Loop Alignment for Memory Accesses Optimization

Antoine Fraboulet, Guillaume Huard, Anne Mignotte

Laboratoire de l'Informatique du Parallélisme
École Normale Supérieure de Lyon
Unité Mixte de Recherche CNRS-INRIA-ENS LYON n° 5668

CENTRE NATIONAL DE LA RECHERCHE SCIENTIFIQUE

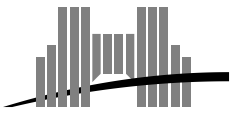
SPI

Loop Alignment for Memory Accesses Optimization

Antoine FRABOULET
Guillaume HUARD
Anne MIGNOTTE

April 1999

Research Report N° 1999–26



École Normale Supérieure de Lyon
46 Allée d'Italie, 69364 Lyon Cedex 07, France
Téléphone : +33(0)4.72.72.80.37
Télécopieur : +33(0)4.72.72.80.80
Adresse électronique : lip@ens-lyon.fr

INRIA

Loop Alignment for Memory Accesses Optimization

Antoine Fraboulet
Guillaume Huard
Anne Mignotte

April 1999

Abstract

Portable or embedded systems allow more and more complex applications like multimedia today. These applications and submicronic technologies have made the power consumption criterium crucial. We propose new techniques thanks to which we can optimize the behavioral description of an integrated system before the hardware/software partitioning (*Codesign*). These transformations are performed on "for" loops that constitute the main parts of the multimedia code which handle the arrays. We present in this paper two new (polynomial) techniques for minimizing memory accesses in loop nests by data temporal locality optimization.

Keywords: Memory Optimization, Code Transformation, Codesign, Loop Alignment (Folding)

Résumé

Les systèmes portables ou embarqués supportent des applications toujours plus complexes comme aujourd'hui le multimédia. Ces applications et les technologies submicroniques ont rendu le critère de la consommation incontournable. Nous proposons de nouvelles techniques permettant d'optimiser la description comportementale d'un système intégré avant le partitionnement matériel-logiciel (*Codesign*). Ces transformations sont effectuées sur les boucles "for" qui sont les principales parties du code multimédia manipulant les tableaux. Nous présentons dans ce rapport deux nouvelles techniques (polynomiales) pour minimiser les accès à la mémoire dans les nids de boucles par optimisation de la localité temporelle des données.

Mots-clés: optimisation mémoire, transformation de code, conception conjointe (codesign), alignement de boucles

Contents

List of Figures

1 Introduction

The design of embedded or integrated systems has become more and more complex, for instance with the appearance of multimedia and data dominated applications. This type of applications consumes a lot of memory for multidimensional data storage like images, sound or video. Thus more than half of the surface of the integrated systems of this kind of application is filled by memory. This massive memory usage combined with submicronic technologies have made power consumption criteria control compulsory. Manual experimentations [Bro98] have shown important consumption gains by code transformations on the algorithmic description of the design (MPEG4 experimentations have allowed a decrease of a factor 4 on average consumption and of a factor 10 on peak power). Experiments have also shown the relative cost of a memory operation compared to arithmetic computations (for example, a transfer from an external memory consumes 33 times more than a 16 bits addition).

Figure 1 shows where in the development flow global memory optimizations can be applied on a design. Once the Hardware/Software partitioning is done, the memory is already divided. It is therefore very important to make optimizations before this partitioning in order to deal with all the memory in homogeneous vision. We want here to optimize both types of memories, the one that will be included in hardware and the one controlled by software.

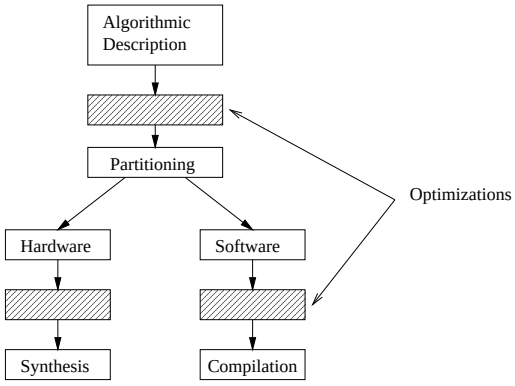


Figure 1: Loop transformations in the Codesign flow

The handling of data is done mainly through "for" loops in this kind of design. These loops form the critical part of the optimizations we want to apply at this stage. We thus propose to transform the algorithmic description of a design by using techniques similar to the ones used in automatic parallelisation [Wol96, BGS94] so as to reduce the consumption in power and size due to memory.

2 Memory Optimization Criteria and Associated Techniques

Target architectures and applications impose a complex memory hierarchy: registers, hardware and/or software caches, on-chip or off-chip memory [Cat98, PDN99]. A simplified view of the target architecture is shown on figure 2. The power consumption of a memory access increases with the level from which the data has to be fetched. An access to an external memory consumes more power than an access to an on-chip memory. Memory hierarchies are well exploited if we can achieve a good data *temporal locality*. This locality represents the amount of time between two successive accesses to the same memory location (either write-read or read-read). At the level of abstraction at which we apply loop transformations, we can only represent this parameter in an abstract manner. We can see on the figure 3(a) a source code composed with 3 different loops. The first loop computes the values stored in the array `b`, these values are then read in the third loop. We can have different approaches to measure temporal locality.

