Универзитет у Београду
Математички факултет

Системи масовног опслуживања са Пуасоновим улазним потоком и приоритетним опслуживањем

мастер рад

Студент:
Јулијана Јевђовић 1054/2014

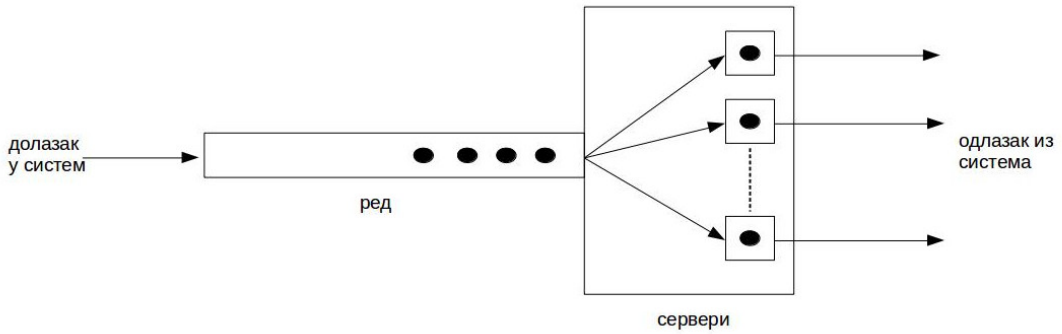
Ментор:
доц. др Ленка Главаш

септембар, 2017.

Садржај

1 Увод

У данашње време чекање је неминовни део свакодневнице. Чека се у реду у пошти или у банци, на каси у продавници, на семафору, у саобраћају, на улазу у паркинг, за коришћење лифта, на указивање помоћи у болници, бродови чекају да уплове у луку, захтеви који стижу до процесора рачунара чекају да буду обрађени. До стварања редова долази јер је потреба за услугом већа од капацитета који систем (банка, пошта, . . .) може да пружи. То би могло да се реши тако што се отвори више шалтера у пошти, прошири број паркинг места, запосли више касирки у маркету али то није економски исплативо. Чак и када би се повећао капацитет, редови би се и даље формирали јер доласци клијената нису савршено усклађени и трајање пружања услуге варира од случаја до случаја. Због тога је потребно пронаћи начин да се смањи чекање у реду без повећања капацитета и да се постигне максимални учинак. Теорија редова чекања (queueing theory) или теорија масовног опслуживања је математичка дисциплина која покушава да одговори на ова питања коришћењем теорије вероватноће и случајних процеса. Она се бави изучавањем система масовног опслуживања, односно система заснованог на односу клијент - сервер (објекат који упућује захтев за известан вид услуге - објекат који пружа услугу). Клијент улази у систем и ако су сви сервери заузети мора да чека, у супротном његово опслуживање почиње одмах. Након што је опслужен, клијент напушта систем. Пример једног таквог основног модела је приказан на Слици 1.



Слика 1: Систем масовног опслуживања

Први значајни допринос теорији редова дао је A. K. Erlang[1] 1909. године у свом делу „The Theory of Probabilities and Telephone Conservations ” [3]. Он се бавио проучавањем телефонског саобраћаја и његова достигнућа представљају основ у теорији редова. Његов рад наставили су E.C.Molina[2], Félix Pollaczek[3], Kolmogorov[4], Khinchin[5] и многи други. Све до средине осамдесетих година прошлог века највећи део резултата није имао практичну примену. Проблеми у реалном свету не одговарају у потпуности математичком моделу и због тога је сада фокус на апроксимацијама, компјутерским прорачунима а не на стриктним теоријским резултатима.

Из свега претходно наведеног закључује се да је ово област која је јако разноврсна са огромном применом и у константном развоју. Пре него што пређемо на описивање

[1] Agner Krarup Erlang (1878-1929), дански математичар и инжењер
[2] Edward Charles Molina (1877-1964), амерички инжењер
[3] Félix Pollaczek (1892-1981), аустријски инжењер и математичар јеврејског порекла
[4] Andrey Nikolaevich Kolmogorov (1903-1987), руски математичар
[5] Aleksandr Yakovlevich Khinchin (1894-1959), руски математичар

система масовног опслуживања, објаснићемо појмове из теорије вероватноће и процесе који су неопходни за разумевање даљег текста.

1.1 Експоненцијална расподела

Дефиниција 1.1. *Случајна величина X има експоненцијалну расподелу $\varepsilon(\lambda)$ са параметром $\lambda, \lambda > 0$, ако је њена густина расподеле вероватноћа $f(x)$ облика*

$$f(x) = \begin{cases} \lambda e^{-\lambda x}, & x \geqslant 0 \\ 0, & x < 0 \end{cases}$$

или еквивалентно, ако је функција расподеле $F(x)$ дата са

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

Математичко очекивање експоненцијално расподељене величине је $EX = 1/\lambda$ и дисперзија $DX = 1/\lambda^2$.

Најзначајнија особина експоненцијалне расподеле је одсуство сећања *(memoryless property)*

$$P\{X > t + s | X > t\} = P\{X > s\} \text{ за } t, s \geqslant 0.$$

Наиме, ако случајна променљива X представља дужину рада у сатима неког уређаја без квара, тада неједнакост $X > t$ значи да је уређај исправан после t сати рада. Вероватноћа да уређај ради исправно бар још s сати једнака је вероватноћи да је уређај исправан s сати од укључења. Односно, уређај не памти шта се раније десило, као да „не зна"да је радио t сати.

Експоненцијална расподела има велику примену, у теорији поузданости, метеорологији, биологији, системима масовног опслуживања и повезана је са многим другим расподелама вероватноћа.

1.2 Пуасонов процес *(Poisson process)*

Дефиниција 1.2. *Случајни процес $\{N(t), t \geqslant 0\}$ је Пуасонов*[6] *ако има следећа својства:*

1. *$N(0) = 0$ скоро сигурно;*
2. *процес $N(t)$ има независне прираштаје;*
3. *постоји неопадајућа фунцција $\mu : [0, +\infty) \to [0, +\infty)$ која је непрекидна са десне стране и за коју важи $\mu(0) = 0$. За произвољне s и $t, 0 < s < t$ важи да $N(t) - N(s)$ има Пуасонову расподелу са параметром $\mu(t) - \mu(s)$, тј.*

$$N(t) - N(s) \sim \mathcal{P}\big(\mu(t) - \mu(s)\big).$$

   Функција μ назива се функција средње вредности Пуасоновог процеса N.

4. *са вероватноћом 1 трајекторије $\big(N(t, w)\big)_{t \geqslant 0}$ случајног процеса N су непрекидне са десне стране за $t \geqslant 0$ и имају леву граничну вредност за $t > 0$.*

Независност прираштаја значи да су за произвољне $0 = t_0 < t_1 < \ldots < t_n$ случајне величине $N(t_k) - N(t_{k-1})$ независне за свако $k = 1, 2, \ldots n$.

Ако је функција средње вредности апсолутно непрекидна онда се може представити у облику $\mu(s, t] = \int_s^t \lambda(u)du, s < t$ за неку ненегативну функцију λ која се назива *функција интензитета.*

[6] Siméon Denis Poisson (1781-1840), француски математичар

Дефиниција 1.3. *Ако је функција средње вредности μ Пуасоновог процеса $\{N(t), t \geqslant 0\}$ линеарна функција времена, односно $\mu(t) = \lambda t$ за $t \geqslant 0$ и $\lambda > 0$, онда се тај процес назива хомоген Пуасонов процес са интензитетом λ. Тада*

$$N(t) - N(s) \sim \mathcal{P}\big(\lambda(t-s)\big).$$

Уколико је $\lambda = 1$ онда се процес назива стандардни хомоген Пуасонов процес.

Дефиниција 1.4. *Поток догађаја $\{V(t), t \geqslant 0\}$ је случајни процес који узима ненегативне целобројне вредности, чије трајекторије не опадају и за који важи $V(0) = 0$.*

Процес $V(t)$ има смисао броја догађаја који су наступили у интервалу $[0, t]$. Хомоген Пуасонов процес је најпростији поток догађаја код кога је вероватноћа да у интервалу дужине t наступи тачно k догађаја

$$P\{N(t) = k\} = \frac{(\lambda t)^k e^{-\lambda t}}{k!},\ k \in \mathbb{N}_0.$$

Случајна величина $N(t)$ има Пуасонову расподелу са параметром λt и њено очекивање, као и дисперзија износи λt.

Услови који доводе до Пуасоновог потока су стационарност прираштаја, одсуство после-дејства и ординарност.

Стационарност прираштаја значи да вероватноћа појављивања k догађаја у интервалу $(s, s+t)$ не зависи од s, већ само од k и t (зависи само од дужине интервала, а не од његовог положаја на временској полуоси).

Одсуство после-дејства означава да појављивање догађаја у интервалу $(s, s+t)$ не зависи од броја догађаја и стања система до тренутка s.

Ординарност потока је услов да је немогуће појављивање два или више догађаја истовремено, тј. у истом тренутку времена.

Пуасонов процес има врло битно својство у литератури познато као PASTA својство (на енглеском језику акроним синтагме Poisson Arrivals See Time Averages). Оно показује да клијент који дође у систем у складу са Пуасоновим процесом види систем који је достигао равнотежно стање на исти начин као да је дошао у произвољном тренутку. Равнотежно стање или еквилибријум представља ситуацију када систем достигне стационарну расподелу, тј. када престану да се мењају вероватносна својства система са протоком времена. Ако је p_j стационарна вероватноћа да је систем у стању E_j у било ком тренутку (посматра клијент који је ван система) и p_j^* вероватноћа да је систем у стању E_j непосредно пре доласка клијента, из PASTA својства следи да је $p_j = p_j^*$.

1.3 Ланци Маркова

Дефиниција 1.5. *Процес Маркова*[7] *$\{X(t), t \in T\}$ је стохастички процес са својством да уколико је позната вредност $X(t)$, на вредности $X(s)$, $s > t$ не утичу вредности $X(u)$, за $u < t$. Односно, када је познато тренутно стање процеса, вероватноћа понашања процеса у будућности не зависи од понашања процеса у прошлости.*

Колекција свих могућих вредности које променљива $X(t)$ може да узме је простор стања S. Уколико је тај скуп коначан или пребројив, $\{X(t), t \in T\}$ је *процес са дискретним простором стања*, за који се користи и назив *ланац*. Ако вредности које променљива $X(t)$ може да има припадају интервалу, било коначном или бесконачном, онда је у питању *процес са непрекидним скупом стања*. У овом раду биће разматрани само процеси Маркова са дискретним простором стања.

[7] Andrey Andreyevich Markov (1856-1922), руски математичар

1.3.1 Ланци Маркова са дискретним временом

Ако се преласци из стања ланца Маркова догађају у дискретним временским тренуцима, односно ако је параметарски скуп T пребројив онда се процес $\{X(t), t \in T\}$ назива ланац Маркова са дискретним временом.

Дефиниција 1.6. *Низ случајних променљивих $X_0, X_1, X_2, \ldots$ са пребројивим скупом стања $S = \{x_0, x_1, x_2, \ldots\}$ зове се ланац Маркова са дискретним временом ако важи*

$$P\{X_{n+1} = x_{n+1} \mid X_0 = x_0, X_1 = x_1, \ldots, X_n = x_n\} = P\{X_{n+1} = x_{n+1} \mid X_n = x_n\}.$$

За ланац Маркова са дискретним временом користи се ознака $\{X_n, n \in \mathbb{N}_0\}$.

Дефиниција 1.7. *Вероватноћа прелаза из стања i у стање j у једном кораку ако прелаз почиње у временском тренутку n је*

$$p_{ij}(n) = P\{X_{n+1} = j \mid X_n = i\}$$

за свако $i, j \in S$ и за свако $n \geqslant 0$.

Уколико вероватноћа $p_{ij}(n)$ не зависи од n (од временског тренутка), онда је *ланац хомоген* и следи да је

$$p_{ij} = P\{X_{n+1} = j \mid X_n = i\} = p_{ij}(n) \text{ за свако } n \geqslant 0.$$

У даљем раду биће речи искључиво о хомогеним ланцима Маркова.

Вероватноће p_{ij}, где $i, j \in S$, чине матрицу прелаза за један корак $\mathbf{P} = [p_{ij}]_{i,j}$. За матрицу $\mathbf{P}$ важи да је $\sum_{j \in S} p_{ij} = 1$ за свако $i \in S$, односно нису дозвољени преласци у стање које не припада простору стања S.

Слично, може да се говори о вероватноћи прелаза из стања i у стање j али у n корака.

Дефиниција 1.8. *Вероватноћа прелаза из стања i у стање j у n корака је*

$$p_{ij}^{(n)} = P\{X_{n+k} = j \mid X_k = i\}.$$

Матрица вероватноћа прелаза за n корака је

$$\mathbf{P_n} = \left[p_{ij}^{(n)}\right]_{i,j}.$$

Израчунавање вероватноћа прелаза у n корака омогућава једначина Kolmogorov-Chapman[8]

$$p_{ij}^{(m+n)} = \sum_{k \in S} p_{ik}^{(m)} p_{kj}^{(n)} = \sum_{k \in S} p_{ik}^{(n)} p_{kj}^{(m)}.$$

Прелаз из стања i у стање j за $m + n$ корака може да се постигне тако што се из стања i за m корака пређе у међустање k са вероватноћом $p_{ik}^{(m)}$. Затим се из стања k пређе у стање j за n корака.

Стање j је достижно из стања i ако је $p_{ij}^{(n)} > 0$ за неко $n \in \mathbb{N}$. Ланац Маркова је *несводљив (irreducible)* ако је свако стање достижно из било ког другог стања.

За свако стање i са f_i означимо вероватноћу да ће процес крећући из стања i икада поново ући у стање i. Стање i је *повратно(reccurent)* ако је $f_i = 1$, односно *пролазно (transient)* ако је $f_i < 1$.

Стање i има *период* d ако је $p_{ii}^{(n)} = 0$ кад год n није дељиво са d, а d је највећи природан број са овим својством. Ако је период један ($d = 1$) стање је *апериодично(aperiodic)*.

[8] Sydney Chapman (1888-1970), британски математичар

Дефиниција 1.9. *Ланац Маркова је ергодичан ако постоје граничне вредности* $\lim_{n\to\infty} p_{ij}(n) = p_j^*$, *за свако* $i, j \in S$, *које не зависе од* i *и при чему је* $\sum_{j\in S} p_j^* = 1$. *Расподела* $\{p_j^*, j \in S\}$ *представља граничну расподелу ланца Маркова.*

Уколико су сва стања несводљивог ланца повратна и апериодична ланац је *ергодичан.*

Дефиниција 1.10. *Расподела вероватноћа* $\pi = \{\pi_j, j \in S\}$ *је стационарна расподела ланца Маркова са матрицом прелаза* $\mathbf{P}$ *ако важи да је* $\pi = \pi\mathbf{P}$, *односно по компонентама* $\pi_j = \sum_{k\in S} \pi_k p_{kj}$ *за свако* $j \in S$, *при чему је* $\pi_j \geqslant 0$ *за свако* $j \in S$ *и* $\sum_{j\in S} \pi_j = 1$.

Теорема 1.1. *За несводљив, апериодичан ланац Маркова следећа својства су еквивалентна:*

1. *ланац је ергодичан;*
2. *ланац има стационарну расподелу;*
3. *ланац има граничну расподелу.*

У овом случају стационарна и гранична расподела су једнаке и позитивне су.

1.3.2 Ланци Маркова са непрекидним временом

Код ланаца Маркова са непрекидним временом прелази из стања се дешавају у било ком временском тренутку и време које систем проведе у датом стању је случајна променљива која има експоненцијалну расподелу.

Дефиниција 1.11. *Случајни процес* $\{X(t), t \geqslant 0\}$ *је ланац Маркова са непрекидним временом и пребројивим скупом стања* $S = \{x_0, x_1, x_2, \ldots\}$ *ако за свако* $n \in \mathbb{N}$ *и низ временских тренутака* $t_0, t_1, \ldots, t_{n+1}$ *тако да је* $0 < t_0 < t_1 < \ldots < t_{n+1}$ *важи*

$$\begin{aligned} &P\{X(t_{n+1}) = x_{n+1} \mid X(t_0) = x_0, X(t_1) = x_1, \ldots, X(t_n) = x_n\} \\ &= P\{X(t_{n+1}) = x_{n+1} \mid X(t_n) = x_n\}. \end{aligned}$$

Дефиниција 1.12. *Вероватноћа прелаза за један корак из стања* i *у стање* j *је*

$$p_{ij}(s,t) = P\{X(t) = j \mid X(s) = i\}$$

за свако $i, j \in S$ *и за* $0 < s < t$.

Код хомогеног ланца Маркова вероватноће прелаза се дефинишу као

$$p_{ij}(\tau) = P\{X(s+\tau) = j \mid X(s) = i\}$$

где је τ заправо $t - s$.

