Néhány gyakoribb várakozósoros modell rendszertervezéshez.

Dr. Gyarmati G. Péter

1976. július.

Tartalomjegyzék

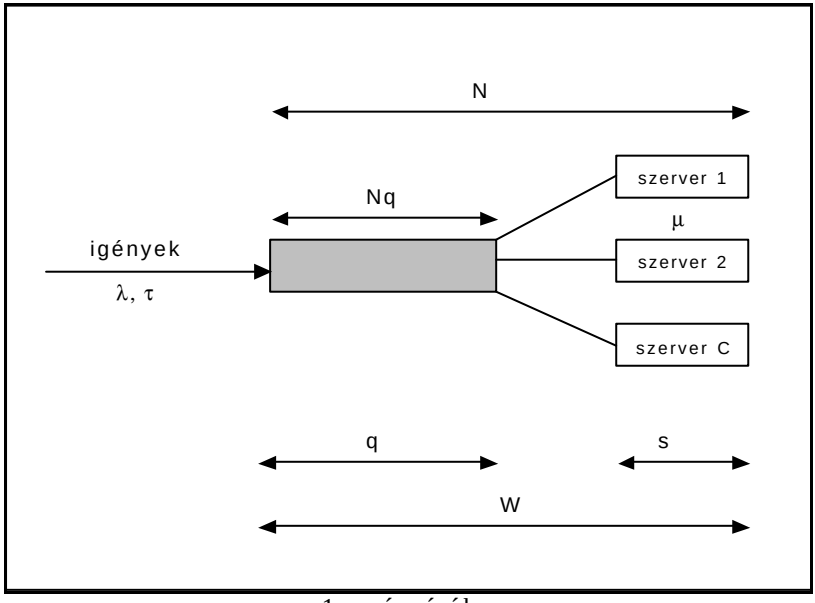
Néhány gyakoribb várakozósoros modell rendszertervezéshez.

1. Bevezetés

A sorbanállási elmélet, --amelyet telefon hálózatok tervezéséhez A. K. Erlang dán matametikus dolgozott ki-- a számítógépes rendszerek tervezésének és vizsgálatának hasznos és nélkülözhetetlen eszközévé vált a ráfordított adaptációs és fejlesztési munka eredményeként.
Az elmélet segitségével elore meghatározhatók a számítógép teljesítoképességének jellemzo adatai, például a várakozási ido egy on-line terminálnál, vagy a tárolóigények üzenetközvetíto rendszerekben, vagy a prioritás hozzárendelések hatása, stb.
Ez a tanulmány ilyen és más hasonló kérdések tárgyalásához szükséges elméleti eszközök bemutatásával --a gyakorlati alkalmazás síkján-- kíván foglalkozni.
A sor a kiszolgálásra várakozó igények tárolója például egy kommunikációs-, vagy számítástechnikai rendszerben, mig a sorbanállási elmélet ezek kezelésének a tanulmányozására, tervezésére alkalmas eszköz.
A feldolgozásra várakozó igények egy rendszerben sort alkotnak. Az ilyen kiszolgálásra várakozó sorok további sorokat alkothatnak, mint például az input-output igények sora egy feldolgozási programban.

2. A várakozósor leírása

Az elméletben az alapelem az *igénylo*, aki, vagy ami igényli és élvezi a kiszolgálást, ezért az ilyen rendszereket –-legyen az kommunikációs, számítástechnikai, informatikai, vagy kereskedelmi, pénzügyi stb. természetu-- *kiszolgálási rendszer*eknek nevezzük és a tervezéshez modellezzük.
Az igénylo által keltett igénylések sorban állnak és várakoznak a kiszolgálásra, ezért *várakozósoros modell*nek nevezzük.
Az informatikai rendszerekben az igénylo lehet például egy átviendo üzenet, egy feldolgozási program, egy választ váró kérdés, egy adatkérés, egy eredményközlés, vagy ezek bármilyen ismétlése, kombinációja, stb.



1. számú ábra

Az igénylo kiszolgálást kér --üzenet átvitelét egy csatornán, egy program utasitásainak végrehajtását, egy kifejezés értelmezését, egy adat megkeresését, adat kérését, átvételét más forrásokból, eredmények közlését, stb.-– egy erre alkalmas kiszolgáló-, szolgáltató forrásból.

A kiszolgáló forrás egy vagy több kiszolgáló egységbol áll, amelyek a kivánt szolgáltatást nyújtják az igényloknek.
Ha az összes kiszolgáló egység foglalt, amikor az igénylo belép a rendszerbe, akkor az igény egy várakozó sorba kerül és várakozik addig, ameddig egy kiszolgáló egység a számára fel nem szabadul.
Az 1.sz ábrán látható az általános modell a használatos kifejezésekkel, a változókkal és jelentésükkel.
A sorbanállási rendszert, vagy modellt matematikai tanulmányozásához az alábbi módon írhatjuk le.

A forrás

A forrás, ahonnan az igények származnak, illetve maguk az igények.Az igények forrása lehet véges, vagy végtelen.

Véges forrású rendszerben nem lehet a kiszolgálásra várakozó sor tetszolegesen hosszú és az igények száma a rendszerben befolyásolja a beérkezési arányt.

Extrém esetben, ha minden igény vár, vagy kiszolgálás alatt van, akkor a beérkezési arány nulla értékü lesz.

