P 93/82

ZESZYTY
NAUKOWE
POLITECHNIKI
ŚLĄSKIEJ

INFORMATYKA

Z. 4
GLIWICE
1982

P.4193/82

POLITECHNIKA ŚLĄSKA

ZESZYTY NAUKOWE

Nr 729

ADAM WOLISZ

WYBRANE NIEKONWENCJONALNE SYSTEMY MASOWEJ OBSŁUGI

GLIWICE 1982

OPINIODAWCY

Prof. dr hab. Zdzisław Pawlak
Doc. dr hab. inż. Andrzej Gościński

KOLEGIUM REDAKCYJNE

Wiesław Gabzdyl (redaktor naczelny), *Stanisław Kozielski* (redaktor działu),
Elżbieta Stinzing (sekretarz redakcji)

OPRACOWANIE REDAKCYJNE

Kazimiera Rymarz

Wydano za zgodą
Rektora Politechniki Śląskiej

PL ISSN 0208-7286

Dział Wydawnictw Politechniki Śląskiej
ul. Kujawska 3, 44-100 Gliwice

Nakł. 140+85 Ark. wyd. 18,10 Ark. druk. 19,25 Papier drukowy kl. V 70x100, 70 g
Oddano do druku 21.06.1982 Podpis. do druku 4.11.1982 Druk ukończ. w grudniu 1982
Zamówienie 616/82 B-23 Cena zł 181,–

Skład, fotokopie, druk i oprawę
wykonano w Zakładzie Graficznym Politechniki Śląskiej w Gliwicach

Q 201/82

SPIS TREŚCI

Str.

WSTĘP 5

I. WPROWADZENIE DO TEORII NIEKONWENCJONALNYCH SYSTEMÓW MASOWEJ OBSŁUGI 11

I.1. Systemy jednostanowiskowe 12
I.1.1. Własności systemów konwencjonalnych 21
I.1.2. Typowe odstępstwa od założeń definiujących konwencjonalne systemy jednostanowiskowe 25
I.1.3. Niekonwencjonalne systemy jednostanowiskowe 32

I.2. Systemy wielostanowiskowe 49
I.2.1. Sieci konwencjonalnych stanowisk obsługi 55
I.2.2. Niekonwencjonalne sieci stanowisk obsługi 65

II. SYSTEMY MASOWEJ OBSŁUGI Z CZASEM OBSŁUGI ZALEŻNYM OD CZASU OCZEKIWANIA I NATURALNYM REGULAMINEM OBSŁUGI 79

II.1. Warunki stabilności 81
II.2. Analiza modelu podstawowego 87
II.3. Przypadki szczególne modelu podstawowego 96
II.4. Inne modele 110
II.5. Niektóre własności systemów WAD z naturalnym regulaminem obsługi 114

III. SYSTEMY MASOWEJ OBSŁUGI Z CZASEM OBSŁUGI ZALEŻNYM OD CZASU OCZEKIWANIA I DOWOLNYM REGULAMINEM OBSŁUGI 137

III.1. Warunki stabilności 138
III.2. Metoda szacowania wydajności granicznej w oparciu o modelowanie cyfrowe 152
III.3. Analiza modelu pomocniczego 169
III.4. Przypadki szczególne i własności modelu pomocniczego . 180
III.5. Dobór parametrów 188
III.6. Wskazania dla doboru regulaminu obsługi 205

IV. MAKSYMALIZACJA WYDAJNOŚCI WIELOSTANOWISKOWYCH, WIELOETAPOWYCH SYSTEMÓW MASOWEJ OBSŁUGI 224

IV.1. Metody wyznaczania wydajności maksymalnej 227
IV.2. Wpływ zrównoważenia systemu na jego wydajność 23

Str.

IV.3. Wpływ rozkładu czasu obsługi oraz pojemności buforów pośrednich na wydajność WSO 241
IV.4. Wpływ wprowadzenia stanowisk wielokanałowych na wydajność WSO 248
IV.5. Wskazania dla maksymalizacji wydajności WSO 255
IV.6. Dodatek - analiza wybranych systemów dwuetapowych 258

ZAKOŃCZENIE 273

DODATEK 275

LITERATURA 283

WSTĘP

Teoria masowej obsługi wyrosła z problemów modelowania obciążenia central telefonicznych, co zaczęło nabierać znaczenia przy występującym z początkiem bieżącego stulecia wzroście liczby abonentów.

Zapoczątkowane pracami A.K. Erlanga [ERLA 09] badania doprowadziły do stworzenia podstawowych pojęć tej teorii, jakimi są: źródła zgłoszeń, kolejka, stanowisko obsługi, regulamin obsługi oraz zdefiniowania podstawowego analizowanego układu - jednostanowiskowego systemu obsługi złożonego z tych elementów.

Sens tych podstawowych pojęć wyszedł wkrótce daleko poza pierwotną dziedzinę opisu działania central telefonicznych i są one chętnie stosowane przy modelowaniu procesów zachodzących we wszystkich dziedzinach techniki, organizacji i gospodarki, w których procesy charakteryzują się dużą liczbą jednostek - "klientów" - ubiegających się o przydzielenie im - na określony czas - ograniczonej liczby zasobów "stanowisk obsługi". Można tu przykładowo wymienić tak różne dziedziny, jak: systemy komputerowe, sieci teletransmisji danych, konserwacja maszyn, gospodarka zapasami, praca placówek służby zdrowia, produkcja stali czy też transport w kopalniach.

Zapoczątkowane w latach sześćdziesiątych stosowanie teorii masowej obsługi dla modelowania i oceny efektywności złożonych systemów komputerowych, sieci teletransmisji danych a następnie sieci komputerowych, a także dyskretnych procesów produkcyjnych, prowadząc do formułowania modeli o bardziej złożonej strukturze aniżeli poprzednio rozpatrywane, stanowiło szczególnie silny bodziec dla notowanego w ostatnich latach burzliwego rozwoju tej dziedziny.

Przystępując do analizy systemów masowej obsługi zakładano pierwotnie z reguły niezależność losową procesów stochastycznych opisujących strumienie zgłoszeń oraz przebieg obsługi poszczególnych zgłoszeń, zakładano też niezależność "zapotrzebowania zgłoszeń na obsługę" od regulaminu obsługi czy też czasu oczekiwania zgłoszeń w kolejce.

Tak rozumiane, konwencjonalne systemy masowej obsługi doczekały się już wielu opracowań monograficznych, systematyzujących olbrzymi dorobek osiągnięty w zakresie metod analizy, pewnych ogólnych praw oraz badania własności licznych wariantów specjalnych.

W szczególności, w odniesieniu do konwencjonalnych systemów masowej obsługi określone zostały precyzyjnie warunki stabilności (tj. istnienia stacjonarnego stanu granicznego rozkładów czasów oczekiwania zgłoszeń,

długości kolejki itp.), co pozwala na odwoływanie się do metod przybliżonych, czy też nawet modelowania cyfrowego w przypadku, gdy dla rozpatrywanego systemu brak rozwiązań analitycznych (których obszar istnienia jest już również dość dobrze określony) bez obawy popełnienia drastycznych błędów o charakterze jakościowym. Udowodnione relacje pomiędzy poszczególnymi charakterystykami systemów (np. pomiędzy czasami oczekiwania zgłoszeń i liczbą zgłoszeń w kolejce, zasady zachowania pracy w systemie) pozwalają ograniczać się do bezpośredniego wyznaczania tylko niektórych spośród nich, co również istotnie upraszcza analizę. Z tego względu aktualne badania w zakresie systemów konwencjonalnych koncentrują się głównie na formułowaniu algorytmów obliczeniowych o możliwie małej złożoności obliczeniowej w odniesieniu do przypadków objętych ścisłymi metodami analizy oraz na problematyce uzyskiwania dalszych, możliwie coraz bardziej dokładnych metod przybliżonych w pozostałych przypadkach (nie w pełni zbadane są również zagadnienia stanów nieustalonych).

Warunkiem efektywnego wykorzystania teorii masowej obsługi dla szerszej liczby zastosowań staje się jednak coraz częściej rezygnacja z tych założeń dodatkowych, zmieniających własności rozpatrywanych systemów zarówno pod względem ilościowym, jak i jakościowym, ale równocześnie istotnie upraszczających ich analizę.

Pojawia się tu podobny problem, jaki wystąpił np. w elektrotechnice czy teorii regulacji, gdy wyniki i metody teorii obwodów liniowych starano się rozszerzyć na układy nieliniowe. Był to problem trudny, gdyż przy rezygnacji z założeń o liniowości nieliczne stają się twierdzenia i metody, które mogłyby stosować się do szerokiej klasy systemów. Stąd też rozwój teorii układów nieliniowych charakteryzował się raczej wyodrębnieniem pewnych klas systemów i określeniem własności i metod odnoszących się do tych klas.

W zakresie niekonwencjonalnych systemów masowej obsługi, gdzie zakładając, że regulamin obsługi i rozkład czasów obsługi mogą być dowolne oraz wzajemnie od siebie uzależnione, a oprócz tego zakładając jeszcze, że charakterystyki strumieni zgłoszeń mogą być dowolne, trudno jest uzyskać jakieś ogólne zależności, które muszą występować w układzie, nie mówiąc już o zależnościach prowadzących do efektywnych metod obliczeniowych.

Z tego względu, podobnie jak to ma miejsce w przypadku teorii układów nieliniowych, teorię niekonwencjonalnych systemów masowej obsługi należy raczej rozwijać na drodze wyodrębnienia pewnych klas niekonwencjonalnych systemów obsługi i opracować dla nich metody analizy, a także badać ich własności.

W szczególności należy znaleźć odpowiedź na następujące, zasadnicze pytania:

- Jak określić warunki stabilności systemu, które w odniesieniu do systemów niekonwencjonalnych nie są jednolite, a wręcz przeciwnie, mogą zależeć od wielu czynników, np. regulaminu obsługi.

Jest to problem bardzo istotny, bowiem dla zastosowania większości metod analitycznych, a także modelowania cyfrowego wymagane jest spełnienie warunków stabilności (lub przynajmniej informacja o tym czy dla konkretnego zestawu parametrów system jest stabilny);

- Jak wyznaczać charakterystyki systemu. W odniesieniu do systemów konwencjonalnych, dla których brak było rozwiązań ścisłych, opracowano liczne metody przybliżone, których dokładność była wielokrotnie testowana. Można łatwo sprawdzić, że metody te, przeniesione do systemów niekonwencjonalnych, z reguły zawodzą.

Istnieje więc pilna potrzeba zarówno rozszerzenia zakresu rozwiązań ścisłych, jak i zbadania możliwości szacowania charakterystyk systemów, gdy rozwiązania te są nieosiągalne;

- Jak dobierać parametry systemu w celu zoptymalizowania jego charakterystyk. Na charakterystyki systemów niekonwencjonalnych z reguły wpływa znacznie większa liczba parametrów i to w sposób częstokroć intuicyjnie nieoczekiwany. Istotne dla poprawy wskaźników pracy systemu może być też aktywne wpływanie na stopień skorelowania zmiennych losowych charakteryzujących system (np. celowe zwiększanie tego skorelowania);

wytyczające kierunki badań nad niekonwencjonalnymi systemami masowej obsługi - a w odniesieniu do wybranych ich klas - zakres niniejszej pracy.

Rozpatrując literaturę przedmiotu można stwierdzić, że jakkolwiek w ostatnich latach coraz liczniejsze są publikacje poświęcone badaniom poszczególnych, bardzo wąsko zdefiniowanych niekonwencjonalnych systemów masowej obsługi, to jednak uzyskiwane zarówno na drodze analitycznej, jak też symulacyjnej wyniki, nawet w odniesieniu do pojedynczych klas takich systemów nie tworzą - jak dotąd - spójnej teorii; nie zapewniają kompletnych charakterystyk ani też wskazań odnośnie zasad projektowania rozpatrywanych systemów (spotyka się np. sprzeczne wskazania odnośnie do reguł optymalizacji).

W czasie swoich prac związanych z komputeryzacją technologicznych procesów przemysłowych autor spotkał się z licznymi obiektami, których modelowanie prowadziło do różnych, niekonwencjonalnych systemów masowej obsługi (jak np. odcinek produkcyjny stalownia-walcownia zgniatacz czy też dwuklatkowe, nawrotne walcowanie blach).

Problemom analizy takich systemów były poświęcone prace: [WOLI 73], [CZAC 74], [WOLI 75], [WOLI 76], [WOLI 76B], [WOLI 78], w których rozpatrywano wpływ pracy własnej na charakterystyki priorytetowych systemów obsługi [WOLI 79], [WOLI 80B], [WOLI 81B], w których analizowano systemy z czasem obsługi zależnym od czasu oczekiwania oraz [WOLI 80A], [WOLI 81A], [WOLI 81C], [PUJO 80A], w których poruszano zagadnienie analizy i optymalizacji wieloetapowych systemów obsługi.

Wyniki przedstawione w tych pracach zachęciły autora do podjęcia dalszych, uwieńczonych powodzeniem, wysiłków w zakresie uzupełnienia metod analizy oraz badania własności dwóch klas niekonwencjonalnych systemów

masowej obsługi: systemów z czasem obsługi zależnym od czasu oczekiwania na jej rozpoczęcie oraz wieloetapowych systemów obsługi z ograniczonymi buforami pośrednimi.

Celem niniejszej pracy było zebranie uzyskanych przez autora oryginalnych rezultatów w zakresie metod analizy oraz badania własności tych dwóch klas niekonwencjonalnych systemów masowej obsługi (w tekście pracy zamieszczono każdorazowo odsyłacze do ewentualnych publikacji lub opracowań wewnętrznych zawierających omawiane wyniki cząstkowe) i przedstawienie ich w ujednolicony sposób na tle całokształtu literatury przedmiotu.

W rozdziale I przedstawiono - obok pojęć i definicji niezbędnych dla dalszych wywodów - próbę sprecyzowania pojęcia i klasyfikacji niekonwencjonalnych systemów masowej obsługi. Przegląd ten, obejmujący zarówno systemy jednostanowiskowe, jak i wielostanowiskowe (sieci stanowisk obsługi), zawiera też praktyczne uzasadnienie celowości rozpatrywania danej klasy systemów, podstawowe ich własności zilustrowane wynikami obliczeń numerycznych oraz omówienie opublikowanej literatury przedmiotu.

Rozdziały II i III poświęcone są rozwinięciu metod analizy oraz projektowania systemów masowej obsługi z czasem obsługi zależnym od czasu oczekiwania na jej rozpoczęcie.
Jest to jedna z najtrudniejszych do analizy klas systemów obsługi, ponieważ nie tylko charakterystyki systemu ale również jego warunki stabilności zależą bezpośrednio od regulaminu obsługi.
Ze względu na różny zakres możliwych do uzyskania na drodze analitycznej wyników, oddzielnie potraktowany został przypadek naturalnego (FIFO) regulaminu obsługi (rozdział II) oraz przypadek dowolnego regulaminu obsługi (rozdział III).

Przy założeniu naturalnego regulaminu obsługi przedyskutowano warunki stabilności systemu a następnie zdefiniowano pewną podklasę tych systemów - tzw. system podstawowy, dla którego uzyskano analityczne rozwiązania.

Rozwiązania te są o tyle wartościowe, że:
- system podstawowy mieści w sobie, jako przypadki szczególne, prawie wszystkie dotąd rozpatrywane przypadki systemów z czasem obsługi zależnym od czasu oczekiwania;
- umożliwiają badanie charakterystyk wielu dalszych, dotąd nie rozpatrywanych przypadków (np. wyjście poza obowiązujące w pracach innych autorów założenie o tym, że zgłoszenia napływają strumieniem Poissona),
- przez odpowiedni dobór systemu podstawowego można, z dowolną dokładnością, przybliżać parametry bardzo szerokiej gamy (prawie wszystkich) systemów z czasem obsługi zależnym od czasu oczekiwania i naturalnym regulaminem obsługi.

Tym samym problem analizy takich systemów można uznać za zamknięty. Podano wyspecyfikowanie uzyskanych rozwiązań dla szeregu przypadków szczególnych (w tym niektórych dotąd nie rozpatrywanych) i zilustrowano ich własności w oparciu o obliczenia numeryczne.

Obok innych, interesujących obserwacji, prowadzą one do wniosku, iż fakt istnienia zależności czasu obsługi od czasu oczekiwania zasadniczo zmienia własności systemu i próby ich szacowania w oparciu o modele pomijające tę zależność prowadzą do istotnych błędów.
Co więcej, również nieścisłość opisu systemu (np. zastąpienie rzeczywistej postaci strumienia zgłoszeń strumieniem Poissona) prowadzi do większych, aniżeli w systemach konwencjonalnych błędów w ocenie charakterystyk systemów.

Zagadnieniom badania systemów z czasem obsługi zależnym od czasu oczekiwania przy różnych od naturalnego regulaminach obsługi poświęcona była - jak dotąd - tylko jedna publikacja [BUZA 74], gdzie rozpatrzono 3 specyficzne, proste przypadki. W rozdziale III niniejszej pracy przeanalizowano warunki stabilności takich systemów. Dla pewnej ich klasy sformułowano i udowodniono warunek stabilności, dla innej, znacznie szerszej grupy, podano metody szacowania (od dołu i od góry) warunku stabilności w oparciu o modelowanie cyfrowe. Metody szacowania warunku stabilności (które mogą być stosowane również w odniesieniu do innych niekonwencjonalnych systemów masowej obsługi), zostały przetestowane na przykładach. Dla pewnej klasy rozpatrywanych systemów uzyskano analityczne rozwiązania i przebadano ich własności. Zaproponowano też i przetestowano na przykładach metodę doboru parametrów systemów przy łącznym wykorzystaniu symulacji oraz modelu analitycznego, w którym przyjęto naturalny regulamin obsługi.
Zaproponowano (w oparciu o rozważania przeprowadzone dla pewnych szczególnych przypadków) heurystyczne reguły doboru regulaminu obsługi z punktu widzenia maksymalizacji wydajności systemu.

Rozdział IV pracy poświęcono opracowaniu reguł projektowania wieloetapowych systemów obsługi z ograniczonymi buforami pośrednimi, z punktu widzenia maksymalizacji ich wydajności.
W formułowaniu tych reguł oparto się na wynikach opublikowanych w literaturze przedmiotu, a także na obliczeniach numerycznych, wykonanych w oparciu o przeanalizowane przez autora trzy modele zestawione w dodatku do tego rozdziału, i rozszerzające zakres wcześniejszych badań.

Uwzględniono wpływy sposobu rozdziału pracy pomiędzy poszczególne etapy, postaci rozkładów czasów obsługi, liczby równoległych kanałów na poszczególnych etapach oraz pojemności buforów pośrednich, ilustrując je licznymi danymi numerycznymi.
Takie uporządkowanie reguł optymalizacji wydajności wydaje się wartościowe wobec pojawienia się licznych, wyrywkowych opracowań zawierających czasem nawet wręcz sprzeczne wskazania.

Omówiono także możliwość uzyskiwania - na drodze analitycznej - charakterystyk takich systemów.

W dołączonym do pracy dodatku wskazano, odwołując się do pewnego przemysłowego systemu produkcyjnego jako przykładu, iż wykorzystanie cech rozpatrzonych uprzednio klas niekonwencjonalnych systemów masowej obsługi

jako modeli cząstkowych pozwala na wyjaśnienie przyczyn występowania pewnych - pozornie zaskakujących i niepożądanych - stanów obiektu, a także na zaproponowanie efektywnych sposobów usprawnienia jego działania.

Autor chciałby w tym miejscu podziękować Prof. dr inż. Stefanowi Węgrzynowi za zachętę do podjęcia niniejszej problematyki i cenne uwagi przedstawione w trakcie pracy nad nią oraz Kolegom z Zakładu Systemów Automatyki Kompleksowej PAN a specjalnie dr inż. Tadeuszowi Czachórskiemu oraz mgr inż. Józefowi Izydorczykowi za liczne wartościowe dyskusje i komentarze.

I. WPROWADZENIE DO TEORII NIEKONWENCJONALNYCH SYSTEMÓW MASOWEJ OBSŁUGI

Gdy przy modelowaniu procesów rozpatrywanych jako systemy masowej obsługi okazuje się, iż nie są spełnione założenia tradycyjnie przy analizie takich systemów przyjmowane, jak na przykład niezależność zmiennych losowych definiujących system masowej obsługi, wówczas mówi się (np. [CALO 78]), iż rozpatrywany przypadek wykracza poza klasę tzw. konwencjonalnych systemów masowej obsługi.

Mimo wyraźnie uwidoczniającego się w ostatnich latach wzrostu zainteresowania takimi nietypowymi systemami, których znaczenie w polepszaniu zgodności modeli z opisywanymi przez nie realiami jest często podkreślane [ROSE 75], [KOBA 77], nie doczekały się one jak dotąd (z jednym skromnym wyjątkiem [BHAT 69]) opracowań wprowadzających systematykę, czy też zawierających przegląd dotychczasowych kierunków i rezultatów badań.

Niniejszy rozdział, zawierający przegląd dorobku w tym zakresie, oparty na pewnej, zaproponowanej przez autora klasyfikacji rozpatrywanych systemów, podzielony jest na dwie części poświęcone odpowiednio systemom jednostanowiskowym i wielostanowiskowym.

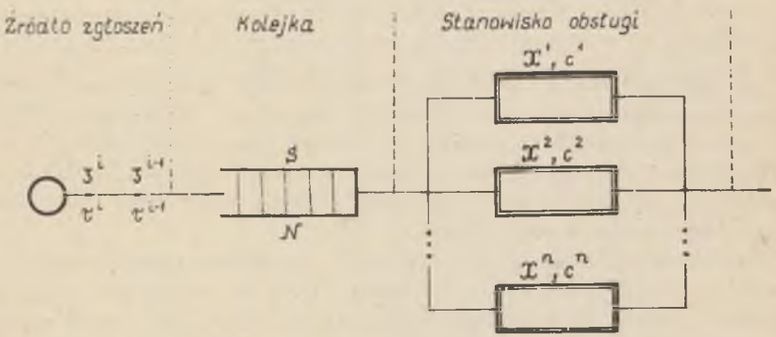
W każdej z tych części, po wprowadzeniu pojęć i oznaczeń niezbędnych dla jednoznacznego opisu rozpatrywanych systemów, a także charakterystyk pozwalających na dokonanie oceny ich działania, podjęte próbę sprecyzowania używanego dotąd dość dowolnie terminu "konwencjonalny system masowej obsługi", podając równocześnie podstawowe własności systemów zaliczonych do tej klasy.

Na tym tle omówione zostaną następnie zasadnicze odstępstwa od takiego, uproszczonego modelu, pozwalające zbliżyć się do cech badanych procesów rzeczywistych, wskazując na podstawowe prace traktujące o poszczególnych przypadkach oraz przytaczając ich najważniejsze cechy.

Cytowane pozycje literaturowe, aczkolwiek nie pretendujące do roli kompletnej bibliografii przedmiotu, umożliwiają orientację w zakresie kierunków badań, osiągniętych wyników oraz perspektyw dalszego rozwoju omawianej dziedziny.

I.1. Systemy jednostanowiskowe

Schemat jednostanowiskowego systemu obsługi przedstawia rysunek I.1.



Rys. I.1. Jednostanowiskowy system masowej obsługi

Zgłoszenia $\zeta^0, \zeta^1, \ldots, \zeta^i, \ldots$ napływają do systemu w kolejnych chwilach czasu $\tau^0, \tau^1, \ldots, \tau^i, \ldots$ i są obsługiwane w n-kanałowym stanowisku obsługi $\mathfrak{X}$ (poszczególne kanały obsługi oznaczać będziemy $\mathfrak{X}^1, \mathfrak{X}^2, \ldots \mathfrak{X}^n$). Nadchodzące zgłoszenia, które zastają wszystkie kanały stanowiska obsługi zajęte (bądź też np. niesprawne) mogą tworzyć kolejkę w przewidzianym dla tego celu buforze S o pojemności N zgłoszeń (wartość N może być skończona lub nieskończona). Przy skończonej pojemności bufora konieczne jest, przy definiowaniu systemu, określenie jego zachowania w przypadku przepełnienia bufora (np. strata zgłoszenia).

Odstępy między chwilami napłynięcia zgłoszeń określają zmienne losowe $a^i = \tau^i - \tau^{i-1}$; $i = 1,2$. Zmienna losowa β^i oznacza zapotrzebowanie zgłoszenia ζ^i na obsługę, wyrażone w jednostkach pracy właściwych dla danego zastosowania [KOBA 79], np. w liczbie instrukcji maszynowych, jeśli obsługa zgłoszenia polega na wykonaniu programu lub w liczbie byte'ów (słów, rekordów), które należy przesłać, jeśli obsługa zgłoszenia polega na przesłaniu pakietu danych w sieci teletransmisji.

Z kolei każdy kanał obsługi charakteryzuje moc c^i, określona w odpowiednich jednostkach pracy na jednostkę czasu, może to więc być np. liczba instrukcji maszynowych wykonywanych w ciągu sekundy bądź liczba byte'ów przesyłanych w ciągu sekundy przez łącze.

Jeżeli wszystkie kanały stanowiska obsługi mają jednakową moc, a więc $c^1 = c^2 \ldots = c^n = c$, wówczas mówimy o jednorodnym stanowisku obsługi.

W dalszym ciągu niniejszej pracy, jeśli nie zostanie wyraźnie zaznaczone inaczej, będziemy mieli na uwadze zawsze jednorodne stanowisko obsługi.

Przez pojęcie mocy stanowiska obsługi c_1 rozumieć będziemy łączną moc zystkich kanałów $c_1 = \sum_{i=1}^{n} c^i$. W pewnych przypadkach może ona zależeć od czby zgłoszeń λ znajdujących się aktualnie w systemie (tj. łącznie w lejce i na stanowisku obsługi) i jest wówczas określona funkcją $c_1(\lambda)$. Tak więc czas obsługi zgłoszenia przez stanowisko obsługi można określić przez podanie zmiennej losowej b^i

$$b^i = \frac{\beta^i}{c}. \qquad (I.1.1)$$

Przy omawianiu zmiennych losowych opisujących proces obsługi zgłoszeń sować będziemy następującą konwencję oznaczeń:
Niech r będzie zmienną losową. Dystrybuantę tej zmiennej losowej oznaić będziemy R(x), funkcję gęstości rozkładu r(x), natomiast transforma- Laplace'a - Stieltjesa dystrybuanty $\bar{r}(s)$, $\bar{r}(s) = \int_0^\infty e^{-sx} dR(x)$.

dpowiednio wartość oczekiwaną rozpatrywanej zmiennej losowej oznaczymy ez E(r), wariancję przez V(r) i współczynnik zmienności rozkładu przez gdzie:

$$c_r = \frac{V(r)}{\left[E(r)\right]^2}. \qquad (I.1.2)$$

oli strumień zgłoszeń napływających do systemu jest stacjonarny, wów- wartość oczekiwaną liczby zgłoszeń generowanych w jednostce czasu λ,

$$\lambda = \left[E(a)\right]^{-1}, \qquad (I.1.3)$$

wać będziemy intensywnością strumienia zgłoszeń.

Rozkłady zmiennych losowych

teorii masowej obsługi rozpatruje się różne rozkłady zmiennych loso- (często przy rozpatrywaniu konkretnych systemów konieczne jest odej- od założenia o dowolnej postaci rozkładów zmiennych losowych a^i czy ako zbyt ogólnego i uniemożliwiającego uzyskanie efektywnych rozwią- .

szczególną rolę odgrywa tu rozkład wykładniczy ze względu na cenną ność "braku pamięci" [GNED 71], co znacznie ułatwia analizę systemu. ówimy, że zmienna losowa r ma rozkład wykładniczy z parametrem ν, i

$$R(x) = 1 - e^{-\nu x} \qquad (I.1.4)$$

i "brak pamięci" wyraża się równością:

$$\Pr\{r > t + h/r > t\} = \Pr\{r > h\}, \quad t>0, \quad h>0. \tag{I.1.5}$$

Rozkład wykładniczy charakteryzujący się równym jedności współczynnikiem zmienności oraz wyraźną asymetrią, jako że

$$\Pr\{r < E(r)\} \cong 0{,}63212, \tag{I.1.6}$$

często nie odpowiada charakterystykom zmiennych losowych obserwowanych w praktyce.

Co więcej, postać rozkładu odstępów między kolejnymi zgłoszeniami czy też rozkładu czasu obsługi ma zwykle istotny wpływ na wartość takich charakterystyk systemu, jak czas oczekiwania czy długość kolejki.

I tak np. średni czas oczekiwania w systemie M/M/1 jest dwukrotnie większy aniżeli w systemie M/D/1 o tym samym średnim czasie obsługi i identycznym strumieniu zgłoszeń. Bardziej szczegółową dyskusję tego wpływu można znaleźć np. w pracach [GROS 75], [CZAC 79] (istnieją nieliczne systemy masowej obsługi, w których postać rozkładu niektórych zmiennych losowych nie wpływa na charakterystyki systemu - patrz np. [KONI 79]).

Z tego względu podjęte zostały próby znalezienia takich rozkładów prawdopodobieństwa, które umożliwiałyby lepsze dopasowanie do własności modelowanych zmiennych losowych (np. drogą uwzględnienia wyższych momentów), zachowując - przynajmniej częściowo - właściwą rozkładowi wykładniczemu łatwość analizy.

A.K. ERLANG [BROC 60] zauważył, iż rozpatrując sumy niezależnych zmiennych losowych o rozkładach wykładniczych można łatwo modelować zmienne losowe o współczynnikach zmienności rozkładu mniejszych od jedności (rozkłady Erlanga [ALLE 78] [GROS 74]), przy czym poszczególne składowe zmienne losowe cechuje nadal "brak pamięci".

W dalszych rozważaniach posługiwać się będziemy często rozkładami Coxa, stanowiącymi uogólnienie tego podejścia. Mówimy, że zmienna losowa φ ma rozkład Coxa k-tego rzędu, jeżeli jej dystrybuanta $\Phi(x)$ ma transformatę Laplace'a - Stieltjesa o postaci:

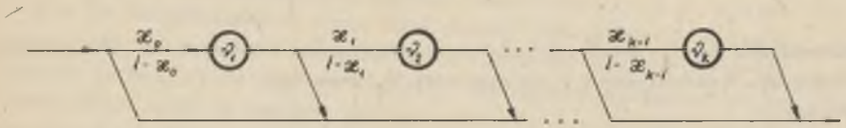
$$\overline{\varphi}(s) = 1 - \alpha_0\left[1 - \sum_{i=1}^{k}(1-\alpha_i)\psi_i \prod_{j=1}^{i}\frac{\vartheta_j}{s+\vartheta_j}\right], \tag{I.1.7}$$

gdzie:

$$\psi_i = \prod_{j=1}^{i-1}\alpha_i$$

$$\vartheta_i > 0, \quad 0 \leq \alpha_i \leq 1 \quad i = 0,1,\ldots k$$

Strukturę tej zmiennej losowej, w rozbiciu na tzw. fazy - tj. elementarne zmienne losowe o rozkładzie wykładniczym - z parametrem ϑ_j, ilustruje rys. I.2 (każdą z faz cechuje brak pamięci!).



Rys. I.2. Struktura zmiennej losowej o rozkładzie Coxa k-tego rzędu, α_i - prawdopodobieństwa dołączenia dalszych faz.

Cox [COXD 55] wykazał, że każda zmienna losowa, posiadająca wymierną transformatę Laplace'a - Stieltjesa dystrybuanty, może być przedstawiona w taki właśnie sposób. Ponieważ każda funkcja może być aproksymowana z żądaną dokładnością przez funkcję wymierną, możliwe jest przy użyciu rozkładu Coxa przybliżenie dowolnego rozkładu prawdopodobieństwa, stąd nazywa się go też rozkładem prawie dowolnym.

W pracy [SCHA 73] wykazano, że dowolną dystrybuantą $F(x)$ nieujemnej zmiennej losowej można przybliżyć dystrybuantą Coxa postaci I.7 również w przypadku założenia, że $\vartheta_1 = \vartheta_2 = \ldots = \vartheta_k = \vartheta$. W [BUXW 77] podano natomiast efektywny algorytm wyznaczenia wartości dystrybuanty $F(x)$ w skończonym zbiorze punktów $x_1, x_2, \ldots, x_m$, tak aby zachować zgodność wartości oczekiwanej i wariancji oraz nie przekroczyć z góry zadanej różnicy wartości dystrybuanty rzeczywistej i poszukiwanej w żadnym z tych punktów.

Ścisłe odwzorowanie, empirycznie określonych na obiekcie rzeczywistym rozkładów czasów obsługi czy też odstępów między chwilami generacji kolejnych zgłoszeń, może prowadzić do zidentyfikowania rozkładów Coxa wysokiego rzędu, a więc zadań o stosunkowo dużej złożoności. Z tego względu oraz z uwagi na znacznie mniejszy wpływ wartości wyższych momentów [GROS 75] w praktyce często ogranicza się wymagania do zachowania zgodności wartości dwóch pierwszych momentów tych zmiennych losowych. Dowolną dystrybuantę $R(x)$ zmiennej losowej r o wartości oczekiwanej $E(r) > 0$ oraz współczynniku zmienności rozkładu $C_r \geq 0{,}5$ można, spełniając powyższy warunek, przybliżyć rozkładem Coxa drugiego rzędu o dystrybuancie $\Phi(x)$ posiadającej transformatę Laplace'a - Stieltjesa

$$\overline{\varphi}(s) = \frac{1 + h_0 s}{1 + h_1 s + h_2 s^2}. \tag{I.1.8}$$

Współczynniki h_0, h_1, h_2 można przy tym wyznaczyć jednoznacznie w funkcji $E(r)$ oraz C_r według następujących reguł:

i) Przy $0{,}5 \leq C_r < 1$

$$h_0 = 0, \quad h_1 = E(r), \quad h_2 = 0{,}5(1 - C_r)\left[E(r)\right]^2. \tag{I.1.9}$$

Dystrybuanta $\Phi(x)$ przybiera wówczas postać:

$$\Phi(x) = 1 + \frac{\nu_1}{\nu_2 - \nu_1} e^{-\nu_2 x} - \frac{\nu_2}{\nu_2 - \nu_1} e^{-\nu_1 x},$$

odpowiadającą sumie dwóch niezależnych zmiennych losowych o rozkładzie wykładniczym (odpowiednio z parametrami ν_1 oraz ν_2). Oczywiście:

$$E(r) = \frac{1}{\nu_1} + \frac{1}{\nu_2},$$

$$C_r = 1 - \frac{2\nu_1\nu_2}{(\nu_1 + \nu_2)^2}.$$

ii) Przy $C_r = 1$

$$h_o = 0, \quad h_1 = E(r), \quad h_2 = 0, \tag{I.1.10}$$

co odpowiada rozkładowi wykładniczemu,

$$\Phi(x) = 1 - e^{-\nu x},$$

gdzie:

$$E(r) = \frac{1}{\nu}.$$

iii) Przy $C_r > 1$,

$$h_o = C_r E(r); \quad h_1 = (C_r + 1)E(r); \quad h_2 = 0{,}5(C_r + 1)\left[E(r)\right]^2, \tag{I.1.11}$$

co odpowiada rozkładowi hiperwykładniczemu o dystrybuancie

$$\Phi(x) = 1 - k\, e^{-2k\sigma x} - (1 - k)e^{-2(1-k)\sigma x}, \quad 0{,}5 < k < 1,$$

gdzie:

$$E(r) = \frac{1}{\sigma}, \quad C_r = \frac{1}{2k(1-k)} - 1.$$

Po przedyskutowaniu charakterystyk rozkładów zmiennych losowych przejdziemy obecnie do omówienia sposobów organizacji obsługi.

Regulaminy obsługi

Zgłoszenia oczekujące na rozpoczęcie obsługi tworzą kolejkę w buforze S o pojemności ograniczonej, ogólnie rzecz biorąc, do N miejsc (N skończone lub nieskończone), przy czym reguły szeregowania zgłoszeń określa regulamin obsługi R. W szczególności regulamin obsługi określa, w każdej chwili czasu, które spośród zgłoszeń obecnych w systemie ma być obsługiwane przez każdy z kanałów obsługi, określa też momenty wykonywania przez te kanały pewnych operacji dodatkowych.

Ogólnie rzecz biorąc każdy kanał stanowiska obsługi może znajdować się w jednym z trzech stanów:

- zajęty - gdy realizuje obsługę zgłoszenia,
- przygotowywany - gdy kanał nie obsługuje żadnego zgłoszenia mimo niepustej kolejki. W tym stanie kanał obsługi może znajdować się w przypadku np. wymuszonego oczekiwania na napłynięcie nowego zgłoszenia, wykonywania pracy własnej, związanej ze zmianą typu obsługiwanego zgłoszenia czy też naprawy kanału,
- bezczynny - gdy kanał nie obsługuje zgłoszenia i kolejka jest pusta.

Będziemy mówili, że stanowisko obsługi jest bezczynne, gdy wszystkie jego kanały są bezczynne. Okres czasu pomiędzy przerwaniem bezczynności stanowiska obsługi a jego powrotem do stanu bezczynności nazywać będziemy okresem czynnym (tak więc przygotowywanie stanowiska obsługi wchodzi w skład okresu czynnego).

Wprowadzimy definicje pewnych, specyficznych klas regulaminów obsługi.

Będziemy mówili o regulaminach nieprzerywalnych, te jest takich, że zgłoszenie wprowadzone na stanowisko obsługi może je opuścić dopiero po całkowitym zakończeniu obsługi, w odróżnieniu od regulaminów przerywalnych, przy których obsługa zgłoszenia może być czasowo przerwana a następnie kontynuowana (bądź wznowiona od początku). Zauważmy, że na czas oczekiwania zgłoszeń w kolejce, w przypadku stosowania regulaminów przerywalnych, składa się łączny czas wszystkich pobytów zgłoszenia w kolejce.

Regulaminem obsługi zachowującym pracę nazywać będziemy regulamin obsługi, w którym:

i) zapotrzebowanie każdego zgłoszenia na obsługę jest określone w chwili jego wygenerowania (chociaż z reguły nie znane w sposób jawny), nie jest więc zależne od przebiegu procesu obsługi, a w szczególności od regulaminu obsługi,

ii) zgłoszenie, które zostało umieszczone w kolejce, może opuścić system obsługi dopiero po całkowitym zakończeniu obsługi,

iii) dopuszczalne są tylko dwa stany każdego z kanałów obsługi: zajęty lub bezczynny. Nie dopuszcza się więc wstrzymania rozpoczęcia obsługi zgłoszeń, jeśli kolejka zgłoszeń jest niepusta.

Systemy obsługi z takim regulaminem obsługi nazywać będziemy zachowawczymi. Dla krótkiego oznaczenia rodzaju rozpatrywanego systemu masowej

obsługi będziemy często używać tzw. notacji Kendalla postaci [KLEI 75] [GROS 74]

$$\mathcal{A}/\mathcal{B}/n/L/\mathcal{R},$$

gdzie $\mathcal{A}$ określa charakterystykę odstępów między wygenerowaniem kolejnych zgłoszeń, $\mathcal{B}$ określa charakterystykę czasów obsługi, n określa liczbę kanałów obsługi, L definiuje pojemność systemu, to jest maksymalną liczbę zgłoszeń, które mogą znajdować się w systemie (L = N≥n) i wreszcie $\mathcal{R}$ określa regulamin obsługi.

Przyjmuje się następujące symbole dla oznaczania $\mathcal{A}$ i $\mathcal{B}$

G(t) - oznacza, że zmienne losowe a^i (lub b^i) są zmiennymi losowymi o dowolnych, zmiennych w czasie rozkładach,

G - oznacza, że zmienne losowe a^i (lub b^i) są zmiennymi losowymi o identycznych, dowolnych rozkładach,

GI - oznacza, że zmienne losowe a^i (lub b^i) są zmiennymi losowymi niezależnymi o identycznych, dowolnych rozkładach.

W szczególności dla konkretnych rozkładów stosuje się oznaczenia:

M - rozkład wykładniczy,

D - rozkład regularny,

E_k - rozkład Erlanga k-tego rzędu,

H_k - rozkład hiperwykładniczy k-tego rzędu,

K_l - rozkład Coxa l-tego rzędu.

Opuszczenie symbolu L w tej notacji oznacza $L = \infty$, natomiast opuszczenie symbolu $\mathcal{R}$ oznacza regulamin naturalny (FIFO), tj. obsługę zgłoszeń w kolejności ich napływania do systemu.

Charakterystyki systemu

Wprowadzimy obecnie kilka pojęć i oznaczeń związanych z oceną działania systemu masowej obsługi.

Łączny czas spędzony przez zgłoszenie ζ^i w kolejce (czas oczekiwania) określa zmienna losowa w^i, natomiast czas pobytu zgłoszenia w systemie, tj. czas od momentu wygenerowania zgłoszenia ze źródła do chwili całkowitego zakończenia jego obsługi określać będzie zmienna losowa g^i.

Niech $z_{\mathcal{R}}^k(t)$, k = 1,2,...,n określa zapas pracy k-tego kanału obsługi w chwili t, to jest czas pracy k-tego kanału obsługi - począwszy od chwili t - niezbędny, przy przyjętym regulaminie obsługi $\mathcal{R}$, dla obsłużenia wszystkich zgłoszeń ζ^i takich, że $\tau^i < t$, przy założeniu braku napływu dalszych zgłoszeń.

Wprowadźmy wektor

$$Z_{\mathcal{R}}(t) = \left\{ z_{\mathcal{R}}^{[1]}(t),\ z_{\mathcal{R}}^{[2]}(t),\ldots,z_{\mathcal{R}}^{[n]}(t) \right\}$$

oraz wyrażenie określające łączny zapas pracy

$$\mathcal{Z}(\mathcal{R},t) = \sum_{k=1}^{n} z_{\mathcal{R}}^{[k]}(t).$$

Łączny zapas pracy w dowolnej chwili t jest przy wszystkich regulaminach obsługi zachowujących pracę identyczny.

Oznaczmy przez

$$\mathcal{Z}_{\mathcal{R}}^i = \mathcal{Z}_{\mathcal{R}}(\tau^i - 0).$$

Można łatwo sprawdzić, że przy n=1 i regulaminie naturalnym zachodzi:

$$w^i = \mathcal{Z}_{FIFO}^i = z_{FIFO}^{[1]}(\tau^i - 0). \qquad (I.1.12)$$

System masowej obsługi nazywamy systemem stabilnym, gdy istnieją takie F(t), H(t), że

$$\bigvee_{i=0,1,\ldots} \bigvee_{t\geq 0} \left\{ F(t) \leq W^i(t) \leq H(t) \right\},$$

gdzie F(t) oraz H(t) są pewnymi dystrybuantami (prawidłowymi, tj. takimi, że $\lim_{t\to\infty} F(t) = \lim_{t\to\infty} H(t) = 1$) i niestabilnym w przeciwnym razie. W większości przypadków systemy stabilne posiadają graniczny rozkład czasu oczekiwania.

$$W(x) = \lim_{i\to\infty} W^i(x). \qquad (I.1.13)$$

Systemy stabilne, nie spełniające (I.1.13), nazywać będziemy systemami substabilnymi. W dalszych rozważaniach, jeśli nie zostanie wyraźnie zaznaczone inaczej, mówiąc o systemach stabilnych będziemy rozpatrywać systemy spełniające (I.1.13).

W systemach stabilnych zwykle wartość średnia czasu oczekiwania jest ograniczona, tj.

$$\bigvee_{W^0(x)} \exists_{M<\infty} \left\{ \limsup_{i\to\infty} E(w^i) < M \right\}$$

Dla szerokiej klasy systemów obsługi można wykazać, że warunki istnienia granicznego rozkładu, a także warunki stabilności sformułowane w odniesieniu do czasu oczekiwania są identyczne z analogicznymi warunka-

mi sformułowanymi w odniesieniu do innych charakterystyk systemów, np. liczby zgłoszeń oczekujących w kolejce czy też zapasu pracy w systemie, np. [MIYA 79].

Warunki stabilności, a także charakterystyki systemów stabilnych przy $t \to \infty$, czyli przy tzw. pracy w stanie ustalonym (stacjonarnym stanie granicznym), należą do najistotniejszych kryteriów oceny działania systemów masowej obsługi.

Systemy masowej obsługi mogą być zwykle stabilne bądź niestabilne, w zależności od intensywności λ strumienia zgłoszeń, tj. wartości oczekiwanej liczby zgłoszeń nadchodzących do systemu w jednostce czasu, lub też - w przypadku niestacjonarnych strumieni zgłoszeń - w zależności od postaci tych strumieni.

Przez pojęcie warunku stabilności będziemy rozumieli warunek narzucony na strumień zgłoszeń, przy spełnieniu którego rozpatrywany system obsługi jest stabilny. Systemy masowej obsługi, które są niestabilne dla dowolnej, różnej od zera intensywności strumienia zgłoszeń, nazywać będziemy systemami strukturalnie niestabilnymi.

Wydajnością V systemu masowej obsługi nazywać będziemy wartość oczekiwaną liczby całkowicie obsłużonych zgłoszeń opuszczających stanowisko obsługi w jednostce czasu, natomiast wydajnością graniczną T nazywać będziemy kres górny zbioru wydajności systemu obsługi, przy których system jest stabilny. W stabilnym systemie obsługi z $N = \infty$ zachodzi $V = \lambda$.

Niech w stabilnym systemie obsługi $\eta_{\mathcal{R}}^{i}(\lambda)$ oznacza stopień wykorzystania i-tego kanału obsługi, tj. ułamek czasu, w którym kanał ten nie jest bezczynny, natomiast $\Xi_{\mathcal{R}}^{i}(\lambda)$ stopień efektywnego wykorzystania kanału, tj. ułamek czasu, w którym kanał jest zajęty efektywną obsługą zgłoszeń, w przypadku gdy intensywność strumienia zgłoszeń napływających wynosi λ i obowiązuje regulamin obsługi $\mathcal{R}$.

W oparciu o te definicje wprowadzimy pojęcia:

- stopnia wykorzystania stanowiska obsługi

$$\eta_{\mathcal{R}}(\lambda) = \frac{1}{n}\sum_{i=1}^{n}\eta_{\mathcal{R}}^{i}(\lambda), \quad \eta_{\mathcal{R}}(\lambda) \leq 1;$$

- stopnia efektywnego wykorzystania stanowiska obsługi

$$\Xi_{\mathcal{R}}(\lambda) = \frac{1}{\eta}\sum_{i=1}^{n}\Xi_{\mathcal{R}}^{i}(\lambda) \leq \eta_{\mathcal{R}}(\lambda);$$

- jednostkowego czasu obsługi zgłoszenia

$$\eta_{\mathcal{R}}^{*}(\lambda) = \frac{n\eta_{\mathcal{R}}(\lambda)}{\lambda},$$

określającego średni czas pracy stanowiska obsługi przypadający na indywidualne zgłoszenie;

- jednostkowego efektywnego czasu obsługi zgłoszenia

$$\Xi_{\mathcal{R}}^{*}(\lambda) = \frac{n\Xi_{\mathcal{R}}(\lambda)}{\lambda} \leq \eta_{\mathcal{R}}^{*}(\lambda).$$

W dalszych rozważaniach, w przypadku gdy jednoznacznie określono jaki regulamin obsługi jest rozpatrywany, indeks określający regulamin obsługi będziemy opuszczać.

Wielkości charakteryzujące pracę systemu masowej obsługi można podzielić na trzy grupy:

- charakterystyki związane z poszczególnymi zgłoszeniami (np. czas oczekiwania zgłoszenia w kolejce, czas pobytu zgłoszenia w systemie),
- charakterystyki związane z czasem (np. liczba zgłoszeń oczekujących w kolejce, liczba zgłoszeń obecnych w systemie, zapas pracy w systemie),
- charakterystyki określające tryb pracy stanowiska obsługi (stopień wykorzystania czy też efektywnego wykorzystania, jednostkowy czas obsługi).

Metody prowadzące do wyznaczania tych trzech rodzajów charakterystyk są często różne. Stąd wiele uwagi poświęcono wykryciu powiązań między wartościami charakterystyk należących do różnych grup, szczególnie grupy pierwszej i drugiej (do tego zagadnienia wrócimy poniżej). Charakterystyki grupy trzeciej są często wyznaczalne w sposób bardzo prosty.

I.1.1. Własności systemów konwencjonalnych

Jednostanowiskowy system masowej obsługi nazywać będziemy systemem konwencjonalnym, jeśli:

i) stanowisko obsługi ma stałą moc,

ii) a^i oraz b^i, $i = 1,2,\ldots$, są ciągami niezależnych zmiennych losowych o jednakowych rozkładach, posiadających odpowiednio dystrybuanty $A(x)$ oraz $B(x)$ o skończonych wartościach oczekiwanych $E(a) = \frac{1}{\lambda}$, $E(b) = \frac{1}{\mu}$,

iii) zgłoszenia obsługiwane są zgodnie z regulaminem zachowującym pracę,

iv) bufor S ma nieograniczoną pojemność ($N = \infty$).

Stosując notację Kendalla mówimy więc o systemach $GI/GI/n/\infty/\mathcal{R}$ z regulaminem obsługi zachowującym pracę.

Dla jednostanowiskowych konwencjonalnych systemów masowej obsługi znany jest warunek stabilności [KIEF 56] o postaci:

$$\rho = \frac{\lambda}{n\mu} < 1, \qquad (I.1.14)$$

z wyjątkiem jednego tylko, specjalnego przypadku, gdy zarówno zmienne losowe a^i, jak też b^i mają rozkłady regularne (przypadek deterministryczny), kiedy to warunek stabilności przybiera postać:

$$\rho \leqslant 1.$$

Tak więc wydajność graniczna T w systemach konwencjonalnych jest określona jako:

$$T = n\mu.$$

Dla każdego jednokanałowego systemu konwencjonalnego zapas pracy $\mathcal{Z}(\mathcal{R},t)$ nie zależy od regulaminu obsługi, a więc w przypadku systemów stabilnych zachodzi też:

$$\lim_{t\to\infty} E\left\{\mathcal{Z}(\mathcal{R},t)\right\} = \alpha = \text{const.} \qquad \text{(I.1.15)}$$

bez względu na stosowany regulamin obsługi [GELE 80A]. Zależność ta nosi nazwę zasady zachowania pracy.

Dla wszystkich stabilnych systemów konwencjonalnych prawdziwe są też następujące równości:

$$\left.\begin{aligned} \Xi(\lambda) &= \eta(\lambda) = \rho = \frac{\lambda}{\mu}, \\ \Xi^*(\lambda) &= \eta^*(\lambda) = E(b), \end{aligned}\right\} \qquad \text{(I.1.16)}$$

dla każdego $0 < \lambda < T$.

Definicja regulaminu obsługi zachowującego pracę zawierała wymóg, by zapotrzebowanie zgłoszenia na obsługę (a więc, przy stałej mocy stanowiska obsługi, również czas obsługi) było określone w chwili jego wygenerowania. Różne mogą być jednak sytuacje, jeśli chodzi o znajomość tego czasu. W pewnych przypadkach jest on znany precyzyjnie, w innych przyjmujemy, iż wyróżnia się pewne klasy zgłoszeń cechujące się różnymi rozkładami czasów obsługi. W zależności od stopnia szczegółowości tych informacji można podać pewne dodatkowe własności systemów konwencjonalnych.

Ograniczmy się czasowo do regulaminów obsługi, w których szeregowanie zgłoszeń odbywa się przy braku jakiejkolwiek informacji o wymaganym czasie obsługi indywidualnego zgłoszenia. Dowodzi się [GITT 78] [VASI 77], iż dla nieprzerywalnych regulaminów obsługi w wielokanałowych systemach konwencjonalnych średni czas oczekiwania jest niezależny od regulaminu obsługi. Przez zmianę regulaminu obsługi (np. zastosowanie losowego wyboru kolejności oczekujących zgłoszeń) można zmienić jedynie wyższe momenty rozkładu czasu oczekiwania. Wykazano przy tym [VASI 77], że w systemach

typu GI/GI/n/∞/$\mathcal{R}$ wariancja czasu oczekiwania przy użyciu regulaminu $\mathcal{R}$ jest nie większa aniżeli wariancja czasu obsługi w tym systemie przy użyciu regulaminu LIFO, natomiast nie mniejsza aniżeli przy użyciu regulaminu FIFO. W pracy [COHE 69] podano, dla systemu jednokanałowego, zestawienie wartości wariancji czasów oczekiwania uzyskiwanych przy sześciu różnych regulaminach obsługi, zachowujących stałość wartości średniej czasu oczekiwania.

Definiując systemy zachowawcze przyjęliśmy, że żadne stanowisko obsługi nie może pozostawać bezczynne w przypadku istnienia niepustej kolejki. W pracy [GITT 78] wykazano, że usunięcie tego warunku z definicji wielokanałowego systemu konwencjonalnego nie może prowadzić do zmniejszenia średniego czasu oczekiwania zgłoszeń, który byłby wówczas większy lub równy wartości osiąganej w systemie konwencjonalnym.

Wcześniej i przy nieco słabszych założeniach dowody stałości średniego czasu oczekiwania uzyskano dla systemów jednokanałowych [KLEI 76], np. również przy stosowaniu regulaminów przerywalnych, pod warunkiem, że czasy obsługi zgłoszeń mają rozkład wykładniczy. Co więcej, dopuszczalne jest rozdzielenie strumienia napływających zgłoszeń na klasy; zachowany jest też rozkład liczby zgłoszeń w systemie. W przypadku regulaminów nieprzerywalnych i p klas zgłoszeń w systemach typu GI/GI/1/∞/$\mathcal{R}$ zachodzi relacja [KLEI 65], [WOLF 70] [SCHR 70]:

$$\sum_{i=1}^{p} \lambda_i E(b_i) E(w_i) = \alpha - 0{,}5 \sum_{i=1}^{p} \lambda_i E(b_i^2), \qquad \text{(I.1.17)}$$

gdzie: λ_i, $E(b_i^2)$, $E(w_i)$ oznaczają odpowiednio: intensywność strumienia zgłoszeń, drugi moment czasu obsługi oraz średni czas oczekiwania zgłoszeń należących do i-tej klasy, i = 1,2,...p; natomiast α jest stałą z zależności (I.1.15).

Tak więc dla uzyskania niektórych interesujących parametrów wystarczy przeanalizować dowolny regulamin obsługi z omawianej klasy, np. stosunkowo najprostszy regulamin naturalny.

Nawet jednak dla tego regulaminu, w przypadku jednokanałowym wykazano, że przy przyjęciu dowolnych rozkładów (system GI/GI/1) nie istnieją zależności określające bezpośrednio wartości oczekiwane (nie mówiąc już o rozkładach) czasu oczekiwania czy też liczby zgłoszeń w systemie.

Bogata literatura przedmiotu obejmuje wiele prostszych postaci rozkładów (patrz np. monografia [COHE 69] poświęcona wyłącznie systemom jednokanałowym).

W przypadku systemów wielokanałowych brak analitycznych rozwiązań nawet klasy M/GI/n.

Ze względu na stosowanie różnych metod analizy, prowadzących bezpośrednio tylko do niektórych charakterystyk systemów, istotne okazały się

prace pozwalające na określenie, dla naturalnego regulaminu obsługi, wzajemnych związków między różnymi charakterystykami; a specjalnie związków między czasem oczekiwania a liczbą zgłoszeń oczekujących w kolejce.

Najstarszy wynik, tzw. wzór Little'a [LITT 61] $\tilde{L} = \lambda\tilde{W}$, określający związek średniego czasu oczekiwania zgłoszeń $\tilde{W}$ oraz średniej długości kolejki $\tilde{L}$, został uogólniony [BRUM 72] na wyższe momenty tych charakterystyk. Zależności między tymi charakterystykami były dalej badane m.in. w pracach [BRUM 71], [MIYA 79], [MORI 80], [BORO 72] oraz [HEYM 80] (wyniki uzyskane w tym zakresie dotyczą systemów wielokanałowych).

Zastosowanie teorii masowej obsługi do modelowania systemów komputerowych spowodowało potrzebę głębszej analizy różnych regulamów obsługi.

W szczególności rozwinięto istotnie (przy założeniu, że zgłoszenia napływają strumieniem Poissona) teorię systemów priorytetowych, które doczekały się nawet odrębnych monografii [JAIS 68], [GNED 73], [BRON 76] [MOWA 77]. Na specjalne wymienienie zasługują wśród nich omówione obszernie w [KLEI 76] systemy z podziałem czasu (patrz też [ADIR 69] [SAKA 71], [WEGR 74]), w tym systemy posiadające złożoną wielopoziomową strukturę kolejek, względnie z dynamicznym przydziałem kwantów [BHAT 79B], [POTI 76], [BLEV 76], systemy z rozbudowanymi regulaminami o stałych priorytetach z wyróżnieniem przerywalnych i nieprzerywalnych faz obsługi [SCHR 69], [WOLI 76], [MISZ 74] oraz systemy z dynamiczną zmianą priorytetu [KLEI 76], [GOLD 77].

Tematem wielu prac, ograniczających się jednak do przypadku jednokanałowego, było zagadnienie syntezy regulaminu obsługi tak, aby uzyskać wymagane własności systemu (np. czasy oczekiwania) drogą doboru odpowiedniego regulaminu obsługi bądź też parametrów takiego regulaminu. Liczne wyniki w tym zakresie zawiera [GELE 80A] i bibliografia [CRAB 77], na uwagę zasługują też [SIMO 73], [MICH 72], [WOLI 76A].

Aktualnie podejmowane są próby opracowania jednolitego, sformalizowanego opisu możliwie szerokiej klasy regulaminów obsługi [RUSC 81], co ułatwia dobór odpowiedniego regulaminu dla konkretnych wymagań.

Możliwości kształtowania charakterystyk systemów masowej obsługi są ograniczone więzami wynikającymi z zasady zachowania pracy, np. typu (I.1.17).

Obszerne omówienie warunków określających możliwość osiągnięcia wybranych wartości wektora $[w_1, w_2, \ldots, w_p]$, tj. średnich czasów oczekiwania zgłoszeń różnych klas podano w [GELE 80A].

Omawiane dotąd wyniki bazowały wyłącznie na znajomości ogólnych, statystycznych charakterystyk wymaganego czasu obsługi poszczególnych zgłoszeń.

Przy założeniu dokładniejszej informacji, mimo iż dalej zachodzi (I.1.15), zmiana regulaminu obsługi wpływa - ogólnie rzecz biorąc - na średni czas oczekiwania. Wykazano (np. [SCHR 68B], [SEVC 74]) optymalność (z punktu widzenia minimalizacji średniego czasu oczekiwania) regulaminu SPT (obsługa w kolejności wzrastających czasów obsługi), przy obsłudze

nieprzerywalnej bądź SRPT (obsługa w kolejności wzrastających pozostałych czasów obsługi) przy obsłudze przerywalnej, w przypadku deterministycznej znajomości czasów obsługi poszczególnych zgłoszeń. W monografii [KLEI 76] przytoczono w ślad za [SCHR 74] tabelę zawierającą wskazania dla optymalnego doboru regulaminu obsługi, przy różnym poziomie informacji o czasie obsługi (brak informacji, znany rozkład czasu obsługi, znany dokładnie czas osługi) wraz z podaniem źródłowych prac zawierających odpowiednie dowody.

W problematyce wyznaczania charakterystyk konwencjonalnych systemów jednostanowiskowych wiele jest jeszcze problemów otwartych.

Stąd też obserwuje się wyraźną koncentrację prac nad przybliżonymi metodami określenia tych charakterystyk (przegląd takich metod zawiera praca [BHAT 79A]).

Biorąc pod uwagę istnienie analitycznie sformułowanych warunków stabilności, rozwijane są też teoretyczne podstawy metod symulacyjnych dla systemów stabilnych, mające na celu opracowanie takiej organizacji eksperymentu, aby uzyskać wymaganą dokładność przy możliwie małej złożoności obliczeniowej [CRAN 77], [IGLE 80].

Należy sądzić, iż obok rozpatrywania nowych, wynikających z konkretnych potrzeb regulaminów obsługi, wymienione wyżej nurty tematyczne będą dominowały w dalszych pracach dotyczących jednostanowiskowych, konwencjonalnych systemów masowej obsługi.

I.1.2. Typowe odstępstwa od założeń definiujących konwencjonalne systemy jednostanowiskowe

Konwencjonalne systemy masowej obsługi stanowią znacznie uproszczony model procesów rzeczywistych. Wprowadzone uproszczenia pozwoliły jednak na osiągnięcie znacznego dorobku w zakresie ich analizy i optymalizacji, dorobku, z którego można bezpośrednio korzystać - przynajmniej w pierwszym przybliżeniu - dla zgrubnego określenia własności analizowanych obiektów. Co więcej, można wskazać przypadki, gdy taki, znacznie uproszczony model, dostarcza informacji o charakterze nie tylko jakościowym, ale również ilościowym. Często cytowanym przykładem są tu prace z zakresu modelowania systemów wielodostępnych [SCHR 68A], gdzie pomiary prowadzone na pracującym zestawie wykazały zaskakująco dobrą zgodność z wynikami analizy stosunkowo "prymitywnego" modelu.

W związku z istotnym rozszerzeniem zakresu zastosowań teorii masowej obsługi (obszerną dyskusję tradycyjnych i nowszych zastosowań można znaleźć np. w [GROS 74], [ROSE 75]), okazało się, że wierne zamodelowanie zjawisk zachodzących w rozpatrywanych procesach wymaga uwzględniania istotnych odstępstw od założeń określających konwencjonalne systemy masowej obsługi [KOBA 77], [ROSE 75].

Tabela I.1

Typowe odstępstwa od założeń określających konwencjonalne jednostanowiskowe systemy masowej obsługi

Źródło zgłoszeń	Obsługa zgłoszeń	Regulamin obsługi
a) Grupowe nadchodzenie zgłoszeń [SAAT 61], [KUCZ 73], [COHE 80]	a) Grupowa obsługa zgłoszeń [SAAT 61], [GROS 74], [COHE 80]	a) Systemy z ograniczeniem liczby zgłoszeń, $N < \infty$ [GNED 71], [TRUS 75], [SCHM 78]
b) Zależność intensywności strumienia zgłoszeń od liczby zgłoszeń w systemie [JAIS 68], [KOPO 73], [WOLI 76], [GORD 74], [CHOO 79], [FALI 80]	b) Zależność mocy stanowiska obsługi od liczby zgłoszeń w systemie [HARR 67], [HARR 70], [GORD 74]	b) Regulaminy z czasem obsługi zależnym od kolejności obsługi [WELC 64], [FINC 59], [MINH 80]
c) Niestacjonarny strumień zgłoszeń [PURD 74], [NEUT 78A], [NEUT 78B]	c) Skorelowanie czasów obsługi [LOYN 62], [BORO 72], [JACO 80], [NISH 77]	c) Regulaminy z uwzględnieniem pracy własnej [SCHR 69], [MISZ 74], [WOLI 76], [EISS 74], [KONH 74], [KLEI 70], [BHAT 79]
e) Skorelowanie czasów między chwilami generacji kolejnych zgłoszeń [LOYN 62], [BORO 72], [JACO 80]	d) Zależność zapotrzebowania na obsługę od czasu potrzebnego na generację zgłoszenia [CONO 74], [CONO 79], [MITC 79], [JACO 80]	d) Regulaminy z czasową niedostępnością stanowiska obsługi [SKIN 67], [GELE 80], [GELE 79], [BACE 79]
	e) Zależność zapotrzebowania na obsługę od czasu oczekiwania zgłoszenia w kolejce [SUGA 65], [BUZA 74], [POSN 73], [WOLI 79], [BARR 57], [GNED 71], [BACE 81]	

Biorąc pod uwagę istotne utrudnienie w analizie modelu, jakie z reguły pociągają za sobą takie odstępstwa, zagadnieniem podstawowym stało się określenie, które spośród nich powodują najistotniejsze zmiany własności rozpatrywanego systemu obsługi w stosunku do systemu konwencjonalnego. Odpowiedzi na to pytanie udziela się często w oparciu o badania symulacyjne bądź analityczne rozpatrzenie możliwie prostego strukturalnie przypadku (system jednokanałowy, rozkłady wykładnicze bądź regularne), w którym badane, pojedyncze odstępstwo może być uwzględnione. Dopiero stwierdzenie istnienia takich znaczących zmian własności uzasadnia - z punktu widzenia

praktycznego znaczenia ewentualnych rezultatów - bardziej dogłębne badania odpowiedniego, niekonwencjonalnego systemu.

Bardzo wyraźna jest aktualnie w literaturze przedmiotu tendencja do rozpatrywania różnorodnych, różnych od konwencjonalnych, systemów obsługi, zmierzających do "urealnienia" rozpatrywanych modeli [ROSE 75] i kierunek ten można uznać za jeden z najintensywniej rozwijanych.

Autorowi znana jest jednak tylko jedna praca [BHAT 69], w której podjęto próbę częściowego usystematyzowania takich modeli i podsumowania (wówczas jeszcze skromnego) dorobku w zakresie ich analizy.

W tabeli I.1 zestawiono typowe odstępstwa od założeń określających jednostanowiskowe, konwencjonalne systemy masowej obsługi wraz z podaniem podstawowych pozycji literaturowych, poświęconych badaniu odpowiednich modeli.

Niektóre z tych przypadków (oznaczone literą Z w tabeli) rozpatrywane już od dawna, nie wymagające specjalnych, nowych metod analizy i omawiane w opublikowanych monografiach, nazywać będziemy zmodyfikowanymi systemami konwencjonalnymi i skwitowane zostaną one tylko krótkimi wzmiankami. Więcej uwagi poświęcimy natomiast bardzo słabo dotąd rozpoznanej grupie systemów niekonwencjonalnych, wynikających z odstępstw związanych z wprowadzeniem zależności losowych pomiędzy poszczególnymi procesami stochastycznymi, opisującymi system masowej obsługi oraz z wprowadzeniem regulaminów obsługi, nie zachowujących pracy. Przedstawimy przykłady systemów uzasadniających rozpatrywanie takich właśnie przypadków odstępstw i omówimy ich skutki określone dotąd na drodze analitycznej bądź symulacyjnej.

Zmodyfikowane systemy konwencjonalne

W większości systemów rzeczywistych mamy do czynienia z ograniczonym buforem, w którym gromadzi się zgłoszenia oczekujące na rozpoczęcie obsługi. Zgłoszenia, które zastają bufor zapełniony, są tracone. Taki system masowej obsługi jest zawsze stabilny. Jeśli chodzi o wyznaczanie charakterystyk takich systemów (do których dołącza się zwykle również prawdopodobieństwo utraty zgłoszenia), to dobrze rozpoznany jest w zasadzie tylko, analizowany od dawna, przypadek M/M/n/N oraz klasyczny przypadek M/G/n/n, tj. system ze stratami ([GNED 71]).

Dopiero w ostatnich latach wyznaczono wiele podstawowych charakterystyk systemów M/G/1/L, GI/M/1/L oraz niektórych bardziej złożonych przypadków i to z reguły dla regulaminu naturalnego [BHAT 73], [HOKS 75], [LAVE 75], [TRUS 75].

Rozszerzono też na systemy z $N < \infty$ niektóre zależności pomiędzy charakterystykami związanymi z czasem a charakterystykami związanymi ze zgłoszeniami [SCHM 78].

Ograniczając się do rozkładów Coxa można działanie każdego systemu $K_1/K_1/n/L$ przedstawić w postaci ergodycznego łańcucha Markowa o skończonej liczbie stanów.

Taka reprezentacja często nie prowadzi jednak do rozwiązań analitycznych określających odpowiednie rozkłady ergodyczne i wymagane jest rozwiązanie bardzo dużych układów równań liniowych o specyficznej strukturze.

Efektywne metody numeryczne dla rozwiązywania takich zadań są aktualnie intensywnie rozwijane ([MARI 78A], [MARI 78B], [RAJU 77], [STEW 78]). Metody te obejmować mogą też przypadki złożonych regulaminów obsługi, np. regulamin z dynamicznym priorytetem [RAMA 80].

W systemach konwencjonalnych zakłada się całkowitą niezależność procesu generacji zgłoszeń od liczby zgłoszeń aktualnie znajdujących się w systemie. W praktyce często występują istotne odchyłki od tego założenia. Mogą one wynikać na przykład z faktu, iż łączna liczba zgłoszeń możliwych do wygenerowania jest ograniczona, a więc mamy do czynienia ze źródłem o skończonej wymiarowości.

Przypomnimy tu, że dla przypadku wykładniczego rozkładu (z parametrem λ) czasu pobytu poszczególnych zgłoszeń w źródle prawdopodobieństwo wygenerowania w odcinku czasu $(t_o, t_o + \Delta)$ liczby j zgłoszeń przez k-wymiarowe źródło, w którym w chwili t_o znajdowały się wszystkie zgłoszenia, określa zależność:

$$p_r\left\{\text{liczba zgłoszeń wygenerowanych} = j\right\} = \binom{k}{j}\left[1-e^{-\lambda\Delta}\right]^j e^{-(k-j)\lambda\Delta} \qquad (I.1.18)$$

Odpowiada to na przykład systemowi wielodostępnemu o ustalonej, ograniczonej liczbie terminali. Przy pracy w trybie konwersacyjnym liczba zgłoszeń (zleceń wyemitowanych przez użytkowników) nie może przekroczyć liczby terminali.

Typowym problemem prowadzącym do analizy źródeł o skończonej wymiarowości są problemy napraw parku maszynowego, od których analiza takich systemów wzięła często spotykaną nazwę "zadania konserwatora" [TAKA 62], [KOPO 73A], [KOPO 73B].

Na uwagę zasługują, rozwijane głównie z myślą o modelowaniu algorytmów szeregowania zadań w systemach operacyjnych, prace w zakresie jednokanałowych systemów priorytetowych ze źródłami o skończonej wymiarowości [JAIS 68] i jednowymiarowymi [WOLI 73], [CZAC 74], [TOMK 75], [WOLI 76A], [WOLI 78].

Wprowadzenie źródeł o ograniczonej wymiarowości istotnie zmienia własności systemu masowej obsługi[x]. Przede wszystkim system taki jest zawsze stabilny. Również zasadniczej zmianie ulegają wartości charakterystyk systemu. W pracy [BUZE 74] przedstawiono analizę błędu w określeniu średniego czasu oczekiwania, popełnionego przy rozpatrywaniu konwencjonalnego systemu M/G/1 zamiast analogicznego systemu z ograniczoną wymiarowością źródła zgłoszeń. Wpływ ten ilustruje zaczerpnięta z cytowanej pracy tabela I.2.

[x] Należy zaznaczyć, że każdy jednostanowiskowy system masowej obsługi z ograniczoną wymiarowością źródła zgłoszeń jest równoważny odpowiedniej zamkniętej sieci stanowisk obsługi, o których będziemy szerzej mówić w punkcie I.2.

Tabela I.2

Oszacowanie dokładności określenia średniego czasu pobytu $E(\varepsilon_{\mathcal{N}})$ zgłoszenia w systemie typu M/G/1 z $\mathcal{N}$ wymiarowym źródłem zgłoszeń przy użyciu średniego czasu pobytu $E(\varepsilon)$, wyznaczonego w konwencjonalnym systemie M/G/1 o identycznym rozkładzie czasów obsługi oraz identycznym stopniu wykorzystania stanowiska obsługi $\eta(\lambda)$. Dane wg [BUZE 74]

a) Błąd względny $\left[E(\varepsilon) - E(\varepsilon_{\mathcal{N}})\right]/E(\varepsilon_{\mathcal{N}})$ w funkcji stopnia wykorzystania stanowiska obsługi $\eta(\lambda)$ oraz wymiarowości źródła zgłoszeń N, przy wykładniczym rozkładzie czasów obsługi

$\mathcal{N}$ \ $\eta(x)$	.1	.3	.5	.7	.9
1	.117	.430	1.016	2.344	8.750
4	.028	.131	0.334	0.809	3.159
7	.017	.077	0.212	0.566	2.145
10	.012	.057	0.158	0.433	1.762
40	.003	.015	0.045	0.151	0.786
70	.002	.009	0.027	0.093	0.519
100	.001	.006	0.020	0.066	0.415
200	.001	.003	0.010	0.035	0.279

b) Minimalne wartości $\mathcal{N}$, przy których zdefiniowany wyżej błąd względny nie przekracza 5% w funkcji $\eta(\lambda)$ oraz współczynnika C_B zmienności rozkładu Coxa drugiego rzędu, określającego czas obsługi.

C_B \ $\eta(x)$	.1	.2	.3	.4	.5	.6	.7	.8
0	2	3	6	10	17	32	81	160
0,5	2	5	9	16	27	43	101	>200
1	3	6	12	21	38	70	142	>200
3/2	4	8	16	27	48	93	188	>200

Wskazując na fakt, że błąd rośnie ze wzrostem stopnia wykorzystania stanowiska obsługi oraz współczynnika zmienności rozkładu (przy stałej wymiarowości źródła zgłoszeń), osiągając w skrajnych przypadkach wartość kilkuset procent, Autorzy przestrzegają przed pomijaniem ograniczeń wymiaru źródła.

Poza tym, w przypadku źródeł skończenie wymiarowych, stopień wykorzystania stanowiska obsługi jest zależny od regulaminu obsługi.

Również w systemach ze źródłami nieskończenie wymiarowymi może wystąpić zjawisko zmiany intensywności źródła zgłoszeń w funkcji liczby zgłoszeń już w systemie obecnych (odpowiada to zmniejszaniu się liczby "klientów" dołączających do kolejki w jednostce czasu, przy wydłużaniu się tej kolejki).

W sensie formalnym efekt taki osiąga się bądź to drogą zmiany parametrów opisujących źródło zgłoszeń, bądź też przyjęcia, iż każde wygenerowane przez (nie modyfikowane) źródło zgłoszenie, może z pewnym, zależnym od liczby zgłoszeń w systemie, prawdopodobieństwem dołączyć do kolejki lub nie, co też zmienia efektywny strumień zgłoszeń. Szczególnym przypadkiem jest tu więc sytuacja, w której osiągnięcie odpowiednio dużej liczby zgłoszeń w systemie powoduje zablokowanie źródła do czasu, gdy chociaż jedno całkowicie obsłużone zgłoszenie opuści stanowisko obsługi. Taka szczególna sytuacja ograniczająca maksymalną liczbę zgłoszeń znajdujących się w systemie jest równoważna odpowiednio zdefiniowanej, zamkniętej sieci stanowisk obsługi (zauważmy, że przy strumieniu zgłoszeń różnym od strumienia Poissona przypadek taki jest różny od systemu GI/GI/n/L!). W literaturze przedmiotu rozpatruje się również często przypadek zmiany mocy stanowiska obsługi zależnie od liczby zgłoszeń obecnych w systemie. Zmiana ta może polegać na modyfikacji liczby działających kanałów obsługi lub mocy poszczególnych kanałów, łącznie z całkowitym "wyłączeniem" stanowiska obsługi przy małej liczbie zgłoszeń w systemie [HARR 67], [HARR 70].

Często rozpatruje się oddziaływanie na źródło zgłoszeń, w celu zmiany jego intensywności, stosownie do stanu systemu oraz oddziaływanie na moc stanowiska obsługi jako środki do optymalizacji efektywności działania systemu masowej obsługi, definiowanej w postaci funkcji (czasami o złożonej postaci) uwzględniającej koszty oczekiwania zgłoszeń, koszty utrzymania stanowiska obsługi w stanie gotowości do działania, koszty pracy stanowiska obsługi z różną mocą itp. Obszerną bibliografię z tego zakresu można znaleźć w pracy [CRAB 77].

Modele uwzględniające zmianę intensywności źródeł zgłoszeń oraz mocy stanowiska obsługi w zależności od liczby zgłoszeń w systemie prowadzą, w przypadku założenia strumienia zgłoszeń w postaci strumienia Poissona i wykładniczych czasów obsługi, do rozpatrywania procesów urodzin i śmierci, a więc procesów stochastycznych dobrze znanych. Przypadek ten jest szczególnie dobrze rozpoznany [GROS 74]. W szczególności konieczny i wystarczający warunek stabilności (zaczerpnięty z teorii procesów urodzin i śmierci) sprowadza się do wymagania, aby szereg

$$\sum_{j=1}^{\infty} \prod_{i=1}^{j} \frac{\lambda_{i-1}}{\mu_i} \qquad \text{(I.1.19)}$$

był zbieżny, gdzie $\lambda_i \Delta$, $\mu_i \Delta$ oznaczają odpowiednio prawdopodobieństwo zwiększenia bądź zmniejszenia (o jeden) liczby zgłoszeń w systemie w odcinku czasu $(t_o, t_o + \Delta)$, pod warunkiem że w chwili t_o w systemie znajdowało się "i" zgłoszeń. W szczególności więc system jest stabilny, gdy stosunek intensywności strumienia zgłoszeń do mocy stanowiska obsługi przybiera wartość mniejszą od jedności przy długości kolejki dążącej do nieskończoności.

W tym miejscu wspomnimy jeszcze o pewnej specyficznej postaci modyfikacji charakterystyk źródła zgłoszeń rozpatrywanych w pracach [ALEK 74], [CHOO 79]. Założono tam, że zgłoszenia napływają ze źródła strumieniem Poissona o intensywności λ do jednokanałowego systemu obsługi o zerowej pojemności bufora. Zgłoszenia, które zastają stanowisko obsługi zajęte, umieszczane są w źródle pomocniczym, w którym przebywają przez czas posiadający rozkład wykładniczy o parametrze ν. Zgłoszenie może powracać do źródła pomocniczego wielokrotnie. Charakterystyka źródła pomocniczego zawierającego k zgłoszeń zadana jest więc zależnością (I.1.18). Taki przypadek odzwierciedla tryb pracy centrali telefonicznej (ponawianie próby uzyskania połączenia) oraz kanałów transmisji danych w sieciach komputerowych. Łączna intensywność źródła zgłoszeń zależy więc od liczby zgłoszeń znajdujących się w źródle pomocniczym, pełniącym rolę swoistej kolejki.

Dalsze wyniki prac dotyczących takich systemów wraz z listą dodatkowych pozycji literaturowych można znaleźć w pracach [FALI 79A], [FALI 79B], [FALI 80].

W konwencjonalnych systemach masowej obsługi zakłada się pojedyncze napływanie zgłoszeń i pojedynczą obsługę. Takie założenie jest często niesłuszne, np. jeżeli w systemie komputerowym (czy sieci komputerowej) jednostką transmisji informacji jest znak (słowo, rekord, pakiet), to zgłoszenie żądania transmisji programu (zbioru danych) jest równoznaczne z jednoczesnym wygenerowaniem wielu zgłoszeń o transmisję. Możemy też mieć do czynienia z sytuacją odwrotną, gdy pojedyncze zgłoszenia (np. informacje o stanie różnych czujników pomiarowych) są zbierane i buforowane lokalnie, a następnie przesyłane grupowo do jednostki centralnej.
Grupowe pojawianie się bądź obsługa zgłoszeń są typowe dla procesów transportowych i niektórych procesów produkcyjnych, np. procesów termicznych (nagrzewanie, wypalanie, suszenie), którym poddawana jest równocześnie grupa wyrobów.

Systemy masowej obsługi uwzględniające tę specyfikę rozpatrywane są od dawna (głównie dla przypadku jednokanałowego) przy użyciu standardowych metod, najczęściej metody włożonych łańcuchów Markowa. Zachowują one (w odpowiednio zmodyfikowanej postaci) warunek stabilności właściwy systemom konwencjonalnym, tj. wymaganie, by ilość pracy dostarczona w jednostce czasu ze źródła była mniejsza od łącznej mocy stanowiska obsługi. Rozważane były różne warianty reguł obsługi (np. założenie, że obsługa może

rozpocząć się dopiero, gdy grupa oczekujących zgłoszeń będzie odpowiednio liczna). W związku z istnieniem dość kompletnych rezultatów (patrz [SAAT 61], [COHE 69], [GROS 74], [MURA 72B], [KUCZ 73]) w ostatnich latach zainteresowanie tymi systemami maleje - z wyjątkiem prac związanych z zastosowaniem złożonych regulaminów obsługi [DIMI 79], [WISZ 79], ukierunkowanych na aplikacje (np. [KADO 76], [HALA 74]) względnie łączących różne odstępstwa od konwencjonalnych założeń [MURA 72A], [HARR 70].

I.1.3. Niekonwencjonalne systemy jednostanowiskowe

Przejdziemy obecnie do omówienia zbioru niekonwencjonalnych systemów obsługi wymienionych w tabeli I.1, dzieląc je wstępnie na trzy grupy: systemy o parametrach zmiennych w czasie oraz systemy o parametrach stacjonarnych, zachowawcze bądź też niezachowawcze.

a) Systemy o parametrach zmiennych w czasie

Przypadki systemów o parametrach zmiennych w czasie spotyka się bardzo często, szczególnie, gdy chodzi o parametry źródeł zgłoszeń. Typowymi przykładami mogą tu być: intensywność ruchu lotniczego, połączeń telefonicznych, czy też liczba wykorzystywanych terminali systemu wielodostępnego, przy czym w wielu aplikacjach zmiany tych parametrów mają charakter periodyczny (dobowy, tygodniowy, miesięczny). Podobnie znaleźć można przykłady zmiennej z czasem mocy stanowiska obsługi (np. w wyniku zmian liczebności personelu, ograniczeń energetycznych). Ze względu na identyczną metodę badania, przypadki te omówimy łącznie.

Można wyróżnić dwa zasadnicze kierunki prac w tym zakresie:

- Jeśli uwzględnić periodyczny charakter zmian charakterystyk źródła, wówczas rozpatrywany system masowej obsługi może być niestabilny bądź też substabilny, dążąc do swoistego "cyklu granicznego", jeśli chodzi o jego charakterystyki (przy nieperiodycznych, deterministycznych zmianach parametrów systemu brak stacjonarnego stanu granicznego).

  W tym przypadku w analizie mogą być przydatne takie metody, które w systemach konwencjonalnych służą do badania stanów nieustalonych. Nie zagłębiając się w bardziej szczegółowe omawianie wyników w tym zakresie, odsyłamy zainteresowanych do przykładowych pozycji, wybranych z niezbyt zresztą bogatej literatury przedmiotu [SAAT 61], [NEWE 68], [NEUT 71], [KOOP 72], [MOOR 75].

- Rozpatrywany był również inny przypadek, w którym zakładano, że dany jest pewien nieprzywiedlny, ergodyczny łańcuch Markowa o skończonej liczbie "m" stanów i macierzy przejść $Q = [q_{ij}]$, $i,j=1,2,\ldots,m$. Przejście do stanu "i" łączyło się przy tym z ustaleniem wartości intensywności strumienia zgłoszeń λ_i oraz mocy stanowiska obsługi c_i, ogólnie rzecz biorąc różnych dla każdego i. Każdemu stanowi "i" odpowiada więc średni czas obsługi zgłoszeń $[\mu_i]^{-1}$.

Dla takich systemów, których badania zainicjowano w [EISE 63] a dalsze wyniki przedstawiono m.in. w [NEUT 71], [YECH 71], [PURD 74], [NEUT 78A], [NEUT 78B], może być osiągany stacjonarny stan graniczny.

Dla przypadku systemu jednokanałowego z wykładniczymi czasami obsługi oraz zgłoszeniami nadchodzącymi strumieniem Poissona wykazano, że warunek stabilności ma postać [NEUT 78A]:

$$\rho = (\Pi\Lambda)(\Pi M)^{-1} < 1,$$

gdzie:

Π - jest m-wymiarowym wektorem ergodycznych prawdopodobieństw stanów wprowadzonego łańcucha Markowa (zawsze istniejącym),

Λ, M - są m-wymiarowymi wektorami odpowiednio o składowych λ_i oraz μ_i $i = 1,2,\ldots,m$.

Rozpatrywany model pozwala więc badać przypadki czasowego (w pewnych stanach) przeciążenia systemu, tj. przypadki, gdy

$$\exists_{i \in [1,\ldots,m]} \left\{\lambda_i > \mu_i\right\}.$$

Określono algorytmy pozwalające na stosunkowo łatwe, numeryczne wyznaczanie charakterystyk systemów. Dostępne są też analogiczne wyniki dla systemów wielokanałowych przy pewnych dodatkowych uproszczeniach, np. [NEUT 79].

Na przykładach numerycznych wskazano [NEUT 78A], że takie charakterystyki systemu, jak np. wartość oczekiwana czy wariancja liczby zgłoszeń oczekujących w kolejce, przyjmują w rozpatrywanym przypadku wartości większe (kilku- lub kilkunastokrotnie) aniżeli identyczne wskaźniki dla konwencjonalnych systemów M/M/1, o identycznej wartości stopnia wykorzystania stanowiska obsługi ρ.

Tak więc, gdyby zaniedbać fakt zmienności intensywności strumienia zgłoszeń (i mocy stanowiska obsługi) i drogą długich obserwacji wyznaczyć tylko ich wartości oczekiwane, a następnie w oparciu o zależności prawdziwe dla systemów konwencjonalnych prognozować charakterystyki omawianego, niekonwencjonalnego przypadku, to prawidłowo oceniony zostałby warunek stabilności i stopień wykorzystania stanowiska obsługi, natomiast popełniony zostałby duży błąd w ocenie długości kolejki, czasów oczekiwania itp.

b) Systemy zachowawcze o stałych parametrach

W wielu systemach rzeczywistych nie jest uzasadnione przyjmowanie założenia, że elementy ciągu $\{a^i\}$ i elementy ciągu $\{b^i\}$ są wzajemnie niezależne losowo, a także niezależne losowo w ramach każdego z tych ciągów.

I tak na przykład, w tekście programu (traktowanego jako zbiór zgłoszeń - wierszy tekstu - które należy kolejno obsłużyć - przesłać) występują zawsze sekwencje definiujące zadanie o podobnej, ściśle określonej,

strukturze. Przy transmitowaniu podprogramów pisanych przez różnych programistów uwidoczniają się ich upodobania do umieszczenia większej lub mniejszej liczby znaków w wierszu.

W przypadku pracy z końcówki interaktywnej, dłuższemu czasowi generacji zgłoszenia (pisania komunikatu) odpowiada zwykle większa jego długość, a więc większy czas obsługi. W rozdziale I.2 omówimy zagadnienie transmisji pakietów w sieciach komputerowych prowadzącej w najprostszym przypadku również do analizy systemu jednostanowiskowego z zależnością wartości b^i od odpowiadających im wartości a^i.

Niech w dyskretnym procesie produkcyjnym generacja zgłoszeń oznacza przygotowanie półproduktu (w oparciu o zawsze niezerowy zapas surowców), natomiast ich obsługa niech odpowiada pracom wykończeniowym.

Łatwo wskazać przykłady, gdy lepsze (a więc bardziej oczekiwane) przygotowanie półproduktu skraca operacje wykończeniowe lub odwrotnie, zmienna jakość surowca (np. odchyłki wymiarów) powoduje równocześnie bądź to wydłużenie, bądź skrócenie czasu trwania obu operacji.

Podwaliny analizy stabilności systemów obsługi typu G/G/n położył R.M. Loynes [LOYN 62], rozpatrując jednostanowiskowe systemy obsługi z regulaminem naturalnym. Przyjmując, iż w przedmiotowym systemie masowej obsługi dwuwymiarowy proces stochastyczny $Y = \{a^i, b^i\}_{i=1,2,\ldots}$; $a^i, b^i \in R^+$ jest stacjonarny, gdzie $A^i(x) = A(x)$, $B^i(x) = B(x)$ wykazano, że warunek stabilności ma postać:

$$\rho = \frac{E(b)}{nE(a)} < 1. \qquad (I.1.20)$$

W cytowanej pracy wskazano na przykładzie, iż w przypadku wielokanałowym możliwa jest, przy spełnieniu warunku (I.1.20), substabilność rozpatrywanego systemu masowej obsługi, co było wykluczone w przypadku systemów GI/GI/n. Postawiona została też hipoteza, że jest to możliwe tylko w przypadku deterministycznych ciągów wartości a^i i b^i.

W [BORO 72] wykazano prawdziwość warunku stabilności (I.1.20) również dla grupowego napływania zgłoszeń, natomiast w [MIYA 77] rezultat ten uogólniony został na przypadek, gdy (przy pewnych założeniach dodatkowych) stosowany jest dowolny, zachowujący pracę regulamin obsługi.

W cytowanych pracach wykazano też, że jeśli $\rho > 1$, to system masowej obsługi jest niestabilny, nie rozstrzyga się natomiast problemu stabilności przy $\rho = 1$.

Przypomnijmy, że systemy konwencjonalne były stabilne przy $\rho = 1$ jedynie i wyłącznie w przypadku D/D/1, a więc dla regularnych rozkładów zmiennych losowych a^i oraz b^i.

Dla systemów niekonwencjonalnych nie należy spodziewać się, przy $\rho = 1$, ogólnych reguł ważnych bez względu na postać zależności losowej występującej w procesie stochastycznym Y i oddzielnym badaniom muszą podlegać różne rodzaje tej zależności.

Reprezentatywnym przykładem może tu być rozpatrzony w pracy [CALO 77] (jako część analizy bardziej złożonego modelu) przypadek jednokanałowy, w którym zmienne losowe $\{a^i\}$ $i = 1,2,\ldots$, są niezależne, o identycznym rozkładzie, posiadającym dystrybuantę $A_c(x)$, natomiast poszczególne realizacje zmiennej losowej b^i określone są jako:

$$b^i = \psi a^i, \quad i = 1,2,\ldots, \quad \psi \leq 1. \qquad (I.1.21)$$

Systemy takie noszą nazwę systemów z samoregulacją [COMO 68]. Rozpatrywany jest więc przypadek deterministycznej zależności pomiędzy wartościami a^i oraz b^i. Dla $\psi = 1$ wykazano, że jeśli zmienne losowe a^i są ograniczone, tj.:

$$\exists_{\alpha < \infty} \; \forall_{x > \alpha} \{A_c(x) = 1\}, \qquad (I.1.22)$$

wówczas system jest stabilny i rozkład czasu oczekiwania w stacjonarnym stanie granicznym posiada dystrybuantę

$$W(x) = \begin{cases} 1 - A_c(\alpha - x) & 0 \leq x < \alpha, \\ 1 & x \geq \alpha, \end{cases} \qquad (I.1.23)$$

o wartości oczekiwanej $E(w) = \alpha - E(a_c)$.

Jeżeli natomiast zmienne losowe a^i są nieograniczone, wówczas system jest niestabilny.

Oczywiście w przypadku $\psi < 1$ system z samoregulacją jest zawsze stabilny, natomiast w przypadku $\psi > 1$ zawsze niestabilny.

Dla systemów obsługi G/G/n udowodnione zachodzenie zależności między charakterystykami związanymi ze zgłoszeniem oraz charakterystykami związanymi z czasem, analogicznych do zależności omawianych w przypadku systemów GI/GI/n, aczkolwiek przybierających zwykle nieco bardziej złożoną postać. Problematyka ta została obszernie potraktowana w [BORO 72], [MIYA 79], a także w [BRUM 72], [MORI 80], [HEYM 80].

Co więcej, zarówno zależność (I.1.16), jak też wyrażona wzorem (I.1.17) stałość średniego czasu oczekiwania bez względu na stosowany regulamin obsługi (nieprzerywalny, nie posiadający informacji o czasie obsługi indywidualnych zgłoszeń) pozostaje w mocy ([SCHR 70]) również dla systemów typu $G/G/1/\infty/\mathcal{R}$, określając, podobnie jak to miało miejsce w przypadku systemów konwencjonalnych, więzy na klasę możliwych do uzyskania rozkładów czasów obsługi. Dla systemów $G/G/c/\infty/\mathcal{R}$ stałość sumy średnich czasów oczekiwania zgłoszeń napływających w ciągu okresu czynnego wykazano w [VASI 77].

Można więc powiedzieć, że wprowadzenie zależności losowych w ciągach $\{a^i\}$, $\{b^i\}$ bądź między elementami tych ciągów nie prowadzi do zasadniczych jakościowych zmian w stosunku do systemów konwencjonalnych.

Przejdziemy teraz do omówienia opublikowanych wyników badań charakterystyk takich systemów, dotyczących jak dotąd niemal wyłącznie przypadków jednokanałowych i naturalnego regulaminu obsługi.

W cyklu prac [CONO 68], [CONO 69], [CONO 74] przebadano analitycznie własności systemów z samoregulacją, przy założeniu wykładniczego rozkładu odstępów między chwilami generacji kolejnych zgłoszeń, porównując je z konwencjonalnymi systemami M/M/1. W pracach [CONO 79] na drodze analitycznej (z dokładnością do współczynników określanych przy użyciu procedury numerycznej) oraz na drodze symulacyjnej [MITC 79] badano natomiast przypadek ogólniejszy, gdy zależność czasu obsługi i czasu generacji określona jest rozkładem dwuwymiarowym o funkcji gęstości f(x,y) takiej, że:

$$\bar{f}(\Theta,s) = \int_0^\infty\int_0^\infty e^{-sx}e^{-\Theta y}f(x,y)dxdy = (1-\varkappa)\left[(1-\frac{\Theta}{\lambda})(1+\frac{s}{\mu})-\varkappa\right]^{-1}, \qquad (I.1.24)$$

gdzie: $0 \leq \varkappa < 1$ jest współczynnikiem korelacji czasu obsługi i czasu generacji, $\mu, \lambda > 0$.
Rozpatrywany system jest stabilny, jeśli zachodzi:

$$\rho = \frac{E(b)}{E(a)} = \frac{\int_0^\infty\int_0^\infty yf(x,y)dxdy}{\int_0^\infty\int_0^\infty xf(x,y)dxdy} < 1. \qquad (I.1.25)$$

W badaniach symulacyjnych przedział zmienności współczynnika korelacji rozszerzono dodatkowo o zakres $-0.25 \leq \varkappa < 0$.

Uzyskane w cytowanych pracach wyniki wskazują jednoznacznie na fakt silnego zmniejszania się wartości średniej czasu oczekiwania ze wzrostem współczynnika korelacji $\varkappa$. Co więcej, równocześnie występuje jeszcze silniejszy spadek wariancji czasu oczekiwania. Zjawiska te ilustruje zaczerpnięta z [CONO 79] tabela I.3.
Fakt zmniejszania się średniego czasu oczekiwania w przypadku dodatniej korelacji zmiennych losowych b^i oraz a^i stwierdzono też, dla innej postaci uzależnienia, w [KREJ 79].

W pracy [JACO 80] użyto następującego algorytmu dla generowania ciągu skorelowanych zmiennych losowych:

Tabela I.3

Średni czas obsługi i wariancja czasu obsługi zgłoszeń

ρ / ϰ		0.1		0.3		0.5		0.7		0.9		0.99	
		E(w)	√V(w)	E(w)	√V(w)	E(w)	√V(w)	E(w)	√V(w)	E(w)	√V(w)	E(w)	√V(w)
0	a)	.111	.111	.429	.429	1.000	1.000	2.333	2.333	9.000	9.000	99.00	99.00
	b)												
0,1	a)	.111	.110	.425	.413	.975	.939	2.220	2.138	8.275	8.118	89.36	89.10
	b)	.111	.110	.426	.419	.983	.959	2.246	2.182	8.340	8.198	89.40	89.20
0,3	a)	.111	.108	.418	.385	.929	.824	2.005	1.760	6.852	6.366	69.95	69.31
	b)	.111	.109	.422	.402	.951	.877	2.075	1.882	7.039	6.596	70.23	69.60
0,5	a)	.111	.105	.411	.359	.887	.721	1.807	1.411	5.483	4.640	50.68	49.52
	b)	.111	.108	.417	.384	.919	.798	1.911	1.589	5.773	5.004	51.15	50.00
0,7	a)	.110	.103	.405	.337	.850	.633	1.632	1.107	4.207	2.981	31.58	29.74
	b)	.111	.107	.413	.368	.889	.724	1.754	1.310	4.564	3.433	32.23	30.41
0,9	a)	.110	.101	.399	.317	.817	.563	1.487	0.871	3.116	1.547	13.03	10.04
	b)	.111	.106	.409	.352	.860	.654	1.610	1.054	3.456	1.939	13.78	10.85
0,99	a)	.110	.100	.397	.309	.805	.536	1.432	0.791	2.744	1.114	5.706	1.781
	b)	.111	.105	.407	.345	.848	.626	1.549	0.952	3.022	1.376	6.23	2.26

a) w jednokanałowym systemie obsługi, w którym proces obsługi jest skorelowany z procesem nadchodzenia zgłoszeń (wg [CONO 79]),

b) w drugim węźle dwuetapowego pojedynczego systemu obsługi, z zależnymi losowo czasami obsługi, omawianego w punkcie I.2.b (dane wg [CHOO 80]).

Niech $\{I_l\}$ oraz $\{K_n\}$ będą niezależnymi ciągami zmiennych losowych, przyjmujących wartości 0 lub 1, takich, że:

$$P_r\{I_j = 1\} = 1 - \xi, \quad 0 \leq \xi \leq 1; \qquad j = 0,1,\ldots,$$

$$P_r\{K_j = 1\} = 1 - \psi, \quad 0 \leq \psi < 1;$$

natomiast $\{\varphi_j\}$ będzie ciągiem niezależnych zmiennych losowych o rozkładzie wykładniczym z parametrem ν.
Ciąg zmiennych losowych $\{x_j\}$ taki, że

$$\left.\begin{aligned} x_j &= \xi\varphi_j + I_j\alpha_{j-1} \\ \alpha_j &= \psi\alpha_{j-1} + K_j\varphi_j \end{aligned}\right\} \quad j = 1,2,\ldots \qquad (I.1.26)$$

ma tę własność, iż współczynnik korelacji

$$Corr(x_1, x_{j+1}) = \psi^j(1-\xi)\left[\xi(1-\psi) + \psi(1-\xi)\right]. \qquad (I.1.27)$$

Dalsze własności tego ciągu omówiono szczegółowo w pracy [JACO 77]. Dla takiej postaci zależności rozpatrywano wpływ zarówno korelacji elementów ciągów $\{a^i\}$ oraz $\{b^i\}$, jak też wzajemnych korelacji ich elementów ze szczególnym uwzględnieniem pracy przy dużym stopniu wykorzystania stanowiska obsługi.

Wykazano, że przy $\eta(\lambda) = \varrho \to 1$ rozkład zmiennej losowej $u=(1-\varrho)w$ (gdzie "w" jest zmienną losową określającą czas oczekiwania w stacjonarnym stanie granicznym) jest w przybliżeniu wykładniczy, co stanowi uogólnienie znanego w przypadku systemów konwencjonalnych wyniku Kingmana [KING 62]. Wartość E(u) jest przy tym różna od wartości uzyskiwanej przy identycznych parametrach E(a) oraz E(b) w systemie konwencjonalnym.

Stwierdzono, iż wartość średnia czasu oczekiwania rośnie ze wzrostem korelacji elementów każdego z ciągów $\{a^i\}$ $\{b^i\}$, natomiast maleje ze wzrostem korelacji wzajemnej między elementami tych ciągów, co jest zgodne z wynikami omawianymi powyżej.

Przedstawione, przybliżone, rozwiązanie analityczne wykazało pełną poprawność jakościową i stosunkowo dobrą dokładność w porównaniu z wynikami uzyskanymi na drodze symulacji cyfrowej [BOOM 78].

Wspomnimy jeszcze o kilku dalszych pracach dotyczących podobnych systemów masowej obsługi. Stosunkowo wcześnie wzbudził zainteresowanie model uwzględniający podział napływających zgłoszeń na K klas przy założeniu, że czas obsługi zgłoszenia zależy od jego klasy oraz od klasy zgłoszenia poprzednio wygenerowanego [NEUT 66], [CINL 67A], [PURD 75] bądź też, że odstęp czasu pomiędzy wygenerowaniem dwóch kolejnych zgłoszeń zależy od typu obu tych zgłoszeń [CINL 67B].

Inny przypadek specyficznego typu zależności rozpatrywano w [GRAN 75], analizując zmodyfikowany system D/G/1, w którym zgłoszenia mogą nadchodzić w sposób niezupełnie regularny. W szczególności, o ile w systemie D/G/1 chwila nadejścia zgłoszenia $\tau^{[i]}$ byłaby określona jako $\tau^{[i]} = (i-1)E(a)$, o ile pierwsze zgłoszenie nadeszło w chwili $\tau^{[1]} = 0$, to w rozpatrywanym systemie $\tau^{[i]} = (i-1)E(a) + \delta_i$, gdzie δ_i, $i \geq 0$ są zmiennymi losowymi niezależnymi, o jednakowych rozkładach, ograniczonymi $|\delta_i| < d < \infty$, o zerowej wartości oczekiwanej $E(\delta_i) = 0$. Wyznaczono w sposób przybliżony niektóre parametry rozkładu czasu oczekiwania dla systemu stabilnego oraz oszacowania parametrów rozkładu czasu oczekiwania i-tego zgłoszenia w przypadku niestabilnym.

W pracach [PEAR 67], [PEAR 68] wykorzystano jeszcze inną postać zależności losowej. Przyjęto mianowicie, iż mając do dyspozycji ciąg zmiennych losowych niezależnych, o identycznych rozkładach, u_i, $i=1,2,\ldots$, tworzono zmienne losowe skorelowane v_i jako

$$v_i = \sum_{j=0}^{p} f_{p-j}(u_{i+j}), \quad p \geq 0 \qquad (I.1.28)$$

gdzie f_i są dowolnymi funkcjami zapewniającymi nieujemność zmiennej losowej v_i.

Rozpatrzono modyfikację systemu M/M/1, w którym rozkład odstępów między zgłoszeniami (lub czasów obsługi zgłoszeń) określony był w powyższy sposób.

Na zakończenie omawiania systemów jednokanałowych wspomnimy jeszcze o pracach [MURA 72A], [MOHA 72], gdzie rozpatrzono stany nieustalone w specyficznym systemie obsługi, w którym zgłoszenia mogą napływać tylko w pewnych, dyskretnych chwilach czasu t_i, $i = 1,2,\ldots$, przy czym prawdopodobieństwo napłynięcia zgłoszenia w chwili t_i zależy od tego czy w chwili t_{i-1} zgłoszenie napłynęło, czy też nie oraz pracy [BLOM 73], gdzie modelowano pewien wariant pracy zbiornika.

Wszystkie prezentowane wyżej zadania mogą oczywiście pojawić się również w przypadku wielokanałowych systemów obsługi; jak dotąd brak jednak informacji o wynikach prac w tym zakresie.

W publikacji [NISH 77] podjęto natomiast bardzo interesujący problem wielokanałowego systemu obsługi, w którym występuje skorelowanie czasów obsługi zgłoszeń obsługiwanych równolegle w różnych kanałach obsługi, co może być wynikiem wspólnego korzystania przez te kanały z pewnych wspólnych (ograniczonych) zasobów bądź też współpracy tych kanałów.

Rozpatrzono przypadek zgłoszeń napływających strumieniem Poissona oraz zarówno skończonej, jak i nieskończonej pojemności bufora wejściowego (N skończone lub nieskończone) przy specyficznym, wielowymiarowym, wykładniczym, rozkładzie czasów obsługi, uzyskując na drodze analitycznej charakterystyki systemu.

c) <u>Niezachowawcze systemy o stałych parametrach</u>

Wszystkie omawiane dotąd systemy masowej obsługi były systemami zachowawczymi. Stąd też, mimo różnych modyfikacji założeń dotyczących pracy źródeł zgłoszeń, czy też stanowiska obsługi, praktycznie nie ulegały zmianie warunki stabilności (tam, gdzie stacjonarny stan graniczny mógł być osiągany tylko przy pewnych parametrach systemu obsługi), zawsze też zachodziła równość:

$$\Xi_R(\lambda) = \varrho_R(\lambda). \qquad (I.1.29)$$

Co więcej, w przypadku, gdy praca źródła zgłoszeń była niezależna od przebiegu obsługi i $N = \infty$, wówczas wartość $\eta_R(\lambda)$ była niezależna od zastosowanego regulaminu obsługi i była funkcją liniową intensywności strumienia zgłoszeń

$$\eta(\lambda) = \lambda E(b). \qquad (I.1.30)$$

Obecnie przedyskutujemy takie odstępstwa od założeń poczynionych przy definiowaniu konwencjonalnych systemów masowej obsługi, które prowadzą do systemów nie posiadających podanych wyżej własności. Rozpatrzmy problem zależności czasu obsługi zgłoszenia od czasu oczekiwania na jej rozpoczęcie. Zjawiska takie obserwuje się w procesach metalurgicznych, gdzie wytapiana w procesie konwertorowo-tlenowym (lub martenowskim) stal jest rozlewana do wlewnic, a następnie poddawana krystalizacji przez czas określony wskazaniami technologicznymi. Po okresie krystalizacji wlewnice zostają usunięte a zakrzepłe, przynajmniej częściowo, bloki stalowe (wlewki) transportowane są do oddziału pieców wgłębnych w celu podgrzania, tak by uzyskać jednakową (teoretycznie) temperaturę w całej objętości, niższą od temperatury topnienia stali, ale pozwalającą na prawidłowe przewalcowanie wlewków na walcarce-zgniataczu.

Ze względu na stochastyczny charakter strumienia wlewków dostarczanych do oddziału pieców wgłębnych, podobnie jak czasów nagrzewania wlewków w piecach, tworzy się przed nimi kolejka wlewków oczekujących na załadunek. Wlewki czekając, tracą ciepło, co powoduje konieczność wydłużenia czasu ich nagrzewania [BUZA 74], [SUGA 65].

Na uzależnienie czasu obsługi od czasu oczekiwania zwrócono uwagę [LIBU 71] w przypadku systemów informatycznych, np. systemów kierowania ruchem lotniczym, gdzie wydłużenie czasu oczekiwania na wykonanie korekcji lotu wymaga policzenia w procesie obsługi dodatkowych programów interpolacji położenia lub dodatkowych odwołań do stacji obserwacyjnych. Wskazuje się też na zmianę czasu potrzebnego dla udzielenia pierwszej pomocy medycznej w funkcji czasu oczekiwania na nią oraz wydłużenie czasu gaszenia pożaru (np. lasu [POSN 73]) przy opóźnieniu początku akcji gaśniczej.

Wypada wreszcie wspomnieć o pewnej uproszczonej wersji zależności czasu obsługi od czasu oczekiwania, jaką jest rezygnacja zgłoszenia z obsługi przy wydłużaniu się czasu oczekiwania na jej rozpoczęcie, częsta w przypadku systemów usługowych, transportowych (zepsucie się towaru).

Przedstawimy pokrótce sposób definiowania systemów obsługi z czasem obsługi zależnym od czasu oczekiwania oraz wyniki ich badań.

Postać zależności czasu obsługi od czasu oczekiwania na jej rozpoczęcie jest określona poprzez zadanie warunkowej dystrybuanty $B(x/w)$ definiującej rozkład czasu obsługi zgłoszenia pod warunkiem, iż oczekiwało ono w kolejce przez czas w. Stosunkowo najwcześniej rozpatrzono jednokanałowe systemy ze wspomnianą wyżej uproszczoną zależnością czasu obsługi od czasu oczekiwania, które są zawsze stabilne przy ograniczeniu dopuszczalnego czasu oczekiwania. Analizę przypadku, w którym zgłoszenia tworzą strumień Poissona, dopuszczalny czas oczekiwania jest zadany w postaci stałej D, oraz $B(x/w)$ ma postać:

$$B(x/w) = \begin{cases} 1-e^{-\mu x} & w \leq D, \quad \mu > 0, \quad D > 0, \\ \mathbf{H}(x) & w > D, \end{cases} \qquad (I.1.31)$$

gdzie $\mathbf{H}(x)$ jest jedynką Heaviside'a

przedstawiono w [BARR 57]. Uogólnienie tego modelu na przypadki:

- dowolnego rozkładu czasu obsługi zgłoszeń, których czas oczekiwania nie przekroczy D lub
- wielokanałowego stanowiska obsługi przy regularnym, a także wykładniczym rozkładzie wartości dopuszczalnego czasu oczekiwania można znaleźć w [GNED 71].

Przypadek dowolnego rozkładu czasu obsługi zgłoszeń, które nie opuściły kolejki oraz wykładniczego rozkładu wartości dopuszczalnego czasu oczekiwania rozpatrzono w [SUBB 66], a wyniki przytoczono w [JAIS 68]. W [GNED 71] można też znaleźć ogólne sformułowanie równania całkowego spełnianego przez dystrybutantę rozkładu czasu oczekiwania dla przypadku dowolnego rozkładu czasu obsługi zgłoszeń pozostających w kolejce oraz dowolnego rozkładu dopuszczalnego czasu oczekiwania w systemie jednokanałowym, natomiast efektywną metodę znajdowania charakterystyk w takim ogólnym przypadku podano w [BACE 81].

Przegląd rezultatów uzyskanych dla bardziej ogólnych postaci zależności czasu obsługi od czasu oczekiwania do roku 1978 zawiera [WOLI 79] i w tym miejscu podamy tylko zasadnicze cechy rozpatrywanych systemów.

W pracy [BUZA 74] rozpatrzono system jednokanałowy ze zgłoszeniami tworzącymi strumień Poissona i zależnością czasu obsługi od czasu oczekiwania o postaci:

$$b_w = \begin{cases} B_s = D & \text{dla} \quad w \leq D, \\ B_L > B_s & \text{dla} \quad w > D. \end{cases} \qquad D > 0 \qquad (I.1.32)$$

Analizując regulamin naturalny (FIFO) oraz regulamin odwrotny (LIFO) stwierdzono, że wydajności graniczne systemu przy obu tych regulaminach są zasadniczo różne ($T_{LIFO} < T_{FIFO}$). Co więcej, wydajność graniczna systemu, w którym obowiązuje jeden z tych regulaminów obsługi, może ulec powiększeniu drogą wprowadzenia możliwości oczekiwania stanowiska obsługi na napłynięcie nowego zgłoszenia przez pewien okres czasu Δt, w przypadku gdy w kolejce znajdują się wyłącznie zgłoszenia, których czas oczekiwania jest większy aniżeli D.

