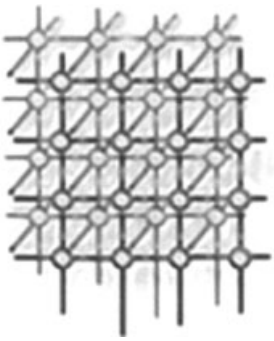
CONCURRENCY AND COMPUTATION: PRACTICE AND EXPERIENCE
Concurrency Computat.: Pract. Exper. (2008)
Published online in Wiley InterScience (www.interscience.wiley.com). DOI: 10.1002/cpe.1334



LC-GRFA: global register file assignment with local consciousness for VLIW DSP processors with non-uniform register files

Chia-Han Lu, Yung-Chia Lin, Yi-Ping You and Jenq-Kuen Lee*, †

Department of Computer Science, National Tsing Hua University, Hsinchu 30013, Taiwan

SUMMARY

Embedded processors developed within the past few years have employed novel hardware designs to reduce the ever-growing complexity, power dissipation, and die area. Although using a distributed register file architecture is considered to have less read/write ports than using traditional unified register file structures, it presents challenges in compilation techniques to generate efficient codes for such architectures. This paper presents a novel scheme for register allocation that includes global and local components on a VLIW DSP processor with distributed register files whose port access is highly restricted. In the scheme, an optimization phase performed prior to conventional global/local register allocation, named global/local register file assignment (RFA), is used to minimize various register file communication costs. A heuristic algorithm is proposed for global RFA to make suitable decisions based on local RFA. Experiments were performed by incorporating our schemes on a novel VLIW DSP processor with non-uniform register files. The results indicate that the compilation based on our proposed approach delivers significant performance improvements, compared with the solution without using our proposed global register allocation scheme.
Copyright © 2008 John Wiley & Sons, Ltd.

Received 20 October 2007; Revised 8 January 2008; Accepted 12 February 2008

KEY WORDS: register allocation; ping-pong register file; DSP; VLIW

*Correspondence to: Jenq-Kuen Lee, Department of Computer Science, National Tsing Hua University, Hsinchu 30013, Taiwan.
†E-mail: jklee@cs.nthu.edu.tw

Contract/grant sponsor: NSC; contract/grant numbers: 96-2218-E-007-013, 96-2220-E-007-030 and 96-2220-E-007-031
Contract/grant sponsor: Ministry of Economic Affairs; contract/grant number: 96-EC-17-A-01-S1-034

Copyright © 2008 John Wiley & Sons, Ltd.

1. INTRODUCTION

Embedded processors developed in the recent years have attempted to employ novel hardware designs to reduce the ever-growing complexity, power dissipation, and die area. Although using a distributed register file architecture with hierarchical access constraints is considered to have less read/write ports than using traditional unified register file structures, conventional compilation techniques cannot produce the optimal performance from such new register file organizations. Our research work is focused on compiler schemes for a VLIW DSP with distributed register files, known as parallel architecture core (PAC) architectures [1].

The appearance of multiple banks of register files, distributed register clusters, and ping-pong architectures on embedded VLIW DSPs (e.g. PAC architectures) presents a great challenge for compilers attempting to generate efficient codes for multimedia applications. Current research results on compiler optimizations for such problems in the literature have been limited to addressing issues related to cluster-based architectures. This includes work on partitioning register files to work with instruction scheduling [2], loop partitions for clustered register files [3], and cluster register files [4]. This complex optimization issue for embedded DSPs has been addressed previously [5] only in the layer of copy propagation optimizations, and attempts have been made to deal with software pipelining issues with distributed register files [6]. In this paper, we address the issues related to global register allocation (GRA).

This study relates to a compiler based on a PAC DSP [1,7–9] designed with distinctively banked register files where port access is highly restricted. The PAC DSP employs a heterogeneous design with a single scalar unit (for light-weight arithmetic, address calculation, and program flow control) plus two data-stream-processing clusters, each of which contains a load/store unit and an ALU/MAC unit with powerful SIMD capabilities; every unit in the clusters collocates three diverse types of register files, providing different accessing methods and constraints, and the scalar unit has its own accessible register file. The major specialization of the register file architectures in the PAC DSP is the incorporation of a so-called *ping-pong register file structure* [10]. This is divided into two banks, each of which can only be accessed mutually exclusively, with a semi-centralized register file among clusters and functional units within a cluster. This design decreases the power consumption because there are fewer port connections, but its clustered nature makes register access across clusters problematic, and the switched access nature of the ping-pong register file prompts the investigation of further register assignment methods for increasing the degree of instruction-level parallelism.

This paper presents a novel scheme for register allocation comprising GRA and local register allocation (LRA) on a VLIW DSP processor with distributed register files whose port access is highly restricted. In the scheme, a phase performed prior to conventional GRA/LRA, named global/local register file assignment (RFA), is introduced to minimize various register file communication costs. Our study is based on the open research compiler (ORC) software framework, in which register allocation comprises GRA and LRA. There are two kinds of temporary name (TN): local and global. LRA manages to allocate registers for local TNs, whose liveness is bounded within a single basic block, and GRA does this for global TNs (GTNs), whose liveness will cross basic blocks. GRA proceeds before LRA in the back-end of the ORC. For featured register file structures where each cluster contains heterogeneous register files, the conventional register allocation scheme with the cluster assignment has to be enhanced to cope with both inter-cluster and intra-cluster

Copyright © 2008 John Wiley & Sons, Ltd.

