Badania operacyjne

Systemy masowej obsługi

1. Opis i działanie

Jerzy Konorski
PON 16-17, CZW 12-13 pok. 139 EA
jekon@eti.pg.gda.pl

Literatura

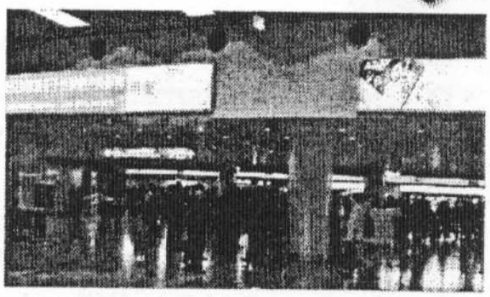
- L. Kleinrock: *Queuing systems*, vol. I, II, Wiley 1975-1976
- D. Gross, C.M. Harris: *Fundamentals of Queuing Theory*, Wiley 1998
- Joti Lal Jain, W. Boehm, Sri Gopal Mohanty: *A Course on Queuing Models*, Chapman & Hall 2006
- G. Bolch, S. Greiner, H. de Meer, K. S. Trivedi: *Queueing Networks and Markov Chains. Modeling and Performance Evaluation with Computer Science Applications*, 2nd Ed., Wiley-Interscience 2006

- D. Koenig, D. Stoyan: *Metody teorii obsługi masowej*, WNT, Warszawa 1979
- G.P. Klimow: *Procesy obsługi masowej*, WNT, Warszawa 1979
- J. Błażewicz i in.: *Badania operacyjne dla informatyków*, WNT, Warszawa 1983, rozdziały 3 i 4
- B. Filipowicz: *Modele stochastyczne w badaniach operacyjnych. Analiza i synteza systemów obsługi i sieci kolejkowych*, WNT, Warszawa 1996
- T. Czachórski: *Modele kolejkowe w ocenie efektywności sieci i systemów komputerowych*, Wyd. J. Skalmierski, Gliwice 1999

Systemy masowej obsługi

- **system informatyczny** (komputer / *call center* / baza danych / serwer WWW) : przerwania / uaktualnienia / zapytania / transakcje oczekują w kolejce na zwolnienie się operatorów / jednostek obliczeniowych / obszarów danych
- **urządzenie telekomunikacyjne** (karta sieciowa / centrala telefoniczna / koncentrator szybkiego łącza danych): ramki danych / wywołania abonenckie oczekują na zwolnienie się łącza
- **infrastruktura transportowa** (wjazd na autostradę / stacja benzynowa / nabrzeże portowe / pas startowy): pojazdy oczekują na "szczelinę obsługi" (*slot*)
- **punkt świadczenia usług** (bankomat / kasa w supermarkecie / urząd): klienci / petenci oczekują na zwolnienie się osób / urządzeń obsługujących

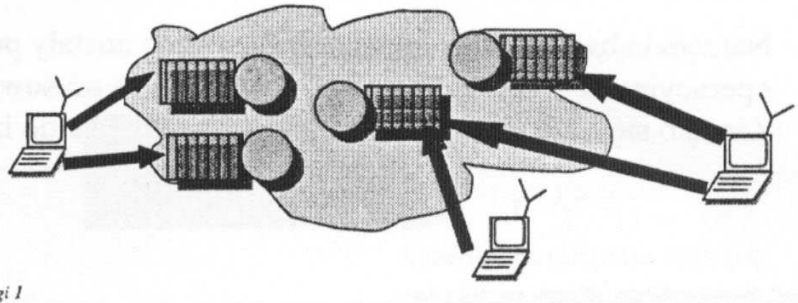
Systemy masowej obsługi (2)



Wszędzie kolejki!
Trzeba je uwzględniać w projektowaniu systemów.

Systemy masowej obsługi (3)

- **system** posiada **zasoby** – ograniczone, zwrotne
- reaguje na zdarzenia w postaci przybywających **zgłoszeń** = żądań dostępu do zasobów
- w odpowiedzi przydziela zgłoszeniu wybrane zasoby, umożliwiając ich konsumpcję przez pewien skończony czas = **obsługa zgłoszenia**
- zasoby systemu zdolne do świadczenia obsługi = **procesory**
- zgłoszenie może natrafić na procesory zajęte obsługą innych zgłoszeń
- w takim wypadku zostaje umieszczone w **pamięci buforowej**, gdzie oczekuje w **kolejce** na zwolnienie się procesora i rozpoczęcie obsługi



Misja teorii kolejek

Populacja zgłoszeń / źródeł zgłoszeń jest zwykle bardzo liczna.

Pozbawia to sensu optymalizację konkretnych scenariuszy przybywania zgłoszeń, np. harmonogramowanie zbioru zadań minimalizujące czas zakończenia lub liczbę wykorzystanych procesorów.

Można jedynie analizować i projektować zasady obsługi przybywających "na bieżąco" **strumieni zgłoszeń**. W tym celu studiujemy przebiegi czasowe rozmaitych charakterystyk kolejek = **procesy kolejkowe** (**procesy obsługi**)

Misja teorii kolejek (2)

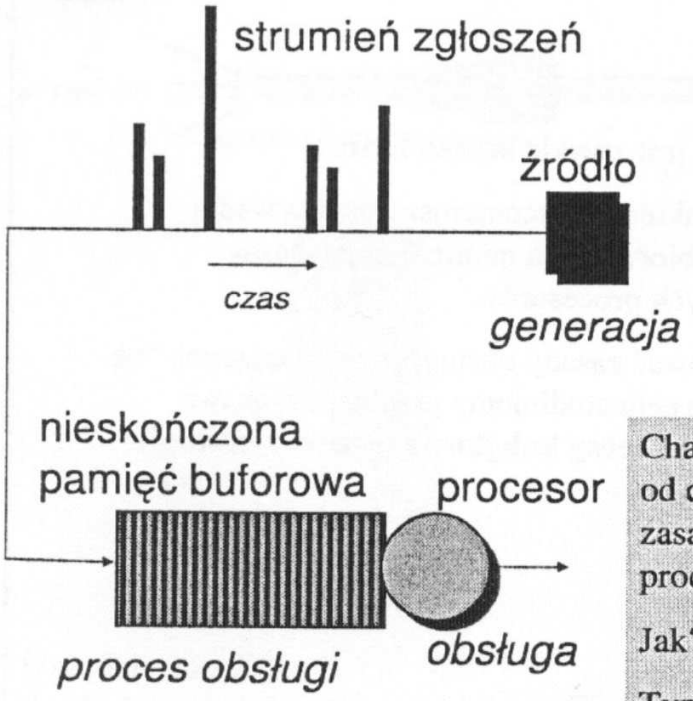
Przy licznej populacji zgłoszeń chwilowy popyt na obsługę często przekracza chwilową podaż obsługi – wtedy tworzą się kolejki. Należy ograniczać ich negatywne skutki, np.

- niezadowolenie klientów (opóźnienia, odmowa / rezygnacja z obsługi),
- koszty buforowania i zarządzania,

dbając jednocześnie o ekonomię wykorzystania procesorów.

Narzędzia badawcze po temu wypracowane zostały przez ważny dział badań operacyjnych – **teorię kolejek** (inaczej **teorię masowej obsługi**). Zaczęło się od bombowców krążących w kolejce do lądowania...

Najprostszy model



Elementy najprostszego modelu systemu masowej obsługi:

- pamięć buforowa
- procesor
- strumień zgłoszeń

Charakterystyki procesu obsługi zależą od charakterystyk strumienia zgłoszeń i zasad pracy systemu pamięć buforowa-procesor.

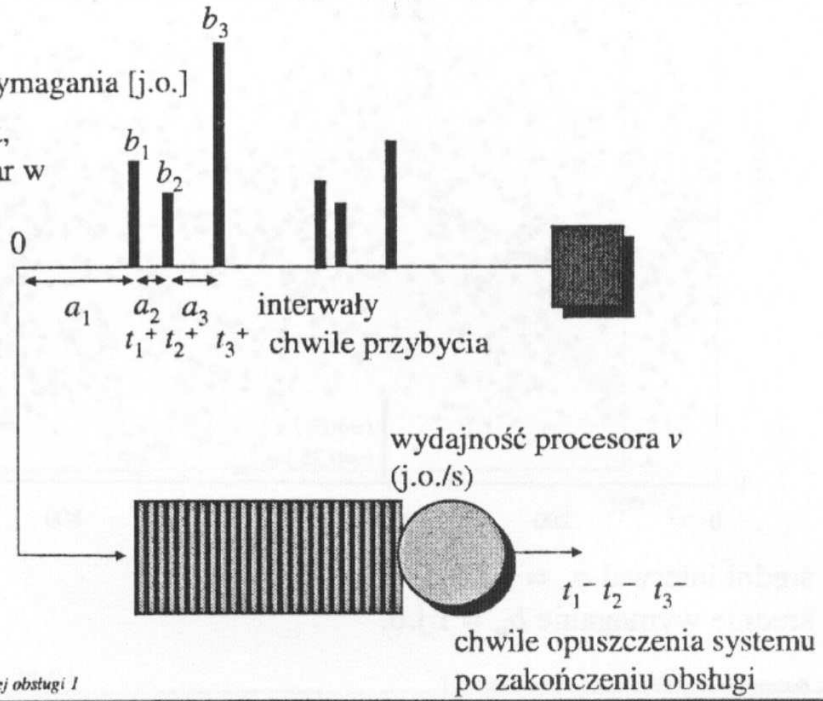
Jak?

Tym właśnie zajmuje się teoria kolejek.

Najprostszy model (2)

Które charakterystyki strumienia zgłoszeń i zasady obsługi są istotne?

j.o.: jednostka obsługi (bajt, operacja elementarna, towar w koszyku, ...)



Najprostszy model (3)

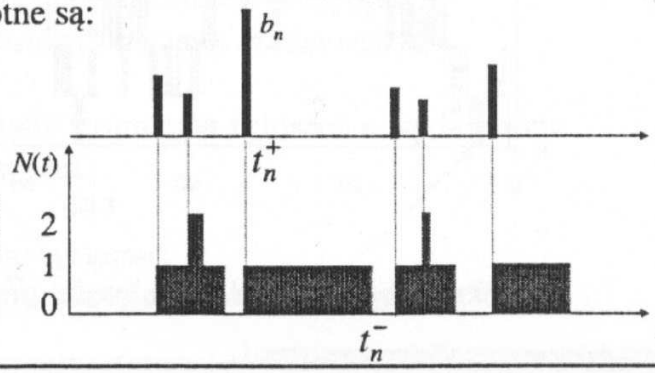
Uwagi:

- $t_n^+ = a_1 + \ldots + a_n$
- $b_n/v = \tau_n$ = wymagany czas obsługi zgłoszenia n
- $d_n = t_n^- - t_n^+$ = **opóźnienie systemowe** zgłoszenia n
- $w_n = d_n - b_n/v$ = **opóźnienie buforowania** zgłoszenia n (stracony czas)
- $N(t) = \#\{n \mid t_n^+ \le t \le t_n^-\}$ = **liczba zgłoszeń w systemie** w chwili t

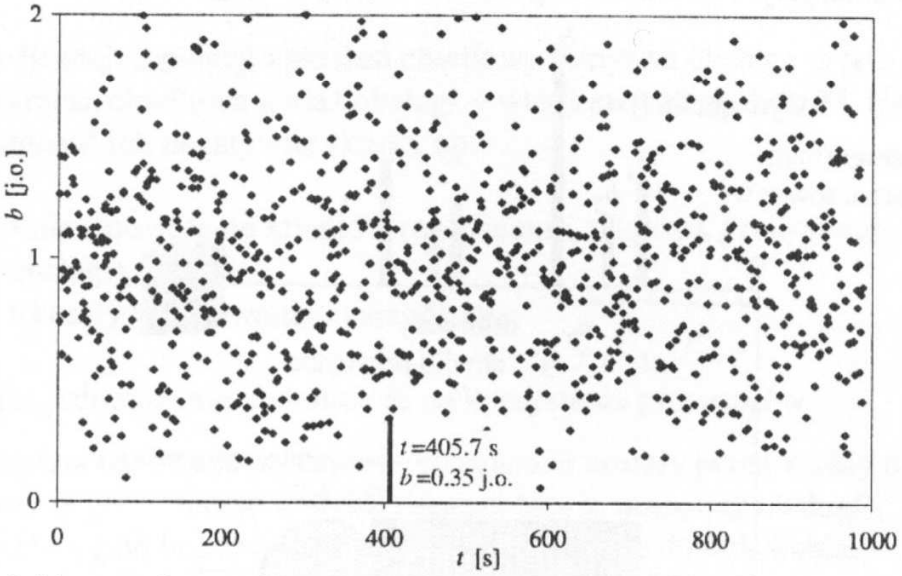
Dla konstrukcji procesów $N(t)$ i w_n istotne są:
- wydajność procesora v,
- **dyscyplina** (kolejność) **obsługi**!

FIFO: $t_n^- = \max\{t_n^+, t_{n-1}^-\} + b_n/v$

$v = 1$ j.o./s



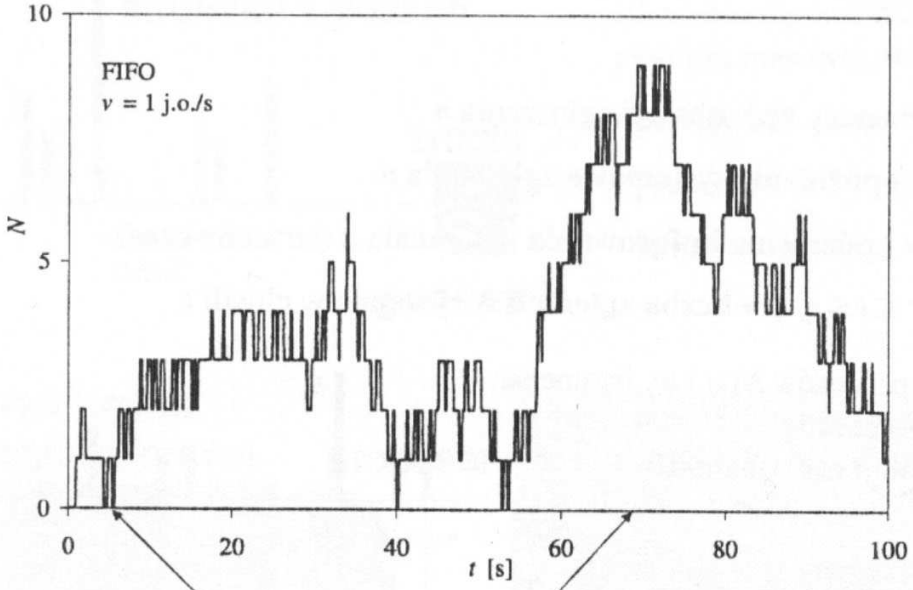
Strumień zgłoszeń i proces obsługi – przykład



średni interwał $a_{śr} = 1$ s

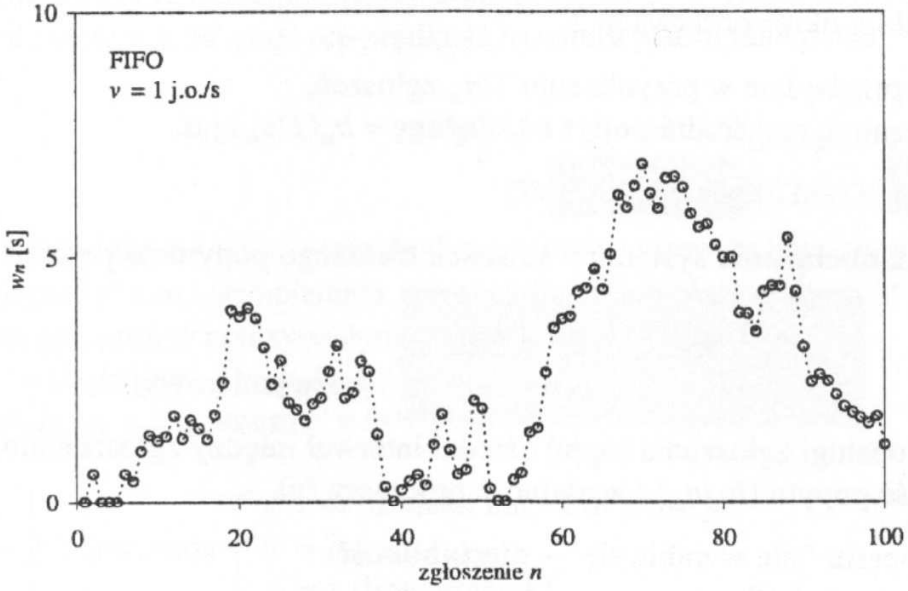
średnie wymaganie $b_{śr} = 1$ j.o.

Strumień zgłoszeń i proces obsługi – przykład (2)



okresy bezczynności i zajętości procesora

Strumień zgłoszeń i proces obsługi – przykład (3)



Jakie własności ma "dobry" proces obsługi?

- z punktu widzenia zgłoszeń: niewielkie opóźnienia buforowania, rzadkie odmowy przyjęcia zgłoszenia wskutek zbyt długiej kolejki
- z punktu widzenia operatora systemu: wysokie wykorzystanie procesora (rzadkie okresy bezczynności)

Sprzeczne oczekiwania! Coraz rzadszym okresom bezczynności towarzyszy:

- występowanie kolejek
- wzrost długości kolejek
- systematyczne narastanie kolejki / lawina odmów przyjęcia
  – procesor "nie wyrabia się z obsługą" w czasie rzeczywistym!

Relacja pomiędzy charakterystykami strumienia zgłoszeń a wydajnością procesora określa **obciążenie systemu.**
Jak je ilościowo zdefiniować?
Jaką to daje informację o działaniu systemu?
Przy jakim obciążeniu procesor przestaje "wyrabiać się z obsługą"?

Współczynnik obciążenia – ruch oferowany

Rozpatrzmy długi okres obserwacji T.

W tym czasie przybędzie w przybliżeniu $T/a_{śr}$ zgłoszeń,
łącznie reprezentują one średni **popyt na obsługę** = $b_{śr}(T/a_{śr})$ j.o.

Podaż obsługi w tym czasie wyniesie vT.

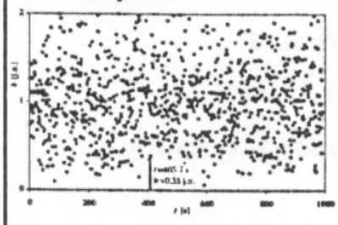
Współczynnik obciążenia systemu = stosunek średniego popytu do podaży:

$$r = \frac{b_{śr}(T/a_{śr})}{vT} = \frac{b_{śr}/v}{a_{śr}} = \frac{b_{śr}/a_{śr}}{v} \quad \text{(bezwymiarowy)}$$

= średni czas obsługi zgłoszenia ($b_{śr}/v$) / średni interwał między zgłoszeniami ($a_{śr}$),
= intensywność popytu ($b_{śr}/a_{śr}$) / wydajność procesora (v).

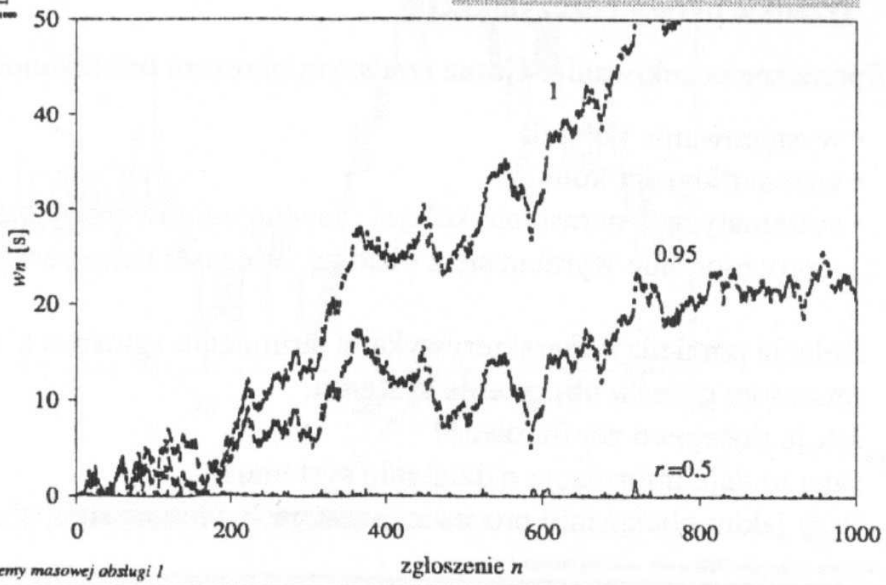
Przy $r > 1$ procesor "nie wyrabia się" – **niestabilność**.
Co się dzieje przy $r \leq 1$?

Współczynnik obciążenia – ruch oferowany (2)



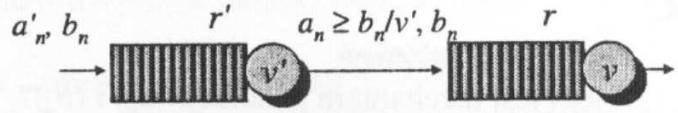
ten sam strumień zgłoszeń ($a_{śr}$ = 1 s, $b_{śr}$ = 1 j.o.)
zmniejszanie v

niestabilność przy $r = 1$!



Wpływ prędkości transferu źródło-system

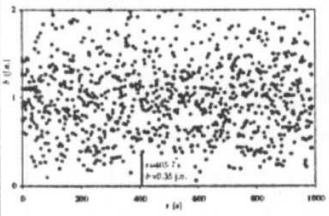
Dotąd zakładaliśmy natychmiastowy transfer zgłoszeń ze źródła do systemu (dowolnie małe a_n). W praktyce prędkość transferu jest ograniczona.



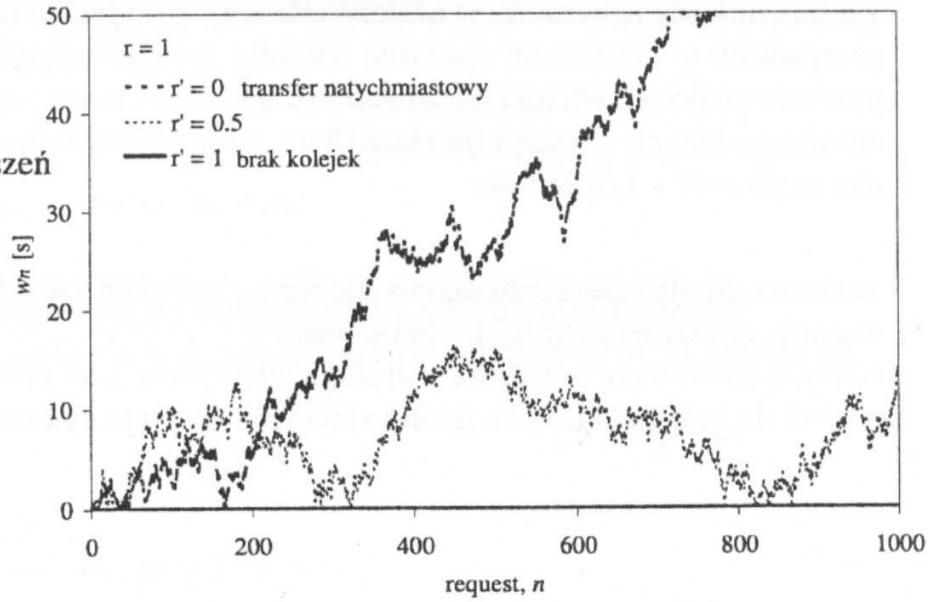
Można to modelować jako "wirtualny" wejściowy system kolejkowy z procesorem o wydajności $v' < \infty$ i strumieniem zgłoszeń (b_n) i dowolnie małymi (a'_n).
Strumień zgłoszeń w rzeczywistym systemie ma $a_n \in [b_n/v', \infty)$.
Niech $r' = (b_{śr}/a'_{śr})/v'$, $r = (b_{śr}/a_{śr})/v$.
Oczywiście, $a_{śr} = a'_{śr}$, zatem $r' = (v/v')r$.

$$\left.\begin{array}{l} r' = 0 \\ r' = 0.5r \\ r' = r \end{array}\right\} \text{odpowiada} \left\{\begin{array}{l} v' = \infty \text{ (transfer natychmiastowy)}, \\ v' = 2v \\ v' = v \text{ (brak kolejek)}. \end{array}\right.$$

Wpływ prędkości transferu źródło-system (2)



ten sam strumień zgłoszeń
($a_{śr}$ = 1 s, $b_{śr}$ = 1 j.o.)



