Polskie Towarzystwo Badań Operacyjnych i Systemowych
Instytut Badań Systemowych Polskiej Akademii Nauk
Wojskowa Akademia Techniczna

Redaktorzy:
Zbigniew Nahorski
Marian Chudy
Andrzej Straszak

OPTYMALIZACJA

Warszawa 1991

POLSKIE TOWARZYSTWO
BADAŃ OPERACYJNYCH I SYSTEMOWYCH

INSTYTUT BADAŃ SYSTEMOWYCH
POLSKIEJ AKADEMII NAUK

WOJSKOWA AKADEMIA TECHNICZNA

OPTYMALIZACJA

ZADANIA, METODY, ALGORYTMY

Redaktorzy

Zbigniew Nahorski, Marian Chudy, Andrzej Straszak

WARSZAWA 1991

OPTYMALIZACJA:
Zadania, metody, algorytmy
PTBOiS, Warszawa 1991

KOSZTY CZASOWE KOMUNIKACJI W MASZYNACH RÓWNOLEGŁYCH

Leon Słomiński
Instytut Badań Systemowych PAN
ul.Newelska 6, 01-447 Warszawa

Przedstawiono sposoby modelowania,analizy i szacowania nakładów czasu na komunikację wewnętrzną i zewnętrzną w sieciach złożonych z wielu procesorów.Dla modelu maszyny równoległej **MP-RAM** i modelu **SIMD-Network-VLSI** przedyskutowano podstawowe funkcje, które musi spełniać sieć w związku z wykonywaniem obliczeń. Na kilku przykładach sieci pokazano jak można szacować czas niezbędny na komunikację wewnętrzną i zewnętrzną poprzez parametry sieci.

1.MODELE OBLICZEŃ RÓWNOLEGŁYCH

Czas zużyty na komunikację między procesorami sieci i na komunikację sieci z otoczeniem odgrywa istotną rolę w obliczeniach równoległych. Znane są przykłady pokazujące, że czas ten może decydować o łącznej złożoności czasowej algorytmu [5]. Z punktu widzenia metod analizy i technik budowy efektywnych algorytmów równoległej optymalizacji dyskretnej ważną rolę odgrywa wybór modelu maszyny równoległej. W pracy tej uwagę skupiamy na modelach sieciowych z nieograniczonym zasobem pamięci i z nieograniczoną liczbą procesorów. Ważnym wyróżnikiem modeli sieciowych jest zołożenie o lokalności pamięci (każdy procesor ma własny moduł pamięci) i zasada komunikowania się (synchronicznego lub niesynchronicznego) procesorów przez łącza tworzące sieć (komutatorową lub stałą).Modele sieciowe obejmują szeroki wachlarz maszyn, spośród których wyróżnimy model **MP-RAM** (Message Passing - Random Access Machine) [3], oraz model **SIMD-Network** (Single Instruction stream Multiple Data stream Network type machine) [2], w tym model **SIMD-Network-VLSI** [7].

Przyjmiemy konwencję przedstawiania sieci za pomocą grafu $G(V,E)$, którego wierzchołki (zbiór V) reprezentują

procesory (wraz z modułami pamięci), zaś krawędzie (elementy zbioru E) przedstawiają bezpośrednie połączenia między procesorami. Podobną konwencję stosujemy do przedstawienia algorytmu, z tym że reprezentuje go graf skierowany, którego wierzchołki są procesami połączonymi łukami ukazującymi następstwa zachodzące między procesami.

Model **MP-RAM** składa się z następujących elementów:

1.p,($p \geq 2$),procesorów. Każdemu procesorowi jest przyporządkowany proces.Procesy są to odrębne samo-modyfikowalne programy,utworzone z ciągów dowolnie ponumerowanych (zaetykietowanych) instrukcji.

2.p modułów pamięci, po jednym module dla każdego procesora.

3.p akumulatorów po jednym dla każdego programu.

4.Jednej dowolnie długiej taśmy-wyłącznie do czytania.

5.Jednej,dowolnie długiej,taśmy- wyłącznie do pisania.

6.Sieci (dynamicznej lub statycznej),którą tworzą łącza służące do komunikacji między procesorami.

Model **SIMD-Network** różni się od modelu **MP-RAM** postacią założenia 1 i dodatkowym założeniem 7:

1'.p, ($p \geq 2$) identycznych programów wykonywanych synchronicznie na strumieniu niekoniecznie identycznych danych.

7.Instrukcja wektorowa może być użyta z maską, która powoduje zawieszenie pracy niektórych procesorów dla tej instrukcji.

