3D GPU-based Image Reconstruction Algorithm for the Application in a Clinical Organ-targeted PET Camera

Borys Komarov

Supervisors: Dr. Zubair Md. Fadlullah

Dr. Alla Reznik

Department of Computer Science

Lakehead University

August 2022

Abstract

Functional medical imaging is unique in its ability to visualize molecular interactions and pathways in the body. Organ-targeted Positron Emission Tomography (PET) is a functional imaging technique that has emerged to meet the demands of precision medicine and has shown advantages in terms of sensitivity and image quality compared to whole-body (WB) PET. A common application for organ-targeted PET is oncology, particular breast cancer imaging. In this work we present the application of Graphics Processing Unit (GPU) to significantly accelerate reconstruction of clinical breast images acquired with an organ-targeted PET camera and reconstructed using the Maximum Likelihood Estimation Maximization (MLEM) algorithm. The PET camera is configured with two planar detector heads with a sensing area of 232mm×174mm. Acquired raw image data are converted into list mode format and reconstructed by a GPU-based 3D MLEM algorithm that was developed specifically for the limited-angle capabilities of the planar PET geometry. The algorithm applies corrections including attenuation and scatter to provide clinical grade image quality. We demonstrate that a transition from originally developed Central Processing Unit (CPU) to newly developed GPU-based algorithm improves single iteration speed by more than 400 times, while preserving image quality. The latter has been assessed on clinical data and through phantom tests performed according to the National Electrical Manufacturers Association (NEMA) NU-4 standards. The gain in reconstruction speed is expected to result in improved patient throughput capabilities of the clinical organ-targeted PET. Indeed, GPU-based image reconstruction reduces time needed for a typical breast image reconstruction to less than 5 minutes thus making it shorter than the image acquisition time. This is of particular importance in improving patient throughput and clinical adaptation of the PET system.

Table of Contents

List of Figures

List of Tables

Acknowledgments

I would like to express my immense gratitude to my supervisors Dr. Fadlullah and Dr. Reznik for their support and guidance throughout my Master’s degree. The outcome of this work was made possible through their great expertise and ability to convey an abundance of knowledge to me from their respective domains.

Dr. Olexander Bubon and Dr. Edward Anashkin were the sources of technical and experimental inspiration. I am grateful to both for their expertise and assistance throughout my research work.

I am also thankful to my examination committee, Dr. Bajwa and Dr. Fouda for agreeing to examine me and for dedicating time to review my work.

I want to thank our industry partner Radialis for their collaboration, and for providing the PET technology and data for my research and experimentation.

Finally, I would like to extend my gratitude to my wife Anhelina Komarova, who has been my best friend and main driver during this journey. Without her I would not have succeeded.

1 Introduction

Breast cancer, like others, is a disease which is most successfully treated with early diagnosis. The most common tool for breast cancer screening in women over 40 is x-ray mammography. It has been shown that x-ray mammography is most effective in entirely fatty breasts, while it performs poorly in heterogeneous and extremely dense breasts [1] which are associated with nearly 50% of women in certain populations [2]. Increased breast density is associated with reduced sensitivity in mammography and is a strong risk factor in the development of breast cancer [1]. It has been recognized as the greatest contributor to the failure of mammography, resulting in false-negative findings delay diagnosis towards advanced stage cancers and result in poorer outcomes. These factors define a need to provide women with dense breasts with a supplemental screening technique which addresses the shortcomings of other modalities. The most widely used supplemental tests are digital breast tomosynthesis (DBT), whole breast ultrasound (US), and breast Magnetic Resonance Imaging (MRI). The best results in terms of sensitivity and specificity are shown by the breast MRI method, however this method is also prone to disadvantages such as high false-positive rate, high sensitivity to hormonal changes, and large associated cost.

The domain of functional medical imaging provides additional breast screening methods which allow observation of molecular interactions and pathways within the patient's body rather than relying on morphological differences. Organ-targeted Positron Emission Tomography (PET) with fluorodeoxyglucose (^{18}F-FDG/FDG) for breast, sometimes referred to as Positron Emission Mammography (PEM), offers greater specificity, lower associated cost, and a smaller list of contraindications compared to breast MRI. A summary of peer-reviewed studies illustrates the ability of PEM to reliably detect early forms of cancer with high specificity and independent of breast density [3]. PET relies on the properties of radioactive positron decay and annihilation which may be detected by groups of detectors surrounding a region of interest. With the development of technology various capabilities of PEM were investigated in several clinical studies and has shown great promise in preoperative evaluation [4] and treatment planning. However, more widespread clinical adaptation is limited due to high associated radiation dose and limited standardization of the technology.

These challenges are currently being addressed by Dr. Reznik and her research team as they develop and investigate a novel solid-state technology for organ-targeted PET with improved sensitivity, capable of significant dose reduction (factor of 10) in comparison to commercial whole-body (WB) PET scanners and with a potential reduction of the dose to the DBT level or even x-ray mammography level. The technology is commercialized by Radialis, Thunder Bay, Ontario, and is now called Radialis organ-targeted PET camera. The initial application of the Radialis system is in breast imaging. The first system prototype has been built and is currently used in clinical trials at University Health Network-Princess Margaret Cancer Centre (UHN-PMCC) in Toronto, Ontario. The second system is currently being finalized in Thunder Bay, Ontario and will undergo clinical trials at the University of Pittsburgh Medical Center (UPMC), Pittsburg USA. Radialis PET camera is a state-of-the art device that has shown great results in terms of system sensitivity and spatial resolution, thus allowing imaging with reduced dose and lower acquisition duration while still producing high-quality images. The goal of Radialis PET is to ensure that these improvements can progress from theoretical research to a widely used clinical tool.

Despite the hardware technology and imaging performance, a primary bottleneck associated with Radialis system workflow is the image reconstruction speed. In a clinical setting, reconstruction of image data should not take longer than the data acquisition in order to ensure a high rate of patient throughput. Clinical protocols often dictate that a patient cannot be released until all images are produced. Image reconstruction speed is a common bottleneck due to the high computational intensity of the iterative algorithms most commonly used in PET. The most common iterative reconstruction algorithm for PET is Maximum Likelihood Expectation Maximization (MLEM) and its subsets version: Ordered Subsets Expectation Maximization (OSEM). They have proven to be able to accurately enough solve PET image reconstruction problem. In its current implementation, the Radialis system uses a Central Processing Unit-based (CPU) MLEM reconstruction that is optimized for the detector geometry and adopted data format. It can produce clinical images of good quality, but for a price of high reconstruction time (20 minutes – up to hours). Iterative reconstruction speedup has been an active research area. The Graphics Processing Unit-based (GPU) reconstruction acceleration methods have become a gold standard for various PET systems. Slow CPU-based image reconstruction in Radialis system must be addressed to comply with high patient throughput requirements of the clinical environment.

1.1 Contributions

In this work, we present our effort in addressing this issue by reformulating the clinically suitable MLEM image reconstruction algorithm from CPU to GPU architecture. In doing this, a special consideration had to be taken for the nature of the intrinsically parallel GPU hardware and its thread and memory organization. We show our incremental progress and implemented optimizations that include reducing data transfer overhead, coalesced global memory access, effective shared memory use, and other methods to maximize GPU utilization, that were employed to make a successful transition to GPU. We emphasize that making such transition shall enhance the reconstruction throughput and shall not result in any deterioration in clinical imaging performance. The resulting GPU-based algorithm is compared to the CPU counterpart and evaluated using both clinical images and standardized phantom tests performed according to National Electrical Manufacturers Association (NEMA) NU-4 [5]. Results reported in this thesis demonstrate the superiority of the GPU-based algorithm in terms of reconstruction speed, thus ensuring improved patient throughput capability of the Radialis PET system. The novelty of this method lies in the analysis of its application to the system as a whole, considering the effects of this work onto clinical operation, and specific system characteristics.

1.2 Thesis Organization

The content of this dissertation is organized in the following way. Chapter 2 introduces underlying physics of PET, different types of PET devices, and a short literature overview on the topic of PET image reconstruction and current developments in the field. Chapter 3 dives deep into Radialis PET system and various techniques for building a modern clinical organ-targeted PET system. The main research problem is formulated at the end of the Chapter 3. In Chapter 4, we propose a fast and efficient GPU-based image reconstruction method for dual-head planar PET. Finally, Chapters 5 & 6 conclude the thesis and provide an insight into the future work that can be applied in the area, respectively.

2 Background and Related Works

2.1 General Overview of Positron Emission Tomography

Positron Emission Tomography (PET) is a functional medical imaging technique that allows to observe molecular interactions and pathways within the patient body [1]. This makes PET more specific for a wide range of diseases than anatomical imaging techniques such as Magnetic Resonance Imaging (MRI), X-Ray, and Computed Tomography (CT). In basic terms, one may look at anatomical imaging as something providing an information on the structure of organs [2] whereas PET gives an access to the functional, or molecular, processes happening inside the organs. It is important to mention that it is very common to combine PET with anatomical imaging techniques such as CT or MRI in dual-modality scanners to improve the diagnostic capabilities. In fused PET/CT or PET/MRI images, CT or MRI provides anatomical landmarks and morphologic information while PET pin-points abnormal functioning of an organ.

The specificity of PET arises from the range of positron-emitting radioisotopes which can be easily substituted directly into biomolecules without changing (or significantly disturbing) their biological function thus making a radioactive tissue biomarker. PET imaging provides a quantitative information on the distribution of tissue biomarkers (or PET radiopharmaceuticals, or PET tracers) in the body. Among the currently used radioisotopes ^{18}F takes a special place as it is used to substitute a hydroxyl group in a glucose molecule producing a FDG for identifying the regions of abnormal glucose metabolism. ^{18}F-FDG is currently the most widely used PET radiopharmaceutical because of its use in clinical oncology as well as in cardiology and neurology [3]. Although newest studies show the applicability of PET in the diagnosis of parkinsonism, drug discovery, and many other fields [4], [5], the most common clinical application for PET remains in cancer diagnosis and treatment in particular in detection of primary tumors; accurate staging of the disease and detection of lymph node metastases; treatment planning and assessment of the effectiveness of the therapy.

2.2 Basic Physics of PET

Being a nuclear imaging technique, PET relies on the physical properties of radioactive decay. PET imaging procedure starts from intravenously administration of PET radiopharmaceutical that binds to a targeted tissue or is up-taken in a lesion thus making this tissue or lesion more radioactive than surrounding areas of patient body. This excessive radioactivity can be detected with PET scanner as outlined below.

A radioisotope within a PET tracer decays emitting a positron ($\beta+$). A positron has only a transient existence: after traveling short distance called positron range, and losing its kinetic energy, positron annihilates with an electron from the surrounding tissue emitting two 511 KeV photons in the opposite directions (at roughly 180°) [6]. A PET scanner is developed to detect and localize the simultaneous back-to-back annihilation photons. Detecting both photons from a single annihilation event simultaneously (within a very short time interval which is referred to as a “coincidence window”) is called a “true coincidence” or a “true event” detection.

PET scanner consists of large number of detectors surrounding the object to be imaged. It is very common to place the surrounding detectors in a ring-like shape that encircles the patient (refer to Figure 2.1). The coincident photons are collected over many angles around the body axis of the patient allowing to draw multiple “lines of response” (LORs) that are lines connecting coordinates of two detected photons. Since 18F’s positrons have a short range (<1 mm), the point of positron emission can be approximated with the position of the annihilation event. Therefore, it is safe to assume that an annihilation event occurs along a LOR. Therefore, LOR data can be used to reconstruct the image of the activity distribution in slice or tomographic form.

In addition to a ring detector design shown in Figure 2.2, that is used in WB PET scanners, there are alternative designs with planar detector geometries which are employed mainly in organ-targeted imaging systems.

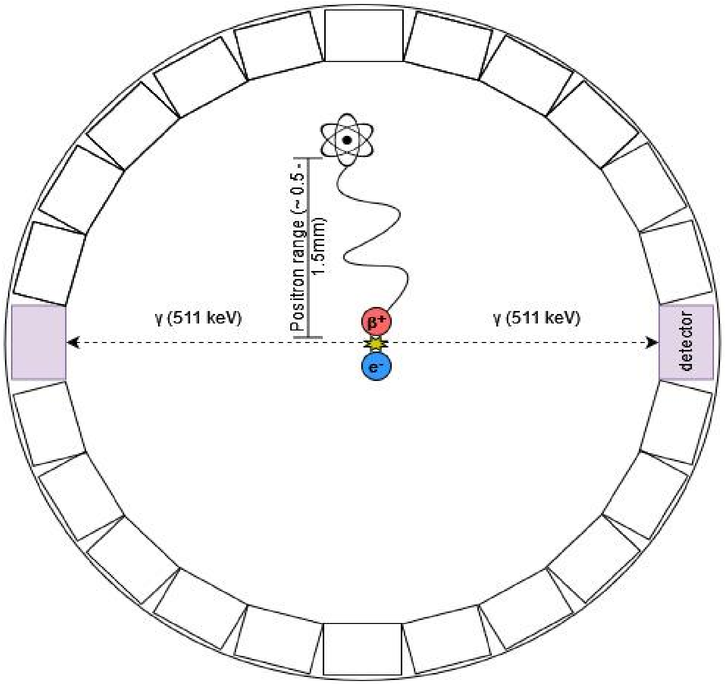


Figure 2.1 - Positron annihilation

2.2.1 Event Types

The positron annihilation detection becomes even more complex when we consider that there is happening hundreds of thousands of annihilations per second in-between PET scanner detectors. The rapid pace of the annihilations creates additional challenges for data collection in PET.

One of the consequences of this effect is that some of the detected coincidences might not be the desired "true" events thus affecting quantitative accuracy and the resulting image quality of the system [6]. Figure 2.2 schematically shows different event types. In case of the "true" events, it is a fact that a LOR for such event passes through the location where the annihilation of a positron occurred. On the other hand, if two gamma photons from two different annihilations reach detectors within the same coincidence window such event is categorized as random coincidence. Random coincidences do not possess any spatial information about the activity distribution, are hard to filter out and make a significant contribution to image noise. The other type of undesirable coincidences are scatter coincidences that involve scattering of one of the photons mainly via Compton scatter; Compton scatter reduces photon energy and deflect a photon from its initial direction. In conventional PET reconstruction algorithms, the scattered data cause the blurring of images and thus should be estimated and filtered out using an energy filtering of the events based on a concept of “energy window”. Finally, the last common event type we are going to discuss is multiple coincidences when more than 2 photons are detected within a coincidence window, in this case, we would normally discard such events.

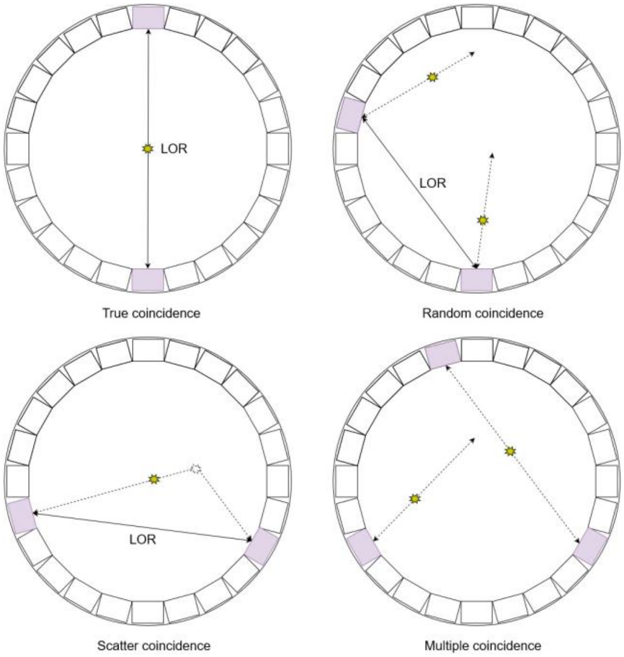


Figure 2.2 - Event types

2.3 Data Format

There are 2 commonly used data formats for PET: sinogram and list mode. In sinogram format every acquired coincidence is histogrammed into a 2-D matrix where each element is equal to the number of events detected in a particular LOR (between a pair of detectors). This matrix is arranged in the following way: each column corresponds to the radial offset from the center of the field of view (FOV) and each row is a projection of the coincidence at a particular angle. The visual representation of a sinogram is presented in Figure 2.3. Although sinogram format remains the most commonly used in conventional WB PET/CT, list mode format is more suitable for newer Time-of-Flight (ToF) PET technology since it provides a new aspect to use time of flight and energy information of each coincidence in the reconstruction. Specifically, list mode format records each event individually including information on the two locations (coordinates), where annihilation photons were detected, energy, and time when the coincidence happened. List mode format is easily extendable to include any additional details associated with the events which is especially important for the ToF PET/CT scanners. Because of its ability to preserve all available information some of the data corrections can be efficiently implemented only on the list mode data, for example the respiratory motion correction for cardiac PET [7].

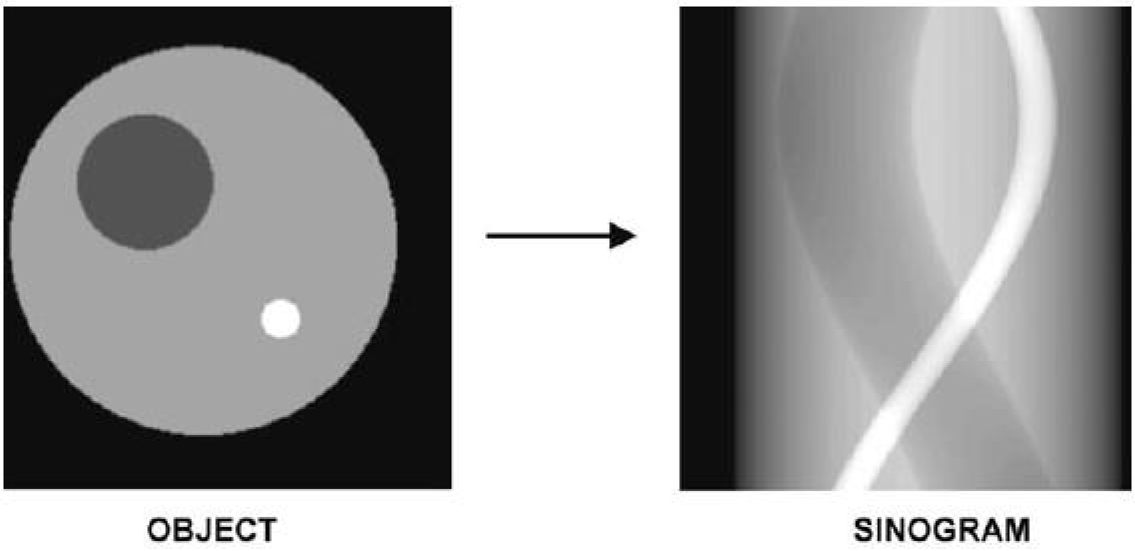


Figure 2.3 - Sinogram representation for a simple object that would result from generating projection views around it. Used with permission from [8]

2.4 Whole-body vs Organ-targeted PET

When dealing with ^{18}F-FDG PET examinations involving systemic injection of radiopharmaceuticals into the patient, the issue of radiation exposure is becoming one of the most important aspects limiting molecular imaging procedures for use in undiagnosed patients and the pediatric population. Significant dose reduction can be achieved with the development and clinical use of PET technology with an increased sensitivity capable to work at a fraction of the standard FDG dose. Improvement in sensitivity is possible with dedicated PET scanners with optimized geometry, that attains the highest possible angular coverage of the dedicated organ. This provides more efficient gamma-ray detection than with WB ring detector design permitting a lower dosage of the radiotracer.

