

Debrecen Egyetem
Informatika Kar

Call centerek matematikai modellezése

Diplomamunka

Témavezető: Dr Sztrik János
MTA doktora
egyetemi tanár

Készítette: Kovács József
Programtervező matematikus
szakos hallgató

Debrecen
2009

1 Tartalomjegyzék

1	Tartalomjegyzék.....	2
2	Bevezetés.....	3
3	Alapvető tudnivalók a valószínűségszámítás köréből.....	5
3.1	Poisson eloszlás definíciója és tulajdonságai	5
3.2	Az exponenciális eloszlás definíciója és tulajdonságai.....	6
3.3	A normális eloszlás definíciója és tulajdonságai. A standard normális eloszlás.....	7
4	Összefüggések a sorbanállási modellek és a Call Centerek között	10
4.1	Sorbanállási rendszerek jellemző fogalmai és a Kendall jelölés.....	10
4.2	A Call Centerek, mint sorbanállási rendszerek	11
5	A Call Centerek hatékonyságmutatói, és alapvető összefüggések.....	13
5.1	A Call Centerek hatékonyságmutatói.....	13
5.2	Little Formula.....	15
5.3	PASTA tulajdonság.....	16
6	Erlang C (M/M/N rendszerek)	17
6.1	Az Erlang C formula használata.....	19
6.2	Az Erlang C formula tulajdonságai	21
6.3	A négyzetgyökös létszámfeltöltő szabály	24
6.4	Mennyire jó az Erlang formula?.....	25
6.5	Az Erlang formula implementálása.....	27
7	M/G/N Modell.....	29
7.1	Négyzetgyökös biztonsági létszámfeltöltés (Square-root safety staffing)	30
7.2	Működtetési rendszerek, egyesítés (pooling) és méretgazdaságosság	33
8	Erlang B modell (M/G/N/N sorok)	37
8.1	A blokkolás a gyakorlatban.....	37
8.2	Az Erlang C kiterjesztése	38
8.3	Az Erlang B formula	39
8.4	Gyakorlatok az Erlang formulára	40
9	Erlang A Modell (M/M/N+M sor)	43
9.1	Tűrelmetlen ügyfelek	45
9.2	Formalizálás és jelölések.....	48
9.3	Az M/M/N/B+M Modell pontos számítása.....	50
9.4	Az eloszlások közelítése és a működtetési rezsimek	54
9.5	Implementáció	59
9.5.1	Paraméterek becslése.....	60
9.5.2	A kiszolgálási színvonal.....	61
9.5.3	Közelítések	62
9.5.4	Gyakorlati szabályok.....	64
10	Összefoglalás.....	68
11	Irodalomjegyzék.....	70
12	Függelék	72
13	Köszönetnyilvánítás	75

2 Bevezetés

A telefonos ügyfélszolgálatok sok vállalat nélkülözhetetlen részét képezik, a gazdasági szerepük jelentős, és folyamatosan növekszik. Emellett nagyon érdekes szociotechnikai rendszerek is ([9]), amelyekben az ügyfelek, és a munkatársak viselkedése szorosan összefonódik a fizikai hatékonyságmutatókkal. Ilyen rendszerekben a tradicionális működtetési modellek nagyon hasznosak, ugyanakkor fundamentális okok miatt korlátozottan alkalmazhatóak egy rendszer teljesítményének jellemzésekor.

A telefonos ügyfélszolgálatok vagy a követőik, az ügyfélkapcsolati központok egyre inkább a vállaltok ügyfelekkel történő kommunikációjának preferált és elterjedt eszközeivé válnak. A legtöbb szervezet, amelynek egyéni ügyfelei vannak, már átalakította az infrastruktúráját, hogy egy vagy több ügyfélszolgálatot hozzon létre. Ez az átalakítás nem csak a magánvállalatokra jellemző, hanem a kormányzati vagy a sürgősségi szolgáltatásokra is. Előfordul, hogy nem egy belső szervezeti egységet hoznak létre az ügyfélszolgálat számára, hanem kiszervezik azt. Sok esetben, például a légitársaságoknál, kereskedelmi bankoknál és hitelkártya társaságoknál, ez az elsődleges kapcsolat az ügyfelek felé.

E telefonos szolgáltatások minőségével és működtetési hatékonyságával szemben támasztott elvárások rendkívül nagyok is lehetnek. Egy hatalmas, úgynevezett best-practice telefonos ügyfélszolgálaton, akár több száz ügyfélszolgálatos szolgálhat ki, óránként több ezer hívást. A munkaerő kihasználtság foka átlagosan 90-95% között van, egyetlen ügyfél sem kap foglalt jelet a hívásakor, viszont az 50 százalékuk hívását azonnal felveszik. A várakozó ügyfelek várakozási ideje néhány másodperc, és 1-2 százalék azoknak az aránya, akik a várakozás közben leteszik a telefont.

Ezzel egyidejűleg a best-practice telefonos ügyfélszolgálat inkább a kivételek közé tartozik, mintsem az uralkodó irányzathoz. A legtöbb ügyfélszolgálatnál, még a jól működőket is ideszámítva, nem sikerül folytonosan magas minőségű és hatékonyságú szolgáltatást elérni. Ez részben annak is köszönhető, hogy nem vagy kevésbé ismertek azok a tudományos elméletek, amelyek a best-practice működtetéshez szükségesek.

A teljesítmény probléma valószínűsíthetően az ügyfélszolgálati központok növekvő komplexitásával is kapcsolatba hozható. A hálózati technológiák újdonságai, a képesség alapú hívásirányítás, a multimédia mind egyre nagyobb kihívások elé állítják az ügyfélszolgálatok menedzsmentjét. Amíg a szimpla analitikus modellek tradicionálisan fontos szerepet játszanak a telefonos ügyfélszolgálatok menedzsmentjében, sok kívánnivalót hagynak maguk

után. Kifinomultabb megközelítések kellenek, amelyek pontosan leírják működtetést a valóságban, és a valóság e modelljei jelentősen javíthatják a telefonos ügyfélszolgálat hatásfokát.

A diplomamunkámban ilyen modellekkel szeretnék foglalkozni a témával foglalkozó szakirodalomból szemezgetve. Azonban mindenekelőtt a témakör valószínűségszámítási alapjait veszem sorra, majd a Call Centerek struktúráját, fogalmi rendszerét és hatékonyságmutatóit ismertetem, amely elengedhetetlen az olvasónak a modellek működésének és céljainak megértéséhez. Az elméleti bevezető után elsőként az Erlang C modellt tárgyalom, amely általában a Call Centerrel kapcsolatos vizsgálódások kiindulópontja. Sajnálatos hiányosságai miatt kerül sor először az M/G/N modell vizsgálatára, amely az Erlang C modell által feltételezett exponenciális kiszolgálási időket általános hosszúságúakra kicserélve próbálja a valóság pontosabb becslését adni. Ezután az Erlang B modell kerül sorra, amely abból a valós feltételezésből indul, hogy a Call Centerben végződtetett vonalak száma korlátozott, illetve további korlátozások is alkalmazhatóak az aktuális szituációban. Az Erlang C modellnél ilyen feltételezés nincs. Végül az Erlang A-t tárgyaljuk, amely az Erlang C azon megszorítását próbálja feloldani, hogy az ügyfelek nem várnak végtelen hosszú ideig a sorban, hanem otthagyják azt, amennyiben az elvárttól nagyban eltér a várakozási idejük.

3 Alapvető tudnivalók a valószínűségszámítás köréből

A következő részben a Poisson és az exponenciális eloszlás lényeges tulajdonságait ismertetem, amelyek alapvető jelentőséggel bírnak a sorbanállási elméletek körében. Ezután a normális és a standard normális eloszlásokat ismertetem, amelyek az $M/G/N$ és az Erlang A modell elméletében és gyakorlatában játszik szerepet. Í A standard normális eloszlás eloszlásfüggvényének közelítésére is felvázolok egy algoritmust. További részletek megtudhatóak az [1]-ből.

3.1 Poisson eloszlás definíciója és tulajdonságai

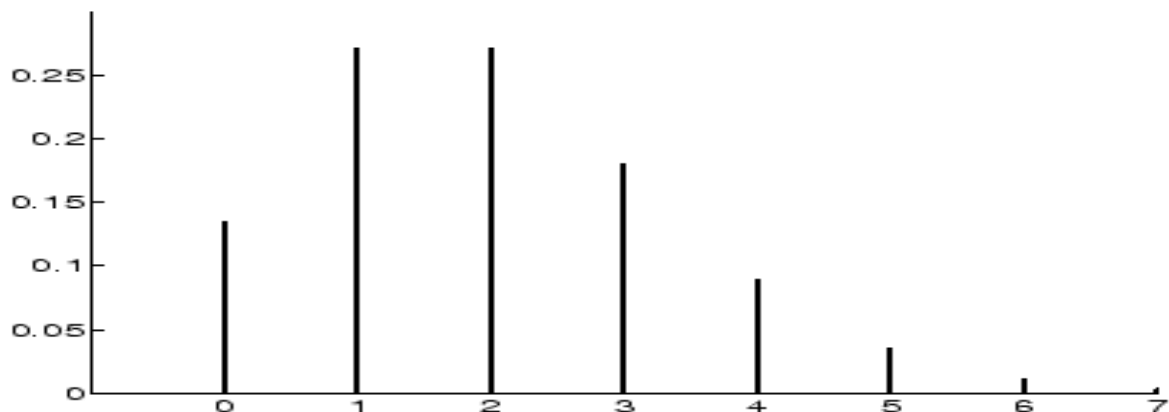
Legyen $\lambda > 0$. A $P(\xi = k) = p_k = \frac{\lambda^k}{k!} e^{-\lambda}$, $k = 0, 1, 2, \dots$ által definiált eloszlást, ahol $\lambda > 0$

állandó, Poisson-eloszlásnak nevezzük. Nyilván $p_k > 0$, és a $\sum_{k=0}^{\infty} \frac{\lambda^k}{k!} = e^{\lambda}$ képlet alapján

$\sum_{k=0}^{\infty} p_k = 1$. Így a fenti számok tényleg eloszlást alkotnak.

A $P(\xi = k)$ valószínűségek növekvőek, amíg k eléri λ -t (egészrész), utána csökkenőek (ha λ egész, akkor $k = \lambda - 1$ és $k = \lambda$ esetén is maximum van).

$\lambda = 2$ esetén a $p_k = \frac{\lambda^k}{k!} e^{-\lambda}$ értékei a következő ábrán láthatóak.



Ábra 1 A Poisson eloszlás

A karakterisztikus függvény: $\varphi(t) = \exp[\lambda(\exp(it) - 1)]$

A generátorfüggvény: $G(z) = \exp[\lambda(\exp(z) - 1)]$

A momentumok:
 $E\xi = \lambda$,
 $E\xi^2 = \lambda + \lambda^2$,

$$D^2\xi = \lambda,$$

$$\text{A centrált momentumok: } E(\xi - E\xi)^3 = \lambda,$$

$$E(\xi - E\xi)^4 = \lambda + 3\lambda^2,$$

$$\text{A ferdeség: } \beta_1(\xi) = \lambda^{-\frac{1}{2}}.$$

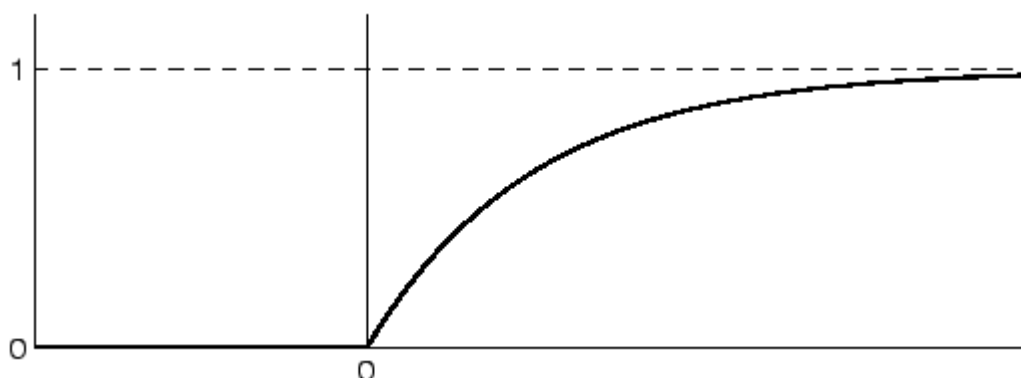
$$\text{A lapultság: } \beta_2(\xi) = \lambda^{-1}.$$

Ha ξ_1 és ξ_2 független Poisson-eloszlásúak λ_1 , illetve λ_2 paraméterrel, akkor $\xi_1 + \xi_2$ is Poisson-eloszlású $\lambda_1 + \lambda_2$ paraméterrel. Rajkov belátta, hogy igaz ennek a megfordítása is: ha ξ_1 és ξ_2 független valószínűségi változók, és $\xi_1 + \xi_2$ Poisson-eloszlású, akkor ξ_1 és ξ_2 is Poisson-eloszlású.

3.2 Az exponenciális eloszlás definíciója és tulajdonságai

A ξ valószínűségi változót λ paraméterű exponenciális eloszlásúnak nevezzük, ha eloszlásfüggvénye: $F(x) = \begin{cases} 0, & x \leq 0 \\ 1 - e^{-\lambda x}, & x > 0 \end{cases}$

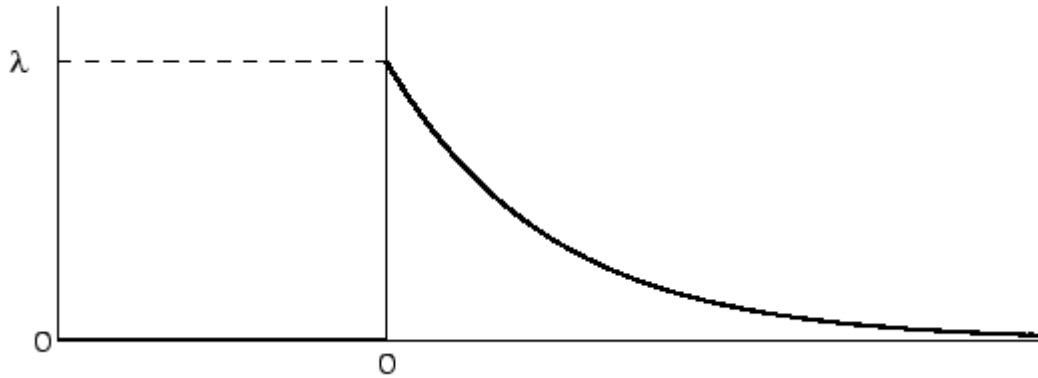
Ahol $\lambda > 0$ rögzített.



Ábra 2 Az exponenciális eloszlás

Az exponenciális eloszlás élettartamok és várakozási idők eloszlásaként lép fel. Az exponenciális eloszlás és a vele kapcsolatos más eloszlások a sorbanállás-elméletben és a megbízhatóság-elméletben használatosak.

$$\text{Az exponenciális eloszlás sűrűségfüggvénye: } f(x) = \begin{cases} 0, & x \leq 0 \\ \lambda e^{-\lambda x}, & x > 0 \end{cases}$$



Ábra 3 Az exponenciális eloszlás sűrűségfüggvénye

Az exponenciális eloszlás jellemző mennyiségei

A karakterisztikus függvény: $\varphi(t) = \left(1 - \frac{it}{\lambda}\right)^{-1}$

A momentumok: $E\xi^k = \frac{k!}{\lambda^k}$, $k = 1, 2, \dots$

Speciálisan, a várható érték és a szórásnégyzet: $E\xi = \frac{1}{\lambda}$, $D^2\xi = \frac{1}{\lambda^2}$.

A ferdeség: $\beta_1(\xi) = 2$. A lapultság: $\beta_2(\xi) = 6$.

Az exponenciális eloszlás „örökifjú”: $P(\xi < t + s \mid \xi \geq t) = P(\xi < s)$, $t > 0, s > 0$.

3.3 A normális eloszlás definíciója és tulajdonságai. A standard normális eloszlás.

A normális eloszláson alapul a statisztika klasszikus elméletének túlnyomó része.

Az ξ valószínűségi változót normális eloszlásúnak nevezzük, ha sűrűségfüggvénye:

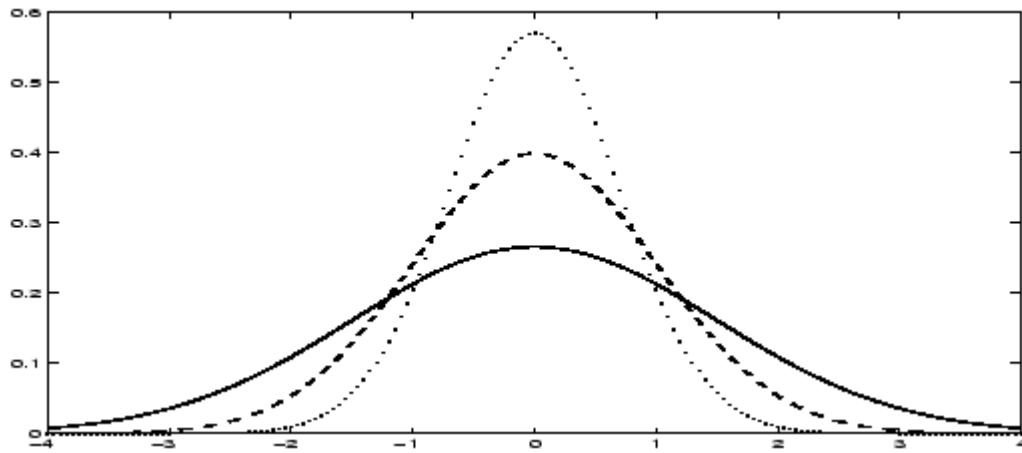
$$f(x) = \frac{1}{\sqrt{2\pi}\sigma} \exp\left(-\frac{(x-m)^2}{2\sigma^2}\right),$$

ahol $m \in \mathbb{R}$, $\sigma > 0$.

Jelölése: $\xi \sim N(m, \sigma^2)$

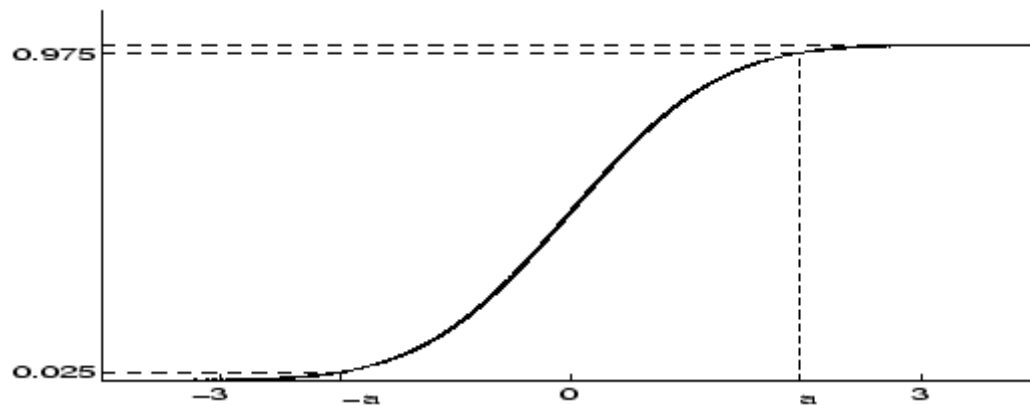
Az f grafikonja az úgynevezett haranggörbe (Gauss-görbe). Az f függvény m -re szimmetrikus, f szigorúan monoton növekvő a $(-\infty, m]$ intervallumon. $m \pm \sigma$ -ban f -nek inflexiós pontja van. m -ben f -nek maximumhelye van, a maximum értéke $\frac{1}{\sqrt{2\pi}\sigma}$. σ

növelésével a harang alakú görbe „laposabbá” válik, a σ csökkentésével pedig „csúcsosabbá”. Az ábrán normális sűrűségfüggvények láthatóak $m=0$ esetén.



Ábra 4 A normális eloszlás sűrűségfüggvénye

A normális eloszlásfüggvényre nincs zárt formula, de vannak jó közelítések.



Ábra 5 A normális eloszlás

Ha $\xi_0 = N(0,1)$, akkor ξ_0 -at standard normális eloszlásúnak nevezzük. A sűrűségfüggvénye $\phi(x) = \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{x^2}{2}\right)$.

A standard normális eloszlásfüggvény egy approximációját Abramowitz és Stegun írták le a [2]-ben. Ezt a későbbiekben fel fogom használni az $M/G/N$ és az Erlang A modell közelítéseiben. Jelölje Φ a standard normális eloszlásfüggvényt. Ekkor

$$\Phi(x) = \begin{cases} 0 & \text{ha } x < -6 \\ 1-n & \text{ha } -6 < x < 0 \\ n & \text{ha } 0 \leq x \leq 6 \\ 1 & \text{ha } 6 < x \end{cases}$$

ahol $n = 1 - b((((b_5 t + b_4)t + b_3)t + b_2)t + b_1)t$

$$\text{ahol} \quad b = c_2 \cdot \exp\left(-x\left(\frac{x}{2}\right)\right), \quad t = \frac{1}{1+|x|p}, \quad b_1 = 0,31938153, \quad b_2 = -0,356563782, \\ b_3 = 1,781477937, \quad b_4 = -1,821255978, \quad b_5 = 1,330274429, \quad p = 0,2316419 \text{ és } c_2 = 0,3989423.$$

4 Összefüggések a sorbanállási modellek és a Call Centerek között

Ebben a fejezetben ismertetjük a sorbanállási modellek fogalmi rendszerét, amelyről további információkhoz juthatunk a témához kapcsolódó egyéb tankönyvekben, például itt: [[3], [4]]. Ezután a Call Centereket, mint sorbanállási modelleket vizsgálom, amelyről több is olvasható ebben a cikkben: [11].

4.1 Sorbanállási rendszerek jellemző fogalmai és a Kendall jelölés

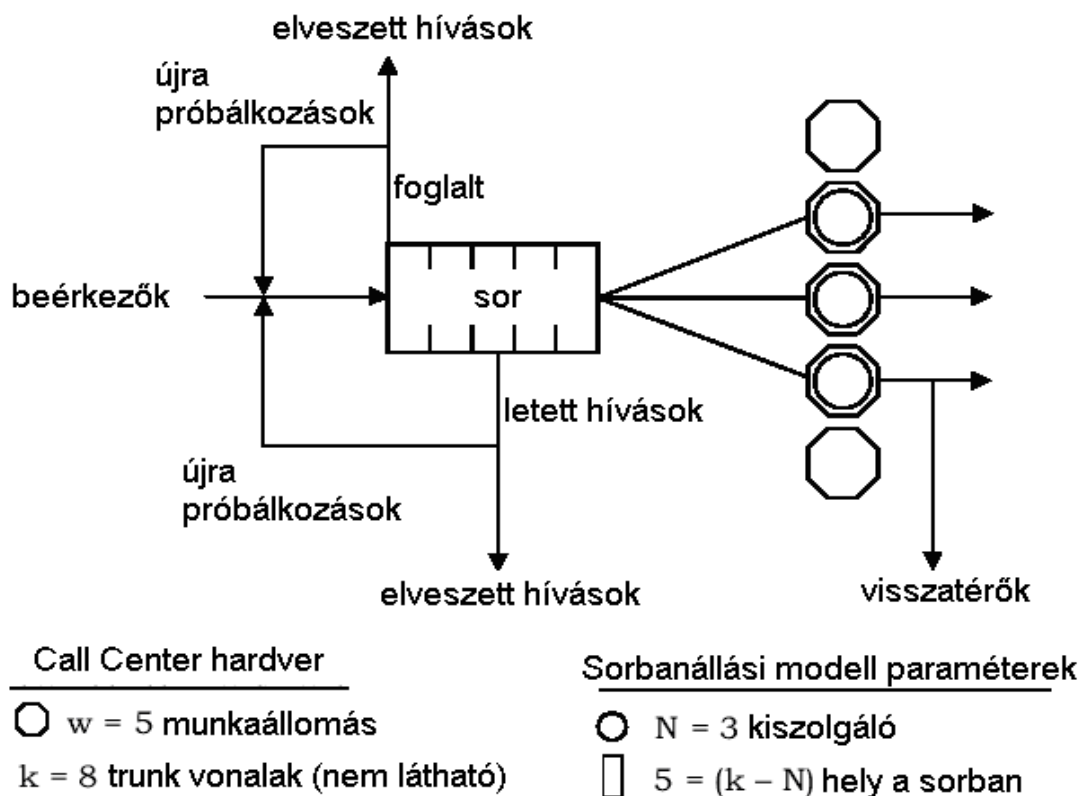
- a). Ügyfelek beérkezési folyamata: Legtöbbször feltételezzük, hogy a beérkezési időközök függetlenek és azonos eloszlásúak. Sok, gyakorlatban előforduló esetben az ügyfelek Poisson folyamat szerint érkeznek (azaz a beérkezési idők exponenciális eloszlásúak). Az ügyfelek egyenként vagy csoportosan érkezhetnek. A telefonos ügyfélszolgálatoknál a hívók egyenként érkeznek a sorba.
- b). Ügyfelek viselkedése: Az ügyfelek lehetnek türelmesek és hajlandók a várakozásra. Vagy lehetnek türelmetlenek, és elmehetnek. A telefonos ügyfélszolgálatoknál a hívó ügyfelek viselkedése speciális, mert nem láthatják az előttük álló sor hosszát. Ezért ők kezdetben türelmesebbek, majd az idő múlásával válnak egyre türelmetlenebbekké. Szokás az ilyen helyzetek feloldására a telefonba bementani a várható várakozási időt.
- c). Kiszolgálási idők: Általában feltesszük, hogy a beérkezési és a kiszolgálási idők egymástól függetlenek, és a kiszolgálási idők független, azonos eloszlású valószínűségi változók. A kiszolgálási idők lehetnek determinisztikusak, vagy exponenciális eloszlásúak. Ebből az is következhet, hogy a kiszolgálási idők függenek a sor hosszától.
- d). Kiszolgálási elv: Az igények kiszolgálása történhet egyenként, vagy csoportokban. A kiszolgálás sorrendjét tekintve beszélhetünk:
 - a) Az első érkező távozik elsőként (first come first served – FCFS)
 - b) Véletlenszerűen kiválasztott
 - c) Utolsóként érkező távozik elsőnek
 - d) Prioritásos (gyors, vagy legrövidebb kiszolgálási idejűek távoznak elsőként)
 - e) Processzor megosztásos (a processzor egyszerre szolgál ki, teljesítménye megoszlik)
- e). A kiszolgáló befogadóképessége: Egy vagy több egység végezheti az igények kiszolgálását.
- f). Várakozási terület: Korlátozásokat vezethetünk be, figyelembe véve a rendszerben tartózkodó igények számát.

Kendall gyorsírási jelölésrendszert vezetett be a különböző sorbanállási modellek jellemzésére, a háromrészes $a/b/c$ jelölést. Az első betű a beérkezési időközök eloszlását, a második betű a kiszolgálási idők eloszlását jelenti. Például a G jelenti az általános, az M az exponenciális, a D pedig a determinisztikus időket. Az harmadik, és egyben utolsó betű jelenti a kiszolgálók számát. Hogy más sorbanállási modelleket is lefedjen ez a rendszer, kiegészíthetjük egy új betűvel. Például egy exponenciális beérkezési és kiszolgálási idejű, N kiszolgálóval rendelkező, és N darab igény befogadására alkalmas várakozási területtel rendelkező rendszer rövidítése $M/M/N/N$.

