УДК 519.68

В.Я. Лісовець, Г.Г. Цегелик
Львівський національний університет імені Івана Франка,
кафедра математичного моделювання соціально-економічних процесів

ОПТИМАЛЬНІ СТРАТЕГІЇ ПАРАЛЕЛЬНОГО ПОШУКУ ІНФОРМАЦІЇ У ПОСЛІДОВНИХ ФАЙЛАХ БАЗ ДАНИХ

© Лісовець В.Я., Цегелик Г.Г., 2008

Будуються оптимальні стратегії пошуку записів у послідовних файлах, які зберігаються у зовнішній пам'яті багатопроцесорної ЕОМ, для таких законів розподілу ймовірностей звертання до записів, як рівномірний, "бінарний", Зіпфа та узагальнений, частковим випадком якого є розподіл, який наближено задовольняє правило "80 – 20". За критерій ефективності взято математичне сподівання загального часу, необхідного для пошуку запису у файлі.

Optimal strategist of record searching in sequenced files stored in external memory is made for different probability distribution of record request frequency (discrete uniform, binomial, Zipf and generalized the partial occasion of witch is the probability distribution approximately satisfying the rule "80 – 20"). The mathematical expectation of total time needed for search of a record in file is taken as a criterion of effectiveness.

Вступ

Застосування паралельних обчислювальних систем є стратегічним напрямком розвитку обчислювальної техніки. Це викликано обмеженістю максимально можливої швидкодії звичайних послідовних ЕОМ, а також наявністю обчислювальних задач, для розв'язування яких можливості існуючих засобів обчислювальної техніки недостатні.

Проблема створення високопродуктивних обчислювальних систем належить до переліку найскладніших науково-технічних задач. Організація паралельних обчислень здійснюється переважно за рахунок уведення надлишкових функціональних пристроїв (декількох процесорів). Якщо поділити застосовані алгоритми на інформаційно незалежні частини й виконувати кожну частину обчислень на різних процесорах, то можна прискорити процес обчислень. Такий підхід дає змогу виконувати необхідні обчислення з меншими затратами часу. Одержання максимально можливого прискорення обмежується тільки кількістю наявних процесорів і "незалежних" частин алгоритму.

Однак треба зазначити, що сьогодні паралельні обчислення не є настільки поширеними, як могли очікувати дослідники. Однією з можливих причин такої ситуації була донедавна висока вартість високопродуктивних багатопроцесорних систем. Та ситуація кардинально змінилася з впровадженням порівняно дешевих багатоядерних процесорів, які вже набули масового застосування. Інша причина стримування поширення паралельних систем полягає в тому, що для проведення обчислень необхідно замінити традиційну послідовну технологію розв'язування задач на ЕОМ паралельною.

Завдяки високій надійності та продуктивності багатопроцесорні ЕОМ широко використовують для підтримки й організації великих баз даних (БД). Під час розв'язування різноманітних задач із використанням БД основний акцент переноситься з процедур опрацювання інформації на процедури організації збереження та пошуку інформації в них. Тому продуктивність обчислювальних систем, орієнтованих на роботу з великими БД, значною мірою визначається ефективністю методів пошуку інформації.

У [3–5] проаналізовано методи m-паралельного послідовного перегляду та m-паралельного блочного пошуку записів для різних законів розподілу ймовірностей звертання до записів.

У цій статті, використовуючи метод m-паралельного послідовного перегляду, побудуємо оптимальні стратегії пошуку записів в послідовних файлах, які зберігаються у зовнішній пам'яті багатопроцесорної ЕОМ.

Побудову оптимальних стратегій проведемо для рівномірного розподілу ймовірностей звертання до записів і таких законів нерівномірного розподілу ймовірностей, як [1–8]:

- “бінарний” розподіл

$$p_i = \frac{1}{2^i},\ i = 1,2,\dots,N\text{-}1,\ p_N = \frac{1}{2^{N-1}},$$

де p_i – ймовірність звертання до i-го запису, N – кількість записів у файлі;

- закон Зіпфа

$$p_i = \frac{1}{iH_N},\quad i = 1,2,\dots,N,$$

де $H_N = \sum_{k=1}^{N} \frac{1}{k}$;

- узагальнений закон розподілу

$$p_i = \frac{1}{i^{(c)} H_N^{(c)}},\ i = 1,2,\dots,N,$$

де c ($0<c<1$) – будь-який параметр і $H_N^{(c)} = \sum_{k=1}^{N} \frac{1}{k^c}$.

