STUDIA INFORMATICA 2001
Volume 22 Number 2 (44)

Franciszek GRABOWSKI, Wiesław STRĘCIWILK
Politechnika Rzeszowska, Zakład Systemów Rozproszonych

SKALOWALNOŚĆ WYDAJNOŚCI E-WĘZŁA BIZNESU

Streszczenie. Artykuł dotyczy nowej metody pomiaru jakości obsługi klienta w systemach klient-serwer bazującej na charakterystyce $\log(t_o)=f[\log(r)]$, gdzie t_o jest czasem odpowiedzi systemu na żądanie klienta, zaś r jest pozycją serwera WWW w rankingu popularności. Czas odpowiedzi systemu $t_o=t_w+t_k+t_s$, gdzie t_w jest czasem obsługi przez serwer, t_k czasem kolejkowania, zaś t_s czasem obsługi przez sieć. W modelu przyjęto, że t_w i t_s są niezależne od r. Z drugiej strony t_k, który zależy od r, określa rozkład czasu kolejkowania, tj. kształt charakterystyki $1/r$. Analiza wykazuje, że dyspersja t_k jest bardzo duża. Superpozycja t_k prowadzi do uzyskania liniowej charakterystyki $\log(t_o)=f[\log(r)]$ o nachyleniu równym -1. Gdy $r\rightarrow 1$, wtedy wykorzystanie serwera $U\rightarrow 1$, zaś $t_o\rightarrow\infty$. Natomiast, gdy pozycja serwera w rankingu jest niska, wtedy, $t_k\rightarrow 0$, zaś $t_o\rightarrow t_w+t_s$. Odchylenie od liniowej charakterystyki o nachyleniu -1 w dół oznacza skrócenie czasu obsługi klienta przez system, a tym samym poprawę jakości obsługi.

PERFORMANCE SCALABILITY OF BUSINESS E-SITE

Summary. The paper deals with a new measurements method of QoS in the client-server system based on measurement of the $\log(t_o)=f[\log(r)]$ characteristic, where t_o is the system response time but r is rank of WWW server. Response time $t_o=t_w+t_k+t_s$, where t_w is the server service time, t_k the waiting time, but t_s is the network service time. In this model has been assumed that network service time is constant, and independent on r. On the second hand the queue t_k, which depends on r, determines excess time i.e., $1/r$ characteristic form. Analysis shows that t_k dispersion is very large. Superposition of t_k leads to the linear characteristic $\log(t_o)=f[\log(r)]$ at slope equals -1. When $r\rightarrow 1$, then the utilization of the server $U\rightarrow 1$, but $t_o\rightarrow\infty$. On the other hand, when WWW server rank is low, $t_k\rightarrow 0$, but $t_o\rightarrow t_w+t_s$. Taking into account of the $\log(t_o)=f[\log(r)]$ characteristic it can be noticed that tendency to deviate down in comparison to slope equals -1 can be measure of QoS improve.

1. Wprowadzenie

Analizując systemy rozproszone można sobie wyobrazić idealny przypadek, kiedy wszystkie elementy mają nieskończenie wielką wydajność. Wtedy zadania byłyby realizowane w czasie równym zero. W innym, również idealistycznym, rozwiązaniu można by przyjąć, że wszystkie elementy mają identyczne, aczkolwiek skończone wydajności, zaś obciążenie jest rozłożone równomiernie. W rzeczywistości wszystkie elementy systemu mają różne i skończone wydajności, zaś obciążenie jest nierównomierne. Przyjmując to za punkt wyjścia, można wskazać dwa specyficzne obszary zastosowań systemów rozproszonych.

Pierwszy obejmuje systemy obliczeniowe [1,2]. Charakteryzują się one tym, że w przypadku jednorodności sprzętowej i programowej, czas realizacji każdego procesu, tj. czas obliczeń plus czas komunikacji, byłby identyczny. W ogólnym przypadku, tzn. dla niejednorodności sprzętowej i programowej, czas realizacji każdego procesu jest różny i dynamicznie się zmienia w bardzo krótkim czasie i jest trudny do przewidzenia. Ponadto rozkład czasów wykonania cechuje się niezbyt dużą dyspersją. Problem ten rozwiązuje się m.in. poprzez dynamiczne wyrównywanie obciążeń [1,2].

