

Low Cost FPGA Implementation of Fresnel Transform for Digital Holography

Huan-Yuan Chen¹, Song-Fu Shih¹, Wen-Jyi Hwang¹, Jung-Gen Wu¹, Chau-Jern Cheng²

¹Department of Computer Science and Information Engineering
National Taiwan Normal University
Taipei, 117, Taiwan

ray1308714@gmail.com; frank3130610@gmail.com
whwang@ntnu.edu.tw; jgwu@csie.ntnu.edu.tw

²Institute of Electro-Optical Science and Technology
National Taiwan Normal University
Taipei, 117, Taiwan
cjcheng@ntnu.edu.tw

Abstract - A novel FPGA architecture is presented for fast three-dimensional (3D) image reconstruction of digital holograms in this paper. The architecture is based on Fresnel transform for the 3D rendering. The implementation features low on-chip hardware resource consumption for the large size digital holograms. It uses the off-chip memory for buffering the intermediate results for subsequent computation. The adoption of the off-chip memory is realized in a network-on-chip (NOC) platform for efficient data accessing.

Keywords: FPGA; Digital Holography; Network on Chip; 3D Renderin

1. Introduction

Three dimensional (3D) digital imaging is gaining importance in applications such as metrology, biology, industrial inspection, and consumer electronics. Digital holography (DH) [6, 7] has been found to be effective for 3D imaging by recording the wavefront of a 3D object via charged-coupled devices (CCDs). The 3D image reconstruction of an object under observation can be carried out by digital diffraction computation on the recorded hologram. Different techniques can be used for diffraction computation, including Fresnel transform method, convolution method and angular spectrum method. Although these methods have been found to be effective, they share a common drawback of requiring high computational complexities. The fast Fourier transform (FFT) can be used to accelerate the computation. However, realtime 3D reconstruction may still be difficult for computers with limited computation capacities.

The general purpose graphic computation units (GPUs) can be used to accelerate the diffraction computation. A number of GPU-based implementations [1, 8, 9] for diffraction computation have been proposed. These implementations exploit the many-core capability of the GPUs to offer a significant enhancement of throughput at the expense of higher power consumption. The implementations therefore may not be well-suited for the mobile/embedded devices with strict limitation on power dissipation.

To reduce the power consumption, field programmable gate arrays (FPGAs) have been adopted to implement diffraction computation. The architectures in [2, 3, 10] are designed based on the convolution approach. The work in [4] implements the angular spectrum method in hardware. The study in [5] realizes the Fresnel transform architecture. Although these implementations have been found to be effective for providing high throughput computation, area costs may not be an important concern in the architectures. For example, circuits in [5] consumes large on-chip memory for efficient pipelining operations for Fresnel transform. The size of on-chip memory consumption may grow enormously for large digital holograms. Therefore, it would be difficult to implement diffraction computation circuits for large digital holograms in FPGA devices with limited hardware resources.

One way to solve the problem is to use the off-chip memory as the buffers. However, some designs focusing only on the stand alone FPGA implementation without considering the integration of the design into the system-on-chip (SOC) or network-on-chip (NOC) platforms. Therefore, off-chip memory access may not be efficient, especially for the column-wise one dimensional (1D) FFT operations where access of non-consecutive memory locations are required.

The objective of this paper is to present a novel FPGA-based coprocessor for the 3D image reconstruction. The circuit is able to operate real-time Fresnel transform without large hardware resource utilization. Moreover, the circuit is

capable of actively accessing the off-chip memory without the aids of other processors and/or direct memory access (DMA) controllers. Digital holograms with large sizes can then be processed by FPGA devices with limited hardware resources. The proposed circuit has been implemented in Altera DE4 development board. It acts as a hardware accelerator in the Qsys-based NOC platform. Experimental results reveal that the proposed architecture is effective for applications where high speed computation, low hardware costs, and low power consumption are desired.

