Teoria masowej obsługi

Badania operacyjne

prof. dr hab. inż. Joanna Józefowska

Instytut Informatyki

Poznań, 2016/2017

Notatki 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.

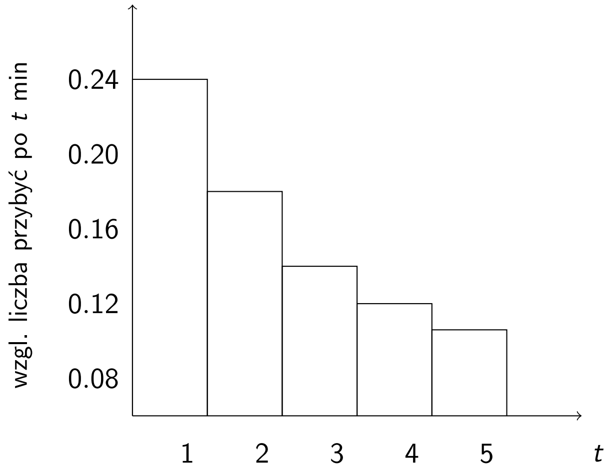
Analiza procesu przybywania klientów

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.

odstęp	liczba klientów	względna liczba klientów
1 minuta	72	24
2 minuty	54	18
3 minuty	42	14
4 minuty	36	12
5 minut	30	10
$>$ 5 minut	54	18

Histogram czasów przybycia do systemu



Rozkład wykładniczy

funkcja gęstości prawdopodobieństwa

$$f(t) = \lambda e^{-\lambda t}, \text{ dla } t \geqslant 0$$

dystrybuanta

$$F_U(t) = P(U \leqslant t) = 1 - e^{-\lambda t}$$

wartość oczekiwana

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

Liczba klientów przybywających w przedziale Δt

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

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

Stan procesu w chwili $t + \Delta t$

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

Układ równań różnicowo-różniczkowych

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

warunki początkowe:

$$P_{-1}(t) = 0, P_0(0) = 1, P_n(0) = 0, n = 0, 1, \ldots$$

oraz

$$P_0(t) + P_1(t) + P_2(t) + \ldots = 1$$

Rozwiązanie

$$P_n(t) = \frac{(\lambda t)^n}{n!} e^{-\lambda t}, n = 1, 2, \ldots$$

Rozkład Poissona

funkcja gęstości prawdopodobieństwa

$$P(X(t) = n) = P_n(t) = \frac{(\lambda t)^n e^{-\lambda t}}{n!}, \text{ dla } n = 0, 1, \ldots$$

wartość oczekiwana

$$E(X(t)) = \lambda t$$

WAŻNE!

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

Wniosek
Dla stacjonarnego procesu Poissona odstęp czasowy między kolejnymi zgłoszeniami ma rozkład wykładniczy.

Proces Markova

Proces Markova jest to proces stochastyczny, dla którego

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

$$= P(X_n = k_n | X_{n-1} = k_{n-1})$$

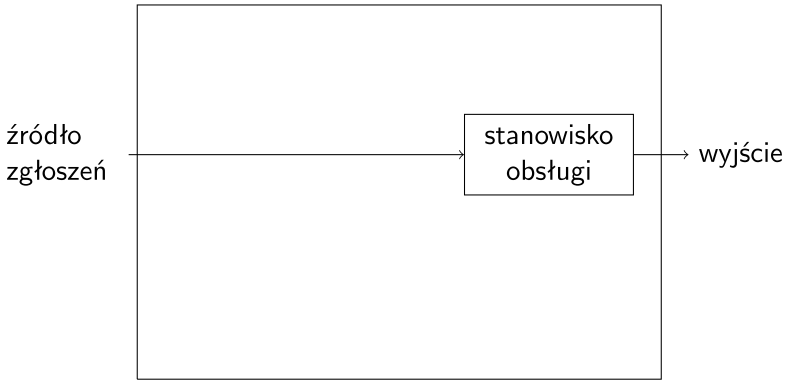
czyli proces bez pamięci.

