Abstract—The increasing use of programmable processors as IP blocks in embedded system design creates a need for C/C++ compilers capable of generating efficient machine code. Many of today’s compilers for embedded processors suffer from insufficient code quality in terms of code size and performance. This violates the tight chip area and real-time constraints often imposed on embedded systems. The reason is that embedded processors typically show architectural features which are not well handled by classical compiler technology. This paper provides a survey of methods and techniques dedicated to efficient code generation for embedded processors. Emphasis is put on DSP and multimedia processors, for which better compiler technology is definitely required. In addition, some frontend aspects and recent trends in research and industry are briefly covered. The goal of these recent efforts in embedded code generation is to facilitate the step from assembly to high-level language programming of embedded systems, so as to provide higher productivity, dependability, and portability of embedded software.

1 Introduction

Due to the increasing complexity of embedded systems and availability of deep submicron VLSI technology, there is a shift towards more abstract system specification and implementation methods. Today’s VHDL or Verilog based specification methods are step by step being replaced by C/C++ based languages, which offer both a convenient abstraction level and high simulation speed. Examples are the SystemC [1] and SpecC [2] initiatives. In addition, hardware synthesis from C/C++ is becoming common, e.g. in products by Synopsys (CoCentric) and C Level Design.

C/C++ also offer an ideal interface to software synthesis for embedded systems. The building blocks of today’s and future systems are complex intellectual property (IP) components, or cores, many of which are programmable processors. Obviously, this IP based implementation methodology requires compilers capable of mapping C specifications into assembly code for embedded processors.

This contribution mainly deals with efficient code generation for embedded processors. Efficiency of the generated code is very important for embedded systems, due to limited system-on-a-chip memory sizes, real-time constraints of embedded applications, and the need to minimize power consumption of mobile devices. The following processor classes are common in embedded systems:

**Microcontrollers:** These are tailored for control-intensive applications and typically show a CISC architecture. Microcontrollers allow for a high code density, but computational resources are usually very limited. Examples are the 8051 and 6502 CPUs.

**RISC processors:** These show a load-store architecture and a large file of general-purpose registers. Due to the simplified instruction set, the most effective code optimization technique for RISCs is global register allocation [3]. A well-known example for RISC cores is the ARM family.

**DSP processors:** These are tuned for arithmetic intensive applications and allow for fast execution of DSP routines such as FIR filters or FFT. This is achieved by dedicated hardware support (e.g. multipliers and address generation units) and DSP-specific data path architectures. DSPs exist in a very large variety of domain-specific architectures. Major vendors include Texas Instruments, Motorola, Analog Devices, and NEC.

**Multimedia processors:** These are a recent mixture of RISC and DSP processors. They use the VLIW programming paradigm, i.e., multiple functional units working in parallel with statically determined schedules. Thus, multimedia processors allow for very high performance, however at the expense of low code density and high power consumption. Frequently, there is support for vectorized (SIMD) instructions (see section 3). Examples are the TI C6x [5] and the Philips Trimedia [6] families.

**Application-specific processors:** ASIPs are a compromise between off-the-shelf processors and ASICs. They show application-specific data paths which sometimes can be customized w.r.t. register file sizes and word
In this paper, we primarily deal with code generation for the latter three processor classes, since these are the most challenging ones from a compiler design viewpoint. In fact, current compilers for many standard DSPs and multimedia processors have been empirically shown to produce significantly less efficient code (as much as 1000% overhead) in terms of performance and code size, as compared to hand-optimized reference code [9, 10, 11, 12]. The reason is that their special instruction set architectures are not well exploited by conventional compiler technology. Therefore, a large part of the software for such processors still has to be developed in assembly language. This implies time-consuming programming, extensive debugging, and low code portability. The requirements of short time-to-market and dependable code are obviously much better met by using C/C++ instead of assembly.

The need for more efficient compilers for embedded processors has been recognized about a decade ago. The purpose of this contribution is to provide a survey of important techniques that have been developed for embedded code optimization, while also highlighting their key methodologies which reach beyond the scope of classical compiler technology for general-purpose systems.