2. Fresnel Transform for Digital Holography

The proposed architecture is able to perform diffraction computation for a DH microscopic imaging system. The resulting hologram, denoted by η , can be captured by CCDs and stored in digital computer. Given the hologram η , an object's image B in a plane parallel to the hologram plane at distance z can be reconstructed by Fresnel transform as follows:

$$B(r,s) = \frac{-j}{\lambda z} e^{j\frac{2\pi}{\lambda}z} e^{j\frac{\pi}{\lambda z}(r^2+s^2)} \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} \eta(p,q) e^{j\frac{\pi}{\lambda z}(p^2+q^2)} e^{-j\frac{2\pi}{\lambda z}(pr+qs)} dpdq \quad (1)$$

where λ is the wavelength of light source, and (p,q) and (r,s) are the coordinates on the hologram and image planes, respectively.

Since the hologram η is discretized in a CCD, the discrete representations of Fresnel transform is necessary for DH. Suppose the digital recording/sampling operations produce $N \times N$ samples for η with sampling interval Δ_f in both the x and y directions. Direct discretization of the Fresnel integral gives the following:

$$B_{u,v} = \frac{-j}{\lambda z} e^{j\frac{2\pi}{\lambda}z} e^{j\frac{\pi}{\lambda z}\Delta_g^2(u^2+v^2)} \sum_{x=0}^{N-1} \sum_{y=0}^{N-1} [\eta_{x,y} e^{j\frac{\pi}{\lambda z}\Delta_f^2(x^2+y^2)}] e^{-j2\pi(\frac{xu}{N} + \frac{yv}{N})} \quad (2)$$

where $\{B_{u,v}, 0 \leq u, v \leq N-1\}$ is the object's image in digital form,

$$\eta_{x,y} = \eta(x\Delta_f, y\Delta_f),$$

is the (x,y) -th sample of the discretized hologram η , $0 \leq x, y \leq N-1$, and

$$\Delta_g = \frac{\lambda z}{N\Delta_f},$$

is the inverse of Δ_f scaled by $\frac{\lambda z}{N}$.

3. The Proposed Architecture

The proposed architecture aims to compute eq.(2) by FPGA. There are three units in the proposed architecture: pre-transform unit, FFT unit, and post-transform unit.

The goal of pre-transform unit is to compute

$$\rho_{x,y} = \eta_{x,y} \times \mu_x \times \mu_y, \quad (3)$$

where

$$\mu_x = e^{j\frac{\pi}{\lambda z}\Delta_f^2 x^2}, \quad \mu_y = e^{j\frac{\pi}{\lambda z}\Delta_f^2 y^2}. \quad (4)$$

The FFT unit then takes Fourier transform on $\rho_{x,y}$. The result produced by FFT unit, termed $\tau_{u,v}$, is given by

$$\tau_{u,v} = \sum_{x=0}^{N-1} \sum_{y=0}^{N-1} \rho_{x,y} e^{-j2\pi(\frac{xu}{N} + \frac{yv}{N})}. \quad (5)$$

Define

$$\alpha = \frac{-j}{\lambda z} e^{j\frac{2\pi}{\lambda}z}, \quad \omega_u = e^{j\frac{\pi}{\lambda z}\Delta_g^2 u^2}, \quad \omega_v = e^{j\frac{\pi}{\lambda z}\Delta_g^2 v^2}. \quad (6)$$

By substituting eqs.(5)(6) into eq.(2), it follows that

$$\varepsilon_{u,v} = \alpha \times \omega_u \times \omega_v \times \tau_{u,v}. \quad (7)$$

Therefore, when $\tau_{u,v}$ is available, the post transform unit computes $\alpha \times \omega_u \times \omega_v \times \tau_{u,v}$ to find $\varepsilon_{u,v}$. In addition, $\phi_{u,v}$, the phase of $\varepsilon_{u,v}$, is also computed in the unit for hologram reconstruction.

