A 249Mpixel/s HEVC Video-Decoder Chip for Quad Full HD Applications

C-T. Huang¹, M. Tikekar¹, C. Juvekar¹, V. Sze², A. P. Chandrakasan¹

¹Massachusetts Institute of Technology
²Texas Instruments
High Efficiency Video Coding (HEVC)

- Decoding Complexity
- Coding Efficiency

**H.265/HEVC** (2013)
- 64x64 LCU (Largest Coding Unit)
- Hierarchical Coding Structure

- 16x16 Macroblock

**MPEG-4** (1999)
- Macroblock

**MPEG-2** (1994)
- Macroblock

- Pipeline SRAM ↑
- 16x
Hierarchical Coding Structure

Three Possible LCU Sizes

- **64x64**
- **32x32**
- **16x16**

Coding Unit (CU) Tree with mixed intra/inter CUs

- CU size up to 64x64
- CU tree depth up to 3
- TU tree depth up to 2

- Intra CU
- Inter CU
- Prediction Unit (PU)
- Transform Unit (TU)
Design Challenges

• For an HEVC decoder, the following issues, besides complexity, need to be addressed:
  – On-chip memory size
  – Various coding flows
  – PU and TU diversity
  – DRAM throughput

  → Variable-Size System Pipeline
  → Unified Processing Engines
  → Four-Parallel MC Cache
Support Quad Full HD decoding with all coding structures
Variable-Size System Pipeline

Accommodate the uncertain MC cache latency

Group I
- Entropy Decoder
- MC Dispatch

Group II
- Inverse Transform
- Prediction
- Deblock
- REC DMA

Coefficients in TU FIFO

Variable-size Pipeline Block (VPB)

LCU
- 64x64
- 32x32
- 16x16
Variable-Size System Pipeline

Accommodate the uncertain MC cache latency

Group I
- Entropy Decoder
- MC Dispatch

Group II
- Inverse Transform
- Prediction
- Deblock
- REC DMA

Coefficients in TU FIFO

64 Y U V

32

Save 20KB Coeff SRAM
Two-Stage Sub-Pipelining

Sub-PPB Pipelining Saves 36KB Ref Pixel SRAM

PPB: Prediction Pipeline Block

MC Dispatch

MC Cache Output

Prediction

Cache Latency

Sub-PPB Pipelining

64x64 64x32 64x16

PPB
(Stage 1)

Sub-PPB
(Stage 2)

0 1 2 3 4 5

Y U/V

0 1 2 3

4 5

0 1 2 3

0 1 2 3

0 1 2 3

PPB_0 PPB_1 PPB_2 PPB_3

PPB_0 PPB_1 PPB_0

PPB_0 PPB_1 PPB_0

PPB_0 PPB_1 PPB_0
Unified Prediction Engine

Mode-adaptive scheduling for 36 modes

Unified Y 8-tap and U/V 4-tap filter

Inter CU

Intra CU

Intra Core

Inter Core

VPB Boundary

LMChroma Buffer

REC Core

Ref Pixel

Residue

Unified Reconstruction Path

4 Row Pixels

REC Pixel
HEVC Inverse Transform (WD 4)

Transform unit (TU) sizes:
- \( TU_{32 \times 32} \)
- \( TU_{8 \times 32} \)
- \( TU_{16 \times 16} \)
- \( TU_{16 \times 4} \)
- \( TU_{4 \times 16} \)
- \( TU_{32 \times 8} \)
- \( TU_{8 \times 8} \)
- \( TU_{4 \times 4} \)

Transpose memory size:
- HEVC 32x32
- H.264 8x8

1-D transform complexity:
- HEVC
  - 4, 8, 16, 32-point transform
  - 8b integer precision
- H.264
  - 4, 8-point transform
  - 5b integer precision

1-D transform sizes:
- 4, 8, 16, 32-point IDCT
- 4-point IDST
Transform Top-Level Architecture

- One 4-point IDCT or IDST per clock
- Partial 8, 16, 32-point IDCT with accumulator
- Instead of register array of 125 Kgates, four single-port SRAMs are used for transpose memory (2KB)
Variable-Size Partial 1D Transform

Recursive Even-Odd Matrix Decomposition
## Area Optimization for Partial Matrix

### Partial 4x4 matrix

Observation: 4x4 matrix has only 4 unique coefficients

\[
\begin{bmatrix}
  y_0 \\
  y_1 \\
  y_2 \\
  y_3 \\
\end{bmatrix} =
\begin{bmatrix}
  89 & 50 & 75 & 18 \\
  75 & -89 & -18 & -50 \\
  50 & 18 & -89 & 75 \\
  18 & 75 & -50 & -89 \\
\end{bmatrix}
\begin{bmatrix}
  x_0 \\
  x_2 \\
  x_1 \\
  x_3 \\
\end{bmatrix}
\]

