Understanding the Design Space of DRAM–Optimized Hardware FFT Accelerators

Berkin Akın, Franz Franchetti, James C. Hoe

Electrical and Computer Engineering Department
Carnegie Mellon University

(This work was sponsored by DARPA under agreement HR0011-13-2-0007. The content, views and conclusions presented in this document do not necessarily reflect the position or the policy of DARPA.)
Motivation

- Technology scaling -> memory wall and power wall
  - High-performance AND energy-efficiency is challenging

- In data intensive computing memory subsystem is the key for high-performance/energy-efficiency
  - DRAM-optimized hardware FFT accelerators

- Memory-optimized block-layout FFT algorithms
  - Reshapes inefficient memory access patterns
  - Algorithm-architecture mapping is required to be effective
  - Various platforms and memory subsystems, different tradeoffs

- Automated techniques to evaluate the large design space
  - Find the best parameter configuration for a certain design goal
Contribution

- **DRAM-optimized hardware FFT accelerators**
  - 1D, 2D and 3D FFTs sizes ranging up to 1 billion data elements
  - 3D–stacked and off–chip DRAM, ASIC, FPGA

- **Automated design generation & exploration tool**
  - Extension of Spiral algorithm&architecture co–optimization framework
  - Given problem/platform parameters and constraints:
    - What are the design possibilities with varying power/performance?
    - Which design is the best for a certain metric?

- **Analysis of design space tradeoffs**
  - What are the internal parameters that control certain tradeoff curves?
  - How they differ for different platforms/problems?

- **DRAM–specific optimizations and design exploration leads to:**
  - >10x bandwidth and 5.5x energy improvement in DRAM
  - 6x performance and 6.5x energy improvement in overall system
Background: DRAM Architecture

![DRAM Architecture Diagram]

- **Rank 1**: 
  - DRAM Chip 0: Bank b, Bank 1, Bank 0
  - DRAM Chip 1: Bank b, Bank 1, Bank 0
  - DRAM Chip c: Bank b, Bank 1, Bank 0

- **Rank 0**: 
  - DRAM Chip 0: Bank b, Bank 1, Bank 0
  - DRAM Chip 1: Bank 1, Bank 0
  - DRAM Chip 0: Bank 1, Bank 0

- **I/O bus**

**DDR3-1600 Bandwidth & Energy Consumption**

- Bandwidth Util. [GB/s]
- Energy Cons. [J/GB]

- **Refresh**, **Rd/Wr**, **Act/Pre**, **Static**

- Bandwidth Util. [GB/s]
- log(block size)

- 1 bank
- 2 banks
- 4 banks
- 8 banks
Background: 3D–stacked DRAM

- 3D–stacked DRAM architecture

- 3D–stacked DRAM operation: Abundant parallelism
  - Each rank (vault) and each bank in a layer operate in parallel
  - Each layer within a rank operates in pseudo–parallel
  - Dense, high bandwidth, low latency TSVs

- Other highlights:
  - Peripherals in the logic layer → Much better timing & energy
  - Custom logic & DRAM closely coupled
Background: Large FFT & Blocked Layout

2D-FFT:

Logical view of the dataset:

Memory address space:

Blocked layout

FFT:

Tiled layout

Cubic layout

(Akin et. al. ICASSP’14)
Algorithm/Architecture/Platform Mapping

- **Design Parameters**
  - Throughput: FFT radix (r), streaming width (w), frequency (f).
  - Bandwidth: Type of tiling (2D, 3D), tile size (T)

- **Platform/Problem Constraints**
  - Problem: FFT type (1D, 2D, 3D), size (n), precision (p)
  - Platform: DRAM banks (b), row buffer size (R), max bandwidth (B), max power (P), max on-chip SRAM (S)
Design Space Exploration

There is no structured way of finding the best parameter set

- Parameters are dependent to each other and to problem/platform
- Tradeoff curves vary based on parameter set
- Conflicting tradeoffs construct an optimization problem

Can we explore the design space automatically?

