A Fully-Integrated Energy-Efficient H.265/HEVC Decoder with eDRAM for Wearable Applications

Mehul Tikekar, Prof. Vivienne Sze,
Prof. Anantha Chandrakasan

Massachusetts Institute of Technology

Motivation for Fully-Integrated Video Decoder

- **50mW** power budget [1]
- Off-chip memory access power is **2.8x-6x** processing power [2,3]
- Need to reduce board footprint for wearables

[1] M. Aleksic, Qualcomm, VLSI 2017 Short Course
[2] C.-T. Huang, ISSCC 2013, [3] D. Zhou, ISSCC 2012

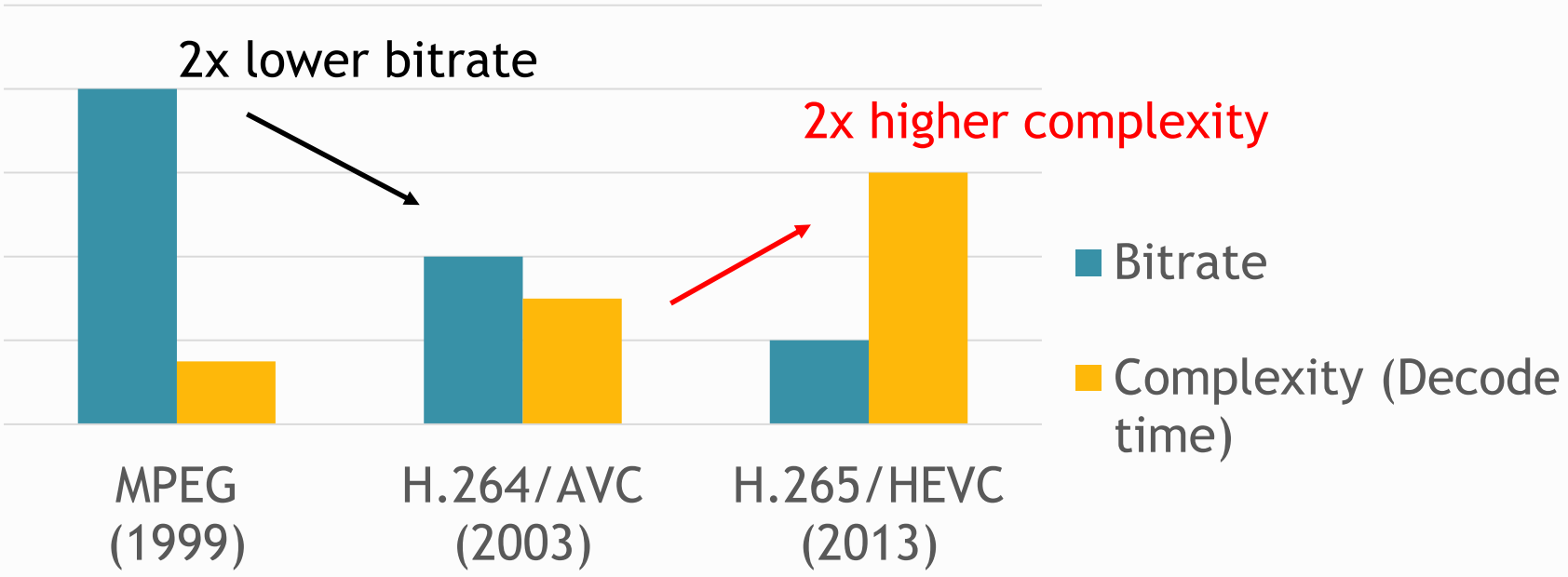
Previous Work

	ISSCC 2012	ISSCC 2013	A-SSCC 2013	ESSCRIC 2014	ISSSCC 2016
Standard	H.264/AVC MP/MVC	H.265/ HEVC WD4	H.265/ HEVC	H.265/HEVC, multistandard	H.265/ HEVC
Technology	65nm/1.2V	40nm/0.9V	90nm/1V	28nm/0.9V	40nm/1V
Max Throughput	7640x4320 @60fps	3840x2160 @30fps	1920x1080 @35fps	3840x2160 @60fps	7640x4320 @120fps
Frame buffer Storage	64b DDR3	32b DDR3	n/a	32b LPDDR3	64b DDR3
Core Power [mW]	410	76	36.9	104	690
Frame buffer Power [mW]	2520	219	n/a	n/a	n/a

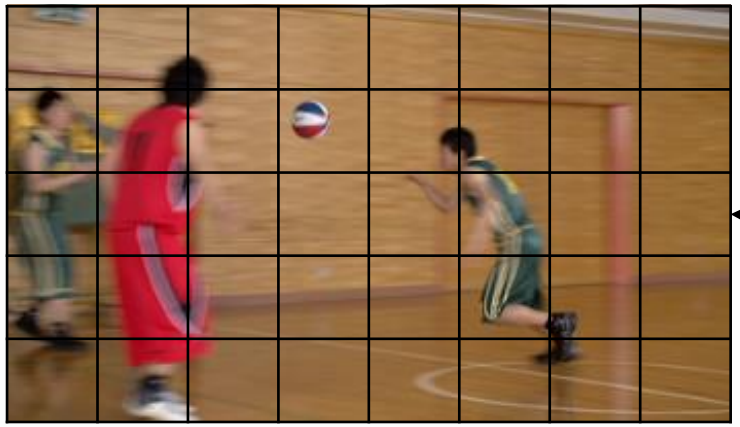
Difficult to meet **50mW** power budget for wearables with DRAM-based decoders

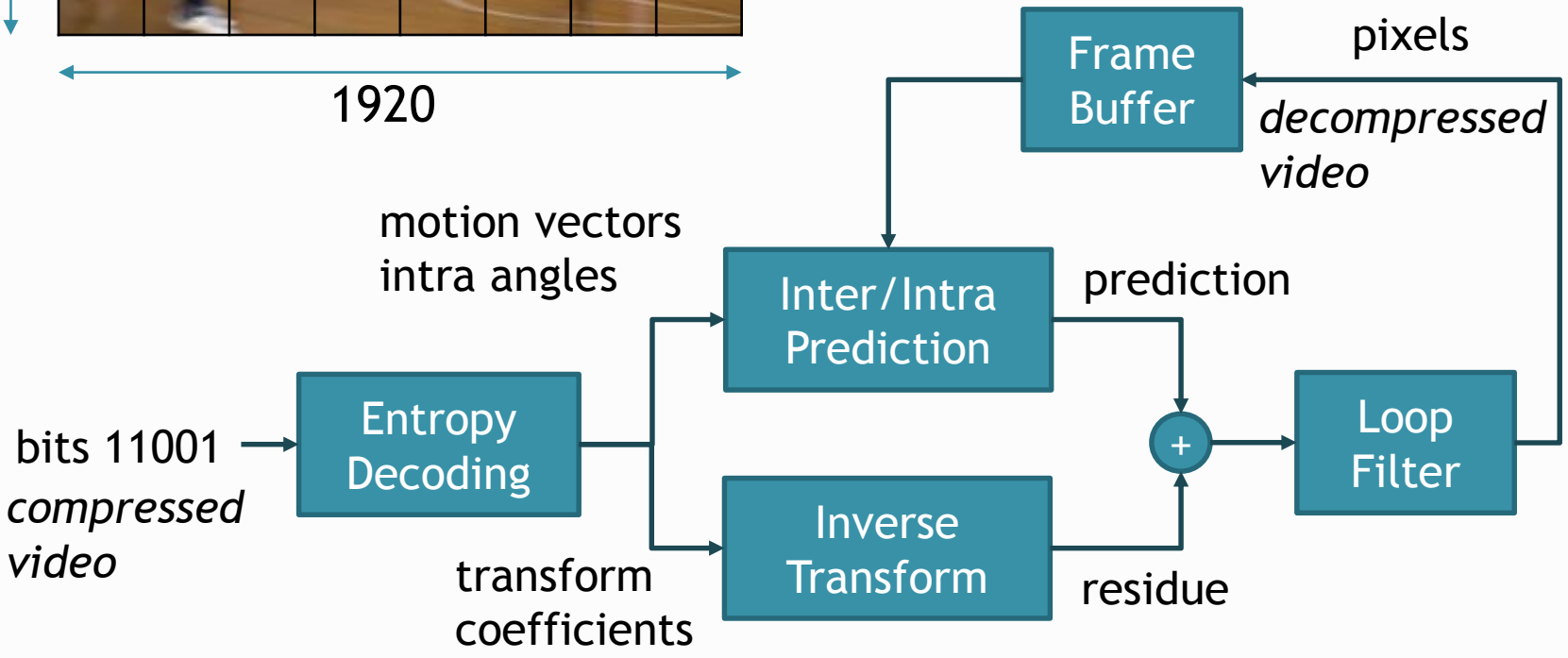
Video Coding Standard: H.265/HEVC

- High-Efficiency Video Coding (H.265/HEVC)
- 2x better compression vs. H.264/AVC
- System power savings from wireless RX
  - WiFi RX energy = 2x video decoding energy



