GPU Acceleration of the HEVC Decoder Inter Prediction Module

technology
from seed

Diego F. de Souza, Aleksandar Ilic, Nuno Roma and Leonel Sousa

INESC-ID, IST, Universidade de Lisboa

Lisboa – Portugal

3rd IEEE Global Conference on Signal & Information Processing (GlobalSIP)

Orlando, Florida, USA, December 14-16, 2015

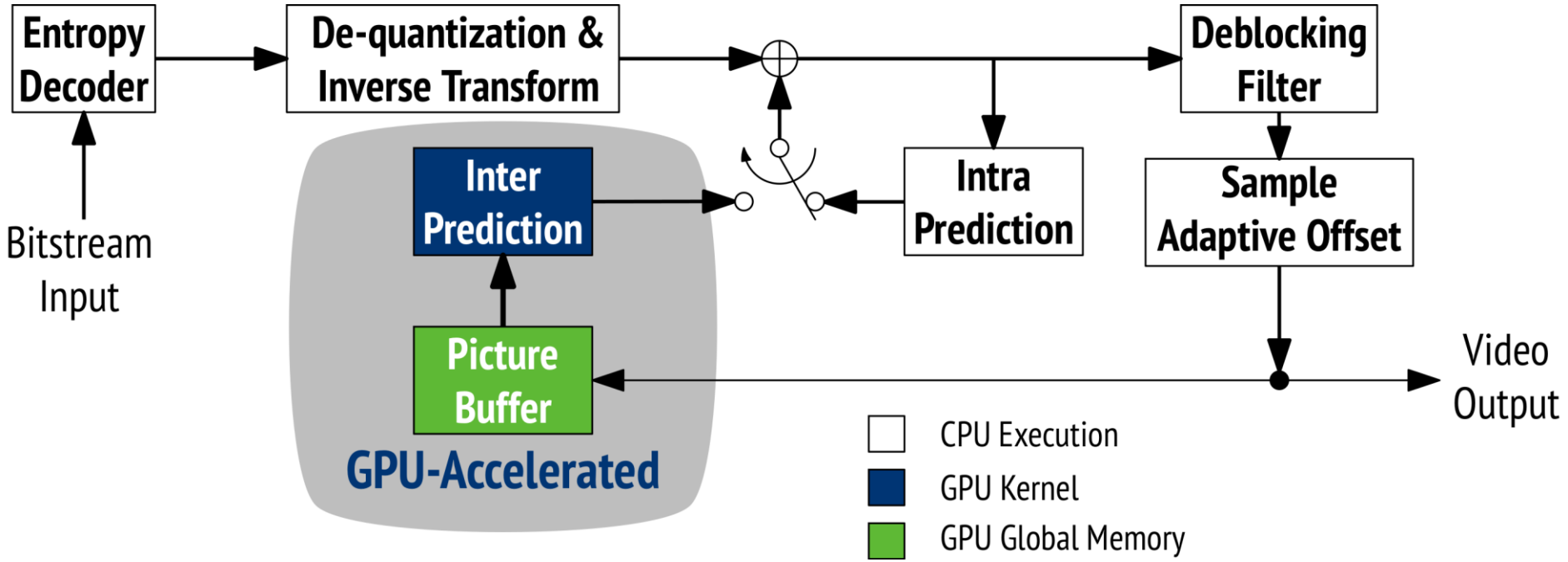
Outline

- HEVC Decoder
  - Inter Prediction
    - Partitioning Structure
    - Interpolation
- Proposed Parallel Inter Prediction Algorithm
  - GPU Thread Assignment
  - Packed Motion Data
  - Framework
- Experimental Evaluation
- Conclusions

Outline

- HEVC Decoder
  - Inter Prediction
    - Partitioning Structure
    - Interpolation
- Proposed Parallel Inter Prediction Algorithm
  - GPU Thread Assignment
  - Packed Motion Data
  - Framework
- Experimental Evaluation
- Conclusions

