An Improvement on the Processing Speed of Computed Tomography with GPU

Jiun-Ching Chen, Ta-Hua Lai, Shih-Chieh Lin

Dep. of Power Mechanical Engineering, National Tsing Hua University, Taiwan

johnlaidejp@yahoo.co.jp

Abstract – It is apparent that taking fewer images for tomography computing can reduce the absorption of radiation. It was noted that the Simultaneous Algebraic Reconstruction Technique (SART) can results in much better reconstructed image quality than the Filtered Back-Projection (FBP) method can when limited images were taken.

In this study, the SART approach was adopted to insure a better image quality when limited images were taken. Furthermore, for the efficiently computing purpose, we crystallize our research goal that aimed at an in-depth investigation of several related domestic in the scope of CT image reconstruction with CUDA GPU parallel computing technology, both in theory and experiment to reduce processing time.

Keywords - Computed Tomography, Simultaneous Algebraic Reconstruction, GPU coprocessor, Parallel Computing.

I. INTRODUCTION

Automatic optical inspection (AOI), which combines interdisciplinary technology, such as optics, electronics, mechanics, and computed image process, can be applied to assess the process quality in a production line. Merits of adopting AOI in product inspection are not only improving the efficiency of quality inspection but also reducing the inspection cost. Nevertheless, traditional optical inspection is limited to detect defects appeared on the surface of the examined objects. In order to detect defects below the object surface, the X-ray Computed Tomography (CT) technique [1] can be adopted to obtain information about the size and shape of the defect at any desired section.

X-ray CT had been a popular tool to identify the 3D detailed structure of examined objects for many years. In general, more than hundreds X-ray images were taken, and then three dimension images were computed based on these images by using Algebraic Reconstruction Technique (ART) or Filtered Back Projection (FBP) algorithm.

It is apparent that taking fewer images for tomography computing can reduce the absorption of radiation. At the study of using FDK reconstruction technique on BGA inspection, Liang [2] described that the FBP method has good computational efficiency for CT but also accompanying the inferior quality of reconstructed images. So as to improvement the performance of ART, Xu [3] introduced a simultaneous ART (SART) methodology that inherited some advantages from ART and SIRT both. For all practical purposes, in the further experiment [4], SART certainly had a better image quality then traditional ART when limited images were taken.

Generally, although SART has a good reconstructed quality but it requires many renewing steps of weights, which included many trivially computational iterations. This computational inefficiency restricts their use in many applications, especially when real-time performance is crucial.

For the efficiently computing purpose, an adopted SART method was introduced in this study to insure a better image quality when limited images were taken. In Practical, we crystallize our research goal that aimed at an in-depth investigation of several related domestic in the scope of CT image reconstruction with CUDA GPU parallel computing technology, both in theory and experiment. Experiments are conducted to demonstrate the validity of the proposed method.

In this paper, the fundamental concept of the SART technique is briefly described. A new procedure for weight estimation is then proposed to ease the handling of weight matrix in computing process. A series of simulation study are then conducted to study the improvement of using the new procedure for weight estimation in processing time and to study the performance of CUDA GPU parallel computing coprocessor to that of Intel CPU. The tomography image quality is also compared. Finally, conclusions were made based on these simulation results.

II. SIMULTANEOUS ALGEBRAIC RECONSTRUCTION

SART is a type of iterative reconstruction methods. Techniques for using SART to reconstruction the original 3D object are based on multiple 2D projected images, which obtained from the detector array, with little algebraic tricks. The first step of SART is to discretelize a section of the analyte where we are interested, as a regular lattice of image. The regular lattice can be seen as a matrix that consists of a set of graphical units. The 2D or 3D object to be reconstructed here can be represented by the finite number of the regular lattice, which corresponding to the pixels or voxels respectively. Second, the main linear algebraic function of SART can be derived based on the regular lattice and the projected data obtained by the detector array together. The major purpose of SART is to find the solution of the linear algebraic function to accomplish the job of image reconstruction.