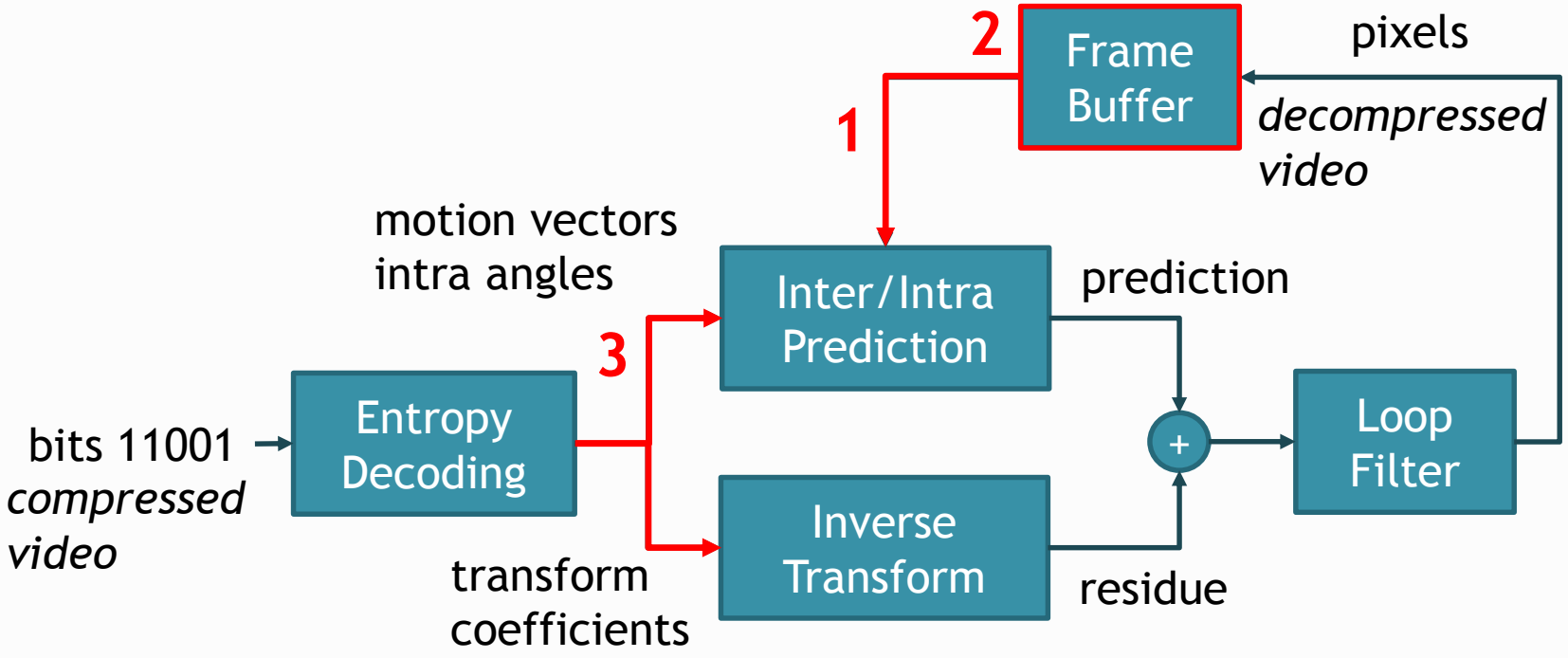
HEVC Decoder Pipeline





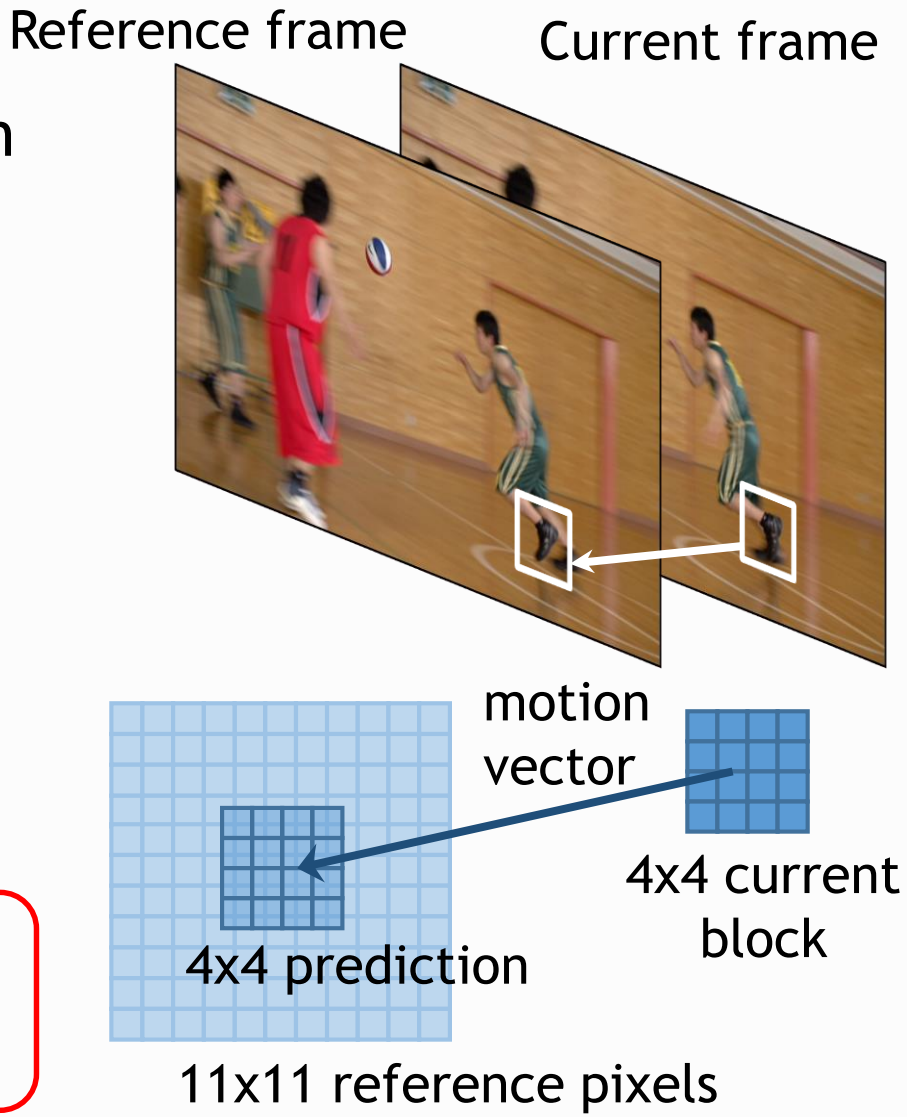
Focus of This Talk

1. Frame Buffer to Inter Prediction
2. On-demand Power-up of eDRAM
3. Data movement of Syntax Elements



Frame Buffer and Inter Prediction

- Inter-frame prediction provides most compression
- 50% processing time
- Dominates memory bandwidth requirements
  - 8-tap filter: 11x11 pixels read for 4x4 prediction
  - Prediction from 2 frames
- Frame buffer needs to store several older frames

