Lecture 7:

H.264 Video Compression

Visual Computing Systems
CMU 15-769, Fall 2016

Example video

Go Swallows!

30 second video: 1920 x 1080, @ 30fps

After decode: 8-bits per channel RGB → 24 bits/pixel → 6.2MB/frame

(6.2 MB * 30 sec * 30 fps = 5.2 GB)

Size of data when each frames stored as JPG: 531MB

Actual H.264 video file size: 65.4 MB (80-to-1 compression ratio, 8-to-1 compared to JPG)

Compression/encoding performed in real time on my iPhone 5s

H.264/AVC video compression

- AVC = advanced video coding
- Also called MPEG4 Part 10
- Common format in many modern HD video applications:
  - Blue Ray
  - HD streaming video on internet (Youtube, Vimeo, iTunes store, etc.)
  - HD video recorded by your smart phone
  - European broadcast HDTV (U.S. broadcast HDTV uses MPEG 2)
  - Some satellite TV broadcasts (e.g., DirecTV)
- Benefit: much higher compression ratios than MPEG2 or MPEG4
  - Alternatively, higher quality video for fixed bit rate
- Costs: higher decoding complexity, substantially higher encoding cost
  - Idea: trades off more compute for requiring less bandwidth/storage

Hardware implementations

- Support for H.264 video encode/decode is provided by fixed-function hardware on many modern processors (not just mobile devices)

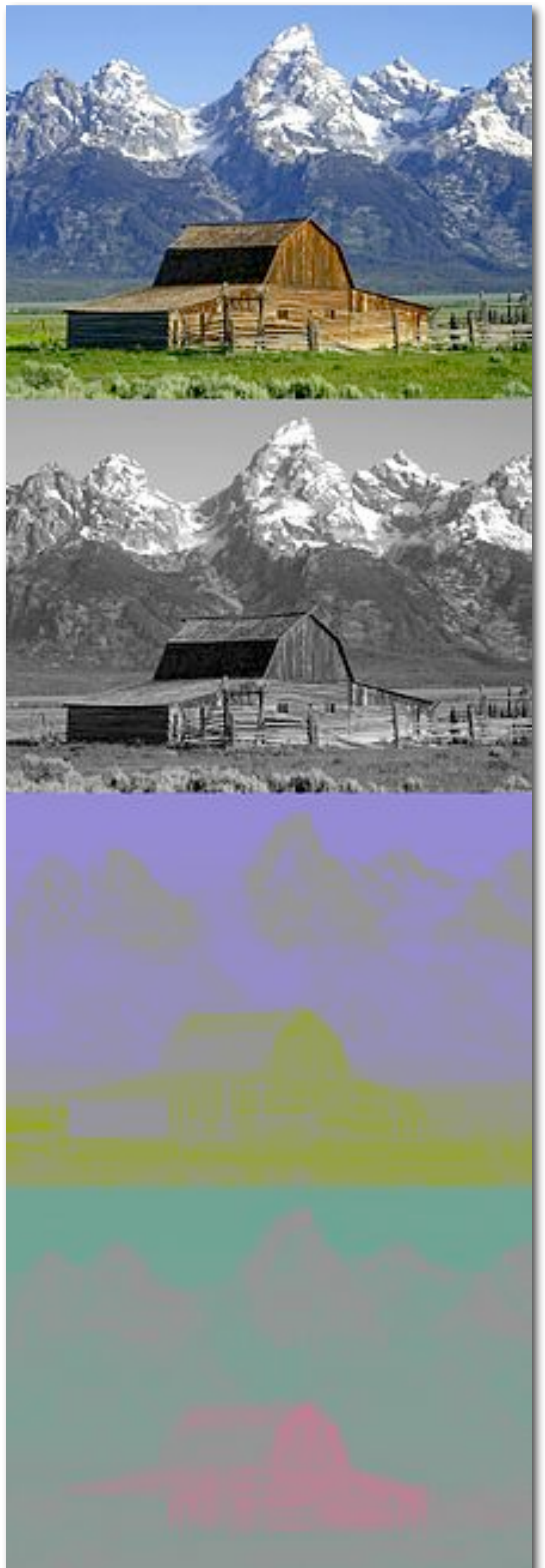
- Hardware encoding/decoding support existed in modern Intel CPUs since Sandy Bridge (Intel "Quick Sync")

- Modern operating systems expose hardware encode decode support through APIs
  - e.g., DirectShow/DirectX (Windows), AVFoundation (iOS)

Video container format versus video codec

- Video container (MOV, AVI) bundles media assets
- Video codec: H.264/AVC (MPEG 4 Part 10)
  - H.264 standard defines how to represent and decode video
  - H.264 does not define how to encode video (this is left up to implementations)
  - H.264 has many profiles
    - High Profile (HiP): supported by HDV and Blue Ray

Review: Y’CbCr 4:2:0



Y’

Cb

Cr

Y’ = perceived brightness (“luma”)

Cb = blue-yellow deviation from gray

Cr = red-cyan deviation from gray

4:2:0 representation (subsampled chroma):

- Store Y’ at full resolution
- Store Cb, Cr at half vertical and horizontal resolution (1/4 as many chroma samples as luminance samples)

Y'_{00} Cb_{00} Cr_{00}	Y'_{10}	Y'_{20} Cb_{20} Cr_{20}	Y'_{30}
Y'_{01}	Y'_{11}	Y'_{21}	Y'_{31}

X:Y:Z notation:

