UDK 004.254 (045)
Мельник В.М., Мельник К.В., Багнюк Н.В., Гринюк С.В.
Луцький національний технічний університет

ANDROID-BUILT CODE GENERATION MODELLING FOR HETEROGENEOUS ARCHITECTURE

Melnyk V., Melnyk K., Bagniuk N., Hryniuk S. Android-built code generation modelling for heterogeneous architecture. The success of Android is based on its unified Java programming model that allows to write platform-independent programs for a variety of other target platforms. However, this comes at the cost of performance. As a consequence, Google introduced APIs that allow to write native applications and to exploit multiple cores as well as embedded GPUs for compute-intensive parts. Here is proposed code generation techniques in order to target the Renderscript and Filterscript APIs. Render script harnesses multi-core CPUs and unified shade GPUs, while the more restricted Filterscript also supports GPUs with earlier shade models. Our techniques are focused on image processing applications that allow to target these APIs and OpenCL from to a common description. We further supersede memory transfers by sharing the same memory region among different processing elements on HSA platforms. As reference, we use an embedded platform hosting a multi-core ARM CPU and an ARM Mali GPU. We show that our generated source code is faster than native implementations in OpenCV as well as the pre-implemented script intrinsic provided by Google for acceleration on the embedded GPU.

Мельник В.М., Мельник К.В., Багнюк Н.В., Гринюк С.В. Модель генерації вбудованого Андроїд-коду для гетерогенної архітектури. Переваги Android базуються на його уніфікованій моделі програмування Java, що дозволяє писати незалежні від платформи програми для різних цільових платформ. Проте це відбувається за рахунок зниження продуктивності. Як наслідок, Google представив API-інтерфейси, що дозволяють писати власні програми та використовувати кілька ядер, а також вбудовані графічні процесори для ресурсомістких частин. В статті пропонується методи генерації коду з метою виявлення API, Renderscript і Filterscript. Renderscript використовує багатоядерні процесори і графічні процесори уніфікованого відтінку, в той час як більш обмежена Filterscript також підтримує графічні процесори з більш ранніми тіньовими моделями. Наша технологія орієнтована на додатки для обробки зображень, які дозволяють призначати подібні інтерфейси і OpenCL з загального опису. Крім того, здійснюється витіснення перекладів пам'яті, розділяючи одну і ту ж область пам'яті між різними елементами обробки на HSA-платформах. В якості порівняння використовується вбудована хостинг-платформа багатоядерного процесора ARM і ARM Mali GPU. Продемонстровано, що згенерований вихідний код є швидшим, ніж власна реалізація в OpenCV, а також попередньо реалізований сценарій, характерно наданий Google для прискорення на вбудованому GPU.

Introduction

The steady desire for new applications and higher display resolutions drives the development of embedded platforms and the need for faster, more powerful processors at the same time. Today's mobile platforms found in smartphones and tablets host multi-core Central Processing Units (CPUs) and even programmable embedded Graphics Processing Units (GPUs) to deliver the demanded performance. This raises the question how to harness the processing power of these parallel and heterogeneous platforms. As a remedy, Google proposed two new parallel programming concepts for Android that allow to target CPUs as well as GPUs: Renderscript and Filterscript.

These programming models were designed for the predominant application domain of image processing with portability in mind. While applications in Android use Java as programming language, Renderscript and Filterscript are based on C99. Hence, Java programmers have to write low-level C code in order to benefit from the high performance that is provided by these new programming models. This work proposes code generators that allow to automatically generate Renderscript and Filterscript code. The generated target code is derived from a Domain-Specific Language (DSL) for image processing algorithms [1]. The proposed code generators for Renderscript and Filterscript are based on the existing compiler infrastructure, which provides back ends for CUDA and OpenCL on discrete, standalone GPU accelerators.

The focus in this work is on the code generation that allows to target the different components in today's heterogeneous embedded platforms.

