OCEANS: Optimising Compilers for Embedded ApplicatioNS*

Michel Barreteau[1], François Bodin[2], Peter Brinkhaus[3], Zbigniew Chamski[4], Henri-Pierre Charles[1], Christine Eisenbeis[5], John Gurd[6], Jan Hoogerbrugge[4], Ping Hu[5], William Jalby[1], Peter M. W. Knijnenburg[3], Michael O'Boyle[7], Erven Rohou[2], Rizos Sakellariou[6], André Seznec[2], Elena A. Stöhr[6], Menno Treffers[4], and Harry A. G. Wijshoff [3]

[1] Laboratoire PRiSM, Université de Versailles, 78035 Versailles, France.
[2] IRISA, Campus Universitaire de Beaulieu, 35042 Rennes, France.
[3] Department of Computer Science, Leiden University, P.O. Box 9512, 2300 RA Leiden, The Netherlands.
[4] Philips Research, Information and Software Technology, Prof. Holstlaan 4, 5656 AA Eindhoven, The Netherlands.
[5] INRIA, Rocquencourt, BP 105, 78153 Le Chesnay Cedex, France.
[6] Department of Computer Science, The University, Manchester M13 9PL, U.K.
[7] Department of Computer Science, The University, Edinburgh EH9 3JZ, U.K.

Abstract. This paper presents an overview of the activities carried out within the ESPRIT project OCEANS whose objective is to investigate and develop advanced compiler infrastructure for embedded VLIW processors. This combines high and low-level optimisation approaches within an iterative framework for compilation.

1 Introduction

Embedded applications have become increasingly complex during the last few years. Although sophisticated hardware solutions, such as those exploiting instruction level parallelism, aim to provide improved performance, they also create a burden for application developers. The traditional task of optimising assembly code by hand becomes unrealistic due to the high complexity of hardware/software. Thus the need for sophisticated compiler technology is evident.

Within the OCEANS project, the consortium intends to design and implement an optimising compiler that utilises aggressive analysis techniques and integrates source-level restructuring transformations with low-level, machine dependent, optimisations [1, 14, 16]. A major objective of the project is to provide a prototype framework for iterative compilation where feedback from the low-level is used to guide the selection of a suitable sequence of source-level transformations and *vice versa*. Currently, the Philips TriMedia (TM1000) VLIW processor [8] is used for the validation of the system.

In this paper, we present the work that has been carried out during the first 15 months since the project started (September 1996). This has largely

* This research is supported by the ESPRIT IV reactive LTR project OCEANS, under contract No. 22729.

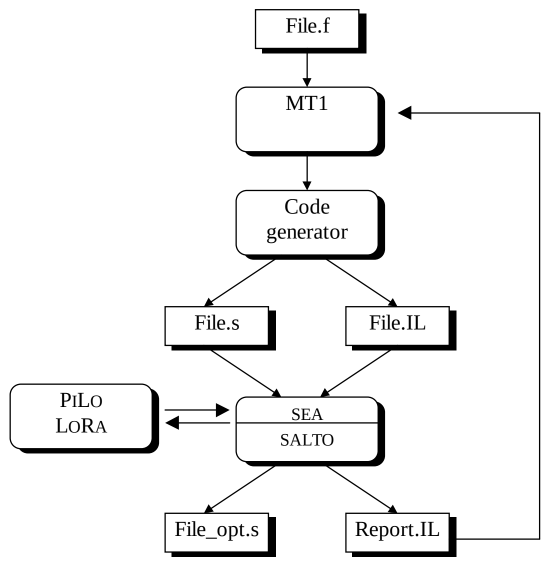


Fig. 1. The Compilation Process.

concentrated on the development of the necessary compiler infrastructure. An overall description of the system is given in Section 2. Sections 3 and 4 present the high-level and the low-level subsystems respectively, while the steps that have been taken towards their integration are highlighted in Section 5. Finally, some results from the initial validation of the system are shown in Section 6, and the paper is concluded with Section 7.

