Designing Reconfigurable Hardware for a Scalable Wavelet Video Decoder

Prof. Dirk Stroobandt

Dr. Mark Christiaens, ir. Hendrik Eeckhaut, ir. Harald Devos

Research Group PARIS
Department ELIS
Ghent University

- The RESUME project: motivation and goals
- Scalable video characteristics
- Software transformations
- Hardware implementations of basic components
- QTL and hardware: the conflicts
- A new Wavelet Entropy Codec
- Results
- Conclusions and future work

Confidential

RESUME

Reconfigurable **E**mbedded **S**ystems

for **U**se in scalable **M**ultimedia **E**nvironments

Theme: reconfigurable hardware

Theme: multimedia systems negotiation

Theme: operating system/middleware layer.

Theme: video information (de)coding.

http://www.elis.UGent.be/resume/

Confidential

RESUME users' committee

Confidential

Multimedia applications (on mobile terminals)

Demand flexibility, changes during execution

Demand huge computational power (real time)

Software (SW)

Conflict

Hardware (HW)

Very flexible
Low parallelism
Too slow

Flexible
High parallelism
Ideal

Fixed
High parallelism
Too expensive

Other option: FPGAs

"*Field Programmable Gate Array*": Large array of **hardware** blocks with reconfigurable functions and interconnections

Confidential

- Why is so much flexibility needed?
  - Time multiplexing of hardware resources
  - Application Specific HW accelerator is useless if applications are not known beforehand
  - New applications emerge and terminal has to be able to run them
  - HW is tailored to each application that runs: efficiency
  - Scalable applications to change QoS to the needs of the moment

Confidential

- Main goal:
  - To demonstrate the feasibility and the benefits of using reconfigurable hardware to deliver scalable quality of service to multimedia handheld devices

- Requirements
  - Content provider must know the expected performance of the HW (negotiation protocol)
  - Accurate management of the resources (middleware layer)
  - Introduction of scalability of application and multimedia content
  - Reconfigurability of the underlying hardware (with performance estimation)

Confidential

- Success of the project depends on the integration of the different research aspects
  - WP1: Definition of the architecture/interfaces
    - Spec. of reconfigurable HW
    - Spec. API platform/MW
    - Spec. API application/MW
    - Terminal description
    - Profiling video codecs
  - WP2: Negotiation
    - Content description framework
    - Negotiation protocol definition
  - WP3: Design and management of reconfigurable applications
    - Design scalable video codecs
    - Development MW services
    - Management FPGA reconfiguration
  - WP4: Demonstrator

Confidential

- Specification and evaluation of reconfigurable HW
  - Enable execution of scalable video codecs on the HW, therefore
  - Study the performance requirements of the FPGA hardware
    - From the application's footprint
    - Performance, memory, throughput, etc.
  - Evaluation of the benefits of reconfigurability
    - Trade-off between improved flexibility and overhead of reconfiguration / performance loss
    - Use performance estimates for evaluating different FPGA implementaqtions
  - Study the required amount of reconfigurability
    - For different tasks
    - For downscaling/upscaling performance (negotiation)
    - Partial reconfigurability

Confidential

- Reconfigurable HW parameterisation and evaluation
  - Goal: set of configuration bit strings, together with parameters indicating performance capabilities (used for task allocation)
  - Hardware parameterisation
  - Study relations between configurations for scaling of tasks
  - Study Pareto-optimal configurations

Confidential

- Management of FPGA reconfigurations
  - What is the best way to reconfigure the FPGAs?
  - Reconfiguration by choosing a pre-computed bit stream
    - Limits use of FPGA to pre-known set of applications (better than 1!)
  - Partial reconfiguration
  - Also study possibility to include FPGA configuration software on the processor
    - Not for run-time reconfiguration but for pre-reconfiguration for a new task
    - Should only be done once for each additional task (added to database)

Confidential

- The RESUME project: motivation and goals
- Scalable video characteristics
- Software transformations
- Hardware implementations of basic components
- QTL and hardware: the conflicts
- A new Wavelet Entropy Codec
- Results
- Conclusions and future work

Confidential

Quality ~ deployed hardware resources

Confidential

Confidential

- Videostream = sequence of frames
- Split and process per GOP (*Group Of Pictures*) of 16 frames

Confidential

=> vectors + error frame

Confidential

Motion compensation

Scalability in time

time

Confidential

Overview video codec

Confidential

Scalability in resolution

Confidential

Overview video codec

Confidential

Reference frame

Wavelet Transform

Entropy Encoding

Scalability in quality

Confidential

Tree is encoded in a symbol stream (0 of 1) and compressed with an arithmetic encoder

Confidential

Quadtree limited algorithm

- Quadtree does not descend to pixel level, but stops at a certain blocksize
- All significant pixels in a signifance list (linked list)

Confidential

Quadtree limited algorithm

- Pixels that are already significant (from higher bitplanes) are refined
- Information from surrounding pixels in higher bit planes is used to encode lower bit planes

Confidential

Quadtree limited algorithm

Confidential

Quadtree limited algorithm

Confidential

Quadtree limited algorithm

Confidential

Quadtree limited algorithm

Confidential

- The RESUME project: motivation and goals
- Scalable video characteristics
- Software transformations
- Hardware implementations of basic components
- QTL and hardware: the conflicts
- A new Wavelet Entropy Codec
- Results
- Conclusions and future work

Confidential

How to measure quality?