We present the first code generator for Renderscript and Filterscript on Android platforms starting from an abstract high-level representation. The generated implementations are even faster compared to the target-specific implementations in the *Open Source Computer Vision* (OpenCV) framework. At the moment, the

algorithm description is compact and requires only a fraction compared to available highly optimized implementations.

The target code is generating for embedded *Heterogeneous System Architecture* (HSA) platforms. With HSA, CPU and GPU share the same physical memory (Fig. 1). This allows us to avoid extensive memory transfers and enables the employment of heterogeneous resources where the same data has to be accessed frequently from different compute resources.

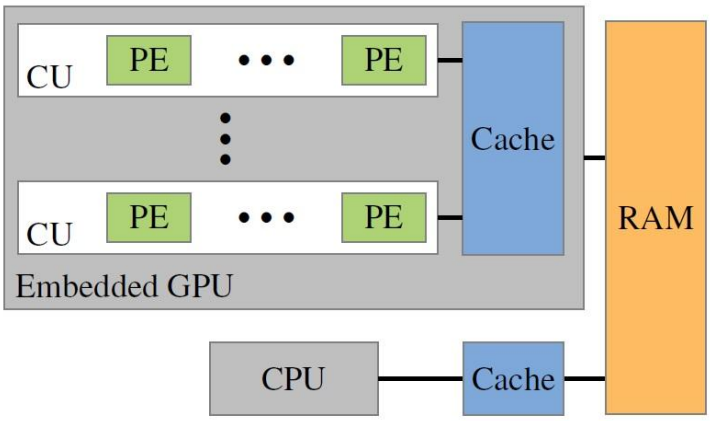

Fig. 1. Structure of typical HSA platform with an embedded GPU

Fig. 1: Structure of a typical HSA platform with an embedded GPU

Programming models on embedded devices

The Android operating system is widespread on end user mobile devices but it becomes more and more interesting to use it also on other embedded devices as found in the industrial automation or the automotive sector. For example, compute-intensive advanced driver assistance systems such as self- and convoy-driving technology, parking guidance, and better infotainment systems are becoming more and more important for automotive.

For such compute-intensive tasks, energy efficient embedded GPUs are of particular interest. However, there exists no common programming language that allows to harness also embedded GPUs. Android, for example, provides multiple programming models that allow to target a huge variety of devices with different usage scenarios in mind:

a) SDK: The Android Software Development Kit (SDK) is based on the Java programming language. The SDK source code is compiled to bytecode that is executed by the Dalvik V.M., which is either directly interpreting the bytecode or compiling it into machine-specific instructions and executing it. Hardware-specific characteristics are hidden and not exposed to the programmer. For that reason it is not possible to gain benefit from certain hardware specific optimizations like vectorization. This dramatically limits the overall achievable performance.

b) NDK: The Android Native Development Kit (NDK) promises much better execution performance. Even though complete Android applications can be developed using C/C++ in the NDK, native code is usually executed by an application that is written using the SDK through the Java Native Interface (JNI).

Using the NDK, hardware-specific characteristics are transparent to the programmer and can be utilized, for instance, by compiler intrinsics. For efficiently supporting a wide range of different CPU architectures, native code needs to be rewritten beforehand in a way that exploits particular architecture features to achieve high performance. For that reason, usually only compute-intensive code segments are realized in native parts of an Android application.

c) Renderscript: Google first introduced the Renderscript programming language in 2011. The aim of this new language is to provide a programming model that avoids performance issues of the SDK without introducing portability problems the NDK suffers from. Renderscript is based on C99 and provides additional support for vector types. The Renderscript front end compiler generates an intermediate representation which

is then further compiled to native code that is optimized for a specific available target architecture like CPUs, Digital Signal Processors (DSPs), and GPUs. Unlike other parallel computing Application Programming Interfaces (APIs) such as OpenCL or CUDA, Renderscript was not designed with performance as the primary goal. Hardware-specific features like local memory are hidden from the programmer and, hence, target-specific optimizations like tiling may not be exploited.

