A REVIEW OF TECHNIQUES FOR OPTIMIZING CACHE ENERGY EFFICIENCY

HINA JILANI1, YASIR DANI AL KHAN2
School of Systems and Technology, University of Management And Technology, Lahore, Pakistan
Email:1hinakhan5454@gmail.com
yaser.khan@umt.edu.pk

ABSTRACT: In today’s computer there are larger sizes of the cache are using on the chip. Moreover, there is significant change has happen in the technology. Because of that change cache power efficiency has become the important issue in the processors. To solve this problem most of the researchers have proposed different methodology to improve the cache energy efficiency. This paper review different techniques that are proposed for improving cache energy efficiency. There are different techniques discussed in this survey paper that are used to improve the cache energy at different level of the cache. The main focus of the paper of survey is to urge the different researcher to propose different methodology so they can make the cache more efficient and energy saving.

Keywords: Cache energy; Efficiency; Dynamic cache; Static cache

1. Introduction. We are living in the era, in which technology is rapidly increasing day by day. Because of that the importance of energy efficient design in IT has become the important key factor. Recently, the main objective in designing the chip design is moving from achieving the highest performance and as well as highest energy efficiency. To achieve the energy efficiency, we have to design all of the processors, different portable devices, personal computers and etc energy efficient.

To overcome the problem of energy efficiency in cache different researchers have proposed different methodology at architecture design. The energy efficient techniques have been proposed at different processors, caches and memory. To managing the energy at cache level is crucial because of their new technology processors. Today different devices include multi core processors which have cache and large size of memory for better performance which needed the high power of consumption. There are different schemes have proposed at different level of cache [1]. In the coming years, the size of the chip is increasing that will change devices performance and decrease power consumption. The main thing that is different between the memory and processors are cache because processors use cache to increase their size and also their performance. For example, today’s personal computer 8 MB last level cache and the server side it’s up to 32 MB last level cache (LLC) [2]. However, cache is the important factor that consumes the most of the energy of the processors. To address these problems of energy efficiencies in the cache different researchers have proposed different techniques at cache level. The main thing of this paper is to address these techniques.

In this paper different techniques have been discussed that have been proposed for energy efficiency at cache level. There are different models and principles that are use for saving energy at cache level. There are different techniques at dynamic and static level of cache that are use for energy efficiency. There are some leakage energy techniques that are helpful in this regard. There are several techniques that are use for saving energy at cache level but few of them are discussed in this survey paper. In this paper some circuits level and architecture level techniques have discussed that are proposed by different researchers. In this the main focus is to discuss those techniques which are use for energy efficiency at cache level. In this paper different techniques are being addressed that are useful for managing the power efficiency of cache. There are multiple solutions on different level of cache like first level cache, last level cache and instruction level cache etc. All these techniques are discussed here that are proposed by different researchers. There are
two techniques that are discussed here is dynamic energy saving techniques and static energy savings techniques. The main focus of the paper is the review of the techniques that are use for saving cache energy. This survey paper is characterized in the following manner. The section I discussed the introduction of the problem and in Section II the background of the problem is presented in brief that which modeling and what principles are use for optimizing the cache energy efficiency. In the section III discussed the different techniques in detail that are used by dynamic and static energy techniques that how these techniques are helpful in minimizing the energy in the cache. In that section showing that which techniques is used more in that manner. In the rest of the paper the comparison of the both techniques have been discussed and also showing that which technique are more helpful in this regard. In the last the conclusion of the paper are present.

2. Related Work And Background: As discussed in the introduction that the energy efficiency of caches can be solved at different level like at last level, first level and instruction data cache etc. The two main approaches that are discussed is static and dynamic techniques that are use to optimize the energy efficiency in the cache. In dynamic approach the cache energy is adjust through frequency and voltage and can also be achieve by adjusting the activity like decreasing the access of cache and also decreasing the number of bits access through cache. On the other hand, in static techniques the main thing is to design the circuits like they use less power cells and also decrease some of the transistors that are use so that the power is adjustable. Hence the survey structured the following level and both the techniques under the study of different papers and the optimization techniques they are used is discussed here.

