

УДК 681.327.2(045)

В. М. Синеглазов, д-р техн. наук, проф.,
А. С. Юрченко, канд. техн. наук

МИНИМИЗАЦИЯ СРЕДНЕГО ВРЕМЕНИ ОБРАЩЕНИЯ К ИЕРАРХИЧЕСКОЙ СИСТЕМЕ ПАМЯТИ

Институт электроники и систем управления НАУ, email: iesy@nau.edu.ua

На основе аналитической модели иерархической памяти современных ЭВМ решается задача выбора оптимальной конфигурации иерархической системы памяти. Показывается выполнимость необходимых и достаточных условий минимума среднего времени обращения к иерархии памяти.

Ключевые слова: архитектура современных ЭВМ, модель системы памяти, среднее время обращения, памяти.

Введение и постановка задачи. Архитектура современных ЭВМ предусматривает организацию многоуровневой иерархической памяти, обеспечивающую доступ центрального процессора к каждому уровню памяти. При этом производительность ЭВМ во многом определяется производительностью памяти, т. е. средним временем обслуживания одного обращения к памяти. Это среднее время обслуживания зависит от объема каждого уровня иерархической памяти, размера страниц, времени цикла устройств памяти на каждом уровне иерархии. Современные ЭВМ, оперирующие массивами данных большого объема, требуют использования устройств памяти с различным быстродействием и, следовательно, с различной стоимостью. Такие иерархические системы памяти очень дороги, и оценка производительности этих систем становится важной уже на стадии разработки. За последние годы появился ряд исследований, в которых оценивалось эффективность многоуровневых систем памяти [1 – 6].

В работе приводится корректное решение сформулированной в [6] задачи выбора оптимальной конфигурации иерархической системы памяти, работающей под управлением одного процессора.

Модель системы памяти. Пусть иерархия памяти состоит из различных уровней M_i , где $1 \leq i \leq N$. При этом предполагается, что память M_1 имеет наименьший размер и наименьшее время обращения к этой памяти, а память M_N – наибольший размер и наибольшее время обращения к этой памяти. Каждый уровень памяти M_i характеризуется [1] следующими параметрами: B_i – стоимость памяти M_i ; b_i – стоимость хранения единицы данных в памяти M_i (например, стоимость хранения одного байта); c_i – размер памяти M_i в единицах данных, $B_i = b_i c_i$; t_i – среднее время обращения к единице данных памяти M_i , точнее, время задержки вычислений, вызванное обращением к памяти M_i ; P_i – вероятность обращения к памяти M_i ; $P_i = p_i c_i$.

Системные характеристики, представляющее интерес, полная стоимость иерархии памяти и среднее время обращения к этой иерархии памяти и среднее время обращения к этой иерархии памяти описываются соответственно формулами:

$$T_{avg} = \sum_{i=1}^N t_i P_i, \quad (1)$$

$$S = \sum_{i=1}^N B_i. \quad (2)$$

Описанная задача при проектировании иерархии памяти заключается в минимизации среднего времени обращения при данной стоимости S посредством выбора параметров этой иерархии памяти. Такая минимизация требует анализа двух функций: 1) характеристики оборудования, т. е. соотношения между t_i и b_i ; 2) характеристики использования, которая указывает распределение обращений к иерархии памяти и тем самым задает связь P_i и c_i . Как и в работе [5], будем считать заданными число и размеры различных видов памяти M_i , а так же вероятности P_i обращения к памяти M_i .

Известно, что стоимость единицы памяти является убывающей функцией от времени обращения [5]. Поэтому модель памяти может быть задана функцией.

$$b_i(t) = b_0 t_i^{-\beta}, \quad t_i > 0. \quad (3)$$

Для реальных иерархических систем памяти значения параметра β лежат в окрестности 0,5 [4]. Использование модели памяти в виде (3) согласуется с реальными данными при изменении времени обращения на восемь порядков. Значения параметров b_0 и β могут быть различными в зависимости от рассматриваемых приложений.

Таким образом, задача выбора оптимальной конфигурации иерархической системы памяти сводится к минимизации среднего времени обращения в иерархической памяти, т. е. сводится к оптимизации формулы (1) при ограничении (2).

Оптимизация конфигурации памяти. Для решения задачи минимизации среднего времени обращения T_{avg} при ограничении на общую стоимость системы памяти S используем метод множителей Лагранжа. С этой целью построим функцию Лагранжа

$$F = \sum_{i=1}^N c_i p_i t_i + \lambda (\sum_{i=1}^N c_i b_0 t_i^{-\beta} - S).$$

Необходимым условием минимума среднего времени обращения T_{avg} является

$$\frac{dF}{dt_i} = c_i p_i + \lambda (-c_i b_0 \beta t_i^{-\beta-1}) = 0.$$

Из этого выражения получаем:

$$p_i = \lambda b_0 \beta t_i^{-\beta-1}; \quad (4)$$

$$t_i = (\lambda b_0 \beta / p_i)^{1/(\beta+1)}; \quad (5)$$

$$t_i^{-\beta} = (\lambda b_0 \beta / p_i)^{-\beta/(\beta+1)}. \quad (6)$$

Подставляем (6) в формулу (2), предполагая, что суммирование в этой формуле выполняется с индексом j (во избежание в дальнейших выкладках путаницы с использованием индекса i):

$$\sum_{j=1}^N c_j b_0 (\lambda b_0 \beta / p_j)^{-\beta/(\beta+1)} = S;$$

$$b_0 (\lambda b_0 \beta)^{-\beta/(\beta+1)} \sum_{j=1}^N c_j p_j^{\beta/(\beta+1)} = S;$$

$$\lambda b_0 \beta = \left(\frac{b_0}{S} \sum_{j=1}^N c_j p_j^{\beta/(\beta+1)} \right)^{(\beta+1)/\beta} = S. \quad (7)$$

