

JCTVC-F499

Temporal QP memory compression

Muhammed Coban, Marta Karczewicz (Qualcomm)

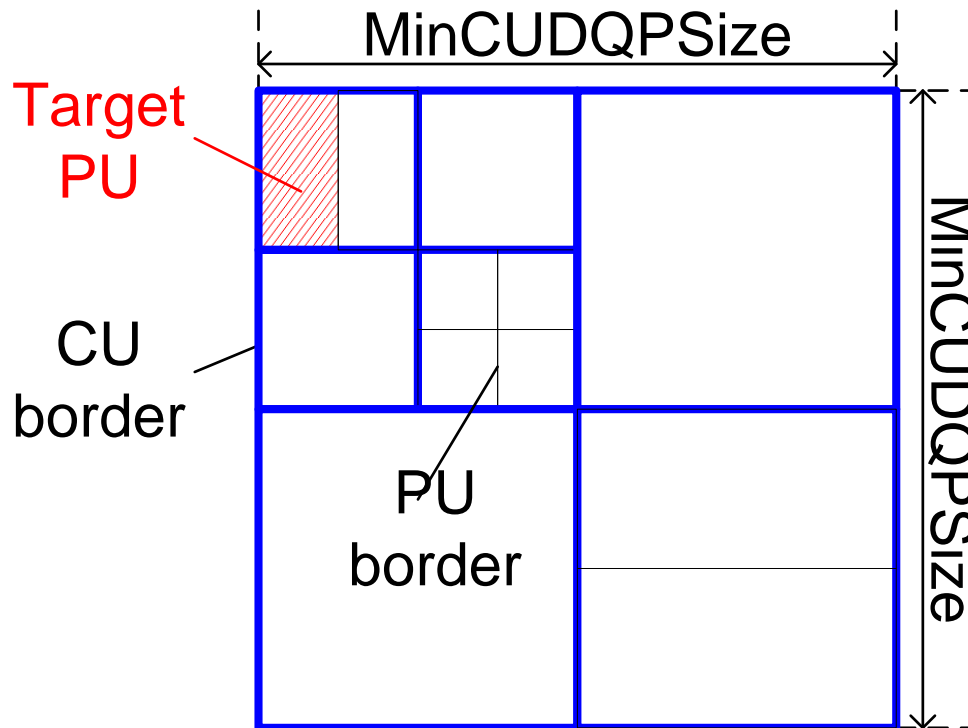
Hirofumi Aoki, Keiichi Chono (NEC)

Motivation

- QP prediction based on intra/inter prediction (JCTVC-F103)
 - Tested as 2.4.b in CE4 Subtest 2
 - Outstanding gain among proposals in CE4 Subtest 2
- Problem: Additional buffer memory is needed for temporal QP prediction
 - QP values in reference frames should be stored
- Solution: **QP buffer compression**
 - Buffer compression is also employed for temporal MV prediction
 - Storage requirements for QP values could be significantly reduced

CE4.2.4.b overview (1)

- CE4.2.4.b: QP prediction based on intra/inter prediction information
- Top-left-most CU and PU in QU (Quantization group of coding units) are employed for extracting prediction information

