Research Article

Pipeline FFT Architectures Optimized for FPGAs

Bin Zhou,1,2 Yingning Peng,1 and David Hwang2

1 Department of Electronic Engineering, Tsinghua University, Beijing 100084, China
2 Department of Electrical and Computer Engineering, George Mason University, 4400 University Drive, Fairfax, VA 22030, USA

Correspondence should be addressed to David Hwang, dhwang@gmu.edu

Received 28 February 2009; Accepted 23 June 2009

This paper presents optimized implementations of two different pipeline FFT processors on Xilinx Spartan-3 and Virtex-4 FPGAs. Different optimization techniques and rounding schemes were explored. The implementation results achieved better performance with lower resource usage than prior art. The 16-bit 1024-point FFT with the R22SDF architecture had a maximum clock frequency of 95.2 MHz and used 2802 slices on the Spartan-3, a throughput per area ratio of 0.034 Msamples/s/slice. The R4SDC architecture ran at 123.8 MHz and used 4409 slices on the Spartan-3, a throughput per area ratio of 0.028 Msamples/s/slice. On Virtex-4, the 16-bit 1024-point R22SDF architecture ran at 235.6 MHz and used 2256 slice, giving a 0.104 Msamples/s/slice ratio; the 16-bit 1024-point R4SDC architecture ran at 219.2 MHz and used 3064 slice, giving a 0.072 Msamples/s/slice ratio. The R22SDF was more efficient than the R4SDC in terms of throughput per area due to a simpler controller and an easier balanced rounding scheme. This paper also shows that balanced stage rounding is an appropriate rounding scheme for pipeline FFT processors.

Copyright © 2009 Bin Zhou et al. This is an open access article distributed under the Creative Commons Attribution License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

1. Introduction

The Fast Fourier Transform (FFT), as an efficient algorithm to compute the Discrete Fourier Transform (DFT), is one of the most important operations in modern digital signal processing and communication systems. The pipeline FFT is a special class of FFT algorithms which can compute the FFT in a sequential manner; it achieves real-time behavior with nonstop processing when data is continually fed through the processor. Pipeline FFT architectures have been studied since the 1970’s when real-time large scale signal processing requirements became prevalent. Several different architectures have been proposed, based on different decomposition methods, such as the Radix-2 Multipath Delay Commutator (R2MDC) [1], Radix-2 Single-Path Delay Feedback (R2SDF) [2], Radix-4 Single-Path Delay Commutator (R4SDC) [3], and Radix-22 Single-Path Delay Feedback (R22SDF) [4]. More recently, Radix-22 to Radix-2k SDF FFTs were studied and compared in [5]; in [6] an R22SDF was implemented and shown to be area efficient for 2 or 3 multipath channels. Each of these architectures can be classified as multipath or single-path. Multipath approaches can process M data inputs simultaneously, though they have limitations on the number of parallel data-paths, FFT points, and radix. This paper focuses on single-path architectures.

From the hardware perspective, Field Programmable Gate Array (FPGA) devices are increasingly being used for hardware implementations in communications applications. FPGAs at advanced technology nodes can achieve high performance, while having more flexibility, faster design time, and lower cost. As such, FPGAs are becoming more attractive for FFT processing applications and are the target platform of this paper.

The primary goal of this research is to optimize pipeline FFT processors to achieve better performance and lower cost than prior art implementations. In this paper, two comparative implementations (R4SDC and R22SDF) of pipeline FFT processors targeted towards Xilinx Spartan-3 and Virtex-4 FPGAs are presented. Different parameters such as throughput, area, and SQNR are compared.

The rest of the paper is organized as follows. Section 2 discusses the methodology used to select the two architectures. Section 3 describes the implementation tools and optimization methods used to improve performance and reduce resource utilization. Section 4 explains the balanced rounding schemes that were implemented and their impact on the signal-to-quantization noise ratio (SQNR). Section 5 presents the results, and Section 6 presents some brief conclusions.
2. Pipeline FFT Architectures

2.1. Architecture Selection. The major characteristics and resource requirements of several pipeline FFT architectures are listed in Table 1. Computational efficiency is measured by resource utilization percentage—how often the resources are in an active state versus an idle state. As shown in the table, the radix-4 Single-Path Delay-Commutator (R4SDC) and radix-22 Single-Path Delay Feedback (R22SDF) architectures provide the highest computational efficiency and were selected for implementation. The R4SDC architecture is appealing due to the computational efficiency of its addition; however the controller design is complex. The R22SDF architecture has a simple controller but a less efficient addition scheme. These designs are both radix-4 and scalable to an arbitrary FFT size \( N \) (\( N \) is a power of 4).

