Poglavje 2

Teorija strežbe

Področje teorije strežbe (angl. *queueing theory*[1]) se je začelo razvijati v začetku 20. stoletja zaradi potreb obvladovanja načrtovanja zmogljivosti infrastrukture ob vpeljavi klasične telefonije, kasneje pa tudi zaradi potreb hitrega razvoja računalniških sistemov in omrežij. Pionirsko delo na področju formalizacije strežbe zahtev v klasični telefoniji je opravil danski inženir Agner K. Erlang (1878–1929), njegovo delo pa je z natančnejšo formalizacijo teorije strežbe za področje računalniških sistemov in omrežij nadaljeval ameriški raziskovalec Leonard Kleinrock (1934-). Izčrpna formalizacija teorije strežbe in raziskave na tem področju so predstavljene v Kleinrockovem delu [3], povzetek področja teorije strežbe pa v delu [1], po katerem je deloma povzeto tudi pričujoče poglavje.

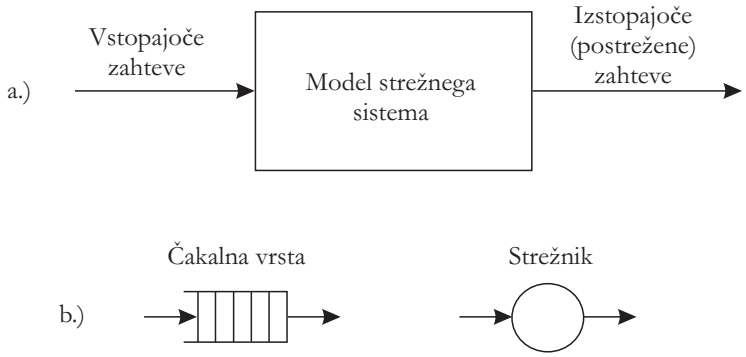
Temeljni pojmi področja teorije strežbe so *zahteve*, *čakalne vrste* in *strežniki*, katerih pomene v domeni računalniških omrežij smo spoznali že v prejšnjem poglavju. Teorija strežbe nam bo v kontekstu pričujočega dela v pomoč predvsem pri izvajanju *analize zmogljivosti delovanja* računalniških omrežij.

2.1 Model strežnega sistema

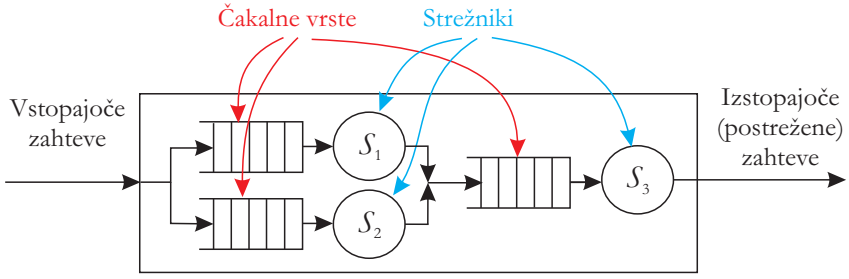
Osnovni objekt opazovanja v pričujočem poglavju je *model strežnega sistema*. Zahteve vanj na vhodni strani vstopajo, na izhodni strani pa iz njega postrežene (obdelane, servisirane itd.) izstopajo. V domeni računalniških in komunikacijskih sistemov vstopajoče zahteve imenujemo tudi za *vhodno breme sistema* [4]. Notranjost modela strežnega sistema je sestavljena iz čakalnih vrst in strežnikov, med njimi pa vodijo povezave, po katerih potujejo zahteve. Na gornjem delu slike 2.1 je predstavljena osnovna shema modela strežnega sistema, na spodnjem delu iste slike sta predstavljena grafična primitiva za ponazarjanje obeh tipov gradnikov, na sliki 2.2 pa je predstavljen primer modela strežnega sistema

[1] Ustreznejši neposredni slovenski prevod pojma „*queueing theory*“ bi bil sicer „teorija čakalnih vrst“, a se je avtor pričujočega dela odločil za uporabo pomensko ustreznejšega termina *teorije strežbe*. Slovenski računalniški slovar DIS (http://dis-slovarcek.ijs.si/) na tem mestu svetuje uporabo termina *teorija množične strežbe*.

s tremi strežniki in tremi čakalnimi vrstami.



Slika 2.1: Osnovna shema modela strežnega sistema (a) z grafičnima primitivoma za označevanje čakalnih vrst in strežnikov (b).



Slika 2.2: Grafična ponazoritev modela strežnega sistema s tremi strežniki (S_1,S_2,S_3) in tremi čakalnimi vrstami.

Vsak model strežnega sistema praviloma vsebuje eno ali več *strežnih poti*, ki vodijo od vstopne do izstopne točke modela strežnega sistema. Posamezno strežno pot sestavlja zaporedje parov, pri čemer posamezni par sestavljata čakalna vrsta in strežnik. Povedano drugače vsaka strežna pot vsebuje zaporedje strežnikov (storitev strežb), pred vsakim strežnikom pa je nameščena čakalna vrsta. Model strežnega sistema s slike 2.2 tako vsebuje dve strežni poti. Prva pot vodi zahteve skozi strežnika S_1 in S_3, druga pot pa skozi strežnika S_2 in S_3. V primeru, da bo zahteva ubrala prvo pot, bo postrežena v strežnikih S_1 in S_3, v primeru da bo zahteva ubrala drugo pot, pa bo postrežena v strežnikih S_2 in S_3.

Za strežbo zahteve na izbrani strežni poti v modelu strežnega sistema veljajo naslednje zakonitosti:

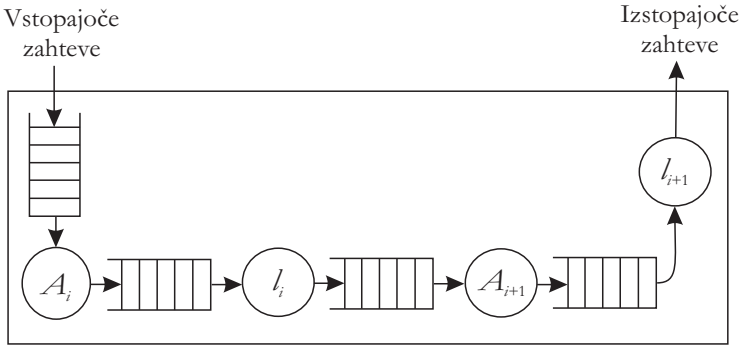
- z izbiro strežne poti bo ob normalnem delovanju strežnega sistema zahteva

po nekem določenem času postrežena na vseh strežnikih, ki se nahajajo na izbrani strežni poti in bo nato zapustila model strežnega sistema;

- v posameznem strežniku se lahko istočasno obdeluje (streže) največ ena zahteva; v primeru, da v času strežbe neke zahteve pride na vhod strežnika novoprispela zahteva, se slednja uvrsti v čakalno vrsto pred strežnikom; tam se bo zadrževala po principu neke vnaprej določene *strežne discipline* vse do tedaj, dokler ne bo strežnik prost, opazovana zahteva pa izbrana za vstop v strežbo; najpogosteje uporabljana strežna disciplina je disciplina FIFO (angl. *first in, first out*);
- strežba posamezne zahteve v strežniku ima venomer neko časovno trajanje, ki je v splošnem odvisno tako od hitrosti delovanja strežnika, kot tudi od velikosti in vrste zahteve;

Zmogljivost posameznega strežnika običajno podajamo s parametrom *intenzivnosti strežbe* (v domeni računalništva tudi s parametrom intenzivnosti procesiranja), s katero se obdelujejo vstopajoče zahteve. Merska enota intenzivnosti strežbe se podaja v številu obdelanih zahtev na časovno enoto, matematično pa se bomo na intenzivnost strežbe i-tega strežnika sklicevali s spremenljivko μ_i $(i = 1, ..., n)$, pri čemer je n število strežnikov v opazovanem modelu strežnega sistema.

Na tem mestu lahko s predstavljeno *grafično strežno notacijo* natančneje ponazorimo primer modela vezave omrežnih naprav in fizičnih prenosnih medijev s slike 1.2. Grafično je model prikazan na sliki 2.3, z razliko od prvotne slike pa v modelu strežnega sistema tudi prenosni mediji postanejo strežniki.



Slika 2.3: Grafična ponazoritev modela strežnega sistema s slike 1.2.

V kontekstu realnih računalniških omrežij in sistemov zahteve v modelih igrajo vloge paketov v računalniških omrežjih, vloge zahtev po procesiranju podanih s strani aplikacij proti procesorju v opazovanem računalniškem sistemu itd. Tipični primeri strežnikov v realnih računalniških sistemih so procesorji, registri, vodila, komunikacijski kanali, usmerjevalniki v računalniških omrežjih itd.

2.2 Čakalne vrste v modelih strežnih sistemov

Osnovne tri značilnosti čakalne vrste v modelu strežnega sistema so njena *dolžina* (število čakalnih mest v vrsti), njena politika *zavržbe zahtev* (paketov) ter njena *strežna disciplina* (metodologija jemanja zahtev iz čakalne vrste).

S teoretičnega vidika so dolžine čakalnih vrst lahko *končne* (omejene na m čakalnih mest, kar pomeni, da je lahko v čakalni vrsti največ m zahtev) ali pa *neskončne* (neomejene). Pojem neskončne čakalne vrste nam pomaga pri poenostavitvah matematičnih izračunov, moramo pa se zavedati, da v praksi neskončne čakalne vrste ne moremo realizirati.

Končna čakalna vrsta se lahko s časom zapolni (v njej se nahaja m čakajočih zahtev), kar nas ob nadaljnjem vstopanju zahtev v čakalno vrsto pripelje do *izgubljanja zahtev*. Pri tem se lahko izgubljajo novoprispele zahteve (se ne uvrščajo v že zasedeno čakalno vrsto, temveč zavržejo) ali pa zahteve, ki so bile predhodno že uvrščene v čakalno vrsto (predhodno uvrščene zahteve v vrsti se zavržejo in nadomestijo z novoprispelimi zahtevami). Opisani strategiji predstavljata dve osnovni politiki zavržbe zahtev (angl. *drop policy*) in s tem posledično izgubljanja zahtev (paketov). Na področju računalniških omrežij se večinoma srečujemo s prvo politiko, ko se v primeru polne čakalne vrste zavržejo novoprispeli paketi (angl. *tail drop*).

Strežna disciplina (angl. *queueing discipline*) ali vrstni red jemanja zahtev iz vrste določa, katera od zahtev v čakalni vrsti bo izbrana za strežbo, ko se bo strežnik, pred katerim se nahaja čakalna vrsta, sprostil. Najbolj razširjene osnovne strežne discipline so sledeče [1]:

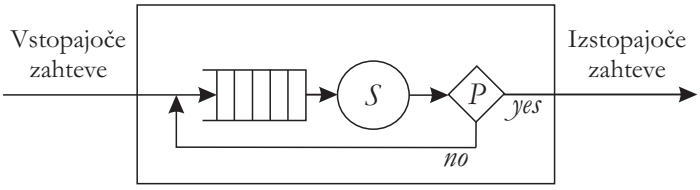
- *FIFO disciplina* (angl. *first in, first out*): vrstni red jemanja zahtev iz vrste je enak vrstnemu redu prihajanja zahtev v vrsto; omenjeni način strežne discipline je najpogostejši in ga na področju modeliranja računalniških sistemov in omrežij jemljemo kot privzet način discipline strežbe;

- *LIFO disciplina* (angl. *last in, first out*): vrstni red jemanja zahtev iz vrste je obraten vrstnemu redu prihajanja zahtev v vrsto; tipičen primer tovrstne uporabe v računalništvu najdemo pri implementaciji sklada (angl. *stack*);

- *prioritetna disciplina* (angl. *priority queueing*): vrstni red jemanja zahtev iz vrste je pogojen z njihovimi prioritetami; iz vrste se venomer jemlje zahtevo z največjo prioriteto; tipičen primer tovrstne uporabe v računalniških komunikacijah je v različnih prioritetah paketov, ki jih generirajo visokonivojski protokoli; v ta namen imajo nekatere strežniki v omrežnih napravah več FIFO čakalnih vrst, pri čemer se paketi uvrščajo v vrste glede na svojo prioriteto; strežnik jemlje pakete iz čakalne vrste z najvišjo prioriteto, šele pri praznosti omenjene vrste pa začne prevzemati pakete iz vrst, ki imajo nižje prioritete; na omenjeni način dosežemo prioritetizacijo in zvečanje hitrosti prometa določene vrste paketov, po drugi plati pa se upočasni promet paketov z nižjimi prioritetami; upočasnitev in pospešitev

sta odvisni od razmerja med prioritetami večje populacije prispelih paketov; tako lahko govorimo o dveh skupinah paketov („*low-priority packets*“, „*high-priority packets*“), kar pa je v nasprotju z iniciativo o *nevtralnosti interneta*; prioritetna disciplina prinaša tudi dilemo, ali se strežba zahteve z manjšo prioriteto ob prihodu zahteve z večjo prioriteto prekine in se vzame v strežbo novoprispelo zahtevo, ali pa se strežbo zahteve z manjšo prioriteto dokonča do konca, šele potem pa v strežbo vstopi zahteva z večjo prioriteto; v prvem primeru govorimo o *strežbi s prekinitvami* (angl. *preemtive queueing*), v drugem primeru pa o *strežbi brez prekinitev* (angl. *non-preemtive queueing*);

- *naključna disciplina* (angl. *service in random order*): vrstni red jemanja zahtev iz vrste je naključen;
- *SJF disciplina* (angl. *shortest job first*): iz vrste se venomer vzame zahtevo, ki bo imela najkrajši čas obdelave;
- *LJF disciplina* (angl. *largest job first*): iz vrste se venomer vzame zahtevo, ki bo imela najdaljši čas obdelave;
- *dodeljevanje časovnih rezin strežbe* (angl. *time sharing, processor sharing*): v tem primeru se razpoložljiva vnaprej določena enotska rezina strežnega časa strežnika venomer dodeli prvočakajoči zahtevi; ni nujno, da bo ta zahteva po dodelitvi časovne rezine in izvedbi strežbe dokončno postrežena; če ne bo, se vrne na konec čakalne vrste pred strežnikom in čaka po FIFO principu na dodelitev nove časovne rezine, če pa je s tem strežba zahteve dokončana, zahteva zapusti strežnik in odpotuje naprej po svoji strežni poti; na sliki 2.4 je predstavljen model strežnega sistema z dodeljevanjem časovnih rezin strežbe; ključna gradnika modela sta vejanje glede na izpolnjenost pogoja P in povratna povezava, ki vodi nedokončno postreženo zahtevo nazaj v čakalno vrsto; pomen pogoja P bi bil npr. „*ali je zahteva postrežena v celoti*“;



Slika 2.4: Grafična ponazoritev modela strežnega sistema z dodeljevanjem časovnih rezin strežbe.

Čakalne vrste v računalniških sistemih in omrežjih realiziramo s pomnilniki ali tako imenovanimi *vmesniki* (angl. *buffer*). Z vidika realizacije pristajamo tako zgolj na vrste s končno dolžino. Še enkrat poudarimo, da je privzeta strežna

disciplina v čakalnih vrstah tipa FIFO, privzeta politika zavržbe paketov pa strategija „*tail drop*“. Pri obeh privzetih značilnostih govorimo o čakalni vrsti „*FIFO with tail drop*“.

Na področju računalniških omrežij najdemo mnogo izvedenk predhodno navedenih osnovnih strežnih disciplin. Ena od novejših strategij pri usmerjevalnikih je npr. dinamično[2] formiranje vrst FIFO pred izvajanjem usmerjanja (strežbo), pri čemer se vsaka od vrst formira za posamezen „požrešen izvor prometa“, v nadaljevanju pa se paketi iz teh vrst v strežbo (usmerjanje) jemljejo po principu „poštenosti“ oziroma krožnem (angl. *round robin*) principu. Slednje pomeni, da v usmerjanje vzamemo en paket iz 1. vrste, nato en paket iz 2. vrste, ..., nato en paket iz zadnje n-te vrste, nato zopet paket iz 1. vrste itd. Omenjeni koncept v angleškem jeziku imenujemo s terminom „*Fair queueing*“, njegova prednost pa je v njegovi zmožnosti razlikovanja med različnimi izvori prometa.

2.3 Strežna enota in strežna mreža

Doslej smo v pričujočem poglavju govorili o modelih strežnih sistemov. Slednje delimo na dve skupini in sicer na *strežne enote* (angl. *single queueing node, single queueing system*) in na *strežne mreže* (angl. *multi queueing node, multi queueing system*). V obeh primerih gre za modela (model strežne enote in model strežne mreže), a bomo izraz „model“ zaradi preglednosti v nadaljevanju opuščali. V nadaljevanju definiramo pojem strežne enote.

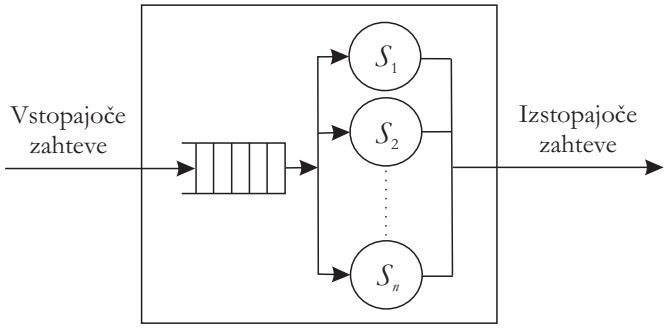
Definicija 1 *Strežna enota je sestavljena iz n ($n \geq 1$) vzporedno vezanih strežnikov S_i ($i = 1, ..., n$) s skupno čakalno vrsto. Od strežnikov zahtevamo, da so funkcionalno in zmogljivostno ekvivalentni.*

Definicija 2 *Funkcionalna ekvivalentnost paralelno vezanih strežnikov pomeni, da vsi strežniki ponujajo enak tip strežbe, njihova zmogljivostna ekvivalentnost pa pomeni, da imajo vsi paralelno vezani strežniki enako intenzivnost strežbe ($i = 1, ..., n : \mu_i = \mu$).*

Vstopajoča zahteva je v strežni enoti deležna natanko ene strežbe v natanko enem od n strežnikov. V vsakem strežniku se lahko istočasno obdeluje ali streže le ena zahteva. Privzeta strežna disciplina čakalne vrste pred paralelno vezavo

[2] Pod pojmom dinamičnosti imamo v mislih sprotno kreacijo novih čakalnih vrst glede na porajanje novih *požrešnih izvorov prometa*. Ob usahnitvi posameznega požrešnega izvora prometa se njegova čakalna vrsta nemudoma izbriše.

strežnikov je tipa FIFO. Slednje pomeni, da se ob prostosti poljubnega od n strežnikov vzame zahtevo iz čakalne vrste tipa FIFO in dostavi v strežbo v prost strežnik. Posamezna zahteva nima možnosti izbire strežnika, v katerem bo postrežena. Slednje tudi nebi imelo pomena, ker so strežniki funkcionalno in zmogljivostno ekvivalentni - vsi ponujajo enak tip in hitrost strežbe. Grafična predstavitev opisane strežne enote se nahaja na sliki 2.5.



Slika 2.5: Grafična ponazoritev modela strežne enote.

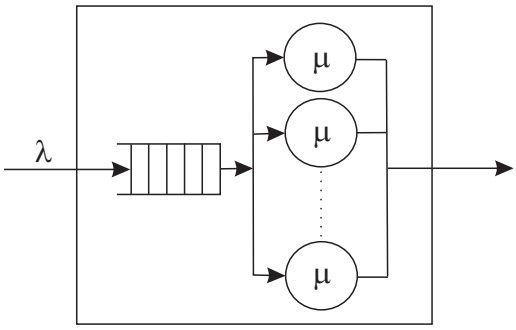
Paralelna vezava strežnikov nam omogoča *razširljivost* ali *skalabilnost* strežne enote, saj lahko z zvečevanjem števila paralelno vezanih strežnikov dosežemo v realnih strežnih sistemih[3] dovolj hitro obdelavo poljubne količine vstopajočih zahtev (vhodnega bremena).

Predhodno smo že omenili, da je za vsak strežnik v strežni enoti karakteristična *intenzivnost strežbe* μ. Poleg te kvantitativne spremenljivke je za dovolj veliko prepustnost strežne enote pomembna tudi druga kvantitativna spremenljivka λ, ki jo imenujemo za *intenzivnost porajanja* ali *vstopanja* zahtev v strežno enoto. Omenjena mera je karakteristična za vstopno točko strežne enote. Izražamo jo z mersko enoto števila zahtev, ki vstopijo v strežno enoto v nekem časovnem intervalu. Na sliki 2.6 je predstavljena grafična ponazoritev modela strežne enote z njenimi osnovnimi kvantitativnimi spremenljivkami μ in λ.