Ha a források száma véges, de nagy, akkor végtelen forrást feltételezhetünk az alkalmazott matematikai módszer kezelhetosége érdekében.

Az igények beérkezése

Feltételezzük, hogy az igények $t_0 < t_1 < t_2 < t_3 ... < t_k ...$ idoben érkeznek a rendszerbe. A beérkezések közötti idot a $t \approx t_k - t_{k-1}$ (ahol $k \geq 1$) valószínuségi változóval írjuk le, amelyrol feltételezzük hogy független és azonos eloszlású valószínuségi változók sorát képezi.

Tetszoleges beérkezések közötti ido esetén a τ jelet használjuk. A beérkezéseket a beérkezési közötti ido eloszlásfüggvényével $A(t) = P(t \leq t)$ jellemezzük.

Ha a beérkezési eloszlásfüggvény exponenciális, –-a következo igény beérkezésének valószínusége az ido múlásával exponenciálisan növekszik-- tehát minden t-ra $P(t \leq t) = 1 - e^{-lt}$, akkor annak valószínusége, hogy bármely *t* idoben *n* igény érkezik be: $e^{-lt}(lt)^n / n!$, ahol *n=0,1,2,...* és l

az átlagos beérkezési arány. Ebben az esetben a beérkezések a Poisson eloszlást követik.
Gyakran használatos még az Erlang-*k* és a konstans eloszlás.
A használatos eloszlásfüggvények leírásai megtalálhatók az [1., 2.] irodalomban.

A kiszolgálási ido

Legyen s_k a *k*-adik igény kiszolgálásához szükséges ido és feltételezzük, hogy az s_k egy független, azonos eloszlású valószínuségi változó. (Tetszoleges kiszolgálási ido esetén az s jelölést használjuk.)

A kiszolgálási ido eloszlásfüggvénye legyen $W_s(t) = P(s \le t)$.

Az átlagos kiszolgálási arányt m -vel jelöljük.

Véletlen kiszolgálás esetén --exponenciális eloszlás-- az eloszlásfüggvény: $W_s(t) = 1 - e^{-mt}$, ha $t \ge 0$.

A gyakoribb eloszlásfüggvények még az Erlang-*k* és a konstans eloszlás.
Leírásuk az [1., 2.] irodalomban található.

Egy paraméter

Egy hasznos statisztikai paraméter az *A(t)* és W(t) eloszlásfüggvények jellemzésére az u.n. négyzetes variációs együttható: $C_x^2 = \frac{\mathrm{var}(x)}{E(x^2)}$.

Ha *x* egy konstans valószínuségi változó, akkor $C_x^2 = 0$,

ha *x* exponenciális eloszlású, akkor $C_x^2 = 1$,

ha *x* Erlang-*k* eloszlású, akkor $C_x^2 = 1/k$.

Következtetések:

a. Ha C_t^2 megközelitoleg nulla, akkor a beérkezéseknek szabályos rendje van, ha közel van az egyhez, akkor a beérkezések véletelenszerunek jellemezhetok és végül, ha nagyobb egynél, akkor a beérkezések csoportosulására lehet következtetni.

b. Ha C_s^2 megközelitoleg nulla, akkor a kiszolgálási ido megközelitoleg állandó, nagy értékei pedig a kiszolgálási ido nagy mértéku változásait jelzik

A várakozósoros rendszer kapacitása

Egyes rendszerekben a sor kapacitását *végtelen*nek feltételezik, vagyis minden igény addig várakozhat, amig kiszolgáláshoz nem jut.

Más rendszerekben a sor kapacitása *nulla*, vagyis, ha minden kiszolgáló egység foglalt amikor egy igény beérkezik, akkor az igény visszautasításra kerül.

Megint más rendszerek *véges* kapacitással rendelkeznek, vagyis bizonyos mennyiségu igény felvételére képesek.

A kiszolgáló egységek száma

A legegyszerubb várakozósoros rendszer *egy* kiszolgáló egységgel rendelkezik, egyidoben egy igényt szolgál ki.

A több kiszolgáló egységu rendszerek C számu azonos kiszolgáló egységgel rendelkeznek és egyidoben *C* számu igényt elégítenek ki.

Végtelen számú kiszolgálási egységu rendszerben minden igény azonnal kiszolgálásra kerül.

A sorbanállási rend

A sorbanállási rend, vagy kiszolgálási rend az a szabály, amely szerint kiválasztásra kerül a kiszolgálásban sorrakerülo következo igény.

A leggyakrabban alkalmazott ilyen szabály az „elso be - elso ki”, a *FIFO* (first-in, first-out), vagy másnéven *FCFS* (first-come, first-served). Ez nem más, mint a mindennap emlegetett sorszám.

Egy másik másik gyakori rend az „utolsó be - elso ki”, a *LIFO* (last-in, first-out), vagy más néven *LCFS* (last-come, first-served). Ilyen modell például a közismert veremautomata.

Megemlítheto még a „véletlen kiválasztás”, az *RSS* (Random Selection for Service, más nevén *SIRO* (Service In Random Order) és a „fontossági sorrend”, a *PRI* (priority service).

A röviditések kombinálásával további modellek is létrehozhatók, amelyeknek a mi vizsgálataink szempontjából kevéssé van jelentosége.

