Determination of Optimal Sizes for a First and Second Level SRAM-DRAM On-Chip Cache Combination

Rupinder Hundal
Vojin G. Oklobdzija
Electrical and Computer Engineering Department
University of California, Davis
Davis, California 95616

Abstract

In this paper a SRAM-DRAM combination memory is used as an on-chip cache memory. The SRAM portion is considered to be the first level cache and the DRAM portion is considered to be the second level cache. Curves are derived for the optimal SRAM and DRAM sizes with memory access time optimized. These curves do vary with the amount of fixed on-chip area available, with different cell size ratios of DRAM to SRAM and with varying DRAM access times. The optimal amount of SRAM, or first level cache, is consistently seen to be much smaller then the general trend of a large first level cache. For instance, if the on-chip area available for cache could contain 256K of SRAM then only about 16K of SRAM should exist along with 2.5M of DRAM as second level cache for optimal usage of that area.

1 Introduction

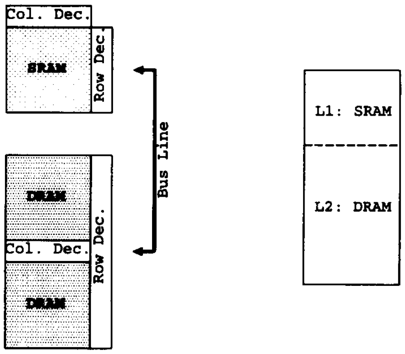
In the past years processor speeds have increased much more than memory access speeds. Even in low-end systems DRAM access speeds need to be increased in order to keep pace with processor speeds. In the last few years Cache DRAMs have been introduced in order to decrease the DRAM access times [1, 2, 3, 4, 11]. A portion of the DRAM chip contains cache memory in the form of SRAM. Approximately 90% of the time when an access is made to the CDRAM, only the SRAM is accessed [3] which results in a fast access time. Ramtron has developed an enhanced DRAM [5], EDRAM, which also contains some SRAM on a DRAM memory chip.

With a beneficial combination of fast access time due to the SRAM and high density due to the DRAM, the CDRAMs designed until now have been aimed at replacing existing DRAMs as either main memory [5] or secondary caches [3, 11]. Their application as microprocessor on-chip memory also seems feasible with SRAM portion being the first level cache and the DRAM portion being the second level cache [11, 12].

In this multilevel cache system the block size of the first level cache can be increased for a higher hit rate without the penalty of a higher transfer time. Due to the proximity of the first and second level cache the data transfer bus can be made very wide ranging from 32 bits [1, 2, 3, 11] to 2048 bits [5] in the existing SRAM-DRAM combination memories.

Using DRAM as an on-chip memory does have some drawbacks. Along with a slower access time, it has a higher susceptibility to soft errors and there is the requirement for dynamic refresh. The LISP processor [7] successfully employs DRAM as an on-chip memory by using the 4T dynamic cell and a refreshing scheme. Lee and Katz [8] also describe a technique to manage DRAM as an on-chip cache without the need for refreshing.

This paper looks at the SRAM-DRAM combination memory and considers it for use as an on-chip multilevel cache(Figure 1) citer3. It attempts to determine the amount of SRAM and DRAM necessary for minimal average access time. Next section looks at the mean memory-access time per reference equation, which is used to mathematically model the memory system. Sections 3, 4 and 5 look at the optimal SRAM and DRAM combinations as they vary with area, cell size ratios, and DRAM access times. Section 6 draws some conclusions.



(a) (b)

Figure 1: *SRAM-DRAM combination memory as on-chip multilevel cache*

2 Mean Memory-Access Time per Reference

Since processor speeds have increased much more than memory access speeds, either a single or multi-level cache is required to reduce the mean time to access information from memory.

