Rozdział 1

Systemy masowej obsługi

Systemy masowej obsługi (lub kolejkowe) występują w wielu praktycznych sytuacjach, np. samoloty na lotnisku oczekują na start lub lądowanie, klienci w banku oczekują na obsługę przy okienku, zadania obliczeniowe oczekują na procesor lub zadania wydruku na drukarkę wielodostępną itd. Wynikiem analizy takich systemów jest ustalenie optymalnej liczby stanowisk obsługi lub zasady wyboru zgłoszeń oczekujących w kolejce do obsługi. Celem może być zapewnienie przez system założonego średniego czasu oczekiwania na obsługę lub określonej długości kolejki. Ze względu na fakt, że dokładne czasy przybycia zadań do systemu, jak i czasy obsługi zadań przez system nie są znane a priori przyjmuje się, że są one zmiennymi losowymi i analizuje się je za pomocą modeli niedeterministycznych.

Przykład 1.1. *Hrabina Hermenegilda Kociubińska prowadziła mały pensjonat. W pensjonacie na stałe mieszkała baronowa von und zu Allesgute, mecenas Paragrafek i profesor Rozumek, emerytowany matematyk. Goście chwalili zwłaszcza pyszne desery, które hrabina przyrządzała na bazie lodów domowej roboty wg receptury babki Hiacynty. Hrabina ledwo wiązała koniec z końcem, a tymczasem potrzeby związane z wykształceniem kilkunastoletniej córki ciągle rosły. Hrabina postanowiła rozszerzyć swoją działalność i otworzyć kiosk z lodami licząc na zainteresowanie turystów licznie odwiedzających piękna okolicę pensjonatu. Realizacja tego pomysłu praktycznie nie wymagała inwestycji. Niestety, już na początku pojawił się problem: ile lodów należy przygotować? Następnie: o której godzinie rozpocząć sprzedaż, a kiedy zakończyć? O jakiej porze będzie najwięcej klientów? Hrabina poprosiła o pomoc profesora Rozumka, który orzekł, że problem jest niedeterministyczny i podjął się obserwacji podobnego punktu działającego w pobliżu dworca PKP.*

Jako, że profesor był już na emeryturze codziennie rano udawał się w okolice dworca i w swoim grubym notesie zapisywał dokładnie godzinę, o której przy-

bywał i wychodził każdy klient. Uznał, że warto analizować godziny od 10:00 do 18:00. Notatki profesora zawiera Tablica 1.1

Tabela 1.1. Obserwacje profesora Rozumka

godzina	zdarzenie
10:01	przybywa klient 1
10:04	przybywa klient 2
10:05	wychodzi klient 1
10:06	przybywa klient 3
10:08	wychodzi klient 3
10:10	przybywa klient 4
10:15	wychodzi klient 2
10:17	wychodzi klient 4
10:19	przybywa klient 5
10:22	przybywa klient 6
10:25	przybywa klient 7
10:26	przybywa klient 5
10:27	wychodzi klient 6
10:29	przybywa klient 8

itd.

Tabela 1.2. Analiza procesu przybywania klientów do systemu

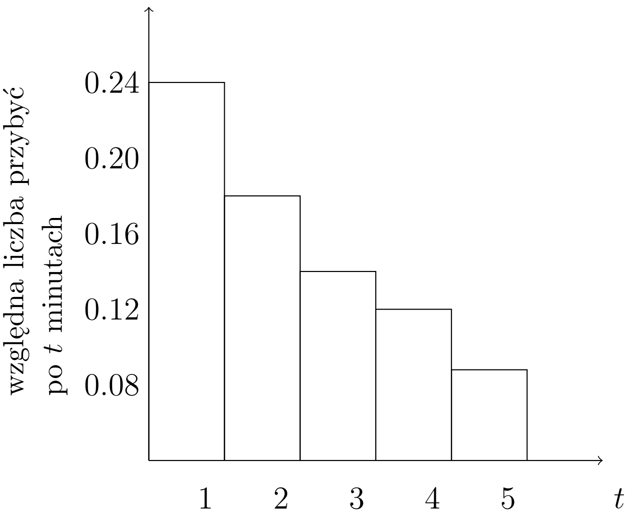
przedział czasu	odstęp
10:01 – 10:04	3 minuty między klientem 1 i klientem 2
10:04 – 10:06	2 minuty między klientem 2 i klientem 3
10:06 – 10:10	4 minuty między klientem 3 i klientem 4
10:10 – 10:19	9 minut między klientem 4 i klientem 5
10:19 – 10:22	3 minuty między klientem 5 i klientem 6
10:22 – 10:25	3 minuty między klientem 6 i klientem 7
10:25 – 10:29	4 minuty między klientem 7 i klientem 8

itd.

1.0.1. Analiza procesu przybywania zadań do systemu

Po tygodniu tablica 1.1 znacznie się rozrosła, a po miesiącu nie sposób było ją analizować w tej postaci. Na początek profesor skupił się na analizie procesu przybywania klientów przyjmując za parametr odstęp czasu między kolejnymi klientami przybywającymi do pensjonatu. Wyniki przedstawia Tablica 1.2.

Dane zebrane w ciągu miesiąca profesor podsumował w postaci histogramu (Rys. 1.8), na którym przedstawił względną liczbę klientów przybywających po



Rysunek 1.1. Histogram czasów przybycia do systemu

upływie t minut po poprzednim. W tablicy 1.2, na przykład, po 3 minutach po poprzednim przybyło 3 z 9 klientów, a zatem $1/3$ *wszystkich ujętych w tablicy.*

1.0.2. Rozkład wykładniczy

Analizując histogram z rysunku 1.8 można zauważyć, że udział klientów przybywających do systemu t minut po porzednim maleje wykładniczo ze wzrostem t. Zjawisko to można opisać funkcją $f(t) = \lambda e^{-\lambda t}$ dla $\lambda > 0$. Metodą prób i błędów profesor ustalił, że w przypadku klientów domowych obiadów pani hrabiny współczynnik $\lambda = \frac{1}{3}$.

Jak pamiętamy, funkcja $f(t) = \lambda e^{-\lambda t}$ dla $t \geq 0$ jest funkcją gęstości prawdopodobieństwa rozkładu wykładniczego. Oznaczmy przez U zmienną losową oznaczającą czas między kolejnymi klientami.

Możemy zatem wyznaczyć prawdopodobieństwo zdarzenia, że odstęp między kolejnymi klientami będzie mniejszy niż t (minut) znajdując dystrybuantę $F(t)$ rozkładu zmiennej losowej U.

$$P(U \leq t) = F(t) = \int_0^t f(x)dx \tag{1.1}$$

Przyjmując, że U ma rozkład wykładniczy otrzymujemy:

$$P(U \leq t) = \int_0^t \lambda e^{-\lambda x} dx = \left[-e^{-\lambda x}\right]_0^t = -e^{-\lambda t} - (-1) = 1 - e^{-\lambda t} \tag{1.2}$$

Wartość oczekiwaną zmiennej losowej U obliczamy jako:

$$E(U) = \int_0^\infty x f(x) dx = \int_0^\infty t\lambda e^{-\lambda t} dt \tag{1.3}$$

$$\begin{aligned} g(t) &= t \qquad & f'(t)dt &= \lambda t e^{-\lambda t} dt \\ g'(t)dt &= dt \qquad & f(t) &= -e^{-\lambda t} \end{aligned}$$

$$\int_0^\infty t\lambda e^{-\lambda t} dt = \left[-te^{-\lambda t}\right]_0^\infty + \int_0^\infty (-e^{-\lambda t}) dt = \left[-te^{-\lambda t}\right]_0^\infty + \left[\frac{1}{\lambda} e^{-\lambda t}\right]_0^\infty = \frac{1}{\lambda} \tag{1.4}$$

$$E(U) = \frac{1}{\lambda} \tag{1.5}$$

Fakt, że odstęp między kolejnymi zgłoszeniami ma rozkład wykładniczy wykorzystamy w dalszej analizie problemu.

Oznaczmy przez $X(t)$ zmienną losową oznaczającą liczbę klientów, którzy przybyli w przedziale $[0, t)$ i załóżmy, że
— klineci przybywają niezależnie od siebie (niezależne przyrosty),
— częstotliwość przybywania klinetów nie zależy od pory dnia (stacjonarny),
— zdarzenia zachodzą pojedynczo,
— zdarzenia zachodzą błyskawicznie.

Niech $X(t)$ oznacza liczbę zgłoszeń przybywajacych w przedziale $[0, t)$. Oznaczymy $P\{X(t) = n\} = P_n(t)$.