Other advantages of dedicated systems over WB PET scanners include reduced signal from elsewhere in the body that improves image contrast recovery and further increases the sensitivity in the organ of interest; lower cost; and higher spatial resolution allowing small lesion detection. Indeed, use of small field-of-view (FOV) organ-targeted PET detectors rather than large WB rings allow the use of finer pitch segmentation of the scintillator (used to convert gamma-photons into light, which is subsequently converted into a measurable signal by front-end electronics) without increasing the number of elements. This helps to achieve higher spatial resolution due to reduced light dispersion. Table 2.1 summarizes major advantages and disadvantages of WB and organ-targeted PET technologies.

Table 2.1 - Organ-targeted vs Whole-body PET

Modality	Advantages	Disadvantages
Whole-body PET	● Whole-body coverage	● Higher radiation dosage ● Lower sensitivity ● Higher cost
Organ-targeted PET	● Better sensitivity ● Better image quality ● Lower cost of the system ● Close proximity to the organ under study ● Use cases related to assisting during other medical procedures (such as biopsy, surgery, etc...)	● Single organ examination ● Dedicated detector geometry design and technology development required

Organ-targeted systems emerge for imaging in breast, brain, heart, and prostate [2]. However, widespread clinical adaptation of the organ-targeted technology requires advances in data acquisition and image reconstruction techniques to drive improvements in PET diagnostic capabilities and reduce the dose associated with PET imaging.

2.4.1 Positron Emission Mammography

One of the most successful clinical applications for organ-targeted PET nowadays is in breast cancer detection and treatment planning. A special term was created for breast-dedicated PET systems - Positron Emission Mammography (PEM). A recent summary of peer-reviewed studies illustrates the ability of PEM to reliably detect early forms of cancer with high specificity independent of breast density [9] and hormonal changes. PEM has been shown to have a higher sensitivity than MRI for the smallest cancers (in part since it is not angiogenesis-dependent) and can be used to screen high-risk women at any age [9] [10].

Clinical PEM has come a long way: from its invention and the first feasibility studies in 1994 [11]–[13] to the first commercially available PEM scanner (PEM Flex, CMR Naviscan Corporation, San Diego, CA), which was declared a "Leader in the Future of Molecular Breast Imaging" by Frost & Sullivan [14]. The success of the Naviscan PEM system is in its pioneering breast-dedicated, compact design, as opposed to general purpose, large footprint, WB PET systems. Figure 2.4 (left) shows the commercially available Naviscan PEM-FLEX Solo II system with two planar PET detector heads,

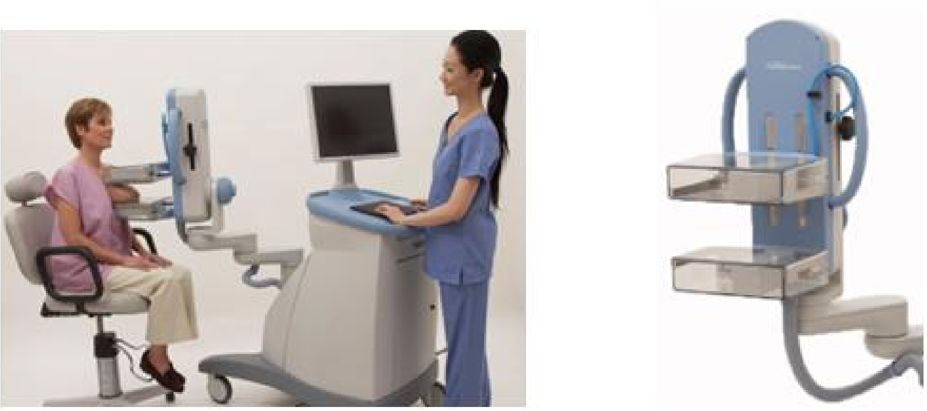


Figure 2.4 - Left: Naviscan PEM-FLEX Solo; Right: Naviscan's two detector heads with strip detectors inside

positioned on both sides of the slightly compressed breast. In comparison with WB PET scanners, the detector heads are located much closer to the imaged organ, thus improving the sensitivity in the area under investigation. As this is evident from Figure 2.4 (left), the breast-dedicated design makes gamma-ray detection more efficient since a greater solid angle is subtended and the collection efficiency is improved. In addition, Figure 2.4 (left) demonstrates that because the detector heads are placed around the breast, they only detect signal from that part of the body, thereby increasing the signal-to-noise ratio and improving image contrast. Because of finer pitch segmentation of the detector components spatial resolution and detection of smaller lesion are better than in current WB systems. The Naviscan system can detect lesions with resolution down to around 1.6 mm. Figure 2.4 (right) shows Naviscan's two detector heads; as shown, each head consists of a clear casing containing a gamma-photon sensor in the form of a strip. Images are acquired by scanning the two strip sensors simultaneously within these casings across the breast. Each strip sensor employs scintillators (to convert gamma photons into visible light) optically coupled to vacuum photomultiplier tubes (PMTs) (to convert scintillation light into a measurable electrical signal). PMT-based technology does not allow the construction of a planar detector head large enough to cover the entire breast; this in turn requires the use of a scanning technique, which reduces the amount of time that the sensors are exposed to a particular breast region. Since moving detectors collect less of the injected radiation at a given time, this method results in longer acquisition times, decreased sensitivity and higher dose exposures. As a result, Naviscan technology limits the clinical use of PEM to cancer staging and treatment follow up. Also, scanning detectors require precise coordination between the movements of the two detectors or motion artifacts will be introduced into the image.

Overall, widespread PEM application requires significant improvement in PEM technology to reduce radiation dose during PEM imaging (while keeping high spatial resolution) and to suppress radiation-induced cancer risk to that of digital mammography [15], [16].

Typically, breast imaging protocols with a dedicated PET system (including Naviscan PEM-FLEX Solo) require an injection of 370 MBq (10 mCi) of ^{18}F-FDG [17], resulting in an effective dose to the breast of 3.4 mGy and an effective whole-body dose of up to 6.2–7.1 mSv [15], [18]. This effective dose is more than 10 times the average effective dose of 0.5 mSv for digital mammography [15], [18] and poses a significant risk of radiation-induced cancer from annual PET scans. From radiation-induced cancer risk standpoint, the activity of the administered ^{18}F-FDG must be reduced to 70 MBq or less for a nuclear medicine scanner to be considered for applications in screening or in other procedures that involves undiagnosed patients [18]. These issues are addressed in the Radialis organ-targeted PET technology reported below.

2.5 PET Image Reconstruction

The goal of PET studies is to produce a quantitatively accurate image visualizing radiotracer distribution in an object of interest from which diagnostic decision or computable characteristics can be found. From high-level application point of view, a PET system works in 3 steps (refer to Figure 2.5).

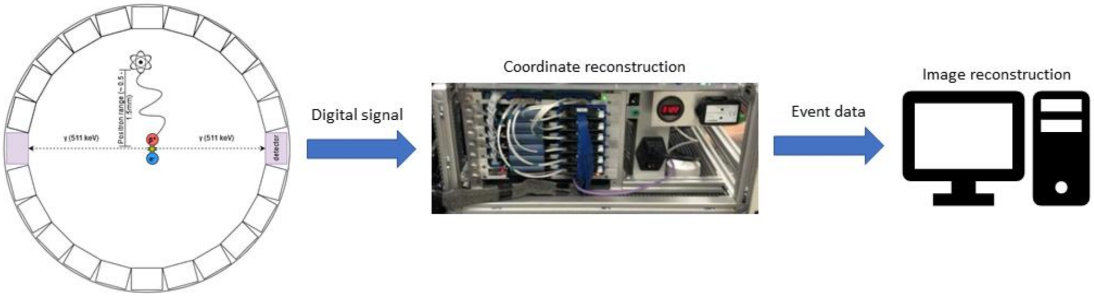


Figure 2.5 - High level overview of PET work process

In this work, we are focusing on the image reconstruction part of the PET process. Once a coincidence is detected a dedicated electronics unit computes actual coordinate of the coincidence and saves them for future reconstruction either in the form of list mode data or sinogram. To get an image a further processing of the coincidence coordinates data are needed. The role of image reconstruction is to convert these measurements into a 2-D or 3-D image that would quantitatively correspond to the distribution of the radiotracer. There are two primary approaches to image reconstruction: mathematical and statistical. Mathematical algorithms aim to utilize mathematics of the PET process and find a mathematical connection that would relate measured data and activity distribution. They are often called “Analytical”, common examples are filtered back projection (FBP) and Fourier reconstruction. The statistical approach uses iterative methods to find the most accurate image representing the measured data. A short summary for both approaches is present below, more detailed discussion on reconstruction methods in PET can be found in [19].

Analytical image reconstruction for PET has the advantage of less computational burden on the system thus generating resulting images faster as well as simpler implementation relying on well-known mathematical constructs. The most popular algorithm in this group is FBP that has gained a lot of popularity for various tomographic reconstructions and is still often used in a clinical setting. However, there are a couple of limitations related to the inability to account for various degrading factors present in any PET scanner such as photon scattering, positron range, etc. In addition, statistical characteristics of the acquired data are not taken into consideration. Some of the shortcomings can be improved using a reconstruction filter; however, such methods normally result in decreasing the overall image resolution.

2.5.1 Iterative Solutions

The second common approach to PET image reconstruction are iterative reconstruction methods. In the first step we need to make an initial guess for the radioactivity distribution. Then we compare the measured distribution with the initial guess using an appropriate cost function and update the initial guess based on the result of the comparison using a dedicated update function. Then the process is repeated until our estimate does not match closely the true image. Due to limits imposed on the data quality usually it is very hard to reach the exact true distribution and in practice iterative algorithms have a stop condition based on the number of iterations to run. Most of the iterative solutions differ in either the cost function, the update function or both. The most widely used approach is MLEM method proposed in [20]. Nowadays, iterative algorithms have become a gold standard for clinical image reconstruction in PET systems [21], [22]. A common limitation with iterative reconstruction is their higher computational intensity compared to analytical solutions. This is caused by the repetitive nature of the computations and the need to perform forward and back projection operations in each iteration.

2.5.2 Maximum Likelihood Expectation Maximization (MLEM)

Among different iterative algorithms for PET image reconstruction, the MLEM and its faster approximation version, the OSEM [23], stand out as the most common solution for clinical and research systems. Maximum likelihood is a statistical function which is maximized when we minimize the difference between the measured and estimated data. The Expectation Maximization (EM) algorithm maximizes likelihood using a Poisson data model [24]. It assumes that the measured data has a Poisson distribution identified by the counting statistics in each LOR hence considers the statistical noise in the data. Each iteration update of the MLEM algorithm for PET can be written as:

$$\lambda_j^{(k+1)} = \frac{\lambda_j^k}{\sum_i A_{i,j}} \sum_i A_{ij} \frac{y_i}{\sum_i A_{ij} \lambda_j^k} \tag{2.1}$$

where the voxel value λ in a voxel j for the iteration $k+1$ is calculated using the previous voxel value λ_j^k, measured projection data y_i, and the matrix A_{ij} that corresponds to the probability that a positron emission in a voxel j will be detected in the projection i. Matrix A is often referred to as System Matrix (SM) or System Response Matrix (SRM). The algorithm works in the following way: the estimate image λ^k is forward projected into the projection space. Then the ratio between measured and estimated projections is calculated to see how well the estimate describes the measured data. This ratio is back projected back to the image space and is known as a correction term. Finally, the correction term is multiplied by the estimate (λ_j^k) and normalized by the sensitivity matrix $\sum_i A_{i,j}$ generating a new estimate $\lambda_j^{(k+1)}$. These steps are repeated until either the algorithm converges, or the stop condition is met. It is important to mention that the initial estimate $\lambda^0$is required to start computations. Normally a uniform image with values > 0 is selected as an initial estimate. Figure 2.6 shows a basic flowchart describing internal work of the MLEM algorithm.

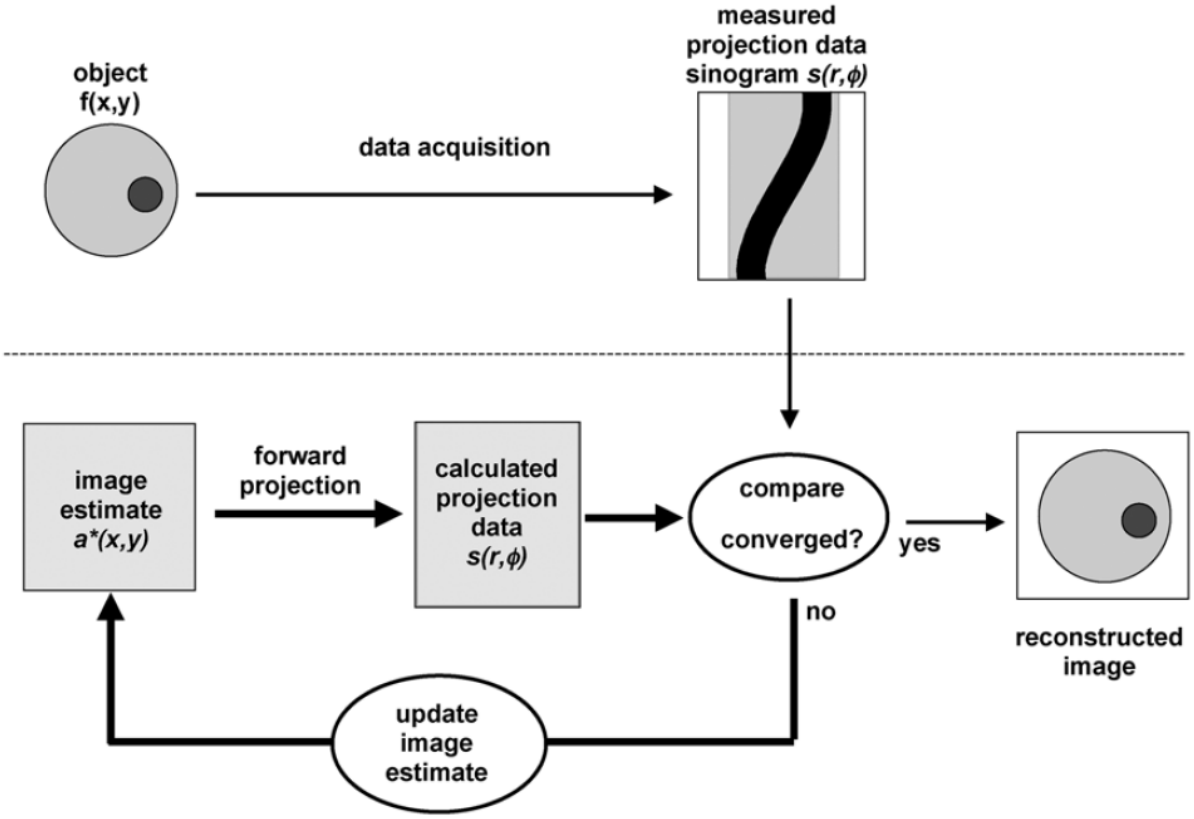


Figure 2.6 - Iterative PET reconstruction flow chart. Used with permission from [8]

There are a couple of limitations that had to be addressed for the MLEM algorithm. First, the resulting images end up being rather noisy due to the bad conditioning of the problem [20]. Therefore, in practice the algorithm is stopped before convergence, different stopping rules were developed [25], [26]. Another solution to the image noise might be applying a smoothing filter [27]. Long reconstruction time of the MLEM algorithm due to high computational complexity and slow convergence rates (normally MLEM requires 30 - 100 iterations for a practical result) has led to a lot of research on possible optimizations.

For example, the OSEM algorithm[23] is an acceleration of the MLEM that uses only a subset of measured data in each iteration. OSEM can be described as:

$$\lambda_j^{(k+1,b)} = \frac{\lambda_j^{(k,b-1)}}{\sum_{i \in S_b} A_{i,j}} \sum_{i \in S_b} A_{ij} \frac{y_i}{\sum_i A_{ij} \lambda_j^k} \tag{2.2}$$

where B denotes the number of subsets, and S_b is the current subset in processing. As the result image updates are happening roughly B times more often than in the traditional MLEM algorithm. It is an important condition to carefully choose B size and good data sampling is required. One downside of the OSEM algorithm is that it does not guarantee convergence into an ML solution.

Overall, in order to make computational complexity of MLEM feasible for the clinical devices a shift from Central Processing Unit (CPU) to Graphics Processing Unit (GPU) architecture had to happen. What made PET image reconstruction a perfect candidate for General-Purpose computing on Graphics Processing Units (GPGPU) is the easily parallelizable and the most time-consuming operations of forward and back projection that are repeated on every iteration of MLEM. It turned out that they can be implemented extremely efficiently on the GPU architecture [28]. First attempts in adopting GPU for the iterative image reconstruction for PET started as early as 2003 by accelerating OSEM algorithm [29]. With the development of the GPU hardware a numerous development in GPU-based iterative PET image reconstruction were made. Transition to GPU allowed researchers not only to significantly improve reconstruction speed, sometimes up to 200 times [30], but also to investigate and adopt more sophisticated physical models such as shift variant point spread function (PSF) [31], made use symmetry characteristics of imaging systems [32] and even attempts on image guided intervention using a PEM system [33].

2.5.3 General-purpose Computing on Graphics Processing Units

Graphics processing unit is a great example of a massively parallel architecture. GPUs were specifically designed to handle image processing and 3-D graphics related computations. With the advancement of GPU technology, it was soon realized that the same hardware can be used efficiently to for many scientific computational problems that fall under Single Instruction Multiple Data (SIMD) or similar Single Instruction Multiple Thread (SIMT) paradigm where the same set of instructions is executed on different data inputs or by different threads. The usage of GPUs for general computations was later named as GPGPU. At first, general-purpose usage of the GPU hardware required problems to be rewritten using graphics primitives. Luckily, everything changed when Compute Unified Device Architecture [34] (CUDA) and OpenCL frameworks were developed to simplify problem reformulation and accessibility of GPUS as general-purpose processing units. CUDA is now by far the most popular framework for GPU programming, it has a downside of only working with NVIDIA GPUs. However, the great thing about CUDA code is once it is written it can be utilized on the future GPUs without requiring code reformulation thus inherently taking advantage of the developing technology.

Computing on GPUs has been successfully applied in medical imaging for different applications including CT, MRI, PET [35] to name a few. What makes GPGPU so powerful are some architectural choices made along the way of development of the hardware technology.

Figure 2.7 visualizes architectural differences between CPU and GPU. The GPU emphasis on parallel processing is clearly visible comparing the number of cores or arithmetic logic units (ALU) between CPU and GPU. This allows GPUs to run order of magnitude more threads than a typical CPU. It is also clear that each individual CPU core has more control over its execution and memory resources available. A high-level comparison between features of CPU vs GPU architectures is present in the Table 2.2. It becomes apparent that CPUs win in terms of programmability and single core processing,

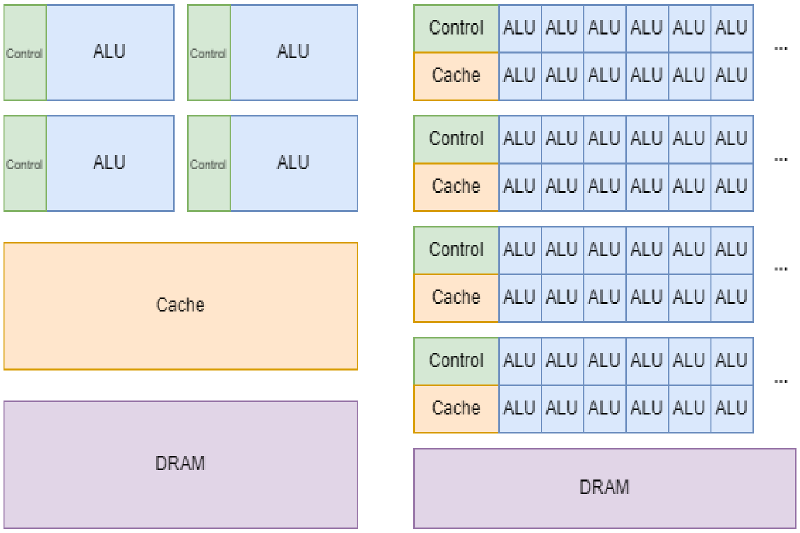


Figure 2.7 - CPU vs GPU architecture

and that the key difference is how GPU makes use of high number of available cores. In our work we are utilizing CUDA framework for GPU programming and therefore further descriptions will closely relate to NVIDIA GPU programming approach, even though basic principles will remain the same and would not depend on a particular GPU manufacturer.

