Summary of changes

First, we would like to say thank you to the reviewers for the useful comments to improve the paper. We have addressed all the comments as explained below.

Referee 1

- *The term "irregular" in the paper used to describe the access patterns is not accurate and needs to be better defined. Even though there are access holes in the given example, the access patterns can still be well defined using simple mathematical formula.*

  The term irregular has been replaced with more descriptive terms, i.e. application with holes in their access pattern.

- *In the experimental evaluation, have you also compared the proposed methodology with the state-of-art energy optimization techniques for embedded systems? If so, what is improvement of your approach over theirs?*

  It is difficult to reproduce the results from other groups only by reading their published work. The set of benchmarks is also different. Thus, related work is presented but not directly compared with the current approach. Many of the related work is also not directly applicable on the target benchmarks as shown in the paper. This work compares the current approach with the previous published work from our groups.

- *Interleaving is only one type of data layout optimization. Have you also tried other more complex optimizations? What is the main reason that you think interleaving is the most suitable/efficient one?*

  Interleaving is an energy optimization technique for the target set of benchmark applications, which are characterized by holes in their access patterns. The applications that benefit most from the proposed methodology are characterized by having access patterns with holes. Interleaving is a widely used technique that fits the goal of generating more compact sets of data. Other optimization techniques can be complementary to the interleaving step of the methodology, if they are applied prior to the interleaving step. Otherwise, the methodology needs to be modified in order to be compatible with more complex optimizations.

  This has been clarified on Section 6.1

- *Do you think improving temporal locality of data accesses in cache is also important, if the cache is available? In embedded systems, there should also be a wide range of applications in which the same data need to be accessed multiple times? Thus if temporal locality is exploited, both stalls and number of accesses to main memory can be reduced.*

  The proposed architecture uses scratch-pad memories and no cache memory is included in the current study. The main reason is that the data can be allocated on the scratchpad memory during compilation by the designer. Software controlled allocation is a significant feature for the current methodology, as the allocation of data can be fully determined by the designer during design-time. On the other hand, the hardware control automatically copies the data into the cache and resolves any hits or misses. The basic principles of the methodology are still applicable, but some modifications would be needed to deal with cache memories. Temporal locality and data reuse are taken into consideration during the interleaving exploration.

  This has been clarified on Section 4

- *Can you add references to the following: 1) simple energy model for evaluating the energy consumption for the motivation example; 2) the commercial memory compiler for SRAM memory models; 3) the XML based language used to describe the architecture.*

1. A quick estimation for the difference in the energy consumption between the approaches presented in Fig. 1 can be calculated using a simple energy model for the memory banks. The model is based on simple, yet realistic, estimations for the dynamic and static energy on the memory banks. The scope is to illustrate the motivation for this work, without extensively describing the detailed memory model characteristics. The target architecture and the energy models are described in depth in Sec. 4. For the simple model in this chapter, we assume that the static energy is 30% of the dynamic energy, which is a rational approximation for small memory banks with sizes between 1KB and 16KB. The static energy increases linearly with the memory size, which is a good approximation based on the detailed models. Similarly the approximation for the correlation between the dynamic energy and the memory size is an approximation, which is simple but sufficiently accurate for the motivational example.

This has been clarified on Section 2

2. The commercial memory compiler is one of the state-of-the-art compilers. The logic libraries supporting a wide range of foundries and process technologies from 250nm to 28nm. The memories are optimized for low power, high performance and high density. The 40nm library was chosen for the current work. The datasheet can be found in [3].

   This has been clarified on Section 4.1

3. An XML based language is used to describe the architecture, and a cycle-accurate simulator of the processor is used to simulate the generated code on the architecture. The XML provides a structural way of describing the architecture presented in Fig. 2 including the different components, the parameters of each component and the relationship between them. The XML description generates a graphical representation of the architecture and is the input for the simulator presented in [6]. The chosen simulator is developed for coarse-grained reconfigurable architectures and is suitable in our case, because of the dynamic parameters of our architecture.

   This has been clarified on Section 4.2

- *At the end of page 3, there is such description "The interleaving of the arrays A, B, C and D is shown in Fig.1(c)"? Do you actually mean Fig.1(b)?*

  Yes. The error has been fixed.

- *In Page 9, please rephrase the redundant description, "For efficient utilization of the vector FU, the register file ..."*

  The appropriate changes have been made.

