

Da die Auslastung stets kleiner als Eins ist, läßt sich die Wartezeit folglich mindestens um den Faktor c reduzieren. Im Fall von zwei Warteschlangen müßte man also nur die halbe Zeit in der Warteschlange verbringen.

Übrigens fast zeitgleich mit einer Wissenschaftssendung des WDR, in der wir über diese Fragestellungen berichten durften, hat die Deutsche Bahn AG die Warteordnung in ihren Reisezentren geändert. Sie läßt ihre Kunden jetzt nicht mehr in separaten Warteschlangen sondern genauso wie am Flughafen in einer gemeinsamen Warteschlange zusammenlaufen. Dieses Beispiel zeigt allerdings auch, wie mühsam und langwierig der Prozeß der Umsetzung wissenschaftlicher Resultate in die Praxis ist.

Ausblick

Die moderne Mathematik ist inzwischen in der Lage, nicht nur einzelne Warteschlangensysteme, sondern auch Netzwerke von vielen hundert Knoten, wie sie in der industriellen Produktion oder der Telekommunikation vorkommen, zu rechnen. Auf der Basis dieser Methoden hat die Arbeitsgruppe Stochastische Modelle in den Ingenieurwissenschaften an der TU Clausthal Software-Lösungen entwickelt, die bereits bei der Dillinger Hütte AG und der IBM Storage Systems Division im Einsatz sind.

Prof. Dr. Thomas Hanschke
Institut für Mathematik
Erzstraße 1
38678 Clausthal-Zellerfeld
Tel.: 05323/72-2401
Fax: 05323/72-2304