Teoria aşteptării

mgorun@inf.ucv.ro

A înțelege şi a învăța cum să gestionezi o linie de aşteptare, (*a waiting line*), o "coadă" (*a queue*) este una din provocările timpurilor noastre.

Rezolvare ştiințifică a acestei probleme este dată de ***teoria aşteptării.***

Termeni specifici teoriei aşteptării

- ***Clienții potențiali*** sunt acele persoane care se presupune că se vor aşeza la rândul de aşteptare pentru a fi servite;
- ***Service***-ul este servirea într-o anumită manieră a acestor clienți potențiali îcare sosesc la rând;
- ***Server***-ul desemnează pe cel ce efectuează service-ul;

- ***Canal de servire*** sau ***canalele de servire*** corespund în principiu numărului de servere.

Menționăm că poate exista un singur rând care se bifurcă în fața mai multor servere sau fiecare server are propria coadă.

- ***Populația*** din care provin clienții reprezintă sursa clienților potențiali și poate fi finită sau nu.

Identitatea unui sistem de aşteptare este dată de cele mai importante relații dintre elementele sale.

Prezentăm criteriile după care plasăm un sistem de aşteptare într-o anumită categorie.

- ***Repartiția timpilor inter-sosire*** este repartiția timpilor care se scurg între două sosiri consecutive ale clienților la rând, aşa numiții ***timpii inter-sosire***.

În cadrul acestei prezentări considerăm că sosirea clienților la coadă se face unul câte unul.

- ***Repartiția timpului de service*** este repartiția timpului în care un client este servit.

- ***Capacitatea maximă a sistemului de aşteptare*** este numărul maxim de clienți cărora li se permite să se aşeze la coadă.

Capacitatea maximă poate fi finită, caz în care orice client ce soseşte după ce a fost atinsă nu mai are voie să se aşeze la rând sau infinită

- ***Numărul de servere*** . Există modele de aşteptare cu un singur server, ***single server*** sau cu mai multe servere, ***multiple server***.

In general şi în cazul mai multor servere există tot un singur rând de aşteptare.

- ***Disciplina aşteptării*** este ordinea de servire a clienților aşezați la coadă: primul sosit, primul servit (*first come, first served*), ultimul sosit, primul servit (*last come, first served*) sau selecție aleatoare pentru service (*random selection for service*)

Exemplu:
În general disciplina cozii cea mai des întâlnită este natural primul sosit, primul servit, dar în cazul unui lift, ultimul ce intră în lift coboară primul, adică ultimul sosit, primul servit.

Notația consacrată pentru diferitele lase de modele de aşteptare se datorează lui *M.* Kendall şi anume avem modele de aşteptare de tipul *A*/B/*c*/K/*m*/Z, unde:

- *A* este tipul de repartiție a sosirilor;
- B este tipul de repartiție a timpului de service;
- *c* este numărul de server
- K reprezintă capacitatea maximă a sistemului de aşteptare
- *m* este mărimea populației
- *Z* reprezintă tipul de disciplină a cozii.

Cele mai frecvente repartiții ale timpilor inter-sosire sunt: ***sosiri Poisson***, notate M (M provine de la proprietatea markoviană a lipsei memoriei), ***sosiri k-Erlang*** notate E_k, ***sosiri generale***, notate GI.

Repartiția Poisson

Repartiția Poisson este o repartiție discretă.

Repartiția unei variabile aleatoare X care se supune legii Poisson de parametru $\lambda > 0$ este dată de

$$p_X(k) = P(X = k) = \frac{e^{-\lambda} \cdot \lambda^k}{k!},\ k = 0, 1, 2 \ldots$$

Media şi dispersia vor fi date de

$$E(X) = \lambda \cdot \sum_{i=0}^{\infty} k \cdot e^{-\lambda} \cdot \frac{\lambda^{k-1}}{k!} = \lambda \ \text{ şi } \ D^2(X) = 1$$

Repartiția Poisson guvernează în anumite ipoteze numărul de evenimente produse într-un anumit interval de timp.

Exemplu

- La secția de urgență a unui spital vin zilnic accidentați.

 în urma accidentelor rutiere, numărul acestora fiind modelat după o variabilă Poisson cu $\lambda = 3$.

 Pentru a calcula probabilitatea ca astăzi sa nu vină niciun accidentat auto la urgență ținem seama că este vorba de o variabilă Poisson si astfel $P(X=0) = e^{-3} = 0.0498$

Analiza timpilor dintre aparițiile succesive ale evenimentelor (sosiri), numiți ***timpi inter-sosire*** este deosebit de importantă.
Astfel vom considera T_1 timpul până la apariția primului eveniment, T_2 timpul între apariția primului eveniment şi apariția celui de al doilea eveniment, T_k timpul între apariția al evenimentului $k-1$ şi apariția evenimentului k.

Analiza timpilor dintre aparițiile succesive ale evenimentelor (sosiri), numiți ***timpi inter-sosire*** este deosebit de importantă.

Astfel vom considera T_1 timpul până la apariția primului eveniment, T_2 timpul între apariția primului eveniment și apariția celui de al doilea eveniment, T_k timpul între apariția al evenimentului $k-1$ și apariția evenimentului k.

Şirul $(T_n)_n$ se numeşte ***şirul timpilor inter-sosire***, iar şirul $(S_n)_n$ definit prin $S_n = T_1 + T_2 + ... + T_n$ se numeşte ***şirul timpilor de aşteptare***.

Un rezultat cunoscut afirmă că timpii inter-sosire corespunzători unui proces Poisson de parametru $\lambda > 0$ sunt variabile aleatoare exponențiale, independente şi identic repartizate având media $\frac{1}{\lambda}$.

De cele mai multe ori ***timpii de service*** au o repartiție exponențială
(notație *M* identică cu cea a sosirilor Poisson).

Repartiția exponențială

Repartiția exponențială este repartiție continuă

Densitatea repartiției exponențiale de parametru λ este

dată de $f_X(x) = \begin{cases} 0, & x < 0 \\ \lambda \cdot e^{-\lambda x}, & x \geq 0 \end{cases}$

având funcția de repartiție corespunzătoare

$$F_X(x) = \begin{cases} 0, & x < 0 \\ 1 - e^{-\lambda x}, & x \geq 0 \end{cases}$$

Media și dispersia corespunzătoare sunt:

$E(X) = \int_0^\infty x\lambda e^{-\lambda x} dx = \frac{1}{\lambda}$ și $D^2(X) = \frac{1}{\lambda^2}$

Analiza timpilor dintre aparițiile succesive ale evenimentelor (sosiri), numiți ***timpi inter-sosire*** este deosebit de importantă.

Astfel vom considera T_1 timpul până la apariția primului eveniment, T_2 timpul între apariția primului eveniment și apariția celui de al doilea eveniment, T_k timpul între apariția al evenimentului $k-1$ și apariția evenimentului k.

Şirul $(T_n)_n$ se numeşte ***şirul timpilor inter-sosire***, iar şirul $(S_n)_n$ definit prin $S_n = T_1 + T_2 + ... + T_n$ se numeşte ***şirul timpilor de aşteptare***.

Un rezultat cunoscut afirmă că timpii inter-sosire corespunzători unui proces Poisson de parametru $\lambda > 0$ sunt variabile aleatoare exponențiale, independente și identic repartizate având media $\frac{1}{\lambda}$.

Repartiția gamma este o repartiție continuă.

O variabilă aleatoare continuă având densitatea dată de

$$f_X(x) = \begin{cases} \dfrac{\lambda^k \cdot x^{k-1} \cdot e^{-\lambda x}}{\int_0^\infty e^{-x} \cdot x^{k-1} dx}, & x \geq 0, \\ 0, & x < 0 \end{cases}$$

, unde $\Gamma(k) = \int_0^\infty e^{-x} \cdot x^{k-1} dx$

este funcția gamma ,se numeşte variabila gamma de parametrii $\lambda > 0$ şi $k > 0$

Media este dată de formula $E(X) = \dfrac{k}{\lambda}$

Se poate arăta că dacă $X_1, X_2, \ldots, X_k$ sunt variabile aleatoare exponențiale, independente şi identic repartizate de parametru λ, atunci variabila sumă $X_1 + X_2 + \ldots + X_k$ va fi o variabilă gamma de parametrii λ şi k.

Exemplu

Sosirile pacienților unui cabinet medical au repartiție Poisson și rata de sosire este de 2 persoane pe oră.

Să calculăm care este timpul mediu până când sosește al 7-lea client: deoarece timpii de așteptare sunt variabile aleatoare având repartiția gamma, avem $E(S_7)=\frac{7}{\lambda}=\frac{7}{2}=3\frac{1}{2}$ ore .