Autor wskazał, iż zmiany wydajności granicznej wynikające ze zmiany regulaminu obsługi są duże (w zamieszczonym przykładzie obliczeniowym sięgały one 50%) i najgorszy z rozpatrywanych okazał się regulamin naturalny.

Z wyjątkiem specyficznych przypadków wskazanych w [BUZA 74] brak rozwiązań analitycznych dla regulaminów obsługi innych aniżeli naturalny, a nawet warunków stabilności, co podważa możliwość efektywnego poszukiwania rozwiązań stacjonarnych na drodze symulacyjnej (jako że ogólnie rzecz biorąc, nie wiadomo czy przy zadanych dla danego przebiegu symulacyjnego parametrach rozwiązanie stacjonarne w ogóle istnieje).

Również wyniki w zakresie regulaminu naturalnego były fragmentaryczne. Wyznaczone, przy stosunkowo silnych ograniczeniach dodatkowych, warunki stabilności [SUGA 65], [CALL 73], [TWEE 75] sprowadzają się do wymagania, aby przy $w \to \infty$ średni czas obsługi był mniejszy aniżeli średni odstęp pomiędzy chwilami generacji kolejnych zgłoszeń. Analityczne wyniki w zakresie charakterystyk systemów z czasem obsługi zależnym od czasu oczekiwania, dotyczące wyłącznie przypadku jednokanałowego oraz zgłoszeń nadchodzących strumieniem Poissona, uzyskane zostały przy specyficznych postaciach dystrybuanty $B(x/w)$ [MUDR 61], [POSN 73], [LIBU 71]. Przy użyciu metod przybliżonych badano też przypadek źródła o skończonej wymiarowości [SWEN 59].

Wprost z definicji rozpatrywanej klasy systemów widać, że jednostkowy efektywny czas obsługi przypadający na indywidualne zgłoszenie nie jest tu wartością stałą i w zależności od postaci dystrybuanty $B(x/w)$ oraz zastosowanego regulaminu obsługi może być funkcją o różnych własnościach. Dodatkowo, w przypadku wprowadzenia oczekiwania stanowiska obsługi na napłynięcie nowego zgłoszenia mimo niepustej kolejki, uzyskujemy $\Xi_R(\lambda) \neq \eta_R(\lambda)$.

Systemy masowej obsługi z czasem obsługi zależnym od czasu oczekiwania na jej rozpoczęcie różnią się więc zasadniczo od systemów konwencjonalnych i wobec stosunkowo skromnych wyników ich analizy zasługiwały na dokładniejsze przebadanie.

Szereg dalszych rezultatów w tym zakresie przedstawiono w pracach [WOLI 80B], [ZSAK 79], [ZSAK 80].

Analiza systemów z czasem obsługi zależnym od czasu oczekiwania, łącznie z pewnymi metodami przybliżonymi, przydatnymi dla badania przypadków, w których niedostępne są rozwiązania analityczne, przedstawiona zostanie w rozdziałach II i III niniejszej pracy.

Czas obsługi zgłoszenia może też zależeć od kolejności zgłoszeń. Najprostszy przykład takiej zależności polega na tym, że pierwsze zgłoszenie obsługiwane w czasie każdego okresu czynnego posiada czas obsługi różny od wszystkich dalszych zgłoszeń, co może wynikać np. z mniejszej sprawności stanowiska obsługi w czasie pierwszej, po bezczynności, obsługi. Analizę takiego przypadku przeprowadzono w [PINC 59A] oraz [WELC 64]. Doczekał się on też wielu dalszych rozważań, z których najnowsze i chyba najogólniejsze zawiera [MINH 80] (w pracy tej założono, że również rozkład

czasu między napłynięciem pierwszego i drugiego zgłoszenia w każdym okresie czynnym może być inny aniżeli rozkład następnych odstępów między chwilami generacji zgłoszeń).

Rozpatrywana prawidłowość nie prowadzi do zmiany warunku stabilności i analityczne wyznaczenie charakterystyk systemu możliwe jest praktycznie we wszystkich przypadkach systemów jednokanałowych, w których przeanalizowany został analogiczny system konwencjonalny (łatwo sprawdzić, iż przypadek taki jest również specjalnym, uproszczonym wariantem systemu z czasem obsługi zależnym od czasu oczekiwania). Model taki jest często wykorzystywany jako pomocniczy, wprowadzony w sposób sztuczny, przy analizie systemów konwencjonalnych o złożonych regulaminach obsługi, np. regulaminach priorytetowych [JAIS 68], istotnie upraszczając ich rozpatrywanie.

Jednokanałowy system obsługi, w którym rozkłady czasów obsługi wszystkich zgłoszeń zależą od kolejności ich obsługi w ramach okresu czynnego omawiano w [MINE 76], motywując zainteresowanie takim wariantem obsługi np. sposobem zachowania się pojazdów przejeżdżających przez skrzyżowanie w czasie, gdy zapalone jest zielone światło. Wykazano, że o stabilności systemu decydują parametry i-tego zgłoszenia w okresie czynnym, gdy $i \to \infty$.

W dotychczasowych rozważaniach zakładano, że implementacja regulaminu obsługi nie wiąże się z żadnymi stratami czasu pracy stanowiska obsługi. W praktyce założenie takie jest często niespełnione. I tak np. załóżmy, [COFF 78], że obsługa zgłoszenia polega na odczycie zbioru danych z pamięci dyskowej o ruchomych głowicach. Na czas obsługi składa się więc łącznie czas naprowadzenia głowicy na wymaganą ścieżkę (zależny od jej poprzedniego położenia), czas rotacji i czas właściwej transmisji.

Praca własna związana z naprowadzeniem głowicy zależy więc od sekwencji, w jakiej zgłoszenia są obsługiwane.

Również w systemach operacyjnych maszyn cyfrowych zmiana realizowanego programu łączy się z wykonaniem pewnej pracy własnej. Przykładowo w [DAYP 73] podano, że maksymalny czas pracy własnej, związanej z przyjęciem przerwania priorytetowego i realizacją funkcji szeregowania zadań oraz wymuszeniem rozpoczęcia realizacji innego programu, w pewnym systemie opartym na nowoczesnym minikomputerze jest rzędu 300 cykli dostępu do pamięci operacyjnej, a więc nie może być pominięty, zwłaszcza w stosunku do czasu trwania krótkich programów użytkowych (np. obsługi czujników pomiarowych).

Przy bardziej złożonych regułach szeregowania programów czas pracy własnej może zależeć w sposób istotny od liczby zadań ubiegających się o czas procesora [BHAT 79B].

W dyskretnych procesach produkcyjnych często zachodzi potrzeba uwzględnienia pracy własnej (wymiana elementów roboczych, konserwacja bieżąca, zmiana nastaw) przy zmianie rodzaju produkcji lub wykonaniu pewnej partii wyrobów. Podobnie jak w zakresie obsługi właściwej, praca własna może być nieprzerywalna bądź rzadziej przerywalna, a w tym drugim przypadku z zaliczeniem pracy wykonanej do przerwania, bądź jej powtórzeniem od początku.

Literatura w zakresie systemów obsługi uwzględniających różne typy zgłoszeń i pracę własną, związaną ze zmianą typu obsługiwanego zgłoszenia, jest bogata, szczególnie w zakresie regulaminów priorytetowych o stałych priorytetach (np. [SKIN 67], [MEVE 68], [SCHR 69], [EISE 71], [MISZ 74], [MISZ 78], [WOLK 80], [NEUT 77] oraz [WOLI 76A], [WOLI 78], w których, jako jedynych, rozpatrzono zbiór źródeł jednowymiarowych), regulaminów cyklicznych, zwanych też regulaminami o priorytetach przemiennych (np. [SYKE 70], [LIBU 72], [EISE 74], [KOMH 74], [SWAR 80]) oraz regulaminów z podziałem czasu (np. [KLEI 70], [WEGR 74], [BHAT 79B]).

Wielokrotnie rozpatrywane też (np. [GAVE 63], [WOLI 76A], [KREJ 80]) pracę własną poprzedzającą pierwszą obsługę w każdym okresie czynnym, a wynikającą z konieczności przeprowadzenia "rozruchu" stanowiska obsługi, przy czym czas trwania tej pracy własnej może być [KREJ 80] zależny od czasu trwania ostatniego okresu bezczynnego. Przypadek ten jest zbliżony pod względem metod analizy oraz wyników do wzmiankowanego wyżej modelu z różnym od pozostałych rozkładem czasu obsługi pierwszego zgłoszenia w każdym okresie czynnym [WELC 64].

Należy podkreślić, iż praktycznie wszystkie wyniki uzyskane w tym zakresie dotyczą systemów jednokanałowych.

Nie dyskutując szczegółowo różnych, przyjmowanych w poszczególnych pracach założeń, zwrócimy uwagę na zasadnicze skutki uwzględnienia pracy własnej.

W systemach z regulaminami priorytetowymi, przerywalnymi uwzględnienie pracy własnej związanej z przerwaniem prowadzi z zasady do zmiany (a ściślej - wzrostu) jednostkowego czasu obsługi przypadającego na pojedyncze zgłoszenie, ze wzrostem intensywności strumienia zgłoszeń, jako że konieczne jest wówczas wykonanie większej liczby przerwań.

Dla ilustracji przytoczono za [WOLI 76A] wyniki numeryczne obrazujące zmianę stopnia wykorzystania stanowiska obsługi, jednostkowego czasu obsługi oraz średnich czasów oczekiwania w systemie priorytetowym, przy założeniu strumienia Poissona zgłoszeń napływających, czterech klas priorytetu, wykładniczego rozkładu czasu obsługi, stałego efektywnego stopnia wykorzystania stanowiska obsługi oraz stałej (nieprzerwalnej) pracy własnej, związanej z przerwaniem obsługi, a także ze wznowieniem obsługi przerwanej (tabela I.4).

Często spotykane założenie regularnego rozkładu czasu pracy własnej jest oczywiście także pewnym uproszczeniem. Analizę wpływu zmienności czasu pracy własnej (m.in. w funkcji liczby zgłoszeń ubiegających się o obsługę) w przypadku systemu z podziałem czasu zawiera [BHAT 79B].

W związku ze specyfiką różnych regulaminów obsługi łączny czas poświęcony na pracę własną zmienia się z zasady przy zmianie regulaminu obsługi. W wyniku tego zmienia się też opłacalność stosowania różnych regulaminów obsługi w porównaniu z systemami konwencjonalnymi [SEVC 72], oraz zmieniają się warunki stabilności. Przykładowo w [WOLI 76A] wskazano, jak dro-

Tabela I.4

Wybrane charakterystyki systemu z regulaminem priorytetowym bezwzględnym i z uwzględnieniem czasu pracy własnej. Dane wg [WOLI 76A]

	$R_0\mu_2$ [%]	0	2.5%	5%	10%
$\lambda_1 = 0{,}234$	$\eta(\lambda)$	0.600	0.620	0.640	0.684
	$\eta^*(\lambda)$	2.560	2.650	2.740	2,920
	$E(w_1)$	0.141	0.196	0.254	0.383
$\Xi(\lambda_1) = 0{,}6$	$E(w_2)$	1.210	1.319	1.438	1.710
	$E(w_3)$	9.662	10.475	11.385	13.558
	$E(w_4)$	85.991	96.047	109.153	144.449
$\lambda_2 = 0{,}312$	$\eta(\lambda)$	0.800	0.835	0.873	0.955
	$\eta^*(\lambda)$	2.560	2.680	2.800	3.060
	$E(w_1)$	0.200	0.278	0.363	0.558
$\Xi(\lambda_2) = 0.8$	$E(w_2)$	2.000	2.225	2.480	3.098
	$E(w_3)$	20.667	23.750	27.539	38.271
	$E(w_4)$	312.000	427.881	629.737	2331.910

Przyjęto strumień Poissona zgłoszeń o parametrze λ_1 lub λ_2, zapewniającym efektywny stopień wykorzystania stanowiska obsługi $\Xi(\lambda_1) = 0.6$, $\Xi(\lambda_2) = 0.8$. Zgłoszenia należą do i-tej klasy priorytetowej z prawdopodobieństwem p_i takim, że

$$\bigvee_{i=1,2,3,4} \lambda p_i \frac{1}{\mu_i} = \text{const} = \frac{1}{4}\Xi(\lambda),$$

gdzie: μ_i - parametry rozkładu wykładniczego, określającego czas obsługi $\mu_1 = 5\,\mu_2 = 25\,\mu_3 = 125\,\mu_4$, $\mu_1 = 1.25$, R_0 - czas pracy własnej przy przerwaniu i wznowieniu obsługi, wartości R_0 określono przez podanie stosunku czasu pracy własnej do średniego czasu obsługi zgłoszeń drugiej klasy priorytetowej, wyrażonego w procentach.

gą zmniejszenia liczby przerwań priorytetowych można znacznie obniżyć stopień wykorzystania stanowiska obsługi (zachowując stały stopień efektywnego wykorzystania stanowiska obsługi), przy stosunkowo niewielkich modyfikacjach rozkładów czasów oczekiwania. Równocześnie uzyskuje się też zwiększenie wydajności granicznej systemu.

Wynikające z uwzględnienia pracy własnej zjawisko obniżania się wydajności systemu przy wzroście intensywności strumienia zgłoszeń budzi obecnie coraz większe zainteresowanie [SHOR 80], przy czym jego eliminacja

wymaga z reguły wprowadzenia dynamicznych, zależnych od stanu regulaminów obsługi.

Rozpatrzmy często omawiany przypadek, w którym wyróżnia się K-typów zgłoszeń i zakłada się, iż czas obsługi zgłoszeń i-tego typu i=1,2,...,K określa zmienna losowa b_i, natomiast w przypadku, gdy po zgłoszeniu typu i obsługiwane ma być zgłoszenie typu j, niezbędna w tym celu praca własna zajmuje kanał obsługi na czas określony zmienną losową y_{ij} (przypadek taki odpowiada np. wzmiankowanemu wyżej trybowi korzystania z pamięci dyskowej i rozpatrywany był np. w [COFF 78], gdzie rozszerzono przedstawione w [NEUT 76] wyniki analizy regulaminu naturalnego, a także omówiono sposób określenia parametrów modelu dla różnych rodzajów pamięci obrotowych). Drogą doboru regulaminu obsługi można znacznie zwiększyć wydajność graniczną takiego systemu, stąd stosowane są różne regulaminy, np. wybór zgłoszenia, którego obsługa poprzedzona byłaby najmniejszą pracą własną (np. [KONS 75]).

W pracy [MARD 80], niezależnie od omówienia dokładnej i przybliżonej metody analizy takich systemów, stwierdzono, że również w tym przypadku oczekiwanie na napłynięcie nowego zgłoszenia należącego do typu ostatnio obsługiwanego, mimo niepustej kolejki zawierającej zgłoszenia innych typów, może prowadzić do poprawy sprawności systemu, mierzonej liniową funkcją czasu oczekiwania zgłoszeń różnych typów.

Podobnie jak w przypadku systemów z czasem obsługi zależnym od czasu oczekiwania uwzględnienie pracy własnej sprawia, iż trudne jest wyznaczenie warunków stabilności systemu, które zależą w istotny sposób od stosowanego regulaminu obsługi.

Wymaga to na ogół kompletnego przeanalizowania rozpatrywanego regulaminu obsługi, co w wielu przypadkach np. dynamicznej zmiany priorytetu, różnych od strumienia Poissona charakterystyk źródeł zgłoszeń w regulaminach priorytetowych itd. udało się jak dotąd w bardzo ograniczonym zakresie, nawet przy pominięciu istnienia pracy własnej.

Często podejmuje się próby uwzględnienia wpływu pracy własnej w sposób przybliżony, np. zakładając zależność mocy stanowiska obsługi od liczby zgłoszeń w systemie. Zwiększenie zaabsorbowania stanowiska obsługi pracą własną przy dużej liczbie zgłoszeń w systemie modeluje się więc jako zmniejszenie efektywnej mocy stanowiska obsługi, co prowadzi do rozpatrywania łatwiejszej do analizy, omawianej wcześniej klasy systemów masowej obsługi [PUJO 80A], [SHOR 80].

Wszystkie omawiane dotąd niekonwencjonalne systemy masowej obsługi cechowały się tym, że zakończenie obsługi ostatniego zgłoszenia w danym okresie czynnym powodowało przejście kanału obsługi do stanu bezczynności, natomiast bezczynny kanał obsługi mógł rozpocząć obsługę zgłoszenia (lub przynajmniej poprzedzający ją rozruch) w dowolnym momencie, a także tym, że obsługa zgłoszenia mogła być przerwana jedynie w celu podjęcia obsługi innego zgłoszenia.

W praktyce częste są przypadki występowania okresów, w których stanowisko obsługi jest niedostępne.

Wyróżnimy tu trzy zasadnicze warianty:

i) Sytuację, gdy okresy niedostępności występują wyłącznie po zakończeniu obsługi zgłoszenia i wiążą się na przykład z koniecznością wykonania konserwacji stanowiska obsługi, jego testowania czy też uzupełnienia materiałów eksploatacyjnych. Możemy też mieć do czynienia z sytuacją, gdy przerwanie bezczynności stanowiska obsługi możliwe jest tylko w pewnych, określonych chwilach czasu, przy czym odstępy między tymi chwilami mogą być - ogólnie rzecz biorąc - określone jako zmienne losowe. Takie modele zostały dość dokładnie przebadane [SKIN 67], [GELE 80A].

ii) Sytuację, gdy chwile rozpoczynania obsługi kolejnych zgłoszeń są ściśle określone i następują po sobie w stałych odstępach (tzw. praca synchronizowana).
Z takimi przypadkami mamy do czynienia w przypadku transmisji informacji w sieciach komputerowych [KOBA 77] lub modelowania taśmy produkcyjnej poruszającej się z określonym taktem [BUXE 73].
Rozpatrywanie takich sytuacji prowadzi do specjalnego działu tzw. dyskretnej teorii masowej obsługi (patrz np. [MEIS 58], [KOBA 77]), którego omawianie wykracza poza ramy niniejszej pracy.

iii) Sytuację, gdy okres niedostępności może się rozpocząć w dowolnym momencie okresu czynnego (wymagając przerwania obsługi aktualnie realizowanej), ewentualnie również w dowolnym momencie okresu bezczynnego. Taki przypadek może być np. wynikiem awarii kanału obsługi. Modele uwzględniające strumień takich nieprzewidzianych przerw w działaniu stanowiska obsługi są z reguły pod względem formalnym identyczne z systemami priorytetowymi, w których na najwyższym stopniu priorytetu wprowadza się dodatkowy strumień zgłoszeń reprezentujących takie przerwy. Należy jednak podkreślić, iż przy aktualnym rozwoju teorii priorytetowych regulaminów obsługi przyjęcie strumieni zgłoszeń różnych od strumieni Poissona natrafia na spore trudności.

Jeszcze bardziej złożone w analizie są przypadki, gdy przerwa w pracy stanowiska obsługi wymaga, po jej zakończeniu, ponownego wykonania nie tylko obsługi przerwanej, ale być może także pewnej liczby obsług wcześniejszych. Klasycznym przykładem są tu bazy danych w systemach czasu rzeczywistego, modyfikowane po przetworzeniu kolejnych transakcji (obsłudze kolejnego zgłoszenia). W przypadku awarii sprzętu cała zawartość bazy mogłaby ulec zniszczeniu, co jest niedopuszczalne ze względów funkcjonalnych. Dlatego też przeprowadza się w pewnych odstępach czasu kopiowanie zawartości bazy danych oraz rejestrowanie wszystkich transakcji przetworzonych od ostatniego kopiowania. W przypadku wykrycia awarii następuje rekonfiguracja systemu, odtworzenie stanu bazy danych, polegające na odczytaniu jej stanu w chwili ostatniego kopiowania i ponowne przetworzenie

wszystkich transakcji od ostatniego kopiowania do momentu awarii. Oczywiście zarówno kopiowanie, jak i odtwarzanie stanowią okresy, gdy stanowisko obsługi jest wyłączone z normalnego działania, przy czym czas odtwarzania zależy też od liczby transakcji przetworzonych między momentem kopiowania i momentem wystąpienia awarii. Analizę takiego przypadku przeprowadzono w [GELE 79], inne modele związane z poruszonym tu problemem omawiano też np. w [CHAN 75C], [BACE 79], [DUDA 81].

W oparciu o omówione przypadki niekonwencjonalnych systemów obsługi możemy bardziej precyzyjnie określić wprowadzone na wstępie niniejszego rozdziału pojęcie stanu przygotowywania stanowiska obsługi. Jak widzimy, przygotowywanie to może polegać na oczekiwaniu na napłynięcie zgłoszeń wybranego typu, pracy własnej związanej z implementacją regulaminu obsługi (czy też zmiana typu zgłoszenia) bądź też okresowej niedostępności stanowiska obsługi.

d) Podsumowanie

Własności omawianych systemów obsługi uzasadniają ich podział na cztery grupy:

i) Systemy, w których dopuszcza się zależności losowe między elementami ciągów $\{a^i\}$ oraz $\{b^i\}$. Wszystkie przypadki należące do tej grupy cechuje podobny jak w systemach konwencjonalnych warunek stabilności (różnice występują tylko w granicznych przypadkach, gdy $\frac{\lambda E(b)}{n} = 1$), różna mogą być natomiast wartości charakterystyk,

ii) Systemy, w których na skutek ograniczeń (w trybie generacji zgłoszeń, długości kolejki, czy też czasu oczekiwania) niemożliwa jest niestabilność. Przypadek ten różni się zasadniczo pod względem własności od systemów konwencjonalnych.

iii) Systemy, w których możliwa jest niestabilność, przy czym warunki stabilności zależą ogólnie rzecz biorąc od regulaminu obsługi.

iv) Systemy, które nie osiągają nigdy stacjonarnego stanu granicznego.

Przypadek iv) wymaga stosowania metod analizy właściwych stanom nieustalonym. Przypadki i) oraz ii) pozwalają na rozpatrywanie systemu, którego stabilność (bądź - być może - substabilność) jest udowodniona, tak więc możliwe jest, w przypadku braku rozwiązań analitycznych, odwołanie się do modelowania cyfrowego (jakkolwiek np. w [NEUT 78A] zwrócono uwagę na praktyczną nierealność stosowania tej metody w odniesieniu do systemów z parametrami zmiennymi w czasie i czasowymi przeciążeniami, ze względu na bardzo duże wartości wariancji charakterystyk, przekraczające wielokrotnie wartość oczekiwaną).

Najwięcej trudności sprawiają z reguły systemy należące do trzeciej grupy, ze względu na brak ogólnych metod pozwalających na określenie warunków ich stabilności przy różnych regulaminach obsługi, co w znacznym stopniu ogranicza możliwość wykorzystania nawet metod symulacyjnych. Przedmiotem rozważań przedstawionych w rozdziałach II i III będzie typowy - należący do tej klasy - system z czasem obsługi zależnym od czasu oczekiwania, przy czym niektóre metody przedstawione w rozdziale III mogą znaleźć również szersze zastosowanie.

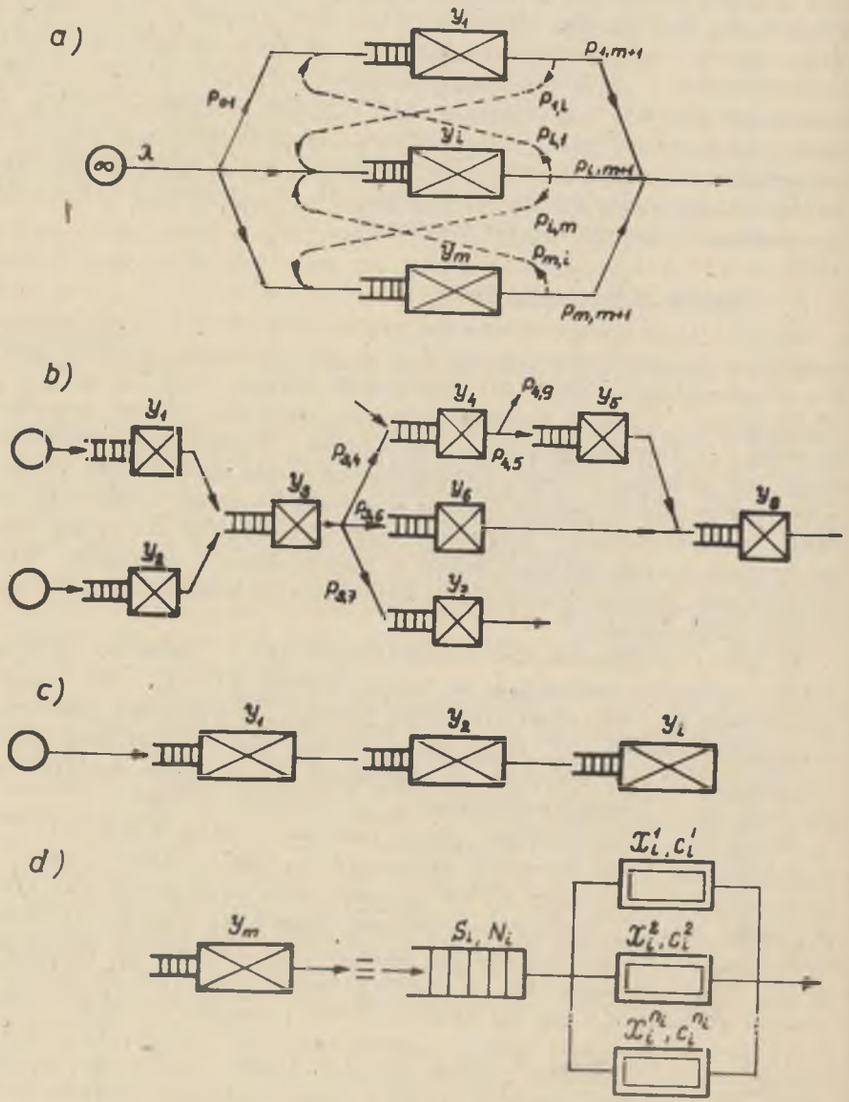
W związku z trudnościami występującymi przy analizie niekonwencjonalnych systemów masowej obsługi większość dotąd przeprowadzonych badań ogranicza się do przypadku jednokanałowego. Nawet jednak i w tym zakresie, jak to wynika z przedstawionego omówienia, uzyskane wyniki są jedynie fragmentaryczne i pozwalają na ogół określić wyłącznie ogólne własności różnych rozpatrywanych przypadków. Stąd istnieje wyraźna potrzeba dalszych prac w tym zakresie, a szczególnie prac dotyczących metodyki badania pewnych klas systemów niekonwencjonalnych, a nie tylko wybranych, specyficznych przypadków charakteryzujących się na przykład specjalną postacią zależności losowej czy też rozkładów.

I.2. Systemy wielostanowiskowe

Omawiane dotąd jednostanowiskowe systemy obsługi nie są wystarczającym narzędziem dla modelowania bardziej złożonych procesów, w których zgłoszenie wymaga kolejno obsługi na stanowiskach różnego typu (i to być może wielokrotnie). I tak na przykład modelując działanie systemu komputerowego można - w dużym uproszczeniu - przyjmować, iż jedynym stanowiskiem obsługi jest procesor liczący, zaniedbując przy tym ograniczenia wynikające z dostępu do innych zasobów (pamięć operacyjna, pamięci zewnętrzne, urządzenia wejścia-wyjścia). Znacznie precyzyjniej odtwarza się jednak własności takiego systemu używając modelu w postaci wielu stanowisk obsługi różnego typu, między którymi krążą zgłoszenia reprezentujące poszczególne procesy obliczeniowe.

Na rys. I.3a przedstawiono schematycznie sieć M stanowisk obsługi, w której zgłoszenie napływające ze źródła zgłoszeń kieruje się z prawdopodobieństwem $p_{0,i}$ do i-tego stanowiska obsługi, a następnie odpowiednio z prawdopodobieństwem p_{ij}, j=1,2,..., M kieruje się do następnego stanowiska obsługi (w tym, być może, do stanowiska i-tego, które właśnie opuściło) bądź też z prawdopodobieństwem $p_{i,M+1}$ opuszcza system.

Oznaczenie źródła zgłoszeń jako stanowiska obsługi z numerem "0" nie jest wyłącznie pewną konwencją notacyjną, ale kryje w sobie głębszy sens. Łatwo sprawdzić, że każde źródło zgłoszeń można potraktować jako strumień wyjściowy pewnego, odpowiednio dobranego stanowiska obsługi. W szczególności zgłoszenia, których odstępy określone są niezależnymi zmiennymi losowymi o dystrybuancie A(x), stanowią strumień wyjściowy jednokanałowego systemu obsługi, z nigdy nie pustym buforem i czasami obsługi w postaci niezależnych zmiennych losowych o rozkładzie A(x). Taką sieć stanowisk obsługi nazywać będziemy siecią otwartą. Gdyby, zachowując identyczną strukturę przepływów zgłoszeń wewnątrz sieci, przyjąć, iż źródło ma zerową intensywność generacji zgłoszeń i $\bigvee_{i=1,2,\ldots,M} [p_{i,M+1} = 0]$, wówczas mówilibyśmy o zamkniętej sieci stanowisk obsługi, w której krąży pewna liczba N zgłoszeń.



Rys. I.3. Schematy sieci stanowisk obsługi
a) sieć dowolna, b) sieć acykliczna, c) wieloetapowy system obsługi, d) postać pojedynczego węzła sieci

Na rys. I.3b oraz I.3c przedstawiono pewne szczególne przypadki sieci stanowisk obsługi, wyróżniające się specyficznymi regułami przepływu zgłoszeń. W sieciach acyklicznych (rys. I.3b) zgłoszenie może w czasie pobytu w sieci tylko jeden raz być obsługiwane na którymkolwiek ze stanowisk obsługi.

W wieloetapowych systemach obsługi (rys. I.3c) każde zgłoszenie musi zostać obsłużone (tylko jeden raz) na każdym ze stanowisk obsługi i to w ściśle określonej kolejności. Jeśli wszystkie stanowiska obsługi są jednokanałowe, to system nazywamy pojedynczym. Systemy takie, będące najprostszym przypadkiem sieci stanowisk obsługi, mają duże znaczenie praktyczne, w szczególności dla modelowania linii produkcyjnych [BUXE 73] oraz transmisji informacji w sieciach komputerowych [KLEI 76].

Jak to zaznaczono na rys. I.3d, każde rozpatrywane w ramach sieci stanowisko obsługi $\mathcal{X}_i$ ma strukturę identyczną jak stanowisko obsługi rozpatrywane w poprzednim punkcie (porównaj rys. I.1) i jest poprzedzone buforem S_i o pewnej, skończonej lub nieskończonej pojemności N_i. Stanowisko obsługi X_i wraz z buforem S_i będziemy nazywali węzłem sieci i oznaczali Y_i. W dalszych rozważaniach ograniczymy się do przypadku, gdy wszystkie stanowiska obsługi są jednorodne. Tak więc pozostają ważne wszystkie wprowadzone w punkcie I.1 oznaczenia, przy czym dla prawidłowego rozróżnienia będziemy oznaczenia parametrów węzła sieci Y_i uzupełniać indeksem "i" (np. $B_i(x)$ oznaczać będzie dystrybuantę czasu obsługi w dowolnym kanale stanowiska obsługi $\mathcal{X}_i$, c_i - oznaczać będzie łączną moc wszystkich kanałów obsługi stanowiska X_i, natomiast w_i^j oznaczać będzie czas oczekiwania zgłoszenia ζ^j na rozpoczęcie obsługi w stanowisku obsługi $\mathcal{X}_i$).

Dla oceny efektywności działania sieci stanowisk obsługi używać będziemy, w odniesieniu do każdego z nich, charakterystyk wprowadzonych w punkcie I.1.

W szczególności będziemy mówili, że sieć stanowisk obsługi jest niestabilna, jeśli relacja (I.13) jest niespełniona w odniesieniu do czasu oczekiwania na przynajmniej jednym stanowisku obsługi.

Sieć stanowisk obsługi nazywać będziemy siecią konwencjonalną, gdy każde ze stanowisk obsługi spełnia warunki wymienione w punkcie I.1.1 a więc:

i) posiada stałą moc,
ii) czasy obsługi kolejnych zgłoszeń są ciągiem zmiennych losowych o identycznych rozkładach i skończonej wartości oczekiwanej,
iii) zgłoszenia obsługiwane są zgodnie z regulaminem zachowującym pracę,
iv) bufor poprzedzający stanowisko obsługi ma nieograniczoną pojemność oraz gdy zgłoszenia napływają do systemu (w przypadku sieci otwartych) w chwilach czasu, których odstępy są niezależnymi zmiennymi losowymi o jednakowych rozkładach.

Biorąc pod uwagę duży dorobek w zakresie analizy systemów jednostanowiskowych, w sposób naturalny nasuwa się pytanie czy i w jakim zakresie możliwe jest oddzielne rozpatrywanie poszczególnych węzłów sieci, z u-

względnieniem ich powiązań z innymi węzłami sieci wyłącznie poprzez odpowiednie zdefiniowanie strumienia zgłoszeń napływających do tych wydzielonych węzłów.

Na przykład wprost z rys. I.3c widać, że o ile tylko wszystkie bufory w systemie wieloetapowym mają nieograniczoną pojemność, to na pracę dowolnego węzła Y_i nie mają wpływu procesy zachodzące w węzłach $Y_{i+1}, Y_{i+2}, \ldots, Y_M$, a procesy zachodzące w węzłach $Y_1, Y_2, \ldots, Y_{i-1}$ oddziaływują wyłącznie poprzez postać strumienia zgłoszeń napływających do węzła Y_i. Aby orzec czy postać tego strumienia zgłoszeń umożliwia analizę pracy stanowiska, konieczne jest zbadanie własności strumieni zgłoszeń opuszczających jednostanowiskowe systemy obsługi.

Liczne prace z tego zakresu zostały obszernie omówione w [BURK 72], [DALE 76], [MELA 79], [LABE 79A]. Niektóre najważniejsze wyniki charakteryzujące własności strumieni wyjściowych przedstawimy poniżej jako punkt wyjścia do omówienia metody analizy sieci stanowisk obsługi.

Jakkolwiek pierwsze prace w zakresie analizy sieci stanowisk obsługi opublikowane zostały dopiero w latach 1954-1956 to jednak tematyka ta wzbudziła ogromne zainteresowanie szczególnie w latach siedemdziesiątych i posiada już obecnie bardzo bogatą literaturę.

W zakresie konwencjonalnych sieci stanowisk obsługi (a także pewnych ich modyfikacji) uzyskano cenne wyniki umożliwiające ich analityczne rozpatrywanie zarówno przy użyciu metod ścisłych, jak i przybliżonych. Omówienie dorobku tych prac można znaleźć w opracowaniach przeglądowych [CHAN 78], [DISN 75], [LEMO 77], [FUJO 80B], i fragmentach nowych monografii o szerszym profilu [KLEI 76], [KOBA 79], [GELE 80A]. Ukazała się też pierwsza monografia poświęcona wyłącznie problematyce sieci stanowisk obsługi [GELE 80B].

W pracach tych, a także [KLEI 76], [KOBA 77], [MON 78] wskazywano wielokrotnie, iż realia modelowanych procesów wymagają naruszenia założeń definiujących konwencjonalne sieci stanowisk obsługi.

Poniżej podana zostanie jedynie krótka charakterystyka rodziny konwencjonalnych sieci stanowisk obsługi, dla których aktualnie dostępne są dokładne rozwiązania analityczne (ze wskazaniem efektywnych metod analizy oraz perspektyw uogólnienia tych wyników), a następnie przedyskutowane zostaną, znacznie słabiej rozpoznane, własności i metody analizy niekonwencjonalnych sieci stanowisk obsługi.

Potraktujemy stanowisko obsługi jako układ transformujący strumień zgłoszeń napływających ze źródła na, ogólnie rzecz biorąc, różny od niego strumień wyjściowy. Strumienie zgłoszeń, w których odstępy między kolejnymi zgłoszeniami stanowią niezależne zmienne losowe o jednakowych rozkładach, nazywane są strumieniami odnowień.

Ponieważ (z nielicznymi, omówionymi w punkcie I.1.3 wyjątkami) możliwości analitycznego badania systemów obsługi ograniczają się do takich strumieni zgłoszeń, będziemy w pierwszym rzędzie zainteresowani określeniem klasy stanowisk obsługi transformujących (przy pracy w stacjonarnym stanie granicznym) wejściowy strumień odnowień w strumień odnowień, o być może różnym rozkładzie odstępów między zgłoszeniami, czyli klasą stanowisk obsługi posiadających tzw. własności $R \rightarrow R$. Podzbiór tej klasy stanowią stanowiska obsługi posiadające tzw. własność $M \rightarrow M$, polegającą na transformowaniu strumienia Poissona w strumień Poissona. W stacjonarnym stanie granicznym intensywność strumienia wyjściowego i wejściowego muszą być oczywiście równe.

Rozpatrzmy na początek systemy obsługi z buforem o nieograniczonej pojemności.

W [BURK 56] udowodniono, że strumień wyjściowy systemu M/M/n jest strumieniem Poissona oraz że stan tego systemu w dowolnej chwili t_0 jest niezależny losowo od postaci strumienia wyjściowego w czasie $t < t_0$. Własności te noszą nazwę twierdzenia o strumieniu wyjściowym. Tę samą własność posiadają też ([HADI 72]) systemy M/M/1 z mocą stanowiska obsługi zależną od liczby zgłoszeń w systemie. W [FINC 59B] wykazano, iż strumień zgłoszeń opuszczających system M/GI/1 jest strumieniem odnowień wyłącznie w przypadku, gdy G = M, w przeciwnym razie strumień ten jest skorelowany. Podobnie dla systemów GI/M/1 wykazano ([DALE 68]), iż warunkiem koniecznym i wystarczającym, aby strumień wyjściowy był strumieniem odnowień jest, aby GI= M.

W systemie GI/M/1 (GI $\neq$ M), przy dowolnym rozkładzie odstępów między nadchodzącymi zgłoszeniami, posiadającym dystrybuantę A(x) na to, aby odstępy między zgłoszeniami opuszczającymi system były nieskorelowane potrzeba i wystarcza spełnienie równości:

$$\lambda E(b) = \delta ,$$

gdzie δ jest jedynym, leżącym w przedziale $[0,1)$ rozwiązaniem równania

$$\delta = \bar{a}\left[\mu(1-\delta)\right], \qquad \mu = \left[E(b)\right]^{-1}.$$

Mimo iż dowodu takiego stwierdzenia dotąd nie przeprowadzono, powszechnie uważa się, iż jedynym systemem typu GI/GI/1 posiadającym własność $R \rightarrow R$ jest system M/M/1 (cechujący się silniejszą własnością $M \rightarrow M$).

Strumieniem Poissona jest również strumień wyjściowy systemu M/GI/∞ [MIRA 63], [NEWE 66]. W tym przypadku strumień wyjściowy pozostaje strumieniem Poissona, ale o parametrze zmiennym w czasie, jeśli rozpatrywany jest stan nieustalony bądź strumień wejściowy posiada intensywność zmienną w czasie.

Wymienione wyżej rezultaty dotyczyły naturalnego regulaminu obsługi względnie dowolnego innego regulaminu obsługi, nie uwzględniającego rzeczywistego zapotrzebowania indywidualnych zgłoszeń na obsługę.

Muntz [MUNT 72] udowodnił, iż własnością M— M charakteryzują się również systemy $M/K_1/1$ z dwoma specjalnymi regulaminami obsługi, a mianowicie:

i) PS - regulaminem z podziałem procesora (będącym granicznym przypadkiem regulaminu cyklicznego z podziałem czasu, przy wielkości kwantu dążącej do zera [KLEI 76]).

ii) PLIFO - regulaminem z priorytetem bezwzględnym, przerywającym, przyznawanym każdorazowo zgłoszeniu, które ostatnie napłynęło do systemu.

W oparciu o wprowadzone w [CHAN 72B] pojęcie równowagi lokalnej (o równowadze lokalnej mówimy wówczas, gdy intensywność strumieni zdarzeń wyprowadzających stanowisko obsługi z dowolnego stanu równa jest intensywności strumieni zdarzeń przeprowadzających je do tego stanu; definicja stanu może być przy tym różna w zależności od regulaminu obsługi i rozkładu czasów obsługi - patrz np. [BASK 75]) oraz podany w [MUNT 72] dowód, iż zachowanie przez stanowisko obsługi równowagi lokalnej jest warunkiem wystarczającym, aby rozpatrywane stanowisko obsługi posiadało własność M— M można było w ostatnich latach rozszerzyć klasę regulaminów obsługi, które mogą posiadać tę własność przy różnych od wykładniczego rozkładach czasów obsługi.

W [CHAN 77] udowodniono, że własność M— M posiadają wszystkie regulaminy obsługi charakteryzujące się tzw. "równowagą stanowiska", tj. takie, w których stopień, w jakim zgłoszenie zajmujące pewną pozycję w kolejce angażuje moc stanowiska obsługi, jest proporcjonalny do prawdopodobieństwa napłynięcia zgłoszenia na tę pozycję (uwaga: każdy regulamin posiadający tę własność cechuje się natychmiastowym rozpoczęciem obsługi nowo nadchodzącego zgłoszenia) i to dla dowolnych rozkładów czasów obsługi o różniczkowalnych dystrybuantach.

W [NOET 79] wprowadzono natomiast rodzinę regulaminów obsługi LBPS (których nie będziemy tu szczegółowo opisywać) dowodząc, że posiadają one własność M— M dla dowolnych rozkładów czasów obsługi oraz iż w pewnej ogólnej klasie tzw. symetrycznych regulaminów obsługi warunkiem koniecznym posiadania własności M— M dla rozkładów czasów obsługi różnych od wykładniczego jest przynależność do tej rodziny.

Prace te mają istotne znaczenie teoretyczne, jak to się uwidoczni wyraźnie w dalszych rozważaniach, nie rozszerzają jednak zbioru regulaminów obsługi posiadających własność M— M o regulaminy mające istotne znaczenie praktyczne.

We wszystkich omawianych wyżej przypadkach zakładano nieograniczoną pojemność bufora związanego z rozpatrywanym stanowiskiem obsługi. Omówimy obecnie niektóre interesujące warianty stanowisk z ograniczoną pojemnością bufora.

Uogólniając rezultaty [FINC 59B], [DISN 73] wykazano, że jedynymi systemami typu M/G/1/L, przy $L < \infty$ i niezerowych czasach obsługi, posiadającymi własność IR— IR są systemy M/G/1/1 (a więc o zerowej pojemności bufora) oraz systemy M/D/1/2 (a więc o pojemności bufora równej jedności). Należy podkreślić, iż w obu przypadkach odstępy między zgłoszeniami w strumieniu wyjściowym mają rozkład różny od wykładniczego.

Strumieniem Poissona jest natomiast <u>łączny</u> strumień zgłoszeń obsłużonych oraz traconych w systemach typu M/GI/n/n; własność ta pozostaje w mocy również dla niektórych systemów M/M/n o nieograniczonych buforach ze stratami zgłoszeń wynikającymi z ograniczonego czasu oczekiwania w kolejce bądź niezerowego prawdopodobieństwa niedołączenia zgłoszenia do kolejki (omówienie wyników takich prac zawiera [DALE 76]).

Dla uzupełnienia należy dodać, że jak to zauważono w [LABE 79], własnością M— M mogą cechować się również systemy obsługi, w których czasy kolejnych obsług nie są niezależnymi zmiennymi losowymi o identycznych rozkładach.

Przykładem może być jednokanałowy system obsługi równoważny systemowi przedstawionemu na rys. I.3, który posiadałby własność M— M (z definicji) i różny od wykładniczego rozkład czasu obsługi (a więc na mocy wyżej cytowanych twierdzeń czasy obsługi nie mogą być zmiennymi losowymi niezależnymi o identycznych rozkładach).

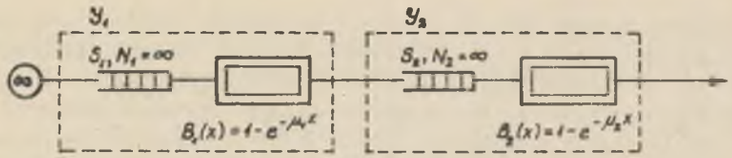
Wielu autorów podejmowało problematykę określenia rozkładów odstępów między zgłoszeniami, a także wartości współczynników korelacji, charakteryzujących strumienie wyjściowe nie będące strumieniami odnowień. Nie relacjonując wyników w zakresie określania zależności prowadzących do wyznaczenia rozkładów odstępów między zgłoszeniami opuszczającymi system, ani też wyznaczenia współczynników korelacji, których omówienie zawierają cytowane prace przeglądowe (odnośnie do systemów jednokanałowych patrz też [COHE 69]; nowsze wyniki zawierają też prace [PACK 75] w zakresie systemów M/D/1, [PACK 77] w zakresie D/M/1, [DISN 76] w zakresie systemów M/GI/1 oraz [PACK 78] w zakresie systemów wielokanałowych), przytoczymy tu dla ilustracji charakterystyczny rezultat. Otóż w odniesieniu do systemów typu M/GI/1 o ustalonej wartości oczekiwanej czasu obsługi maksymalna wartość współczynnika autokorelacji odstępów między zgłoszeniami w strumieniu wyjściowym osiągana jest w przypadku, gdy GI = D i wynosi około 0.18.

I.2.1. Sieci konwencjonalnych stanowisk obsługi

Omówione własności strumieni wyjściowych posłużą nam obecnie do scharakteryzowania analitycznych rozwiązań szerokiej klasy sieci stanowisk obsługi.

Rozpatrzmy na wstępie dwuetapowy system obsługi z jednokanałowymi stanowiskami o stałej mocy, przedstawiony na rys. I.4.

Stan takiego systemu określa w każdym momencie jednoznacznie wartość $P_t(x_1,x_2,t)$ prawdopodobieństwa tego, że w chwili t w węźle Y_1 znajduje się x_1, zgłoszeń, natomiast w węźle Y_2 znajduje się x_2 zgłoszeń.



Rys. I.4. Prosty, dwuetapowy system obsługi ze zgłoszeniami napływającymi strumieniem Poissona i wykładniczymi rozkładami czasów obsługi

W pracy [JACK 54] udowodniono naturalny w świetle własności strumienia wyjściowego stanowiska M/M/1 fakt, że przy $\lambda < \min(\mu_1, \mu_2)$ w stacjonarnym stanie granicznym zachodzi:

$$\lim_{t\to\infty} P_t(x_1,x_2,t) = P(x_1,x_2) = P_1(x_1)P_2(x_2) = (1-\frac{\lambda}{\mu_1})(\frac{\lambda}{\mu_1})^{x_1}(1-\frac{\lambda}{\mu_2})(\frac{\lambda}{\mu_2})^{x_2},$$

$$x_1, x_2 = 0,1,\ldots \qquad (I.2.1)$$

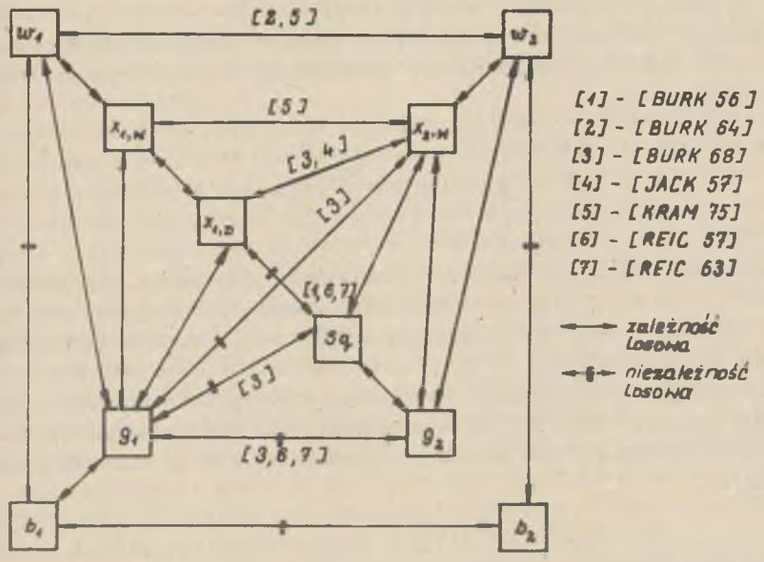
a więc prawdopodobieństwa $P_i(x_i)$ tego, że w węźle Y_i znajduje się x_i zgłoszeń ($i = 1,2$) są niezależne i określane wg zależności obowiązujących dla węzłów Y_1, Y_2 w izolacji, rozpatrywanych jako oddzielne systemy M/M/1.

O ile w taki prosty sposób można wyznaczyć stan rozpatrywanego systemu, to znacznie większe trudności nastręcza próba określenia charakterystyk związanych z przebiegiem obsługi indywidualnego zgłoszenia. Niezależność stanów poszczególnych stanowisk obsługi nie implikuje bowiem bezpośrednio niezależności innych zmiennych opisujących proces obsługi, np. dowodzi się [BURK 64], że czasy oczekiwania zgłoszenia w buforach S_1 i S_2 są zależne losowo.

Na rys. I.5 przytoczono w ślad za [KRAM 75] schemat przedstawiający zależności losowe różnych zmiennych losowych, opisujących proces obsługi zgłoszeń w systemie obsługi przedstawionym na rys. I.4 (podane relacje pozostają w mocy również dla wielokanałowych stanowisk obsługi). Cytowana praca zawiera też interesujące wywody dotyczące możliwości wnioskowania o sposobie obsługi zgłoszenia w węźle Y_2 (z wielokanałowym stanowiskiem obsługi), przy różnym poziomie informacji o historii pobytu zgłoszenia w jednokanałowym węźle Y_1.

Wspomniane zależności losowe tłumaczą podany wyżej fakt, że równoważny wieloetapowemu systemowi obsługi przedstawionemu na rys. I.3 system typu M/G/1 (w którym na czas obsługi składałby się czas obsługi na stanowisku $\mathfrak{X}_1$, czas oczekiwania w buforze S_2 oraz czas obsługi na stanowisku $\mathfrak{X}_2$), cechowałby się czasami obsługi nie będącymi niezależnymi zmiennymi losowymi o jednakowych rozkładach.

Przechodząc do omówienia bardziej złożonych sieci stanowisk obsługi należy podkreślić, iż analityczne określenie rozkładu prawdopodobieństw poszczególnych stanów sieci jest możliwe w zasadzie wyłącznie tam, gdzie



Rys. I.5. Schemat zależności losowych różnych charakterystyk systemu obsługi przedstawionego na rys. I.4 (wg [KRAM 75])

b_i, μ_i, s_i - czasy obsługi, oczekiwania i pobytu zgłoszeń w Y_i, $x_{i,w}$ - liczba zgłoszeń, które zastaje zgłoszenie testowe, $x_{1,D}$ - liczba zgłoszeń pozostawiona przez zgłoszenie testowe w chwili opuszczenia węzła Y_1, s_q - sekwencja chwil zakończenia obsługi w węźle Y_1 przez zgłoszenia obsługiwane przed zgłoszeniem testowym

prawdopodobieństwa te mają specyficzną, tak zwaną iloczynową postać typu:

$$\psi(\underline{\alpha}_1,\underline{\alpha}_2,\ldots,\underline{\alpha}_M) = K\phi_1(\underline{\alpha}_1)\phi_2(\underline{\alpha}_2)\ldots\phi_M(\underline{\alpha}_M), \qquad (I.2.2)$$

gdzie: $\underline{\alpha}_i$ - wektory stanu poszczególnych węzłów, jako że przy złożonych regulaminach obsługi oraz różnych od wykładniczego rozkładach czasów obsługi stan węzła nie daje się z reguły scharakteryzować wyłącznie przez podanie liczby zgłoszeń znajdujących się w danym węźle; $\phi_i(\underline{\alpha}_i)$ - funkcje, których dziedziną są wektory stanu, natomiast K jest stałą normalizacyjną.

Korzystając z omówionych wcześniej własności strumieni wyjściowych można wprost stwierdzić, że w systemach wieloetapowych, a także w sieciach acyklicznych, złożonych wyłącznie ze stanowisk charakteryzujących się własnością M — M, w których wszystkie źródła generują Poissonowskie strumienie zgłoszeń, również wszystkie strumienie zgłoszeń wewnątrz sieci są niezależnymi losowo strumieniami Poissona, o ile tylko dla każdego stanowiska obsługi, rozpatrywanego oddzielnie, spełniony jest warunek stabilności.

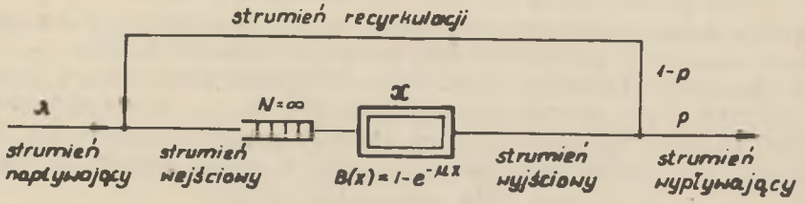
Wynika stąd wprost niezależność stanów poszczególnych węzłów i możliwość uzyskania rozwiązań o postaci iloczynowej.

Bardziej złożona jest natomiast sytuacja w przypadku dowolnych sieci stanowisk obsługi. Posługując się oznaczeniami z rys. I.2 można sformułować układ równań

$$\gamma_i = \lambda p_{oi} + \sum_{j=1}^{M} \gamma_j p_{ji}, \qquad (I.2.3)$$

określający, dla sieci otwartej, intensywności strumieni zgłoszeń w węzłach Y_i, i = 1,2,... M albo, dla sieci zamkniętej, względne częstości pobytu zgłoszeń w węzłach Y_i; nie można natomiast bezpośrednio określić charakteru tych strumieni.

Jackson [JACK 57] wykazał, że również w przypadku dowolnej struktury połączeń w sieci otwartej ze stanowiskami typu M/M/n obowiązuje iloczynowa postać rozwiązań (pod warunkiem zapewnienia stabilności każdego ze stanowisk).



Rys. I.6. Schemat systemu masowej obsługi ze sprzężeniem zwrotnym

Własność ta jest zaskakująca w świetle wyników badań własności strumieni wewnętrznych w takiej sieci. I tak w pracy [LABE 79A] badając prosty system przedstawiony na rys. I.6 wykazano, że aczkolwiek strumienie napływający i wypływający są strumieniami Poissona, to jednak strumień wejściowy oraz strumień wyjściowy charakteryzują się hiperwykładniczymi rozkładami odstępów między zgłoszeniami. Co więcej, wykazano, że zarówno strumień wyjściowy [POLE 77], jak i strumień recyrkulacji [LABE 79B] nie są strumieniami odnowień.

W [BEUT 78], [MELA 79] wykazano, iż w sieci stanowisk obsługi typu M/M/n wyłącznie strumienie zgłoszeń opuszczających sieć oraz strumienie nie zawierające zgłoszeń "recyrkulujących" są strumieniami Poissona, co dowodzi, iż taka właśnie postać strumieni wewnętrznych w sieci nie jest warunkiem koniecznym iloczynowej postaci rozwiązań.
Z kolei w [JACK 63], [GORD 67] wykazano, iż iloczynową postacią rozwiązań charakteryzują się również zamknięte sieci stanowisk obsługi typu M/M/n.

W ostatnim dziesięcioleciu wiele wysiłku poświęcono na określenie: warunków, przy których możliwe jest uzyskanie iloczynowej postaci rozwiązań w sieciach stanowisk obsługi, własności tych rozwiązań oraz metod efektywnego ich uzyskiwania.

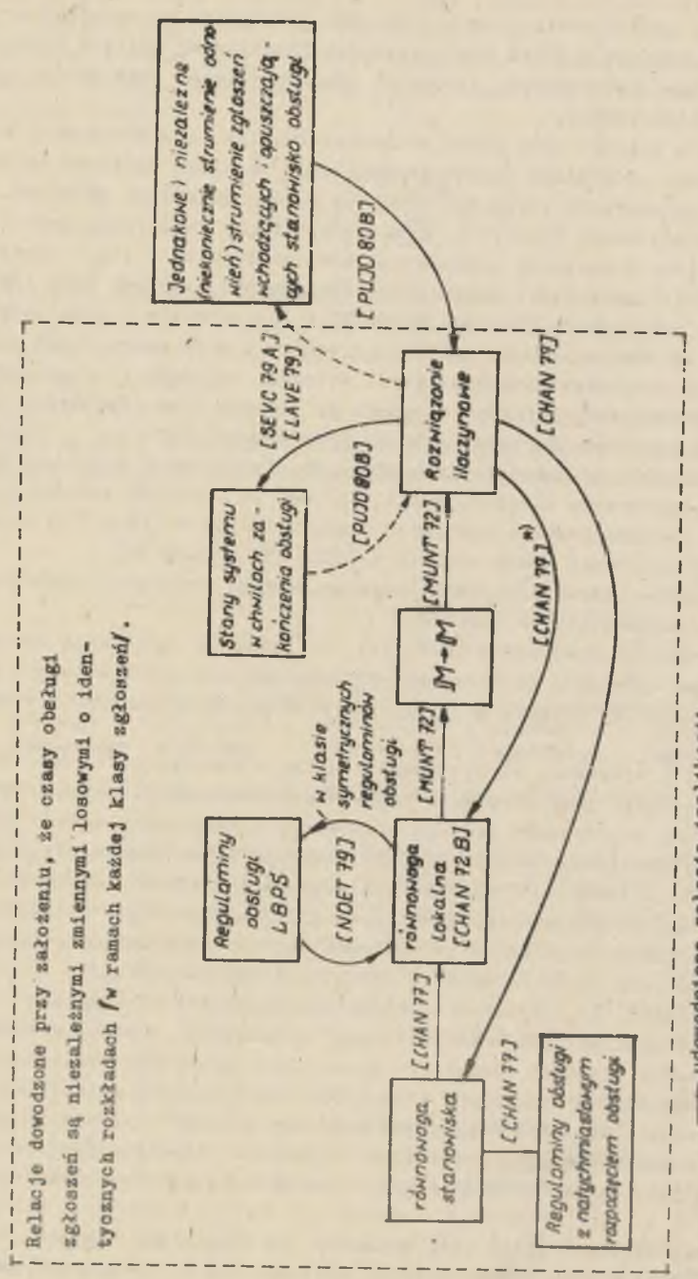
Szczególnie istotne były próby rozbudowy rodziny dopuszczalnych regulaminów obsługi, rozkładów czasów obsługi oraz podziału zgłoszeń na klasy, co znacznie rozszerzyło klasę modelowanych reguł przepływu zgłoszeń. Wymienić tu należy pracę [BASK 75], gdzie udowodniono iloczynową postać rozwiązań dla sieci otwartych, zamkniętych oraz mieszanych (tj. otwartych dla pewnych klas zgłoszeń i zamkniętych dla innych) z węzłami typu $\cdot/K_1/\infty$ $\cdot/K_1/1/\infty$/PS, $\cdot/K_1/1/\infty$/PLIFO oraz $\cdot/M/1/\infty$/FIFO (to ostatnie z mocą stanowiska obsługi zależną od liczby zgłoszeń w węźle, a więc można w ten sposób modelować też przypadek wielokanałowy). Autorzy założyli, iż zgłoszenia napływają strumieniami Poissona i należą do różnych klas. Marszruty zgłoszeń w sieci określone są przez zadanie prawdopodobieństw $p_{ij,kl}$ tego, że zgłoszenie klasy i po zakończeniu obsługi na stanowisku Y_j przejdzie do węzła Y_l jako zgłoszenie klasy k. Model ten nosi od nazwisk autorów nazwę modelu BCMP (bardzo podobne systemy rozpatrywane też w [KELL 75], niewielkie dalsze uogólnienia można znaleźć w [KELL 76], [BARB 76]).
Analiza warunków zapewniających iloczynową postać rozwiązań doprowadziła do szeregu fundamentalnych wyników.

W szczególności wykazano w [MUNT 72], że cechę tę posiadają wszystkie sieci złożone wyłącznie ze stanowisk obsługi posiadających własność M—M, a więc tym bardziej złożone ze stanowisk obsługi cechujących się równowagą lokalną.

W [CHAN 77] wykazano, że równowaga lokalna w odniesieniu do każdego stanowiska obsługi jest warunkiem koniecznym iloczynowej postaci rozwiązań, co więcej w przypadku każdego stanowiska o rozkładzie czasów obsługi różnym od wykładniczego warunkiem koniecznym jest też równowaga stanowiska. Wyklucza to możliwość uzyskania iloczynowej postaci rozwiązań przy założeniu regulaminów obsługi nie zapewniających równowagi stanowiska, a więc między innymi regulaminu naturalnego czy też regulaminów priorytetowych przy różnych od wykładniczego rozkładach czasów obsługi.

Spostrzeżenie, iż w znanych sieciach posiadających iloczynową postać rozwiązań, w chwilach zakończenia obsługi zgłoszenia rozkład prawdopodobieństwa liczby zgłoszeń obecnych w opuszczonym przez nie węźle jest identyczny z rozkładem stacjonarnym (dla sieci otwartych, natomiast identyczny z rozkładem stacjonarnym przy zmniejszonej o jedno liczbie zgłoszeń dla sieci zamkniętych) doprowadziło do wysunięcia hipotezy [PUJO 80B], iż taka cecha jest warunkiem koniecznym i wystarczającym iloczynowej postaci rozwiązań.

Z drugiej strony w [LABE 79A] wykazano, iż w dowolnej otwartej sieci stanowisk obsługi typu M/M/n rozkłady odstępów czasu pomiędzy chwilami



Rys. I.7. Schemat związków zachodzących pomiędzy różnymi pojęciami i własnościami sieci stanowisk obsługi posiadających rozkłady różne od wykładniczego

rozpoczynania obsług oraz rozkłady odstępów czasu pomiędzy chwilami zakończenia obsług są identyczne.
W odniesieniu do systemu z rys. I.6 można nawet wykazać, że procesy wejścia i wyjścia zgłoszeń są jednakowe i niezależne.

Sugeruje się [LABE 79A], [PUJO 80B], iż być może posiadanie tej własności jest poszukiwanym warunkiem istnienia iloczynowej postaci rozwiązań. Sformułowanie warunku koniecznego i wystarczającego jest tu jak widać sprawą otwartą.

Relacje pomiędzy różnymi własnościami węzłów sieci stanowisk obsługi a istnieniem iloczynowej postaci rozwiązań ilustruje schemat przedstawiony na rys. I.7.

Rozpatrzmy niektóre własności charakterystyczne dla wszystkich uzyskanych rozwiązań o postaci iloczynowej. Otóż, co wskazano np. w rozwiązaniach modelu BCMP, łączny rozkład prawdopodobieństwa liczby zgłoszeń w poszczególnych węzłach w sieci otwartej ma postać analogiczną do (I.2.1):

$$p(x_1,x_2,\ldots,x_M) = p_1(x_1)p_2(x_2)\ \ldots\ p_M(x_M), \qquad \text{(I.2.4)}$$

gdzie prawdopodobieństwa $p_j(x_i)$ wyznaczone są zgodnie z zależnościami obowiązującymi w systemie M/M/1, jeśli węzeł zawiera stanowisko obsługi różne od M/G/∞ lub też zgodnie z zależnościami obowiązującymi w systemie M/G/∞; przy założeniu, iż strumień zgłoszeń w takim izolowanym systemie jednostanowiskowym jest strumieniem Poissona o intensywności γ_i wyznaczonej z układu równań (I.2.3).

Tak więc w systemach otwartych (oddzielnie można wykazać, iż dotyczy to również systemów zamkniętych) dla określenia rozkładu zgłoszeń w systemie istotne są jedynie wartości oczekiwane czasów obsługi a nie ich rozkłady!
Ta własność niewrażliwości rozwiązań o postaci iloczynowej na postać rozkładu została udowodniona, przy bardzo ogólnych założeniach, w pracach [SCHA 78] oraz [STOY 78].

W przypadku analizy sieci zamkniętych układ równań (I.2.3) staje się nieoznaczony i poważnym zagadnieniem staje się wyznaczenie stałej normalizacyjnej K w (I.2.2).
Ze względu na bardzo dużą liczbę stanów niemożliwe jest wyznaczenie tej stałej wprost z definicji jako odwrotności sumy prawdopodobieństw wszystkich możliwych stanów sieci. Efektywne algorytmy wyznaczenia stałej normalizacyjnej były przedmiotem wielu prac [BUZE 73], [REIS 75A], [REIS 76], [REIS 77], [BALB 77], [BRUE 78], których omówienie oraz ujednoliconą charakterystykę zawiera [ZAHO 79].

Ze względu na fakt, iż nawet użycie tych algorytmów może prowadzić do pewnych trudności numerycznych (dokładność, zakres liczb zmiennoprzecinkowych) atrakcyjna wydaje się, przedstawiona w [REIS 78A], [REIS 78B] metoda, pozwalająca na bezpośrednie wyznaczanie charakterystyk zamkniętych

sieci stanowisk obsługi przy użyciu wyłącznie intensywności strumieni zgłoszeń oraz średnich długości kolejek.

Tak więc, ścisła analiza konwencjonalnych sieci stanowisk obsługi jest możliwa przy stosunkowo ogólnych założeniach, nie odzwierciedlających jednak w pełni realiów wielu modelowanych systemów (szczególnie często akcentuje się brak możliwości uwzględnienia różnych od wykładniczego rozkładów czasów obsługi przy naturalnym regulaminie obsługi). Stąd w ostatnich latach można zauważyć burzliwy rozwój przybliżonych metod analizy, które mogą być stosowane dla dowolnych sieci stanowisk obsługi.

Metody te - aczkolwiek na pozór bardzo różnorodne - sprowadzają się w ogólnym zarysie do rozdzielenia rozpatrywanej sieci stanowisk obsługi na pewne oddzielnie rozpatrywane fragmenty (podsystemy).

Przyjmowane są natomiast różne założenia upraszczające odnośnie sposobu opisu poszczególnych podsystemów (np. modelując je w uproszczeniu jako systemy jednostanowiskowe), a także odnośnie opisu łączących je strumieni zgłoszeń (np. przybliżając je strumieniami odnowień).

Próbę zestawienia założeń przyjmowanych w opublikowanych metodach przybliżonych podjęto w pracy [TRIP 79], charakteryzując każdą z nich przez podanie:

a) przyjmowanej charakterystyki rozkładów czasów obsługi w węzłach,
b) przyjmowanej charakterystyki strumieni łączących podsystemy,
c) założonej postaci modeli podsystemu i sposobu jego analizy,
d) sposobu łączenia rozwiązań poszczególnych podsystemów,

co prowadzi do zaczerpniętej z cytowanej pracy tabeli I.5.

Ostatnio w [PUJO 79A] opublikowano jeszcze jedną propozycję odnośnie przybliżonej analizy sieci otwartych, polegającą na przedstawieniu rozpatrywanej sieci w postaci acyklicznej sieci odpowiednio dobranych podsystemów - co w niektórych przypadkach jest możliwe.

Należy podkreślić, że w zależności od tego czy rozpatrywana sieć jest otwarta, czy zamknięta, konieczne jest rozpatrzenie jednego z następujących problemów:

i) w przypadku sieci otwartej konieczne jest upewnienie się co do jej stabilności. W tym celu rozwiązuje się układ równań I.2.3 (w którym występują jedynie intensywności strumieni zgłoszeń bez względu na ich charakterystykę) przyjmując, iż warunkiem wystarczającym stabilności jest, aby intensywność strumienia zgłoszeń γ_i zapewniała stabilność węzła Y_i zgodnie z zależnością (I.1.14), tj. aby zachodziło:

$$\rho_i = \frac{\gamma_i}{n_i \mu_i} < 1; \quad \mu_i = \left[B(b_i)\right]^{-1}, \quad i = 1,2,\ldots, M, \qquad (I.2.5)$$

Przedmiotem oddzielnych rozważań [LOYN 64] były systemy wieloetapowe z jednokanałowymi stanowiskami obsługi, w odniesieniu do których dowiedziono stabilności również w przypadku regularnego strumienia zgłoszeń

Tabela I.5

Zestawienie założeń przyjmowanych w różnych metodach przybliżonych (według [TRIP 79])

Metoda przybliżona oparta na analogii twierdzenia Nortona [SAUE 75]
- a) Wartość oczekiwana i współczynnik zmienności rozkładu
- b) Strumień Poissona z intensywnością zależną od obciążenia
- c) Dowolne podsystemy; przybliżenie strukturą o rozwiązaniu iloczynowym
- d) Analiza równowagi globalnej z węzłami odwzorowującymi przepływy

Metoda przybliżona oparta na analogii twierdzenia Nortona [SEVC 77]
- a) Wartość oczekiwana i współczynnik zmienności rozkładu
- b) Wartość oczekiwana i współczynnik zmienności rozkładu
- c) Dowolne podsystemy; przybliżenie strukturą o rozwiązaniu iloczynowym
- d) Analiza równowagi globalnej z węzłami odwzorowującymi przepływy

Metoda iteracyjna [CHAN 75B]
- a) Wartość oczekiwana
- b) Strumień Poissona z intensywnością zależną od obciążenia
- c) Pojedyncze stanowiska; przybliżenie rozwiązaniami modelu M/G/1/L
- d) Iteracyjne dopasowanie dla spełnienia przyjętych niezmienników

Metoda rozszerzonej postaci iloczynowej [SHUM 76]
- a) Wartość oczekiwana i współczynnik zmienności rozkładu
- b) Strumień Poissona
- c) Pojedyncze stanowiska; przybliżenie rozwiązaniami modelu M/G/1/L
- d) Postać iloczynowa, stała normalizacyjna wyznaczona przy użyciu iteracyjnego algorytmu uwzględniającego ograniczenie na przepływy

Metoda aproksymacji dyfuzyjnej [REIS 74]
- a) Wartość oczekiwana i współczynnik zmienności rozkładu
- b) Wartość oczekiwana i współczynnik zmienności rozkładu
- c) Pojedyncze stanowiska; rozwiązanie przybliżone modelu G/G/1 metodą aproksymacji dyfuzyjnej
- d) Postać iloczynowa rozwiązania ze stałą normalizacyjną

Metoda aproksymacji Kuhna [KUHN 76]
- a) Wartość oczekiwana i współczynnik zmienności rozkładu
- b) Wartość oczekiwana i współczynnik zmienności rozkładu
- c) Pojedyncze stanowiska; rozwiązanie przybliżone modelu G/G/1 metodą numeryczną
- d) Bezpośrednio - dotyczy wyłącznie systemów otwartych

Metoda Mariego [MARI 78]
- a) Rozkład prawie dowolny
- b) Strumień Poissona z intensywnością zależną od obciążenia
- c) Dowolne podsystemy; rozwiązanie przybliżone metodą numeryczną
- d) Iteracyjny algorytm dla spełnienia przyjętych niezmienników

(nadchodzących co t_o jednostek czasu), przy ograniczonych czasach obsługi na k pierwszych stanowiskach obsługi $k \leq M$ (tzn. istnieje takie α_j, że $B_j(\alpha_j) = 1$, $j = 1,2,\ldots,k$), jeśli $\alpha_j \leq t_o$ dla $j = 1,2,\ldots,k$ oraz spełniona jest zależność (I.2.5) dla pozostałych stanowisk obsługi.

ii) w przypadku sieci zamkniętej stabilność jest oczywiście zapewniona zawsze, jeśli tylko wszystkie czasy obsługi zgłoszeń mają skończone war-

tości średnie. Pojawia się natomiast zagadnienie normalizacji uzyskanych rozwiązań, czyli mówiąc precyzyjniej, zagadnienie wyznaczenia rzeczywistych intensywności strumieni zgłoszeń a nie tylko proporcji między nimi, do czego ogranicza się - w odniesieniu do systemów zamkniętych - sens nieoznaczonego układu równań I.3.3. W tym celu mogą być użyte bądź metody identyczne, jak w przypadku rozwiązań o postaci iloczynowej, bądź to metody iteracyjnego dopasowywania rozwiązań.

Wadą większości przybliżonych metod analizy sieci stanowisk obsługi jest brak możliwości określenia popełnionego błędu na drodze teoretycznej. Stąd też przeprowadza się zwykle testy metody na pewnych przykładach uznanych za reprezentatywne i próbuje ekstrapolować wyniki tych testów również na inne przypadki. Ponieważ taka, niczym głębiej nie uzasadniona, ekstrapolacja może być niebezpieczna w skutkach, z zasady przy stosowaniu metod przybliżonych do konkretnych systemów porównuje się uzyskane wyniki (przynajmniej w kilku punktach) z wynikami modelowania cyfrowego.

Dla ułatwienia analizy sieci stanowisk obsługi metodami ścisłymi oraz przybliżonymi, a także przy użyciu modelowania cyfrowego, opracowywane były w ostatnich latach w wielu ośrodkach pakiety programów umożliwiające zdefiniowanie struktury oraz parametrów rozpatrywanej sieci, wybór metody jej badania, sprawdzenie czy spełnione są założenia wymagane dla tej metody, przeprowadzenie analizy oraz wyemitowanie jej wyników w pewnych standaryzowanych formatach [REIS 78C].

Najbardziej rozwiniętymi spośród takich pakietów są QNAP [MERL 78], opracowany w instytucie INRIA (Francja) i RESQ [REIS 75A], [REIS 75B], [SAUE 77], opracowany przez ośrodek badawczy IBM.
W literaturze można też znaleźć informacje o skromniejszych, jeśli chodzi o zakres funkcjonalny, pakietach QSOLVE [ZAHO 77] Uniwersytetu w Toronto, SNAP [KRZE 77] Uniwersytetu w Stellenbosch, oraz ASQ [KELL 74] Uniwersytetu w Austin.

W odniesieniu do pojedynczych stanowisk obsługi wspominaliśmy o pewnych, najczęściej formułowanych zadaniach optymalizacji, których rozwiązanie prowadziło do wyboru regulaminu obsługi - jedynej cechy systemu, na którą zwykle można oddziaływać (w rozdziale II omówimy celowość oddziaływania na źródło zgłoszeń w odniesieniu do systemów z czasem obsługi zależnym od czasu oczekiwania).

W sieciach stanowisk obsługi można z reguły oddziaływać na znacznie większą liczbę parametrów, np. na rozdział łącznej mocy pomiędzy poszczególne stanowiska obsługi, marszrutę zgłoszeń w sieci oraz oczywiście regulaminy obsługi. Jak dotąd nie opracowano jednak żadnych ogólnych reguł w tym zakresie, z wyjątkiem przypadku systemów wieloetapowych, którym poświęcimy więcej uwagi w rozdziale IV.

I.2.2. Niekonwencjonalne sieci stanowisk obsługi

Jako niekonwencjonalne określać będziemy takie sieci stanowisk obsługi, w których bądź niektóre stanowiska nie spełniają założeń w definicji konwencjonalnych stanowisk obsługi, bądź też narzucone są pewne dodatkowe ograniczenia na sposób funkcjonowania sieci (np. sieć otwarta z ograniczoną maksymalną liczbą zgłoszeń, które mogą się znajdować w sieci, uzależnienie losowe działania różnych stanowisk obsługi).

Charakteryzując sieci stanowisk obsługi posiadające iloczynową postać rozwiązań zaznaczyliśmy możliwość uwzględnienia modyfikacji, polegających na dopuszczeniu zależności czasu obsługi od liczby zgłoszeń w węźle typu ·/M/1 oraz dopuszczeniu zależności intensywności strumieni zgłoszeń (jeśli są to strumienie Poissona) od liczby zgłoszeń w sieci. Wspomnieliśmy też o pewnych, specyficznych niekonwencjonalnych stanowiskach obsługi (np. ze stratami zgłoszeń), posiadających własność M—M, a więc mogących być elementami sieci o iloczynowej postaci rozwiązań.

Z nielicznymi dalszymi wyjątkami, które zostaną omówione nieco później, odstępstwo od założeń określających sieć konwencjonalnych stanowisk obsługi przekreśla możliwość uzyskania rozwiązań analitycznych.

Mogą natomiast być stosowane w odniesieniu do takich systemów metody przybliżone, omówione wyżej, z tym tylko, że rozpatrzenia wymaga w każdym przypadku indywidualnie stabilność sieci (można przy tym posługiwać się, w odniesieniu do poszczególnych stanowisk warunkami stabilności podanymi w punkcie I.1), a także dokładność zastosowanej metody.

Ze względu na znaczenie praktyczne takich modeli podejmowane prace, mające na celu rozwój metod analizy oraz badanie własności niekonwencjonalnych sieci stanowisk obsługi, obejmowały głównie różne warianty sieci z uzależnionymi losowo czasami obsługi w różnych węzłach oraz sieci z ograniczoną maksymalną liczbą zgłoszeń. Scharakteryzujemy najważniejsze kierunki i wyniki tych prac.

a) <u>Sieci stanowisk obsługi z uzależnionymi losowo czasami obsługi w różnych węzłach</u>

1) Wieloetapowe systemy obsługi z zależnymi losowo czasami obsługi

W pracy [KLEI 64], rozpatrując modele systemów telekomunikacyjnych modelowanych jako sieci stanowisk obsługi zwrócono uwagę, iż czasy transmisji pojedynczego komunikatu (pakietu danych) przez różne fragmenty sieci są zależne losowo (zwykle proporcjonalne). Z tego względu konieczne stało się zbadanie czy i na ile odstępstwo od założeń niezależności czasów obsługi w poszczególnych węzłach utrudnia wykorzystanie teorii sieci konwencjonalnych stanowisk obsługi dla modelowania układów komunikacyjnych.

Przy użyciu modelowania cyfrowego stwierdzono, że w przypadku gdy przez pojedyncze łącze transmitowane są komunikaty, których marszruty tylko na tym <u>jednym</u> łączu są identyczne, wówczas możliwe jest zaakceptowanie zało-

żenia o niezależności, co można uzasadnić faktem obsługiwania "mieszanki" zgłoszeń o czasach obsługi skorelowanych z różnymi wcześniejszymi procesami obsługi, na różnych stanowiskach. Jeśli natomiast fragment sieci komunikacyjnej ma charakter wieloetapowego systemu obsługi, w którym ciąg komunikatów przechodzi kolejno przez te same łącza, wówczas skutki zależności losowej czasów transmisji są istotne.

Równocześnie jednak autor podkreślił, iż nawet w najprostszym przypadku systemów dwuetapowych, pojedynczych nie był w stanie uzyskać rozwiązań analitycznych i wskazał trudności pojawiające się przy próbach analizy. Zupełnie oczywiste jest też występowanie zależności czasów obsługi zgłoszeń w różnych węzłach sieci w przypadku dyskretnych procesów produkcyjnych.

Dopiero w ciągu kilku ostatnich lat udało się uzyskać pewne wyniki w zakresie analizy takich systemów.

Rozpatrując sieci stanowisk obsługi posiadające iloczynową postać rozwiązań, przy stosunkowo ogólnych założeniach odnośnie marszruty zgłoszeń, w [KELL 76] wskazano, iż obejmują one również przypadek identycznych czasów obsługi zgłoszeń we wszystkich węzłach sieci przy stosowaniu regulaminu z podziałem procesora (PS), który w odniesieniu do transmisji danych (a tym bardziej procesów produkcyjnych), wydaje się mało realistyczny.

W [BOXM 78A] rozpatrywano dwuetapowy, pojedynczy system obsługi ze zgłoszeniami napływającymi strumieniem Poissona o intensywności λ i identycznymi czasami obsługi zgłoszeń na obu etapach (charakteryzującymi się dowolnym rozkładem o dystrybuancie B(x)). Czasy obsługi kolejnych zgłoszeń uważano przy tym za niezależne zmienne losowe.

Udowodniono, że warunkiem koniecznym i wystarczającym na to, by liczba zgłoszeń obsłużonych w okresie czynnym była skończona (z prawdopodobieństwem równym jedności) jest $\lambda E(b) \leq 1$, przy czym nierówność ostra zapewnia również skończoną wartość oczekiwaną czasów między kolejnymi chwilami rozpoczęcia okresów bezczynnych, a więc stabilność systemu.

Podano też zależności pozwalające na wyznaczenie rozkładów czasów oczekiwania w obu buforach oraz czasu pobytu zgłoszenia w drugim węźle, których praktyczne stosowanie wydaje się jednak być dosyć skomplikowane.

Część druga cytowanej pracy zawiera wyniki obliczeń numerycznych charakteryzujących rozpatrywany system. W szczególności stwierdzono, iż w porównaniu z analogicznym konwencjonalnym systemem dwuetapowym istotnemu obniżeniu - z wyjątkiem pewnych nieregularności przy małym obciążeniu - ulega wartość średnia (oraz jeszcze silniej wariancja) czasu pobytu zgłoszeń w drugim węźle, podczas gdy charakterystyki przebiegu obsługi w węźle pierwszym pozostają naturalnie bez zmian. Prawidłowość tę ilustruje tabela I.6.

Tabela I.6

Wpływ wprowadzenie identycznych czasów obsługi na obu etapach $B(x) = 1-e^{-\mu x}$ (wg [BOXM 78A])

Stopień wykorzystania stanowisk obsługi $\lambda E(b)$	Stosunek wartości średniej E(d) oraz wariancji V(d) w rozpatrywanym systemie do analogicznych wartości w systemie konwencjonalnym	
	$E(d)/E_k(d)$	$V(d)/V_k(d)$
0.1	1.06	1.05
0.75	0.80	0.22
0.99	0.09	0.0006

W pracach [CALO 77] oraz [CALO 78] rozpatrywano ogólniejszy przypadek wieloetapowych, pojedynczych systemów obsługi, przy założeniu proporcjonalności czasów obsługi b_j^i na poszczególnych etapach

$$b_j^i = \psi_j b^i, \quad i = 0,1,\ldots \quad j = 1,2,\ldots,M. \tag{I.2.6}$$

Przy założeniu, iż bufor S_1 nigdy nie jest pusty oraz $\psi_1=1$, $l=1,2,\ldots,M-1$, podano warunek stabilności (przytoczony w odniesieniu do systemu dwuetapowego z interpretacją pierwszego węzła jako źródła zgłoszeń jako zależność (I.1.22)) sprowadzający się do wymaganie, aby czasy obsługi były ograniczone. W tym przypadku zachodzi też równość czasów oczekiwania zgłoszeń we wszystkich buforach z wyjątkiem pierwszego.

$$w_j^i = w_2^i, \quad i = 0,1,\ldots \quad j = 2,3,\ldots,M. \tag{I.2.7}$$

Natomiast zakładając dowolny strumień zgłoszeń oraz dowolne rozkłady czasów obsługi można wykazać, że

$$w_2^i \leq w_3^i \leq \ldots \leq w_k^i \leq \ldots \leq w_M^i. \tag{I.2.8}$$

Rozpatrzmy obecnie przypadek ogólniejszy proporcjonalnych (a nie jak poprzednio identycznych) czasów obsługi zgłoszeń w kolejnych węzłach

$$\psi_j = \alpha^j, \quad j = 1,2,\ldots,M; \tag{I.2.9}$$

kiedy to, przy nigdy nie pustym buforze S_1 mamy:

$$w_1^i = \alpha^{1-2} w_2^i, \quad l = 2,3,\ldots,M. \tag{I.2.10}$$

Można też wykazać, że jeśli zachodzi:

$$\psi_1 \geq \alpha_o,\ \psi_M \geq \alpha_o,\ \psi_k = \alpha_o \quad \text{dla} \quad k = 2,3,\ldots,M-1;$$

to wówczas, przy dowolnym strumieniu zgłoszeń i dowolnych rozkładach czasów obsługi, nierówność (I.2.8) pozostaje w mocy.

Powyżej odwoływaliśmy się do przypadków, gdy rozpatrywany system wieloetapowy cechował się nigdy nie pustym buforem S_1. Taki system będziemy nazywać systemem nasyconym. Charakter systemu nasyconego ma wiele rzeczywistych procesów, np. dyskretne procesy produkcyjne, dla których elementy pobierane są z magazynu zaopatrywanego właściwie, tj. tak, aby nie zachodziła potrzeba zatrzymania produkcji z braku elementów.

Zauważmy jeszcze, że dwuetapowy, pojedynczy system nasycony jest identyczny z systemem jednokanałowym.

Przedmiotem przedstawionych w [CHOO 80] rozważań były dwuetapowe, pojedyncze systemy obsługi ze zgłoszeniami napływającymi strumieniem Poissona, w którym zależność czasów obsługi w obu węzłach określa rozkład dwuwymiarowy typu (I.1.24).

Autorzy podali metodę pozwalającą na numeryczne wyznaczanie charakterystyk rozpatrywanego systemu, którego własności ilustruje tabela I.3b (α oznacza współczynnik korelacji czasów obsługi).

Dodatnie skorelowanie czasów obsługi wpływa na zmniejszenie wartości średniej (oraz wariancji) czasu oczekiwania w drugim węźle. Porównując wartości zestawione w wierszach a) i b) tabeli I.3 widać, że zysk wynikający ze skorelowania czasów obsługi w rozpatrywanym systemie dwuetapowym jest mniejszy aniżeli miałoby to miejsce w przypadku, gdy bufor S_1 nigdy nie był pusty, co odpowiada omawianemu w rozdziale I.1 systemowi jednostanowiskowemu. Jest to naturalne, bowiem w tym drugim przypadku ze względu na gwarantowaną ciągłość pracy stanowiska $\mathfrak{X}_1$ skorelowanie jest w istocie silniejsze.

W omawianej pracy wyznaczono też charakterystyki rozpatrywanego systemu dwuetapowego, przy założeniu zerowej pojemności bufora pośredniego ($N_2 = 0$).

Badania wieloetapowych, pojedynczych systemów obsługi z uzależnionymi losowo czasami obsługi, prowadzone przy użyciu modelowania cyfrowego, przedstawiono w [MITC 77].

Przy $M = 2$, $N_2 = \infty$ uzyskano wyniki zgodne z wyżej omawianymi (pod względem jakościowym, wyraźne są natomiast różnice wartości średniego czasu oczekiwania w drugim węźle). Zakładając identyczność czasów obsługi na poszczególnych etapach autorzy stwierdzili też, że:

- przy $M = 2$ oraz $N_2 = 0$ średni czas oczekiwania w węźle drugim jest większy aniżeli w analogicznym systemie z nieskorelowanymi czasami obsługi,

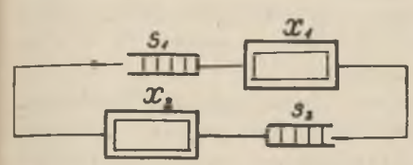
- przy $M = 2$ oraz $N_2 = 1$ skorelowanie czasów obsługi wpływa na średni czas oczekiwania negatywnie przy niskim stopniu wykorzystania stanowiska obsługi i pozytywnie przy wysokim stopniu wykorzystania,
- przy $N_2 = \infty$ oraz $2 \leq M \leq 25$ stosunek wartości średniej łącznego czasu oczekiwania w rozpatrywanym systemie do tej wartości w systemie bez zależności losowej czasów obsługi jest unimodalną, posiadającą minimum funkcją liczby etapów M. Położenie minimum przesuwa się w kierunku wyższych wartości M (i wartość ilorazu maleje) ze wzrostem stopnia wykorzystania stanowiska obsługi.

Niemniej, np. przy $\eta(\lambda) = 0.7$ już dla $M > 5$ wartość rozpatrywanego ilorazu przekracza 1. Wyniki symulacji są przy tym zgodne (w odniesieniu do wartości średnich) z relacją (I.2.8), orzekającą, iż czasy oczekiwania w kolejnych buforach są uporządkowane niemalejąco.

Jak dotąd, brak podobnych prac w odniesieniu do sieci otwartych o bardziej złożonej strukturze.

ii) Zamknięte systemy obsługi z zależnymi losowo czasami obsługi

Wpływ skorelowania czasów obsługi badano również w przypadku najprostszego strukturalnie wariantu sieci zamkniętej, złożonej z dwóch jednokanałowych węzłów, przez które kolejno przechodzi $\mathcal{N}$ zgłoszeń (patrz rys. I.8). System taki może reprezentować na przykład wariant zadania konserwatora, w którym spośród $\mathcal{N}$ urządzeń jedno pracuje (stanowisko $\mathfrak{X}_1$), a pozostałe pozostają w rezerwie. W przypadku uszkodzenia (zakończenie obsługi na stanowisku $\mathfrak{X}_1$) urządzenie zajmuje miejsce w kolejce do naprawy (odbywającej się na stanowisku $\mathfrak{X}_2$).



Rys. I.8. Prosta, zamknięta sieć stanowisk obsługi z $\mathcal{N}$ zgłoszeniami

W [SANK 77] można znaleźć kilka szczegółowych modeli uwzględniających różne rodzaje awarii urządzeń (oraz możliwości ich unikania przy odpowiedniej eksploatacji) prowadzących do losowego uzależnienia czasu pracy urządzenia oraz czasu jego naprawy, wraz z wyznaczeniem stopnia skorelowania tych czasów.

Przedstawiony system zamknięty może też posłużyć do modelowania systemów wieloprogramowych [GAVE 73], przy czym wskazywano na fakt uzależnienia losowego czasów obsługi na jednym ze stanowisk [LEWI 73].

Analizie przedstawionego na rys. I.8 systemu zamkniętego, przy założeniu losowej niezależności czasów trwania kolejnych obsług na stanowisku $\mathfrak{X}_2$ oraz skorelowania czasów obsługi kolejnych zgłoszeń na stanowisku $\mathfrak{X}_1$, poświęcono pracę [JACO 78]. Przy założeniu, iż uzależnienie rozpatrywanych zmiennych losowych odpowiada relacjom (I.1.26), rozpatrzono analitycznie przypadki $\mathcal{N} = 1,2$, przy czym stwierdzono silny wpływ przyjętego skorelowania czasów obsługi na charakterystyki systemu.

Niestety, przedstawiona metoda prowadzi przy $N > 2$ do układu równań, którego analityczne rozwiązanie nie jest możliwe. W cytowanej pracy zamieszczono więc wyniki badań symulacyjnych rozpatrywanego systemu przy $N = 5$.

Obszerne rezultaty badań symulacyjnych identycznego systemu można znaleźć także w pracy [MERL 78].

Wpływ skorelowania czasów obsługi (na jednym ze stanowisk) na charakterystyki systemu zamkniętego, aczkolwiek wyraźny, nie ma tak jednoznacznego charakteru jak w przypadku systemów wieloetapowych i nie będziemy go szczegółowo charakteryzować.

iii) Hierarchiczne systemy masowej obsługi

Szczególną postacią zależności losowej czasów obsługi zgłoszeń na różnych stanowiskach obsługi jest wymaganie, aby w pewnych przypadkach zgłoszenie było obsługiwane jednocześnie na kilku stanowiskach obsługi.

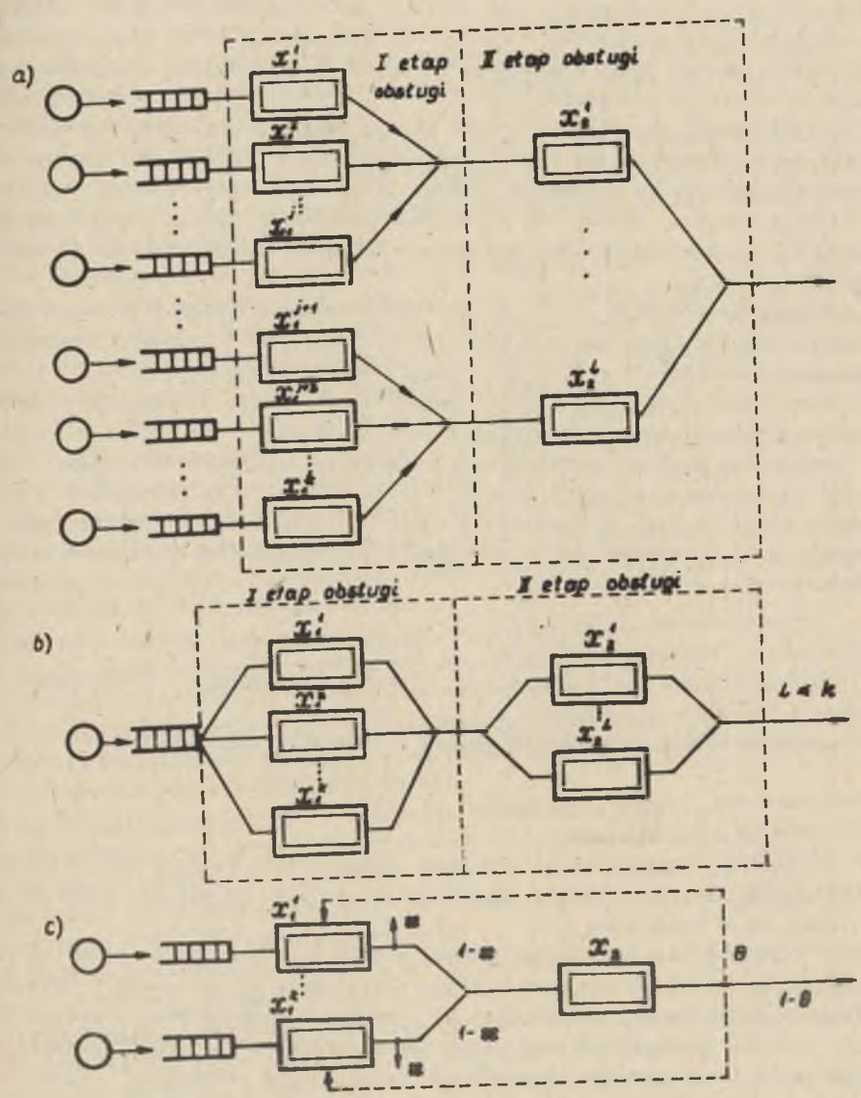
Jest to typowa sytuacja, jeśli chodzi o przebieg procesów obliczeniowych w maszynach cyfrowych, bowiem np. do wykonania programu niezbędne jest zapewnienie dostępu zarówno do pamięci operacyjnej, jak i do procesora, podobnie dla dokonania transmisji z pamięci dyskowej konieczne jest równoczesne przydzielenie właściwemu zadaniu zarówno kanału (obsługującego kilka jednostek pamięci), jak i jednostki pamięci dyskowej (i to w momencie odpowiedniego ustawienia głowic [SAUE 78]).

Systemy tej klasy są, ogólnie rzecz biorąc, poza zasięgiem dokładnych metod analitycznych i mogą być badane przy użyciu modelowania cyfrowego (np. struktura pakietu RESQ umożliwia rozpatrywanie tzw. rozszerzonych sieci stanowisk obsługi, uwzględniających możliwość obsługi zgłoszenia równocześnie na tzw. aktywnym stanowisku obsługi oraz jednym lub kilku tzw. pasywnych stanowiskach obsługi) bądź też przy użyciu metod przybliżonych.

W pracy [SAUE 78] omówiono metodę przybliżoną, opartą na dekompozycji systemu i zastosowano ją - jak to wykazały porównania z wynikami modelowania cyfrowego - z dobrym skutkiem do analizy przykładowego systemu.

Bardziej wnikliwej analizy doczekały się wieloetapowe systemy hierarchiczne cechujące się tym, iż zgłoszenie ζ_k może ubiegać się o przydzielenie kanału obsługi etapu "i" tylko wówczas, jeśli posiadało ono przez czas nie mniejszy aniżeli b^k_{i-1} (gdzie b^k_{i-1} czas obsługi na etapie "i-1") po jednym kanale obsługi etapów 1,2,...,i-1. Dopiero po zakończeniu obsługi na ostatnim etapie wszystkie zajęte przez rozpatrywane zgłoszenie kanały obsługi zostają zwolnione.

Wyróżnia się przy tym dwa podstawowe warianty zilustrowane w odniesieniu do systemów dwuetapowych na rys. I.9a, b. Są to systemy z dedykowanymi kanałami obsługi (w których na każdym etapie określony jest ściśle kanał - lub podzbiór kanałów - przeznaczony do obsługi zgłoszenia pochodzącego z konkretnego źródła) oraz systemy z uniwersalnymi kanałami obsługi.



Rys. I.9. Dwuetapowe, hierarchiczne systemy masowej obsługi
a) z dedykowanymi kanałami obsługi, b) z uniwersalnymi kanałami obsługi,
c) z dedykowanymi kanałami obsługi i niejednoznaczną marszrutą

Rozpatrywane przypadki ograniczają się w zasadzie w obu wariantach do zgłoszeń napływających strumieniami Poissona (bądź zawsze niepustych buforów wejściowych) oraz wykładniczych rozkładów czasów obsługi.

W pracach [FOST 78], [FOST 79], [FOST 80] badano systemy z dedykowanymi kanałami obsługi. Przeanalizowano (częściowo w sposób przybliżony) system dwu- oraz trzyetapowy z regulaminami naturalnymi na każdym etapie. Szczególną uwagę poświęcono systemom wieloetapowym, w których obsługa zgłoszeń w ramach etapu drugiego i dalszych przebiegała wg regulaminu obsługi z podziałem procesora.

Rozpatrywane były również - przy użyciu metod przybliżonych - systemy dwuetapowe, dopuszczające rezygnację zgłoszenia z dalszej obsługi (po zakończeniu obsługi na pierwszym etapie - z prawdopodobieństwem α) bądź też wymaganie ponownej obsługi (z prawdopodobieństwem Θ) przez zgłoszenie, którego obsługa została zakończona. Ogólny schemat takiego systemu przedstawia rys. I.9c.

Przypadek, w którym $\Theta = 0$, był przedmiotem pracy [WILL 74] (dopuszczono dowolny rozkład czasu obsługi zgłoszeń na pierwszym etapie), natomiast przypadek, w którym $k = 2$, $\alpha = 0$, analizowano w [PERR 81].

Zastosowane metody zostały, w oparciu o porównanie z wynikami modelowania cyfrowego, uznane za wystarczająco dokładne.

Dwuetapowe systemy hierarchiczne z uniwersalnymi kanałami obsługi (rys. I.8b) rozpatrywano w sposób ścisły w [WILE 73]. Dla liczby kanałów pierwszego etapu obsługi k większej od dwóch nie można uzyskać bezpośrednich rozwiązań, a przedstawiona metoda wymaga znalezienia pierwiastków wielomianu stopnia p

$$p = \begin{cases} 1.5\,k + 1 & \text{dla k parzystych} \\ 1.5\,k + 1.5 & \text{dla k nieparzystych} \end{cases}$$

a następnie rozwiązania układu równań liniowych o wymiarowości

$$0.5\,(k + 3).$$

W [WILE 75] zaproponowano natomiast drugą metodę analizy takich systemów, opartą na wykorzystaniu transformacji Z, nie prowadzącą również do bezpośrednich rozwiązań.

W [GWOZ 81] stwierdzono, iż żadnej z tych metod nie można uznać za zadowalającą, ponieważ pierwsza z nich, niezależnie od złożoności obliczeniowej przy większych wartościach k, prowadzi do rozwiązywania układu równań, który - szczególnie przy małej intensywności strumienia zgłoszeń - jest słabo uwarunkowany. Stosowalność drugiej, nie nadającej się do implementacji komputerowej, limituje duża pracochłonność przekształceń.

W cytowanej pracy przedstawiono też metodę przybliżonej analizy takich systemów (również o większej liczbie etapów), ilustrując na przykładach jej dokładność.

b) Sieci stanowisk obsługi z ograniczoną maksymalną liczbą zgłoszeń

Rozpatrując otwarte sieci konwencjonalnych stanowisk obsługi przyjmowano założenie o braku jakichkolwiek ograniczeń na pojemność buforów w poszczególnych węzłach. W praktyce taka sytuacja zdarza się niestety stosunkowo rzadko.

Również i sieci zamknięte, zakładające stałą liczbę rozpatrywanych zgłoszeń, ale dopuszczające możliwość zgromadzenia się wszystkich tych zgłoszeń w jednym buforze, odzwierciedlają bezpośrednio jedynie własności pewnej specyficznej grupy zadań.

W praktyce często konieczne jest modelowanie układów, w których pojemność buforów w poszczególnych węzłach sieci jest ograniczona (np. liczba programów mieszczących się w pamięci operacyjnej maszyny cyfrowej lub liczba pakietów możliwych do zapamiętania w procesorze komunikacyjnym sieci komputerowej) bądź też takich w których ograniczona (lecz niekoniecznie stała) jest ogólna liczba zgłoszeń krążących w całej sieci, względnie wydzielonych jej fragmentach (jak to ma miejsce w niektórych sieciach komputerowych). Ograniczenia obu tych typów mogą też występować łącznie.

Wprowadzenie ograniczeń na liczbę zgłoszeń akceptowanych przez węzeł sieci (lub jej fragment) wymaga zdefiniowania sposobu postępowania w przypadku przepełnienia, tj. próby naruszenia ograniczenia, przez skierowanie zgłoszenia do zapełnionego bufora.

Wyróżnimy następujące zasadnicze warianty reakcji na przepełnienie:

- blokowanie: Zgłoszenie powodujące przepełnienie pozostaje na stanowisku obsługi (w źródle), z którego zostało wyemitowane, uniemożliwiając mu dalszą działalność, tj. blokując je (w przypadku gdy zgłoszenie powodujące przepełnienie napłynęło z wielokanałowego stanowiska obsługi zablokowaniu ulega tylko jeden kanał).
- blokowanie prewencyjne: W przypadku zapełnienia bufora zgłoszeń następuje zatrzymanie pracy stanowisk obsługi, które obsługują zgłoszenia kierowane z dodatnim prawdopodobieństwem do zapełnionego bufora (tzw. blokowanie selektywne), względnie zatrzymanie pracy wszystkich stanowisk obsługi (tzw. blokowanie globalne).
- straty: Zgłoszenie powodujące przepełnienie jest bezpowrotnie tracone (wariant dopuszczalny tylko w sieci otwartej).
- rezygnacja: Zgłoszenie powodujące przepełnienie rezygnuje z obsługi w niedostępnym węźle (uzyskuje bez czekania w kolejce obsługę o zerowym czasie trwania), a następnie kieruje się dalej, zgodnie z obowiązującą w sieci marszrutą.
- powtarzanie obsługi: Zgłoszenie powodujące przepełnienie poddawane jest ponownej obsłudze na stanowisku obsługi, które właśnie opuściło (lokal-

ne powtarzanie obsługi) bądź też poddawane jest od początku obsłudze w całej sieci (globalne powtarzanie obsługi - wariant dopuszczalny tylko w sieci otwartej).

Scharakteryzujemy obecnie te nieliczne przypadki, w których możliwa jest ścisła analiza sieci stanowisk obsługi z ograniczoną liczbą zgłoszeń oraz omówimy pewne ogólne twierdzenie pozwalające wnioskować o wydajności oraz stabilności takich systemów.

i) Przypadki charakteryzujące się iloczynową postacią rozwiązań

W pracy [LAMS 77] wprowadzono pojęcie dwuwartościowych funkcji blokowania oraz funkcji doładowania, zdefiniowanych dla każdego stanu sieci BCMP (ściślej: funkcje takie definiowane są dla każdego stanu sieci i dla każdego ergodycznego podłańcucha Markowa, na które można podzielić łańcuch Markowa określający reguły zmiany klas zgłoszeń oraz marszrutę zgłoszeń). W przypadku, gdy dla pewnego stanu wartość funkcji blokowania jest równa zeru, wówczas napływające zgłoszenia są tracone, natomiast równa zeru wartość funkcji doładowania oznacza, iż w przypadku opuszczenia sieci przez zgłoszenie następuje natychmiast uzupełnienie liczby zgłoszeń do poprzedniego poziomu. Wartości tych funkcji równe jeden oznaczają brak odpowiednich ograniczeń.

W cytowanej pracy udowodniono, że jeśli dla każdego stanu funkcja doładowania jest równa jedności jedynie i wyłącznie w przypadku, gdy funkcja blokowania dla liczby zgłoszeń mniejszej o jedno jest też równa jedności (z odpowiednim uogólnieniem na podłańcuchy), wówczas rozpatrywany system charakteryzuje się iloczynową postacią rozwiązań.

Przedstawiono również przykłady zadań związanych z pracą węzła komunikacyjnego w sieci transmisji danych oraz gospodarką pamięcią operacyjną maszyn cyfrowych, które prowadzą do modeli o scharakteryzowanych wyżej ograniczeniach.

Ostatnio w [LENY 80] wykazano natomiast, że iloczynową postać rozwiązań zachowują również zamknięte sieci typu BCMP, w których bufory w pewnej liczbie węzłów mają ograniczoną pojemność, o ile reakcją na przepełnienie jest globalne blokowanie prewencyjne lub rezygnacja.
Analogicznie, jeśli obsługiwane zgłoszenia podzielone są na klasy i ograniczeniom podlega liczba zgłoszeń danej klasy akumulowana w węźle, wówczas globalne blokowanie prewencyjne, stosowane do wszystkich stanowisk obsługi obsługujących zgłoszenia należące do krytycznej klasy, również zapewnia iloczynową postać rozwiązań.

Poza wymienionymi wyżej systemami, z wyjątkiem prostych przypadków specyficznych systemów złożonych z 2-3 stanowisk obsługi, brak ścisłych rozwiązań analitycznych. Intensywnie rozwijane są natomiast prace nad metodami przybliżonymi, głównie w odniesieniu do wielostopowych systemów obsługi. Z rozwiązań, które - jak to sprawdzono na przykładach posługując się modelowaniem cyfrowym - wykazują dobrą dokładność oraz pewną uniwersalność, należy tu wymienić metody zaproponowane w [LABE 77], [BOXM 81].

ii) Wyznaczenie wydajności granicznej

W odniesieniu do sieci konwencjonalnych systemów masowej obsługi określony został warunek wystarczający stabilności, przy spełnieniu którego wydajność sieci - to jest wartość oczekiwana liczby zgłoszeń opuszczających sieć w jednostce czasu - jest równa łącznej intensywności strumieni zgłoszeń napływających do sieci. Wskazuje się na przykładach ([SCHW 76]), iż wydajność takiej sieci, w przypadku gdy jest ona niestabilna, może przekroczyć wydajność graniczną; fakt ten ma jednak niewielkie znaczenie praktyczne.

W przypadku ograniczonej pojemności buforów pośrednich brak takich jednoznacznych zależności.

Jeśli reakcją na przepełnienie jest strata zgłoszenia, wówczas sieć jest zawsze stabilna. Wydajność $V(\lambda,\mathcal{N})$ otwartej sieci typu BCMP z ograniczoną łączną liczbą zgłoszeń $\mathcal{N}$, intensywnością strumienia zgłoszeń λ i stratami w przypadku przepełnienia badana była w [SCHW 76].

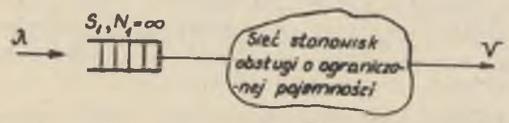
Udowodniono, iż $V(\lambda,\mathcal{N})$ jest rosnącą, ograniczoną funkcją obu argumentów, której kres górny równy jest wydajności granicznej identycznej (pod względem struktury i charakterystyk stanowisk obsługi) sieci, posiadającej bufory o nieograniczonej pojemności ($\mathcal{N}=\infty$).

Wynika stąd, iż w rozpatrywanym przypadku wprowadzenie ograniczeń na liczbę zgłoszeń powoduje obniżenie wydajności sieci, rozumianej jako wartość oczekiwana liczby całkowicie obsłużonych zgłoszeń, opuszczających sieć w jednostce czasu.

Prawidłowości tej nie można uogólnić na wszystkie sieci stanowisk obsługi z ograniczoną liczbą zgłoszeń. Modelowanie sieci komputerowych [KOBA 77], [WONG 78] jako sieci stanowisk obsługi spowodowało duże zainteresowanie określeniem wydajności sieci stanowisk obsługi z ograniczoną pojemnością buforów i stratami, względnie powtarzaniem obsługi.
Stwierdzono przy tym, iż w wielu konfiguracjach takich sieci można zauważyć początkowo wzrost wydajności sieci ze wzrostem intensywności strumieni zgłoszeń, podczas gdy dalsze zwiększenie tej intensywności prowadzi do silnego obniżenia wydajności sieci [RUDI 76], [WONG 78].

Aby przeciwdziałać temu zjawisku wprowadza się reguły sterowania przepływami w sieciach, mające na celu bądź to ograniczenie łącznej liczby zgłoszeń przyjmowanych do sieci, bądź rozdział buforów pomiędzy zgłoszenia różnych typów.
Nie komentując bardziej szczegółowo tych zjawisk, zwrócimy tylko uwagę, iż ograniczenie liczby zgłoszeń w węźle stanowisk obsługi (lub w większym jej fragmencie) jest skuteczną metodą zwiększenia wydajności również w przypadku, gdy rozpatrywana sieć charakteryzuje się innymi odstępstwami od postaci konwencjonalnej. Przykładem może tu być np. zjawisko wzrostu pracy własnej ze wzrostem liczby zgłoszeń we fragmencie sieci, występujące między innymi przy modelowaniu systemów komputerowych z pamięcią wirtualną [GELE 80A], gdzie brak efektywnych ograniczeń liczby zadań prowadzi do obniżenia sprawności systemu komputerowego.

W przypadku reguł reakcji na przepełnienie różnych aniżeli straty zgłoszeń, sprawą podstawową staje się wyznaczenie warunków stabilności. Fundamentalny wynik w tym zakresie zawiera praca [LAVE 78], w której rozpatrzono systemy masowej obsługi, złożone ze źródła zgłoszeń, bufora wejściowego o nieograniczonej pojemności oraz sieci stanowisk obsługi o dowolnej strukturze ale o ograniczonej pojemności, przyjmującej zgłoszenia wyłącznie z bufora wejściowego (rys. I.10).



