UHASSELT

KNOWLEDGE IN ACTION

Faculteit Bedrijfseconomische Wetenschappen

master handelsingenieur

Masterthesis

Minimum-cost staffing in queueing systems with abandonments: a simulation study

Caro Beyens

Scriptie ingediend tot het behalen van de graad van master handelsingenieur, afstudeerrichting operationeel management en logistiek

PROMOTOR :

Prof. dr. Inneke VAN NIEUWENHUYSE

2022
2023

UHASSELT

KNOWLEDGE IN ACTION

www.uhasselt.be
Universiteit Hasselt
Campus Hasselt:
Martelarenlaan 42 | 3500 Hasselt
Campus Diepenbeek:
Agoralaan Gebouw D | 3590 Diepenbeek

UHASSELT

KNOWLEDGE IN ACTION

Faculteit Bedrijfseconomische Wetenschappen

master handelsingenieur

Masterthesis

Minimum-cost staffing in queueing systems with abandonments: a simulation study

Caro Beyens

Scriptie ingediend tot het behalen van de graad van master handelsingenieur, afstudeerrichting operationeel management en logistiek

PROMOTOR :

Prof. dr. Inneke VAN NIEUWENHUYSE

Simulatie-gebaseerd onderzoek naar de trade-off tussen gemiddelde wachttijd en verlatingspercentage

Caro Beyens
HI (OML)
Faculteit Bedrijfseconomische Wetenschappen, UHasselt

Deze masterproef is een simulatiestudie naar wachtrijsystemen met verlatingsgedrag waarbij het geduld van de klant afhankelijk is van de tijd. Meer specifiek wordt de trade-off tussen gemiddelde wachttijd en verlatingspercentage onderzocht aan de hand van discrete-event simulatie. Dit wordt gedaan onder verschillende systeemcondities, waaronder variabele servicetijden, verschillende kansverdelingen voor geduld en verschillende servicecapaciteiten. Ook is er in deze studie onderzocht of de verkregen inzichten kunnen gelinkt worden aan één of meerdere capaciteitsregimes uit de literatuur (QR, ER en/of QER).

Kernwoorden: simulatie, wachtrijsystemen, wachtrijen met verlatingsgedrag, gemiddelde wachttijden, bezettingsgraad, geduld, variabele servicetijden, servicecapaciteit, kwaliteit vs. efficiëntie, regimes (QR, ER & QER), callcenter, Erlang-modellen, niet-tijdsafhankelijk aankomstproces en dienstverleningsproces

1. Inleiding

De grote uitdaging bij wachtrijsystemen is om een evenwicht te bereiken tussen de kwaliteit van dienstverlening en de operationele efficiëntie van het systeem (Garnett, Mandelbaum, & Reiman, 2002). Hierbij is het doel vaak om de operationele kosten te verlagen tot het niveau waarbij nog een aanvaardbaar niveau van dienstverlening kan worden bereikt. Zo wil men niet meer personeel hebben dan nodig en tegelijkertijd de gewenste kwaliteit van dienstverlening leveren (Whitt, 2007). Veel studies op het vlak van wachtrijsystemen gaan daarom op zoek naar de laagste servicecapaciteit waarbij de kwaliteit van dienstverlening nog wordt bereikt (Defraeye & Van Nieuwenhuyse, 2016), waarbij deze kwaliteit wordt gemeten aan de hand van een prestatiemaatstaf gelinkt aan congestie (bijv. de wachtrijlengte (Kim & Ha, 2012) en/of de wachttijd van bediende klanten (Izady & Worthington, 2012)). Deze masterproef focust zich op wachtrijsystemen waarbij klanten de neiging hebben de wachtrij te verlaten wanneer ze ongeduldig worden (ook wel “verlatingsgedrag” genoemd (Liu & Whitt, 2012)) en heeft tot doel om inzichten te geven in de trade-off tussen 2 verschillende performantiemaatstaven: de verwachte wachttijd in de wachtrij (m.b.t. de klanten die effectief service hebben gekregen), en het

verlatingspercentage (d.w.z. het percentage klanten dat de wachtrij verlaten heeft omdat de wachttijd te hoog opliep). Om deze inzichten te genereren, wordt gebruik gemaakt van simulatie, waarbij verschillende systeemcondities worden bestudeerd met betrekking tot servicecapaciteit, variabiliteit van de servicetijden en kansverdeling voor het geduld van de klant.

De invloed van variabele servicetijden op de prestaties van het systeem is interessant om te onderzoeken, omdat de congestie normaliter zal toenemen wanneer de variabiliteit in het systeem stijgt. In een realistisch systeem zijn er twee soorten variabiliteit aanwezig, namelijk voorspelbare en stochastische variabiliteit. De meeste algoritmes voor het bepalen van de servicecapaciteit houden enkel rekening met stochastische variabiliteit, i.e., variabiliteit als gevolg van het willekeurige (en dus onvoorspelbare) gedrag van klanten en personeel. Dit is in reële systemen altijd aanwezig (Whitt, 2007). Zo zal bijvoorbeeld niet elke klant even lang bediend worden, en zal ook niet elke klant exact hetzelfde geduld hebben. Voorspelbare variabiliteit daarentegen is niet altijd aanwezig. Deze variabiliteit is afhankelijk van de tijd, zoals bijvoorbeeld de stijging van de vraag in een piekseizoen versus de daling in een laagseizoen. Voorspelbare variabiliteit is daarom enkel van toepassing indien men tijdsafhankelijke aankomst- en/of serviceprocessen veronderstelt. In deze wachtrijsystemen probeert men dan ook de servicecapaciteit te veranderen doorheen de tijd om zo tegemoet te komen aan die voorspelbare variabiliteit (Gans, Koole, & Mandelbaum, 2003). In deze masterproef wordt enkel de *stochastische*, niet-voorspelbare variabiliteit in de servicetijden beschouwd, om zo de invloed van deze variabiliteit op beide performantiemaatstaven te kunnen analyseren.

Enkel bij een systeem *zonder* variabiliteit is het mogelijk om de servers continu te laten werken, zonder dat er wachtrijen ontstaan. We spreken dan van een bezettingsgraad van 100%. Zelfs indien de klanten zeer weinig geduld hebben, zullen er in zulk systeem nooit klanten verloren gaan als gevolg van verlatingsgedrag (aangezien er nooit een wachtrij ontstaat). Echter is er in een realistisch systeem altijd variabiliteit in de servicetijden aanwezig. Dit zorgt ervoor dat sommige klanten zullen moeten wachten op hun beurt, en dat er dus congestie in het systeem zal ontstaan. Ook zal een bezettingsgraad van honderd procent niet langer haalbaar zijn, aangezien de congestie dan zou blijven toenemen (wat theoretisch gezien leidt tot wachtrijen die groeien naar oneindig). Bijgevolg is het ook interessant om de servicecapaciteit op te nemen in het model. Het variëren van de servicecapaciteit zal namelijk de bezettingsgraad in het systeem beïnvloeden, en bijgevolg ook de congestie in het systeem. Hoe meer servers het systeem bevat, hoe lager de bezettingsgraad zal zijn, en bijgevolg hoe lager de congestie.

Daarnaast speelt ook het geduld, en het daaruit voortvloeiende verlatingsgedrag van de klant, een belangrijke rol bij de prestaties van het systeem. Zo zal, in een realistisch systeem, de klant de wachtrij verlaten wanneer de wachttijd zijn geduld overschrijdt (Garnett et al., 2002). Verlatingsgedrag gaat gepaard met een gemiste omzet aangezien men klanten verliest. Het verlatingsgedrag van klanten vertegenwoordigt dus een kost voor het systeem. Om de kwaliteit van dienstverlening te beoordelen dient men dus niet alleen te kijken naar de gemiddelde wachttijd in een systeem, maar ook naar het verlatingspercentage. Als men dit laatste niet mee zou opnemen, dan zouden systemen met weinig geduld (en dus meer verlatingsgedrag) op papier steeds blijk lijken te geven van een betere dienstverlening: door het hoge verlatingsgedrag reguleert de wachtrij immers zichzelf, zodat de gemiddelde wachttijd (van de klanten die effectief werden bediend, dus de wachtrij niet verlaten) laag blijft. Daarbovenop zal er ook minder servicecapaciteit nodig zijn, aangezien er minder klanten bediend moeten worden dan het aantal klanten die aankomen (Garnett et al., 2002). Bijgevolg is het makkelijker om de wachttijden laag te houden in een systeem met weinig geduld. Echter staat hier wel een kost tegenover, namelijk het omzetverlies veroorzaakt door de vele klanten die het systeem verlaten.

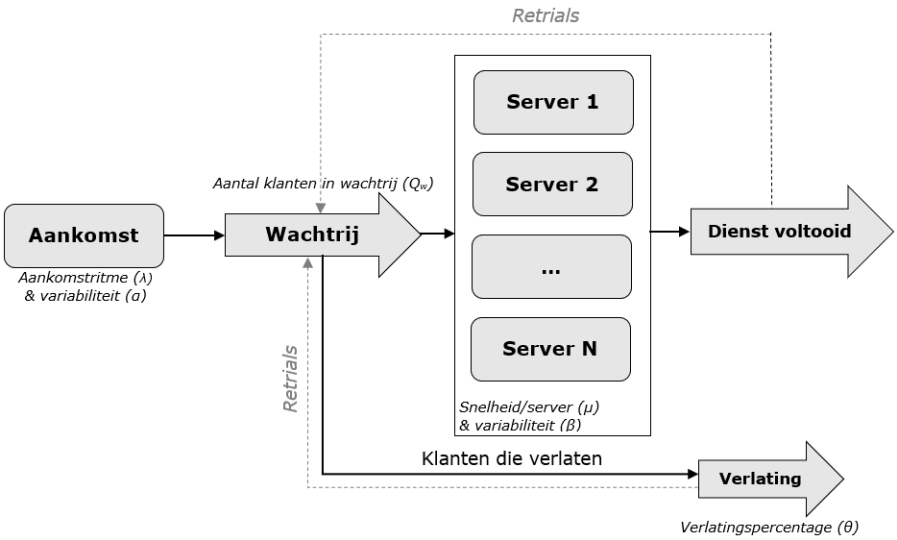
Bedrijven kunnen afhankelijk van hun beleid meer inzetten op kwaliteit of efficiëntie. Wanneer bedrijven streven naar een zo hoog mogelijke bezettingsgraad en dus bijgevolg een zo hoog mogelijke efficiëntie, dan wordt er in de literatuur gesproken van een efficiëntiegedreven regime (ER). Echter als bedrijven vooral inzetten op het verbeteren van de kwaliteit van hun dienstverlening, dus het verlagen van de vertragings- en verlatingskansen, spreekt men van een kwaliteitsgedreven regime (QR). Bedrijven zouden ook een balans kunnen proberen nastreven tussen kwaliteit en efficiëntie, wat dan het kwaliteits- en efficiëntiegedreven regime (QER) wordt genoemd. Verdere details met betrekking tot deze regimes worden besproken in sectie 2.