### Generic Design

\( i = 0 \ldots 3 \)

\[
\begin{array}{cccc}
  89 & 50 & 75 & 18 \\
  75 & -89 & -18 & -50 \\
  50 & 18 & -89 & 75 \\
  18 & 75 & -50 & -89 \\
\end{array}
\]

\( x_i \)
Area Optimization for Partial Matrix

Partial 4x4 matrix

Observation: 4x4 matrix has only 4 unique coefficients

\[
\begin{bmatrix}
  y_0 \\
  y_1 \\
  y_2 \\
  y_3 \\
\end{bmatrix} = \begin{bmatrix}
  89 & 50 & 89 & 50 \\
  75 & -89 & -18 & -50 \\
  50 & 18 & 75 & 75 \\
  18 & 75 & -89 & -50 \\
\end{bmatrix} \begin{bmatrix}
  x_0 \\
  x_1 \\
  x_2 \\
  x_3 \\
\end{bmatrix}
\]

Generic Design

HEVC Specific Design

Permute and Negate

\( i = 0 \ldots 3 \)
Area Optimization for Partial Matrix

- Replace 4 multipliers with 4 adders in Partial 4x4 matrix
- Partial 8x8 matrix uses 8 adders
- Partial 16x16 matrix uses 13 adders

<table>
<thead>
<tr>
<th>Partial Transform</th>
<th>Area for Generic Design (Kgates)</th>
<th>Area for Proposed Design (Kgates)</th>
<th>Savings</th>
</tr>
</thead>
<tbody>
<tr>
<td>4x4</td>
<td>10.7</td>
<td>7.3</td>
<td>32%</td>
</tr>
<tr>
<td>8x8</td>
<td>23.2</td>
<td>13.5</td>
<td>42%</td>
</tr>
<tr>
<td>16x16</td>
<td>46.7</td>
<td>34.4</td>
<td>26%</td>
</tr>
</tbody>
</table>

- Total area reduced from 96K to 71K gates.
High Throughput Requirement for HEVC MC (WD 4)

For 4x4 prediction, reference block increased by 49%

Addressed by Four-parallel MC cache

24 cache lines in 8 cycles

x2 for bi-direction
Four-Parallel Cache-Line Mapping

Each cache line column of four is mapped to one tag register file and one cache SRAM.

Addressed by Four-parallel MC cache
Cache Top-level Architecture

- Command FIFO
- Memory Interface Arbiter
- Read FIFO

- DMA Control
- Reordered Hit/Miss Resolution
- Address Queues
- Two-Port Cache SRAM

- DRAM Address Mapping
- Collison-free mapping and reordering
- Cache Tag Register File
- Address Queue
  - Address Valid
  - Race Hazard Detection
  - Cache Address Index

- Four-Parallel MC Cache

MC Dispatch

To Ref Pixel SRAM
Cache Configurations

- **Cache SRAM Size (LCU 64x64)**
  - Hit Rate:
    - 4KB: 40%
    - 8KB: 50%
    - 16KB: 61%
    - 32KB: 70%

- **Set-Associativity (LCU 64x64)**
  - Way:
    - 1-way: 40%
    - 2-way: 50%
    - 4-way: 60%
    - 8-way: 70%

- **Cache Line Configurations**
  - **Cache Line 8x4**
  - **Cache Line 4x8**
  - LCU 64x64
  - LCU 32x32
  - LCU 16x16

- **16KB Cache SRAM**
  - 4-Way Associative
  - Cache Line 8x4

- Hit Rate:
  - +9%
  - +10%
Twisted 2D DRAM Mapping

Precharge and activate DRAM Bank when changing DRAM row

Reference Frame (Y or U/V)

Cache line sub-tiling

8 Bank Row with 64x64 Pixel Banks

64x64 pixels

8x4 Pixels

0x00 0x01 ... 0x07
0x08
0x10
0x78
Twisted 2D DRAM Mapping

Twisting reduces DRAM bank activation by increasing the distance for the same bank.

