Scheduling affine parameterized recurrences by means of variable dependent timing functions

Christophe Mauras, Patrice Quinton, Sanjay Rajopadhye, Yannick Saouter

INRIA

UNITÉ DE RECHERCHE
INRIA RENNES

Institut National
de Recherche
en Informatique
et en Automatique

Domaine de Voluceau
Rocquencourt
B.P.105
78153 Le Chesnay Cedex
France
tél.(1) 39 63 55 11

Rapports de Recherche

N° 1204

Programme 2
Structures Nouvelles d'Ordinateurs

SCHEDULING AFFINE PARAMETERIZED RECURRENCES BY MEANS OF VARIABLE DEPENDENT TIMING FUNCTIONS

Christophe MAURAS
Patrice QUINTON
Sanjay RAJOPADHYE
Yannick SAOUTER

Avril 1990

RR_1204

IRISA

INSTITUT DE RECHERCHE EN INFORMATIQUE
ET SYSTÈMES ALÉATOIRES

Campus Universitaire de Beaulieu
35042 - RENNES CÉDEX
FRANCE
Téléphone : 99 36 20 00
Télex : UNIRISA 950 473 F
Télécopie : 99 38 38 32

Scheduling Affine Parameterized Recurrences by means of Variable Dependent Timing Functions
Ordonnancement de récurrences affines paramétrées à l'aide de fonctions de temps dépendant des variables *

Christophe Mauras†
Patrice Quinton
Sanjay Rajopadhye
Yannick Saouter

21 février 1990
Publication Interne n° 520 - 14 Pages

Abstract We present new scheduling techniques for systems of affine recurrence equations. We show that it is possible to extend earlier results on affine scheduling to the case when each variable of the system is scheduled *independently of the others* by an affine timing-function. This new technique makes it possible to analyze systems of recurrence equations with variables in different index spaces, and multi-step systolic algorithms. We illustrate our method on dynamic programming and LU decomposition.

Key-words : affine recurrences, systolic architectures, automatic synthesis, affine timing-functions, dynamic programming, LU decomposition

Résumé On présente des techniques d'ordonnancement nouvelles pour les systèmes d'équations récurrentes affines. On montre qu'il est possible d'étendre les résultats connus sur les fonctions de temps affines au cas ou chacune des variables du système est ordonnancée *indépendamment des autres* par une fonction de temps affine. Cette extension permet d'analyser des systèmes récurrents dont les variables ont des espaces d'indices différents, ainsi que des algorithmes multi-étapes. On illustre cette méthode sur l'algorithme de programmation dynamique et la décomposition LU.
Mots-clés : récurrences affines, architectures systoliques, synthèse automatique, fonctions de temps affines, programmation dynamique, décomposition LU

*Recherches partiellement supportées par le PRC C^3 et le projet ESPRIT BRA NANA.
†C. Mauras, P. Quinton and Yannick Saouter sont à l'IRISA, Campus de Beaulieu, 35042 RENNES-CEDEX, FRANCE. S. Rajopadhye est à l'Université d'Oregon, Eugene, OR 97403-1202, USA

CENTRE NATIONAL DE LA RECHERCHE SCIENTIFIQUE
(L. A. 227)
UNIVERSITÉ DE RENNES 1 I. N. S. A. DE RENNES

INSTITUT NATIONAL DE RECHERCHE
EN INFORMATIQUE ET EN AUTOMATIQUE
(LABORATOIRE DE RENNES)

1 Introduction

Systems of recurrence equations (RE) [8] are now a widely accepted starting specification means for the derivation of parallel algorithms in the field of signal processing. Methods to derive systolic architectures, and more generally, parallel implementations of algorithms from RE have been presented by many authors, in the case of Uniform Recurrence Equations (URE) (see among others[9, 10]). More recently, Rajopadhye and Fujimoto considered extensions of URE to affine dependences, called ARE[13], and Quinton and Van Dongen presented results on scheduling parameterized ARE (PARE) [11]. All these methods rely upon the notion of dependence vector, which make implicitly the assumption that all variables are in the same index space. Thus, a preliminary heuristic transformation is necessary to ensure that this condition is met.

In the following, we show that methods for defining affine schedules can be extended to the case when the variables have different index space. In section 2, we introduce the notion of ARE, and of PARE. Section 3 define variable dependent affine timing-functions, and gives necessary and sufficient conditions that the coefficients of these timing-functions must satisfy, for both the "non-rippling" and the "rippling" logic assumptions. In section 4, we apply thes notions to several well-known algorithms.

2 Parameterized affine recurrence equations and algorithms

2.1 Systems of affine recurrence equations

Definition

In the following, $\mathbf{Z}$ denotes the set of integers. A *system of affine recurrence equations* (ARE) is a finite collection of equations of the form

$$z \in D \rightarrow U(z) \;=\; f[V(I(z)),...] \tag{1}$$

where :

- z is a point of $\mathbf{Z}^n$.
- U and V are *variable* belonging to a finite set $\mathbf{U}$.
- D is the set of integral points of a convex polyhedron of $\mathbf{Z}^n$, i.e. a set defined by a finite number of linear inequalities on z (see [16] for notions related to convex polyhedra). D is called the *domain* of the equation.
- I is an affine mapping from $\mathbf{Z}^n$ to $\mathbf{Z}^l$ called *index mapping*. I has the form

  $$I(z) = A.z + B$$

  where the constants M and B are integral matrices: $M \in \mathbf{Z}^l \times \mathbf{Z}^n$, and $B \in \mathbf{Z}^l$.

- f is a single-valued function that depends *strictly* on its arguments; we assume that the function f has complexity $O(1)$.
- the '...' means that there can be other arguments of the same form as $V(I(z))$.
- the domains of two equations having the same variable in the left-hand side are disjoints. This hypothesis ensures that a variable is not defined twice.

Note that all equations need not be indexed in the same subspace, i.e., n is not necessarily the same for all equations.

Example 1

In order to illustrate this definition we consider the following algorithm for optimally parenthesizing a string of length n using dynamic programming (see [6, 12, 1]. The cost for parenthesizing substring i through j (is $c(i,j)$ and the cost for the outermost parentheses is $w(i,j)$.

$$
\begin{aligned}
0 < i < j \leq n \quad \rightarrow \quad & c(i,j) = \textbf{case} \\
& \quad j - i = 1 \rightarrow w(i,j) && (2) \\
& \quad j - i > 1 \rightarrow \min_{i<k<j} \{c(i,k) + c(k,j)\} + w(i,j) && (3) \\
& \textbf{esac}
\end{aligned}
$$

The fan-in of operator min in equation (4) is not bounded. The above system is therefore not an ARE. The following is an equivalent ARE obtained by serializing the minimization in equation (4) starting by the middle:

$$
\begin{aligned}
0 < i < j \leq n \quad \rightarrow \quad & c(i,j) = \textbf{case} \\
& \quad j - i = 1 \rightarrow w(i,j) && (4) \\
& \quad j - i > 1 \rightarrow C(i,j,1) + w(i,j) && (5) \\
& \textbf{esac} \\
0 < i < j \leq n \quad \rightarrow \quad & C(i,j,k) = \textbf{case} \\
& \quad 2k > j - i \rightarrow \infty && (6) \\
& \quad 1 \leq k, 2k \leq j - i \rightarrow \min\{c(i,i+k) + c(i+k,j), \\
& \qquad c(i,j-k) + c(j-k,j), C(i,j,k+1)\} && (7) \\
& \textbf{esac} \\
& && (8)
\end{aligned}
$$

2.2 Parameterized ARE

The notion of size parameter is a fundamental one when deriving parallel programs or systolic arrays. Indeed, the notions of *modularity* and *locality* can only be defined in

relationship with the size of the problem. In [11], the notion of parameterized system of recurrence equation was introduced. Here, we take a slightly different approach, and introduce parameterized ARE as a special case of ARE.

Definition

A *parameterized ARE* (PARE) is an ARE whose equations have the form

$$z \in D, p \in \mathbf{P} \rightarrow U(z,p) \;=\; f[V(I(z),p),...] \tag{9}$$

where, for all equations, $p = (p_1, p_2, ..., p_m)$ is a point of $\mathbf{Z}^m$ named *the size parameter* of the system. We assume that $\mathbf{P} \subset \mathbf{Z}^m$ is a convex set.

Note that the form of the right-hand side variable index in the definition of a PARE does not make it possible to have dependences between variables having different size parameter values. Therefore, a PARE is a collection of *independent ARE.*

Example 1 (continued)

The PARE for to dynamic programming algorithm is obtained by adding the parameter n as index for all variables:

$$\begin{aligned}
0 < i < j \leq n, n \geq 0 \;\rightarrow\; & c(i,j,n) = \textbf{case} \\
& \quad j - i = 1 \rightarrow w(i,j,n) && (10)\\
& \quad j - i > 1 \rightarrow C(i,j,1,n) + w(i,j,n) && (11)\\
& \textbf{esac} \\
0 < i < j \leq n, n \geq 0 \;\rightarrow\; & C(i,j,k,n) = \textbf{case} \\
& \quad 2k > j - i \rightarrow \infty && (12)\\
& \quad 1 \leq k, 2k \leq j - i \rightarrow \min\{c(i,i+k,n) + c(i+k,j,n), \\
& \qquad c(i,j-k,n) + c(j-k,j,n), C(i,j,k+1,n)\} && (13)\\
& \textbf{esac} \\
& && (14)
\end{aligned}$$

2.3 Reduced dependence graph

Given an ARE, we say that a variable instance $U(x)$ *depends directly* on $V(y)$ if there exists an equation

$$z \in D \rightarrow U(z) \;=\; f[V(I(z)),...]$$

such that $x \in D$ and $y = I(x)$. We say that $U(x)$ depends on $V(y)$, and we denote $U(x) \succ V(y)$, iff there exists a finite sequence of directly dependent variable instances $U_1(x_1)$, ... $U_k(x_k)$ such that $U(x) = U_1(x_1)$ and $V(y) = U_k(x_k)$.

The reduced dependence graph (RDG) of an ARE is the graph whose vertices are the variables $U \in \mathbf{V}$ of the system, and whose edges are tuples (U, V, D, I), with origin vertex U, extremity vertex, V, domain D and index I. There is such an edge for all pair U and V of all equation $z \in D \rightarrow U(z) = f[V(I(z)), ...]$.

We can define a *composition of edges* in the following way. Given two edges (U, V, D, I), and (V, W, E, J) with common vertex V, their composition is $(U, W, D \cap I^{-1}(E), J \circ I)$. The composition of edges summarizes the information needed to find out dependent instances of variables.

Finally, given a variable U, we define the domain D_U of U as the convex hull of the set of points x such that $U(x)$ is defined.

3 Affine variable-dependent timing functions

The scheduling problem is to find a function that associates each variable instance $U(z)$ with a given non negative instant of time t, in such a way that the arguments needed for the calculation of $U(z)$ are already calculated at time t. If such a mapping exists, the system is said to be *explicit* or *computable*. In [15], it is shown that the computability is in general undecidable for parameterized ARE.

A particular case of timing function is an *affine timing function*, that is to say, a timing function which is affine in the index of the variables. This case is interesting, because it is a good model for VLSI array. Here, we extend the usual definition to the case when the variables do not have the same index space.

3.1 Affine timing-function of an ARE

Define for each variable U, a function t_U from $\mathbf{Z}^n$ to $\mathbf{Z}$, where n is the index dimension of U, of the form :

$$t_U(z) = \lambda_U^1 z_1 + ... + \lambda_U^n z_n + \alpha_U$$

where λ_U^1, ..., λ_U^n are integers dependent of U. In the following, we let $\lambda_U.z = \lambda_U^1 z_1 + ... + \lambda_U^n z_n$.

