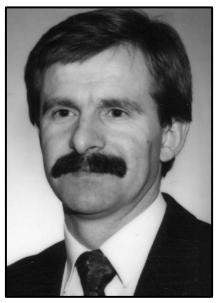
Emil MICHTA
UNIWERSYTET ZIELONOGÓRSKI, INSTYTUT METROLOGII ELEKTRYCZNEJ

Teoria szeregowania zadań w analizie dotrzymania ograniczeń czasowych w systemach pomiarowo-sterujących

Dr inż. Emil MICHTA



Adiunkt w Instytucie Metrologii Elektrycznej Uniwersytetu Zielonogórskiego. Stopień doktora nauk technicznych uzyskał w 1989 r. na Politechnice Wrocławskiej. W latach 1991-1992 przebywał na stażach w University of Minho w Bradze i Bristol Uniwersity a w roku 1993 na stażu w Advanced Research Center w Hajfie. Jego zainteresowania koncentrują się na zagadnieniach związanych z inteligentną aparaturą pomiarowo-sterującą, sieciami przemysłowymi oraz bezprzewodowymi sieciami czujników.

e-mail: e.michta@ime.uz.zgora.pl

Streszczenie

W artykule zarysowano elementy teorii szeregowania zadań, które mogą być przydatne do analizy dotrzymania ograniczeń czasowych w systemach pomiarowo-sterujących. Zaprezentowano trzy metody szeregowania zadań ze statycznym i dynamicznym przydziałem priorytetu. Przedstawiono podstawowe zależności do sprawdzenia warunku realizowalności zadań w projektowanym systemie dla szeregowania zadań metodami RM, DM i EDF.

Słowa kluczowe: systemy pomiarowo-sterujące. szeregowanie zadań.

Task Scheduling Theory in Time Deadline Analysis of Measurement-Control Systems

Abstract

In this paper essentials of task scheduling theory, which can be helpful to time deadline analysis in measurement-control systems are outlined. Three task scheduling methods with static and dynamic priority assignment are presented. Basic relations to task utilization condition testing in system being design for task scheduling based on RM, DM and EDF methods are presented.

Keywords: Measurement-Control Systems, Task Scheduling.

1. Wstęp

Współczesne systemy pomiarowo - sterujące (SPS) zbudowane są z inteligentnych węzłów, które połączone są systemem komunikacyjnym. Każdy z węzłów systemu potrafi realizować zadania pomiarowe i/lub sterujące, komunikacyjne oraz przetwarzanie danych. Wykonywanie poszczególnych zadań w węzłach jest nadzorowane przez proste systemy operacyjne takie jak np.: TinyOS, Telos, Mantis itp. Jedną z pożądanych funkcji takiego systemu jest ich zdolność do szeregowania zadań okresowych, aperiodycznych i sporadycznych realizowanych w poszczególnych węzłach [2]. Cechą charakterystyczną SPS jest to, że informacja pomiarowa lub informacja o zdarzeniu w jednym z węzłów jest przesyłana przez system komunikacyjny to innego węzła np. w celu wykonania zadania sterowania. Na ten łańcuch zdarzeń nałożone są często ograniczenia czasowe tzn. czas, jaki upływa od chwili wystąpienia zdarzenia w węźle pomiarowym do czasu wykonania reakcji na to zdarzenie nie może przekroczyć zadanego czasu wynikającego z wymogów nadzorowanego obiektu lub procesu. Z praktycznego punktu widzenia, projektantów SPS najczęściej interesuje to, czy dany zbiór zadań jest realizowalny i jaki jest najgorszy przypadek czasu reakcji na zdarzenie? Odpowiedzi na te pytania można uzyskać, wykorzystując elementy teorii szeregowania zadań [3, 4].

