Energy- and Area-Efficient Parameterized Lifting-Based 2-D DWT Architecture on FPGA

Yusong Hu and Viktor K. Prasanna

Ming Hsieh Department of Electrical Engineering

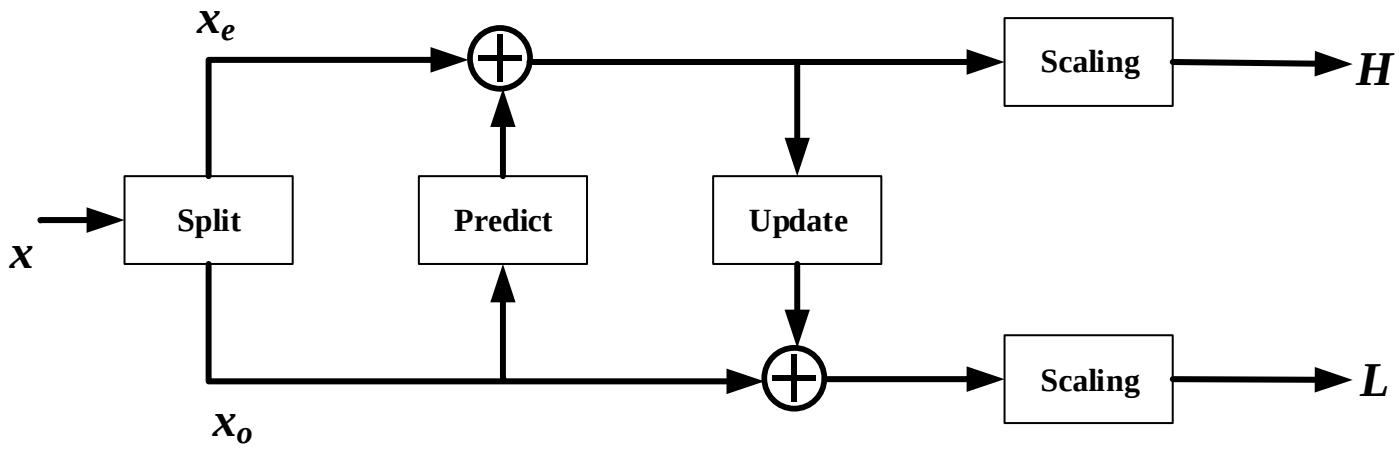
University of Southern California

Outline

- Introduction
- Contributions
- Architecture and Implementation
- Experimental Results and Analysis
- Conclusion and Future Work

Discrete Wavelet Transform (DWT) (1)

- DWT: representing the signal by wavelet bases
- Lifting-based DWT
  - Input signal *x* is processed by several lifting steps
  - Lifting steps: 1 multiplier and 2 adders



Discrete Wavelet Transform (DWT) (2)

- Lifting-based 5/3 and 9/7 DWT algorithm
  - 5/3 and 9/7 are widely used wavelet base
  - *K* lifting steps
  - *K*=2,4 for 5/3 and 9/7 filter, respectively
- **External and on-chip memory dominate the area and energy consumption!**

Related Work

- Many designs focus on memory efficiency
  - Memory efficiency: size of on-chip memory
  - Line-based architectures
    - Read image in a line-by-line order
    - Large transposition memory
  - Overlapped stripe-based
    - Some pixels are read multiple times
    - Significant image memory read overhead
- Energy efficiency not considered as a key performance metric

Outline

- Introduction
- Contributions
- Architecture and Implementation
- Experimental Results and Analysis
- Conclusion and Future Work

Contributions (1)

- An overlapped block-based image scanning method
  - Optimizes the number of external accesses and on-chip memory size
  - On-chip memory size reduced to constant
  - Improves the energy efficiency
- An energy-efficient parameterized 2-D DWT architecture
  - Implementation of 5/3 and 9/7 filters on FPGAs

Contributions (2)