- *In the experimental part, can you comment on why you choose 5 banks? And what will be effect with different number of banks?*

  There are two main reason for exploring architectures up to five memory banks. Firstly, the energy gains achieved by increasing the number of memory banks in the memory architecture are nearly saturated even for five banks. In [4] a group of different applications were studied with regard to their energy consumption on a clustered memory architecture

consisting of up to five memory banks. The results shows that depending on the application, the energy gains started to saturate after adding a third or a fourth bank and become far smaller when adding a fifth bank. Thus, for most applications a memory architecture with five memory banks already provides more than necessary reconfiguration options. Secondly, the overhead increases exponentially with the number of memory banks, due to the increased complexity of the memory architecture. Therefore, memory architectures with six or more banks are typically not efficient options due to the high overhead and the very low energy gain.

This has been clarified in the first paragraph on Section 5.3.

- *Section 6 and Section 7 can be combined as single Section for experimental evaluation.*

  The appropriate changes have been made.

- *Typos:*
  Page 4: we assume here that he memory →we assume here that the memory
  Page 4: using four banks is presented in Fig.1(b) →using four banks is presented in Fig.1(c)
  Page 4: The data-to-memory mapping for the constructed →The optimized data-to memory mapping for the constructed
  Page 6: that is always accesses →that is always accessed
  Page 7: must developed →must be developed
  Page 8: is higher compared to sleep mode →is higher compared to light sleep mode
  Page 14: This application is an representative →This application is a representative

  The typos have been fixed.

Referee 2

- *However, the paper does not clearly distinguish the main contribution and the difference from the previous work. Original contribution needs to stand out clearly.*

  The contribution of the proposed work is the development of a combined approach that investigates the interleaving and memory mapping

options for a reconfigurable SIMD architecture. The current work combines and expands in a non-trivial way, the interleaving exploration presented in [7] and the data to memory mapping methodology presented in [4]. The current work is more than a simple application of the two approaches, one after the other. Such an approach cannot always guarantee a viable solution. The reason is that for each different interleaving solutions, there is a number of constraints on the placement of the data into the memory due to hardware limitations. If the constraints are not propagated into the data-to-memory mapping step, the final solutions suffers from data conflicts. Different interleaving solutions introduce different constraints. Therefore, there is a need to develop an integrated methodology to achieve the improvements of both approaches.

This has been clarified at the end of Section 3.

- *It would be interesting to see more workloads analyzed in the framework.*

  Unfortunately, it was impossible to analyze more benchmarks for the current revision. The chosen applications are representative candidates for the multimedia and the wireless domain. The results are representative for other applications with the same characteristics.

- *The paper focuses on SIMD architectures but it does not make any reference memory mapping optimization on GPUs. Also it would be interesting to see if the framework can capture more irregular access patterns and data mapping schemes (e.g. permutations, indirect addresses, etc.).*

  The presented methodology should be applicable on these cases but some modifications would be needed.

- *Some references worth discussing: Dymaxion: optimizing memory access patterns for heterogeneous systems, SC'11 Data reorganization in memory using 3D-stacked DRAM, ISCA'15 DL: A data layout transformation system for heterogeneous computing, InPar'12*

  In [2] the authors tackle the problem of sub-optimal data structure layouts in GPUs with a large number of parallel cores, especially for programs that are designed with a CPU memory interface in mind. An API is presented, that allows programmers to improve CUDA programs by optimizing memory mappings in order to increase the efficiency of memory accesses. The main differences of the current work is the different platform and the different type of code transformations. We

focus on an SIMD CPU and a dynamic clustered scratchpad memory compared to multicore GPUs and a static memory. We differentiate by focusing on interleaving as the preferred code transformation, as it is more suitable for the target applications. Another work that discusses the memory layout for GPUs is presented in [9]. Although the authors focus on off-chip DRAM memory optimization while the current work discusses the closest memory to CPU, they use a number of data layout transformations. We differentiate by studying data interleaving while the main focus in [9] is the transformations that increase data parallelism which is more important for their multicore GPU architecture. In [1] a number of common data reorganization operations such as shuffle, pack/unpack, swap, transpose, and layout transformations are presented. The goal is to study the cost of applying these operations in the memory during run-time. The target memory is 3D-stacked DRAM and additional hardware is employed in order to efficiently perform the reorganization operations with a low overhead. Apart from the different type of platform, the current work differentiates in the type of data reorganizations and the mapping of the reorganized data at the scratchpad memory at compilation time.

This has been clarified on Section 3.