2. Szeregowanie zadań

Szeregowanie obejmuje alokację czasu i zasobów do zadania, w taki sposób, że wymagania czasowe lub inne wymagania wydajnościowe są spełnione. Do analizy dotrzymania ograniczeń czasowych w SPS o topologii magistralowej mogą być wykorzystane analizy prowadzone dla systemów z jednym procesorem. W systemach jednoprocesorowych, zbiór zadań współdzieli wspólne zasoby, takie jak procesor, pamięci i urządzenia we/wy. W SPS zbiór zadań w postaci komunikatów do przesłania przez n węzłów współdzieli jedną magistralę. Podstawowym celem analizy szeregowania zadań jest formalne wykazanie, że realizowane zadania o znanych parametrach zostaną wykonane w każdych warunkach i w zadanym czasie. Procedurę szeregowania zadań można wykonać *online* lub *offline*. *Online* oznacza prowadzenie analizy szeregowania zadań w trakcie pracy systemu natomiast szeregowanie zadań typu *offline* oznacza, że analiza szeregowania zadań została przeprowadzona przed uruchomieniem systemu.

Innym, bardziej elastycznym podejściem do statycznego szeregowania zadań, możliwym do wykorzystania w SPS jest wykorzystanie mechanizmu priorytetu, w którym nie występuje sprecyzowane uszeregowanie kolejności wykonywanych zadań. Podczas pracy systemu zadania są wykonywane w kolejności zależnej od przypisanego im wcześniej priorytetu. Rozwiązania bazujące na wykorzystaniu mechanizmu priorytetu są bardziej elastyczne i lepiej przystosowują się do potrzeb [1, 3, 4].

3. Zasady przydziału priorytetu

Najczęściej wykorzystywaną zasadą przydziału priorytetu danemu zadaniu jest uwzględnianie okresu jego występowania według zasady: *krótszy okres występowania wyższy priorytet*. Stosowanie tej zasady wynika z tego, że zadania występujące częściej zazwyczaj są ważniejsze od zadań występujących rzadziej. Ponadto, ograniczenia czasowe i/lub najgorszy przypadek czasu odpowiedzi dla zadań występujących częściej są krótsze, co dodatkowo uzasadnia stosowanie tej zasady. Taki sposób przydziału priorytetu określany jest jako *Rate Monotonic* (*RM*).

Zastosowanie podobnej zasady w odniesieniu do zadań sporadycznych nie wydaje się sensowne. Zadania występujące sporadycznie mogą być również istotne dla systemu, zatem ich obsługa powinna być realizowana według innej zasady, umożliwiającej dotrzymanie ograniczeń czasowych stawianych poszczególnym zadaniom występującym sporadycznie. W takiej sytuacji priorytet może być przydzielany na podstawie względnego ograniczenia czasowego według zasady: *mniejsza wartość ograniczenia czasowego, większy priorytet*. Ten sposób przydziału priorytetu określany jest jako *Deadline Monotonic* (*DM*).

Jeżeli priorytety przypisywane poszczególnym zadaniom zgodnie z zasadą *RM* lub *DM* nie są zmieniane podczas pracy systemu, to uważa się je za systemy ze statycznym przydziałem priorytetu. Jeżeli w trakcie pracy systemu priorytet przypisywany zadaniu może zostać zmieniony to systemy te nazywamy systemami z dynamicznym przydziałem priorytetu. Przykładem systemu z dynamicznym przydziałem priorytetu jest system wykorzystujący zasadę *EDF* (ang. *earliest deadline first*), zgodnie z którą najwyższy priorytet przydzielany jest zadaniu, któremu najwcześniej kończy się ograniczenie czasowe.

W systemach z dynamicznym przydziałem, przydzielenie priorytetu i wysłanie zadania do wykonania jest realizowane wówczas kiedy pojawia się nowe zadanie do wykonania lub jeżeli kończy się wykonanie aktualnie wykonywanego zadania. Zaletą metody dynamicznego przydziału priorytetu *EDF*, w porównaniu do metod statycznego przydziału priorytetu (*RM, DM*) jest lepsze wykorzystanie procesora. Jej wadą jest większe obciążenie procesora zadaniami szeregowania podczas pracy. W systemach czasu rzeczywistego o dużych wymaganiach czasowych występują zadania mniej istotne, które w systemach wykorzystujących metodę *EDF* mogą blokować wykonanie ważnych zadań. Ponadto metoda *EDF* jest wrażliwa na chwilowe przeciążenia systemu wynikające z wystąpienia sytuacji wyjątkowych lub podczas poprawiania błędów, które mogą doprowadzić do przekroczenia

ograniczeń czasowych. W systemach ze statycznym przydziałem priorytetu niezależnie od sytuacji, zawsze w pierwszej kolejności będą wykonywane zadania o wyższym priorytecie.