The first one would be to consider loops as atomic groups of instructions. As arrays are manipulated through loops, this implies that we do not consider locality between memory locations but we use a coarser

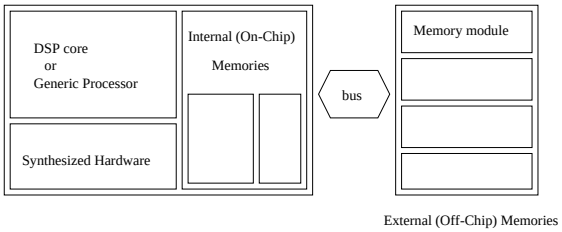


Figure 2: Target Architecture

grain represented by complete arrays. This level of granularity is used for *global* code transformations such as *moving code* or *loop merging* [Wol96, BGS94]. The figure 3(b) shows the first transformation we can apply on the source 3(a) in order to improve temporal locality. The last loop has been shifted up in order to tighten the production and consumption of the array `b`. The next step is to merge the two first loops to have a common iteration space where the consumption of a value `b(i)` can be made nearer from its production. The figure 3(c) shows the three loops merged into one. The third loop has been also merged because it uses values from the array `a` which are also in use in the second loop.

```
for i=1,n
 b[i]=a[i]
for i=1,n
 c[i]=a[i+1]
for i=1,n
 d[i]=b[i-1]
```

(a) source code

```
for i=1,n
 b[i]=a[i]
for i=1,n
 d[i]=b[i-1]
for i=1,n
 c[i]=a[i+1]
```

(b) moving code

```
for i=1,n
 b[i]=a[i]
 d[i]=b[i-1]
 c[i]=a[i+1]
end for
```

(c) loop merging

```
d[1]=b[0]
for i=2,n
 b[i-1]=a[i-1]
 d[i]=b[i-1]
 c[i-1]=a[i-1]
end for
b[n]=a[n]
c[n]=a[n+1]
```

(d) loop alignment

Figure 3: Example of code transformations: moving, merging, loop alignment.

The second way to represent temporal locality is to look inside loops. This representation allows us to consider subsets of the arrays handled by the loop. This level of abstraction can be used to perform *local* loop transformations such as *interchange*, *skewing*, *folding* or *alignment* [Wol96]. On the example 3(c) for each iteration, the loop produces the value `b[i]`, `c[i]`, `d[i]` and uses the values `b[i-1]`, `a[i]` and `a[i+1]`. The next transformation step will take into account values that are produced and consumed in different iterations of the loop. We call this gap of iterations a *distance*. A new value is produced at each iteration and must be kept into a separate foreground (on-chip) memory buffer until its last use by another statement of the loop. The number of "memories" needed to store a value computed and used in different iterations is given by the amount of iterations the value has to cross.

Figure 3(d) shows the loop once aligned to optimize the use of the arrays `b` and `a`. We can see that values of the array `b` are consumed as soon as they are produced. This optimization increases the probability we have to find the value `b[i-1]` in a very high level of the memory hierarchy. Optimization has also been performed in the use of the array `a`: the value `a[i-1]` has to be fetched from distant memory only once per loop iteration. We will use and develop this measure of temporal locality for loop alignment in the next section.

Memory is by itself a source of power consumption. It is also important to reduce the size of the memory needed by an application. A reduction of the amount of needed memory can decrease the number of levels in the memory hierarchy. A significant reduction would ideally allow to store everything in the on-chip memory, thus enabling the removal of the off-chip memory. This optimization can be done only if the consumption of a value appears right after its production. The array `b` on figure 3(d) can be completely removed from

memory if the array is not used elsewhere in the code. This optimization of memory size is also associated with loop alignment.

Loop transformations at this stage of the codesign flow can do a lot by themselves. But they cannot perform all the needed transformations. More powerful optimizations—in terms of power and memory size gain—can be achieved in later steps of the compiling flow, once the design has been partitioned. Optimizations like *in-place mapping* [De 98], *memory distribution* across modules [PDN99], *cache level* optimizations [KCM98] and many others [Cat98, PDN99] have to be done afterwards. These optimizations are *enabled* by high level transformations done *before* the hardware-software partitioning. Optimization criteria developed in the next section have been defined considering that they are performed afterwards.

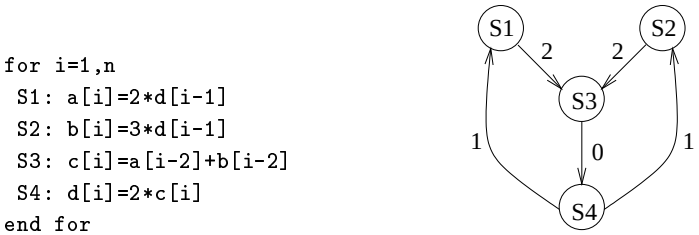
We present in the next two sections 3 and 4 two different algorithms for memory accesses optimizations. This first one minimizes the number of buffers needed between iterations of a loop by loop alignment. The second one finds a minimal bound for all dependencies of the graph.