Megjegyzés: A két M , Markov miatt van, az exponenciális eloszlás tulajdonságára az emlékezet nélküliségre utal. Mind a beérkezés, mind a kiszolgálási idők exponenciális eloszlásúak.

4.2 A Call Centerek, mint sorbanállási rendszerek

Tekintsük az alábbi ábrát.



Ábra 6 Call Centerek struktúrája

Ez a Call Center a következőképpen van felépítve. k darab trunk vonalon keresztül érkeznek be a hívások a Call Centerbe. Van $w \leq k$ munkaállomás, amelyeket gyakran neveznek üléseknek is. $N \leq w$ számú kiszolgáló szolgál ki bejövő hívásokat. Ha egy bejövő

hívás mind a k trunk vonalat foglaltnak találja, foglalt (busy) jelzést kap, és a rendszer további használata nem folytatódhat. Egyéb esetekben a hívás kapcsolva lesz a Call Centerhez, és lefoglal egyet a szabad vonalakból. Ha kevesebb, mint N kiszolgáló foglalt, akkor a hívás kapcsolva lesz az egyikhez, a kiszolgálás megkezdődhet. Ha a hívás több mint N kiszolgálót talál foglaltnak, de kevesebb, mint k hívás van folyamatban a rendszerben, akkor egy sorban (queue) kezd várakozni arra egy kiszolgálóra, hogy az szabad legyen. Ha az ügyfél türelmetlen lesz, akkor lehet, hogy feladja, mielőtt kiszolgálásra kerülne (abandon). Azok számára, akik végül is kiszolgálásra kerülnek, a kiszolgálási sorrend FCSF.

Ha egy hívás elhagyja a rendszert, akkor a birtokolt erőforrások, például a trunk vonal, a munkaállomás, az kiszolgáló, felszabadításra kerülnek, és elérhetővé válnak új beérkező hívások számára. Egyes hívások valamilyen okból nem kerülnek kiszolgálásra, ezek egy részét újra próbálják (retrials). A többi hívás, amelyet feladtak vagy foglalt jelzést kapott, számít elvesztett hívásnak (lost). Végül már kiszolgált ügyfelek is visszatérhetnek (return). Vagy azért mert újabb szolgáltatásokat is igénybe vennének, vagy azért mert az előző híváskor nyújtott szolgáltatással kapcsolatos problémák adódtak. Az első eset pozitívnak számít, míg a másodikat általában rosszabbnak tartják.

A trunk vonalak száma, a k , a rendszerben található összes, egyidejűleg kiszolgálás alatt lévő, vagy várakozó hívások számának felső korlátja. Hasonlóan, a hívások fogadására alkalmas kiszolgálók száma, az $N \leq w$ a felső korlátja a párhuzamosan kiszolgálható hívások számának. A beérkező hívások által okozott terhelés változásainak követésére, a nap folyamán, a Call Center menedzserek dinamikusan változtatják az aktuálisan dolgozó kiszolgálók számát.

Bármilyen rögzített N esetén konstruálható egy sorbanállási modell, amelyben a hívók az ügyfelek és N kiszolgáló dolgozik, a sor hívókból áll, amelyek kiszolgálásra várnak, és ehhez egy kiszolgálót kell kapniuk. Ha az N változik, akkor a sorban lévő helyek is változnak, $k-N$ -re. A rendszer bemenetei beérkező hívások, letett hívások és a kiszolgálási folyamatok statisztikái lennének. A modell alapvető kimenetei a feladott hívások számának hosszú távú alakulása, a sorban való várakozás idejének egyensúlyi (steady-state) eloszlása valamint a kiszolgálók foglaltságának a hosszú távú alakulása.

5 A Call Centerek hatékonyságmutatói, és alapvető összefüggések

5.1 A Call Centerek hatékonyságmutatói

Ebben a fejezetben megbeszéljük a Call Center menedzsment általános céljait. A rendszer olyan hatékonyságmutatóiról lesz szó, amelyek jó értékei esetén jelenthetjük ki, hogy a céljaink teljesülnek. Egy rendszer hatékonyságának vizsgálatakor, bizonyos dolgok rövid távon adottnak tekinthetők. Ilyenek például a költség, a munkaállomások száma, az ICT (Information and Communication Technology) infrastruktúra színvonala és a rendelkezésre álló megfelelően képzett munkaerő létszáma. Hosszú távon ezek bármelyike növelhető, de az ezzel kapcsolatos vizsgálódások nem képezik a dolgozatom tárgyát. Alapvető feltételezésem, hogy a vizsgált időszak rövid, maximum 30 perc. Ezen adott feltételek mellett kell vizsgálni a modellben a hatékonyságmutatók alakulását. Összefoglalva a menedzsment célja és feladata az, hogy az adott feltételek mellett a hatékonyságmutatók legjobb értékeit ériék el.

A következőekben iparági fogalmakról lesz szó, ezért megemlítem a szokásos iparági rövidítéseket is a matematikai jelölés mellett. A továbbiakban minkét jelölés szerepelni fog. Az alábbiakról több is megtudható itt: [10],[11] és [13].

Mielőtt azonban tárgyalnánk a hatékonyságmutatót meg kell említeni egy általánosan használt küszöbértéket, amelynek segítségével minden hívásról eldönthető, hogy az jó szolgáltatást kapott-e vagy sem. Ez az érték az **Elfogadható Várakozási Idő (Average waiting time - AWT)**. Gyakran használt értéke a 20 másodperc, ami azt jelenti, hogy amennyiben 20 másodpercnél hamarabb jutott a hívás a kiszolgálóhoz, akkor az ügyfél jó szolgáltatást kapott, ellenkező esetben nem. Ezek után tekintsük át a különböző hatékonyságmutatókat.

a). $P\{W > t\}$: Telefonos szolgáltatási faktor (Telephony Service Factor - TSF) vagy Szolgáltatási Színvonal (Service Level - SL). Ez azt mutatja meg, hogy mekkora a valószínűsége annak, hogy a várakozási idő több mint a t idő. Ez a t idő gyakran az AWT. Általában a jelölése x/t , ami fordított logikát követ. Az x azoknak az aránya százalékban, akik kevesebbet várokoztak mint t hosszúságú idő. A $(1 - P\{W > t\}) \cdot 100\%$ képlettel az x megkapható.

b). $E[W]$: Átlagos válaszolási sebesség (Average speed of answer - ASA). Az átlagos idő, amit a hívó várakozásban tölt, mielőtt az kiszolgáló felveszi a hívást.

- c). $E[W - t]^+$: Átlagos időtúllépés (Average excess time - AET). Ez a t idő gyakran az AWT.
- d). $P\{Bl\}$: Annak a valószínűsége, hogy az ügyfél foglalt jelzést kap. (Bl – Blocked)
- e). $P\{W > 0\}$: Annak a valószínűsége, hogy a várakozási idő nagyobb, mint 0. Ez azokat az ügyfeleket jelenti, amelyeket egyáltalán a várakozási sorba állítunk. A legjobb kiszolgálást mindig azok kapják, akik egyből kiszolgálóhoz jutnak.
- f). $P\{Ab\}$: Annak a valószínűsége, hogy egy hívó feladja a hívást (Ab – Abandoned). Ennek egy alacsony értéknek kell lennie. Körülbelül 3% még elfogadott iparági szabvány.
- g). $P\{W > t; Ab\}$: Annak a valószínűsége, hogy a hívás többet vár t időnél, és feladja a hívást. Itt különböző esetek lehetségesek. Az ügyfél feladja egy rövid t időn belül, például téves hívás esetén. Az a gyakorlat, hogy nem számolják azokat a hívásokat, amelyek egy minimális időt sem hajlandóak várakozni. A következő t időpont általában az AWT, ami előtt a hívás feladása még nem rontja a szolgáltatási színvonalat, de fontos lehet, például későbbi visszahívás miatt.
- h). $E[W; Ab]$: Azoknak a várakozási időknek az átlagos hossza, ami a hívás feladásával fejeződik be.
- i). $E[W; W > t; Ab]$: Azoknak a várakozási időknek az átlagos hossza, amelyek hosszabbak egy t időnél és a hívás feladásával fejeződnek be. Itt is érvényesek a g). pontban közölt esetek.
- j). $P\{W > t; \neg Ab; \neg Bl\}$: Azoknak az ügyfelek az aránya, akik hosszabb várakozási idő után (például AWT) kaptak csak kiszolgálást.

1. megjegyzés: A dolgozatomban a hatékonyságmutatók által vizsgált időszak mindig rövid lesz. Ebben biztosítható, hogy a hívásszám nem ingadozik jelentősen.

2. megjegyzés: A dolgozatom nagyobb részében feltételezzük, hogy a beérkező hívások azonos típusúak. Ez nagyban befolyásolja a beérkező hívások kiszolgálásának időtartamát, amelynek közel azonosnak kell lennie. Egyedüli kivétel az $M/G/N$ modell, amely az Erlang C modell általános kiszolgálási időkre való kiterjesztése.

Az előbbi hatékonyságmutatók alapvetően a sorban történő várakozással voltak kapcsolatosak. Azonban meg kell említeni más mutatókat is, amelyek szintén nagy hatással lehetnek a szolgáltatás minőségére. Ezek közvetett módon befolyásolhatják a többi hatékonyságmutató paramétereit is, amely az egész modell viselkedését megváltoztathatja. A

következő mutatók közvetlenül a kiszolgálás minőségével kapcsolatosak. Itt megemlítem őket, azonban később már nem foglalkozom velük a dolgozatomban.

a). Első Alkalommal Megoldott Ügyek Aránya (First time resolution - FTR). A legtöbb Call Centerben az ügyfelek valamilyen probléma megoldásáért telefonálnak, hogy az erre specializált kiszolgálók kezeljék az ügyeiket. Különösen hatékony, ha az ügyfelek problémáját az első alkalommal megoldjuk, és egy adott hívás nem generál visszahívásokat a későbbiekben, vagy hosszabb kezelési időket egy ideges ügyféllel kapcsolatban. Ez ugyanis elronthatja azt a feltételezésünket, hogy két vizsgált időszak között nincsen összefüggés.

b). Átlagos tartásban töltött idő (Average holding time - AHT). Ez a mutató a Call Centerben működő folyamatok, technikai eszközök és a kiszolgálók kiképzésének hatékonyságáról nyújthat képet. Az elérendő cél ennek a mutatószámnak a minimalizálása.

Vannak további hatékonyságmutatók, amelyeket meg kell említenem azért, hogy láthassuk, mennyire komplex lehet a modell. Például, mennyire barátságos a kiszolgáló az ügyféllel, vagy mennyire magabiztos az ügyfél problémájának kezelése közben. Ezek egy további rendszerelem a QC (Quality Control) modul hatáskörébe tartozik, és például a hívások visszahallgatásával, majd értékelésével mérhető.

5.2 Little Formula

A Little formula megadja az összefüggést $E[L]$, a rendszerben tartózkodó hívások átlagos száma, $E[S]$, az átlagos tartózkodási idő, és λ , az egységnyi idő alatt a rendszerbe érkező igények átlagos száma között. Az összefüggés:

$$E[L] = \lambda E[S]$$

Feltételezzük, hogy a rendszer megfelelő mennyiségű igény kiszolgálására alkalmas (azaz a rendszerben tartózkodó igények száma nem válhat végtelenné).

A fenti egyenlőtlenséget szemléletesen a következőképpen értelmezhetjük: tételezzük fel, hogy minden ügyfél 1 dollárt fizet a minden rendszerben eltöltött egységnyi időért. Ez kétféleképpen történhet. Az egyik lehetőség, hogy az ügyfelek folyamatosan fizetnek. Ekkor a rendszer átlagos bevétele $E[L]$ dollár. A másik lehetőség, hogy az ügyfelek az egész összeget a rendszer elhagyásakor fizetik ki, minden egységnyi időért 1 dollárt. Egyensúlyi helyzetben a rendszert egységnyi idő alatt elhagyó ügyfelek átlagos száma megegyezik az egységnyi idő alatt a rendszerbe belépő ügyfelek átlagos számával. Tehát a rendszer bevétele egységnyi idő

alatt $\lambda E[S]$. Nyilvánvaló, hogy a bevétel minkét esetben ugyanaz. E pontos bizonyítást megtalálható itt: [5], [6].

5.3 PASTA tulajdonság

A kizárólag Poisson beérkezésű, tehát $M/\cdot/\cdot$ rendszerekre vonatkozó PASTA (Poisson Arrivals See Time Averages – Poisson Beérkezések Időbeni Átlagokat Látnak) tulajdonság szerint a beérkező hívások nagy átlagban ugyanazt a helyzetet találják a rendszerben, mint amit tetszőleges időpillanatban egy külső megfigyelő észlel. Pontosabban, ha a beérkező igények egy hányada a rendszert A állapotban találja, az ugyanazt jelenti, mint az, hogy a rendszer az idő egy hányadában A állapotban van. Ez a tulajdonság csak Poisson beérkezésekre vonatkozik. A szemléletes magyarázat az, hogy a Poisson beérkezések teljesen véletlenszerűen történnek. A PASTA tulajdonság bizonyítása megtalálható az itt hivatkozott művekben: [7], [8].

6 Erlang C (M/M/N rendszerek)

Ebben a fejezetben bemutatjuk a híres Erlang C formulát azaz Erlang „késési” formulát, amely a dán matematikusról lett elnevezve, aki a formulát a 20. század elején levezette. További aspektusairól a [10][11] és a [12]-ben olvashat. Ebben a modellben egy fajta hívás van és nincsen feladott hívás, azaz minden hívó addig vár, amíg egy kiszolgáló fel nem veszi. Az egy időegység alatt átlagosan beérkező hívások számát jelöljük a görög betűvel λ_i -val. Az átlagos kiszolgálási intenzitást jelöljük μ_i -vel, és mérjük ugyanazzal az időegységgel. Az ajánlott terhelést jelöljük R_i -vel amit az $R_i = \lambda_i / \mu_i = \lambda_i E[S_i]$ képlettel számolunk ahol az $E[S_i]$ a kiszolgálás időtartamának várható értéke. Az egységnyi terhelés mértékegysége az Erlang. A kiszolgálók foglaltsága $\rho_i = \lambda_i / (N\mu_i) = R_i / N$. Megjegyzendő, hogy a mennyiségek többsége az i . időtartamban van értelmezve, amelyek kellően rövidek ahhoz, hogy azokban ne következzen be jelentős változás. A kiszolgálók számára az N hosszabb távon adott.

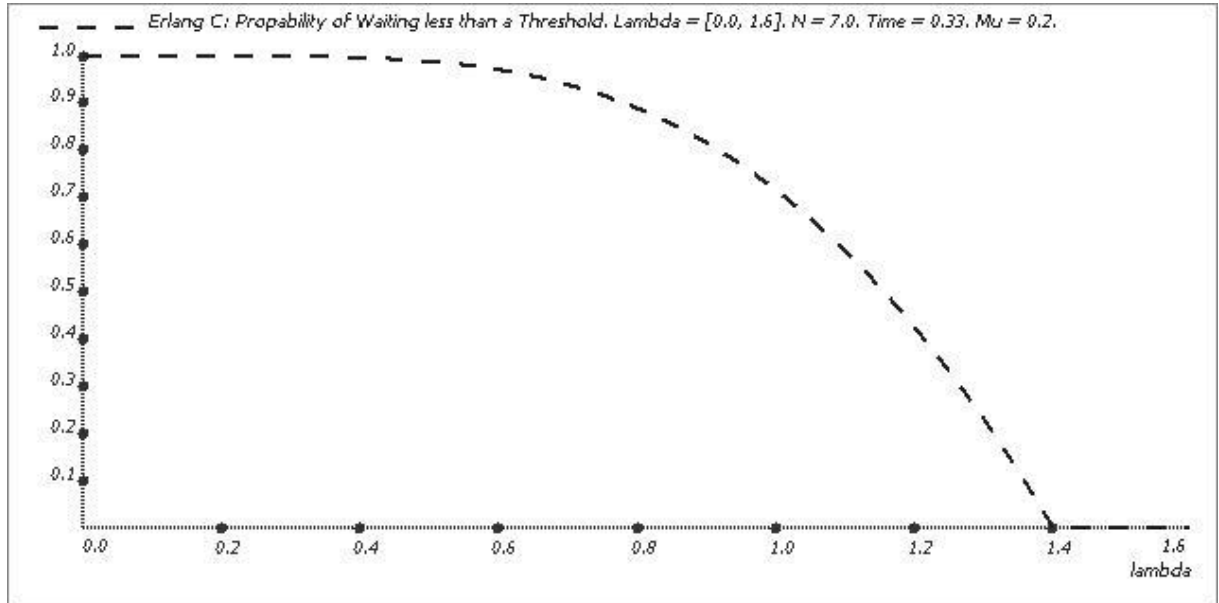
Az adott forgalmat kiszolgálók egy csoportja kezeli. Feltételezzük, hogy a kiszolgálók száma magasabb, mint a terhelés (emiatt $N > R_i$). Máskülönben, általában több hívás jön be, mint amennyit letesznek, emiatt a várakozó hívások száma folyamatosan nő, ami 0 TSF-et eredményez. Emiatt az N és az R_i közötti különbséget a rendszer többletkapacitásának nevezzük. Ez biztosítja, hogy az ajánlott terhelés változásait elnyelje a rendszer. Ezek a változások nem a λ_i és a μ_i változásai miatt vannak, hanem a hívás érkezések és szolgáltatási időtartamok belső véletlen viselkedéséből eredeztethetők. Emlékezzünk arra, hogy a λ_i és a μ_i átlagok, egy rövid időtartamra számolódnak. Sokszor azonban nagyobb hívásérkezési intenzitás van vagy a szolgáltatási időtartam olyan hosszú, hogy a rendszer alulteljesít. Az Erlang formula ereje abban rejlik, hogy képes számszerűsíteni a TSF-et (és más várakozási idővel kapcsolatos hatékonyságmutatókat) ebben a véletlenszerű környezetben, rövid alulteljesítési periódusokkal, sorok használatával.

Az Erlang formula megadja a TSF-et egy adott λ_i , μ_i , N és AWT mellett. A matematikai szempontból érdeklődő olvasó számára a pontos formula $R_i < N$ esetén a következő:

1

$$TSF = P\{Wait \leq T\} = 1 - P\{Wait > T\} \cdot P\{Wait > T | Wait > 0\} = 1 - C(N, R_i) \cdot e^{-N\mu_i(1-\rho_i)T}$$

A $C(N, R_i)$ annak a valószínűsége, hogy egy tetszőleges hívó, minden kiszolgálót foglaltnak talál, azaz késleltetés valószínűsége. Ha $R_i \geq N$ akkor a $TSF = 0$. A következő ábrán az Erlang formula látható rögzített μ_i , N és AWT esetén. A λ_i változó.



Ábra 7 A TSF változása

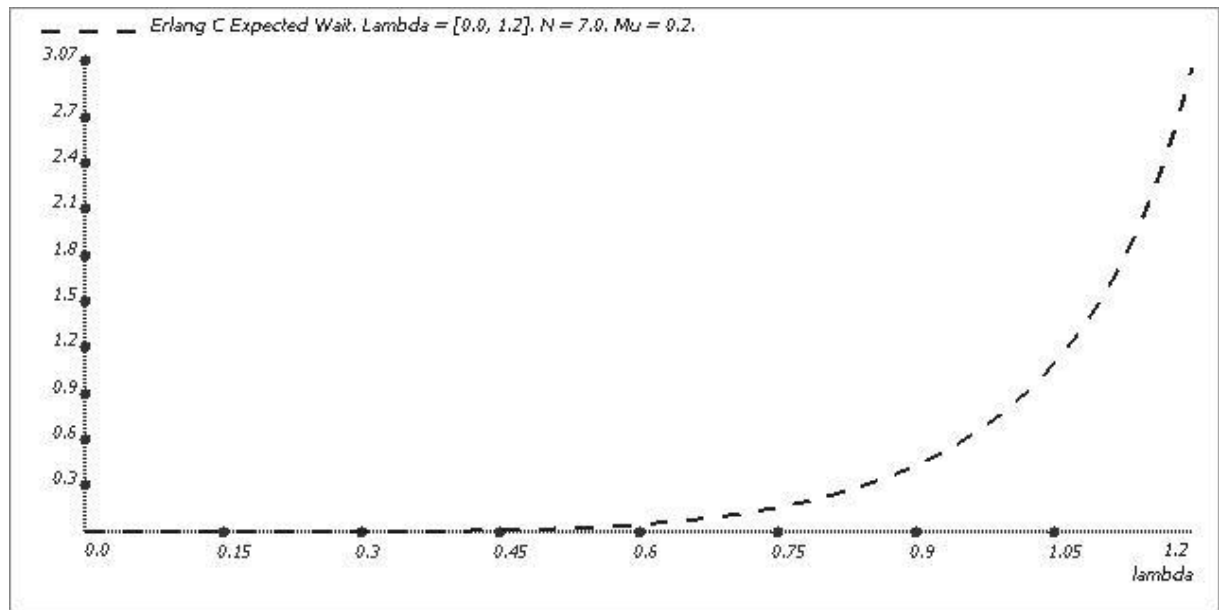
100%-ról indulva a TSF közel marad ehhez a felső szinthez még relatíve magas λ_i értékek esetén is. Ahogy a λ_i eljut ahhoz, hogy $R_i = \lambda_i / \mu_i$ megközelíti az N -t akkor a TSF elkezd csökkenni meredekebben addig, amíg eléri 0-át $\lambda = N\mu = 7/5 = 1,4$. Ettől a ponttól kezdve a TSF marad 0%.

Az SL mellett, ki tudjuk számítani az átlagos válaszolási időt, az ASA-t is. Az, hogy hogyan függenek egymástól, meg van adva az Erlang formula ASA számára készült változatával.

2

$$ASA = E[Wait] = P\{Wait > 0\} \cdot E[Wait|Wait > 0] = C(N, R_i) \cdot \left(\frac{1}{N} \right) \left(\frac{1}{\mu_i} \right) \left(\frac{1}{1 - \rho_i} \right)$$

A következő diagram az ASA változásait mutatja meg az előbb használt paraméterek használatával. Látható, hogy amikor a λ megközelíti az $N\mu = 1,4$ -et, akkor a várakozási idő meredeken nőni kezd.



Ábra 8 Az ASA változása a λ hatására.

A késleltetés valószínűségének meghatározása nem egy köztes lépés a TSF vagy az ASA számolásakor. Az önmagában is érdekes. Elmondja, hogy mennyi hívó kerül be a sorba és mekkora hányaduk fog kiszolgálót találni azonnal. Azáltal, hogy kiszámoljuk a TSF-et egy 0 AWT-hez, megkapjuk a késleltetési százalékot 100%-ból kivonva. Ezt elosztva százszal a késleltetési valószínűséghez jutunk. Azaz: $100 \times \text{a késleltetési valószínűség} = 100 - \text{TSF}$.

Eddig csak a szolgáltatási színvonallal kapcsolatos aspektusait tárgyaltuk az Erlang formulának. Szerencsére a kiszolgáló oldala relatíve egyszerű. Vegyük tekintetbe azt az esetet, hogy az $R_i < N$, emiatt $N - R_i$ a többletkapacitás. Mivel minden hívó elér egy kiszolgálót valamikor, az egész ajánlott terhelés az R_i , el van osztva N számú kiszolgáló között. Ez $R_i / N \times 100\%$ produktivitást eredményez mindnek, ha azt feltételezzük, hogy a terhelés egyenlően van elosztva köztük. Ha az $R_i \geq N$, akkor telítettség (saturation) következik be, azaz a kiszolgálók egyből hívást kapnak, amint felszabadulnak. Elméletben, ilyen magas produktivitás csak rövid ideig tartható fenn.

6.1 Az Erlang C formula használata

Az előző részben láttuk, hogy az Erlang C formula az átlagos várakozási idő kiszámítására használható, adott számú kiszolgáló, kiszolgálási és forgalom intenzitás esetén. A formula alkalmazható arra is, hogy más típusú kérdéseket válaszoljon meg, úgymint: adott μ_i és R_i , valamint a maximálisan elfogadható ASA vagy SL mellett, mi a maximum hívás mennyiség időegységenként, azaz mi az a λ_i , amit az ügyfélszolgálat még kezelni tud? A

$C(N, R_i)$ komplexitása miatt, nem tudjuk a formulát visszafordítani, de próba-hiba módszerrel megválaszolhatjuk az ilyen típusú kérdéseket.

A kérdés, amelyet természetesen a leggyakrabban feltesznek, az, hogy mi a legkevesebb számú kiszolgáló egy bizonyos terhelés szinthez és szolgáltatási színvonalhoz. Ez szintén megválaszolható próba-hiba módszerrel, de szoftverek gyakran automatikusan képesek megválaszolni.

A legtöbb eszköz egy egész számot a kiszolgálók számát adja vissza, mint választ. Ennek van is értelme, hiszen nem alkalmazhatunk például fél kiszolgálót. Azonban, alkalmazhatunk kiszolgálót fél munkaidőre. Emiatt, ha a szoftveres eszköz azt kívánja, hogy alkalmazzon 17,4 kiszolgálót fél órára, akkor 17 kiszolgálót kell alkalmazni 18 percre és a 18-at a fennmaradó 12 percre. 17 kiszolgálóval alatta vagyunk a szolgáltatási színvonalnak, 18-cal felette. Emiatt a rossz szolgáltatási színvonal 18 percre kompenzálódik azzal a 12 perccel, amikor az átlag felett szolgáltattunk.

Ha szemét megy be, szemét is jön ki. Ez a jól ismert frázis jellemző az Erlang formulára is. Emiatt az input paramétereket elővigyázatosan kell meghatározni. Leginkább a várt szolgáltatási intenzitásra becsült értékkel a μ_i -vel lehet hibát elkövetni. Ennek az az oka, hogy az gyakran az összes idővel számolni kell, amelyben a kiszolgáló nem vehet fel hívást. Ez azonban téves, mert az Erlang C modell számára a kiszolgálás akkor kezdődik, amikor az ACD hozzárendeli a hívást a kiszolgálóhoz. És akkor fejeződik be, amikor a kiszolgáló újra elérhetővé válik, például a hívásirányító rendszer ismét hozzá tud rendelni hívást a kiszolgálóhoz. Ezért a $E[S_i]$ nem csak a hívás időtartamát tartalmazza, hanem a reakció időt (akár 10 másodperc is lehet), és a szükséges utómunka időtartamát is (ami akár olyan hosszú is lehet, mint maga a hívás). Jegyezzük meg, hogy a reakció idő a hívó számára, mint várakozási idő jelenik meg. Ezt figyelembe kell venni a szolgáltatási színvonal kalkulációjánál, azaz csökkenteni kell az elfogadható várakozási időt az átlagos reakció idővel.

