A GENERATOR OF MEMORY-BASED, RUNTIME-RECONFIGURABLE $2^N 3^M 5^K$ FFT ENGINES

Angie Wang, Jonathan Bachrach, Borivoje Nikolić

University of California, Berkeley

ABSTRACT

Runtime-reconfigurable, mixed-radix FFT/IFFT engines are essential for modern wireless communication systems. To comply with varying standards requirements, these engines are customized for each modem. The Chisel hardware construction language has been used in this work to create a generator of runtime-reconfigurable $2^N 3^M 5^K$ FFT engines targeting software-defined radios (SDR) for modern communications, but with flexibility to support a wide range of applications. The generator uses a conflict-free, in-place, multi-bank SRAM design, and exploits the duality of decimation-in-frequency (DIF) and decimation-in-time (DIT) FFTs to support continuous data flow with only 2N memory blocks. DFT decomposition using the prime-factor algorithm (PFA) followed by the Cooley-Tukey algorithm (CTA) reduces twiddle ROM sizes. A programmable Winograd’s Fourier Transform (WFTA) butterfly supporting radix-2/3/4/5/7 operations reuses radix-7 hardware to support reconfigurability with minimal area penalty. The generated FFTs use 50% less memory than iterative FFTs from Spiral. The twiddle ROM size of the generated LTE/WiFi FFT engine is 16% smaller than that of a 2048-pt Spiral design.

Index Terms— Fast Fourier Transform, Cooley-Tukey, Winograd’s Fourier Transform, Prime Factor Algorithm, Reconfigurable Hardware Generator.

1. INTRODUCTION

OFDM-based wireless transceivers need reconfigurable IFFT/FFT engines to support a multitude of channel bandwidths and modulation schemes. The need for adaptation to a wide range of channel bandwidths requires support for many different FFT sizes, which are often not decomposable to $2^n$, making reuse of these blocks difficult. For example, in order for an SDR to support both WiFi and LTE with single-carrier frequency-division multiple access (SC-FDMA), the maximum powers of (2, 3, 5) that must be supported are (2048, 243, 25) [1]. The need for a plethora of FFT engines has been addressed by building commercial reconfigurable, application-specific FFT cores. However, they are generally not resource efficient, which is problematic when users only need a subset of FFT sizes and are otherwise severely hardware constrained. Additionally, they cannot be easily modified to support new standards. The terrestrial DTV system in China, for example, requires a 3780-pt FFT and a radix-7 butterfly [2],[21]. A step forward is the development of FFT generators; however, all generators available in the open literature are of fixed size and not applicable to SDR.

FFT generators such as Spiral exploit design regularity, enabling usage in diverse applications with ideally no design overhead [3],[20]. Highly parameterizable generators also enable rapid and extensive design exploration, allowing throughput/area tradeoffs [17], memory access methods, fixed point optimizations (as in [4]), etc. to be studied in detail. However, no generator in open literature supports all these requirements. Thus, we have designed a generator of memory-based, runtime-reconfigurable $2^N 3^M 5^K$ FFT engines. The design is hardware-optimized: it relies on butterfly reuse, supports continuous data flow with 2N memory blocks, and stores fewer twiddle factors than designs from state-of-the-art generators like Spiral.

2. DESIGN METHODOLOGY

2.1. Memory-Based Architecture

Hardware FFT designs often use either in-place memory-based or pipelined architectures. For large FFT sizes, memory-based designs are smaller in area—usually relying on a single iterating butterfly—but support lower throughput [2]. Pipelined designs use more processing elements to improve throughput and more easily support continuous data flow [16], but often over provision butterflies (i.e. log2N radix-2 butterflies for 2N-pt FFTs) to achieve this.

Memory-based designs with conflict-free scheduling schemes ease area/throughput tradeoffs and can compute an N-pt FFT in

$$C = SP + \sum_{i=0}^{N-1} \left\lfloor \frac{n_i}{r_i} \right\rfloor$$

(1)
clock cycles, where $B$ is the # of parallel butterflies, $r_i$ is the radix at the $i^{th}$ calculation stage, $S$ is the # of stages, and $P$ represents the pipelining between memory accesses (butterfly pipelines + sequential memory read delay). The first term shows that only two radix-2/4 butterflies are needed to complete a 2048-pt FFT in <2048 cycles. Memory-based architectures require more complex control logic, but adapt more easily to reconfigurable, non-$2^n$ flavors.