Definicija 3 *Strežna mreža je skupek zaporedno, vzporedno, ali mešano vezanih m ($m > 1$) strežnih enot.*

V strežni mreži se tajo pojavita vsaj dve strežni enoti, kar pomeni, da imamo tudi najmanj dva strežnika in najmanj dve čakalni vrsti. Zahteve so lahko v strežni mreži postrežene več kot enkrat (gredo skozi več strežnikov). Slika 2.2

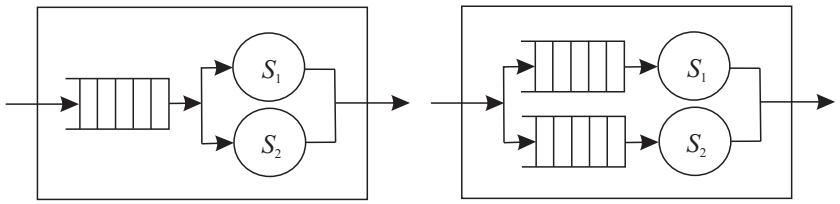
[3] Pod pojmom *realnih strežnih sistemov* imamo v mislih implementacije računalniških omrežij in sistemov.



Slika 2.6: Grafična ponazoritev modela strežne enote z njenimi osnovnimi kvantitativnimi spremenljivkami μ in λ.

predstavlja primer takšne strežne mreže s tremi strežnimi enotami, slika 2.3 pa primer strežne mreže s štirimi strežnimi enotami.

Na sliki 2.7 sta prikazana na prvi pogled podobna strežna sistema. Levi predstavlja strežno enoto z dvema strežnikoma, desni pa strežno mrežo z dvema strežnima enotama. V prvem primeru zahteva strežne poti ne more izbrati (z uvrstitvijo v privzeto FIFO čakalno vrsto bo dodeljena v strežbo prvemu prostemu strežniku), v drugem primeru pa zahteva po nekem kriteriju (npr. verjetnostnem) izbere eno od dveh vzporednih strežnih poti (eno od dveh vzporedno vezanih strežnih enot), v kateri bo postrežena.



Slika 2.7: Grafični ponazoritvi modela strežne enote (levo) in modela strežne mreže (desno).

V nadaljevanju poglavja o teoriji strežbe bomo večino časa namenili obravnavi strežnih enot, na samem koncu pa se bomo osredotočili tudi na strežne mreže. Pomen strežnih enot je z vidika modeliranja računalniških omrežij pomemben, saj v modelih običajno posamezne fizične prenosne medije in posamezne omrežne naprave intepretiramo s posameznimi strežnimi enotami, z njihovim povezovanjem v celovit model računalniškega omrežja pa pridemo do končnih modelov – strežnih mrež.

2.4 Kendallova notacija strežnih enot

Britanski statistik David George Kendall (1918–2007) je leta 1953 vzpostavil klasifikacijo strežnih enot, ki je na področju teorije strežbe v uporabi še danes [1]. Kendallova notacija temelji na šestorčku

$$A/B/m/k/P/Q, \tag{2.1}$$

s katerim enolično opišemo poljubno strežno enoto (angl. *queueing node*). Pomeni posameznih parametrov šestorčka so sledeči:

- A - vrsta porazdelitve *medprihodnih časov* vstopajočih zahtev v strežno enoto;
- B - vrsta porazdelitve *strežnih časov* zahtev;
- m - število paralelno vezanih strežnikov v strežni enoti;
- k - kapaciteta strežne enote; predstavlja maksimalno število zahtev, ki se lahko nahajajo v strežni enoti istočasno; kapaciteto strežne enote zapišemo z izrazom
$$k = m + len(queue), \tag{2.2}$$
kjer $len(queue)$ predstavlja dolžino čakalne vrste, m pa število strežnikov v strežni enoti; privzeta vrednost kapacitete je neskončna ($k = \infty$), kar ob končnem številu strežnikov neposredno implicira hipotetično neskončno dolžino čakalne vrste ($len(queue) = \infty$);
- P - velikost populacije zahtev, ki vstopajo iz zunanjega okolja v strežno enoto; privzeta vrednost velikosti populacije je neskončna ($P = \infty$);
- Q - tip strežne discipline v čakalni vrsti strežne enote; privzeta vrednost je strežna disciplina FIFO;

Za *medprihodne čase* (angl. *inter arrival times*) štejemo čase med prihodi sosednjih zahtev v strežno enoto, za *strežne čase* (angl. *service times*) pa čase, ki se porabijo za strežbo posameznih zahtev v strežniku strežne enote. Poudarimo, da imamo pri zapisu v Kendallovi notaciji, v katerem izpustimo zadnje tri parametre, v mislih privzete vrednosti .../∞/∞/FIFO ($k = \infty, P = \infty, Q =$FIFO). Parametre m, k in P v primeru njihove končnosti zapisujemo s pozitivnimi celoštevilčnimi vrednostmi, parametra k in P pa opcijsko tudi kot neskončni (privzeti) vrednosti. Zadnje tri parametre k, P in Q navajamo le, če so različni od privzetih vrednosti.

Za parametra A in B iz Kendallove notacije uporabljamo oznake z naslednjimi pomeni:

- D: deterministična, degenerirana ali izrojena porazdelitev;
- M: eksponentna porazdelitev;
- E_k: Erlangova porazdelitev k-tega reda;

- G: splošna verjetnostna porazdelitev;

Proti koncu obsežnega poglavja o teoriji strežbe si bomo natančneje ogledali primere strežnih enot $M/M/1$, $M/M/m$, $M/M/1/s$, $M/M/m/m$ in $M/M/m/m/c$, pred tem pa si bomo ogledali nekaj značilnosti strežnih sistemov.

2.4.1 Vrste porazdelitev

V pričujočem razdelku si bomo ogledali *deterministično* (D) in *eksponentno porazdelitev* (M). Prva je dokaj idealizirana, a vseeno uporabna za modeliranje deterministčnih realnih strežnih sistemov, druga pa najpogosteje uporabljana na področju modeliranja strežbe zahtev v računalniških sistemih in omrežjih. Obe temeljita na pojmu *slučajne spremenljivke*[4] in njene *verjetnostne porazdelitve*[5]. Uporaba Erlangove porazdelitve in splošne verjetnostne porazdelitve je na področju modeliranja računalniških omrežij redka, zato jih bomo pri opisih v pričujočem razdelku izpustili. Za natančnejše razlage različnih vrst verjetnostnih porazdelitev bralcu svetujemo vpogled v osnovno literaturo s področja verjetnosti.

Deterministična porazdelitev

Deterministična porazdelitev je primer izrojene verjetnostne porazdelitve (angl. *degenerate distribution*) slučajne spremenljivke, kjer ima slednja vedno enako vrednost ali enak izid. Primer sistema, kjer vlada tovrstna deterministična porazdelitev, bi bilo štetje izidov metanja kovanca, pri čemer bi imel kovanec na obeh straneh vgraviran grb. Vsak dogodek tako rezultira v enak izid - met grba. Za medprihodne in strežne čase z deterministično porazdelitvijo tako velja sledeče:

- časi med prihodi zahtev v strežno enoto so skozi čas enaki oziroma konstantni (angl. *fixed inter arrival times*);

- časi zadrževanja posamezne zahteve v strežniku ali časi njene strežbe so za vse vstopajoče zahteve enaki oziroma konstantni (angl. *fixed service times*);

Predhodno smo že omenili intenzivnost prihajanja zahtev λ in intenzivnost strežbe zahtev μ. V primeru determinističnega porajanja zahtev velja izraz

$$t_{interarrival} = \frac{1}{\lambda}, \tag{2.3}$$

[4]Slučajna spremenljivka je količina, katere vrednost je odvisna od slučaja [5] in nastopi kot rezultat poskusa (dogodka), kjer je možnih več izidov. Pri tem pojavitev katerekoli vrednosti iz danega območja predstavlja slučajno vrednost (vir: Wikipedia).

[5]Verjetnostna porazdelitev določa verjetnost, da slučajna spremenljivka zavzame neko vrednost. Porazdelitev verjetnosti opisuje območje, ki ga slučajna spremenljivka lahko zavzame, in verjetnost, da je vrednost slučajne spremeljivke v tem območju (vir: Wikipedia).

pri čemer $t_{interarrival}$ predstavlja časovno nespremenljiv ali *konstanten medprihodni čas*, v primeru deterministične intenzivnosti strežbe pa velja izraz

$$t_{service} = \frac{1}{\mu}, \tag{2.4}$$

kjer $t_{service}$ predstavlja časovno nespremenljiv ali *konstanten strežni čas*.

Sistem, kjer sta tako prihajanje kot tudi strežba deterministično opredeljena, v splošnem zapišemo kot $D/D/1$ sistem. Deterministični porazdelitvi medprihodnih in servisnih časov sta v realnih sistemih izredno redki. Z vidika računalniških omrežij paketi v strežne enote (omrežne naprave in na fizične prenosne medije) ne vstopajo s konstantnimi medprihodnimi časi, niti niso konstantni časi servisiranja ali strežbe paketov. S tega vidika je uporaba deterministične porazdelitve za parametra A in B pri modeliranju računalniških omrežij redko uporabna.

Eksponentna porazdelitev

Za opisovanje medprihodnih in strežnih časov v modelih strežnih enot najpogosteje uporabljamo *eksponentno porazdelitev*. Slednja se ne uporablja zgolj za modeliranje računalniških omrežij in sistemov, temveč tudi za ostale dinamične sisteme iz realnega okolja, ki niso deterministični in so do neke mere "*naključni*"ali "*slučajni*".

Eksponentna porazdelitev je v tesni povezavi s *Poissonovo porazdelitvijo*. Prva nam daje odgovor na vprašanje "*kolikšen je čas med porajanjema dveh sosednih zahtev*", druga pa odgovor na vprašanje "*koliko zahtev se bo porodilo v opazovanem časovnem intervalu*". Več bomo o Poissonovi verjetnostni porazdelitvi povedali v enem od sledečih razdelkov. Eksponentna porazdelitev torej opisuje časovne intervale med posameznimi dogodki v Poissonovi porazdelitvi. Ti dogodki so med seboj praviloma neodvisni in naključni.

Funkcija gostote verjetnosti (angl. *probability density function* - PDF) eksponentne porazdelitve za slučajno spremenljivko X se izračuna po izrazu

$$f(x, \lambda) = \begin{cases} \lambda e^{-\lambda x}, & x \geq 0, \\ 0, & x < 0, \end{cases} \tag{2.5}$$

zbirna funkcija verjetnosti (angl. *cumulative distribution function* - CDF) pa po izrazu

$$F(x, \lambda) = \begin{cases} 1 - e^{-\lambda x}, & x \geq 0, \\ 0, & x < 0. \end{cases} \tag{2.6}$$

V obeh primerih λ označuje *parameter stopnje funkcije* (angl. *rate parameter*). V njegovi vlogi lahko nastopata tako nam že znani λ (intenzivnost porajanja zahtev) in μ (intenzivnost servisiranja zahtev). Zbirno funkcijo verjetnosti (CDF) lahko zapišemo tudi z izrazoma

$$P(X > x) = e^{-\lambda x}, \tag{2.7}$$

ki ponazarja verjetnost, da je vrednost slučajne spremenljivke X (medprihodni ali strežni čas) večja od neke podane vrednosti x in

$$P(X \leq x) = 1 - e^{-\lambda x}, \tag{2.8}$$

ki ponazarja verjetnost, da je vrednost slučajne spremenljivke X (medprihodni ali strežni čas) manjša ali enaka neki podani vrednosti x.

V nadaljevanju prikažemo zgled uporabe eksponentne porazdelitve na primeru strežne enote.

Zgled 1 *Opravka imamo s strežno enoto z enim strežnikom, v katerem je čas strežbe porazdeljen eksponentno z intenzivnostjo strežbe $\mu = \frac{1}{10}$ zahteve na časovno enoto. Zahteva vstopi v prazno strežno enoto in zanimajo nas sledeče verjetnosti:*

(i) kolikšna je verjetnost, da bo strežni čas zahteve manjši ali enak 5 časovnim enotam?

(ii) kolikšna je verjetnost, da bo strežni čas zahteve daljši od 10 časovnih enot?

(iii) kolikšna je verjetnost, da bo strežni čas zahteve večji od 5 časovnih enot in istočasno manjši ali enak 10 časovnim enotam?

Rešitev: *Intenzivnost strežbe μ nam služi kot parameter za evaluacijo zbirne funkcije verjetnosti iz izraza (2.8). Odgovore na vprašanja dobimo po sledečih izrazih:*

(i) uporabimo izraz

$$P(X \leq x) = 1 - e^{-\mu x}, \tag{2.9}$$

pri čemer predstavlja X vrednost slučajne spremenljivke časovnega trajanja, x pa neko konkretno vrednost, ki je v našem primeru 5 časovnih enot; numerični odgovor dobimo po izrazu

$$P(X \leq 5) = 1 - e^{-\frac{1}{10} * 5} = 1 - e^{-\frac{1}{2}} = 1 - 0,6066 = 0,3934. \tag{2.10}$$

(ii) uporabimo izraz

$$P(X > x) = 1 - P(X \leq x) = 1 - (1 - e^{-\mu x}) = e^{-\mu x}, \tag{2.11}$$

numerični odgovor pa dobimo po izrazu

$$P(X > 10) = e^{-\frac{1}{10} * 10} = e^{-1} = 0,3678. \tag{2.12}$$

(iii) uporabimo izraz

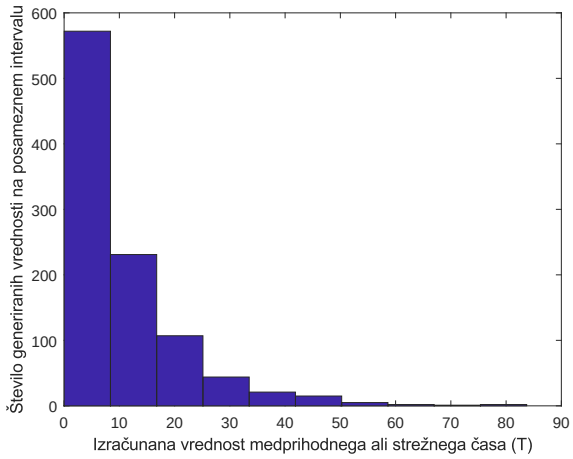
$$P(5 < X \leq 10) = 1 - (P(X \leq 5) + P(X > 10)) = 0.2388. \tag{2.13}$$

V domeni eksponentne porazdelitve velja, da tako medprihodni, kot tudi strežni časi niso več konstantni, temveč *med seboj neodvisni*, *naključni* ter *po-*

razdeljeni skladno z omenjeno porazdelitvijo. Do njihove vsakokratne vrednosti pridemo z izračunom inverza eksponentne zbirne funkcije verjetnosti (CDF). V avtomatiziranih simulacijskih okoljih so nam za izračune medprihodnih in strežnih časov običajno na voljo posebne funkcije, ki temeljijo na izračunu inverza CDF funkcije. V matematičnem orodju MATLAB je tako v ta namen na razpolago funkcija `exprnd(T)` (njej sorodna je funkcija `expinv(T,p)`), ki ji kot argument podamo inverzno vrednost λ ali μ (pričakovani medprihodni ali strežni čas t), kot rezultat pa dobimo izračunani čas r (medprihodni čas $t_{interarrival}$ ali strežni čas $t_{service}$) porazdeljen po eksponentni porazdelitvi. Z izrazom

$$r = -t * ln(1 - p) \tag{2.14}$$

ponazorimo način izračuna spremenljivke r (simulirani medprihodni ali strežni čas), kot ga izvede funkcija `exprnd(T)`, pri čemer p predstavlja naključno izbrano vrednost z intervala $[0, 1]$, t pa pričakovani čas ali inverzno vrednost ene od intenzivnosti (λ ali μ). Na sliki 2.8 je predstavljen histogram[6] tisočih ponovitev generiranja vrednosti r pri povprečni vrednosti t=10 časovnih enot. Iz histograma je lepo razvidna eksponentna porazdelitev generiranih vrednosti r.



Slika 2.8: Histogram tisočih ponovitev generiranja naključno pogojene eksponentne vrednosti pri vhodnem podatku t=10 časovnih enot.

Z opisoma deterministične in eksponentne verjetnostne porazdelitve smo povedali nekaj osnov o *vhodnem procesu* pred strežno enoto in o *strežnem procesu* v sami strežni enoti.

Brez porajanja zahtev, ki predstavljajo breme modeliranega strežnega sistema, bi bil slednji ves čas nezaseden (nezaposlen) in kot tak nezanimiv za opazovanje. Osnovni značilnosti *vhodnega procesa* sta intenzivnost porajanja zahtev λ

[6] Vrsta grafikona, ki ga v statistiki uporabljamo za prikaz porazdelitve opazovane slučajne spremenljivke.

in njena porazdelitev, na osnovi obeh pa znamo izračunavati medprihodne čase. Do sedaj smo govorili v okviru deterministič nih in eksponentnih porazdelitev o časovno nespremenljivi intenzivnosti porajanja λ. V nekaterih sistemih se lahko ta spreminja (tako dobimo časovno odvisno $\lambda(t)$), kar je mnogokrat pogojeno z velikostjo populacije. Manjša je populacija P po Kendallovi notaciji, večje nihanje $\lambda(t)$ skozi čas lahko pričakujemo. Inverzna vrednost intenzivnosti porajanja zahtev λ nam v primeru deterministič nega porajanja zahtev poda enolični medprihodni čas med sosednjimi zahtevami, v primeru eksponentnega porajanja zahtev pa zgolj izhodišče za vsakokratni izračun novega medprihodnega časa.

Strežni proces strežne enote temelji na m strežnikih, pri čemer je vsakemu od njih lastna intenzivnost strežbe μ. Posamezen strežnik lahko obdeluje istočasno le eno zahtevo. Inverzna vrednost intenzivnosti strežbe zahtev μ nam v primeru deterministič nega časa strežbe zahteve poda strežni čas strežbe posamezne zahteve enolično, v primeru eksponentne porazdelitve časa strežbe zahtev pa zgolj izhodišče za vsakokratni izračun novega strežnega časa. V splošnem bomo inverzno vrednost intenzivnosti strežbe μ zapisovali z oznako x. V primeru deterministič ne strežbe predstavlja x konstantni strežni čas, v primeru eksponente pa njegovo pričakovano vrednost, ki v konkretnih primerih odstopa od te vrednosti po eksponentni porazdelitvi.

Do sedaj smo spoznali osnovne kvantitativne metrike[7] za oceno strežne enote in sicer λ, μ in x. V naslednjem razdelku bomo naredili obširnejši pregled kvantitativnih metrik za ocenjevanje strežnih enot.

2.4.2 Kvantitativne metrike za ocenjevanje zmogljivosti strežnih enot

Osnovne kvantitativne metrike za ocenjevanje zmogljivosti opazovane strežne enote, ki so z vidika modeliranja zanimive za analitika, so po [1] sledeče:

- družina metrik *števila zahtev* N:
  - *število zahtev v strežni enoti*:
    - $N(t)$: število zahtev v strežni enoti v opazovanem času t;
    - N: povprečno število zahtev v strežni enoti v daljšem časovnem intervalu;
  - *število zahtev v čakalni vrsti*:
    - $N_q(t)$: število zahtev v čakalni vrsti v opazovanem času t;
    - N_q: povprečno število zahtev v čakalni vrsti v daljšem časovnem intervalu;
  - *število zahtev v strežbi*:
    - $N_s(t)$: število zahtev v strežnikih v opazovanem času t;

[7] Kvantitativna metrika ali mera je običajno merljiva v obliki številčno ali numerično izraženih podatkov, kvalitativna metrika pa v obliki opisnih ali lingvistično izraženih podatkov.

    - N_s: povprečno število zahtev v strežnikih v daljšem časovnem intervalu;

- družina metrik *časov prebivanja zahtev*:
  - *čas prebivanja zahteve v strežni enoti*:
    - T_i: čas prebivanja i-te zahteve v strežni enoti (angl. *response time*);
    - T: povprečni čas prebivanja zahteve v strežni enoti (angl. *average response time*);
  - *čas prebivanja zahteve v čakalni vrsti*:
    - W_i: čas prebivanja i-te zahteve v vrsti strežne enote (angl. *queueing delay*);
    - W: povprečni čas prebivanja zahteve v vrsti strežne enote (angl. *average queueing delay*);
  - *čas strežbe zahteve v strežniku*:
    - x_i: čas strežbe i-te zahteve v strežniku (angl. *service time*);
    - x: povprečni čas strežbe zahteve v strežniku (angl. *average service time*);

- *verjetnostno pogojeni metriki*:
  - $P_k(t)$: verjetnost nahajanja k zahtev v strežni enoti v opazovanem času t;
  - P_k: stacionarna verjetnost prebivanja k zahtev v strežni enoti;

Skupno število zahtev v strežni enoti je odvisno od števila zahtev v čakalni vrsti in števila zahtev v strežbi po izrazih

$$N(t) = N_q(t) + N_s(t), \tag{2.15}$$

$$N = N_q + N_s, \tag{2.16}$$

Prvi izraz odraža stanje v opazovani časovni točki t, drugi pa odraža časovno povprečje. Čas prebivanja zahteve v strežni enoti določimo na osnovi časa prebivanja zahteve v čakalni vrsti in časa strežbe po izrazih

$$T_i = W_i + x_i, \tag{2.17}$$

$$T = W + x, \tag{2.18}$$

Prvi izraz odraža čas glede na opazovano i-to zahtevo, drugi pa odraža povprečje glede na večje število opazovanih zahtev. S tem smo močno razširili paleto kvantitativnih metrik za ocenjevanje zmogljivosti opazovane strežne enote.

