

AhG7: Overflow Prevention in HEVC inverse transform

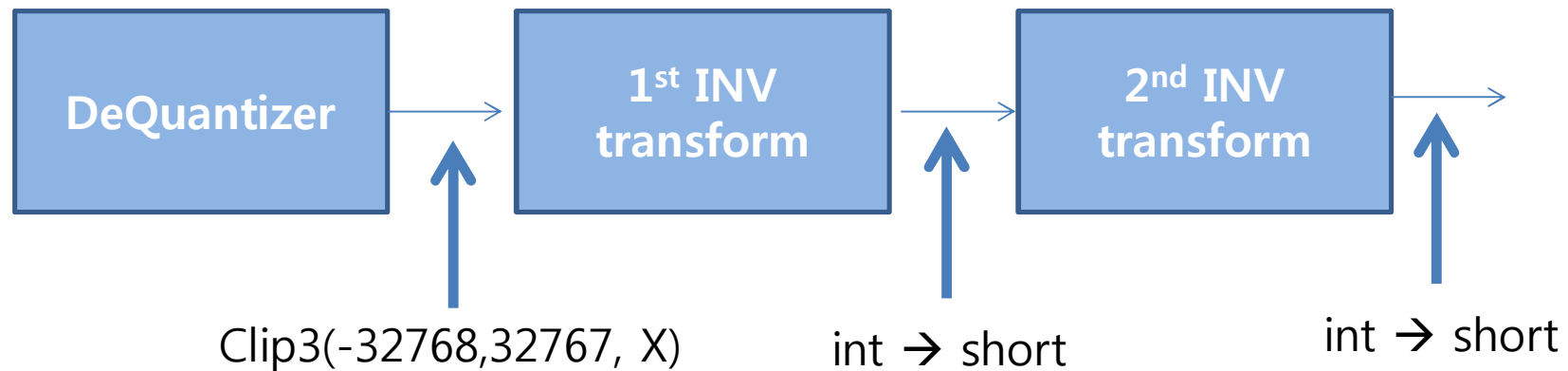
JCTVC-G782

Elena Alshina

Alexander Alshin

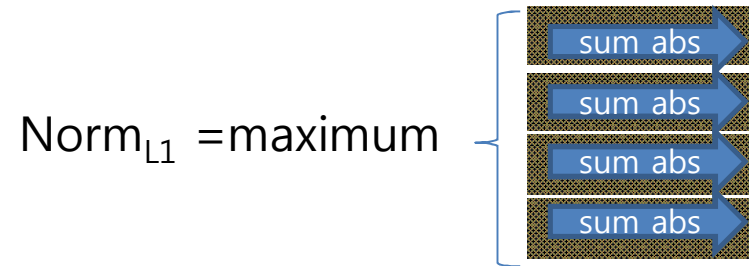
JeongHoon Park

Inverse transform implementation in HM



De facto we have clipping both after 1st and 2nd inverse transforms

Dynamic range for matrix product



If

$$A = B \times C \quad (X = \text{Max}_{ij} |C_{ij}|)$$

then

$$\text{Max}_{ij} |A_{ij}| = N * X \quad (***)$$

Here

$$N = \text{Norm}_{L1}(B) = \text{Max}_i (\sum_j |B_{ij}|)$$

(Sum along rows; Maximum across all rows)

This is always possible to find C which will realize (***)

Dynamic range in inverse transform

Internal Bit-Depth 8

JVTCV E243 Inverse transform

HM	Size	L1Norm	MaxAbs coeff	reg ₁ bits	temp ₂ bits	reg ₂ bits	rec res bits
DST	4	242	32767	24	17	25	13
DCT	4	247	32767	24	17	25	13
	8	479	32767	25	18	27	15
	16	940	32767	26	19	29	17
	32	1862	32767	27	20	31	19

4pt example of 16 bits overflow(1/2)

$$\begin{bmatrix} X & X & X & X \\ X & X & X & X \\ X & X & X & X \\ X & X & X & X \end{bmatrix} = \text{Clip}(\text{coeff}, \text{Min}, \text{Max})$$

Min = -32768
 Max = 32767
 X = 32767 (within 16 bits!)

$$\text{register}_2 = \begin{bmatrix} 64 & 83 & 64 & 36 \\ 64 & 36 & -64 & -83 \\ 64 & -36 & -64 & 83 \\ 64 & -83 & 64 & -36 \end{bmatrix} \times \begin{bmatrix} X & X & X & X \\ X & X & X & X \\ X & X & X & X \\ X & X & X & X \end{bmatrix} = \begin{bmatrix} 247X & 247X & 247X & 247X \\ -47X & -47X & -47X & -47X \\ 47X & 47X & 47X & 47X \\ -119X & -119X & -119X & -119X \end{bmatrix}$$

$$\text{temp}_2 = \text{register}_2 / 128 = \begin{bmatrix} 1.9X & 1.9X & 1.9X & 1.9X \\ -0.4X & -0.4X & -0.4X & -0.4X \\ 0.4X & 0.4X & 0.4X & 0.4X \\ -0.9X & -0.9X & -0.9X & -0.9X \end{bmatrix} = \begin{bmatrix} 63230 & 63230 & 63230 & 63230 \\ -12032 & -12032 & -12032 & -12032 \\ 12032 & 12032 & 12032 & 12032 \\ -30463 & -30463 & -30463 & -30463 \end{bmatrix}$$

$$\begin{bmatrix} X & X & X & X \\ X & X & X & X \\ X & X & X & X \\ X & X & X & X \end{bmatrix}$$