Egy lehetséges konklúziója az előbbieknak az lehet, hogy az kiszolgálókat stimulálni kell, hogy gyorsabban reagáljanak azért, hogy elkerüljük egy extra kiszolgáló ütemezését. Azonban, ezek a mechanizmusok, melyek a mennyiségi aspektusait javítják az ügyfélszolgálatoknak, a minőség romlásához vezethetnek, a munkahelyi nyomás növekedése miatt. Nem kezeljük a humán aspektusait sem az ügyfélszolgálati munkának. Jegyezzük meg, hogy 100%-os produktivitás egyáltalán nem lehetséges, és az a többletkapacitás, amit az

Erlang formula számítja ki, annak az egyik lehetséges eszköze, hogy a kiszolgálók megkapják a szükséges, rövid szüneteket a hívások között.

6.2 Az Erlang C formula tulajdonságai

Az Erlang formula ismerete egy dolog, az értése egy másik. Az Erlang formulának van egy pár tulajdonsága fontos menedzseri következményekkel. Ebben a részben megvitatjuk ezeket.

Robosztusság: Egy kiszolgálóval több vagy kevesebb, nagy eltérést okozhat az SL-ben, még a nagy ügyfélszolgálatokban is. Ez jó hír azoknak az ügyfélközpontoknak, amelyeknek alacsonyabb elvárásokat támasztó szolgáltatási színvonala van. Az SL relatíve kevés erőfeszítéssel növelhető egy elfogadható szintre. Más részről egy valamennyivel nagyobb terhelés figyelemre méltóan leronthatja az SL-t. Általában kijelenthetjük, hogy az Erlang formula nagyon érzékeny az input paraméterek (λ_i , μ_i és az N) egészen kis változásaira. Különösen abban az esetben, amikor az R_i közel van az N -hez, amint az előző ábrákon is látható volt. A görbe egyre meredekebb lesz, amint a λ_i megközelíti az $N\mu_i$ -t, és ezek a kis változások a vízszintes tengelyen, nagy változást okoznak a függőleges tengely mentén. Ez az érzékenység nagyon nehézzé teszi a Call Center menedzser feladatát. A hívások érkezésének kis, előre nem megjósolható változásai vagy néhány kiszolgáló nem várt hiányzása tönkre teheti az SL-t. Az N által a TSF-re gyakorolt hatás megfigyelhető itt [Ábra 21] is.

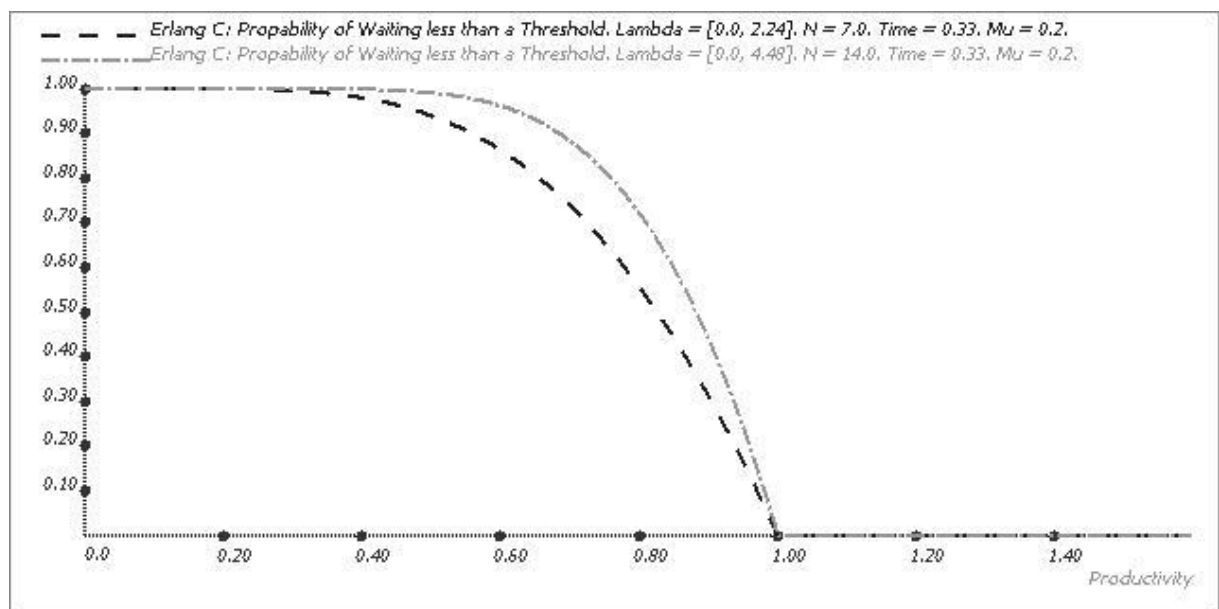
Az idő nyújthatósága. A második tulajdonság a hívás sajátosságaival, például a λ_i és a μ_i abszolút és a relatív értékeivel kapcsolatos. Emlékezzünk, hogy a terhelés az $R_i = \lambda_i / \mu_i$ összefüggéssel írható le. Ha mind a λ_i mind a μ_i megkétszereződik, akkor a terhelés ugyanaz marad. Ez nem jelenti, hogy ugyanazon számú kiszolgáló kell a egy meghatározott szolgáltatási színvonal eléréséhez.

Ha a λ_i -t megszorozzuk ugyanazzal a számmal, amivel a μ_i -t, akkor a terhelés ugyanaz marad, de ez olyan mintha a rendszer lelassulna. A várakozási idő bizonyíthatóan nő. Ha az AWT-t megszorozzuk ugyanezzel a számmal, akkor a TSF nem változik. Az ASA és az idő nyújtása közötti kapcsolat komplikáltabb.

Kis túlzással elmondhatjuk, hogy a terhelés nem érzékeny az idő nyújtására. Egyes hatékonyságmutatók csak az R_i és az N értékétől függenek, de nem függenek λ_i vagy a μ_i értékeitől külön-külön. A késleltetés valószínűsége, a $C(N, R_i)$ jó példa erre. Ugyanaz nem

tartható a TSF-re, mert a μ_i és a λ_i aktuális értéke fontos szerepet tölt be. Valójában, adott R_i és N értékre, az SL csak az $T\mu$ értékétől függ. Ezért, ha az időt nyújtom, és az AWT is hasonlóan nyújtva van, akkor a TSF ugyanaz marad. Természetesen ez csak az elmélet, habár az a gyakorlat, hogy az AWT magasabb olyan Call Centerekben, ahol hosszú beszélgetési idők vannak, összehasonlítva olyanokkal, ahol csak rövid ideig tartanak a beszélgetések. Az ASA számára az idő nyújtása szimpla. Az ASA-t ugyanazzal a faktorral kell nyújtani.

Méretgazdaságosság. Másik jól ismert tulajdonság, hogy a nagy Call Centerek hatékonyabban dolgoznak. Ez a méretgazdaságosság hatása: ha N -et megduplázzuk, akkor a λ_i -t több mint kétszeresére növelhetjük, mialatt megőrizzük ugyanazt az SL-t, feltételezve, hogy a μ_i és az AWT konstans marad.

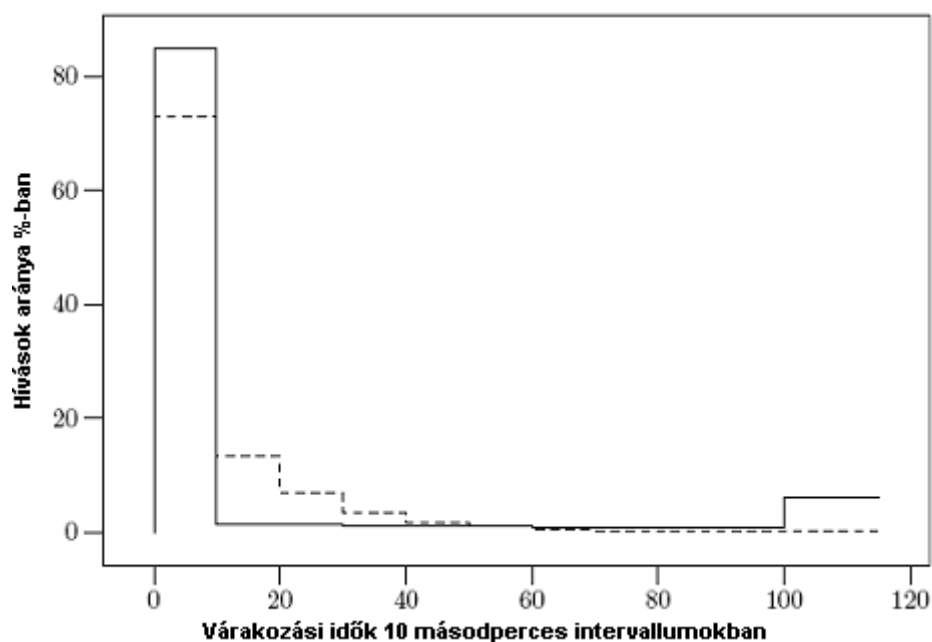


Ábra 9 A TSF alakulása a produktivitás függvényében a két különböző Call Centerben

Hogy mélyebb betekintést nyújtsunk a méretgazdaságosság témakörében, két szituációt rajzoltunk fel a fenti ábrára. A TSF értékeit rajzoltuk fel 7 és 14 kiszolgáló esetén. A produktivitást tettük a vízszintes tengelyre, és a TSF-et a függőlegesre. Látható, hogy a görbe meredeksége nagyobb, olyan produktivitás értékek esetén, amelyek 1-hez közel vannak, ami az Erlang formula érzékenységét fejezi ki a paraméterek kicsiny változtatása esetén. Fontos azonban megjegyezni, hogy a relatív javulás méretükben növekvő Call Centerekben csökken, ahogy a méret egyre növekszik. Az abszolút javulás azonban lassan de, növekszik.

Variációk a várakozási időknél. Vegyünk két Call Centert. Az egyiknél $\lambda_i = 1$, $\mu_i = \frac{1}{5}$ és $N = 8$, míg a másikonál $\lambda_i = 20$, $\mu_i = 3$, és $N = 8$ szintén. Mindkét Call Centernél a TSF

86% körül alakul, AWT=20 másodperc esetén. Akkor ez azt jelenti, hogy a két Call Center várakozási idői összehasonlíthatóak? Nem. Ennek tisztázására, vessen egy pillantást a következő ábrára, ahol a várakozási idők hisztogramjait ábrázoljuk, a két Call Centerben. A jobb oldalon azok a hívók vannak, amelyeknél a várakozási idő meghaladja a 100 másodpercet. Látjuk, hogy az első Call Centerben, amelyet folytonos vonallal jelölünk, az ügyfeleknek vagy egyáltalán nem kell várakozniuk vagy nagyon hosszú ideig várakoznak. Tulajdonképpen alig van hívó, amely ebbe a két kategóriába nem esik bele. A második Call Centerben kevesebb hívó kap kiszolgálót egyből, de nagyon kevesüknek kell sokat várni.



Ábra 10 A várakozási idők hisztogramja két különböző Call Centerben

Két konklúzió vonható le ebből a példából. Először: a TSF nem mond el mindent. De sokkal fontosabb az, hogy a Call Center karakterisztikájától függően, több vagy kevesebb variációja lehet a várakozási időknek. Csak például a TSF több AWT-val való alapos megvizsgálása tudja felfedni egy bizonyos Call Center karakterisztikáját.

A fennmaradó várakozási idő. Amikor beállunk egy sorba (például a postánál, vagy a supermarketben), akkor meg tudjuk becsülni a fennmaradó kiszolgálási időt az előttünk a sorban lévő ügyfelek számára alapozva. Általában a fennmaradó várakozás hosszának becsült értéke csökken, miközben várakozunk, amiatt, hogy látunk ügyfeleket távozni előttünk. De mi van fennmaradó várakozási idővel láthatatlan sorokban, mint amilyen a Call Centerekben is van? A matematika szerint az Erlang C modellben a fennmaradó várakozási idő állandó. Ezért nem számít, mennyi ideig várakozunk, az átlagos fennmaradó várakozási idő mindig ugyanaz. Hogyan lehet magyarázni ezt az első látásra nem logikus jelenséget? Amikor bekerülünk a

sorba, arra számítunk, hogy bizonyos számú hívás várakozik a sorban előttünk. Amikor kicsit várakoztunk, azt a következtetést vonjuk le, hogy a sor láthatóan hosszabb, mint amire számítottunk. Az Erlang formulából az következik, hogy ameddig várakozunk, az ügyfelek remélt száma mindig ugyanaz marad.

Lehetséges következmény lehet az ügyfelek számára az, hogy senkinek sem szabad letenni a kagylót mialatt várakozik. Miért tenné le valaki 1 perc után, ha a fennmaradó várakozási idő ugyanaz marad, mint akkor, amikor kezdte a hívást. A gyakorlatban azonban vannak jó okok is arra, hogy bizonyos idő után lerakjuk a telefont, mint ahogy vannak jó okok a vonalban maradásnak is. A hívás feladása mellett szóló érv, hogy az ügyfelek nem tudják a Call Center paramétereit, és emiatt nem tudják azt sem, hogy mi az átlagos várakozási idő abban a Call Centerben. Minél tovább várakozik valaki, annál nagyobb az esélye annak, hogy egy olyan Call Centerbe telefonál, amelynek nem kedvezőek a paraméterei, és ezért a fennmaradó várakozási idő még nő is! Másrészről, az Erlang C formula nem számít a hívások feladására. Ha valakinek a türelme nagyobb, mint azoké, akik előtte vannak a sorban, akkor ők valószínűleg előtte leteszik, és az illető végül ki lesz szolgálva. Egy olyan rendszerben, ahol a hívásokat tesznek le a sorban, az átlagos fennmaradó várakozási idő csökken a várakozás során.

6.3 A négyzetgyökös létszámfeltöltő szabály

Eddig láthattuk, hogy a méret növelése előnyökkel járhat a produktivitás és/vagy a szolgáltatási színvonal javítására. Ezek az előnyök az Erlang C formula használatával számszerűsíthetők. Hogy egy általánosabb érteshez jussunk, egy gyakorlati szabályt formálunk, amely rögzített szolgáltatási színvonal mellett teremt kapcsolatot a hívásmennyiség és a kiszolgálók száma között. Ez a kapcsolat a következőképpen néz ki:

$$\text{többletkapacitás \% -ban} \times \sqrt{N} = \text{állandó.}$$

Az állandó a formulában kapcsolódik a szolgáltatási színvonalhoz, a formula emiatt csak a többletkapacitástól és a kiszolgálók számától függ. A többletkapacitás aránya a formulában a $100 \times (1 - R_i/N)$ összefüggéssel adható meg. A gyakorlati szabályból így kapjuk meg a következő eredményt. Ha a Call Center négyszer akkora lesz, akkor a többletkapacitás körülbelül fele akkora marad.

Másfelől, könnyen benyomást szerezhethetünk arról, hogy mekkora lehet a beérkező hívásmennyiség, ha változtatjuk a foglaltság szintjét. Gyakrabban azonban a kiszolgálók

számát szeretik meghatározni egy megnövekedett hívásmennyiség kezeléséhez. Ennek a kiszámítása azonban komplexebb. Ha c -vel jelöljük a konstanst, amely a szolgáltatási színvonalhoz kapcsolódik, akkor a formula N -re a következő:

$$N = \left(\frac{c + \sqrt{c^2 + 4R_i}}{2} \right)^2$$

Ha a c egy nagyon kicsi érték, akkor látjuk, hogy az N arányos az R_i -vel. Ez azt jelenti, hogy a méretgazdaságosság növekedési potenciálja kisebb lesz nagyon nagyméretű Call Centerekben, mivel ezeknél a méretgazdaságosság már most is a lehető legnagyobb. Hogy mi számít nagynak ebben a kontextusban, az a szolgáltatási színvonalától függ.

6.4 Mennyire jó az Erlang formula?

Ebben a részben az Erlang formula gyenge pontjait vesszük sorra, és annak a mélyén meghúzódtó feltételezéseket. Ez fog motiválni más kifinomultabb modelleket, amelyekről szó esik majd a későbbiekben.

Meglepő lehet, hogy az ASA nagyobb, mint 0, holott többletkapacitás van. Az ok a hívások érkezésének és a kiszolgálás időtartamának változékonyságában keresendő. Ha az érkezési időpontok között egyforma időablak lenne, és a hívások hossza állandó lenne, akkor soha nem lenne várakozás. Azonban egy véltetlen környezetben, mint amilyen a Call Center is, kapacitáshiány mutatkozhat egy rövid ideig. Emiatt van a sorbanállás is. A sor végül kiürül mindig, ha megfelelő ideig többletkapacitás van. Az Erlang formula számszerűsíti a várakozás mennyiségét (ASA-ként vagy TSF-ként) egy bizonyos típusú random érkezéshez és kiszolgálási időhöz. Emiatt a matematikai véletlenül alapuló folyamatok, amelyek modellezik az érkezéseket és a távozásokat, nem mások, mint közelítések. A közelítés minősége és a formulának, a modell különböző aspektusainak megváltozására való érzékenysége eldönti, hogy a képlet elfogadható eredményeket ad-e. Most foglalkozunk a háttérben levő feltételezésekkel, és beszéljük meg a közelítések következményeit.

Hívás feladása. Egy jól dimenzionált Call Centerben néha lerakják a telefont a sorban. Nem modellezni ezeket azért nem durva egyszerűsítés. De vannak olyan Call Centerek, amelyek azért mutatnak teljesen más viselkedést, mint amit az Erlang formula jósol, mert nem számolnak a modellben, a hívások feladásával. Általában kimondható, hogy a hívások feladása csökkenti a várakozási időt mások számára, ami jó az SL számára. Olyan Call Centerekben, ahol az R_i közel van, vagy esetleg meghaladja az N -t, kritikus, hogy

modellezzék a hívások feladását. Ezek modellezésével foglalkozik a később tárgyalt Erlang A modell.

Újratárcsázás. Míg a hívások feladása relatíve jól értett dolog, az Erlang C formula igazán nagy probléma nélkül kiterjeszthető, hogy számoljon azzal. Ez nem igaz arra az esetre, amikor az ügyfelek, akik egyszer letették, újratárcsáznak. Kevés tudott azon ügyfelek viselkedéséről, akik újra próbálkoznak csakúgy, mint ezt a viselkedést matematikailag jól modellező rendszerekről sem.

Csúcsok a terhelésben. Formálisan, az Erlang formula nem enged ingadozásokat a beérkező terhelésben. Azonban, minden Call Centerben vannak napon belüli változások is. Addig, amíg ezek a változások limitáltak, és ami a fő, hogy nincsenek kapacitáshiányos periódusok, az Erlang C formula jól teljesít azokban a periódusokban, ahol csak kis fluktuáció jelentkezik a terhelésben vagy a kiszolgálók számában. Az Erlang C formula különböző idő intervallumokban történő használatakor, a teljes képet átlag számításával kaphatjuk meg. Azonban akkor, amikor a kapacitáshiány történik, a felgyülemlett hívások áttolódnak egyik periódusból a másikba. Ezt a lemaradást explicit módon modellezni kell, de ez nem lehetséges az Erlang C formula keretén belül. Ezért az Erlang C formula nem alkalmazható kapacitáshiány esetén. A forgalom rövid csúcsainál egyértelmű kapacitás kalkulációk alkalmazhatók, amelyek figyelmen kívül hagyják a rendszer random viselkedését.

A dolgozatban kisebb időperiódusokat feltételezek, amelyek nincsenek hatással egymásra, azaz függetlenek.

A hívás időtartamok típusai. Az Erlang C formula azon alapul, hogy a szolgáltatási idők az exponenciális eloszlásból jönnek. Anélkül, hogy a matematikai részleteibe mennénk, megjegyezhetjük, hogy a minden pozitív érték elképzelhető hívás időtartamként, emiatt nagyon hosszú és nagyon rövid értékek is. Azonban a legtöbb az átlag alatt van. Bizonyos mérések a standard telefonos forgalmon mutatják, hogy a hívástartamok megközelítően exponenciálisak, habár az irodalom eredményei nem teljesen értenek egyet ebben a tárgyban. Tipikus eset, amikor több hívástípus van különböző híváshossz átlagokkal, vagy amikor a hívás mindig eltart legalább egy meghatározott minimális ideig. Ilyenkor a hívástartamok nem exponenciálisak. Ebben az esetben, érdekes lehet tudni, milyen hatással vannak a különböző kiszolgálási idők az Erlang C formulára. Bizonyos elővigyázatossággal ki lehet következtetni, hogy csak az átlagos hívástartamnak van fontos hatása a Call Center teljesítményére.

Az általános eloszlású kiszolgálási időkkel kiterjesztett Erlang C modellel, az $M/G/N$ modellel foglalkozó fejezet foglalkozik.

Emberi viselkedés. Eddig figyelmen kívül hagytuk a kiszolgálók viselkedését, a hívás felvételénél felhasznált reakcióidőt leszámítva. Azonban a kiszolgáló viselkedése nem olyan egyszerű. Az alkalmazottak kivehetnek rövid szüneteket, hogy kávézzanak, megbeszéljenek dolgokat stb. Az emberi viselkedést modellezni, nagyon nehéz feladat, leírni és számszerűsíteni pedig még inkább. A legtöbb szituációban ezek a kis szünetek akkor vannak, amikor nincs hívás a sorban. Emiatt várható, hogy kis hatása van az SL-re. Más szituációkban nagyobb hatása van arra, és komolyan limitálhatja a mennyiségi modellezés lehetőségeit.

6.5 Az Erlang formula implementálása

A korábbiakban adtunk egy formulát az ASA és a TSF kiszámítására az Erlang modellben. A késési valószínűség fontos szerepet töltött be ebben a formulában. Ez a valószínűség a következő formulában adott:

$$C(s, a) = \frac{a^s}{(s-1)!(s-a)} \left[\sum_{j=0}^{s-1} \frac{a^j}{j!} + \frac{a^s}{(s-1)!(s-a)} \right]^{-1}$$

$$C(N, R_i) = 1 - \frac{\sum_{m=0}^{N-1} (R_i^m / m!)}{\sum_{m=0}^{N-1} (R_i^m / m!) + (R_i^N / N!)(1/(1 - R_i / N))}$$

Emlékezzünk, hogy az N a kiszolgálók száma, és az R_i a beérkező terhelés: a beérkezési ráta és az átlagos híváskezelési idő hosszának szorzata, azaz $R_i = \lambda_i / \mu_i$. A formula csak akkor működik, ha $N > R_i$, egyébként minden hívásnak a sorban kell várakoznia, emiatt a késés valószínűsége 1. Ez nem az, amit a formula ad, ezért az $N > R_i$ feltételt kell először vizsgálni. Ha az $N > R_i$, akkor a fenti formulát kell kiszámolni. Azonban ez gyakran numerikus túlcsoorduláshoz vezet, ami miatt teljesen rosszak lesznek az eredmények. Például, ha az $\frac{R_i^m}{m!}$ -t számítjuk, ami a fenti összegzés része, és már $R_i = m = 50$ esetén is extrém nagy számok osztásához vezet. Ezt kritikus elkerülni a korrekt eredmény szempontjából.

Emiatt, átírjuk a formulát a következőképpen, egyértelmű átalakításokkal:

$$C(R_i, N) = \left[1 + \frac{N - R_i}{R_i} \sum_{m=0}^{N-1} \frac{(N-1) \dots (m+1)}{R_i^{N-m-1}} \right]^{-1}$$

Koncentráljunk a $\sum_{m=0}^{N-1} \frac{(N-1) \dots (m+1)}{R_i^{N-m-1}}$ összegre. Ezt máshogyan kifejezve a következőt

kapjuk:

$$\left(\dots \left(\left(\frac{1}{R_i} + 1 \right) \frac{2}{R_i} + 1 \right) \dots \right) \frac{N-1}{R_i} + 1 \text{ amit már könnyű implementálni.}$$

Ezt a késleltetési valószínűséget a legnehezebb megkapni az Erlang formulában. Ezután az ASA és az SL már könnyen meghatározható. Megismételjük a formulát az ASA-ra, amelyet matematikai jelöléssel $E[W]$ -nek jelölünk:

$$E[W] = \frac{C(N, R_i)}{(N - R_i) \mu_i}$$

Az SL kissé komplexebb, mert az e matematikai állandót is magában foglalja.

$$P\{t < W\} = C(N, R_i) e^{-(N-R_i)t\mu_i}$$

Itt a P -t úgy kell olvasni, hogy „annak a valószínűsége, hogy ...”, és a t az AWT szerepét tölti be. A Call Centerekben sokszor az a kérdés, hogy egy adott szolgáltatási színvonalhoz mi az elvárt mennyiségű kiszolgáló, azaz az N ? Ez a típusú kalkuláció úgy történik, hogy az SL-t ismétlődően kiszámolják különböző N -hez, egy minimális N -t vesznek kiindulásként, például, amikor $N > R_i$. Ezután N -t 1-gyel növeljük addig, amíg elérjük a kívánt szintet. Ekkor megtaláltuk a megfelelő N -t.

Beszéltünk egy alternatív hatékonyságmutatóról is: amit átlagos időtúllépésnek neveztünk. A matematikai jelölése $E[W - t]^+$, ahol a t megint egyenlő az AWT-vel. Az AET-et a következő formulával adjuk meg.

$$E[W - t]^+ = \frac{C(R_i, N) e^{-(N-R_i)t\mu_i}}{\mu_i(N - R_i)}$$

A ehhez tartozó ábra a függelékben található: [Ábra 22].

7 M/G/N Modell

Erlang C alapú jóslatokról gyakran derül ki, hogy nagyon pontatlanok az alapul szolgáló feltételezések megsértése miatt, és e sérülések modellezése nem egyértelmű. Például a nem exponenciális szolgáltatási idők az $M/G/N$ sorhoz vezetnek, amely erősen különbözve az $M/M/N$ soroktól analitikusan konok. Erről a témáról többet is megtudhat itt [11], vagy az ebben a fejezetben található további hivatkozásoknál.

Emiatt hasznosak lesznek az approximációk mind a megértés, mind a modell robusztussága miatt. Amikor Call Centereket modellezünk, a leginkább hasznos approximációk tipikusan azok, amelyek nagy forgalmú rendszerek számára készültek, azok ahol a kiszolgáló kihasználtság magas. A magas forgalom feltételezése természetesen tükrözi a magas fokon kihasznált, hatalmas Call Centerek működtetésének természetét, a gyakorlatban a csúcsidőszakok feltételeit, amelyek meghatározzák az egész rendszer nagyságát.

Tekintsük az $M/G/N$ sort. Kis és közepes számú magas fokon kihasznált kiszolgálónál alkalmazható Kingman klasszikus „Torlódási törvénye” (Law of Congestion) [14] amely azt állítja, hogy a késés a sorban megközelítőleg exponenciális, amelynek a középértéke a következő:

3

$$E[\text{várakozás az } M/G/M \text{ soroknál}] \approx E[\text{várakozás az } M/M/M \text{ soroknál}] \times \frac{1+c_s^2}{2}$$

A $c_s = \sigma(S)/E[S]$ (A várakozási idők szórása osztva a várakozási idők várható értékével) jelöli a kiszolgálási idő variációs együtthatóját, egy mértékegység nélküli mennyiség, amely természetes módon méri a sztochasztikus változékonyságot. Továbbá ez a magas forgalmú rendszerhez tartozik, hogy alapvetően minden ügyfél tapasztal némi késleltetést mielőtt kiszolgálásra kerül.