Niezależne zmienne losowe

Mówimy, że zmienne losowe są niezależne, gdy dla każdych liczb rzeczywistych A, B zachodzi równość:

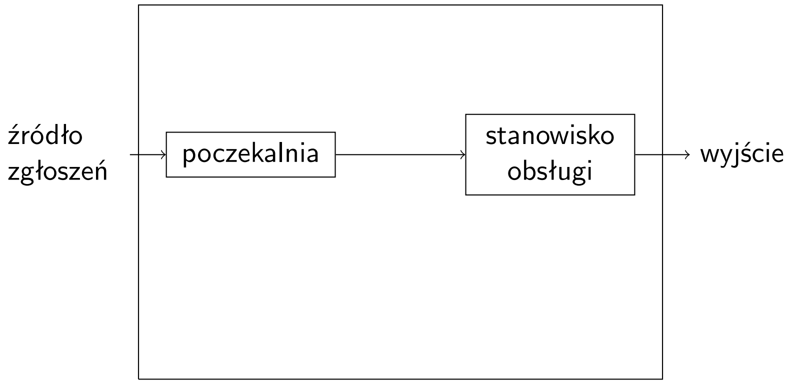
$$P(X \leqslant A)P(Y \leqslant B) = P(X \leqslant A \wedge Y \leqslant B).$$

Jednostanowiskowy system obsługi



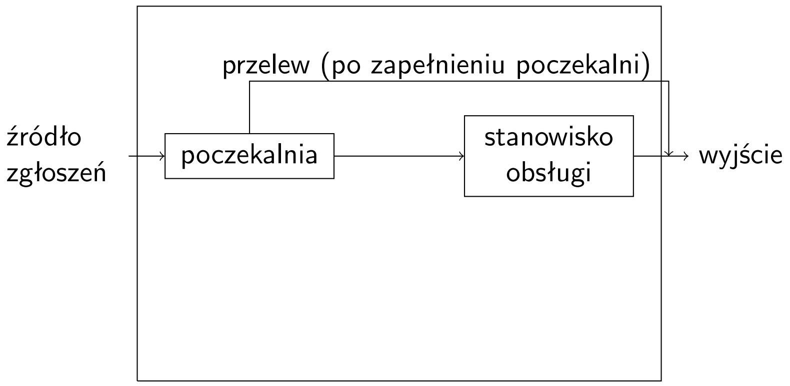
Rysunek: Jednostanowiskowy system obsługi.

Jednostanowiskowy system obsługi



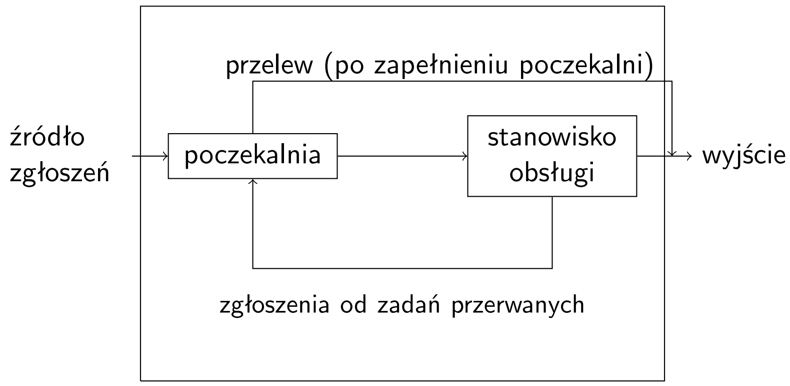
Rysunek: Jednostanowiskowy system obsługi.

Jednostanowiskowy system obsługi



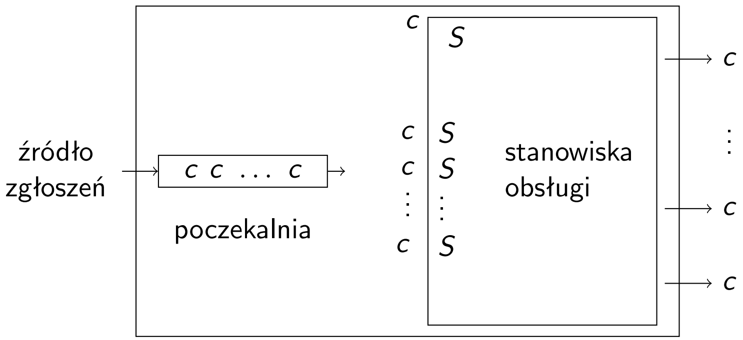
Rysunek: Jednostanowiskowy system obsługi.

