Rétroprojection 2D sur plateforme SOPC

Stéphane Mancini, Nicolas Gac , Michel Desvignes

Laboratoire des Images et des Signaux
46 avenue Félix Viallet, 38031 Grenoble Cedex 1, France
{stephane.mancini, nicolas.gac, michel.desvignes}@lis.inpg.fr

Résumé − Le développement et la diffusion des équipements TEP passent par la réduction des temps de calcul de la reconstruction des images acquises. Aussi cet arcticle présente une solution mixte logicielle/matérielle pour l'accélération de la reconstruction 2D sur une plateforme SoPC (System on Programmable Chip), la nouvelle génération de circuits reconfigurables. Le verrou technologique posé par la latence des accès mémoire est levé grâce au cache 2D Adaptatif et Prédictif (cache 2D-AP).

Abstract − Reduction of image reconstruction time is a key point for the development and spreading of PET scans. Thus this article presentes a hardware/software architecture which aims at accelerating the 2D reconstruction on a SoPC (System on Programmable Chip) plateform, the new generation of reconfigurable chip. Issue posed by the latency of memory accesses has been solved thanks to the 2D Aptative and Predictive cache (2D-AP cache).

1 Introduction

La reconstruction d'images en tomographie à émissions de positons (TEP) s'effectue à l'aide de processeurs classiques et demande un temps de calcul important. Elle est ainsi souvent découplée du processus d'acquisition. Or une reconstruction rapide des données acquises en imagerie TEP faciliterait le positionnement du patient et aiderait à détecter des problèmes potentiels survenus lors de l'acquisition. L'obtention rapide d'une image permettrait de réduire sensiblement la durée des examens cliniques TEP pour la détection précoce du cancer, l'évaluation de la propagation de la maladie et de la réponse au traitement. Cela permettrait de réduire les coûts de ces examens et favoriserait la diffusion de cette technique efficace d'imagerie médicale moléculaire. De plus, son utilisation pour la mammographie ou pour l'étude des petits animaux (micro PET) nécessite un système simple, flexible et rapide.

La reconstruction d'image s'effectue généralement en deux étapes : acquisition des données puis rétroprojection filtrée. Les travaux de ces dernières années portant sur l'acquisition en TEP humaine ont augmenté la qualité et la vitesse d'acquisition en utilisant du matériel dédié [1]. En effet, les processeurs de traitement numérique (DSP) et/ou les nouvelles technologies de logique programmable (FPGA) offrent une solution modulaire, adaptable et programmable, qui réduit les coûts et le temps en recherche et developpement. La rétroprojection n'est pas uniquement utilisée en TEP, elle l'est également dans d'autres domaines de la tomographie que ce soit en tomodensitométrie à rayons X ou en Tomographie à Emission Monophotonique (TEMP). Une autre famille d'algorithme se développe de plus en plus : les algorithmes itératifs. Ils augmentent sensiblement la qualité des images (artéfacts, ratio signal sur bruit) mais nécessitent un temps de reconstruction notablement supérieur aux techniques classiques de rétroprojection filtrée.

Il existe plusieurs implémentations de rétroprojection sur des clusters de PC [2, 3], sur du matériel dédié comme les ASICS ou les FPGA [4] (pour les scanners X) ou avec une architecture logicielle/matérielle [5]. Une autre stratégie est d'utiliser les processeurs 3D classiques comme les GPU (Graphic Processor Unit) pour accélérer la reconstruction [6]. Le principal goulot d'étranglement de tels systèmes est l'accès à la mémoire stockant les sinogrammes. En effet, pour reconstruire un pixel il faut parcourir entièrement tous les sinogrammes selon une courbe sinusoïdale. La recherche de parallélisation des calculs est rendu difficile par la limitation des accès mémoire par la bande passante de la mémoire principale.

Cet article présente un système de reconstruction par rétroprojection 2D implémenté sur une plateforme SOPC (System On Programmable Chip). L'originalité de ce système réside dans son architecture, solution efficace au problème soulevé par les accès aux mémoires externes. En effet, ces mémoires de type SDRAM sont peu coûteuses mais lentes et créent ainsi un goulot d'étranglement. Le cache 2D adaptatif et prédictif (cache 2D-AP) décrit en [7] constitue la base cette architecture. Cette stratégie s'est avérée efficace : elle permet de réduire le temps de reconstruction d'un ordre de magnitude par rapport aux solutions logicielles.