Dalej wykorzystajmy błyskawiczność zachodzenia zdarzeń:

$$P_0(\Delta t) = P(U > \Delta t) = 1 - P(U \leq \Delta t) = e^{-\lambda \Delta t}$$

Korzystając z rozwinięcia funkcji e^x mamy:

$$e^x = 1 + x + \frac{x^2}{2!} + \frac{x^3}{3!} + \dots$$

$$P_0(\Delta t) = 1 - \lambda\Delta t + o(\Delta t)$$

gdzie

$$\lim_{\Delta t \to 0} \frac{o(\Delta t)}{\Delta t} = 0$$

Klienci przybywają w sposób niezależny, a zatem jeżeli oznaczę przez p prawdopodobieństwo, że w czasie Δt przybędzie jeden klient, to prawdopodobieństwo, że przybędzie w tym czasie dwóch klientów wynosi p^2, trzech - p^3, itd. Ponadto wiadomo, że:

$$\sum_{k=0}^{\infty} P_k(\Delta t) = 1$$

Zatem

$$\sum_{k=0}^{\infty} P_k(\Delta t) = P_0(\Delta t) + p + p^2 + p^3 + \dots = 1$$

$$1 - \lambda\Delta t + o(\Delta t) + 1 + p + p^2 + p^3 + \dots = 1 + 1$$

$$1 - \lambda\Delta t + o(\Delta t) + \frac{1}{1-p} = 2$$

$$-\lambda\Delta t + o(\Delta t) + \frac{1}{1-p} = 1$$

$$-\lambda\Delta t(1-p) + o(\Delta t)(1-p) + 1 = 1 - p$$

$$-\lambda\Delta t + p\lambda\Delta t + o(\Delta t)(1-p) = p$$

$$p(\lambda\Delta t + 1) + o(\Delta t)(1-p) = \lambda\Delta t$$

$$p = \frac{\lambda\Delta t}{(\lambda\Delta t + 1)} + o(\Delta t)$$

Zastosujemy ponownie wzór na sumę szeregu geometrycznego ($-1 < \lambda\Delta t <$ 1)

$$p = \frac{1}{1 - (-\lambda\Delta t)}\lambda\Delta t + o(\Delta t)$$

$$p = \lambda\Delta t - (\lambda\Delta t)^2 + (\lambda\Delta t)^3 - (\lambda\Delta t)^4 + \ldots + o(\Delta t)$$

$$P_1(\Delta t) = \lambda\Delta t + o(\Delta t)$$

mamy zatem średnią liczbę zadań przybywających w jednostce czasu (szybkość przybywania zgłoszeń):

$$\lim_{\Delta t \to 0} \frac{P_1(\Delta t)}{\Delta t} = \lambda$$

1.1. Proces przybywania klientów jako proces losowy

Wprowadzimy teraz zmienną losową $X(t)$ oznaczającą liczbę klientów, którzy przybyli w okresie $\langle 0, t)$. Zgodnie z definicją 1.1, proces przybywania zgłoszeń do systemu można traktować jako proces sygnałowy.

Definicja 1.1. *Proces losowy* $\{X(t), t \in \langle 0; \infty)\}$*, gdzie* $X(t)$ *jest liczbą zdarzeń losowych (zgłoszeń, sygnałów) w przedziale* $\langle 0, t)$ *nazywamy* procesem sygnałowym.

Zauważmy, że:
— $X(t)$ jest liczbą naturalną,
— jeżeli $t_1 < t_2$, to $X(t_1) \leq X(t_2)$,
— zatem $X(t_2) - X(t_1)$ jest liczbą naturalną.

Ponadto przyjmiemy, że

$$P[X(0) = 0] = 1. \tag{1.6}$$

Definicja 1.2. *Proces sygnałowy nazywamy* procesem sygnałowym o przyrostach stacjonarnych *jeśli rozkład prawdopodobieństwa liczby zgłoszeń w przedziale o dowolnej długości* Δt *nie zależy od położenia tego przedziału na osi* *czasu. W szczególności oznacza to, że średnia szybkość przybywania zgłoszeń dla takiego procesu jest stała.*

Definicja 1.3. *Proces sygnałowy nazywamy* procesem sygnałowym o przyrostach niezależnych *jeśli liczby zgłoszeń w dowolnych rozłącznych przedziałach czasowych są zmiennymi losowymi niezależnymi. Zauważmy, że oznacza to, iż również odstępy czasowe pomiędzy kolejnymi zgłoszeniami są zmiennymi losowymi niezależnymi.*

Na mocy założenia 1.6 proces sygnałowy o przyrostach naturalnych jest procesem Markova (szczególny przypadek warunku $P[X(t_0) = c] = 1$).

Definicja 1.4. *Proces Markova jest to proces stochastyczny, dla którego*

$$P(X_n = k_n | X_{n-1} = k_{n-1}, X_{n-2} = k_{n-2}, \ldots, X_0 = k_0) = P(X_n = k_n | X_{n-1} = k_{n-1})$$

czyli proces bez pamięci.

Definicja 1.5. *Mówimy, że zmienne losowe są niezależne, gdy dla każdych liczb rzeczywistych zachodzi równość* $P(X \leq A)P(Y \leq B) = P(X \leq A \wedge Y \leq B)$.

Załóżmy, że nasz proces losowy znajduje się w stanie $X(t) \geq 0$. Rozważmy prawdopodobieństwo możliwych przejść stanu procesu $X(t) \to X(t+\Delta t), \Delta t \to 0$. Zauważmy, że ze względu na pojedynczość zachodzenia zdarzeń mamy tylko 2 możliwości:

$$X(t + \Delta t) = X(t) + 1$$

$$X(t + \Delta t) = X(t)$$

Zatem

$$P[X(t+\Delta t) = X(t)+1] = P(\text{w czasie } \Delta t \text{ przybyło 1 zgłoszenie}) = \lambda\Delta t + o(\Delta t)$$

$$P[X(t + \Delta t) = X(t)] = P(\text{w czasie } \Delta t \text{ przybyło 0 zgłoszeń}) = 1 - \lambda\Delta t - o(\Delta t)$$

Przeanalizujemy zatem przypadek ogólny:

$P_n(t + \Delta t) =$

P(w przedziale o dł. t przybyło n−1 zgłoszeń i w przedziale o dł. Δt przybyło 1 zgłoszenie)+
P(w przedziale o dł. t przybyło n zgłoszeń i w przedziale o dł. Δt przybyło 0 zgłoszeń)

zatem

$$P_n(t+\Delta t) = P_{n-1}(t)P_1(\Delta t) + P_n(t)(1 - P_1(\Delta t)) =$$
$$P_{n-1}(t)(\lambda\Delta t + o(\Delta t)) + P_n(t)(1 - (\lambda\Delta t + o(\Delta t))) =$$
$$\lambda P_{n-1}(t)\Delta t + P_n(t) - \lambda P_n(t)\Delta t + o(\Delta t)$$

$$\lim_{\Delta t\to 0} \frac{P_n(t+\Delta t)}{\Delta t} = \lambda P_{n-1}(t) + \lim_{\Delta t\to 0} \frac{P_n(t)}{\Delta t} - \lambda P_n(t)$$

$$\lim_{\Delta t\to 0} \frac{P_n(t+\Delta t)-P_n(t)}{\Delta t} = \lambda P_{n-1}(t) - \lambda P_n(t)$$

$$P_n'(t) = \lambda P_{n-1}(t) - \lambda P_n(t), n = 0, 1, 2, \ldots$$

Otrzymaliśmy zatem nieskończony układ równań różniczkowo-różnicowych, z którego możemy wyznaczyć dowolne $P_n(t), n = 0, 1, \ldots$ ponieważ znamy warunki początkowe: $P_{-1}(t) = 0, P_0(0) = 1, P_n(0) = 0, n = 0, 1, \ldots$ oraz $P_0(t) + P_1(t) + P_2(t) + \ldots = 1$. Istnieją ogólne metody rozwiązywania takich równań, których nie będziemy tu omawiać. Rozwiążemy natomiast nasze równanie dla początkowych wartości n.