Modern computer have different level of cache. In the first level data and the instruction are spate and in the last level of the cache these both are unified. These levels have some different methodologies that are use for cache energy efficiency. The more detail of these levels are presented in this section.

The first level cache has the strong impact on the overall performance of the cache [3]. The last level caches are designed according to handle the cache misses and the level one of the cache are designed to control the latency. The size of the last cache level are getting increase more in the multi core system that’s why it will impact more effect the performance of the processors this level of the cache multiple program at the same time that’s why it will consume more power of cache [4]. There are some techniques that are proposed by the researcher for the both FLC (first level cache) and LLC (last level cache). The LLC are much larger than the FLC. The FLC are use in the form of dynamic energy and the LLC are use in the form of the static energy consumption [5]. In the next section these both techniques are discussed in more detail.

In this survey paper there are different techniques are presented here that propped by different researcher to optimize the cache energy efficiency at the both level of the like FLC and the LLC. There are also some techniques that are helpful in this regard but these two techniques cover the more than any other techniques. There are some hardware techniques that can also reduce the energy efficiency but those techniques have more cost and also effect the performance of the processors so that’s why those techniques are not discussed here. Moreover, there are some software level techniques but these are also complex to implement and they are not directly impact on the cache energy efficiency. So because of that reason the two main techniques are discussed here is dynamic and static energy efficient cache techniques.

Dynamic power (dynamic) is when the transistors change their voltage to a specific mode, while static power or leakage power is when the flow of the leakage happen when the power off, mathematically, they are given by,

\[
\text{Dynamic} = \alpha \times C_{\text{eff}} \times V^2 \times D \times F \quad (32)
\]

\[
\text{Static} = V_{\text{DD}} \times I_{\text{leak}} \times N \times k_{\text{design}} \quad (33)
\]

3. Dynamic Energy Sabing Techniques: There are different methodologies that are proposed by different researcher for saving cache energy at this level. There are different techniques available in terms of their structure and dimensions. Some tag base [4] [6] [7] [8] [9] [10] techniques are used to improve the dynamic energy optimization for saving the energy consumption in the cache. Some techniques use for saving the energy by reducing the cache access by associating some extra memory at different level of cache.

In [4] an approach is proposed at last level of cache because the size of that level is larger as compared to other level. In this away-based-filtering (WF) and logical last level of the cache architecture proposed in which they reduce the consumption of energy in the LLC. Furthermore, they also present a tag base approach WF use that reduces the tag base energy consumption. The other technique that use in this paper is sequential base accessing and indexing logical proposed that enhance the multiple LLC in which
multiple logic ways hits happen on the single physical medium and that control by the tag base filters. In this scheme almost 30% to 34% of the LLC energy consumption is reduced.

V-way cache technique [6] in which increase the number of tags store other than the number of data lines and using this it reduce the link between the per set bases. It reduces the mapping between the tag and data. In this [8] an architectural technique have been proposed in which that is based on the hypothesis of same distribution of the least significant bit in the cache set associatively.

In [7] a tag base approach is proposed that is eTag cache architecture in which a comparison base tag in cache that allow the access of data. It reduces the overall energy by using the tag comparison by means of using the amplifier. In this paper, a technique that is base on tag is proposed that optimize the LLC and DRAM in the equal manner. The results show that this approach reduces the cache energy by 15.4% to 33.9% in different types of the processors. Furthermore, the merged of the Tag approach reduces 32.1% to 48.7% of the cache power.

In this paper [9] a tag base approach has been approach proposed that decrease the power of the cache. The approach that is use in this paper is tag base TLB Index-Based Tagging in which index base tagging has been happened that main focus on to reduce the energy efficiency at the dynamic and static both of the level of the cache. This paper presents [10] the technique Way-Halting that uses the cache in the filters on the lines that are use for a specific set by adjusting the number of bits that are in size of four to the tag at time of indexing. In this technique it is also helpful in overcome the problem of the cache power that how you can achieve this.

