## Generating High-Performance General Size Linear Transform Libraries Using Spiral

#### Yevgen Voronenko

Franz Franchetti Frédéric de Mesmay Markus Püschel Carnegie Mellon University



This work was supported by DARPA DESA program, NSF-NGS/ITR, NSF-ACR, and Intel



# **The Problem: Example DFT**

Discrete Fourier Transform (DFT) on 2xCore2Duo 3 GHz (single precision) Performance [Gflop/s]



- Standard desktop computer
- Same operations count ≈4nlog<sub>2</sub>(n)
- Similar plots can be shown for all numerical problems

# **DFT Plot: Analysis**



- High performance library development = nightmare
- Automation?

## **Idea: Textbook to Adaptive Library**



# **Goal:** Teach Computers to Write Libraries

### Input:

- Transform:DFT<sub>n</sub>
- Algorithm:  $DFT_{km} \rightarrow (DFT_k \otimes I_m)T_m^{km}(I_k \otimes DFT_m)L_k^{km}$
- Hardware: 2-way SIMD + multithreaded

Spiral

### **Output:**

- FFTW equivalent library
- For general input size
- Vectorized and multithreaded
- Performance competitive

### **Key technologies:**

- Layered domain specific language
- Algorithm manipulation via rewriting
- Feedback-driven search

### **Result:**

Full automation



# **Contribution: General Size Library**



Fundamentally different problems

# **Beyond Fourier Transform and FFTW**



## **Examples of Generated Libraries**



**Total:** 300 KLOC / 13.3 MB of code generated in < 20 hours *from a few simple algorithm specs* 

Intel IPP library 6.0 will include Spiral generated code

### I. Background

- **II.** Library Generation
- **III.** Experimental Results
- **IV.** Conclusions and Future Work

## **Linear Transforms**

Mathematically: matrix-vector product



10

## Fast Algorithms, Example: 4-point FFT



- SPL = mathematical, declarative specification
- Space of algorithms generated using breakdown rules  $DFT_{mn} \rightarrow (DFT_m \otimes I_n)D(I_m \otimes DFT_n)P$

## **Examples of Breakdown Rules**

$$\begin{array}{c} \mathbf{DFT}_{n} \longrightarrow (\mathbf{DFT}_{k} \otimes I_{m}) D_{k,m} (I_{k} \otimes \mathbf{DFT}_{m}) L_{k}^{n} \\ \mathbf{DFT}_{n} \longrightarrow V_{m,k}^{-1} (\mathbf{DFT}_{k} \otimes I_{m}) (I_{k} \otimes \mathbf{DFT}_{m}) V_{m,k} \\ (2.2) \\ \mathbf{DFT}_{n} \longrightarrow W_{n}^{-1} (I_{1} \oplus \mathbf{DFT}_{n-1}) E_{n} (I_{1} \oplus \mathbf{DFT}_{n-1}) W_{n} \\ (2.3) \\ \mathbf{DFT}_{n} \longrightarrow B_{n,m}^{\top} D_{m} \mathbf{DFT}_{m} D'_{m} \mathbf{DFT}_{m} D''_{m} B_{n,m}, \quad m \geq 2n-1 \\ (2.4) \\ \mathbf{DFT}_{n} \longrightarrow P_{k/2,2m}^{\top} (\mathbf{DFT}_{2m} \oplus (I_{k/2-1} \otimes_{i} C_{2m} \mathbf{rDFT}_{2m} ((i+1)/k))) (\mathbf{RDFT}_{k}' \otimes I_{m}) \\ (2.5) \\ \end{array}$$

$$\begin{array}{c} \left| \begin{array}{c} \mathbf{RDFT}_{n} \\ \mathbf{PDTT}_{n} \\ \mathbf{DHT}_{n} \\ \mathbf{DHT}_{n}' \\ \mathbf{DH$$

"Teach" Spiral domain knowledge of algorithms. Never obsolete. Each rule leads to a library