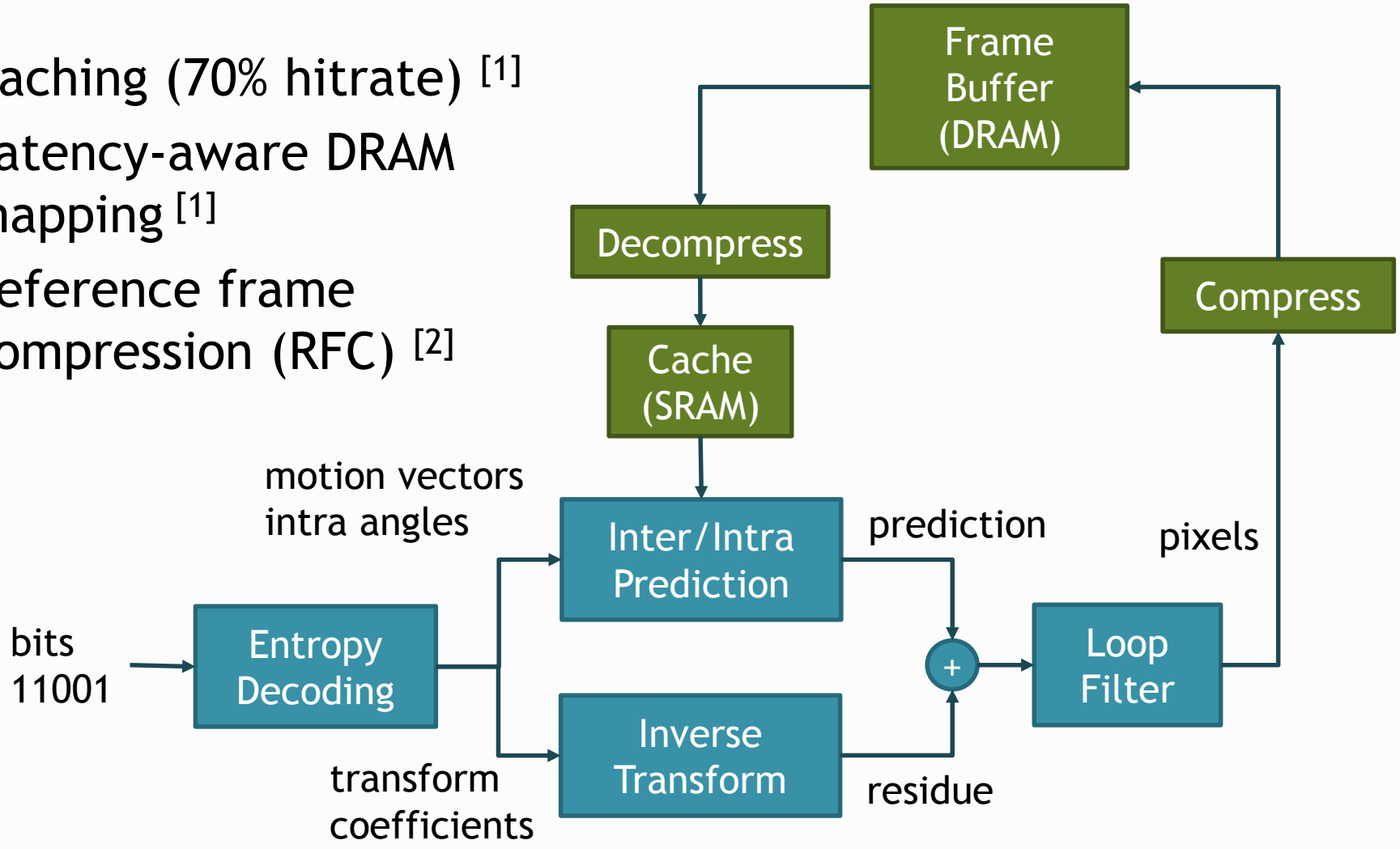


Frame buffer requirements

Size: 10 – 50 MB
Bandwidth: 0.5 – 1 GB/s

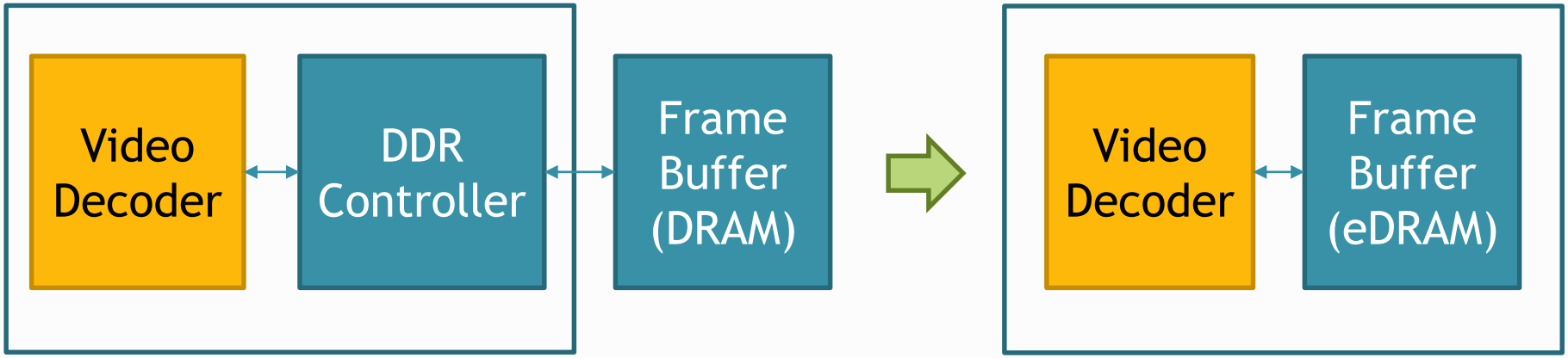
Memory Optimization Techniques

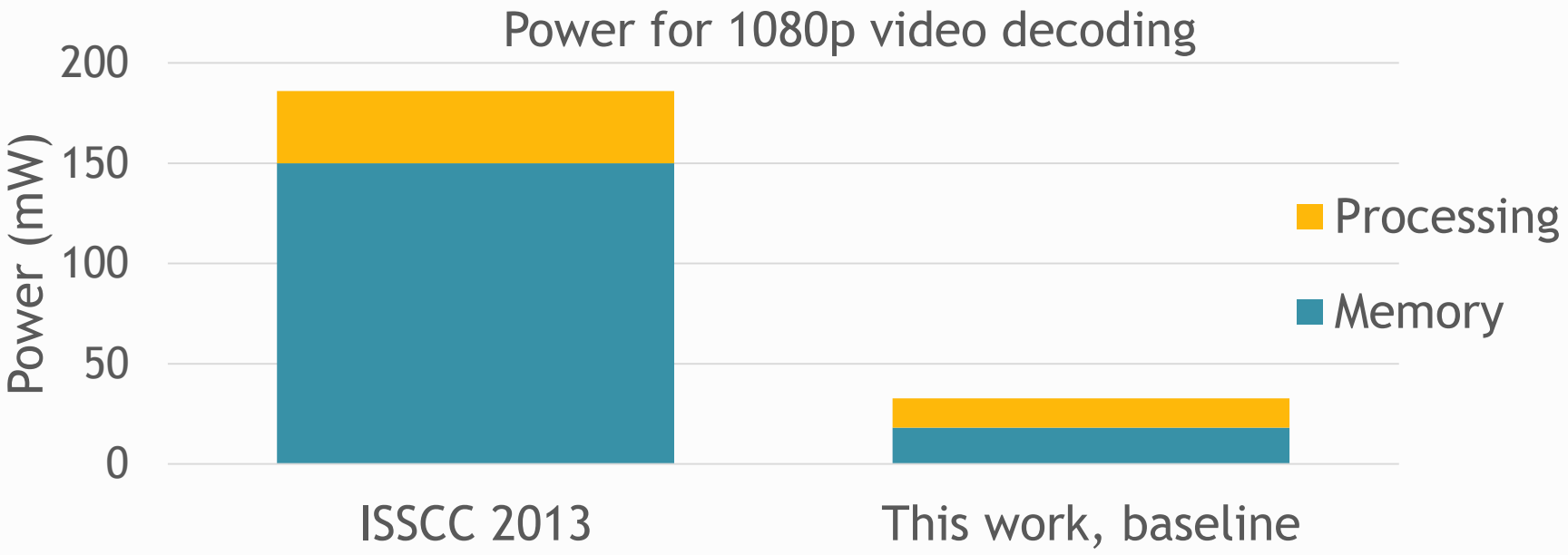
- Caching (70% hitrate) [1]
- Latency-aware DRAM mapping [1]
- Reference frame compression (RFC) [2]



[1] C.-T. Huang, ISSCC 2013 [2] Guo, TMM 2014

Motivation for Fully-Integrated Video Decoder





eDRAM vs. DRAM

Pros

- Lower energy/access
- Lower latency, higher bandwidth
- Smaller board footprint on wearable devices
- Smaller sized macros can be individually powered down

Cons

- Lower density
- More frequent refresh

In video decoder,
eDRAM refresh power = **4x read/write power**

eDRAM Operating Modes

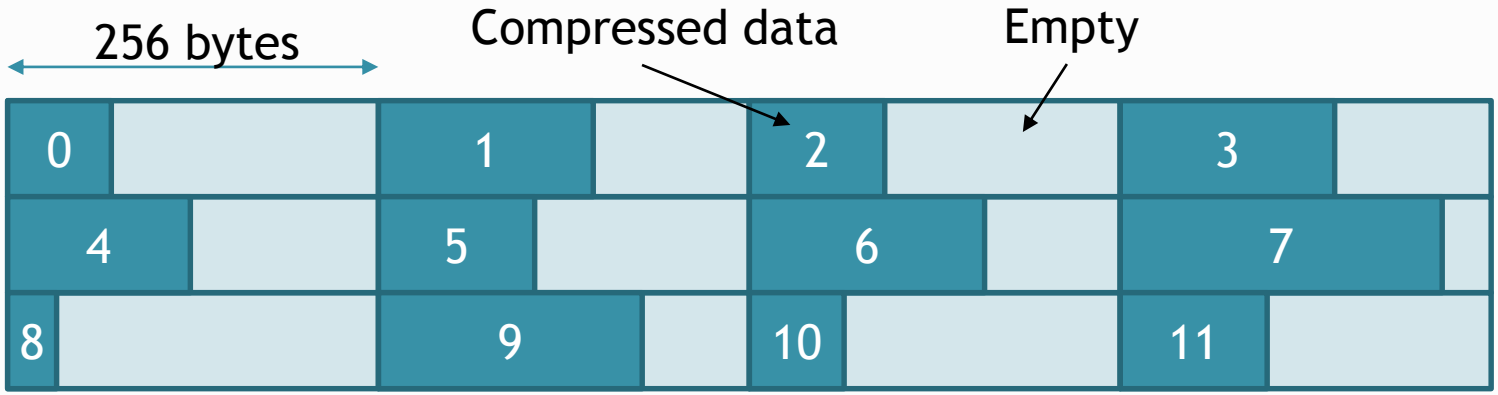
Active mode	Read, write, refresh
Self-refresh mode	Refresh
Deep power-down mode	(No data retention)

Lower power

Maximize use of Deep power-down mode
to reduce refresh power

RFC to Reduce eDRAM Refresh Power

- RFC techniques for DRAM use direct addressing
- For DRAM, bandwidth is more important than capacity

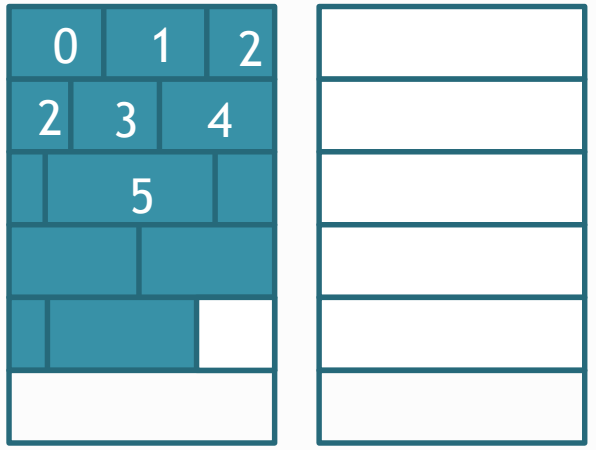


- Memory size and refresh power remain unchanged

Traditional RFC techniques do not reduce eDRAM refresh power

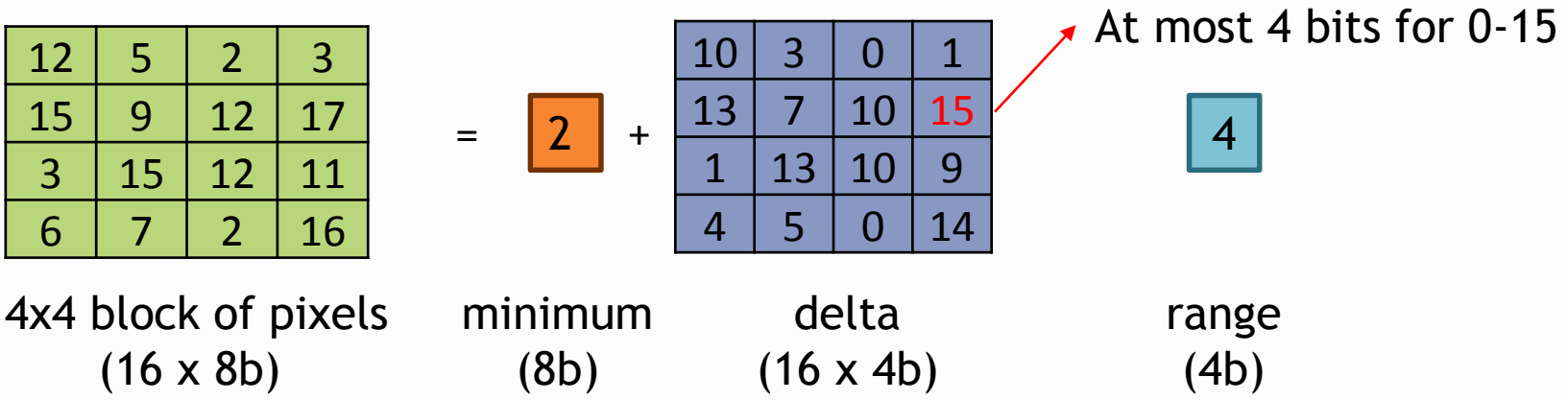
RFC for eDRAM with Indirect Addressing

- For eDRAM, reducing memory usage is more important than bandwidth
- Fully packed format: indirect addressing
- Address look-up memory is needed
- Exploits low latency and low energy/access cost of eDRAM



Proposed method exploits key benefits of eDRAM to reduce refresh power

Proposed RFC Scheme Example



No. of bits = 8 (minimum) + 4 (range) + 16*range

Compression achieved = 128/76 = 1.7x

range	range of deltas
0	0
1	0-1
2	0-3
…	
8	0-255 (compression off)

Average compression: 2x*

* Over HEVC Common Test Conditions (384 video sequences)

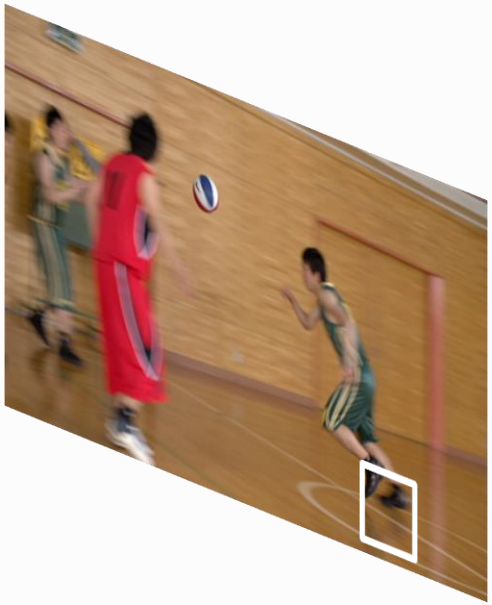
Comparison with Prior Work

	This work	Guo, TMM 2014
Compression method	Min-delta	Intra-prediction + DPCM + coding
Data saving	50%	60%
Area	8 kgate	80 kgate
Throughput	32 pixel/cycle	32 pixel/cycle