4. Szeregowanie zadań ze statycznym przydziałem priorytetu metodą *RM* i *DM*

Dla przypisania priorytetu zgodnie z metodą *RM* parametry szeregowanych zadań powinny spełnić warunek gwarantujący, że *N* zadań zostanie wykonanych przed upływem ich ograniczenia czasowego [1]:

$$\sum_{i=1}^{N} \frac{C_i}{T_i} \le N \times \left(2^{1/N} - 1\right), \tag{1}$$

gdzie: C_i jest maksymalnym czasem wykonania *i-tego* zadania a T_i jest okresem jego występowania (w odniesieniu do zadań sporadycznych T_i jest minimalnym czasem pomiędzy kolejnymi wystąpieniami zadania).

Przyjęto założenie, że ograniczenie czasowe jest równe okresowi występowania zadania. Takie założenie jest naturalnym i najczęściej przyjmowanym założeniem podczas analiz dotrzymania ograniczeń czasowych w projektowanych SPS. Analiza dotrzymania ograniczeń czasowych w systemach bez wywłaszczania zadań ma w przypadku SPS duże znaczenie praktyczne, ponieważ taka sytuacja występuje na poziomie systemu komunikacyjnego i obsługi stosu protokołowego w zdecydowanej większości sieci przemysłowych stosowanych do przesyłania informacji w SPS.

W systemach bez wywłaszczania zadań mogą wystąpić sytuacje, w których zadania o niższym priorytecie blokują te zadania o wyższym priorytecie, które pojawiły się po rozpoczęciu wykonywania zadania o niższym priorytecie. Oznaczając przez B_i maksymalny czas blokowania zadania *i* nierówność (1) można zmodyfikować do postaci [3]:

$$\sum_{i=1}^{N} \left(\frac{C_i}{T_i}\right) + \frac{B_i}{T_i} \le i \times \left(2^{1/i} - 1\right), \forall_{i,1 \le i \le N} \tag{2}$$

5. Czas odpowiedzi na zdarzenia

Czas odpowiedzi na zdarzenia jest jednym z podstawowych parametrów SPS. Jego znaczenie uwidacznia się zwłaszcza w tych systemach, w których istotne jest dotrzymanie warunków czasu rzeczywistego. Potrzeba projektowania przewidywalnych SPS, tzn. takich, w odniesieniu do których można na etapie projektowania określić jego parametry i przewidzieć jego zachowanie się w sytuacjach krytycznych (najgorszy przypadek) wymaga opracowania takiej metodyki postępowania, która zapewni osiągnięcie żądanych parametrów na drodze formalnej. W tym celu opracowano zależności wspomagające proces projektowania. W SPS analiza jego parametrów może dotyczyć poszczególnych węzłów, systemu komunikacyjnego oraz całego systemu obejmującego zarówno węzły jak i system komunikacyjny.

Dowiedziono [3], że najbardziej niekorzystną sytuacją dla oszacowania czasu odpowiedzi R_i zadania *i* jest synchroniczne, jednoczesne uaktywnienie wszystkich zadań z ich maksymalną częstotliwością występowania. Czas ten wyznacza się z następującego równania rekurencyjnego:

$$R_i^{n+1} = \sum_{j \in hp(i)} \left(\left\lceil \frac{R_i^n}{T_j} \right\rceil \times C_j \right) + C_i \tag{3}$$

gdzie: n jest kolejnym krokiem iteracji.

Rozwiązanie równania rekurencyjnego uzyskujemy jeżeli $R_i^{n+1} = R_i^n$. Jeżeli zadania *1...N* uszeregowane są według rosnącego priorytetu i są niezależne, to liczba kroków iteracji, po której uzyskujemy rozwiązanie jest równa *(N-i)+1*. Iterację rozpoczynamy przyjmując $R_i^0 = 0$. W pierwszym kroku iteracji uzyskujemy minimalną wartość czasu odpowiedzi R_{imin}, która jest równa czasowi C_i wykonania zadania *i*. Sytuacja taka występuje, jeżeli w czasie wykonywania zadania *i* nie pojawi się żadne zadanie o priorytecie wyższym od zadania *i*. Jeżeli wynik rekurencji jest zbieżny do okresu T_i występowania zadania *i* lub przekroczy jego wartość to zadanie nie jest szeregowalne. Oznacza to, że podczas funkcjonowania systemu może wystąpić sytuacja, podczas której zadanie nie zostanie wykonane przed upływem okresu jego występowania.

