

# FREE RIDER: A Tool for Retargeting Platform-Specific Intrinsic Functions



# FREE RIDER: A Tool for Retargeting Platform-Specific Intrinsic Functions

Stanislav Manilov Björn Franke
Institute for Computing Systems Architecture
School of Informatics, University of Edinburgh
Informatics Forum, 10 Crichton Street, Edinburgh, EH8 9AB, United Kingdom
s.z.manilov@sms.ed.ac.uk, bfranke@inf.ed.ac.uk

Anthony Magrath Cedric Andrieu
Cirrus Logic International (UK) Ltd.
Cirrus Logic Inc.
Westfield House, 26 Westfield Road, Edinburgh, EH11 2QB, United Kingdom
anthony.magrath@cirrus.com
cedric.andrieu@cirrus.com

## Abstract

Short-vector SIMD and DSP instructions are popular extensions to common ISAs. These extensions deliver excellent performance and compact code for some compute-intensive applications, but they require specialised compiler support. To enable the programmer to explicitly request the use of such an instruction, many C compilers provide platform-specific intrinsic functions, whose implementation is handled specially by the compiler. The use of such intrinsics, however, inevitably results in non-portable code. In this paper we develop a novel methodology for retargeting such non-portable code, which maps intrinsics from one platform to another, taking advantage of similar intrinsics on the target platform. We employ a description language to specify the signature and semantics of intrinsics and perform graph-based pattern matching and high-level code transformations to derive optimised implementations exploiting the target's intrinsics, wherever possible. We demonstrate the effectiveness of our new methodology, implemented in the FREE RIDER tool, by automatically retargeting benchmarks derived from OPENCV samples and a complex embedded application optimised to run on an ARM CORTEX-M4 to an INTEL EDISON module with SSE4.2 instructions. We achieve a speedup of up to 3.73 over a plain C baseline, and on average 96.0% of the speedup of manually ported and optimised versions of the benchmarks.

***Categories and Subject Descriptors*** D.3.4 [*Programming Languages*]: Processors—Code generation, run-time environments, optimization

***General Terms*** Languages, Performance

***Keywords*** Retargeting, intrinsics, compiler-known functions, graph pattern matching

## 1. Introduction

Instruction set extensions are computer architects' favourite choice of weapon to add domain-specific acceleration to processor cores with their mature, proven software and hardware eco-systems, respectively. For example, INTEL have devised various streaming SIMD extensions (first MMX, then SSE to SSE4) to speed up graphics and digital signal processing. Similar capabilities are offered by the ALTIVEC floating point and integer SIMD extensions designed by APPLE, IBM and FREESCALE SEMICONDUCTOR. In the embedded space, ARM offers DSP and multimedia support through their SIMD extensions for multimedia and NEON extensions. Whilst conceptually similar, these different instruction set extensions differ significantly in detail, e.g. in their word and subword size, supported data types, and use of processor registers.

Despite improvements in compiler technology, including automatic vectorisation [14, 15], short-vector instructions offered by the architecture are typically accessed through platform-specific compiler *built-in functions*. This is due to the superior performance of hand-tuned vector code, which often outperforms auto-vectorised code [12]. Built-in functions, also called *intrinsics*, are functions available for use in C, but their implementation is handled specially in the compiler: the original intrinsic call is directly substituted by a machine instruction. For example, MICROSOFT and INTEL's C/C++ compilers as well as GCC and LLVM implement intrinsics that map directly to the X86 SIMD instructions. The use of intrinsics enables programmers to exploit the underlying instruction set extensions and to increase the efficiency of their programs, but their use inevitably results in non-portable code. Obviously, this seriously restricts the re-use and porting of software components such as libraries, which have been heavily optimised for one particular instruction set extension and where no plain C sources are available.

In this paper we develop a novel technique for cross-platform retargeting of code comprising platform-specific intrinsics. The **key idea** is to accept the presence of intrinsics as an opportunity and a source of information, rather than an obstacle. We develop a graph based matching approach, which aims at substituting existing intrinsics with those available on the target machine and possibly additional code providing compatibility. We provide descriptions of intrinsics for a number of different instruction set extensions using a custom description language, covering the syntactic and semantic specification of intrinsics. These descriptions are translated to graph representations by our FREE RIDER tool, which then translates any C program written using one set of intrinsics (e.g. those for an ARM CORTEX-M4 core) to make use of intrinsics of any other platform (e.g. INTEL SSE). Any pair of the available architectures can be used, in either direction. This translation process might also include additional source code transformations such as loop unrolling to account for different SIMD word sizes of the source and target platforms, respectively.



*LCTES'15* June 18 - 19, 2015, Portland, OR, USA
© ACM. ISBN 978-1-4503-3257-6//...$15.00
DOI: http://dx.doi.org/10.1145/2670529.2754962

```
1  char A[128], B[128], C[128];
2
3  // ... initialize A and B ...
4
5  // Packed vector access
6  #define PV(x) (*((uint32_t*)(&x)))
7
8  // Compute loop with ARM UADD8 intrinsic
9  for (int i = 0; i < 128; i+=4) {
10    PV(C[i]) = __UADD8(PV(A[i]), PV(B[i]));
11 }
12
```

(a) Platform-specific code using ARM UADD8 intrinsic.

```
1  char A[128], B[128], C[128];
2
3  // ... initialize A and B ...
4
5  ...
6  ...
7
8  // Compute loop
9  for (int i = 0; i < 128; ++i) {
10        C[i] = A[i] + B[i];
11 }
12
```

(b) Portable, but **unavailable** plain-C implementation.

```
1  char A[128], B[128], C[128];
2
3  // ... initialize A and B ...
4
5  // Compute loop with Intel SSE intrinsic
6  for (int i = 0; i < 128; i+=16) {
7        SV(C[i], _mm_add_epi8(LV(A[i]), LV(B[i])));
8  }
9
```

(c) Platform-specific code using INTEL _mm_add_epi8 intrinsic.

```
1
2  // Load vector
3  #define LV(x)     (_mm_loadu_si128 ((__m128i*)(&x)))
4
5  // Store vector
6  #define SV(x, y) (_mm_storeu_si128 ((__m128i*)(&x),y))
7
8
9
```

(d) Auxiliary load/store macros for INTEL SSE vectors.

Figure 1: Motivating example illustrating the use of the intrinsics to speed up a vector addition loop. The code in Figure 1a is optimised for an ARM CORTEX-M4. This code makes use of the ARM-specific UADD8 intrinsic and **will not compile** for e.g. an INTEL platform. Equivalent plain-C code as shown in Figure 1b is often **not available**. Figure 1c shows the vector addition loop from Figure 1a **translated** to an INTEL platform, now using the INTEL _mm_add_epi8 SSE intrinsic. This translation requires not only substitution of the ARM intrinsic, but additional code transformations. These comprise the introduction of suitable **short vector accesses** (Figure 1d), further **loop unrolling** to match the wider SIMD word size of the INTEL architecture and **dead store elimination** of redundant flag setting operations implicitly contained in the original ARM UADD8 intrinsic, which are not used in this example, but need to be emulated where required.