2.2. Cooley-Tukey & Prime-Factor Algorithms

The widely-used Cooley-Tukey algorithm (CTA) computes $2^n$-pt FFTs with a complexity of $O(N\log N)$ [2],[14]:

$$X[k_1, k_2] = \sum_{n_2} W_N^{n_2 k_2} \left\lfloor \sum_{n_1} x[n_1, n_2] W_N^{n_1 k_1} \right\rfloor$$

(2)

As shown in [15], The DIF FFT is constructed with ordered inputs, digit-reversed outputs, and twiddle multiplication ($W_N^{n_1 k_1}$) at the butterfly output. The opposite is true for a DIT FFT.
The CTA can be used for coprime DFT decompositions, but the prime-factor algorithm (PFA) achieves better efficiency by eliminating twiddle multiplications [5],[19]:

\[ X[k_1, k_2] = \sum_{n_1} \sum_{n_2} x[n_1, n_2] W^{{n_1}k_1} W^{{n_2}k_2} \]  

(3)

In [2], Hsiao et al. propose the following input/output index-mapping scheme, which we adopt in our design:

\[
\begin{align*}
    n &= (N_2 n_1 + A_1 n_2) \bmod N, \quad A_1 = p_1 N_1 = q_1 N_2 + 1 \\
    k &= (B_1 k_1 + N_1 k_2) \bmod N, \quad B_1 = p_2 N_2 = q_2 N_1 + 1
\end{align*}
\]  

(4) (5)

When \( N_1 \) and \( N_2 \) are coprime, (4-5) simplify to one of several possible PFA mappings. Otherwise, using \( A_1, B_1 = 1 \), the equations reduce to the CTA mapping. Ordering the factorization of \( N \) so that the PFA decomposition occurs before the CTA minimizes twiddle multiplications [5]. This allows the twiddle ROMs to be partitioned by coprimes to ease reconfigurability, and reduces the necessary ROM size. Without further optimizations, the # of twiddle factors stored for radix-2/3/4/5 support is:

\[ T = 3 \times 2^n \times a_{\max}^2 + 2 \times 3^n \times a_{\max}^1 + 4 \times 5^n \times a_{\max}^1 \]  

(6)

Only 1718 twiddle factors are needed for all Wi-Fi/LTE FFTs, so the twiddle ROM can be 16% smaller than that of a 2048-pt Spiral design [3].

### 2.3. DIF ↔ DIT Duality for Memory Reduction

Typically, calculations can be done in-place, but I/O cannot because input/output mappings are digit-reversed. A naive ping-pong memory-based FFT architecture thus requires at least 3N memory entries to support continuous data flow. The duality of the DIF/DIT decompositions (and extensions to the PFA) enables 2N memory usage [6]. As shown in Fig. 1, reversing the decomposition order every 2\( \alpha \)th symbol allows \( x_{2\alpha-2}[n], x_{2\alpha+2}[n] \) to be written to memory as \( x_{2\alpha}[k], x_{2\alpha+1}[k] \) read out, in order, with \( n=k \). The generated design uses 50% less BRAM (in an FPGA) than a Spiral equivalent.

To extend this duality to the PFA [2], first assume \( N \) is decomposed in the order \( N_1, N_2, N_3 \). Then, from (4-5),

\[
\begin{align*}
    n &= (N_2 n_1 + A_1 n_2) \bmod N, \quad A_1 = p_1 N_1 = q_1 N_2 + 1 \\
    k &= (B_1 k_1 + N_1 k_2) \bmod N, \quad B_1 = p_2 N_2 = q_2 N_1 + 1
\end{align*}
\]  

(7) (8)

When \( N_1, N_2, N_3 \) are coprime (GCD=1),

\[
\begin{align*}
    A_1 &= p_1 N_1 = Q_1 N_2 + 1, \quad Q_1 = q_1 \\
    A_2 &= p_2 N_2 = Q_2 N_3 + 1, \quad Q_2 = q_2 \\
    A_3 &= p_3 N_3 = Q_3 N_1 + 1, \quad Q_3 = q_3
\end{align*}
\]  

(9) (10) (11)

\( Q_i \) is calculated via the Scala implementation of the extended Euclidean algorithm \( \gcd(a, b) = ax + by \).