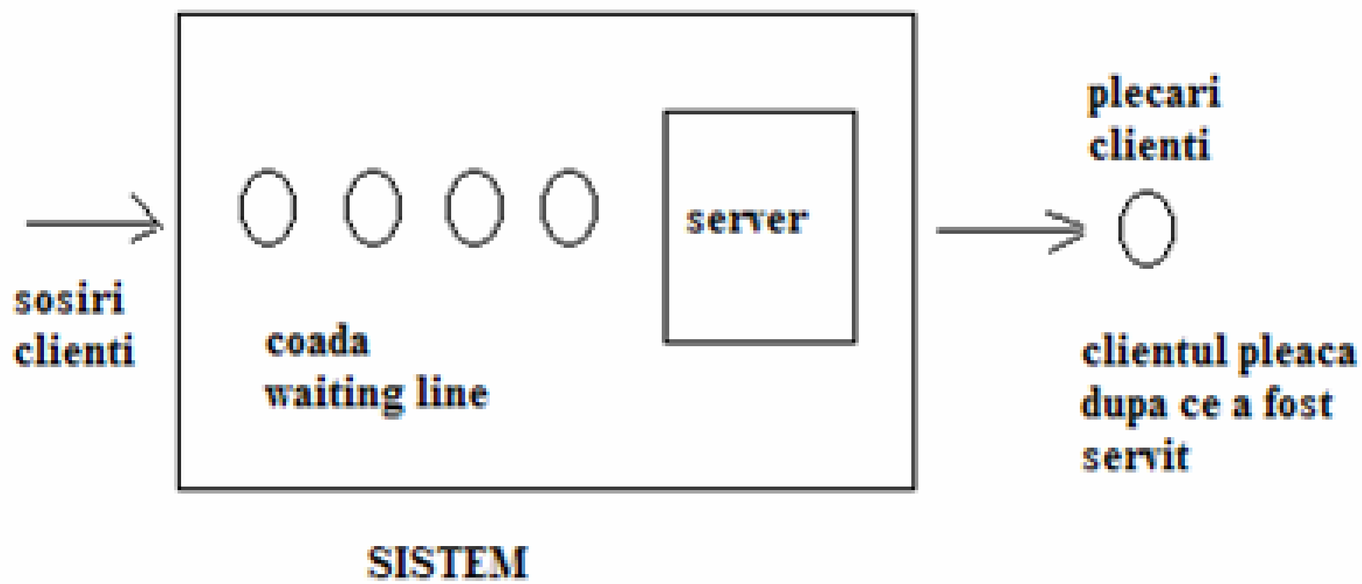
Calculăm probabilitatea ca timpul scurs între sosirea celui e-al 7-lea pacient și sosirea celui de-al 8-lea sa depășească $2\frac{1}{2}$ ore:

$$P(T_8>\frac{5}{2})=1-P(T_8\le\frac{5}{2})=e^{-2\cdot\frac{5}{2}}=e^{-5}=0.0067$$

In cazul *sosirilor Poisson,* timpii inter-sosire sunt independenți și identic repartizați exponențial de densitate $f(t) = \lambda \cdot e^{-\lambda t}, t \geq 0$, unde λ este parametrul sosirilor Poisson, cunoscut sub numele de *rata de sosire*. Reamintim că media sosirilor este $\frac{1}{\lambda}$.

In ceea ce priveşte *timpii de service* , dacă este vorba de un service exponențial *rata de service* va fi μ, în timp ce rata de sosire va fi λ, în cazul unei sosiri Poisson.

diagrama modelului ***M/M/1/FCFS.***



In orice tip de model de aşteptare sosirile sunt întâmplătoare, ceea ce înseamnă că o sosire este independentă de alta şi că nu putem prognoza când are loc o sosire.

Repartiția Poisson dă o bună descriere a pattern-ului sosirilor.

Probabilitatea ca să existe *k* sosiri într-un interval de timp precizat este

$$p_X(k) = P(X = k) = \frac{e^{-\lambda} \cdot \lambda^k}{k!},\ k = 0, 1, 2...$$

unde *k* este numărul sosirilor în intervalul de timp precizat şi λ este media numărului de sosiri în intervalul de timp precizat.

Exemplu

Considerăm că la sucursala unei bănci există o singură casierie pentru persoane fizice. Casierul preia numărul de cont al clientului şi efectuează operațiunea de plată cerută. După plecarea acestui client, va prelua operațiunea de plată a următorului client.
Dacă vor sosi mai mulți clienți se va forma o coadă, „a waiting line".

Analizând datele s-a stabilit că rata sosirilor este de 45 sosiri pe oră. Considerând intervalul de timp precizat ca fiind de un minut, media sosirilor este $\lambda = \frac{45}{60} = 0.75$ /minut.

Probabilitatea ca să existe k sosiri într-un minut este

$$P(X=k)=\frac{e^{-0.75}\cdot 0.75^{k}}{k!}=0.4724\cdot\frac{0.75^{k}}{k!},\ k=0,1,2...$$

şi astfel:

$$P(X=0)=0.4724\cdot\frac{0.75^{0}}{0!}=0.4724$$

$$P(X=1)=0.4724\cdot\frac{0.75^{1}}{1!}=0.3543$$

$$P(X=2)=0.4724\cdot\frac{0.75^{2}}{2!}=0.1329$$

$$P(X=3)=0.4724\cdot\frac{0.75^{3}}{3!}=0.0332$$

În practică se înregistrează numărul de sosiri într-o perioadă de câteva zile sau săptămâni şi se compară funcția de repartiție a sosirilor cu repartiția Poisson, pentru a verifica dacă această descrie distribuția sosirilor.

Timpul de service este intervalul de timp care începe când clientul a dat numărului contului, a spus ce plată bancară doreşte şi se încheie o dată cu primirea şi semnarea chitanței.

Plățile bancare diferă de la client la client, o operație simplă este rezolvată în mai puțin de un minut, în timp ce o operație complexă poate dura mai mult de două minute.

S-a demonstrat că repartiția exponențială dă în general o bună aproximare a timpului de service într-un model de aşteptare. probabilitatea ca durata timpului de service sa fie mai mică egală cu t este:

$$P(\text{timp service} \leq t) = 1 - e^{-\mu t}$$

unde μ este media unităților ce pot fi servite intr-o unitate de timp precizată.

În cazul nostru s-a constatat că sunt serviți 60 clienți pe oră, ceea ce înseamnă că rata medie a service-ului este $\mu = 1$ client/minut.
Să calculăm probabilitatea ca o plată bancară să fie procesată în 30sec, respectiv1 min, respectiv 2 min.

$$P(\text{timp service} \leq 0.5\text{min}) = 1 - e^{-1 \cdot 0.5} = 0.3935$$

$$P(\text{timp service} \leq 1\text{min}) = 1 - e^{-1 \cdot 1} = 0.6321$$

$$P(\text{timp service} \leq 2\text{min}) = 1 - e^{-1 \cdot 2} = 0.8647$$

În practică se studiază datele pentru a stabili dacă repartiția exponențială descrie timpii de service.

O proprietate a repartiției exponențiale este că variabila aleatoare ia o valoare mai mică egală cu media cu o probabilitate de 0.6321.
În aplicațiile teoriei așteptării repartiția exponențială indică că 63% din timpii de service sunt mai mici decât media.

Dimineața, la prima oră când se deschide banca nu avem clienți.
Gradual activitatea se amplifică, până la stadiul normal de desfăşurare, ***stadiu de echilibru statistic, „steady state"***,

Această perioada de timp de la deschidere până la atingerea stadiului normal se numeşte ***perioada de tranziție.***

Modelele de aşteptare descriu caracteristicile numerice ale stadiului de echilibru statistic

Caracteristicile numerice ale modelului M/M/1//FCFS

Vom studia modelul de aşteptare cu un singur server, sosiri Poisson, timp de service exponențial, disciplina cozii fiind primul venit, primul servit, calculând caracteristicile numerice ale acestuia.

Repartiția timpului inter-sosire este exponențială de parametru λ,

$F(t) = 1 - e^{-\lambda t}, t \geq 0$,

iar service-ul este repartizat exponențial, de parametru μ

$g(t) = 1 - e^{-\mu t}, t \geq 0$.

Astfel rata de sosire este $\lambda_k = \lambda, k = 0, 1, 2, \ldots$, rata de service este $\mu_{k+1} = \mu, k = 0, 1, 2, \ldots$.

Este necesar ca media timpului inter-sosire să depăşească media timpului de service, adică $\frac{\lambda}{\mu} < 1$, altfel coada devine infinită.

Considerând procesul în stare de echilibru statistic, notăm cu p_j probabilitatea ca să existe j clienți în sistem ($j-1$ la rând şi unul la service) avem

$$p_j = \left(1 - \frac{\lambda}{\mu}\right) \cdot \left(\frac{\lambda}{\mu}\right)^j \,,\; j = 0, 1, 2, \ldots$$

Caracteristici numerice ale modelului

- probabilitatea ca să nu existe clienți în sistem este $P_0 = 1 - \frac{\lambda}{\mu}$
- numărul mediu de clienți la rând (waiting line) este $L_q = \frac{\lambda^2}{\mu \cdot (\mu - \lambda)}$
- numărul mediu de clienți în sistem este $L = L_q + \frac{\lambda}{\mu}$
- timpul mediu de aşteptare (timpul pe care un client îl petrece la rând)

este $W_q = \frac{L_q}{\lambda}$

- timpul mediu pe care un client îl petrece în sistem este $W = W_q + \frac{1}{\mu}$
- probabilitatea ca service-ul să fie ocupat (probabilitatea ca un client sosit să aştepte pentru a fi servit) este $P_w = \frac{\lambda}{\mu}$, raportul $\frac{\lambda}{\mu}$ fiind cunoscut sub numele de factor de utilizare a service-ului
- probabilitatea să fie *n* clienți în sistem este $P_n = \left(\frac{\lambda}{\mu}\right)^n \cdot P_0$

Dacă $\frac{\lambda}{\mu} \geq 1$, coada va creşte nemărginit în timp ce service-ul nu va avea capacitatea de a servi clienții noi sosiți.