Model **SIMD-Network-VLSI** jest stosowany do analizy nakładów w sieciach wielomikroprocesorowych (liczba mikroprocesorów - może mieć wartość milion i więcej) realizowanych w jednym mikroukładzie, w technologii Bardzo Dużej Skali Scalenia (BDSS, ang.VLSI). W modelu tym złożoność obliczeń jest wyrażana iloczynem czasu operacji obliczeniowych i komunikacyjnych przez powierzchnię mikroukładu.

W naszych rozważaniach korzystamy z modyfikacji modelu obliczeń zaproponowanego przez Thompsona [7]:

- Mikroukład jest dwuwarstwowy;

- Każdy mikroprocesor zajmuje stałą powierzchnię (np. jeden bit logiki lub pamięci zajmuje O(1) jednostek powierzchni);
- Przewody łączące mikroprocesory mają stałą szerokość i krzyżują się pod kątem prostym;
- Przewód o długości l wnosi opóźnienie czasowe proporcjonalne do $log_2 l$, (w oryginalnym modelu Thompsona zakłada się niezależność opóźnienia od długości przewodu).Na opóźnienie to mają wpływ: czas transmisji ($O(log_2 l)$ dla jednostki transmitowanych danych), oraz czas opóźnienia wnoszony przez układ wzmocnienia sygnału.Przujmuje się, że przewód o długości l jednostek ma l-stopniowy wzmacniacz kaskadowy, który zajmuje powierzchnię $O(l)$ i wnosi opóźnienie $O(log_2 l)$ jednostek czasu.

Przedstawiony model,i inne tego typu modele, dają oceny ogólnego nakładu obliczeń typu: $AT^2 = \Omega(N^2)$,gdzie A jest powierzchnią mikroukładu, T - czasem wykonywania wszystkich działań, N - rozmiarem rozwiązywanego zadania, wyrażonym np. liczbą bitów potrzebną do przedstawienia danych.

2.PODSTAWOWE FUNKCJE KOMUNIKACYJNE W SIECI PROCESORÓW

Punkt ten opieramy w dużym stopniu na pracy [1],ograniczymy się przy tym do modelu **MP-RAM** opisanego grafem $G(V,E)$ reprezentującym sieć. Dane początkowe, wyniki pośrednie i wyniki końcowe są gromadzone w pamięciach lokalnych węzłów sieci. Wymiana informacji między procesorami odbywa się w trybie wymiany komunikatów, które są pakietami utworzonymi z ciągów bitów. Łączny czas zużyty na przekazanie pakietu od nadawcy do odbiorcy składa się z czasów:

- przygotowania transmisji (tworzenie pakietów, przypisanie adresów i sygnałów sterujących, wybór drogi wprowadzenia pakietu do bufora początkowego, drogi transmisji itp.);
- oczekiwania w kolejce na wolne łącze i korekty błędów transmisji;

- samej transmisji;
- propagacji, tzn. czasu upływającego od chwili wysłania ostatniego bitu przez źródło, do chwili odebrania tego bitu przez procesor-odbiornik.

Komunikację w sieciach procesorów można opisać za pomocą dobrze zdefiniowanych powtarzalnych, działań, nazywanych *funkcjami komunikacyjnymi*. Wyróżnia się następujące funkcje komunikacyjne:

- rozsyłanie jednorodne pakietów (identyczny pakiet jest wysyłany z jednego, wyróżnionego, procesora do wielu, lub wszystkich pozostałych procesorów sieci);
- odbiór jednorodny pakietów (identyczny pakiet jest wysyłany jednocześnie przez wiele, lub wszystkie procesory sieci, do jednego, wyróżnionego procesora-odbiornika);
- rozsyłanie niejednorodne pakietów (jeden procesor wysyła odrębny pakiet do wielu, lub do wszystkich pozostałych procesorów);
- odbiór niejednorodny pakietów (wiele procesorów, być może wszystkie, z wyjątkiem jednego, chce przekazać jednocześnie odrębne pakiety do jednego procesora).

Wyliczone funkcje mają wariant ogólniejszy:

- rozsyłanie/odbiór jednorodny może się dokonywać w układzie: każdy procesor do/od wszystkich pozostałych procesorów;
- rozsyłanie/odbiór różnych pakietów może polegać na wymianie informacji między wszystkimi procesorami.