### 1.1 Motivating Example

Consider the example in Figure 1, which illustrates the steps involved in translating a vector addition loop using intrinsics for an ARM CORTEX-M4 to an INTEL SSE-enabled processor.

In Figure 1a ARM-optimised code is shown, which exploits the UADD8 intrinsic available on the CORTEX-M4 platform and which provides convenient C-level access to a quad 8-bit unsigned addition instruction implemented in the processor's ISA. Using the UADD8 intrinsic four pairs of one-byte values are added using a single processor instruction (line 10). To account for this implicit loop unrolling the surrounding loop is incremented in steps of four (in line 9), whilst also enabling 32-bit data accesses (rather than four individual 8-bit accesses). This is achieved by the access macro PV, which performs the necessary 32-bit cast operation. The measurable benefit of using the UADD8 intrinsic in Figure 1a is a speedup of about four over a plain C implementation such as shown in Figure 1a (on a FREESCALE KINETIS K70 implementation of the ARM CORTEX-M4 core). However, higher performance for the platform-specific code comes at a price – the code in Figure 1a is **not portable** and does not work on platforms other than the ARM CORTEX-M4.

Porting of the code in Figure 1a to another platform is hindered by the fact that a plain C version such as shown in Figure 1b is often **not available**. In this situation, the user could (a) **manually derive** the plain C implementation and then try to vectorise this code, either manually or using an auto-vectoriser, or (b) use our FREE RIDER tool and methodology for **automatic retargeting**.

Now consider the automatically retargeted code, optimised for an INTEL processor with SSE extensions, in Figure 1c. It exploits the _mm_add_epi8 intrinsic, which provides access to an 8-bit addition instruction that operates on two groups of sixteen elements. Using _mm_add_epi8 intrinsic sixteen pairs of one-byte values are added in a single processor instruction (line 7). Accordingly, the loop increment has been adjusted to sixteen (line 6), and 128-bit data accesses are provided by the access macros LV and SV, shown in Figure 1d. Without user intervention platform-specific ARM code has been retargeted to an INTEL platform whilst **retaining the performance benefit** of the original ARM intrinsic. Compared to a plain C baseline (such as the one in Figure 1b) the code in Figure 1c is about ten times faster[1].

FREE RIDER does not require a plain C implementation such as the one in Figure 1b, but directly retargets platform-specific code *where no plain C implementation exists*. Translation of intrinsics involves a number of processing steps briefly outlined in Figure 2. We start with the code in Figure 1a, but we do not have access to a plain C implementation such as the one shown in Figure 1b. As a first step of the transformation process the UADD8 intrinsic is expanded in the internal representation of FREE RIDER – it is essentially expressed as a vector of four additions followed by a vector of four compare-and-set operations. This is shown in Figure 2b and follows closely the specification of the UADD8 intrinsic from Figure 2a. The next step is to analyse which output of the intrinsic is actually used by the program. In the case of the motivating example the result of the addition is later used (for outputting the result, further computations, etc.), but the APSR register is never read. This

[1] Memory access overheads prevent the speedup to reach its ideal value of sixteen.

The __uadd8 intrinsic returns:

- the addition of the first bytes in each operand, in the first byte of the return value
- the addition of the second bytes in each operand, in the second byte of the return value
- the addition of the third bytes in each operand, in the third byte of the return value
- the addition of the fourth bytes in each operand, in the fourth byte of the return value.

Each bit in APSR.GE is set or cleared for each byte in the return value, depending on the results of the operation. If res is the return value, then:

- if res[7:0] $\geq$ 0x100 then APSR.GE[0] = 1 else 0
- if res[15:8] $\geq$ 0x100 then APSR.GE[1] = 1 else 0
- if res[23:16] $\geq$ 0x100 then APSR.GE[2] = 1 else 0
- if res[31:24] $\geq$ 0x100 then APSR.GE[3] = 1 else 0.

(a) Specification of the ARM CORTEX-M4 __uadd8 intrinsic according to [1].

```
// Compute loop, abstract vector form
for (int i = 0; i < 128; i+=4)
{
    vector_4 [j = 0 .. 3]{
        C[i + j] = A[i + j] + B[i + j];
        APSR.GE[j] = (C[i + j] > 0x100) ? 1 : 0
    }
}
```

(b) Code in abstract vector representation.

```
// Reduced abstract representation
for (int i = 0; i < 128; i+= 16) {
    (C[i] = A[i] + B[i]) x 16
}
```

(c) Unnecessary writes to APSR removed. Unroll factor increased.

```
// Compute loop with Intel SSE intrinsic
for (int i = 0; i < 128; i+=16) {
        SV(C[i], _mm_add_epi8(LV(A[i]), LV(B[i])));
}
```

(d) Equivalent INTEL SSE code.

Figure 2: Motivating example illustrating the transformation from the use of the UADD8 intrinsic on the ARM CORTEX-M4 core (in Figure 2a) to the use of the INTEL _mm_add_epi8 SSE intrinsic (in Figure 2d). The transformation requires not only substitution of the ARM intrinsic, but additional code transformations, which comprise suitable short vector accesses and further loop unrolling to match the wider SIMD word size of the INTEL architecture. In addition, the overflow checking logic is removed, as the APSR register is not read later in the program, making the writes to it unnecessary.

register exists in the ARM CORTEX-M4 core to set flags indicating different program status - zero result, negative result, overflow (as is the case of UADD8), and others. Since the register is not read, writing to it is a waste of processing resources, so the compare-and-set operations are removed altogether. Another operation performed at this step is to find appropriate target SIMD intrinsic that consists of the remaining core operation – addition. In the case of INTEL SSE this is the _mm_add_epi8 intrinsic, which has internal representation (c = a + b) x 16 - it is a vector of sixteen additions. Since the widths of the vector operations do not match, the loop is unrolled to fit an _mm_add_epi8 operation. This step is shown in Figure 2c. Finally, when the resulting abstract representation matches exactly a target instruction it is replaced by that instruction together with appropriate access macros. The resulting code after retargeting, shown in Figure 2d, matches the INTEL SSE implementation from Figure 1d.

We have implemented this methodology in the FREE RIDER tool and demonstrate its effectiveness using a set of compute-intensive OPENCV computer vision benchmarks [3]. Automatically retargeting these benchmarks from an ARM NEON platform to an INTEL EDISON module with short-vector SSE4.2 instructions, we achieve on average 96.0% of the performance of manually retargeted and optimised ports. Furthermore, an evaluation against a full-scale robotic application [11], which implements the computer vision component of a high-end autopilot for unmanned aerial vehicles (UAV) delivers a speedup of 3.73 over a plain C baseline, when ported from an ARM CORTEX-M4 platform to INTEL EDISON using our methodology.

### 1.2 Contributions

This paper makes the following contributions:

1. We develop a novel, automated methodology for retargeting C code containing platform-specific intrinsics, whilst making efficient use of those intrinsics offered by the target platform,
2. we combine in our approach high-level descriptions of intrinsics, graph based matching and source-level code transformations to account for differences in the SIMD word sizes between machines, and
3. we evaluate our methodology using compute-intensive OPENCV benchmarks as well as full applications and demonstrate performance levels competitive with manual retargeting efforts.