X = width of block

Y = number of chroma samples in first row

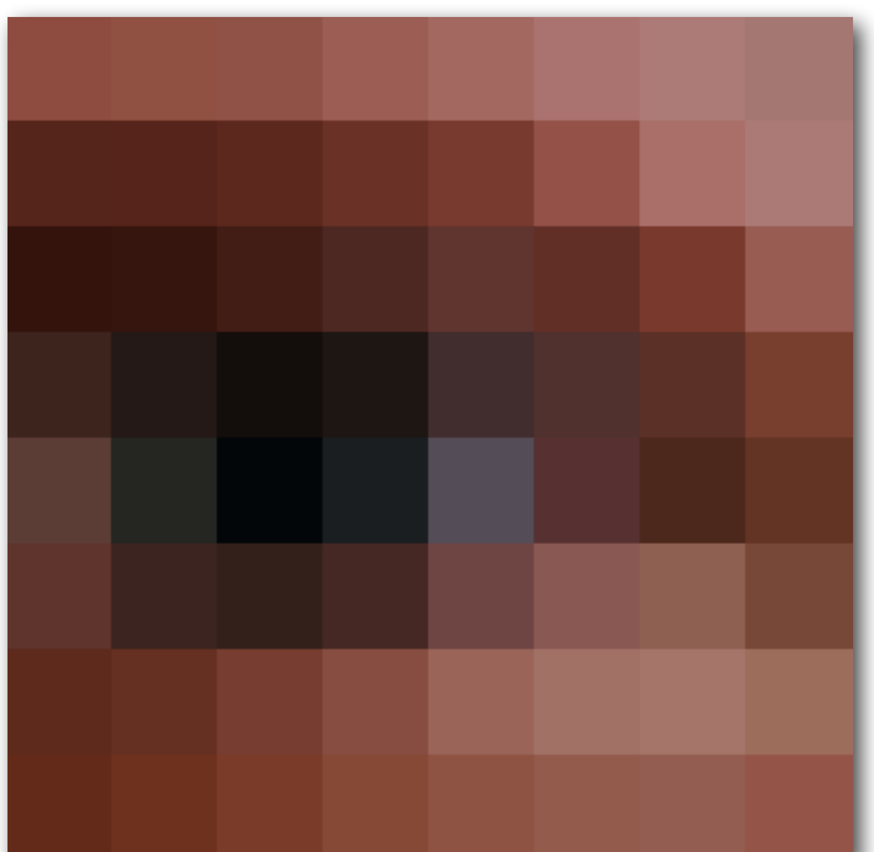
Z = number of chroma samples in second row

Image credit: Wikipedia

Review: image transform coding via DCT

(recall: JPEG compression segment of camera pipeline lecture)

8x8 pixel block
(64 coefficients of signal in "pixel basis")

