PFPC: An Infrastructure for Research on Parallelizing Compilers*

Chao-Tung Yang† Shian-Shyong Tseng‡ Yun-Woei Fann
Ming-Huei Hsieh

Ground Section Section
National Space Program Office
Hsinchu, Taiwan 300, ROC

Dept. Computer & Information Science
National Chiao Tung University
Hsinchu, Taiwan 300, ROC

Abstract

As we know, the execution efficiency of a loop can be enhanced if the loop is executed in parallel or partially parallel, like a DOALL or DOACROSS loop. This paper also reports on a practical parallel loop detector (PPD) that is implemented in PFPC on finding the parallelism in loops. The PPD can extract the potential DOALL and DOACROSS loops in a program by verifying array subscripts. In addition, a new model by using knowledge-based approach is proposed to exploit more loop parallelisms in this paper. The knowledge-based approach integrates existing loop transformations and loop scheduling algorithms to make good use of their ability to extract loop parallelisms. Two rule-based systems, called the KPLT and IPLS, are then developed using repertory grid analysis and attribute ordering tables respectively, to construct the knowledge bases. Finally, a runtime technique based on inspector/executor scheme is proposed in this paper for finding available parallelism on loops. Our inspector can determine the wavefronts of a loop with any complex indirected array indexing pattern by building a DEF-USE table. Experimental results show that the new method can handle any complex data dependence pattern that cannot be handled by the previous research.

1 Introduction

The last decade has seen the coming of age of parallel computing. Many different classes of multiprocessor systems have been designed and implemented in industry and academia, for example, IBM RP3, Cray T3D, NEC SX-3, CONVEX C4, CONVEX SPP, and IBM SP2. To achieve high speedup of such systems, it requires decomposition of tasks into several sub-tasks which can be executed on different processors in parallel. Unfortunately, it possesses several difficulties for the users to write explicitly parallel programs. First, they had to rewrite their existing sequential programs into parallel programs. Second, most of the resulting explicitly parallel programs were not portable. Third, writing efficient parallel programs often required optimizations that need intimate knowledge of the machine's architecture and the program's access patterns, e.g., data distribution, prefecting, or blocking.

To address these difficulties, parallelizing compilers were developed to transform sequential programs into parallel ones [14, 1, 7]. Parallelizing compilers can be broken into two components: a component that identifies parallelism in a program, and a component that exploits this parallelism. The component that identifies parallelism attempts to determine what parts of a program can be run in parallel. The component that exploits parallelism determines which of these parallel parts should be run in parallel, as well as how to generate efficient codes for them. Therefore, design of efficient parallelizing compiler is an important part of achieving maximum parallelism on multiprocessors. However, the generation processes of parallel object codes by parallelizing compilers are very difficult and complicated. Most investigations of parallelizing compiler still focus on source-to-source transformation, for example, *Parafrase-2 and Polaris* developed at UIUC [1, 8], *ParaScope* developed at Rice University [3], and *SUIF* developed at Stanford University [6].

In addition to the advance in computer architecture, some operating systems also support parallelism. Multithreading support seems to be the most obvious approach for helping programmers to take the advantage of parallelism by operating system. For example, Mach, OSF/1, Solaris, Microsoft Windows NT are operating systems that support multithreading. These operating systems usually have packages for handling multithreads [2], e.g., the C Threads package in Mach and P Threads package in OSF/1. Although a multithreading operating system for a multiprocessor system can be powerful, it still needs good parallelizing compilers to help programmers exploit parallelism and gain performance benefit. So, we wanted to design and implement a portable parallelizing compiler for multithreading operating system. Our compiler can generate parallel object codes for running on multiprocessor systems rather than being just a source-to-source restructurer [10, 12].

This paper describes the design and implementation of an efficient parallelizing compiler to parallelize loops and achieve high acceleration rates on multiprocessor systems. In this paper we introduce how to design and implement a portable FORTRAN parallelizing compiler (PFPC) on a shared-memory multiprocessor machine running multi-

*This work was supported in part by National Science Council of Republic of China under Grant No. NSC89-2213-E009-100.

†Corresponding author. E-mail: ctyang@nspo.gov.tw

‡E-mail: sstseng@cis.nctu.edu.tw.

threading operating system OSF/1. Our compiler is highly modularized so that porting to other platforms will be very easy. Furthermore, the compiler can partition parallel loops into multithreaded codes based on several DOALL loop-partitioning algorithms. Then, this paper reports on the practical parallelism detector (PPD) that is implemented in PFPC at NCTU to concentrate on finding available the parallelism on loops [13]. The PPD is used on extracting the potential DOALL and DOACROSS loops in a program. Moveover, if DOACROSS loops are available, an optimization of synchronization statements were made.