Exemplu

La sucursala unei bănci există o singură casierie pentru persoane fizice. Rata sosirilor este $\lambda = 0.75$ clienți pe minut în timp ce rata de service este $\mu = 1$, clienți pe minut, așadar $\frac{\lambda}{\mu} < 1$.

Caracteristicile numerice ale modelului:

- probabilitatea ca să nu existe clienți în sistem este

$$P_0 = 1 - \frac{\lambda}{\mu} = 1 - \frac{0.75}{1} = 0.25$$

- numărul mediu de clienți la rând este

$$L_q = \frac{\lambda^2}{\mu \cdot (\mu - \lambda)} = \frac{0.75^2}{1 \cdot (1 - 0.75)} = 2.25 \text{ clienți}$$

- numărul mediu de clienți în sistem este

$$L = L_q + \frac{\lambda}{\mu} = 2.25 + \frac{0.75}{1} = 3 \text{ clienți}$$

- timpul mediu de aşteptare este

$$W_q = \frac{L_q}{\lambda} = \frac{2.25}{0.75} = 3 \text{ minute}$$

- timpul mediu petrecut de un client în sistem este

$$W = W_q + \frac{1}{\mu} = 3 + \frac{1}{1} = 4 \text{ minute}$$

- probabilitatea de aşteptare este $P_w = \frac{\lambda}{\mu} = \frac{0.75}{1} = 0.75$,

- probabilitatea ca în sistem să fie *n* clienți în sistem este:

numărul de clienți în sistem	probabilitatea ca *n* clienți să fie în sistem $P_n = \left(\frac{\lambda}{\mu}\right)^n \cdot P_0$
0	0.2500
1	0.1875
2	0.1406
3	0..1055
4	0.0791
5	0.0593
6	0.0445
≥ 7	$0.25 \cdot \sum_{n=7}^{\infty} (0.75)^n = 0.25 \cdot (0.75)^7 \cdot \frac{1}{1-0.75} = 0.1$

Probabilitatea de 0.1335 ca să avem 7 sau mai mulți clienți în sistem indică faptul că deseori vor fi cozi mari la această casierie.
Se pot face modificări prin folosirea unei noi tehnologii pentru creşterea ratei service-ului sau prin adăugarea unui nou server, astfel încât să fie serviți mai mulți clienți în perioada de timp considerată.

Dacă prin modificările făcute vor fi serviți 75 clienți pe oră, înseamnă că rata service-ului revizuit este $\mu = \frac{75}{60} = 1.25$ clienți pe minut.
In aceste condiții $\lambda = 0.75$ clienți pe minut şi $\mu = 1.25$ clienți pe minut .

probabilitatea ca serverul să fie liber	0.400
numărul mediu de clienți la rând	0.900
numărul mediu de clienți în sistem	1.500
timpul mediu de aşteptare pentru un client	1.200 min
timpul mediu petrecut de un client în sistem	2.000 min
probabilitatea de aşteptare	0.600
probabilitatea ca în sistem să fie *n* clienți în sistem	0.028

Să reținem:

Presupunerea că repartiția sosirilor este Poisson este echivalentă cu presupunerea ca timpii inter-sosiri au o repartiție exponențială.

De exemplu dacă sosirile într-o linie de așteptare sunt repartizate Poisson cu o medie de 20 de sosiri pe oră, timpii inter-sosiri vor avea o repartiție exponențiala cu o medie de $\frac{1}{20} = 0.05$ ore.

Exemple

1. *Camioanele au un singur punct de încărcare în docuri;* sosesc la acest punct după o repartiție Poisson, având rata de sosire de 12 camioane pe zi. Timpii ceruți pentru încărcarea /descărcarea unui camion au o repartiție exponențială, cu media timpului de service de 18 camioane pe zi.

- probabilitatea să nu existe camioane în sistem

$$P_0 = 1 - \frac{\lambda}{\mu} = 1 - \frac{12}{18} = 0.333$$

- numărul mediu de camioane ce aşteaptă la coadă

$$L_q = \frac{\lambda^2}{\mu \cdot (\mu - \lambda)} = \frac{12^2}{18 \cdot (18 - 12)} = 1.333 \text{ clienți}$$

- timpul mediu de aşteptare pentru începerea acțiunii de încărcare /descărcare, al unui camion

$$W_q = \frac{L_q}{\lambda} = \frac{1.333}{12} = 0.1111 \text{ ore} = 6.67 \text{ minute}$$

- probabilitatea ca un nou sosit să nu aştepte

$$P_w = \frac{\lambda}{\mu} = \frac{12}{18} = 0.6667,$$

2. O firmă de decorațiuni interioare oferă și consultanță pentru clienții săi printr-un specialist în design de interioare, care răspunde la întrebări și dă sfaturi.

Media sosirii clienților la acest birou este de 2.5 clienți pe oră.

Consultantul alocă în medie 10 minute fiecărui client.

Presupunem că sosirile au repartiție Poisson și timpii de service au repartiție exponențială caracteristicile de operare ale acestei linii de aşteptare:

Avem $\lambda = 2.5$ şi $\mu = \frac{60}{10} = 6$ clienți pe oră;

probabilitatea ca să nu existe clienți în sistem

$$P_0 = 1 - \frac{\lambda}{\mu} = 1 - \frac{2.5}{6} = 0.5833;$$

numărul mediu de clienți de la coadă

$$L_q = \frac{\lambda^2}{\mu\cdot(\mu-\lambda)} = \frac{2.5^2}{6\cdot(6-2.5)} = 0.2976 \text{ clienți;}$$

numărul mediu de clienți în sistem

$$L = L_q + \frac{\lambda}{\mu} = 0.7143 \text{ clienți;}$$

μ

timpul mediu de aşteptare pentru un client

$$W_q = \frac{L_q}{\lambda} = 0.1190 \text{ ore} = 7.14 \text{minute};$$

timpul mediu petrecut de un client în sistem

$$W = W_q + \frac{1}{\mu} = 0.2857 \text{ ore} = 17.14 \text{ minute};$$

probabilitatea ca specialistul să fie ocupat este

$$P_w = \frac{\lambda}{\mu} = \frac{2.5}{6} = 0.4167 \,,$$

- Scopul managerului este ca un client să nu aştepte mai mult de 5 minute, ceea ce nu este posibil deoarece $W_q = 7.14$ minute.

Soluțiile ar fi creşterea ratei de service μ sau angajarea celui de-al doilea specialist în design interior.

- În cazul în care consultantul reduce timpul de service la 8 minute problema expusă va fi rezolvată deoarece avem

$\lambda = 2.5$ și $\mu = \dfrac{60}{8} = 7.5$ clienți pe oră;

numărul mediu de clienți din rând este

$$L_q = \frac{\lambda^2}{\mu \cdot (\mu - \lambda)} = \frac{2.5^2}{7.5 \cdot (7.5 - 2.5)} = 0.1667 \text{ clienți;}$$

timpul mediu de așteptare pentru un client

$$W_q = \frac{L_q}{\lambda} = 0.0667 \text{ ore} = 3.96 \text{ minute;}$$

numărul mediu de clienți în sistem

$$L = L_q + \frac{\lambda}{\mu} = 0.5 \text{ clienți.}$$

3. Firma T ce vinde anvelope a hotărât să angajeze un mecanic care să schimbe anvelopele clienților ce cumpără un set nou de la această firmă. La concurs s-au prezentat doi mecanici:

- X, fără experiență, cere un salariu de 14$ pe oră şi poate servi în medie 3 clienți pe oră
- Y, cu câțiva ani de experiență cere un salariu de 20$ pe oră şi serveşte în medie 4 clienți pe oră

Presupunând că la acest punct de lucru doi clienți sosesc în medie pe oră, că sosirile sunt Poisson și timpii de service sunt exponențiali să calculăm caracteristicile acestei linii de așteptare pentru fiecare mecanic