Na tem mestu navedimo nekaj primerov uporab kvantitativnih metrik pri analizi dinamike modela računalniškega omrežja. Ciljni postulati analitika modela omrežja z vidika zagotavljanja zadovoljstva uporabnikov bi bili npr. tako lahko sledeči:

- spremljajmo vrednosti metrike $N_q(t)$ in preverjajmo, ali njena maksimalna vrednost v modelu presega velikost dolžine predvidene čakalne vrste; če jo, bo po vsej verjetnosti v realnem omrežju prihajalo do izgubljanja paketov;
- želimo si čim krajših časov zadrževanja paketov na fizičnih prenosnih medijih in v omrežnih napravah (želimo si čim manjše vrednosti metrike T, s tem pa posredno tudi čim manjše vrednosti metrik W in x);
- razmerje med vrednostima metrik W in x naj bi bilo čimmanjše (čas čakanja v vrsti W naj bi bil zanemarljiv proti času strežbe zahteve x, če le x ni zanemarljivo majhen);
- izračunajmo verjetnost nahajanja k zahtev v omrežju na osnovi metrike $P_k(t)$ in na osnovi te verjetnosti izračunajmo verjetnost izgubljanja paketov itd.;

S tem smo navedli nekaj iztočnic za delo analitika modela računalniškega omrežja.

2.4.3 Littlovo pravilo

Ena od osnovnih značilnosti strežne enote je veljavnost izraza

$$N = T * \lambda, \tag{2.19}$$

ki ga imenujemo za *Littlovo pravilo* [1]. Omenjena relacija je ob pravilni interpretaciji spremenljivk N, λ in T veljavna za vse vrste strežnih enot.

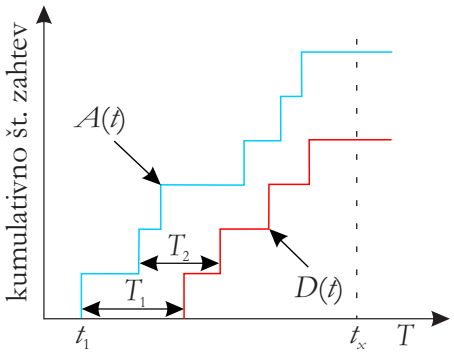
Poizkusimo dokazati veljavnost relacije iz izraza (2.19). Opazujoč število vstopajočih in izstopajočih zahtev iz strežne enote pridemo do stopničastih funkcij $A(t)$, ki predstavlja kumulativno število prispelih zahtev v časovnem intervalu $[0, t_x]$ in $D(t)$, ki predstavlja kumulativno število zahtev, ki so v časovnem intervalu $[0, t_x]$ strežno enoto zapustile. Za navedeni funkciji velja relacija ponazorjena z izrazom

$$\forall t \in T : A(t) \geq D(t), \tag{2.20}$$

saj do neke poljubne časovne točke t iz strežne enote ne more izstopiti več zahtev, kot jih je vanjo vstopilo. Po definiciji strežne enote se namreč v njej zahteve ne morejo porojevati. Če obe kumulativni funkciji ponazorimo grafično (glej sliko 2.9) ugotovimo, da lahko $N(t)$ izračunamo na osnovi izraza (2.21) kot razliko površin obeh funkcij.

$$\int_0^{t_x} N(t)dt = \int_0^{t_x} A(t)dt - \int_0^{t_x} D(t)dt =$$
$$= \sum_{k=1}^{D(t_x)} T_k * 1 + \sum_{k=D(t_x)+1}^{A(t_x)} (t_x - t_k) * 1, \tag{2.21}$$

Pri tem t_k predstavlja čas rojstva posamezne zahteve, T_k pa celotni čas njenega prebivanja v strežni enoti. Odtod pridemo do izraza (2.22), kjer obe strani



Slika 2.9: Kumulativni funkciji prihodov $A(t)$ in odhodov $D(t)$ zahtev iz strežne enote.

enačbe delimo z dolžino opazovanega časovnega intervala t_x. Tako leva stran tega izraza določa povprečno vrednost števila zahtev v strežni enoti N skozi čas opazovanja. Na desno stran izraza (2.22) prenesemo vsoto iz izraza (2.21), ki jo utežimo z enoto $(\frac{A(t)}{A(t)})$ in delimo z dolžino časovnega intervala opazovanja kot na levi strani izraza.

$$\frac{1}{t_x} * \int_0^{t_x} N(t)dt = \left(\sum_{k=1}^{D(t_x)} T_k + \sum_{k=D(t_x)+1}^{A(t_x)} (t_x - t_k) \right) * \frac{1}{A(t)} * \frac{A(t)}{t_x}. \quad (2.22)$$

Levi del izraza (2.22) je enak N, skrajno desni člen $\frac{A(t)}{t_x}$ intenzivnosti prihajanja zahtev λ, preostanek desnega dela enačbe pa povprečnemu času bivanja zahteve v strežni enoti T na časovnem intervalu $[0, t_x]$. Vsota, ki definira povprečni čas bivanja zahteve v strežni enoti je sestavljena iz levega dela (vsote časov prebivanj zahtev, ki so strežno enoto že zapustile) in desnega dela (vsote časov prebivanj zahtev, ki svojega bivanja v strežni enoti do časovne točke t_x še niso zaključile). S tem je dokaz veljavnosti Littlovega pravila končan.

Littlovo pravilo iz izraza (2.19) se nanaša na strežno enoto v celoti, lahko pa ga projeciramo le na čakalno vrsto ali strežnik. Tako pridemo do izrazov

$$N_q = \lambda * W, \quad (2.23)$$

$$N_s = \lambda * x. \quad (2.24)$$

2.4.4 Uporabnostni faktor strežne enote

Uporabnostni faktor ρ (angl. *utilization factor*) strežne enote nam pove, kolikšna je odstotkovna zasedenost strežnikov v strežni enoti. Izračunamo ga kot razmerje med resničnim časom zasedenosti ali delovanja enega ali več strežnikov v strežni enoti in časom opazovanja (časovnim intervalom) strežne enote.

Uporabnostni faktor strežne enote z enim strežnikom izračunamo z izrazom

$$\rho = \frac{\lambda}{\mu}, \tag{2.25}$$

uporabnostni faktor strežne enote z m strežniki pa z izrazom

$$\rho = \frac{\lambda}{m * \mu}. \tag{2.26}$$

Strežne enote, za katere velja relacija $\rho < 1$ imenujemo za *stabilne*, vse ostale, za katere ta relacija ne velja, pa za *nestabilne*.

Če imamo v strežni enoti več strežnikov, slednje lahko opazujemo skozi njim lastne uporabnostne faktorje ρ_i $(i = 1, ..., m)$. V tem primeru za opazovani i-ti strežnik velja psevdo izraz

$$\rho_i = \frac{\text{Čas zasedenosti } i\text{-tega strežnika}}{\text{Celotni čas opazovanja } i\text{-tega strežnika}}. \tag{2.27}$$

Če je časovni interval opazovanja m strežnikov dovolj dolg lahko pričakujemo, da bodo ρ_i med seboj dokaj podobni, v večini primerov pa vseeno ne enaki.

Na tem mestu skušajmo z vidika uporabnostnih faktorjev strežnih enot (resursov) v omrežju osvetliti temeljna, med seboj nasprotujoča si cilja *uporabnika* in *lastnika* računalniškega omrežja. Cilj prvega je doseganje čim manjših uporabnostih faktorjev omrežnih strežnih enot, kar se manifestira v maksimalnih možnih hitrostih prenosov, cilj drugega pa doseganje čim večjih uporabnostih faktorjev omrežnih strežnih enot, kar se manifestira z manjšo ceno omrežja (z manj zmogljivimi strežnimi enotami in/ali manjšim številom strežnikov v strežnih enotah). Osnovno vodilo področja *upravljanja z omrežji* je zagotavljanje takšnih konfiguracij strežnih enot, ki kompromisno zadovoljujejo tako uporabnika, kot tudi lastnika omrežja.

Izkušnje kažejo (angl. *rule of thumb*), da se naj uporabnostni faktor posamezne strežne enote v modelu omrežja ali v realnem omrežju ne bi povzpel preko vrednosti 0.7[8]. V primeru, da do te prekoračitve v neki strežni enoti pride, dotična strežna enota postane „**ozko grlo omrežja**“, uporabniki pa to občutijo kot izrazito upočasnjeno (*degradirano*) delovanje ali kot odpoved delovanja celotnega omrežja. Slednje vodi v nezadovoljstvo uporabnikov in s tem posledično do eventuelnih finančnih posledic (npr. odstopa uporabnikov od naročniške pogodbe).

Omenjeni situaciji se izgonemo s skalabilnostjo (z zvečevanjem števila enako zmogljivih strežnikov v dotični strežni enoti) ali s povečanjem zmogljivosti posameznih strežnikov v strežni enoti (z zamenjavo dotične strežne enote z bolj zmogljivo strežno enoto).

[8] https://www.starstandard.org/images/guidelines/DIG2012v1/ch02s02.html
https://www.johndcook.com/blog/2009/01/30/server-utilization-joel-on-queuing/

Sorodna metrika uporabnostnemu faktorju je *intenzivnost prometa* (angl. *traffic intensity)* v strežni enoti. Tudi ta se izračuna kot kvocient med intenzivnostjo prihajanja zahtev in intenzivnostjo servisiranja zahtev. Z razliko od uporabnostnega faktorja ima mersko enoto. V primeru intenzivnosti prometa uporabljamo mersko enoto *Erlang*. Enota enega *Erlanga* ponazarja metrično značilnost strežne enote, v katero vstopa na časovno enoto natanko toliko zahtev, kot jih je le ta v tem času zmožna postreči. Strežne enote z intenzivnostjo prometa več *Erlangov* zahtevajo vključevanje dodatnega števila strežnikov, da bi mera padla pod 1 *Erlang*. Strežna enota z enim strežnikom in intenzivnostjo prometa 12,4 *Erlangov* bi tako zahtevala dodajanje nadaljnjih dvanajstih strežnikov v njeno paralelno vezavo, da bi strežba potekala nemoteno in bi bil dosežen kriterij stabilnosti ($\rho < 1$).

2.4.5 Zakon o ohranitvi pretoka

Zakon o ohranitvi pretoka (angl. *flow conservation law*) pravi, da je intenzivnost prihajanja zahtev v *stabilno strežno enoto* na daljši rok opazovanja enaka intenzivnosti odhajanja zahtev iz takšne strežne enote.

Ob dokazovanju veljavnosti zakona izhajamo iz upoštevanja intenzivnosti prihajanja a in odhajanja b iz strežne enote (označbi omenjenih točk sta predstavljeni na sliki 2.10). Ob tem lahko naredimo naslednja zaključka [1]:

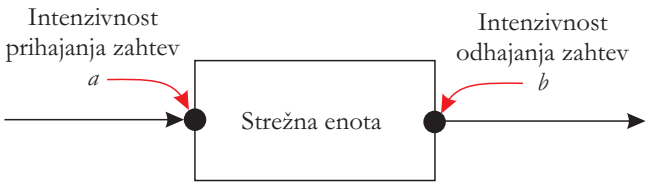
- če bi veljala relacija $a > b$, potem bi imeli v strežni enoti vse več zahtev (število zahtev bi v čakalnih vrstah vseskozi naraščalo) in s časom bi postala nestabilna, saj poljubno dolgih čakalnih vrst v realnih strežnih sistemih ne moremo realizirati; pravimo, da taka strežna enota preide v *nasičenje*, saj število zahtev v njej narašča preko vseh meja; ker zakon o ohranitvi pretoka velja le za stabilne strežne enote lahko pridemo do sklepa, da relacija $a > b$ za stabilne enote ne velja;

- če bi veljala relacija $a < b$, bi to pomenilo, da se nam v notranjosti strežne enote porajajo nove zahteve, česar v svoji definiciji strežna enota ne predvideva; odtod pridemo do sklepa, da relacija $a < b$ za strežne enote ne more veljati;

Ob neveljavnosti predpostavk iz prve in druge alineje pridemo do sklepa, da pri dovolj dolgem času opazovanja stabilne strežne enote velja kot edina možna relacija $a = b$, kar potrjuje veljavnost zakona o ohranitvi pretoka v stabilnih strežnih enotah.

2.5 Poissonov proces

Omenili smo že, da za mnoge realne strežne sisteme velja, da so njihovi medprihodni in strežni časi porazdeljeni eksponentno [1]. Povedano drugače, sta tako prihajalni kot tudi strežni proces Poissonova procesa, zato si bomo v pričujočem razdelku ogledali njegove osnove.



Slika 2.10: Strežna enota s ponazorjenima točkama opazovanja intenzivnosti prihajanja a in odhajanja b zahtev.

Poissonov proces si najlažje interpretiramo kot proces štetja naključno porajajočih se k dogodkov na časovnem intervalu $[0,t]$. Naključna ali slučajna spremenljivka $X(t)$, ki predstavlja število med seboj neodvisno porajajočih se dogodkov na časovnem intervalu $[0,t]$, je porazdeljena po *Poissonovi verjetnostni porazdelitvi* na osnovi izraza

$$P(X(t)=k)=\frac{(\lambda t)^k}{k!}e^{-\lambda t}. \tag{2.28}$$

Pri tem produkt $\lambda * t$ imenujemo za povprečje Poissonove slučajne spremenljivke, pomensko pa predstavlja povprečno število pojavitev dogodkov v časovnem intervalu $[0,t]$. $P(X(t)=k)$ predstavlja verjetnost porajanja k zahtev na opazovanem časovnem intervalu $[0,t]$, parameter λ pa intenzivnost porajanja zahtev na vhodu (λ) ali intenzivnost strežbe (μ). Slika 2.11 prikazuje funkcije gostote verjetnosti (angl. *probability density function*) izračunane s funkcijo `poisspdf(X,λ)` v okolju Matlab za različne intenzivnosti porajanja λ.

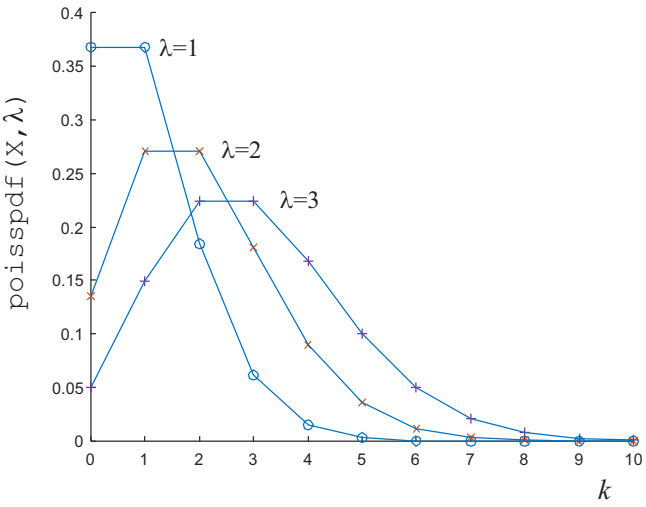
Poizkušajmo izraz (2.28) matematično še sami izpeljati, pri čemer bomo upoštevali dejstvo, da je Poissonov proces poseben primer binomske porazdelitve Bernoullijevega eksperimenta [1]. Predpostavimo, da je časovni interval $[0,t]$ razdeljen na n podintervalov, ki so tako kratki, da se v njih lahko porodi le en dogodek. Upoštevajoč k porajanj dogodkov se binomska porazdelitev izračuna po izrazu

$$P\left[k\ dogodkov\ v\ n\ podintervalih\right]=\left(\begin{array}{c}n\\k\end{array}\right)p^k(1-p)^{n-k}, \tag{2.29}$$

pri čemer je p verjetnost porajanja dogodka v posameznem podintervalu. Če na intervalu $[0,t]$ povečamo število podintervalov n in istočasno zmanjšamo verjetnost p, tako da povprečno število porajanih dogodkov ostane nespremenjeno (veljati mora $n*p=\lambda*t$), pridemo z nekaj koraki izpeljave do Poissonove porazdelitve. Omenjena izpeljava je podana z izrazom

$$P\left[k\ dogodkov\ v\ n\ podintervalih\right]=\lim_{n\to\infty}\left(\begin{array}{c}n\\k\end{array}\right)\left(\frac{\lambda t}{n}\right)^k\left(1-\frac{\lambda t}{n}\right)^{n-k}=$$

$$=\frac{(\lambda t)^k}{k!}\left[\lim_{n\to\infty}\frac{n(n-1)...(n-k+1)}{n^k}\right]\left[\lim_{n\to\infty}\left(1-\frac{\lambda t}{n}\right)^n\right]=$$



Slika 2.11: Funkcije gostote verjetnosti za različne vrednosti parametra λ ($\lambda = 1$, $\lambda = 2$, $\lambda = 3$).

$$= \frac{(\lambda t)^k}{k!} \left[\lim_{n\to\infty} \left\{ \left(1 - \frac{\lambda t}{n}\right)^{\frac{n}{\lambda t}} \right\}^{-\lambda t} \right] = \frac{(\lambda t)^k}{k!} e^{-\lambda t}. \tag{2.30}$$

Pred pregledom lastnosti Poissonove porazdelitve si oglejmo zgled njene uporabe v domeni strežnih enot.

Zgled 2 *Opravka imamo s strežno enoto, v katero vstopajo zahteve po Poissonovi porazdelitvi z intenzivnostjo prihajanja $\lambda = 10$ zahtev na časovno enoto. Zanima nas, kolikšne so naslednje verjetnosti:*

(i) kolikšna je verjetnost, da je v 1 časovni enoti v strežno enoto vstopilo 0 zahtev?

(ii) kolikšna je verjetnost, da je v 1 časovni enoti v strežno enoto vstopilo 10 zahtev?

(iii) kolikšna je verjetnost, da je v 1 časovni enoti v strežno enoto vstopilo 100 zahtev?

Rešitev: *Intenzivnost prihajanja λ nam služi kot parameter za evaluacijo funkcije Poissonove porazdelitve iz izraza (2.28). Odgovore na vprašanja dobimo po sledečih izrazih:*

(i) uporabimo izraz

$$P(X(t) = 0) = \frac{(10 * 1)^0}{0!} e^{-10*1} \approx 0,00005; \tag{2.31}$$

(ii) uporabimo izraz

$$P(X(t) = 10) = \frac{(10 * 1)^{10}}{10!} e^{-10*1} \approx 0,1251; \tag{2.32}$$

(iii) uporabimo izraz

$$P(X(t) = 100) = \frac{(10 * 1)^{100}}{100!} e^{-10*1} \approx 0,0049 * 10^{-60}; \tag{2.33}$$

2.5.1 Povezanost Poissonove in eksponentne porazdelitve

V predhodnih razdelkih smo si ogledali eksponentno in Poissonovo verjetnostno porazdelitev, sedaj pa bomo opisali še njuno medsebojno povezanost.

Poissonova porazdelitev sodi v družino *diskretnih porazdelitev*, za katere je značilno, da je zaloga vrednosti slučajne spremenljivke diskretna, eksponentna porazdelitev pa v družino *zveznih porazdelitev*, za katere je značilno, da je zaloga vrednosti slučajne spremenljivke zvezna. V prvem primeru je pomen diskretne slučajne spremenljivke diskretno število dogodkov, v drugem primeru pa pomen zvezne slučajne spremenljivke zvezna časovna vrednost.

Eksponentna verjetnostna porazdelitev podaja čase med porajanji dogodkov, ki se porajajo po Poissonovi porazdelitvi. Predpostavimo, da po Poissonovi porazdelitvi izračunamo verjetnost, da v časovnem intervalu $[0, t]$ ni prišlo do porajanja niti ene zahteve. Na levo stran omenjenega izraza postavimo člen $P(X > t)$, ki opisuje natanko to - čas porajanja dogodka bo nastopil šele po pretečenem času t. Slednje strnimo v izraz

$$P(X > t) = P(X(t) = 0) = \frac{(\lambda t)^0}{0!} e^{-\lambda t} = e^{-\lambda t}. \tag{2.34}$$

Odtod lahko neposredno sklepamo na veljavnost izraza

$$P(X \leq t) = 1 - P(X > t) = 1 - e^{-\lambda t}, \tag{2.35}$$

s čimer smo prišli do izrazov (2.7), (2.8) in s tem povezali obe porazdelitvi.

V splošnem velja, da če so medprihodni časi med dogodki porazdeljeni eksponentno, se potemtakem dogodki porajajo po Poissonovi porazdelitvi in obratno, da če se dogodki porajajo po Poissonovi porazdelitvi, so potemtakem njihovi medprihodni časi porazdeljeni eksponentno.