The mean memory-access time per reference(MAR) for a two-level cache system can be determined using

$$MAR = h_1T_1 + (1 - h_1)h_2T_2 + (1 - h_1)(1 - h_2)M \quad (1)$$

where h_1 and h_2 are hit rates of cache levels 1 and 2 respectively, T_1 and T_2 are access times of cache levels 1 and 2 respectively, and M is the access time of main memory.

Hit rate h_1 is dependent on the size of cache level 1, $h_1 = f(c_1)$, and h_2 is dependent on both the sizes of cache levels 1 and 2, $h_2 = f(c_1, c_2)$. Since hit rate h is $h = 1 - m$, where m is the miss rate, then the hit rate functions can be evaluated from miss rates. There are three kinds of miss rates: global, local and solo [9]. In a two level cache system the *global miss rate*, m_{12}, is measured as

$$m_{12} = \frac{num\ of\ L2\ misses}{num\ Of\ CPU\ references}$$

and the *local miss rate*, m_k, is measured as

$$m_k = \frac{num\ of\ Lk\ misses}{num\ of\ requests\ reaching\ Lk\ cache}$$

1063-6404/94 $4.00 © 1994 IEEE

and the *solo miss rate* is measured by removing the other caches and looking at the misses of the Lk level only and dividing it by the number of CPU references. Therefore,

$$m_{12} = \frac{num\ of\ L2\ misses}{requests\ reaching\ L2} \times \frac{num\ of\ L1\ misses}{num\ of\ CPU\ references}$$

and

$$m_{12} = m_2 \times m_1$$

The global miss rate is also approximately equal to the miss rate of a single cache system with a cache size equal to the sum of c_1 and c_2, i.e. $m_{12} = m_{1+2}$, for systems where the second level cache is much larger than the first level cache [9, 10], therefore,

$$m_{12} = m_{1+2} = m_2 m_1$$

and

$$m_2 = m_{12}/m_1 = m_{1+2}/m_1$$

Hence, the hit rate of level 2 is

$$h_2 = 1 - m_2 = 1 - (1 - h_{12})/(1 - h_1)$$

$$h_2 = (h_{12} - h_1)/(1 - h_1)$$

where h_{12} is the global hit rate.

If the hit rate of level 1 is [11]

$$h_1 = 1 - 10^{-1/6(4+\log_2 c_1)} \quad (2)$$

where c_1 is in KBytes, then

$$h_{12} = 1 - 10^{-1/6(4+\log_2(c_1+c_2))}$$

and

$$h_2 = 1 - 10^{1/6(\log_2 c_1 - \log_2(c_1+c_2))} \quad (3)$$

Figure 2 shows the plot of L2 cache size vs. the hit rate for varying L1 cache sizes. The graph is not valid below the line $c_1 = c_2$ since equation 3 is derived from the assumption that $L2 > L1$.

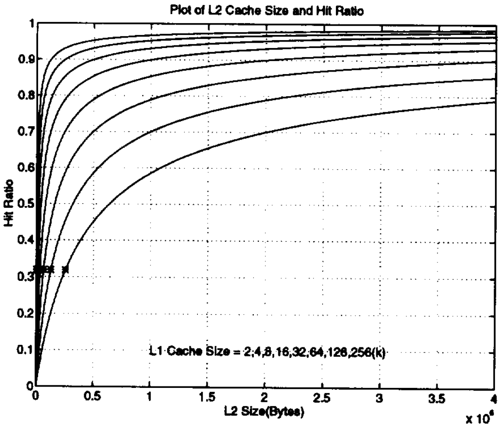


Figure 2: *Second Level Cache Size vs. Hit Ratio*

The cycle time of the first level cache, T_1, can be determined from the access time of the level 1 cache. The access time of the first level cache very often determines the cycle time. The access time of first level on-chip cache memory can be determined using [13]