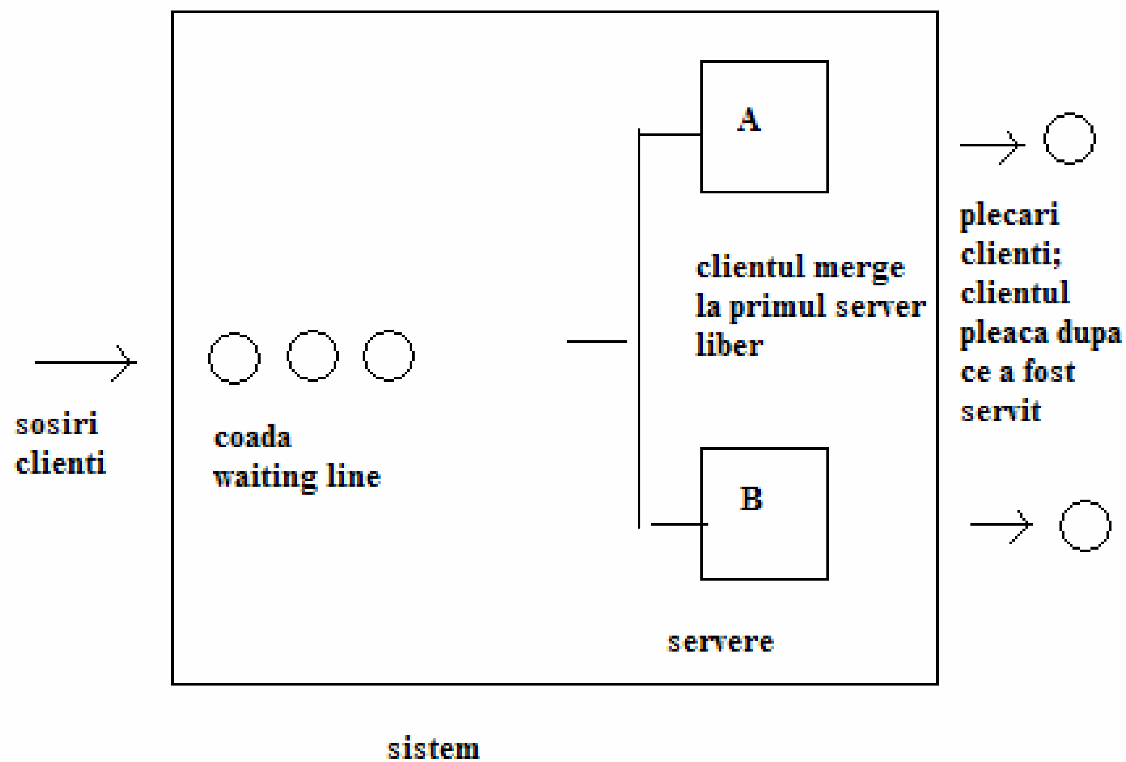
$\lambda = 2$	$\mu = 3$	$\mu = 4$
numărul mediu de clienți din rând (L_q)	1.3333	0.5000
numărul mediu de clienți în sistem (L)	2.0000	1.0000
timpul mediu de așteptare pentru un client (W_q)	0.6667	0.2500
timpul mediu petrecut de un client în sistem (W)	1.000	0.5000
probabilitatea de așteptare este P_w	0.6667	0.5000

M/ M/k/ FCFS

Vom studia un model de aşteptare multi-server, un singur rând, cu sosiri Poisson şi timp de service exponenţial, rata service-ului fiind aceeşi pentru toate serverele.
În acest model clienţii care aşteaptă într-un singur rând se duc la primul server care este liber.

Exemplu

modelul casieriei unice, pentru persoane fizice din sucursala unei bănci, poate fi modificat prin deschiderea unui al doilea ghişeu bancar:



caracteristicile numerice ale modelului *M/M/k/FCFS*

avem k servere, sosirile au o repartiție Poisson, rata medie a sosirilor fiind λ, service-ul este repartizat exponențial, rata medie de service fiind aceeași μ pentru fiecare server.

Raportul $\frac{\lambda}{\mu}$ este măsura densității traficului, numită ***erlang***, după numele unuia dintre întemeietorii teoriei așteptării A.K Erlang. Această măsură indică numărul minim de servere necesar pentru a gestiona sosirile astfel încât să nu se piardă clienți.

- Probabilitatea ca să nu existe clienți în sistem este:

$$P_0 = \frac{1}{\sum_{n=0}^{k-1} \frac{(\lambda/\mu)^n}{n!} + \frac{(\lambda/\mu)^k}{k!} \cdot \left(\frac{k\mu}{k\mu - \lambda} \right)};$$

- Numărul mediu de clienți din rând este $L_q = \frac{(\lambda/\mu)^k \lambda\mu}{(k-1)!(k\mu - \lambda)^2} \cdot P_0$
- Numărul mediu de clienți în sistem este $L = L_q + \frac{\lambda}{\mu}$
- Timpul mediu pe care un client îl petrece la coadă este $W_q = \frac{L_q}{\lambda}$
- Timpul mediu pe care un client îl petrece în sistem este $W = W_q + \frac{1}{\mu}$

- Probabilitatea de aşteptare este $P_w = \frac{1}{k!} \cdot \left(\frac{\lambda}{\mu}\right)^k \cdot \left(\frac{k\mu}{k\mu - \lambda}\right) \cdot P_0$

- Probabilitatea să fie *n* clienți în sistem este $P_n = \begin{cases} \frac{(\lambda / \mu)^n}{n!} \cdot P_0,\ n \le k \\ \frac{(\lambda / \mu)^n}{k!\, k^{(n-k)}} \cdot P_0,\ n > k \end{cases}$

Exemplu

Caracteristicile numerice în cazul exemplului cu două ghişee bancare-casierii, pentru persoane fizice la sucursala unei bănci. Rata sosirilor este $\lambda = 0.75$ clienți pe minut, rata de service este $\mu = 1$ client pe minut.

- Probabilitatea ca să nu existe clienți în sistem este:

$$P_0 = \frac{1}{\frac{(\lambda/\mu)^0}{0!} + \frac{(\lambda/\mu)^1}{1!} + \frac{(\lambda/\mu)^2}{2!} \cdot \left(\frac{2\mu}{2\mu - \lambda}\right)} =$$

$$= \frac{1}{\frac{(0.75)^0}{0!} + \frac{(0.75)^1}{1!} + \frac{(0.75)^2}{2!} \cdot \left(\frac{2}{2 - 0.75}\right)} = 0.4545$$

- Numărul mediu de clienți din rând: $L_q = \frac{(0.75)^2 \cdot 0.75}{1!(2-0.75)^2} \cdot 0.4545 = 0.1227$
- Numărul mediu de clienți în sistem: $L = 0.1227 + 0.75 = 0.8727$
- Timpul mediu pe care un client îl petrece la coadă $W_q = \frac{0.1227}{0.75} = 0.16\,\text{min.}$
- Timpul mediu pe care un client îl petrece în sistem $W = 0.16 + 1 = 1.16\,\text{min.}$
- Probabilitatea de așteptare este $P_w = \frac{1}{2!} \cdot (0.75)^2 \cdot \left(\frac{2 \cdot 1}{2 \cdot 1 - 0.75}\right) \cdot 0.4545 = 0.2045$

- Probabilitatea să fie *n* clienți în sistem:

număr de clienți	probabilitatea
0	0.4545
1	0.3409
2	0.1278
3	0.0479
4	0.0180
≥ 5	0.0109

Să comparăm caracteristicile numerice ale modelului cu 2 servere cu cele ale modelului cu un server:

- timpul mediu pe care clientul îl petrece în sistem (timpul
- mediu de sta la rând plus timpul mediu de service)
- s-a redus de la $W = 4$ minute .la $W = 1.16$ minute
- numărul mediu de clienți la coadă s-a redus de la $L_q = 2.25$ clienți

la $L_q = 0.1227$ clienți

- timpul mediu pe care un client îl petrece la coadă s-a redus de la $W_q = 3$ minute la $W_q = 0.16$ minute.
- probabilitatea ca un client să aştepte s-a redus de la $P_w = 0.75$, adică 75%, la $P_w = 0.2045$, adică 20.45%.

Modelul multi-server prezentat are un singur rând.
S-a demonstrat ştiinţific că în acest caz caracteristicile operaţionale sunt mai bune decât dacă fiecare server avea propriul rând.
Pe de altă parte cazul prezentat nu mai apar frustrări de genul:
„ X a venit mai târziu, dar stând la alt rând a a fost servit mai repede”.
Astfel în bănci, în aeroporturi, în fast-food-uri se utilizează de obicei modelul multi-server cu un singur rând (waiting-line).

Exemplu

Un local de tip fast-food are un serviciu de „drive -up window”. Sosirile clienților sunt Poisson, cu media de 24 de mașini pe oră și timpii de service au o repartiție exponențială.

Clienții dau comanda la o stație intercom, la intrarea în parcare și conduc spre fereastra de servire unde plătesc și primesc mâncarea.

Vom lua în considerare trei variante de servicii:

- A: varianta cu un singur rând, în care acelaşi angajat pregăteşte comanda şi încasează banii, caz în care timpul mediu de service este de 2 minute.
- B: varianta cu un singur rând, în care un angajat pregăteşte comanda şi altul încasează banii, caz în care timpul mediu de service este de 1.25 minute.
- C: o variantă cu două servere, la fiecare fereastră existând un singur angajat care pregăteşte comanda şi încasează banii, caz în care timpul mediu de service pentru fiecare server este de 2 minute.

caracteristica numarică	A: $k=1, \lambda=24, \mu=30$	B: $k=1, \lambda=24, \mu=48$	C: $k=2, \lambda=24, \mu=30$
P_0	0.2000	0.5000	0.4286
L_q	3.2000	0.5000	0.1524
L	4.0000	1.0000	0.0524
W_q	0.1333	0.0200	0.0063
W	0.1667	0.0417	0.0397
P_w	0.8000	0.5000	0.2286

Analiza economică a problemei

Din tabel reiese că cel mai bun service este asigurat de al sistemul C.