DOI: 10.1002/cpe

communications. Owing to the potential heavy influences of global RFA (GRFA) on local RFA (LRFA), a heuristic algorithm is proposed where GRFA makes suitable decisions on communication for LRFA. Experiments were performed with a compiler developed based on the ORC, with the results indicating that compilation based on proposed approach delivers significant performance improvements, compared with the solution without using our proposed GRA scheme.

The remainder of this paper is organized as follows. In Section 2, we introduce the processor architecture and register file structure of the PAC VLIW DSP. The proposed register allocation scheme involving RFA is covered in Section 3. Section 4 compares register allocation schemes, and Section 5 presents related work on cluster assignment and our previous work on the processor. Finally, conclusions are drawn in Section 6.

2. PAC DSP ARCHITECTURE

The PAC DSP features a clustered VLIW architecture that boosts scalability, and a large number of registers arranged as innovative heterogeneous and distinct partitioned register file structures. Unlike the symmetric architectures of most DSPs available nowadays, the PAC DSP is constructed as a heterogeneous five-way issue VLIW architecture, comprising two integer ALUs (I-unit), two memory load/store units (M-unit), and the program sequence control unit/scalar unit (B-unit) that is mainly in charge of control-flow instructions such as branch and jump. Each unit has its own executable subset of the instruction set, and each executable instruction is subject to its own register accessibility constraints. The M- and I-units are organized in pairs, with each pair containing one M-unit and one I-unit in a cluster with associated register files. Each cluster is logically appropriate for processing one data stream, and the current design of the PAC DSP consists of two clusters to support a maximum workload capacity of two concurrent data streams. However, the scalability of the cluster design in the PAC DSP could allow the processor to use more clusters to increase the data-processing workload. The B-unit consists of two subcomponents—the program sequence control unit and the scalar unit—due to the hierarchical decoder design for variable-length instruction encoding in the PAC DSP. The program sequence control unit is primarily responsible for the operations of control-flow instructions. The scalar unit, which is capable of simple load/store and address arithmetic, is placed separately from data-stream-processing clusters, with its own register file. The overall architecture is presented in Figure 1.

Registers in the PAC DSP are organized into several distinct partitioned register files and placed as cluster structures to reduce wire connections between functional units and registers so as to minimize chip area and power consumption. Local register files are attached to all units in the processor, including the R-register file, AC-register file, and A-register file, which are only accessible by the B-, I-, and M-units, respectively. Global register file named the D-register file is designed to be shared by the pair of M- and I-units in each cluster. The D-register file is further partitioned into two banks to utilize instructional port switching technology in order to further reduce wire connections between the M- and I-units. This technology is the *ping-pong register file structure*, which decreases the register bank port connections that limit the accessibility of the two banks; in each cycle, the two functional units must access different banks. Each instruction bundle encodes the information of which bank is to be accessed by each functional unit in the cycle, so that the hardware can perform port switching between D-register file banks and functional units to facilitate data sharing

Copyright © 2008 John Wiley & Sons, Ltd.

DOI: 10.1002/cpe

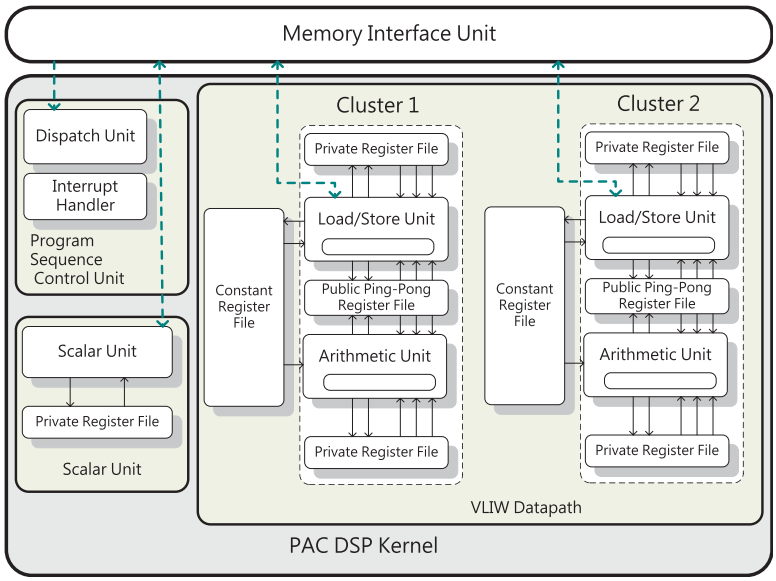


Figure 1. PAC DSP architecture.

within a cluster. Overlapping two data-stream operations in a cluster minimizes occasions when M- and I-units access the same data at the same time, and hence the access constraints of the ping-pong register file structure should have little impact on the performance. The ping-pong register file structure is designed to consume less power because of the reduced number of read/write ports [11] while retaining the data communication capability. Each cluster also contains an additional constant register file that is shared by both M- and I-units as one of the read-only operand sources usable by certain instructions. Only M-units can initialize the data in the constant register file.

3. LOCAL-CONSCIOUS GRFA

This section describes our proposed local-conscious GRFA (LC-GRFA) employing LRFA to guide the assignment. We also introduce our previously proposed ping-pong-aware-local-favorable LRFA (PALF-LRFA) [12,13], acting as LRFA in the later examples and experiments.

3.1. Motivation and RFA

Because the PAC DSP has a clustered organization with non-uniform distributed register files, the more common 'cluster assignment' problem often seen in code generation for VLIW processors [14–16] does not directly apply to this design. Communications in the PAC DSP can be divided into intra-cluster and inter-cluster communications. Intra-cluster communication, which occurs within a single cluster, uses the public register file as the communication channel. For functional unit FU_a, if there is an operand in the same cluster that it cannot access (i.e. belonging to the private register file of another functional unit, called FU_b), intra-cluster communication is required,

Copyright © 2008 John Wiley & Sons, Ltd.