In deze masterproef wordt, naast de verwachte wachttijd, het verlatingspercentage mee opgenomen als kwaliteitsindicator om de kwaliteit van dienstverlening te beoordelen. Dit leidt tot de volgende onderzoeksvragen:

- Hoe ziet de trade-off tussen verlatingspercentage en verwachte wachttijd eruit, bij verschillende assumpties m.b.t. de kansverdeling van het geduld van de klanten?
- Hoe verandert deze trade-off bij wijzigingen in de servicecapaciteit?
- Hoe verandert deze trade-off bij wijzigingen in de variabiliteit van de servicetijden?
- Is er een link tussen de resultaten en de drie regimes (QR, ER & QER)?

Dit artikel is als volgt opgebouwd. In sectie 2 wordt de probleemstelling van deze studie toegelicht samen met de gangbare terminologie en notaties m.b.t. wachtrijsystemen. Sectie 3 geeft een literatuuroverzicht en sectie 4 focust op de methodologie van deze studie. In sectie 5 worden de resultaten besproken, gevolgd door de conclusie van deze masterproef in sectie 6.

2. Probleemstelling



Figuur 1: *Single-stage* wachtrijsysteem

Het wachtrijsysteem weergegeven in Figuur 1 bestaat uit drie processen: het aankomstproces, het wachtrijproces en het dienstverleningsproces. Klanten komen aan volgens een bepaald aankomstproces waarna ze in de wachtrij terechtkomen. Wanneer de klant de eerste in de wachtrij is (we veronderstellen een *first in first out* wachtrijdiscipline) en er één of meerdere servers vrij zijn, start het dienstverleningsproces. Wanneer de dienst voltooid is, verlaat de klant het systeem. Echter kan het ook zijn dat de klant de wachtrij verlaat, namelijk als de wachttijd het (beperkte) geduld van de klant overschrijdt. In principe zou deze klant later opnieuw kunnen proberen om alsnog de dienst te verkrijgen (in de literatuur wordt dit beschreven als *retrials;* zie bijvoorbeeld Defraeye and Van Nieuwenhuyse (2016); Feldman, Mandelbaum, Massey, and Whitt (2008)), en ook bediende klanten

zouden opnieuw in de wachtrij kunnen komen, als ze de desbetreffende dienst meerdere malen zouden willen verkrijgen (Defraeye and Van Nieuwenhuyse (2016); Gans, Koole, and Mandelbaum (2003); Garnett, Mandelbaum, and Reiman (2002)). Met terugkerende klanten na dienstverlening of na verlatingsgedrag wordt in deze masterproef geen rekening gehouden. Er wordt enkel gefocust op het feit dat klanten de wachtrij verlaten als gevolg van hun beperkte geduld, waardoor men klanten verliest en dus een deel van de omzet verloren zal gaan (Defraeye & Van Nieuwenhuyse, 2016).

De ervaring van de klant met betrekking tot het wachten, en bijgevolg zijn verlatingsgedrag, is verschillend in een fysieke wachtrij versus een virtuele wachtrij (Gans et al., 2003). In een fysieke rij kan de klant effectief zien hoeveel mensen er voor hem in de rij staan; bij een virtuele wachtrij heeft men deze informatie niet (tenzij hierover expliciet informatie wordt verstrekt). In deze masterproef gaan we ervan uit dat de klant de wachtrij verlaat omdat de wachttijd zijn geduld overschrijdt; het aantal klanten in de wachtrij is dus irrelevant (zie o.a. Cheng and Huo (2013); ; Green, Soares, Giglio, and Green (2006)).

Bij wachtrijsystemen is er een duidelijke trade-off tussen efficiëntie (gemeten a.d.h.v. bijv. de bezettingsgraad van de servers) en de behaalde servicekwaliteit (gemeten a.d.h.v. bijv. de gemiddelde wachttijd). Hoe hoger de bezettingsgraad, hoe hoger de kans op congestie en dus hoe hoger de gemiddelde wachttijden zullen zijn (Gans et al., 2003). In tegenstelling tot Gans et al. (2003), neemt Garnett et al. (2002) ook het verlatingsgedrag mee op als kwaliteitsmaatstaf, waarbij hogere bezettingsgraden niet alleen zullen zorgen voor hogere gemiddelde wachttijden maar ook voor hogere verlatingspercentages. Er wordt een onderscheid gemaakt tussen drie verschillende regimes die gehanteerd kunnen worden, weergegeven in Tabel 1 (Gans et al., 2003; Garnett et al., 2002).

	Kwaliteitsgedreven regime	Efficiëntiegedreven regime	Kwaliteits- en efficiëntiegedreven regime
Bezettingsgraad	Laag	Dichtbij 100%	Tussen laag en 100%
Kans op vertraging	Dichtbij 0	Dichtbij 1	Tussen 0 en 1
Kans op verlatingsgedrag	Dichtbij 0	Dichtbij ε*	Tussen 0 en 1
Aantal servers	> nominale behoeften R	< nominale behoeften R	≥/≤ nominale behoeften R
Streefdoel	Kwaliteit van dienstverlening	Efficiëntie	Kwaliteit van dienstverlening & efficiëntie

*voor verdere details over ε zie sectie 5.4

Tabel 1: Overzicht van drie regimes met een focus op kwaliteit, efficiëntie of beide

Het kwaliteitsgedreven regime (*QR*) is een regime dat de kwaliteit van dienstverlening als doel vooropstelt. Het streeft dus naar zo laag mogelijke wachttijden, met hogere bezettingsgraden tot gevolg. Men gaat eigenlijk meer personeel in dienst nemen dan de nominale behoeften. Met nominale behoeften bedoelt men het gemiddeld aantal klanten dat aanwezig zou zijn in het systeem als het aantal servers oneindig zou zijn. Met andere woorden is dit de gemiddelde servicecapaciteit nodig om aan de vraag te voldoen indien er geen enkele capaciteitsbeperking is (Duffield, Massey, & Whitt, 2001). Door een hogere servicecapaciteit te nemen dan de nominale behoeften probeert men, zelfs in de piekperiodes, de kans op vertraging voor alle klanten naar nul te brengen. Indien de kans op verlatingsgedrag mee opgenomen wordt als kwaliteitsmaatstaf, probeert men ook deze voor alle klanten naar nul te brengen (Garnett et al., 2002). Daartegenover staat het efficiëntiegedreven regime (*ER*) waarbij het zo optimaal mogelijk benutten van alle servers centraal staat. Hierbij heeft men minder personeel in dienst dan de nominale behoeften, waardoor het personeel honderd procent van de tijd bezet zal zijn, met hogere kansen op vertraging als gevolg (indien er verlatingsgedrag aanwezig is, gaat dit ook gepaard met hogere verlatingskansen). Tussen deze twee uiterste regimes ligt het kwaliteits- en efficiëntiegedreven regime (*QER*). Hierbij gaat men een balans zoeken tussen het maximaliseren van de bezettingsgraad van de servers en het minimaliseren van de vertragingskansen (en de verlatingskansen, indien verlatingsgedrag mee opgenomen wordt in het wachtrijsysteem). Enerzijds, als een organisatie te veel personeel in dienst heeft (met het oog op de minimalisatie van vertragings- en verlatingskansen), dan kan dit leiden tot onnodige kosten en inefficiëntie. Anderzijds, als een organisatie te weinig personeel in dienst heeft (met het oog op de maximalisatie van de bezettingsgraad van de servers), dan kan dit leiden tot lange wachttijden en ontevreden klanten. Door de juiste afweging te maken tussen deze factoren, kan een organisatie de optimale bezettingsgraad en vertragingskans (en eventueel verlatingskans) berekenen die nodig zijn om de gewenste kwaliteit van dienstverlening te bieden zonder onnodig veel personeel in dienst te hebben. Deze combinatie kan zelfs bereikt worden met het aantal servers kleiner dan de nominale waarde, door de efficiëntie van de servers te optimaliseren en de wachttijden voor klanten te beheersen (Gans et al., 2003; Garnett et al., 2002).

Het doel van deze studie is om de relatie tussen de gemiddelde wachttijd en het verlatingspercentage te bestuderen, bij verschillende servicecapaciteit, en onder variërende systeemcondities (variabiliteit van de servicetijden en kansverdeling voor het geduld van de klanten). Er wordt niet specifiek gefocust op een bepaald regime, maar er wordt wel nagegaan of de verkregen inzichten kunnen gelinkt worden aan één of meerdere regimes.

3. Literatuuroverzicht

In deze sectie wordt een overzicht gegeven van de relevante literatuur. De hiervoor gebruikte databanken zijn Google Scholar en Web Of Science. Initieel werd gezocht aan de hand van kernwoorden zoals *staffing for abandonment queues, abandonment queue models* en *staffing in queue models*. Vervolgens werd aan de hand van referenties in de gevonden artikels op zoek gegaan naar andere relevante literatuur. Er werd gefocust op artikels met publicatiejaar na 2000, om zo het risico op gedateerde literatuur te vermijden. Zo heb ik geen artikels ouder dan 30 jaar gebruikt en slechts drie artikels ouder dan 20 jaar. Ik heb ervoor gekozen deze drie papers toch te gebruiken omdat deze papers met minder voorkomende assumpties werken die moeilijk terug te vinden waren in papers na 2000 (zie ook Tabel 2). Zo heeft Brandt et al. (1999) drie minder voorkomende assumpties toegepast: geen personeelsintervallen, een wachtrijdiscipline gebaseerd op prioriteiten en een tijdsafhankelijk aankomstproces. Jennings et al. (1996) heeft dan weer gebruikgemaakt van een minder voorkomend dienstverleningsproces: een tijdsafhankelijke algemene verdeling. Tot slot heeft Thompson (1993) de minder voorkomende *exhaustive* servicediscipline toegepast.

3.1. Assumpties

In de literatuur bestaan er vooral analytische, wiskundige modellen die het beheer van de servicecapaciteit ondersteunen. Echter, om die wiskundige modellen bruikbaar te maken voor interpretatie en capaciteitsbeslissingen worden er vaak veel assumpties gemaakt. Tabel 2 geeft een overzicht van de assumpties die regelmatig worden gemaakt, inclusief verwijzingen naar papers die gebruik maken van de desbetreffende assumptie. Uit deze tabel kan dan ook afgeleid worden welke assumpties vaak voorkomen (kolom 1) en welke assumpties eerder zeldzaam zijn (kolom 2).

