ROBUSTNOST VYBRANÝCH SYSTÉMU HROMADNÉ OBSLUHY!

J.Michálek, KAM PřF UJEP Brno

V teorii hromadné obsluhy jsou zkoumány matematické modely reálných front požadavků či zákazníků, kteří přicházejí k lince obsluhy a zde čekají na svoji obsluhu. Bohužel, aby vůbec bylo možné popsat složité systémy reálných front matematickým modelem, je potřeba do modelu zabudovat celou řadu doplňujících a omezujících předpokladů. Jsou to zejména předpoklady kladené na tvar rozdělení a na nezávislost náhodných veličin, s nimiž pracujeme, na stacionaritu a ergodičnost procesů, které popisují systém apod. Jde o předpoklady, které v jednotlivých praktických situacích mnohdy nebývají splněny, a přesto aproximace reálných situací poskytnuté matematickým modelem dávají užitečné představy o chování reálného systému a o jeho sledovaných charakteristikách. Velmi důležitou úlohou teorie hromadné obsluhy je nalezení robustních metod, které by umožnily odhadnout práci složitých systémů jednoduchými metodami pomocí snadno zjistitelných charakteristik.

Ve svém příspěvku bych si chtěl všimnout robustnosti systémů hromadné obsluhy (dále SHO) a sítí hromadné obsluhy z hlediska závislosti na tvaru rozdělení náhodných veličin, které popisují jednak proces obsluhy a jednak proces příchodů požadavků.

1. Robustnost systému G/G/1

V tomto odstavci bych chtěl uvést některé přístupy z hledání robustních odhadů základních charakteristik, zejména pak doby čekání v systému G/G/1. Ve svém příspěvku budu předpokládat, že všechny uvažované SHO se nacházejí ve statistické rovnováze. Budeme uvažovat G/G/1 s režimem obsluhy v pořadí příchodů požadavků (FIFO) a označíme $A(t)$ distribuční funkci náhodné veličiny $\tilde{t}$, udávající délku časového intervalu mezi příchody požadavků, $B(x)$ distribuční funkci doby obsluhy $\tilde{x}$ libovolného požadavku. Potom doba čekání $\tilde{w}$ požadavku na obsluhu má rozdělení s distribuční funkcí $W(y)$ (viz.(7)), která je dána integrální rovnicí

$$W(y) = P(\tilde{w} \leq y) = \int_{-\infty}^{y} W(y-u)\, dC(u) \qquad y \geq 0$$
$$= 0 \qquad y < 0 \qquad (1)$$

kde $C(u)$ je distribuční funkce náhodné veličiny $\tilde{u} = \tilde{x} - \tilde{t}$.

Přestože vzorec (1) nebo některé jeho modifikace umožňují

v jednotlivých konkrétních případech stanovit rozdělení doby čekání $\tilde{w}$, je v obecném případě obtížné odvodit i základní charakteristiky tohoto rozdělení, např. stanovit jeho střední hodnotu apod. V obecném systému G/G/1 není vyjádření střední hodnoty $E\tilde{w}$ v obecném tvaru známé, lze ji vyjádřit pouze přes některé jiné charakteristiky, jejichž přímé stanovení je rovněž obtížné (viz (7)). Obecně lze říci, že jak rozdělení $\tilde{w}$, tak $E\tilde{w}$ silně závisí na A(t) a B(x). Bylo by proto velmi užitečné nalezení obecných postupů pro odhady těchto charakteristik, jež by na tvar A(t) a B(x) byly málo citlivé. Zejména je potřeba znát robustnost těchto odhadů. Ke stanovení odhadů lze uvést podle typu aproximace dva základní přístupy:

I. Aproximace hledaných charakteristik.

II. Aproximace zkoumaného systému.

I. přístup vychází ze základních vlastností náhodné veličiny $\tilde{w}$ a veličin, na nichž $\tilde{w}$ bezprostředně závisí a potom rozdělení $\tilde{w}$ nebo jeho střední hodnotu $E\tilde{w}$ vhodným způsobem aproximuje, nezávisle na tvaru distribučních funkcí A(t) a B(x). Tímto způsobem byly nalezeny robustní odhady distribuční funkce doby čekání W(y) pro případ, že koeficient využití systému $\rho = E\tilde{x}/E\tilde{t} \to 1$, tj. při velkém zatížení systému, a byla nalezena aproximace W(y) pro velké hodnoty y (viz.(6),(4)). Získané odhady jsou značně robustní a závisí pouze na prvních dvou momentech rozdělení A(t) a B(x). Dále v (5) a (10) byly nalezeny horní a dolní hranice pro střední dobu čekání $E\tilde{w}$, opět značně robustní.