In the following three sections, we give examples for code optimization techniques at different levels in the compilation flow, reaching from source level (section 2) to assembly level techniques (section 4). In section 5, we briefly discuss compiler frontend related issues, while sections 6 and 7 provide an overview of recent compiler trends in academia and industry. Finally, conclusions are given.

## 2 Source level optimization

The compilation process, i.e., the mapping of a given C program to a behaviorally equivalent assembly program, starts with source code analysis and source level optimization. Optimizing already at the C level is attractive, because the result is still machine-independent in the sense that it can be compiled to different target processors. In addition, also C based hardware synthesis can benefit from source level optimization. Several examples are given in the following.

### Standard optimizations

All reasonable compilers perform machine-independent standard optimizations, such as constant folding, common subexpression elimination, or jump optimization [13, 14, 15]. These techniques need only a minimum of machine-specific information and are beneficial for most programs. Frequently, such standard optimizations are also performed at the intermediate representation (IR) level, where complex source code constructs have already been split into a simple form, such as three-address code.

### Address code transformation

The high-level address code transformation techniques described in [16] can be regarded as an extension of the above standard optimizations, targeted towards memory-intensive applications. Here the goal is the simplification of array index expressions beyond the classical induction variable elimination technique (see e.g. [13]). At the expense of larger code size, this leads to a reduction of up to 50% in instruction cycles for array-intensive nested loops.

### Loop transformations

In case of multimedia applications mapped to VLIW processors, loop transformations are a very effective means of code optimization. A simple example is loop unrolling, where loop iterations are duplicated, resulting in larger basic blocks and thereby in a higher potential for parallelization of instructions during scheduling. Its counterpart is loop folding or software pipelining [18, 19], where loop iterations are restructured in such a way, that the critical path length within the loop body is reduced. Again, these loop optimizations come at the price of an increased code size. Further loop restructuring techniques are described in [20].

### Function inlining

Function inlining is a well-known technique, in which function calls are replaced by copies of function bodies, so as to reduce the calling overhead. Usually, compilers use local heuristics in order to identify suitable candidate functions for inlining, while mostly neglecting code size constraints. In contrast, the inlining technique described in [17] aims at a maximum program speedup for a given global code size constraint, and thus better meets the demands of embedded processors. It is based on profiling information, code growth estimation, and a branch-and-bound optimization procedure. This technique also exemplifies a common concept in code generation for embedded processors: By the use of time-intensive optimization techniques, a larger search space can be explored, which in turn leads to better code than in traditional compilers. This approach is valid, since for embedded systems high code quality is much more important than high compilation speed.

## 3 Optimized instruction set mapping

After source level or IR level optimization, the machine-independent IR statements are mapped to assembly instructions. At this point all machine-specific features, such as special-purpose registers, complex instruction patterns, and inter-instruction constraints need to be taken into account. This is what makes efficient code generation for embedded processors generally difficult.
Tree pattern matching  A very important basic technique for instruction set mapping, or code selection, is tree pattern matching [21]. IR statements are represented in the form of data flow trees (DFTs), where tree nodes correspond to variables, constants, and operations, while edges denote data dependencies. Also the machine instructions can be considered as small tree patterns. Thus, an optimal mapping is given by a minimum cover of a DFT by cost-attribute instruction patterns (fig. 1). A number of tools are available for automatic generation of tree pattern matchers from instruction set grammar specifications. These include IBUG [22], BEG [23], and OLIVE [21]. The generated matchers are given as C source code, and they compute optimum tree covers in linear time by means of dynamic programming.

Figure 1: Code selection: a) Data flow tree, b) available instruction patterns, c) optimal tree cover