Przy $r = 1$ lecz skończonej prędkości transferu, system może pozostać stabilny!

W stronę realistycznych modeli

Jaki wpływ na procesy obsługi mają inne charakterystyki strumienia zgłoszeń, zachowanie się zgłoszeń w systemie i zasady obsługi?

- **strumień zgłoszeń**
  jaki jest mechanizm generacji (a_n) i (b_n)?
  zmienność w czasie? zależność od procesu obsługi? grupowość?
  wymagania b_n – znane/nieznane w chwili przybycia?
- **pamięć buforowa**
  skończona pojemność Q?
  ograniczenia dostępności? kiedy odmowa przyjęcia – *drop-tail*?

W stronę realistycznych modeli (2)

- **zachowanie się zgłoszenia w obliczu odmowy przyjęcia / długiej kolejki**
  przepadek?
  ponowienie po odczekaniu (*timeout*)?
  eliminacja innych zgłoszeń (*pushout*)?
  niecierpliwość – I, II rodzaju
- **zachowanie się zgłoszenia po rozpoczęciu / zakończeniu obsługi**
  warunki opuszczenia kolejki – blokowanie?
  warunki zwolnienia procesora – obsługa także przez inne procesory?
  powrót do kolejki? kiedy? z jakimi zmianami (kreacja zgłoszeń potomnych)?

W stronę realistycznych modeli (3)

- **dyscyplina obsługi – co decyduje o kolejności obsługi**
  chwila przybycia (FIFO, LIFO)?
  przypadek (RANDOM)?
  wielkość wymagania / bieżące zaawansowanie obsługi (SJF, LASF)?
  karuzela (RR)?
  dodatkowe wymagania zgłoszeń – progi niecierpliwości?
  zewnętrzna klasyfikacja zgłoszeń – priorytety, sprawiedliwość (HOL, WFQ)?
- **procesor – tryb obsługi**
  zachowawczy (okres zajętości trwa $\Sigma b_n/v$)?
  niezachowawczy (awarie / strumień tła / "wakacje",
  obsługa porcjami, zaniechania / wznowienia obsług)?
  bez przeplotu? z przeplotem?
  więcej niż 1 procesor ? dostępność? wymagania zgłoszeń?
  obsługa umiejscowiona / zrównoleglona?

Czy uda się stworzyć uniwersalny symulator systemów masowej obsługi?

Stan ustalony

W **stacjonarnym** systemie masowej obsługi:

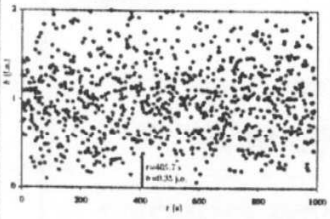
- mechanizm generujący strumień zgłoszeń pozostaje niezmienny w czasie (charakterystyki strumienia zgłoszeń uśredniane za długi okres obserwacji nie zależą od długości okresu obserwacji)
- wydajność procesora v pozostaje stała.

W *dobrze zaprojektowanym* systemie ($r < 1$) występuje **stan ustalony** (charakterystyki procesu obsługi uśredniane za coraz dłuższy okres obserwacji ustalają się).

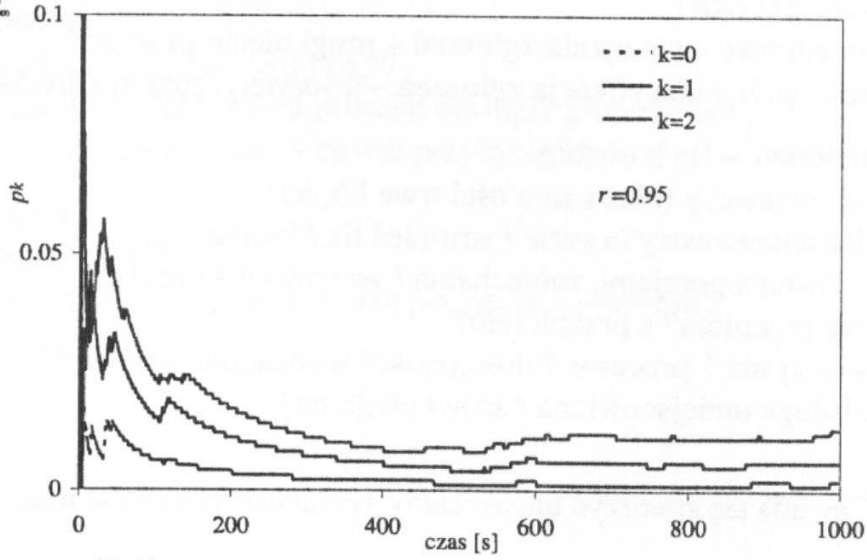
W szczególności

$$\frac{|\{t \le T \mid N(t) = k\}|}{T} \to p_k \text{ przy } T \to \infty$$

Stan ustalony (2)

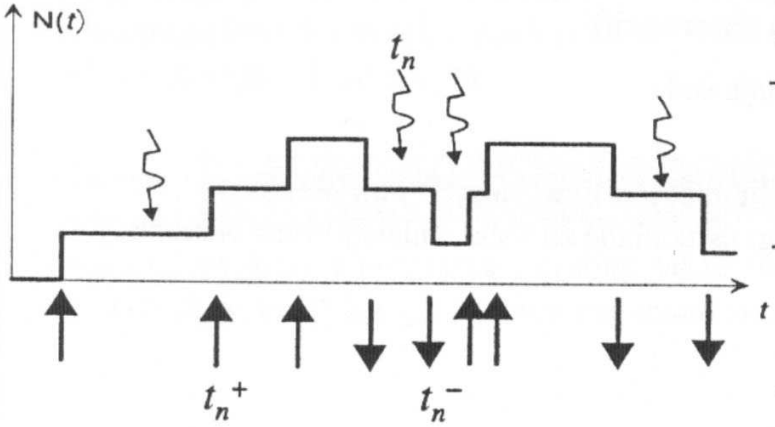


ten sam strumień zgłoszeń
($a_{śr}$ = 1 s, $b_{śr}$ = 1 j.o.)



Stan ustalony (3)

Obserwacja $N(t)$ – różne punkty widzenia



Przy $M \to \infty$:

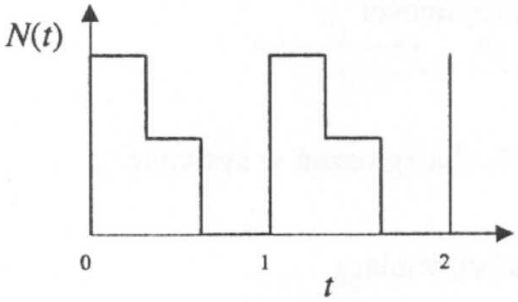
$$\frac{\#\{n \le M \mid N(t_n) = k\}}{M} \to p_k$$

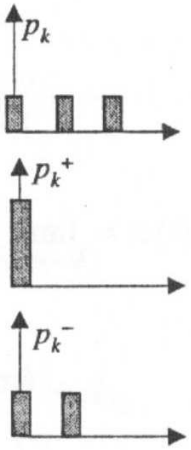
$$\frac{\#\{n \le M \mid N(t_n^+ - 0) = k\}}{M} \to p_k^+$$

$$\frac{\#\{n \le M \mid N(t_n^- + 0) = k\}}{M} \to p_k^-$$

Stan ustalony (4)

Uwaga, "praktycy"!

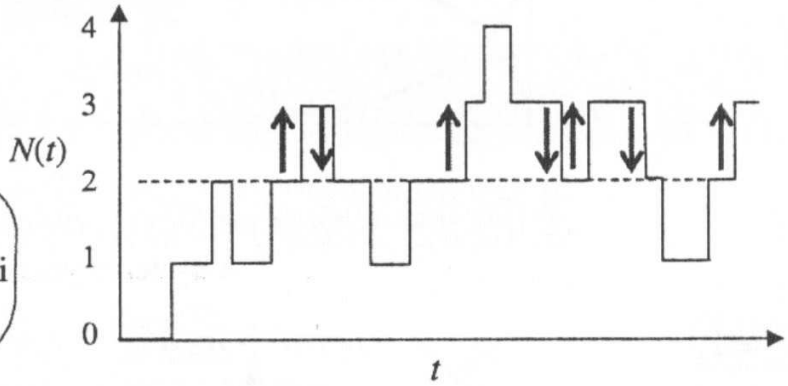




Stan ustalony (5)

Argument skoków ±1: jeżeli $N(t)$ zmienia się tylko o ±1, to $p_k^+ \equiv p_k^-$

Różnica między liczbą skoków w góre i w dół = długość kolejki (< ∞ z prawd. 1), zaś różnica #↑ i #↓ jest ≤ 1 w dowolnym przedziale czasu!



Uwaga: Jeżeli możliwe są odmowy przyjęcia, to liczą się tylko zgłoszenia przyjęte. Np. gdy $Q < \infty$ (odmowy wskutek przepełnienia pamięci buforowej), to argument skoków ±1 daje

$$\frac{p_k^+}{1-p_Q^+} = p_k^-, \; k = 0,\ldots,Q-1$$

Stan ustalony (6)

$p_1^+ + \ldots + p_{Q-1}^+$ — frakcja zgłoszeń buforowanych

$L = p_Q^+$ — frakcja zgłoszeń utraconych wskutek przepełnienia pamięci buforowej (bez ponowień)

$1 - p_0$ — współczynnik zajętości (wykorzystania) procesora = 1 – współczynnik bezczynności

$$N_{śr} = \lim_{T\to\infty} \frac{1}{T}\int_0^T N(t)dt = \lim_{M\to\infty} \frac{1}{M}\sum_{n=1}^{M} N(t_n)$$

średnia liczba zgłoszeń w systemie

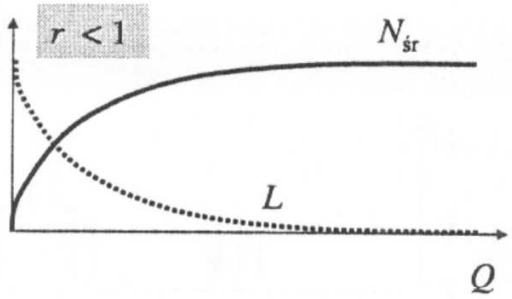
$$\frac{w_{śr}}{\frac{b_{śr}}{v}} = \frac{w_{śr}}{\tau_{śr}}$$

średnie opóźnienie buforowania,

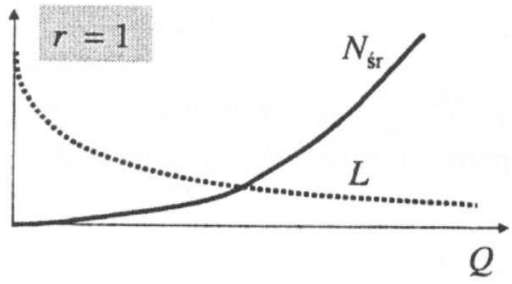
$$\frac{d_{śr}}{\tau_{śr}}$$

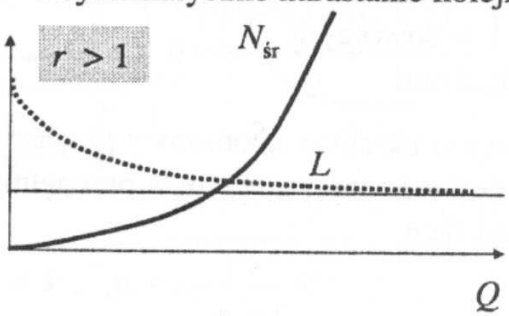
...systemowe (znormalizowane względem średniego czasu obsługi – podstawa oceny systemu!)

Stan ustalony (7)

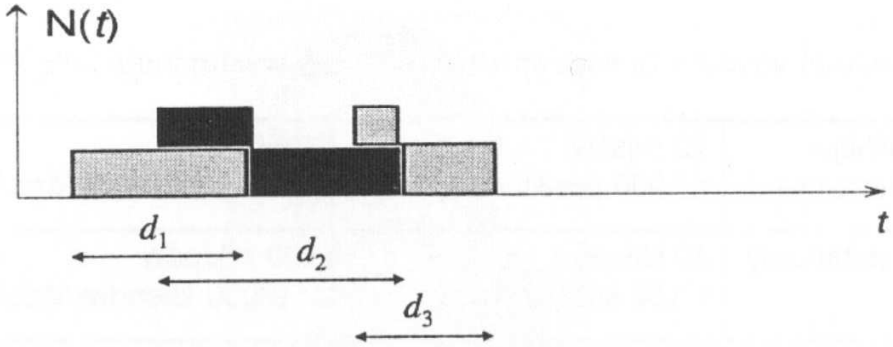


+ systematyczne narastanie kolejki





Prawo Little'a



$$\text{przy } M,T \to \infty: \underbrace{\int_0^T N(t)dt}_{TN_{śr}} = \underbrace{d_1 + \ldots + d_M}_{(T/a_{śr})(1-L)d_{śr}}$$

Prawo Little'a (2)

$$N_{śr} = \frac{1-L}{a_{śr}} d_{śr}$$

średnia liczba zgłoszeń w systemie =
(średnia **przepływność** systemu) x (średni czas przebywania w systemie)

średnia populacja = (średnia cyrkulacja) x (średni czas życia)

Obowiązuje dla dowolnego fragmentu systemu:

- procesor: $1 - p_0 = \frac{1-L}{a_{śr}} \tau_{śr} = (1-L)r$ **równanie ciągłości przepływu**
- pamięć buforowa: $N_{śr} - (1 - p_0) = \frac{1-L}{a_{śr}} w_{śr}$

Teoria kolejek straci znaczenie?

W przyszłości wzrost v (a więc spadek r do zera) wyeliminuje kolejki?

komunikacja transatlantycka	10 rejsów x 1000 pasażerów/20 dni	1000 lotów x 250 pasażerów/dzień	x500
handel detaliczny	10 sklepów x 100 klientów/dzień	100 sklepów x 10000 klientów/dzień	x1000
łącze dostępowe	10 kb/s	100 Mb/s	x10 000
RAM	10 KB	2 GB	x200 000

Mimo wszystko odpowiedź: Nie.

Teoria kolejek straci znaczenie? (2)

Po pierwsze, samoloty na trasach transatlantyckich, kasy w supermarketach, łącza internetowe, centra komputerowe są zatłoczone jak nigdy.

Popyt na obsługę rośnie w ślad za podażą – r wcale nie spada!

Teoria kolejek straci znaczenie? (3)

Po drugie, skąd właściwie biorą się kolejki?

Nie tylko z powodu $v < \infty$ and $b_{śr}/a_{śr} > 0$, lecz przede wszystkim z powodu zmienności a_n (arytmia strumienia zgłoszeń) oraz b_n (kapryśne wymagania zgłoszeń)!

Arytmia powoduje, że *chwilowe* obciążenie systemu waha się od 0 do ∞. Eliminacja choćby tylko okazjonalnych kolejek wymaga procesora z $v = \infty$.

Przy $b_m/a_m < \infty$ daje to $r = 0$ tj. zerowe wykorzystanie procesora!!! Bardzo nieeekonomiczne, niezależnie od nieprzewidywalnych postępów w technologii i zarządzaniu.

Co *jest* ekonomiczne? Utrzymywanie $r < 1$, tj. $v > b_m/a_m$, *ale nie bardzo!* Oznacza to dopuszczenie do powstawania kolejek.

Komentarze do teorii kolejek

Teoria kolejek to analiza matematyczna. Lepszy byłby prototyp / symulator?

- wiarygodne oceny kłopotliwych charakterystyk
  zdarzenia rzadkie – przekroczenie progu długości kolejki,
  długi okres zajętości – czy to już niestabilność?
- jakościowy (a nie swoisty względem określonych ustawień) wpływ parametrów modelu na charakterystyki procesów obsługi – oszczędza wielu niepotrzebnych eksperymentów
- ! - uniwersalny (jakościowo, a często i ilościowo) charakter wyników dla nieskomplikowanych modeli systemów masowej obsługi

Wbrew pozorom analiza matematyczna jest bardzo kosztowna. Opłaca się, gdy udziela odpowiedzi, o które inaczej byłoby trudno.

Komentarze do teorii kolejek (2)

Agner Krarup Erlang (1878-1929)

duński matematyk i inżynier,
pierwszy zrozumiał, że nowoczesna telefonia nie może się obejść bez matematyki

jednostka ruchu telekomunikacyjnego = 1 *erlang*

Komentarze do teorii kolejek (3)

Jak teoria kolejek "ma się do" analizy:

szeregowania zadań
tam: optymalny harmonogram dla ustalonego zbioru zadań
tutaj: ciągły nieprzewidywalny strumień przybywających zadań

procesów współbieżnych
tam: deterministyczna analiza konkretnych scenariuszy zdarzeń
tutaj: w dużych populacjach zdarzeń (obsługa *masowa...*) widać tylko charakterystyki statystyczne

procesów przypadkowych
wykorzystuje podobny aparat obliczeniowy
(proces obsługi = inercyjne przekształcenie strumienia zgłoszeń)

Uffffff...

Dziękuję za uwagę (na razie).

Badania operacyjne III

Systemy masowej obsługi

2. Modele i charakterystyki stochastyczne

Jerzy Konorski
jekon@eti.pg.gda.pl

II

Zmienne losowe i strumienie losowe

Dla danego scenariusza zgłoszeń można wyznaczyć $a_{śr}$, $b_{śr}$ i inne charakterystyki. Jednak potrzeba modelu generacji nieprzewidywalnych (a_n) i (b_n).

Dobrym modelem zjawisk nieprzewidywalnych są **zmienne losowe** – przyjmują konkretne wartości z pewnymi prawdopodobieństwami, tj. posiadają pewne **rozkłady prawdopodobieństwa**. W ramach powszechnego w nauce i technice **podejścia bayesowskiego** przyjmujemy te rozkłady za znane / dające się wyliczyć.

Dalej interesować się będziemy wyłącznie **strumieniami losowymi**. Dlaczego nielosowymi nie?

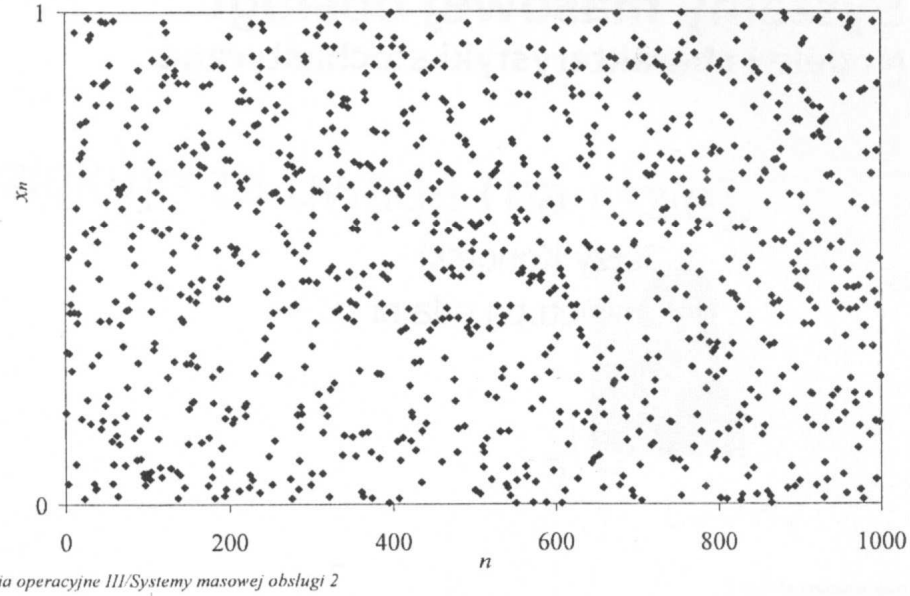
- praktycznie nie występują,
- nie niosą informacji, nie stanowią problemu projektowego (!)

Przedmiotem dalszych rozważań będą więc **stochastyczne** modele i charakterystyki systemów masowej obsługi.

Zmienne losowe i strumienie losowe (2)

Przykładem zmiennej losowej może być liczba zwracana przez `random` (o ile nie znamy algorytmu generacji liczb pseudolosowych). Posiada ona o rozkład równomierny w przedziale [0,1).

Pierwsze 1000 **realizacji** (zaobserwowanych wartości):

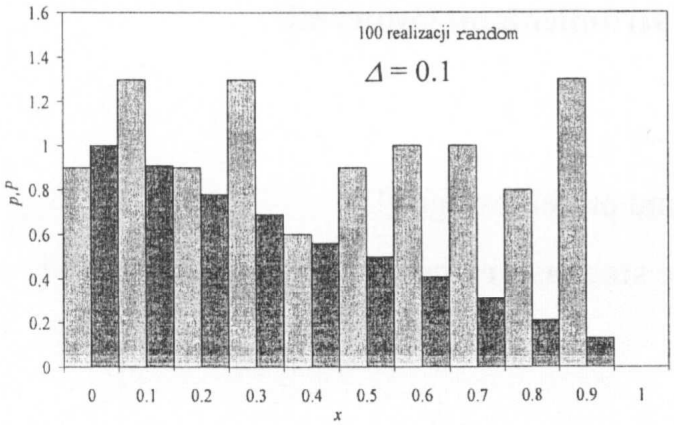


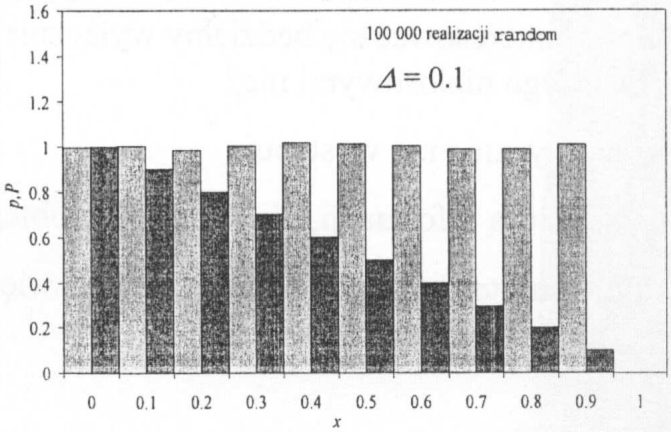
Rozkłady empiryczne

Na podstawie realizacji $x_1,\ldots,x_N$ tworzymy **histogram**:

- dzielimy zakres możliwych realizacji na przedziały o szerokości Δ,
- zliczamy realizacje w i-tym przedziale, $k_i = \#\{n \mid i\Delta \le x_n \le (i+1)\Delta\}$
- dla argumentu $i\Delta$ rysujemy słupek o szerokości Δ i wysokości $p_i = (k_i/N)/\Delta$,

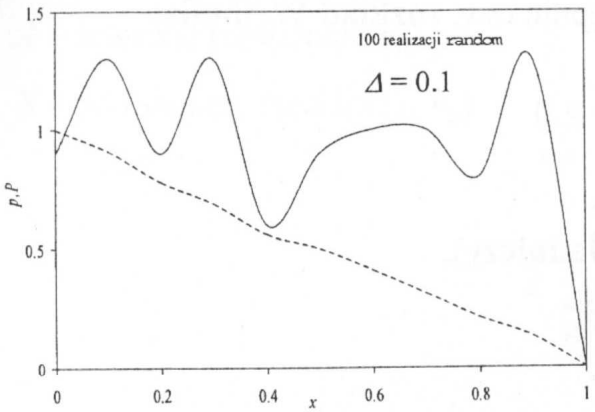
Histogram kumulacyjny powstaje analogicznie, słupki o wysokości $P_i = \Sigma_{j\ge i} p_j$.