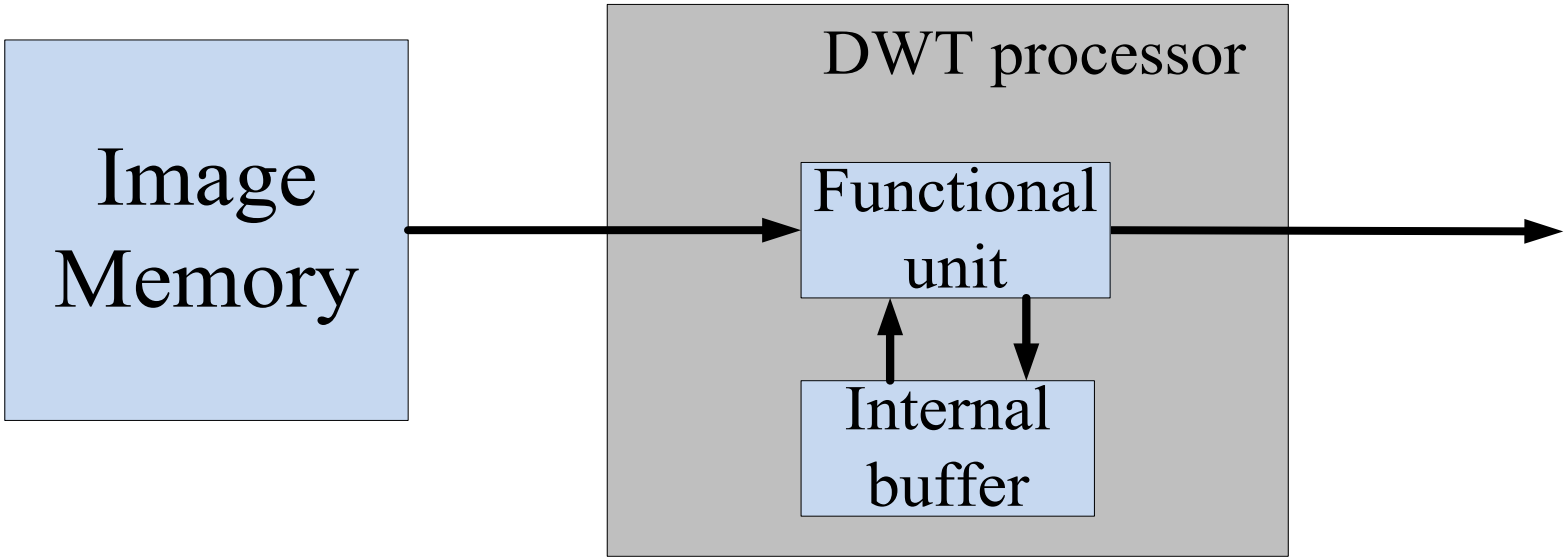
- DRAM activation schedule
  - Reduce the energy consumption of external memory by minimizing the number of row activations
- Implementations show significant improvement
  - Energy efficiency and composite metric (EAT) compared with the state-of-the-art

Outline

- Introduction
- Contributions
- Architecture and Implementation
- Experimental Results and Analysis
- Conclusion and Future Work

Design (1)

- DWT architecture
  - Composed of image memory (DRAM) and DWT processor (FPGA)

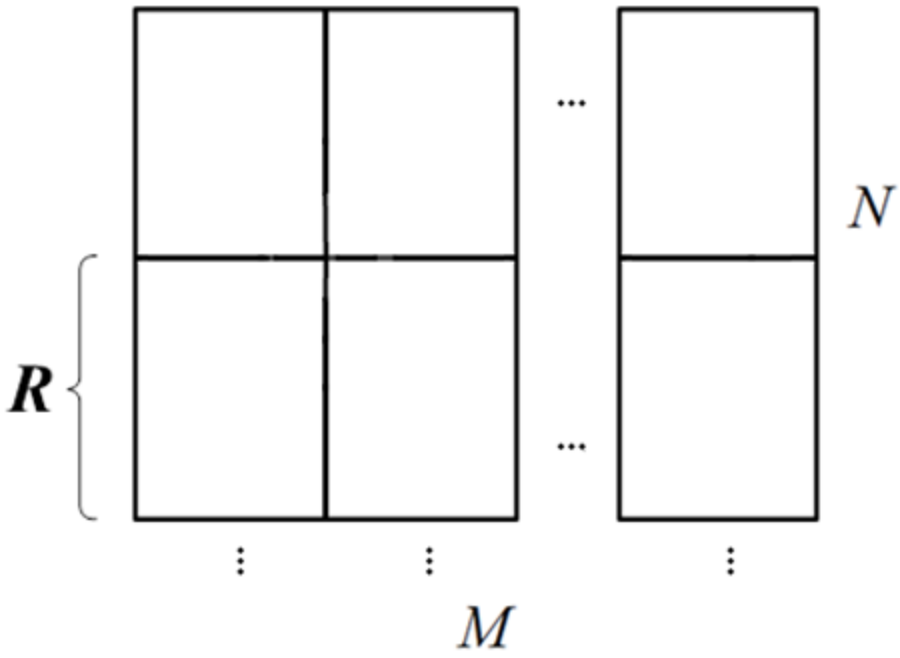


Design (2)