DOI: 10.1002/cpe

and hence data must be copied from the private register file of FU_b into the public register file to make it accessible to FU_a. Furthermore, inter-cluster communication will also be required if an operand needed by FU_a is on the other cluster. In the case of inter-cluster communication, one cluster must issue a *send* instruction and the other one must issue a *receive* instruction in the same bundle (one of the hardware constraints) to complete the inter-cluster communication.

Figure 2 illustrates the complicated communication in the PAC DSP. In Figure 2(a) the original instruction is an addition and is unbundlized, whereas in Figure 2(b) the instruction has its TN register files assigned. Figure 2(c) has a bundle where the instruction is bound to **I2**. However, because **I2** cannot access *1A*, inter- and intra-cluster communication is required. Thus, in Figure 2(d), a bundle that contains *send*/*receive* instructions is inserted for inter-cluster communication, which copies from *1A* to *2A*, while a *nop* bundle has to follow to meet the latency requirement of these instructions. Although there is an inter-cluster copy, intra-cluster communication is necessary because *2A* is still out of reach of **I2**. This leads to a bundle including a copy, as shown in Figure 2(e).

Complicated communication scheme in the PAC DSP makes it desirable to use a new phase, RFA, to handle communications. RFA includes three major subphases: CA, register file class assignment, and copy insertion. The first two subphases handle inter- and intra-cluster communications and assign suitable register files to TNs, whereas the third subphase inserts and unfolds copy instructions

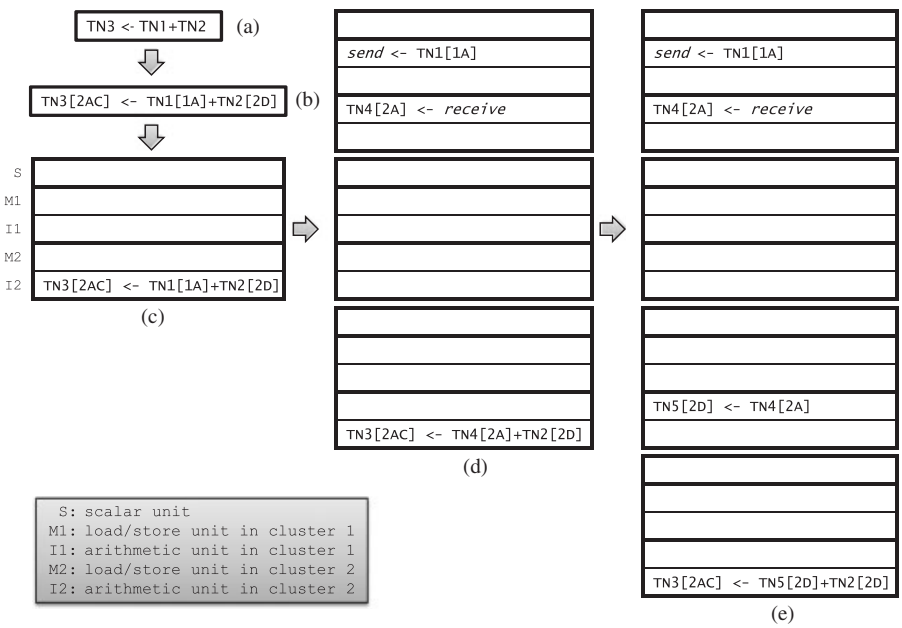


Figure 2. Example of inter-/intra-cluster communication in the PAC DSP: (a) original instruction, (b) each TN has its register file assigned, (c) bundlized, (d) after inter-cluster communication insertion, and (e) after intra-cluster communication insertion. In the figure, *A* is the private register file of the M-unit, *AC* is the private register file of the I-unit, and *D* is the public register file, and *A*, *AC*, and *D* are together defined as the *register file class* of TN, which are prefixed with a number representing the cluster. Thus, *1D* is the public register file in cluster 1, and *2AC* is the private register file of the I-unit in cluster 2.

Copyright © 2008 John Wiley & Sons, Ltd.

DOI: 10.1002/cpe

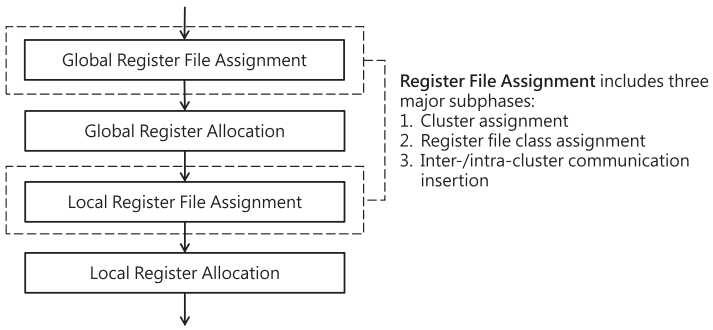


Figure 3. Register allocation scheme involving register file assignment with three major subphases.