Referee 3

- *In particular, what microarchitecture enhancement is included and how access patterns are extracted (in SW/HW) ? How address mapping is implemented and its policy? How interleaving is handled when there are bank conflicts and imperfect coalescing for SIMDs. In general, many details are missing. Fig 3 is not very helpful.*

  The application is fully analyzed during design-time, because it is a time consuming task. The access patterns are extracted in software and the possible interleaving options are explored. The mapping in the specific target architecture is also fully decided during design-time. The data-to-memory mapping exploration takes into consideration the platform and code limitation and proposes a mapping without data and bank conflicts. The employment of a scratchpad memory architecture is crucial for this. In contrast to cache memory systems, in which the mapping of data elements is done at run-time, in scratchpad memory systems the mapping is performed by the programmer or the compiler [5]. Unlike the cache memory, the scratchpad memory does not need tag search operations and it is the responsibility of the programmers or

compilers to correctly allocate code and data on the scratchpad memory [8]. Of course, this is possible for a small embedded system designed to run a specific application. When data allocation is a time consuming task, a cache memory is the preferred solution as the detection of a cache hit or miss is done automatically and any conflicts can also be resolved by the platform at run-time. In our case, the application and the memory system are fully analyzed and the allocation of data to a scratchpad memory can be easily done and offer a more energy efficient solution.

This has been clarified on Section 4. Smaller changes have been made to better explain the platform and the flow presented in Fig. 3

- *Also, the authors may look into the recent work by Akin, et al. "Data Reorganization in Memory Using 3D-stacked DRAM", ISCA 2015.*

  In [1] a number of common data reorganization operations such as shuffle, pack/unpack, swap, transpose, and layout transformations are presented. The goal is to study the cost of applying these operations in the memory during run-time. The target memory is 3D-stacked DRAM and additional hardware is employed in order to efficiently perform the reorganization operations with a low overhead. Apart from the different type of platform, the current work differentiates in the type of data reorganizations and the mapping of the reorganized data at the scratchpad memory at compilation time.

  This has been clarified on Section 3.

- *A minor point is the chosen benchmarks are not very irregular if appropriate data layouts are chosen. This is fine, but the authors may consider other possibilities such as graphs and sparse matrices.*

  The term irregular has been replaced with more descriptive terms, i.e. application with holes in their access pattern, as also noted by Referee 1.

References

[1] Berkin Akin, Franz Franchetti, and James C Hoe. Data reorganization in memory using 3d-stacked dram. In *Proceedings of the 42nd Annual International Symposium on Computer Architecture*, pages 131–143. ACM, 2015.

[2] Shuai Che, Jeremy W Sheaffer, and Kevin Skadron. Dymaxion: optimizing memory access patterns for heterogeneous systems. In *Proceedings of 2011 international conference for high performance computing, networking, storage and analysis*, page 13. ACM, 2011.

[3] DesignWare Memory Compilers. *40LP-TSMC Datasheet*.

[4] Iason Filippopoulos, Francky Catthoor, and Per Gunnar Kjeldsberg. Exploration of energy efficient memory organisations for dynamic multimedia applications using system scenarios. *Design Automation for Embedded Systems*, pages 1–24, 2013.

[5] Yuriko Ishitobi, Tohru Ishihara, and Hiroto Yasuura. Code placement for reducing the energy consumption of embedded processors with scratchpad and cache memories. In *Embedded Systems for Real-Time Multimedia, 2007. ESTIMedia 2007. IEEE/ACM/IFIP Workshop on*, pages 13–18. IEEE, 2007.

[6] Bingfeng Mei, Serge Vernalde, Diederik Verkest, Hugo De Man, and Rudy Lauwereins. Dresc: A retargetable compiler for coarse-grained reconfigurable architectures. In *Field-Programmable Technology, 2002.(FPT). Proceedings. 2002 IEEE International Conference on*, pages 166–173. IEEE, 2002.

[7] Namita Sharma, TV Aa, Prashant Agrawal, Praveen Raghavan, Preeti Ranjan Panda, and Francky Catthoor. Data memory optimization in lte downlink. In *Acoustics, Speech and Signal Processing (ICASSP), 2013 IEEE International Conference on*, pages 2610–2614. IEEE, 2013.

[8] Stefan Steinke, Lars Wehmeyer, Bo-Sik Lee, and Peter Marwedel. Assigning program and data objects to scratchpad for energy reduction. In *Design, Automation and Test in Europe Conference and Exhibition, 2002. Proceedings*, pages 409–415. IEEE, 2002.

[9] I-Jui Sung, Geng Daniel Liu, and Wen-Mei W Hwu. Dl: A data layout transformation system for heterogeneous computing. In *Innovative Parallel Computing (InPar), 2012*, pages 1–11. IEEE, 2012.