IDCT pruning

(JCTVC-E386)

Madhukar Budagavi

Texas Instruments

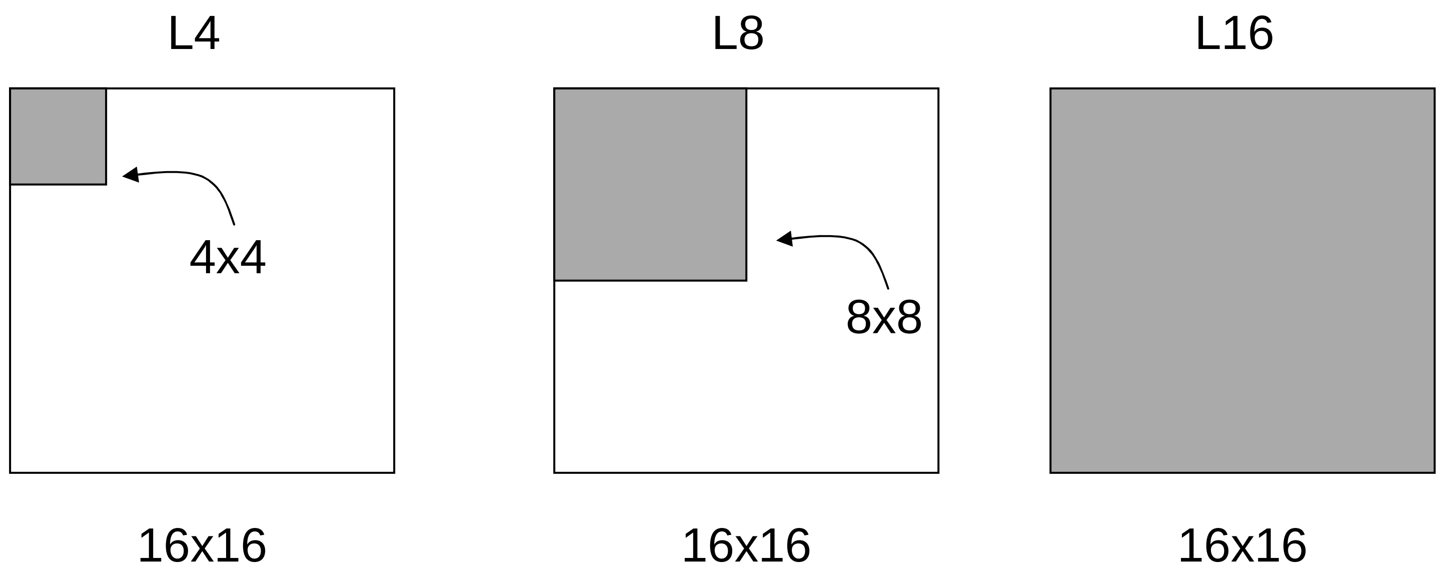
Joint Collaborative Team on Video Coding (JCT-VC)
of ITU-T SG16 WP3 and ISO/IEC JTC1/SC29/WG11
5th Meeting: Geneva, CH, 16-23 March, 2011

IDCT pruning

- High frequency region of large transforms is typically zero due to quantization and energy compaction properties of transform
- One can use pruning to eliminate IDCT computations that have zero input
  - This was highlighted by D. Flynn (BBC) during transform breakout activity in Daegu
- Pruning leads to:
  - Reduction in SIMD computational complexity
  - And also corresponding power savings in hardware transform engines
- Non-normative contribution
- Goal of the contribution is to motivate that pruning properties of large transforms be considered in design of HEVC transforms
- *Note: IDCT abbreviation used to denote inverse transform*