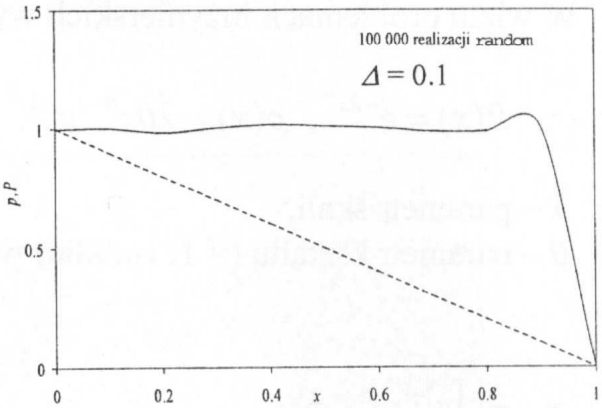




Rozkłady empiryczne (2)

...albo bardziej czytelnie, przy pomocy linii ciągłych





Wartość średnia: $x_{śr} = \dfrac{x_1 + \ldots + x_N}{N}$

Odchylenie standardowe (rozrzut wokół średniej): $\sigma_x = \sqrt{\dfrac{1}{N-1}\sum_{n=1}^{N}(x_n - x_{śr})^2}$

Rozkłady teoretyczne

Gęstość prawdopodobieństwa i **komplementarna dystrybuanta** – histogramy, jakie otrzymalibyśmy w granicy przy $N \to \infty$ i $\Delta \to 0$.

Gęstość prawdopodobieństwa: $p(x) = \lim_{\Delta \to 0} \dfrac{\Pr[x \le X < x + \Delta]}{\Delta}$

Dla dowolnych x' i x'': $\Pr[x' \le X < x''] = \int_{x'}^{x''} p(x)dx$,

Komplementarna dystrybuanta: $P(x) = \int_{x}^{\infty} p(y)dy$

$$x_{śr} = \int_{-\infty}^{\infty} x p(x)dx, \quad \sigma_x = \sqrt{\int_{-\infty}^{\infty}(x - x_{śr})^2 p(x)dx}$$

Rozkłady teoretyczne (2)

W wielu problemach inżynierskich występuje tzw. **rozkład Weibulla**:

$$P(x) = e^{-\lambda x^{\theta}}, \quad p(x) = \lambda\theta x^{\theta-1} e^{-\lambda x^{\theta}}, \quad x \geq 0$$

λ – parametr skali,
θ – parametr kształtu (= 1: **rozkład wykładniczy).**

$$x_{śr} = \int_0^{\infty} \sqrt[\theta]{\frac{y}{\lambda}} e^{-y} dy$$

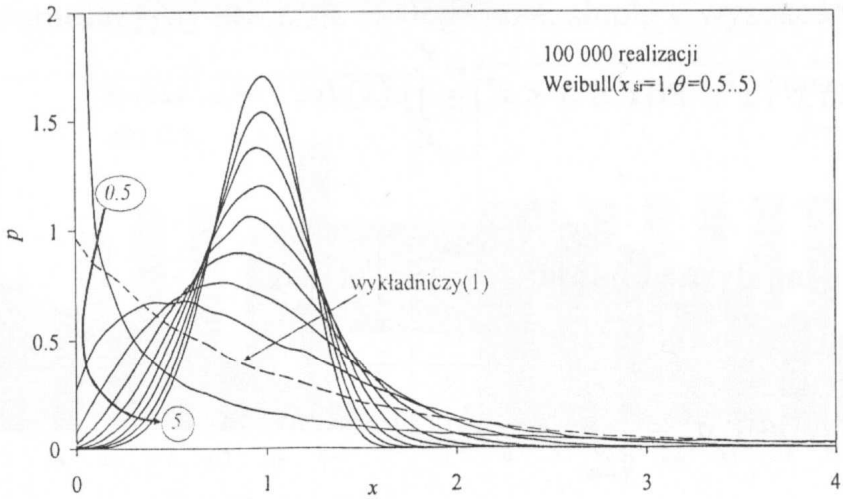
$$\sigma_x = \sqrt{\int_0^{\infty} \sqrt[\theta]{\frac{y^2}{\lambda^2}} e^{-y} dy - x_{śr}^2}$$

Sztuczna generacja strumieni zgłoszeń

Mamy generator `random` liczb pseudolosowych (z_n).
Jak generować liczby pseudolosowe (x_n) o dowolnej funkcji $P(x)$?

x_n jest rozwiązaniem równania $P(x) = z_n$, np. dla r. Weibulla: $x_n = \sqrt[\theta]{-\frac{\ln z_n}{\lambda}}$

(tzw. **metoda odwracania dystrybuanty**; jest też wiele innych).

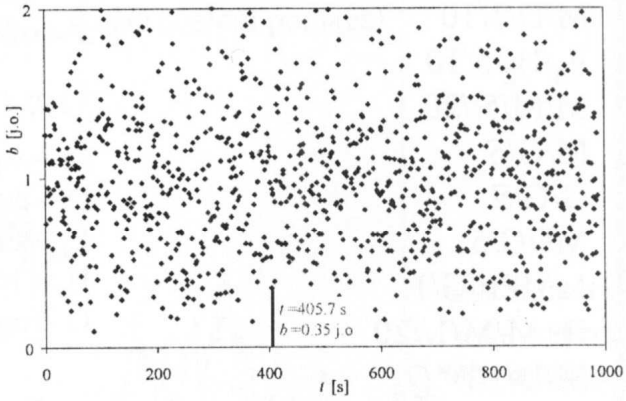


Sztuczna generacja strumieni zgłoszeń (2)

Wprowadźmy zmienne losowe:

A – interwał (realizacje a_n),

B – wymaganie (realizacje b_n)

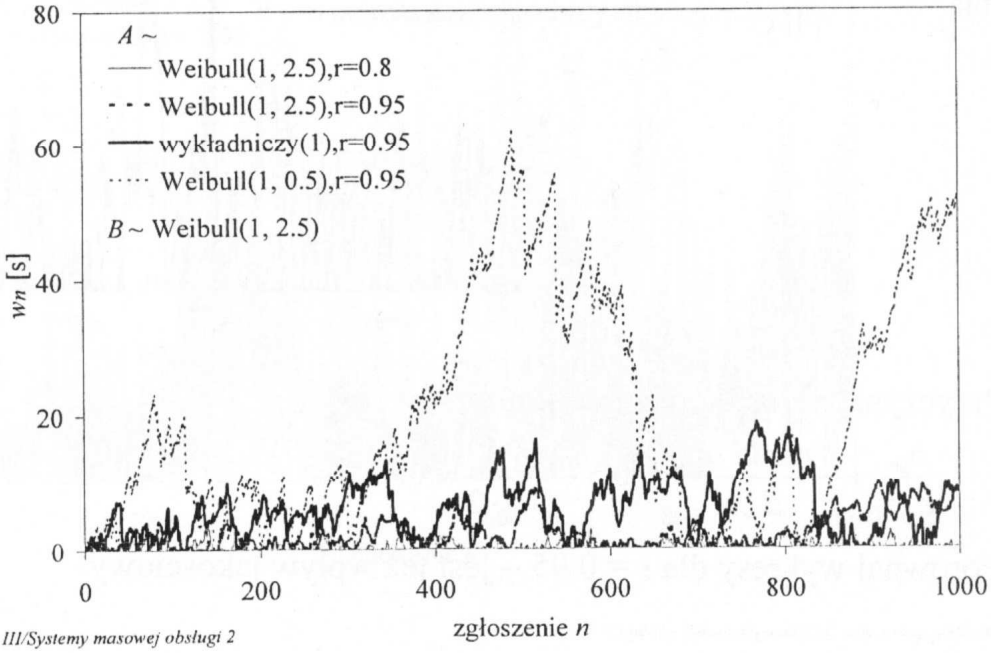


Znany nam już strumień został wygenerowany jako A, B ~ Weibull(1, 2.5).

Wpływ kształtu rozkładu *A*, *B*

Widzieliśmy, że przebieg procesu obsługi silnie zależy od wartości średnich $a_{śr}$, $b_{śr}$. Czy także od *kształtu* rozkładów A i B?

nawet bardzo!



Notacja Kendalla

$A/B/S/Q/J$ ...

Przykłady:

M/M/1
M/D/1/10
$E_3/H_2/5/10$
M/M/5//20
M/G/S
G/G/S
M^X/G/5
MMPP/G/1
D+M/M/1//20
M/D/1/P*Q
M/G/1 LIFO, HOL, ...

S – #procesorów
Q – pojemność pamięci buforowej
J – rozmiar populacji źródeł zgłoszeń

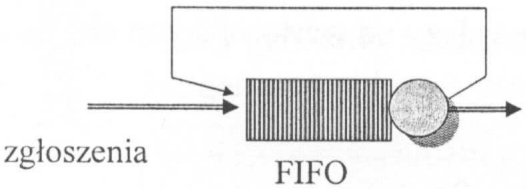
Typy rozkładów:

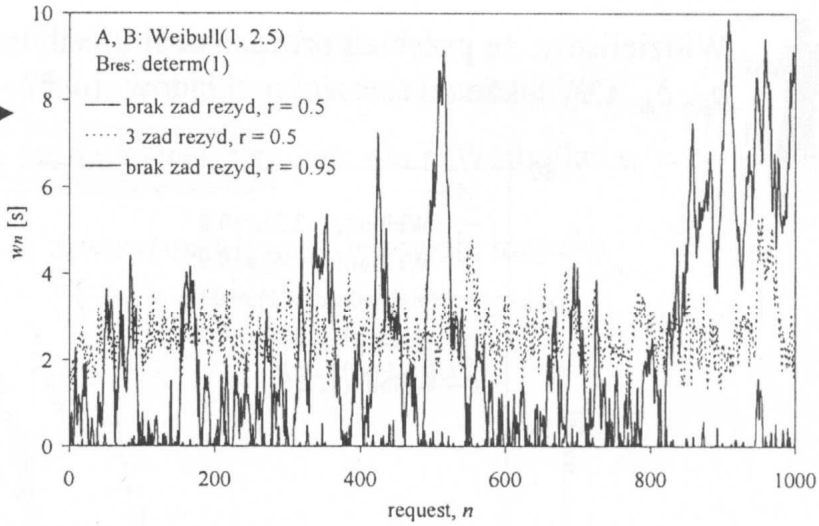
M – wykładniczy
D – deterministyczny
E_k – Erlanga rzędu k
H_k – hiperwykładniczy rzędu k
G – ogólny

Jaki jest wpływ innych czynników – S, Q, zachowań zgłoszeń, zasad obsługi?

Zadania rezydentne

zadania rezydentne





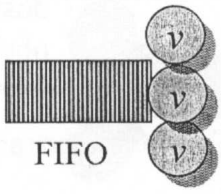
Przyczyną wzrostu opóźnień jest wzrost r?

$$\frac{b_{śr}}{a_{śr}v} + \frac{\#rezyd \cdot b_{śr}}{d_{śr}v} \approx 0.5 + \frac{3 \cdot 1}{3.5 \cdot 2} \approx 0.95$$

Porównaj wykresy dla $r = 0.95$ – jest też wpływ jakościowy.

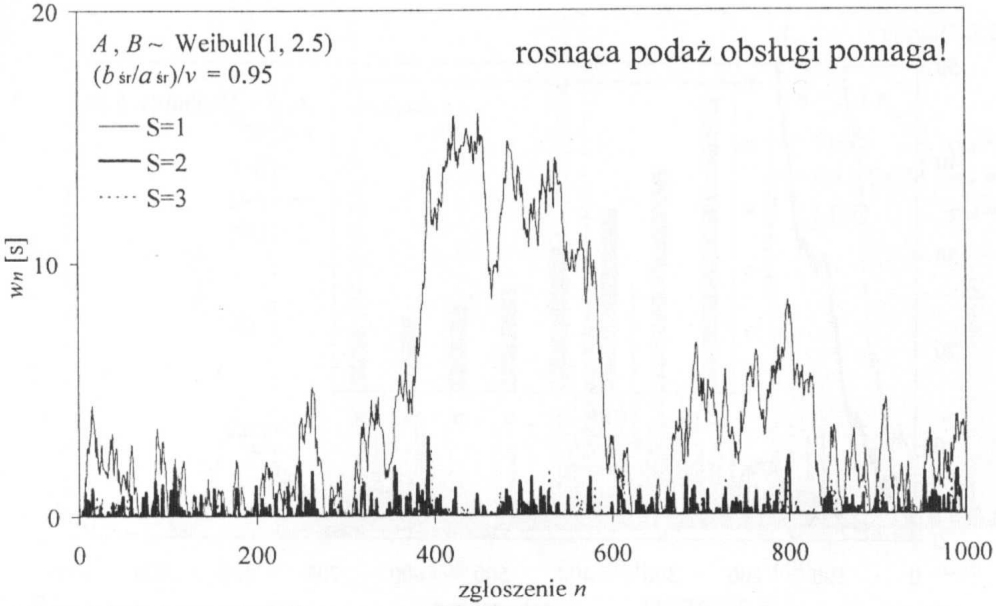
System wieloprocesorowy: *A/B/S*

wymagania:
dowolny procesor



dostępność: pełna
obsługa: umiejscowiona

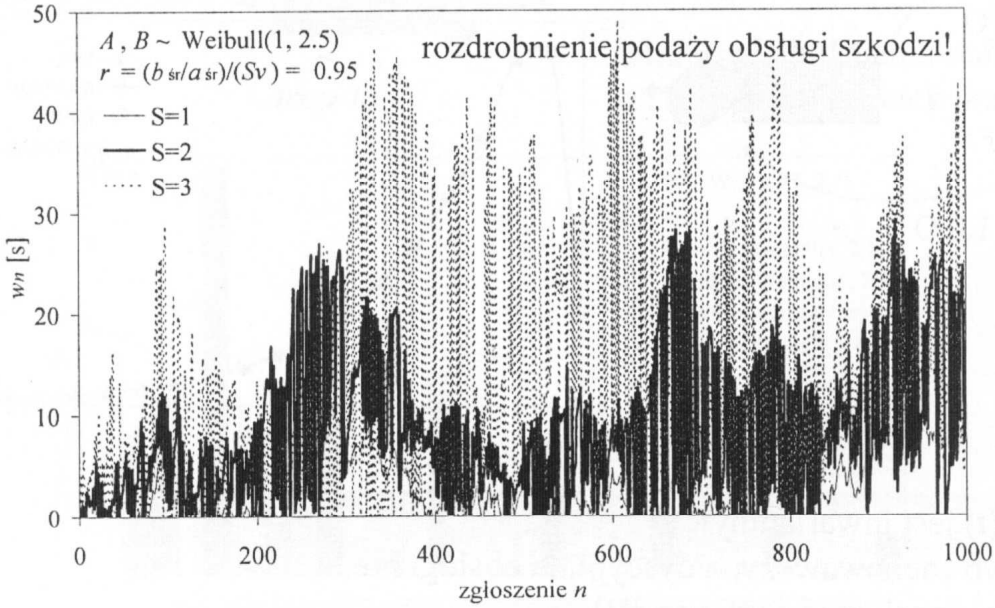
Stała intensywność popytu:



normatywny współczynnik obciążenia: cały popyt skierowany na 1 procesor

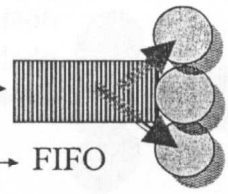
System wieloprocesorowy: *A/B/S* (2)

Sprawiedliwsze porównanie: stały współczynnik obciążenia. I co?!



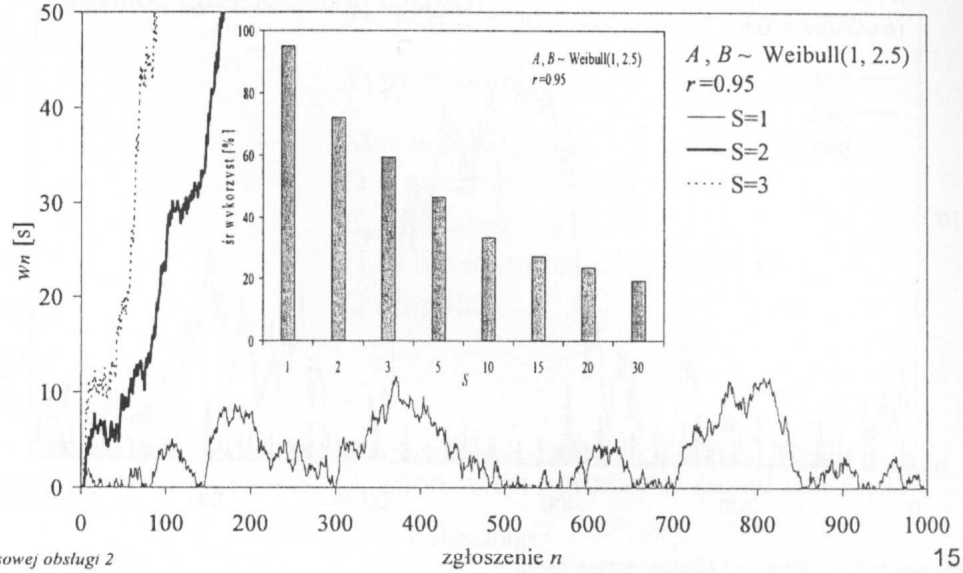
System wieloprocesorowy: *A/B/S* (3)

wymagania:
wybrany procesor

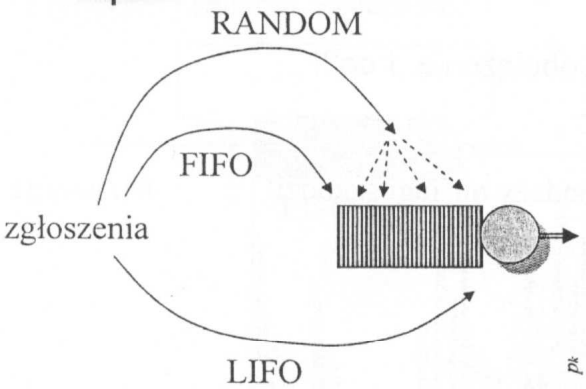


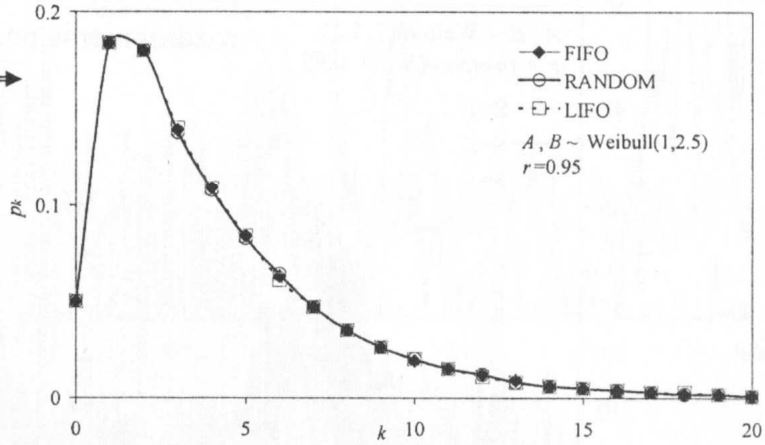
dostępność: pełna
obsługa: umiejscowiona

motywacja:
procesor pośredniczący
(pole komutacyjne,
kontroler równoważący
obciążenie,...)



Dyscyplina obsługi

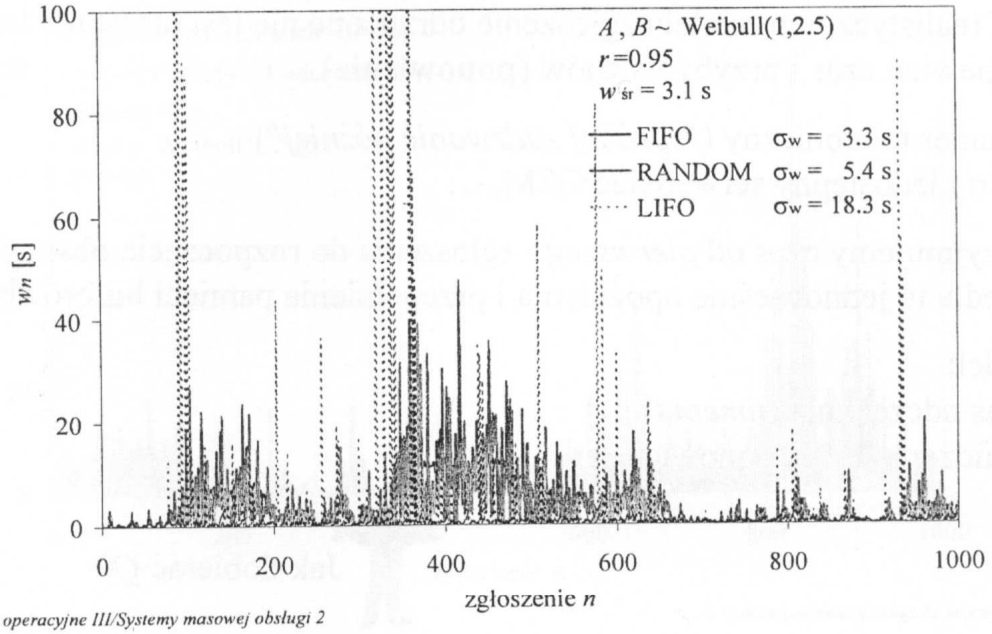




Rozkład $N(t)$ jest inwariantny!
(Dopóki tryb zachowawczy, a dyscyplina obsługi nie wykorzystuje znajomości wymagań!)

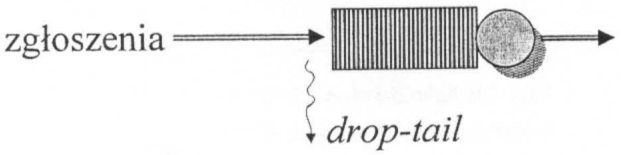
Dyscyplina obsługi (2)

Która dyscyplina powoduje największe $w_{śr}$? A która największe σ_w?



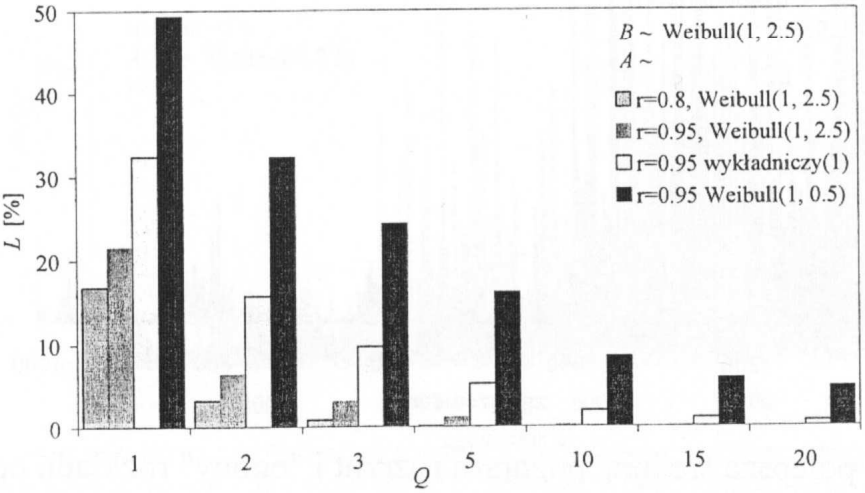
Skończona pojemność bufora: *A/B/1/Q*

pojemność Q zgłoszeń
(łącznie z obsługiwanym)

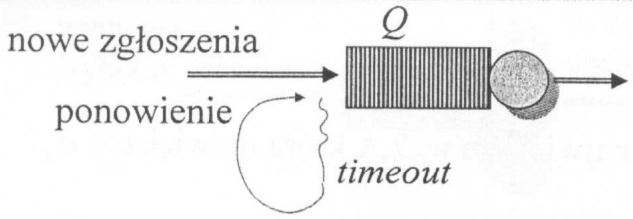


Motywacja:
- koszt pamięci?!
- zarządzanie, opóźnienia.