Подставляя выражение (7) в формулу (5), имеем

$$t_i^0 = \left(\frac{b_0}{S} \sum_{j=1}^N c_j p_j^{\beta/(\beta+1)} \right)^{1/\beta} p_i^{-1/(\beta+1)}. \quad (8)$$

Отсюда среднее время обращения T_{avg} будет описываться выражением

$$\begin{aligned} T_{avg} &= \sum_{i=1}^N t_i^0 P_i = \sum_{i=1}^N c_i t_i^0 p_i = \sum_{i=1}^N c_i \left(\frac{b_0}{S} \sum_{j=1}^N c_j p_j^{\beta/(\beta+1)} \right)^{1/\beta} p_i^{\beta/(\beta+1)} = \\ T_{avg} &= \left(\frac{b_0}{S} \sum_{j=1}^N c_j p_j^{\beta/(\beta+1)} \right)^{1/\beta} \sum_{i=1}^N c_i p_i^{\beta/(\beta+1)} = S^{-1/\beta} b_0^{1/\beta} \left(\sum_{i=1}^N c_i p_j^{\beta/(\beta+1)} \right)^{(1+\beta)/\beta}. \end{aligned}$$

Теперь покажем, что найденное решение t_1^0, \dots, t_N^0 удовлетворяет достаточному условию минимума функции (1) при ограничениях (2), т. е.

$$d^2 F = \sum_{i=1}^N \sum_{k=1}^N \frac{d^2 F}{dt_i dt_k} \Big|_{(t_1^0, \dots, t_N^0)} dt_i dt_k > 0.$$

Так как

$$dF / dt_i = c_i p_i - \lambda c_i \beta b_0 t_i^{-\beta-1},$$

то справедливо $d^2 F / dt_i dt_j$ для всех $i \neq j$ и $d^2 F / dt_i^2 = \lambda c_i b_0 \beta (\beta + 1) t_i^{-\beta-2}$.

Поэтому выполняется соотношение

$$d^2 F / \lambda b_0 \beta (\beta + 1) \sum_{i=1}^N t_i^{-\beta-2} dt_i^2 > 0.$$

Поскольку для реальных иерархических систем памяти параметры p_i, c_i, b_0, β имеют положительные значения, то из формулы (4) и (8) следует, что и $\lambda > 0$.

Таким образом, формула (8) является решением задачи минимизации среднего времени обращения к иерархии памяти T_{avg} при ограничении на общую стоимость системы памяти S , а следовательно, и решением задачи выбора оптимальной конфигурации иерархической системы памяти.

Заключение. На основе математической модели иерархической системы памяти решена задача минимизации среднего времени обращения к иерархии памяти при ограничении на общую стоимость системы памяти.

Список литературы

1. *Catthoor K., Danckaert, C., Kulkarni, E., Brockmeyer P. G., Kjeldsberg T., Van Achteren and T. Omnes* Data Access and Storage Management for Embedded Programmable Processors. Kluwer Academic, – P. 356 – 364, – 2002.
2. *Budiu M., Goldstein S. C.* Optimizing Memory Accesses For Spatial Computation, in International Symposium on Code Generation and Optimization. – 2003 – P. 234 – 242.
3. *Baradaran N., Diniz P. C.* A register allocation algorithm in the presence of scalar replacement for fine-grain architecture, in the 2005 Conferenc on Design Automation and Testing in Europe, 2005 – P. 337 – 345.
4. *Benini A., Macii E., Macii, and M. Pancino.* Synthesis of application-specific Memories for power optimization in embedded systems. In Proc. of 37th Design Automation Conference, June 2000. – P. 300 – 303.

5. *Alex Settle and Daniel A. Connors*, Optimization for the Intel Architecture Register Stack, International Conference of in Proceedings of International Symposium of Code Generation and Optimization (CGO), San Francisco, March 2002 – P. 245 – 252.
6. *Welch T. A.* Memory hierarchy configuration analysis // IEEE. Trans. Comput. – V.27, №5. – 1998. – P. 408 – 413.

В. М. Синеглазов, О. С. Юрченко

До мінімізації середнього часу звернення до ієрархічної пам'яті

На основі математичної моделі ієрархічної системи пам'яті вирішено завдання мінімізації середнього часу звернення до ієрархії пам'яті з обмеженням на загальну вартість системи пам'яті.

V. M. Sineglazof, A. S. Yurchenko

To minimization of average time of access to hieratical memory

On the basis of mathematical model of the hierarchical system of memory the task of minimization of mean time of address to the hierarchy of memory is decided at a limit on the total worth of the system of memory.