Rys. I.10. Struktura systemów masowej obsługi spełniających twierdzenie Lavenberga

Założono przy tym, że zgłoszenia napływające ze źródła tworzą strumień odnowień o intensywności λ i mogą opuszczać sieć stanowisk obsługi wyłącznie po całkowitym zakończeniu obsługi, niedopuszczalny jest powrót (recykl) zgłoszeń z sieci do bufora wejściowego oraz że czasy obsługi zgłoszeń na poszczególnych stanowiskach mają rozkłady prawie dowolne (nie będziemy przytaczać dalszych, bardziej szczegółowych założeń, nie zawężających w istotny sposób klasy rozpatrywanych systemów).

Weźmy pod uwagę system nasycony odpowiadający omawianemu systemowi otwartemu (tj. system obsługi identyczny z wyżej określonym, przy założeniu, że bufor wejściowy S_1 nigdy nie może być pusty) i oznaczmy przez Λ wydajność tego systemu.

Lavenberg udowodnił, że wydajność V systemu otwartego można określić jako:

$$V = \begin{cases} \lambda & \text{gdy} \quad \lambda < \Lambda \\ \Lambda & \text{gdy} \quad \lambda > \Lambda \end{cases}$$

oraz że przy $\lambda < \Lambda$ system jest stabilny (tj. czasy oczekiwania zgłoszeń w buforze wejściowym są skończone), natomiast przy $\lambda > \Lambda$ niestabilny. Tak więc wydajność graniczna pokrywa się z wydajnością maksymalną. Przytoczony warunek stabilności jest bardzo użyteczny. O ile bowiem analiza systemu przedstawionego na rys. I.10 jest - ogólnie rzecz biorąc - niemożliwa przy użyciu metod dokładnych, to wyznaczenie Λ dla systemu nasyconego jest realne zarówno przy użyciu metod numerycznych (należy w tym celu określić stacjonarne prawdopodobieństwa ergodycznego łańcucha Markowa o skończonej liczbie stanów), jak i na drodze modelowania cyfrowego (jako że modelowaniu podlega system stabilny).

Wartość wydajności granicznej Λ zależy od wielu parametrów sieci stanowisk obsługi, przy czym charakter tych zależności jest stosunkowo słabo przebadany.

Najwięcej wysiłku skierowano dotąd na badanie wpływu różnych parametrów na wydajność wieloetapowych systemów obsługi z ograniczonymi buforami pośrednimi; problemowi określenia wskazań dla projektowania takich systemów poświęcony jest w całości rozdział IV. Tam też wskazane zostaną te przypadki systemów wieloetapowych z ograniczonymi buforami pośrednimi, dla których dostępne są rozwiązania analityczne. Rozpatrywano również inne specjalne przypadki systemów złożonych z niewielkiej liczby stanowisk obsługi, rozmieszczonych w różnych specyficznych komfiguracjach, np. [ARYA 72], [SHAR 73], [KONH 76], [REIS 78D], [WISZ 80], których nie będziemy jednak szczegółowo omawiać.

W niniejszym rozdziale przedstawiliśmy próbę klasyfikacji niekonwencjonalnych systemów masowej obsługi, omawiając jednocześnie podstawowe ich własności oraz aktualny dorobek w zakresie ich analizy, która powinna ułatwić orientację w coraz liczniejszych pracach dotyczących takich systemów.

Jak wynika z przedstawionego wyżej przeglądu, niewiele jest tu wyników o charakterze ogólnym, a badania zmierzały raczej do rozpoznania wybranych klas takich systemów.

Wskazane zostały zasadnicze trudności występujące przy analizie systemów niekonwencjonalnych. Zasygnalizowano też najważniejsze kierunki dalszych prac, wynikające bądź to z potrzeb aplikacyjnych, bądź to ze specyficznych własności rozpatrywanych systemów.

W dalszej części niniejszej pracy omówione zostaną w sposób szczegółowy dwie klasy niekonwencjonalnych systemów masowej obsługi:

- jednostanowiskowe systemy z czasem obsługi zależnym od czasu oczekiwania,
- wielostanowiskowe, wieloetapowe systemy obsługi,

na przykładzie których przedyskutowane zostaną precyzyjnie liczne zagadnienia, potraktowane w tym rozdziale - z konieczności - w sposób opisowy.

II. SYSTEMY MASOWEJ OBSŁUGI Z CZASEM OBSŁUGI ZALEŻNYM OD CZASU OCZEKIWANIA I NATURALNYM REGULAMINEM OBSŁUGI

Niniejszy rozdział poświęcony jest omówieniu własności systemów masowej obsługi z czasem obsługi zależnym od czasu oczekiwania (które będziemy w skróceniu nazywać systemami WAD), przy założeniu obsługi zgłoszeń w kolejności ich napływania.

W dalszych rozważaniach ograniczymy się do przypadku, gdy zgłoszenia napływają pojedynczo i zmienne losowe a^i, $i = 1,2,\ldots$, określające odstępy między chwilami napłynięcia kolejnych zgłoszeń, są nieujemne, niezależne losowo, o jednakowych rozkładach, posiadających dystrybuantę $A(x)$ i intensywność strumienia zgłoszeń

$$0 < \lambda = \left[E(a) \right]^{-1} < \infty$$

Przyjmiemy też, że stanowisko obsługi, przed którym może tworzyć się kolejka o nieograniczonej długości, ma stałą moc, obsługuje zgłoszenia pojedynczo i czas obsługi zgłoszenia ζ^i jest nieujemną zmienną losową b^i_w o identycznym dla wszystkich zgłoszeń, zależnym jedynie od czasu oczekiwania w^i rozkładzie, posiadającym warunkową dystrybuantę

$$B(x/w) = \Pr\left\{ b^i_w < x \;/\; w^i = w \right\}, \quad w \geq 0,\ i = 0,1,\ldots \qquad \text{(II.1)}$$

Dodatkowo założymy, jak to się tradycyjnie przyjmuje w systemach wielokanałowych z regulaminem FIFO, że kolejne zgłoszenie kierowane jest do tego kanału obsługi, który posiada najmniejszy zapas pracy.

Rozpatrzmy zapas pracy w dyskretnych chwilach czasu, bezpośrednio poprzedzających moment nadejścia kolejnego zgłoszenia, podobnie jak się to robi w systemach konwencjonalnych [GNED 71], [KIEF 56].

$$z^{[k]}_{FIFO}(\tau^{i+1} - 0) = \left[z^{[k]}_{FIFO}(\tau^i - 0) - a^{i+1} \right]^+, \quad \text{dla } k \neq 1,\ 1 \leq k \leq n;$$

$$z^{[1]}_{FIFO}(\tau^{i+1} - 0) = \left[z^{[1]}_{FIFO}(\tau^i - 0) - a^{i+1} + b^i_w \right]^+, \qquad \text{(II.2)}$$

gdzie:

$$[y]^+ = \max[0,y],$$

$$Z^{[l]}_{FIFO}(\tau^i - 0) = \min_{1\le l\le n} Z^{[j]}_{FIFO}(\tau^i - 0).$$

Tak więc ciąg $\left\{Z^i_{FIFO}\right\}$, dla którego określenia potrzebne jest oczywiście zadanie postaci Z^0_{FIFO}, jest mimo uwzględnienia zależności losowej b^i od w^i, podobnie jak w systemie konwencjonalnym, łańcuchem Markowa z ciągłą przestrzenią stanów.

W przypadku jednokanałowym równanie II.2 sprowadza się, przy uwzględnieniu (I.12), do postaci:

$$w^{i+1} = \max\left[0,\ w^i + b^i_w - a^{i+1}\right] = \left[w^i + b^i_w - a^{i+1}\right]^+,\ i=0,1,\dots, \qquad (II.3)$$

w oparciu o którą można łatwo określić dystrybuantę czasu oczekiwania i-tego zgłoszenia na rozpoczęcie obsługi (np. [POSN 73]),

$$W^{i+1}(x) = \Pr\left\{w^i + b^i_w - a^{i+1} \le x\right\} =$$

$$= \int_{y=0^-}^{\infty}\int_{z=0}^{\infty} \Pr\left\{a^{i+1} \ge y+z-x / w^i = y,\ b^i_w = z\right\} d_z B(z/y)\, d W^i(y) =$$

$$= \int_{y=0^-}^{\infty}\int_{z=0}^{\infty} \left[1 - A(y+z-x)\right] d_z B(z/y)\, d W^i(y). \qquad (II.4)$$

Jeśli więc w rozpatrywanym systemie istnieje dystrybuanta graniczna $W(x) = \lim_{i\to\infty} W^i(x)$, to spełnia ona równanie całkowe [POSN 73], [SZNE 68]:

$$W(x) = \int_{y=0^-}^{\infty}\int_{z=0}^{\infty} \left[1 - A(y+z-x)\right] d_z B(z/y)\, d W(y). \qquad (II.5)$$

Wśród jednokanałowych systemów WAD wyróżnimy specjalną klasę, którą określać będziemy mianem modelu podstawowego.

W modelu podstawowym:

i) rozkład odstępów między chwilami napłynięcia kolejnych zgłoszeń posiada gęstość a(x) taką, że $\bigvee_{x>0}\left\{a(x) > 0\right\}$,

ii) dystrybuanta warunkowa B(x/w) ma postać:

$$B(x/w) = \begin{cases} B_0(x) & \text{dla } w = 0, \\ B_i(x) & \text{dla } w_{i-1} < w \le w_i,\ i=1,2,\dots J, \end{cases} \qquad (II.6)$$

gdzie

$$0 = w_0 < w_1 < \dots < w_J = \infty.$$

iii) $E(a) < \infty$, $E(b_i) < \infty$ dla $i=0,1,\dots,J$.

W dalszym ciągu niniejszego rozdziału omówimy warunki stabilności systemów WAD z naturalnym regulaminem obsługi, a następnie przejdziemy do dyskusji metod pozwalających na efektywne wyznaczenie dystrybuanty W(x) w przypadku jednokanałowym.

Szczególnie wiele uwagi zwrócimy na system podstawowy, za pomocą którego można aproksymować dowolny system WAD. Analiza takiego modelu prowadzona była wyłącznie przy założeniu, że zgłoszenia napływają strumieniem Poissona oraz że J=2 [LIBU 71] lub $B_i(x) = 1-\mathrm{Exp}(-\mu_i x)$ [POSN 73].

Omówiony poniżej sposób osłabienia tych ograniczeń został po raz pierwszy wskazany i efektywnie zastosowany w [WOLI 81 A].

W niniejszej pracy przedstawiona zostanie szczegółowo ogólna metoda analizy systemów WAD w stacjonarnym stanie granicznym, rozpatrzonych zostanie także kilka przypadków szczególnych.

Posłużą one następnie do wskazania, przy użyciu przykładów numerycznych, własności systemów WAD z naturalnym regulaminem obsługi, prowadząc, między innymi do stwierdzenia, że stosowane we wcześniejszych pracach uproszczenia istotnie zawężały zakres stosowalności wyników.

Omówione zostaną również możliwości bezpośredniej analizy niektórych specyficznych, różnych od modelu podstawowego, systemów WAD, a także możliwości zwiększenia wydajności takich systemów obsługi.

II.1. Warunki stabilności

Orzekanie o stabilności rozpatrywanych systemów obsługi sprowadza się do sklasyfikowania łańcucha Markowa $\left\{Z^i_{FIFO}\right\}$ (metody badania łańcuchów Markowa z ciągłą przestrzenią stanów są ostatnio intensywnie rozwijane, np. [TWEE 75], [TWEE 76], [LASL 78]).

W szczególności rozpatrywany system masowej obsługi jest stabilny, gdy łańcuch $\left\{Z^i_{FIFO}\right\}$ jest ergodyczny.

W dalszym ciągu, mówiąc o warunkach stabilności, będziemy mieli na uwadze warunki ergodyczności odpowiedniego łańcucha Markowa.

Przed przystąpieniem do omówienia ogólniejszych reguł określających warunki stabilności systemów WAD przedstawimy pewien prosty przykład systemu jednokanałowego.

Niech zależność czasu obsługi od czasu oczekiwania ma postać:

$$B(x/w) = \mathbb{1}(x - cw - d), \tag{II.1.1}$$

co oznacza, iż czas obsługi kolejnego zgłoszenia zależy liniowo od czasu oczekiwania

$$b_w^i = cw^i + d, \quad c,d > 0.$$

Określmy, w sposób analogiczny do zaproponowanego przez Lindleya w odniesieniu do systemu GI/GI/1 ([LIND 52], patrz też [KLEI 75], [GNED 71]), czasy oczekiwania kolejnych zgłoszeń, wprowadzając dodatkowe oznaczenie

$$u_i = d - s^{i+1}, \tag{II.1.2}$$

$$w^1 = \left[w^o + cw^o + d - s^1\right]^+ = \max\left[0,(1+c)w^o + u_o\right],$$

$$w^2 = \left[w^1+cw^1+d-s^2\right]^+ = \max\left\{0,\ u_1+(1+c)\max\left[0,(1+c)w^o+u_o\right]\right\} =$$

$$= \max\left[0,u_1,u_1+(1+c)u_o+(1+c)^2w^o\right],$$

$$w^3 = \left[w^2+cw^2+d-s^3\right]^+ = \max\left\{0,u_2+(1+c)\max\left[0,u_2,u_2+(1+c)u_o+(1+c)^2w^o\right]\right\} =$$

$$= \max\left[0,u_2,u_2+(1+c)u_1,u_2+(1+c)u_1+(1+c)^2u_o+(1+c)^3w^o\right]$$

$$\vdots$$

$$w^i = \max\Big[0,u_{i-1},u_{i-1}+(1+c)u_{i-2},u_{i-1}+(1+c)u_{i-2}+(1+c)^2u_{i-3},\ \ldots,$$

$$u_{i-1}+(1+c)u_{i-2}+\ldots+(1+c)^{i-2}u_1,u_{i-1}+(1+c)u_{i-2}+\ldots+(1+c)^{i-1}u_1 +$$

$$+ (1+c)^{i-1}u_o+(1+c)^iw^o\Big].$$

Przy założeniu, że zmienne s^i, i=1,2,..., są losowo niezależne, niezależne losowo są również zmienne u_i, i=0,1,..., a więc rozkład zmiennej losowej w^i jest identyczny z rozkładem zmiennej losowej $\tilde{w}^i$,

$$\tilde{w}^i = \max\Big[0,u_o,u_o+(1+c)u_1,u_o+(1+c)u_1+(1+c)^2u_2,\ldots,u_o+(1+c)u_1 +$$

$$+ (1+c)^2u_2+\ldots+(1+c)^{i-2}u_{i-2}, \tag{II.1.3}$$

$$u_o+(1+c)u_1+(1+c)^2u_2+\ldots+(1+c)^{n-2}u_{i-2}+(1+c)^{i-1}u_{i-1}+(1+c)^iw^o\Big]$$

Oznaczmy

$$\left.\begin{aligned} U_i &= \sum_{j=0}^{i-1}(1+c)^ju_j \\ U_o &= 0. \end{aligned}\right\} \tag{II.1.4}$$

Tak więc

$$\tilde{w}^i = \max\left[U_o,U_1,U_2,\ldots,U_i+(1+c)^iw^o\right]. \tag{II.1.5}$$

W szczególnym przypadku $w^o = 0$ ciąg $\tilde{w}^i$ jest monotonicznie rosnący ze wzrostem "i", tak więc istnieje graniczna zmienna losowa (przyjmująca być może wartości nieskończone).

$$\tilde{w} = \lim_{i\to\infty}\tilde{w}^i \tag{II.1.6}$$

taka, że:

$$\tilde{w} = \sup_{i\geq 0} U_i. \tag{II.1.7}$$

W oparciu o (II.1.4) i (II.1.7) można łatwo stwierdzić, że skończoność wartości $\tilde{w}$, a ściślej, zachodzenie relacji

$$\tilde{W}(\infty) = 1$$

ma miejsce wyłącznie przy specyficznych założeniach dodatkowych narzuconych na zmienne losowe s^i, na przykład w przypadku, gdy $\bigvee_i\left\{u_i<0\right\}$, kiedy to $w^i = 0$.

Ogólnie rzecz biorąc, rozpatrywany system jest niestabilny bez względu na intensywność strumienia zgłoszeń (niestabilność strukturalna).

Na podstawie tego przykładu można się spodziewać, że dla stabilności systemów WAD wymagane będą pewne ograniczenia na maksymalny dopuszczalny czas obsługi, co istotnie znajduje odzwierciedlenie w podanych niżej warunkach.

Przypadek dowolnej, deterministycznej zależności czasu obsługi od czasu oczekiwania, przy ograniczeniu maksymalnego dopuszczalnego czasu obsługi, rozpatrzono w [SUGA 65] w odniesieniu do systemu wielokanałowego.

Przy założeniu, że:

i) $E(a) < \infty$

ii) $B(x/w) = H(x - \varphi(w))$,

gdzie: $\varphi(w)$ jest funkcją monotonicznie rosnącą, taką, że $\varphi(0) > 0$ oraz istnieje $\lim_{w\to\infty} \varphi(w) = \varphi(\infty) < \infty$

iii) stanowisko obsługi nie może być bezczynne w przypadku niepustej kolejki zgłoszeń,

wykazano, że rozpatrywany system obsługi jest stabilny, jeżeli zachodzi:

$$nE(a) > \varphi(\infty). \qquad \text{(II.1.8)}$$

Jeśli natomiast $nE(a) < \varphi(\infty)$, wówczas system jest niestabilny z wyjątkiem specjalnego przypadku, w którym

$$\bigvee_{i>0} \Pr\left\{n.a^i < \varphi(0)\right\}. \qquad \text{(II.1.9)}$$

kiedy to kolejka pozostaje pusta (autorzy nie wypowiedzieli się o przypadku równości $nE(a) = \varphi(\infty)$).

Warunki wystarczające stabilności jednokanałowych systemów z niedeterministyczną zależnością czasu obsługi od czasu oczekiwania rozpatrywano następnie w pracach [CALL 73], [SZNE 68]. W pierwszej z nich przyjęto upraszczające założenia, że zmienne losowe a^i, w^i, b_w^i są zmiennymi dyskretnymi, przyjmującymi wartości całkowitoliczbowe, nieujemne (będziemy je oznaczać $\tilde{a}^i$, $\tilde{w}^i$, $\tilde{b}_w^i$). Pozwoliło to autorowi skorzystać z dobrze znanych warunków ergodyczności łańcuchów Markowa z dyskretną przestrzenią stanów.

Wykazano, że w systemie, w którym

i) $\bigvee_i \Pr\left\{\tilde{b}_w^i - \tilde{a}^{i+1} = j\right\} > 0$ dla $j = -1,0,1$;

ii) $\bigvee_w E(\tilde{b}_w^i) < \infty$

warunkiem wystarczającym stabilności jest, aby

$$\lim_{w\to\infty} \sup E(\tilde{b}_w^i) < E(\tilde{a}). \qquad \text{(II.1.10)}$$

Postawiono też hipotezę, że analogiczny warunek powinien obowiązywać w przypadku systemu z ciągłymi zmiennymi losowymi.

W [SZNE 68] podano (bez dowodu) twierdzenie, iż jeśli dystrybuanta $A(x)$ jest funkcją absolutnie ciągłą i istnieją skończone wartości oczekiwane $E(a)$ oraz $E(b_w)$ dla każdego w, wówczas spełnienie warunku

$$\exists_{y_o} \bigvee_{y>y_o} \left\{\int_0^\infty u \, d_u B(u/y) < E(a)\right\} \qquad \text{(II.1.11)}$$

zapewnia istnienie rozkładu granicznego $W(x)$, będącego rozwiązaniem równanie całkowego (II.5).

Zagadnienie to rozpatrywano w sposób ścisły w [TWEE 75], a następnie przy ogólniejszych założeniach w [LASL 78].

Wykazano, że jeśli łańcuch Markowa $\{w^i\}$ określony zależnością (II.3) jest nieprzywiedlny oraz dystrybuanta $B(x/w)$ jest ciągłą funkcją w, wówczas warunkiem wystarczającym stabilności jest, aby:

$$\left.\begin{aligned} &\sup_w E(b_w) < \infty \\ &\lim_{w\to\infty} \sup E\left[b_w\right] < E(a) \end{aligned}\right\} \qquad \text{(II.1.12)}$$

W pracy tej podano też dwa, stosunkowo ogólne warunki zapewniające nieprzywiedlność, z których jeden ma postać:

$$\bigvee_{w\geq 0} \exists_{\substack{\delta>0 \\ \gamma>0}} \left[\Pr\left\{b_w^{i-1} - a^i < -\delta\right\} \geq \gamma\right] \qquad \text{(II.1.13)}$$

Tak więc wymaganie to jest w szczególności spełnione, gdy istnieje $a(x)>0$ dla każdego $x\geq 0$, to znaczy np. w przypadku gdy odstępy między chwilami napłynięcia kolejnych zgłoszeń mają rozkład Coxa.

W [MORO 80] udowodniono natomiast, że w przypadku systemów, do których zgłoszenia napływają strumieniem Poissona, spełnienie warunku

$$\sup_{0<w<\infty} E\left[b_w\right] < \infty \qquad \text{(II.1.14)}$$

zapewnia istnienie dla każdego $x>0$, $y>0$ granic

$$\lim_{i\to\infty} F^i(x/y) = \lim_{t\to\infty} F(t, x/y) = F(x)\delta(y), \qquad \text{(II.1.15)}$$

gdzie:

$$F^i(x/y) = \Pr\left\{w^i < x \;/\; \chi(0) = y\right\},$$

$$F(t, x/y) = \Pr\left\{\chi(t) < x \;/\; \chi(0) = y\right\},$$

$$\delta(y) = \Pr\left\{X_o < \infty \;/\; \chi(0) = y\right\}, \quad X_o = \inf\left\{t>0 : \chi(t) = 0\right\},$$

przy czym, jeżeli dodatkowo $\delta(0) < \infty$, wówczas $F(+\infty) = 1$ i $F(x)$ jest funkcją absolutnie ciągłą.

Zacytowany wynik przesądza o identyczności granicznych rozkładów zapasu pracy w systemie oraz czasu oczekiwania.

W [BACE 81] ograniczając się również do zgłoszeń napływających strumieniem Poissona rozpatrzono przypadek, gdy:

$$B(x/w) = \left[1 - H(w)\right] \mathbb{H}(x) + H(w)B(x), \qquad \text{(II.1.16)}$$

gdzie: 1-H(x) określa prawdopodobieństwo tego, że zgłoszenie opuści kolejkę (a więc wymagać będzie zerowego czasu obsługi) w czasie oczekiwania x,

przyjmując iż

$$\lim_{x\to\infty} H(x) > 0. \qquad \text{(II.1.17)}$$

Przy założeniu, że $\lambda E(b) < \infty$ udowodniono, iż warunkiem koniecznym i wystarczającym stabilności jest spełnienie nierówności:

$$\lambda E(b)\, H(\infty) < 1. \qquad \text{(II.1.18)}$$

W pracy [IZYD 81B] pokazano, że nierówność:

$$E(b_J) < E(a) \qquad \text{(II.1.19)}$$

jest warunkiem koniecznym i wystarczającym stabilności modelu podstawowego, w którym dodatkowo założono $E(b^i) > 0$ dla $i = 0,1,\ldots,J$. Tamże rozpatrzono też przypadek wielokanałowy przy założeniu, że:

i) postać B(x/w) jest dowolna, spełniająca warunek $\exists_{\psi<\infty} \forall_{w\geq 0} \left\{B(\psi/w) = 1\right\}$,

ii) $E(a) < \infty$

iii) $\exists_{\delta>0} \forall_{\alpha\geq 0} \forall_{i\geq 0} \left[\Pr\left\{-\delta \leq b^i - a^{i+1} \leq \delta \;/ w^i = \alpha\right\} > 0\right]$,

wykazując

- warunek wystarczający stabilności postaci[x)]

$$\limsup_{w\to\infty} E(b_w) < nE(a) \qquad \text{(II.1.20)}$$

oraz

- warunek konieczny stabilności postaci:

$$\liminf_{w\to\infty} E(b_w) \leq nE(a). \qquad \text{(II.1.21)}$$

x) Dla dowodu tego warunku wystarczy przyjąć tylko pierwszą część nierówności wymienionej w założeniu iii).

Tak więc w uproszczeniu można by powiedzieć, że warunki stabilności systemu WAD sprowadzają się do wymagania, aby z prawdopodobieństwem równym jedności czas obsługi zgłoszenia był zawsze skończony, bez względu na długość czasu oczekiwania oraz aby spełnione były nierówności uwzględniające postać granicznych rozkładów czasów obsługi przy $w \to \infty$.

Należy jednak podkreślić, że dodatkowe warunki, związane z odpowiednią nieprzywiedlnością łańcucha Markowa $\mathbb{Z}^i_{FIFO}$ (czy też jego nieokresowością, jeśli wymagane jest uniknięcie substabilności), nie mogą być pominięte.

W [LASL 78] można znaleźć prosty, aczkolwiek - jak sami autorzy przyznają - sztuczny przykład, wskazujący iż warunki typu II.1.19 mogą nie wystarczyć dla zapewnienia ergodyczności rozpatrywanego łańcucha Markowa.

II.2. Analiza modelu podstawowego

Przedstawimy obecnie (wg [WOLI 81D]) metodę wyznaczenia dystrybuanty stacjonarnego rozkładu czasu oczekiwanie zgłoszeń w modelu podstawowym, to znaczy metodę rozwiązywania równania całkowego (II.5) przy założeniu, że system jest stabilny, a więc, że spełniona jest nierówność (II.1.19).

W dalszych rozważaniach ograniczymy się do prawie dowolnych rozkładów odstępów między chwilami napłynięcia kolejnych zgłoszeń, to jest przypadku, gdy $\bar{a}(s)$ dane jest zależnością (I.1.7), przy $\alpha_0 = 1$.

Zauważmy, że $\bar{a}(s)$ można przedstawić w równoważnej postaci

$$\bar{a}(s) = (1 + \sum_{k=1}^{m-1} d_k s^k)/(1 + \sum_{k=1}^{m} c_k s^k) = \frac{N(s)}{D(s)}. \qquad \text{(II.2.1)}$$

Na wstępie udowodnimy, że poszukiwana dystrybuanta $W(x)$ ma postać:

$$W(x) = W_0\left[1 + \int_0^x v(y)dy\right]; \quad x \geq 0 \qquad \text{(II.2.2)}$$

gdzie v(y) jest funkcją nieujemną. (Dowód dla przypadku m = 1, zawiera [MORO 80] a przy dodatkowych ograniczeniach również [LIBU 71]). W tym celu oszacujemy, korzystając z (II.5), iloraz różnicowy

$$\frac{W(x+h)-W(x)}{h} = \int_0^\infty \int_0^\infty \frac{A(y+z-x)-A(y+z-x-h)}{h}\, d_z B(z/y)\, dW(y) \qquad \text{(II.2.3)}$$

dla $x+h,\ x \in [0,\infty)$.

W przypadku gdy A(x) jest dystrybuantą rozkładu Coxa, funkcja gęstości rozkładu a(x) jest ograniczona, tj.

$$\exists_{M} \forall_{x,x+h} \frac{A(x) - A(x+h)}{h} < M, \qquad \text{(II.2.4)}$$

stąd:

$$\frac{W(x+h) - W(x)}{h} < \int_{0^-}^{\infty}\int_{0}^{\infty} M d_z B(z/y)\, d\mathcal{W}(y) = M, \qquad \text{(II.2.5)}$$

tak więc badany iloraz różnicowy jest ograniczony, a więc $\mathcal{W}(x)$ jest absolutnie ciągła w przedziale otwartym $(0,\infty)$. Z drugiej strony, wprost z II.3 wynika, że $\mathcal{W}(x)$ jest też absolutnie ciągła (i tożsamościowo równa zeru) dla $x < 0$.

Tak więc $\mathcal{W}(x)$ posiada co najwyżej jeden punkt nieciągłości dla $x = 0$ i może być przedstawiona w postaci:

$$\mathcal{W}(x) = \mathcal{W}_0 \mathbb{H}(x) + \mathcal{W}_1(x), \qquad \text{(II.2.6)}$$

gdzie $\mathbb{H}(x)$ jest jedynką Heaviside'a, natomiast $\mathcal{W}_1(x)$ jest funkcją absolutnie ciągłą, taką że $\mathcal{W}_1(x) = 0$ dla $x<0$.

Zauważmy, że wobec faktu, iż $\forall_{x>0}\{a(x)>0\}$, wprost z (II.5) wynika[x] $\mathcal{W}_0>0$, a więc (II.2.6) można przedstawić w postaci:

$$\mathcal{W}(x) = \mathcal{W}_0\left[\mathbb{H}(x) + \frac{1}{\mathcal{W}_0}\mathcal{W}_1(x)\right], \qquad \text{(II.2.7)}$$

co przy wykorzystaniu twierdzenia Radona - Nikodyma [PELL 80] prowadzi wprost do (II.2.2). Zależność ta jest prawdziwa również dla wszystkich innych rozkładów zmiennej losowej a, posiadających ograniczoną gęstość rozkładu.

Przejdziemy obecnie do omówienia metody wyznaczania $\mathcal{W}_0$ oraz v(t).

Z ergodyczności rozpatrywanego procesu wynika, że równanie (II.5) posiada dokładnie jedno rozwiązanie o własnościach dystrybuanty. W dalszym ciągu, wychodząc z (II.5), wyprowadzimy układ równań różniczkowo-całkowych, wśród rozwiązań którego znajdować się będzie funkcja v(t).

---

x) Dla każdej skończonej wartości $w^n + b_w^n$ zachodzi bowiem

$$\exists_{\delta>0}\left[\Pr\left\{a^{n+1} \geq w^n + b_w^n\right\} = \delta\right]$$

W tym celu zróżniczkujemy n-krotnie obie strony równania (II.5) i rozpatrzymy sumę:

$$\mathcal{W}(x) + \sum_{k=1}^{m}(-1)^k c_k \mathcal{W}^{(k)}(x). \qquad \text{(II.2.8)}$$

Zauważmy, że wszystkie dystrybuanty są dystrybucjami, posiadają więc pochodne dystrybucyjne wszystkich rzędów [ANTO 73]. W dalszych rozważaniach oznaczać będziemy przez $\varphi^{(k)}(x)$ k-tą pochodną dystrybucji $\varphi(x)$. Po prostych przekształceniach otrzymamy:

$$\mathcal{W}(x)+\sum_{k=1}^{m}(-1)^k c_k \mathcal{W}^{(k)}(x) = \int_{0^-}^{\infty}\int_{0}^{\infty}\left[1 - A(y+z-x) - \sum_{k=1}^{m} c_k A^{(k)}(y+z-x)\right] d_z B(z/y)\, d\mathcal{W}(y) \qquad \text{(II.2.9)}$$

Z drugiej strony na podstawie (II.2.1) mamy:

$$\bar{a}(s) + \sum_{k=1}^{m} c_k s^k \bar{a}(s) = 1 + \sum_{k=1}^{m-1} d_k s^k. \qquad \text{(II.2.10)}$$

Przeprowadzając odwrotną transformację obu stron (II.2.10) otrzymujemy:

$$A(x) + \sum_{k=1}^{m} c_k A^{(k)}(x) = \mathbb{H}(x) + \sum_{k=1}^{m-1} d_k \mathbb{H}^{(k)}(x). \qquad \text{(II.2.11)}$$

Odejmując obie strony (II.2.11) od jedności otrzymujemy:

$$1 - A(x) - \sum_{k=1}^{m} c_k A^{(k)}(x) = 1 - \mathbb{H}(x) - \sum_{k=1}^{m-1} d_k \mathbb{H}^{(k)}(x). \qquad \text{(II.2.12)}$$

Korzystając z (II.2.12) można przedstawić (II.2.9) w postaci:

$$\mathcal{W}(x)+\sum_{k=1}^{m}(-1)^k c_k \mathcal{W}^{(k)}(x) = \int_{0^-}^{\infty}\int_{0}^{\infty}\left[1 - \mathbb{H}(y+z-x) - \sum_{k=1}^{m-1} d_k \mathbb{H}^{(k)}(y+z-x)\right] d_z B(z/y)\, d\mathcal{W}(y). \qquad \text{(II.2.13)}$$

Stosując znane w teorii dystrybucji zależności:

$$f^{(k)}(x) * g(x) = f(x) * g^{(n-k)}(x),$$

$$\left\{f(ax+b)\right\}^{(k)} = a^k f^{(k)}(ax+b),$$

gdzie: f(x), g(x) są dystrybucjami,
* oznacza działanie splotu,
otrzymujemy:

$$\mathcal{W}(x)+\sum_{k=1}^{m}(-1)^k c_k \mathcal{W}^{(k)}(x) = \int_{0^-}^{x} B(x-y/y)\,d\mathcal{W}(y)+\sum_{k=1}^{m-1}(-1)^k d_k \int_{0^-}^{x} B^{(k)}(x-y/y)\,d\mathcal{W}(y), \tag{II.2.14}$$

Wprowadźmy ciąg funkcji:

$$v_p(x) = v(x) \quad \text{dla} \quad 0 < x \leq w_p, \quad p = 1,2,\ldots J \tag{II.2.15}$$

Oczywiście $v_J(x) = v(x)$.
Korzystając z tej definicji,(II.2.2) i dzieląc obie strony przez $\mathcal{W}_0$ otrzymujemy:

$$1 + \int_0^x v_p(x) + \sum_{k=1}^{m}(-1)^k c_k v_p^{(k-1)}(x) = \left[B_o(x) + \sum_{k=1}^{m-1}(-1)^k d_k B_o^{(k)}(x)\right] +$$

$$+\sum_{i=1}^{p-1}\left\{\int_{w_{i-1}}^{w_i} B_i(x-y)v_i(y)dy + \sum_{k=1}^{m-1}(-1)^k d_k \int_{w_{i-1}}^{w_i} B_i^{(k)}(x-y)v_i(y)dy\right\} +$$

$$+\int_{w_{p-1}}^{x} B_p(x-y)v_p(y)dy + \sum_{k=1}^{m-1}(-1)^k d_k \int_{w_{p-1}}^{x} B_p^{(k)}(x-y)v_p(y)dy \quad \text{(II.2.16)}$$

dla $w_{p-1} < x \leq w_p$, $p = 1,2,\ldots,J$.
Wprowadźmy dodatkowo

$$u_p(x) = \begin{cases} v(x) & \text{dla} \quad w_{p-1} < x \leq w_p, \quad p = 1,2,\ldots,J \\ 0 & \text{dla} \quad x > w_p, \quad x \leq w_{p-1} \end{cases} \tag{II.2.17}$$

Zauważmy, że $\mathcal{W}_0 \int_0^{\infty} u_p(x)dx$ określa prawdopodobieństwo tego, że czas oczekiwania mieści się w przedziale $(w_{p-1}, w_p]$.
Oczywiście zachodzi:

$$\frac{1}{\mathcal{W}_0}\mathcal{W}(x) = \left[1 + \int_0^x v(x)dy\right] = \left[1 + \sum_{i=1}^{l}\int_0^{\infty} u_i(x)dx + \int_0^x u_{l+1}(y)dy\right], \tag{II.2.18}$$

gdzie l wyznacza się z nierówności:

$$w_l < x \leq w_{l+1}.$$

Na podstawie (II.2.16) i (II.2.17)

$$1 + \int_0^x v_p(y)dy + \sum_{k=1}^{m}(-1)^k c_k v_p^{(k-1)}(x) = \left[B_o(x) + \sum_{k=1}^{m-1}(-1)^k d_k B_o^{(k)}(x)\right] +$$

$$+\sum_{i=1}^{p-1}\left\{\int_0^x \left[B_i(x-y) - B_p(x-y)\right]u_i(y)dy + \sum_{k=1}^{m-1}(-1)^k d_k \int_0^x \left[B_i^{(k)}(x-y) - B_p(x-y)\right]u_i(y)dy\right\} +$$

$$+\int_0^x B_p(x-y)v_p(y)dy + \sum_{k=1}^{m-1}(-1)^k d_k \int_0^x B_p^{(k)}(x-y)v_p(y)dy \quad \text{(II.2.19)}$$

dla $0 < x \leq w_p$, $p = 1,2,\ldots,J$.

Rozszerzmy dziedzinę funkcji $v_p(x)$, $p = 1,3,\ldots,J$. Przyjmiemy, że wartości funkcji $v_p(x)$ dla $x \geq w_p$ określone są odpowiednim p-tym równaniem układu (II.2.18), (II.2.19), natomiast $v_p(x) = 0$ dla $x < 0$.
Oznaczmy przez $v_p(x, \acute{o}_1, \ldots, \acute{o}_{m-1})$, $p = 1,\ldots,J$ dowolne rozwiązanie układów równań (II.2.18), (II.2.19), zależne od warunku początkowego $\acute{o}_1 \ldots \acute{o}_{m-1}$.
Dokonując transformacji Laplace'a obu stron równania (II.2.19) i wprowadzając oznaczenie

$$f^*(s) = \int_0^{\infty} e^{-st} f(t)dt,$$

otrzymujemy po prostych przekształceniach:

$$v_p^*(s, \acute{o}_1, \ldots \acute{o}_{m-1}) = \left[1 + \sum_{k=1}^{m}(-s)^k c_k - (1 + \sum_{k=1}^{m-1}(-s)^k d_k)\bar{b}_p(s)\right]^{-1} \cdot$$

$$\cdot \left\{-1 + \bar{b}_o(s) + \sum_{k=1}^{m-1}(-s)^k d_k \bar{b}_o(s) + \sum_{k=1}^{m}(-1)^k c_k \sum_{i=1}^{k-1} s^{k-i}\acute{o}_i + \right.$$

$$\left. + \sum_{k=1}^{j-1}\left[\bar{b}_i(s) - \bar{b}_j(s)\right]u_i^*(s, \acute{o}_1, \ldots \acute{o}_{m-1})\left[1 + \sum_{k=1}^{m-1}(-1)^k d_k s^k\right]\right\}, \quad \text{(II.2.20)}$$

gdzie:

$$v_i^*(s,\sigma_1,\ldots,\sigma_{m-1}) = \int_0^\infty e^{-st} v_i(t,\sigma_1,\ldots,\sigma_{m-1})dt,$$

$$u_i^*(s,\sigma_1,\ldots,\sigma_{m-1}) = \int_0^\infty e^{-st} u_i(t,\sigma_1,\ldots,\sigma_{m-1})dt = \int_{w_{i-1}}^{w_i} e^{-st} v_i(t,\sigma_1,\ldots,\sigma_{m-1})dt, \tag{II.2.21}$$

skąd po dalszych przekształceniach otrzymujemy ostatecznie:

$$v_p^*(s,\sigma_1,\ldots,\sigma_{m-1}) =$$

$$= \frac{-1+\bar{b}_0(s)N(-s)+\sum_{k=2}^{m}(-1)^k c_k \sum_{i=1}^{k-1}\sigma_i s^{k-i}+\sum_{i=1}^{p-1}\left[\bar{b}_i(s)-\bar{b}_p(s)\right]u_i^*(s,\sigma_1,\ldots,\sigma_{m-1})N(-s)}{D(-s)-N(-s)\bar{b}_p(s)} =$$

$$= \frac{\varphi(s,\sigma_1,\ldots,\sigma_{m-1})}{D(-s)-N(-s)\bar{b}_p(s)}. \tag{II.2.22}$$

Na podstawie (II.2.2) oraz (II.2.16) widać, że transformata Laplace'a - Stieltjesa dystrybuanty $W(x)$ należy do rodziny funkcji $\bar{f}(s,\sigma_1,\ldots,\sigma_{m-1})$ takich, że:

$$\bar{f}(s,\sigma_1,\ldots,\sigma_{m-1}) = w_o\left[1 + v_j^*(s,\sigma_1,\ldots,\sigma_{m-1})\right]. \tag{II.2.23}$$

Tak więc określiliśmy transformatę poszukiwanej dystrybuanty z dokładnością do m stałych $(w_o,\sigma_1,\ldots,\sigma_{m-1})$.
Oznaczmy przez $\hat{\sigma}_1, \ldots, \hat{\sigma}_{m-1}$ poszukiwane wartości stałych $\sigma_1, \ldots, \sigma_{m-1}$. Dla wyznaczenia $\hat{\sigma}_1, \ldots, \hat{\sigma}_{m-1}$ wykorzystamy warunek ograniczoności transformaty Laplace'a-Stieltjesa dystrybuanty, a wówczas wartość w_o będzie wynikała wprost z warunku normalizacji.

Rozpatrzmy pierwiastki równania

$$D(-s) - N(-s)\bar{b}_j(s) = 0. \tag{II.2.24}$$

Dowodzi się (patrz [COHE 69]), że równanie (II.2.24) ma dokładnie (m-1) pierwiastków s_r, r = 1, ..., m-1 takich, że $Re(s_r) > 0$.

Ponieważ moduł transformaty Laplace'a-Stieltjesa dowolnej dystrybuanty jest zawsze mniejszy od jedności przy $Re(s) > 0$, więc musi zachodzić:

$$\varphi_j(s_r,\sigma_1,\sigma_2,\ldots,\sigma_{m-1}) = 0, \quad r=1,2,\ldots,m-1. \tag{II.2.25}$$

Układ równań (II.2.25) jest, jak łatwo sprawdzić, liniowy ze względu na $\sigma_1,\ldots,\sigma_{m-1}$.

Jeżeli wszystkie równania należące do układu(II.2.25)są liniowo niezależne, to możliwe jest jednoznaczne określenie stałych $\hat{\sigma}_1,\ldots,\hat{\sigma}_{m-1}$ jako pierwiastków tego układu równań.

Zauważmy, że w przypadku gdy niektóre pierwiastki s_r równania (II.2.24) są wielokrotne, wówczas odpowiednie równania układu (II.2.25) są identyczne, co nie przekreśla jednak możliwości znalezienia $\hat{\sigma}_1,\ldots,\hat{\sigma}_{m-1}$. Załóżmy, że wśród s_r występuje jeden pierwiastek l-krotny, podczas gdy wszystkie pozostałe są pojedyncze i ponumerujmy rozpatrywane pierwiastki tak, aby:

$$s_i = s_{i+1} = \ldots = s_{i+l-1}.$$

W takim przypadku należy w układzie (II.2.25) zastąpić równania:

$$\varphi_j(s_k,\sigma_1,\ldots,\sigma_{m-1}) = 0, \quad k = i+1,i+2,\ldots,i+l-1 \tag{II.2.26}$$

równaniami:

$$\left(\frac{\partial^k \varphi_j(s,\sigma_1,\ldots\sigma_{m-1})}{\partial s^k}\right)_{s=s_k} = 0 \tag{II.2.27}$$

$$k=i+1,i+2,\ldots,i+l-1$$

i znaleźć pierwiastki $\hat{\sigma}_1,\ldots,\hat{\sigma}_{m-1}$ tak zmodyfikowanego układu równań. Znając stałe $\hat{\sigma}_1,\hat{\sigma}_2,\ldots,\hat{\sigma}_{m-1}$ można łatwo wyznaczyć w_o z warunku normalizacji jako

$$w_o = \frac{E(a) - E(b_j)}{c_1 = \sum_{k=1}^{m}(-1)^k c_k \hat{\sigma}_{k-1} + \sum_{i=0}^{j-1}\left[E(b_i) - E(b_j)\right]p_i},$$

gdzie:

$$p_i = \begin{cases} u_i^*(0,\hat{\sigma}_1,\ldots,\hat{\sigma}_{m-1}) & \text{dla } i > 0, \\ 1 & \text{dla } i = 0. \end{cases} \tag{II.2.28}$$

Po łatwych, ale pracochłonnych przekształceniach można też wykazać, że:

$$E(w) = \frac{c_2 - d_2 - d_1 E(b_J) - 0{,}5\, E(b_J^2)}{E(b_J) - E(a)} -$$

$$- \frac{c_2 + \sum_{k=3}^{m} (-1)^k c_k \hat{\sigma}_{k-2} + \sum_{i=1}^{J-1} \alpha_i u_i^{(1)}(0, \hat{\sigma}_1, \ldots, \hat{\sigma}_{m-1}) - \sum_{i=0}^{J-1} (0{,}5\beta_i + d_1 \alpha_i) p_i}{-c_1 + \sum_{k=2}^{m} (-1)^k c_k \hat{\sigma}_{k-1} + \sum_{i=0}^{J-1} \alpha_i p_i}, \quad \text{(II.2.29)}$$

gdzie:

$$\alpha_i = E(b_J) - E(b_i) \qquad u_i^{(1)}(0, \hat{\sigma}_1, \ldots, \hat{\sigma}_{m-1}) = \frac{d}{ds} u_i(s, \hat{\sigma}_1, \hat{\sigma}_2, \ldots, \hat{\sigma}_{m-1})\Big|_{s=}$$

$$\beta_i = E(b_J^2) - E(b_i^2)$$

Rozpatrzmy też dalsze charakterystyki systemu.

Wyznaczmy $\eta^*(\lambda)$ - średni jednostkowy czas obsługi zgłoszeń, bezwzględu na ich czas oczekiwania

$$\eta^*(\lambda) = \int_0^{\infty} \int_0^{\infty} x \, dB(x/w) dW(x). \quad \text{(II.2.30)}$$

Łatwo widać, że:

$$\eta^*(\lambda) = \mathcal{W}_o \sum_{i=0}^{J} p_i E(b_i). \quad \text{(II.2.31)}$$

Oczywiście stopień wykorzystania stanowiska obsługi

$$\eta(\lambda) = \lambda \eta^*(\lambda) = \lambda \mathcal{W}_o \sum_{i=0}^{J} p_i E(b_i). \quad \text{(II.2.32)}$$

Niech zmienna losowa g^i oznacza łączny czas pobytu i-tego zgłoszenia w systemie:

$$g^i = b_w^i + w^i. \quad \text{(II.2.33)}$$

Tabela II.1

Wyznaczanie dystrybuanty $W(x)$ dla modelu podstawowego WAD

$$\bar{a}(s) = \frac{1 + \sum_{k=1}^{m-1} d_k s^k}{1 + \sum_{k=1}^{m} c_k s^k} = \frac{N(s)}{D(s)}.$$

$$B(x/w) = \begin{cases} B_o(x) & \text{dla } w = 0 \\ B_i(s) & \text{dla } w_{i-1} < w < w_i, \quad i = 1, 2, \ldots, J; \end{cases}$$

gdzie:

$$0 < w_o < w_1 < \ldots < w_{J-1} < w_J = \infty.$$

$$\bar{w}(s) = \mathcal{W}_o \left[1 + v_J^*(s, \hat{\sigma}_1, \hat{\sigma}_2, \ldots, \hat{\sigma}_{m-1})\right],$$

gdzie funkcję $v_J^*(s, \sigma_1, \sigma_2, \ldots, \sigma_{m-1})$ wyznacza się iteracyjnie z równań

$$v_p^*(s, \sigma_1, \sigma_2, \ldots, \sigma_{m-1}) = \frac{-1 + \bar{b}_o(s) N(-s) + \sum_{k=2}^{m} (-1)^k c_k \sum_{i=1}^{k-1} s^{k-i} \sigma_i}{D(-s) - N(-s) \bar{b}_p(s)} +$$

$$+ \frac{\sum_{i=1}^{p-1} \left[\bar{b}_i(s) - \bar{b}_p(s)\right] u_i^*(s, \sigma_1, \sigma_2, \ldots, \sigma_{m-1}) N(-s)}{D(-s) - N(-s) \bar{b}_p(s)} = \frac{\varphi_p(s, \sigma_1, \ldots, \sigma_{m-1})}{D(-s) - N(-s) \bar{b}_p(s)},$$

$$u_p^*(s, \sigma_1, \sigma_2, \ldots, \sigma_{m-1}) = \int_{w_{p-1}}^{w_p} e^{-st} v_p(t, \sigma_1, \sigma_2, \ldots, \sigma_{m-1}) dt,$$

$$v_p^*(s, \sigma_1, \sigma_2, \ldots, \sigma_{m-1}) = \int_0^{\infty} e^{-st} v_p(t, \sigma_1, \sigma_2, \ldots, \sigma_{m-1}) dt,$$

natomiast stałe $\hat{\sigma}_1, \hat{\sigma}_2, \ldots, \hat{\sigma}_m$ są rozwiązaniami układu równań liniowych z ewentualnymi modyfikacjami - patrz (II.2.26) oraz (II.2.27).

$$\varphi_J(s_r, \sigma_1, \sigma_2, \ldots, \sigma_{m-1}) = 0 \quad r = 1, 2, \ldots, m-1,$$

przy czym wartości s_r są pierwiastkami równania $D(-s) - N(-s)\bar{b}_J(s) = 0$ takimi, że $\mathrm{Re}(s_r) > 0$. Stałą $\mathcal{W}_o$ wyznacza się ze wzoru:

$$\mathcal{W}_o = \frac{E(a) - E(b_J)}{c_1 - \sum_{k=2}^{m} (-1)^k c_k \hat{\sigma}_{k-1} + \sum_{i=1}^{J-1} \left[E(b_i) - E(b_J)\right] u_i^*(0, \hat{\sigma}_1, \hat{\sigma}_2, \ldots, \hat{\sigma}_{m-1}) + E(b_o) - E(b_J)}.$$

Dystrybuantę rozkładu granicznego $G(x) = \lim_{i\to\infty} G^{i}(x)$ można wyznaczyć wprost na podstawie (II.2.33) jako:

$$G(x) = \int_0^x B(x-y/y)d\mathcal{W}(y). \qquad (II.2.34)$$

Wyprowadzone wyżej relacje, określające dystrybuantę czasu oczekiwania w modelu podstawowym WAD, zebrano w tabeli II.1. Biorąc pod uwagę przyjęte założenie o prawie dowolnym rozkładzie odstępów między chwilami napłynięcia kolejnych zgłoszeń oraz postaci zależności czasu obsługi od czasu oczekiwania, można, dla każdego jednokanałowego systemu WAD, dobrać model podstawowy przybliżający z dowolną dokładnością parametry zarówno procesu napływania, jak i obsługi zgłoszeń. Tak więc, teoretycznie rzecz biorąc, możemy rozpatrzyć z dowolną dokładnością każdy jednokanałowy system WAD z regulaminem naturalnym. Użyta do analizy badanego modelu metoda odcinkowego określania funkcji gęstości rozkładu czasu oczekiwania była pierwotnie stosowana do systemów z ograniczonym czasem oczekiwania (patrz np. [GNED 71]).
Na tej drodze uzyskano też w [LIBU 71], [POSN 73] wyniki analityczne dla niektórych specjalnych przypadków ze zgłoszeniami napływającymi strumieniem Poissona i ograniczeniami na postać zależności $B(x/w)$. Pierwszy przypadek analizy systemu WAD ze strumieniem zgłoszeń różnym od strumienia Poissona zawierała dopiero praca [WOLI 80B].

W następnym punkcie przedstawimy rozwiązania kilku szczególnych przypadków modelu podstawowego. Łatwo przy tym zauważyć, że praktyczne wykorzystanie przedstawionej metody wiąże się jednak ze znaczną pracochłonnością, wzrastającą silnie ze wzrostem J (określającej liczbę przedziałów, w których różne są rozkłady czasów obsługi), ponieważ J razy konieczne jest znajdowanie oryginału funkcji $v_p(t,\delta_1,\ldots,\delta_{m-1})$, a więc dokonywanie odwrotnej transformacji Laplace'a oraz całkowanie jej w przedziale, dla określenia $u_p^*(s,\delta_1,\ldots,\delta_{m-1})$, przy czym złożoność tej funkcji rośnie ze wzrostem p. Niezależnie od tego przy rzędzie rozkładu Coxa wyższym niż trzeci ($m > 3$) konieczne jest numeryczne poszukiwanie pierwiastków równania (II.2.24).

II.3. Przypadki szczególne modelu podstawowego

Omówiony w poprzednim punkcie model podstawowy umożliwił nam bardzo ogólne ujęcie szerokiego wachlarza różnych systemów cechujących się zależnością czasu obsługi od czasu oczekiwania. Poniżej podamy jego wyspecyfikowanie dla pewnych, wybranych przypadków szczególnych, z których wie-

le wykorzystamy w dalszych częściach niniejszej pracy dla lepszego poznania własności systemów WAD.

a) Zgłoszenia napływające strumieniem Poissona

W przypadku, gdy zgłoszenia napływają strumieniem Poissona, zależności przedstawione w tabeli 1 znacznie się upraszczają i przyjmują postać:

$$\bar{s}(s) = \frac{1}{1 + c_1 s}, \qquad (II.3.1)$$

gdzie: $c_1 = E(a) = \frac{1}{\lambda}$ (λ - intensywność strumienia zgłoszeń).

$$\bar{w}(s) = \mathcal{W}_0\left[1 + v_J^*(s)\right], \qquad (II.3.2)$$

gdzie funkcję $v_j^*(s)$ wyznacza się iteracyjnie z równań:

$$v_p^*(s) = \frac{-1 + \bar{b}_0(s) + \sum_{i=1}^{p-1}\left[\bar{b}_i(s) - \bar{b}_p(s)\right]u_i^*(s)}{1 - \frac{1}{\lambda}s - \bar{b}_p(s)} \qquad (II.3.3)$$

przy

$$u_p^*(s) = \int_{w_{p-1}}^{w_p} e^{-st}v_p(t)dt, \qquad v_p^*(s) = \int_0^{\infty} e^{-st}v_p(t)dt.$$

Stała $\mathcal{W}_0$ wyraża się w tym przypadku wzorem:

$$\mathcal{W}_0 = \frac{1 - \lambda E(b_J)}{1+\lambda\left[\sum_{i=1}^{J-1}\left[E(b_i)-E(b_J)\right]u_i^*(0)+E(b_0)-E(b_J)\right]}, \qquad (II.3.4)$$

natomiast

$$E(w) = \frac{1}{2}\frac{\lambda E(b_J^2)}{1-\lambda E(b_J)} - 0{,}5\lambda\mathcal{W}_0\frac{\sum_{i=0}^{J-1}p_i\left[E(b_J^2)-E(b_i^2)\right]-2\sum_{i=1}^{J-1}\left[E(b_J)-E(b_i)\right]u_i^{(1)}(0)}{1-\lambda E(b_J)}, \qquad (II.3.5)$$

gdzie:

$$u_i^{(1)}(0) = \frac{d}{ds}\, u_i^{*}(s)\Big|_{s=0},$$

$$p_i = \begin{cases} 1 & \text{dla } i = 0 \\ u_i^{*}(0) & \text{dla } i > 0. \end{cases}$$

Analiza takiego przypadku nie wymaga więc nigdy obliczeń numerycznych. W [LIBU 71] rozpatrywano też dla takiego systemu rozkład liczby zgłoszeń w kolejce.

Niech π_k oznacza prawdopodobieństwo tego, że w chwili zakończenia obsługi pozostaje w systemie k zgłoszeń.

Wprowadźmy funkcję tworzącą $\pi(\Theta)$,

$$\pi(\Theta) = \sum_{k=0}^{\infty} \pi_k \Theta^k. \qquad (II.3.6)$$

Wykazano, że:

$$\pi(\Theta) = \bar{g}(\lambda - \lambda\Theta), \qquad (II.3.7)$$

gdzie: $\bar{g}(s)$ jest transformatą Laplace'a-Stieltjesa dystrybuanty rozkładu czasu pobytu zgłoszeń w systemie, zadanej wzorem (II.2.34).

Bardziej szczegółowe wyniki można uzyskać po przyjęciu dodatkowych założeń upraszczających.

a1) <u>Wykładnicze rozkłady czasów obsługi</u>

W pracy [POSN 73] rozpatrywano przypadek, gdy:

$$B_0(x) = B_1(x) = 1 - e^{-\mu_1 x},$$

$$B_i(x) = 1 - e^{-\mu_i x}; \quad i=2,3,\ldots,J, \qquad (II.3.8)$$

uzyskując następujące wyrażenie na określone zależnością (II.2.17) funkcje $u_p(x)$:

$$u_p(x) = \begin{cases} f_p(x) & \text{dla } w_{p-1} < x < w_p, \quad p=1,2,\ldots,J \\ 0 & \text{dla } x > w_p, \quad x \le w_{p-1}. \end{cases} \qquad (II.3.9)$$

$$f_1(x) = \lambda e^{(\lambda-\mu_1)x},$$

$$f_j(x) = \frac{\lambda}{\mu_j-\mu_1-\lambda}\left[(\mu_j-\mu_1)e^{-\mu_1 x} - \lambda e^{(\lambda-\mu_j)x}\, e^{(\mu_j-\mu_1-\lambda)w_{j-1}}\right] +$$

$$+ \sum_{r=1}^{j-1} \frac{\lambda K_r}{\mu_j-\mu_r-\lambda}\left[(\mu_j-\mu_r)e^{-\mu_r x} - \lambda e^{(\lambda-\mu_j)x}\, e^{(\mu_j-\mu_r-\lambda)w_{j-1}}\right],$$

dla

$$\mu_j - \mu_r - \lambda \neq 0; \quad j=2,3,\ldots,J; \quad r=1,2,3,\ldots,J-1. \qquad (II.3.10)$$

gdzie:

$$K_1 = e^{\lambda w_1} - 1,$$

$$K_j = \frac{\lambda}{\mu_j-\mu_0-\lambda}\left[e^{(\mu_j-\mu_1)w_j} - e^{(\mu_j-\mu_1-\lambda)w_{j-1}}\, e^{\lambda w_j} + \right.$$

$$\left. + \sum_{r=1}^{j-1} \frac{\lambda K_r}{\mu_j-\mu_r-\lambda}\left[e^{(\mu_j-\mu_r)w_j} - e^{(\mu_j-\mu_r-\lambda)w_{j-1}}\, e^{\lambda w_j}\right]\right. . \quad j=2,3,\ldots,J.$$

Przy czym dla kompletnego określenia dystrybuanty $W(x)$ należy wyznaczyć W_0 z warunku normalizacji:

$$W_0\left[1 + \sum_{i=1}^{J} \int_0^{\infty} u_i(x)dx\right] = 1. \qquad (II.3.11)$$

Tak więc rozkład czasu oczekiwania jest w tym przypadku kombinacją liniową rozkładów wykładniczych. W przypadku gdyby istniała taka para j,r ($j = 2,3,\ldots,J$; $r=1,2,3,\ldots,J-1$), dla której $\mu_j-\mu_r-\lambda = 0$, wówczas zależności określające funkcje $f_j(x)$ skomplikowałyby się do postaci kombinacji liniowej rozkładów wykładniczych oraz rozkładów Erlanga drugiego rzędu.

W cytowanej pracy wyznaczono też przedziałami funkcję gęstości rozkładu czasu pobytu zgłoszenia w systemie $g(x)$

$$g(x) = W_0 \sum_{r=1}^{j-1} \mu_r K_r \exp(-\mu_r x) + \mu_j \exp(-\mu_j x) \int_{w_{j-1}}^{x} e^{\mu_j y} f_j(y)dy, \qquad (II.3.12)$$

gdzie j wyznacza się jako $w_{j-1} < x \le w_j$.

a2) Przypadek J = 2

Analizę modelu podstawowego ze zgłoszeniami napływającymi strumieniem Poissona oraz J = 2 przedstawiono w [LIBU 71]. Ponieważ dla przypadku tego wzory (II.3.1)-(II.3.5) nie ulegają istotnym uproszczeniom więc nie będziemy ich w odpowiednio zmodyfikowanej postaci przytaczać.
Zamieścimy natomiast zależność definiującą transformatę Laplace'a-Stieltjesa dystrybuanty czasu pobytu zgłoszenia w systemie

$$\bar{g}(s) = \mathcal{W}_0\left\{\bar{b}_2(s)v_2^*(s) + u_1^*(s)\left[\bar{b}_1(s) - \bar{b}_2(s)\right] + \bar{b}_0(s)\right\}, \quad \text{(II.3.13)}$$

gdzie, zgodnie z poprzednio wprowadzonymi oznaczeniami

$$v_2^*(s) = \int_0^\infty e^{-sx}v_2(x)dx = \int_0^\infty e^{-sx}v(x)dx,$$

$$u_1^*(s) = \int_0^{w_1} e^{-sx}v(x)dx.$$

Na podstawie (II.3.7) oraz (II.3.13) określa się funkcję tworzącą liczby zgłoszeń oczekujących w kolejce

$$\mathcal{K}(\theta) = \mathcal{W}_0 \frac{\bar{b}_2(\lambda-\lambda\theta)-\theta\bar{b}_0(\lambda-\lambda\theta)-\theta u_1^*(\lambda-\lambda\theta)\left[\bar{b}_1(\lambda-\lambda\theta) - \bar{b}_2(\lambda-\lambda\theta)\right]}{b_2(\lambda-\lambda\theta) - \theta} \quad \text{(II.3.14)}$$

W dalszym ciągu podamy wyspecyfikowanie zależności określających charakterystyki systemu WAD dla przypadków wybranych postaci rozkładów $B_i(x)$; i = 0,1,2.
Będziemy przy tym zakładać, że $B_0(x) = B_1(x)$ i wprowadzimy oznaczenie $D = w_1$.
Tak więc:

$$B(x/w) = \begin{cases} B_0(x) = B_1(x) & \text{dla} \quad 0 \leq w \leq D, \\ B_2(x) & \text{dla} \quad w > D. \end{cases} \quad \text{(II.3.15)}$$

Możemy mieć więc do czynienia z obsługą o rozkładzie czasu trwania $B_1(x)$ z prawdopodobieństwem $\mathcal{W}(D)$ lub o rozkładzie czasu trwania $B_2(x)$ w przeciwnym razie.

Tak więc zachodzi

$$\gamma^*(\lambda) = E(b_2)\left[1-\mathcal{W}(D)\right] + E(b_1)\mathcal{W}(D) = E(b_2)-\left[E(b_2)-E(b_1)\right]\mathcal{W}(D). \quad \text{(II.3.16)}$$

i) Regularny rozkład czasu obsługi [WOLI 79]
Niech

$$\left.\begin{aligned} B_0(x) = B_1(x) &= \mathbb{H}(x - B_s), \\ B_2(x) &= \mathbb{H}(x - B_L). \end{aligned}\right\} \quad \text{(II.3.17)}$$

W [WOLI 79] wykazano, że:

$$v_1^*(s) = \frac{\lambda\left[1 - e^{-sB_s}\right]}{s-\lambda\left[1-e^{-sB_s}\right]}, \quad \text{(II.3.18)}$$

$$v^*(s) = \frac{\lambda\left[1 - e^{-sB_s}\right] - \lambda\left[e^{-sB_s} - e^{-sB_L}\right]u_1^*(s)}{s - \lambda\left[1 - e^{-sB_L}\right]}, \quad \text{(II.3.19)}$$

gdzie:

$$u_1^*(s) = \int_0^D e^{-st}v_1(t)dt,$$

natomiast

$$\mathcal{W}_0 = \frac{1 - \lambda B_L}{1 - \lambda(B_L-B_s)\zeta(D)}, \quad \text{(II.3.20)}$$

gdzie:

$$\zeta(x) = \sum_{j=0}^{k} e^{\lambda(x-jB_s)} \frac{\left[-\lambda(x-jB_s)\right]^j}{j!},$$

dla $jB_s \leq x < (j+1)B_s$. (II.3.21)

W szczególności

$$\mathcal{W}(x) = \mathcal{W}_0\,\zeta(x) \quad \text{dla} \quad x \leq D, \quad \text{(II.3.22)}$$

a stąd prawdopodobieństwo wykonania obsługi o rozkładzie $B_1(x)$ wynosi $\mathcal{W}(D)$,

$$\mathcal{W}(D) = \mathcal{W}_o \zeta(D) = \frac{1 - \lambda B_L}{[\zeta(D)]^{-1} - \lambda(B_L - B_s)}. \qquad (II.3.23)$$

Po żmudnych przekształceniach wyznaczono także średni czas oczekiwania zgłoszenia na rozpoczęcie obsługi

$$E(w) = \frac{\lambda B_L^2}{2(1-\lambda B_L)} + \lambda(B_L - B_s)\frac{2\nu + (B_L + B_s)\zeta(D)}{2[1-\lambda(B_L - B_s)\zeta(D)]}, \qquad (II.3.24)$$

gdzie:

$$\nu = \zeta(D)\, D + \frac{1}{\lambda}\left\{n+1 - \sum_{j=0}^{k}\sum_{i=0}^{j} e^{\lambda(D-jB_s)}\frac{[-\lambda(D-jB_s)]^i}{i!}\right\},$$

zaś k wyznacza się z nierówności $kB_s \leq D < (k+1)B_s$.

ii) Wykładniczy rozkład czasu obsługi [LIBU 71], [POSN 73]

Niech

$$B_o(x) = B_1(x) = 1 - e^{-\mu_1 x}, \qquad B_2(x) = 1 - e^{-\mu_2 x}. \qquad (II.3.25)$$

Wykazano, że w tym przypadku zachodzi:

$$v^*(s) = \frac{\lambda}{\mu_1 - \lambda + s} + \frac{\lambda^2(\mu_1 - \mu_2)e^{-(\mu_1-\lambda)D}}{(\mu_1+\lambda)(\mu_1-\lambda+s)(\mu_2-\lambda+s)} e^{-sD}, \qquad (II.3.26)$$

$$\mathcal{W}_o = \begin{cases} \dfrac{\mu_1(\mu_2-\lambda)(\mu_1-\lambda)}{\mu_1\mu_2(\mu_1-\lambda) - \lambda(\mu_1-\mu_2)\left\{\mu_1 - \lambda + \lambda\left[1-e^{-(\mu_1-\lambda)D}\right]\right\}} & \text{dla } \mu_1 \neq \lambda \\ \dfrac{\mu_1(\mu_2-\lambda)}{\mu_1\mu_2 - \lambda(\mu_1-\mu_2)(1+\lambda T)} & \text{dla } \mu_1 = \lambda. \end{cases} \qquad (II.3.27)$$

W szczególności

$$\mathcal{W}(x) = \mathcal{W}_o\left\{1 + \frac{\lambda}{\mu_1-\lambda}\left[1-e^{-(\mu_1-\lambda)x}\right]\right\} \quad \text{dla} \quad x \leq D \qquad (II.3.28)$$

natomiast wartości $u_1^*(0)$, $u_1^{(1)}(0)$ niezbędne do określenia $E(w)$ według wzoru (II.3.5) wyrażają się zależnościami:

$$\left.\begin{aligned} u_1^*(0) &= \frac{\lambda}{\mu_1-\lambda}\left[1-e^{-(\mu_1-\lambda)D}\right], \\ u_1^{(1)}(0) &= \frac{\lambda}{\mu_1-\lambda} D\, e^{-(\mu_1-\lambda)D} - \left[1-e^{-(\mu_1-\lambda)D}\right]\frac{\lambda}{(\mu_1-\lambda)^2}. \end{aligned}\right\} \qquad (II.3.29)$$

iii) Czasy obsługi o rozkładzie Coxa II rzędu [WOLI 80B]

Niech

$$\left.\begin{aligned} \overline{b}_o(s) = \overline{b}_1(s) &= \frac{1 + d_{11}s}{1 + c_{11}s + c_{21}s^2}, \\ \overline{b}_2(s) &= \frac{1 + d_{12}s}{1 + c_{12}s + c_{22}s^2}. \end{aligned}\right\} \qquad (II.3.30)$$

Zgodnie z konwencją omówioną w rozdziale I parametry d_{1i}, c_{1i}, c_{2i} wyrazimy jednoznacznie przez wartość oczekiwaną $E(b_i)$ oraz współczynnik zmienności rozkładu c_{Bi} jak następuje:

	$0,5 \leq c_{Bi} \leq 1$	$c_B > 1$
c_{1i}	$E(b_i)$	$E(b_i)\left[1+c_{Bi}\right]$
c_{2i}	$0,5\left[E(b_i)\right]^2(1-c_{Bi})$	$0,5(1+c_{Bi})\left[E(b_i)\right]^2$
d_{1i}	0	$E(b_i)c_{Bi}$

i oczywiście

$$E(b_i^2) = \left[E(b_i)\right]^2\left[1 + c_{Bi}\right]. \qquad (II.3.31)$$

Łatwo sprawdzić, że zachodzi relacja

$$c_{1i} - d_{1i} = E(b_i). \qquad (II.3.32)$$

Załóżmy

$$c_{B1} = c_{B2} = c_B. \qquad (II.3.33)$$

W rozpatrywanym przypadku

$$v_1^*(s) = \frac{l_o+sl_1}{m_o+m_1s+s^2} = \frac{l_o+\hat{s}_1 l_1}{\hat{s}_1-\hat{s}_2}\,\frac{1}{s-\hat{s}_1} + \frac{l_o+\hat{s}_2 l_1}{\hat{s}_2-\hat{s}_1}\,\frac{1}{s-\hat{s}_2}, \qquad \text{(II.3.34)}$$

a stąd

$$v_1(t) = \frac{\hat{s}_1 l_1+l_o}{\hat{s}_1-\hat{s}_2}\,e^{\hat{s}_1 t} + \frac{\hat{s}_2 l_1+l_o}{\hat{s}_2-\hat{s}_1}\,e^{\hat{s}_2 t}, \qquad \text{(II.3.35)}$$

$$\mathcal{W}(x) = \mathcal{W}_o\left[1 + \frac{\hat{s}_1 l_1+l_o}{\hat{s}_1(\hat{s}_1-\hat{s}_2)}\,e^{\hat{s}_1 x} + \frac{\hat{s}_2 l_1+l_o}{\hat{s}_2(\hat{s}_2-\hat{s}_1)}\,e^{\hat{s}_2 x}\right] \quad \text{dla } x \leq D \qquad \text{(II.3.36)}$$

$$\mathcal{W}_o = \frac{1-\lambda E(b_2)}{1+\lambda\left[E(b_1)-E(b_2)\right]\left[1+u^*(0)\right]} \qquad \text{(II.3.37)}$$

oraz

$$E(W) = \frac{E(b_2)\left[1+c_B\right]}{2\left[\frac{1}{\lambda}-E(b_2)\right]} - \frac{E(b_2)-E(b_1)\left\{u_1^{(1)}(0)-0,5\left[E(b_2)+E(b_1)\right](1+c_B)\left[1+u_1^*(0)\right]\right\}}{\frac{1}{\lambda}+\left[E(b_2)-E(b_1)\right]\left[1+u_1^*(0)\right]}, \qquad \text{(II.3.38)}$$

$$\eta^*(\lambda) = \mathcal{W}_o\left[E(b_1)-E(b_2)\right]\left[1+u_1^*(0)\right] \qquad \text{(II.3.39)}$$

gdzie:

$$u_1^*(0) = \frac{l_o}{m_o} - \frac{H_2-H_1}{\sqrt{\Delta}}, \qquad \text{(II.3.40)}$$

$$u_1^{(1)}(0) = \frac{l_1 m_o - l_o m_1}{m_o^2} - \frac{(1-D\hat{s}_2)H_2 - (1-D\hat{s}_1)H_1}{\sqrt{\Delta}}, \qquad \text{(II.3.41)}$$

przy

$$\left.\begin{aligned} \Delta &= m_1^2 - 4m_o, \\ \hat{s}_{1,2} &= 0,5(-m_1 \pm \sqrt{\Delta}), \\ H_i &= \frac{l_i\hat{s}_i+l_o}{s_i}\,e^{D\hat{s}_i}, \quad i=1,2. \end{aligned}\right\} \qquad \text{(II.3.42)}$$

$$l_o = \frac{\lambda E(b_1)}{c_{21}}, \quad l_1 = \lambda, \quad m_o = \frac{1-\lambda E(b_1)}{c_{21}}, \quad m_1 = \frac{c_{11}-\lambda c_{21}}{c_{21}}. \qquad \text{(II.3.43)}$$

Przytoczone zależności określają wprost charakterystyki systemu i umożliwiają bezpośrednie określenie $\mathcal{W}(x)$ dla $x>D$ na mocy wzorów (II.3.2) oraz (II.3.3).

b) <u>Odstępy między chwilami napłynięcia kolejnych zgłoszeń posiadające rozkład Coxa II rzędu [WOLI 80B]</u>

Rozpatrzmy przypadek zgłoszeń nadchodzących w odstępach o rozkładzie Coxa drugiego rzędu, tj.:

$$\bar{a}(s) = \frac{1+d_1 s}{1+c_1 s+c_2 s^2}, \qquad \text{(II.3.44)}$$

przy czym parametry d_1, c_1, c_2 wyrażone są przez wartość oczekiwaną E(a) oraz współczynnik zmienności rozkładu C_A w sposób analogiczny do podanego w rozdziale I.
Jeśli chodzi o charakterystykę czasów obsługi, to przyjmiemy:

$$\left.\begin{aligned} B_o(x) = B_1(x) &= 1 - e^{-\mu_1 x} \\ B_2(x) &= 1 - e^{-\mu_2 x} \end{aligned}\right\} \qquad \text{(II.3.45)}$$

Można sprawdzić, że w tym przypadku zachodzi:

$$v_1^*(s,6) = \frac{l_1 s + l_o}{s^2 + m_1 s + m_o}, \qquad \text{(II.3.46)}$$

gdzie:

$$l_1 = 6, \quad l_o = \frac{\mu_1 6 c_2 - d_1\mu_1 - 1}{c_2}, \quad m_1 = \frac{\mu_1 c_2 - c_1}{c_2}, \quad m_o = \frac{1-E(a)\mu_1}{c_2}. \qquad \text{(II.3.47)}$$

Zauważmy, że (II.3.46) ma postać identyczną z (II.3.24) przy zastąpieniu (II.3.43) przez (II.3.46). Tak więc pozostają w mocy zależności (II.3.35) i (II.3.36) oraz (II.3.39), określające $v_1(t,\delta)$, $W(x)$ dla $x \leqslant D$ oraz $\eta^*(\lambda)$. W rozpatrywanym modelu wylicza się natomiast inaczej.

$$W_0 = \frac{E(s) - \frac{1}{\mu_2}}{c_1 - c_2\delta + \left[E(b_1) - E(b_2)\right]\left[1 + u^*(0,\delta)\right]}, \qquad \text{(II.3.48)}$$

$$E(w) = \frac{\left(\frac{1}{\mu_2}\right)^2 + d_1\frac{1}{\mu_1} - c_2}{E(s) - \frac{1}{\mu_2}} +$$

$$+W_0\,\frac{c_2+\left[E(b_2)-E(b_1)\right]\left\{u_1^{(1)}(0,\delta)-\left[E(b_2)+E(b_1)+d_1\right]\left[1+u_1^*(0,\delta)\right]\right\}}{E(s) - \frac{1}{\mu_2}} \qquad \text{(II.3.49)}$$

przy czym $u_1^*(0,\delta)$ oraz $u_1^{(1)}(0,\delta)$ wyznacza się z relacji (II.3.40) oraz (II.3.41) przy uwzględnieniu (II.3.42) oraz (II.3.47).

Dla kompletnego wyznaczenia charakterystyk badanego modelu należy jeszcze wyznaczyć stałą δ.

W rozpatrywanym przypadku układ równań (II.2.25) redukuje się do pojedynczego równania jednoznacznie określającego δ,

$$\delta = -\frac{\alpha_2(s_1)}{\alpha_1(s_1)}, \qquad \text{(II.3.50)}$$

gdzie:

$$\alpha_2(s_1) = -(1 + \frac{s_1}{\mu_1})(1 + \frac{s_1}{\mu_2}) +$$

$$+ (1-d_1s_1)\left\{(1 + \frac{s_1}{\mu_1}) - s_1(\frac{1}{\mu_2} - \frac{1}{\mu_1})\,\frac{d_1\mu_1 + 1}{c_2}\left[S_0 + \frac{S_2-S_1}{\sqrt{\Delta}}\right]\right\},$$

$$\alpha_1(s_1) = (1 + \frac{s_1}{\mu_1})(1 + \frac{s_1}{\mu_2}) +$$

$$+ \mu_1 c_2 s_1^2(1-d_1s_1)(\frac{1}{\mu_2} - \frac{1}{\mu_1})\left\{(s_1 + \frac{1}{\mu_1})S_0 + \frac{(\hat{s}_2 + \frac{1}{\mu_1})S_2-(\hat{s}_1 + \frac{1}{\mu_1})S_1}{\sqrt{\Delta}}\right\}$$

przy:

$$S_0 = \left[(s_1 - \hat{s}_1)(s_1 - \hat{s}_2)\right]^{-1},$$

$$S_i = \frac{e^{-D(s_1-\hat{s}_i)}}{s_1-\hat{s}_i}, \quad i=1,2,$$

natomiast s_1 jest dodatnim pierwiastkiem równania (II.2.24), które [COHE 69] posiada w rozpatrywanym przypadku dwa pierwiastki, dla których $\mathrm{Re}(s) \geqslant 0$, w tym jeden równy zeru

$$s_1 = \frac{\mu_2}{2c_2}\left\{\frac{c_1}{\mu_2} - c_2 + \sqrt{(\frac{c_1}{\mu_2} - c_2)^2 - 4\,\frac{c_2}{\mu_2}\left[\frac{1}{\mu_2} - E(s)\right]}\right\}$$

Na podstawie zamieszczonych tu rozwiązań przypadków szczególnych modelu podstawowego widać, że jego analiza może być dość pracochłonna nawet wówczas, gdy nie zachodzi konieczność odwoływania się do obliczeń numerycznych.

c) Analiza pewnego systemu ze stratami

Dotychczasowe rozważania ograniczyliśmy do systemów z czasem obsługi zależnym od czasu oczekiwania i to bez ograniczania tego ostatniego.

Warto zwrócić uwagę na fakt, iż przeanalizowany model podstawowy obejmuje jako szczególne przypadki również:

- systemy typu $K_n/G/1$ z rozkładem czasu obsługi pierwszego zgłoszenia w każdym okresie czynnym różnym od pozostałych (przy przyjęciu $J = 1$), a więc istotny (jak to podkreślono w rozdziale I) dla wielu zastosowań przypadek niekonwencjonalnego systemu obsługi;
- systemy typu $K_n/G/1$ z ograniczonym w sposób deterministyczny czasem oczekiwania na rozpoczęcie obsługi (i być może różnym od pozostałych czasem obsługi pierwszego zgłoszenia w okresie czynnym), przy przyjęciu $J = 2$ oraz $B_2(x) = H(x)$;
- systemy WAD z ograniczonym w sposób deterministyczny czasem oczekiwania na rozpoczęcie obsługi, po upływie którego zgłoszenie jest tracone (przy $B_J(x) = H(x)$).

Poniżej wyznaczymy rozkład czasu oczekiwania w pewnym systemie WAD z ograniczonym czasem oczekiwania na rozpoczęcie obsługi. Przypadek ten będzie podstawą do analizy w rozdziale III pewnych bardziej złożonych regulaminów obsługi.

Rozpatrzmy system, w którym zgłoszenia nadchodzą strumieniem Poissona o intensywności λ, natomiast czas obsługi zależy od czasu oczekiwania w sposób opisany dystrybuantą.

$$B(x/w) = \begin{cases} \mathcal{K}(x,m,\mu_x) & \text{dla} \quad w = 0 \\ 1 - e^{-\mu_1 x} & \text{dla} \quad 0 < w \leq D \\ 1 - e^{-\mu_2 x} & \text{dla} \quad 0 < w \leq \tau \\ \mathbf{H}(x) & \text{dla} \quad w > \tau \end{cases} \qquad (II.3.51)$$

gdzie:

$$\mathcal{K}(x,m,\mu_x) = m\frac{\mu_1}{s+\mu_1} + (1-m)\left\{\frac{\mu_1}{s+\mu_1}\frac{\mu_x}{s+\mu_x}\left[1-e^{-\mu_x D}e^{-sD}\right] + \right.$$

$$\left. + \frac{\mu_2\mu_x}{(s+\mu_2)(s+\mu_x)}\left[e^{-\mu_x D}e^{-sD} - e^{-\mu_x\tau}e^{-s\tau}\right] + \frac{\mu_x}{s+\mu_x}e^{-\mu_x\tau}e^{-s\tau}\right\}, \quad 0 \leq m \leq 1;$$

$$E(k) = \frac{m}{\mu_1} + (1-m)\left[(1-e^{-\mu_x D})\left(\frac{1}{\mu_x} - \frac{1}{\mu_1}\right) + (e^{-\mu_x D} - e^{-\mu_x\tau})\left(\frac{1}{\mu_2} + \frac{1}{\mu_x}\right) + e^{-\mu_x\tau}\frac{1}{\mu_x}\right].$$

Korzystając z zależności (II.3.2)-(II.3.4) otrzymujemy:

$$v_1^*(s) = \frac{\lambda}{s+\mu_1-\lambda} + \frac{\lambda\mu_1(1-m)}{\mu_1-\mu_x-\lambda}\left[\frac{1}{s+\mu_1-\lambda} - \frac{1}{s+\mu_x}\right] + \varkappa_1(s)e^{-sD} + \varkappa_2(s)e^{-s\tau},$$

dla $\lambda \neq \mu_1$

$\lambda \neq \mu_1 - \mu_x$,

gdzie: $\varkappa_1$, $\varkappa_2$ - pewne funkcje.

$$v_1(t) = \lambda e^{-(\mu_1-\lambda)t} + \frac{\lambda\mu_1(1-m)}{\mu_1-\mu_x-\lambda}\left[e^{-\mu_x t} - e^{-(\mu_1-\lambda)t}\right] \quad \text{dla} \quad t \leq D. \qquad (II.3.52)$$

$$u_1^*(s) = \frac{1}{s+\mu_1-\lambda}\left[1-e^{-(\mu_1+s-\lambda)D}\right]\left[\lambda - \frac{\lambda\mu_1(1-m)}{\mu_1-\mu_x-\lambda}\right] + \frac{1}{s+\mu_x}\left[1-e^{-(\mu_x+s)D}\right]\frac{\lambda\mu_1(1-m)}{\mu_1-\mu_x-\lambda}.$$

(II.3.53)

$$v_2^*(s) = R_1\frac{1}{s+\mu_1-\lambda} + R_2\frac{1}{s+\mu_x} + R_3\frac{1}{s+\mu_2-\lambda}e^{-(\mu_1-\lambda)D}e^{-sD} +$$

$$+ R_4\frac{1}{s+\mu_1}e^{-(\mu_1-\lambda)D}e^{-sD} - R_1\frac{1}{s+\mu_1-\lambda}e^{-(\mu_1-\lambda)D}d^{-sD} +$$

$$+ R_5\frac{1}{s+\mu_2-\lambda}e^{-\mu_x D}e^{-sD} + R_5\frac{1}{s+\mu_1}e^{-\mu_x D} + R_5\frac{1}{s+\mu_x}e^{-\mu_x D}e^{-sD} + \varkappa_3(s)e^{-s\tau},$$

gdzie:

$$R_1 = \left[1 - \frac{\mu_1(1-m)}{\mu_1-\mu_x-\lambda}\right]\lambda, \quad R_2 = \frac{\lambda\mu_1(1-m)}{\mu_1-\mu_x-\lambda},$$

$$R_3 = \frac{\lambda}{\lambda+\mu_1-\mu_2}R_1, \qquad R_4 = \frac{\mu_1-\mu_2}{\lambda+\mu_1-\mu_2}R_1,$$

$$R_5 = \lambda(\mu_1-\mu_2)(1-m)\frac{\mu_x(\mu_1-\mu_x-\lambda)+\lambda\mu_1}{(\lambda-\mu_2+\mu_1)(\lambda-\mu_2+\mu_x)(\mu_1-\mu_x-\lambda)},$$

dla $\lambda \neq \mu_1$

$\lambda \neq \mu_1 - \mu_x$

$\lambda \neq \mu_2 - \mu_1$

$\lambda \neq \mu_2 - \mu_x$

$\lambda \neq \mu_2$.

$$v_2(t) = R_1e^{-(\mu_1-\lambda)t} + R_2e^{-\mu_x t} + R_3e^{-(\mu_2-\lambda)t}e^{-(\mu_1-\mu_2)D}\mathbf{H}(t-D) +$$

$$+ R_4e^{-\mu_1 t}e^{-\lambda D}\mathbf{H}(t-D) - R_1e^{-(\mu_1-\lambda)t}\mathbf{H}(t-D) +$$

$$+ R_5e^{-(\mu_2-\lambda)t}e^{-(\lambda+\mu_x-\mu_2)D}\mathbf{H}(t-D) + R_5e^{-\mu_1 t}e^{-(\mu_x-\mu_1)D}\mathbf{H}(t-D) -$$

$$- R_5e^{-\mu_x t}\mathbf{H}(t-D) \quad \text{dla} \quad t < \tau \qquad (II.3.54)$$

$$u_2^*(s) = \frac{R_2}{s+\mu_x}\left[e^{-(\mu_x+s)D} - e^{-(\mu_x+s)\tau}\right] +$$

$$+ \frac{R_3}{s+\mu_2-\lambda}\left[e^{-(\mu_2-\lambda+s)D} - e^{-(\mu_2-\lambda+s)\tau}\right]e^{-(\mu_1-\mu_2)D} +$$

$$+ R_4\frac{1}{s+\mu_1}\left[e^{-(s+\mu_1)D} - e^{-(s+\mu_1)\tau}\right]e^{\lambda D} +$$

$$+ R_5\frac{1}{s+\mu_2-\lambda}\left[e^{-(\mu_2-\lambda+s)D} - e^{-(\mu_2-\lambda+s)\tau}\right]e^{-(\mu_x-\mu_2+\lambda)D} +$$

$$+ R_5 \frac{1}{s+\mu_1}\left[e^{-(s+\mu_1)D} - e^{-(s+\mu_1)t}\right] e^{-(\mu_x-\mu_1)D} -$$

$$- R_5 \frac{1}{s+\mu_x}\left[e^{-(s+\mu_x)D} - e^{-(s+\mu_x)t}\right]. \qquad \text{(II.3.55)}$$

$$v_3(t) = v_2(t) \quad \text{dla} \quad t < \tau \qquad \text{(II.3.56)}$$

Oznaczmy przez $w^{\Phi}(x)$ dystrybuantę czasu oczekiwania, gdzie $\Phi(x)=K(x,m,\mu_x)$ będzie określało, przy jakiej postaci dystrybuanty $K(x,m,\mu_x)$ była prowadzona analiza modelu, łatwo widać, że:

$$w^{\Phi}(0) = \left[1 + \lambda E(k) + \lambda u_1^*(0)/\mu_1 + \lambda u_2^*(0)/\mu_2\right]^{-1}, \qquad \text{(II.3.57)}$$

$$w^{\Phi}(t) = w^{\Phi}(0)\left[1 + \int_0^t v_2(t)dt\right] \quad \text{dla} \quad t < \tau$$

$$w^{\Phi}(\tau) = w^{\Phi}(0)\left[1 + u_1^*(0) + u_2^*(0)\right],$$

$$w^{\Phi}(D) = w^{\Phi}(0)\left[1 + u_1^*(0)\right].$$

II.4. Inne modele

Niezależnie od omówionych w punkcie II.2 różnych wariantów modelu podstawowego rozpatrywane były jeszcze dwa specyficzne przypadki systemów z czasem obsługi zależnym od czasu oczekiwania na jej rozpoczęcie i naturalnym regulaminem obsługi, które tu pokrótce przedstawimy.

a) Liniowa zależność czasu obsługi od czasu oczekiwania

W pracy [MUDR 61] analizowano jednokanałowy system WAD ze zgłoszeniami napływającymi strumieniem Poissona (o intensywności λ) oraz zależnością czasu obsługi od czasu oczekiwania określoną jako:

$$B(x/w) = H(x-\psi(w)), \quad \psi(w) = \vartheta - cw, \quad c > 0. \qquad \text{(II.4.1)}$$

Czas obsługi był więc deterministyczną, malejącą, funkcją czasu oczekiwania, przy czym dodatkowo czas oczekiwania ograniczony był do wartości $w_{max} = \frac{\vartheta}{1-c}$, po przekroczeniu której zgłoszenia opuszczały kolejkę, rezygnując z dalszego oczekiwania (były tracone). Uzyskano wyrażenie na

jednostkowy czas obsługi zgłoszenia $\eta^*(\lambda)$ oraz prawdopodobieństwo utraty zgłoszenia na skutek nadmiernie długiego oczekiwania P_{str},

$$\left.\begin{aligned} \eta^*(\lambda) &= \vartheta_m + \lambda^{-1} e^{-\lambda w_{max}} + c\lambda^{-1}(1 - e^{-\lambda w_{max}}) \\ P_{str} &= 1 - \left[\lambda\, \eta^*(\lambda)\right]^{-1} \quad \text{gdzie } \vartheta_m = \vartheta - cw_{max} \end{aligned}\right\} \qquad \text{(II.4.2)}$$

b) Wielokanałowy system z ograniczoną wymiarowością źródeł zgłoszeń

W pracy [SWEN 59] rozpatrzono n-kanałowy system masowej obsługi w którym zgłoszenia generowane są przez K wymiarowe źródło zgłoszeń takie, że czas pobytu każdego ze zgłoszeń w źródle jest zmienną losową o rozkładzie wykładniczym z parametrem λ (patrz wzór (I.1.18)).

Przyjęto następującą postać zależności czasu obsługi od czasu oczekiwania

$$\bar{b}(s/w) = \frac{\mu}{s+\mu} \exp\left[-\lambda_d w\left(1 - \frac{\mu_d}{s+\mu_d}\right)\right], \qquad \text{(II.4.3)}$$

gdzie:

$$\bar{b}(s/w) = \int_0^{\infty} e^{-sx} d_x B(x/w).$$

Taka postać zależności czasu obsługi od czasu oczekiwania posiada następującą interpretację: Proces obsługi zgłoszenia, którego czas oczekiwania jest równy zeru, składa się z jednej fazy o czasie będącym zmienną losową b_1, posiadającą rozkład wykładniczy z parametrem μ. Oczekiwanie na rozpoczęcie obsługi powoduje konieczność rozbudowy tej obsługi o następne fazy, przy czym liczba tych faz jest zmienną losową o rozkładzie Poissona z parametrem $(\lambda_d w)$, natomiast długość każdej z kolejnych faz jest zmienną losową o rozkładzie wykładniczym z parametrem μ_d. Łączne rozpatrzenie czasu obsługi prowadzi w takim modelu do zależności przedstawionej wzorem (II.4.3).

Analizę tak zdefiniowanego modelu przeprowadzono metodą przybliżoną. Założono mianowicie, że analizowany będzie, jako aproksymacja systemu wyjściowego, system zastępczy, w którym czas obsługi zgłoszenia ma rozkład wykładniczy o parametrze μ_i zależnym od liczby zgłoszeń "i" aktualnie znajdujących się poza źródłem (tzn. obsługiwanych lub oczekujących w kolejce). Przyjęto, że zależność μ_i od "i" ma postać:

$$\mu_i = \begin{cases} \mu & \text{dla} \quad i \leqslant n \\ \mu - \frac{i-n}{n} \circledcirc \lambda_d & \text{dla} \quad \mathcal{K} \geqslant i > n \end{cases} \tag{II.4.4}$$

gdzie: $\circledcirc = \frac{\mu_d}{\mu}$.

Przy zapewniającym stabilność systemu założeniu, że

$$n\mu > (\mathcal{K} - n) \circledcirc \lambda_d,$$

wyznaczono prawdopodobieństwo π_i tego, że w stacjonarnym stanie granicznym poza źródłem przebywa "i" zgłoszeń:

$$\pi_o = \left[\sum_{i=1}^{n-1} (\frac{\lambda}{\mu})^i (\frac{\mathcal{K}}{i}) + \sum_{i=n}^{\mathcal{K}} \gamma_i \right]^{-1},$$

$$\pi_i = p_o (\frac{\lambda}{\mu})^i (\frac{\mathcal{K}}{i}) \quad \text{dla} \quad i < n, \tag{II.4.5}$$

$$\pi_i = p_o \gamma_i \quad \text{dla} \quad \mathcal{K} > i \geqslant n,$$

gdzie:

$$\gamma_i = (\frac{\lambda}{\mu})^i \frac{(\mathcal{K})_i}{n! n^{i-n}} \left\{ \prod_{k=0}^{i-1} (1 - \frac{k \circledcirc \lambda_d}{n\mu}) \right\}^{-1},$$

$$(\mathcal{K})_i = \frac{K!}{(K-i)!}.$$

Na tej podstawie przy upraszczającym założeniu $\lambda = \lambda_d$ oraz $\mu = \mu_d$ wyznaczono szereg charakterystyk systemu, między innymi średni czas oczekiwania zgłoszeń w kolejce, wyrażający się wzorem:

$$E(w) = \frac{\mathcal{K} - n\eta(\lambda)}{\mathcal{K}\left[1 - \eta(\lambda)\right]} \frac{1}{n\mu}, \tag{II.4.6}$$

gdzie:

$$\eta(\lambda) = \frac{\mathcal{K}\lambda}{\mu + \lambda} \tag{II.4.7}$$

i porównano je z wynikami badań symulacyjnych, uzyskując dobrą zgodność.

Ze względu na użycie bardzo krótkich serii zdarzeń przy symulacji nie można tu jednak mówić o oszacowaniu błędu metody.

c) Przypadek dyskretnych rozkładów zmiennych losowych definiujących system WAD

W pracy [SZNE 68] rozpatrzono przypadek, gdy zmienne losowe s^i, w^i, b_w^i są zmiennymi dyskretnymi przyjmującymi wartości całkowitoliczbowe nieujemne (a więc identyczny z założeniami, przy których w [CALL 73] dowodzono warunku stabilności). Przyjęto, że oś czasu podzielona jest na odcinki czasu będące całkowitą wielokrotnością kwantu Δt. Założono, że w odcinku Δt do systemu napływa co najwyżej jedno zgłoszenie. Niech α_i oznacza prawdopodobieństwo tego, że odstęp czasu między napłynięciem kolejnych zgłoszeń wynosi i kwantów ($\alpha_o = 0$), $q_k(i)$ oznacza prawdopodobieństwo tego, że czas oczekiwania zgłoszenia, które napłynęło w chwili $i\Delta t$ wyniesie $k\Delta t$ oraz β_{ij} oznacza prawdopodobieństwo tego, że czas obsługi wynosi $i\Delta t$ pod warunkiem, że czas oczekiwania wynosi $j\Delta t$.

Graniczny rozkład czasu oczekiwania q_k, $k = 0,1,\ldots$, można znaleźć na podstawie iteracyjnej zależności:

$$\left[\sum_{j=2}^{\infty} \alpha_j + \alpha_1 \beta_{0,k+1} \right] q_{k+1} +$$

$$+ (\alpha_1 \beta_{1k} - 1) q_k + \alpha_1 \sum_{j=0}^{k-1} \beta_{k+1-j,j} q_j = 0; \quad k = 1,2,\ldots, \tag{II.4.8}$$

przy czym q_o wyznacza się z warunku normalizacji

$$\sum_{k=1}^{\infty} q_k = 1. \tag{II.4.9}$$

W szczególności wykazano, że jeśli założyć:

i)

$$\left.\begin{aligned} &\alpha_o = 0, \\ &\alpha_i = \alpha(1-\alpha)^{i-1} \quad i = 1,2,\ldots \quad 0 < \alpha < 1. \end{aligned}\right\} \tag{II.4.10}$$

ii)

$$\left.\begin{aligned} &\beta_{oj} = 1 - \beta^j, \\ &\beta_{ij} = ip^{2(j+1)}(1 - p^{j+1})^{i-1}\beta^j, \\ &i \geqslant 1, \quad j \geqslant 0; \\ &0 < \beta < 1, \quad 0 < p < 1; \end{aligned}\right\} \tag{II.4.11}$$

to znaczy scharakteryzować strumień zgłoszeń jako dyskretną analogię rozkładu wykładniczego, natomiast czas obsługi jako dyskretną analogię rozkładu Erlanga drugiego rzędu, wówczas rozwiązanie graniczne ma postać:

$$q_{k+1} = \frac{1}{\alpha\beta^{k+1} - 1}\left\{\left[\alpha p^{2(k+1)}\beta^k - 1\right]q_k + \right.$$

$$\left. + \alpha\sum_{j=0}^{k-1}(k+1-j)p^{2(j+1)}(1-p^{j+1})^{k-j}\beta^i q_i\right\}, \quad k=1,2,\ldots$$

gdzie: $q_{-1} = 0$, natomiast q_0 wyznacza się z warunku normalizacji. Zilustrowano również rozwiązanie dla przykładowych parametrów systemu.

II.5. Niektóre własności systemów WAD z naturalnym regulaminem obsługi

Korzystając z przedstawionych poprzednio analitycznych rozwiązań, wyznaczających charakterystyki szerokiej klasy systemów obsługi z czasem obsługi zależnym od czasu oczekiwania oraz regulaminem naturalnym, przedyskutujemy obecnie niektóre własności takich systemów.

W szczególności postaramy się odpowiedzieć na pytanie, czy i na ile charakterystyki systemów WAD, a także ich wrażliwość na poszczególne parametry systemu, różnią się od podobnych relacji w systemach konwencjonalnych. Zastanowimy się też, czy i na ile dokładnie można byłoby wnioskować o pewnych charakterystykach systemu WAD na podstawie analizy odpowiednio dobranego systemu konwencjonalnego.

a) Porównanie własności prostego systemu WAD i systemu M/G/1 [WOLI 79]

Rozpatrzmy system WAD z regularnym rozkładem czasu obsługi i zgłoszeniami napływającymi strumieniem Poissona przeanalizowany w punkcie II.3.a2, którego charakterystyki zadane są zależnościami (II.3.18)-(II.3.24). Tak więc czas obsługi i-tego zgłoszenia $b^i_{w^i}$ określony jest jako:

$$b^i_{w^i} = \begin{cases} B_s & \text{dla } w^i \leq 0, \\ B_L & \text{dla } w^i > 0 \end{cases} \quad i = 0,1,\ldots \qquad \text{(II.5.1)}$$

przy czym założymy dodatkowo, że $B_L > B_s$.
System ten będziemy w dalszym ciągu nazywać systemem bazowym.
Niech będzie dana rodzina systemów M/G/1 takich, że:

i) zgłoszenia napływają strumieniem Poissona o intensywności λ identycznej jak w badanym systemie bazowym,

ii) czasy obsługi zgłoszeń są niezależnymi zmiennymi losowymi o rozkładzie zależnym od wartości λ, posiadającym dystrybuantę

$$B_\lambda(x) = \mathcal{W}_\lambda(D)\,\mathbf{H}(x - B_s) + \left[1 - \mathcal{W}_\lambda(D)\right]\mathbf{H}(x - B_L), \qquad \text{(II.5.2)}$$

gdzie: $\mathcal{W}_\lambda(D) = \mathcal{W}(D)$ jest prawdopodobieństwem wykonania w systemie bazowym obsługi o czasie trwania B_s, zadanym wzorem (II.3.23),
które nazywać będziemy dalej systemami zastępczymi.

Zauważmy, że z (II.4.2) wynika równość wartości oczekiwanej czasu obsługi zgłoszenia w systemie zastępczym $E(b_\lambda)$ oraz jednostkowego czasu obsługi zgłoszenia w systemie bazowym, wyznaczonego przez zależność (II.3.16). Tak więc dla dowolnej intensywności źródła zgłoszeń zdefiniowaliśmy system zastępczy identyczny z systemem bazowym pod względem sposobu wykorzystania stanowiska obsługi, jako że rozkłady czasów obsługi w obu tych przypadkach są identyczne.

Porównanie parametrów tych systemów pozwoli stwierdzić, czy popełniony zostałby duży błąd, gdyby przy badaniu obiektu rzeczywistego o charakterze systemu WAD zamiast analizować jako jego model system bazowy, wziąć pod uwagę wyznaczony na przykład eksperymentalnie, drogą pomiarów na obiekcie, rozkład $B_{\lambda_0}(x)$ (przy pewnej intensywności strumienia zgłoszeń λ_0) i wnioskować o własnościach obiektu na podstawie własności systemu zastępczego.

Wprowadźmy oznaczenia:

$V(\lambda) = E(w)$ - średni czas oczekiwania w systemie bazowym (określony wzorem (II.3.24)),

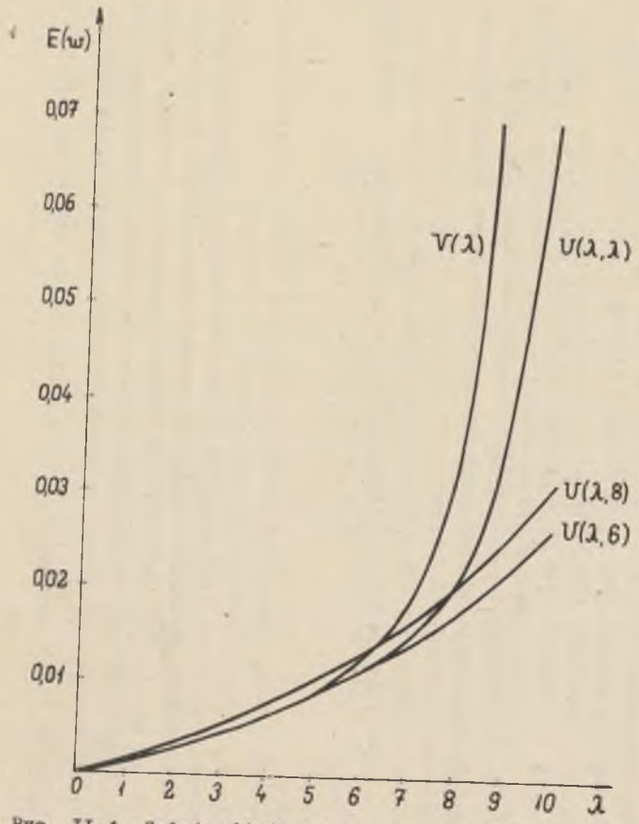
$U(\lambda, \lambda_0)$ - średni czas oczekiwania w systemie zastępczym, dla którego rozkład czasu obsługi został określony równaniem (II.4.2) przy $\lambda = \lambda_0$.

Korzystając z elementarnych zależności dla systemów M/G/1 [KLEI 75] mamy:

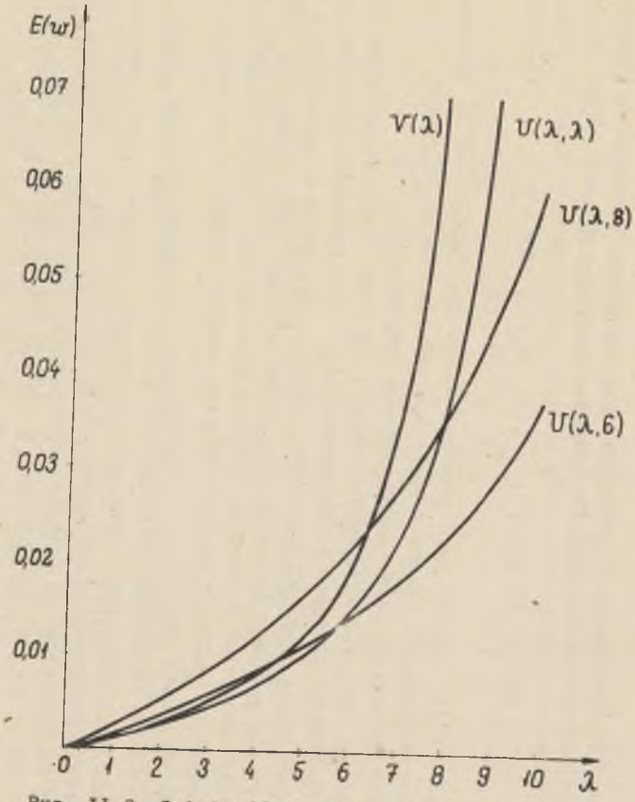
$$U(\lambda, \lambda_0) = \frac{\lambda\left[\mathcal{W}_{\lambda_0}(D)(B_s^2 - B_L^2) + B_L^2\right]}{2\left\{1 - \lambda\left[\mathcal{W}_{\lambda_0}(D)(B_s - B_L) + B_L\right]\right\}}. \qquad \text{(II.5.3)}$$

Przykładowe wykresy funkcji $V(\lambda)$ oraz $U(\lambda,\lambda_0)$ dla dwóch zestawów parametrów B_L, B_s, D przedstawiono na rysunkach II.1 oraz II.2.

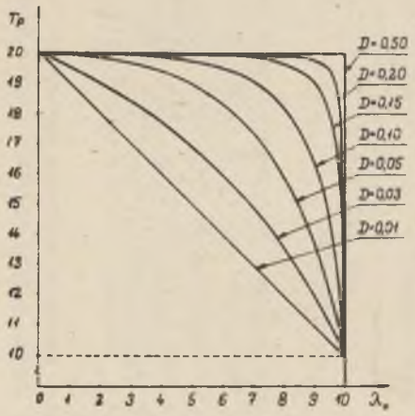
Porównanie wartości $V(\lambda)$ oraz $U(\lambda,\lambda)$ pozwala ocenić różnicę średnich czasów oczekiwania zgłoszeń w systemie bazowym i zastępczym przy założeniu, iż dla każdej intensywności λ strumienia zgłoszeń rozkłady czasów obsługi w obu tych systemach są jednakowe. Jak widzimy, uzależnienie czasu obsługi od czasu oczekiwania w systemie bazowym powoduje, iż $\bigvee_\lambda[V(\lambda) - U(\lambda,\lambda) > 0]$, przy czym wartość tej różnicy silnie rośnie ze wzrostem λ.



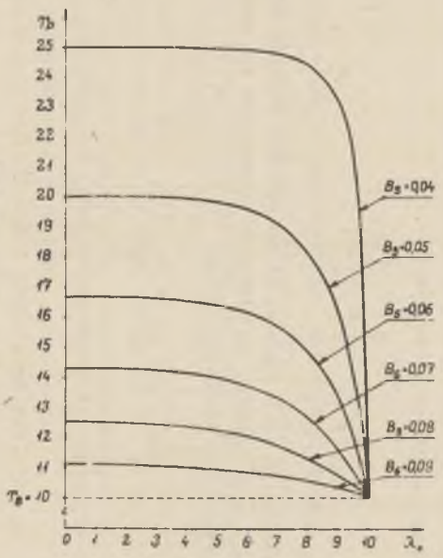
Rys. II.1. Zależność średniego czasu oczekiwania w systemie bazowym i zastępczym przy $B_L = 0.1$, $B_S = 0.05$, $D = 0.1$ od intensywności strumienia zgłoszeń λ



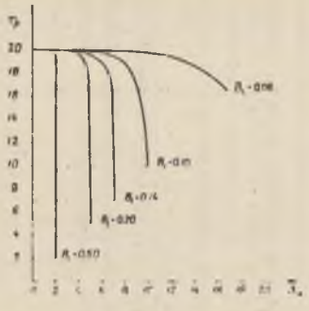
Rys. II.2. Zależność średniego czasu oczekiwania w systemie bazowym i zastępczym przy $B_L = 0.1$, $B_S = 0.05$, $D = 0,05$ od intensywności strumienia zgłoszeń λ



Rys. II.3. Wydajność graniczna systemu zastępczego przy $B_L = 0.1$, $B_S = 0.05$ w funkcji λ_o i D



Rys. II.4. Wydajność graniczna systemu zastępczego przy $B_L = 0.1$, $D = 0.01$ w funkcji λ_o i B_S



Rys. II.5. Wydajność graniczna systemu zastępczego przy $B_S = 0.05$, $D = 0.1$ w funkcji λ_o i B_L

W zamieszczonych przykładach wartość tej różnicy może nawet kilkakrotnie przewyższać wartość $U(\lambda,\lambda)$.

Krzywe $U(\lambda,\lambda_o)$ przy $\lambda \neq \lambda_o$ przedstawiają prognozę czasu oczekiwania dla różnych λ, wyznaczoną przy użyciu systemu zastępczego na podstawie znajomości rozkładu czasu obsługi $B_{\lambda_o}(x)$. Błąd prognozy $V(\lambda) - U(\lambda,\lambda_o)$ jest ujemny dla małych λ i ze wzrostem λ rośnie (maleje pod względem wartości bezwzględnej) do zera w punkcie $\lambda_x < \lambda_o$, przyjmując następnie coraz większe wartości dodatnie.

Przejdźmy obecnie do przedyskutowania najistotniejszego z punktu widzenia użytkownika systemu parametru, jakim jest wydajność graniczna. W przypadku systemu bazowego wynosi ona (patrz punkt II.1):

$$T_B = \left[B_L\right]^{-1}. \tag{II.5.4}$$

Wydajność graniczna systemu zastępczego, definiowanego dla intensywności strumienia zgłoszeń λ_o, wynosi natomiast:

$$T_z(\lambda_o) = \left\{ E(b_{\lambda_o}) \right\}^{-1}. \tag{II.5.5}$$

Na rysunkach II.3, II.4 i II.5 przedstawiono wykresy reprezentujące przebieg funkcji $T_z(\lambda_o)$ przy różnych wartościach parametrów systemu bazowego. Jak widać z tych wykresów, zachodzi zawsze

$$R(\lambda_o) = T_z(\lambda_o) - T_B > 0, \tag{II.5.6}$$

przy czym różnica ta maleje ze wzrostem wartości λ_o, jako że

$$\left.\begin{aligned} &\lim_{\lambda_o \to 0} R(\lambda_o) = \frac{1}{B_s} - \frac{1}{B_L}, \\ &\lim_{\lambda_o \to T_B} R(\lambda_o) = 0. \end{aligned}\right\} \tag{II.5.7}$$

Funkcja $R(\lambda_o)$ ściśle malejąca jest jednak w praktycznie interesującym zakresie zmienności parametrów B_L, B_s, D silnie nieliniowa. Zachowuje ona przy wzrastających wartościach λ_o dość długo prawie stałą wartość, aby gwałtownie zmaleć przy λ_o bliskich T_B. Jeśli chodzi o wpływ parametrów systemu bazowego na wartość błędu $R(\lambda_o)$, to z załączonych rysunków widać, że dla ustalonego λ_o wartość R rośnie ze wzrostem D oraz B_L, natomiast maleje ze wzrostem B_s.

Zastąpienie systemu bazowego systemem zastępczym typu M/G/1 prowadziłoby więc z reguły do bardzo dużych błędów w ocenie wydajności granicznej.

Przykładowe wartości błędu względnego w ocenie wydajności granicznej dla przypadku $B_L = 0.1$, $B_s = 0.05$, $D = 0.1$ zamieszczono w tabeli II.2.

