**A Comparative Analysis of Implementing a Deconvolution of a Diffused Image in real-time as a front-end for Calcium Imaging**

Ray Duran  
Department of Electrical Engineering and Computer Science  
Univertisy of North Dakotaray.duran@und.edu

*Abstract*- **There have been many advancements in miniaturized microscopes that produce in-vivo calcium extraction imaging in the study of neuroscience. Typically these small form factor cameras use small lenses but are restricted by the normal physical tradeoffs of resolution and field of view. In place of small lenses there has also been a lot of development for on-chip fluorescence microscopy using diffuser lenses that can allow for better tradeoffs. While diffuser lenses provide a low-cost and simple alternative to traditional lenses, they do require the solution of an ill-posed inverse problem that for in-vivo calcium extraction require real-time processing. We aim to provide in this paper a comprehensive comparison between the latency computation time using a gradient descent optimization minimizing a convex function for three processing devices, a Graphical Processing Unit (GPU), Central Processing Unit (CPU), and Field Programmable Gate Array (FPGA).**

I. Background

With the acknowledgement of the Nobel Prize in Chemistry in 2008 for Shimomura, Chalfie, and Tsien for the discovery and development of the green fluorescent protein GFP, the past decade has touched off a stream of advances and papers in the extraction of Ca2+ in-vivo imaging of large number of neural cells. At its’ core calcium imaging requires the processing of four major areas: background filtering, motion correction, neural enhancement and finally calcium trace extraction for processing. [ reference /fluor\_mico NIPS-2017 Giovannucci). On way to reduce the dimensionality of the data and provide noise reduction is by the ubiquitous principal component analysis (PCA). As a first step this filtered data would then be fed into the next stage of our calcium image processing, which would usually entail a way to identify the actual neurons themselves. One method tried by [ reference /calcium imaging Automated…Mukamel). was to use Independent Component Analysis, which used the FastICA( See Reference Hyvrinen and Oja 2000 ), which would be of aid in the motion correction portion of the processing to help compensate for any motion of the brain. Finally, ideas from image processing, involving image segmentation could help with the final extraction of neural images for study and analysis.

< A figure like /calcium\_imaging Automated paper by Mukamel Figure 1 >

While calcium imaging defines the approach technique with which we intend to study neural activity, it remains how best to gather the light field, fluorescence light, considering the maximum field of view, resolution, and efficiency in collecting light. Here, we investigate the use of diffusers as our front-end system for a calcium imaging

Diffusers provide a light field representation of 4-D information into a 2-D format, and in the process capture spatial and angle information. While in this paper diffusers provide our principal optical component, they have been successfully used in many other application (3D neural activity [3], compressive radar imaging [4], synthetic aperture [5], and visual odometry [6].). As a highly diffractive medium, the speckles (caustic patterns) from interference in our diffuser create a type of signature that can encode plenoptic information about our illuminated object and with proper assumptions allows us linearity in reconstruction that lends itself to well-established inverse problem definition and optimization techniques. To aid us in solving diffuser inverse problems with less than full rank we make use of the randomness they provide is to see if this fits within the theoretical framework of compressive sensing. Mathematically, if the combination of the measurement and representation matrix satisfy the property of mutual coherence then full reconstruction of the sparse system is possible.

The literature of using diffusers for lens less applications has been explored by many(/diffusers). In addition to their devicway( amplitude mask Flatscope Paper by Rice univ and Jesse K Adams found in folder (/fluor\_micro) Finally there are other ways of approaching the reconstruction problem tht do not entail, one example is the study of phase space itself, of light as it passes through a scattering medium, [ Optical Phases Space meas, Liu from UC Berkeley}

Having briefly mentioned the overall system modality, calcium imaging, and our optical component choice, diffusers, we now outline the three technologies we intend to compare for the solving of our inverse problem: Graphical Processing Unit (GPU), Central Processing Unit (CPU), and Field Programmable Gate Array (FPGA). Here, it is best to investigate these three computing paradigms by studying three aspects of computing: 1. The programming model: As the language that these computing engines use highlights their internal architecture and will allow us to gain insight into how latency can be computed. 2. How memory is partitioned and used in the device. And 3., how we achieve concurrency in our computation.