$$\begin{aligned} T_{access} = & (122 \times 10^{-12}) * D^{1/\ln D} * \ln D \\ & +(20.36 \times 10^{-12}) * (BA) * (1/Ndwl) \\ & +(1.58 \times 10^{-9}) \\ & +(3.69 \times 10^{-12}) * (c_1/(BA)) * (1/Ndbl) \\ & +(632.6 \times 10^{-12}) \\ & +(60.6 \times 10^{-12}) * (A \times Ndbl) \\ & +(3.64 \times 10^{-12}) * (B \times A \times Ndbl) \quad (4) \end{aligned}$$

where

$$D = (1/2) * (c_1/BA) * (\log_2(c_1/BA)) * (Ndwl + Ntwl)$$

c_1	Level 1 cache size(bytes)
B	Block size(bytes)
A	Associativity
$Ndwl$	Number of segments per word line(data)
$Ndbl$	Number of segments per bit line(data)
$Ntwl$	Number of segments per word line(tag)

This equation assumes a 0.8μm CMOS technology based SRAM. In order to determine T_{access} as a function of cache size, $T_{access}(c_1)$, the variables are equated to the following values,

$$\begin{aligned} B &= 4\text{Bytes} = 32\text{bits} \\ A &= 1 \\ Ndwl &= Ntwl = 1 \\ Ndbl &= 8 \end{aligned}$$

Therefore,

$$\begin{aligned} T_1 &= T_{access}(c_1) \\ &= (122 \times 10^{-12}) * D^{1/\ln D} * \ln D \\ &= +(115.31 \times 10^{-15}) * c_1 \\ &= +(2.895 \times 10^{-9}) \end{aligned} \quad (5)$$

where

$$D = (360.67 \times 10^{-3}) * c_1 * \ln(c_1/4)$$

Figure 3 shows the plot of L1 cache size vs. the access time on a semilog graph.

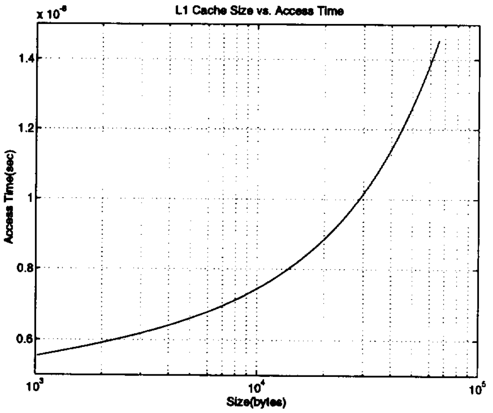


Figure 3: *L1 Cache Size vs. Access Time*

The second level cache is based on DRAM. In an attempt to determine the access time(or cycle time) of a 0.8μm CMOS technology based DRAM, some data was gathered from conference proceedings of the last few years. The results are as yet unclear since the data is based mostly on high-speed DRAM.

Using the cycle time data of the existing CDRAMs the following linear approximation can be made for the DRAM cycle times.

$$T_2 = T_{cycle}(c_2) = (9.5238 \times 10^{-15}) * c_2 + (5.0 \times 10^{-8}) \quad (6)$$

Figure 4 shows the plot of L2 cache size vs. the cycle time on a semilog graph.

Main memory cycle time consists of both transfer time and latency. Since main memory consists of DRAM there is also a recovery time derived from the difference between a DRAM access and cycle times. L2 cache also consisted of DRAM but

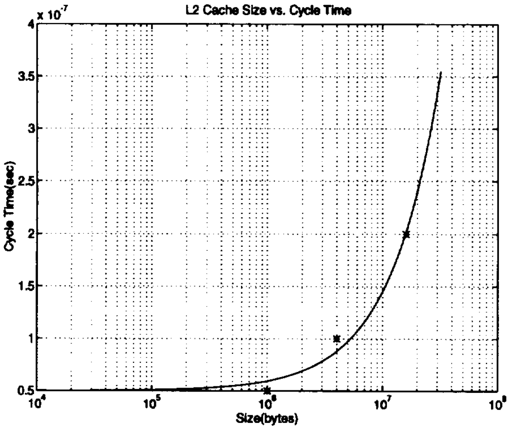


Figure 4: *L2 Cache Size vs. Cycle Time*

both the latency and recovery times were combined into the cycle time.

