Abstract - Intense computations in engineering and science, especially bioinformatics have been made practical by the recent advances in Graphical Processing Unit (GPU) computing technology. In this study, implementation and performance evaluations for a GPU-accelerated dynamic functional connectivity (DFC) analysis, which is an analysis method for investigating dynamic interactions among different brain networks, is presented. Open Computing Library (OpenCL), which provides a general framework for GPU computing, is utilized, and it is shown to reduce the DFC analysis computation time. The parallel implementation with OpenCL provides up to 10x speed-up over sequential implementation.

I. INTRODUCTION

Advances in Graphical Processing Unit (GPU) technology enabled the utilization of video cards for accelerating computationally-intensive and parallelizable algorithms in science and engineering. Most of the modern video devices provide a significant alternative as computing facilities to accelerate algorithms which can be implemented in parallel. Large number of lightweight threads which can be executed concurrently in a GPU enable properly designed parallel algorithms to be executed relatively faster than sequential implementation, for general purpose computing applications [1-3]. Despite the fact that exploitation of a GPU for computing purposes can be quite complicated, Open Computing Library (OpenCL) provides a good abstraction for programmers to utilize GPUs efficiently [4]. Since their introduction, OpenCL-based applications attracted researchers to use this tool in various scientific and engineering applications [5-7]. OpenCL-based parallel programming techniques are also utilized for speeding up computing methods in bioinformatics [8-13]. An example of parallelizable computing methods used in medical imaging is dynamic functional connectivity (DFC) analysis, which basically performs sliding-window time-series analysis of temporal neuroimaging data from the brain [14-17]. With the help of functional magnetic resonance imaging (fMRI) data, researchers can reveal dynamic interactions among various brain networks using DFC, which can also reveal different brain interaction dynamics between normal and diseased brain [14-17]. FMRI data from multiple subjects can be quite large in size. In addition, the number of distinct brain networks, and number of different combinations of interactions among these brain networks can also be very large. DFC analysis can be computationally intensive depending on the input data size, resolution, and analysis parameters such as window width, and window step size. Accelerating DFC computations is obviously useful for performing analysis rapidly and extracting the interactions among the regions of interest. With acceleration, dynamic interactions can be tracked near-real-time or real-time during fMRI experiments, and the interactions can be potentially used as neurofeedback, which can be potentially used to learn voluntary self-regulation of brain activity [18-40] and voluntary regulation of functional interactions between different brain regions [30]. This leads to improved different sensory perceptions based on different targeted brain areas [34, 35], faster motor response [18] based on neurofeedback training of the primary motor cortex, therapeutic effects in chronic patients [23], Parkinson’s disease [36], tinnitus [24], and depression [41]. Hence, the focus in this study is the utilization of parallel computing capabilities of video card that have a GPU with OpenCL support to speed up the DFC analysis on a multi-subject fMRI dataset. In this work, 234 de-identified fMRI scans from 83 subjects were used. Each had 30 brain networks and corresponding time-courses. These networks and their time-courses were revealed by a group spatial independent components analysis (ICA), which is a data-driven blind source-separation technique that provides independently behaving brain networks [18, 19].

Experiments were realized on NVidia GeForce GTX 660 which supports OpenCL and is widely available at present; therefore, the results show the efficiency for practical cases. Sequential results which were referenced for performance evaluations were also obtained on the same host. The paper is organized as follows; the following section provides a brief introduction about the DFC analysis and its sequential implementation; the third section explains the implementation of the algorithm using OpenCL; detailed performance measurements are presented in the fourth section.
II. Dynamic Functional Connectivity (DFC) Analysis

DFC analyses explore interactions among brain regions of interest using fMRI data. These analyses help identify pairs of brain areas that are functionally active together, and the analyses track this co-activity dynamically in time. Basically, functional connectivity among different networks or components from the fMRI data are calculated dynamically by computing the correlations among time-courses using a sliding time-window [14-17], as shown in Fig. 1. Implementation of the DFC is straightforward as shown by the pseudo code in Fig. 2. Initially, the subject number to be analyzed is determined by the first for-loop. Then all the combinations of time-courses which belong to the selected subjects are swept by the second for-loop. Correlations for the selected parts by sliding-windows with specified steps on time-courses are realized by the while-loop. Therefore, the resulting DFC coefficients provide the information about the dynamic evolution of correlation among two brain regions/networks.