2 An Overview of the OCEANS Compiler System

The OCEANS compiler is centred around two major components: a high-level restructuring system, MT1, and a low-level system for supporting assembly language transformations and optimisations, SALTO, which is coupled with SEA, a set of classes that provide an abstract view of the assembly code, and tools for software pipelining (PILO) and register allocation (LORA). Their interaction is illustrated in Figure 1 which shows the overall organisation of the OCEANS compilation process. In particular, a program is compiled in three main steps:

- First, MT1 performs lexical, syntactical and semantic analysis of a source FORTRAN program (`File.f`). Also, a sequence of source program transformations can be applied.
- The restructured source program is then fed into the code generator which generates sequential assembly code that is annotated with instruction identifiers used to identify common objects in MT1 and SALTO, and a file written in an *Interface Language* (`File.IL`) that provides information on data dependences and control flow graphs.

– Finally, SALTO (coupled with SEA) performs code scheduling and register allocation. At this step guarded instructions are created and resource constraints are taken into account.

The above process is repeated iteratively until a certain level of performance is reached. Thus, different optimisations, both at the source-level and the low-level, are checked and evaluated. An important feature of the system is the existence of a client-server protocol that has been implemented in order to provide easy access to the compiler over the Internet, for all members of the consortium. MT1 and the code generator are located at Leiden, and SALTO, SEA, PILO and LORA are located at Rennes.

3 High-Level Transformations

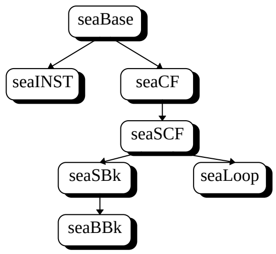
Optimizing and restructuring compilers incorporate a number of program transformations that replace program fragments by semantically equivalent fragments to obtain more efficient code for a given target architecture. The problem of finding an optimum order for applying them is commonly known as the *phase ordering problem*. Within the MT1 compilation system [5] this problem is solved by providing a *Transformation Definition Language (TDL)* [3] and a *Strategy Specification Language (SDL)* [2]. Transformations and strategies specified in these languages can be loaded dynamically into the compiler.

3.1 Transformation Definition Language

The TDL is based on pattern matching. The user can specify an *input pattern*, a transformed *output pattern* and a *condition* when the transformation can be legally and/or beneficially applied. Patterns may contain *expression* and *statement variables*. When a pattern is matched against the code these variables are bound to actual expressions and code fragments, respectively. The expression and statement variables can be used in turn in the specification of the output pattern and the condition. This mechanism allows one to specify a large number of transformations, such as loop interchange, loop distribution or loop fusion. However, it is not powerful enough to express other important transformations, such as loop unrolling. Therefore, the TDL also allows for *user-defined functions* in the output pattern. User-defined functions are the interface to the internal data structures of the compiler. In this way, any algorithm for transforming and testing code can be implemented and made accessible to the TDL.

3.2 Strategy Specification Language

The order in which transformations have to be applied is specified using a Strategy Specification Language (SSL). It allows the specification of an optimising strategy at a more abstract level than the source code level. This language contains sequential composition of transformations, a choice construct and two repetitive constructs.



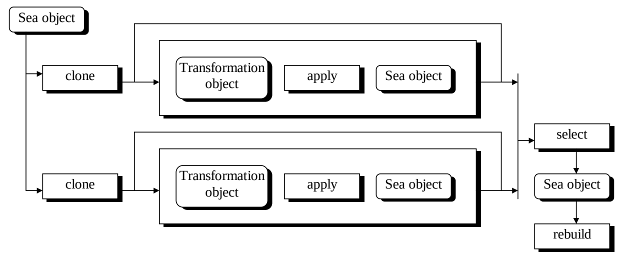


Fig. 2. SEA class hierarchy.

Fig. 3. Typical usage of SEA classes.

An IF statement consists of a transformation that acts as a condition, a THEN part and an optional ELSE part. The transformation in the condition can be applied successfully or not. If it is successful, the transformations in the THEN part are to be executed. Optionally, in the ELSE part a list of transformations can be given which should be executed in case the transformation matched but was not applied successfully due to failing conditions.