Using a model similar to that used by Przybylski [9] the following is obtained. The latency includes one L2 cycle time to send address to memory and 180 ns to retrieve data. The transfer time is a multiple of one L2 cycle time to transfer a set of data which is the size of the bus between L2 and main memory. The recovery time is a constant at 120 ns. Therefore,

$$\begin{aligned} \text{latency} &= T_2 + (180 \times 10^{-9}) \\ \text{transfer} &= TS * T_2 \\ \text{recovery} &= (120 \times 10^{-9}) \end{aligned}$$

This results in a main memory cycle time of

$$M = T_2 + (180 \times 10^{-9}) + (TS * T_2) + (120 \times 10^{-9})$$

$$M = T_2(1 + TS) + (300 \times 10^{-9}) \qquad (7)$$

where TS is the transfer size in terms of the number of transfers required to transfer the complete L2 block size from memory. If the bus width = 4 Bytes = 32 bits and if the block size of second level cache is 8 Bytes then the transfer size, TS, is 2. If the block size of the second level cache is 64 Bytes then the transfer size is 16.

Figure 5 shows the plot of L2 cache size vs. main memory access time M using a TS value of 2.

Using the results of equations 2, 3, 5, 6, and 7 and substituting them into equation 1 gives the mean memory-access time as a function of cache sizes c_1 and c_2. This is then used to model the memory system using Matlab.

3 Varying Amounts of Area Available

Suppose that in a processor a certain amount of fixed area is designated for cache memory. Then the question arises as to what type of memory should be used and how much. Should the entire space be filled with SRAM? or DRAM? or a specific combination of the two?

Assume that the amount of chip area allotted to cache memory is measured in terms of the amount of SRAM that it can hold. Then the possible ways to configure that area is to have all SRAM at one extreme or to decrease the SRAM amount and replace the unoccupied area with DRAM, which could be continued until the other extreme is reached where the entire area consists of DRAM. Figure 6 shows the SRAM size vs. MAR for total fixed cache areas of 128K, 256K, 512K, 1M, 2M, and 4M of SRAM where DRAM size is varying appropriately to maintain the constant area.

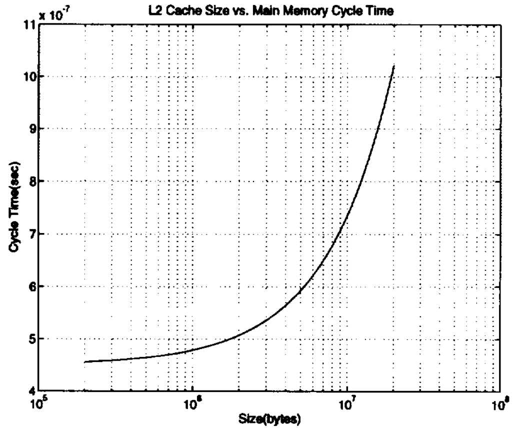


Figure 5: *L2 Cache Size vs. Main Memory Access Time*

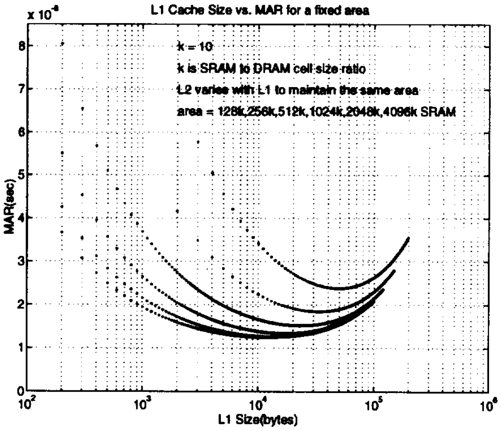


Figure 6: *First Level Cache Size vs. MAR for a fixed area*

The minimum of each of the curves gives the L1, or SRAM, cache size which is optimal since the mean memory-access time(MAR) is the smallest. The total area is also known to be between 128K and 4M for each of the curves. In this particular simulation the SRAM to DRAM cell size ratio is assumed to be 10. From this information the optimal L2, or DRAM, cache size can be determined as follows.