Ekkor adott $C(N, R) \approx 1$ esetén az Erlang C ASA a következőképpen alakul.

$$E[\text{várakozás az } M/G/M \text{ soroknál}] \approx \left(\frac{1}{N}\right) E[S] \left(\frac{\rho}{1-\rho}\right) \left(\frac{1+c_s^2}{2}\right)$$

Ebből tisztán látható, hogy mind a ρ kihasználtságnak, és a c_s sztochasztikus változékonyságnak hatása van a torlódásra, a valóságban ez nem-lineáris, konvex módon növekvő. Valóban még a kihasználtságban (ρ -ban) bekövetkező kicsiny változásoknak is

lehet nyomasztóan negatív hatása a magas fokon kihasznált rendszerekben. A teljesítmény lerontható hosszabb és változékonyabb kiszolgálási időkkel, $E[S]$ -el és c_s^2 -el, de javítható megnövelt párhuzamossággal azaz N -nel.

7.1 Négyzetgyökös biztonsági létszámfeltöltés (Square-root safety staffing)

A kiszolgálók csapatának nagysága nap közben nagy, gyakran 100-300 fő. A terheltségi cél nagy, de nem a maximális forgalmú időszakokban. Miközben a közelítések a maximális vagy könnyű forgalmú időszakokhoz készültek addig az általunk vizsgált régió a kettő közé esik. Tipikus az, hogy a kiszolgálók 90-95%-a foglalt az zsúfolt időszakokban, de a kiszolgálók nagy számának köszönhetően csak az ügyfelek fele van késleltetve (Sze [15]).

A (3)-es közelítés a gyakorlatban azonosítja az exponenciális kiszolgálási időket, minden más kiszolgálási idővel együtt, ekkor $a c_s = 1$. Később látható lesz, hogy ez az azonosítás pontatlanná válik sok, magas fokon használt kiszolgáló esetében.

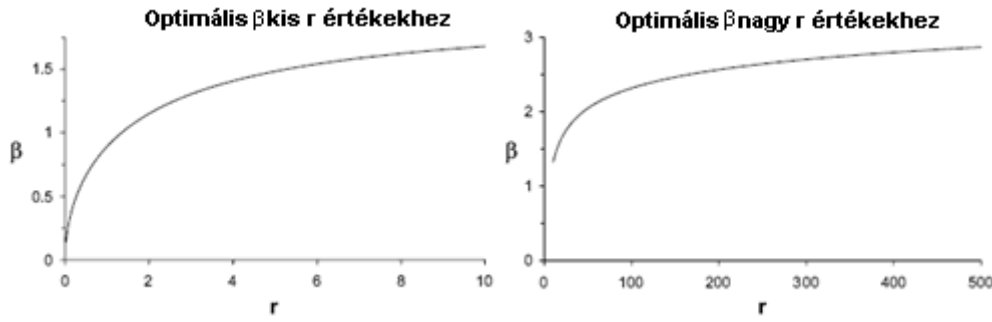
Valóban, a legtöbb Call Centernél az N nem 10 alatti, hanem több 10 vagy 100. És a nagyobb N teret ad aszimptotikus¹ rendszerek számára, amelyek különböznek a Kingman törvényében megfogalmazottól, mert az ügyfelek jelentős hányada nem szeret várni, és a szolgáltatás minőségének óvatos egyensúlyban kell lennie a kiszolgálók hatékonyságával. Az ilyen elvárásokkal rendelkező rendszereket QED (Quality and Efficiency Driven) azaz Minőség és Hatékonyság Vezérelt működtetési rendszereknek nevezzük. Később röviden egy másik irodalomban használt fogalommal, Racionalizált rezsimnek nevezzük [13].

A QED rendszer a $M/M/N$ rendszerek számára először Halfin és Whitt [16] által lett analizálva. Formálisan ebben a rendszerben a kiszolgálási intenzitás, a μ fix úgy, mint a $\alpha \in (0,1)$ a $P\{\text{Wait} > 0\}$ számára. Ez úgy van definiálva, hogy némelyik, de nem minden ügyfél várakozik a kiszolgálásra. Ha a λ és az N is tart a végtelenhez, akkor Halfin és Whitt demonstrálják, hogy:

4

$$P\{\text{Wait} > 0\} \rightarrow \alpha \Leftrightarrow \sqrt{N}(1 - \rho_N) \rightarrow \beta$$

¹ Ha egy görbének (lehet az függvénygörbe is) a végtelenbe vesző vége tetszőlegesen megközelít egy egyenest, de azt soha el nem éri, akkor e görbének ez a vége aszimptotikus tulajdonságú, és az az egyenes, amihez közelít, az a görbe aszimptótája



Ábra 11 Optimális β a lineáris várakozási és létszámfeltöltési költségekre (Borst et al. [17])

bizonyos fix szolgáltatási gradiensekre $\beta \in (0, \infty)$ úgy hogy a $\rho_N = \lambda / N\mu \uparrow 1$. Ekkor ők a következő aszimptotikus kifejezést adták az Erlang C formulához:

5

$$P\{\text{Wait} > 0\} \approx P(\beta) = \left[1 + \frac{\beta \Phi(\beta)}{\beta \phi(\beta)} \right]^{-1}$$

ahol a $\alpha = P(\beta)$ a (4)-ben. Itt a Φ és a ϕ egyenként a standard normális eloszlás- és sűrűségfüggvénye (középértéke 0 szórása 1).

Egy fix β szolgáltatási gradienchez a (4) sugallja a négyzetgyökös biztonsági létszámfeltöltést, amely azt ajánlja, hogy a kiszolgálók száma az N legyen:

6

$$N = R + \Delta = R + \beta \sqrt{R} \quad 0 < \beta < \infty$$

ahol, ismét $R = \lambda / \mu$ az ajánlott terhelés. Jegyezzük meg, hogy a (4) precízebben ekvivalens a $N \approx R + \beta \sqrt{N}$, de a $\beta \ll N$, ezért, a két kapcsolat alapvetően ugyanazt jelenti.

A $\Delta = \beta \sqrt{R}$ a biztonsági létszám a sztochasztikus változékonyság kiküszöbölésére. Jegyezzük meg, hogy a $\beta \leq 0$ -hoz 100%-os vagy annál nagyobb kihasználtság tartozik, emiatt egy instabil rendszer. Ahogy a β növekszik úgy nő a biztonsági létszám szintje. Megfordítva, $P\{\beta\} \approx P\{\text{Wait} > 0\}$ csökken a csökkenő β esetén.

Emlékezzünk arra, hogy a $\{\text{Wait} | \text{Wait} > 0\}$ exponenciális eloszlású $(N\mu - \lambda)^{-1}$ középértékkel. Ki lehet következtetni (1), (2) és (6)-ból, hogy a négyzetgyökös biztonsági létszámfeltöltés $\Delta = \beta \sqrt{R}$ esetén az ASA:

7

$$E[\text{Wait}] = P\{\text{Wait} > 0\} \cdot E[\text{Wait} | \text{Wait} > 0] \approx P\{\text{Wait} > 0\} \cdot \frac{E[S]}{\Delta}$$

összefüggést nyerjük. Csakúgy, mint a következő egyszerű kifejezést a késés eloszlására (TSF).

8

$$P\{\text{Wait} > T\} \approx P\{\text{Wait} > 0\} \cdot e^{-(T/E[S])\Delta}$$

Végtelen számú kiszolgálóval rendelkező rendszerek esetén egy adott hívás által foglaltan talált kiszolgálók számát Poisson eloszlással írhatjuk le, és a heurisztika feltételezi, hogy hatalmas véges rendszerek esetén ez közel Poisson, ha a késleltetések nem uralkodóak. Megfordítva a Poisson valószínűségi változó R középértékkel megközelítően egy normál eloszlású valószínűségi változó R középértékkel és $\sqrt{R}$ szórással. Ekkor legyen egy határ valószínűségi érték a késéshez, amit jelöljünk α -val, és válasszunk egy β -t úgy, hogy:

$$\alpha = 1 - \Phi(\beta) \equiv \bar{\Phi}(\beta)$$

Ez bizonyított a következővel:

$$P\{\text{Wait} > 0\} \approx P\{\text{A foglalt kiszolgálók száma} > N\} = P\{R + Z\sqrt{R} > R + \beta\sqrt{R}\} = \bar{\Phi}(\beta)$$

Itt a Z a standard normális eloszlás valószínűségi változóját jelöli, és a PASTA tulajdonság biztosítja, hogy a $P\{\text{Wait} > 0\} \approx P\{\text{A foglalt kiszolgálók száma} > N\}$. Egy kicsiny $P\{\text{Wait} > 0\}$ esetén, $\bar{\Phi}^{-1}(\beta) \approx P^{-1}(\beta)$ és a heurisztika ajánlata alapvetően illeszkedik arra, amit Halfin és Whitt leír.

Borst et al. [17] bizonyítja, hogy a négyzetgyökös elméleten alapuló különböző természetes késési költségfüggvények és a létszámfeltöltés, valójában jól működnek hatalmas, erősen terhelt rendszerekben. Hogy minimalizáljuk a költségeket, optimális, hogy a QED rendszerben dolgozzunk. Ugyanezt a következtetést vonhatjuk le, amikor minimalizáljuk a létszámfeltöltés szintjét, amely a teljesítmény mérés megszorításainak tárgya, ami elterjedtebb a gyakorlatban.

A négyzetgyökös létszámfeltöltés kivételesen pontos és robusztus. Tesztelve lett Borst et al. által minden rendszerben, nagyon gyenge forgalmútól a nagyon erős forgalmúig, és nagyon ritkán tér el többel, mint egyetlen kiszolgáló a pontosan optimális létszámtól.

Meghatároz egy eszközt az optimális β meghatározására, amelyet ők dimenziózásnak (dimensioning) neveznek. Az Ábra 11 (feljebb) bemutatja az optimális β -t, amikor a várakozás költsége és a munkaerő költsége egyaránt lineáris függvénye az időnek. Ebben az

esetben jelölje az r az óránkénti késleltetési költség és a kiszolgáló óránkénti költségének arányát. Ekkor, az optimális β látható, hogy kivételes módon lassan növekszik az r -rel:

$$\beta(r) \approx \begin{cases} \sqrt{r/(1+r(\sqrt{\pi/2}-1))} & 0 \leq r < 10 \\ \sqrt{2\ln(r/\sqrt{\pi/2}) - \ln(2\ln(r/\sqrt{\pi/2}))} & 10 \leq r \leq \infty \end{cases}$$

A Ábra 11 baloldali részén látható, hogy $r=10$ esetén – ami azt jelzi, hogy a késleltetési költség 10-szer nagyobb, mint a munkaerő költsége – az optimális β körülbelül 1,68. Egy olyan Call Centerben, ahol az ajánlott terhelés $R=400$, a számítások alapján a biztonsági létszám 34 fő ($1,68 \cdot \sqrt{400} \approx 33,6$) és a Call Center ekkor 92,2%-os kihasználtsággal ($400/434$) üzemel.

7.2 Működtetési rendszerek, egyesítés (pooling) és méretgazdaságosság

A négyzetgyökös biztonsági létszámfeltöltés elmélete újabb rálátást biztosít a méretgazdaságosság természetéről szempontjából az $M/N/N$ rendszerekben. A gyakorlatban, Borst et al. analízise a három aszimptotikus esetet vizsgálja, és mindegyik különböző méretgazdaságosságot mutat.

Az első esetben az ügyfelek várakozási költségei dominálnak a kapacitás költségeivel szemben, és az optimális létszámpolitika aszimptotikusan fix kihasználtsági rátát alkalmaz. A létszámfeltöltés szintje lineárisan nő az ajánlott terheléssel, és nincs méretgazdaságosság. Egy hatalmas rendszerben a hívók túlnyomó része mindenféle késedelem nélkül kiszolgálásra kerül. Ezt minőségvezérelt rendszernek nevezik.

Egy példa lehet a minőségvezérelt rendszerre az IVR működtetése. Itt a kapacitás relatíve olcsó a kiszolgáló költségéhez képest. Hogy motiválják az ügyfeleket az önkiszolgálás irányába, a vállalatok biztosítják, hogy a kapacitás bőséges legyen, hogy a hívók soha ne észleljenek torlódást.

A másik véglet az, amikor a létszámmal kapcsolatos költségek dominálnak az ügyfelek késleltetésével szemben. Ebben az esetben, az optimálisan felüli többlet létszám aszimptotikusan fix az optimális rendszerben. Ez az, amely szükséges az ajánlott terhelés csúcsainak a kezeléséhez. Emiatt, ha a terhelés növekszik, akkor nagyon hamar eléri a 100%-ot a rendszer kihasználtsága.

Ezt hatékonyságvezérelt rendszernek nevezzük. Példákat találhatunk e-mail csatornán keresztül kommunikáló kontakt centerekben, és olyan help-desk üzemeltetésekben, ahol ingyenesen szolgáltatnak olyanoknak, akik nem régen vásároltak hardvert vagy szoftvert. Ezekben a rendszerekben, alapvetően minden ügyfél késleltetve van a sorban, az ASA majdnem annyi, mint a várt kiszolgálási idő, és a kiszolgálók majdnem 100%-os kihasználtsággal dolgoznak.

Ezek között a végletek között vannak azok a Call Centerek, amelyek a minőség és hatékonyságvezérelt (QED) rendszerek közé esnek, ahol a minőség és a hatékonyság megfelelően egyensúlyozott. Ahogy növekednek, ezek a rendszerek mutatják mind a méretgazdaságosság (hatékonyságvezérelt rendszerek), mind pedig a magas fokú elérhetőség jeleit (minőségvezérelt rendszerek).

Emlékezzünk arra, hogy a QED rendszer fő jellemzője, hogy van az ügyfeleknek egy része, amelyet késleltetünk, de ez a rész sem nem 0-hoz közeli, sem pedig az egyhez. A méretgazdaságosság az, ami miatt a QED rendszer megkerülheti a tradicionális egyezményt a szolgáltatási színvonal és az erőforrások hatékony felhasználása között. Hogy javítsuk ennek a megértését, tekintsük a következő problémát, amely gyakran kerül a Call center menedzserek elé. Ez pedig a földrajzilag szétszórta Call centerek egyesítése. Ezt az egyesítést lehet, hogy fizikailag érik el, azáltal, hogy bezárnak egyeseket, és kiterjesztenek másokat. Vagy virtuálisan, a hálózati technológiák használatával, amit lehetővé teszi a hívások irányítását különböző területek között. Ehhez a problémához összehasonlíthatjuk az egyes rendszereket, milyen a hatásuk a méretgazdaságosságra az egyesítés következtében.

Első lépésként, használjuk a (1) és (2)-öt hogy definiáljuk a következő (7)-(8)-tel analóg összefüggéseket.

$$A\tilde{S}A = E\left[\frac{Wait}{E[S]} \middle| Wait > 0\right] \approx \frac{1}{\Delta},$$

és

$$T\tilde{S}F = P\left\{\frac{Wait}{E[S]} > T \middle| Wait > 0\right\} \approx e^{-T\Delta}$$

Jegyezzük meg, hogy ezek a definíciók megváltoztatják az ASA és a TSF standard verzióit, mégpedig kétféle módon: új feltétel, hogy a késleltetés nem nulla, és a várakozási idő a várható kiszolgálási időtartamban, mint mennyiségi egységben van mérve. Ez teret ad

egyszerű kifejezéseknek, hogy elvégezhessük az egyes rendszerek közötti egyértelmű összehasonlítást.

Észrevesszük, hogy a mindhárom rendszerben az egyik hatékonyságmutató fix, amely ezután meghatározza a többi.

- a). A hatékonyságvezérelt rendszerben a többletkapacitás a Δ fix és emiatt az $\tilde{A\bar{S}A}$ és a $\tilde{T\bar{S}F}$ is.
- b). A minőségvezérelt rendszerben, a rendszerkihasználtság $\rho = R/(R + \Delta)$ konstans
- c). A QED rendszerben a szolgáltatási gradiens a β és $P\{\beta\} \approx P\{\text{Wait} > 0\}$ a fix.

Ezután nézzük meg, mi a hatása m statisztikailag azonos Call center egyesítésének egyetlen közös működtetés alá. Minden Call centerben ugyanaz a λ és a μ . A beérkezési ráta $m\lambda$ lesz, míg a μ nem változik.

A hatékonyságvezérelt létszámfeltöltésnél, a szolgáltatási gradiens csökken β -ról $\beta/\sqrt{m}$ -re, és a késleltetési valószínűség növekszik $P\{\beta\}$ -ről, $P\{\beta/\sqrt{m}\}$ -re (ami jelentős lehet már akár kis m értékek esetén is). Jegyezzük meg, hogy az $\tilde{A\bar{S}A}$ és a $\tilde{T\bar{S}F}$ is ugyanaz marad. Ahogy m tart a végtelenhez, a rendszer gyorsan konvergál ahhoz az állapothoz, ahol minden kiszolgáló 100%-osan kihasznált. Így a rendszer egy darab kiszolgálóhoz hasonlít, amely m -szer gyorsabban dolgozik, és alapvetően minden ügyfél késleltetve lesz.

Egy minőségvezérelt rendszerben jelentős általános szolgáltatási minőségjavulás van. $\tilde{A\bar{S}A}$ csökken $\tilde{A\bar{S}A}/m$ -re, a $\tilde{T\bar{S}F}$ csökken $(\tilde{T\bar{S}F})^m$ -re és a késési valószínűség csökken $P\{\beta\}$ -ről, $P\{\beta/\sqrt{m}\}$ -re. Ahogy m tart a végtelenhez, alapvetően minden ügyfél azonnal kiszolgálásra kerül.

Végül a QED rendszerben, a szolgáltatási gradiens és a várakozás valószínűsége definíció szerint konstans marad. Kontrasztképpen $\tilde{A\bar{S}A}$ csökken $\tilde{A\bar{S}A}/\sqrt{m}$ -re, a $\tilde{T\bar{S}F}$ csökken $(\tilde{T\bar{S}F})^{\sqrt{m}}$ -re. Jegyezzük meg, hogy ez a rendszer hatékonyságvezérelt (a kihasználtság nő 100% felé), és minőségvezérelt is egyszerre (egy jelentős része az ügyfeleknek, nevezetesen $1 - P\{\beta\}$, azonnal kiszolgálásra kerül).

Összefoglalásképpen tekintsük át a következő ábrát.

Méretgazdaságosság

Alap eset: M/M/N rendszer λ, μ és N paraméterekkel

Forgatókönyv: $\lambda \rightarrow m \lambda$ ($R \rightarrow mR$)

	Alap eset	Hatékonyságvez.	Minőségvezérelt	QED
Ajánlott terhelés	$R = \frac{\lambda}{\mu}$	mR	mR	mR
Biztonsági létszám	Δ	Δ	$m\Delta$	$\sqrt{m}\Delta$
Kiszolgálók száma	$N = R + \Delta$	$mR + \Delta$	$mR + m\Delta$	$mR + \sqrt{m}\Delta$
Szolgáltatási gradiens	$\beta = \frac{\Delta}{\sqrt{R}}$	$\frac{\beta}{\sqrt{m}}$	$\beta\sqrt{m}$	$\boxed{\beta}$
Erlang C - $P\{\text{Wait} > 0\}$	$P(\beta)$	$P\left(\frac{\beta}{\sqrt{m}}\right) \uparrow 1$	$P(\beta\sqrt{m}) \downarrow 0$	$\boxed{P(\beta)}$
Kihasználtság	$\rho = \frac{R}{R + \Delta}$	$\frac{R}{R + \frac{\Delta}{m}} \uparrow 1$	$\boxed{\rho = \frac{R}{R + \Delta}}$	$\frac{R}{R + \frac{\Delta}{\sqrt{m}}} \uparrow 1$
$\widetilde{ASA} = E\left[\frac{\text{Wait}}{E(S)} \mid \text{Wait} > 0\right]$	$\frac{1}{\Delta}$	$\boxed{\frac{1}{\Delta} = \widetilde{ASA}}$	$\frac{1}{m\Delta} = \frac{\widetilde{ASA}}{m}$	$\frac{1}{\sqrt{m}\Delta} = \frac{\widetilde{ASA}}{\sqrt{m}}$
$\widetilde{TSF} = P\left\{\frac{\text{Wait}}{E(S)} > T \mid \text{Wait} > 0\right\}$	$e^{-T\Delta}$	$\boxed{e^{-T\Delta} = \widetilde{TSF}}$	$e^{-mT\Delta} = (\widetilde{TSF})^m$	$e^{-\sqrt{m}T\Delta} = (\widetilde{TSF})^{\sqrt{m}}$

Ábra 12 Az Erlang C, a hatékonyság-, minőségvezérelt és QED rendszerekben [18]

8 Erlang B modell (M/G/N/N sorok)

8.1 A blokkolás a gyakorlatban

A Call Centerekben néha adott a technika [10], hogy a menedzserek változtatni tudják az aktív trunk vonalak számát, a k -t. Például a kisebb k érték csúcsidőszakban csökkenti a feladott és a várakozó hívások számát. Ez azzal jár, hogy például zöld számok használata esetén a kommunikációs költség csökkenthető. Hátrány azonban, hogy a foglalt jelzést kapó ügyfelek száma növekszik.

Elméletileg az Erlang C modell tetszőleges számú ügyfelet enged sorba rendezni a modellben. Ez nem csak azért nem valósulhat meg, mert a gyakorlatban az ügyfelek feladják a hívást, de a vonalak száma is limitált, amelyekkel a Call Centerhez lehet csatlakozni. Ezt az úgynevezett blokkolást soha nem lehet figyelmen kívül hagyni. Néha még akkor is érdemes blokkolni ügyfeleket, ha vannak szabad vonalak. Ez növeli a blokkolási rátát, de csökkenti a sorba állított ügyfelek átlagos várakozási idejét. Ezen kívül a produktivitást is csökkenti.

Hogy a vonalak megfelelő számát kiszámítsuk (vagy ekvivalens módon a rendszerben található ügyfelek maximális számát) ki kell számítanunk a produktivitást, blokkolási százalékot és a várakozási időt különböző mennyiségű vonalhoz. Az eredményekből kiválasztható a leginkább kompromisszumos megoldás, amely mind a háromnak megfelelően kedvez. A produktivitás és a várakozási idők kiszámításához szükséges egy feltételezés az ügyfél viselkedését illetően, hogy felépítsük a matematikai modellt. Figyelmen kívül hagyjuk azokat az eseteket, amikor a hívások elvesztek, esetleg ugyanezek újra próbálják hívni a Call Centert vagy visszahívjuk őket, amikor a terhelés engedi. Ezek az esetek mindegyike más modellt követel meg, és más hatékonyságmutató értékekhez vezet.

Egy jól kihasznált rendszerben, átlagban minden $N\mu$ időegység után egy hívás befejeződik. Ha egy hívás az n -edik a sorban, akkor ennek a várakozási ideje az az idő, amely addig telik el, amíg az n -edik kiszolgálás befejeződik, amely átlagban $nN\mu$ időegység. Egy nagy Call Centerben (vagy egy hosszú sorban) ez a szám egész pontos. Használható arra, hogy a vonalak számát meghatározzuk. Például egy hívás befejeződik minden 6. másodpercben. Ha azt szeretnénk, hogy blokkoljuk őket, ahelyett, hogy 1 percnél több ideig várakoztatnánk, akkor 9 hely kell a sorban, hogy elkerüljük, hogy egy 10-ik hívás is bejöjjön, aminek már 1 percig kell várnia.

A fentiekben leírtak alapján tiszta, hogy a blokkolás egy tökéletes mód annak megelőzésére, hogy egy hívás túl sokáig váráksozzon. (Meg kell jegyeznünk, hogy a tovább váráksozó hívók tendenciózusan tovább is beszélnek) Alternatíva lehet az, hogy az ügyfeleket informáljuk a hosszú váráksozási időkről, és megkérjük őket, hívjanak később.

8.2 Az Erlang C kiterjesztése

Az Erlang C modell egy különlegesen egyszerű eszköz a kapacitás és a hozzáférhetőség közötti középút meghatározásának. A nagy forgalom melletti határértékei [11] bepillantást nyújtanak azokba a megegyezésekbe, amelyek segítik a Call centerek méretgazdaságosságának megértését és menedzsmentjét. Azonban eléggé jelentős megszorításai is vannak az Erlang C modellnek.

A gyakorlatban, emlékezhünk, hogy három módon hagyhatja el a rendszert a beérkező hívás. Az a hívás, amely mind a k trunk vonalat foglaltan találja, foglalt jelzést kap. A hívó türelmetlenné is válhat, miközben a sorban váráksozik, és lerakhatja, mielőtt kiszolgálhatnák. A harmadik esetben a hívó kiszolgálóhoz kerül, kiszolgálják, majd távozik. Az Erlang C modell figyelmen kívül hagyja az első két esetet a hároomból. Az Erlang B modell foglalkozik az első esettel, amely ennek a fejezetnek a tárgya.

Egy Call Center képes mindenféle késleltetés megszüntetésére, ha a vonalak számát a kiszolgálók számával egyenlően állítja be. Ebben az esetben az úgynevezett Erlang B formula (B „blocking”-ot azaz „blokkolást” jelent) jellemzi a blokkolás (foglalt jel) valószínűségét a hozzátartozó $M/M/N/N$ rendszerben. Nincsenek sorok, és a hozzáférés lehetősége kizárólag azon ügyfelek arányával van mérve, akik foglalt jelzést kapnak. Ebben az lehet értékes, hogy a blokkolási valószínűség nem érzékeny a szolgáltatási idő eloszlására. Ez magában foglalja inkább az általános, mintsem csak az exponenciális eloszlásokat (ezért inkább $M/G/N/N$, mint $M/M/N/N$).

A QED rendszerben, az Erlang B ugyanolyan négyzetgyökös eredményeket ad, mint az Erlang C rendszerben. Hatalmas $M/G/N/N$ rendszerben az $N = R + \beta\sqrt{R}, -\infty < \beta < \infty$ -el a blokkolási valószínűség a következő: $1/\sqrt{N} : \sqrt{N} \cdot P\{\text{minden vonal foglalt}\} \rightarrow \phi(\beta)/\Phi(\beta)$. Ezért még sorba állítás képessége nélkül is a hozzáférés esélye magas marad a QED rendszerben.

$\beta > 0$ esetén, kicsi a hívók azon része, amely foglalt jelzést kap az Erlang B rendszerben. Továbbá (5) megmutatja, hogy ugyanazon feltételek mellett ($\beta > 0$), a töredék elég kicsi

ahhoz, hogy nem terhelné túl a rendszert, ha engedélyoznánk a sorokat. Természetesen az Erlang C rendszerben vázolt végtelen számú hely a sorban (a trunk vonalak száma) a gyakorlatban nem elérhető.