Wprowadzimy następujące oznaczenia dla funkcji komunikacyjnych:

- rozsyłanie jednorodne z jednego źródła $1 \longrightarrow p$,
- rozsyłanie jednorodne z wielu źródeł $p \longrightarrow p$,
- odbiór jednorodny z wielu źródeł $1 \longleftarrow p$,
- odbiór jednorodny przez wiele odbiorników $p \longleftarrow p$,
- rozsyłanie niejednorodne z jednego źródła $1 \xrightarrow{n} p$,
- odbiór niejednorodny z wielu źródeł $1 \xleftarrow{n} p$,
- pełna wymiana niejednorodna $1 \xleftrightarrow{n} p$.

Teraz omówimy ogólne związki zachodzące między wyliczonymi funkcjami. W tym celu uściślimy założenia dotyczące sieci. Każda krawędź jest traktowana jako dwukierunkowe,asynchroniczne,wolne od błędów transmisji łącze, przekazujące bity informacji. Komunikacja może być zainicjowana równocześnie, wzdłuż wszystkich łącz (w obydwu kierunkach) incydentnych z danym procesorem. Co więcej, komunikację tę można zainicjować jednocześnie we wszystkich procesorach. Pomijając czasy oczekiwań można zakładać, że pakiety jednakowej długości potrzebują jednakowego czasu komunikacji, niezależnie od łącza, przy czym przekazanie pakietu wzdłuż dowolnego łącza zajmuje jedną jednostkę czasu.

Rozsyłanie jednorodne,w układzie $1 \longrightarrow p$ i odbiór jednorodny w układzie $1 \longleftarrow p$, można traktować jako parę funkcji dualnych. Oznacza to, że algorytm który rozwiązuje jedno zadanie,po prostych modyfikacjach, rozwiązuje drugie z nich,przy czym nakłady czasu są identyczne. Algorytm optymalny dla funkcji $1 \longrightarrow p$ polega na ustanowieniu drzewa skierowanego dróg najkrótszych (dendrytu skierowanego) z wybranego wierzchołka do pozostałych wierzchołków. Wykonanie tej funkcji zajmuje $O(r)$ jednostek czasu, gdzie r jest najdłuższą drogą we wspomnianym drzewie. Odwrócenie orientacji łuków w drzewie, pozwala użyć to samo drzewo do wykonania funkcji $1 \longleftarrow p$. Założenie o identycznym czasie wykonania każdej funkcji pozostaje prawdziwe jeżeli przyjmiemy, że w drzewie odwróconym pakiety są łączone w jeden pakiet na wejściu procesora, i tak powstały pakiet nadal potrzebuje jednej jednostki czasu na komunikację.

Rozsyłanie/odbiór jednorodny w układzie $p \longrightarrow p \; / p \longleftarrow p$ polega na wielokrotnym wykonaniu funkcji $1 \longrightarrow p \; / \; 1 \longleftarrow p$. Pojawia się jednak istotne utrudnienie, polegające na możliwości wystąpienia równoczesnego zapotrzebowania na dane łącze przez wiele procesorów. Konieczny jest algorytm rozwiązywania konfliktów.

Komunikacja niejednorodna,w każdym z układów : $1 \xrightarrow{n} p$,

$1 \xleftarrow{n} p$, $p \xleftrightarrow{n} p$, jest funkcją bardziej złożoną od funkcji jednorodnych. Problem łącza-wąskiego gardła i konieczność rozwiązania zadania obsługi kolejki może się pojawić w każdym z tych przypadków.

3.PARAMETRY SIECI ISTOTNE DLA KOMUNIKACJI

Sieć opisujemy nadal grafem nieskierowanym $G(V,E)$. Podamy definicje następujących parametrów: średnica, stopień spójności (indeks spójności), elastyczność.

Średnica sieci - jest to maksymalna, ze względu na pary wierzchołków, odległość mierzona najmniejszą liczbą krawędzi, niezbędnych do połączenia pary. Dla sieci o średnicy r czas opóźnienia komunikacji jest w najgorszym przupadku $O(r)$.

Stopień sieci - jest to najmniejsza liczba krawędzi incydentnych z wierzchołkami. Stopień sieci będzie oznaczany następująco: $\deg_G = \min_{i \in V} \deg(i)$, gdzie $\deg(i)$ jet *stopniem wierzchołka* - liczbą krawędzi incydentnych z wierzchołkiem i. ! Wyższy stopień sieci zapewnia lepsze parametry czasowe komunikacji.