Ținând seama că:

- timpul de aşteptare al unui client este evaluat la $c_0 = 25\$$ pe oră, ceea ce înseamnă că timpul de aşteptare este un cost pentru fast-food;
- fiecare angajat este plătit cu $6.50\$$ pe oră;
- echipamentele şi spațiul de lucru ale unui server costă firma $20\$$ pe oră

să hotărâm care este arhitectura variantei ce are cel mai mic cost:

Calculăm c_1, costul service-ului pe oră pentru fiecare server:

A: $c_1 = 6.50 + 20 = 26.50\$$

B: $c_1 = 2 \cdot 6.50 + 20 = 33.00\$$

C: $c_1 = 6.50 + 20 = 26.50\$$

Calculăm costul total: $TC = c_0 \cdot L + c_1 \cdot k$

A: $TC_A = 25 \cdot 4 + 26.50 \cdot 1 = 126.50\$$

B: $TC_B = 25 \cdot 1 + 33 \cdot 1 = 58\$$

C: $TC_C = 25 \cdot 0.9524 + 26.50 \cdot 2 = 76.81\$$

Sistemul B este cel mai economic.

Ecuațiile lui Little,

$$L = \lambda \cdot W$$

$$L_q = \lambda \cdot W_q$$

Importanța ecuațiilor lui Little rezultă din faptul că sunt aplicabile oricărui model, chiar dacă sosirile nu sunt Poisson sau timpul de service nu este exponențial.

Din ecuațiile lui Little rezultă că $W_q = \dfrac{L_q}{\lambda}$, ecuație pe care am folosit-o deja în cazul modelelor de așteptare prezentate.

O relație naturală, ce se aplică în toate modelele de așteptare spune că timpul mediu petrecut în sistem este suma timpului mediu petrecut la rând și a timpului mediu de service.

Dacă rata medie a service-ului este μ , avem:

$$W = W_q + \frac{1}{\mu}$$

Exemple

1. S-a calculat că la raionul de cofetărie al unui super-market sosirile clienților sunt Poisson, cu media de 24 de sosiri pe oră, adică $\lambda = \frac{24}{60} = 0.40$ clienți pe minut, în timp ce timpii de service au o repartiție normală, timpul mediu de service fiind de 30 clienți pe oră, adică $\mu = \frac{30}{60} = 0.5$ clienți pe minut.

S-a observat că un client petrece în medie 4.5 minute în sistem, deci $W = 4.5$ minute.

caracteristicile numerice ale acestui model:

- timpul mediu pe care un client îl petrece la coadă este:

$$W_q = W - \frac{1}{\mu} = 4.5 - \frac{1}{0.5} = 2.5 \text{ minute},$$

- numărul mediu de clienți în sistem este $L = \lambda \cdot W = 0.4 \cdot 4.5 = 1.8$ clienți
- numărul mediu de clienți de la rând este $L_q = \lambda \cdot W_q = 0.4 \cdot 2.5 = 1$ client.

2. Pe un stadion, la un fast-food cu mai multe puncte de servire, fiecare având rândul propriu, (model multi-channel); s-a calculat că timpul petrecut de un client în sistem (de la sosire până la plecarea cu comanda rezolvată) este de 10 minute.

În timpul meciului, rata medie a sosirilor este de 4 clienți pe minut.

Timpul mediu de rezolvare a unei comenzi/client este de 2 minute.

În concluzie $\lambda = 4$, $W = 10$ minute.

- Rata medie a service-ului pe server în termeni de clienți /minut

$$\mu = \frac{1}{2} = 0.5 .$$

- Timpul mediu pe care un client îl petrece la coadă, până a da comanda

$$W_q = W - \frac{1}{\mu} = 10 - \frac{1}{0.5} = 8 \text{ minute.}$$

- Numărul mediu de clienți în sistem: $L = \lambda \cdot W = 4 \cdot 10 = 40$

Pentru a face o analiză economică a modelului de aşteptare este necesar studiul unui model de cost total care include costul aşteptării şi costul service-ului efectiv. Avem:

$$TC = c_0 \cdot L + c_k \cdot k$$

unde

- TC costul total pentru o anumită perioadă de timp,
- c_0 costul aşteptării pe aceea perioadă de timp pentru fiecare client,
- L numărul mediu de clienți în sistem,
- c_k costul service-ului pentru fiecare server în acea perioadă de timp,
- k numărul de servere.

Estimarea lui c_k este relativ simplă, dar estimarea lui c_0 este complicată, fiind de cele mai multe ori subiectivă.
Dacă nu se ține seama de acest cost al aşteptării şi se permite formarea de cozi lungi se vor pierde clienți, ceea ce înseamnă diminuarea vânzărilor.
Adaugarea de servere duce la:

- îmbunătățirea caracteristicilor numerice ale modelului
- reducerea costului aşteptării
- creşterea costului service-ului.

Analiza economică se face în scopul găsirii numărului optim de servere care vor minimiza costul total.

Exemplu

- Reluăm exemplul cu firma de decorațiuni interioare ce oferă și consultanță pentru clienții săi, printr-un specialist în design de interioare.

Media sosirii clienților la acest birou este de 2.5 clienți pe oră.

Consultantul alocă în medie 10 minute fiecărui client.

Se presupune că sosirile au repartiție Poisson și timpii de service au repartiție exponențială. Avem două variante:

1. Folosim un consultant ce alocă în medie 8 minute fiecărui client
2. Folosim doi consultanți ce alocă în medie 10 minute fiecărui client.

În ipoteza că salariul pe oră al consultanților este de 10$ şi ora de aşteptare pentru un client este evaluată la 25$, să evaluăm din punct de vedere economic două variante.

1. Timpul de service fiind de 8 minute avem $\mu = \frac{60}{8} = 7.5$

$$L_q = \frac{\lambda^2}{\mu \cdot (\mu - \lambda)} = \frac{(2.5)^2}{7.5 \cdot (7.5 - 2.5)} = 0.1667 \text{ clienți}$$

$$L = L_q + \frac{\lambda}{\mu} = 0.50 \text{ clienți}$$

$$TC = 25 \cdot L + 16 = 25 \cdot 0.50 + 16 = 28.50 \; \$$$

2. Două severe, $\lambda = 2.5,\ \mu = \frac{60}{10} = 6$

$$P_0 = 1 - \frac{\lambda}{\mu} = 0.6552$$

$$L_q = \frac{\left(\frac{\lambda}{\mu}\right)^2 \cdot \lambda \cdot \mu}{1! \cdot (2\mu - \lambda)^2} \cdot P_0 = 0.0189 \text{ clienți}$$

$$L = L_q + \frac{\lambda}{\mu} = 0.4356 \text{ clienți}$$

$$TC = 25 \cdot 0.4356 + 2 \cdot 16 = 42.89\$$$

Concluzia este că este preferabilă utilizarea unui consultant ce alocă în medie 8 minute fiecărui client.

$M/G/1$.

Vom studia un model cu un singur server, cu sosiri Poisson, cu timpi de service ce au o repartiție arbitrară generală $G(x)$, de medie $\frac{1}{\mu}$, adică un model $M/G/1$.

În calculul caracteristicilor numerice vom folosi următoarele notații:

i. λ - rata medie a sosirilor
ii. μ - rata medie a service-ului
iii. $\frac{1}{\mu}$ - timpul mediu de service
iv. σ - deviația standard a timpului de service.

formule de calcul:

- Probabilitatea ca să nu existe clienți în sistem: $P_0 = 1 - \frac{\lambda}{\mu}$;
- Numărul mediu de clienți din rând este $L_q = \frac{\lambda^2 \cdot \sigma^2 + \left(\frac{\lambda}{\mu}\right)^2}{2\left(1 - \frac{\lambda}{\mu}\right)}$
- Numărul mediu de clienți în sistem este $L = L_q + \frac{\lambda}{\mu}$

- Timpul mediu pe care un client îl petrece la coadă este $W_q = \dfrac{L_q}{\lambda}$
- Timpul mediu pe care un client îl petrece în sistem este $W = W_q + \dfrac{1}{\mu}$
- Probabilitatea de aşteptare este $P_w = \dfrac{\lambda}{\mu}$

Exemple

1. La un supermarket, la raionul de Peşte şi Fructe de mare este un singur vânzător. Clienții sosesc aleator, în medie 21 clienți pe oră, adică $\lambda = \frac{21}{60} = 0.35$ clienți pe minut. Timpul de service este în medie de 2 minute cu o deviație standard $\sigma = 1.2$ minute. Vânzătorul are rata medie de service $\mu = \frac{1}{2} = 0.5$ clienți pe minut.

caracteristici numerice:

- Probabilitatea ca să nu existe clienți în sistem

$$P_0 = 1 - \frac{\lambda}{\mu} = 1 - \frac{0.35}{0.60} = 0.30;$$

- Numărul mediu de clienți din rând:

$$L_q = \frac{\lambda^2 \cdot \sigma^2 + \left(\frac{\lambda}{\mu}\right)^2}{2\left(1 - \frac{\lambda}{\mu}\right)} = \frac{0.35^2 \cdot 1.2^2 + \left(\frac{0.35}{0.50}\right)^2}{2\left(1 - \frac{0.35}{0.50}\right)} = 1.11 \text{ clienți}$$

- Numărul mediu de clienți în sistem

$$L = L_q + \frac{\lambda}{\mu} = 1.11 + \frac{0.35}{0.50} = 1.81 \text{ clienți}$$

- Timpul mediu pe care un client îl petrece la coadă

$$W_q = \frac{L_q}{\lambda} = \frac{1.1107}{0.35} = 3.17 \text{ minute}$$

- Timpul mediu pe care un client îl petrece în sistem

$$W = W_q + \frac{1}{\mu} = 3.17 + \frac{1}{0.5} = 5.17 \text{ minute}$$

- Probabilitatea de aşteptare: $P_w = \frac{\lambda}{\mu} = \frac{0.35}{0.50} = 0.70$

2. Firma X se ocupă de mici reparații în construcții. Sosirea comenzilor are o repartiție Poisson cu o rată medie de 2 comenzi la ziua de lucru de 8 ore. Timpii de efectuare a reparațiilor au o repartiție normală, timpul mediu necesar pentru efectuarea unei reparații fiind de 3.2 ore și deviația standard de 2 ore.