Table 2.2 - CPU vs GPU architecture comparison

CPU	**GPU**
Tens of cores in parallel	Thousands of cores in parallel
Good programmability	Bad programmability (but there are tools: CUDA, OpenCL)
Extremely fast single thread	Single thread is on a slower side and not very capable
Very well optimized	Requires a lot of manual optimizations

2.5.4 CUDA Thread & Memory Hierarchy

CUDA is a GPU parallel computing platform developed by NVIDIA to provide a C/C++ like language extension with an appropriate application programming interface (API) for the GPU hardware. It is important to mention here a couple of key concepts that are natural for CUDA programming model. A GPU can run thousands of threads, these threads are organized in blocks and blocks are executed in parallel by the Streaming Multiprocessors (SM). All threads within a block share limited resource, there is a cap of maximum 1024 threads per block. There is another useful abstraction that organizes blocks on the programmer’s side which is grid. Figure 2.8 summarizes thread hierarchy in CUDA.

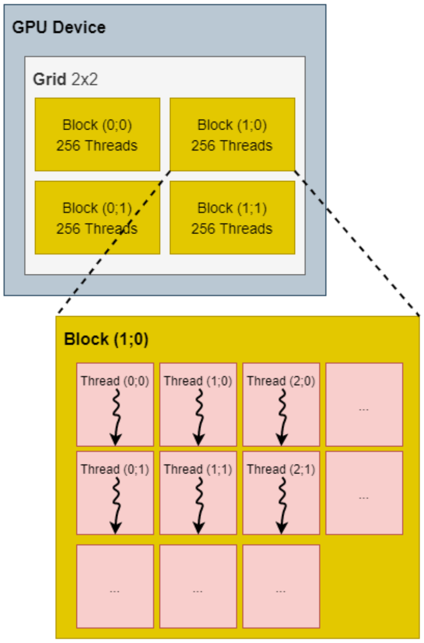


Figure 2.8 - GPU thread hierarchy

The second important resource on the GPU is memory, there are 5 primary memory types summarized in the Table 2.3.

Table 2.3 - GPU Memory Hierarchy

Memory type	Scope	Speed
Register	Thread	Very Fast
Local	Thread	Slow
Shared	Block	Fast
Global	Grid	Slow
Constant	Grid	Fast Read

Memory usage is a common place for optimization since speed for different memory types vary significantly. The fastest memory available is register memory, however it's characterized by very limited size and programmers do not have control over register memory allocation. Second fastest memory is shared memory it is physically located on-chip that allows it to have high bandwidth and extremely low latency. The size of the shared memory is configurable up to 96 KB per block on the latest GPU devices. Shared memory is shared between every thread running within a common block. Main bulk of GPU memory is also referred to as global memory is the place where any input data is copied from the host computer. Global memory is much slower than the shared memory and register memory therefore we aim to maximize coalescing of data access operations. Local memory is just an abstraction on top of the global memory that is only accessible on a single thread level. Finally, constant memory resides together with the global memory but provides a faster read access because it is cached in a constant cache.

3 Radialis PET - State-of-the-art Clinical Organ-Targeted System

After covering general PET information as well as making an introduction to the image reconstruction algorithms used in PET, I would like to introduce the Radialis PET Camera as a state-of-the-art clinical organ-targeted PET system.

3.1 System Characteristics

Dr. Reznik and her research team invented and patented a novel solid-state technology for organ-targeted PET with improved sensitivity, capable of significant dose reduction (factor of 10) in comparison to commercial WB PET scanners. The improvement was made possible by the development of a new type of four-side tillable sensor modules shown in Figure 3.1 (left). Each module employs an advanced scintillation crystal, Cerium doped Lutetium Yttrium Orthosilicate (LYSO), in combination with a matching array of high-gain photodetectors - Silicon Photomultipliers (SiPMs). LYSO and SiPMs are optically coupled through a specially designed light guide with slightly slanted edges so that the front face and the back face of the light guide have the exact dimensions of the scintillating crystal array and the photodetector layer, respectively. All the modules' components and front-end electronics are mounted in such a way that none of the components is larger than a specially designed scintillating crystal. As a result, the scintillating crystal has an overhang over the photosensor array to maintain 100% tileability: all four sides of the modules are tiled against each other so that they can be seamlessly combined into a sensor area of the needed size. Figure 3.1 (right) shows an

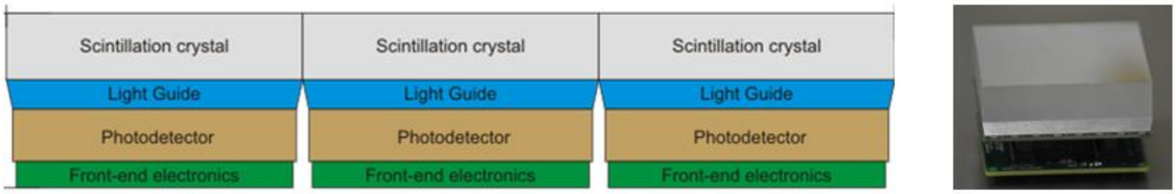


Figure 3.1 - Left: Schematic presentation of the cross-section of three tiled detector blocks; Right: the photo of a block detector with an electronic board underneath.

actual detector block with the slanted-edge light-guide mounted onto a PCB board with the front-end electronics.

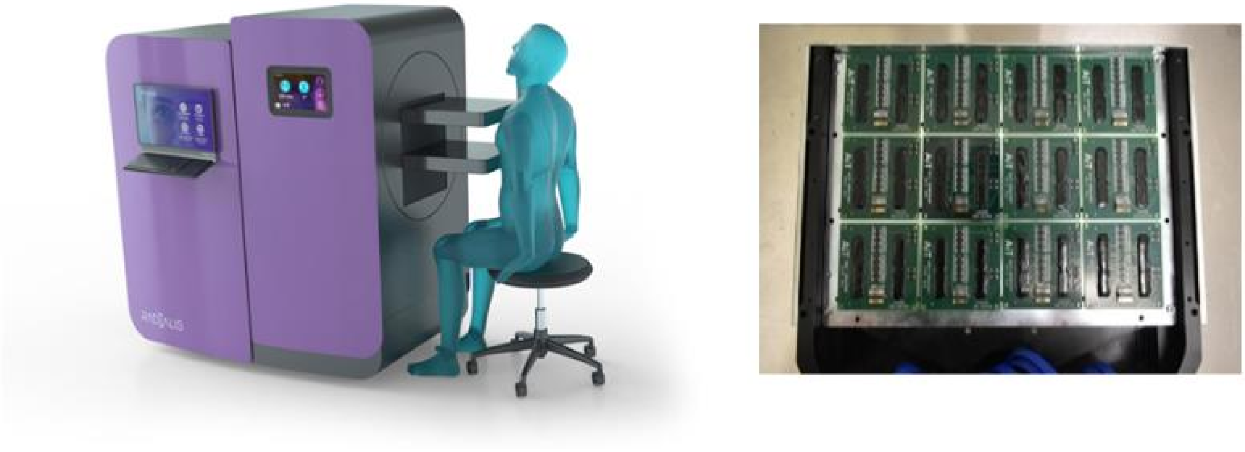


Figure 3.2 - Left: Clinical prototype of PEM camera with two planar detector heads; Right: 3×4 array of sensor modules inside a detector head (bottom).

The technology is commercialized by Radialis, Thunder Bay, Ontario, and is now called Radialis organ-targeted PET camera. The first application of the Radialis organ-targeted PET camera is in a breast imaging. The breast-targeted design is similar to that of PEM (Figure 3.2 left). For breast imaging, 12 sensor modules are arranged in a 4x3 array to make a planar detector head (Figure 3.2 right). For image acquisition, two detector heads are positioned on either side of the immobilized breast. The size of an individual module is 58×58 mm2, which results in a sensor area of 232x174 mm2 that provides a sufficient field-of-view (FOV) to cover the entire breast. The developed breast-targeted system has been tested with a comprehensive set of standardized experiments outlined in the NEMA NU-4 standards [36]. These tests demonstrated an in-plane spatial resolution of 2.3 ± 0.1 mm that is comparable to best-in-class organ-targeted PET scanners, combined with a peak sensitivity of 3.5%, which is much better than the sensitivity of any WB or organ-targeted PET scanner on the market or in clinical trials.

Another parameter that characterizes the efficiency of activity detection in PET imaging is Noise Equivalent Count Rate (NECR). NECR describes the true coincidence rate that would give the observed signal to noise ratio (SNR), or the same level of statistical noise, if there were no randoms and no scattered events. NECR performance is normally evaluated over a clinically relevant activity range using a specialised phantom filled with ^{18}F solution, and efficiency at peak noise equivalent count rate is determined as the peak NECR normalized to the activity at the peak. Table 3.1 presents the efficiency at peak count rate for several PET systems including organ-targeted, whole-body, and total-body systems. It can be seen that the Radialis PET camera exhibits much higher efficiency at peak count rate when compared to current WB systems. The SiPM-based total body PET technology of uExplorer also provides superior sensitivity in comparison to the WB systems, achieved with detectors that completely cover the axial length of a patient's body. Radialis' SiPM-based organ-targeted technology uses the same approach: the coverage of the Radialis system is larger than the organ being imaged and improves the sensitivity from the increased axial extent of the detectors.

Table 3.1 - Values for efficiency at peak count rate are calculated from the peak NECR data reported for each system.

PET System	Efficiency at Peak Count Rate (cps/MBq)	Peak NECR (kcps)	Concentration at Peak NECR (kBq/mL)	Phantom Volume (mL)	Activity at Peak NECR (MBq)
Radialis PET Camera (NU-4)	5,650	17.8	10.5	300	3.15
uExplorer [37] (NU-2) (Total Body)	3,790	1440	16.8	22,600	380
Oncovision Mammi PEM Dual Ring (NU-4) [38] (PEM)	1,260	34.0	31.2	866	27.0
GE Discovery IQ [39] (PET/CT)	618	123.6	9.1	22,000	200
GE Discovery MI (NU-2) [40] (PET/CT)	581	266	20.8	22,000	458
Phillips Vereos (NU-2) [41](PET/CT)	556	646	52.8	22,000	1,160
GE Signa PET [42] (PET/MR)	524	218	17.8	22,600	402

Siemens Biograph Vision (NU-2) [43] (PET/CT)	435	306	32	22,000	704
Naviscan PEM Flex Solo II [44] (NU-4) (PEM)	393	10.6	90	300	27.0

In the clinical setting, the higher sensitivity, and the fact that the counting rate peaks at relatively low activity values allow the activity of the administered radiopharmaceutical to be reduced. In fact, clinical demonstration with imaging in breast revealed that the Radialis PET technology is well-suited to identifying cancers even at a 10-fold dose reduction in comparison with standard WB PET dose (see the selected results from the clinical trials (ClinicalTrials.gov ID: NCT03520218) at the UHN-PMCC in Figure 3.3). An injection of 37 MBq (1 mCi) of ^{18}F-FDG corresponds to an effective radiation dose of less than 1 mSv, which is estimated to be equivalent to the effective dose from the standard breast screening procedures with Full-Field Digital Mammography (FFDM) or Digital Breast Tomosynthesis (DBT) [18].

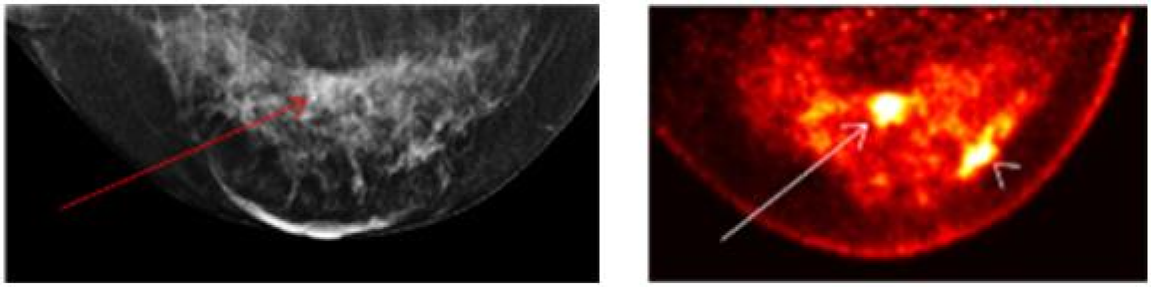


Figure 3.3 - Images of the same breasts. Left: full-field digital mammography (FFDM); B: Radialis PEM image acquired with 37 MBq of activity that is 10 times lower than the standard dose used in WB PET imaging. The focal uptakes on PEM image (arrow and arrowhead in right image) corresponds to one mass (arrow in left image) detected on FFDM, however the other mass that was also histopathology proven was detected only in PEM images despite using a low dose [45].

The demonstrated capability for imaging at low dose of administered radiotracer suggests that Radialis organ-targeted PET technology could be used in low-dose clinical applications such as breast and prostate cancer screening, the multiple examinations required for prostate cancer patients on active surveillance, and neuro degenerative examinations. High-quality organ-targeted imaging may also be particularly well-suited to applications with emerging targeted radiotracers.

3.2 Radialis Imaging Process

Being a clinical system brings more complexity into Radialis PET' imaging process. The Figure 3.4 provides a high-level overview of all the steps involved in generating a clinical image. The operation starts with a coincidence detection; a dedicated electronics unit is responsible for detector signal readout and event/coincidence coordinates reconstruction. Once coordinates for the coincidence are reconstructed, they are being transferred, along with energy and time information, to the main computer via UPD connection in a binary format. More in-depth discussion of the acquisition process and data type generated by electronics is out of scope of this work. During a patient scan the main computer is constantly monitoring the UDP connection and saving all the incoming data. As soon as the amount of saved data passes a certain threshold the system starts to generate a series of preview images (top branch on the Figure 3.4) using a small subset of events acquired thus far in order to ensure correctness of patient's position and monitor any patient' movements during scan. Once the acquisition is over (a typical acquisition duration is 5 – 10 minutes) the main computer stops saving the coincidence information to disk. Immediately after that a separate pipeline for image reconstruction is started in the background (bottom branch on the Figure 3.4).

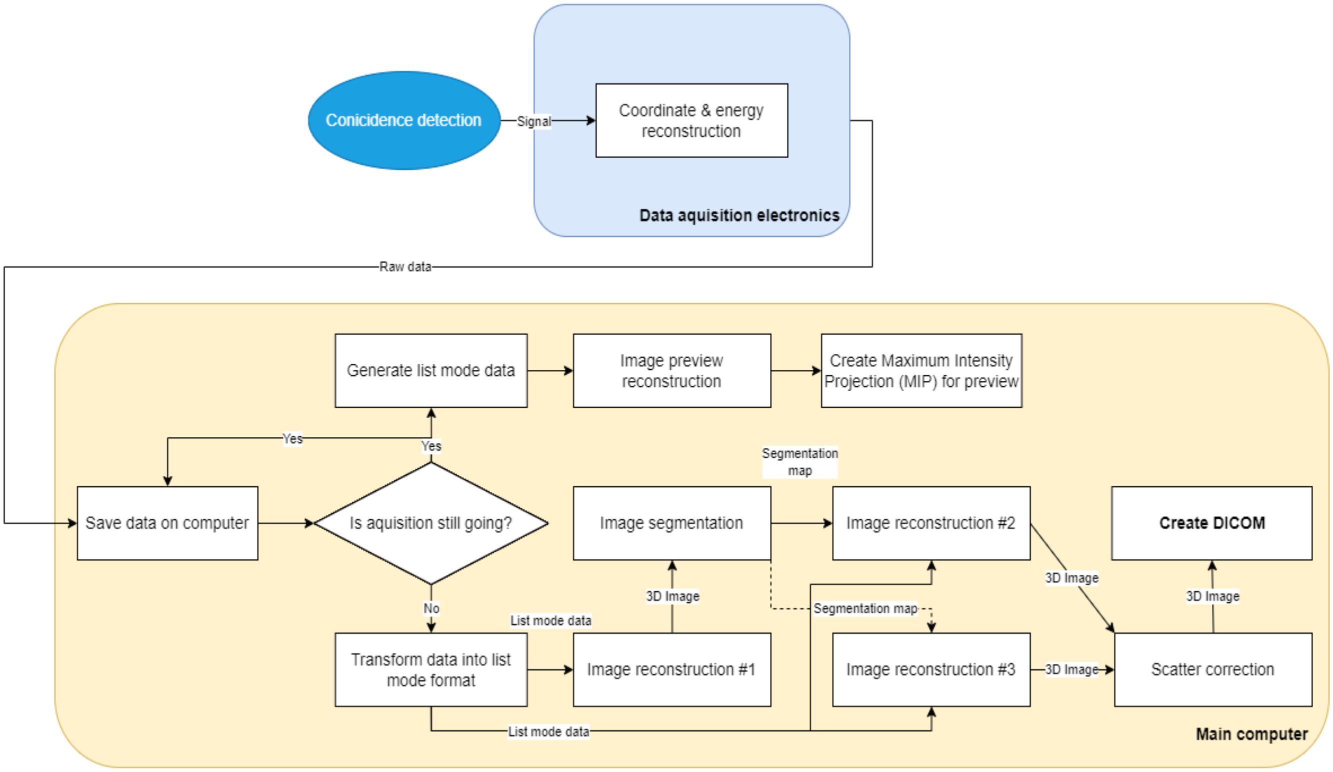


Figure 3.4 - Hight level overview of the imaging process used by Radialis PET.

It starts with transforming the raw data into the list mode format suitable for reconstruction. Then, the first image reconstruction is made to get preliminary image that is later used in the attenuation and scatter corrections. Scatter correction is needed to reduce image blur while attenuation correction (AC) is required for quantification of PET images in terms of an accurate correlation between image counts and true tissue activity. Before a Digital Imaging and Communications in Medicine (DICOM) file can be generated the system needs to perform image segmentation map and use it to apply attenuation correction during 2 more reconstruction runs one with upper energy window and one with regular parameters. Then the results of the reconstructions can be combined to apply scatter correction. Finally, a 3D image is converted into a DICOM file which ultimately is the result of the procedure. Each image reconstruction is a separate 15 iteration Radialis MLEM run. Such approach yields great results for clinical image quality; however, it makes any inefficiencies in the image reconstruction to propagate 3 times affecting the overall performance.

A single patient scan consists of multiple data acquisitions for different detector positions. For each acquisition a separate 3D image is generated. This and the need to generate preview images would not allow to allocate all the system's resources to the image reconstruction task. Image reconstruction is by far the most time-consuming part of the imaging process. Ideally the time needed to reconstruct an image should not be longer than acquisition, so it is very important to make sure that reconstruction is fast. Currently, image reconstruction time by far exceeds image acquisition time.

3.3 List Mode Reconstruction

The traditional MLEM algorithm had to be reworked to be suitable for the data in list mode format that differs from projection data of sinograms that the MLEM was initially designed to work with. The required modifications were straightforward [46], [47]. The updated MLEM algorithm can be formulated by replacing y_i in Equation 2.1 with the sum over all recorded events (see Equation 3.1).

$$\lambda_j^{(k+1)} = \frac{\lambda_j^k}{\sum_i A_{i,j}} \sum_{w \in events} A_{i_w j} \frac{1}{\sum_j A_{i_w j} \lambda_j^k} \tag{3.1}$$

The main difference with the sinogram based MLEM is that instead of evaluating the sum on the right-hand side over the whole set of projections we do it only for the list of measured coincidences. Thus, in practice we no longer need to process every possible LOR in the FOV that comes handy for high-resolution systems. In case of Radialis PET list mode reconstruction was beneficial also due to a limited angle coverage of the dual-head planar detectors. If we were to construct sinograms based on measured data, we would not be able to cover every angle therefore leaving parts of it empty. Additional advantages of the list mode data were discussed in the Data Format section.