- Operations
  - Computation operations: one multiplication
  - Memory-access operations: read/write to the on-chip memory or image memory
  - Input image of size *MN*
  - At least 2*MN* and 4.5*MN* computations required for 5/3 and 9/7 filters
  - At least *MN* image memory accesses for an image of size *MN*

Design (3)

- Algorithm-mapping parameters
  - Parallelism (*L*): number of computations performed every clock cycle
  - Height of image block (*R*)



Performance metrics (1)

- Energy efficiency
  - Number of operations performed per unit of energy (GOPS/sec/W)

$$\eta = C/E = C/Pt$$

  - C: minimum required computation operations
  - E: energy consumed for processing one frame
  - P: average power dissipated by the architecture
  - t: time for processing one frame

Performance metrics (2)

- Composite metric
  - Evaluate the impact of energy efficiency on area and throughput
- Example: Energy × Area × Time (EAT)
  - Energy: energy consumed per frame
  - Area: area of the architecture
  - Time: computation time per frame

Performance metrics (3)

- Sustained energy efficiency
  - Ratio of energy efficiency of the design to the peak performance of the target platform
- Peak Performance
  - Energy consumed to perform the required operations
  - Ignore overheads such as I/O, control logic, routing, on-chip buffers

$$\eta_{peak} = C/(P_c + P_t)t$$

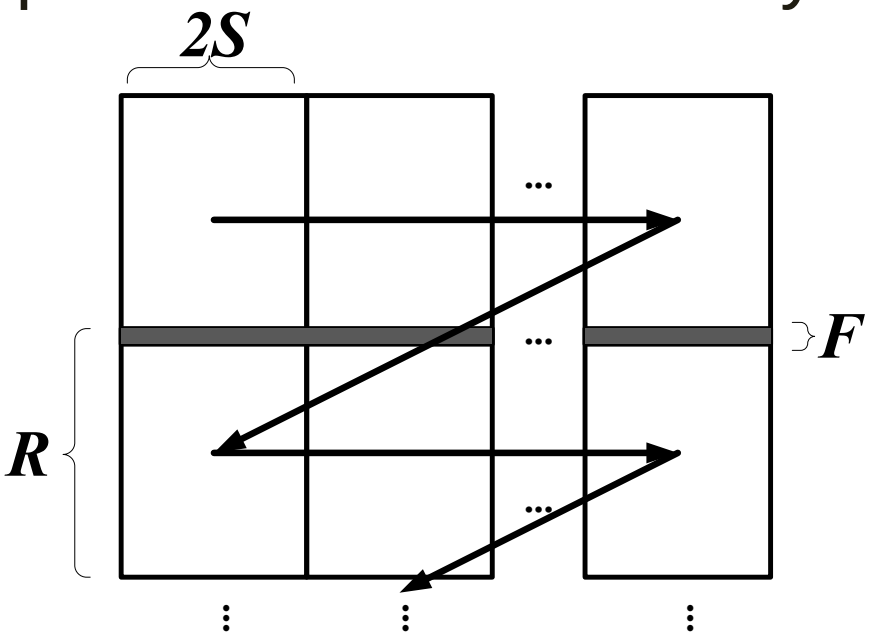
  - P_c and P_t are the average power dissipated by computation and memory-access operations

Overlapped block-based scanning method (1)

- Partition the image into blocks
  - Blocks of width 2*S* and height *R*
  - *F* rows overlap between vertically adjacent blocks
  - *F* = 3 and 7 for the 5/3 and 9/7 filter
  - Blocks are processed one-by-one in a row major order

Overlapped block-based scanning method (2)

- Within each block, the pixels are scanned in column major order
- 2*S* pixels input to architecture every clock cycle