Једначина Kolmogorov-Chapman за прелаз за $t + h$ корака, $t, h > 0$, је

$$p_{ij}(t+h) = \sum_{k\in S} p_{ik}(t) p_{kj}(h).$$

За описивање ланаца Маркова са непрекидним временом потребно је дефинисати и интензивност (брзину) прелаза q_{ij}. Претпоставимо да је

$$\lim_{t\to 0+} p_{ij}(t) = \begin{cases} 1, & \text{за } i = j \\ 0, & \text{за } i \neq j \end{cases}.$$

Дефиниција 1.13. *Интензитет прелаза из стања i у стање j за хомоген ланац Маркова са непрекидним временом је дат са*

$$q_{ij} = p'_{ij}(0) = \lim_{t \to 0+} \frac{p_{ij}(t)}{t} \text{ за } i \neq j$$

$$q_{ii} = p'_{ii}(0) = \lim_{t \to 0+} \frac{p_{ii}(t)}{t} \text{ за } i = j.$$

Матрица $\mathbf{Q} = [q_{ij}]_{i,j \in S}$ назива се *инфинитезимални генератор* или *матрица интензивности*. Збир елемената по врстама матрице $\mathbf{Q}$ је нула, $\sum_{j \in S} q_{ij} = 0$ за свако $i \in S$. Дужина боравка процеса у неком стању, тј. време до првог изласка из неког стања има експоненцијалну расподелу чији параметар зависи од тог стања.

Теорема 1.2. *Гранична расподела ергодичног хомогеног ланца Маркова са непрекидним временом једнака је стационарној расподели и израчунава се решавањем система једначина*

$$q_{jj}\pi_j + \sum_{k \neq j} q_{kj}\pi_k = 0$$

или у матричном запису

$$\pi \mathbf{Q} = 0$$

уз услов да је $\sum_{i \in S} \pi_i = 1$ *и* $\pi = (\pi_0, \pi_1, \pi_2, \ldots)$.

2 Карактеристике система масовног опслуживања

Систем масовног опслуживања састоји се од клијената или муштерија који у случајним временским тренуцима стижу на место где добијају одређени вид услуге и онда одлазе. При томе долази до стварања редова уколико је већа потражња за услугом него што су могућности сервера који дају услугу. На време чекања и дужину реда утичу пре свега интензитет пристизања и интензитет обраде захтева, али и начин опслуживања и други параметри. На основу свих тих параметара извршена је карактеризација и класификација система масовног оплслуживања.

2.1 Улазни поток *(Arrival process)*

Улазни поток клијената је случајни процес и моменти доласка клијената у систем су ненегативне случајне величине $\tau_1, \tau_2, \tau_3, \ldots$ уређене монотоно неопадајуће. Клијенти могу да долазе појединачно (у једном тренутку само један клијент) или у групама произвољне величине (batch или bulk arrivals). Случајна величина τ_i представља долазак i-тог клијента, односно i-те групе.

Интензитет улазног потока λ је просечан број клијената који дођу у систем у јединици времена. Може бити временски зависан, што је случај у ресторанима када је повећан број долазака у вечерњим сатима или да зависи од стања система, на пример када је дужина реда ограничена.

Након доласка у систем у коме се већ формирао ред, клијент може да одлучи да чека колико год је неопходно или да напусти систем пре него што буде опслужен. Ако пристигли клијент одбије да чека или му није дозвољено (јер је систем пун) кажемо да је *одбијен (balked)*. Клијент може да чека, при чему после извесног времена изгуби стрпљење и одлучи да оде. У овом случају клијент је *одустао (reneged)*. Када постоји два или више паралелних редова клијенти могу да прелазе из једног у други (у нади да

ће што мање чекати) односно *бирају најбољу позицију (jockey for position)*. Претходна три наведена случаја представљају пример тзв. нестрпљивих клијената.

За описивање система масовног опслуживања јако је битно одредити расподелу времена које протекне између доласка клијената у систем. Претпоставља се да су времена између узастопних долазака $T_1 = \tau_1 - 0, T_2 = \tau_2 - \tau_1, T_3 = \tau_3 - \tau_2, ...$ независне и једнако расподељене случајне величине. Уколико су и експоненцијално расподељене онда је улазни поток *хомоген Пуасонов процес*. У том случају вероватноћа доласка n клијената у интервалу дужине t је

$$P\{N(t) = n\} = \frac{(\lambda t)^n e^{-\lambda t}}{n!}.$$

2.2 Процес опслуживања *(Service process)*

Као и улазни поток и процес опслуживања је случајни процес. Нека је B_i време које је потребно да сервер пружи услугу i-том клијенту. Дужине опслуживања $B_1, B_2, B_3, ...$ су ненегативне, независне и једнако расподељене случајне промењиве.

Интензитет опслуживања μ је просечан број опслужених клијената у јединици времена када је сервер заузет. Трајање опслуживања може да зависи од стања система и од времена. Сервер може да ради брже уколико има више захтева али са друге стране може да постане и успорен. Временска зависност подразумева да се временом побољшавају способности оних који врше опслуживање.

Иако је уобичајено да се на једном серверу опслужује само један клијент, може се десити да истовремено један сервер опслужује више клијената. То је случај код рачунара код којих је могуће паралелно извршавање задатака. Још један пример су туристи на вођеној тури за обилазак различитих дестинација, где улогу сервера има туристички водич.

Чак и при великом интензитету опслуживања опет долази до формирања редова јер клијенти у систем долазе насумично у произвољним временским тренуцима. Стога, расподела вероватноћа дужине реда зависи од улазног процеса и процеса опслуживања, који су углавном међусобно независни.

2.3 Број сервера *(Number of servers)*

Систем може да се састоји од једног или више паралелних сервера, чак и бесконачно много. Нека је s ознака за број сервера. Различити сервери могу да имају различит интензитет опслуживања. У систему са више сервера може да постоји само један ред тако да клијент који је на реду за опслуживање приступа првом слободном серверу (пример је ред у пошти) или да за сваки сервер постоји засебан ред (као на каси у хипермаркетима).

2.4 Капацитет система *(System capacity)*

Максимални број корисника у систему (у реду и на опслуживању) представља капацитет система и означава се са c. Може се сматрати да је капацитет бесконачан уколико није другачије назначено. Ако је капацитет коначан и број клијената у систему је једнак c, новим клијентима није дозвољен улазак у систем.

2.5 Популација *(Population)*

Величина популације представља укупан број клијената у систему и ван система опслуживања, и означава се са n. Обично је величина популације толико велика у поређењу са бројем клијената у систему па се може сматрати бесконачном. За популацију

кажемо да је хомогена уколико су клијенти међусобно слични у својим потребама и начину на који се врши њихово опслуживање.

2.6 Дисциплина опслуживања *(Queue discipline)*

Дисциплина опслуживања представља начин на који се бира који ће клијент из реда бити следећи опслужен. Најчешћа је FIFO (first in, first out), позната и под акронимом FCFS (first come, first serve), где је следећи клијент који се опслужује онај који је најдуже у реду, који је први дошао од присутних. Код LIFO (last in, first out) или LCFS (last come, first served) дисциплине, последњи клијент који се прикључио реду биће први услужен. SIRO (service in random order) или RSS (random selection for service) је метод код кога се клијент бира из реда на опслуживање на случајан начин и сви клијенти имају једнаку шансу да буду изабрани. У информационационим технологијама и комуникацији највише се користи PS (processor sharing) дисциплина. За PS методу карактеристично је да опслуживање свих клијената одмах почиње (нема чекања) али је интензитет опслуживања пропорционалан броју клијента у систему. Код приоритетног опслуживања или PR (priority) дисциплине, клијенти су подељени у групе различитог приоритета. Клијенти са највишим приоритетом се први опслужују независно од тога када су ушли у систем, а потом они са нижим приоритетом. Типичан пример ове методе је на одељењу хитне помоћи у болницама, где се пацијентима који су у критичном стању прво указује помоћ.

2.7 Кендалова нотација *(Kendall notation)*

Кендал[9] је 1953. године у свом чланку „Stochastic Processes Occurring in the Theory of Queues and their Analysis by the Method of the Imbedded Markov Chain" [4] увео нотацију која се и данас користи за описивање система масовног опслуживања. Наиме, систем се може описати помоћу низа симбола $A|B|s|c|p|Z$, при чему словне ознаке имају следеће значење:

A - природа улазног потока односно тип расподеле интервала између узастопних долазака у систем;

B - процес опслуживања, тј. тип расподеле дужине опслуживања клијената;

s - број сервера;

c - капацитет система;

p - величина популације;

Z - дисциплина опслуживања.

Симболи који се најчешће користе за опис улазног потока и процеса опслуживања су:

- M - Марковљев процес, означава да су времена између узастопних долазака клијената у систем или дужине трајања опслуживања експоненцијално расподељене;
- E_k - Ерлангова расподела реда k, значи да времена између узастопних долазака клијената у систем или дужине трајања опслуживања имају Ерлангову расподелу;
- D - детерминистичка расподела, тј. константно време које протекне између долазака клијената, односно константно време потребно за опслуживање сваког клијента;

[9] David George Kendall (1918-2007), британски математичар

- G - генерална, било која расподела.

О значењима осталих симбола било је речи у претходном делу.

Запис $M|D|2|\infty|\infty|FIFO$ представља систем са Пуасоновим улазним потоком, детерминистичком расподелом опслуживања, два паралелна сервера, неограниченим капацитетом и неограниченом величином популације и FIFO дисциплином. Ако је популација бесконачно велика ($p = \infty$) и дисциплина опслуживања FIFO ($Z = FIFO$), онда се ови симболи изостављају из нотације. Такође, ако је и капацитет система бесконачан ($c = \infty$) и та ознака се изоставља. У већини ситуација користе се само прва три симбола. Тако $M|M|3$ означава систем са Пуасоновим улазним потоком, експоненцијалном расподелом времена опслуживања, три сервера, неограниченим капацитетом и популацијом и FIFO дисциплином.

2.8 Перформансе система

У анализи ефикасности система битно је утврдити колико времена клијент мора да проведе у реду, шта доводи до стварања реда, као и колико времена су сервери неактивни. Како су у питању стохастички системи, ове мере су случајне величине чије се расподеле, или бар очекиване вредности могу израчунати.

Када је реч о времену које клијент утроши разликују се два типа, време које проведе у реду и укупно време које проведе у систему. У зависности од циља анализе, значајно је једно или друго или чак оба. Ако се посматра забавни парк, треба се фокусирати на време које је изгубљено чекањем у реду, док је за поправку машина значајно целокупно време док се машина поново не оспособи за рад. Сходно томе, битно је проценити и просечан број клијената у реду, као и у читавом систему. Неактивност сервера може се мерити за сваки сервер појединачно или на нивоу целог система.

Нека је λ интензитет улазног потока, X случајна величина која представља дужину опслуживања и s број сервера у систему. Искоришћеност система ρ је мера система која се израчунава као $\rho \overset{\text{деф}}{=} \lambda E[X]/s$. Ако је $\rho > 1 (\lambda E[X] > s)$, просечан број долазака клијената у систем већи је од просечног броја клијената које сви сервери могу да опслуже и временом долази до стварања све већег реда. Да би се достигло равнотежно стање система ρ мора да буде строго мање од 1. Када је $\rho = 1$ и уколико улазни поток и процес опслуживања нису детерминистички, не достиже се еквилибријум, пошто никад неће доћи до пражњења система.

У системима масовног опслуживања основно је наћи број клијената у систему N. Тај број представља суму клијената који чекају у реду N_q, и оних који су на опслуживању N_s. Како се након дужег временског периода систем уравнотежи (достигне еквилибријум) може се наћи стационарна расподела $\pi_n = P\{N = n\}$. Просечан број клијената у систему је

$$L = E[N] = \sum_{n=0}^{\infty} n\pi_n,$$

и просечан број клијената у реду

$$L_q = E[N_q] = \sum_{n=s+1}^{\infty} (n - s)\pi_n,$$

при чему је s број сервера. Нека T_q представља време које клијент проведе чекајући у реду и T укупно време проведено у систему ($T = T_q + X$). Две основне мере перформансе су просечно време чекања у реду $W_q = E[T_q]$ и просечно време које клијент проведе у

систему $W = E[T]$ и за њих важи следећа веза

$$L = \lambda W \tag{2.8.1}$$

и

$$L_q = \lambda W_q. \tag{2.8.2}$$

Претходне две формуле представљају Little-ов закон[10] (*Little's law*). Little је ове формуле развио шездесетих година прошлог века и то је један од најважнијих резултата у теорији редова. Нигде се не претпоставља какав је систем у питању, који је улазни поток, дисциплина опслуживања, већ резултати важе за све системе. Довољно је наћи једну од четири очекиване вредности из формула и могу се наћи остале. При томе, јасно је да важи $E[T] = E[T_q] + E[X]$, односно $W = W_q + E[X]$.

Још један занимљив резултат добија се на основу Little-овог закона

$$L - L_q = \lambda(W - W_q) = \lambda E[X].$$

3 Математичке трансформације

Понекад није једноставно добити експлицитне резултате за различите перформансе система као што су време чекања у реду, број клијената у систему, просечна дужина реда итд. У том случају за њихово израчунавање користе се математичке трансформације. Поред тога што се намећу као природно решење проблема, имају и бројне особине које омогућавају лакша израчунавања и добијање других значајних резултата. Трансформације које су највише коришћене у раду објашњене су у тексту који следи.

3.1 z-трансформација

Нека је $f_n, n \geqslant 0$, реалан низ. За f_n z-трансформација се дефинише на следећи начин:

$$F(z) \overset{\text{деф}}{=} \sum_{n=0}^{\infty} f_n z^n.$$

За дату функцију f_n z-трансформација је јединствена. $F(z)$ је функција комплексне променљиве z и постоји све док низ функција f_n не расте брже него геометријски, тј. све док постоји $a > 0$ тако да је

$$\lim_{n\to\infty} \frac{|f_n|}{a^n} = 0.$$

Једна од најбитнијих особина z-трансформације је конволуција. Конволуција функција f_n и g_n се дефинише на следећи начин

$$f_n * g_n \overset{\text{деф}}{=} \sum_{k=0}^{n} f_{n-k} g_k.$$

z-трансформација конволуције функција је

$$f_n * g_n = F(z)G(z),$$

при чему су $F(z)$ и $G(z)$ одговарајуће z-трансформације функције f_n, односно g_n. Доказ се може наћи у [1].

[10] John Little (1928-), амерички инжењер

Нека је X дискретна случајна величина са датом расподелом вероватноћа. Функција генератриса вероватноћа, односно генераторна функција од X је

$$F(z) = \sum_{n=0}^{\infty} P\{X = n\} z^n = E[z^X].$$

Функција $F(z)$ је заправо математичко очекивање случајног елемента z^X.

Ако је $|z| \leq 1$ онда је $|F(z)| \leqslant 1$, односно

$$|F(z)| \leqslant \sum_{n=0}^{\infty} |P\{X = n\}||z^n| \leqslant \sum_{n=0}^{\infty} P\{X = n\} = 1.$$

Први извод функције $F(z)$ у тачки 1 једнак је првом моменту случајне величине X, $F'(1) = EX$. Други извод од $F(z)$, такође у тачки 1 износи $F''(1) = E[X^2] - EX$. Уопштено важи да је

$$E[X(X-1)\dots(X-n+1)] = \frac{d^n}{dz^n} F(z)\Big|_{z=1}.$$

На основу формуле долази се до облика z-трансформације неке функције f_n. Међутим, обрнут случај, из z-трансформације добити f_n није увек једноставно. Постоје формуле инверзије, а често се на њих примењују нумеричке методе. Значај ових трансформација је у томе што се из њиховог облика може доћи до оригиналне функције. Преглед неких функција и одговарајућих z-трансформација дат је у књизи [1], Appendix I.

3.2 Лапласова трансформација

Лапласова трансформација непрекидне функције $f(t), t \geqslant 0$, је

$$F^*(s) \stackrel{\text{деф}}{=} \int_{-\infty}^{\infty} f(t) e^{-st} dt.$$

$F^*(s)$ је функција комплексне променљиве s и има слична својства као и z-трансформација. За функцију $f(t)$ Лапласова трансформација је јединствена и постоји све док $f(t)$ нема раст бржи од експоненцијалног, односно док постоји реалан број b тако да је

$$\lim_{\tau \to \infty} \int_0^{\tau} |f(t)| e^{bt} dt < \infty.$$

Лапласова трансформација конволуције функција $f(t)$ и $g(t)$ је

$$f(t) * g(t) = F^*(s) G^*(s),$$

при чему су $F^*(s)$ и $G^*(s)$ одговарајуће Лапласове трансформације функције $f(t)$, односно $g(t)$. Доказ ове једнакости може се наћи у [1].

Уколико је $f(x)$ функција густине неке случајне променљиве X, Лапласова трансформација представља математичко очекивање елемента e^{-sX}

$$F^*(s) = E(e^{-sX}).$$

Моменат реда n случајне величине X може се изразити преко Лапласове трансформације

$$E[X^n] = (-1)^n F^{*(n)}(0),$$

где $F^{*(n)}$ означава n-ти извод Лапласове трансформације.

Као и код z-трансформације, тако је и овде могуће из облика Лапласове трансформације одређеним методама доћи до оригиналне функције. Још неке особине Лапласове трансформације као и неке битније функције са одговарајућим трансформацијама налазе се у књизи [1], Appendix I.

4 $M|G|1$ системи

Централни део овога рада посвећен је $M|G|1$ системима. То су системи који немају марковско својство и то отежава описивање својстава система. Развијене су различите методе за ову класу процеса, а овде је коришћен приступ помоћу уметнутих ланаца Маркова (*imbedded Markov chain*). У пракси се користи и метод стања (*method of stages*), али његова највећа мана је што резултати нису дати експлицитно па се не могу испитивати својства. Трећи приступ је помоћу Линдлове интегралне једначине (*Lindley's integral equation*) која је погодна за $G|G|1$ системе.