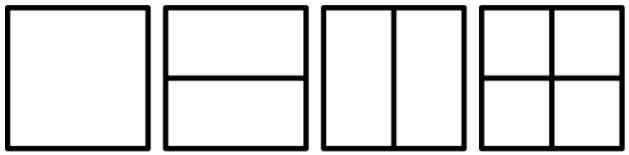
HEVC Decoder: Simplified Block Diagram



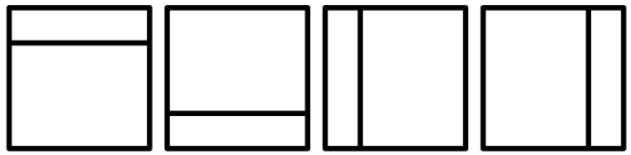
- The **Inter Prediction module** of the HEVC decoder is responsible for **43-49% of the total decoding time** in both **ARM** and **x86** instruction set architectures

Inter Prediction: Partitioning Structure

Symmetric Partitioning



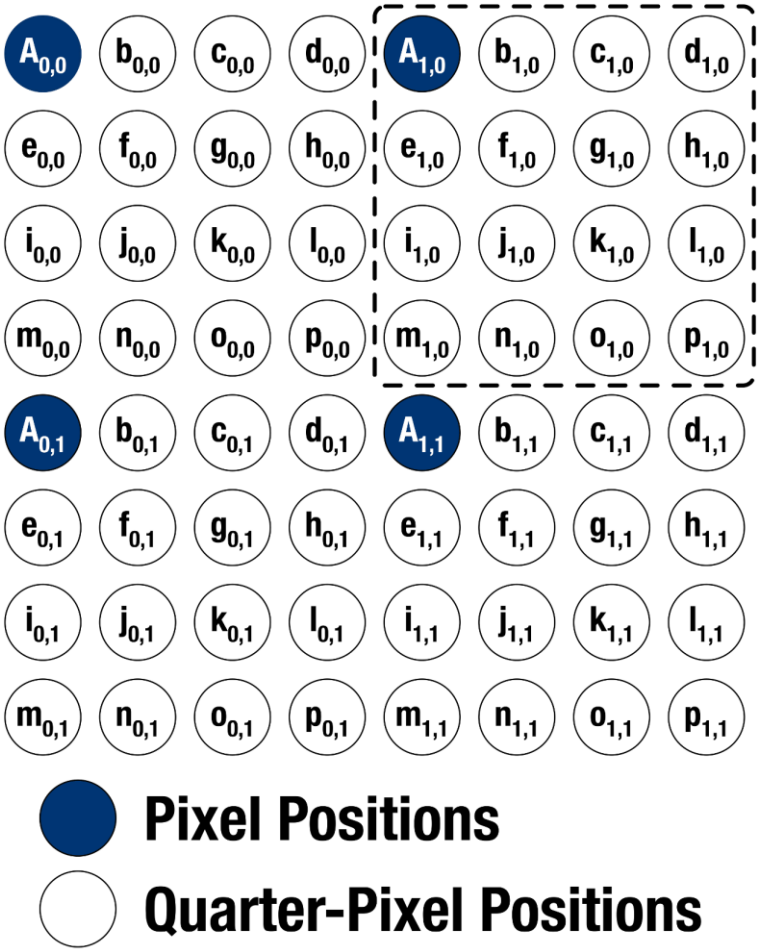
Asymmetric Partitioning



- The **symmetric partitioning** is restricted to the **quadtree structure**, where a PU is split in up to four blocks

