Large size 2D Fast Fourier Transform

- Used in image processing, scientific computing
- Typical datasets are large and high precision!
  - e.g., 2K-by-2K double precision 2D-FFT:
    - Input dataset: 64 MB
    - # of operations: ~461.4 MFlop
- Does not fit on-chip
- Stored off-chip

Memory access pattern and achieved bandwidth

- Have large strided DRAM access pattern
- Does not exploit DRAM row-buffer locality
- Results in low memory bandwidth utilization!
- Memory bandwidth becomes bottleneck for achieving high performance
- Effective bandwidth orchestration is required for:
  1. Performance
  2. Bandwidth Efficiency
  3. Power Efficiency

2D-FFT algorithms

- Row column algorithm:
  \[ DFT_{m,n} = (DFT_a \circ I_a) (I_b \circ DFT) \]
  Row-wise and column-wise accesses!
- \( DFT_{m,n} = \prod_{l=0}^{N-1} (I_{l,b} \circ DFT_{a,l}) \)
  2D FFT operates on 2D data, e.g., images

DRAM operation

- Need to make use of every row touched to maximize bandwidth
- Large strides result in small packets of transferred data
- Double-buffering:
  - Overlapped computation and I/O
  - All modules kept busy

Solution: Algorithm and Architecture

- Data is accessed as tiles, not row and column-wise
- Row-buffer misses are minimized!

From algorithm to hardware

- Double-buffering:
  - Overlapped computation and I/O
  - All modules kept busy
- Matching throughput to memory bandwidth:
  - Achieved via fine-grain control over datapath parallelism
  - Results in balanced design
- Ensuring continuous dataflow:
  - Buffers are used to smooth the flow of data.

Evaluation

Target application:
- Double-buffering complex 2D-FFT
- Data sizes up to 2,048-by-2,048

Target platforms:

<table>
<thead>
<tr>
<th>2D-FFT Raw performance (double precision)</th>
<th>2D-FFT Bandwidth Efficiency (double precision)</th>
<th>2D-FFT Power Efficiency (double precision)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Performance</td>
<td>Problem Size</td>
<td>Performance</td>
</tr>
<tr>
<td>64x64x128</td>
<td>GTX 460</td>
<td>Core i7 960</td>
</tr>
<tr>
<td>256x256x128</td>
<td>GTX 460</td>
<td>Core i7 960</td>
</tr>
<tr>
<td>512x512x128</td>
<td>GTX 460</td>
<td>Core i7 960</td>
</tr>
<tr>
<td>1Kx1Kx1K</td>
<td>GTX 460</td>
<td>Core i7 960</td>
</tr>
<tr>
<td>2Kx2Kx2K</td>
<td>GTX 460</td>
<td>Core i7 960</td>
</tr>
</tbody>
</table>

The authors acknowledge the support of the C2S2 Focus Center, one of six research centers funded under the Focus Center Research Program (FCRP), a Semiconductor Research Corporation entity.