System-on-Chip Design Using High-Level Synthesis Tools

Erdal Oruklu1*, Richard Hanley1, Semih Aslan2, Christophe Desmouliers1, Fernando M. Vallina3, Jafar Saniie1

1Department of Electrical and Computer Engineering, Illinois Institute of Technology, Chicago, USA
2Ingram School of Engineering, Texas State University, San Marcos, USA
3Xilinx Inc., San Jose, USA
Email: *erdal@ece.iit.edu

Received August 13, 2011; revised November 20, 2011; accepted November 30, 2011

ABSTRACT

This paper addresses the challenges of System-on-Chip designs using High-Level Synthesis (HLS). HLS tools convert algorithms designed in C into hardware modules. This approach is a practical choice for developing complex applications. Nevertheless, certain hardware considerations are required when writing C applications for HLS tools. Hence, in order to demonstrate the fundamental hardware design concepts, a case study is presented. Fast Fourier Transform (FFT) implementation in ANSI C is examined in order to explore the important design issues such as concurrency, data recursions and memory accesses that need to be resolved before generating the hardware using HLS tools. There are additional language constraints that need to be addressed including use of pointers, recursion and floating point types.

Keywords: System Level Design; High Level Synthesis; Field Programmable Gate Arrays; Fourier Transform

1. Introduction

In the past decade, there has been a substantial increase in the level of hardware abstraction that High-Level Synthesis (HLS) [1-5] tools offer, which has made designing a complete System-on-Chip (SoC) much more practical. By designing at the system level, it has become possible for hardware engineers to avoid gate-level semantics. HLS tools work by taking applications written in a subset of ANSI C, and translating it into a Register Transfer Level (RTL) module for Application-Specific Integrated Circuit (ASIC) or Field Programmable Gate Arrays (FPGAs) chip design. The design workflow requires knowledge of both software to write C applications and hardware to parallelize tasks, resolve timing and memory management issues. There has been significant previous work that discusses how to teach RTL concepts to students and design simple applications for SoCs [6,7]. Nevertheless, the learning curve for software engineers is relatively high since they need to use Hardware Descriptive Languages (HDL) such as Verilog and VHDL. By using HLS tools, software engineers can use their programming skills along with hardware knowledge to create complex embedded hardware/software co-design systems.

To demonstrate the critical hardware and software design issues, a Fast Fourier Transform (FFT) [8] case study is used as a guideline. In order to generate hardware modules satisfying predefined constraints such as throughput and area, different modifications of the code required by HLS tools are presented step by step.

Section 2 of this paper provides a brief background of HLS tools, and the current pedagogical techniques. Section 3 presents an introduction of the FFT algorithm along with a software implementation. This software based FFT is then deconstructed in Section 4, where a fully synthesizable product is created. Section 5 analyzes the different results that can be produced depending on the constraints selected by the user such as speed, area, throughput, and targeted system.

For this paper, all designs are targeted for the Xilinx Virtex-5 FPGA platform [9] using the HLS tool called PICO, provided by Synfora Inc [10] (currently known as Synphony C Compiler by Synopsys [11]). However, the different code modifications presented in this paper are applicable to other HLS tools such as AutoESL [12] which targets primarily Xilinx FPGAs with architecture aware synthesis and Catapult C [13] which provides full-chip high-level synthesis for both ASIC and FPGA devices and automatic RTL verification.

2. High Level Synthesis Tools

This section outlines the important concepts that software developers need to know before entering the field of HLS. There is a special emphasis on how these concepts differ from the contemporary software environment familiar to the software engineers. Design of SoCs has historically
been accomplished using Hardware Descriptive Languages such as VHDL or Verilog. Each expression in HDL represents a group of gates that operate in parallel, as opposed to machine instructions executed sequentially. This concept of instruction level parallelism is one of the first major hurdles when introducing hardware concepts.

Once an RTL module is designed, it can be compiled and simulated. The simulation is done by creating a series of pre-defined inputs, known as a testbench, and recording the outputs. If a module passes the simulation then a low level implementation can be created. This low level implementation then enters the verification process to ensure that all timing dependencies are met. In practice, simulating and verifying an implementation can take 50% - 60% of the development time, increasing the time-to-market (TTM) [14]. By automating the simulation and verification process, it is possible to greatly reduce the development time.