2 Objectifs

2.1 L'algorithme

L'algorithme implémenté effectue la rétroprojcetion des données acquises par le scanner. Ces dernières, appelées sinogrammes, forment la transformée de Radon de la fonction f représentant la densité du taux d'émission radioactive du volume observé. Le sinogramme est une matrice

image à deux dimensions à K colonnes. Chaque colonne k correspond à la projection orthogonale de f sur r détecteurs placés orthogonalement sur une ligne inclinée de l'angle $\theta_k = \frac{k*\pi}{K}$ par rapport à l'axe x. Ainsi, le point du sinogramme $S(\theta_k, r)$ est la somme des pixels sur une Ligne De Réponse (LDR) du scanner qui est perpendiculaire à l'axe des détecteurs et passe par le détecteur r. A partir de ce sinogramme, l'algorithme reconstruit l'image f^* en rétroprojetant les K lignes du sinogramme dans l'espace image.

$$f^*(x,y) = \sum_{k=0}^{K} S(k, r_k) \quad (1)$$

$$r_k = x * \cos(\frac{k\pi}{K}) - y * \sin(\frac{k\pi}{K}) + \text{offset} \quad (2)$$

Cette méthode n'est pas l'exacte transformée de Radon inverse. En effet, elle produit des artefacts en étoile et l'image reconstruite devient floue. Pour améliorer la reconstruction, on peut filtrer le sinogramme mais cette étape étant indépendante de la rétroprojection 2D, nous ne la développerons pas dans cet article.

2.2 Le "challenge"

Le sinogramme étant stocké en mémoire externe de type SDRAM, nous devons avoir une gestion efficace de la mémoire pour compenser la latence et permettre un haut degré de parallélisme. Cette stratégie diffère de celle de M. Leeser [8] qui utilise plusieurs bancs de mémoires SRAM indépendants avec une latence nulle pour améliorer le débit mémoire. Cette dernière constitue cependant une solution coûteuse et il est possible d'obtenir de meilleures performances avec des mémoires SDRAM à faibles coûts malgré une latence plus importantes. Les caches standards ayant pour principe un accès linéaire à la mémoire ne peuvent être une solution satisfaisante étant donné leur complexité technologique et leur faible taux de charge de données utiles. En effet, les accès mémoires nécessaires pour reconstruire un pixel image $f^*(x,y)$ suivent une sinusoïde dans le sinogramme ce qui constitue une faible localité spatialle pour les adresses mémoire. Pour accélérer ces accès mémoire, un nouveau mécanisme de cache est nécessaire. Une prédiction des points du sinogramme dont a besoin l'unité de calcul permettra au cache de charger les données pendant le processus de calcul.

3 Architecture

3.1 Architecture système

Une plateforme SoPC (System on Programmable Chip) a l'avantage de permettre une implémentation efficace et économique de la rétroprojection 2D. Pour davantage de flexibilité, et à coût réduit, le sinogramme et l'image reconstruite sont stockées en SDRAM, ou DDR-SDRAM, et la hiérarchie mémoire indispensable au recouvrement des calculs et accès mémoires, à grande latence, est construite à l'aide des blocs de mémoire embarquée dans le circuit SoPC.

L'architecture présentée figure 1 est évaluée sur une carte qui dispose d'une interface PCI avec une station hôte pour permettre l'échange de données avec le système. Cette carte dispose d'un circuit Virtex 2 Pro 2VP20, de 32 MO de SDRAM et 128 MO de DDR-SDRAM et d'un bridge PCI, réalisé sur un FPGA Spartan. Les données sont échangées entre le PC hôte et la carte grâce à un mécanisme de synchronisation.