Spójność sieci - sieć jest spójna jeżeli między dowolną parą jej wierzchołków jest połączenie. Najmniejsza liczba wierzchołków/ krawędzi sieci, usunięcie których zrywa jej spójność nosi nazwę *wierzchołkowego* (*krawędziowego*) *indeksu spójności sieci*, s. Wysoka wartość indeksu spójności krawędziowej sieqi jest pożądana ze względu na niezawodność i możliwość skrócenia czasu transmisji. Dodatkowe połączenia można wykorzystać w przypadku uszkodzenia jednego z łącz. Wiele krawędziowo rozłacznych dróg między parą wierzchołków pozwala zwiększyć przepustowość. Niech k będzie liczbą krawędziowo rozłącznych dróg miezy daną parą wierzchołków.Jeżeli pominiemy czas oczekiwania na transmisję i straty czasu wynikające z dzielenia pakietów między rozłączne drogi i odtwarzania informacji u odbiorcy, to możemy oczekiwać k-krotnego zwiększenia przepustowości, w stosunku do pojedyńczej drogi.

Zauważmy, że $k \leq \deg(i)$, a $s \leq \deg_G(i)$.

Elastyczność sieci - definiuje się jako jej zdolność do efektywnej realizacji szerokiej gamy algorytmów. Niech będzie dany algorytm realizowany efektywnie przez sieć daną grafem $G'(V',E')$. Chcemy wykorzystać ten sam algorytm, z niezmienioną efektywnością, w sieci danej innym grafem - $G(V,E)$. Możemy to wykonać, jeżeli potrafimy odwzorować graf G' w graf G, w ten sposób, że dla każdej pary wierzchołków $\sigma(i)-\sigma(j), i \neq j, \sigma(i) \neq \sigma(j)$,g dzie σ jest funkcją odwzorowującą V' w V,zachodzi relacja: $(\sigma(i),\sigma(j)) \in A$ jeżeli $(i,j) \in A'$. Opisane odwzorowanie jest ważne m.in. wówczas, gdy chcemy odwzorować, w sieć, graf powiązań między podzadaniami zdekomponowanego zadania, w ten sposób aby podzadania połączone bezpośrednio krawędzią znalazły się w procesorach połączonych łączem. Takie odwzorowanie minimalizuje czas komunikacji

Ważną cechą sieci jest liczba portów wejścia/wyjścia do komunikacji z otoczeniem zewnętrznym. Generalnie pożądane są sieci z liczbą portów większą od dwóch.

4.PRZYKŁADY SIECI I NAKŁADY CZASU NA WYKONANIE FUNKCJI KOMUNIKACYJNYCH

Zatrzymamy się na trzech sieciach : *tablica jednowymiarowa, hipersześcian* i *las drzew ortogonalnych*. Dwie pierwsze sieci są przedstawione w modelu **MP-RAM**, ostatnia sieć - w modelu **SIMD-Network-VLSI**. Przykłady analizy złożoności obliczeniowej algorytmów optymalizacji dyskretnej, w której są uwzględnione idee zaprezentowane w punkcie 3 i w punkcie bieżącym, można znaleźć w pracy [6].

Tablica jednowymiarowa. Sieć tego typu składa się z p procesorów ponumerowanych od 1 do p, połączonych ze sobą w ten sposób,że $(i,i+1) \in E$, dla $i = 1,\ldots,p-1$. Średnica sieci $r = p-1$,oraz indeks spójności krawędziowej $s=1$, nie są zachęcające. Niemniej, tablicę jednowymiarową można odwzorować w wiele innych sieci (np. pierścień, siatka, hipersześcian). Oznacza to,że parametry komunikacyjne

tablicy tej nie mogą być lepsze od paramtrów sieci,w którą daje się ona odwzorować. Rozsyłanie/odbiór jednorodny w relacjach: $1 \longrightarrow p$, $1 \longleftarrow p$, $p \longleftrightarrow p$, można realizować w czasie optymalnym, proporcjonalnym do $(p-1)$. Identyczna ocena obowiązuje dla algorytmów optymalnych rozsyłania/odbioru niejednorodnego w relacjach $1 \xrightarrow{n} p$, $1 \xleftarrow{n} p$. Algorytm optymalny pełnej wymiany (niejednorodnej) ma nakład $O(p^2)$ jednostek czasu.