$M|G|1$ је систем у коме је улазни поток хомоген Пуасонов процес, дужина трајања опслуживања је случајна величина која има општу (неодређену) расподелу вероватноћа и опслуживање врши један сервер по редоследу пристизања клијената у систем. Расподела времена између долазака дата је функцијом расподеле

$$A(t) = 1 - e^{-\lambda t}, t \geqslant 0, \lambda > 0.$$

Интензитет којим клијенти долазе је λ клијената по јединици времена, просечно време између долазака $1/\lambda$ одговарајуће јединице времена, варијанса $\sigma^2 = 1/\lambda^2$. Нека је функција расподеле дужине трајања опслуживања $B(x)$ и функција густине $b(x)$.

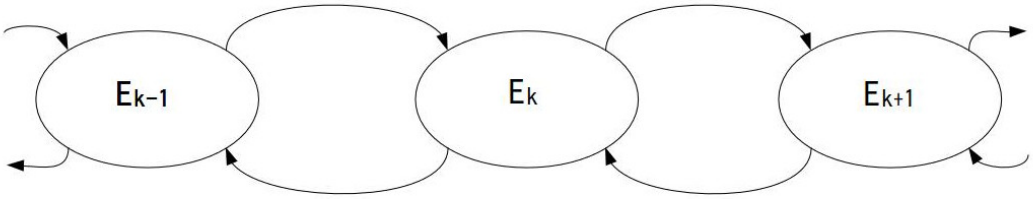
За описивање стања система у неком тренутку t поред броја клијената у систему $N(t)$, потребно је одредити и $X_0(t)$, време до тренутка t утрошено на опслуживање тренутног клијента. Ово је неопходно јер је расподела дужине трајања опслуживања произвољна и у општем случају нема својство одсуства меморије, за разлику од улазног процеса. Стога, случајни процес $\{N(t), t \geqslant 0\}$ није марковски процес док фамилија вектора $\{(N(t), X_0(t)), t \geqslant 0\}$ јесте. Овај вектор је вектор стања и садржи све информације о прошлости система које су релевантне за понашање $M|G|1$ система у будућности. Систем се може анализирати и коришћењем овог вектора.

4.1 Уметнути ланац Маркова

Пошто је компликовано и непрактично за израчунавање користити дводимензионални вектор, идеја је да се вектор $(N(t), X_0(t))$ упрости на једнодимензионални вектор $N(t)$. У том случају мора се задати потрошено време клијента који се опслужује. Ово се постиже тако што се гледа само одређени дискретан скуп тачака са својством да уколико се утврди број клијената у систему у једној таквој тачки моћи ће да се израчуна и у следећој таквој тачки. Постоји много скупова таквих тачака али је најпогоднији скуп тачака у тренутку напуштања система. У тим тачкама време потрошено на опслуживање је нула за клијента који тек треба да буде услужен (јер још није почело његово опслуживање), а и за све остале клијенте у систему. На тај начин одређено је потрошено време $X_0(t)$ и смањена је димензија вектора стања. Овде је заправо описан полу-марковски процес *(semi-Markov process)*, јер се прелази из стања дешавају у тренуцима одласка клијената из система. У тим инстанцама се дефинише уметнути ланац Маркова који представља број клијената у систему након одласка. Вероватносна расподела времена које протекне између промене стања је иста као расподела времена опслуживања $B(x)$, сем у случају када након одласка остане празан систем.

У уметнутим тачкама посматра се број клијената који је остао у систему и на основу тога испитује понашање система. Резултати добијени у овим тачкама одласка преносе се на све тачке током времена. Биће објашњено зашто се то може учинити.

Нека је E_k стање система када је k клијената у њему. Једини преласци који су могући из стања E_k су за један корак у E_{k+1} и E_{k-1}, $E_k \to E_{k+1}$ и $E_k \to E_{k-1}$, при томе да је последњи прелаз могућ само када је $k > 0$.



Слика 2: Могући прелази система из стања E_k

Прелази типа $E_k \to E_{k+1}$ одговарају доласку клијента у систем и дешавају се у тачкама доласка, док $E_k \to E_{k-1}$ прелази представљају одласке из система и догађају се у тренуцима одласка. Број ових прелаза може се разликовати највише за један (више долазака него одлазака) али након дужег времена овај број се изједначава. Зато је гранична расподела стања система одређена доласком у систем једнака граничној расподели стања система одређеној одласком из система, односно важи следеће:

1. када је улазни поток Пуасонов процес важи да је

$$\pi_k = P\{\text{долазак у тренутку } t \text{ затиче } k \text{ клијената у систему}\};$$

2. ако број клијената има само дискретне промене у величини и ако било која од следећих граничних расподела постоји, онда постоје и остале расподеле и једнаке су:

$r_k := \lim_{n\to\infty} P\{\text{долазак у тренутку } t \text{ затиче } k \text{ клијената у систему}\}$
$d_k := \lim_{n\to\infty} P\{\text{одлазак у тренутку } t \text{ оставља } k \text{ клијената у систему}\}$

$$r_k = d_k$$

и за $M|G|1$ важи

$$r_k = \pi_k = d_k.$$

4.2 Вероватноће прелаза

У тренуцима одласка клијената дефинише се уметнути ланац Маркова као број клијената који остаје у систему након одласка. На овај начин су комплетно описана стања система јер је време потрошено на опслуживање клијента у том моменту једнако нули и време доласка последњег клијента не утиче на будућност због својства одсуства меморије. Да би описали овај ланац потребно је наћи вероватноће прелаза. Уведимо следеће ознаке које ће се користити даље у раду:

C_n представља n-тог клијента који је ушао у систем;

τ_n је тренутак доласка клијента C_n;

$t_n = \tau_n - \tau_{n-1}$ време које протекне између доласка клијента C_{n-1} и C_n;

x_n време потребно за опслуживање клијента C_n;

q_n је број клијената који остану у систему након одласка клијента C_n;

v_n је број клијената који дођу у систем док се C_n опслужује.

Вероватноћа прелаза за један корак дефинише се као

$$p_{ij} \overset{\text{деф}}{=} P\{q_{n+1} = j \mid q_n = i\}$$

при чему прелаз није могућ ако је $q_{n+1} < q_n - 1$. Могући су прелази када је испуњен услов да је $q_{n+1} \geqslant q_n - 1$. Вероватноће прелаза се рачунају на следећи начин:

$$p_{00} = P\{q_{n+1} = 0 \mid q_n = 0\} = P\{v_{n+1} = 0\},$$

$$p_{01} = P\{q_{n+1} = 1 \mid q_n = 0\} = P\{v_{n+1} = 1\},$$

$$p_{02} = P\{q_{n+1} = 2 \mid q_n = 0\} = P\{v_{n+1} = 2\},$$

односно вероватноћа да је након одласка C_n остало i клијената, а након одласка C_{n+1} j клијената једнака је вероватноћи да је за време опслуживања C_{n+1} дошло $j - i + 1$ клијената

$$p_{ij} = P\{q_{n+1} = j \mid q_n = i\} = P\{v_{n+1} = j - i + 1\}.$$

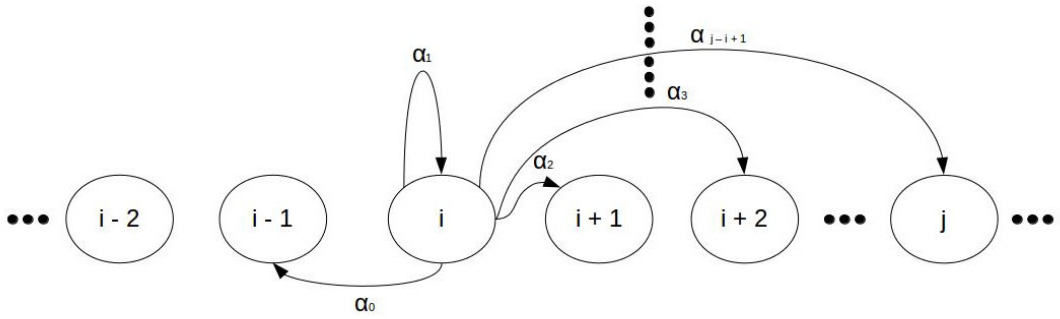
Ако са α_k означимо вероватноћу да је током опслуживања C_n дошло k клијената, $\alpha_k \overset{\text{деф}}{=} P\{v_n = k\}$, за свако $(i, j = 0, 1, 2, \ldots)$ вероватноће прелаза су

$$p_{ij} = \begin{cases} 0, & j < i - 1 \\ \alpha_{j-i+1}, & i \geqslant 1,\ j \geqslant i - 1 \\ \alpha_j, & i = 0 \end{cases}.$$

Матрица вероватноће прелаза $\mathbf{P} = [p_{ij}](i, j = 0, 1, 2, \ldots)$ има следећи облик

$$\mathbf{P} = \begin{bmatrix} \alpha_0 & \alpha_1 & \alpha_2 & \alpha_3 & \ldots \\ \alpha_0 & \alpha_1 & \alpha_2 & \alpha_3 & \ldots \\ 0 & \alpha_0 & \alpha_1 & \alpha_2 & \ldots \\ 0 & 0 & \alpha_0 & \alpha_1 & \ldots \\ 0 & 0 & 0 & \alpha_0 & \ldots \\ \vdots & \vdots & \vdots & \ddots & \vdots \end{bmatrix}$$

На Слици 3 приказан је граф промене стања из стања i.



Слика 3: Могући прелази из стања i

Интензитет улазног потока не зависи од стања система, ни дужина трајања опслуживања за C_n не зависи од n (x_n има произвољну расподелу $B(x)$). Број клијената v_n, који је дошао током x_n зависи само од дужине трајања x_n, не и од n. Због тих независности можемо x_n заменити са $\tilde{x}$ и v_n са $\tilde{v}$. Вероватноће α_k рачунамо на следећи начин коришћењем закона тоталне вероватноће:

$$\begin{aligned}\alpha_k &= P\{v_n = k\} = P\{\tilde{v} = k\} \\ &= \int_0^\infty P\{\tilde{v} = k \mid \tilde{x} = x\} b(x) dx.\end{aligned}$$

Током времена $\tilde{x}$ дошло је k клијената и пошто је улазни поток Пуасонов процес, важи

$$\alpha_k = \int_0^\infty \frac{(\lambda x)^k}{k!} e^{-\lambda x} b(x) dx. \tag{4.2.1}$$

Вероватноће α_k су позитивне за свако $k \geqslant 0$, па је свако стање достижно из било ког стања и Марковљев ланац је несводљив и апериодичан.

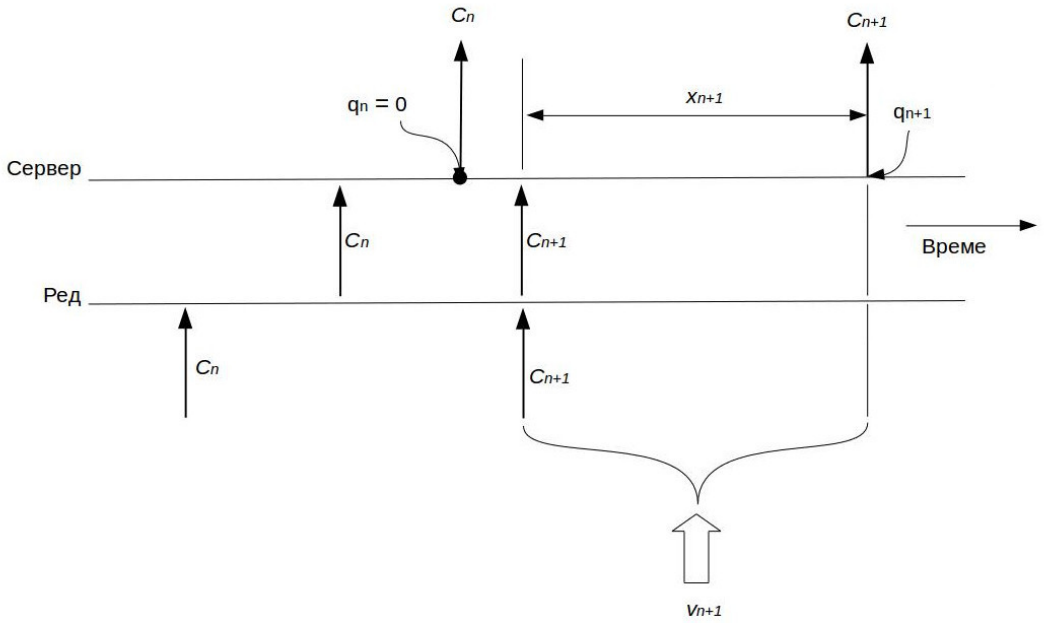
Систем достиже стационарну расподелу ако је интензитет опслуживања клијената већи од интензитета доласка клијената. У супротном ред се никада не би смањивао, константно би растао. Стационарна расподела постоји ако важи $\pi = \pi\mathbf{P}$ и $\sum_{j \in S} \pi_j = 1$, где је вектор $\pi = (\pi_0, \pi_1, \pi_2, \ldots)$ и $\mathbf{P}$ матрица вероватноћа прелаза. Како је ланац Маркова несводљив и апериодичан, из теореме (1.1) следи да је стационарна расподела уједно и гранична расподела, односно $\pi_k = d_k = P\{\tilde{q} = k\}$.

4.3 Pollaczek-Khinchin-ова формула

На основу Pollaczek-Khinchin (ен. mean value formula) формуле може се израчунати број клијената у систему, у реду, време које клијент проведе чекајући и укупно време које проведе у систему. За израчунавање ових величина потребни су основни параметри, интензитет доласка клијената и познавање расподеле дужине трајања опслуживања.

Претпоставимо да је уметнути ланац Маркова ергодичан и да систем у неком тренутку достиже еквилибријум. У том случају на основу стационарне расподеле можемо добити граничне вредности појединих величина. Почећемо са израчунавањем броја клијената у систему. Иако је уметнути ланац Маркова формиран у тачкама одласка, број клијената у тим тренуцима представља број клијената у систему због тога што систем видимо на исти начин и у тачкама доласка и у тачкама одласка и самим тим у свим тачкама (PASTA својство).

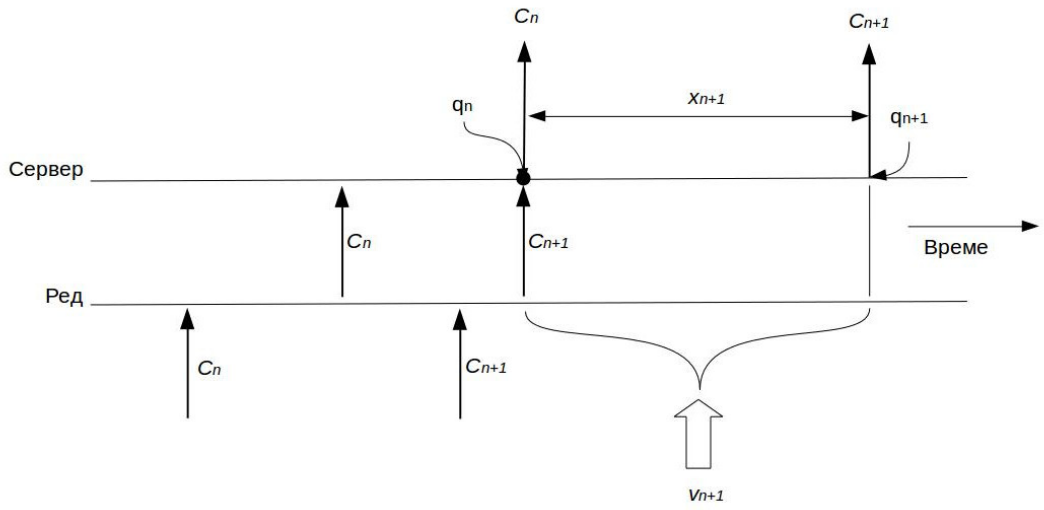
Најпре је потребно пронаћи везу између q_{n+1} и q_n, где је q_n број клијената који остану у систему након одласка клијента C_n. Разликују се две ситуације: када након одласка клијента C_n остане празан систем, $q_n = 0$, и када након одласка C_n постоје клијенти у систему, $q_n > 0$. Први случај је приказан на дијаграму (Слика 4).



Слика 4: Систем када након одласка клијента C_n нема других клијената у систему

У овом случају је број клијената који је остао након одласка C_{n+1} заправо број клијената који је дошао током опслуживања C_{n+1}, $q_{n+1} = v_{n+1}$. Напоменимо да се овде и даље у раду подразумева FIFO метода опслуживања која утиче на дужину чекања у реду, али не на дужину реда и периоде када је сервер заузет.

Када је $q_n > 0$, број клијената који остане у систему након одласка C_{n+1} једнак је суми клијената који су остали након одласка C_n, изузимајући самог клијента C_{n+1}, и клијената који су дошли током опслуживања C_{n+1}, $q_{n+1} = q_n - 1 + v_{n+1}$. Пример ове ситуације приказан је на Слици 5.



