AN ENERGY SAVING SET ASSOCIATIVE CACHE ALGORITHM WITH IMPROVED PERFORMANCE

S.SUBHA
SITE, Vellore Institute of Technology
Vellore, India
E-mail: ssubha@rocketmail.com

ABSTRACT

Enabling one way in set associative cache during operation is proposed in literature. However, this architecture degrades the average memory access time. This paper proposes an algorithm to map to certain way in set associative cache improving the performance. The address is mapped to certain way by certain transformations involving XOR'ing and shifting and bit selection. The line is accessed or placed/replaced in the mapped way. One way is enabled during this operation. A mathematical model is developed for the proposed model. Conditions for energy saving and performance improvement are derived. Simulations with SPEC2K benchmarks are performed. Energy improvement of 38% with performance improvement of 20% is observed for the chosen parameters.

Keywords: *Average Memory Access Time, Cache Algorithm, Cache Performance, Energy Saving, Set Associative Cache*

1. INTRODUCTION

Caches are denoted by (C,k,L) where C is the capacity with associativity k and linesize L [1,2]. Caches are of three kinds-direct mapped, set associaitve and fully associaitve. A line is placed in fixed location in direct mapped cache while it can occupy any of n blocks in fully associative cache of n blocks. It can occupy any of w ways of mapped set in w-way set associative cache. Placing subset of mapped cache ways in low energy mode with performance degradation is proposed in [3]. Partial tag comparison for cache access is proposed in [4,6]. A method to access cache ways selectively is proposed in [5]. Energy savings is obtained. Algorithms to save energy in set associative cache operation is proposed in [7,8]. Algorithm to enable one cache way with average memory access improvement is proposed in [9]. The results are based on simulations run on earlier version of GNU C compiler.

This paper proposes algorithm to place a line in fixed cache way resulting in improved average memory access time. The tag portion of the address is XOR'd with sixteen times the associativity of the cache. The result is shifted right four times. The mapped way is obtained by bit selection of resultant tag value with cache associativity. An address is mapped to one way in this algorithm. Simulations with SPEC2K benchmarks are performed run with GNU compiler version 4.4.7. An improvement of 20% in average memory access time with energy saving of 38% is observed.

The rest of the paper is organized as follows. Section 2 gives motivation, section 3 the proposed model section 4 simulations section 5 conclusion followed by references.

2. MOTIVATION

Consider the SPEC2K benchmarks. Consider two level inclusive cache hierarchy. Let level one cache be 8-way set associative with 2048 sets and level two be 16-way set associative cache with 4096 sets. The benchmarks are run against Simplescalar toolkit. The data addresses are collected with line size of 32B using load-load-op-store policy. The collected addresses are run against the chosen cache. The hits and miss are shown in Table 1. Next consider the following algorithm for any address a in level one. Let L1SIZE = 2048, NUM_WAYS=8

```
1. calculate set1=address % L1SIZE tag1=addresss / L1SIZE, temp = NUM_WAYS
2. Let temp<<= 4
3. tag1 = tag1 XOR temp
4. tag1 >>= 4
5. wayno = tag1 % NUM_WAYS
6. stop.
```

Table 1 Hit And Miss Values In Traditional Cache System

name	hits1	miss1	hits2	miss2	total
256.bzip2	372116	95207	91111	4095	467323
181.mcf	3544	420	0	419	3964
197.parser	28930	3513	57	3455	32443
300.twolf	4076	437	0	436	4513
255.vortex	11497	549	54	494	12046
175.vpr	8207	549	0	548	8756

The line is mapped to wayno calculated above in the mapped set. If the line is found, increment the hits, access the line and stop. If it is absent, the number of misses is incremented and the line is placed in wayno of mapped set respecting the inclusive property of the cache system. The above algorithm has time complexity of two shift operations and two bit extractions and one XOR operation per address. The above algorithm is simulated with SPEC2K benchmarks. The hits and misses are shown in Table 2. As seen from Table 1 and Table 2 on average there is increase in number of level one hits.

Table 2 Hits and miss values in proposed system

name	hits1	miss1	hits2	miss2	total
256.bzip2	275908	191415	187319	4095	467323
181.mcf	3961	3	0	2	3964
197.parser	32425	18	8	9	32443
300.twolf	4458	55	43	11	4513
255.vortex	11033	1013	900	112	12046
175.vpr	8753	3	0	2	8756

Only one way is enabled in the proposed algorithm in contrast to whole set enabled in traditional cache. Assume the cache operates in two modes - high power mode and low power mode. It is in low power mode during non-operation. The accessed ways are in high power mode during operation. Assuming the cache consumes 5J per way during non-operation and 10J per way during operation, the energy calculations are shown in Table 3. As seen from Table 3 there is 38% improvement in energy consumption. This improvement in energy saving with improved hits is the motivation of this paper.