3.2 Finding affine timing-function

Assume that the evaluation of each function of the system takes at least one unit of time. The following theorem gives a constructive means for obtaining the functions t_U :

Theorem 1 The numbers λ_U, $\alpha_U, U \in \mathbf{V}$ define a timing function iff:

- for all edge (U, V, D, I) of the RDG, we have:
  - *(i)* for all vertex σ of D, $\lambda_U.\sigma - \lambda_V.(I(\sigma)) + \alpha_U - \alpha_V > 0$,
  - *(ii)* for all ray ρ of D, $\lambda_U.\rho - \lambda_V.A.\rho \geq 0$

- for all variable U,
  - *(iii)* for all vertex σ of the domain D_U of U, $\lambda_U.\sigma + \alpha_U \geq 0$,
  - *(iv)* for all ray ρ of D_U, $\lambda_U.\rho \geq 0$.

Proof

- $\boxed{\Rightarrow}$ Let (U, V, D, I) be an edge of the RDG. Consider a vertex σ of D. By definition of a timing-function, and since $\sigma \in D$, $t_U(\sigma) > t_V(\sigma)$. Therefore, $\lambda_U.\sigma - \lambda_V.(I(\sigma)) + \alpha_U - \alpha_V > 0$, which proves *(i)*.

  Let ρ be a ray of D. Note that, given any point x of D, there exist infinitely many points in D of the form $x + k\rho$, where $k \in \mathbf{N}$. By definition of a timing-function, $t_U(x) - t_V(I(x)) > 0$ and $t_U(x + k\rho) - t_V(I(x + k\rho)) > 0$. But,

  $$t_U(x + k\rho) - t_V(I(x + k\rho)) \;=\; t_U(x) - t_V(I(x)) + k(\lambda_U.\rho - \lambda_V.A.\rho).$$

  If we had $\lambda_U.\rho - \lambda_V.A.\rho < 0$, there would obviously exist k such that $t_U(x + k\rho) - t_V(I(x + k\rho)) < 0$, which is a contradiction. Therefore, $\lambda_U.\rho - \lambda_V.A.\rho \geq 0$, which proves *(ii)*.

  As t_U is non-negative for all instances $U(x)$, it should obviously be non-negative for the vertices of the domain D_U, which proves *(iii)*. Finally, *(iv)* can be proved by using an argument similar to the demonstration of *(ii)*.

- $\boxed{\Leftarrow}$ We prove first that $t_U(x) > t_V(y)$ whenever $U(x)$ depends on $V(y)$ by induction on the length of the dependence path between $U(x)$ and $V(y)$. Consider two directly dependent variable instances $U(x)$ and $V(y)$. By definition, there exists an equation

  $$z \in D \rightarrow U(z) \;=\; f[V(I(z)), ...]$$

  such that $x \in D$ and $y = I(x)$. However, As $x \in D$, x is the sum of a convex combination of the vertices of D and of a positive combination of rays of D. Thus

  $$x = \sum_i \mu_i \sigma_i + \sum_j \nu_j \rho_j$$

  with $\sum_i \mu_i = 1$ and $\nu_j > 0$ for all j. We have

  $$\begin{aligned} t_U(x) &= \lambda_U.(\sum_i \mu_i \sigma_i + \sum_j \nu_j \rho_j) + \alpha_U \\ &= \sum_i \mu_i (\lambda_U.\sigma_i) + \sum_j \nu_j (\lambda_U.\rho_j) + \alpha_U \\ &> \sum_i \mu_i (\lambda_U.I(\sigma_i) + \alpha_V) + \sum_j \nu_j (\lambda_V.A(\rho)) \\ &= \lambda_V.I(\sum_i \mu_i \sigma_i + \sum_j \nu_j \rho_j) + \alpha_V \\ &= t_V(y) \end{aligned}$$

which proves the property.

On the other hand, for all U, $t_U(x)$ is obviously non negative because of conditions *(iii)* and *(iv)*. ■

3.3 Timing-functions under the rippling logic assumption

The above definition of timing-function maps directly the structure of the equations: it is implicitly assumed that the evaluation of the system is to be performed on an architecture where the combinational operators are separated by at least one delay. However, this hypothesis exclude practical cases when one may wish to cascade several combinational elements. The minimal condition to obtain a systolic implementation is that there is one delay at least on any elementary cycle of the RDG (see [14, 3]).

For this new model, the following result gives a constructive means for obtaining the functions t_U :

Theorem 2 The numbers λ_U, $\alpha_U, U \in \mathbf{V}$ define a timing function iff:

- for any edge (U, V, D, I) of the RDG
  - *(i)* for all vertex σ of D, $\lambda_U.\sigma - \lambda_V.(I(\sigma)) + \alpha_U - \alpha_V \geq 0$
  - *(ii)* for all ray ρ of D, $\lambda_U.\rho - \lambda_V.A.\rho \geq 0$
- *(iii)* for all elementary cycle of the RDG with composition (U, U, D, I), for all vertex σ of D, $\lambda_U.(\sigma - I(\sigma)) > 0$,
- for all variable U
  - *(iv)* for all vertex σ of the domain D_U, $\lambda_U.\sigma + \alpha_U \geq 0$,
  - *(v)* for all ray ρ of D_U $\lambda_U.\rho \geq 0$

We omit the proof which is similar to that of Theorem 1.

The following remarks should be made. First of all, both Theorems provide a set of linear inequalities that the coefficients of the functions t_U should satisfy. The number of such inequalities is finite, as the constraints derived by the rays can be reduced to a finite set of extremal rays (see [16]). The problem can thus be solved by Integer Linear Programming. We do not consider here the problem of finding optimal timing-functions: this problem was tackled by several authors, among which [17, 5, 4]. Secondly, if Theorem 1 gives a solution, then the system is computable, but the converse is not true (see [15]). Thirdly, is is simple to see that Theorem 2 has a solution iff Theorem 1 has one, although the sets of solutions are not identical.

3.4 Affine timing-function for a PARE

Timing-functions for a PARE can be derived directly from the one of the corresponding ARE, using both theorem. Indeed, the functions t_U of the ARE provide a parameterized expression of the timing-functions of the instances of the PARE. It should be noted however, that given a particular instance of PARE when p is fixed, there may exist timing-functions which can not be obtained by this means. In general, what is sought is a schedule which is *modular*, i.e. that has the same characteristics for all instances of the algorithm. This is exactly what we provide here. Moreover, if we are interested in a subset of problem instances, it suffices to add constraints on the value of the parameter.

3.5 Bounded delay timing-functions

As each variable in the system has its own timing-function, some of the solutions given by Theorems 1 and 2 may be such that two arguments of the same equation may be computed at instants of time which differ by an unbounded number. It is possible to modify the conditions of both Theorems in order to exclude such situations. To do so, it suffices to replace condition *(ii)* of Theorem 1 or 2 by:

- *(ii')* for all ray $\lambda_U.\rho - \lambda_V.A.\rho = 0$.

Theorem 3 *When condition (ii) of Theorem 1 is replaced by (ii'), then all solutions are such that for all edge (U, V, D, I) of the dependence graph, $t_U(x) - t_V(I(x))$ is bounded when $x \in D$.*

Proof Obvious. ■

This condition restricts the solution space of the timing-function, and is very often useful in practice (see 2D-convolution above).

4 Examples

In the following section, we apply the above theory to three examples.

4.1 Dynamic programming

The usual way of handling this example is to replace the occurrence of $c(i,j)$ by $C(i,j,1)$, in order to have all variables in the same space. In [12], it is shown that this system has the timing-function $t(i,j,k) = j - i - k$.

Theorem 1 provides a way to handle this example *without doing a substitution.*

To do so, we need to find out the generating system of the domains D_C and D_c. The domain D_C has one vertex $(1,3,1,3)$ and extremal rays $(0,0,0,1)$, $(0,2,1,2)$, $(0,1,0,1)$, and $(1,1,0,1)$. The domain D_c has one vertex $(1,3,3)$ and extremal rays $(1,1,1)$, $(0,0,1)$ and $(0,1,1)$.

Denote the timing-function of C by $\lambda_C^1 i + \lambda_C^2 j + \lambda_C^3 k + \lambda_C^4 n + \alpha_C$, and the timing-function of c by $\lambda_c^1 i + \lambda_c^2 j + \lambda_c^3 n + \alpha_c$. To illustrate the application of Theorem 1, consider the dependence between $(C(i,j,k,n), c(i,i+k,n))$ in equation (13). Its domain is D_C. Therefore, we must have:

$$\lambda_C^1 + 3\lambda_C^2 + \lambda_C^3 + 3\lambda_C^4 - \lambda_c^1 - 2\lambda_c^2 + 3\lambda_c^3 + \alpha_C - \alpha_c \geq 1 \quad (15)$$

$$2\lambda_C^2 + \lambda_C^3 + \lambda_C^4 - \lambda_c^2 - \lambda_c^3 \geq 0 \quad (16)$$

$$\lambda_C^2 + \lambda_C^4 - \lambda_c^3 \geq 0 \quad (17)$$

$$\lambda_C^1 + \lambda_C^2 + \lambda_C^4 - \lambda_c^1 - \lambda_c^2 - \lambda_c^3 \geq 0 \quad (18)$$

$$\lambda_C^4 - \lambda_c^3 \geq 0 \quad (19)$$

Inequality (15) comes from the vertex $(1,3,1,3)$ of D_C. Inequalities (16 - 19) come respectively from rays $(0,2,1,2)$, $(0,1,0,1)$, $(1,1,0,1)$, and $(0,0,0,1)$.

Applying this to all dependence pairs, and adding the conditions (*iii*) and (*iv*) for obtaining a positive timing-function gives a system of 23 inequalities. The single vertex of this domain is $\lambda_C^1 = \lambda_C^3 = \lambda_c^1 = -1$, $\lambda_C^2 = \lambda_c^2 = 1$, $\lambda_C^4 = \lambda_c^4 = 0$, $\alpha_C = -1$, and $\alpha_c = -2$. It represents a possible solution, actually, the one that minimizes the total amount of time. In other words, $C(i,j,k,n)$ should be computed at time $j - i - k - 1$, and $c(i,j,n)$ at time $j - i - 2$, which is, up to an additive constant, the solution provided by [12]. Note that here, the timing-function is independant of n.

4.2 LU decomposition

Consider a square matrix A of size n. The LU-decomposition algorithm aims at finding an upper triangular matrix U, and a lower triangular matrix L such that $A = LU$. The most "natural" equations for LU decomposition are the following ones;

$$0 \leq k \leq i \leq n, k \leq j \leq n \rightarrow A(i,j,k,n) = \textbf{case}$$

$$k = 0 \rightarrow a(i,j) \quad (20)$$

$$k > 0 \rightarrow A(i,j,k-1,n) - L(i,k,n) \times U(k,j,n) \quad (21)$$

$$\textbf{esac}$$

$$1 \leq i \leq n, 1 \leq j < i \rightarrow L(i,j,n) = A(i,j,j-1,n)/U(j,j,n) \quad (22)$$

$$1 \leq i \leq n, i \leq j \leq n \rightarrow U(i,j,n) = A(i,j,i-1,n) \quad (23)$$

Let $t_A(i,j,k,n) = \lambda_A^1 i + \lambda_A^2 j + \lambda_A^3 k + \lambda_A^4 n + \alpha_A$, $t_L(i,j,n) = \lambda_L^1 i + \lambda_L^2 j + \lambda_L^3 n + \alpha_L$, and $t_U(i,j,n) = \lambda_U^1 i + \lambda_U^2 j + \lambda_U^4 n + \alpha_U$. Consider as an example the dependance between $A(i,j,k,n), L(i,k,n)$ in equation (21). The domain of this equation is $\{0 < k \leq i \leq n, k \leq j \leq n\}$. This domain has one vertex $(2,2,1,2)$, and five rays $(1,0,0,1)$, $(0,1,0,1)$, $(1,1,0,1)$, $(0,0,0,1)$, and $(1,1,1,1)$. By applying the conditions of Theorem 1, we obtain the following set of inequalities:

$$2\lambda_A^1 + 2\lambda_A^2 + \lambda_A^3 + 2\lambda_A^4 + \alpha_A - 2\lambda_L^1 - \lambda_L^2 - 2\lambda_L^3 - \alpha_L \geq 1 \quad (24)$$

$$\lambda_A^1 + \lambda_A^4 - \lambda_L^1 - \lambda_L^3 \geq 0 \quad (25)$$

$$\lambda_A^2 + \lambda_A^4 - \lambda_L^3 \geq 0 \quad (26)$$

$$\lambda_A^1 + \lambda_A^2 + \lambda_A^3 + \lambda_A^4 - \lambda_L^1 - \lambda_L^2 - \lambda_L^3 \geq 0 \quad (27)$$
$$\lambda_A^1 + \lambda_A^2 + \lambda_A^4 - \lambda_L^1 - \lambda_L^3 \geq 0 \quad (28)$$
$$\lambda_A^4 - \lambda_L^3 \geq 0 \quad (29)$$

Applying the same to all dependence pairs, provides a system with 10 variables and 38 inequalities. Once solved, this system has one vertex, which corresponds to the solution $t_A(i,j,k,n) = i + j + k - 2$, $t_L(i,j,n) = i + 2j - 2$, and $t_U(i,j,n) = 2i + j - 2$. This solution corresponds to the usual schedule when one wants to design a systolic array for this algorithm.