Drugi dotyczy szeroko postrzeganych internetowych systemów klient-serwer [3], np. typu audio i video na żądanie, telewizja interaktywna, nauczanie na odległość, systemy telekomunikacyjne czy też telemedycyna [4] wpisujących się w kategorię B2C e-biznesu. Równie rozlegle są pozostałe kategorie e-biznesu typu B2B, C2C, G2B i G2C. Wszystkie te kategorie charakteryzują się bardzo wielkim zróżnicowaniem obciążenia poszczególnych serwerów WWW w zależności od ich popularności. Dynamika zmian obciążeń poszczególnych serwerów w czasie stosunkowo długim, nie jest tak wielka jak w systemach obliczeniowych i jest do przewidzenia. Odnosi się to szczególnie do aplikacji e-biznesowych, np. kiedy po akcji promocyjnej danego produktu serwer oferujący go jest oblegany przez klientów. Z drugiej strony można wskazać serwery WWW, które są obciążone minimalnie. Jednakże w tym przypadku, w odróżnieniu od systemów obliczeniowych rozproszonych, obciążenia nie można wyrównać pomiędzy różnymi serwerami WWW. Można oczywiście zwiększać przepustowość danego węzła stosując wpisujące się w formuły przetwarzania równoległego czasowego lub przestrzennego różnorakie metody [5]. W tym przypadku czas odpowiedzi systemu na żądanie klienta można wyraźnie podzielić na czas obsługi przez zasoby systemu typu sieć oraz serwery i czas kolejkowania. Zakładając niezależność czasu obsługi od rankingu serwera WWW oraz zależność czasu kolejkowania od rankingu serwera WWW, można zaproponować metodę pomiaru jakości obsługi klienta.

Jako cel artykułu stawiamy sobie pokazanie podstawowych relacji rządzących obsługą masową klientów przez e-węzeł. W związku z tym przyjmujemy najprostszy model pojedynczej kolejki, który w kolejnych badaniach będzie rozwijany. Model uwzględnia

napływające procesy, obsługę procesów, przestrzeń przechowywania kolejki oczekujących żądań oraz liczbę zasobów, do których się procesy odwołują. Artykuł dotyczy nowej metody pomiaru jakości obsługi klienta w systemie klient-serwer bazującej na charakterystyce $\log(t_o)=f[\log(r)]$, gdzie t_o jest czasem odpowiedzi systemu na żądanie klienta, zaś r jest pozycją serwera WWW w rankingu popularności. Czas odpowiedzi systemu $t_o=t_w+t_k+t_s$, gdzie t_w jest czasem obsługi przez serwer, t_k czasem kolejkowania, zaś t_s czasem obsługi przez sieć. W modelu przyjęto, że t_w i t_s są niezależne od r. Z drugiej strony t_k, który zależy od r, określa rozkład czasu kolejkowania, tj. kształt charakterystyki $1/r$. Analiza wykazuje, że dyspersja t_k jest bardzo duża. Superpozycja t_k prowadzi do uzyskania liniowej charakterystyki $\log(t_o)=f[\log(r)]$ o nachyleniu równym -1. Gdy $r\rightarrow 1$, wtedy wykorzystanie serwera $U\rightarrow 1$, zaś $t_o\rightarrow\infty$. Natomiast, gdy pozycja serwera w rankingu jest niska, wtedy $t_k\rightarrow 0$, zaś $t_o\rightarrow t_w+t_s$. Odchylenie od liniowej charakterystyki o nachyleniu -1 w dół oznacza skrócenie czasu obsługi klienta przez system, a tym samym poprawę jakości obsługi.

W związku z tym w dalszej części artykułu omówiono kolejno istotę zagadnienia (punkt 2.1), czas obsługi klienta przez sieć (punkt 2.2) oraz czas obsługi klienta przez serwer (punkt 2.3).