The number of threads is directly inferred from the output buffer where each element is processed by a single thread. Another limitation is that the actual execution target (CPU, GPU, DSP) cannot be specified and is automatically chosen by the runtime system. These limitations ensure portability at the expense of absolute performance.

d) Filterscript: Filterscript is a subset of Renderscript with certain limitations and stricter constraints to ensure a wider compatibility among CPUs, GPUs, and DSPs. Filterscript files are used and compiled the same way as Renderscript.

The major difference is that pointers are not allowed. Therefore, memory cannot be directly read using linear arrays. Instead, the provided access API functions must be used. Only gather reads are supported, which means that only one output value can be written per thread (in contrast to scatter writes). Instead of assigning the value to a buffer, it is returned by the kernel function and written to the global ID of the current thread. Further limitations are relaxed floating point precision and lack of 64-bit built-in types.

e) OpenCL: The Open Compute Language (OpenCL) programming model is not officially supported on Android devices. However, on recent devices, like the Nexus 4 smartphone and the Nexus 10 tablet, a working OpenCL driver can be found within the system libraries. Note that Google pulled the unofficial OpenCL support in Android 4.3.

Providing a well-known common API like OpenCL cannot hide the fact that embedded GPU architectures vary considerably from their desktop counterparts. For instance, the local memory is not dedicated on-chip memory, but is instead mapped to global memory. Another major difference is that compute cores of most embedded GPUs are Very Long Instruction Word (VLIW) architectures and gain a huge benefit from vectorization or the use of vector types.

OpenCL supports the allocation of host accessible memory using map() and unmap(). On desktop GPUs, this feature can be used to define page-locked host memory, which may speed up memory transfers. On HSA platforms like the ARM Mali [2] and AMD Fusion [3], these operations are used to avoid copying buffers. Because the main memory is shared among CPU and GPU, both can access the same memory region without copies.

Target code generation

This section introduces the target code generation for the parallel computing APIs introduced in Section II. We use a DSL for image processing as basis for target code generation and employ source-to-source translation to target different parallel programming models.

The Heterogeneous Image Processing Acceleration (HIPA^{cc}) Framework

The DSL provided by the HIPA^{cc} framework [1] is based on C++ and provides built-in classes for two-dimensional images and other objects such as filter masks. An image in the DSL stores the image pixels and can be initialized from plain C data arrays:

```
uchar *image = readPGM(&width, &height, "lena.pgm");
Image<uchar> in(width, height);
in = image;
```

Similarly, a filter mask can store the stencil used for a convolution:

```
const float filter_mask[] =
{
  0.057118f, 0.124758f, 0.057118f,
  0.124758f, 0.272496f, 0.124758f,
  0.057118f, 0.124758f, 0.057118f
}
Mask<float> mask(3, 3);
mask = filter_mask;
```

Using these data abstraction classes, computations on multidimensional image objects can be defined. A computational kernel is defined as a C++ class that holds a kernel() method describing the computation on a single pixel. Within the kernel method, all memory accesses are relative to the current pixel and only members of the C++ class can be accessed. Considering the Gaussian blur filter as an example, its computation can be expressed using relative memory accesses to the mask filter mask and the input image. The result is written back using the output() method:

```
void kernel()
{
  float sum = 0;
  int range = size/2;
  for (int yf = -range; yf <= range; ++yf)
    for (int xf = -range; xf <= range; ++xf)
      sum += mask(xf, yf) * input(xf, yf);
  output() = (uchar) sum;
}
```

A more concise and expressive syntax for common computational patterns such as convolutions are provided by the HIPA^{cc} DSL as well: the convolve() method describes the convolution of an image with a mask:

```
void kernel()
{
  output() = convolve(mask, SUM, [&] () -> float {
    return mask() * input(mask); });
}
```

Using the convolve method is not only more compact, it gives the source-to-source compiler also more freedom to optimize the code.