Integration of HLS tools into the FPGA or ASIC design flow, as shown in Figure 1, allows software designers to build hardware modules and speed up the TTM significantly. During the generation process of an RTL module from a software implementation, simulation and verification are done automatically by using a formal proof provided during the initial steps. Subsequently, by using synthesis tools, the RTL module is implemented and timing verification is done. An independent evaluation of HLS tools for Xilinx FPGAs has been done by Berkeley Design Technology [15]. It shows that using HLS tools with FPGAs can improve the performances of an application by an order of magnitude compared to DSPs. Moreover, this study shows that for a given application, HLS tools will achieve similar results compared to hand-written HDL code with a shorter development time.

HLS software based approach for simulation and verification is made possible by using SystemC, a language developed by Synopsys, University of California Irvine, Frontier Design and IMEC. SystemC is an extension of C++ that provides additional libraries to design an embedded system. The first version was released in 1999 and in 2005 it became IEEE standardized SystemC [16, 17] as the IEEE-1666-2005. These additional libraries make it possible to specify the hardware and software components in an embedded system using one unified paradigm and to generate testbenches.

Focusing further on HLS, the design flow is shown in Figure 2. Each module of a system is implemented using high level languages such as C, C++, Java, or Matlab [2,18], which can then be tested automatically with testbenches provided by the user. After verification of the complete system, the user can specify in the HLS tool which modules will be converted into hardware accelerators in order to speed up the application. This is one of the core elements of hardware/software co-design that software developers need to understand. There are inherent restrictions in the HDLs that are mirrored in the HLS tool. Therefore, the emphasis for teaching HDL to software developers is on its constraints and how it affects the HLS tools.

After generation of the hardware modules along with testbenches, the system is verified and can be implemented using synthesis tools.

This paper, as mentioned earlier, focuses on designing a Fast Fourier Transform. The concept of HLS is presented by using PICO (Program-In Chip-Out) Extreme from Synfora [10,19,20] to generate the RTL code of an FFT. To be specific, PICO takes a C-based description of an algorithm and generates: performance-driven device-dependent synthesizable RTL code, testbench files, application drivers, simulation scripts as well as SystemC based Transaction Level Models (TLM) [3,17,18,21]. PICO design flow is shown in Figure 3. With integration

![Figure 1. FPGA high level synthesis block diagram.](image1)

![Figure 2. High level synthesis (HLS) design flow.](image2)
of the PICO design tools to their FPGA flow, designers can create complex hardware [20] sub-systems from sequential untimed C algorithms. It allows designers to explore programmability, performance, power, area and clock frequency. This is achieved by providing a comprehensive and robust verification and validation environment. PICO is designed to explore different type of parallelism and will choose the optimal one transparently. Results in terms of throughput and area are given along with detailed reports that will help the user for code optimization. When the synthesized performances are satisfactory, RTL code is generated and can be implemented in the targeted platform. Because the testing is done in C, the verification time of the RTL module can be significantly reduced [20].

3. Fast Fourier Transform

In most cases, the first step when using an HLS tool is to create a reference implementation, which is used to verify the synthesized product. The reference code itself can be compiled using any C compiler, and is purely software-based. This means that no new concepts have to be taught, making the reference implementation a logical starting point when using HLS.

When creating the reference code for FFT, there are few issues that need to be addressed when using HLS tools. The first issue is that arithmetic operations such as division can significantly decrease the performance of the design, and therefore should be avoided whenever possible. Nevertheless, division by a power of two is considered as a bit shift operation and hence can be used at no cost. The second issue, and more fundamental issue, is that pointers and recursion are not supported by the current HLS tools due to the fact that those concepts are purely software and can’t be applied to hardware designs. Finally, HLS tools may not have the capability to synthesize software functions such as cosine and sine. As a result of these constraints, the reference code included in this section does not use divisions, is completely iterative, and has no pointer variables. However, before going into the details of the implementation, the mathematical background of the FFT is presented.