Tabela II.2

Przykłady wartości błędu w ocenie wydajności granicznej

λ_o	2	4	6	8
λ_o/T_B [%]	20	40	60	80
$R(\lambda_o)/T_B$ [%]	99.99	99.48	97.26	87.76
λ_o/T_z [%]	10.0	20.0	30.4	42.6

Zwróćmy uwagę na jeszcze jedną istotną prawidłowość wynikającą z danych zawartych w tabeli II.2.

Otóż w systemach konwencjonalnych zachodziła równość

$$\eta_z(\lambda) = \frac{\lambda}{\lambda_o}, \tag{II.5.8}$$

a więc pomiar stopnia wykorzystania stanowiska obsługi pozwalał wprost określić, na ile możliwe jest zwiększenie intensywności strumienia zgłoszeń.

W przypadku rozpatrywanego systemu bazowego, ze względu na wspomnianą nieliniowość funkcji $\eta^*_B(\lambda)$, takie rozumowanie prowadziłoby do dużych błędów.

W szczególności więc, jeśli przy $\lambda_o/T_B = 0.8$ stwierdzamy, że stopień wykorzystania stanowiska obsługi wynosi 42.6%, to nie można stąd wnioskować, że możliwe jest znaczne, np. dwukrotne zwiększenie intensywności strumienia zgłoszeń, gdyż w istocie już wzrost λ_o o 25% wartości początkowej spowoduje przekroczenie wydajności granicznej.

Wskazane wyżej prawidłowości można uogólnić również na inne systemy WAD (bez strat zgłoszeń) mieszczące się w ramach modelu podstawowego.

Gdyby zachodziło w modelu podstawowym:

$$E(b_i) < E(b_{i+1}), \quad i=1,2,\ldots,J-1,$$

wówczas $\eta^*(\lambda)$ byłaby niemalejącą funkcją λ i wnioski wyżej sformułowane pozostałyby (w sensie jakościowym) słuszne.

Tak więc w szczególności w systemach tych należy z dużą ostrożnością odnosić się do możliwości zwiększenia ich obciążenia w przypadku stwierdzenia niskiego stopnia wykorzystania stanowiska obsługi, które - jak to wskazano - mogą być złudne.

b) Wpływ współczynnika zmienności rozkładu czasów obsługi i odstępów między napływięciem zgłoszeń na charakterystykę systemów WAD [WOLI 80B]

Rozpatrując systemy masowej obsługi dążymy często do ograniczenia się, przy charakteryzowaniu rozkładów zmiennych losowych definiujących te systemy, do podania możliwie małej liczby ich momentów.

W szczególności podejmuje się próby ograniczenia opisu tych zmiennych do wartości oczekiwanych i przyjęcia rozkładów wykładniczych (co zasadniczo upraszcza analizę - w przypadku systemów WAD jest to wyraźnie widoczne na przykładzie systemu omawianego w punkcie II.3.a1 - ale z reguły prowadzi do sporych błędów) bądź też dwóch pierwszych momentów (w wielu przypadkach takie przybliżenie jest wystarczające; faktowi temu zawdzięcza dużą popularność metoda aproksymacji dyfuzyjnej [KLEI 76]).

Jak to już wspomniano w rozdziale I, w przypadku systemów konwencjonalnych prowadzone były prace (np. [GROS 75]) nad wpływem odejścia od założeń o rozkładzie wykładniczym na charakterystyki systemu. W tym miejscu przedyskutujemy wpływ współczynnika zmienności rozkładów definiujących system masowej obsługi z czasem obsługi zależnym od czasu oczekiwania na jego charakterystyki.

Zarówno rozkład odstępów między zgłoszeniami, jak też rozkłady czasów obsługi będziemy charakteryzować przez podanie wartości oczekiwanej oraz współczynnika zmienności rozkładu, definiowanego (patrz I.1.2) jako stosunek wariancji rozkładu do kwadratu jego wartości oczekiwanej. Oba te parametry mogą przyjmować wartości z przedziału $[0, \infty)$.

Tak więc dla celów eksperymentu, który miałby na celu zbadanie wpływu postaci rozkładów na parametry systemu masowej obsługi, należałoby wybrać taką rodzinę rozkładów, w których możliwa byłaby zmiana zarówno wartości średniej, jak i współczynnika zmienności rozkładu w wyżej podanym przedziale. Często do takich eksperymentów wykorzystuje się rozkłady Coxa, tj. takie, dla których transformata Laplace'a - Stieltjesa dystrybuanty jest funkcją wymierną. Pokrywają one pełny zakres zmienności obu interesujących nas parametrów. Uwzględnienie małych - bliskich zera - wartości współczynnika zmienności rozkładu wymaga jednak użycia rozkładów Coxa wysokiego rzędu. Ponieważ wykorzystanie rozkładów Coxa wysokiego rzędu prowadzi do bardzo złożonych obliczeń, więc często ogranicza się w badaniach eksperymentalnych do rozkładów Coxa rzędu drugiego, dla których wartość oczekiwana może przyjmować dowolne wartości nieujemne, natomiast współczynnik zmienności może przyjmować wartości z przedziału $[0.5, \infty)$, co pokrywa większość sytuacji praktycznie interesujących.

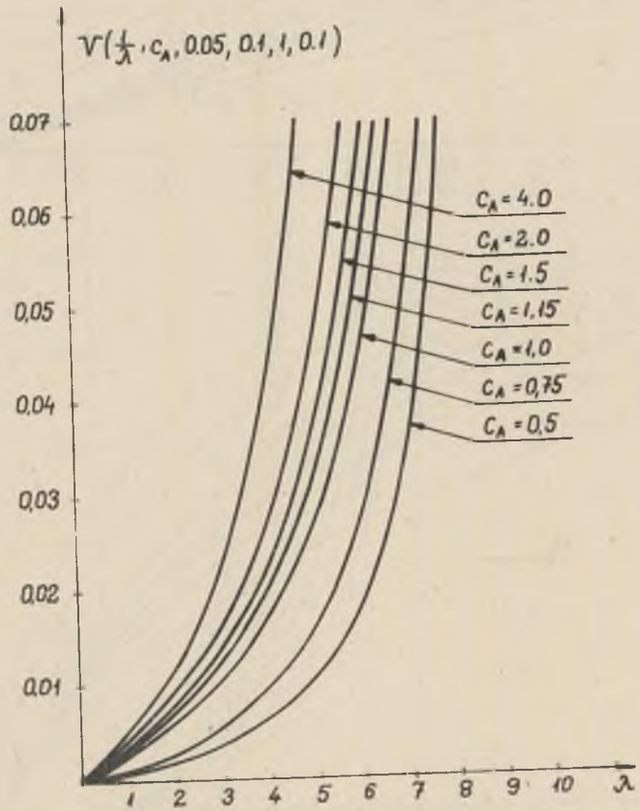
W dalszym ciągu tego rozdziału wykorzystywać będziemy zarówno dla opisu odstępów pomiędzy napływięciem kolejnych zgłoszeń, jak i dla opisu czasu obsługi szczególną postać rozkładu Coxa drugiego rzędu, dla której dystrybuanta R(x) posiada transformatę Laplace'a - Stieltjesa $\bar{r}(s)$:

$$\bar{r}(s) = \frac{1 + \alpha_0 s}{1 + \alpha_1 s + \alpha_2 s^2}$$

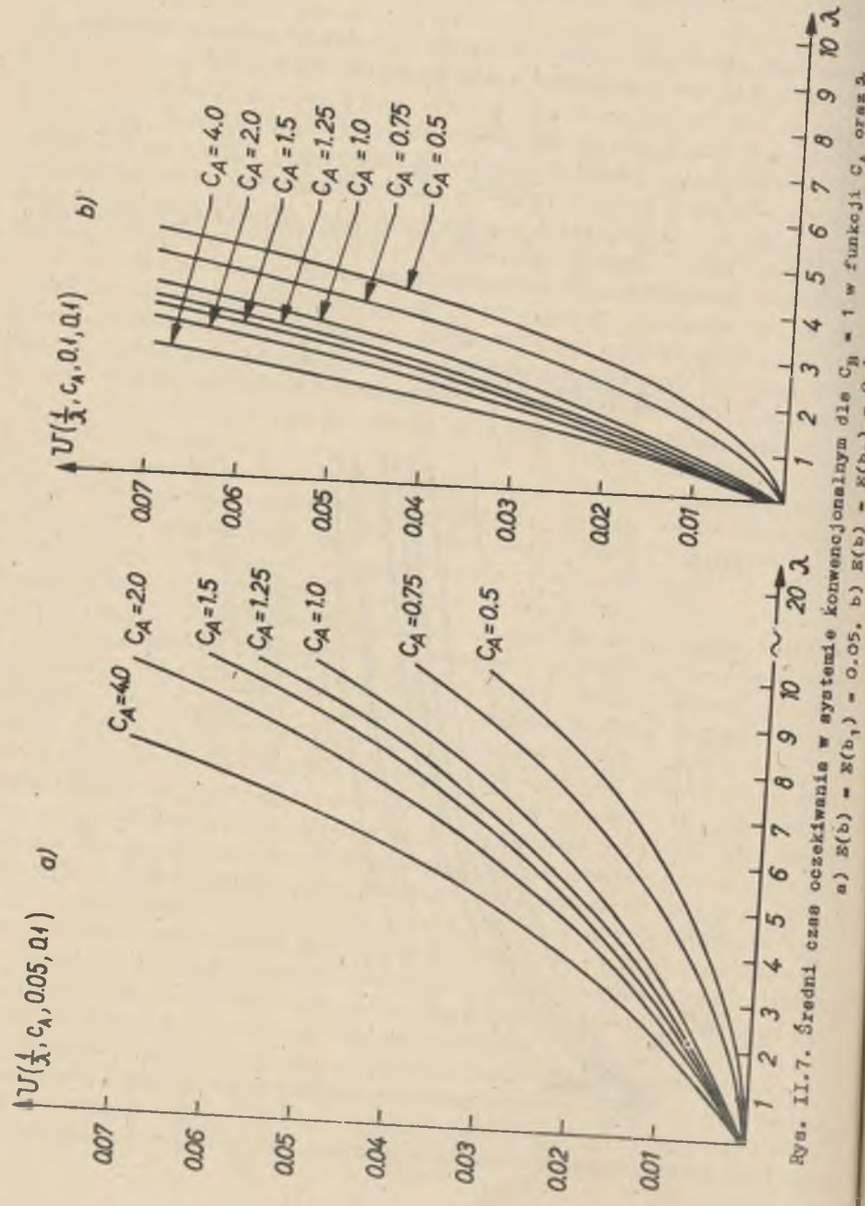
z parametrami określonymi przez E(r) oraz C_r w sposób podany w rozdziale I. Rozpatrzymy przy tym systemy WAD o zależności B(x/w) postaci:

$$B(x/w) = \begin{cases} B_1(x) & \text{dla} \quad w \leq D \\ B_2(x) & \text{dla} \quad w > D \end{cases} \qquad \text{(II.5.9)}$$

i założymy, że $C_{B_1} = C_{B_2}$, tj. oba rozkłady czasów obsługi mają jednakowy współczynnik zmienności.



Rys. II.6. Średni czas oczekiwania w systemie WAD dla $C_B = 1$ w funkcji C_A oraz λ



Rys. II.7. Średni czas oczekiwania w systemie konwencjonalnym dla C_B = 1 w funkcji C_A oraz λ
a) E(b) = E(b_1) = 0.05, b) E(b) = E(b_2) = 0.1

System WAD będzie więc jednoznacznie zdefiniowany przez podanie szóstki $\langle E(a), C_A, E(b_1), E(b_2), C_B, D \rangle$.

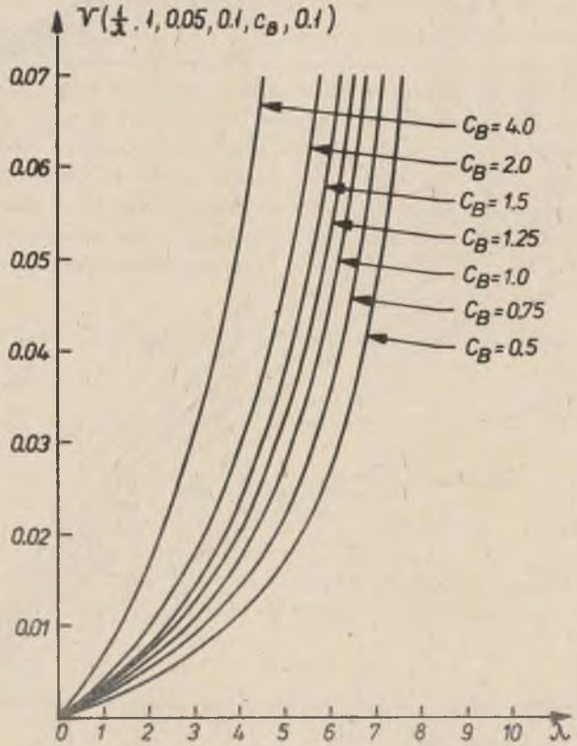
Przy założeniu, że zawsze co najmniej jedna z wartości C_A, C_B będzie równa jedności, możemy określić dla rozpatrywanych systemów WAD charakterystyki, a to:

- średni czas oczekiwania $E(w) = V(E(a), C_A, E(b_1), E(b_2), C_B, D)$,
- jednostkowy czas obsługi $\gamma^{(k)}(\lambda) = E(E(a), C_A, E(b_1), E(b_2), C_B, D)$

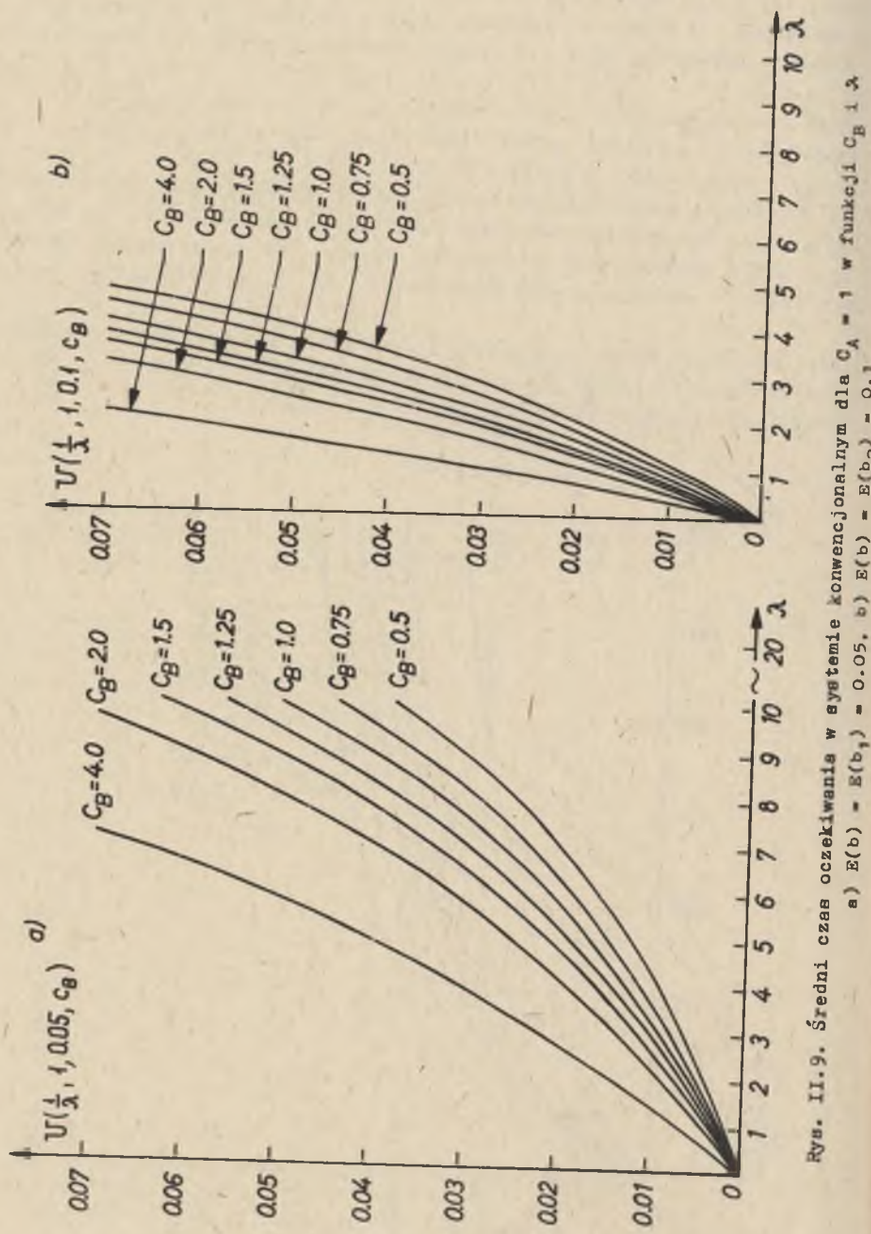
w oparciu o wzory podane w punkcie II.3.

Dla porównania zamieścimy też wyniki analizy konwencjonalnych systemów $K_2/K_2/1$. Takie systemy jednoznacznie definiujemy czwórką $\langle E(a), C_A, E(b), C_B \rangle$, na którą składają się wartości oczekiwane oraz współczynniki zmienności rozkładów odstępów między zgłoszeniami oraz rozkładów czasów obsługi.

Wartości średniego czasu oczekiwania dla takich systemów, które oznaczać będziemy przez $U(E(a), C_A, E(b), C_B)$, można określić w oparciu o znane zależności podane np. w [COHE 69].



Rys. II.8. Średni czas oczekiwania w systemie WAD dla C_A = 1 w funkcji C_B oraz λ



Rys. II.9. Średni czas oczekiwania w systemie konwencjonalnym dla $C_A = 1$ w funkcji C_B i λ
a) $E(b) = E(b_1) = 0.05$, b) $E(b) = E(b_2) = 0.1$

Na rys. II.6 przedstawiono rodzinę krzywych $V(\frac{1}{\lambda}, C_A, 0.05, 0.1, 1, 0.1)$ obrazującą wpływ wartości współczynnika C_A (przy wykładniczym rozkładzie czasów obsługi) na średni czas oczekiwania zgłoszeń. Dla porównania na rys. II.7 przedstawiono średni czas oczekiwania w systemie konwencjonalnym przy tym samym strumieniu wejściowym oraz czasie obsługi o rozkładzie $B_1(x)$ (krzywe $U(\frac{1}{\lambda}, C_A, 0.05, 0.1$ - przypadek a), względnie $B_2(x)$ (krzywe $U(\frac{1}{\lambda}, C_A, 0.1, 0.1$ - przypadek b). Tak więc rys. II.7 dotyczy systemów konwencjonalnych, w których czas obsługi stanowi oszacowanie od dołu bądź oszacowanie od góry systemu WAD, którego charakterystykę przedstawiono na rys. II.6 przy identycznym strumieniu zgłoszeń.

Analogiczny zestaw wykresów, a to: rodzinę krzywych $V(\frac{1}{\lambda}, 1, 0.05, 0.1, C_B, 0.1)$ ilustrującą wpływ wartości współczynnika C_B (przy zgłoszeniach napływających strumieniem Poissona) na średni czas oczekiwania zgłoszeń w systemie WAD oraz zamieszczone dla porównania wykresy $U(\frac{1}{\lambda}, 1, 0.05, C_B)$ i $U(\frac{1}{\lambda}, 1, 0.1, C_B)$, dotyczące dobranych analogicznie jak poprzednio systemów konwencjonalnych, zawierają rysunki II.8 oraz II.9.

W obu wyżej zilustrowanych eksperymentach średnie czasy oczekiwania w systemie WAD mieszczą się pomiędzy czasami oczekiwania porównywanych z nim systemów konwencjonalnych.

Z przedstawionych wykresów wynika duże podobieństwo wpływu współczynników C_A oraz C_B w systemach WAD oraz w systemach konwencjonalnych.
W celu dokładniejszego przebadania tego ewentualnego podobieństwa przeprowadzono następujący eksperyment.

Dla systemów WAD, rozpatrywanych wyżej i charakteryzujących się wartościami C_A lub C_B różnymi od jedności, wyznaczono oszacowanie $\tilde{V}$ wartości średniej czasu oczekiwania postaci

$$\tilde{V}(\tfrac{1}{\lambda}, C_A, E(b_1), E(b_2), C_B, D) = \varkappa V(\tfrac{1}{\lambda}, 1, E(b_1), E(b_2), 1, D), \qquad (II.5.10)$$

gdzie:

$$\varkappa = \frac{U(\frac{1}{\lambda}, C_A, \beta, C_B)}{U(\frac{1}{\lambda}, 1, \beta, 1)},$$

przy czym wartość β jest jednostkowym czasem obsługi zgłoszenia w systemie WAD, w którym $C_A = C_B = 1$ i zachowanymi pozostałymi parametrami, a więc:

$$\beta = E(\tfrac{1}{\lambda}, 1, E(b_1), E(b_2), 1, D).$$

Innymi słowy, szacujemy czas oczekiwania zgłoszeń w rozpatrywanym systemie WAD na podstawie czasu oczekiwania w przypadku identycznego systemu

WAD z wykładniczymi rozkładami A(x), $B_i(x)$; i=1,2 oraz współczynnika korekcyjnego $\varkappa$, określającego wpływ wapółczynnika zmienności rozkładów na czasy oczekiwania w systemach konwencjonalnych.

Zauważmy przy okazji, że w przypadku systemów ze zgłoszeniami nadchodzącymi strumieniem Poissona zachodzi:

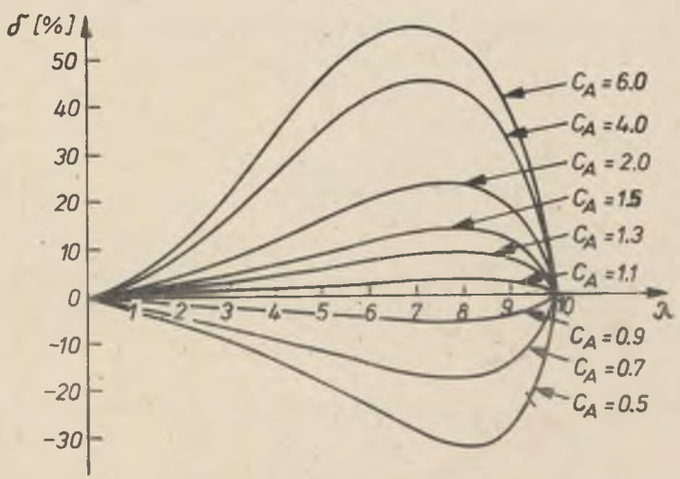
$$\varkappa_A = \frac{U(\frac{1}{\lambda}, 1, \beta, C_B)}{U(\frac{1}{\lambda}, 1, \beta, 1)} = \frac{1 + C_B}{2}, \qquad \text{(II.5.11)}$$

co wynika ze znanych wzorów opisujących system obsługi M/G/1. Rozpatrzmy dla wyżej zdefiniowanego oszacowania wyrażenia:

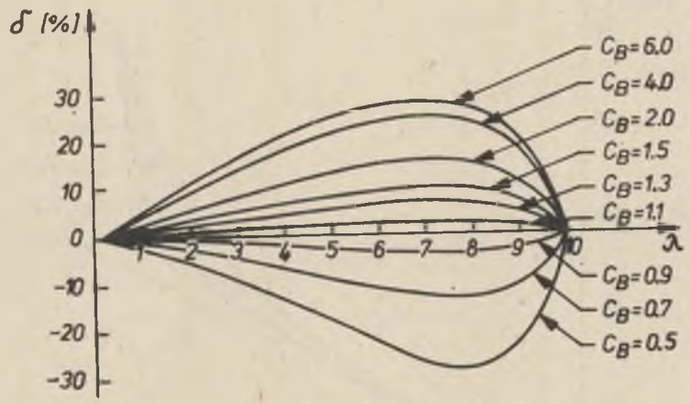
$$\delta = 1 - \frac{\tilde{V}(\frac{1}{\lambda}, C_A, E(b_1), E(b_2), C_B, D)}{V(\frac{1}{\lambda}, C_A, E(b_1), E(b_2), C_B, D)}, \qquad \text{(II.5.12)}$$

będące miarą różnicy wpływu współczynników C_A i C_B na czas oczekiwania w systemach WAD w stosunku do systemów konwencjonalnych. Wykresy wartości δ w funkcji intensywności strumienia zgłoszeń przedstawiono na rysunkach II.10 (przy $C_A = 1$) oraz II.11 (przy $C_B = 1$).

Błąd jest dodatni w przypadku, gdy $C_A > 1$ (lub $C_B > 1$) i ujemny w przeciwnym razie, przy czym dla dowolnego λ wartość $|\delta|$ rośnie ze wzrostem $(C_A - 1)$ względnie $(C_B - 1)$.



Rys. II.10. Błąd względny δ w funkcji λ i C_A, $C_B = 1$



Rys. II.11. Błąd względny δ w funkcji λ i C_B, $C_A = 1$

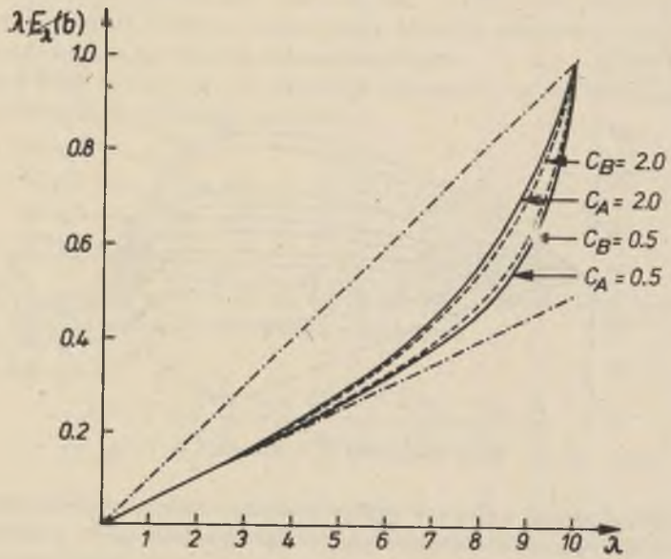
Z przedstawionych wykresów wynika jednoznacznie, że wpływ współczynnika zmienności rozkładu odstępów między zgłoszeniami oraz rozkładu czasu obsługi jest w przypadku systemu WAD istotnie silniejszy aniżeli w przypadku konwencjonalnych systemów obsługi.

Rozpatrując zmienność δ w funkcji λ widzimy, że δ dąży do zera w przypadku, gdy $\lambda \to 0$ oraz gdy $\lambda \to \frac{1}{E(b_2)} = \tau_{FIFO}$ (tj. wydajności granicznej systemu), a więc w tych przypadkach wpływ C_A i C_B jest w systemie WAD identyczny jak w systemie konwencjonalnym. Istotnie zauważmy, że parametry systemu WAD przy takich granicznych przejściach dążą asymptotycznie do parametrów systemów konwencjonalnych o czasie obsługi scharakteryzowanych dystrybuantą $B_1(x)$ lub $B_2(x)$.

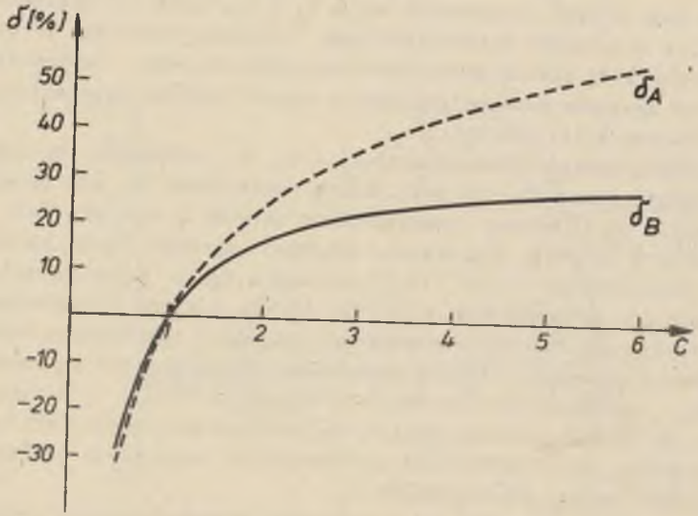
Najsilniejszy wzrost oddziaływania C_A i C_B w porównaniu do systemów konwencjonalnych obserwuje się przy takich wartościach λ, gdy największa jest niepewność co do wyboru rozkładu czasu obsługi w systemie WAD.

Spostrzeżenie to staje się jeszcze bardziej wyraźne, gdy weźmie się pod uwagę przedstawione na rys. II.12 wykresy stopnia wykorzystania stanowiska obsługi w systemie WAD, $\eta(\lambda)$. Na rysunku liniami przerywanymi zaznaczono stopień wykorzystania stanowiska obsługi w tych samych dwóch systemach konwencjonalnych, których używaliśmy do poprzednich porównań (rys. II.7, II.9). rysunków II.12 oraz II.13 wynika też inne, istotne spostrzeżenie, że zmiany współczynnika C_A mają silniejszy wpływ zarówno na czas oczekiwania, jak i na stopień wykorzystania stanowiska obsługi aniżeli identyczne zmiany współczynnika C_B.

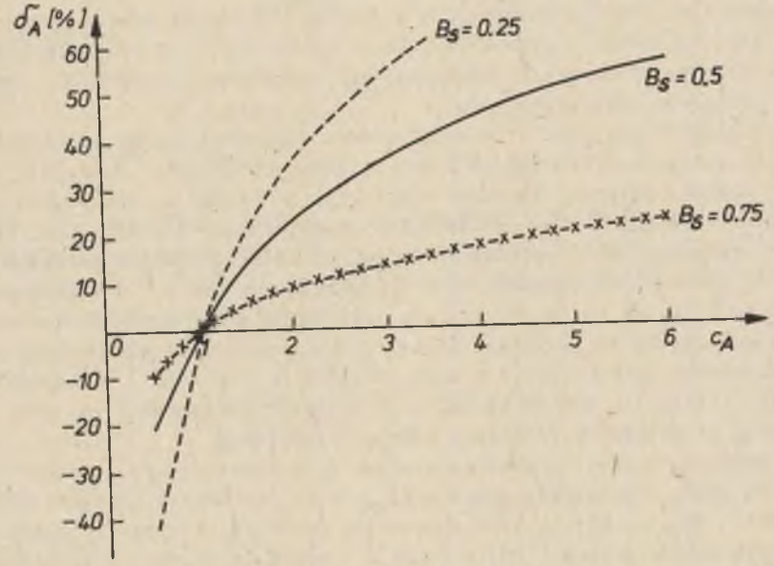
Tak więc widzimy, że o ile już w systemach konwencjonalnych podkreślano, że niecelowe jest ograniczanie się do pierwszego momentu w charakteryzowaniu zmiennych losowych definiujących system masowej obsługi,



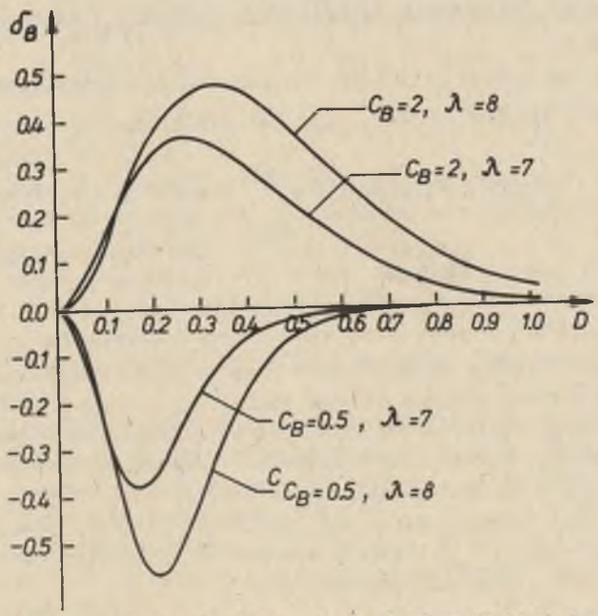
Rys. II.12. Stopień wykorzystania systemu WAD w funkcji λ dla różnych C_A i C_B



Rys. II.13. Porównanie błędów względnych δ_A (przy $C_A = C$, $C_B = 1$) oraz δ_B (przy $C_A = 1$, $C_B = C$) w funkcji C przy $\lambda = 7.5$



Rys. II.14. Błąd względny δ_A ($C_A \neq 1$, $C_B = 1$) w funkcji C_A i B_S dla $\lambda = 7.5$



Rys. II.15. Błąd względny δ_B ($C_A = 1$, $C_B \neq 1$) w funkcji C_B oraz λ

to tym bardziej jest to niewskazane w stosunku do systemów WAD, a specjalnie w odniesieniu do strumienia zgłoszeń.

Dążąc do uzyskania ważnych pod względem ilościowym charakterystyk należy więc korzystać z przedstawionych w punkcie II.2 relacji umożliwiających dokładne odwzorowanie rzeczywistego charakteru strumienia zgłoszeń oraz rozkładów czasów obsługi.

Mimo - podkreślonej już - pracochłonności związanej z ich stosowaniem; w świetle wniosków wynikających z przedstawionych powyżej badań fakt posiadania takich ogólnych rozwiązań analitycznych nabrał - obok czysto teoretycznych wartości - również znaczenia praktycznego. Zauważmy przy okazji, że w rozpatrywanych przykładach największe nieregularności związane z różnymi od jedności wartościami C_A i C_B występowały dla $\lambda = 0.7\text{-}0.9\ T$, a więc w zakresie, gdzie na ogół dobierany jest punkt pracy systemu.

Przedstawione na rysunkach II.6-II.13 wykresy dotyczyły pewnego zestawu parametrów systemu $E(b_1) = 0.05$, $E(b_2) = 0.1$, $D = 0.1$. Wyciągnięte stąd wnioski pozostają, jak to sprawdzono, w całej rozciągłości w mocy (pod względem jakościowym) również w innych przypadkach.

Ilustrację wpływu parametrów systemu (a specjalnie $E(b_1)$ oraz D) na wartość błędu δ zawierają rysunki II.14 i II.15. Wzrost wartości $|\delta|$ przy wzroście $\left[E(b_2) - E(b_1)\right]$ jest oczywisty, natomiast kształt krzywych obrazujących zmiany wartości $|\delta|$ w funkcji D można uzasadnić w sposób analogiczny do wywodu, jaki przeprowadzono w odniesieniu do krzywych przedstawionych na rys. II.10 oraz II.11.

c) <u>Możliwości zwiększenia wydajności systemu WAD z naturalnym regulaminem obsługi</u>

Rozpatrzmy system podstawowy ze zgłoszeniami napływającymi strumieniem Poissona, $J = 2$ oraz

$$\left.\begin{aligned} B_0(x) = B_1(x) &= 1 - e^{-\mu_1 x} \quad \text{dla} \quad w \leq w_1 = D, \\ B_2(x) &= 1 - e^{-\mu_2 x} \quad \text{dla} \quad w > w_1 = D, \\ \mu_2 &< \mu_1 \end{aligned}\right\} \qquad \text{(II.5.13)}$$

Zgodnie z omawianymi szczegółowo w punkcie II.1 warunkami stabilności wydajność graniczna takiego systemu wynosi $T_{FIFO} = \mu_2$.

Wprowadźmy obecnie pewną modyfikację rozpatrywanego systemu zakładając, że zgłoszenia, których czas oczekiwania przekracza zadaną wartość τ, są tracone (ich czas obsługi jest równy zeru).

Przy takim trybie pracy żadne zgłoszenie nie może oczekiwać z definicji dłużej aniżeli τ i system pozostaje stabilny dla dowolnych wartości intensywności strumienia zgłoszeń λ.

Dla analizy tego zmodyfikowanego systemu należy rozpatrzyć dwa przypadki:

i) przy $\tau \leq D$

$$B(x/w) = \begin{cases} B_1(x) = 1 - e^{-\mu_1 x} & \text{dla} \quad 0 \leq w \leq \tau \\ H(x) & \text{dla} \quad w > \tau \end{cases} \qquad \text{(II.5.14)}$$

W oparciu o wyprowadzone dla modelu podstawowego zależności można łatwo wyznaczyć wszystkie charakterystyki tego wariantu, pokrywającego się z rozpatrywanym np. w [BROD 73] systemem z ograniczonym czasem oczekiwania, w tym wydajność systemu V, wynoszącą:

$$V = \lambda \frac{1 - \rho e^{(\lambda-\mu_1)D}}{1 - \rho^2 e^{(\lambda-\mu_1)D}} \quad (\text{dla } \lambda \neq \mu_1), \qquad \text{(II.5.15)}$$

gdzie: $\rho = \frac{\lambda}{\mu_1}$.

ii) przy $\tau > D$

$$B(x/w) = \begin{cases} B_1(x) = 1 - e^{-\mu_1 x} & \text{dla} \quad 0 \leq w \leq D, \\ B_2(x) = 1 - e^{-\mu_2 x} & \text{dla} \quad D < w \leq \tau \\ H(x) & \text{dla} \quad w > \tau \end{cases} \qquad \text{(II.5.16)}$$

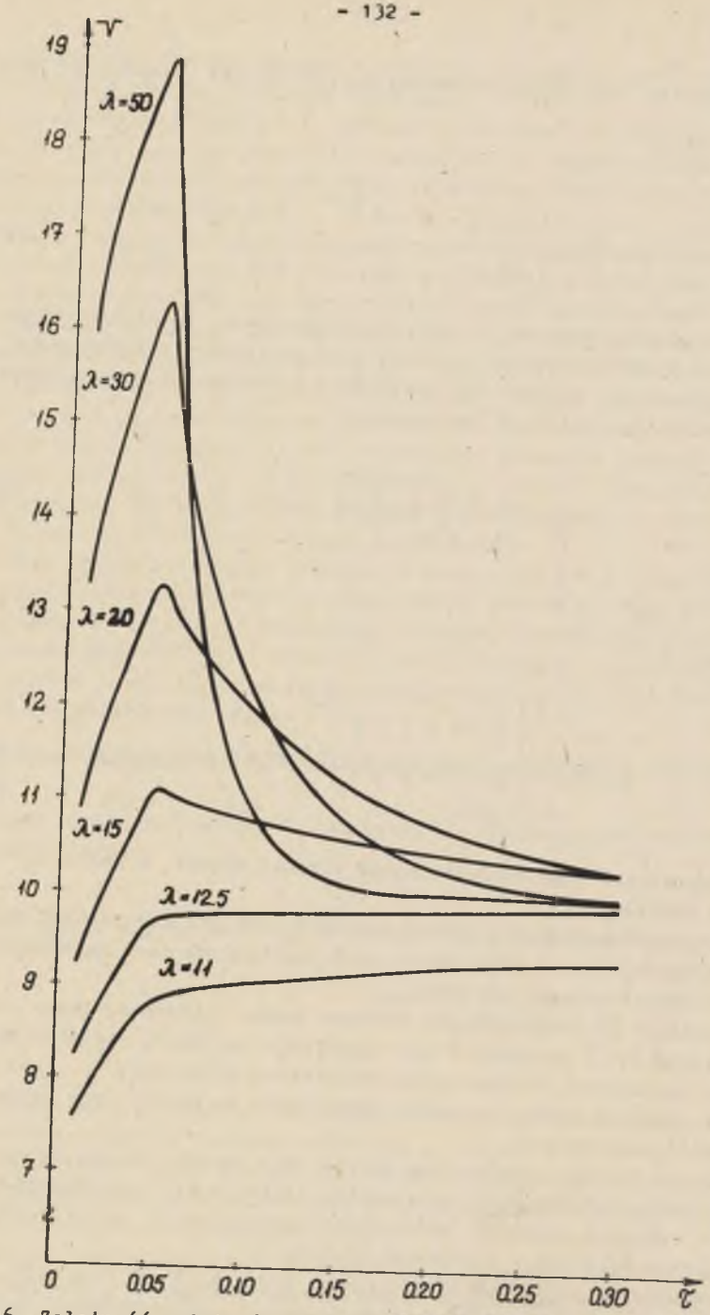
Charakterystyki tego wariantu można uzyskać wprost w oparciu o wzory podane w punkcie III.3c.

Rozpatrzymy wydajność takich systemów obsługi przy różnych wartościach λ, a w szczególności zastanowimy się, czy możliwe jest osiągnięcie wydajności przekraczającej wartości μ_2.

Zauważmy, że rozpatrywanym systemom można też nadać inną, równoważną interpretację. W przypadku, gdy zgłoszenia napływają strumieniem Poissona, strata zgłoszenia, którego czas oczekiwania przekracza τ, jest równoważna zablokowaniu źródła zgłoszeń każdorazowo na okres, gdy zapas pracy w systemie przekroczy τ.

Parametr λ określa wówczas moc źródła (tj. średnią liczbę zgłoszeń, które mogły zostać wygenerowane w jednostce czasu) a nie rzeczywistą intensywność strumienia zgłoszeń, która wobec braku strat i stabilności systemu jest oczywiście równa wydajności systemu V.

Przebadajmy na przykładzie, przy $\mu_1 = 20$, $\mu_2 = 10$, $D = 0.05$, zależność wydajności systemu V od parametrów λ oraz τ, zilustrowaną na rys. II.16 i posiadającą typową postać.



Rys. II.16. Zależność wydajności systemu V od λ oraz τ ; $\mu_1 = 20, \mu_2 = 10$, $D = 0.05$

Przy $\lambda < \mu_2$ wydajność $V < \lambda$ i $V \to \lambda$, gdy $\tau \to \infty$. Zastosowanie wartości λ niewiele przekraczających μ_2 prowadzi do identycznej postaci zależności V od λ i τ jak to ma miejsce przy $\lambda < \mu_2$. Natomiast silniejsze zwiększenie λ prowadzi do uzyskiwania wartości wydajności V przekraczających, i to znacznie, wartość μ_2.

Łatwo zauważyć, że jeśli tylko $V > \mu_2$, to maksymalną wydajność osiąga się przy $\tau = D$, co znalazło potwierdzenie również w szeregu innych przykładach.

Zależność wydajności V od λ przy $\tau = D$ dla rozpatrywanego przykładu zawiera tabela II.3.

Tabela II.3

Zależność wydajności V systemu obsługi od λ przy stratach zgłoszeń (blokowaniu źródła) $\mu_1 = 20$, $\mu_2 = 10$, $D = \tau = 0.05$

λ	10	12	14	16	18	20	25	30
V	8.21	9.46	10.58	11.60	12.51	13.33	15.03	16.31
λ	40	50	60	70	80	90	100	
V	17.97	18.89	19.39	19.67	19.81	19.90	19.94	

Z (II.5.15) wynika, iż $\lim_{\lambda \to \infty} V(\lambda) = \mu_1 > \mu_2$, co określa maksymalną wydajność rozpatrywanego systemu przy $\tau = D$.

Warto podkreślić, iż przy $\tau \leq D$ wydajność systemu rośnie ze wzrostem λ. Dla $\tau > D$ sytuacja jest znacznie bardziej złożona. Jak to wynika z rys. II.16, przy pewnych, wystarczająco dużych wartościach, wydajność systemu maleje ze wzrostem mocy źródła zgłoszeń λ! Ten pozorny paradoks wynika z faktu zwiększenia się udziału procentowego zgłoszeń wymagających obsługi o czasie określonym rozkładem $B_2(x)$. Wynika stąd, że sam tylko dobór mocy źródła λ nie przesądza o wydajności systemu.

Wskazane powyżej na prostym przykładzie prawidłowości można łatwo uogólnić. W szczególności, rozpatrując dowolny system podstawowy, określony zależnością (II.6), można, drogą ograniczenia czasu oczekiwania do D_1 (gdzie: $D_1 = w_k$, natomiast k wyznacza się z relacji $E(b_k) = \min_{i=1,2,\ldots,J} E(b_i)$) uzyskać wydajność graniczną dążącą do $\left[E(b_k)\right]^{-1}$ przy $\lambda \to \infty$. Takie ograniczenie czasu oczekiwania uzyskuje się drogą blokowania źródła zgłoszeń (bądź odrzucania zgłoszeń), co jest efektywnym środkiem kształtowania charakterystyki systemu. Jest to jeszcze jedna cecha różniąca systemy WAD od systemów konwencjonalnych, gdzie wprowadzanie ograniczeń na pobyt zgłoszeń w systemie (bądź blokowanie źródła zgłoszeń) nie może być narzędziem dla zwiększenia wydajności.

Ponieważ oczywiście zwiększanie mocy (przedymensjonowanie) źródła zgłoszeń wiąże się z dodatkowym kosztem, można sformułować w oparciu o zależ-

ności analityczne, wyprowadzone w tym rozdziale oraz wyznaczone dla konkretnego, praktycznie rozpatrywanego przypadku, funkcje kosztów, zadanie optymalizacji, dla określenia czy i w jakim stopniu bardziej opłacalne jest zwiększenie mocy źródła w stosunku do alternatywnego podejścia, jakimi jest zwiększanie mocy stanowiska obsługi (a więc, w efekcie zwiększanie zarówno μ_1 jak i μ_2).

Przykład sformułowania i rozwiązania takiego zadania optymalizacji podamy w rozdziale IV.

Ponieważ zamiast blokować źródło zgłoszeń można odrzucać z kolejki zgłoszenia, których czas oczekiwania przekroczył wartość D, więc naturalne jest rozpatrzenie magazynowania takich odrzuconych zgłoszeń i wprowadzanie ich na stanowisko obsługi w przypadku, gdy kolejka opróżni się.

Takie regulaminy obsługi omówimy szczegółowo w rozdziale III.

d) Podsumowanie

Przedstawione w niniejszym rozdziale wyniki analityczne pozwalają na badanie - w sposób ścisły - bardzo szerokiej klasy jednokanałowych systemów masowej obsługi z czasem obsługi zależnym od czasu oczekiwania, umożliwiając równocześnie rozpatrywanie w sposób przybliżony, z dowolnie dużą dokładnością, w zasadzie wszystkich takich systemów.

Dotyczy to nie tylko systemów z oczekiwaniem, lecz również systemów ze stratami. Wprowadzając bowiem w definicji (II.6) dystrybuanty $B_i(x)$ o postaci:

$$B_i(x) = (1 - \psi)\, H(x) + \psi\phi(x),$$

gdzie: $0 \leq \psi \leq 1$, $\phi(x)$ pewna dystrybuanta,

można, przy odpowiednio dużej liczbie przedziałów J, aproksymować z dowolną dokładnością dystrybuantę warunkową o postaci (II.1.16). Tak więc przedstawione rozwiązanie może być traktowane jako alternatywa w stosunku do ogólnego rozwiązania takiego przypadku przedstawionego w [BACE 81].

Dodatkowo, specjalnymi przypadkami rozpatrywanego modelu są również systemy typu $K_1/G/1$ z różnym od pozostałych rozkładem czasu obsługi pierwszego zgłoszenia w każdym okresie czynnym.

Posługując się uzyskanymi rozwiązaniami, wskazano na przykładach pewne własności systemów WAD. W szczególności stwierdzono brak możliwości racjonalnego prognozowania charakterystyk takich systemów, a nawet oceny ich stabilności w oparciu o rozpatrzenie pewnych systemów konwencjonalnych, a także znaczny (silniejszy jeszcze aniżeli w systemach konwencjonalnych) wpływ drugiego momentu zmiennych losowych opisujących system na jego charakterystyki.

Uproszczenia polegające nie tylko na pominięciu zależności czasu obsługi od czasu oczekiwania ale nawet na przyjęciu rozkładów wykładniczych (co znacznie upraszcza analizę - patrz punkt II.3.a1) nie powinny więc być stosowane, jako prowadzące do istotnych błędów. Wskazano też na dodatkowe, nie występujące w systemach konwencjonalnych, możliwości zwiększenia wydajności systemów WAD. Na krótki komentarz zasługuje przy tym zagadnienie oddziaływania na charakterystykę źródła zgłoszeń, jako że w wielu pracach poświęconych teorii masowej obsługi charakterystyka ta traktowana jest jako nie podlegająca żadnym zmianom część definicji systemu. Takie podejście ulega zmianie przy rozpatrywaniu bardziej rozbudowanych systemów (sieci stanowisk) masowej obsługi.

W rozdziale I.2 wskazano, że pojęcie strumienia zgłoszeń opuszczającego źródło jest równoważne pojęciu strumienia wyjściowego z pewnego stanowiska obsługi (w przypadku źródła nieskończenie wymiarowego jest to stanowisko jednokanałowe z nigdy niepustą kolejką zgłoszeń).

Zagadnienie dopasowania charakterystyki źródła zgłoszeń jest więc w istocie identyczne z pewnymi problemami optymalizacji w sieciach stanowisk obsługi, do których wrócimy szerzej w rozdziale IV.

III. SYSTEMY MASOWEJ OBSŁUGI Z CZASEM OBSŁUGI ZALEŻNYM OD CZASU OCZEKIWANIA I DOWOLNYM REGULAMINEM OBSŁUGI

W poprzednim rozdziale omówiliśmy metody analizy oraz podstawowe własności szerokiej klasy systemów masowej obsługi z czasem obsługi zależnym od czasu oczekiwania i naturalnym regulaminem obsługi. Obecnie rozpatrzymy tę samą klasę systemów obsługi przy zastosowaniu innych regulaminów obsługi.

Należy od razu podkreślić, że przy odejściu od założenia o naturalnym regulaminie obsługi rozpatrywany problem znacznie się komplikuje, co zilustrujemy na prostym przykładzie:

Niech dany będzie jednokanałowy system WAD, do którego zgłoszenia napływają strumieniem Poissona z parametrem λ. Załóżmy, że w chwili $\tau^i - 0$ w kolejce znajduje się l zgłoszeń $\zeta^{i-1}, \zeta^{i-2}, \ldots, \zeta^{i-l}$, które napłynęły do systemu w chwilach czasu $\tau^{i-1}, \tau^{i-2}, \ldots, \tau^{i-l}$, natomiast na stanowisku obsługi znajduje się zgłoszenie ζ^{i-l-1}, którego pozostały czas obsługi wynosi ψ^i.

Zauważmy, że zapas pracy w systemie

$$z_R(\tau^i-0) = \varphi(R,\tau^{i-1},\tau^{i-2},\ldots,\tau^{i-l},\psi^i),$$

zależy w sposób istotny nie tylko od momentów napłynięcia poszczególnych zgłoszeń, ale także od zastosowanego regulaminu obsługi. Co więcej, ogólnie rzecz biorąc, zapas pracy w systemie po nadejściu kolejnego zgłoszenia

$$z_R(\tau^i+0) \neq z_R(\tau^i-0) + b^i_w, \qquad w = z_R(\tau^i-0).$$

Dla jednoznacznego scharakteryzowania stanu systemu w chwili τ^i w oparciu o znajomość regulaminu obsługi i stanu systemu w chwili τ^{i-1} konieczne byłoby wprowadzenie wektora stanu, którego składowymi byłyby momenty nadejścia do systemu wszystkich zgłoszeń oczekujących w kolejce oraz pozostały czas obsługi zgłoszenia aktualnie obsługiwanego.

Wektor stanu mógłby mieć zmienną wymiarowość. Proces zmian takiego wektora stanu byłby łańcuchem Markowa z dyskretnym czasem, ciągłą przestrzenią stanów, wielowymiarowym o zmiennej wymiarowości - co przesądza o praktycznej niemożliwości badania warunków stabilności systemu (tj. ergodyczności

łańcucha Markowa), czy też analizy jego charakterystyk, w sposób analogiczny do przedstawionego w rozdziale II (z wyjątkiem specjalnych prostych przypadków).

Biorąc pod uwagę zależność zapasu pracy od regulaminu obsługi można się przy tym spodziewać, że warunki stabilności też będą różne dla różnych regulaminów obsługi i konieczne będzie oddzielne ich wyznaczanie dla każdego rozpatrywanego przypadku.

Jak dotąd, ukazała się tylko jedna praca [BUZA 74] traktująca o trzech prostych przypadkach systemów WAD z regulaminem obsługi różnym od naturalnego, której wyniki przedyskutowano w punkcie III.1.d.

Można więc powiedzieć, że w zakresie badania systemów WAD, z różnymi od naturalnego regulaminami obsługi, pojawiają się trzy zasadnicze zadania:

a) wyznaczenie warunków stabilności systemów,
b) wyznaczenie charakterystyk systemów stabilnych w stacjonarnym stanie granicznym,
c) określenie wskazań dla doboru regulaminu obsługi (np. pod kątem maksymalizacji wydajności granicznej).

W dalszym ciągu omówimy kolejno metody rozwiązywania tych trzech zadań przy użyciu zarówno metod analitycznych, jak i symulacyjnych, a w zakresie ostatniego z nich również - w pewnym zakresie - metod heurystycznych.

Metody symulacyjne przedstawione w tym rozdziale mogą też być z powodzeniem stosowane do innych przypadków niekonwencjonalnych systemów masowej obsługi.

III.1. Warunki stabilności

W niniejszym punkcie zaproponujemy oraz udowodnimy prawidłowość warunku stabilności dla pewnej klasy systemów z czasem obsługi zależnym od czasu oczekiwania, wskazując też możliwości osłabienia założeń, przy których dowód został przeprowadzony.

Analityczną postać warunku stabilności udało się wcześniej uzyskać dla bardzo nielicznych prostych przypadków.

Pokażemy, że część tych wyników stanowi szczególny przypadek rozpatrywanej tu klasy systemów i omówimy pozostałe.

a) Opis rozpatrywanej klasy systemów

W dalszym ciągu rozpatrywać będziemy jednokanałowe systemy WAD o następujących własnościach.

Zgłoszenia napływają pojedynczo, tworząc strumień Poissona o parametrze λ, a więc:

$$A(x) = 1 - e^{-\lambda x}. \qquad \text{(III.1.1)}$$

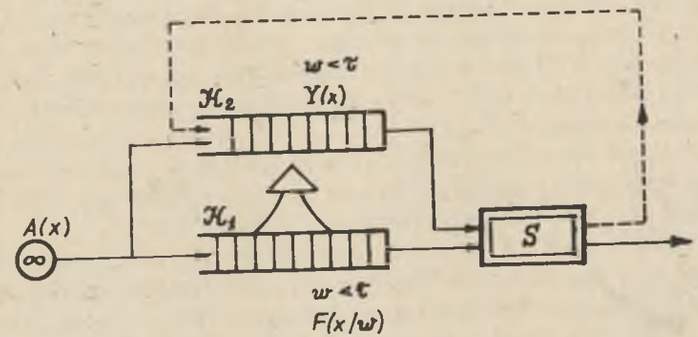
Założymy też, że w każdej chwili t dostępna jest określona w sposób deterministyczny prognoza faktu, czy w odcinku czasu $(t, t+\vartheta)$, $\vartheta \geq 0$ napłynie kolejne zgłoszenie.

Oczywiście, z definicji strumienia zgłoszeń wynika, że prognoza będzie negatywna z prawdopodobieństwem β

$$\beta = e^{-\lambda\vartheta} \qquad \text{(III.1.2)}$$

i pozytywna z prawdopodobieństwem $(1 - \beta)$.

Możliwość uzyskania takiej prognozy wynika zwykle z faktu, iż napływający do rozpatrywanego systemu strumień zgłoszeń jest strumieniem wyjściowym innego stanowiska obsługi, umożliwiającego określenie momentu wyemitowania zgłoszenia.



Rys. III.1. Schemat rozpatrywanej klasy systemów masowej obsługi (linią kreskowaną zaznaczono możliwość powrotu do kolejki Q_2 zgłoszeń, których obsługa została przerwana, jeśli dopuszcza się taki wariant)

Założymy, że czas obsługi każdego zgłoszenia jest zależny od czasu oczekiwania w sposób określony warunkową dystrybuantą

$$B(x/w) = \begin{cases} F(x/w) & \text{dla} \quad w \leq \tau \\ Y(x) & \text{dla} \quad w > \tau \end{cases} \qquad \text{(III.1.3)}$$

$$E(y) < \infty \qquad \exists_{M<\infty} \forall_{w} \left\{ \int_0^{\infty} x\, d_x P(x/w) < M \right\}$$

co stanowi uogólnienie zależności przyjmowanej dla modelu podstawowego.

Przyjmiemy, że system obsługi ma następującą strukturę (zilustrowaną na rys. III.1):

- zgłoszenia, których czas oczekiwania na rozpoczęcie obsługi nie przekroczył wartości τ tworzą kolejkę $\mathcal{K}_1$, obsługiwaną wg dowolnego regulaminu obsługi $\mathcal{R}_1$,
- zgłoszenia, których czas oczekiwania na rozpoczęcie obsługi przekroczył wartość τ, tworzą kolejkę $\mathcal{K}_2$, obsługiwaną wg dowolnego regulaminu obsługi $\mathcal{R}_2$. Zauważmy, że czas obsługi tych zgłoszeń nie zależy od czasu ich oczekiwania,
- zgłoszenia z kolejki $\mathcal{K}_1$ posiadają priorytet względny, bezwzględny lub mieszany bezwzględno-względny w stosunku do zgłoszeń z kolejki $\mathcal{K}_2$,
- dodatkowo, jeśli kolejka $\mathcal{K}_1$ jest pusta, to następuje analiza prognozy. W przypadku prognozy pozytywnej oczekuje się napłynięcia nowego zgłoszenia (którego obsługa rozpoczyna się bezzwłocznie), natomiast w przypadku prognozy negatywnej do obsługi wybierane jest zgłoszenie z kolejki $\mathcal{K}_2$ (zgodnie z regulaminem $\mathcal{R}_2$),
- jeżeli obie kolejki są puste, system oczekuje bezczynnie na napłynięcie zgłoszenia, którego obsługę rozpoczyna się bezzwłocznie.

Przez pojęcie priorytetu mieszanego bezwzględno-względnego [JAIS 68], [CHAC 74], [WOLI 76] rozumiemy przypadek, gdy obsługa zgłoszenia o niższym priorytecie może być przerwana, jeśli zaliczony dotąd czas obsługi nie przekroczył pewnej zadanej wartości $\mathcal{F}$, będącej parametrem regulaminu - w przeciwnym razie przerwanie nie może nastąpić. Regulamin taki obejmuje, jako przypadki szczególne, regulamin priorytetowy względny (przy $\mathcal{F}=0$) oraz regulamin priorytetowy bezwzględny (przy $\mathcal{F}\to\infty$).

W przypadku priorytetu bezwzględnego (bądź pierwszej fazy obsługi przy priorytecie mieszanym) dopuszcza się przy tym konieczność wykonania pracy własnej związanej z opróżnieniem stanowiska obsługi. Nie są natomiast nakładane żadne ograniczenia na wariant stosowanego priorytetu bezwzględnego (z kontynuacją obsługi przerwanej, wznowieniem obsługi przerwanej od początku wg identycznego lub różnego rozkładu czasu obsługi).

Tak więc rozpatrujemy rodzinę systemów, w których zgłoszenia nie posiadające już uwarunkowania czasu obsługi od czasu oczekiwania tworzą odrębną kolejkę, o niższym priorytecie w stosunku do kolejki zgłoszeń, których uwarunkowanie takie jeszcze dotyczy. Wprowadzono też dodatkowe ograniczenie na możliwość rozpoczynania obsługi zgłoszeń o niższym priorytecie, związane z prognozą napływania zgłoszeń o priorytecie wyższym.

b) Dowód warunku stabilności

Niech $\xi_1,\xi_2,\ldots,\xi_i$ oznaczają chwile czasu, w których kończy się obsługa zgłoszenia, kolejka $\mathcal{K}_1$ jest pusta i prognoza nadejścia zgłoszenia w odcinku czasu $(\xi_i, \xi_i+\vartheta)$ jest negatywna.
Innymi słowy, jeśli tylko kolejka $\mathcal{K}_2$ nie jest pusta, wówczas w chwili $\xi_i + 0$ rozpoczyna się obsługa zgłoszenia z tej kolejki - w przeciwnym razie stanowisko obsługi rozpoczyna bezczynne oczekiwanie na napłynięcie zgłoszenia.

Oznaczmy przez X_m liczbę zgłoszeń w kolejce $\mathcal{K}_2$ w chwili ξ_i, określającą jednoznacznie stan systemu w tych chwilach czasu. Łatwo widać, że ciąg $\{X_m\}$ jest łańcuchem Markowa z dyskretnym czasem, przyjmującym wartości całkowitoliczbowe nieujemne.

Niech $[p_{ij}]$ oznacza macierz przejść rozpatrywanego łańcucha Markowa

$$p_{ij} = Pr\{X_{m+1} = j / X_m = i\} \qquad i,j=0,1,\ldots \tag{III.1.4}$$

Wprost z definicji oraz z przyjętej postaci strumienia zgłoszeń wynika, że łańcuch $\{X_m\}$ jest nieokresowy oraz:

$$\left.\begin{array}{lll} \text{i)} & p_{ij} = 0 & \text{dla} \quad i \geqslant 1,\ j < i-1, \\ \text{ii)} & p_{i,i-1} > 0 & \text{dla} \quad i \geqslant 1, \\ \text{iii)} & p_{i,j} > 0 & \text{dla} \quad i \geqslant 0,\ j \geqslant i, \end{array}\right\} \tag{III.1.5}$$

a więc jest on też nieprzywiedlny.
Niech

$$\gamma_i = E\{X_{m+1} - X_m / X_m = i\} = \sum_{j=1}^{\infty} j p_{ij} - i. \tag{III.1.6}$$

O ergodyczności łańcuchów Markowa z dyskretnym czasem, nieprzywiedlnych, nieokresowych, przyjmujących całkowite, nieujemne wartości można rozstrzygać w oparciu o następujące warunki:

Warunek 1 [PAKE 69]

Warunkiem wystarczającym ergodyczności zdefiniowanych wyżej łańcuchów Markowa jest, aby:

$$\text{i)} \quad \bigvee_i |\gamma_i| < \infty \tag{III.1.7}$$

$$\text{ii)} \quad \limsup_{i\to\infty} \gamma_i < 0. \tag{III.1.8}$$

Warunek 2

Jeżeli ciąg $\{\gamma_i\}$ składa się z r podciągów monotonicznie malejących, zbieżnych odpowiednio do granic $\sigma_1,\sigma_2,\ldots,\sigma_r$ i każdy element ciągu $\{\gamma_i\}$ należy do jednego z tych podciągów, wówczas warunkiem koniecznym ergodyczności zdefiniowanych wyżej łańcuchów Markowa jest, aby:

$$\liminf_{i\to\infty} \gamma_i = \min_{i=1,2,\ldots,r} \sigma_i < 0. \tag{III.1.9}$$

Dowód tego warunku zamieszczono w dodatku do niniejszego punktu.

Łatwo widać, że w przypadku rozpatrywanego łańcucha Markowa $\{X_n\}$ zachodzi:

$$\bigvee_{i>0} \{\gamma_i = \gamma_1\}, \tag{III.1.10}$$

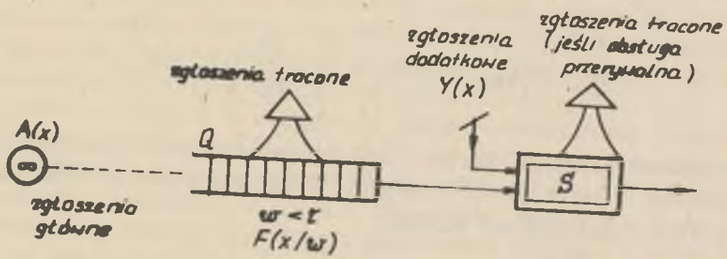
tak więc ostatecznie można sformułować:

Warunek ergodyczności łańcucha Markowa $\{X_n\}$

Jeżeli $|\gamma_1| < \infty$ oraz $|\gamma_0| < \infty$ to nierówność

$$\gamma_1 < 0 \tag{III.1.11}$$

jest warunkiem koniecznym i wystarczającym ergodyczności łańcucha Markowa $\{X_m\}$.



Rys. III.2. Schemat pomocniczego systemu obsługi ze stratami

Udowodniony warunek ergodyczności pozwala nam więc orzekać o stabilności rozpatrywanych systemów WAD w oparciu o następującą procedurę:

- przyjąć, że kolejka $\mathcal{K}_2$ nie może być pusta,
- wyznaczyć wartość oczekiwaną $L(\lambda)$ liczby zgłoszeń przechodzących do kolejki $\mathcal{K}_2$ w czasie pomiędzy momentami rozpoczęcia obsługi dwóch kolejnych zgłoszeń pochodzących z kolejki $\mathcal{K}_2$ przy intensywności strumienia zgłoszeń wynoszącej λ,
- system jest stabilny dla wszystkich wartości λ, przy których

$$L(\lambda) < 1. \tag{III.1.12}$$

Zauważmy, że wartość $L(\lambda)$ jest stosunkiem intensywności strumienia zgłoszeń traconych do strumienia zgłoszeń dodatkowych, w pomocniczym systemie obsługi WAD (rys. III.2) ze stratami, w którym przy pustej kolejce $\mathcal{K}$ i

negatywnej prognozie zawsze rozpoczyna się obsługa zgłoszenia dodatkowego, o dystrybuancie czasu trwania $Y(x)$.

Na strumień zgłoszeń traconych składają się zgłoszenia, których czas oczekiwania w kolejce $\mathcal{K}$ przekroczył τ oraz - w przypadku systemów z obsługą przerywalną wznawianą od początku - również zgłoszenia, których obsługa została przerwana.

Spełnienie sformułowanego w warunku stabilności wymagania, aby zachodziło $|\gamma_0| < \infty, |\gamma_1| < \infty$, sprowadza się wobec skończonej intensywności strumienia zgłoszeń do zapewnienia, by okres czasu pomiędzy kolejnymi chwilami ζ_i, ζ_{i+1} był skończony z prawdopodobieństwem równym jedności, co - jak łatwo sprawdzić - zachodzi zawsze przy przyjętych dla rozpatrywanego systemu założeniach.

Wartość $L(\lambda)$ niezbędna do orzeknięcia o stabilności systemu może być wyznaczona na drodze analitycznej bądź też przy użyciu metod symulacyjnych. Użyteczność wyprowadzonego kryterium wykracza więc znacznie poza klasę przypadków, w których możliwe jest analityczne badanie systemu pomocniczego. Znaczenie posiadania takiego kryterium przy użyciu metod symulacyjnych stanie się jeszcze bardziej oczywiste w świetle rozważań zamieszczonych w następnym punkcie, a dotyczących szerszej klasy systemów WAD.

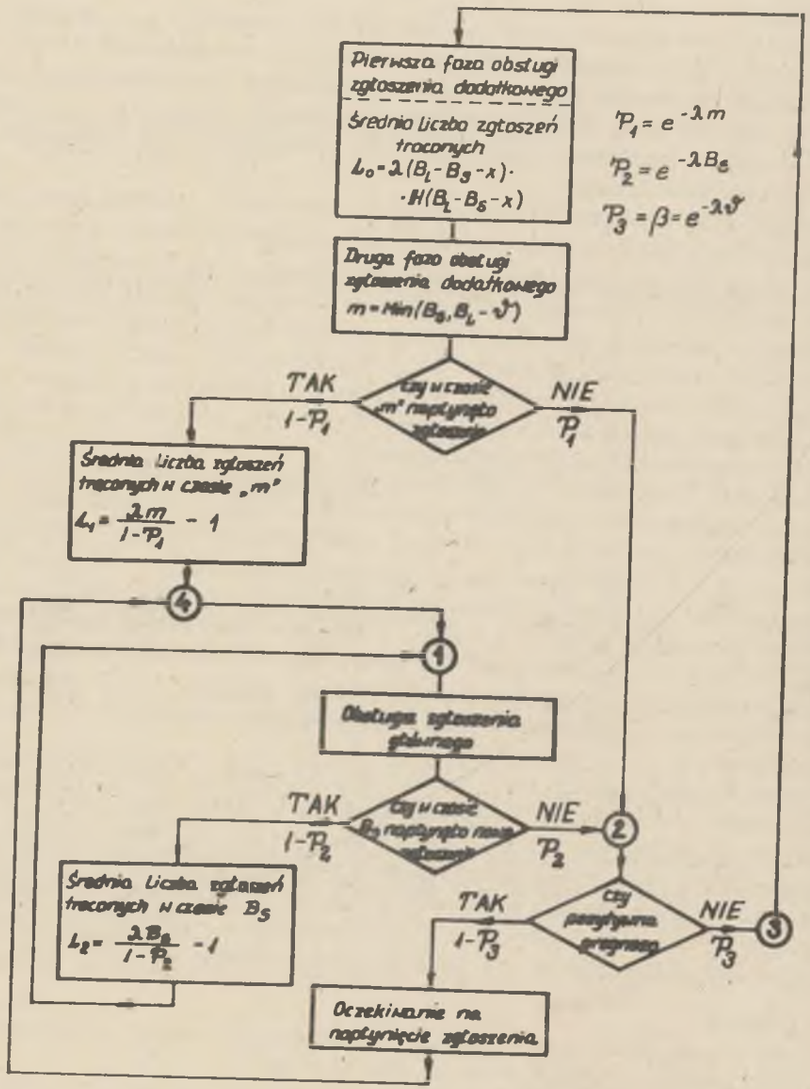
Rozpatrzmy jeszcze krótko możliwości osłabienia założeń, przy których wyżej podany warunek został wyprowadzony.

Zauważmy, że w oparciu o zacytowane warunki (wystarczający oraz konieczny) ergodyczności łańcuchów Markowa można orzekać o stabilności systemów masowej obsługi o zdefiniowanej na wstępie strukturze, w przypadku ogólniejszym, w którym czas obsługi zgłoszeń oczekujących w kolejce $\mathcal{K}_1$ byłby zależny od liczby zgłoszeń "i" oczekujących w kolejce $\mathcal{K}_2$ i posiadał dystrybuantę $F^i(x/w)$ $i = 0,1,\ldots$. Podobnie od liczby zgłoszeń w kolejce $\mathcal{K}_2$ może być zależny czas obsługi zgłoszeń pochodzących z kolejki $\mathcal{K}_2$ i posiadających wówczas dystrybuantę $Y^i(x)$ dla $i=0,1,\ldots,l$ oraz $Y^l(x)$ dla $i > l$. Sposób uogólnienia jest w tym przypadku oczywisty i nie będziemy go omawiać.

Można sformułować hipotezę, że analogiczny warunek stabilności będzie obowiązywał również w przypadku wielokanałowym, także dla prawie dowolnych strumieni zgłoszeń. Z punktu widzenia kolejki $\mathcal{K}_2$ mamy bowiem do czynienia jak gdyby z jednostanowiskowym konwencjonalnym systemem masowej obsługi, w którym obsługa zgłoszenia polega na "rzeczywistej obsłudze" zgłoszenia pobranego z kolejki $\mathcal{K}_2$ oraz opróżnieniu kolejki $\mathcal{K}_1$. Dowodu prawdziwości tej hipotezy nie będziemy tu zamieszczać.

c) Przykład

Dla zilustrowania sposobu użycia przedstawionej metody wyznaczymy wydajność graniczną przykładowego systemu obsługi.



Rys. III.3. Schemat obsługi zgłoszeń w systemie zdefiniowanym zależnością (III.1.13) przy regulaminie obsługi LIFO (ϑ)

Niech zależność czasu obsługi od czasu oczekiwania będzie zadana w postaci:

$$B(x/w) = \begin{cases} H(x - B_S) & \text{dla} \quad w \leq B_S \leq B_L, \\ H(x - B_L) & \text{dla} \quad w > B_S. \end{cases} \tag{III.1.13}$$

Załóżmy, że zarówno regulamin $\mathcal{R}_1$, jak i regulamin $\mathcal{R}_2$ są regulaminami odwrotnymi (LIFO) i obsługa wszystkich zgłoszeń jest nieprzerywalna. Uwzględniając istnienie prognozy napływu zgłoszeń regulamin ten oznaczać będziemy LIFO (ϑ).

W rozpatrywanym przypadku zgłoszenia główne mają więc czas obsługi B_S, natomiast zgłoszenia dodatkowe czas obsługi B_L.

Zauważmy przy tym, że przy przyjętej postaci zależności $B(x/w)$ obsługę zgłoszenia dodatkowego można podzielić na dwie fazy:

i) fazę pierwszą, o czasie trwania $B_L - B_S$. Wszystkie zgłoszenia które napłyną w tym czasie są tracone.

ii) fazę drugą, o czasie trwania B_S. Tylko jedno spośród zgłoszeń, które napłyną w tej fazie (ostatnie), zostanie obsłużone, pozostałe zostaną stracone.

W przypadku $\vartheta > 0$ może okazać się, że - zgodnie z podaną prognozą - prawdopodobieństwo napłynięcia zgłoszenia będzie zerowe przez znaczną część obsługi zgłoszenia dodatkowego.

Na rys. III.3 przedstawiono schemat ilustrujący przebieg zdarzeń w pomocniczym systemie obsługi, odpowiadającym rozpatrywanemu przypadkowi. Zdarzenia oznaczone etykietami reprezentują odpowiednio:

1 - moment rozpoczęcia obsługi zgłoszenia głównego,

2 - moment zwolnienia się stanowiska przy pustej kolejce $\mathcal{K}_1$,

3 - moment rozpoczęcia obsługi zgłoszenia dodatkowego,

4 - moment rozpoczęcia pierwszej obsługi zgłoszenia głównego po zakończeniu obsługi zgłoszenia dodatkowego lub po oczekiwaniu stanowiska obsługi na napłynięcie nowego zgłoszenia.

Średnią liczbę zgłoszeń $L(\lambda)$ traconych w okresie między kolejnymi momentami rozpoczęcia obsługi zgłoszeń dodatkowych wyznaczymy etapami:

W pierwszym kroku wyznaczymy średnią liczbę zgłoszeń traconych pomiędzy kolejnymi zdarzeniami 4 i 2, którą oznaczymy L_1,

$$L_1 = 0P_2 + 1_2(1 - P_2) \cdot P_2 + 21_2(1 - P_2)^2 P_2 + \ldots$$

$$= P_2 1_2 \sum_{i=1}^{\infty} i(1 - P_2)^i = 1_2 \frac{1 - P_2}{P_2}. \tag{III.1.14}$$

$$L_1 = \frac{\lambda B_S - 1 + e^{-\lambda B_S}}{e^{-\lambda B_S}} \tag{III.1.15}$$

W identyczny sposób można też określić średnią liczbę zgłoszeń L_2 traconych pomiędzy kolejnymi zdarzeniami 2 i 3:

$$L_2 = L_1 \frac{1 - P_3}{P_3}. \qquad \text{(III.1.16)}$$

Łączna więc średnia liczba zgłoszeń traconych $L(\lambda)$ wynosi:

$$L(\lambda) = (1 - P_1)\left[l_1 + L_1 + L_2\right] + P_1 L_2 + l_o =$$

$$= L_2 + (1 - P_1)(l_1 + L_1) + l_o. \qquad \text{(III.1.17)}$$

Zależność (III.1.17) wykorzystamy w warunku stabilności (III.1.12), specyfikując ją oddzielnie dla przypadku $\vartheta \leq B_L - B_s$ oraz $\vartheta > B_L - B_s$

i) Niech $\vartheta \leq B_L - B_s$

W tym przypadku zachodzi $m = B_s$, a więc $l_1 = l_2$ oraz $P_1 = P_2$, a także $l_o = \lambda(B_L - B_s - \vartheta)$.
Stąd, po przekształceniach otrzymujemy:

$$L(\lambda) = l_1 \frac{1 - P_1}{P_1} \frac{1}{P_3} + \lambda(B_L - B_s - x). \qquad \text{(III.1.18)}$$

Warunek stabilności ma więc postać:

$$l_1(1 - P_1) + P_1 P_3 \lambda(B_L - B_s - \vartheta) < P_1 P_3, \qquad \text{(III.1.19)}$$

i ostatecznie:

$$\lambda B_s + e^{-\lambda B_s}\left[1 - e^{-\lambda(B_L - B_s)}\right] < 1. \qquad \text{(III.1.20)}$$

ii) Niech $\vartheta > B_L - B_s$

W tym przypadku zachodzi $m = B_L - \vartheta$ oraz $l_o = 0$, gdzie m jest jedynym fragmentem obsługi zgłoszenia dodatkowego (drugiej fazy), w czasie którego mogą napływać nowe zgłoszenia. Z (III.1.17) wynika warunek:

$$L(\lambda) = l_2 \frac{1-P_2}{P_2} \frac{1-P_3}{P_3} + (1-P_1)(l_1 + l_2 \frac{1-P_2}{P_2}) < 1. \qquad \text{(III.1.21)}$$

Po przekształceniach otrzymujemy stąd:

$$\lambda B_s + e^{-\lambda B_s}(1 - e^{-\lambda\vartheta}) + e^{-\lambda B_L}(1 - \lambda B_s) -$$

$$- e^{-\lambda B_s} e^{-\lambda\vartheta}\left[1 - \lambda(B_L - \vartheta)\right] < 1, \qquad \text{(III.2.22)}$$

dla $\vartheta > B_L - B_s$.

d) <u>Inne warunki stabilności</u>

Warunki stabilności (III.1.20) oraz (III.1.22) nie są wynikami oryginalnymi. Zostały one podane po raz pierwszy w pracy [BUZA 74] - jedynej jak dotąd - poświęconej systemom WAD z różnymi od naturalnego regulaminami obsługi.
Rozpatrywano tam jednokanałowy system WAD, ze zgłoszeniami napływającymi strumieniem Poissona oraz zależnością czasu obsługi od czasu oczekiwania określoną wzorem (III.1.13) przy dwóch regulaminach obsługi: FIFO oraz LIFO i dodatkowym założeniu, że przed wprowadzeniem do obsługi zgłoszenia wymagającego obsługi "długiej" (o czasie trwania B_L) analizowana jest prognoza napływu zgłoszeń w pewnym horyzoncie ϑ, a w przypadku prognozy pozytywnej wprowadza się odstępstwo od zasadniczego regulaminu obsługi, wymuszając oczekiwanie stanowiska obsługi na nadejście nowego zgłoszenia. Regulaminy te oznaczać będziemy FIFO (ϑ) oraz LIFO (ϑ). Zauważmy, że regulamin LIFO (ϑ) jest identyczny z przypadkiem rozpatrywanym poprzednio.

W cytowanej pracy przeprowadzono analizę wspomnianych wyżej dwóch regulaminów obsługi (poprzedzoną oddzielnym potraktowaniem przypadków FIFO ≡ FIFO(0) oraz LIFO ≡ LIFO(0)) metodą wymuszonych łańcuchów Markowa, co było możliwe, aczkolwiek pracochłonne, ze względu na prostą strukturę przestrzeni stanów, wynikającą z postaci zależności (III.1.13).
Wyznaczono również warunki stabilności postaci $\lambda < T_R$, gdzie wartości wydajności granicznych T_R są pierwiastkami równań zawartych w tabeli III.1. Jak to pokazano w przypadku wyżej regulaminu LIFO(ϑ), identyczne wyniki można uzyskać w prostszy sposób, bezpośrednio[x)] bez prowadzenia pełnej analizy systemu.

Przedstawioną przez nas metodę cechuje przy tym większa ogólność, co umożliwia wyjście na drodze analitycznej poza wymienione w tabeli III.1 regulaminy obsługi, jedyne, jak dotąd, dla których warunki stabilności zostały określone.

---

x) Dla ścisłości należy wspomnieć, iż Buzacott zauważył w wyprowadzonych warunkach prawidłowość pozwalającą je interpretować jako wymaganie, by średnia liczba zgłoszeń napływających do systemu w czasie między chwilami zakończenia dwóch kolejnych obsług (pod warunkiem obecności w systemie chociaż jednego zgłoszenia o czasie oczekiwania większym niż τ) była mniejsza od jedności. Taka (nie udowodniona) prawidłowość jest jednak też dość trudna do stosowania.

Tabela III.1

Warunki stabilności dla czterech systemów WAD (wg [BUZA 74])

$\mathcal{R}$	
FIFO(0)	$TB_I = 1$
LIFO(0)	$TB_s + (B_L - B_s)e^{-TB_s} = 1$
FIFO(ϑ)	$TB_s + 1 - e^{-T\vartheta} + e^{-T\vartheta}(B_L - B_s - \vartheta) = 1$ dla $0 \leq \vartheta \leq B_L$
LIFO(ϑ)	$TB_s+e^{-T\vartheta}e^{-TB_s}(B_L-B_s-\vartheta)T+e^{-TB_s}(1-e^{-T\vartheta})=1$ dla $0 \leq \vartheta \leq B_L-B_s$
	$TB_s+e^{-TB_s}(1-e^{-TB_s})+e^{-TB_L}(1-TB_s)-e^{-TB_s}e^{-T\vartheta}\left[1-T(B_L-\vartheta)\right] = 1$ dla $B_L-B_s \leq \vartheta \leq B_L$

W punkcie III.3 wyznaczymy wartości $L(\lambda)$ dla pewnych, dalszych przypadków.

Można łatwo sprawdzić, że wydajności graniczne T_R spełniają następujące relacje:

$$\forall_{\vartheta \in [0, B_L]} \; T_{REG(\vartheta)} = T_{REG(B_L-B_s)}, \tag{III.1.23}$$

$$\forall_{\vartheta \in [0, B_L]} \; T_{REG(\vartheta)} > T_{REG(0)}, \tag{III.1.24}$$

$$\forall_{\vartheta_1, \vartheta_2 \; [0, B_L]} \; T_{LIFO(\vartheta_1)} > T_{FIFO(\vartheta_2)}, \tag{III.1.25}$$

gdzie $REG(\vartheta) = LIFO(\vartheta), FIFO(\vartheta)$.

Tabela III.2

Wartości wydajności granicznej różnych regulaminów obsługi w przykładowym systemie obsługi (wg [BUZA 74])

$\mathcal{R}$	TB_s
FIFO(0)	0.500
FIFO (B_L-B_s)	0.568
LIFO(0)	0.659
LIFO (B_L-B_s)	0.751

Zależność (III.1.23) określa optymalny horyzont prognozy dla rozpatrywanych regulaminów obsługi, nierówność (III.1.24) pozwala stwierdzić, że uwzględnienie prognozy zawsze prowadzi do zwiększenia wydajności granicznej, natomiast nierówność (III.1.25) stwierdza, że każdy regulamin bazujący na odwrotnym (w stosunku do kolejności napływania zgłoszeń) porządku obsługi cechuje większa wydajność graniczna aniżeli dowolny z regulaminów bazujących na regulaminie naturalnym.

Przykładowe relacje ilościowe przy parametrach systemu $B_L = 2B_s$ przedstawia zaczerpnięta z [BUZA 74] tabela III.2.

Będziemy mówili, że regulamin $\mathcal{R}_1$ jest lepszy od regulaminu $\mathcal{R}_2$, jeśli:

$$T_{R_1} \geq T_{R_2} \tag{III.1.26}$$

W rozpatrywanym przypadku regulamin naturalny jest więc zdecydowanie najgorszy, przy czym zjawisko to nie jest specjalną własnością klasy systemów zdefiniowanych w punkcie II.1.a.

Rozważmy dowolny system podstawowy, dla którego określimy:

$$\left.\begin{aligned} E(b_k) &= \min_{i \in 1,2,\ldots,J} E(b_i) \\ E(b_1) &= \max_{i \in 1,2,\ldots,J} E(b_i) \end{aligned}\right\} \tag{III.1.27}$$

W przypadku regulaminów obsługi nie dopuszczających bezczynności stanowiska obsługi przy niepustej kolejce zachodzi:

$$E(b_k) \leq \eta^*(\lambda) \leq E(b_1), \tag{III.1.28}$$

a ponieważ

$$\lim_{\lambda \to T_R^{-1}} \eta(\lambda) = 1, \tag{III.1.29}$$

więc:

$$E(b_1)^{-1} \leq T_R \leq E(b_k)^{-1}. \tag{III.1.30}$$

W szczególności, jeżeli wartość oczekiwana czasu obsługi jest niemalejącą funkcją czasu oczekiwania, a więc jeśli zachodzi:

$$E(b_i) \leq E(b_{i+1}), \quad i=0,1,\ldots,J-1, \tag{III.1.31}$$

to regulamin naturalny jest zawsze najgorszy w klasie regulaminów obsługi nie dopuszczających bezczynności stanowiska obsługi przy niepustej kolejce, ponieważ:

$$T_R \geqslant E(b_1)^{-1} = E(b_J)^{-1} = T_{FIFO}. \qquad \text{(III.1.32)}$$

Nierówność (III.1.30) podaje też maksymalną, teoretycznie możliwą do osiągnięcia wydajności graniczną, przy czym - jak to widać wprost w tabeli III.2 - nawet najlepszy z rozpatrywanych regulaminów obsługi cechowała wydajność graniczna niższa o ok. 13% od maksymalnej.

Przytoczone dane pozwalają na stwierdzenie, że drogą doboru regulaminu obsługi można istotnie zwiększyć wydajność graniczną systemu z czasem obsługi zależnym od czasu oczekiwania w stosunku do wartości uzyskiwanych przy stosowaniu regulaminu naturalnego.

Na szczególne podkreślenie zasługuje fakt, że w przeciwieństwie do konwencjonalnych systemów obsługi celowe, z punktu widzenia zwiększenie wydajności granicznej, okazuje się w niektórych przypadkach utrzymanie stanowiska obsługi w stanie bezczynności mimo niepustej kolejki.

Do problematyki doboru regulaminów obsługi pod kątem widzenia maksymalizacji wydajności granicznej wrócimy jeszcze szerzej w punkcie III.5.

e) <u>Dodatek</u>

Dowód warunku koniecznego ergodyczności łańcuchów Markowa.

Niech:

a) $\{X_m\}$ będzie łańcuchem Markowa z dyskretnym czasem, nieprzywiedlnym, nieokresowym, przyjmującym wartości całkowite, nieujemne, o macierzy przejść $[P_{ij}]$,

b) γ_i będzie określone

$$\gamma_i = E\{X_{m+1} - X_m / X_m = i\} = \sum_{j=1}^{\infty} jP_{ij} - i,$$

c) ciąg $\{\gamma_i\}$ składa się z r podciągów, monotonicznie malejących, zbieżnych odpowiednio do granic $\delta_1, \delta_2, \ldots, \delta_r$ (każdy element ciągu $\{\gamma_i\}$ należy do jednego z tych podciągów).

Udowodnimy, że warunkiem koniecznym ergodyczności $\{X_m\}$ jest, aby:

$$\lim_{i \to \infty} \inf \gamma_i < 0 \quad .$$

<u>Dowód</u>

Dla łańcuchów Markowa spełniających warunek a), ergodycznych, zachodzi [KARL 68]:

$$\pi_j = \sum_{i=0}^{\infty} \pi_i P_{ij}, \qquad \text{(III.1.33)}$$

$$\sum_{i=0}^{\infty} \pi_i = 1, \qquad \text{(III.1.34)}$$

$$\pi_i > 0, \qquad \text{(III.1.35)}$$

gdzie π_i - prawdopodobieństwo pobytu w stanie "i" w stacjonarnym stanie granicznym.

Mnożąc obie strony (III.1.33) przez z^j i sumując otrzymujemy:

$$\sum_{j=0}^{\infty} z^j \pi_j = \sum_{j=0}^{\infty} z^j \sum_{i=0}^{\infty} \pi_i P_{ij},$$

i po przekształceniach:

$$\sum_{j=0}^{\infty} z^j \pi_j = \sum_{i=0}^{\infty} \pi_i \sum_{j=0}^{\infty} z^j P_{ij}.$$

Różniczkując obustronnie względem z i kładąc z = 1 otrzymujemy:

$$\sum_{j=1}^{\infty} j\pi_j = \sum_{i=0}^{\infty} \pi_i \sum_{j=1}^{\infty} jP_{ij} \qquad \text{(III.1.36)}$$

Zmieniając wskaźnik sumowania lewej strony mamy:

$$\sum_{i=1}^{\infty} i\pi_i = \sum_{i=0}^{\infty} \pi_i \sum_{j=1}^{\infty} jP_{ij},$$

po uporządkowaniu

$$\sum_{i=0}^{\infty} \pi_i \left(\sum_{j=1}^{\infty} jP_{ij} - i \right) = 0,$$

czyli:

$$\sum_{i=0}^{\infty} \pi_i \gamma_i = 0. \qquad \text{(III.1.37)}$$

Zauważmy jednak, że:

$$\sum_{j=1}^{\infty} j P_{oj} > 0, \qquad \text{(III.1.38)}$$

a więc:

$$\gamma_0 > 0.$$

Z drugiej strony na mocy (III.1.35) po to, aby możliwe było spełnienie równania (III.1.37), musi zachodzić:

$$\exists_{i \in [1,\infty)} \left\{ \gamma_i < 0 \right\}. \qquad \text{(III.1.39)}$$

Jeśli istnieje chociaż jedno i, dla którego zachodzi (III.1.39), to na mocy warunku c) γ_i jest elementem ciągu zdążającego do granicy σ_1, takiej, że $\sigma_1 \leq \gamma_i$.

Widzimy więc, że warunkiem koniecznym spełnienia (III.1.37), a więc i ergodyczności łańcucha $\{X_n\}$ jest, aby zachodziło:

$$\lim_{i \to \infty} \inf \gamma_i < 0 \quad \text{c.b.d.o.}$$

Uwaga

Udowodniona własność stanowi dyskretną analogię twierdzenia 5 podanego w erracie do [PUJO 78] (jako konsekwencja twierdzenia udowodnionego w [TWEE 75]) dla przypadku, gdy łańcuch zdefiniowany jest na ciągłej przestrzeni stanów i prawdopodobieństwa przejść są funkcjami ściśle ciągłymi.

III.2. Metoda szacowania wydajności granicznej w oparciu o modelowanie cyfrowe

Konwencjonalne systemy masowej obsługi, a także systemy niekonwencjonalne, zachowujące pracę, mają tę dogodną cechę, że znane są ich warunki stabilności nawet dla przypadków najogólniejszych G/G/n, dla których można udowodnić brak istnienia dokładnych rozwiązań analitycznych określających ich charakterystyki.

W związku z tym, przy spełnieniu warunków stabilności, charakterystyki systemu można w każdym przypadku określić drogą modelowania cyfrowego, które jest też bezcennym narzędziem dla weryfikacji i oceny dokładności przybliżonych metod analizy. Jak to przedstawiono w rozdziale II, podobnie przedstawia się sytuacja w zakresie systemów WAD z regulaminem naturalnym (gdzie zresztą modelowanie cyfrowe jest jedyną dostępną metodą badania systemów wielokanałowych).

Ponieważ zasadniczą wadą modelowania cyfrowego jest duża czasochłonność, wiele uwagi poświęcono zagadnieniom sposobu organizacji eksperymentów symulacyjnych w systemach stabilnych, tak aby zmniejszyć nakład obliczeń niezbędny dla wyznaczenia poszukiwanych charakterystyk w stacjonarnym stanie granicznym z zadaną dokładnością (z odpowiednio wąskim przedziałem ufności).

W szczególności rozstrzygnięcia wymagają dwa problemy:

- sposób uwzględnienia zależności losowej kolejnych obserwacji (np. czasów oczekiwania kolejnych zgłoszeń), której istnienie skłania do użycia metod analizy szeregów czasowych (patrz np. [BOXG 70]) w miejsce standardowych metod statystyki matematycznej,
- sposób wyboru punktu startowego. W starszych pracach z reguły zalecono dla wyznaczenia charakterystyk systemów stabilnych pominięcie początkowego odcinka symulacji w celu wyeliminowania stanu nieustalonego (i przyjęcie następnie dowolnego punktu startowego). Wielokrotnie wskazywano przy tym na niską efektywność wszystkich znanych kryteriów osiągnięcia stacjonarnego stanu granicznego.

Aktualnie wiele badań koncentruje się na rozpatrzeniu własności alternatywnego podejścia polegającego na rozpoczynaniu symulacji od pewnego odpowiednio wybranego stanu systemu (co pozwala na efektywne wykorzystanie całego przebiegu symulacji).

W ostatnich latach rozwinięto w licznych pracach zapoczątkowanych przez [CRAN 74] metodę ominięcia obu wyżej wymienionych trudności drogą pogrupowania obserwacji systemu w zbiory niezależne losowo oraz charakteryzujące się jednakowymi rozkładami.

Grupowanie takie jest możliwe, gdy proces stochastyczny $X_p(t)$, opisujący zapas pracy w systemie, jest procesem regenerującym się [KOPO 73A], tzn. charakteryzującym się istnieniem pewnego zdarzenia $\mathcal{E}$ (tzw. zdarzenia rekurencyjnego) takiego, że jeśli w jakimś momencie zdarzenie $\mathcal{E}$ zajdzie, to dalszy przebieg procesu po tym momencie nie zależy od historii tego procesu, przy czym okresy czasu pomiędzy dwoma kolejnymi zajściami zdarzenia $\mathcal{E}$ (momentami regeneracji) są zmiennymi losowymi o jednakowych rozkładach (niezależnymi).

Oczywiście, aby zagwarantować efektywność metody opartej na takim grupowaniu (skończony czas trwania eksperymentu), trzeba założyć, że z prawdopodobieństwem jeden zdarzenie $\mathcal{E}$ musi pojawić się nieskończenie wiele razy (jest to tzw. zdarzenie rekurencyjne powracające), a odstępy pomiędzy momentami regeneracji mają skończoną wartość oczekiwaną.

Dla przeważającej większości stabilnych systemów masowej obsługi momenty rozpoczynania kolejnego okresu czynnego (tj. momenty nadejścia zgłoszenia do systemu, w którym kolejka jest pusta a wszystkie kanały obsługi bezczynne) są momentami regeneracji o wymienionych wyżej własnościach. Na przykład w pracy [WHIT 72] wykazano, że wielokanałowe, konwencjonalne systemy masowej obsługi typu GI/G/n posiadają tę własność, o ile zachodzi $\Pr\{a^i > b^{i-1}\} > 0$, co jest wymaganiem bardzo ogólnym, aczkolwiek silniejszym aniżeli warunek stabilności systemu.
Można łatwo sprawdzić, że zachodzi to np. dla wszystkich systemów stabilnych, w których dystrybuanta A(x) posiada gęstość a(x) taką, że $\forall_{x>0} A(x)>0$ na mocy identycznego rozumowania jak przytoczone w punkcie III.2.

Wprost z definicji ergodycznych łańcuchów Markowa z ciągłą przestrzenią stanów [TWEE 75] wynika, że takie stwierdzenie jest też prawdziwe w każdym przypadku, gdy Z_R^i jest ergodycznym łańcuchem Markowa, o ile tylko spełnione są pewne warunki nieprzywiedlności (w szczególności nieprzywiedlność względem dowolnej miary posiadającej pojedynczy atom przy $\mathcal{Z}(R, z^i - 0) = 0$).

Z drugiej strony, jeśli przedmiotowy łańcuch Markowa (przy spełnieniu powyższych warunków) jest chwilowy, to z pewnym prawdopodobieństwem $\alpha > 0$ kolejny moment regeneracji nie nastąpi w skończonym czasie, to znaczy, że z prawdopodobieństwem α okres czynny nie zakończy się w skończonym czasie (własność ta zostanie przez nas dalej wykorzystana).

Omówienie sposobu wykorzystania własności procesów regenerujących się dla symulacji stabilnych systemów masowej obsługi (ze szczególnym uwzględnieniem sposobu wyznaczania przedziałów ufności) można znaleźć w wielu publikacjach. W biuletynach wydawniczych zasygnalizowano też wydanie pozycji monograficznych [CRAM 77], [IGLE 80].

W przypadku systemów WAD z dowolnym regulaminem obsługi (a także wielu innych systemów niekonwencjonalnych, jak np. systemów priorytetowych z uwzględnieniem pracy własnej) sytuacja jest znacznie trudniejsza, bowiem nie można - ogólnie rzecz biorąc - podać warunków stabilności na drodze analitycznej.
Oczywiste jest, że bezpośrednie zamodelowanie pracy badanego systemu obsługi (przy pewnej zadanej intensywności strumienia zgłoszeń) i obserwacja jego parametrów w skończonym czasie nie mogą być podstawą do orzeczenia czy system ten jest stabilny, czy też nie.

Sprawą podstawowej wagi jest więc określenie takiego eksperymentu symulacyjnego, prowadzonego na przykład w oparciu o pewną modyfikację systemu badanego, który pozwoliłby w skończonym czasie oszacować (z odpowiednim poziomem ufności) wydajność graniczną badanego systemu.

Do niedawna znana była w literaturze tylko jedna praca [LAVE 78] (omawiana w rozdziale I), której wyniki mogły posłużyć dla organizacji takiego eksperymentu. W szczególności w odniesieniu do systemów jednostanowiskowych wydajność graniczną (przy pewnych założeniach) można wyznaczać jako intensywność strumienia zgłoszeń opuszczających stanowisko obsługi

przy niepustej kolejce. Nie wdając się w szczegółowe wymienianie przypadków, w których założenia twierdzeń Lavenberga nie są spełnione, z istoty eksperymentu je wykorzystującego, wynika brak możliwości stosowania takiego podejścia w przypadku systemów, w których sposób obsługi jest uzależniony od parametrów związanych z pobytem zgłoszeń w kolejce, a w szczególności w przypadku (ogólnym) występowania zależności czasu obsługi od czasu oczekiwania. Warto zauważyć, że wprowadzony w punkcie III.1 sposób wyznaczania warunku stabilności pewnej klasy systemów przez wyznaczanie strumienia zgłoszeń traconych w systemie pomocniczym jest specyficzną analogią twierdzeń Lavenberga i - jak to wskazano - może też być podstawą dla określenia efektywnego eksperymentu symulacyjnego.

Poniżej przedstawimy efektywną metodę (stanowiącą rozwinięcie propozycji zawartej w [WOLI 81B]), dostarczającą oszacowanie wydajności granicznej dla szerokiej klasy systemów masowej obsługi, posiadających omówiony wyżej charakter procesu regenerującego się w momentach zakończenia bezczynności stanowiska obsługi.

Po sprecyzowaniu rozpatrywanej klasy systemów przedstawimy zasadę omawianej metody na przykładzie jednokanałowych systemów ze zgłoszeniami napływającymi strumieniem Poissona, a następnie udowodnimy poprawność jej uogólnień.

W dalszym ciągu przedstawimy sposób określenia wiarygodności wyników, własności oszacowania, zalecany sposób organizacji eksperymentu symulacyjnego a także przykłady obliczeniowe.

a) <u>Określenie rozpatrywanej klasy systemów</u>

W dalszych rozważaniach ograniczymy się do przypadku jednostanowiskowych, wielokanałowych systemów obsługi z nieograniczonym buforem o następujących własnościach:
- Zgłoszenia napływają pojedynczo i zmienne losowe $\{a^i\}$ są niezależne losowo, o jednakowym rozkładzie i intensywności strumienia zgłoszeń λ;
- Czasy obsługi zgłoszeń $\{b^i\}$ są zmiennymi losowymi takimi, że $0<b^i<\infty$, $i = 0,1,\ldots$; być może zależnymi losowo, zależnymi od czasu oczekiwania zgłoszenia itp.;
- Sposób obsługi zgłoszeń jest jednoznacznie określony (być może z uwzględnieniem pracy własnej na implementację regulaminu obsługi, czasową niedostępnością stanowiska obsługi itp.) tak, że możliwe jest zasymulowanie systemu;
- Jednostkowy czas obsługi zgłoszenia $\eta^*(\lambda)$ (a tym samym i stopień wykorzystania stanowiska obsługi $\eta(\lambda)$ oraz pojedynczego kanału obsługi $\eta^i(\lambda)$) jest niemalejącą funkcją λ.
- Oznaczmy przez $P_e(\lambda)$ prawdopodobieństwo tego, że system jest bezczynny, natomiast przez $e^i(\lambda) = 1 - \eta^i(\lambda)$ prawdopodobieństwo tego, że pojedynczy kanał obsługi jest bezczynny.

Przyjmiemy, że stanowisko obsługi jest jednorodne, a kanały obsługi obciążone jednakowo, tj. $e^{i}(\lambda) = e(\lambda)$, $i = 1,2,\ldots,n$ oraz że zachodzi zwykła w systemach wielokanałowych relacja

$$\left[e(\lambda)\right]^{n} \leq P_{o}(\lambda), \qquad \text{(III.2.1)}$$

gdzie oczywiście przy $n = 1$ mamy $P_{o}(\lambda) = \mathcal{E}(\lambda)$;

- Jeśli stanowisko obsługi stanie się bezczynne w chwili t, to pozostanie bezczynne do chwili napłynięcia najbliższego zgłoszenia;
- Istnieje wartość wydajności granicznej T taka, że dla wszystkich $\lambda < T$ wartość oczekiwana czasu trwania okresu czynnego jest skończona, jeśli natomiast $\lambda > T$, to z pewnym prawdopodobieństwem $\alpha > 0$ okres czynny jest nieskończenie długi.

Oznaczmy przez $\rho_{K}(\lambda)$ prawdopodobieństwo tego, że w czasie pojedynczego okresu czynnego zostanie wykonanych K obsług, przy intensywności strumienia zgłoszeń λ, natomiast

$$\pi_{K}(\lambda) = \sum_{i=K}^{\infty} \rho_{i}(\lambda). \qquad \text{(III.2.2)}$$

Zakładamy więc, że dla rozpatrywanej klasy systemów obsługi istnieje takie $T > 0$ (wydajność graniczna), że dla każdego $\lambda < T$ system jest stabilny i wartość oczekiwana

$$\Theta(\lambda) = \sum_{i=1}^{\infty} i\rho_{i}(\lambda) \qquad \text{(III.2.3)}$$

jest skończona.

Przeciwnie, dla $\lambda > T$ system jest niestabilny, z prawdopodobieństwem $\alpha > 0$ liczba obsług w okresie czynnym jest nieskończenie wielka, a więc i wartość $\Theta(\lambda)$ jest nieskończona.

Oznaczmy przez $\psi(\lambda)$ prawdopodobieństwo tego, że kolejne nadchodzące zgłoszenie zapoczątkuje nowy okres czynny (a więc zastanie stanowisko obsługi bezczynne). Zachodzi przy tym:

$$\psi(\lambda) = \left[\Theta(\lambda)\right]^{-1}. \qquad \text{(III.2.4)}$$

b) <u>System jednokanałowy ze zgłoszeniami napływającymi strumieniem Poissona</u>

Przedstawimy obecnie metodę szacowania wydajności granicznej, ograniczając się chwilowo do jednokanałowych systemów masowej obsługi ze zgłoszeniami napływającymi strumieniem Poissona. Zdefiniujmy elementarny eksperyment symulacyjny, polegający na zamodelowaniu w rozpatrywanym systemie pojedynczego okresu czynnego, ograniczonego co najwyżej do K obsług, przy założeniu intensywności strumienia zgłoszeń równej pewnej, arbitralnie przyjętej wartości Λ. Taki elementarny eksperyment musi zakończyć się w skończonym czasie, na jeden z następujących sposobów:

i) zakończeniem okresu czynnego po wykonaniu liczby obsług mniejszej aniżeli K i będziemy wówczas mówili o negatywnym wyniku eksperymentu elementarnego,
ii) przerwaniem symulacji po wykonaniu K obsług w ciągu pojedynczego okresu czynnego i będziemy wówczas mówili o pozytywnym wyniku eksperymentu elementarnego.

Opisany elementarny eksperyment należy powtórzyć "m"-krotnie dla tej samej wartości Λ (wyniki kolejnych eksperymentów elementarnych są oczywiście losowo niezależne) i na tej podstawie określić oszacowanie $\tilde{\pi}_{K}(\Lambda)$ prawdopodobieństwa $\pi_{K}(\Lambda)$ określonego przez (III.2.2). Sposób wyznaczenia tego oszacowania omówimy szczegółowo poniżej.

Odwołamy się teraz do znanej z rachunku prawdopodobieństwa nierówności Markowa (np. [KLEI 75]), która orzeka, że dla dowolnej zmiennej losowej x, posiadającej wartość oczekiwaną E(x), zachodzi

$$\Pr\left\{ x \geq \xi \right\} \leq \frac{E(x)}{\xi}, \quad \xi > 0. \qquad \text{(III.2.5)}$$

W oparciu o tę nierówność testować będziemy hipotezę, że przy intensywności strumienia zgłoszeń Λ średnia liczba zgłoszeń obsłużonych w okresie czynnym jest równa $c > 1$, a więc, że zachodzi

$$\tilde{\pi}_{K}(\Lambda) \leq \frac{c}{K}. \qquad \text{(III.2.6)}$$

Oczywiście, nierówność (III.2.6) jest spełniona z definicji (dla dowolnego K) w przypadku gdy $c = \Theta(\Lambda)$, przy czym rzecz jasna skończona lub nieskończona wartość $\Theta(\Lambda)$ jest nieznana.

Niech c^{*} oznacza dowolną wartość c, dla której testowana hipoteza została odrzucona.

Ponieważ

$$c^{*} < \Theta(\Lambda), \qquad \text{(III.2.7)}$$

więc uzyskaliśmy pewne oszacowanie wartości $\Theta(\Lambda)$, które będzie obarczone najmniejszym błędem, gdy wybierzemy $c^{*} = c_{o}$, gdzie c_{o} jest największą liczbą rzeczywistą, dla której nierówność (III.2.6) nie jest spełniona.

Przedyskutujmy obecnie możliwość oszacowania wydajności granicznej rozpatrywanego systemu w oparciu o znajomość pary wartości $\langle \Lambda, \Theta(\Lambda) \rangle$.

Dla systemów jednokanałowych ze zgłoszeniami napływającymi strumieniem Poissona zachodzi:

$$\left[\Theta(\lambda)\right]^{-1} = \psi(\lambda) = P_0(\lambda) = e(\lambda) = 1 - \eta(\lambda), \qquad \text{(III.2.8)}$$

a więc

$$\eta(\Lambda) = 1 - \frac{1}{\Theta(\Lambda)}. \qquad \text{(III.2.9)}$$

Gdyby rozpatrywany system obsługi był systemem konwencjonalnym, wówczas (patrz rozdz. I.1) zachodziłoby[x]:

$$\eta_{\mathcal{K}}(\lambda) = \frac{\lambda}{T} \quad \text{dla} \quad \lambda < T$$
$$\eta_{\mathcal{K}}(\lambda) < \frac{\lambda}{T} \quad \text{dla} \quad \lambda \geq T \qquad \text{(III.2.10)}$$

Ponieważ nie wiemy czy stosowana dla eksperymentu intensywność strumienia zgłoszeń Λ jest większa, czy mniejsza od wydajności granicznej T, musimy przyjąć, że:

$$\eta(\Lambda) \leq \frac{\Lambda}{T}. \qquad \text{(III.2.11)}$$

Łatwo sprawdzić, że nierówność (III.2.11) jest również prawdziwa dla systemów niekonwencjonalnych, o ile tylko $\eta^*(\lambda)$ jest niemalejącą funkcją λ.

Na podstawie (III.2.9) oraz (III.2.11) możemy więc powiedzieć, że:

$$\frac{\Lambda}{T} \geq 1 - \frac{1}{\Theta(\Lambda)}, \quad 0 < \Lambda < \infty \qquad \text{(III.2.12)}$$

gdzie równość występowałaby w przypadku systemów konwencjonalnych przy $\Lambda < T$, co prowadzi do oszacowania wydajności granicznej T postaci:

$$T \leq \frac{\Lambda}{1 - \frac{1}{\Theta(\Lambda)}}. \qquad \text{(III.2.13)}$$

Przeprowadzony eksperyment symulacyjny nie prowadził jednak do wyznaczenia wartości $\Theta(\Lambda)$, a tylko do jej oszacowania postaci (III.2.7), więc praktycznie można skorzystać z nierówności:

$$T < \frac{\Lambda}{1 - \frac{1}{c^*}}, \quad 1 < c^* \leq c_0, \quad 0 < \Lambda < \infty \qquad \text{(III.2.14)}$$

[x] Indeks $\mathcal{K}$ oznacza parametry systemu konwencjonalnego.

Ostatecznie więc, biorąc pod uwagę dokładność oszacowania (III.2.7) najlepsze możliwe do uzyskania tą drogą oszacowanie od góry $\tilde{T}$ wartości wydajności granicznej T ma postać:

$$\tilde{T} = \frac{\Lambda}{1 - \frac{1}{c_0}}, \quad 0 < \Lambda < \infty \qquad \text{(III.2.15)}$$

Należy podkreślić, iż dokładność oszacowania, tj. wartość różnicy $(\tilde{T} - T)$, zależy od wybranych dla eksperymentu parametrów Λ, K, m, których wpływ zostanie przedyskutowany nieco później.

c) <u>Uogólnienie oszacowania na przypadek wielokanałowy i dowolny strumień zgłoszeń</u>

Wykażemy obecnie możliwość uogólnienia podanego wyżej oszacowania na pełną klasę systemów masowej obsługi scharakteryzowanych w ustępie a).