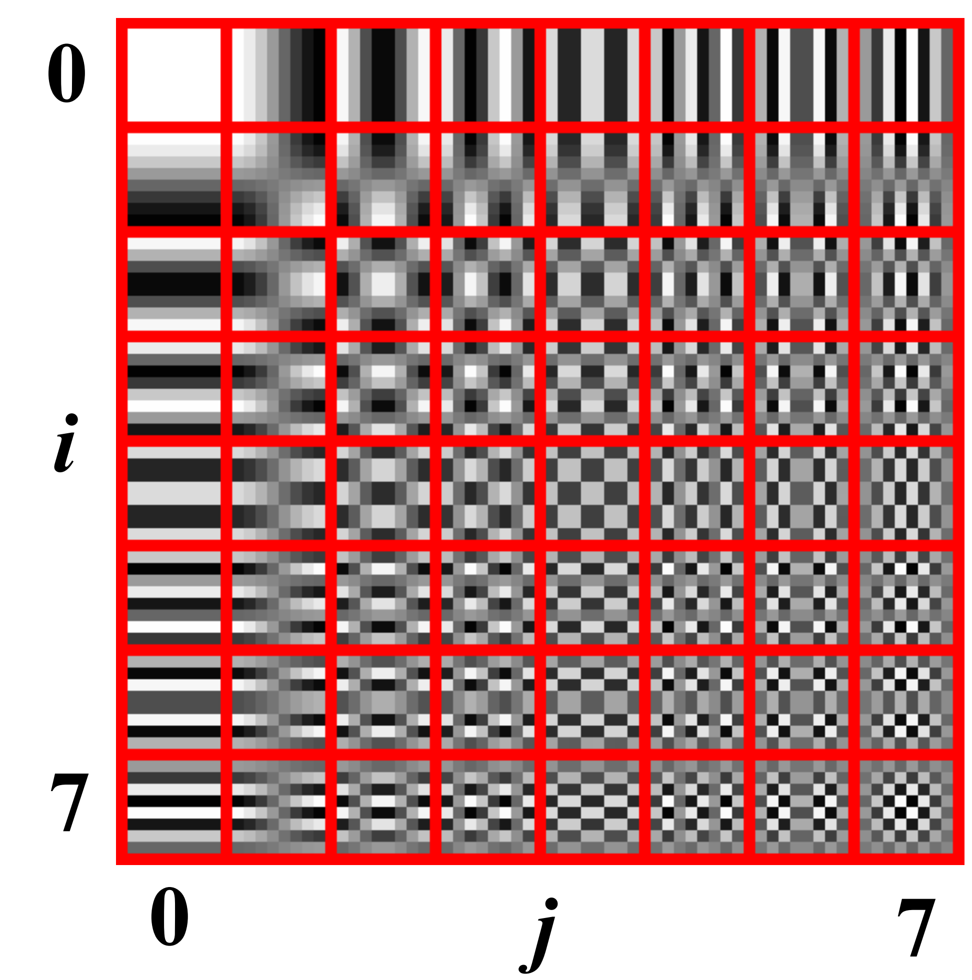


64 basis coefficients

$$
= \begin{bmatrix}
-415 & -30 & -61 & 27 & 56 & -20 & -2 & 0 \\
4 & -22 & -61 & 10 & 13 & -7 & -9 & 5 \\
-47 & 7 & 77 & -25 & -29 & 10 & 5 & -6 \\
-49 & 12 & 34 & -15 & -10 & 6 & 2 & 2 \\
12 & -7 & -13 & -4 & -2 & 2 & -3 & 3 \\
-8 & 3 & 2 & -6 & -2 & 1 & 4 & 2 \\
-1 & 0 & 0 & -2 & -1 & -3 & 4 & -1 \\
0 & 0 & -1 & -4 & -1 & 0 & 1 & 2
\end{bmatrix} \times
$$

64 cosine basis vectors
(each vector is 8x8 image)

$$\cos\left[\pi \frac{i}{N}\left(x + \frac{1}{2}\right)\right] \times \cos_{j}\left[\pi \frac{j}{N}\left(y + \frac{1}{2}\right)\right]$$



Note: only showing coefficients for one channel (e.g., Y') here. Each channel is transformed independently.

Credit: Wikipedia, Pat Hanrahan

Review: quantization and entropy compression

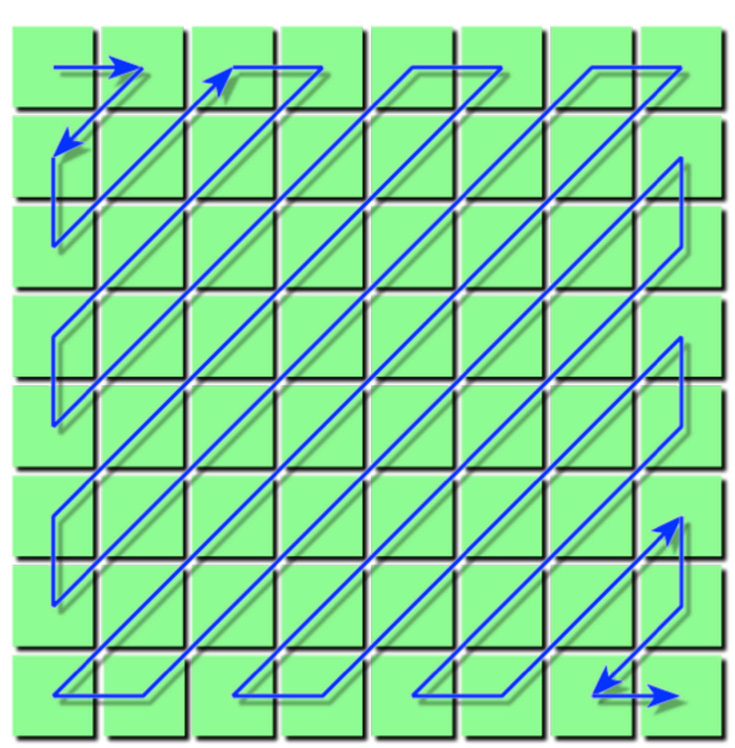
$$\begin{bmatrix} -415 & -30 & -61 & 27 & 56 & -20 & -2 & 0 \\ 4 & -22 & -61 & 10 & 13 & -7 & -9 & 5 \\ -47 & 7 & 77 & -25 & -29 & 10 & 5 & -6 \\ -49 & 12 & 34 & -15 & -10 & 6 & 2 & 2 \\ 12 & -7 & -13 & -4 & -2 & 2 & -3 & 3 \\ -8 & 3 & 2 & -6 & -2 & 1 & 4 & 2 \\ -1 & 0 & 0 & -2 & -1 & -3 & 4 & -1 \\ 0 & 0 & -1 & -4 & -1 & 0 & 1 & 2 \end{bmatrix} \; / \; \begin{bmatrix} 16 & 11 & 10 & 16 & 24 & 40 & 51 & 61 \\ 12 & 12 & 14 & 19 & 26 & 58 & 60 & 55 \\ 14 & 13 & 16 & 24 & 40 & 57 & 69 & 56 \\ 14 & 17 & 22 & 29 & 51 & 87 & 80 & 62 \\ 18 & 22 & 37 & 56 & 68 & 109 & 103 & 77 \\ 24 & 35 & 55 & 64 & 81 & 104 & 113 & 92 \\ 49 & 64 & 78 & 87 & 103 & 121 & 120 & 101 \\ 72 & 92 & 95 & 98 & 112 & 100 & 103 & 99 \end{bmatrix}$$

DCT / **Quantization Matrix**

$$= \begin{bmatrix} -26 & -3 & -6 & 2 & 2 & -1 & 0 & 0 \\ 0 & -2 & -4 & 1 & 1 & 0 & 0 & 0 \\ -3 & 1 & 5 & -1 & -1 & 0 & 0 & 0 \\ -4 & 1 & 2 & -1 & 0 & 0 & 0 & 0 \\ 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \end{bmatrix}$$

Quantized DCT

Quantization loses information (lossy compression!)