As shown in Figure 1, all the three units are attached to the NOC system by Altera Qsys. Each unit has a dedicated network interface for accessing data outside the unit. In the NOC system, a NIOS II embedded processor, and a DRAM controller are also included. The NIOS II processor is used for coordinating the pipelining operations among these units. The DRAM controller is used for the off-chip memory accesses requested by the embedded processor and the three units.

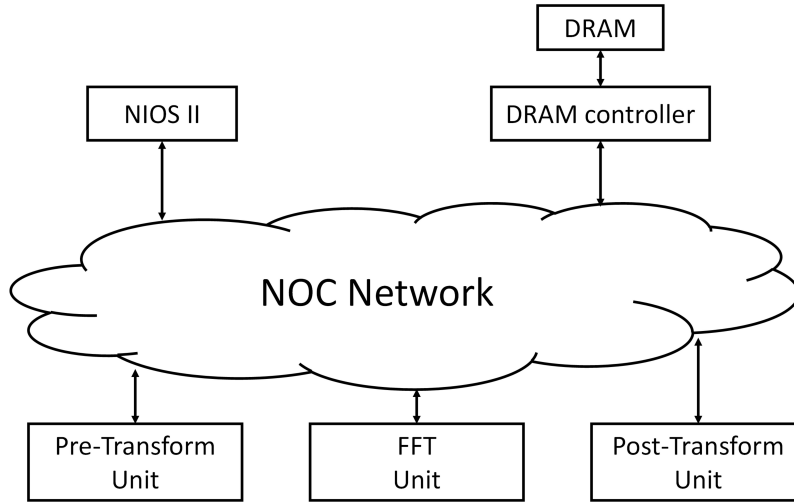


Fig. 1: The proposed architecture in the Qsys NOC platform for Fresnel transform.

3.1. Pre-transform Unit

The operations of pre-transform unit is based on eq.(3). Therefore, the unit involves the computation of μ_x , μ_y and multiplications. To accelerate the computation, the values of μ_x and μ_y can be pre-computed, and stored in tables. Because $0 \leq x, y \leq N - 1$, μ_x and μ_y only take N different values when λ , z , and Δ_f are known. Therefore, each table for the computation of μ_x and μ_y contains N entries.

Figure 2 shows the architecture of the pre-transform unit, which contains an address generation unit (AGU), a controller, two tables, two complex number multipliers, and two buffers. The AGU is responsible for the generation of addresses for reading $\eta_{x,y}$ from off-chip RAM to the read buffer. The controller then generates indices x and y for loading μ_x and μ_y values from the tables. The multipliers in the circuit are then used to compute $\rho_{x,y}$, which is first stored in the write buffer, and then sent back to off-chip RAM for subsequent FFT operations.

Because the multipliers in the architecture are for complex numbers with floating point format, it may be difficult for the multiplications to be completed in a single clock cycle. In our design, all the multipliers perform multiple clock cycles multiplications. To enhance the throughput, they are all fully pipelined. Therefore, in addition to indices generation, the controller also coordinates different components in the circuit for pipelining operations.

3.2. FFT Unit

The goal of FFT unit is to compute $\tau_{u,v}$ given by eq.(5). The FFT unit consists of an AGU, a controller, and a one-dimensional FFT (1D-FFT) module. To perform two-dimensional FFT (2D-FFT) using the 1D-FFT module, rows of the array $\{\rho_{x,y}, 0 \leq x, y \leq N - 1\}$ are loaded from off-chip RAM and operated *one at a time*. The FFT unit then writes the computational results directly back to the same row in the off-chip memory. After the row operations are completed, the column operations will proceed in the same manner. After the completion of all the column operations, the array stored in the on-chip RAM is $\{\tau_{u,v}, 0 \leq u, v \leq N - 1\}$, the 2D-FFT of $\{\rho_{x,y}, 0 \leq x, y \leq N - 1\}$.