Rozpatrzmy przypadek systemu wielokanałowego ze zgłoszeniami napływającymi strumieniem Poissona. Korzystając z nierówności (III.2.1) widzimy, że:

$$e(\lambda) \leq \sqrt[n]{P_0(\lambda)}, \qquad \text{(III.2.16)}$$

a więc równość (III.2.8) przechodzi w tym przypadku w nierówność

$$\frac{1}{\sqrt[n]{\Theta(\lambda)}} = \sqrt[n]{\psi(\lambda)} = \sqrt[n]{P_0(\lambda)} \geq e(\lambda) = 1 - \eta(\lambda). \qquad \text{(III.2.17)}$$

Jeżeli natomiast strumień zgłoszeń jest różny od strumienia Poissona, to z założonego faktu, iż stanowisko obsługi, które stanie się bezczynne w chwili t, pozostaje bezczynne aż do chwili nadejścia najbliższego zgłoszenia wynika, że

$$\psi(\lambda) \geq P_0(\lambda), \qquad \text{(III.2.18)}$$

a więc realcja analogiczna do (III.2.8) przybiera ostatecznie, w przypadku ogólnym, postać:

$$\frac{1}{\sqrt[n]{\Theta(\lambda)}} = \sqrt[n]{\psi(\lambda)} \geq \sqrt[n]{P_0(\lambda)} \geq e(\lambda) = 1 - \eta(\lambda), \qquad \text{(III.2.19)}$$

a stąd zamiast równości (III.2.9) otrzymujemy nierówność:

$$\eta(\Lambda) \geq 1 - \frac{1}{\sqrt[n]{\Theta(\Lambda)}}. \qquad \text{(III.2.20)}$$

Prowadząc dalsze rozumowanie analogicznie jak poprzednio, a w szczególności uwzględniając nierówność (III.2.11) oraz (III.2.7) uzyskujemy dla ogólnego przypadku oszacowanie T postaci:

$$\vec{T} = \frac{\Lambda}{1 - \frac{1}{\sqrt[3]{c_o}}}. \qquad \text{(III.2.21)}$$

d) Wyznaczanie wartości $\vec{\pi}_K(\lambda)$

Przejdziemy obecnie do szczegółowego omówienia sposobu statystycznej oceny wartości $\pi_K(\lambda)$ w oparciu o opisany powyżej eksperyment symulacyjny.

Każdy z elementarnych eksperymentów polegających na symulowaniu pojedynczego okresu czynnego (ograniczonego) kończy się wynikiem pozytywnym z prawdopodobieństwem $\pi_K(\lambda)$ i wynikiem negatywnym z prawdopodobieństwem $\left[1 - \pi_K(\lambda)\right]$, przy czym wyniki ich są niezależne losowo.

Niech $f(j,m,\pi_K(\lambda))$ oznacza prawdopodobieństwo tego, że j spośród m przeprowadzonych eksperymentów elementarnych skończy się wynikiem pozytywnym. Prawdopodobieństwo to określa rozkład Bernoulliego

$$f(j,m,\pi_K(\lambda)) = \binom{m}{j}\left[\pi_K(\lambda)\right]^j\left[1 - \pi_K(\lambda)\right]^{m-j}. \qquad \text{(III.2.22)}$$

W rozpatrywanym przypadku w wyniku modelowania cyfrowego uzyskuje się po przeprowadzeniu "m" eksperymentów elementarnych pewną liczbę "j" sukcesów i zadaniem naszym jest oszacować wartość $\pi_K(\lambda)$ w oparciu o znajomość "m" i "j".

Jak to wykazano np. w [SMIR 65], wartość:

$$E_\pi = \frac{j}{m} \qquad \text{(III.2.23)}$$

jest nieobciążonym, najbardziej efektywnym i asymptotycznie normalnym estymatorem wartości $\pi_K(\lambda)$.

Rozpatrzmy przedział ufności dla tego estymatora. Ponieważ poszukiwane oszacowanie $\vec{\pi}_K(\lambda)$ będzie użyte w nierówności (III.2.6), nie jesteśmy zainteresowani w uzyskaniu formowanego zwykle symetrycznego przedziału ufności, a przeciwnie, dążymy do wyznaczenia takiej wartości $\vec{\pi}_K(\lambda)$, dla której zachodziłoby:

$$\Pr\left\{\pi_K(\lambda) > \vec{\pi}_K(\lambda) = E_\pi - \delta\right\} = \gamma \qquad \text{(III.2.24)}$$

gdzie γ jest przyjętym poziomem ufności.

Wartości δ związane z poziomem ufności γ można wyznaczyć (patrz [SMIR 65]) w następujący sposób:

i) Jeśli spełniona jest nierówność

$$m\pi_K(\lambda)\left[1 - \pi_K(\lambda)\right] > 9, \qquad \text{(III.2.25)}$$

wówczas rozkład Bernoulliego może być, z praktycznie wystarczającą dokładnością, przybliżony rozkładem normalnym o wartości średniej $\pi_K(\lambda)$ i odchyleniu standardowym σ

$$\sigma = \sqrt{\frac{\pi_K(\lambda)\left[1 - \pi_K(\lambda)\right]}{m}}, \qquad \text{(III.2.26)}$$

posiadającym estymator E_σ

$$E_\sigma = \sqrt{\frac{E_\pi(1 - E_\pi)}{m}}. \qquad \text{(III.2.27)}$$

Ponieważ

$$\Pr\left\{\frac{E_\pi - \pi_K(\lambda)}{\sigma} \leq h\right\} = \Phi(h), \qquad \text{(III.2.28)}$$

gdzie:

$$\Phi(t) = \frac{1}{\sqrt{2\pi}}\int_{-\infty}^{t} e^{-\frac{y^2}{2}}\,dy;$$

więc

$$\Pr\left\{\pi_K(\lambda) > E_\pi - t_\gamma E_\sigma\right\} = \gamma, \qquad \text{(III.2.29)}$$

gdzie: $\Phi(t_\gamma) = \gamma$,

natomiast E_σ wyznacza się ze wzoru (III.2.27).

Dla przyjmowanego zwykle poziomu ufności $\gamma = 0{,}95$ zachodzi:

$$\Pr\left\{\pi_K(\lambda) > E_\pi - 1.645\ E_\sigma\right\} = 0{,}95. \qquad \text{(III.2.30)}$$

Ze względu na czas trwania eksperymentu dążymy do wykonania możliwie małej liczby eksperymentów elementarnych m_δ, zapewniającej uzyskanie pewnej zadanej wielkości przedziału ufności δ^* na poziomie ufności γ.

Wprost z (III.2.29) oraz (III.2.26) wynika, że m_δ musi spełniać nierówność:

$$m_\delta \geq \pi_K(\lambda)\left[1 - \pi_K(\lambda)\right]\left(\frac{t_\gamma}{\delta^*}\right)^2, \qquad \text{(III.2.31)}$$

a ponieważ:

$$\pi_K(\lambda)\left[1 - \pi_K(\lambda)\right] \leq \frac{1}{4}, \qquad \text{(III.2.32)}$$

więc wymaganą dokładność można zawsze uzyskać po przeprowadzeniu

$$m_\delta = 0.25\left(\frac{t_\gamma}{\delta^*}\right)^2 \qquad \text{(III.2.33)}$$

eksperymentów (oczywiście przy równoczesnym spełnieniu (III.2.25)).

ii) W przypadku gdy nierówność (III.2.25) nie jest spełniona, wówczas omówiona wyżej aproksymacja nie może być użyta i należy wykorzystać następującą procedurę:
Niech $S_{l,m}(\pi_K(\lambda))$ oznacza sumę

$$S_{l,m}(\pi_K(\lambda)) = \sum_{i=l}^{m} f(i,m,\pi_K(\lambda)). \qquad \text{(III.2.34)}$$

Dowodzi się [DUNI 55], że:

$$\Pr\left\{\pi_K(\lambda) > \xi\right\} = \gamma, \qquad \text{(III.2.35)}$$

gdzie ξ jest jedynym rozwiązaniem równania

$$S_{l,m}(\xi) = 1 - \gamma, \qquad \text{(III.2.36)}$$

leżącym w przedziale $[0,1]$.

W cytowanej pracy zamieszczono też tablice umożliwiające bezpośrednio wyznaczenie wartości ξ dla różnych m, l, γ.

Tak więc potrafimy w każdym przypadku uzyskać, z wymaganym poziomem ufności γ, oszacowanie $\tilde{\pi}_K(\Lambda)$, a biorąc pod uwagę ostre nierówności wykorzystywane przy wyprowadzeniu (III.2.21) można stwierdzić, iż $\tilde{T} > T$ na poziomie ufności wyższym niż γ.

e) Wpływ parametrów eksperymentu na wartość oszacowania $\tilde{T}$

Dokładność uzyskanego oszacowania $(\tilde{T} - T)$ zależy w sposób istotny od doboru parametrów eksperymentu Λ, K, m, których wpływ obecnie przedyskutujemy. Dla uproszczenia ograniczymy się chwilowo do przypadku jednokanałowego (n = 1).
Na wstępie przedyskutujemy zależność wartości $\tilde{T}$ od Λ, rozpatrując oddzielnie przypadek $\Lambda < T$ oraz $\Lambda > T$.
i) Załóżmy, że $\Lambda > T$. W tym przypadku, z prawdopodobieństwem $\alpha > 0$, na okres czynny składa się nieskończenie wiele obsług, czyli

$$\bigvee_{K>1} \pi_K(\Lambda) \geq \alpha > 0. \qquad \text{(III.2.37)}$$

Ponieważ jednak wartość c_o występującą we wzorze (III.2.21) wyznacza się jako:

$$c_o = \tilde{\pi}_K(\Lambda)K - \varepsilon \approx \pi_K(\Lambda)K - \varepsilon, \qquad \text{(III.2.38)}$$

gdzie ε jest małą dodatnią;
gdyż przy $m \to \infty$ wartość oszacowania $\tilde{\pi}_K(\Lambda) \to \pi_K(\Lambda)$, więc:

$$c_o > \alpha K - e_o, \qquad \text{(III.2.39)}$$

a co za tym idzie $c_o \to \infty$, gdy $K \to \infty$, $m \to \infty$.
Innymi słowy, jeśli tylko $\Lambda > T$, wówczas:

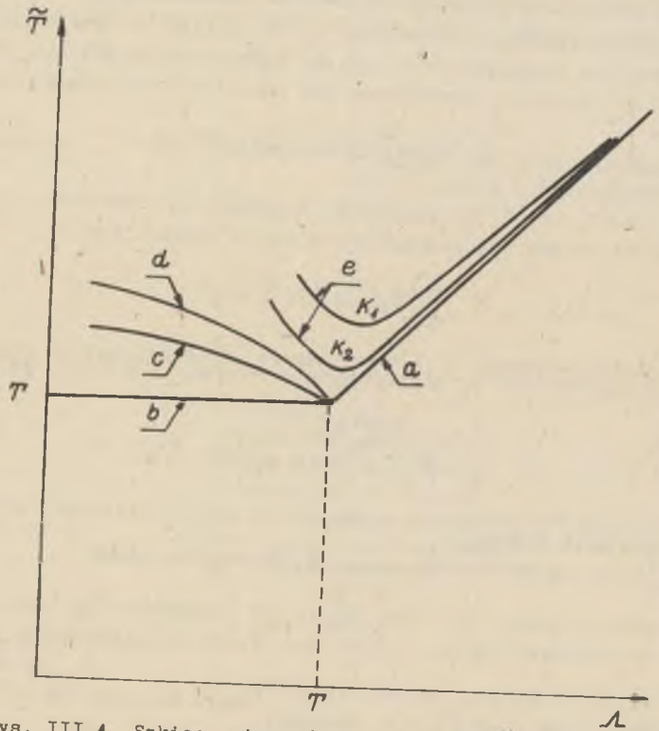
$$\lim_{K\to\infty} \frac{\Lambda}{1 - \frac{1}{c_o}} = \Lambda, \qquad \text{(III.2.40)}$$

czyli możliwe jest (teoretycznie rzecz biorąc) uzyskanie wartości oszacowania $\tilde{T} \to T$ przy $\Lambda \to T$ przez wartości większe od T.
ii) W przypadku $\Lambda < T$ zachodzi natomiast dla systemów ze zgłoszeniami napływającymi strumieniem Poissona (III.2.8)

$$\left[\Theta(\Lambda)\right]^{-1} = 1 - \eta(\Lambda). \qquad \text{(III.2.41)}$$

Tak więc, gdyby zachodziło $\Theta(\Lambda) = c_o$, wówczas mielibyśmy:

$$\tilde{T}^* = \frac{\Lambda}{\eta(\Lambda)}, \qquad \text{(III.2.42)}$$



Rys. III.4. Szkice zmienności oszacowania $\tilde{T}$ w funkcji Λ : a) wartości osiągane asymptotycznie (przy $K \to \infty$, $m \to \infty$) dla $\Lambda > T$, b) oszacowanie idealne $\tilde{T}^*_{\infty}$ w systemie konwencjonalnym, c) oszacowanie idealne $\tilde{T}^*$ w systemie niekonwencjonalnym ($\Theta(\lambda) = c_o$), d) oszacowanie $\tilde{T}$ przy $c_o < \Theta(\lambda)$, e) spodziewane rzeczywiste przebiegi zależności $\tilde{T}(\Lambda)$, $K_1 < K_2$

a w szczególności dla systemów konwencjonalnych, spełniających (III.2.10) byłoby

$$\bigvee_{\Lambda > T} \left\{ \tilde{T}^*_{\infty} = T \right\}. \qquad \text{(III.2.43)}$$

W przypadku systemów niekonwencjonalnych o niemalejącej funkcji $\eta^*(\lambda)$ zachodzi natomiast, jak łatwo sprawdzić:

$$\bigvee_{0 \le \Lambda_1 < \Lambda_2 < T} \left\{ \tilde{T}^*(\Lambda_1) \ge \tilde{T}^*(\Lambda_2) > T \right\}. \qquad \text{(III.2.44)}$$

Jeśli natomiast $c_o < \Theta(\Lambda)$ to oczywiście oszacowanie $\tilde{T}$ będzie dla każdej wartości Λ obarczone dodatkowym błędem w stosunku do omawianych wyżej war-

tości $\tilde{T}^*$. Dodatkowy błąd wprowadza też przyjęcie ogólniejszej postaci strumienia zgłoszeń.

Omawiane wyżej prawidłowości ilustruje rys. III.4.

Wartość oszacowania $\tilde{T}$ osiąga więc minimum (teoretyczne) równe T w punkcie $\Lambda = T + 0$. Wynika stąd bezpośrednio sugestia dotycząca doboru wartości Λ, przy których należy wyznaczać oszacowanie $\tilde{T}$. Celowe jest rozpoczęcie eksperymentów od dużych wartości Λ, przekraczających T (co łatwo sprawdzić, ponieważ powinny one prowadzić do $\tilde{T}(\Lambda) \approx \Lambda$), a następnie zmniejszając stopniowo Λ, dążyć do znalezienia minimalnej wartości $\tilde{T}$.

Celowe wydaje się przy tym, po zlokalizowaniu minimum, powtórzenie eksperymentu przy użyciu większych wartości K oraz m dla zwiększenia dokładności. Zauważmy, że uzyskana po przeprowadzeniu opisanego eksperymentu wartość Λ_o minimalizująca funkcje $\tilde{T}(\Lambda)$ jest również oszacowaniem od góry poszukiwanej wydajności granicznej, co więcej z postaci zależności (3.2.21) wynika, iż jest to przybliżenie dokładniejsze!

Z przedstawionych wyżej rozważań wynika jasno, że zarówno wzrost wartości K, jak i m wpływają na zwiększenie dokładności metody. Spróbujemy sformułować tę zależność w sposób bardziej precyzyjny.

W szczególności, korzystając z (III.2.29) oraz (III.2.27) można przedstawić $\tilde{\pi}_K(\lambda)$ w postaci:

$$\tilde{\pi}_K(\lambda) = G_1 - G_2 \frac{1}{\sqrt{m}}, \qquad \text{(III.2.45)}$$

gdzie G_1, G_2 - stałe zależne od parametrów modelowanego systemu i przyjętego poziomu ufności oraz K, a stąd:

$$\tilde{T} = \frac{1}{1 - \left[K \tilde{\pi}_K(\lambda)\right]^{-1}} = \frac{1}{1 - \left[K\left(G_1 - G_2 \frac{1}{\sqrt{m}}\right)\right]^{-1}} T. \qquad \text{(III.2.46)}$$

Zależność $\tilde{T}$ od m jest tu widoczna bezpośrednio. Zależność $\tilde{T}$ od K ma natomiast charakter bardziej złożony i ze względu na brak znajomości wpływu K na wartość $\pi_K(\lambda)$ (z wyjątkiem ogólnej prawidłowości, iż $\pi_K(\lambda)$ jest z definicji nie rosnącą funkcją K) musimy ograniczyć się do rozpatrywania wyłącznie dużych K takich, dla których zachodzi:

$$\bigvee_{\Lambda > T} \bigvee_{K_1 > K_2} \left\{ \pi_{K_1}(\Lambda) \approx \pi_{K_2}(\Lambda) \approx \alpha \right\} \qquad \text{(III.2.47)}$$

Przy tym założeniu zachodzi:

$$\tilde{T} = \frac{\Lambda}{1 - \frac{G_3}{K}}, \qquad \text{(III.2.48)}$$

gdzie G_3 - stała, zależna od parametrów systemu, poziomu ufności oraz m

Ponieważ czasochłonność modelowania wzrasta liniowo ze wzrostem m i co najwyżej liniowo (gdy spełnione jest (III.2.47) - w przeciwnym razie wzrost ten jest słabszy) ze wzrostem K, więc wydaje się bardziej celowe zwiększenie wartości K w celu zwiększania dokładności oszacowania $\tilde{T}$.
Biorąc pod uwagę postać (III.2.21) można łatwo stwierdzić, że wszystkie omówione wyżej prawidłowości pozostają ważne również dla systemów wielokanałowych. Należy jednak, ze względu na pierwiastkowanie wartości c_o, liczyć się z istotnym spadkiem dokładności oszacowania $\tilde{T}$ (uzyskiwanej przy identycznych wartościach Λ, K, m) ze wzrostem liczby kanałów obsługi n. Tak więc uzyskanie odpowiednio dokładnych wyników dla systemów wielokanałowych musi być związane ze wzrostem czasochłonności symulacji.

f) Przykłady ilustrujące dokładność metody

Ścisłe wyznaczenie błędu oszacowania wydajności granicznej wymagałoby każdorazowo pełnego, analitycznego znalezienia parametrów badanego systemu obsługi. Z tego względu, jak to zresztą ma miejsce w przypadku wielu metod przybliżonych, o dokładności metody będziemy wnioskować, ograniczając się do zbadania błędu metody przez jej stosowanie do pewnych, reprezentatywnych systemów obsługi, możliwych do analitycznego przebadania.

i) System typu M/G/1

Dla systemów typu M/G/1 zachodzi:

$$T = \left[E(b)\right]^{-1}, \tag{III.2.49}$$

gdzie E(b) jest wartością oczekiwaną czasu obsługi.
Dla takich systemów można wyznaczyć w sposób analityczny wartości $\pi_K(\lambda)$; oceny własności oszacowania T możemy więc w tym przypadku dokonać z pominięciem błędu związanego z określaniem wyników symulacji (błędu, którego wielkość możemy, jak to wyżej pokazano dokładnie określić).

Oznaczmy

$$F(z) = \sum_{i=1}^{\infty} p_i(\lambda) z^i. \tag{III.2.50}$$

Dowodzi się [KLEI 75], że F(z) spełnia równanie funkcyjne

$$F(z) = z\bar{b}\left[\lambda - \lambda F(z)\right], \tag{III.2.51}$$

gdzie $\bar{b}(s)$ jest transformatą Laplace'a-Stieltjesa dystrybuanty czasu obsługi.
Równanie III.2.51 zostało rozwiązane dla pewnych, szczególnych przypadków tak:

- dla systemów typu M/D/1

$$p_i(\lambda) = \frac{(i\rho)^{i-1}}{i!} e^{-i\rho}, \tag{III.2.52}$$

- dla systemów typu M/M/1

$$p_i(\lambda) = \frac{1}{i}\binom{2i-2}{i-1}\rho^{i-1}(1+\rho)^{1-2i}, \tag{III.2.53}$$

gdzie: $\rho = \lambda E(b)$.

W oparciu o (III.2.52) oraz (III.2.53) można łatwo wyznaczyć wartość $\pi_K(\lambda)$, a co za tym idzie, uzyskać odpowiednie wartości oszacowania $\tilde{T}$ dla dowolnego Λ.

W tabeli III.3 zestawiono wartości $\pi_K(\lambda)$ oraz wartości stosunku $\tilde{T}/T$, oznaczając je odpowiednio indeksem D dla systemu M/D/1 oraz indeksem M dla systemu M/M/1.

Z przedstawionych wyników można wyciągnąć następujące wnioski:
- zgodnie z oczekiwaniami, dla każdego K istnieje takie Λ_o, przy którym oszacowanie $\tilde{T}$ osiąga minimum. Wartość Λ_o stanowi bardzo dobre oszacowanie od góry wydajności granicznej i maleje (choć wolno) ze wzrostem K,
- dokładność oszacowania $\tilde{T}$ silnie rośnie ze wzrostem K,
- zarówno wartości $\tilde{T}$ jak i Λ_o wzrastają ze wzrostem współczynnika zmienności rozkładu czasów obsługi zgłoszeń.

Analiza przedstawionych wyników potwierdza też sformułowaną wyżej metodykę postępowania przy poszukiwaniu najlepszego oszacowania wydajności granicznej:
- w pierwszym rzędzie, używając niewielkich stosunkowo wartości K_1 (np. 400-600), należy zlokalizować wartości $\Lambda_o^{(1)}$,
- następnie, przy istotnie zwiększonej wartości $K_2 > K_1$ należy ponowić eksperyment dla $\Lambda < \Lambda_o^{(1)}$ poszukując nowego minimum $\Lambda_o^{(2)} < \Lambda_o^{(1)}$.

Warto zauważyć, iż przy właściwej organizacji eksperymentu symulacyjnego można w drugim kroku dla wartości Λ, które były już użyte poprzednio, ograniczyć się do modelowania tych okresów czynnych, które w pierwszym kroku zostały przerwane po K_1 obsługach (dały pozytywny wynik eksperymentu elementarnego) i to kontynuując ich modelowanie od punktu przerwania.

Takie postępowanie pozwala istotnie zmniejszyć czasochłonność symulacji.

ii) Jednokanałowy system WAD

W poprzednim punkcie wyznaczyliśmy, w sposób analityczny, wydajność graniczną systemu masowej obsługi z czasem obsługi zależnym od czasu oczekiwania zadanym zależnością (III.1.13), regulaminem obsługi LIFO (ϑ) i zgłoszeniami napływającymi strumieniem Poissona. Dla $\vartheta = 0$ oszacujemy te-

Tabela III.3

Wyniki obliczeń testowych dla systemów M/M/1 oraz M/D/1

K	$\frac{\Lambda}{T}$	$\pi_K^D(\Lambda)$	$\frac{\tilde{T}^D}{T}$	$\pi_K^M(\Lambda)$	$\frac{T^D}{T}$
200	0.98	0.0393	1.1230	0.0309	1.1692
	0.99	0.0473	1.1069	0.0352	1.1538
	1.00	0.0563	1.0974	0.0399	1.1434
	1.01	0.0662	1.0925	0.0448	1.1368
	1.02	0.0769	1.0909	0.0501	1.1332
	1.03	0.0883	1.0918	0.0556	1.1319
	1.04	0.1004	1.0945	0.0613	1.1332
	1.05	0.1130	1.0986	0.0673	1.1343
	1.10	0.1811	1.1312	0.0996	1.1582
	1.20	0.3138	1.2194	0.1679	1.2368
500	0.98	0.0195	1.0922	0.0166	1.1146
	0.99	0.0268	1.0699	0.0206	1.0963
	1.00	0.0357	1.0594	0.0252	1.0861
	1.005	0.0407	1.0570	0.0277	1.0831
	1.01	0.0461	1.0559	0.0303	1.0812
	1.02	0.0578	1.0565	0.0360	1.0800
	1.03	0.0708	1.0599	0.0421	1.0814
	1.04	0.0848	1.0651	0.0486	1.0847
	1.05	0.0995	1.0715	0.0554	1.0893
	1.10	0.1766	1.1126	0.0926	1.1243
	1.20	0.3137	1.2077	0.1667	1.2146
1000	0.98	0.0102	1.0860	0.0096	1.0936
	0.99	0.0166	1.0534	0.0133	1.0703
	1.00	0.0252	1.0413	0.0178	1.0594
	1.005	0.0304	1.0392	0.0204	1.0568
	1.01	0.0360	1.0388	0.0231	1.0556
	1.02	0.0490	1.0413	0.0292	1.0562
	1.03	0.0634	1.0465	0.0359	1.0596
	1.04	0.0790	1.0533	0.0431	1.0647
	1.05	0.0953	1.0611	0.0507	1.0711
	1.10	0.1762	1.1063	0.0912	1.1122
	1.20	0.3137	1.2038	0.1667	1.2072
2000	0.98	0.0458	1.1002	0.0051	1.0876
	0.99	0.0967	1.0340	0.0083	1.0538
	1.00	0.0178	1.0283	0.0126	1.0413
	1.005	0.0232	1.0272	0.0152	1.0391
	1.01	0.0293	1.0277	0.0181	1.0387
	1.02	0.0435	1.0319	0.0247	1.0411
	1.03	0.0595	1.0388	0.0321	1.0463
	1.04	0.0766	1.0468	0.0401	1.0531
	1.05	0.0939	1.0556	0.0484	1.0610
	1.10	0.1761	1.1031	0.0909	1.1061
	1.20	0.3137	1.2019	0.1667	1.2036

raz wydajność graniczną tego systemu, przy wykorzystaniu omówionych wyżej eksperymentów symulacyjnych.

Obliczenia przeprowadzono dla parametrów systemu $B_L = 100$, $B_S = 50$, dla których, posługując się równaniem zamieszczonym w tabeli III.1, można łatwo wyznaczyć dokładną wartość wydajności granicznej $T = 0.01318$. Stosując zaproponowaną wyżej metodę postępowania uzyskano wyniki przedstawione w tabeli III.4.

Tabela III.4

Szacowanie wydajności granicznej prostego systemu WAD

Λ	K	m	$\tilde{T}$	$\frac{\tilde{T}-T}{T}$ [%]
0.0130	200	400	0.020170	53
0.0135	200	200	0.015965	21
0.0140	200	200	0.017562	33
0.0135	2000	400	0.013936	5.3

Również i w tym przypadku metoda okazuje się stosunkowo dokładna, przy czym uzyskana przy małych K oraz m wartość $\Lambda_0 = 0.0135$ stanowi zaskakująco dobre oszacowanie (błąd wynosi tylko 2,4%).

Przedstawiona metoda szacowania od góry wydajności granicznej systemów masowej obsługi cechuje się asymptotyczną zbieżnością wyników do rozważań dokładnych. Zamieszczone przykłady wskazują, iż możliwe jest uzyskanie dobrej dokładności przy stosunkowo niewysokiej czasochłonności. W punkcie III.5 pokażemy, że w wielu przypadkach możliwe jest, na innej drodze, uzyskanie dodatkowo oszacowania poszukiwanej wydajności granicznej od dołu.

Omawiane tu przykłady dotyczyły systemów, w których stanowisko obsługi nie pozostaje bezczynne przy niepustej kolejce. Uwzględnienie tej dodatkowej cechy nie powoduje żadnych dodatkowych trudności - należy po prostu jako zakończenie okresu czynnego przyjąć moment, w którym rozpoczyna się bezczynność stanowiska obsługi przy pustej kolejce (okresy oczekiwania stanowiska obsługi na napłynięcie nowego zgłoszenia przy niepustej kolejce wchodzą w skład okresu czynnego).

Na zakończenie rozważań na temat przedstawionej metody należy podkreślić, iż możliwości jej stosowania wykraczają poza systemy z czasem obsługi zależnym od czasu oczekiwania i może ona być z powodzeniem stosowana również do innych, niekonwencjonalnych systemów masowej obsługi.

III.3. Analiza modelu pomocniczego

Powrócimy obecnie do omawiania klasy systemów jednostanowiskowych zdefiniowanych w punkcie III.1, w którym udowodniono, iż warunek stabilności takich systemów (których strukturę przedstawia rys. III.1) można określić w oparciu o wyniki analizy systemów pomocniczych, przedstawionych na rys. III.2. Analiza takich systemów pomocniczych jest interesująca również z innych względów. Załóżmy, że bufor, w którym tworzy się kolejka $\mathcal{K}_2$, w systemie na rys. III.1 ma dużą pojemność, tak że możliwa jest w długich okresach czasu praca systemu z różną intensywnością napływu i wypływu zgłoszeń z tego bufora, bez jego przepełnienia lub opróżnienia. Identyczna sytuacja wystąpi w przypadku sprzężenia tego bufora z innymi systemami ob-

sługi, tak, że nigdy nie ulegnie on przepełnieniu ani opróżnieniu. Techniczny przykład takiej sytuacji podamy w zakończeniu pracy.
Wówczas nieistotny okazuje się problem stabilności, a analiza systemu pomocniczego staje się zagadnieniem istotnym samym w sobie. Poniżej podamy metodę analizy systemu pomocniczego pod warunkiem, że znane jest analityczne rozwiązanie systemu WAD ze zgłoszeniami napływającymi strumieniem Poissona, obsługiwanymi zgodnie z regulaminem obsługi R_1 oraz zależnością czasu obsługi od czasu oczekiwania,

$$B(x/w) = \begin{cases} \Phi(x) & \text{dla} \quad w = 0 \\ F(x/w) & \text{dla} \quad 0 < w < \tau \\ H(x) & \text{dla} \quad w > \tau \end{cases} \qquad \text{(III.3.1)}$$

gdzie $\Phi(x)$ jest pewną, dowolną dystrybuantą, natomiast $F(x/w)$ jest identyczna jak we wzorze (III.1.3) (tzn. m.in., że istnieje $F(x/0)$), który będziemy dalej nazywać systemem bazowym. System bazowy jest więc systemem obsługi zgłoszeń głównych o zmodyfikowanej charakterystyce czasów obsługi i ze względu na ograniczony do τ czas oczekiwania (zgłoszenia oczekujące dłużej mają czas obsługi równy zeru - są tracone) jest on zawsze stabilny.

W szczególności założymy, że znana jest dystrybuanta czasu oczekiwania w systemie bazowym $W^{\Phi}(x)$, posiadająca transformatę Laplace'a-Stieltjesa $\overline{W}^{\Phi}(s)$ o postaci:

$$\overline{W}^{\Phi}(s) = W^{\Phi}(0)\left[1 + v^{\Phi}(s)\right], \qquad \text{(III.3.2)}$$

gdzie indeks Φ określa, dla jakiej funkcji $B(x/0) = \Phi(x)$ rozwiązanie zostało uzyskane.
Zauważmy, że np. dowolny system podstawowy, analizowany w rozdziale II.3, może być przyjęty jako system bazowy przy założeniu, że zgłoszenia napływają strumieniem Poissona.

a) Struktura okresu czynnego

W dalszym ciągu będziemy mówili, że w systemie pomocniczym trwa okres czynny wówczas, gdy kolejka K_1 jest niepusta lub obsługiwane jest zgłoszenie główne. Okres czynny rozpoczyna się więc z chwilą napłynięcia zgłoszenia głównego do systemu, w którym stanowisko obsługi jest zajęte przez zgłoszenie dodatkowe bądź oczekuje na napłynięcie zgłoszenia głównego.

Przyjmiemy pewne dalsze założenia określające zachowanie się systemu w sytuacji, gdy w czasie obsługi zgłoszenia dodatkowego napłynie zgłoszenie główne inicjujące kolejny okres czynny.
I tak, jeśli zgłoszenie główne napłynęło zanim czas obsługi zgłoszenia dodatkowego przekroczył pewną ustaloną wartość $\mathcal{T}$, wówczas obsługa ta jest przerywana, przy czym z przerwaniem jej związana jest pewna praca własna o dystrybuancie czasu trwania $Q(x)$, natomiast przerwana obsługa musi być następnie wznowiona od początku.

Jeśli natomiast w chwili nadejścia zgłoszenia głównego czas obsługi zgłoszenia dodatkowego przekroczył $\mathcal{T}$, wówczas obsługa ta nie może być przerwana.
Zdefiniowany wyżej sposób uprzywilejowania zgłoszeń głównych w stosunku do zgłoszeń dodatkowych nosi nazwę regulaminu priorytetowego mieszanego, bezwzględno-względnego (discretionary priority discipline).

Tak więc, obsługa zgłoszenia dodatkowego o dystrybuancie czasu trwania $Z(x)$ (posiadającej funkcję gęstości rozkładu $z(x)$) jest w istocie rozbita na dwie fazy [WOLI 76]:
- fazę przerywalną, o funkcji gęstości rozkładu $z_1(x)$

$$z_1(x) = z(x)\left[1 - H(x - \mathcal{T})\right] + \delta(x - \mathcal{T}) \int_{\mathcal{T}}^{\infty} z(t)dt, \qquad \text{(III.3.3)}$$

posiadającej transformatę Laplace'a

$$\overline{z}_1(s) = \int_0^{\mathcal{T}} e^{-sx} z(x)dx + e^{-s\mathcal{T}} \int_{\mathcal{T}}^{\infty} z(t)dt, \qquad \text{(III.3.4)}$$

jako że faza ta obejmuje całą obsługę zgłoszenia dodatkowego, jeśli jest ona krótsza aniżeli $\mathcal{T}$ lub też odcinek czasu $\mathcal{T}$, w przeciwnym razie
- fazę nieprzerywalną, występującą z prawdopodobieństwem p

$$p = \int_{\mathcal{T}}^{\infty} z(t)dt, \qquad \text{(III.3.5)}$$

o funkcji gęstości rozkładu czasu trwania $z_2(x)$,

$$z_2(x) = \frac{z(x + \mathcal{T})}{\int_{\mathcal{T}}^{\infty} z(t)dt}, \qquad \text{(III.3.6)}$$

posiadającej transformatę Laplace'a

$$\overline{z}_2(s) = \frac{e^{s\mathcal{T}} \int_{\mathcal{T}}^{\infty} e^{-sx} z(x)dx}{\int_{\mathcal{T}}^{\infty} z(x)dx} \qquad \text{(III.3.7)}$$

Oczywiście,w przypadkach granicznych otrzymujemy dla $\mathcal{F} = 0$ całkowicie nieprzerywalną, natomiast dla $\mathcal{F} \to \infty$ całkowicie przerywalną obsługę zgłoszeń dodatkowych.

Rozpatrzmy obecnie zależność pomiędzy wartościami horyzontu prognozy ϑ oraz $\mathcal{F}$. Jeśli $\vartheta \leq \mathcal{F}$, wówczas nie dysponujemy żadną informacją o procesie nadchodzenia zgłoszeń podczas nieprzerywalnej fazy obsługi. Jeśli natomiast $\vartheta \geq \mathcal{F}$, wówczas niemożliwe jest napłynięcie zgłoszenia głównego w czasie trwania przerywalnej fazy obsługi i w konsekwencji bez znaczenia jest fakt czy w odcinku czasu $0 \div \vartheta$ obsługa zgłoszenia dodatkowego jest przerywalna, czy też nieprzerywalna. W dalszych rozważaniach będziemy, dla uproszczenia notacji, w każdym przypadku, gdy $\vartheta \geq \mathcal{F}$, przyjmować $\mathcal{F} = \vartheta$, co - jak to wyżej uzasadniono - w niczym nie wpływa na rzeczywisty przebieg obsługi.

Tak więc, nie ograniczając ogólności rozważań, możemy ograniczyć się do przypadku $\vartheta \leq \mathcal{F}$.

Zauważmy, że w rozpatrywanym systemie pomocniczym mogą występować trzy typy okresów czynnych, różniących się sposobem ich rozpoczęcia:

typ y Okres czynny, w którym zgłoszenie główne inicjujące go zastaje stanowisko obsługi puste i w związku z tym rozkład czasu trwania obsługi tego zgłoszenia określa dystrybuanta $Y(x) = F(x/0)$.

typ c Okres czynny, w którym zgłoszenie główne inicjujące go zastaje stanowisko obsługi zajęte obsługą zgłoszenia dodatkowego w jego fazie przerywalnej. Zgłoszenie to oczekuje więc na rozpoczęcie obsługi przez czas określony dystrybuantą $Q(x)$, który, uwzględniając postać dystrybuanty $F(x/\vartheta)$, wpływa też na czas obsługi tego zgłoszenia głównego.

Dla dalszych rozważań wygodnie jest przyjąć, że przez pojęcie czasu obsługi zgłoszenia głównego, inicjującego taki okres czynny, rozumieć będziemy sumę czasu oczekiwania oraz czasu rzeczywistej obsługi tego zgłoszenia.

Założymy, że tak rozumiany, zmodyfikowany czas obsługi zgłoszenia inicjującego okres czynny, określać będzie pewna dystrybuanta $C(x)$.

typ d Okres czynny, w którym zgłoszenie główne inicjujące go zastaje stanowisko obsługi zajęte obsługą zgłoszenia dodatkowego w jego fazie nieprzerywalnej. Zgłoszenia to oczekuje więc na zakończenie obsługi zgłoszenia dodatkowego. Zakładając, podobnie jak poprzednio, że przez pojęcie czasu obsługi zgłoszenia inicjującego okres czynny rozumieć będziemy sumę czasu oczekiwania i czasu obsługi, przyjmiemy, że tak zmodyfikowany czas obsługi określać będzie pewna dystrybuanta $D(x)$.

W dalszym ciągu wskażemy jak można określić charakterystyki systemów, w których występowałyby wyłącznie okresy czynne jednego z wyżej wymienionych typów, a następnie wyznaczymy charakterystyki systemu pomocniczego, uwzględniając fakt występowania, z odpowiednimi prawdopodobieństwami, okresów czynnych różnego typu.

W przypadku, gdyby w systemie obsługi występowały wyłącznie okresy czynne typu y, byłby on identyczny z systemem bazowym, w którym $\phi(x)=Y(x)$, a więc ten przypadek z założenia potrafimy przeanalizować.

Przejdziemy obecnie do systemu, w którym występują wyłącznie okresy czynne typu c.

Niech α_1 oznacza prawdopodobieństwo tego, że rozpoczęła się obsługa zgłoszenia dodatkowego w jego fazie przerywalnej i do czasu zakończenia tej fazy nie napłynęło ani jedno zgłoszenie główne. Ponieważ obsługę zgłoszenia dodatkowego rozpoczyna się wyłącznie przy negatywnej prognozie, więc funkcja gęstości rozkładu czasu pomiędzy chwilą rozpoczęcia obsługi zgłoszenia dodatkowego a chwilą napłynięcia pierwszego zgłoszenia głównego ma postać

$$a_M(t) = \frac{\lambda_e e^{-\lambda t}}{e^{-\lambda\vartheta}} H(t-\vartheta), \tag{III.3.8}$$

a odpowiadająca jej dystrybuanta

$$A_M(t) = \left[1 - e^{\lambda\vartheta - \lambda t}\right] H(t-\vartheta).$$

Wartość $1 - \alpha_1$ można więc wyznaczyć jako:

$$1 - \alpha_1 = \int_0^{\infty} A_M(t)\,dZ_1(t) = \int_{\vartheta}^{\infty} dZ_1(t) - \int_{\vartheta}^{\infty} e^{\lambda\vartheta} e^{-\lambda t}\,dZ_1(t) =$$

$$= 1 - Z_1(\vartheta) - e^{\lambda\vartheta}\left[\bar{z}_1(\lambda) - \int_0^{\vartheta} e^{-\lambda t}\,dZ_1(t)\right],$$

tak więc

$$\alpha_1 = Z_1(\vartheta) + e^{\lambda\vartheta}\left[\bar{z}_1(\lambda) - \int_0^{\vartheta} e^{-\lambda t}\,dZ_1(t)\right]. \tag{III.3.9}$$

W szczególnym przypadku dla $\vartheta = 0$ otrzymujemy znaną zależność

$$\alpha_1\Big|_{\vartheta=0} = \bar{z}_1(\lambda).$$

Ponieważ dystrybuanta czasu oczekiwana zgłoszenia głównego inicjującego okres czynny typu c jest identyczna z dystrybuantą pracy własnej $Q(x)$,

możemy wyznaczyć funkcję rozkładu zmodyfikowanego czasu obsługi tego zgłoszenia c(x) jako

$$c(x) = \int_0^x f(x-\xi/\xi)q(\xi)d\xi. \qquad \text{(III.3.10)}$$

Łatwo sprawdzić, że system obsługi, w którym wszystkie okresy czynne są okresami typu c, posiada dystrybuantę czasu oczekiwania $w_d^C(x)$ o transformacie Laplace'a-Stieltjesa $\bar{w}_d^C(s)$, wyrażającej się wzorem:

$$\bar{w}_d^C(s) = \bar{w}^C(0)\left[\bar{q}(s) + v^C(s)\right], \qquad \text{(III.3.11)}$$

gdzie $w^C(0)$ oraz $v^C(s)$ są rozwiązaniami systemu bazowego przy $\Phi(x) = C(x)$.

Składnik $\bar{q}(s)$ we wzorze (III.3.11) odzwierciedla fakt, że w istocie zgłoszenie inicjujące okres czynny oczekuje przez czas q, który włączyliśmy do czasu obsługi tego zgłoszenia po to, aby możliwe było bezpośrednie wykorzystanie wyników analizy systemu bazowego.

Przejdziemy obecnie do omówienia systemów z okresami czynnymi typu d. Niech α_2 oznacza prawdopodobieństwo tego, że rozpoczęła się obsługa zgłoszenia dodatkowego w jego fazie nieprzerywalnej i do czasu zakończenia tej fazy nie napłynie ani jedno zgłoszenie główne.

Ponieważ z założenia zawsze zachodzi $\nu \leq \mathcal{F}$, więc brak prognozy strumienia zgłoszeń nadchodzących podczas nieprzerywalnej fazy obsługi zgłoszeń dodatkowych i na podstawie rozumowania identycznego jak przedstawione poprzednio zachodzi:

$$\alpha_2 = \bar{\mathcal{X}}_2(\lambda). \qquad \text{(III.3.12)}$$

Rozpatrzmy warunkowy rozkład czasu oczekiwania zgłoszenia inicjującego okres czynny typu d pod warunkiem, że zgłoszenie takie nadeszło podczas trwania nieprzerywalnej fazy obsługi zgłoszenia dodatkowego. Zmienna losowa h stanowi więc różnicę dwóch zmiennych losowych z_2 oraz a (posiadającej dystrybuantę (III.1.1)) przy wyżej sprecyzowanym warunku.

Funkcję gęstości rozkładu takiej zmiennej losowej wyznaczono w [WOLI 75], korzystając ze znanego (patrz np. [PAPO 72]) wzoru na funkcję gęstości rozkładu różnicy dwóch zmiennych losowych, otrzymując:

$$h(t) = \int_{-\infty}^{\infty} \mathcal{X}_2(t-\xi)\,\mathbb{H}(t-\xi)\lambda e^{\lambda\xi}\,\mathbb{H}(-\xi)d\xi\,\frac{1}{1-\alpha_2} =$$

$$= \int_0^{\infty} \mathcal{X}_2(t+v)\lambda e^{-\lambda v}dv = \lambda e^{\lambda t}\left[\bar{\mathcal{X}}_2(\lambda) - \int_0^t e^{-\lambda v}\mathcal{X}_2(u)du\right]\frac{1}{1-\alpha_2}.$$

Stosując transformację Laplace'a uzyskujemy:

$$\bar{h}(s) = \lambda\,\frac{\bar{\mathcal{X}}_2(\lambda) - \bar{\mathcal{X}}_2(s)}{(s-\lambda)(1-\alpha_2)}. \qquad \text{(III.3.13)}$$

Podobnie jak poprzednio, możemy więc wyznaczyć funkcję gęstości rozkładu zmodyfikowanego czasu obsługi zgłoszenia inicjującego okres czynny d(x)

$$d(x) = \int_0^x f(x-\xi/\xi)h(\xi)d\xi \qquad \text{(III.3.14)}$$

oraz transformatę Laplace'a-Stieltjesa dystrybuanty czasu oczekiwania zgłoszeń głównych w systemie, w którym wszystkie okresy czynne byłyby okresami typu d

$$\bar{w}_d^{\mathcal{D}}(s) = w^{\mathcal{D}}(0)\left[\bar{h}(s) + v^{\mathcal{D}}(s)\right], \qquad \text{(III.3.15)}$$

gdzie $w^{\mathcal{D}}(0)$ i $v^{\mathcal{D}}(s)$ są rozwiązaniami systemu bazowego, przy $\Phi(x) = \mathcal{D}(x)$.

b) Charakterystyki systemu

Dla wyznaczenia charakterystyk systemu pomocniczego, a specjalnie średniej liczby zgłoszeń traconych w odcinku czasu pomiędzy dwoma kolejnymi chwilami rozpoczęcia obsługi zgłoszenia dodatkowego $L(\lambda)$, rozpatrzmy przedstawiony na rys. III.5 schemat obsługi zgłoszeń w systemie pomocniczym.

Łatwo zauważyć, że w rozpatrywanym przypadku średnia liczba $L(\lambda)$ zgłoszeń traconych pomiędzy chwilami rozpoczęcia kolejnych obsług zgłoszeń dodatkowych wynosi:

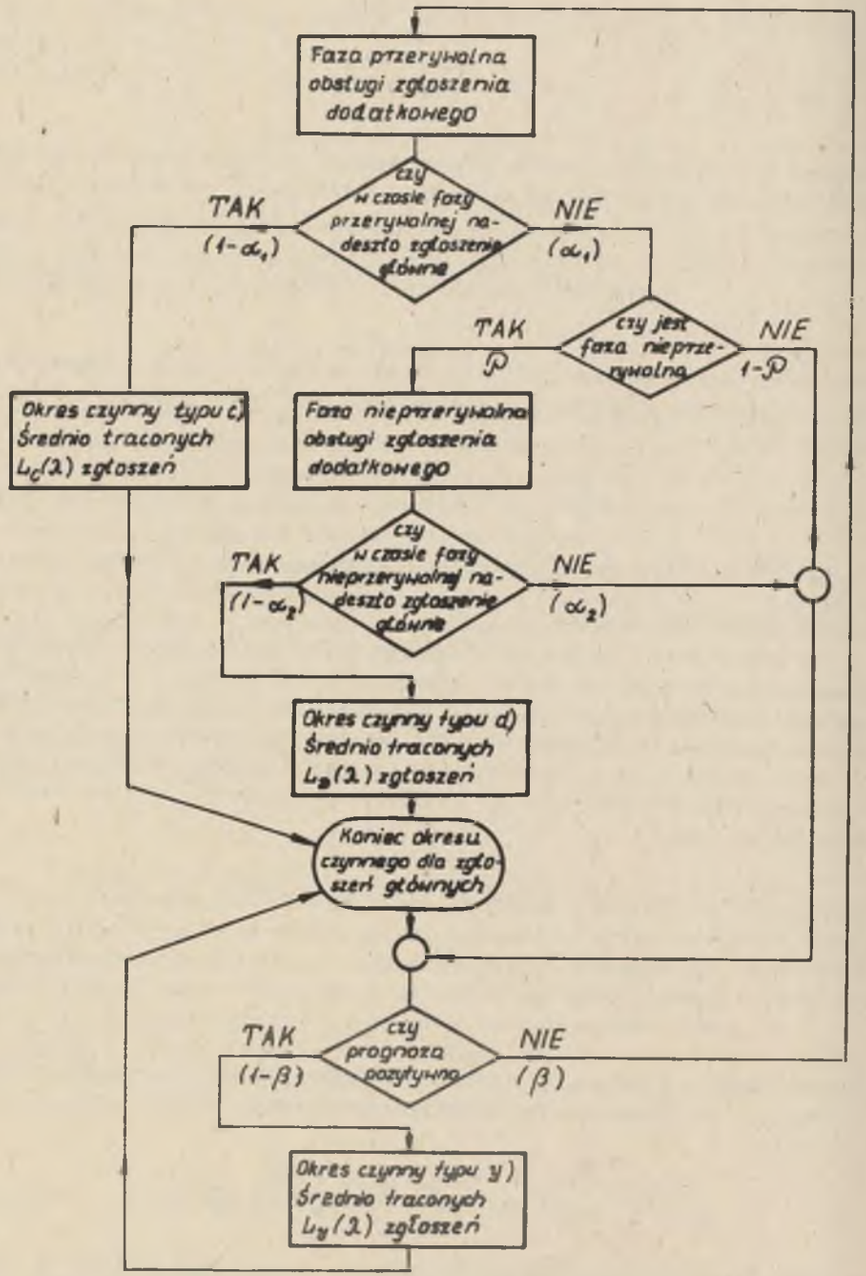
$$L(\lambda) = \alpha_1\left[(1-\mathcal{P}) + \mathcal{P}\alpha_2\right]0 + (1-\alpha_1)L_c(\lambda) + \alpha_1\mathcal{P}(1-\alpha_2)L_{\mathcal{D}}(\lambda) +$$

$$+ \beta\sum_{i=1}^{\infty} i(1-\beta)^{i-1}L_y(\lambda), \qquad \text{(III.3.16)}$$

gdzie: $L_y(\lambda)$, $L_c(\lambda)$, $L_{\mathcal{D}}(\lambda)$ oznaczają średnią liczbę zgłoszeń traconych w czasie pojedynczego okresu czynnego, odpowiednio typu y, c, d.
Tak więc:

$$L(\lambda) = (1-\alpha_1)L_c(\lambda) + \alpha_1\mathcal{P}(1-\alpha_2)L_{\mathcal{D}}(\lambda) + \frac{1-\beta}{\beta}L_y(\lambda). \qquad \text{(III.3.17)}$$

Średnią liczbę zgłoszeń traconych w czasie okresu czynnego typu y znajdziemy na podstawie następującego rozumowania:



Rys. III.5. Schemat obsługi zgłoszeń w rozpatrywanym systemie pomocniczym

Średnia liczba zgłoszeń głównych, traconych w jednostce czasu w systemie, w którym wszystkie okresy czynne byłyby okresami typu y) wynosi $\lambda\left[1 - \mathcal{W}^{Y}(\tau)\right]$ jako że prawdopodobieństwo tego, że czas oczekiwania zgłoszenia głównego na rozpoczęcie obsługi wynosi $\left[1 - \mathcal{W}^{Y}(\tau)\right]$. Z drugiej strony, średnia liczba okresów czynnych w takim systemie wynosi $\lambda\mathcal{W}^{Y}(0)$. Tak więc średnia liczba zgłoszeń traconych w czasie okresu czynnego typu y wyraża się wzorem:

$$L_y(\lambda) = \frac{\lambda\left[1 - \mathcal{W}^{Y}(\tau)\right]}{\lambda\mathcal{W}^{Y}(0)} = \frac{1 - \mathcal{W}^{Y}(\tau)}{\mathcal{W}^{Y}(0)}. \qquad \text{(III.3.18)}$$

Analogicznie

$$L_D(\lambda) = \frac{1 - \mathcal{W}_d^{D}(\tau)}{\mathcal{W}^{D}(0)}, \qquad \text{(III.3.19)}$$

oraz

$$L_C(\lambda) = \frac{1 - \mathcal{W}_d^{C}(\tau)}{\mathcal{W}^{C}(0)} + 1, \qquad \text{(III.3.20)}$$

jako że w przypadku okresu czynnego typu c, niezależnie od straty zgłoszeń głównych (których liczba wyznaczona jest identycznie jak w przypadku okresów czynnych typu y i typu d) tracone jest każdorazowo również przerwane zgłoszenie dodatkowe, którego obsługa musi być, zgodnie z przyjętym założeniem, wznowiona od początku.

Analiza systemu pomocniczego jest interesująca nie tylko ze względu na możliwość wyznaczania wydajności granicznej systemów wyjściowych, zdefiniowanych w punkcie III.1. Często dzięki odpowiedniej gospodarce zapasem zgłoszeń w kolejce $\mathcal{K}_2$ (np. ich wymianie z innymi systemami) zapewniona jest stabilność systemu wyjściowego z jednej strony oraz stała obecność zgłoszeń w kolejce $\mathcal{K}_2$ z drugiej strony.

W takich przypadkach, interesująca jest znajomość innych charakterystyk systemu pomocniczego, takich jak: jego wydajność, strumień zgłoszeń traconych, strumień zgłoszeń dodatkowych itp., do których wyznaczenia obecnie przejdziemy.

Na wstępie rozpatrzymy prawdopodobieństwo tego, że kolejny okres czynny będzie odpowiednio okresem czynnym typu y, c, d, które oznaczymy odpowiednio ψ, γ, ν.

Na podstawie schematu przedstawionego na rys. III.5 można łatwo stwierdzić, że:

$$\psi = (1-\beta)\sum_{i=0}^{\infty}\beta^i\alpha_1^i\left\{(1-p) + p\alpha_2\right\}^i = \frac{1-\beta}{1-\Omega},$$

gdzie:

$$\Omega = \beta\alpha_1\left\{(1-p) + p\alpha_2\right\}; \qquad \text{(III.3.21)}$$

$$\gamma = \beta(1-\alpha_1)\sum_{i=0}^{\infty}\Omega^i = \frac{\beta(1-\alpha_1)}{1-\Omega}, \qquad \text{(III.3.22)}$$

$$\nu = \beta\alpha_1 p(1-\alpha_2)\sum_{i=0}^{\infty}\Omega^i = \frac{\beta\alpha_1 p(1-\alpha_2)}{1-\Omega}, \qquad \text{(III.3.23)}$$

przy czym oczywiście zachodzi:

$$\psi + \gamma + \nu = 1.$$

Uwzględniając wyznaczone prawdopodobieństwa wystąpienia okresu czynnego odpowiedniego typu oraz fakt, iż średnia liczba okresów czynnych dowolnego typu Φ w jednostce czasu wynosi $\lambda\mathcal{W}^{\Phi}(0)$, a więc średnia liczba cykli obsługi złożonych z okresu czynnego danego typu oraz okresu bezczynnego wynosi $\left[\lambda\mathcal{W}^{\Phi}(0)\right]^{-1}$, możemy wyznaczyć średnią liczbę cykli obsługi systemu pomocniczego w jednostce czasu $\mathcal{N}$ jako:

$$\mathcal{N} = \left[\frac{\psi}{\lambda\mathcal{W}_o^{Y}} + \frac{\gamma}{\lambda\mathcal{W}_o^{C}} + \frac{\nu}{\lambda\mathcal{W}_o^{D}}\right]^{-1}. \qquad \text{(III.3.24)}$$

Przez wydajność systemu pomocniczego V będziemy dalej rozumieć intensywność strumienia całkowicie obsłużonych zgłoszeń, opuszczających stanowisko obsługi.
Wprost z definicji mamy więc:

$$V = \mathcal{N}L, \qquad \text{(III.3.25)}$$

gdzie:
$\mathcal{N}$ - średnia liczba cykli obsługi w jednostce czasu,
L - średnia liczba zgłoszeń obsłużonych w jednym cyklu obsługi.

Oczywiście

$$L = \psi L_y^* + \gamma L_c^* + \nu L_d^* = \psi\left[L_y^{(1)} + L_y^{(2)}\right] + \gamma\left[L_c^{(1)} + L_c^{(2)}\right] + \nu\left[L_d^{(1)} + L_d^{(2)}\right], \qquad \text{(III.3.26)}$$

gdzie: L_y^*, L_c^*, L_d^* oznaczają średnią liczbę zgłoszeń obsłużonych w czasie cyklu obsługi, pod warunkiem że w cyklu tym wystąpi okres czynny odpowiedniego typu.

Na łączną liczbę zgłoszeń obsłużonych w czasie pojedynczego cyklu obsługi składa się średnia liczba zgłoszeń dodatkowych, oznaczana odpowiednio przez $L_y^{(1)}$, $L_c^{(1)}$, $L_d^{(1)}$ oraz średnia liczba zgłoszeń głównych, oznaczona odpowiednio przez $L_y^{(2)}$, $L_c^{(2)}$, $L_d^{(2)}$.
Odwołując się ponownie do schematu przedstawionego na rys. III.4, otrzymujemy:

$$L_y^{(1)} = \frac{1}{\psi}\left[(1-\beta)0 + (1-\beta)\Omega\cdot 1 + (1-\beta)\Omega^2\cdot 2 + \ldots\right] = \frac{1-\beta}{\psi}\sum_{i=1}^{\infty} i\Omega^i,$$

ale

$$\sum_{n=0}^{\infty} nX^n = \sum_{n=1}^{\infty} nX^n = \frac{X}{(1-X)^2}$$

tak więc:

$$L_y^{(1)} = \frac{1}{\psi}\frac{(1-\beta)\Omega}{(1-\Omega)^2}, \qquad \text{(III.3.27)}$$

$$L_c^{(1)} = \frac{1}{\gamma}\left[\beta(1-\alpha_1)\sum_{i=1}^{\infty} i\Omega^i\right] = \frac{1}{\gamma}\frac{\beta(1-\alpha_1)\Omega}{(1-\Omega)^2}, \qquad \text{(III.3.28)}$$

$$L_D^{(1)} = \frac{1}{\nu}\left\{\beta\alpha_1 p(1-\alpha_2)\sum_{i=0}^{\infty}(i+1)\Omega^i\right\} = \frac{1}{\nu}\left\{\frac{(1-\alpha_2)p}{(1-p)+p\alpha_2}\sum_{i=0}^{\infty}(i+1)\Omega^{i+1}\right\},$$

a więc:

$$L_D^{(1)} = \frac{1}{\nu}\frac{(1-\alpha_2)p}{(1-p)+p\alpha_2}\frac{\Omega}{(1-\Omega)^2}. \qquad \text{(III.3.29)}$$

Oczywiście

$$L_y^{(2)} = \frac{w^Y(\tau)}{w^Y(0)}, \quad L_c^{(2)} = \frac{w_d^C(\tau)}{w_d^C(0)}, \quad L_{\mathcal{D}}^{(2)} = \frac{w_d^{\mathcal{D}}(\tau)}{w^{\mathcal{D}}(0)}. \tag{III.3.30}$$

Korzystając z zależności (III.3.18)-(III.3.20) można też wprost określić intensywność strumienia traconych zgłoszeń głównych S_g

$$S_g = \lambda\left[\psi\frac{1-w^Y(\tau)}{w^Y(0)} + \gamma\frac{1-w_d^C(\tau)}{w^C(0)} + \vartheta\frac{1-w_d^{\mathcal{D}}(\tau)}{w^{\mathcal{D}}(0)}\right] \tag{III.3.31}$$

i intensywność strumienia zgłoszeń dodatkowych

$$\Theta_d = \lambda\left[\psi L_y^{(1)} + \gamma L_c^{(1)} + \vartheta L_d^{(1)}\right] = \frac{\beta\alpha_1}{1-\Omega} \tag{III.3.32}$$

Na zakończenie warto zwrócić uwagę na fakt, iż w istocie przy analizie systemu pomocniczego korzystamy z rozwiązania systemu bazowego wyłącznie dla specyficznych postaci dystrybuanty $\Phi(x)$, a w szczególności:

$$\Phi(x) = \begin{cases} Y(x) = P(x/0), \\ C(x), \\ \mathcal{D}(x), \end{cases} \tag{III.3.33}$$

gdzie $C(x)$ i $\mathcal{D}(x)$ określone są odpowiednio przez (III.3.10) oraz (III.3.14), co może w praktyce rozszerzyć klasę systemów bazowych, które potrafimy przeanalizować względnie, jak to pokażemy na przykładzie, znacznie uprościć obliczenia.

III.4. Przypadki szczególne i własności modelu pomocniczego

Przy obowiązujących w poprzednim punkcie ogólnych założeniach nie można sformułować zasad doboru wartości horyzontu prognozy ϑ oraz wartości $\mathcal{T}$.

Poniżej wyprowadzimy takie reguły, zakładając wykładniczy rozkład czasu obsługi zgłoszeń dodatkowych, tj.

$$\left.\begin{aligned} Z(x) &= 1 - e^{-\mu_d x} \\ z(x) &= \mu_d e^{-\mu_d x} \end{aligned}\right\} \tag{III.4.1}$$

a następnie przyjmując dalsze uproszczenia zilustrujemy niektóre specyficzne własności rozpatrywanego modelu.

a) Dobór wartości parametrów ϑ oraz $\mathcal{T}$

Korzystając z wyprowadzonych w punkcie III.3 zależności oraz (III.4.1) uzyskujemy (przy uwzględnieniu nierówności $\vartheta \leq \mathcal{T}$):

$$\left.\begin{aligned} \alpha_1 &= 1 - \frac{\lambda}{\lambda+\mu_d} e^{-\mu_d\vartheta} + \frac{\lambda}{\lambda+\mu_d} e^{\lambda\vartheta-\lambda\mathcal{T}-\mu_d\mathcal{T}} \\ \alpha_2 &= \frac{\mu_d}{\lambda+\mu_d}, \qquad p = e^{-\mu_d\mathcal{T}} \end{aligned}\right\} \tag{III.4.2}$$

Wartości $L_c(\lambda)$ oraz $L_y(\lambda)$, występujące we wzorze (III.3.17), nie zależą od ϑ. W przypadku wykładniczego czasu obsługi zgłoszeń dodatkowych zachodzi:

$$\bar{h}(s) = \frac{\mu_d}{s+\mu_d}, \tag{III.4.3}$$

a więc również $L_{\mathcal{D}}(y)$ nie zależy od ϑ.
Uwzględniając (III.4.2) w (III.3.17) otrzymujemy:

$$L(\lambda) = \frac{\lambda}{\lambda+\mu_d}\left[e^{-\mu_d\vartheta} - e^{-(\lambda+\mu_d)\mathcal{T}} e^{\lambda\vartheta}\right]L_c(\lambda) +$$
$$+ \left[1 - \frac{\lambda}{\lambda+\mu_d} e^{-\mu_d\vartheta} + \frac{\lambda}{\lambda+\mu_d} e^{-(\lambda+\mu_d)\mathcal{T}} e^{\lambda\vartheta}\right] e^{-\mu_d\mathcal{T}} \frac{\lambda}{\lambda+\mu_d} L_{\mathcal{D}}(\lambda) + (e^{\lambda\vartheta}-1)L_y(\lambda), \tag{III.4.4}$$

upraszczającą się w szczególnym przypadku $\vartheta = \mathcal{T}$ do postaci:

$$L(\lambda) = \frac{\lambda}{\lambda+\mu_d} e^{-\mu_d\vartheta} L_{\mathcal{D}}(\lambda) + (e^{\lambda\vartheta} - 1)L_y(\lambda). \tag{III.4.5}$$

W celu wyznaczenia wartości $\mathcal{T}_{opt}$ minimalizującej $L(\lambda)$ wyznaczymy pochodną:

$$\frac{\partial L(\lambda)}{\partial\vartheta} = -G_1 e^{-\mu_d\vartheta} + G_2 e^{\lambda\vartheta} \quad \text{dla} \quad \vartheta < \mathcal{T} \tag{III.4.6}$$

$$\frac{\partial L(\lambda)}{\partial\vartheta} = -G_3 e^{-\mu_d\vartheta} + G_4 e^{\lambda\vartheta} \quad \text{dla} \quad \vartheta = \mathcal{T}, \tag{III.4.7}$$

gdzie:

$$G_1 = \frac{\lambda\mu_d}{\lambda+\mu_d} L_c(\lambda) - \frac{\lambda^2\mu_d}{(\lambda+\mu_d)^2} e^{-\mu_d \mathcal{F}} L_{\mathcal{D}}(\lambda),$$

$$G_2 = \frac{\lambda^3}{(\lambda+\mu_d)^2} e^{-(\lambda+2\mu_d)\mathcal{F}} L_{\mathcal{D}}(\lambda) + \frac{\lambda^2}{\lambda+\mu_d} e^{-(\lambda+\mu_d)\mathcal{F}} L_c(\lambda) + \lambda L_y(\lambda),$$

$$G_3 = \frac{\lambda\mu_d}{\lambda+\mu_d} L_{\mathcal{D}}(\lambda),$$

$$G_4 = \lambda L_y(\lambda).$$

Łatwo widać, że dla przypadku $\mathcal{V} = \mathcal{F}$, czyli przy braku fazy przerywalnej, zachodzi $G_3 > 0$, $G_4 > 0$, a więc istnieje jednoznaczne rozwiązanie równania:

$$-G_3 e^{-\mu_d \mathcal{V}} + G_4 e^{\lambda \mathcal{V}} = 0, \qquad \text{(III.4.8)}$$

o postaci:

$$\mathcal{V}_{opt} = \frac{1}{\lambda+\mu_d} \ln\left(\frac{G_3}{G_4}\right), \qquad \text{(III.4.9)}$$

dla którego osiąga się minimalną wartość funkcji $L(\lambda)$.
Z jednoznaczności rozwiązania (III.4.9) oraz ciągłości lewej strony równania (III.4.8) wynika, iż funkcja $L(\lambda)$ posiada ze względu na $\mathcal{V}$ jedno ekstremum w przedziale zmienności $\mathcal{V} \in (0, \infty)$.

Rozpatrzmy obecnie przypadek $\mathcal{V} < \mathcal{F}$ to znaczy równanie

$$-G_1 e^{-\mu_d \mathcal{V}} + G_2 e^{\lambda \mathcal{V}} = 0. \qquad \text{(III.4.10)}$$

Ponieważ $G_2 > 0$, więc możliwe są dwie sytuacje:

i) $G_1 > 0$ i wówczas istnieje dla przypadku $\mathcal{V} < \mathcal{F}$ jednoznacznie określona, optymalna wartość $\mathcal{V}^*_{opt}$, przy której funkcja $L(\lambda)$ osiąga minimum

$$\mathcal{V}^*_{opt} = \frac{1}{\lambda+\mu_d} \ln\left(\frac{G_1}{G_2}\right). \qquad \text{(III.4.11)}$$

ii) $G_1 \leq 0$ i wówczas w przedziale zmienności $\mathcal{V} \in (0, \mathcal{F})$ wartości $L(\lambda)$ rosną monotonicznie ze wzrostem $\mathcal{V}$ jako że $\frac{\partial L(\lambda)}{\partial \mathcal{V}} > 0$ dla każdego $\mathcal{V}$ z rozpatrywanego przedziału zmienności.

Podane wyżej reguły pozwalają na określenie wartości $\mathcal{F}$ i T minimalizujących $L(\lambda)$.

b) <u>System bazowy w postaci systemu podstawowego</u>

Przyjmiemy, że

$$B(x/w) = \begin{cases} \Phi(x) = B_0(x) & \text{dla } w = 0 \\ F(x/w) = B_i(x) & \text{dla } w_{i-1} < w \leq w_i \quad i=1,2,\ldots,J-1 \\ \mathbb{H}(x) & \text{dla } w > w_{J-1} = \tau \end{cases} \qquad \text{(III.4.12)}$$

oraz

$$F(x/0) = B_1(x),$$

gdzie $B_i(x)$ są dowolnymi dystrybuantami, takimi że $E(b_i) < \infty$ $i=0,1,\ldots J-1$ i skorzystamy z wyników analizy systemu podstawowego uzyskanych w rozdziale II.

Zauważmy, że zależność (III.3.10) przybiera wówczas postać:

$$c(x) = \int_0^x f(x-\xi/\xi) q(\xi) d\xi =$$

$$= \sum_{i=1}^{J} \int_0^x b_i(x-\xi) q(\xi) \left[\mathbb{H}(\xi - w_{i-1}) - \mathbb{H}(\xi - w_i)\right] d\xi, \qquad \text{(III.4.13)}$$

a transformata Laplace'a-Stieltjesa tej funkcji gęstości wyraża się wzorem:

$$\bar{c}(s) = \sum_{i=1}^{J} \left\{ \bar{b}_i(s) \int_0^\infty e^{-s\xi} q(\xi) \left[\mathbb{H}(\xi - w_{i-1}) - \mathbb{H}(\xi - w_i)\right] d\xi \right\} \qquad \text{(III.4.14)}$$

Wyznaczenie tej transformaty może w zależności od postaci funkcji $q(\xi)$ stwarzać istotne trudności. Jest ono natomiast bardzo proste w przypadku, gdy $q(\xi) = \mu_q e^{-\mu_q \xi}$, bowiem:

$$\int_0^\infty e^{-s\xi} \mu_q e^{-\mu_q \xi} \mathbb{H}(\xi - w_i) = \frac{\mu_q}{s+\mu_q} e^{-\mu_q w_i} e^{-s w_i}$$

Analogicznie można wyznaczyć

$$d(x) = \int_0^x f(x-\xi/\xi)h(\xi)d\xi =$$

$$= \sum_{i=1}^{J} \int_0^x b_i(x-\xi)h(\xi)\left[H(\xi-w_{i-1}) - H(\xi-w_i)\right]d\xi, \qquad \text{(III.4.15)}$$

oraz

$$\bar{d}(s) = \sum_{i=1}^{J}\left\{ b_i(s) \int_0^\infty e^{-s\xi}h(\xi)\left[H(\xi-w_{i-1}) - H(\xi-w_i)\right]d\xi\right\}, \qquad \text{(III.4.16)}$$

W dalszych rozważaniach ograniczymy się, ze względu na złożoność obliczeń, do rozpatrywania systemu bazowego typu (III.4.12), w którym

$$J = 3,$$

$$F(x/0) = B_1(x) = 1 - e^{-\mu_1 x},$$

$$B_2(x) = 1 - e^{-\mu_2 x},$$

$$w_1 = D. \qquad \text{(III.4.17)}$$

Przeanalizujemy system pomocniczy oparty na tym systemie bazowym przyjmując, że dystrybuanta czasu obsługi zgłoszeń dodatkowych ma postać:

$$Z(x) = 1 - e^{-\mu_d x}, \qquad \text{(III.4.18)}$$

natomiast dystrybuanta czasu pracy własnej

$$Q(x) = 1 - e^{-\mu_q x} \qquad \text{(III.4.19)}$$

i będziemy badać własności tego systemu.

Parametry α_1, α_2, p, a także postać funkcji $\bar{h}(s)$ dla rozpatrywanego przypadku określają zależności (III.4.2) i (III.4.3).

Niech $K(t, 1, \mu_x)$ oznacza dystrybuantę, której transformata Laplace'a-Stieltjesa ma postać:

$$\bar{k}(s,1,\mu_x) = 1\,\frac{\mu_1}{s+\mu_1} +$$

$$+(1-1)\left\{\frac{\mu_1}{s+\mu_1}\,\frac{\mu_x}{s+\mu_x}\left[1-e^{-\mu_x D}e^{-sD}\right] + \frac{\mu_2}{s+\mu_2}\,\frac{\mu_x}{s+\mu_x}\left[e^{-\mu_x D}e^{-sD} - e^{-\mu_x \tau}e^{-s\tau}\right] + \right.$$

$$\left. + \frac{\mu_x}{s+\mu_x}\,e^{-\mu_x\tau}e^{-s\tau}\right\}. \qquad \text{(III.4.20)}$$

Łatwo sprawdzić, że przy uwzględnieniu wymienionych wyżej założeń zachodzi:

$$\bar{c}(s) = \bar{k}(s,0,\mu_q), \qquad \text{(III.4.21)}$$

$$\bar{d}(s) = \bar{k}(s,0,\mu_d) \qquad \text{(III.4.22)}$$

oraz oczywiście

$$\bar{y}(s) = \int_0^\infty e^{-st}dF(t/0) = \bar{k}(s,1,\mu_1). \qquad \text{(III.4.23)}$$

Tak więc dla znalezienia dystrybuanty czasu oczekiwania w systemie bazowym, przy wystąpieniu dowolnego spośród trzech wymienionych w poprzednim punkcie okresów czynnych, wystarczy przeanalizować przypadek, dla którego:

$$\Phi(t) = K(t,1,\mu_x), \quad 1 \geqslant 0. \qquad \text{(III.4.24)}$$

Wyniki analizy systemu podstawowego z wyżej określoną funkcją $B_o(x)$, prowadzonej zgodnie z zasadami określonymi w rozdziale II, zawiera ustęp II.3.c.

W oparciu o nie możemy w szczególności wyznaczyć:

$$\left.\begin{aligned} w_i^Y(\tau) &= w^Y(0)\left[1 + u_1^*(0) + u_2^*(0)\right]\Big|_{Y(x)} = K(x,1,\mu_1), \\ w_d^C(\tau) &= w^Y(0)\left[1-e^{-\mu_q\tau} + u_1^*(0) + u_2^*(0)\right]\Big|_{C(x)} = K(x,0,\mu_q), \\ w_d^D(\tau) &= w^D(0)\left[1-e^{-\mu_d\tau} + u_1^*(0) + u_2^*(0)\right]\Big|_{D(x)} = K(x,0,\mu_d), \end{aligned}\right\} \qquad \text{(III.4.25)}$$

gdzie wartości $w^{\Phi}(0)$ oraz $u_1^{\Phi}(0)$, $u_2^{\Phi}(0)$ (zachowujące znaczenie omówione w rozdziale II) wyznacza się wg wzorów z ustępu II.3.c przy odpowiedniej postaci funkcji $\Phi(x)$.

Korzystając z (III.4.25) można już bezpośrednio w oparciu o zależności uzyskane w punkcie III.3 określić różne parametry systemu. W omawianych niżej przykładach numerycznych wyznaczono jeszcze dodatkowo dla przypadku nieprzerywalnej obsługi zgłoszeń dodatkowych jednostkowy czas obsługi zgłoszeń $\eta^*(\lambda)$, wyrażający się zależnością:

$$\eta^*(\lambda) = \frac{1}{L}\left\{\psi\frac{w^Y(\tau)}{w^Y(0)}\,BS1+\vartheta\frac{w_d^{\mathcal{D}}(\tau)}{w_d^{\mathcal{D}}(0)}\,BS2+\left[\psi L_y^{(1)}+\vartheta L_{\mathcal{D}}^{(1)}\right]E(z)\right\}, \qquad \text{(III.4.26)}$$

gdzie:

$$BS1 = \frac{\lambda w^Y(D)E(b_1) + \lambda\left[w^Y(\tau) - w^Y(D)\right]E(b_2)}{\lambda w^Y(\tau)}$$

$$BS2 = \frac{\lambda w_d^{\mathcal{D}}(D)E(b_1) + \lambda\left[w_d^{\mathcal{D}}(\tau) - w_d^{\mathcal{D}}(D)\right]E(b_2)}{\lambda w_d^{\mathcal{D}}(\tau)}.$$

c) <u>Przykłady numeryczne i własności modelu</u>

Wyniki obliczeń numerycznych przedstawionych poniżej nie mają na celu systematycznego badania własności rozpatrywanej klasy systemów a jedynie zilustrowanie na przykładach niektórych cech.

Obliczenia prowadzone były dla systemu pomocniczego, dla którego zależność czasu obsługi od czasu oczekiwania ma postać (III.4.17) z całkowicie nieprzerywalną obsługą zgłoszeń dodatkowych, tj. przy $\mathcal{F} = \vartheta$.

Jako podstawę do obliczeń przyjęto model o parametrach:

$$\mu_1 = 2, \quad \mu_2 = 1, \quad \mu_d = 0.5, \quad D = 5, \quad \tau = 20.$$

Proporcje między poszczególnymi parametrami tego modelu są podobne do proporcji pomiędzy parametrami pewnego procesu technologicznego [ZSAK 80].

W tabeli III.6 przedstawiono wpływ horyzontu prognozy ϑ w rozpatrywanym systemie na wybrane jego charakterystyki (łatwo sprawdzić, że przy założonych parametrach $\vartheta_{opt} = 1.83$).

Jak widać, wybór horyzontu prognozy wpływa na sposób istotny na wszystkie rozpatrywane charakterystyki systemu.

Na uwagę zasługuje silna zależność jednostkowego czasu obsługi $\eta^*(\lambda)$ od horyzontu prognozy. W przypadku gdy, jak to zwykle ma miejsce, koszt pracy stanowiska obsługi jest proporcjonalny do $\eta(\lambda) = \lambda\eta^*(\lambda)$, istnieje realna możliwość znaczącego obniżenia kosztu pracy stanowiska obsługi dro-

gą odpowiedniego doboru ϑ. W szczególności, ponieważ $\eta^*(\lambda)$ maleje monotonicznie ze wzrostem ϑ, natomiast wydajność systemu V posiada maksimum, możliwe jest istotne obniżenie wartości $\eta^*(\lambda)$ przy stosunkowo niewielkim zmniejszeniu wydajności systemu.

Tabela III.5

Wybrane charakterystyki systemu pomocniczego przy $\vartheta = 1$

λ^{-1}	$L(\lambda)$	V	$V-S_g$	Θ_d	$\eta^*(\lambda)$
.75	52.67930	1.02274	1.01673	0.00601	0.96997
.80	11.47092	1.05236	1.03349	0.01889	0.92831
.85	3.11287	1.08722	1.04498	0.04224	0.87560
.90	1.00671	1.11063	1.03908	0.07155	0.83183
.94	0.40962	1.11564	1.01935	0.09729	0.80847
.98	0.18013	1.10824	0.98854	0.11970	0.79884

W tabeli III.5 zilustrowano wpływ zmian intensywności strumienia zgłoszeń na charakterystyki rozpatrywanego modelu.

Na szczególną uwagę zasługuje fakt, iż przy ustalonych wartościach innych parametrów wzrost intensywności strumienia zgłoszeń powyżej pewnej wartości powoduje wyraźny spadek wydajności V przy równoczesnym wzroście jednostkowego czasu obsługi $\eta^*(\lambda)$.

Ten pozorny paradoks, którego skutki dla pracy systemu mogą być poważne, wynika z faktu, iż w strumieniu obsługiwanych zgłoszeń głównych zwiększa się ze wzrostem λ, udział zgłoszeń, których czas oczekiwania w kolejce przekroczył D i jest analogiczny do zjawiska zaobserwowanego w punkcie II.5.c.

W tabeli III.7 pokazano, iż przeciwdziałać temu zjawisku można drogą obniżania wartości granicznego, dopuszczalnego czasu oczekiwania τ przy wzroście wydajności źródła zgłoszeń. Również i tu najlepsze wyniki osiąga się przy $\tau = D$.

Tak więc, w pewnych przypadkach celowe może być odrzucenie zgłoszeń, których czas oczekiwania nie przekroczył jeszcze wartości granicznej.

Ponieważ w przypadku, gdy zgłoszenia napływają strumieniem Poissona, pod względem formalnym rozpatrywanie strat zgłoszenia jest równoważne rozpatrywaniu blokowania źródła zgłoszeń, omawianym wynikom można nadać również tę drugą interpretację.

Przedstawione tu wzory i rozważania stanowią uzupełnienie uwag co do celowości oddziaływania na źródło zgłoszeń zawartych w ustępie II.5.c.

Pozostałe, występujące w przedstawionych tabelach prawidłowości są oczywiste i nie wymagają odrębnego komentarza.

Tabela III.6

Wybrane charakterystyki systemu pomocniczego przy $\lambda^{-1} = 0.9$

ϑ	$L(\lambda)$	V	$V-S_g$	Θ_d	$\eta(\lambda)$
0.00	1.53763	1.07611	1.01101	0.06510	0.92756
0.25	1.36860	1.08531	1.01533	0.06999	0.91635
0.50	1.22452	1.09468	1.02148	0.07320	0.89691
0.75	1.10410	1.10342	1.02952	0.07390	0.86844
1.00	1.00671	1.11063	1.03908	0.07155	0.83183
1.25	0.93250	1.11558	1.04938	0.06620	0.78987
1.50	0.88248	1.11798	1.05943	0.05857	0.74656
1.75	0.85877	1.11814	1.06838	0.04976	0.70583
2.00	0.86483	1.11664	1.07576	0.04088	0.67035
2.25	0.90580	1.11419	1.08146	0.03273	0.64127
2.50	0.98897	1.11139	1.08568	0.02572	0.61848
2.75	1.12446	1.10863	1.08868	0.01994	0.60118
3.00	1.32598	1.10611	1.09078	0.01533	0.58831
3.25	1.61202	1.10394	1.00922	0.01172	0.57977
3.50	2.00729	1.10212	1.09320	0.00892	0.57199
3.75	2.54470	1.10064	1.09386	0.00678	0.56700

Tabela III.7

Wybrane charakterystyki systemu pomocniczego przy $\lambda^{-1} = 0.8$ $\vartheta = 1$ w funkcji maksymalnego dopuszczalnego czasu oczekiwania zgłoszenia

τ	$L(\lambda)$	V	$V-S_g$	Θ_d	$\eta^*(\lambda)$
18	6.99142	1.07912	1.05060	0.02852	0.89436
16	4.27436	1.11466	1.07333	0.04133	0.85178
14	2.62612	1.15761	1.10080	0.05682	0.80383
12	1.62578	1.20399	1.12046	0.07353	0.75590
10	1.01696	1.24848	1.15891	0.08957	0.71333
8	0.64235	1.28700	1.18354	0.10346	0.67900
6	0.38630	1.91954	1.20435	0.11519	0.65201

III.5. Dobór parametrów

We wcześniejszych rozważaniach zwrócono uwagę, że poza omówionymi wyżej specjalnymi przypadkami niemożliwe jest analityczne rozpatrzenie systemów WAD z dowolnym regulaminem obsługi i dla wyznaczenia charakterystyk systemu konieczne jest użycie modelowania cyfrowego. Na tej drodze, dla systemów stabilnych, można w każdym przypadku uzyskać oszacowanie charakterystyk i to - teoretycznie rzecz biorąc - z dowolnie dużą dokładnością.

Znaną niedogodnością modelowania cyfrowego jest duża chasochłonność obliczeń, silnie uzależniona od parametrów modelowanego systemu, a szczególnie od stopnia wykorzystania stanowiska obsługi $\eta(\lambda)$. Ze wzrostem $\eta(\lambda)$ rośnie bowiem średnia długość kolejki, wzrasta korelacja pomiędzy poszczególnymi realizacjami badanych zmiennych losowych, a więc uzyskanie założonych przedziałów ufności dla poszukiwanych charakterystyk systemu masowej obsługi wymaga istotnego wydłużenia stosowanych ciągów symulacyjnych.

W tabeli III.8a podano w ślad za [HEID 80] minimalną liczbę obsług l, którą należy zasymulować dla uzyskania z poziomem ufności 90% przedziału ufności równego 10% wartości $E(w)$, przy określaniu średniego czasu oczekiwania w systemie M/M/1, przy różnych wartościach stopnia wykorzystania systemu $\eta(\lambda) = \rho$.

Tabela III.8a

Minimalna liczba obsług pozwalająca na określenie na poziomie ufności 90% średniego czasu oczekiwania z przedziałem ufności 10% wartości średniej [HEID 80]

ρ	0.1	0.2	0.3	0.4	0.5	0.6	0.7	0.8	0.9	0.95	0.99
$E(w)$	0.11	0.250	0.429	0.667	1.00	1.50	2.33	4.00	9.00	19.0	99.0
l [tys]	8.20	6.02	5.83	6.43	7.85	10.60	16.70	33.4	129	455	10^4

Wspomniane w punkcie III.2 metody symulacji, wykorzystujące własności procesów regenerujących się, również nie pozwalają na wyeliminowanie tego zjawiska jako że ze wzrostem $\eta(\lambda)$ rośnie także średnia liczba zgłoszeń obsługiwanych w ciągu pojedynczego okresu czynnego.

Zjawisko to zilustrujemy danymi dotyczącymi systemu M/M/1. W [CRAN 74] podano, iż szacując średni czas oczekiwania w tym prostym systemie masowej obsługi przy użyciu 10 000 okresów czynnych, w zakresie zmienności $\eta(\lambda) \in (0.3, 0.5)$ (przy $E(b) = 1$), uzyskano na poziomie ufności 0.98 przedział ufności wynoszący około 19% dokładnej wartości poszukiwanej charakterystyki.

Ponieważ jednak wartość oczekiwana liczby zgłoszeń obsłużonych w okresie czynnym $\Theta(\lambda)$ określona jest zależnością (III.2.4), więc korzystając ze znanych wyników analizy systemu M/M/1 otrzymujemy:

$$\Theta(\lambda) = \frac{1}{1 - \eta(\lambda)} = \frac{1}{1 - \lambda E(b)}. \qquad \text{(III.5.1)}$$

Tabela III.8b ilustruje przebieg funkcji $\Theta(\lambda)$ przy $E(b) = 1$.

Uzyskanie jednakowego błędu względnego wymagało więc w cytowanej pracy przy $\lambda = 0.5$ nakładu obliczeniowego 1.4-krotnego wyższego aniżeli przy $\lambda = 0.3$ i wynoszącego około 20 000 obsług. Zamodelowanie 10 000 okresów czynnych przy $\eta(\lambda) = 0.8$ wymagałoby już około 50 000 obsług.

Tabela III.8b

Zależność średniej liczby zgłoszeń w okresie czynnym systemu M/M/1 od λ, $E(b) = 1$

λ	.1	.2	.3	.4	.5	.6	.7	.8	.9	.95	.98
$\Theta(\lambda)$	1.11	1.25	1.43	1.67	2.00	2.50	3.33	5.00	10.00	20.00	50.00

Jeszcze gorzej przedstawia się sytuacja w przypadku systemów wielokanałowych M/M/n, dla których zachodzi:

$$\Theta(\lambda) = \sum_{i=0}^{n} \frac{\rho^i}{i!} + \frac{n^n}{n!} \frac{(\frac{\rho}{n})^{n+1}}{1 - \frac{\rho}{n}}, \qquad \text{(III.5.2)}$$

gdzie $\rho = \lambda E(b)$.

Przytoczone przykłady wskazują jednoznacznie na duży nakład czasu maszynowego niezbędny dla uzyskania wyników w oparciu o modelowanie cyfrowe w odniesieniu do jednego zestawu parametrów systemu.
Tak wysoka pracochłonność stawia pod znakiem zapytania możliwość efektywnego wykorzystania modelowania cyfrowego jako narzędzia dla projektowania systemów masowej obsługi, polegającego na doborze zestawu parametrów, przy których charakterystyki systemu spełniają pewne zadane warunki.

Przykładami takich zadań projektowania mogą być (dla założonej postaci zależności czasu obsługi od czasu oczekiwania):
- dobór maksymalnej intensywności strumienia zgłoszeń, przy której stopień wykorzystania stanowiska obsługi nie przekracza pewnej wartości (przy określonej mocy stanowiska obsługi),
- dobór minimalnej mocy stanowiska obsługi, przy której stopień wykorzystania stanowiska obsługi nie przekracza pewnej wartości (przy określonej intensywności strumienia zgłoszeń),
- dobór regulaminu obsługi cechującego się maksymalną intensywnością graniczną (przy określeniu pewnego zbioru rozpatrywanych regulaminów obsługi).

Rozwiązanie każdego z tych zadań w oparciu o modelowanie cyfrowe wymaga przebadania wielu wariantów, w tym często również wariantów cechujących się stosunkowo wysokim stopniem wykorzystania systemu (jak to widać np. z rys. II.3-II.5, stopień wykorzystania systemu może gwałtownie wzrosnąć nawet przy niewielkich zmianach parametrów systemu).

Oczywiście odpowiedni dobór procedur optymalizacji czy też zastosowanie metod planowania eksperymentu może zmniejszyć liczbę rozpatrywanych wariantów, natomiast użycie identycznych ciągów liczb losowych w przypadku porównywania wariantów alternatywnych może zmniejszyć łączny czas symulacji niezbędny dla uzyskania wymaganej dokładności.

Środki te, których dokładniejsze omówienie wykracza poza ramy niniejszej pracy, nie zapewniają jednak radykalnego obniżenia pracochłonności związanej z rozwiązaniem zadań projektowania.

Istotnym krokiem naprzód może natomiast być, jak się wydaje, zastosowanie tzw. metod kombinowanych, tj. łączących modelowanie cyfrowe z wykorzystaniem pewnych, uproszczonych rozwiązań analitycznych. Omówienie szeregu takich metod, pozwalających na znaczne zmniejszenie nakładów obliczeniowych związanych z wyznaczaniem charakterystyk systemów, zawierają prace [PUGA 73], [LAVE 79B], [HEID 80].

W niniejszym punkcie przedstawimy koncepcję wykorzystania do rozwiązywania niektórych zadań powstających przy projektowaniu systemów WAD z dowolnym regulaminem obsługi, uzyskanych w sposób omówiony w rozdziale II rozwiązań analitycznych, dotyczących systemu WAD z identyczną jak w projektowanym systemie postacią zależności czasu obsługi od czasu oczekiwania i naturalnym regulaminem obsługi. Proponowaną metodę omówimy a przykładzie rozwiązywania jednego z zadań projektowania, tj. wyznaczania maksymalnej intensywności strumienia zgłoszeń Λ_{α}, przy której system osiąga zadany stopień wykorzystania α.

Niech będzie dany jednokanałowy system podstawowy z pewnym regulaminem obsługi R taki, że $\eta_R^*(\lambda)$ jest niemalejącą funkcją λ.
Założenie to jest spełnione np., gdy zachodzi:

$$E(b_i) \leq E(b_{i+1}), \quad i=0,1,\ldots,J-1. \qquad \text{(III.5.3)}$$

Przyjmiemy, że potrafimy zamodelować rozpatrywany system, a w szczególności, że potrafimy określić dla każdego $\lambda < T_R$ przedział $[\psi^-(\lambda), \psi^+(\lambda)]$, do którego z wymaganym poziomem ufności γ należy wartość stopnia wykorzystania stanowiska obsługi $\eta_R(\lambda)$.
Oznaczmy wartość oczekiwaną tego parametru przez $\psi(\lambda)$. Zgodnie z regułami omawianymi w rozdziale II potrafimy też wyznaczyć dla każdego $\Lambda < T_{FIFO}$ wartość $\eta^*_{FIFO}(\Lambda)$.
Dla znalezienia oszacowania $\tilde{\Lambda}_{\alpha}$ wartości Λ_{α} proponuje się użycie następującego algorytmu:

Algorytm 1

i) Przyjąć pewną wartość $\lambda_0 < \lambda_1 < T_R$ i wyznaczyć na drodze symulacyjnej $\psi(\lambda_0)$, $\psi(\lambda_1)$.
ii) Stosując podane w rozdziale II metody wyznaczyć wartości $\lambda_{0\alpha}$ oraz $\lambda_{1\alpha}$ takie, że:

$$\eta_{FIFO}(\lambda_{0\alpha}) = \psi(\lambda_0), \qquad \text{(III.5.4)}$$

$$\eta_{FIFO}(\lambda_{1\alpha}) = \psi(\lambda_1), \qquad \text{(III.5.5)}$$

iii) Znaleźć wartość y, spełniającą równanie:

$$\eta_{FIFO}(y) = \alpha. \qquad (III.5.6)$$

iv) Wyznaczyć oszacowanie $\tilde{\Lambda}_\alpha$ jako:

$$\tilde{\Lambda}_\alpha = \lambda_1 + (y - \lambda_{1z}) \frac{\lambda_1 - \lambda_o}{\lambda_{1z} - \lambda_{oz}}. \qquad (III.5.7)$$

Algorytm 1 bazuje na liniowym przekształceniu krzywej $\eta_R(\lambda)$ i prowadziłby do dokładnego określenia wartości Λ_α w przypadku, gdyby $\eta_R(\lambda)$ oraz $\eta_{FIFO}(\lambda)$ były funkcjami liniowymi o znanych dokładnie wartościach.

Zastosowanie algorytmu 1 wymaga, w przypadku ogólnym, zamodelowania rozpatrywanego systemu dla dwóch zestawów parametrów. Przyjmując $\lambda_o = 0$ (a więc $\lambda_{oz} = 0$) możemy użyć tego algorytmu w oparciu o wynik modelowania dla jednego zestawu parametrów. Przyjmując $\alpha = 1$ możemy uzyskać oszacowanie wydajności granicznej.

Rozpatrzmy błąd oszacowania $e_\alpha = \tilde{\Lambda}_\alpha - \Lambda_\alpha$
Na błąd ten składają się:
- błąd wynikający z nieliniowości funkcji $\eta_R(\lambda)$ oraz $\eta_{FIFO}(\lambda)$,
- błąd w określeniu na drodze modelowania cyfrowego, wartości $\eta_R(\lambda_o)$ oraz $\eta_R(\lambda_1)$,
- błąd w rozwiązaniu równań (III.5.4), (III.5.5) oraz (III.5.6), które wymaga użycia metod numerycznych.

Omówimy kolejno te trzy składniki błędu e_α. Na wstępie założymy, że $\psi^+(\lambda) = \psi^-(\lambda) = \psi(\lambda) = \eta(\lambda)$.
Jeżeli $\eta^*_R(\lambda)$ oraz $\eta^*_{FIFO}(\lambda)$ są funkcjami niemalejącymi, to funkcje $\eta_R(\lambda) = \lambda\eta^*_R(\lambda)$ są silnie nieliniowe.
Zauważmy, że przy spełnieniu warunku:

$$\bigvee_{\lambda_o \in [0, T_R)} \left\{ \left. \frac{\partial[\eta_{FIFO}(\lambda)]}{\partial\lambda} \right|_{\lambda=\lambda_{oz}} > \left. \frac{\partial[\eta_R(\lambda)]}{\partial\lambda} \right|_{\lambda_o} \right\}, \qquad (III.5.8)$$

gdzie λ_{oz} spełnia równanie:

$$\eta_{FIFO}(\eta_{oz}) = \eta_R(\lambda_o),$$

zachodzi:

$$\bigvee_{\alpha > \eta_R(\lambda_1)} \{\varepsilon_\alpha < 0\}, \qquad (III.5.9)$$

a więc w szczególności przyjęcie $\alpha = 1$ pozwalałoby wówczas uzyskać oszacowanie od dołu wydajności granicznej systemu.

W takim przypadku celowe jest iteracyjne powtarzanie obliczeń podanych w algorytmie 1. Należy przy tym rozpocząć od $\lambda_o = 0$ oraz małej wartości λ_1 dla uzyskania pewnego oszacowania $\tilde{\Lambda}^{(1)}_\alpha$, a następnie przyjąć $\lambda_o = \lambda_1$, $\lambda_1 = \lambda^{(1)}_\alpha$ i wyznaczyć kolejne oszacowanie $\tilde{\Lambda}^{(2)}_\alpha$. Ze względu na (III.5.9) oraz charakter funkcji $\eta_R(\lambda)$ iteracje te są asymptotycznie zbieżne do dokładnego rozwiązania.

Prowadzone badania pozwoliły na stwierdzenie, iż zależność III.5.8 jest typowa dla systemów WAD z rosnącym w funkcji czasu oczekiwania średnim czasem obsługi. Niemniej jednak nie ma możliwości sprawdzenia czy zachodzi ona dla konkretnego rozpatrywanego regulaminu obsługi. Stąd też w każdym przypadku konieczne jest dodatkowe zamodelowanie systemu przy $\lambda = \tilde{\Lambda}^{(i)}_\alpha$ i sprawdzenie, że istotnie wartość $\eta_R(\tilde{\Lambda}^{(i)}_\alpha)$ jest wystarczająco bliska α.

Biorąc pod uwagę liniowy charakter przekształcenia zastosowanego w algorytmie 1 można oczekiwać, iż błąd e_α będzie tym większy, im silniejsza będzie nieliniowość funkcji $\eta_R(\lambda)$ oraz $\eta_{FIFO}(\lambda)$.

Stąd też na zmniejszenie błędu oszacowania powinno wpłynąć wykorzystanie relacji:

$$\eta_R(\lambda) = \lambda\eta^*_R(\lambda)$$

i odwołanie się do podobieństwa funkcji $\eta^*_R(\lambda)$ i $\eta^*_{FIFO}(\lambda)$, a więc użycie algorytmu 2.

Algorytm 2

i) Przyjąć pewne wartości $\lambda_o < \lambda_1 < T_R$ i wyznaczyć na drodze symulacyjnej $\psi(\lambda_o)$, $\psi(\lambda_1)$.
ii) Stosując podane w rozdziale II metody wyznaczyć wartości λ^*_{or} oraz λ^*_{1r} takie, że:

$$\eta^*_{FIFO}(\lambda_{or}) = \frac{1}{\lambda_o}\psi(\lambda_o), \qquad (III.5.10)$$

$$\eta^*_{FIFO}(\lambda_{1r}) = \frac{1}{\lambda_1}\psi(\lambda_1). \qquad (III.5.11)$$

iii) Znaleźć wartość x spełniającą równanie:

$$x\,\eta^*_{FIFO}\left[\lambda_{1r} + (x-\lambda_1)\frac{\lambda_{1r} - \lambda_{or}}{\lambda_1 - \lambda_o}\right] = \alpha. \qquad (III.5.12)$$

iv) Przyjąć oszacowanie $\tilde{\Lambda}_{\alpha}$ jako:

$$\tilde{\Lambda}_{\alpha} = x. \tag{III.5.13}$$

Przy stosowaniu algorytmu 2 zachodzenie relacji (III.5.9) zapewnia spełnienie warunku:

$$\forall_{\lambda_0 \in [0,T_R]} \left\{ \left| \frac{\partial \eta^*_{FIFO}(\lambda)}{\partial \lambda} \right|_{\lambda=\lambda_{or}} > \left| \frac{\partial [\eta^*_R(\lambda)]}{\partial \lambda} \right|_{\lambda_0} \right\}, \tag{III.5.14}$$

gdzie λ_{or} spełnia równanie:

$$\eta^*_{FIFO}(\lambda_{or}) = \eta_R(\lambda_o).$$

Zarówno przy stosowaniu algorytmu 1 jak i algorytmu 2 nie jest możliwe, ogólnie rzecz biorąc, określenie błędu metody.

Dla zilustrowania dokładności tych metod zamieszczono wyniki testów prowadzonych przy użyciu systemu WAD ze zgłoszeniami napływającymi strumieniem Poissona, zależnością czasu obsługi od czasu oczekiwania postaci (III.1.13) i regulaminem obsługi LIFO.

Wydajność graniczną takiego systemu T_{LIFO} określa (patrz tabela III.1) pierwiastek równania:

$$x\left[B_s + (B_L - B_s)\exp(-xB_s)\right] = 1, \tag{III.5.15}$$

natomiast jednostkowy czas obsługi, zgodnie z analizą podaną w [BUZA 74], można wyznaczyć jako:

$$\eta^*_{LIFO}(\lambda) = B_s + (B_L - B_s)\frac{\lambda B_s - 1 + \exp(-\lambda B_s)}{1 - \lambda(B_L - B_s)}. \tag{III.5.16}$$

Tabela III.9 zawiera wyniki badania błędu względnego oszacowania wydajności granicznej $\tilde{T}^a_{LIFO}$ i $\tilde{T}^b_{LIFO}$ takiego systemu, uzyskane przy użyciu odpowiednio algorytmów 1 i 2 (jak zaznaczono wyżej, wymaga to przyjęcia $\alpha = 1$) przy założeniu, że $\lambda_o = 0$, a więc w oparciu o znajomość dokładnej wartości funkcji $\eta^*(\lambda_1)$ w jednym punkcie $\lambda = \lambda_1$. Przyjęto różne wartości λ_1 oraz trzy różne wartości stosunku $\frac{B_s}{B_L}$, definiującego charakterystykę badanego systemu. Dla porównania podano też błąd względny oszacowania wydajności granicznej uzyskiwany w oparciu o te same informacje przy potraktowaniu badanego systemu jako systemu zachowującego pracę, a więc oszacowanie o postaci:

$$\tilde{T}^c_{LIFO} = \frac{1}{\eta_{LIFO}(\lambda_1)}. \tag{III.5.17}$$

Tabela III.9

Błąd względny oszacowania wydajności granicznej.
A - algorytm 1, B - algorytm 2, C - analogia do systemu klasycznego.
Wyniki podano w procentach wartości dokładnej

λ_1/T_R	B_S=0.03 T_R=12.896			B_S=0.05 T_R=13.181			B_S=0.07 T_R=12.064		
	A	B	C	A	B	C	A	B	C
0.05	-22.46	-21.74	158.35	-24.13	-23.09	51.65	-17.11	-16.06	18.37
0.10	-22.45	-20.98	157.98	-24.12	-22.03	51.39	-17.10	-15.03	18.24
0.15	-22.44	-20.20	157.33	-24.09	-20.95	50.94	-17.07	-14.02	18.01
0.20	-22.40	-19.38	156.34	-24.02	-19.85	50.29	-17.01	-13.03	17.69
0.25	-22.32	-18.52	154.99	-23.29	-18.73	49.43	-16.90	-12.06	17.27
0.30	-22.20	-17.63	153.22	-23.69	-17.60	48.35	-16.75	-11.10	16.76
0.35	-21.99	-16.69	150.97	-23.36	-16.44	47.02	-16.49	-10.17	16.16
0.40	-21.68	-15.72	148.16	-22.89	-15.26	45.44	-16.15	-9.26	15.46
0.45	-21.22	-14.70	144.71	-22.25	-14.06	43.58	-15.70	-8.37	14.67
0.50	-20.56	-13.63	140.51	-21.40	-12.85	41.44	-15.12	-7.50	13.79
0.55	-19.67	-12.52	135.43	-20.31	-11.62	39.00	-14.38	-6.65	12.81
0.60	-18.50	-11.35	129.32	-18.96	-10.38	36.23	-13.48	-5.83	11.74
0.65	-17.02	-10.14	121.96	-17.35	-9.11	33.12	-12.41	-5.02	10.58
0.70	-15.23	-8.87	113.11	-15.48	-7.84	29.65	-11.15	-4.24	9.33
0.75	-13.15	-7.54	102.44	-13.36	-6.55	25.79	-9.71	-3.48	7.99
0.80	-10.82	-6.15	89.52	-11.02	-5.26	21.54	-8.09	-2.74	6.56
0.85	-8.29	-4.71	73.77	-8.49	-3.95	16.85	-6.30	-2.02	5.05
0.90	-5.61	-3.20	54.42	-5.79	-2.64	11.72	-4.35	-1.33	3.45
0.95	-2.83	-1.63	30.36	-2.95	-1.32	6.11	-2.24	-0,65	1.77
0.975	-1.42	-0.82	16.09	-1.49	-0.66	3.12	-1.14	-0.32	0.89

Jest to odpowiednik oszacowania, którego dokładność badaliśmy w punkcie III.5.a w odniesieniu do regulaminu naturalnego.

Zastosowanie obu podanych algorytmów prowadziło w rozpatrywanym przypadku do oszacowania od dołu wydajności granicznej i to oszacowania stosunkowo dokładnego. Zgodnie z oczekiwaniami algorytm 2 dostarczał wyniki obarczone mniejszym błędem. Błąd oszacowania malał też w miarę zwiększania wartości λ_1.

W tabeli III.10 zilustrowano wpływ doboru wartości λ_o na błąd oszacowania wydajności granicznej według algorytmu 2. Jak widać, w przypadku założenia dokładnej znajomości wartości $\eta^*_R(\lambda_o)$ oraz $\eta^*_R(\lambda_1)$, oszacowanie jest tym dokładniejsze im mniejsza jest wartość różnicy $\Delta = \lambda_1 - \lambda_o$.

Wyniki zamieszczone w tabelach III.9 oraz III.10 dotyczyły $\alpha = 1$, a więc przypadku, gdy wartość różnicy $\delta = \alpha - \lambda\eta^*_R(\lambda)$ była największa. Błąd oszacowania wartości Λ_{α} przy $\alpha < 1$ jest zawsze mniejszy i maleje ze zmniejszeniem wartości δ. Prawidłowość tę ilustruje tabela III.11, zawie-

Tabela III.10

Błąd względny oszacowania wydajności granicznej przy użyciu algorytmu 2
$\lambda_o = \lambda_1 - \Delta$; $B_L = 0.1$. Wartości T_R jak w tabeli III.9.
Wyniki podano w procentach wartości ścisłej

$\frac{\lambda_1}{T_R}$	B_S = 0.03	B_S = 0.05	B_S = 0.07	B_S = 0.05					
	Δ = 0.05 T_R			Δ = 0.1 T_R	Δ = 0.2 T_R	Δ = 0.3 T_R	Δ = 0.4 T_R	Δ = 0.5 T_R	Δ = 0.6 T_R
0.05	-21.74	-23.09	-16.06	-23.09	-23.09	-23.09	-23.09	-23.09	-23.09
0.10	-20.31	-21.09	-14.11	-22.03	-22.03	-22.03	-22.03	-22.03	-22.03
0.15	-18.88	-19.14	-12.32	-20.05	-20.95	-20.95	-20.95	-20.95	-20.95
0.20	-17.45	-17.25	-10,63	-18.14	-19,85	-19,85	-19.85	-19.85	-19.85
0.25	-16.02	-15.43	-9,20	-16.28	-17.93	-18.73	-18.73	-18.73	-18.73
0.30	-14.59	-13.69	-7.85	-14.50	-16.08	-17.60	-17.60	-17.60	-17.60
0.35	-13.17	-12.02	-6.64	-12.79	-14.29	-15.74	-16.44	-16.44	-16.44
0.40	-11.77	-10.43	-5.55	-11.10	-12.58	-13.95	-15.26	-15.26	-15.26
0.45	-10.39	-8.93	-4.58	-9.62	-10.95	-12.23	-13.47	-14.07	-14.04
0.50	-9.03	-7.53	-3.73	-8.17	-9.40	-10,60	-11.75	-12.85	-12.85
0.55	-7.72	-6.23	-2.88	-6.82	-7.95	-9.05	-10.11	-11.13	-11.62
0.60	-6.45	-5.04	-2.33	-5.57	-6.60	-7.60	-8.56	-9.49	-10.38
0.65	-5.24	-3.96	-1.77	-4.43	-5.35	-6.24	-7.11	-7.93	-8.73
0.70	-4.11	-3.00	-1.30	-3.41	-4.21	-4.99	-5.75	-6.47	-7.17
0.75	-3.07	-2.16	-0.91	-2.51	-3.19	-3.85	-4.49	-5.11	-5.71
0.80	-2.15	-1.45	-0.59	-1.74	-2.29	-2.83	-3.35	-3.86	-4.34
0.85	-1.35	-0.88	-0.35	-1.09	-1.52	-1.92	-2.33	-2.71	-3.08
0.90	-0.71	-0.44	-0.17	-0.59	-0.87	-1.15	-1.42	-1.69	-1.94
0.95	-0.25	-0.15	-0.06	-0.22	-0.37	-0.51	-0.65	-0.78	-0.91
0.975	-0.1	-0.06	-0.02	-0.09	-0.17	-0.24	-0.31	-0.37	-0.44

rająca błąd względny oszacowania Λ_α w oparciu o algorytm 2 przy różnych wartościach δ.

W dotychczasowych rozważaniach zakładaliśmy precyzyjną znajomość wartości $\eta_R(\lambda_o)$ oraz $\eta_R(\lambda_1)$, co jest oczywiście nieosiągalne na drodze modelowania cyfrowego. W rzeczywistości będziemy mieli do dyspozycji jedynie informacje, że na poziomie ufności γ zachodzi $\eta_R(\lambda_o) \in [\psi^-(\lambda_o), \psi^+(\lambda_o)]$, $\eta_R(\lambda_1) \in [\psi^-(\lambda_1), \psi^+(\lambda_1)]$ (z wyjątkiem przypadku gdy $\lambda_o = \eta_R^*(\lambda_o) = 0$). Na podstawie tych danych możemy wyznaczyć błąd wnoszony przez niedokładną znajomość wartości $\eta_R(\lambda_o)$ oraz $\eta_R(\lambda_1)$.

Tabela III.11

Błąd względny oszacowania wartości Λ_α przy użyciu algorytmu 2
$\alpha = \eta^*(\lambda_1)\lambda_1 + \delta$, $B_S = 0.05$, $B_L = 0.1$, $\Delta = 0.1\ T_R$

$\frac{\lambda_1}{T_R}$	δ = 0.1		δ = 0.2		δ = 0.3		δ = 0.4	
	λ_o=0	$\lambda_o=\lambda_1-\Delta$	λ_o = 0	$\lambda_o=\lambda_1-\Delta$	λ_o = 0	$\lambda_o=\lambda_1-\Delta$	λ_o = 0	$\lambda_o=\lambda_1-\Delta$
0.05	-0.11	-0.11	-0.80	-0.80	-2.57	-2.57	-5.54	-5.54
0.10	-0.18	-0.18	-1.04	-1.04	-3.03	-3.03	-6.08	-6.08
0.15	-0.27	-0.22	-1.31	-1.15	-3.47	-3.14	-6.56	-6.34
0.20	-0.38	-0.26	-1.60	-1.24	-3.89	-3.10	-6.96	-5.92
0.25	-0.49	-0.30	-1.88	-1.31	-4.28	-3.21	-7.28	-5.75
0.30	-0.62	-0.33	-2.10	-1.35	-4.61	-3.18	-7.52	-5.52
0.35	-0.75	-0.35	-2.43	-1.38	-4.89	-3.10	-7.69	-5.26
0.40	-0.89	-0.38	-2.67	-1.39	-5.10	-3.00	-7.78	-4.96
0.45	-1.03	-0.39	-2.88	-1.37	-5.26	-2.86	-7.79	-4.63
0.50	-1.16	-0.40	-3.04	-1.32	-5.35	-2.69	-7.72	-4.28
0.55	-1.27	-0.40	-3.17	-1.27	-5.38	-2.51	-7.60	-3.92
0.60	-1.36	-0,39	-3.24	-1.20	-5.34	-2.31	-7.41	-3.56
0.65	-1.43	-0.38	-3.27	-1.11	-5.24	-2.09	-7.16	-3.20
0.70	-1.48	-0.36	-3.25	-1.02	-5.09	-1.88	-6.85	-2.83
0.75	-1.49	-0.33	-3.18	-0.92	-4.88	-1.66	-6.50	-2.48
0.80	-1.48	-0.31	-3.07	-0.82	-4.63	-1.45	-	-
0.85	-1.43	-0.27	-2.90	-0.71	-	-	-	-
0.90	-1.36	-0.24	-	-	-	-	-	-
0.95	-1.26	-0.21	-	-	-	-	-	-
0.975	-	-	-	-	-	-	-	-

W szczególności dla algorytmu 1 należy wyznaczyć wartości λ_{oz}^+, λ_{oz}^-, λ_{1z}^+, λ_{1z}^-, takie, że:

$$\left.\begin{aligned} \eta_{FIFO}(\lambda_{oz}^+) &= \psi^+(\lambda_o) \\ \eta_{FIFO}(\lambda_{oz}^-) &= \psi^-(\lambda_o) \\ \eta_{FIFO}(\lambda_{1z}^+) &= \psi^+(\lambda_1) \\ \eta_{FIFO}(\lambda_{1z}^-) &= \psi^-(\lambda_1) \end{aligned}\right\} \qquad \text{(III.5.18)}$$

na podstawie których oraz rozwiązania równania (III.5.6), określimy:

$$\tilde{\lambda}^{+}_{\alpha} = \lambda_1 + (\gamma - \lambda^{-}_{1z}) \frac{\lambda_1 - \lambda_o}{\lambda^{-}_{1z} - \lambda^{+}_{oz}} \qquad \text{(III.5.19)}$$

$$\tilde{\lambda}^{-}_{\alpha} = \lambda_1 + (\gamma - \lambda^{+}_{1z}) \frac{\lambda_1 - \lambda_o}{\lambda^{+}_{1z} - \lambda^{-}_{oz}} \qquad \text{(III.5.20)}$$

Możemy więc powiedzieć, iż na poziomie ufności γ wartość oszacowania $\tilde{\lambda}_{\alpha}$ wyznaczona przy znajomości dokładnych wartości $\eta_R(\lambda_o)$ i $\eta_R(\lambda_1)$ mieści się w przedziale $(\tilde{\lambda}^{-}_{\alpha}, \lambda^{+}_{\alpha})$.