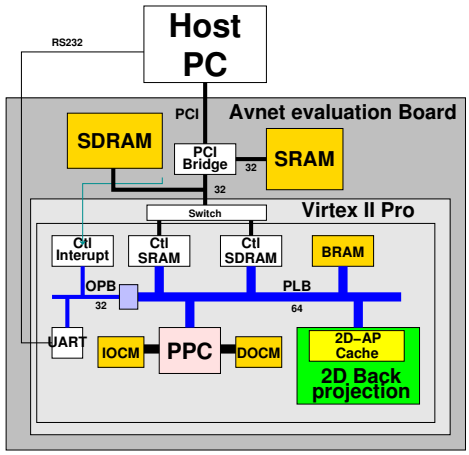

FIG. 1 – Architecture Système

L'architecture système est constituée :
- d'un processeur PPC en charge de la synchronisation des calculs et des communications
- d'une unité de rétro-projection avec son cache 2D-AP illustré en figure 2.
- de mémoire externe

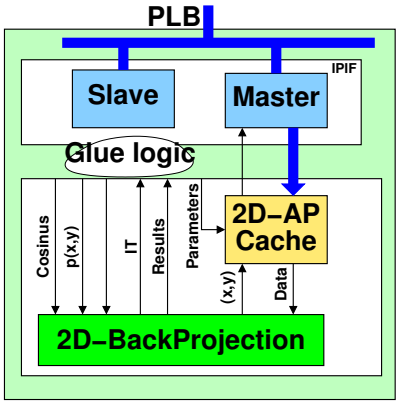

FIG. 2 – Unité de rétro-projection

Pour bénéficier au maximum de la cohérence spatiale 2D et temporelle des calculs, l'unité de rétroprojection reconstruit un bloc de forme quelconque de pixels voisins de f^*. Par souci de simplicité les blocs sont des carrés de pixels mais des hexagones pourraient être utilisés. Les données du sinogramme sont fournies à l'unité de rétroprojection par le cache 2D-AP qui se charge de leur transfert depuis la mémoire externe, le module étant maître sur le bus PLB.

Pour réduire les temps de calculs, l'unité de rétroprojection est parallélisée et une hiérarchie mémoire fournit les

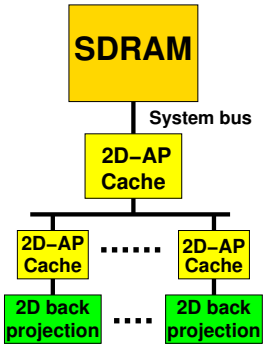

FIG. 3 – Architecture parallèle et Cache 2D-AP hierarchique

données aux modules : chaque module obtient ses données d'un cache 2D-AP qui lui même les met à jour à partir d'un cache 2D-AP de niveau supérieur. Dans notre exemple, nous nous arrêtons à un cache de niveau 2, comme illustré figure 3.

Cette architecture mémoire originale permet de n'utiliser qu'une seule mémoire externe qui contient tout le sinogramme, contrairement à la solution proposée par [8], tout en autorisant :

- le recouvrement de la latence de la mémoire externe et du bus système
- la réduction du débit à la mémoire externe

Les performances de cette hiérarchie mémoire sont améliorées lorsque la latence se réduit et une connexion directe de la mémoire à l'unité de rétro-projection permettrait de paralléliser massivement les unités de rétroprojection.

3.2 Le cache 2D-AP

3.2.1 Objectifs du cache 2D-AP

Le cache 2D-AP est l'élément original de l'architecture proposée en permettant aux unités de calculs d'accéder aux données sans temps mort. Ceci est rendu possible grâce à un mécanisme générique de prédiction des prochaines données utilisées. Le cache 2D-AP anticipe les accès au sinogramme 2D par une analyse de la séquence des coordonnées de pixels requis pour prédire les prochains accès à partir d'une hypothèse à vitesse de déplacement constante sur l'image. La reconstruction d'un bloc de pixels permet de produire une séquence d'accès au sinogramme qui tire partie au mieux des caractéristiques du cache 2D-AP.

La séquence produite par l'unité de rétroprojection pour reconstruire un bloc est l'union de toutes les sinusoïdes nécessaires à la reconstruction de chaque pixel. Dans le cas d'un carré de pixels, la séquence obtenue est une sorte de "tube" dont la figure 4 donne un exemple. Pour chaque angle θ, on accède à tous les points du sinogramme, sur une même colonne, nécessaires à la reconstruction de chaque pixel du bloc.