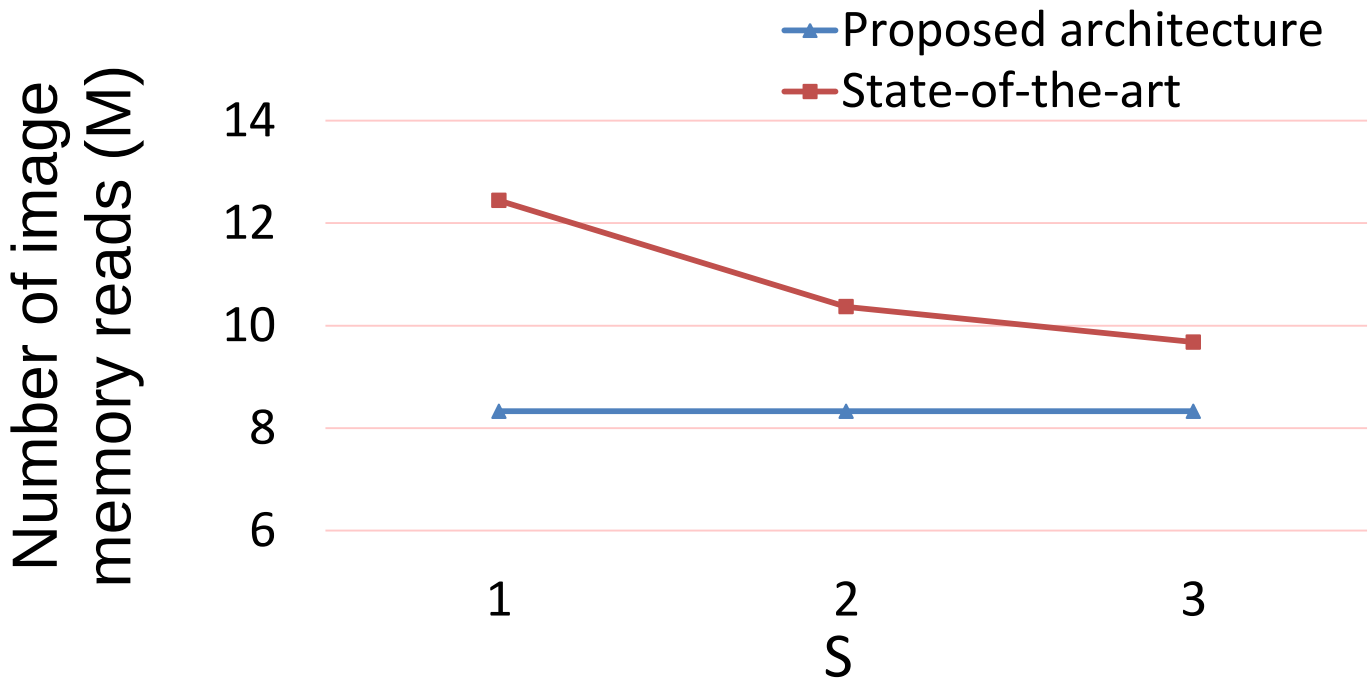


Overlapped block-based scanning method (3)

- Reduces the number of memory accesses
  - Slightly increased number of computation operations
  - $MN + MF\frac{N-R}{R-F}$ image memory read for image of size MN
  - Number of image memory reads reduced by up to 35% compared with state-of-the-art design

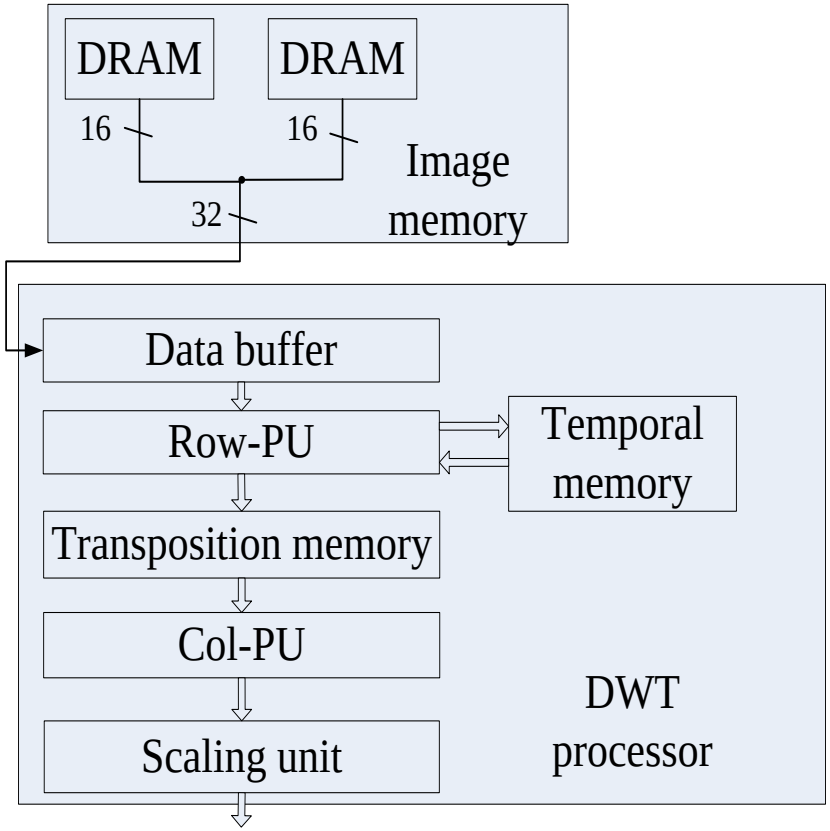
Overlapped block-based scanning method (4)