2. Metoda pomiaru jakości obsługi klienta

2.1. Istota zagadnienia

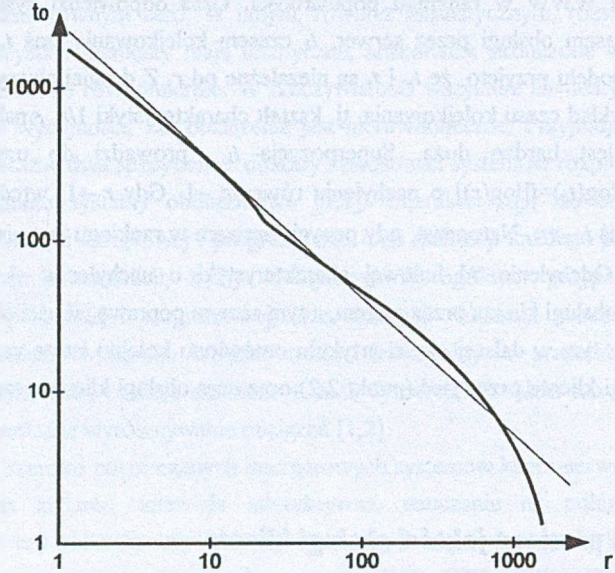
Transakcje w systemach e-biznesu są obsługiwane przez wiele różnych zasobów, takich jak serwery: WWW, FTP, baz danych, identyfikacji, płatności itp. W przypadku gdy zasób ma wystarczającą przepustowość w stosunku do zadań, które realizuje, wtedy wykonuje je w czasie charakterystycznym dla niego, zwanym czasem obsługi. W naszym przypadku dotyczy to serwera z czasem obsługi t_w i sieci z czasem obsługi t_s. Kiedy liczba zadań wzrasta i przekracza przepustowość zasobu, wtedy tworzy się kolejka, opóźniając wykonanie zadania o czas oczekiwania w kolejce t_k. Zatem czas odpowiedzi systemu t_o może być określony jako:

$$t_o = t_w + t_k + t_s \tag{1}$$

Czas $t_w+t_k=t_e$ jest czasem odpowiedzi serwera. Spójrzmy na system klient-serwer e-biznesu. Analiza danych eksperymentalnych sugeruje, że wiele zjawisk w WWW może być zamodelowanych prawem Zipfa [6]. Silnie zróżnicowana popularność rozmaitych dokumentów WWW jest dobrze znanym zjawiskiem, które można określić następującą zależnością:

$$P = \frac{k}{r} \tag{2}$$

gdzie P jest liczbą odwołań do dokumentu, którego pozycja w rankingu popularności jest r, zaś k jest stałą. Odchylenie od liniowej charakterystyki o nachyleniu –1 w dół oznacza skrócenie czasu obsługi klienta przez system, a tym samym poprawę jakości obsługi. Pokazano to na rys. 1.



Rys. 1. Zależność czasu odpowiedzi systemu klient-serwer, t_o, od rankingu serwera WWW, $\log(t_o)=f[\log(r)]$. Dane eksperymentalne [8]

Fig. 1. Dependence of the client-server system response time, t_o, versus rank of the WWW server, $\log(t_o)=f[\log(r)]$. Experimental data [8]

2.2. Czas odpowiedzi e-węzła – maksymalny czas odpowiedzi

Jako cel artykułu stawiamy sobie pokazanie podstawowych relacji rządzących obsługą masową klientów przez e-węzeł. W związku z tym przyjmujemy najprostszy model pojedynczej kolejki, który w kolejnych badaniach będzie rozwijany. Model uwzględnia napływające procesy, obsługę procesów, przestrzeń przechowywania kolejki oczekujących żądań oraz liczbę zasobów, do których się procesy odwołują. Zakładamy, że żądania wpływają do serwera z szybkością λ. Serwer obsługuje je z szybkością μ i odsyła. Proces ten można scharakteryzować grafem o k stanach, gdzie k=0, 1, ... Pomiędzy sąsiednimi stanami grafu są dwa przejścia. Jedno z nich skierowane ku serwerowi ilustruje nadchodzące żądania, drugie skierowane od serwera ku wejściu odpowiada obsłużonym żądaniom. Zakładamy równowagę operacyjną systemu, tzn. liczba wpływających żądań ze średnią szybkością λ jest równa liczbie obsłużonych żądań ze średnią szybkością μ. Kolejka jest scharakteryzowana

przez funkcję $S(n)$, która reprezentuje średni czas obsługi przypadający na jedno żądanie, przy czym długość kolejki jest n. Weźmy pod uwagę serwer oblegany przez bardzo dużą liczbę klientów, np. serwer poczytnego dziennika o zasięgu światowym. Wtedy proces przybywania klientów może być scharakteryzowany średnią częstością przybywania żądań λ. Ze względu na prosty model zakładamy, że czas obsługi przez zasób nie zależy od obciążenia systemu. Funkcja obsługi specyfikuje szybkość, z jaką są obsługiwane żądania przez serwer, tzn. jaka jest jego przepustowość. Tak więc średnia przepustowość serwera jest określona jako $X_O = \mu$. Należy zauważyć, że szybkość obsługi przez serwer nie jest wyłącznie funkcją ich charakterystyk fizycznych, np. liczby i typów procesorów, szybkości dysków, lecz zależy również od obciążenia systemów zarządzających bazami danych, złożoności operacji przeszukiwania itd.