Az Erlang B és C rendszerek közötti lehetséges középutak egyike lehet a blokkolás késleltetéssel. Az első csökken a sorban rendelkezésre álló helyek számának csökkenésével párhuzamosan, míg a másik növekszik. Mekkora számú hely kell a sorban? Feinberg [19] szimulálja ezt az egyik tanulmányában a $M/N/N/k$ rendszerekről, ahol $k \geq N$ szisztematikusan váltakozik. Kimutatja, hogy csak 10%-kal kell több vonal, mint kiszolgáló, ahhoz, hogy a rendszer jó teljesítménnyel működjön. Több vonal túl sok várakozást okoz, kevesebb túl sok foglalt jelzést.

8.3 Az Erlang B formula

Az Erlang B (vagy „loss” azaz „veszteség”) formula megadja a blokkolási valószínűségét az Erlang B modellben, máshogyan az $M/M/N/0$ modellben. Ez a modell N egyforma kiszolgálót tartalmaz, amelyek párhuzamosan dolgoznak, és nincs extra hely a várakozásra. Azok az ügyfelek, amelyek akkor érkeznek, amikor mind az N kiszolgáló foglalt, foglalt jelet kapnak, elvesznek, anélkül, hogy befolyásolhatnák a jövőbeni érkezéseket, például, nincsenek újrapróbálkozások. Ez a modell Poisson érkezési folyamat, és IID (independent and identically distributed – független és azonos eloszlású) exponenciális eloszlással véges középértékkel. A következőekben leírt formuláról és a annak az Erlang C formulával való kapcsolatáról itt olvashat: [12].

A konvenciók követésével, legyen a beérkezési ráta jelölése λ , és a kiszolgálási idő középértékének jele az $1/\mu$. Emiatt, az egyedi kiszolgálási ráta μ . Mivel N számú lehet a legtöbb ügyfél, ami egyszerre a rendszerben tartózkodik, a sztochasztikus folyamat, amely a kiszolgálók számát az idő függvényében reprezentálja, rendelkezik egy megfelelő stacionárius eloszlással minden pozitív λ és μ paraméterre. Az Erlang veszteség modellnek van egy érzéketlenségi tulajdonsága (insensitivity), amely azt jelenti, hogy a blokkolási valószínűség független a szolgáltatás-idő eloszlástól annak középértéke körül. Emiatt a blokkolási valószínűség ugyanaz, az $M/G/N/0$ modellben, mint az $M/M/N/0$ modellben, ami biztosítja, hogy a szolgáltatás-idő eloszlásnak véges középértéke van.

A foglalt kiszolgálók számának a stacionárius eloszlású értéke nem függ az idő mérésének mértékegységétől. Ezért a blokkolási valószínűség csak a beérkezési és a kiszolgálási rátától függ, és csakis az arányukon keresztül, ami $R \equiv \lambda / \mu$ az ajánlott terhelés.

Közeli viszonyban van ezzel a kiszolgálónkénti ajánlott terhelés, amelyet kiszolgáló kihasználtságnak vagy forgalom intenzitásnak nevezünk. $\rho \equiv \lambda / N\mu = R / N$

Mind a ajánlott terhelés, mind a forgalom intenzitás (ρ) dimenzió nélküli mennyiségek.

Jelölje az s az egyensúlyi foglalt kiszolgálószámot egy tetszés szerinti időben. Az s eloszlását az alapvető születés és halál folyamatból nyerjük. Kiderül, hogy az s egy csonkított Poisson eloszlással rendelkezik.

$$P\{s = j\} = \frac{R^j / j!}{\sum_{k=0}^{k=N} R^k / k!}, \quad 0 \leq j \leq N;$$

Ahogy fentebb jeleztük, az Erlang B formula megadja egy tipikus beérkezés stacionárius blokkolási valószínűségét. Az Erlang B formulát a PASTA tulajdonság alkalmazásával a $P\{s = N\}$ valószínűségből kapjuk. Az Erlang B formula tehát a következő összefüggés:

$$B \equiv B(N, R) = P\{s = N\} = \frac{R^N / N!}{\sum_{k=0}^{k=N} R^k / k!}.$$

Egy fontos kapcsolódó mennyiség az összefüggés reciproka. Fontosságát egyszerűbb analízálhatósága adja: $Re \equiv Re(N, R) \equiv \frac{1}{B(N, R)}.$

8.4 Gyakorlatok az Erlang formulára

a). Rekurzió az R -re

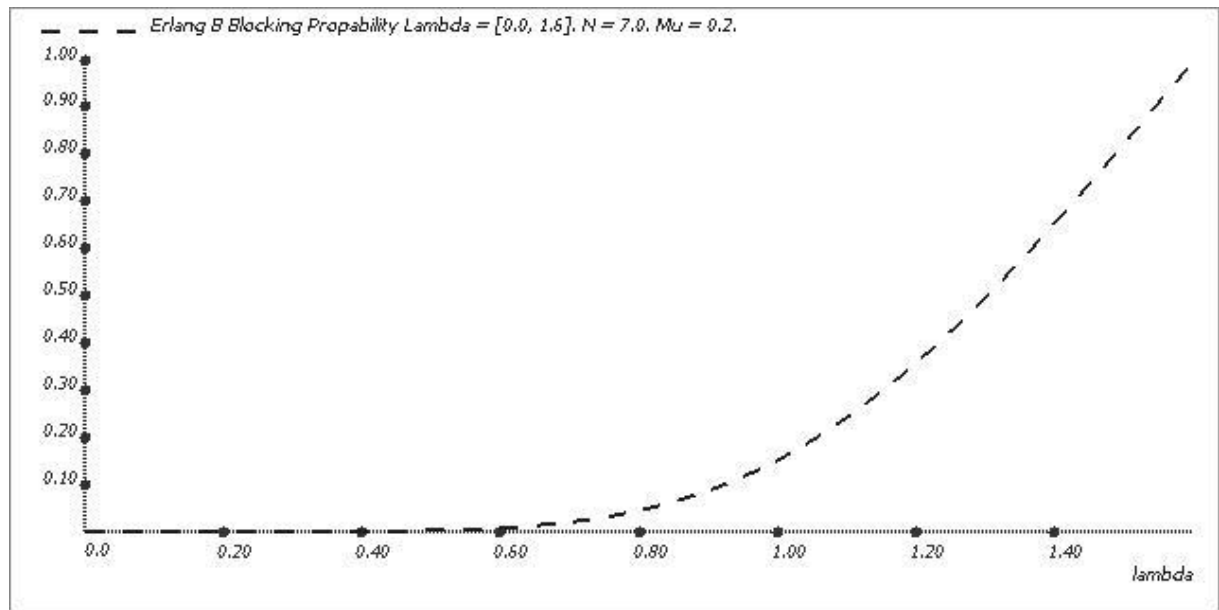
$$Re(N, R) = \frac{N Re(N-1, R)}{R} + 1, \quad N \geq 1, \text{ ahol } Re(0, R) \equiv 1.$$

b). Ebből következik a rekurzió a B -re, ami a $B(N, R) = \frac{RB(N-1, R)}{N + RB(N-1, R)}$ összefüggéssel

írható le. Itt $B(0, R) \equiv 1$ és az ajánlott terhelés az R , amelyet az értelmezés szerint rögzítünk.

A jobboldal átalakításával a $\rho = R / N$ felhasználásával a következő

$$B(N, R) = \frac{\rho B(N-1, R)}{1 + \rho B(N-1, R)} \text{ összefüggéshez jutunk.}$$



Ábra 13 Erlang B Blokkolási valószínűség alakulása a beérkezési intenzitás függvényében

c). Számolás. Miért vonzó a rekurzió a blokkolási valószínűség számolásakor nagyon nagy N -nél az explicit formulával szemben? Mert a nagyon nagy hatványok és faktoriálok destabilizálják az algoritmust.

d). A b).-ben leírt formula segítségével megmutatható, hogy a $B(N, R)$ monoton növekszik R -ben.

e). Alsó korlát. A Little törvénye ($L = \lambda W$) segítségével megállapítható, hogy $B(N, R)$ alulról korlátos, azaz $B(N, R) \geq \max\{0, 1 - \rho^{-1}\}$, ahol $\rho \equiv R / N$.

f). A b).-ben és e).-ben leírtak alapján megmutatható, hogy a $B(N, R)$ monoton csökken N -ban, ahogy N felveszi a pozitív egész számokat.

g). Határérték analízis. A terhelés, amelyet az utolsó kiszolgáló vesz fel az $F_B(N, R) \equiv R[B(N-1, R) - B(N, R)]$ összefüggéssel írható le. Látható az f). bizonyítása alapján, hogy az F_B csökken az N -ben. Feltételezzük, hogy van egy költsége minden kiszolgálónak, amely c egység minden egyes perc után, és egy bevételi ráta, amely r egység a felvett terhelés minden egysége után. Hogyan használhatjuk a legutolsó kiszolgáló által felvett terhelést, az F_B -t arra, hogy meghatározzuk a kiszolgálók optimális számát, amely maximalizálja a profitot (ami a bevétel - költség)? Az optimális N megtalálásához elég, ha meghatározzuk a legnagyobb N -t, amelyre az $F_B(N, R) \geq c/r$. Legyen az N^* ez az N érték. Ekkor az utolsó kiszolgáló által hozzáadott bevétel nagyobb, mint a költség, de ez a tulajdonság már nem érvényes a $N^* + 1$ -re.

h). Hatalmas rendszerek. Tekintsük a B -t, mint a ρ függvényét, úgy hogy $B(N, N\rho)$ (ahol $R = N\rho$) fix. Mi lehet a $B(N, 3N)$ értéke megközelítőleg, ha az N nagyon nagy? Ábrázoljuk, hogyan néz ki a $B(N, N\rho)$, mint a ρ függvénye nagy N esetén ($N \rightarrow \infty$)?

Ha a $\rho \leq 1$, $B(N, N\rho) \rightarrow 0$ amikor $N \rightarrow \infty$. Ha a $\rho > 1$, a beérkező hívások $(N\rho - N)/s\rho$ összefüggéssel megadott hányadát blokkolni kell. Ebből következik, hogy a $B(N, 3N)$ körülbelül $2/3$ nagy N esetén.

9 Erlang A Modell (M/M/N+M sor)

Az $M/M/N/k+G$ az a modell, amely magában foglalja a foglalt jelzéseket és a hívásfeladást is. Ebben a modellben a türelem úgy van definiálva, hogy az a maximális mennyiségű idő, amelyet az ügyfél hajlandó várni a kiszolgálásra. Ha nem kerül kiszolgálásra ennyi idő alatt, akkor leteszi a kagylót. A „+G” jelzés azt jelenti, hogy a türelem általános eloszlású, az független, azonos eloszlású (IID) ügyfelektől függ, de semmi mástól. Ez a fejezet elsősorban a következő cikkekre épül: [13] és [11].

Az uralkodó gyakorlat az, hogy bőséges számú vonalat installálnak annyit, hogy a foglalt jelzés nagyon ritka esemény lesz. Ebben az esetben egy $M/M/N + M$ (Ekvivalens módon írva $M/M/N/\infty + M$) amire úgy hivatkozunk, hogy Erlang A. Az „A” az angol „Abandonment” szót jelöli, annak a ténynek köszönhetően, hogy az Erlang A interpolál az Erlang B és Erlang C között.

A nagy forgalmú rendszerekkel analóg módon három működtetési rendszer létezik az Erlang A modellhez. A hatékonyságvezérelt, a minőségvezérelt, és a QED. Ahogyan ezelőtt, a működtetési rendszerek a késleltetési valószínűség szerint vannak jellemezve. Az első esetben ez közel van egyhez, a második a nullához, végül az utolsó valahol a $(0,1)$ intervallumon értelmezett. És ugyanúgy, ahogy eddig, a QED rendszer $N \approx R + \beta\sqrt{R}$ kiszolgálóval, elég robusztus, hogy lefedje az egész működtetési spektrumot. Itt, azonban a szolgáltatási gradiens felvehet negatív és pozitív értékeket egyaránt, mivel a hívásfeladás stabilizálja a rendszert minden létszámfeltöltési szintnél. A QED rendszer működtetési jellemzői eléggé megnyerőek, és összefoglalhatóak a következőképpen. A kiszolgáló tétlensége, az ASA, és legfontosabbként a hívást feladók aránya mind-mind $1/\sqrt{N}$ rendűek.

6/13/00 - Tue

Charlotte - Center										
Time	Recvd	Answ	Abn %	ASA	AHT	Occ %	On Prod%	On Prod FTE	Sch Open FTE	Sch Avail %
0	20,577	19,860	3.5%	30	307	95.1%	85.4%	222.7	234.6	95.0%
8:00	332	308	7.2%	27	302	87.1%	79.5%	59.3	66.9	88.5%
8:30	653	615	5.8%	58	293	96.1%	81.1%	104.1	111.7	93.2%
9:00	886	796	8.1%	63	308	97.1%	84.7%	140.4	145.3	96.6%
9:30	1,152	1,138	1.2%	28	303	90.8%	81.6%	211.1	221.3	95.4%
10:00	1,330	1,286	3.3%	22	307	98.4%	84.3%	223.1	229.0	97.4%
10:30	1,364	1,338	1.9%	33	296	99.0%	84.1%	222.5	227.9	97.6%
11:00	1,380	1,280	7.2%	34	306	98.2%	84.0%	222.0	223.9	99.2%
11:30	1,272	1,247	2.0%	44	298	94.6%	82.8%	218.0	233.2	93.5%
12:00	1,179	1,177	0.2%	1	306	91.6%	88.6%	218.3	222.5	98.1%
12:30	1,174	1,160	1.2%	10	302	95.5%	93.6%	203.8	209.8	97.1%
13:00	1,018	999	1.9%	9	314	95.4%	91.2%	182.9	187.0	97.8%
13:30	1,061	961	9.4%	67	306	100.0%	88.9%	163.4	182.5	89.5%
14:00	1,173	1,082	7.8%	78	313	99.5%	85.7%	188.9	213.0	88.7%
14:30	1,212	1,179	2.7%	23	304	96.6%	86.0%	206.1	220.9	93.3%
15:00	1,137	1,122	1.3%	15	320	96.9%	83.5%	205.8	222.1	92.7%
15:30	1,169	1,137	2.7%	17	311	97.1%	84.6%	202.2	207.0	97.7%
16:00	1,107	1,059	4.3%	46	315	99.2%	79.4%	187.1	192.9	97.0%
16:30	914	892	2.4%	22	307	95.2%	81.8%	160.0	172.3	92.8%
17:00	616	615	0.0%	2	328	83.0%	93.6%	135.0	146.2	92.3%
17:30	420	420	0.0%	0	328	73.8%	95.4%	103.5	116.1	89.2%
18:00	49	49	0.0%	14	180	84.2%	59.1%	5.8	1.4	416.2%

Ábra 14 Félórás lebontású összefoglaló jelentés az ACD-ből (Példa, a Wharton Call Center fórum jóvoltából).

Tehát az Erlang A modell matematikai módon demonstrálja, hogy a magas szinten terhelt Call centerek teljesítményét jelentősen befolyásolja az ügyfelek hívásfeladása. A fenti ábrán egy jelentés látható egy létező Call Centerről. Egy-egy sorban a vizsgált időszak ajánlott terhelése (Recvd), a megválaszolt hívások száma (Answ) és az AHT² össze van rendelve az ASA-val és a hívásfeladási aránnyal (Abn%). Például látható az is, hogy 10 és 11:30 között a dolgozó kiszolgálók száma nem változik jelentősen.

Ahhoz, hogy lássuk, hogyan hat a hívásfeladás a teljesítményre, az Erlang A modellt önálló, egymástól független félórás szakaszokban vizsgáljuk. A táblázatból használjuk a beérkező hívások számát és az AHT-t mint a λ és a μ becslését. Felkerekítjük a „On Production FTE³” számait, hogy hozzájussunk az N közelítéséhez. Három paraméter adott a négyből az $M/M/N + M$ modellben. Azt a hívásfeladási rátát keressük, ami a vizsgált fél órában észlelt ASA-hoz és hívásfeladási százalékhöz közelít.

Például a 10:30 és a 11 óra közötti periódusban, az előbbi eljárás produkálja a hívásfeladásig eltelt idő becsült középértékét. Adottnak tekintve ezt a becslést,

² AHT – Average handling time – Átlagos híváskezelési idő

³ FTE – Full time equivalent – Az ütemezett kiszolgálók száma osztva teljes munkaidő hosszával

megállapítható, hogy a 223-ból 5 kiszolgáló hiánya is eredményezheti az ASA, valamint hívásfeladó hányad kétszeresre növekedését.

Érdekes és lényeges megfigyelni, hogy egy olyan modell, ahol az átlagos türelem 30 perc drámaian különbözik attól a modelltől, amely nem figyeli a hívásfeladást („végtelen a türelem”). A 10:30-tól 11-ig tartó időszakban az utóbbi egy nem stabil rendszert jósol, amelyben a kiszolgálók idejük több mint 100%-ában foglaltak. Megjegyzendő, hogy a stabilitás már két további kiszolgáló hozzáadásával elérhető, de ebben az esetben az ASA közel lenne a 7 perchez. Ha a hívásfeladást figyelmen kívül hagynák, az több lenne, mint közelítési hiba a jósolt teljesítményben.

Nagy forgalomnál, még a kisszámú hívásfeladási (vagy blokkolási) arány is jelentős hatással lehet a rendszer teljesítményére, emiatt számolni kell vele a minimális létszámszint meghatározásakor. Ezért, ajánlatos az Erlang A használata standardként gyakran használt Erlang C helyett.

Gyakori panasz a Call Center menedzserek között, hogy a munkaerő menedzsment szoftverek mindig többletlétszám ütemezését erőltetik. Míg néhányuk intuitív érzéküket használják arra, hogy hogyan nyomják le a létszámszintet, sokkal jobb megközelítés az, ha a hívásfeladásokat modellezzük elsőször.

9.1 Türelmetlen ügyfelek

A szolgáltató központokban általában, ha a szolgáltatás telefonon keresztül történik, az ügyfelek hajlamosak a türelmetlenségre, és általában néhányan, akik a sorban állnak úgy döntenek, hogy leteszik a telefont, mielőtt a szolgáltatás elkezdődhetne. Az a modell, ami ezzel számol emiatt a Call Center pontosabb leírását szolgáltatja. Ha hozzáadjuk a modellünkhöz az ügyfelek távozását a sorból, mint tényezőt, akkor csökkentjük a torlódást, hiszen nem minden beérkező hívás igényelt szolgáltatást. Következésképpen, a sorok hossza és a várakozási idő lecsökken. Ezért ha csak az $M/M/N/B$ (Erlang B késleltetéssel) modellt használjuk, az többletlétszámot eredményez.

A hívásfeladás figyelembe vételével az átlagos várakozási időben benne van a kiszolgált és a hívást feladó ügyfelek ideje is. Azonban a „hűségese” ügyfelek átlagos várakozási ideje külön is fontos, ami ezáltal egy külön hatékonyságmutató ebben a modellben.

Az $M/M/N$ modell exponenciális hívásfeladásokkal való kiegészítésének az is egy fontos előnye, hogy az eredmény robusztusabb, mindig van egy egyensúlyi állapota, nem számít

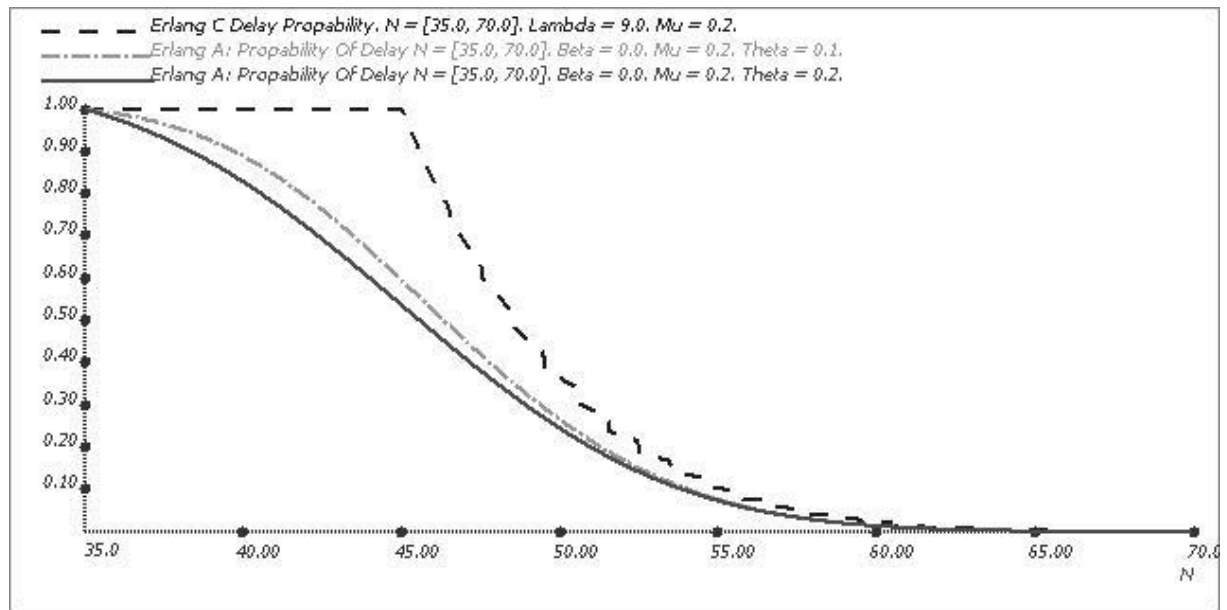
hány kiszolgálónk (N) van, és mik a beérkezési (λ) és kiszolgálási (μ) rátáink. Ez szembeállítható azzal, hogy a normál $M/M/N$ modell a $\lambda < N\mu$ feltételt kívánja meg a stabilitáshoz. A robusztusság kritikus lehet akkor, ha a rendszereket erőteljes forgalom alatt vizsgáljuk, amely akár csak átmenetileg is túlterhelt lesz ($\lambda \geq N\mu$).

Talán kevésbé nyilvánvaló hátránya a $M/M/N/B$ modellnek az, hogy az bizonyos lényeges információkat egyáltalán nem szolgáltat a Call Center menedzsereinek. Amikor egy Call Centert erőteljes forgalom alatt kell menedzselni, tekintetbe kell venni a hívásfeladó ügyfelek hatását is a szolgáltatási színvonalra. Nem elég a várakozási időket és azokat az ügyfeleket tekinteni, akik foglalt jelzést kaptak, mivel a hívásfeladási statisztika az egyetlen olyan ACD⁴ adat, amely leleplezi az ügyfelek észlelését a szolgáltatás minőségéről. A szolgáltatási színvonal háromdimenziós, három külön aspektusa van, a várakozás, a blokkolás, és a hívásfeladás. Az $M/M/N/B$ modell ezek közül csak kettőt szolgáltat. A gyakorlatban a Call Centerek nagy százaléka kitűzi célul a hívásfeladók arányának egy alacsony felső korlátját, de a legtöbb esetben felülmúlják azt.

Az elosztási határok (diffusion limits) fontos eszközt képeznek a sorbanállási modellek elemzésekor. A fontossága háromszoros: az elosztási folyamatokat használják az közelítések származtatásához, a szükséges rálátást biztosítja, és gyakorlati szabályok kidolgozását teszi lehetővé.

Ennek a dokumentumnak a fő eredménye az 5. tétel, amely a $M/M/N + M$ sorok sorozataival foglalkozik erős forgalomnál. Az elméleti megfontolásokon túl, a gyakorlatban is jelentősek ezek az eredmények, mert úgy paraméterezik a rendszereket, hogy azok a leginkább hasonlítsanak a hatalmas Call Centerekre. Ezek az eredmények biztosítják az sorba állított ügyfelek arányát közelítő eljárások tényesítését.

⁴ ACD – Automatic call distributor. – Automatikus híváselosztó. A hívások sorba rendezését, a sorok kezelését végzi. Számunkra igen fontos az ez által mért adat, és az abból generált jelentés, amely inputként szolgál modelljeinknek, vagy épp ellenkezőleg modelljeink paramétereit állítjuk be úgy, hogy ezekhez az adatokhoz vagy jelentésekhez közelítsenek.



Ábra 15 A késleltetés valószínűségének alakulása az Erlang C és az Erlang A modellekben

Az $M/M/N$ és az $M/M/N + M$ modellek összehasonlítása látható az előző ábrán is.

Vegyük észre, hogy minden $N \leq \frac{\lambda}{\mu}$ esetén (ahol N , λ és μ egyenként a kiszolgálók számát, a beérkezési rátát, valamint a kiszolgálási rátát jelöli) a közelítés abban a modellben, ahol nincs használva a hívásfeladás, 1, azaz mindenki, aki a rendszerbe érkezik sorba lesz állítva. Ez azért van, mert ez egy túlterhelt rendszer, ahol nincsen egyensúlyi állapot. Az N nagy értékeire, a rendszerekben többletkapacitás keletkezik. Ezekben az esetekben a közelítések egybeesnek az elhanyagolható számú hívásfeladások miatt. A sorba állított ügyfelek számának kiszámítására szolgáló kifejezés a C metódusból és a 9.3-as fejezet első megjegyzéséből származtatható. A következő összefüggés adja ezt meg:

$$1 - \frac{1 - P\{Bl\}}{1 + P\{Bl\} \left[\frac{50 \times 1}{0,5} \left(\frac{0,5}{48} \right)^{50 \times 1/0,5} e^{48/0,5} \gamma \left(\frac{50 \times 1}{0,5}, \frac{48}{0,5} \right) - 1 \right]}$$

ahol a $P\{Bl\}$ a blokkolt ügyfelek száma az $M/M/N/N$ modellben, ahol N darab trunk vonal van.

Az erős forgalomnál számolt eredmények egy újabb gyakorlati szabályhoz vezetnek. A következő, egy olyan lehetséges forgatókönyv, amelyben az felhasználható. Egy Call Center ahol N kiszolgáló van, a kiszolgálási ráta a μ , a beérkezési ráta λ , és van egy szolgáltatási gradiense is, amit ζ -val jelölünk (A ζ nagy értékei magas szolgáltatási színvonalhoz kapcsolódnak). A következő ünnepekkor magasabb érkezési arányt jósolnak amit $\hat{\lambda}$ -pal

jelöljük. A Call Center menedzsere, a jelenlegi szolgáltatási színvonalat akarja biztosítani akkor is, emiatt döntést kell hoznia a kiszolgálók számáról, ami $N = \hat{N}(\zeta)$ az ünnepi műszakokban. Ahogy azt majd később látni fogjuk, három működtetési rendszert javasolhatunk a Call Centernek, amit a szolgáltatási minőség és/vagy a működtetési hatékonyság fog meghatározni. Ezek a rendszerek a következők: minőségvezérelt, amelyet ritka várakozás és hívásfeladás jellemez; hatékonyságvezérelt, amely a kiszolgálók hatékonyságára helyezi a hangsúlyt, és az ügyfelek többsége várakozik, jelentős részüket feladja a hívást; és a racionális, amelynél a minőség és a hatékonyság egyensúlyban van, továbbra is jól kihasználja az erőforrásokat, de csak az ügyfelek egy kontrollált töredékének kell várakoznia és csak néhányának kell feladni a hívását.

