Factory Physics – Zusammenfassung

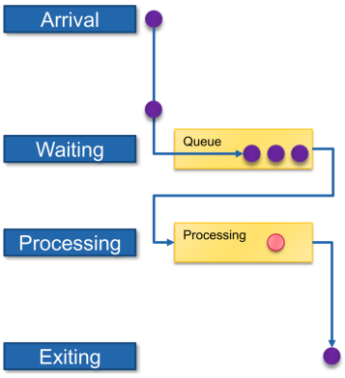
Was machen wir in der Factory Physics?

- Wollen ein grundlegendes Verständnis für Warteschlangensituationen und deren Ursachen entwickeln
- Wollen Gesetze ableiten und anwenden, um betriebliche Abläufe zu analysieren.
- Wir wollen Warteschlangenphänomene von betrieblichen Abläufen in Industrie- oder Dienstleistungsumgebungen mit Hilfe eines naturwissenschaftlichen Ansatzes beschreiben.

Ursache der Warteschlangenbildung

- Ressourcen (physisch, technisch oder menschlich) werden benötigt, um eine Aktivität an Entitäten (Objekten / Einheiten) - gemeinhin als Kunden oder Objekte bezeichnet – durchzuführen.
- Warteschlangenphänomene entstehen durch den Verbrauch begrenzter Ressourcen
  - Begrenzte Kapazitäten (z.B. Mangel an Mitarbeitern oder Maschinen)
  - Zeitaufwendige Dienstleistungen
  - Nachfragespitzen
- Selbst wenn im Durchschnitt genügend Ressourcen zur Verfügung stehen, kann es zu Konflikten kommen, wenn mehrere Kunden gleichzeitig dieselbe Ressource benötigen: Wartephänomene mit Kunden, die Zugang zu einer belegten Ressource sind die Folge.

Grundlegendes Modell



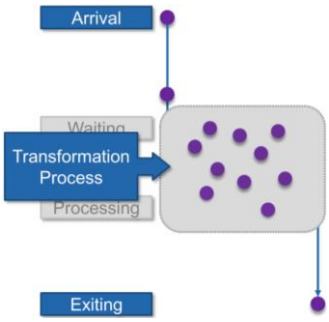
Das grundlegende Warteschlangenmodell besteht aus 4 Teilen:

- Prozess der **Ankunft** von Kunden/Objekten
- Eine **Warteschlange** von angekommenen Kunden/Objekten, die darauf warten, bearbeitet zu werden
- Eine **Verarbeitungseinheit** (Server, Schalter, Arbeitsplatz, ...)
- Prozess des **Verlassens**
- Der Prozess des Verlassens ist interessant, wenn weitere Arbeitsstationen, Server, ... berücksichtigt werden (siehe Variabilität)

Leistungsmessungen und Little's Law

- Wir wollen für eine sehr allgemeine Klasse von Systemen interessante/relevante langfristige Leistungsmaße ableiten. Wir nehmen eine weitere Abstraktion für den Prozess vor und betrachten ihn als Blackbox.

Transformations-/Umwandlungsprozess



Wir betrachten sehr allgemeine Transformationsprozesse:

- Wir nehmen an, dass nur ein Objekt pro Zeiteinheit am System ankommt
- Im Prozess wird das Eingangsobjekt in einen Ausgang transformiert
- Nach der Transformation verlassen die Objekte das System
- **Beispiel für Transformationsprozess:** Kranke Patienten (Objekte) kommen ins Krankenhaus (System) und werden behandelt (transformiert).

Folgende Leistungskennzahlen sind für Entwurf eines Systems interessant

- **Umwandlung/Transformation:**
  - **Wie viele Objekte** können innerhalb eines bestimmten **Zeitfensters maximal transformiert** werden?
  - Oder auf **lange Sicht**?
- **Zeit:**
  - **Wie viel Zeit** wird im System **durchschnittlich** für die **Durchführung** der Transformation **benötigt**?
  - Was ist die **maximale/minimale Transformationszeit**?
- **Objekte:**
  - Wie hoch ist die **durchschnittliche Anzahl** der **Objekte** im System?
  - Was ist die **maximale/minimale Anzahl** von **Objekten** im System?

Vier wichtige Messgrößen für die Leistung eines Prozesses

Zeichen	Bedeutung	Einheit
• $\boldsymbol{\lambda}$	Durchschnittliche Ankunftsrate	$\frac{1}{T} = \frac{1}{Zeit}$, $\frac{Anzahl\ Objekte}{Zeit}$
• $\boldsymbol{\lambda'}$	Durchschnittliche Abgangsrate	$\frac{1}{T} = \frac{1}{Zeit}$, $\frac{Anzahl\ Objekte}{Zeit}$
• $\boldsymbol{N}$	Durchschnittliche Anzahl Objekte im System (:= warten und bearbeiten)	Einheitslos oder Anzahl
• $\boldsymbol{W}$	Durchschnittliche Aufenthaltszeit eines Objekts im System	$T = Zeit$

Fliessgleichgewicht, wenn $\lambda = \lambda'$ gleich sind.
Transient = dynamisch/instationär (Einschwingvorgang), intransient = stationär/im Gleichgewicht

Vorhandensein von Leistungsmaßen
Wir gehen davon aus, dass der Prozess bei $t_0 = 0$ beginnt:

Zeichen	Bedeutung	Einheit
• $\boldsymbol{\lambda_t}$	Durchschnittliche Ankunftsrate **im Intervall** $[\mathbf{0}, \boldsymbol{t}]$	$\frac{1}{T} = \frac{1}{Zeit}$, $\frac{Anzahl\ Objekte}{Zeit}$
• $\boldsymbol{\lambda_t'}$	Durchschnittliche Abgangsrate **im Intervall** $[\mathbf{0}, \boldsymbol{t}]$	$\frac{1}{T} = \frac{1}{Zeit}$, $\frac{Anzahl\ Objekte}{Zeit}$
• $\boldsymbol{N_t}$	Durchschnittliche Anzahl der Objekte im System **im Intervall** $[\mathbf{0}, \boldsymbol{t}]$	Einheitslos oder Anzahl
• $\boldsymbol{W_t}$	Durchschnittliche Aufenthaltszeit eines Objekts im System **Intervall** $[\mathbf{0}, \boldsymbol{t}]$	$T = Zeit$

Stationärer Zustand
- Wir sind nicht an den Leistungsmaßen in der Übergangsphase interessiert. Wir suchen nach Situationen, in denen die Leistungsmaße nicht mehr von der Zeit t abhängen.
- System befindet sich dann im stationären Zustand, wenn $\lim_{t\to\infty} W_t = W$, $\lim_{t\to\infty} N_t = N$, $\lim_{t\to\infty} \lambda_t = \lambda$
- Folglich gilt sie im stationären Zustand: $\lim_{t\to\infty} \lambda_t = \lim_{t\to\infty} \lambda_t' = \lambda$

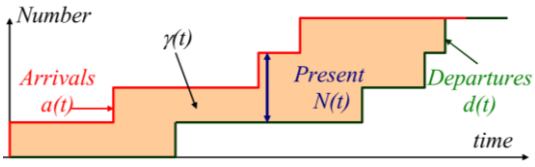
Beziehung zwischen Leistungskennzahlen: Little's Law
- Wenn ein System die Bedingungen des **stationären Zustands** erfüllt, gilt folgendes: $\boldsymbol{N = \lambda \cdot W}$, wobei:
  - $\boldsymbol{\lambda}$ ist die **langfristige** Ankunftsrate von Objekten
  - $\boldsymbol{N}$ die **langfristige** durchschnittliche Anzahl der Objekte im System ist
  - $\boldsymbol{W}$ ist die **langfristige** durchschnittliche Aufenthaltszeit im System

Beispiele für die Anwendung von Little's Law
- Die ZHAW vergibt pro Jahr durchschnittlich 70 Bachelor-Abschlüsse in WI. Ein durchschnittlicher Student braucht drei Jahre, um den Abschluss zu erreichen. Mit wie vielen Studierenden kann Richard Bödi im Durchschnitt (mindestens) rechnen im gesamten Studiengang rechnen?
  $N = \lambda \cdot W \Leftrightarrow N = 70 \cdot 3 \Leftrightarrow N = 210$
- Betrachten Sie einen Arbeitsplatz mit einer durchschnittlichen Ankunftsrate von 0,25 [Teilen/min] und einer durchschnittlichen Anzahl von 1,35 von Teilen am Arbeitsplatz. Wie viel Zeit wird im Durchschnitt benötigt für den Umwandlungsprozess?
  $\lambda = \lambda' = 0.25, \quad N = \lambda \cdot W \Leftrightarrow 1.35 = 0.25 \cdot W \Leftrightarrow 5.4 = W$

Herleitung Little's Law
Betrachten wir die folgenden zeitabhängigen Parameter:

Zeichen	Bedeutung	Einheit
• $\boldsymbol{a(t)}$	kumulative Anzahl der im Intervall $[0, t]$ **ankommenden** Objekte	Einheitslos oder Anzahl
• $\boldsymbol{d(t)}$	kumulative Anzahl von Objekten, die im Intervall $[0, t]$ **abreisen**	Einheitslos oder Anzahl
• $\boldsymbol{N(t)}$	Anzahl der zum Zeitpunkt t vorhandenen Objekte, $N(t) \neq N_t$	Einheitslos oder Anzahl
• $\boldsymbol{\gamma(t)}$	Anzahl der "Objektzeiteinheiten" im Intervall $[0, t]$	Einheitslos oder Anzahl

Bemerkung: Befindet sich ein Objekt im System im Intervall $[0{,}5]$, dann addiert es 5 Objektzeiteinheiten zu $\gamma(t)$.
- Die Fläche zwischen $a(t)$ und $d(t)$ entlang $[0, t]$ ist $\gamma(t)$.
- $N(t) = a(t) - d(t), \quad \lambda_t = \frac{a(t)}{t}, \quad W_t = \frac{\gamma(t)}{a(t)}, \quad N_t = \frac{\gamma(t)}{t}$
- $N_t = \lambda_t W_t, \qquad N = \lim_{t\to\infty} N_t = \lim_{t\to\infty} \lambda_t W_t = \lim_{t\to\infty} \lambda_t \cdot \lim_{t\to\infty} W_t = \lambda W$



Einige Bemerkungen
- Von nun an gehen wir davon aus, dass sich die betrachteten Systeme im stationären Zustand befinden. Die Leistungsmaße sind (daher) unabhängig von der Zeit.
- Der **Kehrwert** der durchschnittlichen Ankunftsrate λ ist die durchschnittliche Zeit zwischen zwei Ankünften, d. h. die durchschnittliche **Zwischenankunftszeit (IAT – inter-arrival-time)**.
- **Im Operations Management wird die folgende Notation bzw. Formulierung verwendet:**
  - $\lambda \rightarrow TH$: Durchschnittliche Leistung
  - $N \rightarrow WIP$ Durchschnittliche Objekte in Arbeit
  - $W \rightarrow CT$ Durchschnittliche Zyklusdauer
  - $WIP = TH \cdot CT \rightarrow$ Little's Law
- Little's Law gibt keine Auskunft über absolute Größe von N und W. Es sagt nur, dass sie **proportional** sind.

Weitere Leistungskennzahlen: Kapazität (capacity)

- Wir betrachten die Kapazität als quantitative Leistung.
- Die Kapazität eines Prozesses (Ressource) in Bezug auf einen bestimmten Objekttyp (Produkt) entspricht **maximal möglichen durchschnittlichen Durchsatz** TH^{max} des Prozesses in Bezug auf den Objekttyp.
- Masseinheit der Kapazität: [Anzahl der Objekte / Zeiteinheit]
- Im **stationären Zustand** gilt per Definition immer: $\lambda \leq TH^{max}$

Weitere Leistungskennzahlen: Auslastung (utilization)

- Die Auslastung u eines Prozesses (einer Ressource) ist allgemein definiert durch: $u = \frac{Verwendete\ Kapazität}{Verfügbare\ Kapazität}$
- Masseinheit für die Auslastung: einheitenlos (z. B. %)
- Die Auslastung ist der Prozentsatz der verfügbaren Zeit, zu der eine Ressource genutzt wird.
- Die Auslastung u eines Prozesses (einer Ressource) in Bezug auf einen bestimmten Objekttyp ist gegeben durch: $u = \frac{\lambda}{TH^{max}}$, **da im stationären Zustand** $\lambda \leq TH^{max}$ **gilt, ist** u **immer** $u \leq 1$

Zeichen	Bedeutung	Einheit
• TH^{max}	Kapazität, maximal möglichen durchschnittlichen Durchsatz des Prozess	$\frac{1}{T} = \frac{1}{Zeit}$, $\frac{Anzahl\ Objekte}{Zeit}$
• u	Masseinheit für die Auslastung eines Prozesses	einheitenlos oder %

Warteschlangentheorie

- Die Warteschlangentheorie ist eine Stufe unter dem Little'schen Gesetz, sie ist weniger **aggregiert**.
- Mit der Warteschlangentheorie verstehen wir die **Variabilität** der **Ankunfts**- und **Serviceprozesse** und sind in der **Lage**, Fragen wie die **Wahrscheinlichkeit**, länger als 3 Minuten zu warten, zu beantworten.
- Wir erhalten **tiefere Einblicke** in das Verhalten unserer Systeme.
- Wir verwenden die folgenden Warteschlangen-Notationen:

Zeichen	Bedeutung	Einheit
• N	langfristige durchschnittliche **Anzahl** von Objekten im **System**	$\frac{1}{T} = \frac{1}{Zeit}$, $\frac{Anzahl\ Objekte}{Zeit}$
• N_q	langfristige durchschnittliche **Anzahl** von Objekten in **Warteschlange**	$\frac{1}{T} = \frac{1}{Zeit}$, $\frac{Anzahl\ Objekte}{Zeit}$
• W_q	langfristige durchschnittliche **Wartezeit** in der **Warteschlange**	$T = Zeit$
• u	Masseinheit für die Auslastung eines Prozesses	einheitenlos oder %

Grundlegendes Modell

- Oft **fehlendes Wissen** über den genauen **Ankunftsprozess**
- Oft **fehlendes Wissen** über die genauen **Bearbeitungszeiten** für jedes Objekt/jeden Kunden.
- Dennoch **wissen** wir **vielleicht** etwas über die **Verteilung** und können daher:
  - **Informationen** über **Ankunftsschwankungen** nutzen
  - **Informationen** über **Verarbeitungsschwankungen** nutzen
- **Warteschlange** kann als Ergebnis der **Interaktion zwischen** den beiden stochastischen **Prozessen** betrachtet werden:
  - Prozess der Ankunft
  - Bearbeitung von Objekten/Kunden

Definition eines Warteschlangensystems

Um ein Warteschlangensystem zu definieren, müssen die folgenden Elemente beschrieben werden:

Serviceorientiert

- **Ankunftsprozess** der Kunden
- **Serviceprozess**
- **Anzahl Server** und ihre **Bedienungsrate**
- **Anzahl Plätze** im System
- **Bevölkerungsbeschränkungen**
- **Warteschlangendisziplin** der Kunden

Herstellung/Industrie

- **Ankunftsprozess** der Objekte/Teile/Aufträge
- **Verarbeitungseinheit**
- **Anzahl Arbeitsstationen** und deren **Verarbeitungsrate**
- **Anzahl Plätze** im System
- **Gesamtzahl** der zu berücksichtigenden **Objekte**
- **Reihenfolge** der **Bearbeitung**

→ **Im Grunde sind beide Systembeschreibungen identisch!**

Kendall: Eine Standard-Warteschlangen-Notation

Warteschlangensystem wird durch folgende Parameter beschrieben (Kendall-Notation): $A|B|c|N^{max}|K|P$

Zeichen	Bedeutung	Einheit
• A	**Verteilung** der **Zwischenankunftszeit**	
• B	**Verteilung** der **Service-Zeit/Bearbeitung**	
• c	**Anzahl** der **parallelen Server**	Einheitslos oder Anzahl
• N^{max}	**Anzahl** der **Plätze** im **System (Anzahl Server/Schalter + Warteplätze)**	Einheitslos oder Anzahl
• K	**Größe** der **Population**	Einheitslos oder Anzahl
• P	**Warteschlangendisziplin**	

Für die **Verteilung** der **Zwischenankunfts**- und **Servicezeit** A bzw. B gibt es gängige Typen:

- M steht für **exponentiell** oder **Markov**
- D deterministisch/konstant (nicht zufällig)
- E_k Erlang der Ordnung k
- N Normalverteilung
- G Allgemeine (willkürliche) Verteilung, kann jede der oben genannten einschließen

Manchmal werden N^{max}, K, und/oder P weggelassen. Es wird dann angenommen, dass

- N^{max} **unendlich** • K **unendlich** oder zumindest **groß genug** • P First in first out (FIFO)

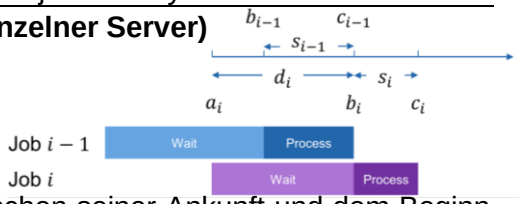
Beispiele

- $M|D|2$**:** Markov'scher Ankunftsprozess, deterministischer Serviceprozess, 2 Arbeitsstationen
- $M|G|1|4$**:** Markov'scher Ankunftsprozess, allgemeine Verteilung für die Bearbeitungszeit, eine Arbeitsstation und vier Plätze (1 am Arbeitsplatz, 3 in der Warteschlange)

Einige elementare Erkenntnisse:

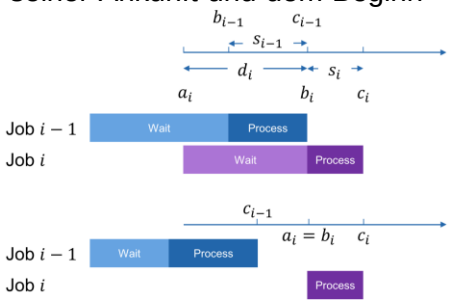
- Wenn N^{max} **endlich** ist, dann **bleibt** die **Anzahl** der **Objekte** in einem **System endlich.**
- Sei λ die Ankunftsrate und μ die Verarbeitungsrate der Objekte. Wenn N^{max} unendlich ist, aber **Ankunftsrate größer** ist **als Verarbeitungsrate** ($\lambda \geq \mu$), dann wächst Anzahl Objekte im System **unaufhaltsam**.