W tym punkcie stawiamy sobie za cel określenie równania pozwalającego wyznaczyć najdłuższy czas odpowiedzi serwera na żądanie obsługi klienta. Kolejno więc będziemy wyznaczać współczynnik prawdopodobieństwa obciążenia serwera p, średnią liczbę żądań $\overline{N}$, współczynnik wykorzystania serwera U, przepustowość X, by wreszcie określić średni czas odpowiedzi serwera t_e na żądanie klienta.

Rozpoczynamy od określenia wartości p_k. Dla danego stanu zakładamy równowagę pomiędzy żądaniami wpływającymi i wypływającymi. Tak więc [7]

$$\begin{aligned} \mu p_1 &= \lambda p_0 \\ \mu p_2 &= \lambda p_1 \\ &\ldots\ldots \\ \mu p_k &= \lambda p_{k-1} \end{aligned} \tag{3}$$

Łącząc wyrażenie (3), otrzymujemy

$$p_k = \frac{\lambda}{\mu} p_{k-1} = \frac{\lambda}{\mu}\left(\frac{\lambda}{\mu} p_{k-2}\right) = \ldots = p_0\left(\frac{\lambda}{\mu}\right)^k, \; k=1, 2, 3 \ldots \tag{4}$$

Weźmy pod uwagę p_k jako funkcję p_0 dla wszystkich wartości k=1, 2, 3 ... . W tym celu określmy p_0. Serwer może być w danej chwili w jednym z możliwych dopuszczalnych stanów. Tak więc suma ułamków czasu, w których serwer jest w jakimkolwiek możliwym stanie, od 0 do ∞, jest równa 1. Stąd

$$p_0 + p_1 + p_2 + \cdots + p_k + \cdots = \sum_{k=0}^{\infty} p_k = \sum_{k=0}^{\infty} p_0 \left(\frac{\lambda}{\mu}\right)^k = 1 \tag{5}$$

Prowadzi to do postaci

$$p_0 = \left[\sum_{k=0}^{\infty}\left(\frac{\lambda}{\mu}\right)^k\right]^{-1} = 1 - \frac{\lambda}{\mu} \tag{6}$$

Z drugiej strony wykorzystanie serwera U jest określone jako

$$U = 1 - p_0 = \frac{\lambda}{\mu} \tag{7}$$

Oznacza to, że

$$p_k = (1-U)U^k \text{ dla } k=0, 1, \ldots \tag{8}$$

Zauważmy, że znając p_k, możemy łatwo obliczyć średnią liczbę żądań $\overline{N}$

$$\overline{N} = \sum_{k=0}^{\infty} k \times p_k = \sum_{k=0}^{\infty} k \times (1-U)U^k = (1-U)\sum_{k=0}^{\infty} k \times U^k \tag{9}$$

$$\text{gdzie } \sum_{k=0}^{\infty} k \times U^k = U/(1-U)^2 \text{ dla } U<1. \tag{10}$$

Tak więc otrzymujemy

$$\overline{N} = \frac{U}{1-U} \tag{11}$$

Średnia przepustowość X serwera jest określona równaniem:

$$X = U \times \mu = \frac{\lambda}{\mu} \times \mu = \lambda \tag{12}$$

W stanie równowagi średnia szybkość przybywania i odsyłania żądań z serwera jest równa. Tak więc biorąc pod uwagę przepustowość X, określoną równaniem (12) i średnią liczbą żądań $\overline{N}$ (równanie (11)), ostatecznie otrzymujemy średni czas odpowiedzi serwera na żądanie klienta:

$$t_e = t_w + t_k = \frac{\overline{N}}{X} = \frac{\frac{U}{\lambda}}{1-U} = \frac{\frac{1}{\mu}}{1-U} = \frac{S}{1-U} \tag{13}$$

2.3. Czas obsługi przez sieć – minimalny czas obsługi

Komunikat od klienta do serwera lub od serwera do innego serwera ma do przebycia wiele warstw protokółów i może być transmitowany przez jedną lub wiele sieci różnymi trasami. Komunikaty wygenerowane przez aplikacje mają do przejścia stos protokołów warstwy transportowej, np. TCP lub UDP, internetowych, np. IP lub IPX, i sieciowych (np. Ethernet, Token Ring). Protokoły istnieją na poziomie każdej warstwy komunikacyjnej.

W przypadku kiedy nie wystąpi fragmentaryzacja, wysyłający host (klient lub serwer) generuje pakiety danych, których rozmiar jest mniejszy lub równy minimum maksymalnej jednostki transmisji (MTU) wszystkich N sieci.