A highly optimized CPU-based implementation of the List mode MLEM was developed. Some of the key features included: computation of the SM on-the-fly, single pass through the list of events, Forward projection and Back projection combined within the main events processing loop for a single iteration. Due to the nature of the algorithm each event could be processed independently. OpenMP API was used to utilize multiple CPU cores for the events processing part of the algorithm.

Apart from OpenMP for computations parallelization, OSEM algorithm was considered at first; but after comparing potential gains from OSEM and GPU-based implementation approaches, we decided to focus on the GPU-based approach with an ability to extend MLEM implementation to OSEM in case further speedup would be needed. The potential gains from OSEM are normally 5-10 times faster than MLEM counterparts whereas GPU-based solutions demonstrate 100+ times speedup for image reconstruction speed. In addition, the implementation of OSEM would have required additional elaborate image quality testing to ensure that clinical system's performance would remain intact.

To improve image quality and quantitative accuracy of radioactivity distribution determination in reconstructed PET images two data corrections are applied: attenuation and scatter as mentioned in Section 3.2.

It is typical to use anatomical imaging (i.e., CT transmission scan data) to estimate attenuation of photons in WB PET/CT [48]. However, being a standalone PET device, we do not have access to the anatomical data, so an alternative approach is used. Attenuation correction in Radialis PET is similar to the one proposed in [49]. Knowing which organ is under study we know an approximation of the attenuation coefficient in tissue of the given organ. Then we build a segmentation map based on the preliminary reconstruction (See Figure 3.4). The map contains linear attenuation coefficients that are used in the next reconstructions to adjust weight for each LOR.

The scatter correction is performed by the Estimation of Trues Method (ETM). This method was proposed in [50]. The general concept behind this method is that the percentage of scattered events in high energy window is negligible relative to this percentage in the standard energy window. We implement it by first reconstructing an image with standard energy filter (0.35 MeV – 0.7 MeV), then the second image is reconstructed but only considering events with higher energy (0.5 MeV – 0.7 MeV). Second image is scaled to match the number of unscattered events in the first image. Scaled second image is subtracted from the first image, essentially leaving only the image of scattered events. Scattered events image is smoothed and subtracted from the first image.

The result of the image reconstruction is a so-called raw image containing the 3D distribution of calculated voxel values. The final step is to convert such 3D image into a DICOM file that can be assessed by medical professionals.

3.3.1 Image Examples

The following images in Figure 3.5 represent various steps of the Radialis imaging process. The data was acquired and reconstructed using Radialis PET camera. A series of preview images (Figure 3.5 top-right) are generated during image acquisition. Then a basic reconstruction with no corrections is applied (Figure 3.5 bottom-right). Finally, once both attenuation and scatter corrections are applied a final image is reconstructed (Figure 3.5 bottom-left).

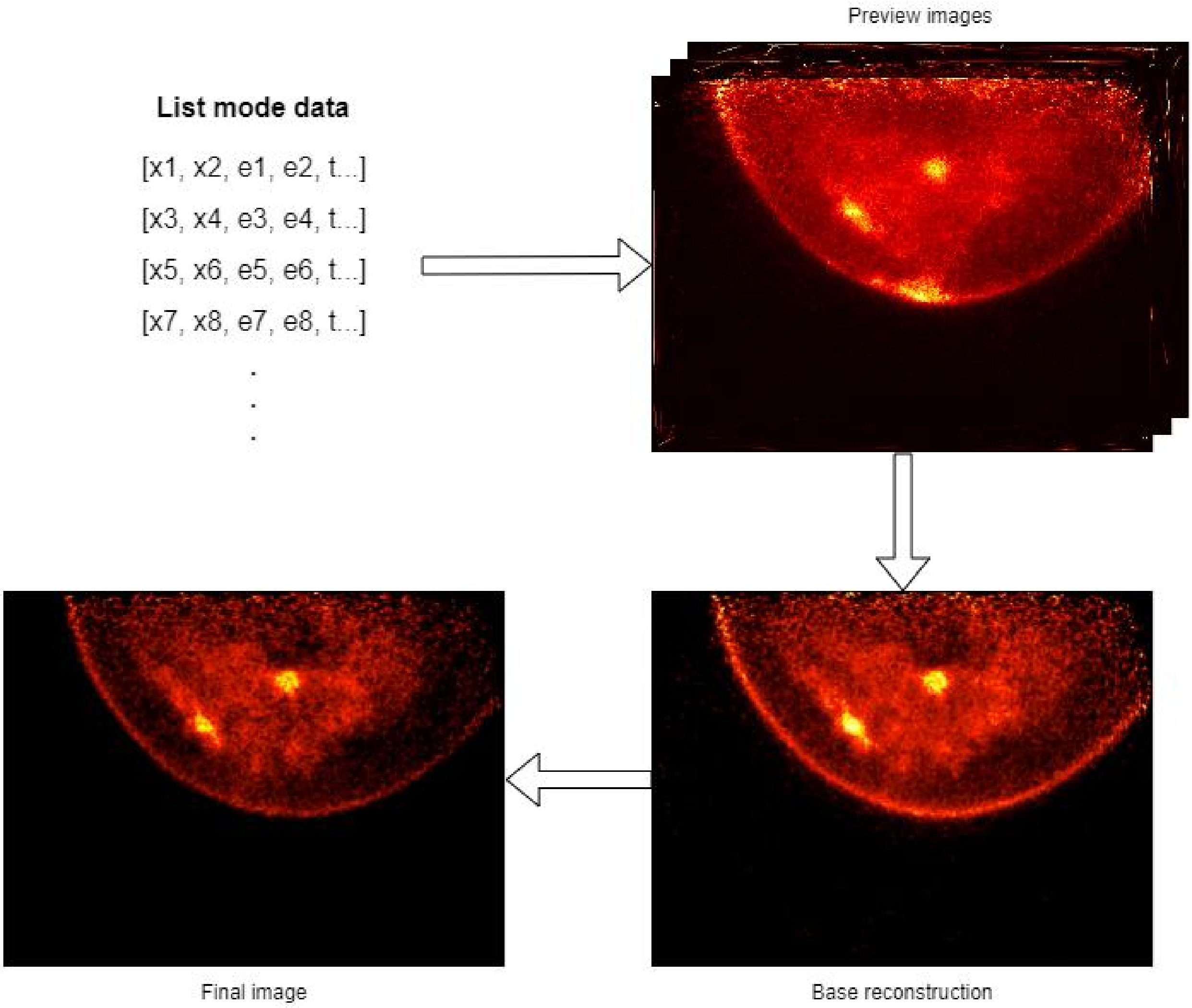


Figure 3.5 - Different steps in Radialis imaging process

3.4 Problem Formulation

Organ-targeted PET is novel in its clinical implementation, and ensuring widespread adoption requires addressing several key issues. These include device performance in terms of sensitivity, spatial resolution and image quality, standardization of image quality quantification methods, clinical trial protocol development and adjustment for lower dose imaging, high patient throughput requirements. Radialis PET is a clinical organ-targeted PET system that successfully addresses the needs for imaging performance and sensitivity to radiotracers. There are two key components which determine the rate of patient throughput: acquisition duration and image reconstruction speed. Radialis PET was able to successfully lower the acquisition time 3 - 6 times (compared to WB systems) by improving overall system sensitivity, however image reconstruction speed remains less than desirable. At present, an average image reconstruction takes between 20 minutes to a couple of hours to generate an image, which is significantly more than the acquisition time of 5-10 minutes. The goal of this work is to address the issue of image reconstruction speed by identifying bottlenecks in the current implementation and to develop a novel formulation of highly optimized MLEM algorithm for PET image reconstruction on GPU that would be suitable for the clinical setting of the dual-head planar organ-targeted Radialis PET Camera.

4 Application of GPU-based Reconstruction for a Clinical Organ-Targeted PET Scanner

4.1 Introduction

Clinical and laboratory evaluation of the organ-targeted Radialis PET Camera indicate that it is a promising technology for high-image-quality, low-dose PET imaging. High-efficiency radiotracer detection also offers an opportunity to reduce administered doses of radiopharmaceuticals and, therefore, patient exposure to radiation. However, widespread clinical use of the Radialis PET camera imposes new requirements in terms of patient throughput. Patient throughput is directly correlated with both the acquisition duration and time needed to reconstruct an image that depends on image reconstruction speed. Higher sensitivity allows Radialis PET to successfully lower the acquisition time to about 5 min (to be compared with about 30 minutes needed to complete a WB PET/CT scan), however, currently developed CPU-based image reconstruction takes too much time to be effectively used in a clinical setting. The current reconstruction method produces desirable image quality while requiring tens of minutes and up to hours to reconstruct a typical clinical image using multiple CPU cores. A clinical scan is considered complete when both image acquisition and reconstruction have concluded, at which point the patient can be released. Therefore, image reconstruction should not take longer than the time required to acquire the image data. Minimizing image reconstruction time for improved patient throughput with the Radialis PET system is the motivating factor for our research.

An obvious candidate for the reconstruction speedup is a GPU-based approach. It has been shown [35] that reformulation of tomographic reconstruction using GPUs in nuclear medicine and medical imaging enables ultrafast image reconstruction. For PET this creates possibilities for adaptation of interactive reconstruction parameters adjustments for noise/resolution balancing [51], [52], intra-operative imaging for tumor localization [33], and even attempts at dynamic motions correction in cardiac PET [7]. And, of course, speeding up image reconstruction addresses the problem of patient throughput for clinical PET procedures. GPU usage has already become a “gold standard” for image reconstruction in PET. Our goal is to complement the advanced hardware technology with a GPU-based image reconstruction to enhance processing speed.

From a high-level view, computational work associated with forward and back projection operations (the most time-consuming component of the MLEM algorithm) is highly parallelizable as each LOR can be processed independently. Due to specifics of the GPU architecture a complete reformulation of a typical CPU-based reconstruction solutions is required to achieve maximum performance increase. There has been a lot of research in terms of MLEM algorithm adaptation for GPU, however, due to the significant number of parameters that characterize a PET system (detector geometry, data format, application domain, presence of other imaging modalities, etc.) it becomes hard to find a one-fits-all solution for a specific system's requirements. In particular, [32] proposed exploiting symmetry properties of the dual-head planar PET during GPU-based reconstruction in order to reduce system matrix (SM) size. The proposed solution works well for the sinogram data; however, it is not efficient for the sparce list mode data. Ref [30] proposed a CUDA algorithm utilizing several unique GPU properties for list-mode OSEM implementation. They exploited LOR partitioning according to predominant direction, image slice caching in fast shared memory for slice-by-slice processing of the measured events, and Tube of Response (TOR) approach to accurately model the SM. Then, [53] expanded on their work utilizing a double-GPU hardware setup. In addition, [53] improved on slice-by-slice processing by further splitting the image into cubes and processing these cubes independently on GPU using a faster resolution blurring algorithm (RBL) version of the OSEM algorithm. Both [30] and [53] were used as a foundation for the algorithm we propose in the current work, however, both works target ring detector geometry and have specific assumptions and optimizations based on this. Moreover, some of the existing CPU-based image reconstruction features must be incorporated to ensure stability of the image quality. A great emphasis was placed on further parallelization of the iterative reconstruction algorithms for multiple GPU devices [53]–[55] considering different solutions for device synchronization and iterative image update schemes. The Radialis scanner has a single NVIDIA GTX 1070 GPU available, which limits the solutions that are suitable for this scanner. However, the multiple GPU approach is a straightforward extension and can be explored for future systems.

In this chapter, we present our gradual process on reformulating a clinical 3D list-mode MLEM image reconstruction algorithm from CPU to GPU architecture using CUDA framework. The process involves step-by-step transition from CPU-based reconstruction to naïve GPU-based implementation, later expanding it with a complete reformulation to take advantage of the GPU architecture. This work is based on previously developed best practices which are optimized for the Radialis PET and aimed at improving patient throughput. We emphasize that such reformulation shall enhance the reconstruction throughput and should not result in any deterioration in clinical imaging performance. To evaluate this, the resulting GPU-based algorithm was compared to the CPU counterpart and evaluated in terms of speed, clinical image quality and system performance indicators obtained with standardized tests performed according to NEMA NU-4. It should be specified that Radialis PET is a non-ToF scanner [45], and therefore our proposed implementation does not include ToF reconstruction.

4.2 Methods & Tools

This section introduces methods and tools that are applied to reformulate MLEM reconstruction algorithm for the GPU architecture. Outlined are details on the Radialis PET scanner including its coordinate system, in addition to key GPU programming principles. The primary focus of this section is a progressive description of the algorithm reformulation process. Naïve GPU-based solution without any optimizations was developed initially as a baseline for future improvements.

4.2.1 Radialis PET Detectors

Sensor modules in Radialis PET detectors use high segmentation of the LYSO scintillator to achieve the required spatial resolution: LYSO is pixilated to make a 24 × 24 grid with each pixel being 2.32 mm × 2.32 mm × 13 mm. Schematics of the detector heads along with the adopted coordinate system is presented in Figure 4.1.

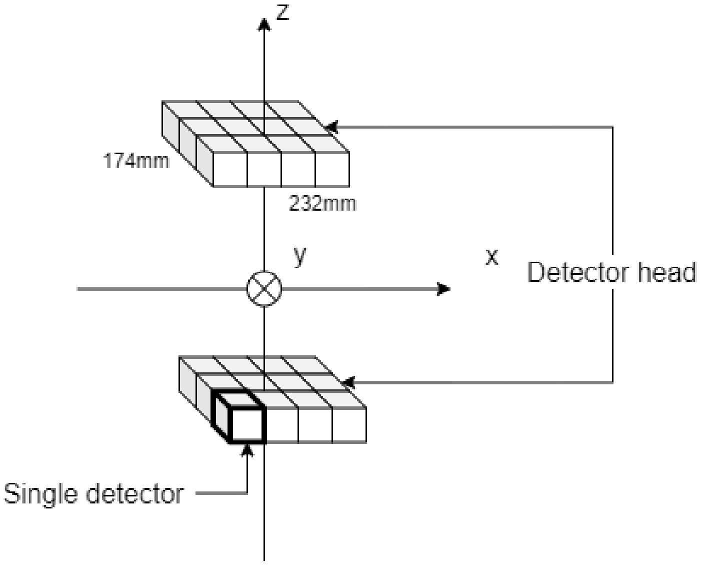


Figure 4.1 - Radialis PET detectors high-level schematics and coordinate system

The total size of the detector FOV is 232 mm × 174 mm. Dedicated acquisition electronics allow data to be saved on the primary computer as a binary file for further processing by the reconstruction software. Reconstruction generates results as a 3D DICOM image with default resolution 577 x 433 x 24 voxels. Individual voxel size is 0.4 mm x 0.4 mm and z dimension is calculated based on the detector heads separation by dividing the distance between detectors by the default number of Z slices. Radialis detectors design is described in detail in [45].

4.2.2 List Mode MLEM

The iterative list mode MLEM algorithm [46] can be summarized as

$$\lambda_j^{(k+1)} = \frac{\lambda_j^k}{\sum_i A_{i,j}} \sum_{w \in events} BP_w \frac{1}{FP_w \lambda_j^k} \tag{4.1}$$

where λ_j^k is a voxel value of voxel j of the 3D image λ at iteration k, FP and BP correspondingly represent forward and back projection operations over the list of all detected coincidences. FP computes contribution of every voxel to every LOR and BP counts contributions of every measured LOR to every voxel. The system matrix A corresponds to a distribution of probability that a particular line of response (LOR) i is detected in a voxel j. In practice, the size of the SA is significant [56]. Considering high resolution of Radialis PET it would be very computationally demanding and impractical to precompute the system matrix. Therefore, other methods for computing SM on-the-fly were proposed in [57] and [58]. Radialis PET utilizes a custom algorithm based on [59] and [57] to calculate system matrix on-the-fly. The $\frac{1}{\sum_i A_{i,j}}$ is a so-called normalization term or a sensitivity matrix. Initially, Radialis PET MLEM was implemented as highly optimized CPU-based solution. To improve the performance, the FP and BP were computed serially for each event within a single pass over the list of events. A high-level flow chart of the steps involved in CPU-based implementation is shown in Figure 4.2.

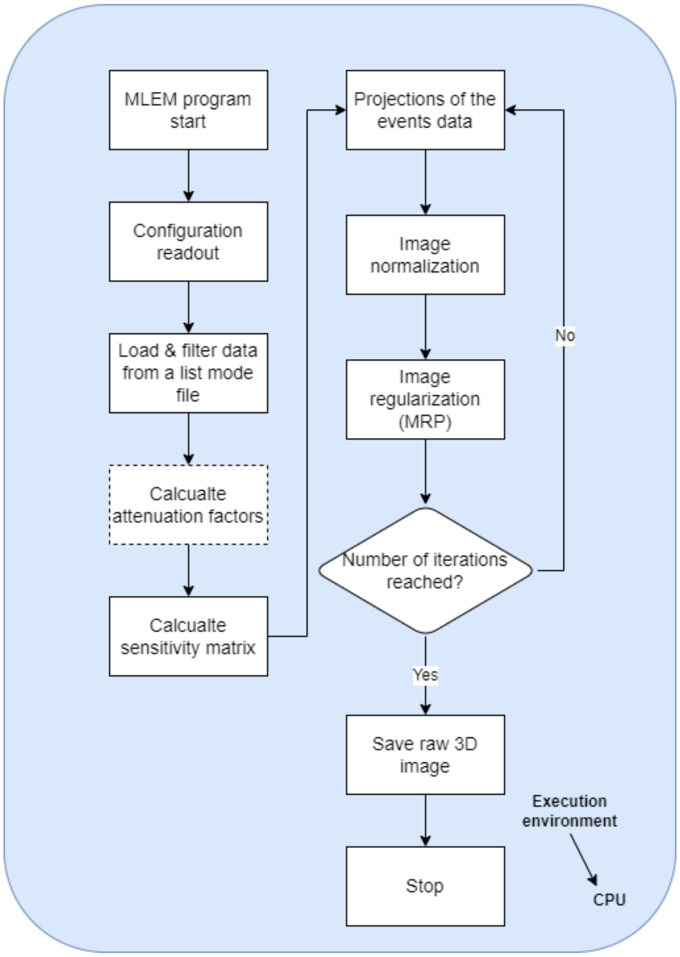

MLEM program start
Configuration readout
Load & filter data from a list mode file
Calcualte attenuation factors
Calcualte sensitivity matrix
Projections of the events data
Image normalization
Image regularization (MRP)
Number of iterations reached?
No
Yes
Save raw 3D image
Stop
Execution environment
CPU

Figure 4.2 – High-level flow chart for CPU implementation of MLEM

The program starts with reading the configuration file and the list of events from the computer drive. Optionally, attenuation factors are calculated if an appropriate segmentation map is passed as an argument to the program. The next step is to calculate the sensitivity matrix, after which the most time-consuming component of reconstruction, being FP and BP, is complete. Finally, division by the sensitivity matrix and image regularization using median root prior (MRP) [60] method finalize a single iteration. An MRP filter is used to stabilize the noise in the MLEM algorithm. The main advantage of a median-based filter is that it does not blur the edges of bright objects in the image. This summarizes the CPU-based approach. The following sections describe the steps taken to reformulate and efficiently implement this algorithm on GPU.

4.2.3 Efficient GPU Programming

Although GPU hardware comes with a great computational power, it requires careful planning and algorithm design to fully utilize its potential. Key considerations include: avoiding thread divergence, minimizing lower bandwidth global memory usage and ensuring coalesced memory access, improving the ratio of arithmetic to memory operations, and usage of faster memory types such as shared memory, constant memory, caches [34].