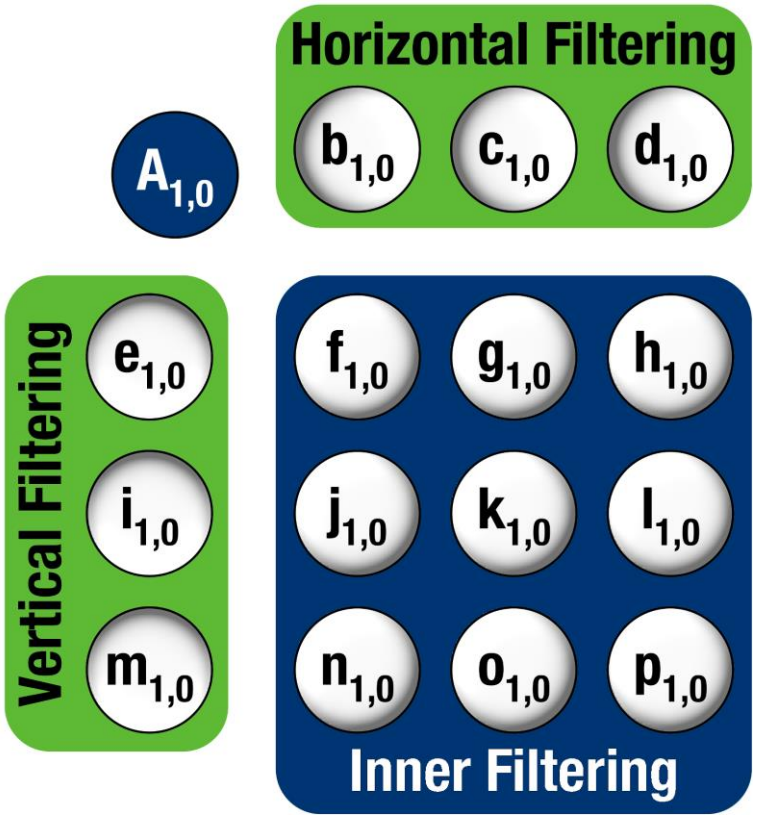
- The **HEVC** standard also **introduced asymmetric partition** modes for Inter prediction, which allow more **accurate predictions** and offer up **to 2.8% of bit-rate reduction**

Inter Prediction: Interpolation



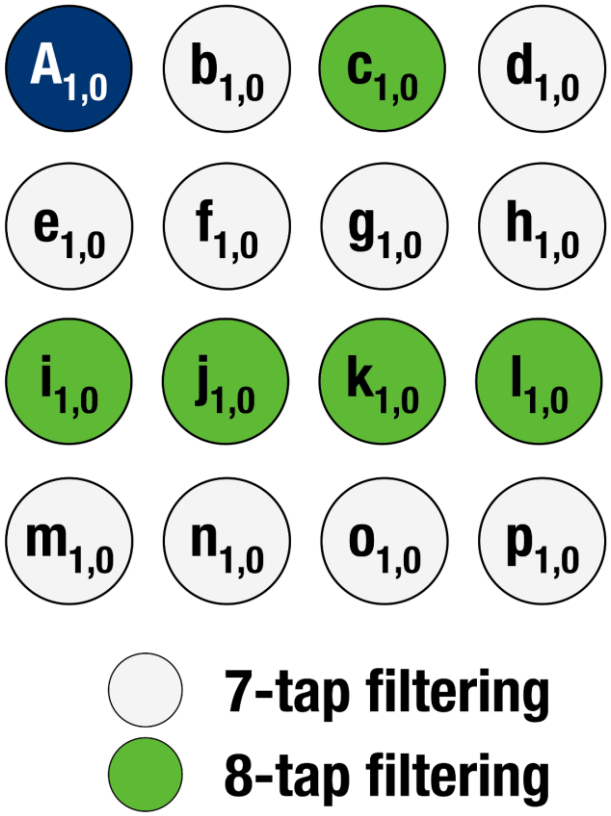
- The **HEVC** standard allows **motion vectors** at luma **quarter-pixel** resolution
- The **pixel** samples ($A_{x,y}$) are **directly obtained** from the **reference frame**

Inter Prediction: Interpolation



- ***Horizontal Filtering***: from the **pixels** from the **same row**
- ***Vertical Filtering***: from the **pixels** in the **same column**
- ***Inner Filtering***: performing the **vertical filtering** on the **sub-samples** from the **same column**