Зауважимо, що у випадку однопроцесорних ЕОМ ефективність методів пошуку, а також побудова оптимальних стратегій пошуку інформації ву послідовних файлах, для різних законів розподілу ймовірностей звертання до записів розглянуті в [1–2, 6–10].

Побудова оптимальних стратегій пошуку інформації

Розглянемо послідовний упорядкований файл, що зберігається у зовнішній пам’яті ЕОМ, до складу якого входять m процесорів, які працюють паралельно і мають спільне поле пам’яті. Припустимо, що файл, який містить N записів, поділений на n блоків, у кожному з яких є ml записів. Нехай $a_0 = b_0 + d_0ml$ – час читання блоку записів в основну пам’ять, де b_0, d_0 – деякі сталі; t_0 – час виконання операції m-паралельного послідовного перегляду записів в основній пам’яті; p_i – ймовірність звертання до i-го запису файла, E – математичне сподівання загального часу, необхідного для пошуку запису у файлі. Приймаємо, що для пошуку запису спочатку відбувається послідовне читання блоків записів в основну пам’ять і їх m-паралельний послідовний перегляд. Тоді

$$E = \sum_{k=1}^{n}\sum_{i=1}^{l}\sum_{j=1}^{m}\{ka_0 + [(k-1)l + i]t_0\}p_{(k-1)ml+(i-1)m+j}\cdot$$

Знайдемо явний вираз для E у випадку різних законів розподілу ймовірностей звертання до записів і визначимо значення параметрів n і l, за яких математичне сподівання досягає мінімуму.

Рівномірний розподіл. Якщо розподіл імовірностей звертання до записів є рівномірний, то для E одержуємо вираз

$$E = \frac{1}{2}\left[(n+1)\cdot\left(b_0 + \frac{d_0N}{n}\right) + \left(\frac{N}{m}+1\right)t_0\right].$$

Функція E досягає мінімуму, якщо $n = (d_0N/b_0)^{1/2}$.

“**Бінарний” розподіл.** Нехай імовірності звертання до записів задовольняють “бінарний” розподіл. Тоді для E, аналогічно як в [8], матимемо формулу

$$E = \frac{2^{ml}}{2^{ml}-1}\left(1-2^{-N}\right)\cdot a_0 + \left[\frac{2^m}{2^m-1}\left(1-2^{-N}\right) + \frac{n-l}{2^N}\right]t_0.$$

Нехтуючи нескінченно малою величиною 2^{-N}, із достатньо високою точністю можемо прийняти

$$E = \frac{2^{ml}}{2^{ml}-1}a_0 + \frac{2^m}{2^m-1}t_0,$$

або

$$E = \frac{2^{ml}}{2^{ml}-1}(b_0 + d_0 ml) + \frac{2^m}{2^m - 1} t_0 .$$

Для визначення параметра l, за якого функція E досягає мінімуму, отримаємо рівняння

$$2^{ml} = 1 + \left(ml + \frac{b_0}{d_0} \right) \ln 2 .$$

Одержане рівняння має розв'язок $l_0 \geq 1$, якщо для m виконується умова

$$2^m \leq \left(m + \frac{b_0}{d_0} \right) \ln 2 + 1 .$$

Закон Зіпфа. Нехай імовірності звертання до записів задовольняють закон Зіпфа. Тоді для E, аналогічно як в [8], одержуємо вираз

$$E = \frac{1}{H_N} \{ [(n+1)H_N - S_{ml}(n)](a_0 + t_0 l) + [(1-l)H_N + lS_{ml}(n) - S_m(nl)] \cdot t_0 \},$$

де

$$S_{ml}(n) = \sum_{k=1}^{n} H_{kml} , \quad S_m(nl) = \sum_{k=1}^{nl} H_{km} .$$

Якщо для сум $S_{ml}(n)$ і $S_m(nl)$ використати відповідно апроксимації [8]:

$$S_{ml}(n) \approx n(H_N - 1) + \frac{1}{2} \ln n + C_1 , \qquad S_m(nl) \approx nl(H_N - 1) + \frac{1}{2} \ln nl + C_1 ,$$

де $C_1 = 0{,}5 \ln 2\pi$, то з достатньо високою точністю можемо прийняти

$$E = \frac{1}{H_N} \left[\left(H_N + n - \frac{1}{2} \ln n - C_1 \right) \left(b_0 + \frac{d_0 N}{n} \right) + \left(H_N + \frac{N}{m} - \frac{1}{2} \ln \frac{N}{m} - C_1 \right) t_0 \right].$$