- 2 frames: PSNR (Peak Signal Noise Ratio):

$$10\log_{10}\left(\frac{255^2.\frac{3}{2}.\text{rows.cols}}{\sum(Y_0-Y_1)^2+\sum(U_0-U_1)^2+\sum(V_0-V_1)^2}\right)$$

(0: original frame; 1: decoded frame)

- 2 sequences: mean PSNR

But what if sequences are scaled differently?

e.g. Different frame rate

- Scale down original rate
  ➡ Reduced Frame Rate PSNR (RFR)
- Scale up decoded rate
  ➡ Full Frame Rate PSNR (FFR)

Confidential

- Scale down original rate
  ➡ Reduced Frame Rate PSNR (RFR)
- Scale up decoded rate
  ➡ Full Frame Rate PSNR (FFR)

1 GOP (Group Of Pictures) = 16 frames

Confidential

Two ways for PSNR

Confidential

Two ways for PSNR

Confidential

External requirements

- We want to be able to fully (up to lossless) decode video:
  - Frame rate: 30 Hz
  - Frame size: QCIF (176 x 144) and possibly CIF (352 x 288)
  - Corresponds to video bandwidth of up to 1.09 MiB/s for QCIF and 4.35 MiB/s for CIF
  - Seamless support for higher performance once/if better hardware is available
- Cost
  - **Memory**: maximum amount of memory required at one time
  - **Computations**: number of operations per second
  - **Bandwidth**: communication between components

Confidential

Relative execution time

Absolute execution time

- Motion compensation
- Inverse wavelet transform
- Entropy decoding

On a Pentium IV 2.0 GHz

Confidential

- Generic:
  - Floating point → integer operations:
    no loss in quality
  - More efficient memory use
- Quadtree:
  - Better datatype choice
  - Code transformations towards a hardware implementation
- Wavelet:
  - Remove irrelevant calculations
  - Use symmetry
- Motion Compensation:
  - Completely rewritten towards a hardware implementation
  - Scalable in time and resolution
  - Compensation of CIF-frames real time (Pentium IV 2 GHz)

Confidential

Memory requirements

QCIF: 0.5 MiB, CIF: 2.11 MiB: not a problem

Confidential

Computational requirements

- AD: very small
- QT: small but probably memory bound
- IWT: 261.12 10^6 additions/s and 274.03 10^6 multiplications/s for CIF
- MC: 6.08 10^6 additions/s for CIF
- Other components are negligible

Confidential

Bandwidth to the component

QCIF: 19.46 MiB/s and CIF 77.81: MiB/s

- Bandwidth reorder buffer largest
- Worst case internal bandwidth of QT very bad: 426.36 MiB/s

Confidential

- The RESUME project: motivation and goals
- Scalable video characteristics
- Software transformations
- Hardware implementations of basic components
- QTL and hardware: the conflicts
- A new Wavelet Entropy Codec
- Results
- Conclusions and future work

Confidential

Our board

Confidential

Altera EP1S25

- 25660 Logic Elements
- Memory hierarchy: (1.944.576 bits)
  - 224 M512 RAM blocks (512 bit + parity)
  - 138 M4K RAM blocks (128 x 36 bits)
  - 2 MegaRAM blocks (4Ki x 144 bits)
- 10 DSP blocks (4 18bit multipliers each)
- 6 PLLs

Confidential

Modular design

33 MHz

133 MHz

PCI bus

pci_mt64 MegaCore

PCI Control logic

registers

66 MHz

IDWT

SDRAM Controller

SDRAM

Confidential

- Independent access to PCI and RAM
- New modules can be added
- Communication with PC is memory-mapped (via two Linux drivers)
- Simple to change from a hardware to a software implementation
- Interrupts, DMA

Confidential

Overview video codec

Confidential

- Floating point → integer operations: no quality loss
- Scalable in time and resolution

  Quality of lower resolution decoded sequences strongly depends on quality of Motion Estimation
- Sub pixel interpolation with shorter filters: less computations and memory accesses (but also reduced quality)
- Better than real time on Pentium 2 GHz

Confidential

- Used to test
  - modular design
  - design flow (SystemC → VHDL → FPGA)
- First implementation
  - short implementation time: not optimal (row/col-wise)

    (we are planning a line-based implementation)

Confidential

- Used to test
  - modular design
  - design flow (SystemC → VHDL → FPGA)
- First implementation
  - short implementation time: not optimal (row/col-wise)