Of the three technologies, Graphical…. (See pic 1a)We will focus on the GPU by Nvidia, that uses a language called CUDA, that stands for Compute Unified Device Architecture( CUDA) . ( see reference 24 by Lancouture in folder solving\_inverse/non\_real\_time), which is really just an extension of the C/C++ language, with key-words used by the NVIDIA processor. The CUDA code breaks down the programming workflow into three steps: i) Allocating and internalizing data. This means since a CPU is usually the starting point of a program that the CPU copies data to be used by the GPU appropriate memory. ii) Calls the GPU kernel and launches threads. Iii) Upon completion of computing task, copies data back into the CPU memory. Memory usage in the NVIDIA GPU consist of ….. Concurrency in GPUs is achieved by the architecture and a programming construct. GPUs group threads in a hierarchy called warps . Here threads are controlled intrinsically to make sure that there are no data hazards to computing, A programming construct called barrier synchronization allows synchronization across warps that pauses all threads in a block when needed.

For the technology of CPUs we choose to compare the Intel i7 multicore processor which is Talk about CPUs…( pic 1b). List some references for CPU here ref( <https://www.intel.com/content/www/us/en/products/sku/227853/intel-core-i71265ue-processor-12m-cache-up-to-4-70-ghz/specifications.html> CPUs are programmed by C/C++ and often can include librarties that are needed for the math of linear algebra such as Blaze, GNU scientific library , and librsb to name a few. See Referencr for a longer list of math libraries for linear algebra.

<https://en.wikipedia.org/wiki/Comparison_of_linear_algebra_libraries>

Memory in CPUs ….. Concurrency in CPUs is achieved with the use of multicores, of which in our comparison the i7 Intel processor has four.

Field Programmable Gate Arrays (FPGAs) are dominated by two companies, AMD( Xilinx FPGAs) and Intel(Altera FPGAs). Both families of FPGAs are ubiquitous in the embedded comm are widely used in embedded products that require speed (MHz), latency(time) and allow for the flexibility of being able to re-program an algorithm even when the product is deployed in the field. Here is a small sampling of the use of FPGAs in…..

<https://www.sciencedirect.com/topics/engineering/field-programmable-gate-arrays#:~:text=FPGA%20is%20an%20integrated%20circuit,switches%20in%20the%20interconnect%20matrix>.

FPGAs use a family of languages called Hardware Description Language (HDL), of which there are two main dialects, VHSIC Hardware Description Language (VHDL) and Verilog, with System Verilog a newer version of Verilog. HDL allows capture of a design down to a level called Register Transfer Logic (RTL) which breaks down the computation into fundamental digital logic elements such as flip-flops(registers), state-machines, shifters, muxes, adders, AND and OR gates, etc. This flexibility of course permits almost any kind of computation and data transfer with the only limitations being speed and resources of the device. The memory of the FPGA exists throughout the device and can therefore be allocated as shared or global memory depending on the exact application. Additionally, modern FPGAs also consist of interfaces, DDR memory as one example, that allow for the expanse of external memory, thus greatly enlarging the memory available for logic to process. Finally, since an FPGA is essentially a sea of gates that operate with many clocks, concurrency is implicit as a result.

<Show picture of GPU, CPU and FPGA>

II. Methods

In this section below, we present a variation of the gradient descent algorithm called the Fast Iterative Shrinkage-Thresholding Algorithm (FISTA). Here we assume that our calibrated point spread function is invariant( see figure) and that allows us to reconstruct the image using linear deconvolution, <see figure below>

Diagram

Description automatically generated

In theory a lens less system has all the information available at the sensor. However, the problem is ill-posed. Introducing a diffuser gives structure to our point spread function, or random matrix, allowing us to solve an otherwise intractable problem. If we enforce sparsity as a prior, and non-negativity (no negative pixels), minimize for the least squares we are trying to solve the equation below

Text

Description automatically generated

Here in (1) we do not show the regularization term that helps enforce sparsity, instead we want to show the main thrust of how we approach minimization overall

Text

Description automatically generated

Here, the significance of (4) is that we have reduced the problem of computing A which could be a very large matrix to the adjunct of M and C, a much more manageable problem. (C is introduced as a cropping matrix)

A screenshot of a computer

Description automatically generated with low confidence

Here there are a few points to note. The first is that the M matrix in (5) has already been decoupled from A, and this is because if we included the cropping effect into A, the matrix would be ill-conditioned. Continuing with (5), we have just expressed the convolution of “Mv” as a product of Fourier and the inverse of that dot product but have expressed the equation as products of matrices.