2.5.2 Lastnosti Poissonovega procesa

Ključni lastnosti Poissonovega procesa sta njegova *superpozicija* in njegova *nezmožnost pomnjenja*.

Pod pojmom superpozicije smatramo, da ob predpostavki, da imamo k neodvisnih Poissonovih procesov, ob njihovi združitvi v nov enoten proces velja,

da je tudi nov proces Poissonov proces. Intenzivnost porajanja zahtev novega procesa je enaka vsoti intenzivnosti porajanj posameznih Poissonovih procesov.

Poissonov proces je *brez pomnjenja*. Če takšen proces opazujemo v neki časovni točki t in poznamo časovno točko zadnjega porajanja zahteve $t - \Delta t$, ne moremo določiti, kdaj v prihodnosti bo prispela nova zahteva, poznamo pa verjetnost njenega porajanja. Značilnost nezmožnosti pomnjenja (angl. *memoryless*) pomeni to, da na osnovi stanja sistema (npr. časovne točke porajanja zadnje zahteve), ne moremo napovedati novega stanja sistema (npr. časovne točke porajanja nove zahteve). Omenjena lastnost ne velja samo za časovne točke porajanja posameznih zahtev, temveč tudi za medprihodne čase porajajočih se zahtev.

2.6 Stohastičen proces

Stohastičen, *slučajen* ali *naključen* proces (angl. *stochastic process*) je dinamičen proces, čigar dinamika je vsaj do neke mere pogojena z naključnostjo (naključnimi dogodki), ki jo opišemo s pomočjo verjetnosti. V nadaljevanju podajamo definicijo stohastičnega procesa, povzeto po viru [6].

Definicija 4 *Stohastičen proces je dinamičen proces, čigar dinamiko pogojuje družina slučajnih spremenljivk* $\{X(t),\ t \in T\}$ *definiranih v verjetnostnem prostoru, pri čemer* X *predstavlja slučajno spremenljivko,* T *pa indeksno množico posamezne slučajne spremenljivke, s katero slučajno spremenljivko indeksiramo.*

Posamezno vrednost slučajne spremenljivke $X(t)$ imenujemo za **stanje** opazovanega procesa, množico vseh možnih vrednosti slučajne spremenljivke pa za **prostor stanj** opazovanega procesa. Prostor stanj je lahko *diskreten* ali *zvezen*. Če je prostor stanj diskreten (angl. *discrete-state process*) stohastični proces poimenujemo za **stohastično verigo** (angl. *stochastic chain*), če pa je prostor stanj zvezen stohastični proces poimenujemo za **stohastični proces z zvezno zalogo stanj**[9] (angl. *continuous-state process*).

Parameter t predstavlja časovni parameter, ki izhaja iz indeksne množice T ($t \in T$, $T \subseteq \mathbb{R}_+ = [0, \infty)$), uporabljamo pa ga za časovno referenciranje vrednosti slučajnih spremenljivk. V primeru, da je indeksna množica T diskretna, stohastični proces poimenujemo za **diskretni časovni stohastični proces** (angl. *discrete-time process*), v primeru da je indeksna množica T zvezna, pa stohastični proces poimenujemo za **zvezni časovni stohastični proces** (angl. *continuous-time process*). Diskretni časovni stohastični proces imenujemo tudi za **stohastično sekvenco** (angl. *stochastic sequence*). V primeru,

[9] Žal za angleški pojem "*continuous-state process*"v slovenskem jeziku nimamo krajšega izraza, kot je "*proces z zvezno zalogo stanj*".

	Diskretni prostor stanj	**Zvezni prostor stanj**
Diskretna indeksna množica T	*diskretna časovna stohastična veriga*	*diskretni časovni stohastični proces z zvezno zalogo stanj*
Zvezna indeksna množica T	*zvezna časovna stohastična veriga*	*zvezni časovni stohastični proces z zvezno zalogo stanj*

Tabela 2.1: Klasifikacija stohastičnih procesov glede na vrsto prostora stanj in vrsto indeksnega parametra T.

da je množica T diskretna, za indeksiranje slučajne spremenljivke uporabljamo notacijo $\{X_n, n \in T\}$, v primeru da je množica T zvezna, pa uporabljamo notacijo $\{X(t), t \in T\}$. V prvem primeru proces spreminja svoje stanje v diskretnih časovnih korakih, v drugem primeru pa lahko proces svoje stanje spremeni kadarkoli na zvezni časovni osi. Spremembo ali menjavo stanja procesa imenujemo *tranzicija* ali *prehajanje*.

Iz obrazložitve obravnave indeksne množice T (časa) in zaloge vrednosti slučajne spremenljivke pridemo do zaključka, da stohastične procese delimo na štiri različne skupine, kot je ponazorjeno v tabeli 2.1. Le te so *diskretne časovne stohastične verige*, *diskretni časovni stohastični procesi z zvezno zalogo stanj*, *zvezne časovne stohastične verige* in *zvezni časovni stohastični procesi z zvezno zalogo stanj*.

Z vidika stohastičnosti zveznih časovnih procesov nas največkrat zanima verjetnost $P[X(t) = i]$, da bomo v času t dosegli stanje i, z vidika stohastičnosti diskretnih časovnih procesov pa verjetnost $P[X_n = i]$, da bomo na n-tem koraku dosegli stanje i.

Stohastičnost je pri modeliranju prometa v računalniških omrežjih vsekakor prisotna. Slučajna spremenljivka nastopa tako v domeni prihajanja zahtev in medprihodnih časov, kot tudi v domeni strežnih časov.

Pri gradnji modelov omrežij se v domeni časa odločamo med uporabo diskretnega ali zveznega časa, v domeni prostora stanj modelov pa med uporabo diskretne ali zvezne zaloge stanj. Najpogosteje velja, da modele prometa zahtev (paketov) v računalniških omrežjih zasnujemo kot stohastične verige (uporabimo diskretni prostor stanj).

V nadaljevanju poglavja si bomo glede na povedano ogledali diskretne in zvezne časovne Markovske verige.

2.7 Diskretna časovna Markovska veriga

Markovski[10] *proces* je posebna oblika stohastičnega procesa za katerega velja, da "*nima pomnjenja*". Slednje pomeni, da je vpliv na naslednje stanje procesa ali menjavo njegovega stanja pogojen le s trenutnim stanjem procesa, ne pa s predhodnimi stanji tega procesa.

Diskretna časovna Markovska veriga predstavlja stohastični proces z diskretnim časom in diskretnim prostorom stanj. Za diskretne časovne Markovske verige velja definicija, povzeta po viru [1].

Definicija 5 *Stohastično zaporedje* $\{X_k, k \in T\}$ *imenujemo za diskretno časovno Markovsko verigo, če za vsak nabor indeksov* i, j *in* k *velja pogojna verjetnost iz izraza*

$$P\left[X_{k+1} = j \mid X_0 = i_0, X_1 = i_1, ..., X_k = i_k\right] = \\ = P\left[X_{k+1} = j \mid X_k = i_k\right] = p_{ij}. \tag{2.36}$$

Pri tem izraz $X_{k+1} = j$ pomeni, da je veriga na $k+1$ koraku v j-tem stanju, p_{ij} pa predstavlja verjetnost neposrednega prehajanja iz i-tega v j-to stanje procesa. Le ta je skozi čas lahko konstantna ali spremenljiva. V primeru nespremenljive verjetnosti p_{ij} govorimo o konstantnih prehajalnih verjetnostih, takšno verigo pa imenujemo za *časovno homogeno* Markovsko verigo. V pričujočem delu bomo obravnavali samo časovno homogene Markovske verige. Običajno množico stanj Markovske verige označujemo z vrednostmi nenegativnih celih števil $\{0, 1, 2, ...\}$.

Zgled 3 *Kot primer diskretne časovne Markovske verige navedimo sekvenco desetih metov kovanca. V spremenljivko* X *(*$x_i \in X, i = 1, .., 10$*) beležimo izide metov (1 za grb in 0 za cifro), v spremenljivko* Y *(*$y_i \in Y, i = 1, .., 10$*) pa kumulativo vrednosti spremenljivke* X *po izrazu*

$$y_0 = 0; y_i = y_{i-1} + x_i, i = 1, ..., 10. \tag{2.37}$$

Spremenljivka X *je tako stohastična, spremenljivka* Y *pa izkazuje Markovsko lastnost procesa.*

Verjetnost nahajanja verige v stanju j na k-tem koraku bomo v nadaljevanju označevali z izrazom

$$\pi_j^{(k)} \equiv P\left[X_k = j\right]. \tag{2.38}$$

[10] Ime procesa je povzeto po ruskem matematiku Andreju Andrejeviču Markovu (1856-1922), ki je bil eden od pionirjev raziskav na področju teorije stohastičnih procesov.

Veriga skozi diskretne časovne korake menja svoje stanje (prihaja do prehajanja med stanji), kar lahko grafično ponazorimo z *diagrami prehajanja stanj*. Prehajanja med stanji so možna le ob vnaprej določenih diskretnih korakih. Ob predpostavki, da poznamo začetno stanje verige (začetne verjetnosti posameznih stanj verige), lahko na osnovi izraza

$$P\left[X_{k+1}=j\right]=\sum_{i=0}^{\infty}P\left[X_k=i\right]*P\left[X_{k+1}=j\mid X_k=i\right]=\sum_{i=0}^{\infty}\pi_i^{(k)}*p_{ij} \quad (2.39)$$

izračunamo verjetnost nahajanja verige v kateremkoli stanju na poljubnem časovnem koraku.

2.7.1 Matrični zapis verjetnosti stanj

V primeru, da je Markovska veriga časovno homogena, se njene prehajalne verjentosti p_{ij} skozi čas ne spreminjajo. Ob predpostavki, da je število različnih možnih stanj verige končno (predpostavimo, da je to število n), lahko prehajalne verjetnosti zapišemo v obliki matrike reda $n*n$. Imenovali jo bomo za *matriko verjetnosti prehajanj* M

$$M=(p_{ij}),\ 1\leq i\leq n,\ 1\leq j\leq n. \quad (2.40)$$

Za tovrstno matriko veljata relaciji v izrazih

$$\forall i:\sum_{j=1}^{n}p_{ij}=1, \quad (2.41)$$

$$\forall i,j:0\leq p_{ij}\leq 1. \quad (2.42)$$

Poleg matrike verjetnosti prehajanj potrebujemo za izračune še matematični zapis začetnega stanja verige, za katerega smo predhodno že povedali, da se izraža kot porazdelitev verjetnosti glede na opazovani prostor stanj. Izhajajoč iz predhodno omenjenega izraza $\pi_j^{(k)}$, ki odraža verjetnost nahajanja verige v stanju j na k-tem koraku, ob n stanjih sistema začetno porazdelitev verjetnosti posameznih stanj verige zapišemo v obliki vektorja $\pi^{(0)}$ z izrazom

$$\pi^{(0)}=(\pi_1^{(0)},\pi_2^{(0)},...,\pi_n^{(0)}). \quad (2.43)$$

Z zapisom vektorja $\pi^{(k)}$ lahko pridemo do porazdelitve verjetnosti stanj na poljubnem diskretnem časovnem koraku k, iz njega pa lahko venomer odčitamo verjetnost stanja j ($\pi_j^{(k)}$) na tem istem časovnem koraku. Porazdelitev verjetnosti stanj verige na k-tem časovnem koraku lahko izračunamo upoštevajoč enačbi (2.38) in (2.39) z zaporedjem izrazov

$$\pi^{(1)}=\pi^{(0)}M,$$

$$\pi^{(2)}=\pi^{(1)}M,$$

.....

$$\pi^{(k)} = \pi^{(k-1)} M. \tag{2.44}$$

S povratno substitucijo [1] posameznih členov π tako pridemo do izraza

$$\pi^{(k)} = \pi^{(0)} M^k, \tag{2.45}$$

kjer M^k predstavlja k-to potenco matrike M, M^0 pa enotsko matriko ($M^0 = I$). Ker velja pri matričnih operacijah relacija

$$M^{k+1} = M^k * M, \tag{2.46}$$

velja tudi relacija v izrazu

$$p_{ij}^{k+1} = \sum_{k=0}^{n} p_{ik}^k p_{kj}, \tag{2.47}$$

ki jo imenujemo za Chapman-Kolmogorovo enačbo.

2.7.2 Stacionarna stanja v diskretnih časovnih Markovskih verigah

V večini sistemov iz realnega okolja, ki jih modeliramo z diskretnimi časovnimi Markovskimi verigami, obstajajo neke limitne vrednosti verjetnosti stanj, ki niso odvisne od njihove začetne porazdelitve. V tem primeru pravimo, da imajo sistemi *stacionarne verjetnosti stanj* ali *stacionarna stanja*. Če obstaja limitna vrednost $\lim_{k\to\infty} M_{ij}^k = \pi_j$, potem lahko naredimo sledečo izpeljavo

$$\lim_{k\to\infty} \pi_j^{(k)} = \sum_{i=1}^{n} \pi_i^{(0)} M_{ij}^k = \pi_j \sum_{i=1}^{n} \pi_i^{(0)} = \pi_j \tag{2.48}$$

in zapišemo definicijo o stacionarnih stanjih v diskretnih časovnih Markovskih verigah povzeto po viru [1].

Definicija 6 *Diskretna časovna Markovska veriga $\{X_k\}$, ki je aperiodična, nereducibilna in časovno homogena, je tudi ergodična. Za ergodične Markovske verige vedno obstajajo limitne verjetnosti po izrazu*

$$\pi_j = \lim_{k\to\infty} \pi_j^{(k)} = \lim_{k\to\infty} P\left[X_k = j\right], j = 0, 1, ..., n, \tag{2.49}$$

ki so neodvisne od začetnega stanja verjetnostne porazdelitve $\pi^{(0)}$.

Za stacionarno verjetnost π_j veljata izraza

$$\sum_{j=1}^{n} \pi_j = 1, \tag{2.50}$$

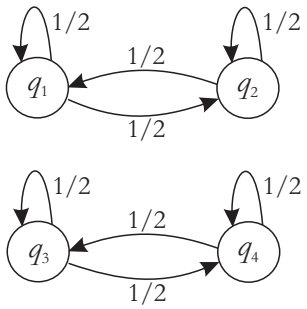
$$\pi_j = \sum_{i=1}^{n} \pi_i p_{ij}. \tag{2.51}$$

Doslej smo že spoznali lastnost *časovne homogenosti* diskretne časovne Markovske verige, za določanje ergodičnosti sistema in s tem posredno obstoja limitnih verjetnosti pa moramo spoznati še njeni lastnosti *nereducibilnosti* in *aperiodičnosti*.

Markovska veriga je *reducibilna*, če vsebuje več kot eno izolirano podmnožico stanj. Podmnožica stanj je izolirana, kadar iz poljubnega stanja te podmnožice ni mogoče preiti v kako drugo stanje sistema, ki ni v tej podmnožici in ko v nobeno stanje te podmnožice ne vodi povezava iz ostalih stanj sistema izven opazovane podmnožice. Primer reducibilne Markovske verige je podan z matriko verjetnosti prehajanj M po izrazu

$$M = \frac{1}{2} * \begin{bmatrix} 1 & 1 & 0 & 0 \\ 1 & 1 & 0 & 0 \\ 0 & 0 & 1 & 1 \\ 0 & 0 & 1 & 1 \end{bmatrix}, \tag{2.52}$$

možna prehajanja med stanji opisane Markovske verige pa z *diagramom prehajanja stanj* prikazanim na sliki 2.12. Iz slike je razvidno, da je Markovska veriga



Slika 2.12: Diagram prehajanja stanj na osnovi podane matrike verjetnosti prehajanj M.

M sestavljena iz dveh izoliranih podmnožic $\{q_1, q_2\}$, $\{q_3, q_4\}$. Veriga je torej reducibilna, s čimer ni izpolnjen pogoj njene nereducibilnosti in s tem posledično veriga tudi ni ergodična (nima stacionarnih verjetnosti). Za zadostitev pogoju nereducibilnosti bi morala imeti veriga le eno izolirano podmnožico stanj, v kateri bi bila vsa stanja sistema.

Markovska veriga je *periodična* s periodo τ, če se po $k * \tau$ $(k = 1, 2, ...)$ korakih vrača v isto stanje sistema. V nereducibilni Markovski verigi so vsa stanja bodisi aperiodična ali pa periodična z isto periodo.

Od izraza (2.41) naprej smo se v vsotah sklicevali na končno število stanj Markovske verige n. V splošnem bi namesto oznake n lahko uporabljali tudi znak ∞.

Na tem mestu se sprašujemo, kakšno povezavo ima diskretna časovna Markovska veriga s področjem modeliranja in teorije strežbe. V primeru diskretnega časa vlogo takšnega procesa prevzamejo njegove opazovane diskretne spremenljivke ali njegova diskretna stanja. Takšne spremenljivke bi tako lahko bile število zahtev v strežni enoti (N), število zahtev v čakalni vrsti (N_q), število zahtev v strežniku (N_s) itd.

2.8 Zvezna časovna Markovska veriga

Če v Markovsko verigo namesto diskretnega časa vpeljemo zvezni čas, se prehajanje med stanjema lahko izvrši v poljubni točki zveznega časovnega intervala. Še vedno velja, da je takšen proces stohastičen proces $\{X(t)\}$, v katerem je prehod v novo stanje odvisen le od trenutnega stanja in ne od stanj v preteklosti - torej je proces brez pomnenja. Slednje zapišemo z izrazom

$$P\left[X(t_{k+1}) = j \mid X(t_1 = i_1), X(t_2 = i_2), ..., X(t_k = i_k)\right] =$$

$$= P\left[X(t_{k+1}) = j \mid X(t_k = i_k)\right],\ t_1 < t_2 < ... < t_k < t_{k+1}. \tag{2.53}$$

Predpostavimo, da se Markovska veriga nahaja v stanju i. Verjetnost, da takšen proces izvede prehod v stanje j v infinitezimalnem času[11] Δt, ne glede na to koliko časa je bil proces v stanju i, je

$$p_{ij}(t, t + \Delta t) = q_{ij}\Delta t, \tag{2.54}$$

kjer q_{ij} predstavlja intenzivnost prehajanja iz stanja i v stanje j. Celotna intenzivnost zapuščanja stanja i se izraža po izrazu

$$\sum_{j \neq i} q_{ij}. \tag{2.55}$$

Čas, v katerem se sistem nahaja v stanju i, imenujemo za *čas zadrževanja procesa* v stanju i.

Naj bo τ_i čas prebivanja procesa v stanju i. Predpostavimo, da časovni interval $[0, t]$ razdelimo na k enako dolgih podintervalov dolžine Δt $(t = k * \Delta t)$. Če je t manjši od τ_i, potem v opazovanem intervalu v procesu ni prišlo do spremembe stanja. Verjetnost, da do menjave stanja v posameznem podintervalu

[11] Infinitezimala ali infinitezimalno majhna količina je v matematiki oznaka za količino, ki je po absolutni vrednosti zelo majhna, vendar ni enaka 0 (vir: Wikipedia).

ne pride, se izračuna po izrazu

$$1 - \sum_{j \neq i} q_{ij} \Delta t. \tag{2.56}$$

Ob izražavi intenzivnosti zapuščanja stanja i po izrazu

$$q_i = \sum_{j \neq i} q_{ij} \tag{2.57}$$

lahko izpeljemo izraz

$$P\left[\tau_i > t\right] = \lim_{k \to \infty} \left[1 - \sum_{j \neq i} q_{ij} \Delta t\right]^k = \lim_{k \to \infty} \left[1 - \sum_{j \neq i} q_{ij} \frac{t}{k}\right]^k = e^{-q_i t}. \tag{2.58}$$

Iz njega lahko ugotovimo, da je porazdelitev časov zadrževanja v posameznih stanjih med prehajanji podana z eksponentno porazdelitvijo po izrazu

$$P\left[\tau_i \leq t\right] = 1 - e^{-q_i t}. \tag{2.59}$$

2.8.1 Porazdelitev verjetnosti stanj

Verjetnost zadrževanja sistema v času t v stanju j zapišemo z izrazom

$$\pi_j(t) = P\left[X(t) = j\right]. \tag{2.60}$$

Sprememba verjetnosti v infinitezimalnem času Δt je potemtakem podana z izrazom

$$\pi_j(t + \Delta t) = \sum_{i \neq j} \pi_i(t) q_{ij} \Delta t + \pi_j(t) \left[1 - \sum_{k \neq j} q_{jk} \Delta t\right]. \tag{2.61}$$

Levi del vsote na desni strani izraza (2.61) ponazarja verjetnost stanja i v času t in kasnejši prehod v stanje j v časovnem intervalu Δt, desni del vsote pa verjetnost stanja j v času t, pri čemer do prehajanja v časovnem intervalu Δt ne pride. Ob predpostavki veljavnosti izraza

$$q_{jj} = -\sum_{k \neq j} q_{jk}, \tag{2.62}$$

ki definira vsoto vseh intenzivnosti zapuščanj stanja j, pri deljenju izraza (2.61) z Δt pridemo do izraza

$$\frac{d\pi_j(t)}{dt} = \sum_{i \neq j} \pi_i q_{ij} - \pi_j \sum_{k \neq j} q_{jk}, \tag{2.63}$$

preko njega pa do izrazov

$$\frac{d\pi_j(t)}{dt} = \sum_{i} \pi_i(t) q_{ij}, \tag{2.64}$$

$$\frac{d\widetilde{\pi}(t)}{dt} = \widetilde{\pi}(t)\widetilde{Q}. \tag{2.65}$$

Pri tem veljata izraza

$$\widetilde{\pi}(t) = (\pi_1(t), \pi_2(t), ...), \tag{2.66}$$

$$\frac{d\widetilde{\pi}(t)}{dt} = (\frac{d\pi_1(t)}{dt}, \frac{d\pi_2(t)}{dt}, ...), \tag{2.67}$$

matrika $\widetilde{Q}$ pa predstavlja intenzivnosti prehajanj ali infinitezimalni generator ($\widetilde{Q} = [q_{ij}]$).