### 1.3 Overview

The remainder of this paper is structured as follows. In Section 2 we briefly introduce the background on compiler intrinsics, target platforms and applications. In Section 3 we present our methodology for retargeting of platform-specific intrinsics involving a high-level description of intrinsics, a graph-based matching algorithm and source-level code transformations. The results of our evaluation on benchmarks and full applications are presented in Section 4, before we establish the context of related work in Section 5. Finally, in Section 6 we summarise and conclude.

## 2. Background

### 2.1 Target Platforms

The specific target platforms used in our research are the ARM CORTEX-M4 core and INTEL X86 processors with short-vector SSE4.2 instructions (in particular, INTEL EDISON). However, without modification our work applies to any X86 architecture that supports SSE4.2) and ARM NEON enabled processors. By

providing further target descriptions other platforms such as POWERPC/ALTIVEC could be supported, but this is beyond the scope of this paper.

The CORTEX-M4 processor is specifically developed to address digital signal control markets. It is designed so that it has low power consumption while offering high-efficiency signal processing functionality, provided instruction set extensions accessible through ARM specific intrinsics.

ARM NEON is ARM's SIMD extension that targets more computationally demanding tasks, e.g. video processing, voice recognition, or computer graphics rendering. Architectures that support NEON have higher computational performance compared to the CORTEX-M4 processor and are typically found as application processors within mobile devices such as smartphones or tablets.

INTEL X86 on the other hand includes a huge family of processors, from embedded low-power chips to high-end server CPU's offering a one-size-fits-all instruction set. SSE4.2 is an instruction set extension that allows INTEL X86 processors to execute SIMD instructions on vectors up to 128-bit wide. This allows such processors to be used efficiently for multimedia and graphics processing.

SIMD operations, both for ARM and INTEL, are accessible to the C programmer by means of intrinsic functions. An intrinsic function is not explicitly defined by the programmer, but is provided (as a built-in function) by the compiler, which replaces a intrinsic function call with a hard-coded sequence of low-level instructions [2]. Examples for intrinsic functions are the `UADD8` intrinsic for the ARM CORTEX-M4 processor and the `_mm_add_epi8` intrinsic for the SSE instruction set extension, both of which are part of the motivating example (Figure 1).

In general, the operands of CORTEX-M4 SIMD instructions are 32-bit wide fields. Depending on the instruction each operand is treated as a single 32-bit number, two 16-bit numbers, or four 8-bit numbers. The available operations that can be performed range from simple operations like addition, to very specialised operations like the `SMLALDX` instruction which performs dual 16-bit exchange and multiply with addition of products and 64-bit accumulation. A list of the groups of available operations are: addition (13 instructions), subtraction (13 instructions), sum of absolute differences with or without accumulation (2 instructions), halfword multiply with addition or subtraction, with or without exchange, and with or without accumulation (12 instructions), parallel add and subtract halfwords with exchange (12 instructions), sign-extend byte, with or without addition (4 instructions), half-word saturation (2 instructions), status register based selection (1 instruction).

At the same time, the operands of INTEL SSE4.2 instructions are 128-bit wide fields when they signify a vector, or any other type from the C programming language when they signify vector elements, bitmasks, or shift values. The 128-bit wide fields can be treated as vectors of two, four, eight, or sixteen elements of 64, 32, 16, or 8 bit values, respectively, depending on the instruction. The available operations are not as specialised as those of the CORTEX-M4 SIMD processor, but rather resemble standard processor instructions that operate on vectors instead of single elements.

While there are also miscellaneous utility (e.g. cache control) instructions for the INTEL SSE instruction set we primarily target arithmetic instructions. The integer instructions can be grouped in the following categories: addition (8 instructions), subtraction (8 instructions), sum of absolute differences (1 instruction), halfword multiply with addition (1 instruction), multiplication (5 instructions), maximum, minimum and average (6 instructions), shifts and bitwise operations(22 instructions), comparison (9 instructions). The miscellaneous instructions that are of our interest are shuffle instructions and pack/unpack instructions. They can be used to implement more complicated SIMD operations that include exchanging of vector elements.

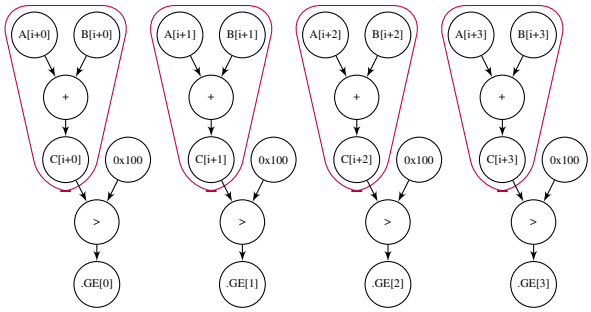


Figure 3: Graph representation of the `__UADD8` intrinsic

Finally, the NEON extension is similar to SSE. Vectors can be either 64-bit or 128-bit wide and can contain signed or unsigned 8-bit, 16-bit, 32-bit, 64-bit integers, or single precision float numbers. The operations that are supported include standard arithmetic and logical operations, comparison operations, memory operations and shuffling operations. The more specialised operations which we do not take into consideration include table lookup and complicated mathematical operations, like reciprocal square-root estimate.

### 2.2 Target Application

There are no readily available standard benchmarks, which make explicit use of intrinsic functions due to the resulting undesirable restriction to a single platform. Therefore, we use a compute-intensive, open-source application extensively used in the academic, hobby and industrial communities. This application is PX4 [11] – a high-end autopilot for unmanned aerial vehicles (UAV) using computer vision algorithms – jointly developed by researchers from the Computer Vision and Geometry, the Autonomous Systems and the Automatic Control Labs at ETH Zurich (Swiss Federal Institute of Technology). PX4 has been developed and optimised for an ARM CORTEX M4F CPU and, in particular, the optical flow module makes extensive use of SIMD intrinsics. Among the most computational intensive functions in the computer vision component are those for the calculation of the Hessian at a pixel location, the average pixel gradient of all horizontal and vertical steps, the SAD distances of subpixel shift of two $8 \times 8$ pixel patterns, the SAD of two $8 \times 8$ pixel windows, and the pixel flow between two images. We have extracted these functions and use them in isolation (to avoid system benchmarking involving the whole UAV) for our empirical evaluation. A single function (*absdiff*) is written entirely using ARM assembly, for which we provide a portable C implementation.

In addition, we use a number of benchmarks extracted from the popular OPENCV computer vision library [3]. This provides with reference implementations, manually ported and optimised by a independent third party, supporting both ARM and INTEL through platform-specific intrinsics. We use these benchmarks to evaluate the performance and capabilities of our FREE RIDER retargeting tool in comparison to a manual effort.

## 3. FREE RIDER Methodology

### 3.1 Overview

The FREE RIDER tool performs of four major transformation steps as shown in the overview diagram in Figure 4: Header generation, data-flow extraction, graph matching, and source-level code transformation.