Для наближеного обчислення значення параметра n, за якого E досягає мінімуму, маємо рівняння

$$2n^2 - n = \frac{d_0 N}{b_0} (2H_N + 1 - \ln n - 2C_1).$$

Узагальнений закон. Нехай імовірності звертання до записів задовольняють узагальнений закон розподілу. Тоді для визначення математичного сподівання E, аналогічно як в [8], маємо формулу

$$E = \frac{1}{H_N^{(c)}} \left\{ \left[(n+1)H_N^{(c)} - S_{ml}^{(c)}(n) \right] \cdot (a_0 + t_0 l) + \left[(1-l)H_N^{(c)} + lS_{ml}^{(c)}(n) - S_m^{(c)}(nl) \right] \cdot t_0 \right\},$$

де

$$S_{ml}^{(c)}(n) = \sum_{k=1}^{n} H_{kml}^{(c)} , \quad S_m^{(c)}(nl) = \sum_{k=1}^{nl} H_{km}^{(c)} .$$

Якщо для $S_{ml}^{(c)}(n)$ і $S_m^{(c)}(nl)$ використати відповідно апроксимації [8]:

$$S_{ml}^{(c)}(n) \approx nH_N^{(c)} + \frac{N^{1-c}}{1-c} \left(\frac{c-1}{2-c} n + \frac{a^{(c)}(n)}{n^{1-c}} \right), \quad S_m^{(c)}(nl) \approx nlH_N^{(c)} + \frac{N^{1-c}}{1-c} \left(\frac{c-1}{2-c} nl + \frac{a^{(c)}(nl)}{(nl)^{1-c}} \right),$$

де

$$a^{(c)}(k) = H_k^{(c-1)} - \frac{1}{2-c} k^{2-c}$$

повільно зростаюча функція, з достатньо високою точністю можемо прийняти

$$E = \frac{1}{H_N^{(c)}} \left\{ \left[H_N^{(c)} - \frac{N^{1-c}}{1-c} \left(\frac{c-1}{2-c} n + \frac{a^{(c)}(n)}{n^{1-c}} \right) \right] \left(b_0 + \frac{d_0 N}{n} \right) + \right.$$

$$+\left[H_N^{(c)}+\frac{N^{1-c}}{1-c}\left(\frac{c-1}{2-c}\frac{N}{m}+\frac{a^{(c)}(N/m)}{(N/m)^{1-c}}\right)\right]t_0\Bigg\}.$$

Обчислимо похідну від функції E по n, підставляючи замість похідної від $\alpha^{(c)}(n)$ відповідно різницю $\alpha^{(c)}(n+1)-\alpha^{(c)}(n)$. Одержимо рівняння для наближеного обчислення значення параметра n

$$n^{3-c}+(2-c)\cdot\left(n+\frac{2-c}{1-c}\frac{d_0}{b_0}N\right)a^{(c)}(n)=$$

$$=(2-c)\frac{d_0}{b_0}N^c n^{1-c}H_N^{(c)}+\frac{2-c}{1-c}n\left(n+\frac{d_0}{b_0}N\right)\left(a^{(c)}(n+1)-a^{(c)}(n)\right).$$

Порівняння результатів

Оптимальні значення параметра l, за яких математичне сподівання загального часу, необхідного для пошуку запису у файлі, досягає мінімуму для $N = 10^6$, $d_0/b_0 = 20$, деяких m і різних законів розподілу ймовірностей звертання до записів наведено в таблиці.

Оптимальні значення параметра l для різних законів розподілу ймовірностей та різної кількості процесорів

m	Рівномірний	Узагальнений				Зіпфа	"Бінарний"
		c=0.2	c=0.4	c=0.6	c=0.8		
1	4464,29	4201,68	3831,42	3278,69	2421,31	1367,99	4
2	2232,14	2100,84	1915,71	1639,34	1210,65	684,00	2
4	1116,07	1050,42	957,85	819,67	605,33	342,00	1
5	892,86	840,34	766,28	655,74	484,26	273,60	1
10	446,43	420,17	383,14	327,87	242,13	136,80	1
20	223,21	210,08	191,57	163,93	121,07	68,40	1
40	111,61	105,04	95,79	81,97	60,53	34,20	1
50	89,29	84,03	76,63	65,57	48,43	27,36	1
100	44,64	42,02	38,31	32,79	24,21	13,68	1

Як бачимо з таблиці, із збільшенням кількості процесорів в k разів оптимальні значення параметра l зменшуються в k разів для всіх розглянутих законів розподілу ймовірностей звертання до записів, окрім "бінарного".