Jednostanowiskowy system obsługi



Rysunek: Jednostanowiskowy system obsługi.

System kolejkowy



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

Charakterystyka systemu kolejkowego

- rozkład wejścia (zmienna losowa U o średniej $\frac{1}{\lambda}$),
- rozkład czasu obsługi zadania (zmienna losowa V o średniej $\frac{1}{\mu}$),
- liczba stanowisk obsługi (m),
- wymiar źródła zgłoszeń (k),
- pojemność poczekalni (L),
- regulamin (algorytm) obsługi.

intensywność ruchu

$$\rho = \frac{\lambda}{\mu}$$

Notacja Kendalla

algorytm FIFO (naturalny)

$$X|Y|m|L|k$$

- X – rozkład wejścia (zmiennej U),
  - D - rozkład jednopunktowy,
  - M - rozkład wykładniczy,
  - G - rozkład dowolny,
  - E_k - rozkład Erlanga rzędu k,...
- Y – rozkład czasu obsługi zadania (zmiennej V),
- m – liczba stanowisk obsługi,
- L – pojemność poczekalni,
- k – wymiar źródła zgłoszeń – maksmalna liczba zgłoszeń przebywających równocześnie w systemie.

Podstawowe pojęcia

- *stacjonarny strumień prosty* – strumień zgłoszeń będący stacjonarnym procesem Poissona;
- *intensywność strumienia zgłoszeń* λ – średnia liczba 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

$$\overline{U} = \frac{1}{\lambda} = EU$$

Stan równowagi statystycznej

zmienne losowe:

- $X(t)$ – liczba klientów, którzy przybyli w przedziale $[0, t)$,
- $N(t)$ – liczba zadań w systemie w chwili t (*stan systemu w chwili* t) zatem $P_n(t) = P(N(t) = n)$

statistical equilibrium, steady state, stan ustalony

$$\lim_{t \to \infty} N(t) = N$$

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

$$p_n, n = 0, 1, \ldots$$

nazywamy prawdopodobieństwem stacjonarnym.

Stan równowagi statystycznej

warunek konieczny i dostateczny osiągnięcia stanu równowagi statystycznej

$$\rho < m$$

$$\lambda < m\mu$$

dla $m = 1$

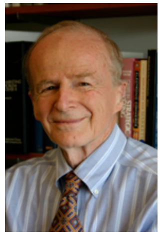
$$\lambda < \mu$$

System, dla którego warunek równowagi jest spełniony nazywamy **systemem stabilnym**.

Analiza systemu kolejkowego

- czas oczekiwania na obsługę - zmienna losowa W,
- czas odpowiedzi systemu - zmienna losowa $T = W + V$,
- $\overline{T} = \overline{W} + \overline{V} = \overline{W} + \frac{1}{\mu}$,
- $\overline{N} = \lambda \overline{T}$ (wzór Little'a)

John Little (1928)



Professor of Marketing, MIT Sloan School of Management
Research: in operations research methodology, traffic signal control, decision support systems, and especially marketing. Member of National Academy of Engineering, received the Parlin and Converse Awards of the American Marketing Association, Kimball Medal of the Institute for Operations Research and the Management Sciences (INFORMS), Fellow of INFORMS and the American Association for the Advancement of Science.