3 Buffers Minimization in Loop Nests by Loop Alignment

The algorithm we present in this section minimizes the size of the foreground memory needed to store values that are computed and used in the same loop. This minimization can also be seen as optimizing the average distance in terms of temporal locality between read and write accesses to the same variable in different loop iterations. This technique is not only useful to keep values in a memory near the top of the hierarchy (where memories are smaller and less power consuming) but it can also decrease the memory size needed by the application (a dimension of an array can be reduced to a scalar value for example).

3.1 Modeling the Problem for a Single Loop (Monodimensional Case):

We use a Reduced Dependence Graph ($G = (V, E, w)$) representation for modeling the problem. Graph nodes (V) represent the statements of the loop, edges (E) represent data dependences between these statements. Each dependence edge is weighted by a distance w which corresponds to the number of iterations between the two accesses. These distances are positive as a program cannot use a value before its computation. We restrict ourselves to the case of uniform (constant) dependence distances [Wol96] over the loop to be able to use *retiming* [LS91, DH98] techniques.



(a) source code (b) dependence graph

Figure 4: Modeling dependences in a loop

Example: we can see on figure 4 that there are two dependences of distance 2 in the first loop to statement `S3` from statement `S1` and `S2`. There is also a dependence of distance 1 in the inner loop from statement `S4` to statements `S1` and `S2`. The value produced by the statement `S3` (`c[i]`) is consumed in the same iteration by the statement `S4`. There is a dependence distance of 0 between these last two statements.

The number of buffers needed for a statement represented by a node u depends on the dependence length $w(e)$ of all its out-edges e. This amount is given by the following relation.

Cost per node:

$$C(u) = \max_{e=(u,v)\in E} w(e) \qquad (u \xrightarrow{e} v)$$

The total number of buffers across iterations in the graph is thus:

$$Cost(G) = \sum_{u\in V} C(u).$$

Minimizing $Cost(G)$ can be solved in polynomial time, as we will see in the section 3.3, by using *retiming* techniques. A retiming value $r(u)$ (integer) is associated with each node u. This weight represents a shift (or a delay) in a number of iterations for the associated statement. Therefore applying a retiming on a graph modifies dependence distances. The graph after retiming can be rewritten into a code, functionally equivalent, but with new dependence distances w_r given by the following relation:

$$w_r(e) = w(e) + r(v) - r(u), \qquad (u \xrightarrow{e} v)$$

We must define a constraint in order to obtain a *legal* retiming on the graph, dependence distances after retiming must be positive (we cannot use a value before it is computed)

$$w_r(e) \geq 0, \ \forall e \in E$$

3.2 Integer Linear Program Formulation:

In this section we present the ILP formulation for the problem of minimizing $Cost(G)$.

$$\min \sum_{u\in V} C(u) \tag{1}$$

$$w(e) + r(v) - r(u) \geq 0, \quad \forall e = (u,v) \in E \tag{2}$$

$$C(u) \geq w(e) + r(v) - r(u), \quad \forall e = (u,v) \in E \tag{3}$$

The objective function of our ILP formulation is given by the relation 1. The constraints 2 ensure that we have a legal retiming. The cost of a node after retiming is given by the equation 3. As we cannot use a max function in the constraint—the problem would not be linear—we must define the cost of a node u to be greater or equal to the cost of each out-edge. Minimizing 1 ensures that the maximal value is reached by $C(u)$, giving the expected cost for each node. The ILP formulation given by (1), (3) and (2) minimizes the number of buffers needed across iterations of the loop.

Values $r = 0$ and $C(u) = \max_{e=(u,v)\in E} w(e)$ are always a feasible solution for the problem. Furthermore any feasible solution has a cost $\sum_{u\in V} C(u) \geq 0$ which ensures that an optimal solution always exists, because of this lower bound.

Size of the formulation:

- variables (r and C): $2|V|$;
- constraints: $2|E|$.

3.3 A Polynomial Algorithm

The matrix representation of the previous ILP formulation is given by the relation 4.

$$\min \left\{ \begin{pmatrix} r & C \end{pmatrix} \begin{pmatrix} 0 \\ 1 \end{pmatrix} \middle| \begin{pmatrix} r & C \end{pmatrix} \begin{pmatrix} -A & A \\ 0 & A^+ \end{pmatrix} \geq \begin{pmatrix} -w & w \end{pmatrix} \right\}, \tag{4}$$

where the matrix A is the nodes-edges incidence matrix (each column has one and only one +1 and -1, see [GM95]) of the reduced dependence graph $G = (V, E, w)$ and the matrix A^+ is defined as follow:

$$\begin{cases} a^+_{i,j} = 1 \text{ if } a_{i,j} = 1 \text{ (where the } a_{i,j} \text{ are the elements of } A) \\ a^+_{i,j} = 0 \text{ otherwise} \end{cases} \tag{5}$$

The matrix A^+ has the same dimensions as A but we keep only its positive values. This corresponds to the fact that we define the cost $C(u)$ only for out-edges of a node and not for in-edges.