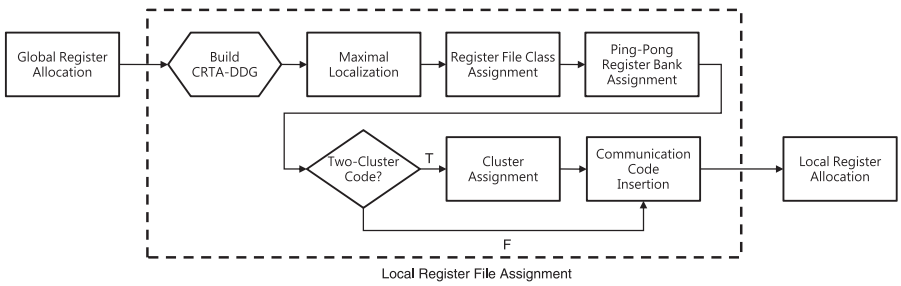


Figure 4. Flowchart of PALF scheme.

to establish the communications derived in the other subphases. The PAC DSP compiler, derived from the ORC, also divides register allocation into GRA and LRA, which are register allocations for GTNs and TNs, respectively. Owing to various differences between GRA and LRA, the mentioned RFA has to be extended into GRFA and LRFA, as shown in Figure 3.

3.2. PALF-LRFA

This section reviews our previously proposed PALF-LRFA that, given a dependency directed acyclic graph (DAG) [17] describing the compilation regions, heuristically determines the appropriate register file/bank assignment and employs state-of-the-art graph-coloring register allocation for each assigned register file/bank in PAC architectures.

The flow of the PALF-LRFA is shown in Figure 4. Our approach requires building an extended data dependence DAG, which preserves the information of the execution and storage relationship for hierarchical access constraint analysis, in addition to the original partial order imposed by instruction precedence constraints. Nodes in the extended data dependence DAG represent instructions of the input code block, with the component-type association (that indicates which functional unit should preferentially be scheduled for this node) and the register-type association (that annotates

Copyright © 2008 John Wiley & Sons, Ltd.

DOI: 10.1002/cpe

the appreciated physical register file/bank, to where the operands/results will be allocated); the edges linked between the nodes represent data dependency that serializes the execution order to be followed in the scheduled code sequence. The PALF-PALF could be organized into the following five phases:

1. Perform the preferable functional unit assignment to the default execution type of each instruction using *maximal localization* analysis.
2. Assign operands/results (required to be allocated to physical registers) of each node in the extended data dependence DAG to the desirable register files.
3. Partition the operands/results assigned to the global ping-pong register files to the preferred register banks according to the strategy of optimizing ping-pong parallelism.
4. Optionally partition the nodes in the extended data dependence DAG into two clusters properly if we would like to compile for two clusters.
5. Insert nodes of the required communication code to avoid invalid register file/bank assignment and cluster partitioning, followed by the physical register allocation for each register file.

3.3. Proposed method for GRFA

Owing to the variety of distributed register files involved and hierarchical access constraints in the design of PAC DSP architectures, we developed the PALF-LRFA before the actual LRA, allowing state-of-the-art graph-coloring-based methods for allocating homogeneous registers in the same register file. This design simplifies the optimization of register allocation for the non-uniform distributed register files used in the PAC DSP, but achieves adequate performance.

Although the proposed PALF-LRFA performs well with the local optimizations and instruction scheduling, GRFA cannot be directly derived from the PALF-LRFA as the original design decisions in the heuristics are based on the data dependence DAG. Therefore, we add a new phase of GRFA, named LC-GRFA, which refers to the local preference of the PALF-LRFA for each basic block and determines the appropriate RFA for GTNs across multiple blocks. Our proposed LC-GRFA extends the PALF-LRFA to obtain RFA of GTNs that is advantageous for both LRFA and local instruction scheduling. We now introduce two LC-GRFA methods: one-pass and two-pass.

The one-pass GRFA is proposed in Algorithm 1, which includes the following steps:

1. *Basic block prioritization*: The prioritization will prioritize basic blocks in the programming unit based on static or dynamic score. The static score is calculated based on factors such as the loop depth and scheduling length, whereas the dynamic score is based on profiling. Basic blocks with higher scores have a greater impact on overall performance, so they should be given higher priority to be processed earlier.
2. *RFA*: This step employs LRFA virtually on basic blocks in the order of priority until all basic blocks are traversed or all GTNs have register files. For each iteration, if there are previous assignments to GTN register files, the iteration uses them as a precondition for the LRFA to be employed. After the LRFA, the step preserves GTN register files given by LRFA but resets TN register files to unassigned. Because some phases in GRA will produce new TNs, such as spilling, it is preferable to leave all TN register files to be assigned in LRFA after GRA.
3. *Communication insertion*: The third step involves simply inserting intra- or inter-cluster communication codes so as to make all operations legal.

Copyright © 2008 John Wiley & Sons, Ltd.

DOI: 10.1002/cpe

Algorithm 1 One-pass LC-GRFA

Require: $\mathsf{BB\_List_{all}}$ and $\mathsf{GTN\_List_{all}}$
Ensure: Register file assignment of all GTNs
  {*Step 1: Basic block prioritization*}
  $\mathsf{BB\_List_{prio}} \leftarrow$ Prioritize the basic blocks in $\mathsf{BB\_List_{all}}$;
  {*Step 2: Register file assignment*}
  while ($\mathsf{BB\_List_{prio}}$ is not empty) or ($\mathsf{GTN\_List_{all}}$ is not empty) **do**
    $\mathsf{BB_{prio}} \leftarrow$ Pop the basic block from $\mathsf{BB\_List_{all}}$ with the highest priority;
    $\mathsf{GTN\_List_{fixed}} \leftarrow$ All the global temporary names whose register files have been fixed in $\mathsf{GTN\_List_{all}}$;
    $\mathsf{Reg\_File\_Assign_{prio}} \leftarrow$ *Local_Reg_File_Assign*($\mathsf{BB_{prio}}$, $\mathsf{GTN\_List_{fixed}}$);
    for all $\mathsf{GTN_{prio}} \in$ *GTN_List*($\mathsf{Reg\_File\_Assign_{prio}}$) **do**
      Fix register file of $\mathsf{GTN_{prio}}$;
    end for
  end while
  {*Step 3: Communication insertion*}
  for all $\mathsf{BB} \in \mathsf{BB\_List_{all}}$ **do**
    Insert_Comm(BB);
  end for