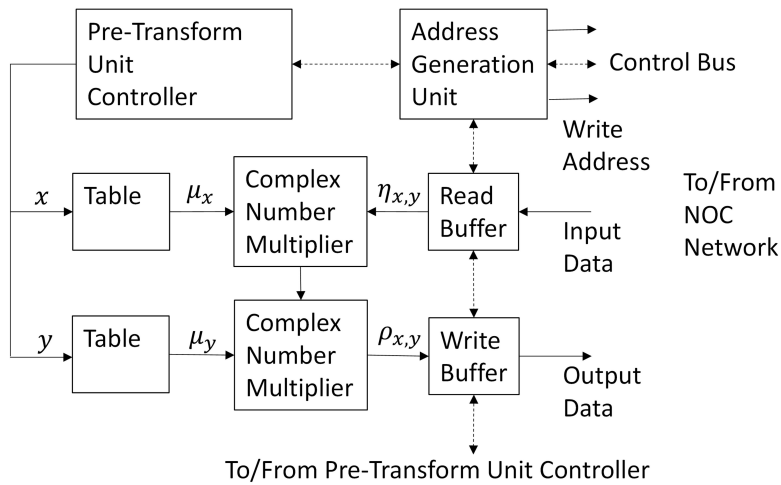


Fig. 2: The architecture of pre-transform unit.

We use Altera FFT MegaCore function to implement the 1D-FFT module. Because one row or one column is operated at a time, the transform length of the FFT is N . The 1D-FFT module has single data input and single data output. The module is fully pipelined. In addition, the input/output dataflow of the module is able to operate in streaming mode, allowing the continuous process of input data stream, as well as producing the continuous output data stream.

To perform 2D FFT using the 1D FFT module, the AGU in the FFT unit generates addresses for loading the source data from off-chip RAM and writing the results produced by 1D-FFT module to the off-chip RAM. Because the 1D-FFT has single data input and single data output, two addresses are generated in each clock cycle: one for loading data, and another for writing result. In addition, because the 1D FFT module is fully pipelined, and is able to operate in streaming mode, consecutive rows (or columns) can be loaded to the module in a seamless way. This can be accomplished by the employment of read and write buffers, each holding a row (or a column) of the source data (or results).

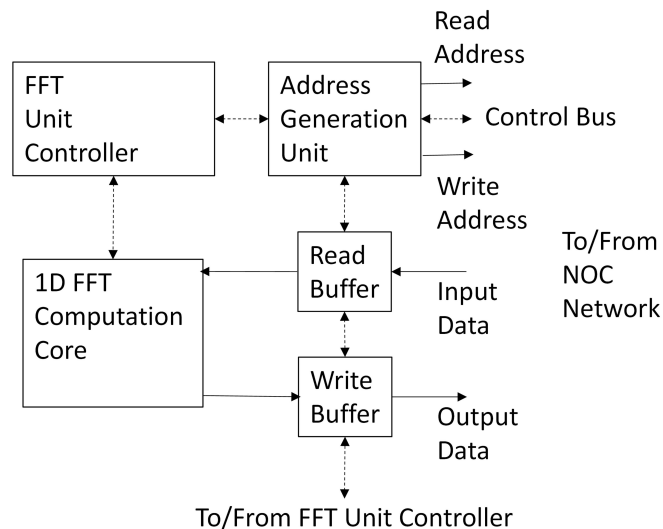


Fig. 3: The architecture of FFT unit.

3.3. Post-transform Unit

The post-transform unit is responsible for reconstructing the object image $\varepsilon_{u,v}$ using eq.(7). As depicted in Figure 3, the architecture of the post-transform unit is similar to that of the pre-transform unit, comprising of an AGU, a controller, two tables, three multipliers, two buffers. Additional arctan circuit for phase computation is also required.