3.1. FFT Algorithm

The Fourier transform takes a signal $x$ in time $t$ and transforms it into a function $X$ in frequency $\omega$:

$$X(\omega) = \int_{-\infty}^{\infty} x(t) e^{-j\omega t} \, dt$$  \hspace{1cm} (1)

The transform can be computed using a Discrete Fourier Transform (DFT).

$$X_k = \sum_{n=0}^{N-1} x_n e^{-2\pi j nk/N},$$  \hspace{1cm} (2)

where $k = 0, \ldots, N-1$.

The direct realization of DFT algorithm requires $O(N^2)$ computational time. To make this computation faster, an entire class of Fast Fourier Transforms (FFT) were developed [8]. However, in this paper a radix-2 FFT decimated in time is implemented. This algorithm divides the original DFT into two DFTs with half the length (i.e. decimation). The first step in decimation is shown below:

$$X_k = \sum_{m=0}^{N-1} x_{2m} e^{-2\pi j (2^k) m/N} + \sum_{m=0}^{N-1} x_{2m+1} e^{-2\pi j (2^k) m/N}$$  \hspace{1cm} (3)

Then the algorithm is recursively applied to each term until each DFT’s length is 1. This recursive deconstruction of the DFT makes the computational time of $O(N\log(N))$ [8].

3.2. Software Implementation of the FFT

In Figure 4, a 16-point radix-2 FFT is shown. A signal is
The particular implementation chosen for this reference FFT was provided by [22]. The exact code used is shown in Figure 5. \(N\) represents the length of the FFT and must be a power of 2. Before using the function \(fft\_ref\), the function \(table\_setup\) must be executed in order to compute the twiddle factors and store them in the array \(fix\_float\). The FFT of an input \(z\) can then be executed. The first phase is the bit-reverse operation where the input data are rearranged as shown in Figure 4. Then, for each pass, the butterfly operations are performed until the FFT is completed. In the next section this code will be made fully synthesizable by applying four modifications to it.

4. Code Modification for HLS

The objective of this section is to generate the hardware of a FFT block based on the reference C code using HLS tools. Multiple modifications are needed in order to generate an optimal hardware in terms of resource usage and throughput. As an example, we generate an 8-bit 1024-point radix-2 FFT. The output is on 18 bits and will be available in natural order. The size of the data width inside the FFT has been chosen so that the HLS FFT gives the same results as the Xilinx FFT core [23].

4.1. Floating Point to Fixed Point Implementation

Since the reference C code is using floating point numbers, a fixed-point library is needed. For example, PICO, the HLS used in this demonstration, provides such library. The PICO fixed-point arithmetic library derives its semantics from the SystemC fixed-point library and it supports signed and unsigned arithmetic operations. Hence, the previous floating point complex structure must be modified as followed:

```c
typedef struct
{
  float x;
  float y;
} s_complex;
```

Moreover, in order to perform the butterfly operation, the \(W^k_N\) terms need to be calculated. Since we assume that HLS library does not support cosine and sine functions, the twiddle factors are pre-computed and stored in a table using the code below:

```c
#include "fft.h"
#include <math.h>
#define pi2 ((double)6.28318530717958647692528676655901
extern s_complex fix_float[N / 2];
void table_setup(void)
{
  double a = 0.0;
  double e = -pi2 / N;
  float cos_val, sin_val;
  int i;
  for(i = 0; i < N / 2; i++)
  {
    cos_val = cos(a);
    sin_val = sin(a);
    fix_float[i].x = cos_val;
    fix_float[i].y = sin_val;
    a = a + e;
  }
}
```

The butterfly operation requires complex number arithmetic additions and multiplications. Because of the programming constraints placed on the reference code, most complex number libraries are not usable. Hence, this reference code uses its own complex number representation shown below:

\[
W^k_N = e^{-2\pi k/N} \quad \text{(called the Twiddle factor)}
\]

\[
F = f + W^k_N \times g
\]

\[
G = f - W^k_N \times g \tag{4}
\]

The objective of this section is to generate the hardware of a FFT block based on the reference C code using HLS tools. Multiple modifications are needed in order to generate an optimal hardware in terms of resource usage and throughput. As an example, we generate an 8-bit 1024-point radix-2 FFT. The output is on 18 bits and will be available in natural order. The size of the data width inside the FFT has been chosen so that the HLS FFT gives the same results as the Xilinx FFT core [23].