![Figure 1. Dynamic Functional Connectivity Analysis](image)

Computational demand in DFC calculations is directly proportional to the number of fMRI subjects, the number of fMRI time-courses, and the number of combinations of distinct brain networks or regions. In a typical multi-subject fMRI experiment, number of subjects can range anywhere from a few to hundreds, and the number of brain networks or regions can range from a few to over a hundred. Two DFC analysis parameters, window width and step size, are other factors that the computational load is depends on. Smaller window width or smaller step size means an increase in the number of windows. Therefore, realization of the analysis may take a long time depending on the fMRI data size and the DFC parameters. The fMRI data that we utilized have 234 scans from 83 subjects, 30 brain networks/regions which results in 435 combinations, 316 time-points, and the number of windows ranged from 31 to 284 depending on various window width and window step size values we have used.

III. OpenCL Implementation

In comparison with the CPU architecture, GPU has a large number of computing units and these units can be utilized to run concurrent lightweight threads. In addition to various hardware supports for parallel computing, OpenCL is an important tool for utilizing computing capabilities of GPUs [4]. It provides a general framework based on C language for reducing the programming complexity of GPUs.

![Figure 3. OpenCL memory model](image)
A parallel program using OpenCL is made up of host code and kernel code that works on work items which are the basic processing element of OpenCL. As shown by Fig. 3, each work item has a private memory that can communicate with the local memory within a workgroup. Workgroups are also connected to global memory, and the data transfer between the host and the device can be realized through the global memory. Therefore, before data are transferred from host memory to private memory, it goes through global and local memory.

Parallelization with GPU involves using a host, where any parallel codes are placed within serial codes executed on the host. During execution, GPU interacts with the host to transfer analysis data and kernel codes into its memory and to make necessary pre- and post-operations, such as memory allocation and freeing as shown by Fig. 4. Before running the kernel using OpenCL, first the environment is set. Since OpenCL can be executed on various types of devices; in our case we execute it on a GPU. Then a memory area is allocated on device according to the input and output parameters of the algorithm. After building the kernel code, it is executed on parallel running work items. When all work items finish computing, processed data can be moved from device buffer to host memory for another operation.

Figure 4. Fundamental operations in OpenCL based computation

The kernel algorithm given by Fig. 5 depicts the basic idea for DFC analysis on GPU. In our case, we implemented each work item to run a DFC operation for a pair of selected time-courses. Each work item determines its allocated time-courses according to a global ID number of the resident work item. For this purpose, subject number and combination numbers are determined. According to the number of subjects and number of combinations and time points in each combination, the starting positions of the time-courses are determined from the input data, which, in our experiments, is a single dimensional array. After assigning the time-courses, then the algorithm given by Fig. 2 is evaluated. Afterwards, the computed results, which are the averages of sliding window operations, are written to the shared output array in global memory.

Computation complexity of parallel version of DFC is found as follows. Let N be subjects, T be number of time courses for each subjects, and M be data points in each time course. Two parameters of DFC are notated as W and S, window and step sizes, respectively. Given these notations, the serial version of DFC takes

\[
N \frac{T(T - 1)(T - W)}{2S} \approx e
\]

correlation calculations and complexity is \(O(n)\) assumedly. On the other hand, the parallel version we proposed takes

\[
\frac{ek}{P} + r
\]

where \(k\) is scaling factor, \(r\) is overhead due to data transfer and \(P\) is number of streaming processors in GPU. Therefore parallel version of DFC is \(O(n \log n)\).

