A Taxonomy for Cache Memory Misses

José Luis Hamkalo[1], Bruno Cernuschi-Frías[2]

[1] Universidad de Buenos Aires
Depto. de Electrónica
Paseo Colón 850 (1063)
Buenos Aires, Argentina
{jhamkal@fi.uba.ar}
[2] Universidad de Buenos Aires and CONICET

Abstract—

One way to understand the causes of cache memory misses is to use a classification for them. Usually statistical models such as the 3C model are used to make the classification. In the present work a new definition for the 3C model: compulsory, capacity and conflict misses are given. The corresponding operational definitions are given, which are based on the use of the LRU stack distances. The proposed model is called a deterministic 3C model or D3C. The D3C model classifies the memory references in an individual way, conforming a taxonomy, and then it is possible to analyze when a memory reference belongs to a given category. Also the passage of a given memory reference from one category to another when some cache parameter is modified may be studied. The D3C model does not present anomalies such as negative conflict miss rates, as in the 3C model. Several patterns for memory access are theoretically analyzed for the 3C and D3C models, showing that the results given by the D3C model are intuitive and have easy interpretation. The 3C model underestimates the conflict misses and overestimates the capacity misses when compared with the D3C model. This difference comes from the references that hit in the cache under study and miss in a fully associative cache of the same size with LRU replacement policy. The amount of these references was measured using SPEC95 benchmarks in trace driven simulations. It is shown that high percent values are obtained for these references for usual cache configurations, and therefore these references have an important participation in the cache statistics.

Keywords— **Cache, Model, Conflict.**

I. Introduction

The cache memory is a key computer element, which compensates the difference of speeds that exist between the processor and the main memory [SMI 82]. The understanding of the possible causes of the misses in cache memories is of great interest and can help in more effective designs of cache memory organizations. A way to understand the causes of misses is through a classification for them. A classification of very extended use is the 3C model [HIL 89], [PAT 95]. This model classifies cache memory misses in three types: compulsory, capacity and conflict. Each of the previous miss types is quantified in the following way: 1) conflict: it is the miss rate of the cache under study minus the miss rate of a LRU fully associative cache of the same size; 2) capacity: it is the miss rate of the LRU fully associative cache minus the miss rate for a LRU cache of infinite size; 3) compulsory: it is the miss rate of a LRU cache of infinite size. Figure 1 shows a possible situation for the representation of the misses for a set associative cache. In figure 1, U is the set of all the references to memory made by the program, B is the set of references that miss in the cache under analysis, A is the set of references that miss in a fully associative cache of the same size than the cache under study and use the LRU replacement policy, and C is the set corresponding to the memory references that miss in the cache of infinite size.

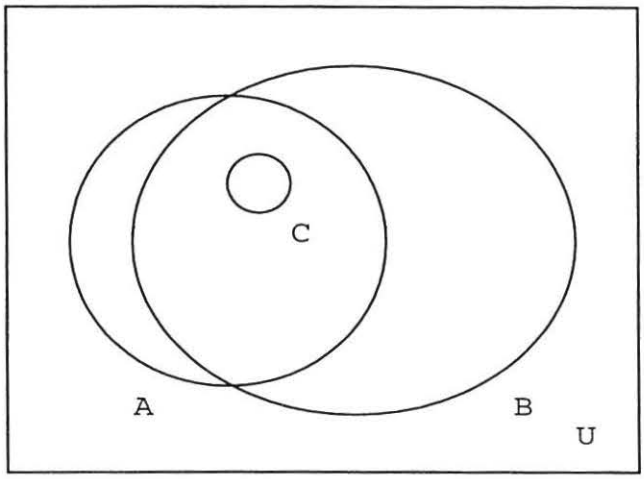

Fig. 1. Representation by sets of the hits and misses in a cache memory.