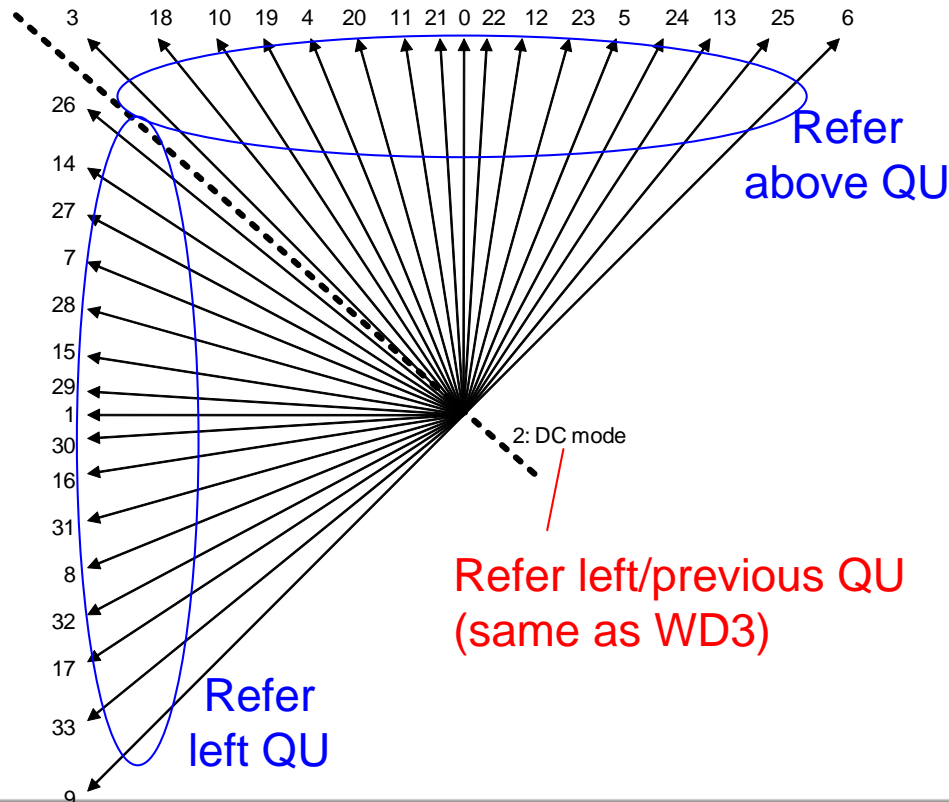


CE4.2.4.b overview (2): QP prediction in intra CU

Spatial QP prediction (SQPP) as in JCTVC-F159 is employed

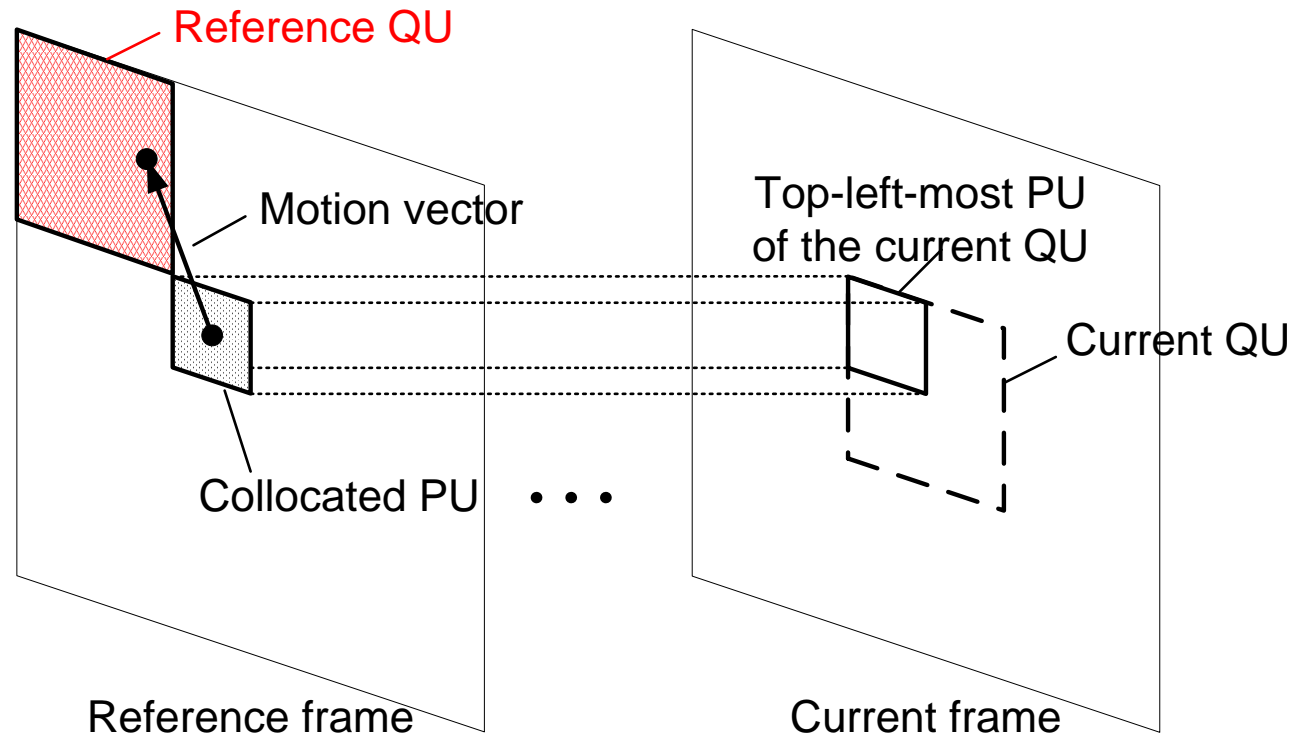
Prediction modes are classified into 3 types