W systemach pracujących bez wywłaszczania zadań, podstawowym zagadnieniem jest uwzględnienie w prowadzonych analizach skutków inwersji priorytetu, polegające na blokowaniu wykonania zadań z wyższym priorytetem przez zadania o niższym priorytecie. Wyniki tych analiz mają bezpośrednie zastosowanie do analizy systemu komunikacyjnego i stosu protokołowego SPS, w którym zastosowano protokoły komunikacyjne ze zdecentralizowanym dostępem do magistrali, czyli protokoły klasy „*peer-to-peer*". Popularnymi przedstawicielami protokołów tej klasy są: CAN, LonWorks, Modbus Plus, Profibus (tylko węzły aktywne).

Uwzględniając czynnik blokowania, wyrażenie (3) na najgorszy przypadek czasu odpowiedzi przyjmie następującą postać:

$$R_i^{n+1} = B_i + \sum_{j \in hp(i)} \left(\left\lceil \frac{R_i^n}{T_j} \right\rceil * C_j \right) + C_i \tag{4}$$

6. Szeregowanie zadań z dynamicznym przydziałem priorytetu metodą EDF

W metodzie *EDF* dynamicznego przydzielania priorytetu wykonywanym zadaniom, konieczne jest sprawdzenie zachowania się systemu w fazie projektowej. Podobnie, jak dla poprzednio przeprowadzonych analiz dla metody *RM* szeregowania zadań, warunkiem podstawowym dla metody *EDF* w systemach bez wywłaszczania dla zadań okresowych z ograniczeniem czasowym równym okresowi występowania zadania jest:

$$\sum_{i=1}^{N} \frac{C_i}{T_i} \le 1 \tag{5}$$

Dla niezależnych i niewywłaszczanych zadań w nierówności (5) należy uwzględnić czynnik blokowania B_i i wówczas przyjmie ona postać:

$$\sum_{i=1}^{N} \left(\frac{C_i}{T_i}\right) + \frac{B_i}{T_i} \le 1 \tag{6}$$

gdzie: B_i jest maksymalnym czasem blokowania zadania *i* przez zadania o niższym priorytecie.

Jeżeli ograniczenia czasowe stawiane poszczególnym zdaniom są mniejsze niż okres ich występowania, to wówczas nierówność (6) przyjmie postać:

$$\sum_{i=1}^{N} \left(\frac{C_i}{D_i}\right) + \frac{B_i}{D_i} \le 1 \tag{7}$$

Zależności (5) i (6) są proste do przeprowadzenia analiz ale jednocześnie są one dość pesymistyczne, gdyż uwzględniają najbardziej niekorzystne przypadki. W kolejnym punkcie wprowadzone zostaną pewne modyfikacje pozwalające na uzyskanie zależności dających bardziej realistyczne wyniki.

7. Szeregowanie zadań metodą EDF z wywłaszczaniem

W wielu rzeczywistych rozwiązaniach występują zadania z ograniczeniem czasowym D_i krótszym od okresu T_i występowania zadania. Ich uwzględnienie prowadzi do następującej zależności [3]:

$$\sum_{i=1}^{N} \left\lceil \frac{t - D_i}{T_i} \right\rceil^{+} * C_i \le t \tag{8}$$