Coefficient reordering

RLE compression of zeros

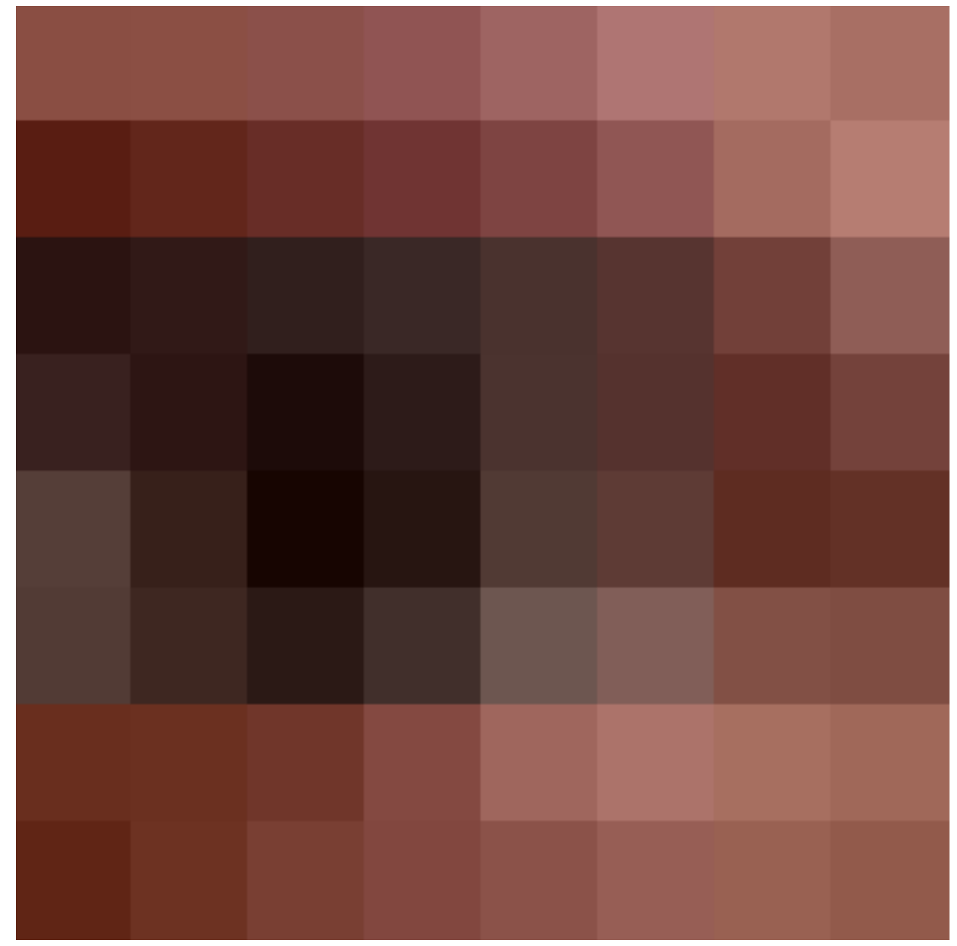
Entropy compression of non-zeros

Lossless compression!