2.2. R4SDC Architecture.

R4SDC Algorithms. The R4SDC was proposed by Bi and Jones [3] and uses an iterative architecture to calculate the radix-4 FFT. The key to the algorithm is splitting the FFT into different stages by using different radices. In this paper, the radix is always 4.

The derivation starts from the fundamental DFT equation for an \( N \)-point FFT:

\[
X(k) = \sum_{n=0}^{N-1} x(n) W_N^{nk} \quad (k = 0, 1 \ldots N - 1; W_N = e^{-j(2\pi/N)}) \tag{1}
\]

\( N \) can be represented as composite of \( v \) numbers \( N = r_1 r_2 \cdots r_v \) and defined as

\[
N_t = \frac{N}{r_1 r_2 \cdots r_t}, \quad 1 \leq t \leq v - 1, \tag{2}
\]

where \( t \) is the stage, and \( r_t \) is the stage radix. After putting (2) into (1) and applying the relationship \( W_{N,N_t}^{N_{t,1}} = W_{N_{t,1}}^{N_{t,1}} \), (1) becomes

\[
X(k) = \sum_{q_1=0}^{N_{t-1}} \sum_{p=0}^{r_{t-1}} x(N_1 p + q_1) W_{r_{t-1}}^{p k}. \tag{3}
\]

Indeces \( k_1 \) and \( m_1 \) can be defined by \( k = r_1 k_1 + m_1 \), where \( 0 \leq k_1 \leq N_1 - 1, \ 0 \leq m_1 \leq r_1 - 1 \). Equation (3) becomes

\[
X(r_1 k_1 + m_1) = \sum_{q_1=0}^{N_{t-1}} x_1(q_1, m_1) W_{N_{t,1}}^{q_1 k_1}, \tag{4}
\]

\[
x_1(q_1, m_1) = W_{N_{t,1}}^{q_1 m_1} \sum_{p=0}^{r_{t-1}} x(N_1 p + q_1) W_{r_{t-1}}^{p m_1}. \tag{5}
\]
2.3. R2^2-SDF Architecture. The R2^2-SDF architecture was proposed by He and Torkelson [4] and also begins from (1). He applies a 3-dimensional index map:

\[ n = \left\lfloor \frac{N}{2} \right\rfloor n_1 + \frac{N}{4} n_2 + n_3 \right\rfloor ; \quad k = (k_1 + 2k_2 + 4k_3)_N. \]  

Using the Common Factor Algorithm (CFA) to decompose the twiddle factor, the FFT can be reconstructed as a set of 4 DFTs of length N/4:

\[ X(k_1 + 2k_2 + 4k_3) = \sum_{n_3=0}^{N/4-1} \left[ H(k_1, k_2, n_3) W_{N/4}^{n_3 k_3} \right] W_{N/4}^p k_3 \]  

\[ H(k_1, k_2, n_3) \] can be expressed as

\[ H(k_1, k_2, n_3) = \left[ x(n_3) + (-1)^{k_1} x(n_3 + (N/2)) \right] + (-j)^{k_2} x(n_3 + (3/4)N), \]  

3. FPGA-Based Implementations and Optimizations

3.1. Specifications, Tool Flow, and Verification. Both of these FFT architectures were implemented with generic synthesizable VHDL code and verified with simulation against Matlab scripts using Modelsim. Synplify or XST was used to perform the synthesis, and ISE was used for place and route and implementation. The architectures were optimized to achieve maximum throughput with minimal area (slices). The tools and development environment used are shown in Table 2.