Text

Description automatically generated

Here we substituted (7) back into (4) to give us (8) and taken the adjunct of (8) to give us (9). At this point (8) and (9) we can implement into code to solve our iteration (10).

Text

Description automatically generated

A picture containing text, gallery, different, clock

Description automatically generated

Here we include a figure with a picture that shows what the PSF would look like( Antipa Reference) and how the FISTA as it operates would iterate on our original diffused images.

( see figure PSF from our paper last semester Fig 7.)

( see figure reconstruction from our paper last semester Fig 8/)

III.System Implementation

To implement the FISTA algorithm from our method section II into our different computing platforms (CP): GPU, CPU and FPGA, we must consider the structural differences of each CP. The main components of our algorithm consist of six major sections: i) padding/down sampling, ii) FFT/IFFT, iii) Pre-computations of convolutional matrix and adjoint iv) Cropping, v) gradient-descent updates, vi) Data flow processing of new images into and out of memory. (See Figure below)

Diagram

Description automatically generated

Here owing to the architecture of the GPU, we can define the threads in our blocks and grids as (2D) to process our 2D FFT/IFFT functions (See figure below ( see reference 24 by Lancouture in folder solving\_inverse/non\_real\_time figure 2.2>). Data movement in an out of the GPU is usually assisted by one-board CPUs that act as direct memory access (DMA) controllers. The rest of the FISTA functions, like down-sampling, padding and cropping can be implemented in threadsT he table below shows the possible specs from a GPU that we can use.

< Figure 2.2 from See figure below ( see reference 24 by Lancouture in folder solving\_inverse/non\_real\_time figure 2.2>).

< add a figure with specs from . ( see reference 24 by Lancouture in folder solving\_inverse/non\_real\_time table 4.1>

Owing to the obfuscation of our C/C++ implementation of the FISTA algorithm by the C compiler, where the scheduling of the cores and threads is handled at the OS level, the practice of extracting maximum concurrency in operation can be challenging. The figure X shows how hyperthreading is handled in the i7 processor (see reference for hyperthreading here) Here the i7 FPU will handle the DSP elements of our algorithm. This includes both FFT and IFFT operations in our FISTA algorithm along with multiplies in the Fourier domain along with any additions and subtractions that are required from the gradient update operations. (See Ogata reference). Padding and down-sampling and cropping can be burst operations and are repetitive and can be instruction cached. Finally, when possible, the pre-computed convolutional matrix and adjoint matrix can be data cached and the high-speed external DDR4 memory can be used for data movement of new diffused images.

< Add a table that shows then specs of our own PC>

< Add picture of hyperthreading that is in /non\_real\_time/cpu/hyperthreading

< add reference <https://appuals.com/how-does-hyper-threading-work-in-intel-core-i7-processors/> >

( Reference paper is in file : IEEE under FPGa folder/ “An Efficient, Model-Based CPU-GPU Heterogeneous FFT Library”, by Ogata, et al. CUDA can help in some areas here.( See reference [11] from Ogata paper above)For the Cores (CPU and GPU)we see ( from Ogata [5] and [8]))

In an FPGA, using fixed point 16 bits, with flexibility offered by the fine-grain architecture of logic, memory, and DSP elements we breakdown our FISTA algorithm as follows. We buffer up the several frames of our original image that received from our CMOS sensor of 3280(H) x 2464(V) into external DDR4 @ 32 bits wide. In the next step we would down sample our original image, this would help us save in total computational time, into an image of size 256x324 pixels. Down sampling in digital signal processing is not an intensive operation and just really involves discarding N samples from a row or column. Proceeding, we would next need to pad our data so that we can streamline the processing into an FFT, as most FFTs are radix 2. So, for our FPGA implementation we would then convert our 256x324 image into a 512x1024 pixel image. While FFT cores of radix 2 and N size up to 4K are just IP cores of an FPGA vendor, 2-D cores need to be constructed out of those libraries in general. Here in our FPGA implementation, we would use internal block ram to buffer for the transposition of row FFT into our column FFT data flow. Figure (For 2-D FFT) highlights the basic strategy of construction. (See figure /fpga 2d fft as an example of a 2-D implementation). With the computation of the FFT, we multiply our pre-computed convolution matrix before computing the inverse FFT. We then performing the cropping operation in the time-domain where we again truncate to a 256x354 pixel image. At this point we have the convolution matrix “A” matrix multiplied by our first estimate of the recovered image “v”. Proceeding, now with the first step of the gradient subtraction operation. All the FPGA processing post the FFT do not require heavy operations and can be done with high-speed and low latency in DSP multiplier blocks where we use the arithmetic component of the DSP. Having up to this point effectively half of the processing for one image, we repeat the process from padding to output of inverse FFT, but now use the adjunct convolution matrix as our constant multiplicative factor. Finally, we stream this updated prediction of our reconstruction back into memory to be repeated all over again by all the elements above. Based in some Python experimentation with the FISTA algorithm we estimate that we would have to repeat the process above approximately 80-100 times for a reasonably valid image of the deconvolution.