    (we are planning a line-based implementation)

Confidential

Inverse wavelet transform

Confidential

Inverse wavelet transform

- Decoupled I/O and calculations (in parallel at different speeds with different clocks)

- Calculations (3 levels) max f=75MHz
  - CIF: 850000 cycl. (17 ms @ 50 MHz)
  - QCIF: 235000 cycl. (4.7 ms @ 50 MHz)
- FPGA usage
  - IDWT: 2953 LE (11%), 55kb (2%), 12 DSP block 9-bit elements (15%)
  - modular design + IDWT: 7800 LE (30%)
- Currently I/O-bound

Confidential

- The RESUME project: motivation and goals
- Scalable video characteristics
- Software transformations
- Hardware implementations of basic components
- QTL and hardware: the conflicts
- A new Wavelet Entropy Codec
- Results
- Conclusions and future work

Confidential

Reference frame

Wavelet Transform

Entropy Encoding

Confidential

Quadtree limited algorithm

Confidential

- Tree encoded in a symbol stream (consisting of 0 and 1) and then compressed using an arithmetic coder
- CIF: 31×10^6 symbols/second (mobile)
- QCIF: 6×10^6 symbols/second (foreman)

Confidential

Quadtree Algorithm

- Good compression
- Most quality enhancing bits come first
- But very complex:
  - Arithmetic coder with division
  - Linked lists
  - Very irregular memory access pattern
  - Recursive Quadtree structure
  - Large memory footprint

Confidential

A typical Altera Stratix (EP1S25) FPGA:

Name	Number	Size (byte + parity)
MRAM	2	64k
M4K	138	512
M512	224	64
Registers	25660	1/8
		214kiB

Quadtree:

- Current bitplane (352x288): 101376bit
- Quadtree Structure: ~540 bit
- Significance lists : 101376 17bit entries (~198 kiB), probably streamable

Model Selection:

- Maps for model selection: 4 x ~12.5 kiB (more than 200 kiB in software implementation)

Confidential

- The RESUME project: motivation and goals
- Scalable video characteristics
- Software transformations
- Hardware implementations of basic components
- QTL and hardware: the conflicts
- A new Wavelet Entropy Codec
- Results
- Conclusions and future work

Confidential

Consider a simpler variant

Wish list:

- Scalability:
  - Quality → bit layers
  - Resolution → wavelet bands
- Acceptable compression
- Economical with memory
- Sweet and simple
- High degree of parallelism

Confidential

A new Wavelet Entropy Codec!

Wavelet Transform

Entropy Encoding

Confidential

Confidential

Bitplane

Significance bitmap

Sign bitmap

Confidential

Bitvalue + model → (CABAC-based) encoder

Confidential

454 different models

This high number is not a problem:

each model needs only a 9 bit state

Models are initialised (warmed up) with values obtained from training a number of representative sequences

Confidential

- The RESUME project: motivation and goals
- Scalable video characteristics
- Software transformations
- Hardware implementations of basic components
- QTL and hardware: the conflicts
- A new Wavelet Entropy Codec
- Results
- Conclusions and future work

Confidential

Lower is better (smaller)

Confidential

Results

Canoa (CIF)

Foreman (CIF)

Mobile (QCIF)

Stefan (QCIF)

Mobile CIF

Foreman CIF

Confidential

- Algorithm is simpler
- No recursive behaviour
- Small memory footprint

  [2 bitmaps of size(resolutionlayer)]
- Memory access is more regular
- Arithmetic coder must process more symbols, but is simpler and faster (modified arithmetic encoder of CABAC)
- Compression is better for all bitrates

Confidential

Name	Number	Size (byte + parity)
MRAM	2	64k
M4K	138	512
M512	224	64
Registers	25660	1/8
		214kiB

- Arithmetic Decoder:
  - State Lookup table: 256B
  - Range Lookup table: 2048B
  - Model state table: 545B
  - Buffer

  ~7 M4K block

- Model Selector
  - Significance and sign bitmap: 2 x 9504B
  - Buffers
  - (1 MRAM block is large enough to contain 6 significance and sign bitmaps)

  ~47 M4K blocks

Confidential

- The RESUME project: motivation and goals
- Scalable video characteristics
- Software transformations
- Hardware implementations of basic components
- QTL and hardware: the conflicts
- A new Wavelet Entropy Codec
- Results
- Conclusions and future work

Confidential

- Scalable video codecs have benefits but are complex, real-time performance is tough goal
- Reconfigurable implementation is the only way
- How to compare quality for scalable video?
- Hardware implementation is bandwidth-limited
- Some things come for free (simpler can be better)
- In search of implementation methodology

Confidential

- Line based FPGA implementation of IDWT
- Use rate distortion for efficient rate allocation
- Fine tune the CABAC arithmetic coder and WEC
- Implementation in FPGA
- Learn from implementation results for a priori estimation/exploration
- Investigate reconfigurability issues

Confidential