ARCHITECTURES VLSI POUR L'ESTIMATION DE MOUVEMENT EN CODAGE D'IMAGES

VLSI Architectures for Motion Estimation in Image Coding

G. PRIVAT, M. RENAUDIN

CENTRE NATIONAL D'ETUDES DES TÉLÉCOMMUNICATIONS, CNET-Grenoble
BP 98, 38243 Meylan Cedex, France

RÉSUME

La recherche de vecteurs mouvement correspondant à des translations sur des blocs, utilisés comme prédicteurs dans une boucle de codage différentiel inter-images, constitue la tâche la plus intensive en calcul dans les algorithmes de codage d'images les plus performants actuellement préconisés, dans une gamme de débits couvrant visiophonie numérique, visioconférence, télédistribution numérique et télévision haute-définition. L'utilisation de circuits VLSI spécialisés à architecture massivement parallèle y est donc fortement valorisée. Les spécifications d'interface externes pour l'insertion d'un tel circuit dans un codeur sont précisées, le spectre des options architecturales est passé en revue, et une solution à parallélisme maximal, adaptable à un grand nombre de configurations de calcul , basée sur une partie opérative semi-systolique implémentant l'algorithme de recherche exhaustive (25Gop/s), est présentée en détail.

SUMMARY

Block-matching motion estimation using exhaustive search is the most computation intensive task in state of the art image coding algorithms, where displacement vectors are used as predictors in DPCM inter-frame coding loops ; hence the interest of designing special purpose chips based on highly parallel architectures, for use in videophone, videoconference, digital TV distribution and HDTV codecs. External interface specifications for embedding such a circuit in coders are detailed. A comprehensive overview of all possible architectural solutions for the problem is given, based on high-level synthesis methods. A maximally parallel solution is presented in full detail. It offers the possibility to adapt to various external configurations and parameters with a flexible and cascadable chip, based on a semi-systolic (25 Gop/s) operative part and involving minimal control overhead.

I. INTRODUCTION

Les algorithmes d'estimation de mouvement peuvent être regroupés en deux classes [1-5]:

- les algorithmes récursifs qui extrapolent une connaissance du déplacement à l'image "i" pour estimer le déplacement correspondant à l'image "i+1", en opèrant soit au niveau pixel soit au niveau bloc.
- les algorithmes ayant recours à la corrélation de blocs (block-matching) qui estiment un déplacement, en général limité à des translations, en calculant la position du pic d'une fonction d'intercorrélation bidimensionnelle entre blocs de deux images successives.

Ces techniques d'estimation de mouvement peuvent être mises en oeuvre dans différents types de codeurs : par transformée, codeurs interpolatifs, codeurs prédictifs. Les codeurs *prédictifs à compensation de mouvement* ont été les plus largement étudiés. Ils peuvent être regroupés en deux classes, caractérisées par le type de prédicteur à compensation de mouvement utilisé :

- les prédicteurs à compensation de mouvement *a priori* qui mettent en oeuvre les algorithmes récursifs d'estimation et qui ne transmettent que l'erreur de prédiction au décodeur,
- les prédicteurs à compensation de mouvement *a posteriori* (schéma très simplifié figure 1) qui utilisent généralement des algorithmes à corrélation de blocs pour estimer le déplacement d'objets. Ces prédicteurs imposent de transmettre au récepteur le vecteur caractéristique du mouvement estimé en plus de l'erreur de prédiction, car ce schéma de prédiction utilise une fenêtre non causale au voisinage des points codés. Il est bien évident que le coût en termes de débit occasionné par la transmission du vecteur mouvement est largement compensé par les performances du prédicteur en terme de réduction de l'entropie du signal d'erreur de prédiction. Les algorithmes de ce type sont actuellement considérés comme permettant d'atteindre le meilleur compromis complexité-performances pour les codeurs bas débits.

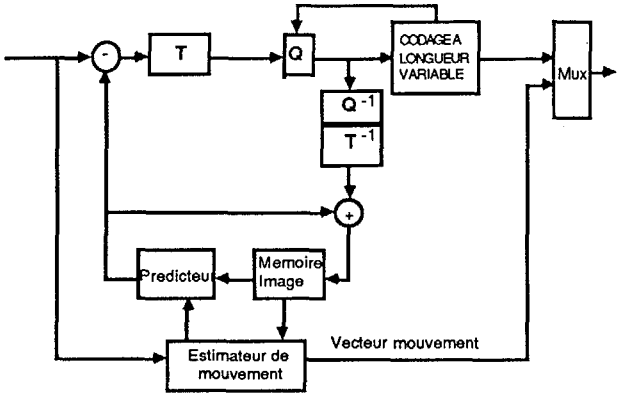

Figure 1 : Codeur hybride à compensation de mouvement.