Initially descriptions of the source and target intrinsics are taken as inputs and emulation C header files (in the style of [19]) are generated. These header files declare and define portable C inline

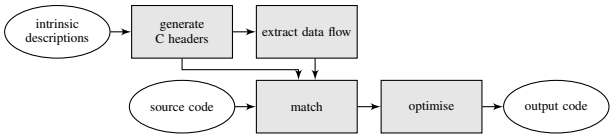


Figure 4: Stages of execution of the FREE RIDER tool

functions for the intrinsics of the source platform. We show in Section 4 that the use of these "emulated" intrinsics results in portable code, but yields a low performance level of only 70% of a plain C implementation of the corresponding functionality on the target platform. This means that emulation of intrinsics through inline C functions provides compatibility, but results in a performance penalty.

In a second step the header files are used as input to the next stage, in which we generate data flow graphs for each intrinsic (see Figure 3 for example). These graphs, annotated with the types of inputs and outputs, serve as intermediate representation. Nodes of the graphs are also annotated with the operations performed, for example vector addition or vector sum reduction. We will use these data flow graphs for graph based pattern matching.

In the next step the C header files, the data flow graphs, and the source code of the program under consideration are all fed to the matching stage of FREE RIDER. It employs a greedy subgraph isomorphism algorithm (similar to [10]) to match the data flow graph of each intrinsic encountered in the source code with data flow graphs of target intrinsics. The graphs of two target intrinsics can connect into a single graph, by connecting the output nodes of one of them to the input nodes of the other by an assignment edge. In this way multiple target intrinsics can cover completely or partially a source intrinsic. For source intrinsics, which can only be partially covered by target intrinsics, scalar C code is generated for the remaining, non-covered parts of the data flow graphs. The output of this stage is C code of the target application with its source intrinsics partly or fully replaced with those of the target platform, wherever possible, and additional plain C code where a direct match is not possible.

Finally, the resulting code after substituting source intrinsics with target intrinsics is further optimised. Checks are performed to remove dead computations and variables (e.g. introduced as part of the flag setting operations in ARM intrinsics, see also Figure 5). Additionally, loop unrolling might be performed to adjust the possibly different SIMD word sizes of the two platforms (also shown in Figure 5).

### 3.2 Description of Intrinsics

Intrinsics are described by the user in a high-level, human readable format. The description comprises the following items: name of native platform, list of operand names and types, output name and type, and the behaviour (a snippet of restricted C code). An abbreviated example of such a description is provided in figure 6, which shows the specification of the ARM `UADD8` intrinsic.

Operand and result types can be standard C types, or vector types which should also be described in a format comprising the name of the native platform, total size in bits, and the type of a single element (atom type). While this allows for nested types of vectors of vectors, this feature is not used as there are no instructions that operate on such complex types. Thus, the atom type is a standard C type.

Behaviours of intrinsics, i.e. semantic actions, are expressed in a restricted subset of the C language. During generation of header files behaviours are used as the function body of the generated inline function for the intrinsic.

```
define intrinsic UADD8
{
    platform ARM_CORTEX_M4
    operands val0:uint8x4_t@32, \
             val1:uint8x4_t@32
    result   res:uint8x4_t@32
    behaviour {
        uint8x4_t res;

        // Load data and cast to prepare for
        // main operation
        uint16_t a0 =
            (uint16_t)UINT8X4_T_READ(val0,0);
        uint16_t a1 =
            (uint16_t)UINT8X4_T_READ(val0,1);
        uint16_t a2 =
            (uint16_t)UINT8X4_T_READ(val0,2);
        uint16_t a3 =
            (uint16_t)UINT8X4_T_READ(val0,3);

        uint16_t b0 =
            (uint16_t)UINT8X4_T_READ(val1,0);
        uint16_t b1 =
            (uint16_t)UINT8X4_T_READ(val1,1);
        uint16_t b2 =
            (uint16_t)UINT8X4_T_READ(val1,2);
        uint16_t b3 =
            (uint16_t)UINT8X4_T_READ(val1,3);

        // Perform additions
        // Need 16-bit intermediate results
        // to determine overflow flags
        uint16_t c0 = a0 + b0;
        uint16_t c1 = a1 + b1;
        uint16_t c2 = a2 + b2;
        uint16_t c3 = a3 + b3;

        // Assign results, casting to 8-bit
        UINT8X4_T_WRITE(res,0,(uint8_t)c0);
        UINT8X4_T_WRITE(res,1,(uint8_t)c1);
        UINT8X4_T_WRITE(res,2,(uint8_t)c2);
        UINT8X4_T_WRITE(res,3,(uint8_t)c3);

        // Flag setting, depending on
        // 16-bit intermediate results
        if (c0 >= 0x100) APSR_GE_SET(0);
        else APSR_GE_RESET(0);
        if (c1 >= 0x100) APSR_GE_SET(1);
        else APSR_GE_RESET(1);
        if (c2 >= 0x100) APSR_GE_SET(2);
        else APSR_GE_RESET(2);
        if (c3 >= 0x100) APSR_GE_SET(3);
        else APSR_GE_RESET(3);

        // Return result
        return res;
    }
}
```

Figure 6: Example showing the high-level description of the ARM `UADD8` intrinsic.

Finally, platform-specific special registers can be described if they are used as part of the side effects of an intrinsic function execution. An example of such register is the ARM CORTEX-M4 `APSR` (Application Program Status Register), which is used e.g. for indicating arithmetic overflow.

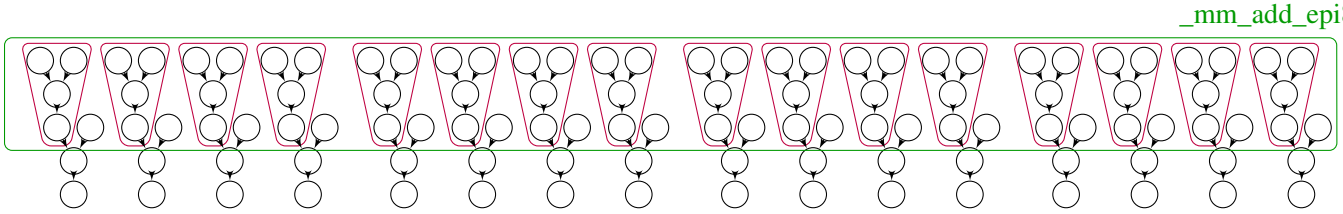


Figure 5: Matching of four __UADD8 intrinsics (in red) and one _mm_add_epi8 intrinsic (in green) resulting from subgraph isomorphism detection, loop unrolling and dead variable/code elimination. Redundant flag setting computations have no counterpart in SEE and either require additional scalar C code or can be eliminated if there are no further uses of the flags (see also Figure 3).

### 3.3 Generation of C Header Files

After the intrinsic descriptions are provided they are used to generate one C header file per platform. Definitions of the custom types are output first together with macro functions to read and write separate elements of a vector. Then, special registers are implemented as bit field structures and access macros for them are generated.