Substitution of (9-11) into (7-8) results in the mappings [2]:

\[
\begin{align*}
    n &= (N_2 n_1 + p_1 N_2 n_2 + p_1 p_2 N_3 n_2) \bmod N \\
    k &= (p_2 p_3 N_2 k_1 + N_2 p_3 N_3 k_2 + N_3 p_3 k_2) \bmod N
\end{align*}
\]  

(12) (13)

A similar decomposition in the reversed order \( N_3, N_2, N_1 \) results in:

\[
\begin{align*}
    n &= (N_3 n_1 + p_1 N_3 n_2 + p_1 p_2 N_2 n_2) \bmod N \\
    k &= (p_2 p_3 N_3 k_1 + N_3 p_3 N_2 k_2 + N_2 p_3 k_2) \bmod N
\end{align*}
\]  

(14) (15)

As in the CTA case, to achieve in-place I/O with 2N memory, \( k_3, k_2, k_1 \) to \( n_1, n_2, n_3 \) in (16):

\[
(\begin{array}{c}
    k_3, k_2, k_1 \\
\end{array})_{2a} \rightarrow (n_1, n_2, n_3)_{2a+2}
\]  

(16)
To support coprime factorization and then CTA, the vector $(n_{3,0}, n_{3,1}, ..., n_1, n_{2,0}, n_{2,1}, ..., n_2, n_{1,0}, n_{1,1}, n_{1,2})$ can be obtained by using the $(n_1, n_2, n_3)$ values to address corresponding decimal-to-base-r LUTs. However, we have chosen instead to store the Q's in base-r notation and implement base-r adders. The coprime modulus operations are obtained “for free” by simply masking out the appropriate base-r digits, and the final index vector is directly obtained without additional LUTs. As an example, if $N_1$ is 5$, $n_3$ is decomposed into the base-5 digits $(n_{3,0}, n_{3,1}, ..., n_{3,0})$. For an $N_1 N_2 N_3$ decomposition, $n_1$ uses base-5 and the index vector is reversed for memory address/bank generation: $(n_{3,0}, ..., n_{3,0}, n_{2,0}, ..., n_{2,0}, n_{1,0}, ..., n_{1,2})$. The individual coprimes are vectorized in digit-reversed order.

### 3.3. Calculation + Memory Access Control

The DIF FFT uses $S$ radix stages. Fig. 1 shows that the number of radix-$r_l$ butterflies at stage $i$ is $N/r_l$. For $l > 0$, butterflies occur in $\prod_{k=1}^{l-1} r_k$ groups of $\prod_{k=1}^{l} r_k$ ordered operations. For the DIT-equivalent FFT, the DIF stages are traversed in reverse. A set of $n_x$ counters is used to index operands and track the butterfly iterations per stage:

$$n_x, = [n_{x-1} + 1 \text{ if } n_{x-1} \neq r_{x-1} - 1 \text{ and } S \neq 1; \text{ else } 0]$$

$$n_x, = [n_{x-1} + 1 \text{ if } n_{x-1} \neq r_{x-1} - 1 \text{ and } S \neq 1; \text{ else } 0]$$