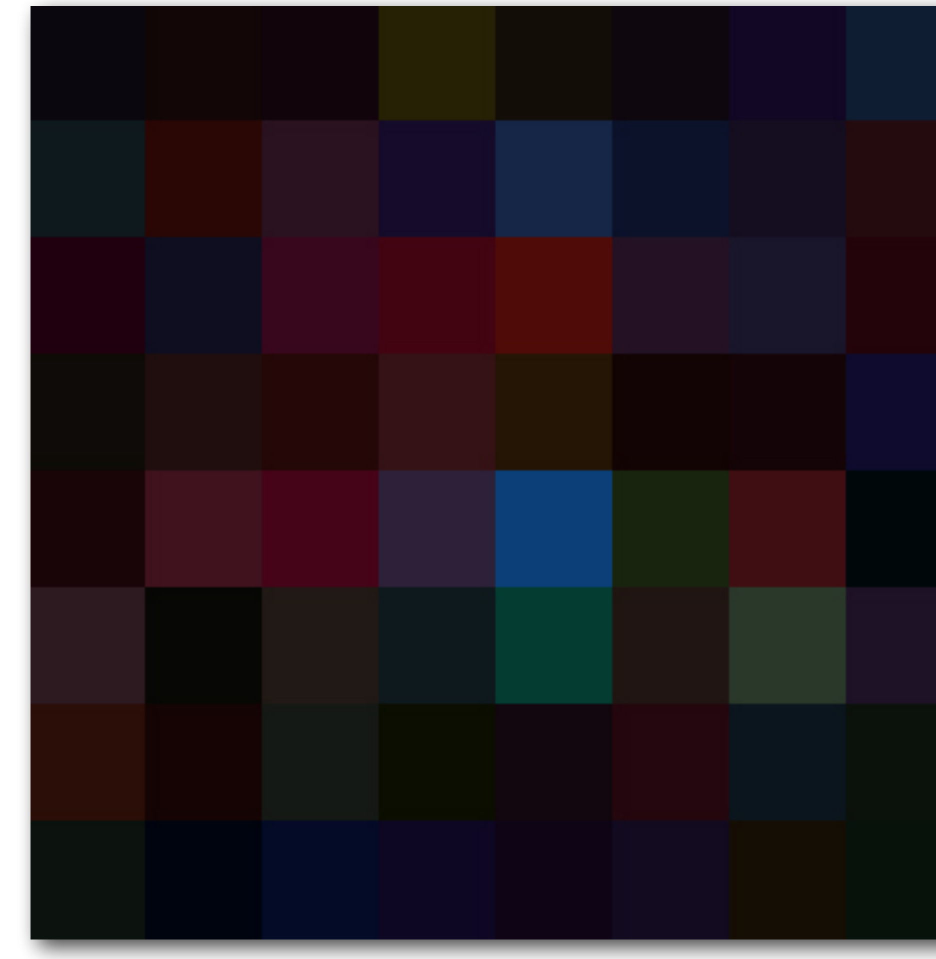
Compressed bits

Residual: difference between compressed image and original image

Original pixels

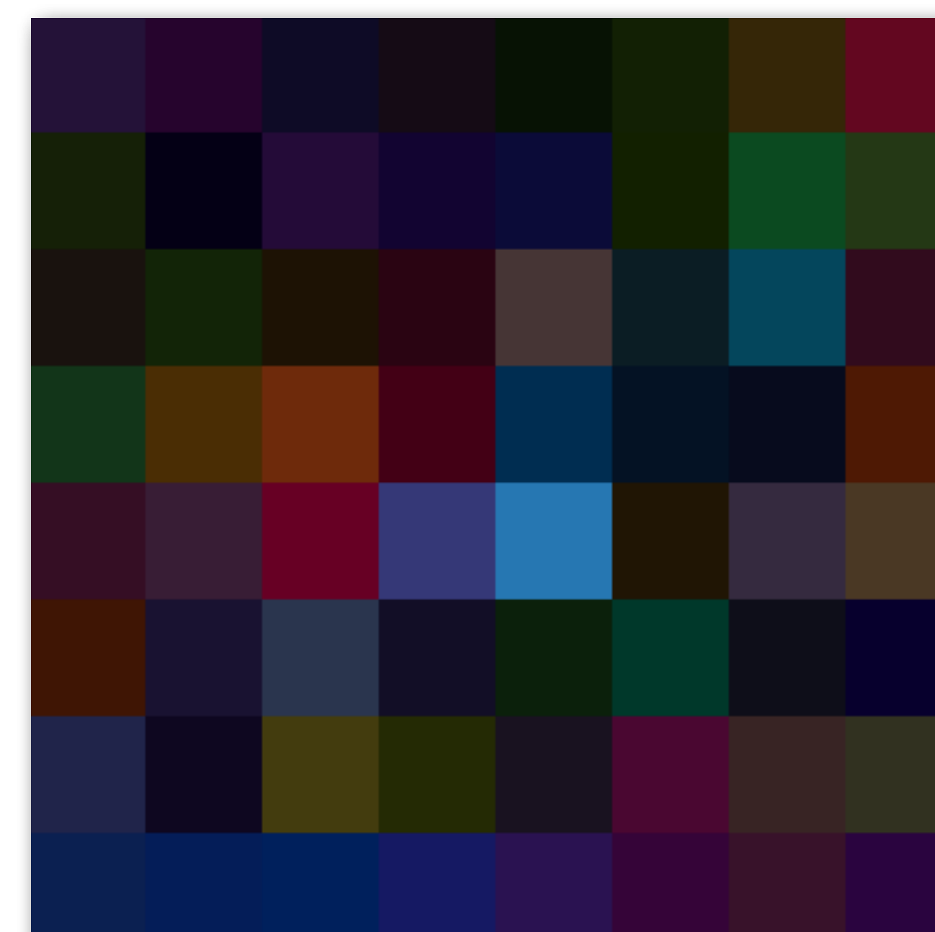


Compressed pixels
(JPEG quality level 6)



Residual
(amplified for visualization)

Compressed pixels
(JPEG quality level 2)

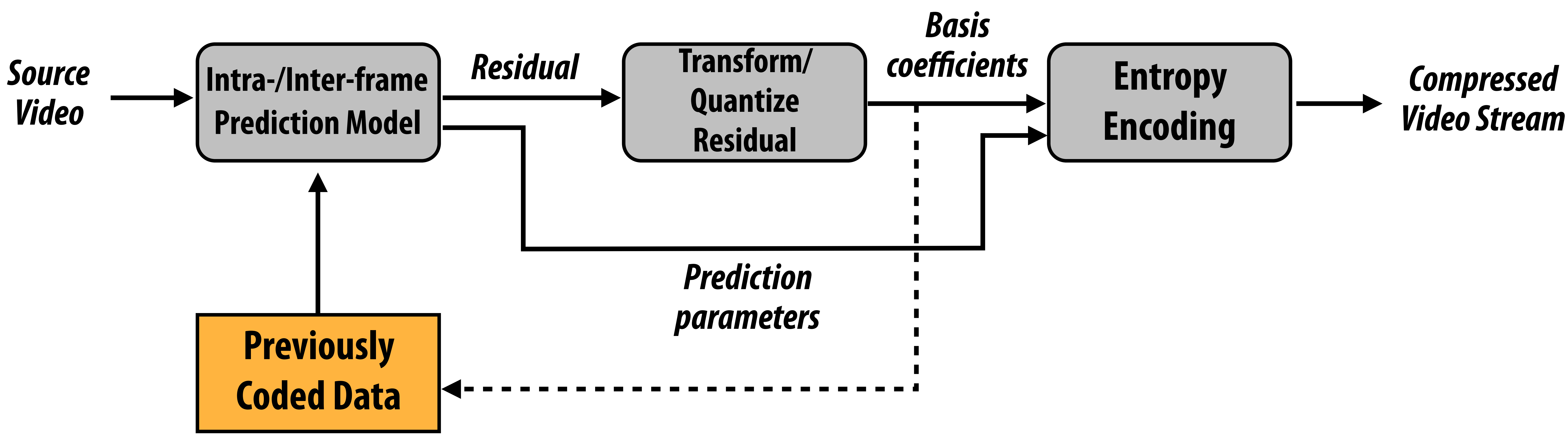


Residual
(amplified for visualization)