II. přístup vychází z aproximace systému G/G/1 jiným systémem hromadné obsluhy, který lze popsat jednodušším nebo jiným způsobem známým matematickým modelem. Tyto modely bývají dvojího druhu:

a) Aproximace rozdělení A(t) a B(x) diskrétním rozdělením. Pro takto aproximovaný systém je snadné řešit rovnici (1). Těžiště problému je potom ve volbě aproximace rozdělení A(t) a B(x), aby se při této aproximaci nezměnila podstata řešení rovnice (1). Otázka takovéto aproximace je složitá, její řešení pro uvedené potřeby se teprve začíná zkoumat. Zatím lze doporučit provést aproximaci tak, aby souhlasilo pokud možno co nejvíce momentů diskrétního a aproximovaného rozdělení. V (15) a (8) jsou uvedeny některé výsledky, které ukazují na adekvátnost takovéhoto postupu. Kvalita této aproximace posuzovaná kvalitou odhadu střední doby čekání $E\tilde{w}$ je velmi dobrá,

b) Aproximace spojitými procesy.
Otázka této aproximace se v poslední době velmi intensivně zkoumá (8), kde lze nalézt bohaté odkazy na literaturu. Ukazuje se, že některé výsledky takovéto aproximace jsou velmi blízko výsledkům, získaným ad a). Pozornost je v tomto směru věnována zejména difuzním procesům.

2. Markovské sítě hromadné obsluhy a jejich robustnost.

Síť hromadné obsluhy budeme chápat jako několik vzájemně propojených SHO, tak, že požadavky, které po zpracování jedním SHO přicházejí ke zpracování do jiného SHO. Jednotlivé systémy budeme nazývat uzly sítě. Jednotlivé požadavky nemohou v žádném uzlu sítě ze sítě odcházet ani do ní přicházet - mluvíme o uzavřené síti. Je-li v každém uzlu sítě doba obsluhy požadavku exponenciélně rozdělená náhodná veličina, mluvíme potom o markovské síti.

Dále si všimneme jedné speciální sítě hromadné obsluhy, kterou lze dobře popsat pradikci výkonnosti počítačového systému (viz.(11)). Budeme předpokládat, že markovská síť je uzavřená a obsahuje právě N uzlů, i-tý uzel je tvořen m_i nezávislými paralelními obslužnými linkami, doba obsluhy na každé lince i-tého uzlu má exponenciální rozdělení se střední hodnotou $1/\mu_i$ $i=1,\dots,N$, stejné pro všechny linky daného uzlu. Dále budeme předpokládat, že v síti je právě K požadavků, každý požadavek po obsluze v i-tém uzlu přichází k obsluze na j-tém uzlu s pravděpodobností r_{ij} a $\Sigma_{j=1}^{N} r_{ij} = 1 \; \forall i$. Doba potřebná pro přechod požadavku z i-tého na j-tý uzel je nulová a požadavky, které při příchodu do uzlu naleznou všechny jeho linky obsazené, zaujmou místo ve frontě v pořadí příchodů (FIFO).

Označme $S(N,K) = \{(k_1,\dots,k_N): \Sigma_{i=1}^{N} k_i = K, k_i \in [0,\dots,K], i=1,..,N\}$ a $p(k_1,\dots,k_N)$ pravděpodobnost, že ve statistické rovnováze bude v popsané síti právě k_i požadavků v i-tém uzlu, $i=1,\dots,N$. Potom (viz.(2))

$$p(k_1,\dots,k_N) = \frac{1}{G(N)} \prod_{i=1}^{N} \frac{x_i^{k_i}}{\beta_i(k_i)} \qquad (2)$$