Mniejsze opóźnienia buforowania za cenę strat wskutek przepełnienia.



A/B/1/Q z ponowieniami



W bardziej realistycznym modelu zgłoszenie odrzucone nie jest utracone, lecz odczekuje pewien czas i przybywa znów (**ponowienie**).

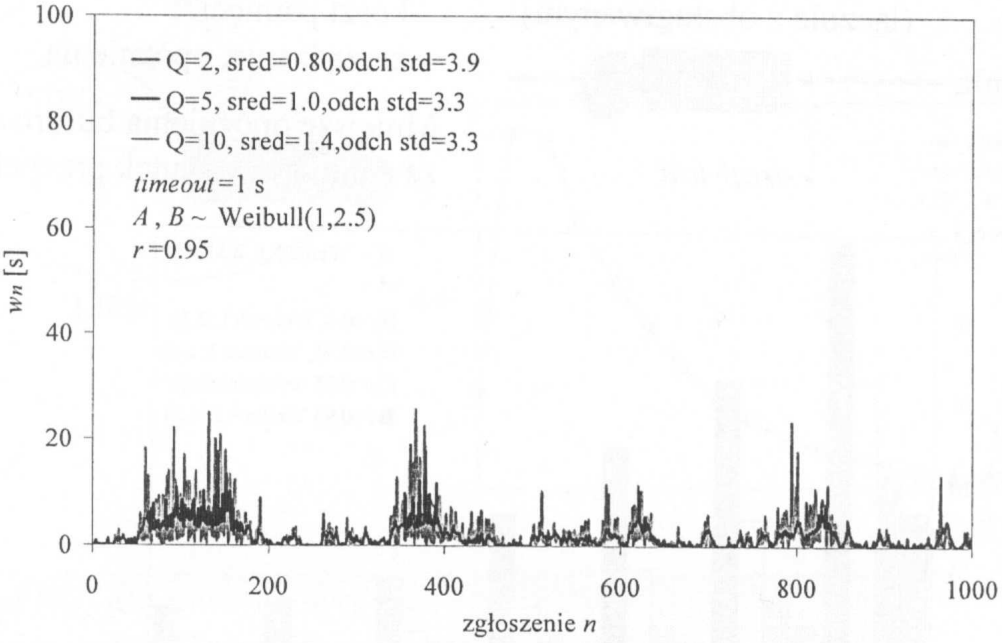
- zajęty numer telefoniczny (*"spróbuj zadzwonić później"*)
- chwilowo niedostępny serwer/sieć GSM, ...

Jako w_n przyjmujemy czas od *pierwszego* zgłoszenia do rozpoczęcia obsługi. Odzwierciedla to jednocześnie opóźnienia i przepełnienia pamięci buforowej.

Prosty model:
- stały czas odczekania (*timeout*)
- nieograniczona liczba ponowień.

Jak dobierać Q?

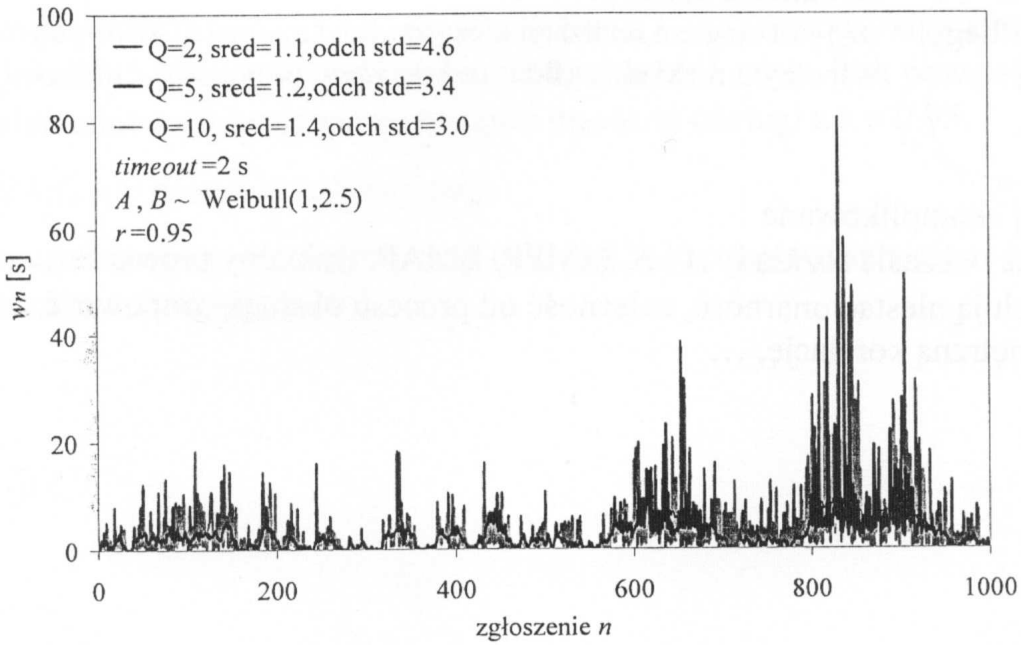
A/B/1/Q z ponowieniami (2)



Zmniejszanie Q polepsza średnią, pogarsza rozrzut i "ogony" rozkładu opóźnień.

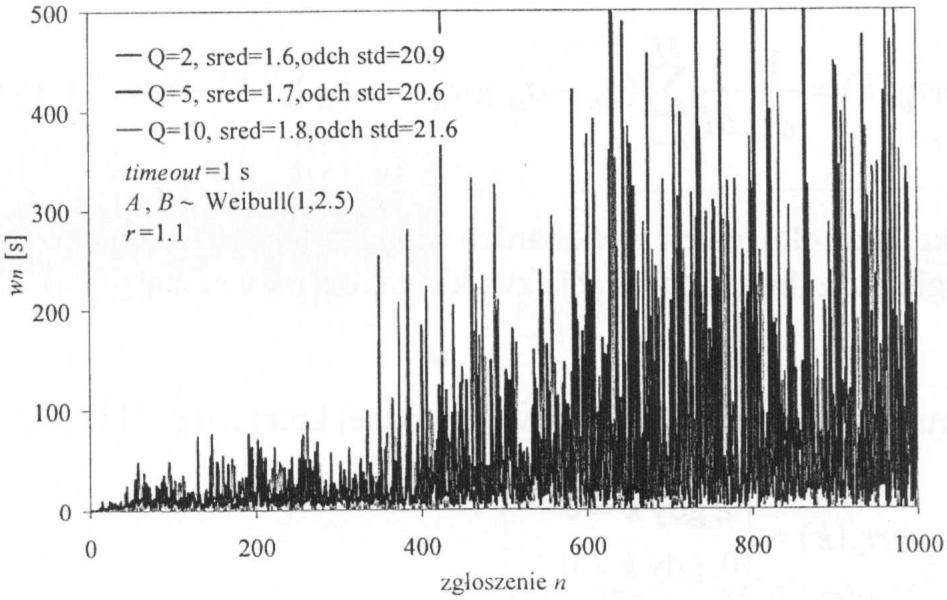
A/B/1/Q z ponowieniami (3)

Staje się to bardziej widoczne przy "cierpliwszych" zgłoszeniach:



A/B/1/Q z ponowieniami (4)

Przy nieco większym obciążeniu dramatycznie widoczne są "ogony":



Ważniejsze modele strumieni zgłoszeń

- Weibulla
- Bernouilliego
- Erlanga
- gamma
- ...

strumienie odnowy:
(a_n) jest ciągiem realizacji niezależnych zmiennych losowych o wspólnym rozkładzie (**iid** – *independent, identically distributed*)

- bardziej skomplikowane
  zmienne w czasie rozkłady A i B, MMPP, BMAP, fraktalny proces Browna, ...
  – modelują niestacjonarność, zależność od procesu obsługi, grupowość, wewnętrzną korelację, ...

Wewnętrzna korelacja

Czy o przebiegu procesu kolejkowego przesądzają same rozkłady A i B, czy też rolę odgrywa **wewnętrzna korelacja** (**autokorelacja**) w ciągu (a_n)?

$$corr_a(k) = \frac{1}{\sigma_a^2}\frac{1}{M}\sum_{n=1}^{M}(a_n - a_{śr})(a_{n+k} - a_{śr}), \quad M \to \infty, k = 0, 1, 2, ...$$

(**funkcja autokorelacji** = jak bardzo skorelowane są ze sobą interwały odległe o k zgłoszeń; korelacja zwykle zanika przy rosnącym k)

W strumieniu odnowy brak jest wewnętrznej korelacji (~ "biały szum"):

$$corr_a(k) = \begin{cases} 1, \text{ gdy } k = 0 \\ 0, \text{ gdy } k \neq 0 \end{cases}$$

Wewnętrzna korelacja (2)

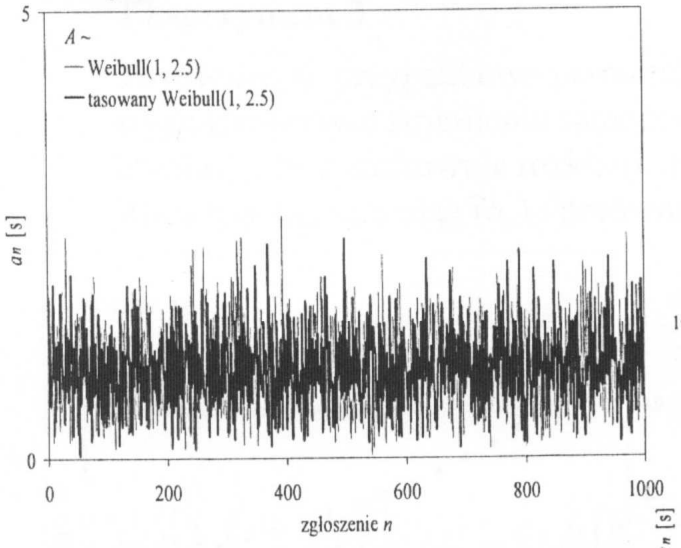
Eksperyment 1

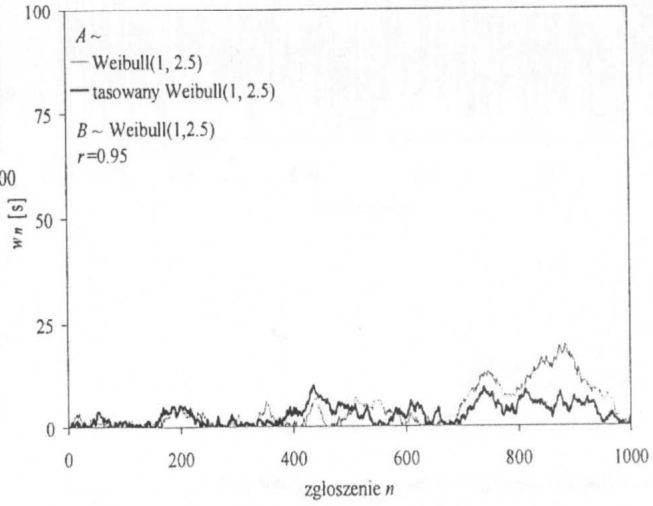
Generujemy ciągi (a_n) i (b_n) zgodnie z Weibull(1, 2.5) metodą odwracania dystrybuanty, wykorzystując kolejne liczby pseudolosowe. Tak powstały strumień odnowy kierujemy do systemu masowej obsługi z $r = 0.95$.

Obserwujemy proces kolejkowy (w_n).

Następnie **tasujemy** wygenerowany ciąg (a_n), bierzemy ten sam ciąg (b_n) i powtarzamy obserwację (w_n).

Wewnętrzna korelacja (3)





Wewnętrzna korelacja (4)

Eksperyment 2

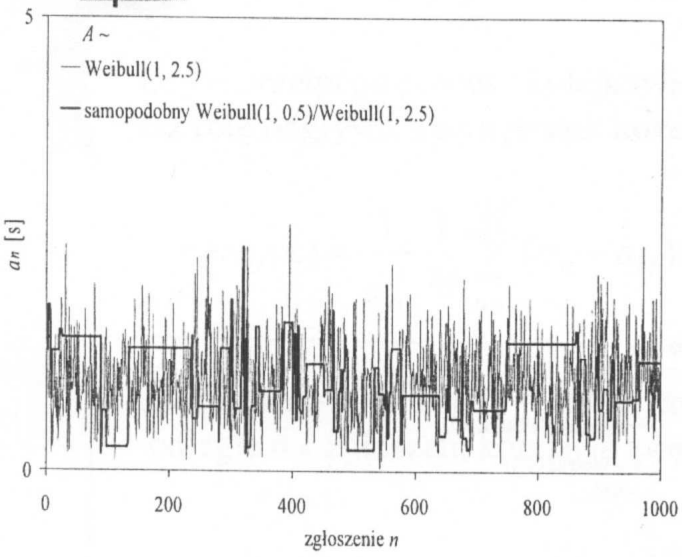
Generujemy ciąg (a_n) metodą odwracania dystrybuanty, $A \sim$ Weibull(1, 2.5),

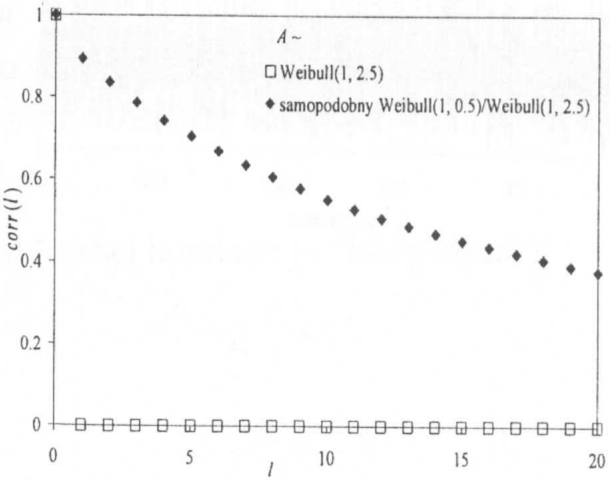
- wykorzystując kolejne liczby pseudolosowe – strumień odnowy
- powtarzając interwał liczbę razy wylosowaną z rozkładu Weibull(1, 0.5)

W obu wariantach rozkład interwału jest taki sam. Jednak drugi wariant daje (a_n) z dalekosiężną wewnętrzną korelacją – tzw. **strumień samopodobny**.

Bierzemy ten sam ciąg (b_n) i znów kierujemy tak powstały strumień zgłoszeń do systemu masowej obsługi, w którym $r = 0.95$.

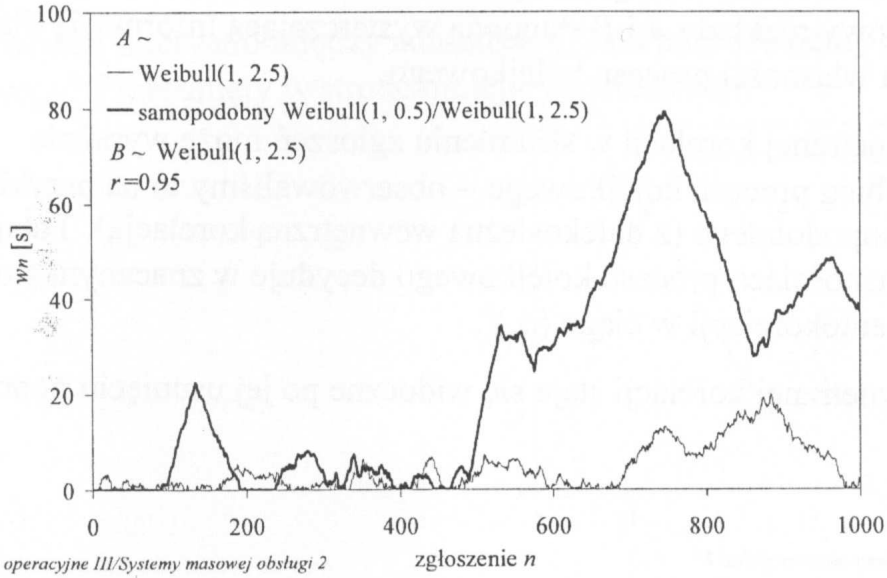
Wewnętrzna korelacja (5)





Wewnętrzna korelacja (6)

Porównanie procesów kolejkowych dla strumienia odnowy i samopodobnego wypada dramatycznie…

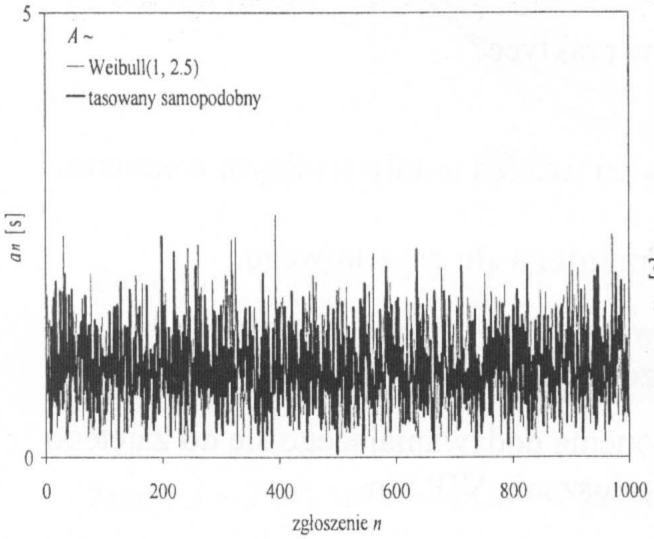


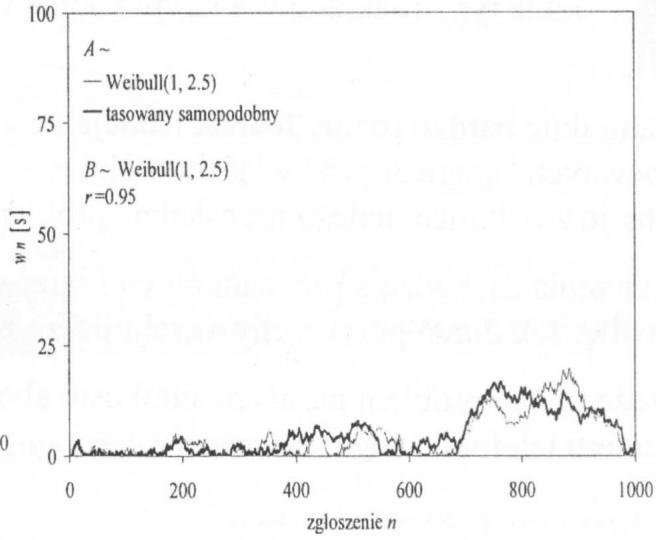
Wewnętrzna korelacja (7)

Eksperyment 3

Tasujemy, tj. przypadkowo permutujemy ciąg (a_n) we wcześniej wygenerowanym strumieniu samopodobnmy, co niszczy wewnętrzną korelację, lecz zachowuje rozkłady A i B.

Bierzemy ten sam ciąg (b_n) i porównujemy ze strumieniem odnowy…





Wewnętrzna korelacja (8)

Wnioski:

- Tasowanie strumienia odnowy nie zmienia (braku) wewnętrznej korelacji, nie wpływa też istotnie na przebieg procesu kolejkowego – można się o tym przekonać np. konstruując histogram opóźnień buforowania. W przypadku strumienia odnowy rozkłady A i B stanowią wystarczającą informację dla przewidywania własności procesu kolejkowego.
- Obecność wewnętrznej korelacji w strumieniu zgłoszeń może wyraźnie pogarszać przebieg procesu kolejkowego – obserwowaliśmy to na przykładzie strumienia samopodobnego (z dalekosiężną wewnętrzną korelacją). Tutaj o przebiegu i własnościach procesu kolejkowego decyduje w znacznym stopniu postać funkcji autokorelacji w ciągu (a_n).
- Znaczenie wewnętrznej korelacji staje się widoczne po jej usunięciu poprzez tasowanie.

Strumień odnowy: interwał resztkowy

Z wyjątkiem szczególnych zastosowań (rytm serca, wylewy Nilu, ruch WWW,...) strumienie odnowy stanowią dobry model rzeczywistych strumieni zgłoszeń. Przez chwilę skupimy uwagę tylko na interwałach.

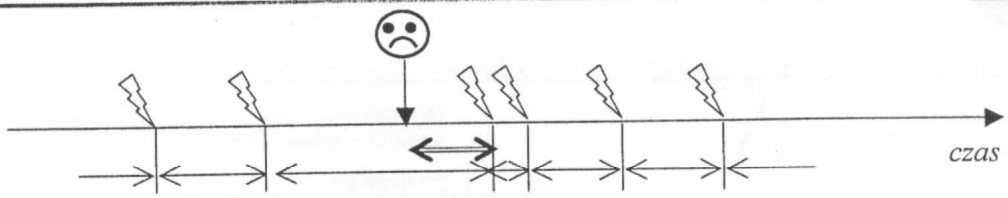
Jakie typy rozkładów A są spotykane w praktyce?

Naturalnie bardzo różne. Jednak istnieje pewien rozkład o uniwersalnym znaczeniu i pewnych "magicznych" właściwościach.
Aby je zrozumieć, należy prześledzić problem **interwału resztkowego**.

Zdarzenia zachodzą z przypadkowymi interwałami. Wybieramy przypadkową chwilę. Jak długo poczekamy na najbliższe zdarzenie?

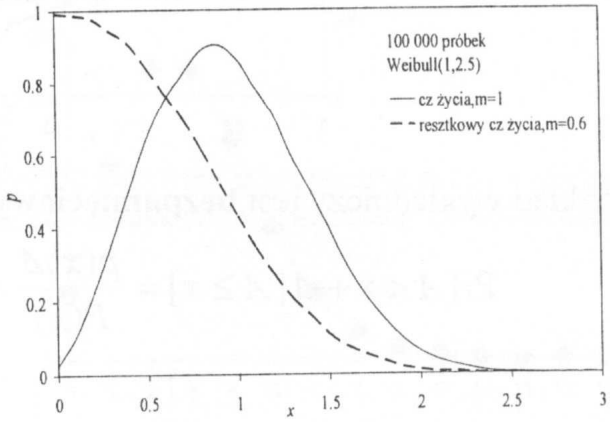
Może to być problem pasażera autobusu, abonenta dodzwaniającego się do zajętego numeru telefonicznego, nieprzerywającego zgłoszenia VIP i in.

Strumień odnowy: interwał resztkowy (2)

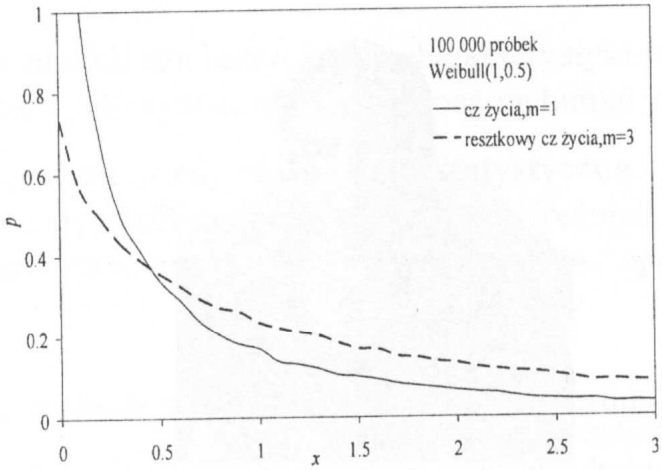


Autobus przyjeżdża średnio co godzinę. "Statystyczny pasażer" czeka pół godziny?

Znamy rozkład interwału między autobusami. Jaki przewidujemy rozkład interwału resztkowego? Przesunięty w stronę mniejszych realizacji?