Using figure 1, the 3C model quantifies the misses in the following way:

$$\begin{cases} Compulsory & = \#C \\ Capacity & = \#A - \#C \\ Conflict & = \#B - \#A \end{cases} \quad (1)$$

Where #C, #A and #B are the number of elements in the sets C, A and B respectively. It is known that using the 3C model, in some cases it is possible to obtain negative conflict miss rates. This is a result of difficult interpretation

for the evaluation of cache memories. These cases appear when a set associative cache has a smaller miss rate than a fully associative cache ($\#A > \#B$). In the 3C model, when calculating the conflict misses, these are obtained as the total of references that miss in the cache under study ($\#B$), minus the total of references that miss in a fully associative cache ($\#A$), independently if some of these references hit in the cache under study (these references are given by $A - B$, where $A - B$ is set subtraction). Therefore a negative conflict miss rate implies that the number of references that miss in a fully associative cache and hit in the cache under study ($\#(A - B)$), is greater than the number of references that hit in the fully associative cache and miss in the cache under study ($\#(B - A)$).

Although the situation when a cache memory has a negative conflict miss rate has been catalogued as rare [LEB 94], it is necessary to evaluate the influence in the 3C model for the references that hit in the cache under study and miss in the fully associative. Such type of references have been called in [LEB 94] anticonflict misses, and were proposed as a possible correction term for the 3C model, although no operational definition was given for the evaluation and use of this correction term.

The statistical character of the 3C model, presents the additional disadvantage of not be able to permitting the individual classification of the memory misses.

Another classification for cache memory misses was proposed in [SUG 93]. The miss types for this classification are: compulsory, capacity, mapping and replacement policy. The Sugumar model uses the miss rates statistics of different cache organizations with optimal replacement to calculate the amount of each type of miss. Like the 3C model, the Sugumar model is a statistical model where the miss rate for each type of miss is calculated from arithmetic operations using the miss rate statistics of different cache organizations. Compulsory misses in the Sugumar model are equal to the corresponding compulsory misses in the 3C model. Capacity misses are defined as the extra misses occurring in a fully associative cache of the same size than the analyzed cache, simulated using an optimal replacement strategy. Mapping misses are due to a small degree of associativity in the cache and are defined as the additional misses that produces a cache similar to the cache under study but using an optimal replacement. Finally, replacement misses are the extra misses due to the use of a sub-optimal replacement strategy in the cache under study.

In the present work and on the basis of the 3C model, a new definition for each type of miss is given. The corresponding operational definitions are given based on the use of the LRU stack distances. The proposed model is called deterministic 3C model or D3C and is defined in section II. In section III the 3C and D3C models are analyzed for three paradigmatic models of memory reference. In section IV the D3C model is analyzed and in section V an experimental analysis is made. Finally in section VI the conclusions are given.

II. Definition of the D3C model

A. Model Definition

The 3C model is well known and it has been extensively used. The classification of misses in the types compulsory, capacity and conflict is conceptually clear and intuitive. It is advisable therefore, to maintain these types for the classification proposed here. Below are given the new definitions for the misses types given by the 3C model.

Definition: A compulsory miss is a memory reference that misses in the cache under study and also misses in a LRU fully associative cache of infinite size. A capacity miss is a memory reference that produces a noncompulsory miss in the cache under study, that also produces a miss in a LRU fully associative cache of the same size. The rest of the noncompulsory misses are defined as conflict misses, i.e. those references that miss in the cache under study and hit in the fully associative cache.

Using figure 1 again, for the D3C model it is:

$$\begin{cases} Compulsory & = C \\ Capacity & = (A \cap B) - C \\ Conflict & = B - A \end{cases} \qquad (2)$$

Note that in (2) the operations are performed between sets and therefore the misses are classified in an individual way.

In the following section an operational definition for each type of miss is given.

B. Operational Definitions

In this section the operational definitions for the D3C model are given. They are based on the use of the LRU stack distance. The LRU stack distance for a reference to a given memory block, is defined as the number of references to different memory blocks that were made from the last reference to the given block.

A reference that is a miss in a set associative cache of total size equal to L blocks, is a miss in a LRU fully associative cache of the same size if the LRU distance for such reference is greater than L. This fact is used to provide the operational definition for the D3C model.

Let L be the size in blocks units of the cache under study, and D the LRU distance of a reference that produces a miss. Then, the miss references are classified in the following way:

$$\begin{cases} Compulsory & : D \text{ not computable} \\ Capacity & : D > L \\ Conflict & : D \leq L \end{cases} \quad (3)$$

C. General Observations on the Model

The D3C model establishes the relations that allow to know when a reference to memory is compulsory, capacity or conflict. All the instructions or data that are evicted from the cache and then returned to it upon CPU request (i.e. producing a miss) without the intermediate request of more than L different blocks of memory are conflict misses. The references that produce misses of the type "ping pong" (i.e. the references that alternate in the cache, excluding one to the other) are clearly classified as conflict misses. This is because the LRU distance D for such type of references are very small and for typical cache sizes is $D < L$. In the same way a capacity miss implies that the reference that misses is requested after an intense activity in the cache, quantitatively, more than L different blocks have been required by the CPU from the time that the given reference was evicted from the cache.