In the post-transform unit, the tables are used to store the pre-computed values of ω_u and ω_v . Similar to the cases for μ_x and μ_y , because $0 \leq u, v \leq N - 1$, each table for the computation of ω_u and ω_v contains N entries. The controller in the post-transform unit operates in the similar fashion to that of the pre-transform unit. The controller produces indices (i.e., u and v values) for loading ω_u and ω_v values from the tables. The AGU generates addresses to the off-chip RAM for loading $\tau_{u,v}$. The result of multiplication, $\varepsilon_{u,v}$, is then used for computation of the phase $\phi_{u,v}$ using arctan circuit. After that, the phase is stored back to off-chip RAM.

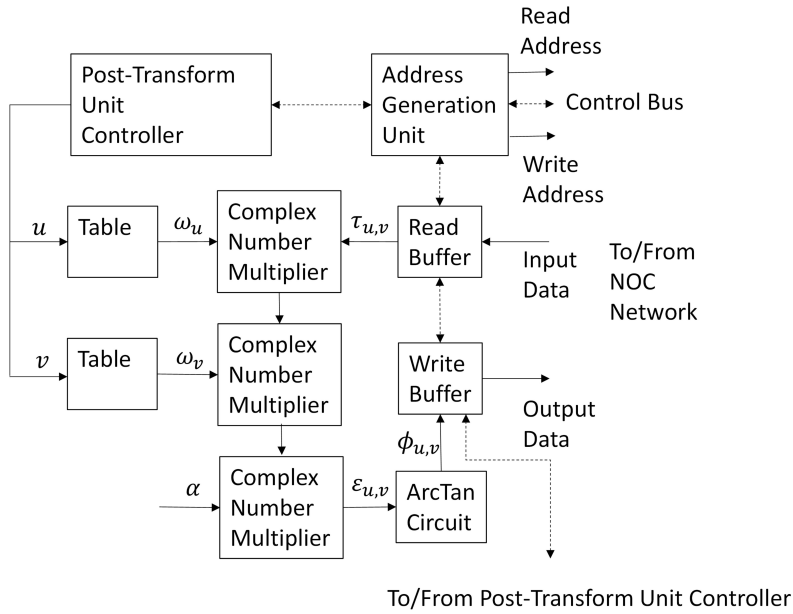


Fig. 4: The architecture of post-transform unit.

4. Experimental Results

Some experimental results of the proposed architecture are presented in this section. The design platform is Altera Quartus II [11, 12] with Qsys. The target FPGA device is Altera Stratix IV EP4SGX230. All the computations in the proposed architecture are the single-precision floating point computations. Therefore, the numbers of the proposed architecture are represented by IEEE 754 single-precision format, where the length of a number is 32 bits. The off-chip memory is 800 MHz DDR II memory with size of 1 Gbytes. The circuit operates at 200 MHz.

Table 1 shows the consumption of the hardware resources of the proposed architecture for holograms with dimensions 256×256 and 512×512 . There are four types of area costs considered in the experiment: adaptive logic lookup tables (ALUTs), dedicated logic registers, embedded memory bits, and DSP blocks. To reduce the consumption of general purpose hardware resources such as ALUTs and logic registers, the embedded logic registers and DSP blocks are used for the implementation of on-chip memory and arithmetic operators, respectively. Because the number of arithmetic operators is independent of dimensions, it can be observed from Table 1 that the number of DSP blocks is independent of the size of holograms. In addition, because only the read buffer and write buffer in each unit are implemented by the embedded memory blocks, the consumption of embedded memory bits is small.

The area costs of the entire NOC system are summarized in Table 2. It can be observed from Table 2 that the hardware utilization of the entire NOC system is small as compared with the hardware capacity of the target FPGA device. In fact, the

Table 1: The consumption of hardware resources of each unit in the proposed architecture for various image sizes.