To exploit more parallelism, a new model by using knowledge-based techniques is proposed in this paper [9]. The knowledge-based approach integrates existing loop transformations and loop scheduling algorithms to make good use of their ability to extract loop parallelisms. Two rule-based systems, called the KPLT and IPLS, are then developed using repertory grid analysis and attribute ordering tables respectively, to construct the knowledge bases. For instance, IPLS can choose an appropriate algorithm and then apply the resulting algorithm to assigning parallel loops on multiprocessor systems to achieve high speedup rates [4]. Finally, a runtime technique based on inspector/executor scheme is proposed in this paper for finding available parallelism on loops. Our inspector can determine the wavefronts of a loop with any complex indirected array indexing pattern by building a DEF-USE table [11]. The inspector is fully parallel without any synchronization. Experimental results show that the speedup delivered by our compiler is high. Furthermore, for system maintenance and extensibility, our approach is obviously superior to others. As an ultimate goal, a high-performance and portable FORTRAN parallelizing compiler on shared-memory multiprocessors will be constructed.

2 The Model of Parallelizing Compilers

2.1 An Overview of PFPC

Multithreading support may be the most obvious approach to help programmers take the advantage of parallelism by operating systems. Therefore, we propose a new model of parallelizing compiler for exploiting potential power of multiprocessors and gaining performance benefit on multithreaded operating systems OSF/1 [2]. The portable FORTRAN parallelizing compiler (PFPC) intended to produce parallel object codes rather than just acting as a source-to-source restructurer is shown in Figure 1 [10, 12].

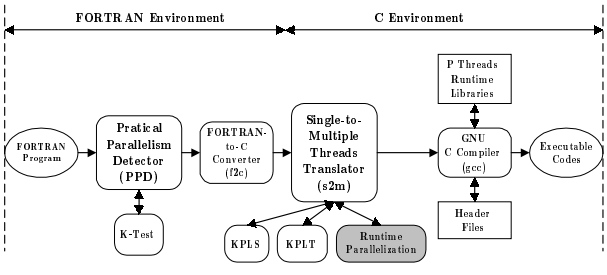

Figure 1: The **PFPC** model running on OSF/1.

First, a practical parallelism detector (PPD) is used to test the data dependences of array references and then restructure a sequential FORTRAN source program into a parallel form at compile-time [13], i.e., if a loop can be parallelized or partially parallelized, then PPD marks that loop with DOALL loop or DOACROSS loop by comments. If the access patterns of some arrays cannot be determined at compile-time or have non-constant dependence vector, then PPD marks that loops with DOCONSIDER loop by comments. The flow of loops parallelization is shown in Figure 2. The PPD (practical parallelism detector) will analyze the loop's array access patterns to find the data dependences of array references. As we know, if the information of data dependence is not available until the program is running, i.e., defy the static analysis, then PPD will mark it as a DOCONSIDER loop. If there is no dependence between statements in a loop, or these dependences are loop-independent dependences, different iterations can be executed in parallel on separate processors as DOALL loops. If dependence is occurring across different iterations, i.e., is a loop-carried dependence, it is called a DOACROSS loop. The iterations are executed either sequentially, or partially in parallel by means of enforced synchronization instructions within the bodies of the concurrent loops, and incur some run-time overhead will be incurred. Otherwise, if the loop dependency patterns are too complex to analyze by current algorithms , for example, with non-linear array index expressions or with non-constant dependence distance, then we also can mark it as a DOCONSIDER loop.

Second, because OSF/1 has no FORTRAN compiler and because multithreading only supports C programming, a FORTRAN-to-C (f2c) converter is used to convert the FORTRAN program output by PPD into its C equivalent. Third, the single-to-multiple threads translator (s2m) takes the program obtained from f2c as input, and then generates the output in which the parallel loops (DOALL or DOACROSS) are translated into sub-tasks by replacing them with multithreaded codes. For run-time parallelization, the s2m will generate the inspector and executor codes for DOCONDISER loops at compile-time.

Finally, The resulting multithreaded program is then compiled and linked with the P Threads or C Threads run-time libraries by using the native C compiler, e.g., GNU C compiler. Then, the generated parallel object codes can be scheduled and executed in parallel on the multiprocessors to achieve high performance. Based upon this model, we implemented a FORTRAN parallelizing compiler to help programmers take advantage of multithreaded parallelism on AcerAltos 10000 multiprocessor system, running OSF/1.

2.2 Using Knowledge-based Techniques for Loop Parallelization

Knowledge system is a system that depends on a vast base of knowledge to perform difficult tasks. The knowledge is saved in a knowledge base separately from the inference component. This makes it convenient to append now knowledge or update existing knowledge easily. The rule-based approach is one of the commonly used form in many knowledge-based systems. The primary difficulty in building a knowledge base is how to acquire the desired knowledge. To ease acquisition of knowledge, one primary technique among them is Reper-