OVERZICHT ASSUMPTIES	Meest voorkomende assumptie		Andere assumpties
KLANT	**Homogeen** [1] [4] [8] [9] [12] [17] [19] [22] [24]		**Heterogeen** [13] [15] [16] [20] [28]
SERVERS	**Homogeen** [4] [8] [9] [12] [17] [19] [22] [24]		**Heterogeen** [1] [13] [15] [16] [20] [28]
	Tijdsafhankelijk [4] [8] [9] [12] [15] [16] [17] [19] [20] [22] [24] [28]		**Niet-tijdsafhankelijk** [1] [2] [13] [26]
PERSONEELSINTERVALLEN	**Ja** [4] [9] [12] [15] [16] [17] [19] [24] [28]		**Nee** [2] [8] [20]
WACHTRIJDISCIPLINE	***FIFO*** [1] [4] [8] [9] [12] [17] [19] [20] [22] [24]		**Gebaseerd op prioriteiten** [2] [16] [28]
SERVICEDISCIPLINE	***Preemptive*** [12] [13] [15] [19]		***Exhaustive*** [16] [24]
STRUCTUUR VAN HET SYSTEEM	***Single-stage*** [1] [2] [4] [8] [9] [12] [13] [15] [17] [19] [20] [22] [24]		***Multi-stage*** [14] [16] [28]
METHODE	**Analytisch**		**Simulatie** [4] [8] [28]
	Erlang [10] [11] [25]	**Benadering** [12] [15] [16] [20] [24]	
AANKOMSTPROCES	**Exponentiële verdeling** [1] [2] [8] [11] [12] [13] [15] [16] [19] [20] [22] [24] [26] [28]		**Algemene verdeling** [4] [9] [17]
	Tijdsafhankelijke verdeling [1] [4] [8] [9] [12] [13] [15] [16] [17] [19] [20] [22] [24] [28]		**Niet-tijdsafhankelijke verdeling** [2] [11] [26]
DIENSTVERLENINGSPROCES	**Exponentiële verdeling** [1] [2] [9] [11] [12] [13] [19] [20] [24] [26] [28]		**Algemene verdeling** [4] [8] [15] [16] [17] [22]
	Niet-tijdsafhankelijke verdeling [1] [2] [4] [8] [9] [11] [12] [13] [15] [16] [19] [20] [22] [24] [26] [28]		**Tijdsafhankelijke verdeling** [17]
Geduld	**Exponentiële verdeling** [11] [13] [19] [26]		**Algemene verdeling** [2] [8] [15] [22]
	Niet-tijdsafhankelijke verdeling [2] [8] [11] [13] [15] [19] [22] [26]		**Tijdsafhankelijke verdeling** /

Tabel 2: Overzicht assumpties

Zoals weergegeven in Tabel 2 zijn er iets meer papers die klanten als homogeen (d.w.z. er is één soort klantklasse) veronderstellen. Ook bij servers wordt er iets meer gebruikgemaakt van de homogene assumptie wat wil zeggen dat servers meestal verondersteld worden dezelfde vaardigheden te bezitten

en bijgevolg dezelfde servicesnelheid te leveren (Defraeye & Van Nieuwenhuyse, 2016). Daarnaast bepalen de meeste studies de servicecapaciteit via personeelsintervallen (en niet via *continue staffing*). Ook is de FIFO-wachtrijdiscipline (*first in first out*) veel populairder dan de wachtrijdiscipline gebaseerd op prioriteiten. Prioriteiten zorgen ervoor dat bepaalde klanten voorrang krijgen op andere klanten, zoals bijv. het geval is op een spoedafdeling (Defraeye & Van Nieuwenhuyse, 2016). Vervolgens kan men een *preemptive* of *exhaustive* servicediscipline toepassen, wat aangeeft wat er gebeurt met de klant die nog service ondergaat op het ogenblik dat de werktijd (vaak gedefinieerd aan de hand van de shiftduur) van de server afloopt. Bij een *preemptive* servicediscipline zal de service onderbroken worden, waarna de klant weer in de wachtrij zal komen (zij het wel als eerste). Bij een *exhaustive* discipline daarentegen zal de klant eerst verder geholpen worden en zal de server het systeem pas verlaten als de klant volledig is bediend (Defraeye & Van Nieuwenhuyse, 2016). Dit laatste is realistischer, maar wordt in de literatuur minder vaak gebruikt. Verder komen *single-stage* wachtrijsystemen (d.w.z. één dienstverleningsfase) veel meer voor dan *multi-stage* systemen (met meerdere dienstverleningsfases). Wat betreft het aankomstproces en dienstverleningsproces kan men een onderscheid maken tussen tijdsafhankelijke en niet-tijdsafhankelijke processen. Het aankomstproces veronderstelt men meestal als een tijdsafhankelijk proces (waarbij de kansverdeling van de tussenaankomsttijden verandert doorheen de tijd), het dienstverleningsproces als een niet-tijdsafhankelijk proces (de kansverdeling van de servicetijden blijft dus doorgaans hetzelfde doorheen de tijd). Daarnaast wordt er in Tabel 2 ook een onderscheid gemaakt tussen de exponentiële kansverdeling, en andere (i.e., algemene) verdelingen. Exponentiële verdelingen worden vaker gebruikt bij zowel het aankomst- als dienstverleningsproces, mede vanwege hun eenvoudige toepasbaarheid in wiskundige modellen en berekeningen. Dit maakt ze bijzonder nuttig voor het analyseren van wachtrijsystemen en het voorspellen van hun prestaties (Defraeye & Van Nieuwenhuyse, 2016).

Het geduld van de klanten is een belangrijke factor voor het verlatingsgedrag van klanten. Uit Tabel 2 wordt duidelijk dat verschillende papers het geduld van klanten op een andere manier modelleren. Zoals weergegeven in de tabel wordt het geduld van de klanten nooit als een tijdsafhankelijke variabele beschouwd, maar kan er wel een onderscheid gemaakt worden wat betreft de verdeling. Zo definiëren enkele papers geduld als een eenvoudige exponentiële verdeling (Garnett et al. (2002); Harrison and Zeevi (2005); Kim and Ha (2012); Whitt (2006)). Helber et al. (2010) heeft dan weer gewerkt met een Poisson verdeling voor het aantal verlatende klanten per tijdsinterval. Echter zijn er ook papers die een meer algemene verdeling gebruiken. Zo kan men gebruik maken van een individuele willekeurige

maximale wachttijd, waarbij elke klant een willekeurig geduld toegewezen krijgt zoals bij Brandt and Brandt (1999), Feldman, Mandelbaum, Massey, & Whitt (2008) en Whitt (2006).

Wat betreft de methodes om de wachtrijsystemen te analyseren, wordt er vaker gewerkt met analytische, wiskundige methodes in plaats van met simulatie-gebaseerde methodes. Bij de analytische, wiskunde methodes kiest men meestal voor het gebruik van benaderingen, maar ook de meest eenvoudige modellen, de Erlang-modellen, worden af en toe gebruikt. De Erlang-modellen worden hieronder verder toegelicht om inzicht te geven in het feit dat verlatingsgedrag opnemen in het wachtrijsysteem van cruciaal belang is. Echter zijn de Erlang-modellen niet realistisch en niet flexibel doordat ze gebruikmaken van strikte assumpties zoals exponentiële verdelingen. Door die assumpties wordt het mogelijk om berekeningen te doen aan de hand van eenvoudige formules, en dat is dan ook de belangrijkste reden dat papers hier nog gebruik van maken (zie o.a. Gans et al. (2003), Garnett et al. (2002) en Whitt (2004))

Het Erlang C-model, ook wel bekend als het M/M/N-wachtrijmodel, is het meest eenvoudige model dat vaak wordt gebruikt in de wachtrijtheorie. De Kendall-notatie wordt gebruikt om de kenmerken van het wachtrijsysteem te beschrijven. Hierbij staat de eerste letter voor het aankomstproces, de tweede letter voor het dienstverleningsproces en de N staat voor het aantal servers in het systeem. De letter "M" wordt gebruikt om een exponentiële verdeling aan te duiden, wat betekent dat zowel de tussenaankomsttijden als de procestijden exponentieel verdeeld zijn in het geval van het Erlang C-model. Gans et al. (2003) maken gebruik van het Erlang C-model, dat echter geen rekening houdt met het geduld van klanten. Het veronderstelt dus impliciet dat klanten oneindig veel geduld hebben, waardoor er geen verlatingsgedrag zal zijn. De kwaliteit van dienstverlening is daarom ook enkel gerelateerd aan de kans dat een klant moet wachten (Gans et al., 2003). Daarnaast is het systeem alleen stabiel als de servicecapaciteit N groter is dan de nominale behoeften (Garnett et al., 2002). Indien dit niet het geval is, zal de bezettingsgraad van de servers stijgen tot 100%, waarbij de wachtrij toeneemt naar oneindig. Om dergelijk onstabiel systeem onder controle te krijgen, moet men meer personeel in dienst nemen (i.e., meer servers inzetten) zodat de bezettingsgraad onder de 100% ligt. Echter bevat elk systeem in realiteit wel verlatingsgedrag, waardoor het Erlang C-model altijd een te hoge servicecapaciteit zal aanbevelen (Gans et al., 2003).

Het Erlang A-model (M/M/N+M-wachtrij), gebruikt door o.a. Garnett et al. (2002) en Whitt (2004), is nog steeds een sterk vereenvoudigd model, maar het is wel al iets realistischer dan het Erlang C-model aangezien het rekening houdt met het feit dat klanten een beperkt geduld hebben, en als gevolg de

wachtrij kunnen verlaten (in de Kendall-notatie wordt dit aangeduid met de term +M die verwijst naar exponentieel geduld). De kwaliteit van dienstverlening is dus niet enkel meer gebaseerd op de kans op wachten, maar ook op de verlatingskans. Een groot verschil met het Erlang C systeem is dat het Erlang A systeem automatisch een stabiele wachtrij oplevert, ook als de nominale behoeften groter zijn dan de servicecapaciteit N. Dit is een gevolg van het verlatingsgedrag, waardoor de grootte van de wachtrij zichzelf reguleert, en dus nooit tot oneindig zal toenemen. Een dergelijk systeem zal uiteindelijk minder klanten bedienen, wat betekent dat er minder servicecapaciteit nodig is om de gewenste gemiddelde wachttijd te bereiken. Echter is het ook belangrijk om het aantal verlatende klanten te beperken om een hoog verlatingspercentage te voorkomen. Een hoog verlatingspercentage kan de gemiddelde wachttijd weliswaar verbeteren, maar zal leiden tot een hoger omzetverlies waardoor de kwaliteit van dienstverlening als geheel lager kan zijn. In een studie naar een systeem met hoge bezettingsgraad, heeft Garnett et al. (2002) het verschil tussen het Erlang C-model en het Erlang A-model onderzocht. De resultaten tonen duidelijke verschillen in gemiddelde wachttijd en bezettingsgraad aan, wat benadrukt dat het meenemen van verlatingsgedrag als performantiemaatstaf in wachtrijsystemen van cruciaal belang is.

