Algorithm Architecture Adequacy for High Speed 3D Tomography

Nicolas Gac, Stéphane Mancini, Michel Desvignes, Dominique Houzet

Algorithm Architecture Adequacy for High Speed 3D Tomography

Nicolas GAC, Stéphane MANCINI, Michel DESVIGNES and Dominique HOUZET

***Abstract*— Backprojection is a computational costly step in tomography image reconstruction such as Positron Emission Tomography (PET). In this purpose, this paper presents a Pipelined, Prefetch and Parallelized Architecture for PET backprojection (3PA-PET). The main strength of this architecture comes from its original memory access strategy, masking the high memory latency of the external memory. The 3PA-PET architecture is implemented on a System on Programmable Chip (SopC). Time performances are compared with a desktop PC, a workstation and a GPU. We prove that the exploitation of the intrinsic temporal and spatial locality by the 3D Predictive and Adaptative (3D-AP) memory cache succeeds to run efficiently several pipelines of backprojection : each reaches a computational throughput close to 1 operation per cycle.**

I. Introduction

Reconstruction of images in tomography is cpu intensive and usually postponed. However, real-time reconstruction of the acquired data in PET imaging would facilitate positioning of the subject and detection of potential problems during the acquisition. Real time reconstruction is also needed for large scale diffusion of clinical PET examinations, which are used for early detection of cancer, evaluation of disease spread and treatment response. Then, minimization of examination duration can decrease the cost of PET to make its powerful technology more widely available. However Image reconstructions in the field of tomography, including SPECT, CT, multislide CT need to process a costly backprojection step.

There are several implementations of reconstruction systems on PC clusters [1][2], on DSPs [3], on adapted CPUs for vector processing like Cray or Cell [4][5], on specific hardware like ASICs or FPGAs [6][7][8][9] or on Graphics Processor Units (GPUs) [10]. All these implementations have to face the memory bottleneck due to the limited bandwidth of the main memory. Without an adapted memory strategy, one rule applies to all these architectures : the greater the computing power is, the more the reconstruction time is penalized by the memory wall. Works previously mentioned are based on Algorithm Architecture Adequacy methods to overcome this issue. For instance, the implementation of 3D cone-beam backprojection done by M.Kachelriess [4] has coupled high computing speed on a cell processor (code parallelization, incremental algorithm, loop unrolling, interpolation pre-processing) with software memory prefetching techniques.

In this paper, we present a Pipelined, Prefetch and Parallelized Architecture for 3D PET backprojection (3PA-PET) implemented on a System on Programmable Chip. This architecture has a computation throughput of about one cycle per operation for each pipeline. It overcomes the memory access bottleneck thanks to a prefetching memory strategy. Results in accuracy of reconstruction and in time performance are evaluated and compared to a desktop PC, a workstation and a GPU.

Nicolas GAC : GIPSA-lab, INPG-CNRS, 38031 Grenoble, France, nicolas.gac@lis.inpg.fr
Dr. Stéphane MANCINI : GIPSA-lab, stephane.mancini@lis.inpg.fr
Prof. Michel DESVIGNES : GIPSA-lab, michel.desvignes@lis.inpg.fr
Prof. Dominique HOUZET : GIPSA-lab, dominique.houzet@lis.inpg.fr

II. Algorithm Architecture Adequacy

A. 3D PET Backprojection

Data acquired by PET scanners is the Radon Transform of the observed body and is called sinogram [11]. One sinogram element called a bin represents the number of coincidence events counted on two detectors of the scanner. The line what connects two detectors is called a line of response (LOR). The sinogram is a 4D space along $(\Delta, \lambda, \psi, \rho)$. Coordinates (Δ, λ) represent a couple of rings : Δ the distance between the two rings (segment number) and λ the mean axial coordinate of the two rings (plane number). Then the coordinates (ψ, ρ) represent one particular LOR : ψ is the azimuthal angle (angle number) and ρ is the tangential coordinate (bin number). Backprojection computes the estimated distribution of radio-tracer f^* for each voxel $\vec{\mathtt{x}}$ by summing up all the bins corresponding to the voxel projection :