- rata medie de sosire a comenzilor este de $\lambda = \frac{2}{8} = 0.25$ comenzi pe oră
- rata medie a reparațiilor (a service-ului) este $\mu = \frac{1}{3.2} = 0.3125$ reparații

pe oră

- numărul mediu de comenzi - reparații în așteptare este

$$L_q = \frac{\lambda^2 \cdot \sigma^2 + \left(\frac{\lambda}{\mu}\right)^2}{2\left(1-\frac{\lambda}{\mu}\right)} = \frac{0.25^2 \cdot 2^2 + \left(\frac{0.25}{0.3125}\right)^2}{2\left(1-\frac{0.25}{0.3125}\right)} = 2.225$$

- timpul mediu de așteptare până când se începe reparația (timpul mediu de așteptare al comenzii) este $W_q = \frac{L_q}{\lambda} = \frac{2.225}{0.25} = 8.9$ ore

- timpul mediu de așteptare din clipa în care s-a dat comanda până s-a efectuat reparația este $W = W_q + \frac{1}{\mu} = 8.9 + \frac{1}{0.3125} = 12.1$ ore

- probabilitatea ca firma să aibă de lucru și comanda să stea în așteptare este $P_w = \frac{\lambda}{\mu} = \frac{0.25}{0.3125} = 0.80$,

ceea ce înseamnă că 80% din timp firma are de lucru.

- numărul mediu de comenzi - reparații în așteptare este

$$L_q = \frac{\lambda^2 \cdot \sigma^2 + \left(\frac{\lambda}{\mu}\right)^2}{2\left(1 - \frac{\lambda}{\mu}\right)} = \frac{0.25^2 \cdot 2^2 + \left(\frac{0.25}{0.3125}\right)^2}{2\left(1 - \frac{0.25}{0.3125}\right)} = 2.225$$

- timpul mediu de așteptare până când se începe reparația (timpul mediu de așteptare al comenzii) este $W_q = \frac{L_q}{\lambda} = \frac{2.225}{0.25} = 8.9$ ore

- timpul mediu de așteptare din clipa în care s-a dat comanda până s-a efectuat reparația este $W = W_q + \frac{1}{\mu} = 8.9 + \frac{1}{0.3125} = 12.1$ ore

- probabilitatea ca firma să aibă de lucru și comanda să stea în așteptare este $P_w = \frac{\lambda}{\mu} = \frac{0.25}{0.3125} = 0.80$,

ceea ce înseamnă că 80% din timp firma are de lucru.

Dacă managerul nu este mulțumit de caracteristicile numerice ale modelului, încearcă să le îmbunătățească, de cele mai multe ori **crescând rata medie a service-ului** μ.

Numărul mediu de clienți la rând este influențat de variația timpilor de service, adică de deviația standard a timpului de service. **Micşorarea variațiilor timpilor de service** este o alternativă pentru scăderea numărul mediu de clienți la rând, îmbunătățind şi celelalte caracteristici numerice ale modelului.

M/D/1

Este un model de aşteptare în care sosirile sunt aleatoare, de repartiție Poisson şi timpul de service este constant, situație care apare în producția manufacturieră, când timpul de service al maşinilor este constant.

Singura modificare a caracteristicilor numerice ale modelului este in cazul numărului mediu de clienți din rând, deoarece deviația standard a timpului de service. este nulă, $\sigma = 0$

$$L_q = \frac{\left(\frac{\lambda}{\mu}\right)^2}{2\left(1-\frac{\lambda}{\mu}\right)}$$

Model de tip *BCC*- (blocked customers cleared).

Este un model în care nu sunt permise cozile. Clienții care sosesc sunt serviți la unul sau mai multe servere.
Dacă toate serverele sunt ocupate este interzis accesul clienților în sistem, ceea ce în terminologia teoriei aşteptării înseamnă dacă sistemul este complet, sosirile sunt blocate şi eliminate din sistem.
Clienții ce nu sunt admişi în sistem renunță, ceea ce înseamna că sunt pierduți de către sistem, sau vin altă dată.

$M/G/k$, de tip *BCC*

Ne situăm în următoarele ipoteze:

- sistemul are k servere
- sosirile sunt un proces Poisson de rată λ
- timpii de service al fiecărui server pot avea o repartiție probabilistă oarecare
- rata medie a service-ului este aceeași pentru toate serverele, μ
- un client nou sosit intră în sistem dacă cel puțin unul din cele k servere este liber; este interzis accesul clienților în sistem dacă toate serverele sunt ocupate.

Întrebare: de câte servere este nevoie.

O aplicație a acestui model este în sistemele de comunicație, în care apelurile sunt sosirile iar severele sunt reprezentate de liniile de comunicație. Dacă serverele sunt ocupate apelul primeşte un semnal de „ocupat” şi nu îi este permis accesul în sistem.

Probabilitatea ca exact *j* din cele *k* servere să fie ocupate este:

$$P_j = \frac{\frac{\lambda}{\mu} \cdot \frac{1}{j!}}{\sum_{i=0}^{k} \left(\frac{\lambda}{\mu}\right)^i \cdot \frac{1}{i!}}$$

- λ este rata medie a sosirilor,
- μ este rata medie a service-ului pentru fiecare server,
- *k* este numărul serverelor din sistem
- P_j este probabilitatea ca exact *j* din cele *k* servere să fie ocupate, $j = 0, 1, 2, \ldots k$.

Probabilitatea P_k , probabilitatea ca cele k servere să fie ocupate, adică procentul de sosiri blocate, a căror intrare în sistem este interzisă, este cea mai importantă caracteristică.

Cum este influențată probabilitatea P_k de numărul de servere?

Altă caracteristică ce prezintă interes este numărul mediu de clienți în sistem, ceea ce poate fi interpretat şi ca numărul mediu de servere ce lucrează.:

$$L = \frac{\lambda}{\mu} \cdot \left(1 - P_k\right).$$

În acest caz caracteristicile numerice L_q şi W_q sunt nule, deoarece în acest sistem nu este permisă aşteptarea în coadă.

Exemple

1. Firma Y utilizează comanda prin telefon a produselor sale, clienții apelând în acest sens numărul 800.
Rata medie a apelurilor telefonice este 12 apeluri pe oră.
Timpul necesar procesării comenzii diferă de la o comandă la alta, dar în general sunt rezolvabile 6 comenzi pe oră.
Telefonul cu numărul 800 are 3 linii interne, la fiecare dintre acestea lucrând un vânzător.
Apelurile primite sunt transferate automat liniei ce este liberă.
Dacă toate liniile sunt ocupate, clientul va primi semnalul de ocupat.
Foarte puțini dintre clienții refuzați astfel, revin mai târziu.

Astfel, managerul firmei este interesat de câte linii interne este nevoie pentru ca să fie rezolvabile comenzile a 90% din posibilii clienți.

Calculăm procentul de clienți refuzați dacă ar exista 3 linii telefonice interne:

$$P_3 = \frac{\left(\frac{12}{6}\right)^3 \cdot \frac{1}{3!}}{\left(\frac{12}{6}\right)^0 \cdot \frac{1}{0!} + \left(\frac{12}{6}\right)^1 \cdot \frac{1}{1!} + \left(\frac{12}{6}\right)^2 \cdot \frac{1}{2!} + \left(\frac{12}{6}\right)^3 \cdot \frac{1}{3!}} = \frac{1.3333}{6.3333} = 0.2105$$

ceea ce înseamnă că 21% din clienți sunt refuzați.

Calculăm procentul de clienți refuzați dacă ar exista 4 linii telefonice interne:

$$P_4 = \frac{\left(\frac{12}{6}\right)^4 \cdot \frac{1}{4!}}{\left(\frac{12}{6}\right)^0 \cdot \frac{1}{0!} + \left(\frac{12}{6}\right)^1 \cdot \frac{1}{1!} + \left(\frac{12}{6}\right)^2 \cdot \frac{1}{2!} + \left(\frac{12}{6}\right)^3 \cdot \frac{1}{3!} + \left(\frac{12}{6}\right)^4 \cdot \frac{1}{4!}} = \frac{0.6667}{7} = 0.0952$$

În acest caz fiind blocați 9% dintre clienți, avem 91% clienți a căror comandă va fi primită.