4.3 2D-Convolution

The final example we consider aims at showing that our approach makes it possible to compute directly the parameterized form for the timing-function, when it depends on the parameter. Consider an image whose pixel intensity is denoted by $x(i,j)$, and a coefficient window $w(k,l)$, where $\{-p \leq i \leq p, -p \leq j \leq p\}$. The equation for computing the 2D-Convolution is:

$$0 \leq i < N, 0 \leq j < N \quad \rightarrow \quad y(i,j,N,p) = \sum_{k=-p}^{p} \sum_{l=-p}^{p} w(k,l)x(i+k, j+l). \quad (30)$$

After serialization of the $\sum$ operator, introducing new intermediate variables Y and Z, we get:

$$0 \leq i < N, 0 \leq j < N \quad \rightarrow \quad y(i,j,N,p) = Y(i,j,p,N,p) \quad (31)$$

$$0 \leq i < N, 0 \leq j < N, -p-1 \leq k \leq p \quad \rightarrow \quad Y(i,j,k,N,p) = \textbf{case}$$
$$k = -p - 1 : 0 \quad (32)$$
$$-p \leq k \leq p : Y(i,j,k-1,N,p)$$
$$+Z(i,j,k,p,N,p) \quad (33)$$
$$\textbf{esac}$$

$$0 \leq i < N, 0 \leq j < N,$$
$$-p \leq k \leq p, -p-1 \leq l \leq p \quad \rightarrow \quad Z(i,j,k,l,N,p) = \textbf{case}$$
$$l = -p - 1 : 0 \quad (34)$$
$$-p \leq l \leq p : Z(i,j,k,l-1,N,p)$$
$$+w(k,l,N,p)x(i+k, j+l, N, p) \quad (35)$$
$$\textbf{esac}$$

Once applied to this system, Theorem 3 provides a convex set, defined by 34 inequalities, with two vertices. After adding constraints on the coefficients of t_Y and t_Z, so that the input data $x(i,j)$ are treated sequentially for increasing value of i and j, the solutions are $t_Y(i,j,k,N,p) = i + j + k + 3p + 2$, $t_Z(i,j,k,l,N,p) = i + j + k + l + 2p + 1$, or $t_Y(i,j,k,N,p) = i + j + 2k + 4p + 2$, and $t_Z(i,j,k,l,N,p) = i + j + 2k + l + 2p + 1$. The first one is faster.

5 Conclusion

The natural recurrence equations for most of the algorithms involves variables with different index spaces, which does not make it possible to use the notion of dependence vector. We have shown here that one can overcome this problem by associating different affine timing-functions to the variables, we have given necessary and sufficient conditions that the coefficients of these timing-functions must satisfy, for both the "non-rippling" and the "rippling" logic assumptions. This theory applies directly to many problems, such as dynamic programming, LU decomposition, and 2D-convolution which are presented here, and avoids in particular preliminary heuristic rewriting of the equations. An important consequence of this work is the possibility to handle directly multi-step algorithms (see [2]).

The notion of system of parameterized affine recurrence equation introduced in [11] and used here is also of great importance, as it captures the notion of problem for which a modular and uniform solution should be found. We have shown that it amounts to consider a system of affine recurrence equations, the variables of which are indexed by the parameters.

Finally, let us mention that all the examples treated here have been solved using the ALPHA du CENTAUR environment presented in [7]. In particular, the calculation of the generator systems of the convex sets was done using the "geometric engine" of this environment, which is a package of routines for calculations on convex polyhedra.

Open problems that are not tackled here are the extension of usual projections methods for obtaining processor allocations for a given algorithms, and also extensions of uniformizations techniques.

References

[1] M.C. Chen. A design methodology for synthesizing parallel algorithms and architectures. *Journal of Parallel and Distributed Computing*, 461–491, December 1986.

[2] J.M. Delosme and I.C.F. Ipsen. Efficient systolic arrays for the solution of Toeplitz systems: an illustration of a methodology for the construction of systolic architectures for VLSI. In W. Moore, A. McCabe, and R. Urquhart, editors, *International Workshop on Systolic Arrays*, pages 37–46, Adam Hilger, University of Oxford, UK, July 2-4 1986.

[3] J.M. Delosme and I.C.F. Ipsen. Systolic array synthesis : computability and time cones. In *Parallel Algorithms and Architectures*, pages 295–312, North-Holland, 1986.

[4] V. Van Dongen. *The minimal rays projection problem and its application to the design of low-cost systolic circuits.* Manuscript M 203, Philips Research Laboratory, July 1987.

[5] P. Gachet. Conception d'algorithmes et d'architectures systoliques. Thèse de l'Université de Rennes I, Sept 1987.

[6] P. Gachet, B. Joinnault, and P. Quinton. Synthesizing systolic arrays using DIASTOL. In W. Moore, A. McCabe, and R. Urquhart, editors, *International Workshop on Systolic Arrays*, pages 25–36, Adam Hilger, University of Oxford, UK, July 2-4 1986.

[7] P. Gachet, C. Mauras, P. Quinton, and Y. Saouter. ALPHA du CENTAUR: an environment for the design of regular algorithms. In *1989 International Conference on Supercomputing*, Crete, Greece, June 1989.

[8] R.M. Karp, R.E. Miller, and S. Winograd. The organization of computations for uniform recurrence equations. *Journal of the Association for Computing Machinery*, 14(3):563–590, July 1967.

[9] D.I. Moldovan. On the analysis and synthesis of VLSI algorithms. *IEEE Transactions on Computers*, C-31(11), November 1982.

[10] P. Quinton. Automatic synthesis of systolic arrays from recurrent uniform equations. In *11th Annual Int. Symp. Computer Arch., Ann Arbor*, pages 208–214, June 1984.

[11] P. Quinton and V. Van Dongen. The mapping of linear recurrence equations on regular arrays. *The Journal of VLSI Signal Processing*, 1:95–113, 1989.

[12] S.V. Rajopadhye. Synthesizing systolic arrays with control signals from recurrence equations. *Distributed Computing*, 3, 1989.

[13] S.V. Rajopadhye and R.M. Fujimoto. Systolic array synthesis by static analysis of program dependencies. In J.W. de Bakker, A.J. Nijman, and P.C. Treleaven, editors, *Parallel Architectures and Languages Europe*, pages 295–310, Springer-Verlag, June 1987.

[14] S.K. Rao. *Regular Iterative Algorithms and their Implementations on Processor Arrays.* PhD thesis, Standford University, U.S.A., October 1985.

[15] Y. Saouter and P. Quinton. *Computability of Recurrence Equations.* Technical Report, IRISA - Rennes (France), 1989.

[16] A. Schrijver. *Theory of Linear and Integer Programming. Wiley-Interscience series in Discrete Mathematics*, John Wiley and Sons, 1986.

[17] Y. Wong. *Algorithms for Systolic Array Synthesis.* PhD thesis, Yale University, New Haven, Connectictut, December 1989.

LISTE DES DERNIERES PUBLICATIONS INTERNES IRISA

PI 514 **PARALLELISATION D'UN RESEAU NEURONAL**
Krysztof WOLINSKI
Février 1990, 20 Pages.

PI 516 **COMMENT INTRODUIRE LA CONTIGUITE EN ANALYSE DES CORRESPONDANCES ? Application en segmentation d'image.**
Brigitte ESCOFIER, Habib BENALI, Kaddour BACHAR
Février 1990, 26 Pages.

PI 517 **MACHINE MODELING AND LOOP OPTIMIZATION FOR HORIZONTAL MICROCODED MACHINES**
François BODIN, François CHAROT
Février 1990, 24 Pages.

PI 518 **MULTISCALE SYSTEM THEORY**
Albert BENVENISTE, Ramine Nikoukhah, Alan S. Willsky.
Février 1990, 30 Pages.

PI 519 **PANDORE : A SYSTEM TO MANAGE DATA DISTRIBUTION**
Françoise ANDRE, Jean-Louis PAZAT, Henry THOMAS
Février 1990, 14 Pages.

PI 520 **SCHEDULING AFFINE PARAMETERIZED RECURRENCES BY MEANS OF VARIABLE DEPENDENT TIMING FUNCTIONS**
Christophe MAURAS, Patrice QUINTON
Sanjav RAJOPADHYE, Yannick SAOUTER
Février 1990, 14 Pages.

PI 521 **COMPUTABILITY OF RECURRENCE EQUATIONS**
Yannick SAOUTER, Patrice QUINTON
Février 1990, 28 Pages.

Imprimé en France
par
l'Institut National de Recherche en Informatique et en Automatique

ISSN 0249-6399

INRIA

UNITE DE RECHERCHE
INRIA-RENNES

Institut National
de Recherche
en Informatique
et en Automatique

Domaine de Voluceau
Rocquencourt
B.P.105
78153 Le Chesnay Cedex
France
Tel.:(1) 39 63 55 11

Rapports de Recherche

N° 1205

Programme 3
Réseaux et Systèmes Répartis

PROGRAMMING BY MULTISET TRANSFORMATION

Jean-Pierre BANATRE
Daniel LE METAYER

Avril 1990

IRISA

INSTITUT DE RECHERCHE EN INFORMATIQUE
ET SYSTÈMES ALÉATOIRES

Campus Universitaire de Beaulieu
35042 - RENNES CÉDEX
FRANCE
Téléphone : 99 36 20 00
Télex : UNIRISA 950 473 F
Télécopie : 99 38 38 32

Programming by multiset transformation

Programmation par transformation de multi-ensembles

Jean-Pierre Banâtre
Daniel Le Métayer

Publication Interne n°522 - mars 1990 - 26 pages

Abstract: *We present a new formalism called GAMMA in which programs are described in terms of multiset transformations. A distinguishing property of GAMMA is the possibility of expressing algorithms in a very abstract way, without any artificial sequentiality. The expressive power of the formalism is illustrated through a series of examples chosen from a wide range of domains (string processing problems, graph problems, geometric problems ...).*

Résumé: *Nous présentons un nouveau formalisme, appelé GAMMA, dans lequel les programmes sont décrits comme des transformateurs de multi-ensembles. La caractéristique majeure de GAMMA est de pouvoir exprimer des algorithmes sous une forme très abstraite, sans séquentialité artificielle. La puissance d'expression du formalisme est illustrée par une série d'exemples choisis dans divers domaines d'applications (traitement de textes, problèmes de graphes, problèmes de géométrie, ...).*

Ce travail a été effectué à l'IRISA, dans le cadre de l'équipe *Langages et Systèmes Parallèles.*

Adresse des auteurs: IRISA, Campus de Beaulieu, 35042 RENNES CEDEX, FRANCE

CENTRE NATIONAL DE LA RECHERCHE SCIENTIFIQUE
(L. A. 227)
UNIVERSITÉ DE RENNES 1 I. N. S. A. DE RENNES

INSTITUT NATIONAL DE RECHERCHE
EN INFORMATIQUE ET EN AUTOMATIQUE
(LABORATOIRE DE RENNES)

It has been argued for a long time that parallelism in a programming language makes the programming task far more difficult. This statement is undoubtedly true in the context of imperative languages because the programmer has to mentally manage several threads of control. Let us remark however that even sequential imperative programming entails an operational reasoning which is difficult to master. This suggests that the origin of the problem lies in the imperative paradigm rather than in parallelism itself. We go even further: we believe that parallelism is a powerful structuring facility that could profitably be exploited in program construction. We should make a distinction here between *logical parallelism* and *physical parallelism*. Physical parallelism is related to the implementation: it corresponds to the distribution of tasks on several processors. By logical parallelism, we mean the possibility of describing a program as a composition of several independent tasks. Of course, a particular implementation can turn logical parallelism into physical parallelism but these two notions have very different natures: the former is a program structuring tool whereas the latter is an implementation technique. Unfortunately, the term *parallelism* is often used without distinction for both of these concepts, which makes any discussion about parallelism very confusing. In this paper we use the term *parallelism* for *logical parallelism*; we are not concerned with implementation issues.

