Proceedings of the Federated Conference on Computer Science and Information Systems pp. 679–683
DOI: 10.15439/2016F344
ACSIS, Vol. 8. ISSN 2300-5963

Acceleration of image reconstruction in 3D Electrical Capacitance Tomography in heterogeneous, multi-GPU system using sparse matrix computations and Finite Element Method

Paweł Kapusta*,Michał Majchrowicz†,Dominik Sankowski‡ and Lidia Jackowska-Strumiłło§
Lodz University of Technology
Institute of Applied Computer Science
ul. Stefanowskiego 18/22, Łódź, Poland
* Email: pawel.kapusta@p.lodz.pl † Email: mmajchr@iis.p.lodz.pl ‡ Email: dsan@iis.p.lodz.pl § Email: lidia_js@iis.p.lodz.pl

***Abstract*—3D Electrical Capacitance Tomography provides a lot of challenging computational issues that have been reported in the past by many researchers. Image reconstruction using deterministic methods requires execution of many basic operations of linear algebra. Due to significant sizes of matrices used in ECT for image reconstruction and the fact that best image quality is achieved by using algorithms of which significant part is FEM and which are hard to parallelize or distribute. In order to solve these issues a new set of algorithms had to be developed.**

I. Introduction

ELECTRICAL Capacitance Tomography (ECT) is a relatively new imaging technique that can be used for non-invasive visualization in industrial applications in 2D, 3D and even 4D dynamic mode. ECT is performing the task of imaging of materials with a contrast in dielectric permittivity by measuring capacitance from a set of electrodes (Fig. 1). Among other non-invasive imaging techniques, ECT is characterized by much higher temporal resolution than Magnetic Resonance Imaging, Computed Tomography etc.

Unfortunately to achieve best image quality in 3D image reconstruction complex algorithms have to be used, especially ones that use large sensitivity matrices, Finite Element Method as well as neural networks approach [3].

In this article the authors have focused on accelerating non-linear image reconstruction algorithms, that are based on Finite Element Method and use sparse matrices to store data. We show that it is indeed possible to parallelize such algorithm and achieve significant speed-up, as well as develop them in such a way, to be able to use them in a distributed, heterogeneous computational system.

A. Image reconstruction in ECT

The scheme of image synthesis in Electrical Capacitance Tomography is called image reconstruction. It is based on solving the so called inverse problem, in which the spatial distribution of electric permittivity from the measured values of capacitance C is approximated. We can distinguish two types of image reconstruction algorithms. Firstly there are linear algorithms, which, because of higher temporal resolution, are used for monitoring fast-varying industrial process applications, like oil-gas flows in pipelines [1] or gravitational flows and discharging of silo [9] and non-linear algorithms, which allow reconstructing images with higher quality. Afterwards, reconstructed images can be analysed using either state of the art algorithmic approach, such as fuzzy-logic based classification [1] or by using a novel method of applying crowdsourcing [2], in order to determine, for example, flow characteristics.

II. Non-linear reconstruction algorithms

Non-linear three-dimensional image reconstruction in 3D capacitance tomography is a complex numerical problem, saturated with linear algebra transformations. During this iterative calculation process a set of parameters is determined, that is necessary for proper reconstruction of three-dimensional tomographic image optimization. The general idea of the algorithm is presented in Figure 2. One of the three key stages of the iterative process of reconstruction is a forward problem involving setting up a simulated vector based on a given spatial distribution of dielectric permittivity. The accuracy of the forward problem solution has a significant impact on the quality and speed of image reconstruction, and depends on the method of its determination. Most often forward problem is determined numerically using the Finite Element Method (FEM) based on a numerical model of a capacitance sensor. The authors have focused primarily on developing methods for accelerating the

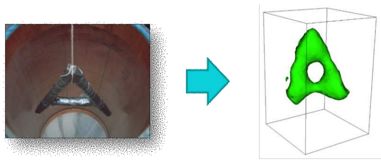

Fig. 1. Object and 3D reconstruction obtained using Electrical Capacitance Tomography

calculations using algorithms developed specifically for use with sparse matrices (CULA library, CUSP). This made it possible to develop proprietary paralel computing algorithms (as a set of functions and procedures), dedicated to specific processing of tomographic data. Developed methods allow reconstructing three-dimensional images by using relatively fast methods of solving sparse matrix equations (AMG method - Algebraic Multi Grid, the Jacobi method and the Conjugate Gradient algorithm), which are computed on graphic processors.