In addition to the one-pass method, we propose the two-pass method that employs separate two passes of register file class assignment and CA. As either passes refer LRFA to make their assignment, the method is considered to be advantageous for local optimizations in some cases. This method is displayed in Algorithm 2, which mainly includes the following passes:

1. *Register file class assignment*: In the first pass, the method manages to assign register file classes to all GTNs, such as *A*, *AC*, and *D*. After collecting LRFA of each basic blocks, the method assigns each GTN register file class based on whether a GTN is 'consistent' or not. A GTN is considered 'consistent' in register file class assignment if LRFA of each basic block assigns the GTN to the same register file class. If this condition occurs, it indicates that LRFA may agree with the consistent register file class, and hence the method assigns the GTN to the class. Otherwise, it assigns the GTN to the public register file class, as the disagreement suggests that intra-communication may be required.
2. *Cluster assignment*: After all GTN register file classes are assigned, CA is performed in the second pass, which operates like the one-pass method.

3.4. Examples of LC-GRFA

This section provides simple examples to illustrate the current version of the proposed LC-GRFA methods. Figure 5(a) shows an example of the one-pass method with 10 GTNs in the programming unit, which are listed in the left most column. And each assignment of GTN comprises two parts: the cluster and the register file class. For example, an assignment of a GTN to *D* in cluster 2 is represented by *2D*. The one-pass method is based on the basic block priority. Referring to Figure 5(a), in Step 1 the method chooses BB5 (which has the highest priority), which references

Copyright © 2008 John Wiley & Sons, Ltd.

DOI: 10.1002/cpe

Algorithm 2 Two-pass LC-GRFA

Require: $\mathsf{BB\_List_{all}}$ and $\mathsf{GTN\_List_{all}}$
Ensure: Register file assignment of all GTNs
{*Pass 1: Register file class assignment*}
{*1.1: Local register file assignment collection of basic blocks*}
Reg_File_Class_Assign_List ← *Empty*;
for all BB ∈ $\mathsf{BB\_List_{all}}$ **do**
Reg_File_Class_Assign_List [BB]
← *Reg_File_Class_Assign*(*Local_Reg_File_Assign*(BB));
end for
{*1.2: Register file class assignment*}
for all GTN ∈ $\mathsf{GTN\_List_{all}}$ **do**
Reg_File_Class ← *Consistent_Reg_File_Class*(GTN, Reg_File_Class_Assign_List);
if (Reg_File_Class is *Nothing*) **then**
Reg_File_Class(GTN) ← Reg_File_Class;
else
Reg_File_Class(GTN) ← *Public Register File Class*;
end if
Fix register file class of GTN;
end for
{*Pass 2: Cluster assignment*}
{*2.1: Basic block prioritization*}
$\mathsf{BB\_List_{prio}}$ ← Prioritize the basic blocks in $\mathsf{BB\_List_{all}}$;
{*2.2: Cluster assignment*}
while ($\mathsf{BB\_List_{prio}}$ is not empty) or ($\mathsf{GTN\_List_{all}}$ is not empty) **do**
$\mathsf{BB_{prio}}$ ← Pop the basic block from $\mathsf{BB\_List_{all}}$ with the highest priority;
$\mathsf{GTN\_List_{fixed}}$ ← All the global temporary names whose clusters or cluster file classes have been fixed in $\mathsf{GTN\_List_{all}}$;
$\mathsf{Cluster\_Assign_{prio}}$ ← *Cluster_Assign*(*Local_Reg_File_Assign*($\mathsf{BB_{prio}}$, $\mathsf{GTN\_List_{fixed}}$));
for all $\mathsf{GTN_{prio}}$ ∈ *GTN_List*($\mathsf{Cluster\_Assign_{prio}}$) **do**
Fix cluster of $\mathsf{GTN_{prio}}$;
end for
end while
{*Communication insertion*}
for all BB ∈ $\mathsf{BB\_List_{all}}$ **do**
Insert_Comm(BB);
end for

four GTNs: GTN4, GTN5, GTN6, and GTN12. The method performs LRFA and assigns the GTNs to *1D* and *2D*. BB6 chosen in Step 2 references GTN5, GTN7, GTN8, GTN9, and GTN11. Because GTN5 has been assigned in Step 1, the GTN cannot be modified and used as a precondition for the subsequent LRFA. Hence, LRFA is performed in Step 2 using the assignment of GTN5 as a precondition, with GTN7, GTN8, GTN9, and GTN11 being assigned. This process continues in Steps 3 and 4. Finally Step 4, there are no more unassigned GTNs, and hence the method terminates.

Copyright © 2008 John Wiley & Sons, Ltd.

