Serviceniveau en kosten

Eigenlijk zou er meteen ingegaan moeten worden op de vier basiselementen mensen, cultuur, systemen en processen. Alvorens echter daar iets over te zeggen worden in de eerstvolgende twee hoofdstukken eerst een aantal technisch getinte aspecten besproken, die van belang zijn bij planningsprocessen en bij keuzes die moeten worden gemaakt. Noem het maar een stukje theorie. Misschien niet voor iedereen even leuk maar wel belangrijk voor het begrip. Allereerst iets over de relatie tussen serviceniveau en kosten.

Onder serviceniveau wordt verstaan: het percentage interacties dat binnen een gewenste tijdslimiet wordt aangegaan. Bijvoorbeeld het percentage telefonische oproepen dat binnen 20 seconden wordt beantwoord of het percentage bezoekers dat binnen vijf minuten aan de beurt is.

Telefonische oproepen komen evenals baliebezoekers in het algemeen willekeurig binnen. Het enige wat meestal te achterhalen is, is een bepaalde trend over de dag. Te lang op de beurt wachten wekt irritatie maar daar staat tegenover dat je ook niet “rücksichtslos” een grote afdeling gaat opzetten om altijd maar mensen beschikbaar te hebben om meteen de telefoon aan te nemen of balieklanten te woord te staan. Dat zou niet erg economisch zijn omdat je dan de kans hebt dat dit verspilling van tijd zou betekenen omdat de medewerkers te lang op telefoontjes of bezoekers zitten te wachten.

Erlang

Nu heeft ongeveer een eeuw geleden de Deense wiskundige Agner Krarup Erlang een theorie ontwikkeld op basis waarvan je statistisch kunt bepalen wat de kans is dat bijvoorbeeld een telefonische oproep binnen 20 seconden wordt beantwoord.

Om hier iets over te kunnen zeggen moeten we rekening houden met een aantal variabelen, dat de service kan beïnvloeden en onderdeel uitmaakt van de zogenaamde Erlang-C formules (wachtrij theorie):

- Het aantal interactie aanvragen (bijvoorbeeld het aantal telefoontjes) per uur
- Het aantal mensen dat hiervoor beschikbaar is
- Gemiddelde behandeltijd
- Gewenste reactietijd

Op basis hiervan is het percentage oproepen dat statistisch binnen de reactietijd wordt beantwoord te berekenen. Overigens mag je in plaats van het aantal oproepen ook het aantal binnenkomende bezoekers bij een balie nemen.

Nu kun je de formule ook anders gebruiken. Je kunt bijvoorbeeld met een gewenst serviceniveau (laten we zeggen: ik wil dat 80% van de oproepen binnen 20 seconden wordt beantwoord) en met inzicht in het aanbod aan telefoontjes en de gemiddelde gespreksduur, berekenen hoeveel medewerkers je daarvoor nodig hebt.

Een te lage capaciteit aan medewerkers maakt dat weliswaar de bezettingsgraad van de medewerkers hoog is (kosten effectief) maar de bereikbaarheid en kwaliteit sterk wordt aangetast wat vaak leidt tot negatieve berichtgeving (reputatie schade). Bij een te grote capaciteit geldt het omgekeerde. De bereikbaarheid zou dan prima kunnen zijn maar de bezettingsgraad wordt te laag waardoor de kosten per telefoontje hoger zijn dan noodzakelijk.

Je kunt het vergelijken met de exploitatie van een tennispark. Als je te weinig velden hebt ten opzichte van de vraag, dan zal de wachttijd voordat je kunt gaan spelen, vervelend lang worden. Wil je die wachttijd verkorten door meer velden aan te leggen, dan loop je kans dat de bezettingsgraad te laag wordt en daarmee ook de opbrengsten.

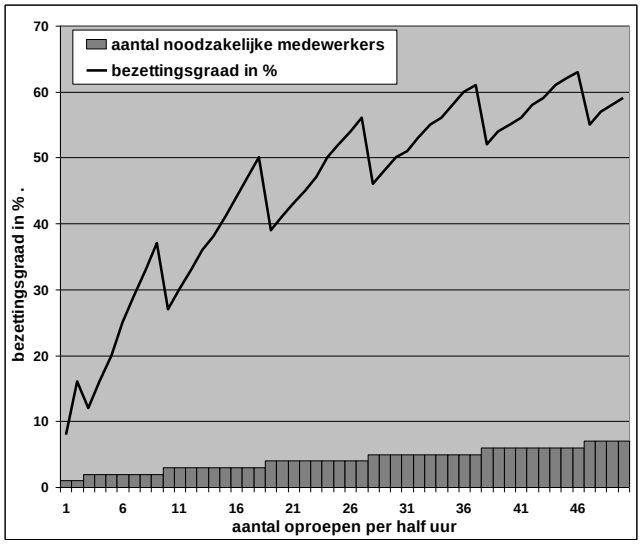
Zoals hierboven al is aangegeven kun je met Erlang-berekeningen ook het serviceniveau bepalen van een fysiek serviceloket (bijvoorbeeld een balie in een gemeentehuis of postkantoor) en omgekeerd kun je uitrekenen hoeveel loketten je nodig hebt om aan een bepaald serviceniveau te kunnen voldoen.