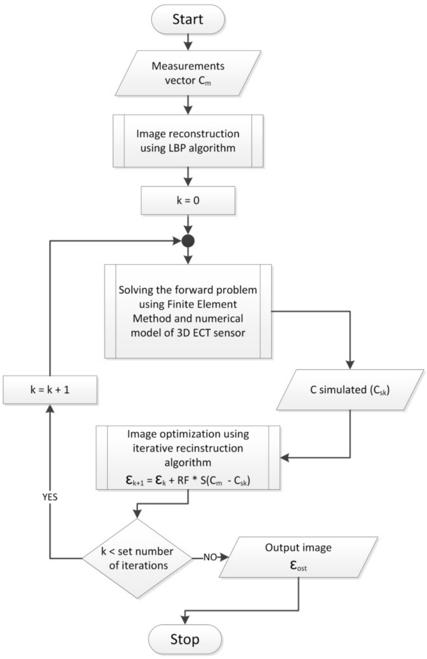
Start
Measurements vector C_m
Image reconstruction using LBP algorithm
k = 0
Solving the forward problem using Finite Element Method and numerical model of 3D ECT sensor
C simulated (C_{sk})
k = k + 1
Image optimization using iterative recinstruction algorithm
$\varepsilon_{k+1} = \varepsilon_k + RF * S(C_m - C_{sk})$
YES
k < set number of iterations
NO
Output image
ε_{ost}
Stop

Fig. 2. General non-linear reconstruction algorithm in ECT

A. Finite Element Method

Image reconstruction algorithms using the Finite Element Method are often used in 3D Electrical Capacitance Tomography because of the possibility of obtaining a more accurate solution to the forward problem than linear algorithms, which in turn can improve spatial resolution of resulting 3D images. A major drawback of this method, however, is its large computational complexity. The authors have developed a number of proprietary software algorithms, which are designed to significantly reduce the time of image reconstruction using the Finite Element Method, through the implementation of parallel computing for sparse matrices and calculations in a heterogeneous and distributed environments.

Main idea of the developed algorithm is to obtain the solution (electric field distribution) given in the form of equation:

$$\varphi = \boldsymbol{Y}^{-1}\boldsymbol{F} \tag{1}$$

where:

φ - is a sought distribution of the electric field - represented by the spatial distribution of nodal potential - partial solution of the forward problem in capacitance tomography;

$\boldsymbol{Y}$ - is a transformation matrix, built according to the geometric dependencies of sensor model mesh and Neumann boundary conditions;

$\boldsymbol{F}$ - is the extortion vector, defining the given Dirichlet boundary conditions

The first step of the algorithm is to pre-process the input data and store it as a set of sparse matrices. Then, in order to obtain $\boldsymbol{Y}$, matrix decomposition is performed as described by the equation:

$$\boldsymbol{Y} = \boldsymbol{A}^{\boldsymbol{T}}\boldsymbol{B}\boldsymbol{A} \tag{2}$$

where:

$\boldsymbol{A}$ - shape functions gradients matrix

$\boldsymbol{B}$ - matrix of normalized mesh volumes

$\boldsymbol{A}^{\mathrm{T}}$ - transposed shape functions gradients matrix

The next step of the algorithm is processing of the input data matrix and selecting the rows corresponding to each electrode - known potential in nodes describing the electrode. Then the matrix is preconditioned using either Jacobi or Algebraic Multi-Grid method. This makes it possible to solve the equation (1) using a Conjugate Gradients Method. Once this is done the resulting matrix is supplemented with data from the Dirichlet boundary conditions. The last step of the algorithm is to determine the vector of simulated measurements using Gauss' law described by the formula:

$$C_{eg} = \frac{\iiint_{\Omega} \varepsilon\,(x, y, z)\; grad[\varphi(x, y, z)]d\Omega}{\varphi_e - \varphi_g} \tag{3}$$

where:

C_{eg} – Capacitance between electrodes e and g

ε(x,y,z) – distribution of electric permitivity

φ(x,y,z) – distributon of potential

φ_e – electric potential on electrode e

φ_g – electric potential on electrode g

x,y,z – cartesian coordinates

B. Computations using sparse matrices

The operation of multiplying three matrices, represented by the formula (2), is an integral part of the Finite Element Method for 3D ECT. This action, however, is characterized by high computational complexity. Moreover, the stiffness matrices $\boldsymbol{Y}$, generated by numerical models of 3D ECT sensors, are too large to fit entirely in RAM of graphics cards. However, the number of non-zero elements is relatively small in relation to the dimensions. Thus, it is possible to treat them as sparse matrices to reduce memory usage.