$$(\mathit{Total\ Area} - \mathit{L1Size}) * 10 = \mathit{L2Size}$$

Figure 7 shows SRAM sizes and DRAM sizes that are optimal for different amounts of total area available on chip. This data can also be viewed, in Figure 8, as the DRAM to SRAM size ratios that are optimal.

4 Varying Cell Size Ratios

In the previous section the simulations are based on the assumption that the SRAM to DRAM cell size ratio is 10. Actually that ratio can vary depending on the type of cells used and the feature size of the device. Mitsubishi's existing CDRAMs use a stacked capacitor cell for the DRAM and a 6T-SRAM cell. Their feature sizes have also varied resulting in SRAM to

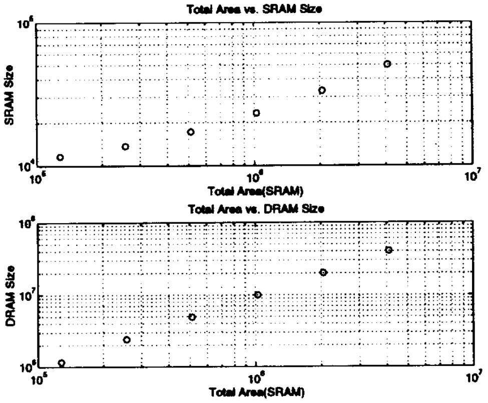


Figure 7: *Total on-chip area vs. the optimal SRAM and DRAM Sizes*

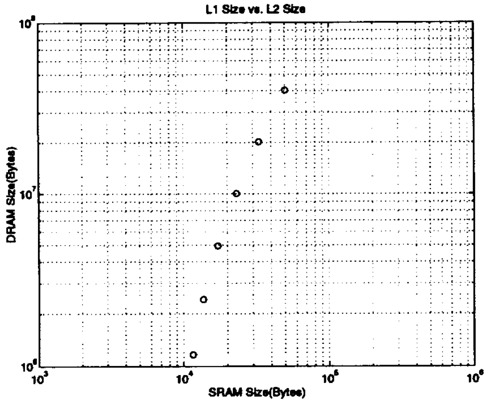


Figure 8: *Optimal SRAM and DRAM Sizes*

DRAM cell size ratios of approximately 10, 30 and 40(see table 1).

The results of rerunning the simulations of the previous sections for cell size ratios of 30 and 40 can be seen in figure 9. It shows the optimal SRAM and DRAM sizes as they vary with cell size ratio for a specific amount of area allotted to cache memory on a chip. The same SRAM vs. DRAM curve of Figure 8 is followed with the increased cell size ratios(see figure 10).

5 Varying DRAM Cycle Times

The effect of the DRAM access time on the optimal DRAM and SRAM curves is significant. The equation used in the previous simulations was a linear approximation based on the cycle times of the existing CDRAMs. If the cycle times are some multiple of that curve then the SRAM vs. DRAM curve of figure 8 shifts to the left or right as seen in Figure 11.

If the DRAM cycle time is higher then the approximated time, e.g. by a factor of 5, 10 or 15, then the curve shifts to the right and a larger size of SRAM is optimal. On the other hand, if the DRAM cycle time is lower then the estimated time then the curve shifts to the left and a smaller size of SRAM becomes optimal. (The right shifting of the curve eventually reaches a limit at DRAM cycle time equal to 15 * T_2. Any higher multiplication factor does not shift the curve further.)