$n = 0$
$P_0'(t) + \lambda P_0(t) = 0$
$\frac{dP_0}{dt} = -\lambda P_0(t)$
$\frac{dP_0}{P_0} = -\lambda dt$
$lnP_0 = -\lambda t$
$P_0 = e^{-\lambda t}$
$P_0(t) = e^{-\lambda t}$

$n = 1$
$P_1'(t) = \lambda P_0(t) - \lambda P_1(t)$
$P_1'(t) + \lambda P_1(t) = \lambda e^{-\lambda t}$
mnożymy obustronnie przez *czynnik całkujący* $e^{\lambda t}$
$P_1'(t)e^{\lambda t} + \lambda P_1(t)e^{\lambda t} = \lambda e^{-\lambda t}e^{\lambda t}$
$P_1'(t)e^{\lambda t} + \lambda P_1(t)e^{\lambda t} = \lambda$
$(P_1(t)e^{\lambda t})' = \lambda$
$P_1(t)e^{\lambda t} = \lambda t$

$P_1(t) = \lambda t e^{-\lambda t}$
itd.

$$P_n(t) = \frac{(\lambda t)^n}{n!} e^{-\lambda t}, n = 0, 1, \ldots \tag{1.7}$$

Definicja 1.6. *Proces sygnałowy o przyrostach niezależnych nazywamy* procesem Poissona *jeżeli* $\forall t \in \langle 0; t)$ $X(t)$ *ma rozkład Poissona, tzn. istnieje taka funkcja* $\lambda(t)$, *że*

$$\forall n \in \mathcal{N} \;\; P[X(t) = n] = e^{-\lambda(t)} \frac{\lambda^n(t)}{n!}.$$

Jak wiadomo dla rozkładu Poissona $EX(t) = \lambda(t)$ stąd

$$E[X(t_2) - X(t_1)] = \lambda(t_2) - \lambda(t_1), t_2 > t_1$$

zatem

$$\forall t_2 > t_1 \forall n \in \mathcal{N} \;\; P[(X(t_2) - X(t_1)) = n] = \frac{[\lambda(t_2) - \lambda(t_1)]^n}{n!} e^{(-[\lambda(t_2) - \lambda(t_1)])}$$

W szczególności niech $\lambda(t) = \lambda t$, to mówimy o *stacjonarnym jednorodnym procesie Poissona*. Wówczas:

$$P[X(t) = n] = \frac{(\lambda t)^n}{n!} e^{(-\lambda t)}$$

$$P(X(t_2) - X(t_1) = n) = \frac{\lambda^n (t_2 - t_1)^n}{n!} e^{(-\lambda(t_2 - t_1))}$$

Wykazaliśmy następujące Twierdzenie 1.1

Twierdzenie 1.1. *Stacjonarny proces sygnałowy o przyrostach niezależnych, w którym zgłoszenia przybywają pojedynczo i błyskawicznie jest stacjonarnym procesem Poissona.*

Bezpośrednim wnioskiem z Twierdzenia 1.1 jest następujące Twierdzenie 1.2.

Twierdzenie 1.2. *Dla stacjonarnego procesu Poissona odstęp czasowy między kolejnymi zgłoszeniami ma rozkład wykładniczy.*

Dowód

Wyznaczmy dystrybuantę zmiennej losowej U oznaczającej odstęp czasowy między kolejnymi zgłoszeniami:

$$F_U(t) = P(U < t) = 1 - P(U \geq t) = 1 - P_0(t) = 1 - e^{-\lambda t} \quad \square$$

UWAGA 1: Można wykazać, że założenie błyskawiczności wynika z pozostałych trzech założeń Twierdzenia 1.1.

UWAGA 2: Ponieważ rozkład wykładniczy jest jedynym ciągłym rozkładem o własności braku pamięci, więc twierdzenie odwrotne do Twierdzenia 1.2 jest również prawdziwe.

Oczywiście nie zawsze czasy przybycia zgłoszeń mają rozkład wykładniczy. W takim przypadku jednak analiza systemu jest znacznie bardziej złożona. W przypadku rozkładu wykładniczego można natomiast wykazać wiele pożytecznych własności i wykorzystać je do analizy systemu.

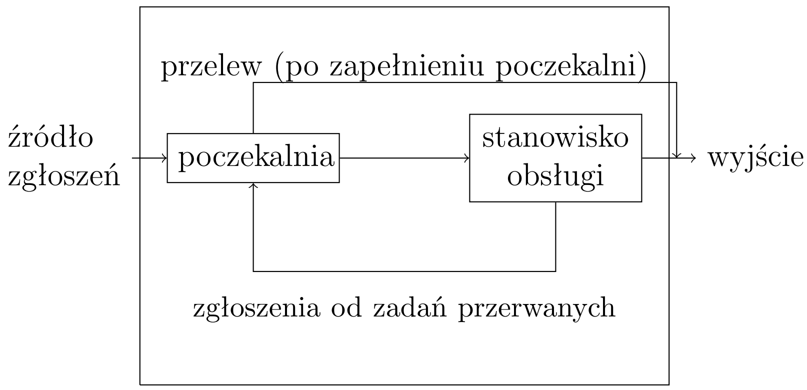
1.2. Charakterystyka systemu masowej obsługi

Analizą ilościową systemów, w których zgłoszenia napływają do punktów obsługi i oczekują na obsłużenie w kolejkach zajmuje się *teoria masowej obsługi* (teoria kolejek). Celem tej analizy jest ustalenie takiej organizacji systemu obsługi, aby czas oczekiwania w kolejce był jak najmniejszy. Badania w tej dziedzinie zapoczątkował duński inżynier pracujący w centrali telefonicznej A.K. Erlang ok. roku 1900 [?]. Agner Krarup Erlang (1.01.1878 – 3.02.1929) duński matematyk, statystyk i inżynier, twórca inżynierii ruchu i teorii kolejek.

Opracowane metody pozwalają ocenić jakość obsługi za pomocą takich parametrów, jak prawdopodobieństwo, że nowe zgłoszenie będzie musiało oczekiwać na obsługę, średni czas oczekiwania na obsługę, średnia długość kolejki itp. [?]. Takie informacje pozwalają menedżerom porównać różne warianty organizacyjne systemu czy ocenić opłacalność inwestowania w dodatkowe stanowiska obsługi.

Jednostki przybywające do systemu (zadania, samoloty, klientów itd.) nazywamy zwykle *zgłoszeniami.*

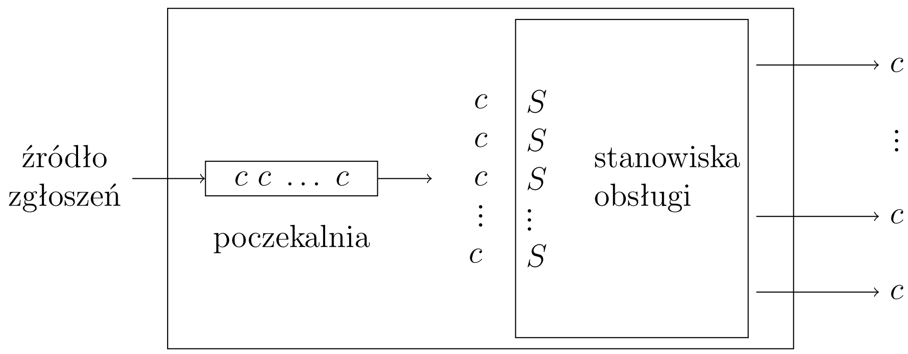
Najprostszy system obsługi przedstawiono na rysunku 1.2, gdzie występuje tylko jedno stanowisko obsługi. Zadania, które nie mieszczą się w poczekalni są tracone, a zadania przerwane wracają do kolejki.



Rysunek 1.2. Jednostanowiskowy system obsługi.

W praktyce często rozważa się bardziej złożone systemy składające się z połączonych szeregowo i równolegle systemów jednostanowiskowych tworzących w ogólności zamknięte lub otwarte sieci kolejkowe (*queuing networks*). System szeregowy nazywa się *kanałem obsługi*. Jeżeli w sieci krąży stałą liczba zadan, to sieć jest zamknięta, w przeciwnym razie jest otwarta.

Elementy systemu masowej obsługi przedstawiono na rysunku 1.3.



Rysunek 1.3. System kolejkowy: c - zgłoszenie, S - stanowisko obsługi.

System przedstawiony na rysunku 1.8 składa się ze źródła zgłoszeń, kolejki (em queue) oraz stanowisk obsługi.

W systemie kolejkowym nie tylko czasy przybycia zgłoszeń są zmiennymi losowymi, czasy obsługi są również niedeterministyczne i charakteryzuje się je przez podanie rozkładu zmiennej losowej V oznaczającej czas wykonania zadania.

Ponadto, nie zawsze nowe zgłoszenia mogą w nieoograniczonej ilości napływać do systemu. Mówimy, że źródło zgłoszeń ma wymiar k, gdy aktualna liczba zgłoszeń w systemie (w poczekalni i na stanowiskach obsługi) jest ograniczona przez k.

System obsługi definiuje się przez scharakteryzowanie:
— liczby stanowisk obsługi (m),
— wymiaru źródła zgłoszeń (k),
— pojemności poczekalni (L),
— rozkładu wejścia (zmiennej U),
— rozkładu czasu obsługi zadania (zmiennej V),
— regulaminu (algorytmu) obsługi.