$$f^*(\vec{x}) = \int\int \mathtt{bin}(\Delta, \lambda, \psi, \rho) \cdot J_\Delta \cdot d\psi \cdot d\Delta \quad (1)$$

$$\begin{cases} \rho(\psi) = x \cdot \cos\psi + y \cdot \sin\psi \\ \lambda(\psi, \Delta) = z - (-x \cdot \sin\psi + y \cdot \cos\psi) \cdot \dfrac{\Delta}{2R_a} \end{cases} \quad (2)$$

B. Memory access strategy

As the sinogram is kept in a SDRAM like external memory, we need an efficient memory management to overcome its latency and allow a high level of parallelism. Standard caches based on a linear accesses to memory can't be a satisfactory solution because of their complexity and their weakness to load the needed data. Indeed, memory accesses needed to reconstruct one single voxel $f(\vec{x})$, follow a 3D sinusoid in the 4D sinogram as shown on figure 1. Such a pattern is of poor address locality. Moreover because of the additional lambda dimension, 3D backprojection accesses are greater and more distributed in the memory space than in the 2D backprojection case [12]. The challenge is to speed-up these 4D memory accesses.

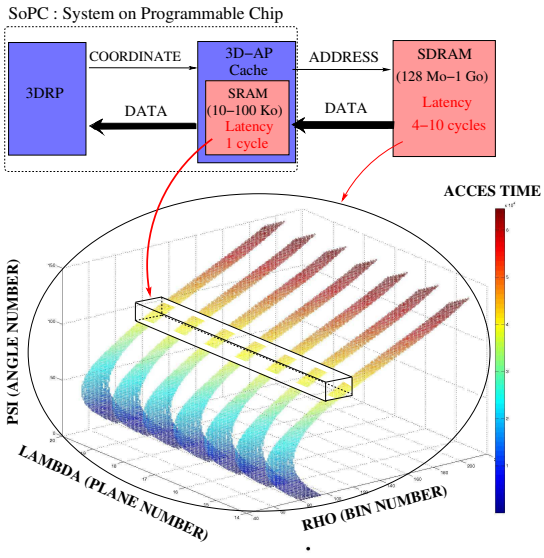


Fig. 1. The memory access strategy is based on a fast and small cache memory inside the SoPC. The cache predicts the needs of the 3D Back-Projection (3D-BP) unit and therefore succeeds to mask the 4-10 cycles latency of the slower and bigger external SDRAM memory.

Therefore a new cache mechanism is needed. Predicting which bins the processing unit will use, would help the cache to download the needed bins during the computing process. Thus, we could balance the computational throughput with the access memory throughput.

C. Improvement of spatial and temporal locality

A volume reconstruction by the Voxel-driven Bilinear Interpolation (VBI) standard backprojection is made of three loops : the first on the voxels $\vec{\mathtt{x}}$, the second on the segments `Delta` and the third on the angles `psi`. Since voxels can be reconstructed independently, the loop on voxels can be split by two, one on blocs of voxels (0 .. $\mathtt{n}_{\mathtt{max}}$) and one on the voxels of a bloc `n` ($\vec{\mathtt{x}}_{\mathtt{min}}(\mathtt{n})$ .. $\vec{\mathtt{x}}_{\mathtt{max}}(\mathtt{n})$).