The HIPA^{cc} compiler [1] generates target CUDA and OpenCL code from programs written in this DSL. Using source-to-source translation, instances of DSL C++ classes are replaced by corresponding API calls to the CUDA and OpenCL runtime library provided by HIPA^{cc}. Compute kernels, in contrast, are not mapped one-to-one to corresponding CUDA and OpenCL. Instead, the Abstract Syntax Tree (AST) of a kernel is analyzed and optimizations are applied such as staging image pixels into local memory or mapping multiple iterations to one GPU thread (loop unrolling).

Memory accesses are then redirected to memory fetches from global memory, texture memory, or local memory-depending on the target device. Similarly, Mask accesses get mapped to constant memory or propagated as constants in case the operator is described using the convolve() function.

Renderscript and Filterscript Support for HIPA^{cc}

We have extended HIPA^{cc} for Renderscript and Filterscript support, adding a new back end for each API. Program parts of the DSL responsible for resource management are mapped to corresponding commands in the runtime library, which we provide. The compute-intensive kernels, however, are translated into Renderscript and Filterscript kernels. These get initialized at program start and can be executed afterwards.

1) Memory Access Mapping: As highlighted in Section II, memory accesses are handled differently in OpenCL, Renderscript, and Filterscript. Therefore, the introduced back ends map reads and writes to an Image to corresponding API calls and memory array accesses. To illustrate this, we consider a simple read from an Image, followed by a write in HIPA^{cc}:

```
Image<uchar> input;
...
void kernel()
{
  uchar val = input();
  output() = val;
}
```

In OpenCL, these memory accesses are mapped to 1D memory arrays that are added to the signature of the kernel function with corresponding attributes indicating that the arrays reside in global CPU or GPU memory. In case neighboring pixels are read, the x and y index is adjusted accordingly:

```
__kernel void kernel(__global const uchar *input,
__global uchar *output, ...)
{
   uchar val = input[y*width + x];
   output[y*width + x] = val;
}
```

Renderscript provides data buffers for storing image data (rs_allocation) and rsGetElementAt API calls for reading/writing data elements. Rather than writing the result to the current iteration point (i. e., writing to the first kernel parameter), the result is stored to _iter, the currently processed pixel:

```
rs_allocation input;
...
void kernel(uchar *_iter, uint32_t x, uint32_t y)
{
   uchar val = rsGetElementAt_uchar(input, x, y);
   *_iter = val;
}
```

In Filterscript, data buffers are defined and read as in Renderscript, but no API calls are provided for storing results. Instead, the result for the current thread (pixel) is returned by the kernel:

```
rs_allocation input;
...
uchar __attribute__((kernel)) kernel(uint32_t x,
uint32_t y)
{
   uchar val = rsGetElementAt_uchar(input, x, y);
   return val;
}
```

In order to map the execution of a kernel in Renderscript and Filterscript either to the CPU or to the GPU, environment variables are used.

In order to achieve high performance, the kernel has to be mapped to the memory hierarchy of the target architecture. GPUs provide multiple memory types apart from the global memory that are optimized for different access patterns such as constant memory, texture memory, or local memory. OpenCL allows to explicitly use these memory types in the source program. For example, filter masks are typically mapped to constant memory and read-only images with high spatial or temporal locality to texture memory or local memory. In contrast to this, Renderscript and Filterscript do not support any explicit mapping to the memory hierarchy.

2) Iteration Space Mapping: HIPA^{cc} allows the programmer to define a Region of Interest (ROI) in the output image to be computed. Similarly, only an ROI on input images can be read. This allows to work on images of different size in one kernel and to process only image regions of interest. The ROI on the output image defines also the iteration space and the number of threads required for kernel execution. This iteration space size is used as launch configuration in parallel compute APIs such as CUDA and OpenCL and offsets to the image are passed to the kernel in order to process only the ROI. However, the native API of Renderscript and Filterscript does not provide launch configurations up to Android 4.4. Instead, the buffer holding the image data defines also the launch configuration. That is, for each pixel in the image a thread is started resulting in an index space that is larger than the iteration space defined by the programmer.