X= 32767

Cannot be output of forward 4pt transform,
 but can come from "non -confirmed" bit-stream

4pt example of 16 bits overflow(2/2)

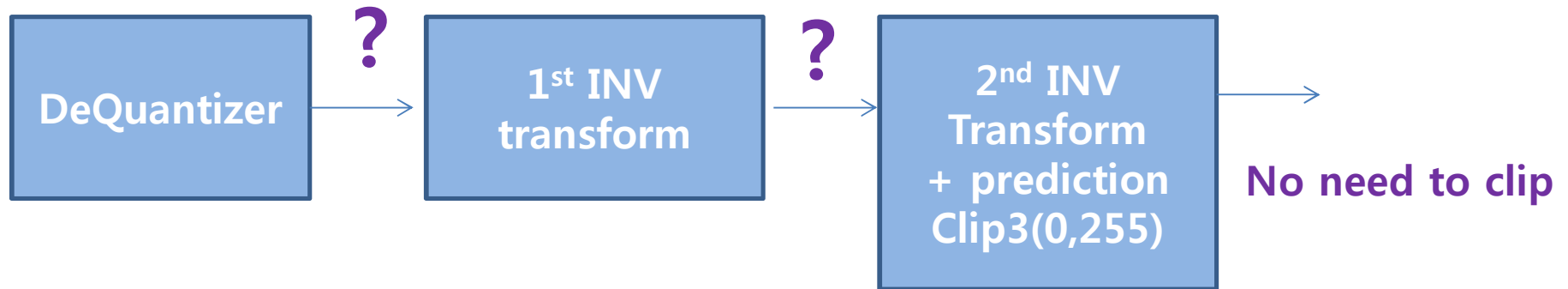
$$\text{Register3} = \begin{bmatrix} 63230 & 63230 & 63230 & 63230 \\ \dots & \dots & \dots & \dots \\ \dots & \dots & \dots & \dots \\ \dots & \dots & \dots & \dots \end{bmatrix} \times \begin{bmatrix} 64 & 64 & 64 & 64 \\ 84 & 36 & -36 & -83 \\ 64 & -64 & -64 & 64 \\ 36 & -83 & 83 & -36 \end{bmatrix} =$$

$$= \begin{bmatrix} 15617810 & \dots & \dots & \dots \\ \dots & \dots & \dots & \dots \\ \dots & \dots & \dots & \dots \\ \dots & \dots & \dots & \dots \end{bmatrix}$$

$$\text{Rec_res} = \text{register}_3 / 256 = \begin{bmatrix} 61007 & \dots & \dots & \dots \\ \dots & \dots & \dots & \dots \\ \dots & \dots & \dots & \dots \\ \dots & \dots & \dots & \dots \end{bmatrix}$$

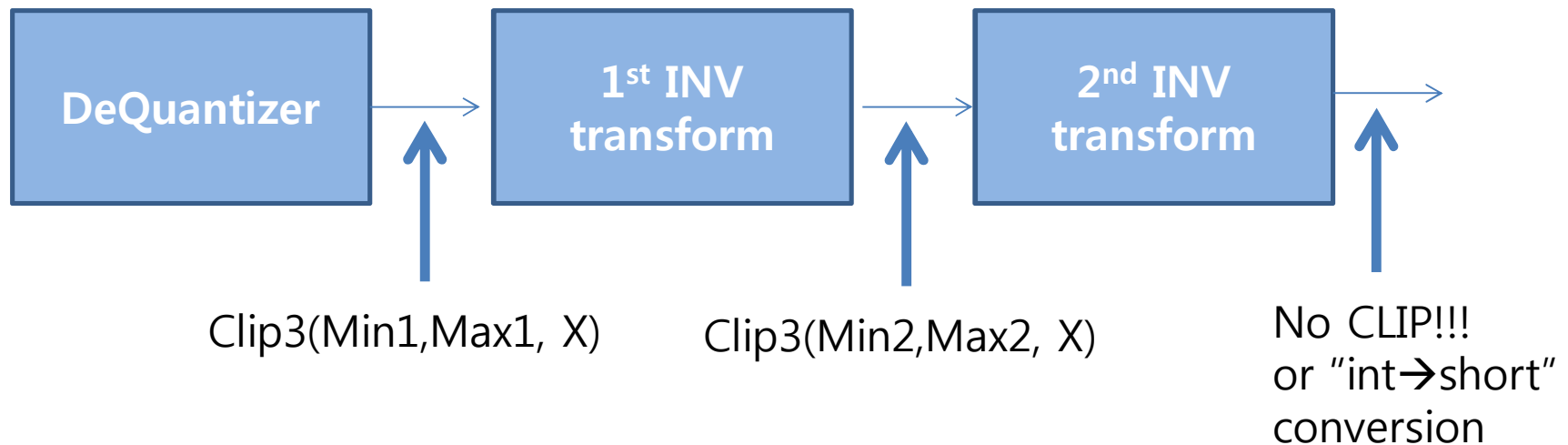
De-scaling factor '256" corresponds "internal bit depth" 12