Because $r_i \leq r_{\text{max}}$, each butterfly operand indexed by $j < r_i$ is guaranteed to come from a different bank $b_j$ [2], given by:

$$b_0 = \left( \sum_{k=0}^{j} n_x \right) \mod r_{\text{max}}$$

$$b_{0 < j < r_i} = (b_0 + j) \mod r_{\text{max}}$$

Mod operations are more complex than the XOR logic in [8], but they simply extend XOR (add mod 2) to more general base-$r$'s, which is useful in mixed-radix designs. Because the $N$ data are split amongst $r_{\text{max}}$ banks, data in each bank is mapped to $[0, N/r_{\text{max}}]$. A possible addressing scheme if $r_0 = r_{\text{max}}$ (where $a_j$ is the address of the $j$th butterfly operand) is:

$$A_{j-1} = 1, A_{x < j-1} = A_{x-1} \times r_{x-1} + 1$$

$$a_0 = \sum_{x=0}^{n_x} A_x n_x$$

$$a_{j > 0} = a_{j-1} + A_j$$

3.4. Twiddle Address Generation

For an $N$-pt DIF FFT, the non-trivial $i$th stage twiddles $W_i N^{r_i}$, $j \in [1, r_i]$ repeated in each butterfly group are addressed by:

$$n_{\text{tw}, j > 0} = \left( \prod_{k=1}^{i} r_k \right) \times [0, N/(\prod_{k=0}^{i} r_k) - 1]$$

As shown in [15], the first CTA DIF stage requires the most twiddle factors. Because the DFT is decomposed into smaller sizes, each subsequent stage uses fewer unique twiddles, and the last stage requires none. The twiddle index is renormalized back to the full ROM range (i.e. $W_i^{'d} = W_{i,d}^{'d}$). Since the range of twiddle addresses for a given radix-$r$ is set by its largest associated coprime $r$, all values of $2^p < 2^n$, $2^m < 3^n$, $3^m < 5^n$, etc. are supported by only 9 ROMs with address renormalization (i.e. the renormalization factor for an initial radix-4 stage is $2^{n_{\text{max}}}/2^n$).

To extend this addressing scheme to combined PFA/CTA decompositions, the twiddle address of a radix-$r_i$ stage is held for $r_i$ butterflies. If $r_i$'s corresponding coprime is $C_i$, $t \in [0, D)$ where $D$ represents the number of coprimes in the decomposition, then

$$\begin{align*}
\tilde{z}_i &= \left( \prod_{k=1}^{i} r_k \right) \times C_i \\
\text{Note that when } t = D - 1, \tilde{z}_i &= 1.
\end{align*}$$

### 3.5. WFTA Butterfly Supporting Hardware Reuse

The WFTA butterfly uses fewer multipliers at the expense of extra adders [18]. To minimize area, the adders/multipliers of a radix-7 butterfly can be reused for radix-2/3/4/5, as in [9]. Fig. 5 shows a modified version of [9], used here. It supports optional logic pipelining per computation stage as required by the FPGA/ASIC and bypasses adders/multipliers when radices are unused. To support full reconfigurability, constant multipliers are not used, but
only products of complex inputs and a few real/imaginary values are needed. As Fig. 4 shows, the processing element (butterfly + twiddle multipliers) supports DIF/DIT reconfiguration using one set of twiddle multipliers with at least one stage of pipelining between memory accesses.

4. DISCUSSION

4.1. Scheduling Limitations

For continuous data flow with LTE/WiFi FFTs, butterflies are computed at 2x the I/O clock rate. By using a single butterfly at this clock rate, the N=648 and N=964 calculations (requiring >1350 and >1728 computation cycles respectively) cannot be completed in time. Since the radix-2 stage (no twiddles) requires the most clock cycles (CCs), halving the iterations during that stage remedies this problem. This is achieved by operating two radix-2 butterflies in parallel, reusing radix-4 hardware, and shuffling butterflies to prevent bank conflicts. For the radix-2 stage of an N=24 FFT, Fig. 2 shows that the by butterflies in groups 0/2 and 1/3 use the same addresses and non-conflicting memory banks. Additional logic derived from this observation is used to guarantee FFT computation in N I/O CCs for LTE/WiFi.

Only one butterfly may be used with the implemented banking scheme. In order to support area/throughtput flexibility and an arbitrary number of butterflies in parallel, it is necessary to extend butterfly scheduling further. This is especially difficult when one desires a general solution for runtime-reconfigurable, mixed-radix FFTs without adding considerable complexity to the I/O (as would be incurred in [10]). However, it is not necessary to change operand locations in order to prevent bank conflicts. Instead, butterfly operations within a radix-r stage may be reordered to minimize conflicts.