– **Much less energy consumption for image memory access**



Energy-efficient lifting-based 2-D DWT architecture (1)

- Overall architecture
  - DWT Processor (on FPGA)
  - Image memory (two DRAM chips)

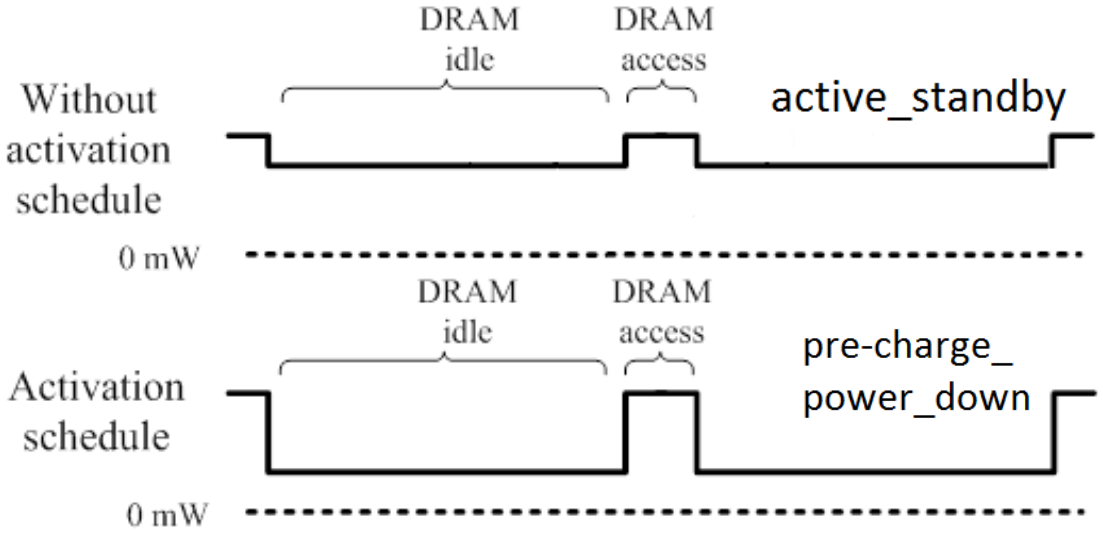


Energy-efficient lifting-based 2-D DWT architecture (2)

- DRAM activation schedule
  - DRAM power = Active power + read/write term power + background power
  - Background power depends on power mode
    - Pre-charge_power-down: lowest power dissipation
    - Pre-charge_standby
    - Active_power-down
    - Active_standby: highest power dissipation