Wie kommt es zu Verzögerungen bei Aufträgen im Allgemeinen? (einzelner Server)



- Sei a_i, a_{i-1} die **Ankunftszeit** von Aufträgen i und $i-1$
- Sei s_i, s_{i-1} die **Bearbeitungszeit** der Aufträge i und $i-1$
- Sei b_i, b_{i-1} der **Beginn** der **Bearbeitung** von Aufträgen i und $i-1$
- Sei c_i, c_{i-1} das **Ende** der **Bearbeitung** der Aufträge i und $i-1$
- Sei d_i die **Verzögerung** des **Starts** von Auftrag i, d. h. die Zeit zwischen seiner Ankunft und dem Beginn der Verarbeitung (**Wartezeit**).

Daraus folgt:



- Wenn **Ankunft** von Auftrag i vor der Beendigung von Auftrag $i-1$ liegt, dann muss Auftrag i warten, bis Auftrag $i-1$ beendet ist, d. h.
- **Wenn**, $a_i < c_{i-1}$ **dann**
  - $d_i = c_{i-1} - a_i$ ○ $b_i = c_{i-1}$ ○ $c_i = a_i + d_i + s_i$

  Sonst
  - $d_i = 0$ ○ $b_i = a_i$ ○ $c_i = a_i + s_i$

Beispiel Wenn $a_1 = 15$, $a_2 = 47$ und $s_1 = 43$. Berechne folgende Angaben:

- $d_1 = 0$, • $c_1 = 58$, • $d_2 = 11$, • $b_2 = 58$, • $s_2 = x$, • $c_2 = 58 + x$

Statistik

- $\bar{r} = \frac{a_n}{n}$ **Durchschn. Zwischenankunftszeit** • $\bar{s} = \frac{1}{n}\sum_i s_i$ **Durchschn. Bearbeitungszeit**
- Abgeleitet von der **durchschnittlichen Ankunftsrate** $\bar{\lambda}$: **Inverse** von $\bar{r}$, d.h. $\bar{\lambda} = \frac{1}{\bar{r}}$
- Abgeleitet von der **durchschnittlichen Bearbeitungsrate** $\bar{\mu}$: **Inverse** von $\bar{s}$, d.h. $\bar{\mu} = \frac{1}{\bar{s}}$

Beispiel – Berechnen Sie die

i	1	2	3	4	5	6	7	8	9	10
ai	15	47	71	111	123	152	166	226	310	320
di	0	11	23	17	35	44	70	41	0	26
si	43	36	34	30	38	40	31	29	36	30

- Durchschnittliche Zwischenankunftszeit: $\bar{r} = \frac{a_n}{n} = \frac{320}{10} = 32$
- Durchschnittliche Bearbeitungszeit: $\bar{s} = \frac{1}{n}\sum_i s_i = \frac{1}{10}347 = 34.7$
- Ankunftsrate: $\bar{\lambda} = \frac{1}{\bar{r}} = \frac{1}{32} = 0.03125$
- Bearbeitungsrate: $\bar{\mu} = \frac{1}{\bar{s}} = \frac{1}{34.7} = 0.029$

Noch mehr Statistik

- **Durchschn. Verspätung** (**Wartezeit**) $\bar{d} = \frac{1}{n}\sum_i d_i = \widehat{W}_q$
- **Durchschn. Verweildauer** im **System** (Wartezeit + Bearbeitung) $\bar{w} = \frac{1}{n}\sum_i w_i = \frac{1}{n}\sum_i (d_i + s_i) = \bar{d} + \bar{s} = \widehat{W}$

Fortsetzung Beispiel: $\bar{d} = \frac{1}{n}\sum_i d_i = \widehat{W}_q = \frac{1}{10}(267) = 26.7,\ \bar{w} = \frac{1}{n}\sum_i (d_i + s_i) = \frac{1}{10}(267 + 347) = 61.4 = \widehat{W}$

Pollaczek-Khinchin-Formel, 1930

Kommen die Objekte unabhängig voneinander an einer Bedientheke (**M | G | 1 -System**) an, so errechnet sich die durchschnittliche Anzahl der Kunden N im System aus der Ankunfts- bzw. Bedienungsrate λ bzw. μ und zusätzlich der Varianz der Bedienungs-/Bearbeitungszeiten.

$\boldsymbol{N = u + \frac{u^2 + \lambda^2 Var(S)}{2(1-u)}}$ wobei

- $\boldsymbol{\lambda}$ **Ankunftsrate**
- $\boldsymbol{\mu}$ **Bearbeitungsrate**, $\frac{1}{\mu}$ ist die durchschnittliche Bearbeitungszeit
- $\boldsymbol{u = \frac{\lambda}{\mu}}$ ist **Auslastung** (erforderlich $u < 1$ → stationärer Zustand!!!)
- $\boldsymbol{Var(S)}$ die **Varianz** der **Bearbeitungszeitverteilung**

Bemerkungen

- Wir können Warteschlangensystem beschreiben, bevor wir irgendwelche Werte beobachten, wenn wir etwas über die Verteilungen wissen.
- Die Durchschnittswerte hängen nur
  - von der **durchschnittlichen Bedienungszeit** $\boldsymbol{\frac{1}{\mu} = E(S)}$
  - der **durchschnittlichen Ankunftsrate** $\boldsymbol{\lambda}$
  - und die **Varianz** $\boldsymbol{Var(S)}$ der **Bedienungszeitverteilung**

  aber **nicht** auf **höhere Momente**!
- **Durchschnittswerte steigen linear** mit der **Varianz**
- Mehr **Zufälligkeit** (höhere Varianz) in Servicezeiten führt zu **höheren Wartezeit/Warteschlangenlänge**

M|M|1-Warteschlangen

Systemeinstellung für tiefergehende Analysen eines Kundenservice-Schalter:

- **Ankunft:** Kunden kommen am Service Desk an
- **Warten:** Kunden warten, um bedient zu werden
- **Bedienung:** Kunden werden bedient
- **Verlassen:** Kunden verlassen das System

➔ Analyse nur des Einzelarbeitsplatzes "Service Desk"

Ankunftsprozess

- Zwei Gesichtspunkte:
  - Zeit zwischen zwei Ankünften (**Zwischenankunftszeiten**): Folgen sie einer **statistischen Verteilung**?
  - **Anzahl Ankünfte** innerhalb eines bestimmten **Zeitraums**: Folgt sie einer **statistischen Verteilung**?
- Beide Sichtweisen verlangen nach Statistiken, doch
  - Zwischenankunftszeiten: **kontinuierlich**
  - Anzahl der Ankünfte: **diskret**
- Bezeichne $\boldsymbol{\lambda(t)}$ die Ankunftsrate (Anzahl der Kunden pro Zeiteinheit), die von der Zeit $\boldsymbol{t}$ abhängen kann.

Ankunftsprozess: Zwischenankunftszeiten

- Zum Zeitpunkt 0 startet das System. Frage: Wie lange dauert es bis zur ersten Ankunft?
- $\boldsymbol{T}$ sei der Zeitpunkt der ersten Ankunft
- Suche nach der Wahrscheinlichkeitsverteilung von $\boldsymbol{P(T \leq t)}$
- Verwandte Frage: Wie groß ist Wahrscheinlichkeit, dass nächste Ankunft innerhalb Intervalls $I = [t, t+h]$?

  ➔ Wahrscheinlichkeit der ersten Ankunft ist **exponentialverteilt** mit Parameter $\boldsymbol{\lambda}$ ➔ $\boldsymbol{P(T \leq t) = 1 - e^{-\lambda t}}$
- Wir müssen nicht nur die erste Ankunft betrachten. Ein ähnliches **Argument zeigt** schließlich, dass die **Zwischenankunftszeiten exponentialverteilt** sind, **wenn** einige **Annahmen erfüllt** sind:
  - **Ankunftsrate konstant** über die Zeit ($\boldsymbol{\lambda(t) = \lambda}$) und
  - **Unabhängigkeit** der **Ankünfte**: Neue Ankunftsereignisse hängen nicht von vergangenen Ankünften ab

Poisson-Prozess

- Markov'sche Ankunftsprozesse erfordern (abgekürzt):
  - Die **Wahrscheinlichkeitsverteilungen** der **Anzahl** der **Ankünfte** in einem bestimmten **Zeitintervall** hängen von der **Länge** des **Intervalls**, aber nicht von seinem **Startpunkt** ab.
  - Die **Wahrscheinlichkeit** für **genau eine Ankunft** innerhalb eines kurzen **Intervalls** kann von der **Anzahl** der **bereits beobachteten Ankünfte abhängen.**
- **Poisson-Ankunftsprozess** ist Spezialfall Markov-Prozesses, bei dem die Wahrscheinlichkeit einer Ankunft in einem kleinen Zeitintervall unabhängig von der Anzahl der bereits beobachteter Ankünfte ($\boldsymbol{\lambda_n = \lambda}$) ist.

Ankunftsprozess: Anzahl der Ankünfte

- Betrachtet nur das System "Ankunft": Keine Abgänge, also steigt die Anzahl der Teile im System "Ankunft"
- Sei $\boldsymbol{a(t)}$ die Anzahl der angekommenen Teile bis zum Zeitpunkt $\boldsymbol{t}$.
- $\boldsymbol{P(a(t) = n)}$ bezeichne die Wahrscheinlichkeit, dass bis $\boldsymbol{t}$ $\boldsymbol{n}$ Ankünfte erfolgt sind.
- Was ist nun $\boldsymbol{P(a(t+\epsilon) = n)}$?
- Wie hoch ist die Wahrscheinlichkeit, dass zwischen $\boldsymbol{t}$ und $\boldsymbol{t+\epsilon}$ keine Ankunft erfolgt?
- Sei $\boldsymbol{\lambda(t)}$ die Ankunftsrate zum Zeitpunkt $\boldsymbol{t}$. Dann gilt für kleine $\boldsymbol{\epsilon}$: $\boldsymbol{P(a(t+\epsilon) - a(t) \geq 1) = \lambda(t) \cdot \epsilon + o(\epsilon)}$
  Nur eine Ankunft im Intervall t und $t + \epsilon$, $\boldsymbol{o(\epsilon)}$ bedeutet einen Ausdruck, der schneller als linear gegen 0 geht
- Nehmen wir an, dass die Ankunftsrate unabhängig von der Zeit $\boldsymbol{t}$ ist, also ($\boldsymbol{\lambda(t) = \lambda}$) für alle t
- Wenn $\epsilon \to \infty$ ergibt sich daraus $\boldsymbol{P(a(t) = n) = \frac{(\lambda t)^n}{n!} e^{-\lambda t}}$ → Die Anzahl der Ankünfte $\boldsymbol{a(t)}$ im System "Ankunft" im Zeitintervall [$\boldsymbol{0}$, $\boldsymbol{t}$] folgt also einer **Poisson-Verteilung**.

Zusammenfassung:

- Wenn die Ankünfte
  - **unabhängig** voneinander einzeln und
  - unabhängig von der Zeit sind, ($\boldsymbol{\lambda(t) = \lambda}$) für alle $\boldsymbol{t}$

  → dann ist Wahrsch'keit, dass $\boldsymbol{n}$ Kunden bis zum Zeitpunkt $\boldsymbol{t}$ eintreffen, **poissonverteilt** mit Parameter $\boldsymbol{\lambda}$
  → die **Zwischenankunftszeit** ist **exponentialverteilt** mit dem **Parameter** $\boldsymbol{\lambda}$.
- Ein bisschen allgemeiner: Wenn die Zwischenankunftszeit exponentialverteilt ist mit dem Erwartungswert $1/\lambda$, dann ist die Anzahl der Ankünfte ein Poisson-Prozess mit dem Parameter λ und umgekehrt.
  → Dies erklärt M | M | 1

Repetition Verteilungen

Poisson-Verteilung: $\boldsymbol{N \sim Poi(\lambda)}$, für $\boldsymbol{\lambda \geq 0}$

- $\boldsymbol{P(N = n) = \frac{\lambda^n}{n!} e^{-\lambda}}$ für $\boldsymbol{n = 0, 1, 2, \ldots}$
- **Erwartungswert** $\boldsymbol{E[N] = \lambda}$
- **Varianz** $\boldsymbol{Var(N) = \lambda}$
- **Standardabweichung** $\boldsymbol{\sigma(N) = \sqrt{\lambda}}$

Exponential-Verteilung $\boldsymbol{T \sim exp(\lambda)}$ für $\boldsymbol{\lambda \geq 0}$

- $\boldsymbol{P(T \leq t) = 1 - e^{-\lambda t}}$ **für** $\boldsymbol{t \geq 0}$
- **Dichte:** $\boldsymbol{p(t) = \lambda e^{-\lambda t}}$
- **Erwartungswert** $\boldsymbol{E[T] = \frac{1}{\lambda}}$
- **Varianz** $\boldsymbol{Var(T) = \frac{1}{\lambda^2}}$
- **Standardabweichung** $\boldsymbol{\sigma(T) = \frac{1}{\lambda}}$

➔ Exponentialverteilung ist die einzige kontinuierliche Verteilung, die Gedächtnislosigkeit besitzt!

Warten

- Informationen über den Serviceprozess.
- Wenn ein anderer Kunde gerade bedient wird, wartet der ankommende Kunde in einem Wartebereich
- Wie groß ist die Kapazität des Wartebereichs/der Warteschlange?
- Gehen die Kunden, bevor sie bedient werden?
- Was passiert, wenn die Kapazität bereits voll ausgeschöpft ist?
- Überholen Kunden andere Kunden?
- Im Allgemeinen müssen wir einige Regeln und Annahmen aufstellen. Häufig verwendet: Wenn weitere Informationen weggelassen werden, dann überholen die Kunden weder, noch kehren sie um. Es wird angenommen, dass die Kapazität unendlich ist.

Bearbeitungszeiten

- Informationen zum Ablauf der Dienstleistung.
- Wie viele Kunden werden parallel bedient?
- Müssen wir den Serviceschalter jedes Mal aufräumen, bevor ein neuer Kunde bedient wird?
- Ist die Bearbeitungszeit von den verschiedenen Kundenanfragen abhängig?
- Wissen wir etwas über die Bearbeitungszeit zusätzlich zu ihrer durchschnittlichen Dauer?
- Wenn paralleler Schalter genutzt wird, hat jeder Mitarbeiter die gleiche Leistung/Geschwindigkeit/Qualität?
- Annahme: es wird jeweils nur ein Kunde bearbeitet (keine parallele Bearbeitung am gleichen Arbeitsplatz)

Exponentielle Bearbeitungszeiten

- Wenn wir die Bearbeitungszeiten als Markov'sche Prozesse behandeln, **nehmen** wir **bewusst** eine **Abweichung** in **Kauf**. Wir **approximieren** die **Servicezeiten** mit **Exponentialverteilungen** und nehmen an, dass die Servicezeiten für die Kunden (oder Objekte) unabhängig voneinander sind (jeder hat einen anderen Servicebedarf), und nehmen darüber hinaus eine **konstante Service-Rate an** (nicht abhängig von der Zeit oder der Länge der Warteschlange, ...).
- In der **Realität** werden diese **Abweichungen** oft **akzeptiert**, und Exponentialverteilungen werden üblicherweise für erste Erkenntnisse verwendet, da diese Modelle "besser" funktionieren.
- **Für die weitere Analyse gehen wir von exponentiellen Service-/Bearbeitungszeiten aus**

Verlassen

- Es gibt keine weiteren Informationen über den Ausstiegsprozess.
- Noch nicht behandelt: Die Verweildauer eines einzelnen Kunden im System wird hier gestoppt
- Beendigung des Prozesses ist wichtig, wenn eine weitere Bearbeitungs-/Servicestation folgt.
- Daraus folgt: Der Ausstiegsprozess muss berücksichtigt werden, wenn sich Arbeitsplätze/Server in einer Linie oder in einem Netzwerk befinden.