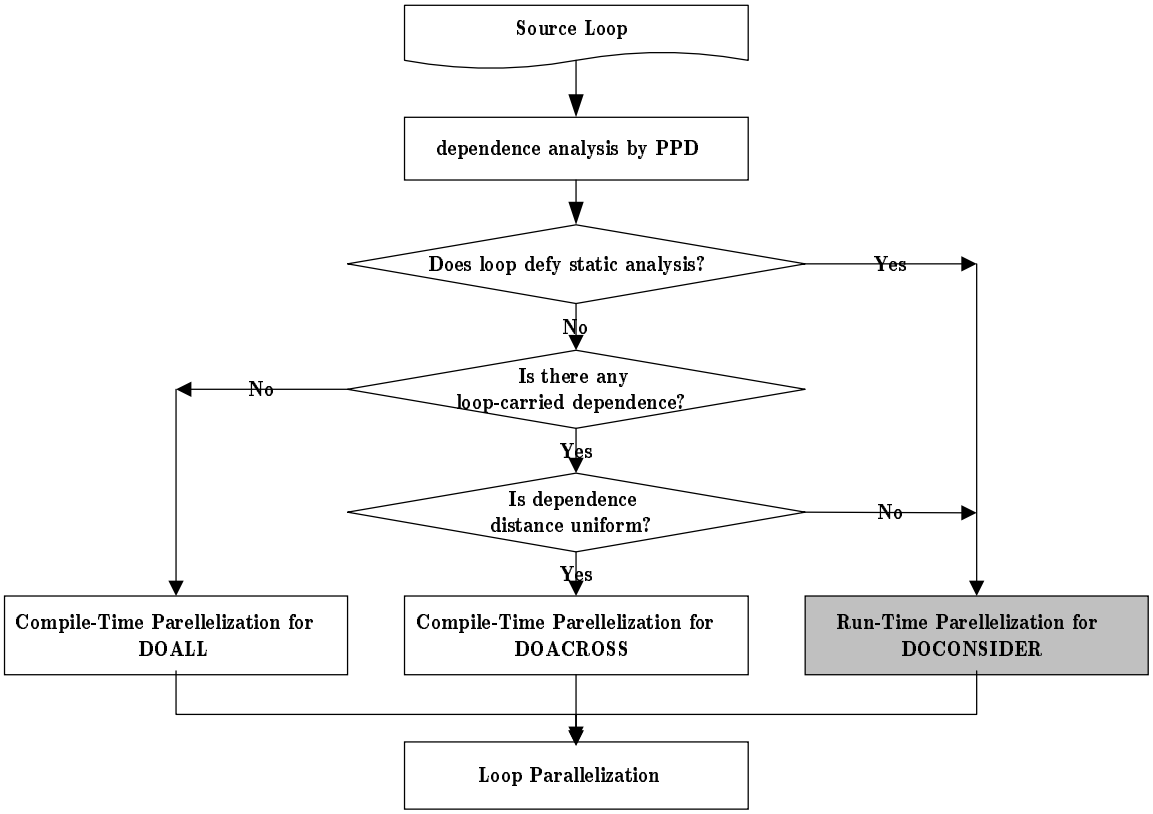

Figure 2: The flow of loop parallelization.

tory Grid Analysis (RGA). RGA is easy to use, but it suffers from the problem of missing embedded meanings. For example, when a doctor expresses the features of catching a cold are headache, cough and sneeze, he may have those features. However, in RGA, a person is not considered to catch a cold except that he gets all of the features. To overcome the problem, the concept of Attribute Ordering Table (AOT) is employed to elicit embedded meanings by recording the importance of each A knowledge-based system is composed of two parts: the development environment and the runtime environment. The former is used to build the knowledge base, while the latter is used to solve the problem. In this paper, the development environment is not discussed here. The runtime environment contains five components, which are briefly described as follows:

The *runtime environment* which using knowledge-based techniques for loop parallelization contains three components as shown in Figure 3, which are briefly described as follows.

- **Knowledge Base:** This component contains knowledge required for solving the problem of determining an appropriate test, scheduling, or transformation to be applied. The knowledge can be organized in many different schemes, and can be encoded into many different forms. Therefore, there exist many choices of building the knowledge base. In our implementation, the knowledge base is constructed as a rule base, i.e., the knowledge is expressed in the form of production rules. These rules can be coded by hand or generated by a translator. In our system, the latter method is used. A translator, GRD2CLP, translates the repertory grid and attribute ordering table to CLIPS's production rules. This approach has great flexibility as we can add new scheduling algorithms to the repertory grid and attribute ordering table, and then use GRD2CLP to convert the tables into CLIPS rules.

- **Inference Engine:** The inference engine is the interpreter of the knowledge stored in the knowledge base. It examines the contents of the knowledge base and the data including the system characteristics and the loop attributes provided by machine architecture and programmers to derive a conclusion, an appropriate parallel loop-scheduling algorithm. The inference engine attempts to find connections between the input attributes stated in section three and the selected loop-scheduling algorithm according to RGA and AOT. An example of applying RGA/AOT is shown in Table 1. 'X' means that the attribute has no relation with the scheduling algorithm. 'D' means that the attribute dominates the scheduling algorithm, i.e., if the attribute is not equal to the entry value, it is impossible for the scheduling algorithm to be implied. For those entries that are not labeled 'X' or 'D', integer numbers are used represented

the relative degree of importance for attribute does not dominate the object but is of some degree of importance relative to other attributes. Larger integer number implies the attribute being more important to the object. According to the table, four rules can be generated. As we observe, [A1, S1]=1,5,6, [A2, S1]=YES, [A3, S1]=X; hence the resulting rule will be generated.
RULE:
If (A1 is in 1,5,6) and (A2=YES) Then Choose S1

Table 1: The repertory grid and the attribute ordering table.