The reordering of loops increases the temporal and spatial locality of memory accesses. Indeed, for given `psi` and `Delta` values, a $\mathtt{bin}(\mathtt{psi}, \vec{\mathtt{x}})$ will be used several times since the projection

Algorithm 1 Reordering loops of 3D Backprojection improving spatial and temporal locality

for $\mathtt{n} = 0$ to $\mathtt{n}_{max}$ **do**
  for $\mathtt{Delta} = 0$ to $\mathtt{Delta}_{max}$ **do**
    for $\mathtt{psi} = 0$ to $\mathtt{psi}_{max}$ **do**
      for $\vec{\mathtt{x}}_{min}(\mathtt{n})$ to $\vec{\mathtt{x}}_{max}(\mathtt{n})$ **do**
        $f(\vec{\mathtt{x}}) = f(\vec{\mathtt{x}}) +$
        $\mathtt{bin}\big(\mathtt{RHO}(\mathtt{psi}, \vec{\mathtt{x}}), \mathtt{PSI}, \mathtt{LAMBDA}(\mathtt{psi}, \vec{\mathtt{x}})\big)$
      end for
    end for
  end for
end for

of a 3D Bloc of voxels is a 2D plane in the 4D space of the sinogram.

D. Mean Bin Reuse Rate (MBRR)

The Mean Bin Reuse Rate (MBRR) is defined as the ratio between the number of bins accessed in cache memory and the number of bins loaded in cache memory. The MBRR can be computed analytically. It depends on the shape of the reconstructed bloc of voxels.

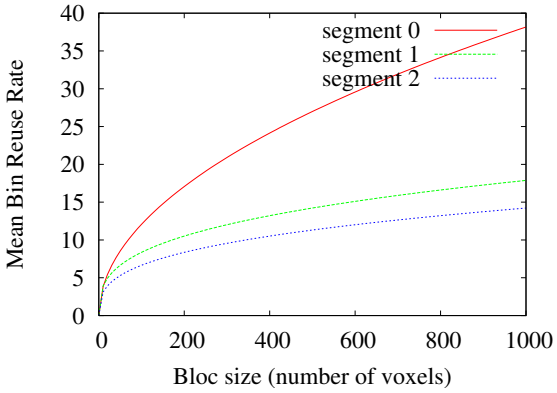


Fig. 2. Mean Bin Reuse Rate (MBRR) estimated versus the size of reconstructed blocs of voxels for each segment of a Siemens HR+ sinogram (span 9 with 96 angles of projection)

Figure 2 presents the optimal MBRR computed for each segment versus the size of the bloc. For a 16*16*3 bloc, we can expect a MBRR of 32 for segment 0, 15 for segment 1 and 13 for segment 2.

III. 3P Architecture for PET

A. Pipelined Architecture

The pipeline implements the different steps of the VBI standard backprojection : the computation of $\mathtt{rho}(\mathtt{psi}, \vec{\mathtt{x}})$ and $\mathtt{lambda}(\mathtt{psi}, \vec{\mathtt{x}})$, the bilinear interpolation of the bin, and finally the accumulation of the voxel value. The forward flow control is done by packets passing through each stage of the pipeline.