There are many formats for storing sparse matrices. Among them the most common formats are CSC (Compressed Sparse Column) and CSR (Compressed Sparse Row). The authors decided to use CSR format because it allows, in most cases, for optimal access to the data stored in GPU memory. Reading and writing data is usually done in a row-major manner, which is optimal for most architectures of CPUs and GPUs. Saving sparse matrix in the CSR format is, for the same reason, not optimal for multiplication, as there needs to be a way of quickly accessing the columns of the matrix without causing uncoalesced reads/writes from GPU memory. This situation arises when data must be read in a manner that does not comply with optimal memory access for specific hardware and cannot be obtained in one transaction. The impact of this phenomenon on the speed of computations is highly dependent on the hardware architecture of the GPU, however, it is always significant.

In order to significantly reduce this problem the authors have introduced a hybrid format, called Hybrid Compressed Sparse Row-Column (H-CSRC), comprised of both the records of CSR and CSC. Depending on the needs, data can be accessed in either row or column-major manner, while minimizing memory operation and maintaining compatibility with other algorithms.

Multiplication of three sparse matrices has been implemented as a single operation. This approach allows for optimal use of local and private memory on the GPU, in order to increase the speed of calculations. In this algorithm, it is necessary to use the local memory, shared by thread groups, to minimize the number of global memory accesses.

In 3D ECT it is particularly important to optimize each of the algorithms for the speed of execution. Hence the authors have developed a special version of three matrix multiplication algorithm, which takes into account all the specific properties of the matrix calculations in the 3D ECT, as defined by equation (2). There are three main properties of the input data, specific to the ECT, that allow for further optimizations:

- Items in the matrix $\boldsymbol{B}$ have a non-zero values on the main diagonal only. In addition, they repeat in sets of three, which is due to the specificity of the input data.
- The output array is symmetrical along its main diagonal, which, using proper element indexing, can reduce the number of operations almost by half.
- Due to the nature of the calculations, the amount of output elements and their position, does not change during the execution of the program, assuming the immutability of input data distribution. Hence this can be determined before the execution of the program and put into the algorithm as a map of elements to instantly skip the input matrix elements, which are known a priori to not produce results.

C. Parallelization

The first variant of image reconstruction algorithm using the Finite Element Method is the reconstruction in the local system. As it constitutes a platform for further modifications it was necessary to design and implement a solution, that would be also applicable in multi-GPU [4], as well as distributed systems. To ensure efficient 3D image reconstruction the proposed algorithm includes data caching solutions. This issue is particularly important in the case of heterogeneous systems. In most 3D ECT systems measurement data is collected with higher frequency than it can be reconstructed. Moreover, because of the asynchronous nature of the developed solution, based on the commissioning of tasks to local GPUs using CUDA technology, as well as remote computing nodes, delays can accumulate, therefore there is a need for their elimination by buffering systems. All the algorithms have been designed, implemented and optimized from the start as a solution suited to multi-GPU and distributed systems. Due to the specific nature of the computations the most optimal solution is to start a separate thread for each GPU in the system, that are synchronised when reading the results.