The two repetitive constructs consist of a transformation to be checked and a statement list to be executed if the condition is true or false, respectively. They consist of a WHILE-ENDWHILE and a REPEAT-UNTIL construct.

Examples of how to specify strategies in SSL can be found in [2].

4 Low-Level Optimisations

Low-level optimisations are built on the top of SALTO, a retargetable system for assembly language transformation and optimisation [15]. To facilitate the implementation of optimisations, a set of classes has been designed, SEA (SALTO Enhanced Abstraction), that provides an abstract view of the assembly code which is more pertinent to the code scheduling and register allocation problems. The most important features of SEA are that it allows the evaluation of various code transformations before producing the final code, and that it separates the implementation of the global low-level optimisation strategy from the implementation of individual optimisation sequences.

The SEA model contains two kinds of objects:

code fragments The following objects can be used. `seaINST`: an instruction object; `seaCF`, an unstructured set of code fragments; `seaSCF`, a structured subset of control flow graph with a unique entry point; `seaSBk`, a superblock; `seaBBk`, a basic block; and finally, `seaLoop`, a structured piece of code that has loop properties. Figure 2 illustrates the corresponding class hierarchy.

transformations to be applied to subgraphs. All transformations are characterized by the following main methods: `preCond()` returns the set of control flow subgraphs that qualifies for the transformation; `apply()` applies the transformation to a given subgraph, and finally, `getStatus()` that returns the status of the transformation after application (*success* or *failure*) and the reason for the failure.

The usage of the SEA objects is shown in Figure 3. A transformation is tried on a cloned piece of code, then according to performance or size criteria one of the solutions found is chosen and propagated to the low-level program representation using the `rebuild()` method.

The optimisations currently available within SEA are: *register renaming*, *superblock construction* [12], *guard insertion* [11], *loop unrolling* (also available at the high-level), *local/superblock scheduling* [12], *software pipeline*, and *register allocation*. The implementation of software pipeline is based on the tools PILO [17] and LORA [9] which generate a modulo scheduling of the loop body.

5 Integration

5.1 The Interface Language

In order to transmit information between the various components of the compiler, an *Interface Language* (IL) was designed. This allows the propagation of information, such as data dependences and loop control data, from MT1 to SALTO, as well as feedback information from the scheduled code back to MT1.

An IL description consists of three sections: *a list of keywords* that specifies the list of attributes that can apply to an object; *a default level setting* that indicates the type of code the objects belong to; and *a list of object references* which specify the nature, contents and attributes of an object. More details on the IL can be found in [7].

5.2 Information Forwarded and Feedback

Data dependence information is propagated from MT1 to SALTO and is used for memory disambiguation. The feedback from SEA to MT1 (file `Report.IL` in Figure 1) contains information on the code structure, the basic blocks, as well as a record of the transformations that were applied. Data related to each basic block include the total number of assembly instructions, the critical path for scheduling the code, the number of cycles of the scheduled code, and a grouping depending on the nature of the instructions. Examples can be found in [7].

MT1 uses the feedback from SALTO in order to build an internal data structure that can be accessed by the TDL and the SSL by means of user-defined functions in the condition that can check for the identity of a code fragment and suggestions made by SALTO. When such a transformation is used as the condition for an IF construct in the SSL, we are able to select the transformations we want to apply to this fragment.

Initially, MT1 compiles the program without performing any restructuring and the compiled program is scheduled by SALTO. SALTO identifies the code fragments that can be improved. It reports its diagnostics to a cost model that makes a decision on what kind of restructuring could be performed next. Then, MT1 reads the suggestions for restructuring and performs these. It is intended that a transformation sequence for a given program fragment is selected by following a systematic approach for searching through a domain of possible transformations. First, each different transformation is applied once and then the same follows for each branch of the tree. The search space is minimised by using a threshold condition for terminating branches that are not likely to yield an optimum result in their descendants. Some preliminary experiments using this strategy can be found in [10]; further work is in progress.

6 Validation of the Initial System

In order to validate the compiler, four public domain multimedia codes have been selected [4]. These are a low bit-rate encoder/decoder for H.263 bitstreams, an MPEG2 encoder/decoder, an implementation of the CCITT G.711, G.721 and G.723 voice compression standards, and the Persistence of Vision Ray-Tracer for creating 3D graphics.