pro $(k_1,\dots,k_N) \in S(N,K)$, kde parametry $x_1,\dots,x_N$ musí vyhovovat soustavě lineárních rovnic

$$G(K) = \Sigma_{(k_1,\dots,k_N) \in S(N,K)} \prod_{i=1}^{N} \frac{x_i^{k_i}}{\beta_i(k_i)}$$

$$x_i \mu_i = \Sigma_{j=1}^{N} \mu_j x_j r_{ji} \qquad i=1,\dots,N .$$

kde $\beta_i(k_i) = k_i!$ pro $k_i \leq m_i$

$\beta_i(k_i) = m_i!\, m_i^{k_i - m_i}$ pro $k_i > m_i$, $i = 1,2,\ldots,N$

Ze vzorce (2) lze potom snadno stanovit další charakteristiky sítě. Např. střední počet požadavků v i-tém uzlu je roven

$$E\,\tilde{k}_i = \sum_{k_i=1}^{K} x_i^{k_i} \frac{G(K-k_i)}{G(K)} .$$

Otázkou je, jak je uvedená síť robustní z hlediska exponenciálního rozdělení doby obsluhy. V práci [1] byla robustnost této sítě sledována metodou Monte Carlo a bylo zjištěno, že jednotlivé pravděpodobnosti $p(k_1,\ldots,k_N)$ pro generovanou síť značně závisí na tvaru rozdělení doby obsluhy v jednotlivých uzlech (bylo generováno exponenciální, rovnoměrné a logaritmicko-normální rozdělení doby obsluhy, N = 3, K = 7).

Na druhé straně při aplikacích uvedené markovské sítě v počítačových systémech se ukázalo, že modelem získané analytické výsledky se dobře shodují s výsledky naměřenými, ačkoliv vesměs nebyly dodrženy předpoklady pro použití modelu (viz [14,] [12]). To by ukazovalo na dobrou robustnost popsané sítě vzhledem k uvedeným aplikacím.

Tak např. v práci [12] je ukázáno, že výsledky predikce střední doby odezvy u dvouprocesorového systému IBM 360/67 získané na základě popsané markovské sítě se od naměřených neliší o více než o 10%. Při tom byly hrubě porušeny předpoklady uvedené markovské sítě ve třech směrech:

P1 : všechny požadavky si vedou stochasticky stejně (předpoklad, že všechny pravděpodobnosti r_{ij} a rozdělení doby obsluhy v uzlech jsou pro všechny požadavky stejné)

P2 : předpoklad exponenciálního rozdělení doby obsluhy v uzlech

P3 : režim obsluhy v uzlech je režim v pořadí příchodů (FIFO)

Proto si dále všimneme sítě, která by porušení uvedených předpokladů respektovala a uvedeme, za jakých podmínek je tato síť robustní. Před popisem takové sítě uvedeme některé režimy obsluhy

užívané v počítačových systémech a modelování obecné doby obsluhy pomocí kaskády exponenciálních linek obsluhy.

3. Režimy obsluhy

V počítačových systémech, které pracují v režimu sdílení procesoru, jsou používány zejména takové režimy, které preferují požadavky, jež potřebují na svou obsluhu "krátký" čas před požadavky, které potřebují na svou obsluhu "dlouhý" čas. Takovouto prioritu je možné vyjádřit například tak, že doba čekání ve frontě $\tilde{w}(x)$ požadavku, který na svou obsluhu vyžaduje x časových jednotek, je úměrná x. Často užívané režimy obsluhy s touto vlastností jsou round robin (RR) a inversní pořadí obsluhy s absolutní prioritou a doobsloužením (LIFO).

Režim RR spočívá v tom, že požadavky se před linkou obsluhy (procesorem) staví do fronty v pořadí svých příchodů, každý požadavek, který je v čele fronty přechází pak na linku obsluhy, kde obdrží q časových jednotek obsluhy ($q > 0$ je daná konstanta tzv. kvantum) a potom se opět staví na konec fronty, není-li jeho obsluha ještě dokončena nebo v opačném případě odchází ze systému. V limitním případě $q \to 0$ dochází ke sdílení procesoru všemi požadavky přítomnými v systému.