8192x8192 2D-FFT Design Space

Constraints and fixed parameters:
- 8kx8k 2D-FFT
- Single precision, complex values
- 4 layer DRAM + 1 layer logic
- 8 banks/layer, 512 TSVs/bank
- Row buffer (R) = 1KB
- Max bandwidth (B) = 305GB/s

Design space parameters:
- \( r = 2 \) cpx words
- \( w = 2 \rightarrow 16 \) cpx words
- \( T = 0.125x \rightarrow 2x \) row-buffer (R)
- \( f = 0.4 \rightarrow 2 \) GHz
Overall Toolchain including Spiral

Inputs:
- Problem: DFT\textsubscript{512} complex single prec.
- Platform: # mem bank = 8, # of cores = ... etc

Main memory optimizations:
- Block data layouts
- Abstract machine model
  (Akin et al ICASSP’14)

Algorithm Derivation

HW/SW Code Generation

Performance Analysis
- Computation: Cycle-accurate
- DRAM: DRAMSim2, CACTI-3D

Power & Energy Analysis
- Logic/FP: Synopsys DC synthesis
- RAM/ROM: CACTI 6.5
- DRAM Controller: McPAT
- DRAM: DRAMSim2 & CACTI-3D

Actual implementation concerns:
- Fine tuning for DRAM (3D/reg.)
- Custom HW gen. backend
  (This work)

Outputs:
- Final system implementation
- Power & performance estimates
Design Tradeoffs: Tile Size

(a) Tile Size and Performance/Power Tradeoff (Off-chip DRAM, conf-A)

- **Perf. [GFLOPS]**
- **Power [W]**
- **Power Eff. [GFLOPS/W]**

- **Tile width (FFT size)**
  - (512x512 2D-FFT)
  - (2kx2k 2D-FFT)
  - (8kx8k 2D-FFT)
  - (32kx32k 2D-FFT)

- **Power efficiency**

- **DRAM power (off-chip)**
- **On-chip power**
- **Performance**
Design Tradeoffs: Tile Size

(b) Tile Size and Performance/Power Tradeoff (3D-stacked DRAM, conf-E)

- DRAM power (3D-stacked)
- On-chip power
- Performance

Power [W]

<table>
<thead>
<tr>
<th>Tile width (FFT size)</th>
<th>4</th>
<th>8</th>
<th>16</th>
<th>32</th>
<th>64</th>
</tr>
</thead>
<tbody>
<tr>
<td>(512x512 2D-FFT)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>(2kx2k 2D-FFT)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>(8kx8k 2D-FFT)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>(32kx32k 2D-FFT)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Power Eff. [GFLOPS/W]

- Power efficiency

<table>
<thead>
<tr>
<th>Tile width (FFT size)</th>
<th>4</th>
<th>8</th>
<th>16</th>
<th>32</th>
<th>64</th>
</tr>
</thead>
<tbody>
<tr>
<td>(512x512 2D-FFT)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>(2kx2k 2D-FFT)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>(8kx8k 2D-FFT)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>(32kx32k 2D-FFT)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Perf. [GFLOPS]

- DRAM power (3D-stacked)
- On-chip power
- Performance
Design Tradeoffs: Pareto Optimal Designs

(a) 2D, 3D and 1D-FFT (Off-chip DRAM, conf-A)

FFT size (log(N) for NxN 2D-FFT, NxNxN 3D-FFT and N-point 1D-FFT)

(b) 2D, 3D and 1D-FFT (3D-stacked DRAM, conf-E)

FFT size (log(N) for NxN 2D-FFT, NxNxN 3D-FFT and N-point 1D-FFT)
Design Tradeoffs: Closer look into DRAM

(a) DRAM Energy and Bandwidth Improvement (Off-chip)

- Energy Consumption [J/GB]
- Bandwidth Utilization [GB/s]
- Refresh
- Rd/Wr
- Act/Pre
- Static
- Bandwidth Util.