Рис. 1. Оптимальні значення параметра n у випадку $N = 10^6$ та $d_0/b_0 = 20$ для таких законів розподілу ймовірностей звертання до записів: рівномірного (c=0), узагальненого та Зіпфа (c=1)

На відміну від параметра l, оптимальні значення параметра n, за яких математичне сподівання загального часу, необхідного для пошуку запису у файлі, досягає мінімуму, залежать не від кількості процесорів, а лише від зміни закону розподілу ймовірностей звертання до записів для всіх законів розподілу, крім "бінарного". Залежність оптимального значення параметра n від зміни закону розподілу ймовірностей звертання до записів та залежність параметра n від зміни кількості процесорів у випадку "бінарного" закону розподілу для $N = 10^6$ і $d_0/b_0 = 20$ зображено відповідно на рис. 1 і 2.

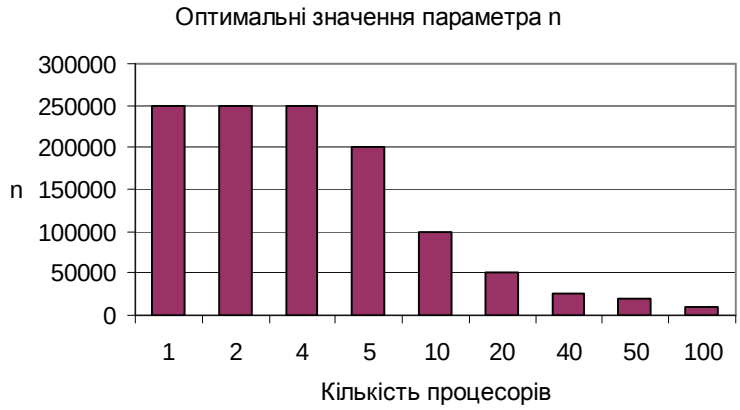

Рис. 2. Оптимальні значення параметра n у випадку "бінарного" закону розподілу ймовірностей звертання до записів, $N = 10^6$, $d_0/b_0 = 20$ та різної кількості процесорів

Висновки

Побудовано оптимальні стратегії пошуку записів в послідовних файлах, які зберігаються у зовнішній пам'яті багатопроцесорної ЕОМ, для таких законів розподілу ймовірностей звертання до записів, як рівномірний, "бінарний", Зіпфа та узагальнений, частковим випадком якого є розподіл, що наближено задовільняє правило "80 – 20". За критерій ефективності взято математичне сподівання загального часу, необхідного для пошуку запису у файлі. Знайдено оптимальні значення параметрів, за яких математичне сподівання досягає мінімуму.

1. Кнут Д. Искусство программирования для ЭВМ. Т. 3: Сортировка и поиск. – М.: Изд. дом "Вильямс", 2000. – 832 с. 2. Мартин Дж. Организация баз данных в вычислительных системах. – М.: Мир, 1980. – 644 с. 3. Лісовець В.Я., Цегелик Г.Г. Метод m-паралельного послідовного перегляду записів та його використання для пошуку інформації у послідовних файлах баз даних // Фізико-математичне моделювання та інформаційні технології. – 2007. – Вип. 5. – С. 109–119. 4. Лісовець В., Цегелик Г. Метод m-паралельного послідовного пошуку записів у файлах баз даних і його ефективність // Вісн. Львів. ун-ту. Сер. прикл. матем. та інформ. – 2006. – Вип. 13.– С. 177–186. 5. Лісовець В.Я., Цегелик Г.Г. Метод m-паралельного блочного пошуку записів у файлах баз даних та його ефективність // Відбір та обробка інформації. – 2007. – Вип. 27(103). – С. 87–92. 6. Мельничин А.В., Цегелик Г.Г. Аналіз методів пошуку інформації в файлах баз даних для різних законів розподілу ймовірностей звертання до записів // Комп'ютерні технології друкарства. – 2006. – № 15. – С. 95–112. 7. Мельничин А.В., Цегелик Г.Г. Методи пошуку інформації у файлах баз даних та їх ефективність для різних законів розподілу ймовірностей звертання до записів // Комп'ютерні технології друкарства. – 2006. – № 16. – С. 41–52. 8. Цегелик Г.Г. Организация и поиск информации в базах данных. – Львов: Вища шк., 1987. – 176 с. 9. Цегелик Г.Г., Мельничин А.В. Порівняльний аналіз ефективності методів пошуку інформації у файлах баз даних // Відбір і обробка інформації. – 2005. – № 23. – С. 135–142. 10. Цегелик Г.Г., Філяк М.І., Дороцька Х.С. Порівняльний аналіз ефективності методу блочного пошуку для різних законів розподілу ймовірностей звертання до записів // Комп'ютерні технології друкарства. – 2000. – № 5. – С. 320–326.