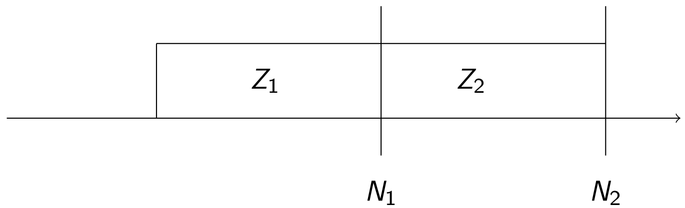
In a 1954 paper Little law was assumed true and used without proof. The form $L = \lambda W$ was first published by Philip M. Morse where he challenged readers to find a situation where the relationship did not hold. Little published in 1961 his proof of the law, showing that no such situation existed. Little proof was followed by a simpler version by Jewell and another by Eilon. Shaler Stidham published a different and more intuitive proof in 1972. Little, J. D. C. (1961). A Proof for the Queuing Formula: $L = \lambda W$. Operations Research 9 (3): 383–387.

System $M|G|1$

- rozkład wejścia (zmiennej U – wykładniczy),
- **rozkład czasu obsługi zadania (zmiennej V – dowolny)**,
- liczba stanowisk obsługi ($m = 1$),
- wymiar źródła zgłoszeń ($k = \infty$),
- pojemność poczekalni ($L = \infty$),
- regulamin (algorytm) obsługi - FIFO.

Wyznaczymy średnie wartości zmiennych losowych N, W, T systemu znając tylko średni czas obsługi $\frac{1}{\mu}$ (nie znając jego rozkładu).

Wartość $\overline{N}$ w systemie $M|G|1$



- N_1, N_2 – liczba zadań w systemie odpowiednio bezpośrednio po zakończeniu obsługi zadania Z_1, Z_2,
- R – liczba zadań, które przybyły do systemu w czasie obsługi zadania Z_2

Jeżeli $N_1 = 0$, to $N_2 = R$.

Wartość $\overline{N}$ w systemie $M|G|1$

$$\Gamma = \begin{cases} 1 & N_1 = 0 \\ 0 & wpp \end{cases} \quad (1)$$

$$N_2 = N_1 + R - 1 + \Gamma \quad (2)$$

$$\overline{N_2} = \overline{N_1} + \overline{R} - 1 + \overline{\Gamma} \quad (3)$$

$$\overline{R} = \rho \quad (4)$$

$$\overline{\Gamma} = 1 - \rho \quad (5)$$

$$\overline{N} = \frac{\overline{R^2} - \rho}{2(1-\rho)} + \rho \quad (6)$$

Wzór Pollaczka

$$\frac{\overline{W}}{\overline{V}} = \frac{\rho}{2(1-\rho)}(1 + \mu^2\sigma_V^2) \quad (7)$$

Średni czas oczekiwania i odpowiedzi

$$\sigma^2 = E(X^2) - [E(X)]^2$$

czyli

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

$$\overline{W} = \frac{\lambda \overline{V^2}}{2(1-\rho)} \quad (8)$$

$$\overline{T} = \overline{V} + \overline{V} = \frac{1}{\mu} + \frac{\lambda \overline{V^2}}{2(1-\rho)} \quad (9)$$

System $M|D|1$

$$P(V = v_0) = 1$$

$$\sigma_V^2 = 0$$

a więc:

$$\overline{W} = \frac{v_0 \rho}{2(1-\rho)}$$

System $M|M|1$

$$\overline{V} = \frac{1}{\mu}$$

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

a więc:

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

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

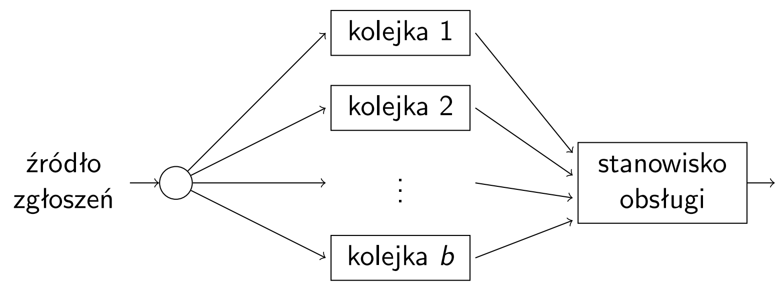
Systemy jednostanowiskowe z poissonowskim strumieniem zgłoszeń