Lightweight compression method achieves good cost-performance tradeoff

Reading Pixels for Motion Compensation

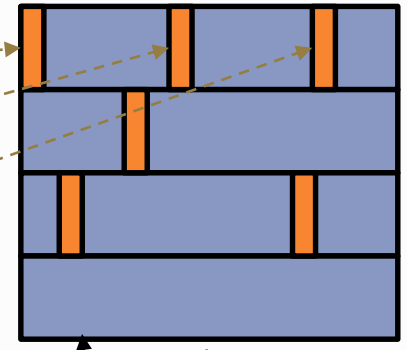
Reference frame



Address eDRAM

range 4-bit	Addr 22-bit
2	0
4	5
3	12
…	…

Data eDRAM



Position in Reference frame

4	5

2	10	3	0	1
	13	7	10	15
	1	13	10	9
	4	5	0	14

12	5	2	3
15	9	12	17
3	15	12	11
6	7	2	16

Address lookup adds 20% storage overhead

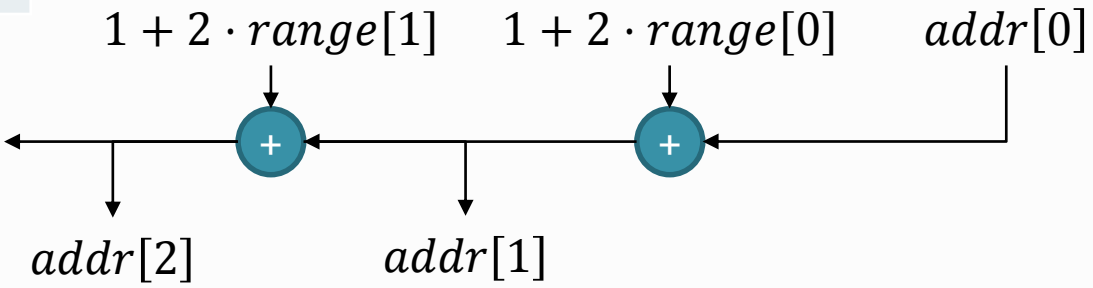
Efficient Address Storage

range 4-bit	addr 22-bit
2	0
4	5
3	12
…	…
4	145
7	154

24 entries

range [23]				range [1]	range [0]	addr [0]
		…	3	4	2	0
				7	4	145
						…

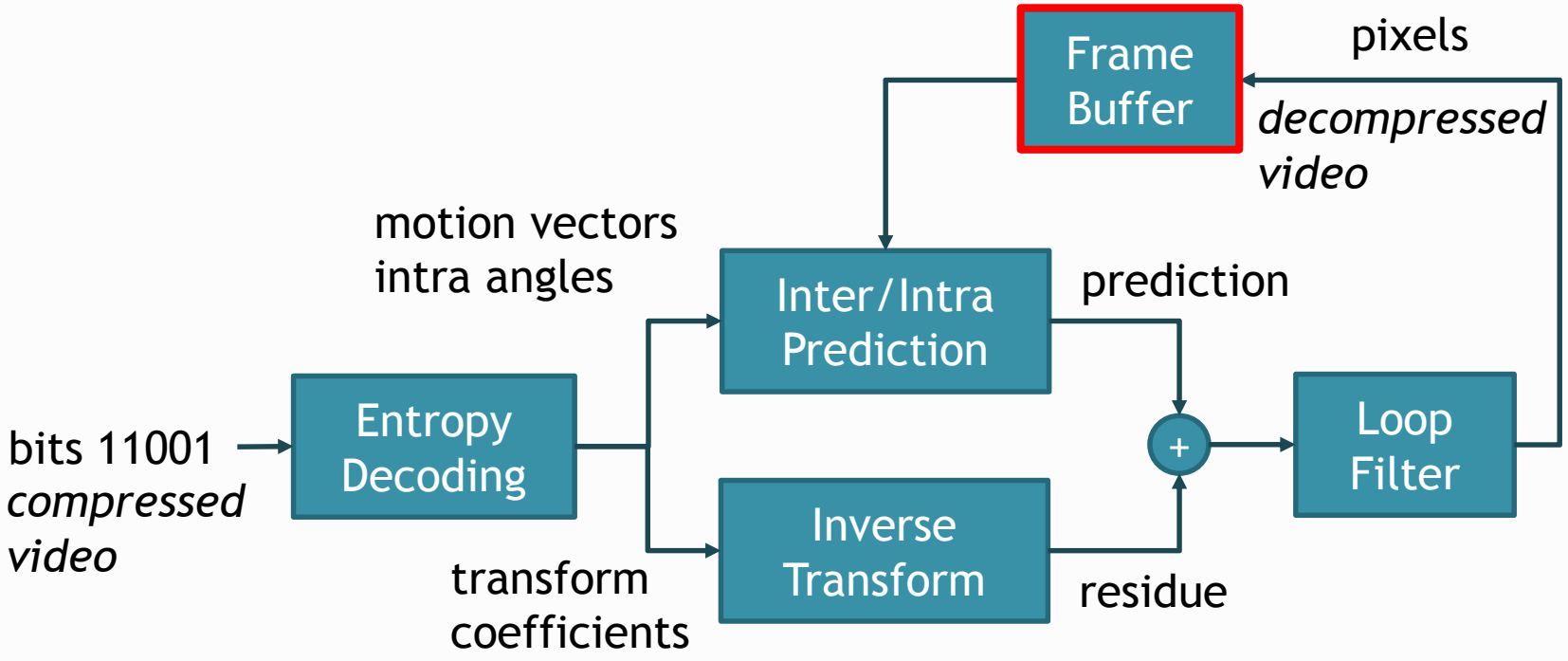
$$addr[1] = addr[0] + (8 + 16 \cdot range[0])/8$$



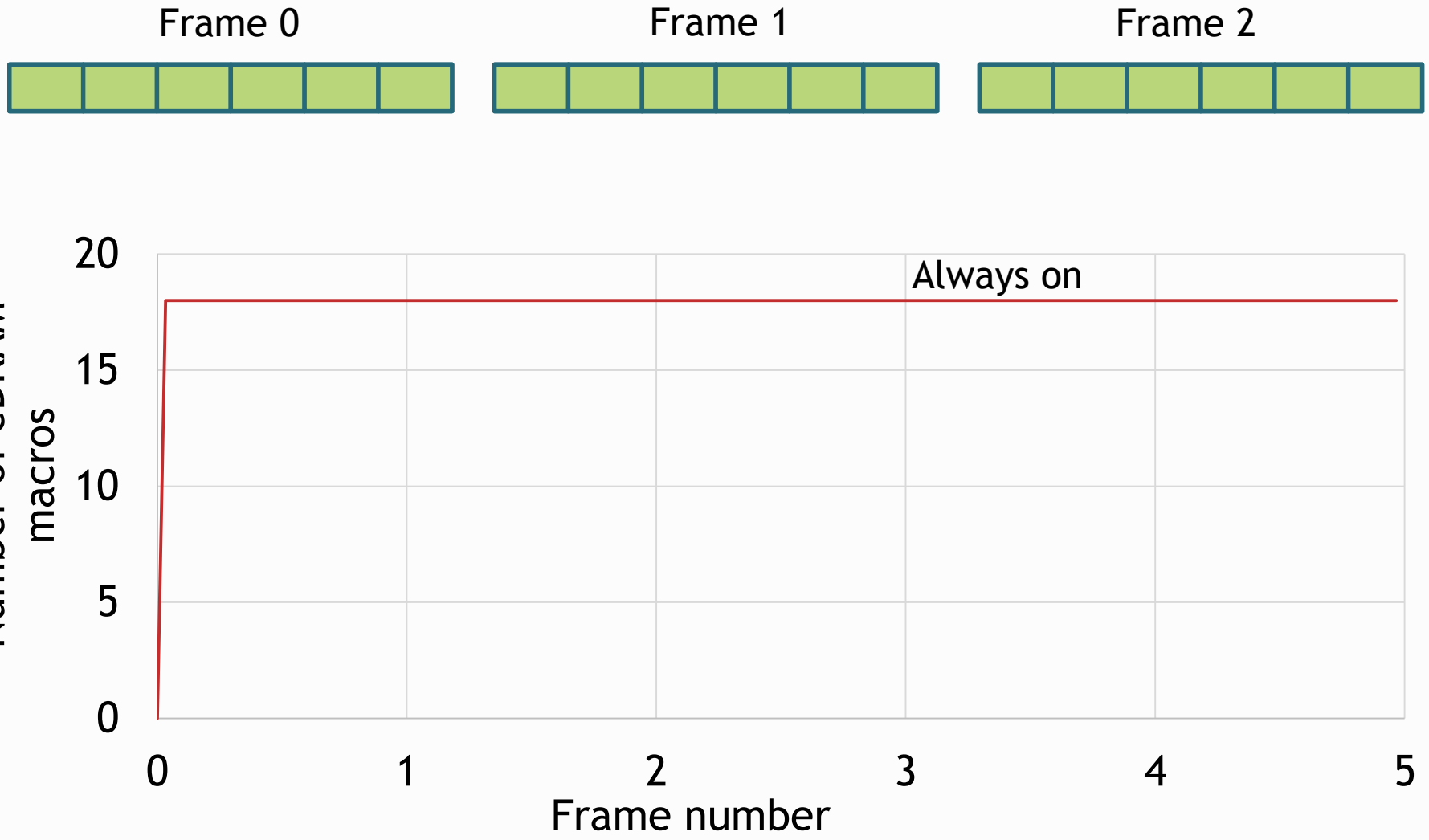
Address lookup overhead reduced from 20% to 4%

Focus of This Talk

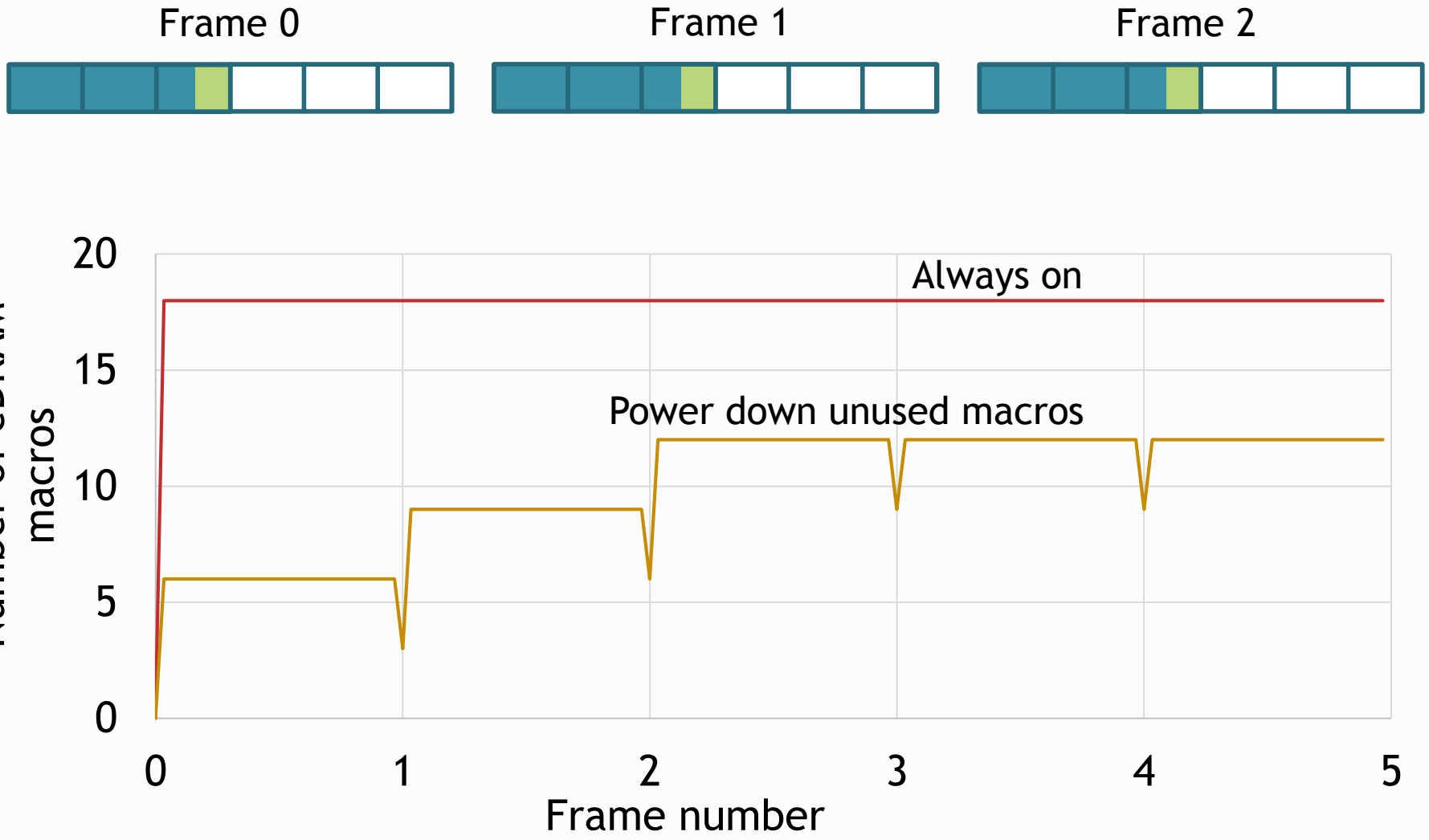
1. Frame Buffer to Motion Compensation
2. **On-demand Power-up of eDRAM**
3. Data Movement of Syntax Elements