Przyjmijmy, że w czasie $t = 0$, nie ma zadań oczekujących na wykonanie. Zatem, warunkiem koniecznym gwarantującym dotrzymanie ograniczeń czasowych jest taka ilość czasu, która jest potrzebna

do wykonania wszystkich zadań wygenerowanych w przedziale [*0, t*] z bezwzględnym ograniczeniem czasowym nie większym niż t. Ponieważ minimalny czas pomiędzy kolejnymi wystąpieniami zadania *i* wynosi T_i , więc w przedziale czasu [*0, t*] zadanie to wystąpi co najwyżej $[(t - D_i)/T_i]$ razy. Zatem czas potrzebny na wykonanie wszystkich wystąpień zadania *i* w rozpatrywanym przedziale wyniesie:

$$\left\lceil \frac{t - D_i}{T_i} \right\rceil^{+} * C_i \tag{9}$$

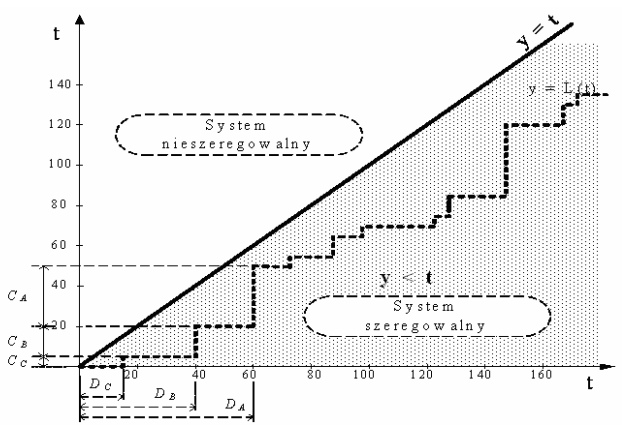
Rozpatrując wszystkie *N* zadań w przedziale [*0, t*], czas potrzebny do ich wykonania jest sumą czasów zajmowanych przez kolejne zadania, co prowadzi do zależności (8). Łatwo można wykazać, że jeżeli ograniczenia czasowe dla poszczególnych zadań są równe okresowi ich występowania, tzn. $D_i = T_i$, to nierówność (8) jest spełniona, jeżeli spełniona jest nierówność (5).

W celu ilustracji graficznej omawianego zagadnienia rozważmy następujący przypadek, w którym przeprowadzono analizę dotrzymania warunków czasowych dla trzech zadań pracujących w systemie z wywłaszczaniem zadań (tab. 1).

Tab. 1. Parametry szeregowanych zadań dla metody *EDF*
Tab. 1. A scheduled tasks parameters for *EDF* method

Zadanie	C_i	T_i	D_i	C_i/D_i	C_i/T_i
A	30	80	60	0.5	0.375
B	10	40	40	0.25	0.25
C	5	25	15	0.3	0.2
	$\sum_{i=1}^{3} C_i = 45$			$\sum_{i=1}^{3} \frac{C_i}{D_i} = 1.05$	$\sum_{i=1}^{3} \frac{C_i}{T_i} = 0.825$

Stosując do oceny realizowalności zadań, w systemie szeregującym metodą *EDF*, wyrażenie pesymistyczne $(C_i/D_i) < 1$, uzyskujemy wynik negatywny. Natomiast stosując do analizy realizowalności zadań, wyrażenia uwzględniające zarówno ograniczenia czasowe jak i okres występowania uzyskujemy wynik pozytywny, co zostało przedstawione w postaci graficznej na rys. 1.



Rys. 1. Ilustracja graficzna pracy systemu z metodą *EDF*
Fig. 1. A graphical presentation of a system with *EDF* method

Otwartą i ważną kwestią jest określenie długości przedziału czasu [*0, t*], w którym powinny być prowadzone analizy. Wobec tego, że wyrażenie (8) zmienia się jedynie w chwilach czasowych $k*T_i + D_i$, zatem weryfikacja poprawności tej nierówności została przeprowadzona dla tych chwil czasowych, co zostało wykorzystane podczas konstrukcji wykresu przedstawionego na rys. 1. W którym momencie należy zakończyć konstrukcję wykresu, ażeby mieć pewność, że zadania są szeregowalne? Można wykazać, że jeżeli spełniona jest nierówność (5), to istnieje taki przedział czasu $[0, t_{max}]$, że nierówność (8) spełniona jest dla każdego $t > t_{max}$.