M|M|1 und M|M|c Warteschlangen: Eine und mehr als eine Servicestelle/Abarbeitungsstelle

Leistungskennzahlen im M|M|1-System Systembeschreibung

- λ ist die durchschnittliche Ankunftsrate ($\frac{1}{\lambda}$ damit die durchschnittliche Ankunftszeit, Erwartungswert Exp.-V.)
- Die **Zwischenankunftszeiten** sind **voneinander unabhängig**.
- μ ist die durchschnittliche Service-/Bearbeitungsrate μ ($\frac{1}{\mu}$ durchschnittliche Bearbeitungszeit, wieder Exp.)
- Wir haben eine Station (Bearbeitungseinheit / Serviceschalter)
- Die Kunden/Waren werden in FIFO-Reihenfolge bearbeitet
- Wir **analysieren** den **stationären Prozess**, d.h. die Anzahl der ankommenden Artikel/Kunden ist identisch mit der Anzahl der auslaufender Artikel/Kunden.

Kennzahlen für M|M|1-Systeme

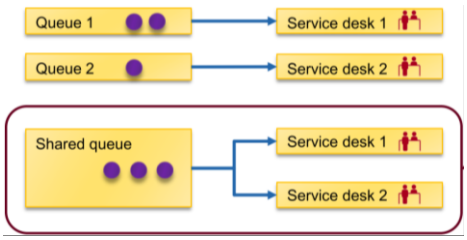
Kennzahl	Little's Law	Auslastung	Raten
$E(N)$ **Anzahl** Objekte im **System**	$\lambda \cdot E(W)$	$\frac{u}{1-u}$	$\frac{\lambda}{\mu-\lambda}$
$E(W)$ **Aufenthaltszeit** im **System**	$\frac{E(N)}{\lambda}$	$\frac{1}{1-u} \cdot \frac{1}{\mu}$	$\frac{1}{\mu-\lambda}$
$E(W_q)$ **Wartezeit** in der **Warteschlange**	$E(W) - \frac{1}{\mu}$	$\frac{u}{1-u} \cdot \frac{1}{\mu}$	$\frac{\lambda}{\mu(\mu-\lambda)}$
$E(N_q)$ **Anzahl** Objekte in der **Warteschlange**	$\lambda \cdot E(W_q)$	$\frac{u^2}{1-u}$	$\frac{\lambda^2}{\mu(\mu-\lambda)}$

Auslastung	$u = \frac{\lambda}{\mu}$
Wahrscheinlichkeit einer **Zeit im System größer als** x	$P(W > x) = e^{-\mu(1-u)x}$
Wahrscheinlichkeit der **Wartezeit größer als** x	$P(W_q > x) = u \cdot e^{-\mu(1-u)x}$
Wahrscheinlichkeit, dass sich n **Objekte** im **System** befinden	$n = 0, 1, \ldots: P(N = n) = u^n(1-u)$

Beispiel dazu:

- Im Durchschnitt kommen 15 Kunden pro Stunde an einer Station an. Wenn ein anderer Kunde gerade bedient wird, wartet der ankommende Kunde in einem Wartebereich. Dieser Wartebereich hat eine unendliche Kapazität. Die Bedienung eines Kunden dauert im Durchschnitt 3 Minuten.
- Ankunftsrate: $\lambda = 15 \frac{Personen}{Stunde}$
- Bearbeitungsrate: $\mu = 20 \frac{Personen}{Stunde}$
- Auslastung: $u = \frac{\mu}{\lambda} = \frac{15}{20} = 0.75$
- Durchschn. Anzahl der Kunden im System $E(N) = \frac{15}{20-15} = 3p$
- Durchschnittliche Verweildauer eines Kunden im System $E(W) = \frac{1}{20-15} = \frac{1}{5} Stunden = 12\ Min$
- Durchschnittliche Wartezeit eines Kunden $E(W_q) = \frac{15}{20(20-15)} = \frac{3}{20} Stunden = 9\ Min$
- Durchschnittliche Anzahl der Kunden, die in Warteschlange warten $E(N_q) = \frac{15^2}{20(20-15)} = 2.25\ Personen$
- Wie groß ist die Wahrscheinlichkeit, dass sich mehr als 2 Kunden in der Warteschlange befinden?
  Wir suchen nach $P(N_q > 2)$. Wir haben einen Serviceschalter, d.h. mehr als 2 Kunden in der **Warteschlange** ist gleichbedeutend mit mehr als 3 Kunden im **System**:
  $P(N_q > 2) = P(N > 3) = 1 - P(N \leq 3) = 1 - (P(N = 0) + P(N = 1) + P(N = 2) + (P(N = 3)) =$
  $= 1 - (1-u)(u^0 + u^1 + u^2 + u^3) = 1 - (1 - u^4) = u^4 = 0.75^4 = 0.3164$
- Wahrscheinlichkeit, dass ein ankommender Kunde innerhalb von 5 Minuten aussteigt?
  $P\left(W \leq \frac{1}{12}\right) = 1 - P\left(W > \frac{1}{12}\right) = 1 - e^{-\frac{20(1-0.75)}{12}} = 0.341$
- Wahrscheinlichkeit, dass ein ankommender Kunde höchstens 2 Minuten lang in der Warteschlange bleibt?
  $P\left(W_q \leq \frac{1}{30}\right) = 1 - P\left(W_q > \frac{1}{30}\right) = 1 - 0.75e^{-\frac{20(1-0.75)}{30}} = 0.365$

Erweiterung des Systems



Mehr als eine Station/Server

- Schema sieht fast gleich aus: Jetzt mit 2 Bedien-/Abarbeitungsstellen
- Sobald ein Kunde an einem der Tische bedient wurde, wird der nächste Kunde bedient (keine Leerlaufzeit)
- Natürliche Frage: Eine Warteschlange für 2 Service-Desks oder 2 getrennte Warteschlangen?

Leistungskennzahlen im M|M|c-System

- Die Ankunftsrate $\boldsymbol{\lambda}$ und die Abarbeitungsrate $\boldsymbol{\mu}$ sind wieder wie oben definiert.
- Die **Zwischenankunftszeiten** sind **voneinander unabhängig**.
- Wir haben $\boldsymbol{c}$ Stationen (Bearbeitungseinheiten / Leistungszähler)
- Die $\boldsymbol{c}$ Servicestationen sind identisch, d.h. die Bearbeitungsrate ist $\boldsymbol{\mu}$ für alle Stationen
- Die Kunden/Güter werden in FIFO-Reihenfolge bearbeitet.
- **Analysieren stationären Prozess**, d. h. Anzahl ankommenden Objekte = Anzahl ausgehenden Objekte

Vorbemerkungen

Wir können nicht direkt zu den Formeln übergehen; wir brauchen einige Vorbereitungen:

- Auslastung $\boldsymbol{u} = \frac{\lambda}{c \cdot \mu}$ ist durchschnittliche Auslastung **einer Station** und Auslastung **aller Arbeitsstellen**
- Daher ist: $\boldsymbol{c \cdot u} = \frac{\lambda}{\mu}$ (im Durchschnitt) die Anzahl der **aktiven/besetzten Arbeitsplätze.**
- $\boldsymbol{\zeta = P(N \geq c)}$ **(Zeta)** ist **Wahrscheinlichkeit**, dass ein ankommender **Kunde warten** muss (**Verzögerungswahrscheinlichkeit**), d.h. alle Stationen sind besetzt**.**
- **Erlang C-Formel (muss man einmal erlebt haben…):** $\zeta = \frac{\frac{(cu)^c}{c!(1-u)}}{\frac{(cu)^c}{c!(1-u)} + \sum_{k=0}^{c-1} \frac{(cu)^k}{k!}} = \frac{(cu)^c}{c!}\left((1-u)\sum_{k=0}^{c-1}\frac{(cu)^k}{k!} + \frac{(cu)^c}{c!}\right)^{-1}$
- Mit $\boldsymbol{P_0} = \left(\frac{(cu)^c}{c!(1-u)} + \sum_{k=0}^{c-1}\frac{(cu)^k}{k!}\right)^{-1}$ können wir schreiben als $\boldsymbol{\zeta = P_0 \frac{(cu)^c}{c!(1-u)}}$
  Wahrscheinlichkeit, dass Server/Mitarbeiter untätig sind (null Einträge im System)

Kennzahlen für M|M|c-Systeme

- Die Berechnung von $\boldsymbol{\zeta}$ bei $\boldsymbol{P(W_q > x)}$ führt aufgrund des Terms $\frac{(cu)^c}{c!}$ zu numerischen Problemen, weshalb wir die Funktion
  $\boldsymbol{\psi(n,x)} = \frac{\frac{x^n}{n!}}{\sum_{k=0}^{n}\frac{x^k}{k!}}$ mit $\boldsymbol{n \geq 0 \wedge n \in \mathbb{N}, x > 0}$
  und $\boldsymbol{\psi(0,x) = 1}$ definieren.
- Dies können wir zu $\boldsymbol{\zeta = \frac{u\psi(c-1,cu)}{1-u+u\psi(c-1,cu)}}$ umformen, die **rekursiv** und **numerisch stabile** Formel ist

Kennzahl	Little's Law	ζ-Formel
$\boldsymbol{E(N)}$ **Anzahl** Objekte im **System**	$\boldsymbol{\lambda \cdot E(W)}$	$\boldsymbol{\frac{u\zeta}{1-u} + cu}$
$\boldsymbol{E(W)}$ **Aufenthaltszeit** im **System**	$\boldsymbol{\frac{E(N)}{\lambda}}$	$\boldsymbol{\frac{\zeta}{c\mu(1-u)} + \frac{1}{\mu}}$
$\boldsymbol{E(W_q)}$ **Wartezeit** in der **Warteschlange**	$\boldsymbol{E(W) - \frac{1}{\mu}}$	$\boldsymbol{\frac{\zeta}{c\mu(1-u)}}$
$\boldsymbol{E(N_q)}$ **Anzahl** Objekte in der **Warteschlange**	$\boldsymbol{\lambda \cdot E(W_q)}$	$\boldsymbol{\zeta \frac{u}{1-u}}$

Auslastung	$\boldsymbol{u = \frac{\lambda}{c\mu}}$
Wahrscheinlichkeit einer **Aufenthaltszeit im System größer als x, Durchlaufzeitverteilung**	$\boldsymbol{P(W > x) = \begin{cases} (1+\zeta\mu x)e^{-\mu x} & \frac{\lambda}{\mu} = c-1 \\ e^{-\mu x}\left(1 + \frac{\zeta}{c-1-cu}\left(1 - e^{-\mu x(c-1-cu)}\right)\right), & \frac{\lambda}{\mu} \neq c-1 \end{cases}}$
Wahrscheinlichkeit Wartezeit größer als x	$\boldsymbol{P(W_q > x) = \zeta \cdot e^{-c\mu(1-u)x}}$
Wahrscheinlichkeit, dass sich $\boldsymbol{n}$ **Objekte** im **System** befinden	$\boldsymbol{n = 0, 1, \ldots c: P(N = n) = P_n = \frac{(cu)^n}{n!}P_0}$
	$\boldsymbol{n = c+1, \ldots: P(N = n) = P_n = u^n \frac{c^c}{c!}P_0}$

Beispiel M|M|c-System – Mindestanzahl von Agenten Problem

Ein Call Center verwendet die folgenden Daten:

- Durchschnittlich eingehende Anrufe: 31 in einem Zeitraum von 30 Minuten
- Durchschnittliche Bearbeitungszeit: 100sec
- Aktueller TSF-Satz (Telephone Service Factor): $x = 75\%$ und $n = 20\ Sekunden$

Fragen:

- Wie hoch ist die Mindestanzahl von Agenten, um den aktuellen TSF zu erfüllen?
- Wie viele Bearbeiter werden benötigt, um ein TSF mit $x = 95\%$ und $n = 20\ Sekunden$ zu erfüllen?
- Management erlaubt nur eine zusätzliche Mitarbeiterzahl. Wie können Sie einen TSF mit $x = 95\%$ erfüllen?

Validierung, ob M|M|c Modell gültig

- Ankunft = M: man kann unabhängige Anrufe annehmen, Annahme einer konstanten Rate ok
- Service = M: Annahme des unabhängigen Servicebedarfs ist gültig
- Agenten = c: vom Modell her ok
- Wenn Prioritätsregeln vernachlässigt werden (alle Kunden haben die gleiche Priorität), ist FIFO in Ordnung
- Unendliche Kapazität der Warteschlange: Infrastruktur kann als ausreichend groß angenommen werden
  → M/M/c Modell ist deshalb gültig → Frage jetzt in der Modellsprache: Wie groß ist c?

Anzahl der benötigten Agenten:

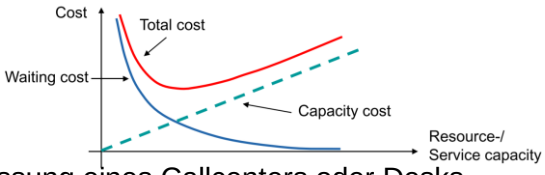
- λ und μ sind gegeben. Unbekannt ist c.
- Benutzen M|M|c-Formeln, um die Wahrscheinlichkeit zu berechnen, nicht länger als n Sekunden zu warten. Erinnerung: $P[W_q \leq x] = 1 - \zeta e^{-c\mu(1-u)x}$
- c **erhöhen**, bis TSF erfüllt ist

Lösungen: (ausprobieren mit M|M|c-Rechner in Excel)

- Wie hoch ist Mindestanzahl von Agenten, um TSF 75% zu erfüllen?
  ➔ 3 Agenten, $P(W_q \leq 20\ Sec) = 75.05\ \%$
- Wie viele Agenten werden benötigt, um TSF $x = 95\%$ und $n = 20\ Sekunden$ zu erfüllen?
  ➔ 5 Agenten, $P(W_q \leq 0\ Sec) = 98.22\ \%$
- Management erlaubt nur eine zusätzliche Mitarbeiterzahl. Wie kann man einen TSF mit $x = 95\%$ erfüllen?
  ➔ 4 Agenten, aber $n = 36\ Sec$: $P(W_q \leq 36\ Sec) = 95.03\ \%$

Abriss: Abwägung zwischen Kapazität und Kosten

Frage: Wie viele Call Agents brauchen wir, um einen guten Service zu bieten, ohne unglaublich hohe Kosten zu haben?