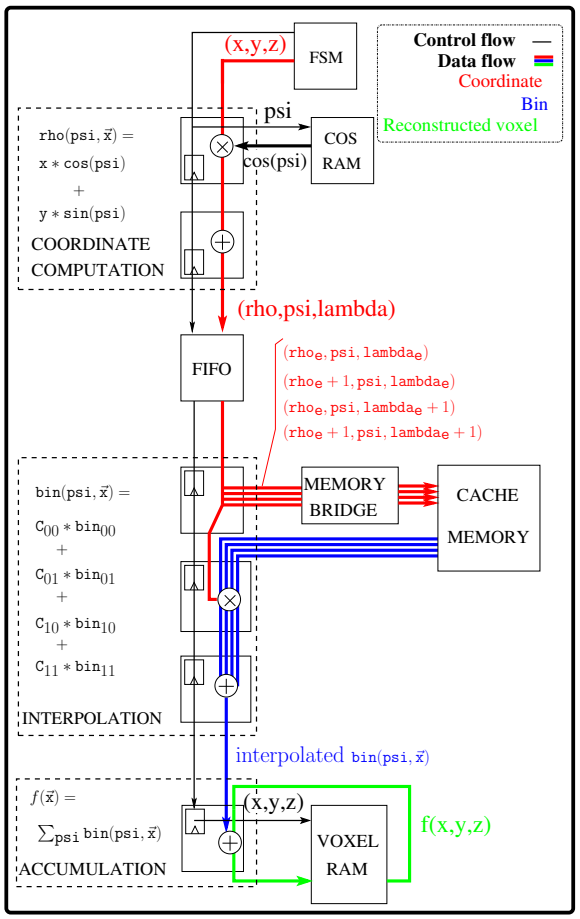


Fig. 3. Pipeline of the 3PA-PET Architecture

The accesses to the 4 bins needed for the bilinear interpolation are done through the bridge memory. This bridge controls the cache memory and freezes or not the pipeline whether the data

requested is available or not. A backward flow control synchronizes the pipeline and the cache memory.

B. Prefetch Architecture

The 3D-AP cache [13] masks the latency of the external memory. Thanks to it, the pipeline is no more systematically stalled. The bridge memory gets four bins from the cache at each clock cycle.

The 3D-AP cache is a generic cache memory mechanism that prefetches memory access sequences following a continuous path into a 3D memory space. The requests of data from the pipeline are done with spatial coordinates, here (`rho`, `psi`, `lambda`). From the previous coordinate requests, the cache estimates dynamically which data is likely to be requested in the future. This is done by a statistical analysis on each spatial dimension. Moreover, cache data transfers are masked. Indeed, while the new data grabbed by a cache update is transfered from external memory, the data shared by the old and the new cache zone stay available in the cache.

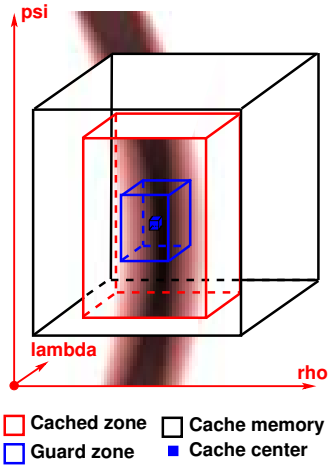


Fig. 4. 3A-AP Cache zones

Cache parameters have to be set beforehand by the user to fit to the 3D memory path as close as possible. In this study, we set for each dimension the value of five parameters :

- *cut-off frequency and sampling frequency* : the mean coordinate is computed by a first order low-pass IIR filter configured by these two frequencies.
- *cached zone size* : this zone is declared to the bridge memory to be available in cache. In this study, this size is a static parameter.
- *guard zone size* : each time the mean coordinate is out of the guard zone, the cache zone is updated.
- *cache speed* : The cache speed has to be set according to the speed of the data accesses performed by the application on each spatial dimension.

C. Parallelized Architecture

To increase the computing power, several pipelines are parallelized. A hierarchical cache reduces as much as possible the memory bus occupation when backprojection units work in parallel. In this hierarchical design, one leaf cache is associated to one 3D backprojection unit while a root cache is feeding each of these leaf caches.

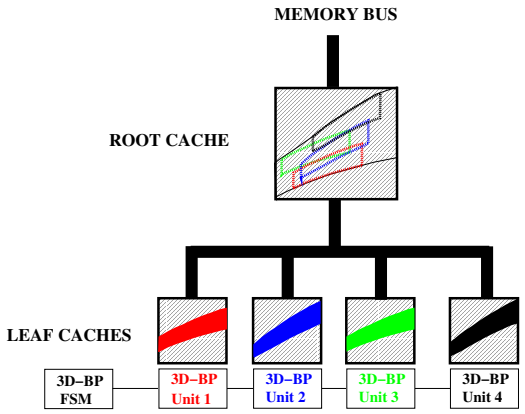


Fig. 5. Each leaf cache is feeding by the root cache. All backprojection units are synchronized by one single finite state machine (FSM).