Слика 5: Систем када након одласка клијента C_n има клијената у систему

Ове две ситуације могу се груписати у једну чиме се добија следећи израз за q_{n+1}

$$q_{n+1} = \begin{cases} q_n - 1 + v_{n+1}, & \text{за } q_n > 0 \\ v_{n+1}, & \text{за } q_n = 0 \end{cases}$$

Уведимо функцију Δ_k

$$\Delta_k \overset{\text{деф}}{=} \begin{cases} 1, & \text{за } k = 1, 2, 3 \dots \\ 0, & \text{за } k \leqslant 0 \end{cases}$$

Користећи функцију Δ_k, q_{n+1} се може записати на следећи начин

$$q_{n+1} = q_n - \Delta_{q_n} + v_{n+1}. \tag{4.3.1}$$

Израз (4.3.1) представља везу између q_{n+1} и q_n и основ за проучавање $M|G|1$ система. Претпоставимо да постоји гранична вредност j-тог момента од q_n, односно

$$\lim_{n\to\infty} E[q_n^j] = E[\tilde{q}^j].$$

Нађимо математичко очекивање леве и десне стране израза (4.3.1)

$$E[q_{n+1}] = E[q_n] - E[\Delta_{q_n}] + E[v_{n+1}].$$

Када $n \to \infty$ гранична вредност је

$$E[\tilde{q}] = E[\tilde{q}] - E[\Delta_{\tilde{q}}] + E[\tilde{v}].$$

Индекс n код случајне величине $\tilde{v}$ је занемарен јер број клијената који дођу током опслуживања не зависи од n.

Из последњег израза добија се да је

$$E[\Delta_{\tilde{q}}] = E[\tilde{v}] \tag{4.3.2}$$

Десна страна ове једнакости $E[\tilde{v}]$ представља просечан број клијената који дође у систем током опслуживања клијента. На основу дефиниције математичког очекивања, лева страна се израчунава

$$E[\Delta_{\tilde{q}}] = \sum_{k=0}^{\infty} \Delta_k P\{\tilde{q} = k\} = \Delta_0 P\{\tilde{q} = 0\} + \Delta_1 P\{\tilde{q} = 1\} + \dots$$

Како је $\Delta_k = 1$ за $k > 0$, а у свим осталим случајевима је нула, претходни израз се своди на

$$E[\Delta_{\tilde{q}}] = \sum_{k=1}^{\infty} P\{\tilde{q} = k\} = P\{\tilde{q} > 0\}.$$

Вероватноћа $P\{\tilde{q} > 0\}$ представља вероватноћу да је систем заузет, односно да након одласка у систему увек има клијената. Та вероватноћа је заправо искоришћеност система ρ

$\rho \overset{\text{деф}}{=}$ (просечан интензитет доласка клијената)·(просечно време опслуживања)$= \lambda E[x]$. Ово се може протумачити као да је у систем донето $\lambda E[x]$ посла јер сваки клијент донесе количину посла за чије обављање је потребно просечно $E[x]$ времена. Када је у питању систем са s сервера искоришћеност је $\rho = \lambda E[x]/s$.

Из једнакости (4.3.2) долази се до закључка да је

$$E[\tilde{v}] = \rho. \tag{4.3.3}$$

Да би систем достигао еквилибријум потребно је да важи $\rho < 1$, иначе долази до преоптерећености система и нагомилавања клијената у систему. Такве системе није могуће анализирати.

Претходним рачунањем математичког очекивања обеју страна једнакости (4.3.1) није добијен резултат за $E[\tilde{q}]$. Исти поступак се може применити на квадрирану једнакост (4.3.1)

$$q_{n+1}^2 = q_n^2 + (\Delta_{q_n})^2 + v_{n+1}^2 - 2q_n\Delta_{q_n} + 2q_n v_{n+1} - 2\Delta_{q_n} v_{n+1}.$$

Из дефиниције функције Δ_k следи да је $(\Delta_{q_n})^2 = \Delta_{q_n}$ и $q_n\Delta_{q_n} = q_n$. Примењујући ово добија се да је

$$E[q_{n+1}^2] = E[q_n^2] + E[\Delta_{q_n}] + E[v_{n+1}^2] - 2E[q_n] + 2E[q_n v_{n+1}] - 2E[\Delta_{q_n} v_{n+1}].$$

Број клијената v_{n+1} који је дошао током опслуживања C_{n+1}, и број клијената q_n који је остао у систему након одласка C_n, независне су случајне величине и очекивање њиховог производа је једнако производу очекивања. Уколико $n \to \infty$ добија се

$$0 = E[\Delta_{\tilde{q}}] + E[\tilde{v}^2] - 2E[\tilde{q}] + 2E[\tilde{q}]E[\tilde{v}] - 2E[\Delta_{\tilde{q}}]E[\tilde{v}].$$

Користећи израз (4.3.2), а потом и (4.3.3), следи да је

$$E[\tilde{q}] = \frac{\rho + E[\tilde{v}^2] - 2\rho^2}{2 - 2\rho}.$$

Потребно је још наћи $E[\tilde{v}^2]$.

Помоћу z-трансформације налази се било који моменат случајне величине $\tilde{v}$. Како је $\alpha_k = P\{\tilde{v} = k\}$, z-трансформација од $\tilde{v}$ има следећи облик

$$V(z) = E[z^{\tilde{v}}] = \sum_{k=0}^{\infty} \alpha_k z^k.$$

У претходном поглављу одређена је експлицитна формула $\alpha_k = \int_0^\infty \frac{(\lambda x)^k}{k!} e^{-\lambda x} b(x)dx$.

Применом Beppo-Levi теореме добија се

$$\begin{aligned} V(z) &= \sum_{k=0}^{\infty} \int_0^\infty \frac{(\lambda x)^k}{k!} e^{-\lambda x} b(x)dx z^k \\ &= \int_0^\infty e^{-\lambda x} \sum_{k=0}^{\infty} \frac{(\lambda x z)^k}{k!} b(x)dx \\ &= \int_0^\infty e^{-\lambda x} e^{\lambda x z} b(x)dx \\ &= \int_0^\infty e^{-x(\lambda - \lambda z)} b(x)dx. \end{aligned}$$

Лапласова трансформација за расподелу времена опслуживања је

$$B^*(s) = \int_0^\infty b(x) e^{-sx} dx.$$

Израз за $V(z)$ је заправо Лапласова трансформација $B^*(s)$ само за вредност $\lambda - \lambda z$

$$V(z) = B^*(\lambda - \lambda z). \tag{4.3.4}$$

Последња једнакост представља веома корисну и битну везу између z-трансформације расподеле случајне величине $\tilde{v}$ и Лапласове трансформације густине случајне величине $\tilde{x}$.

Из особина Лапласове и z-трансформације следи да је

$$B^{*(k)}(0) = \left.\frac{d^k B^*(s)}{ds^k}\right|_{s=0} = (-1)^k E[\tilde{x}^k] \tag{4.3.5}$$

$$V'(1) = \left.\frac{dV(z)}{dz}\right|_{z=1} = E[\tilde{v}] \tag{4.3.6}$$

$$V''(1) = \left.\frac{d^2V(z)}{dz^2}\right|_{z=1} = E[\tilde{v}^2] - E[\tilde{v}]. \tag{4.3.7}$$

Ради лакшег означавања уместо $E[\tilde{x}^k]$ и $E[\tilde{v}^k]$ користићемо $\overline{x^k}$ и $\overline{v^k}$, респективно. Да би израчунали $E[\tilde{v}^2] = \overline{v^2}$ потребан нам је други извод функције $V(z)$ у тачки $z = 1$. Из једнакости (4.3.4) следи да је

$$\frac{d^2V(z)}{dz^2} = \frac{d^2B^*(\lambda - \lambda z)}{dz^2}.$$

Други извод функције $B^*(\lambda - \lambda z)$ је

$$\begin{aligned}\frac{d^2B^*(\lambda - \lambda z)}{dz^2} &= \frac{d}{dz}\left[\frac{dB^*(\lambda - \lambda z)}{dz}\right] \\ &= \frac{d}{dz}\left[-\lambda\frac{dB^*(\lambda - \lambda z)}{d(\lambda - \lambda z)}\right] \\ &= \lambda^2\frac{d^2B^*(\lambda - \lambda z)}{d(\lambda - \lambda z)^2}.\end{aligned}$$

Добили смо да је $V''(1) = \lambda^2 B^{*(2)}(0)$ и користећи формуле (4.3.5) и (4.3.7) важи да је

$$\overline{v^2} = \bar{v} + \lambda^2\overline{x^2}.$$

Вратимо се на израз $E[\tilde{q}] = \dfrac{\rho + E[\tilde{v}^2] - 2\rho^2}{2 - 2\rho}$. Сада су све величине у десном делу ове једнакости познате и добија се

$$E[\tilde{q}] = \rho + \frac{\lambda^2 E[x^2]}{2(1-\rho)}. \tag{4.3.8}$$

Последња формула је *Pollaczek-Khinchin P-K mean value formula* којом се израчунава просечан број клијената у $M|G|1$ систему. Ако у израз (4.3.8) уврстимо формулу за квадратни коефицијент варијације расподеле времена опслуживања $C_b^2 = \sigma_b^2/E[x]^2$ добија се

$$E[\tilde{q}] = \rho + \rho^2\frac{(1 + C_b^2)}{2(1-\rho)}. \tag{4.3.9}$$

На основу Little-овог закона и једнакости (4.3.8) и (4.3.9) може се израчунати просечно време проведено у систему W, просечно време проведено у реду W_q и просечан број клијената у реду L_q.

$$W = E[x] + \frac{\lambda E[x^2]}{2(1-\rho)} = E[x] + \frac{\rho E[x](1 + C_b^2)}{2(1-\rho)}$$

Први члан суме је просечно време опслуживања $E[x]$, док други члан представља просечно време проведено у реду (време чекања)

$$W_q = \frac{\lambda E[x^2]}{2(1-\rho)} = \frac{\rho E[x](1+C_b^2)}{2(1-\rho)}.$$

Просечан број клијената у реду је

$$L_q = \frac{\lambda^2 E[x^2]}{2(1-\rho)} = \frac{\rho^2(1+C_b^2)}{2(1-\rho)}.$$

У добијеним формулама фигуришу само интензитет улазног потока и први и други моменат расподеле опслуживања. Значај P-K формула је у томе да знајући само расподелу улазног потока и расподелу опслуживања можемо израчунати основне перформансе система.

4.4 Расподела броја клијената у систему

P-K формуле иако значајне не дају никакву информацију о расподели случајних величина. Како у опису $M|G|1$ система користимо уметнути ланац Маркова и посматрамо стање система у тренуцима одласка клијената, расподела броја клијената је расподела случајне величине q_n, односно њене граничне вредности $\tilde{q}$. Ову расподелу ћемо наћи помоћу z-трансформација. Нека су $Q_n(z)$ и $Q(z)$ z-трансформације од q_n и $\tilde{q}$, респективно

$$Q_n(z) = \sum_{k=0}^{\infty} P\{q_n = k\} z^k = E[z^{q_n}], \tag{4.4.1}$$

$$Q(z) = \lim_{n\to\infty} Q_n(z) = \sum_{k=0}^{\infty} P\{\tilde{q} = k\} z^k = E[z^{\tilde{q}}]. \tag{4.4.2}$$

За израчунавање $Q_n(z)$ користимо раније изведену једнакост $q_{n+1} = q_n - \Delta_{q_n} + v_{n+1}$. Најпре рачунамо степен са основом z

$$z^{q_{n+1}} = z^{q_n - \Delta_{q_n} + v_{n+1}},$$

а потом и математичко очекивање

$$E[z^{q_{n+1}}] = E[z^{q_n - \Delta_{q_n} + v_{n+1}}].$$

Лева страна претходне једнакости је заправо $Q_{n+1}(z)$. Случајне величине q_n и v_{n+1} су независне, самим тим и њихове функције, и очекивање производа једнако је производу очекивања. Добија се да је

$$Q_{n+1}(z) = E[z^{q_n - \Delta_{q_n}}] E[z^{v_{n+1}}].$$

Број долазака током опслуживања клијента зависи само од дужине опслуживања, тако да v_{n+1} не зависи од n и може се заменити са $\tilde{v}$. Израз $E[z^{\tilde{v}}]$ је z-трансформација од $\tilde{v}$ која је у поглављу *4.3* дефинисана као $V(z)$. У једнакости

$$Q_{n+1}(z) = E[z^{q_n - \Delta_{q_n}}] V(z) \tag{4.4.3}$$

непознат је још само фактор $E[z^{q_n - \Delta_{q_n}}]$. На основу дефиниције математичког очекивања следи да је

$$E[z^{q_n - \Delta_{q_n}}] = \sum_{k=0}^{\infty} P\{q_n = k\} z^{k-\Delta_k} = P\{q_n = 0\} z^0 + \sum_{k=1}^{\infty} P\{q_n = k\} z^{k-1}.$$

Сума у последњој једнакости слична је изразу (4.4.1), али се разликује у експоненту и индексу сумације. То може да се трансформише на следећи начин:

$$\sum_{k=1}^{\infty} P\{q_n = k\} z^{k-1} = \frac{1}{z} \sum_{k=0}^{\infty} P\{q_n = k\} z^k - \frac{1}{z} P\{q_n = 0\} z^0$$
$$= \frac{1}{z}(Q_n(z) - P\{q_n = 0\}).$$

Уврстимо последњи резултат у израз (4.4.3)

$$Q_{n+1}(z) = \left(P\{q_n = 0\} + \frac{Q_n(z) - P\{q_n = 0\}}{z} \right) V(z).$$

Нађимо лимес ове једнакости када n тежи бесконачно

$$Q(z) = \left(P\{\tilde{q} = 0\} + \frac{Q(z) - P\{\tilde{q} = 0\}}{z} \right) V(z). \tag{4.4.4}$$

Вероватноћа $P\{\tilde{q} > 0\}$ је вероватноћа да у систему увек има клијената и износи ρ, па је $P\{\tilde{q} = 0\} = 1 - \rho$. Подсетимо се везе између z-трансформације просечног броја клијената који дођу у систем током опслуживања и Лапласове трансформације просечног времена опслуживања $V(z) = B^*(\lambda - \lambda z)$. Заменом ових вредности у (4.4.4) добија се

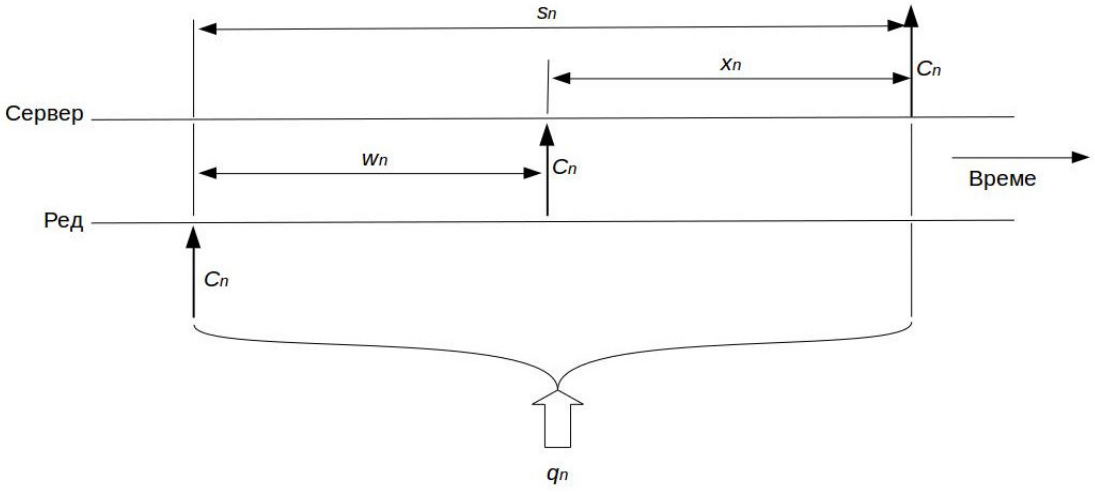
$$Q(z) = B^*(\lambda - \lambda z) \frac{(1 - \rho)(1 - z)}{B^*(\lambda - \lambda z) - z} \tag{4.4.5}$$

Ово је Pollaczek-Khinchin-ова једнакост (*Pollaczek-Khinchin P-K transform equation*) којом је дата расподела броја клијената у систему. Она је представљена у виду трансформација и због комплексности израчунавања није увек једноставно доћи до експлицитног облика расподеле за q_n. Ситуација се компликује када вредност P-K формуле покушамо да израчунамо за $z = 1$. Јавља се неодређени облик и потребно је користити Л'Опиталово правило иако је познато из дефиниције $Q(z)$ да је $Q(1) = 1$.

4.5 Расподела времена које клијент проведе у систему и у реду

Дисциплина опслуживања није имала утицаја на досадашње резултате за $M|G|1$ системе. За одређивање расподеле времена које клијент проведе чекајући битан је начин опслуживања. Претпоставимо да се опслуживање врши у складу са FIFO (first in, first out) методом.

На следећој слици приказано је време које клијент C_n проведе у систему, од тренутка уласка у систем до момента када га напусти.



Слика 6: Време које клијент C_n проведе у систему

Случајна величина w_n представља време које протекне од када C_n уђе у систем док не почне његово опслуживање. То је временски период који C_n потроши чекајући у реду. Дужина трајања опслуживања је x_n, док је s_n укупно време које C_n проведе у систему, тј.

$$s_n = w_n + x_n.$$

Клијенти који се налазе у систему у тренутку одласка C_n су заправо они који дођу током боравка клијента C_n у систему. Ово следи из чињенице да је у питању FIFO дисциплина опслуживања и да клијенти напуштају систем тек када буду услужени. Нека је S_n функција расподеле вероватноћа укупног времена које клијент C_n проведе у систему

$$S_n(y) = P\{s_n \leqslant y\}, \quad \text{за } y \geqslant 0.$$

Пошто претпостављамо ергодичност уметнутог ланца Маркова, низ случајних величина s_n *конвергира у расподели*[11] ка $\tilde{s}, s_n \xrightarrow{\mathrm{D}} s$, односно низ функција расподеле S_n конвергира ка функцији расподеле S, $S_n \rightrightarrows S$ када $n \to \infty$.