Amikor a menedzser eldönti, melyik legyen az a működtetési rendszer a Call Centerben, amely legjobban reprezentálja a kívánt egyensúlyt a minőség és a hatékonyság között. A szabályaink előállítják a $\hat{N}(\zeta) = \left\lceil \frac{\hat{\lambda}}{\mu} + \zeta \sqrt{\frac{\hat{\lambda}}{\mu}} \right\rceil$ összefüggést, ahol $\zeta = \sqrt{\frac{\mu}{\lambda} \left(N - \frac{\lambda}{\mu} \right)}$ (amely a többletkapacitás és a terhelés arányának a négyzetgyöke). Továbbá, a késleltetett és hívást feladó ügyfelek arányának becsült értékei a következők:

$$P\{\text{Wait} > 0\} \approx \left[1 + \frac{h(\zeta \sqrt{\mu/\theta})}{\sqrt{\mu/\theta} h(-\zeta)} \right]^{-1}$$

$$P\{\text{Abandon}\} \approx \frac{1}{\sqrt{\hat{N}}} \left[\sqrt{\theta/\mu} \cdot h(\zeta \sqrt{\mu/\theta}) - \zeta \right] \cdot \left[1 + \frac{h(\zeta \sqrt{\mu/\theta})}{\sqrt{\mu/\theta} h(-\zeta)} \right]^{-1}$$

Itt a $h(x) = \phi(x)/[1 - \Phi(x)]$ amely a standard normális eloszlás rizikó rátája, amelyek definíciója $\phi(x) = \frac{1}{\sqrt{2\pi}} e^{-x^2/2}$, és $\Phi(x) = \int_{-\infty}^x \phi(y) dy$.

Jegyezzük meg még a következőt is az előbbiekkal kapcsolatban. Ha növeljük a kiszolgálók számát, az N -t és a késleltetett ügyfelek arányát ugyanolyan szinten tartjuk, a kiszolgálók kihasználtsága fokozódik, de a hívást feladó ügyfelek száma az átlagos várakozási idővel együtt csökken.

9.2 Formalizálás és jelölések

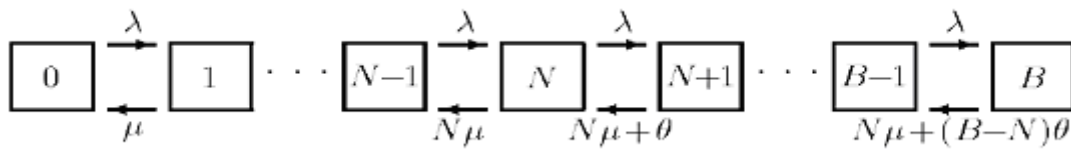
Kényelmi szempontokból nézzük át először a $M/M/N/B+M$ modell feltételezéseit és folyamatait. A rendszernek egyetlen sora van, amely N független és statisztikai szempontból

azonos kiszolgálót lát el. Az ügyfelek egy λ rátával Poisson folyamat szerint érkeznek, és az érkezésük sorrendjében (FCSF) lesznek kiszolgálva. A szolgáltatás idők (μ) exponenciális valószínűségi változók. Minden ügyfél türelme (az az időperiódus, amelyet a sorban akar tölteni, különben elhagyja azt) egy θ exponenciális valószínűségi változó, minden másától független. A rendszer kapacitása B számú ügyfél (ez alapján legtöbb $K = B - N$ lehet a sorban). Azok, akik a kapacitáson felül érkeznek, el lesznek utasítva a rendszer használatától.

A rendszer állapota egy t időpillanatban a rendszerben levő ügyfelek számával definiált (éppen kiszolgálás alatt vannak, vagy várakoznak a sorban) és $Q(t)$ -vel van jelölve. Mivel a szolgáltatási idők, az ügyfelek türelme, a beékezési idők mind független exponenciális eloszlású változók, $Q = \{Q(t), t \geq 0\}$ egy születési-halálozási folyamat, a következő születési és halálozási rátákkal:

$$\lambda_k = \begin{cases} \lambda, & 0 \leq k \leq B-1 \\ 0, & \text{egyébként} \end{cases}; \quad \mu_k = \begin{cases} (N \wedge k)\mu + [k - N]^+ \theta, & 1 \leq k \leq B \\ 0 & \text{egyébként} \end{cases}$$

A folyamat átmenet diagramja a következő ábrán látható.



Ábra 16 $\{Q(t), t \geq 0\}$ – Átmeneti diagram

Megjegyzés: A Q folyamatnak mindig van egyensúlyi állapota, a legjelentősebb az, amikor $B = \infty$ ahol a folyamatnak végtelen állapottere van. Ez az ellentéte az $M/M/N$ modellnek, amelyekben az egyensúlyi állapot hatékonyságmutatói nem kalkulálhatóak túlterhelt rendszer esetén (ahol $\lambda \geq N\mu$).

Az ügyfelek általános F eloszlású türelme esetén, és egy olyan rendszerben, ahol végtelen kapacitás van, az egyensúlyi állapot létezésének a kritériuma:

$\lambda(1 - F(\infty)) < N\mu$, mivel a kiszolgálóknak le kell küzdeniük azt a forgalmat, amelyet a végtelen türelmű ügyfelek okoznak.

Megjegyzés: A modellünk feltételezi, hogy az ügyfél türelme nem függ a sorban elfoglalt helyétől. Ez a feltételezés nem ok nélküli a telefonon keresztüli szolgáltatások esetén, mivel a sor nem látható. Az ügyfeleknek általában nincs információjuk a sorról.

További jelölések: A következőekben használni fogjuk a gamma függvényt, amit az alábbi összefüggés definiál:

$$\Gamma(x) = \int_0^{\infty} t^{x-1} \exp(-t) dt$$

Valamint a nem teljes gamma függvényt, amit a $\gamma(x, y)$ jelöl:

$$\gamma(x, y) = \int_0^y t^{x-1} \exp(-t) dt.$$

A gyenge konvergencia a standard jelölés, amely formalizálja a valószínűségi eloszlások közelítéseit. Ez a stacionárius eloszlások és a sztochasztikus folyamatok konvergenciája esetén jön elő. Minkét esetben, a következőképpen jelöljük az $\{X_N\}$ sorozat gyenge konvergenciáját X -hez: $X_N \xrightarrow{d} X$, ahol a d azt jelenti, hogy eloszlásban konvergál (convergence in distribution).

9.3 Az M/M/N/B+M Modell pontos számítása

Ebben a fejezetben bevezetésre kerül több hatékonyságmutató számítási módszere egy M/M/N/B+M modell egyensúlyi állapotához. Az alatta meghúzó szünetelési-halálozási folyamat miatt, ezek a kalkulációk majdnem triviálisak (nem teljesen a numerikus problémák következtében). Ezek az eredmények később fel lesznek használva a közelítések származtatásakor.

A hatékonyságmutatók kalkulációjakor, alapul vesszük a feltételezést, hogy a rendszer elérte az egyensúlyi állapotát. Habár a beérkezési ráta sok Call Centerben időben változó (időpontról-időpontra változhat egy napon belül, vagy napról-napra változhat egy héten belül, ünnepek, szezonális hatások stb.), más paraméterek, mint a kiszolgálók száma egy műszakban is változtatható. Feltételezzük, hogy egy rövid időintervallumon belül (például egy óra) ezek a változások elég kicsik ahhoz, hogy figyelmen kívül hagyjuk őket, és „lassúak” ahhoz képest, amilyen gyorsan a rendszer a következő egyensúlyi állapotba kerül.

Minket a tipikus ügyfél érdekel, aki az egyensúlyi állapotban lévő rendszerbe érkezik. Legyen a V valószínűségi változó a tipikus ügyfél potenciális várakozási ideje (az az idő, amelyet a sorban kellene várakoznia, hogy a kiszolgálás elkezdődjön, ha a türelme végtelen volna). Legyen X ennek az ügyfélnek a türelme (jegyezzük meg, hogy X és V függetlenek) és legyen a W az aktuális várakozási ideje. Tisztán látható, hogy a $W = V \wedge X$.

Végül legyen a $\{Bl\}$ az az esemény, hogy az ügyfél foglalt jelzést kapott, és $\{Ab\}$ esemény jelentse azt, hogy az ügyfél feladta a hívást ($\{Ab\} = \{V > X\}$).

$f(v, x)$	$E[f(V, X)]$
$1_{\{v>x\}}$	$P\{Ab\}$
$1_{\{t,\infty\}}(v \wedge x)$	$P\{W > t\}$
$1_{\{t,\infty\}}(v \wedge x)1_{\{v>x\}}$	$P\{W > t; Ab\}$
$(v \wedge x)1_{\{v>x\}}$	$E[W; Ab]$
$(v \wedge x)1_{\{t,\infty\}}(v \wedge x)1_{\{v>x\}}$	$E[W; W > t; Ab]$
$g(v \wedge x)$	$E[g(W)]$

Ábra 17 Hatékonyságmutatók $E[f(V, X)]$ alakban

Mit is jelent a tipikus ügyfél? Vegyük tekintetbe a $\{w_n, n \in \mathbb{N}\}$ ahol a w_n az n -edik ügyfél potenciális várakozási ideje. Legyen az F_w a stacionárius eloszlása ennek a sorozatnak. F_w a stacionárius eloszlása a $v(t)$ folyamatnak is, amely a virtuális várakozási idő a t időpontban (az az idő, amit várakozással tölt el egy hipotetikusán végtelen türelmű ügyfél a sorban, aki a t időpontban érkezett). Ezért egy tipikus ügyfél potenciális várakozási ideje, a V , szintén rendelkezik egy F_w eloszlásfüggvénnyel. Hasonlóan érdekes a V_n , amely olyan valószínűségi változó, amelynek az eloszlása ugyanaz, mint a V -nek, amikor n darab ügyfél van a sorban érkezéskor, és minden kiszolgáló foglalt, $n = 0, 1, \dots$ V_n eloszlásfüggvénye F_n .

A Call Center menedzsereket érdeklő sok hatékonyságmutató kifejezhető, mint V és az X egyszerű függvényeinek várható értékei. A fenti tábla egy reprezentatív listát közöl belőlük.

Megjegyzés: 1. Ebben a fejezetben a π -t használjuk a $Q(t)$ folyamat stacionárius eloszlásának jelölésére, nevezetesen $\lim_{t \rightarrow \infty} P\{Q(t) = n\} = \pi_n, n = 0, 1, 2, \dots, B$.

Egy általános kifejezés ezekre a valószínűségekre:

$$\pi_k = \begin{cases} \frac{(\lambda/\mu)^k}{k!} \pi_0, & 0 \leq k \leq N \\ \prod_{j=N+1}^k \left(\frac{\lambda}{N\mu + (j-N)\theta} \right) \frac{(\lambda/\mu)^N}{k!} \pi_0 & N < k \leq B \end{cases}$$

ahol

$$\pi_0 = \left[\sum_{k=0}^N \frac{(\lambda/\mu)^k}{k!} + \sum_{k=N+1}^B \prod_{j=N+1}^k \left(\frac{\lambda}{N\mu + (j-N)\theta} \right) \frac{(\lambda/\mu)^N}{N!} \right]^{-1}.$$

2. A V eloszlása előzetesen nem ismert, hanem a modell analizálásán keresztül származtatható. Másfelől, V_n kifejezhető $n+1$ független valószínűségi változó összegeként, $N\mu, N\mu + \theta, \dots, N\mu + n\theta$ paraméterekkel. Az i -edik ezek közül azt az időt jelenti, amelyet az ügyfél a sor i -edik helyén töltött el, mielőtt az $i-1$ -edik helyre léphetett volna (előtte a kiszolgálás befejeződött, vagy valaki feladta a hívást).

3. Egy foglalt jelzést kapó ügyfél számára (mert a sor tele volt érkezőkkel) a $V=0$ konvenció kerül bevezetésre.

4. Bizonyos fontos hatékonyságmutatók nem számolhatók közvetlenül a javasolt módszerrel, de a $E[f(V, X)]$ típusú hatékonyságmutatók hányadosaként előállítható. Például, azoknak az ügyfeleknek a száma egy nagyon fontos hatékonyságmutató, amelyek feladják a hívást azok közül, akik továbbra is a sorban várakoznak. De néhány tapasztalt Call Center menedzser mostanság kihagyja azokat az ügyfeleket a számításból, amelyek nem akarnak egy rövid időperiódust sem várakozással tölteni. Náluk a

$$P\{Ab|W > t\} = \frac{P\{V \wedge X; V > X\}}{P\{V \wedge X > t\}} = \frac{E[1_{(t, \infty)}(v \wedge x) 1_{\{v > x\}}]}{E[1_{(t, \infty)}(v \wedge x)]}$$

összefüggés használatos.

Az $E[f(V, X)]$ számítását kezdjük a következő dekompozícióval:

$$\begin{aligned} E[f(V, X)] &= E[f(V, X) \cdot 1_{\{V > 0\}}] + E[f(V, X) \cdot 1_{\{V=0\}}] \\ &= E[f(V, X) \cdot 1_{\{V > 0\}}] + E[f(0, X)] \cdot \left(\pi_B + \sum_{k=0}^{N-1} \pi_k \right). \end{aligned}$$

Minden f függvény számára, amely érdekes esetünkben, $E[f(0, X)]$ értéke 0 vagy 1. Emiatt, mi az első kifejezés kiszámításával megyünk tovább. Három különböző módszert prezentálunk a kalkuláció elvégzéséhez, mindegyiknél megemlítjük előnyeiket és hátrányaikat is. Ezek a módszerek azok, amelyeket kisebb Call Centerek analizálásához használunk.

Az A módszer. Legyen a paraméterünk az ügyfelek száma, amelyek a sorban vannak érkezéskor. $\bar{F}_n(t) = 1 - F_n(t)$ -re legyen az alábbi összefüggés

$$E[f(V, X) \cdot 1_{\{V > 0\}}] = c \pi_N \sum_{k=0}^{B-N-1} (-1)^k \frac{(\lambda/\theta)^k}{k!} I(k) \sum_{n=k}^{B-N-1} \frac{(\lambda/\theta)^{n-k}}{(n-k)!},$$

$$\text{ahol } I(k) = \frac{1}{c+k} \int_0^\infty \int_0^\infty f(t, x) (c+k) \theta e^{-(c+k)\theta t} \theta e^{-\theta x} dt dx = \frac{1}{c+k} E[f(X_k, X)] \text{ ahol } c = N\mu/\theta,$$

és az X_k egy X -től független valószínűségi változó, $\exp(N\mu + k\theta)$ eloszlással.

Az $I(k)$ értékeinek a kalkulációja általában egyszerű feladat. A legnagyobb hátránya ennek a módszernek, hogy az első összeg az előjeleit változtatgatja, amely numerikusan instabil. Emiatt mutatjuk be a következő módszert, amely elkerüli ezt a problémát.

A B módszer. Az A módszerhez hasonlóan kezdve, és az egyik összeg eliminálásával valamint a $\sum_{k=0}^n \binom{n}{k} (-e^{-\theta t})^k = (1 - e^{-\theta t})^n$ reláció használatával elérkezünk a

$$E[f(V, X), 1_{\{V>0\}}] = \theta^2 c \pi_N \sum_{n=0}^{B-N-1} \frac{(\lambda/\theta)^n}{n!} J(n) \text{ összefüggéshez, ahol}$$

$$J(n) = \int_0^\infty \int_0^\infty f(t, x) e^{-(x+ct)\theta} (1 - e^{-\theta t})^n dx dt. \text{ Itt a } J(n) \text{ kalkulációja eléggé költséges, mivel az}$$

integrálok megoldása általában numerikusan történik.

Ezek a módszerek elvesztik vonzerejüket, amikor végtelen pufferral dolgozunk ($B = \infty$). Ekkor mindkét módszerben az összegek végtelenné válnak, és le kell csonkítani azokat valamely ponton, az implementáció működésekor (az A módszerben váltakozó előjelek is problémásak lehetnek a csonkításkor). Mivel ez az eset arra kényszerít minket, hogy figyelembe vegyük a pontosság kérdését, bemutatjuk a harmadik módszert, ami egy egyértelmű numerikus integrál.

A C módszer. Az általánosabb megoldás érdekében nézzük az f_V^+ függvényt ahol $\frac{f_V^+}{P\{V>0\}}$ egy sűrűségfüggvény. f_V^+ adott a

$$f_V^+(t) = N\mu\pi_N \left[1 - \frac{\gamma\left(B-N, \frac{\lambda}{\theta}(1-e^{-\theta t})\right)}{\Gamma(B-N)} \right] \cdot \exp\left\{\frac{\lambda}{\theta}(1-e^{-\theta t}) - N\mu t\right\}, t > 0 \text{ kifejezés által.}$$

Most már csak ki kell számolnunk a dupla integrált.

$E[f(V, X) \cdot 1_{\{V>0\}}] = \int_0^\infty \int_0^\infty f(t, x) \theta e^{-x\theta} f_V^+(t) dx dt$. Az integrál az x -re gyakran analitikusan és könnyen számolható (az f -től függően), ezután nekünk csak egy numerikus integrállal kell foglalkoznunk (a t -re).

Néhány plusz megjegyzés szükséges még a végtelen puffer esethez.

1. Az egyensúlyi állapot szükséges egyenletek megoldásához még mindig szükség van egy végtelen összegre. Palm [20] adott egy megoldást, amelyik kifejezi a stacionárius eloszlást, mint az egyszerűen kiszámolt blokkolási valószínűség függvényét egy $M/M/N/N$ rendszerben (itt $P\{Bl\}$ -lel lesz jelölve), ugyanazzal a beérkezési és kiszolgálási rátákkal:

$$\pi_n = \begin{cases} \frac{P\{Bl\}}{1 + \left(A\left(\frac{\lambda}{N\mu}, \frac{N\mu}{\theta} \right) - 1 \right) P\{Bl\}} \cdot \frac{N!}{n! \left(\frac{\lambda}{\mu} \right)^{N-n}}, & n < N \\ \frac{P\{Bl\}}{1 + \left(A\left(\frac{\lambda}{N\mu}, \frac{N\mu}{\theta} \right) - 1 \right) P\{Bl\}} \cdot \frac{\left(\frac{\lambda}{\theta} \right)^{n-N}}{\left(\frac{N\mu}{\theta} + 1 \right) \dots \left(\frac{N\mu}{\theta} + (n-N) \right)}, & n \geq N \end{cases}$$

ahol $A(x, y) = \frac{ye^{xy}}{(xy)^y} \cdot \gamma(xy, y)$.

2. A $B = \infty$ esetén, az f_V^+ sűrűségfüggvény egy speciális esete lesz Baccelli és Hebuterne [21] egy $M/M/N + G$ modellre, az ügyfelek türelmének F eloszlásával, nevezetesen:

$$f_V^+(t) = N\mu\pi_N \exp \left\{ \lambda \int_0^t (1 - F(u)) du - N\mu t \right\} \quad t > 0.$$

9.4 Az eloszlások közelítése és a működtetési rezsimek

A jelen cikk fő témája, hogyan lehet a hatékonyságmutatókat közelíteni, a paramétereiktől való függés felderítésével a folyamatok mélyebb megismeréséhez jutni. Az első lépés a Q folyamat és annak a π stacionárius eloszlásának a megbecslése, amelyet majd $Q(\infty)$ -nel jelölünk. Az elméleti eredményeinkből, amelyek az uralkodó gyakorlattal is támogatottak, az következik, hogy a hatalmas Call Centerek szintén képesek magas szolgáltatási szintet hozni, mialatt magas kihasználtsággal működnek. A nagy forgalomhoz kapcsolódó közelítésekre igazolására fókuszálunk, amikor az $N \uparrow \infty$. Továbbá a legtöbb nagy telefonos Call Centernek van elég trunk vonala, hogy lényegében ki tudja szűrni az ügyfelek blokkolását (a foglalt jelzéseket). Ezért feltételezhető, hogy a továbbiakban a puffer végtelen, azaz a $B = \infty$.

Megjegyzés. Szeretnénk kiemelni, hogy a semmiképpen nem akarjuk védelmezni itt a „nincs foglalt jelzés” feltétel nélküli használatát. Igazából a foglalt jelzés a legegyszerűbb mód a torlódás továbbadására. A zöld számos szolgáltatásoknál ez a legegyszerűbb mód a működési költségek csökkentésére. Ezért egy megegyezés szüksége a blokkolási és a hívásfeladási arányok között.

Vegyük figyelembe a folyamatok sorozatát a $\{Q_N, N = 1, 2, \dots\}$ -t, ahol a $Q_N = \{Q_N(t), t \geq 0\}$ a sorhossz folyamat, amely az $M/M/N + M$ modellhez kapcsolódik (N kiszolgáló). A N indexet hozzáadtuk a jelölésrendszerünkhöz, hogy az N -edik rendszer paramétereit jelöljük.

Jellemezzük most a modell paramétereinek függőségét az N -től. Specifikusan, minket az a sorozat érdekel, amelyben a, $\lambda_N \uparrow \infty$ ahogy a $N \uparrow \infty$ és $\mu_N \equiv \mu$. Ez a sorozat megfelel a Call Center méretének növelésének ($N \uparrow$) amely bírja az új terhelést ($\lambda_N \uparrow$) mialatt biztosítjuk, hogy a kiszolgálási ráta (μ) nem változik a mérettel (N).

Használjuk a következő két hatékonyságmutatót: azon ügyfelek aránya, amelyek feladják a hívást ($P_N\{Ab\}$), és azok, akik késleltetve lettek a sorban ($P_N\{W > 0\}$). Ezek legyenek az irányelvek a helyes működtetési rezsim kiválasztásához. (Meg kell jegyezni, hogy a sorban való átlagos várakozás lineárisan kapcsolódik a $P_N\{Ab\}$ -hez a $P_N\{Ab\} = \theta_N \cdot E[W]$ -én keresztül).

A legtöbb telefonos Call Center el akarja kerülni, hogy a hívásfeladási arány túl magas legyen. Ez általában lefordítható arra, hogy az ügyfelek nem elhanyagolható százalékának sorba kell állnia, és csak egész kicsiny százalék hagyhatja el a sort. A következőekben egy analitikus eredményben bevezetjük a forgalom intenzitás fogalmát, amit a $\rho_N = \frac{\lambda_N}{N\mu}$ kifejezés definiál.

Tétel 1. Feltételezzük, hogy a $\lim_{N \rightarrow \infty} \rho_N = \rho_\infty$ bizonyos $0 \leq \rho_\infty \leq \infty$ esetén. Ekkor a hívást feladó ügyfelek arányának valószínűségének határértéke adott a

$$\lim_{N \rightarrow \infty} P_N\{Ab\} = \begin{cases} 0 & \rho_\infty \leq 1 \\ 1 - \frac{1}{\rho_\infty} & \rho_\infty \geq 1 \end{cases}$$

összefüggéssel.

Az 1. tételt alapul véve tisztán látható, hogy a hívásfeladások szempontjából, nincs okunk, hogy $\rho_\infty < 1$ -el működjünk, a $\rho_\infty = 1$ már elenyésző hívásfeladási valószínűséget szolgáltat. Másfelől, ha a $\rho_\infty \gg 1$, a hívásfeladási valószínűség határértéke nagyobb, mint ami általában kívánatos. Következtethető az 1-es tételből, hogy a $\rho_\infty = 1 + \varepsilon$ lehet a megfelelő, mivel ez egy $\varepsilon/(1 + \varepsilon)$ határértéket nyújt. Azonban a kiszolgálók kihasználtságának szempontjából (másképpen annak az időmennyiségnek az aránya, amelyet telefonbeszélgetéssel töltenek, adott az $\frac{\lambda_N(1 - P_N\{Ab\})}{N\mu}$ kifejezéssel) a kihasználtság maximális határértékét már elértük a $\rho_\infty = 1$ -el. Ezért a $\rho_\infty = 1$ egy speciális egyensúlyi pont a Call Centerek hatékonysága és minősége között. Kissé alulterhelt ($\rho = 1 - \varepsilon$) vagy túlterhelt ($\rho = 1 + \varepsilon$) Call Centerek tekinthetők ennek a rezsimnek, a zavarainak. A rezsim közelítéseinek magas pontossága az Ábra 15-ről is látható.

A $\rho_\infty = 1$ -ra tett megszorítás, konzisztens Halfin és Whitt [16] munkájával, akik az $M/M/N$ modellt analizálták és megtalálták azt az érdekes határértéket, amely akkor fordul elő ha $\rho_N \sim 1 - \beta/\sqrt{N}$, $0 < \beta < \infty$ (érdekes abban az értelemben, hogy csak ekkor nem degenerálódik a várakozásra kényszerült ügyfelek arányának határértéke). Mivel a $M/M/N + M$ modell nagyon türelmes ügyfelekkel ($\theta_N \downarrow 0$) közel van egy $M/M/N$ modellhez, (ez formálisan támogatott a 2-es tétellel), mi szintén korlátozzuk magunkat a $\lim_{N \rightarrow \infty} \sqrt{N}(1 - \rho_N)$, $-\infty < \beta < \infty$ esethez. Ahogy lehet majd később látni, az 5-ik tétel bizonyítja, hogy ez az érdeklődésünkre érdemes rezsim.

Ami a türelem paramétert illeti, feltételezni fogjuk, hogy a $\lim_{N \rightarrow \infty} \theta_N = \theta$ ahol $1 \leq \theta \leq \infty$. Ez természetszerűleg vezet három rezsimhez.

- a). $\theta = 0$: „Türelmes” ügyfelek
- b). $\theta = \infty$: „Türelmetlen” ügyfelek
- c). $0 < \theta < \infty$: „Kiegyensúlyozott” hívásfeladás.

A „kiegyensúlyozott” esetet Fleming, Stolyar és Simon [22] analizálta. Ez tűnik a megfelelő rezsimnek – indokolt, hogy feltételezzük, az ügyfelek türelme független a kiszolgálók számától, mivel ez a szám általában ismeretlen a hívó számára. Ez azonban vitatható feltételezés, mert amikor hatalmas Call Centereket hívunk, az ügyfélként hajlamosak vagyunk azonnali kiszolgálást elvárni. Annak ellenére, hogy nem tudjuk az egy műszakban

dolgozó kiszolgálók pontos számát, van egy mennyiségi elképzelésük arról. A következő tétel a „kiegyensúlyozott” rezsimeket támogatja, és később magyarázat is lesz hozzáfűzve.

Vegyük tekintetbe a sztochasztikus folyamatok sorozatát, a $\{q_N\}$ -t ahol a $q_N(t) = \frac{Q_N(t) - N}{\sqrt{N}}$, a stacionárius eloszlásuk, a $q_N(\infty)$.

A következőekben egy tételt ismertetünk a $\{q_N\}$ sorozat konvergenciájáról, amelyet $\{Q_N\}$ -ből nyertünk központosítás és arányosítás által. Specifikusan, N körül központosítva, létrehoz egy olyan folyamatot, amelynek az abszolút értéke vagy a sor hossza ($q_N \geq 0$) vagy a tétlen kiszolgálók száma ($q_N \leq 0$). Az arányosító faktor a $\sqrt{N}$ kiemelkedik, mint a helyes nagyságrend, amely létrehozza a nem-triviális folytonos határfolyamatot, a q -t. Ez utóbbit fogjuk használni a saját eredeti születési-halálozási folyamatunk közelítésére ($\{Q_N\}$), a $Q_N \stackrel{d}{\approx} N + \sqrt{N}q$ segítségével. A határtétel támogatja a három határértéket, amelyek egyenként megfelelnek a háromféle ügyfél-viselkedéstípusnak. A matematikai részletei ennek a tételnek nem az előfeltétele a következmények követésének, amelyeket a tétel után magyarázunk azonnal.