4.1. Floating Point to Fixed Point Implementation

Since the reference C code is using floating point numbers, a fixed-point library is needed. For example, PICO, the HLS used in this demonstration, provides such library. The PICO fixed-point arithmetic library derives its semantics from the SystemC fixed-point library and it supports signed and unsigned arithmetic operations. Hence, the previous floating point complex structure must be modified as followed:

```c
typedef struct
{
  float x;
  float y;
} s_complex;
```
FFT is computed using 22-bit data width with 18 bits for the integer part and 4 bits for the fractional part. Rounding and saturation configuration is used. The effect of the number of bits allocated to the fractional part on the precision and resource usage of the FFT HLS is presented in Section V. The twiddle factors are pre-calculated with a precision of 16 bits and stored in an array eliminating the need of trigonometric functions.

### 4.2. Input Array to Stream of Input Data

In the reference C code, the input data are passed to the function as an array. This will be translated into memory accesses by the HLS tool which is not optimal for hardware implementation. Hence, a stream of input data is used. PICO supports two types of streams: external and internal. External streams are used to stream data from/to global memory and/or other blocks in the system. Internal streams are used to stream data between loops within a multi-loop accelerator designed by PICO. In PICO, streams are specified using explicit procedure calls that transmit a scalar value to an output stream or receive a scalar value from an input stream. These procedures are converted into special opcodes that receive (transmit) data from (to) actual streams. For the FFT application, four streams are needed: input/output streams for real and imaginary parts:

![Figure 5. FFT reference C code.](image)

```c
typedef struct
    floatP x;
    floatP y;
}s_complexP;

typedef pico::s_fixed<22,18,pico::S_RND,pico::S_SAT,0> floatP;

int fft_ref (int n, int m, s_complex + z)
{
    int i, j, k, n1, n2;
    float c, s, r1, r2;
    j = 0;
    n2 = n / 2;
    for (i = 1; i < n - 1; i += 2)
        if (j >= n1) {
            j = j - n1;
            n1 = n2 / 2;
        }
        j = j + n1;
    s = fix_float((index) * y);
    index = index + incr;
    if (index >= N/2) { (index = 0; )
        float temp,temp1;
        for (k = j; k < n; k = k + n2) {
            r1 = z[k + n1].x * c + z[k + n1].y * s;
            temp = z[k + n1].y - r1;
            r2 = temp;
            r2 += temp;
            z[k + n1].x = z[k].x - r1;
            z[k + n1].y = z[k].y + r1;
            z[k].y = z[k].x + r1;
        }
    }
    return 0;
}
```

Finally after the FFT is computed, the unloading phase is performed:

```c
for(p = 0;p < N; p += 1)
    pico_stream_output_yout(z[p].y);
```

### 4.3. Bit-Reverse Operation

If we look at the reference C code, the next step would be the bit-reverse stage; this operation takes 1024 cycles. However, it can be integrated in the radix-2 FFT block, hence reducing the total number of cycles required to perform the calculations. This can be done using the `bit_swap` function:

```c
for(h = 0,h < N;k += 1)
    z[h].x = (floatP)pico_stream_input_zin();
    z[h].y = (floatP)pico_stream_input_zin();
```