Sizes	Hardware Resources	Pre-transform Unit	FFT Unit	Post-Transform Unit	Total
256×256	ALUTs	4892	14649	9752	29293
	Dedicated Logic Registers	8250	17139	12065	37454
	Embedded Memory Bits	29224	80836	31740	141800
	DSP Blocks	32	48	88	168
512×512	ALUTs	5348	19995	10532	35875
	Dedicated Logic Registers	8284	22124	12092	42500
	Embedded Memory Bits	61992	234270	56316	352584
	DSP Blocks	32	64	88	184

utilization of ALUTs, dedicated logic registers, embedded memory bits, and DSP blocks are 26 %, 32 %, 11 % and 15 % of those provided by Altera Stratix IV EP4SGX230, respectively.

To further evaluate the performance of the proposed architecture, comparisons with the work in [5] are made, as shown in Table 3. We see from the table that, as compared with its counterpart, the proposed architecture consumes significantly less embedded memory bits at the expense of higher latency. In fact, the embedded memory bits consumption is only 2% (352584 vs. 16936864) of that of the architecture in [5]. The significant reduction in embedded memory bits is beneficial for allowing the fast diffraction computation to be implemented in smaller FPGA devices with lower costs.

Finally, Figure 5 and Figure 6 show the 3D image reconstruction results of the proposed architecture. The image considered in the experiment are produced by the digital holographic microscopies (DHMs). It is the microlens with radius of curvature 120 microns. The error of 3D reconstruction is only 0.1 micron. Therefore, while having high speed computation, the architecture is also able to achieve high accuracy for 3D reconstruction

Table 2: The consumption of hardware resources of the NOC.

Components	ALUTs	Dedicated Logic Registers	Embedded Memory Bits	DSP Blocks
Proposed Arch.	35875	42500	352584	184
NIOS II Processor	1364	1519	64512	4
SDRAM Controller	4440	5062	159946	0
Total	46803	58034	1634974	188

Table 3: Comparisons of various architectures.

Arch.	ALUTs	Dedicated Logic Registers	Embedded Memory Bits	DSP Blocks	Latency
Proposed	35875	42500	352584	184	258.3 ms
[5]	27890	25048	16936864	144	2.1 ms

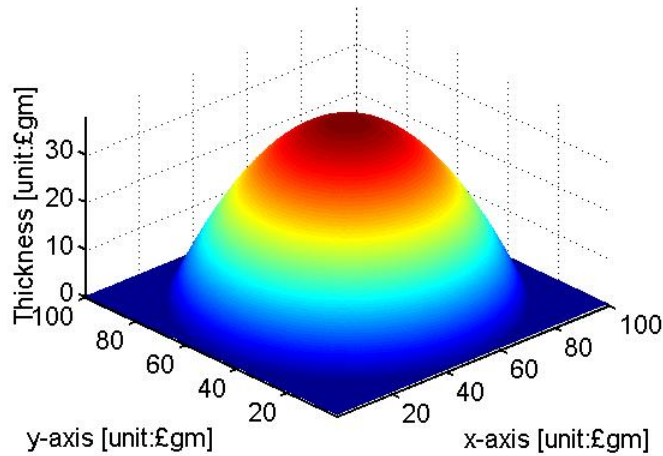


Fig. 5: The 3D reconstruction of a single microlens by the proposed architecture with image size 512×512 .

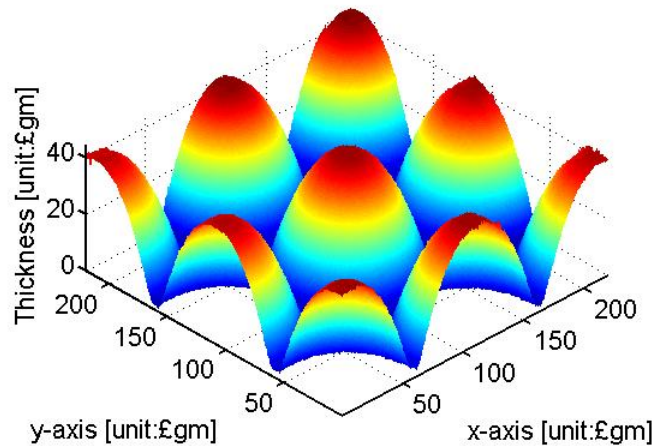


Fig. 6: The 3D reconstruction of a microlens array by the proposed architecture with image size 512×512 .