II. Spécifications Externes

Le principe de l'estimation de mouvement par corrélation de blocs est présenté figure 2. Pour chaque bloc *de référence* de taille bxb pixels de l'image courante, un bloc de même taille est recherché dans l'image précédente qui minimise un critère de distance par rapport au bloc de référence, à l'intérieur d'une fenêtre de recherche de taille (b+2d)(b+2d)dans l'image précédente, centrée sur la projection du bloc de référence.

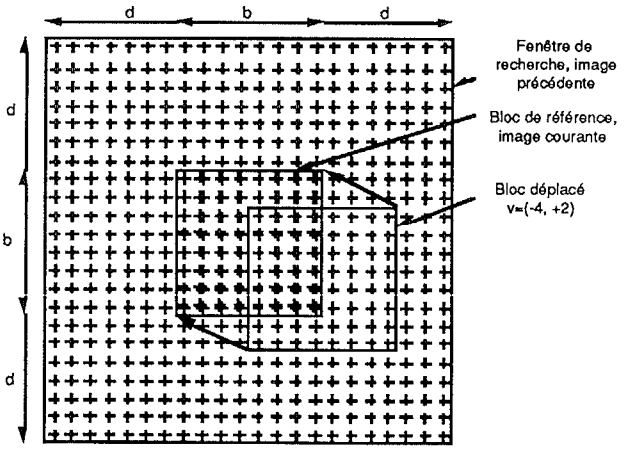

Figure 2 : Principe de l'estimation de mouvement par corrélation de blocs .

Les prédicteurs du codeur et du décodeur utiliseront alors les valeurs des pixels de ce bloc de l'image précédente comme estimation des pixels correspondants du bloc de référence de l'image courante, si ce bloc est *à mouvement compensé* .

Il faut noter que, contrairement à d'autres fonctions de l'algorithme de codage, l'estimation de mouvement utilisant la corrélation de blocs n'a pas à être normalisée dans le cadre d'un standard, le décodeur utilisant seulement la valeur du vecteur déplacement fournie par le codeur, sans qu'il lui soit nécessaire de connaître la manière dont ce vecteur a été obtenu. Ce ne serait pas le cas si l'on utilisait un algorithme récursif. Exceptées la valeur du déplacement maximum dans les deux directions (des valeurs typiques étant ±7 et ±15, correspondant respectivement à 8 et 10 bits alloués pour la transmission), et la taille du bloc de référence, (8x8, 8x16, 16x16), entière latitude est laissée pour trouver la meilleure association algorithme-architecture, sans interférer avec les normalisations à venir.

Critère de distance

Le critère d'erreur absolue moyenne, qui correspond à une distance L1, représente le meilleur compromis complexité-performance. Distance euclidienne et intercorrélation normalisée ont été également préconisées, mais correspondent à une amélioration des performances marginales au prix d'un considérable accroissement de la complexité de calcul

Algorithme de recherche

Un certain nombre de techniques de recherche sous-optimales (logarithmique, 3 pas, OTS, CDS, interpolation [3]) ont été introduites pour réduire la taille de l'espace de recherche. Elles trouvaient leur justification pour des implémentations sur processeur séquentiel, où une réduction du nombre total d'opérations était d'importance primordiale. Pour une réalisation VLSI spécialisée, par contre, il est bien connu que la régularité des calculs et la minimisation de la complexité de contrôle sont prioritaires. La technique de recherche exhaustive (par "force brutale") malgré le nombre très élevé de calculs qu'elle comporte, est parfaitement adaptée à une réalisation sur partie opérative massivement parallèle de type systolique.

Calculs annexes

Le seul résultat nécessaire demandé dans tous les cas est le vecteur de déplacement correspondant au bloc le plus fortement corrélé. Certaines sorties complémentaires peuvent être requises d'un estimateur de mouvement selon le codeur particulier dans lequel il peut être utilisé.