Figure 5. Kernel structure

```c
#include

_kernel void computeDFC(double *A, double *X, int *otherParams, ...)
{
    // Get global id
    int id = get_global_id(0);

    // Determine Subject number and TCs
    int subjectNo = int(id/numOfCombs);
    int combsNo = id%numOfCombs;

    // Determine start positions for TCs
    posTC1 = combs[2*combsNo]*nTimePoints+subjectNo*nTCs;
    posTC2 = combs[2*combsNo+1]*nTimePoints+subjectNo*nTCs;

    // sliding windows
    for (k=0;k<M;k++)
    {
        // Determine sliding window position
        posTC1 = posTC1+stepSize;
        posTC2 = posTC2+stepSize;

        // mean of TCs
        meanA = 0;
        for (i=0;i<windowSize;i++)
            meanA = X[1][start1+i];

        // Perform other DFC computations
        ...
        // Accumulate correlations
        sum += corr;
    }
    // Save average of correlations to shared output
    C[idx] = sum/M;
}
```
IV. EXPERIMENTAL RESULTS

Sample fMRI data from 234 scans from 83 and 30 different time-courses of brain networks revealed by group spatial independent component analysis (ICA) were used. The number of combination pairs to be analyzed per subject is $C(30,2) = 435$. Each pair involves a number of sliding window operations (correlations) determined by the window width, window step size and the number of time points within a time-course. Total number of sliding window operations for all subjects and network pairs, for different number of window width and different number of window step size used in the experiments, are listed in Table I.

<table>
<thead>
<tr>
<th>Window width</th>
<th>1</th>
<th>2</th>
<th>4</th>
<th>8</th>
</tr>
</thead>
<tbody>
<tr>
<td>32</td>
<td>28908360</td>
<td>14454180</td>
<td>7227090</td>
<td>3562650</td>
</tr>
<tr>
<td>64</td>
<td>25651080</td>
<td>12825540</td>
<td>6412770</td>
<td>3155490</td>
</tr>
</tbody>
</table>

Each time-course in the fMRI data contains 316 time-points. Analyses were repeated for window widths of 32 and 64 and step sizes of 1, 2, 4, and 8. The results were obtained on a computer with AMD FX 8320 processor running at 3.5 GHz, has eight-core CPU, and 16MB total CPU cache memory. The video card used in the experiments is an NVidia GeForce GTX 660, which has 5 multiprocessors, 2GB total GPU memory, and 192-bit memory bus width. The memory size of the host computer is 2×4GB in unganged memory mode. The algorithm codes were written in Microsoft Visual C++ and run on Windows 7 64-bit operating system.

Computation times (running durations) using GPU for two test cases, with window widths 32 and 64, for sequential and GPU (OpenCL) implementations are shown in Fig. 6a and Fig. 6b, respectively. The results show that GPU implementation is faster when compared to the sequential implementation. There are also two results for GPU execution durations; one is for the first execution and the other is the second execution. First GPU execution of the program is a slightly slower than the second GPU execution, due to the initialization and allocation issues of the operating system [4]. These initializations and allocations are not needed after the first GPU execution. We included the GPU memory-allocation times and GPU memory-freeing times in our measurements of execution duration in Fig 5. Further improvements can be obtained by discarding these operations for the second and subsequent GPU runs for different parameters.

Fig. 7a and 7b shows the speed-up achievements for the two test cases in terms of speed-up for sequential and GPU implementations. GPU implementations show speed-up improvement ranging from 3x to 10x over sequential implementation. For both first and second GPU runs, the speed-up is more for smaller window step sizes, where the number of correlation operations are more, as mentioned in Table I. It is also notable that, second GPU run is much faster than the first one for higher window step size, which is expected since the overhead is more prominent for smaller number of correlation operations for higher window step sizes.

V. CONCLUSIONS

OpenCL simplifies the utilization of the GPU devices by providing efficient abstractions for computing with GPU. Based on data size and analysis parameters, execution of DFC analyses may result in high computational demand. Our experimental results show that OpenCL implementation is rather useful in reducing the execution time of DFC analysis of fMRI data and in the utilization of parallel computing.
capabilities of GPU devices. For our experiments obtained using NVidia GeForce GTX 660, speed-up varies from 3x to 10x over the sequential execution durations run on CPU. Further performance increases can be obtained for the same algorithm using high-end NVidia GPU devices. Future studies will be focus of other type of medical imaging problems such as skin lesions [44, 45] and pathological virtual slides [46, 47].

ACKNOWLEDGMENTS

This work was supported by National Institute of Drug Addiction (NIDA) 1R03DA031292-01, Office of Graduate Studies and Office of Research and Sponsored Programs at Texas A&M University at Commerce, Turkish Higher Education Institution (YÖK), and Research Fund of Sakarya University, Turkey.

REFERENCES