The D3C Model is called deterministic since the references to memory are classified individually and not as the result of arithmetic operations between the statistics of the cache under study and a fully associative cache.

When classifying the misses in an individual way a taxonomy is obtained. In the same way it is possible to obtain the evolution of a reference within the classification when the parameters for the cache under study are changed.

Being the D3C model a taxonomy, it does not give rise to the possibility of anomalies such as negative conflict miss rates, independently of the load and the configuration of the cache under study. Also, some paradigmatic models of references to memory give results of very simple intuitive interpretation, as it is discussed next.

III. Application of the D3C Model to Some Paradigmatic Memory Reference Models

A. Introduction

In this section three paradigmatic memory reference models are analyzed, comparing the results that are obtained for the 3C and D3C models. The first model is a pattern of totally deterministic accesses to memory. The second model is a totally random pattern. Finally a sequence of accesses to memory that produces an anomalous behavior in FIFO caches is analyzed.

B. The Simple Loop Model

The simple loop model has been proposed and analyzed in [SMI 85] and further analyzed in [HAM 99]. This model reference iteratively a contiguous zone of memory. Each block is referenced at the same byte, being the stride between references exactly one cache block (except for the last reference that jumps backwards in memory to close the loop). The number of iterations of the loop is infinite.

For loops smaller than the cache, the miss rate is nearly zero, independently of the cache configuration. When the loop is greater than the cache, misses begin to take place. For loop sizes between L and $2L$, the set associative caches under LRU perform worse as the associativity increases. Therefore a fully associative cache has the worse miss rate and set associative caches perform better.

For example, for a two way set associative cache with 128 blocks (64 sets) and a loop of 160 blocks, the following results hold:

TABLE I
MISS RATIOS FOR A SIMPLE LOOP MODEL EXAMPLE FOR THE 3C AND D3C MODELS

	Compulsory	Capacity	Conflict	Total
$3C$	0	1	−0.4	0.6
$D3C$	0	0.6	0	0.6

In table I it is obtained that compulsory misses are zero in both models (strictly they tend to zero). This is because the number of dynamic references to memory is infinite over a finite zone of memory. It can be observed that for the D3C model all the misses are classified as capacity, whereas the 3C model presents a great anomaly with a negative conflict miss ratio. The results given by the D3C model can be explained in the following way: the simple loop model is the most ordered way to fill a LRU cache. This is so because the contiguous memory blocks referenced by the model map to contiguous sets, until the cache is completely filled. Therefore the existence of misses only can be attributed to a lack of capacity of the cache and not to some kind of interference between the references.

C. The Random Memory Access Model

Here a model that accesses memory totally at random is considered. In this model a contiguous region of memory is referenced at random, an infinite number of times. All the memory blocks have the same probability of being referenced at every moment. For the calculation of the hit ratio in the steady state an urn model is used. If the memory blocks (M blocks) are considered as white balls within a box and the balls whose corresponding memory block have a copy in the cache (L blocks) are marked with a red point, then if a

ball of the box is extracted at random (memory reference at random) the probability that the extracted ball has a red point (hit probability) is L/M. This urn model shows in a simple way the no dependency of the hit rate on the cache organization. Direct mapped, set associative and fully associative caches produce the same hit rate.

The application of the 3C and D3C models are analyzed next for the random memory access model for a direct mapped cache. The theoretical results obtained are given in table II, where N is the quotient $\frac{M}{L}$.

TABLE II
MISS RATIOS FOR A DIRECT MAPPED CACHE UNDER THE RANDOM MEMORY ACCESS MODEL FOR THE 3C AND D3C MODELS.