(b) DRAM Energy and Bandwidth Improvement (3D-stacked)
Performance Model Verification

- FPGA based implementation
  - Altera DE4 w/ 2xDDR2–800
  - @200MHz, 2.5MB SRAM
  - Test case & model verification

- Performance results from actual hardware implementations

<table>
<thead>
<tr>
<th>FFT</th>
<th>Prec. (bits)</th>
<th>TP (GFLOPS)</th>
<th>Perf (% of TP) (GFLOPS)</th>
<th>Model Est. (Error) (GFLOPS)</th>
</tr>
</thead>
<tbody>
<tr>
<td>256 × 256</td>
<td>32</td>
<td>32</td>
<td>23.25 (72.6%)</td>
<td>23.15 (-0.41%)</td>
</tr>
<tr>
<td>512 × 512</td>
<td>32</td>
<td>36</td>
<td>29.23 (81.2%)</td>
<td>29.31 (+0.27%)</td>
</tr>
<tr>
<td>1k × 1k</td>
<td>32</td>
<td>40</td>
<td>34.42 (86.0%)</td>
<td>34.74 (+0.94%)</td>
</tr>
<tr>
<td>2k × 2k</td>
<td>32</td>
<td>44</td>
<td>38.35 (87.2%)</td>
<td>39.54 (+3.10%)</td>
</tr>
<tr>
<td>4k × 4k</td>
<td>32</td>
<td>48</td>
<td>42.10 (87.7%)</td>
<td>43.89 (+4.25%)</td>
</tr>
<tr>
<td>256 × 256</td>
<td>64</td>
<td>16</td>
<td>12.47 (77.9%)</td>
<td>12.53 (+0.48%)</td>
</tr>
<tr>
<td>512 × 512</td>
<td>64</td>
<td>18</td>
<td>14.97 (83.2%)</td>
<td>15.10 (+0.85%)</td>
</tr>
<tr>
<td>1k × 1k</td>
<td>64</td>
<td>20</td>
<td>17.19 (86.0%)</td>
<td>17.40 (+1.19%)</td>
</tr>
<tr>
<td>2k × 2k</td>
<td>64</td>
<td>22</td>
<td>19.23 (87.4%)</td>
<td>19.50 (+1.40%)</td>
</tr>
<tr>
<td>128 × 128 × 128</td>
<td>32</td>
<td>28</td>
<td>23.40 (83.6%)</td>
<td>23.69 (+1.22%)</td>
</tr>
<tr>
<td>256 × 256 × 256</td>
<td>32</td>
<td>32</td>
<td>26.89 (84.0%)</td>
<td>27.24 (+1.30%)</td>
</tr>
<tr>
<td>512 × 512 × 512</td>
<td>32</td>
<td>36</td>
<td>30.30 (84.2%)</td>
<td>30.70 (+1.32%)</td>
</tr>
</tbody>
</table>

73–88% max 4%
Summary

- **Hardware FFT accelerators**
  - Memory subsystem is the key -> DRAM-optimized FFT algorithms

- **Careful mapping algorithms to architectures**
  - Crucial for high performance and energy efficiency
  - Various platforms different tradeoff relations
  - Automated design optimization and exploration

- **Pareto-optimal designs**
  - 10x bandwidth, 5.5x energy eff. in DRAM
  - 6x performance, 6.5x energy eff. in system

- **Actual hardware implementations**
  - Efficient utilization of the platform
  - High accuracy of the performance model
Understanding the Design Space of DRAM–Optimized Hardware FFT Accelerators

Berkin Akın, Franz Franchetti, James C. Hoe

Electrical and Computer Engineering Department
Carnegie Mellon University

(This work was sponsored by DARPA under agreement HR0011-13-2-0007. The content, views and conclusions presented in this document do not necessarily reflect the position or the policy of DARPA.)