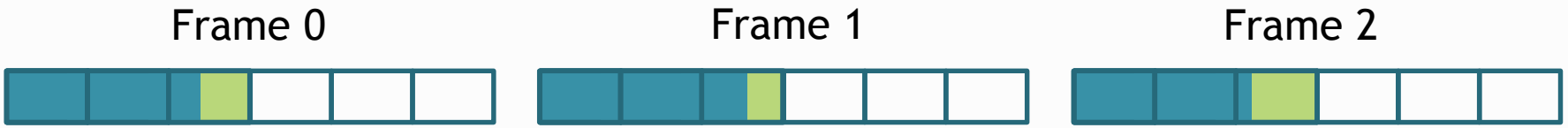
Always On Scheme

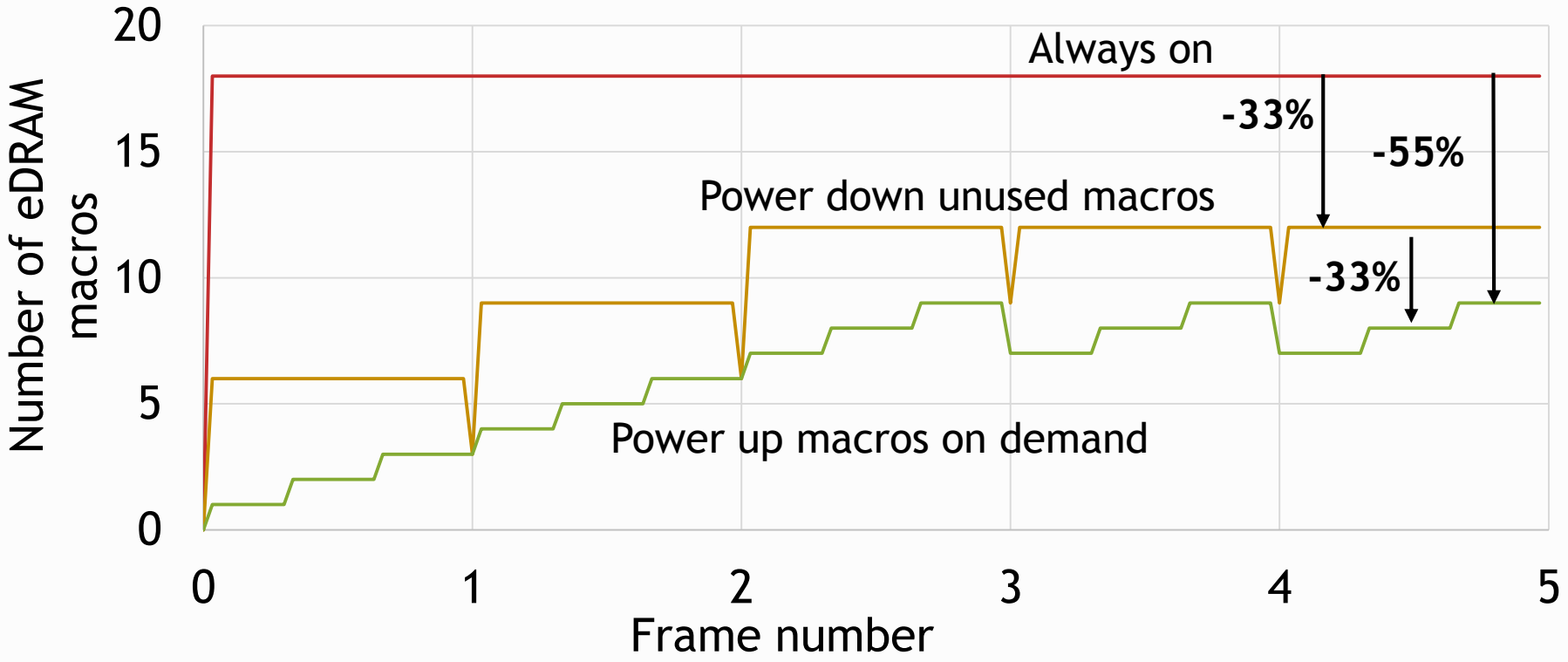


Power Down Unused Macros

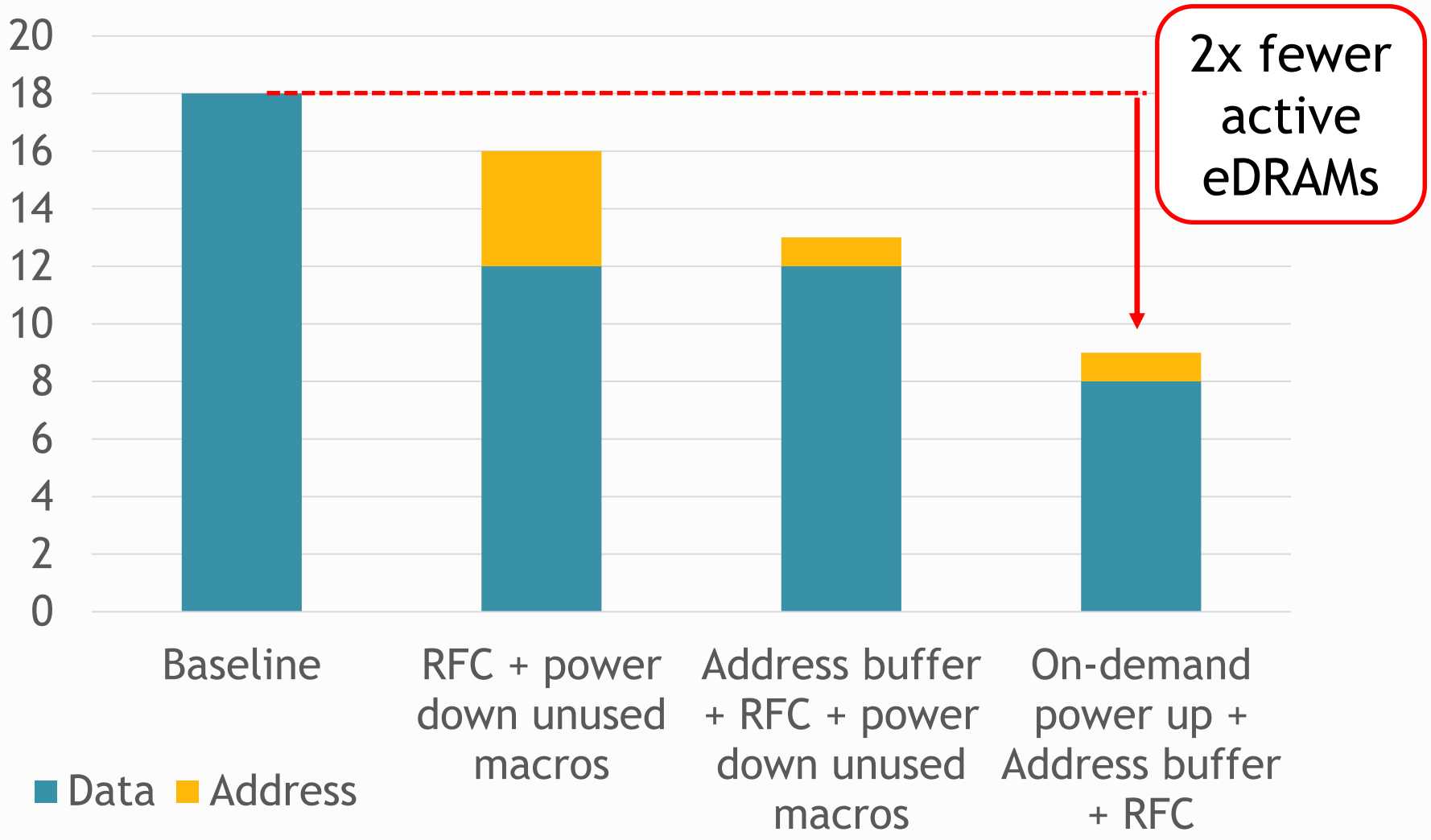


Power Up Macros *On Demand*





Reduction in Number of Active eDRAMs

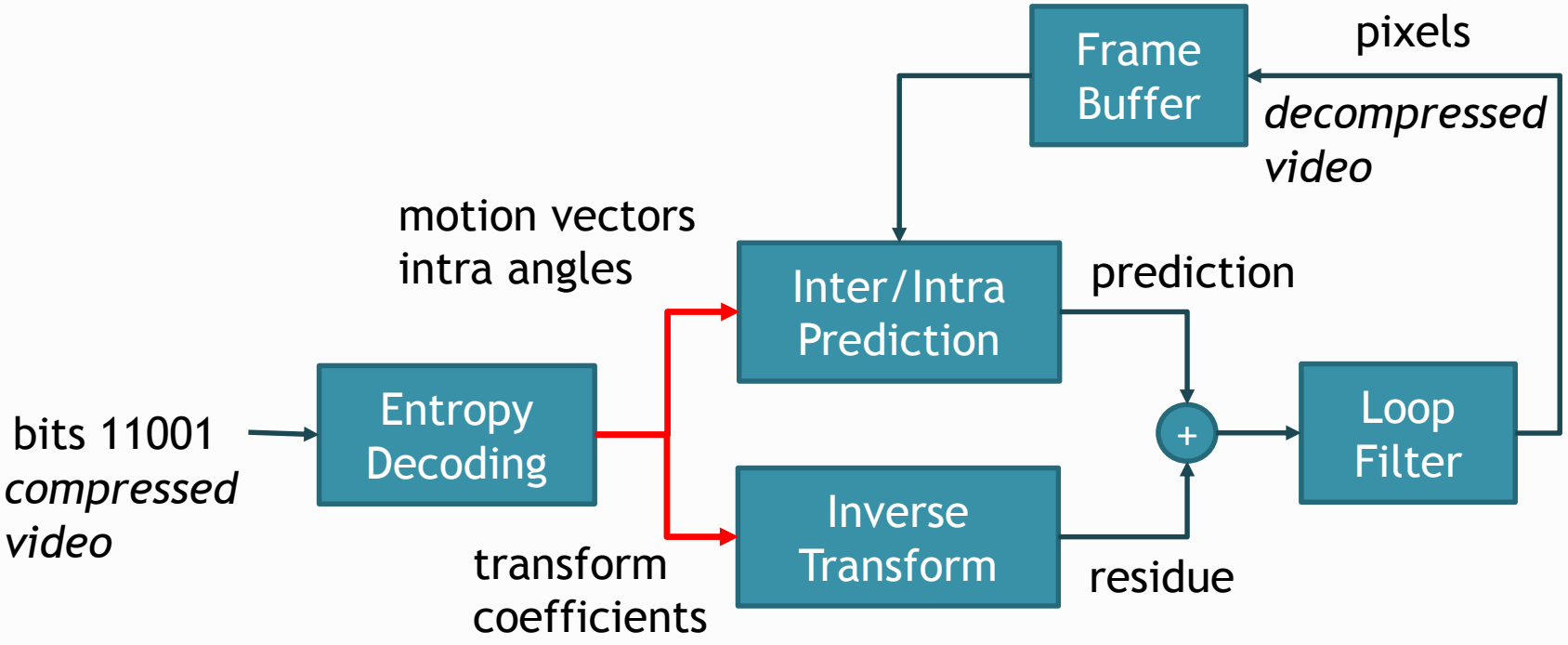


Frame Buffer Energy Savings

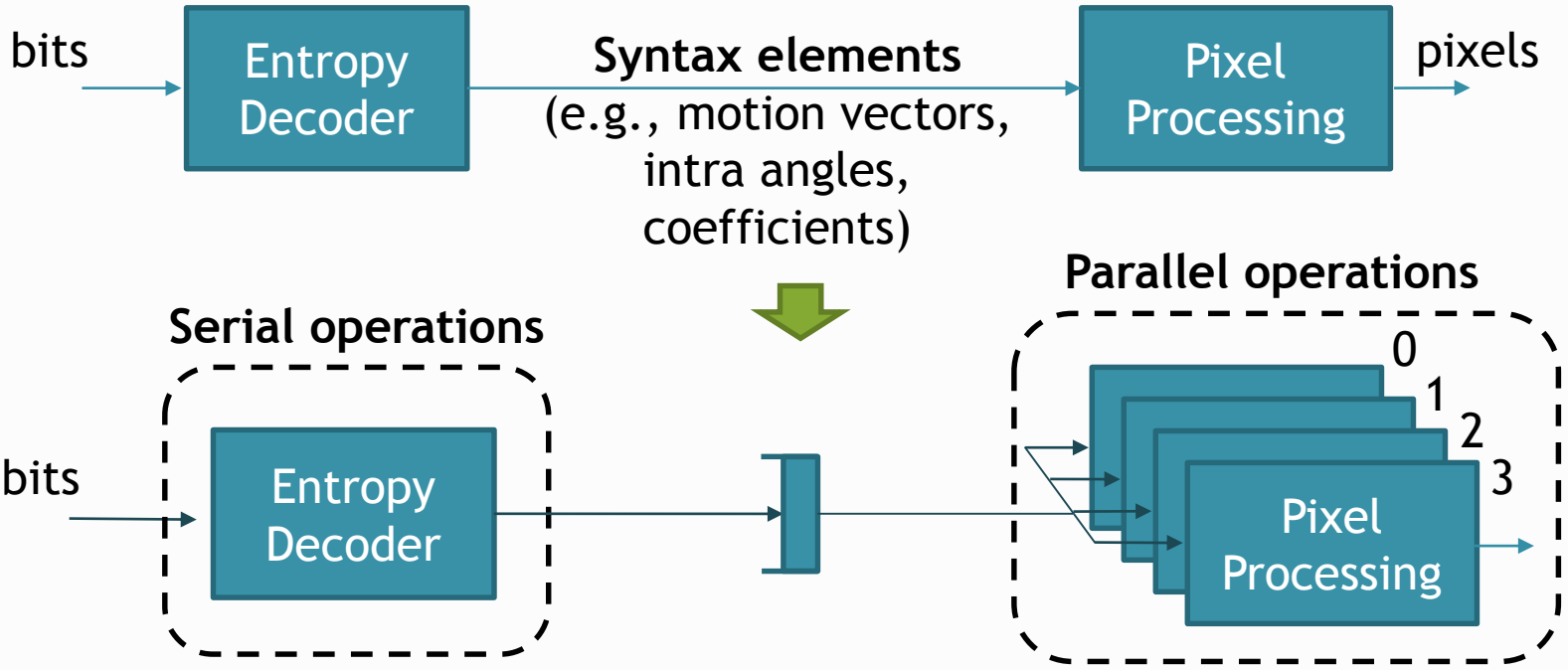
- Refresh power is major challenge for using eDRAM
- RFC compression + decompression in 8 kgates
  - < 1% total gate count of decoder
- Compression achieved: 20% - 80%