Дакле,

$$S(y) = P\{\tilde{s} \leqslant y\}, \quad \text{за } y \geqslant 0.$$

Лапласова трансформација укупног времена проведеног у систему је

$$S^*(s) = \int_0^\infty e^{-sy} dS(y) = E[e^{-s\tilde{s}}].$$

Подсетимо се једнакости (4.3.4)

$$V(z) = B^*(\lambda - \lambda z).$$

$V(z)$ је z-трансформација расподеле броја клијената који долазе у складу са Пуасоновим процесом (са интензитетом λ) током опслуживања клијента. Десни део једнакости

[11] Низ случајних величина $(X_n)_{n\in\mathbb{N}}$, са функцијом расподеле $(F_n)_{n\in\mathbb{N}}$, конвергира у расподели ка случајној величини X ако за сваку ограничену и непрекидну функцију $f : \mathbb{R} \to \mathbb{R}$ важи $\lim\limits_{n\to\infty} E[f(X_n)] = E[f(X)]$, ознака $X_n \xrightarrow{\mathrm{D}} X$, када $n \to \infty$. Ако конвергира у расподели низ случајних величина, конвергира и одговарајући низ функција расподеле $F_n \rightrightarrows F$ и обратно.

је Лапласова трансформација расподеле опслуживања $B(x)$ у конкретној тачки. Ова веза ће важити за било које две променљиве од којих једна представља број долазака клијената где је улазни поток Пуасонов, а друга временски интервал током кога се ти доласци броје. Ако је тај временски интервал укупно време које клијент C_n проведе у систему s_n, q_n је променљива којом се описује број долазака клијената. Из аналогије са формулом (4.3.4) добија се следеће

$$Q(z) = S^*(\lambda - \lambda z). \tag{4.5.1}$$

Pollaczek-Khinchin-ова једнакост из претходног поглавља (4.4.5) даје расподелу за q_n. Када ту вредност уврстимо у (4.5.1) добија се

$$S^*(s) = B^*(s)\frac{s(1-\rho)}{s - \lambda + \lambda B^*(s)}. \tag{4.5.2}$$

Ово је друга P-K једнакост којом је дата расподела укупног времена проведеног у систему у облику Лапласове трансформације. У формули фигуришу само почетни параметри, интензитет Пуасоновог процеса и расподела времена опслуживања.

Функција расподеле вероватноћа времена које клијент проведе чекајући дефинише се као

$$W_n(y) = P\{w_n \leqslant y\}, \quad \text{за } y \geqslant 0.$$

Као и у случају укупног времена проведеног у систему, w_n конвергира у расподели ка $\tilde{w}, w_n \xrightarrow{\mathrm{D}} \tilde{w}$ и $W_n \rightrightarrows W$, када $n \to \infty$. Функција расподеле за $\tilde{w}$ је

$$W(y) = P\{\tilde{w} \leqslant y\}, \quad \text{за } y \geqslant 0.$$

Одговарајућа Лапласова трансформација је

$$W^*(s) = \int_0^\infty e^{-sy} dW(y) = E[e^{-s\tilde{w}}].$$

Време опслуживања клијента не зависи од времена које клијент проведе у реду. Дакле, укупно време проведено у систему је сума две независне променљиве. Пошто посматрамо граничне расподеле запис је следећег облика

$$\tilde{s} = \tilde{w} + \tilde{x}. \tag{4.5.3}$$

Лапласова трансформација израза (4.5.3) је

$$S^*(s) = W^*(s)B^*(s).$$

Ово следи из Лапласове трансформације конволуције функција. Из једнакости (4.5.2) долази се до резулата за $W^*(s)$

$$W^*(s) = \frac{s(1-\rho)}{s - \lambda + \lambda B^*(s)}. \tag{4.5.4}$$

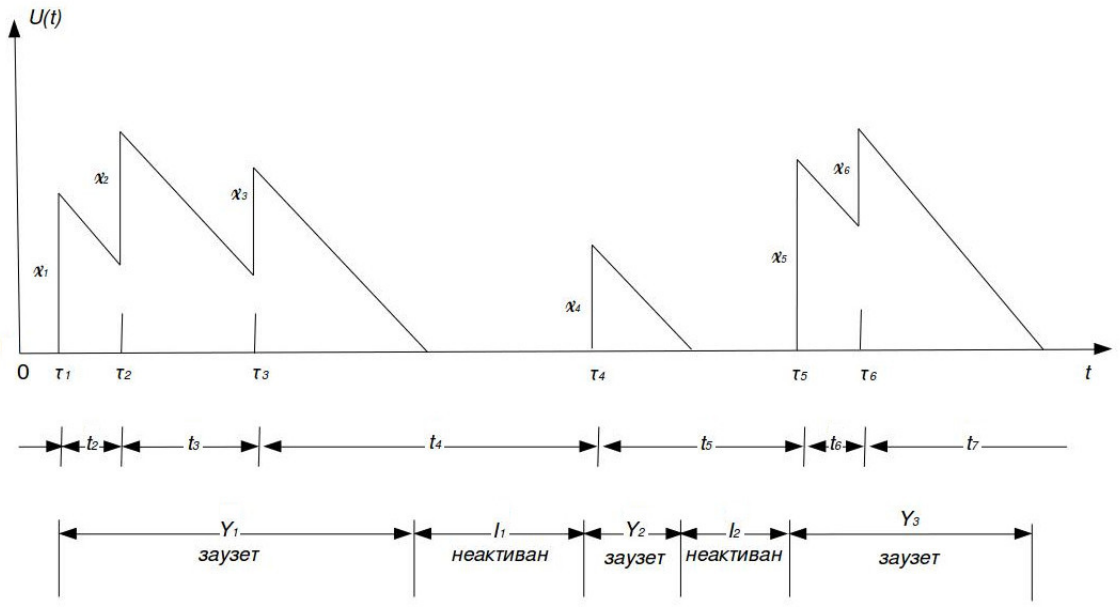
Овим је одређена и расподела времена чекања клијента у систему. Резулатат (4.5.4) је познат као трећа P-K једнакост.

4.6 Период када је сервер заузет

Систем масовног опслуживања пролази наизменично кроз периоде када је заузет (бар један клијент је у систему) и када је неактиван (у систему нема клијената). Потребно је наћи расподелу дужине трајања периода заузетости и периода када је систем неактиван.

Нека је $U(t)$ функција која представља незавршен посао у систему у тренутку t. Може се интерпретирати и као време потребно да се услуже сви корисници који су у тренутку t присутни у систему. Када је у питању FIFO дисциплина опслуживања, $U(t)$ је и време чекања (време које проведе у реду) клијента који је у моменту t ушао у систем. Сама функција $U(t)$ не зависи од дисциплине опслуживања. Из дефиниције функције следи да је систем заузет када је $U(t) > 0$, а неактиван када је $U(t) = 0$.

На Слици 7 приказано је како се смењују заузети и неактивни временски интервали. Означимо трајање периода заузетости са $Y_1, Y_2, Y_3, \ldots$ и неактивних периода са $I_1, I_2, I_3, \ldots$



Слика 7: График функције $U(t)$, периоди заузетости и неактивни периоди

Клијент C_1 улази у систем у тренутку τ_1 и пошто не затиче друге клијенте одмах почиње његово опслуживање. Тиме је завршен један неактивни интервал и почео је период заузетости система дужине Y_1. C_1 је донео одређену количину посла за коју је потребно време опслуживања x_1. Функција $U(t)$ има скок величине x_1. Како време протиче, функција $U(t)$ опада јер се опслуживање приводи крају. Након t_2 јединица времена, у моменту τ_2, у систем улази клијент C_2. $U(t)$ има скок величине x_2, тј. времена потребно за опслуживање клијента C_2. Функција $U(t)$ поново опада, све док у систем не уђе C_3 када долази до скока величине x_3. Период заузетости траје све до тренутка $\tau_1 + Y_1$, када је завршено опслуживање свих клијената и систем је празан. Тада почиње нови неактивни период I_1. Неактивни период траје све до доласка клијента C_4 када се иницира почетак новог периода заузетости. На овај начин се смењују периоди заузетости и неактивни периоди. Функција $U(t)$ има скокове када клијенти долазе у систем и опада све док је позитивна. Дефинишимо функције расподеле дужине неактивног периода и

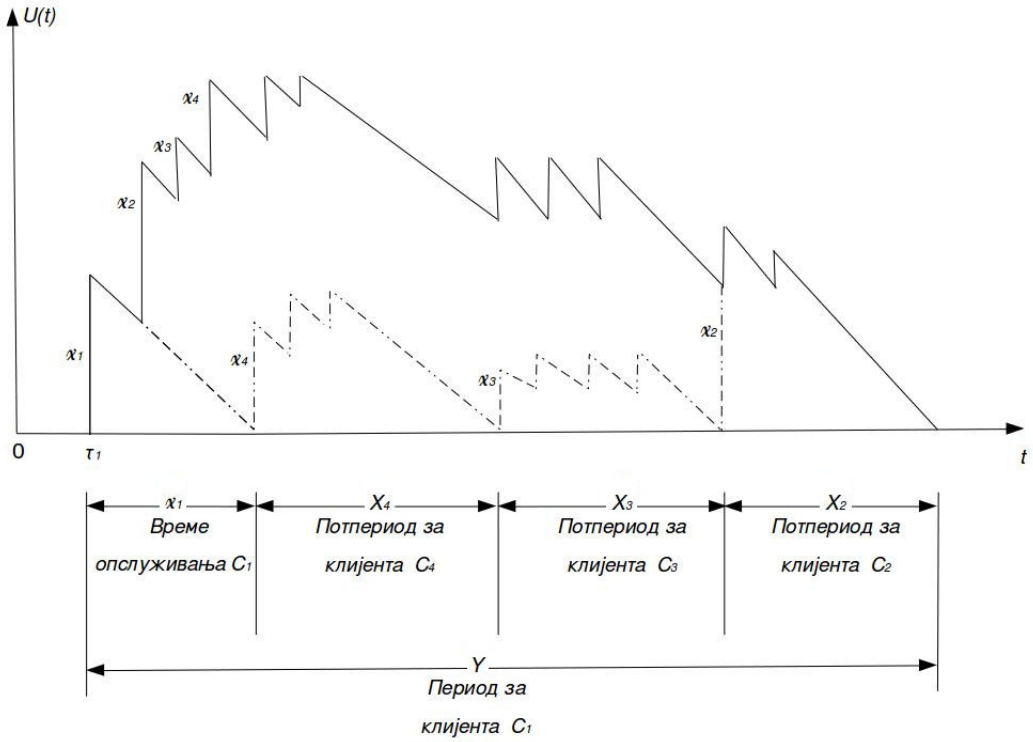
периода заузетости система

$$F(y) \stackrel{\text{деф}}{=} P\{I_n \leqslant y\} \stackrel{\text{деф}}{=} \text{расподела дужине неактивног периода,}$$
$$G(y) \stackrel{\text{деф}}{=} P\{Y_n \leqslant y\} \stackrel{\text{деф}}{=} \text{расподела дужине периода заузетости.}$$

Када се заврши период заузетости, започиње нови неактивни период који траје све док клијент не уђе у систем. Пошто расподела времена до доласка следећег клијента има својство одсуства меморије, расподела дужине неактивног периода је

$$F(y) = 1 - e^{-\lambda y}, \text{ за } y \geqslant 0. \tag{4.6.1}$$

Расподелу дужине периода када је систем заузет није једноставно израчунати. Посматрајмо Слику 8.



Слика 8: Периоди када је систем заузет

Клијент C_1 улази у систем у тренутку τ_1 и почиње период заузетости дужине Y. Током његовог опслуживања (дужине x_1) у систем долазе други клијенти C_2, C_3 и C_4 који продужавају период заузетости. Како је дужина периода заузетости независна од дисциплине опслуживања, може се изабрати last in, first out (LIFO) метода, односно да клијент који дође последњи буде први опслужен. Када се заврши опслуживање C_1 следећи на реду је C_4 који је дошао после C_2 и C_3. За клијенте C_2 и C_3 може се сматрати да су тренутно изван система. Опслуживање C_4 иницира нови потпериод заузетости чију дужину чемо означити са X_4. X_4 обухвата опслуживање C_4 и свих оних који дођу током његовог опслуживања, C_5 и C_6. У тренутку $\tau_1 + x_1 + X_4$ завршен је потпериод и наставља се опслуживање у складу са LIFO методом. Предност над C_2 има C_3 и он генерише нови потпериод током кога ће бити опслужени C_3, C_7, C_8 и C_9 (C_7, C_8 и C_9 су дошли током опслуживања x_3). На крају је на реду C_2 са потпериодом дужине X_2, након чега је систем празан и завршава се главни период заузетости дужине Y.

Са Слике 8 се види да су контуре потпериода идентичне контурама главног периода, само су транслиране на временској оси (за вредност која је једнака суми опслуживања клијената чији потпериоди још нису генерисани). Потпериоди се статистички понашају на исти начин као главни период дужине Y. Случајне величине $\{X_k\}$ су независне и једнако расподељене и имају исту расподелу као дужина главног периода заузетости. Променљива Y је сума $1 + \tilde{v}$ случајне променљиве, где је $\tilde{v}$ број клијената који дође у систем током опслуживања клијента C_1

$$Y = x_1 + X_{\tilde{v}+1} + X_{\tilde{v}} + \ldots + X_3 + X_2.$$

Функција расподеле дужине периода заузетости дефинисана је са $G(y)$, а Лапласова трансформација је облика

$$G^*(s) = \int_0^\infty e^{-sy} dG(y) = E[e^{-sY}].$$

У израчунавању расподеле дужине заузетог периода користићемо условно математичко очекивање. Посматраћемо два услова за Y: дужину опслуживања клијента C_1 и број клијената који дође током тог периода. Из независности X_k следи да је

$$\begin{aligned} E[e^{-sY} \mid x_1 = x, \tilde{v} = k] &= E[e^{-s(x+X_{k+1}+\ldots+X_2)}] \\ &= E[e^{-sx} e^{-sX_{k+1}} \ldots e^{-sX_2}] \\ &= E[e^{-sx}] E[e^{-sX_{k+1}}] \ldots E[e^{-sX_2}] \end{aligned}$$

Вредност x је константа и из особина математичког очекивања важи да је $E[e^{-sx}] = e^{-sx}$. Потпериоди дужине X_k имају функцију расподеле $G(y)$ и Лапласову трансформацију $G^*(s)$. Добија се да је

$$E[e^{-sY} \mid x_1 = x, \tilde{v} = k] = e^{-sx}[G^*(s)]^k.$$

Број долазака $\tilde{v}$ током интервала x има Пуасонову расподелу са параметром λx. Уклонимо услов за $\tilde{v}$

$$\begin{aligned} E[e^{-sY} \mid x_1 = x] &= \sum_{k=0}^{\infty} E[e^{-sY} \mid x_1 = x, \tilde{v} = k] P\{\tilde{v} = k\} \\ &= \sum_{k=0}^{\infty} e^{-sx}[G^*(s)]^k \frac{(\lambda x)^k}{k!} e^{-\lambda x} \\ &= e^{-x(s+\lambda-\lambda G^*(s))}. \end{aligned}$$

На сличан начин може се склонити и други услов за x_1

$$\begin{aligned} G^*(s) &= \int_0^\infty E[e^{-sY} \mid x_1 = x] dB(x) \\ . &= \int_0^\infty e^{-x(s+\lambda-\lambda G^*(s))} dB(x) \end{aligned}$$

Интеграл на десној страни једнакости је Лапласова трансформација расподеле дужине трајања опслуживања у тачки $s + \lambda - \lambda G^*(s)$

$$G^*(s) = B^*(s + \lambda - \lambda G^*(s)). \tag{4.6.2}$$

Ово је главни резултат за расподелу периода заузетости $M|G|1$ система. Дат је у облику Лапласове трансформације коју је најчешће немогуће инвертовати како би се добила

експлицитно расподела. Међутим, за дате вредности s могуће је израчунати $G^*(s)$ помоћу следеће итеративне једнакости

$$G^*_{n+1}(s) = B^*(s + \lambda - \lambda G^*_n(s))$$

где је $0 \leqslant G^*_0(s) \leqslant 1$. За $\rho = \lambda E[x] < 1$ израз ће конвергирати ка $G^*(s)$ и могуће је нумерички доћи до густине расподеле.

Из једнакости (4.6.2) могу се израчунати моменти дужине периода заузетости. Нека је $E[Y^k]$ k-ти моменат расподеле дужине периода заузетости. Наћи ћемо први моменат $E[Y]$ а слично се израчунавају и остали моменти (за било које k). Из особина Лапласове трансформације важи да је

$$E[Y^k] = (-1)^k G^{*(k)}(0) \tag{4.6.3}$$

$$E[x^k] = (-1)^k B^{*(k)}(0). \tag{4.6.4}$$

Директно из израза (4.6.3) добија се

$$\begin{aligned} E[Y] = -G^{*(1)}(0) &= -B^{*(1)}(0)\frac{d}{ds}[s + \lambda - \lambda G^*(s)]\Big|_{s=0} \\ &= -B^{*(1)}(0)[1 - \lambda G^{*(1)}(0)] \\ &= E[x](1 + \lambda E[Y]). \end{aligned}$$

На крају се добија да је просечна дужина периода заузетости за $M|G|1$ системе

$$E[Y] = \frac{E[x]}{1 - \rho}. \tag{4.6.5}$$

Ова вредност зависи само од интензитета улазног потока и просечне вредности дужине опслуживања.