Video compression: main ideas

- Compression is about exploiting redundancy in a signal
  - Intra-frame redundancy: value of pixels in neighboring regions of a frame are good <u>predictor</u> of values for other pixels in the frame (spatial redundancy)
  - Inter-frame redundancy: pixels from nearby frames in time are a good <u>predictor</u> for the current frame's pixels (temporal redundancy)

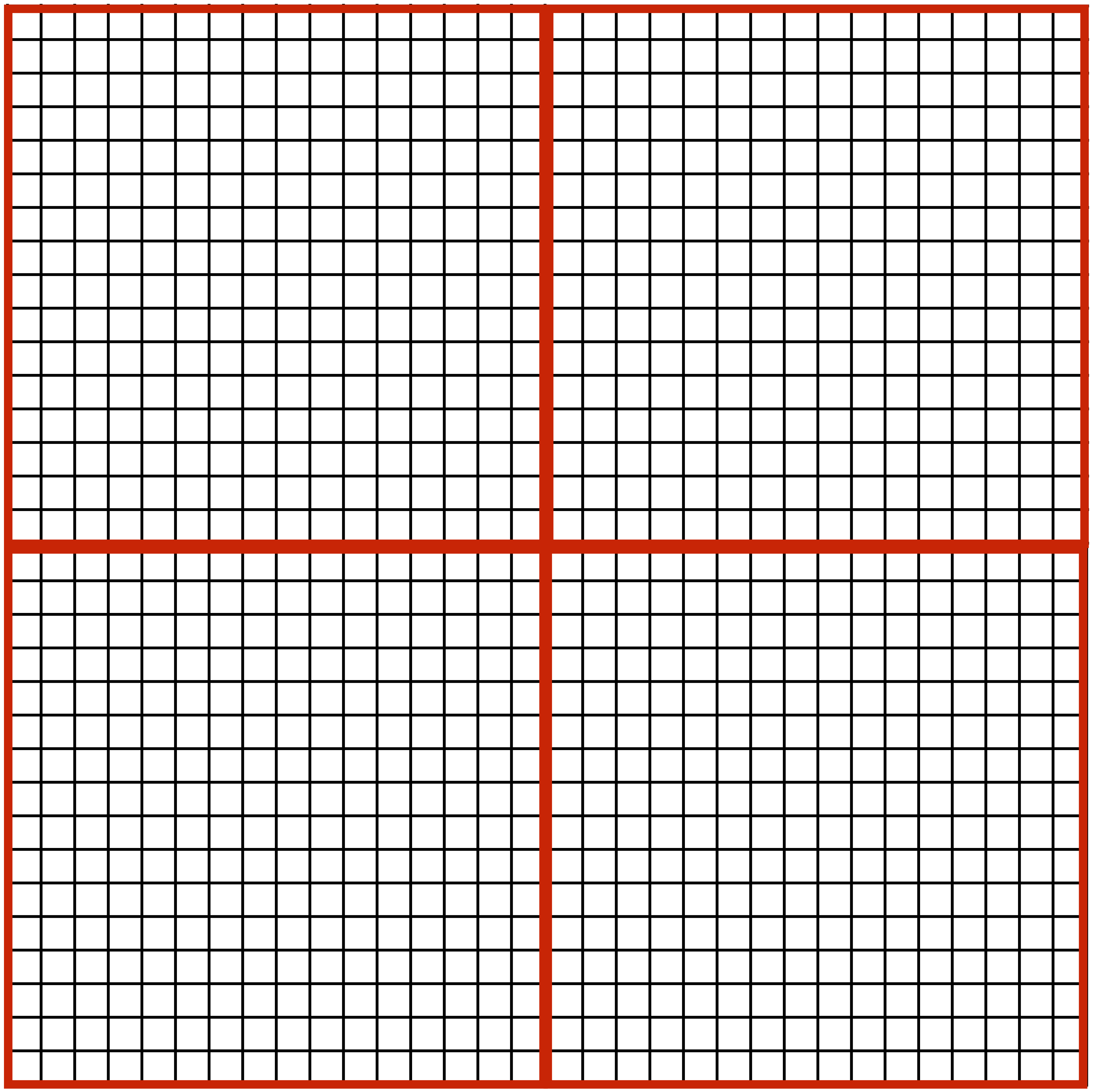
H.264/AVC video compression overview



Residual: difference between predicted pixel values and input video pixel values

Credit: Figure derived from H.264 Advanced Video Compression Standard, I. Richardson, 2010

16 x 16 macroblocks



Video frame is partitioned into 16 x 16 pixel macroblocks

Due to 4:2:0 chroma subsampling, macroblocks correspond to 16 x 16 luma samples and 8 x 8 chroma samples