DOI: 10.1002/cpe

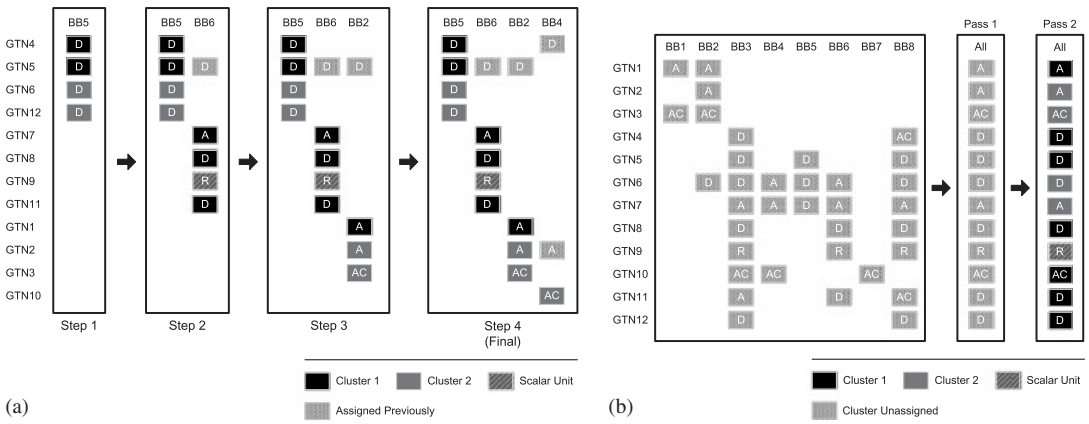


Figure 5. Examples of LC-GRFA. (a) Example of one-pass LC-GRFA. In the figure, each column is headed by the basic blocks involved in each step and (b) example of two-pass LC-GRFA.

Figure 5(b) presents an example of the two-pass method. First of Pass 1 involves performing LRFA on each basic block independently and collecting the assignment of the register file class of each basic block as a reference. For example, BB2 references four GTNs: GTN1, GTN2, GTN3, and GTN6. LRFA produces an assignment of *A* for GTN1 and GTN2, *AC* for GTN3, and *D* for GTN6. Pass 1 also makes a compromise among these assignments. For example, based on our method, Pass 1 assigns GTN1 to *A*, GTN6 to *D*, and GTN10 to *AC*. This is followed by Pass 2 performing CA based on Pass 1, which assigns GTN1 to *1A*, GTN6 to *2D*, and GTN10 to *1AC*.

4. EXPERIMENTS AND DISCUSSION

This section describes our preliminary experiments on the proposed LC-GRFA methods. Figure 6(a) compares various prioritization methods and (b) compares various register allocation schemes, using the DSPstone benchmark suite [18] in the PAC DSP. Figure 6(c) compares register allocation schemes on applications, including G.723, G.727, and programs from MiBench [19]. All the experiments were performed with our developed compiler based on ORC infrastructures and evaluated with a cycle-accurate instruction set simulator of the PAC DSP. The simulator is provided by SoC Technology Center of Industrial Technology Research Institute (STC/ITRI) in Taiwan.

Our method is based on prioritization of basic blocks, which influences the performance. Figure 6(a) compares among three prioritization methods: block length, random, and frequency. The results for the random method are the averages of 10 runs of the benchmarks. The frequency method refers to using the frequency utility provided by the ORC. Related to block-length method, the random method exhibits impaired performance, whereas the frequency method improves performance by 7% on average.

In Figure 6(b), the following register allocation schemes with different RFA combinations were compared: (1) O1 with PALF-LRFA, (2) O2 with PALF-LRFA plus naïve GRFA, (3) O2 with PALF-LRFA plus the one-pass LC-GRFA, and (4) O2 with PALF-LRFA plus the two-pass LC-GRFA.

Copyright © 2008 John Wiley & Sons, Ltd.

DOI: 10.1002/cpe

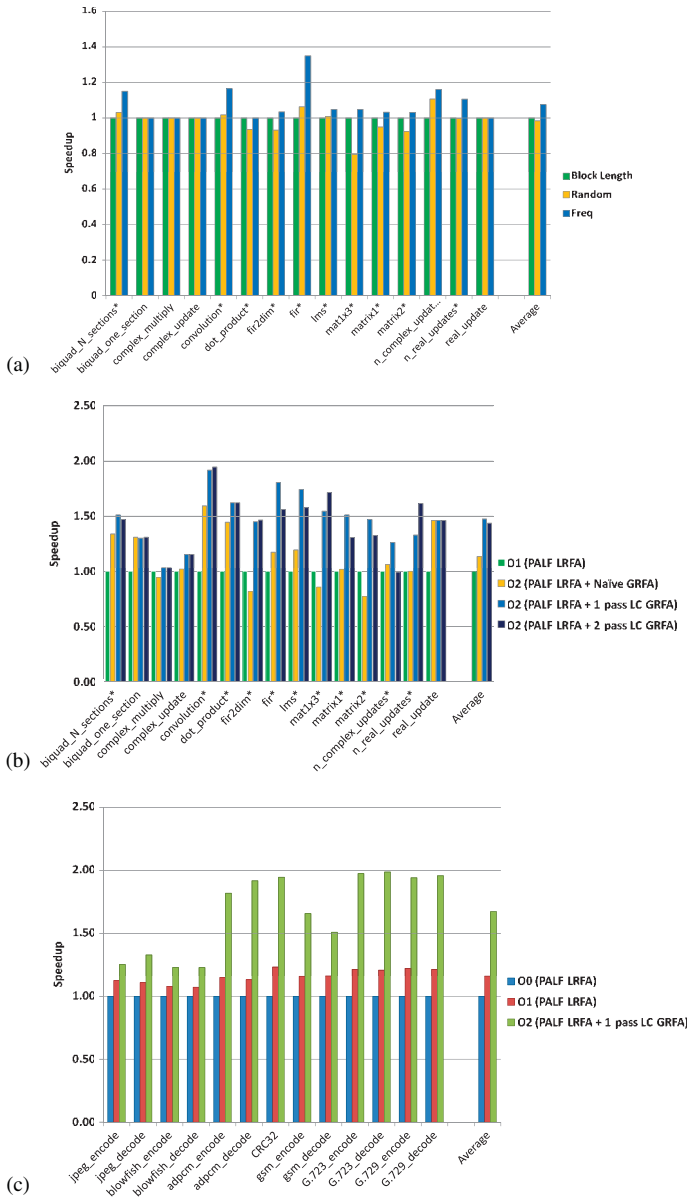


Figure 6. Experiment results. (a) Comparison of various prioritization methods; (b) comparison of register allocation schemes on DSPStone; and (c) comparison of register allocation schemes on applications.