Hypothetical implementation



Overflow after 2nd transform may be resolved by implementation

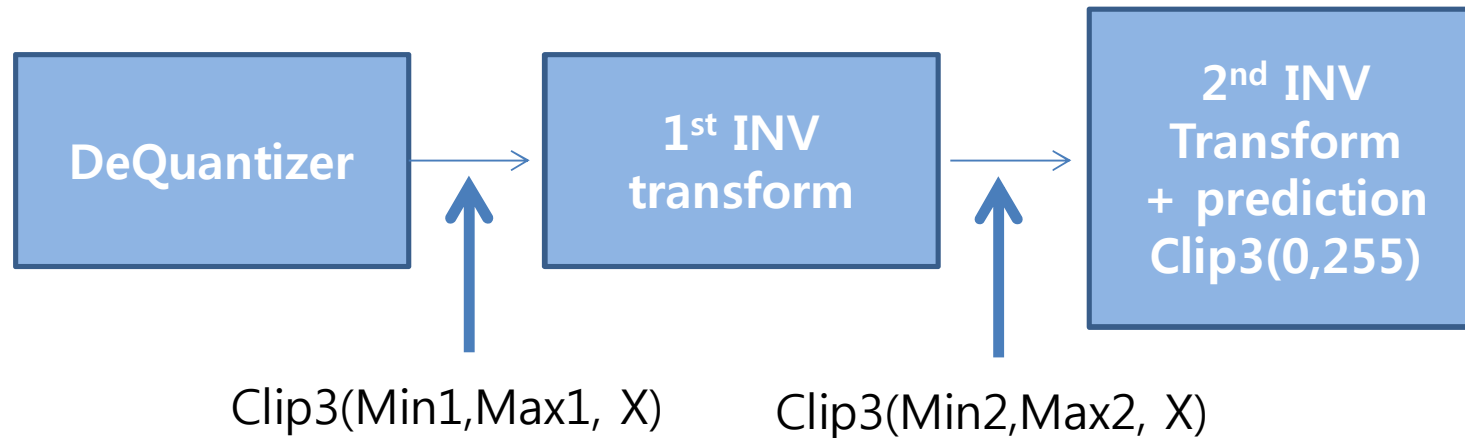
Proposed design (1)



Need to put to the specification:

- Clip to Min1...Max1 range after de-Quantization
- Clip to Min2...Max2 range after 1st INV transform
- No clipping or "int→short" conversion

Proposed design (2)



Need to put to the specification:

- Clip to Min1...Max1 range after de-Quantization*
- Clip to Min2...Max2 range after 1st INV transform
- Add residual and clip to original signal range before output of 2nd transform

(*)Preferable to move clipping from De-Quantizer to earlier stage (JCT VC-G719)

Min1, Max1? Min2, Max2?

JVTCV E243 Inverse transform						internal bit-depth		10
	Size	L1Norm	coeff <= Max1	reg ₂ bits	temp2 <= Max2	temp ₂ bits	reg3 bits	rec res bits
DST	4	242	32767	24	32767	16	24	14
DCT	4	247	32767	24	32767	16	24	14
	8	479	32767	25	32767	16	25	15
	16	940	32767	26	32767	16	26	16
	32	1862	32767	27	32767	16	27	17

Min1=-32767; Max1= 32767

Min2=-32767; Max2= 32767

Min1=Min2 = -32768 (if it is preferable somehow)

Min1, Max1? Min2, Max2?

JVTCV F251 Inverse transform				internal bit-depth				8
	Size	L1Norm	coeff	reg ₂ bits	temp2	temp ₂ bits	reg3 bits	rec res bits
DST	4	15488	32767	30	32767	16	30	12
DCT	4	15808	32767	30	32767	16	30	12
	8	30622	32767	31	32767	16	31	13
	16	60326	32767	32	32767	16	32	14
	32	119262	18006	32	18006	16	32	15

Size < 32 Min1=-32767; Max1= 32767
 Min2=-32767; Max2= 32767
 Min1=Min2 = -32768 (if it is preferable somehow)

Size = 32 Min1=Min2 = -18006
 Min1=Max2 = 18006

Conclusions:

We would like to suggest following inverse transform framework for HEVC:

- Restrict dynamic range of de-quantized coefficients (preferably move restriction prior de-quantizer JCTVC-G719)
- Restrict dynamic range of the out-put of 1st inverse transform
- Add residual and clip to out-put signal range before output of 2nd transform