Hipersześcian (k-wymiarowy sześcian) - jest siecią, którąformalnie można traktować jako siatkę k-wymiarową o $n_i = 2$ dla każdego i. Średnica hipersześcianu wynosi $r = k = \log_2 p = \log_2 2^k$, indeks spójności wierzchołkowej $s = k$. Dwa procesory łączy krawędź jeżeli ich numery identyfikacyjne $(l=0,1,\ldots,p-1)$, przedstawione k-bitowymi liczbami binarnymi, różnią się dokładnie na jednej pozycji. Liczba łącz na drodze między dowolną parą wierzchołków jest niemniejsza niż liczba pozycji, na których różnią się ich reprezentacje binarne. Hipersześcian jest siecią elastyczną, można pokazać, [1], że w sieć tę odwzorowuje się m.in.pierścień i siatka. Nie można w k-sześcian odwzorować pełnego drzewa binarnego o 2^k-1 wierzchołkach, dla $k \geq 3$. Niemniej w hipersześcian daje się odwzorować drzewo nazywane *dwukorzeniowym pełnym drzewem binarnym*, gdy liczba jego wierzchołków jest 2^k. Jeżeli numery binarne pary wierzchołków różnią się na δ miejscach, to δ wierzchołkowo rozłącznych dróg, łączących tę parę, ma długość δ (łącz) każda, a pozostałe $k-\delta$ dróg ma długość $\delta+2$.

Las binarnych drzew ortogonalnych (Binarne drzewa ortogonalne - BDO),to sieć zaproponowana w [5]. Jej główną zaletą jest wieloportowość. Binarne drzewa ortogonalne można traktować jak tablicę o wymiarze $n*n$, $n \geq 2$, której węzły są liśćmi drzew binarnych, rozpiętych na kolumnach i wierszach (łącznie $2n$ drzew binarnych). Korzeń każdego drzewa i wszystkie jego wierzchołki wewnętrzne są także procesorami. Zadanie procesorów nie-liści polega na przekazywaniu danych od i do liści. Średnica drzewa ortogonalnego $r = 2n$, a indeks spójności $s = 2$. Nakłady czasu na wykonanie podstawowych

funkcji komunikacyjnych podamy dla modelu obliczeń **BDSS** (patrz punkt 1). Rozsyłanie/odbiór pakietów jednorodnych w układzie: $1 \longrightarrow p/1 \longleftarrow p$, kosztuje $O(\log_2 n)$ jednostek czasu. Można uzasadnić to następująco. Założyliśmy, że pakiety mają długość $O(\log_2 n)$ bitów i że transmisja odbywa się bit po bicie. Droga od korzenia do liścia ma, w przypadku najgorszym, $O(\log_2 n)$ krawędzi (decyduje średnica sieci). Mamy więc $O(\log_2 n)$ łączy o długości $O(n)$ każde. Opóźnienie na jednym łączu wynosi $O(\log_2 n)$, zatem opóźnienie łączne wyniesie $O(\log_2^2 n)$ jednostek czasu. Transmisja $O(\log_2 n)$ bitów zajmuje, dzięki konwejeryzacji, tylko $O(\log_2 n)$ jednostek czasu ($O(1)$ *jednostek na 1 bit*). Tym samym komunikacja po drodze: korzeń - najbardziej oddalony liść innego drzewa - korzeń, kosztuje $O(\log_2^2 n) + O(\log_2 n) = O(\log_2^2 n)$. Jednoczesne rozsyłanie pakietu, z korzenia do wszystkich liści zajmuje $O(\log_2^2 n)$ jednostek czasu Nakłady na pozostałe funkcje komunikacyjne można oszacować podobnie.

Literatura

[1] Bertsekas D.P., Tsitsiklis J.N.: Parallel and Distributed Computation. Numerical Methods. Prentice-Hall, Englewood Cliffs,1989.

[2] Dekel E., Nassimi D., Sahni S.: Parallel Matrix and Graph Algorithms. SIAM Journal on Computing, Vol.10, No.4, 1981, pp.657-675

[3] Kindervater G.A.P., Lenstra J.K.: Parallel Computing in Combinatorial Optimization. Report OS - R8720. CWI, Amsterdam, November 1987.

[4] Nassimi D., Sahni S.: An Optimal Routing Algorithm for Mesh Connected Parallel Computers. Journal of the ACM, Vol.27, No.1, 1980, pp.6-29.

[5] Nath D., Maheshwari S.N., Bhatt P.C.P.: Efficient VLSI Networks for Parallel Processing Based on Orthogonal Trees. IEEE Transactions on Computers, Vol.C-3, No.6, 1983, pp.569-581.

[6] Słomiński L.:Komunikacja w systemach wieloprocesorowych i jej wpływ na efektywność obliczeń. Raport ZPM-22/A1530/90. IBS PAN, Warszawa 1990.

[7] Thompson C.D.: Area-Time Complexity for VLSI. Proc. of the ACM 11th Annual ACM Symp. on Theory of Computing. 1979, pp.81-88.

ISBN 83-900412-1-9.