< See figure for a 2-D FFT>

< Add a table with specs of the dev board Ultrascale+ KCU116>

IV.Results

The frequency of the operations are repeated for 100 iterations

Operation Size Frequency Portion of Code ( Note 6) Cycles/Time(Notes1,3)

Normalization 256x324 ( Note 10) pre-processing raw data 1 ( Note 4)

Down Sample 2048x2592 ( Note 10) pre-processing raw data 1 ( Note 4)

FFT/IFFT 512x1024 4 during gradient update ( Note 9)

Pad/Crop 512x1024 3 during gradient update 1 ( Note 4)

Add/Sub 512x1024 1 vk – alpha\*gradient ( Note 5) 1 ( Note 4)

2 xk+((tk-1/t\_k1)\*(xk-x\_k1) 1 ( Note 4)

Sqrt N/A N/A N/A ( Note 7) N/A

Mult(element-wise) 512x1024 vk – alpha\*gradient 1 ( Note 2)

512x1024 H.\*vk 1 ( Note 2)

512x1024 H\_adj.\*X 1 ( Note 2)

Square N/A N/A N/A ( Note 7) N/A

Divide N/A N/A N/A ( Note 7) N/A

Non-negativity 512x1024 1 during update of xk 1 ( Note 8)

Notes/Assumptions: 1) GPU and FPGA takes 1 clock cycle. 2) CPU takes 3 clock cycles 3) All cycles are evaluated at 300 MHz which is a reasonable speed for off-core logic. 4) Normalization, Down sampling, pad/crop and adds takes 1 clock cycle 5) Ignore calculation of alpha term. 6) Ignore initialization of PSF matrices in calculation 7) For scalar calculations we ignore computation because of negligible size compared to matrices.8) non-negativity assume one clock 9) See FFT/IFFT Equations for calculating time of computation. 10) The frequency of the operations is repeated for 100 iterations except for Normalization and Down sample which are done once for block.

Here are the FFT calculations that we referenced in our table(X) above. Note that we are not including any of the data transfer times for CPU calculation. Our reasoning for this is that the GPU requires this as an extra step for processing and that data transfer times for the CPU and FPGA can be “baked into” the calculations that we made for the down sampling. In other words we just say that we are preforming the fetches of data from memory and that that counts as our data transfer.

Now, for the CPU and GPU FFT we compute as

%% FFT specific System Parameters

Ktr = 1.11e-8; % Matrix transpose on CPU

KgMa = 5.95e-10; % Memory allocation on GPU

Kc2g = 6.16e-9; % Data transmission from CPU to GPU

Kg2c = 5.68e-9; % Data transmission from CPU to GPU

KgP = 2.73e-11; % Pre-processing of 1D-FFT on GPU

KdDP = 7.28e-12; % Post-processing of 1D-FFT on GPU

KcFFT = 2.61e-9; % 1D-FFT on CPU

KgFFT = 1.65e-10; % 1D-FFT on GPU

KgMF = 3.01e-9; % Memory releasing on GPU

Mtr = Ktr; % Note!!: Assumption From our understanding

% that Mtr=Ktr=Ctr

Ctr = Mtr;

%% Other parameters

n = 512; % Rows

m = 1024; % Cols

%% CPU-only 2D FFT computational time

r = 0;

C\_col = (1-r)\*n\*m\*2\*Ktr\*Ctr + (1-r)\*m\*n\*log(m)\*KcFFT;

C\_row = (1-r)\*m\*n\*log(n)\*KcFFT;

time\_cpu = C\_col + C\_row;

%% GPU-only 2D FFT computational time

r = 1;