Strumień odnowy: interwał resztkowy (3)

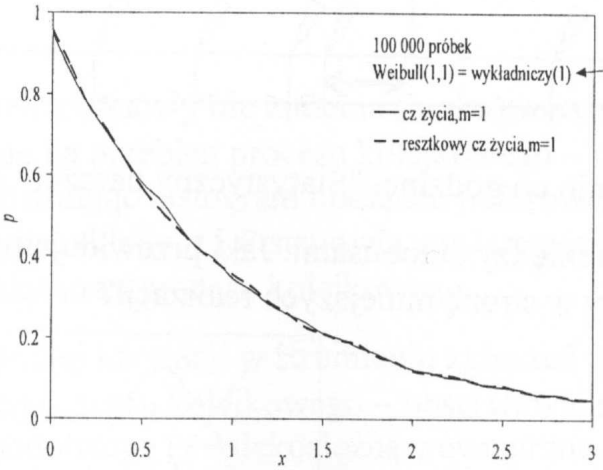


Paradoks resztkowego czasu życia:

$$\tilde{a}_{śr} = \frac{a_{śr}}{2}(1+c_a^2) \qquad c_a = \frac{\sigma_a}{a_{śr}}$$ – **współczynnik rozproszenia** rozkładu A

Jeżeli $A \sim P(x)$, to $\tilde{A} \sim \tilde{p}(x) = P(x)/a_{śr}$

Strumień odnowy: interwał resztkowy (4)



$$P(x) = e^{-x/a_{śr}}$$

Rozkład wykładniczy jest **bezpamięciowy**:

$$\Pr[A < x + \Delta \mid A \geq x] = \frac{p(x)\Delta}{P(x)} = \frac{\Delta}{a_{śr}} \quad \text{bez względu na } x$$

Strumień Poissona

Simeon-Denis Poisson (1781-1840)

matematyk, fizyk i astronom francuski

badał bezpamięciowe procesy stochastyczne – *procesy Poissona* – dobrze modelujące współczesny ruch telekomunikacyjny i komputerowy

Strumień Poissona (2)

W każdej chwili nowe zgłoszenie przybywa ze stałym prawdopodobieństwem.
W notacji Kendalla: **systemy M/...**

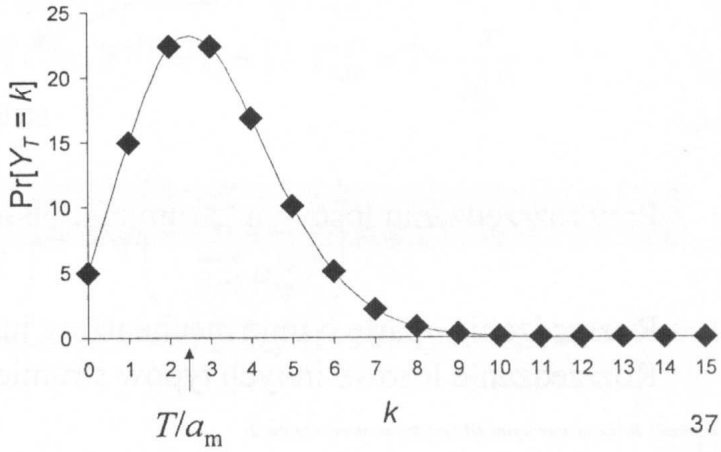
$$\Pr[\text{zgłoszenie w } (t, t+\Delta)] = \Delta / a_{śr} + o(\Delta)$$

Y_T – liczba zgłoszeń przybywających w przedziale długości T

$$Y_{T śr} = T / a_{śr}$$

$$\Pr[Y_T = k] = \frac{(T / a_{śr})^k}{k!} e^{-T / a_{śr}}$$

rozkład Poissona



Strumień Poissona (3)

Dla stacjonarnych modeli stochastycznych charakterystyki procesu obsługi mające sens średnich czasowych wyznaczamy z pomocą rachunku prawdopodobieństwa.

Stan ustalony nazywa się wtedy **równowagą statystyczną**.
Polega na ustalaniu się w czasie interesujących nas średnich, np. prawdopodobieństw stanów systemu / utraty zgłoszenia / opóźnień buforowania /...

PASTA (*Poisson Arrivals See Time Averages*): dla strumienia Poissona $p_k^+ \equiv p_k$

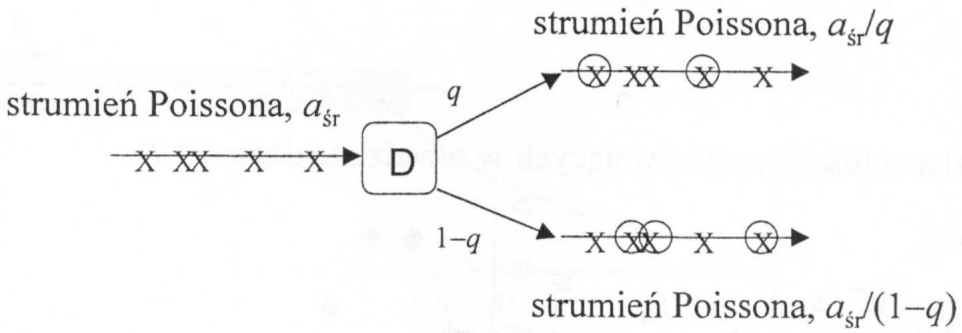
(przybywające zgłoszenia tworzące strumień Poissona "widzą" taki sam rozkład długości kolejki jak przypadkowy obserwator)

Dowód: $$\Pr[N(t) = k \mid \text{przybycie w } (t, t+\Delta)] = \frac{\Pr[N(t) = k]\frac{\Delta}{a_{śr}}}{\frac{\Delta}{a_{śr}}} = \Pr[N(t) = k]$$

M/G/S/Q: straty wskutek przepełnienia pamięci buforowej $L = p_Q^+ \equiv p_Q$.

Strumień Poissona (4)

Rozrzedzanie losowe:

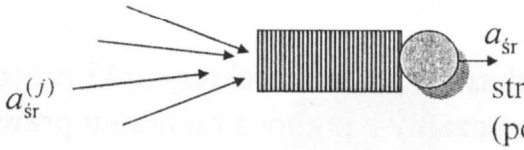


Przy rozrzedzaniu losowym strumień Poissona zachowuje swój charakter!

Rozrzedzanie oparte o inne mechanizmy już tej własności nie ma.
Rozrzedzanie losowe innych typów strumienia – także nie.

Agregacja strumieni odnowy

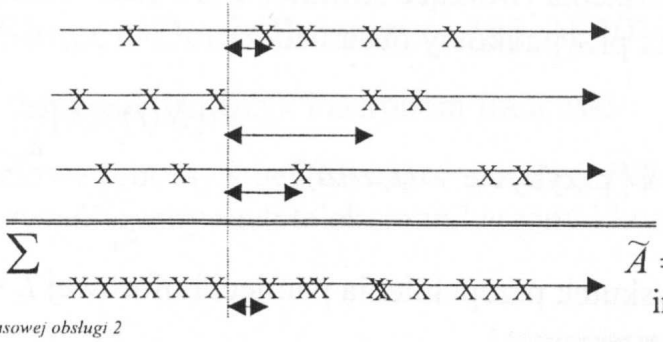
J niezależnych strumieni odnowy (internauci, abonenci, terminale, …)



strumień zagregowany (podawany na łącze szkieletowe, superkomputer / serwer WWW…)

$$\frac{T}{a_{śr}} = \sum_{j=1}^{J} \frac{T}{a_{śr}^{(j)}} \quad \text{(przy jednakowych strumieniach składowych } a_{śr} = \frac{a_{śr}^{(j)}}{J}\text{)}$$

Jaki rozkład interwału "widzi" system? Czy przynajmniej strumień odnowy?
W ogólności nie. Obliczenia b. trudne.



$\widetilde{A} = \min\{\widetilde{A}^{(1)}, \ldots, \widetilde{A}^{(J)}\}$
interwały resztkowe

Agregacja strumieni odnowy (2)

$$\widetilde{A} = \min\{\widetilde{A}^{(1)},\ldots,\widetilde{A}^{(J)}\}, \text{ zatem } \Pr[\widetilde{A} \geq x] = \prod_{j=1}^{J} \Pr[\widetilde{A}^{(j)} \geq x]$$

$J \to \infty$, ale $a_{\text{śr}} > 0$ (praktyczna sytuacja). Stąd $a_{\text{śr}}^{(j)} \to \infty$.

Dla dowolnego skończonego x: $x / \widetilde{a}_{\text{śr}}^{(j)} \to 0$ i możemy zaniedbać $Y_x^{(j)} > 1$:

$$\Pr[\widetilde{A}^{(j)} \geq x] = 1 - \Pr[Y_x^{(j)} > 0] \approx 1 - \Pr[Y_x^{(j)} = 1] \approx 1 - Y_{x,\text{śr}}^{(j)} \approx 1 - \frac{x}{a_{\text{śr}}^{(j)}}$$

Ostatecznie

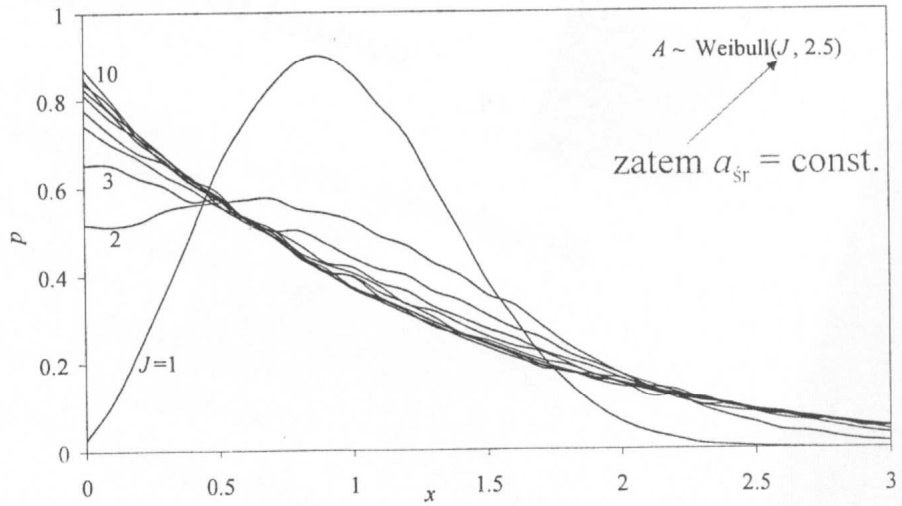
$$\Pr[\widetilde{A} \geq x] \approx \prod_{j=1}^{J}\left(1 - \frac{x}{a_{\text{śr}}^{(j)}}\right) \approx \exp\left(-\sum_{j=1}^{J}\frac{x}{a_{\text{śr}}^{(j)}}\right) = e^{-x/a_{\text{śr}}}$$

gdyż przy $x, y, z, \ldots << 1$ mamy $(1-x)(1-y)(1-z)\ldots \approx e^{-(x+y+z+\ldots)}$

Agregacja strumieni odnowy (3)

Resztkowy interwał w strumieniu zagregowanym ma rozkład wykładniczy, więc interwał też!

Strumień zagregowany jest strumieniem Poissona (**twierdzenie Palma**).



Uffffff...

Dziękuję za uwagę (na razie).

Badania operacyjne III

Systemy masowej obsługi

3. Modele markowowskie

Jerzy Konorski
jekon@eti.pg.gda.pl

Systemy markowowskie

Przypomnienie: teoria kolejek zajmuje się problemami masowej obsługi, tj. systemami i procesami, które można obserwować, mierzyć i symulować.

Analiza matematyczna systemów masowej obsługi też jest przydatna, ale tylko wówczas, gdy przynosi *proste i pouczające* wyniki.

Rozpatrzmy model *A/B/...* Czy charakterystyki procesu obsługi łatwo jest przewidzieć teoretycznie?

Tak – o ile wprowadzimy niezbędne uproszczenia:
- niezbyt drastyczne - związek z rzeczywistością!
  inaczej zarzut "szukania pod latarnią, bo tam jest jasno"
- znaczne – prostota analizy! uniwersalne walory poznawcze!

Przykład: **systemy markowowskie**.

Systemy markowowskie (2)

Mają niezwykłą i niezwykle przydatną w modelowaniu **własność Markowa.**

(Którą co prawda dzielą z ogromną liczbą występujących w rzeczywistości systemów dynamicznych: technicznych, fizycznych, społecznych, biologicznych, ekonomicznych, ...)

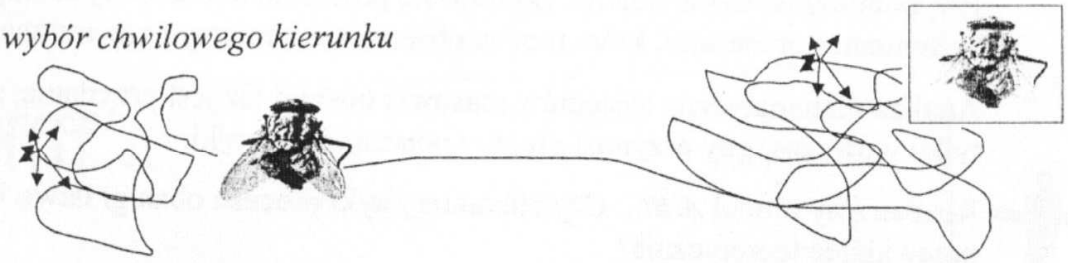
$$stan(t + \Delta) = f[stan(t), \phi]$$

losowe "pobudzenie zewnętrzne" w chwili t
na ogół zależne od bieżącego stanu,
jednak niezależne od stanów wcześniejszych

Systemy markowowskie (3)

- Mucha…

wybór chwilowego kierunku



- Tasowanie talii kart: kolejność kart (*wybór karty do przełożenia*)
- Kapitał gracza / populacja (*bieżąca stopa zysku / przyrost naturalny*)
- $trend(t + \Delta) = (1 - c) \cdot trend(t) + c \cdot \phi(t)$ (*kolejna obserwacja*)
- $udział_w_rynku(t + \Delta) = \phi \cdot udział_w_rynku(t)[1 - udział_w_rynku(t)]$ (*bieżąca jakość zarządzania*)
- topologia Internetu (*liczba nowych sieci i miejsca ich dołączenia*)

ϕ

Systemy markowowskie (4)

Andriej Aleksiejewicz Markow (1856-1922)

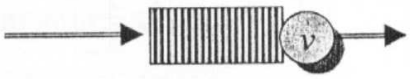
matematyk rosyjski

badał procesy przypadkowe o ograniczonej pamięci – *procesy Markowa* – doskonale modelujące wiele zjawisk przyrodniczych i technicznych

Systemy markowowskie (5)

Są nimi systemy obsługi M/M/…

strumień Poissona, $a_{śr}$
wykładniczy rozkład wymagań, $b_{śr}$



Znaczenie:

- strumień Poissona:
  - twierdzenie Palma
  - rozrzedzanie losowe
  - PASTA
  - pesymistyczne oceny charakterystyk procesów obsługi
- wykładniczy rozkład wymagań:
  - rozmowa telefoniczna
  - transfer P2P
  - …

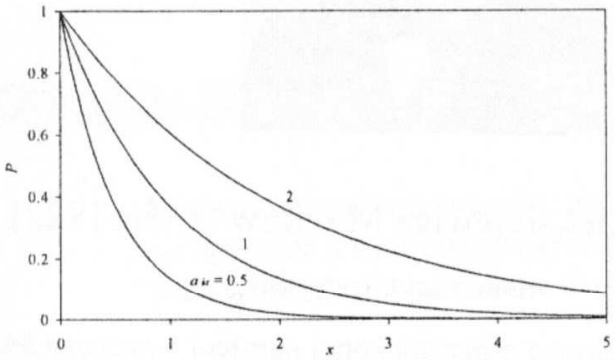
Systemy markowowskie (6)

v – (stała) wydajność procesora

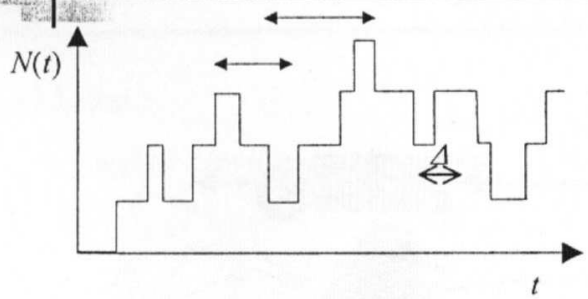
$A \sim \mathrm{M} = \text{wykładniczy}(a_{śr})$: $\Pr(A \geq x) = e^{-x/a_{śr}}$

$B \sim \mathrm{M} = \text{wykładniczy}(b_{śr})$: $\Pr(B \geq x) = e^{-x/b_{śr}}$

$\Pr[\text{czas obsługi} \geq x] = e^{-x/(b_{śr}/v)}$



System M/M/1



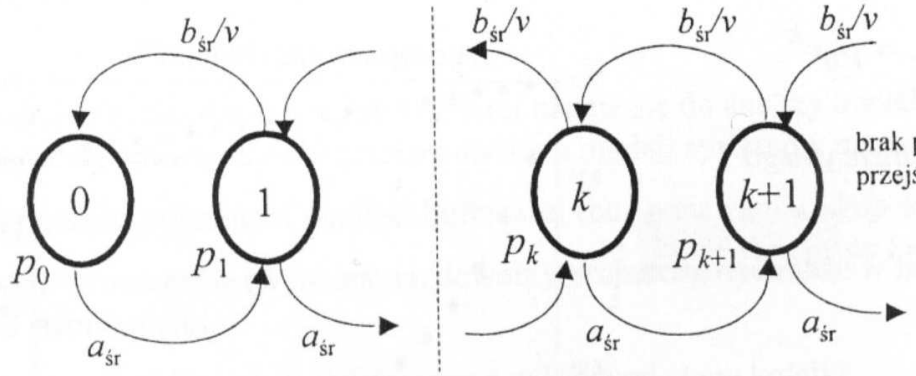
$N(t)$ – **proces urodzin i śmierci**
$a_{śr}$ – średni **interwał urodzin**
$b_{śr}/v$ – średni **interwał śmierci**

Niech $p_k(t) = \Pr[N(t) = k]$. Wiemy, że $\lim_{t \to \infty} p_k(t) = p_k$ (**równowaga statystyczna**). Podstawą wyznaczenia rozkładu (p_k) jest wypisanie **równań urodzin i śmierci.**

- Rozkład wykładniczy nie posiada "atomu" w punkcie 0
  $\Rightarrow$ Obserwujemy stan $N(t) = k$. Co może zdarzyć się między t a $t + \Delta$?
  praktycznie tylko: nic / 1 urodziny / 1 śmierć (gdy $k > 0$),
  (przejścia jedynie między sąsiednimi stanami)

- Rozkład wykładniczy jest bezpamięciowy (własność Markowa)
  $\Rightarrow$ interwał resztkowy do najbliższego zdarzenia urodziny / śmierć
  jest statystycznie taki sam jak interwał urodzin / śmierci

System M/M/1 (2)



brak pętli własnych i przejść z interwałem ∞

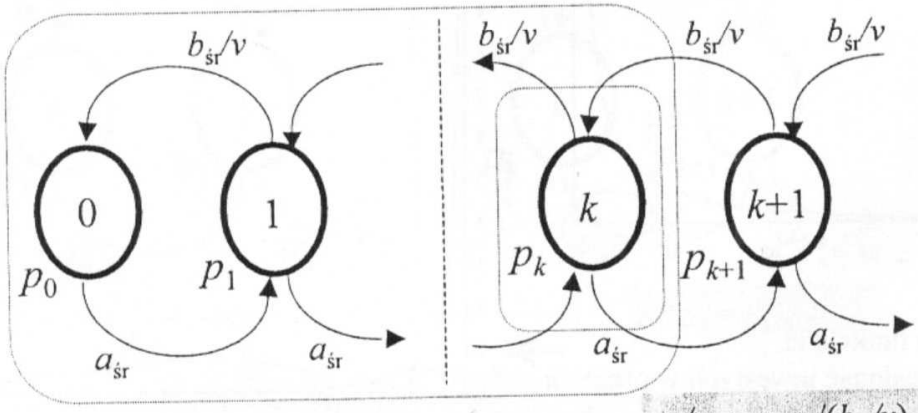
Sugestywna interpretacja "hydrauliczna" :

- ciecz = "masa prawdopodobieństwa"
- stan = zbiornik cieczy
- p_k = ciśnienie cieczy w zbiorniku
- przejście stanów = rura o oporze przepływu ~ średni interwał odpowiedniego zdarzenia

System M/M/1 (3)

W równowadze statystycznej obowiązuje bilans przepływu cieczy przez *dowolny* kontur zamknięty. (To są nasze równania urodzin i śmierci.)

Wybieramy kontury przecinające jak najmniej strzałek!



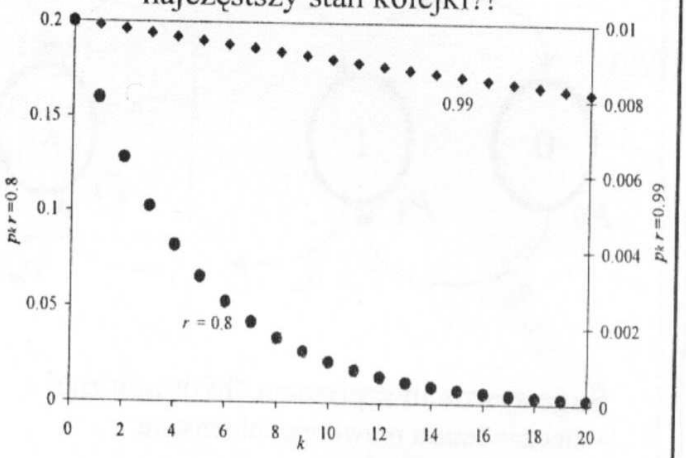
$$p_k/a_{śr} = p_{k+1}/(b_{śr}/v)$$

$$\Rightarrow \frac{p_{k+1}}{p_k} = \frac{b_{śr}}{a_{śr}v} = r$$

System M/M/1 (4)

Rozwiązanie: $p_k = p_0 r^k$

najczęstszy stan kolejki?!



Z warunku normalizacyjnego wyznaczamy p_0:

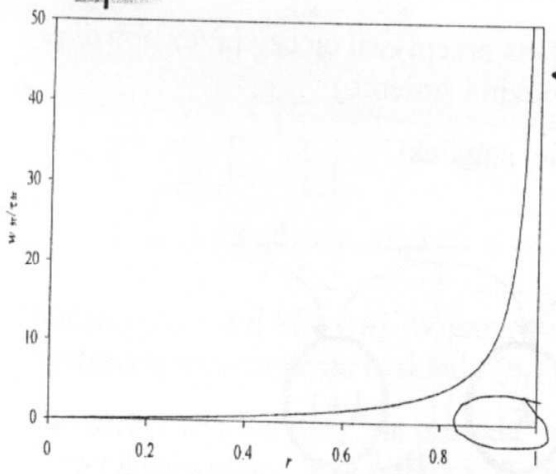
$$p_0 + p_1 + p_2 + \ldots = 1 \Rightarrow p_k = (1-r)r^k$$

$$N_{śr} = p_1 + 2p_2 + 3p_3 + \ldots = \frac{r}{1-r}$$

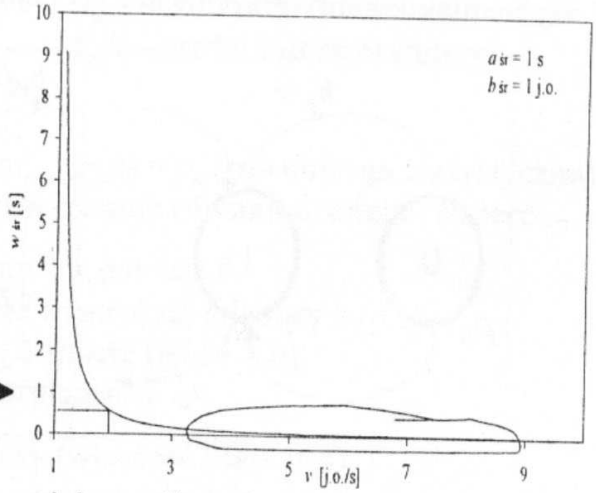
Z prawa Little'a: $d_{śr} = \dfrac{\frac{r}{1-r}}{\frac{1}{a_{śr}}} = \tau_{śr}\dfrac{1}{1-r}$ $\quad w_{śr} = d_{śr} - \tau_{śr} = \tau_{śr}\dfrac{r}{1-r}$

$r = 90\%$: zgłoszenie średnio dostaje 10% procesora

System M/M/1 (5)



$w_{śr}/\tau_{śr}$ jako funkcja r:
niebezpieczny zakres obciążeń



$w_{śr}$ jako funkcja v:
nieopłacalność inwestycji w procesor

Rozkład opóźnienia buforowania zależy od dyscypliny obsługi.
Dla FIFO wyznaczymy po wzbogaceniu aparatu matematycznego… niebawem.