In fact the confusion between these two kinds of parallelism is part of the heritage of several decades of imperative programming. Another aspect of this lack of discrimination is the pervasive use of the sequencing operator ";" in imperative programming. This operator has been introduced because imperative programming languages were supposed to be "abstract" models of traditional von Neumann machines. As a consequence, most ";" in imperative programs reflect the need to model sequential machines, whereas few are really imposed by the logic of the program.

For example let us consider the following program fragment:

```
ma := max_array(a);
mb := max_array(b);
m := max(ma,mb)
```

where *max_array* is supposed to be a user-defined function yielding the maximum element of an array. The first occurrence of ";" is not related to the logic of the program (the first two statements could be exchanged without altering the meaning of the program); it is imposed by the (supposed) underlying architecture. The second occurrence of ";" however expresses a logical dependency between the first two statements and the last one (exchanging them would alter the meaning of the program).

This distinction is related to the more general issue of the separation of correctness and efficiency concerns: it is now admitted that correctness should be the primary concern in program development; it is only in a second stage that provably correct solutions should be used to derive more efficient versions. To this respect let us quote [6]: *"The basic problem in programming is managing complexity. We cannot address that problem as long as we lump together concerns about the core problem to be solved, the language in which the program is to be written, and the*

hardware on which the program is to execute. Program development should begin by focusing attention on the problem to be solved and postponing considerations of architecture and language constructs". In order to achieve this goal, one should be able to build in the first place an abstract version of the program in a high level language. In particular, these abstract programs should be free of artificial sequentiality. Unfortunately, to our knowledge, there is no available formalism allowing such an abstract description of programs. Let us take a simple example to illustrate this point. The problem is to find the maximum element of a set. In an imperative language the set can be represented as an array *a[1:n]* and a possible program is:

```
maxset1:  m:= a[1];
          i := 1;
          * [i < n  →
                i := i+1;
                m:= max(m, a[i])]
```

While the condition $i < n$ holds, index i is incremented and a new value of m is computed.
In a functional language, the set would be represented as a list and the program would be:

$$maxset_2(l) = \textbf{if } tail(l) = nil \textbf{ then } head(l) \textbf{ else } max\ (head(l), maxset_2(tail(l)))$$

In both cases the program imposes a strict ordering between the comparisons of the elements: the first element is compared with the second, then their maximum is compared with the third, and so forth... In both formalisms one could imagine a less constraining solution involving implicit parallelism. For example, in the functional language a divide-and-conquer version of the above program would be:

$$maxset_3(l) = \textbf{if } tail(l) = nil \textbf{ then } head(l) \textbf{ else}$$
$$\textbf{let } (l_1, l_2) = split(l) \quad \textbf{in}$$
$$max(maxset_3(l_1), maxset_3(l_2))$$

where $split\ (l_1, \ldots, l_n) = ((l_1, \ldots, l_{n/2}), (l_{n/2+1}, \ldots, l_n))$. Here again a (non-strict) ordering is imposed on the comparisons: for example the first and last elements of the list will not be compared (except in the case where they are the maxima of their respective sublists).

In fact the maximum of a set can be computed by performing the comparisons of the elements in any order. So we would like an abstract algorithm of the form:

***while** there are at least two elements in the set*
***select** two elements of the set, **compare** them and **remove** the smaller one*

This is almost a GAMMA program. In GAMMA such a statement can be written as follows:

$maxset_4\ (s) = \Gamma((R,A))\ (s)$ **where**

$R(x,y) = x \leq y$

$A(x,y) = \{y\}$

Function R specifies a property to be satisfied by the selected elements; these elements are replaced in the set by the result of the application of function A. Nothing is said in this definition about the order of evaluation of the comparisons; if several disjoint pairs of elements satisfy property R, the comparisons and replacements can even be performed in parallel. An intuitive way of describing the meaning of a GAMMA program is the metaphor of the chemical reaction: the set can be seen as a chemical solution, function R (called the reaction condition) is a property to be satisfied by reacting elements and A (the action) describes the product of the reaction. The computation terminates when a stable state is reached, that is to say when no elements of the set satisfy the reaction condition.

Let us now give a more formal presentation of GAMMA. The basic data structure in GAMMA (General Abstract Model for Multiset mAnipulation) is the *multiset*, which is the same as a set except that it may contain multiple occurrences of the same element [16]; the multiset is sometimes referred to as a *bag*. The benefit of using multisets is the possibility of describing compound data without any form of constraint or hierarchy between its components. This is not the case for recursively defined data structures such as lists which impose an ordering on the examination of elements (function $maxset_2$ above is an illustration of this constraint). The control structure associated with multisets is the Γ operator; as we have seen on the above example, Γ reflects the absence of hierarchy in the data structure and entails some kind of chaotic model of execution. Its formal definition can be stated as follows:

$\Gamma((R_1,A_1),...,(R_m,A_m))\ (M) =$

if $\quad \forall i \in [1,m],\ \forall x_1,...,x_n \in M,\ \neg R_i\ (x_1,...,x_n)$

then $\quad M$

else $\quad$ *let* $x_1,...,x_n \in M$, *let* $i \in [1,m]$ *such that* $R_i\ (x_1,...,x_n)$ *in*

$\Gamma((R_1,A_1),...,(R_m,A_m))\ ((M - \{x_1,...x_n\}) + A_i\ (x_1,...,x_n))$

The notation $\{..\}$ is used to represent multisets. There is no ambiguity here since we never use simple sets. The basic operations on multisets are the following:

- *union*: the number of occurrences of an element in $M_1 + M_2$ is the sum of its numbers of occurrences in M_1 and M_2.
- *difference*: the number of occurrences of an element in $M_1 - M_2$ is the difference between its numbers of occurrences in M_1 and M_2 (if this difference is greater than or equal to zero, otherwise it is zero).
- *product*: $M \times N$ is the cartesian product of M and N.
- *maximum*: the number of occurrences of an element in $M \cup N$ is the maximum of its numbers of occurrences in M and N.
- *minimum*: the number of occurrences of an element in $M \cap N$ is the minimum of its numbers of occurrences in

M and N.

- *cardinal*: *card(M)* yields the sum of the numbers of occurrences of all the elements of the multiset M.

We use $\exists\, x_1,...x_n \in M$ as a shorthand notation for $\exists\, \{x_1,...x_n\} \subseteq M$, which means that $x_1,...x_n$ are different elements of the multiset (even if some of them may possibly possess the same value); this is in contrast to $\exists\, x_1 \in M,...,\ \exists\, x_n \in M$, where $x_1,...x_n$ are not necessarily different elements.

(R_i,A_i) are pairs of closed functions (functions whose definition does not involve global variables) specifying reactions. The effect of a reaction (R_i,A_i) on a multiset M is to replace in M a subset of elements $\{x_1,...x_n\}$ such that $R_i\,(x_1,...x_n)$ is true by the elements of $A_i\,(x_1,...x_n)$. If no elements of M satisfy any reaction condition ($\forall\, i \in [1,m],\ \forall\, x_1,...x_n \in M\ ,\ \neg R_i\,(x_1,...x_n)$) then the result is M; otherwise the result is obtained by carrying out one reaction $((M - \{x_1,...x_n\}) + A_i(x_1,...x_n))$ and repeating the same process. The above definition implies that if one or several reaction conditions hold for several subsets at the same time, the choice which is made among them is not deterministic. The importance of the *locality property* of GAMMA cannot be overemphasized: if the reaction condition holds for several disjoint subsets, the reactions can be carried out independently (and simultaneously). This property is the basic reason why GAMMA programs do generally exhibit a lot of potential parallelism.

GAMMA programs can be composed using the traditional functional notation. Furthermore, functions R_i and A_i can be defined by pattern matching on the form of their arguments; pattern matching can also be used to extract elements from a multiset (for example *m* ***where*** $\{m\} = \Gamma((R,A))(G)$ is the extraction of the unique element of the result multiset). It should be noticed that the arguments of R_i and A_i must be of the same form since A_i is applied to arguments satisfying R_i. The notation used above to describe GAMMA programs may seem slightly awkward since it requires two separate definitions for R_i and A_i which entails the duplication of the text of their formal arguments. An alternative syntax could be:

replace $(x_1,...x_n)$ ***such that*** *"text of the condition"*
by *"text of the action"*

Nevertheless we shall keep the original syntax here because it emphasizes the functional nature of reaction conditions and actions.

We discuss in next section the programming style entailed by the GAMMA formalism. Then we illustrate the expressive power of the language with a variety of problems: numerical problems, sorting problems, string processing problems, graph problems, geometric problems, and process synchronization problems. The conclusion contains a comparison with related work and a discussion of the role of a high level language like GAMMA in the program construction process.

THE GAMMA STYLE OF PROGRAMMING

In this section we illustrate three programming styles through a simple example: the imperative style, the functional style and the GAMMA style. Our objective is to emphasize the distinguishing properties of GAMMA and to exhibit the basic programming methodology in GAMMA.

Prime number generation

The goal is to produce the prime numbers less than a given N.

Imperative solutions

These solutions are based on standard sieve techniques, which rely on the fact that the $(i+1)^{th}$ prime number (with $i \geq 1$) is the smallest integer exceeding the i^{th} prime that is not divisible by the first i primes. Integers greater than $\sqrt{N}$ which have not been eliminated by divisions by integers less than $\sqrt{N}$ are prime. We present two solutions in a traditional imperative language with guarded commands: a sequential program and a program with explicit parallelism.

Sequential solution

Initially array t contains 2 and all odd integers less than N, and eventually the first k_max elements of t contain the primes less than N. $\lceil X \rceil$ represents the smallest integer greater than X.

```
begin
   var max: integer init ⌈(N+1)/2⌉;
   t: array[1: max] of integer;
   var k_max: integer init max;
   var n: integer init 3;
   var i: integer init 2;
   var current_odd: integer init 2;
    var p, k, m: integer;
   t[1] := 2;
   *[ i ≤ max → t[i] := n ; i:= i+1; n:= n+2];
   *[ t[current_odd] < √N →
         i := current_odd+1; k := i;
         p:=t[current_odd];
         *[ i≤ k_max →
               m := t[i];
               [multiple(m,p) → skip □ not(multiple(m,p)) → t[k] := m; k:= k+1];
               i:= i+1
         ];
         k_max := k;
         current_odd := current_odd+1
   ]
end
```

To simplify this solution, we have used function *multiple(x,y)* which returns *True* if x is a multiple of y and *False* otherwise. The invariant maintained by this program is: $(\forall i \in [1: current\_odd], t[i] \in \{primes\ less\ than\ N\})$. The program is made of two embedded loops; iteration eliminates numbers between *t[current_odd + 1]* and *t[k_max]* which are multiple of the prime *t[current_odd]*; *t[current_odd]* corresponds to the last detected prime, i is the index of the element of the array currently being processed and k is the index of the last element in the current array which is not a multiple of *t[current_odd]*.