Najprostszym (naturalnym) algorytmem obsługi jest algorytm FIFO (*First In First Out*), który zakłada, że zgłoszenia sa obsługiwane w kolejności wejścia do systemu (poczekalni). Dla systemów masowej obsługi z algorytmem naturalnym (FIFO), złożonych z identycznych równoległych stanowisk obsługi stosuje się często notację Kendalla, w której system jest scharakteryzowany za pomoca uporządkowanej piątki pól $X|Y|m|L|k$, gdzie:

— X oznacza rozkład zmiennej U,
— Y oznacza rozkład zmiennej V,
— m oznacza liczbę stanowisk obsługi,
— L oznacza pojemność poczekalni,
— k oznacza wymiar źródła zgłoszeń.

Jeżeli $L = k = \infty$, to notacja Kendalla ogranicza się do trzech pierwszych pól. Rozkłady zmiennych losowych oznacza się następującymi symbolami:

— D - rozkład jednopunktowy,
— M - rozkład wykładniczy,
— G - rozkład dowolny,
— itd.

UWAGA: Notacja Kendalla może być uogólniona do opisu większej klasy systemów, ale wymaga to dodatkowego komentarza.

Zdefiniujemy następujące terminy:

— *(stacjonarny) strumień prosty* – strumień zgłoszeń będący stacjonarnym procesem Poissona,

— *intensywność strumienia zgłoszeń* λ – oznacza średnią liczbę zgłoszeń przybywających do systemu w jednostce czasu, dla strumienia prostego $\lambda = lim_{\Delta t \to 0} P_1(\Delta t)(\Delta t)^{-1}$; łatwo zauważyć, że $\bar{U} = \frac{1}{\lambda} = EU$
— średnia liczba zgłoszeń obsłużonych w jednostce czasu μ ($\bar{V} = \mu^{-1}$),
— *intensywność ruchu* $\rho = \frac{\lambda}{\mu}$,
— *stan systemu w chwili t* – liczba zgłoszeń w systemie w chwili t ($N(t)$); jest to oczywiście proces losowy, który zastąpi oznaczenie $X(t)$; $P_1(t) = P[N(t) = n]$,
— *czas oczekiwania na obsługę* – zmienna losowa oznaczana przez W (*waiting*,
— *czas odpowiedzi systemu* – zmienna losowa oznaczana przez $T = V + W$ *total*.

W analizie systemów kolejkowych wyróżnia się tzw. analizę wartości chwilowych oraz analizę w stanie równowagi statystycznej. Będziemy się zajmować tylko tą drugą analizą. System znajduje się w stanie równowagi statystycznej (*statistical equilibrium, steady state*) jeżeli osiągnięte są następujące granice:

$$\lim_{t \to \infty} N(t) = N, N < \infty \tag{1.8}$$

$$\lim_{t \to \infty} P_1(t) = p_n, n = 0, 1, \ldots \tag{1.9}$$

gdzie p_n nazywamy *prawdopodobieństwem stacjonarnym.*

Można wykazać, że przy bardzo ogólnych założeniach (spełnionych przez wszystkie rozpatrywane przez nas systemy) warunek dostateczny osiągnięcia stanu równowagi statystycznej ma postać: $\rho < m$ (lub $\lambda < m\mu$). W szczególności, dla $m = 1$ mamy $\rho < 1$ (lub $\lambda < \mu$).

Systemy obsługi, dla których ten warunek jest spełnony nazywa się *stabilnymi.* W systemach tych interesują nas 3 zmienne losowe: N, W, T, a w szczególności ich wartości oczekiwane, między którymi zachodzą następujące związki:

$$\overline{T} = \overline{W} + \overline{V} = \overline{W} + \frac{1}{\mu} \tag{1.10}$$

$$\overline{N} = \lambda \overline{T} \tag{1.11}$$

Wzór 1.11 nazywa się wzorem Little'a. Ze wzorów wynika, że w systemie stabilnym wartości średnie wszystkich interesujących nas zmiennych losowych są skończone.

W systemach komputerowych ważne jest zapewnienie preferencji zadaniom krótkim, czyli o małych wartościach $V = t$. Można opisać taką preferencję definiując warunkowe rozkłady zmiennych losowych, których dystrybuanty mają następującą postać:

$$F_W(x|t) = P(W < x|V = t)$$

$$F_T(x|t) = P(T < x|V = t)$$

Między wartościami średnimi tych rozkładów zachodzi związek: $\overline{T}(t) = \overline{W}(t) + t$.

1.3. Systemy z algorytmem naturalnym (FIFO)

1.3.1. System $M|G|1$

System oznaczony w notacji Kendalla jako $M|G|1$ jest to system jednostanowiskowy z algorytmem FIFO, nieskończenie wymiarowym źródłem zgłoszeń, poczekalnią o nieograniczonej pojemności oraz wykładniczym rozkładem czasów przybycia do systemu (U) i dowolnym rozkładem czasu obsługi (V).

Założenie o wykładniczym rozkładzie czasów wejścia oznacza (twierdzenie 1.2), że na wejściu systemu mamy strumień prosty.

Algorytm FIFO (First In First Out), inaczej FCFS (First Come First Served) oznacza, że zadania są wykonywane w kolejności przybywania do systemu i bez przerywania.

Pokażemy, że wartości średnie wartości interesujących nas zmiennych losowych N, W, T możemy wyznaczyć dla dowolnego rozkładu zmiennej V znając jedynie jej wartość średnią $\frac{1}{\mu}$.

Średnia liczba zadań w systemie $\overline{N}$

Oznaczmy przez N_1, N_2 liczby zgłoszeń w systemie bezpośrednio po obsłużeniu kolejnych zgłoszeń z_1, z_2 oraz przez R liczbę zgłoszeń, które nadeszły w czasie obsługiwania zgłoszenia z_2. Łatwo zauważyć, że:

$$N_1 = 0 \to N_2 = R$$

$$N_1 \neq 0 \to N_2 = N_1 + R - 1$$

Wprowadźmy zmienną boolowską Γ:

$$\Gamma = \begin{cases} 1 & N_1 = 0 \\ 0 & wpp \end{cases} \tag{1.12}$$

Korzystając z tej zmiennej losowej możemy zapisać wzory:

$$N_2 = N_1 + R - 1 + \Gamma \tag{1.13}$$

Średnia obu stron wzoru 1.17:

$$\overline{N_2} = \overline{N_1} + \overline{R} - 1 + \overline{\Gamma} \tag{1.14}$$

Korzystając z założenia, że system jest w stanie równowagi statystycznej, czyli $N_2 = N_1$ wyznaczymy teraz wartość $\overline{R}$. Ponieważ wartość $\overline{R}$ zależy od zmiennej losowej V wyznaczymy najpierw $\overline{R}$ dla ustalonego $V = v$, a następnie uśrednimy po rozkładzie zmiennej losowej V.

$$\overline{R}(v) = \lambda v \tag{1.15}$$

$$\overline{R} = \int_0^\infty \lambda v dF(v) = \lambda \int_0^\infty v dF(v) = \lambda \overline{V} = \frac{\lambda}{\mu} = \rho \tag{1.16}$$

Wracając do wzoru:

$$\overline{\Gamma} = 1 - \overline{R} = 1 - \rho$$

Podnieśmy do kwadratu wzór 1.17:

$$N_2^2 = N_1^2 + R^2 + 1 + \Gamma^2 + 2N_1R - 2N_1 + 2N_1\Gamma - 2R + 2R\Gamma - 2\Gamma \tag{1.17}$$

Z definicji $2N_1\Gamma = 0$ oraz $\Gamma^2 = \Gamma$, zatem po uproszczeniu wartość średnia obu stron:

$$\overline{N_2}^2 = \overline{N_1}^2 + \overline{R}^2 + 1 + 2\overline{N_1R} - 2\overline{N_1} - 2\overline{R} + 2\overline{R\Gamma} - \overline{\Gamma} \tag{1.18}$$

W stanie równowagi statystycznej $\overline{N_1^2} = \overline{N_2^2}$. Ponadto zmienne losowe R i N_1 oraz R i Γ są niezależne. Otrzymujemy:

$$0 = \overline{R}^2 + 1 + 2\overline{N_1 R} - 2\overline{N_1} - 2\overline{R} + 2\overline{R\Gamma} - \overline{\Gamma} \tag{1.19}$$

Podstawiając wzory na $\overline{R}$ i $\overline{\Gamma}$ mamy:

$$0 = \overline{R^2} + 1 + 2\overline{N_1}\rho - 2\overline{N_1} - 2\rho + 2\rho(1-\rho) - (1-\rho) \tag{1.20}$$

$$2\overline{N_1}(1-\rho) = \overline{R^2} - \rho + 2\rho(1-\rho) \tag{1.21}$$

zatem:

$$\overline{N} = \overline{N_1} = \frac{\overline{R^2} - \rho}{2(1-\rho)} + \rho \tag{1.22}$$