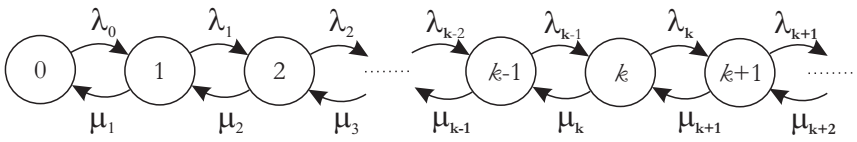
Za nereducibilno in časovno homogeno Markovsko verigo z zveznim časom velja, da zanjo obstaja limitna vrednost, ki je neodvisna od začetnega stanja verige.

2.8.2 Matriki intenzivnosti prehajanj in verjetnosti prehajanj

V razdelku o diskretnih časovnih Markovskih verigah smo spoznali matriko verjetnosti prehajanja stanj M, v pričujočem razdelku o zveznih časovnih Markovskih verigah pa matriko intenzivnosti prehajanj $\widetilde{Q}$. Prvo matriko tvorijo verjetnosti p_{ij}, drugo pa intenzivnosti prehajanj q_{ij}. Da bi tudi v drugi matriki prišli do verjetnosti, moramo posamezne intenzivnosti prehajanj q_{ij} pomnožiti s časovnim intervalom Δt ($\Delta t * q_{ij}, \Delta t \to 0$). Matriki M in $\widetilde{Q}$ enolično določata Markovsko verigo posameznega tipa.

2.9 Rojstno smrtni proces

Rojstno smrtni proces (angl. *birth-dead process*) je posebna oblika zvezne časovne Markovske verige, v kateri so možni prehodi iz trenutnega stanja verige le v njemu sosednja stanja. Termin "*rojstno smrtni proces*"izvira iz opazovanja števila zahtev v sistemu v času t. Predpostavimo, da imamo v času t opravka s številom k zahtev v sistemu. Tovrstno stanje sistema v času t bomo označevali s k. Prehod iz stanja k v stanje $k+1$ predstavlja rojstvo (v sistem je vstopila nova zahteva), prehod v stanje $k-1$ pa smrt zahteve (zahteva je bila uspešno servisirana in je zapustila sistem). Predpostavljamo, da se rojstvo in smrt ne moreta zgoditi hkrati, istočasno pa se ne more zgoditi več smrti ali več rojstev hkrati. Stanja sistema ali procesa označujemo s celimi pozitivnimi števili $\{0, 1, 2, ...\}$, ki označujejo število zahtev v sistemu. Diagram prehajanja stanj v rojstno smrtnem sistemu je predstavljen na sliki 2.13, pri čemer λ_i predstavlja *intenzivnost rojevanja* (prihajanja) novih zahtev v stanju sistema i, μ_i pa *intenzivnost umiranja* (strežbe) zahtev v stanju sistema i. Iz slike 2.13 je razvidno, da rojstva zahtev vodijo iz stanja i v stanje $i+1$ ali v stanje z indeksom, ki je večji za ena od indeksa izhodiščnega stanja, smrti pa iz stanja i v stanje $i-1$ ($i = 0, 1, 2, ...$) ali v stanje z indeksom, ki je manjši za ena od indeksa izhodiščnega stanja, če izhodiščno stanje le ni stanje z indeksom 0.



Slika 2.13: Diagram prehajanja stanj rojstno smrtnega procesa.

Predpostavimo, da imamo v sistemu (procesu) k zahtev, kar pomeni, da se nahajamo v stanju z indeksom k. Velja izraz

$$\lambda_k = q_{k,k+1},\ \mu_k = q_{k,k-1}, \tag{2.68}$$

pri čemer smo si pomena intenzivnosti prehajanj $q_{k,k+1}$ in $q_{k,k-1}$ ogledali že v prejšnjem razdelku. Na osnovi izraza (2.68) lahko matriko intenzivnosti prehajajanj za rojstno smrtni proces zapišemo z izrazom

$$Q = \begin{bmatrix} -\lambda_0 & \lambda_0 & 0 & 0 & ... \\ \mu_1 & -(\lambda_1+\mu_1) & \lambda_1 & 0 & ... \\ 0 & \mu_2 & -(\lambda_2+\mu_2) & \lambda_2 & ... \\ 0 & 0 & \mu_3 & -(\lambda_3+\mu_3) & ... \\ ... & ... & ... & ... & ... \end{bmatrix}. \tag{2.69}$$

Tako veljata relaciji

$$\frac{dP_k(t)}{dt} = -\left(\lambda_k+\mu_k\right)P_k(t) + \lambda_{k-1}P_{k-1}(t) + \mu_{k+1}P_{k+1}(t),\ k \geq 1, \tag{2.70}$$

$$\frac{dP_0(t)}{dt} = -\lambda_0 P_0(t) + \mu_1 P_1(t),\ k = 0. \tag{2.71}$$

Poseben primer rojstno smrtnega procesa je *čisti rojstni proces*, kjer ni strežbe in zaradi tega posledično ne prihaja do umiranja zahtev. Zanj velja izraz

$$\forall k : \lambda_k = \lambda > 0,\ \mu_k = 0, \tag{2.72}$$

tako da se izraza (2.70) in (2.71) poenostavita v izraza

$$\frac{dP_k(t)}{dt} = -\lambda P_k(t) + \lambda P_{k-1}(t),\ k \geq 1, \tag{2.73}$$

$$\frac{dP_0(t)}{dt} = -\lambda P_0(t), k = 0. \tag{2.74}$$

2.10 Vzorčne strežne enote

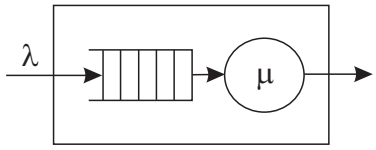
V pričujočem razdelku si bomo ogledali nekaj primerov vzorčnih strežnih enot s sledečimi značilnostmi:

- proces v strežni enoti je stohastične narave;
- prihajalni proces v strežni enoti je Poissonov proces;
- strežni časi v strežni enoti so porazdeljeni eksponentno;
- proces v posamezni strežni enoti je Markovskega tipa ali brez zmožnosti pomnjenja;
- prihajanje in strežbo zahtev ponazorimo z rojstno smrtnim procesom;

Za posamezne vzorčne strežne enote bomo podali tudi načine izračunov osnovnih zmogljivostnih metrik. Njihove izpeljave so natančneje opisane v viru [1].

2.10.1 Strežni enota $M/M/1$

Najosnovnejši model sistema iz realnega okolja je strežna enota tipa $M/M/1$. Temelji na Poissonovem prihajalnem procesu, eksponentni porazdelitvi strežnih časov in enem strežniku. Manjkajoči parametri iz Kendallove notacije so povzeti po privzetih vrednostih in sicer predpostavljamo, da sta čakalna vrsta in populacija zahtev neskončni, strežna disciplina pa temelji na principu FIFO ($k = \infty, P = \infty, Q = FIFO$). Strežna enota $M/M/1$ je grafično predstavljena na sliki 2.14.



Slika 2.14: Grafična ponazoritev $M/M/1$ strežne enote.

Za $M/M/1$ strežno enoto običajno predpostavljamo, da sta λ in μ neodvisni od števila zahtev v njej in se skozi čas tako ne spreminjata. Slednje pomeni, da sta neodvisni od stanja, v katerem se strežna enota nahaja, kar ponazorimo z izrazom

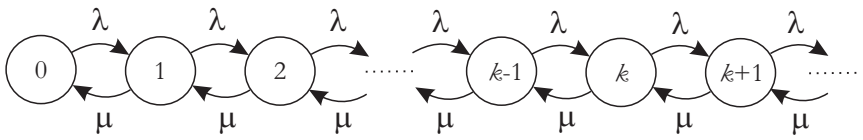
$$\lambda_k = \lambda, \ \mu_k = \mu, \tag{2.75}$$

diagram prehajanja stanj v takšni strežni enoti pa je predstavljen na sliki 2.15. Iz diagrama prehajanja stanj je razvidno, da diagram na desni strani ni omejen, kar pomeni da število zahtev v strežni enoti ali njen indeks stanja lahko narašča preko vseh meja.

Ob predpostavki, da v strežni enoti obstaja stabilno ali ravnovesno stanje (angl. *steady state*) [1], veljata izraza

$$\frac{dP_k(t)}{dt} = 0, \tag{2.76}$$

$$P_k = \lim_{t \to \infty} P_k(t), \tag{2.77}$$



Slika 2.15: Diagram prehajanja stanj $M/M/1$ strežne enote.

na osnovi katerih lahko izraza (2.70) in (2.71) poenostavimo v izraza

$$(\lambda + \mu)P_k = \lambda P_{k-1} + \mu P_{k+1}, k \geq 1, \tag{2.78}$$

$$\mu P_1 = \lambda P_0, \ k = 0. \tag{2.79}$$

Ob vseh podanih predpostavkah lahko izpeljemo sledeče kvantitativne zmogljivostne metrike strežne enote tipa $M/M/1$:

- z uporabo z-transformacije[12] nad izrazoma (2.78) in (2.79) pridemo do izrazov

$$P_0 = 1 - \frac{\lambda}{\mu} = 1 - \rho, \tag{2.80}$$

$$P_k = (1 - \rho)\rho^k = P_0\rho^k; \tag{2.81}$$

- verjetnost, da imamo v strežni enoti n ali več zahtev izpeljemo z izrazom

$$P[i \geq n] = \sum_{k=n}^{\infty} P_k = (1 - \rho) \sum_{k=n}^{\infty} \rho^k =$$

$$= (1 - \rho) \left[\sum_{k=0}^{\infty} \rho^k - \sum_{k=0}^{n-1} \rho^k \right] = (1 - \rho) \left[\frac{1}{1 - \rho} - \frac{1 - \rho^n}{1 - \rho} \right] = \rho^n; \tag{2.82}$$

- povprečno število zahtev v strežni enoti N je pogojeno z verjetnostmi nahajanj v posameznih stanjih po izrazu

$$N = \sum_{k=0}^{\infty} kP_k = \sum_{k=0}^{\infty} k(1 - \rho)\rho^k =$$

$$= (1 - \rho)\rho \sum_{k=0}^{\infty} k\rho^{k-1} = \frac{\rho}{1 - \rho} = \frac{\lambda}{\mu - \lambda}; \tag{2.83}$$

- povprečni čas prebivanja zahteve v strežni enoti T se izračuna po izrazu

$$T = \frac{N}{\lambda} = \frac{\rho}{\lambda(1 - \rho)} = \frac{1}{\mu - \lambda}; \tag{2.84}$$

[12] z-transformacija je ekvivalent Laplaceove transformacije v diskretnem prostoru stanj [7].

- povprečno število zahtev v strežniku N_s, kar odraža zasedenost strežnika (faktor uporabnosti), je definirano z izrazom

$$N_s = \frac{\lambda}{\mu} = \rho = 1 - P_0; \tag{2.85}$$

- povprečni čas zadrževanja zahteve v čakalni vrsti (W - čakalni čas) je podan kot razlika med časom zadrževanja zahteve v sistemu in časom strežbe po izrazu

$$W = T - \frac{1}{\mu} = \frac{\rho}{\mu - \lambda}; \tag{2.86}$$

- povprečno število zahtev v čakalni vrsti N_q se izračuna po izrazu

$$N_q = \lambda W = \frac{\rho^2}{1 - \rho}; \tag{2.87}$$

Poseben primer $M/M/1$ strežne enote je *omahljiv strežni sistem* (angl. *discouraged arrival process*), pri katerem se intenzivnost prihajanja zahtev spreminja v odvisnosti od števila zahtev v strežni enoti, intenzivnost procesiranja pa se ne spreminja. V splošnem omahljivi strežni sistem opišemo z izrazom

$$\lambda_k = \lambda/(k+1),\ \mu_k = \mu,\ k = 0, 1, ... \tag{2.88}$$

Zgled 4 *V usmerjevalnik prihajajo paketi po Poissonovem procesu, pri čemer je njihova intenzivnost prihajanja 100.000 paketov/sekundo, čas usmerjanja posameznega paketa pa je porazdeljen po eksponentni porazdelitvi in v povprečju traja* $5 * 10^{-6}$ *sekunde na posamezno zahtevo. Izračunaj osnovne zmogljivostne metrike* N, N_s, N_q, W *in* T *za podani strežni sistem, pri čemer predpostavljamo, da je kapaciteta strežnega sistema neomejena.*

Rešitev: *Glede na opis procesa prihajanja in usmerjanja (strežbe) za model sistema povzamemo strežno enoto tipa* $M/M/1$. *Na začetku deklariramo osnovne zmogljivostne metrike po izrazih*

$$\lambda = 10^5\ paketov/sekundo,\ x = 5 * 10^{-6}\ sekunde/zahtevo, \tag{2.89}$$

$$\mu = \frac{1}{x} = 2 * 10^5\ paketov/sekundo, \rho = \frac{\lambda}{\mu} = \frac{1}{2}. \tag{2.90}$$

Iskane metrike izračunamo po vrsti po izrazih

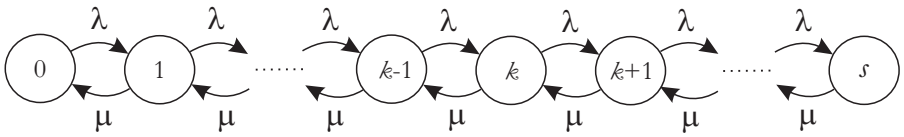
$$N = \frac{\rho}{1 - \rho} = 1\ zahteva,\ N_s = \rho = \frac{1}{2}\ zahteve,\ N_q = \frac{\rho^2}{1 - \rho} = \frac{1}{2}\ zahteve, \tag{2.91}$$

$$W = \frac{\rho}{\mu - \lambda} = 5 * 10^{-6}\ sekunde,\ T = \frac{N}{\lambda} = 1 * 10^{-5}\ sekunde. \tag{2.92}$$

2.10.2 Strežna enota $M/M/1/s$

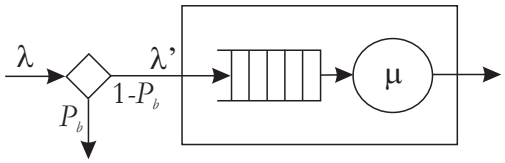
Za modele strežnih sistemov iz realnega okolja, kjer ne moremo zagotoviti realizacije njihove neskončne kapacitete, je v primerjavi z $M/M/1$ primernejša izbira modela $M/M/1/s$ strežne enote, ki ima glede na Kendallovo notacijo končno kapaciteto ($k = s$) in ob enem strežniku čakalno vrsto končne dolžine $s-1$ mest. V tovrstni strežni enoti se tako lahko nahaja največ s zahtev, pri čemer je lahko največ $s-1$ zahtev v čakalni vrsti in največ ena v procesu strežbe (v strežniku). Značilnosti prihajalnega in strežnega procesa ostajata nespremenjeni - sta enaki kot v strežni enoti $M/M/1$. V primeru nahajanja s zahtev v strežni enoti se novoprispela zahteva ali ena od zahtev v čakalni vrsti zavrže in v tem primeru rečemo, da je strežna enota *zasičena*. Tako pridemo do sklepa, da model $M/M/1/s$ strežne enote predstavlja **izguben sistem**. Tudi za $M/M/1/s$ strežno enoto predpostavljamo, da sta λ in μ neodvisni od števila zahtev v njej in se skozi čas tako ne spreminjata.

Na sliki 2.16 je predstavljen diagram prehajanja stanj strežne enote $M/M/1/s$. Iz njega je razvidno, da je na desni strani omejen s stanjem, čigar indeks predstavlja kapaciteto sistem s.



Slika 2.16: Diagram prehajanja stanj $M/M/1/s$ strežne enote.

Na sliki 2.17 je predstavljena grafična ponazoritev strežne enote $M/M/1/s$, ki se od slike 2.14 razlikuje le v vejanju na vhodni strani modela. Slednje ponazarja verjetnostno pogojeno izbiro poti, ki vodi bodisi v strežno enoto (v strežni enoti je manj kot s zahtev), ali v izgubo zahteve (v strežni enoti se predhodno že nahaja s zahtev). Izbira prve poti je pogojena z verjetnostjo $1 - P_b$, izbira druge pa z verjetnostjo P_b, pri čemer P_b predstavlja verjetnost izgube zahteve. Verjetnost P_b enačimo z verjetnostjo P_s, pri čemer P_s predstavlja verjetnost nahajanja s zahtev v sistemu ali verjetnost popolne zasedenosti strežne enote.



Slika 2.17: Grafična ponazoritev $M/M/1/s$ strežne enote.

Na sliki 2.17 nastopata tudi intenzivnosti porajanja zahtev λ in λ'. Pri tem λ predstavlja intenzivnost porajanja zahtev iz okolja, λ' pa intenzivnost

vstopajočih zahtev. Slednja je zmanjšana za število zahtev, ki jih strežna enota zaradi zasedenosti zavrne (izgubi). Tako lahko zapišemo izraz

$$\lambda' = \lambda * (1 - P_b). \tag{2.93}$$

Iz veljavnosti zakona o ohranitvi pretoka lahko sklepamo, da mora biti izhodna intenzivnost γ iz strežne enote enaka vhodni intenzivnosti λ', kar zapišemo z izrazom

$$\gamma = \lambda(1 - P_b). \tag{2.94}$$

Po drugi plati lahko izhodno intenzivnost za opisano strežno enoto izračunamo tudi po izrazu

$$\gamma = \sum_{k=1}^{s} \mu P_k = \mu(P_1 + P_2 + ... + P_s) = \mu(1 - P_0). \tag{2.95}$$

Če izenačimo oba predhodna izraza tako dobimo nov izraz

$$\lambda(1 - P_b) = \mu(1 - P_0). \tag{2.96}$$

Osnovne kvantitativne zmogljivostne metrike strežne enote $M/M/1/s$ so sledeče:

- verjetnosti nahajanja strežne enote v stanjih 0 in k se izračunata po izrazih

$$P_0 = \frac{1 - \rho}{1 - \rho^{s+1}}, \tag{2.97}$$

$$P_k = \frac{(1 - \rho)\rho^k}{1 - \rho^{s+1}}; \tag{2.98}$$

- verjetnost zasičenosti strežne enote je pogojena z izrazom

$$P_s = P_b = \begin{cases} \frac{(1-\rho)\rho^s}{1-\rho^{s+1}}, & \rho < 1, \\ \frac{1}{s+1}, & \rho = 1; \end{cases} \tag{2.99}$$

- povprečno število zahtev v strežni enoti N je pogojeno z verjetnostmi nahajanj v posameznih stanjih strežne enote in se izračuna po izrazu

$$N = \begin{cases} \sum_{k=0}^{s} kP_k = \frac{\rho}{1-\rho} - \frac{\rho}{1-\rho}(s+1)P_b, & \rho < 1, \\ \frac{s}{2}, & \rho = 1; \end{cases} \tag{2.100}$$

- povprečno število zahtev v strežniku se izračuna po izrazu

$$N_s = \rho(1 - P_b); \tag{2.101}$$

- povprečno število zahtev v čakalni vrsti se izračuna po izrazu

$$N_q = N - N_s = \frac{\rho^2}{1 - \rho} - \rho\frac{\rho + s}{1 - \rho}P_b; \tag{2.102}$$

- upoštevajoč izraz (2.93) se povprečni čas bivanja zahteve v strežni enoti T in povprečni čas čakanja v njeni čakalni vrsti W izračunata po izrazih

$$T = \frac{N}{\lambda'} = \frac{1}{\mu - \lambda} - \frac{s\rho^{s+1}}{\lambda - \mu\rho^{s+1}}, \tag{2.103}$$

$$W = \frac{N_q}{\lambda'} = \frac{\rho}{\mu - \lambda} - \frac{s\rho^{s+1}}{\lambda - \mu\rho^{s+1}}. \tag{2.104}$$

Natančnejše izpeljave matematičnih izrazov za strežno enoto $M/M/1/s$ si bralec lahko ogleda v delu [1].