Režim LIFO s absolutní prioritou a doobsloužením spočívá v tom, že každý nově příchozí požadavek do systému jde ihned na linku obsluhy, vytlačí z obsluhy požadavek právě obsluhovaný (za předpokladu, že linka obsluhy není volná). Vytlačený požadavek se postaví do čela fronty. Jakmile se linka obsluhy uvolní pro jeho obsluhu, začíná jeho obsluha z místa, kde byla přerušena.

Pro systém M/G/1 s režimy obsluhy RR a LIFO s absolutní prioritou a doobsloužením lze odvodit (viz [8]), že v době statistické rovnováhy platí

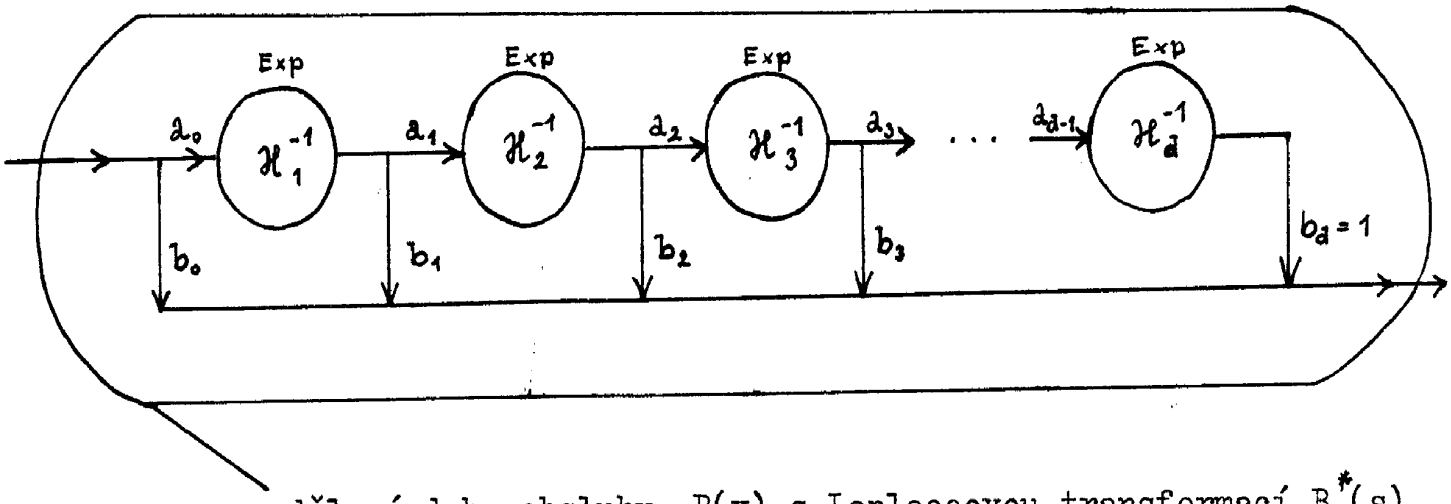
$$\tilde{w}(x) = \frac{\rho^x}{1-\rho}$$

4. Modelování obecné doby obsluhy pomocí kaskády exponenciálních linek obsluhy

Jestliže doba obsluhy v jednotlivých uzlech sítě nemá exponenciální rozdělení pravděpodobností, potom za předpokladu, že Laplaceova transformace hustoty rozdělení doby obsluhy (označme ji $B^*(s)$ pro daný uvažovaný uzel) je racionální funkcí, lze ukázat (viz např. [9]), že ji lze nahradit kaskádou exponenciálních rozdělení. Uvedená náhrada vychází z vyjádření $B^*(s)$ ve tvaru

$$B^*(s) = b_0 + \sum_{j=1}^{d} a_0 a_1 \dots a_{j-1} b_j \prod_{i=1}^{j} \frac{1}{1 + s\mu_i} \qquad (3)$$

kde d je stupeň polynomu ve jmenovateli $B^*(s)$, $-\mu_j^{-1}$, $j = 1,2,\dots d$, jsou jeho kořeny, polynom v čitateli $B^*(s)$ musí být stupně nejvýše d, $a_i \geq 0$, $b_i \geq 0$ jsou dané konstanty, $a_i + b_i = 1$ pro $i = 0,1,\dots d$, $b_d = 1$. Kaskáda d exponenciálních linek obsluhy s parametry μ_i^{-1}, $i = 1,2,\dots,d$ je potom znázorněna na obrázku 1.