Inter Prediction: Interpolation

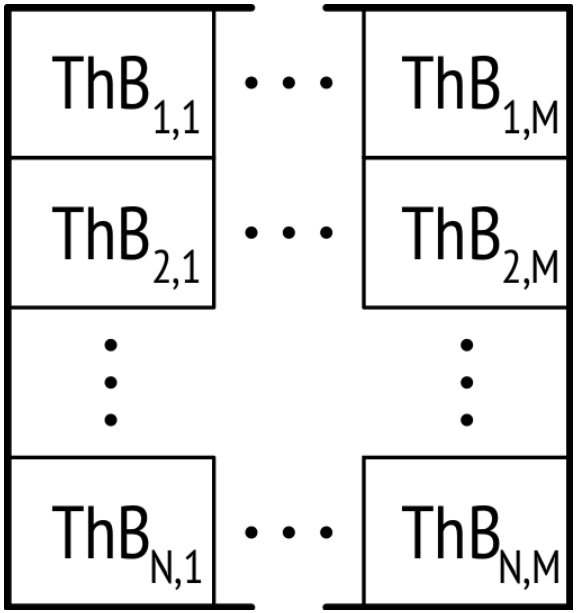


- **8-tap** and **7-tap filters** are adopted, according to each **sub-pixel position**
- **7-tap filtering** is applied to create the **sub-pixel** samples that are **close to the pixels**
- In the **chroma interpolation**, only **4-tap filters** are used

Outline

- HEVC Decoder
  - Inter Prediction
    - Partitioning Structure
    - Interpolation
- Proposed Parallel Inter Prediction Algorithm
  - GPU Thread Assignment
  - Packed Motion Data
  - Framework
- Experimental Evaluation
- Conclusions

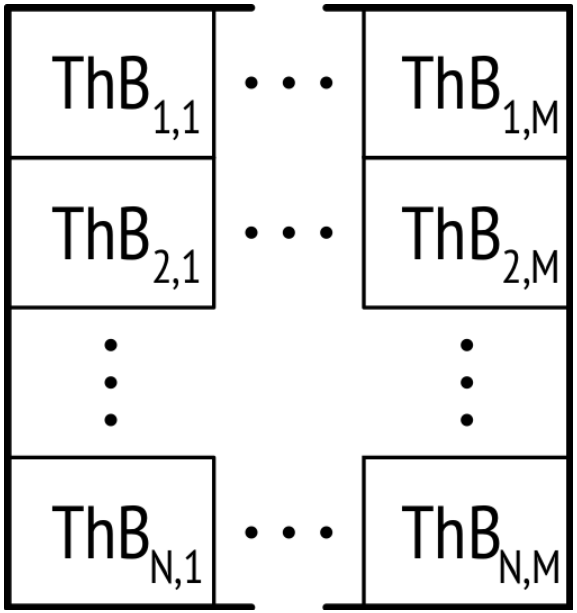
Proposed Inter Prediction Decoding Parallelization

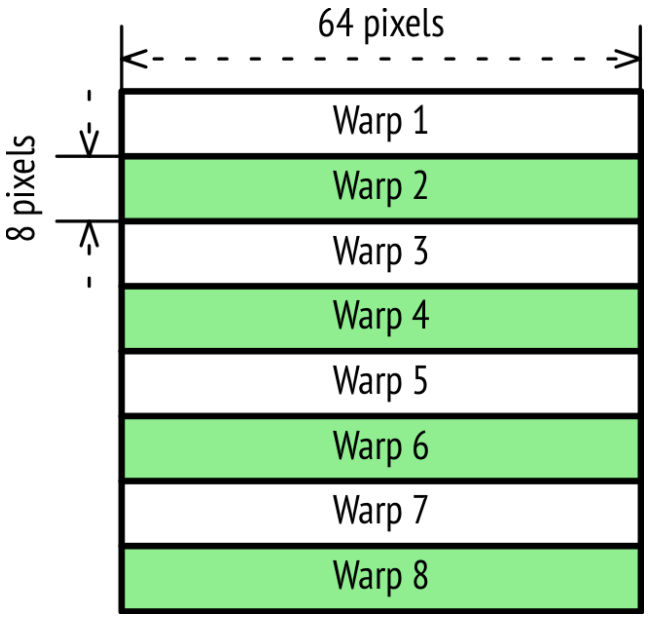


- **Each Thread Block** (ThB) performs the motion compensation for **a 64×64 CTU block**

