Message Routing Mechanisms

Asist. Felician ALECU
Catedra de Informatică Economică, A.S.E. Bucureşti

The main goal of any routing method is to minimize the total communication delay between sender and receiver. If the message traffic in the network is relatively low, then minimal delay is obtained by choosing a path with a minimum number of connecting links. Dynamic routing has the advantage that it can respond to changes in the network traffic in order to avoid congestion.
Keywords: *routing mechanism, interconnection network, parallel computer, store and forward, cut-through.*

Introducere

Pentru conectarea procesoarelor şi a blocurilor de memorie ce compun un sistem paralel se folosesc reţelele de interconectare. O astfel de reţea este formată din noduri conectate între ele prin intermediul legăturilor. În nodurile reţelei se pot afla procesoare sau blocuri de memorie. De asemenea pot exista şi noduri complexe formate din procesoare şi blocurile de memorie ataşate. Topologia reţelei are o mare influenţă asupra performanţelor de ansamblu şi asupra costurilor.

O reţea de interconectare reprezintă în esenţă un graf G = (V, M), unde V este mulţimea vârfurilor (nodurilor) iar M mulţimea muchiilor (conexiunilor). O muchie între două noduri există numai dacă cele două noduri sunt conectate. Legăturile dintre nodurile grafului pot fi direcţionate (orientate) sau nedirecţionate (neorientate).

Reţelele de interconectare utilizate în cadrul calculatoarelor paralele pot fi:

- *reţele statice* – sunt formate din conexiuni fixe, punct la punct, între nodurile reţelei. Reţelele statice sunt în general utilizate în cazul calculatoarelor cu transfer de mesaje.
- *reţele dinamice* – sunt compuse din conexiuni variabile în timp. Nodurile sunt conectate între ele în mod dinamic cu ajutorul comutatoarelor. Reţele dinamice sunt utilizate cu preponderenţă pentru construirea calculatoarelor cu memorie partajată.

Mesajul reprezintă unitatea logică de comunicare între două noduri. Mecanismele de rutare determină calea pe care o urmează un mesaj pentru a ajunge de la nodul sursă la cel destinaţie. Pe baza adreselor nodurilor sursă şi destinaţie, mecanismul de rutare va genera una sau mai multe căi prin care mesajul ar putea fi transferat în cadrul reţelei statice de interconectare.

În funcţie de lungimea drumului generat, mecanismele de rutare se pot clasifica în două categorii:

- *mecanisme minimale* – aceste mecanisme selectează întotdeauna calea cea mai scurtă dintre nodurile sursă şi destinaţie. La fiecare pas de transfer mesajul va fi adus mai aproape de nodul destinaţie, cu toate că o astfel de abordare poate conduce la congestia unor părţi din reţea.
- *mecanisme neminimale* – acestea pot selecta o cale mai lungă din dorinţa de a evita aglomerarea reţelei. Nodurile care sunt congestionate vor fi evitate chiar dacă acest lucru atrage după sine obţinerea unei rute care trece prin mai multe noduri.

Adaptarea mecanismelor de rutare la starea curentă a reţelei reprezintă un alt criteriu de clasificare a acestora:

- *mecanisme deterministe* – determină calea de urma pe baza adresei nodului sursă şi a celui destinaţie fără a ţine cont de starea curentă a reţelei şi de aglomerările înregistrate. Folosirea unui astfel de mecanism poate conduce la o utilizare neechilibrată a reţelei de interconectare.
- *mecanisme adaptive* – rutarea unui mesaj se face ţinând cont atât de adresele nodurilor sursă şi destinaţie cât şi de starea curentă a reţelei de interconectare. Astfel, transferul mesajului se va face astfel încât să se evite