5. Conclusion

The experimental results reveal that the proposed architecture is well suited for low cost FPGA implementation of 3D image reconstruction for digital holograms. The architecture has low consumption of ALUTs, dedicated logic registers, embedded memory bits and DSP blocks. In particular, the architecture only consumes 352584 bits of embedded memory. The architecture is then beneficial for applications requiring both low hardware resource utilization and high speed computation.

References

- [1] M. Dogar, H. A. Ilhan and M. Ozcan, "Real-time, auto-focusing digital holographic microscope using graphics processors," *Review of Scientific Instruments*, vol. 84, 2013.
- [2] Y. Abe, N. Masuda, H. Wakabayashi, Y. Kazo, T. Ito, S. Satake, T. Kunugi and K. Sato, "Special purpose computer system for flow visualization using holography technology," *Opt. Express*, vol. 16, pp. 7686-7692, 2008.
- [3] N. Masuda, T. Sugie, T. Ito, S. Tanaka, Y. Hamada, S. Satake, T. Kunugi and K. Sato, "Special purpose computer system with highly parallel pipelines for flow visualization using holography technology," *Computer Physics Commun.*, vol. 181,

- pp. 1986-1989, 2010.
- [4] R. Veitch, D. C. Hendry and J. Watson, "Reconfigurable hardware applied to holographic reconstruction," in Proc. IEEE Oceans, Aberdeen, U.K., 2007.
 - [5] C. J. Cheng, W. J. Hwang, C. T. Chen and X. J. Lai, "Efficient FPGA-Based Fresnel Transform Architecture for Digital Holography," *IEEE Journal of Display Technology*, vol. 10, pp.272-281, 2014.
 - [6] P. Marquet, B. Rappaz, P. J. Magistretti, E. CuChe, Y. Emery, T. Colomb, and C. Depeursinge, "Digital holographic microscopy: a noninvasive contrast imaging technique allowing quantitative visualization of living cells with subwavelength axial accuracy," *Opt. Letters*, vol. 30, pp. 468-470, 2005.
 - [7] C. J. Mann, L. Yu, C. M. Lo and M. K. Kim, "High-resolution quantitative phase-contrast microscopy by digital holography," *Opt. Express*, vol. 13, pp.8693-8698, 2005.
 - [8] T. Shimobaba, T. Ito, N. Masuda, Y. Abe, Y. Ichihashi, H. Nakayama, N. Takada, A. Shiraki and T. Sugie, "Numerical calculation library for diffraction integrals using the graphic processing unit : the GPU-based wave optics library, *J. Opt. A: Pure Appl. Opt.*, vol. 10, 075308, 2008.
 - [9] Y. Pan, X. Xu, S. Solanki, X. Liang, R. B. A. Tanjung, C. Tan and T. C. Chong, "Fast CGH computation using S-LUT on GPU," *Opt. Express*, vol. 17, pp. 18543-18555, 2009.
 - [10] K. Kamegai, T. Kakue, T. Shimobaba, N. Masuda and T. Ito, "Simulation of the Special-Purpose Computer for High-Speed Imaging by Digital holography," *Proceedings of 3DSA*, pp. 4-9, 2013.
 - [11] Altera Corporation. (2016). FFT IP core user guide [Online]. Available: http://www.altera.com/en_US/pdfs/literature/ug/ug_fft.pdf
 - [12] Altera Corporation. Quartus Prime Pro Handbook, v.16.0.0 [Online]. Available: <http://www.altera.com/products/design-software/fpgadesign/quartus-prime/support.html>