3.2. General Optimization Methods. Some general optimization measures were performed, including FSM encoding, retiming, and CAD-related optimizations. Since the FFT processors were targeted to Xilinx Spartan-3 and Virtex-4 FPGAs (as well as synthesized for Virtex-E FPGAs), the SRL16 component, which can implement a 16-bit shift register within a single LUT, was inferred as much as possible to preserve LUTs. This particularly helped the R2^2-SDF architecture because of the large number of shift registers. R4SDC also benefited from SRL16 components in its commutator registers. Block RAMs were used to store twiddle factors, which dramatically reduced the combinational logic utilization.

3.3. Architecture-Specific Optimization. A number of architecture-specific optimizations were used. For both architectures, a complex multiplication technique was used. Usually, a complex multiplication is computed as:

\[ (a + bi) \times (c + di) = a \times c - b \times d + (a \times d + b \times c)i. \]  

This requires 4 multiplications and 2 add/suboperations. As is well known, the equation is simplified to save one multiplier:

\[ (a + bi) \times (c + di) = [a \times (c + d) - (a \times b) + c)i \]  

\[ + [a \times (c + d) + (a - b) \times c)i. \]  

This requires only 3 multiplications and 5 add/suboperations. Pipeline registers were also added in order to avoid the long critical path brought by the connection of real adders and multipliers. Figure 7 shows the pipeline stages inserted which were effective to reduce the critical path (REG means pipeline register).

3.3.1. R4SDC Optimization. The R4SDC has a complex controller, which creates a long critical path. By observing that all stages have the same control bits but have different sequences, using a ROM with an incremental address was
a simpler solution than using a complex FSM. Pipeline registers were also added to the butterfly elements, multipliers, and between stages. Figure 8 illustrates the addition of pipeline registers to cut the critical path efficiently within the butterfly element. Since two continuous add/subelements bring about a long propagation path, they were split using pipelining. Figure 9 shows the addition of pipeline registers between majority elements and between stages. For timing purposes, the applicable control signals were also buffered.

There were some special measures taken into account within controller in order to keep proper timing of signals. Twiddle factors should also be delayed to cope with the delayed sequence.

### 3.3.2. R2^2SDF Optimization

Due to its simple control requirements, a simple counter was sufficient as the entire controller for the R2^2SDF. To speed up the controller, a fast adder could potentially be faster than a simple ripple-carry adder. However, due to the small number of stages (log₂N), no substantial savings were found for a fast adder. Pipeline registers were added between major elements and also between stages. Note that the R2^2SDF is not suited for adding pipeline registers within individual butterfly elements, because this would break the timing for the data feedback path. Figure 10 presents the pipelined stages. Note that registers were only added between element units; in addition, registers were added as necessary to keep the control signals properly timed.

### 4. Rounding Scheme and SQNR

Due to finite wordlength effects, the implemented FFTs always scaled by 1/N at the output of the design. This scaling factor was distributed as divide-by-two operations throughout each stage to reduce error propagation. As is well known, truncation or conventional rounding (which is denoted as round-half-up) will bring a notable quantization error bias in divide-by-two operations, and this bias will accumulate throughout the processing chain [5]. To alleviate the bias, three unbiased rounding methods are investigated for division by two.

**Sign Bit-Based Rounding.** In this scenario, if the MSB of the number to be divided is 0 (i.e., positive number) it is rounded-half-up. This will have a positive bias. On the other hand, if the MSB is 1 (i.e., negative number) it is truncated, leaving a negative bias. Assuming that the positive and negative numbers are uniformly distributed, this approach will lead to an unbiased rounding scheme. However, selecting the bias based on the MSB implies that these two rounding methods coexist in a single rounding position, which requires extra hardware. This increases the critical path, harming the performance. So it is not chosen.

**Randomized [7].** In this scenario, if the bit to be rounded is 1, a random up or down rounding is performed. If it is 0, the same rounding scheme as done previously is performed. From the statistical point of view, no bias exists. But this method requires a random bit generator and a long accumulation time, requiring big extra hardware resources and significantly affects the performance. So it is not implemented.

**Balanced Stages Rounding [11].** This rounding method explores balancing between stages. Round-half-up and truncation are used in an interlaced fashion, as shown in Figure 11.
### Table 3: SQNR with different FFT sizes.