Macroblocks in an image are organized into slices

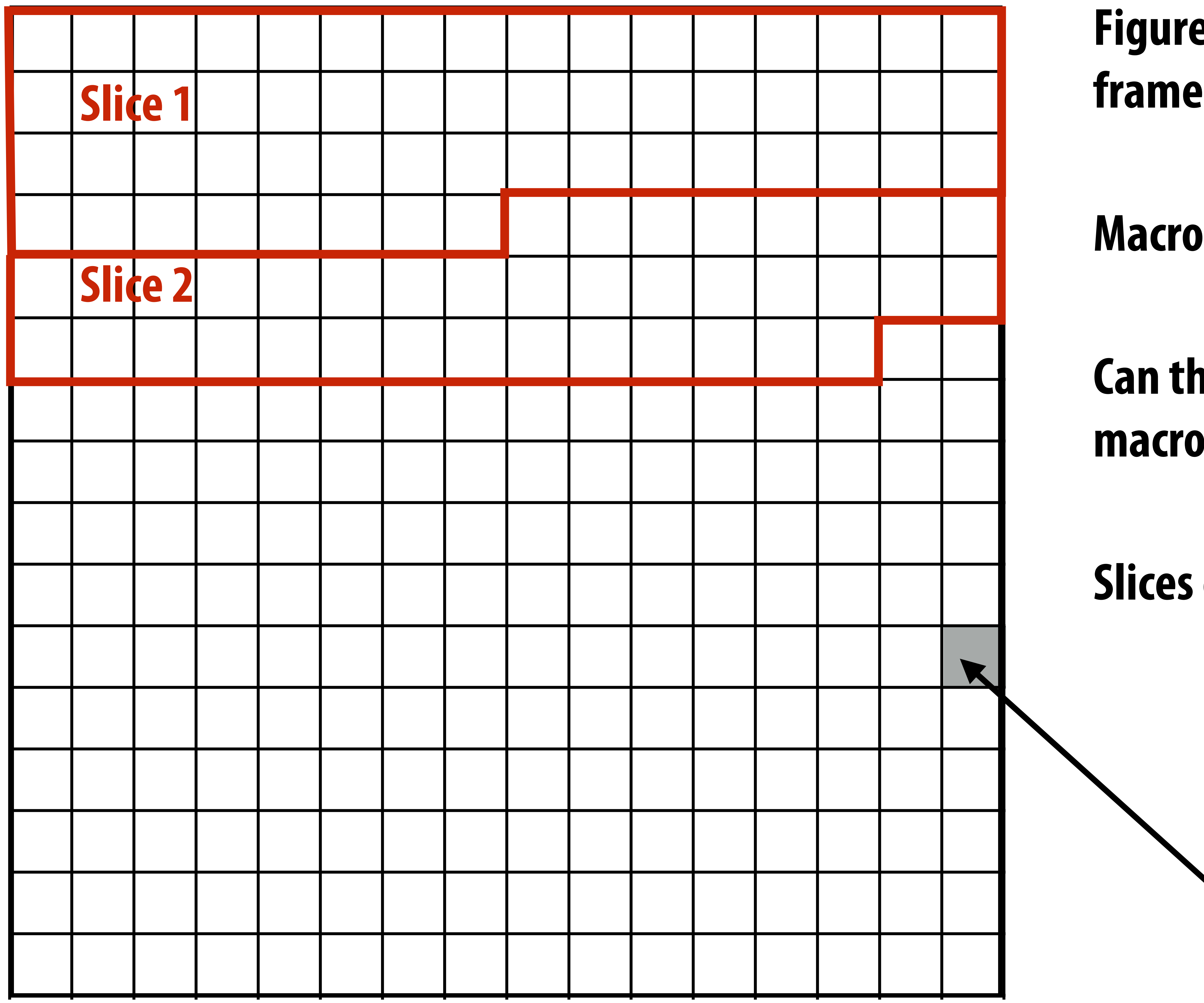


Figure to left shows the macro blocks in a frame

Macroblocks are grouped into slices

Can think of a slice as a sequence of macroblocks in raster scan order *

Slices can be decoded independently **

* H.264 also has non-raster-scan order modes (FMO), will not discuss today.

** Final "deblocking" pass is often applied to post-decode pixel data, so technically slices are not fully independent.

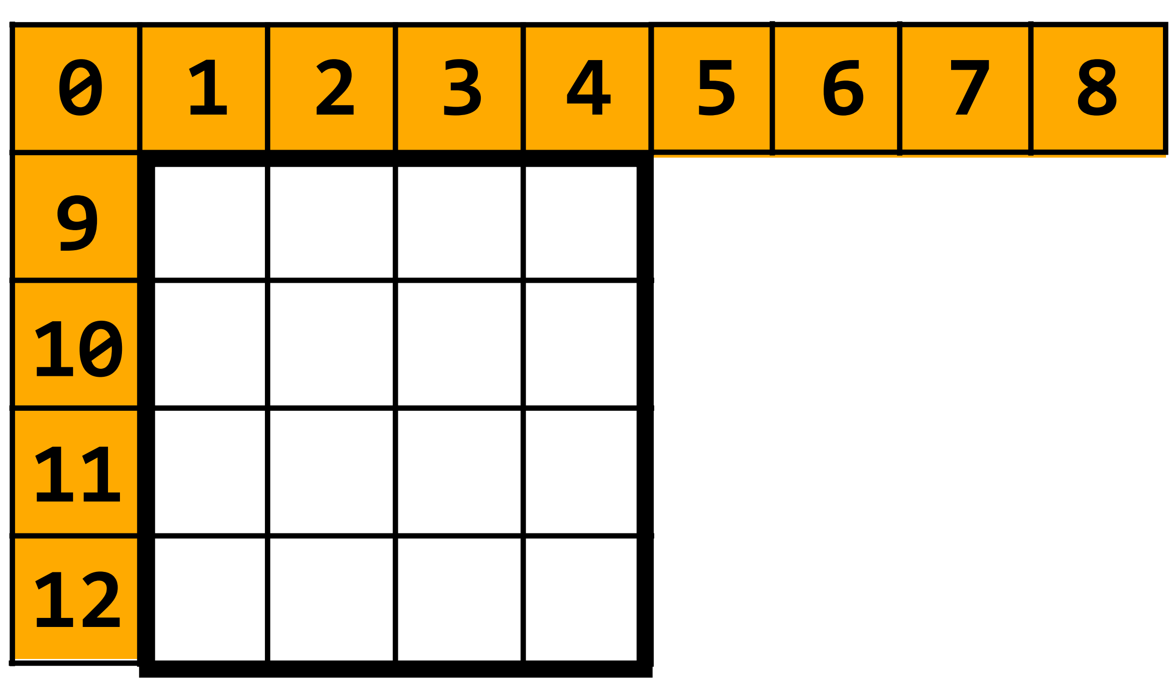
Decoding via prediction + correction

- **During decode, samples in a macroblock are generated by:**
  1. **Making a prediction based on already decoded samples in macroblocks from the same frame (<u>intra-frame prediction</u>) or from other frames (<u>inter-frame prediction</u>)**
  2. **Correcting the prediction with a residual stored in the video stream**

