О целесообразности изменения количества уровней в многоуровневой памяти

Д. В. Биматов
Томский государственный университет, Томск

Рассмотрим многоуровневую подсистему памяти с количеством уровней в иерархии равным $U > 1$ (рис. 1). На каждом уровне $u = \overline{1...U-1}$, кроме последнего, находится кэш; последний уровень $u = U$ занимает оперативное запоминающее устройство (ОЗУ).

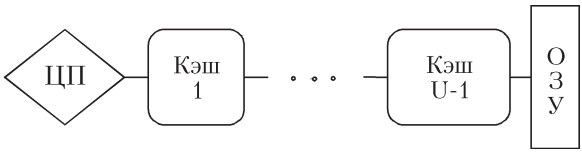


Рис. 1. Схема многоуровневой памяти

Среднее время доступа к адресуемым объектам многоуровневой памяти с U уровнями блокирующего типа определяется соотношением [1]:

$$T_U\left(K_1,\ R_1,\ K_2,\ R_2,\ ...,\ K_U,\ R_U\right) = \sum_{u=1}^{U} K_u \prod_{i=1}^{u-1} R_i \qquad (1)$$

где R_i — вероятность промаха в кэш уровня i, а K_i — время выбора блока из кэша уровня i, выраженное в тактах. В рамках рассматриваемой модели очевидно, что $R_U = 0$.

Предположим, что в подсистему иерархической памяти добавляется еще один уровень памяти перед уровнем с номером a. Пусть время доступа к добавляемому уровню памяти — K_*, а вероятность промаха — R_*. Естественно, что параметры добавляемого уровня памяти должны соответствовать логике построения многоуровневой памяти: кэш более высокого уровня характеризуется большим временем доступа и меньшей вероятностью промаха:

$$K_{a-1} \le K_* \le K_a$$
$$R_{a-1} \ge R_* \ge R_a$$

Целесообразность добавления уровня памяти в существующую систему перед уровнем с номером a выражается неравенством [2]:

$$\Delta T^{(a)} = T_U\left(K_1, R_1, ..., K_U, R_U\right) -$$
$$-T_{U+1}^{(a)}\left(K_1, R_1, ..., K_{a-1}, R_{a-1}, K_*, R_*, K_a, R_a, ..., K_U, R_U\right) \geq 0$$

Здесь $T_{U+1}^{(a)}\left(K_1, R_1, ..., K_{a-1}, R_{a-1}, K_*, R_*, K_a, R_a, ..., K_U, R_U\right)$ — среднее время доступа к подсистеме памяти с дополнительно введенным на a-ом уровне кэшем. Подставляя в последнее неравенство соотношение (1), после приведения подобных получим:

$$\frac{K_*}{1-R_*} \leq \sum_{u=a}^{U} K_u \prod_{i=a}^{u-1} R_i \qquad (2)$$

Неравенство (2) означает, что добавление еще одного уровня памяти в иерархию перед уровнем с номером а целесообразно, если отношение скорости обращения к добавляемому уровню памяти к вероятности попадания в добавляемый уровень памяти меньше среднего времени доступа к уровням памяти более высокого уровня. Для удобства выражение (2) можно переписать в виде

$$K_* \leq \Pi_* \cdot \left(\sum_{u=a}^{U} K_u \prod_{i=a}^{u-1} R_i \right)$$

или в виде

$$\Pi_* \geq \frac{K_*}{\sum_{u=a}^{U} K_u \prod_{i=a}^{u-1} R_i}$$

где $\Pi_* = 1 - R_*$ — вероятность попадания в добавляемый кэш.

Рассмотрим двухуровневую подсистему памяти «Кэш-ОЗУ». Целесообразность добавления кэша первого уровня будет выражаться неравенством:

$$\Delta T^{(1)} = T_2\left(K_1, R_1, \mathrm{K_{ОЗУ}}\right) - T_3^{(1)}\left(K_*, R_*, K_1, R_1, K_{ОЗУ}\right) \geq 0$$

Отсюда получаем:

$$\frac{K_*}{1-R_*} \leq K_1 + R_1 K_{ОЗУ}$$

Целесообразность добавления кэша второго уровня к исходной подсистеме памяти задается условием:

$$\Delta \mathrm{T}^{(2)} = T_2\left(\mathrm{K}_1, R_1, K_{ОЗУ}\right) - T_3^{(2)}\left(K_1, R_1, K_*, R_*, K_{ОЗУ}\right) \geq 0$$