	S1	S2	S3	S4
A1	1,5,6/D	X/X	3/D	2,4/D
A2	YES/D	X/X	YES/D	X/X
A3	X/X	NO/2	NO/D	X/X

- **Algorithm Library:** The library collects several representative tests, transformations, and schedules, either proposed by others or designed by ourselves. The question of how these tests, transformations, and schedules are chosen in the *development environment*, so here we assume that it has been built. For example, we have included eight scheduling algorithms in the library for loop scheduling, that are static scheduling, SS, CSS, GSS, Factoring, TSS, AFS (MAFS, DAFS), and LDS. This is another advantage of using knowledge-based system; we can easily modify the rules and add any new scheduling strategy.

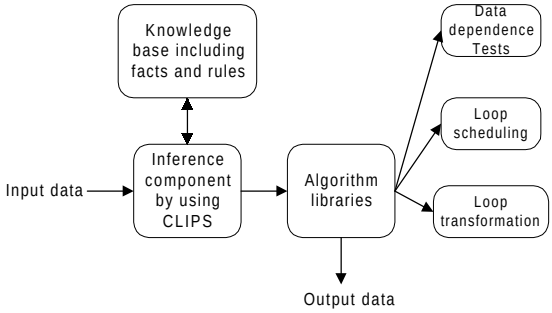

Figure 3: Components of our new model

3 Main Results

3.1 PPD: A Practical Parallel Loop Detector

PPD takes the traditional FORTRAN 77 source program as input and yields the corresponding prompted parallel code. The framework of PPD is divided into two phases, *analysis phase* and *codegen phase*. In analysis phase, a single-subscript testing algorithm, the I test, is used for checking if the linear equation formed by array subscript has an appropriate integer solution. Instead of linearizing the subscript of an array, we check it subscript-by-subscript since there is no certainty that either of them overrides the other in precision. The effect of analysis phase is the determination of the execution modes of all loops. The execution mode of a loop may be the one of the following three types: DOALL, DOACROSS, and DOSEQ, where the former two ones point out that a loop can be executed in a fully or partial parallel manner respectively, and the last one is the normal sequential style. In codegen phase, the outcome of analysis phase is referred to produce the prompted parallel codes. The optimizations for synchronized statements of DOACROSS loops are also taken.

3.2 S2m: A Single-to-Mutiple Threads Translator

The component, single-to-multiple threads translator (s2m), takes the program obtained from f2c as input, and then generates the output in which the parallel loops (DOALL or DOACROSS) are translated into sub-tasks by replacing them with multithreaded codes. The structure of single-to-multiple threads translator (s2m) [10] consists of five modules. The kernel module is written to be portable; it calls functions in thread-code generating module and calls functions in DOALL loop-partition module. and calls functions in DOACROSS loop-partition module through the config module. The thread-code generating module contains several functions that are used to generate different thread specific codes; P Threads or C Threads. The DOALL loop-partition and DOACROSS loop-partition modules contain routines partitioning DOALL and DOACROSS loops, respectively.

We now describe how the s2m converts specific types of conventional sequential programs, i.e., DOALL loops, into their parallel equivalents with the P Threads runtime library codes embedded in them. The general form of a DOALL loop program to s2m is shown in Figure 4. In this figure, there is one for-loop enclosed in "`/* /$DOALL$/ L???: */`" and "`/* /$END_DOALL$/ L???: */`" comments, these two comments are used to indicate the for-loop enclosed by them is a DOALL loop. The `???` here stands for the loop label used in the original FORTRAN program.

```
main()
{
    Variables declaration area
   ...
/* /$DOALL$/???:  */
   for (i= ....  ){
      ...
/* L???:  */
   }
/* /$END_DOALL$/ ???:  */
   ...
}
```

Figure 4: The DOALL loop of input program to s2m.

The output of the main program has the form shown in Figure 5 produced by s2m. There are six rectangles in this figure, each corresponds to a session that performs a specific job. The first session, thread-related definition, outputs thread-related definitions. Some variables for using the

thread package are defined in this session. The `loop` variable is an array of `loop_args`, which is used to pass the begin iteration, end iteration, and the iteration step for each pthread created later on. The `ThCount` variable records the number of threads; this number is decreased by one when a thread is going to be terminated.

```
Thread related definition
Variables declaration area
main()
{
    Mutex & condition variable initialization
    ...
    Iterations calculation
    Fork threads
    Synchronization
    ...
}
```

Figure 5: Main program of the general output produced by s2m.

```
void DOALL??(loop)
struct loop_args *loop;
{
int i;
   for (i=loop->begin; i<=loop->end; i++){
      ...
   }
    Decrease the thread count by 1
}
```

Figure 6: The DOALL function definition for an output thread produced by s2m.

3.3 IPLS: An Intelligent Parallel Loop Scheduling

In PFPC, we propose a system as shown in Figure 7, called intelligent parallel loop scheduling (IPLS), by using knowledge-based techniques to select an appropriate loop-scheduling algorithm. The approach will make good use of the advantages of the algorithms for loop parallelism. By the resulting algorithms for assigning parallel loop on multiprocessor systems, it is believed that the applications can save execution time and achieve high speedup.