This sequential program is rather hard to understand. It involves several state variables used to manage the computation of primes and the necessary updatings of t. Shorter imperative programs can be proposed but they are not significantly simpler and we have chosen to present this one because it is closer to the functional and GAMMA solutions.

Parallel solution

This solution is expressed in CSP [15] and it is inspired by a program given in [15]. It is defined in terms of an array of processes, *sieve*, in which each process first inputs a prime p from its predecessor, prints it, and then transmits to its successor all the numbers received from its predecessor that are not multiples of p. There is no dynamic creation of processes in CSP; so it is necessary to determine statically the exact number of processes necessary to carry out the computation. In fact, $\lceil\sqrt{N}\rceil + 2$ processes are necessary: one process computes the even integers, $\lceil\sqrt{N}\rceil$ processes compute the primes less than $\lceil\sqrt{N}\rceil$ and one process collects integers greater than $\lceil\sqrt{N}\rceil$ and not eliminated. An extra process named *print* is supposed to be available to gather the primes and to print them.

```
begin
   sieve[0] ::
      print ! 2; n: integer; n:= 3;
      *[ n ≤N → sieve[1] ! n; n:= n+2]
   ||
   sieve(i:1.. ⌈√N⌉) ::
      p: integer;
      sieve(i-1) ? p;
      print ! p;
      *[ m: integer; sieve(i-1) ? m →
          [multiple(m,p) → skip□ not(multiple(m,p)) → sieve(i+1)!m]
       ]
   ||
   sieve[⌈√N⌉ + 1] ::
      *[n: integer; sieve[ ⌈√N⌉] ? n → print ! n]
end
```

Functional solution

This recursive solution uses two auxiliary functions: *int_between(m, n)* which produces the list of integers *(m..n)* and *filter(p, x)* which eliminates all multiples of *p* from list *x*.

```
primes (N) = prime_numbers (int_between (2, N))
filter (p, y) = if multiple (head (y), p)  then filter (p, tail (y))
                    else  cons ( head(y), filter(p, tail(y)))
int_between (m, n)  = if m > n then nil
                    else cons (m, int_between (m+1, n))
prime_numbers (l) = if null(l) then nil
                    else cons (head(l), prime_numbers( filter(head(l), tail(l)))
```

GAMMA solution

In GAMMA the solution consists in removing multiple elements from the multiset *{2,...,N}*. The result multiset contains exactly the prime numbers less than *N*: a prime number cannot be eliminated from the initial multiset and eliminated elements are obviously not prime numbers.

```
prime_numbers(N) = Γ((R, A)) ({2, ..., N})   where
      R(x, y) = multiple(x, y)
      A(x, y) = {y}
```

Compared with the previous solutions, the GAMMA program is far more concise and easier to understand. The reason is that many computational details are left unspecified in GAMMA. Both imperative solutions are more intricate because they give a precise description of execution order (expressed through the ";" operator in the sequential program and through explicit process communication in the parallel program) and memory management (through the use of a collection of variables). The functional solution is also lower level because list construction and decomposition have to be made explicit. In fact, one could say that the GAMMA program captures in a natural way the basic strategy applied in the imperative and functional solutions, namely to eliminate multiple elements from the set *{2,...,N}*.

As a consequence of the absence of commitment to a particular execution order, the GAMMA program can be implemented naturally in a parallel way [2]. This is also the case for the parallel imperative program but the major difference is that the programmer is responsible for parallelism management and he must have a precise knowledge of the underlying architecture to write a program suitable for parallel execution. This is a case where conflicts may arise between logical parallelism used as a programming technique and the physical parallelism necessary to exploit a particular architecture. The functional solution can also be implemented in a parallel way: functions *int_between*, *filter* and *prime_numbers* can be executed in a pipeline but the parallelism is far more constrained than in the GAMMA program where all comparisons between different elements can potentially be performed simultaneously. It should be clear that we do not claim here that a realistic parallel implementation of GAMMA is straightforward. In

fact any reasonable implementation of GAMMA must rely on sophisticated derivation techniques. What we claim is that our high level of abstraction eliminates unfortunate sequential biases.

Program construction in GAMMA

The GAMMA model of computing puts forward a quite unusual approach to program design. A program is no longer a sequence of instructions modifying a state, or a function applied to its arguments but rather a multiset transformer operating on all the data at once. The development of GAMMA programs entails a choice in data representation and a choice in the type of transformation applied to this data.

Data decomposition

The unique data structure provided in GAMMA is the multiset. Elements of multisets may be basic or composed values (even multisets) but there is no recursive data structure definition in GAMMA. So the first task to perform when designing a program in GAMMA is to find a representation of data as a multiset of items. If the program has to operate on basic values such as integers, a suitable decomposition of these values has to be found. For example the *prime_numbers* program above decomposes its argument N into a multiset $\{2,...,N\}$. Next section contains another decomposition of integers in terms of prime factors. Complex data such as sequences, trees or graphs can be represented as multisets in a straightforward way: for example, the components of a multiset representing a tree are the nodes and the leaves of the tree associated with parenthood information, the components of a sequence multiset are pairs *(index,value)* ... It is a distinguishing property of GAMMA to view all data structures as flat multisets: all the components of a data structure are directly accessible, independently of their position in the structure. If the position in the structure is relevant to the reaction, it has to be expressed via the reaction condition ($R(x,y)$ = *leftson(x,y)* for example, in a tree manipulating program). We can say that GAMMA has a *topological view of data types*; this contrasts with the traditional *recursive view of data types* where a walk through the data is necessary to access a particular component. This property has deep effects on the GAMMA programming style.

Relaxation

Relaxation is a method used in mathematics to solve systems of equations by iteration; first an estimate vector solution is produced (guessed), and the errors in the initial estimation are decreased and relaxed as calculation continues. The same method may be applied for solving many problems with the GAMMA paradigm. The initial multiset represents a (possibly rough) estimate of the solution and a series of actions refine this estimate until a proper solution is found. In the prime number generator example, the initial multiset $\{2, ..., N\}$ can be seen as a rough estimate of the set of prime numbers less than N, and the computation refines the result by eliminating the anomalies (non primes).

Data expansion and reduction

Programming by relaxation is suitable when the basic structure of the result is known. Let us now take a new example to illustrate two other programming techniques: data expansion and reduction. The following GAMMA program computes *Fibonacci(n)* where *Fibonacci(n)* = ***if*** $n \leq 1$ ***then*** *1* ***else*** *Fibonacci(n-1)+Fibonacci(n-2)*:

$fib(n) = m$ **where**

$\{m\} = sigma(gen(\{n\}))$

$gen(N) = \Gamma((R_1, A_1), (R_2, A_2))(N)$ **where**

$R_1(n) = n > 1$

$A_1(n) = \{n - 1, n - 2\}$

$R_2(0) = true$

$A_2(0) = \{1\}$

$sigma(M) = \Gamma(R, A)(M)$ **where**

$R(x, y) = true$

$A(x, y) = \{x + y\}$

The initial number n is decomposed into a number of ones which are then summed up to produce the expected result. By *data expansion*, we mean the decomposition of values into a collection of items. Computation stops when the multiset is a collection of indivisible elements. Data expansion involves unary reaction conditions and actions. In the *fib* example above, *gen* performs a data expansion, indivisible elements are ones. *Reduction* corresponds to the case where a multiset of items is reduced to a singleton by successive applications of the action. In the Fibonacci example, the *sigma* operator proceeds by reduction: a multiset $\{1, ..., 1\}$ is transformed into a singleton multiset by a series of sums.

Data expansion and data reduction are dual programming techniques: the former decomposes values into collections of simpler components and the latter gathers individual elements to build more complex values. A comparison of the GAMMA program with the original functional version of Fibonacci shows that data expansion corresponds to the recursive function calls and reduction to function returns. On the other hand, relaxation does not transform the structure of the multiset but proceeds by successive refinements. The three programming techniques described here are put into practice in the rest of the paper to solve a wide range of problems.

NUMERICAL PROBLEMS

Numerical problems are often used to test the expressiveness of linguistic constructs. We have chosen two examples here: the classical factorial problem and the prime factorization of an integer.

Factorial

The following GAMMA program computes $n!$:

$fact(n) = \Gamma((R,A))(\{1, ..., n\})$ **where**

$R(x, y) = true$

$A(x, y) = \{x*y\}$

This very simple program is an illustration of the reduction technique. It should be noticed that no constraint is put on the order in which multiplications are performed. No imperative or functional solution exhibits such a freedom in

the execution order.

Prime factorization of an integer

A fundamental theorem of arithmetic states that every positive integer n can be written as a product of primes, and that this decomposition is unique. This fact gives a one-to-one correspondance between positive integers and multisets of prime numbers; for example if $n = 2^2 * 3^3 * 11$, the corresponding multiset is $\{2, 2, 3, 3, 3, 11\}$.

Given a positive integer n, the set of prime numbers less than or equal to n, can be obtained using the program *primes* presented earlier. The notation $P \otimes \{(n, 0)\}$ represents the set of triples $(p_i,n,0)$, where $p_i \in P$.

$factorization(n) = P_2\ (P_1(\ prime\_numbers(n) \otimes \{(n, 0)\}))$ **where**

$P_1(M) = \Gamma((R_1, A_1))(M)$

$P_2(M) = \Gamma((R_2, A_2))\ (M)$

$R_1((n_1, n_2, k)) = multiple(n_2, n_1)$

$A_1((n_1, n_2, k)) = \{(n_1, n_2/n_1, k+1)\}$

$R_2((n_1, n_2, k)) = true$

$A_2((n_1, n_2, k)) = \{n_1,...,n_1\}$ *(multiset with k occurrences of n_1)*

P_1 proceeds by relaxation to evaluate the coefficient associated with each prime number and P_2 removes unnecessary information from the triples. The prime factorization of the greatest common divisor, the least common multiplier and the product of two numbers can be computed very easily from their prime factorization:

$gcd\ (M,N) = M \cap N$

$lcm(M,N) = M \cup N$

$product(M,N) = M + N$

SORTING PROBLEMS

We illustrate the relaxation principle through three instances of the general problem of sorting a collection of values according to a particular criterion.

Multiset partitioning

Given two multisets of integers S and T, the problem consists in partitioning $S + T$ into two multisets S' and T' such that $S + T = S' + T'$, $card(S) = card(S')$ and $card(T) = card(T')$, and every element of S' is smaller than or equal to every element of T'. This is a generalization to multisets of the traditional set partitioning problem. The solution in GAMMA consists in gathering all the elements into a multiset $\{(x,inS) / x \in S\} + \{(x,inT) / x \in T\}$ and exchanging values (x,inS) and (y,inT) such that $x > y$ until the solution is reached (inS and inT are tags representing the origin of the value).

$sp(S,T) = \Gamma((R,A))\ (S \times \{inS\} + T \times \{inT\})$ **where**

$R((x,inS),(y,inT)) = x > y$

$A((x,inS),(y,inT)) = \{(x,inT),(y,inS)\}$

Traditional solutions to this problem proceed sequentially by selecting the greatest element of S and the smallest element of T and exchanging them until the former is smaller than or equal to the latter. In contrast, our solution performs exchanges in a chaotic way and the tag associated with a value may switch several times.