- Vertical: Predicted QP is set equal to the above QP
- Horizontal: Predicted QP is set equal to the left QP
- Others: Default prediction of WD3 is used



CE4.2.4.b overview (3): QP prediction in inter CU

- Temporal QP prediction (TQPP) is employed
- Predicted QP = QP of the center of the reference block for MC



- Slice-level QP offset is used to account for inter-frame QP control

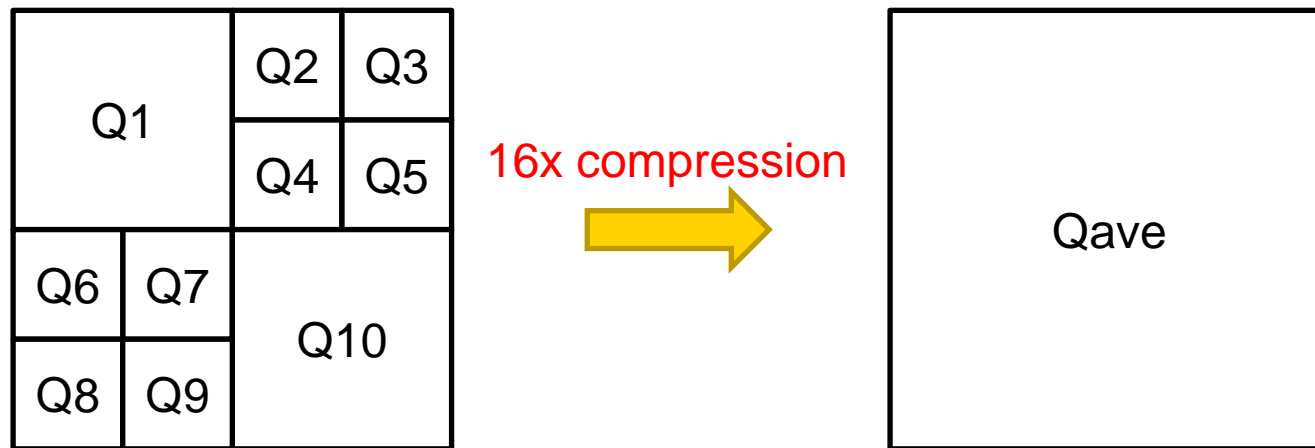
$$Q_{\text{Pred}} = Q(\mathbf{F}_{\text{Ref}}, \mathbf{x}_{\text{Ref}}) + Q_{\text{Slice}}(\mathbf{F}_{\text{Curr}}, \mathbf{x}_{\text{TL}}) - Q_{\text{Slice}}(\mathbf{F}_{\text{Ref}}, \mathbf{x}_{\text{Ref}})$$

QP buffer memory compression

- Stored QP values are calculated by averaging QP values within a block
 - Block size is predefined (fixed or signaled in PPS)

$$\text{RefQP} = \frac{1}{\text{MinCUTQPSize}^2} \left(\left(\sum_{j=0}^{\text{MinCUTQPSize}-1} \sum_{i=0}^{\text{MinCUTQPSize}-1} Q(x_{\text{QRefTl}+i}, y_{\text{QRefTl}+j}) \right) + \text{MinCUTQPSize}^2 / 2 \right)$$

Example: MinCUSize=8x8, MinCUDQPSize=8x8, MinCUTQPSize=32x32



$$\text{Qave} = (4 \cdot \text{Q1} + \text{Q2} + \text{Q3} + \dots + \text{Q9} + 4 \cdot \text{Q10} + 8) / 16$$

Summary of experimental results vs. CE4 anchor

Gain by the CE4.2.4.b method is retained while significantly reducing storage requirement

- BD-rate increases 0.1% with every 4x compression

QP signaling (MinCUDQP)	QP buffer (MinCUTQP)	Y BD-rates				dQP bits			
		RA HE	RA LC	LB HE	LB LC	RA HE	RA LC	LB HE	LB LC
8x8	8x8	-0.6	-0.6	-0.8	-1.0	-18.3%	-18.7%	-20.3%	-21.2%
	16x16	-0.5	-0.5	-0.8	-0.9	-16.0%	-16.2%	-18.1%	-18.4%
	32x32	-0.4	-0.4	-0.7	-0.7	-12.9%	-12.5%	-14.4%	-13.6%
	64x64	-0.3	-0.3	-0.5	-0.4	-9.0%	-7.4%	-10.0%	-7.7%
16x16	16x16	-0.4	-0.5	-0.8	-0.9	-23.0%	-23.0%	-28.2%	-27.8%
	32x32	-0.3	-0.3	-0.6	-0.6	-16.1%	-15.6%	-19.0%	-18.3%
	64x64	-0.2	-0.2	-0.4	-0.4	-10.8%	-9.0%	-12.7%	-10.4%
32x32	32x32	-0.3	-0.3	-0.6	-0.7	-29.2%	-28.8%	-41.2%	-38.6%
	64x64	-0.2	-0.2	-0.3	-0.4	-16.1%	-15.4%	-21.1%	-20.0%
64x64	64x64	-0.2	-0.2	-0.4	-0.4	-37.3%	-36.0%	-57.1%	-51.0%