The reconstruction of a macro-bloc of neighbor blocs is distributed over the backprojection units. The cache concept presented previously with one unit, applies here in the same manner.

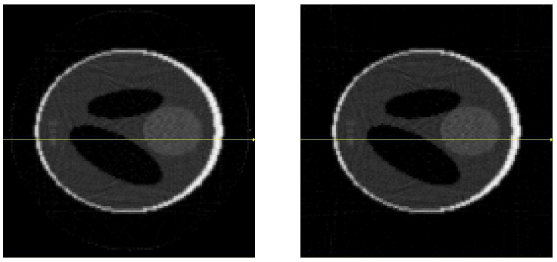
The bin needed during a macro-bloc reconstruction draws a 3D sinusoid. Each leaf cache stores a smaller 3D sinusoid needed for its bloc reconstruction as presented on figure 5.

IV. Results

A. Accuracy of reconstruction

The implemented VBI standard backprojection is a fixed point version of the original algorithm. Moreover the sinogram data is converted from float to short int (16 bits). Accuracy of reconstruction of 3PA-PET is compared to software reconstruction thanks to a bit true software implementation of the FPGA-based architecture.

The reference data set used is a sinogram analytically computed from a 3D Shepp Logan volume of 128*128*63 voxels. This phantom is a standard volume used in tomography to measure the accuracy of reconstruction. The sinogram is obtained from the STIR open source tool kit [14].



STIR 3PA-PET

Fig. 6. A slice of the 3D Shepp Logan phantom (128*128*63 voxels) reconstructed by STIR and 3PA-PET backprojection.

On figures 6 and 7, one can compare the Shepp Logan volumes reconstructed by the STIR voxel-driven backprojection and by the 3PA-PET backprojection.

The accuracy of reconstruction of the 3PA-PET backprojection is measured with two metrics : the mean absolute percentage error (MAPE) and the peak signal-to-noise ratio (PSNR). Both compare a volume f_1 with a volume of reference f_{ref} corresponding to the original volume or to a volume reconstructed

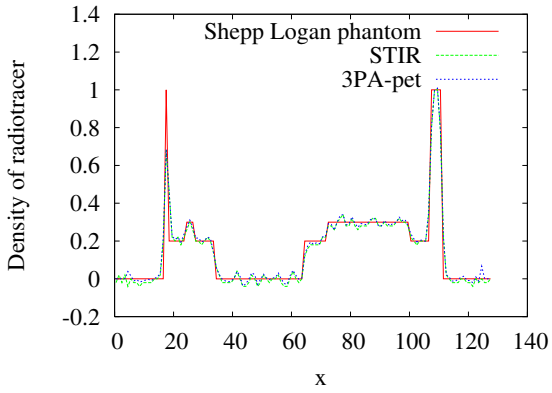


Fig. 7. Profile of the 3D Shepp Logan phantom slices corresponding to the lines on figure 6

with a reference 3D-BP implementation. The PSNR corresponds to the ratio between the maximum of f_{ref} (dynamic range) and the mean squared error (MSE) of f_1 compared to f_{ref} :

$$PSNR = 20 \cdot log_{10} \frac{max\big(f_{ref}(\vec{x})\big)}{\sqrt{MSE}} \qquad (3)$$

On table I, volumes reconstructed by STIR, by VBI with floating point arithmetic (VBI-flt) or VBI with fixed point arithmetic (VBI-fix) are compared to the original phantom and to each other.

compared volumes			data	MAPE	PSNR
Accuracy of reconstruction					
STIR	/	original	float	3.89 %	10.5 dB
VBI-Flt	/	original	float	3.88 %	10.5 dB
VBI-Fix	/	original	float	3.88 %	10.5 dB
VBI-Flt	/	original	int16	3.97 %	10.5 dB
VBI-Fix	/	original	int16	3.97 %	10.5 dB
Compared reconstructions					
STIR	/	VBI-Flt	float	0.35 %	21.5 dB
VBI-Fix	/	VBI-Flt	float	0.13 %	26.2 dB
VBI-Fix	/	VBI-Flt	int16	0.13 %	23.0 dB
VBI-Fix	/	VBI-Flt	int16/flt	1.1 %	19.0 dB