To overcome this deficiency, there exist three approaches. First, we can define a buffer with the size of the iteration space. This temporary buffer stores the result of the kernel and is copied back to the ROI in the output image as specified by the programmer. This approach requires additional memory of ROI_{width} _ROI_{height} and requires one additional memory transfer of the same size. Second, we can define a dummy

buffer with dimensions equal to the iteration space and use this buffer to provide the index pace. Result pixels are not stored to this buffer, but to the buffer associated with the output image of the kernel. This approach requires in theory no additional memory and no additional memory transfers, but can only be used for Renderscript. 1However, it turns out that the Renderscript runtime still allocates memory for the buffer although no data is associated with the buffer. In Filterscript, the output pixel is not written to a buffer, but returned within the kernel. Third, we can use an index space with a size of the whole output image and add guards to the kernel so that only threads of the index space calculate pixels that are also part of the iteration space. This launches (IMG_{width} _IMG_{height})-(ROI_{width} _ROI_{height}) additional threads that do not compute pixels. While this approach is valid for Renderscript, the behavior is undefined in Filterscript in case no return statement is executed. Hence, we read the corresponding pixel value from the output image for index points outside of the iteration space and return those. This requires (IMG_{width}_IMG_{height})-(ROI_{width}_ROI_{height}) additional memory reads and writes for the Filterscript implementation, but has no memory allocation overhead. This approach is the only one that provides a valid iteration space mapping for Filterscript with only little overhead. Thus, this approach is followed by the code generator in this work.

3) *Vector Support*: HIPA^{cc} supports scalar data types for discrete GPUs from AMD and NVIDIA. These GPUs schedule scalar instructions to single lanes of their Single Instruction, Multiple Data (SIMD)-like architecture. Adjacent threads are mapped to adjacent lanes. The embedded CPUs and GPUs considered here require vector instructions, though. Otherwise only a fraction of peak performance can be achieved. Therefore, we add vector type support to HIPA^{cc}, which is compatible with the syntax in OpenCL and Renderscript.

Support for HSA Memory Management

The HIPA^{cc} DSL was designed for desktop systems and therefore has no particular support for HSA platforms. Device memory is abstracted from the developer by the Image class in HIPA^{cc}. Memory transfers to the device are handled implicitly by the framework. On HSA targets, it is possible to share memory between CPU and GPU by using host accessible memory that must be allocated using OpenCL API calls.

We added support for HSA platforms to HIPA^{cc} by extending its memory management to abstract host memory as well and implicitly manage map() and unmap() operations. This additional abstraction has the benefit that a) faster page-locked memory can be utilized on desktop systems, and b) the same memory region can be used for the CPU and GPU on HSA platforms in order to avoid memory copies. In case memory is allocated by third party frameworks (e. g., OpenCV or FreeImage), the programmer can still manage host memory explicitly.

Related works

While there exist a wide range of frameworks and compilers that generate low-level assembly code for embedded architectures, there is only little related work on targeting the new compute APIs Renderscript and Filterscript.

For example, the Portland Group introduced the PGCL framework [6], which adds support for OpenCL to Android on the ST-Ericsson NovaThor platform. Their compiler supports the NEON instruction set and vectorization for ARM multi-core CPUs. Halide [7], a DSL for image processing provides a back end for ARM NEON, and the OpenCV [8] library utilizes also the ARM NEON instruction set. These frameworks have all in common to target only the CPU. More recently, a source-to-source translator for mapping annotated Java code (using pragmas similar to OpenMP within source code comments) to Renderscript and OpenCL was presented [9]. Hence, the developer requires profound knowledge of OpenMP, while our domain-specific approach abstracts from low-level architecture details.