Zgled 5 *Predpostavimo, da imamo opravka z enakim strežnim procesom, kot v zgledu 4, pri čemer ima usmerjevalnik omejeno čakalno vrsto, v katero je mogoče uvrstiti 9 čakajočih paketov. Za omenjeni strežni sistem izračunaj verjetnost izgubljanja paketov P_b in njegovo zmogljivostno metriko T.*

Rešitev: *Glede na dodatno omejitev omejenosti čakalne vrste za model sistema povzamemo strežno enoto tipa $M/M/1/s$, pri čemer je kapaciteta sistema $s = 10$. Na začetku deklariramo osnovne zmogljivostne metrike po izrazih*

$$\lambda = 10^5 \ paketov/sekundo, \ \mu = 2 * 10^5 \ paketov/sekundo, \rho = \frac{1}{2}, \tag{2.105}$$

v nadaljevanju pa izračunamo odgovore na vprašanja po izrazih

$$P_b = P_s = \frac{(1-\rho) * \rho^s}{1 - \rho^{s+1}} = \frac{(\frac{1}{2})^{11}}{1 - (\frac{1}{2})^{11}} = 0,00049, \tag{2.106}$$

$$T = \frac{1}{\mu - \lambda} - \frac{s\rho^{s+1}}{\lambda - \mu\rho^{s+1}} =$$

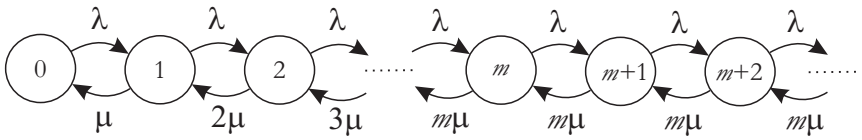
$$= \frac{1}{2 * 10^5 - 10^5} - \frac{10 * (\frac{1}{2})^{11}}{10^5 - 2 * 10^5 * (\frac{1}{2})^{11}} = 9,9 * 10^{-6} \ sekunde. \tag{2.107}$$

Primerjajoč rezultate iz zgleda 4 in pričujoče rezultate je razvidno, da se v primeru omejene kapacitete strežne enote čas zadrževanja paketa v usmerjevalniku malenkostno zmanjša na račun malenkonstno manjšega čakalnega časa zahtev, ki izvira iz 0,049% izgubljenih paketov.

2.10.3 Strežna enota $M/M/m$

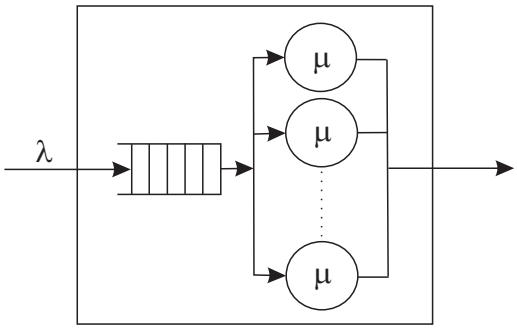
Strežna enota $M/M/m$ se od strežne enote $M/M/1$ razlikuje po tem, da vsebuje m paralelno vezanih funkcionalno ekvivalentnih strežnikov ($m > 1$). Vsaka zahteva je v takem sistemu obdelana natanko enkrat na poljubnem od m razpoložljivih strežnikov. Značilnosti prihajalnega in strežnega procesa sta enaki kot pri $M/M/1$ strežni enoti, imamo pa zopet opravka z neskončno vrsto in s tem posledično z neskončno kapaciteto strežne enote.

Na sliki 2.18 je predstavljen diagram prehajanja stanj strežne enote $M/M/m$. Zaradi neskončne kapacitete strežne enote diagram na desni strani ni omejen, ključne pa so intenzivnosti strežbe, ki od leve proti desni v stanjih naraščajo vse do $m * \mu$ v stanju z indeksom m, odtod naprej pa ostajajo nespremenjene.



Slika 2.18: Diagram prehajanja stanj $M/M/m$ strežne enote.

Na sliki 2.19 je predstavljena grafična ponazoritev strežne enote $M/M/m$, ki se od slike 2.14 razlikuje le po številu strežnikov.



Slika 2.19: Grafična ponazoritev $M/M/m$ strežne enote.

Karakteristično za strežno enoto $M/M/m$ je to, da ji s povečevanjem števila strežnikov m lahko poljubno povečujemo intenzivnost strežbe, s čimer imamo na nivoju modeliranja omogočeno *skalabilnost*. Če je μ intenzivnost strežbe posameznega strežnika, imamo tako pri vsaj m zahtevah v strežni enoti opravka z strežno intenzivnostjo $m * \mu$, v primeru pa da je teh zahtev manj ($k < m$), pa z intenzivnostjo strežbe $k * \mu$.

Upoštevajoč sliko 2.18 lahko nastavimo sistem ravnotežnih enačb v izrazih

$$k < m : k\mu P_k = \lambda P_{k-1}, \tag{2.108}$$

$$k \geq m : m\mu P_k = \lambda P_{k-1}. \tag{2.109}$$

Ob vpeljavi nove spremenljivke a po izrazu

$$a = \frac{\lambda}{\mu}, \tag{2.110}$$

ki nam omogoči preglednejši matematični zapis in ravnotežnih enačb po izrazih

$$k \leq m: \ P_k = \frac{a^k}{k!} P_0, \tag{2.111}$$

$$k \geq m: P_k = \frac{a^k}{m! m^{(k-m)}} P_0, \tag{2.112}$$

za strežno enoto $M/M/m$ veljata izraza

$$\rho = \frac{\lambda}{m\mu}, \tag{2.113}$$

$$P_0 = \left[\sum_{k=0}^{m-1} \frac{a^k}{k!} + \frac{a^m}{m!(1-\rho)} \right]^{-1}, \tag{2.114}$$

kjer ρ predstavlja uporabnostni faktor strežne enote, P_0 verjetnost prazne strežne enote, P_k pa verjetnost nahajanja k zahtev v strežni enoti tipa $M/M/m$.

Osnovne kvantitativne zmogljivostne metrike strežne enote $M/M/m$ so sledeče:

- verjetnost, da novoprispela zahteva naleti na vse zasedene strežnike in bo morala čakati v vrsti, se izračuna po izrazu

$$P_d = \frac{P_0 a^m}{m!(1-\rho)}; \tag{2.115}$$

- povprečno število zahtev v čakalni vrsti se izračuna po izrazu

$$N_q = \frac{\rho}{1-\rho} P_d; \tag{2.116}$$

- povprečen čas čakanja zahteve v čakalni vrsti se izračuna po izrazu

$$W = \frac{N_q}{\lambda}; \tag{2.117}$$

- povprečen čas prebivanja zahteve v strežni enoti se izračuna po izrazu

$$T = W + \frac{1}{\mu}; \tag{2.118}$$

- povprečno število zahtev v strežni enoti se izračuna po izrazu

$$N = \lambda T; \tag{2.119}$$

Natančnejše izpeljave matematičnih izrazov za strežno enoto $M/M/m$ si bralec lahko ogleda v delu [1].

Zgled 6 *Opravka imamo s skalabilnim usmerjevalnikom, v katerega lahko vstavimo največ 10 vezij, ki vsako posebej izvaja usmerjanje paketov v usmerjevalniku. Intenzivnosti strežbe posameznega vezja in prihajanja paketov sta enaki in sicer* 10.000 *paketov na sekundo. Prihajanje paketov je pogojeno s Poissonovim procesom, strežni časi pa so eksponentno porazdeljeni. Predpostavimo, da kapaciteta sistema ni omejena. Kolikšna je razlika v času čakanja posameznega paketa med konfiguracijama z dvemi ali tremi vezji v usmerjevalniku?*

Rešitev: *Glede na opis skalabilnega usmerjevalnika za njegov model povzamemo strežno enoto $M/M/m$, pri čemer za prvo konfiguracijo velja $m = 2$, za drugo pa $m = 3$. Nastavimo najprej enačbe za osnovne zmogljivostne metrike, ki so podane v izrazu*

$$\lambda = 10^4 \ paketov/sekundo, \ \mu = 10^4 \ paketov/sekundo, \ \rho = \frac{1}{m}, \ a = 1. \tag{2.120}$$

Odtod izračunamo čakalna časa za obe konfiguraciji po izrazih

- *primer konfiguracije z dvema vezjema ($m = 2$, $\rho = \frac{1}{2}$):*

$$P_0 = \left[\sum_{k=0}^{m-1} \frac{a^k}{k!} + \frac{a^m}{m!(1-\rho)} \right]^{-1} = \frac{1}{3}; \tag{2.121}$$

$$P_d = \frac{P_0 a^m}{m!(1-\rho)} = \frac{1}{3}, \ N_q = \frac{\rho}{1-\rho} P_d = \frac{1}{3}; \tag{2.122}$$

$$W = \frac{N_q}{\lambda} = \frac{\frac{1}{3}}{10.000} = \frac{1}{3} * 10^{-4} \ sekunde/zahtevo; \tag{2.123}$$

- *primer konfiguracije s tremi vezji ($m = 3$, $\rho = \frac{1}{3}$):*

$$P_0 = \left[\sum_{k=0}^{m-1} \frac{a^k}{k!} + \frac{a^m}{m!(1-\rho)} \right]^{-1} = \frac{4}{11}; \tag{2.124}$$

$$P_d = \frac{P_0 a^m}{m!(1-\rho)} = \frac{1}{11}, \ N_q = \frac{\rho}{1-\rho} P_d = \frac{1}{22}; \tag{2.125}$$

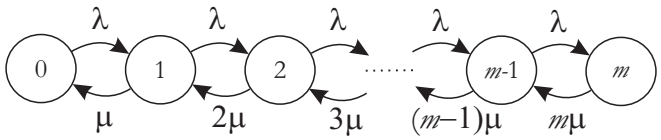
$$W = \frac{N_q}{\lambda} = \frac{\frac{1}{22}}{10.000} = \frac{1}{22} * 10^{-4} = 0,29 * 10^{-4} \ sekunde/zahtevo; \tag{2.126}$$

Iz izračunov je razvidno, da pade čas čakanja ob prehodu iz konfiguracije z dvema vezjema na konfiguracijo s tremi vezji za $0,04 * 10^{-4}$ sekunde/zahtevo ali za 4 μs.

2.10.4 Strežna enota $M/M/m/m$

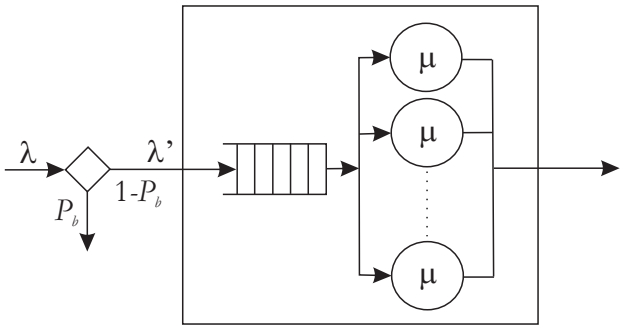
Za strežno enoto $M/M/m/m$ je karakteristično, da je kapaciteta sistema enaka številu strežnikov v sistemu ($k = m$), kar pomeni, da v strežni enoti nimamo čakalne vrste ali povedano drugače imamo opravka s čakalno vrsto, katere dolžina je 0. Slednje nas vodi do zaključka, da imamo zopet opravka z modelom strežne enote, ki je izguben.

Diagram prehajanja stanj strežne enote $M/M/m/m$ je predstavljen na sliki 2.20. Na desni strani je omejen s stanjem m, zanj pa so karakteristične tudi naraščajoče intenzivnosti strežbe od leve proti desni.



Slika 2.20: Diagram prehajanja stanja $M/M/m/m$ strežne enote.

Na sliki 2.21 je predstavljena grafična ponazoritev strežne enote $M/M/m/m$ kot kombinacija gradnikov iz predhodno predstavljenih $M/M/m$ in $M/M/1/s$ strežnih enot. Zaradi splošnosti ponazoritve strežne enote v predstavitvi ohranjamo grafični primitiv čakalne vrste, pri čemer se zavedamo, da je njena dolžina enaka 0.



Slika 2.21: Grafična ponazoritev $M/M/m/m$ strežne enote.

Iz veljavnih ravnotežnih izrazov

$$\lambda P_{k-1} = k\mu P_k, \tag{2.127}$$

$$P_k = P_0 \frac{a^k}{k!}, a = \frac{\lambda}{\mu}, \tag{2.128}$$

lahko izpeljemo enačbi v izrazih

$$P_0 = \left[\sum_{k=0}^{m} \frac{a^k}{k!}\right]^{-1}, \tag{2.129}$$

$$P_k = \frac{a^k}{k! \left[\sum_{k=0}^{m} \frac{a^k}{k!}\right]}, \tag{2.130}$$

na osnovi slednjih pa tudi *izgubno enačbo*

$$P_b = \frac{a^m}{m! \left[\sum_{k=0}^{m} \frac{a^k}{k!}\right]}, \tag{2.131}$$

pri čemer P_b predstavlja verjetnost izgubljanja zahtev. Veljata tudi izraza

$$N = \sum_{k=0}^{m} kP_k = ... = a(1 - P_b), \tag{2.132}$$

$$T = \frac{1}{\mu}, N_q = 0, W = 0. \tag{2.133}$$

Natančnejše izpeljave matematičnih izrazov za strežno enoto $M/M/m/m$ si bralec lahko ogleda v delu [1].

Zgled 7 *Mobilni operater postavlja novo bazno postajo za potrebe mobilne telefonije, pri čemer v njeni konfiguraciji lahko nastavi število aktivnih sej (mobilnih telefonskih pogovorov) m, ki jih je bazna postaja zmožna zagotavljati med področjem, ki ga pokriva (celico) in zunanjim svetom. Na področju posamezne celice se zadržuje 300 uporabnikov (naročnikov), vsak od njih pa opravi v povprečju za 20 minut klicev dnevno. Operater bi rad zagotovil, da bi bila verjetnost izgubljanja klicev manjša ali enaka 2% ($P_b \leq 0,02$). Kolikšna naj bo nastavitev parametra m, če je maksimalna obremenitev v posamezni uri dneva 14% dnevnega prometa (angl. peak hour) in pribitek na nezanesljivost linij 10%? Pri tem predpostavljamo, da klici ne morejo čakati v čakalni vrsti. Primer je povzet po viru [1].*

Rešitev: *Ker prihajajoči klici ne morejo čakati v čakalni vrsti, lahko sklepamo, da imamo opravka z izgubnim sistemom brez čakalne vrste, pri čemer iskani parameter m predstavlja v domeni strežnih enot število strežnikov. Ker je populacija naročnikov relativno velika, povzamemo privzeto vrednost populacije za neskončno ($k = \infty$), tako da bazno postajo z vidika njenih strežnih značilnosti smatramo za strežno enoto tipa $M/M/m/m$.*

Najprej izračunamo maksimalno breme strežne enote v posamezni uri (angl. total traffic load - TTL) po izrazih

$$TTL = \frac{n_{customers} * time_{per-day} * traffic_{max}}{60\ min}, \tag{2.134}$$

$$TTL = \frac{300 * 20 * 0,14}{60} = 14 \ Erlangov, \tag{2.135}$$

*v nadaljevanju pa TTL utežimo še s pribitkom na nezanesljivost (14*1,1=15,4 Erlanga), s čimer pridemo do končne vrednosti TTL. Slednjo namesto spremenljivke a vstavimo v izraz*

$$P_b = \frac{a^m}{m! \left[\sum_{k=0}^{m} \frac{a^k}{k!}\right]} = \frac{(15,4)^m}{m! \left[\sum_{k=0}^{m} \frac{(15,4)^k}{k!}\right]}. \tag{2.136}$$

Z vstavljanjem različnih vrednosti spremenljivke m se izkaže, da verjetnost P_b pade pod željeno vrednost 2% pri nastavitvi 23 aktivnih sej ($m = 23 \rightarrow P_b = 0,0167, m = 22 \rightarrow P_b = 0,0254$). S tem smo zagotovili ustrezno prepustnost pogovorov v obdobju največjega števila porajanja klicev, v preostalih delih dneva pa je ta še večja.

Zgled 8 *Satelitski sistem mobilne telefonije ponuja na posameznem prenosnem kanalu pasovno širino 1200 bps (angl. bits per second) za izvajanje klicnih storitev in 2400 bps za podporo podatkovnih storitev (prenosov sporočil). Pri tem je sistem zasnovan skalabilno z m_v klicnimi prenosnimi kanali in m_d podatkovnimi prenosnimi kanali. Klici in podatkovna sporočila se porajajo po Poissonovem procesu, prvi z intenzivnostjo 200 klicev na sekundo in drugi z intenzivnostjo 40 prenosov sporočil na sekundo. Oboji so porazdeljeni eksponentno in sicer prvi z dolžino 54 bitov/sekundo in drugi z dolžino 240 bitov/sekundo. Ob zasedenosti prenosnih kapacitet se klici zavračajo (izgubljajo), podatkovna sporočila pa odhajajo v vmesnik hipotetične neskončne dolžine. Izračunaj število potrebnih klicnih kanalov (m_v), da bo verjetnost izgube klica manjša od 2% in število potrebnih podatkovnih kanalov (m_d), da bo zamik (angl. message delay) dostave manjši od 0,115 sekunde. Primer je povzet po viru [1].*

Rešitev: *Klicni del strežnega sistema lahko zopet obravnavamo kot $M/M/m/m$ strežno enoto (m predstavlja število klicnih kanalov m_v), podatkovni del strežnega sistema pa kot $M/M/m$ strežno enoto (m predstavlja število podatkovnih kanalov m_d).*

- *klicni del: iščemo ustrezno število klicnih kanalov (m_v), da zadostimo podanim pogojem; intenzivnost prihajanja, inverzna vrednost intenzivnosti strežbe in spremenljivka a imajo po vrsti vrednosti*

  $$\lambda_v = 200 \ zahtev/sek, \frac{1}{\mu_v} = x = 54/1200 = 9/200 \ sek/zahtevo, a = \frac{\lambda_v}{\mu_v} = 9; \tag{2.137}$$

  z uporabo izraza

  $$P_b = \frac{a^{m_v}}{m_v! \left[\sum_{k=0}^{m_v} \frac{a^k}{k!}\right]} = \frac{(9)^{m_v}}{m_v! \left[\sum_{k=0}^{m_v} \frac{9^k}{k!}\right]} \leq 0,02 \rightarrow m_v \geq 15, \tag{2.138}$$

  in vstavljanjem vrednosti za spremenljivko m_v ugotovimo, da podanemu

pogoju zadostimo pri vsaj 15 klicnih kanalih;

- *podatkovni del: iščemo ustrezno število podatkovnih kanalov (m_d), da zadostimo podanim pogojem; intenzivnost prihajanja, inverzna vrednost intenzivnosti strežbe in spremenljivka a imajo po vrsti vrednosti*

$$\lambda_d = 40 \ zahtev/sek, \frac{1}{\mu_d} = x = \frac{240}{2400} = 0,1 \ sek/zahtevo, a = \frac{\lambda_d}{\mu_d} = 4; \tag{2.139}$$

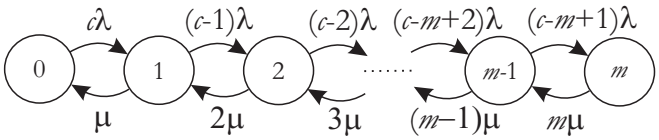
z uporabo izraza

$$T = \frac{1}{\mu_d} + \frac{P_d}{m_d\mu_d - \lambda_d} = 0,1 + \frac{P_d}{10m_d - 40} \leq 0,115 \ sek \rightarrow m_d \geq 6 \tag{2.140}$$

in vstavljanjem vrednosti za spremenljivko m_d ugotovimo, da podanemu pogoju zadostimo pri vsaj 6 podatkovnih kanalih;

2.10.5 Strežna enota $M/M/m/m/c$

Strežno enoto $M/M/m/m/c$ imenujemo tudi za *Engsetov izgubni sistem*. Od predhodno obravnavanega modela strežne enote $M/M/m/m$ se razlikuje zgolj v velikosti populacije zahtev, ki ni več neskončna, temveč omejena ter dovolj majhna, da na vhodnem delu strežne enote intenzivnost prihajanja glede na število zahtev v sistemu niha ($P = c, c < \infty$). Diagram prehajanja stanj strežne enote $M/M/m/m/c$ je predstavljen na sliki 2.22.



Slika 2.22: Diagram prehajanja stanj $M/M/m/m/c$ strežne enote.