W odniesieniu do algorytmu 2 należy wyznaczyć λ^{+}_{or}, λ^{-}_{or}, λ^{+}_{1r}, λ^{-}_{1r} takie że:

$$\left.\begin{aligned} \eta^{*}_{FIFO}(\lambda^{+}_{or}) &= \frac{1}{\lambda_o}\psi^{+}(\lambda_o) \\ \eta^{*}_{FIFO}(\lambda^{-}_{or}) &= \frac{1}{\lambda_o}\psi^{-}(\lambda_o) \\ \eta^{*}_{FIFO}(\lambda^{+}_{1r}) &= \frac{1}{\lambda_o}\psi^{+}(\lambda_1) \\ \eta^{*}_{FIFO}(\lambda^{-}_{1r}) &= \frac{1}{\lambda_o}\psi^{-}(\lambda_1) \end{aligned}\right\} \qquad \text{(III.5.21)}$$

na podstawie których należy znaleźć wartość x^{+} oraz x^{-} jako rozwiązanie równań:

$$x^{+}\eta^{*}_{FIFO}\left[\lambda^{-}_{1r} + (x - \lambda_1)\frac{\lambda^{-}_{1r} - \lambda^{+}_{or}}{\lambda_1 - \lambda_o}\right] = \alpha \qquad \text{(III.5.22)}$$

$$x^{-}\eta^{*}_{FIFO}\left[\lambda^{+}_{1r} + (x - \lambda_1)\frac{\lambda^{+}_{1r} - \lambda^{-}_{or}}{\lambda_1 - \lambda_o}\right] = \alpha \qquad \text{(III.5.23)}$$

W tym przypadku wartość oszacowania $\tilde{\lambda}_{\alpha}$ wyznaczona przy znajomości dokładnych wartości $\eta_R(\lambda_o)$ i $\eta_R(\lambda_1)$ mieści się na poziomie ufności γ w przedziale $(\tilde{\lambda}^{+}_{\alpha} = x^{+}, \tilde{\lambda}^{-}_{\alpha} = x^{-})$.

W tabeli III.12 przedstawiono wartości $\tilde{\lambda}^{+}_{\alpha}, \tilde{\lambda}^{-}_{\alpha}$ uzyskane przy użyciu algorytmu 2. Jak widać, uwzględnienie błędu określenia wartości $\eta_R(\lambda)$ powoduje istotny wzrost błędu oszacowania szczególnie w przypadku małych wartości $\Delta = \lambda_1 - \lambda_o$.

W obu proponowanych algorytmach zachodzi potrzeba numerycznego rozwiązywania równań nieliniowych. Ze względu na założenie o niemalejącym charakterze funkcji $\eta^{*}_{FIFO}(\lambda)$, występujące we wszystkich tych równaniach, posiadają one jednoznaczne rozwiązanie, o ile tylko zachodzi:

$$\lambda^{-}_{1z} > \lambda^{+}_{oz} \qquad \text{(III.5.24)}$$

lub

$$\lambda^{-}_{1r} > \lambda^{+}_{or} \qquad \text{(III.5.25)}$$

Wynika stąd ograniczenie na minimalne wartości $\Delta = \lambda_1 - \lambda_o$, których stosowanie jest dopuszczalne.

Przy spełnieniu nierówności (III.5.24) względnie (III.5.25) znane metody numeryczne zapewniają osiągnięcie praktycznie dowolnie dużej dokładności rozwiązania i błąd wnoszony na tym etapie algorytmu może być pominięty.

W ten sposób omówiliśmy i zilustrowaliśmy przykładami wyników numerycznych wszystkie składowe błędu rozpatrywanego oszacowania. Dla jego praktycznego wykorzystania pozostaje określenie wartości, od której rozpoczyna się analizę.

Ze względu na konieczność zapewnienia, by $\lambda_1 < T_R$ oraz łatwość uzyskiwania dokładnych oszacowań $\eta_R(\lambda_1)$ celowe jest przyjmowanie małych wartości początkowych λ_1.

Ponieważ z reguły rozpatrywać będziemy regulaminy obsługi lepsze aniżeli regulamin naturalny, więc przyjęcie $\lambda_1 = 0.6\text{-}0.8\ T_{FIFO}$ powinno zapewnić stabilność systemu (w przypadkach wątpliwych należy stosować jeszcze mniejsze wartości λ_1).

W punkcie III.1 wyznaczono oszacowanie od góry wydajności granicznej systemu identycznego z systemem służącym do prowadzonych tu badań, przy $B_S = 0.05$, $B_L = 0.1$ (przeskalowanego).

W tym miejscu, posługując się algorytmem 2, podamy oszacowanie od dołu tej wielkości. Tabela III.12 zawiera wyniki uzyskane przy kilku różnych wartościach λ_1 oraz dokładności wyznaczania $\eta_R(\lambda)$ w oparciu o modelowanie badanego systemu w dwóch punktach.

Jak widać, uzyskane oszacowanie cechuje duża dokładność. Przedstawione wyżej testy wykorzystywały wyniki analityczne, dostępne wyłącznie dla nielicznych przypadków specjalnych. Dla określenia przydatności zaproponowanej metody w stosunku do innych systemów przedstawimy wyniki niektórych dalszych eksperymentów.

Dla potrzeb tych eksperymentów opracowane zostały programy symulacyjne umożliwiające wyznaczanie wartości oczekiwanej $\eta_R(\lambda)$ oraz przedziału ufności na poziomie ufności 0.95. Dla wyznaczenia przedziału ufności wartości oczekiwanej jednostkowego czasu obsługi na podstawie "k" kolejnych realizacji czasu obsługi wykorzystano metodę podziału wyników eksperymentu na bloki, podaną w [GROS 74], z uwzględnieniem wpływu skorelowania wyników.

Tabela III.12

Błąd względny oszacowań wydajności granicznej Λ_α^+ oraz Λ_α^- uzyskanych przy użyciu algorytmu 2 i niedokładnej znajomości wartości $\eta_R(\lambda)$. Przyjęto $B_s = 0.05$, $B_L = 0.1$, $\lambda_0 = \lambda_1 - \Delta$

λ_1/T_R		$\lambda_0 = 0$	Δ=0.5 T_R	Δ=0.4 T_R	Δ=0.3 T_R	Δ=0.2 T_R	$\lambda_0 = 0$	Δ=0.5 T_R	Δ=0.4 T_R	Δ=0.3 T_R	Δ=0.2 T_R
		$\psi^+(\lambda) = 1.001\,\eta_R(\lambda)$	$\psi^-(\lambda) = 0.999\,\eta_R(\lambda)$				$\psi^+(\lambda) = 1.005\,\eta_R(\lambda)$	$\psi^-(\lambda) = 0.995\,\eta_R(\lambda)$			
0.75	+	-6.453	-4.732	-4.149	-3.502	-2.782	-6.043	-3.205	-2.730	-2.030	-1.047
	-	-6.653	-5.492	-4.835	-4.199	-3.590	-7.044	-7.027	-6.166	-5.527	-5.100
0.80	+	-5.177	-3.593	-3.104	-2.573	-1.982	-4.850	-2.524	-2.084	-1.509	-0.736
	-	-5.336	-4.119	-3.598	-3.082	-2.582	-5.649	-5.162	-4.554	-4.055	-3.693
0.85	+	-3.887	-2.532	-2.152	-1.746	-1.306	-3.627	-1.797	-1.436	-0.984	-0.426
	-	-4.015	-2.892	-2.499	-2.107	-1.722	-4.266	-3.598	-3.171	-2.799	-2.510
0.90	+	-2.587	-1.566	-1.305	-1.029	-0.733	-2.380	-1.078	-0.822	-0.521	-0.149
	-	-2.689	-1.803	-1.539	-1.273	-1.011	-2.890	-1.266	-1.992	-1.743	-1.540
0.95	+	-1.280	-0.706	-0.572	-0.431	-0.282	-1.115	-0.406	-0.270	-0.115	+0.072
	-	-1.361	-0.851	-0.718	-0.584	-0.450	-1.522	-1.133	-1.000	-0.877	-0.772
		$\psi^+(\lambda) = 1.01\,\eta_R(\lambda)$	$\psi^-(\lambda) = 0.99\,\eta_R(\lambda)$				$\psi^+(\lambda) = 1.02\,\eta_R(\lambda)$	$\psi^-(\lambda) = 0.98\,\eta_R(\lambda)$			
0.75	+	-5.510	-1.260	-0.859	-0.059	-1.369	-4.373	-	+3.256	+4.449	+7.171
	-	-7.515	-9.067	-7.765	-7.074	-6.807	-8.404	-	+10.848	-9.847	-9.721
0.80	+	-4.428	-1.149	-0.742	-0.083	-0.991	-3.534	+1.771	+2.197	+3.127	+5.057
	-	-6.028	-6.465	-5.697	-5.195	-4.962	-6.748	-9.403	-7.845	-7.251	-7.166
0.85	+	-3.291	-0.845	-0.494	+0.080	+0.771	-2.587	+1.186	+1.560	+2.241	+3.543
	-	-4.571	-4.462	-3.973	-3.614	-3.419	-5.153	-6.172	-5.466	-5.092	-5.019
0.90	+	-2.113	-0.444	-0.187	-0.152	-0.640	-1.556	+0.917	+1.194	+1.644	+2.442
	-	-3.135	-2.825	-2.531	-2.297	-2.154	-3.605	-3.889	-3.529	-3.306	-3.245
0.95	+	-0.903	-0.012	+0.128	+0.305	+0.546	-0.463	+0.840	+0.998	+1.232	+1.622
	-	-1.718	-1.469	-1.335	-1.222	-1.146	-2.096	-2.090	-1.948	-1.850	-1.814

Rozpatrzmy jednokanałowy system obsługi z regulaminem LIFO, do którego zgłoszenia napływają strumieniem Poissona, przy założeniu zależności czasu obsługi od czasu oczekiwania postaci:

$$B(x/w) = \begin{cases} \mathbb{H}(x-B_s) & \text{dla} \quad x \leq D \\ \mathbb{H}(x-B_L) & \text{dla} \quad x > D \end{cases} \qquad \text{(III.5.26)}$$

gdzie: $B_s = 0.05$, $B_L = 0.1$, $D = 0.2$.

Przy użyciu algorytmu 2 będziemy dla takiego systemu poszukiwali takiej wartości intensywności strumienia zgłoszeń Λ_α, przy której $\eta_{LIFO}(\Lambda_\alpha) = \alpha = 0.8$.

Przyjmijmy $\lambda_0 = 0$, $\lambda_1 = T_{FIFO} = 10$.

Po zrealizowaniu 20 tys. obsług uzyskano przy $\lambda = 10$:

$$\psi(10) = 50.65 * 10^{-2}, \quad \psi^+(10) = 50.73 * 10^{-2}, \quad \psi^-(10) = 50.57 * 10^{-2},$$

co prowadzi do oszacowania

$$\tilde{\Lambda}_\alpha = 10.98, \quad \Lambda_\alpha^+ = 11.06, \quad \Lambda_\alpha^- = 10.90.$$

Przeprowadzenie eksperymentu symulacyjnego przy intensywności strumienia zgłoszeń $\lambda = 11$ prowadziło, po wykonaniu 20 tys. obsług, do wartości

$$\psi(11) = 56.05 * 10^{-2}, \quad \psi^+(11) = 56.16 * 10^{-2}, \quad \psi^-(11) = 55.94 * 10^{-2}.$$

Przyjmując $\lambda_0 = 10$, $\lambda_1 = 11$ uzyskano kolejne oszacowanie

$$\tilde{\Lambda}_\alpha = 13.95 \quad \Lambda_\alpha^+ = 15.40 \quad \Lambda_\alpha^- = 12.84.$$

Eksperyment symulacyjny przy $\lambda = 14$ pozwolił uzyskać, po wykonaniu 20 tys. obsług, wartości:

$$\psi(14) = 73.57 * 10^{-2}, \quad \psi^+(14) = 73.77 * 10^{-2}, \quad \psi^-(14) = 73.76 * 10^{-2}$$

natomiast uzyskane przy założeniu wartości $\lambda_0 = 11$, $\lambda_1 = 14$ oszacowania miały wartości:

$$\tilde{\Lambda}_\alpha = 14.83, \quad \Lambda_\alpha^+ = 14.95, \quad \Lambda_\alpha^- = 14.73.$$

Biorąc pod uwagę fakt, iż modelowanie rozpatrywanego systemu przy $\lambda = 15$ prowadziło do wyników:

$$\psi(15) = 0.800, \quad \psi^+(15) = 0.805, \quad \psi^-(15) = 0.794,$$

uzyskane w oparciu o wykonanie trzech, a nawet tylko dwóch eksperymentów symulacyjnych, wyniki można uważać za dobre. Zauważmy, że w rozpatrywanym systemie wydajność graniczna T_{LIFO} przekracza w sposób istotny wartość T_{FIFO}, skoro $?_{LIFO}(1.5\ T_{FIFO}) = 0.8$ a więc różni się on w sposób bardzo silny od systemu stanowiącego podstawę do oszacowań.

Biorąc pod uwagę brak jakichkolwiek wyników analitycznych w przypadku wielokanałowych systemów WAD, warto zastanowić się, czy uzyskane w rozdziale II wyniki mogłyby w jakimś stopniu ułatwić dobór parametrów również i w tym przypadku.

Problem oszacowania charakterystyk systemów wielokanałowych przy użyciu odpowiednio dobranych systemów jednokanałowych doczekał się wielu analiz (np. [ARJA 78], [SOND 79] zawierające też informacje o wcześniejszych pracach), prowadząc na ogół do pewnych stochastycznych nierówności. Nie można przy tym zwykle określić ścisłości uzyskanych oszacowań. Bardzo bogaty i wartościowy zbiór wyników badań symulacyjnych określających dokładność oszacowań omawianych w [ARJA 78] zawiera [LEHT 78].

W szczególności badając szeroką klasę systemów typu G/G/n stwierdzono, że ich charakterystyki można stosunkowo dobrze przybliżać charakterystykami systemów jednokanałowych o identycznym strumieniu zgłoszeń i dystrybuancie czasu obsługi $B(t) = \mathcal{E}(nt)$, gdzie $\mathcal{E}(t)$ dystrybuanta czasu obsługi pojedynczego kanału obsługi. W szczególności stwierdzono, że czas oczekiwania w tak zdefiniowanym systemie jednokanałowym jest większy aniżeli czas oczekiwania w systemie wielokanłowym (stanowi to uogólnienie prawidłowości zaobserwowanych dla systemów M/M/n i omawianych w odniesieniu do wartości oczekiwanych, np. w [WEGR 75]).

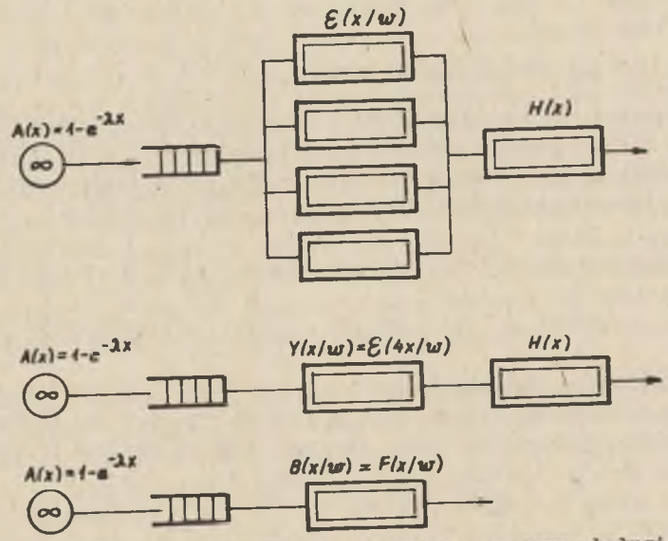
Tak więc w przypadku wielokanałowych systemów WAD z czasem obsługi rosnącym w funkcji czasu oczekiwania zadanym dystrybuantą warunkową $\mathcal{E}(x/w)$ można się spodziewać, że czas oczekiwania w_n będzie nie większy aniżeli czas oczekiwania w systemie jednokanałowym o identycznym strumieniu zgłoszeń i regulaminie obsługi oraz dystrybuancie warunkowej czasu oczekiwania $B(x/w) = \mathcal{E}(nx/w)$.

W efekcie stopień wykorzystania systemu n-kanałowego nie powinien dla żadnej wartości intensywności strumienia zgłoszeń λ przewyższać stopnia wykorzystania zdefiniowanego wyżej systemu jednokanałowego, co uzyskało potwierdzenie w przeprowadzonych, wyrywkowych eksperymentach symulacyjnych.

Tym samym przy wykorzystaniu wskazanego wyżej sposobu konstrukcji systemu jednokanałowego można wyniki jego analizy wykorzystać dla usprawnienia doboru intensywności strumienia zgłoszeń zgodnie z zaproponowanymi algorytmami, uzyskując (pod warunkiem spełnienia nierówności (III.5.8) względnie (III.5.14)) oszacowania od dołu poszukiwanych wartości, w tym też wydajności granicznej.

Nie analizując szczegółowo dokładności metody w odniesieniu do systemów wielokanałowych, zamieścimy przykład wskazujący na opłacalność jej

stosowania nawet w bardziej jeszcze rozbudowanych systemach. Na rys. III.6a przedstawiono schemat dwuetapowego systemu obsługi bez bufora pośredniego.



Rys. III.6. Schemat przykładowego, dwuetapowego systemu obsługi z zależnością czasu obsługi od czasu oczekiwania na pierwszym etapie i jego uproszczenia

$$E(x/w) = \begin{cases} H(x-0.2) & \text{dla} \quad w \leq 0.2 \\ H(x-0.4) & \text{dla} \quad w > 0.2 \end{cases} \qquad H(x) = H(x-0.05)$$

$$F(x/w) = \begin{cases} \Phi_0(x) & \text{dla} \quad w = 0 \\ \Phi_1(x) & \text{dla} \quad 0 < w \leq 0.2 \\ \Phi_2(x) & \text{dla} \quad w > 0.2 \end{cases}$$

gdzie $\Phi_0(x)$, $\Phi_1(x)$, $\Phi_2(x)$ określone są, zgodnie z zależnościami (IV.6.12) i (IV.6.13), przy $B_s = 0.05$, $B_L = 0.1$, $\alpha = 0.05$

Wprowadzając, zgodnie z wyżej omówionymi zasadami, stanowisko jednokanałowe na pierwszym etapie, uzyskujemy system jak na rysunku III.6b, który - jak to łatwo wykazać korzystając z metod redukcji dwuetapowych, pojedynczych systemów obsługi omawianych bardziej szczegółowo w rozdziale IV - jest równoważny pod względem wydajności systemowi przedstawionemu na rys. III.6c.

Jako podstawę do oszacowania stopnia wykorzystania systemu przedstawionego na rys. a przyjęto uproszczony wariant systemu jednostanowiskowego o zależności czasu obsługi od czasu oczekiwania w postaci:

$$B(x/w) \approx F(x/w),$$

$$B(x/w) = \begin{cases} \Phi_1(x) = \mathbb{H}(x-0.05) & \text{dla} \quad w \leq 0.2 \\ \Phi_2(x) = \mathbb{H}(x-0.1) & \text{dla} \quad w > 0.2 \end{cases}$$

zachowujący wartość wydajności granicznej.

Zakładając, że w rozpatrywanym systemie dwuetapowym stosowany jest regulamin LIFO, postaramy się wyznaczyć intensywność strumienia zgłoszeń, przy której stopień wykorzystania stanowisk pierwszego etapu (z uwzględnieniem blokowania) wynosić będzie $\eta_{LIFO}(\Lambda_\alpha) = \alpha = 84.4$.

Modelowanie cyfrowe tego systemu przy $\lambda_0 = 10$ pozwoliło wyznaczyć:

$$\psi(10) = 53.54 \quad \psi^+(10) = 53.61 \quad \psi^-(10) = 53.48$$

a stąd, stosując algorytm 2 przy $\lambda_0 = 0$ uzyskano oszacowanie $\Lambda_\alpha = 10.24$. Ponieważ modelowanie cyfrowe przy $\lambda = 10.24$ wykazało, iż uzyskana wartość Λ_α była zdecydowanie zaniżona, przyjęto arbitralnie $\lambda = 12.5$, uzyskując

$$\psi(12.5) = 68.22 \quad \psi^+(12.5) = 68.36 \quad \psi^-(12.5) = 68.10$$

a stąd, stosując algorytm 2 przy $\lambda_0 = 10$, $\lambda_1 = 12.5$ uzyskano wartości $\tilde{\Lambda}_\alpha^+ = 15.16$ oraz $\tilde{\Lambda}_\alpha^- = 14.87$, które okazały się zaskakująco trafne wobec wyników modelowania cyfrowego systemu przy $\lambda = 15$, kiedy to

$$\psi(15) = 84.37 \quad \psi^+(15) = 84.86 \quad \psi^-(15) = 83.87$$

Dla ścisłości należy dodać, iż dokładność prognozy nie zależy w tym przypadku w sposób istotny od wyboru λ_1, ponieważ bardzo zbliżone wartości oszacowań uzyskano przy $\lambda_1 = 13$, $\lambda_1 = 12{,}5$, $\lambda_1 = 14$.

Widzimy więc, iż przedstawiona metoda może istotnie ułatwiać dobór parametrów nawet w przypadkach bardziej złożonych aniżeli jednokanałowe systemy obsługi.

Podstawą dla uzyskania oszacowań poszukiwanych parametrów jest w zaproponowanej metodzie liniowe odwzorowanie charakterystyk systemu WAD z regulaminem naturalnym. Można byłoby oczywiście podjąć próbę uogólnienia takiego sposobu postępowania na odwzorowanie nieliniowe z wykorzystaniem klasycznych metod identyfikacji dla określania ich parametrów, co wymagałoby jednak zwiększenia liczby eksperymentów niezbędnych dla określenia takiej prognozy i wobec zadowalających wyników zyskiwanych przy użyciu odwzorowania liniowego nie wydaje się celowe.

W analogiczny sposób, jak to omówiono wyżej w odniesieniu do doboru intensywności strumienia zgłoszeń, można też realizować dobór mocy stanowiska obsługi zapewniającej wymagany stopień jego wykorzystania. W celu zachowania kolejności eksperymentów zapewniających wzrost stopnia wykorzystania stanowiska obsługi należy rozpoczynać je od dużych wartości mocy stanowiska obsługi. Podobnie można też dobierać parametry systemu pod kątem uzyskania właściwych wartości innych charakterystyk.

Specjalną rolę w projektowaniu systemów WAD odgrywa problematyka doboru regulaminu obsługi, a szczególnie zagadnienie maksymalizacji na tej drodze wydajności granicznej. Udowodnienie optymalności regulaminu obsługi z punktu widzenia maksymalizacji wydajności granicznej jest zwykle niemożliwe nawet w przypadku najprostszych systemów WAD (problematyką tą zajmiemy się bardziej szczegółowo w następnym punkcie). Dostępne są jedynie pewne wskazania dla doboru regulaminu, przy czym w pełni uzasadnione decyzje w tym zakresie mogą być podjęte jedynie drogą bezpośredniego porównania wydajności granicznych różnych, branych pod uwagę regulaminów obsługi.

Ponieważ, jak to wcześniej omawiano, wyznaczenie wydajności granicznej dla konkretnego regulaminu obsługi napotyka na znaczne trudności, cenna wydaje się możliwość uzyskiwania dla szerokiej klasy regulaminów obsługi oszacowania od dołu wydajności granicznej systemów WAD w oparciu o algorytm 2 (lub algorytm 1).

W połączeniu z metodą szacowania wydajności granicznej od góry, zaprezentowaną w punkcie III.2, uzyskuje się możliwość stosunkowo dokładnego (jak to widać na przykładzie systemu z zależnością czasu obsługi od czasu oczekiwania zadaną wzorem (III.1.13), dla którego oba te oszacowania zostały określone) wyznaczenia tego parametru kolejno dla różnych regulaminów obsługi i wyboru najlepszego spośród nich.

III.6. Wskazania dla doboru regulaminu obsługi

Z zamieszczonych w niniejszym rozdziale rozważań wynika jednoznacznie, że warunki stabilności systemów WAD zależą w sposób istotny od przyjętego regulaminu obsługi. Należy więc oczekiwać, iż w praktyce często stawiane będzie pytanie o możliwość wskazania, dla pewnego rozpatrywanego systemu WAD, regulaminu obsługi charakteryzującego się największą wydajnością graniczną. Problem ten został po raz pierwszy postawiony w [BUZA 74], w odniesieniu do rozpatrywanego tam prostego przypadku systemu z zależnością czasu obsługi od czasu oczekiwania określoną wzorem (III.1.13), dla którego w punkcie III.1d przytoczono wynikające z bezpośredniej analizy systemów nierówności, wskazujące, który z rozpatrywanych dwóch regulaminów obsługi FIFO(ψ^*), LIFO(ψ^*) jest lepszy, oraz jak należy dla tych regulaminów dobierać horyzont prognozy τ^* (relacje (III.1.23)-(III.1.25)). W cytowanej pracy wykazano, że nie jest możliwe określenie optymalnego regula-

minu obsługi (przy zadanym poziomie informacji o strumieniu zgłoszeń) przy użyciu programowania dynamicznego - w szczególności na przykład nie można wykazać, że przy braku prognozy optymalny jest regulamin LIFO.

Poniżej przedyskutujemy możliwość określenia wskazań dla doboru regulaminu obsługi w nieco ogólniejszym aniżeli rozpatrywany w [BUZA 74] modelu, wykorzystując warunek wystarczający ergodyczności łańcucha Markowa określającego proces zmian stanu systemu w pewnych wybranych chwilach czasu. Wnioski wynikające z tych rozważań w połączeniu z innymi prawidłowościami posłużą następnie do sformułowania pewnych, mających częściowo charakter heurystyczny, wskazań dla ogólniejszych p. zypadków.

W szczególności rozważania rozpoczniemy od przedyskutowania przypadku systemu jednokanałowego ze zgłoszeniami napływającymi strumieniem Poissona (o intensywności λ) i zależnością czasu obsługi od czasu oczekiwania postaci:

$$B(x/w) = \begin{cases} B(x-B_s) & \text{dla} \quad w \leq D, \\ B(x-B_L) & \text{dla} \quad w > D, \end{cases} \qquad \text{(III.6.1)}$$

przy

$$B_s \leq D \leq (B_L - B_s), \qquad B_s < B_L. \qquad \text{(III.6.2)}$$

a) Proces zmian stanu systemu

Rozpatrzmy zdefiniowany zależnościami (III.6.1) i (III.6.2) system w chwilach ξ_i, i=0,1,... następujących bezpośrednio po opuszczeniu stanowiska obsługi przez kolejne zgłoszenie.

Niech $\mathcal{K}_i$ oznacza zbiór zgłoszeń oczekujących w kolejce w chwili ξ_i, takich, których czas oczekiwania nie przekroczył D, natomiast $\mathcal{U}_i$ - zbiór pozostałych zgłoszeń. Liczebności tych zbiorów oznaczymy odpowiednio przez K_i i U_i.

Obsługę trwającą przez czas B_s nazywać będziemy "krótką", natomiast trwającą przez czas B_L nazywać będziemy "długą".

Ponieważ wszystkie zgłoszenia należące do zbioru $\mathcal{U}_i$ są między sobą identyczne, więc sposób ich uszeregowania nie wpływa na wydajność systemu.

Regulamin obsługi powinien więc rozstrzygać:

i) o rozpoczęciu obsługi lub oczekiwaniu na napłynięcie nowego zgłoszenia (przy niepustej kolejce),

ii) o wyborze do obsługi zgłoszenia ze zbioru $\mathcal{U}_i$ lub $\mathcal{K}_i$,

iii) o uszeregowaniu zgłoszeń należących do $\mathcal{K}_i$.

Ponieważ $D \leq B_L$, więc wybór w chwili ξ_i do obsługi zgłoszenia należącego do zbioru $\mathcal{U}_i$ oznacza, iż wszystkie zgłoszenia należące do zbioru $\mathcal{K}_i$ zostaną zaliczone do zbioru $\mathcal{U}_{i+1}$.

Z drugiej strony, niezależnie od zastosowanego regulaminu obsługi, co najwyżej

$$M = \text{entier}\left(\frac{D}{B_s}\right) + 1 - \text{sgn}\left[B_s \,\text{entier}\left(\frac{D}{B_s}\right) - D\right] \qquad \text{(III.6.3)}$$

$$\text{sgn}(x) = \begin{cases} 1 & \text{gdy} \quad x \geq 0 \\ 0 & \text{gdy} \quad x < 0 \end{cases}$$

spośród zgłoszeń należących do zbioru $\mathcal{K}_i$ może uzyskać obsługę krótką, natomiast pozostałe wymagają obsługi długiej.

Utwórzmy w każdej chwili ξ_i zbiór $\mathcal{S}_i$ o liczebności S_i będący podzbiorem $\mathcal{K}_i$. Zaliczymy do niego te spośród zgłoszeń należących do $\mathcal{K}_i$, które skorzystałyby zgodnie z obowiązującym regulaminem obsługi z obsługi krótkiej, gdyby założyć że począwszy od chwili ξ_i obsługiwane są wyłącznie zgłoszenia należące do $\mathcal{K}_i$ i proces ten nie jest zakłócany napływem nowych zgłoszeń.

Oczywiście

$$S_i \leq \min\left\{K_i, M\right\}. \qquad \text{(III.6.4)}$$

Utwórzmy też zbiór $\mathcal{L}_i$

$$\mathcal{L}_i = \mathcal{U}_i \vee (\mathcal{K}_i / \mathcal{S}_i), \qquad \text{(III.6.5)}$$

o liczebności

$$L_i = K_i + U_i - S_i. \qquad \text{(III.6.6)}$$

Przy braku dodatkowych założeń dotyczących regulaminu obsługi i możliwości napływania nowych zgłoszeń do zbioru $\mathcal{K}_i$ bądź to do zbioru $\mathcal{S}_i$ bądź to do zbioru $\mathcal{K}_i$ zaliczenie zgłoszenia należącego do zbioru $\mathcal{S}_i$ bądź to do zbioru $\mathcal{L}_i$ nie musi przesądzać o czasie jego rzeczywistej obsługi.

Przyjmiemy obecnie dodatkowe założenia ograniczające klasę rozpatrywanych regulaminów obsługi.

- Przyjmiemy, że prawdopodobieństwo wyboru w chwili ξ_i do obsługi zgłoszenia wymagającego obsługi "krótkiej" (tj. należącego do zbioru $\mathcal{S}_i$) jest zależne wyłącznie od regulaminu obsługi $\mathcal{R}$ oraz wartości S_i, L_i i wynosi $r_{\mathcal{R}}(S_i, L_i)$, przy czym

$$\bigvee_{\substack{S_i = j \\ j=0,1,2,\ldots,M}} \left\{ \lim_{p \to \infty} r_{\mathcal{R}}(j,p) = r^*_{\mathcal{R}}(j) \right\}. \qquad \text{(III.6.7)}$$

- Przyjmiemy też, że wszystkie zgłoszenia zaliczone w chwili ξ_i do zbioru S_i będą obsługiwane w kolejności wyznaczonej regulaminem obsługi w chwili ξ_i, bez względu na proces nadchodzenia dalszych zgłoszeń, o ile tylko w którejś z chwili $\xi_i, \xi_{i+1} \dots \xi_{i+S_i}$ nie zostanie podjęta decyzja o wykonaniu obsługi długiej.
- W chwilach czasu ξ_i, i=0,1,..., dostępna jest prognoza napłynięcia nowego zgłoszenia w przedziale czasu $(\xi_i, \xi_i + \vartheta)$. Oczywiście prognoza jest negatywna ze stałym prawdopodobieństwem $\beta = e^{-\lambda\vartheta}$. Decyzja o rozpoczęciu obsługi jednego ze zgłoszeń oczekujących w kolejce bądź oczekiwania na napłynięcie nowego zgłoszenia podejmowane jest wyłącznie w oparciu o znajomość tej prognozy, S_i oraz L_i.
- W przypadku, gdy zbiór S_i jest niepusty, nie dopuszcza się bezczynnego oczekiwania stanowiska obsługi na napłynięcie nowego zgłoszenia.

Zdefiniujmy stan systemu w chwili ξ_i jako

$$X_i = S_i B_S + L_i B_L, \qquad \text{(III.6.8)}$$

a więc równy zapasowi pracy w systemie w przypadku wykonania najpierw kolejno S_i obsług "krótkich".

Z (III.6.2) oraz (III.6.4) wynika wprost, że X_i określa jednoznacznie S_i oraz L_i.

Stan systemu w kolejnych chwilach czasu tworzy jednowymiarowy łańcuch Markowa $\{X_i\}$ przyjmujący wartości rzeczywiste, nieujemne należące do przeliczalnego zbioru $\mathcal{B}$

$$\mathcal{B} = \left\{ x_k : x_k = m.B_S + lB_L \quad \begin{matrix} k = 1,2,\dots \\ m = 0,1,\dots,M \\ l = 0,1,\dots \end{matrix} \right\} \qquad \text{(III.6.9)}$$

i posiadający macierz przejść $[pij]$,

$$p_{ij} = Pr\left\{X_{k+1} = x_j / X_k = x_i\right\}, \quad x_i, x_j \in \mathcal{B} \qquad \text{(III.6.10)}$$

Łatwo widać, że łańcuch ten jest nieprzywiedlny i nieokresowy.

Wartości poszczególnych elementów macierzy przejść $[pij]$ są istotnie zależne od postaci przyjętego regulaminu obsługi zgłoszeń (oraz, oczywiści, od parametru λ strumienia zgłoszeń).

Dla podkreślenia tej zależności oraz określenia, jaki regulamin mamy na myśli, w przypadku porównywania różnych regulaminów obsługi będziemy dalej pisać $\{X_i(\mathcal{R})\}$ oraz $[p_{ij}(\mathcal{R})]$.

Rozpatrując stan systemu w chwilach następujących bezpośrednio po zakończeniu obsługi dwóch kolejnych zgłoszeń ξ_i, ξ_{i+1} uzyskujemy następującą zależność $X_{i+1}(\mathcal{R})$ od $X_i(\mathcal{R})$

$$X_{i+1}(\mathcal{R}) = \varkappa_{\mathcal{R}} \left\{ X_i(\mathcal{R}) - B_s + \psi(\xi_{i+1} - \xi_i) B_s + \alpha_{\psi}(S_i)\left[B_L - B_S\right] \right\} +$$

$$+ (1-\varkappa_{\mathcal{R}}) \left\{ X_i(\mathcal{R}) - B_L + S_i(B_L - B_S) + \psi(\xi_{i+1} - \xi_i - D) B_L + \right.$$

$$\left. + \psi(D) B_S + \Gamma_{\psi}(B_L - B_s) \right\} \qquad \text{(III.6.11)}$$

gdzie:

- $L_i = \text{entier}\left[\frac{X_i(\mathcal{R})}{B_L}\right]$, $S_i = \text{entier}\left[X_i(\mathcal{R}) - L_i B_L\right]$,
- $\varkappa_{\mathcal{R}}$ jest zmienną losową przyjmującą wartość 1 z prawdopodobieństwem $r_{\mathcal{R}}(S_i, L_i)$ lub wartość 0 z prawdopodobieństwem $1 - r_{\mathcal{R}}(S_i, L_i)$.
- Zmienna losowa $\psi(\Delta)$ opisuje liczbę zgłoszeń napływających do systemu w przedziale czasu Δ. Zgodnie z założeniem o postaci strumienia zgłoszeń zmienna ta posiada rozkład Poissona

$$Pr\left\{\psi(\Delta) = k\right\} = p_k^{\Delta} = \frac{(\Delta\lambda)^k}{k!} e^{-\lambda\Delta} \qquad \text{(III.6.12)}$$

- Zmienna losowa $\alpha_{\psi}(S_i)$ oznacza - w przypadku obsługi krótkiej - liczbę zgłoszeń, które nadeszły w odcinku czasu $[\xi_i, \xi_{i+1}]$ i nie zostały zaliczone do zbioru S_{i+1}. Liczba takich zgłoszeń zależy od S_i, ponieważ zgodnie z założeniem o klasie regulaminów obsługi w zbiorze S_{i+1} musi znaleźć się $S_i - 1$ zgłoszeń ze zbioru S_i.
- Zmienna losowa Γ_{ψ} oznacza - w przypadku obsługi długiej - liczbę zgłoszeń, które mimo nadejścia w odcinku czasu $[\xi_{i+1} - D, \xi_{i+1}]$ włączone zostaną do zbioru $\mathcal{A}_{i+1}$ (zależna od regulaminu obsługi zgłoszeń).

b) <u>Zalecane regulaminy obsługi dla rozpatrywanego, prostego systemu jednokanałowego</u>

Dla porównania różnych regulaminów obsługi korzystać będziemy z następującego warunku wystarczającego ergodyczności łańcuchów Markowa nieprzerywalnych, nieokresowych, przyjmujących wartości rzeczywiste należące do przeliczalnego zbioru $\mathcal{B} = \{x_i,\ i=1,2,\dots\}$.

<u>Warunek 3</u> [ZSAK 79]

Niech γ będzie określone zależnością

$$\gamma_i = E\left\{X_{j+1} - X_j / X_j = x_i\right\} = \sum_j x_j p_{ij} - x_i \qquad \text{(III.6.13)}$$

Warunkiem wystarczającym ergodyczności łańcucha $\{X_i\}$ jest, aby

i) $$\bigvee_{x_i \in \mathcal{B}} |\gamma_i| < \infty$$

ii) $$\lim_{x_i \to \infty} \sup \gamma_i < 0.$$

$$X_{i+1}(\mathcal{A}) = \varkappa_{\mathcal{A}}\left\{X_i(\mathcal{A}) - B_s + \psi(\xi_{i+1}-\xi_i)B_s + \varkappa_\psi(S_i)\left[B_L - B_S\right]\right\} +$$

$$+ (1-\varkappa_{\mathcal{A}})\left\{X_i(\mathcal{A}) - B_L + S_i(B_L-B_S) + \psi(\xi_{i+1} - \xi_i - D)B_L + \right.$$

$$\left. + \psi(D)B_S + \Gamma_\psi(B_L-B_s)\right\} \qquad \text{(III.6.11)}$$

gdzie:

- $L_i = \text{entier}\left[\frac{X_i(\mathcal{A})}{B_L}\right]$, $S_i = \text{entier}\left[X_i(\mathcal{A}) - L_i B_L\right]$,
- $\varkappa_{\mathcal{A}}$ jest zmienną losową przyjmującą wartość 1 z prawdopodobieństwem $r_{\mathcal{A}}(S_i, L_i)$ lub wartość 0 z prawdopodobieństwem $1 - r_{\mathcal{A}}(S_i, L_i)$.
- Zmienna losowa $\psi(\Delta)$ opisuje liczbę zgłoszeń napływających do systemu w przedziale czasu Δ. Zgodnie z założeniem o postaci strumienia zgłoszeń zmienna ta posiada rozkład Poissona

$$\Pr\left\{\psi(\Delta) = k\right\} = p_k^\Delta = \frac{(\Delta\lambda)^k}{k!}\, e^{-\lambda\Delta} \qquad \text{(III.6.12)}$$

- Zmienna losowa $\varkappa_\psi(S_i)$ oznacza - w przypadku obsługi krótkiej - liczbę zgłoszeń, które nadeszły w odcinku czasu $\left[\xi_i, \xi_{i+1}\right]$ i nie zostały zaliczone do zbioru S_{i+1}. Liczba takich zgłoszeń zależy od S_i, ponieważ zgodnie z założeniem o klasie regulaminów obsługi w zbiorze S_{i+1} musi znaleźć się S_i-1 zgłoszeń ze zbioru S_i.
- Zmienna losowa Γ_ψ oznacza - w przypadku obsługi długiej - liczbę zgłoszeń, które mimo nadejścia w odcinku czasu $\left[\xi_{i+1} - D, \xi_{i+1}\right]$ włączone zostaną do zbioru $\mathcal{C}_{i+1}$ (zależna od regulaminu obsługi zgłoszeń).

b) Zalecane regulaminy obsługi dla rozpatrywanego, prostego systemu jednokanałowego

Dla porównania różnych regulaminów obsługi korzystać będziemy z następującego warunku wystarczającego ergodyczności łańcuchów Markowa nieprzerywalnych, nieokresowych, przyjmujących wartości rzeczywiste należące do przeliczalnego zbioru $\mathcal{B} = \left\{x_i,\ i=1,2,\ldots\right\}$.

Warunek 3 [ZSAK 79]

Niech γ będzie określone zależnością

$$\gamma_i = E\left\{X_{j+1} - X_j / X_j = x_i\right\} = \sum_j x_j p_{ij} - x_i \qquad \text{(III.6.13)}$$

Warunkiem wystarczającym ergodyczności łańcucha $\left\{X_i\right\}$ jest, aby

i) $$\bigvee_{x_i \in \mathcal{B}} |\gamma_i| < \infty$$

ii) $$\lim_{x_i \to \infty} \sup \gamma_i < 0.$$

Warunek ten, stanowiący uogólnienie udowodnionego przez Pakesa Warunku 1, wykorzystywanego w punkcie III.1 (do którego sprowadza się natychmiast, gdy B_L jest wielokrotnością B_s), wynika z twierdzenia pracy [TWEE 76], którego spełnienie założeń można łatwo sprawdzić.

Rozpatrzmy dwa regulaminy obsługi $\mathcal{A}_1$ oraz $\mathcal{A}_2$.
Niech $\Theta_{\mathcal{A}_1}$ i $\Theta_{\mathcal{A}_2}$ oznaczają suprema zbiorów wartości λ, dla których spełniony jest Warunek 3 przy zastosowaniu odpowiednio regulaminu obsługi $\mathcal{A}_1$ lub $\mathcal{A}_2$. Ponieważ korzystamy wyłącznie z warunku wystarczającego, więc $\Theta_{\mathcal{A}_1} \leq T_{\mathcal{A}_1}$, $\Theta_{\mathcal{A}_2} \leq T_{\mathcal{A}_2}$.
Jeżeli zachodzić będzie nierówność $\Theta_{\mathcal{A}_1} \geq \Theta_{\mathcal{A}_2}$, to będziemy mówili, że regulamin obsługi $\mathcal{A}_1$ jest zalecany w porównaniu z regulaminem obsługi $\mathcal{A}_2$, nie dysponując bowiem możliwością bezpośredniego porównania wydajności granicznych $T_{\mathcal{A}_1}$, $T_{\mathcal{A}_2}$, nie możemy stwierdzić, czy $\mathcal{A}_1$ jest lepszy od $\mathcal{A}_2$.

Na podstawie (III.6.11) oraz (III.6.13) mamy:

$$\gamma_i = r_{\mathcal{A}}(S_i, L_i)\left\{-B_s + E\left[\psi(\xi_{i+1}-\xi_i)\right]B_S + E\left[\varkappa_\psi(S_i)\right](B_L - B_S)\right\} +$$

$$+ \left[1 - r_{\mathcal{A}}(S_i, L_i)\right]\left\{-B_L + S_i(B_L - B_S) + E\left[\psi(\xi_{i+1}-\xi_i - D)\right]B_L + E\left[\psi(D)\right]B_S + E(\Gamma_\psi)(B_L - B_S)\right. \qquad \text{(III.6.14)}$$

gdzie: $L_i = \text{entier}\left(\frac{X_i}{B_L}\right)$, $S_i = \text{entier}\left[X_i - L_i B_L\right]$,

przy czym, na podstawie (III.6.12) zachodzi

$$E\left[\psi(\xi_{i+1} - \xi_i)\right] = \lambda(\xi_{i+1} - \xi_i),$$

$$E\left[\psi(D)\right] = \lambda D,$$

$$E\left[\psi(\xi_{i+1} - \xi_i - D)\right] = \lambda(\xi_{i+1} - \xi_i - D).$$

Tak więc, biorąc pod uwagę (III.6.7), łatwo widać, iż dla dowolnego regulaminu obsługi $\mathcal{A}$ ciąg $\left[\gamma_i\right]$ posiada (M+1) punktów skupienia $\sigma_j(\mathcal{A})$

$$\sigma_j(\mathcal{A}) = \lim_{i \to \infty} \gamma_i\Big|_{x_i = jB_S + lB_L}, \quad j=0,1,\ldots,M; \qquad \text{(III.6.15)}$$

$$\sigma_j(\mathcal{A}) = r_{\mathcal{A}}^*(j)Y_j + \left[1 - r_{\mathcal{A}}^*(j)\right]V_j, \qquad \text{(III.6.16)}$$

gdzie:

$$Y_j = -B_S + \lambda(\xi_{i+1} - \xi_i)B_S + E\left[\varkappa_\psi(j)\right]\left[B_L - B_S\right], \qquad \text{(III.6.17)}$$

$$V_j = -B_L + j(B_L - B_S) + \lambda(\xi_{i+1} - \xi_i - D)B_L + \lambda D B_S + E(\Gamma_\psi)(B_L - B_S). \qquad \text{(III.6.18)}$$

Warunek ten, stanowiący uogólnienie udowodnionego przez Pakesa Warunku 1, wykorzystywanego w punkcie III.1 (do którego sprowadza się natychmiast, gdy B_L jest wielokrotnością B_S), wynika z twierdzenia pracy [TWEE 76], którego spełnienie założeń można łatwo sprawdzić.

Rozpatrzmy dwa regulaminy obsługi $\mathcal{R}_1$ oraz $\mathcal{R}_2$.
Niech $\Theta_{\mathcal{R}_1}$ i $\Theta_{\mathcal{R}_2}$ oznaczają suprema zbiorów wartości λ, dla których spełniony jest Warunek 3 przy zastosowaniu odpowiednio regulaminu obsługi $\mathcal{R}_1$ lub $\mathcal{R}_2$. Ponieważ korzystamy wyłącznie z warunku wystarczającego, więc $\Theta_{\mathcal{R}_1} \leq T_{\mathcal{R}_1}$, $\Theta_{\mathcal{R}_2} \leq T_{\mathcal{R}_2}$.
Jeżeli zachodzić będzie nierówność $\Theta_{\mathcal{R}_1} \geq \Theta_{\mathcal{R}_2}$, to będziemy mówili, że regulamin obsługi $\mathcal{R}_1$ jest <u>zalecany</u> w porównaniu z regulaminem obsługi $\mathcal{R}_2$, nie dysponując bowiem możliwością bezpośredniego porównania wydajności granicznych $T_{\mathcal{R}_1}$, $T_{\mathcal{R}_2}$, nie możemy stwierdzić, czy $\mathcal{R}_1$ jest lepszy od $\mathcal{R}_2$.

Na podstawie (III.6.11) oraz (III.6.13) mamy:

$$\gamma_i = r_{\mathcal{R}}(S_i,L_i)\left\{-B_S+E\left[\psi(\zeta_{i+1}-\zeta_i)\right]B_S+E\left[\varkappa_\psi(S_i)\right](B_L-B_S)\right\} +$$

$$+\left[1-r_{\mathcal{R}}(S_i,L_i)\right]\left\{-B_L+S_i(B_L-B_S)+E\left[\psi(\zeta_{i+1}-\zeta_i-D)\right]B_L+E\left[\psi(D)\right]B_S+E(\Gamma_\psi)(B_L-B_S)\right. \tag{III.6.14}$$

gdzie: $L_i = \text{entier}\left(\frac{X_i}{B_L}\right)$, $S_i = \text{entier}\left[X_i - L_i B_L\right]$,

przy czym, na podstawie (III.6.12) zachodzi

$$E\left[\psi(\zeta_{i+1}-\zeta_i)\right] = \lambda(\zeta_{i+1}-\zeta_i),$$

$$E\left[\psi(D)\right] = \lambda D,$$

$$E\left[\psi(\zeta_{i+1}-\zeta_i-D)\right] = \lambda(\zeta_{i+1}-\zeta_i-D).$$

Tak więc, biorąc pod uwagę (III.6.7), łatwo widać, iż dla dowolnego regulaminu obsługi $\mathcal{R}$ ciąg $[\gamma_i]$ posiada (M+1) punktów skupienia $\sigma_j(\mathcal{R})$

$$\sigma_j(\mathcal{R}) = \lim_{l\to\infty}\gamma_i\Big|_{x_i = jB_S+lB_L}, \quad j=0,1,\ldots,M; \tag{III.6.15}$$

$$\sigma_j(\mathcal{R}) = r^*_{\mathcal{R}}(j)Y_j + \left[1-r^*_{\mathcal{R}}(j)\right]V_j, \tag{III.6.16}$$

gdzie:

$$Y_j = -B_S+\lambda(\zeta_{i+1}-\zeta_i)B_S+E\left[\varkappa_\psi(j)\right]\left[B_L-B_S\right], \tag{III.6.17}$$

$$V_j = -B_L+j(B_L-B_S)+\lambda(\zeta_{i+1}-\zeta_i-D)B_L+\lambda DB_S+E(\Gamma_\psi)(B_L-B_S). \tag{III.6.18}$$

W rozpatrywanym przypadku zachodzi więc

$$\limsup_{x_i\to\infty}\gamma_i = \sigma_I(\mathcal{R}) = \max_{j=1,2,\ldots,M}\sigma_j(\mathcal{R}). \tag{III.6.19}$$

Udowodnimy, korzystając z Warunku 3 oraz zależności (III.6.16) i (III.6.19), pewne wskazania dotyczące doboru regulaminu obsługi w rozpatrywanym systemie. Ponieważ zwykle nie potrafimy określić wartości I występującej we wzorze (III.6.19), będziemy dążyli do wykazania, że regulamin obsługi $\mathcal{R}_1$ jest zalecany w porównaniu z regulaminem obsługi $\mathcal{R}_2$, dowodząc, że:

$$\bigvee_{j\in[1,2,\ldots,M]}\left\{\sigma_j(\mathcal{R}_1) \leq \sigma_j(\mathcal{R}_2)\right\}. \tag{III.6.20}$$

<u>Stwierdzenie 1</u>

Niech klasa regulaminów obsługi $\mathcal{K}$ obejmuje regulaminy, które nie dopuszczają oczekiwania stanowiska obsługi na napłynięcie nowego zgłoszenia przy niepustej kolejce.

W porównaniu z dowolnym regulaminem $\mathcal{R}_1 \in \mathcal{K}$; $1 = 1,2,\ldots$, dla którego $\exists_j\, r^*_{\mathcal{R}_1}(j) < 1$, zalecany jest regulamin obsługi $\mathcal{R}^0_1$, charakteryzujący się identycznym jak $\mathcal{R}_1$ sposobem szeregowania zgłoszeń należących do zbioru $\mathcal{K}_i$ i taki, że

$$\bigvee_j r^*_{\mathcal{R}^0_1}(j) = 1, \quad j=1,2,\ldots,M.$$

Dla rozpatrywanej klasy regulaminów obsługi przy $j \geq 1$ zachodzi

$$\zeta_{i+1}-\zeta_i = \begin{cases} B_S & \text{jeśli wykonywana jest obsługa krótka} \\ B_L & \text{jeśli wykonywana jest obsługa długa} \end{cases}$$

a stąd

$$\left.\begin{aligned} Y_j &= -B_S+\lambda B_S^2+E\left[\varkappa_\psi(j)\right]\left[B_L-B_S\right] \\ V_j &= -B_L+j(B_L-B_S)+\lambda(B_L-D)B_L+\lambda DB_S+E(\Gamma_\psi)(B_L-B_S) \end{aligned}\right\} \quad j=1,2,\ldots,M \tag{III.6.21}$$

$$\sigma_0(\mathcal{K}) = V_0 = -B_L+\lambda\left[B_L-D\right]B_L+\lambda DB_S+E(\Gamma_\psi)(B_L-B_S) \tag{III.6.22}$$

Wykażemy, że dla każdego $j > 0$, $\lambda > 0$ zachodzi

$$V_j - Y_j > 0.$$

Korzystając z (III.6.19) mamy:

$$V_j - Y_j = -(B_L - B_S) + j(B_L - B_S) + \lambda(B_L - D)B_L + D\lambda B_S - \lambda B_s^2 +$$

$$+ E(\Gamma_\psi)(B_L - B_S) - E[\sigma_\psi(j)](B_L - B_S) =$$

$$= (j - 1)(B_L - B_S) + \lambda B_L(B_L - D) + \lambda B_S(D - B_s) +$$

$$+ E(\Gamma_\psi)(B_L - B_S) - E[\sigma_\psi(j)](B_L - B_S). \qquad \text{(III.6.23)}$$

Zauważmy jednak, że:

$$\left\{[\sigma_\psi(j)] < \psi(B_S)\right\} \Rightarrow \left\{E[\sigma_\psi(j)] < \lambda B_S\right\},$$

oraz

$$\lambda B_L(B_L - B_S - D) + \lambda B_s(D - B_s) > 0,$$

tak więc:

$$V_j - Y_j > (j-1)(B_L-B_S) + \lambda B_L(B_L - B_S - D) + \lambda B_S(D-B_s) +$$

$$+ E(\Gamma_\psi)(B_L-B_S) + \lambda B_s^2 > 0. \qquad \text{(III.6.24)}$$

Widzimy więc, że z (III.6.23) wynika wprost dla każdego λ nierówność

$$\bigvee_{l>0}\left\{\sigma_j(\varkappa_l) > \sigma_j(\varkappa_l^o)\right\}, \quad j=1,2,\ldots,M; \qquad \text{(III.6.25)}$$

co łącznie z zależnością

$$\bigvee_{l>0}\left\{\sigma_o(\varkappa_l) = \sigma_o(\varkappa_l^o)\right\}. \qquad \text{(III.6.26)}$$

kończy dowód stwierdzenia 1.

Zauważmy, że na mocy (III.6.21) oraz (III.6.22)

$$\sigma_o < V_p < V_{p+1}, \quad p=1,2,\ldots,M-1; \qquad \text{(III.6.27)}$$

a więc warunek stabilności związany jest istotnie z wartościami σ_i, $i \geq 1$, i należy unikać rozpoczynania obsługi "długiej" w sytuacji, gdy w kolejce znajdują się zgłoszenia oczekujące na obsługę "krótką".

W dotychczasowych rozważaniach nie braliśmy pod uwagę sposobu wzajemnego szeregowania zgłoszeń wymagających obsługi "krótkiej". Wskazania odnośnie do tej części regulaminu obsługi zawiera stwierdzenie 2.

Stwierdzenie 2

W porównaniu z dowolnym regulaminem obsługi $\varkappa_1^o$, zdefiniowanym wyżej, zalecany jest regulamin $\varkappa_N^o$, narzucający w każdej chwili ζ_i kolejność obsługi zgłoszeń należących do zbioru $\mathcal{K}_i$ zgodną z kolejnością ich napływania do systemu.

Stwierdzenie to wynika bezpośrednio ze znanych reguł szeregowania zadań przed liniami krytycznymi [COFF 80], [BLAZ 77], ponieważ regulamin naturalny minimalizuje liczbę zgłoszeń przekraczających linie krytyczne. Podamy tu prosty dowód tego stwierdzenia.

Niech w przedziale czasu $[t_o-\Omega, t_o]$ napłynie do systemu obsługi "m" zgłoszeń $\zeta_1,\zeta_2,\ldots,\zeta_m$ (odpowiednio w chwilach czasu $\tau_1,\tau_2,\ldots,\tau_m$), których obsługa może się rozpocząć dopiero w chwili t_o.
Załóżmy też, że obsługa każdego ze zgłoszeń trwa przez czas B_s=const, z tym że zgłoszenie, którego czas oczekiwania przekracza wartość $D > \Omega$, jest tracone.

Sprawdzimy, że obsługa zgłoszeń w kolejności ich nadchodzenia minimalizuje liczbę zgłoszeń traconych.

Istotnie, przy zastosowaniu takiego sposobu obsługi chwile rozpoczęcia obsługi kolejnych zgłoszeń t_i, i=1,2,...,m można wyznaczyć z relacji:

$$t_1 = t_o,$$

$$t_i = \begin{cases} t_{i-1} + B_s & \text{jeśli } \tau_{i-1} + D \geq t_{i-1} \quad i=1,2,\ldots,m, \\ t_{i-1} & \text{w przeciwnym razie,} \end{cases} \qquad \text{(III.6.28)}$$

(zgłoszeniom traconym przypisujemy zerowy czas obsługi).

Niech pierwszym zgłoszeniem traconym będzie zgłoszenie ζ_j. Łatwo widać, że żadne inne uszeregowanie zgłoszeń nie może zapobiec stracie co najmniej jednego zgłoszenia ze zbioru $\{\zeta_1,\zeta_2,\ldots,\zeta_j\}$. Gdyby bowiem zamienić momenty rozpoczynania obsługi zgłoszenia ζ_j oraz ζ_i, i<j, to oczywiście, ponieważ $\tau_i < \tau_j$, więc:

$$(\tau_j + D > t_j) \Rightarrow (\tau_i + D > t_j)$$

i wówczas stracone zostałoby zgłoszenie ζ_i.
Uszeregowanie zgłoszeń zgodnie z kolejnością ich napływania rzeczywiście minimalizuje więc liczbę zgłoszeń traconych.

Zauważmy, że jeśli M = entier $(\frac{D}{B_s})$, wówczas liczbę zgłoszeń traconych L można oszacować jako:

$$\begin{aligned} m - \mathrm{Min}[m,M] \leq L \leq m-1 \qquad & \text{gdy} \quad \Omega > B_s \\ L = m - \mathrm{Min}[m,M], \quad & \text{gdy} \quad \Omega = B_s \end{aligned} \tag{III.6.29}$$

co zostanie wykorzystane w dalszych rozważaniach.

Dla dowolnego regulaminu $\mathcal{R}_1^o$ ciąg $\{\gamma_i\}$ posiada punkty skupienia

$$\left.\begin{aligned} \sigma_0(\mathcal{R}_1^o) &= V_o \\ \sigma_j(\mathcal{R}_1^o) &= Y_j \qquad j=1,2,\ldots,M \end{aligned}\right\} \tag{III.6.30}$$

gdzie V_o oraz Y_j określone są wzorami (III.6.21) oraz (III.6.22).

Zgodnie z wykazaną wyżej regułą dla dowolnej realizacji strumienia zgłoszeń

$$\left.\begin{aligned} &\bigvee_i \left\{ \alpha_{\psi(j)}\Big|_{\mathcal{R}_i^o} \geq \alpha_{\psi(j)}\Big|_{\mathcal{R}_M^o} \right\} \\ &\bigvee_i \left\{ \Gamma_\psi\Big|_{\mathcal{R}_i^o} \geq \Gamma_\psi\Big|_{\mathcal{R}_M^o} \right\} \end{aligned}\right. \tag{III.6.31}$$

takie same nierówności zachodzą więc dla wartości oczekiwanych, czyli:

$$\sigma_p(\mathcal{R}_i^o) \geq \sigma_p(\mathcal{R}_M^o), \qquad p=0,1,\ldots,M; \tag{III.6.32}$$

i istotnie regulamin obsługi $\mathcal{R}_M^o$ jest zalecany w porównaniu z każdym innym regulaminem $\mathcal{R}_i^o$ należącym do rozpatrywanej klasy.

W ten sposób określiliśmy wskazania co do doboru regulaminu obsługi w przypadku, gdy nie jest dopuszczalne oczekiwanie stanowiska obsługi na napłynięcie nowego zgłoszenia przy $X_i > 0$.

Łatwo sprawdzić, iż w przypadku, gdy D = B_S, zmiany stanu systemu w przypadku zalecanego regulaminu $\mathcal{R}_M^o$ oraz regulaminu LIFO są identyczne, a więc nasze wskazania potwierdzają sugestię [BUZA 74] co do doboru regulaminu obsługi.

Przejdziemy teraz do rozpatrzenia celowości oczekiwania stanowiska obsługi w przypadku $X_i > 0$.

Wprowadzając założenia o klasie rozpatrywanych regulaminów obsługi wykluczyliśmy, dla prostoty rozważań, możliwość oczekiwania stanowiska obsługi na napłynięcie nowego zgłoszenia, w przypadku gdy $S_i > 0$. Przypadek ten można by również uwzględnić, modyfikując nieco interpretację zmiennej losowej $\alpha_\psi(j)$, a następnie przy użyciu rozumowania analogicznego jak w dowodzie stwierdzenia 2 wykazać, iż nie można na tej drodze uzyskać regulaminu zalecanego w porównaniu z $\mathcal{R}_M^o$. Dowodu tej, również intuicyjnie oczywistej własności, nie będziemy tu zamieszczać.

W dalszym ciągu rozpatrzymy natomiast celowość oczekiwania na napłynięcie nowego zgłoszenia w sytuacji, gdy $S_i = 0$, $L_i > 0$, ograniczając się do rozważenia przypadków, w których M = entier $(\frac{D}{B_S})$, to znaczy, gdy D jest całkowitą wielokrotnością B_S.

Stwierdzenie 3

Oznaczmy przez $\mathcal{E}_M$ regulamin obsługi o następującej charakterystyce:
- w przypadku gdy $S_i \geq 0$, do obsługi wybierane są zgłoszenia należące do zbioru $\mathcal{S}_i$ i to zgodnie z regulaminem naturalnym,
- zakłada się dostępność prognozy momentu nadejścia kolejnego zgłoszenia w przedziale czasu τ_i, $\tau_i + \vartheta$ (prognoza negatywna występuje z prawdopodobieństwem $\beta = e^{-\lambda\vartheta}$),
- jeśli przy $S_i = 0$, $L_i > 0$, prognoza jest negatywna, wówczas rozpoczyna się obsługę długą, w przeciwnym razie oczekuje się na nadejście nowego zgłoszenia.

Wykażemy, że w przypadku gdy M = entier $(\frac{D}{B_s})$, regulamin obsługi $\mathcal{E}_M$ jest zalecany w porównaniu z regulaminem obsługi $\mathcal{R}_M^o$.

Rozpatrzmy punkty skupienia ciągów $\{\gamma_i\}$ dla obu rozpatrywanych regulaminów obsługi zgłoszeń.

Oczywiście

$$\sigma_j(\mathcal{R}_M^o) = \sigma_j(\mathcal{E}_M) = -B_s + \lambda B_s^2 + E[\alpha_\psi(j)](B_L - B_s) \quad \text{dla} \quad j=1,2,\ldots,M \tag{III.6.33}$$

Istotnie różne są natomiast wartości σ_o jako że:

$$\sigma_o(\mathcal{R}_M^o) = -B_L + \lambda(B_L - D)B_L + \lambda DB_s + E(\Gamma_\psi)(B_L - B_S), \tag{III.6.34}$$

natomiast

$$\begin{aligned} \sigma_o(\mathcal{E}_M) = (1-e^{-\lambda\vartheta})\left\{\lambda B_s^2 + E[\alpha_\psi(0)](B_L - B_S)\right\} + \\ + e^{-\lambda\vartheta}\left[-B_L + \lambda(B_L - \vartheta - D)B_L + \lambda DB_s + E(\Gamma_\psi)(B_L - B_S)\right] \end{aligned} \tag{III.6.35}$$

$$\text{gdy} \quad B_L - \vartheta - D > 0$$

oraz

$$\sigma_0(\mathcal{E}_N) = (1-e^{-\lambda\vartheta})\left\{\lambda B_s^2 + E[\alpha_\psi(0)](B_L-B_S)\right\} + \\ + e^{-\lambda\vartheta}\left\{-B_L + \lambda(B_L-\vartheta)B_S + E(\Gamma_\psi^*)(B_L-B_s)\right\} \qquad \text{(III.6.36)}$$

$$\text{dla} \quad B_L \geq \vartheta \geq B_L - D,$$

gdzie: Γ_ψ^* jest zmienną losową oznaczającą liczbę zgłoszeń, które mimo nadejścia w odcinku czasu $[\xi_i - (B_L - \vartheta), \xi_i]$ wymagać będą obsługi "długiej". Łatwo sprawdzić, że $\sigma_0(\mathcal{E}_N)$ jest funkcją ciągłą dla $\vartheta \in [0, B_L]$.
Rozpatrzmy najpierw przypadek $\vartheta < B_L - D$.

Przekształcimy (III.6.34) do postaci analogicznej z (III.6.35).

$$\sigma_0(\mathcal{H}_N^0) = e^{-\lambda\vartheta}\left\{-B_L + \lambda(B_L-\vartheta-D)B_L + \lambda DB_s + E(\Gamma_\psi)\right\} + \\ + (1-e^{-\lambda\vartheta})\left\{\lambda(B_L-D-\vartheta)B_L + (R_1-1)B_L + \lambda DB_S + E(\Gamma_\psi)\right\}, \qquad \text{(III.6.37)}$$

gdzie:

$$R_1 = \frac{\lambda\vartheta}{1-e^{-\lambda\vartheta}} > 1$$

jest wartością oczekiwaną liczby zgłoszeń, które nadejdą w czasie ϑ, pod warunkiem nadejścia przynajmniej jednego zgłoszenia i zbadamy znak różnicy:

$$\sigma_0(\mathcal{H}_N^0) - \sigma_0(\mathcal{E}_N) = (1-e^{-\lambda\vartheta})\left\{\lambda(B_L-D-\vartheta)B_L + (R_1-1)B_L + \lambda DB_S - \lambda DB_s^2 + \right. \\ \left. + \left[E(\Gamma_\psi) - E(\alpha_\psi(0))\right](B_L-B_S)\right\}. \qquad \text{(III.6.38)}$$

Z uwag zamieszczonych przy dowodzie stwierdzenia 2 wynika, że:

$$E(\Gamma_\psi) \geq E\left\{\max[0,\psi(D) - M]\right\} = \sum_{j=1}^{\infty} jp_{M+j}^{D},$$

$$E[\alpha_\psi(0)] = E\left\{\max[0,\psi(B_S)-M]\right\} = \sum_{j=1}^{\infty} jp_{M+j}^{B_S} \qquad \text{(III.6.39)}$$

Wykażemy, że dla dowolnych $x > 0$, $y \geq x$, $l=1,2,\ldots$, zachodzi nierówność:

$$\sum_{j=1}^{\infty} jp_{l+j}^{y} - \sum_{j=1}^{\infty} jp_{l+j}^{x} > 0, \qquad \text{(III.6.40)}$$

gdzie p_i^{Δ} określone są wzorem (III.6.12).
Istotnie, po prostych przekształceniach uzyskujemy:

$$\sum_{j=1}^{\infty} jp_{l+j}^{y} = \sum_{k=0}^{\infty}\sum_{i=1}^{\infty}(i+k)p_{l+i}^{x}p_k^{y-x} + \sum_{k=0}^{l}\sum_{m=1}^{\infty} mp_k^x p_{k-l+m}^{y-x} = \\ = \sum_{i=1}^{\infty} ip_{l+i}^{x} + \sum_{m=0}^{\infty} mp_m^{y-x}\sum_{i=1}^{\infty} p_{l+i}^{x} + \sum_{k=0}^{l}\sum_{m=1}^{\infty} mp_k^x p_{k-l+m}^{y-x},$$

a więc:

$$\sum_{j=1}^{\infty} jp_{l+j}^{y} - \sum_{j=1}^{\infty} jp_{l+j}^{x} = \lambda(y-x)\sum_{i=1}^{\infty} p_{l+i}^{x} + \sum_{k=0}^{l}\sum_{m=1}^{\infty} mp_k^x p_{k-l+m}^{y-x} > 0. \qquad \text{(III.6.41)}$$

Na mocy (III.6.38) oraz (III.6.41) zachodzi:

$$E(\Gamma_\psi) > E[\alpha_\psi(0)], \qquad \text{(III.6.42)}$$

toteż

$$\sigma_0(\mathcal{H}_N^0) - \sigma_0(\mathcal{E}_N) > (1-e^{-\lambda\vartheta})\left\{\lambda B_L(B_L-D-\vartheta) + (R_1-1)B_L + \lambda B_S(D-B_S)\right\}. \qquad \text{(III.6.43)}$$

Ponieważ $R_1 > 1$ i $D > B_S$, więc różnica $\sigma_0(\mathcal{H}_N^0) - \sigma_0(\mathcal{E}_N)$ jest zawsze dodatnia i regulamin obsługi $\mathcal{E}_N$ jest zalecany w porównaniu z $\mathcal{H}_N^0$ na zasadzie (III.6.20).

Na licznych przykładach numerycznych przy różnych wartościach B_S, B_L, M stwierdzono, że istotnie $\sigma_0(\mathcal{H}_N^0) > \sigma_j(\mathcal{H}_N^0)$, $j=1,2,\ldots,M$.

Rozpatrzmy jeszcze wpływ wartości $\vartheta \in [0, B_L - D]$ na wartość różnicy (III.6.38).
W tym celu obliczymy pochodną

$$\frac{d}{d\vartheta}\left\{\sigma_0(\mathcal{H}_N^0) - \sigma_0(\mathcal{E}_N)\right\} = \lambda e^{-\lambda\vartheta}\left[\lambda B_L(B_L-D) + \lambda B_S(D-B_S) - B_L\right] - \\ - \lambda^2 B_L \vartheta e^{-\lambda\vartheta} + \lambda B_L e^{-\lambda\vartheta} \qquad \text{(III.6.44)}$$

Tak więc dla każdego $0 < \vartheta < B_L - D$,

$$\frac{d}{d\vartheta}\left\{\sigma_o(\mathcal{R}_N^o) - \sigma_o(\mathcal{E}_N)\right\} = \lambda e^{-\lambda\vartheta}\left[\lambda B_L(B_L-D-\vartheta) + \lambda B_S(D-B_S)\right] > 0. \qquad \text{(III.6.45)}$$

W przedziale zmienności $\vartheta \in \left[0, B_L-B\right]$ wyznaczona pochodna jest dodatnia i najbardziej wskazaną wartością horyzontu prognozy jest $\vartheta_o = B_L-D$.

Rozpatrzmy obecnie celowość stosowania wartości $\vartheta > B_L-D$. Wykazanie w całym tym przedziale $\sigma_o(\mathcal{E}_N) > \sigma_o(\mathcal{R}_N^o)$ jest trudne. W tym miejscu sprawdzimy tylko, że stosowanie regulaminu $\mathcal{E}_N$ z wartością $\vartheta_1 = B_L-B_S > \vartheta_o$ jest zalecane w porównaniu z regulaminem $\mathcal{R}_N^o$. Istotnie

$$\sigma_o(\mathcal{E}_N)\Big|_{\vartheta_1} = \lambda B_S^2 - e^{-\lambda\vartheta_1}B_L + E\left[\alpha_\psi(0)\right](B_L-B_S), \qquad \text{(III.6.46)}$$

a więc

$$\sigma_o(\mathcal{R}_N^o) - \sigma_o(\mathcal{E}_N)\Big|_{\vartheta_1} = \lambda(B_L-D)B_L + \lambda B_S(D-B_S) - B_L(1-e^{-\lambda\vartheta_1}) +$$

$$+\left\{E(\Gamma_\psi^*) - E\left[\alpha_\psi(0)\right]\right\}(B_L-B_S) > \varphi_1(\lambda), \qquad \text{(III.6.47)}$$

gdzie:

$$\varphi_1(\lambda) = \lambda(B_L-D)B_L - B_L\left[1 - e^{-\lambda(B_L-B_S)}\right].$$

Ponieważ $\varphi_1(0) = 0$ oraz

$$\frac{d}{d\lambda}\varphi_1(\lambda) = (B_L-D)B_L - (B_L-B_S)B_L\, e^{-\lambda(B_L-B_S)} > 0, \qquad \text{(III.6.48)}$$

a więc funkcja $\varphi_1(\lambda)$ jest stale dodatnia dla $\lambda > 0$, czyli

$$\sigma_o(\mathcal{R}_N^*) - \sigma_o(\mathcal{E}_N)\Big|_{\vartheta_1} > 0. \qquad \text{(III.6.49)}$$

Nie można przy tym ogólnie wskazać, która z wartości ϑ_o czy ϑ_1 jest lepsza (zależy to od parametrów systemu).

W szczególnym przypadku, gdy $D = B_S$, uzyskujemy:

$$\frac{d}{d\vartheta}\left\{\sigma_o(\mathcal{R}_N^o) - \sigma_o(\mathcal{E}_N)\Big|_{\vartheta_o=\vartheta_1}\right\} = 0, \qquad \text{(III.6.50)}$$

a ponieważ dodatkowo

$$\frac{d^2}{d\vartheta^2}\left\{\sigma_o(\mathcal{R}_N^o) - \sigma_o(\mathcal{E}_N)\Big|_{\vartheta_o=\vartheta_1}\right\} =$$

$$= -\lambda^2 e^{-\lambda\vartheta}\left[\lambda B_L(B_L-D) + \lambda B_S(D-B_S)\right] + \lambda^3 B_L\vartheta e^{-\lambda\vartheta} - \lambda^2 B_L e^{-\lambda\vartheta} < 0, \qquad \text{(III.6.51)}$$

więc wartość $T = T_o = T_1$ maksymalizuje różnicę (III.6.49) (można łatwo sprawdzić, że jest to jedyne maksimum tej funkcji).

Ten wynik, oparty wyłącznie na wykorzystaniu stwierdzenia 3, jest zgodny z rezultatami przedstawionymi dla takiego, szczególnego przypadku w [BUZA 74], gdzie właśnie taką wartość ϑ wyznaczono jako optymalną.

Wydaje się więc, iż korzystanie ze stwierdzenia 3 jako kryterium przy ustalaniu wskazań dla doboru regulaminu obsługi jest celowe i niezależnie od wskazań wypracowanych dla rozpatrywanego w tym punkcie prostego systemu obsługi zasługuje na uwagę również w innych przypadkach. Warto podkreślić, iż dla bardziej złożonych systemów można dążyć do wyznaczenia wartości ϑ_i częściowo lub całkowicie przy użyciu modelowania cyfrowego.

c. <u>Propozycje doboru regulaminu obsługi dla dowolnych systemów obsługi z czasem obsługi zależnym od czasu oczekiwania</u>

Powyżej, posługując się przykładem prostego systemu WAD, pokazano, iż posługiwanie się warunkiem wystarczającym stabilności może ułatwić podejmowanie decyzji co do wyboru regulaminu obsługi spośród grupy rozpatrywanych.

Wcześniej już podkreślono, że metody przedstawione w punktach III.2 oraz III.5 zapewniają możliwość określenia wydajności granicznej dla szerokiej klasy systemów WAD, a w konsekwencji również możliwość wyboru - drogą bezpośredniego porównywania wydajności granicznych różnych wariantów - regulaminu obsługi maksymalizującego pewną funkcję celu (na przykład wydajność graniczną). Sprawą otwartą pozostaje oczywiście określenie zbioru regulaminów obsługi, w którym porównania takie będą prowadzone.

Poniżej przedstawimy pewne propozycje regulaminów obsługi, które wydają się godne wzięcia pod uwagę przy poszukiwaniu rozwiązań maksymalizujących wydajność graniczną względnie minimalizujących stopień wykorzystania stanowiska obsługi. Uzasadnienie tych propozycji - mające charakter intuicyjny - nie pretenduje do roli dowodu optymalności i może ono być traktowane jako heureza podlegająca w każdym konkretnym przypadku ocenie oraz ewentualnie porównaniu z innymi, alternatywnymi regulaminami.

Tak jak w poprzednich punktach, ograniczymy się do rozpatrywania systemów (być może wielokanałowych), których stabilność pociąga za sobą skończoność wartości oczekiwanej czasu trwania okresu czynnego. Ograniczymy się też czasowo do przypadku, gdy nie dopuszcza się bezczynności żadnego kanału obsługi przy niepustej kolejce.

W systemach konwencjonalnych czas trwania każdego okresu czynnego jest niezależny od stosowanego regulaminu obsługi i wynika wyłącznie z konkretnej realizacji zmiennych losowych opisujących rozpatrywany system.

W systemach WAD wybór regulaminu obsługi wpływa oczywiście na czas trwania okresu czynnego. Z definicji regulamin obsługi, który zapewniałby minimalizację czasu trwania każdego okresu czynnego, cechowałby się największą wydajnością graniczną oraz najniższym stopniem wykorzystania systemu przy dowolnej intensywności strumienia zgłoszeń.

Tak więc, teoretycznie rzecz biorąc, dysponując możliwością przewidzenia momentów nadejścia poszczególnych zgłoszeń z dowolnie długim wyprzedzeniem oraz możliwością dokładnego określenia a priori czasu obsługi każdego indywidualnego zgłoszenia w funkcji jego czasu oczekiwania, można byłoby pokusić się o porównanie wszystkich możliwych uporządkowań zgłoszeń (przy założeniu, iż obsługa zgłoszenia może rozpocząć się dopiero po jego napłynięciu i stosując - być może - metody mniej pracochłonne aniżeli przegląd zupełny) i znalezienie wśród nich rozwiązania minimalizującego czas trwania kolejnego okresu czynnego.

Praktyczna implementacja takiego regulaminu obsługi jest zwykle niemożliwa. Rozpatrzmy przypadek prostszy, zakładając dostępność prognozy momentów napływania zgłoszeń w pewnym ograniczonym horyzoncie czasowym ϑ.

Niech w chwili ξ_1, w której następuje zwolnienie się kanału obsługi, w kolejce oczekuje n zgłoszeń $\zeta_1,\zeta_2,\ldots,\zeta_m$, które napłynęły odpowiednio w chwilach czasu $\tau_0,\tau_1,\ldots,\tau_m$.

Założymy też, że w odcinku czasu $[\xi_1,\xi_1+\vartheta]$ a ściślej w chwilach czasu $\tau_{m+1},\ldots,\tau_{m+p}(\tau_{m+1}\geq\xi_1,\tau_{m+p}\leq\xi_1+\vartheta)$ napłynie dalszych p zgłoszeń $\zeta_{m+1},\zeta_{m+2},\ldots,\zeta_{m+p}$, a także, że znane są chwile zwalniania się kolejnych kanałów obsługi $\xi_2,\xi_3,\ldots,\xi_n$.

Tak więc, znając czasy obsługi zgłoszeń w funkcji ich oczekiwania można wyznaczyć chwile zakończenia dalszych obsług.

Rozpatrzmy

Regulamin $\mathcal{R}_1$

W każdej chwili ξ_1, w której następuje zwolnienie stanowiska obsługi, należy określić uporządkowanie zgłoszeń $\zeta_1,\ldots,\zeta_{m+p}$ minimalizujące łączny czas obsługi tych zgłoszeń i rozpocząć obsługę zgłoszenia ζ_i, $i\in(1,2,\ldots,m)$ występującego jako pierwsze w tym uporządkowaniu.

Efektywna implementacja takiego regulaminu obsługi jest - ogólnie rzecz biorąc - uciążliwa obliczeniowo. Stąd też wydaje się, że praktyczne zastosowanie regulaminu $\mathcal{R}_1$ może mieć miejsce głównie w przypadkach szczególnych, gdy zastosowane mogą być w tym celu pewne specjalne metody obliczeniowe.

Dobrym przykładem może tu być system wielokanałowy o takich parametrach, że nigdy dwa zgłoszenia objęte szeregowaniem w chwili τ_1 nie mogą być przewidziane do obsługi na tym samym stanowisku obsługi, a więc w szczególności np. system z ograniczoną, niewielką w stosunku do liczby kanałów

obsługi długością kolejki oraz pewnymi ograniczeniami na rozkład czasu obsługi. Załóżmy chwilowo, że horyzont prognozy napływu zgłoszeń $\vartheta = 0$.

Oznaczmy przez q_{ij} czas obsługi zgłoszenia ζ_i, pod warunkiem, iż jego obsługa rozpocznie się w chwili ξ_j $(i,j=1,2,\ldots,m)$.

Zgodnie z wprowadzonym wyżej ograniczeniem zachodzi przy tym

$$\min_{i,j} q_{ij} > \zeta_m - \zeta_1.$$

Niech $\mathcal{K}$ będzie wektorem

$$\mathcal{K} = \begin{bmatrix} k_1 \\ \vdots \\ k_1 \\ \vdots \\ k_m \end{bmatrix}$$

takim, że:

$$\left.\begin{array}{l} k_1 \in [1,2,\ldots,m] \\ i,j,l = 1,2,\ldots,m \end{array}\right\} \quad \text{oraz} \quad \bigvee_i \bigvee_j \left\{(i \neq j) \Rightarrow (k_i \neq k_j)\right\}.$$

Wartość $k_1 = i$ oznacza, że obsługa zgłoszenia ζ_i ma być rozpoczęta w chwili ξ_1.

Należy znaleźć taką realizację wektora $\mathcal{K}$ która minimalizuje funkcję F_1

$$F_1 = \sum_{j=1}^{m} q_{k_j,j}.$$

Można łatwo sprawdzić, że rozpatrywane zagadnienie daje się sprowadzić do klasycznego w teorii badań operacyjnych wariantu programowania całkowitoliczbowego, zwanego "zadaniem komiwojażera" polegającego na minimalizacji funkcji

$$\Phi_1 = \sum_{i=1}^{m}\sum_{j=1}^{m} q_{ij}x_{ij},$$

gdzie:

$$x_{ij} = \begin{cases} 1 & \text{oznacza } k_j = i, \\ 0 & \text{oznacza } k_j \neq i, \end{cases}$$

przy ograniczeniach:

$$\sum_{i=1}^{m} x_{ij} = 1,$$

$$\sum_{j=1}^{m} x_{ij} = 1,$$

$$i,j=1,2,\ldots,m.$$

Algorytm efektywnego rozwiązywania tego zadania metodą podziału i ograniczeń podał Little [LITT 63][x].

Zadanie komiwojażera jest zagadnieniem o dużej złożoności obliczeniowej i należy do rodziny problemów tzw. silnie NP-zupełnych [COFF 80]. Stąd w praktyce można ograniczyć się do pewnych, uzyskanych przy użyciu metody podziału i ograniczeń wyników pośrednich jako dość dobrze przybliżających rozwiązanie optymalne (i uzyskiwanych w dużo krótszym czasie).

Propozycja zastosowania takiego regulaminu obsługi, z wykorzystaniem rozwiązywania zadania komiwojażera jako metody jego implementacji, do pewnego problemu technicznego (szeregowanie wlewków ładowanych do pieców wgłębnych w walcowni półwyrobów kombinatu metalurgicznego) została przedstawiona w opracowaniu [ZSAK 79].

Wyniki przedstawionych tam testów przeprowadzonych na danych rzeczywistych wykazały przewagę takiego regulaminu obsługi nad stosowanym aktualnie, intuicyjnym sposobem szeregowania.

Przedstawiona powyżej metoda szeregowania zadań może być rozszerzona również na przypadek $\vartheta > 0$. Mimo że z założenia o niemożliwości oczekiwania na nadejście nowego zgłoszenia przy niepustej kolejce wynika, iż żadne ze zgłoszeń $\zeta_{m+1},\ldots,\zeta_{m+p}$ nie może być wytypowane do obsługi w chwili ζ_1, to jednak uwzględnienie ich napływu może wpływać na wartość k_1.

Dla znalezienia właściwej wartości k_1 niezbędna jest dodatkowo znajomość czasów zwalniania się dalszych p kanałów obsługi $\zeta_{m+1},\ldots,\zeta_{m+p}$, po czym możliwe jest sformułowanie zadania komiwojażera o wymiarze (m+p) przy zdefiniowaniu macierzy $[q_{ij}]$ jak następuje:

i) Jeśli zachodzi $\zeta_j \geq \tau_i$, wówczas q_{ij} wyznacza się w sposób omówiony wyżej.

ii) Jeśli zachodzi $\zeta_j < \tau_i$, wówczas należy podstawić $q_{ij} = \mathcal{M} + \tau_i - \zeta_j$, gdzie $\mathcal{M}$ jest liczbą przewyższającą wszystkie wartości q_{ij} wyznaczone w kroku i.

---

x) W dalszych pracach np. [LOKI 78] podano metody wprowadzania pewnych dodatkowych ograniczeń, np. wymagań, aby - mówiąc w kategoriach problemu szeregowania zadań - niektóre z nich były obsługiwane bezpośrednio po sobie.

W przypadku natomiast, gdyby dopuścić możliwość oczekiwania stanowiska obsługi na napłynięcie nowego zgłoszenia, rozwiązywać należałoby również zadanie komiwojażera o wymiarze (m+p) przy zdefiniowaniu elementów macierzy $[q_{ij}]$ w sposób następujący (przy założeniu, iż możliwe jest tylko oczekiwanie na zgłoszenia, które nadejdą w rozpatrywanym horyzoncie).

i) Jeśli zachodzi $\zeta_j \geq \tau_i$, wówczas q_{ij} wyznacza się jak poprzednio.

ii) Jeśli zachodzi $\zeta_1 < \tau_i$, wówczas $q_{ij} = \tau_i - \zeta_j + b^i_o$, gdzie b^i_o jest czasem obsługi zgłoszenia ζ^i, gdy czas jego oczekiwania na rozpoczęcie obsługi jest równy zeru.

W przypadku gdy czas obsługi jest niemalejącą funkcją czasu oczekiwania, godny polecenia jest prostszy regulamin R_2 o postaci:

Regulamin R_2

W każdej chwili ζ_1, w której następuje zwolnienie stanowiska obsługi, należy wyznaczyć, dla każdego spośród zgłoszeń $\zeta_1,\ldots,\zeta_m$ czas jego obsługi $b^i(\zeta_1)$ jakiego wymagałoby to zgłoszenie w przypadku rozpoczęcia jego obsługi w chwili ζ_1. Dla zgłoszeń $\zeta_{m+1},\ldots,\zeta_{m+p}$ wyznaczyć należy $b^i(\zeta_1) = \tau^i - \zeta_1 + b^i_o$, gdzie b^i_o jest czasem obsługi zgłoszenia ζ^i pod warunkiem, że jego czas oczekiwania jest równy zeru.

Do obsługi należy wytypować zgłoszenie, dla którego wartość $b^i(\zeta_1)$ jest najmniejsza.

W dotychczasowych rozważaniach zakładano precyzyjną znajomość momentów nadchodzenia kolejnych zgłoszeń, a także czasów obsługi.

W przypadku, gdy znane są tylko rozkłady tych wielkości, możliwe jest stosowanie omówionych regulaminów obsługi w oparciu o wartości oczekiwania odpowiednich zmiennych losowych.

Zaproponowane wyżej regulaminy obsługi, nie będące regulaminami optymalnymi, mogą, ze względu na łatwość implementacji oraz stosunkowo dużą (jak to stwierdzono na przykładach) efektywność, znaleźć zastosowanie w wielu przypadkach.

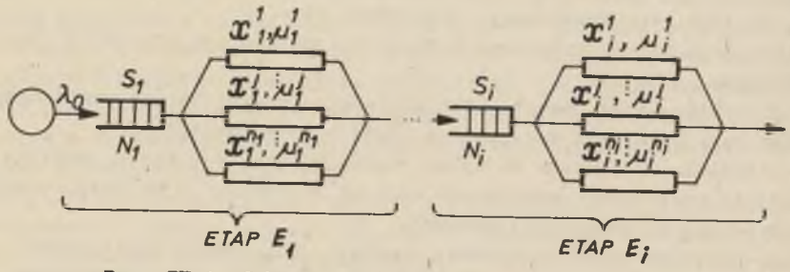
d) Uwagi końcowe

W niniejszym rozdziale podjęto zagadnienie analizy systemów WAD z dowolnymi regulaminami obsługi. Wskazano iż, ogólnie rzecz biorąc nie można wyznaczyć dla takich systemów na drodze analitycznej nie tylko charakterystyk ale nawet warunków stabilności. Stąd szczególnego znaczenia nabiera rozwinięcie metod łączących w sobie pewne podstawy analityczne, które wsparte wynikami odpowiednio zaplanowanych eksperymentów symulacyjnych umożliwiłyby rozpatrywanie przedmiotowych systemów. Metody takie są, jak dotychczas, słabo rozwinięte szczególnie w zakresie szacowania warunków stabilności, ponieważ w konwencjonalnych systemach masowej obsługi problem taki nie występował.

Przedstawione wyżej metody pozwalają na określenie warunków stabilności dla szerokiej klasy systemów WAD, ułatwiają także dobór parametrów tych systemów z punktu widzenia optymalizacji pewnych wskaźników jakości.

IV. MAKSYMALIZACJA WYDAJNOŚCI WIELOSTANOWISKOWYCH WIELOETAPOWYCH SYSTEMÓW MASOWEJ OBSŁUGI

Wieloetapowe systemy obsługi (WSO) zostały zdefiniowane w rozdziale I jako klasa sieci stanowisk obsługi, w których każde zgłoszenie musi zostać obsłużone (i to tylko jeden raz) na każdym ze stanowisk obsługi, w ściśle określonej kolejności. Przed przystąpieniem do omawiania metod analizy i własności WSO przypomnimy i rozszerzymy podaną poprzednio ich klasyfikację, wprowadzimy także pewne, dodatkowe pojęcia związane z tą klasą systemów.



Rys. IV.1. Schemat wieloetapowego systemu obsługi
λ_0 - intensywność strumienia zgłoszeń, n_i - liczba kanałów stanowiska obsługi $\mathcal{X}_i$, zainstalowanego na i-tym etapie, S_i - bufor poprzedzający i-ty etap, N_i - pojemność bufora S_i, μ_i - odwrotność średniego czasu obsługi zgłoszeń w kanale x_i^j, i=1,2,...,M

Założymy, że rozważane WSO składać się będą z M stanowisk obsługi $\mathcal{X}_i$, z których każde zawiera n_i kanałów o stałej mocy i=1,2,...,M (rys. IV.1). Czas trwania obsługi w kanale x_i^j stanowiska obsługi $\mathcal{X}_i$ określa zmienna losowa b_i^j o dystrybuancie $B_i^j(x)$ i wartości oczekiwanej $E(b_i^j) = [\mu_i^j]^{-1}$.

W dalszych rozważaniach zakładać będziemy, że czasy trwania obsługi zgłoszenia na różnych stanowiskach obsługi, a także czasy trwania obsługi różnych zgłoszeń są losowo niezależne i stosowane są regulaminy obsługi zachowujące pracę (nieliczne wyjątki dotyczące systemów z czasem obsługi zależnym od czasu oczekiwania zostaną wyraźnie wyodrębnione).

Według wprowadzonej w rozdziale I klasyfikacji stanowisk obsługi wszystkie WSO są sieciami otwartymi. Ze względu na charakterystykę źródła zgłoszeń i bufora wejściowego S_1 można je podzielić na trzy grupy:

- <u>właściwe</u>, jeśli bufor S_1 ma nieograniczoną pojemność ($N_1 = \infty$), natomiast zgłoszenia napływają strumieniem odnowień o intensywności λ,
- <u>niewłaściwe</u>, jeśli bufor S_1 ma ograniczoną pojemność lub zgłoszenia napływają strumieniem różnym od strumienia odnowień,
- <u>nasycone</u>, jeśli strumień zgłoszeń ma taką charakterystykę, że bufor S_1 poprzedzający pierwszy etap obsługi nie może nigdy zostać opróżniony. Oczywiście dla systemu nasyconego nie jest istotna ani pojemność bufora S_1, ani też dokładna znajomość postaci strumienia zgłoszeń.

Systemy: właściwy, niewłaściwy oraz nasycony <u>odpowiadają sobie wzajemnie</u>, jeśli posiadają identyczne charakterystyki wszystkich stanowisk obsługi a także pojemności <u>buforów pośrednich</u>, tj. buforów S_1, l=2,3,...,M, a różnią się jedynie sposobem napływu zgłoszeń do bufora S_1.

Rozpatrzmy dowolny, nasycony WSO oraz identyczny z nim pod względem wszystkich parametrów system, w którym zgłoszenia przepływają w odwrotnym kierunku, tj. rozpoczynając obsługę od stanowiska $\mathcal{X}_M$, a kończąc na stanowisku $\mathcal{X}_1$.

Takie systemy nazywać będziemy systemami <u>sprzężonymi</u>.

Wprowadzimy też pojęcie <u>pojemności WSO</u> definiowanej jako maksymalna liczba zgłoszeń mogących równocześnie przebywać w systemie, nie wliczając w to zgłoszeń oczekujących w buforze S_1. Tak więc pojemność WSO, oznaczoną L, można określić jako:

$$L = \sum_{i=1}^{M} n_i + \mathcal{N}, \qquad \text{(IV.1)}$$

gdzie: $\mathcal{N} = \sum_{i=2}^{M} N_i$ jest łączną pojemnością buforów WSO.

Oczywiście, pojemności sprzężonych WSO, a także odpowiadających sobie wzajemnie WSO są zawsze jednakowe. W przypadku systemów o ograniczonej pojemności mogą wystąpić przepełnienia buforów, powodując reakcje opisane w rozdziale I.

Będziemy nazywać WSO <u>pojedynczym</u>, gdy każdy jego etap składa się z jednokanałowego stanowiska obsługi ($n_i = 1$, $i = 1,2,...,M$), bądź <u>wielokanałowym</u> w przeciwnym razie.

Dla systemów pojedynczych przyjmiemy oznaczenie $B_i^1(x) = B_i(x)$; $\mu_i^j = \mu_i$.

O wielokanałowym WSO mówimy, że jest on <u>jednorodny</u>, gdy jednorodne są wszystkie stanowiska obsługi, tj. gdy zachodzi $B_i^j(x) = B_i^o(x)$; $j=1,2,...,n_i$; $i=1,2,...,M$.

W niniejszej pracy rozpatrywać będziemy wyłącznie systemy jednorodne.

Jeżeli wydajność każdego stanowiska obsługi jest jednakowa, a więc zachodzi:

$$\bigvee_i \sum_{j=1}^{n_i} \mu_i^j = \text{const}, \qquad \text{(IV.2)}$$

wówczas WSO nazywamy <u>zrównoważonym</u>, w przeciwnym razie mówimy o systemie <u>niezrównoważonym</u>.

Aby umożliwić porównywanie różnych WSO przyjmiemy (nie zmniejszając ogólności rozważań), że:

$$\sum_{i=1}^{M}\left[\sum_{j=1}^{n_i}\mu_i^j\right]^{-1} = M, \qquad (IV.3)$$

a więc dla systemów zrównoważonych zachodzi:

$$\sum_{j=1}^{n_i}\mu_i^j = 1, \qquad (IV.4)$$

przy czym w przypadku zrównoważonych systemów jednorodnych

$$\mu_i^j = \mu_i^o, \quad j=1,2,\ldots,n_i; \qquad (IV.5)$$

$$\mu_i^o \cdot n_i = 1, \quad i=1,2,\ldots,M.$$

Jako miarę niezrównoważenia systemu przyjmować będziemy ([HILL 79]) wskaźnik I

$$I = \frac{1}{M}\sum_{i=1}^{M}\left|\left(\sum_{j=1}^{n_i}\mu_i^j\right)^{-1} - 1\right|. \qquad (IV.6)$$

System wieloetapowy nazywać będziemy systemem <u>synchronizowanym</u>, jeśli momenty rozpoczęcia obsługi na różnych stanowiskach narzucone są pewnym taktem zewnętrznym względnie systemem <u>swobodnym</u>, gdy rytm pracy poszczególnych etapów jest niezależny od czynników zewnętrznych. Analogicznie jak w poprzednich rozdziałach, przez pojęcie <u>wydajności WSO</u> rozumieć będziemy wartość oczekiwaną liczby zgłoszeń opuszczających system w jednostce czasu po całkowitym zakończeniu obsługi na wszystkich stanowiskach.
Szczególnie interesować nas będzie <u>wydajność maksymalna</u> możliwa do osiągnięcia w rozpatrywanym systemie.

W niniejszym rozdziale, stanowiącym rozszerzoną i zmodyfikowaną wersję pracy [WOLI 81A], omówione zostaną zasady optymalizacji wydajności maksymalnej swobodnych, wieloetapowych systemów obsługi, przy czym koncentrować się będziemy na systemach z ograniczonymi buforami pośrednimi i blokowaniem. W szczególności wszędzie tam, gdzie nie będziemy precyzowali sposobu reakcji na przepełnienie założono, że ma miejsce blokowanie.

Dyskusję tych zasad rozpoczniemy od krótkiego omówienia metod wyznaczenia maksymalnej wydajności WSO oraz pewnych, ogólnych prawidłowości zachodzących dla rozpatrywanej klasy systemów masowej obsługi.
Następnie przedstawiona zostanie systematyczna analiza wpływu poszczególnych parametrów WSO na ich wydajność maksymalną. Analiza ta, przeprowadzona w oparciu o bogatą literaturę przedmiotu (w tym prace własne autora), pozwala ustosunkować się do różnych (czasem wręcz sprzecznych) zaleceń, formułowanych często w oparciu o wyrywkowe badania.

IV.1. <u>Metody wyznaczania wydajności maksymalnej</u>

W odniesieniu do WSO z nieograniczonymi buforami wydajność maksymalna, równa wydajności granicznej, może być określona wprost jako:

$$Y = \min_{i=1,2,\ldots,M}\sum_{j=1}^{n_i}\mu_i^j. \qquad (IV.1.1)$$

Jak dotąd brak ogólnych reguł analizy wydajności maksymalnej przypadków, w których część buforów miałaby ograniczoną pojemność, podczas gdy pozostałe byłyby nieograniczone.
W odniesieniu do właściwych WSO z ograniczonymi buforami pośrednimi możemy natomiast skorzystać z omawianego w rozdziale I twierdzenia Lavenberga i poszukiwać wydajności maksymalnej rozpatrywanego systemu, także i w tym przypadku równej jego wydajności granicznej jako wydajności odpowiadającego mu systemu nasyconego.

Odrębnego potraktowania wymagają WSO z regularnymi rozkładami czasów obsługi na wszystkich etapach, do których nie odnosi się bezpośrednio twierdzenie Lavenberga.
Systemy takie były szczegółowo rozpatrywane w pracach [BUZA 72], [FRIE 65], [LABE 76], [MUTH 73], [SUZU 74], [AVII 65B].

Wykazano, że dla dowolnego, jednorodnego WSO, w którym $E(b_i^o) = \alpha_i$, zarówno czasy pobytu zgłoszeń w systemie, jak i postać strumienia wyjściowego zgłoszeń nie zależą od pojemności buforów pośrednich (ani też od faktu czy są one ograniczone, czy też nieograniczone). Obie te charakterystyki nie uległyby też zmianie, gdyby dokonać modyfikacji kolejności, w jakiej zgłoszenia przechodzą przez różne stanowiska obsługi, przy czym podane relacje pozostają w mocy również dla dowolnej postaci strumieni zgłoszeń.
Dla efektywnego wyznaczania wydajności maksymalnej przydatne okazuje się wprowadzone w [FRIE 65] pojęcie dominacji. Będziemy mówili, że stanowisko obsługi x_k dominuje nad stanowiskiem obsługi x_p, jeśli w przypadku gdy x_k poprzedza x_p (niekoniecznie bezpośrednio), zgłoszenia nigdy

nie oczekują na rozpoczęcie obsługi na stanowisku x_p. Udowodniono, że warunkiem koniecznym i wystarczającym dominacji jest, aby $\alpha_p \leq K \alpha_k$, gdzie $K = \text{Entier}(n_p/n_k)$, a także podano pewne własności relacji dominacji.

Tak więc, jeśli w rozpatrywanym WSO istnieje stanowisko obsługi dominujące nad wszystkimi pozostałymi, wówczas kolejka zgłoszeń może tworzyć się wyłącznie przed tym stanowiskiem obsługi. Używając tego pojęcia zaproponowano w [FRIE 65] a następnie rozwinięto w [SUZU 74] metodę redukcji WSO do równoważnego, pod względem wydajności maksymalnej, systemu o mniejszej liczbie etapów - w pewnych przypadkach nawet do systemu dwuetapowego ze stanowiskiem obsługi typu G/D/n na etapie pierwszym oraz stanowiskiem obsługi typu ·/D/∞ na etapie drugim. W takich, specjalnych przypadkach można łatwo określić wydajność maksymalną, równą wydajności granicznej rozpatrywanego systemu, używając znanego kryterium stabilności dla jednostanowiskowych systemów G/D/n.

Analitycznie wyznaczono natomiast [BUZA 67], [BUZA 71A] maksymalną wydajność pojedynczych systemów z regularnymi rozkładami czasów obsługi o postaci:

$$V = \left[\max(\alpha_1, \alpha_2, \ldots, \alpha_M)\right]^{-1}. \qquad \text{(IV.1.2)}$$

Przedyskutujemy obecnie metody wyznaczania wydajności nasyconych WSO. Zauważmy, że analiza M-etapowego systemu nasyconego jest równoważna - pod względem stopnia złożoności - rozpatrywaniu właściwego lub niewłaściwego systemu (M-1) etapowego. Ściśle rozwiązania analityczne są dostępne dla niektórych systemów dwuetapowych oraz bardzo nielicznych przypadków, gdy $M \geq 2$. Ogólne rozważania dotyczące wyznaczania wydajności takich systemów, zawarte w [HILD 67] i [HILD 68], jakkolwiek istotne z punktu widzenia teorii (i stanowiące inspirację dla opracowania pewnych metod przybliżonych), nie prowadzą jednak, niestety, do efektywnych algorytmów.

Dyskutując - w dalszej części tego rozdziału - różne warianty nasyconych WSO, będziemy każdorazowo zwracać uwagę na te przypadki, dla których dostępne są ścisłe rozwiązania analityczne, wskazując równocześnie odpowiednie prace źródłowe (trzy przypadki modeli, prowadzących do oryginalnych wniosków odnośnie do projektowania systemów wieloetapowych, załączono w dodatkach do niniejszego rozdziału).

W stosunku do nasyconych WSO, o prawie dowolnych rozkładach czasów obsługi i ograniczonych buforach pośrednich, naturalne jest przedstawienie ich w postaci łańcucha Markowa o skończonej liczbie stanów i wyznaczanie prawdopodobieństw ergodycznych - tam gdzie analitycznie jest to niemożliwe - przy użyciu metod numerycznych (a więc rozwiązania odpowiedniego układu równań liniowych).

Procedurę numerycznego wyznaczenia wydajności pojedynczych WSO z identycznymi pod względem pojemności buforami pośrednimi $N_i = N$, $i=2,3,\ldots,M$ i rozkładami Erlanga k-tego rzędu omówiono w pracy [HILL 67]. Niestety wymagała ona rozwiązywania układów równań liniowych o wymiarowości rosnącej bardzo szybko ze wzrostem M, k oraz N. Z tego powodu praktyczne możliwości jej użycia przy $k > 1$ są ograniczone do systemów najwyżej 4-etapowych (przy k = 1, N = 0,1 rozpatrywano jeszcze system 6 etapowy) a i to przy niewielkich (rzędu 3-5) wartościach k oraz N.

Analizę liczby stanów w takich systemach zawiera [HAYD 72], a przykładowe dane zaprezentowano w [MUTH 73]. I tak przy k=1, N=0 liczba stanów wyraża się funkcją $[0.5*(3+\sqrt{5})]^M$, co np. przy M = 10 daje 67bb stanów. Uwzględnienie niezerowej pojemności buforów i $k > 1$ prowadzi więc szybko do rozmiaru zadania uniemożliwiającego efektywne rozwiązanie.

W tej samej pracy zaproponowano również numeryczną metodę przybliżoną, nie posiadającą tak ostrych ograniczeń, wskazując, iż we wszystkich rozwiązanych w sposób dokładny przypadkach błąd jest niewielki, maleją cy monotonicznie do zera ze wzrostem N.

W oparciu o ogólniejsze rozważania przedstawione w [MUTH 73] metody numerycznego wyznaczenia wydajności trzyetapowych, pojedynczych systemów nasyconych o-praktycznie rzecz biorąc - dowolnych rozkładach czasów obsługi zaproponowano niezależnie w [RAON 76A] oraz [MUTH 77], jednak również i w tym przypadku obowiązuje bardzo silne ograniczenie na wartość N.

W literaturze (np. [CASE 79]) można też znaleźć informacje o wykorzystaniu do wyznaczania wydajności maksymalnej WSO (a także innych charakterystyk) metody zaproponowanej w [CASE 77], przy czym rozwiązywane przykłady ograniczały się do przypadków systemów pojedynczych, z wykładniczymi rozkładami czasów obsługi.

W zakresie systemów wielokanałowych najnowsze wyniki [LATO 80] ograniczają się do systemu nieograniczonego, M = 2, oraz wykładniczych rozkładów czasów obsługi (wprowadzono też możliwość recyrkulacji zgłoszeń do kolejki wejściowej po zakończeniu obsługi na drugim etapie, jak również uwzględnienia bardziej rozbudowanych metod blokowania, stąd też rozpatrywany tam przypadek jest ogólniejszy aniżeli rozwiązanie podane w Dodatku c).

Tak więc również metody numeryczne mają w odniesieniu do WSO z ograniczonymi buforami pośrednimi ograniczone możliwości zastosowania. Z tego względu podejmowane są starania w kierunku wypracowania rozwiązań przybliżonych, np. [PUJO 79C], [PANW 79], [BOXM 81].

Podstawową metodą precyzyjnego wyznaczania wydajności nasyconych WSO z ograniczoną pojemnością pozostaje jednak modelowanie cyfrowe. W odniesieniu do niektórych klas WSO możliwe jest podanie pewnych, ogólnych własności oraz oszacowań wydajności bez jej wyznaczania[x)].

---

x) Przy dowodzeniu takich własności korzysta się często z faktu, że rozpatrywanie WSO przedstawionych na rys. IV.1 można zastąpić badaniem systemów, w których wszystkie bufory mają pojemność równą zeru ([AVII 65A], [MUTH 73]) jako że dowolny bufor o pojemności N można zastąpić ciągiem N jednokanałowych stanowisk obsługi o zerowych czasach obsługi ($B(x) = H(x)$).

Najważniejszym takim wynikiem jest tzw. zasada odwracalności [YAMA 75], orzekająca, iż sprzężone pojedyncze WSO z dowolnymi rozkładami czasów obsługi mają jednakową wydajność. Tak więc zmiana kierunku przepływu zgłoszeń nie zmienia wydajności takiego systemu. Podana zasada jest o tyle istotna, iż pozwala stwierdzić, że alokacja zasobów (mocy stanowisk obsługi, pojemności buforów), optymalizująca wydajność pojedynczego WSO, musi być "symetryczna".

Aktualnie niezależne dowody zasady odwracalności opublikowano również w [DATT 78] oraz [MUTH 79]. Uogólnienie tej zasady, stwierdzające iż również rozkład czasu pobytu zgłoszenia w systemie oraz liczba zgłoszeń obsłużonych w odcinku czasu (0, t) są identyczne w systemach sprzężonych, podano w [KAWA 75], natomiast w [YAMA 78] wykazano, iż zasada odwracalności pozostaje w mocy również wtedy, gdy niektóre etapy są stanowiskami wielokanałowymi o regularnych rozkładach czasów obsługi.

Dodatkowo w [WOLI 80A] sprawdzono, że zasada ta jest ważna również dla dwuetapowych, wielokanałowych, jednorodnych systemów obsługi o dowolnej (być może różnej na każdym z etapów) liczbie kanałów, charakteryzujących się wykładniczymi czasami obsługi.

Wracając do pojedynczych WSO można podać ogólnie obowiązujące oszacowanie wydajności V_p takich systemów [MUTH 73], a mianowicie:

- Oszacowanie od góry $V_p^+ \geq V_p$ postaci

$$V_p^+ = \min_i \mu_i; \quad i=1,2,\ldots,M, \tag{IV.1.3}$$

które, jak to łatwo sprawdzić, równe jest maksymalnej wydajności etapu o najdłuższym średnim czasie obsługi, a więc wydajności maksymalnej systemu o identycznych z rozpatrywanym stanowiskach obsługi i nieograniczonej pojemności buforów pośrednich.

- Oszacowanie od dołu $V_p^- \leq V_p$ postaci:

$$V_p^- = \left\{ E\left[\max(b_1, b_2, \ldots, b_M)\right] \right\}^{-1}. \tag{IV.1.4}$$

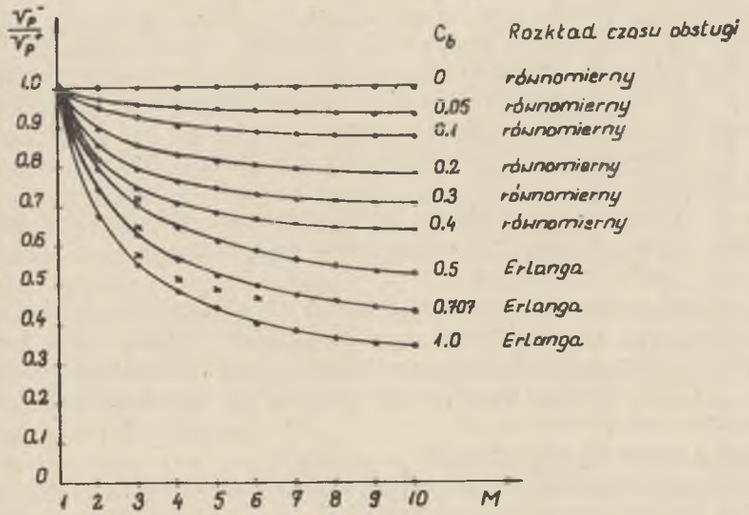
Równość:

$$V_p = V_p^+ = V_p^-$$

zachodzi przy tym dla pojedynczych WSO o regularnych rozkładach czasów obsługi.