### I. Background

### **II.** Library Generation

### **III.** Experimental Results

### **IV.** Conclusions and Future Work

# **How Library Generation Works**



# **Breakdown Rules to Library Code**

Cooley-Tukey Fast Fourier Transform (FFT)

$$\mathbf{DFT}_{km} x = (\mathbf{DFT}_k \otimes I_m) T_m^{km} (I_k \otimes \mathbf{DFT}_m) L_k^{km} x$$
$$\mathbf{DFT} = \begin{bmatrix} \mathbf{DFT}_k \otimes I_m \\ \mathbf{DFT} \end{bmatrix} \begin{bmatrix} \mathbf{FT}_k \otimes I_m \\ \mathbf{FT}_k \end{bmatrix} \end{bmatrix} \begin{bmatrix} \mathbf{FT}_k \otimes I_m \\ \mathbf{FT}_k \end{bmatrix} \begin{bmatrix} \mathbf{FT}_k \otimes I_m \\ \mathbf{FT}_k \end{bmatrix} \end{bmatrix} \begin{bmatrix} \mathbf{FT}_k \otimes I_m \\ \mathbf{FT}_k \end{bmatrix} \begin{bmatrix} \mathbf{FT}_k \otimes I_m \\ \mathbf{FT}_k \end{bmatrix} \end{bmatrix} \begin{bmatrix} \mathbf{FT}_k$$

#### Naive implementation

```
void dft(int n, cplx X[], cplx Y[]) {
    k = choose_factor(n); m = n/k;
    Z = permute(X)
    for i=0 to k-1
        dft_subvec(m, Z, Y, ...)
    for i=0 to n-1
        Y[i] = Y[i]*T[i];
    for i=0 to m-1
        dft_strided(k, Y, Y, ...)
```

#### 2 extra functions needed

# **Breakdown Rules to Library Code**

Cooley-Tukey Fast Fourier Transform (FFT)

$$\mathbf{DFT}_{km} x = (\mathbf{DFT}_k \otimes I_m) T_m^{km} (I_k \otimes \mathbf{DFT}_m) L_k^{km} x$$

$$\mathbf{DFT} = \left[ \begin{array}{c} \mathbf{DFT}_k \otimes I_m \\ \mathbf{DFT}_k \otimes I_m \end{array} \right] \left[ \begin{array}{c} \mathbf{Stride} \\ \mathbf{$$

#### Naive implementation

```
void dft(int n, cplx X[], cplx Y[]) {
    k = choose_factor(n); m = n/k;
    Z = permute(X)
    for i=0 to k-1
        dft_subvec(m, Z, Y, ...)
    for i=0 to n-1
        Y[i] = Y[i]*T[i];
    for i=0 to m-1
        dft_strided(k, Y, Y, ...)
}
```

#### Optimized implementation

```
void dft(int n, cplx X[], cplx Y[]) {
    k = choose_factor(n); m = n/k;
```

```
for i=0 to k-1
   dft_strided2(m, X, Y, ...)
for i=0 to m-1
   dft_strided3_scaled(k, Y, Y, T, ...)
}
```

### How to discover these specialized variants automatically?

# **Library Structure**

 $\bigvee_{k,m}^T (\mathbf{DFT}_k \otimes \mathbf{I}_m) (\mathbf{I}_k \otimes \mathbf{DFT}_m) \vee_{k,m} \\ (\mathbf{DFT}_k \otimes \mathbf{I}_m) \top_m^{km} (\mathbf{I}_k \otimes \mathbf{DFT}_m) \sqcup_k^{km}$ 

#### **Library Structure**

Parallelization / Vectorization Recursion Step Closure

 $\mathbf{DFT}$   $S_z \mathbf{DFT} G_h$   $S_h \mathbf{DFT} G_z$   $S_h \mathbf{DFT} \operatorname{diag} (\operatorname{Dat}) G_h$   $S_h \mathbf{DFT} G_h$   $S_h \mathbf{DFT} G_{h \circ z}$   $S_{h \circ z} \mathbf{DFT} G_h$   $S_h \mathbf{DFT} \operatorname{diag} (\operatorname{Dat}) G_{h \circ z}$   $S_z \mathbf{DFT} \operatorname{diag} (\operatorname{Dat}) G_h$   $S_{h \circ z} \mathbf{DFT} \operatorname{diag} (\operatorname{Dat}) G_h$ 