G\_col = r\*n\*m\*2\*Ktr\*Ctr + r\*m\*n\*log(m)\*KgFFT + ...

r\*n\*m\*(KgMa + Kc2g + KgP + KdDP + Kg2c );

G\_row = r\*m\*n\*log(n)\*KgFFT + ...

r\*n\*m\*(Kc2g + KgP + KdDP + Kg2c + KgMF );

time\_gpu = G\_col + G\_row;

For FPGA FFT times we use the ( ref /fpga 10-1244 paper by Nguyen) where latency is 156us for a 32K FFT much larger than we need for our 512 point matrix. So, as an approximation for the FFT we ignore the transposition( because our FFT point size is much larger than our FFT) and we double up the value for the 2-D. This gives us 312 us time for the FPGA 2-D FFT.

Now computing

clearvars -except time\_cpu time\_gpu

freq = 300e6

cycle\_time = 1/freq;

time\_fpga = 312e-6; % fft time for 2-D FPGA

iter = 50; % Assume that this is reasonable time for reconstruction

%% FISTA time

n = 512; % Rows

m = 1024; % Cols

padding = n\*m\*cycle\_time\*3;

add\_sub = n\*m\*cycle\_time\*3;

mult\_gpu = n\*m\*cycle\_time\*3;

mult\_fpga = n\*m\*cycle\_time\*3;

mult\_cpu = n\*m\*cycle\_time\*3\*3;

non\_neg = n\*m\*cycle\_time;

%% Total times for one iter

gpu\_one\_iter\_time = time\_gpu\*4 + padding + add\_sub + mult\_gpu + non\_neg;

cpu\_one\_iter\_time = time\_cpu\*4 + padding + add\_sub + mult\_cpu + non\_neg;

fpga\_one\_iter\_time = time\_fpga\*4 + padding + add\_sub + mult\_fpga + non\_neg;

%% Total times for reconstruction

gpu\_total\_time = gpu\_one\_iter\_time\*iter;

cpu\_total\_time = cpu\_one\_iter\_time\*iter;

fpga\_total\_time = fpga\_one\_iter\_time\*iter;

Estimate FPGA

In our system we will only compare the 2- dimensional FFTs since they account for the major processing modules in our FISTA deconvolution design.

Estimates with CPU

<ref for in solve\_inverse\_problem/cpu instruction\_tables.pdf page 279/442 MUL I<UL r32 3 cycle latency >

Can we make the assumption that this is what it will be for every single data, or does this get cached and then requires only one clock, not sure……?

Estimates with GPU

< ref nvidia\_tensor\_core in /solve\_inverse\_problem/gpu page 1 can do one mult of matrix 4x4 in one cycle >

WE are assuming here that we are doing single precision arithmetic.

We will assume that additions and subtractions are just as long.

Do the times meet our system requirements for processing 30 frames a second for a size image that has been downsampled from the original full CMOS array.

Remember a reference of why we are using 3 clocks for CPU look at (instructions\_table.pdf )

V.Discussion

??? WE are not sure how images are buffered in memory and also how the cores are partitioned in the CPU

Talk about calcium imaging to be done Other types of processors

The reason we want to use a diffuser is for the tradefoss that we have discussed but also the reason that we are interested in a fast way to do compitations is that the processing of the steps for calcium imaging also requires a certain amount of time

??? Why even try to comnibe in real time the functional processing of diffusion with calcium imaging

Well I would say that threre are applications and papers that discuss real time calcium imaging BCAS(UCLA folk) and that adding the diffusion we are really just chaging the front end of how we collect light to better enhance the overall system. Said otherwise we have just added another pre-processor step tlo the systemn that will allow us to improve the accuracy of detecting and extracting the neurons of intereste…

There are models that can be used to calibrate

Talk about models and variant and invariant

While in this paper we focus on the front-end processing of reconstructing the image,

Talk about ADMM

A little about optogenetics

Double vs single vs fixed…..

Mixing and matching GPU,CPU and FPGA

Memory limitations

Programming limitations that relate to DirectX or OpenGL

VI.Conclusion

Future work could also include wireless

Future work could also include with optogenetics control of a system with a closed loop.

VII..References

FPGAs

<https://www.sciencedirect.com/topics/engineering/field-programmable-gate-arrays#:~:text=FPGA%20is%20an%20integrated%20circuit,switches%20in%20the%20interconnect%20matrix>.