rozdělení doby obsluhy $B(x)$ s Laplaceovou transformací $B^*(s)$

Obr. 1

Požadavek, který projde znázorněnou kaskádou exponenciálních rozdělení, kde parametry a_i jsou pravděpodobnosti přechodu po obsluze na i-tém uzlu k exponenciální obsluze na $i+1$ uzlu a parametry b_i jsou pravděpodobnosti odchodu požadavku ze systému po obsluze na i-tém uzlu, má, jak plyne ze vztahu (3), rozdělení doby obsluhy s distribuční funkcí $B(x)$.

5. Zobecněná robustní síť obsluhy

Popíšeme nyní síť obsluhy, která na základě předních dvou odstavců přihlédne k předpokladům P1 - P3, které byly při empirickém užití markovské sítě posané v odstavci 2 silně porušeny.

Předpokládejme, že daná síť sestává z N uzlů a že požadavky vytvářejí L různých tříd. Označme $r_{ij}(l)$ pravděpodobnost, že požadavek l-té třídy přejde po skončení obsluhy v uzlu i k obsluze do uzlu j. Proti odstavci 2 nebudeme předpokládat, že síť je uzavřená, označme proto $\gamma_i(l)$ parametr Poissonova procesu, jež popisuje proces příchodů požadavků l-té třídy do uzlu i a $r_i(l) = 1 - \sum_{j=1}^{N} r_{ij}(l)$ pravděpodobnost, že požadavek l-té třídy po ukončení obsluhy v i-tém uzlu opustí síť. (Pro uzavřenou síť zřejmě platí $\gamma_i(l) = 0$ a $r_i(l) = 0$, $l = 1,\ldots,L$, $i = 1,\ldots,N$). Dále pro každé $l \in \{1,2,\ldots,L\}$ označíme $e_i(l)$, $i = 1,2,\ldots N$ řešení soustavy rovnic

$$e_i(l) = \gamma_i(l) + \sum_{j=1}^{N} e_i(l) r_{ij}(l) \qquad (4)$$

Budeme rozlišovat 3 různé typy uzlů sítě:

Uzel typu 1: FIFO, /M/1 : Uzly tohoto typu obsahují jednu linku obsluhy s režimem FIFO, doba obsluhy na lince tohoto uzlu má exponenciální rozdělení pravděpodobnosti s parametrem μ_{ik}, který může být i funkcí počtu k požadavků v tomto uzlu (uzlu i).

Uzel typu 2: RR, /G/1 : Uzly tohoto typu obsahují jednu linku obsluhy s režimem RR. Doba obsluhy na lince tohoto uzlu má dané obecné rozdělení pravděpodobností s distribuční funkcí B(x), jejíž hustota má Laplaceovu transformaci $B^*(s)$ a může být zapsána ve tvaru (3).

Uzel typu 3: LIFO, /G/1: Uzly tohoto typu obsahují jednu linku obsluhy s režimem LIFO absolutní prioritou a doobsloužením. Doba obsluhy na lince tohoto uzlu má rozdělení pravděpodobností stejné jako v uzlu typu 2 (tj. vyhovuje (3)).

Pro takto zobecněnou síť bylo v [13] dokázáno, že pravděpodobnosti $p(k_1,\ldots,k_N)$, že v uzlu i je k_i požadavků, $i = 1,2,\ldots,N$ (bez rozlišení tříd), jsou tvaru

$$p(k_1,\ldots,k_N) = C \cdot h_1(k_1)h_2(k_2) \ldots h_N(k_N),$$

kde

$$h_i(k_i) = \left(\sum_1 \frac{e_i(1)}{\mu_{ik_i}} \right)^{k_i} \quad \text{pro uzel typu 1}$$

$$h_i(k_i) = \left(\sum_1 \frac{e_i(1)}{\mu_{i1}} \right)^{k_i} \quad \text{pro uzel typu 2 a 3}$$

přičemž součet je přes všechny třídy požadavků, které vstoupily do i-tého uzlu, $1/\mu_{i1}$ je střední doba obsluhy požadavku třídy 1 v uzlu i, $1/\mu_{ik_i}$ je střední doba obsluhy požadavku v uzlu i při exponenciálním rozdělení doby obsluhy, je-li přítomno k_i požadavků, C je daná konstanta.