- La distance correspondant au bloc choisi pour l'estimation est utilisée par comparaison à un seuil quand une décision codé-non codé est prise pour un bloc à mouvement compensé.
- La distance correspondant au déplacement nul est comparée à la distance minimum pour une décision compensé-non compensé.
- Une résolution au demi-pixel ou quart de pixel peut être utile lorsque l'on opère sur des images préalablement sous-échantillonnées, comme c'est le cas en codage TV numérique. Elle peut être obtenue soit en arrondissant une moyenne pondérée par les distances correspondantes entre les 8 positions entourant immédiatement celle de la distance minimum, ou par un calcul analogue sur les positions correspondant au N plus petites distances (N>8). Cette dernière solution nécessite de réaliser un tri partiel sur les vecteurs de déplacement, les clés de tri correspondantes étant les distances.

III. Etude Architecturale

Partie opérative

On délimitera ici le problème au calcul d'un ensemble de distances correspondant à un couple fenêtre de recherche-bloc de référence. On peut étudier le calcul de distance à une dimension, le cas deux dimensions s'en déduisant très simplement.

Un calcul de ce type fait structurellement partie d'une classe très générale qui comprendrait entre autres : convolution, corrélation, distance L_2 (utilisée par exemple en quantification vectorielle), L_∞ , morphologie mathématique, étiquetage d'objets,etc.... Il s'agit dans tous les cas d'une opération globale (Σ, max, min) sur des résultats partiels résultant de l'application d'un opérateur binaire à un ensemble de couples d'échantillons pris dans deux fenêtres translatées l'une par rapport à l'autre. Une grande part de ce qui va être dit dans la suite peut être étendu à des calculs de cette classe.

En utilisant les notations de la figure 2, on cherche à calculer 2d+1 distances y(i) entre une séquence de b+2d échantillons x(i) et une séquence de b "coefficients" h(i). Une description complète du calcul peut être donnée dans un formalisme de type graphe de dépendance ou équations récurrentes uniformes, [7,8] dont deux variantes possibles sont données à la figure 3 (dans le cas b=2, d=3).

Selon les axes sur lesquels et le long desquels on effectue la projection de ce graphe, [6,7] trois grandes familles d'architectures peuvent être dérivées, qui sont représentées figure 4, et diffèrent principalement par la manière dont les 3 flots de données x, h et y circulent dans le tableau linéaire des processeurs élémentaires, l'un étant diffusé globalement, l'autre décalé en pipeline, et le troisième fixe. Dans chaque

cas, on a représenté la structure transposée qui peut être obtenue soit directement, soit par projection à partir du graphe dual de la figure 3. Un nombre illimité de variantes existe pour chacune de ces structures, par variation du degré de pipelinage ou multiplexage des processeurs élémentaires.

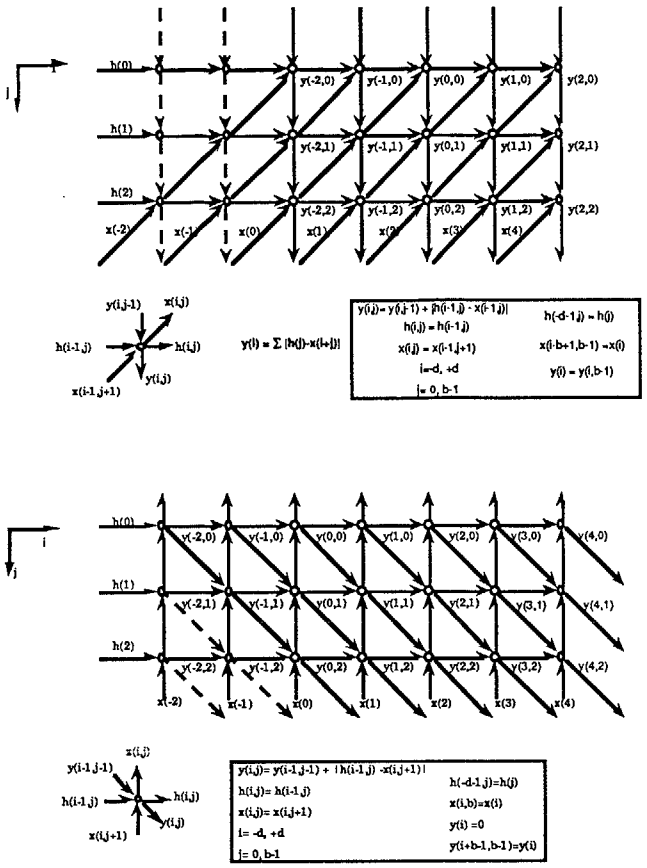

Figure 3 : Graphes de dépendance du calcul de distance dans le cas 1 D.