Cu patru servere target-ul managerului este îndeplinit.

Să calculăm numărul mediu de apeluri, ceea ce înseamnă de fapt gradul de utilizare a celor patru servere

$$L = \frac{\lambda}{\mu} \cdot (1 - P_k) = \frac{12}{6}(1 - 0.0952) = 1.81$$

Prezentăm probabilitățile de a avea *j* linii ocupate în cazul existenței a 4 servere.

număr de linii ocupate	probabilitatea
0	0.1429
1	0.2857
2	0.2857
3	0.1905
4	0.0952

Se poate face o analiza economică, comparând costurile unei noi linii telefonice şi al unui nou angajat cu costul apelurilor blocate, adică a clienților pierduți.

Dacă avem 9.52% apeluri blocate şi $\lambda = 12$ apeluri pe oră, într-o zi de 8 ore lucrătoare vom avea o medie de $8 \cdot 12 \cdot 0.0952 = 9.1$ apeluri blocate pe zi.

2. O companie de asigurări are un sistem central ce conține o varietate de informații despre conturile clienților. Agenții de asigurări, care au locația în 6 regiuni folosesc linii telefonice pentru a avea acces în baza de date cu informații despre clienți.

Sistemul central permite accesarea bazei de date de 3 utilizatori simultan.

În cazul în care sistemul este complet, este blocată accesarea sistemului şi nu este permisă aşteptarea în coadă, situație ce este ineficientă.

În medie sunt 42 de apeluri pe oră, apeluri ce au o repartiție Poisson.

Rata medie a service-ului este de 20 de apeluri pe oră pentru fiecare server.

Calculăm probabilitățile $P_j \quad j = 0, 1, 2, 3$ ca j linii de acces să fie ocupate,

utilizând formula $P_j = \dfrac{\dfrac{\lambda}{\mu} \cdot \dfrac{1}{j!}}{\sum\limits_{i=0}^{k} \left(\dfrac{\lambda}{\mu}\right)^i \cdot \dfrac{1}{i!}}$, cu $\lambda = 42, \mu = 20$

$$P_0 = \frac{\dfrac{\lambda}{\mu} \cdot \dfrac{1}{0!}}{\sum\limits_{i=0}^{3} \left(\dfrac{\lambda}{\mu}\right)^i \cdot \dfrac{1}{i!}} = \frac{1}{6.8485};$$

$$P_1 = \frac{\dfrac{\lambda}{\mu} \cdot \dfrac{1}{1!}}{\sum\limits_{i=0}^{3} \left(\dfrac{\lambda}{\mu}\right)^i \cdot \dfrac{1}{i!}} = \frac{2.1}{6.8485} = 0.3066;$$

$$P_2 = \frac{\frac{\lambda}{\mu} \cdot \frac{1}{2!}}{\sum_{i=0}^{3} \left(\frac{\lambda}{\mu}\right)^i \cdot \frac{1}{i!}} = \frac{2.2050}{6.8485} = 0.3220\,; \qquad P_3 = \frac{\frac{\lambda}{\mu} \cdot \frac{1}{3!}}{\sum_{i=0}^{3} \left(\frac{\lambda}{\mu}\right)^i \cdot \frac{1}{i!}} = \frac{1.5435}{6.8485} = 0.2254$$

Probabilitatea de a fi blocată intrarea unui agent în sistem: $P_3 = 0.2254$.

Numărul mediu de linii ocupate: $L = \frac{\lambda}{\mu} \cdot (1 - P_k) = \frac{42}{20} \cdot (1 - 0.2254) = 1.6267$.

Pe viitor compania se va extinde şi consideră că vor fi în medie $\lambda = 50$ apeluri pe oră. Probabilitatea de a fi respins un apel trebuie să fie sub 22.54%, aşa că se pune problema de câte linii de accesare a bazei de date este nevoie.

Deoarece $P_4 = \dfrac{\frac{\lambda}{\mu} \cdot \frac{1}{4!}}{\sum_{i=0}^{4} \left(\frac{\lambda}{\mu}\right)^i \cdot \frac{1}{i!}} = 0.1499$, 4 linii sunt suficiente, probabilitatea ca accesul să fie respins fiind de 15%..

Model în care mărimea populației este finită

Dacă numărul potențialilor clienți este nelimitat rata medie a sosirilor λ este constantă, indiferent de câți clienți sunt in sistem.
În cazul în care numărul potențialilor clienți este finit rata medie a sosirilor descreşte o dată cu creşterea numărului de clienți din sistem.

Aplicație

Aceste modele se folosesc mai ales în cazul sistemelor de reparații:
Un grup de mașini reprezintă o populație finită de „clienți".
Când o mașină se strică, avem o „sosire", în sensul că este necesară o reparație.
Dacă altă mașină se strică, înainte de a fi terminată reparația celeilalte, aceasta din urmă trebuie să aștepte, formându-se astfel o „coadă".
Termenul de server este înlocuit cu punct de reparații.
Termenul de rată de sosire devine rată de defectare, iar rata de service este rată de reparație.

Modelul de aşteptare cu număr finit de potențiali clienți pe care îl prezentăm este caracterizat prin:

- are un singur server
- numărul potențialilor clienți este finit
- sosirile au o repartiție Poisson, rata standard de sosire λ fiind aceeaşi pentru orice client
- timpii de service au o repartiție exponențială, μ fiind rata standard de service.
- disciplina cozii este primul sosit,primul servit.

Caracteristicile numerice ale modelului

Folosim notațiile:

λ -rata standard de sosire pentru fiecare client,

μ - rata medie a service-ului,

N- mărimea populației (numărul de potențiali clienți)

- Probabilitatea ca sa nu fie clienți în sistem : $P_0 = \dfrac{1}{\sum\limits_{n=0}^{N} \dfrac{N!}{(N-n)!} \cdot \left(\dfrac{\lambda}{\mu}\right)^n}$;

- Numărul mediu de clienți din rând: $L_q = N - \frac{\lambda + \mu}{\lambda} \cdot (1 - P_0)$
- Numărul mediu de clienți în sistem: $L = L_q + (1 - P_0)$
- Timpul mediu pe care un client îl petrece la coadă: $W_q = \frac{L_q}{(N - L) \cdot \lambda}$
- Timpul mediu pe care un client îl petrece în sistem: $W = W_q + \frac{1}{\mu}$
- Probabilitatea de a fi *n* clienți în sistem: $P_n = \frac{N!}{(N-n)!} \cdot \left(\frac{\lambda}{\mu}\right)^n \cdot P_0$

Exemple

1. O firmă A are un grup de 6 maşini identice, care lucrează în medie 20 de ore între două reparații.

Astfel rata de defectare (rata de sosire) este $\lambda = \frac{1}{20} = 0.05$ maşini pe oră.

Apariția defecțiunilor are o repartiție Poisson.

La punctul de reparații (server) un meseriaş se ocupă de repararea celor 6 maşini. Timpii de reparație (timpii de service) sunt exponențial distribuiți şi timpul mediu de reparație a unei maşini este de 2 ore, deci rata service-ului este $\mu = \frac{1}{2} = 0.5$ maşini pe oră.

Calculăm caracteristicile numerice ale acestui model de aşteptare, ținând seama că $\frac{\lambda}{\mu} = \frac{0.05}{0.5} = 0.1$

- Probabilitatea ca să nu fie maşini în sistem (maşini de reparat):

$$P_0 = \frac{1}{\sum_{n=0}^{6} \frac{6!}{(6-n)!} \cdot (0.1)^n} =$$

$$= \frac{1}{1+6\cdot 0.1+30\cdot 0.01+120\cdot 0.001+360\cdot 0.0001+720\cdot 0.00001+720\cdot 0.000001} = 0.4845$$

- Numărul mediu de maşini din rând:

$$L_q = N - \frac{\lambda+\mu}{\lambda}\cdot(1-P_0) = 6 - \frac{0.05+0.5}{0.05}\cdot(1-0.4845) = 0.3295 \text{ maşini}$$

- Numărul mediu de maşini în sistem:

$$L = L_q + (1-P_0) = 0.3295 + (1-0.4845) = 0.845 \text{ maşini}$$

- Timpul mediu pe care o maşină defectă aşteaptă reparaţia:

$$W_q = \frac{L_q}{(N-L)\cdot\lambda} = \frac{0.3295}{(6-0.845)\cdot 0.05} = 1.28 \text{ ore}$$

- Timpul mediu pe care o maşină îl petrece în sistem, adică nu lucrează:

$$W = W_q + \frac{1}{\mu} 1.28 + \frac{1}{0.50} = 3.28 \text{ ore.}$$

- Folosind formula $P_n = \frac{N!}{(N-n)!} \cdot \left(\frac{\lambda}{\mu}\right)^n \cdot P_0$ vom afla probabilitatea ca *n* maşini să fie în sistem, adică să nu lucreze:

număr de maşini în sistem	probabilitatea
0	0.4845
1	0.2907
2	0.1453
3	0.0581
4	0.0174
5	0.0035
6	0.0003

analiză economică: ce este mai profitabil pentru firmă: maşinile să aştepte să fie reparate sau să angajeze încă un meseriaş la punctul de reparații.

