Some Experiences on Dynamic Memory Management Refinement at System-Level for Multimedia Applications

David Atienza*, Marc Leeman†, Jose M. Mendias*, Francky Catthoor‡, Chantal Ykman‡
Vincenzo De Florio†, Geert Deconinck†, Rudy Lauwereins‡

*DACYA/UCM, Avda. Complutense s/n, 28040 Madrid, Spain. Email: {datienza, mendias}@dacya.ucm.es
†ESAT/K.U. Leuven, Kasteelpark Arenberg 10, B3001 Heverlee, Belgium. Email: {name.surname}@esat.kuleuven.ac.be
‡IMEC vzw, Kapeldreef 75, 3001 Heverlee, Belgium. Email: {name.surname}@imec.be

Abstract—Nowadays, 3D multimedia applications have grown rapidly in number and consist of complex systems (e.g. 3D graphical processing or games) that process extensive amounts of data to create 3D images and results. This produces high-cost and high-power consumption systems whereas a superior these multimedia applications, the dynamic memory subsystem is currently one of the main sources of power consumption and its inattentive management can affect severely the performance and power consumption of the whole system. In this paper, we illustrate a new system-level method to explore and refine the dynamic memory management of multimedia systems on current typical case studies, i.e. a relatively new 3D image reconstruction system and a 3D simulation game. This method is based on an analysis of the access pattern, amount of memory used and power consumption estimations. With this information, a phase-wise exploration and refinement flow is used to optimize the system at the different phases of its hardware-oriented design process. As the results in the case studies show, our system-level method achieves great improvements in memory footprint, power consumption and performance for multimedia applications.

I. INTRODUCTION

The complexity of 3D multimedia applications has increased enormously in the last years. Presently, they process large amounts of data and require a good level of performance. Representative examples of their use can be archaeological site reconstruction, film industry and 3D games [6], [13], [1]. These examples need handheld devices as final platforms to achieve quick on-site visualization, processing and real-time interaction with the user. However, the complexity of multimedia applications results in high cost and power consumption systems while current markets demand the opposite. New multimedia algorithms greatly rely on dynamic memory (DM) due to the unpredictability of the input data at compile-time. Furthermore, they demand large amounts of memory that make the DM subsystem one of the main sources of power consumption, memory usage and (due to its latency and delay) heavily affect the global performance of the system.

With the previous characteristics, classical hardware design refinements can only partially compensate for the growing hardware/software gap in power and speed [11]. Recently, the critical design improvements for very large scale integration systems have been shifting toward the software part.

In order to optimize embedded multimedia systems, detailed power consumption, memory use and performance profilings must be available at an early stage of the design flow. For statically allocated data and operations, this information is available from modern analysis and profiling tools. However, for dynamically (de)allocated data types (further called DDTs) implementations, the situation is much worse. At the most, summarized information can be available for the pools of data, but the details about individual DDTs are lost then. Also, simulation data can be generated at a much closer level to the final implementation on a certain platform, e.g. at instruction or cycle accurate hardware level. This is however very CPU-time consuming and requires the complete mapping trajectory, thus we do not include it currently since it is not acceptable for system-level exploration purposes. In dynamic multimedia applications, DDTs are a very relevant power consumption and speed factor (in our case studies, they can consume up to 60% of the total power used in the memory subsystem).

In this paper, we illustrate a new phase-wise system-level method to explore and refine DM management on embedded multimedia systems in an early stage of the system design and integration flow. As the case studies show, it uses a high level (i.e. from C++ code) profiling method [10] to extract the necessary run-time information for a power-aware, performance and memory usage refinement of the DDT implementations and DM managers involved. To this end, the three important factors that influence power consumption and global performance of the memory subsystem are analyzed per DDT (and not for the entire pool of data only), i.e. memory usage pattern over time, memory accesses and data access mechanisms. This analysis allows the optimizations in the latter stages of the method, where it is used to optimize the whole DM management subsystem early on during system integration, enabling large savings in the design-time of the system.