	Compulsory	Capacity	Conflict	Total
$3C$	0	$1-\frac{L}{M}$	0	$1-\frac{L}{M}$
$D3C$	0	$\frac{\binom{M-1}{L}-\frac{\binom{M-N}{L}}{N}}{\binom{M}{L}}$	$\neq 0$	$1-\frac{L}{M}$

From table II it is obtained that the compulsory miss ratio tends to zero in both models like in the simple loop model case. Since in the random memory access model the miss rate for a direct mapped cache is the same than that of a fully associative cache of the same size, for the 3C model all the misses are capacity misses and then the conflict miss ratio is zero. Table 2 results for the D3C model show that the conflict miss ratio is different from zero and depending on the sizes of M and L it may quantitatively turn out to be important (the expression was not written in table II for simplicity, but is obtained as the total miss ratio minus the capacity miss ratio). The presence of conflict misses for the random memory access model, as the D3C model shows, is easily interpretable and intuitive: in a random pattern of references to memory the probability that two or more references near in the time are mutually excluded from the cache ("ping pong" effect), is greater than zero. This behavior is captured and quantified correctly by the D3C model.

D. The Belady Anomaly

Here the 3C and D3C models are applied to a short series of memory references to a FIFO fully associative cache. This sequence produces greater number of misses when the cache size is increased, being this phenomenon known as the Belady anomaly [BEL'69]. The sequence of references to memory blocks is: 0 1 2 3 0 1 4 0 1 2 3 4. The caches analyzed have 3 and 4 blocks respectively and the resulting statistics are given in table III:

From table III there are 9 misses for the 3 blocks cache and for the 4 blocks cache there is an additional miss. The D3C model explains this behavior in the following way: the increase in the cache size produces a reduction in the capacity misses from 4 to 3. Nevertheless fully associative FIFO caches may present conflict misses and for this sequence in particular, the increase in the cache size introduces 2 conflict misses. These conflict misses are not totally compensated with the reduction in one for the capacity misses, giving as result an increase in one in the total amount of misses.

TABLE III
MISSES FOR TWO FULLY ASSOCIATIVE FIFO CACHES UNDER THE BELADY ANOMALOUS SEQUENCE FOR THE 3C AND D3C MODELS.

$L=3$	Compulsory	Capacity	Conflict	Total
$3C$	5	5	−1	9
$D3C$	5	4	0	9
$L=4$	Compulsory	Capacity	Conflict	Total
$3C$	5	3	2	10
$D3C$	5	3	2	10

IV. ANALYSIS OF THE D3C MODEL

A. Comparative Analysis With The 3C Model

Since references with LRU distances greater than L can't hit on a fully associative cache with LRU replacement, it results that all the misses for these type of caches under the D3C model must be of capacity. An important observation is that the capacity misses in the D3C model never can be more than the capacity misses given by the 3C model. From figure 1 it results that the number of elements in the set A is always greater or equal than the number of elements in the set $A \cap B$. From this observation, according to the definitions given for both models in sections I and II, it results that the capacity misses of the 3C model are always greater or equal than those of the D3C model. This shows that systematically the 3C model overestimates the capacity misses and underestimates the conflict misses with respect to the D3C model. The difference between both models comes from the anti-conflict references. In section V quantified results for these references for the SPEC95 benchmarks are given, showing the difference between the 3C and D3C models.

B. Mutation Analysis for Hits and Misses

One of the advantages of using a model that classifies the hits and misses individually, is that the passage from a category to another for an individual reference can be studied when some cache parameter is varied. We call this type of study mutation analysis. Next, a general framework for the mutation analysis is given for a LRU cache with fixed block size.

B.1 Cache size change with constant associativity

A cache size increase implies a greater number of sets and therefore a wider distribution of the memory blocks over the sets. No set of the increased cache will receive more references than its predecessor set in the original cache. Being the LRU policy a stack policy, no hit in the original cache can be a miss in the increased cache but it is possible that a certain number of misses could be transformed into hits. A reference that is a conflict miss in the original cache and continues being a miss in the increased cache, remains as a conflict miss since its LRU distance does not change (the LRU distance does not depend on the cache configuration) and the number of cache blocks is greater than in the original cache. A capacity miss in the original cache that also is a miss in the increased cache could remain as a capacity miss or mutate to a conflict miss if the LRU distance for that reference is less than the new number of blocks in the increased cache.