Pozostaje zatem wyznaczenie $\overline{R^2}$, co zrobimy znowu dwuetapowo. Najpierw policzymy $\overline{R^2}$ dla ustalonego $V = v$. Ponieważ na wejściu mamy proces stacjonarny, prawdopodobieństwo, że w czasie v do systemu nadejdzie r zgłoszeń wyraża się wzorem:

$$\frac{(\lambda v)^r}{r!} e^{-\lambda v}$$

Zatem

$$\overline{R^2}(v) = \sum_{r=1}^{\infty} r^2 \frac{(\lambda v)^r}{r!} e^{-\lambda v} = e^{-\lambda v} \sum_{r=1}^{\infty} r^2 \frac{(\lambda v)^r}{r!} = \lambda v + \lambda^2 v^2 \tag{1.23}$$

Ostatnia równość wynika z następującej obserwacji:

$$\begin{aligned} & \sum_{r=1}^{\infty} r^2 \frac{(\lambda v)^r}{r!} \\ &= \sum_{r=2}^{\infty} r(r-1) \frac{(\lambda v)^r}{r!} + \sum_{r=1}^{\infty} r \frac{(\lambda v)^r}{r!} \\ &= \sum_{r=2}^{\infty} r(r-1) \frac{(\lambda v)^2}{r(r-1)} \frac{(\lambda v)^{r-2}}{(r-2)!} + \sum_{r=1}^{\infty} r \frac{\lambda v}{r} \frac{(\lambda v)^{r-1}}{(r-1)!} \\ &= (\lambda v)^2 \sum_{r=2}^{\infty} \frac{(\lambda v)^{r-2}}{(r-2)!} + \lambda v \sum_{r=1}^{\infty} \frac{(\lambda v)^{r-1}}{(r-1)!} \\ &= \lambda v e^{\lambda v} + \lambda v e^{\lambda v} \end{aligned} \tag{1.24}$$

Na drugim etapie uśrednimy wartość $\overline{R^2}$ względem rozkładu zmiennej losowej V.

$$\overline{R^2} = \int_0^\infty (\lambda v + \lambda^2 v^2) dF(v) = \lambda \overline{V} + \lambda^2 \overline{V^2} = \frac{\lambda}{\mu} + \lambda^2 (\sigma_V^2 + \frac{1}{\mu^2}) = \rho + \lambda^2 \sigma_V^2 + \rho^2 \quad (1.25)$$

Podstawiając do wzoru na $\overline{N}$ otrzymujemy:

$$\overline{N} = \frac{\rho + \lambda^2 \sigma_V^2 + \rho^2 - \rho}{2(1-\rho)} + \rho = \frac{\lambda^2 \sigma_V^2 + \rho^2}{2(1-\rho)} + \rho \quad (1.26)$$

Korzystając z tego wyniku łatwo wyliczymy pozostałe interesujące nas średnie. Ze wzoru Little'a 1.11 mamy:

$$\overline{N} = \lambda \overline{T} = \lambda(\overline{V} + \overline{W}) | : \lambda \overline{V} = \rho \quad (1.27)$$

$$\frac{\overline{W}}{\overline{V}} = \frac{\overline{\mu}}{\rho} - 1 = \frac{\rho}{2(1-\rho)} (1 + \mu^2 \sigma_V^2) \quad (1.28)$$

Powyższy wzór nazywamy wzorem Pollaczka.

Félix Pollaczek (1.12.1892 w Wiedniu – 29.04.1981 w Boulogne-Billancourt) inżynier i matematyk pracujący we Francji znany z wkładu w teorię liczb, analizę matematyczną i teorię prawdopodobieństwa. Najbardziej znany jest jako współautor wzoru Pollaczka–Chinczyna (1930) i wielomianów Pollaczka.

$$\overline{W} = \frac{\rho}{2\mu(1-\rho)} (1 + \mu^2 \sigma_V^2) = \frac{\lambda}{2\mu^2 (1-\rho)} (1 + \mu^2 \sigma_V^2) = \frac{\lambda \overline{V^2}}{2(1-\rho)} \quad (1.29)$$

Ponieważ $\sigma^2 = E(X^2) - (E(X))^2$ czyli $\frac{1}{\mu^2} + \sigma_V^2 = \overline{V^2}$.

Algorytm FIFO nie preferuje żadnych zadań, zatem $\overline{W}(t) = \overline{W}$ natomiast $\overline{T}(t) = \overline{V} + \overline{W} = \frac{1}{\mu} + \frac{\lambda \overline{V^2}}{2(\rho-1)}$ oraz $\overline{T}(t) = \overline{W}(t) + t$

1.3.2. System $M|D|1$

Rozkład zmiennej losowej oznaczającej czas obsługi zadania jest rozkładem jednopunktowym zatem:

$$P(V = v_0) = 1$$
$$\sigma_V^2 = 0$$

a więc:

$$\overline{W} = \frac{\lambda\overline{V^2}}{2(1-\rho)} = \frac{\lambda v_0 \frac{1}{\mu}}{2(1-\rho)} = \frac{v_0\rho}{2(1-\rho)}$$

1.3.3. System $M|M|1$

Rozkład zmiennej losowej oznaczającej czas obsługi zadania jest rozkładem wykładniczym zatem:

$$\overline{V} = \frac{1}{\mu}$$
$$\sigma_V^2 = \frac{1}{\mu^2}$$

a więc:

$$\overline{W} = \frac{\lambda\overline{V^2}}{2(1-\rho)} = \frac{\lambda(\frac{1}{\mu^2} + \sigma_V^2)}{2(1-\rho)} = \frac{\lambda(\frac{1}{\mu^2} + \frac{1}{\mu^2})}{2(1-\rho)} = \frac{\frac{2\lambda}{\mu^2}}{2(1-\rho)} = \frac{\rho}{\mu(1-\rho)}$$
$$\overline{T} = \frac{1}{\mu(1-\rho)}$$

Porównamy teraz systemy $M|D|1$ oraz $M|M|1$, zakładając, że w obu z nich $\overline{V}$ jest taki sam.

Podzielimy obustronnie wzory Pollaczka dla tych rozkładów ($\overline{V} = v_0 = \frac{1}{\mu}$):

$$\frac{\overline{W_{M|M|1}}}{\overline{W_{M|D|1}}} = \frac{\rho}{\mu(1-\rho)}\frac{2(1-\rho)}{v_0\rho} = 2$$

Zatem średni czas oczekiwania na obsługę w systemie ze stałym czasem obsługi v_0 jest dwukrotnie krótszy niż w systemie z wykładniczym czasem obsługi o wartości średniej $\overline{V} = v_0$!

Rysunek 1.4. Porównanie systemów $D|M|1$ i $M|M|1$ dla $\lambda = 1$

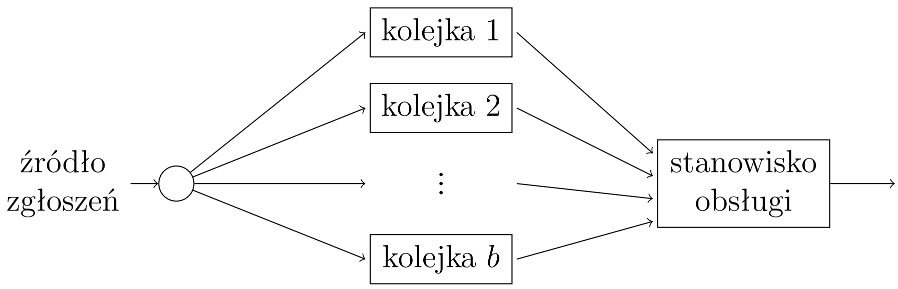
1.4. Algorytmy

1.4.1. Algorytmy ze stałymi priorytetami

Omówimy teraz algorytmy, które pozwalają wyróżnić uprzywilejowane zgłoszenia i obsłużyć je z wyższym priorytetem. W systemach komputerowych dotyczy to głównie zadań krótkich.

W algorytmach ze stałymi priorytetami przyporządkowuje się każdemu zadaniu przybywającemu do systemu priorytet, który nie ulega zmianie do czasu opuszczenia systemu przez zadanie. Zazwyczaj tworzy się odrębne kolejki tak, aby zadania w danej kolejce miały ten sam priorytet.

Zasady nadawania priorytetów mogą być różne, np.:
— Priorytet nadawany przez programistę – najczęściej programiści nadają swoim zadaniom najwyższy priorytet. Użytkownik może obniżyć priorytet swojego zgłoszenia tylko, gdy towarzyszy temu możliwość obniżenia kosztów obsługi.
— Priorytet nadawany przez operatora – dotyczy systemów wsadowych.
— O priorytecie decyduje rozmiar (liczba instrukcji) zadania.
— Priorytet określa się na podstawie historii obsługi zadania – sposób ten może być wykorzystany, gdy zadania wykonywane są wielokrotnie.