Fundamental operating mechanisms of SART had interpreted by Roh and Cho [5]. The physical phenomenon of X-ray decay is simply modeled as (1) according to the intersection length and the material absorption rate for X-ray photons.

$$I = I_0 \exp\left(-\sum_{i=1}^{K} a_i f_i \right) \qquad (1)$$

Where I represents the resultant X-ray intensity, I_0 represents its initial intensity that passes through K different materials. a_i and f_i are the X-ray intersection length and the absorption coefficient for the i th material respectively. For computational conveniences, equation (1) can be rewritten as a linear form in the following

$$h = \log\left(I_0 / I\right) = \sum_{i=1}^{K} a_i f_i \tag{2}$$

Here the variable h , which physically represents a log-scaled measure of decayed X-ray intensity, is defined as the linear combination of a_i and f_i .

Practically, if we consider the projection of the jth X-ray in a 2D object, which holds N pixels as illustrated in Fig.1, we can further rewritten (1) such that

$$h^j = \sum_{i=1}^{N} a_i^j f_i, \qquad j = 1,2,\cdots,M \tag{3}$$

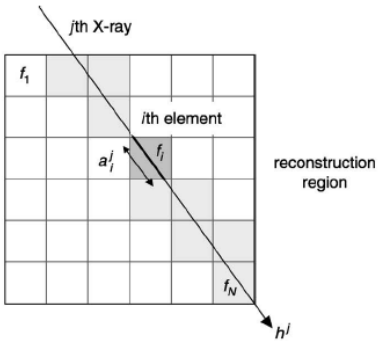

Fig.1 The projection of algebraic methods.

For this statement, the weighting coefficient a_i^j describes the jth X-ray intersection length at the ith pixel of the object. M is the total number of rays. The variable f_i is the absorption rate for X-ray photons, which corresponding to the average gray level of the CT image, of the ith pixel. Geometrically variables f_i can be seen as a point located in an N-dimensional space.

In practical utilization, the equation (3) which includes M linear equations with N unknown values of f_i can be rewritten as a matrix equation

$$\mathbf{h} = \mathbf{A}\mathbf{f} \tag{4}$$

where

$$\mathbf{h} = \left(h^1, h^2, \cdots, h^3\right)^T, \qquad \mathbf{h} \in \mathbb{R}^M$$
$$\mathbf{A}_{j,i} = a_i^j, \qquad \mathbf{A} \in \mathbb{R}^{M \times N}$$
$$\mathbf{f} = \left(f_1, f_2, \cdots, f_N\right)^T, \qquad \mathbf{f} \in \mathbb{R}^N$$

The vector $\mathbf{f}$ represents the absorption rate distribution within the reconstruction volume, and the vector $\mathbf{h}$ represents the X-ray projections that are collected from X-ray images by the detector array. The weighting matrix $\mathbf{A}$ represents the X-ray imaging geometry determined by the relations between the co-ordinates of the X-ray sources, the object and the image planes. The main computational issue of image reconstruction in CT is to solve the matrix equation (4), and then determines the absorption values of the object (gray level of the image). Because of some unavoidable system noise, the solution of the equation (4) frequently has no closed-form solution.

Theoretically, when the total number of rays M and the total number of pixels N are small, it is applicably to find the solution of the equation (4) by some direct algebraic methods such as generalized inverse or the singular value decomposition. However, in many practical applications, high resolution images for precisely analysis is prerequisite, it will lead the sizes of the matrices in (3) are usually too large to be solved by direct methods.

Instead, an iterative reconstruction method, the ART algorithm, is available as follows:

$$f_i\left(t+1\right) = f_i\left(1\right) + \lambda \frac{a_i^j \left(g^j - h^j\left(t\right)\right)}{\sum_i^N \left(a_i^j\right)^2} \tag{5}$$

An update scheme is presented in this equation, where t is the index of iteration step and λ is a relaxation parameter controlling the convergence performance. This algorithm updates at the further step for an ith unit $f_i\left(t\right)$ based on a measurement of the jth ray g^j and its estimation value h^j. The convergence of (5) will be faster as λ increases, but it does not performance well, even diverges, when specified conditions are exceeded, which are problem-dependent and also depends on the initialized density values $f_i\left(0\right)$.