At the high-level, initial experimentation aimed at identifying those transformations that appear to be the most crucial in optimising code scheduling. Inspection of the benchmarks revealed that they contain many imperfectly nested double or triple loops with much overhead due to branch delays. In order to deal with such loops, a transformation that converts the imperfectly nested loop into a single loop has been suggested [13].

At the low-level, the initial validation of the system has been carried out by applying four different optimisation sequences:

- S_0 is the simplest sequence. First, the code is scheduled locally and then register allocation is performed.
- $S_1(u)$ is based on unrolling the loop body u times. The unrolled body is transformed into a superblock by guarding instructions. As in S_0, register allocation is performed after local scheduling.
- $S_2(u)$ is similar to $S_1(u)$ except that register allocation is performed before scheduling. This usually requires less registers, allowing this sequence to succeed when $S_1(u)$ fails due to a lack of registers.
- S_3 consists in applying a software pipelining algorithm.

The above optimisation sequences were validated and indicative results, using the most-time consuming loops of H263, are illustrated in Figure 4. Every optimisation sequence has been applied to each of the six selected loops and the size of the resulting VLIW code and the speed of the loop, i.e. the number of cycles per iteration, were computed. ¿From the table, a well-known result is observed: the more we unroll a loop, the faster it runs — cf. columns $S_1(2)$, $S_1(3)$, $S_1(4)$ — but at the expense of a larger code size. As expected $S_2(2)$ yields too poor performance and large code because of the presence of false dependences. Finally, software pipelining (S_3) gives the best performance but at the expense of a very large increase in code size. Note that this transformation failed with the last loop, due to a lack of registers.

	Optimisation sequences						C code
	S_0	$S_1(2)$	$S_1(3)$	$S_1(4)$	$S_2(2)$	S_3	

speed	8	6	5	5	7	3	`for (i=xa;i<xb; i++)`
							`{ d[i]=s[i]*om[i];`
							`}`

size	8	12	16	20	13	75	

speed	9	7	6	6	10	5	`for (i=xa; i<xb; i++)`
							`{ d[i]+=s[i]*om[i];`
							`}`

size	9	13	18	22	19	55	

speed	12	8	8	7	12	6	`for (i=xa; i<xb; i++)`
							`{ dp[i]+=(((unsigned int)(sp[i]`
							`+sp2[i]+1))>>1)*om[i];`
							`}`

size	12	16	24	28	24	121	

speed	15	10	9	9	16	6	`for (i=xa; i<xb; i++)`
							`{ dp[i]+=(((unsigned int)(sp[i]+`
							`sp[i+1]+1))>>1)*OM[c][j][i];`
							`}`

size	15	20	28	34	31	172	

speed	15	10	10	8	17	7	`for (i=xa; i<xb; i++)`
							`{ dp[i]+=(((uint)(sp[i]+sp2[i]+`
							`sp[i+1]+sp2[i+1]+2))>>2)*om[i];`
							`}`

size	15	19	29	33	33	179	

speed	19	13	12	11	30	–	`for (k=0; k<5; k++)`
							`{ xint[k]=nx[k]>>1;  xh[k]=nx[k] & 1;`
							`yint[k]=ny[k]>>1;  yh[k]=ny[k] & 1;`
							`s[k]=src+lx2*(y+yint[k])+x+xint[k];`
							`}`

size	19	25	36	44	59	–	

Fig. 4. Time consuming loops extracted from H263.

In most embedded applications, it is necessary to answer globally questions such as: "Given a maximum code size, what is the highest performance that can be achieved?", or "Given a performance goal, what is the smallest code size that can be achieved?". Within the OCEANS compiler this trade-off is evaluated quantitatively by applying a novel compiler strategy. Thus, the choice of the most suitable optimisation is made *a posteriori*, when the impact of each possible transformation is known. More details can be found in [6].