Rysunek 1.5. System kolejkowy z algorytmem priorytetowym.

Założymy, że każdemu priorytetowi $\pi_i, i = 1, \ldots, b$, odpowiada jednorodny strumień poissonowski przybywających zadań o intensywności λ_i, a rozkłady czasów obsługi poszczególnych strumieni są niezależne, o wartościach średnich μ_i. Ponadto oznaczymy $\rho_i = \frac{\lambda_i}{\mu_i}, i = 1, \ldots, b, \lambda = \sum_{i=1}^{b} \lambda_i, \rho = \sum_{i=1}^{b} \rho_i$.

W literaturze rozważa się dwa warianty algorytmu ze stałym priorytetem. W pierwszym wariancie zadanie o priorytecie π_i dołącza się na koniec kolejki z tym priorytetem, przed jakimkolwiek zadaniem o niższym priorytecie. Przybycie zadania o wyższym priorytecie od aktualnie wykonywanego nie powoduje przerwania obsługi. W tym wariancie średni czas oczekiwania zadania o priorytecie π_i nie zależy od wartości czasu obsługi, a tylko od przyznanego priorytetu, czyli $\overline{W_i}(t) = \overline{W_i}$, który można obliczyć jako:

$$\overline{W_i} = \frac{\overline{W_o}}{(1 - \sigma_i)(1 - \sigma_{i+1})}, i = 1, \ldots, b \tag{1.30}$$

gdzie:

$$\sigma_i = \sum_{j=i}^{b} \rho_j \tag{1.31}$$

$$\overline{W_o} = \sum_{i=1}^{b} \frac{\lambda_i \overline{V_i^2}}{2} \tag{1.32}$$

jest średnim czasem oczekiwania (opóźnieniem) zdania, spowodowanym przez zadanie aktualnie obsługiwane, $\overline{V_i^2}$ jest momentem zwykłym drugiego rzędu rozkładu czasu obsługi zadań o priorytecie π_i.

Na podstawie wzoru 1.30 można ocenić jaki wpływ na średni czas oczekiwania $\overline{W_i}$ maja zadania o priorytecie π_i i wyższym, obecne w kolejce w momencie przybycia zadania o priorytecie π_i. Wpływ ten wyraża czynnik $(1 - \sigma_i)$ w mianowniku.

Można także ocenić wpływ zadań o priorytetach wyższych, przybywających do systemu w trakcie oczekiwaia w kolejce zadania o priorytecie π_i, który wyraża czynnik $(1 - \sigma_{i+1})$. Zauważmy, że zadaia o priorytetach niższych od π_i, tj. $\pi_1, \pi_2, \ldots, \pi_{i-1}$ wpływają na sredni czas oczekiwania $\overline{W_i}$ jedynie za pośrednictwem licznika $(\overline{W_o})$ wyrażenia 1.30. Zatem ze wzoru tego wynika, że dla zadań o wysokich priorytetach średni czas oczekiwania może być skończony, podczas gdy dla innych grup zadań o niskich priorytetach czas ten rośnie do nieskończoności. Ponadto średni czas oczekiwania da zadań o priorytecie π_i jest skończony, gdy $\rho_i + \rho_{i+1} + \ldots + \rho_b < 1$. Średni czas oczekiwania silnie zależy od priorytetu, a nie zależy od wartości czasu obsługi zadania.

Druga wersja algorytmu ze stałymi priorytetami ma charakter algorytmu cyklicznego i będzie omówiona w kolejnym punkcie.
++
Wykresy - ew.

1.4.2. Algorytmy cykliczne

Dla systemów cylkicznych charakterystyczne jest przerwania zadania. W systemach komputerowych przerwanie następuje najczęściej po żądaniu urządzenia wejścia-wyjścia – system z takim algorytmem preferuje zadania, które rzadko korzystają z tych urządzeń. Procesor przechodzi zatem do wykonywania następnego zadania, gdy obsługiwane zadanie zakończy sie lub zażąda urządzenia wejścia-wyjścia. Wzory analityczne są w tym przypadku trudne do uzyskania.

Częściej stosuje się drugi typ algorytmów cyklicznych – bazujący na technice podziału czasu, zwany RR (*Round Robin*), który przez niektórych autorów jest utożsamiany z algorytmem cyklicznym. W tym systemie przerwanie następuje po kwancie czasu Δ_i, który w systemach komputerowcyh jest parametrem systemu operacyjnego. W analizie teoretycznej przyjmujemy, że długość kwantu dąży do zera, co pozwala zaniedbać inne zdarzenia, które zbiegają się z zakończeniem kwantu. Można udowodnić, że średni czas odpowiedzi nie zależy od rozkładu czasu obsługi. Średnie czasy odpowiedzi i oczekiwania dla dowolnego rozkładu czasu obsługi wynoszą odpowiednio:

$$\overline{T}(t) = \frac{t}{1-\rho} \tag{1.33}$$

$$\overline{W}(t) = \frac{\rho t}{1-\rho} \tag{1.34}$$

Jak widać system z tym algorytmem zapewnia liniową preferencję zadań w funkcji długości, tzn. zadania n-krotnie dłuższe spędzi średnio n-razy więcej czasu w kolejce (systemie).

Zauważmy, że bezwarunkowe średnie czasy odpowiedzi i oczekiwania wynoszą:

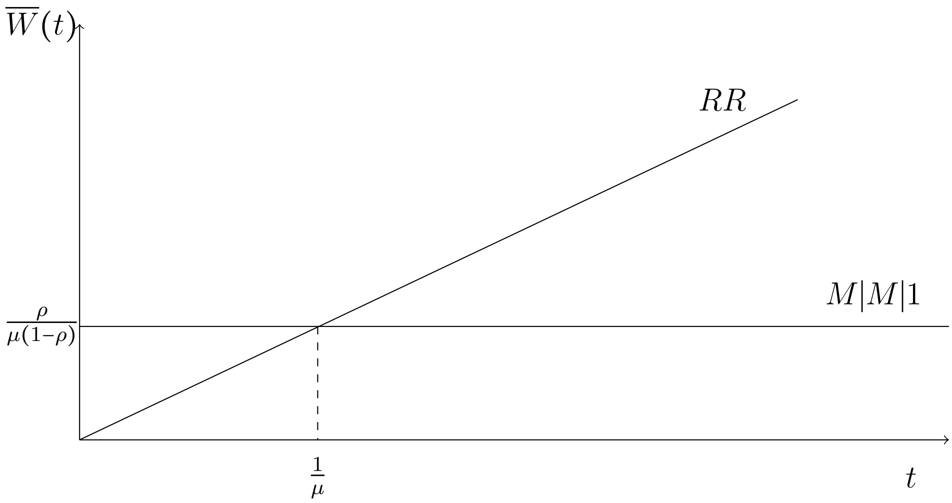
$$\overline{T} = \frac{1}{\mu(1-\rho)} \tag{1.35}$$

$$\overline{W} = \frac{\rho}{\mu(1-\rho)} \tag{1.36}$$

Powyższe średnie są identyczne, jak w systemie $M|M|1$.
Ciekawe jest porównanie tego algorytmu z algorytmem FIFO.

$$\overline{T}_{FIFO}(t) = \frac{\rho}{\mu(1-\rho)} + t = \frac{t}{1-\rho} = \overline{T}_{RR}(t) \tag{1.37}$$

Rozwiązaniem powyższego równiania jest $t = \frac{1}{\mu}$. Zatem algorytm RR preferuje zadania krótkie a dyskryminuje zadania długie, co widać na wykresach.