Dutch national flag

This variation of the partitioning problem has been proposed by Dijkstra. The goal is to sort an array of elements designated *red*, *white* or *blue* so that all the *red* elements appear before the *white*, which in turn appear before all the *blue* elements [10]. We assume that the initial multiset contains at least one element of each colour. In GAMMA, the array is represented by a multiset of pairs *(index,colour)* and the program proceeds again by exchanging ill-sorted elements until the solution is reached.

$dnf(Array) = \Gamma((R,A))\ (Array)$ **where**

$R((i,red),(j,white)) = i > j$

$A((i,red),(j,white)) = \{(i,white),(j,red)\}$

$R((i,white),(j,blue)) = i > j$

$A((i,white),(j,blue)) = \{(i,blue),(j,white)\}$

Sorting

We consider now the general sorting problem: the goal is to organize the elements of an array in increasing order. We use again a multiset of pairs *(index,value)* and the program exchanges ill-ordered values until all values are well-ordered.

$sort(Array) = \Gamma((R,A))\ (Array)$ **where**

$R((i,v),(j,w)) = (i > j)$ **and** $(v < w)$

$A((i,v),(j,w)) = \{(i,w),(j,v)\}$

STRING PROCESSING PROBLEMS

A string can be seen as a linear sequence of characters. Strings are central in many applications of computer science such as word processing systems for example. We present here three well-known string processing examples: the telegram problem, the longest upsequence problem and the majority element problem.

The telegram problem

The telegram analysis problem can be stated as follows [10]: it is required to process a stream of telegrams, each terminated by a string ZZZZ. Words are separated by one or more spaces. The resulting telegram must have a single

space between words and no leading or trailing spaces.

Data items are represented as triples (v, i, s), where v is the value of the data item (a character or a space), i is the index of this item and s is the shift to be used for accessing the next significant data item in the sequence. Initially, s is equal to 1 for all triples. The first character has index 1 and a string ZZZZ is added in the front of the sequence to avoid special treatment for the first telegram.

$$telegram\_analysis\ (M) = \Gamma((R_1, A_1), (R_2, A_2), (R_3, A_3), (R_4, A_4))\ (M + \{(Z,-3,1),(Z,-2,1),(Z,-1,1),(Z,0,1)\})$$

where

$$R_1((space, i_1, s_1), (space, i_2, s_2)) = (i_2 = i_1 + s_1)$$

$$A_1((space, i_1, s_1), (space, i_2, s_2)) = \{(space, i_1, s_1 + s_2\ )\}$$

$$R_2((v_1, i_1, s_1), (v_2, i_2, s_2)) = (i_2 = i_1 + s_1 \ \textbf{and}\ s_1 > 1)$$

$$A_2((v_1, i_1, s_1), (v_2, i_2, s_2)) = \{(v_1, i_1, 1), (v_2, i_1 + 1, s_1 + s_2 - 1)\}$$

$$R_3((Z,i,1),(Z,i+1,1),(Z,i+2,1),(Z,i+3,1),(space,i+4,k)) = True$$

$$A_3((Z,i,1),(Z,i+1,1),(Z,i+2,1),(Z,i+3,1),(space,i+4,k)) =$$
$$\{(Z,i,1),(Z,i+1,1),(Z,i+2,1),(Z,i+3,k+1)\}$$

$$R_4((space,i-1,1),(Z,i,1),(Z,i+1,1),(Z,i+2,1),(Z,i+3,k)) = True$$

$$A_4((space,i-1,1),(Z,i,1),(Z,i+1,1),(Z,i+2,1),(Z,i+3,k)) =$$
$$\{(Z,i-1,1),(Z,i,1),(Z,i+1,1),(Z,i+2,k+1)\}$$

Each reaction corresponds to a particular requirement of the specification: (R_1, A_1) and (R_2, A_2) are used for space elimination ((R_1,A_1) performs the elimination at the price of increasing shifts and (R_2, A_2) compacts the telegram) and (R_3, A_3), (R_4, A_4) for elimination of the first and last spaces. According to our classification, this program performs a relaxation (the initial multiset being a first approximation of the solution).

This problem which is presented in [10] as inherently sequential, is given here a solution with a high potential for parallelism: elimination of extra spaces and compaction may be carried out concurrently and can lead to a fair amount of parallelism.

The longest upsequence problem

A subsequence is obtained from a sequence by deleting some (non necessarily adjacent) values. A sequence is called an upsequence if its values are in non-decreasing order. The problem is to compute the length of the longest upsequence of a sequence.

A sequence is represented as a set of triples (n, x, l_n), where x is the value at index n and l_n is the length of the longest known upsequence ending at index n. We first give a GAMMA program which computes all the triples (n, x, l_n); M_0 is a multiset of pairs representing the initial sequence:

$$lup(M_0) = \Gamma(R, A)\ (\{(k, x_k, 1) \mid (k, x_k) \in M_0\}) \quad \textbf{where}$$

$$R((n, x_n, l_n), (k, x_k, l_k)) = (n < k \ \textbf{and}\ x_n \leq x_k \ \textbf{and}\ l_k < l_n + 1)$$

$$A((n, x_n, l_n), (k, x_k, l_k)) = \{(n, x_n, l_n), (k, x_k, l_n + 1)\}$$

A second program *max* is required to find the maximum of the lengths l_i.

$$max(M) = \Gamma(R_{max}, A_{max})\,(M) \quad \textbf{where}$$
$$R_{max}\,((n, x_n, l_n), (k, x_k, l_k)) = (l_k \leq l_n)$$
$$A_{max}\,((n, x_n, l_n), (k, x_k, l_k)) = \{(n, x_n, l_n)\}$$

The GAMMA program evaluating the longest upsequence is the following:

$$lus(M) = l_k \textbf{ where } \{(k, x_k, l_k)\} = max(lup(M_0))$$

Program *lup* performs a relaxation and *max* is a reduction.

The majority element problem

The majority element of a multiset M is an element occurring more than $card(M)/2$ times in the multiset. We propose first a solution to the problem of finding the majority element, assuming that such an element exists:

$$maj\_elem(M) = \Gamma(R, A)(M) \quad \textbf{where}$$
$$R(x, y) = (x \neq y)$$
$$A(x, y) = \{\}$$

This solution can be seen as an abstract (and parallel) version of the "hands-in-the-pocket" presentation given in [14]. Let us now discharge the assumption of the existence of a majority element. We require a program yielding the majority element if it exists and $\perp$ otherwise. Elements of the original multiset are represented by pairs (v, n), where v is the value and n is the number of occurrences of v represented by the pair; initially $n = 1$ for all values.

$$maj\_elem(M) = P_1(P_2(M \times \{1\}), card(M)) \quad \textbf{where}$$
$$P_2(M) = \Gamma((R_2, A_2))\,(\Gamma((R_1, A_1))(M)) \quad \textbf{where}$$
$$R_1((v_1, n_1), (v_2, n_2)) = (v_1 = v_2)$$
$$A_1((v_1, n_1), (v_2, n_2)) = \{(v_1, n_1 + n_2)\}$$
$$R_2((v_1, n_1), (v_2, n_2)) = (n_1 \geq n_2)$$
$$A_2((v_1, n_1), (v_2, n_2)) = \{(v_1, n_1)\}$$
$$P_1(M, c) = \textbf{if } n > c/2 \textbf{ then } v \textbf{ else } \perp$$
$$\textbf{where } \{(v, n)\} = M$$

This program is an elaborated version of the previous one. The numbers of occurrences are necessary to decide whether the resulting value is a majority element. Reaction (R_1, A_1) performs a relaxation to evaluate the number of occurrences of every value in the multiset and (R_2, A_2) is a reduction yielding an element whose number of occurrences is maximum. This is yet another problem usually presented as inherently sequential which has a nice

GAMMA solution with a high potential for concurrency.

GRAPH PROBLEMS

Many problems are naturally formulated in terms of graphs. We show in this section how three fundamental graph problems can be dealt with in GAMMA: the connectivity problem, the shortest-path problem and the minimum spanning tree problem.

Connectivity

An undirected graph is a collection of *vertices* (or *nodes*) and *edges*. Vertices are simple objects and edges are connections between two vertices. A *path* from vertex x to y is a list of vertices in which successive vertices are connected by edges in the graph. A graph is *connected* if there is a path from every node to every other node. The problem we consider here is to decide whether a graph is connected or not. The idea of the GAMMA program is to build bigger and bigger aggregates of connected nodes: the graph is connected if and only if all the nodes can ultimately be gathered into one aggregate. This program is a typical example of application of the reduction strategy.

Graphs are represented as multisets of vertices and edges: a vertex x is denoted by the singleton $\{x\}$ and an edge connecting x to y is represented by a pair (x,y). We use the boolean function *vertices* to test whether an element of the multiset is a set of vertices. The GAMMA program is the following:

$$connected\ (G) = singleton\ (\Gamma((R_1,A_1),(R_2,A_2))\ (G))\ \textbf{where}$$
$$R_1(v,w,(m,n)) = vertices(v)\ \textbf{and}\ vertices(w)\ \textbf{and}\ m \in v\ \textbf{and}\ n \in w$$
$$A_1(v,w,(m,n)) = \{v+w\}$$
$$R_2(v,(m,n)) = vertices(v)\ \textbf{and}\ m \in v\ \textbf{and}\ n \in v$$
$$A_2(v,(m,n)) = \{v\}$$

Function *singleton* tests whether the multiset is a singleton or not. Reaction (R_1,A_1) consumes three elements of the multiset: two sets of vertices v and w and an edge connecting one element of v to one element of w. It yields a larger set of connected vertices $v+w$. Reaction (R_2,A_2) is just a specialization of (R_1,A_1) which involves only one vertex. It is used to remove edges of the graph that are no longer necessary. It is easy to see that if all the nodes of the graph are connected they will eventually be gathered into one set of vertices and all edges will be removed by (R_2,A_2). If the graph is not connected, the nodes will never be gathered into one set and the resulting multiset cannot be a singleton.

Shortest path

We consider now a weighted directed graph: a cost is associated with each edge and edges are "one-way". The length of a path is the sum of the costs of its edges. The problem consists in finding the length of the shortest path from x to y for all pairs of vertices (x,y). The initial multiset contains one triple (n,m,c) per pair of vertices (n,m) in the graph. If an edge (n,m) is present in the initial graph, then c is the cost of the edge (n,m), otherwise the value of c is ∞ (∞ satisfying the property $\forall n, n < \infty$). The resulting multiset is composed of triples (n,m,c) where c represents

the length of the shortest path from n to m. If there is no path from n to m the value of c is ∞. The GAMMA program is the following:

$$shortest\_path(G) = \Gamma((R,A))\,(G)\ \textbf{where}$$
$$R((v_1,v_2,c_{12}),(v_2,v_3,c_{23}),(v_1,v_3,c_{13})) = c_{13} > c_{12} + c_{23}$$
$$A((v_1,v_2,c_{12}),(v_2,v_3,c_{23}),(v_1,v_3,c_{13})) = \{(v_1,v_2,c_{12}),(v_2,v_3,c_{23}),(v_1,v_3,c_{12}+c_{23})\}$$

The cost associated with a pair of vertices represents the length of the shortest known path between the two vertices. The program performs a relaxation, decreasing costs until they represent the length of the shortest path.