- 50% of eDRAM macros in deep power-down mode
- eDRAM refresh power reduced by 5.3 mW
  - 40% memory power
  - 20% system power

Focus of This Talk

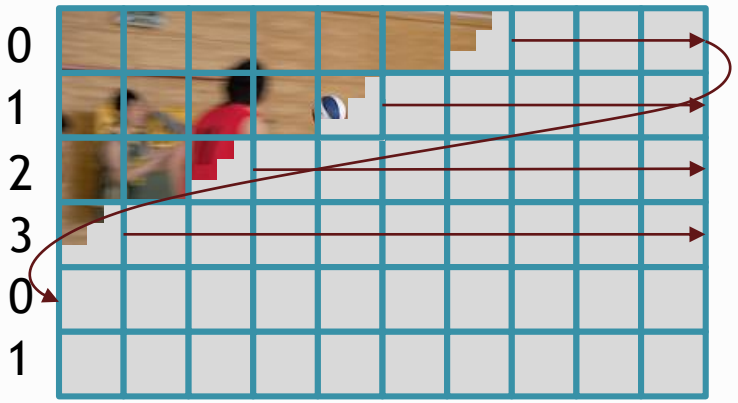
1. Frame Buffer to Motion Compensation
2. On-demand Power-up of eDRAM
3. **Data movement of Syntax Elements**



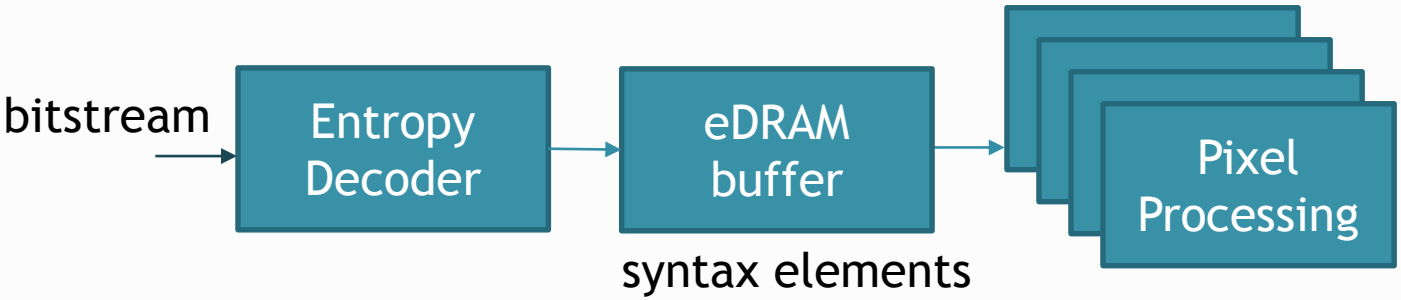
High-level Parallelism in HEVC

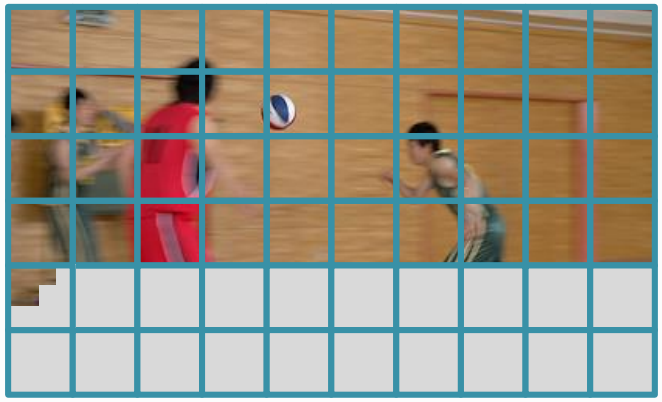


- Each pixel processor operates on 1 row of 64x64 pixel blocks
- Pixel processors are run at **0.25x** clock frequency to reduce power

