Cache Behavior Modelling for Codes Involving Banded Matrices*

Diego Andrade, Basilio B. Fraguela, and Ramón Doallo

Universidade da Coruña
Dept. de Electrónica e Sistemas
Facultade de Informática
Campus de Elviña, 15071 A Coruña, Spain
{dcanosa,basilio,doallo}@udc.es

Abstract. Understanding and improving the memory hierarchy behavior is one of the most important challenges in current architectures. Analytical models are a good approach for this, but they have been traditionally limited by either their restricted scope of application or their lack of accuracy. Most models can only predict the cache behavior of codes that generate regular access patterns. The Probabilistic Miss Equation (PME) model is able nevertheless to model accurately the cache behavior for codes with irregular access patterns due to data-dependent conditionals or indirections. Its main limitation is that it only considers irregular access patterns that exhibit an uniform distribution of the accesses. In this work, we extend the PME model to enable to analyze more realistic and complex irregular accesses. Namely, we consider indirections due to the compressed storage of most real banded matrices.

1 Introduction

Memory hierarchies are essential in current architectures, since they cushion the gap between memory and processor speed. Understanding and improving the usage of caches is therefore absolutely necessary for obtaining good performance both in sequential and parallel computers. There are several methods to study the cache behavior. For example, trace-driven simulations [1] provide accurate estimations of the cache behavior but the required simulations have a high computational cost. Hardware counters [2] yield also accurate estimations but the execution of the real code is needed and their use is limited to the architectures where such registers exist. Both techniques provide a summarized characterization of the cache behavior and little insight about the observed behavior is obtained. As a result, it is difficult to benefit from the information generated in order to improve the cache performance.

* This work has been supported in part by the Ministry of Science and Technology of Spain under contract TIN 2004-07797-C02-02 and Xunta de Galicia under contract PGIDIT 03TIC10502PR.
Analytical models of the source code are the best suited approach to enable compilers to extract the behavior of the memory hierarchy and guide optimizations based on this understanding. Models can obtain an accurate prediction of the cache behavior based on the analysis of the source code to execute. Their main drawback is their limited scope of application. Most of them are restricted to codes with regular access patterns. There have been few attempts to model irregular codes but they are either non-automatable or quite imprecise. The Probabilistic Miss Equation (PME) model is nevertheless able to analyze automatically codes with irregular access patterns originated by data-dependent conditionals or indirections with a reasonable accuracy. In the case of irregular codes due to indirections, some knowledge about the distribution of the values contained in the structure that produces the indirection is required in order to achieve certain precision in the predictions. Until now the PME model could only model with a reasonable accuracy indirect accesses that follow an uniform distribution, that is, access patterns in which every position of the dimension affected by the indirection has the same probability of being accessed. This model extension was fully automated and integrated in a compiler in [9]. In the present work the PME model is extended to be able to model automatically and precisely an important class of non-uniform irregular access patterns. Namely, we consider the indirections generated by the compressed storage of realistic banded matrices, a very common distribution in sparse matrices [10]. Most banded matrices are composed of a series of diagonals with different densities of nonzeros. This way, we have developed a more general model that considers this kind of distribution. The accuracy of this new extension will be evaluated using well-known matrix collections.

The rest of the paper is organized as follows. Section 2 introduces the basics of the PME model. Then, Section 3 discusses the extended scope of the model and the problems that the modeling of non-uniformly distributed irregular accesses implies. The extension of the model to cover indirections using realistic banded matrices is described in Section 4. Section 5 is devoted to the experimental results. Section 6 briefly reviews the related work. Finally, in Section 7 the conclusions of our work are established.

2 Introduction to the Probabilistic Miss Equations (PME) Model

Our model estimates the number of misses generated by a code studying the behavior of each static reference \( R \) separately. Its strategy lies in detecting the accesses of \( R \) that cannot exploit reuse in the cache, and the potential reuse distances for those that can. The reuse distance is the interval in the execution between two consecutive accesses to a same line. During the reuse distance other data structures of the program can be accessed that can interfere in the cache with the studied data structure. These reuse distances are measured in terms of iterations of the loops that enclose the reference, and they generate interference miss probabilities that depend on the cache footprint of the regions accessed.