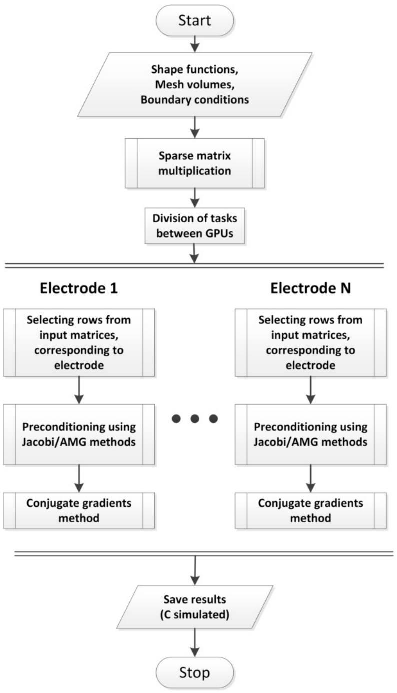
Start
Shape functions, Mesh volumes, Boundary conditions
Sparse matrix multiplication
Division of tasks between GPUs
Electrode 1
Electrode N
Selecting rows from input matrices, corresponding to electrode
Preconditioning using Jacobi/AMG methods
Conjugate gradients method
Selecting rows from input matrices, corresponding to electrode
Preconditioning using Jacobi/AMG methods
Conjugate gradients method
Save results (C simulated)
Stop

Fig. 3. Algorithm for calculating solution to forward problem using multiple GPUs

All the algorithms, developed by the authors, were designed to function in systems with multiple GPUs. Thus a natural direction for obtaining a further acceleration of computations is to perform non-linear 3D image reconstruction in a distributed system, which tends to have a higher degree of heterogeneity than the local systems.

The main idea of the developed algorithm is, that each GPU inside the compute calculates the solution to forward problem for one or more electrodes, by assigning to each GPU calculations specific to the selected electrode of the system, where all the GPU compute the result of a single image reconstruction (Fig. 3). This approach allows for more precise control of tasks allocation. This in turn enables its use in distributed systems with a high degree of heterogeneity. This solution also allows for potential reduction in overall system response time, since all nodes perform calculation for a single output image. Therefore, the task of caching is greatly simplified, and the image can be displayed on the screen without introducing delays larger than the reconstruction time for a single image. The disadvantage of this solution is, however, that it limits the scalability of a distributed system, since the total number of graphics processors cannot be greater than the number of electrodes.

TABLE I
COMPONENTS OF TEST SYSTEMS

Processor	HPC Hal: Intel i7-930 (4 cores, 8 threads) HPC Dave: Intel i7-920 (4 cores, 8 threads)
RAM	HPC Hal: 12 GB (6x2 GB) DDR3 1833 MHz HPC Dave: 8 GB (4x2 GB) DDR3 1833 MHz
GPU accelerators	HPC Hal: NVIDIA Tesla S1070 + Tesla C2070 HPC Dave: 2x GTX 570
Operating systems	Windows 7 64-bit

TABLE II
RESULTS OF NON-LINEAR IMAGE RECONSTRUCTION [IMAGES/SECOND]

Elements in image vector	4 GPUs	2 GPUs	GPU	CPU (BLAS)
8488	0.035	0.024	0.015	0.003
20499	0.021	0.012	0.007	0.002
60896	0.007	0.004	0.002	0.001
87172	0.005	0.003	0.002	<0.001
157264	0.003	0.002	0.001	<0.001

III. RESULTS

The research conducted on ECT algorithms [6] has shown that, although, dynamic development of GPU computing performance and its recent application for image reconstruction in ECT has significantly improved calculations time, in modern systems a single GPU is not enough to perform many tasks [7]. As a result multiple GPUs have to be used to accelerate calculations [5]. Thus, the authors are proposing a distributed, multi-node, multi-GPU heterogeneous system with a software layer that will allow use of multiple computers with fast GPUs to perform calculations across network connection [5]. The developed system, based on the proprietary KISDC networking platform [8], is designed to fully exploit parallel performance of all devices that the nodes are equipped with. Such architecture is very scalable and makes it possible to increase computation performance by adding new network nodes. Reconstruction algorithm verification tests were conducted using real measurement data, recorded during the research under Ministry of Education grant number 4664/B/T02/2010/38, using semi-industrial installation. Due to the nature of calculations using the graphics processors, the stability of execution times is lower than for algorithms executed on the CPU. Therefore, all of the results shown in this paper represent the worst case scenario - the lowest number of reconstructed images per second, achieved during testing.

All the results achieved with GPUs were compared with the performance of algorithms executed on the CPU, implemented using optimized BLAS libraries (Basic Linear Algebra Subprograms), compiled with Intel compiler and optimized for the tested CPU architecture.