Mellékesen megemlitjük a fogyasztói társadalomban kialakult „közgazdasági modellt, a *GIGO*-t (garbage-in, garbage-out)". Ennek tárgyalása nem tárgya jelen tanulmányunknak.

3. A modell leírása

A továbbiakban az eddig bemutatott alapfogalmak szerinti sorbanállási, vagy várakozósoros rendszerek gyakorlati használatra is megfelelo modelljét mutatjuk be.

A rendszerek rövidített jelölése

A sorbanállásos rendszereket az u.n. *Kendall jelölés*sel szokás rövidített formában leírni az alábbi módon:

A/B/c/K/m/Z, ahol

A a beérkezések közötti ido eloszlása,

B a kiszolgálási idö eloszlása,

C a kiszolgáló egységek száma,

K a rendszer kapacitása (a tárolható igények száma),

M a források száma,

Z a sorbanállási rend.

Gyakran csak az *A/B/c* jelölést használjuk, ilyenkor a várakozósornak nincs felso korlátja, a források száma végtelen és a várakozási rend *FIFO*.
Az *A* és *B* helyén a következo eloszlásokat használjuk a leggyakrabban:

D konstans eloszlás (kiszolgálásra, beérkezésre),

E-k Erlang-k eloszlás ((kiszolgálásra, beérkezésre),

Gi általános, független beérkezések közötti ido,

G általános kiszolgálási ido a függetlenség feltételezésével,

M exponenciális eloszlás (kiszolgálásra, beérkezésre).

Például az M/E4/3/20/8/SIRO rendszer exponenciális eloszlású beérkezések közötti idovel, három kiszolgáló

egységgel, azonos Erlang-k kiszolgálási ido eloszlással, 20 igény (ebbol 3 kiszolgálás alatt, 17 a várakozósorban) kapacitással, végtelen számú igényforrással és véletlen kiválasztásos (minden várakozó igény u.a. valószínuséggel kerül sorra) kiszolgálási renddel jellemezheto.

A rendszer forgalma

A rendszer forgalma a forgalom intenzitásával --*(u)*-- jellemezheto, ami a kiszolgálás várható értékének –*E(s)*– és a beérkezések közötti ido várható értékének --*E(t)*-- az aránya. Ez az arány a sorbanállásos rendszer egyik legfontosabb paramétere és az alábbi formulával írható le:

$$u = \frac{E(s)}{E(t)} = lE(s) = \frac{l}{m}$$

A forgalom intenzitása *(u)* határozza meg a kiszolgálási egységek minimális számát, amely szükséges a rendszer egyensúlyához a beérkezo igények számát tekintve.

Ha például *E(t)= 10 sec és E(s)= 15 sec*, akkor legalább két kiszolgáló szükséges, mivel: $\frac{E(s)}{E(t)} = \frac{15}{10} = 1{,}5 \rightarrow 2$.

Az *u* egységét *erlang*nak nevezték el A.K. Erlang után.

A kiszolgálási egység foglaltsága

Egy másik fontos paraméter a kiszolgálási egységenkénti intenzitás: *r*, ami *u/c*, ha a forgalom egyenletesen oszlik meg az egységek között.

A kiszolgálási egységek foglaltsága annak a valószínusége, hogy az adott egység foglalt. A nagy számok törvénye értelmében a *r* annak az idonek a megközelíto része, amikor minden egység foglalt.

Egy valószínuség: *t* idoben *n* igény van a rendszerben

Ez a valószínuség $p_n(t)$ nemcsak *t*-tol függ, hanem a rendszer iniciális állapotától is, vagyis a kiszolgálás indulásakor már fennálló igények számától.

A legtöbb muködo rendszer esetében *t* növekedésével $p_n(t)$ közelíti és eléri a p_n értéket, ami független *t*-tol és így az iniciális állapottól. Az ilyen rendszert *kiegyensúlyozott rendszer*nek nevezzük.

Ebben a tanulmányban csak kiegyensúlyozott rendszerekkel foglalkozunk, mivel az idofüggo és tranziens megoldások általában túl bonyolultak a gyakorlati használatra. De a legfontosabb, hogy a legtöbb esetben -- amelyekrol itt szót ejtünk -- kiegyensúlyozott rendszer kialakítása a célunk. A nem kiegyensúlyozott, vagyis az idofüggo, illetve tranziens rendszerek tervezésével, illetve az ilyen jelenségek felismerésére, elkerülésére irányuló módszerekkel például az Adaptiv irányítások (a [6., 9.] irodalom) foglalkoznak.

További jellemzok

A sorbanállási elmélet a rendszer teljesítményének statisztikus mérését teszi lehetové és így segíti a rendszertechnikust, hogy az igényelt kiszolgálási szinthez szükséges minimális ráfordítású rendszert tervezhessen, alakíthasson ki.

Ezek a statisztikus mértékek és változataik többek között az alábbiak:

- a várakozósorban eltöltött ido várható értéke: *Wq*,
- a rendszerben eltöltött ido várható értéke: *W*,
- a sorban álló igények varható értéke: *Lq*,
- a rendszerben lévo igények várható értéke: *L*.