Service-Level

- Das Service Level ist ein Kriterium von vielen, für die Qualitätsmessung eines Callcenters oder Desks.
- Aufgaben, die «**guten Service**» **positiv/negativ beeinflussen**:
  - Unterschiedliche Fähigkeiten der Agenten
  - Rate/Schnelligkeit der abgeschlossenen Fälle
  - Umgang mit verschiedenen Kundengruppen (unterschiedliche Bedürfnisse)
- Parameter, die das **Kundenerlebnis beeinflussen**:
  - Gleichzeitige Anfragen, gleichzeitige Eingänge, Ankunftsrate
  - Nichtverfügbarkeit von Agenten, Wartezeiten
  - Anzahl der Weiterleitungsanfragen
- Weitere Faktoren: → Prioritäten

Anzahl der Agenten und ihre Einplanung Übergeordneter Bereich:

- Planung der Agenten (**wie viele? wann?**): Entscheidungsproblem mit vielen Aspekten
- Einsatzplan für Agenten so, dass:
  - das **angestrebte Serviceniveau erreicht** wird
  - die Gehälter (und andere Kosten für das Personal) nicht zu hoch sind
  - Gesetze und Verträge eingehalten werden
  - Wünsche der Agenten werden berücksichtigt, Fairness der Zeitpläne
- Komplexes Problem mit Zielkonflikten und Zeithorizonten (→ Operations Management Vorlesungen)
- Hier: Trade-off zwischen Kosten und Leistungsniveau

Modell zur Messung des Servicegrads

- Annahme, dass die Ankunftsrate λ bekannt ist (→ Nachfrage ist bekannt)
- Man nimmt an, dass die Servicerate μ oder die durchschnittliche Servicezeit bekannt ist.
- Definieren Sie das Serviceniveau durch den **Telefon-Service-Faktor** (TSF): TSF von $x\%$ für eine bestimmte Zeitschwelle n bedeutet: Mindestens $x\%$ der Kunden warten nicht länger als n Sekunden
- Übliche Werte in der Praxis
- $n = 20\ Sec - 60\ Sec\ (-180\ Sec)$
- $60\% \leq x \leq 95\%$

M|G|1 Warteschlangen

➔ Nur eine Verarbeitungsstation/Server: M|G|1, M|D|1, M|Lognorm|1, M|Gamma|1

Ausweitung von M|M|1

Es gibt mehrere Möglichkeiten, den M|M|1-Fall zu erweitern:

- Andere Messungen der Ankunfts-/Bearbeitungszeit können zu anderen Verteilungen führen
- Der Wartebereich ist nicht mehr unendlich
- Service Desk hat mehr als einen Mitarbeiter (parallele Bearbeitung) (bereits für M|M|c durchgeführt)
- Vor- und/oder nachgelagerte Stationen werden in die Analyse einbezogen
- Rüstzeit für Maschinen, die Teile bearbeiten
- Maschinenausfälle

➔ **Bearbeitungszeit nun nicht (mehr) exponentialverteilt**. → **Nun mit genereller Verteilungsfunktion**

Parameter und Annahmen für das M|G|1-System

- λ sei durchschnittliche **Ankunftsrate** ($\frac{1}{\lambda}$ durchschnittliche Zwischenankunftszeit, der Erww. E(X) v. Exp.-V.)
- Die **Zwischenankunftszeiten** sind **voneinander unabhängig**.
- Die **Service-/Bearbeitungszeit** hat eine **allgemeine Verteilung**.
- μ sei durchschnittliche **Bearbeitungsrate** ($\frac{1}{\mu}$ ist damit die durchschnittliche **Bearbeitungszeit**)
- Wir haben eine Station (Bearbeitungseinheit/Serverzähler)
- Die Kunden/Einzelstücke werden in **FIFO-Reihenfolge** bearbeitet.
- **Analysieren stationären Prozess: Anzahl ankommenden Kunden = Anzahl abgehenden Kunden**

Variationskoeffizient

Wir betrachten eine **Zufallsvariable** X und führen **Maß** für die **Größe** der **Variabilität** von X ein: Der Variationskoeffizient (CV) $\boldsymbol{cv}$ ist definiert durch: $\boldsymbol{cv = \frac{\sigma(X)}{E(X)}}$, wobei

- $\boldsymbol{E(X)}$: Erwartungswert von X
- $\boldsymbol{\sigma(X)}$: Standardabweichung von X
- Wir verwenden häufig den quadrierten Variationskoeffizienten $\boldsymbol{cv^2}$ (SCV).
- Der Variationskoeffizient der Prozesszeiten bzw. der Zwischenzeiten wird mit $\boldsymbol{cv_\mu}$ und $\boldsymbol{cv_\lambda}$ bezeichnet.
- Bei Exponentialverteilung sind $E(X) = \frac{1}{\lambda}$ und $\sigma(X) = \frac{1}{\lambda}$. Wir erhalten immer als Spezialfall: $\boldsymbol{cv = \frac{\frac{1}{\lambda}}{\frac{1}{\lambda}} = 1}$

Klassifizierung der Variabilität

- Geringe Variabilität (LV): $0 \leq cv \leq 0.75$
- Mäßige Variabilität (MV): $0.75 < cv \leq 1.33$
- Hohe Variabilität (HV): $1.33 < cv$

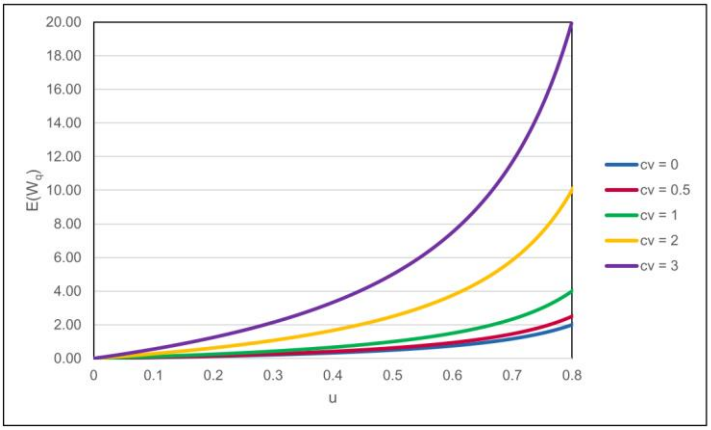
Leistungskennzahlen für M|G|1 mit Pollaczek-Khinchin-Formel mit CV

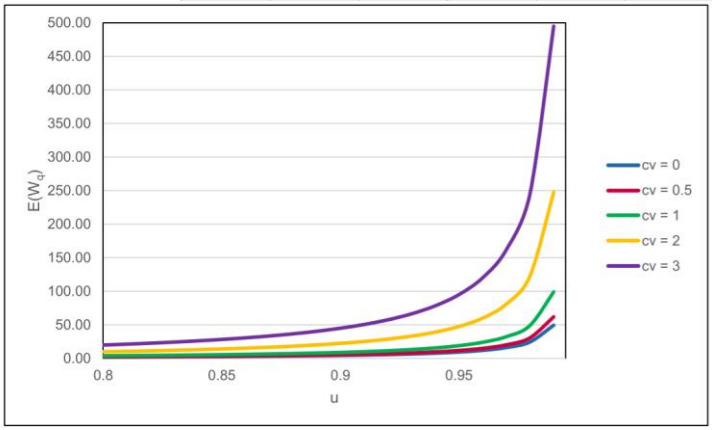
Kennzahl	Little's Law	Auslastung, $var(S)$	Auslastung, cv
$E(N)$ **Anzahl** Objekte im **System**	$\boldsymbol{\lambda \cdot E(W)}$	$\boldsymbol{u + \frac{u^2+\lambda^2 \cdot Var(S)}{2(1-u)}}$	$\boldsymbol{u + \frac{u^2(1+cv_\mu^2)}{2(1-u)}}$
$E(W)$ **Aufenthaltszeit** im **System**	$\boldsymbol{\frac{E(N)}{\lambda}}$	$\boldsymbol{\frac{1}{\mu} + \frac{\lambda \cdot \mu \cdot Var(S)+u}{2(\mu-\lambda)}}$	$\boldsymbol{\frac{u(1+cv_\mu^2)}{2(1-u)} \cdot \frac{1}{\mu} + \frac{1}{\mu}}$
$E(W_q)$ **Wartezeit** in der **Warteschlange**	$\boldsymbol{E(W) - \frac{1}{\mu}}$	$\boldsymbol{\frac{\lambda \cdot \mu \cdot Var(S)+u}{2(\mu-\lambda)}}$	$\boldsymbol{\frac{u(1+cv_\mu^2)}{2(1-u)} \cdot \frac{1}{\mu}}$
$E(N_q)$ **Anzahl** Objekte in der **Warteschlange**	$\boldsymbol{\lambda \cdot E(W_q)}$	$\boldsymbol{\frac{\lambda^2 \cdot \mu \cdot Var(S)+u \cdot \lambda}{2(\mu-\lambda)}}$	$\boldsymbol{\frac{u^2(1+cv_\mu^2)}{2(1-u)}}$

Qualitatives Verhalten von $E(W_q)$ für M|G|1-Warteschlange

Durchschnittliche Wartezeit $E(W_q)$ für verschiedene Größen der Prozesszeitvariabilität cv_μ mit Prozesszeit gleich $\frac{1}{\mu} = 1$

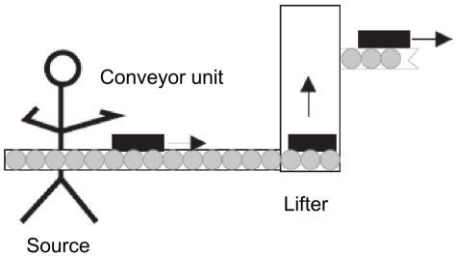
	E(W$_q$) (Process time: 1)				
u	cv = 0	cv = 0.5	cv = 1	cv = 2	cv = 3
0.5	0.5	0.625	1	2.5	5
0.8	2	2.5	4	10	20
0.9	4.5	5.625	9	22.5	45
0.95	9.5	11.875	19	47.5	95
0.99	49.5	61.875	99	247.5	495





Beispiel: Transportsysteme

Warteschlangenmodelle spielen eine wichtige Rolle bei der Analyse von Transportanlagen. Für diese Systeme sind die Prozesszeiten oft (nahezu) deterministisch. Nehmen wir an, dass die Fördereinheiten von verschiedenen unabhängigen Arbeitsstationen zum Heber geliefert werden. Eine Fördereinheit hat eine quadratische Grundfläche (40 cm x 40 cm). Die durchschnittliche Ankunftsrate beträgt 1,8 Fördereinheiten pro Minute. Der Heber braucht genau 15 Sekunden für einen Weg.



- Wie lang ist die Warteschlange vor dem Lift im Durchschnitt? $\lambda = \frac{1.8\ E}{Min}$

  Lift braucht 2x 15 Sek, um nach oben zu fahren und anschliessend wieder nach unten → $\frac{1\ E}{30\ Sec} = \frac{2\ E}{Min} = \mu$

  $u = \frac{\frac{1.8\ E}{Min}}{\frac{2\ E}{Min}} = 0.9$, Warteschlangenlänge → $E(N_q)$ Anzahl Objekte in der Warteschlange $E(N_q) = \frac{u^2(1+cv_\mu^2)}{2(1-u)}$

  Da keine Variabilität $cv_\mu^2 = 0$ → $E(N_q) = \frac{0.9^2(1+0)}{2(1-0.9)} = 4.05$, ein Paket = 40 cm → $4.05 \cdot 0.4 = 1.62\ m$

- Wie viel Zeit verbringt eine Fördereinheit im Durchschnitt in der Warteschlange vor dem Lift?

  $E(W_q) = \frac{u(1+cv_\mu^2)}{2(1-u)} \cdot \frac{1}{\mu} = \frac{0.9(1+0)}{2(1-0.9)} \cdot \frac{1}{2} = 2.25\ Min$ oder $E(N_q) = \lambda \cdot E(W_q) \rightarrow E(W_q) = \frac{E(N_q)}{\lambda} = \frac{4.05}{1.8} = 2.25$

Repetition: Gamma und Log-Normal Verteilung

Verteilung	Log-Normal: $X \sim logN(\mu, \sigma^2)$ $x > 0, \sigma > 0, \mu \in \mathbb{R}$	Gamma: $X \sim G(k, \theta)$ $x, k, \theta > 0$	Exp. $X \sim Exp(\lambda)$
Dichte	$\frac{1}{x} \cdot \frac{1}{\sigma\sqrt{2\pi}} \cdot \exp\left(-\frac{(\ln(x)-\mu)^2}{2\sigma^2}\right)$	$\frac{\theta^{-k}x^{k-1} \cdot \exp\left(-\frac{x}{\theta}\right)}{\Gamma(k)}$	$\begin{cases} \lambda e^{-\lambda x} & \mid x \geq 0 \\ 0 & \mid x < 0 \end{cases}$
Kumulative Verteilungsfunktion	$\Phi\left(\frac{\ln(x)-\mu}{\sigma}\right)$, Φ Verteilungsfunktion von $N(0,1)$	Keine allgemeine analytische Lösung, nur für $k \in \mathbb{N}$	$\begin{cases} 1 - e^{-\lambda x} & \mid x \geq 0 \\ 0 & \mid x < 0 \end{cases}$
Erwartungswert $E(X)$	$e^{\left(\mu+\frac{\sigma^2}{2}\right)}$	$k \cdot \theta$	$\frac{1}{\lambda}$
Standardabweichung σ	$e^{\left(\mu+\frac{\sigma^2}{2}\right)} \cdot \sqrt{e^{\sigma^2} - 1}$	$\sqrt{k} \cdot \theta$	$\frac{1}{\lambda}$
Variationskoeffizient cv	$\sqrt{e^{\sigma^2} - 1}$	$\frac{1}{\sqrt{k}}$	$1 = \frac{1}{\lambda} / \frac{1}{\lambda}$
Erstes Moment $E(X)$	$e^{\left(\mu+\frac{\sigma^2}{2}\right)}$	$k \cdot \theta$	$\frac{1}{\lambda}$
Zweites (raw) Moment $E(X^2)$	$e^{(2\mu+2\sigma^2)}$	$(k+1) \cdot k \cdot \theta^2$	$\frac{2}{\lambda^2}$
Drittes (raw) Moment $E(X^3)$	$e^{\left(3\mu+\frac{9}{2}\sigma^2\right)}$	$(k+2) \cdot (k+1) \cdot k \cdot \theta^3$	$\frac{6}{\lambda^3}$
n-tes (raw) Moment $E(X^n)$	-	$\frac{\Gamma(k+n)}{\Gamma(k)} \cdot \theta^n$, Γ Gammafunktion	$\frac{n!}{\lambda^n}$

Anmerkung: Der Variationskoeffizient kann jeden Wert innerhalb des Intervalls $(0, \infty)$ annehmen.

Leistungskennzahlen für spezielle M|G|1 Systeme

Kennzahl	$M\|D\|1$	$M\|Gamma\|1$	$M\|Lognorm\|1$
$E(N)$ **Anzahl** Objekte im **System**	$u + \frac{u^2}{2(1-u)}$	$u + \frac{u^2\left(1+\frac{1}{k}\right)}{2(1-u)}$	$u + \frac{u^2 \cdot exp(\sigma^2)}{2(1-u)}$
$E(W)$ **Aufenthaltszeit** im **System**	$\frac{u}{2(1-u)}\frac{1}{\mu} + \frac{1}{\mu}$	$\frac{u\left(1+\frac{1}{k}\right)}{2(1-u)} \cdot \frac{1}{\mu} + \frac{1}{\mu}$	$\frac{u \cdot exp(\sigma^2)}{2(1-u)} \cdot \frac{1}{\mu} + \frac{1}{\mu}$
$E(W_q)$ **Wartezeit** in der **Warteschlange**	$\frac{u}{2(1-u)} \cdot \frac{1}{\mu}$	$\frac{u\left(1+\frac{1}{k}\right)}{2(1-u)} \cdot \frac{1}{\mu}$	$\frac{u \cdot exp(\sigma^2)}{2(1-u)} \cdot \frac{1}{\mu}$
$E(N_q)$ **Anzahl** Objekte in der **Warteschlange**	$\frac{u^2}{2(1-u)}$	$\frac{u^2\left(1+\frac{1}{k}\right)}{2(1-u)}$	$\frac{u^2 \cdot exp(\sigma^2)}{2(1-u)}$

Verteilung der Wartezeit für M|G|1-Warteschlange

Für die Verteilungsfunktion für die Wartezeit $\boldsymbol{P(W_q \leq x)}$ und für die Durchlaufzeit $\boldsymbol{P(W \leq x)}$ gibt es **keine geschlossene analytische Lösung** wie z.B. im M|M|c-Fall. Es gibt jedoch **verschiedene Näherungen** für diese Verteilungen, hier stellen wir eine Näherung vor, die auf

- Gamma-Verteilung und
- ersten drei Momenten der Servicezeit S, d.h. $E(S), E(S^2), E(S^3)$

beruhen. Die **Approximation** ist **sehr genau** für $(cv_\mu, u) \in [0.05,4] \times [0,0.95]$.

Methode der Approximation:

- **Schritt 1:** Berechnen Sie das (genaue) erste und zweite Moment der Wartezeit W_q durch,
  $\boldsymbol{E(W_q) = \frac{u(1+cv_\mu^2)}{2(1-u)} \cdot \frac{1}{\mu}}$ und $\boldsymbol{E(W_q^2) = 2 \cdot E(W_q)^2 + \frac{\lambda \cdot E(S^3)}{3(1-u)}}$
- **Schritt 2:** Sei $\boldsymbol{F_{(k,\theta)}(x)}$ die Gamma-Verteilungsfunktion, dann $\boldsymbol{P(W_q \leq x) \approx F_{(k,\theta)}(x)}$ mit
  $\boldsymbol{k = \frac{E(W_q)^2}{E(W_q^2) - E(W_q)^2}}$, $\boldsymbol{\theta = \frac{E(W_q^2) - E(W_q)^2}{E(W_q)} = \frac{E(W_q)}{k}}$

Anmerkungen:

- Approximation ist **analytischer Ausdruck** für $\boldsymbol{P(W_q \leq x)}$ unter Verwendung der Gamma-Verteilung. Sie ist **schnell implementiert** und die besprochenen Vorteile der analytischen Ansätze bleiben erhalten.
- Konstruktionsbedingt ist der **Mittelwert** der durch die **Approximationsverteilung** $\boldsymbol{F_{(k,\theta)}(x)}$ repräsentierten Zufallsvariablen X die **wahre durchschnittliche Wartezeit** $\boldsymbol{E(W_q)}$.