которое приводит к неравенствам:

$$\frac{K_*}{1-R_*} \le K_{ОЗУ}$$

В общем случае вероятность промаха в кэш R (или вероятность попадания П = 1 – R) не известна и определяется как параметрами подсистемы памяти (объем кэша — V_i, объем ОЗУ — $V = V_{ОЗУ}$, коэффициент ассоциативности — A_i, где i — уровень кэша), так и выполняемой на вычислительной системе смеси задач. В данной работе используется модель с идеальной стратегией вытеснения блоков [3], т. е. предполагается, что заранее известно распределение вероятностей востребованности блоков основной памяти вычислителем:

$$p(i),\ \ i = \overline{0, \mathrm{V}-1},\ \ \sum_{i=0}^{V-1} p(i) = 1. \qquad (3)$$

Полагаем, что вероятности востребованности блоков основной памяти (3) определены на интервале $m = \overline{0, \alpha A_1 - 1}$ и убывают с ростом m по усеченному геометрическому закону (заданному на конечной области определения):

$$p(g + G_1 m) = p(g) \cdot q^m,\ g = \overline{0,\ G_1 - 1}$$

где G_1 — количество групп в кэше уровня 1.

Здесь $\frac{1}{A_1} \le \alpha \le \frac{M_1}{A_1}$ — доля ОЗУ, занятая востребованными вычислителем приложениями, выраженная в количестве объемов кэша первого уровня, а $0 \le q \le 1$ — параметр усеченного геометрического распределения.

Для усеченного геометрического распределения получены аналитические зависимости для вычисления вероятности промаха в кэш уровня 1

$$R_1 = 1 - \frac{1 + q - 2q^{A_1} + q^{\alpha A_1}(1-q)}{(1+q)\cdot\left(1 - q^{\alpha A_1}\right)},\ \ 0 \le q < 1,$$

и в кэш произвольного уровня u:

$$R_u = 1 - \frac{1 + q^{e_u} - 2q^{e_u A_u} + q^{\alpha A_1}\left(1 - q^{e_u}\right)}{\left(1 + q^{e_u}\right)\cdot\left(1 - q^{\alpha A_1}\right)},\ \ 0 \le q < 1,$$

где $e_u = G_u / G_1$ — отношение количества групп в кэше уровня u и кэше уровня 1.

Для равномерного распределения ($q = 1$) выражения для вычисления вероятности промаха в кэш многоуровневой памяти выглядят следующим образом:

$$R_1 = 1 - \frac{1}{\alpha},\ R_u = 1 - \frac{V_u}{\alpha V_1},\ u = \overline{2, U-1}.$$

Рис. 2 иллюстрирует в пространстве быстродействия и вероятности промаха в кэш добавляемого уровня области целесообразности добавления кэша к двухуровневой подсистеме памяти с параметрами: $R_1 = 0{,}5$, $K_1 = 8$, $K_{\text{ОЗУ}} = 64$. Область $0 \le R_* \le R_1$ соответствует добавлению уровня между центральным процессором и существующим кэшем (добавление кэша первого уровня), а область $R_1 \le R_* \le 1$ — добавлению кэша между существующим кэшем и ОЗУ (добавление кэша второго уровня).

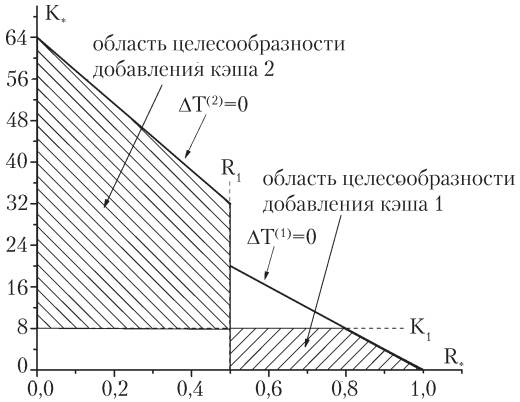


Рис. 2. Области целесообразности введения кэша первого и второго уровней

Рис. 3—5 показывают эффект от добавления кэша второго уровня к двухуровневой системе памяти с параметрами: $V_1 = 128$, $V_{\text{ОЗУ}} = 512$ ($\alpha = 4$), $K_1 = 8$, $K_{\text{ОЗУ}} = 64$ при равномерном распределении востребованности блоков ОЗУ вычислителем. На рис. 3 изображена зависимость среднего времени доступа от времени обращения к добавляемому кэшу K_*. Горизонтальная линия соответствует среднему времени доступа к исходной двухуровневой памяти. Видно, что чем больше объем добавля-

емого кэша, тем шире диапазон значений K_*, при которых трехуровневая память имеет меньшее среднее время доступа. Рис. 4 иллюстрирует влияние объема добавляемого кэша на среднее время доступа к модифицированной подсистеме памяти. Горизонтальная линия соответствует среднему времени доступа к исходной подсистеме памяти. Видно, что чем меньше время обращения к добавляемому кэшу, тем шире диапазон объемов добавляемого кэша V_*, при которых имеет смысл использовать трехуровневую память вместо двухуровневой. Наконец, рис. 5 представляет область целесообразности добавления кэша с объемом V_* и временем доступа K_*.