Ezek a mértékek nem függetlenek egymástól, ezért valamelyik ismeretében a többi már kiszámítható. A kapcsolatok az alábbi formulákban fejezhetok ki:

1. $u = \frac{E(s)}{E(t)} = lE(s) = \frac{l}{m}$
2. $r = \frac{u}{c}$
3. $W = q + s$

4. $W = E(W) = E(q) + E(s) = W_q + W$

5. $N(t) = N_q(t) + N_s(t)$

6. $N = N_q + N_s$

7. $L = E(N) = lW = E(N_q) + E(N_s)$

8. $L_q = E(N_q) = lW_q$

Ha például *Wq*-t ismerjük, akkor a többi már számolható:

Lq= *l*W*q*, *W*=*Wq*+*Ws*, *L*= *l*W.

További hasznos teljesítési mérték a rendszer válaszadási idejének értéke, illetve statisztikusan az u.n. 90%-hoz tartozó érték, a $p_w(90)$. Ez az az ido, amelynél nem többet tölt el a rendszerben a beérkezo igények 90%-a. Formálisan a $p(w \le p_n(90)) = 0{,}9$ egyenlettel fejezheto ki. A 90%-os idoérték a várakozósorra is definiálható ugyanígy, jele $p_q(90)$.

4. Modellek egy kiszolgáló egységgel.

Ebben a fejezetben néhány gyakrabban elofor duló modell leírását és formuláit --a levezetések mellozésével-- adjuk meg gyakorlati használatra.

Az M/M/1 modell

A leggyakrabban alkalmazott modell az M/M/1 tipus, az egyszeru forma és a fontosabb változók eloszlásának pontos meghatározhatósága miatt.

A modell exponenciális eloszlású beérkezések közötti- és kiszolgálási idovel rendelkezik.

Jelentosen különbözik ez a modell azoktól, amelyeknél –-a legtöbb modell ilyen-– a fontosabb változók átlagos, vagy várható értéke, esetleg még a szórása számítható „csak" ki. Egy másik elonye, hogy gyakran célszeru véletlenszeru beérkezésekkel számolni, ugyanakkor a véletlenszeru kiszolgálás feltételezése a gyakorlati szokásokkal találkozik.

A CPU tipusú kiszolgálási ido eloszlásoknál a szórás sokkal nagyobb lehet a várható értéknél és a modell esetenként túl optimisztikus következtetésekre vezethet.

Az alábbiakban az M/M/1 modell kiegyensúlyozott rendszerekre vonatkozó formuláit adjuk meg a levezetések mellozésével:

-a rendszer foglaltsága

$$r = u = \frac{E(s)}{E(t)} = lE(s) = \frac{l}{m} \quad mert\ c = 1$$

-annak valószínusége, hogy a rendszerben n darab igény van

$$P_n = P(N = n) = (1 - r)r^n \quad n = 0,1,2,...$$

-annak valószínusége, hogy a rendszerben n-nél több igény van

$$P(N \geq n) = r^n \quad n = 1,2,...$$

-a rendszerben lévo igények várható száma

$$L = E(N) = \frac{r}{1-r}$$

-a rendszerben lévo igények szórásnégyzete

$$s_N^2 = \frac{r}{(1-r^2)}$$

-a rendszerben tartózkodási ido eloszlásfüggvénye (annak a valószínusége, hogy az igény legfeljebb t idot tolt el a rendszerben)

$$W(t) = P(w \le t) = 1 - e^{-m(1-r)t} = 1 - e^{-t/E(w)}$$

-a rendszerben eltöltött ido várható értéke

$$W = E(w) = \frac{E(s)}{1-r} = \frac{1}{m(1-r)}$$

-a rendszerben eltöltött ido szórásnégyzete

$$s_w^2 = E(w)^2$$

-a várakozósor hosszának várható értéke

$$L_q = E(N_q) = \frac{r^2}{1-r} \quad ha \;\; N_q \ge 0$$

$$= \frac{1}{1-r} \quad ha \;\; N_q > 0 \quad (nem \;\; üres \;\; sor)$$

-a várakozósor hosszának szórásnégyzete

$$s_{Nq}^2 = \frac{r^2(1+r-r^2)}{(1-r)^2} \quad ha \; N_q \ge 0$$

$$= \frac{r}{(1-r)^2} \qquad ha \; N_q > 0 \quad (nem \; üres \; sor)$$

-a várakozósorban eltöltött ido eloszlásfüggvénye (annak a valószínusége, hogy az igény legfeljebb t idöt tölt el a várakozósorban)

$$W_q(t) = P(q \le t) = 1 - re^{-m(1-r)t} = 1 - re^{-t/E(w)}$$

-a várakozósorban eltöltött ido várható értéke

$$W_q = E(q) = \frac{rE(s)}{1-r} = \frac{r}{m(1-r)} \quad ha\ q \ge 0$$

$$= \frac{E(s)}{1-r} = \frac{1}{m(1-r)} \quad ha\ q \quad 0 \quad (nem\ üres\ sor)$$

-a várakozósorban eltöltött ido szórásnégyzete

$$s_q^2 = \frac{r(2-r)(E(s))^2}{(1-r)^2} = \frac{r(2-r)}{m^2(1-r)^2} \quad ha\ N_q \ge 0$$

$$= \left(\frac{(E(s))^2}{1-r}\right)^2 = \left(\frac{1}{m(1-r)}\right)^2 \quad ha\ N_q > 0 \quad (nem\ üres\ sor)$$

-az r százalékos ido a rendszerben

$$p_w(r) = \frac{E(s)}{1-r}\log\left(\frac{100}{100-r}\right) = E(w)\log\left(\frac{100}{100-r}\right)$$