TABLE I
Accuracy of reconstruction and Compared Reconstructions for the Shepp Logan phantom

All implementations have an error of reconstruction around 3.9% with a PSNR of 10.5 dB

compared with the original volume. This error is intrinsic to the method. The difference between floating point and fixed point implementation corresponds to a MAPE of 0.13% and to a PSNR of 23 dB. With different data type, the difference has a MAPE of 1.1% and a PSNR of 19 dB. Thus we can conclude that the 3PA-PET implementation of VBI backprojection leads to an accurate reconstruction system.

B. 3PA-PET complexity

Hardware resources used by the 3PA-PET architecture are presented on table II. The main FSM backprojection and the root cache control are shared between the n units of the 3PA-PET architecture, therefore the cost of an additional unit is only 800 slices. Sizes of leaf and root caches are respectively 2 kB and 18 kB. Hence, 9 backprojection units fit in a Xilinx Virtex 2 Pro VP30 chip.

	1 unit	**4 units**	**9 units**
3D Backprojection			
CLB slices	573 (4.2%)	1817 (13.3%)	3924 (28.6%)
Multipliers	12 (9%)	48 (35%)	108 (79%)
3D-AP Cache			
CLB slices	672 (4.9%)	2830 (20.6%)	4804 (35.1%)
RAMs	2 kB (0.6%)	24 kB (7.8%)	36 kB (11.7%)
3D Backprojection + 3D-AP Cache			
CLB slices	1245 (9.1%)	4637 (32.9%)	8728 (63.7%)

TABLE II

HARDWARE RESOURCES USED BY THE 3PA-PET ARCHITECTURE IN PARENTHESIS THE PERCENTAGE OF OCCUPATION OF THE XILINX VIRTEX 2 PRO VP30 RESOURCES

C. Time performance

The 3PA-PET time performances are compared with STIR and our software VBI backprojection on a desktop PC (Pentium 4 Prescott), a workstation (bi-Xeon dual core) and a GPU (Nvidia GTS88800). The efficiency of the architecture is evaluated by the clock cycles per operation where an operation corresponds to a voxel update. The number of voxel updates is equal to the number of voxels multiplied by the number of segments times the number of angles.

The 3PA-PET cycle efficiency is measured on an Avnet development board. A simulated memory bus is implemented to observe how the 3PA-PET architecture reacts with respect to the memory bandwidth and latency. The results presented in figure 8 are achieved with one backprojecting unit for segments 0 and +2. As we can expect, segment 0 is backprojected more efficiently than segment +2, as the path into the 3D memory space is bounded on lambda dimension. The cache miss rate stays low (about 0.05% with a memory latency of 5 cycles and a memory delay of 1 cycle). This robustness to high latency and low bandwidth is due to the high spatial and temporal locality of the application as presented on section 2. It explains why the architecture can potentially be highly parallelized.

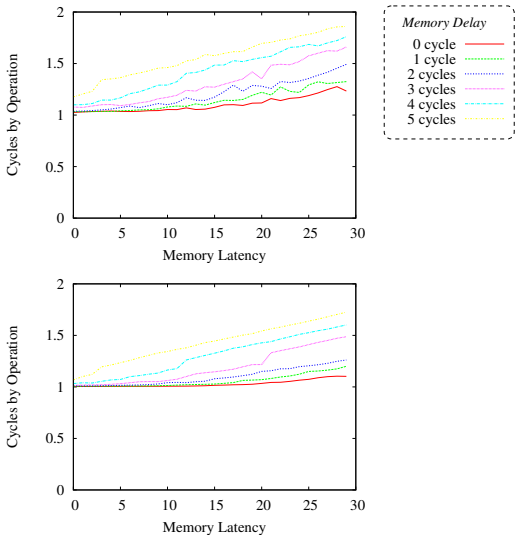


Fig. 8. Performance of 3PA-PET with one unit for backprojection of segment +2 (on the top) and segment 0 (at the bottom) versus the latency and bandwidth of the memory. Bandwidth is represented by the delay between the delivery of two memory words on the memory bus.