7 Conclusion and Future Work

The previous sections outlined the current status of the OCEANS compiler. Although the results obtained so far, using the initial prototype, are satisfactory (comparing with a production compiler), the implementation work still continues on both the high and low levels. A major part of the work during the next months and until the end of the project is devoted to the integration of the two levels, the development of a prototype framework for iterative compilation, and its experimental evaluation and tuning. Finally, it is intended that the system be made publically available in due time (at the moment SALTO is available on request).

References

1. B. Aarts, *et al.* OCEANS: Optimizing Compilers for Embedded Applications. In C. Lengauer, M. Griebl, S. Gorlatch (Eds.), *Proceedings of Euro-Par'97*, Lecture Notes in Computer Science 1300, Springer-Verlag, 1997, pp. 1351–1356.
2. R. A. M. Bakker, F. Bregt, P. M. W. Knijnenburg, P. Touber, and H. A. G. Wijshoff. Strategy Specification Language. OCEANS Deliverable D1.2a, 1997.
3. A. J. C. Bik, P. J. Brinkhaus, P. M. W. Knijnenburg, P. Touber, and H. A. G. Wijshoff. Transformation Definition Language. OCEANS Deliverable D1.1, 1997.
4. A. J. C. Bik, P. J. Brinkhaus, P. M. W. Knijnenburg, P. Touber, H. A. G. Wijshoff, W. Jalby, H.-P. Charles, M. Barreteau. Identification of Code Kernels and Validation of Initial System. OCEANS Deliverable D3.1c, 1997.
5. A. J. C. Bik and H. A. G. Wijshoff. MT1: A Prototype Restructuring Compiler. Technical Report 93-32, Department of Computer Science, Leiden University, 1993.
6. F. Bodin, Z. Chamski, C. Eisenbeis, E. Rohou, and A. Seznec. GCDS: A Compiler Strategy for Trading Code Size Against Performance in Embedded Applications. Research Report 1153, IRISA, 1997.
7. F. Bodin and E. Rohou. High-level Low-level Interface Language. OCEANS Deliverable D2.3a, 1997.
8. B. Case. Philips Hope to Displace DSPs with VLIW. *Microprocessor Report*, 8(16), 5 Dec. 1994, pp. 12–15. See also `http://www.trimedia-philips.com/`
9. C. Eisenbeis, S. Lelait, and B. Marmol. The meeting graph: a new model for loop cyclic register allocation. *Proceedings of PACT'95* (Cyprus, June 1995).
10. J. Gurd, A. Laffitte, R. Sakellariou, E. A. Stöhr, Y. T. Chu, and M. F. P. O'Boyle. On Compile-Time Cost Models. OCEANS Deliverable 1.2b, 1997.
11. W. Hwu, R. E. Hank, D. M. Gallagher, S. A. Mahlke, D. M. Lavery G. E. Haab, J. C. Gyllenhaal, and D. I. August. Compiler Technology for Future Microprocessors. *Proceedings of the IEEE*, 83(12), Dec. 1995, pp. 1625–1639.
12. W. Hwu, *et al.* The Superblock: An Effective Technique for VLIW and Superscalar Compilation. *The Journal of Supercomputing*, 7(1), May 1993, pp. 229–248.
13. P.M.W. Knijnenburg. Flattening: VLIW Code Generation for Imperfectly Nested Loops. *Proceedings CPC'98*, 1998. To appear.
14. OCEANS Web Site at `http://www.wi.leidenuniv.nl/Oceans/`
15. E. Rohou, F. Bodin, A. Seznec, G. Le Fol, F. Charot, F. Raimbault. SALTO: System for Assembly-Language Transformation and Optimization. Technical Report 1032, IRISA, June 1996. See also `http://www.irisa.fr/caps/Salto/`
16. R. Sakellariou, E. A. Stöhr, and M. F. P. O'Boyle. Compiling Multimedia Applications on a VLIW Architecture. *Proceedings of the 13th International Conference on Digital Signal Processing (DSP97)* (Santorini, July 1997), vol. 2, IEEE Press, 1997, pp. 1007–1010.
17. J. Wang, C. Eisenbeis, M. Jourdan, and B. Su. Decomposed Software Pipelining: a New Perspective and a New Approach. *International Journal on Parallel Processing*, 22(3), 1994, pp. 357–379.