Systemy M/M/...

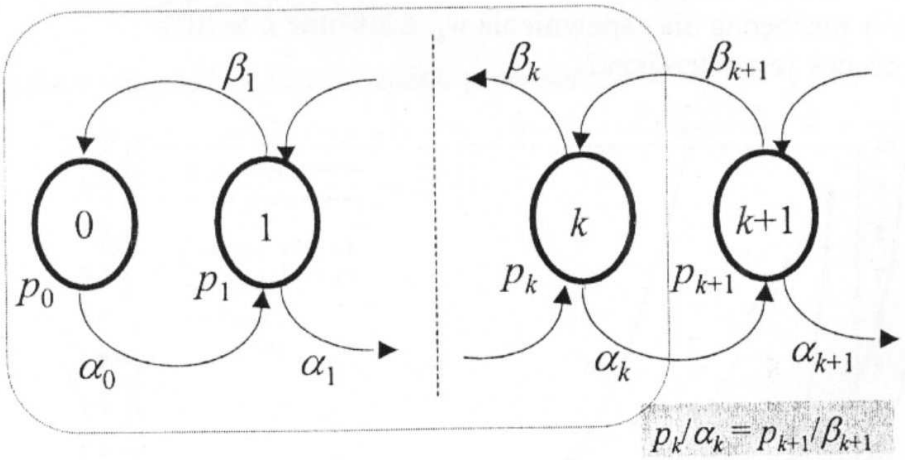
Okazuje się, że proces urodzin i śmierci nadaje się do analizy o wiele bogatszych (i bardziej realistycznych) markowowskich modeli systemów masowej obsługi, np:

- skończona pojemność pamięci buforowej (bez ponawiania, *drop-tail*)
- wiele procesorów (wymagania: dowolny procesor), być może w liczbie zależnej od stanu kolejki
- intensywność przybywania zgłoszeń zależna od stanu kolejki (źródła zgłoszeń modelowane jako inteligentne końcówki)
- różne specjalne modele – postój taksówek / wiadro tokenowe, niecierpliwość, ...

...i to praktycznie *bez komplikacji analizy*!

Systemy M/M/... (2)

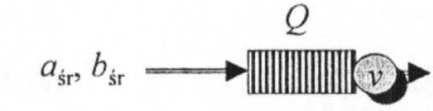
Średni interwał przejścia może zależeć od stanu wyjściowego:



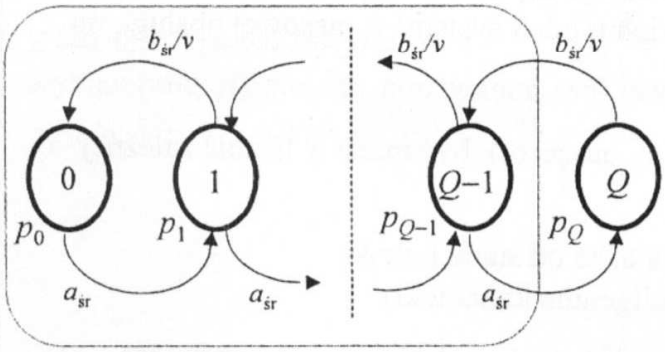
Rozwiązanie: $p_k = p_0 \dfrac{\beta_1 \ldots \beta_k}{\alpha_0 \ldots \alpha_{k-1}}$, $k = 0,1,2,\ldots$ p_0 z warunku normalizacyjnego

Można stąd wyznaczyć wszystkie interesujące charakterystyki: L, p_0, $N_{śr}$, $d_{śr}$, $w_{śr}$,...

System M/M/1/Q



$r = b_{śr}/(a_{śr}v)$



Przybycia ustają przy $k = Q$:

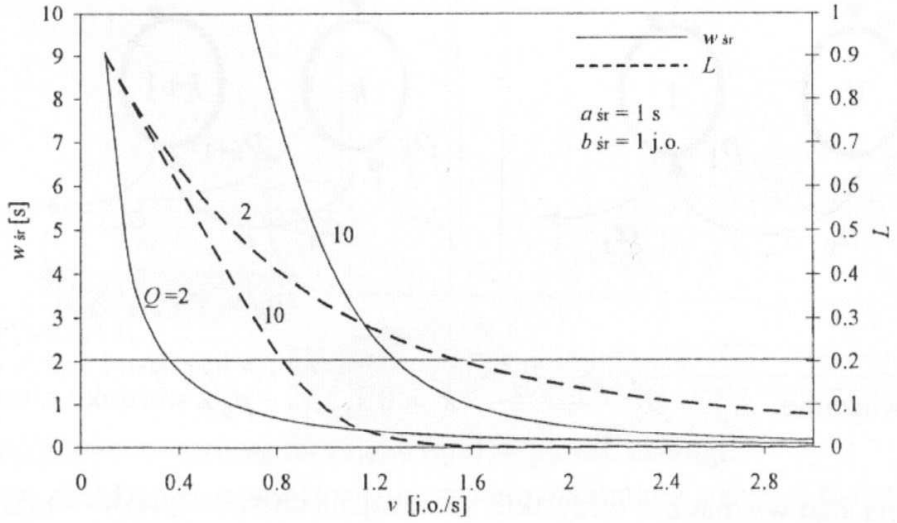
$$\alpha_k = \begin{cases} a_{śr}, \text{ gdy } k < Q \\ \infty, \text{ gdy } k = Q \end{cases} \qquad \beta_k \equiv b_{śr}/v$$

$$p_k = p_0 r^k, \; k = 0,\ldots,Q$$

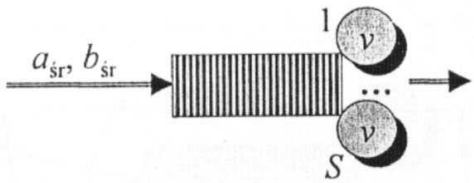
$$L = p_Q^{+} \underset{\text{PASTA}}{=} p_Q = \begin{cases} \dfrac{1-r}{1-r^{Q+1}} r^Q, \text{ gdy } r \neq 1 \\ \dfrac{1}{Q+1}, \text{ gdy } r = 1 \end{cases} \qquad d_{śr} = \frac{N_{śr}}{\frac{1-L}{a_{śr}}} \qquad w_{śr} = d_{śr} - \tau_{śr}$$

System M/M/1/Q (2)

Planowanie wydajności procesora:
- wielkość v niezbędna dla zapewnienia $w_{śr} \leq 2$ s oraz $L \leq 20\%$
- który warunek jest ostrzejszy?

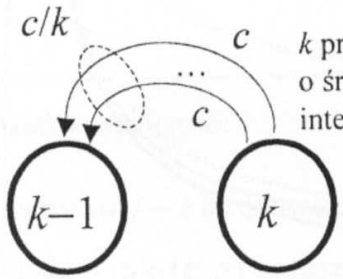


System M/M/S



$$r = \frac{b_{śr} / a_{śr}}{Sv}$$

$\alpha_k \equiv a_{śr}$



k przejść o średnich interwałach $\equiv c$

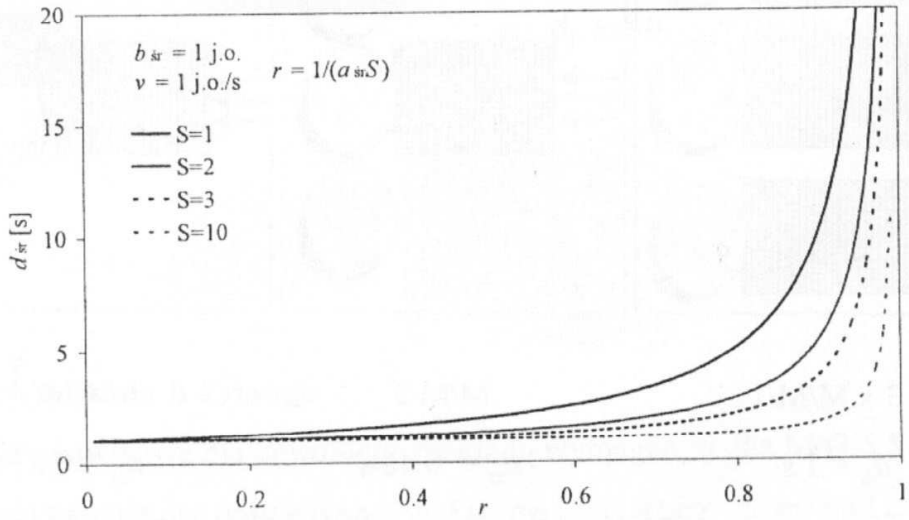
Spośród k zgłoszeń w systemie w trakcie obsługi jest $\min\{k, S\} = k_S \Rightarrow \beta_k = \dfrac{b_{śr} / v}{k_S!}$

System M/M/S (2)

$$p_k = p_0 r^k \frac{S^{k_S}}{k_S!}$$

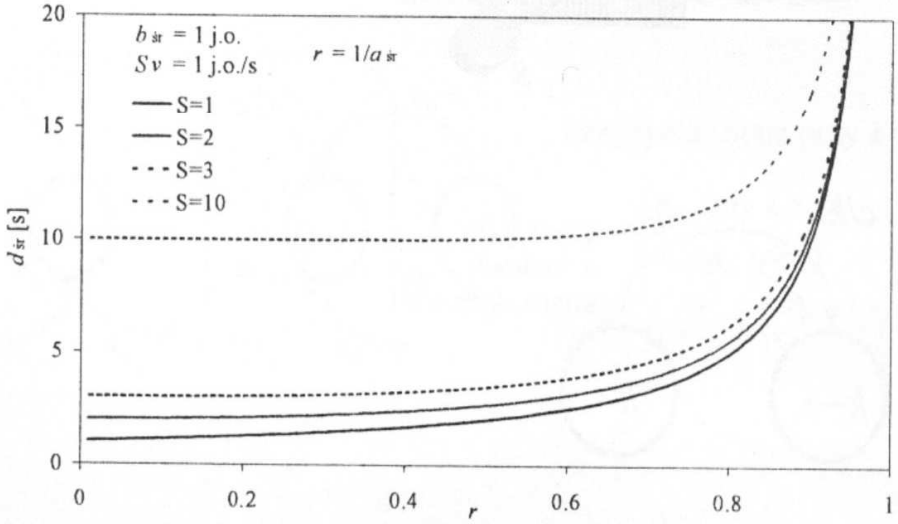
Stąd $N_{śr}$ i dalej,z prawa Little'a, $d_{śr}$.

Stała wydajność obsługi każdego procesora:



System M/M/*S* (3)

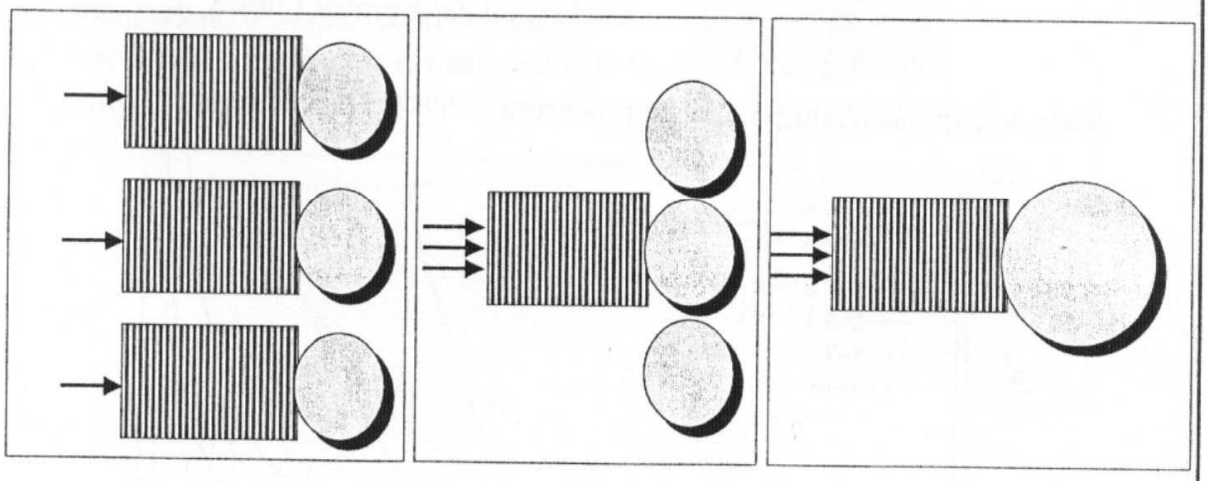
Stała *łączna* wydajność obsługi:



(Widzieliśmy już poprzednio, że rozdrobnienie podaży obsługi szkodzi!)

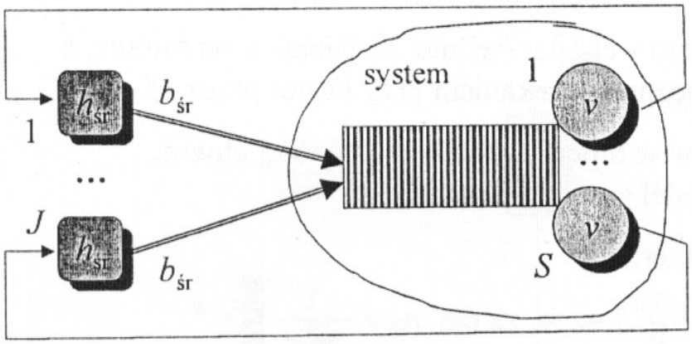
O pożytku ze współdzielenia bufora i komasacji podaży obsługi

$r = 0.667$



3 x M/M/1	M/M/3	M/M/1
$d_{śr} = 1$ s	$d_{śr} = 0.48$ s	$d_{śr} = 0.33$ s

System M/M/S//J



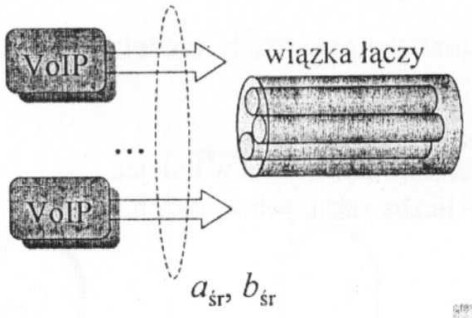
Model

- Pr[czas namysłu $\geq x$] = $e^{-x/h_{śr}}$
- populacja końcówek $J > S$

Gdy k zgłoszeń w systemie:

- w trakcie namysłu $J - k$ końcowek $\Rightarrow$ $\alpha_k = \dfrac{h_{śr}}{J-k}$
- w trakcie obsługi min$\{k, S\}$ zgłoszeń $\Rightarrow$ β_k jak dla M/M/S
- $J \to \infty$, $h_{śr} \to \infty$, $h_{śr}/J \to const.$: M/M/S

System M/M/S/S



brak oczekiwania w kolejce
v – prędkość transmisji w 1 łączu
$\rho = b_{śr}/(a_{śr}v)$
– normatywny współczynnik obciążenia
α_k, β_k, $k = 0,\ldots,S$ jak dla M/M/S

$$p_k = p_0 \frac{\rho^k}{k!}$$

$$p_S = L = \frac{\dfrac{\rho^S}{S!}}{\displaystyle\sum_{k=0}^{S} \frac{\rho^k}{k!}}$$

- słynna **formuła B Erlanga**
- magia: prawdziwa dla *dowolnego* rozkładu wymagań, tj. dla M/G/S/S (!)
- istnieją kalkulatory online (`www.voip-calculator.com/calculator/`)

System M/M/∞

Do bardzo dużego hipermarketu wchodzi średnio 20 klientów na minutę, a każdy średnio robi zakupy (łącznie z czekaniem przy kasie) przez 15 minut.

Oblicz rozkład liczby klientów w hipermarkecie w dowolnej chwili, przyjmując markowowski model systemu obsługi.

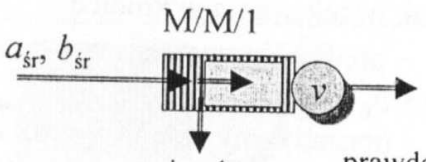
$a_{śr} = 3$ s, $b_{śr}/v = 900$ s, $r = 300$ erlang

Szczególny przypadek M/M/*S*/*S*: $S \to \infty$, zatem $p_0 = \dfrac{1}{\sum_{k=0}^{\infty} \dfrac{\rho^k}{k!}} = e^{-\rho}$

$p_k = e^{-\rho} \dfrac{\rho^k}{k!}$ – rozkład Poissona (!)

$N_{śr} = \rho = 300$ (to wiedzieliśmy już przedtem z prawa Little'a)

Niecierpliwe zgłoszenia



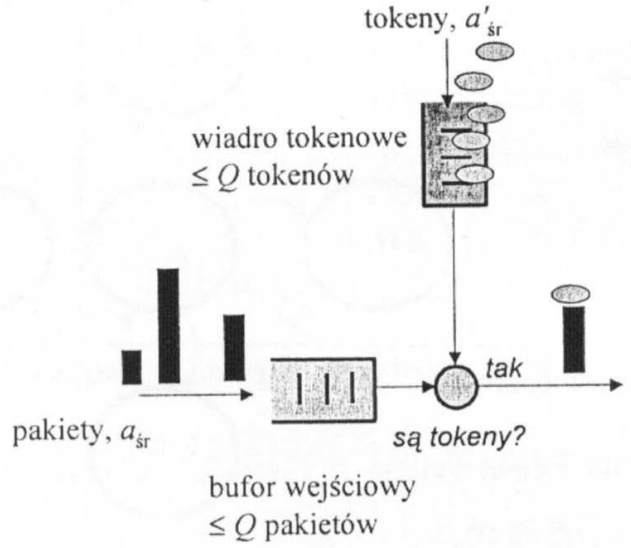
prawdopodobieństwo pozostania w kolejce $= 1/(k+1)$, k – liczba zastanych zgłoszeń

Przyjmując model markowowski, oblicz frakcję ucieczek.

$\alpha_k = a_{śr}(k+1), \quad \beta_k = b_{śr}/v \quad \Rightarrow \quad (p_k)$ – rozkład Poissona (!)

$$L = 1 - \left(p_0 + p_1 \frac{1}{2} + p_2 \frac{1}{3} + \ldots + p_k \frac{1}{k+1} + \ldots \right)$$

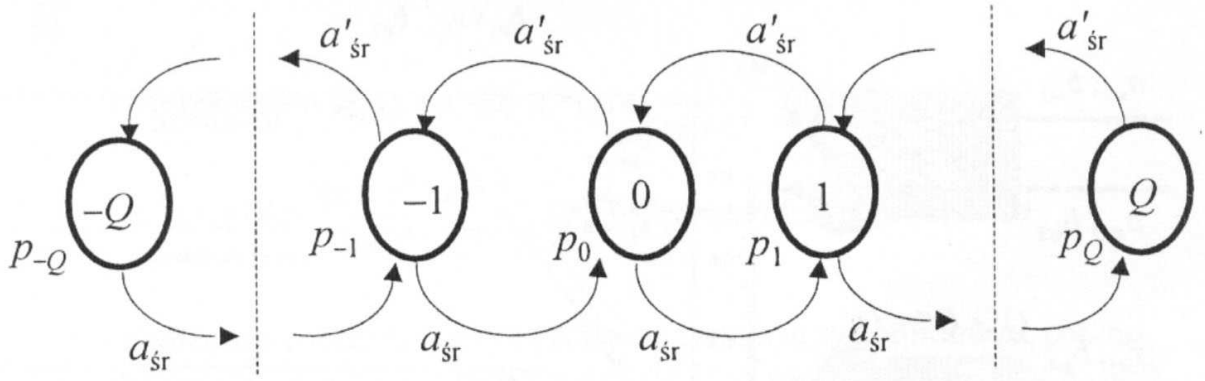
Wiadro tokenowe



Przyjmując model markowowski, dla stanu równowagi statystycznej oblicz prawdopodobieństwo, że tokeny czekają na pakiety / pakiety czekają na tokeny.

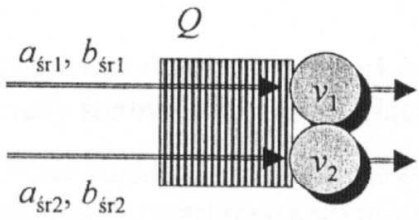
Wiadro tokenowe (2)

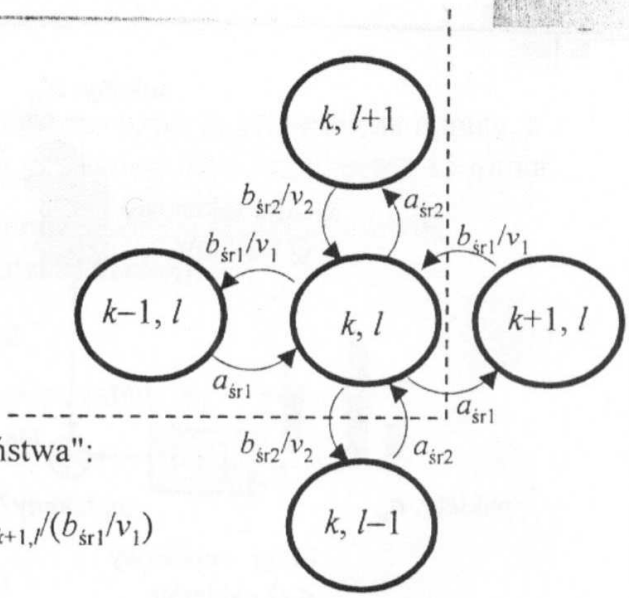
Nowy pakiet zwiększa stan systemu, nowy token zmniejsza:



Czy przy $Q = \infty$ wystąpi równowaga statystyczna?

Procesory dedykowane





Bilans przepływu "masy prawdopodobieństwa":

$$p_{k,l}/a_{śr1} + p_{k,l}/(b_{śr2}/v_2) = p_{k,l-1}/a_{śr2} + p_{k+1,l}/(b_{śr1}/v_1)$$

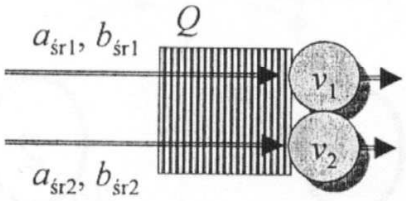
Rozwiązanie przez odgadnięcie:

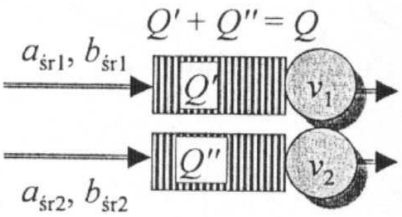
$$p_{k,l} = p_{0,0} r_1^k r_2^l \quad \text{gdzie} \quad r_1 = \frac{b_{śr1}}{a_{śr1} v_1}, r_2 = \frac{b_{śr2}}{a_{śr2} v_2}$$

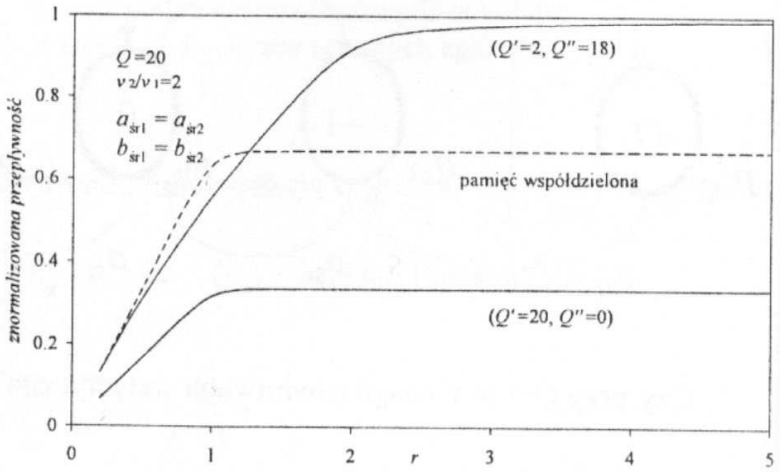
O pożytku z rozdziału pamięci buforowej

Znormalizowana przepływność:

$$\frac{\frac{1}{b_{śr1}/v_1}\left(1-\sum_{l=0}^{Q} p_{0,l}\right) + \frac{1}{b_{śr2}/v_2}\left(1-\sum_{k=0}^{Q} p_{k,0}\right)}{\frac{1}{b_{śr1}/v_1} + \frac{1}{b_{śr2}/v_2}}$$







Transformata Laplace'a

Wyznaczanie *rozkładu prawdopodobieństwa* opóźnień jest trudniejszym zadaniem, wymaga dodatkowego aparatu matematycznego – **transformata Laplace'a**.