<table>
<thead>
<tr>
<th>FFT size</th>
<th>Input data width</th>
<th>Twiddle factor width</th>
<th>Stage number</th>
<th>SQNR (dB)</th>
</tr>
</thead>
<tbody>
<tr>
<td>R4SDC</td>
<td>16</td>
<td>16</td>
<td>2</td>
<td>82.29</td>
</tr>
<tr>
<td></td>
<td>64</td>
<td>16</td>
<td>3</td>
<td>73.49</td>
</tr>
<tr>
<td></td>
<td>256</td>
<td>16</td>
<td>4</td>
<td>67.47</td>
</tr>
<tr>
<td></td>
<td>1024</td>
<td>16</td>
<td>5</td>
<td>61.25</td>
</tr>
<tr>
<td>R2²SDF</td>
<td>16</td>
<td>16</td>
<td>2</td>
<td>81.82</td>
</tr>
<tr>
<td></td>
<td>64</td>
<td>16</td>
<td>3</td>
<td>74.47</td>
</tr>
<tr>
<td></td>
<td>256</td>
<td>16</td>
<td>4</td>
<td>68.22</td>
</tr>
<tr>
<td></td>
<td>1024</td>
<td>16</td>
<td>5</td>
<td>62.68</td>
</tr>
</tbody>
</table>

### Table 4: Implementation results on Spartan-3 devices.

<table>
<thead>
<tr>
<th>Point size</th>
<th>Input data width</th>
<th>Twiddle factor width</th>
<th>Slices</th>
<th>Block RAM</th>
<th>Max. speed (MHz)</th>
<th>Latency (cycles)</th>
<th>Transform time Cycles</th>
<th>Time (μs)</th>
<th>Throughput (MS/s)</th>
<th>Throughput/area (MS/s/slice)</th>
</tr>
</thead>
<tbody>
<tr>
<td>R4SDC</td>
<td>16</td>
<td>16</td>
<td>468</td>
<td>2</td>
<td>108.20</td>
<td>21</td>
<td>16</td>
<td>0.15</td>
<td>108.20</td>
<td>0.231</td>
</tr>
<tr>
<td></td>
<td>64</td>
<td>16</td>
<td>952</td>
<td>2</td>
<td>107.23</td>
<td>73</td>
<td>64</td>
<td>0.60</td>
<td>107.23</td>
<td>0.113</td>
</tr>
<tr>
<td></td>
<td>256</td>
<td>16</td>
<td>1990</td>
<td>3</td>
<td>111.98</td>
<td>269</td>
<td>256</td>
<td>2.76</td>
<td>111.98</td>
<td>0.056</td>
</tr>
<tr>
<td></td>
<td>1024</td>
<td>16</td>
<td>4409</td>
<td>8</td>
<td>123.84</td>
<td>1041</td>
<td>1024</td>
<td>8.27</td>
<td>123.84</td>
<td>0.028</td>
</tr>
<tr>
<td>R2²SDF</td>
<td>16</td>
<td>16</td>
<td>427</td>
<td>2</td>
<td>121.24</td>
<td>22</td>
<td>16</td>
<td>0.13</td>
<td>121.24</td>
<td>0.284</td>
</tr>
<tr>
<td></td>
<td>64</td>
<td>16</td>
<td>810</td>
<td>2</td>
<td>98.14</td>
<td>74</td>
<td>64</td>
<td>0.65</td>
<td>98.14</td>
<td>0.121</td>
</tr>
<tr>
<td></td>
<td>256</td>
<td>16</td>
<td>1303</td>
<td>3</td>
<td>98.73</td>
<td>270</td>
<td>256</td>
<td>2.59</td>
<td>98.73</td>
<td>0.076</td>
</tr>
<tr>
<td></td>
<td>1024</td>
<td>16</td>
<td>2802</td>
<td>8</td>
<td>95.25</td>
<td>1042</td>
<td>1024</td>
<td>10.75</td>
<td>95.25</td>
<td>0.034</td>
</tr>
</tbody>
</table>

