Interleaved sectored caches: reconciling low tag volume and low miss ratio

André Seznec

INRIA

INSTITUT NATIONAL DE RECHERCHE EN INFORMATIQUE ET AUTOMATIQUE

Interleaved Sectored Caches: reconciling low tag volume and low miss ratio

André Seznec

N° 2084

Octobre 1993

PROGRAMME 1

Architectures parallèles,

bases de données,

réseaux et systèmes distribués

ISSN 0249-6399

Interleaved Sectored Caches: reconciling low tag volume and low miss ratio

André Seznec *

Programme 1 — Architectures parallèles, bases de données, réseaux et systèmes distribués
Projet Calcpar

Rapport de recherche n ° 2084 — Octobre 1993 — 28 pages

Abstract: Sectored caches have been used for many years in order to reduce the tag volume needed in a cache. In a sectored cache, a single address tag is associated with a sector consisting in several cache lines, while validity, dirty and coherency tags are associated with each of the inner cache lines.

Using a sectored cache is a design trade-off between a low volume for cache tags allowed by a large line size and a low memory traffic induced by using a small line size. This technique has been used in many cache designs including small on-chip microprocessor caches and large external second level caches. Unfortunately on a sectored cache, a large part of the cache space may be wasted by invalid cache lines, then the miss ratio on a sectored cache is often significantly higher than the miss ratio on a non-sectored cache.

Usually in a cache, a cache line location is statically linked to one and only one address tag word location. In the interleaved sectored cache we propose in this paper, this monolithic association is broken; the address tag location associated with a cache line location is dynamically chosen at fetch time among several possible locations.

Trace driven simulations show that the hit ratio on an interleaved sectored cache is very close to the hit ratio on a non-sectored cache while the tag volume is very low as on traditional sectored caches.

Key-words: caches, sectored caches, cache tag volume, interleaved sectored caches

(Résumé : tsvp)

This work was partially supported by CNRS (PRC-ANM)

*seznec@irisa.fr

Unité de recherche INRIA Rennes
IRISA, Campus universitaire de Beaulieu, 35042 RENNES Cedex (France)
Téléphone : (33) 99 84 71 00 – Télécopie : (33) 99 38 38 32

Antémémoire à secteurs entrelacés : ou comment concilier un volume réduit d'étiquettes et un faible taux d'échec

Résumé : Les antémémoires à secteurs ont été utilisées depuis plus de 20 ans pour réduire le volume des étiquettes. Ce type d'antémémoire offre un compromis de conception entre le volume réduit d'étiquettes permis par l'utilisation de lignes longues et le trafic mémoire réduit permis par l'utilisation de lignes courtes au pris d'une dégradation du comportement global.

Les antémémoires à secteurs entrelacés que nous proposons dans ce rapport réconcilient volume réduit d'étiquettes et faible taux d'échec.

Mots-clé : Antémémoires, secteurs, volume d'étiquettes, antémémoires à secteurs entrelacés

1 Introduction

With the increasing of integration density, the gap between main memory access time and instruction issue frequency is always deepening: up to 400 millions instructions per second on a DEC 21064 [4], while the main memory access time remains around 250 ns. Then the performance of a microprocessor system depends highly on its memory hierarchy behavior and particularly on caches behavior, primary caches as well as second-level caches.

As the impact of cache design on the performance of processor is particularly important, there has been a lot of studies on these designs including studies on influence of associativity on the cache miss ratio [7, 13], impact of cache size [5, 1] or impact of the line size [17] on performance.

Caches contain two kinds of informations, data and tags. Address tags are needed for determining whether or not the accessed data is desired one or not, validity, dirty and coherency tags are needed for maintaining coherency of data among the distinct levels in the memory hierarchy or among the caches associated with distinct processors in a multiprocessor.

The hardware cost for storing these tags may be quite high when the cache line size is in the range of 16 to 64 bytes. Surprisingly, for the ten last years, there has been very few studies on limiting the tag cost and volume for caches [14, 15]. In order to reduce this tag storage cost, sectors have been used in many cache designs for more than 20 years. A sector consists in several contiguous cache lines associated with a single address tag. Using sectored caches instead of a classical cache structure significantly decreases the tag volume. Unfortunately, when using a sectored cache, a large part of the cache space may be wasted by invalid cache blocks, and this may dramatically increase the miss ratios.

The aim of the interleaved sectored cache we propose in this paper is to allow both low tag volume and high hit ratio. Usually in a cache, a cache line location is statically linked to one and only one address tag location. In the interleaved sectored cache, this monolithic association is broken; the address tag location associated with a cache line location is chosen among several.

The remainder of the paper is organized as follows. In section 2, we show why maintaining a low tag volume is an important issue in many cache designs and we comment on the use of sectored caches. In section 3, we introduce the interleaved sectored cache organization and we consider different possible implementations depending on whether the cache is direct-mapped or set-associative. In section 4, trace driven simulation results are given; using interleaved sectored caches for L1 caches as well as for L2 caches leads to hit ratios very close to the hit ratios obtained with non-sectored caches while the tag volume is maintained very low. Section 5 concludes this study.