-a 90%-os ido

$$p_w(90) = 2{,}3E(w)$$

-a 95%-os ido

$$p_w(95) = 3E(w)$$

-az r százalékos ido a várakozósorban

$$p_q(r) = E(w)\log\left(\frac{100r}{100-r}\right) = \frac{E(q)}{r}\log\left(\frac{100r}{100-r}\right)$$

-a 90%-os ido

$$p_q(90) = E(w)\log(10r)$$

-a 95%-os ido

$$p_q(95) = E(w)\log(20r)$$

A valószínuségi változók többsége az M/M/1 modellben közismert formájú:

- az igények száma a rendszerben (*N*) geometrikus eloszlású,

- a rendszer válaszadási ideje (*w*) exponenciális eloszlású,

- a várakozósorban eltöltött ido (*q*) vegyes eloszlású

- (diszkrét belépéskor: *P(q=0)*=1-*r*, *folytonos* később).

Kiegyensúlyozott rendszerben a várakozósorban eltöltött ido eloszlása az alábbi: $W_q(t) = 1 - r^{-t/E(w)} \quad ha \quad t \geq 0$

A modell formuláiból kitunik néhány változó erosen nem-lineáris függése a rendszer foglaltságától (ρ). Ilyen a rendszer különbözo részeiben eltöltött ido. A non-linearitás azt mutatja, hogy magas rendszerfoglaltság esetén a várakozási ido nagyon megnövekszik és minden határon túl no $r \to 1$ esetén. Mintegy $r=0,8$ értéknél a foglaltság egészen kis növekedése dramatikusan lerontja a rendszer átbocsájtó képességét.
Ennek elkerülésére az igények prioritásának bevezetésére van szükség. Egy alkalmasan megtervezett prioritásos rendszer magas foglaltság esetén is jól muködik. Erre a késobbiekben még visszatérünk.

Az M/M/1 modell alkalmas az u.n. léptékhatás bemutatására: Ha az átlagos beérkezési ido aránya l és az átlagos kiszolgálási ido arány m megkétszerezodnek (változatlan $r=l/m$ mellett !), az eltöltött ido várható értéke úgy a várakozósorban *E(q)*, mint a rendszerben *E(W)* egyaránt megfelezodik. Az igények számának várható értéke mindkét helyen változatlan marad.

Általában ha az új rendszerben l helyett nl és m helyett nm áll, akkor a lépték-viszony az alábbi lesz:

$$\frac{E(q)_{új}}{E(q)_{régi}} = \frac{\left(\frac{r}{(1-r)nm}\right)}{\left(\frac{r}{(1-r)m}\right)} = \frac{1}{n}$$

és

$$\frac{E(w)_{új}}{E(w)_{régi}} = \frac{\left(\frac{1}{(1-r)nm}\right)}{\left(\frac{1}{(1-r)m}\right)} = \frac{1}{n}$$

A léptékhatás egy következtetésre ad lehetoséget. Ha egy nagyszámítógép terhelését egyenloen elosztjuk n darab olyan kisebb számítógép között, amelyek sebessége a nagy gép $1/n$-szerese, akkor a válaszadási ido nem változik és a felhasználók alkalmasabban elhelyezheto és üzembiztosabb rendszerhez jutnak.
Errol részletesebben a [12., 14.] irodalomban olvashatnak.

Az M/M/1/K modell

($K^3 1$ és az igények száma $n£K$)
Ez a modell az M/M/1 modelltol annyiban különbözik, hogy a rendszer igényfelvételi kapacitása véges és legfeljebb K lehet $(n£K)$.
Ilyen rendszereknél az igények visszautasításra kerülnek, ameddig $n=K$-val, vagyis a rendszer teljes igénytároló kapacitása foglalt.
A modell az alábbi formulákkal jellemezheto a levezetések mellozésével:

-a rendszer foglaltsága

$$r = (1 - p_k)u$$

-a valószínuség, hogy a rendszerben n darab igény van

$$P_n = \frac{(1-m)u^n}{1-u^{K+1}} \qquad ha \;\; l \neq m \quad és \quad n = 0,1,2,..,K$$

-a valószínuség, hogy a beérkezo igény visszautasításra kerül

$$P_k = P(n = K)$$

-a tényleges beérkezési arány

$$l_a = (1 - p_k)\, l$$

-a rendszerben lévo igények várható száma

$$L = E(n) = \frac{u(1-(K+1)u^K + Ku^{K+1})}{(1-u)(1-u^{K+1})} \qquad ha \quad l \neq m$$

$$= \frac{K}{2} \qquad ha \quad l = m$$

-a rendszerben eltöltött ido várható értéke

$$W = E(w) = \frac{L}{l_a} = \frac{E(n)}{(1-p_k)\, l}$$

-a várakozósor hosszának várható értéke

$$L_q = E(N_q) = L - (1 - p_0)$$

-a várakozósorban eltöltött ido várható értéke

$$W_q = E(q) = \frac{L_q}{l_a} \qquad ha \quad q \geq 0$$

$$= \frac{W_q}{1-p_0} \qquad ha \quad q > 0$$

Az M/M/1/K modell bizonyos szempontból jobb, mint az M/M/1 modell, amint az a formulákból is látszik a rendszer stabil marad még akkor is, ha a beérkezési arány l meghaladja a kiszolgálási arányt m, mivel a rendszer telitettsége esetén a beérkezo igények visszautasításra kerülnek.