After this supporting code has been created the implementation of the intrinsic functions as inline C functions is generated. Signatures are generated using the type information for the operands and the result, and the body of the functions are copied from the behaviour descriptions.

Using the generated header files, a data flow graph is derived for each intrinsic using standard data flow analysis techniques. These data flow graphs, together with the input/output type information of each intrinsic are used in the matching stage as descriptions of the intrinsics.

### 3.4 Graph Matching and Source-level Transformation

The process of matching intrinsics is outlined in Figure 7. Given the data flow graph of a source intrinsic, an intrinsic from the target platform is searched for using the greedy (sub)graph isomorphism algorithm described in [5]. The overhead that is added to the compilation time by using it is unmeasurably small for all our test cases. The authors of the algorithm evaluate that it can match graphs up to 2000 nodes in under a tenth of a second. Since the graphs that we use to represent the intrinsics are quite small in comparison (under 20 nodes for the most complicated intrinsics) we are not concerned about a potential added overhead.

When a structural and operational match is found, the type information of the found operation is compared with the type information at the source location of the match. If the vector widths of the operands and the result match, the matching part of the graph is directly replaced with the found intrinsic and the process is repeated until no further unmatched parts of the source dataflow graph can be found or there are no more target operations that can cover the remaining graph.

There are two reasons for possible mismatches of vector widths: (a) The target intrinsic is too narrow (i.e. it contains fewer operands than the corresponding source intrinsic), or (b) it is too wide (i.e. it contains more operands than the source intrinsic). In the first case, the target intrinsic can be invoked as many times as it takes to match the width of the source vector (e.g. using four 4-element additions to implement one 16-element addition). In the second case, loop unrolling is required in order to enable a match (e.g. unroll a loop containing a 4-element addition in order to use a 16-element addition to implement four 4-element additions from four iterations).

In case loop unrolling is required it is performed alongside further data alignment. The latter might be necessary if arrays are not accessed in order, but some elements are skipped over. If either the unrolling or the data alignment steps fails, the matching fails

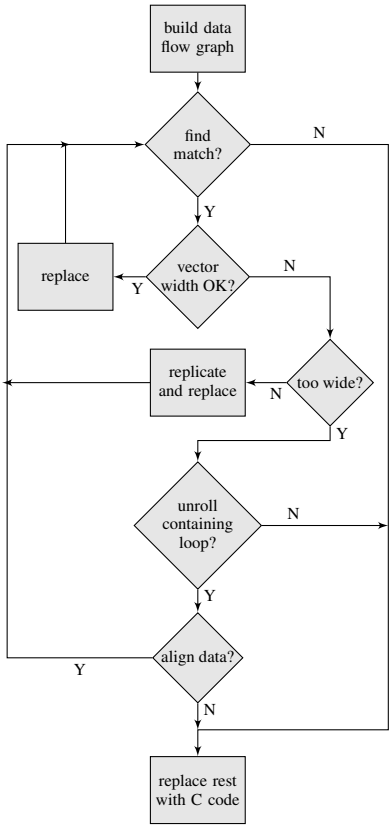


Figure 7: High-level algorithm for matching source and target intrinsics including loop unrolling for adjusting different SIMD word sizes and aligmnents.

and the default replacement with plain C code is performed to ensure correctness of the result. However, if they succeed the whole process is repeated again, until no further matches can be found.

Substitution of intrinsics as well as optimisation (loop unrolling, alignment, dead code/variable elimination) are implemented as source-level transformations. This means that C code enhanced with target intrinsics is generated, which can be compiled with the standard compiler for the target platform.

### 3.5 Limitations

As described in Section 2, intrinsics available for one platform cannot always be expressed by intrinsics available on another platform, or even in standard C code. Examples of such intrinsics are cacheability and synchronisation operations. We limit our approach to

| Benchmark | Summary |
|---|---|
| calib | Calibrates a camera given a sequence of images. |
| bgfg | Split of background and foreground of video. |
| edge | Canny edge detection on an image. |
| align | Automatic alignment of an image. |
| polar | Polar transformation on a video. |
| segm | Automatic segmentation of objects in a video. |
| stitch | Stitching multiple images into a mosaic. |
| vstab | Automatic stabilisation of a video. |

Table 1: OPENCV applications used as benchmarks.

standard data processing operations and do not consider complex intrinsics whose behaviours cannot be expressed in C.

## 4. Empirical Evaluation

### 4.1 Evaluation Methodology

For our evaluation we have applied the FREE RIDER methodology to automatically retarget ARM-specific benchmarks and applications to an INTEL SSE enabled platform. The system used for performance evaluation is an INTEL EDISON module running at 500 MHz. The available physical memory is 1GB and the operating system is YOCTO LINUX, kernel version 3.10.17. All the benchmarks run on a single processor core.

Performance is measured by using the UNIX program `time` to retrieve the total execution time of each benchmark. This is repeated up to 100 times and the reported times are recorded in a log. This log is then analysed to verify that the values roughly fit in a normal distribution, supporting the model that the difference in the measurements is just gaussian white noise. The average of all runtimes per benchmark is considered to be the representative runtime for that benchmark. Error bounds are not included, as they are too small to plot (less than 0.5 % for all benchmarks).

The OPENCV benchmarks are selected from the default sample programs that are provided with the OPENCV library, version `3.0.0-beta`. We have prepared them by removing the user interaction and substituting it with command line arguments and `stdout` messages. Our eight benchmark programs each contain a significant part ($>$ 10% of CPU time) executing vectorisable code. This was computed by compiling OPENCV with and without the included manual vector optimisations and comparing the runtimes of each benchmark for the two cases. Each of the benchmarks makes heavy use of functions provided by the OPENCV library. Many of these OPENCV functions have been manually ported and optimised for different target architectures, including ARM and INTEL. For our evaluation we take the ARM ports of these functions, automatically retarget them to INTEL and then evaluate the performance of these automatically retargeted codes in comparison to the manual INTEL port provided with OPENCV.

Table 1 provides descriptions of the benchmarks. All of these benchmarks are real-world examples of programs from the computer vision domain.

The SSE intrinsics that we implemented include 14 arithmetic operations, 15 logical and comparison operations, 16 memory and initialisation operations, 4 conversion operations, and 8 shuffling operations. These correspond roughly to the NEON intrinsics that we implemented which include 21 arithmetic operations, 13 logical and comparison operations, 19 memory and initialisation operations, 12 conversion operations, and 8 shuffling operations. The greatest discrepancy is in the amount of conversion operations. There are more NEON conversion operations because NEON allows for two vector widths (64- and 128-bit), and can convert between them, adding 8 operations.