- All Thread Blocks **are independent of each other**

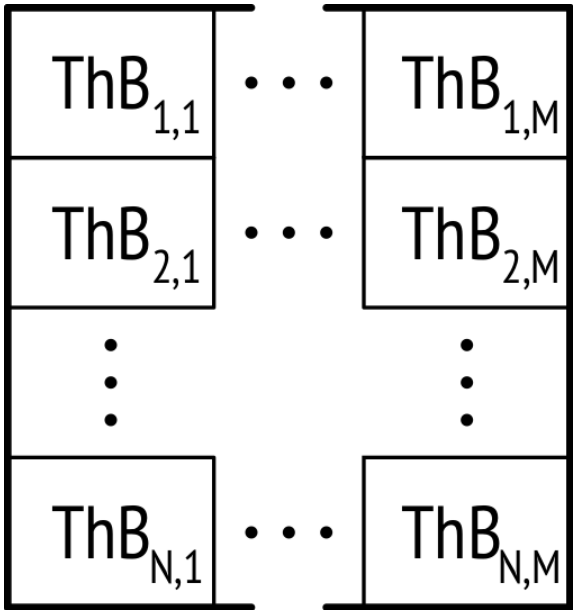
Proposed Inter Prediction Decoding Parallelization

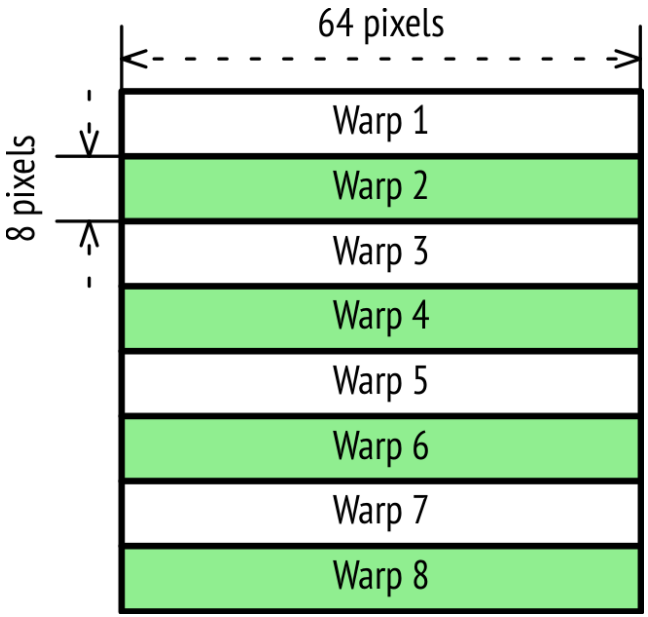


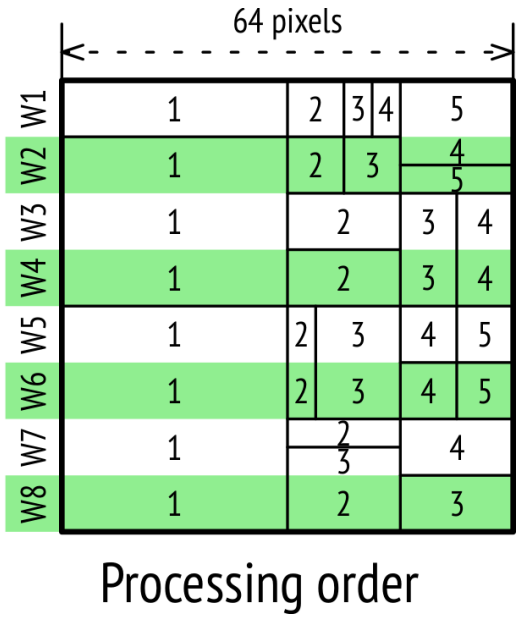


- Each **ThB** contains **8 warps**
- **Each warp** performs the motion compensation of **all sub-blocks** in an **eight-pixel row**
- All warps are **independent of each other**