rozkad V	G	D	M	E_k
$\overline{V}$	$\frac{1}{\mu}$	v_0	$\frac{1}{\mu}$	$\frac{k}{\mu}$
σ_V^2	σ^2	0	$\frac{1}{\mu^2}$	$\frac{k}{\mu^2}$
$\overline{W}$	$\frac{\lambda}{2(1-\rho)}\left(\frac{1}{\mu^2}+\sigma^2\right)$	$\frac{v_0\rho}{2(1-\rho)}$	$\frac{\rho}{\mu(1-\rho)}$	$\frac{\lambda(k+1)}{2k\mu^2(1-\rho)}$
$\overline{L}$	$\frac{\lambda^2\sigma^2+\rho^2}{2(1-\rho)}$	$\frac{v_0\lambda}{2(1-v_0\lambda)}$	$\frac{\rho^2}{(1-\rho)}$	$\frac{\rho^2(k+1)}{2(1-\rho)}$
P_0	$1-\rho$	$1-v_0\lambda$	$1-\rho$	$1-\frac{k\lambda}{\mu}$

Zasady nadawania priorytetów

- priorytet nadawany przez programistę,
- priorytet nadawany przez operatora,
- priorytet zależny od rozmiaru zadania,
- priorytet zależy od historii obsługi zadania.

System z algorytmem priorytetowym



Założymy, że każdemu priorytetowi π_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.

Oznaczenia

$$\rho_i = \frac{\lambda_i}{\mu_i}, i = 1, \ldots, b$$

$$\lambda = \sum_{i=1}^{b} \lambda_i$$

$$\rho = \sum_{i=1}^{b} \rho_i$$

Wariant sekwencyjny

- Przybywające zadanie o priortecie π_i dołącza się na koniec kolejki zadań z tym samym priorytetem.
- Przybycie zadania o priorytecie wyższym od aktualnie wkonywanego nie powoduje przerwania obsługi.
- Średni czas oczekiwania na obsługę nie zależy od wartości czasu obsługi, a tylko od przyznanego priorytetu.

$$W_i(t) = \overline{W_i} = \frac{\overline{W_o}}{(1-\sigma_i)(1-\sigma_{i+1})}, i = 1, \ldots, b \tag{10}$$

gdzie:

$\sigma_i = \sum_{j=i}^{b} \rho_j, \quad \overline{W_o} = \sum_{i=1}^{b} \frac{\lambda_i \overline{V_i^2}}{2},$

$\overline{V_i^2}$ jest momentem zwykłym drugiego rzędu rozkładu czasu obsługi zadań o priorytecie π_i.

Obserwacje

- wpływ zadań o priorytecie π_i i wyższym obecnych w kolejce w chwili przybycia zadania o priorytecie π_i na średni czas oczekiwania $\overline{W_i}$ wyraża czynnik $(1-\sigma_i)$ w mianowniku,
- wpływ zadań o priorytecie wyższym niż π_i przybywających do systemu w czasie oczekiwania zadania o priorytecie π_i w kolejce wyraża czynnik $(1-\sigma_{i+1})$ w mianowniku,
- wpływ zadań o priorytetach niższych od π_i, tj. $\pi_1, \pi_2, \ldots, \pi_{i-1}$ na średni czas oczekiwania $\overline{W_i}$ zawiera licznik $(\overline{W_o})$,
- dla zadań o niskich priorytetach czas oczekiwania może rosnąć do nieskończoności,
- średni czas oczekiwania da zadań o priorytecie π_i jest skończony, gdy $\rho_i + \rho_{i+1} + \ldots + \rho_b < 1$.

Przerwanie zadania

Urządzenia wejścia-wyjścia

Procesor przechodzi do wykonywania kolejnego zadania, gdy zadanie zakończy się lub zażąda urządzenia wejścia-wyjścia (przerwanie).

Wyniki analityczne są trudne do uzyskania.

Podział czasu

Round Robin
Procesor przechodzi do wykonywania kolejnego zadania po upływie określonego kwantu czasu Δ_i (przerwanie).

Można udowodnić, że średni czas odpowiedzi nie zależy od rozkładu czasu obsługi.