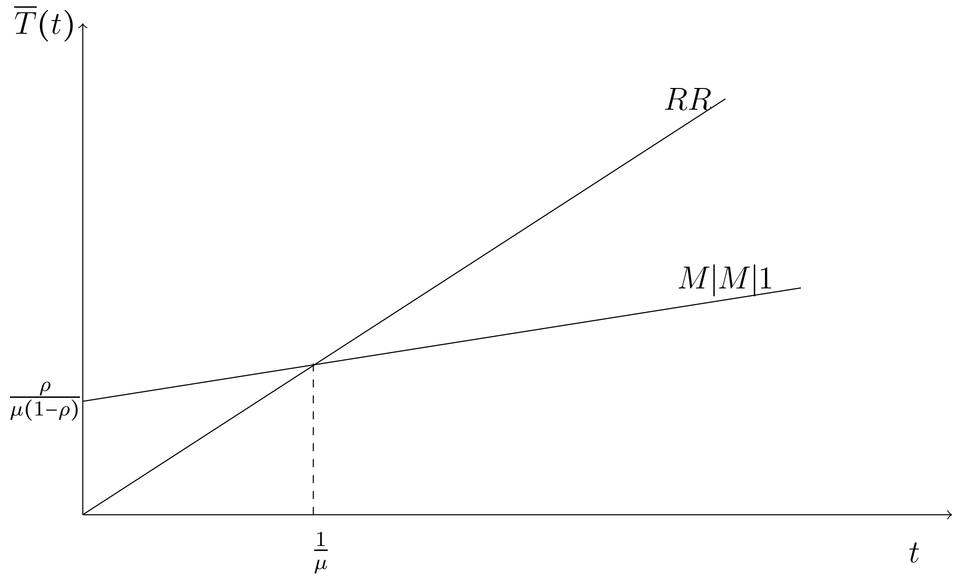
Rysunek 1.6. Porównanie $\overline{W}$ w systemach z algorytmami FIFO i RR

Widzimy zatem, że system z algorytmem RR preferuje zadania o długości mniejszej niż $\frac{1}{\mu}$.

Załóżmy teraz, że kolejka jest podzielona na podkolejki o priorytetach $\pi_i, i = 1, \ldots, b$, i jest obsługiwana cyklicznie przy czym kwant czasu w podkolejce o priorytecie π_i wyraża się wzorem:

$$\Delta_i = \frac{c_i}{\sum_{j=1}^{b} c_j \overline{N}_j} \tag{1.38}$$

gdzie c_j jest stałą dodatnią (tym wiekszą im wyższy jest priorytet π_i) a $\overline{N}_j$ jest średnią liczbą zadań o priorytecie π_j w systemie.



Rysunek 1.7. Porównanie $\overline{T}$ w systemach z algorytmami FIFO i RR

Można wykazać, że w systemie z takim algorytmem słuszne są następujące wzory:

$$\overline{W}_i(t) = \frac{\rho t}{1-\rho}\left[1 + \sum_{j=1}^{b}\left(\frac{c_j}{c_i} - 1\right)\frac{\rho_j}{\rho}\right], i = 1, \ldots, b \tag{1.39}$$

$$\overline{T}_i(t) = \frac{t}{1-\rho}\left[1 + \sum_{j=1}^{b}\left(\frac{c_j}{c_i} - 1\right)\frac{\rho_j}{\rho}\right], i = 1, \ldots, b \tag{1.40}$$

1.4.3. Algorytm LIFO

Algorytm LIFO (LCFS) oznacza obsługę w kolejności odwrotnej do kolejności zgłoszeń. Zadanie wykonywane jest wywłaszczane, jeżeli w systemie pojawi się zadanie o wyższym priorytecie. Kolejka ma postać stosu. Można wykazać, że średnie czasy odpowiedzi i oczekiwania sa takie same, jak w algorytmie RR, ale wariancje są większe.

UWAGA Zauważmy, że w dotychczas rozpatrywanych algorytmach preferencja zadań w funkcji długości (jeśli istniałą) była liniowa. Nie zawsze pozwala to uzyskać maksymalne wykorzystanie mocy obliczeniowych systemu. Obszerną klasę algorytmów zapewniających nieliniową zależność stanowią algorytmy wielokolejkowe (*multilevel processor sharing scheduling algorithms*). Zadanie przybywające do systemujest umieszczane w kolejce o najwyższym priorytecie, a po upływie tego czasu spada do następnej kolejki (jeżeli się nie zakończy). Często stosowany jest algorytm dwukolejkowy (*foreground-background*), w którym pierwsza kolejka jest obsługiwana algorytmem FIFO, a druga RR. Można wykazać, że otrzymujemy wówczas typ preferencji opisany poniższą krzywą S.

Zauważmy, że w systemach z algorytmami wielokolejkowymi zadania długie są silnie dyskryminowane i nie zawsze wystarcza fakt, że kwant czasu w tych systemach rośnie wraz z malejącym priorytetem. Żeby zapobiec temu stosuje się pewne dodatkowe mechanizmy, np. generalną amnestię. Inna metoda, nazywana metodą łaski polega na tym, że dla każdego zadania liczymy czas czekania w określonej kolejce i jeśli przekracza on założoną wartość, to przesuwamy zadanie do kolejki o priorytecie o jeden wyższym.

1.5. Systemy masowej obsługi z momentami przybycia lub czasami obsługi zależnymi od stanu systemu

Zacznijmy od wyprowadzenia wzrou na prawdopodobieństwo stacjonarne dla systemu $M|M|1$.

Załóżmy, że system jest w stanie $N(t) > 0$ i obliczmy prawdopodobieństwo przejścia $N(t+\Delta t) \to N(t)$ dla $\Delta t \to 0$ przy założeniu, że prawdopodobieństwo przybycia zadania dla $\Delta t \to 0$ wynosi $\lambda\Delta t + o(\Delta t)$ oraz, że prawdopodobieństwo opuszczenia systemu przez zadanie w tym przedziale wynosi $\mu\Delta t + o(\Delta t)$. Należy rozważyć następujące sytuacje:

a) $N(t+\Delta t) = N(t)$ – w przedziale Δt 1 zadanie przybyło do systemu i 1 zadanie opuściło system lub 0 zadań przybyło do systemu i 0 zadań opuściło system.
b) $N(t+\Delta t) = N(t)+1$ – w przedziale Δt 1 zadanie przybyło do systemu i 0 zadań opuściło system,
c) $N(t+\Delta t) = N(t)-1$ – w przedziale Δt 0 zadań przybyło do systemu i 1 zadanie opuściło system,

Obliczamy prawdopodobieństwa poszczególnych zdarzeń:

$$\begin{aligned} P[N(t+\Delta t) = N(t)] = \\ [\lambda\Delta t + o(\Delta t)][\mu\Delta t - o(\Delta t)] + [1-\lambda\Delta t + o(\Delta t)][1-\mu\Delta t - o(\Delta t)] = \\ 1-\lambda\Delta t - \mu\Delta t + o(\Delta t) \end{aligned} \tag{1.41}$$

$$\begin{aligned} P[N(t+\Delta t) = N(t)+1] = \\ [\lambda\Delta t + o(\Delta t)][1-\mu\Delta t + o(\Delta t)] = \\ \lambda\Delta t + o(\Delta t) \end{aligned} \tag{1.42}$$

$$\begin{aligned} P[N(t+\Delta t) = N(t)-1] = \\ [1-\lambda\Delta t + o(\Delta t)][\mu\Delta t - o(\Delta t)] = \\ \mu\Delta t + o(\Delta t) \end{aligned} \tag{1.43}$$

Zatem z uwagi na brak pamięci i rozkład wykładniczy na wejściu i na wyjściu systemu mamy:

$$\begin{aligned} P_n(t+\Delta t) = \\ P_{n-1}(t)[\lambda\Delta t + o(\Delta t)] + \\ P_n(t)[1-\lambda\Delta t - \mu\Delta t + o(\Delta t)] + \\ P_{n+1}(t)[\mu\Delta t + o(\Delta t)] = \\ \lambda\Delta t P_{n-1}(t) + P_n(t) - (\lambda+\mu)\Delta t P_n(t) + \mu\Delta t P_{n+1}(t) + o(\Delta t) \end{aligned} \tag{1.44}$$

$$\frac{P_n(t+\Delta t) - P_n(t)}{\Delta t} = \lambda P_{n-1}(t) - (\lambda+\mu)P_n(t) + \mu P_{n+1}(t) + \frac{o(\Delta t)}{\Delta t} \tag{1.45}$$

Przechodząc do granicy przy $\Delta t \to 0$:

$$P_n'(t) = \lambda P_{n-1}(t) - (\lambda+\mu)P_n(t) + \mu P_{n+1}(t), n = 1, 2, \ldots \tag{1.46}$$

Dla $n = 0$ musimy ominąć możliwość opuszczenia systemu w drugim składniku, a zatem: $P_0'(t) = \mu P_1(t) - \lambda P_0(t)$.

Zauważmy, że dla $\mu = 0$ (tylko przybywanie) otrzymamy to samo, co dla $M|G|1$.