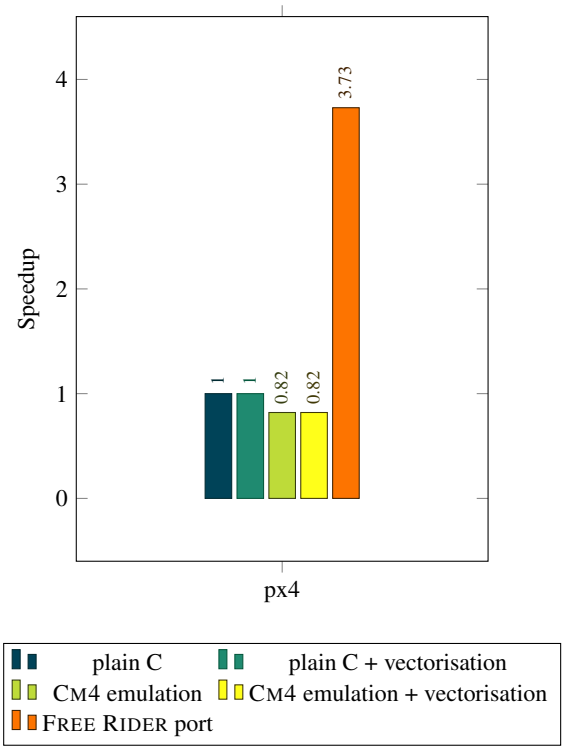


Figure 9: Relative performance of the ported PX4 application on the INTEL platform relative to a plain C baseline. Compiler vectorisation is not effective as it fails to detect and exploit vectorisation opportunities. Emulation of ARM intrinsics through inline functions, implemented in plain C, eases portability, but degrades performance with and without compiler vectorisation efforts. The automatically generated FREE RIDER port is capable of exploiting SIMD parallelism on the INTEL platform and delivers an almost four-fold speedup over the plain C baseline.

The arithmetic operations comprise different versions of additions, subtractions, and multiplications, in addition to a single operation for division, maximum, minimum, and square root. The logical operations comprise different versions of logical ands, ors, xors, and shifts, whereas the comparison operations are comparisons for equality and strict inequalities.

The implemented memory operations comprise different loads and stores, whereas the initialisation operations are generating vectors, all depending on the data type of the given argument. The conversion operations convert the elements of the vector between different datatypes. Finally, the shuffling operations comprise instructions that reorder the elements of a vector in different ways.

On the target `x86` system we use the CLANG/LLVM compiler to produce executable binaries. For comparison, we also have conducted an experiment where plain C sources are presented to the compiler for auto-vectorisation of the PX4 application.

### 4.2 Benchmark Performance Results

Figure 8 shows the results from running the automatically ported OPENCV ARM benchmarks on the INTEL evaluation system. While "native" SSE code delivers a speedup of 1.26 over a plain C baseline, the FREE RIDER ports approach this performance delivering a speedup of 1.21. On average, FREE RIDER produced code that delivers a performance of about 96% of manually ported and optimised INTEL SSE implementation. For every single benchmark

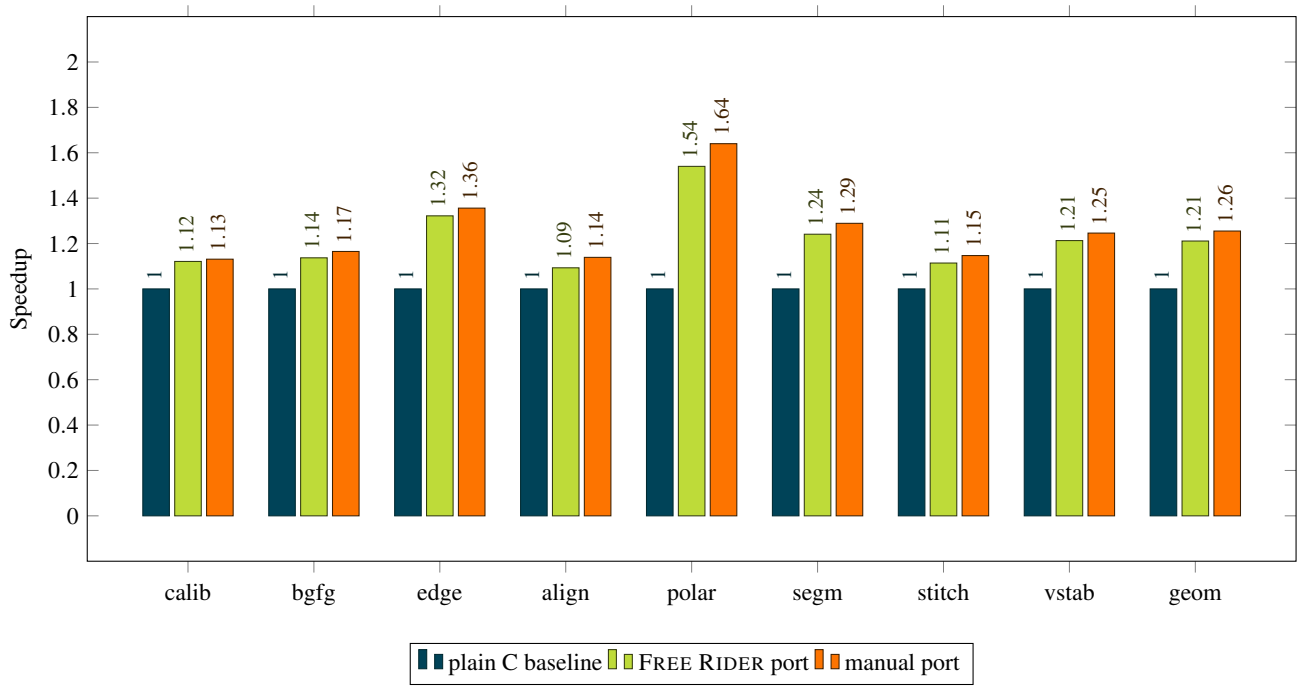


Figure 8: OPENCV applications have been ported automatically from ARM NEON to INTEL SSE. Each bar reports the speedup over a plain C baseline. In each case, the automatically retargeted version produced by FREE RIDER outperforms the plain C baseline. In fact, the performance of the auto-generated ports approaches that of the manually ported OPENCV applications, tuned extensively by OPENCV community developers. On average, we achieve 96% of the speedup of a manual port, however, without the cost involved in manual code rewriting.

the FREE RIDER port outperforms its plain C baseline, despite attempts of the compiler to auto-vectorise this plain C code. Even for the worst performing benchmark (`polar`) the automatically retargeted implementation outperforms the plain C baseline and delivers a speedup just about 6% lower than that of the manual port.

Closer inspection of the generated code revealed that the remaining performance differences between the manual and the auto-generated ports are mainly due to additional code restructuring performed by the expert programmers and optimisations to the scalar code surrounding the intrinsics.

### 4.3 Application Performance Results

Next we focus on porting a larger application – the PX4 computer vision system – from ARM CORTEX-M4 to INTEL SSE. Figure 9 shows a set of results from running the target application in different configurations on the INTEL evaluation system.

The first configuration is a plain C baseline derived from the application's original sources. All further results relative to this baseline configuration. As shown in Figure 9 the compiler fails to automatically vectorise this application, hence performance levels with and without compiler vectorisation are the same.

If ARM intrinsics are emulated by inline C functions, following an approach outlined in [19], performance suffers resulting in a drop of execution speed of about 18%. Again, the LLVM compiler's auto-vectoriser fails to exploit any vectorisation opportunities.