Az M/G/1 modell

Ennél a modellnél általában nem nyerhetok az *N, Nq, W, q* eloszlási függvényei úgy, mint az M/M/1 esetében és így csak a várható értékek kiszámítására szoritkozhatunk.

Ha a kiszolgálási ido várható értékének elso három momentuma -- $E(s),\ E(s^2),\ E(s^3)$ -- ismert, akkor több valószínuségi változó várható értéke és szórása számítható.

A modell az alábbi formulákkal jellemezheto:

-a várakozósor hosszának várható értéke

$$L_q = E(N_q) = \frac{l^2 E(s^2)}{2(1-r)} = \frac{l^2 s_s^{\ 2} + r^2}{2(1-r)}$$

-a várakozósorban eltöltött ido várható értéke

$$W_q = E(q) = \frac{L_q}{l} \qquad ha \quad q \geq 0$$

$$= E(q \mid q > 0) = \frac{W_q}{r} \qquad ha \quad q > 0$$

$$= E(q^2) = \frac{l E(s^3)}{3(1-r)} + \frac{1}{2}\left(\frac{l E(s^2)}{1-r}\right)^2$$

a várakozósorban eltöltött ido szórása

$$s_q^{\ 2} = E(q^2) - W_q^2$$

-a rendszerben lévo igények számának várható értéke

$$L = E(N) = L_q + r$$

-a rendszerben eltöltött ido várható értéke

$$W = E(w) = \frac{L}{l}$$

$$E(w^2) = E(q^2) + \frac{E(s^2)}{1-r}$$

-a rendszerben eltöltött ido szórása

$$s_w^2 = E(w^2) - W^2$$

-a rendszerben lévo igények számának a szórása

$$s_N^2 = \frac{l^3 E(s^3)}{3(1-r)} + \left(\frac{l^2 (E(s^2)}{2(1-r)} \right)^2 + \frac{l^2 (3-2r) E(s^2)}{2(1-r)} + r(1-r)$$

Az $E(q^2)$, s_q^2, $E(w^2)$, s_w^2, s_N^2 csak akkor számítható, ha a kiszolgálási ido elso három momentuma szerinti értékek ismertek.
A kiegyensúlyozott rendszerre vonatkozó *L* és *W* értékek egyenletei az irodalomban <u>*Pollacsek-Khintchine*</u> egyenletekként ismertek.
Adott *E(s)* esetén az *L, Lq, W és Wq* értékek akkor minimálisak, ha a kiszolgálási ido szórása nulla. (s_s^2), azaz a kiszolgálási ido *konstans*. Ilyenkor a momentumok kiszámíthatók és ez az M/D/1 modellhez vezet

<u>Az M/D/1 modell</u>
Az M/D/1 modell az M/G/1 speciális esetének tekintheto, amikor
- a kiszolgálási ido: *E(s)=konstans*,
- a momentumok: $E(s^2) = E(s)^2$ illetve $E(s^3) = E(s)^3$,
- a szórás természetesen nulla ($s_s^2 = 0$).

A modell formulái az alábbiakra egyszerusödnek:

-a várakozósor hosszának várható értéke

$$L_q = E(N_q) = \frac{r^2}{2(1-r)}$$

-a várakozósorban eltöltött ido várható értéke

$$W_q = E(q) = \frac{L_q}{l} = \frac{l E(s)}{2(1-r)} \qquad ha \quad q \geq 0$$

$$= E(q \mid q > 0) = \frac{W_q}{r} = \frac{E(s)}{2(1-r)} \quad ha \quad q > 0$$

a várakozósorban eltöltött ido szórása

$$s_q^2 = \frac{rE(s)^2}{3(1-r)} + \frac{r^2 E(s)^2}{4(1-r)^2}$$

-a rendszerben lévo igények számának várható értéke

$$L = E(N) = L_q + r = \frac{r^2}{2(1-r)} + r$$

-a rendszerben eltöltött ido várható értéke

$$W = E(N) = \frac{L}{l} = W_q + E(s) = \frac{rE(s)}{2(1-r)} + E(s)$$

-a rendszerben eltöltött ido szórása

$$s_w^2 = s_q^2$$

-a rendszerben lévo igények számának a szórása

$$s_N^2 = \frac{r^4}{4(1-r)^2} + \frac{r^3}{3(1-r)} + \frac{r^2(3-2r)}{2(1-r)} + r(1-r)$$

Az *M/Ek/1* modell

Az M/Ek/1 modell egy fontos modell, mert számos esetben a kiszolgálási ido jól közelítheto az Erlang eloszlással. További elonye az Erlang eloszlás alkalmazásának, hogy a

momentumok ismertek és így a *q*, *w*, *N* változók várható értékei és a szórás számíthatók. Ezek ismeretében a 90%-os és 95%-os értékek is becsülhetokké válnak.Ez a modell végülis az M/G/1 modell speciális eseteként is tekintheto. A kiszolgálási ido eloszlási függvényének a momentumai az alábbi módon számíthatók ki:

$$E(s^2) = \frac{(k+1)}{k}(E(s))^2$$

$$E(s^3) = \frac{(k+1)(k+2)}{k^2}(E(s))^3$$

Ezek behelyettesítésével az M/G/1 modell formuláiba az $E(q^2), L_q, s_q^2$ értékek már kiszámolhatóak.