### Table 5: Implementation results on Virtex-4 devices.

<table>
<thead>
<tr>
<th>Point size</th>
<th>Input data width</th>
<th>Twiddle factor width</th>
<th>Slices</th>
<th>Block RAM</th>
<th>Max. speed (MHz)</th>
<th>Latency (cycles)</th>
<th>Transform time Cycles</th>
<th>Time (μs)</th>
<th>Throughput (MS/s)</th>
<th>Throughput/area (MS/s/slice)</th>
</tr>
</thead>
<tbody>
<tr>
<td>R4SDC</td>
<td>16</td>
<td>16</td>
<td>530</td>
<td>1</td>
<td>236.7</td>
<td>21</td>
<td>16</td>
<td>0.07</td>
<td>236.7</td>
<td>0.447</td>
</tr>
<tr>
<td></td>
<td>64</td>
<td>16</td>
<td>803</td>
<td>1</td>
<td>236.4</td>
<td>73</td>
<td>64</td>
<td>0.27</td>
<td>236.4</td>
<td>0.294</td>
</tr>
<tr>
<td></td>
<td>256</td>
<td>16</td>
<td>1370</td>
<td>3</td>
<td>218.9</td>
<td>269</td>
<td>256</td>
<td>1.17</td>
<td>218.9</td>
<td>0.160</td>
</tr>
<tr>
<td></td>
<td>1024</td>
<td>16</td>
<td>3064</td>
<td>8</td>
<td>219.2</td>
<td>1041</td>
<td>1024</td>
<td>4.67</td>
<td>219.2</td>
<td>0.072</td>
</tr>
<tr>
<td>R2²SDF</td>
<td>16</td>
<td>16</td>
<td>517</td>
<td>1</td>
<td>237.9</td>
<td>22</td>
<td>16</td>
<td>0.07</td>
<td>237.9</td>
<td>0.460</td>
</tr>
<tr>
<td></td>
<td>64</td>
<td>16</td>
<td>779</td>
<td>1</td>
<td>236.7</td>
<td>74</td>
<td>64</td>
<td>0.27</td>
<td>236.7</td>
<td>0.304</td>
</tr>
<tr>
<td></td>
<td>256</td>
<td>16</td>
<td>1234</td>
<td>3</td>
<td>236.7</td>
<td>270</td>
<td>256</td>
<td>1.08</td>
<td>236.7</td>
<td>0.192</td>
</tr>
<tr>
<td></td>
<td>1024</td>
<td>16</td>
<td>2256</td>
<td>8</td>
<td>235.6</td>
<td>1042</td>
<td>1024</td>
<td>4.35</td>
<td>235.6</td>
<td>0.104</td>
</tr>
</tbody>
</table>

### Table 6: Performance comparison versus prior art on Virtex-E devices.

<table>
<thead>
<tr>
<th>Design</th>
<th>FFT size</th>
<th>Point size</th>
<th>Input data width</th>
<th>Twiddle factor width</th>
<th>Slices</th>
<th>Block RAM</th>
<th>Max. speed (MHz)</th>
<th>Latency (Cycle)</th>
<th>Transform time Cycles</th>
<th>Time (μs)</th>
<th>Throughput (MS/s)</th>
<th>Throughput/area (MS/s/slice)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Amphion  [8]</td>
<td>1024</td>
<td>16</td>
<td>13</td>
<td>13</td>
<td>1639</td>
<td>9</td>
<td>57</td>
<td>5097</td>
<td>4096</td>
<td>71.86</td>
<td>14.25</td>
<td>0.009</td>
</tr>
<tr>
<td>Xilinx   [8, 9]</td>
<td>1024</td>
<td>16</td>
<td>16</td>
<td>16</td>
<td>1968</td>
<td>24</td>
<td>83</td>
<td>4096</td>
<td>4096</td>
<td>49.35</td>
<td>20.75</td>
<td>0.011</td>
</tr>
<tr>
<td>Sundance [10]</td>
<td>1024</td>
<td>16</td>
<td>10</td>
<td>10</td>
<td>8031</td>
<td>20</td>
<td>49</td>
<td>1320</td>
<td>1320</td>
<td>27.00</td>
<td>49.00</td>
<td>0.006</td>
</tr>
<tr>
<td>Suksawas R2²SDF [8]</td>
<td>1024</td>
<td>16</td>
<td>16</td>
<td>16</td>
<td>7365</td>
<td>28</td>
<td>82</td>
<td>1099</td>
<td>1024</td>
<td>12.49</td>
<td>82.00</td>
<td>0.011</td>
</tr>
<tr>
<td>Our R2²SDF</td>
<td>1024</td>
<td>16</td>
<td>16</td>
<td>16</td>
<td>5008</td>
<td>32</td>
<td>95.0</td>
<td>1042</td>
<td>1024</td>
<td>10.78</td>
<td>95.00</td>
<td>0.019</td>
</tr>
<tr>
<td>Our R4SDC</td>
<td>1024</td>
<td>16</td>
<td>16</td>
<td>16</td>
<td>7052</td>
<td>32</td>
<td>94.2</td>
<td>1041</td>
<td>1024</td>
<td>10.87</td>
<td>94.20</td>
<td>0.013</td>
</tr>
</tbody>
</table>
5. Results and Analysis