3.2. Prestatiemaatstaven

In de literatuur worden verschillende prestatiemaatstaven gebruikt om de kwaliteit van de dienstverlening en de efficiëntie van het systeem te beoordelen. Het is belangrijk dat deze prestatiemaatstaven zorgvuldig gekozen en berekend worden, zodat de prestatie van het systeem op een correcte manier kan worden weerspiegeld. Tabel 3 geeft een overzicht van de prestatiemaatstaven die gebruikt worden in de huidige literatuur, inclusief verwijzingen naar papers waarin de desbetreffende prestatiemaatstaven voorkomen.

Om de kwaliteit van de dienstverlening te beoordelen wordt typisch gekeken naar prestatiemaatstaven met betrekking tot congestie en/of verlatingsgedrag. Merk op (in kolom 2) dat de wachttijd en wachtrijlengte heel sterk gecorreleerd zijn met de verblijfsduur en het aantal klanten in het systeem. Verder kan uit Tabel 3 afgeleid worden dat het beoordelen van de performantie aan de hand van een kost, gerelateerd aan de wachttijd of aan verlatingsgedrag, slechts zelden voorkomt. Prestatiemaatstaven gebaseerd op kansen (bijv. de kans op wachten, de kans op een wachttijd hoger of lager dan een bepaalde drempelwaarde) zijn echter wel populair bij het beoordelen van de kwaliteit van dienstverlening. Voor de efficiëntie kan men kijken naar maatstaven zoals de bezettingsgraad, het aantal

bezette servers en het aantal bediende klanten per tijdseenheid. Zoals weergegeven in Tabel 3 is de bezettingsgraad de meeste voorkomende prestatiemaatstaf voor het bepalen van de efficiëntie van het systeem.

VERLATINGSGEDRAG	CONGESTIE	EFFICIËNTIE
Kans op verlating/ verlatingspercentage [3] [5] [10] [11] [12] [22] [23]	**Wachtrijlengte** [5] [11] [17] [19] [21] [22] [23]	**Bezettingsgraad servers** [5] [10] [11] [16] [28]
Kans op verlating gegeven een wachttijd > X minuten [11]	**Kans op wachten/percentage klanten die moeten wachten** [3] [4] [5] [8] [10] [11] [17] [22] [23] [27]	**Aantal bezette servers** [21] [22]
Verlatingskost [13] [23] [26]	**Wachttijd van bediende klanten (=*ASA*)** [5] [10] [11] [15] [16] [21] [22]	**Aantal bediende klanten/tijdseenheid** [13] [15] [26]
	Kans op langer/minder lang dan X minuten wachten [10] [11] [12] [16] [20] [24] [28]	
	Wachtkost [1] [23] [26]	
	Aantal klanten in systeem [9] [21]	
	Verblijfsduur in systeem [16] [28]	

Tabel 3: Overzicht prestatiemaatstaven

3.3. Focus masterproef

In deze studie wordt, in tegenstelling tot vele andere studies, niet specifiek op zoek gegaan naar de laagst mogelijke servicecapaciteit om een vooraf bepaalde, gewenste kwaliteit van dienstverlening te bereiken. De focus ligt op het verkrijgen van inzicht in de *trade-off* tussen de verwachte wachttijd en het verlatingspercentage in een systeem waarbij klanten een beperkt geduld hebben. Hiervoor maakt deze studie geen gebruik van de vaak voorkomende analytische modellen, maar wel van discrete-event simulatie. Dankzij simulatie is het mogelijk om betrouwbare resultaten te verkrijgen zonder beperkt te worden door onrealistische assumpties (zoals het gebruik van exponentiële verdelingen voor *alle* kansvariabelen). Met simulatie kunnen dus complexere systemen gemodelleerd worden dan via de wiskundige modellen (zoals bijv. de Erlang-modellen). Daarnaast wordt er in deze studie gefocust op niet-tijdsafhankelijke aankomst- en dienstverleningsprocessen, ook al wordt er in de literatuur vaak gebruik gemaakt van tijdsafhankelijke processen (zie o.a. Feldman et al. (2008) en Cheng and Huo (2013)). Indien men tijdsafhankelijke processen zou veronderstellen, zou men namelijk verschillende effecten door elkaar aan het meten zijn, wat uiteraard niet de bedoeling is.

4. Methodologie

In deze studie is er gebruik gemaakt van een discrete-event simulatiemodel, opgesteld met behulp van Arena, om de trade-off tussen de gemiddelde wachttijd en het verlatingspercentage te onderzoeken. Tabel 4 geeft een overzicht van de gebruikte assumpties in dit simulatiemodel. Het is belangrijk om op te merken dat deze aannames gebaseerd zijn op een callcenter.

KLANT	SERVERS	PERSONEELSINTERVALLEN	WACHTRIJDISCIPLINE	SERVICEDISCIPLINE
Homogeen	Homogeen	Nee	FIFO	*Exhaustive*
STRUCTUUR SYSTEEM	**METHODE**	**AANKOMSTPROCES**	**DIENSTVERLENINGSPROCES**	**GEDULD**
Single-stage	Simulatie	Exponentiële verdeling - Gemiddelde = 1 minuut	Lognormale verdeling - Gemiddelde = 5 minuten - Standaarddeviatie = 2.5 , 5 of 7.5	Triangulaire verdeling - Minimum = 0.5 minuten - Gemiddelde = 2.5 , 5, 10, 15, 20, 25 of 30 minuten - Maximum = 2 keer het gemiddelde

Tabel 4: Overzicht gebruikte assumpties

Zoals weergegeven in Tabel 4 wordt in deze studie het aankomstproces gemodelleerd als een exponentiële verdeling met een gemiddelde tijd van 1 minuut tussen twee aankomsten. Het serviceproces daarentegen is gemodelleerd met behulp van een lognormale verdeling, om zo de variabiliteit in de servicetijden te kunnen manipuleren. Om het effect van meer of minder variabele servicetijden op de trade-off tussen de gemiddelde wachttijd en het verlatingspercentage te onderzoeken, werd de standaarddeviatie van de lognormale verdeling gevarieerd. De gebruikte lognormale verdeling heeft een gemiddelde van 5 minuten en drie verschillende standaarddeviaties van 2.5, 5 en 7 minuten. Door deze standaarddeviaties te kiezen, worden er variatiecoëfficiënten (i.e., standaarddeviatie gedeeld door het gemiddelde) van respectievelijk 0.5, 1 en 1.5 verkregen, wat overeenkomt met een lage, middelmatige en hoge mate van variatie ten opzichte van het gemiddelde. Een lage variatiecoëfficiënt betekent dat de servicetijden weinig variatie hebben in verhouding tot het gemiddelde, terwijl een hoge variatiecoëfficiënt aangeeft dat de servicetijden aanzienlijke variatie vertonen ten opzichte van het gemiddelde. Ook met de (homogene) servicecapaciteit is geëxperimenteerd om het effect op de trade-off te kunnen achterhalen, waarbij simulaties werden uitgevoerd met 1 t.e.m. 7 servers. Tot slot werd het geduld gedefinieerd als een triangulaire verdeling met een minimum van 30 seconden, maar een veranderlijk maximum en gemiddeld geduld, waardoor het effect van weinig en veel geduld op de trade-off tussen gemiddelde wachttijd en verlatingspercentage onderzocht kon worden. Er zijn simulaties

uitgevoerd met een gemiddeld geduld van 2.5, 5, 10, 15, 20, 25 en 30 minuten, waarbij het maximum geduld telkens het dubbele van het gemiddeld geduld was.

Bij simulatiemodellen moet men ook het aantal replicaties en de replicatielengte bepalen. In dit simulatiemodel werd de replicatielengte gelijkgesteld aan de duur van een werkdag (i.e., 12 uur). De klanten die op het einde van de werkdag in het wachtrijsysteem zitten, zullen nog worden verder geholpen (*exhaustive* servicediscipline), waardoor de werkdag iets langer dan twaalf uur kan duren. Aanvankelijk was het plan om de *sequential sampling* techniek te gebruiken voor het bepalen van het aantal replicaties (Kelton, Sadowski, & Zupick, 2015). Met deze techniek zou de simulatie stoppen zodra de gewenste precisie bereikt werd. Deze precisie werd bepaald aan de hand van de *half width* (wat de breedte van het betrouwbaarheidsinterval aangeeft) en het gemiddelde van het 95% betrouwbaarheidsinterval (berekend aan de hand van de t-verdeling) van de gemiddelde wachttijd en het aantal verlatende klanten. Door problemen met de uitvoering van de *sequential sampling* techniek door de software, is besloten om een vast aantal replicaties (i.e., 200) te gebruiken. Met deze hoeveelheid replicaties kan een precisie van vijf procent of minder worden gegarandeerd in elk scenario, wat resulteert in een hoge betrouwbaarheid en accuraatheid van de resultaten.

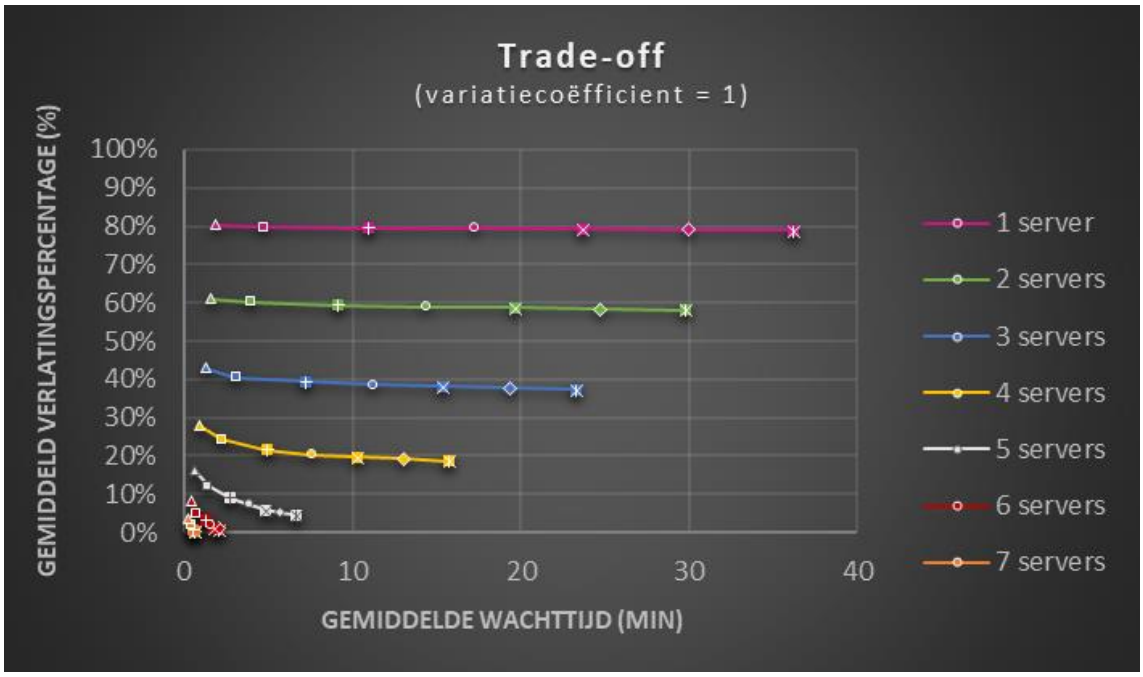
5. Resultaten

Deze resultatensectie is opgebouwd volgens de volgorde van de vier onderzoeksvragen besproken in de inleiding.