Table 3 Energy Comparison

name	Energy(p)	Energy(t)	%improve
256.bzip2	18059415	26678120	32.30627
181.mcf	429660	560800	23.38445
197.parser	573255	1947400	70.56306
300.twolf	436565	584120	25.26108
255.vortex	550870	894400	38.40899
175.vpr	453620	762800	40.53225
			38.40935

3. ARCHITECTURE OF THE PROPOSED MODEL

Consider cache system. The energy consumed in the system depends on the number of active cache components. In set associative cache, the energy consumption depends on the number of active cache ways. The desired function of cache sub-system can be formulated as

min cache energy consumption
subject to
min average access time. (1)

This paper proposes an algorithm to this extent. The tag portion of the address is XOR'd with sixteen times the associativity of the cache. The result is shifted right four times. The mapped way is obtained by bit selection of resultant tag value with cache associativity. If the line is present in the mapped way, the number of hits is incremented, the line is accessed. In case of miss, the number of misses is incremented, the line is fetched respecting the inclusive nature of the cache.

Consider two level inclusive cache system. Let level one be w_1-set associative cache of S_1 sets. Let level two be w_2-way set associative cache of S_2 sets. Let the line size be L bytes. Consider the algorithm described above. Denote the system implementing this algorithm as C_{prop}. The architecture of the proposed system is shown in Fig.2. Consider the traditional cache system with same configuration. Denote it as C_{trad}.In traditional cache let $H_1, H_2, T_1, T_2, T_{12}, M$ be level one hits, level two hits, level one access time, level two access time, transfer time between level one and level two, miss penalty for address trace of R references respectively. The average memory access time in the traditional cache is given by

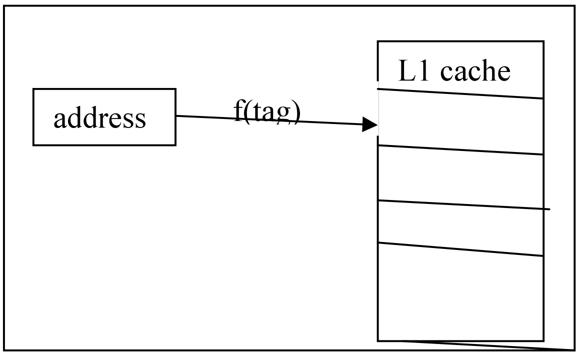

Figure 2 Architecture Of Proposed System. The Tag Portion Of The Address Is Subjected To The Transformations Described In This Section And Mapped To One Cache Way Of The Mapped Set.

$$AMAT\ (C_{trad}) = \frac{1}{R}\begin{pmatrix} H_1T_1 + \\ H_2(T_1 + T_2 + T_{12}) + \\ (R - H_1 - H_2)M \end{pmatrix} \quad (2)$$

Let $h_1, h_2, t_1, t_2, t_{12}, t_f, m$ be level one hits, level two hits, level one access time, level two access time, transfer time between level one and level two, time to perform the address mapping to certain way, miss penalty respectively in the proposed system. The AMAT of proposed system is

$$AMAT\ (C_{prop}) = \frac{1}{R}\begin{pmatrix} Rt_f + h_1t_1 + \\ h_2(t_1 + t_2 + t_{12}) + \\ (R - h_1 - h_2)m \end{pmatrix} \quad (3)$$

The first term in (3) is the time to perform the address mapping to specific way. The second term is the time taken during level one hits. The third term is the time taken for level two hits. The fourth term is the time taken for misses. An improvement in AMAT is observed if

$$\frac{1}{R}(H_1T_1 + H_2(T_1 + T_2 + T_{12}) + (R - H_1 - H_2)M)$$
$$>=$$
$$\frac{1}{R}\begin{pmatrix} Rt_f + h_1t_1 + h_2(t_1 + t_2 + t_{12}) \\ + (R - h_1 - h_2)m \end{pmatrix} \quad (4)$$

Consider the energy consumption. Assume the cache system operates in two energy modes-high power mode and low power mode. During non-operation the cache is in low energy mode. During operation, in the traditional set associative cache the mapped set is in high energy mode. In the proposed model, one way (mapped way) in the mapped set is in high energy mode. Let E_{low}, E_{high} be the energy consumed per way in low energy mode and high energy mode respectively in the cache system. The total energy consumed in the traditional cache is given by

$$\begin{aligned} &E(C_{trad}) = R(w_1S_1 + w_2S_2)E_{low} + \\ &H_1w_1(E_{high} - E_{low}) + \\ &H_2(w_1 + w_2)(E_{high} - E_{low}) + \\ &(R - H_1 - H_2)(w_1 + w_2)(E_{high} - E_{low}) \end{aligned} \quad (5)$$

The first term in (5) is the energy consumed in the traditional cache system during non-operation. All the sets in both cache levels are in low energy mode. The second term is the energy consumed during level one hits. The mapped level one set is in high energy mode. The third term is the energy consumed during level one miss and level two hits. The mapped sets in both cache levels are in high energy mode. The fourth term is the energy consumed during miss. The mapped sets in both cache levels are in high energy mode. The energy consumed in the proposed model is given by

$$\begin{aligned} &E(C_{prop}) = R(w_1S_1 + w_2S_2)E_{low} + \\ &H_1(E_{high} - E_{low}) + \\ &H_2(1 + w_2)(E_{high} - E_{low}) + \\ &(R - H_1 - H_2)(1 + w_2)(E_{high} - E_{low}) \end{aligned} \quad (6)$$