- Profile Information - After the program applying the selected loop scheduling algorithm is executed, some information about number of iterations, maximal time of iteration, minimal time of iteration, total time of program, number of synchronization, number of remote memory accesses, and the workload distribution of each processor will be recorded and saved in a profile file. The profile file will be referred to modify the attributes by refining system.
- Refining System - When a program is embedded with some parallel loop scheduling algorithm, if we can refine some attributes, such as the values of factors in the loop-scheduling algorithm by using the profile information derived from the record of executing process of the program. Refining procedure in order to get ideal values will modify the factors. It is obvious that this make the parallelism of program higher and performance better.

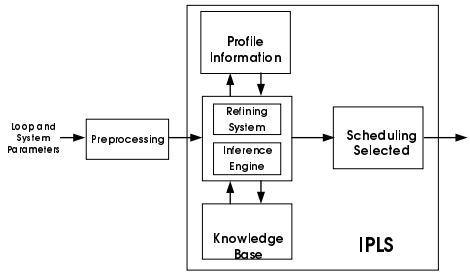

Figure 7: The system architecture of IPLS.

Tables 2 and 3 show the relationships between seventeen attributes and parallel loop scheduling algorithms in UMA and NUMA models respectively. Besides, each table is districted by different kind of loops, i.e., DOALL and DOACROSS loops. The features mentioned above are the attributes upon which we constructed our attribute grid. 'Machine model' is classified into UMA and NUMA. 'Memory access ratio' means the speed ratio of cache, memory and network. 'CPU number' denotes the system size, which can be classified into three levels, small, medium, or large. 'Loop style' includes four kinds of loop, such as U(uniform), I(increasing), D(decreasing) and R(random). 'Program size' shows the appropriate scale that algorithms fit. 'Data locality' determines if loop data behavior has affinity or not. 'Loop boundary' determines if it must be known at compile time. 'Loop level' determines if nested loop is profitable to algorithms. 'Loop carried dependence' is classified into DOALL and DOACROSS. 'Easiness' describes if the implementation of algorithm is easy. 'Factor' means the variables, which can dynamically influence the performance due to loop information and system states. The overheads of synchronization, communication and thread management are roughly classified into four levels, none, light, normal, or heavy. 'Start time' determines whether all each processor starting time need to be equal or not.

In many parallel loop-scheduling algorithms, there are some attributes, such as factors, which influence the performance of executing program. For example, the adaptive hybrid scheduling algorithm has two factor, β and γ, determining the fetching processor whether or not to fetch more iterations form work queue in dynamic level after executing the iterations coming from static level. These two factors, β and γ, should be adjusted by the programmers according to the properties of parallel computers. However, how to select appropriately the value of β and γ on different system is difficult. If we can refine values of the factors in the loop-scheduling algorithm by using the profile information derived

Table 2: The attributive table for UMA models.

UMA Model									
	DOALL								DOACROSS
	Static	SS	CSS	GSS	TSS	Factoring	AHS	SSS	Enhanced CSS
UMA/NUMA	UMA	UMA	UMA	UMA	UMA	UMA	UMA/NUMA	UMA/NUMA	UMA
No of Processor	X	X	X	X	X	X	X	X	X
Memory Access Rate	1:10:200	1:10:200	1:10:100	1:10:200	1:10:200	1:10:200	1:10:200	1:10:200	1:10:200
Loop Style	U, D, I	X	U, R	U, I, R	X	X	X	X	U
Program Size	X	X	Large	X	X	Large	X	X	Large
Loop Type	1-10	X	1-2, 5-7, 9-10	2-3, 7-8	1, 3-4, 7, 11	3-4, 7-8	X	1-3, 5-11	1, 6-7
Data Locality	X	No	No	No	No	No	Yes	Yes	X
Loop Boundary	Yes	X	No	X	X	X	Yes	Yes	X
LCD	DOALL	DOALL	DOALL/(0,1)	DOALL	DOALL	DOALL	DOALL	DOALL	Doacross (>1)
Easiness	X	X	X	No	X	No	No	No	No
Factor	-	-	k	-	N_S, N_F	$x = 2$	β, γ	α, k	K
Thread Overhead	l, n	h	l, n	h	n	n	n, h	n, h	l, n
Comm. Overhead	X	l	l	l	l, n	l	l, n	l, n	l
Sync. Overhead	X	l	2, 3, 4	2	3, 4	3, 4	4, 5	4, 5	3, 4, 5
Start Time	Yes	X	Yes	X	X	X	Yes	X	Yes

Table 3: The attributive table for NUMA models.