At the further study of CT image reconstruction, the modified simultaneous ART scheme had introduced [6]. SART is a new computational recursively approach based on traditional ART considering the update rule of the form

$$f_i\left(t+1\right) = f_i\left(1\right) + \lambda \frac{\sum_{j=1}^{M} a_i^j \dfrac{g^j - h^j\left(t\right)}{\sum_{i=1}^{N} a_i^j}}{\left\| \sum_{j=1}^{M} a_i^j \right\|} \tag{6}$$

Because each iteration of SART reduces the value of the error terms $g^j - h^j\left(t\right)$, convergence of the estimated vector $\mathbf{f}$ of the algorithm is assured to a real one. The main difference between ART and SART is that the ART algorithm can only update the values on the one ray at each time. In fact, as shown in Fig.2, there is not the only one ray that passes through an image unit (as a voxel in this paper) at each iteration. For the purpose of efficient computation, the SART algorithm makes estimations of all rays which pass through an image unit simultaneously.

III. A NEW PROCEDURE FOR WEIGHT ESTIMATION

The SART technique is restricted to the storage and arithmetic operations of weighting matrix $\mathbf{A}$. The high-resolution reconstruction of CT used SART requires enormous system resources to access the weighting matrix,

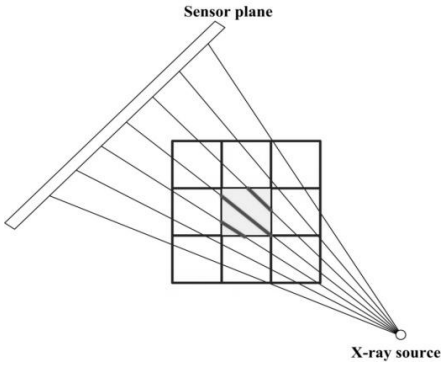

Fig.2 Rays which pass through an image unity at a time

which can be computationally infeasible during updating the equation (6) and leads to excessive computation times when making estimation of the absorption vector **f**.

In this study, a new weighting estimation procedure is adopted to simplify the calculation process of the weighting matrix **A**. The geometrical meaning such as its direction and magnitude of the weighting coefficient a_i^j adopted in this study is interpreted with Fig.3. Assuming a predetermined reconstruction area was constructed by 8 voxel units with the particular corresponding indices $i = 1, 2, \cdots, 8$. The size of the detector array was 2×2, which involves sensor units ***A, B, C, D***. Note that, the weighting coefficient a_i^j adopted in this paper describes the *j*th X-ray intersection length at the *i*th voxel of the object. If the total number of sensor units is *M(j=1, 2, ..., M)* and the total number of voxels is *N(j=1, 2, ..., N)* the corresponding computations for the weighting matrix **A** will be equal to $M \times N$.

Because the definition of the weighting coefficient is the intersection length of a ray which passed through a voxel unit, so the voxel index, the ray index and the intersection length of the path will all be recorded. In previous study, voxel indices and ray indices are preallocated to the register addresses first, and then the intersection lengths are assigned. In the weighting accessing process, the system needs to find the proper addresses of registers and access the values at the correct registers iteratively. This will decrease the computation inefficiency of the system when M and N are large.

The new procedure used to reduce the computations of the weighting matrix is based on the collimation of X-ray and the geometric correlation between each adjacent voxel. When a ray passed through an examined voxel, the adjacent voxel which the ray will also pass through can be determined by using the direction information of the ray.

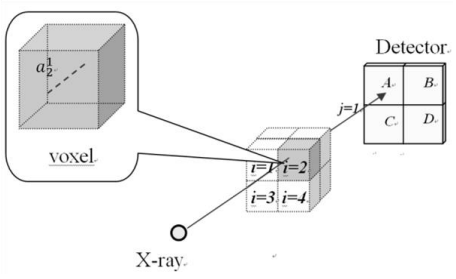

Fig.3 Geometrical meaning of weighting coefficients

Through the proposed rocedure, the iterations of finding the adjacent voxel decreased significantly. The corresponding number of iterations to identify the adjacent voxel the ray passed through using this procedure is equal to the number of voxels which the ray will pass.