Рис. 6, 7 показывают эффект от добавления кэша второго уровня к двухуровневой подсистеме памяти при усеченном геометрическом распределении востребованности блоков ОЗУ вычислителем. Исходная подсистема памяти характеризуется параметрами: $V_1 = 16$, $V_{ОЗУ} = 64$ ($\alpha = 4$); $A_1 = 2$, $K_1 = 8$, $K_{ОЗУ} = 64$. Добавляемый кэш имеет объем V_2=32 и ассоциативность $A_2 = 2$. На рис. 6 приведен график среднего времени доступа исходной (T_2) и модифицированной ($T_3^{(2)}$) подсистем памяти от параметра усеченного геометрического распределения q при различных временах доступа к добавляемому кэшу K_2. Рис. 7 показывает выигрыш от добавления кэша при заданном параметре q и различных временах обращения к добавляемому кэшу K_2.

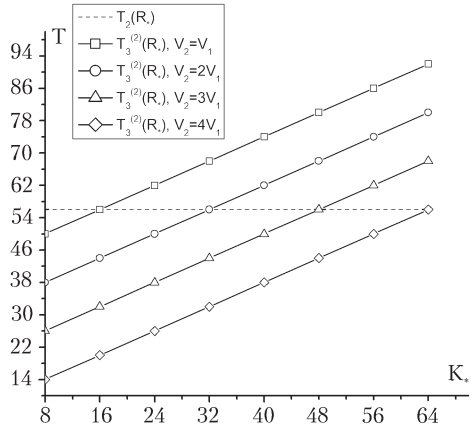


Рис. 3. Влияние времени обращения K_* к добавляемому кэшу второго уровня на среднее время доступа T при равномерном распределении

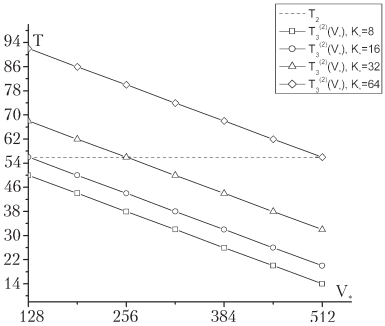


Рис. 4. Влияние объема V_* добавляемого кэша второго уровня на среднее время доступа T при равномерном распределении

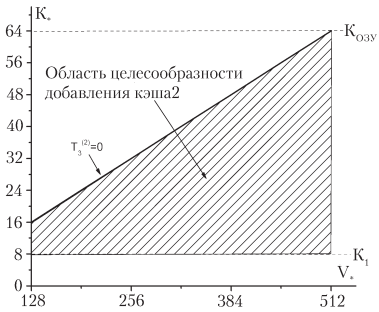


Рис. 5. Область целесообразности добавления кэша второго уровня с объемом V_* и временем доступа K_* при равномерном распределении

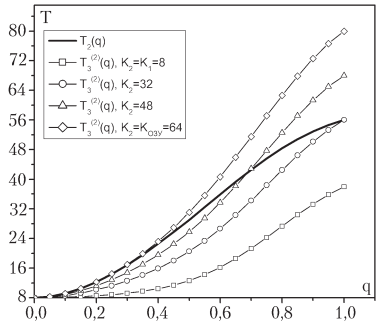


Рис. 6. Зависимость среднего времени доступа к исходной и измененной подсистеме памяти от параметра распределения q при различных временах доступа к добавляемому кэшу второго уровня K_2

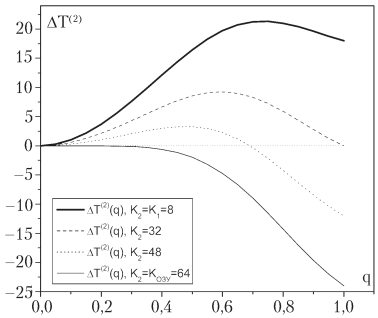


Рис. 7. Выигрыш от добавления кэша второго уровня с различными временами обращения K_2

Литература

1. **Сущенко М. С., Сущенко С. П.** Анализ эффективности многоуровневой памяти вычислительных систем // Обозрение прикл. и промышл. матем. — 2001. — Т. 8. — В. 1. — С. 336—337.

2. **Сущенко М. С., Сущенко С. П.** Условия целесообразности увеличения числа уровней иерархической памяти // Мат-лы Всерос. науч.-практ. конф. «Наука и практика: диалоги нового века». — Томск : Твердыня, 2003. — Ч. 3 Информационные технологии и математическое моделирование. — С. 194—196.

3. **Сущенко М. С., Сущенко С. П.** Анализ производительности множественного ассоциативного кэша // Вестник Томского гос. ун-та. — 2002. — № 275. — С. 218—223.