5.1. Hoe ziet de trade-off tussen verlatingspercentage en verwachte wachttijd eruit, bij verschillende assumpties m.b.t. de kansverdeling van het geduld van de klanten?

Bij de eerste onderzoeksvraag is de invloed van geduld op de trade-off tussen verlatingspercentage en verwachte wachttijd onderzocht. In een systeem met weinig geduld is het makkelijker om de wachttijden laag te houden, maar hier tegenover staat wel het omzetverlies als gevolg van de vele klanten die verlaten. Bijgevolg kan meer geduld leiden tot meer congestie binnen het systeem, waardoor er een lager verlatingspercentage en een hogere gemiddelde wachttijd ontstaat (aangezien meer mensen in de wachtrij zullen staan). Dus anders verwoord zou meer geduld een positieve invloed hebben op het verlatingspercentage en een negatieve invloed op de gemiddelde wachttijd. Echter is deze veronderstelling niet helemaal waar, of toch niet voor elk wachtrijsysteem. Figuur 2 toont aan dat meer

geduld gepaard gaat met een hogere gemiddelde wachttijd ongeacht het aantal servers, wat in lijn is met de verwachting. Elke curve stelt een wachtrijsysteem voor met een bepaalde servicecapaciteit, waarbij elke waarneming op de curve staat voor een andere hoeveelheid geduld. Zo blijkt bijvoorbeeld uit de paarse curve voor één server dat het verschil in wachttijd tussen een systeem met weinig geduld (Δ) en veel geduld (⊠) meer dan dertig minuten kan bedragen. Echter toont Figuur 2 ook aan dat er bij het verlatingspercentage niet altijd zo'n duidelijk effect van geduld aanwezig is. Bij de wachtrijsystemen met weinig servers zijn de curves namelijk bijna horizontaal, wat impliceert dat geduld geen of bijna geen invloed heeft op het verlatingspercentage in systemen met weinig servers. Vanaf 4 à 5 servers is er echter een dalende curve zichtbaar, waarbij de verwachte invloed van geduld op het verlatingspercentage (i.e., meer geduld zorgt voor een lager verlatingspercentage) bevestigd wordt.



Symbool	Gemiddeld geduld
Δ	2.5
□	5

Symbool	Gemiddeld geduld
✚	10
○	15

Symbool	Gemiddeld geduld
⊠	20
◇	25

Symbool	Gemiddeld geduld
⊠	30

Triangulaire verdeling van geduld (in minuten): TRIA(0.5 ; Gemiddeld geduld ; 2*Gemiddeld geduld)

Figuur 2: Trade-off bij verschillend geduld

Figuur 3 zoomt in op de trade-off curves (weergegeven in Figuur 2) van de wachtrijsystemen met 5, 6 en 7 servers om deze beter te kunnen bestuderen. De trade-off-curves in deze figuur weerspiegelen

de verwachte uitkomsten van het simulatie-experiment, namelijk dat meer geduld gepaard gaat met zowel een lager verlatingspercentage als een hogere gemiddelde wachttijd. Zo is bijvoorbeeld bij 5 servers (witte curve) het verschil in verlatingspercentage tussen een systeem met weinig geduld (Δ) en veel geduld (⊠) meer dan tien procent is, en het verschil in gemiddelde wachttijd meer dan vijf minuten. Merk hierbij ook op dat de afname van zowel het verlatingspercentage als de gemiddelde wachttijd groter is bij weinig geduld (waarnemingen liggen verder uit elkaar aan de linkerkant) dan bij veel geduld (waarnemingen liggen dichter bij elkaar aan de rechterkant). Dit gaat gepaard met de bevinding dat de bezettingsgraad bij lagere geduldwaarden sterker afneemt, zoals geïllustreerd in Figuur 4.

Symbool	Gemiddeld geduld
Δ	2.5
□	5

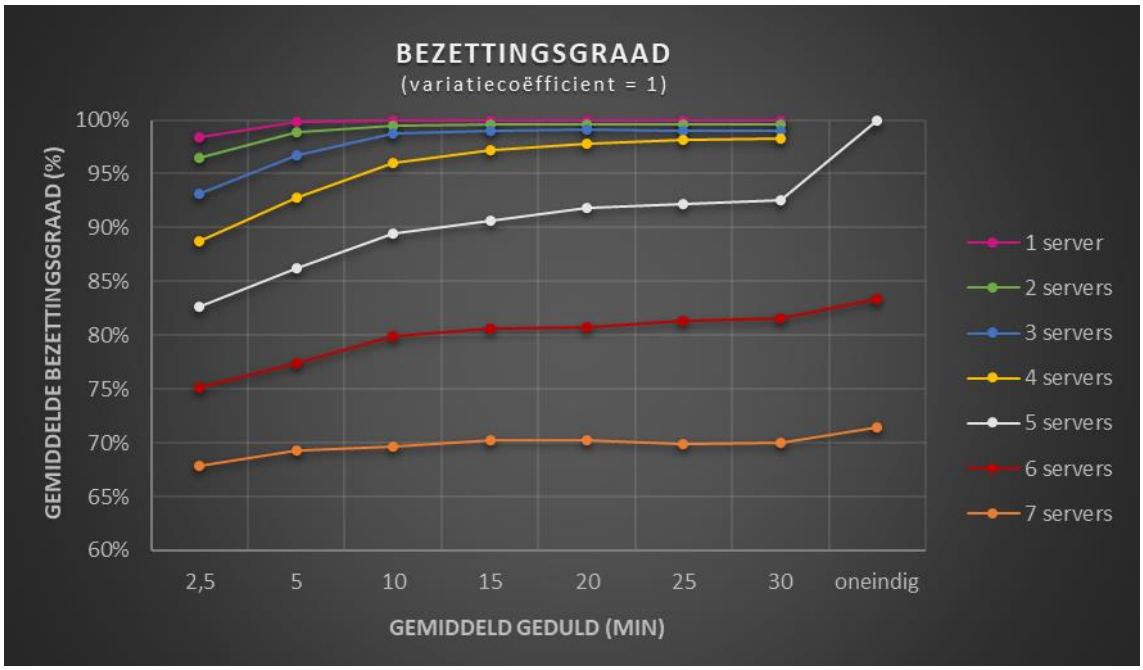
Symbool	Gemiddeld geduld
✚	10
○	15

Symbool	Gemiddeld geduld
⊠	20
◇	25

Symbool	Gemiddeld geduld
⊠	30

Triangulaire verdeling van geduld (in minuten): TRIA(0.5 ; Gemiddeld geduld ; 2*Gemiddeld geduld)

Figuur 3: Trade-off bij verschillend geduld (5 t.e.m. 7 servers)



Figuur 4: Bezettingsgraad

Aangezien de bezettingsgraad de meest gebruikte maatstaf is in een wachtrijsysteem voor het bepalen van de efficiëntie van het wachtrijsysteem, is deze in kaart gebracht in Figuur 4. Zoals eerder besproken zal een hogere bezettingsgraad zorgen voor meer congestie binnen het systeem. Bijgevolg kan verondersteld worden dat hogere bezettingsgraden gepaard gaan met hogere gemiddelde wachttijden en verlatingspercentages. Figuur 4 geeft de bezettingsgraden van de simulatie-experimenten weer waarbij elke curve staat voor een verschillende servicecapaciteit, en waarbij de x-as het gemiddeld geduld aangeeft. Deze figuur bestaat uit allemaal stijgende curves, wat aantoont dat de bezettingsgraad hoger zal liggen wanneer het geduld groter is. Dit is uiteraard logisch want bij meer geduld zullen minder klanten het systeem verlaten, waardoor er meer klanten bediend moeten worden.

Bij het vergelijken van de bezettingsgraad in Figuur 4 met de trade-off-curves in Figuur 2, blijken er enkele interessante verbanden te zijn. Om deze verbanden beter te begrijpen, wordt eerst aan de hand van Tabel 5 de theoretische bezettingsgraad (= aankomstritme/procesritme) van elke systeem met oneindig geduld berekend, om bijgevolg te kunnen bepalen welke systemen intrinsiek (in)stabiel zijn. Uit Tabel 5 kunnen we afleiden dat de theoretische bezettingsgraad boven de 100% ligt bij de systemen met 1, 2, 3 en 4 servers. Deze systemen zijn bijgevolg intrinsieke instabiele systemen waarbij wachtrijen

naar oneindig zullen groeien indien klanten (oneindig) veel geduld hebben. De systemen met 6 en 7 servers hebben een theoretische bezettingsgraad onder de 100% en zijn bijgevolg intrinsieke stabiele systemen waarbij de wachtrij niet tot oneindig zal toenemen. Het wachtrijsysteem met 5 servers is echter een randgeval. In een systeem zonder enige vorm van variabiliteit zouden 5 servers net alle klanten kunnen bedienen, wat leidt tot een theoretische bezettingsgraad van exact 100%. Echter bevat een systeem in realiteit altijd variabiliteit, maar omdat uit de resultaten blijkt dat de wachtrijsystemen met 5 servers zich meer gedragen zoals de systemen met 6 en 7 servers, zullen de systemen met 5 servers in het vervolg van deze paper worden beschouwd als intrinsieke stabiele systemen.