Round Robin

$$\overline{T}(t) = \frac{t}{1-\rho} \tag{11}$$

$$\overline{W}(t) = \frac{\rho t}{1-\rho} \tag{12}$$

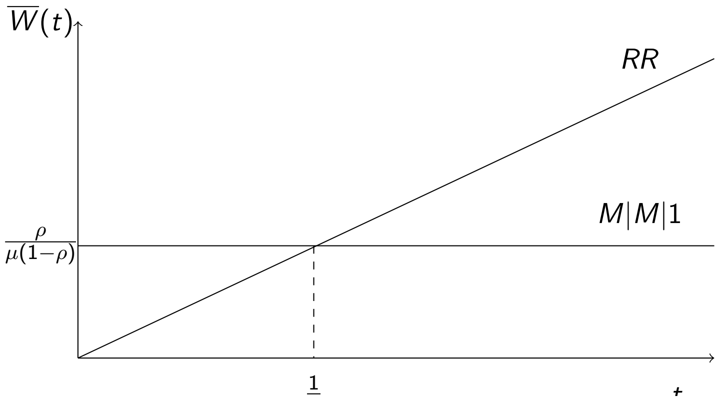
$$\overline{T} = \frac{1}{\mu(1-\rho)} \tag{13}$$

$$\overline{W} = \frac{\rho}{\mu(1-\rho)} \tag{14}$$

Powyższe średnie są identyczne, jak w systemie $M|M|1$.

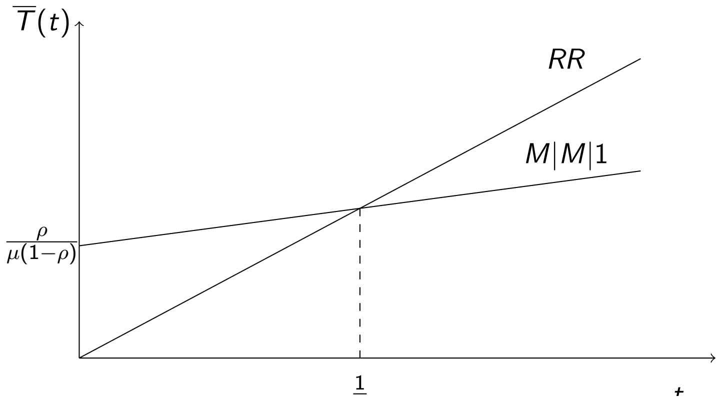
Round Robin vs. FIFO (czas oczekiwania)

$$\overline{W}_{FIFO}(t) = \frac{\rho}{\mu(1-\rho)} = \frac{\rho t}{1-\rho} = \overline{W}_{RR}(t) \tag{15}$$



Round Robin vs. FIFO (czas odpowiedzi)

$$\overline{T}_{FIFO}(t) = \frac{\rho}{\mu(1-\rho)} + t = \frac{t}{1-\rho} = \overline{T}_{RR}(t) \tag{16}$$



Round Robin z priorytetami

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:

$$\Delta_i = \frac{c_i}{\sum_{j=1}^{b} c_j \overline{N}_j} \qquad (17)$$

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.

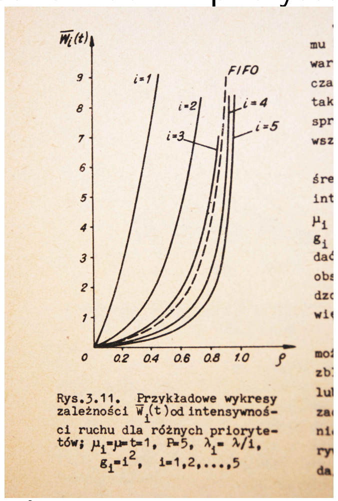
Round Robin z priorytetami

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

$$\overline{W}_i(t) = \frac{\rho t}{1-\rho}\left[1 + \sum_{j=1}^{b}\left(\frac{c_j}{c_i} - 1\right)\frac{\rho_j}{\rho}\right], i = 1, \ldots, b \qquad (18)$$

$$\overline{T}_i(t) = \frac{t}{1-\rho}\left[1 + \sum_{j=1}^{b}\left(\frac{c_j}{c_i} - 1\right)\frac{\rho_j}{\rho}\right], i = 1, \ldots, b \qquad (19)$$