The matrix $M = \begin{pmatrix} -A & A \\ 0 & A^+ \end{pmatrix}$ can be transformed into the matrix $M' = \begin{pmatrix} -A & A^- \\ 0 & A^+ \end{pmatrix}$ by an unimodular transformation (by subtracting the last $2|V|$ rows from the first $2|V|$). The matrix M' is a nodes-edges incidence matrix and is totally unimodular [GM95, p. 5] (every square regular submatrix of M' has a determinant of -1, 0 or +1). As we have transformed M to M' by an unimodular operation the matrix M is also totally unimodular and we can conclude that the ILP formulation 4 admits an integral optimal solution in the rationals and that it can be solved by a polynomial algorithm [dW90].

Although this problem can be solved very efficiently by any ILP solver we use the dual form of the problem 4 to reduce it to a *minimal cost flow* problem.

Interpretation of the Dual Problem: The dual form of the problem 4 is given by the problem 6 [dW90, p. 33].

$$\max \left\{ \begin{pmatrix} -w & w \end{pmatrix} \begin{pmatrix} x & y \end{pmatrix} \middle| \begin{pmatrix} -A & A \\ 0 & A^+ \end{pmatrix} \begin{pmatrix} x & y \end{pmatrix} = \begin{pmatrix} 0 & 1 \end{pmatrix}, \begin{pmatrix} x & y \end{pmatrix} \geq 0 \right\} \tag{6}$$

We first change the problem to have a minimization problem instead of a maximization one. The transformed problem is given in equation 7.

$$\min \left\{ \begin{pmatrix} w & -w \end{pmatrix} \begin{pmatrix} x & y \end{pmatrix} \middle| \begin{pmatrix} -A & A \\ 0 & A^+ \end{pmatrix} \begin{pmatrix} x & y \end{pmatrix} = \begin{pmatrix} 0 & 1 \end{pmatrix}, \begin{pmatrix} x & y \end{pmatrix} \geq 0 \right\} \tag{7}$$

The cost function on the unknown variables x and y to minimize is $w(x-y)$. This minimization is controlled by two sets of constraints which are given in equations 8 and 9.

$$A.(x-y) = 0, \ \begin{pmatrix} x & y \end{pmatrix} \geq 0 \tag{8}$$

$$A^+.y = 1, \ y \geq 0 \tag{9}$$

The first set of constraints 8 imposes that $(x-y)$ be a flow over the graph [GM95, p. 156]. The other set of constraints given by the equation 9 means that the y part of the flow on a node must be directed through one and only one of the out-edges of this node. These constraints can be taken into account by constructing a new graph $G'(V', E', w')$ from $G(V, E, w)$ in the following way:

Out-edges $\{e_i\}$ of the original graph are kept in the transformed graph with their respective weight. We then introduce a *virtual node* v_u. For each edge $e_i = (u, v)$, we build an edge $e'_i = (v, v_u)$. The edge e'_i is weighted by $-w(e_i)$ and has a maximal flow capacity c set to 1. Another edge e_u is added from the virtual node v_u to the node u. This edge has both minimal l and maximal c flow capacity set to 1 and is null weighted.

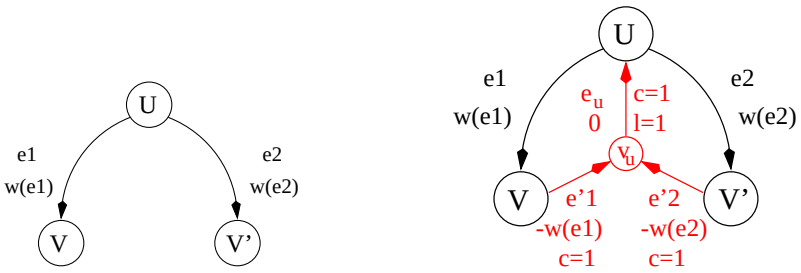
An example of transformation for a node with two out-edges is given on figure 5.

Let f be a flow of G', we define for each edge e a couple $(x(e), y(e))$ in the following way:

$$x(e) = f(e) \tag{10}$$

$$y(e) = f(e') \tag{11}$$

Proposition 1 *There is a bijection between the flows of G' and the feasible solutions of the dual problem. Furthermore minimal cost flows of G' correspond to optimal solutions for the dual problem.*



(a) original graph (b) modified graph

Figure 5: Graph transformations for flow algorithm resolution

Proof For each node u of G we have the flow equation on f:

$$\sum_{e=(u,?)\in E} f(e) + \sum_{e'=(u,?)\in E'\setminus E} f(e') = \sum_{e=(?,u)\in E} f(e) + \sum_{e'=(?,u)\in E'\setminus E} f(e') \tag{12}$$