Oneindig geduld							
Aantal servers	1	2	3	4	5	6	7
Theoretische bezettingsgraad	500,00%	250,00%	166,67%	125,00%	100,00%	83,33%	71,43%

Tabel 5: Theoretische bezettingsgraad in systemen met oneindig geduld (1 t.e.m. 7 servers)

Eerder is al benadrukt dat er interessante verbanden bestaan tussen Figuur 2 en Figuur 4. Zo blijkt dat zowel de gemiddelde wachttijd als de bezettingsgraad zullen dalen bij een daling in het geduld, ongeacht het aantal servers. De impact van geduld op het verlatingspercentage wordt nochtans wel beïnvloed door het aantal servers. Als het aantal servers zo laag is dat het systeem intrinsiek instabiel is, zal geduld geen of slechts een minimale invloed hebben op het verlatingspercentage. Daarentegen wordt bij een intrinsiek stabiel systeem (waarbij de bezettingsgraad onder de honderd procent daalt) steeds meer invloed van geduld op het verlatingspercentage waargenomen. Geduld zal (bijna) geen invloed hebben op het verlatingspercentage in intrinsiek instabiele systemen, doordat deze systemen zo overbezet zijn dat de wachtrij eigenlijk nooit leegloopt, zelfs als de hoeveelheid geduld zou afnemen. Hierdoor zullen er steeds klanten in de wachtrij staan wanneer een server vrijkomt, en bijgevolg zal de bezettingsgraad van de server dichtbij die 100% komen te liggen. In ons simulatiemodel kan elke server gemiddeld gezien 20% (aankomsttijd / servicetijd = 1 minuut / 5 minuten) van de klanten bedienen. De wachttijden van de andere klanten zullen uiteindelijk hun geduld overschrijden, waardoor bij een wachtrijsysteem met één server ongeveer 80% van de klanten de wachtrij zal verlaten, bij 2 servers ongeveer 60%, bij 3 servers ongeveer 40% en bij 4 servers ongeveer 20%. Dit kan vervolgens de vlakke curves waargenomen in Figuur 2 verklaren. Echter, hoe minder overbezet het systeem wordt, hoe meer invloed geduld zal hebben op het verlatingspercentage. Dat is dan ook de reden dat intrinsieke stabiele systemen wel een invloed ondervinden van die hoeveelheid geduld. In deze systemen zal beperkt geduld vaker zorgen voor een lege wachtrij (het aankomstpatroon van nieuwe klanten is niet snel genoeg om de

lengte van de wachtrij constant boven nul te houden), waardoor een onbezette server even zal moeten wachten op een nieuwe klant en er bijgevolg geen sprake meer is van een honderd procent bezettingsgraad. Uit deze bevindingen kunnen we dus concluderen dat de wachtrijsystemen met 5, 6 en 7 servers (de intrinsieke stabiele systemen) zowel een significante invloed van geduld ondervinden op de gemiddelde wachttijd als op het verlatingspercentage. Bij de wachtrijsystemen met 1, 2, 3 en 4 servers heeft geduld enkel een significante invloed op de gemiddelde wachttijd.

Symbool	Gemiddeld geduld
Δ	2.5
□	5

Symbool	Gemiddeld geduld
✚	10
○	15

Symbool	Gemiddeld geduld
⊠	20
◇	25

Symbool	Gemiddeld geduld
⊠	30
●	Oneindig geduld

Triangulaire verdeling van geduld (in minuten): TRIA(0.5 ; Gemiddeld geduld ; 2*Gemiddeld geduld)

Figuur 5: Trade-off bij verschillend geduld, inclusief oneindig geduld (5 t.e.m. 7 servers)

Tot slot werd er nog een klein experimentje gedaan met de wachtrijsystemen van 5, 6 en 7 servers (de intrinsieke stabiele systemen) om te achterhalen of de wachtrijsystemen zonder verlatingsgedrag (dus oneindig geduld) significant verschillen van de eerder gesimuleerde systemen. De resultaten van dit experiment hebben bijgevolg de resultaten in Figuur 3 een beetje uitgebreid, namelijk er is bij elke curve 1 extra waarneming (weergegeven via ●) bijgekomen, wat Figuur 5 oplevert. Hierbij valt op dat de gemiddelde wachttijd bij 6 en 7 servers niet veel groter is geworden ten opzichte van het scenario met

een gemiddeld geduld van dertig minuten. De gemiddelde wachttijd bij een systeem van 7 servers kan dus in dit geval maximum 0.8 minuten bedragen en bij 6 servers 2.6 minuten. Deze gemiddelde wachttijd zal afnemen wanneer het gemiddeld geduld daalt, aangezien meer klanten zullen gaan lopen en er bijgevolg minder mensen in de wachtrij zullen staan. Een groter verschil is echter waarneembaar bij 5 servers. Bij 5 servers en een gemiddeld geduld van 30 (⊠) is het verlatingspercentage nog bijna vijf procent, terwijl dit bij 6 en 7 servers al veel dichter bij nul procent ligt. Bijgevolg ligt de gemiddelde wachttijd bij 5 servers met oneindig geduld meer dan tien minuten hoger dan de situatie waarin het gemiddelde geduld dertig minuten is, wat toch een aanzienlijk verschil is. Dus hoe groter de servicecapaciteit en hoe kleiner de bezettingsgraad, hoe kleiner het effect zal zijn van het al dan niet opnemen van geduld op de gemiddelde wachttijd in het systeem. Echter, wanneer de bezettingsgraad tegen de honderd procent ligt (wachtrijsysteem met minder dan 5 servers en het geduld groot genoeg of oneindig), zal de gemiddelde wachttijd altijd naar oneindig groeien. Merk hierbij op dat het van groot belang is om het verlatingspercentage mee op te nemen als prestatiemaatstaf, omdat verlatingsgedrag gepaard gaat met verloren omzet wat men natuurlijk ten allen tijde wil minimaliseren.

5.2. Hoe verandert deze trade-off bij wijzigingen in de servicecapaciteit?

De tweede onderzoeksvraag gaat kijken naar de invloed van een verschillende servicecapaciteit op de trade-off tussen verlatingspercentage en gemiddelde wachttijd. Er wordt verondersteld dat een hogere servicecapaciteit gepaard gaat met een betere kwaliteit van dienstverlening, maar dit gaat ten koste van de efficiëntie van het systeem. Daarom is bij deze onderzoeksvraag de verwachting: hoe meer servers het wachtrijsysteem bevat, hoe beter de kwaliteit van het wachtrijsysteem zal zijn, dus hoe lager de gemiddelde wachttijd en het verlatingspercentage. Figuur 6 bevestigt deze verwachting. Elke curve stelt een ander geduld voor en elke waarneming op zo'n curve staat voor een andere servicecapaciteit. In Figuur 6 zijn alleen stijgende curves te zien, dus ongeacht de hoeveelheid geduld, bij een lagere servicecapaciteit zullen er langere wachttijden zijn en meer klanten zullen het systeem verlaten dan bij een hogere servicecapaciteit. Neem bijvoorbeeld de gele curve met een gemiddeld geduld van vijftien minuten. Wanneer men op deze curve het wachtrijsysteem van 4 servers (○) met het systeem van 2 servers (◇) vergelijkt, is er een duidelijk verschil te zien in zowel het verlatingspercentage als de gemiddelde wachttijd. Het wachtrijsysteem met 2 servers presteert duidelijk slechter op beide kwaliteitsmaatstaven dan het systeem met 4 servers. Het verhogen van de servicecapaciteit zal dus altijd

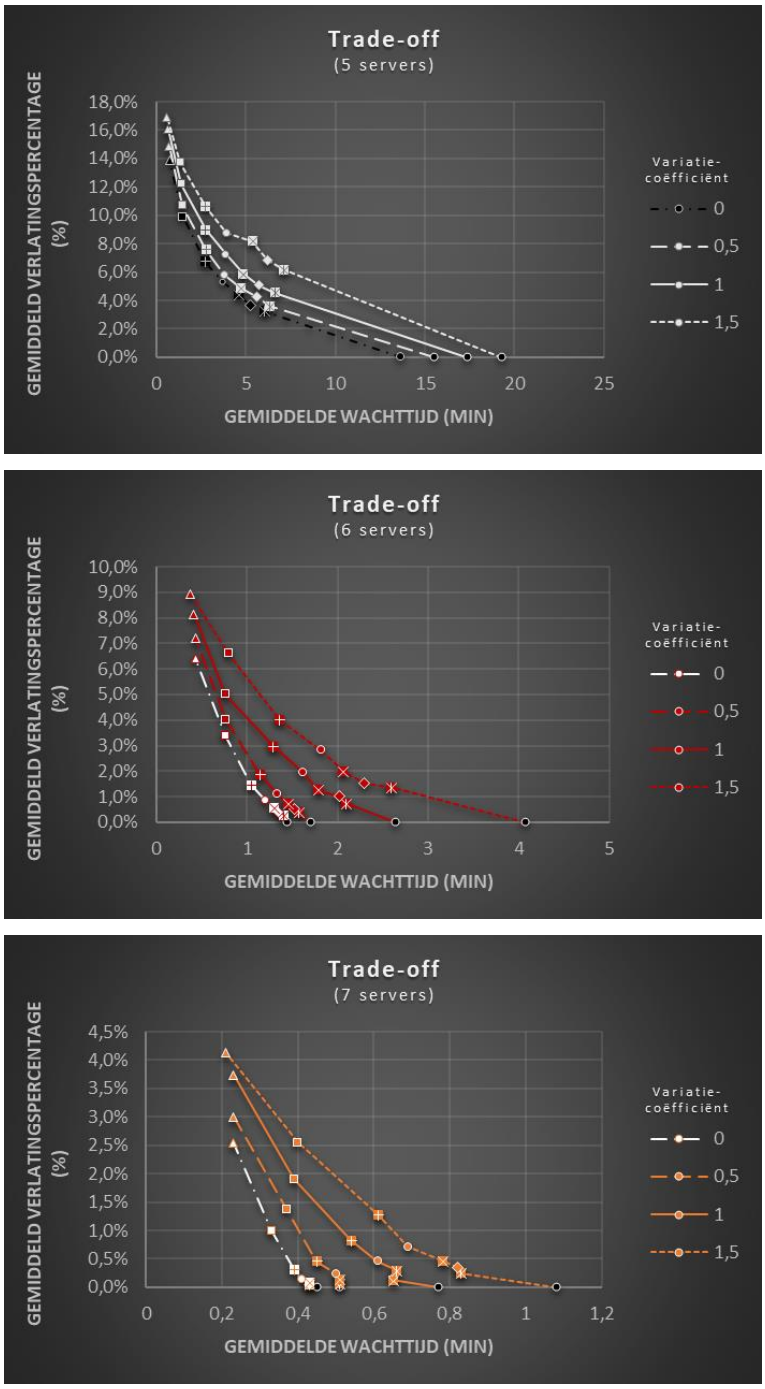
een positieve invloed hebben op de kwaliteit van het wachtrijsysteem, zowel op het verlatingspercentage als op de gemiddelde wachttijd. Het is belangrijk op te merken dat het vergroten van de servicecapaciteit een negatieve invloed heeft op de bezettingsgraad van een wachtrijsysteem, wat kan leiden tot een lagere efficiëntie. Uit Figuur 4 kan worden afgeleid dat dit effect veel minder groot zal zijn bij een intrinsiek instabiel systeem.