Statistics measured for 16x16 block

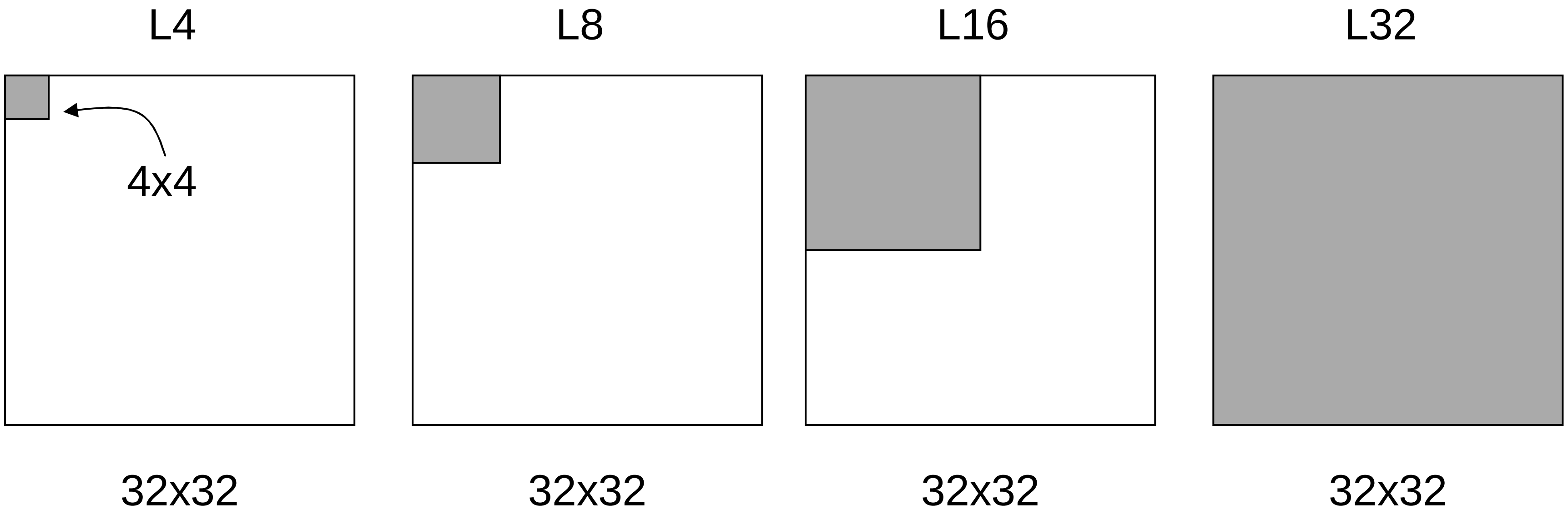
- L*K* denotes a transform block that has only the lowest frequency *K*x*K* sub-block to be non-zero
- Number of L4, L8, L16 occurrences were measured in HM 2.0 anchor bitstreams



Gray color indicates non-zero sub-block

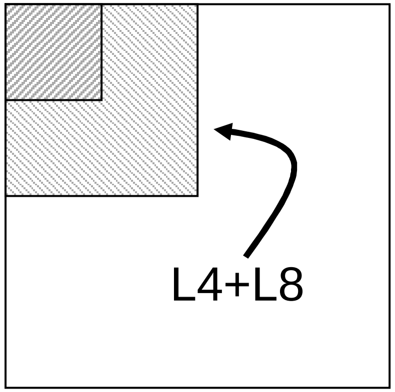
Statistics measured for 32x32 block

- L*K* denotes a transform block that has only the lowest frequency *KxK* sub-block to be non-zero
- Number of L4, L8, L16 occurrences were measured in HM 2.0 anchor bitstreams