NUMA Model								
	DOALL							
	AFS	MAFS	CAFS	LAFS	DAFS	LDS	GDCS	ASS
UMA/NUMA	NUMA	NUMA	NUMA	NUMA	NUMA	NUMA	NUMA	NUMA
No of Processor	S, M	S, M	S, M, L	S, M, L	S	S, M	S, M	S
Memory Access Rate	1:10:200	1:10:200	1:10:200	1:10:200	1:10:200	1:10:200	1:10:200	1:10:200
Loop Style	X	D, I, R	X	X	X	X	X	X
Program Size	-	-	-	X	-	-	-	-
Loop Type	2-3, 9-10	2, 4-5	1, 4, 8	1, 4, 8	2 , 5, 9, 11	1, 5, 6, 8, 10	2, 5, 9, 11	2, 4, 5, 6
Data Locality	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Loop Boundary	X	X	X	X	X	X	Yes	X
Loop Level	X	X	X	X	X	X	X	X
LCD	DOALL	DOALL	DOALL	DOALL	DOALL	DOALL	DOALL	DOALL
Easiness	No	No	No	No	No	No	No	No
Factor	$0.5 < \alpha < 1$	-	K	k	-	B	α, β	$\alpha = \frac{N}{P2}$
Thread Overhead	l, n	l	l, n	l, n	l	l, n	l, n	l
Comm. Overhead	l	l	l	l	l	l, n, h	l, n, h	h
Sync. Overhead	3, 4	5	3, 4	3, 4	5	3, 4	4, 5	3, 4
Start Time	X	X	X	X	X	X	X	X

from the record of executing process of the program, it is obvious that the new factors cyclically modified by refining procedure will make the parallelism of program more clear and make the performance better. And we say this method stated above has feedback-learning ability and is intelligent. In the paper, a refining system based upon the profile information consisting of the following seven items will be included into our model.

- The number of iteration
- Maximal time of iteration
- Minimal time of iteration
- Total time of program
- The number of synchronization
- The number of remote memory access
- The workload distribution of each processor

How to refine attributes and not to modify rules in the knowledge base is a problem, which is solved in our refining system by storing attribute data into a file called Attri_file and using data type of structure (record) as condition testing of antecedent of if statement in rules. When a loop is executed and profile information is generated, the refining system will input profile information to modify the attributes in Attri_file; therefore, the rules in knowledge base does not need to be changed and the inference engine does not need to be recompiled.

There are several situations at which the refining system is suggested. Firstly, when IPLS is constructed completely, maybe the attributes in knowledge base are crude that an optimal loop-scheduling algorithm to transform a sequential program into an efficient multithread program can not be selected. Secondly, when IPLS is ported on a new environment, some attributes about system states, such as memory access rate, need to be changed to influence the selection of scheduling method. In addition, perhaps an appropriate loop-scheduling algorithm is selected by inference engine, but the bad values of factors in algorithm, such as chunk size in CSS, will result in larger execution time. The factors had better be refined to reduce the wasted considerable execution time if the executable code will be executed repeatedly. It seems that the overhead from refining the attributes can be neglected because of its advantage. After all, to increase the accuracy is to increase efficiency. The programmer can determine whether to use refining system before deriving an ideal loop-scheduling algorithm for the program or not. When using the refining system, the programmer can also decide the number of loop-scheduling algorithms selected by inference engine.

3.4 Run-Time Parallelization

The way of parallelizing our general inspector is by partitioning the entire range of iterations into consecutive segments and each segment is assigned to a different thread. Each thread computes a valid parallel schedule for iterations in its segment and ignoring any dependences with other iterations outside of its segment. After all segments have finished, we have a schedule for each segment. Every such schedule,

we called a *sub-schedule*, is a mapping from the iterations in the corresponding segment to the wavefronts of that segment. The overall schedule is formed by concatenating the sub-schedules with the order of the segments in entire range of iterations. The number of segments can be set to a appropriate number, in intuitively, we will set the number of processors to it. But, if the number of segments is larger, then it will increase the total number of wavefronts (i.e., *depth*) in overall schedule. The larger number of wavefronts implies that there are fewer iterations in each wavefront, and then it will decrease the speedup of run-time parallelization.

The executor performs the overall schedule extracted by the general parallel inspector. As a rule of thumb, the executor performs the sub-schedule of each segment in order, i.e., visits the first wavefront till the last wavefront in a segment, then does for next segment's first wavefront closely, go on until the last wavefront of the last segment have been visited. Every wavefront is sequentially executed and ideally, all iterations in the same wavefront are executed concurrently. In practice, iterations in the same wavefront are partitioned into equal-sized chunks and every chunk is enclosed in one thread, the number of threads are automatically adapted according to the number of iterations in each wavefront by calling function `auto-adapted`, and then the threads scheduled by OSF/1 can be executed in parallel manner.

We now compare the methods described in this paper to several other techniques that have been proposed for analyzing and scheduling DO loops at run-time. Most of this work has concentrated on developing inspectors. A high level comparison of the various methods is given in Table 4. Since the process of inspector for finding the wavefronts can be parallelized fully without any synchronization. Our executor can perform the loop iterations concurrently. In addition, for each wavefront in a loop, the `auto-adapted` function is used to get a tailored thread number for optimizing execution.