	Feature Size(μm)	Cell Size(μm^2) SRAM	Cell Size(μm^2) DRAM	Size Ratio SRAM/DRAM
Mitsubishi 11/92	0.7	247.18	9.025	27.39
Mitsubishi 4/91	0.6	265.6	6.75	39.35
Mitsubishi 2/90	1.2	294	29.92	9.83

Table 1: SRAM cell size to DRAM cell size ratio.

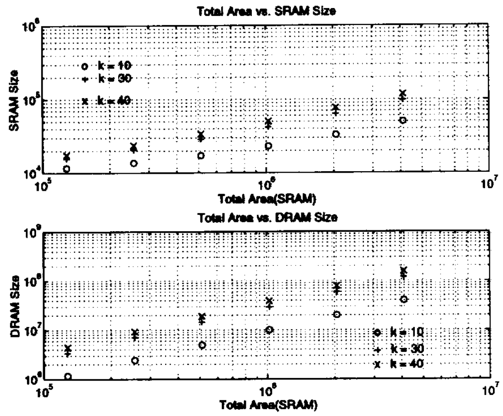


Figure 9: *Total on-chip area vs. the optimal SRAM and DRAM Sizes*

6 Conclusion

Using SRAM-DRAM combination memory as an on-chip two level cache has both the advantages of fast access and high density. But for maximum benefit of the space allocated for cache memory the amount of SRAM should be relatively small compared to the DRAM. As the previous simulations show, the SRAM size should be between 10 and 50 KBytes(see figure 7) depending on the available area where the SRAM to DRAM cell size ratio is 10. Similarly, the DRAM size should be between 1 and 40 MBytes.

The optimal SRAM size will continue to decrease in the future as sections 4 and 5 indicate. The data gathered from current experimental DRAM and SRAM devices show that the feature size of these devices is reaching about 0.3μm and the SRAM to DRAM cell size ratio may be reaching about 5. Therefore, the curves of figure 9 will be even lower. The access times of both DRAM and SRAM memories are also constantly improving therefore, the curves of Figure 11 will be shifting to the left. These trends indicate that the SRAM size should be decreasing further below 10 KBytes for available area of 128K SRAM and correspondingly for other areas.

The sizes of the existing SRAM-DRAM combination memories are shown in Table 2. Figure 12 shows a comparison between the existing memories and the optimal SRAM and DRAM sizes.

Acknowledgment

Assistance of Antonio de la Serna in preparation of this paper is gratefully acknowledged.

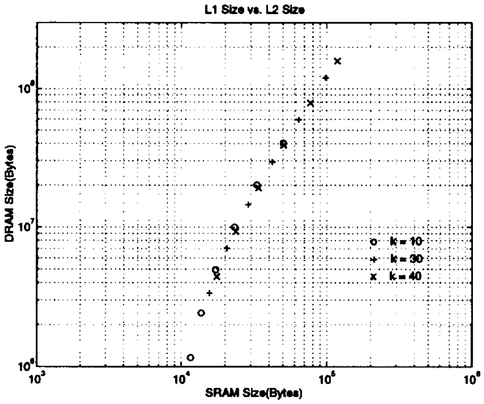


Figure 10: *Optimal SRAM and DRAM Sizes for varying cell size ratios*

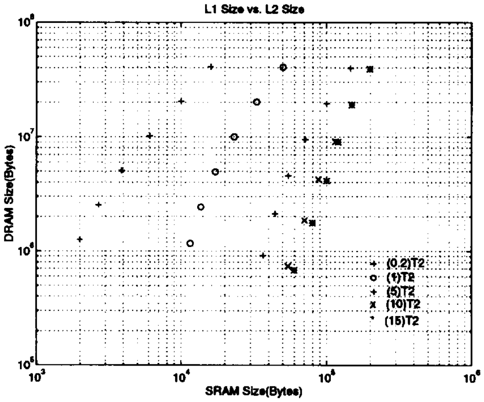


Figure 11: *Optimal SRAM and DRAM Sizes for varying DRAM cycle times*

References

[1] Asakura, M., *et al.*, "An Experimental 1Mb Cache DRAM with ECC", *IEEE Journal of Solid-State Circuits*, vol. 25, no. 1, February 1990, pp. 5-10.