Another common issue in parallel programming is memory writing conflicts. Fortunately, CUDA provides API for atomic operations starting in Fermi architecture [34]. The atomic operations are guaranteed to execute without interruption from other threads, thus they are suitable for situations when simultaneous write operations might happen into the same memory location from different executing threads. However, it is still the programmer's responsibility to ensure appropriate usage of such operations to maintain integrity of the results. Atomic operations serialize execution and decrease efficiency of the GPU and should only be used when necessary.

Finally, optimizing occupancy of streaming multiprocessors is important to ensure maximum throughput. Streaming multiprocessors occupancy is determined by 2 factors; number of threads that run simultaneously and amount of shared memory usage. Once the amount of shared memory utilization is established, one must balance the number of threads/blocks for individual CUDA kernel launches.

4.2.4 Naïve GPU Implementation

We present the first attempt on reformulating Radialis MLEM for GPU. It is called “naïve” because none of the GPU-specific optimization techniques are used in this implementation and the primary focus was on transferring the execution of the most time-consuming event projections operation to GPU. Figure 4.3 shows an updated flow chart for the naïve GPU algorithm reformulation. Two new operations for copying data from CPU to GPU and copying result of the projections back to CPU were added. Execution environment for projections of events data was changed to from CPU to GPU.

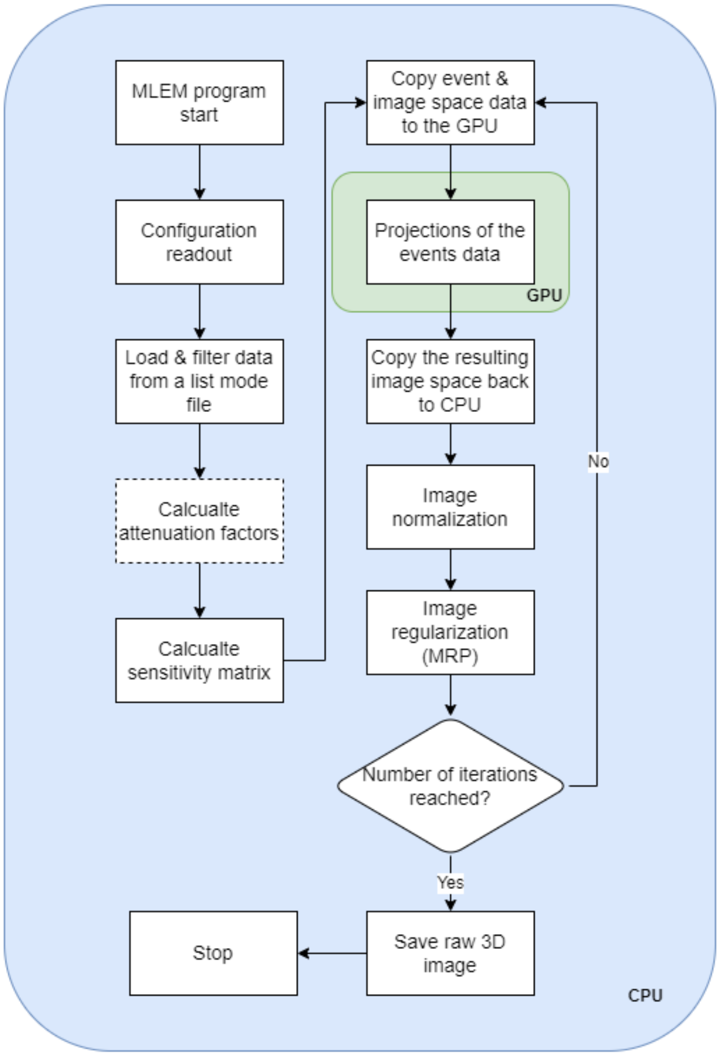

MLEM program start
Configuration readout
Load & filter data from a list mode file
Calcualte attenuation factors
Calcualte sensitivity matrix
Copy event & image space data to the GPU
Projections of the events data
GPU
Copy the resulting image space back to CPU
Image normalization
Image regularization (MRP)
Number of iterations reached?
No
Yes
Save raw 3D image
Stop
CPU

Figure 4.3 - Naive GPU

In the foundation of projection operation itself lies a line tracing algorithm, Radialis PET uses a fast and accurate voxel traversing algorithm proposed in Ref. [59]. The same algorithm was ported to GPU. All LORs are divided by the number of GPU threads in use. Then, each thread is assigned a subset of all LORs. Each LOR is processed separately via a single thread, and atomic operations are used when a thread needs to update image space after back projection. The aim of this implementation is to establish a baseline for GPU reconstruction performance that would help in assessing effects of future optimization techniques.

4.2.5 Reducing Data Transfer Overhead

Memory transfer between CPU and GPU is slow and inefficient [34] and should be minimized where possible. The naïve implementation does not account for this and requires memory transfer between CPU and GPU for every iteration. It is a straightforward change to reduce the number of data transfer operations to 2: before starting the first iteration, upload all necessary data such as list of events, FOV properties, image space, estimate space, etc. Once the required number of iterations is reached, the results of the execution are copied and saved to the disk. This change led to reformulation of all additional operations (normalization and MRP image regularization) that are repeated with each iteration. Normalization is done by multiplying our estimate by the $\frac{1}{\sum_i A_{i,j}}$ where ÷ is an element-wise division and it is easily portable to GPU since each image matrix element is calculated separately. The MRP regularization being a 3D median filter in its core can be efficiently reimplemented on GPU. In addition, attenuation factors calculation was also ported to GPU, because it is very similar to the event projections operation and only minor modifications were required to the function. Figure 4.4 presents these differences in the float chart.

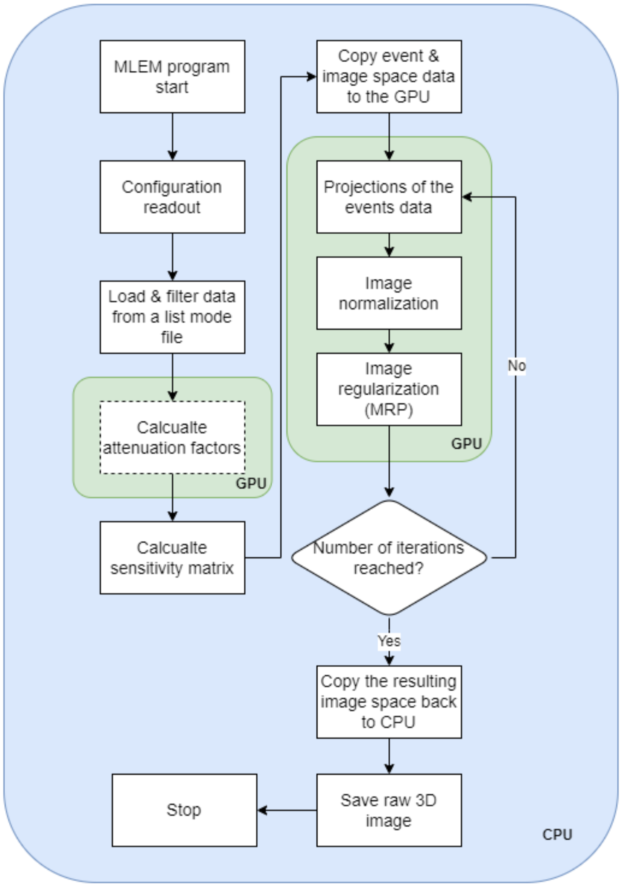

MLEM program start
Configuration readout
Load & filter data from a list mode file
Calcualte attenuation factors
GPU
Calcualte sensitivity matrix
Copy event & image space data to the GPU
Projections of the events data
Image normalization
Image regularization (MRP)
GPU
No
Number of iterations reached?
Yes
Copy the resulting image space back to CPU
Save raw 3D image
Stop
CPU

Figure 4.4 – Data transfer optimized GPU reconstruction flow chart

4.2.6 Coalesced Memory Access and Compiler Optimizations

Global memory layout plays an important role in CUDA program' performance. Due to slower speed, it is important to ensure coalesced access to the global memory so that memory transactions within a warp of threads (each thread block is divided into 32 thread warps) are combined into a single transaction [34]. In our case, a bottleneck was found in the organization of the image space in memory. We vectorize our image space into 1D array for convenient processing on GPU. In the initial implementation the vectorization was done starting with the Z axis, following by Y and X (See Figure 4.5 top). However, given planar detectors, the majority of measured LORs go along the Z axis (from one detector to another) and considering a limited number of Z slices most of adjacent voxels found during projection algorithm are X or Y adjacent, but when having the memory alignment described before these voxels end up located far from each other in the memory thus resulting in non-coalesced access. Rearranging data in the global memory helped to improve global memory performance (See Figure 4.5 bottom).

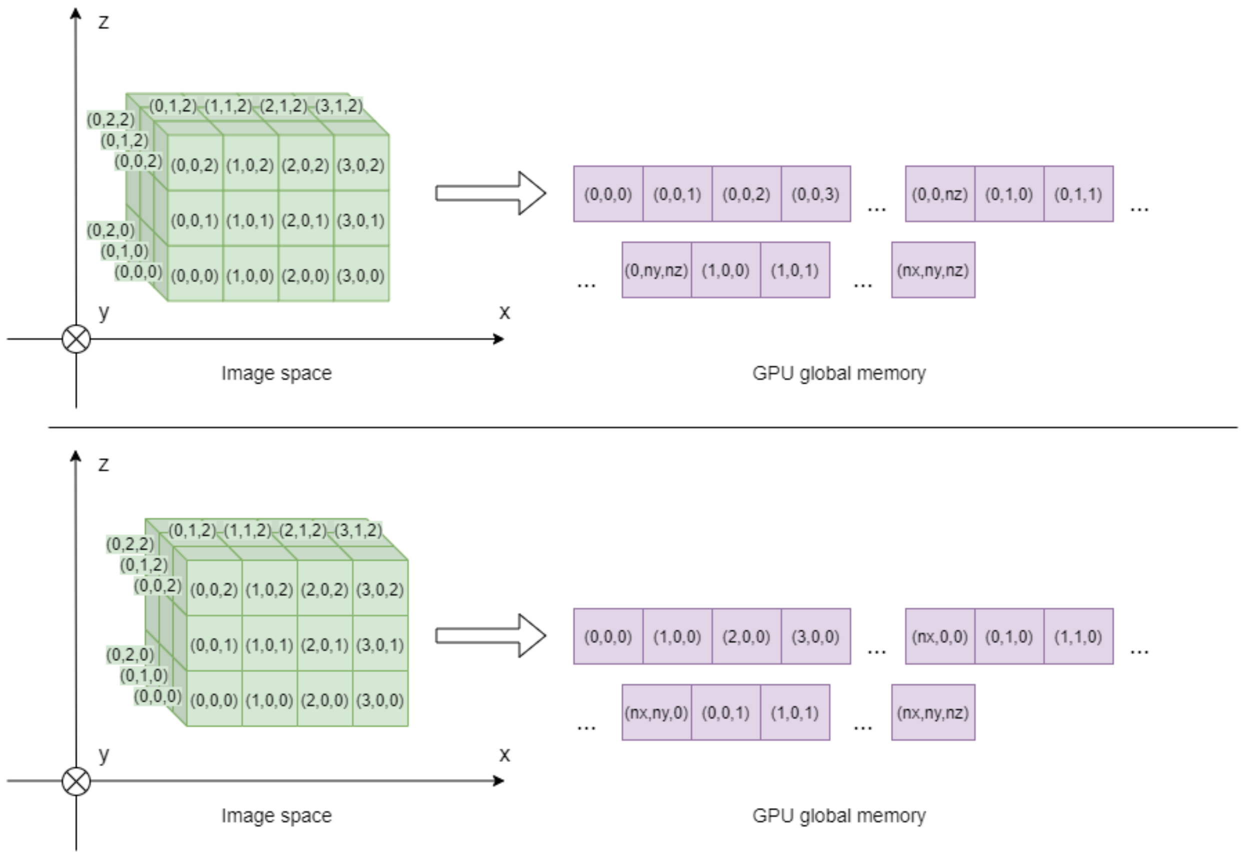


Figure 4.5 - Adjusting image space vectorization approach to improve coalesced thread access to the GPU' global memory. Top: Vectorization is happening along Z axis first then Y then X. Bottom: Revisited approach, vectorization along X axis first then Y then Z. nx, ny, nz are constants representing a total number of voxels along the x, y, z axis correspondingly.

NVIDIA GPUs and CUDA provides intrinsic alternatives for some of standard mathematical functions implemented on the hardware level. Such functions work faster and can be enabled via compiler flag *-use_fast_math*. This may result in a lower precision accuracy; however, our experiments show that enabling intrinsic functions does not impact the accuracy of results.

These optimizations mark a milestone in our reformulation process since we have not changed the conceptual approach to the reconstruction. Both FP and BP are done within a single projection operation (See Algorithm 4.1).

Algorithm 4.1 - Main loop for the projection operations, both FP (lines 3-10) and BP (lines 11-17) are done within the same event processing loop via a single thread

```
1:  for each line assigned to the thread do
2:      Read coordinates and weight $lw$ for line $l$ from global memory
3:      Find the list of voxels $V$ that $l$ intersects
4:      for $voxel = 1, 2, \ldots, V$ do
5:          Find the length of the intersection $LineLength$ of $l$ and $voxel$
6:          if $LineLength > 0$ then
7:              $weight[voxel] \leftarrow LineLength * previousVoxelValue$
8:              Add the contribution $weight[voxel]$ to the $TotalSum$ for the current line
9:          end if
10:     end for
11:     if $TotalSum > 0$ then
12:         for $voxel = 1, 2, \ldots, V$ do
13:             Calculate the contribution $w$ of the $l$ to the $voxel$
14:             $w \leftarrow (weight[voxel] * lw)/TotalSum$
15:             Atomically add $w$ to the image space in global memory
16:         end for
17:     end if
18: end for
```

4.2.7 Shared Memory & Slice-by-slice Processing

Even with all the GPU-specific optimizations added thus far the nature of the computations implemented in Algorithm 4.1 are intrinsically more suitable for CPU architecture. In order to implement both FP and BP within the same iteration over the list of events, the voxel traversing algorithm must take into account the whole FOV while traversing a LOR from one detector to another. The need to store all intersected voxels, to be later used in FP and BP calculations, creates an extensive memory overhead and in case of GPU implementation the only memory type that can handle its size is slow global memory. Having this memory overhead when threads are simultaneously reading and writing excessive amount of data to/from the global memory affects the performance. This leads to a situation when memory transactions take most of the computation time. To address this issue, we propose an approach similar to other work in [30], [53] when, in order to avoid excessive global memory consumption, FOV may be split into smaller pieces and processed individually. Such method requires extracting FP and BP operations from Algorithm 4.1 to use them independently. The difference of our approach is that we do not cache a slice of the image space itself into the fast shared memory, but instead we propose caching the list of voxels which a particular LOR intersects within a slice. Changes to the high-level flow chart diagram of the MLEM process are presented in the Figure 4.6.

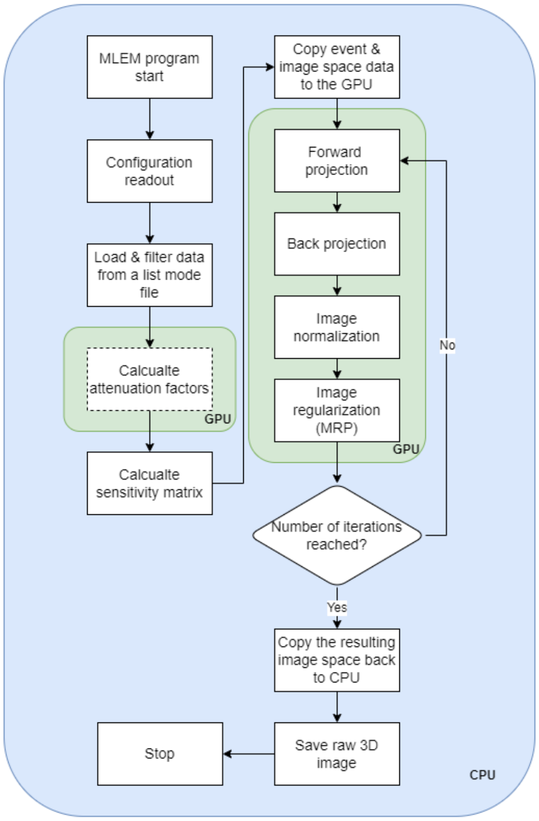

MLEM program start
Configuration readout
Load & filter data from a list mode file
Calcualte attenuation factors
GPU
Calcualte sensitivity matrix
Copy event & image space data to the GPU
Forward projection
Back projection
Image normalization
Image regularization (MRP)
GPU
No
Number of iterations reached?
Yes
Copy the resulting image space back to CPU
Save raw 3D image
Stop
CPU

Figure 4.6 - GPU reconstruction including all presented optimizations

The main difference is that the projections of the event data is now split into separate FP and BP operations.

4.2.7.1 Forward projection

The idea of FP remains to calculate the contribution of a given voxel along an LOR to the total projection value of the LOR. Implementation details are present in Algorithm 4.2. We first initialize enough shared memory to store the list of intersected voxels for each thread. Then we process the image slice-by-slice along the Z axis. Voxel traversing algorithm is applied using the previously allocated to store the intersected voxels and speed up the memory access. Finally, total sum for an event within current slice is added to the total sum of the event in the global memory. Atomic operations are not necessary here because each event is processed by its own individual thread.

Algorithm 4.2 - Forward projection

```
Initialize shared memory to hold intersected voxels
for each slice along Z axis do
    for each line assigned to the thread do
        Read coordinates and weight $lw$ for line $l$ from global memory
        Find the intersection voxels $V$ between the $l$ and the current slice.
        Save $V$ in the shared memory
        Find the length of the intersection $LineLength$ of $l$ and current
        slice
        $VoxelLength \leftarrow LineLength/length(V)$
        for $voxel = 1, 2, \ldots, V$ do
            if $VoxelLength > 0$ then
                $weight[voxel] \leftarrow VoxelLength * previousVoxelValue$
                Add the contribution $weight[voxel]$ in the current slice to
                the $SliceSum$ for the current line
            end if
        end for
        Atomically add $SliceSum$ to the $lorSum[l]$ in the global memory
    end for
end for
```

4.2.7.2 Back projection

BP differs from FP in that it calculates the contribution of a measured LOR to every voxel it passes through. Implementation details are present in Algorithm 4.3. The approach is very similar to the one in FP in terms of shared memory usage. Even though with this approach, the program needs to calculate voxels that a line intersects with twice (once in FP once in BP) doubling the computations. This is still more than excessive global memory transactions in the previous implementation.

Algorithm 4.3- Back projection

```
Initialize shared memory to hold intersected voxels
for each slice along Z axis do
    for each line assigned to the thread do
        Read coordinates and weight $lw$ for line $l$ from global memory
        $TotalSum \leftarrow lorSum[l]$
        Find the intersection voxels $V$ between the $l$ and the current slice. Save $V$ in the shared memory
        Find the length of the intersection $LineLength$ of $l$ and current slice
        $VoxelLength \leftarrow LineLength/length(V)$
        for $voxel = 1, 2, \ldots, V$ do
            if $V oxelLength > 0$ then
                Calculate the contribution $w$ of the $l$ to the $voxel$
                $weight[voxel] \leftarrow VoxelLength * previousVoxelValue$
                $w \leftarrow (weight[voxel] * lw)/TotalSum$
                Atomically add $w$ to the image space in global memory
            end if
        end for
    end for
end for
```

4.2.8 Evaluation Methods

In order to evaluate the proposed algorithm, computer hardware identical to the one deployed in the clinical system was used: Intel Core i7-6800K CPU with 6 cores/12 threads and 3.4 GHz clock speed and NVIDIA GTX 1070 GPU with 1920 CUDA cores grouped into 15 streaming multiprocessors, and 48 kB of shared memory per block. In comparisons between CPU and GPU-based method, unless specified otherwise, 8 CPU threads are utilized while GPU reconstruction only relies on a single CPU thread. Default clinical image resolution (577x433x24) and number of iterations in the MLEM reconstruction (15) were used. For the reconstruction speed evaluation, we measured execution time of the CPU and GPU-based methods. The effect of different optimization techniques presented in sections 4.2.4 - 4.2.7 was analyzed. The optimal GPU-specific parameters, such as number of blocks were found and used for execution of GPU-based algorithms.