3.2.2 Fonctionnement

La prédiction de la zone en cache permet de réduire le taux de défaut de cache lorsque les accès pixels suivant se

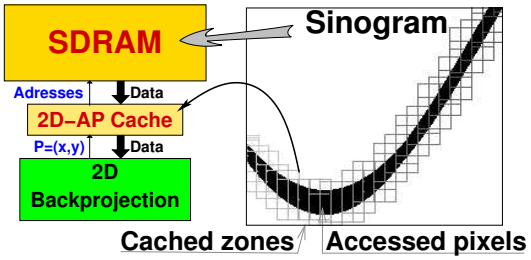

FIG. 4 – Concept du cache 2D-AP

déplacent sur l'image. Le centre et la taille de la zone sont déterminés par le calcul de la moyenne et du pseudo écart type (PSD=pseudo standard deviation) des accès pixels. La moyenne et le PSD sont calculés à l'aide de filtres passe-bas récursifs du premier ordre. En supposant une distribution uniforme des accès pixels autour de la moyenne, nous pouvons estimer que, pour une courte durée, ils sont dans un rectangle centré sur la moyenne et dont les demi-largeurs valent deux fois le PSD sur chaque axe. La zone ainsi calculée est mise à jour lorsque la moyenne varie.

Un mécanisme de prédiction permet d'anticiper la position du centre du cache lorsque les accès pixels suivent un chemin complexe dans l'image et réduit ainsi le coût de la latence des accès mémoire. La mise à jour ne se produit que lorsque la moyenne a suffisament variée et la nouvelle zone est centrée sur une anticipation de la moyenne, supposée se déplacer à vitesse constante. Ainsi, une zone de garde est définie autour du centre de la zone en cache et la mise à jour se fait lorsque la moyenne calculée sort de cette zone de garde, dans la direction de sortie, comme illustrée figure 5.

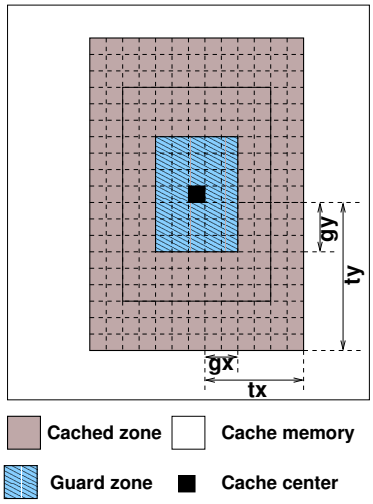

FIG. 5 – Zones du cache 2D-AP

Pour réduire le coût matériel du contrôle et obtenir de bonnes performances les filtres passes bas pour le calcul des moyennes sont réalisés à l'aide de simples additions et décalages. Ce sont des filtres RII du premier ordre dont l'équation est $s_n = ce_n + (1-c)s_{n-1}$, e signal d'entrée et s signal filtré, et nous notons $s = \mathrm{f}^c(e)$. Le paramètre $1-c$ correspond à la constante de temps du filtre RC équivalent. Lorsque les suites s et e sont représentées en virgule fixe et c une puissance de $\frac{1}{2}$, nous ne réalisons plus que des additions et décalages.

4 Résultats

Les mesures effectuées sur la plateforme et les résultats de simulation nous montrent que la hiérarchie mémoire choisie est efficace et nous permet des accélérations importantes par la parallélisation des opérateurs et ceci en n'ayant qu'une mémoire externe qui contient toutes les données. Les métriques obtenues sont données dans le tableau 4 et montrent une accélération quasi-linéaire avec le nombre d'opérateurs. Elles correspondent à la reconstruction d'une image $x_{max} * y_{max} = 320 * 320$ à partir d'un sinogramme de résolution angulaire $K = 512$.

Système	**Cycles**	**Temps**
Logiciel		
Pentium 3 (1 GHz)		3,5 s
PPC (VirtexII-Pro)		94 s
Matériel simple		
Idéal	52.10^6 $(320 * 320 * 512)$	1,1 s
Sans cache	52.10^6*28 (Idéal*Latence)	30 s
1 unité	78.10^6	1,5 s
Matériel parallélisé		
2 unités	42.10^6	0,8 s
4 unités	21.10^6	0,42 s
9 unités	11.10^6 (simulé)	0,22 s

Tab. 1 – Accélération par le module de rétroprojection@50 Mhz

Idéalement, nous devrions pouvoir réaliser une reconstruction sans temps mort, c'est à dire en $x_{max} * y_{max} * K = 320 * 320 * 512$ cycles d'horloge. Ces performances idéales sont altérées du fait d'une part de la synchronisation entre l'opérateur matériel et le PPC et d'autre part des performances du Cache 2D-AP sur le bus système. Les mesures effectuées nous montrent que les simulations réalisées pour la reconstruction d'un bloc sont fiables et peuvent être extrapolées à la reconstruction d'une image complète.