Predpostavimo, da je velikost populacije c zahtev. Če je le ta manjša od m, imamo opravka z obvladljivo strežno enoto, v kateri imamo ves čas vsaj en prost strežnik, kar pomeni, da je sistem z vidika zmogljivosti predimenzioniran. Za strežno enoto $M/M/m/m/c$ lahko na osnovi nastavkov podanih z izrazoma

$$0 \leq k \leq c-1 : \lambda_k = \lambda(c-k), \tag{2.141}$$

$$0 \leq k \leq m : \mu_k = k\mu, \tag{2.142}$$

izpeljemo verjetnosti stanj

$$P_k = \binom{c}{k} a^k P_0, \tag{2.143}$$

$$P_0 = \frac{1}{\sum_{k=0}^{m} \binom{c}{k} a^k}, \tag{2.144}$$

$$P_b = P_m = \frac{\binom{c}{m} a^m}{\sum_{k=0}^{m} \binom{c}{k} a^k} \tag{2.145}$$

ter povprečno intenzivnost porajanja zahtev na vhodni strani

$$\overline{\lambda} = \sum_{k=0}^{m-1} \lambda(c-k) \left[P_0 \binom{c}{k} a^k \right] = ... = \lambda c(1-P_b) - \lambda(N - mP_b). \tag{2.146}$$

Povprečno število zahtev v sistemu se izračuna po izrazu

$$N = \sum_{k=0}^{m} kP_k = \overline{\lambda} T = \frac{\overline{\lambda}}{\mu}. \tag{2.147}$$

Natančnejše izpeljave matematičnih izrazov za strežno enoto $M/M/m/m/c$ si bralec lahko ogleda v delu [1].

Zgled 9 *Opravka imamo s strežniškim sistemom, na katerega je vezanih 10 odjemalcev njegovih storitev. Vsak od odjemalcev občasno od strežniškega sistema zahteva vzpostavitev aktivne seje za izvedbo strežbe. Strežniški sistem lahko nudi največ 3 aktivne seje istočasno. V primeru, da pride do porajanja zahteve za aktivno sejo, pri čemer so 3 seje že v teku, se takšno novo zahtevo za sejo zavrne in je zahteva tako posledično izgubljena. Posamezen odjemalec poraja 6 zahtev na uro, vsaka od zahtev pa se streže v povprečju 5 minut. Porajanje zahtev poteka po Poissonovem procesu, čas strežbe pa je porazdeljen eksponentno. Kolikšna je verjetnost izgubljanja zahtev in kolikšno je povprečno število vzpostavljenih aktivnih sej? Primer je povzet po viru [1].*

Rešitev: *Opisani strežniški sistem je karakterističen za realne sisteme, na katere so vezani odjemalci kot so npr. bančni bankomati, POS (angl. point of sale) prodajna mesta, terminali na bančnih okencih itd. Glede na povedano lahko predpostavimo, da je sistem tipa $M/M/m/m/c$, pri čemer je populacija zahtev omejena ($P = c$) s številom naprav, preko katerih dostopajo odjemalci (npr. komitenti banke preko omejenega števila bankomatov). Tako sklepamo, da je $m = 3$ in $c = 10$.*

Intenzivnost prihajanja, inverzna vrednost intenzivnosti strežbe in spremenljivka a imajo po vrsti vrednosti

$$\lambda = 6/60 = 0,1 \text{ zahteve/minuto}, \mu^{-1} = 5 \text{ minut/zahtevo}, a = \frac{\lambda}{\mu} = 0,5. \tag{2.148}$$

Na osnovi izraza

$$P_b = P_m = \frac{\binom{c}{m} a^m}{\sum_{k=0}^{m} \binom{c}{k} a^k} = \frac{\binom{10}{3} (0,5)^3}{\sum_{k=0}^{3} \binom{10}{k} (0,5)^k} = 0,4651 \tag{2.149}$$

izračunamo verjetnost izgubljanja zahtev P_b, ki je 0,4651, na osnovi izrazov

$$P_0 = \frac{1}{\sum_{k=0}^{m} \binom{c}{k} a^k} = \frac{1}{\sum_{k=0}^{3} \binom{10}{k} \frac{1}{2}^k} = 0,03100775, \tag{2.150}$$

$$P_k = \binom{c}{k} a^k P_0, \;\; P_1 = \binom{10}{1} \frac{1}{2}^1 P_0 = 0,15503875, \tag{2.151}$$

$$P_2 = \binom{10}{2} \frac{1}{2}^2 P_0 = 0,3488371875, \;\; P_3 = \binom{10}{3} \frac{1}{2}^3 P_0 = 0,46511625, \tag{2.152}$$

$$N = \sum_{k=0}^{m} kP_k = \sum_{k=0}^{3} kP_k = 1 * P_1 + 2 * P_2 + 3 * P3 = 2,248061875. \tag{2.153}$$

pa povprečno število vzpostavljenih aktivnih sej N, ki je približno 2,248.

2.10.6 Strežna enota $M/G/1$

Strežna enota $M/G/1$ se od njej podobne strežne enote $M/M/1$ razlikuje le v procesu strežbe, kjer strežni časi niso več eksponentno porazdeljeni, ampak je njihova porazdelitev poljubna ali splošna (angl. *general distribution*). Slednje pomeni, da G predstavlja katerokoli verjetnostno porazdelitev časa strežbe z njegovim znanim povprečjem x in varianco[13] σ_x^2. Slednja je definirana po izrazu

$$\sigma_x^2 = \frac{\sum_{j=1}^{N}(x_j - x)^2}{N}, \tag{2.154}$$

pri čemer x_j predstavlja čas strežbe zahteve, ki je kot j-ta vstopila v strežnik, N pa število zahtev. Ob uporabi G porazdelitve dopuščamo možnost, da vnaša odvisnost od preteklih vrednosti slučajne spremenljivke (strežnega časa), kar pomeni, da proces strežbe v primeru $M/G/1$ ni več nujno Markovski proces, pri katerem tovrstne odvisnosti od „zgodovine“ dogodkov ni [8].

Predpostavimo, da je v $M/G/1$ strežno enoto s FIFO strežno disciplino vstopila i-ta zahteva in nas zanima njen čakalni čas. Le ta bo sestavljen kot vsota časov strežbe zahtev, ki se nahajajo v vrsti pred opazovano zahtevo in *preostalega časa strežbe* (angl. *residual service time*) zahteve, ki je že v obdelavi v strežniku. Slednjega bomo označili z r_i. r_i bo imel vrednost 0, če je strežna enota ob vstopu i-te zahteve prazna. Ob dodatni predpostavki, da je ob vstopu i-te zahteve v čakalni vrsti natanko n zahtev, lahko za njen čakalni čas zapišemo izraz

$$W_i = u(k)r_i + \sum_{j=i-n}^{i-1} x_j, \tag{2.155}$$

kjer je $u(k)$ definirana po izrazu

$$u(k) = \begin{cases} 1, & k \geq 1 \\ 0, & k = 0, \end{cases} \tag{2.156}$$

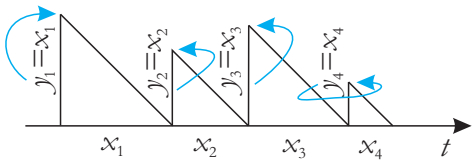
k pa predstavlja število vseh zahtev v strežni enoti. V primeru, da je strežni sistem ergodičen (v tem primeru obstaja ravnovesno stanje sistema), lahko izraz

[13] Varianca v statistiki predstavlja mero razprešnosti opazovane naključne spremenljivke X.

(2.155) zapišemo v posplošeni obliki z izrazom

$$W = \rho R_s + xN_q = \ \rho R_s + x\lambda W = \frac{\rho R_s}{1-\rho}. \tag{2.157}$$

Pri tem N_q predstavlja pričakovano število zahtev v čakalni vrsti strežne enote, R_s pa povprečni preostali čas strežbe zahteve, ki se nahaja v strežniku. Edina preostala neznanka v izrazu (2.157) je tako spremenljivka R_s ali preostali čas zahteve v strežbi. Predpostavimo, da imamo opravka z zaporedjem strežnih časov zahtev x_1,x_2,..., kot je to ponazorjeno na sliki 2.23. Zaporedje strežnih



Slika 2.23: Zaporedje strežnih časov x_1,x_2,... na časovni osi t v strežni enoti $M/G/1$.

časov x_i na časovni osi tako ponazorimo z množico trikotnikov, katerih višina je enaka času trajanja posameznega servisnega časa. Med trikotniki ni vrzeli, ker predpostavljamo, da je v strežni enoti ves čas vsaj ena zahteva, kar nam zagotovi definicija izraza $u(k)$. Tako lahko zapišemo izraz

$$R_s = \lim_{t\to\infty} \frac{povrsina\ trikotnikov}{t} =$$

$$= \lim_{t\to\infty} \frac{1}{2} \frac{\sum_{k=1}^{m(t)} x_k^2}{\sum_{k=1}^{m(t)} x_k} = \frac{E(X^2)}{2E(X)} = \frac{1+c_b^2}{2}x, \tag{2.158}$$

kjer $m(t)$ predstavlja število izvršenih strežb v časovnem intervalu t, $E(X)$ matematično upanje strežnega časa[14] in $E(X^2)$ drugi moment strežnega časa[15]. c_b^2 predstavlja *kvadrat koeficienta variacije*, ki predstavlja razmerje med varianco strežnega časa in kvadratom njegovega povprečja, definiran po izrazu

$$c_b^2 = \frac{\sigma_x^2}{x^2}. \tag{2.159}$$

Iz izraza (2.158) za preostali čas strežbe zahteve v strežniku je razvidno, da je ta čas z vidika novoprispele zahteve večji od polovice povprečnega strežnega časa

[14] Matematično upanje ali pričakovana vrednost $E(X)$ slučajne spremenljivke X je število, proti kateremu limitira povprečna vrednost spremenljivke X, ko število poskusov narašča proti neskončnosti (http://wiki.fmf.uni-lj.si/wiki/Matematično_upanje).

[15] Drugi centralni moment slučajne spremenljivke X se imenuje varianca, ki ga označujemo z σ_x^2, kjer je σ_x standardni odklon vrednosti slučajne spremenljivke X (https://sl.wikipedia.org/wiki/Centralni_moment).

($R_s > \frac{x}{2}$), ali kvečjemu enak polovici povprečnega strežnega časa v primeru, ko velja ${c_b}^2 = 0$.

Z uporabo izraza (2.158) v izrazu (2.157) pridemo do izraza

$$W = \frac{\lambda E(X^2)}{2(1-\rho)E(X)}, \tag{2.160}$$

ki ga imenujemo za Pollaczek-Khincinovo enačbo. Natančnejši razvoj predhodnih izrazov bralec najde v virih [1], [9]. Ob upoštevanju izraza za kvadrat koeficienta variacije lahko Pollaczek-Khincinovo formulo zapišemo tudi z izrazom

$$W = \frac{\rho x}{2(1-\rho)}(1 + c_b^2). \tag{2.161}$$

Osnovne kvantitativne zmogljivostne metrike strežne enote $M/G/1$ podamo z izrazi

$$N_q = \lambda * W = \frac{\lambda^2 * E(X^2)}{2(1-\rho)} = \frac{\rho^2}{2(1-\rho)}(1 + c_b^2), \tag{2.162}$$

$$T = x + \frac{\lambda * E(X^2)}{2(1-\rho)} = x + \frac{\rho x}{2(1-\rho)}(1 + c_b^2), \tag{2.163}$$

$$N = \rho + \frac{\lambda^2 * E(X^2)}{2(1-\rho)} = \rho + \frac{\rho^2}{2(1-\rho)}(1 + c_b^2). \tag{2.164}$$

V primeru, da je $c_b^2 = 1$, se nam izrazi od (2.162) do (2.164) poenostavijo v izraze za zmogljivostne metrike strežne enote $M/M/1$, v primeru pa da je $c_b^2 = 0$, pa se nam izrazi od (2.162) do (2.164) poenostavijo v izraze za zmogljivostne metrike strežne enote $M/D/1$. Same izpeljave zmogljivostnih metrik si bralec lahko ogleda v viru [1].

Zgled 10 *V preklopno vozlišče (angl. switch) prihajajo po Poissonovi porazdelitvi paketi s povprečno dolžino L z intenzivnostjo λ in čakajo na prenos po odhodnem kanalu s prepustnostjo D bps. Izračunaj povprečni čas zadrževanja paketa v preklopnem vozlišču T in povprečno število paketov N v preklopnem vozlišču za eksponentno porazdeljene L in konstantne L. Primer je povzet po viru [1].*

Rešitev: *V prvem primeru (eksponentno porazdeljene dolžine L) preklopno vozlišče obravnavamo kot $M/M/1$ strežno enoto, v drugem primeru (konstantne dolžine L) pa kot $M/D/1$ strežno enoto. V obeh primerih nam odhodni kanal predstavlja strežnik strežne enote, vozlišče pa zgolj čakalno vrsto. Velja izraz*

$$x = L/D. \tag{2.165}$$

Izračuna rešitev sta podana v nadaljevanju:

- *$M/M/1$ - eksponentna porazdelitev dolžin L - strežni časi ali časi prenosa po kanalu so porazdeljeni eksponentno:*

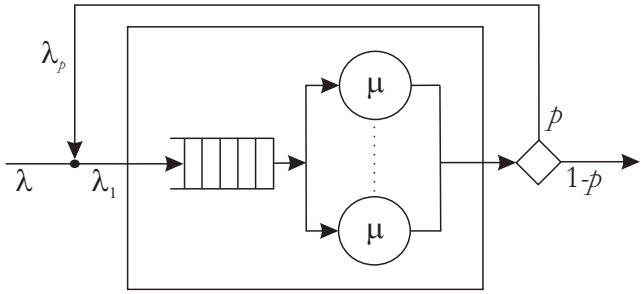
$$T = \frac{1}{\mu - \lambda}, \; N = \frac{\lambda}{\mu - \lambda}, \tag{2.166}$$

- *$M/D/1$ - konstantna porazdelitev dolžin L - strežni časi ali časi prenosa po kanalu so konstantni, tako da uporabimo nastavek za $M/G/1$ strežno enoto z robnim pogojem $c_b^2 = 0$:*

$$T = x + \frac{\rho * x}{2(1-\rho)},\ N = \rho + \frac{\rho^2}{2(1-\rho)}. \tag{2.167}$$

2.10.7 Strežna enota z delnim vračanjem zahtev v strežbo

Strežna enota z delnim vračanjem zahtev v strežbo se od predhodno opisanih razlikuje glede na strežno disciplino, ki se v tem primeru izvaja po konceptu dodeljevanja časovnih rezin strežbe (angl. *time sharing*, *processor sharing*, *time slicing* [6]). V tem primeru zahteva vstopi v strežnik, se obdeluje neko vnaprej predvideno število časovnih enot, potem pa se s strežbo zaključi ne glede na to, ali je bila zahteva postrežena do konca, ali ne. Na sliki 2.24 je predstavljena grafična ponazoritev strežne enote z delnim vračanjem zahtev v strežbo. Za



Slika 2.24: Grafična ponazoritev strežne enote z delnim vračanjem zahtev v strežbo.

omenjeno strežno enoto je specifično verjetnostno pogojeno vejanje na izhodni strani strežne enote, pri čemer se z verjetnostjo p zahteve z intenzivnostjo λ_p vračajo v čakalno vrsto strežne enote, z verjetnostjo $1-p$ pa zahteve zapuščajo strežno enoto. Velja izraz

$$\lambda_1 = \lambda + \lambda_p, \tag{2.168}$$

pri čemer λ predstavlja intenzivnost porajanja zahtev iz zunanjega okolja, λ_p intenzivnost vračanja zahtev v strežbo, λ_1 pa intenzivnost vstopanja zahtev v strežno enoto. Ob upoštevanju veljavnosti zakona o ohranitvi pretoka lahko nastavimo izraz

$$\lambda_p = \lambda_1 * p, \tag{2.169}$$

ob vstavitvi izraza (2.169) v izraz (2.168) pa pridemo do veljavnosti izraza

$$\lambda_1 = \frac{\lambda}{1-p}. \tag{2.170}$$

Za strežno enoto z delnim vračanjem zahtev v strežbo smo si ogledali zgolj izpeljavo njene intenzivnosti vstopanja zahtev λ_1. Za izpeljavo ostalih metrik bi morali imeti več informacij o značilnostih povratne povezave λ_p, pri čemer se bomo tem izpeljavam v pričujočemu delu zaradi splošnosti izognili.

2.10.8 Povzetek strežnih enot $M/M/*$

Doslej smo predpostavljali, da je narava vhodnega procesa enaka Poissonovemu procesu, za strežni proces pa smo razen v primeru strežne enote $M/G/1$ povzeli eksponentno verjetnostno porazdelitev strežnih časov. Obe predpostavki izhajata iz področij klasične in mobilne telefonije ali natančneje iz narave zahtev, ki se porajajo v klicnih omrežjih. Mednje sodijo telefonski klici, pošiljanje faksov itd. Z razliko od telefonskih klicnih omrežij imamo v današnjem času večinoma opravka z digitaliziranim podatkovnim prometom, pri katerem poteka paketni prenos podatkov. S tega vidika govorimo o *podatkovnih* ali *paketnih omrežjih*. Za slednje mnogokrat velja, da se proces strežbe ne deklarira po eksponentni porazdelitvi strežnih časov, temveč po drugačnih porazdelitvah. Ena od najbolj idealiziranih predpostavk je, da so dolžine paketov konstantne dolžine, kar vodi v *deterministične strežne čase*, manj ekstremna predpostavka pa, da predpostavimo za strežbo *splošno porazdelitev strežnih časov*. Dolžine paketov, ki neposredno vplivajo na porazdelitev strežnih časov, so odvisne od narave podatkov v paketih in komunikacijskega protokola, ki nadzoruje promet v paketnem omrežju.

Eksponentna verjetnostna porazdelitev strežbe, ki smo jo uporabljali v predhodno opisanih strežnih enotah, je z vidika izračunanih zmogljivostnih metrik v svojih napovedih pesimistična in na osnovi drugačnih modelov lahko pridemo do natančnejših in bolj optimističnih simulacijskih rezultatov. V primeru, da ostajamo pri uporabi eksponentne verjetnostne porazdelitve strežbe, lahko objektivno pričakujemo, da rezultati dinamike v realnem sistemu ne bodo slabši od simulacijskih rezultatov.

Poissonov proces in eksponentna porazdelitev strežnih časov sta z vidika primerjave simulacijskih rezultatov z dinamiko v realnih dinamičnih sistemih navkljub pesimističnosti izračunov dovolj dober približek za praktično rabo, tako da ju uporablja večina analitičnih metod in orodij.

2.11 Prioritetna strežba v strežnih enotah

V razdelku o čakalnih vrstah in strežnih disciplinah smo predhodno že omenili, da strežna disciplina v strežni enoti lahko temelji na različnih prioritetah posameznih zahtev. Predpostavimo, da imamo opravka s strežno enoto, v katero vstopajo zahteve z različnimi prioritetami p iz množice

$$p \in \{1, ..., P\}\,, \tag{2.171}$$

pri čemer $p = 1$ predstavlja najmanjšo možno prioriteto, $p = P$ pa vnaprej znano največjo možno prioriteto zahteve. V proces strežbe iz čakalne vrste venomer

vzamemo zahtevo z največjo prioriteto, če pa je takšnih zahtev več, izmed njih vzamemo iz vrste v strežbo tisto, ki je v vrsti najdlje, s čimer upoštevamo FIFO princip med zahtevami z enakimi prioritetami.

Pri tem ločujemo med *neprekinitvenim* (angl. *non-preemptive queueing*) in *prekinitvenim* (angl. *preemptive queueing*) modelom prioritetne strežbe. Pri prvem se ne glede na prioriteto novoprispele zahteve najprej dokonča strežba obdelovane zahteve, nato pa se prevzame iz vrste zahtevo z najvišjo prioriteto. V drugem primeru se ob vstopu zahteve z višjo prioriteto, kot je prioriteta trenutno obdelovane zahteve, obdelava prekine in v strežbo se prevzame novoprispelo zahtevo.

S prehodom na različne prioritete se spremeni tudi obravnava zmogljivostnih metrik, ki smo jih obravnavali v predhodnih razdelkih. Tako lahko zapišemo sledeče splošne izraze

$$\lambda = \sum_{p=1}^{P} \lambda_p, \;\; x = \sum_{p=1}^{P} \frac{\lambda_p}{\lambda} x_p, \tag{2.172}$$

$$\rho_p = \lambda_p * x_p, \;\; \rho = \lambda * x = \sum_{p=1}^{P} \rho_p, \tag{2.173}$$

$$T_p = W_p + x_p, \tag{2.174}$$

kjer indeks p pri posamezni spremenljivki določa prioritetno skupino ali prioritetni razred zahtev, P pa največjo možno prioriteto zahteve. λ, x, ρ, T in W nam predstavljajo že znane numerične metrike in sicer po vrsti intenzivnost prihajanja zahtev, povprečni strežni čas, faktor uporabnosti, čas zadrževanja zahteve v strežni enoti in čakalni čas zahteve v čakalni vrsti. Za uporabnostni faktor zopet predpostavljamo veljavnost relacije $\rho < 1$. Omenjeni izrazi so venomer veljavni zgolj za $M/G/1$ strežno enoto, ker je ob možnosti prekinjanja strežbe veljavnost eksponentne porazdelitve strežnih časov mnogokrat vprašljiva.