Dla zmiennej losowej X o gęstości prawdopodobieństwa $p(x)$ definiujemy:

$$X^*(s) = \int_0^\infty e^{-sx} p(x)dx$$

Np. dla zmiennej losowej o rozkładzie wykładniczym $p(x) = e^{-x/c}/c$: $X^*(s) = \dfrac{1}{cs+1}$

Jeżeli zmienne losowe X_1 i X_2 są statystycznie niezależne i $X = X_1 + X_2$, to

$$X^*(s) = X_1^*(s)X_2^*(s).$$

Dla sumy 2, 3, … zmiennych losowych o rozkładach wykładniczych :

$$X^*(s) = \left(\frac{1}{cs+1}\right)^2, \left(\frac{1}{cs+1}\right)^3, \ldots$$

Transformata Laplace'a (2)

Mając $X^*(s)$, jak znaleźć $p(x)$? Czyli jak **odwracać** transformatę Laplace'a?

- Można analitycznie, np. gdy $X^*(s) = \left(\dfrac{1}{cs+1}\right)^M$ (M całkowite), to

$$p(x) = \frac{(x/c)^{M-1} e^{-x/c}}{c(M-1)!} \quad (x \geq 0) \quad \textbf{rozkład Erlanga rzędu } \boldsymbol{M}$$

- Pomocne są obszerne tablice i software rachunku symbolicznego, por. np. `www.educypedia.be/education/calculatorsalgebra.htm`
- Dla pewnych typów $X^*(s)$ powyższe zawodzi. Zostają algorytmy numeryczne. Nie ma jednego uniwersalnego, bo problem jest numerycznie niestabilny! Bezpośrednie korzystanie ze wzoru na transformatę odwrotną jest kłopotliwe, gdyż wymaga operowania na liczbach zespolonych.

Transformata Laplace'a (3)

www.pe.tamu.edu/blasingame/data/P620_reference/P620_Lectures_(pdf)/P620_Mod1_Math/P620_Mod1_ML_05_LaplaceTrans.pdf

M/M/1 FIFO: rozkład opóźnień systemowych

Opóźnienie systemowe zgłoszenia zastającego k zgłoszeń w chwili przybycia składa się z $k + 1$ czasów obsługi (w tym jednego resztkowego, gdy $k > 0$) o rozkładach wykładniczych $P(x) = e^{-x/(b_{śr}/v)}$.

Zatem jego transformata Laplace'a: $\left(\dfrac{1}{\frac{b_{śr}}{v}s+1}\right)^{k+1}$

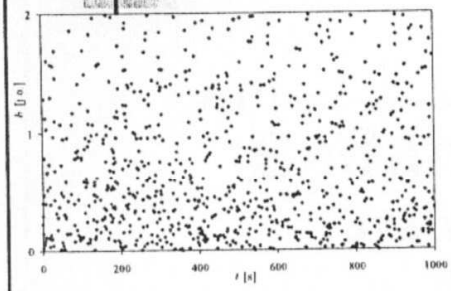
Działa PASTA i argument skoków ± 1, zatem $p_k^+ \equiv p_k$.

Uśredniając na różne liczby zastanych zgłoszeń:

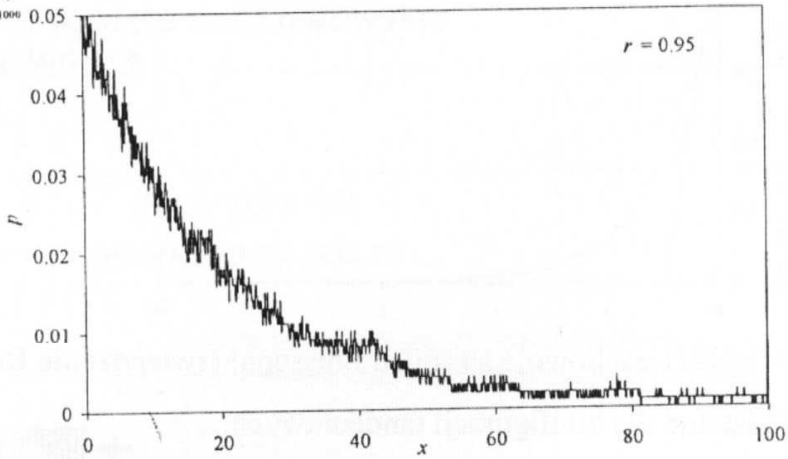
$$D^*(s) = \sum_{k=0}^{\infty} p_k \left(\frac{1}{\frac{b_{śr}}{v}s+1}\right)^{k+1} = \sum_{k=0}^{\infty}(1-r)r^k\left(\frac{1}{\frac{b_{śr}}{v}s+1}\right)^{k+1} = \frac{1}{\frac{\tau_{śr}}{1-r}s+1}$$

Rozkład wykładniczy o znanej już wcześniej wartości średniej!

M/M/1 FIFO: rozkład opóźnień systemowych (2)



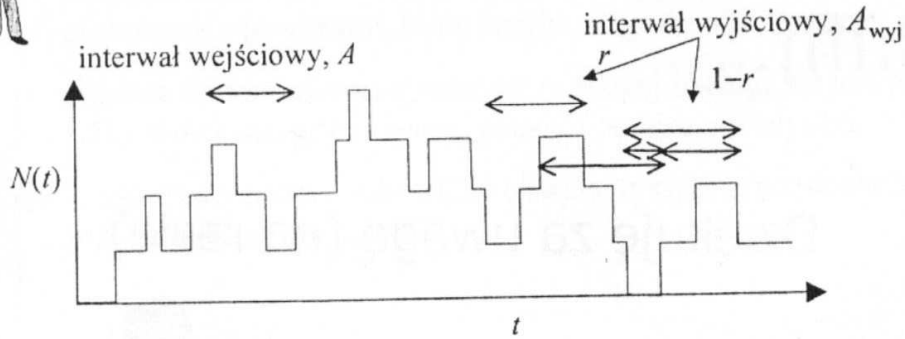
strumień M/M, $a_{śr} = b_{śr} = 1$



histogram
opóźnień systemowych

M/M/1: strumień wyjściowy

Jak wygląda strumień *wyjściowy* z M/M/1?



Działa PASTA i argument skoków ±1, zatem $p_k \equiv p_k^+ \equiv p_k^-$. Czyli $p_0^- = 1 - r$.

A_{wyj} jest z prawdopodobieństwem r jednym czasem obsługi, zaś z prawd. $1-r$ sumą jednego czasu obsługi i resztkowego interwału wejściowego.

Wyznaczenie rozkładu A_{wyj} jest nietrudne, gdy już się opanowało transformatę Laplace'a.

Badania operacyjne III

Systemy masowej obsługi 4. Modele niemarkowowskie

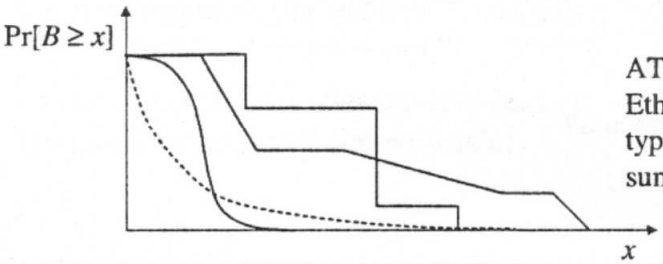
Jerzy Konorski
jekon@eti.pg.gda.pl

Motywacja

Dalej rozważamy strumień zgłoszeń Poissona: $A \sim \Pr[A \geq x] = e^{-x/a_{\text{śr}}}$.

O wielu jego zaletach w modelowaniu mowa była już wcześniej:
- losowe rozrzedzanie
- twierdzenie Palma
- PASTA
- pesymistyczne oszacowania dla strumieni odnowy

Jednak w wielu praktycznie spotykanych systemach masowej obsługi wymagania zgłoszeń posiadają rozkłady znacznie odbiegające od wykładniczego!
Prowadzi to do modeli **M/G/...** – ogólny (dowolny) rozkład wymagań.



ATM,
Ethernet,
typowe zakresy wymagań,
sumy wielu wymagań cząstkowych,...

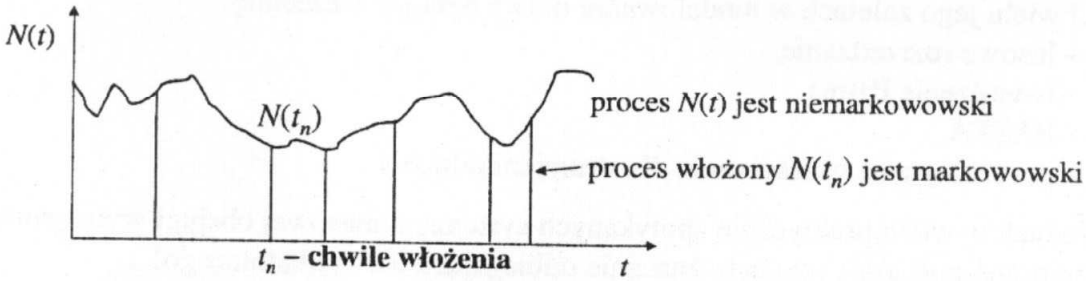
M/G/1 – analiza

strumień Poissona, $a_{śr}$
dowolny rozkład wymagań, $b_{śr}$

- **M/G/1**: 1 procesor, nieskończona pamięć buforowa
- $B \sim P(x)$, dowolny rozkład o gęstości $p(x)$ i średniej $b_{śr}$
- $r = b_{śr}/(a_{śr}v)$, v – wydajność procesora
- $N(t)$ nie ma własności Markowa
- ma tę własność proces dwuwymiarowy [$N(t)$, resztkowy_czas_obsługi (t)], lecz jego analiza jest znacznie bardziej skomplikowana!

M/G/1 – analiza (2)

Sytuacja nie jest beznadziejna: metoda **włożonego procesu Markowa**.



Jak wybierać chwile włożenia?

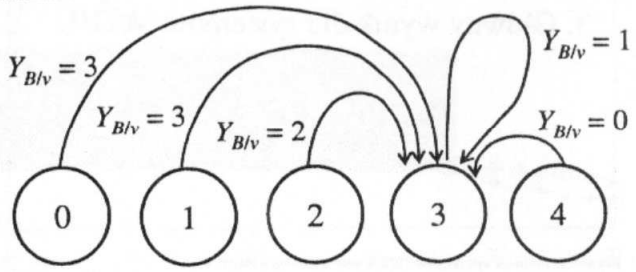
M/G/1 – analiza (3)

Dla procesu obsługi w systemie M/G/1 wybór jest oczywisty:

$t_n \equiv t_n^- + 0$ tj. tuż po opuszczeniu systemu przez kolejne zgłoszenie
$N(t_n)$ = # zgłoszeń pozostawionych przez n-te zgłoszenie

Niech $Y_{B/v}$ = # zgłoszeń przybyłych w trakcie obsługi zgłoszenia (trwającej B/v).
Oczywiście $Y_{B/v,\text{śr}} = (b_{\text{śr}}/v)/a_{\text{śr}} = r$.

$$N(t_n) = \max[N(t_{n-1}) - 1, 0] + Y_{B/v}$$ **(równanie stochastyczne)**



M/G/1 – analiza (4)

Równanie stochastyczne należy "przetłumaczyć" na zależność pomiędzy rozkładami występujących w nim zmiennych losowych.

Najwygodniej to zrobić dla równowagi statystycznej:

$$p_k^- = \lim_{n\to\infty} \Pr[N(t_n) = k]$$

Zauważmy, że

$$p_k^- \equiv p_k^+ \equiv p_k$$

$Q = \infty$ i argument przyrostów ± 1 PASTA

czyli w równowadze statystycznej rozkład $N(t_n)$ jest zarazem rozkładem $N(t)$ dla przypadkowych chwil próbkowania!

M/G/1 – analiza (5)

Aparat obliczeniowy transformaty Laplace'a pozwala uzyskać (p_k) w postaci zamkniętej, choć nie tak strawnej jak dla M/M/...

1. Obliczamy transformatę Laplace'a rozkładu wymagań: $B^*(s) = -\int_0^\infty e^{-sx} dP(x)$
2. Poszukiwany rozkład prawdopodobieństwa wyobrażamy sobie w postaci wielomianu symbolicznej zmiennej z:

$$p_0 + p_1 z + p_2 z^2 + p_3 z^3 + \ldots$$

Jego pochodna dla $z = 1$ to $N_{śr}$ – łatwo sprawdzić.

3. Główny wynik dla systemów M/G/1:

$$p_0 + p_1 z + p_2 z^2 + p_3 z^3 + \ldots = (1-r)\frac{1-z}{1-z/B^*(\frac{1-z}{a_{śr}v})}$$

M/G/1 – przykłady analizy

Dla sprawdzenia obliczmy rozkład p_k dla M/M/1, który przedtem otrzymaliśmy metodami urodzin i śmierci:

$$P(x) = e^{-x/b_{śr}}, \quad B^*(s) = \frac{1}{b_{śr}s+1},$$

$$(1-r)\frac{1-z}{1-z/B^*(\frac{1-z}{a_{śr}v})} = (1-r)\frac{1}{1-rz} = (1-r) + (1-r)rz + (1-r)r^2z^2 + \ldots$$

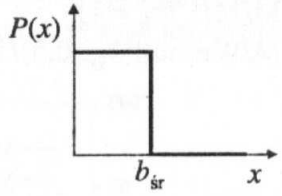
rozwijamy w szereg potęgowy

Zatem $p_k = (1-r)r^k$, zgodnie z analizą markowowską.

M/G/1 – przykłady analizy (2)

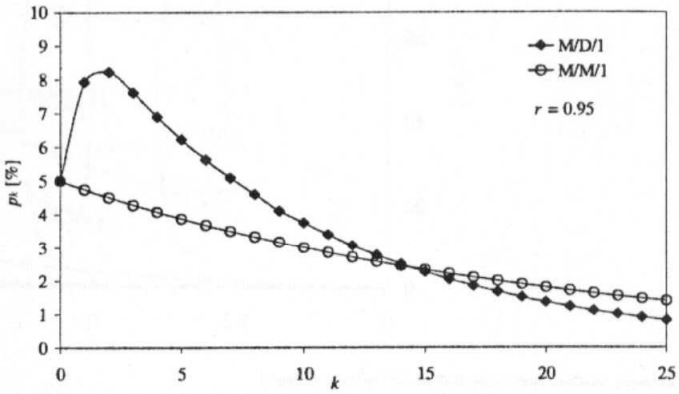
Dla M/D/1 analiza markowowska zawodzi. Metoda transformaty Laplace'a – nie:

$$P(x) = \begin{cases} 1, \text{gdy } x < b_{śr} \\ 0, \text{gdy } x \geq b_{śr} \end{cases}, \quad B^*(s) = e^{-b_{śr}s},$$



$$(1-r)\frac{1-z}{1-z/B^*(\frac{1-z}{a_{śr}v})} = (1-r)\frac{1-z}{1-ze^{r(1-z)}}$$

Numeryczne rozwijanie w szereg nie sprawia trudności – por. `www.educypedia.be/education/calculatorsalgebra.htm` bądź liczne narzędzia softwarowe.



Wzory Pollaczka-Chinczyna

Różniczkując w punkcie $z = 1$ główny wynik dla systemów M/G/1, otrzymujemy $N_{śr}$ i następnie $w_{śr}$ z prawa Little'a:

M/M/1

$$w_{śr} = \frac{1+c_b^2}{2}\tau_{śr}\frac{r}{1-r}$$

$c_b = \sigma_b/b_{śr}$ – współczynnik rozproszenia rozkładu wymagań

Walory poznawcze:

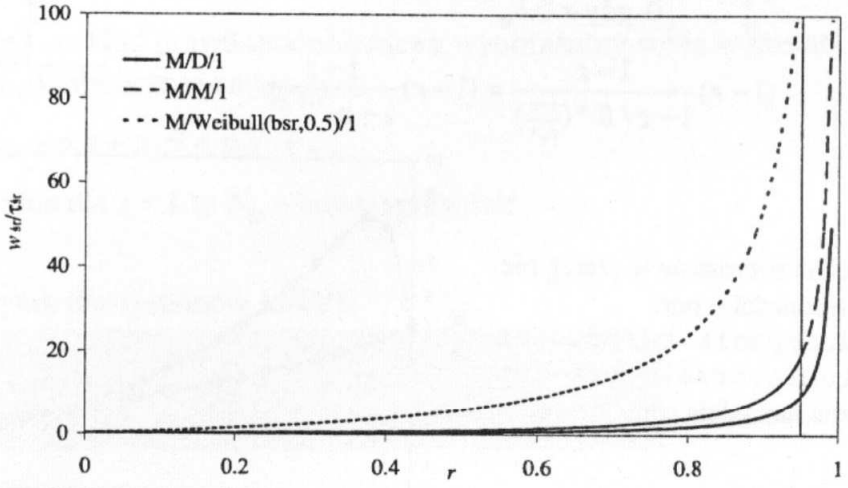
- magiczne $1 - r$ w mianowniku, to nie przypadek
- wpływ *kształtu* rozkładu wymagań tylko poprzez odchylenie standardowe!
- rozproszenie rozkładu wymagań pogarsza średnie opóźnienie
- z wyrażenia na średni resztkowy czas życia wynika alternatywny wzór, niezwykle sugestywny:

$$w_{śr} = \tau_{śr}\frac{r}{1-r}$$

np. $r = 0.9$: 90% szans na odczekanie 10*dokończeń obsługi

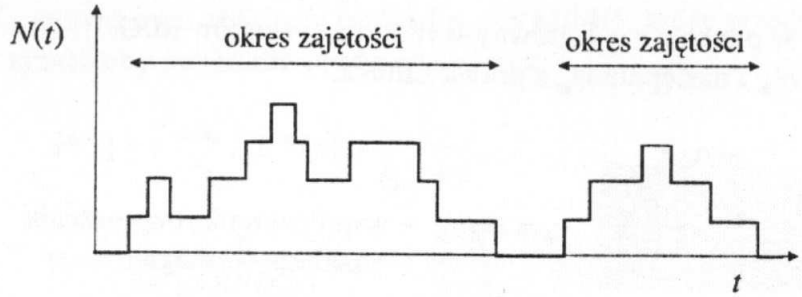
Wzory Pollaczka-Chinczyna (2)

Np. dla $r = 0.95$ porównaj
- M/D/1 ($c_b = 0$)
- M/M/1 ($c_b = 1$)
- M/Weibull($b_{śr}$, 0.5)/1 ($c_b \approx 3.317$) !



Okres zajętości

Zagadnienie, które trudno rozstrzygnąć eksperymentalnie:



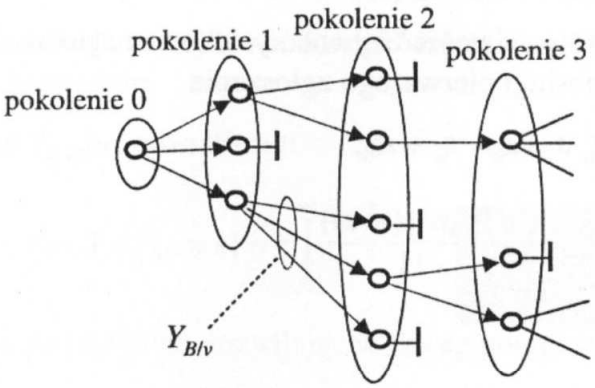
Czy rozpoczęty okres zajętości kiedykolwiek zakończy się?

Niekoniecznie – z pewnym prawdopodobieństwem zależnym od r:

ω = Pr[okres zajętości zakończy się]

Okres zajętości (2)

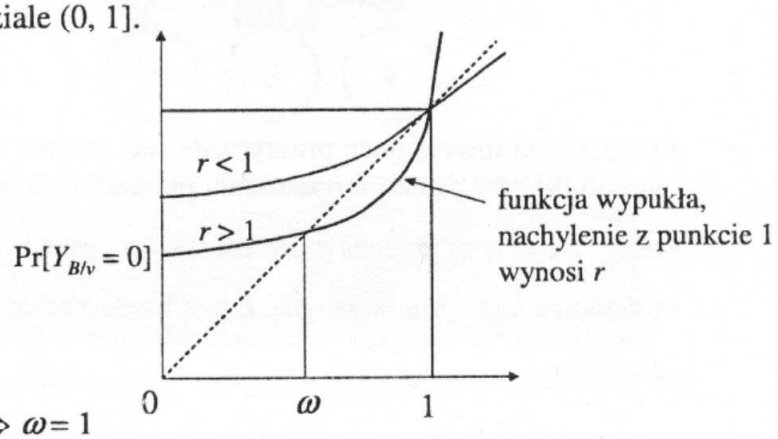
ω = Pr[wymarcie populacji, w której każdy osobnik generuje $Y_{B/v}$ potomków niezależnie od pozostałych osobników (a sam przy tym umiera)].



Okres zajętości (3)

$$\omega = B^*\left(\frac{1-\omega}{a_{śr}v}\right)$$

Z teorii dynamiki populacji wynika, że chodzi o *najmniejszy* pierwiastek tego równania w przedziale (0, 1].



$r > 1 \Rightarrow \omega < 1$, $r < 1 \Rightarrow \omega = 1$

$r < 1 \Rightarrow \omega = 1$ – zaskakujące, bo $r = Y_{B/v,śr}$
populacja ze średnią rozrodczością 1/1, wymrze z prawdopodobieństwem 1

Okres zajętości (4)

Gdy $r < 1$, *średni* okres zajętości $z_{śr}$ jest skończony i łatwo go wyliczyć:

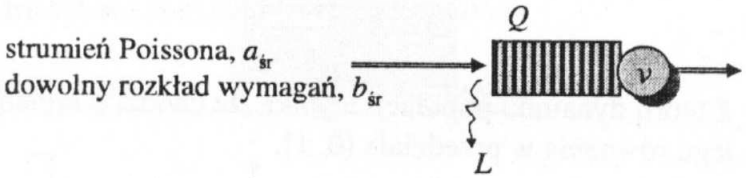
$z_{śr}$ = średni czas obsługi pierwszego zgłoszenia

+ suma średnich okresów zajętości "wywołanych" przez zgłoszenia przybyłe w trakcie obsługi pierwszego zgłoszenia

$z_{śr} = \tau_{śr} + Y_{B/v,śr} z_{śr} = \tau_{śr} + (\tau_{śr}/a_{śr}) \cdot z_{śr} = \tau_{śr} + r z_{śr}$

$$z_{śr} = \frac{\tau_{śr}}{1-r}$$

M/G/1/Q



- najbardziej interesujące praktycznie jest wyznaczanie L – frakcji zgłoszeń utraconych wskutek przepełnienia pamięci buforowej
- brak zamkniętej postaci rozwiązania, jedynie obliczenia numeryczne
- przewaga nad symulacją, gdy L jest bardzo małe

M/G/1/Q (2)

Punktem wyjścia znów jest nasze równanie stochastyczne:

$$N(t_n) = \min\{\max[N(t_{n-1}) - 1, 0] + Y_{B/v}, Q\}$$

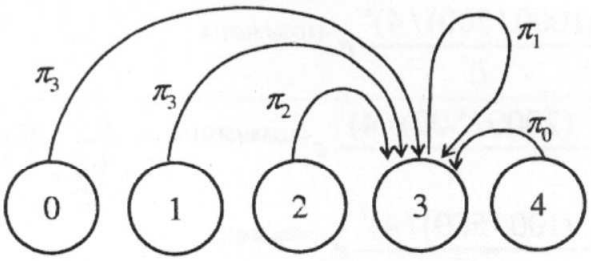
"Tłumaczenia" na rozkład N nie dokonujemy metodą transformaty Laplace'a, lecz bezpośrednio.