On table III, time performance in seconds

is normalized to the reconstruction of a $128 * 128 * 63$ volume. Indeed, STIR reconstructs a cylindrical FOV (Field of View) of $64^2\pi * 63$ voxels. Moreover, to fairly compare our FPGA-based architecture with others technologies, the time measured on a Virtex 2 Pro has been scaled to a Virtex 4. Indeed, the Virtex 4 technology is the same generation as the CPU and GPU used in this study.

The software code of the VBI backprojection is carefully optimized. For instance, time performance is improved by a factor two thanks to an incremental computation of coordinates as done by Kachelriess [4] for 3D cone beam backprojection. Thus the implemented backprojection algorithm is competitive with the one used by STIR. Both have a reconstruction times about 10 s. Then, this code is parallelized to use the four cores of a bi-Xeon dual core workstation. One thread is associated to the reconstruction of one bloc. One strength of the Pentium 4 and Xeon cores is their wide L2 Cache memory of 2 MBytes.

The Nvidia GPU has 12 vector processors, each one having 8 stream processors. A non incremental code is parallelized to run efficiently on these $12 * 8$ stream processors. One thread is associated to one voxel reconstruction. Interpolation is hard-wired and each vector processor is associated with a L1 cache memory.

3D-BP Algorithm	**PE (threads)**	**Time**	**Cycles/Op /PE**	**Cycles/Op total**
Desktop PC : Pentium 4 (3.2 Ghz)				
STIR	1	11.13 s	70.4	70.4
VBI-float	1	9.7 s	63	63
Workstation : bi-Xeon dual core (3 Ghz)				
STIR	1 (1)	5.74 s	34.5	34.5
VBI-float	1 (1)	5.9 s	36	36
VBI-float	2 (2)	3.1 s	38	19
VBI-float	4 (4)	2 s	48	12
VBI-float	4 (8)	1.67 s	40	10
GPU : GTS8800 (1.35 Ghz)				
VBI-float(*)	96 (192)	0.12 s	43.2	0.45
FPGA : Virtex 2 Pro (35 Mhz)				
VBI-fix	1	14.86 s	1.05	1.05
VBI-fix	4	6.37 s	1.8	0.45
VBI-fix	8 (**)	3.89 s	2.2	0.275
FPGA : Virtex 4 (200 Mhz)				
VBI-fix	1(***)	2.6 s	1.05	1.05
VBI-fix	4 (***)	1.11 s	1.8	0.45
VBI-fix	8 (***)	0.68 s	2.2	0.275

(*) non optimized code
(**) simulation
(***) 35 Mhz results scaled to 200 Mhz

TABLE III

Compared Time Performance for the 3D PET backprojection of a 128*128*63 volume from a Siemens HR+ sinogram (5 segments, span 9, 96 angles of projection). FPGA measures are done with a simulated bus (memory latency of 5 cycles and memory bandwidth of 1 cycle/memory word).

D. Discussion

On one hand, GPU seems to be the best adapted solution to speed up 3D backprojection with a final reconstruction time of 0.12 s and a computational throughput of 0.45 C/Op. This good performance is mainly due to its high degree of parallelization. Parallelization works also well on the bi-Xeon processor. From 36 cycles per operation with one core, the Xeon reaches an efficiency of 10 cycles per operations with 4 cores. As a consequence, the workstation is six times faster than the desktop PC. Our FPGA-based architecture on a Virtex 4 is 5 times slower than the GPU but it remains as in our 2D study [12] a faster solution than classical CPUs : 4.5 times faster than a mono-Xeon dual core and 15 times faster than a Pentium 4.