The first term in (6) is the energy consumed by the proposed cache system during non-operation. The second term is the energy consumed during level one hits. One cache way is enabled during this operation. The third term is the energy consumed during level one miss and level two hits. One cache way in level one and one set in level two are in high energy mode in this case. The fourth term is the energy consumed during level one and level two miss. One cache way in level one and one set in level two are in operational mode in this case. An improvement in energy consumption is observed if

$$\begin{aligned} &R(w_1S_1 + w_2S_2)E_{low} + H_1(E_{high} - E_{low}) \\ &+ H_2(1 + w_2)(E_{high} - E_{low}) + \quad <= \\ &(R - H_1 - H_2)(1 + w_2)(E_{high} - E_{low}) \\ &R(w_1S_1 + w_2S_2)E_{low} + H_1w_1(E_{high} - E_{low}) \\ &+ H_2(w_1 + w_2)(E_{high} - E_{low}) + \\ &(R - H_1 - H_2)(w_1 + w_2)(E_{high} - E_{low}) \end{aligned}$$

$$(7)$$

4. SIMULATION

The proposed model is validated with traditional model. The cache parameters are as mentioned in section 2 of this paper. The number of hits and misses are shown in Table 1 and Table 2 in section 2 of the paper. Assuming six cycles for level one access, eighteen cycles for level two access, eighteen cycles for level one to level two access, sixty five cycles for miss penalty the AMAT of traditional cache is shown in Table 4. For the proposed system assuming the address translation takes five cycles, it takes one cycle to access the mapped way. Assuming eighteen cycles for level two access and level one to level two transfer time with sixty five cycles for miss penalty the AMAT is shown in Table 4. The AMAT value comparison is shown in Fig.2. As seen from Fig. 2 there is improvement in AMAT of 20%.

Table 4 AMAT comparison

name	AMAT(p)	AMAT(t)	%improve
256.bzip2	20.99098	13.53582	-55.07729
181.mcf	6.048436	12.25126	50.6301
197.parser	6.028604	12.34824	51.17844
300.twolf	6.513184	11.71305	44.39379
255.vortex	9.290055	8.585838	-8.202079
175.vpr	6.021928	9.699292	37.91374
		average	20.13945

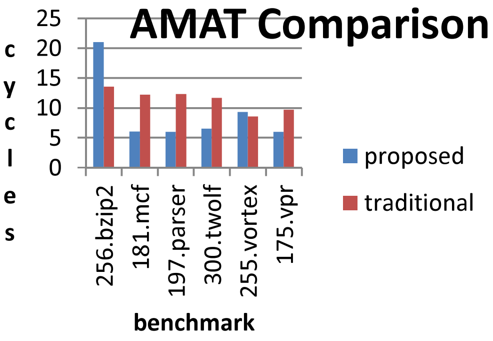

Figure 2 Amat Comparison

As discussed in section 2 ot this paper there is improvement in energy consumption of the proposed system as only one way is in high energy mode during operation in level one cache as opposed to all cache ways enabled in the mapped set in the traditional cache.

5. CONCLUSION

A set associative cache with energy saving and improved performance is proposed in this paper. The tag portion of the address is XOR'd with sixteen times the associativity of the cache. The result is shifted right four times. The mapped way is obtained by bit selection of resultant tag value with cache associativity.A mathematical model is derived for the energy consumption and performance of the proposed model. Simulations with SPEC2K benchmarks was performed. An energy saving of 38% with performance improvement of 20% over traditional set associative cache is observed.

ACKNOWLEDGMENTS

The author expresses thanks to Santa Clara University, Santa Clara, CA, USA for providing Simplescalar Toolkit and SPEC2K benchmarks.

REFERENCES

[1] Alan Jay Smith, Cache Memories, Computing Surveys, Vol.14, No.3, September 1982, pages 473-530

[2] David .A. Patterson, John L Hennessey: Computer System Architecture : A Quantitative Approach, 3rd edition, Morgan Kaufmann Publishers Inc., 2003

[3] David.H. Albonesi: Selective Cache Ways: On Demand Cache Resource Allocation, Proceedings of International Symposium on Microarchitecture, 1999, pages 248-259

[4] Jian Chen, Ruihua Peng, Yuzhuo Fu, Low Power Set Associative Cache with Single-Cycle Partial Tag Comparison, Proceedings of ASICON 2005, pp. 73-77

[5] Michael D. Powell, Amit Agarwal, T. N. Vijaykumar, Babak Falsafi, and Kaushik Roy, Reducing Set-Associative Cache Energy via Way-prediction and Selective Direct-Mapping, Proceedings of MICRO, 2001

[6] RuiMin, Zhiyong Xu, Yiming Hu, Wen-ben Jone, Partial tag comparison: a new technology for power-efficient set-associative cache designs, Proceedings of 17th International Conference on VLSI Design, 2004, pp. 183-188

[7] S.Subha, A Set Associative Cache Model with Energy Saving, Proceedings of ITNG, 2008, pp. 1249-1250