Summary of experimental results vs. CE4.2.3.g (based on intra prediction)

Proposed compression keeps effectiveness of TQPP

QP signaling (MinCUDQP)	QP buffer (MinCUTQP)	Y BD-rates				dQP bits			
		RA HE	RA LC	LB HE	LB LC	RA HE	RA LC	LB HE	LB LC
8x8	8x8	-0.4	-0.4	-0.7	-0.9	-14.7%	-14.8%	-18.8%	-19.4%
	16x16	-0.4	-0.4	-0.7	-0.8	-12.4%	-12.3%	-16.5%	-16.5%
	32x32	-0.3	-0.2	-0.6	-0.6	-9.1%	-8.1%	-12.7%	-11.6%
	64x64	-0.2	-0.1	-0.4	-0.3	-5.1%	-3.1%	-8.3%	-5.6%
16x16	16x16	-0.4	-0.4	-0.7	-0.8	-20.5%	-20.4%	-27.1%	-26.5%
	32x32	-0.2	-0.3	-0.5	-0.6	-13.4%	-12.7%	-17.8%	-16.9%
	64x64	-0.1	-0.1	-0.4	-0.3	-8.0%	-5.9%	-11.4%	-8.8%
32x32	32x32	-0.3	-0.3	-0.7	-0.7	-27.8%	-27.2%	-40.6%	-37.8%
	64x64	-0.1	-0.1	-0.3	-0.3	-14.6%	-13.6%	-20.4%	-19.1%
64x64	64x64	-0.2	-0.2	-0.5	-0.4	-36.7%	-35.3%	-56.8%	-50.7%

Results vs. CE4 anchor: 8x8 DQP, 16x16 QP buffer

	Random Access HE				Random Access LC			
	Y BD-rate	U BD-rate	V BD-rate	dQP incr.	Y BD-rate	U BD-rate	V BD-rate	dQP incr.
Class A	-0.4	-0.6	-0.4	-14.2%	-0.4	-0.4	-0.5	-14.2%
Class B	-0.5	-0.5	-0.5	-15.9%	-0.5	-0.6	-0.6	-15.4%
Class C	-0.7	-0.7	-0.6	-17.8%	-0.7	-0.8	-0.6	-18.2%
Class D	-0.4	-0.6	-0.5	-16.3%	-0.6	-0.7	-0.6	-17.2%
Class E								
All	-0.5	-0.6	-0.5	-16.0%	-0.5	-0.6	-0.6	-16.2%
Enc Time[%]	100%				100%			
Dec Time[%]	101%				99%			

	Low delay B HE				Low delay B LC			
	Y BD-rate	U BD-rate	V BD-rate	dQP incr.	Y BD-rate	U BD-rate	V BD-rate	dQP incr.
Class A								
Class B	-0.7	-0.8	-0.7	-16.8%	-0.7	-0.7	-0.8	-16.0%
Class C	-0.8	-0.8	-0.7	-18.0%	-0.9	-0.9	-0.9	-18.6%
Class D	-0.7	-0.9	-0.4	-17.2%	-0.9	-0.9	-1.2	-18.1%
Class E	-1.1	0.2	-0.1	-21.4%	-1.3	-1.5	-1.3	-22.4%
All	-0.8	-0.6	-0.5	-18.1%	-0.9	-1.0	-1.0	-18.4%
Enc Time[%]	100%				99%			
Dec Time[%]	97%				99%			