Phase-coupled code generation  Besides code selection, the task of machine code generation also comprises register allocation and instruction scheduling. Traditionally, these phases, each of which is an NP-hard optimization problem in itself, are solved sequentially and heuristically. However, there is a cyclic dependence, since each of the three phases may impose possibly unnecessary and obstructive restrictions on the remaining ones. Therefore, one key method in embedded code generation is phase coupling. Early techniques in this area include [24, 25, 26], each of which integrated two out of three code generation phases. In [27], the above tree pattern matching technique has been coupled with register allocation and scheduling for a family of TI DSPs. Later, an additional heuristic for handling (more general) data flow graphs (DFGs) has been presented [28]. An improved approach, based on simulated annealing, has been described in [29]. Mutation scheduling [30] is another approach to achieve phase coupling, where also algebraic transformations are exploited in order to explore alternative instruction set mappings. In [31, 32], phase-coupled compilers for certain classes of VLIW processors have been described. A constraint logic programming technique for DSSPs with irregular architectures has been presented in [33]. In that approach, all binding decisions are delayed until they are really required, which yields a maximum degree of freedom in code generation. For several DSSPs the code quality has been shown to be very close to assembly programs, however at the expense of high compilation times. Further, more theoretically-oriented, contributions include [34, 35, 36].

Figure 2: SIMD instruction “ADD2” in the TI C6x processor

Multimedia instruction sets  Multimedia processors require different mapping techniques than DSPs. On one hand, their architecture is more regular and VLIW/RISC-like, on the other hand they show special new instruction types. An example are SIMD (single instruction multiple data) instructions (fig. 2). These instructions perform multiple identical operations in parallel, while splitting the data registers into subregisters, each of which then contains a separate data item. The power of SIMD instructions lies in faster execution of algorithms operating on "short" (16 or 8 bit) values. However, most current compilers for multimedia processors can exploit SIMD instructions only by means of compiler-known functions, i.e., C level macros. The reason is that SIMD instructions do not fit into the usual tree based code selection approach. This results in non-portable code, and the responsibility of exploiting the SIMD mode is still with the human programmer. In order to circumvent this problem, special language constructs [37] or C++ class libraries [38] have been proposed. The technique described in [39] is capable of generating SIMD instructions also from plain C code, based on an extension of the above tree pattern matching technique towards general graphs.

4 Assembly level optimization

Once assembly code has been generated, a significant optimization potential is still left w.r.t. memory access organization and instruction scheduling. This concerns the consideration of the concrete memory architecture, as well as exploitation of address generation hardware and instruction-level parallelism.

Memory access optimization  Several DSPs, e.g. Motorola 56k, Analog Devices 210x, and AMS Gepard show two memory banks (usually called X and Y), which are accessible in parallel. This raises the problem of partitioning the program variables between X and Y in such a way, that potential parallelism is maximized. Naive compilers use only one of the two banks, or leave the partitioning decision to the programmer by means of C language extensions. More advanced approaches use a dedicated optimization phase for
registration file sizes and bound on size, which differ in the concrete AGU configurations (e.g., already available, including [44, 45, 46, 47, 48, 49, 50, 51, 52], or adding the contents of a modify register. The latter two modes should be preferred, since they translate to efficient parallel auto-increment addressing and fast access modes is exploited in combination with processor pipeline timing information. Due to this "memory-aware" approach, the scheduling of instructions can be adapted to the concrete memory module in use. This generally results in fewer pipeline stalls, and for a TI C6X processor an average speedup of 24% as compared to a traditional approach has been measured.

Address code optimization One area in DSP code optimization that has been thoroughly investigated is the utilization of dedicated address generation units (AGUs, fig. 3). Such AGUs generally contain address register and modify register files, where address register updates can take place in three ways: Adding a "wide" immediate constant c, adding a "short" immediate constant d, or adding the contents of a modify register. The latter two modes should be preferred, since they translate to efficient parallel auto-increment address computations. Exploitation of auto-increment modes depends on the layout of variables in memory. A relatively large number of techniques for computing good layouts are available, including [44, 45, 46, 47, 48, 49, 50, 51, 52], which differ in the concrete AGU configurations (e.g., register file sizes and bound on d) that can be handled, as well as in the optimization method (e.g., heuristics or genetic algorithms). The achievements in this area can be considered satisfactory, and the techniques have already found their way into commercial compilers.

Instruction scheduling Instruction scheduling assigns generated machine instructions to control steps, which is important for VLIW-like multimedia processors as well as for DSPs with limited instruction-level parallelism. Local scheduling (or code compaction) algorithms, such as the list scheduling heuristic, are limited to the scope of a single basic block [53]. For DSPs also exact local scheduling techniques have been developed [54, 55]. In contrast, global techniques operate on an entire function [56, 57]. The latter result in a more effective scheduling of global critical paths, but frequently at the expense of larger code size. Recent scheduling techniques [58] aim at an extensive pre-analysis of scheduling constraints, so as to achieve shorter schedules as compared to pure heuristics.