A *q*, *w*, *N* százalékos valószínuségei becslésére gamma-disztribúciós táblázat alkalmazását ajánljuk, mivel az Erlang-k a gamma-disztribúció speciális esete, ahol a k paraméter diszkrét egész értéku. A táblázatok az u..n. négyzetes variációs együttható: $C_x^2 = \frac{\mathrm{var}(x)}{E(x^2)}$ különbözo értékeihez adják meg a *p*-t.

Ha *x* Erlang-*k* eloszlású, akkor $C_x^2 = 1/k$.

5. Elsobbségi, prioritásos modellek

Az elsobbségi, vagy prioritásos modellek lényege, hogy az igényeket, vagy forrásaikat számozott osztályokba sorolják 1-tol *n*-ig. Az alacsonyabb sorszámúnak elsobbsége van a magasabb számúval szemben, azaz az *i* osztályba tartozónak elsobbsége van a *j* osztálybeli elott, ha *i<j*, továbbá az 1 osztályba tartozó az összes elott van. Az azonos osztályba tartozók kiszolgálása FIFO, beérkezési sorrend szerint történik.

Az *i* igény a beérkezésekor a már várakozó *i* igények mögé kerül, ha ilyen nincs, akkor a *j* igény elé kerül ahol i<j.

A j kiszolgálásának befolyásolásától függoen kétféle taktika lehetséges:

Kivárásos rendszer, non-preemptive
vagy HOL (head of line)
i csak akkor kap kiszolgálást, amikor *j* kiszolgálása befejezodött.

Megszakitásos rendszer, preemptive
j kiszolgálása azonnal megszakad és *i* kiszolgálása megkezdodik.

Ezután a visszatérés a *j* kiszolgálásához kétféle módon lehetséges:
- megszakitás ujraindítással, preemptive-repeat:
  a *j* kiszolgálása teljes egészében újraindul,
- megszakitás folytatással, preemptive-resume:
  a *j* kiszolgálása a megszakitástól folytatódik.

Az M/G/1 prioritásos modell fobb összefüggései

Általános összefüggések:

- az átlagos beérkezési arány

$$l = l_1 + l_2 + ... + l_n$$

- a kiszolgálásban eltöltött ido várható értéke

$$E(s) = \frac{l_1}{l} E(s_1) + \frac{l_2}{l} E(s_2) + ... + \frac{l_n}{l} E(s_n)$$

$$E(s^2) = \frac{l_1}{l} E(s_1^2) + \frac{l_2}{l} E(s_2^2) + ... + \frac{l_n}{l} E(s_n^2)$$

- a rendszer forgalma

$$u_j = l_1 E(s_1) + l_2 E(s2) + ... + l_j E(sj) \quad \text{ahol} \quad j = 1,2,...,n$$

$$u_n = u = l E(s)$$

Nonpreemptive esetben:

- a várakozósorban eltöltött ido várható értéke

$$E(q_j) = \frac{l E(s^2)}{2(1-u_{j-1})(1-u_j)} \qquad \text{ahol} \quad j = 1,2,...,n \quad \text{és} \quad u_0 = 0$$

$$E[q] = \frac{l_1}{l} E[q_1] + \frac{l_2}{l} E[q_2] + ... + \frac{l_n}{l} E[q_n]$$

- a rendszerben eltöltött ido várható értéke

$$E[w_j] = E[q_j] + E[s_j] \qquad \text{ahol} \quad j = 1,2,...,n$$

$$E[w] = E[q] + E[s]$$

- a várakozósor hosszának várható értéke

$$L_q = E[N_q] = l E[q]$$

$$L = E[N] = l E[w]$$

Preemptive-resume esetben:

- a várakozósorban eltöltött ido várható értéke

$$E[q_j] = E[w_j] - E[s_j] \qquad \text{ahol} \quad j = 1,2,...,n$$

$$w_q = E[q] = \frac{l_1}{l} E[q_1] + \frac{l_2}{l} E[q_2] + ... + \frac{l_n}{l} E[q_n]$$

- a rendszerben eltöltött ido várható értéke

$$E[w_j] = \frac{1}{1-u_{j-1}} \left[E[s_j] + \frac{\sum_{i=1}^{j} l_i E[s_i^2]}{2(1-u_j)} \right] \quad \text{ahol} \quad j = 1,2,...,n\,;\ u_0 = 0$$

$$W = E[w] = E[q] + E[s]$$

- a várakozósor hosszának várható értéke

$$L_{q_j} = E\left[N_{q_j}\right] = l_j E\left[q_j\right] + E\left[s_j\right] \quad \text{ahol} \quad j = 1,2,\ldots,n$$

$$L_q = E\left[N_q\right] = lE\left[q\right] = lw_q$$

$$L = E\left[N\right] = lE\left[w\right] = lw$$

6. Összefoglalás

Az itt összefoglalt technika a számítóközpontok üzemének tervezéséhez használatos módszerek egyik készlete. Segitségével tervezési és döntési értéku válasz nyerheto --a számítógépekkel kapcsolatos elvárások alapján-- a rendszer konfigurációs adatainak meghatározására.