Rozkład $Y_{B/v}$ już poznaliśmy:

$$\pi_k = \Pr[Y_{B/v} = k] = -\int_0^\infty \frac{[(x/v)]/a_{śr}]^k}{k!} e^{-[(x/v)]/a_{śr}} dP(x)$$

i można go otrzymać, rozwijając w szereg potęgowy funkcję $B^*\left(\frac{1-z}{a_{śr}v}\right)$.

M/G/1/Q (3)



Patrząc na znajomy rysunek, wypisujemy równania z niewiadomymi p_k^-:

$$p_k^- = p_0^-\pi_k + p_1^-\pi_k + p_2^-\pi_{k-1} + \ldots + p_{k+1}^-\pi_0,\ k = 0,\ldots,Q-2$$

$$p_0^- + \ldots + p_{Q-1}^- = 1$$

Dalej mamy: $p_k \equiv p_k^+$ (PASTA) oraz $p_k^- \equiv p_k^+/(1-p_Q^+)$ (argument skoków ± 1). Stąd

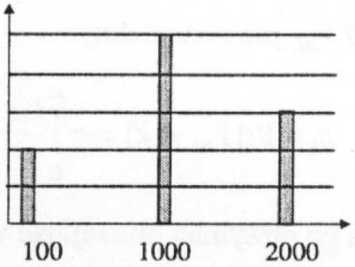
$$L = p_Q = 1 - \frac{1}{r + p_0^-}$$

M/G/1/*Q* – przykład

Zgłoszenia: strumień Poissona, $a_{śr} = 4$ s

typ	udział [%]	wymaganie [j.o.]
`http get`	30	1000
`telnet open`	50	2000
`snmp SetRequest`	20	100

Wydajność procesora = 500 j.o./s



Czy dla zadanego $L = 10\%$ wystarczy $Q = 3$?

M/G/1/*Q* – przykład (2)

Rozkład $Y_{B/v}$:

$$\pi_i = 30\% \cdot \frac{(1000/500)/4)^i}{i!} e^{-(1000/500)/4}$$

$$+ 50\% \cdot \frac{(2000/500)/4)^i}{i!} e^{-(2000/500)/4}$$

$$+ 20\% \cdot \frac{(100/500)/4)^i}{i!} e^{-(100/500)/4}$$

$\pi_0 = 0.5561$
$\pi_1 = 0.2844$

Współczynnik obciążenia: $r = 30\% \cdot \frac{1000}{4 \cdot 500} + 50\% \cdot \frac{2000}{4 \cdot 500} + 20\% \cdot \frac{100}{4 \cdot 500} = 0.65$

Równania:

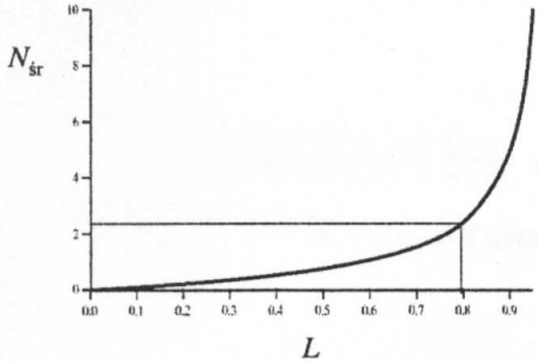
$$p_0^- = p_0^- \pi_0 + p_1^- \pi_0$$

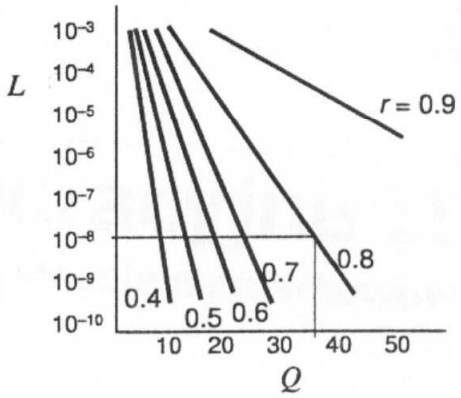
$$p_1^- = p_0^- \pi_1 + p_1^- \pi_1 + p_2^- \pi_0$$

$$p_0^- + p_1^- + p_2^- = 1$$

$$L = 1 - \frac{1}{r + p_0^-} = 7.6\%$$

M/G/1/Q – projektowanie





- określ akceptowalne $N_{śr}$ i L (np. 2.5 i 10^{-8})
- odczytaj dopuszczalne r z lewego wykresu (dla M/D/1 przy pomijalnie małym L)
- odczytaj wymagane Q z prawego wykresu (dla M/D/1/Q)

Uffff...

Dziękuję za uwagę (definitywnie).

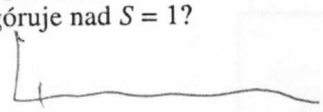
Jerzy Konorski

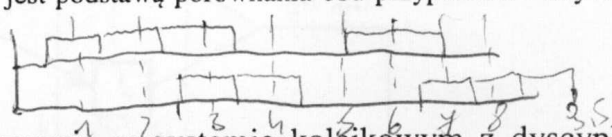
Badania Operacyjne / SMO_1 – Ćwiczenia

Zadanie 1

Narysuj przebieg procesów kolejkowych $N(t)$ oraz $U(t)$ = **praca do zakończenia** w systemie masowej obsługi przyjmującego strumień zgłoszeń: (t_n^+) = (0.5, 2, 5, 6.5), (b_n) =(3, 2, 2, 3) dla dwóch przypadków: a) $S = 2$ procesory o wydajności $v = 1$ j.o./s każdy (pełna dostępność, obsługa umiejscowiona), b) $S = 1$ procesor o wydajności $v = 2$ j.o./s. Porównaj te przypadki.

Uwaga na nachylenie opadających części wykresu! Co jest podstawą porównania obu przypadków? Czy $S = 2$ pod jakimś względem góruje nad $S = 1$?





Zadanie 2

Porównaj średnie opóźnienie systemowe zgłoszenia w systemie kolejkowym z dyscypliną obsługi FIFO oraz *Round Robin* z **kwantem obsługi** 2 s (niewykorzystany kwant obsługi przechodzi na następne zadanie). Wydajność procesora wynosi 1 j.o./s. Trzy zgłoszenia: X, Y i Z, o wymaganiach odpowiednio 7 j.o., 1 j.o. i 3 j.o., przybywają jednocześnie i ustawiają się w kolejności a) XYZ, b) YZX, c) XZY.

Czy z porównania opóźnień *poszczególnych* zgłoszeń wynikają ogólne cechy FIFO i RR?

Zadanie 3

1-procesorowy system masowej obsługi posiada pamięć buforową o pojemności $Q < \infty$ zgłoszeń i pracuje ze współczynnikiem obciążenia $r > 1$. W takim systemie pomimo wzrostu Q frakcja utraconych zgłoszeń L nigdy nie spadnie poniżej pewnego progu – jaką ma on wartość?

Należy przypomnieć sobie o równaniu ciągłości przepływu. Jak często procesor jest bezczynny, gdy $Q \to \infty$?

Zadanie 4

Każda z 50 końcówek sieciowych dołączonych do wspólnego nadajnika pracuje w następujący sposób: faza namysłu trwa średnio $^2/_3$ s, po czym generowane jest zgłoszenie; w 80% przypadków jest to wiadomość (średnio 1000 B), zaś w 20% – raport diagnostyczny (średnio 160 B). Nadajnik pracuje w trybie transmisji półdupleksowej z prędkością 1 Mb/s i w każdej sekundzie średnio przez 750 ms powinien być przełączony na odbiór (w tym czasie jest niedostępny dla naszych zgłoszeń). Jaka będzie średnia frakcja zgłoszeń utraconych?

Dane pozwalają na identyfikację $a_{śr}$, $b_{śr}$ oraz współczynnika zajętości procesora. Należy teraz powrócić do równania ciągłości przepływu.

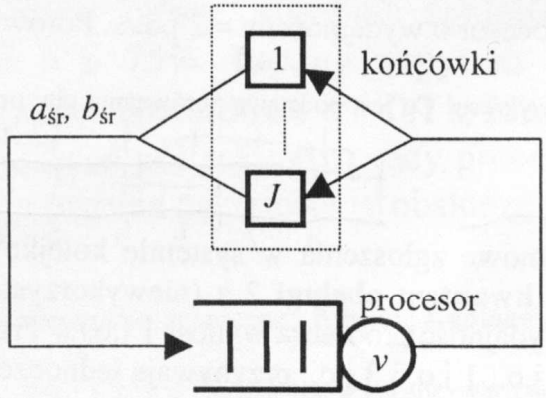
Zadanie 5

O 1-procesorowym systemie masowej obsługi z pamięcią buforową o pojemności $Q = 2$ wiadomo, że w stanie ustalonym mamy $p_0 \geq p_1 \geq p_2$ oraz że $r = 0.75$. Znajdź przedział możliwych wartości p_1.

Należy postąpić tak, jak w poprzednim zadaniu.

Zadanie 6

1-procesorowy system masowej obsługi współpracuje z J =10 inteligentnymi końcówkami w trybie konwersacyjnym zapytanie-odpowiedź. Po otrzymaniu odpowiedzi końcówka generuje nowe zapytanie po czasie namysłu wynoszącym średnio $a_{śr}$ = 4 s. Średnia liczba operacji niezbędnych do wygenerowania odpowiedzi wynosi $b_{śr}$ = 15000, zaś procesor w systemie posiada wydajność v = 5000 operacji/s. Wyznacz zależność pomiędzy współczynnikiem bezczynności procesora a średnim opóźnieniem buforowania.



Znów przyda się równanie ciągłości przepływu.

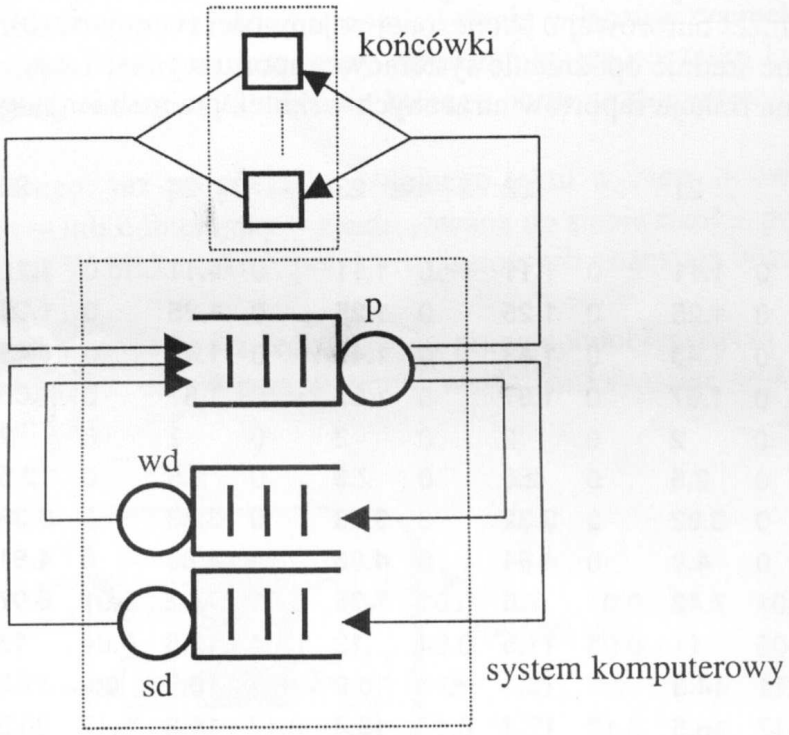
Zadanie 7

System masowej obsługi realizuje średnio w ciągu sekundy 800 transakcji, wymagających wykonania średnio 5000 operacji. Każda przybywająca transakcja otrzymuje do swej dyspozycji procesor o średniej wydajności 4 000 000 operacji na sekundę. Oblicz średnią liczbę transakcji w systemie.

Wykorzystaj prawo Little'a.

Zadanie 8

Każda z $J = 30$ końcówek systemu komputerowego generuje zgłoszenia wymagające sekwencyjnego przetwarzania w procesorze, stacji wolnych dysków i stacji szybkich dysków – rysunek. Średnie liczby wizyt zgłoszenia w urządzeniach wynoszą $l_p = 21$, $l_{wd} = 12$, $l_{sd} = 8$, zaś średnie czasy przetwarzania w trakcie wizyty wynoszą $\tau_p = 0.05$ s, $\tau_{wd} = 0.07$ s, $\tau_{sd} = 0.02$ s. Po zakończeniu przetwarzania zgłoszenia końcówka przechodzi do fazy namysłu trwającej średnio $a_{śr} = 15$ s, po czym generuje kolejne zgłoszenie. a) Gdzie jest wąskie gardło systemu, a gdzie występuje największe przewymiarowanie? Jak się to zmieni, gdy przyśpieszymy procesor tak, że $\tau_p = 0.03$ s? b) Jakie przyśpieszenie procesora niezbędne jest dla uzyskania średniego opóźnienia systemowego (czasu przebywania zgłoszenia w systemie komputerowym) $d^*_{śr} = 12$ s, a jakie dla uzyskania $d^*_{śr} = 9$ s?



Załóż jakąś prędkość przepływu [zgł/s] na styku systemu komputerowego i zbioru końcówek. Wyraź przy jej pomocy współczynnik obciążenia dowolnego urządzenia w systemie. Czy uda się wykorzystać prawo Little'a?

Zadanie 9

Do 1-procesorowego systemu masowej obsługi przybywa strumień niecierpliwych zgłoszeń ze średnim interwałem $a_{śr}$ i średnim wymaganiem $b_{śr}$. Wydajność procesora wynosi v. W chwili przybycia zgłoszenie określa swój indywidualny "próg cierpliwości" (średnia wartość wynosi $c_{śr}$). Jeżeli po jego upływie ciągle jeszcze oczekuje na rozpoczęcie obsługi, ucieka z kolejki. W przeciwnym razie opuszcza system po zakończeniu obsługi. a) Udało nam się pomierzyć współczynnik bezczynności p_0 procesora – jak stąd wyznaczyć frakcję L zgłoszeń, które uciekły z kolejki? b) Udało się też pomierzyć średnią liczbę $N_{śr}$ zgłoszeń w systemie – jak stąd wyznaczyć średnie opóźnienie buforowania zgłoszeń, których obsługa została zrealizowana?

Przedyskutuj zastosowanie prawa Little'a. W punkcie b) ostrożnie!

$N_{śr}/[(1 - L)/a_{śr}] - b_{śr}/v$?

$\ldots = (1 - L)X + L\, c_{śr}$?

Zadanie 10

1-procesorowy system przetwarzania raportów z końcówek telemetrycznych posiada skończoną pamięć buforową o pojemności Q raportów. W zależności od Q i współczynnika obciążenia r zbadano charakterystyki: średniego opóźnienia systemowego raportu, znormalizowanego względem średniego czasu przetwarzania przez procesor (wytłuszczone w tabeli) oraz frakcji raportów utraconych wskutek przepełnienia pamięci buforowej.

Wyznacz maksymalną liczbę jednakowych końcówek telemetrycznych oraz niezbędną pojemność pamięci buforowej przy następujących założeniach:

- końcówka generuje średnio 20 raportów na minutę,
- raport zawiera średnio 1800 rekordów,
- dysponujemy procesorem o wydajności przetwarzania 12000 rekordów na sekundę,
- podział procesora pomiędzy obsługiwane końcówki ma miejsce na zasadzie wspólnego obszaru pamięci buforowej o skończonej pojemności (w raportach),
- dopuszczalne średnie opóźnienie systemowe raportu wynosi 1.8 s,
- dopuszczalna frakcja raportów utraconych wskutek przepełnienia wynosi 4%.

$Q=$	20		21		22		23		24		25	
$r=$												
0.1	**1.11**	0	**1.11**	0	**1.11**	0	**1.11**	0	**1.11**	0	**1.11**	0
0.2	**1.25**	0	**1.25**	0	**1.25**	0	**1.25**	0	**1.25**	0	**1.25**	0
0.3	**1.43**	0	**1.43**	0	**1.43**	0	**1.43**	0	**1.43**	0	**1.43**	0
0.4	**1.67**	0	**1.67**	0	**1.67**	0	**1.67**	0	**1.67**	0	**1.67**	0
0.5	**2**	0	**2**	0	**2**	0	**2**	0	**2**	0	**2**	0
0.6	**2.5**	0	**2.5**	0	**2.5**	0	**2.5**	0	**2.5**	0	**2.5**	0
0.7	**3.32**	0	**3.32**	0	**3.32**	0	**3.33**	0	**3.33**	0	**3.33**	0
0.8	**4.77**	0	**4.8**	0	**4.84**	0	**4.86**	0	**4.89**	0	**4.91**	0
0.9	**7.23**	0.01	**7.42**	0.01	**7.6**	0.01	**7.76**	0.01	**7.92**	0.01	**8.07**	0.01
1	**10.5**	0.05	**11**	0.05	**11.5**	0.04	**12**	0.04	**12.5**	0.04	**13**	0.04
1.1	**13.5**	0.11	**14.3**	0.1	**15.1**	0.1	**15.9**	0.1	**16.7**	0.1	**17.5**	0.1
1.2	**15.5**	0.17	**16.5**	0.17	**17.4**	0.17	**18.4**	0.17	**19.3**	0.17	**20.3**	0.17
1.3	**16.8**	0.23	**17.8**	0.23	**18.7**	0.23	**19.7**	0.23	**20.7**	0.23	**21.7**	0.23
1.4	**17.5**	0.29	**18.5**	0.29	**19.5**	0.29	**20.5**	0.29	**21.5**	0.29	**22.5**	0.29
1.5	**18**	0.33	**19**	0.33	**20**	0.33	**21**	0.33	**22**	0.33	**23**	0.33
1.6	**18.3**	0.38	**19.3**	0.38	**20.3**	0.38	**21.3**	0.38	**22.3**	0.38	**23.3**	0.38
1.7	**18.6**	0.41	**19.6**	0.41	**20.6**	0.41	**21.6**	0.41	**22.6**	0.41	**23.6**	0.41
1.8	**18.8**	0.44	**19.8**	0.44	**20.8**	0.44	**21.8**	0.44	**22.8**	0.44	**23.8**	0.44
1.9	**18.9**	0.47	**19.9**	0.47	**20.9**	0.47	**21.9**	0.47	**22.9**	0.47	**23.9**	0.47
2	**19**	0.5	**20**	0.5	**21**	0.5	**22**	0.5	**23**	0.5	**24**	0.5

Miniprojekty:

Zadanie 11

Strumień zgłoszeń zapisany w pliku `Cw 1 zad 11.txt` (interwały w sekundach, wymagania w j.o.) należy rozdzielić na dwa systemy kolejkowe z nieograniczonymi pamięciami buforowymi i procesorami o wydajnościach v i cv tak, by średnie opóźnienia systemowe ($w_{śr}$ + $\tau_{śr}$) w tych systemach miały się do siebie jak m : 1. W jaki sposób to osiągnąć poprzez losowe rozrzedzanie strumienia? Przyjmij dane: v = 1.25 j.o./s, c = 1.5 i m = 2. Czy to się powiedzie dla każdych v, c i m?

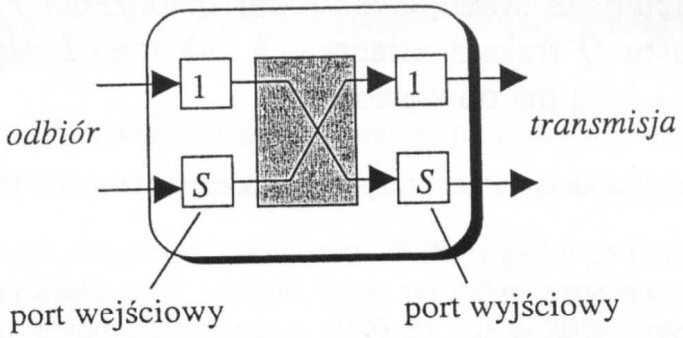
Zadanie 12

Węzeł szybkiej komutacji ramek posiada S portów wejściowych i tyle samo portów wyjściowych, pracujących z taktem równym czasowi transmisji ramki w łączu – rysunek.

Działanie portu wejściowego: O ile nie jest zajęty przez ramkę, na początku kolejnego taktu rozpoczyna odbiór następnej ramki i sprawdza, do którego portu wyjściowego jest on zaadresowany.

Działanie portu wyjściowego: tuż po początku kolejnego taktu wybiera dowolny port wejściowy zajmowany przez – lub odbierający – zaadresowaną do siebie ramkę, po czym rozpoczyna transmisję tej ramki. W braku takich portów wejściowych pozostaje bezczynny do następnego taktu.

Zakładając, że każda ramka może być z jednakowym prawdopodobieństwem zaadresowana do każdego portu wyjściowego, określ przepływność węzła (maksymalne wykorzystanie łączy wyjściowych) w zależności od S.



Zadanie 13

Opracuj narzędzie do wykreślania przebiegu $U(t)$ jak w zadaniu 1 dla strumienia zgłoszeń zadanego w pliku oraz dowolnej liczby procesorów o dowolnych wydajnościach.

Zadanie 14

Dla dyscypliny obsługi *Round Robin* jak w zadaniu 2 oraz strumienia zgłoszeń zadanego w pliku wyznacz eksperymentalnie charakterystykę czasu przebywania w systemie w funkcji wymagania zgłoszenia. Zbadaj wpływ wielkości kwantu obsługi.

Badania Operacyjne / SMO_1 – Ćwiczenia

Zadanie 1

Do 1-procesorowego systemu masowej obsługi z procesorem o wydajności $\nu = 1$ j.o./s przybywa bardzo "gęsty" strumień zgłoszeń, przy czym współczynnik obciążenia wynosi $r = 75\%$. Łączny popyt na obsługę zgłoszony w ciągu sekundy ma odchylenie standardowe $\sigma = 0.1$ s. Jakie są szanse, że w tym okresie uda się wygospodarować pół sekundy pracy procesora dla przetwarzania zadań systemowych bez tworzenia zaległości w obsłudze strumienia zgłoszeń?

Wykorzystaj centralne twierdzenie graniczne i **funkcję Laplace'a**: $\Phi(x) = \frac{1}{\sqrt{2\pi}} \int_0^x e^{-y^2/2} dy$.

Jest tablicowana przeważnie dla $x = 0..5$, a jej wybrane wartości wynoszą: $\Phi(0.5) = 0.191$, $\Phi(1) = 0.341$, $\Phi(1.5) = 0.433$, $\Phi(2) = 0.477$, $\Phi(2.5) = 0.4938$, $\Phi(3) = 0.4987$, $\Phi(\geq 4) \approx 0.5$. Co robić, gdy $x < 0$?

Zadanie 2

1-procesorowy system masowej obsługi posiada pamięć buforową o pojemności $Q < \infty$ zgłoszeń i pracuje ze współczynnikiem obciążenia $r > 1$. W takim systemie pomimo wzrostu Q frakcja utraconych zgłoszeń L nigdy nie spadnie poniżej pewnego progu – jaką ma on wartość?

Należy przypomnieć sobie o równaniu ciągłości przepływu. Jak często procesor jest bezczynny, gdy $Q \to \infty$?

$$\frac{1-p_0}{\nu} = (1-L) - L = \frac{1-p_0}{\nu} - 1$$

$$L = 1 - \frac{1-p_0}{\nu} \to 0$$

$$1 - \frac{1}{\nu}$$