In contrast, the port produced by FREE RIDER substantially outperforms the baseline implementation. The reason for this is that even though FREE RIDER fails to exploit some optimisation opportunities due to irregularity of data accesses, it manages to vectorise the code in the most critical loop of the program and achieve a 3.73 performance speedup. It clearly demonstrates that FREE RIDER is capable of exploiting the source platform's intrinsics and mapping them onto corresponding intrinsics of the target platform. For the INTEL target platform used in this study this is close to the ideal four-fold speedup attainable using four-way SIMD processing.

### 4.4 Coverage and Frequency of Intrinsics

Tables 2 and 3 list those intrinsics that were encountered in translation process of the benchmarks (ARM NEON to INTEL SSE) and the larger application (ARM CORTEX-M4 to INTEL SSE). In addition, for each intrinsic we list its frequency of occurrence in the benchmark sources. The first part of each table lists the intrinsics involved in the translation of the target application, while the second part lists the intrinsics involved in the translation of the benchmarks.

Some of the lines in the table represent multiple intrinsics, which is indicated by the names ending in an asterisk. An example is the `_mm_add*` line from the SSE table, which represents four intrinsics: `_mm_add_epi16`, `_mm_add_epi32`, `_mm_add_ps`, and `_mm_adds_epi16`. Also, some of the lines represent a group of instructions separated by a forward slash, for example `vceq*/vcgt*/vcge*/vlt*/vle*`.

We note a couple of interesting observations in Tables 2 and 3. Firstly, there are a lot more occurrences of INTEL intrinsics when compared to ARM intrinsics. This is because we were targeting INTEL SSE as a destination platform during our experiments and a single source intrinsic might be mapped to one or more destination intrinsics, so we end up with more target intrinsics than source intrinsics after the translation. This is more pronounced in the translation between ARM CORTEX-M4 and INTEL SSE compared to the translation between ARM NEON and INTEL SSE since the first pair of platforms is more dissimilar (as explained in Section 2.1) than the second. Secondly, the tables show that FREE RIDER is capa-

| Intrinsic | Frequency |
|---|---|
| __USADA8 | 36 |
| __UHADD8 | 24 |
| __UADD8 | 3 |
| __USAD8 | 2 |
| vld* | 50 |
| vadd* | 38 |
| vsh* | 35 |
| vdupq_n_* | 27 |
| vget_* | 26 |
| vmov* | 26 |
| vst* | 24 |
| vqmov* | 22 |
| vand* | 16 |
| vmul* | 16 |
| vcombine_* | 15 |
| vceq*/vcgt*/vcge*/vlt*/vle* | 8 |
| vsub* | 7 |
| vmin*/vmax*/vabs* | 5 |
| vqdmulhq_s16 | 4 |
| vmla* | 4 |
| vbslq_f32 | 3 |
| vzip_s16 | 2 |
| vsqrt* | 2 |

Table 2: Frequency of occurrence of ARM CORTEX-M4 and ARM NEON intrinsics in our benchmarks and application. Each line represents a single or multiple intrinsics. In the latter case the name ends with an asterisk to indicate that there are different variants available. The number of multiple intrinsics per line varies from 2 to 5.

| Intrinsic | Frequency |
|---|---|
| _mm_srli_epi32 | 40 |
| _mm_add_epi16 | 40 |
| _mm_loadu_si128 | 34 |
| _mm_store_si128 | 27 |
| _mm_and_si128 | 24 |
| _mm_set1_epi32 | 22 |
| _mm_sub_epi16 | 20 |
| _mm_abs_epi16 | 20 |
| _mm_unpackhi_epi8 | 20 |
| _mm_unpacklo_epi8 | 20 |
| _mm_add_epi8 | 7 |
| _mm_or_si128 | 6 |
| _mm_load* | 66 |
| _mm_sll/sra/srl* | 51 |
| _mm_pack/unpack* | 43 |
| _mm_store* | 40 |
| _mm_add* | 39 |
| _mm_mul* | 30 |
| _mm_set* | 27 |
| _mm_and* | 17 |
| _mm_xor* | 17 |
| _mm_cmp* | 16 |
| _mm_cvt* | 13 |
| _mm_sub* | 11 |
| _mm_andnot_si128 | 7 |
| _mm_sqrt_ps* | 2 |
| _mm_min/max_ps | 2 |
| _mm_div* | 1 |
| _mm_or_si128 | 1 |
| _mm_movemask_epi8 | 1 |

Table 3: Frequency of occurrence of INTEL SSE intrinsics in the retargeted benchmarks and application. Asterisks are used similarly to Table 2. The first part of the table summarises statistics for the automatic translation of the target application, whereas the second part summarises statistics for the benchmarks.

ble to cover wide range of intrinsics, which enable us to automatically process not only isolated benchmarks, but complex real-world applications resulting in performance levels approaching those of manual retargeting and optimisation.

## 5. Related Work

Handling of intrinsic functions by the compiler has found little attention in the academic community, possibly due to their normally straight-forward, but target- and compiler-specific implementation. Among the few publications dealing with various aspects related to intrinsic functions are the following.

Compilation for multimedia instructions has been an active research area for over a decade [6, 7, 9, 16–18]. Krall and Lelait [9] describe basic compilation techniques for multimedia processors. They compare classical vectorisation, borrowed from the age of the vector supercomputers, to using loop unrolling for vectorisation. The mentioned classical vectorisation employs a dependency analysis and might fail if the operations within the loop are not vectorisable. Loop unrolling is more likely to succeed, as the operations of consecutive loop iterations are the same – thus vectorisable. The only reason for failure there might be loop carried dependencies. The authors also explore the problem of unaligned memory accesses. FREERIDER allows alignment to be specified in the description of architectures and honours it during translation.

Pokam et al. propose SWARP [16] – a retargetable preprocessor for multimedia instructions that is extendable by the user. Their work allows taking advantage of vector operations, without the programmer specifying that intention in the source code, i.e. the input provided to SWARP is plain C and it generates C code, which uses SIMD extensions. A flexible idiom recognition phase eases the retargeting of the system to new machines without changing SWARP itself. Our approach is different in that we are retargeting platform-specific intrinsics. We leverage the expertise already invested in optimising the application to one platform and try to maintain this information when translating to another platform. The idiom recognition is replaced by our matching phase, and flexibility is achieved by the intrinsic description language, which is used to describe different targets.

Similar approaches, all operating on plain C input and trying to extract superword level parallelism within an optimising or vectorising compiler are described in [6, 7, 17, 18]. Whilst these techniques are useful for the initial identification of vectorisation opportunities in C code, they fail to process applications, which have already been vectorised for a particular platform using intrinsics.

A graph based instruction selection technique has been developed in [13], where the compiler targets automatically generated instruction set extensions, where instruction patterns are not tree shaped, but highly irregular and sometimes larger (up to 12 inputs and 8 outputs) than typical multimedia instructions. The graph pattern matching approach used in this paper is somewhat comparable to that in [13], however, the purpose of our work is to aid the user retargeting an application optimised for a platform other than the current target platform, whereas graph pattern matching is used in [13] to match highly idiosyncratic instructions.