Minimum spanning tree

We define the cost of a weighted graph as the sum of the weights of its edges. The problem is to find a minimum spanning tree which is defined as a subgraph of minimum cost connecting all the vertices of the graph. The idea for solving this problem in GAMMA is to proceed by reduction, building larger and larger local minimum spanning trees until all the vertices are included into one single tree which is a minimal spanning tree of the graph. Initially local minimum spanning trees are just individual vertices and two local spanning trees are aggregated using the shortest edge between them. We represent a graph as a multiset of quadruples (V,ST,E,M). V is a set of vertices connected by the edges in ST; the graph represented by the edges in ST is a minimum spanning tree connecting the vertices in V. E is the set of triples (n,m,c) such that $n \in V$ and there is an edge of cost c connecting n and m; M is a triple (n,m,c) of E of minimal cost such that $m \notin V$. The initial multiset contains one element $(\{n\},\emptyset,E,(n,m,c))$ per vertex n of the graph. The following GAMMA program computes the minimum spanning tree of a graph:

$$min\_st(G) = ST \quad \textbf{where} \quad \{(V,ST,E,M)\} = \Gamma((R,A))\,(G) \quad \textbf{where}$$
$$R((V_1,ST_1,E_1,(n_1,m_1,c_1)),(V_2,ST_2,E_2,(n_2,m_2,c_2))) = m_1 \in V_2$$
$$A((V_1,ST_1,E_1,(n_1,m_1,c_1)),(V_2,ST_2,E_2,(n_2,m_2,c_2))) =$$
$$\{(V_1+V_2,\ ST_1+ST_2+\{(n_1,m_1,c_1)\},\ E_1+E_2,\ min(V_1+V_2,\ E_1+E_2)\}$$

$$min(V,E) = (n,m,c) \ \textbf{where} \ \{(n,m,c)\} = \Gamma((R'',A''))\,((\Gamma((R',A'))\,(V+E)) - V) \ \textbf{where}$$
$$R'(n,(n',n,c)) = True$$
$$A'(n,(n',n,c)) = \{n\}$$
$$R''((n_1,m_1,c_1),(n_2,m_2,c_2)) = c_1 \leq c_2$$
$$A''((n_1,m_1,c_1),(n_2,m_2,c_2)) = \{(n_1,m_1,c_1)\}$$

The program $min(V,E)$ performs a reduction: it eliminates edges connecting two elements of E (reaction (R',A')) and then removes from the multiset all edges but one (reaction (R'',A'')): the residual element of the multiset is an edge of minimal cost.

GEOMETRIC PROBLEMS

The problems we have studied so far have involved numbers, texts or graphs; we describe in this section two problems involving points: the convex hull problem and an image processing application. Both problems are handled by relaxation, starting with an approximation of the result.

Convex hull

The convex hull of a set of points in the plane is defined to be the smallest convex polygon containing them all. A convex polygon has the property that any line connecting two points inside the polygon must lie entirely inside the polygon [10]. It is easy to show that the vertices of the convex hull of a set of points P are elements of P. The GAMMA program is based on the following property: a point of the original set is a vertex of the convex hull if and only if it is not strictly inside a triangle made of three other points of the set. The initial multiset contains the coordinates of all the points and computation proceeds by throwing out points that fall inside a triangle.

$$convex(Points) = \Gamma((R,A))\ (Points) \quad \textbf{where}$$
$$R((i_1,j_1),(i_2,j_2),(i_3,j_3),(i_4,j_4)) = inside\ ((i_4,j_4),((i_1,j_1),(i_2,j_2),(i_3,j_3)))$$
$$A((i_1,j_1),(i_2,j_2),(i_3,j_3),(i_4,j_4)) = \{(i_1,j_1),(i_2,j_2),(i_3,j_3)\}$$

Function *inside* takes a point and a triangle and returns *true* if the point falls inside the triangle. It involves intricate tests and computations on the coordinates of the points. A generalization of this function to any polygon is given in [10].

An image processing application

A theory called mathematical morphology was proposed some years ago to solve problems in image processing applications [20]. We describe the treatment in GAMMA of edge detection, a classical image processing problem, along these lines. Each point of the image is originally associated with a grey intensity level; then an intensity gradient is computed at each point and edges are defined as the points where the gradient is greater than a given threshold T. The gradient at a point is computed relative to its neighbours: only points at a distance d less than D are considered for the computation of the gradient. The gradient at a point is defined in the following way:

$$G(P) = maximum(neighbourhood) - minimum(neighbourhood)$$
$$\textbf{where}\ neighbourhood = \{intensity(P') \mid distance(P,P') < D\}$$

Functions *maximum* and *minimum* yield respectively the maximum and the minimum of a multiset of values. The GAMMA program uses a multiset of quadruples (P,l,min,max); P is a pair representing the coordinates of a point, l is the intensity level of the point and *min* and *max* are the current values of *minimum(neighbourhood)* and *maximum(neighbourhood)* of the point. The initial value of *min* and *max* is l. The evaluation consists in decreasing *min* and increasing *max* until the limit values are reached. A second GAMMA program removes from the multiset the points where the gradient is less than the threshold.

$edges(Points) = select(\Gamma((R_1,A_1),(R_2,A_2))\,(Points))$ **where**

$R_1((P,l,min,max),(P',l',min',max')) = (distance(P,P')) < D$ **and** $(l' < min)$

$A_1((P,l,min,max),(P',l',min',max')) = \{(P,l,l',max),(P',l',min',max')\}$

$R_2((P,l,min,max),(P',l',min',max')) = (distance(P,P')) < D$ **and** $(l' > max)$

$A_2((P,l,min,max),(P',l',min',max')) = \{(P,l,min,l'),(P',l',min',max')\}$

$select(Points) = \Gamma((R_1,A_1),(R_2,A_2))\,(Points)$ **where**

$R_1((P,l,min,max)) = max - min < T$

$A_1((P,l,min,max)) = \{\}$

$R_1((P,l,min,max)) = max - min \geq T$

$A_1((P,l,min,max)) = \{(P,l)\}$

The treatment of a larger image processing application (namely the recognition of the tridimensional topography of the vascular cerebral network) in GAMMA is described in [7].

PROCESS SYNCHRONIZATION

In this section, we present solutions in GAMMA to some problems usually chosen in concurrent programming as typical challenges to test the expressiveness of synchronization constructs: the dining philosophers problem and a resource allocation problem. Before describing these programs let us emphasize that we use GAMMA in a quite different way in this section: we are not interested in the result of the evaluation (the problems we solve here typically require non terminating programs) but rather in the possible values of the multiset during the computation. In these examples, the multiset is seen as a representation of the state of the system. Furthermore, GAMMA being a very high level formalism, it makes it possible to express very concise solutions to these problems. Of course these solutions must rely on a correct implementation of the formalism (one possible implementation is described in [2]).

The dining philosophers

The formulation of the problem is the following: five philosophers spend their lives thinking and eating. They share a common dining room where there is a circular table surrounded by five chairs, each belonging to one philosopher. In the center of the table there is a bowl of spaghetti which is endlessly replenished, and the table is laid with five forks. However the spaghetti is so hopelessly entangled that two forks are necessary simultaneously to eat.

The problem is to find a protocol ensuring two fundamental properties: (1) there is no deadlock in the system and (2) there is no starvation (any hungry philosopher will be able to eat after a finite amount of time). Each philosopher may only use his two adjacent forks, and may only eat for a finite amount of time.

The state of the system is represented by a multiset containing the available forks and the identities of eating philosophers. The initial multiset $\{F_0, F_1, F_2, F_3, F_4\}$ contains five forks; imagine that philosopher P_1 is allowed to eat with forks F_1 and F_2, then the new multiset will be: $\{F_0, P_1, F_3, F_4\}$. When P_1 has finished eating he returns his two forks so that another philosopher may use them. The following GAMMA program expresses this protocol ($\oplus$ represents addition modulo 5):

$philosophers = \Gamma((R_1, A_1)\ (R_2, A_2))\ (\{F_0, F_1, F_2, F_3, F_4\})$ **where**

$R_1\ (F_i, F_{i\oplus 1}) = True$

$A_1\ (F_i, F_{i\oplus 1}) = \{P_i\}$

$R_2\ (P_i) = True$

$A_2(P_i) = \{F_i, F_{i\oplus 1}\}$

The GAMMA paradigm allows a direct expression of a deadlock-free solution because the basic synchronization facility offered is precisely the *atomic operation* on a collection of items. The absence of starvation is guaranteed by the assumption of *fairness* in the selection process. A parallel implementation of GAMMA exhibiting these properties is described in [2].

Three solutions to this problem are presented in [3]. First, philosophers try to collect independently their left and right forks and this may lead to a deadlock situation because all philosophers may decide to take their left fork simultaneously ... before trying to seize their right forks. A second solution allows a philosopher to seize atomically two forks; this solution uses the monitor concept which offers nice synchronization facilities. However, this solution does not guarantee the absence of starvation because the philosophers can conspire in order to prevent one of their number from getting his two forks. This problem cannot be solved without introducing some asymmetry into the protocol. This can be done by allowing only four philosophers around the table at a given time.

A resource allocation problem

Consider a computing system where n users $U_1, ..., U_n$ share a common pool of resources $\{r_1, ..., r_k\}$. Each user can be in one of the three following states: *passive*, *waiting* for a resource and *busy*. Transitions between these three states can occur as follows: *passive* $\rightarrow$ *waiting* $\rightarrow$ *busy* $\rightarrow$ *passive* .... The state of the system is represented by a multiset containing the identities of free resources, the identities of busy users represented by pairs ($user_i$, *resource allocated*$_j$), the identities of waiting users represented by pairs ($user_i$, *waiting*) and the identities of passive users represented pairs ($user_i$, *passive*). The following GAMMA program solves the problem:

$resource\_management = \Gamma((R_1, A_1), (R_2, A_2), (R_3, A_3))\ (\{r_1, ..., r_k, (U_1, passive), ..., (U_n, passive)\})$

where

$R_1((U_i, passive)) = True$

$A_1((U_i, passive)) = \{(U_i, waiting)\}$

$R_2((U_i, waiting), r_j) = True$

$A_2((U_i, waiting), r_j) = \{(U_i, r_j)\}$

$R_3((U_i, r_j)) = True$

$A_3((U_i, r_j)) = \{(U_i, passive), r_j\}$

We use a single multiset to represent users and resources. It would be possible to extend GAMMA with appropriate syntactic sugar to be able to use a separate multiset for each user and for resources. This would allow us to keep the programming closer to the logic of the problem, but would not change the spirit of the solution.

CONCLUSION

We have shown in this paper that the multiset transformation paradigm can be used to provide elegant solutions to a wide range of programming problems. We have exhibited three basic programming principles (namely *relaxation*, *expansion* and *reduction*) that we have put into practice to build all the programs presented. It may be surprising at first glance that such a simple formalism possesses enough expressive power to solve in a natural way such different problems. We believe that the most fruitful approach to programming language design is to start with a few basic principles and to exploit them as far as possible. In this respect, we cannot resist the temptation to quote E.W.Dijkstra eighteen years ago: "*Another lesson we should have learned from the recent past is that the development of "richer" or "more powerful" programming languages was a mistake in the sense that these baroque monstrosities, these conglomerations of idiosyncrasies, are really unmanageable, both mechanically and mentally. I see a great future for very systematic and very modest programming languages.*" [8].