PICO synthesizes a FIFO (within the RTL) for each internal and external stream in the code. Different parameters such as the length of the FIFO can be configured using pragmas. The first step of the FFT will be the loading phase where input data are stored into a RAM called z as shown below:

```c
for(h = 0,h < N;k += 1)
    z[h].x = (floatP)pico_stream_input_zin();
    z[h].y = (floatP)pico_stream_input_zin();
```

```c
for(p = 0;p < N; p += 1)
    pico_stream_output_yout(z[p].y);
```
In this function, we use the pragma unroll to specify to the HLS tool to unroll the loop and hence parallelize the operations to speed-up the process. This function is used to calculate the new address when performing the butterfly calculation on z as shown below:

```c
for (k = j; k < n; k = k + n2) {
    for (k = j; k < n; k = k + n2) {
        s_complex tmpz_n1,tmpz;
        s_complex outz_n1,outz;
        tmpz = z[k];
        tmpz_n1 = z[k+n1];
        if (%(k)%2) {
            tmpz = z[k];
            tmpz_n1 = z[k+n1];
        } else {
            tmpz = z[k];
            tmpz_n1 = z[k+n1];
        }
        t1 = (tmpz_n1.x + c - tmpz_n1.y * s);
        t2 = (tmpz_n1.x + s - tmpz_n1.y * c);
        outz_n1.x = (tmpz.x + c*1); 
        outz_n1.y = (tmpz.y - t2); 
        outz.x = (tmpz.x + t1); 
        outz.y = (tmpz.y + t2); 
        if (%(k)%2) {
            z[k] = outz;
            z[k+n1] = outz_n1;
        } else{
            z[k] = outz;
            z[k+n1] = outz_n1;
        }
    }
    return 0;
}
```

4.4. Memory Access Reduction

Each array of data in the reference C code will be implemented as a RAM by the HLS tool. We can see that multiple accesses of z are done which is not suitable for hardware implementation since only single or dual port RAMs/ROMs are available. In order to resolve this problem and obtain better performances, the first step is to use temporary variables. This step is shown below:

```c
for (k = j; k < n; k = k + n2) {
    for (k = j; k < n; k = k + n2) {
        s_complex tmpz_n1,tmpz;
        s_complex outz_n1,outz;
        tmpz = z[k];
        tmpz_n1 = z[k+n1];
        if (%(k)%2) {
            tmpz = z[k];
            tmpz_n1 = z[k+n1];
        } else {
            tmpz = z[k];
            tmpz_n1 = z[k+n1];
        }
        t1 = (tmpz_n1.x + c - tmpz_n1.y * s);
        t2 = (tmpz_n1.x + s - tmpz_n1.y * c);
        outz_n1.x = (tmpz.x + c*1); 
        outz_n1.y = (tmpz.y - t2); 
        outz.x = (tmpz.x + t1); 
        outz.y = (tmpz.y + t2); 
        if (%(k)%2) {
            z[k] = outz;
            z[k+n1] = outz_n1;
        } else{
            z[k] = outz;
            z[k+n1] = outz_n1;
        }
    }
    return 0;
}
```

By integrating the modifications presented in this section to the reference C code given in Figure 5, the HLS implementation of the FFT can be obtained as shown in Figure 6.

5. Hardware Synthesis Results

The HLS tool offers different configurations that will have an impact on the hardware generated. For example, the user can specify the desired frequency that may or may not be achieved by the tool depending on the system targeted and the complexity of the C code. As seen in

```
for (k = j; k < n; k = k + n2) {
    for (k = j; k < n; k = k + n2) {
        s_complex tmpz_n1,tmpz;
        s_complex outz_n1,outz;
        tmpz = z[k];
        tmpz_n1 = z[k+n1];
        if (%(k)%2) {
            tmpz = z[k];
            tmpz_n1 = z[k+n1];
        } else {
            tmpz = z[k];
            tmpz_n1 = z[k+n1];
        }
        t1 = (tmpz_n1.x + c - tmpz_n1.y * s);
        t2 = (tmpz_n1.x + s - tmpz_n1.y * c);
        outz_n1.x = (tmpz.x + c*1); 
        outz_n1.y = (tmpz.y - t2); 
        outz.x = (tmpz.x + t1); 
        outz.y = (tmpz.y + t2); 
        if (%(k)%2) {
            z[k] = outz;
            z[k+n1] = outz_n1;
        } else{
            z[k] = outz;
            z[k+n1] = outz_n1;
        }
    }
    return 0;
}
```
this section, increasing the frequency will increase the resources of the hardware generated by the HLS tool. The throughput (number of FFTs that can be done in one second) can also be specified. In order to achieve a high throughput, the HLS tool will parallelize tasks; hence increasing the hardware resources. Finally, the user can specify to implement arrays using block RAMs or look-up tables (LUTs). Hardware implementation results are obtained using Xilinx ISE 12.1 software with either speed or area optimization for Virtex-5 FPGA. The twiddle factors have been implemented using LUTs but can be also implemented using RAMs. By doing this, it will reduce the total number of slices LUTs but increase the number of blocks RAM/FIFO.