Using 10 we can rewrite this equation as:

$$\sum_{e=(u,?)\in E} x(e) - \sum_{e'=(?,u)\in E'\setminus E} f(e') = \sum_{e=(?,u)\in E} x(e) - \sum_{e'=(u,?)\in E'\setminus E} f(e') \tag{13}$$

By construction we have:

- $\forall e = (?,u) \in E,\ \exists!\ e' = (u,?) \in E' \setminus E$ and $f(e') = y(e)$ so

$$\sum_{e'=(u,?)\in E'\setminus E} f(e') = \sum_{e=(?,u)\in E} y(e)$$

- $\forall e = (?,u) \in E' \setminus E,\ e = (v_u, u)$ and

$$\sum_{e'=(?,v_u)\in E'\setminus E} f(e') = f(e_u) = 1,$$

  which is equivalent to the flow conservation on virtual nodes.
  Moreover we have

$$\forall e' = (?,v_u) \in E',\ \exists!\ e = (u,?) \in E \text{ with } f(e') = y(e) \text{ and } \sum_{y=(u,?)\in E} y(e) = 1$$

  which is the constraint on any dual problem solution. So

$$\sum_{e'=(?,u)\in E'\setminus E} f(e') = \sum_{e=(u,?)\in E} y(e).$$

Equation 12 is equivalent to equation 14.

$$\sum_{e=(u,?)\in E} x(e) - \sum_{e=(u,?)\in E} y(e) = \sum_{e=(?,u)\in E} x(e) - \sum_{e=(?,u)\in E} y(e) \tag{14}$$

We can conclude from this last equivalence that a flow f for G' is in bijection with a feasible solution $(x - y)$ for the dual problem according to relations 10 and 11.

Moreover both flows have the same cost by construction of weights on edges of graph G' and because of the relations 10 and 11.

We have shown that a flow f for G' is in bijection with a feasible solution $\begin{pmatrix} x & y \end{pmatrix}$ of G and that both solutions have the same cost. So minimal cost flows f of G' are in bijection with optimal solutions $\begin{pmatrix} x & y \end{pmatrix}$ of the dual problem. □

Computing a Minimum cost flow f on G': we use a standard algorithm for computing minimum cost flows with both capacity and lower bounds (see [dW90, p. 248]). To start the algorithm we trivially construct an admissible flow for G' satisfying capacities on edges (v_u, u) by choosing for each vertex u an arbitrary outgoing edge (u, v) and putting some flow through $\{(u, v), (v, v_u), (v_u, u)\}$. The minimal cost flow algorithm introduces a graph $R^*(f)$ built from the flow f that will be used in the next paragraph.

Solution of the primal problem from the dual one: once we have found an optimal solution for the flow problem we have to compute the corresponding retiming for the primal problem.

We construct the retiming in the following way: we consider the optimal flow with its associated graph $R^*(f)$. We add a source S with a null weighted edge to all nodes of $R^*(f)$ and we compute the shortest path $\pi(u)$ from S to each node $u' \in V'$ by a Bellman-Ford algorithm [GM95].

We chose for each node u

$$r(u) = -\pi(u)$$

as a retiming value and the cost

$$C(u) = \max_{e=(u,v)\in E} w(e) + r(v) - r(u).$$

By definition of the shortest path, the values π provided by the Bellman-Ford algorithm check the following constraint:

$$\forall e' = (u, v) \in E',\ \pi(v) \leq \pi(u) + w(e).$$

These constraints expand to

$$\forall e' = (u, v) \in E',\ (-\pi(v)) - (-\pi(u)) + w(e) \geq 0.$$

Thus we have

$$\forall e = (u, v) \in E,\ r(v) - r(u) + w(e) \geq 0$$

because every edge e in E also belongs to E' with the same weight. The constraints 2 are therefore satisfied. The constraints 3 are also verified by definition of C. So $\begin{pmatrix} r & C \end{pmatrix}$ is a feasible solution for the primal problem.

Proposition 2 *The proposed feasible solution is optimal for the primal problem.*

Proof Since $\begin{pmatrix} r & C \end{pmatrix}$ and $\begin{pmatrix} x & y \end{pmatrix}$ are feasible solutions for the primal and the dual problem, by the *complementary slackness theorem* [dW90, page 39] they are optimal if and only if:

$$r(0 - (-Ax + Ay)) = 0 \tag{15}$$
$$C(1 - A^+ y) = 0 \tag{16}$$
$$(-Ar + w)x = 0 \tag{17}$$
$$(Ar + A^+ C - w)y = 0 \tag{18}$$

- Since $(x - y)$ is a flow of the graph G we have $A.(x - y) = 0$, thus the constraints 15 are verified.
- Since the flow $(x-y)$ deduced from f verifies the constraints 9 the set of constraints 16 are also verified.
- For each edge $e = (u, v)$ of G

- If $x_e = 0$, the constraint $(r(v) - r(u) + w(e)).x(e) = 0$ is trivially verified.
- Otherwise there is an edge (u, v) of weight $w(e)$ and an edge (v, u) of weight $-w(e)$ in $R^*(f)$ because $f(e)$ is neither at minimal nor maximal capacity. Thus the Bellman-Ford algorithm gives the relations

$$\begin{cases} \pi(v) \leq \pi(u) + w(e) \\ \pi(u) \leq \pi(v) - w(e) \end{cases}$$

which implies

$$(-\pi(v)) - (-\pi(u)) + w(e) = 0 \iff r(v) - r(u) + w(e) = 0$$

The constraint $(r(v) - r(u) + w(e)).x(e) = 0$ is verified.