A. Non-linear algorithms - local system

Reconstruction tests using non-linear algorithms and multiple graphics processors at the same time were carried out using NVIDIA Tesla C2070 card and NVIDIA Tesla S1070-400 computing server, which has four graphics cores (Table I). Verification of developed solutions in this case was performed for 1, 2, and 4 GPUs. The division of tasks between the graphics processors was done by creating a new thread for each GPU. As a result, it was possible to separate the control flow of the application from computations, thus allowing for asynchronous commission of tasks to the GPUs. Tests were performed for 10 iterations of non-linear reconstruction algorithm. The test results are presented as the number of images obtained per second.

All the tests were performed using the developed task division algorithm, by assigning each GPU calculations for a specific set of electrodes (Fig. 3). Moreover, verification was conducted using an optimized version of this algorithm, which enables asymmetric division of compute jobs between the units. Based on the known efficiency parameters, each GPU was assigned with solving the forward problem for appropriate number of electrodes. For example, by using two GPUs - Tesla C2070 card and a single GPU from Tesla S1070 accelerator, the C2070 card computes solution for 18 electrodes, and the S1070 GPU for the remaining 14. The results for this configuration are shown in Table II and in Figure 4.

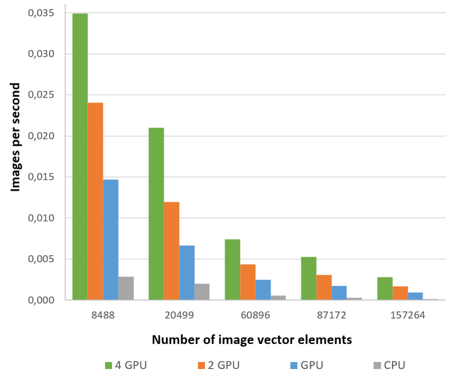

Fig. 4. Speed of non-linear reconstruction

As can be seen from Table II and Fig. 4, the use of multiple GPUs allowed to accelerate a non-linear image reconstruction by up to seventeen times for the two graphics processors and 28-fold in the case of four GPU compared to calculations using traditional algorithms executed on a CPU.

B. Non-linear algorithms - distributed system

Image reconstruction tests in distributed environment were performed using two HPC nodes code named: Dave and Hal. Their full specification is shown in Table I. In this case verification was performed for a total of 6 GPUs - four in HPC Hal (Tesla C2070 + Tesla S1070) and two in HPC Dave (2x Nvidia GTX 570). As it was in the case with computations in the local system all the tests were performed using task allocation based on number of electrodes. The data was sent between the nodes using KISDC networking layer, developed specifically by the authors for use in distributed

TABLE III
RESULTS OF DISTRIBUTED NON-LINEAR IMAGE RECONSTRUCTION [IMAGES/SECOND]

Elements in image vector	**Local system 4 GPUs**	**Distributed system 4+2 GPUs**	**Speed-up**
8488	0.0349	0.0381	1.09
20499	0.0210	0.0198	0.94
60896	0.0074	0.0086	1.16
87172	0.0053	0.0059	1.12
157264	0.0028	0.0034	1.20

image reconstruction in 3D ECT. All the results for these tests are presented in Table III and in Figure 5.

Using the distributed system for image reconstruction purposes the authors were able to speed-up the computations compared to local system by up to 20%. There was however one exception - for the image vector size of 20499. In this case the computations on a distributed system were slower than in local environment. This was caused by a combination of overheads resulting from synchronisation and network delays. Moreover, because of the specifics of GPU computations it is common to come across a combination of input data sizes and algorithm logic that will cause overall slow-down in specific cases. Nevertheless, the authors are sure that further work on the developed algorithms will result in even better results in the future.

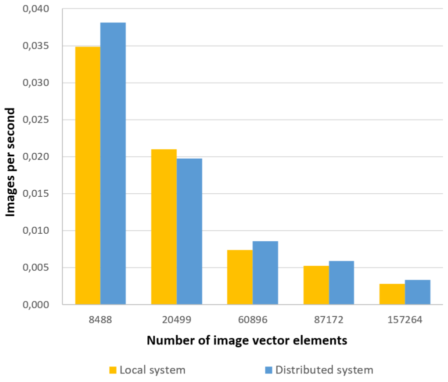

Fig. 5. Speed of non-linear reconstruction

IV. CONCLUSION

As a part of the research authors have developed a flexible, distributed computing system, intended for tomographic image reconstruction process. For both the linear and non-linear reconstruction algorithms parallel architecture developed by the authors is designed in such a way that it can be scaled to any number of computing nodes, assuming that the network medium is not a limiting factor. Performance tests have shown that the practical application of parallel algorithms executed on GPU allows for a 28-fold increase in the rate of performing calculations in the case of non-linear algorithms, compared to the optimized, sequential versions of algorithms.