zonele în care se înregistrează aglomerări.
Rutarea ordonată după dimensiuni reprezintă cea mai utilizată tehnică minimală de rutare deterministă. Astfel, mesajul este deplasat mai întâi pe prima dimensiune, apoi pe cea de-a doua şi tot aşa până la livrarea acestuia.
Pentru transmiterea unui mesaj, acesta este divizat într-un număr arbitrar de pachete care se vor deplasa în mod independent prin reţeaua de interconectare către nodul destinaţie. Din acest motiv este nevoie ca fiecare pachet să conţină adresa nodului destinaţie şi un număr de secvenţă care va fi folosit la reasamblarea mesajului original, datorită faptului că pachetele vor fi recepţionate asincron de către nodul destinaţie. De multe ori pachetele sunt divizate la rândul lor în segmente de lungime fixă iar informaţiile de rutare (adresa nodului destinaţie şi numărul de secvenţă al pachetului) nu sunt conţinute decât în headerul pachetului.
Timpul consumat pentru comunicaţia informaţiilor între procesoare este o componentă importantă a timpului de execuţie a programelor paralele. Timpul de transfer al mesajelor în reţelele statice de interconectare este influenţat de:
- timpul de start (*startup time,* t_s) reprezintă timpul necesar pentru pregătirea mesajului în nodul sursă (pregătirea pachetelor, corecţia erorilor, executarea algoritmului de rutare)
- timpul în nod (*per-hop time,* t_h) este timpul consumat în fiecare nod intermediar pentru executarea funcţiei de rutare care va determina următorul nod spre care va fi transmis mesajul
- timpul de transfer per cuvânt (*per-word time,* t_w) reprezintă timpul necesar unui cuvânt pentru a traversa distanţa dintre două noduri direct conectate.

Acest timp de transfer depinde metoda de comutare utilizată - comutare de pachete sau comutare de circuite - şi de mecanismul de rutare folosit - mecanismul memorează-şi-înaintează (*store and forward*) sau mecanismul prin diviziunea mesajelor (*cut-through*).
Comutarea de pachete a fost folosită încă de la primele tipuri de calculatoare şi presupune împărţirea mesajului în mai multe pachete care vor fi transmise independent unul de altul prin reţea. În antetul pachetului este memorată informaţia de rutare iar datele ce vor fi transmise sunt memorate în corpul acestuia. Mesajul parcurge o serie de noduri intermediare pentru a ajunge la destinaţie. Conform principiului *memorează-şi*-înaintează, în fiecare nod intermediar există un buffer de memorie în care este memorat întregul mesaj înainte de a fi transmis mai departe. Timpul necesar unui pachet pentru a ajunge la destinaţie este $t = \frac{D}{L} d$, unde D reprezintă dimensiunea mesajului, L lăţimea de bandă canalului de comunicaţie iar d este distanţa dintre nodurile sursă şi destinaţie exprimată în număr de conexiuni. Datorită faptului că timpul de transfer al pachetului este direct proporţional cu distanţa dintre noduri, diametrul reţelei devine un factor deosebit de important pentru reţelele care folosesc comutarea de pachete.
Comutarea de circuite este o metodă asemănătoare sistemului telefonic. Transferul unui mesaj se realizează în trei faze: mai întâi se stabileşte legătura între nodurile sursă şi destinaţie prin intermediul unui pachet special, apoi se transmite mesajul dorit după care conexiunea este eliberată. Avantajul major al comutării de circuite faţă de tehnica de comutare de pachete îl reprezintă dispariţia zonelor tampon intermediare necesare stocării pachetelor. Timpul de transfer al mesajului este egal cu $t = \frac{D_1}{L} d + \frac{D_m}{L}$, unde D_1 reprezintă dimensiunea pachetului iniţial folosit pentru realizarea legăturii, D_m este dimensiunea mesajului ce se transmite iar d semnifică distanţa dintre nodurile sursă şi destinaţie exprimată în număr de conexiuni. Timpul de transfer devine independent de distanţa dintre procesoare şi de diametrul reţelei de interconectare atunci când $D_1 << D_m$.
Metoda diviziunii mesajelor combină caracteristicile comutării de circuite şi pe cele ale comutării de pachete. Mesajul este divizat în pachete mai mici care sunt trimise unul după altul (în manieră pipeline) către nodul destinaţie iar nodurile intermediare dispun de buffere în care se pot stoca pachetele ale mesajului.