Copyright © 2008 John Wiley & Sons, Ltd.

DOI: 10.1002/cpe

Note that O2 disables the software pipelining in our compiler, whereas O2 enables it in original ORC. Although PALF-LRFA and LC-GRFA methods have been explained in the earlier part, the naïve GRFA assigns all GTN register files to one single cluster. In other words, it utilizes only one cluster and leaves the other idle. As a whole, the comparison indicates that the last two schemes provide 45% speedup on an average relative to the PALF-LRFA and 30% speedup on an average relative to the PALF-LRFA plus naïve GRFA. Although most of the programs in DSPstone benchmark suite have fine speedup using the scheme, some disclose flaws of our methods. First, in some program, such as matrix1 and matrix2, the one-pass LC-GRFA beats two-pass method though the one-pass LC-GRFA and two-pass LC-GRFA have close speedup on average. One possibility is that RFCA in the two-pass LC-GRFA does not involve priority and it leads to less sensitiveness of the two-pass LC-GRFA to LRFA than the one-pass LC-GRFA. Second, *n_complex_updates*, where the two-pass method degrades, reveals another issue in compiler development for the PAC DSP. For *n_complex_updates*, some register files are highly utilized in the proposed two-pass GFRA and leads to considerable spilling in the LRA, which degrades the performance for the benchmark. This could be solved using register pressure control in every register file in RFA to provide appropriate apportionment between LRA and GRA. As the number of registers in register files on the PAC DSP is limited, the control relies on precise modeling of potential register usage scenarios.

Figure 6(c) presents overall performance of our compiler on the applications. These optimizations are: (1) O0 with PALF-LRFA, (2) O1 with PALF-LRFA, and (3) O2 with PALF-LRFA plus the one-pass LC-GRFA. In the experiment, we used the one-pass LC-GRFA, which is more stable than the two-pass GRFA. The figure shows that O1 and O2 can achieve 116 and 167% speedup on average, respectively.

5. RELATED WORK

Hiser *et al.* have described global register partitioning for CA in [4], where a register component graph (RCG) is built on data dependence DAGs and an optimal schedule is used to model the relationship between registers. Every node in the RCG represents register operands, and the weight of an edge represents the 'affinity' of the connected nodes: positive weight indicates that the related nodes should be put in the same cluster, and *vice versa*. They used a greedy heuristic algorithm to divide these nodes into two clusters. The experimental results show that their algorithm exhibits a performance degradation of only about 10% when compared with an unrealizable monolithic-register-bank architecture with the same level of ILP. However, their architecture is simpler than the PAC DSP, which is subjected to the ping-pong constraint, and hence it is difficult to implement the algorithm in the PAC DSP.

Terechko *et al.* provided another method for CA of global values [20]. Local values have short-lived ranges, whereas global values can be alive throughout the whole programming unit. Those authors provided several methods for performing the task, one of which is involved using the 'affinity' as mentioned above. The affinity between two global values indicates the benefit of assigning them to the same cluster based on the data flow graph. Equations were used to represent the affinity between two global values, rather than the RCG mentioned in [4]. They also provided detailed experimental comparisons.

Copyright © 2008 John Wiley & Sons, Ltd.

DOI: 10.1002/cpe

Besides PALF, we have previously proposed a simultaneous instruction scheduling/RFA method based on simulated annealing [21], which iteratively optimizes the entire RFA state of a single basic block and uses the instruction scheduler itself as the evaluation function. This local approach was borrowed from [2], and can also be globally extended in the same manner as PALF as presented in this paper.

In addition to work on the register allocation scheme, the copy propagation in the original ORC could also be modified [5]. Owing to the non-uniform distributed register file structure in the PAC DSP, conventional copy propagation might degrade the performance. In the present study, a communication cost model was derived for copy propagation, which was based on the cluster distance, register port pressure, and movement type of register sets. The model was used to guide data flow analysis for better performance on the PAC DSP architecture. This scheme is effective in preventing the performance anomaly.

6. CONCLUSION

This paper proposes a new register allocation scheme involving a separate phase named RFA, which comprises LRFA and GRFA. Owing to the non-uniform distributed register file structures on PAC DSP architectures, where the conventional register allocation is not sufficient, RFA could form the core procedure of the scheme to boost performance on the architecture. The preliminary experimental results presented here show that the proposed scheme can efficiently utilize the distributed register file architectures and deliver good performance. Our future work will include a complete exploration of the various issues involved in our proposed scheme.

ACKNOWLEDGEMENTS

This work was supported in part by the NSC under grant Nos. 96-2218-E-007-013, 96-2220-E-007-030 and 96-2220-E-007-031, and by the Ministry of Economic Affairs under grant No. 96-EC-17-A-01-S1-034 in Taiwan.