Further research will also address challenges of calculations in heterogeneous and distributed environments. This work will be aimed at reducing delays, and therefore the response time of the system, due to the transmission of data over the network, as well as overall optimizations to increase the stability of the proposed solution. In addition, the authors are carrying out further work on the system, and the concept of complete reconstruction, utilizing both linear and non-linear algorithms, on the remote nodes. Ultimately, this will enable the transfer of all calculations to remote servers, connected to the data acquisition system over the Internet, thereby allowing monitoring and control of the industrial processes using smartphones or other mobile devices.

REFERENCES

[1] R. Banasiak, R. Wajman, T. Jaworski, P. Fiderek, H. Fidos, J. Nowakowski, D. Sankowski (2014), "Study on two-phase flow regime visualization and identification using 3D electrical capacitance tomography and fuzzy-logic classification," *International Journal of Multiphase Flow, Vol. 58*, 2014, pp. 1-14

[2] Chen, Ch., Woźniak P. W., Romanowski, A., Obaid, M., Jaworski, T., Kucharski, J., and Grudzień, K. and Zhao, S., Fjeld, M., "Using Crowdsourcing for Scientific Analysis of Industrial Tomographic Images," *ACM Trans. Intell. Syst. Technol., Vol. 7, No. 4*, 2016, ACM, pp. 52:1–52:25

[3] Garbaa, H., Jackowska-Strumiłło, L, Grudzień, K., Romanowski, A., "Neural network approach to ECT inverse problem solving for estimation of gravitational solids flow," *In Proc. of the 2014 Federated Conf. on Computer Science and Inf. Systems, AAIA'14, Vol. 2*, Warsaw, Poland, 2014, pp. 19-26

[4] Kapusta, P., Majchrowicz, M, "Accelerating Image reconstruction algorithms in Electrical Capacitance Tomography using Multi-GPU system," *Advanced Numerical Modelling, International Interdisciplinary PhD Workshop, Warsaw, Electrotechnical Institute*, 2011, pp. 47–49.

[5] Kapusta, P., Majchrowicz, M., Sankowski, D., Jackowska-Strumiłło, L., Banasiak, R., "Distributed multi-node, multi-GPU, heterogeneous system for 3D image reconstruction in Electrical Capacitance Tomography - network performance and application analysis," *Przegląd Elektrotechniczny, 89 (2 B)*, 2013, pp. 339-342.

[6] Majchrowicz, M., Kapusta, P., Banasiak, R. , "Applying parallel and distributed computing for image reconstruction in 3D Electrical Capacitance Tomography," *Zeszyty Naukowe AGH - Automatyka, Vol 14, Issue 3/2,* 2010, Kraków, Wydawnictwa AGH, pp. 711–722.

[7] Majchrowicz, M., Kapusta, P., Wąs, Ł., Wiak, S, "Application of General-Purpose Computing on Graphics Processing Units for Acceleration of Basic Linear Algebra Operations and Principal Components Analysis Method," *Man-Machine Interactions 3, Advances in Intelligent Systems and Computing Volume 242,* Springer International Publishing, 2014, pp. 519–527.

[8] Majchrowicz, M., Kapusta, P., Jackowska-Strumiłło, L., Sankowski, D., "Analysis of Application of Distributed Multi-Node, Multi-GPU Heterogeneous System for Acceleration of Image Reconstruction in Electrical Capacitance Tomography," *Image Processing & Communications, vol. 20, Issue 3*, 2015, pp. 5–14.

[9] Sankowski, D., Grudzień, K., Chaniecki, Z., Banasiak, R., Wajman, R., Romanowski, A., "Process tomograhy development at Technical University of Lodz," *Electrical Capacitance Tomography Theoretical Basis and Applications, edited by Dominik Sankowski and Jan Sikora,* Warszawa, 2010, pp. 70-95.