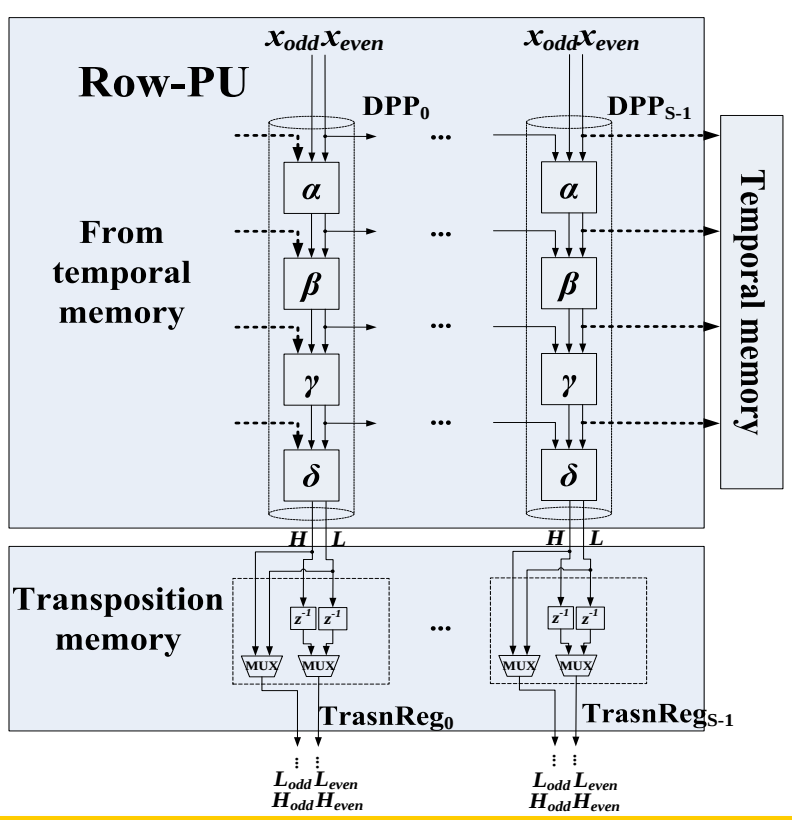
Energy-efficient lifting-based 2-D DWT architecture (3)

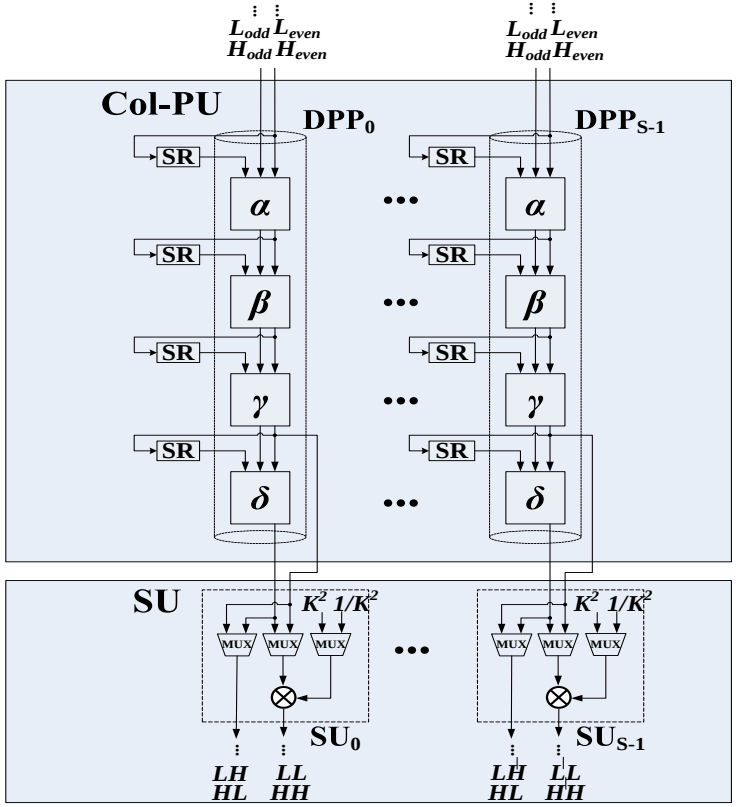
- Minimize the background power dissipation of DRAM
- Switch the DRAMs from active_standby to pre-charge_power-down when DRAMs are not accesse



Energy-efficient lifting-based 2-D DWT architecture (4)

- Parameterized DWT architecture





Energy-efficient lifting-based 2-D DWT architecture (5)

- *S* data processing pipes (DPP) in Row-PU and Col-PU
- 5/3 and 9/7 filter: $L = 4S$ and $9S$
- On-chip memory size of $4R$
- **Less energy consumption of on-chip memory**
- **Area efficiency**
  - On-chip memory size reduced to constant
  - Smaller on-chip memory compared with state-of-the-art architecture

Outline

- Introduction
- Contributions
- Architecture and Implementation
- Experimental Results and Analysis
- Conclusion and Future Work