Round Robin z priorytetami



Rys.3.11. Przykładowe wykresy zależności $\overline{W}_i(t)$ od intensywności ruchu dla różnych priorytetów; $\mu_i=\mu=t=1$, $b=5$, $\lambda_i=\lambda/i$, $g_i=i^2$, $i=1,2,\ldots,5$

$b = 5$

$\mu_i = \mu = t = 1, i = 1, \ldots, b$

$\lambda_i = \frac{\lambda}{i}, i = 1, \ldots, b$

$c_i = i^2, i = 1, \ldots, b$

Źródło: Badania operacyjne dla informatyków, 1984

Algorytm LIFO

- 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.

Algorytmy wielokolejkowe

multilevel processor sharing scheduling algorithms

Zadanie przybywające do systemu jest 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.

Mechanizmy zapobiegające dyskryminacji zadan długich:

- generalna amnestia – powrót do kolejki z najwyższym priorytetem,
- metoda łaski – 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

Prawdopodobieństwo stacjonarne

- system jest w stanie $N(t) > 0$
- prawdopodobieństwo przybycia zadania w czasie $\Delta t \to 0$ wynosi $\lambda \Delta t + o(\Delta t)$
- prawdopodobieństwo opuszczenia zadania w czasie $\Delta t \to 0$ wynosi $\mu \Delta t + o(\Delta t)$

Prawdopodobieństwo stacjonarne

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,

Prawdopodobieństwo stacjonarne

$$p_n = p_0 \frac{\lambda_0 \lambda_1 \ldots \lambda_{n-1}}{\mu_1 \mu_2 \ldots \mu_n} \tag{20}$$

Dodatkowe równanie otrzymujemy ze wzoru:

$$\sum_{n=0}^{\infty} p_n = 1 = p_0 + \sum_{n=1}^{\infty} p_n \tag{21}$$

$$p_0 = \left[1 + \sum_{n=1}^{\infty} \prod_{i=1}^{n} \frac{\lambda_{i-1}}{\mu_i}\right]^{-1} \tag{22}$$

Charakterystyka systemu

$$\overline{N} = \sum_{n=0}^{\infty} n p_n \tag{23}$$

a ze wzoru Little'a możemy wyznaczyć:

$$\overline{T} = \frac{\overline{N}}{\lambda} \tag{24}$$

$L < \inf$

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

$$\lambda_n = \begin{cases} \lambda & n = 0, \ldots, L \\ 0 & n > L \end{cases} \tag{25}$$

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

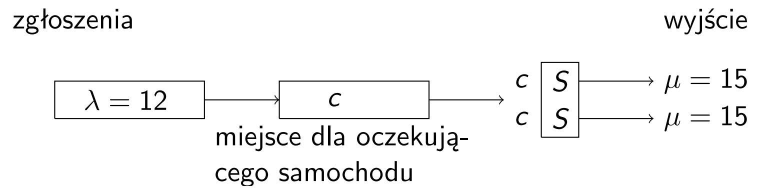
$M|M|m$

W tej sytuacji zwróćmy uwagę na czas obsługi.

$$\mu_n = \begin{cases} n\mu & 1 \leqslant n \leqslant m \\ m\mu & n \geqslant m \end{cases} \tag{26}$$

$M|M|m|L = 3$

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 $n \geqslant 3$, oraz $\mu_n = 0$ dla $n \geqslant 4$.



$M|M|1|\inf|k$

W systemie są dwie końcówki generujące zadania dla procesora, obie z rozkładem normalnym o średniej $\lambda = 1\ /3$ zgłoszenia na godzinę. Obsługa zgłoszenia trwa średnio 30 minut i ma rozkład normalny.

$$\lambda_n = \begin{cases} \frac{2}{3} & n = 0 \\ \frac{1}{3} & n = 1 \\ 0 & n = 2 \end{cases} \tag{27}$$

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

$M|M|1|\inf|k$

Mamy zatem

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

$$p_0 = \left[1 + \frac{1}{3} + \frac{1 \cdot 1}{3 \cdot 6}\right]^{-1} = \frac{18}{25}$$

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

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