The main thing in achieving the dynamic energy efficiency in cache is to decrease the amount of access time of the cache. Dynamic energy techniques are used in the different things like reducing the overall access time of the cache. Like, in [11] proposed a technique that uses cache selective ways, an architecture of the cache in which only the some of the cache subset is active when the there is not the enough cache activity. In [12] reduce the cache access time that also helpful in reducing the energy. In the [13] presents a technique for heavy amount of cache set-associative that use the bloom

<table>
<thead>
<tr>
<th>Table 1 dynamic cache energy saving techniques</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Category</strong></td>
</tr>
<tr>
<td>Other technique</td>
</tr>
</tbody>
</table>

Filters that decrease the dynamic power by eliminating those cache ways that are not have the requested data as the bloom filters. [14] Presents another bloom filter that reduces the complexity and also reduces the access ways. In this paper [15] PS-cache architecture have been proposed that minimize the way cache access even without affecting the performance of the cache. This technique uses the private-shared knowledge of the block that reduces the energy by allowing only those cache ways that have the block. This technique reduces the L1 and L2 cache almost 22% to 40%. In [18] proposed an architectural technique the fixed the size of the cache and issue of the width. They provide a model of the block in which the size of the cache and width coordinate between them. In this a feedback approach is use because of that technique the cache energy efficiency enhances. This experiment reduces almost 10% of the energy consumption and 2.3% cache miss rate achieved.

This paper [19] presents a approach that uses novel data-aware hybrid STT-RAM/SRAM technique
that store the data into two parts that rely on their bits count. These techniques reduce the amount of writes and also reduce the dynamic energy in both. This technique uses the concept of cache write policy and memory swap between the cache partitions. They show the results on UltraSPARC-III processor that utilize the hybrid and architectures in the level 2 cache that save the 42% to 53% energy. In [20] a hybrid cache energy efficiency approach have proposed an architecture that uses the NVM technique that decrease the cache power and also maintain the performance. They design a hybrid cache architecture that locates the data in SRAM and DRAM without any data migration and performance. This technique shows that there is significant amount of the data in the last level and level 2 cache. Hence they use the NVM model that minimizes the power of the cache and also maintain the performance of the cache. Their work on SPEC CPU2000, SPEC CPU2006 and Mibench benchmark have shown the results up to 46% of the power and the 24% of the performance area saving. This technique work well in the hybrid cache architecture and reduce the energy consumption.

This work [22] uses the MultiCopy Cache (MC2) design that increases the efficiency of the cache by keeping multiple copies of the cache data. When the data is ready to use for access then the different copies of the cache are manipulated that keep the errors free cache. These techniques reduce the 60% of the cache energy consumption. This work presents [23] an architecture that are based on the hybrid cache called the prediction hybrid approach has been proposed. The technique is focuses on that how to predict the block for write cache that are happen when the cache miss is occur and also tell that which cache block need to placement. This technique reduces the 26% of the cache energy consumption.

4. Static Energy Sabing Techniques: The static energy saving techniques can be achieved by the only turning off the some of the part of the cache. These techniques also known as leakage energy saving techniques both term can be used. The static energy techniques have two types one is state preserving technique and the other is state destroying technique [28] [34]. In [28] Drowsy Cache technique proposed that preserve the state that act on line granularity that put the cache into the lower energy drowsy mode that reduce the some of the operation that are in the standby mode. In [34] Cache Decay technique proposed that is a state-vanished approach it is also same like a line granularity technique but it turned-off the deadlines which are those lines that are not further accessed.

Many circuits’ level approaches also proposed by different researcher that are used to reduce the static energy consumption [27] [28] [29]. In [27] a gated-Vdd technique has been used that optimize the cache energy efficiency. In [28] drowsy cache approach is used that are used that overcome the problem of the cache energy consumption. This technique simply reduces energy in the unused cache lines that are not disturbing the data of the cache. The DRG-cache scheme [29] proposed in this paper that optimizes the cache according to their power. The complexity of this technique at hardware level is lower because of that reason their energy consumption is also smaller.