Qian, Zhu, and Li [10] compare and analyze the programming model of the SDK, NDK, and Renderscript considering usability aspects such as programmability, but also performance. They conclude that Renderscript provides the best performance while preserving portability at the cost of manual memory management and difficult library extensibility. The solution presented in this work does not have these limitations since the Renderscript code and the supporting files are automatically generated. Moreover, we support also Filterscript and OpenCL as back ends, which allows to map program parts to the GPU and CPU.

GMAC [11] provides similar memory management abstractions for discrete GPUs (CUDA, OpenCL), but do not consider HSA platforms. However, the proposed abstractions can also be integrated into GMAC.

Table II: Lines of code for the OpenCV implementation, HIPA^{cc} DSL code and generated Renderscript code.

	Sobel	**Gaussian**	**Laplase**	**FIR**	**Harris**
OpenCV	1681	1641	1712	982	2247
HIPA^{cc} DSL	15	22	11	11	68
Renderscript	1915	1951	8575	3680	4265

Conclusion

A code generator model has been presented for Renderscript and Filterscript in the domain of image processing by utilizing the HIPA^{cc} DSL. All code variants should be automatically generated from a common description that is highly portable and performs well on both: CPU and GPU. The code generator has to be integrated into the HIPA^{cc} framework and is available as open-source under http://hipacc-lang.org.

Using this code generator, we'll try to produce the efficient code in our next research for different parallel programming models on embedded devices. We also predict to perform better than the target-optimized OpenCV library on the CPU.

References

1. R. Membarth, F. Hannig, J. Teich, M. Körner, W. Eckert, "Generating Device-specific GPU Code for Local Operators in Medical Imaging", in International Parallel & Distributed Processing Symposium (IPDPS), IEEE, May 2012, pp. 569–581.
2. ARM. (2013). Mali-T600 Series GPU OpenCL – Developer Guide.
3. B. A. Hechtman, D. J. Sorin, "Evaluating Cache Coherent Shared Virtual Memory for Heterogeneous Multicore Chips", in International Symposium on Performance Analysis of Systems and Software (ISPASS), Jun. 2013.
4. Arndale Board, Samsung Exynos 5 Dual Arndale Board, http: //www.arndaleboard.org, 2012–2013.
5. C. Harris and M. Stephens, "A Combined Corner and Edge Detector", in Alvey Vision Conference, 1988, pp. 147–151.
6. The Portland Group, PGI OpenCL Compiler for ARM, http://www.pgroup.com/products/pgcl.htm, 2011–2013.
7. J. Ragan-Kelley, A. Adams, S. Paris, M. Levoy, S. Amarasinghe, F. Durand, "Decoupling Algorithms from Schedules for Easy Optimization of Image Processing Pipelines", ACM Transactions on Graphics (TOG), vol. 31, no. 4, 32:1–32:12, Jul. 2012.
8. Willow Garage, Open Source Computer Vision (OpenCV), http://opencv.willowgarage.com/wiki, 1999–2013.
9. Acosta, F. Almeida, "Towards a Unified Heterogeneous Development Model in Android", in International Workshop on Algorithms, Models and Tools for Parallel Computing on Heterogeneous Platforms (HeteroPar), Springer, Aug. 2013.
10. X. Qian, G. Zhu, X.-F. Li, "Comparison and Analysis of the Three Programming Models in Google Android", in Asia-Pacific Programming Languages and CompilersWorkshop (APPLC), ACM, Beijing, China, Jun. 2012, pp. 1–9.
11. Gelado, J. E. Stone, J. Cabezas, S. Patel, N. Navarro, W.-M.W. Hwu. "An Asymmetric Distributed Shared Memory Model for Heterogeneous Parallel Systems", in International Conference on Architectural Support for Programming Languages and Operating Systems (ASPLOS), ACM, Mar. 2010, pp. 347–358.