G|G|c Warteschlangen

- Aufbau auf M|G|1 Systemen
- Weiterhin **nicht exponentiell** verteilte **Ankunftszeit** und **nun mehr als ein Server/Maschine**

Parameter und Annahmen für das G|G|c-System

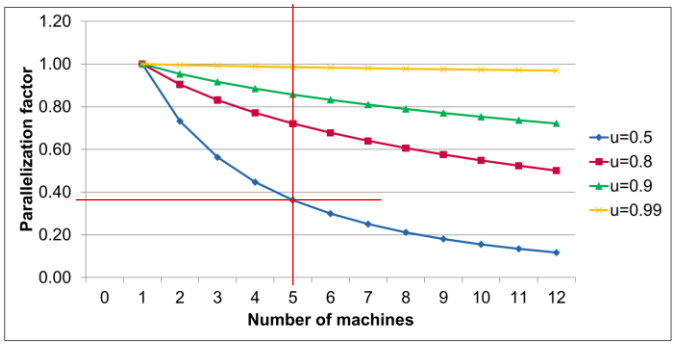
- Die **Ankunftszeit** hat eine **allgemeine Verteilung**.
- λ sei durchschnittliche **Ankunftsrate** ($\frac{1}{\lambda}$ durchschnittliche Zwischenankunftszeit)
- Die **Zwischenankunftszeiten** sind **voneinander unabhängig**.
- Die **Service-/Bearbeitungszeit** hat eine **allgemeine Verteilung**.
- μ sei durchschnittliche **Bearbeitungsrate** ($\frac{1}{\mu}$ ist damit die durchschnittliche **Bearbeitungszeit**)
- Wir haben eine oder **mehrere Station (**Bearbeitungseinheit/Serverzähler)
- Die Kunden/Einzelstücke werden in **FIFO-Reihenfolge** bearbeitet.
- Wir **analysieren** den **stationären Prozess**, d.h. die **Anzahl** der **ankommenden Kunden** ist identisch mit der **Anzahl** der **abgehenden Kunden**.

Näherungen für G|G|c-Warteschlangen

- Für die M|G|c- bzw. G|G|c-Warteschlange mit (c > 1) gibt es (im Allgemeinen) **keine geschlossene analytische Lösung** für die Leistungsmaße $E(N), E(W), E(W_q), E(N_q)$ wie z.B. im M|M|c oder im M|G|1 Fall.
- Es gibt **zwei** verschiedene **Approximationen**:
  - 1. Approximation auf der Grundlage von **Kingman**
  - 2. Approximation auf der Grundlage von **Whitt, Langenbach und Krämer**
- Es gibt **viele** verschiedene **spezielle Approximationen** für **ausgewählte Verteilungen** (z. B. M|D|c, M|Lognorm|c,...). Im **Allgemeinen verwenden** wir die **spezielle Formel**, **wenn** es eine **gibt**.

Annäherungen für die G|G|c-Warteschlange auf der Grundlage von Kingman

- Kingman liefert die folgende Näherungsformel für die G|G|c-Warteschlange:
  $\boldsymbol{E(W_q) = \frac{(cv_\lambda^2 + cv_\mu^2)}{2}\left(\frac{u}{1-u}\right)\cdot\left(u^{\left(\sqrt{2(c+1)}\right)-2}\right)\cdot\frac{1}{c\mu}}$
- Die anderen drei Maße $E(N), E(W), E(N_q)$ berechnet nach Little's Law.
- Die Kingman-Approximation lässt sich leicht in eine Tabellenkalkulation implementieren.
- Intuition : **VUT-Formel** $\boldsymbol{E(W_q) = Variabilität \cdot Auslastung(Utiliziation) \cdot Zeit(Time)}$
- Die **Approximation** ist **exakt** für **M|G|1**, **aber nicht für M|M|c** (c > 1).
- Die **Approximation** ist eine **bewiesene obere Schranke** für **G|G|1-Warteschlangen**, numerische Tests unterstützen die Vermutung, dass sie auch eine **obere Schranke** für **G|G|c-Warteschlangen** ist.
- Gute Approximation für $u \in [0.7, 0.95]$ und $cv_\lambda, cv_\mu \in [0, 3]$ (Annäherung an Schwerverkehr).
- **Analyse der Kingman Formel:** Würden wir statt einer Maschine mit Bearbeitungszeit $\frac{1}{\mu}$ mehrere Maschinen (c) parallel nehmen, die dann eine c-mal größere Bearbeitungszeit haben, dann beschreibt der neue Faktor $\boldsymbol{u^{\left(\sqrt{2(c+1)}\right)-2}}$ den **Effekt** der **Parallelisierung**.
- Graph des **Parallelisierungsfaktors** $\boldsymbol{u^{\left(\sqrt{2(c+1)}\right)-2}}$ für verschiedene Auslastungsgrade → → →
- Wenn man nun beispielsweise 5 Maschinen anstatt einer hat und diese neuen Maschinen fünfmal langsamer sind als eine alte, ist die **Wartezeit** bei einer **Auslastung** von beispielsweise **50 %** (blaue Linie in Grafik) noch **38 %** von **ursprünglicher Wartezeit**.



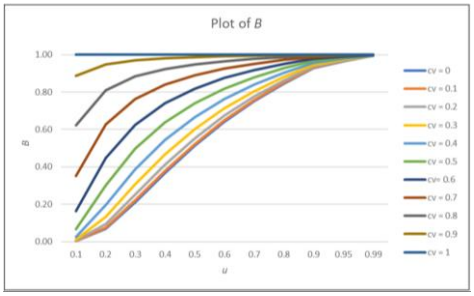
Beispiel Parallelisierungsfaktor

ZHAW hat in einen neuen Kopierer investiert, der 3-mal schneller ist als die drei alten Kopierer. Der neue Kopierer hat eine durchschnittliche Auslastung von 50%. Die Studierenden beklagen sich über lange Wartezeiten aufgrund der grossen Schwankungen im Prozess. Die alten Geräte befinden sich noch im Lager der ZHAW. Wie stark könnten Sie die Wartezeit reduzieren, wenn Sie das neue Gerät durch 3 der alten Geräte ersetzen?

- Auslastung $u = 0.5$
- cv_μ^2 bei beiden Maschinentypen gleich (Annahme).
- Anzahl Maschinen: neue Maschine $c_1 = 1$, drei alte Maschinen $c_2 = 3$ → M|G|1 vs. M|G|3
- Parallelisierungsfaktors c_1: $\boldsymbol{u^{\left(\sqrt{2(1+1)}\right)-2} = 0.5^{(2-2)} = 1}$ und c_2: $\boldsymbol{u^{\left(\sqrt{2(3+1)}\right)-2} = 0.5^{\sqrt{8}-2} = 0.5631}$
  → Bei einer Maschine bleibt der Pararellisierungsfaktor bei 1 und somit gibt es keinen Effekt, während bei drei alten Maschinen sich die Wartezeit um $\boldsymbol{1 - 0.5631 = 0.4369}$ verkürzt (ca. $-44\,\%$).
- Die Verringerung ist durch den Parallelisierungsfaktor gegeben.
- Wenn die durchschnittliche Auslastung mehr als 90 % beträgt, hilft die Ersetzung nicht allzu viel.

Approximation nach Whitt, Langenbach, Krämer

- Die größte Quelle von Ungenauigkeiten in der Approximationsformel von Kingman für die Leistungsmaße der G|G|c-Warteschlange liegt in der einheitlichen Behandlung von Ankunftszeiten mit Variationskoeffizienten cv_λ kleiner bzw. größer als 1. Deshalb Einführung Korrekturterm B, der für beide Fälle unterschiedlich berechnet wird $(0 < u < 1)$
  $$\boldsymbol{B = \begin{cases} \exp\left(-\frac{2(1-u)}{3u}\cdot\frac{(1-cv_\lambda^2)^2}{cv_\lambda^2+cv_\mu^2}\right), & f\ddot{u}r\ cv_\lambda \le 1 \\ 1, & f\ddot{u}r\ cv_\lambda > 1 \end{cases}}$$
- Analyse Korrekturterm B mit $\boldsymbol{cv_\mu = 1}$ und $\boldsymbol{cv_\lambda \in \{0, 0.1, \dots, 1\}}$. → →
- **Je kleiner Variationskoeffizient ist, desto mehr wird korrigiert.**
- Whitt, Langenbach und Krämer liefern die folgende Näherungsformel für G|G|c Warteschlange: $\boldsymbol{E(W_q) = B\cdot\frac{(cv_\lambda^2+cv_\mu^2)}{2}\cdot\frac{\zeta}{c\mu(1-u)}}$, wobei $\frac{\zeta}{c\mu(1-u)}$ durchschnittliche Wartezeit im M|M|c



- Die anderen drei Maße $E(N), E(W), E(N_q)$ werden wiederum nach Little's Law berechnet.
- **Approximation** ist **exakt** für **M|M|c** und **M|G|1**. **Implementierung** ist im Vergleich zur Approximation von Kingman **komplexer**, aber wir haben dennoch analytische Ausdrücke für die Leistungsmaße.
- Sehr gute Näherungsergebnisse für $u > 0.5$ und $cv_\lambda, cv_\mu \in [0,3]$. **Höhere Genauigkeit als Kingman**.
- **Implementiert** in **Queueing Network Analyzer (QNA)**, einem Softwarepaket, das in den Bell Laboratories entwickelt wurde, um **ungefähre Überlastungsmaße** für Netzwerke von **Warteschlangen** zu berechnen.
- Für die **Verteilungsfunktion** für die **Wartezeit** $P(W_q \le x)$ in der **G|G|c-Warteschlange** gibt es **keine** geschlossene **analytische Lösung** wie z.B. im M|M|c-Fall. Es gibt jedoch verschiedene Näherungen für diese Verteilungen. Ein Ansatz beruht auf der Beobachtung, dass die Verteilung der Wartezeit wie folgt angenähert werden kann $P(W_q \le x) \approx 1 - \alpha\cdot\exp(-\eta x)$, mit geeigneten Konstanten α und η

M|G|1-Warteschlangen mit Prioritäten – M|G|1|∞|∞|PQ

- Die **Ankunftszeit** hat eine **Exponentialverteilung**.
- λ sei durchschnittliche **Ankunftsrate** ($\frac{1}{\lambda}$ durchschnittliche **Zwischenankunftszeit**)
- Die **Service-/Bearbeitungszeit** hat eine **allgemeine Verteilung**.
- μ sei durchschnittliche **Bearbeitungsrate** ($\frac{1}{\mu}$ ist damit die durchschnittliche **Bearbeitungszeit**)
- Wir haben **eine Station** (Bearbeitungseinheit/Serverzähler)
- Die **Kunden/Artikel** werden nach einer **Prioritätsregel** bearbeitet
- Wir **analysieren stationären Prozess → Anzahl Ankommenden** ist identisch mit **Anzahl Abgehenden**

Annahmen und Definition der Prioritätsregeln

Die Gegenstände werden in **endlich viele**, **nummerierte Klassen** $\boldsymbol{j = A, B, \dots, n}$ unterteilt, wobei $\boldsymbol{A}$ die höchste Priorität hat. Für jede Klasse $\boldsymbol{j = A, B, \dots, n}$ definieren wir:

- $\boldsymbol{\pi_j}$ ist die **Wahrscheinlichkeit**, dass ein ankommender Auftrag zur Klasse j gehört**.** $\boldsymbol{\sum_j \pi_j = 1}$
- $\frac{1}{\mu_j}$ ist die durchschnittliche Bearbeitungszeit für Klasse j und $\boldsymbol{E(S_j^2)}$ das **zweite Moment** der Bearbeitungszeit (für eine allgemeine, möglicherweise unterschiedliche Verteilung).
- $\boldsymbol{u_j}$ ist die **Auslastung** des Systems in Bezug auf die Klasse j, sie ist gegeben durch $\boldsymbol{u_j = \pi_j \frac{\lambda}{\mu_j}}$
- $\boldsymbol{E(W_q^j)}$ durchschnittliche Wartezeit für Kunden in Klasse j. Das j bei $\boldsymbol{E(W_q^j)}$ ist ein hochgestellter Index.
- Die **durchschnittliche Bearbeitungszeit** $\frac{1}{\mu}$ und die **Auslastung** $\boldsymbol{u}$ des (gesamten) Systems in Bezug auf alle Klassen ist also $\frac{1}{\mu} = \sum_{k=A}^{n} \pi_k \frac{1}{\mu_k}$, $\boldsymbol{u = \sum_{k=A}^{n} u_k = \frac{\lambda}{\mu}}$

Wir unterscheiden **zwei Arten** von **Prioritätsregeln**:

- **Nicht-unterbrechende Prioritäten (Nonpreemtive priorities):** Wenn das System im Leerlauf ist und die Warteschlange nicht leer ist, wird das nächste Element aus der Klasse mit der höchsten Priorität ausgewählt. Wenn es in der Klasse mit der höchsten Priorität mehr als ein Element gibt, wird die FIFO-Regel innerhalb der Klasse angewendet.
- **Unterbrechende Prioritäten (Preemptive priorities):** Wenn das System im Leerlauf ist, wird das nächste Element wie im nicht-unterbrechenden Fall ausgewählt. Wenn das System nicht untätig ist und ein Element aus einer höheren Prioritätsklasse eintrifft, wird das in Betrieb befindliche Element unterbrochen und das eingetroffene Element bearbeitet. Wir gehen davon aus, dass der unterbrochene Auftrag ohne Verlängerung seiner Bearbeitungszeit fortgesetzt werden kann (Arbeitserhaltung).

Nicht-unterbrechende Prioritäten

- M|G|1|∞|∞|PQ-Warteschlange mit nicht-unterbrechender Prioritätsregel: Die **durchschnittliche Wartezeit** $\boldsymbol{E(W_q^j)}$ resp. **durchschnittliche Durchlauf-/Systemzeit** $\boldsymbol{E(W^j)}$ für Artikel der Klasse j kann berechnet werden durch $\boldsymbol{E(W_q^j) = \frac{\lambda \sum_{k=A}^{n} \pi_k E(S_k^2)}{2(1-\sum_{k<j} u_k)(1-\sum_{k\leq j} u_k)}}$, $\boldsymbol{E(W^j) = E(W_q^j) + \frac{1}{\mu_j}}$ für $\boldsymbol{j = A, B, \dots, n}$ wobei
  - $\boldsymbol{E(S_k^2)}$ zweite Moment der Verarbeitungszeit der Klasse k ist und
  - $\boldsymbol{\lambda}$ ist die durchschnittliche Ankunftsrate.
- Die anderen Maße können mit Hilfe des Little's Law berechnet werden.

Beispiel mit zwei Kundengruppen A und B

- Klasse A: $\boldsymbol{E(W_q^A) = \frac{\lambda E(S^2)}{2(1-u_A)} \leq \frac{\lambda E(S^2)}{2(1-u)} = E(W_q^{FIFO})}$
- Kunden der Klasse A müssen nicht auf Kunden der Klasse B warten, aber die Variabilität der Bearbeitungszeit bezieht sich auf alle Kunden. $\boldsymbol{E(W_q^{FIFO})}$ ist die durchschnittliche Wartezeit mit FIFO.
- Klasse B: $\boldsymbol{E(W_q^B) = \frac{\lambda \sum_{k=A}^{B} \pi_k E(S_k^2)}{2(1-u_A)(1-u_A-u_B)} = \frac{\lambda E(S^2)}{2(1-u)} \cdot \frac{1}{(1-u_A)} > E(W_q^{FIFO})}$
- $\boldsymbol{\frac{\lambda E(S^2)}{2(1-u)}}$: durchschnittliche Wartezeit des gesamten Systems (ohne Prioritäten).
- $\boldsymbol{\frac{1}{(1-u_A)}}$: Erweiterungsterm, der (>1) zusätzliches Warten aufgrund der Priorisierung der Klasse A beschreibt.
- Wenn nur eine Klasse, ist die durchschnittliche Wartezeit der Klasse A gleich der durchschnittlichen Wartezeit bei FIFO (d. h. Pollaczek-Khinchin-Formel).
- $\boldsymbol{\sum_{k=A}^{n} \pi_k E(S_k^2) = E(S^2)}$ ist das zweite Moment der durchschnittlichen Bearbeitungszeit $\boldsymbol{E(S) = \sum_{k=A}^{n} \pi_k S_k = \sum_{k=A}^{n} \pi_k \frac{1}{\mu_k}}$ aller Klassen.
- $E(W_q^{j-1}) < E(W_q^j)$ d.h. die durchschnittliche Wartezeit pro Klasse steigt für die Klassen an.
- Je höher Anteil der Objekte in Klasse A, desto geringer ist Effekt auf die durchschnittliche Wartezeit $E(W_q^A)$

Unterbrechende Prioritäten

- M|G|1|∞|∞|PQ-Warteschlange mit unterbrechender Prioritätsregel. Die **durchschnittliche Wartezeit** $\boldsymbol{E\left(W_q^j\right)}$ bzw. **durchschnittliche Durchlaufzeit** $E(W_j)$ für Elemente der Klasse j wird berechnet
  $\boldsymbol{E\left(W_q^j\right) = \frac{\lambda \sum_{k=A}^{j} \pi_k E(S_k^2)}{2(1-\sum_{k<j} u_k)(1-\sum_{k\leq j} u_k)}}$, $\boldsymbol{E(W^j) = E\left(W_q^j\right) + \frac{1}{\mu_j} \cdot \frac{1}{(1-\sum_{k<j} u_k)}}$ für $\boldsymbol{j = A, B, \dots, n}$ wobei
  - $\boldsymbol{E\left(S_k^2\right)}$ zweite Moment der Verarbeitungszeit der Klasse k ist und
  - $\boldsymbol{\lambda}$ ist die durchschnittliche Ankunftsrate.
- Die anderen Maße können mit Hilfe des Little's Law berechnet werden.