Modelling of instruction semantics in ADL processor descriptions for C compiler retargeting has been presented in [4]. The fo-

cus of this work is more on generating a basic compiler using an architecture description, rather than retargeting of existing, optimised code.

Implementation of intrinsic functions for a DSP compiler is subject of [2]. This paper proposes and implements a new approach to intrinsic functions where the programmer targets a compiler's intermediate representation rather than the assembly language of a particular processor.

A general introduction to intrinsics for vector processing in the GCC compiler is provided in [8].

Possibly most relevant to the work presented in the paper is [19], where a set of hand-coded inline functions compatible with ARM NEON intrinsics is provided for an INTEL platform with SSE. The result is a similar "emulation" layer providing portability for a particular combination of intrinsics (ARM NEON to INTEL SSE), but unlike FREE RIDER this is not automated and retargetable to any platform, but the result of a major manual implementation effort for one specific pair of platforms.

## 6. Summary & Conclusions

In this paper we have developed a new methodology for retargeting platform-specific intrinsics from one platform to another. We use a description language to specify signatures and semantics of intrinsics of both platforms. These descriptions are processed by our FREE RIDER tool, which performs subgraph isomorphism checking to substitute one set of intrinsics with one or more intrinsics of the target platform, plus additional scalar code wherever needed. In addition, FREE RIDER performs source-level loop unrolling in order to account for differences in SIMD word sizes and alignment, and dead variable/code elimination to remove artefacts introduced by the substitution of intrinsics. We have evaluated our methodology by automatically porting OPENCV benchmarks optimised for ARM NEON and a compute-intensive application optimised for the ARM CORTEX-M4 processor to INTEL X86 processors that are SSE4.2 enabled. We demonstrate that we can take advantage of foreign intrinsics, and that automatically retargeted code delivers performance levels comparable to manually optimised code for the target platform. We achieve a speedup of up to 3.73 over a plain C baseline on an INTEL EDISON module for the target application, and on average 96.0% of the speedup of manually ported and optimised versions of the benchmarks.

## References


[1] ARM Ltd., *Cortex™-M4 Devices Generic User Guide*, 2010.

[2] D. Batten, S. Jinturkar, J. Glossner, M. Schulte, and P. D'Arcy, *A new approach to DSP intrinsic functions*, Proceedings of the 33rd Annual Hawaii International Conference on System Sciences, Jan 2000, pp. 10 pp. vol.1–.

[3] G. Bradski, *The OpenCV Library*, Dr. Dobb's Journal of Software Tools (2000).

[4] Jianjiang Ceng, Weihua Sheng, Manuel Hohenauer, Rainer Leupers, Gerd Ascheid, Heinrich Meyr, and Gunnar Braun, *Modeling instruction semantics in ADL processor descriptions for C compiler retargeting*, Journal of VLSI signal processing systems for signal, image and video technology **43** (2006), no. 2-3, 235–246 (English).

[5] L.P. Cordella, P. Foggia, C. Sansone, and M. Vento, *A (sub)graph isomorphism algorithm for matching large graphs*, Pattern Analysis and Machine Intelligence, IEEE Transactions on **26** (2004), no. 10, 1367–1372.

[6] Serge Guelton, *SAC: An efficient retargetable source-to-source compiler for multimedia instruction sets*, 2010.

[7] Weihua Jiang, Chao Mei, Bo Huang, Jianhui Li, Jiahua Zhu, Binyu Zang, and Chuanqi Zhu, *Boosting the performance of multimedia applications using SIMD instructions*, Compiler Construction (Rastislav Bodik, ed.), Lecture Notes in Computer Science, vol. 3443, Springer Berlin Heidelberg, 2005, pp. 59–75.

[8] George Koharchik and Kathy Jones, *An introduction to GCC compiler intrinsics in vector processing*, Linux Journal, September 2012.

[9] Andreas Krall and Sylvain Lelait, *Compilation techniques for multimedia processors*, Int. J. Parallel Program. **28** (2000), no. 4, 347–361.

[10] V. Lipets, N. Vanetik, and E. Gudes, *Subsea: an efficient heuristic algorithm for subgraph isomorphism*, Data Mining and Knowledge Discovery **19** (2009), no. 3, 320–350 (English).

[11] Lorenz Meier, Petri Tanskanen, Lionel Heng, Gim Hee Lee, Friedrich Fraundorfer, and Marc Pollefeys, *Pixhawk: A micro aerial vehicle design for autonomous flight using onboard computer vision*, Autonomous Robots (2012), 1–19, 10.1007/s10514-012-9281-4.

[12] Gaurav Mitra, Beau Johnston, Alistair P. Rendell, Eric McCreath, and Jun Zhou, *Use of SIMD vector operations to accelerate application code performance on low-powered ARM and Intel platforms*, Proceedings of the 2013 IEEE 27th International Symposium on Parallel and Distributed Processing Workshops and PhD Forum (Washington, DC, USA), IPDPSW '13, IEEE Computer Society, 2013, pp. 1107–1116.

[13] Alastair Murray and Björn Franke, *Compiling for automatically generated instruction set extensions*, Proceedings of the Tenth International Symposium on Code Generation and Optimization (New York, NY, USA), CGO '12, ACM, 2012, pp. 13–22.

[14] Dorit Nuzman, Sergei Dyshel, Erven Rohou, Ira Rosen, Kevin Williams, David Yuste, Albert Cohen, and Ayal Zaks, *Vapor SIMD: Auto-vectorize Once, Run Everywhere*, Proceedings of the 9th Annual IEEE/ACM International Symposium on Code Generation and Optimization (Washington, DC, USA), CGO '11, IEEE Computer Society, 2011, pp. 151–160.

[15] Dorit Nuzman and Ayal Zaks, *Outer-loop vectorization: Revisited for short SIMD architectures*, Proceedings of the 17th International Conference on Parallel Architectures and Compilation Techniques (New York, NY, USA), PACT '08, ACM, 2008, pp. 2–11.

[16] Gilles Pokam, Stéphane Bihan, Julien Simonnet, and François Bodin, *SWARP: a retargetable preprocessor for multimedia instructions*, Concurrency and Computation: Practice and Experience **16** (2004), no. 2-3, 303–318.

[17] N. Sreraman and R. Govindarajan, *A vectorizing compiler for multimedia extensions*, Int. J. Parallel Program. **28** (2000), no. 4, 363–400.

[18] Christian Tenllado, Luis Piñuel, Manuel Prieto, Francisco Tirado, and F. Catthoor, *Improving superword level parallelism support in modern compilers*, Proceedings of the 3rd IEEE/ACM/IFIP International Conference on Hardware/Software Codesign and System Synthesis (New York, NY, USA), CODES+ISSS '05, ACM, 2005, pp. 303–308.

[19] Victoria Zhislina, *From ARM NEON to Intel SSE – the automatic porting solution, tips and tricks*, Intel Developer Zone, February 2014.