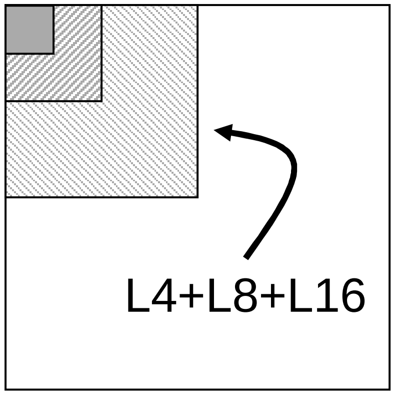
Gray color indicates non-zero sub-block

Statistics summary



16x16

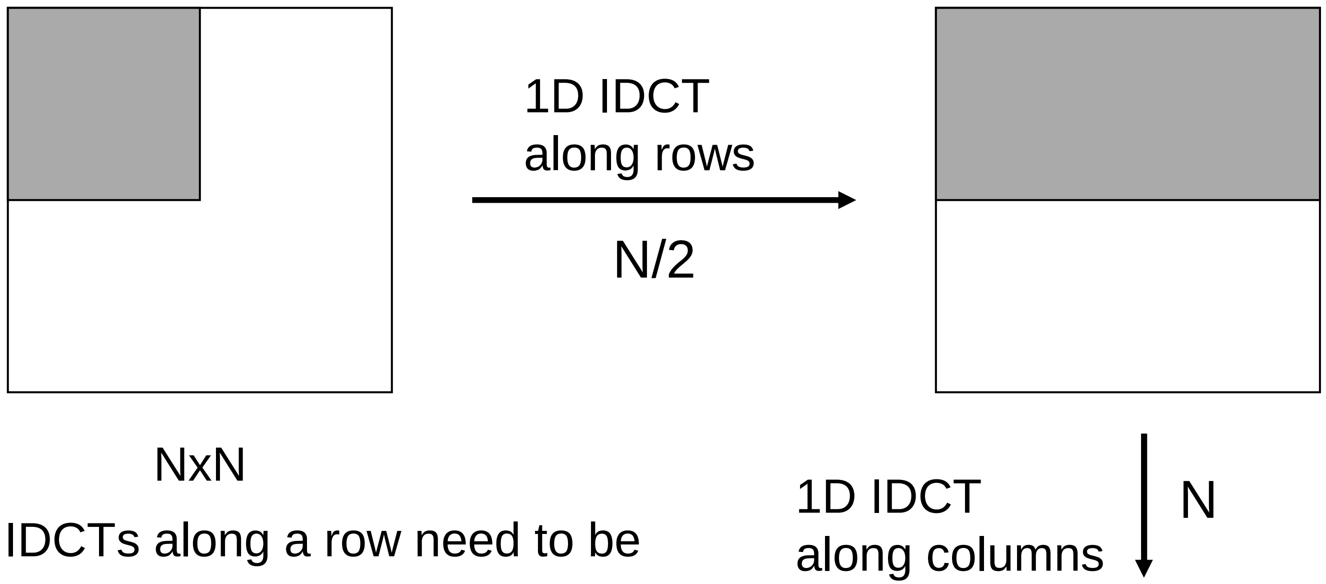
	High efficiency	Low complexity
16x16 IDCT		
Number of L4	59%	64%
Number of L8	18%	29%
Number of L16	23%	7%
Number of L4+L8	77%	93%



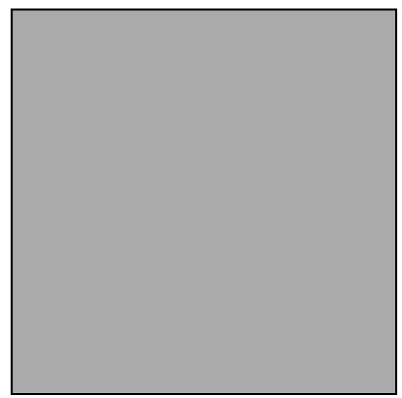
32x32

32x32 IDCT		
Number of L4	44%	61%
Number of L8	15%	29%
Number of L16	17%	10%
Number of L32	24%	0%
Number of L4+L8+L16	76%	100%

Complexity implications: Reduction in number of 1D IDCT



- Only N/2 1D IDCTs along a row need to be carried out
- N 1D IDCTs along columns need to be carried out
- 25% savings in number of 1D IDCTs that need to be carried out
- Idea can be applicable to most IDCTs architectures



Complexity implications: Pruning within 1D IDCT (example using matrix multiplication)

c4	c1	c2	c3	c4	c5	c6	c7	X0	
c4	c3	c6	-c7	-c4	-c1	-c2	-c5	X1	
c4	c5	-c6	-c1	-c4	c7	c2	c3	X2	
c4	c7	-c2	-c5	c4	c3	-c6	-c1	X3	N^2 mults
c4	-c7	-c2	c5	c4	-c3	-c6	c1	X4	
c4	-c5	-c6	c1	-c4	-c7	c2	-c3	X5	
c4	-c3	c6	c7	-c4	c1	-c2	c5	X6	
c4	-c1	c2	-c3	c4	-c5	c6	-c7	X7	