Beispiel mit zwei Kundengruppen A und B

- Klasse A: $\boldsymbol{E\left(W_q^A\right) = \frac{\lambda \pi_A E(S_A^2)}{2(1-u_A)} \leq \frac{\lambda E(S^2)}{2(1-u_A)} \leq \frac{\lambda E(S^2)}{2(1-u)} = E\left(W_q^{FIFO}\right)}$
- Gegenstände der Klasse A haben keine Interaktion mit anderen Klassen. $E\left(W_q^A\right)$ ist kleiner als im nicht-unterbrechenden Fall.
- Klasse B: $\boldsymbol{E\left(W_q^B\right) = \frac{\lambda \sum_{k=A}^{B} \pi_k E(S_k^2)}{2(1-u_A)(1-u_A-u_B)} > E\left(W_q^{FIFO}\right)}$
- Durchschnittliche Wartezeit für Objekt Klasse B (niedrigste Klasse) ist gleich wie bei nicht-unterbrechend
- Wenn es nur eine Klasse gibt, dann ist die durchschnittliche Wartezeit der Klasse A gleich der durchschnittlichen Wartezeit bei FIFO (d. h. Pollaczek-Khinchin-Formel).
- $E\left(W_q^{j-1}\right) < E\left(W_q^j\right)$ für $j = B, .., n-1$ d.h. die durchschnittliche Wartezeit pro Klasse steigt für Klassen an.
- Je höher Anteil der Objekte in Klasse A, desto **geringer** ist **Effekt** auf die durchschnittliche Wartezeit $E\left(W_q^A\right)$
- Die **durchschnittliche Wartezeit** $E\left(W_q^j\right)$ für $j = B, .., n-1$, im **unterbrechenden Fall** ist **kleiner** als im **nicht-unterbrechenden** Fall. Für $E\left(W_q^n\right)$ ist sie in beiden Fällen gleich groß.
- **Durchschn. Bearbeitungszeit** $\frac{1}{\mu_j}$ für Klasse j verlängert sich um die höheren Klassen mit Faktor $\boldsymbol{\frac{1}{(1-\sum_{k<j} u_k)}}$

Erhaltungssatz von Kleinrock

- Bei einer M|G|1|∞|∞|PQ-Warteschlange mit nicht-unterbrechenden Prioritäten und n Prioritätsklassen ergibt sich folgende Gleichung: $\boldsymbol{\sum_{k=A}^{n} \frac{u_k}{u} \cdot E\left(W_q^k\right) = E\left(W_q^{FIFO}\right)}$
- Wenn die **durchschnittliche Bearbeitungszeit** für **alle Klassen gleich** ist, lautet das Erhaltungsgesetz $\boldsymbol{\sum_{k=A}^{n} \pi_k \cdot E\left(W_q^k\right) = E\left(W_q^{FIFO}\right)}$
- **Interpretation des Erhaltungssatzes:** Die gewichtete Summe der durchschnittlichen Wartezeiten der Klassen ist unabhängig von der konkreten Priorisierung.
- Dies bedeutet, dass eine Verbesserung der durchschnittlichen Wartezeit einer Klasse aufgrund einer Priorisierung immer eine Verschlechterung der durchschnittliche Wartezeit einer anderen Klasse zur Folge hat.

Warteschlangen mit begrenzten Warteraum $M|M|c|N^{max}$

- Der Platz im System ist auf N^{max} Plätze begrenzt, d.h. nur N^{max} Kunden/Objekte sind im System erlaubt.
- Wenn das System seine Kapazität erreicht, können Kunden/Aufträge nicht mehr in das System eintreten.
- Es muss festgelegt werden, was mit den abgewiesenen Kunden/Objekten geschieht.
- N^{max} Plätze im System inklusive Bearbeitungsplätze/Serveranzahl c. $N^{max} - c$ Plätze in der Warteschlange

Blockierte Ankünfte aufgrund voll ausgelasteter Warteschlangenkapazität

- Anzahl Kunden/Objekte im System ist begrenzt. Daher hat Modell immer stabilen Zustand und die Warteschlange "explodiert" nie.
- Wir haben eine Ankunftsrate λ **zum** System und eine Ankunftsrate $\bar{\lambda}$ **ins** System. Somit ist unsere Annahme eines **zustandsunabhängigen** Ankunftsprozesses **nicht** mehr **gültig**.
- In einem **stationären System** ist $\bar{\lambda}$ der effektive Zufluss und Durchsatz des Systems
- Notation: Bezeichne λ_n die Ankunftsrate, wenn sich n Kunden/Objekte im System befinden. Bezeichne μ_n die Verarbeitungsrate der Server, wenn n Kunden/Objekte derzeit im System sind.
- **Wir gehen nicht mehr davon aus, dass $\boldsymbol{\lambda < \mu}$**
- $u = \frac{\lambda}{c\mu}$ beschreibt die **Wahrscheinlichkeit**, dass die **Server** bisher (auf lange Sicht) **ausgelastet** sind.
- Nun: $u = \frac{\lambda}{c\mu}$ ist eine **virtuelle Auslastung**, wenn **keine Kunden abgewiesen** werden.
- **Reale effektive Auslastung** beschreibt Serverbelegung $u_{real} = \frac{\bar{\lambda}}{c\mu}$. Berechnung mit effektiv Ankunftsrate $\bar{\lambda}$
- **Spezialfall** für kapazitive Systeme mit einem Server:
  $u_{real} = P(Server\ ist\ beschäftigt) = 1 - \mathrm{P}(Server\ ist\ frei) = 1 - \mathrm{P}_0$

- $P_{N^{max}}$ sei Wahrscheinlichkeit, dass sich N^{max} Objekte/Kunden im System befinden. Dann $\bar{\lambda} = \lambda(1 - P_{N^{max}})$
- $P_{N^{max}}$ ist Blockierungswahrscheinlichkeit, Wahrscheinlichkeit, dass ankommender Kunde abgewiesen wird.
- $\bar{\lambda}$ ist der **effektive Durchsatz** des Systems
- Die Kennzahlen $E(N)$, $E(N_q)$, $E(W)$ und $E(W_q)$ lassen sich im Allgemeinen wie folgt berechnen:
  - Die Wahrscheinlichkeiten P_n, n Kunden/Arbeitsplätze im System zu haben, werden berechnet durch $P_n = \frac{\lambda_{n-1}}{\mu_n} P_{n-1}$ und $\sum_{i \leq N^{max}} P_i = 1$ für alle n bis zu N^{max}
  - $E(N)$ und $E(N_q)$ kann unmittelbar mit Hilfe der Wahrscheinlichkeitsverteilung P_n abgeleitet werden: $E(N) = \sum_i i \cdot P_i$ und $E(N_q) = \sum_i \max(i - c, 0) \cdot P_i$ → (wenn $i - c$ negativ ist, wird 0 genommen…)
  - $E(W)$ und $E(W_q)$ werden mit Littel's Law berechnet. **Hinweis:** Der **Durchsatz** ist $\bar{\lambda}$

Ergebnisse für das $M|M|1|N^{max}$-Model als ersten Spezialfall

- Sei $\lambda_n = \lambda$ für alle $n = 0, \dots, N^{max} - 1$ und $\lambda_n = 0$ sonst die Ankunftsrate **zum** System (potenzielle Kunden) und sei $\mu_n = \mu$ die Bearbeitungsrate des Servers (nur ein Server)
- **Virtuelle Auslastung** $u = \frac{\lambda}{\mu}$
- Die tatsächliche **reale Auslastung** des Servers ist: $u_{real} = \frac{u - u^{1+N^{max}}}{1 - u^{1+N^{max}}}$ für $u \neq 1$ und $u_{real} = \frac{N^{max}}{1+N^{max}}$ für $u = 1$
- Die Wahrscheinlichkeit P_n, genau n Kunden im System zu haben → → → →

Kennzahl	$u \neq 1$	$u = 1$
$n = 0$	$P_0 = \frac{1-u}{1-u^{1+N^{max}}}$	$P_n = \frac{1}{1+N^{max}}$
$1 \leq n \leq N^{max}$	$P_n = u^n P_0 = \frac{1-u}{1-u^{1+N^{max}}} \cdot u^n$	
$n \geq N^{max} + 1$	$P_n = 0$	
$E(N)$ **Anzahl** Objekte im **System**	$\frac{u}{1-u} - \frac{(1+N^{max})u^{1+N^{max}}}{1-u^{1+N^{max}}}$	$\frac{N^{max}}{2}$
$E(W)$ **Aufenthaltszeit** im **System**	$\frac{E(N)}{\lambda(1-P_{N^{max}})} = \frac{E(N)}{\bar{\lambda}}$	
$E(W_q)$ **Wartezeit** in **Warteschlange**	$\frac{E(N_q)}{\lambda(1-P_{N^{max}})} = \frac{E(N_q)}{\bar{\lambda}}$	
$E(N_q)$ **Anz** Objekte in **Warteschlange**	$E(N) - \left(1 - \frac{1-u}{1-u^{1+N^{max}}}\right)$	$\frac{N^{max}}{2} - \frac{N^{max}}{1+N^{max}}$

Ergebnisse für das $M|M|c|N^{max}$-Model

- Nun ist c beliebig. Nehmen wir an, dass $c \leq N^{max}$, d.h. wir haben mindestens so viel Platz wie Server.
- Beginnen wieder mit den Wahrscheinlichkeiten, dass n Kunden im System vorhanden sind.
- Für den Multi-Server-Fall müssen wir verschiedene Fälle betrachten
  - Bis alle Server belegt sind: $n = 0,1, \dots, c$
  - Bis alle Warteplätze belegt sind: $n = c + 1, \dots, N^{max}$
  - System ausgelastet/komplett belegt: $n > N^{max}$
- Wie bei M|M|c bezeichnet $u = \frac{\lambda}{c\mu}$ die (**virtuelle**) Auslastung des Systems, da wir $\lambda < \mu$ **nicht** mehr **fordern**.
- Die Wahrscheinlichkeit P_n, genau n Kunden im System zu haben

$M\|M\|c\|N^{max}$	$u \neq 1$	$u = 1$
$n = 0$	$P_0 = \frac{1}{\sum_{i=0}^{c} \frac{(cu)^i}{i!} + \frac{(cu)^c}{c!} \sum_{i=c+1}^{N^{max}} u^{i-c}}$	
$1 \leq n \leq c$	$P_n = \frac{(cu)^n}{n!} P_0$	
$c + 1 \leq n \leq N^{max}$	$P_n = \frac{(cu)^n}{c!\, c^{n-c}} P_0$	
$n \geq N^{max} + 1$	$P_n = 0$	
$E(N_q)$ **Anz** Objekte in **Warteschlange**	$\frac{P_0(u^{c+1}c^c)}{c!(1-u)^2}\left(1 - u^{N^{max}-c+1} - (1-u)(N^{max} - c + 1)u^{N^{max}-c}\right)$	$P_0 \frac{(cu)^c}{c!} \frac{(N^{max}-c)(N^{max}-c+1)}{2}$
$E(N)$ **Anzahl** Objekte im **System**	$E(N_q) + \frac{(1 - P_{N^{max}})\lambda}{\mu}$	
$E(W_q)$ **Wartezeit** in **Warteschlange**	$\frac{E(N_q)}{(1-P_{N^{max}})\lambda}$	
$E(W)$ **Aufenthaltszeit** im **System**	$E(W_q) + \frac{1}{\mu} = \frac{E(N)}{(1-P_{N^{max}})\lambda}$	

Sonderfall keine Warteschlange: Das $M|M|c|c$-Modell

- In diesem Modell gibt es **keinen Warteraum**, die Kapazität des Systems ist gleich der Anzahl der Server
- Wenn alle Server besetzt sind, werden die neu ankommenden Kunden abgewiesen.
- Hier gibt es keinen Warteprozess, daher sind $E(N_q)$ und $E(W_q)$ beide immer 0.
- Interessante Frage: Wie groß ist Wahrscheinlichkeit, dass ein ankommender Kunde nicht ins System kann?
- Die Wahrscheinlichkeit genau n Kunden im System zu haben
- Die Wahrscheinlichkeit, dass ein ankommender Kunde blockiert wird, beträgt: $P_{N^{max}} = P_c = \frac{(cu)^c}{c!} \frac{1}{\sum_{i=0}^{c} \frac{(cu)^i}{i!}}$
- Die Erlang B-Formel ist sogar für M|G|c|c-Systeme gültig!
- $\bar{\lambda} = \lambda(1 - P_c)$ ist der **effektive Durchsatz** des Systems
- $u_{real} = \frac{\lambda}{c\mu}(1 - P_c) = \frac{\bar{\lambda}}{c\mu}$
- P_c wird rekursiv berechnet: $P_c = B(c,a) = \frac{a \cdot B(c-1,a)}{c + a \cdot B(c-1,a)}$ mit $B(0,a) = 1$ und $a = \frac{\lambda}{\mu}$ → wie Erlang C-Formel

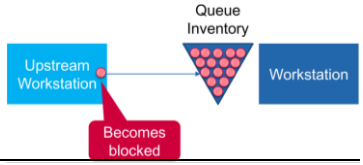
$M\|M\|c\|c$	
$n = 0$	$P_0 = \frac{1}{\sum_{i=0}^{c} \frac{(cu)^i}{i!}}$
$1 \le n \le c$	$P_n = \frac{(cu)^n}{n!} P_0$
$n \ge c + 1$	$P_n = 0$
$E(N_q)$ **Anz** Objekte in **Warteschlange**	0
$E(N)$ **Anzahl** Objekte im **System**	$\frac{(1-P_c)\lambda}{\mu}$
$E(W_q)$ **Wartezeit** in **Warteschlange**	0
$E(W)$ **Aufenthaltszeit** im **System**	$\frac{1}{\mu}$

Verallgemeinerung: Ungeduldige Ankömmlinge

- Ausgelastete Warteschlangen:
  - Die Kapazität schränkt die Aufnahme neuer Objekte/Kunden ein
  - Objekte/Kunden gehen verloren
- In Service-Situationen ist die Blockierung manchmal anders: Blockierte Kunden suchen nach alternativen Service-Möglichkeiten oder sie versuchen es später erneut
- Speziell in Dienstleistungssituationen: Wir können zwischen ungeduldigen Ankömmlingen unterscheiden
  - Kunden sind ungeduldig, weil die Warteschlange zu lang ist
  - Kunden sind ungeduldig, weil die Wartezeit zu lang ist

Balking, Reneging und Blocking

- Ein Kunde ist **balking**, wenn er sich **weigert**, sich in die **Warteschlange einzureihen**. Die Warteschlange ist nicht voll; die Wahrscheinlichkeit, dass ein Kunde sich weigert sich anzustellen, hängt von der Anzahl der Kunden in der Warteschlange $N_q(t)$ oder der erwarteten Wartezeit ab.
- Der Kunde **reneging** sich, wenn er sich in die **Warteschlange einreiht**, diese aber **vor Beginn** der Dienstleistung **verlässt**, weil die beobachtete Wartezeit zu lang geworden ist. Wenn die individuelle Wartezeit aufgebraucht ist, dann verlässt der Kunde die Warteschlange.
- Eine vorgelagerte Arbeitsstation wird blockiert und der Ankunftsprozess zu der nächsten Arbeitsstation wird unterbrochen.



Basics Variabilität

➔ Was passiert, wenn wir mehrere Stationen miteinander verketten?

Beispiel mit exponentiellen Bearbeitungszeiten

- Betrachten wir eine Linie mit zwei Stationen
- **Erste Station:** Durchschn. Bearbeitungszeit 3.25 Minuten pro Auftrag, Exponentielle Bearbeitungszeit
- **Zweite Station:** Durchschn. Bearbeitungszeit 3.0 Minuten pro Auftrag, Exponentielle Bearbeitungszeit
- **Fokus** auf den Bestand der Bohrstation: **Größe**
- Angenommen, Station 1 zieht Aufträge aus unendlichem Bestand. Folglich wird Station 1 niemals hungern.
- Angenommen, Station 2 kann Aufträge in unendliches Inventar schieben. Folglich ist Station 2 nie blockiert.