The remainder of this paper is organized as follows. In Section II we motivate the optimization method and summarize its phases. This is however not the focus of this paper. In Section III we describe the related work. In Section IV we illustrate the large effects of our approach on the cost and performance of modern dynamic multimedia applications using our case studies. This is the core of the paper. Finally, in Section V we draw our conclusions.
II. MOTIVATION AND METHOD

The design flow of new multimedia applications (e.g. 3D graphical processing or games) involves a very time-consuming effort in the correct design of the software and dynamic part of the system. It includes the definition of the interdependent and variable at run-time DDTs of the complex algorithms used. These systems depend on an efficient DM management, which is one of the most difficult design challenges for an efficient mapping on low-power and high-speed platforms without extensive system support for DM.

As mentioned in Section I, the use of DM can have important influence on the overall algorithm performance. Our system design method (see [9], [10] for a more detailed description) analyzes this influence using a phase-wise method, which starts from a high level specification and more information concerning the final implementation is added on each phase. As a result, estimations from previous phases are refined with new and more precise information. Then, an early idea of implementation costs and requirements is achieved and the developers can make changes in the system without expensive re-iterations through the entire design flow.

The illustrated phase-wise method is divided in three main phases according to a virtual structure of typical embedded multimedia systems where data is transferred into memory via DMA, and then the execution is triggered and finished by software interrupts. This structure is depicted in Figure 1. First, the method analyzes the access pattern of the DDTs involved and optimizes their implementations preserving the system constraints, this phase is called Dynamic Data Type Transformation and Refinement (DDTTR), marked as number 1 in Figure 1. Secondly, global DM managers are studied to efficiently tackle the (de)allocation requests of the DDTs. This phase is named Dynamic Memory Management Refinement (DMMR), number 2 in Figure 1. Finally, an additional phase can be used to optimize further considering the physical characteristics of the final platform and memory subsystem, e.g. bank conflicts or cache policies. This final phase is Physical Memory Management (PMM), number 3 in Figure 1.

In the following paragraphs, the first two phases of the method are only briefly summarized because they are not the focus. They will also be clarified in Section IV for the presented case studies, in so far as needed. The PMM optimizations are not applied because our aim is not a specific platform yet and lie out of the scope of this paper.

The first phase of the method is DDTTR, which refines the DDTs of the application under study. Overall, they define the way in which memory is allocated, accessed and freed.

The basis of all DDTs is the combination of a number of basic data types, e.g. trees, arrays or lists, in a single or multi-layered organization [9]. Then, they implement a certain set of basic operations as add, get or set to provide the interface to the user. Our method uses an object oriented programming approach that offers a common interface of operations for the useful multi-layered DDTs implementations in embedded multimedia systems [9]. This allows to design the system in a completely platform and implementation-independent way. Different combined DDTs offer possible trade-offs between performance, memory use and power consumption. For example, an array is rigid in size and memory usage but has fast access, while a simple linked list is the other way around. Finally, these trade-offs for the final solution of the DDT implementations are exploited according to the required restrictions of the system.

After the DDTs have been optimized, a suitable use of the global DM available must be provided according to the design constraints. In fact, an inappropriate management of the DM used by the DDTs reduces enormously the performance of the application and increases its power consumption [9]. Therefore, a well-adjusted DM manager must be selected. This is done in the second phase of the method, i.e. DMMR.

The definition of an appropriate DM manager implies a full analysis of the dynamic behaviour of the application under study. After this specific behaviour is known and we can characterize the actions to be performed by the adjusted DM manager (e.g. (de)allocation, block splitting or coalescing), our method applies high level strategies [17] and a set of possible decision choices that has been specifically defined for embedded multimedia systems [9]. As a result, a set of possible DM managers are available for the developer and he can decide according to the needed constraints, e.g. low-power consumption, performance, etc.

III. RELATED WORK

Currently the foundations of a good DM management in a general-context are already well established and research has been done at its different levels.

Although the number of building blocks in multi-layered data types is limited [9], each developer tends to write his own custom DDTs for each application. Therefore, the number of implementation alternatives is defined by the experience and inspiration of the developers, and the available time to implement them. Presently, suitable access methods and DDT implementations have started to be proposed for multimedia applications considering their features [7], [9]. Also, in the software community much literature is available about DM management implementations and policies [17]. They use the locality of references and other techniques to analyze and minimize fragmentation or enable a fast time-response for systems with speed constraints [17]. However, this earlier work does not provide a complete and useful search space for a systematic exploration in multimedia applications.