The set of constraints 17 is thus verified.

- For each edge $e = (u, v)$ of G
  - If $y(e) = 0$ the constraint $(C(u) - (r(v) - r(u) + w(e))).y(e) = 0$ is trivially verified.
  - Otherwise there is an edge from v_u to v in $R^*(f)$ of weight $w(e)$ because of the value of $y(e)$ and by construction of $R^*(f)$.

    Moreover for all vertex $v' \neq v$ such that $e' = (u, v') \in E$ we have $y(e') = 0$ because of relation 9. Thus we have an edge from v' to v_u of weight $-w(e')$ in $R^*(f)$. So the Bellman-Ford algorithm finds the relations

$$\begin{cases} \pi(v) \leq \pi(v_u) + w(e) \\ \pi(v_u) \leq \pi(v') - w(e') \end{cases}$$

    This lead to the result

$$\pi(v) \leq \pi(v') + w(e) - w(e')$$

    which gives

$$(-\pi(v')) - (-\pi(u))) + w(e') \leq (-\pi(v)) - (-\pi(u)) + w(e)$$

    which is equivalent to

$$r(v') - r(u) + w(e') \leq r(v) - r(u) + w(e)$$

    So, by definition of $C(u)$ we have $C(u) = w(e) + r(v) - r(u)$ which verifies the relation

$$(C(u) - (r(v) - r(u) + w(e))).y(e) = 0.$$

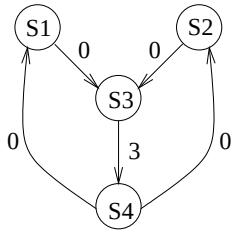
  The constraints 18 are verified.

We have shown that $\begin{pmatrix} r & C \end{pmatrix}$ and $\begin{pmatrix} x & y \end{pmatrix}$ are feasible solutions for the primal and the dual algorithm that verify the complementary slackness theorem so our solutions are optimal for both problems. □

Complexity: The complexity of the algorithm is

$$O(|V||E|.(\sum_{e \in E} w(e)))$$

The example we gave on figure 4, which needs 5 "buffers" is optimized on figure 6 with a cost of only 3 memories.



transformed code

prologue

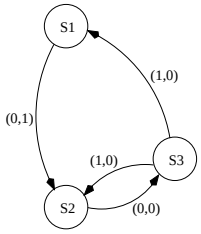
```
for i=1,n
  S4: d[i-1]=2*c[i-1]
  S1: a[i]=2*d[i-1]
  S2: b[i]=3*d[i-1]
  S3: c[i+2]=a[i]+b[i]
end for
```

epilogue

Figure 6: Example of the figure 4 after iteration buffers minimization

```
for i = 1,n
 for j = 1,m
S1: a[i,j]=c[i-1,j]
S2: b[i,j]=a[i-1,j]+c[i-1,j]
S3: c[i,j]=b[i,j]
 end for
end for
```

(a) source code



(b) dependence graph

Figure 7: Modeling dependences in a loop nest

3.4 Extending the Problem to the Multidimensional Case

In this section we are extending the problem to deal with loop nests. In the multidimensional case, dependences are handled by integer vectors. A component w_i of a dependence vector corresponds to the distance carried by the i^{th} loop in the nest, starting from the outer loop. A loop nest composed of n loops will thus carry dependence vectors with at most n components (it can be less than n if the nest is not perfectly nested).

Figure 7 shows the graph representation for the multidimensional case.

In this case, to be correct, the dependences have to be lexico-positive. We denote by $\geq_{lex}$ the lexicographic order. The equation 2 taken from the monodimensional case now becomes:

$$w(e) + r(v) - r(u) \geq_{lex} 0, \ \forall e = (u, v) \in E \tag{19}$$

These constraints are not linear and they cannot be linearized, to our knowledge, without losing total unimodularity on the matrix.

We propose here an efficient heuristic solution for the multidimensional problem by reducing it to the monodimensional one. This reduction is done by applying the monodimensional algorithm several times on the loop nest. Dependences handled by external loops are the more expensive ones as they imply manipulations of complete sub-arrays and also imply longer life time for the buffers we want to minimize, so we will consider them first. Incremental optimizations, like the one we propose, provide the opportunity to stop optimizing memory accesses and size given a tradeoff in order to switch to another optimization problem. For example, memory minimization is often dual with maximizing parallelism and finding an absolute optimal on memory may lead to very poor parallelism detection in the next step of the compilation.