Results vs. CE4.2.3.g: 8x8 DQP, 16x16 QP buffer

	Random Access HE				Random Access LC			
	Y BD-rate	U BD-rate	V BD-rate	dQP incr.	Y BD-rate	U BD-rate	V BD-rate	dQP incr.
Class A	-0.3	-0.4	-0.3	-12.0%	-0.3	-0.4	-0.4	-11.9%
Class B	-0.3	-0.4	-0.4	-12.0%	-0.3	-0.4	-0.4	-11.3%
Class C	-0.4	-0.5	-0.4	-13.0%	-0.4	-0.6	-0.4	-13.0%
Class D	-0.3	-0.6	-0.5	-12.6%	-0.4	-0.5	-0.3	-13.1%
Class E								
All	-0.4	-0.5	-0.4	-12.4%	-0.4	-0.5	-0.4	-12.3%
Enc Time[%]	119%*				116%*			
Dec Time[%]	101%*				106%*			

	Low delay B HE				Low delay B LC			
	Y BD-rate	U BD-rate	V BD-rate	dQP incr.	Y BD-rate	U BD-rate	V BD-rate	dQP incr.
Class A								
Class B	-0.6	-0.8	-0.7	-15.3%	-0.6	-0.7	-0.8	-14.2%
Class C	-0.7	-0.7	-0.6	-15.9%	-0.7	-0.9	-0.6	-15.9%
Class D	-0.6	-0.9	-0.6	-15.7%	-0.7	-0.9	-0.9	-16.4%
Class E	-1.0	-0.8	-1.1	-20.5%	-1.2	-1.5	-1.1	-21.6%
All	-0.7	-0.8	-0.7	-16.5%	-0.8	-1.0	-0.8	-16.5%
Enc Time[%]	119%*				113%*			
Dec Time[%]	98%*				104%*			

* Reference and Tested are run on different platforms

Evaluation of memory requirement for TQPP

Coding gain relative to buffer memory requirement is evaluated and compared to that of temporal MV prediction (TMVP)

- Relative coding gain = BD-rate divided by bits per SCU in reference frame
 - For TMVP: BD-rate divided by 13.75 (= 55/4)
 - For TQPP: BD-rate divided by $6/N^2$, where $N = \text{MinTQPSize}/\text{MinCUSize}$

	TMVP	TQPP 8x8				TQPP 16x16			TQPP 32x32	
	16x16	8x8	16x16	32x32	64x64	16x16	32x32	64x64	32x32	64x64
RA.HE	0.17	0.07	0.23	0.75	1.77	0.24	0.65	1.59	0.44	1.46
RA.LC	0.16	0.07	0.24	0.63	1.06	0.27	0.69	1.21	0.42	1.36
LB.HE	0.17	0.12	0.46	1.50	3.79	0.49	1.42	3.78	1.27	3.71
LB.LC	0.17	0.15	0.51	1.57	3.40	0.56	1.51	3.40	1.14	3.65

When resolution of QP buffer is the same as that of MV buffer, relative coding gain of TQPP is more than that of TMVP

Conclusions

- QP buffer compression scheme for TQPP of CE4.2.4.b is proposed
- Gain is retained while storage requirement is significantly reduced
 - BD-rate roughly increases 0.1% with every 4x compression
 - Gain of the CE4.2.4.b with the proposed compression:
 - AI-HE: 0.4%, AI-LC: 0.4%
 - RA-HE: 0.5%, RA-LC: 0.5%
 - LB-HE: 0.8%, LB-LC: 0.9%(with 8x8 DQP and 16x16 QP buffer)
- Coding gain relative to buffer requirement is more than that of TMVP
- Computational complexity is negligible
- Recommendation:
CE4.2.4.b (intra/inter prediction-based QP prediction) combined with the proposed buffer compression is adopted into the WD4

Empowered by Innovation

NEC

Possible concerns with temporal QP prediction

Temporal QP prediction might not be friendly with rate-control

- As for frame-level rate control, slice-level QP offset accounts for QP variation

Temporal QP prediction still works, in applications with loose buffer constraint

- With CU-level rate control, temporal QP prediction is not friendly

Use of a flag to switch temporal QP prediction on/off would be appropriate

pic_parameter_set_rbsp() {	Descriptor
...	
if(cu_qp_delta_enabled_flag) {	
max_cu_qp_delta_depth	u(4)
temporal_qp_prediction_enabled_flag	u(1)
}	
rbsp_trailing_bits()	
}	

temporal_qp_prediction_enabled_flag equal to 1 specifies that temporal QP prediction is enabled in derivation process for quantisation parameters. **temporal_qp_prediction_enabled_flag** equal to 0 specifies that temporal QP prediction is disabled in derivation process for quantisation parameters. When **temporal_qp_prediction_enabled_flag** is not present, **temporal_qp_prediction_enabled_flag** shall be inferred to be equal to 0.