Beispiel mit exponentiellen Bearbeitungszeiten → Inventar/Wartebereich der Bohrstation: Unendlich

- Welche Art von System haben wir für die zweite Station? → $M \mid M \mid 1$
- Ankunftsrate an zweite Station: Entspricht der Ausgangsrate der 1. Station, daraus folgt: $\frac{1}{3.25} = 0.3077 \frac{jobs}{min}$
- Auslastung der zweiten Station: $u_2 = \frac{\frac{1}{3.25}}{\frac{1}{3}} = \frac{3}{3.25} = 0.9231$
- Leistung der zweiten Station: $WIP_2 = E_2(N) = \frac{u_2}{1-u_2} = 12\ jobs$
  - Durchsatz $\lambda_2 = \frac{1}{3.25} = 0.3077 \frac{jobs}{min}$
  - $CT_2 = E_2(W) = \frac{WIP_2}{\lambda_2} = 39\ min$
- **Total**: Zykluszeit $E_{tot}(W) = 3.25\ min + E_2(W) = 42.25\ min$
- Ware in Arbeit: $WIP_{tot} = \lambda_2 \cdot E_{tot}(W) = 13\ jobs$

Beispiel mit exponentiellen Bearbeitungszeiten → Inventar/Wartebereich der Bohrstation: endlich

- Jetzt: Endlicher Bestand
- Angenommen, es gibt Platz für 4 Aufträge
- Erste Station wird blockiert, wenn alle 4 Plätze belegt sind und die erste Station einen Auftrag beendet hat.
- Aus Sicht der 2. Station ist das gesamte System wie ein $M \mid M \mid 1 \mid 6$-System. Daraus folgt:
  - Durchsatz: $\bar{\lambda}_2 = \frac{1-u_2^{N^{max}}}{1-u_2^{N^{max}+1}}\lambda_2 = \frac{1-0.9231^6}{1-0.9231^7}\frac{1}{3.25} = 0.2736\frac{jobs}{min}$
  - $WIP_2 = E_2(N) = \frac{u_2}{1-u_2} - \frac{(N^{max}+1)\cdot u_2^{N^{max}+1}}{1-u_2^{N^{max}+1}} = 2.682\ jobs$
  - $CT_2 = E_2(W) = \frac{WIP_2}{\bar{\lambda}_2} = \frac{2.682\ jobs}{0.2736\frac{jobs}{min}} = 9.8\ min$
- **Total:**
  - Zykluszeit $E_{tot}(W) = 3.25\ min + \frac{WIP_2}{\bar{\lambda}_2} = 13.05\ min$
  - Ware in Arbeit: $WIP_{tot} = \bar{\lambda}_2 \cdot E_{tot}(W) = 3.57\ jobs$

Vergleich:

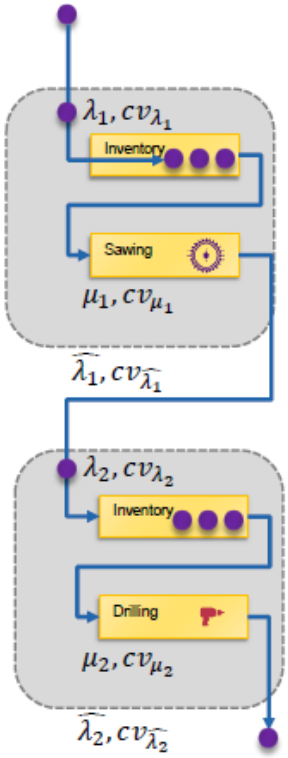
	Ungepuffert, unendlicher Bestand	Begrenztes Inventar auf 4 Plätze	Differenz
$\boldsymbol{WIP, E_{tot}(N)}$	$13\ jobs$	$3.571\ jobs$	$-73\ \%$
$\boldsymbol{CT, E_{tot}(W)}$	$42.25\ min$	$13.05\ min$	$-69\ \%$
$\boldsymbol{TH, \lambda = \lambda_2 = \bar{\lambda}_2}$	$0.3077\frac{jobs}{min}$	$0.2736\frac{jobs}{min}$	$-11\ \%$

Einblicke

- WIP und Durchlaufzeit werden durch Pufferung der Warteschlange zwischen den Stationen reduziert
- Allerdings wird auch der Durchsatz reduziert.
- Frage: Können wir uns diese Reduzierung leisten?
- Abwägung zwischen Umsatzeinbußen und Reduzierung der Lagerkosten
- Kanban kann nicht einfach durch eine Verringerung der Puffergrößen umgesetzt werden! (Der Verlust an Durchsatz wäre in der Regel zu groß)
  → Was können wir tun, um nicht zu viel Durchsatz zu verlieren?

Beispiel mit beliebigen Ankunfts- und Verarbeitungsraten

- Nehmen wir nun beliebige Ankunfts- und Verarbeitungsraten an



Durchfluss-Variabilität

Kennzahlen zum Fluss

- **Ankunft:**
  - Durchschnittliche Ankunftsrate: λ
  - Durchschnittliche Zeit zwischen den Ankünften: $\frac{1}{\lambda}$
  - Standardabweichung der Zeit zwischen den Ankünften: σ_λ
  - Variationskoeffizient der Zeit zwischen den Ankünften: $cv_\lambda = \sigma_\lambda \cdot \lambda$
- **Bearbeitung/Prozess**
  - Durchschnittliche Verarbeitungsrate: μ
  - Durchschnittliche Bearbeitungszeit: $\frac{1}{\mu}$
  - Standardabweichung der Bearbeitungszeit: σ_μ
  - Variationskoeffizient der Bearbeitungszeiten: $cv_\mu = \sigma_\mu \cdot \mu$
- **Abgang**
  - Durchschnittliche Abgangsrate: $\hat{\lambda}$
  - Durchschnittliche Zeit zwischen den Abgängen: $\frac{1}{\hat{\lambda}}$
  - Standardabweichung der Zeit zwischen den Abgängen: $\sigma_{\hat{\lambda}}$
  - Variationskoeffizient der Abgangszeiten: $cv_{\hat{\lambda}} = \sigma_{\hat{\lambda}} \cdot \hat{\lambda}$
- **Es gelten die folgenden Beziehungen:**
  - $\widehat{\lambda_1} = \lambda_2$
  - $cv_{\lambda_2} = cv_{\widehat{\lambda_1}}$
  - $\lambda_2 = \lambda_1$ (stationärer Zustand, Erhaltung der Strömung)
  - cv_{λ_1} kann aus der Verteilung der Ankunftszeiten bestimmt werden
  - cv_{μ_1} kann aus der Verteilung der Bearbeitungszeiten bestimmt werden

Ausbreitung der Variabilität: Näherungsformeln

- Einzelne Maschinen/Server: $cv_{\hat{\lambda}}^2 = cv_\mu^2 \cdot u^2 + (1-u^2)\cdot cv_\lambda^2$
- Mehrere Maschinen/Server: $cv_{\hat{\lambda}}^2 = 1 + (1-u^2)\cdot\left(cv_\lambda^2 - 1\right) + \frac{u^2}{\sqrt{c}}\left(cv_\mu^2 - 1\right)$, wobei $u = \frac{\lambda}{c\cdot\mu}$

Das Beispiel wird wieder aufgegriffen: Reduzierte Variabilität, unendlicher Bestand

- Betrachten wir eine Linie mit zwei Stationen
- Erste Station ("Sägen"): Durchschnittliche Bearbeitungszeit 3.25 min pro Auftrag
- Bearbeitungszeit mit $cv_{\mu_1}^2 = 1$ (exponential!)
- Zweite Station ("Bohren"): Durchschnittliche Bearbeitungszeit 3 min pro Auftrag
- Bearbeitungszeit mit $cv_{\mu_2}^2 = 0.252$
- Inventar der ersten (Bohr-)station: Unendlich
- Ankunftsrate an der zweiten Station: $\frac{1}{3.25} = 0.3077\ \frac{jobs}{min}$ --> entspricht Abgangsrate der ersten Station
- Auslastung der zweiten Station: $u_2 = \frac{\frac{1}{3.25}}{\frac{1}{3}} = \frac{3}{3.25} = 0.9231$

  ➔ Ankunftsrate und Auslastung der zweiten Station bleiben gleich (siehe Rechnung weiter oben)
- Variationskoeffizient der Ankunftsrate 2. Station ist aber höher: $cv_{\lambda_2}^2 = cv_{\bar{\lambda}_1}^2 = cv_{\mu_1}^2 \cdot u_1^2 + (1 - u_1^2) \cdot cv_{\lambda_1}^2 = 1$
- Leistung der zweiten Station:
  - Durchsatz: $\hat{\lambda}_2 = \frac{1}{3.25} = 0.3077\frac{jobs}{min}$
  - $WIP_2 = E_2(N) = u_2 + \frac{u_2^2 \cdot (1+cv_{\mu_2}^2)}{2(1-u_2)} = 6.81\ jobs$
  - $CT_2 = E_2(W) = \frac{WIP_2}{\hat{\lambda}_2} = \frac{6.81\ jobs}{0.3077\frac{jobs}{min}} = 22.125\ min$
- **Total:**
  - Zykluszeit $E_{tot}(W) = 3.25\ min + E_2(W) = 25.375\ min$
  - Ware in Arbeit: $WIP_{tot} = \bar{\lambda}_2 \cdot E_{tot}(W) = 7.81\ jobs$

Vergleich

	$M \mid M \mid 1$	$M \mid M \mid 1 \mid 6$	Differenz		$M \mid G \mid 1$
$WIP, E_{tot}(N)$	$13\ jobs$	$3.571\ jobs$	$-73\ \%$		$7.81\ jobs$
$CT, E_{tot}(W)$	$42.25\ min$	$13.05\ min$	$-69\ \%$		$25.375\ min$
$TH, \lambda = \lambda_2 = \bar{\lambda}_2$	$0.3077\frac{jobs}{min}$	$0.2736\frac{jobs}{min}$	$-11\ \%$		$0.3077\frac{jobs}{min}$

➔ reduzierte Prozessvariabilität

Blockierter Fall $M \mid G \mid 1 \mid N^{max}$

Für die Berechnung im blockierten Fall mit **allgemeinen Verteilungen für Bearbeitungszeiten** benötigen wir die Formel zur Berechnung des Durchsatzes approximativ:

$\lambda < \mu$	$\lambda > \mu$	$\lambda = \mu\ (u = 1)$
$TH \approx \frac{1-u\cdot\rho^{N^{max}-1}}{1-u^2\cdot\rho^{N^{max}-1}}\lambda$		$TH \approx \frac{1+cv_\mu^2+2(N^{max}-1)}{2(1+cv_\mu^2+N^{max}-1)} \cdot \mu$
$\rho = \frac{\frac{(1+cv_\mu^2)}{2}\cdot\frac{u^2}{(1-u)}}{\frac{(1+cv_\mu^2)}{2}\cdot\frac{u^2}{(1-u)}+u}$	$\rho = \frac{\frac{(1+cv_\mu^2)}{2}\cdot\frac{\frac{1}{u^2}}{(1-\frac{1}{u})}+\frac{1}{u}}{\frac{(1+cv_\mu^2)}{2}\cdot\frac{\frac{1}{u^2}}{(1-\frac{1}{u})}}$	

Blockierter Fall $G \mid G \mid 1 \mid N^{max}$

Für die Berechnung im blockierten Fall mit **allgemeinen Verteilungen für Ankunfts- und Bearbeitungszeiten** kann der Durchsatz wie folgt approximiert berechnet werden:

$\lambda < \mu$	$\lambda > \mu$	$\lambda = \mu\ (u = 1)$
$TH \approx \frac{1-u\cdot\rho^{N^{max}-1}}{1-u^2\cdot\rho^{N^{max}-1}}\lambda$		$TH \approx \frac{cv_\lambda^2+cv_\mu^2+2(N^{max}-1)}{2(cv_\lambda^2+cv_\mu^2+N^{max}-1)} \cdot \mu$
$\rho = \frac{\frac{(cv_\lambda^2+cv_\mu^2)}{2}\cdot\frac{u^2}{(1-u)}}{\frac{(cv_\lambda^2+cv_\mu^2)}{2}\cdot\frac{u^2}{(1-u)}+u}$	$\rho = \frac{\frac{(cv_\lambda^2+cv_\mu^2)}{2}\cdot\frac{\frac{1}{u^2}}{(1-\frac{1}{u})}+\frac{1}{u}}{\frac{(cv_\lambda^2+cv_\mu^2)}{2}\cdot\frac{\frac{1}{u^2}}{(1-\frac{1}{u})}}$	

Das Beispiel wird wieder aufgegriffen: Geringere Variabilität, begrenztes Bohrinventar

- Jetzt: Begrenzter Bohrungsbestand / Station 2 → Angenommen, es gibt Platz für 4 Aufträge
- Die erste Station wird blockiert, wenn alle 4 Plätze belegt sind und die erste Station einen Auftrag abschließt.
- Aus Sicht der 2. Station ist das System ein $M \mid G \mid 1 \mid 6$-System. Daraus folgt: Durchsatz $TH_2 \approx 0.2898 \frac{jobs}{min}$

Vergleich

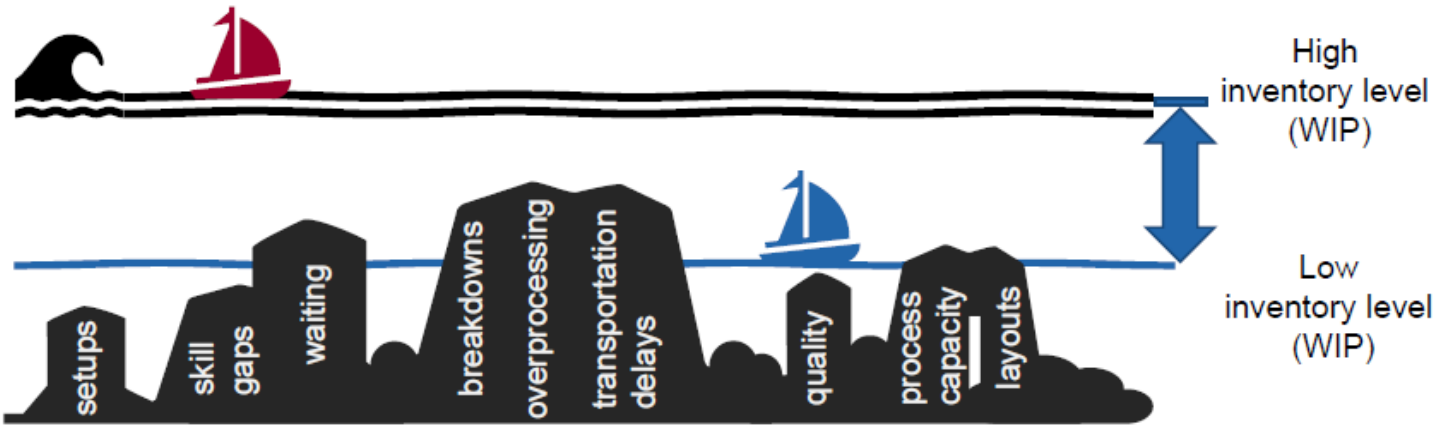
	$M \mid M \mid 1$	$M \mid M \mid 1 \mid 6$	Differenz		$M \mid G \mid 1$	$M \mid G \mid 1 \mid 6$	Differenz
$WIP, E_{tot}(N)$	$13\ jobs$	$3.571\ jobs$	$-73\ \%$		$7.81\ jobs$	$\leq 6\ jobs$	
$CT, E_{tot}(W)$	$42.25\ min$	$13.05\ min$	$-69\ \%$		$25.375\ min$	$\leq 21\ min$	
$TH, \lambda = \lambda_2 = \bar{\lambda}_2$	$0.3077 \frac{jobs}{min}$	$0.2736 \frac{jobs}{min}$	$-11\ \%$		$0.3077 \frac{jobs}{min}$	$0.2898 \frac{jobs}{min}$	$-5.8\ \%$

→ reduzierte Prozessvariabilität

Zusammenfassung

- Mit der Reduzierung der Puffer zwischen den Stationen und der Verringerung der Variabilität der Ausführungsprozesse können wir einen Durchsatz erreichen, der mit dem besten Fall vergleichbar ist, aber mit **begrenztem WIP** und **begrenzter Durchlaufzeit**.
- Mit der **Verringerung der Variabilität** sowohl der Ankunftsprozesse als auch der Ausführungsprozesse können wir sogar noch **bessere Ergebnisse** erzielen.
- Die einzige Möglichkeit, WIP und Durchlaufzeit zu reduzieren, ohne zu viel Durchsatz zu opfern, besteht darin, auch die Variabilität zu reduzieren!
- Darum ist Verringerung der Variabilität ein so wichtiger Bestandteil der JIT-Implementierung (Just in time).

→ **Nicht nur das Wasser ablassen, sondern auch die Felsen entfernen**



Verringerung der Variabilität im System und die 6 Gesetze des Factory Physics

Auslastungsgesetz

- **Erhöht** eine Station die **Auslastung**, ohne andere Änderungen vorzunehmen, so **steigen** sowohl der **durchschnittliche *WIP*** ($E(N)$ und $E(N_q)$) als auch die **durchschnittliche Warte**- und **Durchlaufzeit** ($E(W)$ und $E(W_q)$) in **stark nichtlinearer Weise**.
- Dies fasst die aus der Warteschlangentheorie resultierenden Effekte zusammen.