On the other hand, 3PA-PET is the best cycle efficient architecture per processing element (PE), about twenty times more efficient than the Xeon and the GPU. Indeed, 3PA-PET reaches a computational throughput close to one cycle per PE : 1.05 with one unit, 1.8 with 4 units and 2.2 with 8 units. Besides, as the computational resources are efficiently used, its lower power consumption could become an advantageous solution in another context than medical reconstruction.

V. Conclusion & Future work

The pipelined and parallelized architecture has efficiently sped up 3D PET backprojec-

tion without a significant loss of accuracy. The pipelines are seldomly stalled because the 3D Predictive and Adaptative cache overcomes the limitations due to the latency and bandwidth of the external memory. The encouraging results presented in this paper let us hope to better exploit the memory bandwidth and therefore to run efficiently more than 8 units of backprojection in parallel. Afterwards, we could reach a second level of parallelism by increasing the processing resources and the memory bandwidth. This will be done thanks to a self made board with 7 Virtex 4 (6 processing units and one controller unit), each associated with its own external memory.

VI. Acknowledgement

We would like to thank Anthonin Reilhac from the CERMEP laboratory at Lyon for his help on PET data.

References

[1] G. Kontaxakis et al. Iterative image reconstruction for clinical PET using ordered subsets median root prior and web-based interface. *Mol. Imaging Biol.*, 4(3):219–233, May 2002.

[2] 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.

[3] K.Rajan and L.M. Patnaik. Cbp and art image reconstruction algorithms on media and dsp processors. *Microprocessors and Microsystems*, 25(5):233–238, 2001.

[4] M. Kachelrieß, M. Knaup, and O. Bockenbach. Hyperfast backprojection using the cell broadband engine. In *IEEE Medical Imaging Conference Record*, 2006.

[5] M. Sakamoto, H.Nishiyama, H.Satoh, S.Shimizu, T. Sanuki, K.Kamijoh, A. Watanabe, and A. Asahara. An implementation of the feldkamp algorithm for medical imaging on cell. Technical report, IBM, 2005.

[6] I. Goddard and M. Trepanier. High-speed cone-beam reconstruction : An embedded systems approach. In *Proc. SPIE Medical Imaging Conf.*, pages 483–491, February 2002.

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

[8] J. Li, C. Papachristou, and R. Shekhar. A reconfigurable Soc architecture and caching scheme for 3D medical image processing. In *Proc. IEEE Symp. Field-Programmable Custom Computing Machines (FCCM'04)*, Napa, USA, April 2004.

[9] 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.

[10] K. Mueller, F; Xu, and N. Neophytou. Why do commodity graphics hardware boards (gpus) work so well for acceleration of computed tomography ? In *Proc. SPIE Vol. 6498 [Computational Imaging V]*, 2007.

[11] R. Clackdoyle P. Kinahan, M. Defrise. Analytic image reconstruction methods in emission tomography. In J. Aarsvold M. Wernick, editor, *Emission Tomography : The Fundamentals of PET and SPECT*. Elsevier Academic Press, 2004.

[12] N. Gac, S. Mancini, and M. Desvignes. Hardware/software 2d-3d backprojection on a sopc platform. In *Proceedings of the 2006 ACM Symposium on Applied Computing (SAC), Dijon, France*, pages 222–228, April 2006.

[13] S. Mancini and N. Eveno. An iir based 2d adaptive and predictive cache for image processing. In *Design of Circuits and Integrated Systems*, Bordeaux, France, 2004.

[14] C. Labbé et al. An object-oriented library for 3d pet reconstruction using parallel computing. In *Proceedings of "Bildverarbeitung fuer die Medizin 1999, Algorithmen-Systeme-Anwendungen", Informatik aktuell, Springer*, pages 268–272.