Usually in memory management for embedded systems [12], the DM is partitioned into fixed blocks to store...
the DDTs. Then, the free blocks are placed in a single linked list [12] due to performance constraints. Another recent method is the simulation of the system with different general-purpose DM managers. In this case, [4] proposes an infrastructure to build DM managers using C++ templates. However, these methods focus on performance and do not consider power consumption issues, as embedded systems require.

Work can be found to estimate power consumption based on software instead of at a circuit-level. Most of it use an instruction level analysis [14], but more recently research has been developed for assembly code and a higher abstraction level [15], [16]. Nevertheless, they employ analyses without run-time profiling that are not enough for DM applications.

Due to large scale integration systems, several analytical and abstract power estimation models at the architecture-level have received more attention lately [5]. However, they do not focus on the DM hierarchy of the system and the power consumption from the DDTs at the software level.

In a complete different field, telecom network applications, [18] proposes a power exploration method driven by memory accesses. This work handles systems with independent DDTs and the main focuses are the dynamically created tasks due to network connections. It does not use run-time behaviour analysis because, after the system initialization, a snapshot of the memory at any moment during execution gives a good run-time memory-behaviour image, which can be used to determine the memory contribution of each DDT to the system. The DMMR optimizations we propose are not studied.

New system-level power consumption optimizations for embedded systems are explained in [3]. However, system-level refinements of memory management in applications with complex dynamic behaviour have not received much attention.

IV. DEMONSTRATION OF THE METHOD

The applied method is illustrated using two case studies that represent different modern multimedia application domains: the first case study is part of a new 3D image reconstruction system and the second one is an interactive 3D game.

The first case study forms one of the sub-algorithms of a 3D reconstruction algorithm [13] that works like 3D perception in living beings, where the relative displacement between several 2D projections is used to reconstruct the 3rd dimension. The DMMR optimizations we propose are not studied.

New system-level power consumption optimizations for embedded systems are explained in [3]. However, system-level refinements of memory management in applications with complex dynamic behaviour have not received much attention.

The algorithm uses internally several DDTs whose sizes cannot be determined until the system is running because they depend on factors (e.g. textures in the images) determined outside the algorithm (and uncertain at compile-time). Furthermore, due to the image-dependency related data, the initial DDT implementations do not fit in the internal memory of current embedded processors. These DDTs are the following:

- **ImageMatches (IMatches)** is the list of pairs where one point in the first image, matches another one on the second image based on a neighborhood test [13].
- **CandidateMatches (CMatches)** is the list of candidates that must go through a normalized cross-correlation evaluation of a window around the points [2].
- **MultiMatches (MMatches)** is the list of pairs that pass the aforementioned evaluation criterion. Still one point from the first image can be listed in multiple candidate pairs for a later best match selection.
- **BestMatches (BMatches)** is the subset of pairs where only the best match for a point (according to the test already mentioned) is retained.

These DDTs were originally implemented using double linked lists and exhibit an unpredictable memory behaviour, typical in many state-of-the-art 3D vision systems [2] since they use some sort of **dynamic candidate selection** followed by a criterion evaluation. Figure 2 shows this behaviour in the generation order of the DDTs. First, IMatches is generated after a neighborhood test is applied to pairs of corners detected in two images. Then, CMatches is created using the information acquired from previous images. Finally, MMatches and BMatches are generated with the pairs that pass a normalized cross correlation evaluation [13].

Although the global interaction between these DDTs is defined by the algorithm (see Figure 2), it is not clear how each DDT affects the final system figures (e.g. memory footprint). Therefore, our method is used to explore and refine it.

First of all, DDTTR is applied to optimize the DDTs. After
FIG. 3. Memory footprint over time. All plots mapped on the left axis, except CandidateMatches and CMCopyStatic (right axis). Left graph, original implementation. Right one, optimal implementation suggested by DDTTR with CMCopyStatic removed.

FIG. 4. Block-size allocation bins used in the Kingsley DM manager [17] by the 3D reconstruction system before (left graph) and after DDTTR (right graph) automatically inserting the profiling code and running the tools presented in [10], profiling information is obtained. Then, a memory use graph is generated (left chart in Figure 3) and memory accesses, memory footprint and energy dissipation figures are calculated (see Table I). For the energy estimations, the memory model described in [8] is used. It is a complete energy/delay/area model for embedded SRAMs that is able to scale to different technology nodes (we use the 0.13 μ technology node for the results). This model depends on memory footprint factors (e.g., size, internal structure or leaks) and factors originated by memory accesses (e.g., number of accesses or technology node used). Note that any other model can be used just by replacing that module in the tools.