Kromě toho, je-li systém otevřený, pak lze odvodit (viz [13]), že $p(k_1,\ldots,k_N) = \prod_{i=1}^{N} p_i(k_i)$, kde

$p_i(k_i) = (1 - \varrho_i)\varrho_i^{k_i}$ pro všechny tři uvedené typy uzlů, přičemž

$$\varrho_i = \sum_1 \frac{e_i(1)}{\mu_{i1}} \quad \text{pro } \mu_{i1} = \mu_{i2} = \ldots \text{ pro uzel typu 1}$$

$$\varrho_i = \sum_1 \frac{e_i(1)}{\mu_{i1}} \quad \text{pro uzel typu 2 a 3} \qquad (5)$$

Poslední výsledek je z hlediska robustnosti popisované sítě obzvlášť cenný, neboť ukazuje, že pro všechny tři typy uzlů je rozdělení počtu požadavků v každém uzlu stejné jako v systému M/M/1 pro parametr ϱ_i daný vzorcem (5). Jinými slovy, popsaná síť je vzhledem k rozdělení doby obsluhy značně robustní a chová se jako by uzly byly izolované nezávislé SHO s poissonovským procesem příchodů a exponenciální dobou obsluhy s intenzitou provozu ϱ_i danou vzorcem (5) v i-tém uzlu, i = 1,2,...,N.

Z těchto výsledků pak je zřejmá dobrá shoda empirických výsledků s markovskými sítěmi hromadné obsluhy při porušení předpokladů P1, P2 a P3. Náhradu obecného uzlu G/G/1 v této síti, jehož robustnost byla diskutována v 1.odstavci, za uzel M/M/1 je možno udělat na základě doplňujících předpokladů o režimu obsluhy.

Literatura:

1 Breznicka a.:Simulace vybraných sítí hromadné obsluhy. Diplomová práce UJEP,Brno 1978.

2 Gordon W.J.,Newell G.F.:Closed Queuing Systems with Exponential Servers,Operations Research 15,1967.

3 Harrison J.M.:The Heavy Traffic Approximation for Single Server Queues in Series,J. of Appl. Probability 10,No 3,1973.

4 Kingman J.F.C.:On Queues in Heavy Traffic,J. of the Royal Stat.Soc.,Series B,24,1962.

5 Kingman J.F.C.:Some Inequalities for the Queue GI/G/1, Biometrika 49 (1962).

6 Kingman J.F.C.:Inequalities in the Theory of Queues, J. of the Royal Stat. Soc.,Series B,32,1970.

7 Kleinrock L.:Queuing Systems,Vol. I,Theory,Wiley 1975.

8 Kleinrock L.:Queuing Systems,Vol. II,Computer Applications,Wiley 1975.

9 Kobayashi H.:System Design and Performance Analysis Using Analytic Models,in:Chandy&Yeh:Current Trends in Programming Methodology,Vol. III,New Jersey 1978.

10 Marshall K.T.:Some Inequalities in Queuing,Operations research 16,1968.

11 Michálek J.:Predikce výkonnosti počítačového systému, Informačné systémy 3,1979.

12 Moore C.G.:Network Models for Large-Scale Time-Sharing Systems,Technical Report No 71-1,Dept. of Industrial Engineering,Univ. of Michigan,Ann Arbor,Michigan 1979.

13 Munts R.R.&Baskett F.:Open,Closed and Mixed Networks of Queues with Different Classes of Customers,Techn.Rep. 33, Stanford Elektronics Lab.,Stanford University 1972.

14 Scherr A.A.:An Analysis of Time-Shared Computer Systems, MIT Press,Cambridge,Massatschusets 1967.

15. Wong D.K.:A Discrete Approximation og G/G/1 Queue, M.S.Theses,Computer Science Dept.,School of Engineering and Applied Science,Univ. of California,Los Angeles 1974.