# Programmable Data Parallel Accelerator for Mobile Computer Vision

*T Nyländen, H Kultala[♯], I Hautala, J Boutellier, J Hannuksela, O Silvén*

*University of Oulu,*
*[♯] Tampere University of Technology*

CSE
Department of Computer Science and Engineering
Tietotekniikan Osasto

UNIVERSITY of OULU
OULUN YLIOPISTO

# Outline

- Introduction
- Application use cases
- The proposed accelerator
- Experiments
- Discussion and future work

# Introduction

# Introduction (1/3)

- Performance and power efficiency requirements for mobile devices are constantly increasing
- One of the strongly growing application areas is *computer vision (CV)*
- Hardware acceleration for mobile CV exists, however no programmable solutions
- Mobile GPGPUs offer a possibility for CV acceleration

# Introduction (2/3)

- Mobile GPGPUs are promising for CV
  - \+ High performance
  - \+ Programmable via OpenCL or Cuda
- Unfortunately
  - – Poor performance portability
    - Major architectural differences
  - – Device-specific software optimization due to
    - Arithmetic formats
    - Memory models
  - – High power consumption
    - DVFS enabled quickly in practice

# Introduction (3/3)

- We propose a programmable accelerator for mobile computer vision
- The accelerator is based on the Transport Triggered (TTA) computer architecture
- Experiments show that the proposed solution clearly outperforms commercial mobile GPUs in efficiency

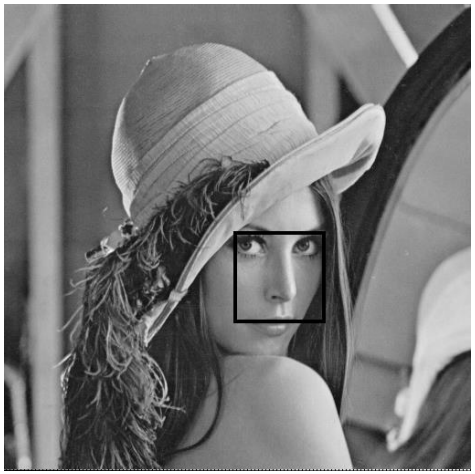
# Application use cases

# Application use cases (1/2)

- Depth Estimation (Stereo disparity)
  - *SAD algorithm → high number of computations*
  - *Highly parallel with very little control code*
  - *Image size: 427x327*
  - *Block size: 8x8*

# Application use cases (2/2)

- Face detection
  - *Viola Jones based Haar-cascade classifier*
  - *Contains code structures that are hard to parallelize*
  - *A lot control code, little room for vectorization*
  - *Image size: 512x512*
  - *Window size: 20*