Program derivation

It should be clear that GAMMA is not a programming language in the usual sense of the term. GAMMA programs are executable but any straightforward implementation would be extremely inefficient. We see GAMMA as a convenient intermediate language between specifications and programs: it is possible to express in GAMMA the *idea* of an algorithm without any detail about the order of execution or the memory management. For example, you can tell in GAMMA that your strategy to sort a sequence is to exchange values, that you want to find the convex hull by removing the points inside a triangle, and so forth. Further program derivation may specialize these abstract programs by choosing a particular data representation and a particular execution order. For example, selection sort, bubble sort or quicksort can be obtained from the abstract exchange sort program presented in this paper by applying different derivation strategies. Actually GAMMA is currently used in the context of program derivation [1,2]. GAMMA programs are first derived from a specification in first order logic; then a second derivation step is performed to obtain traditional programs (for sequential or parallel machines) from this GAMMA program. The derivation is based on the notions of variant and invariant properties [9,13]. Using an intermediate language like GAMMA makes the derivation easier because it allows a nice separation of concerns: the first step (the derivation of the GAMMA program) is related to the logic of the algorithm, whereas the second step expresses lower level choices such as data representation or execution order. In order to show the relevance of GAMMA for program derivation, let us consider an informal specification of a sorting program:

$$
\begin{aligned}
M = sort(M_0) \;<=>\; & \\
& \forall (i,v),(j,w) \in M,\ i > j => v \geq w \quad \textbf{and} && (1)\\
& \{i \mid (i,v) \in M\} = \{1, \ldots, card(M_0)\} \quad \textbf{and} && (2)\\
& \{v \mid (i,v) \in M\} = M_0 && (3)
\end{aligned}
$$

Let us choose *(1)* as the variant property: this implies that the program has to proceed while the negation of *(1)* holds:

$$not(1) = \exists (i,v), (j,w) \in M,\ i > j \textbf{ and } v < w$$

If we look at the definition of GAMMA, we can see that the reaction condition precisely states the condition to be satisfied by some elements of the set for the computation to proceed. So the reaction condition can be derived from the negation in a straightforward way:

$$R((i,v),(j,w)) = (i > j) \textbf{ and } (v < w)$$

According to the invariant *(2)* ***and*** *(3)*, the indexes and values cannot be modified; so the only possible action consists in exchanging values and indexes:

$$A((i,v),(j,w)) = \{(i,w),(j,v)\}$$

The termination of the derived program can be shown using a multiset ordering. The invariant can then be used to derive an array implementation of the multiset (the set of indexes is constant and equal to $\{1,...,n\}$); the most straightforward orders of execution lead to selection sort, insertion sort or bubble sort but quicksort or heapsort can be derived as well (the latter involving a different data representation choice).

Another remarkable benefit of using GAMMA in the derivation is that GAMMA programs do not have any sequential implementation bias. In fact it is often the case that problems which are usually considered as inherently sequential turn out to have a parallel solution in GAMMA (and very often this is indeed the most natural solution). The spanning tree problem, the set partitioning problem and the longest upsequence example are good illustrations of this property.

Related works

It is interesting to note that even in the context of program construction many people have felt the need to describe algorithms in an abstract way very much in the spirit of GAMMA. Let us quote for example Sedgewick in [19], page 489:"*The Ford-Fulkerson method described above can be summarized as follows: "start with zero flow everywhere and increase the flow along any path from source to sink with no full forward edges or empty backward edges, continuing until there are no such paths in the network." But this is not an algorithm in the usual sense, since the method for finding paths is not specified, and any path at all could be used.*" This is followed by the description of a particular implementation of this idea for an algorithm. In the same way Goldberg and Tarjan have recently presented a new algorithm for the maximum-flow problem by giving first an abstract version in the form of two *applicability conditions* and associated *actions* [12].

The term "*production systems*" has been used rather loosely in artificial intelligence to denote systems described in terms of a global database, a set of production rules and a control system [17]. The production rules operate on the global database. Rules are associated with applicability conditions and the control system chooses the next rule to apply. A global termination condition is used to stop the computation. The globality of the production rules and termination conditions and the specification of control make these systems rather different from the GAMMA

formalism. These systems include a control component because the need to express a control strategy is often crucial in artificial intelligence applications. However very restricted forms of production systems called *decomposable production systems* exhibit locality properties allowing certain freedom in the order of application of the rules [17].

Actually several formalisms bearing some similarities to GAMMA have been proposed recently, which seems to denote a current trend towards high level languages of this form. In [6], Chandy and Misra describe a language, called UNITY (for Unbounded Nondeterministic Iterative Transformations), and its associated proof system. A UNITY program is essentially a declaration of variables and a set of multiple-assignment statements. Program execution consists in selecting nondeterministically some assignment statement, executing it and repeating forever. Nondeterministic selection is constrained by the following "fairness" rule: every statement is selected infinitely often. The main objective of UNITY is the systematic development of programs which can be implemented on different (distributed or centralized) architectures. Program development is carried out in two basic steps: first a correct program is derived from specifications, then this program is adapted to the target architecture; this adaptation is achieved by transformations of the original program in order to make control explicit. The multiple-assignment statement is used to express the mapping onto synchronous shared-memory architectures and the mapping onto asynchronous architectures is achieved by the partitioning of the statements of the program.
The major differences between the GAMMA model and UNITY may be summarized as follows:

- UNITY is based on a static data structure, the array, which makes less natural the treatment of problems involving data whose size may evolve dynamically.

- the multiple assignment statement, which is the basis of UNITY, entails an imperative style of programming.

- the notion of locality is not emphasized as it is in GAMMA. Computations which may be carried out in parallel are determined in a special design phase which aims at mapping the UNITY program onto a particular target machine. This mapping phase transforms a UNITY program without explicit parallelism into an explicitly parallel program; this phase is carried out as rigorously as possible but still remains informal.

- the associated proof techniques are more complex and program derivation is more laborious especially when dealing explicitly with parallelism.

However we should mention that the goal of the proponents of UNITY was a bit different to ours since we do not attempt to model within the GAMMA formalism the execution of programs on various kinds of architectures (although GAMMA programs can also be mapped on various architectures).

A programming notation, called *associons*, has been proposed in [18]. Essentially, an associon is a tuple of names defining a relation between entities represented by these names. The state of the computation can be changed by the creation of new associons representing new relations deduced from the already existing ones. Such deductions are described in a closure statement whose execution may be decomposed into several simple activities which may be run in parallel.

The spirit of the proposal is quite similar to the ideas which have led to the GAMMA model. However several important differences may be highlighted:

- the locality principle which is of prime importance in the GAMMA formalism is not emphasized in the associons model. This comes essentially from the fact that negated presence conditions (which correspond to global properties on the set of tuples) are allowed in the associon model.

- unlike GAMMA, the associon model is deterministic; in order to ensure this property, the execution of an action (creation of new associons) cannot invalidate another action; this entails a potential independence between actions but introduces restrictions on the type of actions which are permitted.

- GAMMA is based on multisets while the associon model is based on sets. We find that the extra degree of freedom provided by the use of multisets is very useful as far as program construction is concerned. Furthermore, this freedom is necessary to satisfy the locality property (no global test is needed to check that a produced element is not already present).

Let us also mention the Linda approach [5,11]. Linda contains a few simple commands operating on a tuple space. Adding these tuple-space commands to an existing base language produces a parallel programming dialect. Linda's model is based on generative communications. If two processes need to communicate, the producer adds a tuple to a particular domain, and the consumer may read (destructively or not) this information from the tuple space. Data and program objects are represented in a uniform way as passive or active tuples. Of course, several processes may be active on the same tuple space, thus allowing parallel tuple processing. Linda is a very simple communication model that can easily be incorporated into existing programming languages. As such, Linda is not a computational model. However, in the same way as the GAMMA model, it shows clearly how advanced data structuring facilities such as tuple spaces or multisets may greatly simplify the programming task.

Last but not least, let us mention the Chemical Abstract Machine (or *cham*) proposed by Berry and Boudol [4] to model asynchronous concurrent computations. The *cham* is an elaboration on the original GAMMA formalism introducing the notion of subsolution enclosed in a membrane. It is shown that models of algebraic process calculi can be defined in a very natural way using a *cham*. The fact that concurrency is the primitive built-in notion makes proofs far easier than in the usual process semantics.

REFERENCES

1. Banâtre, J.-P., and Le Métayer D. A new computational model and its discipline of programming. INRIA Research Report 566 (Sept. 1986).
2. Banâtre, J.-P., Coutant, A., and Le Métayer, D. A Parallel Machine for Multiset Transformation and its Programming Style. *Future Generation Computer Systems.* 4, (1988), 133-144.
3. Ben-Ari, M. *Principles of concurrent programming.* Prentice/Hall International, 1982.
4. Berry, G., and Boudol, G. The Chemical Abstract Machine. In *Proceedings of ACM Symp. on Principles of Programming Languages* (San Francisco, Calif.). ACM, New York, 1990, pp. 81-94.
5. Carriero, N., and Gelernter, D. Linda in Context. *Commun. ACM* 32, 4 (April 1989), 444-458.
6. Chandy, M., and Misra, J. *Parallel Program Design: a Foundation.* Addison-Wesley Publishing Company, 1988.
7. Creveuil, C., and Moguerou, G. Dérivation d'un algorithme de segmentation d'images: un exemple d' application du formalisme GAMMA. INRIA Research Report 1049 (June 1989).
8. Dijkstra, E. W. The humble programmer. *Commun. ACM* 15, 10 (October 1972), 859-866.
9. Dijkstra, E. W. *A Discipline of Programming.* Prentice-Hall, Englewood Cliffs, N.J., 1976.
10. Dromey, R. G. *Program Derivation.* Addison-Wesley Publishing Company, International Computer Science Series, 1989.
11. Gelernter, D. Generative Communication in Linda. *ACM Trans. Prog. Lang. Syst.* 7, 1 (Jan. 1985), 80-112.
12. Goldberg, A. V., and Tarjan, R. E. A new approach to the maximum-flow problem. *Journal of ACM* 35, 4, (Oct. 1988), 921-940.
13. Gries, D. *The Science of Programming.* Springer Verlag, New York, 1981.
14. Gries, D. A hands-in-the-pocket presentation of a k-majority vote algorithm, in *Formal Development of Programs and Proofs*, ed. E. W. Dijkstra (University of Texas at Austin Year of Programming series). Addison-Wesley Publishing Company, 1990, pp. 43-46.
15. Hoare, C. A. R. Communicating Sequential Processes, *Commun. ACM* 21, 8 (Aug. 1978), 666-677.
16. Knuth, D. *Seminumerical Algoritms. The Art of Computer Programming.* Addison-Wesley Publishing Company, 1969.
17. Nilsson, N. J. *Principles of Artificial Intelligence.* Tioga publishing company, Palo Alto, 1980.
18. Rem, M. Associons: A Program Notation with Tuples instead of Variables. *ACM Trans. Prog. Lang. Syst.* 3, 3 (July 1981), 251-262.
19. Sedgewick, R. *Algorithms.* Addison-Wesley Publishing Company, 1988.
20. Serra, J. *Image analysis and mathematical morphology.* Academic Press, 1982.

Liste des dernières publications internes parues à l'IRISA

PI 516 **COMMENT INTRODUIRE LA CONTIGUITE EN ANALYSE DES CORRESPONDANCES ? Application en segmentation d'image.**
Brigitte ESCOFIER, Habib BENALI, Kaddour BACHAR
Février 1990, 26 Pages.

PI 517 **MACHINE MODELING AND LOOP OPTIMIZATION FOR HORIZONTAL MICROCODED MACHINES**
François BODIN, François CHAROT
Février 1990, 24 Pages.

PI 518 **MULTISCALE SYSTEM THEORY**
Albert BENVENISTE, Ramine Nikoukhah, Alan S. Willsky.
Février 1990, 30 Pages.

PI 519 **PANDORE : A SYSTEM TO MANAGE DATA DISTRIBUTION**
Françoise ANDRE, Jean-Louis PAZAT, Henry THOMAS
Février 1990, 14 Pages.

PI 520 **SCHEDULING AFFINE PARAMETERIZED RECURRENCES BY MEANS OF VARIABLE DEPENDENT TIMING FUNCTIONS**
Christophe MAURAS, Patrice QUINTON
Sanjay RAJOPADHYE, Yannick SAOUTER
Février 1990, 14 Pages.

PI 521 **COMPUTABILITY OF RECURRENCE EQUATIONS**
Yannick SAOUTER, Patrice QUINTON
Février 1990, 28 Pages.

PI 522 **PROGRAMMING BY MULTISET TRANSFORMATION**
Jean-Pierre BANATRE, Daniel LE METAYER
Mars 1990, 26 Pages.

PI 523 **GOTHIC MEMORY MANAGEMENT : A MULTIPROCESSOR SHARED SINGLE LEVEL STORE**
Béatrice MICHEL
Mars 1990, 20 Pages.

Imprimé en France
par
l'Institut National de Recherche en Informatique et en Automatique

ISSN 0249-6399