Figuur 6: Trade-off bij verschillende servicecapaciteit

5.3. Hoe verandert deze trade-off bij wijzigingen in de variabiliteit van de servicetijden?

In de derde onderzoeksvraag werd er onderzocht of de variabiliteit in de servicetijden een invloed heeft op de trade-off tussen verlatingspercentage en gemiddelde wachttijd. Men zou verwachten dat meer variabiliteit zorgt voor meer congestie en bijgevolg slechtere service kwaliteit van het wachtrijsysteem (i.e., hogere verlatingspercentages en langere wachttijden). Om dit te onderzoeken zijn de curves uit Figuur 3 gesimuleerd voor drie verschillende scenario's, elk met een verschillende variabiliteit in de servicetijden, namelijk een lage variatiecoëfficiënt (0.5), een middelmatige variatiecoëfficiënt (1) en een

hoge variatiecoëfficiënt (1.5) (zie Figuur 7). De wachtrijsystemen met minder dan 5 servers worden buiten beschouwing gelaten, omdat deze systemen intrinsiek instabiel zijn met een bezettingsgraad tegen de honderd procent. Figuur 7 bestaat uit drie grafieken die elk naar een andere servicecapaciteit verwijzen. Elke curve komt overeen met een bepaalde variatiecoëfficiënt voor de servicetijden, en elke waarneming op die curve staat voor een andere hoeveelheid geduld. Aangezien de drie curves niet samenvallen, kan worden geconcludeerd dat variabiliteit in de servicetijden invloed heeft op de kwaliteitsprestaties van het wachtrijsysteem. De curves van de systemen met een hogere (lagere) variantie in de servicetijden gaan gepaard met een hoger (lager) verlatingspercentage (er is een duidelijke verticale verschuiving t.o.v. de Y-as). De invloed van de variabiliteit op de gemiddelde wachttijd is echter beperkt, aangezien de horizontale verschuivingen van de curves miniem zijn. In de meeste gevallen is er een kleine verschuiving naar rechts te zien, wat wijst op een iets hogere gemiddelde wachttijd indien de variabiliteit stijgt. Echter is dit niet altijd het geval. Wanneer men bijvoorbeeld kijkt naar het linkse gedeelte van de witte curves (5 servers) in Figuur 7 (zie Δ en □), is er zelfs een lichte verschuiving naar links te zien. Als de toename in de variabiliteit van de servicetijden het verlatingspercentage zodanig doet stijgen dat de wachttijden zullen dalen, dan zal er een verschuiving naar links zijn. Hogere verlatingspercentages zorgen namelijk voor kortere wachtrijen en bijgevolg voor lagere gemiddelde wachttijden. Een stijging in de variabiliteit heeft dus geen eenduidige invloed op de gemiddelde wachttijd, wat dus verklaard kan worden door die trade-off met het verlatingspercentage. Echter is de invloed van de variabiliteit op de trade-off tussen verlatingspercentage en gemiddelde wachttijd klein. Soms is het terugdringen van die variatie in de servicetijden de enige optie die bedrijven hebben om de servicekwaliteit te verbeteren Daarom is er in Figuur 7 ook een curve toegevoegd met een variatiecoëfficiënt van nul, wat dus de best mogelijke trade-off weergeeft gegeven het vast aantal servers en aankomstpatroon. Aangezien de curves in Figuur 7 niet ver uit elkaar liggen, kan er geconcludeerd worden dat de variabiliteit in een wachtrijsysteem niet veel verschil zal maken voor de prestatiemaatstaven van het systeem, wat toch wel een onverwacht resultaat is.



Symbool	Gemiddeld geduld
Δ	2.5
□	5

Symbool	Gemiddeld geduld
✚	10
○	15

Symbool	Gemiddeld geduld
⊠	20
◇	25

Symbool	Gemiddeld geduld
⊠	30
●	Oneindig geduld

Triangulaire verdeling van geduld (in minuten): TRIA(0.5 ; Gemiddeld geduld ; 2*Gemiddeld geduld)

Figuur 7: Trade-off bij verschillende variabiliteit in de servicetijden (5, 6 en 7 servers)

5.4. Is er een link van de resultaten naar de drie regimes (QR, ER & QER)?

Eerder werden er drie regimes besproken: het kwaliteitsgedreven regime (QR), het efficiëntiegedreven regime (ER) en het kwaliteits- en efficiëntiegedreven regime (QER). In deze sectie wordt er nagegaan of deze drie regimes terugkomen in de gevonden resultaten. In Figuur 8 worden er vier tabellen weergegeven met betrekking tot de gemiddelde bezettingsgraad, de gemiddelde kans op verlatingsgedrag (wat overeenkomt met het verlatingspercentage), de gemiddelde kans op wachten en de gemiddelde wachttijd. Om de wachtrijsystemen onderzocht in deze studie onder te verdelen in de regimes, werd er gebruikgemaakt van de formules toegelicht in Garnett et al. (2002). Zoals eerder aangehaald is er bij wachtrijsystemen een trade-off tussen efficiëntie en kwaliteit. Streeft men naar efficiëntie, dan streeft men naar een hoge bezettingsgraad, wat ten koste zal gaan van de kwaliteit. Bijgevolg benadert de kans op wachten in het efficiëntiegereven regime 100% (=1). De servicecapaciteit die nodig is in het efficiëntiegedreven regime kan bepaald worden aan de hand van volgende formule:

$$N = R - \varepsilon R \text{ (met } \varepsilon > 0)$$

Hierbij staat N voor het aantal servers, R voor de nominale behoefte en ε voor de kans op verlatingsgedrag (Garnett et al., 2002). De nominale behoefte R is in dit simulatiemodel gelijk aan 5 servers, aangezien bij 5 servers de theoretische bezettingsgraad gelijk is aan 100% (zie Tabel 5). Door het aantal servers N in te vullen in de formule, kan de kans op verlatingsgedrag (ε) voor elk systeem berekend worden. Bij 1 server is ε gelijk aan 80%, bij 2 servers aan 60%, bij 3 servers aan 40% en bij 4 servers aan 20% (wat overeenkomt met de vlakke curves in Figuur 2). Vanaf 5 servers zou ε kleiner dan of gelijk zijn aan nul, wat maakt dat deze systemen niet onder het efficiëntiegedreven regime vallen en bijgevolg de formule van het efficiëntiegedreven regime niet meer geldt. In Figuur 8 zijn de wachtrijsystemen met een kans op verlatingsgedrag dichtbij de hierboven berekende epsilons beschouwd als efficiëntiegedreven regimes. Echter mogen we niet alle wachtrijsystemen met minder dan 5 servers aan het ER-regime toewijzen, omdat het beperkte geduld en bijgevolg het verlatingsgedrag ook een rol spelen. Zo wijken de wachtrijsystemen met 3 en 4 servers met de kleinste geduldwaarden (bij 4 servers zelfs de 2 kleinste geduldwaarden) meer dan 3% af van de berekende epsilons en presteren over het algemeen beter op kwaliteit (i.e., kleinere kans op wachten en lagere wachttijden) en slechter op efficiëntie (i.e., hogere bezettingsgraden) dan de andere scenario's die onder het ER-regime vallen. Daarom worden deze drie wachtrijsystemen beter aan het kwaliteits- en efficiëntiegedreven regime toegewezen. De overige wachtrijsystemen moeten dus nog onderverdeeld

worden onder het QR-regime of QER-regime. Volgens Garnett et al. (2002) zou de kans op wachten en verlatingsgedrag bij het QR-regime naar 0 moeten naderen. In Figuur 8 is te zien dat de kans op verlatingsgedrag 0% benadert bij 7 servers en voldoende geduld. Echter is de kans op wachten bij geen enkel wachtrijsysteem onder de 25%. Bijgevolg werden de wachtrijsystemen met 5, 6 en 7 servers toegewezen aan het QER-regime en wordt er geen wachtrijsysteem uit deze studie toegewezen aan het QR-regime. Om de kans op wachten naar nul te brengen, zullen er nog meer servers nodig zijn. Merk hierbij wel op dat de gemiddelde wachttijd bij 7 servers al kleiner is dan 1 minuut, dus ook al is er een kans op wachten van 25% à 30%, dat hoeft niet te betekenen dat de klant ook effectief heel lang zal moet wachten.

ER	QER	QR

Gemiddeld geduld* (minuten)

Aantal servers	30	25	20	15	10	5	2,5	GEMIDDELDE KANS OP VERLATEN
1 server	78,83%	79,02%	79,20%	79,40%	79,58%	79,71%	80,25%	
2 servers	58,02%	58,33%	58,62%	58,95%	59,27%	60,08%	61,08%	
3 servers	37,25%	37,64%	38,10%	38,65%	39,40%	40,49%	43,02%	
4 servers	18,64%	19,20%	19,58%	20,25%	21,51%	24,29%	27,88%	
5 servers	4,57%	5,07%	5,87%	7,22%	8,98%	12,19%	16,06%	
6 servers	0,75%	1,02%	1,26%	1,95%	2,98%	5,00%	8,13%	
7 servers	0,13%	0,18%	0,30%	0,46%	0,83%	1,90%	3,74%	

Variatiecoëfficiënt = 1

Gemiddeld geduld* (minuten)

Aantal servers	30	25	20	15	10	5	2,5	GEMIDDELDE KANS OP WACHTEN
1 server	99,89%	99,90%	99,91%	99,92%	99,92%	99,84%	99,41%	
2 servers	99,66%	99,66%	99,66%	99,66%	99,57%	98,68%	95,80%	
3 servers	99,28%	99,28%	99,25%	99,05%	98,38%	94,49%	87,95%	
4 servers	97,73%	97,26%	96,49%	95,11%	92,09%	84,85%	75,50%	
5 servers	83,69%	82,67%	81,42%	78,97%	75,60%	67,44%	58,96%	
6 servers	56,43%	55,96%	54,34%	53,93%	51,63%	46,13%	41,21%	
7 servers	30,64%	30,77%	31,31%	31,02%	30,12%	28,88%	25,78%	

Variatiecoëfficiënt = 1

Gemiddeld geduld* (minuten)

Aantal servers	30	25	20	15	10	5	2,5	GEMIDDELDE BEZETTINGSGRAAD
1 server	99,99%	99,99%	99,99%	99,99%	99,98%	99,77%	98,36%	
2 servers	99,55%	99,58%	99,54%	99,53%	99,45%	98,89%	96,46%	
3 servers	99,04%	99,03%	99,13%	98,97%	98,72%	96,76%	93,20%	
4 servers	98,27%	98,11%	97,80%	97,19%	95,96%	92,76%	88,68%	
5 servers	92,60%	92,20%	91,80%	90,68%	89,43%	86,27%	82,61%	
6 servers	81,57%	81,39%	80,74%	80,64%	79,85%	77,38%	75,12%	
7 servers	69,95%	69,88%	70,22%	70,26%	69,69%	69,32%	67,88%	