Zagadnienie określenia czasu t_{max} było przedmiotem zainteresowania wielu autorów prowadzących prace badawcze w zakresie oceny wydajności systemów operacyjnych pracujących w oparciu o metodę dynamicznego przydziału priorytetu z wywłaszczaniem i bez wywłaszczania zadań. W pracy [3] wykazano, że wartość czasu t_{max} , dla systemu z dynamicznym przydziałem priorytetów zadaniom okresowym i stosującym metodę wywłaszczania zadań można wyznaczyć z następującej zależności:

$$t_{\max} = \left(\frac{U}{1-U} \right) \times \max_{i=1,\ldots,N} \left\{ T_i - D_i \right\} \tag{10}$$

gdzie:

$$U = \sum_{i=1}^{N} \frac{C_i}{T_i}$$

jest współczynnikiem wykorzystania procesora przez *N* szeregowanych zadań.

Dla analizowanego przykładu (tab. 1) wartość czasu t_{max} określająca górną granicę wykonywania analiz wynosi t_{max} = 94,3. Zależność (10) można zmodyfikować do następującej postaci:

$$t_{\max} = \sum_{i=1}^{N} \left\lceil \frac{1 - D_i}{T_i} \right\rceil^{+} * \; C_i \le \left(1 - U\right) \tag{11}$$

Dla danych z analizowanego przykładu (tab. 1) wartość czasu t_{max} określającą górną granicę wykonywania analiz wynosi t_{max} = 59,2. Z wykresu przedstawionego na rys. 7.4 widać, że dla tej wartości czasu występuje minimalna różnica pomiędzy prostą *y=t* i lewą stroną nierówności (8).

Łatwo można wykazać, że jeżeli współczynnik wykorzystania procesora będzie zbliżał się do 1, to wartość górnego ograniczenia czasu prowadzenia analiz będzie znacznie rosła. Z tego powodu zaproponowano inne rozwiązanie polegające na wyznaczeniu wartości górnego ograniczenia przy założeniu, że punktem wyjścia do analizy jest najbardziej niekorzystny przypadek tzn., że w chwili *t = 0*, wszystkie zadania są gotowe do wykonania. Takie podejście jest stosowane również podczas wyznaczania czasu odpowiedzi dla szeregowania zadań metodą *RM*. Wykorzystując poprzednio opracowane zależności otrzymujemy zależność rekurencyjną na maksymalny przedział czasu t, w którym należy przeprowadzić analizę gwarantującą dotrzymanie warunków szeregowalności można wyznaczyć z zależności rekurencyjnej:

$$L^{n+1} = \sum_{i=1}^{N} \left\lceil \frac{L^n}{T_i} \right\rceil * C_i \tag{12}$$

gdzie: *L* jest przedziałem czasu od chwili wystąpienia krytycznej sytuacji (wszystkie zadania aktywne) do chwili kiedy procesor (magistrala) nie będzie miał zadań oczekujących na wykonanie.

Maksymalna liczba iteracji n_{max} jest nie większa od liczby zadań *N*. Warunkiem uzyskania zbieżności rozwiązania równania (12) jest spełnienie warunku (5). Uzyskana wartość czasu *L* może być jednym ze wskaźników oceny zaprojektowanego systemu.

8. Podsumowanie

Przedstawione w artykule podstawowe zależności wynikające z teorii szeregowania zadań ze statycznym lub dynamicznym przypisaniem priorytetu pozwalają na ocenę dotrzymania ograniczeń czasowych przez zadania realizowane w projektowanym SPS.

9. Literatura

[1] Audsley N., Burns A., Richardson M., Tindell K. and Wellings A.: Applying new scheduling theory to static priority pre-emptive scheduling. Software Engineering Journal, Vol. 8, No. 5,1997, pp. 285-292.
[2] Kim D. and Lee Y.: Periodic and aperiodic task scheduling in strongly partitioned integrated real-time systems. The Computer Journal, Vol. 45, No. 4, 2002, pp. 395 – 409.
[3] Michta E.: Modele komunikacyjne sieciowych systemów pomiarowo-sterujących. Wydawnictwo PZ. Monografia nr 99, 2000.
[4] Sydenham P. and Thorn R.: Handbook of Measuring System Design. John Wiley& Sons, 2005 .

Artykuł recenzowany