Experimental Setup

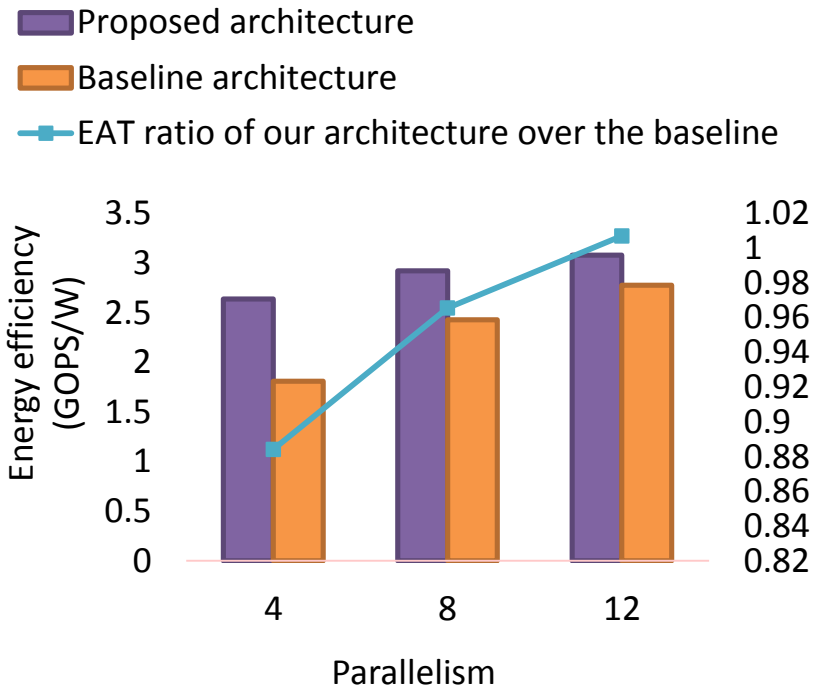
- Target Platform:
  - Xilinx Virtex 7 XC7VX980, -2L speed grade
- Tools
  - Xilinx Virtex 7 XC7VX980, -2L speed grade
  - Xilinx Vivado Power Analysis Tool
- VCD with 50% toggle rate
- Baseline: Design with no optimizations incorporated
- 30 frames per second

Peak energy efficiency

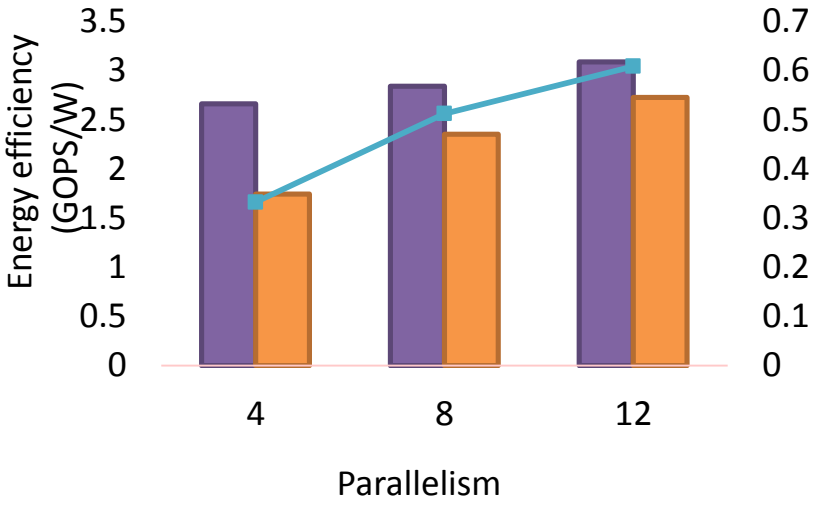
- Image memory accessed at peak bandwidth (1600M 32-bit pixels/sec. & 367.9 mW)
- 32-bit multiplier dissipates 6 mW at 200 MHz
- 16 and 36 multipliers required for 5/3 and 9/7 filter to consume the pixels
- $\eta_{peak\ 9/7} = \frac{4.5MN}{(36\times6+2\times367.9)\frac{4.5MN}{4.5\times1600}} = \mathbf{7.56\ GOPS/J}$
- $\eta_{peak\ 5/3} = \frac{2MN}{(16\times6+2\times367.9)\frac{2MN}{2\times1600}} = \mathbf{3.85\ GOPS/J}$

Performance comparison (1)