Image quality comparison is based on visual analysis of clinical and phantom images along with a set of standardized NEMA-NU-4 tests. NEMA NU-4-based evaluations are very common for small animal and organ-targeted PET devices [61]–[64]. Part of the NEMA NU-4 based analysis of Radialis PET can be found in [45].

4.2.8.1 Visual assessment of clinical images

Comparison for both reconstruction speed and image quality between CPU-based and GPU-based reconstruction was performed using clinical data. The Radialis organ-targeted PET Camera is currently undergoing clinical trial at UHN-PMCC. Participants in the study are women with a newly diagnosed breast cancer; they receive a clinical indication for diagnostic medical imaging tests like full-field digital mammography (FFDM) with or without digital breast tomosynthesis (DBT), or breast MRI, or WB PET/CT scan. This permits to compare the diagnostic capabilities of Radialis PET with standard clinical breast cancer imaging modalities.

Prior to an imaging session with Radialis PET system, participants are injected with ^{18}F-FDG in the range of activities between 37 and 307 MBq (activity is chosen randomly and does not depend on the clinical case). Each participant rests for 60 minutes to allow for the ^{18}F-FDG uptake, followed by Radialis PET image acquisition. Features of known malignancies and additional PET findings were recorded and correlated with histopathology as the ground truth.

The majority of clinical images acquired within the scope of this clinical study were used to assess how the reconstruction speedup affects clinical image quality. Two of the clinical images along with detailed description are provided below to visually compare image quality between the baseline CPU reconstruction and the proposed GPU-based solution.

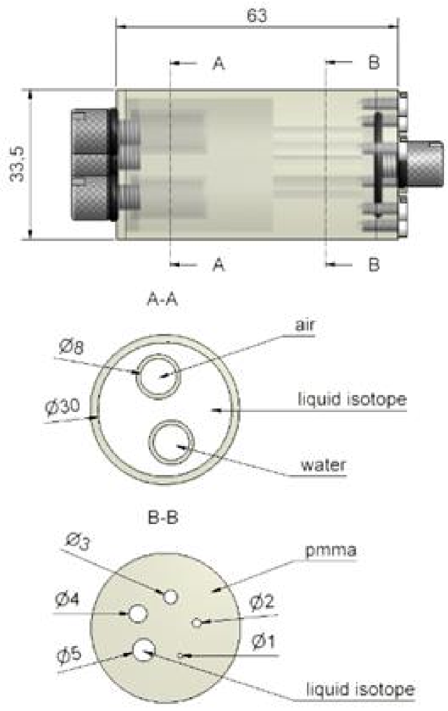
4.2.8.2 Quantitative assessment of phantom data

Evaluation of GPU-based image reconstruction in comparison with CPU-based was performed with acquisition of NEMA NU-4 Image Quality (IQ) phantom and with a Flood Field Uniformity (FFU) test. NEMA NU-4 image quality phantom is composed of two parts: the first part is a fillable cylindric chamber 30 mm in diameter and 30 mm long to be filled with an isotope (hot region). This chamber contains two smaller cavities separated from that volume, which are filled with water and air (cold regions). The second part of the phantom is 20 mm long solid cylinder that houses five fillable “hot” rods with 1, 2-, 3-, 4- and 5-mm diameters. As this is shown in Figure 4.7(B-B) “hot” rods are aligned radially around the phantom length axis, providing a connection to the first half of the cylinder, which is filled with an isotope. Thus, NEMA NU-4 image quality phantom provides hot lesions in the form of hot rods in the cold solid background, as well as uniform hot and cold regions. The phantom is filled with 18.7 MBq of activity and positioned with its axis of symmetry being perpendicular to the surface of the detectors. Three parameters are measured: Uniformity, Recovery Coefficient (RC) and Spill-over ratio (SOR). This metrics is chosen because of its importance for assessing the system’s ability to analyze images quantitively and to apply standardized uptake value (SUV) analysis to lesions of different size and uptake of a radiopharmaceutical. For the uniformity measurement, a cylindrical 22.5x10 mm^2 volume is taken within the central uniform region of the phantom. The mean activity concentration, along with maximum, minimum and standard deviation (STD), are reported. The noise and uniformity measurement in the uniform hot region assesses the signal to noise ratio. Knowing the mean uniformity value, we measure RC (that is the ratio between image-derived and true activity) for the hot rods in the following way: central 10 mm of the hot rods are averaged based on voxel values to generate a single image slice. For each hot rod: a circular region of interest (ROI) with twice the physical diameter of the rod is drawn. ROI is searched for a pixel position with a maximum value. Line profile is drawn along the rod through the maximum pixel. The mean value is taken along this line and divided by the mean activity concentration measured in the uniformity calculation (refer to Equation 4.2).

$$RC = \frac{LineMean}{MeanActivity} \tag{4.2}$$

The measurement of RC in the hot rods is used for assessment of the system’s capability to recover the activity in reconstructed images and is indicative of the spatial resolution of the scanner. Finally, the SOR is measured as a ration between mean value of the activity inside a central 4 x 7.5mm^2 region in each cold chamber and the mean of the activity concentration measured for the uniformity region. SOR assesses the accuracy of the attenuation and scatter corrections (refer to Equation 4.3).

$$SOR = \frac{Mean(cold\ chamber)}{Mean(uniform\ area)} \tag{4.3}$$



Dimensions of the phantom (in mm)

Figure 4.7- NEMA NU-4 Image Quality phantom. Source: https://www.qrm.de/en/products/micro-pet-iq-phantom/

FFU test is not specified by the NEMA standard, but it is a part of reconstruction quality assurance process at Radialis. The method used for FFU assessment is based on the uniformity measurement in gamma cameras [65]. The goal of this method is to assess the degree to which a PET system can render a uniform distribution radioactivity as flat. A flat phantom large enough to cover the whole FOV is positioned parallel to the detectors. At least 5 million coincidences are acquired and reconstructed. Once an image is generated, an 18 cm line profile (an array of inline pixel values) is extracted from the geometrical center of the image for uniformity calculation. The uniformity of a line profile is calculated using the Equation 4.4.

$$U = \frac{Max - Min}{Max + Min} \tag{4.4}$$

where Max and Min are the corresponding maximum and minimal pixel values within the line profile. The experiment is repeated for the line profiles of different width (1 - 9px).

4.3 Results

4.3.1 Complexity Analysis

In this section, the worst case computational complexity of each of the main steps of our adopted MLEM algorithm iteration is listed in Table 4.1. For simplicity we treat our resulting images as cubes with resolution N x N x N voxels, the list of all the events is represented by M. Both the initial CPU-based implementation and the resulting GPU-based solution are the implementations of the same MLEM image reconstruction algorithm applied to list mode data. The primary factors affecting computational complexity of the MLEM algorithm are the forward and back projection operations. The computational complexity associated with these operations is well known to be O(MN^2). The other two steps of MRP and normalization have the complexity of the degree of O(N^3). It seems like every step has a similar computational complexity; however, in reality, the image dimensions are cubic and in particular the Z dimension is much smaller than the number of events M.

Table 4.1 - Computational complexity for different steps of our MLEM algorithm

Step	Computational Complexity
Forward projection	O(MN^2)
Back projection	O(MN^2)
Image normalization	O(N^3)
Image regularization (MRP)	O(N^3)

4.3.2 Reconstruction Speed Effects

In this section we are going to focus on the achieved speedup in terms of the iteration time between the CPU and the latest proposed GPU-based reconstruction. A single iteration includes projection of the events data (FP and BP in case of the latest GPU-based reconstruction), normalization and MRP regularization for the image. Experiments were made on the IQ and the FFU phantoms with 15 iterations and the default image resolution (577x433x24). Voxel size for z dimension was determined by the number of slices and compression. In case of the IQ phantom the compression value is 109mm thus resulting in voxel size 0.4x0.4x4.53mm^3. In case of the FFU phantom compression value was measured at 138mm and the resulting voxel dimension was: 0.4x0.4x5.75mm^3. All experiments were done using the clinical system hardware, a single CPU thread was used for both CPU-based and GPU-based reconstructions with 30 thread blocks allocated on GPU which was measured to be an optimal number. The analysis on finding the optimal number of blocks on the GPU is present in Figure 4.8.

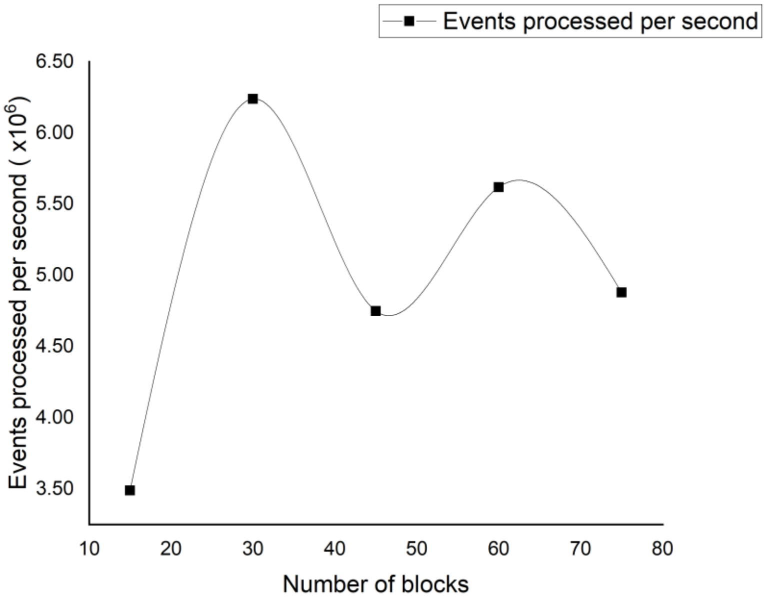


Figure 4.8 - Number of events processed per second based on the selected number of blocks for the GPU

Due to the limitations associated with shared memory used in FP and BP each block size was limited to 128 threads. Figure 4.8 shows that the GPU throughput is maximized when we utilize 30 blocks. The resulting average iteration time speedup from all the proposed optimizations was measured at 436 times between GPU and single thread CPU reconstruction. A breakdown of the contributions during the progressive transition from CPU to GPU is present in Figure 4.9.

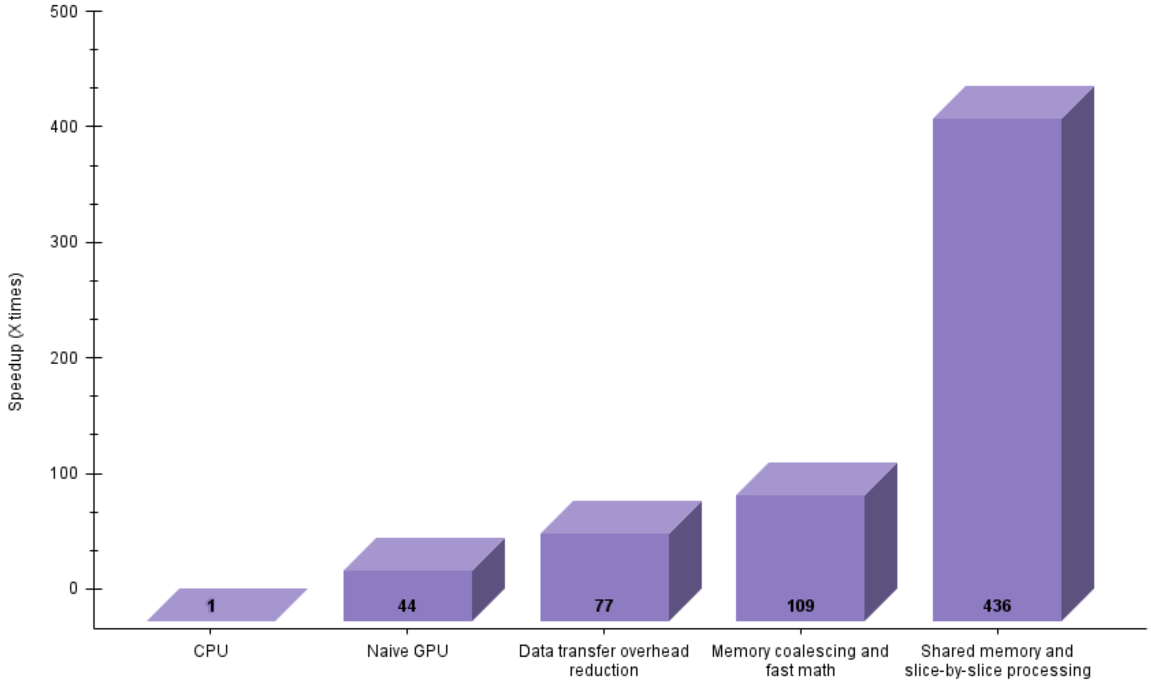


Figure 4.9 - Breakdown of measured speedup for different contributions based for the GPU-based reconstruction compared with the CPU-based reconstruction

It is visible that the most notable improvement was achieved by applying the memory-related optimizations and slice-by-slice processing for the reconstruction.

4.3.3 Clinical Reconstruction

An important step in the analysis of recently proposed GPU-based reconstruction algorithms is to understand their impact on the rate of clinical reconstruction. Clinical reconstruction is a multi-step process that includes every single step, from reading data from a hard drive to creating a DICOM file. The detailed breakdown of the time required for different steps in the clinical reconstruction on GPU is present in the Table 4.2. The table shows the time taken to use the proposed GPU-based image reconstruction algorithm in relation to the full clinical reconstruction workflow. Default clinical parameters for a typical low-dose breast scan with 1623989 events was used.

Table 4.2 - Clinical reconstruction time breakdown for a typical low-dose breast reconstruction file with 1623989 events. Every highlighted column is included in the "Reconstruction time" along with some other things that are not mentioned here explicitly such as attenuation correction, scatter correction, etc

	Make list, s	Loading events, s	Attenuation factors, s	Solid angle, s	All iterations, s	Reconstruction time, s	DICOM generation, s	Total time, s
Image reconstruction #1	37.67	9.86	N/A	1.78	5.02	77	32	146.67
Image reconstruction #2		9.57	0.45	1.68	2.41			
Image reconstruction #3		9.86	1.11	1.67	4.95			

It is worth noting that the total time spent on running all MLEM iterations in this case is 12.48 seconds or 16.2% of the time spent on image reconstruction or only 8.5% of the overall image reconstruction workflow time (including list mode transformation and DICOM generation). Which is a clear indicator that further improvement in image reconstruction speed won't have a significant impact on the overall duration of the imaging procedure.

To further evaluate performance improvements the default clinical reconstruction parameters were applied including 8 threads for the CPU reconstruction and 1 CPU thread for the GPU reconstruction. The data were split into 4 groups based on the number events per acquisition. The average iteration time speedup and total image reconstruction time speedup is reported in the Table 4.3.

Table 4.3 - Average reconstruction and iteration speedup between GPU and CPU reconstruction for different groups within prepared clinical dataset

Group, x10^6 events	Average number of events	Average image total reconstruction speedup, X times		Average iteration speedup, X times	
		GPU vs CPU 1 thread	GPU vs CPU 8 threads	GPU vs CPU 1 thread	GPU vs CPU 8 threads
0-1	383927	20.27	11.71	318.15	168.86
1-3	1979911	40.78	19.48	449.28	206.74
3-8	4859927	57.38	25.54	525.93	227.97

8-20	17125623	103.87	45.39	672.09	290.21

Table 4.3 shows that the overall reconstruction speedup is far from the actual iteration speedup achieved. As MLEM iterations become faster their impact on the overall reconstruction duration decreases and the iterations time starts taking a much smaller proportion of the time giving way to other operations (such as, various corrections, loading data from the drive, etc.) which used to take negligible amount of time in the past. Table 4.4 present a comparison between the average time needed to load events from disk to the time needed to run MLEM on the same number of evets. With the current GPU-based reconstruction it normally takes around 2 times less time to reconstruct the data compared to the time needed to load it into computer memory for a typical clinical file size.

Table 4.4 - Comparison between the time needed to load events from disk and time needed to reconstruct the same events

Number of events	Average time to load events from the drive for processing, s	Average time for 15 iterations using GPU reconstruction, s
18283650	106	48.837
4658729	28.7	11.53
2036642	12.04	6.1
1205429	6.99	3.58

4.3.4 Image Quality Analysis

Finally, it was the task of the utmost importance to ensure that the proposed approach does not sacrifice the resulting image quality. In this section we report phantom data quality analysis first, followed by visual analysis for clinical image quality.

The results of phantom tests for both CPU and GPU reconstruction are presented in Table 4.5 - Table 4.12 and are summarized below. Flood field uniformity results are shown in Table 4.5 for CPU and Table 4.6 for GPU-based reconstruction. Uniformity value is 17.73% for 1-pixel line and slowly improves while increasing line width to 14.53% for 9-pixel line. This is slightly better than the reported CPU results. NEMA NU-4 phantom evaluation is presented in Table 4.7, Table 4.9, and Table 4.11 for the CPU-based reconstruction and in Table 4.8, Table 4.10, and Table 4.12 for the GPU-based reconstruction. GPU reconstruction results show uniformity value of 13.62% within the uniform region, contrast recovery coefficients of 81%, 67%, 47%, 31%, and 14% for the 5, 4-, 3-, 2-, and 1-mm hot rods, respectively. The spill over ratio for the air-filled and water-filled reservoirs were 16% and 24% respectively. The results are on par with the values reported for CPU reconstruction. Views from different slices of the IQ phantom reconstructed with the GPU-based algorithm are illustrated in Figure 4.10. Visual analysis confirmed that there are no visible differences between phantom images reconstructed by CPU and GPU.