Variatiecoëfficiënt = 1

Gemiddeld geduld* (minuten)

Aantal servers	30	25	20	15	10	5	2,5	GEMIDDELDE WACHTTIJD
1 server	36,24	30,01	23,66	17,29	10,95	4,74	1,90	
2 servers	29,81	24,75	19,63	14,40	9,17	4,00	1,59	
3 servers	23,34	19,35	15,35	11,26	7,18	3,10	1,26	
4 servers	15,74	13,04	10,33	7,60	4,88	2,22	0,94	
5 servers	6,60	5,73	4,85	3,90	2,75	1,39	0,65	
6 servers	2,09	2,02	1,78	1,62	1,29	0,76	0,41	
7 servers	0,65	0,65	0,66	0,61	0,54	0,39	0,23	

Variatiecoëfficiënt = 1

** Triangulaire verdeling van geduld (in minuten): TRIA(0.5 ; Gemiddeld geduld ; 2*Gemiddeld geduld)*

Figuur 8: Resultaten onderverdeeld in drie regimes (QR, ER & QER)

6. Conclusies en inzichten

In deze masterproef wordt gefocust op wachtrijsystemen waarbij klanten de neiging hebben om de wachtrij te verlaten wanneer ze ongeduldig worden, wat ook wel bekend staat als verlatingsgedrag. Meer specifiek wordt de trade-off onderzocht tussen de verwachte wachttijd voor klanten die daadwerkelijk service hebben gekregen, en het percentage klanten dat de rij verlaat vanwege de te lange wachttijd. Hiervoor wordt gebruik gemaakt van stochastische simulatie, waarbij diverse systeemcondities worden onderzocht zoals de servicecapaciteit, de variabiliteit van de servicetijden en de kansverdeling van het geduld van de klant.

Om de kwaliteit van de dienstverlening te beoordelen, mag het verlatingspercentage niet genegeerd worden, aangezien systemen met weinig geduld dan altijd een betere service bieden vanwege de lagere gemiddelde wachttijden (minder klanten zullen in de wachtrij blijven staan). Echter gaat weinig geduld gepaard met een hoger omzetverlies doordat meer klanten de wachtrij zullen verlaten (de wachttijd overschrijdt het beperkte geduld van de klant). Echter bevestigt deze studie bovenstaande verwachtingen niet helemaal. Zoals verwacht zal meer (minder) geduld leiden tot hogere (lagere) gemiddelde wachttijden, ongeacht het aantal servers. Maar de hoeveelheid geduld heeft niet altijd invloed op het verlatingspercentage. Het geduld heeft alleen in intrinsieke stabiele systemen een significant effect op het verlatingspercentage, waarbij het verlatingspercentage zal stijgen (dalen) wanneer het geduld afneemt (toeneemt). In de intrinsiek instabiele systemen ondervindt het verlatingspercentage (bijna) geen invloed van de hoeveelheid geduld, wat kan worden toegeschreven aan de overbezetting van deze systemen, waarbij de wachtrij nooit helemaal leeg raakt, zelfs als het geduld zou afnemen. Verder werd onderzocht hoe de variabiliteit in de servicetijden en de servicecapaciteit van invloed zijn op de trade-off tussen de gemiddelde wachttijd en het verlatingspercentage. In systemen met variabiliteit zullen klanten moeten wachten op hun beurt en zal er congestie ontstaan. Uit de resultaten van deze studie blijkt dat de variantie in de servicetijden bijna geen invloed heeft op de trade-off. Dit zou betekenen dat het verlagen van de variantie in de servicetijden de wachtrijprestaties bijna niet zal verbeteren. De servicecapaciteit heeft echter wel een significante invloed op die trade-off. Zo zal het verhogen van het aantal servers de servicekwaliteit verbeteren en efficiëntie verlagen. Tot slot is onderzocht of er een verband bestaat tussen de resultaten en de drie regimes (QR, ER en QER), waarbij gefocust wordt op kwaliteit, efficiëntie of beide. Hieruit blijkt dat in deze studie enkel het efficiëntiegedreven (ER) en kwaliteits- en efficiëntiegedreven (QER) regime is onderzocht. Tijdens deze studie werd ook opgemerkt

dat, ondanks een kans op wachten die groter is dan 0%, de gemiddelde wachttijd verrassend klein blijft, wat aangeeft dat er geen recht evenredigheid bestaat tussen beide.

Dankwoord

Graag wil ik als eerste mijn promotor Prof. dr. Inneke Van Nieuwenhuyse bedanken voor de vlotte communicatie en de uitstekende begeleiding tijdens het schrijven van mijn masterproef. Haar nuttige feedback en aanmoediging hebben mij uitstekend geholpen om mijn onderzoek tot een succesvol einde te brengen. Daarnaast wil ik de Universiteit Hasselt bedanken voor het verstrekken van de nodige middelen en faciliteiten om mijn onderzoek uit te voeren. Hoewel er verder geen specifieke personen zijn die ik kan bedanken, ben ik dankbaar voor de ondersteuning die ik heb gekregen tijdens mijn studie aan deze universiteit. Tot slot wil ik ook mijn familie en vrienden bedanken voor hun aanmoediging en steun gedurende mijn hele academische reis. Mede dankzij hun steun en liefde heb ik mijn masterproef succesvol kunnen afronden.

Bibliografie

[1] Bhandari, A., Scheller-Wolf, A., & Harchol-Balter, M. (2008). An exact and efficient algorithm for the constrained dynamic operator staffing problem for call centers. *Management Science, 54*(2), 339-353.

[2] Brandt, A., & Brandt, M. (1999). On a two-queue priority system with impatience and its application to a call center. *Methodology and Computing in Applied Probability, 1*(2), 191-210.

[3] Cheng, F., & Huo, J. (2013). The Staffing Requirements with Time-Varying Demand and Customer Abandonment in Call Centers. *Innovation and Supply Chain Management, 7*(1), 19-24.

[4] Corominas, A., & Lusa, A. (2012). LETRIS: staffing service systems by means of simulation. *Journal of Industrial Engineering and Management (JIEM), 5*(2), 285-296.

[5] Dai, J. G., & He, S. (2012). Many-server queues with customer abandonment : a survey of diffusion and fluid approximations. *Journal of systems science and systems engineering, 21*(1), 1-36. doi:10.1007/s11518-012-5189-y

[6] Defraeye, M., & Van Nieuwenhuyse, I. (2016). Staffing and scheduling under nonstationary demand for service: A literature review. *Omega, 58*, 4-25.

[7] Duffield, N. G., Massey, W. A., & Whitt, W. (2001). A nonstationary offered-load model for packet networks. Telecommunication Systems, 16(3), 271-296.

[8] Feldman, Z., Mandelbaum, A., Massey, W. A., & Whitt, W. (2008). Staffing of time-varying queues to achieve time-stable performance. *Management Science, 54*(2), 324-338.

[9] Fu, M. C., Marcus, S. I., & Wang, I.-J. (2000). Monotone optimal policies for a transient queueing staffing problem. *Operations research, 48*(2), 327-331.

[10] Gans, N., Koole, G., & Mandelbaum, A. (2003). Telephone call centers: Tutorial, review, and research prospects. *Manufacturing & Service Operations Management, 5*(2), 79-141.

[11] Garnett, O., Mandelbaum, A., & Reiman, M. (2002). Designing a call center with impatient customers. *Manufacturing & Service Operations Management, 4*(3), 208-227.

[12] Green, L. V., Soares, J., Giglio, J. F., & Green, R. A. (2006). Using queueing theory to increase the effectiveness of emergency department provider staffing. *Academic Emergency Medicine, 13*(1), 61-68.

[13] Harrison, J. M., & Zeevi, A. (2005). A method for staffing large call centers based on stochastic fluid models. *Manufacturing & Service Operations Management, 7*(1), 20-36.

[14] He, B., Liu, Y., & Whitt, W. (2016). Staffing a service system with non-Poisson non-stationary arrivals. *Probability in the Engineering and Informational Sciences, 30*(4), 593-621.

[15] Helber, S., & Henken, K. (2010). Profit-oriented shift scheduling of inbound contact centers with skills-based routing, impatient customers, and retrials. *OR spectrum, 32*(1), 109-134.

[16] Izady, N., & Worthington, D. (2012). Setting staffing requirements for time dependent queueing networks: The case of accident and emergency departments. *European Journal of Operational Research, 219*(3), 531-540.

[17] Jennings, O. B., Mandelbaum, A., Massey, W. A., & Whitt, W. (1996). Server staffing to meet time-varying demand. *Management Science, 42*(10), 1383-1394.

[18] Kelton, W. D., Sadowski, R. P., & Zupick, N. B. (2015). *Simulation with Arena* (6th ed. Vol. 635p): New York, N.Y. : McGraw-Hill Education.

[19] Kim, J. W., & Ha, S. H. (2012). Advanced workforce management for effective customer services. *Quality & Quantity, 46*(6), 1715-1726.

[20] Liao, S., Koole, G., Van Delft, C., & Jouini, O. (2012). Staffing a call center with uncertain non-stationary arrival rate and flexibility. *OR spectrum, 34*(3), 691-721.

[21] Liu, Y., & Whitt, W. (2012a). The G t/GI/st+ GI many-server fluid queue. *Queueing Systems, 71*(4), 405-444.

[22] Liu, Y., & Whitt, W. (2012b). Stabilizing customer abandonment in many-server queues with time-varying arrivals. *Operations research, 60*(6), 1551-1564.

[23] Saltzman, R. M. (2005). A hybrid approach to minimize the cost of staffing a call center. *International Journal of Operations and Quantitative Management, 11*(1), 1.

[24] Thompson, G. M. (1993). Accounting for the multi-period impact of service when determining employee requirements for labor scheduling. *Journal of Operations Management, 11*(3), 269-287.

[25] Whitt, W. (2004). Efficiency-driven heavy-traffic approximations for many-server queues with abandonments. *Management Science, 50*(10), 1449-1461

[26] Whitt, W. (2006). Staffing a call center with uncertain arrival rate and absenteeism. *Production and Operations Management, 15*(1), 88-102.

[27] Whitt, W. (2007). What you should know about queueing models to set staffing requirements in service systems. *Naval research logistics, 54*(5), 476-484. doi:10.1002/nav.20243

[28] Zeltyn, S., Marmor, Y. N., Mandelbaum, A., Carmeli, B., Greenshpan, O., Mesika, Y., . . . Lauterman, T. (2011). Simulation-based models of emergency departments: Operational, tactical, and strategic staffing. *ACM Transactions on Modeling and Computer Simulation (TOMACS), 21*(4), 1-25.