- Energy efficiency of 5/3 filter implementation



Small Input



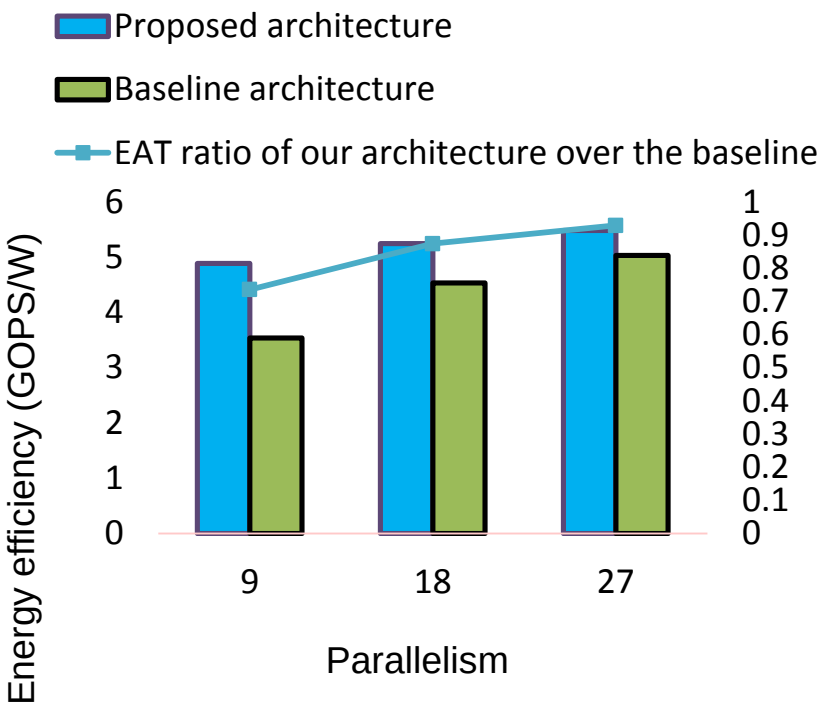
Large Input

Performance comparison (2)

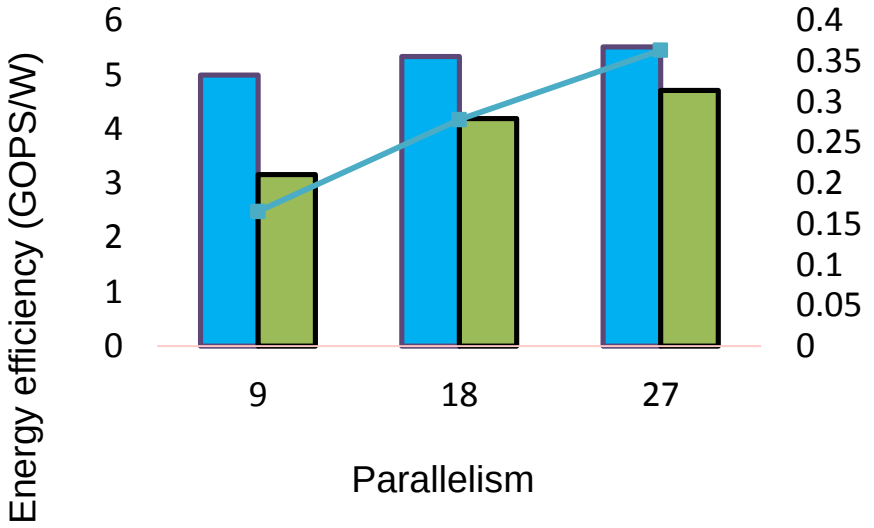
- Better energy efficiency and EAT than state-of-the-art architecture
- Energy efficiency improvement: larger for smaller *S*
- EAT improvement: larger for larger images
- Up to 80.2% of the peak energy efficiency

Performance comparison (3)

- Energy efficiency of 9/7 filter implementation



Small input



Large input

Performance comparison (4)

- Better energy efficiency and EAT than state-of-the-art architecture
- Energy efficiency Improvement: larger for smaller *S*
- EAT Improvement: larger for larger images
- Improvement of EAT is larger than 5/3 filter
- Up to 72.9% of the peak energy efficiency

Conclusion

- Overlapped block-based scanning method
- Parameterized DWT architecture: *L* and *R*
- Optimizations:
  - Reduced number of memory accesses
  - On-chip memory maintained constant
  - DRAM activation schedule
- Energy Efficiency: up to 58% improvement compared with the state-of-the-art design
- Future work
  - Energy efficient multi-level DWT architecture

Thank You!

Questions?

Ganges.usc.edu/wiki/TAPAS