Le problème correspondant à deux dimensions avec un noyau de coefficients bxb et une fenêtre glissante (b+2d)(b+2d) est équivalent à un problème 1-D avec un noyau b^2+ 2dxb et une fenêtre de recherche $(b+2d)^2$, où des processeurs élémentaires vides (réduits aux retards correspondants) seraient insérés dans les intervalles correspondant aux "retours lignes".

Présentation des données

L'enchaînement des calculs relatifs à des blocs de référence adjacents pose un problème pour l'accès aux mémoires d'image externes . Dans le cas à 1 dimension, chaque ensemble de 2d+1 calculs relatifs à une fenêtre de référence donnée comporte une phase d'initialisation de b-1 cycles d'horloge. Il est possible de tirer avantage de ces temps vides pour recouvrir partiellement les calculs adjacents, en évitant la nécessité d'aller chercher de 2 à 5 fois depuis l'extérieur la partie de la fenêtre de recherche qui est commune aux calculs adjacents, réduisant ainsi le goulot d'étranglement d'accès aux mémoires Une élimination complète des recouvrements est possible de cette manière là si 2d<b. Au delà, il est possible d'utiliser E(2d/b) circuits en parallèle partageant l'accès à la mémoire. Ceci est illustré dans le cas 2-D à la figure 5. Malheureusement le bénéfice de cet artifice ne peut être pleinement exploité simultanément dans les deux dimensions. Il en résulte que la fréquence de chargement des pixels de l'image précédente dans le circuit est nécessairement supérieure à la fréquence d'échantillonnage, sauf à utiliser un buffer ligne complet à l'intérieur du circuit

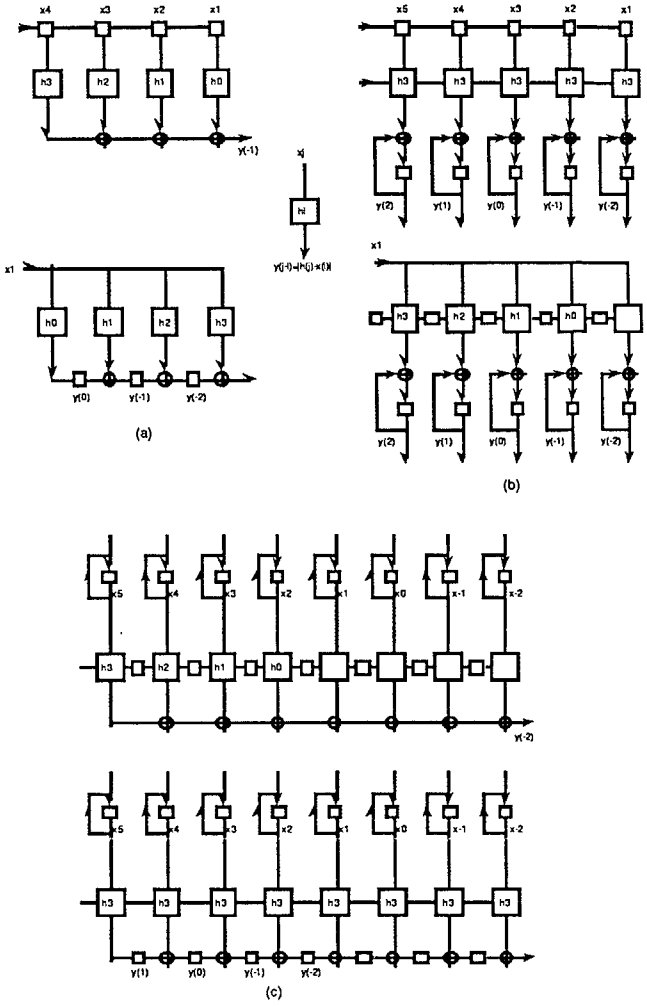

Figure 4 Architecture du calcul de distance dans le cas 1 D.