The analysis of the profiling information (Table I and Figure 3) shows how the DDTs affect the system. First, CMatches is the largest DDT. Secondly, IMatches has frequent accesses. Finally, CMCStatic (an “optimized dynamic array” that reduces the accesses to CMatches) consumes an important part of the energy used by the system.

After the subsequent exploration step of DDTTR (see [10] for more details), our tools suggest an ideal solution for the DDTs trying to minimize power consumption. The final DDT implementations consist of 2-layered dynamic array structures (pointer-arrays to arrays). First, an external dynamic array of 10 positions; then, each position is another array of 756, 1024 or 16384 Bytes (B) depending on which DDT. With these optimized DDTs, CMatches is now fast enough to interact directly with IMatches and Mmatches, thus CMCStatic is removed. This removal improves even more memory footprint, but has no effect on the performance (speed) of the global application, already improved by DDTTR.

TABLE I

<table>
<thead>
<tr>
<th>Variable</th>
<th>memory accesses</th>
<th>memory footprint (B)</th>
<th>energy 13μm tech.(μJ)</th>
</tr>
</thead>
<tbody>
<tr>
<td>IMatches</td>
<td>1.20×10⁵</td>
<td>5.14×10⁴</td>
<td>0.18×10³</td>
</tr>
<tr>
<td>CMatches</td>
<td>8.44×10⁵</td>
<td>2.75×10⁵</td>
<td>3.03×10³</td>
</tr>
<tr>
<td>CMCStatic</td>
<td>6.24×10⁴</td>
<td>1.08×10⁵</td>
<td>4.48×10⁴</td>
</tr>
<tr>
<td>Mmatches</td>
<td>1.84×10⁵</td>
<td>3.62×10⁵</td>
<td>0.02×10⁵</td>
</tr>
<tr>
<td>Bmatches</td>
<td>1.66×10⁵</td>
<td>3.07×10⁵</td>
<td>0.02×10⁵</td>
</tr>
<tr>
<td>Total</td>
<td>2.14×10⁵</td>
<td>3.80×10⁵</td>
<td>4.80×10⁴</td>
</tr>
</tbody>
</table>
The object size and each region uses a double linked list to optimize the time required to traverse and access the data. Different DDTs in the application. Every block allocated stores the relative positions of the remaining ones are updated. As in the previous case, each basic element is 6 floats. Also, the DDT is the list of vertical walls detected in the input images. When the user moves, new walls are detected and the relative positions of the remaining ones are updated. VWalls is the list of horizontal walls detected in the images. It is also updated after a movement of the user. Balls is the list of balls that are currently in the system. It is frequently accessed to update the position of the balls in the scene due to movements of the user and their own actions (e.g., destroyed, created, etc.).

These DDTs were originally implemented using single linked lists and due to their unpredictable DM behaviour, they cannot be placed in current embedded handheld devices.

As in the first case study, the DM management of the system is refined with our method. DDTTR is applied and a detailed timing graph is produced (see Figure 5). Also, detailed profiling information of the DDTs is obtained (see Table III). Then, DDTTR explores possible implementations for each DDT considering trade-offs of power consumption, memory footprint and performance, as Figure 6 shows. Minimizing global power consumption, in the final global solution selected (among the circled points in Figure 6), the DDTs VWalls and HWalls are implemented as 2-layered structures. The first level is a dynamic array of 10 positions. Then, another one of 56 basic elements for VWalls and 26 for HWalls. In both cases, each basic element is 6 floats. Also, the DDT Balls is implemented as a 2-layered structure. A dynamic array of 10 positions stores dynamic arrays of 179 basic elements of 1 float each. As Table IV shows, these optimized DDT implementations reduce significantly energy dissipation and memory footprint compared to the original DDTs.