[2] Hidaka, H., *et al.*, "The Cache DRAM Architecture: A DRAM with an On-Chip Cache Memory", *IEEE Macro*, April 1990, pp. 14-25.

[3] Arimoto, K., *et al.*, " A Circuit Design of Intelligent Cache DRAM with Automatic Write-Back Capability", *IEEE Journal of Solid-State Circuits*, vol. 26, no. 4, April 1991, pp. 560-565.

[4] Katsumi, D., *et al.*, "A 100-MHz 4-Mb Cache DRAM with Fast Copy-Back Scheme", *IEEE Journal of Solid-State Circuits*, vol. 27, no. 11, November 1992, pp. 1534-1539.

[5] Bursky, D., "Combination DRAM-SRAM Removes Secondary Caches", *Electronic Design*, January 23, 1992, pp. 39-43.

[6] Hart, C., *et al.*, "A new era of fast dynamic RAMs", *IEEE Spectrum*, October 1992, pp. 43-49.

	Size		Organization	
	SRAM	DRAM	SRAM	DRAM
Mitsubishi 11/92	16Kb	4Mb	4Kbx4	1Mbx4
Mitsubishi 4/91	32Kb	16Mb	4Kbx8	2Mbx8
Mitsubishi 2/90	8Kb	1Mb	2Kbx4	256Kbx4
Ramtron 1/92	2Kb	4Mb	512Kbx4	1Mbx4

Table 2: Sizes of the existing SRAM-DRAM combination memories.

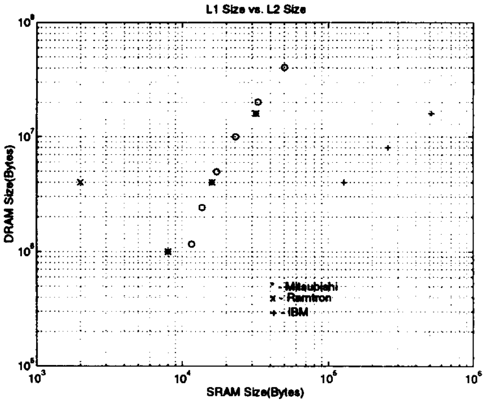


Figure 12: *Sizes of the existing SRAM-DRAM memories in comparison with the optimal SRAM-DRAM sizes*

[7] Bosshart, P., *et al.*, "553K-Transistor LISP Processor Chip", *IEEE Journal of Solid-State Circuits*, vol. SC-22, no. 5, October 1987, pp. 808-819.

[8] Lee, D., Katz, R., "Non-refreshing Dynamic RAM for On-Chip Cache Memories", *Symposium on VLSI Circuits, Digest of Technical Papers*, 1990, pp. 111-112.

[9] Przybylski, S., "Cache and Memory Hierarchy Design", *1990 Morgan Kaufmann Publishers, Inc.* ISBN 1-55860-136-8.

[10] Mekhiel, N., and McCrackin, D., "Performance Analysis For a Two Level Cache System", *26th Asilomar Conference on Circuits, Systems and Computers*, October 27, 1992.

[11] Subramanian, R., Oklobdzija, V., "Optimizing the design of multi-level caches", University of California Berkeley, Computer Science Report, Spring 1990.

[12] Iizuka, T., "Embedded Memory: A Key to High Performance System VLSIs", *Symposium on VLSI Circuits, Digest of Technical Papers*, 1990, pp. 1-4.

[13] Wada, T., Rajan, S., Przybylski, S., "An Analytical Access Time Model for On-Chip Cache Memories", *IEEE Journal of Solid-State Circuits*, vol. 27, no. 8, August 1992, pp. 1147-1156.