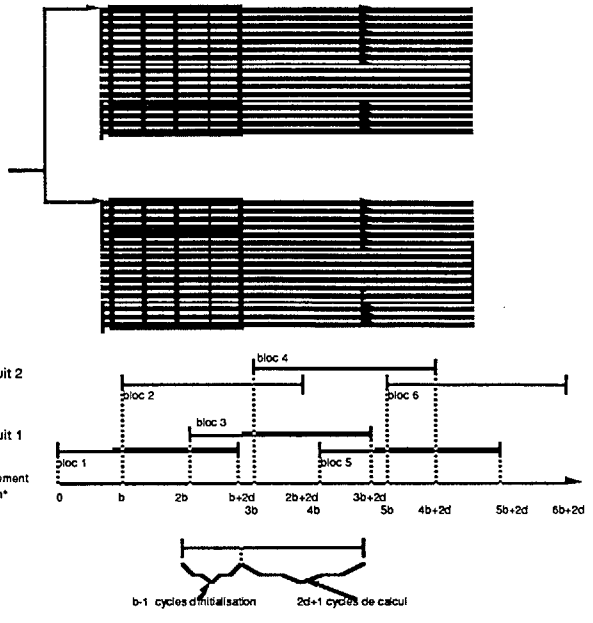

Figure 5 : Exploitation du recouvrement en utilisant E(2d / b) circuits .

IV. Architecture Détaillée

La figure 6 présente un synoptique d'une l'architecture à parallélisme maximal pour le calcul de distance sur des blocs 16x16

Partie opérative

Dérivée de la forme directe de la structure d'un filtre transversal (figure 4-a), la partie opérative est un réseau semi-systolique 16 par 16 de processeurs élémentaires dont chacun calcule la valeur absolue de la différence entre un pixel de l'image courante et le pixel correspondant de l'image précédente. Chaque processeur lit les opérandes dans deux mémoires distinctes. L'une mémorise les pixels du bloc de référence de l'image courante et est organisée en un tableau de 16 par 16 registres statiques distribués dans la partie opérative. L'autre mémorise une partie de la fenêtre de recherche, lue dans l'image précédente reconstruite, et sa structure est un tableau 16 par 16+2d registres dynamiques également distribués dans la partie opérative. L'utilisation d'unités à retard variable (VDU) permet d'agir sur la taille de ce tableau par l'intermédiaire du paramètre "d" qui peut prendre les valeurs comprises entre 1 et 15. La structure à parallélisme maximum de la partie opérative permet d'effectuer un calcul du critère de distance en un seul cycle d'horloge, à la fréquence d'acquisition des pixels de l'image précédente reconstruite.

L'accumulation des 256 valeurs absolue de différence inter-pixel est réalisée par un arbre d'addition pipeliné. Cet arbre est partagé en quatre sous-arbres identiques pour permettre le calcul simultané de quatre distances lorsque les blocs traités sont de taille 8 par 8.

Le traitement de blocs 32 par 32 est également possible en cascadant quatre circuits, chacun d'eux calculant une distance partielle sur un sous bloc de taille 8 par 32.

Opérateurs de tri

Le calcul des vecteurs et distances correspondant aux dix meilleures positions du bloc de référence dans la fenêtre de recherche peut être calculé au moyen d'une queue de priorité bit-parallèle implémentant un algorithme dual du tri de la bulle au moyen d'un tableau d'opérateurs bit de "comparaison-échange". Cette solution très simple est loin d'être optimale selon les critères classiques de complexité asymptotique [9] mais tout à fait adaptée à ce cas particulier d'application. Le réseau d'opérateurs bit est pipeliné en diagonal de façon à équilibrer les chemins critiques horizontaux et verticaux (figure 7).

Le séquencement global du circuit est réduit au minimum en raison du parallélisme maximal et de l'écoulement et interaction semi-systoliques des flots de données dans la structure.

Complexité et performances

Les estimations de complexité et surface (basées sur l'utilisation d'opérateurs disponibles en bibliothèque CMOS 1µ) conduisent à des circuits allant de 190 000 transistors sur 75 mm^2 (déplacement maximum "d" limité à +/- 7, sans tri des dix meilleures distances) à 330 000 transistors dans 114 mm^2 pour la variante la plus complète incorporant les flexibilités maximum.

Une fréquence d'horloge de 33 Mhz permet l'utilisation d'une telle architecture dans des applications allant du visiophone (norme C.I.F) à la télévision numérique (norme C.C.I.R 601). La puissance de calcul nominale du circuit serait de l'ordre de 25 Gop/s. Dans les applications (comme la télévision haute définition), où des débits supérieurs seraient requis, plusieurs circuits pourraient être utilisés en parallèle selon la configuration donnée figure 5.