8 Bank Row with 64x64 Pixel Banks

Reference Frame (Y or U/V)
MC Bandwidth Reduction

3.8GB/s

2.3GB/s

1.2GB/s

-67% by Proposed Mapping and Cache

-56% by Cache

-70% by DRAM Mapping

Cache Line RS & Sub-PPB Share

Cache Line RS & Cache

Twisted 2-D & Cache

RS: Raster Scan

- ACTIVATION

- DATA
DRAM Power Saving

MC

Non-MC

Standby

Cache Line Raster Scan & Sub-PPB Share

This Work

DDR3 Power @ 400MHz, 1.5V for 3840x2160 30fps Decoding

341mW

-122mW by MC Cache and DRAM mapping

219mW

Modeled by [5] with 2x 16b DDR3 DRAM

HEVC Decoder Chip

Technology : 40-nm CMOS
Core Area    : 1.33x1.33mm²
Gate Count   : 715K
On-chip SRAM : 124KB
Power Measurements

Core Power at 0.9V (mW)

- 200MHz, Avg: 76mW (3840x2160 30fps)
- 100MHz, Avg: 51mW (1920x1080 60fps)
- 25MHz, Avg: 31mW (1280x720 30fps)

Similar results for six coding configurations
## Video Decoder Comparison

<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Max Specification</strong></td>
<td>3840x2160 @30fps</td>
<td>7680x4320 @60fps</td>
<td>4096x2160 @24fps</td>
<td>1920x1080 @30fps</td>
</tr>
<tr>
<td><strong>Gate Count</strong></td>
<td>715K</td>
<td>1338K</td>
<td>414K</td>
<td>160K</td>
</tr>
<tr>
<td><strong>On-Chip SRAM</strong></td>
<td>124KB</td>
<td>80KB</td>
<td>9KB</td>
<td>5KB</td>
</tr>
<tr>
<td><strong>Technology</strong></td>
<td>40nm/0.9V</td>
<td>65nm/1.2V</td>
<td>90nm/1.0V</td>
<td>0.18μm/1.8V</td>
</tr>
<tr>
<td><strong>Normalized Core Power</strong></td>
<td>0.31nJ/pixel</td>
<td>0.21nJ/pixel</td>
<td>0.28nJ/pixel</td>
<td>5.11nJ/pixel</td>
</tr>
<tr>
<td><strong>Normalized DRAM Power</strong></td>
<td>0.88nJ/pixel**</td>
<td>1.27nJ/pixel</td>
<td>N/A</td>
<td>N/A</td>
</tr>
<tr>
<td><strong>Normalized System Power</strong></td>
<td>1.19nJ/pixel</td>
<td>1.48nJ/pixel</td>
<td>N/A</td>
<td>N/A</td>
</tr>
<tr>
<td><strong>DRAM Configuration</strong></td>
<td>32b DDR3</td>
<td>64b DDR2</td>
<td>N/A</td>
<td>32b DDR + 32b SDR</td>
</tr>
</tbody>
</table>

* Power for max specification
** Modeled by [5]
*** System Power = Core Power + DRAM Power
# Video Decoder Comparison

<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>Max Specification</td>
<td>3840x2160 @30fps</td>
<td>7680x4320 @60fps</td>
<td>4096x2160 @24fps</td>
<td>1920x1080 @30fps</td>
</tr>
<tr>
<td>Gate Count</td>
<td>715K</td>
<td>1338K</td>
<td>414K</td>
<td>160K</td>
</tr>
<tr>
<td>On-Chip SRAM</td>
<td>124KB</td>
<td>80KB</td>
<td>9KB</td>
<td>5KB</td>
</tr>
<tr>
<td>Technology</td>
<td>40nm/0.9V</td>
<td>65nm/1.2V</td>
<td>90nm/1.0V</td>
<td>0.18μm/1.8V</td>
</tr>
<tr>
<td>Normalized Core Power*</td>
<td>0.31nJ/pixel</td>
<td>0.21nJ/pixel</td>
<td>0.28nJ/pixel</td>
<td>5.11nJ/pixel</td>
</tr>
<tr>
<td>Normalized DRAM Power**</td>
<td>0.88nJ/pixel</td>
<td>1.27nJ/pixel</td>
<td>N/A</td>
<td>N/A</td>
</tr>
<tr>
<td>Normalized System Power***</td>
<td>1.19nJ/pixel</td>
<td>1.48nJ/pixel</td>
<td>N/A</td>
<td>N/A</td>
</tr>
<tr>
<td>DRAM Configuration</td>
<td>32b DDR3</td>
<td>64b DDR2</td>
<td>N/A</td>
<td>32b DDR + 32b SDR</td>
</tr>
</tbody>
</table>

* Power for max specification  
** Modeled by [5]  
*** System Power = Core Power + DRAM Power

Lowest system power achieved for HEVC Decoding
Conclusions

• Variable-size system pipeline provides a unified control flow and on-chip SRAM optimization.
• Unified prediction and transform engines support the diversity of PU and TU in area-efficient ways.
• Four-parallel 16KB cache with a twisted 2D mapping saves 67% DRAM bandwidth for MC.
• Demonstrated an HEVC decoder chip with a low system power 1.19nJ/pixel.

This work was funded by Texas Instruments. The authors thank the TSMC University Shuttle Program for chip fabrication.