Heuristic for the Multidimensional Case: The heuristic we propose for memory accesses in loop nests consist in transforming the nest loop by loop starting from the outermost loop to the innermost one. At each step, we dispose of a graph $G = (V, E, w)$ with weight vectors of dimension n, and we apply our algorithm

in the first dimension to find a retiming r_1 that optimizes it. Then we define $\tilde{G}_{r_1} = (V, E', w')$ from G_{r_1} as follow:

$$E' = \{e \in E \,|w(e) \leq_{lex} (0, +\infty, \ldots, +\infty)\}.$$

and w' is defined from w by restricting it to its $n-1$ last components. Finally we proceed to the next step with $\tilde{G}_{r_1}$.

Proposition 3 *This heuristic produces correct code after retiming.*

Proof Assume that we are given a graph G such that there exists a legal multidimensional retiming r of it. We want to prove that we can produce a new legal multidimensional retiming optimized for the first dimension.

The graph after retiming has the first component of all its dependence positive or null (by legality), and so the retiming r restricted to its first components and the associated cost constitute a feasible solution of the linear program of the monodimensional case. So we can apply our algorithm and find a retiming r_1 of G in its first dimension that minimizes the buffer size of the first loop.

Now we want to prove that this retiming can be completed in the $n-1$ remaining dimension in order to get a full multidimensional retiming of G that is legal and optimized for the first loop. Since G_r is legal, it has no circuit of lexico-negative weight, and so, by conservation of the weight of a circuit by retiming, G, G_{r_1} and by construction $\tilde{G}_{r_1}$ have also no circuit of lexico-negative weight. So given the lexicographic order and the addition on vectors, we can apply an algorithm of Bellman-Ford type on $\tilde{G}_{r_1}$ to find vectors $\pi(u)$ that check:

$$\forall e = (u, v) \in E', \pi(v) \leq \pi(u) + w'(e)$$

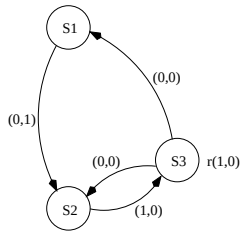
that is

$$w'(e) + (-\pi(v)) - (-\pi(u)) \geq 0,$$

which means that $-\pi$ is a legal multidimensional retiming of $\tilde{G}_{r_1}$ (and since, by construction, the edges not in $\tilde{G}_{r_1}$ are carried by the first dimension in G_{r_1}, the composition of r_1 and $-\pi$ is legal for G).

Finally we can continue the procedure on $\tilde{G}_{r_1}$ which is of dimension $n-1$. Since the starting graph is already legal by feasibility of the original code, the complete procedure is correct and provide an optimized legal retiming (the application of the Bellman-Ford algorithm of the proof is not necessary since only the existence of a legal multidimensional retiming is required, but if the optimization is not conduced down to the innermost loop, we have to apply it on the remaining dimensions to get a legal final retiming). □

Figure 8 shows the optimized code for example on figure 7 where the first loop has been aligned.



transformed code

```
  for j=1,m (i=1)
    a[i,j]=c[0,j]
    b[i,j]=a[0,j]+c[0,j]
  end for
  for i = 2,n-1
    for j = 1,m
S3:   c[i-1,j]=b[i-1,j]
S1:   a[i,j]=c[i-1,j]
S2:   b[i,j]=a[i-1,j]+c[i-1,j]
    end for
  end for
  for j=1,m (i=n)
    c[n,j]=b[n,j]
  end for
```

Figure 8: Example of the figure 7 after iteration buffers minimization

4 Bounding the maximal distance

Minimizing buffers between loop iterations as we have seen in the previous section can increase the dependence distance for some variables while decreasing for others. However it might be more suitable to make dependences more regular to fit better particular hardware design constraints (such as known number and size of cache line). The polynomial algorithm we propose here can modify the dependence distances of a program by retiming in order to find a minimal distance bound.

Given the reduced dependence graph $G = (V, E, w)$ of a loop nest we can bound the maximal dependence distance of the graph.

Problem: Let S be a set of $|V|$ inequalities of the form

$$w(e) + r(v) - r(u) \quad \leq \quad k, \ \forall e = (u, v) \in E \tag{20}$$

on the unknowns $r(u), u \in V$. The problem is to determine feasible values for the $r(u)$ or determine that the system is inconsistent. Let $a_e = k - w(e)$, the system 20 is transformed into the following system:

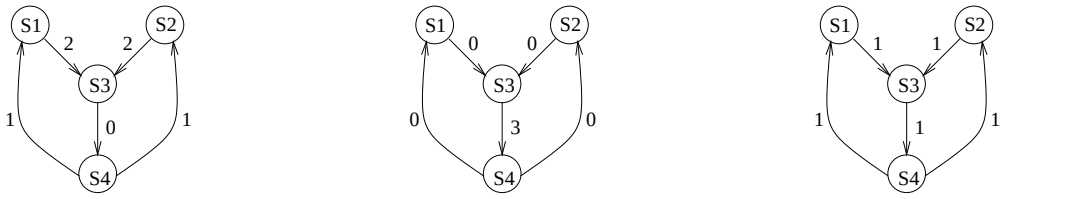
$$r(v) - r(u) \quad \leq \quad a_e, \ \forall e = (u, v) \in E \tag{21}$$

Such system in which each constraint has the form of inequality 21 arises in the shortest path problem that has been extensively studied and can be solved—or determined inconsistent—in $O(|V||E|)$ time by a Bellman-Ford algorithm [GM95, chapter 2].