2. tétel. Feltételezzük, hogy $\lim_{N \rightarrow \infty} \sqrt{N}(1 - \rho_N) = \beta$, $-\infty < \beta < \infty$. Ha $q_N(0) \stackrel{d}{\rightarrow} q(0)$ akkor

a). Gyenge konvergencia: $q_N \xrightarrow{d} q$, ahol q az egyetlen megoldása a sztochasztikus differenciálegyenletnek, a következő rezsimek szerint

$$\theta = 0: \begin{cases} dq(t) = f(q)dt + \sqrt{2\mu}db(t) \\ f(x) = \begin{cases} -\mu(\beta + x) & x \leq 0 \\ -\mu\beta & x > 0 \end{cases} \end{cases}$$

$$\theta < 0 < \infty: \begin{cases} dq(t) = f(q)dt + \sqrt{2\mu}db(t) \\ f(x) = \begin{cases} -\mu(\beta + x) & x \leq 0 \\ -(\mu\beta + \theta x) & x > 0 \end{cases} \end{cases}$$

$$\theta = \infty: \begin{cases} dq(t) = -\mu(\beta + q(t))dt + \sqrt{2\mu}db(t) - dY(t) \\ Y(0) = 0, Y \uparrow 0, qdY = 0 \end{cases}. \text{ (Itt a } b \text{ standard Brown mozgást}$$

jelöl)

b). Felcserélhető határértékek: $\lim_{N \rightarrow \infty} P\{q_N(\infty) \leq x\} = \lim_{t \rightarrow \infty} P\{q(t) \leq x\}$.

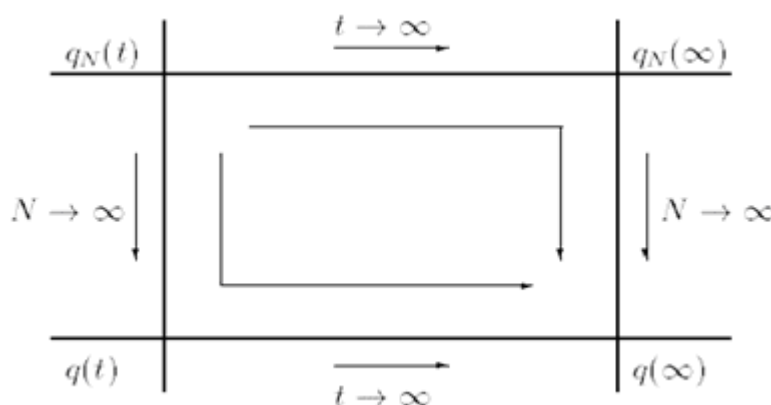
Megjegyzés: Ez utóbbinak itt látható egy ekvivalens reprezentációja, amely specifikusan mutatja, hogy a határértékek fel vannak cserélve: $\lim_{N \rightarrow \infty} \lim_{t \rightarrow \infty} P\{q_N(t) \leq x\} = \lim_{t \rightarrow \infty} \lim_{N \rightarrow \infty} P\{q_N(\infty) \leq x\}$.

A 2. tétel első részéből látható, hogy három különböző határeloszlási folyamat van, azok szerint a rezsimek szerint, amelyeket a θ értékei definiálnak. A „türelmes” ügyfél esetében ($\theta = 0$), a határérték ugyanaz, mint a [16]-ban, amelyik az $M/M/N$ sorok sorozatának nagy forgalom melletti határértékei közül emelkedik ki. Ez azt jelenti, hogy a hívásfeladási jelenség elveszett a határfolyamat során. Hasonlóan, a „türelmetlen” ügyfelek esetében ($\theta = \infty$) a határfolyamat során, maga a sor veszett el – és ugyanahhoz a határértékhez érkeztünk, mint az $M/M/N/N$ („veszteséges” – Erlang B) rendszerek sorozata esetén. Ez ahhoz a következtetéshez vezet, amelyet a számítások is igazolnak:

- a). Különösen türelmes ügyfelek esetén, indokolható az $M/M/N$ modell használata.
- b). Nagyon türelmetlen ügyfelek esetén, az $M/M/N/N$ modellt kell használni.

A „kiegyensúlyozott” esetet ($0 < \theta < \infty$) (egy kissé különböző központosítással) Fleming, Stolyar és Simon [22] következtette ki, és ők adtak bizonyítást a stacionárius eloszlás gyenge határértékére (lásd $q(\infty)$). Általában, ez az eset illeszkedik leginkább az eredeti $M/M/N + M$ rendszerre, mint közelítés. Innentől kezdve ezt használjuk.

A 2. tétel második része fontos, mivel a határértékek ilyen cserélhetősége nem automatikus. A legtöbb alkalmazásban, végül is mindenki a stacionárius eloszlás közelítései iránt érdeklődik, ezért fontos tudni, hogy mindkét útvonal e felé az eloszlás felé tart, ahogy a következő ábrán látható.



Ábra 18 A stacionárius eloszlás $q(\infty)$, mint határeloszlás

A 4. tétel a folytatásban a „kiegyensúlyozott” esethez tartozó közelítést használja.

A 2. tétel használhatósága a Call Center menedzserek számára korlátozott, mivel az ez által nyújtott legtöbb információ a rendszerről szól, nem pedig a kiszolgálásról. Ahhoz, hogy

több információt kapjunk a kiszolgálásról, szükséges a potenciális várakozási idő, vagy ekvivalens módon, a virtuális várakozási idő folyamatának vizsgálata (az előző fejezetben láthattuk, hogy ezeknek a folyamatoknak a határértékei egybeesnek, ahogy az idő korlátlanul növekszik). A Puhalskii [23] által eredményül kapott „Egybeesési Elv”, lehetővé teszi számunkra, hogy létrehozzunk egy egyszerű kapcsolatot a eloszlás határértékei és a virtuális várakozási idő folyamat között (a $0 < \theta < \infty$ esethez). E célból, jelölje ν_N az N -edik rendszer várakozási idő folyamatát.

3. tétel. Feltételezzük, hogy $\lim_{N \rightarrow \infty} \sqrt{N}(1 - \rho_N) = \beta$, bizonyos $-\infty < \beta < \infty$ -ra, és a $\theta_N \equiv \theta$. Ha $q_N(0) \xrightarrow{d} q(0)$ akkor

$$\sqrt{N}\nu_N \xrightarrow{d} \left[\frac{q}{\mu} \right]^+.$$

Ez a központi eredmény heurisztikusan is alátámasztható. Ha vannak tétlen kiszolgálók, a virtuális várakozási idő 0, egyébként a sor hossza körülbelül $q\sqrt{N}$ (a 2. tétel szerint). Milyen sokáig tart egy ügyfélnek keresztüljutni a soron? Az ügyfelek $N\mu$ rátával fogják elhagyni a sort (a kiszolgálás után) plusz $o(N)^5$ (a hívásfeladások; valójában a hívásfeladási ráta a mi megjelölt ügyfelünk előtt nem nagyobb, mint $\theta q\sqrt{N}$). Elosztva a sor hosszát azzal a rátával, amellyel az ügyfelek elhagyják a rendszert, a virtuális várakozási időhöz jutunk, amely emiatt körülbelül $\left[\frac{q}{\sqrt{N}\mu} \right]^+$.

9.5 Implementáció

Az előző két fejezetben az $M/M/N/B + M$ modellel kapcsolatos eredményeket mutattunk be. Mivel hisszük, hogy ennek a modellnek kell a $M/M/N$ és az $M/M/N/B$ modellek helyébe lépni, fényt kell derítenünk arra, hogyan használjuk és értelmezzük ezeket az eredményeket. Az áttekintés kontextusa egy óriási méretű, nagy forgalmú Call Center menedzsment folyamata lesz (a nagy forgalom azt jelenti majd, hogy a hívásfeladás nem elhanyagolható). Elsőként megvitatjuk az modell paraméterek értékeinek becslésének problémáját. Később javaslatot teszünk arra, hogy melyik hatékonyságmutatót kellene a Call Center

⁵ o – kis ordo - A Landautól származó ordó-jelölés (O jelölés) az analízisben és alkalmazásaiban (valószínűség-számítás, analitikus számelmélet) függvények becslését megkönnyítő jelölésmód. Ha nemcsak $|f(x)| \leq Cg(x)$, de $f(x)/g(x) \rightarrow 0$ is teljesül a megadott határátmenetben, azt $f(x) = o(g(x))$ -szel jelöljük és azt mondjuk, hogy „ $f(x)$ egyenlő kis ordó $g(x)$ ”.

menedzsereknek használniuk a szolgáltatási színvonal definiálásához. Végül megvitatjuk a közelítések használatát a gyakorlati szabályok származtatásához.

9.5.1 Paraméterek becslése

Ahhoz, hogy használni tudjuk a modellt és a bemutatott eredményeket, szükséges, hogy beállítsuk a különböző paraméterek értékeit. Bizonyos paraméterek teljesen a Call Center menedzserei ellenőrzése alatt vannak. Például a műszakban dolgozó kiszolgálók száma, az ACD-be befutó trunk vonalak száma. A beérkezési és kiszolgálási ráta általában a historikus ACD adatokból van származtatva. Ahogy az előzőekben megbeszéltük, az időben változó beérkezési ráták esetén, kicsiny időintervallumokat választunk, amelyekben a beérkezési ráta megközelítően konstans.

A legnagyobb problémát a hívásfeladási arány (θ) vagy ekvivalens módon az átlagos türelem ($1/\theta$) becslése okozza. A probléma abból a tényből származik, hogy a közvetlen adat, amelyet gyűjthetünk, ki van válogatva. Csak azoknak az ügyfeleknek a türelmét mérhetjük, akik elhagyják a rendszert, mielőtt a kiszolgálásukat megkezdhetnénk. Azoknál az ügyfeleknél, akik kiszolgáláshoz jutnak, csak az alsó korlátot ismerjük, azt az időmennyiséget, amelyet a sorban töltött várakozással. Vannak olyan statisztikai metódusok, amelyek ilyen kiválogatott adatokkal foglalkoznak, de ezek ismertetésével most nem foglalkozunk. Másik, sokkal alapvetőbb probléma a θ becslésénél, hogy a legtöbb esetben az ACD adat csak átlagokat tartalmaz, a hívásonkénti mérésekkel ellentétben. E célból ajánlunk itt két metódust az átlagos türelem becslésére. Az első a következő egyensúly egyenleten alapul:

$$\theta \cdot E[\text{az ügyfelek száma a sorban}] = \lambda P\{Ab\}$$

Ez az egyenlet leírja az egyensúlyi állapotot azok között az ügyfelek rátája között, akik feladják a hívást (az egyenlet baloldala), és a rendszerbe belépő hívásfeladó ügyfelek aránya között. A Little tétel alapján ($E[\text{az ügyfelek száma a sorban}] = \lambda E[W]$) egy alternatív egyenlethez jutunk

$$\theta \cdot E[W] = P\{Ab\}$$

Az átlagos várakozás a sorban, és a hívást feladó ügyfelek aránya meglehetősen standard ACD adat kimenetek, ezért, eszközt nyújtanak a θ becslésére. Megjegyezzük, hogy az előbbi két egyenlet csak exponenciális türelem esetén érvényes.

A második, egy általánosabb megközelítés. Számoljunk ki egyet a hatékonyságmutatók közül, és hasonlítsuk az eredményt az ACD adatból származtatott értékhez. A cél, hogy addig kalibráljuk a türelem paramétert, amíg a hatékonyságmutató becsült értéke illeszkedni nem kezd a mért értékekre. Az egyik előnye ennek a módszernek, a rugalmassága abban a kérdésben, hogy melyik hatékonyságmutatót válasszuk. Mindegyik megfelelő, amely függ az adott ACD adattól. Továbbá, ez a kalibrálás a modell feltételezéseinek ellenőrzésének egyik formáját is reprezentálja, és kompenzálhat a különbözőségekért. Megjegyezzük azonban, hogy két hatékonyságmutató kiválasztása valószínűleg két becslést idéz elő. További kutatás szükséges, hogy hogyan kombinálható a kettő, hogy egy harmadik, jobb becsléshez jussunk.

9.5.2 A kiszolgálási színvonal

Ahogy korábban már állítottuk, a hívásfeladási jelenség különösen fontos egy Call Center menedzsernek. Ez nem foglalja magába, hogy csak a hívást feladó ügyfelek aránya az érdekes hatékonyságmutató. Sok fontos hatékonyságmutató van, és fontos hogy kiválasszunk néhányat, amely a legjobban tükrözi a kiszolgálási színvonalat az adott Call Centernél, és kiszolgálási célokként vagy kiszolgálási gradiensként szolgálhatnak.

Három alapvető hatékonyságmutatót javasolunk, amelyek a kiszolgálás három fő aspektusát reprezentálják egy Call Centernél:

- a). $P\{Bl\}$ - a blokkolt ügyfelek aránya
- b). $P\{W > t_1; \neg Ab; \neg Bl\}$ - Azok az ügyfelek, amelyek kiszolgálást kaptak „hosszú” várakozási idő után (több mint t_1). Ezek a hűsleges ügyfelek – nem hagyják el a sort.
- c). $P\{Ab; W > t_2\}$ - Azok az ügyfelek, akik feladták a hívást, nem számolva azokkal, akik még egy rövid periódust sem voltak hajlandóak várni (a t_2 rövid).

Ezek a mérőszámok reprezentálják azokat az ügyfeleket, akik hajlandóak voltak egy minimális erőfeszítést tenni, hogy elérjék a Call Centeres kiszolgálót, de rossz szolgáltatást kaptak, vagy semmit sem. Egy Call Center menedzsernek ki kell tűznie célul, hogy ezeket a számokat alacsonyan tartsa, mert ezek olyan ügyfelek, amiket ő meg akar tartani, de elképzelhető, hogy elveszíti őket.

Kíváncos, hogy legyen egy hatékonyságmutató, amely segítségével a kiszolgálás kiértékelhető a Call Centerben. Egy ilyen hatékonyságmutató használható arra, hogy beállítsuk a kiszolgálási színvonal célkitűzéseként. Általában arra használják, hogy beállítsák a kiszolgálók számát, amelyek az adott műszakban szükségesek lehetnek az várt forgalom

kezelésére. Egyetlen egy mérőszám konstruálható több mérőszám (az előbb prezentáltakhoz hasonlóak) súlyozott összegeként (az eredményül kapott gradiens 0 és egy között van, az alacsonyabb gradiens jelent jobb kiszolgálást).

$$\neg SL = a_1 P\{Bl\} + a_2 P\{W > t_1; \neg Ab\} + a_3 P\{W > t_2; Ab\}$$

Itt az $a_1 + a_2 + a_3 = 1$.

Még abban az esetben is, amikor a egy relatíve szimpla modell van prezentálva, egy ilyen mérőszám pontos kiértékelése körülményes abban az értelemben, hogy egy értékekkel feltöltött táblázat produkálása, különösen, ha inverz kalkuláció is szükséges (más szavakkal megtalálni néhány paraméter értékét egy adott hatékonyságmutató értékéhez), nem mindig hajtható végre valós időben. Ezekben az esetekben a közelítések hasznosak lehetnek, ahogy majd látható később. Specifikusan a mi eredményeink alul a következő, nagy forgalom melletti közelítéshez vezetnek (feltételezzük, hogy $P\{Bl\} \approx 0$).

$$\begin{aligned} \neg SL = & a_2 w\left(-\beta, \sqrt{\mu/\theta}\right) \cdot \frac{h\left(\beta\sqrt{\mu/\theta}\right)}{\Psi\left(\beta\sqrt{\mu/\theta} + \sqrt{\theta/(N\mu)}, \sqrt{N\mu\theta}t_1\right)} \cdot e^{-\theta t_1} + \\ & a_3 w\left(-\beta, \sqrt{\mu/\theta}\right) \cdot \frac{h\left(\beta\sqrt{\mu/\theta}\right)}{\Psi\left(\beta\sqrt{\mu/\theta} + \sqrt{\theta/(N\mu)}, \sqrt{N\mu\theta}t_2\right)} \cdot \frac{\Psi\left(\beta\sqrt{\mu/\theta} + \sqrt{\theta/(N\mu)}, \sqrt{N\mu\theta}t_2\right)}{\Psi\left(\beta\sqrt{\mu/\theta}, \sqrt{N\mu\theta}t_1\right)} \cdot e^{-\theta t_2} \end{aligned}$$

Itt $a_2 + a_3 = 1$. Az Ψ , w és a h definíciójához nézze a továbbiakat.

9.5.3 Közelítések

A közelítések akkor hasznosak, ha le akarunk küzdeni számítási nehézségeket, de feltárhatják azt is, hogy a hatékonyságmutatók miként függenek a modell paramétereitől. Ilyen értes szükséges lehet egyszerű gyakorlati szabályok származtatásakor is. A virtuális várakozási idő folyamat eloszlásának közelítését (3. tétel) a hatékonyságmutatók általános reprezentációjával kombinálva, lehetővé tesszük sok hatékonyságmutató közelítésének származtatását. Ezeknek a közelítéseknek olyan, hatalmas Call Centerek esetén kell a legpontosabbaknak lenniük, amelyeket nagy forgalom mellett működtetnek elhanyagolható blokkolási ráta mellett.

Azoknál a hatékonyságmutatóknál, amelyeket kalkulálunk, feltételezzük, hogy a rendszer egyensúlyi állapotban van, emiatt mi a virtuális várakozási idő stacionárius eloszlására vagyunk kíváncsiak. Ilyen közelítés származtatható a következő tétel segítségével.

4. tétel. Legyen a $\nu = \left[\frac{q}{\mu} \right]^+$, ahol a q megoldja azt a sztochasztikus differenciál

egyenletet, amely a $0 < \theta < \infty$ esethez tartozik a 2. tételben (a „kiegyensúlyozott” türelem), és feltételezzük, hogy a $\theta_N \equiv \theta$.

Ekkor:

a). $\nu(\infty) = \lim_{t \rightarrow \infty} \nu(t)$ -nek az eloszlásfüggvénye az F_ν , adott a következő összefüggés által:

$$1 - F_\nu(x) = \begin{cases} 1 - w(-\beta, \sqrt{\mu/\theta}), & x = 0 \\ w(-\beta, \sqrt{\mu/\theta}) \cdot \frac{h(\beta\sqrt{\mu/\theta})}{\Psi(\beta\sqrt{\mu/\theta}, \sqrt{\mu\theta}x)}, & x > 0 \end{cases}$$

b). $\sqrt{N}v_N(\infty) \xrightarrow{d} \nu(\infty)$.

Jegyezzük meg, hogy ebben a tételben megint igazoljuk az megcserélhető határértékeket, most a virtuális várakozási idők sorozatával kapcsolatban $\{v_N\}$. Mivel $\sqrt{N}v_N(\infty) \xrightarrow{d} \nu(\infty)$, a használandó közelítés $V = v_N(\infty) \approx \nu(\infty)/\sqrt{N}$, amely értelmezhető, mint $F_\nu(x) \approx F_\nu(\sqrt{N}x)$.

Példa: Feltételezzük egy hatékonyságmutatót, amely kifejezhető, mint $E[g(W)]$ valamilyen g függvényre. Jusson eszünkbe, hogy a $W \equiv X \wedge V$ és az X és V függetlenek, emiatt:

$$E[g(W)] = \int_0^\infty \int_0^\infty g(x \wedge v) \theta e^{-\theta \cdot x} dF_\nu(v) dx \approx E[g(0)] \cdot (1 - w(-\beta, \sqrt{\mu/\theta})) + \int_0^\infty \int_0^\infty g(x \wedge v) \theta e^{-\theta \cdot x} \sqrt{N\mu\theta} \cdot w(-\beta, \sqrt{\mu/\theta}) \cdot \Psi(\sqrt{N\mu\theta}v + \beta\sqrt{\mu/\theta}, -\sqrt{N\mu\theta}v) dv dx.$$

A következőekben több hatékonyságmutatóhoz eredményül kapott közelítések kerülnek felsorolásra.

$$\begin{aligned} P\{W > 0\} &\approx w(-\beta, \sqrt{\mu/\theta}) \\ P\{Ab|W > 0\} &\approx 1 - \frac{h(\beta\sqrt{\mu/\theta})}{h(\beta\sqrt{\mu/\theta} + \sqrt{\theta/(N\mu)})} \\ P\{Ab\} &\approx \left[1 - \frac{h(\beta\sqrt{\mu/\theta})}{h(\beta\sqrt{\mu/\theta} + \sqrt{\theta/(N\mu)})} \right] \cdot w(-\beta, \sqrt{\mu/\theta}) \\ E[W] &\approx \left[1 - \frac{h(\beta\sqrt{\mu/\theta})}{h(\beta\sqrt{\mu/\theta} + \sqrt{\theta/(N\mu)})} \right] \cdot \frac{w(-\beta, \sqrt{\mu/\theta})}{\theta} \end{aligned}$$

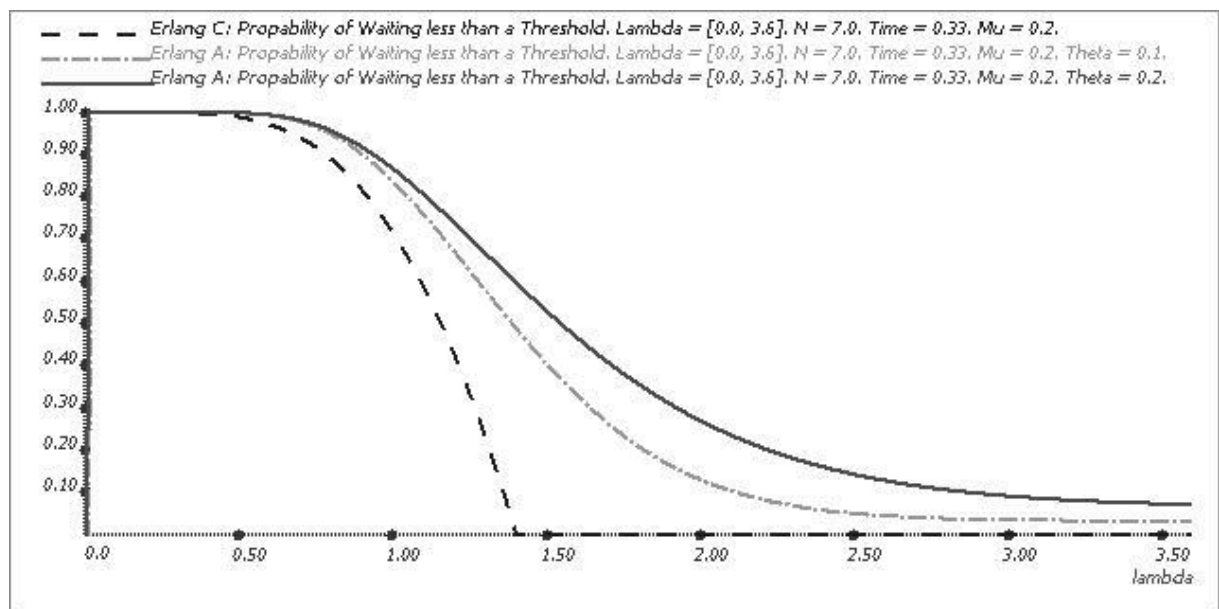
$$P\{W > t\} \approx w\left(-\beta, \sqrt{\mu/\theta}\right) \cdot \frac{h\left(\beta\sqrt{\mu/\theta}\right)}{\Psi\left(\beta\sqrt{\mu/\theta}, \sqrt{N\mu\theta t}\right)} \cdot e^{-\theta \cdot t}$$

$$P\{Ab|W > t\} \approx 1 - \frac{\Psi\left(\beta\sqrt{\mu/\theta}, \sqrt{N\mu\theta t}\right)}{\Psi\left(\beta\sqrt{\mu/\theta} + \sqrt{\theta/(N\mu)}, \sqrt{N\mu\theta t}\right)}$$

$$\text{Itt a } w(x, y) = \left[1 + \frac{h(-xy)}{yh(x)}\right]^{-1}, \text{ a } h(x) = \frac{\phi(x)}{1 - \Phi(x)} \text{ és a } \Psi(x, y) = \frac{\phi(x)}{1 - \Phi(x + y)}.$$

Megjegyzés: Ugyanazokban a sorokban, felépítettünk más hasznos közelítéseket is, érdemes megjegyezni közülük a $E[W|W > t]$ -t.

Végül álljon itt az Erlang A és C modellek összehasonlítása, az előbb ismertetett közelítésekkel számolva. Látható, hogy az Erlang A olyan esetekben is stabil marad, ahol az Erlang C már nem.



Ábra 19 Az Erlang C és az Erlang A TSF-ének alakulása (kisebb és nagyobb türelmü ügyfelekkel)

9.5.4 Gyakorlati szabályok

Fontos, hogy a Call Center menedzser képes legyen érzékelni, mekkora a változtatások hatása a szolgáltatási színvonalra. Ilyen változás lehet a hívásérkezési ráta növekedése egy marketing kampány következtében, vagy változás az ütemezett kiszolgálók számában egy adott műszakban.

A legtöbb hatékonyságmutatóra adott kifejezés, amit az $M/M/N + M$ modellből származtattunk, eléggé komplex. Még az előzőekben megadott közelítések is hajlamosak túlságosan komplexek lenni ahhoz, hogy segítsék annak megértését, ahogyan a paraméterek

értékei befolyásolják a hatékonyságmutatókat. Ez kívánatos, azonban, az egyszerű gyakorlati szabályok származtatásához.

A kiegyensúlyozott hívásfeladási eset megvitatását folytatva, a következő, a Halfin és Whitt [16] $M/M/N$ sorokhoz kapcsolódó összefüggéséhez hasonlító eredményhez jutunk.

5. tétel. Feltételezzük, hogy $\theta_N \equiv \theta$. Ekkor

$$\lim_{N \rightarrow \infty} \sqrt{N}(1 - \rho_N) = \beta, \quad -\infty < \beta < \infty, \text{ akkor és csak akkor, ha}$$

$$\lim_{N \rightarrow \infty} P_N\{W > 0\} = \alpha, \quad 0 < \alpha < 1, \text{ akkor és csak akkor, ha}$$

$$\lim_{N \rightarrow \infty} \sqrt{N}P_N\{Ab\} = \Delta, \quad 0 < \Delta < \infty, \text{ amely esetben az}$$

$\alpha = w(-\beta, \sqrt{\mu/\theta})$ és $\Delta = [\sqrt{\theta/\mu} \cdot h(\beta\sqrt{\mu/\theta}) - \beta] \cdot \alpha$. Itt a w és a h ugyanazok, mint az előzőekben.