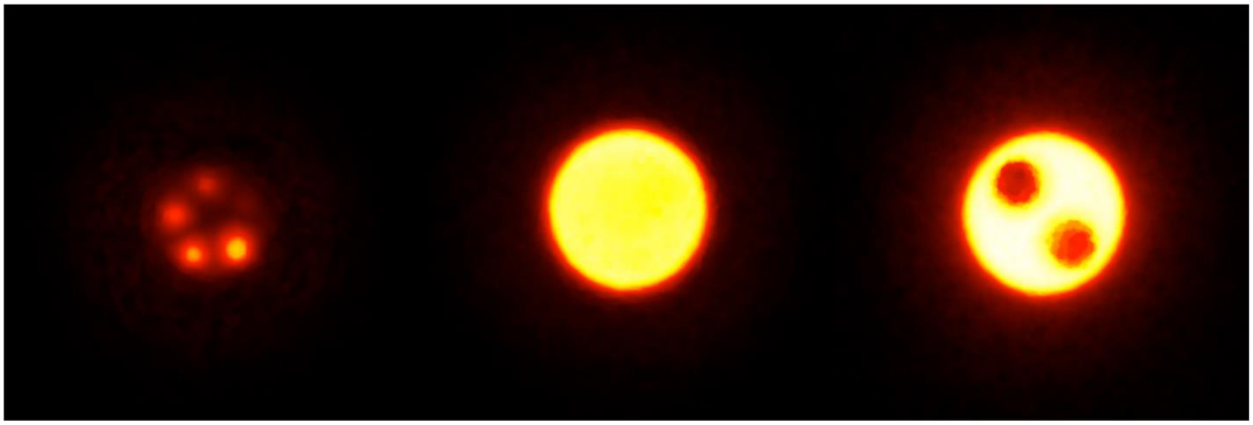


Figure 4.10 - NEMA NU-4 IQ phantom reconstructed using the latest GPU-based reconstruction. Slices displaying the hot rods for recovery coefficient (left), uniform region (center) and air and water reservoirs (right)

Table 4.5 - FFU test results for the CPU-based reconstruction

Line Width (px)	Mean	STD	Min	Max	Uniformity (%)	RMS/Mean
1	481.57	36.87	397.45	576.82	18.41	7.66
2	481.77	33.74	421.56	568.07	14.8	7
3	484.76	33.74	422.43	572.07	15.05	6.96
4	485.19	31.04	430.86	567.7	13.7	6.4
5	484.79	31.67	425.88	578.57	15.2	6.53
6	483.52	30.03	433.13	577.98	14.33	6.21
7	483.75	30.43	429.5	580.35	14.94	6.29
8	483.05	29.53	426.27	576.76	15	6.11
9	483.81	30.2	423.38	576.43	15.31	6.24

Table 4.6 - FFU test results for the GPU-based reconstruction

Line Width (px)	Mean	STD	Min	Max	Uniformity (%)	RMS/Mean
1	494.28	36.64	417.54	597.53	17.73	7.41

2	493.78	34.38	422.85	582.08	15.85	6.96
3	497.68	34.5	423.12	591.87	16.63	6.93
4	498.55	31.54	436.88	583.44	14.36	6.33
5	497.56	32.28	432.68	588.71	15.28	6.49
6	496.1	30.45	440.19	583.85	14.03	6.14
7	496.05	30.92	437.32	584.62	14.41	6.23
8	495.24	29.83	437.44	581.19	14.11	6.02
9	496.26	30.58	434.5	582.22	14.53	6.16

Table 4.7 – Measured uniformity for the IQ test used to evaluate CPU-based reconstruction

	Mean	**Minimum**	**Maximum**	**%STD**
Uniformity	35488	26846	47759	12.38

Table 4.8 - Measured uniformity for the IQ test used to evaluate GPU-based reconstruction

	Mean	**Minimum**	**Maximum**	**%STD**
Uniformity	32730.05	23929.14	45564.82	13.62

Table 4.9 - Measured RC values for the IQ test for the CPU-based reconstruction

Rod Size (mm)	**RC**	**%STD**
5	0.79	12.73
4	0.64	12.43
3	0.45	12.77
2	0.3	12.73
1	0.14	12.91

Table 4.10 - Measured RC values for the IQ test for the GPU-based reconstruction

Rod Size (mm)	**RC**	**%STD**
5	0.81	14.9
4	0.67	14.09
3	0.47	13.75
2	0.31	13.84
1	0.14	14.09

Table 4.11 - Measured SOR values for the IQ test for the CPU-based reconstruction

Chamber	**SOR**	**%STD**
Air	0.13	27.28
Water	0.21	15.58

Table 4.12 - Measured SOR values for the IQ test for the GPU-based reconstruction

Chamber	SOR	%STD
Air	0.16	46.85
Water	0.24	34.22

The results for the FFU and IQ phantom, fall under the benchmark set at Radialis where IQ uniformity must be under 20%, recovery coefficient for the 2mm hot rod must be greater than 20% and the spill over ratio must be no more than 20% and 40% for the air and water reservoirs respectively. Flood field uniformity value must be below 25% for the 10th iteration for 1px line. These are the minimum acceptable values used at Radialis to ensure image quality standards in the system.

In addition to the phantom test, visual analysis of clinical images reconstructed with both algorithms was performed. The selected results are shown in Figures 4.11 and 4.12. Figure 4.11 compares two Radialis PET camera images obtained from a 61years-old female with right breast multifocal invasive and in situ ductal carcinoma and reconstructed with CPU (left) and GPU-based (right) reconstruction algorithms. For Radialis PET acquisition, 178 MBq of ^{18}F-FDG was administrated. Both PET images show a group of multiple distinct masses, which reproduce histopathology findings of multi-focal cancer. Qualitatively, the contrast and detectability of small foci is the same in both images. It took ~1h 14m 45s to reconstruct the image of breast (Figure 4.11, left) with CPU-based reconstruction vs ~4m 55s seconds to get images for the same raw data using GPU reconstruction (Figure 4.11, right).

Figure 4.12 compares between two slices selected from 3D low-dose (37 MBq of ^{18}F-FDG) Radialis PET image reconstructed with CPU-based (Figure 4.12, left) and GPU-based (Figure 4.12, right) reconstructions. The scan was acquired in a 56-year-old female with histopathology diagnosed invasive ductal carcinoma and intermediate-grade ductal carcinoma in situ (DCIS). Two focal uptakes on both images correspond to two histopathology proven masses. Visual analysis shows no differences in the reconstructed images: they look almost identical and clearly present ^{18}F-FDG uptake in both lesions. It took ~25m 45s to produce the image of breast (Figure 4.12, left) with CPU-based reconstruction vs ~3m 38s seconds to get images for the same raw data using GPU reconstruction (Figure 4.12, right).

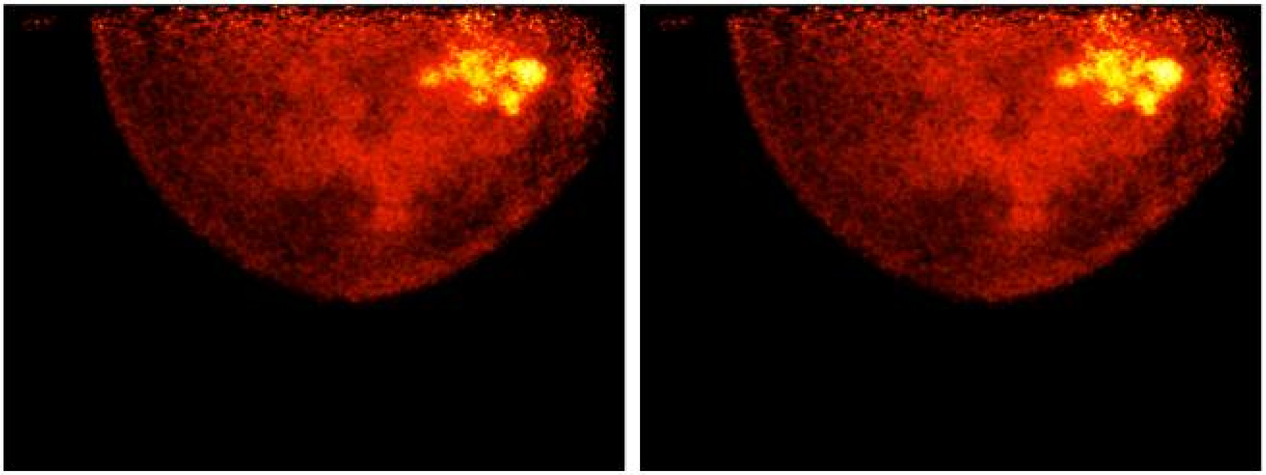


Figure 4.11 - A 61-years old female with right breast multifocal invasive and in situ ductal carcinoma. Breast images reconstructed with CPU (left) and GPU-based (right) reconstruction algorithms. Acquisition time: 10m. Total reconstruction time for the CPU version is ~1h 14m 45s compared to ~4m 55s for GPU reconstruction

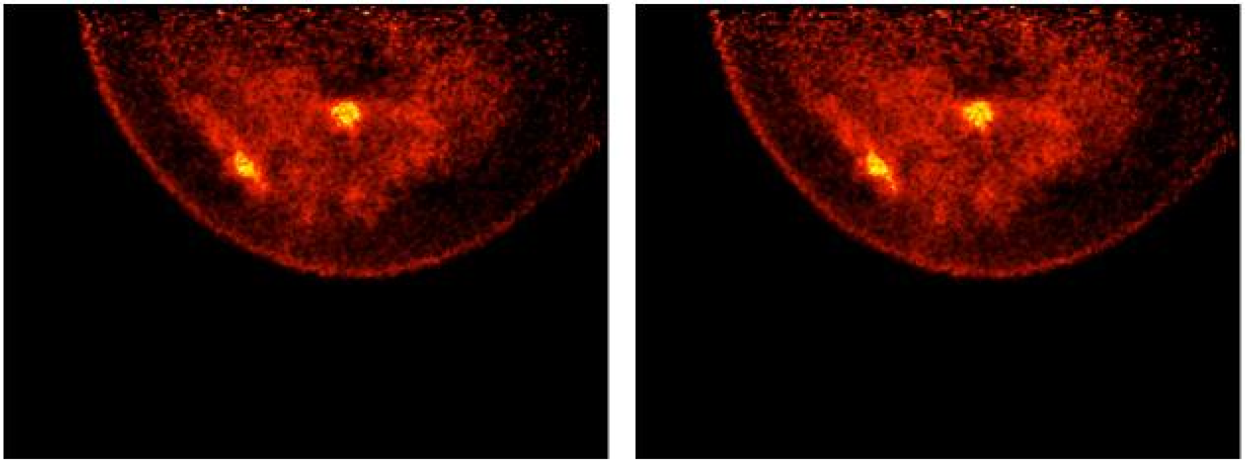


Figure 4.12 - A 56-years old female with invasive ductal carcinoma. Breast images reconstructed with CPU (left) and GPU-based (right) reconstruction algorithms. Acquisition time: 5 m. Total reconstruction time for the CPU version is ~25m 41s compared to ~3m 38s for GPU reconstruction

5 Conclusion

Organ-targeted PET technology is gaining momentum in recent years. New applications are being developed and tested both in clinical applications and in research laboratories. PEM is the most used clinical application for organ-targeted PET. The Radialis organ-targeted PET camera is a great example of a successful PEM system that is actively paving its path to clinical use. The novel approach to the detector architecture significantly improves system sensitivity and spatial resolution in comparison to standard WB PET devices. This, in turn, allowed for imaging with one tenth of a typical WB PET radiation dose, making a significant impact on clinical adaptability of the technology. Radialis PET has successfully demonstrated its capability in addressing the current deficiencies of X-ray and MRI breast screening for dense breast imaging.

This work is devoted to adopting novel GPU-based MLEM image reconstruction algorithm for the clinical organ-targeted PET system with planar detectors to speed up image reconstruction time. Reconstruction speed plays an important role in modeling complexity of a PET system, patient throughput, and research and innovation speed. As shown in section 4.3.4, transition from CPU to GPU did not compromise the image quality which was proven with both phantom and clinical data. This also realized a significant improvement in single iteration speed of at least 436x times compared to the CPU-based reconstruction used previously. Our progressive approach has shown that GPU-specific memory and data processing optimization techniques account for much of the speed gain. However, even the naïve implementation has lowered average iteration time by about 44x times. The resulting clinical reconstruction speedup was more than 24 times on average compared to the clinical CPU-based reconstruction used by Radialis PET. The improvement in reconstruction time is also demonstrated on clinical images presented in Figure 4.11 and Figure 4.12. In both cases, reconstruction time was less than the acquisition time while CPU reconstruction time exceeded acquisition time 7.5 times for Figure 4.11 and 2.66 times for Figure 4.12. Therefore, our goal of making reconstruction time less than the acquisition time has been successfully achieved and validated with both clinical and phantom images, establishing the proposed algorithm as ready for clinical deployment. It should be noted that the speedup becomes larger for files with higher number of events and smaller for files with lower number of events. This can be explained by the fact that higher count projection operations, which are primarily ported to GPU, take proportionally more time compared to CPU-based operations. The percentage of time spent on the MLEM algorithm falls below 15% for most clinical reconstructions. Overall, the novel GPU-based approach shifts the most time-consuming operations away from image reconstruction itself to IO operations, list mode generation, and data corrections. At this stage, it was shown that reading events from the hard drive on average takes 2x longer when compared to the time needed to reconstruct the same number of events.

The proposed solution can be easily extended for more advanced algorithms and physical models that would incorporate Depth of Interaction (DOI) data, shift varying PSF [66], and even ToF PET. One straightforward extension of this work would be to implement a more accurate model to represent the physics of LORs, such as the Tube-of-Response (TOR) approach [67]. This may lead to improvements in the overall image quality; however, this approach was not practical previously due to additional computational overhead it introduces on top of typical MLEM algorithm and slow speed of the CPU-based reconstruction.

GPU technology is constantly developing by increasing number of cores, shared memory size, etc. We believe that the proposed method implemented in CUDA has a potential to yield an improved performance on future generations of the GPUs without any reformulation efforts. A straightforward extension of this work may include a multi-GPU adaptation of the presented method.

Finally, we believe that this reformulation will have a profound impact on the patient throughput capabilities of the clinical Radialis PET system by way of increased image reconstruction speed and freeing of CPU resources by switching the main computation burden to GPU. The vacant resources could be utilized in other applications such as preview image generation, system monitoring, data acquisition, etc. This work ultimately brings us closer to incorporating PEM devices such as Radialis PET in breast cancer screening procedures and increasing adaptation of technology to improve outcomes in broader patient populations. Success of PEM devices may also encourage clinical adaptation of PET for other indications. It is our privilege to work with an industry partner and see the great applicability of the developed solution. We believe that the proposed GPU reconstruction method is ready for the next stage of testing and development that will begin with the clinical deployment occurring this Fall.

6 Future Work

An important topic that was not covered in this work is the comparison of performance between our solution and other GPU-based MLEM / OSEM implementations out there [30], [53]. PET systems come with several different unique characteristics, such as detector geometry, data format, reconstruction algorithm, data corrections, etc; all of which affect the resulting image reconstruction. Here, we focus on the whole clinical reconstruction process with an emphasis on the effects of reconstruction speed affecting it. To compare our solution with other works in the area, we would need to extract the proposed method and come up with some standardized performance metric, using similar hardware, and simulated system characteristics, such as detector geometry. This certainly can be done in the future in order to better understand the applicability of our solutions to other systems; however,it is out of scope of this work.

In addition to that, we anticipate three primary directions for further improvements of Radialis PET: software optimizations and development, image reconstruction improvements, and data pre/post processing. Improvements in these directions are not necessarily exclusive to each other. Advances in GPU hardware and Artificial Intelligence (AI) methods such as Machine Learning (ML) and Deep learning (DL) have recently begun to affect medical imaging through new AI-based solutions to common problems directly or indirectly related to image reconstruction [68]. DL methods in PET image reconstruction have been an active area of research in recent years [69]–[71]. It would be remiss to ignore the application of this emerging technology in considering the future of the Radialis PET. A common issue with DL-based methods in PET stems from reliance on the anatomical modalities in addition to PET data for training purposes as well as the need for high quantity and quality of the training data. In addition, current DL-based methods work best with the 2D data and in many cases are not efficient with the 3D data commonly used in PET. Overall, we do not see AI as the one-and-only method of improvement for the Radialis PET since an efficient GPU-based reconstruction opens doors for adopting more complex physical models for the image reconstruction. Figure 6.1 presents the main directions of development for the technology.

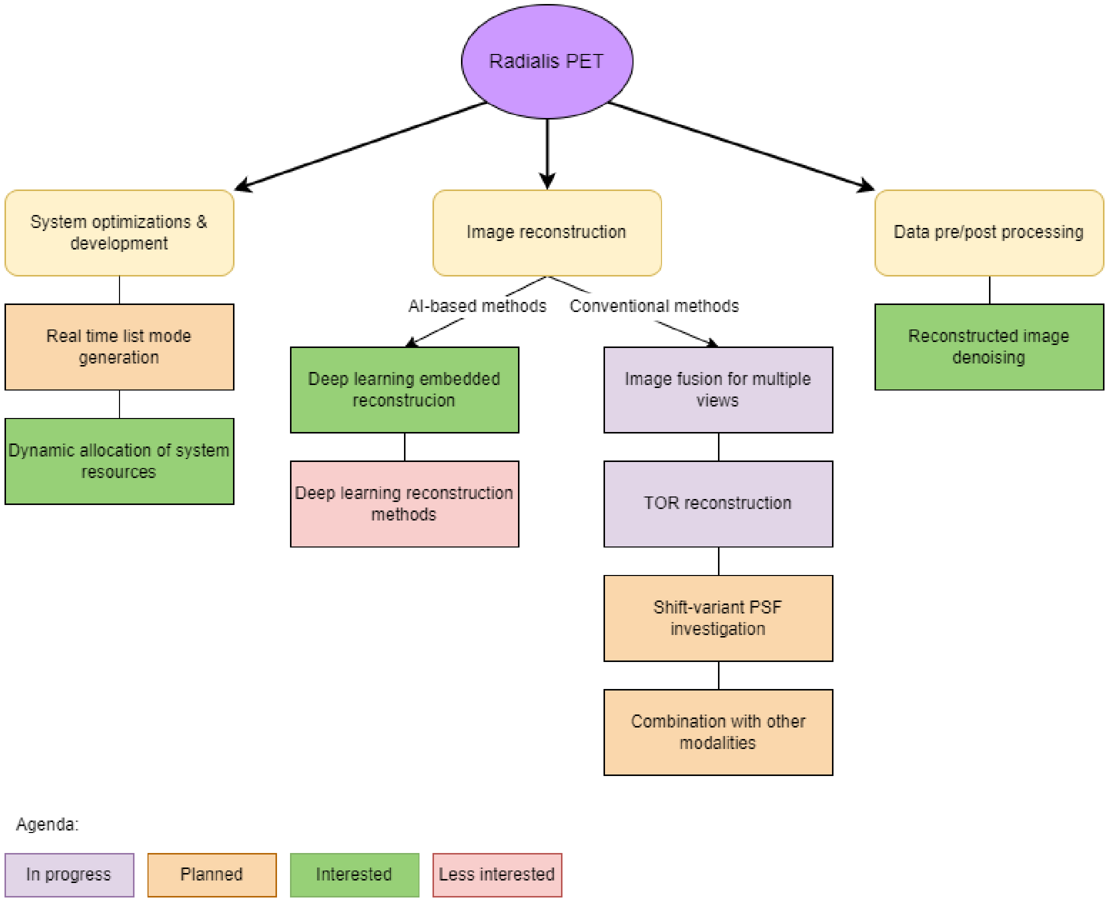

Radialis PET
System optimizations & development
Image reconstruction
Data pre/post processing
Real time list mode generation
Dynamic allocation of system resources
AI-based methods
Conventional methods
Deep learning embedded reconstrucion
Deep learning reconstruction methods
Image fusion for multiple views
TOR reconstruction
Shift-variant PSF investigation
Combination with other modalities
Reconstructed image denoising
Agenda:
In progress
Planned
Interested
Less interested

Figure 6.1 - Future developments for Radialis PET. Current, planned and potential future work

The first direction going forward is “System optimization & development”. The primary focus is to increase the efficiency of the system by improving some aspects which are not directly related to the image reconstruction. Examples of such improvements may include real time list mode generation for increasing speed in data processing and reducing the IO overhead. Employing two hardware architectures in the system places importance on the ability to maximize resources allocation at any given point in time. We continue to investigate possible ways to allocate resources more intelligently to improve overall system utilization.

The second direction is data pre/post processing. Examples for improvements in this area can be motion detection using corrections applied at the list mode level, post reconstruction image denoising or feature enhancing. Some promising work has already been done in this area using the DL methods [72].

The third and most significant area for potential improvements lies in the image reconstruction domain itself. In Figure 6.1 we divided this group into “Conventional” methods and “AI-based” methods. Conventional methods are well known and initially proposed some time ago, with a potential to improve clinical image quality. It may be possible to compensate for low axial resolution inherent to planar detector systems by implementing image fusion of various view angles in clinical situations where greater resolution is required (such as brain scans). Further improvements may be achieved through extending our reconstruction model with TOR approach or a shift variant PSF which are now possible with the GPU reconstruction. Finally, there is an experimental project aiming to combine Radialis PET with MRI.

There are two main directions for AI incorporation in PET image reconstruction: standalone reconstruction where a trained DL model receives measured data and tries to generate the resulting image [71], [73] and embedded approach where DL or ML techniques are incorporated within iterative reconstruction in order to impact a particular aspect of it such as denoising image estimate during iterative reconstruction[74]. While the first approach in its current state is not likely to be implemented with a standalone PET, due to the aforementioned reasons, the second approach shows promise for the Radialis PET.

References

[1] T. Jones, “The role of positron emission tomography within the spectrum of medical imaging,” 1996.