4 Experimental Results

4.1 Performance of PPD

To evaluate the performance of PPD for PFPC, experiments were performed using both practical and contrived data. The practical data included two numerical packages, LINPACK and EISPACK, while the contrived data included several examples that appeared in other papers. Another program parallelization restructurer, Parafrase-2, was also applied to the same testing data, and the results compared with those from our design. LINPACK and EISPACK are two well-known numerical packages. LINPACK is a collection of FORTRAN subroutines that analyze and solve various systems of simultaneous linear algebraic equations, while EISPACK is a collection of subroutines for evaluating the eigenvalues of matrices. Because of their systemization and representatives, the packages have been widely adopted as benchmark programs [14]. There is total of 256 DO loops distributed across the 52 subroutines in LINPACK. PPD was able to exploit 51 DOALL loops and 0 DOACROSS loops, as was Parafrase-2. In the experiments using LINPACK, we have examined all the DOALL loops detecting by PPD and Parefrase-2 carefully. PPD was able to exploit the same 51 DOALL loops as Parafrase-2 was. Because there is no DO loop that can be translated into DOACROSS loop by using our algorithm in the experiments of LINPACK. So, we show the other experiments for demonstrating the DOACROSS loops detected by using PPD.

There are a total of 657 DO loops distributed across the 77 subroutines in EISPACK. PPD was able to exploit 185 DOALL loops and 7 DOACROSS loops, while Parafrase-2 was able to exploit only the 185 DOALL loops. If there is a constant dependence distance in the loop, PPD will record the information for generating the synchronization statements, and translate that loop into DOACROSS loop during codegen phase. In our version, Parafrase-2 cannot detect the DOACROSS loops, so PPD was able to exploit 7 DOACROSS loops in the experiments using EISPACK, while Parafrase-2 was not. PPD translated the loops into DOALL or DOACROSS loops conservatively. So, it is not possible that PPD mistakenly marks non-DOALL loops as DOALL or non-DOACROSS loops as DOACROSS.

4.2 Performance of IPLS

To demonstrate the performance of IPLS, there are two experimentations on UMA system and NUMA system, the first one concerns each execution time and speedup of above ten applications, and the other is a combined program, including ten applications. Under the implementation on UMA system, which is 2-processor machine, the execution time and the corresponding speedup are shown in Table 5.

GSS performs poorly for Adjoint Convolution because the workload of iterations is decreasing, and TSS is the most efficient algorithm for Adjoint Convolution. CSS/2 is suitable for the applications like Gauss Jordan Elimination with random unbalanced workload, LU Decomposition with decreasing unbalanced workload, and SOR with uniform balanced workload respectively. Factoring scheduling algorithm is suitable for Gauss Elimination with random balanced workload. SSS is suitable for the applications like Reverse Adjoint Convolution with increasing unbalanced workload, All Pairs Shortest Paths with random balanced workload, and Transitive Closure with random unbalanced workload respectively. AHS is suitable for Jacobi Iteration with random unbalanced workload. We can find that none of six scheduling algorithms on UMA system is suitable for all applications. Alternatively, IPLS can choose an appropriate scheduling algorithm and get good performance for most applications except Matrix Multiplication and If_Then application.

5 Conclusions and Further Work

This paper describes the design and implementation of an efficient and parallelizing compiler to parallelize loops and achieve high speedup rates on multiprocessor systems. We first introduce how to design a portable FORTRAN parallelizing compiler (PFPC) on a multiprocessor system by multithreading operating system OSF/1. The main contributions of this paper are described as follows. A model of FORTRAN parallelizing compiler on multithreading OSF/1 was also proposed in this paper. This paper also reported on the practical parallel loop detector (PPD) that was implemented in PFPC on finding the parallelism in loops. Furthermore, if DOACROSS loops are available, an optimiza-

Table 4: Characteristics comparison between several methods. The superscripts have the following meanings: 1, Our serial inspector version can perform an optimal schedule. 2, The `bit-vector` atomic operation must be applied to avoid the use of global synchronization. Since most of parallel machines don't provide this operation, the performance of this run-time method is degraded.

Methods	Get optimal schedule	No sequential portions	No global syn.	No restrict type of loops	No merge between pro.	No large local mem. required	Integrate in compiler
Our Method	No[1]	Yes	Yes	Yes	Yes	Yes	Yes
Zhu and Yew	No	Yes	No	Yes	Yes	Yes	No
Midkiff and Padua	Yes	Yes	No	Yes	Yes	Yes	No
Chen *et al.*	No	Yes	No	Yes	Yes	Yes	Yes
Rauchwerger *et al.*	Yes	Yes	Yes	Yes	No	No	Yes
Saltz *et al.*	Yes	No	No	No	Yes	Yes	Yes
Leung and Zahorjan	Yes	Yes	No	No	Yes	Yes	Yes
Sheng *et al.*	Yes	Yes	Yes[2]	Yes	Yes	Yes	No

Table 5: The execution time (ms)/speedup of 11 applications applying different scheduling algorithms.