Input:

Breakdown rules

#### Output:

- Recursion step closure
- Σ-SPL Implementation of each recursion step

- Parallelization/Vectorization
  - Adds additional breakdown rules
  - Orthogonal to the closure generation

# **Computing Recursion Step Closure**

- Input: transform T and a breakdown rule
- **Output:** spawned recursion steps + Σ-SPL implementation



5. Repeat until closure is reached

### Parametrization (not shown) derives the independent parameter set 18 for each recursion step

## **Recursion Step Closure Examples**



#### **17** mutually recursive functions

## **Base Cases**

Base cases are called "codelets" in FFTW

### Why needed:

- Closure is converted into mutually recursive functions
- Recursion must be terminated
- Larger base cases eliminate overhead from recursion

### How many:

- In FFTW 3.2: 183 codelets for complex DFT (21 types) 147 codelets for real DFT (18 types)
- In our generator: # codelet types ‰ # recursion steps

Obtained by using standard Spiral to generate fixed size code

 $\left\{ \mathsf{S}(h_{u_3,1}^{2 \to u_2}) \operatorname{DFT}_2 \mathsf{G}(h_{u_7,u_8}^{2 \to u_6}) \right\} \\ \left\{ \mathsf{S}(h_{u_3,1}^{3 \to u_2}) \operatorname{DFT}_3 \mathsf{G}(h_{u_7,u_8}^{3 \to u_6}) \right\}$ 

# **Library Implementation**

 $\begin{array}{c} \mathbf{DFT} \\ \mathbf{S}_z \, \mathbf{DFT} \, \mathbf{G}_h \\ \mathbf{S}_h \, \mathbf{DFT} \, \mathbf{G}_z \\ \mathbf{S}_h \, \mathbf{DFT} \, \mathrm{diag} \, (\mathrm{Dat}) \, \mathbf{G}_h \\ \mathbf{S}_h \, \mathbf{DFT} \, \mathrm{G}_h \\ \mathbf{S}_h \, \mathbf{DFT} \, \mathbf{G}_{h \circ z} \\ \mathbf{S}_{h \circ z} \, \mathbf{DFT} \, \mathbf{G}_h \\ \mathbf{S}_h \, \mathbf{DFT} \, \mathrm{diag} \, (\mathrm{Dat}) \, \mathbf{G}_{h \circ z} \\ \mathbf{S}_z \, \mathbf{DFT} \, \mathrm{diag} \, (\mathrm{Dat}) \, \mathbf{G}_h \\ \mathbf{S}_{h \circ z} \, \mathbf{DFT} \, \mathrm{diag} \, (\mathrm{Dat}) \, \mathbf{G}_h \end{array}$ 



### Input:

- Recursion step closure
- Σ-SPL implementation of each recursion step (base cases + recursions)

## Output:

- High-performance library
- Target language: C++, Java, etc.

### Process:

- Build library plan
- Perform hot/cold partitioning
- Generate target language code

**High-performance library** 

### I. Background

**II.** Library Generation

### **III.** Experimental Results

### **IV.** Conclusions and Future Work

## **Double Precision Performance: Intel Xeon 5160** 2-way vectorization, up to 2 threads





23

## FIR Filter Performance 2- and 4-way vectorization, up to 2 threads





24

## **2-D Transforms Performance** 2- or 4-way vectorization, up to 2 threads



Performance [Gflop/s]





Performance [Gflop/s]



## **Customization: Code Size**

#### Performance [Gflop/s]



**Generated library JTransforms** 

16k 32k 64k

Generated

Naive

8k

2k

4k

## **Backend Customization: Java**



Portable, but only 50% of scalar C performance

## Summary

 Full automation: Textbook to adaptive library

## Performance

- SIMD
- Multicore
- Customization

## Industry collaboration

 Intel IPP 6.0 will include Spiral generated code