4.7 Број клијената опслужених у периоду када је систем заузет

Нека је N_{zp} број клијената услужених у периоду када је систем заузет. Одговарајући закон расподеле вероватноћа дефинише се као

$$f_n = P\{N_{zp} = n\}. \tag{4.7.1}$$

За расподелу f_n z-трансформација је

$$F(z) = \sum_{n=1}^{\infty} f_n z^n = E[z^{N_{zp}}]. \tag{4.7.2}$$

Члан у суми за $n = 0$ је изостављен јер бар један клијент мора да буде услужен у периоду када је систем заузет. Подсетимо се да променљива $\tilde{v}$ представља број клијената који дођу у систем током периода опслуживања клијента C_{n+1} и да за њену z-трансформацију $V(z)$ важи једнакост (4.3.4)

$$V(z) = B^*(\lambda - \lambda z).$$

У претходном поглављу уведен је појам потпериода периода заузетости система. Сваки долазак у систем током опслуживања клијента C_1 генерише један потпериод. Нека је M_i број клијената услужених у i-том потпериоду. Случајне величине M_i су независне и једнако расподељене са законом расподеле

$$M_i : \begin{pmatrix} 1 & 2 & \dots & k & \dots \\ f_1 & f_2 & \dots & f_k & \dots \end{pmatrix} \text{ за свако } i.$$

Независност величина M_i потиче од независности дужина потпериода X_i. Претпоставимо да је $\tilde{v} = k$ и на основу тога израчунавамо математичко очекивање броја клијената услужених у периоду заузетости

$$\begin{aligned} E[z^{N_{zp}} \mid \tilde{v} = k] &= E[z^{1+M_1+M_2+\ldots+M_k}] \\ &= zE[z^{M_1}]E[z^{M_2}]\ldots E[z^{M_k}] \\ &= z[F(z)]^k. \end{aligned}$$

Ако склонимо услов за број долазака $\tilde{v}$ добија се

$$\begin{aligned} F(z) &= \sum_{k=0}^{\infty} E[z^{N_{zp}} \mid \tilde{v} = k]P\{\tilde{v} = k\} \\ &= z\sum_{k=0}^{\infty}[F(z)]^k P\{\tilde{v} = k\}. \end{aligned}$$

Сума у последњој једнакости је z-трансформација случајне величине $\tilde{v}$ за вредност $F(z)$. Следи да је

$$F(z) = zV[F(z)].$$

На основу једнакости (4.3.4) коначно се добија z-трансфомација броја клијената опслужених у периоду када је систем заузет

$$F(z) = zB^*[\lambda - \lambda F(z)]. \tag{4.7.3}$$

Моменти за број клијената услужених у периоду заузетости израчунавају се помоћу (4.7.3). Означимо са h_k моменат реда k

$$h_k = E[N_{zp}^k].$$

Просечан број клијената опслужених током периода заузетости је

$$\begin{aligned} h_1 = F'(1) &= B^{*(1)}(0)(-\lambda F'(1)) + B^*(0) \\ &= E[x]\lambda h_1 + 1. \end{aligned}$$

Добија се да је

$$h_1 = \frac{1}{1-\rho}.$$

Из анализе периода када је систем заузет може се добити расподела времена које клијент проведе чекајући у реду. Расподела времена чекања зависи од дисциплине опслуживања и не може се користити идеја из пасуса *4.6* о промени редоследа опслуживања клијената. Стога, овде разматрамо само $M|G|1$ системе са FIFO дисциплином опслуживања. Идеја је да се главни период заузетости дужине Y (који започиње доласком клијента C_1) подели на периоде дужина X_k. Први период је X_0, период опслуживања клијента C_1 чија је дужина x_1. Сви клијенти који су дошли током X_0 биће опслужени у току следећег периода дужине X_1. Трајање X_1 једнако је времену потребном за опслуживање клијената који су дошли током X_0. Када истекне интервал у трајању X_1 започиње следећи период дужине X_2 у коме ће бити опслужени сви који су дошли током времена дужине X_1 и тако се процес наставља. Заправо, X_i је дужина временског периода у коме ће се опслужити сви клијенти који су дошли током претходног периода дужине X_{i-1}. Укупна дужина периода заузетости Y је

$$Y = \sum_{i=0}^{\infty} X_i.$$

Ако је $\rho < 1$ систем у неком тренутку достиже еквилибријум и број периода X_i је коначан, односно сервер није константно заузет.

Нека је $N_{X_i}(y)$ функција расподеле за X_i

$$N_{X_i}(y) = P\{X_i \leqslant y\}, \text{ за } y \geqslant 0$$

и одговарајућа Лапласова трансформација

$$X_i^*(s) = \int_0^\infty e^{-sy} dN_{X_i}(y) = E[e^{-sX_i}].$$

Нека је N_i број клијената који дођу у систем током периода дужине X_i. Случајна величина X_{i+1} представља суму N_i трајања периода опслуживања, при чему свака дужина трајања опслуживања има расподелу $B(x)$.

Потребно је пронаћи рекурентну формулу за $X_i^*(s)$ користећи условно математичко очекивање. Претпоставимо дужину периода X_{i-1} и број долазака N_{i-1} током тог периода

$$E[e^{-sX_i} \mid X_{i-1} = y, N_{i-1} = n] = [B^*(s)]^n.$$

Последња једнакост произилази из особине конволуције Лапласове трансформације и чињенице да су дужине трајања опслуживања независне и једнако расподељене. Прво склонимо услов за N_{i-1}

$$\begin{aligned} E[e^{-sX_i} \mid X_{i-1} = y] &= \sum_{n=0}^{\infty} P\{N_{i-1} = n\}[B^*(s)]^n \\ &= \sum_{n=0}^{\infty} \frac{(\lambda y)^n}{n!} e^{-\lambda y}[B^*(s)]^n, \end{aligned}$$

потом и за y

$$\begin{aligned} E[e^{-sX_i}] &= \int_0^\infty \sum_{n=0}^{\infty} \frac{(\lambda y)^n}{n!} e^{-\lambda y}[B^*(s)]^n dN_{X_{i-1}}(y) \\ &= \int_0^\infty e^{-(\lambda - \lambda B^*(s))y} dN_{X_{i-1}}(y). \end{aligned}$$

Из последњег израза добија се веза између Лапласове трансформације случајних величина X_i и X_{i-1}

$$X_i^*(s) = X_{i-1}^*(\lambda - \lambda B^*(s)). \tag{4.7.4}$$

Претпоставимо да је нови клијент дошао у систем током периода заузетости дужине Y, конкретно током интервала дужине X_i. Нека је $\tilde{w}$ време које проведе чекајући у систему. То време се састоји од преосталог времена опслуживања у i-том интервалу и времена опслуживања свих клијената који су дошли пре њега. Израчунаћемо трансформацију $E[e^{-s\tilde{w}} \mid i]$ времена чекања клијента који је дошао током периода дужине X_i помоћу условног математичког очекивања. Означимо са Y_i преостало време периода дужине X_i након доласка посматраног клијента и са N_i број клијената који су дошли током X_i али пре посматраног клијента. Из независности и особине конволуције следи да је

$$E[e^{-s\tilde{w}} \mid i, X_i = y, Y_i = y', N_i = n] = [B^*(s)]^n e^{-sy'}.$$

Улазни поток је Пуасонов процес, и n клијената је дошло у интервалу дужине $y - y'$, па важи

$$\begin{aligned} E[e^{-s\tilde{w}} \mid i, X_i = y, Y_i = y'] &= \sum_{n=0}^{\infty} \frac{(\lambda(y-y'))^n}{n!} e^{-\lambda(y-y')}[B^*(s)]^n e^{-sy'} \\ &= e^{\lambda(y-y')B^*(s)} e^{-\lambda(y-y')} e^{-sy'} \\ &= e^{\lambda(y-y')B^*(s) - \lambda(y-y') - sy'}. \end{aligned}$$

Заједничка расподела периода X_i и Y_i добија се помоћу

$$P\{y' < Y \leqslant y + dy', y < X_i \leqslant y + dN_{X_i}\} = \frac{dN_{X_i}(y)dy'}{E[X_i]}$$

за $0 \leqslant y' \leqslant y \leqslant +\infty$ (извођење ове заједничке расподеле може се наћи у [1]). Заменом ове вероватноће и склањањем услова у изразу за $E[e^{-s\tilde{w}}]$ добија се

$$\begin{aligned} E[e^{-s\tilde{w}} \mid i] &= \int_{y=0}^{\infty}\int_{y'=0}^{y} e^{\lambda(y-y')B^*(s)-\lambda(y-y')-sy'}dN_{X_i}(y)dy'/E[X_i] \\ &= \frac{1}{(-s+\lambda-\lambda B^*(s))E[X_i]}\int_{y=0}^{\infty}(e^{-ys}-e^{-y(\lambda-\lambda B^*(s))})dN_{X_i}(y) \\ &= \frac{X_i^*(s)-X_i^*(\lambda-\lambda B^*(s))}{(-s+\lambda-\lambda B^*(s))E[X_i]}. \end{aligned}$$

Коришћењем рекурентне формуле (4.7.4) за $X_i^*(s)$ последњи израз постаје

$$E[e^{-s\tilde{w}} \mid i] = \frac{X_{i+1}^*(s)-X_i^*(s)}{(s-\lambda+\lambda B^*(s))E[X_i]}.$$

Вероватноћа да је клијент дошао баш током i-тог интервала једнака је $E[X_i]/E[Y]$. Ако претпоставимо да је клијент дошао било када током периода заузетости дужине Y добија се да је

$$\begin{aligned} E[e^{-s\tilde{w}} \mid \text{долазак током } Y] &= \sum_{i=0}^{\infty} E[e^{-s\tilde{w}} \mid i]\frac{E[X_i]}{E[Y]} \\ &= \frac{1}{(s-\lambda+\lambda B^*(s))E[Y]}\sum_{i=0}^{\infty}(X_{i+1}^*(s)-X_i^*(s)). \end{aligned}$$

Како систем достиже еквилибријум сума $\sum_{i=0}^{\infty}(X_{i+1}^*(s)-X_i^*(s))$ се своди на $1-X_0^*(s)$. X_0 је време опслуживања првог клијента и износи x_1, па је $X_0^*(s) = B^*(s)$ и

$$E[e^{-s\tilde{w}} \mid \text{долазак током } Y] = \frac{1-B^*(s)}{(s-\lambda+\lambda B^*(s))E[Y]}.$$

Просечна дужина периода заузетости израчуната је у поглављу *4.6* (једнакост (4.6.5)) и износи $E[Y] = E[x]/(1-\rho)$. Вероватноћа да ће клијент доћи у систем у периоду када је сервер заузет је приближно $\rho = \lambda E[x]$. Коначно се добија да је

$$\begin{aligned} E[e^{-s\tilde{w}}] =& \rho E[e^{-s\tilde{w}} \mid \text{долазак у периоду када је сервер заузет}] \\ &+ (1-\rho)E[e^{-s\tilde{w}} \mid \text{долазак у периоду када је сервер слободан}] \\ &= \rho\frac{(1-B^*(s))(1-\rho)}{(s-\lambda+\lambda B^*(s))E[x]} + (1-\rho)E[e^0] \\ &= \lambda\frac{(1-B^*(s))(1-\rho)}{s-\lambda+\lambda B^*(s)} + (1-\rho) \\ &= \frac{s(1-\rho)}{s-\lambda+\lambda B^*(s)}. \end{aligned} \tag{4.7.5}$$

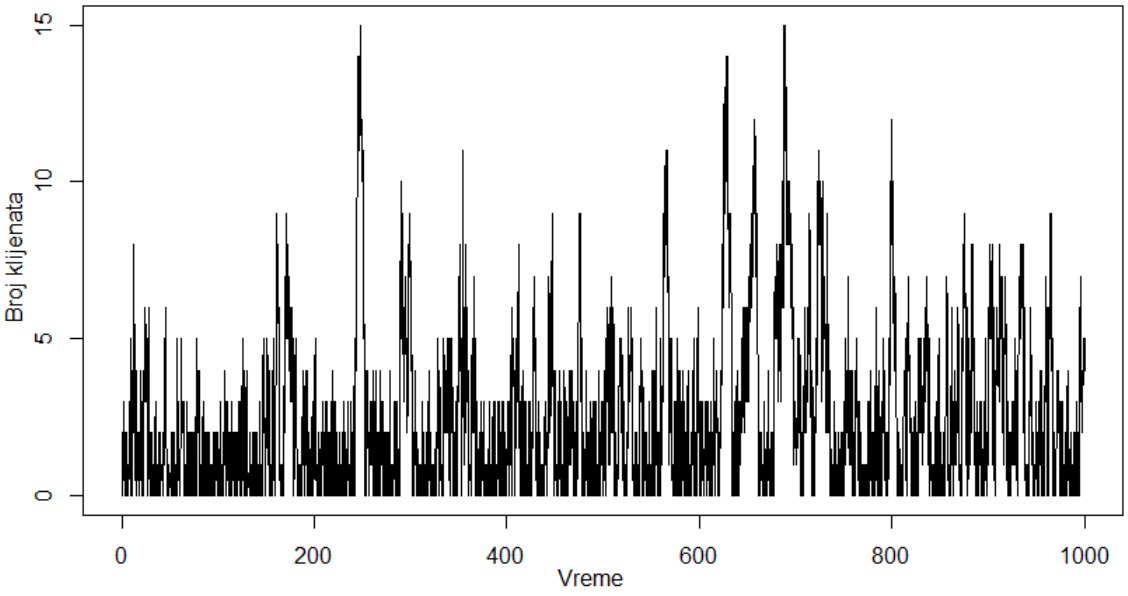
Ово је заправо трећа Pollaczek-Khintchin-ова једнакост за дужину чекања клијента у систему дата формулом (4.5.4), коју смо раније израчунали. Овим је показано да се до расподеле времена чекања може доћи и преко анализе периода када је сервер заузет.

4.8 Симулација

У наредном делу биће приказана симулација система код кога је улазни поток Пуасонов процес, расподела опслуживања експоненцијална, капацитет бесконачан и опслуживање врши један сервер ($M|M|1|\infty$ систем). Систем функционише на следећи начин: клијент уђе у систем, чека у реду ако је сервер заузет, опслужује се и напушта систем. При томе се подразумева да клијенти не напуштају систем пре него што буду комплетно опслужени (нема одустајања), сваки клијент који уђе у систем мора бити опслужен, сервер ради непрекидно, без пауза и нема ограничења у броју клијената у реду.

Најпре је потребно одредити трајање симулације и временске јединице. Да би се постигло равнотежно стање система потребно је да се симулира на довољно великом временском интервалу. У овом случају је то 1000 сати и мерна јединица су сати. Претпоставимо да симулирамо рад неког система, при чему је интензитет улазног потока 3 клијента по сату, а интензитет опслуживања просечно 4 клијента по сату. Клијенти се опслужују у складу са FIFO дисциплином без приоритета међу клијентима. Упоредићемо теоријске и резултате добијене на основу симулације за просечан број клијената у реду и просечно време чекања и тиме проверити исправност симулације.

На Слици 9 приказано је стање система кроз време (број клијената у систему).



Слика 9: Симулација $M|M|1$ система

Систем достиже еквилибријум уколико је искоришћеност система $\rho < 1$. И то је у овом случају испуњено јер је

$$\rho = \frac{\lambda}{\mu} = \frac{3}{4}.$$

Ако случајна величина X представља дужину опслуживања, теоријски резултати за просечно време чекања и просечан број клијената у реду су

$$W_q = \frac{\lambda E[X^2]}{2(1-\rho)} = \frac{3 \cdot \frac{2}{4^2}}{2\left(1-\frac{3}{4}\right)} = 0.75 \text{ сата} = 45 \text{ мин},$$

$$L_q = \frac{\lambda^2 E[X^2]}{2(1-\rho)} = \frac{3^2 \cdot \frac{2}{4^2}}{2\left(1-\frac{3}{4}\right)} = 2.25 \text{ клијената}.$$

Након понављања симулације довољан број пута, добијени резултати су

$$W_q \approx 0.78 \text{ сата} = 46.8 \text{ мин},$$
$$L_q \approx 2.31 \text{ клијент}.$$

Добијени резултати су приближни теоријским вредностима и то иде у прилог исправности модела. Уколико је дужина трајања симулације још већа добијају се све бољи резултати, приближнији теоријским вредностима. Такође, са слике 9 уочава се да се вредности броја клијената групишу око вредности 2.5. У наставку је дат код у програмском језику R којим је одрађена симулација $M|M|1$ система.

```
lambda <- 3 # intenzitet ulaznog potoka
mi <- 4 # intenzitet opsluzivanja
T <- 1000 # trajanje simulacije
t <- 0 # pocetno vreme
red <- 0 # prazan red na pocetku
s <- 0 # broj klijenata u redu
t1 <- rexp(1,lambda) # prvi dolazak u sistem
trenutnired <- 1
vremedogadjaja <- t1
t <- t1
dogadjaj <- 1 # ukupan broj dogadjaja
while (t<T)
{
   dogadjaj <- dogadjaj+1
   if(trenutnired >0) # u redu ima klijenata
   {
      t1 <- rexp(1,lambda+mi) # vreme do sledeceg dogadjaja
      p <- runif(1,0,1)
      red[dogadjaj] <- trenutnired
      trenutnired <- ifelse(p<lambda/(lambda+mi),
      trenutnired+1, # dolazak
      trenutnired-1) # odlazak
   }
   else
   {
      t1 <- rexp(1,lambda) # prazan red
      red[dogadjaj] <- trenutnired
      trenutnired <- 1
   }
   t <- t+t1
   vremedogadjaja[dogadjaj] <- t1
   s <- s+t1*red[dogadjaj]
}
# grafik
plot(cumsum(vremedogadjaja),red,type="l",xlab="Vreme",ylab="Broj klijenata",
main=paste("M/M/1"))
lq=s/t # prosecna duzina
```

5 $M|G|1$ системи са приоритетним опслуживањем

Дисциплина опслуживања представља начин избора клијента који ће бити следећи услужен. Та одлука зависи од времена доласка клијената у систем, дужине опслуживања или од припадности клијената некој групи. Уколико редослед осплуживања зависи од припадности клијената одређеној групи реч је о *системима са приоритетним опслуживањем*. Приоритет се може посматрати као начин да се минимизују трошкови и губици на рачун кашњења одређених клијената.