L'ensemble du module a été décrit en VHDL générique paramétrable pour explorer rapidement les différentes configurations possibles. Le cache est entièrement modulaire de façon a pouvoir mesurer l'efficacité de différents type d'estimateurs prédictifs et pour pouvoir construire une hiérarchie par simple assemblage de blocs. Le tableau 4 donne la complexité du système en nombre de LUT pour une synthèse sur cible Xilinx VirtexII-Pro.

Module	CLB	FG
1 unité	1078	2155
2 unités	2253	4506
4 unités	3877	7753

Tab. 2 – Synthèse de l'IP et son cache 2D-AP.

5 Conclusion

Nous avons présenté dans cet article un système de reconstruction par rétroprojection 2D implémenté sur une plateforme SoPC. Ce système a été conçu de manière à réduire les erreurs de reconstruction dues aux calculs en virgule fixe. Le cache 2D-AP permet de réduire le goulot d'étranglement existant lors des accès en mémoire. Lors de la reconstruction, les données nécessaires sont prédits statistiquement afin que le cache puisse les charger avant que l'unité de rétroprojection les utilise. De plus, l'algorithme de rétroprojection a été implémenté de façon à augmenter la localité spatiale et temporelle, l'utilisation du cache en devient plus pertinante.

La rétroprojection est utilisée par les algorithmes itératifs plus sophistiqués. Ces derniers offrent une meilleure qualité d'image mais en contrepartie ont un temps reconstruction beaucoup plus long. La réalisation présentée dans cet article est une première étape pour construire un système de reconstruction itératif, adaptable et flexible. Implémenté sur un SOPC, le module de rétroprojection 2D peut être dupliqué pour paralleliser les calculs. Un même module de cache peut être partagé ou une hiérarchie de caches mémoire peut être mise en place. Ainsi en s'appuyant sur la localité spatiale, nous pouvons reconstruire simultanément plusieurs pixels et la vitesse de reconstruction est alors notablement améliorée.

Références

[1] M.S. Musrock et al. Performance characteristics of a new generation of processing circuits for pet applications. *IEEE Tr. Nucl. Sci.*, 50(4) :974 – 978, August 2003.

[2] D.W. Shattuck, J. Rapela, E. Asma, A.Chatziioannu, J.Qi, and R.M. Leahy. Internet2-based 3D PET image reconstruction using a PC cluster. *Phys. Med. Biol.*, 47(15) :2785–2795, August 2002.

[3] S. Vollmar, C. Michel, J.T. Treffert, D.F. Newport, M. Casey, C. Knoss, K.Wienhard, X. Liu, M.Defrise, and W.-D. Heiss. Heinzelcluster accelerated reconstruction for FORE and OSEM3D. *Phys. Med. Biol.*, 47(15) :2651–2658, August 2002.

[4] Nikolay Sorokin. *An FPGA-Based 3D Backprojector.* PhD thesis, Universität des Saarlandes, Allemagne, 2003.

[5] J. Müller, D. Fimmel, R. Merker, and R. Schaffer. Hardware- software system for tomographic reconstruction. *J. Circuits Syst. Comp.*, 12(2) :203–229, April 2003.

[6] F. Xu and K. Mueller. Ultra fast 3D filtered back projection on commodity graphics hardware. In *Proc. IEEE Int. Symp. Biomedical Imaging (ISBI'04)*, pages 571–574, Arlington, USA, April 2004.

[7] S. Mancini, N. Gac, and M. Desvignes. Etude d'un cache 2D adaptatif et prédictif pour le traitement d'image. In *Journées Francophones sur l'Adéquation Algorithme Architecture (jFAAA'05)*, Dijon, France, 18-21 Janvier 2005.

[8] M. Leeser, S. Coric, E.Miller, H. Yu, and M.Trepanier. Parallel-beam backprojection an FPGA implementation optimized for medical imaging. *J. VLSI Signal Proc.*, 39(3) :295–311, March 2005.