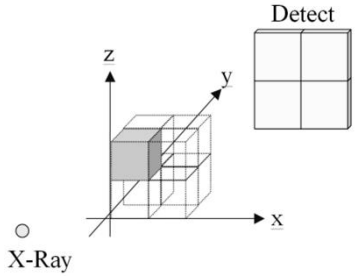

Fig.4 The 3D coordinate diagram of the illustrated system

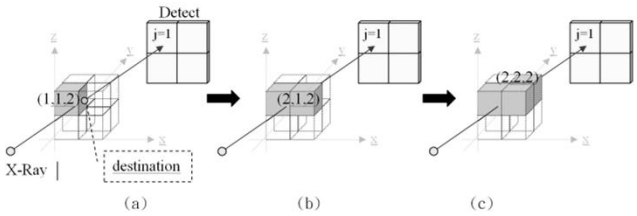

Fig.5 The new procedure for weighting estimation with a ray passed through an examined voxel

The new procedure for weighting estimation can be illustrated in Fig.4 which considering the ray $j = 1$ passed through a examined voxel. The initial conditions of the system are shown in Fig.5a and described as the following:

1) The first voxel that the ray passed through is (1,1,2).

2) Because the position of the X-ray source and the detector are known, thus the direction of the ray is (+, +, +).

3) The destination of the ray that passes through the voxel (1,1,2) is at the voxel's boundary $x = 1$.

Then the ray will also passed through the adjacent voxel (1+1,1,2) = (2,1,2), as shown in Fig.5b. Similarly, the destination of the ray that passes through the voxel (2,1,2) is at the voxel's boundary y=1, the ray will also passed through the adjacent voxel (2,1+1,2)=(2,2,2), as shown in Fig.5c. Furthermore, the trajectory of X-ray source and detector arrays of a CT system is normally in a circular form, when the location of the X-ray source and detector arrays changed, the direction of the ray will also changed. Eight possible directions of the ray in a 3D coordinate space will be (+, +, +), (+, +, −), (+, −, +), (+, −, −), (−, −, +), (−, +, +), (−, +, −), (−, −, −).

Because the definition of the weighting coefficient is the intersection length of the *j*th ray in the *i*th voxel, which is signal independent between each detector in the system (the data-independent property), it is generally a substantial parallelism task suiting for the GPU coprocessor system to improvement the computing efficiency.

A GPU contains a number of multiprocessors, each executing instructions in an SIMT (Single Instruction, Multiple Thread) manner and dividing the workload over a number of smaller processing elements. In many successful practical applications, there are numerous experiment results shown that the novel GPU coprocessor architecture has outstanding performance in extensive engineering and scientific computing problems. In this paper, the Nvidia GTS250 GPU is used as a parallel computing coprocessor and the Nvidia CUDA (Compute Unified Device Architecture) framework is used as a GPU programming language based on C/C++. The improvement results of using GPU coprocessor in SART in our study is described in the following context.

IV. SIMULATION RESULTS AND DISCUSSIONS

The purpose of the experiment in this study is to validate the performance of using GPU as a parallel computing coprocessor to accelerate the SART process in the CT reconstruction task when limited images is taken. The new procedure for weight estimation is also used to reduce the computational complexity of the experiment system. The experimental sample used in previous study [4] is also adopted in this study, which is a BGA (Ball Gird Array) sample composed of 2×2 solder balls as illustrated in Fig.6.

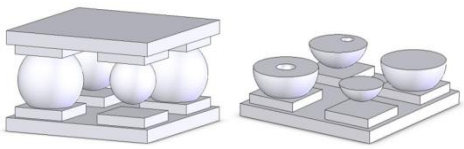

Fig.6 The BGA sample used for simulation study

A. Performance of Using the New Procedure for Weighting Estimation

In this study, the effects of the new procedure for the determination of weighting matrix on the processing time were tested first. The size of detector arrays is assumed to be 181×181, which means that there are 181×181 rays emitted from the X-ray source to the detector arrays. The size of the reconstruction area is assumed to be 61×61×33, which means that there are 61×61×33 voxels of the object should be taken into account. In previous study, it requires almost 4×10^9 ($61^2\times33\times181^2$) iterations to determine the weighting matrix **A**. It is apparent that the previous approach for building the weighting matrix doesn't seem to be particularly useful. In contrast to the previous approach, the new procedure for the determination of weighting matrix needs only $k\times181^2$ iterations.