W cytowanej pracy przedstawiono również, przytoczony na rys. IV.2, wykres charakteryzujący wartość stosunku V_p^-/V_p^+ przy różnej liczbie stanowisk obsługi oraz różnych wartościach współczynnika zmienności rozkładu

czasów obsługi (założono jednakowy, równy jedności, średni czas obsługi na każdym ze stanowisk; a więc przypadek systemu zrównoważonego oraz zerową pojemność buforów pośrednich).



Rys. IV.2. Wartości stosunku V_p^-/V_p^+ dla pojedynczego, M - etapowego WSO o jednakowych rozkładach czasów obsługi na wszystkich etapach, posiadających wartość oczekiwaną E(b) = 1 i współczynnik zmienności rozkładu C_B. Dodatkowo zaznaczono ścisłe wartości wydajności V_p niektórych, specjalnych przypadków, wyznaczone w pracy [HILL 66]. Wykresy przytoczono za [MUTH 73]

Jak to wynika wprost z podanych wyżej wzorów, wartość oszacowania V_p^+, zależna wyłącznie od wartości oczekiwanej czasu obsługi zgłoszeń, jest we wszystkich rozpatrywanych przypadkach stała, natomiast silnym zmianom podlega, zależna od postaci dystrybuanty czasu obsługi, wartość V_p^- (określona wg [SARH 62]).

Na wykresach zaznaczono również dokładne wartości V_p dla pewnych, prostych systemów, które każdorazowo okazały się znacznie bliższe wartości V_p^- aniżeli V_p^+, co wobec zerowej pojemności buforów pośrednich jest naturalne.

W szczególności, jak to wykazano w [AVII 65A], w przypadku systemu dwuetapowego bez kolejki pośredniej (M = 2, N_2 = 0) mamy $V_p = V_p^-$, co pozwala na bezpośrednie, analityczne wyznaczenie wydajności takiego systemu, przy dowolnej postaci rozkładów czasów obsługi $B_1(x)$ oraz $B_2(x)$.

W pracy tej przeprowadzono analizę scharakteryzowanych wyżej WSO w wariancie z nieograniczonym buforem wejściowym (właściwy WSO), poprzez wykazanie, że można przypadek taki sprowadzić do odpowiednio zdefiniowanego

systemu jednostanowiskowego (jednokanałowego) z różnym od pozostałych rozkładem czasu obsługi pierwszego zgłoszenia w każdym okresie czynnym, co pozwala na wyznaczenie wszystkich charakterystyk systemu, a nie tylko wydajności maksymalnej.

W dodatku A do niniejszego rozdziału omówiono sposób analizy takich systemów w przypadku ogólniejszym, gdy stanowisko obsługi pierwszego etapu charakteryzuje się zależnością czasu obsługi od czasu oczekiwania.

W dalszych rozważaniach będziemy starali się określić, które spośród parametrów nasyconych WSO wpływają na ich wydajność i jak silne są wpływy tych parametrów.

Badania te prowadzą w konsekwencji do problemu optymalizacji systemu, formułowanego zwykle w jednej z następujących postaci jako:

i) Zadanie rozbudowy systemu:

Mając dany pewien WSO należy określić, które parametry (i o ile) należy zmienić, aby najniższym kosztem osiągnąć wymagany przyrost wydajności, przy znanych kosztach rozbudowy systemu o dodatkowe elementy, jak np.: koszty jednostkowego zwiększenia pojemności bufora, dołączenia kanału obsługi.

ii) Zadanie reorganizacji systemu:

Mając dany pewien WSO należy zmodyfikować jego strukturę (przy zachowaniu stałości wszystkich zasobów) tak, aby zmaksymalizować wydajność systemu.

Zadanie to można rozwiązać na przykład drogą:

- zmiany wielkości buforów przydzielonych poszczególnym stanowiskom obsługi, przy zachowaniu łącznej pojemności buforów;
- zmiany kolejności, w której zgłoszenia przechodzą przez stanowiska obsługi (co jest dopuszczalne w pewnych przypadkach);
- zmiany rozdziału pracy pomiędzy poszczególne etapy bądź mocy stanowisk obsługi, co prowadzi do modyfikacji średnich czasów obsługi;

a w konsekwencji do sprecyzowania pewnych, ogólnych wytycznych dla projektowania WSO.

Ze względu na stosunkowo bogaty materiał źródłowy omówimy najpierw oddzielnie problematykę pojedynczych WSO a następnie przejdziemy do systemów wielokanałowych i pewnych przypadków specjalnych.

IV.2. Wpływ zrównoważenia systemu na jego wydajność

Rozpatrzmy pojedynczy, wieloetapowy system obsługi spełniający zależność (IV.3), która przybiera wówczas postać:

$$\sum_{i=1}^{M} (\mu_i)^{-1} = M. \qquad \text{(IV.2.1)}$$

Średni czas obsługi zgłoszeń na stanowisku obsługi wynika z ilości pracy, jaką na tym stanowisku należy wykonać oraz z mocy tego stanowiska. Oba te parametry mogą ulegać zmianom. W praktyce często sprawą decyzji organizacyjnej jest rozdział pracy pomiędzy stanowiska obsługi WSO - przykładem może tu być proces walcowania w walcowniach wieloklatkowych. W wielu przypadkach możliwe jest też zwiększenie mocy pewnych stanowisk obsługi kosztem innych, np. drogą przeniesienia części personelu lub przezbrojenia urządzeń.

Ogólnie rzecz biorąc, takie transformacje średnich czasów obsługi charakteryzowane są pewnymi funkcjami $\Phi_i(\mu_1,\mu_2,\ldots,\mu_M)$; i=1,2,...,I (gdzie I - pewna liczba naturalna), określającymi zakres oraz proporcje zmian czasów obsługi. Funkcje te mogą mieć złożoną postać, jak na przykład w rozpatrywanym w pracy [WART 79] zagadnieniu rozdziału pracy pomiędzy dwie klatki nawrotnych walcarek blach.

W dalszych rozważaniach przyjmiemy, że możliwe jest dowolne zmodyfikowanie średnich czasów obsługi zgłoszeń na wszystkich etapach, zachowujące wyrażoną zależnością (IV.2.1) stałość łącznego czasu obsługi, i będziemy poszukiwać rozdziału tej wartości między poszczególne etapy maksymalizującego wydajność systemu.

Na podstawie (IV.1.1) można orzec, iż w przypadku gdyby wszystkie bufory pośrednie miały nieograniczoną pojemność, wówczas rozpatrywany WSO osiągnąłby maksymalną (równą jedności) wydajność jako system zrównoważony o $\mu_i = 1$, i=1,2,...,M.

Taki sam, równomierny, podział czasu obsługi byłby optymalny również (bez względu na pojemności buforów pośrednich), gdyby przyjąć, iż wszystkie stanowiska obsługi posiadają regularne rozkłady czasów obsługi.

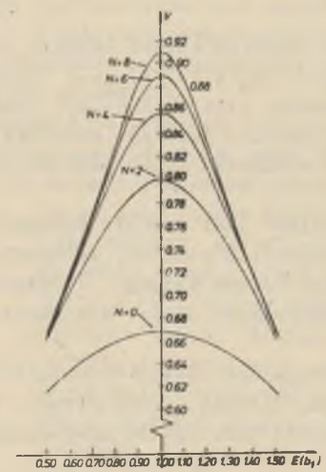
Rozpatrując WSO o ograniczonej pojemności i różnych od regularnych rozkładach czasów obsługi omówimy oddzielnie przypadek jednakowych i oddzielnie przypadek różnych rozkładów czasów obsługi na poszczególnych etapach, zakładając jednak stale jednakowe pojemności wszystkich buforów pośrednich $N_i = N$; i=2,3,...,M.

a) Identyczne rozkłady czasów obsługi na wszystkich etapach

Dwuetapowy system obsługi z wykładniczymi rozkładami czasów obsługi był pierwszym przypadkiem WSO, dla którego wyznaczono w sposób ścisły wydajność [HUNT 56]. Łatwo sprawdzić (rys. IV.3), że największą wydajnością cechuje się system zrównoważony, co wydaje się naturalne z punktu widzenia zasady odwracalności. Jak to wyraźnie widać na rysunku, straty wynikające z niezrównoważenia systemu rosną gwałtownie wraz ze wzrostem pojemności bufora pośredniego.

Hillier i Boling [HILL 66], korzystając z opracowanej przez siebie metody numerycznego wyznaczania wydajności systemów wieloetapowych, rozpatrywali przypadki M = 3,4 przy wykładniczych rozkładach czasów obsług stwierdzając, iż przy większej aniżeli dwa liczbie etapów systemy niezrównoważone mogą osiągać wydajność wyższą aniżeli systemy zrównoważo

Wskazali oni na charakterystyczną prawidłowość, którą nazywać będziemy "zjawiskiem garbu" (ang. - bowl phenomenon), polegającą na tym, iż maksymalizacja wydajności występowała, gdy rozdział pracy był symetryczny (tj. $E(b_1) = E(b_{M+1-i})$ dla i=1,2,...,M) i stanowiska obsługi ulokowane na krańcach systemu cechowały się dłuższymi czasami obsługi. Autorzy postawili hipotezę, iż taki efekt występować będzie również w bardziej rozbudowanych systemach.



Rys. IV.3. Wydajność dwuetapowych, niezrównoważonych systemów masowej obsługi z wykładniczymi rozkładami czasów obsługi (na podstawie [HILL 66])

W odniesieniu do systemu M=3 również w [PATT 64], posługując się także metodami numerycznymi, zauważono tę prawidłowość, wysuwając stąd jednak odmienny wniosek, iż prawidłowo zaprojektowany WSO powinien charakteryzować się rozmieszczaniem na przemian szybkich i powolnych stanowisk obsługi, stawiając też hipotezę, iż zysk wynikający z takiego doboru parametrów systemu będzie malał wraz ze zmniejszaniem współczynnika zmienności rozkładów czasów obsługi.

W literaturze pojawiły się także hipotezy, iż wydajność WSO będzie maksymalna, gdy stanowiska obsługi zostaną uporządkowane według rosnących średnich czasów obsługi, oparte na fakcie stwierdzenia przez Davisa [DAVI 66] takiej prawidłowości w WSO ze stratami zgłoszeń w przypadku przepełnienia.

Ostateczne wyjaśnienie tych rozbieżnych poglądów zawdzięczamy pracy [HILL 79] (wykorzystującej rozwiązania numeryczne) oraz [RAYA 79A] (opartej na modelowaniu cyfrowym). W tabeli IV.1 zamieszczono rozwiązania optymalizujące wydajność systemów wieloetapowych z wykładniczymi rozkładami czasów obsługi przy M = 3,4,5,6 i niektórych wartościach pojemności buforach N_1, które potwierdzają jednoznacznie istnienie zjawiska garbu. Hillier i Boling ograniczyli się przy tym do poszukiwania rozwiązań symetrycznych, jako iż wszystkie przeprowadzone dotąd badania wskazują, że jest to charakterystyczna własność rozwiązań optymalnych. Testy wykonane w [RAYA 79A] zdyskwalifikowały też dwie pozostałe podane wyżej hipotezy, z których "przeplatanie" szybkich i powolnych stanowisk obsługi prowadzi często do wydajności niższej aniżeli wydajność systemu zrównoważonego, natomiast monotoniczne zwiększanie średnich czasów obsługi daje z reguły wyniki znacznie gorsze aniżeli zrównoważenie systemu.

W ramach zalecanej reguły optymalizacji opartej na zjawisku garbu (tj. monotonicznym zwiększaniu średnich czasów obsługi stanowisk bliższych krańców systemu) można w różny sposób formować proporcje średnich czasów

obsługi poszczególnych stanowisk oraz wybierać różne stopnie niezrównoważenia systemu. Hillier i Boling sugerują (w oparciu o przytoczone w tabeli IV.1 wyniki), iż niezrównoważenie systemu ze wzrostem liczby stanowisk obsługi pozostaje mniej więcej stałe, na poziomie około 0.13, przy czym można przyjmować równe średnie czasy obsługi dla wszystkich stanowisk obsługi, z wyjątkiem dwóch skrajnych (x_1 oraz x_M), które powinny cechować się istotnie dłuższymi średnimi czasami obsługi.

Tabela IV.1

Optymalne niezrównoważenie wieloetapowych systemów masowej obsługi z wykładniczymi rozkładami czasów obsługi. Dane dla przypadków M=3,4,5,6 zaczerpnięto z [HILL 79], natomiast dla przypadku M=12 z [RAYA 79A], przy czym zachodzi $E(b_i) = E(b_{12-i})$ i optymalizacja była przeprowadzana przy pewnych dodatkowych ograniczeniach

M	N	$E(b_1)$	$E(b_2)$	$E(b_3)$	$E(b_4)$	$E(b_5)$	$E(b_6)$	V_p	V_p^o	$[V_p/V_p^o]_{100\%}$
3	0	1.086	0.828	1.086	-	-	-	0.56718	0.56410	100.55
	1	1.059	0.882	1.059	-	-	-	0.67376	0.67047	100.49
	2	1.044	0.912	1.044	-	-	-	0.73692	0.73402	100.40
	3	1.035	0.930	1.035	-	-	-	0.77920	0.77671	100.32
	4	1.028	0.944	1.028	-	-	-	0.80963	0.80749	100.27
4	0	1.137	0.863	0.863	1.137	-	-	0.51963	0.51478	100.94
	1	1.093	0.907	0.907	1.093	-	-	0.63617	0.63115	100.80
	2	1.069	0.931	0.931	1.069	-	-	0.70513	0.70071	100.63
	3	1.055	0.945	0.945	1.055	-	-	0.75152	0.74771	100.51
	4	1.045	0.955	0.955	1.045	-	-	0.78509	0.78178	100.42
5	0	1.169	0.898	0.866	0.838	1.169	-	0.49165	0.48580	101.20
	1	1.114	0.928	0.916	0.928	1.114	-	0.61348	0.60758	100.97
	2	1.085	0.946	0.938	0.946	1.085	-	0.68568	0.68049	100.76
6	0	1.192	0.923	0.885	0.885	0.923	1.192	0.47312	0.46671	101.37
	1	1.130	0.943	0.927	0.927	0.943	1.130	0.59817	0.59183	101.07
12	0	1.1	1.06	1.02	0.88	0.94	0.9	0.42168	0.4172	101.06

El-Rayah [RAYA 79A] badał dla M=12 trzy warianty formowania "garbu":
- przyjmowanie równych między sobą średnich czasów obsługi dla stanowisk $x_4 \div x_9$ i równych między sobą, większych niż poprzednie, średnich czasów obsługi dla pozostałych stanowisk,
- przyjmowanie równych między sobą średnich czasów obsługi dla grup stanowisk obsługi $x_1 \div x_4$, $x_5 \div x_8$, $x_9 \div x_{12}$,
- przyjmowanie rozdziału średnich czasów obsługi wg zasady: $E(b_1) = 1 + \Delta$, $E(b_i) = E(b_1) - 0.4(i-1)\Delta$, $E(b_{12} - i) = E(b_i)$, i = 1,2,3,4,5,6,

Tabela IV.2

Wydajności wybranych, niezrównoważonych, wieloetapowych systemów masowej obsługi
o wykładniczych rozkładach czasów obsługi.
Dane wg [HILL 66], [HILL 79] dla M=3,4 oraz [RAYA 79A] dla M=12 (odczyt z wykresu)

	M = 3 $E(b_1) = E(b_3) = 1 + \Delta$ $E(b_2) = 1 - 2\Delta$				M = 4 $E(b_1)=E(b_4)=1+\Delta$ $E(b_2)=E(b_3)=1-\Delta$		M = 12 $E(b_1)=1+\Delta$ $E(b_1)=E(b_1)-0.4(i-1)\Delta$ $E(b_{12-i})=E(b_i),i=1+6$	
	N=0	N=2	N=4		N=0		N=0	
Optymalne niezrównoważenie systemu	Δ = 0.086 $E(b_2)$=0.828	Δ = 0.044 $E(b_2)$=0.912	Δ=0.028 $E(b_2)$=0.944		Δ=0.134 $E(b_2)$=0.866		Δ=0.1 $E(b_6)$=0.9	
Procentowy przyrost wydajności systemu optymalnie niezrównoważonego w stosunku do zrównoważonego $100(V_p/V_p^0)$ [%]	100.55	100.4	100.27		100.94		100.1	
Zakres zmian niezrównoważenia systemu, zachowujący wartość wydajności nie mniejszą aniżeli V_p^x, taką że $100\left[(V_p-V_p^x)/V_p^0\right] < 0.1$	Δ=0.13-0.04 $E(b_2)$=0.74-0.92	Δ=0.07-0.02 $E(b_2)$=0.86-0.96	Δ=0.08-0.01 $E(b_2)$=0.92-0.98		Δ=0.18-0.08 $E(b_2)$=0.82-0.92		Δ=0.145-0.05 $E(b_6)$=0.855-0.95	
Niezrównoważenie systemu (we właściwym kierunku) zachowujące wydajność równą wydajności systemu zrównoważonego	Δ=0.22 $E(b_2)$=0.66	Δ=0.09 $E(b_2)$=0.82	Δ=0.06 $E(b_2)$=0.88		Δ=0.28 $E(b_2)$=0.72		Δ=0.2 $E(b_6)$=0.8	

stwierdzając jednoznacznie wyższość trzeciego z tych wariantów nad pozostałymi (optymalne rozwiązanie w ramach tego wariantu zawiera tabela IV.1).

Zamieszczone wyniki pozwalają stwierdzić, iż aczkolwiek optymalny stopień niezrównoważenia systemu szybko maleje ze wzrostem pojemności buforów pośrednich, to jednak przyrost wydajności uzyskany dzięki temu niezrównoważeniu zmniejsza się znacznie wolniej, uzasadniając celowość stosowania systemów niezrównoważonych również w przypadku użycia buforów o większych pojemnościach [HILL 79].

Tabela IV.3

Optymalne niezrównoważenie trzyetapowego systemu masowej obsługi
wg [HILL 79]
V_p - wydajność systemu optymalnie niezrównoważonego
V_p^0 - wydajność systemu zrównoważonego

k	N	$E(b_1)=E(b_3)$	$E(b_2)$	V_p	V_p^0	$\frac{V_p}{V_p^0}$ 100%
1	2	3	4	5	6	7
1	0	1.086	0.828	0.56718	0.56410	100.55
	1	1.059	0.882	0.67376	0.67047	100.49
	2	1.044	0.912	0.73692	0.73402	100.40
	3	1.035	0.930	0.77920	0.77671	100.32
	4	1.028	0.944	0.80963	0.80749	100.27
2	0	1.067	0.866	0.64360	0.64041	100.50
	1	1.038	0.924	0.76395	0.76131	100.35
	2	1.026	0.948	0.82387	0.82189	100.24
	3	1.019	0.962	0.85950	0.85795	100.18
	4	1.015	0.970	0.88312	0.88186	100.14
3	0	1.056	0.888	0.68731	0.68422	100.45
	1	1.028	0.944	0.81161	0.80947	100.26
	2	1.018	0.964	0.86600	0.86453	100.17
	3	1.013	0.974	0.89603	0.89493	100.12
	4	1.010	0.980	0.91506	0.91418	100.10
4	0	1.050	0.900	0.71695	0.71400	100.41
	1	1.022	0.956	0.84196	0.84017	100.21
	2	1.014	0.972	0.89129	0.89012	100.13
5	0	1.045	0.910	0.73891	0.73609	100.38
	1	1.019	0.962	0.86326	0.86173	100.18
6	0	1.041	0.918	0.75609	0.753384	100.36
7	0	1.039	0.922	0.77005	0.76745	100.34

Zysk wynikający ze stosowania systemów niezrównoważonych rośnie natomiast ze wzrostem liczby etapów.

Bardzo istotnym praktycznie zagadnieniem jest wrażliwość wydajności WSO na prawidłowy dobór średnich czasów obsługi (prawidłowość niezrównoważenia systemu). W tabeli IV.2 w oparciu o dane zawarte w cytowanych pracach określono zakres zmian parametru określającego niezrównoważenie systemu (w prawidłowy sposób), przy którym system niezrównoważony cechuje się niewielkim, w stosunku do przypadku niezrównoważenia optymalnego, obniżeniem wydajności, a także wartości tego parametru, przy których system przesadnie niezrównoważony osiąga wydajność równą wydajności systemu zrównoważonego.

W sposób wyraźny widać, iż elementem decydującym jest prawidłowość reguły niezrównoważenia, natomiast nie jest wymagana duża precyzja w doborze wartości. I tak na przykład w odniesieniu do systemu czteroetapowego bez buforów pośrednich wszystkie wartości $E(b_2) = E(b_3) = 2-E(b_1) = 2-E(b_4)$, leżące w przedziale $0.72 \leq E(b_2) \leq 1$, zapewniają wydajność większą aniżeli system zrównoważony.

Ogólne zasady podane wyżej pozostają w mocy również w przypadku, gdy rozkłady czasów obsługi są różne od rozkładów wykładniczych, co sprawdzono w [HILL 79], rozpatrując rozkłady Erlanga k-tego rzędu, k=1÷7 (tabela IV.3) oraz w [RAYA 79A], rozpatrując rozkład normalny (obcięty) oraz rozkład log - normalny.

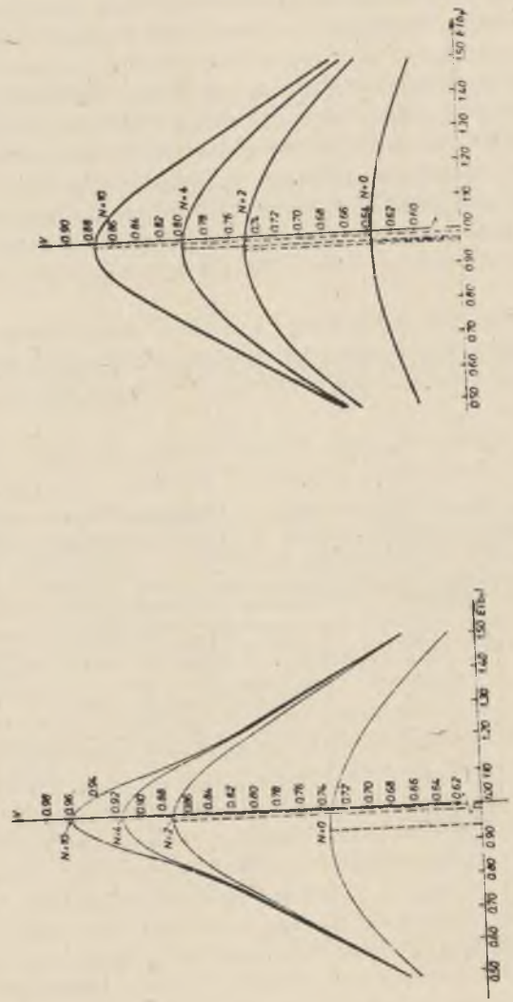
Zmniejszenie współczynnika zmienności rozkładu powoduje obniżenie optymalnego stopnia niezrównoważenia (oraz zysku wynikającego z niezrównoważenia), przy czym efekt ten jest tym silniejszy, im większa jest pojemność buforów pośrednich.

b) Różne rozkłady czasów obsługi

W pracy [RAON 75B] przedstawiono metodę wyznaczania wydajności dwuetapowych systemów obsługi z wykładniczym rozkładem czasów obsługi na jednym ze stanowisk i dowolnym rozkładem czasów obsługi na drugim z nich, przy czym jako integralny krok tej metody wymagane było rozwiązanie układu równań liniowych o rozmiarze zależnym od pojemności bufora pośredniego N.

Badanie przypadków obejmujących rozkłady Erlanga oraz rozkład normalny (obcięty) o współczynniku zmienności $C_B < 1$ doprowadziło do wniosku, że wydajność systemu można zoptymalizować projektując system tak, aby stanowisko obsługi charakteryzujące się większą wartością współczynnika zmienności (a więc stanowisko obsługi o wykładniczym rozkładzie czasów obsługi) miało nieco krótszy średni czas obsługi. Zarówno stopień niezrównoważenia, jak i zysk wynikający z tej modyfikacji systemu wzrastają ze wzrostem różnicy współczynników zmienności rozkładów czasów obsługi na obu etapach, natomiast maleją silnie ze wzrostem N.

Identyczne WSO rozpatrywano w [WOLI 81C], [ZSAK 79], wyznaczając ich wydajność na drodze analitycznej (patrz Dodatek B). Potwierdzono wcześ-



Rys. IV.4. Wydajność dwuetapowego, niezrównoważonego systemu masowej obsługi. Założono wykładniczy rozkład czasów obsługi na pierwszym etapie oraz regularny (rys. a) względnie hiperwykładniczy (rys.b) rozkład czasu obsługi na drugim etapie (według [WOLI 81C])

niejsze wyniki dla $C_B < 1$ i zwrócono uwagę na zaskakujące nieregularności występujące przy rozpatrywaniu przypadku $C_B > 1$ (w oparciu o rozkład hiperwykładniczy drugiego rzędu). Maksymalną wydajność uzyskiwano również przy nieco krótszych średnich czasach obsługi na stanowisku obsługi charakteryzującym się wykładniczym rozkładem czasu obsługi, które jednak w tym przypadku było stanowiskiem o mniejszym współczynniku zmienności rozkładu.

Również wzrost pojemności bufora N powodował początkowo zwiększenie wymaganego stopnia niezrównoważenia (i zysków stąd wynikających), podczas gdy dalsze zwiększenie wartości N prowadziło do efektu odwrotnego.

Przykładowe wyniki dla wartości $C_B = 0$ oraz $C_B = 3$ ilustruje rys. IV.4. Jako typowe należy uznać dla systemów dwuetapowych o różnych rozkładach czasów obsługi wartości $E(b_1) = 0.92 \div 0.96$, przy których można oczekiwać zwiększenia wydajności o około 0.2÷0.3% w stosunku do wariantu zrównoważonego.

W pracy [RAOM 76] rozpatrzono w sposób analityczny przypadek M = 3 z buforami pośrednimi o zerowej pojemności, przyjmując wszystkie możliwe kombinacje regularnego i wykładniczego rozkładu czasu obsługi na poszczególnych etapach.

Tabela IV.4

Optymalne niezrównoważenie systemu trzyetapowego z różnymi rozkładami czasów obsługi na poszczególnych etapach (dane zaczerpnięte z [RAOM 76] oraz [HILL 66]).

Rozpatrywane warianty scharakteryzowano przez podanie rozkładu czasu obsługi na każdym z etapów M - rozkład wykładniczy, D - rozkład regularny

Rozpatrywany wariant	Wydajność systemu zrównoważonego	Zysk wynikający z optymalnego niezrównoważenia	Optymalny sposób niezrównoważenia systemu		
			$E(b_1)$	$E(b_2)$	$E(b_3)$
a) M M D	0.6160	102.80	1.00	0.73	1.27
b) D M M	0.6160	102.80	1.27	0.73	1.00
c) M D M	0.6167	100.35	0.945	1.11	0.945
d) M D D	0.7311	106.79	0.62	1.19	1.19
e) D D M	0.7311	106.79	1.19	1.19	0.62
f) D M D	0.7311	106.79	1.19	0.62	1.19
g) D D D	1	Brak	1	1	1
h) M M M	0.5641	100.5	1.09	0.82	1.09

W tabeli IV.4 zacytowano wyniki określające optymalne niezrównoważenie w takich systemach (wśród których wyróżniają się systemy sprzężone).

Rao zwrócią uwagę na celowość przyporządkowywania krótszego średniego czasu obsługi stanowiskom obsługi posiadającym większą wartość współczynnika zmienności rozkładu ("efekt zmienności rozkładu"), podkreślając, że efekt ten oraz omawiany poprzednio "efekt garbu" mogą się wzajemnie wzmacniać (wariant f) lub mogą wzajemnie sobie przeciwdziałać (wariant e). Niezrównoważenie systemu spowodowane "efektem zmienności rozkładu" powinno być tym silniejsze, im większe są różnice współczynników zmienności rozkładu czasów obsługi na poszczególnych etapach.

Wykazano też, że w przypadku systemu trzyetapowego, w którym skrajne stanowiska obsługi mają wykładnicze rozkłady czasów obsługi, natomiast środkowe stanowisko obsługi ma równomierny rozkład czasów obsługi o wartości $C_B = 0.5$, maksymalną wydajność osiąga się w wariancie zrównoważonym.

Podsumowując, można stwierdzić, iż w pojedynczych WSO maksymalną wydajność uzyskuje się, ogólnie rzecz biorąc, drogą wprowadzenia niezrównoważenia systemu. Wskazano na dwa efekty sugerujące kierunek zróżnicowania wydajności poszczególnych stanowisk obsługi, przy czym efekt związany ze zróżnicowaniem rozkładów czasów obsługi jest silniejszy. Zalecane niezrównoważenie systemów - nawet przy niewielkiej liczbie etapów - może być znaczne i prowadzić do wyraźnych zysków.

IV.3. Wpływ rozkładu czasu obsługi oraz pojemności buforów pośrednich na wydajność WSO

Nieregularny charakter czasów obsługi zgłoszeń wynika w sposób naturalny z nieuniknionych, losowych różnic w zapotrzebowaniu zgłoszeń na obsługę oraz niewielkich wahań mocy stanowisk obsługi. Wynikające stąd zróżnicowanie czasów obsługi jest silnie związane ze specyfiką modelowanego procesu (np. w odniesieniu do linii produkcyjnych obserwuje się z zasady niewielkie wartości współczynnika zmienności rozkładu - jako typową można przyjąć wyznaczoną w [SLAC 72] wartość $C_B = 0.72$; w odniesieniu do systemów teletransmisji danych obserwowano wielokrotnie wartości C bliskie jedności).

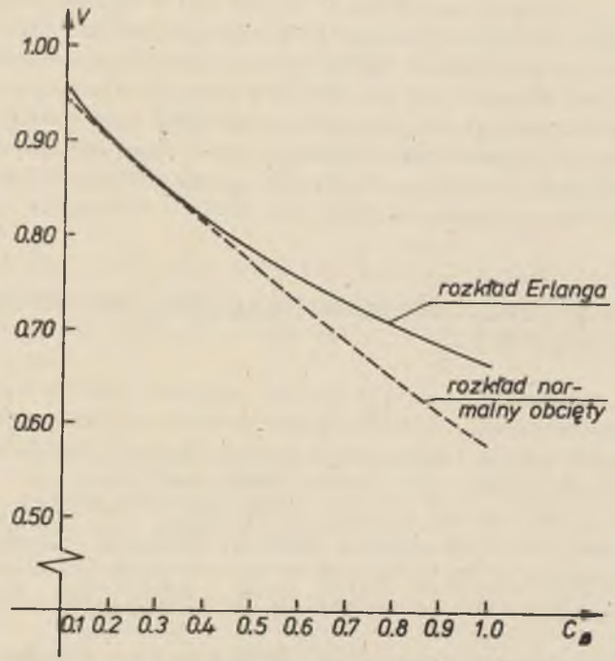
Realistyczny opis WSO wymaga często uwzględnienia możliwości powstawania awarii stanowisk obsługi, przy czym zarówno średni czas pomiędzy awariami, jak i średni czas naprawy stanowiska obsługi przekraczają zwykle wielokrotnie wartości średniego czasu obsługi zgłoszenia przy poprawnej pracy stanowiska obsługi. Łączne rozpatrzenie nieregularności czasów obsługi oraz awarii stanowisk obsługi jest trudne, stąd w literaturze rozpatrywane są z reguły dwa standardowe modele:

- całkowicie niezawodne stanowiska obsługi, w których możliwość awarii uwzględnia się co najwyżej drogą podwyższenia wartości współczynnika zmienności rozkładu,
- regularne rozkłady obsługi z możliwością występowania okresowej niesprawności stanowisk obsługi.

Pozostając w obrębie pierwszego z tych modeli rozpatrzymy wpływ nieregularności czasów obsługi na wydajność WSO.

a) Skutki pominięcia wyższych momentów rozkładów czasów obsługi

Jak na to zwrócono uwagę w rozdziale I, w praktyce często charakteryzując rozkłady czasów obsługi ograniczamy się do określenia ich wartości oczekiwanej oraz współczynnika zmienności rozkładu. Przedyskutujemy obecnie skutki takiego uproszczenia w odniesieniu do WSO.



Rys. IV.5. Wydajność dwuetapowego, zrównoważonego systemu obsługi o identycznych rozkładach czasów obsługi na obu etapach w funkcji współczynnika zmienności rozkładu, przy N=0. Dane zaczerpnięte z [RAON 75A]

Na rys. IV.5 podano w ślad za [RAON 75A] porównanie wydajności dwuetapowych systemów obsługi bez bufora pośredniego przy różnych, cechujących się identycznymi wartościami dwóch pierwszych momentów, rozkładach czasów obsługi, stwierdzając, iż niewielkie ilościowo różnice narastają ze wzrostem C_B.

W [RAON 75B] podkreślono, iż różnice te stają się jeszcze wyraźniejsze, jeśli rozpatrywane są rozkłady o identycznych wartościach E(b) oraz C_B, lecz o silnie różniących się przebiegach dystrybuant, jak na przykład rozkład Erlanga i rozkład równomierny; dodatkowo ich zwiększeniu sprzyja mała pojemność bufora pośredniego (tabela IV.5).

W pracy [ANDE 69] rozpatrywano, przy użyciu modelowania cyfrowego, zrównoważone WSO, posiadające wykładnicze bądź normalne (obcięte) rozkłady czasów obsługi. Badania te, mające na celu wskazanie optymalnego rozdziału pojemności buforów (bardziej szczegółowo omówimy je nieco później), wykazały całkowitą zgodność jakościową oraz duże podobieństwo ilościowe wyników uzyskiwanych w obu przypadkach.

Tabela IV.

Wydajność dwuetapowego, zrównoważonego systemu masowej obsługi z regularnym (C_1=0) lub wykładniczym (C_1=1) rozkładem czasów obsługi na etapie pierwszym i dwoma różnymi rozkładami czasów obsługi (posiadającymi współczynnik zmienności rozkładu C_2) na etapie drugim. Dane zaczerpnięte z [RAON 75B]

C_2	Rozkład czasów obsługi na drugim etapie					
	Równomierny			Erlanga		
	C_1=0, N=0	C_1=1, N=0	C_1=1, N=1	C_1=0, N=0	C_1=1, N=0	C_1=1, N=1
0.00	1.0	0.73106	0.82366	1.0	0.73106	0.82366
0.10	0.95850	0.73008	0.82263	0.91670	0.73008	0.82263
0.20	0.92030	0.72712	0.81954	0.92633	0.72721	0.81960
0.30	0.88503	0.72220	0.81443	0.89372	0.72262	0.81471
0.40	0.85237	0.71530	0.80730	0.86530	0.71635	0.80773
0.50	0.82203	0.70640	0.79824	0.83655	0.70942	0.80029

Identyczne wnioski z badań symulacyjnych WSO posiadających normalne bądź też log - normalne rozkłady czasów obsługi wyciągnięto też w [RAYA 79A] w odniesieniu zarówno do systemów zrównoważonych, jak i niezrównoważonych o liczbie etapów M = 3,4.

Można więc stwierdzić, iż w zakresie $C_b < 1$ nie stwierdzono wpływu wyższych momentów rozkładów czasów obsługi na charakter reguł optymalizacji WSO (sposób równoważenia, rozdział buforów), występują natomiast pewne różnice ilościowe nasilające się w omówionych wyżej sytuacjach.

Z tego względu dla potrzeb rozważań analitycznych można zwykle uznać za uzasadnione ograniczenie się do dobrania jedynie pierwszych dwóch momentów rozkładów. W przypadku modelowania cyfrowego, gdzie bardziej precyzyjne odwzorowanie postaci dystrybuanty nie jest kłopotliwe, korzysta się natomiast często z rozkładów Weibulla, łatwych do uzyskania w generatorach liczb losowych i umożliwiających uwzględnienie również wyższych momentów [CARN 76], [SLAC 72].

b) Wpływ współczynnika zmienności rozkładu czasów obsługi na wydajność WSO

Zamieszczone wyżej rys. IV.3, IV.4, IV.5 oraz tabela IV.3 i IV.4 wskazują jednoznacznie na fakt zmniejszenia się wydajności WSO ze wzrostem współczynnika zmienności rozkładu czasów obsługi chociażby tylko na niektórych stanowiskach obsługi.

Na zjawisko zwrócono uwagę również w [BART 62], gdzie badając na drodze symulacyjnej sześcioetapowe systemy obsługi stwierdzono, iż ulega ono osłabieniu w miarę wzrostu pojemności buforów pośrednich. Podobne spostrzeżenie poczyniono również w [CARN 76] w oparciu o badania symulacyjne, prowadzone przy użyciu rozkładów Weibulla.

Warto zauważyć, iż bardzo podobne zjawisko występuje przy rozpatrywaniu drugiego spośród wymienionych na wstępie tego podrozdziału modelu WSO z regularnymi rozkładami czasów obsługi i podlegającymi awariom stanowiskami obsługi.

Aczkolwiek badania takich systemów omawiane np. w [BUZA 67], [OKAM 77], [OKAM 78] dotyczyły synchronizowanych WSO, to jednak w [BUZA 72] zademonstrowano bezpośrednią stosowalność uzyskanych wyników również w odniesieniu do systemów swobodnych.

W szczególności rozpatrywano na drodze analitycznej system dwuetapowy z niezerową pojemnością bufora pośredniego i wykładniczymi rozkładami czasów obsługi, w którym dopuszczalne były pojawiające się w sposób stochastyczny uszkodzenia o określonym przez pewną zmienną losową czasie trwania.

Wskazano, iż zmniejszenie wydajności takiego systemu w stosunku do systemu dwuetapowego, w pełni niezawodnego i charakteryzującego się regularnymi rozkładami czasów obsługi, można bardzo dokładnie aproksymować drogą określenia sumy strat wynikających z oddzielnego uwzględnienia w takim wzorcowym przypadku bądź to nieregularności czasów obsługi, bądź to awarii stanowisk obsługi. Sformułowano też hipotezę, iż podobna superpozycja jest możliwa również w odniesieniu do systemów o większej liczbie etapów i innych rozkładach czasów obsługi.

c) <u>Zasady rozmieszczenia stanowisk obsługi o różnych rozkładach czasów obsługi w zróżnoważonych WSO</u>

Wyniki eksperymentów symulacyjnych polegających na różnym uszeregowaniu trzech pojedynczych stanowisk obsługi, cechujących się różnymi wartościami współczynników zmienności rozkładu czasu obsługi, podano w [SMIT 77]. Autorzy sugerowali, że umieszczenie jako środkowego tego spośród rozpatrywanych stanowisk, które posiadało najwyższą wartość C_B, było zdecydowanie niekorzystne (w sensie wydajności uśrednionej po różnych wariantach rozdziału buforów).

Przedstawione w tabeli 1 cytowanej pracy wyniki, jako sprzeczne z zasadą odwracalności, nie pozwalają jednak na wyciągnięcie bardziej szczegółowych wniosków.

Wyniki symulacji cztero- oraz dziesięciomtapowych systemów o różnych wartościach N omawiano w [CARN 76]; zakładając, iż część stanowisk obsługi posiada regularne rozkłady czasów obsługi, w pozostałych natomiast czas obsługi określony jest rozkładem Weibulla, przy czym w ramach każdego z tych dwu typów stanowiska obsługi były identyczne.

Stwierdzono, iż dla M = 4 i dwu stanowisk obsługi każdego typu rozmieszczenie stanowisk obsługi o regularnym rozkładzie czasów obsługi w środku prowadziło do istotnego wzrostu wydajności w stosunku do wariantu, w którym byłyby one rozmieszczone na brzegach (tj. jako pierwsze i czwarte).

Zysk wynikający z prawidłowego rozmieszczenia wzrasta ze wzrostem wartości C_B, maleje natomiast ze wzrostem pojemności buforów pośrednich, osiągając przykładowo przy $C_B = 0,5$ oraz N = 1 wartość 4%.

Eksperymenty prowadzone dla M = 10 pozwoliły stwierdzić, że rozmieszczenie sześciu stanowisk obsługi o regularnych rozkładach czasów obsługi jako środkowych prowadziło do wzrostu wydajności rzędu 1.33% w porównaniu z losowym rozmieszczeniem stanowisk oraz 3% w porównaniu z wariantem, w którym stanowiska te zgrupowane byłyby po trzy na początku i końcu WSO. Autorzy wyciągnęli stąd wniosek o istnieniu w zróżnoważonych WSO odpowiednika "zjawiska garbu" w odniesieniu do współczynnika zmienności rozkładu czasów obsługi zgłoszeń.

Szczegółowe badania symulacyjne WSO o normalnych rozkładach czasów obsługi oraz N = 0 opisano w [RAYA 79B]. Przy M = 3 potwierdzono niewłaściwość umiejscowienia jako środkowego, tego spośród stanowisk obsługi, które cechuje najwyższa wartość C_B, potwierdzono też podane wyżej sugestie optymalnej alokacji stanowisk obsługi przy M = 4. Co więcej, autor wykazał, iż wydajność WSO o zróżnicowanych współczynnikach zmienności rozkładu może przewyższyć wydajność WSO o identycznych stanowiskach obsługi (przy założeniu, iż w obu przypadkach suma współczynników zmienności rozkładów wszystkich stanowisk obsługi jest stała).

W rozpatrywanym zakresie zmian $C_B < 0.3$ stwierdzono, iż wzrost wartości C_B prowadził do podobnej zmiany wydajności systemu jak wydłużenie średniego czasu obsługi (niezrównoważenie) tego samego stanowiska obsługi. Jeśli np. w zbiorze 12 stanowisk obsługi cztery posiadały $C_B = 0.15$, natomiast pozostałe charakteryzowały się wartością $C_B = 0.3$, wówczas umieszczenie tych pierwszych w centrum WSO zwiększało wydajność systemu o około 2.5% w stosunku do wariantu, w którym stanowiska te znalazłyby się (w jednej grupie) na początku względnie na końcu systemu.

d) <u>Wpływ pojemności buforów pośrednich na wydajność WSO</u>

Powyżej kilkakrotnie wspominano, iż zależnie od pojemności buforów pośrednich omawiane efekty ulegać mogą wzmocnieniu bądź osłabieniu. Obecnie zilustrujemy bezpośredni wpływ tego parametru na wydajność WSO.

Zadaniem buforów pośrednich jest łagodzenie skutków nierównomierności pracy kolejnych stanowisk obsługi, stąd też ich pojemność nie mogła mieć wpływu na wydajność WSO, w których wszystkie stanowiska obsługi charakteryzowały regularne rozkłady czasów obsługi. Z drugiej strony, z optymalności systemów zrównoważonych w przypadku, gdy wszystkie bufory mają nieograniczoną pojemność, wynika [BUZA 67], [MUTH 73], że wprowadzenie buforów pośrednich nie może zniwelować skutków zróżnicowania średnich czasów obsługi na poszczególnych etapach.

Tabela IV.6

Wartość funkcji $V_p(N)$ oraz $E_p(N)$ w dwuetapowym systemie obsługi o wykładniczych rozkładach czasów obsługi $E(b_1) = E(b_2) = 1$

N	0	1	2	3	4	5	6	7	8	9	10
$V_p(N)$	0.66	0.75	0.8	0.833	0.857	0.876	0.889	0.9	0.901	0.917	0.924
$E_p(N)$	12.5	6.66	4.17	2.86	2.08	1.59	1.25	1.01	0.834	0.7	0.595

Co więcej, jeśli pewien bufor pośredni jest stale pusty (lub stale zapełniony), wówczas nie spełnia on swego zadania. Bez troski o precyzję można więc powiedzieć [BUZA 71A], iż miarą efektywności bufora jest wariancja zmian długości kolejki zgłoszeń tworzonej w tym buforze.

Miarą łącznej efektywności wszystkich buforów WSO może być natomiast stosunek V_p/V_p^+ (przypomnijmy, iż we wszystkich rozpatrywanych systemach zrównoważonych z założenia zachodzi $V_p^+ = 1$). Na rys. IV.3 widać wyraźnie wpływ pojemności N bufora pośredniego na wydajność prostego systemu dwuetapowego. Wpływ ten można wyrazić w postaci analitycznej.

Jak to udowodniono w [HUNT 56], wydajność zrównoważonego, dwuetapowego systemu masowej obsługi, w którym $E(b_1) = E(b_2) = 1$, wyraża się wzorem:

$$V_p = V_p(N) = \frac{N+2}{N+3}. \qquad (IV.3.1)$$

Tak więc zwiększenie o jedność pojemności bufora prowadzi do względnego wzrostu wydajności $E_p(N)$

$$E_p(N) = \frac{V_p(N+1) - V_p(N)}{V_p(N)} = \frac{1}{N^2 + 6N + 8}. \qquad (IV.3.2)$$

Wartości $V_p(N)$ oraz $E_p(N)$ zestawiono w tabeli IV.6, ilustrując fakt, iż jakkolwiek zwiększenie pojemności bufora prowadzi zawsze do wzrostu wydajności rozpatrywanego systemu, to jednak przyrosty te, bardzo wyraźne przy małych N, stają się ze wzrostem N znikome i stosowanie buforów o pojemności większej aniżeli N = 5 staje się mało opłacalne [HUNT 56].

Podobne wnioski można wyciągnąć na podstawie rys. IV.4 w stosunku do dwuetapowych systemów obsługi z innymi rozkładami czasów obsługi, natomiast na podstawie tabeli IV.1 oraz IV.3 w stosunku do systemów wieloetapowych.

Sugestia ta została poparta w [BART 62] wynikami modelowania cyfrowego WSO z normalnymi rozkładami czasów obsługi (przy M = 4,6,10 oraz różnych wartościach C_B), a także wynikami badań opisanych w [SLAC 75].

Oczywiście, jak to wynika np. z tabeli IV.5, minimalna pojemność bufora niezbędna dla osiągnięcia pewnej zadanej wydajności systemu wzrasta ze wzrostem współczynników zmienności rozkładów czasów obsługi.

Sformułowane wyżej sugestie ulegają gwałtownej zmianie w przypadku uwzględnienia możliwości awarii stanowiska obsługi, kiedy to zaleca się, [BUZA 67], stosowanie buforów o pojemności 2-3-krotnie większej aniżeli średnia liczba zgłoszeń obsługiwanych w ciągu średniego czasu naprawy stanowiska obsługi (co prowadzi do pojemności buforów rzędu 30-50).

Zadanie wyboru łącznej pojemności N buforów pośrednich w WSO, a specjalnie problem ich rozdziału pomiędzy poszczególne etapy był przedmiotem wielu prac, np. [FREE 64], [HATC 69], [KNOT 78], [KOEN 56], [KRAE 70], [MASS 74].

Większość tych prac dotyczy jednak synchronizowanych WSO i te wyniki nie mogą być wprost przenoszone na systemy swobodne, co wykażemy na przykładzie:

W odniesieniu do systemów synchronizowanych wskazuje się, [FREE 64], [MASS 74], iż stanowiska obsługi o niskiej niezawodności (a więc innymi słowy, stanowiska obsługi o dużym współczynniku zmienności rozkładu czasów obsługi) umieszczone na końcu WSO wymagają dla zapewnienia odpowiedniej wydajności systemu przydzielenia buforów o większej pojemności aniżeli identyczne stanowiska obsługi ulokowane na początku WSO, co w odniesieniu do systemów swobodnych stałoby w jawnej sprzeczności z zasadą odwracalności.

Szczegółowe, prowadzone przy użyciu modelowania cyfrowego, badania problematyki rozdziału pojemności buforów w WSO opisano w [RAYA 79B]. Rozpatrzono zrównoważone WSO o M = 3,4 przy założeniu, że wszystkie stanowiska posiadają jednakowe współczynniki zmienności rozkładu. Przyjmując pewną, ustaloną wartość $\mathcal{N} = \sum_{i=1}^{M} N_i$ rozważono dobór pojemności poszczególnych buforów optymalizujący wydajność systemu.

Stwierdzono, iż przy $\mathcal{N} \leq (M-1)*4$ największą wydajność uzyskuje się przy równych buforach, czego można było się spodziewać w oparciu o poprzednio omawiane prawidłowości, jako że jednostkowe zmniejszenie pojemności któregokolwiek bufora w stosunku do takiego, równomiernego podziału prowadziłoby do istotnego obniżenia wydajności odpowiedniego etapu.

W przypadku stosowania $\mathcal{N} > (M-1)*4$ równomierny rozdział pojemności buforów był rozwiązaniem bliskim optymalnego, w pewnych przypadkach nieznacznie lepsze rezultaty uzyskiwano przyporządkowując nieco większą pojemność buforom środkowych etapów, co sugeruje pewne, aczkolwiek znacznie słabsze pod względem ilościowym, podobieństwo do zjawiska garbu.

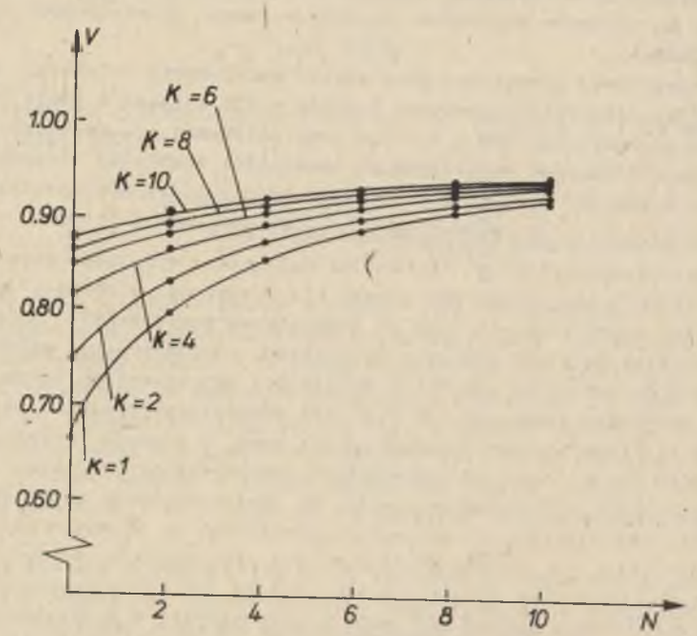
Aczkolwiek nie należy spodziewać się praktycznego wykorzystania tego zjawiska, jako zbyt słabego, w stosunku do WSO o identycznych stanowiskach obsługi, to może ono zostać wykorzystane szerzej w przypadku, gdy poszczególne etapy cechują się różnymi współczynnikami zmienności rozkładów czasów obsługi. W szczególności w [SMIT 77] stwierdzono, iż stosunkowo większe powinna być pojemność buforów otaczających stanowisko obsługi o

większym współczynniku zmienności rozkładu. Odchyłki od równomiernego rozdziału miały przy tym większe znaczenie przy małych wartościach $\mathcal{N}$.

IV.4. Wpływ wprowadzenia stanowisk wielokanałowych na wydajność WSO

W dotychczasowych rozważaniach ograniczyliśmy się wyłącznie do pojedynczych WSO. Obecnie rozpatrzymy niektóre własności systemów wielokanałowych.

W oparciu o modelowanie cyfrowe w pracy [WILD 72] przedstawiono porównanie wydajności układu złożonego z dwóch pojedynczych systemów dwuetapowych z wydajnością systemu dwuetapowego, złożonego z dwóch dwukanałowych stanowisk obsługi.
W obu rozpatrywanych przypadkach przyjęto jednakową pojemność bufora pośredniego przypadającą na pojedynczy kanał obsługi drugiego etapu.
Stwierdzono, że system wielokanałowy posiadał zawsze większą wydajność, przy czym różnica wydajności obu porównywanych wariantów wzrastała ze wzrostem liczby etapów oraz wzrostem współczynników zmienności rozkładów, a także zmniejszaniem pojemności bufora pośredniego.

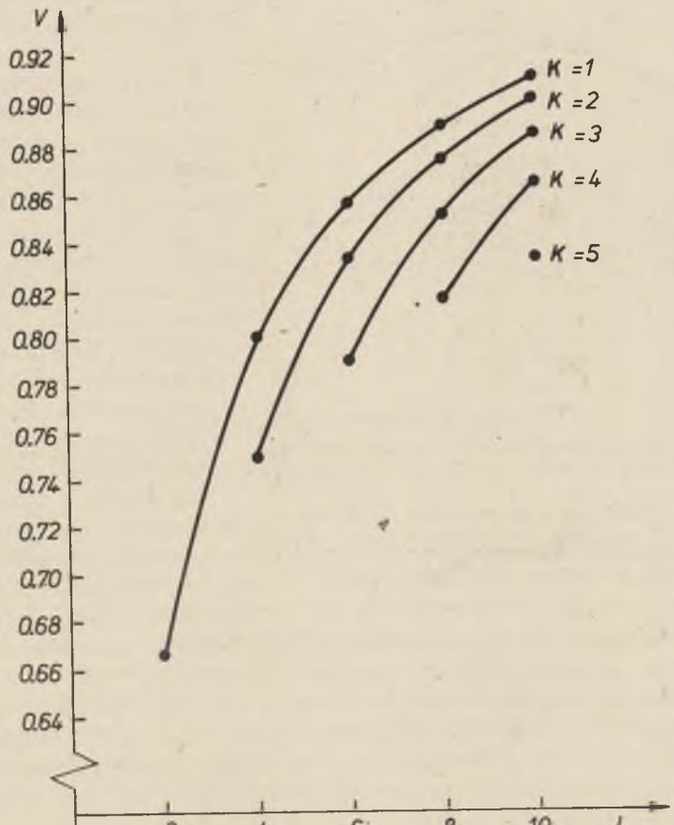


Rys. IV.6. Wpływ pojemności bufora pośredniego N oraz liczby kanałów obsługi $n_1 = n_2 = K$ na wydajność dwuetapowego systemu masowej obsługi (wg [WOLI 80A])

Szczegółowe badania wpływu poszczególnych parametrów na wydajność dwuetapowych, jednorodnych systemów wielokanałowych, przy założeniu wykładniczych rozkładów czasów obsługi, przedstawiono w [WOLI 80A], korzystając z rozwiązań analitycznych omówionych w dodatku c do niniejszego rozdziału. Poniżej przytoczymy wyniki tych badań.

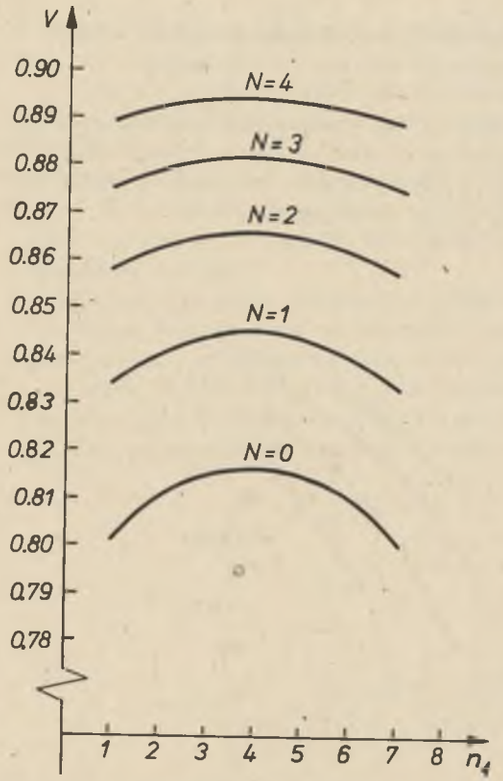
a) System zrównoważony

Wpływ pojemności bufora pośredniego N oraz liczby kanałów obsługi na każdym z etapów $n_1 = n_2 = K$ na wydajność systemu przedstawiono na rys. IV.6.
Uzyskane wyniki wskazują, iż - podobnie jak w przypadku systemów pojedynczych - pojemność bufora pośredniego ma istotny wpływ na wydajność systemu przy małych N i wpływ ten maleje, gdy N wzrasta (oczywiście wydajność rozpatrywanego systemu dąży do wartości V = 1 gdy N→∞).



Rys. IV.7. Porównanie wydajności systemów dwuetapowych o jednakowej, równej L pojemności, przy różnych wartościach N oraz $n_1 = n_2 = K$, dla L = 2k+N (wg [WOLI 80A])

Równocześnie warto zauważyć, że wydajność systemu rośnie ze wzrostem liczby kanałów obsługi K, co wynika z faktu, iż w stanie zablokowania spełniają one również funkcję dodatkowych buforów. Tak więc w przypadku, gdy ograniczona jest pojemność bufora pośredniego, wówczas zwiększenie wydajności rozpatrywanego systemu można osiągnąć drogą użycia na każdym z etapów dużej liczby mało wydajnych kanałów obsługi zamiast małej liczby kanałów wysoko wydajnych. Zysk wynikający z takiej decyzji jest tym większy, im mniejsza jest pojemność bufora pośredniego N.



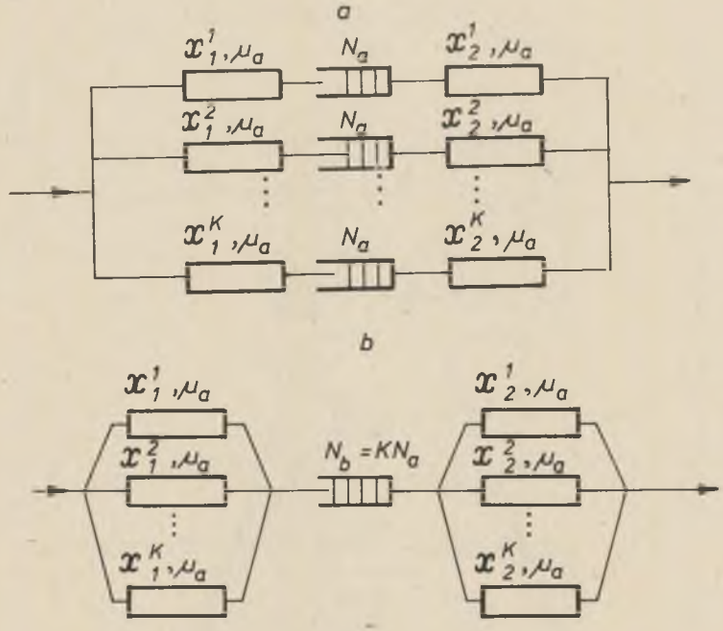
Rys. IV.8. Porównanie wydajności systemów dwuetapowych o różnej liczbie kanałów obsługi na poszczególnych etapach. Przyjęto, że $n_1 = n_2 = \text{const.}$ (wg [WOLI 80A])

W tej sytuacji uzasadnione jest pytanie, czy wydajność systemu zależy jednoznacznie od jego łącznej pojemności L, określonej wzorem (IV.1). Negatywną odpowiedź na to pytanie uzasadniają wyniki obliczeń zilustrowane na rys. IV.7, gdzie zestawiono wydajności kilku systemów o identycznej łącznej pojemności $L = 2K + N$, lecz różniących się wartościami N oraz $N_1 = n_2 = K$.

Jak wynika z przedstawionych wykresów, największą wydajność osiąga się wówczas, gdy na założoną pojemność systemu składa się możliwie duży bufor oraz możliwie mała liczba wysoko wydajnych kanałów obsługi (w szczególności najlepszy jest więc system pojedynczy).

Porównując na przykład przypadek K = 4, N = 2 (dla którego przy L = 10 uzyskuje się V = 0.8655) oraz przypadek K = 2, N = 4 (dla którego przy L = 8 uzyskuje się V = 0.875) widzimy, iż większa łączna pojemność systemu nie zapewnia uzyskania większej wydajności, jeśli nie jest respektowane wyżej sformułowane zalecenie.

Na rys. IV.8 przedstawiono porównanie wydajności dwuetapowych systemów jednorodnych, posiadających jednakową pojemność systemu oraz jednakową pojemność bufora pośredniego, a różniących się jedynie liczbą kanałów obsługi na poszczególnych etapach ($n_1 \neq n_2$, $n_1 + n_2 = \text{const.}$). Jak widać, stosowanie zróżnicowanej liczby kanałów obsługi na obu etapach prowadzi do istotnego zmniejszenia wydajności systemu. Zarówno wyniki wykorzystane na tym rysunku, jak i liczne inne eksperymenty potwierdziły zachodzenie dla rozpatrywanych systemów zasady odwracalności.



Rys. IV.9. Dwa porównywane warianty

a) K systemów pojedynczych, z których każdy posiada bufor pośredni o pojemności N_a, b) system wielokanałowy, posiadający bufor pośredni o pojemności $N_B = K N_a$

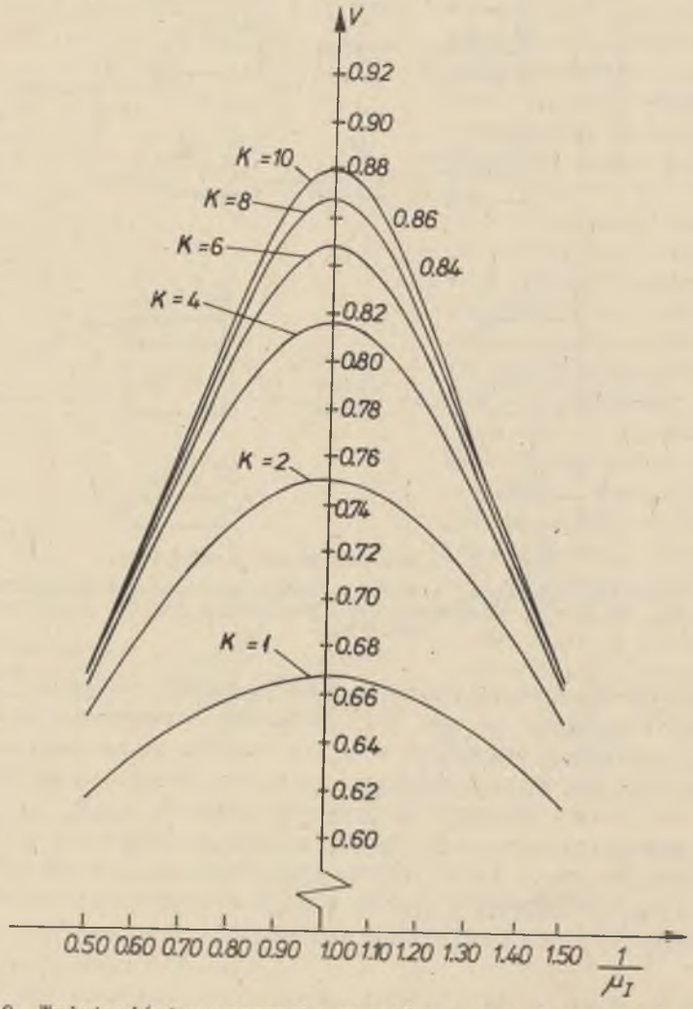
Na zakończenie dyskusji systemów zrównoważonych porównamy wydajność dwóch, przedstawionych na rys. IV.9 wariantów a mianowicie układu K dwuetapowych systemów pojedynczych oraz wielokanałowego systemu dwuetapowego. Założono, iż w obu przypadkach identyczna jest łączna liczba 2K stanowisk obsługi oraz łączna pojemność buforów pośrednich $N_b = K N_a$.

Przedstawione w tabeli IV.7 wyniki analizy porównywanych wariantów wskazują na zdecydowaną wyższość systemu wielokanałowego, tym wyraźniejszą, im większa jest liczba kanałów obsługi oraz im mniejsza jest pojemność buforów pośrednich, co jest całkowicie zgodne z wnioskami podanymi w [WILD 72].

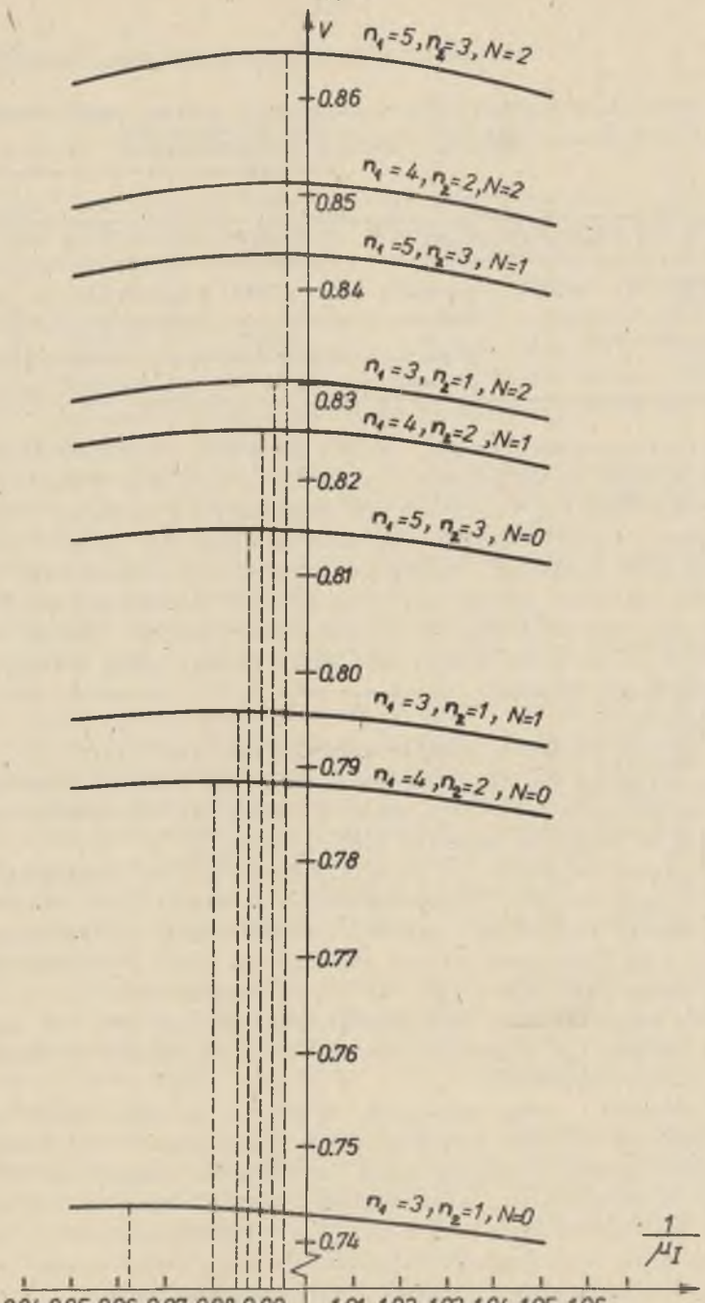
b) System niezrównoważony

Wprowadźmy dla uproszczenia oznaczenia: $\mu_I = n_1 \cdot \mu_1^o$, $\mu_{II} = n_2 \cdot \mu_2^o$, przy czym, zgodnie z wprowadzonymi wcześniej założeniami, rozpatrywać będziemy przypadki spełniające ograniczenie:

$$\frac{1}{\mu_I} + \frac{1}{\mu_{II}} = 2.$$



Rys. IV.10. Wydajność dwuetapowych, niezrównoważonych systemów obsługi o równej liczbie kanałów obsługi na każdym z etapów $n_1 = n_2 = K$, przy $N = 0$ (wg [WOLI 80A])



Rys. IV.11. Wydajność dwuetapowego wielokanałowego systemu niezrównoważonego, przy założeniu różnej liczby kanałów obsługi na poszczególnych etapach (wg [WOLI 80A])

Tabela IV.7

Porównanie wydajności wielokanałowego systemu dwuetapowego oraz zbioru systemów pojedynczych.
Dane zaczerpnięto z [WOLI 80A]

		$N_a = 0$	$N_a = 1$	$N_a = 2$
Zbiór K systemów pojedynczych		0.6667	0.7500	0.8000
Wieloetapowe, wielokanałowe	K = 2	0.7500	0.8333	0.8750
Systemy masowej obsługi	K = 3	0.7900	0.8714	0.9072
	K = 4	0.8161	0.8940	0.9256

Jak to zilustrowano na rys. IV.10 (dla N = 0) w przypadku, gdy liczba kanałów obsługi na obu etapach jest równa $n_1 = n_2 = K$, wówczas maksymalną wydajność osiąga się w systemie zrównoważonym, tj. przy $\mu_I = \mu_{II}$. Straty wynikające z niezrównoważenia systemu rosną przy tym silnie ze wzrostem K. Można łatwo sprawdzić, że reguła ta obowiązuje również dla $N > 0$.

Jeżeli natomiast zachodzi $n_1 \neq n_2$, wówczas dla optymalizacji wydajności systemu wskazane jest, aby łączna wydajność stanowisk obsługi w tym etapie, gdzie są one liczniejsze, była nieco wyższa. Jeśli więc $n_1 > n_2$, wówczas maksymalną wydajność osiąga się przy $\mu_I > 1$, natomiast dla $n_1 < n_2$ - przy $\mu_I < 1$.
Przykłady ilustrujące tę prawidłowość zawiera rys. IV.11.
Zarówno optymalny stopień niezrównoważenia, jak i zysk stąd wynikający maleją ze wzrostem pojemności bufora N, względnie pojemności systemu L, wzrastają natomiast ze wzrostem różnicy $|n_1 - n_2|$.

Można łatwo sprawdzić, iż jakkolwiek przy $n_1 \neq n_2$ optymalną wydajność uzyskuje się w systemie niezrównoważonym, to jednak jest ona niższa aniżeli wydajność uzyskiwana w systemie zrównoważonym, o liczbie kanałów obsługi $n_1 + n_2$ równo rozdzielonej pomiędzy oba etapy. Różnica ta jest tym wyraźniejsza, im mniejsza jest wartość $n_1 + n_2$ oraz N.

Liczne przeprowadzone eksperymenty wykazały przy tym, że zamiana kolejności etapów nie wpływa na wydajność systemu również w przypadku systemów niezrównoważonych.

Podsumowując, można powiedzieć, iż w celu maksymalizacji wydajności rozpatrywanych systemów wielokanałowych należy przede wszystkim dążyć do stosowania systemu o możliwie małej (i równej na każdym etapie) liczbie kanałów, zrównoważonego, zaopatrzonego w możliwie duże bufory pośrednie. Jedynie w przypadku, gdy niemożliwe jest zachowanie równości liczby kanałów obsługi na poszczególnych etapach, wynikający stąd spadek wydajności systemu można zmniejszyć stosując jego niezrównoważenie.

IV.5. Wskazania dla maksymalizacji wydajności WSO

W oparciu o omówione wyżej cząstkowe wyniki można pokusić się o sformułowanie pewnych, ogólniejszych wytycznych dla optymalizacji wydajności WSO.

Wprowadzimy, bez troski o precyzję, pojęcie efektywności pojedynczego etapu obsługi, charakteryzujące wpływ parametrów stanowiska obsługi oraz bufora zainstalowanego na danym etapie na wydajność WSO.
Przyjmiemy, że efektywność identycznych stanowisk obsługi, poprzedzonych buforami o równej pojemności, jest równa. Będziemy natomiast mówili, że efektywność ta wzrasta w wyniku dokonania pewnych modyfikacji parametrów stanowiska obsługi lub bufora zainstalowanego na danym etapie, jeśli modyfikacje te (dokonywane przy niezmiennych parametrach pozostałych etapów) prowadzą do wzrostu wydajności WSO.

W oparciu o omawiane w tym rozdziale wyniki badań WSO można stwierdzić, iż efektywność pojedynczego etapu obsługi wzrasta, gdy zwiększeniu ulega[x]:
- moc stanowiska obsługi (tzn. maleje średni czas obsługi),
- regularność obsługi (tzn. maleje współczynnik zmienności rozkładu),
- niezawodność działania stanowiska obsługi,
- pojemność bufora poprzedzającego stanowisko obsługi,
- liczba kanałów obsługi (przy zachowaniu łącznej mocy stanowiska obsługi).

Ogólnie rzecz biorąc, maksymalną wydajność WSO osiąga się przy zróżnicowaniu efektywności poszczególnych etapów tak, aby najbardziej efektywne z nich znalazły się pośrodku systemu, natomiast najmniej efektywne na jego krańcach. Takie zróżnicowanie efektywności można w zasadzie osiągnąć drogą zmian któregokolwiek z wyżej wymienionych parametrów, ponieważ ich wpływ jest jakościowo równoważny, należy jednak podkreślić, iż czułość wydajności systemu na zmiany poszczególnych parametrów jest silnie zróżnicowana, co zostało zilustrowane wykresami i tabelami zamieszczonymi w poprzednich punktach.

Wszystkie przedstawione reguły optymalizacji wydajności WSO cechowały się tym, iż możliwy do uzyskania w wyniku ich zastosowania zysk był znacznie mniejszy aniżeli straty ponoszone na skutek błędnego doboru parametrów poszczególnych etapów (aczkolwiek w wielu przypadkach zysk ten był również niebagatelny).
Biorąc pod uwagę fakt, iż zwykle niemożliwy jest bardzo precyzyjny dobór parametrów WSO (np. idealne zróżnicowanie czy też ścisłe optymalne niezrównoważenie), podstawowym, praktycznie ważnym rezultatem dyskutowanych badań są wskazania o charakterze jakościowym, pozwalające określić takie

---

[x] Przedstawione wskazania nie dotyczą WSO z regularnymi rozkładami czasów obsługi na wszystkich etapach, gdzie obowiązują zasady omówione w punkcie IV.1

parametry WSO, aby wspomniane, nieuniknione niedokładności nie prowadziły do istotnych strat.

Jest to możliwe nawet przy stosunkowo zgrubnej znajomości charakterystyki obiektu, np. wskazano, iż dla określenia sposobu nierównoważenia systemu nie jest konieczna znajomość rozkładów czasów obsługi, a wystarczające jest określenie dwóch momentów tych rozkładów (o ile $C_b < 1$).

W niniejszej pracy odwoływano się głównie do wyników ścisłych oraz nielicznych, przeprowadzonych bardzo starannie eksperymentów symulacyjnych. Ze względu na niewielkie, na ogół, różnice wydajności występujące w różnych, rozpatrywanych wariantach, szersze wykorzystanie modelowania cyfrowego do wypracowywania reguł optymalizacji jest bardzo pracochłonne i wymaga szczególnie precyzyjnego określania wiarygodności uzyskanych wyników; w oparciu o modelowanie cyfrowe wysuwano już niejednokrotnie błędne wskazania.

Przedstawione rezultaty są dość kompletne, dalszych badań wymagają jednak reguły doboru parametrów w systemach, w których występują stanowiska obsługi cechujące się większym od jedności współczynnikiem zmienności rozkładu czasu trwania obsługi oraz w systemach wielokanałowych, szczególnie w przypadku, gdy rozkłady czasów obsługi są różne od wykładniczego (jeśli chodzi o pojedyncze systemy nasycone, w których czasy obsługi zgłoszeń na obu etapach mają rozkład Coxa, to podstawą do wypracowania reguł optymalizacji mogą być metody numeryczne wykorzystujące wyniki prac [TRUS 75], [GROM 79]).

Warto zauważyć, w oparciu o analizę przeprowadzoną w dodatku A do niniejszego rozdziału, że reguły równoważenia systemów dwuetapowych bez bufora pośredniego, omówione w punkcie IV.2, pozostają ważne również w przypadku, gdy czas obsługi zgłoszenia na pierwszym etapie zależy od czasu oczekiwania zgłoszenia w kolejce.

Zależność wydajności systemu od pojemności bufora pośredniego, dyskutowaną w punkcie IV.3, a w szczególności specyficzny, silnie nieliniowy charakter przyrostu wydajności w funkcji jednostkowego powiększenia pojemności bufora, zilustrowaną w tabeli IV.b, można też zaobserwować w przypadku bardziej złożonych systemów [PUJO 81A].

W niniejszym rozdziale skupiono uwagę na jednym kryterium optymalizacji WSO, którym była maksymalizacja wydajności systemu. Obok tego podstawowego zadania rozpatruje się też inne kryteria optymalizacji, przy czym najczęściej rozważa się (zarówno w odniesieniu do systemów właściwych, jak i nasyconych) problem minimalizacji wartości oczekiwanych liczby zgłoszeń przebywających w systemie oraz łącznego czasu pobytu zgłoszenia w systemie.

Można łatwo sprawdzić, jak to pokazano np. w [RAYA 79B], [SMIT 77], że zmiana kryterium optymalizacji prowadzi do zasadniczo różnych rozwiązań, różna jest też wrażliwość wspomnianych wskaźników na modyfikacje wartości poszczególnych parametrów systemu.

Dla przykładu, w pewnym pojedynczym, nasyconym, systemie trzyetapowym, o łącznej pojemności buforów pośrednich $\mathcal{N} = 8$, rozpatrywanym w [RAYA 79B], zmiana pojemności poszczególnych buforów z $N_2 = 3$, $N_3 = 5$ na $N_2 = 5$, $N_3 = 3$, powoduje wzrost wartości oczekiwanej liczby zgłoszeń przebywających w systemie z 6.3 do 7.6, podczas gdy wydajność obu tych systemów jest, w myśl zasady odwracalności, identyczna. W pracy tej (a także w kilku innych) sformułowano sugestię, iż jeśli w WSO, którego parametry dobrano pod kątem optymalizacji wydajności, rozdział pojemności zostanie zmodyfikowany tak, by pojemność kolejnych buforów monotonicznie wzrastała, wówczas wyraźnemu zmniejszeniu ulegnie wartość oczekiwana liczby zgłoszeń przebywających w systemie przy równoczesnym, stosunkowo niewielkim, spadku wydajności.

Problemy optymalizacji dwuetapowych systemów obsługi, z punktu widzenia minimalizacji czasu pobytu zgłoszeń w systemie, rozpatrywano na drodze analitycznej w [TEMB 74] oraz [BOXM 78B] w odniesieniu do systemów z nieograniczonym buforem pośrednim, natomiast w [KAWA 75] w odniesieniu do systemów z buforem o ograniczonej pojemności. Dalsze prace z tego zakresu, wykorzystujące modelowanie cyfrowe, dotyczyły WSO z nieograniczonymi buforami pośrednimi, przy $M \geq 2$ [PAYN 72], [KALA 73], [MCGE 76], [SHIM 78].

W literaturze można też znaleźć próby optymalizacji parametrów WSO, a szczególnie pojemności buforów pośrednich [ANDE 69], [BASU 77], [KNOT 78], z punktu widzenia bardziej złożonych kryteriów, obejmujących takie elementy, jak:

- koszt utrzymania bezczynnego stanowiska obsługi,
- koszt pojedynczego miejsca w buforze,
- koszt przetrzymywania zgłoszeń w buforze,
- zysk związany z obsłużeniem pojedynczego zgłoszenia.

Bardziej szczegółowe omówienie tych zagadnień wykracza poza ramy niniejszej pracy.

We wszystkich przedstawionych w tym rozdziale rozważaniach ograniczaliśmy się do przypadku, gdy reakcją na przepełnienie bufora jest blokowanie.

Zmiana sposobu tej reakcji wpływa, ogólnie rzecz biorąc, zarówno na wartość wydajności rozpatrywanego systemu, jak też jej wrażliwość na modyfikacje poszczególnych parametrów.

Naturalne jest pytanie, jak istotny jest to wpływ, na które odpowiedzieć można jedynie bardzo wyrywkowo. Na przykład w odniesieniu do pojedynczych WSO z wykładniczymi rozkładami czasów obsługi wykazano w [CASE 79], że blokowanie jest równoważne blokowaniu prewencyjnemu, przy zwiększonej o jedno miejsce pojemności bufora i obie te reguły reakcji na przepełnienie są, w szczególnym przypadku systemów dwuetapowych, równoważne lokalnemu powtarzaniu obsługi. Co więcej, charakterystyki systemu z blokowaniem są zbliżone do charakterystyk systemu z lokalnym powtarzaniem obsługi nawet przy $M > 2$.

Uwagi te nie upoważniają, rzecz jasna, do bezpośredniego przenoszenia uzyskanych wyników na inne reguły reakcji na przepełnienie, pozwalając się jedynie spodziewać dość daleko idących podobieństw w zakresie wymienionych trzech przypadków, podczas gdy wyniki uzyskiwane przy stratach zgłoszeń bądź też rezygnacji z obsługi są silnie zróżnicowane (inne są np. reguły równoważenia systemu).

Wielokrotnie już wcześniej wspomniano o trudnościach w określeniu na drodze analitycznej wydajności WSO z blokowaniem. Podobne problemy występują również w odniesieniu do innych reguł reakcji na przepełnienie oraz innych parametrów systemów (np. czasu pobytu zgłoszenia we właściwych WSO), co stanowiło zachętę do rozwijania metod przybliżonych. Nie omawiając bliżej uzyskanych wyników, wymienimy tylko ciekawsze prace z tego zakresu [HARR 73], [PUJO 76], [SHIM 79] oraz niektóre prace ukierunkowane na modelowanie przesyłu informacji w sieciach komputerowych [CHAN 75A], [IRLA 77], [PUJO 76], [PUJO 79B], [PUJO 80C], rozpatrywanych jako WSO.

W przedstawionych rozważaniach pominęliśmy całkowicie problematykę synchronizowanych WSO. Wyczerpujący (aczkolwiek nie obejmujący najnowszych wyników) przegląd zagadnień optymalizacji wydajności w takich systemach zawierają prace [BUXE 73] oraz [REEV 73].

Zagadnienie optymalizacji parametrów WSO, mimo stosunkowo dużego zainteresowania nie jest więc jeszcze jak widać problemem zamkniętym. Bardziej złożone struktury sieci stanowisk obsługi doczekały się jak dotychczas jedynie nielicznych wyników w zakresie ich optymalizacji i można się spodziewać, że problematyka ta będzie w przyszłości przedmiotem intensywnych badań.

IV.6. Dodatek - analiza wybranych systemów dwuetapowych

W niniejszym dodatku przedstawimy analityczne rozwiązania pozwalające na wyznaczenie wydajności (oraz niektórych innych charakterystyk) trzech różnych dwuetapowych systemów masowej obsługi.

a) Dwuetapowy, pojedynczy system masowej obsługi bez bufora pośredniego

Rozpatrzmy właściwy, pojedynczy system masowej obsługi o następujących własnościach:

- zgłoszenia napływają strumieniem Poissona o intensywności λ,
- czas obsługi zgłoszeń na jednokanałowym stanowisku obsługi pierwszego etapu x_1 jest zależny od czasu oczekiwania na rozpoczęcie obsługi w sposób określony dystrybuantą $Y(x/w)$,

$$Y(x/w) = \begin{cases} Y_0(x) & w = 0 \\ Y_i(x) & w_{i-1} < w \leq w_i \quad i=1,2,\ldots,J \end{cases} \qquad \text{(IV.6.1)}$$

gdzie:

$$0 = w_0 < w_1 < \ldots < w_{J-1} < w_J = \infty,$$

- zgłoszenia obsługiwane są wg regulaminu naturalnego,
- bufor pośredni ma pojemność równą zeru,
- czas obsługi zgłoszeń na jednokanałowym stanowisku obsługi x_2 drugiego etapu określa dystrybuanta $H(x)$,
- zgłoszenia obsłużone na stanowisku obsługi pierwszego etapu blokują je do momentu zwolnienia stanowiska obsługi drugiego etapu.

Tak więc rozpatrujemy system masowej obsługi złożony z podstawowego systemu WAD (analizowanego w rozdziale II) oraz jednokanałowego, konwencjonalnego stanowiska obsługi.

Zwróćmy uwagę na pewną, charakterystyczną prawidłowość, występującą w działaniu takiego systemu. Otóż, jeśli czas oczekiwania w^1 zgłoszenia ζ^1 na rozpoczęcie obsługi był większy od zera (załóżmy, że $w_i < w^1 \leq w_{i+1}$, $i>0$), wówczas rozpoczęcie obsługi zgłoszenia ζ^1 na stanowisku obsługi x_1 zbiega się z rozpoczęciem obsługi zgłoszenia ζ^{1-1} na stanowisku obsługi x_2.

Tak więc zmienną losową φ_i, opisującą czas, przez który zgłoszenie ζ^1 zajmuje stanowisko obsługi x_1 (tj. łączny czas obsługi i blokady), pod warunkiem, że $w_i < w^1 \leq w_{i+1}$, $i>0$, można wyznaczyć jako:

$$\varphi_i = \text{Max}\left\{y_i,\ h\right\}. \qquad \text{(IV.6.2)}$$

Korzystając ze znanych relacji określających funkcję gęstości rozkładu maksimum dwóch zmiennych losowych uzyskujemy:

$$\varphi_i(x) = y_i(x)\,H(x) + Y_i(x)h(x), \quad i=1,2,\ldots,J. \qquad \text{(IV.6.3)}$$

Bardziej złożony jest sposób wyznaczania dystrybuanty zmiennej losowej φ_0 opisującej czas, przez który zgłoszenie ζ^1 zajmuje stanowisko obsługi x_1 w przypadku, gdy $w^1 = w_0 = 0$.

W tym przypadku bowiem równocześnie z rozpoczęciem obsługi zgłoszenia ζ^{1-1} na stanowisku x_2 rozpoczyna się bezczynność stanowiska x_1, trwająca przez czas określony zmienną losową η, a więc:

$$\varphi_0 = y_0 + \text{Max}\left\{h - \eta - y_0,\ 0\right\}. \qquad \text{(IV.6.4)}$$

Ze względu na przyjętą postać strumienia zgłoszeń zmienna losowa η posiada dystrybuantę $\Gamma(x)$ postaci:

$$\Gamma(x) = 1 - e^{-\lambda t}. \qquad \text{(IV.6.5)}$$

Funkcje gęstości rozkładu zmiennej losowej φ_0 można więc zapisać w postaci:

$$\varphi_0(x) = \begin{cases} e^{-\lambda(h-x)}\left[\lambda Y_0(x) + y_0(x)\right] & \text{jeśli} \quad 0 \leq x < h \\ y_0(x) & \text{jeśli} \quad h \leq x < \infty \end{cases} \qquad \text{(IV.6.6)}$$

a uwzględniając rozkład zmiennej losowej h uzyskujemy ostatecznie:

$$\varphi_0(x) = y_0(x)H(x) + e^{\lambda x}\left[\lambda Y_0(x) + y_0(x)\right]\int_x^\infty e^{-\lambda t}h(t)dt \qquad (IV.6.7)$$

Tryb pracy stanowiska obsługi X_1 w rozpatrywanym systemie dwuetapowym jest więc identyczny jak tryb pracy systemu podstawowego WAD ze zgłoszeniami napływającymi strumieniem Poissona i zależnością czasu obsługi od czasu oczekiwania, określoną wzorem II.6, przy czym:

$$B_i(x) = \Phi_i(x) = \int_0^x \varphi_i(t)dt \quad i=0,1,\ldots,J, \qquad (IV.6.8)$$

gdzie: $\varphi_i(t)$ określone są odpowiednio zależnościami (IV.6.7) oraz (IV.6.3).

Wykorzystując zależności podane w rozdziale II można wyznaczyć wszystkie interesujące charakterystyki systemu.

W szczególności można wprost podać warunek stabilności, przybierający postać:

$$\lambda E(\varphi_J) < 1. \qquad (IV.6.9)$$

Przedstawione wyżej rozumowanie wzorowane jest na rozważaniach przedstawionych w pracy [AVII 65A], w której ograniczono się jednak do przypadku, gdy również stanowisko obsługi X_1 miało charakter konwencjonalny (tj. przypadku, gdy $Y(x/w) \equiv Y_0(x)$), co prowadziło w konsekwencji do analizy jednokanałowych systemów obsługi, z różnym od pozostałych rozkładem czasu pierwszej obsługi w każdym okresie czynnym[x]. Przedstawione tam wyniki stanowią szczególny przypadek rezultatów możliwych do uzyskania drogą wykorzystania, w omówiony wyżej sposób, rozwiązań wyprowadzonych w rozdziale II.

Uogólniając rozwiązane w [AVII 65A] przykłady można wprost podać postać funkcji gęstości rozkładów $\varphi_i(x)$ oraz warunki stabilności dla niektórych przypadków specjalnych:

i) Jeśli

$$Y(x/w) = \begin{cases} \mathbb{H}(x-B_s) & \text{dla} \quad w \leq D \\ \mathbb{H}(x-B_L) & \text{dla} \quad w > D \end{cases} \qquad (IV.6.10)$$

[x] System obsługi o identycznej jak w [AVII 65A] charakterystyce sposobu obsługi zgłoszeń napływających strumieniem Erlanga (a więc ogólniejszy przypadek) przeanalizowano w [CHAN 72A].

oraz:

$$H(x) = \mathbb{H}(x-\alpha), \quad \alpha>0, \qquad (IV.6.11)$$

wówczas:

$$\varphi_0(x) = \begin{cases} \delta(x-B_s) & \text{jeśli} \quad B_s>\alpha \\ e^{-\lambda(\alpha-B_s)}\delta(x-B_s)+\lambda e^{-\lambda(\alpha-x)}\left[\mathbb{H}(x-B_s)-\mathbb{H}(x-\alpha)\right] & \text{jeśli} \quad B_s<\alpha \end{cases} \qquad (IV.6.12)$$

gdzie $\delta(t)$ jest pseudofunkcją Diraca

$$\Phi_1(x) = \mathbb{H}\left[x - \mathrm{Max}(B_s, \alpha)\right]$$
$$\Phi_2(x) = \mathbb{H}\left[x - \mathrm{Max}(B_L, \alpha)\right] \qquad (IV.6.13)$$

natomiast warunek stabilności ma postać:

$$\lambda\,\mathrm{Max}\left[B_L, \alpha\right] < 1. \qquad (IV.6.14)$$

ii) Jeśli

$$Y(x/w) = \begin{cases} 1 - \mu_1 e^{-\mu_1 x} & \text{dla} \quad w \leq D, \\ 1 - \mu_2 e^{-\mu_2 x} & \text{dla} \quad w > D, \end{cases} \qquad (IV.6.15)$$

oraz:

$$H(x) = 1 - \mu_B e^{-\mu_B x}, \qquad (IV.6.16)$$

wówczas:

$$\left.\begin{aligned} \varphi_0(x) &= \mu_1 e^{-\mu_1 x} + \frac{\lambda\mu_B}{\lambda+\mu_B} e^{-\mu_B x} - \frac{\lambda(\mu_1+\mu_B)}{\lambda+\mu_B} e^{-(\mu_1+\mu_B)x}, \\ \varphi_1(x) &= \mu_1 e^{-\mu_1 x} + \mu_B e^{-\mu_B x} - (\mu_1+\mu_B) e^{-(\mu_1+\mu_B)x}, \\ \varphi_2(x) &= \mu_2 e^{-\mu_2 x} + \mu_B e^{-\mu_B x} - (\mu_2+\mu_B) e^{-(\mu_2+\mu_B)x}, \end{aligned}\right\} \qquad (IV.6.17)$$

natomiast warunek stabilności ma postać:

$$\lambda\left[\frac{1}{\mu_2}+\frac{1}{\mu_B}-\frac{1}{\mu_2+\mu_B}\right]<1. \qquad \text{(IV.6.18)}$$

b) Dwuetapowy, nasycony, pojedynczy system masowej obsługi

Rozpatrzmy dwuetapowy, nasycony, pojedynczy system masowej obsługi o następujących własnościach:
- czas obsługi zgłoszeń na stanowisku obsługi pierwszego etapu x_1 jest zmienną losową o rozkładzie wykładniczym z parametrem μ_1,
- czas obsługi zgłoszeń na stanowisku obsługi drugiego etapu x_2 jest zmienną losową o dowolnym rozkładzie, scharakteryzowanym dystrybuantą $B_2(x)$ posiadającą skończoną wartość oczekiwaną $E(b_2)$,
- bufor pośredni ma ograniczoną pojemność N,
- w przypadku przepełnienia bufora pośredniego następuje zablokowanie stanowiska x_1,
- stanowisko obsługi x_2 nigdy nie ulega zablokowaniu.

Niech $t_0,t_1,\ldots,t_j,\ldots$ oznaczają odpowiednio chwile zakończenia obsługi zgłoszeń $\zeta^0,\zeta^1,\ldots,\zeta^j,\ldots$ na stanowisku obsługi x_2, natomiast $\mathcal{M}(j)$ niech oznacza liczbę zgłoszeń całkowicie obsłużonych na stanowisku obsługi x_1, znajdujących się w systemie w chwili (t_j+0); $\mathcal{M}(j)=0,1,\ldots,N+1$. Oczywiście $\mathcal{M}(j)=N+1$ oznacza, iż w chwili t_j stanowisko obsługi x_1 było zablokowane.

Wartości

$$p_m(j)=\Pr\left\{\mathcal{M}(j)=m\right\}$$

spełniają układ równań:

$$\begin{cases} p_m(j+1)=p_0(j)f_m+\sum_{i=1}^{m+1}p_i(j)f_{m-i+1}, & m=0,1,\ldots,N \\ \sum_{m=0}^{N+1}p_m(j+1)=1, \end{cases} \qquad \text{(IV.6.19)}$$

gdzie:

$$f_i=\frac{1}{i!}\int_0^\infty(\mu_1x)^ie^{-\mu_1x}b_2(x)dx, \qquad \text{(IV.6.20)}$$

jest prawdopodobieństwem obsłużenia "i" zgłoszeń na stanowisku x_1 podczas obsługi jednego zgłoszenia na stanowisku obsługi x_2.

Proces stochastyczny $\{\mathcal{M}(j)\}$; $j=0,1,\ldots$ jest łańcuchem Markowa z dyskretnym czasem i skończoną liczbą stanów, ergodycznym [COHE 69], tak więc istnieją

$$p_m=\lim_{j\to\infty}p_m(j), \quad m=0,1,\ldots,N+1; \qquad \text{(IV.6.21)}$$

spełniające układ równań

$$\left.\begin{aligned} p_m&=p_0f_m+\sum_{i=1}^{m+1}p_if_{m-i+1} \\ \sum_{m=0}^{N+1}p_m&=1. \end{aligned}\right\} \qquad \text{(IV.6.22)}$$

Dla określenia wydajności V rozpatrywanego systemu wystarczające jest wyznaczenie z układu równań (IV.6.22) wartości p_0.

Niech bowiem zmienna losowa φ reprezentuje odstęp czasu pomiędzy chwilami zakończenia obsługi dwóch kolejnych zgłoszeń. Funkcję gęstości rozkładu tej zmiennej losowej można określić jako:

$$\varphi(x)=(1-p_0)b_2(x)+p_0\left[b_2(x)*\mu_1e^{-\mu_1x}\right], \qquad \text{(IV.6.23)}$$

gdzie $*$ oznacza operację splotu, a stąd

$$E(\varphi)=E(b_2)+\frac{p_0}{\mu_1}. \qquad \text{(IV.6.24)}$$

Wprost z definicji wydajności WSO otrzymujemy więc:

$$V=\left[E(b_2)+\frac{p_0}{\mu_1}\right]^{-1}. \qquad \text{(IV.6.25)}$$

Zauważmy, że układ równań (IV.6.19), a więc oczywiście także (IV.6.22), jest identyczny z układem równań opisujących jednokanałowy system obsługi z buforem o pojemności (N+1) miejsc i stratami zgłoszeń (system M/G/1/N+2) i może być rozwiązany przy użyciu metody przedstawionej w [TRUS 75].

Wprowadźmy funkcję tworzącą

$$f(z) = \sum_{m=0}^{\infty} z^m f_m. \qquad \text{(IV.6.26)}$$

Z równań (IV.6.20) oraz (IV.6.26) wynika:

$$f(z) = \bar{b}_2\left[\lambda(1-z)\right]. \qquad \text{(IV.6.27)}$$

Równość (IV.6.26) można sprowadzić do postaci:

$$\left(\sum_{m=1}^{N+1} z^m p_m\right) f(z) = \sum_{m=0}^{N} z^{m+1} \sum_{i=1}^{m+1} p_i f_{m-i+1} + z^{N+2} R_o(z), \qquad \text{(IV.6.28)}$$

gdzie: $R_o(z)$ jest pewnym szeregiem potęgowym ze względu na z.

Po prostych przekształceniach uzyskuje się:

$$f(z) \sum_{m=1}^{N+1} z^m p_m = z\left\{\sum_{m=0}^{N+1} z^m p_m - p_o f(z)\right\} + z^{N+2} R_1(z), \qquad \text{(IV.6.29)}$$

oraz:

$$\sum_{m=1}^{N+1} z^m p_m = \frac{f(z)(1-z)}{f(z) - z}\, p_o + z^{N+2} R_2(z), \qquad \text{(IV.6.30)}$$

gdzie $R_1(z)$, $R_2(z)$ są szeregami potęgowymi ze względu na z.

Przy rozwinięciu prawej strony równości (IV.6.30) w szereg potęgowy wyrazy związane z potęgami z rzędu 1,2,...,N+1 uzyskiwane będą wyłącznie z pierwszego składnika.

Oznaczając

$$\pi_s(z) = \frac{f(z)(1-z)}{f(z) - z}, \qquad \text{(IV.6.31)}$$

można więc wartości p_m wyznaczyć jako:

$$p_m = \frac{1}{m!}\left[\pi_s(z)\right]^{(m)}\Big|_{z=0} p_o, \qquad \text{(IV.6.32)}$$

i korzystając z równania normalizacji uzyskujemy:

$$p_o \sum_{m=0}^{N+1} \frac{1}{m!}\left[\pi_s(z)\right]^{(m)}\Big|_{z=0} = 1. \qquad \text{(IV.6.33)}$$

Ostatecznie więc

$$p_o = \left\{\sum_{m=1}^{N+1} \frac{1}{m!}\left[\pi_s(z)\right]^{(m)} + 1\right\}^{-1}. \qquad \text{(IV.6.34)}$$

Wstawiając (IV.6.34) do (IV.6.25) uzyskujemy bezpośrednio wydajność systemu, co ilustruje tabela IV.8.

Tabela IV.8

Metoda wyznaczania wydajności
dwuetapowego, pojedynczego, nasyconego systemu masowej obsługi
z wykładniczym rozkładem czasu obsługi o parametrze μ_1
na jednym ze stanowisk
i dowolnym rozkładem czasu obsługi, o dystrybuancie $B_2(x)$
na drugim stanowisku

$$f(z) = \bar{b}_2\left[\mu_1(1-z)\right] \quad \text{gdzie} \quad \bar{b}_2(s) = \int_0^{\infty} e^{-st}\, dB_2(t)$$

$$\pi_s(z) = \frac{f(z)(1-z)}{f(z) - z}$$

$$p_o = \left\{\sum_{m=1}^{N+1} \frac{1}{m!}\left[\pi_s(z)\right]^{(m)}\Big|_{z=0} + 1\right\}^{-1}$$

$$T = \left[E(b_2) + \frac{p_o}{\mu_1}\right]^{-1}$$

Poniżej wyznaczymy w omówiony sposób wydajność systemu dla kilku różnych postaci dystrybuanty $B_2(x)$.

i) rozkład regularny, dla którego:

$$B_2(x) = \mathbb{1}(x-\alpha), \quad \alpha > 0, \qquad E(b_2) = \alpha, \qquad C_B = 0. \qquad \text{(IV.6.35)}$$

W tym przypadku uzyskujemy:

$$\left.\begin{aligned} &\bar{b}_2(s) = e^{-s\alpha}, \quad f(z) = e^{-\mu_1\alpha(1-z)}, \\ &\pi_s(z) = \frac{1-z}{1 - z\, e^{-\mu_1\alpha(1-z)}}. \end{aligned}\right\} \qquad \text{(IV.6.36)}$$

Można łatwo wykazać (patrz np. [GROS 74]), że:

$$\frac{1}{m!}\left[\pi_s(z)\right]^{(m)}\Big|_{z=0} = \begin{cases} e^{\rho}-1 & \text{dla } m=1 \\ \sum_{k=1}^{m}(-1)^{m-k}e^{k\rho}\left[\frac{(k\rho)^{m-k}}{(m-k)!} + (1-\delta_{mk})\frac{(k\rho)^{m-k-1}}{(m-k-1)!}\right] \end{cases} \qquad \text{(IV.6.37)}$$

gdzie:

$$\rho = \mu_1\alpha$$

$$\delta_{mk} = \begin{cases} 0 & \text{dla } m \neq k \\ 1 & \text{dla } m = k. \end{cases}$$

co pozwala na bezpośrednie określenie p_0, a następnie V.
Uzyskane wyniki są zgodne z rezultatami otrzymanymi w [RAON 75B].
ii) rozkład wykładniczy, dla którego:

$$\left.\begin{aligned} &B_2(x) = 1 - e^{-\mu_2 x}, \\ &E(b_2) = \mu_2^{-1} \\ &C_B = 1. \end{aligned}\right\} \qquad \text{(IV.6.38)}$$

W tym przypadku

$$\left.\begin{aligned} &\bar{b}_2(s) = \frac{\mu_2}{s+\mu_2}, \quad f(z) = \frac{\mu_2}{\mu_1(1-z)+\mu_2}, \\ &\pi_s(z) = \frac{1}{1-z\rho}, \quad \rho = \frac{\mu_1}{\mu_2}. \end{aligned}\right\} \qquad \text{(IV.6.39)}$$

Jednakże przy $z\,\frac{\mu_1}{\mu_2} < 1$ zachodzi:

$$\pi_s(z) = \sum_{i=0}^{\infty}(\rho z)^i, \qquad \text{(IV.6.40)}$$

a więc:

$$\frac{1}{m!}\left[\pi_s(z)\right]^{(m)}\Big|_{z=0} = \rho^m. \qquad \text{(IV.6.41)}$$

Korzystając z (IV.6.41) (IV.6.34) oraz (IV.6.25) uzyskujemy po prostych przekształceniach:

$$V = \mu_1\mu_2\frac{\mu_1^{N+2} - \mu_2^{N+2}}{\mu_1^{N+3} - \mu_2^{N+3}}, \qquad \text{(IV.6.42)}$$

co jest identyczne ze znanym wynikiem pracy [HUNT 56].
iii) rozkład hiperwykładniczy, dla którego:

$$\left.\begin{aligned} &b_2(x) = a_1\mu_A e^{-\mu_A x} + a_2\mu_B e^{-\mu_B x} \\ &\text{przy} \quad a_1 = \alpha, \quad \mu_A = 2\alpha\mu_2, \\ &\qquad a_2 = 1-\alpha, \quad \mu_B = 2(1-\alpha)\mu_2, \\ &\qquad 0{,}5 \leq a \leq 1, \quad \mu_2 > 0; \\ &E(b_2) = \mu_2^{-1} \\ &C_B = \left[2\alpha(1-\alpha)\right]^{-1} - 1. \end{aligned}\right\} \qquad \text{(IV.6.43)}$$

W tym przypadku

$$\left.\begin{aligned} &\bar{b}_2(s) = \frac{a_1\mu_A}{s+\mu_A} + \frac{a_2\mu_B}{s+\mu_B} \\ &f(z) = \frac{a_1\mu_A}{\lambda(1-z)+\mu_A} + \frac{a_2\mu_B}{\lambda(1-z)+\mu_B} \end{aligned}\right\} \qquad \text{(IV.6.44)}$$

natomiast $\pi_s(z)$ można przekształcić do postaci:

$$\pi_s(z) = \frac{A_1 - A_2 z}{A_1 - A_3 z + z^2}, \qquad \text{(IV.6.45)}$$

gdzie:

$$A_1 = \left\{ \left[a_1\mu_A + a_2\mu_B \right]\mu_1 + \mu_A\mu_B \right\} / \mu_1^2,$$
$$A_2 = \left[a_1\mu_A + a_2\mu_B \right] / \mu_1,$$
$$A_3 = \left[\mu_1 + \mu_A + \mu_B \right] / \mu_1.$$

Rozpatrzmy pierwiastki mianownika prawej strony (IV.6.45), tj. rozwiązania równania

$$z^2 - \frac{1}{\mu_1}\left[\mu_1+\mu_A+\mu_B\right] z + \frac{1}{\mu_1^2}\left\{ \left[a_1\mu_A+a_2\mu_B \right]\mu_1 + \mu_A\mu_B \right\} = 0. \qquad \text{(IV.6.46)}$$

Uwzględniając wynikające z postaci dystrybuanty równości:

$$\mu_A + \mu_B = 2\mu, \quad \mu_A\mu_B = 4\alpha(1-\alpha)\mu_2^2,$$
$$a_1\mu_A + a_2\mu_B = 2\mu_2\left[2\alpha(\alpha-1) + 1\right], \qquad \text{(IV.6.47)}$$

oraz wprowadzając oznaczenie $\rho = \mu_1/\mu_2$; $\rho > 0$, można sprowadzić (IV.6.46) do postaci:

$$\rho^2 z^2 - (\rho^2+2\rho) z + 2\rho\left[2\alpha^2-2\alpha+1\right] + 4\alpha(1-\alpha) = 0. \qquad \text{(IV.6.48)}$$

Warunkiem koniecznym i wystarczającym istnienia dwóch rzeczywistych pierwiastków tego równania jest spełnienie nierówności:

$$\Delta = (\rho^2+2\rho)^2 - 4\rho^2\left\{ 2\rho\left[2\alpha^2-2\alpha+1\right] + 4\alpha(1-\alpha) \right\} \geq 0. \qquad \text{(IV.6.49)}$$

Wprowadzając

$$d = 4\alpha(1-\alpha) \qquad \text{(IV.6.50)}$$

i dzieląc obie strony (IV.6.49) przez ρ^2 uzyskujemy:

$$\rho^2 + 4(d-1)\rho - 4(d-1) \geq 0. \qquad \text{(IV.6.51)}$$

Jak łatwo sprawdzić, parametr d związany jest ze współczynnikiem zmienności rozkładu C_B zależnością:

$$d = \frac{2}{C_B^2 + 1} \qquad \text{(IV.6.52)}$$

a ponieważ dla rozkładu hiperwykładniczego zachodzi $C_B \in [1, \infty)$, więc odpowiednio $d \in (0, 1]$.
Jednakże

$$\Delta^* = \left[4(d-1)\right]^2 + 16(d-1) = 16d(d-1) \leq 0 \quad \text{dla} \quad 0<d\leq 1, \qquad \text{(IV.6.53)}$$

a więc lewa strona nierówności (II.6.51) jest zawsze dodatnia.

Tak więc wykazaliśmy, że równanie (IV.6.46) posiada dwa pierwiastki rzeczywiste. Obecnie sprawdzimy jeszcze, że są one dodatnie.
W tym celu rozpatrzymy nierówność:

$$\rho^2 + 2\rho > +\sqrt{\Delta} \qquad \text{(IV.6.54)}$$

Wstawiając (II.6.49) do (II.6.54) uzyskujemy, po podniesieniu do kwadratu oraz uporządkowaniu

$$(1+\rho) > (1-\rho)(1-d), \qquad \text{(IV.6.55)}$$

co jest istotnie prawdziwe dla wszystkich $\rho > 0$.

Oznaczmy odpowiednio przez z_1, z_2 oba dodatnie pierwiastki równania (IV.6.46), których istnienie zostało wyżej dowiedzione:

$$z_1 = \frac{A_3 + \sqrt{A_3^2 - 4A_1}}{2}, \quad z_2 = \frac{A_3 - \sqrt{A_3^2 - 4A_1}}{2},$$

co pozwala sprowadzić (IV.6.45) do postaci:

$$\pi_s(z) = \frac{\frac{L_1}{z_1}}{1 - \frac{z}{z_1}} + \frac{\frac{L_2}{z_2}}{1 - \frac{z}{z_2}} = \sum_{n=0}^{\infty}\left\{ \frac{L_1}{z_1}\left(\frac{1}{z_1}\right)^n + \frac{L_2}{z_2}\left(\frac{1}{z_2}\right)^n \right\} z^n, \qquad \text{(IV.6.56)}$$

gdzie:

$$L_1 = \frac{A_1 - A_2 z_1}{z_2 - z_1}, \quad L_2 = \frac{A_1 - A_2 z_2}{z_1 - z_2}.$$

Oczywiście wobec faktu, że $z_1, z_2 > 0$, szeregi potęgowe w (IV.6.56) są zawsze zbieżne dla $z \to 0$, co implikuje $z/z_1 < 1$, $z/z_2 < 1$, który to przypadek jest istotny dla dalszych rozważań.

Wprost z (IV.6.56) wynika, że:

$$\frac{1}{n!}\left[\pi_s(z)\right]^{(n)}\Big|_{z=0} = \frac{L_1}{z_1}\left(\frac{1}{z_1}\right)^n + \frac{L_2}{z_2}\left(\frac{1}{z_2}\right)^n, \qquad \text{(IV.6.57)}$$

skąd po uwzględnieniu równości

$$\frac{L_1}{z_1} + \frac{L_2}{z_2} = 1$$

uzyskuje się:

$$P_o = \left[L_1 \frac{(\frac{1}{z_1})^{N+2} - 1}{1 - z_1} + L_2 \frac{(\frac{1}{z_2})^{N+2} - 1}{1 - z_2} \right]^{-1}, \qquad \text{(IV.6.58)}$$

co pozwala na bezpośrednie wyznaczenie wydajności systemu.
Można łatwo sprawdzić, że przy $\alpha = 1$ uzyskane wyniki są identyczne z otrzymanymi wyżej dla rozkładu wykładniczego.

c) Dwuetapowy, wielokanałowy, nasycony, system masowej obsługi.

Rozpatrzmy dwuetapowy, wielokanałowy, jednorodny system masowej obsługi o następujących własnościach:
- czas obsługi zgłoszeń w każdym z $n_1 = I$ kanałów obsługi pierwszego etapu jest zmienną losową o rozkładzie wykładniczym z parametrem μ_1,
- czas obsługi zgłoszeń w każdym z $n_2 = K$ kanałów obsługi drugiego etapu jest zmienną losową o rozkładzie wykładniczym z parametrem μ_2,
- bufor pośredni ma ograniczoną pojemność N,
- w przypadku przepełnienia bufora pośredniego następuje zablokowanie odpowiedniego kanału obsługi pierwszego etapu,
- żaden z kanałów obsługi wchodzących w skład drugiego etapu nie może zostać zablokowany.