Combineren

Soms worden werkzaamheden gecombineerd met andere taken (bijvoorbeeld telefoon met e-mail) waardoor deze problematiek enigszins kan worden verkleind. De werkinhoud moet overigens wel van vergelijkbaar niveau zijn. Het is niet echt motiverend als medewerkers in de rustige periodes, stopwerk van een dusdanig niveau te doen krijgen dat men gaat uitkijken naar ander werk. Overigens is de bereikbaarheid niet de enige en allesbepalende factor voor de kwaliteitsbeleving van de klant. Daarover later meer.

Hieronder is een grafiek weergegeven waarbij is uitgegaan van een telefonische helpdesk met een gemiddelde gesprekstijd van 2 minuten en een nawerktijd van 30 seconden. We hebben als eis gesteld dat 80% van de oproepen binnen 20 seconden moet worden beantwoord.

Op de horizontale as is het aantal telefoontjes per half uur aangegeven. Op de verticale as het aantal noodzakelijke medewerkers om het serviceniveau van 80/20 te bereiken (grijze staafjes) en het bezettingspercentage van die medewerkers (zwarte lijn). Je ziet dat wanneer er een medewerker (agent) bijgeplaatst wordt, het bezettingspercentage afneemt. Logisch, men krijgt het iets rustiger door de extra collega. Je ziet ook dat hoe groter het aantal binnenkomende telefoontjes en dus ook van het aantal medewerkers, hoe minder de invloed van een extra medewerker op de bezettingsgraad. Als je de zwarte lijn denkbeeldig doortrekt dan zul je tot de ontdekking komen dat de maximaal haalbare bezettingsgraad op ongeveer 70% ligt. Wil je een hogere bezettingsgraad halen, dan moet het serviceniveau worden verlaagd.



figuur 1 Aantal medewerkers en bezettingsgraad als functie van aanbod

Wat nu ook blijkt, is dat kleine groepen erg oneconomisch zijn als je die medewerkers geen aanvullend werk kunt bieden (lage bezettingsgraad) tenzij je bereid bent om een heel slecht serviceniveau te accepteren.

Het zal ook duidelijk zijn dat hier een belangrijke factor ligt om klantenservice efficiënt in te richten. Hoe kun je bij lage hoeveelheden oproepen of bezoekers toch efficiënt werken? En hoe ga je om met een grillig verloop van het aantal telefoontjes of bezoekers over de dag?

Beschikbaarheid

Dat brengt ons ook bij het punt openingstijden (beschikbaarheid). Enerzijds is dit een afgeleide van de visie van de organisatie; goede service of beperkte service. Beperkte openingstijden werken vaak negatief doordat in steeds meer huishoudens mensen pas na normale kantoortijden in de gelegenheid zijn om te bellen. Ook de opkomst van webwinkels geeft enige druk op het ruimer openstellen van een telefonische klantenservice. Als men in de avond op een webwinkel kijkt en er vragen rijzen dan wil een koper graag meteen even navraag doen.

Ruimere openingstijden werken echter kostenverhogend omdat de bezettingsgraad van de medewerkers er negatief door wordt beïnvloed. Het liefst zou je alle telefoontjes binnen een scherp gedefinieerde openingstijd van één uur willen hebben want dan krijg je grote volumes en dat werkt

efficiënter. Maar zo werkt dat in de praktijk natuurlijk niet en bovendien zijn we dan aan het suboptimaliseren. Want buiten dat uur zouden alle bedrijfsmiddelen renteloos staan.

Voor het efficiënt opzetten van een klantenservice, waarbij een goede beschikbaarheid **en** een goed serviceniveau wordt nagestreefd, zijn creatieve oplossingen noodzakelijk. Zo kun je bijvoorbeeld proberen te verwijzen naar een website of de klant vragen om een e-mail te sturen. Een weinig fraaie oplossing als een klant om vijf over vijf met een urgente vraag zit. Ervaringsdeskundigheid om voor- en nadelen van oplossingen goed af te wegen, is hierbij niet overbodig. Met name de volumes spelen hierbij een grote rol. Als hoge volumes worden verwacht dan zijn ruime openingstijden vaak beter haalbaar. In het volgende hoofdstuk wordt hier verder op ingegaan.