Proposed Inter Prediction Decoding Parallelization



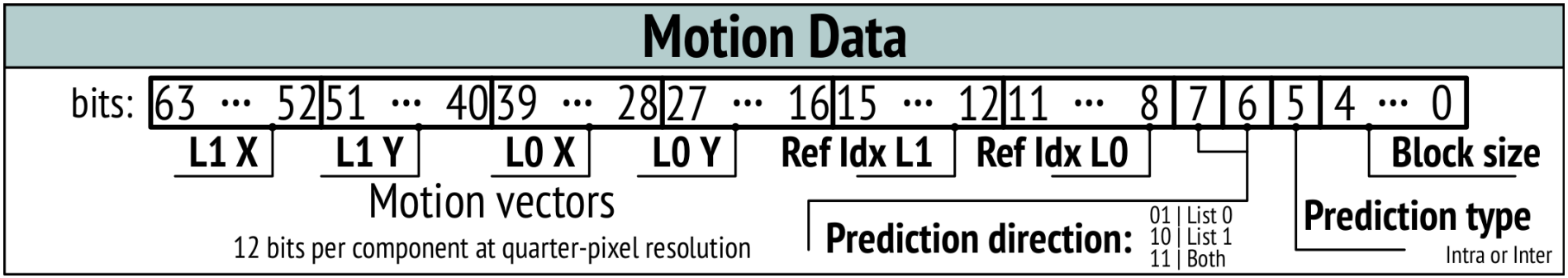




- **Each warp (Wx)** performs the motion compensation of **one sub-block per time**

- Each sub-block is **8×N** or **4×N**

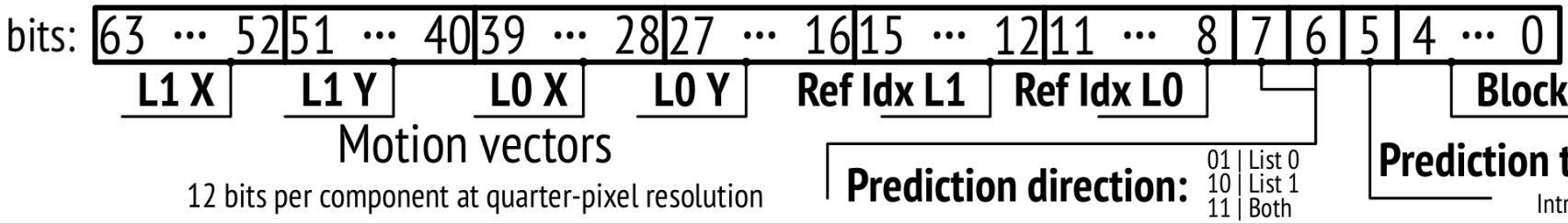
Proposed Inter Prediction Decoding Parallelization



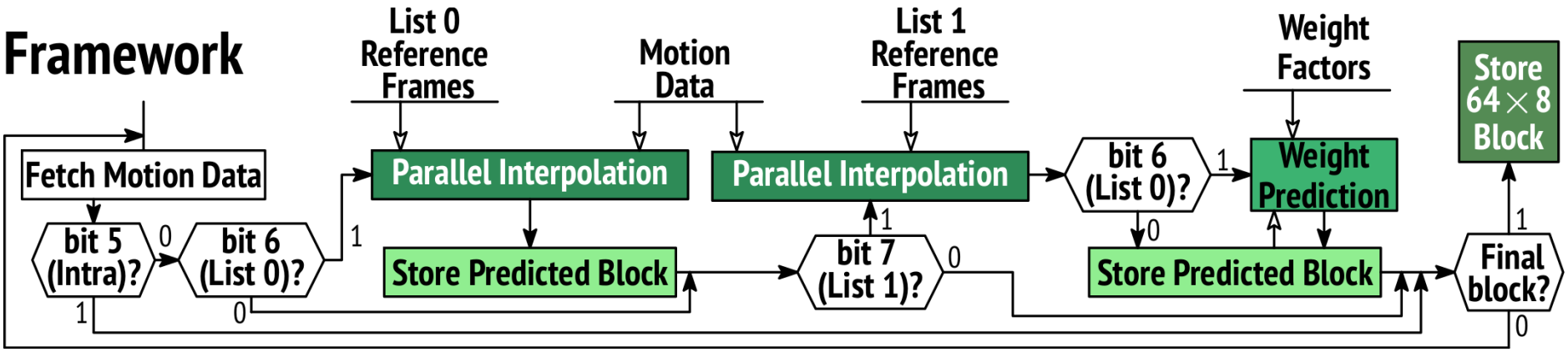
- All the **required motion data** is stored in a **64-bit word**
- For **each 8×8** block, **two 64-bit** words are needed to process asymmetric blocks like 16×4