This paper [21] uses a technique that reduces the leakage power consumption. The technique that is uses Nonuniform cache architecture (NUCA) is an architecture that are based wire that are aware of the delays that are design on the banking of the sub of the cache that allow the bank that are near to the cache controller that are accessed that reduce the latency and with respect to the other bank. These techniques reduce the access data and also enhance the performance of the cache. In [17] proposed a method on the LLC because the size and the

<table>
<thead>
<tr>
<th>Table 2 static cache energy saving techniques</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Category</strong></td>
</tr>
<tr>
<td>State preserving and state destroying technique</td>
</tr>
<tr>
<td></td>
</tr>
<tr>
<td>Circuits level techniques Selective cache</td>
</tr>
<tr>
<td></td>
</tr>
<tr>
<td></td>
</tr>
<tr>
<td>Other techniques</td>
</tr>
<tr>
<td></td>
</tr>
<tr>
<td></td>
</tr>
<tr>
<td></td>
</tr>
</tbody>
</table>
level is larger than the other. They use a slice based approach cache that turned off the some parts of the last level cache. Based on the sliced base architecture some parts of cache can be off according to temporal and spatial locality of the cache. To prevent the data from the lost a data migration technique is also propose in this. Their results show almost 12% of the energy that are decrease in the cache at the last level of the cache. This work [24] proposed a cache filter architecture that is implement data on the cache. They observed that cache misses occur continuously occur when applying the filter on the data cache. These misses affect the cache power and also the overall of things that are related to performance of the cache by optimizing the latency and storing the data that are not used. The proposed filter reduce the cache misses by using three different techniques: early cache hit predictor, locality-based allocation, and no tag matching Write. Their results show that using that technique the energy reduction of the cache is reduces 21% and the ALU energy consumption up to 27.9%.

In this [25] paper use a BTI-induced architecture that are use to overcome the problem of the cache static energy consumption. They develop an analytical model that evaluates the tradeoff between the reliability and the energy consumption. They proposed dynamic voltage scaling (DVS) that check the tradeoff between the energy and reliability. Firstly, they explore that how leakage power can be reduce by implementing the reduction of the sub-threshold with aging. Their results show that it save up to 36 %.

The [30] selective cache way technique is proposed in this paper that overcome the problem of the cache power consumption. This technique disables some cache ways that are selectively disabled and because of that the energy consumption is reduced. Figure1 shows the selective ways cache that how it works.

![Figure 1(a): selective cache ways](image)

The [31] cache selective set scheme that enhance the performance as it turned off some cache set by maintain the cache associativity. The first thing that happen is that cache mapping happens when the cache size is change. The cache indexing is happen on the basis of some operation. In this technique when the cache set size is increase then the mapping happen when the cache set size is decrease then the mapping not happen in this situation.

![Figure 1(b): cache selective set scheme](image)

5. Both Static And Dynamic Energy Saving Techniques: There are some techniques that affect the both of the technique simultaneously because of that reason both dynamic and static energy consumption can be reduced. In [35] Heterogeneous Waysize cache technique proposed in which the number of the set of the cache way can be different. The main thing that can be considered is that the way of the set is the power of the only two values. Because of this method of different cache set that reduce the both cache dynamic and static energy efficiency. This paper [36] presents an approach that reduces the energy consumption at the static and dynamic level. This technique check the cache block of the sub block that tell that which block of the cache line is used and how many times it is used. This technique bring only those sub block of the cache that are necessary and off the block time when it will the time limit. There are some other techniques Drowsy Cache technique [28] Cache Decay technique [34] also uses to save the dynamic level of the cache and at the static level of the cache.
Conclusion: The recent years, there is rapid growth of the technology that uses the multi-core and multichip processor to achieve the better performance. Because of that reason of the performance there is also increase the power consumption of the multiprocessor. So, managing the performance and at the same time managing the energy also becoming the important in that manner because consumption of energy also affects the performance.

In this survey paper different proposed techniques have been discussed that are used to overcome the problem of the cache energy uses. There are two technique discussed that is dynamic and static energy consumption. There are some technique that uses the concept of the both dynamic and static to reduce the energy. This paper enhance the importance that how the cache energy efficiency technique are important to save the most of the problem.

REFERENCES