Kapazitätsgesetz

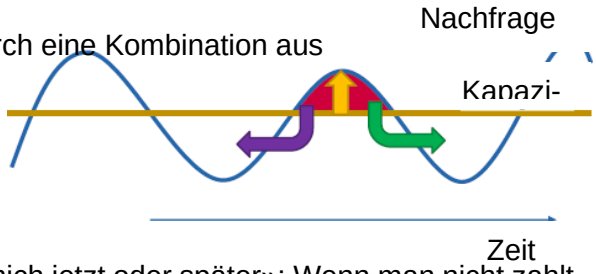
- Im **eingeschwungenen Zustand** werden alle Fabriken im Durchschnitt **Arbeit mit einer Rate produzieren**, die **strikt unter** der **durchschnittlichen Kapazität liegt**.
- **Geplante Strategien** dürfen **niemals Maßnahmen enthalten wie**
  - Überkapazitäten in großem Umfang nutzen
  - Überstunden
  - Vergabe von Unteraufträgen für das Kerngeschäft
- Wenn Manager diese Maßnahmen einbeziehen, dann entscheiden sie sich unbewusst dafür, ihre Fabriken im ständigen «Feuerlöschmodus» zu betreiben.

Variabilitätsgesetz

- **Zunehmende Variabilität verschlechtert immer** die **Leistung** eines **Produktionssystems**.
- Implikation: Höhere Variabilität jeglicher Art muss ein gewisses Maß an Leistung verschlechtern.
- Zunehmende Variabilität wirkt sich auf das System entlang der 3 Dimensionen aus:
  - **Vorrat/Bestand:** $WIP,\ E(N), E(N_q)$
  - **Kapazität:** $TH, TH^{max},\ \mu$
  - **Zeit:** $CT,\ E(W),\ E(W_q)$

Gesetz zur Pufferung der Variabilität

- Die Variabilität in einem Produktionssystem wird gepuffert durch eine Kombination aus
  - **Vorrat/Bestand**
  - **Kapazität**
  - **Zeit**
- Wir haben also die Wahl, wie die Variabilität die Leistung beeinträchtigen wird



Jetzt oder später zahlen

- Das Gesetz der Variabilitätspufferung ist vom Typ «Bezahle mich jetzt oder später»: Wenn man nicht zahlt, um Variabilität zu reduzieren, dann muss man für eine oder mehrere dieser Möglichkeiten bezahlen:
  - Verlorener Durchlauf
  - Vergeudete Kapazität
  - Lange Durchlaufzeiten u/o schlechter Kundenservice (Kunden Auftragsdurchdringungspunkt / Bevorratungsebene)
  - Verlängerte Zykluszeiten
  - Größere Vorräte/Bestände

Gesetz der Variabilitätsplatzierung

- In einer Linie, in der die **Freigaben unabhängig von** der **Fertigstellung** sind, **erhöht** die **Variabilität** zu **Beginn** des Arbeitsplans die Zykluszeit stärker als die entsprechende Variabilität zu einem späteren Zeitpunkt des Arbeitsplans.
- **Bemühungen** zur **Verringerung** der **Variabilität** sollten zuerst **an den Anfang der Linie** gerichtet werden (Maximierung der Auswirkungen von Änderungen)
- Bemerkung: Dieses Gesetz gilt nur, wenn die Freigaben unabhängig von der Fertigstellung sind, d. h. es gilt nicht für CONWIP-Linien oder Kanban-gesteuerte Linien.

Ursachen der Variabilität

Die wichtigsten Quellen der Variabilität in Fertigungssystemen sind:

- «Natürliche» Variabilität, d.h. geringfügige Schwankungen in der Prozesszeit aufgrund von kleinen Unterschieden bei Bedienern, Maschinen, Material, ...
- Zufällige Ausfälle/Pannen
- Umrüstungen
- Verfügbarkeiten von Bedienern und Material
- Nacharbeit (Qualitätsprobleme)

Natürliche Variabilität

- Die natürliche Variabilität ist der natürlichen Prozesszeit inhärent (geschuldet), d.h. alle Quellen **unvorhersehbarer** und **nicht identifizierbarer Schwankungen** in der Prozesszeit wie:
  - Materialzusammensetzung, die eine leicht unterschiedliche Verarbeitungsgeschwindigkeit verursacht
  - Parallele Maschinen, die nicht exakt mit der gleichen Geschwindigkeit arbeiten
  - Geringfügig unterschiedliche Geschwindigkeiten der Bediener, etc.
- **Natürliche Variabilität schließt aus:**
  - Zufällige Ausfallzeiten, Pannen
  - Einrichtungszeit
  - Andere externe Einflüsse
  - Variabilität aufgrund der Art des Auftrags/der Dienstleistung
- Die natürliche Variabilität in Fabriken ist in der Regel gering, in Kundendienst Systemen in der Regel höher

Variabilität durch Pannen

- Pannen sind **ungeplante Ausfallzeiten**, in vielen Systemen ist dies die **Hauptursache** für Schwankungen
- Auswirkungen von Ausfällen: Ein laufender Prozess wird unterbrochen aufgrund von
  - Stromausfall,
  - Notfall des Bedieners,
  - Mangel an Verbrauchsmaterialien,
  - ...
- Unterscheidung zwischen
  - **Unterbrechend Ausfall** (während eines Auftrags: Auftrag geht verloren oder wird gestoppt, Nacharbeit)
  - **Nicht-unterbrechend Ausfall** (zwischen Aufträgen)

Vokabeln

- **MTTF** (Mean Time To Failure) m_f: MTTF ist die durchschnittliche Zeit, die ein System oder eine Komponente in Betrieb ist, bevor es zu einem Ausfall kommt. Sie ist ein Maß für die **erwartete Zeit zwischen aufeinanderfolgenden Ausfällen** in einem System und wird berechnet: $m_f = \frac{Gesamtbetriebszeit}{Anzahl\ der\ Ausfälle}$
- **MTTR** (Mean Time To Repair) m_r: MTTR steht für die durchschnittliche Zeit, die benötigt wird, um ein ausgefallenes System oder eine ausgefallene Komponente zu reparieren und wieder in den Normalbetrieb zu überführen. Sie ist Maß für Wartbarkeit und Reparatureffizienz. Berechnung: $m_r = \frac{Gesamtausfallzeit}{Anzahl\ der\ Ausfälle}$
- cv_r der Reparaturzeiten wird mit 1 angenommen
- Verfügbarkeit $A = \frac{mf}{mf+mr}$
- Angepasste langfristige Bearbeitungszeit $\frac{1}{\mu_e} = \frac{1}{\mu \cdot A}$

Unterbrechender Ausfall: Maschinenausfall, während eines Auftrags

Maschine A

- Prozesszeit $\frac{1}{\mu} = 15\ min$
- Standardabweichung $\sigma_\mu = 3.35\ min$
- $cv_\mu^2 = \sigma_\mu^2 \cdot \mu^2 = 0.0499$
- Mittlere Zeit bis zum Ausfall MTTF $m_f = 12.4\ h$
- Mittlere Zeit bis Reparatur MTTR $m_r = 4.133\ h$
- cv_r der Reparaturzeit wird mit 1 angenommen
- Verfügbarkeit $A = \frac{mf}{mf+mr} = 0.75$
  - ➢ Angepasste langfristige Bearbeitungszeit $\frac{1}{\mu_e} = \frac{1}{\mu \cdot A} = 20\ min$
- Effektive Kapazität $3\ \frac{jobs}{h}$
- Mit Ankunftsrate von $2.25 \frac{jobs}{h}$: $u = 75\ \%$
- $cv_{\mu_e}^2 = 0.05 + 0.25 \cdot 4.133 \cdot 0.75 \cdot 4 \cdot 2 = 6.25$
- $E(W_q) = \frac{u \cdot (1+cv_\mu^2)}{2(1-u)} \cdot \frac{1}{\mu_e} = \frac{u \cdot (1+cv_{\mu_e}^2)}{2(1-u)} \cdot \frac{1}{\mu_e} = \frac{0.75 \cdot (1+6.25)}{2(1-0.75)} \cdot 20 = 217.5\ \text{min} \approx 3.5\ h$
- $E(N_q) = \frac{u^2 \cdot (1+cv_\mu^2)}{2(1-u)} = \frac{u^2 \cdot (1+cv_{\mu_e}^2)}{2(1-u)} = \frac{0.75^2 \cdot (1+6.25)}{2(1-0.75)} = 8.16\ jobs$

Maschine B

- Prozesszeit $\frac{1}{\mu} = 15\ min$
- Standardabweichung $\sigma_\mu = 3.35\ min$
- $cv_\mu^2 = \sigma_\mu^2 \cdot \mu^2 = 0.0499$
- Mittlere Zeit bis zum Ausfall MTTF $m_f = 1.9\ h$
- Mittlere Zeit bis Reparatur MTTR $m_r = 0.633\ h$
- cv_r der Reparaturzeit wird mit 1 angenommen
- Verfügbarkeit $A = \frac{mf}{mf+mr} = 0.75$
  - ➢ Angepasste langfristige Bearbeitungszeit $\frac{1}{\mu_e} = \frac{1}{\mu \cdot A} = 20\ min$
- Effektive Kapazität $3\ \frac{jobs}{h}$
- Mit Ankunftsrate von $2.25 \frac{jobs}{h}$: $u = 75\ \%$
- $cv_{\mu_e}^2 = 0.05 + 0.25 \cdot 0.633 \cdot 0.75 \cdot 4 \cdot 2 = 1$
- $E(W_q) = \frac{u \cdot (1+cv_\mu^2)}{2(1-u)} \cdot \frac{1}{\mu_e} = \frac{u \cdot (1+cv_{\mu_e}^2)}{2(1-u)} \cdot \frac{1}{\mu_e} = \frac{0.75 \cdot (1+1)}{2(1-0.75)} \cdot 20 = 60\ \text{min} = 1\ h$
- $E(N_q) = \frac{u^2 \cdot (1+cv_\mu^2)}{2(1-u)} = \frac{u^2 \cdot (1+cv_{\mu_e}^2)}{2(1-u)} = \frac{0.75^2 \cdot (1+1)}{2(1-0.75)} = 2.25\ jobs$

Berechnen Sie die cv der effektiven Prozesszeit: $cv_{\mu_e}^2 = cv_\mu^2 + (1-A)m_r \cdot \mu_e + cv_r^2(1-A)m_r \cdot \mu_e$

- cv_μ^2: natürliche Variabilität im Prozess
- $(1-A)m_r \cdot \mu_e$: Zufällige Ausfälle unabhängig von der Reparaturvariabilität
- $cv_r^2(1-A)m_r \cdot \mu_e$: Explizite Variabilität der Reparaturzeit

Erste Zusammenfassung

- Maschinen mit **häufigen**, **aber kurzen** Ausfällen sind den **seltenen**, aber **langen Ausfällen vorzuziehen** (vorausgesetzt, die Verfügbarkeiten sind gleich)
- **In der Praxis: Lange, seltene Ausfälle in kürzere, häufigere umwandeln**
  ➢ Vorbeugende und vorausschauende Instandhaltungsmaßnahmen
- Allerdings: **Überhaupt keine Ausfälle sind noch besser**
  ➢Verbesserung der Zuverlässigkeit eines Systems ist immer eine Anstrengung wert

Variabilität durch nicht unterbrechende Ausfälle

- Nicht-unterbrechende Ausfälle sind Ausfallzeiten, die unweigerlich auftreten werden, bei denen wir jedoch eine gewisse Kontrolle über den Zeitpunkt ihres Auftretens haben:
  - Vorbeugende Wartung
  - Pausen
  - Bedienerbesprechungen, Schichtwechsel
  - Einrichten von Maschinen
- Treten typischerweise **zwischen Aufträgen** auf
- (Fast) jede Ursache kann als eine Art von Umrüstung betrachtet werden; daher wird der Begriff Umrüstungen üblicherweise für diese Arten von Ausfällen verwendet.

Nicht-unterbrechende Unterbrechung: Einrichten, zwischen den Aufträgen

Maschine A

- Durchschnittliche Prozesszeit $10\ min$; zusätzlich:
  - Nach durchschn. 4 Jobs ist Setup erforderlich
  - Die Dauer dafür beträgt durchschnittlich $20\ min$

  ➢ Bereinigte mittlere Prozesszeit $\frac{1}{\mu_e} = 15\ min$
- Effektive Kapazität $4\ \frac{jobs}{h}$
- Müssen natürliche Prozessvariabilität anpassen:
- $\sigma_{\mu_e}^2 = \sigma_\mu^2 + \frac{\sigma_s^2}{N_s} + \frac{(N_s-1)}{N_s^2} t_s^2$, wobei
  - N_s = durchschn. Anz. Jobs zwischen 2 Umrüsten
  - t_s = durchschnittliche Zeit für Umrüsten
  - σ_s^2 = Varianz der Umrüstzeiten
- $cv_{\mu_e}^2 = \sigma_{\mu_e}^2 \cdot \mu_e^2$ ist angepasste Variationskoeffizient
- Berechne $cv_{\mu_e}^2$ mit $\sigma_\mu^2 = 11.2225\ min$ und $\sigma_s^2 = 2\ min^2$
- $cv_{\mu_e}^2 = \left(\sigma_\mu^2 + \frac{\sigma_s^2}{N_s} + \frac{(N_s-1)}{N_s^2} t_s^2\right) \cdot \mu_e^2 =$
  $\left(11.2225 + \frac{2}{4} + \frac{(4-1)}{4^2} 20^2\right) \cdot \frac{1}{15^2} = 0.38543$

Maschine B

- Mittlere Prozesszeit $15\ min$; kein Umrüsten
- Effektive Kapazität $4\frac{jobs}{h}$
- Natürliche Prozessvariabilität bleibt

$cv_{\mu_e}^2 = cv_\mu^2 = 11.2225 \cdot \frac{1}{15^2} = 0.0499$

Zweite Zusammenfassung

- Wir haben jetzt gesehen: Prozessvariabilität wird erzeugt durch
  - Schwankungen im Arbeitsablauf (einfache Effekte)
  - Zufällige Ausfälle (komplexere Effekte)
- Mit (nicht-)/unterbrechenden Ausfällen können üblichen Verdächtigen für mögliche hohe Variabilität finden
- Daraus folgt: Die in einem System vorhandene Variabilität ist die Folge von z.B.
  - Prozessauswahl
  - Qualitätskontrolle
  - Systemgestaltung
  - Managemententscheidungen
- Ganz allgemein: **Variabilität wird gepuffert** entweder in
  - Zeit
  - Kapazität
  - Bestand/Inventar
- Die Auswirkungen einer zunehmenden Prozessvariabilität an einer beliebigen Station sind z. B.
  - 1. Erhöhung der Zykluszeit $E(W)$
  - 2. Ausbreitung von mehr Variabilität auf nachgelagerte Stationen
- Frage: Wie können wir mit der Variabilität im Allgemeinen umgehen?

Einführung von Flexibilität und Pufferflexibilitätsgesetz

Pufferflexibilitätsgesetz

- Flexibilität reduziert den Umfang der Variabilitätspufferung in einem Produktionssystem.
- Flexible Puffer sind erforderlich, um die Variabilität zu vertretbaren Kosten zu reduzieren: Nutzen Sie Kapazität, Bestand und/oder Zeit auf mehr als eine Weise!
- Frage: Wie können Puffer flexibler gestaltet werden?

Auswirkungen des Flexibilitätspuffergesetzes

- **Beispiele** für **flexible Puffer**, die in mehr als einer Weise genutzt werden
  - **Flexible Kapazität:** übergreifend geschulte Arbeitskräfte, flexible Maschinen (Mehrzweckmaschinen, weniger Rüsten) im Layout
  - **Flexibler Bestand:** generischer WIP so lange wie möglich im System oder Produktanpassung so spät wie möglich (→ siehe Kundenauftragsdurchdringungspunkt, Bevorratungsebene (D); Make-to-Stock, Assemble-to-Order statt Make-to-Order)
  - **Flexible Zeit:** Angebot variabler Durchlaufzeiten an Kunden je nach Auftragsbestand/Arbeitslast
- **Flexibilität kann erreicht werden durch Produkt-, Anlagen- und Prozessdesign / Personalplanung.**

Ihr Job, Ihre Herausforderung

Kreative Wege zur Flexibilisierung von Ressourcen zu finden, ist die zentrale Herausforderung des Mass Customization-Ansatzes zur Herstellung einer Vielzahl von Produkten zu Massenproduktionskosten.

- *Wallace J. Hopp, Mark L. Spearman, 2008*

Stand der Arbeit

SW	Vorlesung	AB/Aufgabe	Skript	ZF
1	☒	☒	☒	☒
2	☒	☒	☒	☒
3	☒	☒	☒	☒
4	☒	☒	☒	☒
5	☒	☒	☒	☒
6	☒ Case Study	☒	☒	☒
7	☒	☐	☒	☒
8	☒	☒	☒	☒
9	☒	☒	☒	☒
10	☒	☒	☒	☒
11	☒	☒	☒	☒
12	☒	☐	☒	☒
13	☐ Case Study	☐	☒	☐
14	☒	☒	☒	☒

Modulauflagen