Proposed Inter Prediction Decoding Parallelization

Motion Data



Framework



Outline

- HEVC Decoder
  - Inter Prediction
    - Partitioning Structure
    - Interpolation
- Proposed Parallel Inter Prediction Algorithm
  - GPU Thread Assignment
  - Packed Motion Data
  - Framework
- Experimental Evaluation
- Conclusions

Average Frame Processing Time in milliseconds

Class	Sequence	QP	Random Access		Low Delay B	
			HM 15.0	G980	HM 15.0	G980
S (3840×2160) Ultra HD 4K	CrowdRun	22	139.57	17.69	140.77	19.49
		27	115.65	16.10	116.37	17.96
		32	103.66	15.32	98.77	16.80
		37	95.20	14.70	85.67	16.14

A (2560×1600) WQXGA

B (1920×1080) Full HD

- **JCT-VC** common **test conditions** and configurations:
  - **QP: 22, 27, 32** and **37**
  - ***Random Access*** and ***Low Delay B*** configuration
  - HEVC **main profile**
  - Only results of **frames** with **less than 15%** of **Intra blocks**
- **Sequences** – frame resolutions
  - Class S - 3840×2160 – **Ultra HD 4K**
  - Class A - 2560×1600 – **WQXGA**
  - Class B - 1920×1080 – **Full HD**

Average Frame Processing Time in milliseconds

Class	Sequence	QP	Random Access		Low Delay B	
			HM 15.0	G980	HM 15.0	G980
S (3840×2160) Ultra HD 4K	CrowdRun	22	**139.57**	17.69	**140.77**	19.49
		27	**115.65**	16.10	**116.37**	17.96
		32	**103.66**	15.32	**98.77**	16.80
		37	**95.20**	14.70	**85.67**	16.14
A (2560×1600) WQXGA						
B (1920×1080) Full HD						

- **Baseline (HM 15.0)**
  - High Efficiency Video Coding Test Model version 15.0
  - Intel® Core™ i7-5960X @ 3.0GHz

Average Frame Processing Time in milliseconds

Class	Sequence	QP	Random Access		Low Delay B	
			HM 15.0	G980	HM 15.0	G980
S (3840×2160) Ultra HD 4K	CrowdRun	22	139.57	**17.69**	140.77	**19.49**
		27	115.65	**16.10**	116.37	**17.96**
		32	103.66	**15.32**	98.77	**16.80**
		37	95.20	**14.70**	85.67	**16.14**
A (2560×1600) WQXGA						
B (1920×1080) Full HD						

- **Baseline**
  - High Efficiency Video Coding Test Model version 15.0 (**HM 15.0**)
  - Intel® Core™ i7-5960X @ 3.0GHz
- **Proposed GPU Inter Prediction**
  - NVIDIA CUDA version 7.0
  - NVIDIA GeForce GTX 980 @ 1126 MHz (**G980**)

Average Frame Processing Time in milliseconds

Class	Sequence	QP	Random Access		Low Delay B	
			HM 15.0	G980	HM 15.0	G980
S (3840×2160) Ultra HD 4K	CrowdRun	22	139.57	17.69	140.77	19.49
		27	115.65	16.10	116.37	17.96
		32	103.66	15.32	98.77	16.80
		37	95.20	14.70	85.67	16.14
A (2560×1600) WQXGA						
B (1920×1080) Full HD						