B.2 Associativity varies and the cache size remains constant

All the noncompulsory misses in a fully associative cache are capacity misses. If the associativity is reduced, some capacity misses could change to hits or remain as capacity misses, and some hits could change to conflict misses. One second reduction in the associativity could produce again the same kind of changes and it is also possible the remutation of some mutated hits to misses. A similar analysis may be done for the conflict references which can mutate to hits, and to remutate to conflict misses under changes in the associativity.

V. Experimental analysis

A. Methodology

For the evaluation of the D3C model the SPEC95 benchmarks for integer and floating point are used in trace driven simulations. The eight integer and the ten floating point benchmarks were simulated. For the integer benchmarks, an average of 700 million instructions by benchmark were simulated and for the floating point benchmarks an average of 1200 million instructions by benchmark were simulated. The traces were collected with the ATOM tool [SRI 94]. Caches from direct mapped to fully associative were simulated. The simulated caches range from 8Kbytes to 64kbytes and use the LRU replacement policy and a line of 32 bytes. The obtained mean miss rates for instructions and data caches are given in figures 2 to 5.

The horizontal axis of figures 2 to 5 represent the degree of associativity. For each cache size simulated, the maximum degree of associativity reached for the corresponding curve is equal to the cache total number of blocks and occurs when the cache is fully associative. The fully associative cache miss rate is projected horizontally to compare with

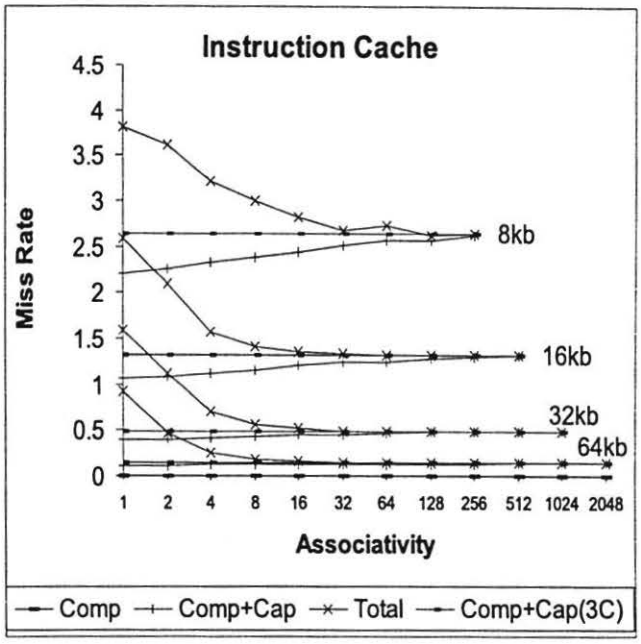

Fig. 2. SPECint95 mean miss rates for instruction caches.

the 3C model. The compulsory misses are obtained directly from the figures and are equal for both models. The capacity misses are obtained as the difference between the respective "Compulsory + Capacity" curves for each model and the curve of compulsory misses (for the 3C model the "Compulsory + Capacity" curve agrees with the miss rate of the fully associative cache, which projects horizontally as it was mentioned previously). Also the total miss rates are plotted. The difference between the values given by the "Total" and the "Compulsory + Capacity" curves indicates the conflict miss rate for each configuration. Finally the anticonflict misses are obtained as the difference between the "Compulsory + Capacity" curves for the two models. This is justified in the following way: the fully associative cache miss rate quantifies the totality of the references that have a LRU distance greater than L (or not computable) since all of them and only them produce misses. The fraction of these references that miss in a set associative cache are capacity or compulsory misses for the D3C model, being the rest of the references of the anticonflict type, since they hit in a set associative cache and miss in a fully associative cache.

B. Analysis of the Results

A first observation from figures 2, 3, 4 and 5 shows that the compulsory miss rate is totally negligible for instruction caches and very low for data caches. The conflict miss rates appear to be high for direct mapped caches and diminish