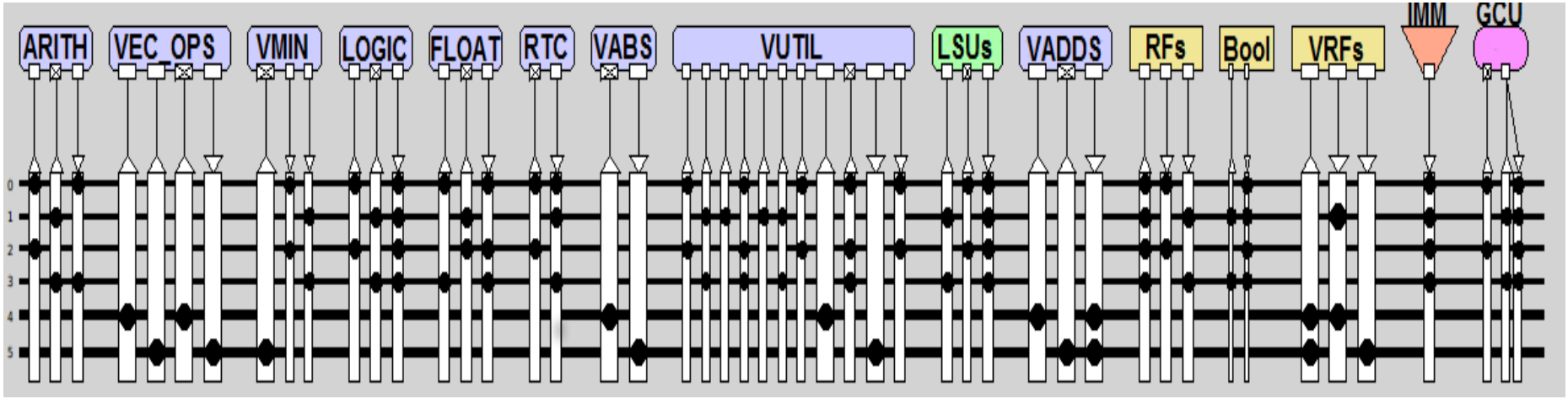


# The proposed accelerator

# The proposed accelerator (1/3)

- The programmable accelerator was designed with the TCE toolset by Tampere University of Technology
- Cycle-accurate profiling based on simulation
  - Allows removing FUs with low utilization
- TTA accelerators have previously been designed for telecommunications, video and audio coding

# The proposed accelerator (2/3)




- Design guideline: maximal performance per Watt
- Function units for (half-precision) floating point, integers and SIMD operations

# The proposed accelerator (3/3)

| Unit | Nro of units | Area (kGE) | Area ($\mu m^2$) | Power consumption (mW) |
|---|---|---|---|---|
| Integer scalar | 9 | 15 | 7023 | 1.2 |
| Float scalar | 6 | 42 | 20690 | 11.27 |
| Half Vector | 4 | 20 | 9732 | 6.73 |
| LSUs | 3 | 24 | 13730 | 2.69 |
| RFs | 4 | 17 | 8475 | 0.69 |
| Vector RFs | 4 | 16 | 8222 | 0.91 |
| Memories | 3 | 273 | 133879 | 6.69 |
| TOTAL | 35 | 436 | 213909 | 39.70 |

28 nm low power technology, post-place & route

# Experiments

# Experiments (1/3): Depth Estimation

| Platform | Throughput (fps) | Throughput/W (fps/W) |
|---|---|---|
| AMD Radeon HD 7750 (Desktop GPU) | 179.5 | 8.6 |
| Intel Core i5-480M (Desktop GPP) | 3.0 | 0.1 |
| Qualcomm  Adreno 330 (Mobile GPU) | 9.8 | 5.4 |
| *Proposed* | 11.6 | 292.2 |

# Experiments (2/3): Face Detection

| Platform | Throughput (fps) | Throughput / W (fps/W) |
|---|---|---|
| Radeon HD 7750 (Desktop GPU) | 31.2 | 0.6 |
| Intel Core i5-480M (Desktop GPP) | 30.7 | 0.9 |
| Qualcomm Adreno 330 (Mobile GPU) | 1.7 | 0.9 |
| Qualcomm Krait 400 (Mobile GPP) | 11.1 | 4.1 (TDP) |
| *Proposed* | 3.6 | 90.7 |

# Experiments (3/3): Generic Applications

CHStone benchmark results: execution time (μS) for each processor @ 100 MHz

| Test | Proposed TTA | Nios II | 5-stage mBlaze |
|---|---|---|---|
| sha | 3298.39 | 6040.46 | 14933.69 |
| blowfish | 4884.98 | 10854.65 | 16714.82 |
| aes | 270.33 | 591.34 | 734.37 |
| motion | 54.18 | n/a | n/a |
| jpeg | 25806.71 | 180017.48 | 231131.94 |
| gsm | 129.3 | 198.77 | n/a |
| adpcm | 463.24 | 1030.78 | 1562.93 |

# Discussion and future work

# Discussion and future work

- Achieved power efficiency of the core outperforms existing commercial solutions
- Achieving real-time performance for video processing would require multiple instances of the proposed accelerator
- Potential future work includes redesigning the accelerator to enable acceleration of baseband processing

Thank you for your attention. Questions?