A comparison of the new procedure for the determination of weighting matrix with the previous approach for performance improvement at different computing platforms, Visual C++ 2008 and Matlab R2008a, are shown in Fig.7. As shown in the figure, no matter how many x-ray images were used for tomography computing,

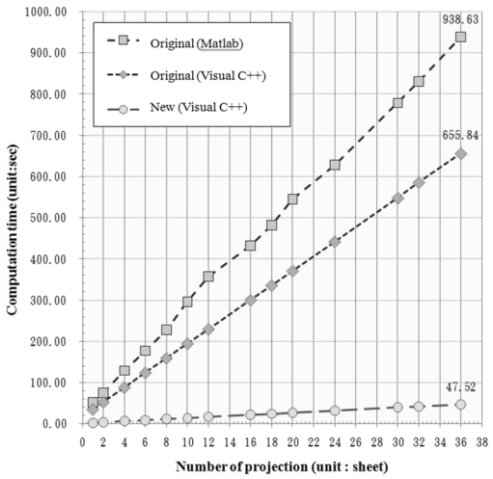

Fig.7 Effects of using the new approach for weighting matrix determination at C++ and Matlab platform on the processing time

the processing time with C++ platform is around 2/3 of that with Matlab platform. In the other hand, when 36 images were used for tomography computing, there is nearly 13-times performance improvement of using the new procedure for weighting matrix than the previous ones. In comparison with the previous study [4], which considering the implementation of SART at Matlab platform, the processing time of using the new procedure at C++ platform is almost 19 times faster. It is shown that the new procedure for the determination of weighting matrix can indeed be used to improve the inefficiency problem of using the previous approach.

B. Performance of Using GPU Coprocessor

A further study is conducted to compare the performance of CUDA GPU parallel computing coprocessor in this BGA reconstruction case to that of Intel Core 2 Quad Q9400 CPU. The test result is shown in Fig.8. Practically, there are 1024 bytes for one operation in the computer system. Because the simulation condition is set as that there has no unnecessary waste of the memory in the GPU coprocessor, thus the total amount of data which input the coprocessor is 1024×1024 .

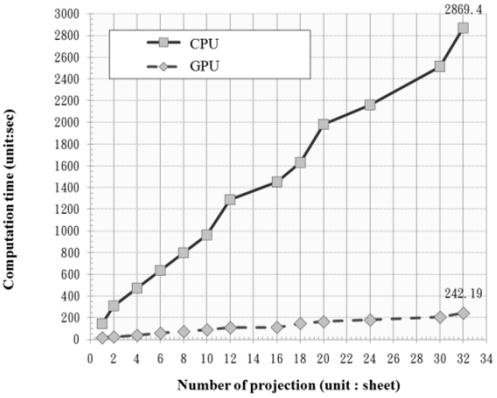

Fig.8 Processing time of using GPU coprocessor

In Fig.8, the advantage of using GPU coprocessor at SART is apparent; when 36 images were used for tomography computation, the processing time of using GPU coprocessor is almost 17.5 times faster than the processing time of using traditional 4-cores CPU. Because of the limited memory size, the maximum data loading capacity is 24 projections in a GPU processing cycle, the phenomenon is also depicted in Fig.8.

C. Image Quality

Finally, for the purpose of validating the image quality, the Correlation Coefficient (CC) [7], is used to quantify the correlative quality of images. In this study the tomography image reconstructed from 32 projected images after 2-cycle iterative reconstruction were compared. Other system parameters are listed in Table 1.

The reconstruction quality of the BGA object at difference layer is shown in Fig.9. The CC index between reconstruction image and reference image in these simulation results are almost all over 0.95 level. There are some difference between the simulation results with those conducted in the previous study [4]. This is so that the results in this study conducted using the GTS250 GPU is absolutely the single-precision (SP) arithmetic result. The traditional CPU structure supports double-precision (DP) computing. Indeed, the word double derives from the fact that a DP number uses twice as many bits as a SP number. The extra bits increase not only the arithmetic precision but also the range of magnitudes that can be represented. This is the main reason for the difference occurred.