References

[1] J.O. Limb and J.A. Murphy, "Measuring the speed of moving objects from television signals", IEEE Trans. Commun., VOL. COM-23, no. 4,pp. 474-478, Apr. 1976.

[2] C. Cafforio and F. Rocca, "Methods for measuring small displacements of television images", IEEE Trans. Inform. Theory, VOL. IT-22, no. 5,pp. 573-579, Sept. 1976.

[3] H.G. Musmann, P. Pirsch, H.J. Grallert, "Advances in Picture Coding", Proceedings of the IEEE, vol. 73, n° 4, April 1985, pp 523-548

[4] A.N. Netravali and J.D. Robbins, "Motion compensated television coding - Part 1", Bell Syst. Tech. J., VOL. 58,pp. 631-670, Mar. 1979.

[5] J.R. Jain and A.K. Jain, "Displacement measurement and its application in interframe image coding", IEEE Trans. Commun., VOL. COM-29, pp. 1799-1806, Dec. 1981.

[6] S.Y. Kung "On supercomputing with Systolic/Wavefront Array Computer" Proceedings of the IEEE, vol 72 n°7, July 1987

[7] G. Privat "Architectures Spécialisées de Circuits VLSI pour le Traitement Numérique du Signal", Thèse de Doctorat, Ecole Nationale Supérieure des Télécomunications, Paris, 1986

[8] P. Quinton, "The systematic Design of Systolic Arrays" MCNC Technical Report, May 1984

[9] C.D. Thompson, "The VLSI Complexity of Sorting", IEEE Trans. on Computers, vol. C-32, N° 12, Dec. 83, pp 1171-1184

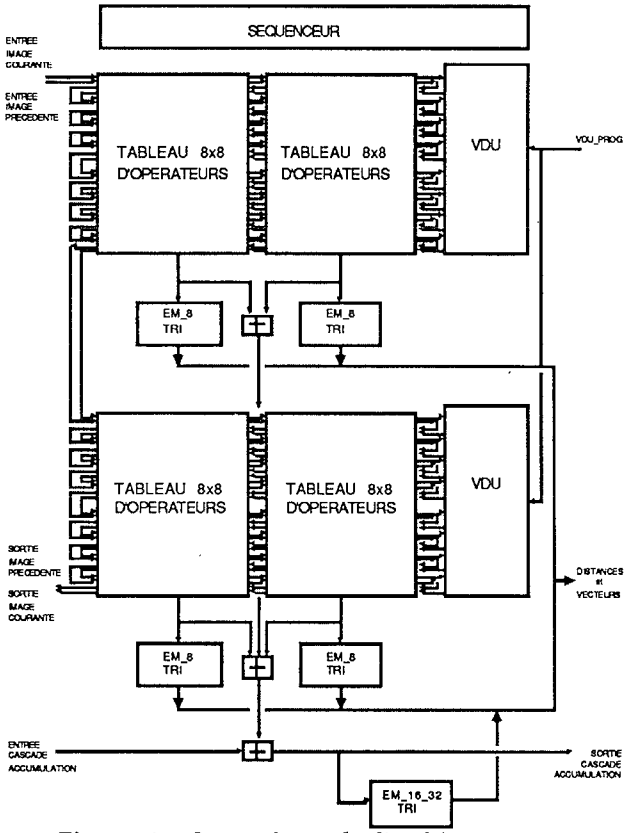

Figure 6 : Synoptique de l'architecture.

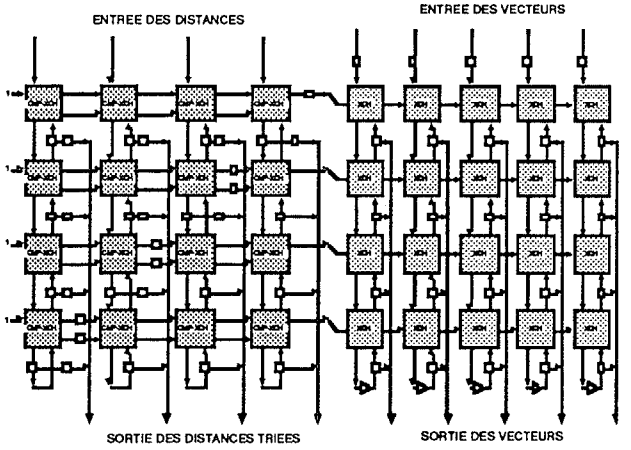

Figure 7 : Réseau de tri à pipelinage diagonal.