Ez az összeállítás megjelent a KSH Rendszertechnikai Közlemények sorozatában. Az ott ismertetett módon került sor ennek a technikának alkalmazására KSH számítóközpontja legújabb gépének konfigurálására és az Üzemeltetési Kézikönyv és -Szabályzat elkészítésére.

A szerzo köszönetet mond munkatársainak a tanulmány elkészítéséhez nyújtott segítségükért, értékes javaslataikért.
Budapest, 1975.

7. Mellékletek

A. Az alkalmazott jelölések és fogalmak

$A(t)$ -a beérkezések között eltelt ido eloszlásfüggvénye, vagyis $A(t) = P(t \le t)$

C -az azonos kiszolgálók száma

D -állandó eloszlás (beérkezés, vagy kiszolgálás)

E_k -Erlang-k eloszlás (beérkezés, vagy kiszolgálás)

$E(N_q | N_q > 0)$ -a nem-üres várakozósor hosszának várható értéke, vagy átlaga

$E(q|q > 0)$ -a nem-üres várakozósorban eltöltött ido várható értéke, vagy átlaga

FCFS -first-come, first-served, lásd még FIFO

FIFO -first-in, first-out, beérkezési sorrend szerint

G -általános eloszlás (kiszolgálás)

GI -általános, független eloszlás (beérkezés, vagy kiszolgálás)

GIGO -garbage-in, garbage-out

K -a rendszer kapacitása: legfeljebb ennyi igény lehet a várakozósorokban és a kiszolgálókban együttesen

L -*E(N)*, a rendszerben lévo igények számának várható értéke

L_q -*E(Nq)*, a várakozósorban lévo igények várható száma, vagy átlaga

LCFS -last-come, first-served, lásd még LIFO

LIFO -last-in, first-out, az utoljára beérkezett következik (verem automata)

l -az átlagos beérkezési ido

M -exponenciális eloszlás (beérkezés, vagy kiszolgálás)

m -az átlagos kiszolgálási ido

N(t) -valószínuségi változó, a rendszerben lévo igények várható száma *t* idopontban

N -valószínuségi változó, a rendszerben lévo igények számának várható értéke

Nq(t) -valószínuségi változó, a várakozósorban lévo igények várható száma t idopontban

Nq -valószínuségi változó, a várakozósorban lévo igények számának várható értéke

Ns(t) -valószínuségi változó, a kiszolgálásban lévo igények várható száma t idopontban

Ns -valószínuségi változó, a kiszolgálásban lévo igények számának várható értéke

Pn(t) -annak a valószínusége, hogy t idopontban n igény van a rendszerben

Pn -annak a valószínusége, hogy n igény van a rendszerben

PRI -elsobbségi kiszolgálási szabály

q -valószínuségi változó, az igénynek a várakozósorban eltöltött ideje a kiszolgálás megkezdése elott

RSS -Random Selection for Service, véletlenszeru kiszolgálás lásd még SIRO

r -a kiszolgáló kihasználtsága: $r = \frac{l}{cm}$

s -valószínuségi változó, a kiszolgálási ido leírására

SIRO -Service In Random Order, minden igény ugyanolyan valószínuséggel kerül kiszolgálásra, lásd még RSS

t -valószínuségi változó, a beérkezések között eltelt ido leírására

u -rendszer forgalom: $u = \frac{E(s)}{E(t)} = lE(s) = \frac{l}{m}$

w -valószínuségi változó, a rendszerben eltöltött ido leírására

W(t) -a *w* (a rendszerben eltöltött ido) eloszlásfüggvénye: $W(t) = P(w \le t)$

W -*E(w)*, az igénynek a rendszerben eltöltött ideje

Wq(t) -a w_q (a várakozósorban eltöltött ido) eloszlásfüggvénye:

Wq -*E(q)*, a várakozósorban eltöltött ido várható értéke

Ws(t) -a w_s (a kiszolgálásban eltöltött ido) eloszlásfüggvénye:

Ws -*E(s)*, a kiszolgálásban eltöltött ido várható értéke.

B. Irodalomjegyzék

1. G.A. Korn: Matematikai kézikönyv muszakiaknak. Muszaki, 1975.

2. Prékopa A.: Valószínuségelmélet, Muszaki, 1974.

3. P. Denning: The working set model of program behaviour. Comm. ACM., 1968.

4. E.G. Coffman: Analysis of two time-sharing algorithms designed for limited swapping. Journal of ACM., 1968.

5. L. Kleinrock: A continuum of time-sharing algorithms. Proc of AFIPS., 1970.

6. Gyarmati P.: Dinamikus eroforrás elosztás vegyes üzem esetén. SZTAKI Közl., 1975

7. A. Brandwajn: A model of time-sharing virtual memory system using equvivalence and decomposition method. Acta info., 1974.

8. T. Beretvas: A simulation model representing OS/VS2 CP. Proc of OS., 1974.

9. P. Gyarmati: On the adaptiv control of Operating Systems., 1975.

10. KSH Rendszertechnikai Közlemények /3. Szerk.: Gyarmati Péter., 1974.

11. G.S. Shedler: A queuing model of multiprogrammed computer with a two level storage system. Comm. ACM., 1973.

12. L. Kleinrock: Queuing systems vol.1 theory., 1975.

13. W. Chang: Single-server queuing processes in computer systems. IBM Sytems Journal., 1970.

14. D. Gross and C. M. Harris: Fundamentals of Queuing Theory. -The McMillan Co., 1974.