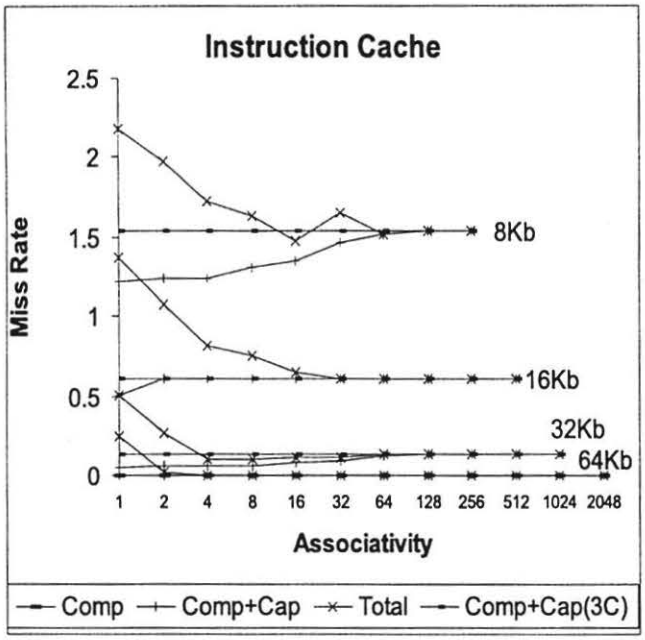

Fig. 3. SPECfp95 mean miss rates for instruction caches.

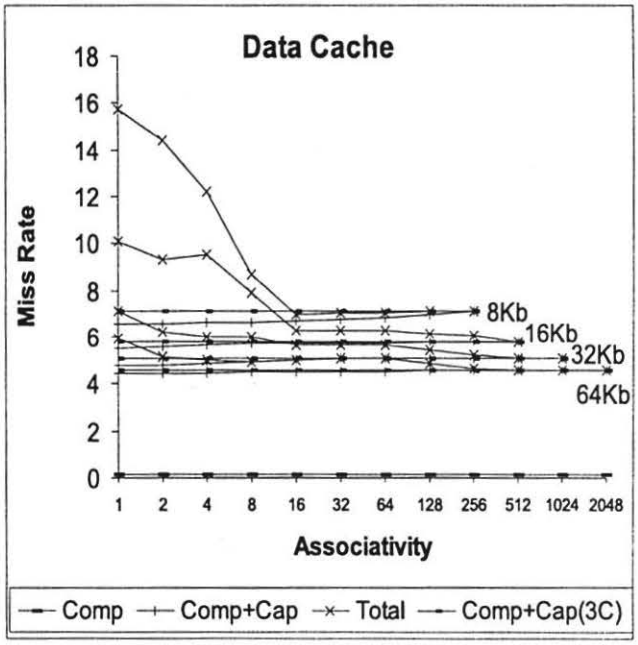

Fig. 5. SPECfp95 mean miss rates for data caches.

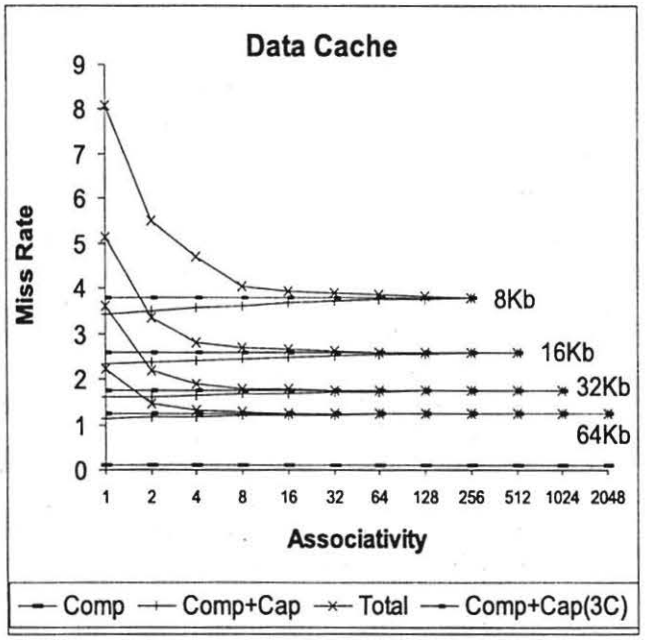

Fig. 4. SPECint95 mean miss rates for data caches.

in sustained form as the associativity increases. Contrary, capacity miss rates for the D3C model have minimum values for direct mapped caches and they increase with greater degrees of associativity equaling those of fully associative caches. This phenomenon is due to the anticonflict references and could be explained in the following way: the references to a given block that are very spaced in time are not retained by fully associative caches because of the use of the LRU policy. A small degree of associativity could introduce asymmetries in the mapping and therefore a number of these references may benefit by having less competitors for a particular set. The direct mapped cache presents the maximum possible number of sets for a given cache size and therefore it is for this organization that the greater number of long term references that remain in the cache can be expected.