The minimal solution for our problem is obtained with a logarithmic binary search for k in $[0, \max_{e \in E}(w(e))]$

Complexity: Each Bellman-Ford verification can be computed in $O(|V||E|)$, this verification is used $O(\ln(\max_{e \in E}(w(e))))$ times during the binary search.
The complexity of the problem is bounded by $O(|V||E| \ln(\max_{e \in E}(w(e))))$.

Starting from the graph 9(a), minimizing the maximal dependence distance of the graph produces the graph 9(c) for which the needed foreground memories are equal to 4. A better solution, as regards to memory size, would have been achieved by the solution 9(b) for which only 3 foreground memories are needed. However the solution 9(c) is more regular and may be best suited for specific architectures.



(a) source (b) minimizing iteration buffers (c) minimizing maximal distance

Figure 9: Minimizing the maximal distance

5 Future Work and Conclusion

We have presented in this paper a polynomial algorithm for memory accesses optimization by loop alignment (folding) in the monodimensional case and a heuristic based on this algorithm for the multidimensional case. A second polynomial algorithm was presented to minimize the maximal dependence distance of a loop nest.

These algorithms will be extended to deal with conditional execution ("if") in order to be able to model real applications. Therefore we want to use the *Program Dependence Graph* defined in [FOW87] by Ferrante et al. to take into account both data flow and control flow dependences for source to source loop transformations.

We have also presented the approach we have taken for automatic loop transformations on data dominated applications in multimedia applications. The interest of this automatic approach is on the one hand to reduce the design time by extracting optimizations for the description and on the other hand to improve the development quality by proposing interactive transformations that a designer could have missed.

We want to go further in the development of new global loop transformation techniques (loop merging, moving code, ...) as well as local transformations . These techniques will be integrated in our transformation engine LOOPING [Loo]. The LOOPING project we have started is a transformation engine prototype for source to source transformations. This engine has to be both automatic and interactive because there are many tradeoffs that only the designer of a system can control at this level of transformations. We want to go into more study on global transformations (loop merging, code moving) as well as local loop transformations (loop interchange, loop alignment, skewing).

References

[BGS94] David F. Bacon, Susan L. Graham, and Oliver J. Sharp. Compiler transformations for high-performance computing. *ACM Computing Surveys*, 26(4):345–420, December 1994.

[Bro98] Erik Brockmeyer. Low power data transfer and storage exploration for mpeg-4 on multi-media processors. Master's thesis, IMEC, April 1998.

[Cat98] Francky Catthor. Power-efficient data storage and transfer methodologies: current solutions and remaining problems. In *CS annual rush on VLSI*, Orlando, April 1998.

[De 98] Eddy De Greef. *Storage Size Reduction for Multimedia Application.* Phd thesis, IMEC, January 1998.

[DH98] Alain Darte and Guillaume Huard. Retiming et parallélisation automatique. Research Report RR1998-33, ENS-Lyon, 1998.

[dW90] Dominique de Werra. *Eléments de programmation linéaire avec applications aux graphes.* Presses polytechniques romandes, 1 edition, 1990. ISBN 2-88074-176-9.

[FOW87] J. Ferrante, K. J. Otteinstein, and J. D. Warren. The program dependence graph and its use in optimization. *ACM Transactions on Programming Languages and Systems*, 9(3):319–349, July 1987.

[GM95] Michel Gondran and Michel Minoux. *Graphes et algorithmes*, volume 37 of *Collection de la direction des études et recherches d'Électricité de France.* Eyrolles, 2 edition, 1995.

[KCM98] C. Kulkarni, F. Catthoor, and H. De Man. Cache optimization for multimedia compilation on embedded processors for low power. In *Proc. Intnl. Parallel Proc. Symp.(IPPS)*, pages 292–297, Orlando FA, April 1998.

[Loo] Projet LOOPING. http://www.ens-lyon.fr/~afraboul/looping/.

[LS91] Charles E. Leiserson and James B. Saxe. Retiming Synchronous Circuitry. In *Algorithmica*, volume 6, pages 5–35. Springer-Verlag, 1991.

[PDN99] Preeti Ranjan Panda, Nikil Dutt, and Alexandru Nicolau. *Memory Issues in Embedded Systems-On-Chip.* Kluwer Academic Publishers, 1999. ISBN 0–7923–8362–1.

[Wol96] Michael Wolfe. *High Performance Compilers for Parallel Computing.* Addison-Wesley Publishing, 1996. ISBN 0–8053–2730–4.