Załóżmy teraz, że w systemie został osiągnięty stan równowagi statystycznej, czyli $P_n(t) = p_n, n = 1, 2, \ldots$ (oczywiście $P_n' = 0$). Mamy zatem:

$$\begin{aligned} \lambda p_{n-1} - (\lambda + \mu) p_n + \mu p_{n+1} = 0, n = 1, 2, \ldots \\ \mu p_1 - \lambda p_0 = 0 \end{aligned} \tag{1.47}$$

Załóżmy teraz, że parametry systemu λ i μ zależą od stanu systemu, tzn. λ_0 oznacza intensywność strumienia zgłoszeń, gdy w systemie nie ma zadań, λ_1, gdy w systemie jest jedno zadanie itd. Analogicznie definiujemy intensywność obsługi zależną od stanu systemu. Rozwiążemy powyższy układ równań.

$$p_1 = \frac{\lambda_0}{\mu_1} p_0 \tag{1.48}$$

Podstawiając do pierwszego:

$$p_2 = \frac{\lambda_0}{\mu_1} \frac{\lambda_1}{\mu_2} p_0 \tag{1.49}$$

Postępując analogicznie można pokazać, że

$$p_n = p_0 \frac{\lambda_0 \lambda_1 \ldots \lambda_{n-1}}{\mu_1 \mu_2 \ldots \mu_n} \tag{1.50}$$

Dodatkowe równanie otrzymujemy ze wzoru:

$$\sum_{n=0}^{\infty} p_n = 1 = p_0 + \sum_{n=1}^{\infty} p_n \tag{1.51}$$

$$p_0 = \left[1 + \sum_{n=1}^{\infty} \prod_{i=1}^{n} \frac{\lambda_{i-1}}{\mu_i} \right]^{-1} \tag{1.52}$$

Wyznaczyliśmy zatem wszystkie prawdopodobieństwa stacjonarne stanu naszego systemu w stanie równowagi statystycznej. Na tej podstawie możemy łatwo wyznaczyć wszystkie parametry tego rozkładu, a więc w szczególności:

$$\overline{N} = \sum_{n=0}^{\infty} n p_n \tag{1.53}$$

a ze wzrou Little'a możemy wyznaczyć:

$$\overline{T} = \frac{\overline{N}}{\lambda} \tag{1.54}$$

UWAGA

Powyższe rozwiązanie jest słuszne dla dowolnych poissonowskich (na wejściu i wyjściu) systemów obsługi z algorytmami obsługi niezależnymi od czasu wykonywania zadania, a zatem dla wszstkich dotąd rozważanych algorytmów z wyjątkiem algorytmu ze stałymi priorytetami, w których priorytety zależą od czasu obsługi zadania.

Zauważmy ponadto, że otrzymane równania na prawdopodobieństwo stacjonarne wyprowadzone dla ogólnego przypadku, gdy momenty przybycia zadań i/lub czasy obsługi zależą od stanu n, dla stałego λ i μ przyjmują postać:

$$p_0 = \frac{1}{1 + \sum_{n=1}^{\infty} \rho^n} = \frac{1}{1 + \frac{\rho}{1-\rho}} = 1 - \rho \tag{1.55}$$

$$p_n = p_0 \rho^n = \rho^n (1 - \rho), n = 1, \ldots \tag{1.56}$$

1.6. Systemy ze skończoną poczekalnią

Załóżmy teraz, że $L < \infty$. Mamy wtedy:

$$\lambda_n = \begin{cases} \lambda & n = 0, \ldots, L \\ 0 & n > L \end{cases} \tag{1.57}$$

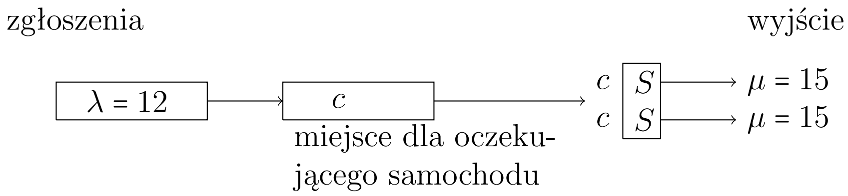
Załóżmy, że $\mu_n = \mu$ dla każdego $n \geq 1$. Wtedy wystarczy podstawić do wzorów ogólnych i wyliczyć charakterystyki.

1.7. System $M|M|m$

W tej sytuacji zwróćmy uwagę na czas obsługi.

$$\mu_n = \begin{cases} n\mu & 1 \leq n \leq m \\ m\mu & n \geq m \end{cases} \tag{1.58}$$

Przykład 1.2. *($M|M|m|L = 3$) Załóżmy, że samochody przybywają do myjni samoobsługowej z częstotliwością 12 na godzinę. Myjnia jest wyposażona w* *dwa stanowiska z myjką ciśnieniową, każde o wydajności 15 samochodów na godzinę. Przed stanowiskiem może czekać co najwyżej jeden samochód (poza aktualnie obsługiwanymi), a zatem system może być w jednym z 4 stanów:* $n = 0, 1, 2, 3$. *Zauważmy, że* $\lambda_n = 0$ *dla* $ngeq3$, *oraz* $\mu_n = 0$ *dla* $n \geq 4$.



Rysunek 1.8. System obsługi z ograniczoną kolejka i dwoma stanowiskami obsługi

Obliczmy:

$$p_0 = \left[1 + \frac{\lambda_0}{\mu_1} + \frac{\lambda_0\lambda_1}{\mu_1\mu_2} + \frac{\lambda_0\lambda_1\lambda_2}{\mu_1\mu_2\mu_3}\right]^{-1} \tag{1.59}$$

$$p_0 = \left[1 + \frac{12}{15} + \frac{12 \cdot 12}{15 \cdot 30} + \frac{12 \cdot 12 \cdot 12}{15 \cdot 30 \cdot 30}\right]^{-1} = \frac{125}{281} = 0.44 \tag{1.60}$$

$$p_1 = \frac{12}{15} p_0 = \frac{100}{281} \tag{1.61}$$

$$p_2 = \frac{12 \cdot 12}{15 \cdot 30} p_0 = \frac{40}{281} \tag{1.62}$$

$$p_3 = \frac{12 \cdot 12 \cdot 12}{15 \cdot 30 \cdot 30} p_0 = \frac{16}{281} \tag{1.63}$$

Oczekiwaną liczbę samochodów w myjni N można obliczyć jako:

$$N = 1p_1 + 2p_2 + 3p_3 = \frac{100 + 2 \cdot 40 + 3 \cdot 16}{281} = \frac{228}{281} \approx 0.81 \tag{1.64}$$

Średni czas odpowiedzi można wyliczyć ze wzoru: $N = \overline{\lambda} T$, gdzie

$$\overline{\lambda} = \sum_{n=0}^{\infty} \lambda_n p_n$$

$$\overline{\lambda} = 12(p_0 + p_1 + p_2) = 12\frac{265}{281} = \frac{3180}{281} \approx 11.3$$

Zatem $T = \frac{281}{3180}\frac{228}{281} = \frac{228}{3180} \approx 0.072$ godz., czlyli ok. 4.3 min.
Natomiast średni czas oczekiwania: $W = \frac{1}{\lambda}L = \frac{281}{3180}\frac{16}{281} = 0.3$ min $(L = 1 \cdot p_3)$.

1.8. Systemy typu $M|M|1|\infty|k$

Rozważmy system komputerowy w którym 1 stanowisko obsługi obługuje k końcowych użytkowników w trybie interaktywnym. Z każdej końcówki może przyjść zgłoszenie, które po obsłużeniu jest odsyłane do końcówki, która po "czasie zastanowienia się" może zgłosić ponowne zadanie. Przyjmując średni czas zastanawiania się $1\ /\lambda$ otrzymamy przykład naszego systemu.

Przykład 1.3. *W systemie są dwie końcówki generujące zadania dla procesora, obie z rozkładem normalnym o średniej $\lambda = 13$ zgłoszenia na godzinę. Obsługa zgłoszenia trwa średnio 30 minut i ma rozkład normalny.*

$$\lambda_n = \begin{cases} \frac{2}{3} & n = 0 \\ \frac{1}{3} & n = 1 \\ 0 & n = 2 \end{cases} \tag{1.65}$$

Oczywiście $\mu_1 = \mu_2 = 2$ (zgłoszenia na godzinę).
Mamy zatem

$$p_0 = \frac{1}{1 + \frac{\lambda_0}{\mu_1} + \frac{\lambda_0 \cdot \lambda_1}{\mu_1 \cdot \mu_2}}$$

$$p_0 = \left[1 + \frac{1}{3} + \frac{1 \cdot 1}{3 \cdot 6}\right]^{-1} = \frac{12}{85}$$

$$p_1 = \frac{1}{3}p_0 = \frac{6}{25}$$

$$p_2 = \frac{1}{18}p_0 = \frac{1}{25}$$

W tym przykładzie:

$$N = 1p_1 + 2p_2 = \frac{8}{25}$$

zgłoszeń.

$$\overline{\lambda} = p_0\lambda_0 + p_1\lambda_1 = \frac{14}{25}$$

$$T = \frac{1}{\lambda} N = \frac{4}{7}$$

godziny.

$$W = \frac{p_2}{\lambda} = \frac{25 \cdot 1}{14 \cdot 25} = \frac{1}{14}$$