Mecanismul de rutare *memorează-şi-înaintează*

Pentru rețelele care implementează acest mecanism de rutare, la nivelul fiecărui nod al rețelei există un buffer în care sunt memorate segmentele unui pachet. Transferul unui pachet de la nodul sursă la cel destinație se face prin tranzitarea unor noduri intermediare. În fiecare astfel de nod intermediar, pachetul este memorat în întregime în buffer-ul local după care este trimis mai departe către nodul următor (figura 1).

Dacă presupunem că se doreşte transmiterea unui mesaj compus dintr-un număr de *m* cuvinte iar nodurile sursă şi destinație sunt separate printr-un număr de *l* conexiuni, atunci timpul total de comunicație pentru mesajul în cauză este: $t_{MI} = t_s + (mt_w + t_h)l$

Pentru marea majoritate a rețelelor statice de interconectare şi a algoritmilor paraleli implementați, timpul în nod (t_h) este neglijabil iar cel de start (t_s) nu depinde de lungimea mesajului sau de distanța dintre noduri. Din acest motiv putem concluziona că timpul de transfer al unui mesaj între două noduri în rutarea memorează-şi-înaintează este de ordinul produsului dintre lungimea mesajului şi distanța dintre noduri: $t_{MI} = O(ml)$

Rutarea de tip memorează-şi-înaintează a fost utilizată cu precădere în cadrul primei generații de calculatoare paralele şi are ca principal dezavantaj faptul că exploatează ineficient resursele de comunicație ale rețelei.

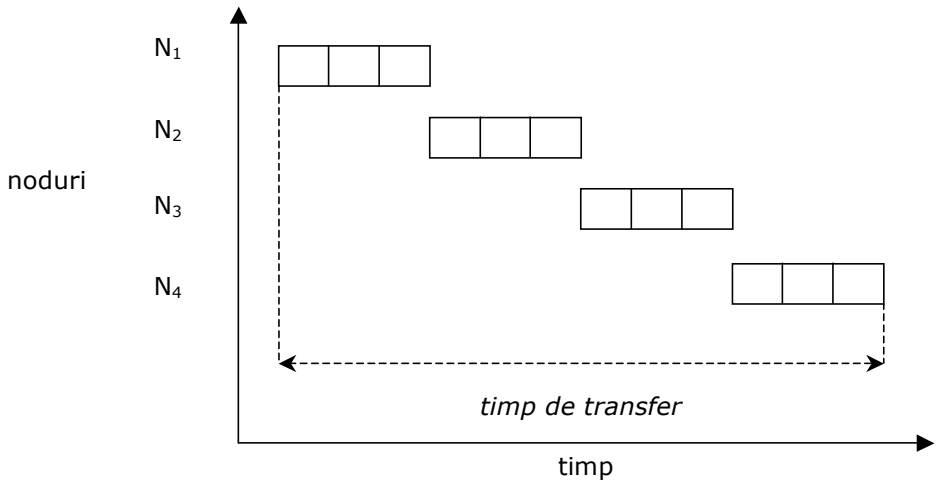

Fig. 1. Rutarea de tip *memorează-şi-înaintează*

Mecanismul de rutare prin diviziunea mesajelor

Acest mecanism nu mai necesită memorarea întregului pachet într-un nod intermediar înainte ca acesta să fie trimis mai departe către nodul următor. Astfel, segmentele recepționate sunt memorate până în momentul în care soseşte segmentul care conține informațiile de direcționare (adresa nodului destinație). Acesta va fi trimis mai departe către următorul nod iar toate celelalte segmente ale pachetului care sosesc nu mai sunt memorate în nodul intermediar ci vor fi imediat trimise nodului următor. Folosind acest mecanism nu mai este nevoie ca întregul pachet să fie recepționat pentru ca acesta să poată fi trimis mai departe către nodul destinație.