5 Frontend issues

Besides the above code optimization techniques, C/C++ compiler development for embedded processors also requires frontend development for embedded processors, which performs lexical, syntactical, and semantical analysis of source programs. A popular commercial product is the EDG frontend [59], which however is relatively expensive. Research frontends include the Triamaran system [60], which applies to a class of VLIW processors, and Stanford’s SUIF system [61]. For the latter, also a backend development system has been announced recently [62]. Also the GNU C compiler [63] is sometimes used in the context of embedded processors, even though its limitations make it rather difficult to adapt it to irregular processors architectures. The LANCE system [64] developed at Dortmund serves as a compiler infrastructure both for research and industrial projects. It comprises an ANSI C frontend, a library of standard IR optimizations, as well as a backend interface for data flow tree generation. One key feature is that new IR optimizations can be quickly added as "plug-and-play" components. An executable IR in a compiled target C code supports validation of IR transformations and simultaneously allows for source-level optimizations.

6 New research directions

Most existing code optimization techniques focus on code size and performance. However, in the context of embedded systems, additional goals may be relevant. These include low power for mobile applications and re-targetability for support of architecture exploration.

Low power Power minimization has been a design goal in hardware synthesis for several years. However, only recently the impact of compiler techniques on power consumption has been investigated. It is important to note, that power optimization in compilers may be contrary to the traditional goals, code size and performance, and therefore requires new techniques. Existing work, e.g., [65, 66], so far mainly focused on the impact of instruction selection and scheduling on power consumption of machine code. However, also more general approaches, such as alternative instruction and data encodings and utilization of power-efficient on-chip memories should be further investigated. A special session at the
Design Automation Conference 2000 also stressed the importance of low power embedded software.

Retargetability Retargetable compilers can generate code for different target processors, based on external (e.g. HDL) machine models (fig. 4). Their application area is twofold: Retargetable compilers are useful in the context of parameterizable ASIPs (see section 1), because a single compiler is sufficient to explore different configurations of a given ASIP, thereby making hardware/software trade-offs. A significant amount of retargetable compiler technology for embedded processors is already available, including MSSQ [67], RECORD [68], SPAM [28], CHESS [69], CodeSyn [70], and AVIV [31]. Currently, retargetable compilers are receiving renewed interest also due to the need for architecture exploration at the system level: Retargetable compilers, in combination with simulators, can provide an early estimation of the performance of different target processors for a given application. A recent research effort into this direction is the EXPRESS project at UCI [71, 43].

7 Industrial trends

With the increasing amount of technology available for embedded code generation, several retargetable optimizing compilers for embedded processors have become commercially available. These include CoSy (ACE, [72]), CHESS (Target Compiler Technologies, [73]), and Archelon [74]. However, these tools still have their limitations w.r.t. certain processor classes, and also the retargeting mechanism should be further improved, so as to better fit into the system design flow. Concerning high-performance “off-the-shelf” processors, currently a trend towards VLIW machines can be observed. This trend is mainly enabled by increasing chip integration scales, which allows for integration of many parallel functional units. The VLIW trend is expected to continue [75], which will also facilitate the construction of more efficient compilers. However, also new problems are introduced, such as mapping to SIMD or conditional instructions, or scheduling for clustered VLIW data paths. Simultaneously, more and more ASIPs and customizable processors are in use. A promising new entry in this area is Tensilica’s Xtensa core [7]. Using a special description language, this RISC core can be extended by application-specific instructions, while both the HDL core model and the corresponding software tools, including a C compiler, are automatically retargeted.

8 Conclusions

We have provided a brief survey of methods and techniques in the area of code generation for embedded processors. More extensive overviews and literature references can be found e.g. in [76, 70, 77]. With the growing importance of compilers for embedded processors, research results are increasingly turned into products. In the future, we expect a higher importance of VLIW and ASIP processors, where also non-conventional compiler design goals like low power and retargetability will be of high interest.

References