5.1. SQNR Results. Figure 13 shows the SQNR results with different rounding schemes (balanced stages, truncation, and round-half-up), for R4SDC and R2^2SDF, respectively, for a 16-bit data width (input data, twiddle factors, and output data are 16 bits). The balanced stage rounding typically improved the SQNR by 1-2 dB. The balanced stages scheme gives better SQNR, because it leverages the randomness between stages. The truncation and round-half-up only reserve half of the information.

Table 3 presents the SQNR results as they vary with FFT size. The larger the FFT, the worse the SQNR due to the longer processing chain. Both architectures gave comparable results in terms of SQNR. It is clear that larger data widths will also give better SQNR but will increase area and critical path. A 16-bit wordlength is a sufficient choice for many signal processing applications.

The FFT architectures with smaller wordlengths than 16 bits are also implemented. The example in the Figure 14 shows the R4SDC architecture for an N = 1024 point FFT. Every bit of word length increment brings about 6 dB of SQNR gain.

5.2. Implementation Results. Table 4 gives the performance results of both architectures with different FFT sizes on Spartan-3 FPGAs (90 nm) [12]. The R2^2SDF achieved a smaller area and better throughput per area than the R4SDC. Due to the pipeline design, the maximum clock frequency did not change drastically with FFT size for either design. As expected the throughput per area decreases for larger FFT sizes, which require more stages and area.

Table 5 presents the results on Virtex-4 FPGAs (90 nm) using a 16-bit wordlength. Virtex-4 FPGAs have hardwired DSP modules called DSP48 blocks, which are high-speed modules optimized for signal processing operations such as multiply-accumulate, and FIR filtering. By utilizing these DSP48 blocks, the maximum clock frequency increased substantially over the Spartan-3 devices.

Comparisons with prior art are shown in Table 6, which shows publicly available pipeline FFT implementations on FPGAs from literature. For fair comparison, since many of the prior art implementations were implemented on Virtex-E FPGAs (180 nm), the designs are also implemented on Virtex-E. The best performance for the R2^2SDF method for a 16-bit 1024-point FFT was published by Sukhsawas and Benkrid in [8]. They used Handel-C as a rapid prototype language and implemented the design on Virtex-E FPGAs. They achieved 82 MHz maximum clock frequency and 7365 slices, giving a throughput per area ratio of 0.011 Msamples/s/slice.

On the Virtex-E, our R2^2SDF achieved better performance of 95 MHz and a smaller area of 5008 slices, giving a superior throughput per area ratio of 0.019 Msamples/s/slice.
Our R4SDC architecture was also superior to prior art, running at 94.2 MHz and using 7052 slices, a throughput per area ratio of 0.013 Msamples/s/slice.

Another point of reference is the Xilinx FFT IP core. For comparison sake, the IP core for Virtex-E is shown in the table. The Virtex-E core shows four times the latency (4096) in cycles due to its internal architecture. Its throughput per area ratio is also only 0.011 Msamples/s/slice. Note that all comparisons for throughput per area do not take into account block RAMs, though each of the designs had a similar number of required block RAMs. However, the Xilinx FFT DSP core could perform better with new Xtreme technology: on Virtex 4 device 4vsx25-10, 1024-point FFT
controller and compatibility with pipelining insertion. Both architectures have comparable maximum clock frequency and SQNR with the balanced stages rounding scheme.

References


Submit your manuscripts at http://www.hindawi.com