Figures 2, 3, 4 and 5 show that the underestimation of the 3C model for the conflict misses can turn out to be important depending on the cache size and associativity, specially for instruction caches. This underestimation comes from the anticonflict references, as was previously mentioned. Taking as reference the miss rate of the fully associative cache, for example for the 8Kbytes direct mapped instruction cache the anticonflict references for the average of the SPECfp95 are 21.1 percent and 8.3 percent for a data cache of the same type. For a 32Kbytes cache the percentages are 59.0 and 5.5 percents respectively. For the average of the SPECint95, a 16Kbytes instruction cache has a 19.7 percent of anticon-

flict references, and for a data cache a 9.9 percent. Also the analysis of the curves reveals the existence of some configurations for which the 3C model has negative conflict miss rates (the points where the "Total" curve is under the "Comp + Cap(3C)" or fully associative curve). For these points the negative conflict miss rate is low, but for some of the benchmarks individually these values were very high. Another interesting observation is that the largest amount of anticonflict references is found for direct mapped caches, for which the 3C model gives the smallest possibility of negative conflict misses.

VI. Conclusions

A new form for the classification of cache memory misses has been proposed here. The new D3C model is a refinement of the extensively used 3C classification that introduces new analysis possibilities. Results of simple and intuitive interpretation for paradigmatic patterns of memory references have been obtained. The proposed model also allows the individual classification of the memory references and its evolution within the classification under changes in the cache configuration. The application of the 3C and D3C models to the benchmarks commonly used for the evaluation of memory hierarchies shows the degree of conflict misses underestimation that the 3C model makes as compared to the D3C model. This numerical difference turns out to be of importance for the configurations more commonly used and are due to the anticonflict references, which therefore have an important participation in the total cache statistics.

Acknowledgments

We gratefully acknowledge Professor Daniel Etiemble from LRI, Université de Paris-Sud, and Professor Alberto Dams from the Universidad de Buenos Aires. This work was partially financed by the Universidad de Buenos Aires, grant TI-09, and the Consejo Nacional de Investigaciones Científicas y Técnicas , CONICET, grant PIP-4030.

References

[BEL 69] BELADY, L. A.; NELSON, R. A.; SHIDLER, G. S. An Anomaly in Space-Time Characteristics of Certain Programs Running in a Paging Environment. **Communications of the ACM**, p.349-353, December 1969.

[HAM 99] HAMKALO, J. L.; CERNUSCHI-FRIAS, B. Theoretical Analysis of Cache Statistics for the Simple Loop Model. **International Journal of Computers & Applications**, v.21, n.1, p.13-18, 1999.

[HIL 89] Hill, M. D.; SMITH, A. J. Evaluating Associativity in CPU Caches. **IEEE Trans. on Computer**, C-38, n.12, p.1612-1630, 1995.

[LEB 94] LEBECK, A. R.; WOOD, D. A. Cache Profiling and the SPEC Benchmarks a Case Study. **IEEE Computer**, v.27, n.10, p.15-26, 1994.

[PAT 95] PATTERSON, D. A.; HENNESSY, J. L. **Computer Architecture a Quantitative Approach**, San Mateo, California: Morgan Kaufmann Publishers, 1995.

[SMI 82] SMITH, A. J. Cache Memories. **ACM Computing Surveys**, v.14, n.3, p.473-530, 1982.

[SMI 85] SMITH, J. E.; GOODMAN, J. R. Instruction Cache Replacement Policies and Organizations. **IEEE Trans. on Computer**, C-34, n.3, p.234-241, 1985.

[SRI 94] SRIVASTAVA, A.; EUSTACE, A. ATOM: A System for Building Customized Program Analysis Tools. In: ACM CONFERENCE ON PROGRAMMING LANGUAGE DESIGN AND IMPLEMENTATION, 1994. **Proceedings...** p.196-205, 1994.

[SUG 93] SUGUMAR, R. A.; SANTOSH, G. A. Efficient Simulation of Caches under Optimal Replacement with Applications to Miss Characterization. In: ACM SIGMETRICS, 1993. **Proceedings...** p.24-35, 1993.