REFERENCES

1. Chang D, Baron M. Taiwan's roadmap to leadership in design. *Microprocessor Report*, In-Stat/MDR, December 2004. Available at: http://www.mdronline.com/mpr/archive/mpr_2004.html.
2. Leupers R. Instruction scheduling for clustered VLIW DSPs. *Proceedings of the 9th International Conference on Parallel Architecture and Compilation Techniques* (*PACT 2000*), IEEE Computer Society Press: Los Alamitos, CA, October 2000; 291–300.
3. Qian Y, Carr S, Sweany P. Optimizing loop performance for clustered VLIW architectures. *Proceedings of the 11th International Conference on Parallel Architectures and Compilation Techniques* (*PACT 2002*), Charlottesville, Virginia, 22–25 September 2002.
4. Hiser J, Carr S, Sweany P. Global register partitioning. *Proceedings of the 9th International Conference on Parallel Architectures and Compilation Techniques (PACT 2000)*, IEEE Computer Society Press: Los Alamitos, CA, October 2000; 13–23.
5. Wu C-J, Chen S-Y, Lee J-K. Copy propagation optimizations for VLIW DSP processors with distributed register files. *Proceedings of the 19th International Workshop on Languages and Compilers for Parallel Computing*, New Orleans, Louisiana, 2–4 November 2006.

Copyright © 2008 John Wiley & Sons, Ltd.

DOI: 10.1002/cpe

6. Chen C-K, Tseng L-H, Chen S-C, Lin Y-J, You Y-P, Lu C-H, Lee J-K. Enabling compiler flow for embedded VLIW DSP processors with distributed register files. *ACM SIGPLAN/SIGBED 2007 Conference on Languages*, *Compilers*, *and Tools for Embedded Systems* (*LCTES'07*), San Diego, June 2007.
7. Lin T-J, Chang C-C, Lee C-C, Jen C-W. An efficient VLIW DSP architecture for baseband processing. *Proceedings of the 21st International Conference on Computer Design*, IEEE Computer Society Press: Los Alamitos, CA, 2003.
8. Lin T-J, Chao C-M, Liu C-H, Hsiao P-C, Chen S-K, Lin L-C, Liu C-W, Jen C-W. Computer architecture: A unified processor architecture for RISC & VLIW DSP. *Proceedings of the 15th ACM Great Lakes Symposium on VLSI*, ACM Press: New York, April 2005.
9. Chang DC-W, Liao I-T, Lee J-K, Chen W-F, Tseng S-Y, Jen C-W. PAC DSP core and application processors. *Proceedings of 2006 IEEE International Conference on Multimedia and Expo*, Toronto, ON, 9–12 July 2006.
10. Lin T-J, Lee C-C, Liu C-W, Jen C-W. A novel register organization for VLIW digital signal processors. *Proceedings of 2005 IEEE International Symposium on VLSI Design*, *Automation*, *and Test*, IEEE Computer Society Press: Los Alamitos, CA, 2005; 335–338.
11. Rixner S, Dally WJ, Khailany B, Mattson P, Kapasi UJ, Owens JD. Register organization for media processing. *International Symposium on High Performance Computer Architecture* (*HPCA*), IEEE Computer Society Press: Los Alamitos, CA, 2000; 375–386.
12. Lin Y-C, You Y-P, Lee JK. PALF: Compiler supports for irregular register files in clustered VLIW processors. *Concurrency and Computation*: *Practice and Experience* 2007; **19**:1–16.
13. Lin Y-C, You Y-P, Lee J-K. Register allocation for VLIW DSP processors with irregular register files. *Proceedings of the 12th Workshop on Compilers for Parallel Computers* (*CPC 2006*), A Coruña, Spain, 9–11 January 2006.
14. Ellis JR. *Bulldog*: *A Compiler for VLIW Architectures*. MIT Press: Cambridge, MA, 1986.
15. Capitanio A, Dutt N, Nicolau A. Design considerations for limited connectivity vliw architectures. *Technical Report TR59-92*, Department of Information and Computer Science, University of California, Irvine, 1993.
16. Özer E, Banerjia S, Conte TM. Unified assign and schedule: A new approach to scheduling for clustered register file microarchitectures. *Proceedings of the 31st Annual International Symposium on Microarchitecture*, Dallas, TX, November 1998; 308–315.
17. Aho AV, Ullman JD, Sethi R. *Compilers Principles*, *Techniques*, *and Tools*. Addison-Wesley: Reading, MA, 1986.
18. Zivojnovic V, Martinez J, Schlager C, Meyr H. DSPstone: A DSP-oriented benchmarking methodology. *Proceedings of the International Conference on Signal Processing and Technology*, DSP Associates: Boston, MA, 1995; 715–720.
19. Guthaus MR, Ringenberg JS, Ernst D, Austin TM, Mudge T, Brown RB. MiBench: A free, commercially representative embedded benchmark suite. *Proceedings of IEEE 4th Annual Workshop on Workload Characterization*, Austin, TX, December 2001.
20. Terechko A, Le Thénaff E, Corporaal H. Cluster assignment of global values for clustered VLIW processors. *Proceedings of the 2003 International Conference on Compilers*, *Architecture*, *and Synthesis for Embedded Systems* (*CASES '03*), 2003; 32–40.
21. Lin Y-C, Tang C-L, Wu C-J, Hung M-Y, You Y-P, Moo Y-C, Chen S-Y, Lee JK. Compiler supports and optimizations for PAC VLIW DSP processors. *Proceedings of the 18th International Workshop on Languages and Compilers for Parallel Computing*, 2005.

Copyright © 2008 John Wiley & Sons, Ltd.

DOI: 10.1002/cpe