2.11.1 Neprekinitveni model strežbe

Za neprekinitveni model strežbe velja, da novoprispela zahteva ne glede na svojo prioriteto ne more prekiniti strežbe zahteve, ki je trenutno v strežniku. Predpostavimo, da je v strežno enoto $M/G/1$ z neprekinitvenim modelom strežbe pravkar vstopila nova zahteva s prioriteto p. Njen pričakovani čas čakanja na strežbo W_p je sestavljen kot vsota naslednjih časov [1]:

(i) povprečni preostali čas strežbe (angl. *residual service time*), ki je potreben za dokončanje strežbe zahteve, ki se že nahaja v strežniku: ob prispetju zahteve s prioriteto p je verjetnost, da je v strežbi zahteva s prioriteto j, pogojena z izrazom

$$\rho_j = \lambda_j x_j; \tag{2.175}$$

povprečni preostali čas strežbe tako izrazimo z vsoto povprečnih preostalih časov uteženo z verjetnostmi ρ_k po izrazu

$$R_1 = \sum_{k=1}^{P} \rho_k \left(\frac{E(X_k^2)}{2 * E(X_k)}\right) = \frac{1}{2} \sum_{k=1}^{P} \lambda_k E(X_k^2) = \frac{1}{2} \sum_{k=1}^{P} \rho_k \frac{1 + c_{b_k}^2}{2} x_k; \quad (2.176)$$

(ii) pričakovani čas, ki je potreben za strežbo vseh zahtev iz čakalne vrste z enako prioriteto p, ki so v čakalno vrsto prispele pred opazovano zahtevo: omenjeni čas izračunamo na osnovi izraza

$$R_2 = x_p N_q^p, \quad (2.177)$$

kjer N_q^p odraža povprečno število zahtev s prioriteto p v čakalni vrsti;

(iii) pričakovani čas, ki je potreben za strežbo vseh zahtev iz čakalne vrste z višjo prioriteto: pri slednjih upoštevamo le tiste zahteve, ki so v čakalno vrsto prispele pred opazovano zahtevo; omenjeni čas izračunamo po izrazu

$$R_3 = \sum_{j=p+1}^{P} x_j N_q^j; \quad (2.178)$$

(iv) pričakovani čas, ki je potreben za strežbo zahtev s prioriteto višjo od p, ki so v strežno enoto vstopile po opazovani zahtevi: omenjeni čas izračunamo po izrazu

$$R_4 = \sum_{j=p+1}^{P} x_j \lambda_j W_j; \quad (2.179)$$

Tako celoten čas čakanja zahteve W_p s prioriteto p lahko zapišemo z izrazom

$$W_p = \sum_{i=1}^{4} R_i = \frac{1}{2} \sum_{k=1}^{P} \rho_k \frac{1 + c_{b_k}^2}{2} x_k + \sum_{j=p}^{P} x_p N_q^p + \sum_{j=p+1}^{P} x_j \lambda_j W_j. \quad (2.180)$$

Čas čakanja zahteve z največjo prioriteto P lahko zapišemo z izrazom

$$W_P = \frac{R_1}{1 - \rho_P}. \quad (2.181)$$

Upoštevajoč izraze (2.180), (2.181) in veljavnost relacije v izrazu

$$N_q^k = \lambda_k W_k, \quad (2.182)$$

pridemo do izraza

$$W_{P-1} = \frac{R_1}{(1 - \rho_P)(1 - \rho_P - \rho_{P-1})}, \quad (2.183)$$

z rekurzivnimi ponovitvami pa do splošnega nastavka za izračun čakalnega časa zahteve s prioriteto i, ki ga zapišemo z izrazom

$$W_i = \frac{R_1}{(1-\rho_P-\rho_{P-1}-...-\rho_{i+1})(1-\rho_P-\rho_{P-1}-...-\rho_i)}. \tag{2.184}$$

Osnovne kvantitativne zmogljivostne metrike neprekinitvenega sistema strežbe v strežni enoti M/G/1 so sledeče [1]:

- povprečno število zahtev s prioriteto i v čakalni vrsti strežne enote se izračuna po izrazu

$$N_{q_i} = \lambda_i W_i = \frac{\lambda_i R_1}{(1-\rho_P-\rho_{P-1}-...-\rho_{i+1})(1-\rho_P-\rho_{P-1}-...-\rho_i)}; \tag{2.185}$$

- povprečno število zahtev v sistemu se izračuna po izrazu

$$N = \sum_{k=1}^{P} N_{q_k} + \rho; \tag{2.186}$$

- v primeru, da so strežni časi vseh prioritetnih razredov eksponentno porazdeljeni s povprečno vrednostjo strežnega časa x_p za posamezni prioritetni razred p, imamo opravka s strežno enoto $M/M/1$ s prioritetno neprekinitveno strežbo; v tem primeru velja izraz

$$R = \frac{1}{2}\sum_{k=1}^{P} \lambda_k (\frac{1}{\mu})^2 = \frac{\rho}{2\mu}; \tag{2.187}$$

Zgled 11 *V omrežju imamo opravka s podatkovnimi in kontrolnimi paketi s fiksnimi dolžinami. Prvi so dolgi 4.800 bitov, drugi pa 192 bitov, pri čemer je zmogljivost posamezne prenosne linije 19.200 bitov na sekundo. V povprečju je 15% paketov kontrolnih, 85% paketov pa podatkovnih. Obe skupini paketov prihajata po Poissonovi porazdelitvi s skupno intenzivnostjo prihajanja 4 paketov na sekundo. Izračunaj povprečni čas čakanja na prenosno linijo (i) in isti čas ob predpostavki, da imamo opravka s prioritetnim neprekinitvenim načinom strežbe, pri čemer imajo kontrolni paketi višjo prioriteto (ii). Primer je povzet po viru [1].*

Rešitev: *Pred reševanjem problema moramo najprej definirati podatke. Veljajo izrazi*

$$x_d = \frac{4800}{19200} = 0,25 \ sekunde, \ \sigma_{x_d} = 0, \ x_d^2 = (0,25)^2 = 0,0625, \tag{2.188}$$

$$x_c = \frac{192}{19200} = 0,01 \ sekunde, \ \sigma_{x_c} = 0, \ x_c^2 = (0,01)^2 = 0,0001, \tag{2.189}$$

$$\lambda_d = 0,85 * \lambda, \ \lambda_c = 0,15 * \lambda. \tag{2.190}$$

Odtod sledita odgovora na vprašanji, ki sta sledeča:

(i) kombinirani promet paketov lahko modeliramo kot $M/G/1$ strežno enoto; čakalni čas W izračunamo po izrazih

$$x^2 = \frac{\lambda_c}{\lambda}x_c^2 + \frac{\lambda_d}{\lambda}x_d^2 = 0,05314 \ sekunde, \tag{2.191}$$

$$x = \frac{\lambda_c}{\lambda}x_c + \frac{\lambda_d}{\lambda}x_d = 0,214 \ sekunde, \tag{2.192}$$

$$W = \frac{\lambda x^2}{2(1-\rho)} = \frac{4 * 0,05314}{2(1-4*0,214)} = 0,738 \ sekunde; \tag{2.193}$$

(ii) kombinirani promet paketov s prioritetami lahko modeliramo kot $M/G/1$ strežno enoto; čakalna časa W_c in W_d izračunamo po izrazih

$$\rho_c = \lambda_c x_c = 0,006, \ \ \rho_d = \lambda_d x_d = 0,85, \tag{2.194}$$

$$R_1 = \frac{1}{2}\sum_{k=1}^{2}\lambda_k x^2 = 0,10628, \tag{2.195}$$

$$W_c = \frac{R_1}{1-\rho_c} = \frac{0,10628}{1-0,006} = 0,10692 \ sekunde, \tag{2.196}$$

$$W_d = \frac{R_1}{(1-\rho_c)(1-\rho_c-\rho_d)} = \frac{0,10628}{0,994 * 0,144} = 0,7425 \ sekunde. \tag{2.197}$$

Iz izračunov je razvidno, da prioritetna strežba z neprekinitvenim modelom ob podanem razmerju prihajajočih paketov močno skrajša čakalni čas paketov z višjo prioriteto, pri čemer se čakalni čas paketov z nižjo prioriteto podaljša minimalno.

2.11.2 Prekinitveni model strežbe

Za omenjeni model velja, da novoprispela zahteva z višjo prioriteto prekine strežbo zahteve z manjšo prioriteto in sama vstopi v proces strežbe. V splošnem velja, da bo zahteva, katere strežba je bila prekinjena, vrnjena v strežbo po času servisiranja prekinjajoče zahteve, pri čemer število ponovitev prekinjanja posamezne zahteve ni omejeno. Tudi za ta primer bomo sestavili izraze po vzoru predhodnega razdelka.

Predpostavimo, da imamo opravka z novoprispelo zahtevo s prioriteto p. Z vidika te zahteve so vse zahteve z nižjimi prioritetami od p irelevantne - opazovana zahteva bo postrežena pred njimi. Za zahteve, ki so prispele pred opazovano zahtevo in imajo višje prioritete, je z vidika opazovane zahteve popolnoma vseeno, ali imamo prekinitveni ali neprekinitveni model strežbe, saj bodo vse postrežene pred njo. Tako velja ocena za čakalni čas opazovane zahteve po

izrazu

$$W_p = \frac{R_p}{(1-\rho_P-\rho_{P-1}-...-\rho_{p+1})(1-\rho_P-\rho_{P-1}-...-\rho_p)}, \tag{2.198}$$

ki smo ga izeljali že v prejšnjem razdelku in izraz za preostali čas strežbe zahtev iz posameznih prioritetnih razredov

$$R_p = \frac{1}{2}\sum_{k=p}^{P}\lambda_k\sigma_{x_k}^2. \tag{2.199}$$

Izraz (2.198) ne upošteva dejstva, da je lahko strežba zahteve s prioriteto p tudi prekinjena. Predpostavimo, da njen čas od začetka do konca strežbe, ki vključuje tudi čas prekinitve strežbe, označimo s T_p'. Tako velja izraz

$$T_p' = x_p + \sum_{j=p+1}^{P} x_j\lambda_j T_p', \tag{2.200}$$

kjer $\lambda_j T_p'$ predstavlja povprečno število prispelih zahtev z višjo prioriteto od opazovane zahteve v času T_p'. Tako lahko zapišemo spremenjeno enačbo za čas zadrževanja zahteve s prioriteto p v strežni enoti po izrazu

$$T_p = W_p + T_p' =$$

$$= \frac{R_p}{(1-\rho_P-\rho_{P-1}-...-\rho_{p+1})(1-\rho_P-\rho_{P-1}-...-\rho_p)} + \frac{x_p}{(1-\rho_P-\rho_{P-1}-...-\rho_{p+1})}. \tag{2.201}$$

Bralec si lahko izpeljavo navedenih izrazov ogleda v viru [1].

Zgled 12 *Predpostavimo, da imamo opravka s strežnim sistemom, kot smo ga opisali v zgledu (11), pri čemer je strežna disciplina realizirana s prekinitvenim modelom strežbe. Izračunaj čase zadrževanja obeh vrst paketov v strežni enoti. Primer je povzet po viru [1].*

Rešitev: *Pred dokončnim izračunom odgovora moramo najprej izračunati izraze*

$$R_1 = \frac{1}{2}\lambda_c x_c^2 = 0,00003, R_2 = \frac{1}{2}(\lambda_c x_c^2 + \lambda_d x_d^2) = 0,10628, \tag{2.202}$$

$$T_c = \frac{x_c(1-\rho_c)+R_1}{1-\rho_c} = 0,01003 \ sekunde, \tag{2.203}$$

$$T_d = \frac{x_d(1-\rho_c-\rho_d)+R_2}{(1-\rho_c)(1-\rho_c-\rho_d)} = 0,994 \ sekunde. \tag{2.204}$$

Če dobljene rezultate časa nahajanja paketov v strežni enoti primerjamo z rezultatoma iz zgleda (11), ki sta po vrsti

$$T_c = W_c + x_c = 0,11692 \ sekunde, \tag{2.205}$$

$$T_d = W_d + x_d = 0,9925 \; sekunde, \tag{2.206}$$

lahko ugotovimo, da prekinitveno servisiranje še dodatno pospeši pretok paketov z višjo prioriteto na račun minimalne upočasnitve paketov z nižjo prioriteto.

2.12 Strežne mreže

Že na začetku pričujočega poglavja smo v definiciji navedli, da *strežno mrežo* sestavljajo poljubne zaporedne, paralelne ali mešane vezave m strežnih enot ($m > 1$). Iz povedanega sledi, da imamo v strežnih mrežah običajno več čakalnih vrst in več funkcionalno različnih strežb.

Strežne mreže ločujemo na dve skupini in sicer na *odprte* in *zaprte* strežne mreže. V prve zahteve vstopajo iz zunanjega okolja in se vanj postrežene vračajo. V drugih kroži konstantno število zahtev, ki okolja sistema zaprte strežne ne zapuščajo, poleg tega pa vanj nove zahteve ne vstopajo. Stanje strežne mreže običajno ponazorimo s številom zahtev v njenih strežnih enotah, ki ga zapišemo z vektorjem populacije zahtev

$$q(t) = (j_1(t), j_2(t), ..., j_n(t)), \tag{2.207}$$

pri čemer $q(t)$ predstavlja stanje strežne mreže v času t, $j_i(t)$ število zahtev v i-ti strežni enoti v času t, n pa število strežnih enot.

V pričujočem razdelku si bomo ogledali primer obravnave zaprte strežne mreže. Za zaprte strežne mreže veljata izraza

$$\forall t : K = \sum_{i=1}^{n} j_i(t), \tag{2.208}$$

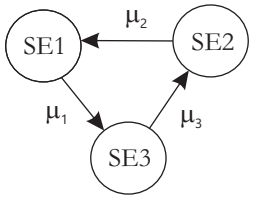
$$M = \binom{n + K - 1}{n - 1}, \tag{2.209}$$

pri čemer K predstavlja število zahtev v zaprti strežni mreži, M pa število različnih možnih stanj zaprte mreže.

Predpostavimo, da imamo opravka z zaprto strežno mrežo sestavljeno iz treh strežnih enot tipa $M/M/1$ (SE1, SE2 in SE3). Intenzivnosti strežnih enot so po vrsti μ_1, μ_2 in μ_3. V strežni mreži se nahajata dve zahtevi. Ti zahtevi prehajata ločeno med strežnimi enotami, kot je prikazano na sliki 2.25.

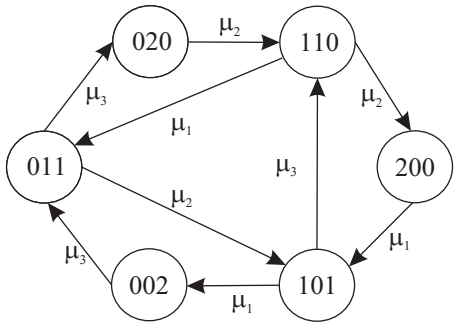
Iz slike je razvidno, da strežna enota SE1 predaja postreženo zahtevo z intenzivnostjo μ_1 strežni enoti SE3, strežna enota SE3 predaja postreženo zahtevo z intenzivnostjo μ_3 strežni enoti SE2 itd. Intenzivnost predaje zahteve naslednji strežni enoti v ciklu je pogojena z intenzivnostjo strežbe enote, ki zahtevo predaja.

Za predhodno opisani primer strežne mreže skušajmo odgovoriti na vprašanje, kolikšna je verjetnost stanja mreže z natanko eno zahtevo v prvi strežni enoti ob podanih intenzivnostih strežbe. Na osnovi izraza (2.209) izračunamo, da je število različnih stanj sistema $M = 6$. Na ta način pridemo do diagrama



Slika 2.25: Primer zaprte strežne mreže.

prehajanja stanj, ki je prikazan na sliki 2.26. V posameznem stanju je zapisan vektor števil, ki ponazarja porazdelitev zahtev po posameznih strežnih enotah. Tako vektor (110) predstavlja po eno zahtevo v prvi in drugi strežni enoti, v tretji strežni enoti pa ni zahtev.



Slika 2.26: Diagram prehajanja stanj v vzorčni zaprti strežni mreži.

Na osnovi diagrama prehajanja stanj nastavimo ravnotežne enačbe, v katerih enačimo vhodne intenzivnosti v opazovana stanja uteženе z verjetnostmi predhodnih stanj z izhodnimi intenzivnostmi opazovanih stanj uteženimi z verjetnostmi opazovanih stanj.

$$\mu_1 * p(2,0,0) = \mu_2 * p(1,1,0), \tag{2.210}$$

$$\mu_2 * p(0,2,0) = \mu_3 * p(0,1,1), \tag{2.211}$$

$$\mu_3 * p(0,0,2) = \mu_1 * p(1,0,1), \tag{2.212}$$

$$\mu_1 * p(1,1,0) + \mu_2 * p(1,1,0) = \mu_2 * p(0,2,0) + \mu_3 * p(1,0,1), \tag{2.213}$$

$$\mu_2 * p(0,1,1) + \mu_3 * p(0,1,1) = \mu_1 * p(1,1,0) + \mu_3 * p(0,0,2), \tag{2.214}$$

$$\mu_3 * p(1,0,1) + \mu_1 * p(1,0,1) = \mu_1 * p(2,0,0) + \mu_2 * p(0,1,1). \tag{2.215}$$

Ob upoštevanju dejstva, da je vsota verjetnosti nahajanj v stanjih gledano preko vseh šestih možnih stanj strežne mreže enaka 1, kar ponazorimo z izrazom

$$p(1,0,1) + p(1,1,0) + p(0,1,1) + p(2,0,0) + p(0,2,0) + p(0,0,2) = 1, \tag{2.216}$$

bi s substitucijami lahko izluščili verjetnosti $p(1,1,0)$ in $p(1,0,1)$. Vsota obeh bi dala odgovor na zastavljeno vprašanje.

2.13 Praktični napotki za modeliranje računalniških omrežij na osnovi teorije strežbe

Na koncu pričujočega poglavja o teoriji strežbe, ki nam omogoča *zmogljivostno modeliranje računalniških omrežij*, navedimo še nekaj pomembnih vodil za snovanje modelov računalniških omrežij na osnovi teorije strežbe. Le ti so sledeči:

- sestavne dele opazovanega sistema, ki ga modeliramo, abstrahiramo na strežne enote in njihove medsebojne povezave; pomembno je, da v model vključimo vse vplivne sestavne dele opazovanega sistema, manj pomembne ali nevplivne pa zaradi eventuelne preobsežnosti modela mnogokrat izpuščamo;
- pri modeliranju opazovanega procesa moramo proces rojevanja in strežbe zahtev abstrahirati na karakteristične porazdelitve, ki so zgolj približek realni dinamiki; pri tem večinoma posegoma po Poissonovi porazdelitvi za modeliranje procesa rojevanja zahtev in po eksponentni porazdelitvi za modeliranje časov strežbe, pri čemer sta oba pristopa v svojih napovedih pesimistična in lahko v praksi pričakujemo pri delovanju omrežja v normalnih pogojih boljše zmogljivostne rezultate;
- v primeru, da strežni ali prihajalni proces ni popolnoma naključen, posegamo po splošni porazdelitvi, v redkem primeru pa da je eden ali sta oba procesa popolnoma deterministična, pa posegamo po deterministični porazdelitvi;
- model opazovanega sistema mnogokrat posplošimo - pristanemo na poenostavitve interpretacije strežnih značilnosti sistema;
- zaradi kompleksnosti matematičnih izračunov, ki z naraščanjem števila strežnih enot v odprti strežni mreži raste izredno hitro, običajno kompleksnejše vezave strežnih enot analiziramo s pomočjo programskih orodij;

Literatura

[1] N. C. Hock, *Queuing Modelling Fundamentals.* John Wiley & Sons, Chichester, Anglija, 1996.

[2] M. Anu, "Introduction to modeling and simulation," in *Proceedings of the 29th conference on Winter simulation* (S. Andradóttir, K. J. Healy, D. H. Withers, and B. L. Nelson, eds.), pp. 7–13, 1997.

[3] L. Kleinrock and R. Gail, *Queuing systems, problems and solutions.* John Wiley & Sons, New York, ZDA, 1996.

[4] N. Zimic and M. Mraz, *Temelji zmogljivosti računalniških sistemov.* Založba FE in FRI, Ljubljana, Slovenija, 2006.

[5] R. Jamnik, *Verjetnostni račun in statistika.* Društvo matematikov, fizikov in astronomov socialistične republike Slovenije, Zveza organizacij za tehnično kulturo Slovenije, Ljubljana, Slovenija, 1986.

[6] K. S. Trivedi, *Probability and Statistics with Reliability, Queueing and Computer Science Applications.* John Wiley & Sons Inc., New York, ZDA, 2002.

[7] H. Stöcker, *Matematični priročnik z osnovami računalništva.* Tehnična založba Slovenije, Ljubljana, Slovenija, 2006.

[8] J. Virant, *Modeliranje in simuliranje računalniških sistemov.* Didakta, Radovljica, Slovenija, 1991.

[9] J. F. Shortle, J. M. Thompson, D. Gross, and C. M. Harris, *Fundamentals of queueing theory.* John Wiley & Sons, Hoboken, ZDA, 2018.