Table 1 System parameters used in this study

X-Ray type	cone beam
Detector size	1810μ m×1810μ m
Image size	181 × 181 pixels
Reconstruction grid	61×61×33
Reconstruction grid resolution	1 μm^3 / voxel
Source distance*	160 mm
Target distance **	80 mm
Relaxation factor(λ)	0.5
Number of iteration(iter)	2
Number of projections	32

* Source distance：Distance between X-Ray source and detector plane

**Target distance：Distance between X-Ray source and object

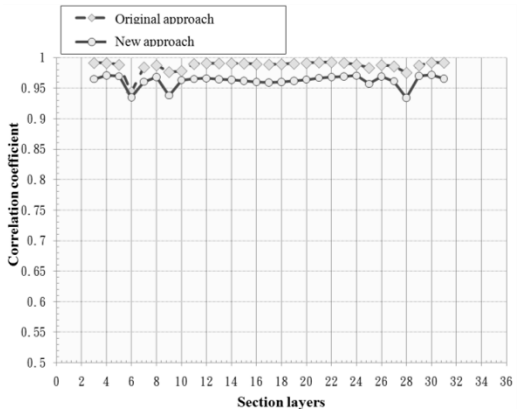

Fig. 9 The reconstruction quality of the BGA case

V. CONCLUSIONS

In this study, the SART approach was adopted to insure a better image quality when limited images were taken. Furthermore, for the efficiently computing purpose, we crystallize our research goal that aimed at an in-depth investigation of several related domestic in the scope of CT image reconstruction with CUDA GPU parallel computing technology, both in theory and experiment to reduce processing time. Based on the simulation results conducted, the following conclusions can be made:

A. CT Image

For the simulation case studied, the correction coefficient of the quality of reconstructed slice images is all up to 0.93 when the parallel computing technology and the GPU is adopted.

B. Reconstruction Efficiency

No matter how many x-ray images were used for tomography computing, the processing time with C++ platform is around 2/3 of that with Matlab platform. In the other hand, when 36 images were used for tomography computing, there is nearly 13-times performance improvement of using the new procedure for weighting matrix than the previous ones.

C. GPU Parallel Computing

When 36 images were used for tomography computation, the processing time of using GPU coprocessor is almost 17.5 times faster than the processing time of using traditional 4-cores CPU.

ACKNOWLEDGEMENTS

The author would also like to convey thanks to the National Science Councils, R.O.C. for their financial support (NSC 100-2221-E-007-021).

REFERENCES

[1] A. C. Kak, and M. Slaney, *Principles of computerized tomographic imaging*, Society for Industrial Mathematics, (1988).

[2] C. C. Liang, *The study of using FDK reconstruction technique on BGA inspection*, M.S. Thesis, National Tsing Hua University, Taiwan (2008).

[3] F. Xu, “Rapid Tomosynthesis with Exact Ray-driven Projector using Graphics Processing Units,” *2nd Workshop on high Performance Image Reconstruction*, pp. 33–36 (2009).

[4] Y. Y. Lin, *The study of using simultaneous algebraic reconstruction technique on BGA inspection*, M.S. Thesis, National Tsing Hua University, Taiwan (2008).

[5] Y. J. Roh, and H.S. Cho, “Implementation of uniform and simultaneous ART for 3D reconstruction in an X-ray imaging system,” *IEE Image Signal Process*, Vol. 151 (2004).

[6] A. H. Andersen, and A.C. Kak, “Simultaneous algebraic reconstruction technique: a superior implementation of the ART algorithm,” *Ultrason. Imaging*, Vol. 6, pp. 81–94 (1984).

[7] K. Mueller, R. Yagel, and J. J. Wheller, “Anti-aliased three-dimensional cone-beam reconstruction of low-contrast objects with algebraic methods,” *IEEE Transactions On Medical Imaging*, Vol. 18, No. 6 (1999).