У тренутку доласка клијента у систем израчунава се функција приоритета и на основу њене вредности одређује се којој групи клијент припада. Обично се виши приоритет означава мањим бројем, тако да је група означена са 1 највишег приоритета. Клијенти са вишим приоритетом се први опслужују у односу на оне са нижим приоритетом без обзира када су стигли у систем. Постоје два типа приоритета: релативни *(nonpreemptive)* и апсолутни *(preemptive)*. У случају апсолутног приоритета, клијент који затекне све линије заузете прекида опслуживање клијента нижег приоритета ако такав постоји на опслуживању. Овде се разликују три случаја, клијент чије је опслуживање прекинуто наставља са опслуживањем *(resume)*, или креће поново из почетка, или добија отказ, тј. мора трајно да напусти систем, при чему његово опслуживање неће бити комплетирано. Код релативног приоритета клијент са вишим приоритетом не прекида опслуживање клијента са нижим приоритетом, већ само стаје у ред испред клијената нижег приоритета. При томе, клијенти истог приоритета обично формирају ред и унутар тог реда опслужују се по FIFO дисциплини.

Системи са приоритетом имају велику примену али их је много теже моделирати него оне без приоритета. Многи резулатати које смо већ добили не зависе од дисциплине опслуживања. До промена долази у расподели времена које клијент проведе чекајући. Израчунаћемо основне перформансе система (просечно време које клијент проведе у систему, просечно време чекања, просечан број клијената у реду и у систему) за $M|G|1$ системе са релативним и апсолутним приоритетом.

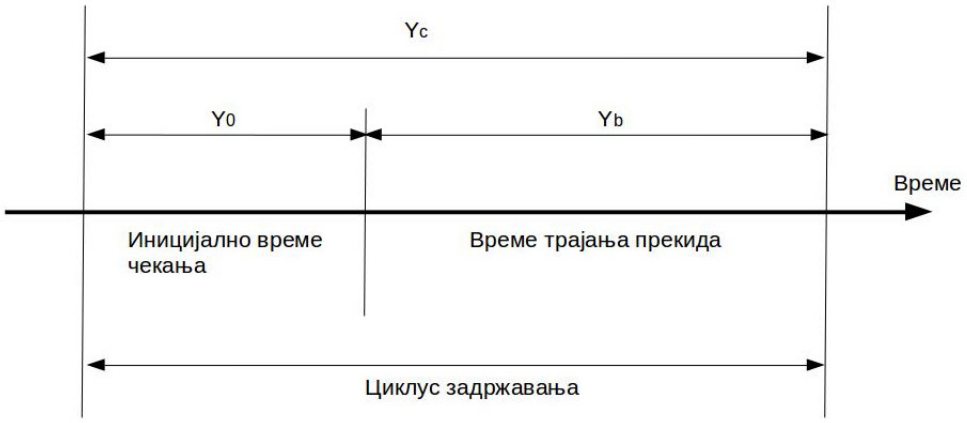
Претпоставимо да клијенти долазе у систем у складу са Пуасоновим процесом и да постоји r различитих група приоритета обележених бројевима од 1 до r. Клијенти означени бројем 1 имају највиши приоритет, док су они са ознаком r најнижег приоритета. Нека је λ_i интензитет доласка у систем клијената који припадају i-тој класи. Дужина трајања опслуживања клијента који има приоритет i је случајна величина X_i која има неодређену функцију расподеле вероватноћа $B_i(x)$.

Случајна величина T_i^q представља време које клијент приоритета i проведе чекајући у реду, док је T_i укупно време које проведе у систему. Просечно време чекања у реду клијента групе i је $W_i^q = E[T_i^q]$ и просечно време проведено у систему је $W_i = E[T_i]$. Случајне величине N_i и N_i^q представљају број клијената приоритета i у систему, односно у реду. Просечан број клијената приоритета i у систему је $L_i = E[N_i]$ и просечан број клијената у реду је $L_i^q = E[N_i^q]$.

5.1 Циклуси задржавања, уопштени периоди заузетости и расподела дужине чекања

Циклуси задржавања *(delay cycles)* састоје се од иницијалног времена чекања клијента у реду и од укупног времена трајања прекида током започетог опслуживања клијента. На основу њих се израчунава Лапласова трансформација густине уопштеног периода заузетости, као и Лапласова трансформација расподеле дужине чекања.

Нека је Y_c циклус задржавања, иницијално време чекања Y_0 и Y_b време трајања прекида током започетог опслуживања. На Слици 10 је приказан један циклус задржавања.



Слика 10: Циклус задржавања

Функције расподеле вероватноћа случајних величина Y_0, Y_b, Y_c су редом $G_0(y), G_b(y)$, $G_c(y)$ и одговарајуће Лапласове трансформације $G_0^*(s), G_b^*(s)$ и $G_c^*(s)$. Иницијално време чекања у реду састоји се од преосталог времена опслуживања клијента који се тренутно опслужује и времена потребног за опслуживање свих клијената који ће се опслужити пре посматраног клијента. Ова вредност се најчешће може израчунати и користи се у изразима за $G_b^*(s)$ и $G_c^*(s)$. Период Y_b чине сва времена опслуживања клијената вишег приоритета који су дошли током опслуживања посматраног клијента и прекинули његово опслуживање. Поступак израчунавања вредности $G_b^*(s)$ и $G_c^*(s)$ је сличан као у поглављу *4.6*, где је одређена расподела периода заузетости сервера коришћењем потпериода заузетости. Нека је N_0 број клијената који дођу у систем током иницијалног времена чекања Y_0. Користећи условно математичко очекивање добија се

$$E[e^{-sY_b} \mid Y_0 = y, N_0 = n] = [G^*(s)]^n,$$

где је $G^*(s)$ Лапласова трансформација густине периода заузетости. Последња једнакост следи из чињенице да су потпериоди независни и имају исту расподелу као главни период заузетости. Уклонимо услов за N_0

$$\begin{aligned} E[e^{-sY_b} \mid Y_0 = y] &= \sum_{n=0}^{\infty} [G^*(s)]^n \frac{(\lambda y)^n}{n!} e^{-\lambda y} \\ &= e^{-(\lambda - \lambda G^*(s))y} \end{aligned}$$

потом и за Y_0

$$G_b^*(s) = E[e^{-sY_b}] = \int_0^{\infty} e^{-(\lambda - \lambda G^*(s))y} dG_0(y).$$

Последња једнакост је Лапласова трансформација густине случајне величине Y_0 у тачки $\lambda - \lambda G^*(s)$. Расподела времена трајања прекида (уопштеног периода заузетости) дата је са

$$G_b^*(s) = G_0^*(\lambda - \lambda G^*(s)). \tag{5.1.1}$$

За израчунавање $G_c^*(s)$ користе се услови за N_0 и Y_0

$$E[e^{-sY_c} \mid Y_0 = y, N_0 = n] = e^{-sy}[G^*(s)]^n$$

одакле следи да је

$$G_c^*(s) = E[e^{-sY_c}] = \int_0^\infty e^{-sy} \sum_{n=0}^{\infty} [G^*(s)]^n \frac{(\lambda y)^n}{n!} e^{-\lambda y} dG_0(y)$$
$$= \int_0^\infty e^{-(s+\lambda-\lambda G^*(s))}.$$

Лапласова трансформација расподеле дужине чекања клијента дата је са

$$G_c^*(s) = G_0^*(s + \lambda - \lambda G^*(s)). \tag{5.1.2}$$

За израчунавање $G^*(s)$ користи се једнакост (4.6.2), $G^*(s) = B^*(s + \lambda - \lambda G^*(s))$.

5.2 Закони одржања *(Conservation laws)*

Системе које клијенти не напуштају пре него што буду комплетно опслужени, као и у којима сервер није неактиван ако има клијената у њему, називамо „*конзервативним*". На овакве системе могу се применити закони одржања. Заправо, док год је дисциплина опслуживања независна од дужине опслуживања, расподела броја клијената у систему и просечно време чекања су инваријантни у односу на редослед опслуживања.

Нека је случајна величина q_n, дефинисана у поглављу *4.2* као број клијената који остану у систему након одласка клијента C_n, у овом случају број клијената који остану у систему након одласка n-тог клијента. Слично је и v_n број клијената који дођу у систем током опслуживања n-тог клијента. На овај начин дозвољена је произвољна дисциплина опслуживања и једнакост (4.3.1) и даље важи. Аналогно као у поглављу *4.4* добија се да Pollaczek-Khinchin-ова једнакост за расподелу броја клијената у систему важи за било коју дисциплину опслуживања независну од дужине опслуживања

$$Q(z) = B^*(\lambda - \lambda z)\frac{(1-\rho)(1-z)}{B^*(\lambda-\lambda z) - z}.$$

Овим је показан први део тврђења закона одржања.

За просечно време чекања система са релативним приоритетом у коме дисциплина опслуживања не зависи од времена опслуживања важи да је

$$\sum_{i=0}^{r} \rho_i W_i^q = \begin{cases} \frac{\rho \sum_{j=1}^{r} \lambda_j E[X_j^2]}{2(1-\rho)}, & \rho < 1 \\ \infty, & \rho \geqslant 1 \end{cases}. \tag{5.2.1}$$

Наиме, помножена сума дужине чекања клијената неће се никада променити. Ако је у тренутку t $N_i(t)$ број клијената приоритета i у реду и ако j-ти од њих ($j = 1, 2, \ldots, N_i(t)$) има дужину опслуживања X_{ij} и ако је X_0 преостало време опслуживања клијента који се у тренутку t опслужује, тада је без обзира на редослед опслуживања незавршени посао $U(t)$

$$U(t) = X_0 + \sum_{i=1}^{r} \sum_{j=1}^{N_i(t)} X_{ij}.$$

Нађимо математичко очекивање претходног израза

$$E[U(t)] = E[X_0] + \sum_{i=1}^{r} \sum_{n_i=0}^{\infty} P\{N_i(t) = n_i\} \sum_{j=1}^{n_i} E[X_{ij}].$$

Када $t \to \infty$, $E[X_{ij}]$ не зависи од j и

$$\lim_{t\to\infty} E[U(t)] = E[X_0] + \lim_{t\to\infty} \sum_{i=1}^{r} \sum_{n_i=0}^{\infty} P\{N_i(t) = n_i\} n_i E[X_i]$$
$$= E[X_0] + \sum_{i=1}^{r} E[X_i]E[N_i].$$

Из Little-овог закона важи да је $E[N_i] = \lambda_i W_i^q$ и

$$\lim_{t\to\infty} E[U(t)] = E[X_0] + \sum_{i=1}^{r} \rho_i W_i^q. \tag{5.2.2}$$

Код система са Пуасоновим улазним потоком незавршени посао у тренутку t може се посматрати као просечно време чекања клијената W^q. Из Pollaczek-Khinchin-ове формуле следи да је $W^q = \dfrac{\lambda E[X^2]}{2(1-\rho)}$. Други моменат дужине опслуживања је

$$E[X^2] = \sum_{i=1}^{r} \frac{\lambda_i}{\lambda} E[X_i^2].$$

Коришћењем израза (5.2.2) за $\lim\limits_{t\to\infty} E[U(t)]$ и $E[X^2]$ добија се да је

$$\sum_{i=0}^{r} \rho_i W_i^q = \frac{\rho \sum\limits_{j=1}^{r} \lambda_j E[X_j^2]}{2(1-\rho)}. \tag{5.2.3}$$

Из закона одржања следи да смањивање дужине једног времена чекања утиче на повећање другог и обратно, како би сума остала непромењена. Специјални случај када је просечна дужина опслуживања једнака за све класе приоритета, закон одржања за $\rho < 1$ има следећи облик

$$\sum_{i=1}^{r} \lambda_i W_i^q = \frac{\lambda \sum\limits_{j=1}^{r} \lambda_j E[X_j^2]}{2(1-\rho)}.$$

Из Little-овог закона важи да је $\lambda_i W_i^q = E[N_i^q]$ и

$$\sum_{i=1}^{r} E[N_i^q] = \frac{\lambda \sum\limits_{j=1}^{r} \lambda_j E[X_j^2]}{2(1-\rho)}.$$

Одавде следи да је просечан број свих клијената у реду и просечно време чекања константно када је просечна дужина опслуживања иста за све класе

$$E[N^q] = \frac{\lambda \sum\limits_{j=1}^{r} \lambda_j E[X_j^2]}{2(1-\rho)}$$
$$W^q = \frac{\sum\limits_{j=1}^{r} \lambda_j E[X_j^2]}{2(1-\rho)}.$$

Заправо, ове мере не зависе од дисциплине опслуживања.

5.3 Опслуживање са релативним приоритетом

Време које клијент утроши чекајући састоји се од три дела: времена које је потребно да се заврши опслуживање клијента кога је у тренутку свог доласка затекао на опслуживању, времена потребног да се опслуже сви клијенти који су испред њега у реду и времена да се опслуже клијенти вишег приоритета који дођу након његовог доласка.

Нека је у систем дошао клијент приоритета i. Просечно преостало време опслуживања клијента који има приоритет j и који се у тренутку доласка посматраног клијента опслужује износи $\frac{E[X_j^2]}{2E[X_j]}$. Задржавање клијента узроковано овом ситуацијом је

$$\sum_{j=1}^{r} \lambda_j E[X_j] \frac{E[X_j^2]}{2E[X_j]} = \frac{1}{2} \sum_{j=1}^{r} \lambda_j E[X_j^2].$$

Нека је N_j број клијената приоритета j који су у реду испред клијента приоритета i и који ће се опслужити пре њега. Просечно време потребно да се заврши опслуживање таквих клијената је

$$\sum_{j=1}^{r} E[X_j]E[N_j].$$

Трећи део времена које клијент проведе чекајући чини опслуживање клијената вишег приоритета који су дошли након њега M_j и који морају бити опслужени пре њега. То време износи

$$\sum_{j=1}^{r} E[X_j]E[M_j].$$

Време чекања у реду клијената приоритета i је

$$\begin{aligned} W_i^q &= \frac{1}{2} \sum_{j=1}^{r} \lambda_j E[X_j^2] + \sum_{j=1}^{r} E[X_j]E[N_j] + \sum_{j=1}^{r} E[X_j]E[M_j] \\ &= \frac{1}{2} \sum_{j=1}^{r} \lambda_j E[X_j^2] + \sum_{j=1}^{r} E[X_j](E[N_j] + E[M_j]). \end{aligned} \tag{5.3.1}$$

Како се у оквиру истог приоритета опслуживање врши у складу са FIFO дисциплином, клијенти нижег приоритета не утичу на чекање посматраног клијента приоритета i и важи да је

$$\begin{aligned} E[N_j] &= 0 \text{ за } j = i+1, i+2, \dots r \\ E[M_j] &= 0 \text{ за } j = i, i+1, \dots r. \end{aligned}$$

Сви клијенти приоритета i и вишег који чекају у реду морају бити опслужени пре посматраног клијента. Из Little-овог закона следи да је просечан број таквих клијената

$$E[N_j] = \lambda_j W_j^q$$

за свако $j = 1, 2, \dots i$. Слично, сви клијенти вишег приоритета који дођу док посматрани клијент чека биће услужени пре њега. Доласци клијената различитог приоритета су међусобно независни и просечно долази $\lambda_j W_i^q$ клијената приоритета j током периода чекања посматраног клијента

$$E[M_j] = \lambda_j W_i^q \text{ за } j = 1, 2, \dots i-1.$$

Заменом вредности у формулу (5.3.1) добија се

$$W_i^q = \frac{1}{2}\sum_{j=1}^{r}\lambda_j E[X_j^2] + \sum_{j=1}^{i} E[X_j]\lambda_j W_j^q + \sum_{j=1}^{i-1} E[X_j]\lambda_j W_i^q \quad \text{за } j = 1, 2, \dots r.$$

Решавањем једначине добија се да је

$$\begin{aligned} W_i^q\Big(1 - \sum_{j=1}^{i} E[X_j]\lambda_j\Big) &= \frac{1}{2}\sum_{j=1}^{r}\lambda_j E[X_j^2] + \sum_{j=1}^{i-1} E[X_j]\lambda_j W_j^q \\ &= W_{i-1}^q - W_{i-1}^q \sum_{j=1}^{i-2} E[X_j]\lambda_j \\ &= W_{i-1}^q\Big(1 - \sum_{j=1}^{i-2} E[X_j]\lambda_j\Big). \end{aligned}$$