O variantă frecvent utilizată a mecanismului de rutare prin divizarea mesajului este algoritmul de rutare prin şerpuire (*wormhole*) prezentat în figura următoare (figura 2).

Dacă se doreşte transferul unui mesaj de lungime *m* cuvinte între două noduri aflate la distanță *l*, atunci timpul total de comunicație este $t_{DM} = t_s + lt_h + mt_w(1 + (l-1)/k)$, unde *k* reprezintă numărul de segmente în care a fost divizat mesajul. Dacă distanța dintre nodurile sursă şi destinație este mult mai mică decât numărul de segmente ale mesajului ($l << k$), atunci timpul total de transfer devine $t_{DM} = t_s + lt_h + mt_w$.

Din acest motiv putem concluziona că timpul de transfer între două noduri al unui mesaj în rutarea prin divizarea mesajelor este de ordinul sumei dintre lungimea mesajului şi distanţa dintre noduri: $t_{MI} = O(l + m)$.

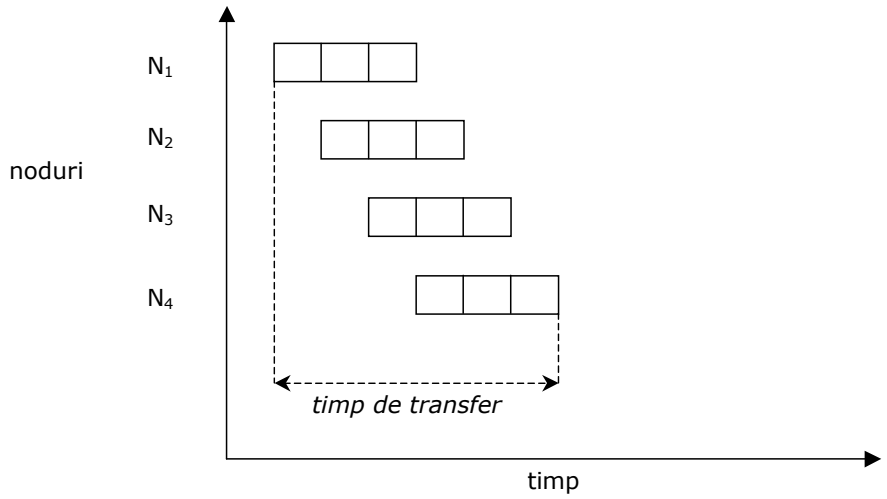

Fig. 2. Rutarea prin diviziunea mesajelor

Transferul de date între două noduri ale unei reţele de interconectare poate fi descris prin intermediul unor operaţii de comunicare de bază, care sunt frecvent utilizate de către algoritmii paraleli: transferul între două noduri, difuziunea, deplasarea circulară, etc. Eficienţa algoritmilor paraleli este influenţată în mod hotărâtor de modul în care sunt implementate aceste operaţii de bază.

Dintre toate aceste operaţii de bază, transferul unui mesaj între două noduri reprezintă o comunicaţie punct la punct în timp ce toate celelalte operaţii implementează o comunicaţie colectivă deoarece implică un grup de noduri.

Bibliografie

[Asm03] S. Asmussen, *Applied Probability and Queues*, Springer, 2003

[Dod02] Gh. Dodescu, B. Oancea, M. Răceanu, *Procesare paralelă*, Editura Economică, Bucureşti, 2002

[Jor02] H. F. Jordan, H. E. Jordan, *Fundamentals of Parallel Computing*, Prentice Hall, 2002

[Tan98] A. S. Tanenbaum, *Reţele de calculatoare*, Computer Press Agora, Bucureşti, 1998