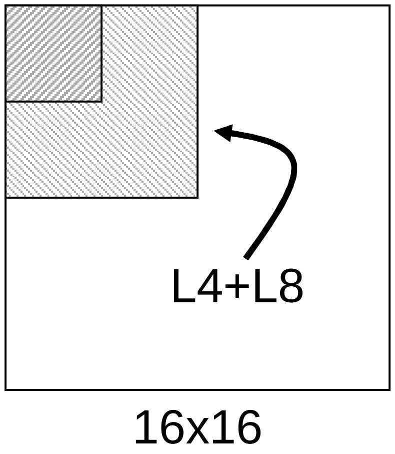
c4	c1	c2	c3	c4	c5	c6	c7	X0	
c4	c3	c6	-c7	-c4	-c1	-c2	-c5	X1	
c4	c5	-c6	-c1	-c4	c7	c2	c3	X2	
c4	c7	-c2	-c5	c4	c3	-c6	-c1	X3	
c4	-c7	-c2	c5	c4	-c3	-c6	c1	0	$N^2/4$ mults
c4	-c5	-c6	c1	-c4	-c7	c2	-c3	0	
c4	-c3	c6	c7	-c4	c1	-c2	c5	0	
c4	-c1	c2	-c3	c4	-c5	c6	-c7	0	

Complexity implications: Pruning within 1D IDCT

- Examples using HM 2.0 IDCT and JCTVC-E243 partial butterfly IDCT (which is equivalent to matrix multiplication)
  - See Appendix in JCTVC-E386 document

Computation comparison: 16x1 IDCT

- Mults counted are 16x8b mults
  - 32x8b mult counted as 2 16x8b mults



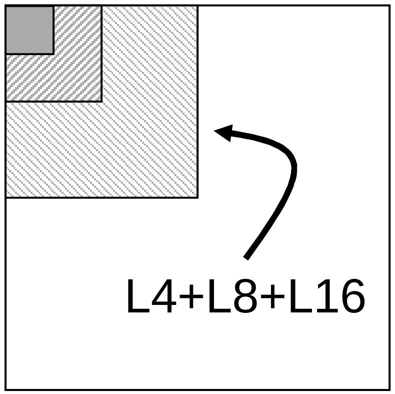
	L16		L8		L4	
	Mult	Add	Mult	Add	Mult	Add
HM 2.0 - 16x16	58	66	44	58	36	42
Partial butterfly - 16x16 (E243)	82	100	44	46	22	32

Note: Operation count might not be a good representative of implementation complexity since SIMD properties of transform architectures are different.

E.g. E243 (matrix multiplication) could have better SIMD properties compared to HM 2.0 transform

Computation comparison: 32x1 IDCT

- Mults counted are 16x8b mults: 32x8b mult counted as 2 16x8b mults



32x32

	L32		L16		L8		L4	
	Mult	Add	Mult	Add	Mult	Add	Mult	Add
HM 2.0 - 32x32	170	194	140	162	124	130	84	88
Partial butterfly - 32x32 (E243)	342	372	172	200	86	112	42	64

Note: Operation count might not be a good representative of implementation complexity since SIMD properties of transform architectures are different.

E.g. E243 (matrix multiplication) could have better SIMD properties compared to HM 2.0 transform

Conclusions

- High frequency region of large transforms is typically zero due to energy compaction properties of transform and quantization
- We can use this information to prune computations
  - Leads to reduction in SIMD computational complexity
  - And also corresponding power savings in hardware transform engines
- Recommend pruning properties of large transforms be considered in design of HEVC transforms

Transform size distribution

- Transform size distribution in term of percentage of number of pixels (luma+chroma) transformed

Intra	**4x4**	**8x8**	**16x16**	**32x32**
Average	30%	27%	26%	16%
Maximum	79%	61%	57%	59%
Minimum	0%	0%	0%	0%

Inter	**4x4**	**8x8**	**16x16**	**32x32**
Average	2%	4%	5%	5%
Maximum	15%	35%	44%	62%
Minimum	0%	0%	0%	0%