Megjegyzés. Ez az eredmény tartható a szélsőértékekre is, nevezetesen

$$\beta = -\infty \text{ akkor és csak akkor, ha } \alpha = 1 \text{ akkor és csak akkor, ha } \Delta = \infty \text{ és}$$

$$\beta = \infty \text{ akkor és csak akkor, ha } \alpha = 0 \text{ akkor és csak akkor, ha } \Delta = 0.$$

A fenti eredmények alapján az $M/M/N + M$ sorokra is kiszámítható a határérték, amikor a $\rho_N \sim 1 - \beta/\sqrt{N}$, de itt a β nem csak pozitív lehet.

Az 5. és az 1. tétel fényében most bemutatjuk a 3 üzemeltetési rezsimet, három illeszkedő gyakorlati szabállyal, amely összeköti a létszámfeltöltési szintet az ajánlott terheléssel, amit az $R = \lambda/\mu$ kifejezés definiál. (Az R Erlangokkal van mérve a telekommunikáció területén, amely egy mértékegység nélküli mennyiség, Egyenlő azzal a munkamennyiséggel ami a rendszerbe érkezik és időegységben mérjük)

- „Minőségvezérelt”: Ilyen Call Centerekben a létszámfeltöltés szintje nagyobb, mint az ajánlott terhelés – hatalmas Call Centerekben ez elhanyagolható hívásfeladási rátához vezet, elhanyagolható várakozási idő mellett. Ilyen Call Centerek analízisének indokolható az $M/M/N$ modell használata. (Ilyenek például a sürgősségi Call Centerek is)
- „Hatékonyságvezérelt”: Ilyenkor a létszámfeltöltés szintje kevesebb, mint az ajánlott terhelés, jelentős számban vannak hívásfeladások, és magas számban várakoznak.
- „Racionalizált”: Azt hisszük, hogy olyan Call Centerekben, ahol telemarketinggel, ügyféltámogatással foglalkoznak, vagy információt szolgáltatnak, az a céljuk, hogy ebben a

rezsimben üzemeltessenek. Itt kevés hívást feladó ügyfél van, és azoknak az ügyfeleknek a száma sem nagy, akinek várakozni kell.

A létszámfeltöltési szint és az ajánlott terhelés közötti mennyiségi kapcsolat az 5. tételből következik. A tétel természetes módon definiál három rezsimeket. A nem degenerált, számunkra érdekes határérték képzése az egyik, valamint két szélsőérték, amire a megjegyzésben hivatkoztunk. Ezek egymás után a racionalizált, hatékonyságvezérelt ($\beta = -\infty$ eset) és a minőségvezérelt ($\beta = \infty$ eset) rezsimekhez tartoznak.

A racionalizált rezsimek létszámfeltöltési szintjét közvetlenül a $\rho_N \sim 1 - \beta/\sqrt{N}$ -ből származtatható $\infty < \beta < -\infty$. Az 5. tétel szélsőértékei csak létszámfeltöltési szint határainak szabnak határt. Minden létszámfeltöltési szint, úgy, mint a $N = \lceil R \pm \varepsilon \cdot R^a \rceil$ $\varepsilon > 0$, $1 \geq a \geq 0.5$ megfelelő, ($+\varepsilon$ a minőségvezéreltnek, és $-\varepsilon$ a hatékonyságvezéreltnek). Azonban mi azt ajánljuk, hogy az a legyen 1, mert ekkor tisztán látható különbség van a kissé kevésbé terhelt (minőségvezérelt) és a kissé túlterhelt (hatékonyságvezérelt) valamint a kritikusan terhelt Call Centerek (racionalizált) között.

A következő táblázatban található irányvonalak közvetlenül az 5. tételből következnek, kivéve a hívást feladók aránya ($P\{Ab\}$) a hatékonyságvezérelt rezsimben, amely az 1. tételen alapul.

Működtetési rezsimek	Létszám szintje	Következmények
Minőségvezérelt	$N = \lceil R \cdot (1 + \epsilon) \rceil$, $\epsilon > 0$	$P\{W > 0\} \rightarrow 0$, $P\{Ab\} = o(1/\sqrt{N})$
Hatékonyságvezérelt	$N = \lceil R \cdot (1 - \epsilon) \rceil$, $\epsilon > 0$	$P\{W > 0\} \rightarrow 1$, $P\{Ab\} \rightarrow \epsilon$
Racionalizált	$N = \lceil R + \zeta\sqrt{R} \rceil$, $\infty > \zeta > -\infty$	$P\{W > 0\} \rightarrow \alpha(\zeta)$, $P\{Ab\} \sim \frac{\Delta(\zeta)}{\sqrt{N}}$

Ábra 20 Gyakorlati szabályok összefoglaló táblázata

Jegyezzük meg, hogy $\alpha(\zeta) = w(-\zeta, \sqrt{\mu/\theta})$ és $\Delta(\zeta) = \left[\sqrt{\theta/\mu} \cdot h(\zeta \sqrt{\mu/\theta}) - \zeta \right] \cdot \alpha(\zeta)$.

Az 5. tételből következik, és Whitt szellemében folytatva, a ζ -t (vagy ε -t) javasoljuk, mint szolgáltatási gradiens. Ennek a gradiensnek a kiemelkedő fontosságát, két rendszer összehasonlíthatósága adja. A gyakorlatban ez egy rendszert jelent, egy remélt változtatás előtt és után. Ha egy menedzser egyszer eldönti, hogy melyik rezsimek lesz az alkalmas a három közül az üzemeltetéshez, meghatározhatja a szolgáltatási gradienst és használhatja a megfelelő gyakorlati szabályt.

Levonjuk a következtetéseket, az első fejezetben megismert forgatókönyv segítségével. Tételezzük fel, hogy a Call Center a racionális rezsimben működik N kiszolgálóval, μ kiszolgálási rátával, és λ beérkezési rátával. A szolgáltatási színvonal mérhető a szolgáltatási gradienssel az ζ -val. Előrejelzés van egy magasabb beérkezési rátára ($\hat{\lambda}$) az ünnepek közeledtével. A Call Center menedzser szeretné ezt a helyzetet kezelni, és eldönteni, hány kiszolgáló dolgozzon a műszakokban ($\hat{N}$). A megfelelő gyakorlati szabály alapján az $\zeta \approx \sqrt{\frac{\mu}{\lambda}} N \left(1 - \frac{\lambda}{N\mu}\right)$, és azt kapjuk, hogy a $\hat{N} = \left\lceil \frac{\hat{\lambda}}{\mu} + \zeta \sqrt{\frac{\hat{\lambda}}{\mu}} \right\rceil$. Továbbá, a hatékonyságmutatók

ünnepekor észlelt értékei:

- a). A várakozók aránya: $P\{W > 0\} \approx \alpha(\zeta)$ (ami ugyanaz, mint az eredeti rendszerben)
- b). A hívást feladók aránya: $P\{Ab\} \approx \Delta(\zeta)/\hat{N}$

10 Összefoglalás

A dolgozatom során a bevezető részen túl, négy különböző, de egymással szoros kapcsolatban álló rendszert elemző forrásokat rendeztem egyfajta logikai sorrendbe. Először az Erlang C került vizsgálódásom középpontjába, amely egyszerűségével igen vonzó a Call Center menedzserek számára. Gyors és pontos algoritmus létezik a kiszámítására, ahogy azt a 6.5. fejezetben idézett forrás is részletesen leírja.

Azonban ennek a rendszernek komoly hiányosságai is vannak. Az első abból adódik, hogy lényegében azonos híváskezelési időkkal számol, amely nem mindig tartható feltételezés. Ennek feloldását tárgyalja a 7. fejezet, amely aszimptotikus közelítést ad a késleltetési valószínűségekre, és ebből származtatja a TSF-et és az ASA-t. Ebben a fejezetben lesz említve először az iparágban létező három létszámfeltöltési rezsím, irányvonal, amelyeket egyenként Hatékonyság-, Minőségvezéreltnek és Racionalizáltnak nevezünk. Fontos megemlíteni a szolgáltatási gradienst, amelynek szinten tartása a Racionalizált rezsím feladata, a többletlétszám (Hatékonyságvezérelt) vagy a rendszerkihasználtság (Minőségvezérelt) szinten tartásával szemben. A Racionalizált rezsímnek ugyanez a célja az Erlang A rendszerben is.

Az Erlang C végtelen türelmet feltételez a hívók részéről, ami torlódás esetén 0 TSF-hez vezet, amelynek elkerülése a Call Center menedzserek elsődleges célja. A trunk vonalak számának beállíthatósága olyan eszköz a kezükben, amellyel elkerülhetik az ilyen túlterhelési szituációkat, cserébe lesznek olyan ügyfelek, amelyek foglalt jelzést kapnak. Ezeknek az arányáról ad pontos képet a 8. fejezet, amely az Erlang B modellel kapcsolatos forrásokból építkezik. Szintén létezik pontos és valós időben számoló algoritmus, amelyet a 8.3. fejezet ismertet olyan rendszerekben, ahol a késleltetés valószínűsége nulla (0 sorhossz).

Végül az Erlang A modellel foglalkozó cikkekből merít a diplomamunkám. A 9.2 fejezetben látható egy olyan modell, amelyben a különböző, eddig tárgyalt rendszerek ismérvei vannak összegyűrva. Ebben a modellben, ha valaki szabad az N számú kiszolgáló közül, akkor a hívás azonnal oda lesz irányítva. Ha nem de $N < B$ ahol B a trunk vonalak száma, és van szabad hely a maximálisan $B - N$ hosszúságú sorban, akkor ott kell várakoznia a hívásnak. A többi foglalt jelzést kap. Ebben a fejezetben a középponti kérdés a sorban várakozó ügyfelek sorsa, amelyek az Erlang C modellel ellentétben nem várnak örökké, hanem feladják a hívást. Ennek az arálynak a számítása fontos hatékonyságmutató ebben a modellben az ASA, TSF és a Blokkolási valószínűség mellett, amelyeket már az előző

modellekben is említettünk. Létezik pontos számítás, amelyeket a források segítségével a 9.3. fejezetben említünk. De ennél is fontosabbak a közelítések, amelyek kellően pontosak, és valós időben számolhatóak (9.5.3. fejezet). A 9. fejezetet a többletlétszám kiszámítására adott gyakorlati szabályok ismertetésével fejezem be a 3 különböző rezsimben.

A dolgozatom eredménye a különböző források szakmai színvonalú fordítása, logikai sorrendbe rendezése, a jelölések normalizálása olyan módon, hogy az eredmények összehasonlíthatóak legyenek. Továbbá kifejlesztésre került egy szoftver az egyes eredmények szemléltetésére. Ilyen ábrák láthatóak a dolgozat egyes részeiben, valamint a függelékben. Ugyanitt a képernyőfelvételek is megtalálhatóak a szoftverről.

A hatékonyságmutatók kritikus fontosságát hivatottak hangsúlyozni a szoftver által végrehajtott számítások. Ezek:

- a) az ASA az Erlang A és Erlang C modellekben
- b) a TSF az Erlang A és az Erlang C modellekben
- c) A blokkolási valószínűség az Erlang B modellben
- d) A hívásfeladók aránya az Erlang A modellben
- e) A várakozási arány a hívásfeladók körében

A számítások az hívásbeérkezési, híváskezelési intenzitás, a kiszolgálószám, az AWT és a türelem paraméterek segítségével paraméterezhetők.

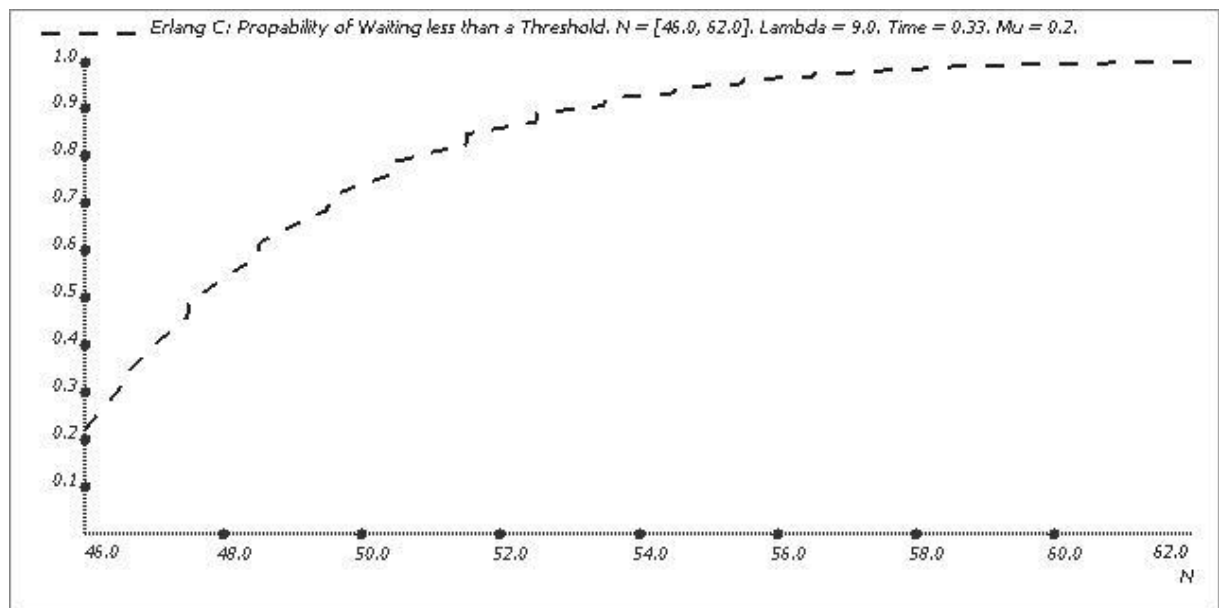
A tárgyalt forrásokban említenek olyan forgatókönyveket, amelyekre nem ad választ az itt tárgyalt egyik modell sem. A legfontosabb ezek közül az újratárcsázási szituáció, amely az eddigi exponenciális hívásbeérkezési intenzitásra gyakorolhat hatást. A dolgozat számításai arra adhatnak választ, hogy egy rövid időperiódusban mekkora az a kiszolgálószám, amely hatékonyan kezelni tudja az akkor jelentkező terhelést, az elvárt minőségi kritériumok teljesítése mellett. Azonban a források tovább olvasásával arra is választ kaphat a kedves olvasó, hogy milyen megfontolások vannak hosszabb távon, a kiszolgálók ütemezésekor.

11 Irodalomjegyzék

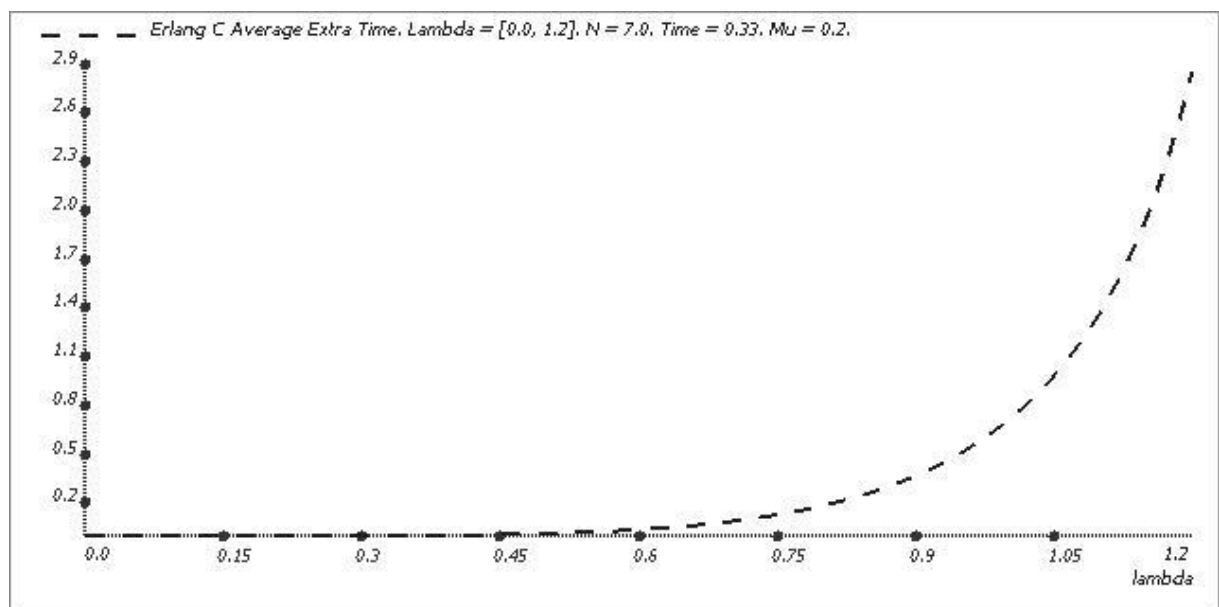
- [1] Valószínűségszámítás – Fazekas, István – 2000
- [2] Handbook of Mathematical Functions – Abramowitz, Stegun – 1972
- [3] Queuing Systems, Vol. 1: Theory. – L. Kleinrock – 1976
- [4] Introduction to Probability Models – S.M. Ross – 1997
- [5] A proof of the queuing formula $L = \lambda W$ – J.D. Little – 1961
- [6] A last word on $L = \lambda W$ – S. Stidham – 1974
- [7] Poisson arrivals see time averages – R.W. Wolf – 1982
- [8] Stochastic modelling and the theory of queues – R.W. Wolf – 1989
- [9] http://en.wikipedia.org/wiki/Sociotechnical_systems_theory#Socio-technical_system
- [10] Call Center Mathematics – Ger Koole – 2007
- [11] Telephone Call Centers: tutorial, review and research prospects. – Noah Gans, Ger Koole, Avishai Mandelbaum – 2003
- [12] Advanced topics in queuing theory: focus on customer contact centers – Ward Whitt – 2002
- [13] Designing a Call Center with impatient customers – O. Garnett – A. Mandelbaum – M. Reiman – Október 8, 1999
- [14] The Heavy Traffic Approximation in the Theory of Queues. – Kingman, J.F.C. – 1965
- [15] A queuing model for telephone operator staffing – Sze D. Y. – 1984
- [16] Heavy-traffic limits for queues with many exponential servers – Halfin, S., W. Whitt – 1981
- [17] Robust algorithms for sharing agents with multiple skills – Borst, S. C., P. Seri – 2000
- [18] Service engineering course material – Mandelbaum, A., S. Zeltyn – 2001
- [19] Performance characteristics of automated call distribution systems – Feinberg M.A. – 1990
- [20] Intensitätsschwankungen im fernsprechverkehr – Palm, C. – 1953
- [21] On queues with impatient customers – Baccelli, F., Hebuterne, G. – 1981

- [22] Heavy traffic limit for a mobile phone system loss model – Fleming, P.J., Stolyar, A., Simon, B. – 1994
- [23] On the invariance principle for the first passage time. – Puhalskii, A. – 1994

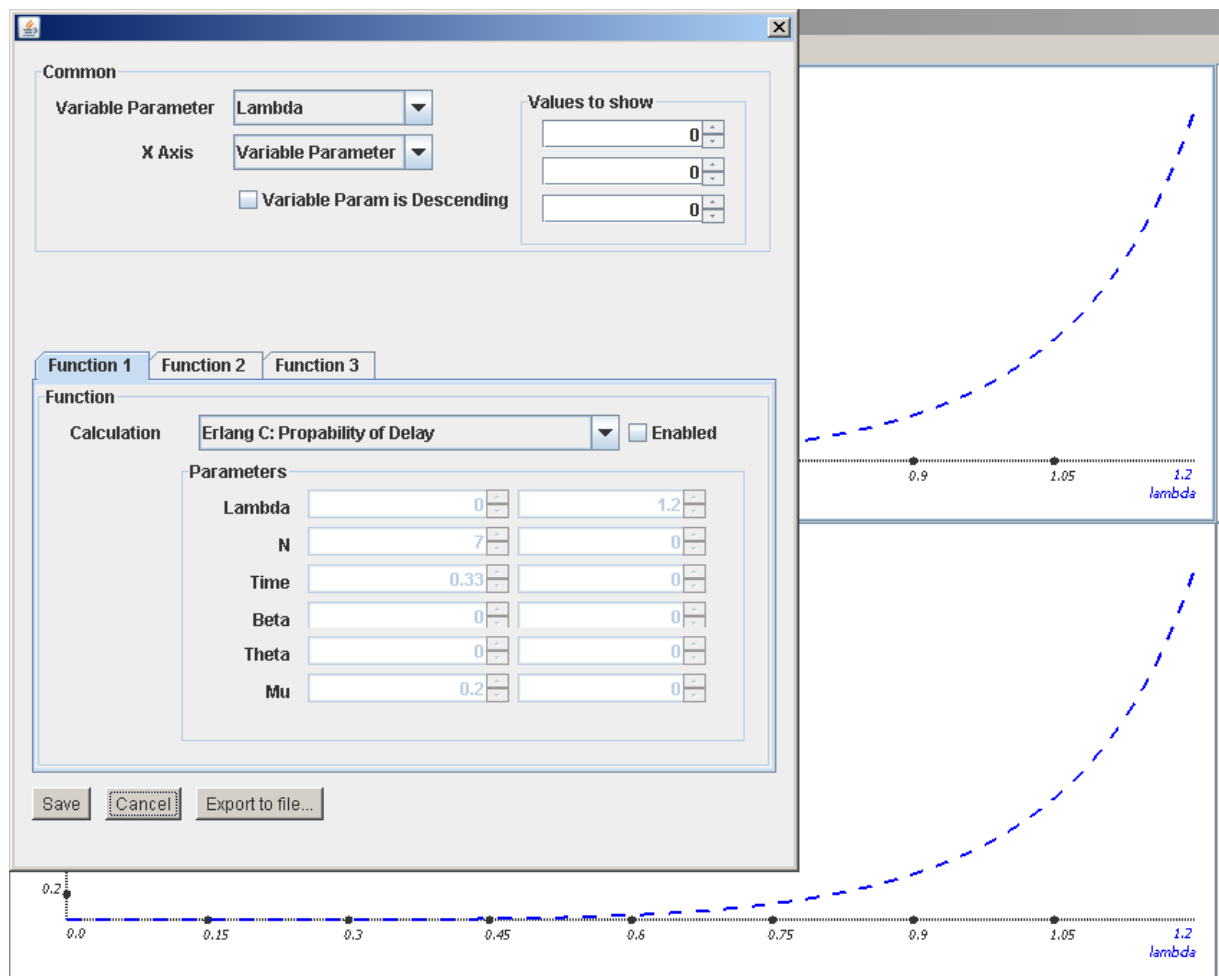
12 Függelék



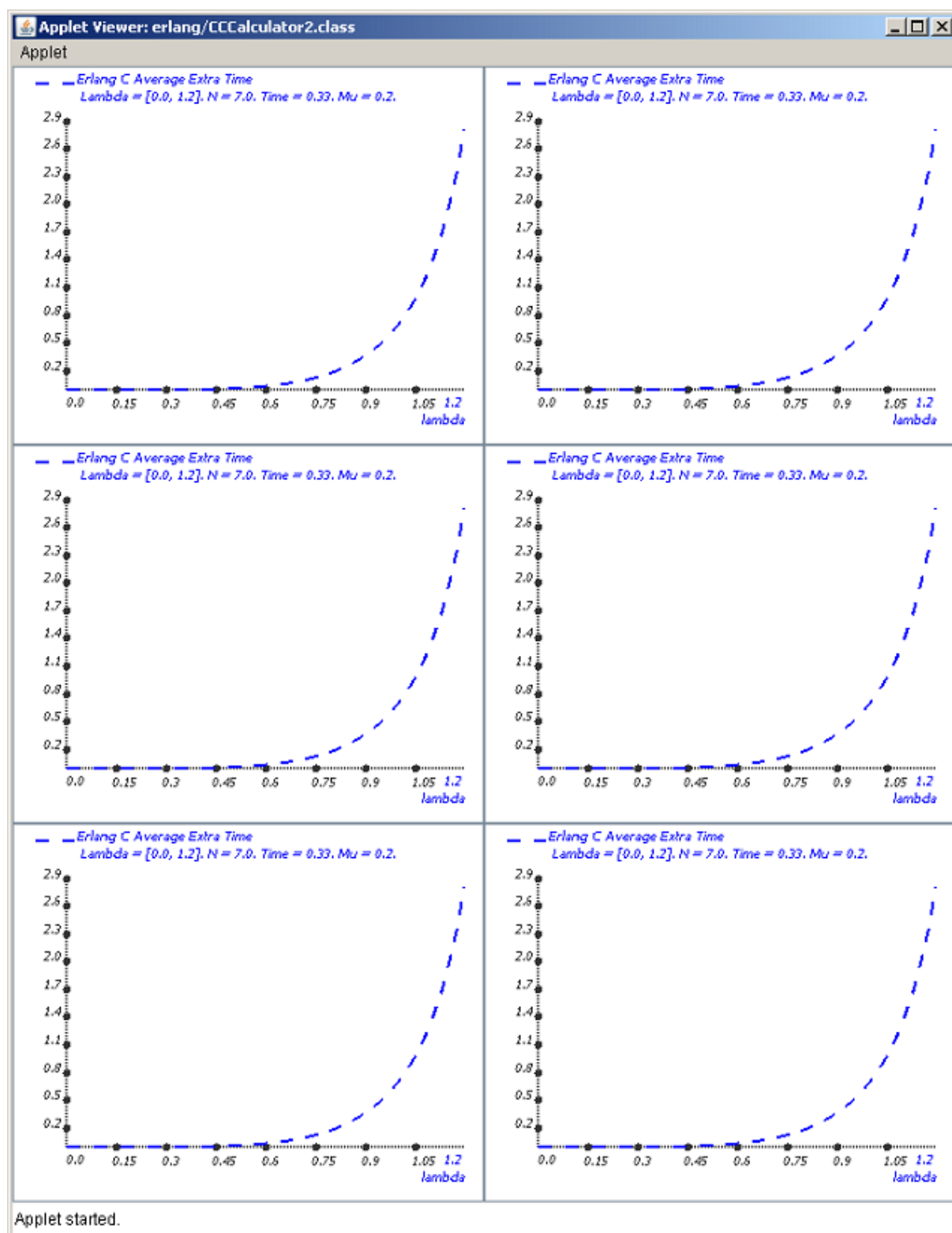
Ábra 21 A TSF változása már akár 2 kiszolgáló hiányzása esetén is jelentős



Ábra 22 Egy alternatív hatékonyságmutató: AET. Az AWT-n felül várakozással eltöltött idő várható értéke.



Ábra 23 Paraméter képernyő



Ábra 24 A program megjelenítő része

13 Köszönetnyilvánítás

Mindenekelőtt köszönet illeti Dr. Sztrik János Professzor Urat a diplomamunka témavezetéséért, elsősorban azért, hogy segített a célok kitűzésében és az eredmények elbírálásában.

Ezen kívül szeretnék köszönetet mondani mindazoknak, akik közvetett módon hozzájárultak a dolgozat megírásához akkor, amikor pályám során, ügyfélszolgálatosként vagy szoftverfejlesztőként közel kerültem a Call Centerek világához.

Végül rokonaimnak és neked is, Beus köszönöm azt a sok türelmet és biztatást, ami nagyban megkönnyítette a helytállást az egyetemi pályafutásom során, csakúgy mint az élet más területein is.