[2] A. J. González, F. Sánchez, and J. M. Benlloch, “Organ-Dedicated Molecular Imaging Systems,” *IEEE Transactions on Radiation and Plasma Medical Sciences*, vol. 2, no. 5. Institute of Electrical and Electronics Engineers Inc., pp. 388–403, Sep. 01, 2018. doi: 10.1109/TRPMS.2018.2846745.

[3] D. L. Bailey, *Positron Emission Tomography Basic Sciences*. Springer, 2005.

[4] A. Verger, S. Grimaldi, M. J. Ribeiro, S. Frismand, and E. Guedj, “Single Photon Emission Computed Tomography/Positron Emission Tomography Molecular Imaging for Parkinsonism: A Fast-Developing Field,” *Annals of Neurology*, vol. 90, no. 5. John Wiley and Sons Inc, pp. 711–719, Nov. 01, 2021. doi: 10.1002/ana.26187.

[5] W. Vaalburg, H. Hendrikse, and E. F. J. de Vries, “Drug development, radiolabelled drugs and PET.”

[6] P. Zanzonico, “Positron Emission Tomography: A Review of Basic Principles, Scanner Design and Performance, and Current Systems,” *Semin Nucl Med*, vol. 34, no. 2, pp. 87–111, 2004, doi: 10.1053/j.semnuclmed.2003.12.002.

[7] S. Ren *et al.*, “Data-driven event-by-event respiratory motion correction using TOF PET list-mode centroid of distribution,” *Phys Med Biol*, vol. 62, no. 12, pp. 4741–4755, Jun. 2017, doi: 10.1088/1361-6560/aa700c.

[8] S. R. Cherry, M. Dahlbom, and M. E. Phelps, *PET: Physics, Instrumentation, and Scanners*. New York, NY: Springer New York, 2006. doi: 10.1007/0-387-34946-4.

[9] C. Caldarella, G. Treglia, and A. Giordano, “Diagnostic Performance of Dedicated Positron Emission Mammography Using Fluorine-18-Fluorodeoxyglucose in Women With Suspicious Breast Lesions: A Meta-analysis,” *Clin Breast Cancer*, vol. 14, no. 4, pp. 241–248, Aug. 2014, doi: 10.1016/j.clbc.2013.12.004.

[10] K. Shilling, “New Breast Imaging Modalities Reveal Cancers as Small as a Single Millimeter,” 2008.

[11] I.N. Weinberg, “Dedicated apparatus and method for emission mammography,” 5252830, Oct. 1993

[12] I.N. Weinberg, “Dedicated apparatus and method for emission mammography,” 5519221, May 1996

[13] C. J. Thompson, K. Murthy, I. N. Weinberg, and F. Mako, “Feasibility study for positron emission mammography,” *Med Phys*, vol. 21, no. 4, pp. 529–538, Apr. 1994, doi: 10.1118/1.597169.

[14] T. Chong, “Frost & Sullivan Names Naviscan a Leader in the Future of Molecular Breast Imaging,” *Frost & Sullivan*, 2008.

[15] R. E. Hendrick, “Radiation Doses and Cancer Risks from Breast Imaging Studies,” *Radiology*, vol. 257, no. 1, pp. 246–253, Oct. 2010, doi: 10.1148/radiol.10100570.

[16] M. J. Yaffe and J. G. Mainprize, “Risk of Radiation-induced Breast Cancer from Mammographic Screening,” *Radiology*, vol. 258, no. 1, pp. 98–105, Jan. 2011, doi: 10.1148/radiol.10100655.

[17] A. M. Fowler, “A Molecular Approach to Breast Imaging,” *Journal of Nuclear Medicine*, vol. 55, no. 2, pp. 177–180, Feb. 2014, doi: 10.2967/jnumed.113.126102.

[18] C. B. Hruska and M. K. O’Connor, “Nuclear imaging of the breast: Translating achievements in instrumentation into clinical use,” *Med Phys*, vol. 40, no. 5, p. 050901, May 2013, doi: 10.1118/1.4802733.

[19] A. J. Reader and H. Zaidi, “Advances in PET Image Reconstruction,” *PET Clinics*, vol. 2, no. 2. pp. 173–190, Apr. 2007. doi: 10.1016/j.cpet.2007.08.001.

[20] L. A. Shepp and Y. Vardi, “Maximum likelihood reconstruction for emission tomography,” *IEEE Trans Med Imaging*, pp. 113–122, 1982.

[21] G. ; Delso, S. ; Fürst, B. ; Jakoby, R. ; Ladebeck, C. ; Ganter, and S. G. Nekolla, “Performance measurements of the Siemens mMR integrated whole body PET/MR scanner,” *The Journal of Nuclear Medicine*, vol. 52, no. 12, pp. 1914–1922, 2011.

[22] T. García Hernández *et al.*, “Performance evaluation of a high resolution dedicated breast PET scanner,” *Med Phys*, vol. 43, no. 5, pp. 2261–2272, May 2016, doi: 10.1118/1.4945271.

[23] H. M. Hudson and R. S. Larkin, “Accelerated Image Reconstruction Using Ordered Subsets of Projection Data,” *IEEE Trans Med Imaging*, vol. 13, no. 4, pp. 601–609, 1994, doi: 10.1109/42.363108.

[24] K. Lange and R. Carson, “EM Reconstruction Algorithms for Emission and Transmission Tomography,” *J Comput Assist Tomogr*, pp. 306–316, 1984.

[25] E. Veklerov and J. Llacer, “Stopping rule for the MLE algorithm based on statistical hypothesis testing,” *IEEE Trans. Med. Imag*, vol. 6, no. 4, pp. 313–319, 1987, [Online]. Available: https://escholarship.org/uc/item/833837hq

[26] A. Gaitanis, G. Kontaxakis, G. Spyrou, G. Panayiotakis, and G. Tzanakos, "PET image reconstruction: A stopping rule for the MLEM algorithm based on properties of the updating coefficients," *Computerized Medical Imaging and Graphics*, vol. 34, no. 2, pp. 131–141, Mar. 2010, doi: 10.1016/j.compmedimag.2009.07.006.

[27] J. Liacer *et al.*, "Results of clinical receiver operating characteristics study comparing filtered backprojection and maximum likelihood estimator images in FDG PET studies.," *The Journal of Nuclear Medicine*, vol. 34, no. 7, pp. 1198–1203, 1993.

[28] G. Pratx, G. Chinn, P. D. Olcott, and C. S. Levin, "Fast, Accurate and Shift-Varying Line Projections for Iterative Reconstruction Using the GPU," *IEEE Trans Med Imaging*, vol. 28, no. 3, pp. 435–445, Mar. 2009, doi: 10.1109/TMI.2008.2006518.

[29] K. Chidlow and T. Möller, "Rapid emission tomography reconstruction," in *Proceedings of the 2003 Eurographics/IEEE TVCG Workshop on Volume graphics - VG '03*, 2003, p. 15. doi: 10.1145/827051.827053.

[30] J. Y. Cui, G. Pratx, S. Prevrhal, and C. S. Levin, "Fully 3D list-mode time-of-flight PET image reconstruction on GPUs using CUDA," *Med Phys*, vol. 38, no. 12, pp. 6775–6786, 2011, doi: 10.1118/1.3661998.

[31] S. Ha, S. Matej, M. Ispiryan, and K. Mueller, "GPU-Accelerated Forward and Back-Projections With Spatially Varying Kernels for 3D DIRECT TOF PET Reconstruction," *IEEE Trans Nucl Sci*, vol. 60, no. 1, pp. 166–173, Feb. 2013, doi: 10.1109/TNS.2012.2233754.

[32] F. Meng, J. Wang, S. Zhu, J. Cheng, J. Liang, and J. Tian, "Comparison of GPU reconstruction based on different symmetries for dual-head PET," *Med Phys*, vol. 46, no. 6, pp. 2696–2708, Jun. 2019, doi: 10.1002/mp.13529.

[33] X. Gu *et al.*, "Real-time reconstruction solution for positron emission mammography imaging-guided intervention," in *2015 IEEE Nuclear Science Symposium and Medical Imaging Conference (NSS/MIC)*, Oct. 2015, pp. 1–5. doi: 10.1109/NSSMIC.2015.7582065.

[34] NVIDIA, "CUDA C++ Programming Guide." 2022. Accessed: Jul. 25, 2022. [Online]. Available: https://docs.nvidia.com/cuda/cuda-c-programming-guide/index.html

[35] P. Després and X. Jia, "A review of GPU-based medical image reconstruction," *Physica Medica*, vol. 42. Associazione Italiana di Fisica Medica, pp. 76–92, Oct. 01, 2017. doi: 10.1016/j.ejmp.2017.07.024.

[36] "NEMA NU 4-2008. Performance measurements of Small Animal Positron Emission Tomographs," 2008.

[37] S. Vandenberghe, P. Moskal, and J. S. Karp, "State of the art in total body PET," *EJNMMI Phys*, vol. 7, no. 1, p. 35, Dec. 2020, doi: 10.1186/s40658-020-00290-2.

[38] A. Soriano *et al.*, "Performance evaluation of the dual ring MAMMI breast PET," in *2013 IEEE Nuclear Science Symposium and Medical Imaging Conference (2013 NSS/MIC)*, Oct. 2013, pp. 1–4. doi: 10.1109/NSSMIC.2013.6829103.

[39] G. Reynés-Llompart, C. Gámez-Cenzano, I. Romero-Zayas, L. Rodríguez-Bel, J. L. Vercher-Conejero, and J. M. Martí-Climent, "Performance Characteristics of the Whole-Body Discovery IQ PET/CT System," *Journal of Nuclear Medicine*, vol. 58, no. 7, pp. 1155–1161, Jul. 2017, doi: 10.2967/jnumed.116.185561.

[40] T. Pan *et al.*, "Performance evaluation of the 5-Ring GE Discovery MI PET/CT system using the national electrical manufacturers association NU 2-2012 Standard," *Med Phys*, vol. 46, no. 7, pp. 3025–3033, Jul. 2019, doi: 10.1002/mp.13576.

[41] I. Rausch, A. Ruiz, I. Valverde-Pascual, J. Cal-González, T. Beyer, and I. Carrio, "Performance Evaluation of the Vereos PET/CT System According to the NEMA NU2-2012 Standard," *Journal of Nuclear Medicine*, vol. 60, no. 4, pp. 561–567, Apr. 2019, doi: 10.2967/jnumed.118.215541.

[42] A. M. Grant, T. W. Deller, M. M. Khalighi, S. H. Maramraju, G. Delso, and C. S. Levin, "NEMA NU 2-2012 performance studies for the SiPM-based ToF-PET component of the GE SIGNA PET/MR system," *Med Phys*, vol. 43, no. 5, pp. 2334–2343, Apr. 2016, doi: 10.1118/1.4945416.

[43] J. van Sluis *et al.*, "Performance Characteristics of the Digital Biograph Vision PET/CT System," *Journal of Nuclear Medicine*, vol. 60, no. 7, pp. 1031–1036, Jul. 2019, doi: 10.2967/jnumed.118.215418.

[44] W. Luo, E. Anashkin, and C. G. Matthews, "Performance Evaluation of a PEM Scanner Using the NEMA NU 4—2008 Small Animal PET Standards," *IEEE Trans Nucl Sci*, vol. 57, no. 1, pp. 94–103, Feb. 2010, doi: 10.1109/TNS.2009.2036847.

[45] J. Stiles *et al.*, "Evaluation of a High-Sensitivity Organ-Targeted PET Camera," *Sensors*, vol. 22, no. 13, p. 4678, Jun. 2022, doi: 10.3390/s22134678.

[46] L. Parra and H. H. Barrett, "List-mode likelihood: EM algorithm and image quality estimation demonstrated on 2-D PET," *IEEE Trans Med Imaging*, vol. 17, no. 2, pp. 228–235, 1998, doi: 10.1109/42.700734.

[47] A. J. Reader, K. Erlandsson, M. A. Flower, and R. J. Ott, “Fast accurate iterative reconstruction for low-statistics positron volume imaging,” *Phys Med Biol*, vol. 43, no. 4, pp. 835–846, Apr. 1998, doi: 10.1088/0031-9155/43/4/012.

[48] T. Xia, A. M. Alessio, and P. E. Kinahan, “Dual energy CT for attenuation correction with PET/CT,” *Med Phys*, vol. 41, no. 1, p. 012501, Dec. 2013, doi: 10.1118/1.4828838.

[49] A. Soriano *et al.*, “Attenuation correction without transmission scan for the MAMMI breast PET,” *Nucl Instrum Methods Phys Res A*, vol. 648, pp. S75–S78, Aug. 2011, doi: 10.1016/j.nima.2010.12.138.

[50] L. Martins *et al.*, “Scatter Correction for Positron Emission Mammography using an Estimation of Trues Method Approach,” *Procedia Technology*, vol. 5, pp. 903–911, 2012, doi: 10.1016/j.protcy.2012.09.100.

[51] B. Ma *et al.*, “Scatter Correction Based on GPU-Accelerated Full Monte Carlo Simulation for Brain PET/MRI,” *IEEE Trans Med Imaging*, vol. 39, no. 1, pp. 140–151, Jan. 2020, doi: 10.1109/TMI.2019.2921872.

[52] G. Pratx and L. Xing, “GPU computing in medical physics: A review,” *Med Phys*, vol. 38, no. 5, pp. 2685–2697, May 2011, doi: 10.1118/1.3578605.

[53] M. Teimoorisichani and A. L. Goertzen, “A Cube-Based Dual-GPU List-Mode Reconstruction Algorithm for PET Imaging,” *IEEE Trans Radiat Plasma Med Sci*, vol. 6, no. 4, pp. 463–474, Apr. 2022, doi: 10.1109/TRPMS.2021.3077012.

[54] Jingyu Cui, G. Pratx, Bowen Meng, and C. S. Levin, “Distributed MLEM: An Iterative Tomographic Image Reconstruction Algorithm for Distributed Memory Architectures,” *IEEE Trans Med Imaging*, vol. 32, no. 5, pp. 957–967, May 2013, doi: 10.1109/TMI.2013.2252913.

[55] Z. Bahi, J. Bert, A. Autret, and D. Visvikis, “High performance Multi-GPU acceleration for fully 3D list-mode PET reconstruction,” in *2012 IEEE Nuclear Science Symposium and Medical Imaging Conference Record (NSS/MIC)*, Oct. 2012, pp. 3390–3393. doi: 10.1109/NSSMIC.2012.6551772.

[56] E. Veklerov, J. Llacer, and E. J. Hoffman, “MLE reconstruction of a brain phantom using a Monte Carlo transition matrix and a statistical stopping rule,” *IEEE Trans Nucl Sci*, vol. 35, no. 1, pp. 603–607, Feb. 1988, doi: 10.1109/23.12795.

[57] R. L. Siddon, “Fast calculation of the exact radiological path for a three-dimensional CT array,” *Med Phys*, vol. 12, no. 2, pp. 252–255, Mar. 1985, doi: 10.1118/1.595715.

[58] P. M. Joseph, “An Improved Algorithm for Reprojecting Rays through Pixel Images,” *IEEE Trans Med Imaging*, vol. 1, no. 3, pp. 192–196, Nov. 1982, doi: 10.1109/TMI.1982.4307572.

[59] Y. K. Liu, H. Y. Song, and B. Žalik, “A General Multi-step Algorithm for Voxel Traversing Along a Line,” *Computer Graphics Forum*, vol. 27, no. 1, pp. 73–80, Mar. 2008, doi: 10.1111/j.1467-8659.2007.01097.x.

[60] S. Alenius, U. Ruotsalainen, and J. Astola, “Using local median as the location of the prior distribution in iterative emission tomography image reconstruction,” *IEEE Trans Nucl Sci*, vol. 45, no. 6, pp. 3097–3104, Dec. 1998, doi: 10.1109/23.737670.

[61] N. Belcari *et al.*, “NEMA NU-4 Performance Evaluation of the IRIS PET/CT Preclinical Scanner,” *IEEE Trans Radiat Plasma Med Sci*, vol. 1, no. 4, pp. 301–309, Jul. 2017, doi: 10.1109/TRPMS.2017.2707300.

[62] Z. Gu *et al.*, “NEMA NU-4 performance evaluation of PETbox4, a high sensitivity dedicated PET preclinical tomograph,” *Phys Med Biol*, vol. 58, no. 11, pp. 3791–3814, Jun. 2013, doi: 10.1088/0031-9155/58/11/3791.

[63] C. Pei *et al.*, “NEMA NU-4 performance evaluation of a non-human primate animal PET,” *Phys Med Biol*, vol. 64, no. 10, p. 105018, May 2019, doi: 10.1088/1361-6560/ab1614.

[64] M. Amirrashedi *et al.*, “NEMA NU-4 2008 performance evaluation of Xtrim-PET: A prototype SiPM-based preclinical scanner,” *Med Phys*, vol. 46, no. 11, pp. 4816–4825, Nov. 2019, doi: 10.1002/mp.13785.

[65] “Computer-Aided Scintillation Camera Acceptance Testing,” 1981. doi: 10.37206/8.

[66] A. M. Alessio *et al.*, “Application and Evaluation of a Measured Spatially Variant System Model for PET Image Reconstruction,” *IEEE Trans Med Imaging*, vol. 29, no. 3, pp. 938–949, Mar. 2010, doi: 10.1109/TMI.2010.2040188.

[67] C. Schretter, “A fast tube of response ray-tracer,” *Med Phys*, vol. 33, no. 12, pp. 4744–4748, Nov. 2006, doi: 10.1118/1.2369467.

[68] J.-G. Lee *et al.*, “Deep Learning in Medical Imaging: General Overview,” *Korean J Radiol*, vol. 18, no. 4, p. 570, 2017, doi: 10.3348/kjr.2017.18.4.570.

[69] G. Zaharchuk, “Next generation research applications for hybrid PET/MR and PET/CT imaging using deep learning,” *Eur J Nucl Med Mol Imaging*, vol. 46, no. 13, pp. 2700–2707, Dec. 2019, doi: 10.1007/s00259-019-04374-9.

[70] J. Cui *et al.*, “PET image denoising using unsupervised deep learning,” *Eur J Nucl Med Mol Imaging*, vol. 46, no. 13, pp. 2780–2789, Dec. 2019, doi: 10.1007/s00259-019-04468-4.

[71] I. Häggström, C. R. Schmidtlein, G. Campanella, and T. J. Fuchs, “DeepPET: A deep encoder–decoder network for directly solving the PET image reconstruction inverse problem,” *Med Image Anal*, vol. 54, pp. 253–262, May 2019, doi: 10.1016/j.media.2019.03.013.

[72] W. Lu *et al.*, “An investigation of quantitative accuracy for deep learning based denoising in oncological PET,” *Phys Med Biol*, vol. 64, no. 16, p. 165019, Aug. 2019, doi: 10.1088/1361-6560/ab3242.

[73] Z. Hu *et al.*, “DPIR-Net: Direct PET Image Reconstruction Based on the Wasserstein Generative Adversarial Network,” *IEEE Trans Radiat Plasma Med Sci*, vol. 5, no. 1, pp. 35–43, Jan. 2021, doi: 10.1109/TRPMS.2020.2995717.

[74] G. Corda-D’Incan, J. A. Schnabel, and A. J. Reader, “Memory-Efficient Training for Fully Unrolled Deep Learned PET Image Reconstruction With Iteration-Dependent Targets,” *IEEE Trans Radiat Plasma Med Sci*, vol. 6, no. 5, pp. 552–563, May 2022, doi: 10.1109/TRPMS.2021.3101947.

List of Publications

Parts of this work were accepted for conference publication or going to be published in a future paper:

- **Application of GPU-based Reconstruction for a Clinical Organ-Targeted PET Scanner** results were accepted to be presented at the 2022 IEEE Nuclear Science Symposium and Medical Imaging Conference
- **Reformulation of 3D MLEM Image Reconstruction Algorithm from CPU to GPU for the Application in a Clinical Organ-targeted PET Camera** (in progress)