Искористимо чињеницу да је $\rho_j = E[X_j]\lambda_j$ (ρ_j је искоришћеност система, односно заузетост сервера клијентима приоритета j.) Помножимо леву и десну страну претходне једнакости са $1 - \sum_{j=1}^{i-1}\rho_j$

$$\Big(1 - \sum_{j=1}^{i}\rho_j\Big)\Big(1 - \sum_{j=1}^{i-1}\rho_j\Big)W_i^q = \Big(1 - \sum_{j=1}^{i-1}\rho_j\Big)\Big(1 - \sum_{j=1}^{i-2}\rho_j\Big)W_{i-1}^q.$$

Овим је дата веза између узастопних величина W_i^q. Итеративним поступком добија се да је

$$\Big(1 - \sum_{j=1}^{i}\rho_j\Big)\Big(1 - \sum_{j=1}^{i-1}\rho_j\Big)W_i^q = (1 - \rho_1)W_1^q. \tag{5.3.2}$$

Просечно време чекања клијента највишег приоритета једнако је времену потребном да се заврши опслуживање тренутног клијента и свих клијената његовог приоритета који већ чекају у реду, тј

$$W_1^q = \lambda_1 W_1^q E[X_1] + \frac{1}{2}\sum_{j=1}^{r}\lambda_j E[X_j^2],$$

и

$$W_1^q = \frac{\sum_{j=1}^{r}\lambda_j E[X_j^2]}{2(1 - \rho_1)}.$$

Када резултат за W_1^q уврстимо у једнакост (5.3.2) добија се да је просечно време које клијент i-тог приоритета ($i = 1, 2, \dots r$) проведе чекајући у реду

$$W_i^q = \frac{\sum_{j=1}^{r}\lambda_j E[X_j^2]}{2\Big(1 - \sum_{j=1}^{i}\rho_j\Big)\Big(1 - \sum_{j=1}^{i-1}\rho_j\Big)}. \tag{5.3.3}$$

На основу (5.3.3) и Little-овог закона израчунава се просечан број клијената i-тог приоритета ($i = 1, 2, \dots, r$) у реду L_i^q, и у читавом систему L_i и просечно време које

проведу у систему W_i

$$L_i^q = \frac{\sum_{j=1}^{r} \lambda_j E[X_j^2]}{2\Big(1-\sum_{j=1}^{i}\rho_j\Big)\Big(1-\sum_{j=1}^{i-1}\rho_j\Big)},$$

$$L_i = \frac{\lambda_i \sum_{j=1}^{r} \lambda_j E[X_j^2]}{2\Big(1-\sum_{j=1}^{i}\rho_j\Big)\Big(1-\sum_{j=1}^{i-1}\rho_j\Big)} + \lambda_i E[X_i],$$

$$W_i = \frac{\sum_{j=1}^{r} \lambda_j E[X_j^2]}{2\Big(1-\sum_{j=1}^{i}\rho_j\Big)\Big(1-\sum_{j=1}^{i-1}\rho_j\Big)} + E[X_i].$$

Најчешће је тешко израчунати преостало време опслуживања сем у случају када је расподела времена опслуживања експоненцијална са истим параметром за све класе приоритета. Просечно време чекања за клијенте вишег приоритета је мање у односу на клијенте нижег приоритета, али је на нивоу система дуже време чекања него код система без приоритетног опслуживања. То се може контролисати тако што се клијентима са краћим временом опслуживања да виши приоритет и укупно време чекања се смањује. Код система са релативним приоритетом чекање у реду за клијенте вишег приоритета зависи и од клијената нижег приоритета (јер се опслуживање не прекида) и може се рећи да класе нижег приоритета нису невидљиве за класе вишег приоритета.

Помоћу циклуса задржавања може се одредити расподела дужине чекања клијента за сваку класу приоритета. Нека је W_i^* Лапласова трансформација расподеле дужине чекања клијента приоритета i [12]

$$W_i^*(s) = \frac{(1-\rho)(s+\lambda_H-\lambda_H G_H^*(s)) + \lambda_L(1-B_L^*(s+\lambda_H-\lambda_H G_H^*(s)))}{s-\lambda_i+\lambda_i B_i^*(s+\lambda_H-\lambda_H G_H^*(s))} \tag{5.3.4}$$

где су индексом L означене мере везане за класе нижег приоритета, са H вишег приоритета од посматраног клијента приоритета i и важи да је

$\lambda_H = \sum_{j=1}^{i-1} \lambda_j$ - интензитет доласка клијената вишег приоритета;

$\lambda_L = \sum_{j=i+1}^{r} \lambda_j$ - интензитет доласка клијената нижег приоритета;

$B_H^*(s) = \sum_{j=1}^{i-1} \frac{\lambda_j}{\lambda_H} B_j^*(s)$ - Лапласова трансформација расподеле дужине опслуживања клијената вишег приоритета;

$B_L^*(s) = \sum_{j=i+1}^{r} \frac{\lambda_j}{\lambda_L} B_j^*(s)$ - Лапласова трансформација расподеле дужине опслуживања клијената нижег приоритета;

$G_H^*(s) = B_H^*(s+\lambda_H-\lambda_H G_H^*(s))$ - однос Лапласових трансформација расподеле дужине периода заузетости и расподеле дужине опслуживања клијената вишег приоритета;

[12] Резултат је преузет из књиге [2] без доказа са истом логиком ознака као у књизи.

$B_j^*(s)$ - Лапласова трансформација дужине опслуживања j-те групе приоритета.

Уколико је у питању систем без приоритетног опслуживања, односно $r = 1$, тада је $\lambda_H = \lambda_L = B_H^*(s) = B_L^*(s) = G_H^*(s) = 0$ и $\lambda_i = \lambda$ и израз (5.3.4) постаје Pollaczek-Khinchin једнакост (4.5.4), што је и очекивано.

5.4 Системи са релативним приоритетом и SPTF дисциплином

Један од најједноставнијих облика система са приоритетом је онај код кога предност има клијент са најкраћим временом опслуживања, познат под називом Shortest Processing Time First (SPTF) систем. Функција приоритета зависи од времена опслуживања. Ако имамо два клијента, један са просечним временом опслуживања x, други са временом опслуживања y и $y > x$, просечно време чекања у читавом систему је мање (у односу на FIFO дисциплину опслуживања) ако се прво опслужи први клијент.

Претпоставимо да је у $M|G|1$ систему интензитет доласка клијената λ, расподела опслуживања $B(x)$, функција густине опслуживања $b(x)$ и да се дужина опслуживања клијента може одредити одмах по уласку у систем како би клијенти могли да буду распоређени у складу са SPTF дисциплином. Нека је у систем ушао клијент чије је време опслуживања x. Он стаје у ред испред свих клијената са временом опслуживања већим од x а иза свих оних са временом опслуживања мањим од x. Клијенти чије је време опслуживања у интервалу $(x, x + dx)$ припадају истој класи приоритета и њихово просечно време чекања може се израчунати на основу формуле (5.3.3)

$$W_i^q = \frac{\sum_{j=1}^{r} \lambda_j E[X_j^2]}{2\Big(1 - \sum_{j=1}^{i} \rho_j\Big)\Big(1 - \sum_{j=1}^{i-1} \rho_j\Big)}.$$

Интензитет доласка оваквих клијената у систем је $\lambda dB(x)/dx = \lambda b(x)$ и искоришћеност система је $\rho(x) = \lambda b(x)x$. Сума $\sum_{j=1}^{i} \rho_j$ понаша се као $\int_0^{x^+} \rho(y)dy$. Просечно време чекања у систему са SPTF дисциплином опслуживања за клијенте чија је дужина опслуживања x дато је са

$$W(x) = \frac{\sum_{j=1}^{r} \lambda_j E[X_j^2]}{2\Big(1 - \lambda \int_0^{x^-} yb(y)dy\Big)\Big(1 - \lambda \int_0^{x^+} yb(y)dy\Big)}. \tag{5.4.1}$$

Када је $B(x)$ апсолутно непрекидна функција израз за $W(x)$ постаје

$$W(x) = \frac{\sum_{j=1}^{r} \lambda_j E[X_j^2]}{2\Big(1 - \lambda \int_0^{x} yb(y)dy\Big)^2}.$$

Често се поставља питање на који начин доделити приоритет клијентима и који су неопходни параметри за одређивање функције приоритета. Најчешће је главни задатак оптимизовати систем и смањити трошкове. Нека је D_i губитак у одабраној валути који претрпи систем за сваку секунду (или неку другу временску јединицу) кашњења сваког

клијента приоритета i. Просечни трошак по секунди на нивоу читавог система који има r класа приоритета је

$$D = \sum_{i=1}^{r} D_i L_i$$

где је L_i просечан број клијената приоритета i у систему. Из Little-овог закона важи да је без обзира на дисциплину опслуживања $L_i = \lambda_i W_i = \lambda_i (W_i^q + E[X_i])$ и

$$D = \sum_{i=1}^{r} D_i \rho_i + \sum_{i=1}^{r} D_i \lambda_i W_i^q. \tag{5.4.2}$$

Потребно је наћи дисциплину опслуживања која минимизује D за системе са релативним приоритетом. Једнакост (5.4.2) се може написати као

$$D - \sum_{i=1}^{r} D_i \rho_i = \sum_{i=1}^{r} \frac{D_i}{E[X_i]} (\rho_i W_i^q).$$

Нека је $f_i = \dfrac{D_i}{E[X_i]}$ и $g_i = \rho_i W_i^q$ и потребно је минимизовати суму $\sum_{i=1}^{r} f_i g_i$. Из закона одржања (5.2.1) сума $\sum_{i=1}^{r} g_i$ је константна без обзира на дисциплину опслуживања. Ако променимо индексе и редослед функција f_i тако да је

$$f_1 \geqslant f_2 \geqslant \ldots \geqslant f_r, \tag{5.4.3}$$

најбољи начин да комбинујемо чланове g_i и f_i је да највећи f_1 спојимо са елементом g_i најмање масе и тако даље. Минимум функције $g_i = \rho_i W_i^q$ добија се смањивањем просечног времена чекања W_i^q пошто је ρ_i константно. Највиши приортет као што је подразумевано имају они клијенти са најмањим временом чекања. Дисциплина опслуживања којом се ово постиже је SPTF додељујући највиши приоритет првој класи. Просечни новчани губитак који настаје у систему због кашњења клијената минимизује се SPTF дисциплином поштујући распоред функција као у изразу (5.4.3).

5.5 Опслуживање са апсолутним приоритетом

Претпоставимо да $M|G|1$ систем са апсолутним приоритетом опслуживања има r различитих класа опслуживања. Када у систем дође клијент C_i приоритета i, ако је сервер слободан и у систему нема клијената вишег приоритета одмах почиње његово опслуживање. Уколико је на опслуживању клијент вишег приоритета, клијент C_i стаје у свој ред на основу приоритета и чека. Ако се тренутно опслужује клијент нижег приоритета $k, k > i$, његово опслуживање се прекида и започиње опслуживање придошлог клијента. Након завршеног опслуживања клијента C_i и свих клијената приоритета $k - i$ који су дошли у међувремену, наставља се опслуживање клијента приоритета k. Овде ћемо разматрати случај да се опслуживање настави тамо где је прекинуто а не да креће из почетка.

За разлику од система са релативним приоритетом овде постоји неколико различитих преосталих времена опслуживања, највише једно по класи приоритета. Због тога је компликовано израчунати перформансе система. На клијенте приоритета i не утичу клијенти нижег приоритета, практично су невидљиви и може се сматрати да је $\lambda_j = 0$ за $j > i$. Просечно време које клијент са приоритетом i проведе у систему једнако је

$$W_i = W_i^q + E[X_i] + E[I_i], \tag{5.5.1}$$

при чему је W_i^q просечно време чекања, X_i случајна величина која представља дужину опслуживања и I_i случајна величина која означава укупно време прекида током опслуживања. Укупна количина посла коју сервер треба да обави не зависи од типа приоритета опслуживања, иста је и код релативног и апсолутног приоритета. Сходно томе, просечно време које клијент проведе чекајући израчунава се на исти начин као код система са релативним приоритетом

$$W_i^q = \frac{\sum_{j=1}^{r} \lambda_j E[X_j^2]}{2\Big(1 - \sum_{j=1}^{i} \rho_j\Big)\Big(1 - \sum_{j=1}^{i-1} \rho_j\Big)}.$$

Време прекида током опслуживања састоји се из два дела: времена опслуживања клијената који су иницијално прекинули опслуживање посматраног клијента и од времена опслуживања клијената који су дошли током првобитног прекида,

$$I_i = x_i \sum_{j=1}^{i-1} \lambda_j X_j + I_i \sum_{j=1}^{i-1} \lambda_j X_j.$$

Следи да је

$$E[I_i] = \frac{E[X_i] \sum_{j=1}^{i-1} \rho_j}{1 - \sum_{j=1}^{i-1} \rho_j}.$$

Заменом вредности за W_i^q и $E[I_i]$ у једнакост (5.5.1) добија се просечно време које клијент приоритета i проведе у систему

$$W_i = \frac{\sum_{j=1}^{r} \lambda_j E[X_j^2]}{2\Big(1 - \sum_{j=1}^{i} \rho_j\Big)\Big(1 - \sum_{j=1}^{i-1} \rho_j\Big)} + \frac{E[X_i]}{1 - \sum_{j=1}^{i-1} \rho_j}.$$

Просечан број клијената i-тог приоритета у реду L_i^q и у читавом систему L_i је

$$L_i^q = \frac{\lambda_i \sum_{j=1}^{r} \lambda_j E[X_j^2]}{2\Big(1 - \sum_{j=1}^{i} \rho_j\Big)\Big(1 - \sum_{j=1}^{i-1} \rho_j\Big)}$$

$$L_i = \frac{\lambda_i \sum_{j=1}^{r} \lambda_j E[X_j^2]}{2\Big(1 - \sum_{j=1}^{i} \rho_j\Big)\Big(1 - \sum_{j=1}^{i-1} \rho_j\Big)} + \frac{\rho_i}{1 - \sum_{j=1}^{i-1} \rho_j}.$$

Поред наведених дисциплина приоритетног опслуживања постоје и многе друге где функција приоритета (функција на основу које се одређује приоритет клијента уколико није константан, тј ако се мења током времена) зависи од времена и дужине опслуживања и планираног распореда опслуживања. Детаљније о овоме може се наћи у књигама [2] и [5].

6 Закључак

Овај рад представља увод у теорију редова са Пуасоновим улазним потоком и приоритетним опслуживањем. Базиран је на већ установљеним и доказаним једнакостима које чине неопходну основу за даље познавање и разумевање ове области. Одређене су расподеле вероватноћа броја клијената, времена које клијент проведе у систему и у реду, периоди заузетости сервера и Pollaczek-Khinchin-ова формула. За системе са приоритетним опслуживањем приказани су циклуси задржавања, закони одржања, расподеле дужине чекања и перформансе система са релативним и апсолутним приоритетом.

Дубља анализа система са приоритетом и различитих функција приоритета може се наћи у књигама [2] и [5].

Теорија редова је област математике која се непрекидно развија и која има јако велику примену. Савремене информационе технологије не могу се замислити без употребе теорије мрежа у чијој је основи теорија редова. Та област је јако комплексна и због оптимизације система и смањивања трошкова. Примери редова јављају се и у болницама и хитној помоћи, где је пацијенту потребно пружити помоћ у зависности од степена повреде. Из свега наведеног настаје потреба за приоритетним опслуживањем, при чему се смањује време чекања клијената вишег приоритета на рачун повећања времена чекања клијената нижег приоритета. У системима са стационарним приоритетом тај однос се не може исконтролисати и зато систем може показати лоше перформансе. Контрола времена чекања може се постићи акумулираним приоритетом, односно приоритетом који није константан већ се мења током времена. Тиме се могу анализирати расподеле чекања и моменти тих расподела. Анализа модела са акумулираним приоритетом представља поље интересовања математичара последњих година, и неки од новијих радова на ту тему су [7], [8] и [9].

Литература

[1] Leonard Kleinrock, *Queueing Systems, Volume 1: Theory*, Wiley & Sons, Inc., New York, 1975.

[2] Leonard Kleinrock, *Queueing Systems, Volume 1: Computer Applications*, Wiley & Sons, Inc., New York, 1977.

[3] A. K. Erlang, *The Theory of Probabilities and Telephone Conversations*, Nyt Tidsskrift for Matematik B, vol 20, 1909.

[4] D. G. Kendall, *Stochastic Processes Occurring in the Theory of Queues and their Analysis by the Method of the Imbedded Markov Chain*, The Annals of Mathematical Statistics, 1953.

[5] R. W. Conway, W. L. Maxwell, L. W. Miller, *Theory of Scheduling*, Courier Corporation, 2003.

[6] D. Gross, J. F. Shortle, J. M. Thompson, C. M. Harris, *Fundamentals of Queueing Theory*, Wiley & Sons, Inc., 2008.

[7] D. A. Stanford, P. Taylor, I. Ziedins, *Waiting time distributions in the accumulating pririty*, Springer, 2013.

[8] V. A. Fajardo, *A Generalization of $M|G|1$ Priority Models via Accumulating Priority*, Waterloo, 2015.

[9] V. A. Fajardo, S. Drekic, *Waiting time distributions in the preemptive accumulating priority queue*, Methodology in Computing and Applied Probability, 2015.