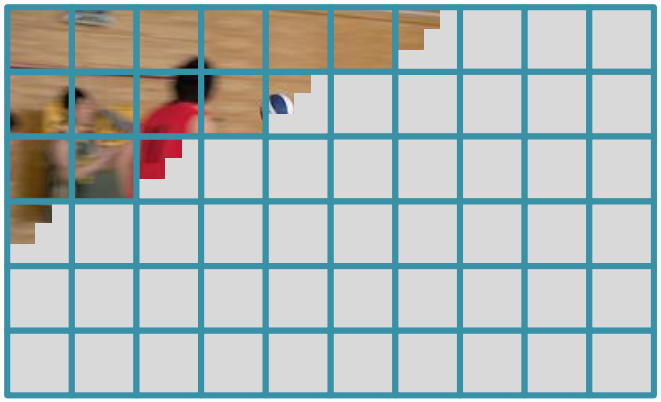


Buffering Requirements





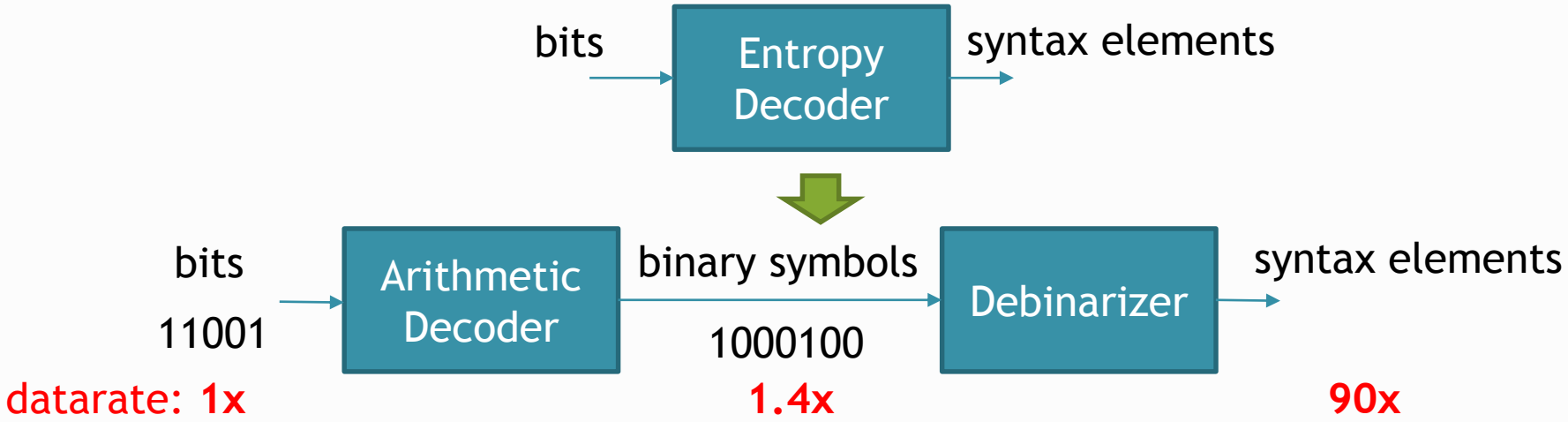
syntax elements in eDRAM buffer



Pixel processors

- A buffer of 8 rows of syntax elements is needed
- Size: **12Mbit** (3 eDRAM macros)
- Bandwidth: **256 MB/s**

Two-stage Entropy Decoding

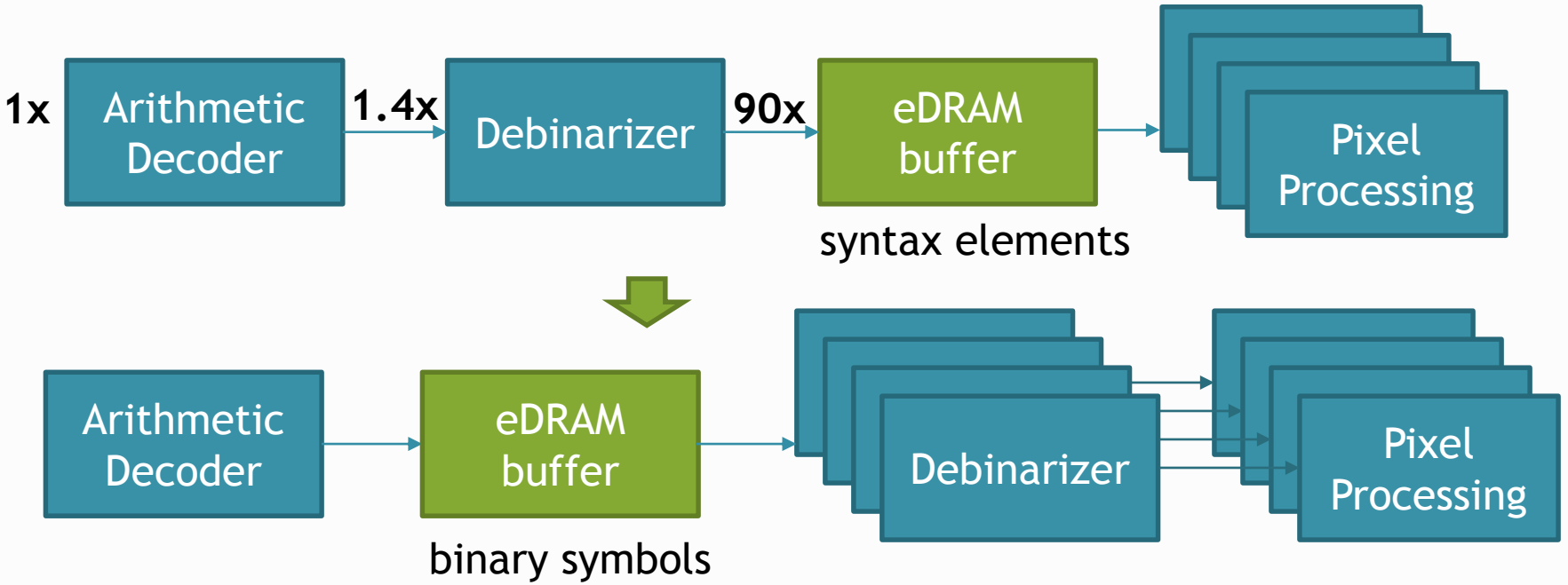


- Arithmetic Decoder[1]
  - Uses probabilities of 0s and 1s
  - Context Adaptive Binary Arithmetic Coding (CABAC)
- Debinarizer
  - Parses stream of binary symbols
  - Huffman Coding, Run Length Coding

Store compact binary symbols in eDRAM to save access and refresh power

[1] Y.-H. Chen, ICIP 2014

Reducing Data Movement of Syntax Elements

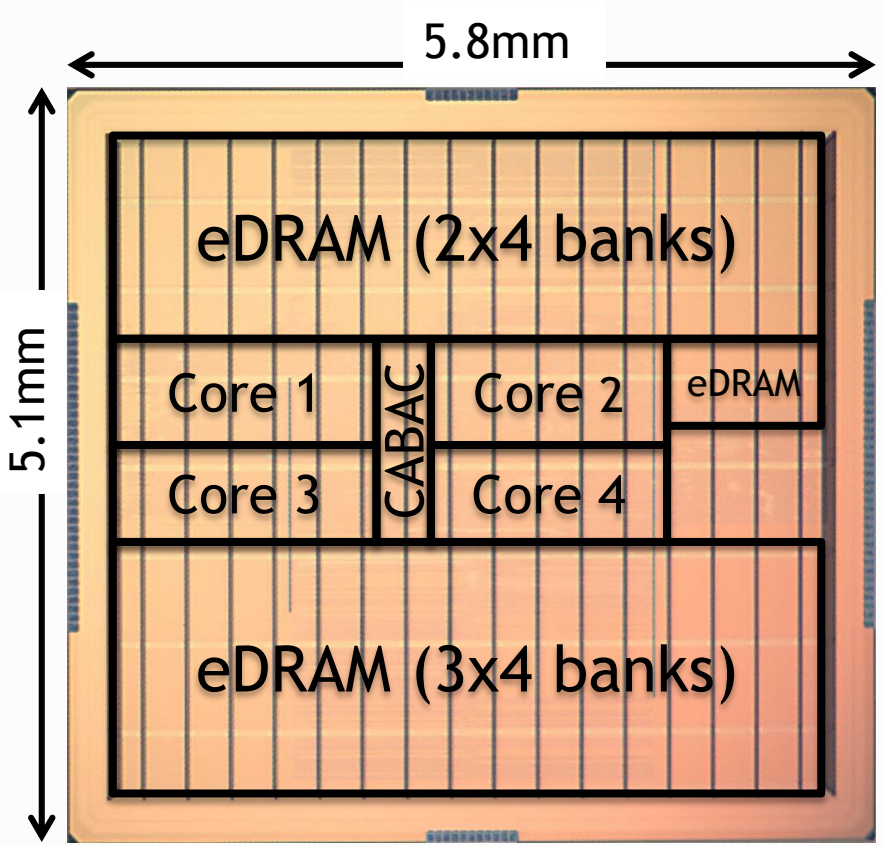


- Bandwidth reduction: **66x** (256MB/s → 3.9MB/s)
- Energy savings: **4.2mW** (16% of total power)
- Chip area reduction: **6%**