4.2. Results & Comparisons with State-of-the-Art

A design has been verified in a cycle-accurate fashion via Chisel’s built-in tester. As shown in Fig. 6, two generated FFT engines (24,24-bit I/O) were synthesized with 28nm standard cells and a clock target of 3.9ns to meet LTE/WiFi requirements. There is a 25% area penalty (compared to a fixed N=2048 engine) to support mixed-radix reconfigurability. The radix-4/2 WFTA butterfly is smaller than the reconfigurable 2/3/4/5 butterfly, and memories occupy a vast majority of the area. Support for all LTE/WiFi FFTs requires 2x(240-length memory + 4 banks of 512-length memory).

The proposed generator can create FFT engines with performance comparable to state-of-the-art reconfigurable FFTs [2,5]. Wang et al. [11] propose a pipelined architecture that can be reconfigured/sequenced, but it isn’t implemented as a generator. While achieving higher throughput, the generated FFT engine uses 50% less data memory and 25% smaller twiddle storage than the radix-2 N=2048 iterative FFT from Spiral’s online generator [3]. Memory savings are even greater relative to Spiral’s streaming FFTs, but with a clear area/throughput tradeoff (Table 1). Designs in Table 2 with fewer multipliers require higher calculation clock rates. Genesis [12] and the fixed pipelined architecture from [13] do not address I/O and its impact on memory.

4.3. Conclusion

This work reports a generator of memory-based, runtime-reconfigurable 2354 FFT engines, developed entirely in Chisel. The FFT design instances are competitive with state-of-the-art reconfigurable architectures, satisfying the WiFi and LTE operating modes while supporting continuous data flow with 2N memories. This work can be extended to support 7 co-primers and more generalized FFT scheduling for exploring area/throughput tradeoffs of reconfigurable architectures.

The authors acknowledge DARPA CRAFT (HR0011-15-1-0010), NSF-GRFP (DGE-1106400), ASPIRE, and BWRC faculty, staff, and students (esp. S. Bailey, P. Rigge, S. Twigg) for their support.

Table 1. Resource comparison for fixed N = 2048

<table>
<thead>
<tr>
<th></th>
<th>This</th>
<th>[5],</th>
<th>[13]</th>
</tr>
</thead>
<tbody>
<tr>
<td>D. Mem</td>
<td>2N</td>
<td>2N</td>
<td>2N</td>
</tr>
<tr>
<td>T. ROM</td>
<td>0.75N</td>
<td>0.99N</td>
<td>0.99N</td>
</tr>
<tr>
<td>CCs</td>
<td>-3600</td>
<td>512</td>
<td>1024</td>
</tr>
<tr>
<td># BFs</td>
<td>1 Rad-4/2</td>
<td>5 Rad-4/2</td>
<td>11</td>
</tr>
<tr>
<td># Muls**</td>
<td>12</td>
<td>56</td>
<td>40</td>
</tr>
</tbody>
</table>

1. Rad-4 streaming; 2: Rad-2 streaming; 3: Rad-2 iterative
* No memory allocated for I/O unscrambling
** Real muls in data path; complex muls = 4 real muls

Table 2. Resource comparison for reconfigurable LTE/WiFi FFTs

<table>
<thead>
<tr>
<th></th>
<th>This</th>
<th>[5]*</th>
<th>Xilinx*</th>
</tr>
</thead>
<tbody>
<tr>
<td>D. Mem</td>
<td>2N</td>
<td>2N</td>
<td>2N</td>
</tr>
<tr>
<td>T. ROM</td>
<td>1718</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>Calc. Clk/IO Clk**</td>
<td>2x</td>
<td>1x</td>
<td>2x</td>
</tr>
<tr>
<td># Muls</td>
<td>26</td>
<td>44</td>
<td>26</td>
</tr>
</tbody>
</table>

* Comparison #’s taken from [5]
** Ratio of calculation to I/O clock rates for continuous data flow

Fig. 6. FFT area breakdown
5. REFERENCES