- **Time increases** with the **decrease** of the **QP**
  - HM 15.0 (**CPU**): **Up to 39%** from the lowest to the highest QP
  - G980 (**GPU**): **Up to 17%** from the lowest to the highest QP
  - The **increase rate is lower** in the **GPU** due to the obtained **parallelism**

Average Frame Processing Time in milliseconds

Class	Sequence	QP	Random Access		Low Delay B	
			HM 15.0	G980	HM 15.0	G980
S (3840×2160) Ultra HD 4K	CrowdRun	22	139.57	17.69	140.77	19.49
		27	115.65	16.10	116.37	17.96
		32	103.66	15.32	98.77	16.80
		37	95.20	14.70	85.67	16.14
	InToTree	22	133.69	17.75	137.20	19.74
	ParkJoy	22	140.39	17.60	152.51	20.39
A (2560×1600) WQXGA	Traffic	22	53.07	7.31	55.14	8.37
	PeopleOnStreet	22	60.32	8.26	60.70	8.89
	NebutaFestival	22	55.61	8.64	57.04	9.01
	SteamLocomotiveTrain	22	44.32	7.06	50.08	8.13
B (1920×1080) Full HD	Kimono1	22	27.97	4.00	28.78	4.72
	ParkScene	22	31.72	4.14	34.64	4.95
	Cactus	22	21.98	3.47	22.63	3.98
	BQTerrace	22	34.65	4.67	39.21	5.22
	BasketballDrive	22	27.44	4.23	28.54	4.65

Average Frame Processing Frame Rate (fps)

Class	Sequence	QP	Random Access		Low Delay B	
			HM 15.0	G980	HM 15.0	G980
S (3840×2160) Ultra HD 4K	CrowdRun	22	7.2	56.5	7.1	51.3
		27	8.6	62.1	8.6	55.7
		32	9.6	65.3	10.1	59.5
		37	10.5	68.0	11.7	62.0
	InToTree	22	7.5	56.3	7.3	50.7
	ParkJoy	22	7.1	56.8	6.6	49.0
A (2560×1600) WQXGA	Traffic	22	18.8	136.8	18.1	119.5
	PeopleOnStreet	22	16.6	121.1	16.5	112.5
	NebutaFestival	22	18.0	115.7	17.5	111.0
	SteamLocomotiveTrain	22	22.6	141.6	20.0	123.0
B (1920×1080) Full HD	Kimono1	22	35.8	250.0	34.7	211.9
	ParkScene	22	31.5	241.5	28.9	202.0
	Cactus	22	45.5	288.2	44.2	251.3
	BQTerrace	22	28.9	214.1	25.5	191.6
	BasketballDrive	22	36.4	236.4	35.0	215.1

Outline

- HEVC Decoder
  - Inter Prediction
    - Partitioning Structure
    - Interpolation
- Proposed Parallel Inter Prediction Algorithm
  - GPU Thread Assignment
  - Packed Motion Data
  - Framework
- Experimental Evaluation
- Conclusions

Conclusions

- **Efficient parallel algorithm is proposed** for the HEVC Inter Prediction module
  - **fully compliant with the HEVC standard**
  - efficiently **exploit GPU: computational capabilities** and **memory hierarchy**
- **Real-time** processing **achieved**
  - for all tested sequences (**Ultra HD 4K, WQXGA** and **Full HD**)
  - for the **most demanding setup (QP = 22)**
- In the **worst case scenario** (**QP=22**), the proposed GPU algorithm in **GeForce GTX 980** allows achieving (on average)
  - **Random Access: 56.6 fps** for class S, **128.8 fps** for class A and **246.1 fps** for class B
  - **Low Delay B: 50.3 fps** for class S, **116.5 fps** for class A and **214.4 fps** for class B

inesc id
lisboa
10 anos

Instituto de Engenharia de Sistemas e Computadores
Investigação e Desenvolvimento em Lisboa