Exploit built-in HEVC compression to reduce data movement

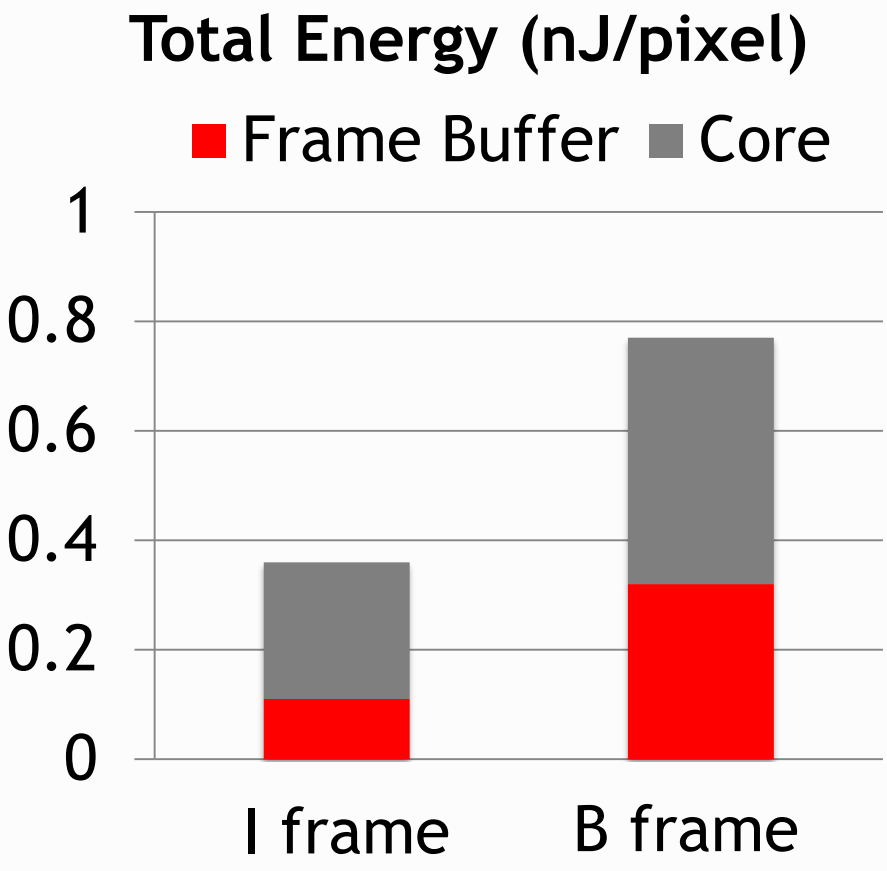
Chip Results

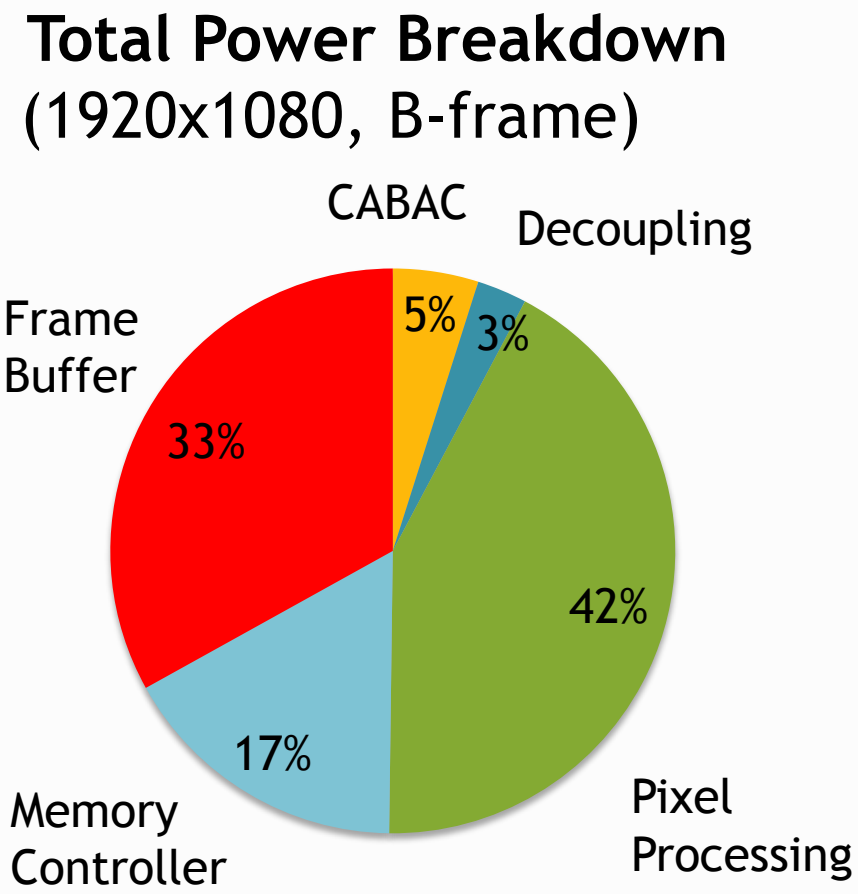
Technology	TSMC 40nm LP
Supply Voltage	Core 0.8 - 1.1V eDRAM 1.1V I/O 2.5V
Standard	H.265/HEVC (Main Profile)
Chip Size	5.8 mm x 5.1 mm
Logic Count	1,122 kgates
On-Chip SRAM	162.75 kB
On-Chip eDRAM	21 x 0.5MB
Max Resolution	1920 x 1080
Max Throughput	47.9Mpixels/s
Power at 1.1V	24.9mW



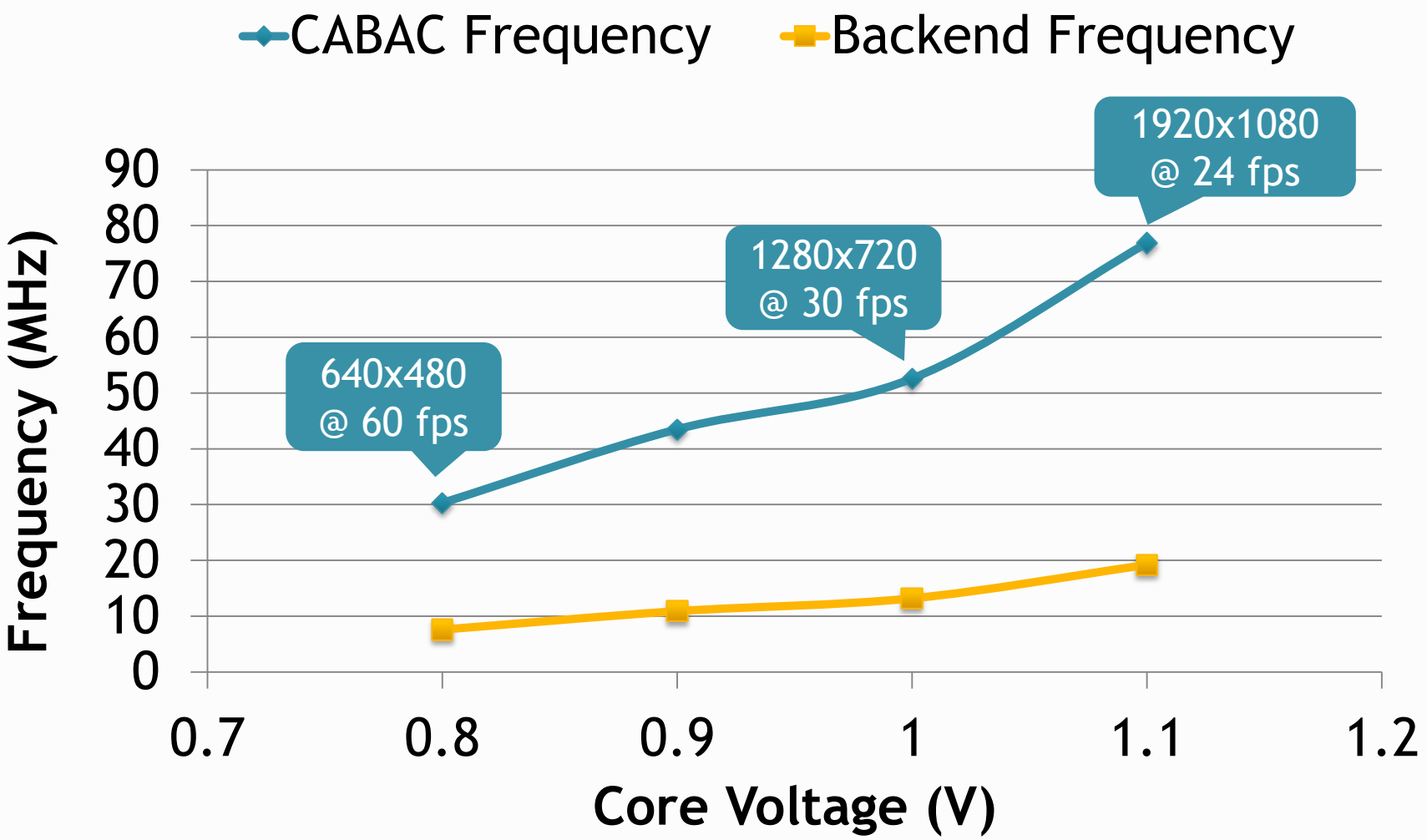
Thanks to TSMC University Shuttle for chip fabrication

Energy and Power breakdown





Voltage-Frequency Scaling

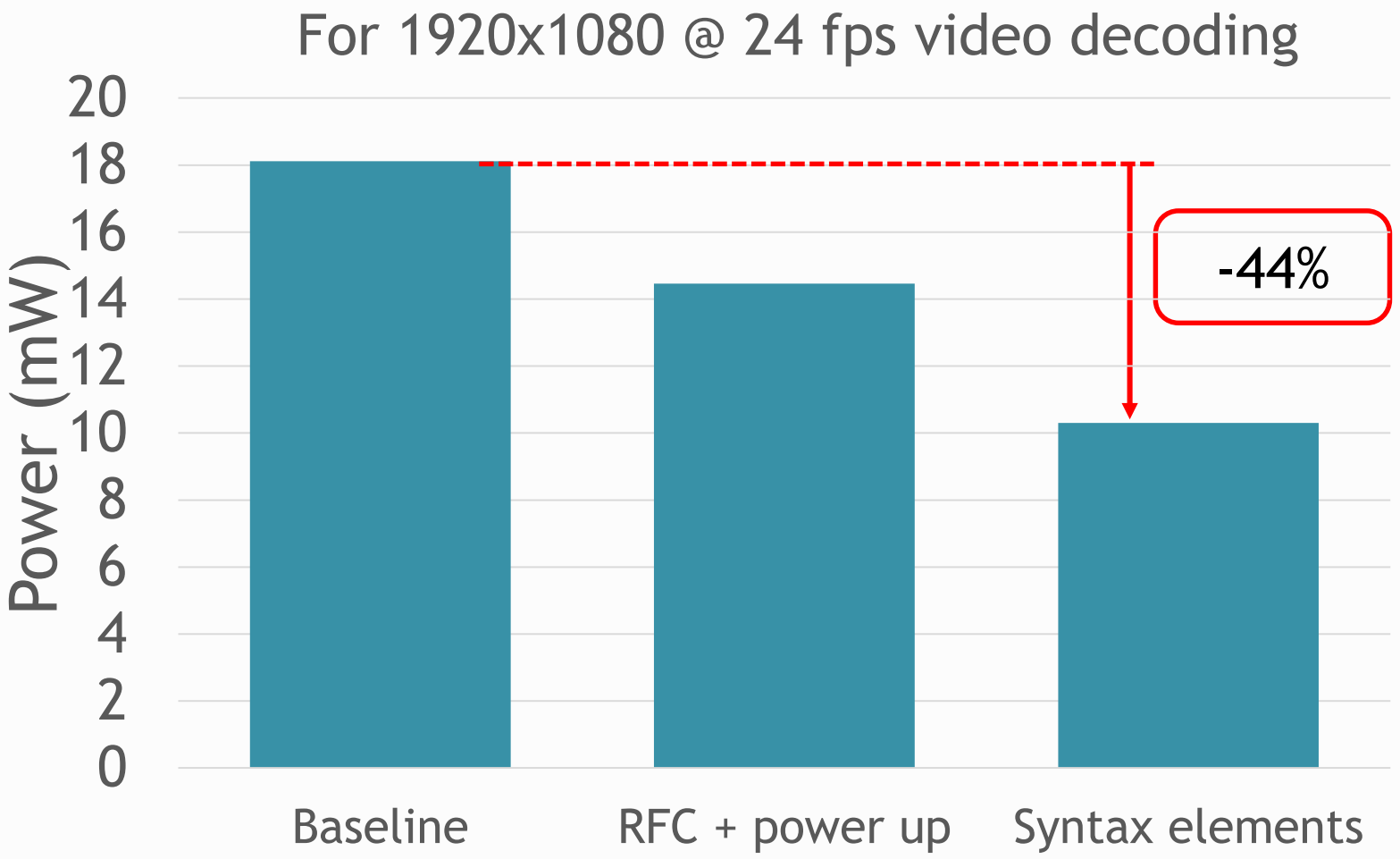


Comparison with previous work

	This Work	ISSCC 2013
Standard	H.265/HEVC	H.265/HEVC WD4
Gate Count	1438K	715K
On-Chip Storage	162.75kB	124kB
Technology	40nm/1.1V	40nm/0.9V
Max Throughput	1920x1080@24fps	3840x2160@30fps
Max Frequency	80MHz/20MHz	200MHz
Frame buffer Storage	128b eDRAM	32b DDR3
1920 x 1080 @ 24 fps decoding power		
Core Power [mW]	14.6	36*
Frame Buffer Power [mW]	10.3	150*
System Power [mW]	24.9	186*

* Estimated by scaling core frequency and memory bandwidth

eDRAM Power Savings



Contributions

Energy-efficient video decoding on wearables

- 1920x1080 at 24fps in **25mW system power**
- Fully-integrated solution minimizes board footprint

Data-dependent energy saving in memory access

- RFC to reduce eDRAM refresh power (20%)
- On-demand power up of eDRAM macros
- Movement of syntax elements (16%)

Energy-efficient use of Embedded DRAM

- **1.8x power saving** in eDRAM

Thanks to TSMC University Shuttle for chip fabrication and NSF for funding