Table 2 shows the hardware usage of the HLS implementation of FFT with 22-bit data width for different targeted frequencies.

One can see a significant increase in terms of logic slices for 150 MHz operational frequency. This is due to the fact that we have selected optimization for speed in ISE in order to achieve the desired operational frequency after place and route. For frequencies lower than 150 MHz, optimization in terms of area has been selected. For frequencies from 50 MHz to 150 MHz, the total number of clock cycles achieved by PICO to perform the 1024-point FFT is 7168 but for 175 MHz it is increased to 12288 clock cycles. 7168 clock cycles is the minimum latency that can be obtained and is calculated as follow:

\[
\text{latency} = \text{loading} + \text{FFT} + \text{unloading}
\]

\[
\text{latency} = N + \frac{N}{2} \cdot \log_2 (N) + N
\]

\[
\text{latency} = 1024 + 512 \cdot 10 + 1024 = 7168 \text{ clock cycles}
\]

Table 1. FFT hardware usage for different frequencies.

<table>
<thead>
<tr>
<th>Targeted frequency</th>
<th>Resource usage</th>
<th>Achieved frequency</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Slices Registers</td>
<td>Slices LUTs</td>
</tr>
<tr>
<td>50 MHz</td>
<td>749</td>
<td>1700</td>
</tr>
<tr>
<td>75 MHz</td>
<td>765</td>
<td>1769</td>
</tr>
<tr>
<td>100 MHz</td>
<td>926</td>
<td>1967</td>
</tr>
<tr>
<td>125 MHz</td>
<td>1042</td>
<td>1714</td>
</tr>
<tr>
<td>150 MHz</td>
<td>1546</td>
<td>2004</td>
</tr>
<tr>
<td>175 MHz</td>
<td>1380</td>
<td>1849</td>
</tr>
<tr>
<td>270 MHz</td>
<td>1457</td>
<td>1989</td>
</tr>
</tbody>
</table>
Table 2. FFT hardware usage for different throughputs.

<table>
<thead>
<tr>
<th>Targeted throughput</th>
<th>Resource usage</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Slices</td>
</tr>
<tr>
<td>20926</td>
<td>1546</td>
</tr>
<tr>
<td>12207</td>
<td>1351</td>
</tr>
<tr>
<td>8616</td>
<td>1186</td>
</tr>
<tr>
<td>6658</td>
<td>1161</td>
</tr>
</tbody>
</table>

Figure 7. Relative error for different bit size for the fractional part.

Table 3. FFT hardware usage for different fractional sizes.

<table>
<thead>
<tr>
<th>Fractional part number of bits</th>
<th>Resource usage</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Slices</td>
</tr>
<tr>
<td>0</td>
<td>767</td>
</tr>
<tr>
<td>2</td>
<td>809</td>
</tr>
<tr>
<td>4</td>
<td>926</td>
</tr>
<tr>
<td>6</td>
<td>968</td>
</tr>
<tr>
<td>8</td>
<td>1188</td>
</tr>
<tr>
<td>10</td>
<td>1279</td>
</tr>
<tr>
<td>12</td>
<td>1400</td>
</tr>
</tbody>
</table>

6. Conclusion

In this paper, we have presented hardware considerations that software engineers need to apply when designing hardware modules using HLS tools. As a demonstration, the implementation of a radix-2 FFT unit has been presented. We have shown the different steps to achieve an optimized C code for HLS tools based on an ANSI C code. Results of the generated FFT for a Virtex-5 FPGA have been presented. FFT has a broad range of applications in digital signal processing, and multimedia. It is a key component that determines most of the design metrics in many signal processing communication applications. HLS tools facilitate complex algorithms to be realized at a higher level. They can reduce the design cycle significantly while successfully generating results very close to handmade HDL design.

7. Acknowledgements

The authors would like to thank the Xilinx, Inc. (www.xilinx.com) and Synopsys (www.synopsys.com) for their valuable support.

REFERENCES

[14] P. Avss, S. Prasant and R. Jain, “Virtual Prototyping In-