Wprowadźmy zmienną losową $\nu(t)$, przyjmującą wartości całkowito-liczbowe, nieujemne, określoną jako liczba zgłoszeń, które znajdują się w systemie i ukończyły całkowicie obsługę w jednym z kanałów obsługi pierwszego etapu.
Zgłoszenia te mogą więc blokować poszczególne kanały obsługi pierwszego etapu, oczekiwać w buforze pośrednim bądź też być obsługiwane w jednym z kanałów obsługi drugiego etapu.
Można łatwo zauważyć, że rozkład zmiennej losowej $\nu(t)$ określony jest przez proces stochastyczny urodzin i śmierci z ciągłym czasem i o skończonej liczbie stanów 0,1,..., I+K+N.

Dla takiego procesu definiuje się prawdopodobieństwo urodzin

$$P_{i,i+1} = Pr\left\{\nu(t+\delta) = i+1 \;/\; \nu(t) = i\right\}$$

oraz prawdopodobieństwo śmierci

$$P_{i,i-1} = Pr\left\{\nu(t+\delta) = i-1 \;/\; \nu(t) = i\right\},$$

$$\delta > 0, \quad i \in [0,1,\ldots,I+K+N],$$

które w naszym przypadku przybierają wartości:

$$P_{i,i+1} = \begin{cases} I\,\mu_1 & i=0,1,\ldots,K+N-1 \\ (I+K+N-i)\mu_1 & i=K+N,K+N+1,\ldots,K+N+I \end{cases} \qquad \text{(IV.6.59)}$$

$$P_{i,i-1} = \begin{cases} i\mu_2 & i=0,1,\ldots,K-1 \\ K\mu_2 & i=K,K+1,\ldots,K+N+I. \end{cases} \qquad \text{(IV.6.60)}$$

Proces taki posiada stacjonarny stan graniczny, a więc istnieje rozkład stacjonarny zmiennej $\nu(t)$,

$$\nu_i = \lim_{t\to\infty} Pr\left\{\nu(t) = i\right\}.$$

i może on być wyznaczony [GROS 74] jako:

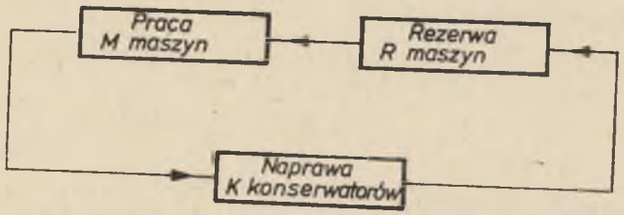
$$\nu_i = \nu_o H_i, \quad 1 \leqslant i \leqslant K+N+I. \qquad \text{(IV.6.61)}$$

$$\nu_o = \left[1 + \sum_{i=1}^{K+N+I} H_i\right]^{-1}, \qquad \text{(IV.6.62)}$$

$$H_i = \frac{\prod_{j=0}^{i-1} P_{j,j+1}}{\prod_{j=1}^{i} P_{j,j-1}}. \qquad \text{(IV.6.63)}$$

Proces urodzin i śmierci o identycznych jak (IV.6.59) i (IV.6.60) prawdopodobieństwach przejść był już w literaturze rozpatrywany, jako opisujący liczbę maszyn zepsutych $\mathcal{M}(t)$ w zadaniu konserwatora sformułowanym przez Palma [PALM 47] i zilustrowanym na rys. IV.12.

W zadaniu tym przyjmuje się, że w chwili początkowej pracuje równocześnie I maszyn, a dalszych R maszyn stanowi rezerwę. W przypadku awarii którejkolwiek z maszyn pracujących jest ona bezzwłocznie zastępowana przez jedną z maszyn rezerwowych.



Rys. IV.12. Zadanie konserwatora (model Palma)

Uszkodzenia maszyny są naprawiane na K równoległych stanowiskach konserwacyjnych. Maszyny znajdujące się w rezerwie nie podlegają uszkodzeniom (tzw. rezerwa zimna). Jeżeli założyć, że czas bezawaryjnej pracy każdej z maszyn ma rozkład wykładniczy z parametrem μ_1, czas naprawy na każdym ze stanowisk ma rozkład wykładniczy z parametrem μ_2 oraz że liczba maszyn rezerwowych R = K+N, wówczas zmienna losowa $\varkappa$ staje się identyczna z określoną poprzednio zmienną losową ϑ.

Można wykazać np. [KOPO 73A], że w rozpatrywanym przypadku

$$\left.\begin{aligned} H_i &= \frac{I^i}{i!}\rho^i && i=0,1,\ldots,K \\ H_i &= \frac{I^i}{K!K^{i-K}}\rho^i && i=K+1,\ldots,K+N \\ H_i &= \frac{I^{K+N}\,I!}{K!K^{i-K}(I+K+N-i)!}\rho^i && i=K+N+1,\ldots,I+K+N \end{aligned}\right\} \qquad \text{(IV.6.64)}$$

gdzie: $\rho = \mu_1/\mu_2$.

Znajomość stacjonarnego rozkładu zmiennej losowej ϑ pozwala na bezpośrednie określenie wszystkich interesujących parametrów systemu. W szczególności wydajność V wynosi:

$$V = \mu_2\left[\sum_{i=1}^{K} i v_i + K\sum_{i=K+1}^{I+K+N} v_i\right]$$

co w przypadku systemu pojedynczego, tj. dla I=K=1 prowadzi do wzoru (IV.6.42), a więc zależności znanej z [HUNT 56].

ZAKOŃCZENIE

W niniejszej pracy wskazano liczne przykłady ilustrujące tezę, iż odzwierciedlenie w formułowanym modelu rzeczywistych zależności występujących między parametrami złożonych systemów komputerowych, sieci komputerowych czy też dyskretnych procesów produkcyjnych prowadzi do wyjścia poza zbiór konwencjonalnych systemów masowej obsługi (dalsze przykłady znaleźć można w literaturze poświęconej modelowaniu takich systemów, np. w [KLEI 76], [KOBA 77], [GRAH 78], [SEID 79]).

Ponieważ analiza takich modeli jest w każdym przypadku utrudniona (w porównaniu do systemów konwencjonalnych), należy postawić podstawowe pytanie, czy uwzględnienie w modelu zależności nie mieszczących się w pojęciu systemu konwencjonalnego jest konieczne, to znaczy czy powoduje istotne zmiany własności modelu.

Na pytanie to trzeba odpowiedzieć twierdząco. Co więcej, zmiany takie mają charakter nie tylko ilościowy lecz w wielu (omówionych w rozdziale I) przypadkach również jakościowy. W załączonym dodatku przedstawiono pewien przykład przemysłowy, dla którego nawet znacznie uproszczony model zachowujący jednak charakterystyczne zależności między parametrami (w tym przypadku zależność czasu obsługi od czasu oczekiwania) pozwala nie tylko zrozumieć mechanizm powstawania pewnych, niepożądanych tendencji zmian charakterystyk obiektu ale również zaproponować efektywne metody przeciwdziałania takim tendencjom, co było niemożliwe przy użyciu modeli w postaci konwencjonalnych systemów masowej obsługi.

Szczególnie kłopotliwe są przypadki, gdy uwzględnienie tych zależności prowadzi do systemów, w których warunek stabilności jest różny dla różnych regulaminów obsługi.

Znaczną część pracy poświęcono badaniom jednego z takich przypadków, systemów z czasem obsługi zależnym od czasu oczekiwania. Jakkolwiek dla tej klasy systemów udało się uzyskać, przedstawione w pracy, stosunkowo kompletne wyniki analizy przy naturalnym regulaminie obsługi oraz znacznie rozwinięto metody analizy również przy innych regulaminach obsługi, to jednak bardzo wyraźnie wystąpiły ograniczenia zakresu stosowalności ścisłych metod analitycznych.

Spowodowało to konieczność sięgnięcia do metod symulacyjnych i to nie tylko w ich tradycyjnym zakresie stosowalności, tj. wyznaczaniu charakterystyk systemów stabilnych, ale również w zakresie wyznaczania warunków stabilności czy też doboru parametrów systemu. Pozytywne wyniki testowania zaproponowanych metod - przedstawione w rozdziale III - zachęcają do

szerszego ich wykorzystania, również w odniesieniu do innych klas niekonwencjonalnych systemów masowej obsługi i należą - zdaniem autora - do najistotniejszych rezultatów prezentowanych w pracy.

Biorąc pod uwagę ograniczenia stosowalności metod analitycznych wydaje się, iż dla badania niekonwencjonalnych systemów masowej obsługi najbardziej przydatne będą metody kombinowane, łączące w sobie modelowanie cyfrowe oraz przybliżone metody analityczne. Połączenie to może mieć różny charakter. Obok podejść zaprezentowanych w tej pracy bardzo interesująca wydaje się koncepcja (przedstawiona np. w [PUGA 73]) wykorzystania zmiennych dodatkowych (np. wynikających z użycia przybliżonych metod analizy) dla zmniejszenia pracochłonności symulacji, co znakomicie zwiększa użyteczność tej metody badania systemów masowej obsługi.

Autor żywi nadzieję, że niniejsza praca przyczyni się do rozwoju teorii niekonwencjonalnych systemów masowej obsługi - jednego z najważniejszych aktualnie zadań w zakresie metod oceny efektywności systemów informatycznych oraz dyskretnych procesów produkcyjnych.

DODATEK

Odwołamy się do przykładu pewnego procesu produkcyjnego - ciągu operacji zachodzących pomiędzy stalownią konwertorową a walcarką zgniatacz w kombinacie metalurgicznym - wskazując, iż przedstawienie go jako systemu masowej obsługi, zawierającego rozpatrywane w poprzednich rozdziałach modele, pozwala nie tylko przewidzieć i uzasadnić niektóre, występujące tam zjawiska, ale również zaproponować praktyczne pociągnięcia organizacyjne pozwalające poprawić jego efektywność.

Ten proces produkcyjny, typowy dla współczesnych kombinatów metalurgicznych, był przedmiotem badań kierowanego przez autora zespołu pracowników ZSAK PAN. Zebrane zostały szczegółowe dane o pracy rzeczywistego obiektu, które stały się podstawą do analiz jego własności, a także badań porównawczych różnych propozycji modyfikacji obowiązującej organizacji przepływu materiału, opisanych szczegółowo w opracowaniach [ZSAK 79] oraz [ZSAK 80].

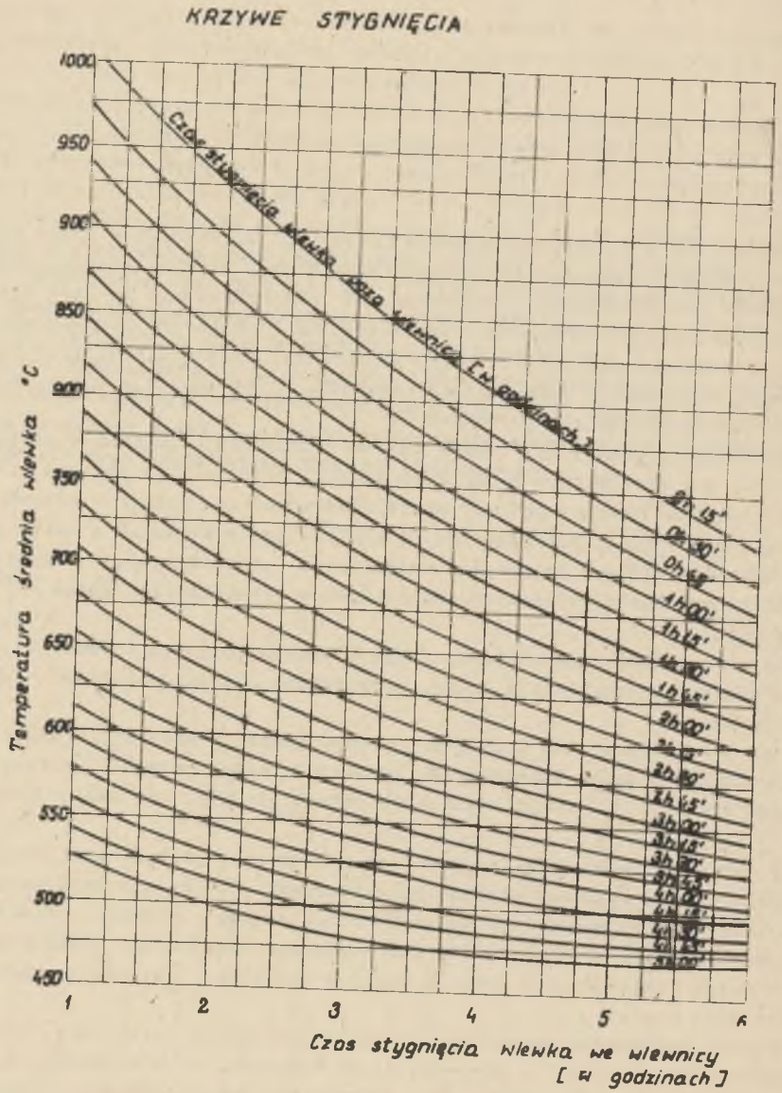
Poniżej, po krótkiej prezentacji rozpatrywanego obiektu i obowiązujących aktualnie reguł organizacji pracy, przytoczone zostaną niektóre wyniki tych badań.

Stal produkowana jest w stalowni konwertorowej, złożonej w analizowanym przypadku z dwóch konwertorów, z których każdy pozwala na uzyskanie jednego wytopu, tj. porcji stali określonego gatunku, co około 35-40 minut.

Wytop, spuszczony do kadzi odlewniczej, przewożony jest do hali lejniczej, w której płynną stalą napełniane są umieszczone na wozach kolejowych zestawy wlewnic. Po całkowitym odlaniu wytopu zestaw wlewnic przetaczany jest na tzw. tory krystalizacji, gdzie podczas narzuconego technologią okresu czasu (zależnego od gatunku stali i kształtu wlewnic) następuje krzepnięcie stali.

Po upływie wymaganego czasu krystalizacji z wlewków usuwa się wlewnice (striperowanie wlewków) i przewozi je do oddziału pieców wgłębnych w celu nagrzania tak, aby uzyskały one jednakową w całej objętości temperaturę (niższą od temperatury topnienia), zapewniającą prawidłowy przebieg walcowania.

Pojedynczy wytop (na który składa się 16-36 wlewków o łącznej masie ok. 350 ton) dzielony jest przy tym pomiędzy dwa - trzy piece wgłębne. Czas operacji nagrzewania wlewków, narzucony przez technologię, zależy od stanu termicznego, przy którym nagrzewanie się rozpoczyna, a więc od czasu trwania wszystkich wcześniejszych operacji technologicznych.



Rys. 1. Przykładowe krzywe stygnięcia wlewków, określające temperaturę w funkcji czasu stygnięcia wlewka we wlewnicy oraz czasu stygnięcia wlewka poza wlewnicą

Tabela 1

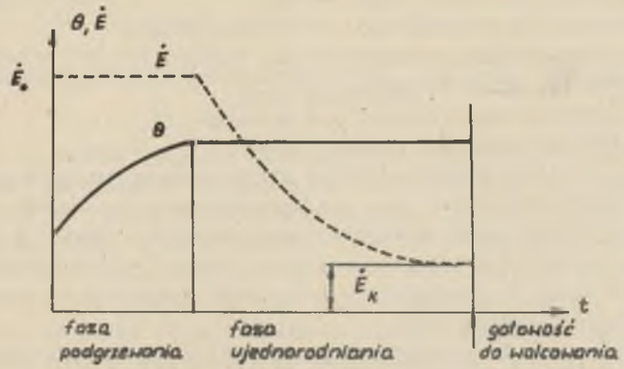
Przykładowa tabela czasów nagrzewania wlewków w piecach wgłębnych

Temperatura wlewków ładowanych [°C]	< 300	300-400	401-500	501-600	601-650	651-700
Całkowity czas nagrzewania [h]	9^{00}-11^{30}	8^{30}-10^{00}	8^{00}-9^{30}	7^{20}-9^{00}	6^{40}-8^{20}	6^{10}-7^{50}
Czas ujednorodnienia [h]	5^{20}-7^{20}	4^{50}-6^{50}	4^{30}-6^{00}	4^{00}-5^{50}	3^{50}-5^{40}	3^{40}-5^{10}
Temperatura wlewków ładowanych [°C]	701-750	751-800	801-850	851-900	901-950	>950
Całkowity czas nagrzewania [h]	5^{20}-7^{00}	4^{30}-6^{00}	3^{50}-5^{10}	3^{30}-4^{30}	3^{10}-4^{10}	2^{50}-3^{30}
Czas ujednorodnienia [h]	5^{30}-4^{50}	3^{20}-4^{10}	3^{00}-3^{50}	2^{40}-3^{30}	2^{20}-3^{10}	2^{00}-2^{50}

Charakter tej zależności ilustrują na przykładzie zaczerpniętym z [TECH 80] rys. 1 oraz tabela 1.

Jak widać, czas nagrzewania waha się bardzo silnie od ok. 3 godzin do ponad 11 godzin, przy czym składają się na niego dwie oddzielne fazy:

- f a z a p o d g r z e w a n i a, w której przy użyciu pełnej mocy cieplnej pieca doprowadza się powierzchnię wlewków do wymaganej temperatury oraz
- f a z a u j e d n o r o d n i e n i a, podczas której przy zachowaniu stałości temperatury powierzchni wyrównywana jest (z pewną tolerancją) temperatura w całej objętości wlewków. Charakter zmian temperatury w piecu wgłębnym oraz strumienie energii cieplnej doprowadzonej do komory pieca naszkicowano na rys. 2.



Rys. 2. Charakter zmian temperatury w komorze θ oraz strumienia energii $\dot{E}$ w czasie nagrzewania wsadu w piecu wgłębnym, $\dot{E}_K \cong 0.1\ \dot{E}_o$

Wygrzane wlewki wyładowywane są pojedynczo na samotok przenoszący je bezpośrednio do walcarki. Ze względu na szybkie stygnięcie rozgrzanego do ok. 1300°C metalu czas pobytu wlewków na samotoku powinien być możliwie krótki. W przypadku jednoczesnego osiągnięcia gotowości do walcowania przez wlewki w kilku piecach występuje efekt blokowania komór przez nagrzany już wsad.

Oddział pieców wgłębnych, złożony z dużej liczby (w rozpatrywanym przypadku 40) pieców, pełni więc dwie funkcje:

- technologiczną - polegającą na podgrzaniu wsadu oraz
- organizacyjną - jako bufor gorącego materiału dla walcarki.

Dla zapewnienia prawidłowej współpracy stalowni oraz walcowni konieczne jest zainstalowanie pomiędzy nimi dodatkowego magazynu materiału wyrównującego długofalowe różnice wydajności (awarie, remonty) w postaci składowiska wsadu zimnego.

Biorąc pod uwagę możliwość uzupełnienia braków wsadu (lub zbycia nadwyżek) drogą wymiany z innymi hutami zapewnia się na drodze organizacyjnej (w normalnych warunkach pracy), że stanowisko to nie ulega nigdy całkowitemu opróżnieniu ani też przepełnieniu.

Przy przestrzeganiu narzuconych norm technologicznych, co jest warunkiem oczywistym, jako nadrzędne kryterium oceny organizacji pracy omawianego odcinka przyjmuje się zapewnienie ciągłości pracy walcarki - zgniatacza. W dalszej kolejności natomiast dąży się do minimalizacji kosztów paliwa (mieszanki gazu koksowniczego i wielkopiecowego), zużywanego dla nagrzewania wlewków w piecach wgłębnych, które w rozpatrywanym przykładzie sięgają rzędu 250 mln zł rocznie (wg cen obowiązujących w latach 1978-79). Na koszty te wpływa wiele czynników technicznych (konstrukcja i stan techniczny pieców a także palników, działanie aparatury kontrolno-pomiarowej zainstalowanej na piecowni), bardzo silny wpływ ma także organizacja przepływu materiału na odcinku stalownia - piece wgłębne. W szczególności muszą być określone reguły wyboru:

- momentu rozpoczęcia striperowania kolejnego wytopu,
- kolejności striperowania wytopów,
- momentu ładowania wsadu do pieca,
- wariantu ładowania wsadu zimnego lub gorącego,
- wytopu, z którego pochodzą ładowane wlewki.

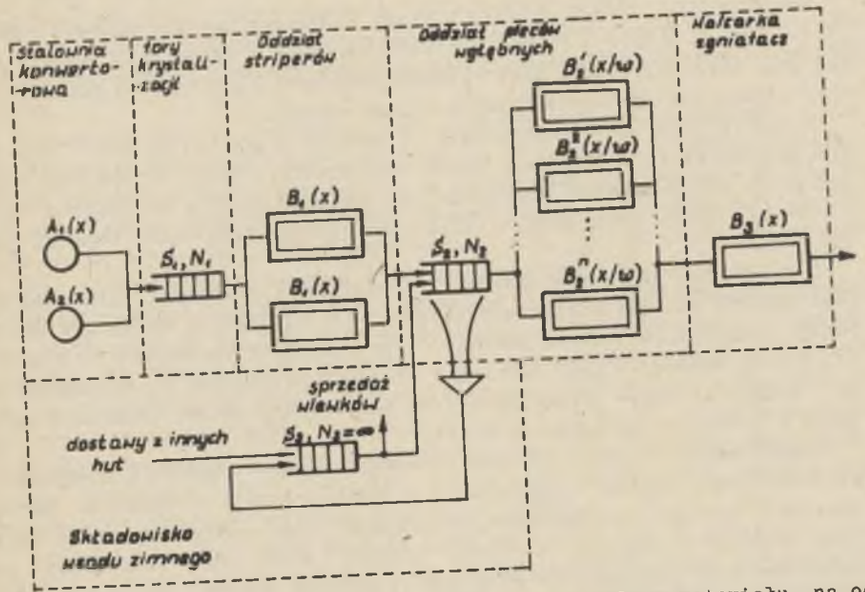
Aktualnie, w rozpatrywanym obiekcie striperowanie wytopów realizowano zawsze, gdy było to możliwe, tzn. gdy dostępne było stanowisko striperowe oraz istniały wytopy, których krystalizacja dobiegła końca. W przypadku obecności na torach krystalizacji większej liczby wytopów gotowych do striperowania wybierano je w zasadzie w kolejności zakończenia krystalizacji (z wyjątkiem stali specjalnych).

Zwolniony (po wyładunku nagrzanych wlewków) piec był natychmiast ładowany wsadem gorącym, jeśli taki był dostępny na piecowni. Przy braku wsadu gorącego dopuszczalne było oczekiwanie na jego podwiezienie, przy czym liczba pustych pieców nie mogła przekroczyć trzech. W takim przypadku bo-

wiem ładowany był wsad zimny (co miało miejsce również bezpośrednio po opróżnieniu pieca, jeśli wiadomo było że nie można się spodziewać dostawy wsadu gorącego).

Reguły wyboru ładowanego wsadu gorącego nie były jednoznaczne, przy zachowaniu ogólnej tendencji do ładowania wytopów w kolejności ich dostawy.

Problem optymalnej organizacji przepływu materiału na tego typu odcinku produkcyjnym rozpatrywano w wielu pracach, stosując zarówno metody deterministyczne, stochastyczne, jak i badania symulacyjne; uzyskane wyniki nie doprowadziły jednak do syntezy efektywnego zbioru reguł optymalnych.



Rys. 3. System masowej obsługi reprezentujący przepływ materiału na odcinku produkcyjnym stalownia konwertorowa - walcarka zgniatacz kombinatu metalurgicznego

Na rys. 3 omawiany odcinek produkcyjny przedstawiono jako system masowej obsługi, zaznaczając odpowiedniość pomiędzy poszczególnymi elementami schematu (źródła zgłoszeń stanowiska obsługi, bufory) a elementami obiektu technologicznego (przy pominięciu operacji transportowych).

Mamy tu, jak widać, do czynienia z wieloetapowym systemem obsługi, przy czym określony jest minimalny czas pobytu zgłoszeń w buforze S_1, natomiast czas obsługi zgłoszeń przez stanowiska obsługi drugiego etapu jest zależny od czasu, jaki upłynął pomiędzy wygenerowaniem zgłoszenia - odlaniem wytopu a załadowaniem go do pieca wgłębnego (ze względu na fakt znacznie szybszego stygnięcia wlewków wystriperowanych można tu w uproszczeniu mówić o zależności czasu obsługi od czasu oczekiwania na jej rozpoczęcie).

Rozumiejąc przez pojęcie zgłoszenia grupę wlewków ładowanych do jednego pieca stwierdzamy, że zarówno generacja zgłoszeń, jak też ich obsługa na pierwszym etapie odbywa się grupowo. Brak bufora pośredniego pomiędzy drugim i trzecim etapem obsługi powoduje występowanie zjawiska blokowania. Organizacja przepływu materiału określona jest przez regulamin obsługi zgłoszeń.

O tym, że zachodzące w rozpatrywanym obiekcie procesy mają istotnie silnie stochastyczny charakter, świadczą np. wyniki analizy strumienia wytopów opuszczających stalownię, jako że współczynnik zmienności rozkładu odstępów między chwilami zakończenia odlewania kolejnych wytopów przekracza wartość 0.6.

Bezpośrednia, ścisła analiza takiego złożonego systemu masowej obsługi jest - jak to wynika z rozważań przedstawionych w niniejszej pracy - niemożliwa, niemniej jednak nawet fragmentaryczne, uproszczone rozpatrzenie tego układu może dostarczyć interesujące wnioski.

I tak na przykład w odniesieniu do analizy pracy striperów (z punktu widzenia określenia średniego czasu oczekiwania w celu doboru np. liczby stanowisk striperowania) wystarczająco dokładne wyniki uzyskiwano drogą stosowania konwencjonalnych, jednostanowiskowych systemów obsługi [MALI 71], [HAJK 73].

Z kolei, potraktowanie układu piece wgłębne - zgniatacz jako dwuetapowego systemu obsługi z blokowaniem [HAJK 73], [HAJK 74] względnie dwuetapowego zamkniętego systemu obsługi (typu zadania konserwatora) [BUZA 71B], [IZYD 81A] pozwala - przy założeniu stałości parametrów dostarczanego wsadu - badać np. wpływ liczby pieców na prawdopodobieństwo przestoju zgniatacza z powodu braku gorącego wsadu.

Przedstawiony w [BUZA 73] model reprezentujący zdarzenia zachodzące w układzie piece wgłębne - zgniatacz jako zmiany stanu w łańcuchu Markowa umożliwia wyciągnięcie wniosków dotyczących reguł załadunku pieców wgłębnych. Chociaż ze względu na rozmiar wektora stanu jego stosowalność jest ograniczona do przypadków o niewielkiej liczbie pieców, to jednak zaobserwowane prawidłowości pozostają w mocy również w stosunku do większych obiektów.

Wszystkie wspomniane wyżej modele bazowały na założeniu, iż wsad ładowany do pieców posiada parametry niezależne od stanu stanowiska obsługi, nie pozwalają one więc na uwzględnienie relacji pomiędzy czasem stygnięcia wsadu a wydajnością piecowni, związaną wprost z temperaturą ładowanego wsadu.

Rozważenie efektów spowodowanych przez te relacje jest możliwe przy przedstawieniu procesów dostawy wsadu jako systemu masowej obsługi z czasem obsługi zależnym od czasu oczekiwania.

Przypomnijmy w tym miejscu przykład przedstawiony w punkcie III.4.c, którego podobieństwo do procesów dostawy wsadu stanie się wyraźne, gdy zamiast stanowisko obsługi "zgłoszenie główne", "zgłoszenie dodatkowe" mówić będziemy: "piecownia", "wsad gorący", "wsad zimny".

Ponieważ w przypadku systemów obsługi ze zgłoszeniami napływającymi strumieniem Poissona pojęcie utraty zgłoszenia jest równoważne pojęciu wstrzymywania pracy źródła zgłoszeń, więc w rozpatrywanej analogii odpowiada ono przesłaniu wlewków na składowisko wsadu zimnego bądź wstrzymaniu pracy striperów (przy pominięciu stygnięcia wlewków we wlewnicach).

Dane liczbowe do obliczeń numerycznych zostały we wspomnianym przykładzie dobrane pod kątem takiej analogii.

W szczególności w oparciu o dane rzeczywiste, zebrane na terenie walcowni półwyrobów, analizowane w [ZSAK 79], wyznaczono średni czas grzania wsadu zimnego, a także średni czas grzania wsadu gorącego, który rozbito na dwie grupy: wsad "bardzo gorący" o temperaturze ładowania powyżej 780° i wsad "ostudzony", o temperaturze ładowania poniżej 780°. Grupy te uzyskano dzieląc zbiór ładunków komorowych nagrzewanych w rozpatrywanym okresie (i pochodzących z ładowania wsadu gorącego) na dwa prawie równoliczne podzbiory.

Wzajemne proporcje średnich czasów obsługi, a także parametrów D i τ określających wartości czasów oczekiwania, po upływie których następuje zmiana wymaganego czasu obsługi, są w przykładzie tym zbliżone do proporcji odpowiednich wartości charakterystycznych dla rzeczywistego procesu. Stąd też do procesów dostawy wsadu odnoszą się, w sensie jakościowym, wnioski sformułowane w oparciu o przytoczone wyniki numeryczne.

I tak na przykład, wskazany efekt pogarszania się wskaźników efektywności stanowiska obsługi (piecowni) przy zwiększaniu intensywności strumienia zgłoszeń (strumienia wsadu gorącego), jeśli zachowane zostają niezmienione reguły organizacji pracy, rzeczywiście jest obserwowany w praktyce.

Nie powtarzając omawianych w punkcie III.4.c sposobów przeciwdziałania temu zjawisku, można w oparciu o przedmiotowy model (jedyny odzwierciedlający wspomniany efekt) stwierdzić konieczność uzależnienia reguł dostawy wsadu na piecownię od aktualnego stanu obiektu.

Uzależnienie takie może mieć na przykład postać ograniczenia liczby wytopów oczekujących na załadunek do pieców wgłębnych drogą odpowiedniego wstrzymywania striperowania. Testy [ZSAK 80], przeprowadzone w oparciu o dane rzeczywiste, pozwoliły stwierdzić, iż takie postępowanie zmniejszyłoby o kilka procent wymagany, łączny czas nagrzewania wlewków.

Biorąc pod uwagę przedstawione w punkcie IV.b uwagi odnośnie do doboru regulaminu obsługi przeprowadzono także eksperyment, polegający na zasymulowaniu zastosowania do szeregowania wlewków ładowanych do pieców sformułowanego tam regulaminu obsługi R_1 z wykorzystaniem algorytmu przybliżonego rozwiązywania odpowiedniego zadania komiwojażera.

Stwierdzono, iż taka organizacja załadunku wsadu prowadziłaby do ok. 2% skrócenia czasu nagrzewania wlewków.

Nie przytaczając szczegółów zawartego w [ZSAK 79] oraz [ZSAK 80] (a w zakresie wcześniej omawianych modeli w [HAJK 73], [HAJK 74], [BUZA 73] opisu przeprowadzonych eksperymentów, można stwierdzić, iż przedstawienie rozpatrywanego procesu produkcyjnego w postaci odpowiedniego niekonwen-

cjonalnego systemu masowej obsługi pozwala na zrozumienie jego specyfiki, a także wskazanie możliwości usprawnienia jego organizacji.

W licznych pracach (np. [KLEI 76], [KOBA 77], [GRAH 78], [WEGR 78], [WOLI 78], [SEID 79]) wskazywano na celowość modelowania w postaci wielostanowiskowych systemów masowej obsługi, złożonych systemów komputerowych oraz sieci komputerowych.

Przy rozpatrywaniu tych modeli napotyka się z reguły na podobne jak w omawianym wyżej przykładzie trudności, w szczególności realia procesów rzeczywistych wykraczają zwykle poza zbiór modeli, które można w sposób ścisły przeanalizować. Wynika to zarówno ze złożonej struktury procesów obsługi w ramach indywidualnych stanowisk obsługi, jak też wyraźnych wzajemnych uzależnień trybu pracy różnych stanowisk obsługi.

Co więcej, z rozważań przedstawionych w niniejszej pracy jednoznacznie wynika, że nie należy się spodziewać opracowania ogólnych metod analitycznych, umożliwiających badanie złożonych sieci niekonwencjonalnych stanowisk obsługi.

Z drugiej strony, jak to można sprawdzić na licznych przykładach, zaniedbanie wspomnianych uzależnień, a więc niekonwencjonalnego charakteru rozpatrywanych systemów uniemożliwia odzwierciedlenie ich podstawowych własności.

Efektywną drogą przezwyciężenia tej sprzeczności jest zastępowanie złożonych, niemożliwych do bezpośredniej analizy sieci stanowisk obsługi zbiorem systemów prostszych, zachowujących jednak zasadnicze cechy układu pierwotnego (drogą doboru systemów najlepiej reprezentujących wydzielone fragmenty tego układu) i rozpatrywanie takich uproszczonych modeli, łączonych następnie w bardziej rozbudowane struktury. W przedstawionym przykładzie prowadziłoby to do analizy a następnie połączenia systemu WAD z systemem dwuetapowym pozbawionym bufora pośredniego.

Jak widzieliśmy, nawet oddzielne rozpatrywanie tych podukładów prowadziło do interesujących wniosków. Podobne podejście okazuje się skuteczne również w wielu innych przypadkach, co stwierdzono np. w pracy [KLEI 80], rozpatrując, z bardzo dobrym skutkiem, pewien system transmisji danych, w stosunku do którego zawiodły klasyczne metody i modele.

Z tego względu można oczekiwać, iż dalsze prace w zakresie niekonwencjonalnych systemów masowej obsługi będą ukierunkowane na dwa niezbędne przy takim podejściu zagadnienia:

- analizę i badanie własności różnych, oddzielnych klas systemów niekonwencjonalnych,
- opracowywanie przybliżonych metod łączenia systemów zarówno konwencjonalnych, jak i niekonwencjonalnych w bardziej rozbudowane sieci, często o złożonych powiązaniach.

Przy rozwiązaniu tych zadań (a szczególnie drugiego z nich) należy oczekiwać znacznego udziału modelowania cyfrowego i to zarówno w charakterze metody weryfikacji dokładności rozwiązań analitycznych, jak też w charakterze metody zasadniczej, uzupełnionej o pewne przybliżone rozwiązanie analityczne, pozwalające w istotny sposób zwiększyć jej efektywność.

LITERATURA

ADIR 69 I. Adiri: A time sharing queue. Management Science 15 (1969), p. 639.

ALLE 78 A.O. Allen: Probability, statistics and queuing theory with computer science applications. Academic Press, New York, 1978.

ANDE 69 D.R. Andersqn, C.L.Moodie: Optimal buffer storage capacity in production line systems. International Journal of Production Research 7 (1969), p. 233.

ANTO 73 P. Antosik, J. Mikusiński, R. Sikorski: Theory of distributions. Elsevier, Amsterdam, 1973.

ARJA 78 E. Arjas, T. Lehtonen: Approximating many server queues by means of single-server queues. Mathematics of Operations Research 3 (1978), p. 205.

ARYA 72 K.L. Arya: Study of a network of serial and non-serial servers with phase type service and finite queuing space. Journal of Applied Probability 9 (1972), p. 198.

AVII 65A B. Avi-Itzhak, M. Yadin: A sequence of two servers with no intermediate queue. Management Science 11 (1965), p. 553.

AVII 65B B. Avi-Itzhak: A sequence of service stations with arbitrary input and regular service times. Management Science 11 (1965), p. 565.

BACE 79 F. Baccelli: Analysis of a server facility with periodic checkpoint. Rapport de Recherche no 58 (1979), Universite de Paris-SUD, Centre d'Orsay, Laboratoire de Recherche en Informatique.

BACE 81 F. Baccelli, G. Hebuterne: On queues with impatient customers - - preliminary draft. Paris, March 1981 (maszynopis pracy udostępniony przez autorów).

BALB 77 G. Balbo, S.C. Bruell, H.D. Schwetman: Customer classes and closed network models - a solution technique. Proceedings IFIP Congress 1977, North Holland Publishing Company, p. 559.

BARB 76 A.D. Barbour: Networks of queues and method of stages. Advances in Applied Probability 8 (1976), p. 584.

BARR 57 D.Y. Barrer: Queueing with impatient customers and indifferent clerks. Operations Research 5 (1957), p. 640.

BART 62 K. Barten: A queueing simulator for determining optimum inventory levels in a sequential process. Journal of Industrial Engineering 13 (1962), p. 245.

BASK 75 F. Baskett, K.M. Chandy, R.R. Muntz, F.G. Palacios: Open closed and mixed networks of queues with different classes of customers. Journal of ACM 22 (1975), p. 248.

BASU 77 R.N. Basu: The interstage buffer storage capacity of nonpowered assembly lines, a simple mathematical approach. International Journal of Production Research 15 (1977), p. 365.

BEUT 78 F.J. Beutler, B. Melamed: Decomposition and customer streams of feedback networks of queues in equilibrum. Operations Research 26 (1978), p. 1059.

BHAT 69 U.N. Bhat: Queueing systems with first-order dependence. Opsearch 6 (1969), p. 1.

BHAT 73 U.N. Bhat: Some problems in finite queues, in Mathematical methods in queueing theory. Lecture Notes in Economics and Mathematical Systems, Springer-Verlag Berlin, 1973, p. 26.

BHAT 79A U.N. Bhat: Approximation techniques in the solution of queueing problems. Naval Research Logistics Quarterly 26 (1979), p. 311.

BHAT 79B U.N. Bhat, R.E. Nance: An evaluation of CPU efficiency under dynamic quantum allocation. Journal of the ACM 26 (1979) p.761.

BLAZ 77 J. Błażewicz: Deadline scheduling of tasks - a survey. Foundations of Control Engineering 1 (1977), p. 4.

BLEV 76 P.R. Blevis, C.V. Ramamoorthy: Aspects of a dynamically adaptive operating system. IEEE Transactions of Computers C-25 (1976), p. 713.

BLOM 73 N. Blomquist: Serial correlation in a simple dam process. Operations Research 21 (1973), p. 966.

BOON 78 P. Boonsong: A simulation study of a class of FCFS queues with EARMA correlation structure. Master Thesis, Naval Postgradual school, Monterey, California, 1978.

BOXG 70 G.E. Box, G.M. Jenkins: Time series analysis-forecasting and control. J. Wiley, New York, 1970.

BOXM 78A C.J. Boxma: On tandem queueing model with identical service times at both counters. Advances in Applied Probability 10 (1978), p. 616 (part. I), p. 644 (part. II).

BOXM 78B O.J. Boxma: Two queues in series with non-overlapping service times. IBM Research Raport RC 7412, 1978.

BOXM 81 O.J. Boxma, A.G. Konheim: Approximate analysis of exponential queueing systems with blocking. Acta Informatica 15 (1981),p.19.

BROC 60 E. Brockmeyer, H.L. Halstrom, A. Jensen: The life and works of A.K. Erlang. Acta Polytechnica Scandinavia, (1960).

BRUE 78 S.C. Bruell: On single and multiple job class queueing network models of computer systems. Ph.D.Thesis, Dept. of Computer Sciences, Purdue University, 1978.

BRUM 71 S.L. Brumelle: On the relation between customer and time averages in queues. Journal of Applied Probability 8 (1971), p. 508.

BRUM 72 S.L. Brumelle: A generalization of $L = \lambda W$ to moments of a queue length and waiting times. Operations Research 20 (1972), p. 1127.

BURK 56 P.J. Burke: The output of a queueing systems. Operations Research 4 (1956), p. 699.

BURK 64 P.J. Burke: The dependence of delays in tandem queues. The Annales of Mathematical Statistics 35 (1964), p. 874.

BURK 68 P.J. Burke: The output of stationary M/M/s queueing system. The Annales of Mathematical Statistics 39 (1968), p. 1144.

BURK 72 P.J. Burke: Output processes and tandem queues. Proceedings of Symposium on Computer Communications Networks and Teletraffic, Polytechnic Institute of Brooklyn (J.Fox, Editor), Polytechnic Press, Brooklyn N.Y. 1972, p. 419.

BUXE 73 M.G. Buxey, N.D. Slack, R. Wild: Production flow line system design - review. AIIE Transactions 5 (1973), p. 37.

BUXW 77 W. Bux, V. Herzog: The phase concept: approximation of measured data and performance analysis; in K.M. Chandy, M.Reiser (editors): Computer Performance, North Holland Publishing Company, Amsterdam, 1977, p. 23.

BUZA 67 J. Buzacott: Automatic transfer lines with buffer stocks. International Journal of Production Research 15 (1967), p. 183.

BUZA 71A J.A. Buzacott: The role of inventory banks in flow-line production systems. International Journal of Production Research 9 (1971), p. 425.

BUZA 71B J.A. Buzacott, J.R. Callahan: The capacity of the soaking pit-rolling mill complex in steel production INFOR 9 (1971), p. 87.

BUZA 72 J.A. Buzacott: The effect of station breakdowns and random processing times on the capacity of flow lines with in-process storage. AIIE Transactions 4 (1972), p. 308.

BUZA 73 J.A. Buzacott, J.R. Callahan: The pit charging problem in steel production. Management Science 20 (1973), p. 665.

BUZA 74 J.A. Buzacott: The effect of queue discipline on the capacity of queues with service times dependent on waiting times. INFOR 12 (1974), p. 174.

BUZE 73 J. P. Buzen: Computational algorithms for closed queueing networks with exponentical servers. Communications of the ACM 17 (1973), p. 527.

BUZE 74 J.P. Buzen, P.S. Goldberg: Guidelines for the use of infinite source queueing models in the analysis of computer system performance. Proceedings of SJCC, 1974, AFIPS Vol. 43, p. 371.

CALL 73 J.R. Calahan: A queue with waiting time dependent service times. Naval Research Logistics Quarterly 20 (1973), p. 321.

CALO 77 S.B. Calo: The message channel, a tandem interconnection of queues: Part I, waiting time preserving families of queues and related realizations of message channel. IBM Research Raport RC 6868, 1977.

CALO 78 S.B. Calo: The message channel, a tandem interconnection of queues: Part II, waiting time properties under general arrivals. IBM Research Report RC 7170, 1978.

CASE 77 P. Caseau: Simulation de processus en serie au moyen de modeles markovien. Rapport Electricite de France, HI (2314) 00, 1972.

CASE 79 P. Caseau, G. Pujolle: Throughput capacity of a sequence of queues with blocking due to finite waiting room. IEEE Transactions on Software Engineering 5 (1979), p. 631.

CARN 76 C.A. Carnall, R. Wild: The location of variable work station and the performance of production flow lines. International Journal of Production Research 14 (1976), p. 703.

CHAN 72A W. Chang: Two servers in sequence with Erlang input. Proceedings of the Symposium of Computer Communications Networks and Teletraffic. Polytechnic Institute of Brooklyn, April 1972, p. 409.

CHAN 72B K.M. Chandy: The analysis and solutions for general queueing networks. Proc. VI-th Annual Conference on Information Sciences and Systems, Princeton University, March 1972.

CHAN 75A W. Chang: Sequential server queues for computer communication systems analysis. IBM Journal of Research and Development 19 (1975), p. 476.

CHAN 75B K.M. Chandy, U. Herzog, L. Woo: Approximate analysis of general queueing networks. IBM Journal of Research and Development 19 (1975), p. 43.

CHAN 75C K.M. Chandy, J.C. Browne, C.W. Dissly, W.R. Uhrig: Analytic models for roll back and recovery strategies in data base systems. IEEE Transactions on Software Engineering SE 1 (1975), p. 100.

CHAN 77 K.M. Chandy, J.W. Howard, D.F. Towsley: Product form and local balance in queueing networks. Journal of the ACM 24 (1977), p. 250.

CHAN 78 K.M. Chandy, C.H. Sauer: Approximate methods for analysing queueing network models of computer systems. Computing Surveys 10 (1978), p. 281.

CHOO 79 Q.H. Choo, B. Conolly: New results in the theory of repeated orders queueing systems. Journal of Applied Probability 16 (1979), p. 631.

CHOO 80 Q.H. Choo, B. Conolly: Waiting time analysis for a tandem queue with correlated service. European Journal of Operational Research 4 (1980), p. 337.

CINL 67A E. Cinlar: Time dependence of queues with semi-markovian services. Journal of Applied Probability 4 (1967), p. 356.

CINL 67B E.Cinlar: Time dependence of queues with semi-markovian arrivals. Journal of Applied Probability 4 (1967), p. 365.

COFF 78 E.G. Coffman, M. Hofri: A class of FIFO queues arising in computer systems. Operations Research 26 (1978), p. 864.

COFF 80 E.G. Coffman, jr (red.): Teoria szeregowania zadań. WNT, Warszawa, 1980.

COHE 69 J.W. Cohen: The single server queue. North Holland Publishing Company, Amsterdam, 1969.

CONO 68 B.W. Conolly: The waiting time process for a certain correlated queue. Operations Research 16 (1968), p. 1006.

CONO 69 B.W. Conolly, N. Hadidi: A correlated queue. Journal of Applied Probability 6 (1969), p. 122.

CONO 74 B.W. Conolly, N. Hadidi: A comparison of the operational features of conventional queues with a self - regulating system. Applied Statistics 18 (1974), p. 41.

CONO 79 B.W. Conolly, Q.H. Choo: The waiting time process for a generalized correlated queue with exponential demand and service. SIAM Journal on Applied Mathematics 37 (1979), p. 263.

COXD 55 D.R. Cox: A use of complex probabilities in the theory of stochastic processes. Proceedings of the Cambridge Philosophical Society 51 (1955), p. 313.

CRAB 77 T.B. Crabill: A classified bibliography of research on optimal design and control of queues. Operations Research 25 (1977), p. 219.

CRAN 74 M.A. Crane, D.L. Iglehart: Simulating stable stochastic systems I - General multiserver queues. Journal of ACM 21 (1974), p.103.

CRAN 77 M.A. Crane, A.J. Lemoine: An introduction to the regenerative method for simulation analysis. Springer Verlag, Lecture Notes in Control and Information Sciences, vol. 4, New York 1977.

CZAC 74 T. Czachórski, A. Wolisz: Priorytetowy, mieszany regulamin obsługi źródeł jednowymiarowych. Podstawy Sterowania 4 (1974), s. 115.

CZAC 79 T. Czachórski: Przybliżone rozwiązania dokładnych modeli i dokładne rozwiązanie modeli przybliżonych w ocenie efektywności złożonych układów komputerowych. Model oparty o symulację. Podstawy Sterowania 9 (1979), s. 125.

DALE 68 D.J. Daley: The correlation structure of the output process of some single server queueing systems. The Annales of Mathematical Statistics 39 (1968), p. 1007.

DALE 76 D.J. Daley: Queueing output processes. Advances in Applied Probability 8 (1976), p. 395.

DATT 78 E.S. Dattatreya: Tandem queueing systems with blocking. Ph. D dissertation, IEOR Department, University of California, Berkeley, January 31, 1978.

DAVI 66 L.E. Davis: Pacing effects on manual assembly lines. International Journal of Production Research 4 (1966), p. 171.

DAYP 73 P. Day, J. Hines: ARGOS - an operating system for a computer utility supporting interactive instrument control. Operating Systems Review 7 (1973), No 4, p. 28.

DISN 73 R.L. Disney, R.L. Farrel, P.R. de Morais: A characterization of M/G/1 queues with renewal departure process. Management Science 19 (1973), p. 1222.

DISN 75 R.L. Disney: Random flow in queueing networks: a review and critique. Transactions of the AIIE 7 (1975), p. 268.

DISN 76 R.L. Disney, P.R. de Morais: Covariance properties for the departure process of M/E/1/N queues. Transactions of the AIIE 9 (1976), p. 169.

DUDA 81 A. Duda: Modele procedur odtwarzania w komputerowych systemach czasu rzeczywistego. Podstawy Sterowania 11 (1981), s. 125.

EISE 63 M. Eisen, M. Tainiter: Stochastic variations in queueing processes. Operations Research 11 (1963), p. 922.

EISE 71 M. Eisenberg: Two queues with changeover times. Operations Research 19 (1971), p. 386.

EISE 74 M. Eisenberg: Queues with periodic service and changeover time. Operations Research 20 (1974), p. 440.

ERLA 09 A.K. Erlang: Probability and telephone calls. Nyt Tidsskr. Mat. ser. B. 20 (1909), p. 33.

FELL 80 W. Feller: Wstęp do rachunku prawdopodobieństwa. Tom 1. PWN, Warszawa 1980 (wyd. 4), Tom 2, PWN, Warszawa 1981 (wyd. 3).

FINC 59A P.D. Finch: A probability limit theorem with application to a generalization of queueing theory. Acta Math. Acad. Sci. Hung. 10 (1959), p. 317.

FINC 59B P.D. Finch: The output process of the queueing system M/G/1. Journal of the Royal Statistical Society B. 21. (1959). p. 375.

FOLE 77 R.D. Foley: Queues with feedback, Ph. D.Dissertation, University of Michigan, 1977.

FOST 78 F.G. Foster, H.H. Perros: Queue networks with blocking: exact and approximate results. Proceedings of the 2 nd International Symposium on Operating Systems IRIA, Rocquencourt, 1978.

FOST 79 F.G. Foster, H.G. Perros: Hierarchical queue networks with partially shared servicing. Journal of the Operational Research Society 30 (1979), p. 157.

FOST 80 F.G. Foster, H.G. Perros: On the blocking process in queue networks. European Journal of the Operation Research 5 (1980), p. 276.

FREE 64 M.G. Freeman: The effect of breakdowns and interstage storage on production line capacity. The Journal of Industrial Engineering 15 (1964), p. 194.

FRLE 65 H.D. Friedman: Reduction methods for tandem queueing systems. Operations Research 13 (1965), p. 121.

GAVE 63 D.P. Gaver, Jr Competitive queueing: idleness probabilities under priority disciplines. Journal of the Royal Statistical Society 25 (1963) Series B, p. 489.

GAVE 73 D.P. Gaver, G.S. Shedler: Processor utilization in multiprogrammed systems via diffusion approximations. Operations Research 21 (1973), p. 569.

GELE 79 E. Gelenbe: On the optimum checkpoint interval. Journal of the ACM 26 (1979), p. 259.

GELE 80A E. Gelenbe, I. Mitrani: Analysis and synthesis of computer systems. Academic Press, London 1980.

GELE 80B E. Gelenbe, L. Labetoulle, R. Marie, M. Metivier, G. Pujolle, W. Stewart: Reseaux de files d'attente - Modelisation et traitement numerique. Hommes et Techniques, Monographies d'informatique de l'AFCET, Paris, 1980.

GITT 78 J.C. Gittins: A comparison of service disciplines for GI/G/n queues. Mathematische Operationsforschung und Statistik, serie Optimization 9 (1978), p. 255.

GNED 71 B.W. Gniedienko, I.N. Kowalenko: Wstęp do teorii masowej obsługi. PWN, Warszawa 1971.

GOLD 77 H.M. Goldberg: **Analysis** of the earliest due date scheduling rule in queueing systems. Mathematics of Operations Research 2 (1977), p. 145.

GORD 67 W.I. Gordon, G.F. Newell: Closed queueing systems with exponential servers. Operations Research 15 (1967), p. 254.

GRAH 78 G.S. Graham: Queueing network models of computer systems performance. Computing Surveys 10 (1978), p. 219.

GRAN 75 D. Granot, F. Granot, A. Lemoine: Approximations for a service system with nonindependent interarrival times. Operations Research 23 (1975), p. 162.

GROS 74 D. Gross, C.M. Harris: Fundamentals of queueing theory. John Wiley & Sons, New York, 1974.

GROS 75 D. Gross: Sensitivity of queueing models to the assumption of exponentiality. Naval Research Logistics Quarterly 22 (1975), p. 271.

GWOZ 81 K. Gwoździ: Metoda przybliżonej analizy hierarchicznych systemów masowej obsługi. (W przygotowaniu-wersja wstępna w [ZSAK 80]).

HADI 72 N. Hadidi: On the output process of a state - dependent queue. Skand. Aktuaridskr 55 (1972), p. 182.

HALA 74 C. Halatsis, D. Maritsas, G. Philokypron: Analysis of blockorganized buffers for Poisson input and periodic block output. International Journal of System Science 5 (1974), p. 29.

HARR 67 C.M. Harris: Queues with state - dependent stochastic service rates. Operations Research 15 (1967), p. 117.

HARR 70 C.M. Harris: Some results for bulk-arrival with state - dependent service times. Management Science 16 (1970), p. 313.

HARR 73 J.M. Harrisor: The diffusion approximation for tandem queues in heavy traffic. Journal of Applied Probability 10 (1973) p. 613.

HATC 69 J.M. Hatcher: The effect of internal storage on the production rate of a series of stages having exponential service times. Transactions of the AIIE 1 (1969), p. 150.

HAYD 72 B.J. Haydon: System states for a series of finite queues, Operations Research 20 (1972), p. 1137.

HEID 80 P. Heidelberger: Variance reduction techniques for the simulation of Markov Processes I - multiple estimates. IBM Journal of Research and Development 24 (1980), No 5, p. 570.

HEYM 80 D.P. Heyman: The relation between customer and time averages in queues. Operations Research 28 (1980), p. 983.

HILD 67 D.K. Hildebrand: Stability of a finite queue, tandem server systems. Journal of Applied Probability 4 (1967), p. 571.

HILD 68 D.K. Hildebrand: On the capacity of tandem server finite queue service systems. Operations Research 16 (1968), p. 72.

HILL 66 F.S. Hillier, R.W. Boling: The effect of some design factors on the efficiency of production lines with variable operation times. The Journal on Industrial Engineering 17 (1966), p. 651.

HILL 67 F.S. Hillier, R.W. Boling: Finite queues in series with exponential or Erlang service times - a numerical approach: Operations Research 15 (1967), p. 286.

HILL 79 F.S. Hillier, R.W. Boling: On the optimal allocations of work in symmetrically unbalanced production line system with variable operation times. Management Science 25 (1979), p. 721.

HOKS 75 P. Hokstad: The G/M/m queue with finite waiting room. Journal of Apolied Probability 12 (1975), p. 779.

HUNT 56 G.C. Hunt: Sequential arrays of waiting lines. Operations Research 4 (1956), p. 674.

IGLE 80 D.L. Iglehart, G.S. Shedler: Regenerative simulation of response times in networks of queues. Springer Verlag 1980.

IRLA 77 M.I. Irland, G. Pujolle: Models of tandem networks of queues motivated by packet switching. IRIA Rapport de Recherche No 256, (1977).

IZYD 71A J. Izydorczyk: Zadanie konserwatora - analiza wpływu zmian parametrów systemu obsługi na jego charakterystyki. Podstawy Sterowania 11 (1981), s. 77.

IZYD 81B J. Izydorczyk: Warunki ergodyczności wielokanałowych systemów obsługi z regulaminem naturalnym i czasem obsługi zależnym od czasu oczekiwania.(w przygotowaniu - wersja wstępna w [ZSAK 79]).

JACK 54 R.R.P. Jackson: Queueing systems with phase type service. Operations Research Quarterly 5 (1954), p. 109.

JACK 57 J.R. Jackson: Networks of waiting lines. Operations Research 5 (1957), p. 516.

JACK 63 J.R. Jackson: Job-shop like queueing systems. Management Science 10 (1963), p. 131.

JACO 77 P.A. Jacobs, P.A. Lewis: A mixed autoregressive - moving average exponential sequence and point process (EARMA 1.1). Advances in Applied Probability 9 (1977), p. 87.

JACO 78 P.A. Jacobs: A cyclic queueing network with dependent exponential service times. Journal of Applied Probability 15 (1978), p. 573.

JACO 80 P.A. Jacobs: Havy traffic results for single - server queues with dependent (EARMA) service and interarrival times. Advances in Applied Probability 12 (1980), p. 517.

JAIS 69 N.K. Jaiswal: Priority queues. Academic Press, New York, 1968.

KADO 76 J.M. Kadosch: Temps d'attente dans les transports urbains en commun. RAIRO 10 (1976), Recherche Operationelle No 2, p. 37.

KALA 73 R. Kala, G.G. Hitchings: The effects of performance time variances on balanced four - station manual assembly line. International Journal of Production Research 11 (1973), p. 341.

KARL 68 S. Karlin: A first course in stochastic processes. Academic Press, New York 1968 (tłum. rosyjskie **MIR**, Moskwa 1971).

KAWA 75 T. Kawashima: Reverse ordering of services in tandem queues. Memoirs of the Defense Academy, Japan 15 (1975), p. 151.

KELL 74 T.W. Keller: ASQ user's manual. University of Texas at Austin, Dept. of Computer Scinces, TR-27, 1974.

KELL 75 F.P. Kelly: Networks of queues with customers of different types. Journal of Applied Probability 12 (1975), p. 542.

KELL 76 F.P. Kelly: Networks of queues. Advances in Applied Probability 8 (1976), p. 416.

KIFF 56 J. Kiefer, J. Wolfowitz: On charakteristic of the general queueing process with application to random walks. The Annales of Mathematical Statistics 27 (1956), p. 147.

KING 62 J.F. Kingman: On queues in heavy traffic. Journal of the Royal Statistical Society ser. B 24 (1962), p. 383.

KLEI 64 L. Kleinrock: Communication nets - stochastic message flow and delay. Mc Graw-Hill Book Company N.Y., 1964.

KLEI 65 L. Kleinrock: A conservation law for a wide class of queueing disciplines. Naval Research Logistics Quarterly 12 (1965), p. 181.

KLEI 70 L. Kleinrock: Swap time considerations in time shared systems. IEEE Transactions on computer C-19 (1970), p. 435.

KLEI 75 L. Kleinrock: Queueing systems - Vol.1: Theory. J. Wiley & Sons. New York 1975.

KLEI 76 L. Kleinrock: Queueing systems - Vol. 2: Computer applications. J. Wiley & Sons New York, 1976.

KLEI 80 L. Kleinrock, Y. Yemini: Interfering queueing processes in packet - switched broadcast communication. Proceedings of the IFIP Congress, 1980 North Holland, Amsterdam, 1980.

KNOT 78 A.D. Knott: The inefficiency of a series of work stations - a simple formula. International Journal of Production Research 8 (1978), p. 108.

KOBA 77 H. Kobayashi, A. Konheim: Queueing models for computer communications system analysis. IEEE Transaction on Communications 25 (1977), p. 2.

KOBA 79 H. Kobayashi: Modeling and analysis - an introduction to system performance evaluation methodology. Addison-Wesly Publishing Company, London 1979.

KOEN 56 E. Koenigsberg: Production lines and internal storage. Management Science 5 (1956), p. 410.

KONH 74 A.G. Konheim, B. Meister: Waiting lines and times in a system with polling. Journal of the ACM 21 (1974), p. 470.

KONH 76 A.H. Konheim, M. Reiser: A queueing model with finite waiting room and blocking. Journal of the ACM 23 (1976), p. 328.

KONI 79 D. König, D. Stoyan: Metody teorii obsługi masowej. WNT, Warszawa, 1979.

KONS 75 M. Konsek: Pamięci obrotowe ze stałymi i ruchomymi głowicami w systemach wieloprogramowych. Podstawy Sterowania 5 (1975), s. 241.

KOOP 72 B.O. Koopman: Air-terminal queues under time-dependent conditions. Operations Research 20 (1972), p. 1089.

KOPO 73A B. Kopociński: Zarys teorii odnowy i niezawodności. PWN, Warszawa 1973.

KOPO 73B I.Kopocińska: The repairman problem. Dissertationes Mathematicae C I PWN, Warszawa 1973.

KRAE 70 S.A. Kraemer, R.F. Love: A model for optimizing the buffer inventory storage size in a sequential production system. Transactions of the AIIE 11 (1970), p. 64.

KRAM 75 W. Krämer: Total waiting time distribution function and the fate of customer in a system with two queues in series. Elektronische Rechenanlagen 17 (1975), p. 18.

KRZE 77 A. Krzesiński, P. Tunissen: Stochastic network analysis program - user's manual. Department of Com. Science, University of Stellenbosch (RSA), Report RW 76-02, March 1977.

KUCZ 73 A. Kuczura: Batch input to a multiserver queue with constant service times. The Bell Systems Technical Journal 52 (1973), p. 83.

KUHN 76 P. Kuhn: Analysis of complex queueing networks by decomposition. Proceeding of the 8-th International Teletraffic Congress, Melbourne, 1976.

LABE 76 J. Labetoulle, P. Pujolle: A study of queueing networks with deterministic service and application to computer networks, Acta Informatica 7 (1976), p. 183.

LABE 77 J. Labetoulle, G. Pujolle: Modeling of packet switching communication networks with finite buffer size at each node; (in: Computer Performance. K.M. Chandy, M. Reiser (edit.)) N. Holland, 1977.

LABE 79A J. Labetoulle, G. Pujolle, C. Soula: Distribution of flows in a general Jackson network. IRIA, Rapport de Recherche No 341, Janvier, 1979.

LABE 79B J. Labetoulle, G. Pujolle, C. Soula: Note on the M/M/1 feedback process. IRIA Rapport de Recherche No 348. Mars 1979.

LAMS 77 S.S. Lam: Queueing networks with population size constraints. IBM Journal of Research and Development, July 1977, p. 370.

LASL 78 G.M. Laslett, O.B. Pollard, R.L. Tweedie: Techniques for establishing ergodic and recurrence properties of constinuous - valued Markov chains. Naval Research Logistics Quarterly 25 (1978), p. 455.

LATO 80 G. Latouche, M. Neuts: "Efficient algorithmic solutions to exponential tandem queues with blocking. SIAM Journal of Algebraic and Discrete Methods 1 (1980), p. 93.

LAVE 75 S.S. Lavenberg: The steady-state queueing time distribution for the M/G/1 finite capacity queue. Management Science 21 (1975), p. 501.

LAVE 78 S.S. Lavenberg: Stability and maximum departure rate of certain open queueing networks having finite capacity constraints. RAIRO, Informatique 12 (1978), p. 353.

LAVE 79A S. Lavenberg, M. Reiser: The state seen by an arriving customer in closed multiple chain queueing networks. IBM Research Report, 1979.

LAVE 79B S.S. Lavenberg, T.L. Moeller, C.H. Sauer: Con-comitant control variables applied to the regenerative simulation of queueing systems. Operations Research 27 (1979), p. 134.

LEHT 78 T. Lehtonen, E. Arjas: Approximating many server queues by means of single server queues II. Research Raport, Helsinki School of Economics, 1978.

LEMO 77 A.J. Lemoine: Networks of queues - a survey of equilibrium analysis. Management Science 24 (1977), p. 464.

LENY 80 L.M. Le Ny: Etude analytique de reseaux de files d'attente multiclasses a routages variables. RAIRO, Recherche Operationelle 14 (1980), p. 331.

LEVI 71 P.A. Levis, G.S. Shedler: Empirically derived micromodels for sequences of page exceptions. IBM Journal of Research and Development 17 (1973), p. 86.

LIBU 71 M. Libura: O jednokanałowym systemie masowej obsługi z zależnością czasu obsługi od czasu oczekiwania na jej rozpoczęcie. Archiwum Automatyki i Telemechaniki 16 (1971), s. 279.

LIBU 72 M. Libura: Dwukanałowy system obsługi z wymianą priorytetów i czasem obsługi zgłoszeń zależnym od stanu procesora. Archiwum Automatyki i Telemechaniki 17 (1972), s. 189.

LIND 52 D.V. Lindley: The theory of queues with a single server. Proceedings of the Cambridge Philosophical Society 48 (1952), p. 277.

LITT 61 J.D. Little: A proof for queueing formula: $L = \lambda W$. Operations Research 9 (1961), p. 383.

LITT 63 J.D.C. Little, K.G. Murty: An algorithm for the travelling salesman problem. Operations Research 11 (1963), p. 972.

LOKI 78 F.C.J. Lokin: Procedures for traveling salesman problems with additional constraints. European Journal of Operational Research 3 (1978), p. 135.

LOYN 62 R.M. Loynes: The stability of a queue with non-independent interarrival and service times. Proceedings of the Cambridge Philosophical Society 58 (1962), p. 497.

LOYN 64 R.M. Loynes: The stability of a system of queues in series. Proceedings of the Cambridge Philosophical Society 60 (1964), p.569.

MARI 78A R. Marie: Methodes iteratives de resolution de modeles mathematiques de systemes informatiques. RAIRO - Informatique 2 (1978), p. 75.

MARI 78B R. Marie: Modelisation par reseaux de files d'attente. These de grade de Docteur - Science Mathematiques, Universite de Rennes, Novembre 1978.

MASS 74 J. Masso, M.L. Smith: Interstage storages for three stage lines subject to stochastic failures. AIIE Transactions 6 (1974), p. 354.

MCGE 76 G.R. Mc Gee, D.B. Webster: An investigation of a two-stage production line with normaly distributed interarrival and service time distributions. International Journal of Production Research 14 (1976), p. 251.

MEIS 58 T. Meisling: Discrete-time queueing theory. Operations Research 6 (1958), p. 96.

MELA 79 B. Melamed: Charakterizations of Poisson traffic streams in Jackson queueing networks. Advances of Applied Probability 11 (1979), p. 422.

MERL 78 D. Merle: Contribution a l'ctude d'un analyseur de modeles a reseaux de files d'attente-algorithmes de calcul et applications. These presente a l'Institut National Polytechnique de Lorraine pour obtenir le diplome de Docteur-ingenieur, 1978.

MEVE 68 P. Mevert: A priority system with setup times. Operations Research 16 (1968), p. 602.

MICH 72 J.A. Michel: Synthesis of priority queueing systems. PhD. Dissertation, Pensylwania State University, 1972.

MINE 76 H. Mine, K. Ohno, T. Koizumi: A single - server queue with service times depending on the order of services. Operations Research 24 (1976), p. 188.

MINH 80 Do le Minh: Analysis of exceptional queueing system by the use of regenerative process and analytical methods. Mathematics of the Operations Research 5 (1980), p. 147.

MIRA 63 N.M. Mirasol: The output of an M/G/∞ queueing system is Poisson. Operations Research 11 (1963), p. 282.

MITC 77 C.R. Mitchell, A.S. Paulson, C.A. Beswick: The effect of correlated exponential service times on single server tandem queues. Naval Research Logistics Quarterly 24 (1977), p. 95.

MITC 79 C.R. Mitchell, A.S. Paulson: M/M/1 queues with interdependent arrival and service processes. Naval Research Logistics Quarterly 26 (1979), p. 47.

MIYA 77 M. Miyazawa: Time and customer processes in queues with stationary inputs. Journal of Applied Probability 14 (1977), p. 349.

MIYA 79 M. Miyazawa: A formal approach to queueing processes in the steady state and their applications. Journal of Applied Probability 16 (1979), p. 332.

MOHA 72 C. Mohan, K. Murari: Time dependent solution to a correlated queueing problem with variable capacity. Metrika 19 (1972), p. 209.

MOOR 75 S.C. Moore: Approximating the behaviour of nonstationary single-server queues. Operations Research 23 (1975), p. 1011.

MORI 80 M. Mori: Relations between queue - size and waiting - time distributions. Journal of Applied Probability 17 (1980), p. 822.

MUNT 72 R.R. Muntz: Poisson departure processes and queueing networks. IBM Research Report RC 4145, December, 1972.

MUNT 78 R.R. Muntz: Queueing networks: A critique of the state of the art and directions for the future. Computing Serveys 10 (1978), p. 353.

MURA 72A K. Murari: Time dependent solution of a queueing problem with correlated batch arrivals and general service time distribution. Metrika 19 (1972), p. 201.

MURA 72B K. Murari: A queueing problem with arrival and service in batches of variable size. Metrika 19 (1972), p. 27.

MUTH 73 E.J. Muth: The production rate of a series of work stations with variable service times. International Journal of Production Research 11 (1973), p. 155.

MUTH 77 E.J. Muth: Numerical methods applicable to a production line with stochastic servers; in: TIMS Studies in the Management Sciences 7 (1977), North Holland, p. 143.

MUTH 79 E.J. Muth: The reversibility property of production lines. Management Science 25 (1979), p. 152.

NEUT 66 M.F. Neuts: The single server queue with Poisson input and Semi-Markov service times. Journal of Applied Probability 3 (1966), p. 202.

NEUT 68 M.F. Neuts: Two queues in series with a finite intermediate waiting room. Journal of Applied Probability 5 (1968), p. 123.

NEUT 71 M.F. Neuts: A queue subject to extraneous phase changes. Advances in Applied Probability 3 (1971), p. 78.

NEUT 76 M.F. Neuts: Some explicit formulas for the steady-state behaviour of the queue with Semi-Markovian service times. Department of Statistics, Perdue University, Mimeograph series No 444, 1976.

NEUT 77 M.F. Neuts: The M/G/1 queue with several types of customers and changeover times. Advances of Applied Probability 9 (1977), p. 604.

NEUT 78A M.F. Neuts: The M/M/1 queue with randomly varying arrival and service rates. Opsearch 15 (1978), p. 140.

NEUT 78B M.F. Neuts: Further results on the M/M/1 queue with randomly varying rates. Opsearch 15 (1978), p. 159.

NEUT 79 M.F. Neuts, D.M. Lucantoni: A markovian queue with N servers subject to breakdowns and repairs. Management Science 25 (1979), p. 849.

NEWE 66 G.F. Newell: The M/G/∞ queue. SIAM Journal of Applied Mathematics 14 (1966), p. 86.

NEWE 68 G.F. Newell: Queues with time - dependent arrival rates. Journal of Applied Probability 5 (1968), Part I p. 436, Part II p. 579, Part III p. 591.

NISH 77 T. Nishida, K. Yoneyama: Correlated multi-server queue with random input. Math. Japonica 22 (1977), p. 395.

NOET 79 A.S. Noetzel: A generalized queueing discipline for product form network solutions, Journal of the ACM 26 (1979), p. 779.

OKAM 77 K. Okamura, H. Yamashina: Analysis of the effect of buffer storage capacity in transfer line systems. AIIE Transactions 9 (1977), p. 127.

OKAM 78 K. Okamura, H. Yamashina: Justification for installing buffer stocks in balanced and unbalanced flow-line production systems. Typescript, to appear in AIIE Transactions.

PACK 75 C.D. Pack: The output of a M/D/1 queue. Operations Research 23 (1975), p. 750.

PACK 77 C.D. Pack: The output of a D/M/1 queue. SIAM Journal of Applied Mathematics 32 (1977), p. 571.

PACK 78 C.D. Pack: The output of a multiserver queueing systems. Operations Research 26 (1978), p. 492.

PAKE 69 A.G. Pakes: Some conditions for ergodicity and recurrence of Markov chains. Operations Research 17 (1969), p. 1058.

PALM 47 C. Palm: The distribution of repairmen in servicing automatic machines. Industridiningen Norden 75 (1947), p. 75.

PANW 79 S.S. Panwalkar, M.L. Smith: A predictive equation for avarage output of stage series systems with finite interstage queues. AIIE Transactions 10 (1978), p. 215.

PAPO 72 A. Papoulis: Prawdopodobieństwo, zmienne losowe i procesy stochastyczne. WNT, Warszawa 1972.

PATT 64 E.L. Patterson: Markov processes occuring in the theory of traffic flow through an N- stage stochastic service system. Journal of Industrial Engineering 15 (1964), p. 188.

PAYN 72 S. Payne, N. Slack, R. Wild: A note on the operating charakteristics of balanced and unbalanced production flow lines. International Journal of Production Research 10 (1972), p. 93.

PEAR 67 C. Pearce: Queues with moving avarage service times. Journal of Applied Probability 4 (1967), p. 553.

PEAR 68 C. Pearce: Imbedded chain approach to a queue with moving average input. Operations Research 16 (1968), p. 1117.

PERR 81 H.G. Perros: A two level open queue network with blocking and feedback. RAIRO, Recherche Operationnelle 15 (1981), p. 27.

POTI 76 D. Potier, E. Gelenbe, J. Lenfant: Adaptive allocation of central processing unit quanta. Journal of the ACM 23 (1976), p. 97.

POSN 73 M. Posner: Single-server queues with service time dependent on waiting time. Operations Research 21 (1973), p. 610.

PUJO 76 G. Pujolle: Files d'attente en serie et application au taux de charge maximale d'un reseau d'ordinateurs. RAIRO 10 (1976), Informatique, No 5, p. 35.

PUJO 78 G. Pujolle: Applications of some Markov chains results to computer system modelling. IRIA Rapport de Recherche No 289 (1978).

PUJO 79A G. Pujolle, Ch. Soula: A study of flows in queueing networks and an approximate method for solution. INRIA Rapport de Recherche No 340, Janvier 1979.

PUJO 79B G. Pujolle: The influence of protocols on the stability conditions in packet - switching networks. IEEE Transactions on communications, 27, (1979).

PUJO 79C G. Pujolle, D. Potier: Reseaux de files d'attente a capacite limitee avec des application aux systemes informatiques. Rapport de Recherche No 219, (1977), to appear in RAIRO, 1979.

PUJO 80 G. Pujolle, A. Wolisz: Performance analysis of a two-computer system. Rapport de Recherche ENST - C - 80002, ENST Department d'Informatique, Paris, 1980.

PUJO 80B G. Pujolle: Reseux de files d'attente à forme produit. RAIRO Recherche Operationelle 14 (1980), p. 317.

PUJO 80C G. Pujolle: Comparison of some end-to-end flow control policies in a packet switching network. Podstawy Sterowania 10 (1980), p. 159.

PURD 74 P. Purdue: The M/M/1 queue in a Markovian environment. Operations Research 22 (1974), p. 562.

PURD 75 P. Purdue: A queue with Poisson input and semi-Markov service times: busy period analysis. Journal of Applied Probability 12 (1975), p. 353.

RAJU 77 S.N. Raju, U.N. Bhat: Recursive relations in the computation of the equilibrium results of finite queues. TIMS Studies in the Management Sciences 7 (1977), North Holland, Publishing Company, Amsterdam. p. 247.

RAMA 80 V. Ramaswami, D.M. Lucantoni: Algorithmic analysis of dynamic priority queue - I. Technical Report 70 B, Applied Mathematics Institute, University of Delaware, Newark, Delaware.

RAON 75A N.P. Rao: On the mean production rate of two-stage production system of the tandem type. International Journal of Production Research 13 (1975), p. 207.

RAON 75B N.P. Rao: Two - stage production systems with intermediate storage, AIIE Transactions 7 (1975), p. 414.

RAON 76A N.P. Rao: A viable alternative to the "method of stages" solution of series production systems with Erlang service times. International Journal of Production Research 14 (1976), p. 699.

RAON 76B N.P. Rao: A generalization of the bowl phenomenom in series production systems. International Journal of Production Research 14 (1976), p. 437.

RAYA 79A T.E. El-Rayah: The efficiency of balanced production lines. International Journal of Production Research 17 (1979), p. 61.

RAYA 79B T.E. El-Rayah: The effect of inequality of interstage buffer capacities and operation time variability on the efficiency of production line systems. International Journal of Production Research 17 (1979), p. 77.

REEV 73 N.R. Reeve, W.H. Thomas: Balancing stochastic assembly lines. AIIE Transactions 5 (1974), p. 223.

REIC 57 R. Reich: Waiting times when queues are in tandem. The Annales of Mathematical Statistics 28 (1957), p. 768.

REIC 63 E. Reich: Note on queues in tandem. The Annales of Mathematical Statistics 34 (1963), p. 338.

REIS 74 M. Reiser, H. Kobayashi: Accuracy of diffusion approximation for some queueing systems. IBM Journal of Research and Development 18 (1974), p. 110.

REIS 75A M. Reiser, H. Kobayashi: Queueing networsk with multiple closed chains: theory and computational algorithms. IBM Journal of Research and Development 19 (1975), p. 283.

REIS 75B M. Reiser: QNET 4 user's guide. IBM Research Report RA 71, New York (1975).

REIS 76 M. Reiser, H. Kobayashi: On the convolution algorithm for separable queueing networks. Proceeding of the International Symposium on Computer Performance Modeling, Measurement and Evaluation, Cambridge Massachusets, March 1976, p. 109.

REIS 77 M. Reiser: Numerical methods in seperable queueing networks. TIMS Studies in the Management Science 7 (1977), p. 113.

REIS 78A M. Reiser, S.S. Lavenberg: Mean value analysis of closed multichain queueing networks. IBM Research Report RC 7023, New York, 1978.

REIS 78B M. Reiser: Mean value analysis of queueing networks, a new look at an old problem. IBM Research Report RC 7228, New York 1978.

REIS 78C M. Reiser, C.H. Sauer: Queueing network models: methods of solution and their program implementation; in: K.M. Chandy, R.T. Yeh editors: Current Trends in Programming Methodology Vol. III: "Software Modeling and Its Impact on Performance", Prentice Hall, 1978.

REIS 78D M. Reiser, A.G. Konheim: Finite capacity queueing systems with applications in computer modeling. SIAM Journal of Computing 7 (1978), p. 210.

ROSE 75 M. Rosenshine: Queueing theory; the state-of-the art. AIIE Transactions 7 (1975), p. 257.

RUDI 76 M. Rudin: An introduction to flow control. Proceedings of the Third International Conference on Computer Communication, Toronto, 3-6 August 1976, p. 463.

RUSC 81 M. Ruschitzka: Policy function scheduling. Performance Evaluation 1 (1981), p. 31.

SAAT 61 T.L. Saaty: Elements of queueing theory with applications. Mc Graw Hill, New York, 1961 (tłum. ros. Moskwa, 1971).

SAKA 71 M. Sakata, S. Noguchi, J. Oizumi: An analysis of the M/G/1 queue under round - robin scheduling. Operations Research 19 (1971) p. 371.

SANK 77 G. Sankavanarayanan, R. Ramanarayanan: On correlated life and repair times. Mathematische Operationsforschung und Statistik, serie Optimization 8 (1977), p. 127.

SARH 62 A.E. Sarhan, B.G. Greenberg: Contributions to order statistics. J. Wiley & Sons, New York, 1962, pp. 343, 431.

SAUE 75 C.H. Sauer, K.M. Chandy: Approximate analysis of central server models. IBM Journal of Research and Development 19 (1975), p. 301.

SAUE 77 C.H. Sauer, E.A. Mac Nair: Computer - communication system modeling with extended queueing networks. IBM Research Report RC-6654 (July 1977).

SAUE 78 C.H. Sauer: Simultaneous resource possesion in queueing models of computers. IBM Research Report RC 6971 (Febr. 1978).

SEVC 72 K.C. Sevcik: Optimal processor scheduling when service times are hyperoxponentially distributed and preemption overhead is not negligible. Technical Report CSRG 16, University of Toronto, 1972.

SEVC 74 K.C. Sevcik: Scheduling for minimum total loss using service time distribution. Journal of the ACM 21 (1974), p. 65.

SEVC 77 K.C. Sevcik, A.I. Levy, S.K. Tripathi, J. Zahorjan: Improving approximations of aggregated queueing network subsystems; in: Computer Performance K.M. Chandy, M. Reiser (editors). North Holland 1977, p. 1.

SEVC 79 K. Sevcik, I. Mitrani: The distribution of queueing network states at input and output instants. Proceedings of the 4 th International Symposium on Performance of Computer System Modeling and Performance Evaluation. Vienna, February 1979.

SCHA 73 R. Schassberger: Warteschlangen. Springer Verlag, Wien-New York, 1973.

SCHA 78 R. Schassberger: The insensitity of stationary probabilities in networks of queues. Advances in Applied Probability 10 (1978), p. 906.

SCHM 78 V. Schmidt: On some relations between stationary time and customer state probabilities for queueing systems G/GI/s/r. Mathematische Operationsforschung und Statistik, Serie Optimization 9 (1978), p. 261.

SCHR 68A A. Scherr: An analysis of time shared computer systems. MIT Press, Cambridge Massachusets, Research Monograph No 38, 1968.

SCHR 68B L. Schrage: A proof of the optimality of the shortest remaining processing time discipline. Operations Research 16 (1968), p. 687.

SCHR 69 L. Shrage: A mixed - priority queue with application to the analysis of real-time systems. Operations Research 17 (1969), p. 728.

SCHR 70 L. Schrage: An alternative proof of a conservation law for the queue G/G/1. Operations Research 18 (1970), p. 185.

SCHR 74 L. Schrage: Optimal scheduling disciplines for a single machine under various degrees of information. Working Paper, Graduate School of Business, University of Chicago, 1974.

SCHW 76 P.J. Schweitzer: Maximum troughput in finite capacity open queueing networks. IBM Research Report RC 5996 (May 1976), również Management Science 24 (1977), p. 217.

SEID 79 J. Seidler: Analiza i synteza sieci łączności dla systemów teleinformatycznych. PWN Warszawa, 1979.

SHAR 73 O.P. Sharma: A model for queues in series. Journal of Applied Probability 10 (1973), p. 691.

SHIM 78 D.G. Shimshak: Optimal sequencing of servers in the two station series queueing system. Journal of the Operational Research Society 29 (1978), p. 779.

SHIM 79 D.G. Shimshak: A comparison of waiting time approximations in series queueing systems. Naval Research Logistics Quarterly 26 (1979), p. 499.

SHOR 80 J.E. Shore: The lazy repairman and other models of performance collapse due to overhead in simple, single server queueing systems. Proceedings of Performance 80, the 7 th IFIP W.G. 7.3. International Symposium on Computer Performance Modelling, Measurement and Evaluation, Toronto, Canada, 1980, p. 275.

SHUM 76 A.W.C. Shum: Queueing models of computer systems with general service time distributions. Ph. D. Thesis, Harvard University, Cambridge Massachusets, 1976.

SIMO 73 A. Simonovitz: Direct comparison of different priority queueing disciplines. Studia Scientiarum Mathematicarum Hungarica 1975, p. 225.

SKIN 67 C.E. Skinner: A priority queueing system with server-walking time. Operations Research 15 (1967), p. 278.

SLAC 72 N.D.C. Slack: The Weibull distribution and its use in describing service time distributions. Production Management Research Group Monograph 72/1, University of Bradford (1972).

SLAC 75 N.D. Slack, R. Wild: Production flow line and collective working: a comparison. International Journal of Production Research 13 (1975), p. 411.

SMIR 65 N.W. Smirnow, I.W. Dunin-Barkowski: Kurs rachunku prawdopodobieństwa i statystyki matematycznej. PWN Warszawa, 1973.

SMIT 77 L.D. Smith, P. Brumbaugh: Allocating interstation inventory capacity in unpaced production lines with heteroscedastic processing times. International Journal of Production Research 15 (1977), p. 163.

SOND 79 D. Sonderman: Comparing multi-server queues with finite waiting rooms II, different numbers of servers. Advances in Applied Probability 11 (1979), p. 448.

STEW 78 W.J. Stewart: A comparison of numerical techniques in Markov modelling. Communications of the ACM 21 (1978), p. 144.

STOY 78 D. Stoyan: Queueing networks - Insensitivity and heurustic approximation. Elektron-Informationsverarbeitung, Kybernetik 11 (1978).

SUBB 66 S. Subba Rao: Balking and reneging in waiting line systems. Ph. D. Thesis, University of Delhi, Delhi, India 1966.

SUGA 65 S. Sugawara, M. Takahashi: On some queues occuring in an intergrated iron and steel-works. Journal of the Operations Research Society of Japan 8 (1965), p. 16.

SUZU 74 T. Suzuki, T. Kawashima: Reduction methods for tandem queueing system. Journal of the Operations Research Society of Japan 17 (1974), p. 133.

SWAR 80 B.G. Swartz: Polling in a loop system. Journal of the ACM 27 (1980), p. 42.

SWEN 59 O. Swensen: An approach to a class of queueing problems. Operations Research 7 (1958), p. 276.

SYKE 70 J.S. Sykes: Simplified analysis of an alternating-priority queueing model with setup times. Operations Research 18 (1970), p. 1182.

TAKA 62 L. Takacs: Introduction to the theory of queues. Academic Press, New York, 1962.

TECH 80 Instrukcje technologiczne walcowni półwyrobów. Kombinat Metalurgiczny Huta Katowice, Dąbrowa Górnicza, Ameks, 1980.

TEMB 74 S.V. Tembe, R.W. Wolff: The optimal order of service in tandem queues. Operations Research 22 (1974), p. 824.

TOMK 75 J. Tomko: Processor utilization study. Computers & Mathematics with Applications 1 (1975), p. 337.

TRIP 79 S.K. Tripathi: On approximate solution techniques for queueing network models of computer systems. Ph.D. Thesis, Technical Report CSRG - 106, University of Toronto, September 1979.

TRUS 75 A.L. Truslove: Queue length for the $E_r/G/1$ queue with finite waiting room. Journal of Applied Probability 12 (1975), p. 215.

TWEE 75 R.L. Tweedie: Sufficient conditons for ergodicity and recurrence of Markov chains on a general state space. Stochastic Processes and their Applications 3 (1975), p. 385.

IWEE 76 R.L. Tweedie: Criteria for classifying general Markov chains. Advances in Applied Probability 8 (1976), p. 737.

VASI 77 O.A. Vasicek: An inequality for the variance of waiting time under a general queueing discipline. Operations Research 25 (1977), p. 879.

WART 79 A. Wartak: Optymalizacja grubości przeniesienia walcowanych pasm według kryterium minimalnego czasu walcowania. Podstawy Sterowania 9 (1979), s. 45.

WEGR 74 S. Węgrzyn: Analiza pracy systemów operacyjnych o regulaminie naturalnym z ograniczeniami na jednorazowy czas zajmowania procesora liczącego. Podstawy Sterowania 4 (1974), s. 219.

WEGR 75 S. Węgrzyn, A. Mrózek, A. Wolisz: Wielokanałowe, równoległe systemy obsługi. Podstawy Sterowania 5, 1975, s. 299.

WEGR 78 S. Węgrzyn: Podstawy Automatyki. Wyd. IV. PWN, Warszawa 1978.

WELC 64 P.D. Welch: On a generalized M/G/1 queueing process in which the first customer of each busy period receives exceptional service. Operations Research 12 (1964), p. 736.

WHIT 72 W. Whitt: Embedded renewal processes in the GI/G/S queue. Journal of Applied Probability 9 (1972), p. 650.

WILD 72 R. Wild, N.D. Slack: The operating charakteristics of "single" and "double" nonmechanical flow systems. International Journal of Production Research 10 (1972), p. 139.

WILL 74 T.R. Willemain: Approximate analysis of general queueing networks. Operations Research 22 (1974), p. 522.

WOLI 73 A. Wolisz, T. Czachórski: Systemy obsługi priorytetowej zbioru źródeł jednowymiarowych. Podstawy Sterowania 3 (1973), s. 291.

WOLI 75 A. Wolisz: A single server priority queueing model of realtime operating systems. Proceedings of the International Conference Informatica 75, Bled 1975.

WOLI 76A A. Wolisz: Szeregowanie zadań czasu rzeczywistego przy ograniczeniu przerwań priorytetowych. Praca doktorska, Wydział Automatyki i Informatyki Pol. Śląskiej, Gliwice 1976.

WOLI 76B A. Wolisz: Jednokanałowy, wielofazowy system masowej obsługi jako model systemu operacyjnego czasu rzeczywistego. Podstawy Sterowania 6 (1976), s. 119.

WOLI 78 A. Wolisz: A detailed system overhead including description of priority algorithms for real-time task scheduling; in: M. Arato, E. Knuth (editors): Selected Paper on Operating Systems. KSH SZAMKI, Budapeszt 1978.

WOLI 79 A. Wolisz, J. Izydorczyk: Systemy masowej obsługi z czasem obsługi zależnym od czasu oczekiwania. Podstawy Sterowania 9 (1979), s. 259.

WOLI 80A A. Wolisz: Optymalizacja wydajności dwuetapowych systemów obsługi z ograniczoną kolejką pośrednią. Podstawy Sterowania 10 (1980), s. 3.

WOLI 80B A. Wolisz, J. Izydorczyk: The influence of interarrival time and service time variability on the characteristics of single server queueing systems with waiting time dependent service times. Podstawy Sterowania 10 (1980), p. 241.

WOLI 81A A. Wolisz: Throughput optimization of multistage systems with finite intermediate storage. Collection Seminaires: Modelisation et Evaluation des Systemes Informatiques, INRIA 1981. p. 165.

WOLI 81B A. Wolisz: Queueing systems stability condition estimation using digital simulation. Proceedings of the III Hungarian Computer Science Conference, Budapest 1981, p. 265.

WOLI 81C A. Wolisz: Przykłady optymalizacji dwuetapowych linii produkcyjnych. Zeszyty Naukowe Politechniki Śląskiej seria Informatyka z. 3, s. 55.

WOLI 81D A. Wolisz, J. Izydorczyk: Analysis of single - server queueing systems with waiting time dependent service times under the FIFO discipline. Maszynopis, wersja wstępna w [ZSAK 79].

WOLF 80 R.W. Wolff: Work conserving priorities. Journal of Applied Probability 7 (1970), p. 327.

WONG 78 J.W. Wong: Queueing network modeling of computer communication networks. Computing Surveys 10 (1978), p. 334.

YECH 71 U. Yechiali, P. Naor: Queueing problems with heterogeneous arrivals and service. Operations Research 19 (1971), p. 722.

YAMA 75 G. Yamazaki, H. Sakasegawa: Properties of duality in tandem queueing system. Annales of the Institut of Statistical Mathematics 27 (1975), p. 201.

YAMA 78 G. Yamazaki, H. Sakasegawa, T. Kawashima: Production rate estimated with flow - shop reversibility. Bulletin of the JSME 21 (1978), p. 161.

ZAHO 77 J. Zahorjan, A.I. Levi: An overview of the QSOLVE system. Technical Report CSRG-83, University of Toronto, Computer System Research Group, Toronto, July 1977.

ZAHO 79 Z. Zahorjan: Computational algorithm for queueing networks with product form solutions; in G. Scott Graham (editor): Topics in Performance Evaluation. Technical Report CSRG-100 University of Toronto, Computer Systems Research Group, Toronto, July 1979.

ZSAK 79 A. Wolisz, K. Gwóźdź, J. Izydorczyk, A. Krawet, J. Mierzwa, A. Wartak: Analiza i optymalizacja systemów obsługi z czasem obsługi zależnym od czasu oczekiwania na przykładzie procesu nagrzewania wlewków w piecach wgłębnych Kombinatu Metalurgicznego Huta Katowice. Opracowanie ZSAK PAN, Gliwice, 1979.

ZSAK 80 A. Wolisz, K. Gwóźdź, J. Izydorczyk, A. Wartak, A. Nowakowski: Podstawy i doświadczalny rozwój metod komputerowego kierowania procesami dyskretnymi na przykładzie rejonu pieców wgłębnych w Hucie Katowice. Opracowanie ZSAK PAN, Gliwice, 1980.

ALEK 74 А.М. Александров: О системе массового обслуживания с повторными требованями. Известия АН СССР, Техническая Кибернетика 1974, Но 2, стр. 86.

BORO 72 А.А. Боровков: Вероятностные процессы в теории массового обслуживания. Наука, Москва, 1972.

BROD 73 С.М. Броди, И.А. Погосян: Вложенные стохастические процессы в теории массового обслуживания. Наукова думка, Киев, 1973.

BRON 76 О.И. Бронштейн, И.М. Духовный: Модели приоритетного обслуживания в информационно-вычислительных системах. Наука, Москва, 1976.

DIMI 79 М.Ц. Димитров: Однолинейная система массового обслуживания с групповым обслуживанием в режиме разделения процессора. Известия АН СССР, Техническая Кибернетика 1979, Но 1, стр. 102.

DUNI 55 И.В. Дунин-Барковский, Н.В. Смирнов: Теория вероятностей и математическая статистика в технике. Москва, 1955.

FALI 79A Г.И. Фалин: Модель спаренного включения при наличии повторных вызовов. Известия АН СССР, Техническая Кибернетика 1979, Но 1, стр. 85.

FALI 79B Г.И. Фалин: Однолинейная система с повторными вызовами. Известия АН СССР, Техническая Кибернетика 1979, Но 2, стр. 107.

FALI 80 Г.И. Фалин: Повторные вызовы в структурно-сложных системах, Известия АН СССР, Техническая Кибернетика 1980, Но 6, стр. 55.

GNED 73 Б.В. Гнеденко, Э.А. Даниелян, Б.Н. Петров, Г.Г. Климов, В.Ф. Матвеев: Приоритетные системы обслуживания. Издательство Московского Университета, Москва 1973.

GROM 79 А.И. Громов: Двухфазная система массового обслуживания с ограниченным промежуточным накопителем и гиперэрланговским обслуживанием специального вида. Известия АН СССР, Техническая Кибернетика 1979, Но 1, стр. 92.

HAJK 73 Б.Е. Хайкин, И.Г. Степанов: Вероятностный анализ функционирования системы стрипперное отделение – нагревательные колодцы – обжимный стан. Известия Высших Учебных Заведений Чёрная Металлургия 1973, Сообщение 1, Но 2 стр. 179; Сообщение 2, Но 10, стр. 177.

HAJK 74 Б.Е. Хайкин, В.М. Зидензон: Исследование работы прокатного комплекса с позиции теории массового обслуживания. Известия Высших Учебных Заведений, Чёрная Металлургия 1974, Сообщение 1, Но 10 стр. 88; Сообщение 2, Но 12, стр. 69.

KREJ 79 А.Я. Крейнин: Об одной системе массового обслуживания. Межвузовский Сборник "Прикладная математика и задачи железнодорожного транспорта", Московский Институт Инж. Железнодорожного Транспорта, Труды Но 640 1979 , стр. 84.

KREJ 80 А.Я. Крейнин: Одноканальная система массового обслуживания с разогревом. Автоматика и Телемеханика 1980. Но 6, стр. 42.

MALI 71 И.И. Маликов: Исследование работы некоторых участков системы сталь-прокат методами теории массового обслуживания. Сталь 1971. Но 11, стр, 1044.

MARD 80 В.В. Мардух: Об однолинейной системе массового обслуживания с учётом времени переключения. Известия АН СССР, Техническая Кибернетика 1980, Но 4, стр. 97.

MISZ 74 Г.К. Мишкой: Система обслуживания с ориентацией и абсолютным приоритетом. Идентичное обслуживание заново прерванного вызова. Известия АН СССР, Техническая Кибернетика 1974, Но 6, стр. 95.

MISZ 78 Г.К. Мишкой: К задаче обслуживания со смешанным приоритетом в системах с потерями на переключение. Автоматика и Телемеханика 1978, Но 2, стр. 15.

MORO 80 Е.В. Морозов: О предельном распределении виртуального и реального времени ожидания в системе M/G/1 с обратной связью, Весці Акадэміі Навук БССР, Серия физ-мат навук, 1980, Но 3, стр. 55.

MOWA 77 В.В. Мова, Л.А. Пономаренко, А.М. Калиновский: Организация приоритетного обслуживания в АСУ, Техника, Киев 1977.

MUDR 61 В.И. Мудров: Очередь с "нетерпливыми" клиентами и переменным временем обслуживания, линейно зависящим от времени пробывания клиента в очереди. Проблемы кибернетики 1961 Но 5, стр. 283.

PUGA 73 В.Н. Пугачёв: Комбинированные методы определения вероятностных характеристик, Москва, Советское Радио", 1973.

SZNE 68 М.А. Шнепс-Шнеппе: Однолинейные системы с зависимостью времени обслуживания от времени ожидания, Новости Медицинского Приборостроения, Выпуск Первый, Москва, 1968.

WILE 73 С.Я. Виленкин, В.С. Елин: Исследование двухфазной системы массового обслуживания с блокировкой. Автоматика и Телемеханика 1973, Но 12, стр. 66.

WILE 75 С.Я. Виленкин, В.С. Елин: Решение задачи теории массового обслуживания для двухфазной системы с блокировкой. Автоматика и Телемеханика 1975, Но 8, стр. 22.

WISZ 79 В.М. Вишневский, В.А. Жарких: Приоритетная система с групповым обслуживанием. Известия АН СССР, Техническая кибернетика 1979, Но 2, стр. 90.

WISZ 80 В. Вишневский, А.С. Твердохлебов: Модели замкнутых сетей с блокировками для анализа мультипрограммных вычислительных систем. Автоматика и Телемеханика 1980, Но 5, стр. 172.

WOLK 80 М.И. Волковинский: Модели n-уровневого приоритетного обслуживания с учётом потерь на переключения. Автоматика и Телемеханика, 1980, Но 5, стр. 23.

NIEKTÓRE WAŻNIEJSZE PUBLIKACJE DOTYCZĄCE TEMATYKI OMAWIANEJ W NINIEJSZEJ PRACY, KTÓRE UKAZAŁY SIĘ PO JEJ ZŁOŻENIU DO DRUKU

BURM 81 D.Y. Burman: Insensitivity in queueing systems, Advances in Applied Probability 13 (1981) p. 846.

CALO 81 S.B. Calo: Delay analysis of a two-queue, nonuniform message channel, IBM Journal of Research and Development 25 (1981) p. 915.

CARO 82 J.L. Carrol, A. van de Liefvoort, L. Lipsky: Solutions of M/G/1//N - type loops with extensions to M/G/1 and GI/M/1 queues, Operations Research 30 (1982) p. 490.

DISN 81 R. Disney: A note on sojourn times in M/G/1 queues with instantaneous, Bernoulli feedback, Naval Research Logistics Quarterly 28 (1981) p. 679.

JACO 82 P.A. Jacobson, E.D. Lazowska: Analyzing queueing networks with simultaneous resource possession, Communications of the ACM 25 (1982) p. 142.

KELL 82 F.P. Kelly: The throughput of a series of buffers, Advances in Applied Probability 14 (1982), p. 633.

KOEN 82 E. Koenigsberg: Twenty five years of cyclic queues and closed queue networks: a review, Journal of the Operational Research Society 33 (1982) p. 605.

REIS 81 M. Reiser: Mean-value analysis and convolution method for queue-dependent servers in closed queueing networks, Performance Evaluation 1 (1981) p. 7.

NIUS 81 S. Niu: On queues with dependent interarrival and service times, Naval Research Logistics Quarterly 28 (1981) p. 497.

NUMM 81 E. Nummelin: Regeneration in tandem queues, Advances in Applied Probability 13 (1981) p. 221.

PERR 81 H.G. Perros: A symmetrical exponential open queue network with blocking and feedback, IEEE Transactions on Software Engineering 7 (1981) p. 395.

PINE 82 M. Pinedo, R.W. Wolff: A comparison between tandem queues with dependent and independent service times, Operations Research 30 (1982) p. 464.

SAUE 81 C. Sauer: Approximate solution of queueing networks with simultaneous resource possession, IBM Journal of Research and Development 25 (1981) p. 894.

SHIM 82 D.G. Shimshak, G.P. Sphicas: Waiting time in a two station series queueing system: the effect of dependent interarrival times, Journal of the Operational Research Society 33 (1982) p. 745.

WOLF 82 R.W. Wolff: Tandem queues with dependent service times in light traffic, Operations Research 30 (1982) p. 619.

WYBRANE, NIEKONWENCJONALNE SYSTEMY MASOWEJ OBSŁUGI

S t r e s z c z e n i e

Zdecydowana większość bardzo obszernej i szybko rosnącej literatury w zakresie systemów masowej obsługi dotyczy bądź to systemów jednostanowiskowych, bądź sieci stanowisk obsługi o charakterze konwencjonalnym, to jest takich, w których odstępy czasu między chwilami napływania kolejnych zgłoszeń oraz czasy ich obsługi tworzą ciągi niezależnych zmiennych losowych o jednakowych rozkładach, i w których pomija się czas pracy własnej potrzebny na implementację regulaminu obsługi.

Jakiekolwiek odstępstwo od tych założeń, a występują one często przy modelowaniu systemów teleinformatycznych oraz dyskretnych systemów produkcyjnych, prowadzi każdorazowo do ilościowych a często również jakościowych zmian we własnościach systemu, włączając w to nawet zmianę tak podstawowej charakterystyki jaką jest warunek stabilności.

Niniejsza praca, traktująca o pewnych, wybranych niekonwencjonalnych systemach masowej obsługi, składa się z czterech rozdziałów.

W rozdziale I sprecyzowano pojęcie "niekonwencjonalnych systemów masowej obsługi" i przedstawiono przegląd takich systemów (zarówno jednostanowiskowych, jak i sieci stanowisk obsługi) w oparciu o zaproponowaną klasyfikację. Uzasadniono celowość rozpatrywania każdego z typów tych systemów, omówiono jego zasadnicze własności (posługując się również wynikami obliczeń numerycznych) i zacytowano literaturę przedmiotu.

W rozdziałach II i III skupiono się na rozwinięciu metod analizy i optymalizacji jednego z najtrudniejszych typów niekonwencjonalnych systemów masowej obsługi - systemów z czasem obsługi zależnym od czasu oczekiwania. Opierając się zarówno na wcześniej opublikowanych pracach (w tym poprzednich pracach autora), jak też na nowych wynikach przedstawiono najbardziej wyczerpujące studium z tego zakresu.

Ze względu na uzależnienie warunku stabilności w takich systemach od regulaminu obsługi oraz na różny zakres możliwych do uzyskania wyników analitycznych, oddzielnie rozpatrzono przypadek naturalnego regulaminu obsługi (rozdział II) i dowolnego regulaminu obsługi (rozdział III).

W pierwszym przypadku przedstawiono zarówno warunki stabilności, jak i metodykę ścisłej analizy przy dosyć ogólnych założeniach. Wydaje się, że to zagadnienie można więc uznać za prawie całkowicie rozwiązane, aczkolwiek użycie przedstawionej metodyki jest pracochłonne.

W drugim przypadku, dla którego dostępne są jedynie bardzo nieliczne wcześniejsze wyniki, rozważano trzy zagadnienia:
- ocenę warunków stabilności,
- analizę w stacjonarnym stanie granicznym,
- optymalizację regulaminu obsługi

przy użyciu zarówno metod analitycznych, ich połączenia z modelowaniem cyfrowym, jak też pewnych metod heurystycznych.
Uzyskano szereg wyników o różnym stopniu ogólności.

W rozdziale IV rozpatrywane jest zagadnienie optymalizacji wydajności wieloetapowych systemów masowej obsługi. Badano wpływ takich parametrów, jak sposób rozdziału pracy między etapy, postać rozkładów czasów obsługi, liczba kanałów obsługi zainstalowanych na poszczególnych etapach i pojemności buforów pośrednich na wydajność systemu, przedstawiając zalecenia dotyczące optymalnego doboru tych parametrów.
Rozstrzygnięto niektóre z istniejących dotąd niejasności w tym zakresie.
Ze względu na dużą liczbę dołączonych wyników numerycznych można łatwo zorientować się w skutkach nieprawidłowego doboru parametrów.

W dołączonym do pracy dodatku wskazano, na przykładzie procesu przemysłowego, iż wykorzystanie omawianych w niniejszej pracy niekonwencjonalnych systemów masowej obsługi umożliwia wyjaśnienie pewnych - pozornie zaskakujących i zdecydowanie niepożądanych - stanów obiektu, a także zaproponowanie sposobów zapobiegania im.

НЕКОТОРЫЕ НЕКОНВЕНЦИОНАЛЬНЫЕ СИСТЕМЫ МАССОВОГО ОБСЛУЖИВАНИЯ

Р е з ю м е

Большинство работ из очень обширной и быстро развивающейся литературы по системам массового обслуживания касается или одностадиовых систем или сети конвенциональных систем обслуживания т.е. таких, в которых интервалы времени между моментами прихождения очередных заявок, а также время их обслуживания являются последовательностями независимых, одинаково распределенных случайных величин, и, в которых минуется собственное время работы, нужное на имплементацию дисциплины обслуживания.

Всякое отклонение от этих положений, а выступают они часто при моделировании сетей информационной связи вычислительных машин, а также в дискретных производственных системах, ведут всегда к количественным, а часто также к качественным отклонениям характеристик системы, включая в то даже изменение такой основной характеристики, какой является условие стабильности.

Данная работа, в которой рассматриваются некоторые неконвенциональные системы массового обслуживания состоит из четырёх глав.

В первой главе определено понятие "неконвенциональных систем массового обслуживания" и представлен обзор этих систем на основании предложенной квалификации.

Доказана целесообразность рассмотра каждого из видов этих систем, обсуждены их основные характеристики при использовании также результатов численных данных и сделаны ссылки на применение в работе ранее опубликованной литературы по данному предмету.

Во второй и третьей главах автор работы сосредотачивает своё внимание на развитие методов анализа и оптимализацию одного из самых сложных типов неконвенциональных систем массового обслуживания - систем с временем обслуживания зависящим от времени ожидания.

Опираясь на ранее опубликованные работы (в том числе на предыдущие работы автора), а также на новейшие исследования в этой области представлен наиболее исчерпывающий анализ этих систем.
В связи с зависимостью условия стабильности в таких системах от дисциплины обслуживания, а также от разной степени получения возможных аналитических результатов, отдельно был рассмотрен случай прямого порядка обслуживания и произвольной дисциплины обслуживания. (Главы вторая и третья).
В первом случае были представлены как условия стабильности так и методы точного анализа при довольно общих предпосылках. Нам кажется, что эту про-

блему можно считать почти целиком решённой, хотя применение представленного метода анализа является трудоёмким процессом.

Во втором случае, при котором имелись в наличае довольно ограниченные, ранее полученные результаты, были рассмотрены три вопроса:
- оценка условий стабильности
- анализ в стационарном предельном состоянии
- оптимализация дисциплины обслуживания

с применением одинаково как аналитических методов, их комбинаций с численным моделированием, так и некоторых евристических методов.
Было получено ряд результатов с разной степенью общности.

В четвёртой главе рассматривалась проблема оптимализации производительности многоэтапных систем массового обслуживания. Исследовано было влияние таких параметров как способ распределения работы по этапам,распределение времени обслуживания, численность каналов обслуживания на любом, произвольном этапе и ёмкость промежуточных буферов, с представлением рекомендаций для оптимального выбора этих параметров. Были выяснены некоторые из, существующих до сих пор, неясности в этих рекомендациях.
В связи с большим количеством подаваемых численных результатов можно легко сориентироваться в результатах неправильного выбора параметров.

В приложении, на примере промышленного производства показано, что использование обсужденых в данной работе неконвенциональных систем массового обслуживания даёт возможность выяснения некоторых, на первый взгляд неожиданных и решительно нежелательных явлений объекта, а также предложены способы для их избежания.

SOME NON-CLASSICAL QUEUEING SYSTEMS

S u m m a r y

The vast majority of the enormously large, and fast growing queueing systems literature is concerned with either single servers or queueing networks of the classical type, meaning such ones in which system parameters like demands interarrival times and service times are assumed to form sequences of independent, identically distributed random variables and no system overhead, due to service discipline implementation is included.
Any violation of these assumptions, which occur frequently while modelling computer-communication systems and production systems, results always in quantitative and usually also in qualitative changes of system behaviour, including possibly even changes of such basic characteristic as stability condition.

This paper dealing with some, chosen types of non-classical queueing systems consists of four chapters.

In chapter I the term "non-classical queueing system" is precisely introduced, and a survey of such systems (single-servers, multi-servers and queueing networks) following an introduced classification is presented. For every type of these systems the justification of it's consideration is given, it's main features are discussed (including numerical data), and the relevant references are mentioned.

In chapters II and III methods of analysis and optimal design for one of the most difficult types of non-classical queueing systems namely queueing systems with waiting time dependent service times are developed. Basing both on so far published results (including earlier papers by the author) and newly reported progress, the most complete study of such systems is given.
Because of the dependence of stability conditions upon the queueing discipline applied, and different scope of possible analytical solutions, the case of FIFO queueing discipline (chapter II) and general queueing discipline (chapter III) are treated seperately.

In the first case both the stability donditions and methodology for exact analysis are presented for fairly general assumptions. Thus this problem seems to be almost completly solved, although the use of presented methodology is laborious.

In the second case, with very limited earlier results available, three problems:

- stability condition estimation
- steady - state analysis
- queueing discipline optimization

have been approached, using equally purely analytical methods, combined analytical and simulational methods as well as some heuristics.

Several results of different generality have been obtained.

In chapter IV the problem of throughput optimization in multistage queueing systems is discussed. The influence of work division among different stages, service time distribution, the number of parallel servers installed at a single stage as well as intermediate buffer capacity on system's throughput is investigated, and recommendations for best choice of these parameters are developed.

Some so far existing ambiguities in this area are cleared. Due to many numerical results the loss of throughput caused by non-optimal choice of parameters is well illustrated.

In an appendix it is shown, using an industrial example, that utilizing the nonclassical queueing system discussed earlier, it is possible to explain some - seemingly surprising and definitly undesirable behaviour patterns of the plant, and suggest ways of their avoiding.

BIBLIOTEKA GŁÓWNA
Politechniki Śląskiej
P 4193 82 4

Cena zł 181,–

WYDAWNICTWA NAUKOWE I DYDAKTYCZNE POLITECHNIKI ŚLĄSKIEJ MOŻNA NABYĆ W NASTĘPUJĄCYCH PLACÓWKACH:

44-100 Gliwice — Księgarnia nr [illegible], ul. Konstytucji [illegible]
44-100 Gliwice — Spółdzielnia Studencka, ul. Wrocławska 4 a
40-950 Katowice — Księgarnia nr 015, ul. Żwirki i Wigury 33
40-096 Katowice — Księgarnia nr 005, ul. 3 Maja 12
41-900 Bytom — Księgarnia nr 048, Pl. Kościuszki 10
41-500 Chorzów — Księgarnia nr 063, ul. Wolności 22
41-300 Dąbrowa Górnicza — Księgarnia nr 081, ul. ZBoWiD-u 2
47-400 Racibórz — Księgarnia nr 148, ul. Odrzańska 1
44-200 Rybnik — Księgarnia nr 162, Rynek 1
41-200 Sosnowiec — Księgarnia nr 181, ul. Zwycięstwa 7
41-800 Zabrze — Księgarnia nr 230, ul. Wolności 288
00-901 Warszawa — Ośrodek Rozpowszechniania Wydawnictw Naukowych PAN — Pałac Kultury i Nauki

Wszystkie wydawnictwa naukowe i dydaktyczne zamawiać można poprzez Składnicę Księgarską w Warszawie, ul. Mazowiecka 9.