Applications	SERIAL	CSS/2	GSS	TSS
Adj_Con	20104/1	15042/1.337	15055/1.335	10398/1.933
Gauss_Eli	365359/1	256945/1.422	197157/1.853	202922/1.8
Gauss_Jor	7765/1	4245/1.829	5587/1.39	5599/1.387
Jacobi_Iter	14047/1	10109/1.39	12836/1.094	12656/1.11
LU	40995/1	28094/1.459	33521/1.223	34356/1.193
Matrix_Mul	23453/1	12281/1.91	12095/1.939	12229/1.918
Radj_Con	27235/1	21274/1.28	14719/1.85	15587/1.747
SOR	109062/1	76891/1.418	82594/1.32	83943/1.299
Spath	63063/1	57032/1.106	58867/1.071	43146/1.462
Tran_Clos	479188/1	298312/1.606	308844/1.552	325430/1.472
If_Then	17125/1	9682/1.769	9693/1.767	8595/1.992

Applications	Factoring	SSS	AHS	IPLS
Adj_Con	13974/1.439	12359/1.627	12352/1.628	as TSS
Gauss_Eli	195016/1.873	208055/1.756	196852/1.856	as Factoring
Gauss_Jor	5266/1.475	4333/1.792	4391/1.768	as CSS/2
Jacobi_Iter	13125/1.07	9802/1.433	9758/1.44	as AHS
LU	33071/1.24	28505/1.438	28432/1.442	as CSS/2
Matrix_Mul	12214/1.92	12187/1.924	12203/1.922	as CSS/2
Radj_Con	15255/1.785	14336/1.9	15477/1.76	as SSS
SOR	86742/1.257	77376/1.41	77680/1.404	as CSS/2
Spath	61547/1.025	38126/1.654	38797/1.625	as SSS
Tran_Clos	310469/1.543	295922/1.619	296078/1.618	as SSS
If_Then	8667/1.976	8656/1.978	8620/1.987	as AHS

tion of synchronization statements are made. Experimental results showed that PPD was more reliable and accurate than previous approaches. In addition, a new model by using knowledge-based techniques was proposed to exploit more loop parallelisms in this paper. The knowledge-based approach integrated existing data dependence tests and loop scheduling algorithms to make good use of their ability to extract loop parallelisms. Experimental results show that the speedup delivered by our compiler was high. As an ultimate goal, a high-performance and portable FORTRAN parallelizing compiler on shared-memory multiprocessors will be constructed. In the study of high-performance parallelizing compilers, results of this paper will be able to deliver theorectical and technical contributions.

References

[1] W. Blume, R. Eigenmann, J. Hoeflinger, and D. Padua, P. Petersen, L. Rauchwerger, P. Tu, "Automatic detection of parallelism: A grand challenge for high-performance computing," *IEEE Parallel & Distributed Technology*, 2(3):37-47, Fall 1994.

[2] J. Boykin, D. Kirschen, A. Langerman, and S. LoVerso, *Programming under Mach*, Addison Wesley, 1993.

[3] K. D. Cooper *et al.*, "The ParaScope parallel programming environment," *Proc. IEEE*, 81(2):244-263, Feb. 1993.

[4] Y. W. Fann, C. T. Yang, C. J. Tsai, and S. S. Tseng, "IPLS: An intelligent parallel loop scheduling for multiprocessor systems," *Proc. of ICPADS'98*, Tainan, Taiwan, pp. 7751-782, Dec. 1998.

[5] L. Rauchwerger, N. M. Amato, and D. Pauda, "Run-time methods for parallelizing partially parallel loops," *Proc. 1995 Int'l. Conf. Supercomputing*, Barcelona, Spain, July 1995.

[6] R. P. Wilson *et al.*, "SUIF: An infrastructure for research on parallelizing and optimizing compilers," *ACM SIGPLAN Notices*, 29(12):31-37, Dec. 1994.

[7] M. Wolfe, *High-Performance Compilers for Parallel Computing*, 137-162, Addison-Wesley Publishing, New York, 1996.

[8] C. T. Yang, S. S. Tseng, and C. S. Chen, "The anatomy of parafrase-2," *Proceedings of the National Science Council Republic of China (Part A)*, 18(5):450-462, Sep. 1994.

[9] C. T. Yang, S. S. Tseng, C. D. Chuang, and W. C. Shih, "Using knowledge-based techniques on loop parallelization for parallelizing compilers," *Parallel Computing*, 23(3):291-309, May 1997.

[10] C. T. Yang, S. S. Tseng, and M. C. Hsiao, "A model of parallelizing compiler on multithreading operating systems," *Int'l. J. of Modelling and Simulation*, 18(1):9-15, 1998.

[11] C. T. Yang, S. S. Tseng, M. H. Hsieh, and S. H. Kao, "An efficient run-time parallelization for do loops," *J. of Info. Sci. and Eng. — Special Issue on Compiler Techniques for High-Performance Computing*, vol. 14, no. 1, pp. 237-253, 1998.

[12] C. T. Yang, S. S. Tseng, M. C. Hsiao, and S. H. Kao, "A portable parallelizing compiler with loop partitioning," *Proc. of the NSC ROC (A)*, vol. 23, no. 6, pp. 751-765, 1999.

[13] C. T. Yang, C. T. Wu, and S. S. Tseng, "PPD: A practical parallel loop detector for parallelizing compilers on multiprocessor systems," *IEICE Trans. Information and Systems*, vol. E79-D, no. 11, pp. 1545-1560, Nov. 1996.

[14] H. P. Zima and B. Chapman, *Supercompilers for Parallel and Vector Computers*, Addison-Wesley Publishing and ACM Press, New York, 1990.