After the DDTs are optimized, DMMR is used to design a custom DM manager minimizing global power consumption for the game controller. This manager discerns two different access patterns and behaviours of the DDTs: (1) VWalls and

<table>
<thead>
<tr>
<th>Variable</th>
<th>memory accesses</th>
<th>memory footprint (B)</th>
<th>energy .13μm tech.(μJ)</th>
</tr>
</thead>
<tbody>
<tr>
<td>IMatches</td>
<td>4.02×10^4</td>
<td>6.84×10^2</td>
<td>2.91×10^2</td>
</tr>
<tr>
<td>CMatches</td>
<td>3.89×10^4</td>
<td>1.27×10^5</td>
<td>7.02×10^2</td>
</tr>
<tr>
<td>VMatches</td>
<td>7.68×10^3</td>
<td>3.81×10^3</td>
<td>0.03×10^1</td>
</tr>
<tr>
<td>HMatches</td>
<td>7.16×10^3</td>
<td>3.78×10^3</td>
<td>0.02×10^1</td>
</tr>
<tr>
<td>Total</td>
<td>8.06×10^4</td>
<td>1.28×10^4</td>
<td>9.98×10^4</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Variable</th>
<th>memory accesses</th>
<th>memory footprint (B)</th>
<th>energy .13μm tech.(μJ)</th>
</tr>
</thead>
<tbody>
<tr>
<td>VWalls</td>
<td>4.39×10^4</td>
<td>1.72×10^4</td>
<td>2.94×10^4</td>
</tr>
<tr>
<td>HWalls</td>
<td>4.01×10^4</td>
<td>1.64×10^3</td>
<td>0.28×10^3</td>
</tr>
<tr>
<td>Balls</td>
<td>2.16×10^3</td>
<td>8.42×10^3</td>
<td>1.91×10^3</td>
</tr>
<tr>
<td>Total</td>
<td>3.02×10^4</td>
<td>1.17×10^4</td>
<td>5.13×10^4</td>
</tr>
</tbody>
</table>
In the pool of walls, the DDT unordered simple link list that keeps a pointer to the last are stored in a double link list with LIFO order. Finally, no combination of DDTs, the circled ones are global solutions (for all DDTs).

Fig. 5. DM behaviour of the game controller module for 6 frames

TABLE IV

<table>
<thead>
<tr>
<th>Variable</th>
<th>Memory accesses</th>
<th>Memory footprint (B)</th>
<th>Energy (J/m)</th>
</tr>
</thead>
<tbody>
<tr>
<td>HWalls</td>
<td>4.16x10^6</td>
<td>8.30x10^7</td>
<td>1.52x10^7</td>
</tr>
<tr>
<td>HWalls</td>
<td>3.62x10^6</td>
<td>6.99x10^2</td>
<td>1.28x10^2</td>
</tr>
<tr>
<td>Balls</td>
<td>2.20x10^6</td>
<td>7.20x10^3</td>
<td>7.54x10^2</td>
</tr>
<tr>
<td>Total</td>
<td>9.98x10^6</td>
<td>8.72x10^4</td>
<td>1.03x10^3</td>
</tr>
</tbody>
</table>

In the end, normalized memory footprint is improved up to 22.48% and power consumption up to 75.3% compared to the original implementation. Furthermore, there is a global speedup factor of one order of magnitude for the 3D game.

V. CONCLUSIONS

As multimedia applications have grown in complexity, the design flow of current multimedia systems require extensive time-consuming efforts to reach a correct design of their software and dynamic part for an efficient system integration. In this refinement process, DM management is one of the crucial and most difficult parts in multimedia applications due to the unpredictability of their input data and events (e.g. images or movements of the user). Furthermore, for portable devices, power consumption is a primary constraint. In this paper, we have illustrated a new system-level exploration and refinement method for DM management on two current multimedia applications, namely a new 3D image reconstruction system and a real world interactive 3D game. Our results in these case studies show that multimedia applications greatly benefit from optimal DM management systems. Places where previous implementations formed clear bottlenecks in the code were completely removed by better DDT implementations and an optimal DM manager. As a result, the optimized applications can be ported efficiently to the final embedded multimedia platforms.

ACKNOWLEDGEMENTS

This work is partially supported by the Spanish Government Research Grant TIC2002/0750, the Fund for Scientific Research - Flanders (Belgium, F.W.O.) through project G.0036.99 and a Postdoctoral Fellowship for Geert Deconinck.

REFERENCES