Liczba pakietów danych IP, N_p, związanych z przesłaniem przez jakąkolwiek sieć n spośród N sieci pomiędzy klientem a serwerem, może być określona zależnością:

$$N_p = \left\lceil \frac{R_k}{MSS} \right\rceil, \tag{14}$$

gdzie R_k jest rozmiarem komunikatu, MSS jest maksymalnym rozmiarem segmentu określonym jako:

$$MSS \leq \min_{n=1}^{N} MTU_n - K_{TCP} - K_{IP}, \tag{15}$$

MTU_n maksymalną jednostką transmisji w sieci n, K_{TCP} kosztami protokołu TCP, K_{IP} kosztami protokołu IP.

Całkowity koszt protokołu transmisji komunikatu w sieci n jest dany zależnością:

$$K_n = N_p \times (K_{TCP} + K_{IP} + K_F) \tag{16}$$

gdzie K_F jest kosztem ramki w sieci n.

Ostatecznie czas obsługi przesłania komunikatu w sieci n może być określony jako:

$$t_s = \frac{R_k + K_n}{B_n}, \tag{17}$$

gdzie B_n jest szerokością pasma sieci n.

3. Wnioski

Przedstawiona w niniejszym artykule analiza pozwala:

- spojrzeć, w odmienny niż dotychczas znany sposób, na zagadnienia jakości obsługi klienta w e-biznesie,
- zróżnicować specyfikę rozproszonych systemów obliczeniowych i klient-serwer,
- zaproponować nową metodę pomiaru jakości obsługi klienta w systemach klient-serwer,
- określić punkt wyjścia do metod analizy i syntezy systemów poprawy jakości obsługi klienta w e-biznesie.

LITERATURA

1. Buyya R.: High Performance Cluster Computing, vol. 1, Architectures and Systems, Prentice Hall, New Jersey 1999.
2. Buyya R.: High Performance Cluster Computing, vol. 2, Programming and Applications, Prentice Hall, New Jersey 1999.
3. Umar A.: Object Oriented Client/Server Internet Environments, Prentice Hall, New Jersey 1997.

4. Perruci L., Re E.D.: An Interactive Multimedia Satellite Telemedicine Service, IEEE Multimedia, ss. 76-83, Apri-June 2000.
5. Ramesh S., Perros. H.G.: A Multilayer Client-Server Queueing Network Model with Synchronous and Asynchronous Messages, IEEE Transaction on Software Engineering, vol.26, 11, ss. 1086-1100, 2000.
6. Breslau L., et al.: Web Caching abd Zipf-like Distributions: Evidence and Implications, Proc. IEEE Infocom'99, New York, March 1999.
7. Mitrani I.: Probabilistic Modelling, Cambridge University Press, 1998.
8. Bestavros A.: WWW Traffic Reduction and Load Balancing through Server-Based Caching, IEEE Concurency, ss. 56-67, January-march 1997.

Recenzent: Prof. dr hab. Ryszard Pregiel

Wpłynęło do Redakcji 27 marca 2001 r.

Abstract

The purpose of this paper is to provide the fundamental analysis to understand the behavior of e-business systems. The quality of service of e-business sites depends on many interrelated factors such architecture, network capacity, and system software structure. The complexity of an e-business site is compounded by unpredictable public behavior. All characteristics of e-business clearly indicate that quantitative techniques are needed to manage behavior of online companies and to guarantee quality of service.

Several empirical measurements indicate that Zipf's Law applies quite strongly to documents served by WWW servers, yielding the relationship [6]:

$$P = \frac{k}{r},$$

where P is the number of references to a document whose rank is r, and k is constant.

The paper deals with a new measurements method of QoS in the client-server system based on measurement of the $\log(t_o)=f[\log(r)]$ characteristic, where t_o is the system response time but r is rank of WWW server. Response time $t_o=t_w+t_k+t_s$, where t_w is the server service time, t_k the waiting time, but t_s is the network service time. In this model has been assumed that network service time is constant, and independent on r. On the second hand the queue t_k, which depends on r, determines excess time i.e., $1/r$ characteristic form. Analysis shows that

t_k dispersion is very large. Superposition of t_k leads to the linear characteristic $\log(t_o)=f[\log(r)]$ at slope equals -1. When $r \rightarrow 1$, then the utilization of the server $U \rightarrow 1$, but $t_o \rightarrow \infty$. On the other hand, when WWW server rank is low, $t_k \rightarrow 0$, but $t_o \rightarrow t_w + t_s$. Taking into account of the $\log(t_o)=f[\log(r)]$ characteristic it can be noticed that tendency to deviate down in comparison to slope equals –1 can be measure of QoS improve.