2. Cinci secretare folosesc acelaşi copiator.

Timpul mediu între două sosiri la copiator este de 40 minute, sosirile au o repartiție Poisson şi rata medie de $\lambda = 1/40 = 0.025$ sosiri pe minut.

O secretară stă în medie 5 minute la copiator, timpii de service au o repartiție exponențială şi rata medie a service-ului este de

$\mu = \frac{1}{5} = 0.20$ utilizatori pe minut.

Calculăm caracteristicile numerice ale acestui model de aşteptare de ținând seama că $\frac{\lambda}{\mu} = 0.125$:

- Probabilitatea ca să nu fie clienți în sistem (secretarele nu au nevoie de copiator):

$$P_0 = \frac{1}{\sum_{n=0}^{5} \frac{5!}{(5-n)!} \cdot (0.125)^n} =$$

$$= \frac{1}{1 + 5 \cdot 0.125 + 20 \cdot (0.125)^2 + 60 \cdot (0.125)^3 + 120 \cdot (0.125)^4 + 120 \cdot (0.125)^5} = 0.4790$$

- Numărul mediu de secretare din rând:

$$L_q = N - \frac{\lambda + \mu}{\lambda} \cdot (1 - P_0) = 5 - \frac{0.225}{0.025} \cdot (1 - 0.4790) = 0.3110$$

- Numărul mediu de secretare la copiator (în sistem) :

$$L = L_q + (1 - P_0) = 0.3110 + (1 - 0.4790) = 0.8320$$

- Timpul mediu pe care o secretară îl petrece la coadă la copiator:

$$W_q = \frac{L_q}{(N-L)\cdot\lambda} = \frac{0.3110}{(5-0.8320)\cdot 0.025} = 2.9846 \text{ minute}$$

- Timpul mediu pe care o secretara îl petrece la copiator:

$$W = W_q + \frac{1}{\mu} = 2.9846 + \frac{1}{0.20} = 7.9846 \text{ minute}$$

- Să calculăm cât timp petrece o secretară la copiator, în timpul unei zile de 8 ore lucrătoare şi cât timp stă la coadă la copiator:

Având o rată medie a sosirilor de $\lambda = 1/40 = 0.025$ sosiri pe minut timpul petrecut la copiator este:

$8ore \cdot 60 min/ora \cdot \lambda \cdot W = 8 \cdot 60 \cdot 0.025 \cdot 7.9846 = 95.8$ min/zi,

în timp ce timpul petrecut la coadă este:

$8ore \cdot 60 min/ora \cdot \lambda \cdot W_q = 8 \cdot 60 \cdot 0.025 \cdot 2.9846 = 35.8$ min/zi

Considerăm că este necesară achiziționarea celui de al doilea copiator, deoarece pe zi cele 5 secretare pierd $5 \cdot 35.8 = 179$ min, adică 3 ore stând în asteptare la copiator.

Pe de altă parte o secretară petrece $\frac{35.8}{480} \cdot 100 = 7.5\%$ din timpul unei zile lucrătoare la coadă la copiator.

Studiu de caz

O linie aeriană a început activitatea de rezervare a biletelor de avion prin telefon, activitate ce poate crea o bună imagine a companiei.

Între orele 10.00 şi 11.00 au fost primite telefoane rezervarea zborurilor la un interval mediu de 3.75 minute.

Se ştie că în general un funcţionar lucrează în medie 3 minute cu fiecare client.

Se poate presupune că telefoanele primite au o repartiţie Poisson iar timpul de service este exponenţial.

Astfel dacă linia telefonică este deseori ocupată şi un client întâmpină greutăţi în a lua legătura cu funcţionarul companiei, se creează o reacţie negativă care duce la pierderi financiare în afaceri.

Un funcționar ce se ocupă de rezervarea telefonică a biletelor de avion este plătit cu 20$ ora.

Managerii companiei doresc o bună funcționare a acestui service, optimizând însă costurile. Compania şi-a propus să se răspundă imediat la cel puțin 85% dintre clienți.

Deoarece timpul mediu de service este mai mic decât rata apelurilor telefonice se pare ca un singur funcționar ar face față.

Dacă serverul este ocupat, clientului i se cere să aştepte, difuzându-se un fundal muzical.

Caracteristicile numerice ale sistemului de rezervare, cu un funcționar:

Media sosirilor este $\lambda = 60/3.75 = 16$ apeluri pe oră.

Rata service-ului este $\mu = 60/3 = 20$ apeluri pe oră.

- probabilitatea ca să nu existe clienți în sistem: $P_0 = 1 - \frac{\lambda}{\mu} = 1 - \frac{16}{20} = 0.20$
- numărul mediu de clienți la rând: $L_q = \frac{\lambda^2}{\mu \cdot (\mu - \lambda)} = \frac{16^2}{20 \cdot (20 - 16)} = 3.2$ clienți
- numărul mediu de clienți în sistem: $L = L_q + \frac{\lambda}{\mu} = 3.2 + \frac{16}{20} = 4$ clienți

- timpul mediu de aşteptare: $W_q = \frac{L_q}{\lambda} = \frac{3.2}{16} = 0.2 \text{ore} = 12$ minute
- timpul mediu petrecut de un client în sistem

$$W = W_q + \frac{1}{\mu} = 0.2 + \frac{1}{20} = 0.25 \text{ore} = 15 \text{ minute}$$

- probabilitatea de aşteptare $P_w = \frac{\lambda}{\mu} = \frac{16}{20} = 0.80$,

A opera cu un singur funcționar este inacceptabil, deoarece probabilitatea de a aştepta este de 80% si timpul mediu de aşteptare este de 12 minute.

Vom analiza care este probabilitatea de aşteptare în cazurile în care exista 2, reapectiv 3 servere.

$k = 2$

Probabilitatea ca să nu existe clienți în sistem:

$$P_0 = \frac{1}{1 + \frac{(\lambda/\mu)^1}{1!} + \frac{(\lambda/\mu)^2}{2!} \cdot \left(\frac{2\mu}{2\mu - \lambda} \right)} = \frac{1}{1 + 0.8 + \frac{(0.8)^2}{2} \cdot \left(\frac{2 \cdot 20}{2 \cdot 20 - 16} \right)} = 0.4286$$

Probabilitatea de aşteptare:

$$P_w = \frac{1}{2!} \cdot \left(\frac{\lambda}{\mu} \right)^2 \cdot \left(\frac{2\mu}{2\mu - \lambda} \right) \cdot P_0 = \frac{1}{2} \cdot (0.8)^2 \cdot \frac{2 \cdot 20}{2 \cdot 20 - 16} \cdot 0.4286 = 0.2286$$

$k = 3$

Probabilitatea ca să nu existe clienți în sistem:

$$P_0 = \frac{1}{1 + \frac{(\lambda/\mu)^1}{1!} + \frac{(\lambda/\mu)^2}{2!} + \frac{(\lambda/\mu)^3}{3!} \cdot \left(\frac{3\mu}{3\mu - \lambda}\right)} = \frac{1}{1 + 0.8 + \frac{(0.8)^2}{2} + \frac{(0.8)^3}{6} \cdot \left(\frac{3 \cdot 20}{3 \cdot 20 - 16}\right)} =$$

$=0.4472$

Probabilitatea de aşteptare este

$$P_w = \frac{1}{3!} \cdot \left(\frac{16}{20}\right)^3 \cdot \left(\frac{3\mu}{3\mu - \lambda}\right) \cdot P_0 = \frac{1}{6} \cdot (0.8)^3 \cdot \frac{3 \cdot 20}{3 \cdot 20 - 16} \cdot 0.4472 = 0.0520$$

Ținând seama de faptul că în cazul a două servere 77% din clienți ar aștepta, iar în cazul a trei servere ar aștepta doar 94,8%, planul companiei fiind de a răspunde imediat la cel puțin 85% dintre clienți vom alege cazul $k = 3$, ceea ce înseamnă angajarea a încă doi funcționari la call-center.

Prezentăm celelalte caracteristici ale modelului cu trei servere:

- Numărul mediu de clienți ce așteaptă :

$$L_q = \frac{(\lambda / \mu)^3 \lambda\mu}{2(3\mu - \lambda)^2} \cdot P_0 = \frac{(0.8)^3 \cdot 16 \cdot 20}{2(3 \cdot 20 - 16)^2} \cdot 0.4472 = 0.0189$$

- Numărul mediu de clienți în sistem: $L = L_q + \frac{\lambda}{\mu} = 0.0189 + 0.8 = 0.8189$

- Timpul mediu pe care un client îl petrece la coadă:

$$W_q = \frac{L_q}{\lambda} = \frac{0.0189}{16} = 0.0012 \text{ore} = 0.07 \text{ minute}$$

- Timpul mediu pe care un client îl petrece în sistem:

$$W = W_q + \frac{1}{\mu} = 0.0012 + \frac{1}{20} = 0.0512 \text{ore} = 3.07 \text{ minute}$$

Am luat în considerare doar apelurile dintre ora 10 şi ora 11, dar numărul apelurilor diferă de la o oră la alta, aşa că este necesar să cunoaştem situația şi să facem pentru fiecare interval orar.

Astfel am putea stabili un orar cu numărul de funcționari necesari în fiecare caz în parte.