- **Three forms of prediction:**
  - **<u>I-macroblock</u>: macroblock samples predicted from samples in previous macroblocks in the same slice of the current frame**
  - **<u>P-macroblock</u>: macroblock samples can be predicted from samples from one other frame (one prediction per macroblock)**
  - **<u>B-macroblock</u>: macroblock samples can be predicted by a weighted combination of multiple predictions from samples from other frames**

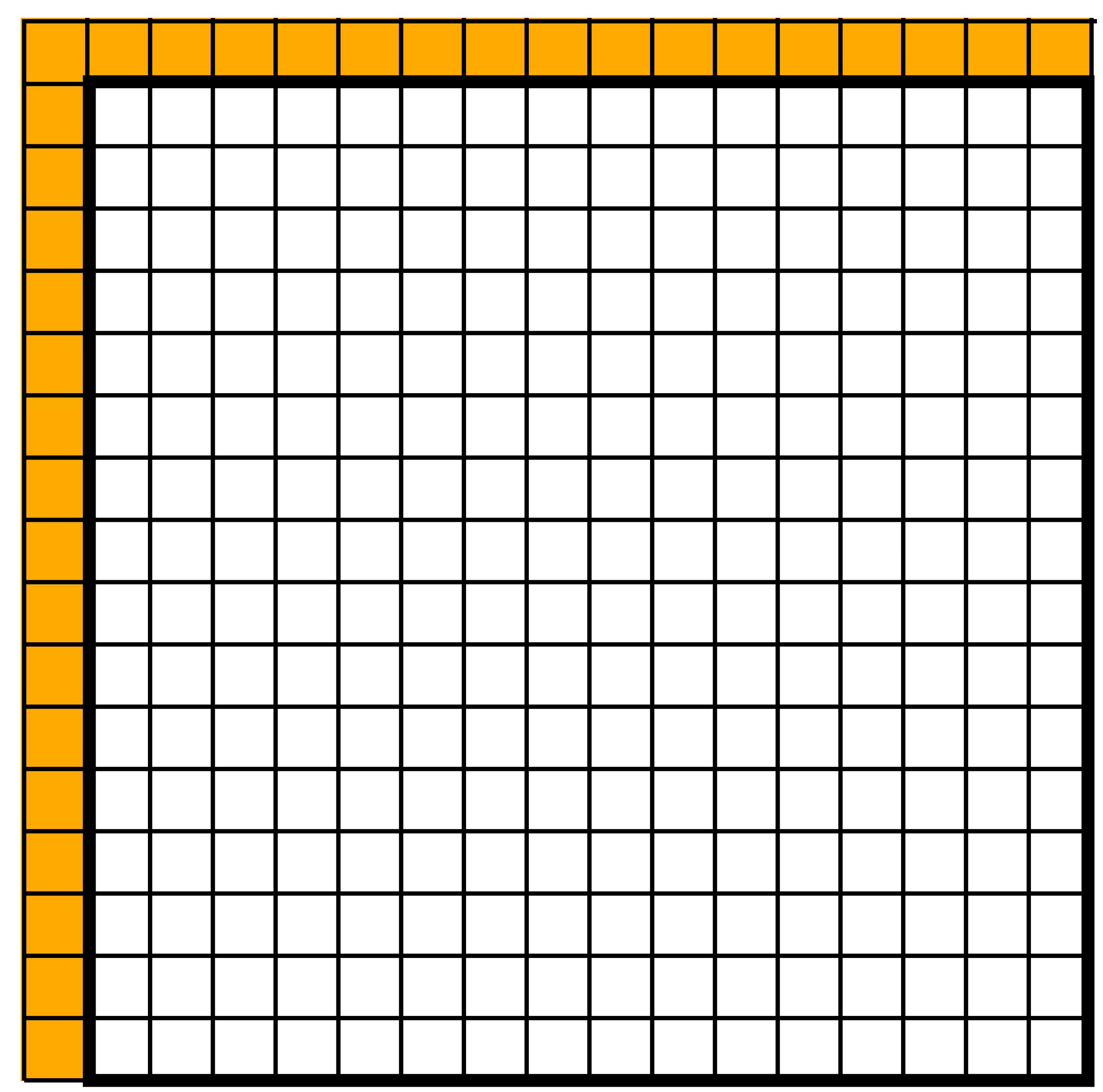
Intra-frame prediction (I-macroblock)

- Prediction of sample values is performed in spatial domain, not transform domain
  - Predicting pixel values, not basis coefficients
- Modes for predicting the 16x16 luma (Y) values: *
  - Intra_4x4 mode: predict 4x4 block of samples from adjacent row/col of pixels
  - Intra_16x16 mode: predict entire 16x16 block of pixels from adjacent row/col
  - I_PCM: actual sample values provided



Intra_4X4

Yellow pixels: already reconstructed (values known)
White pixels: 4x4 block to be reconstructed

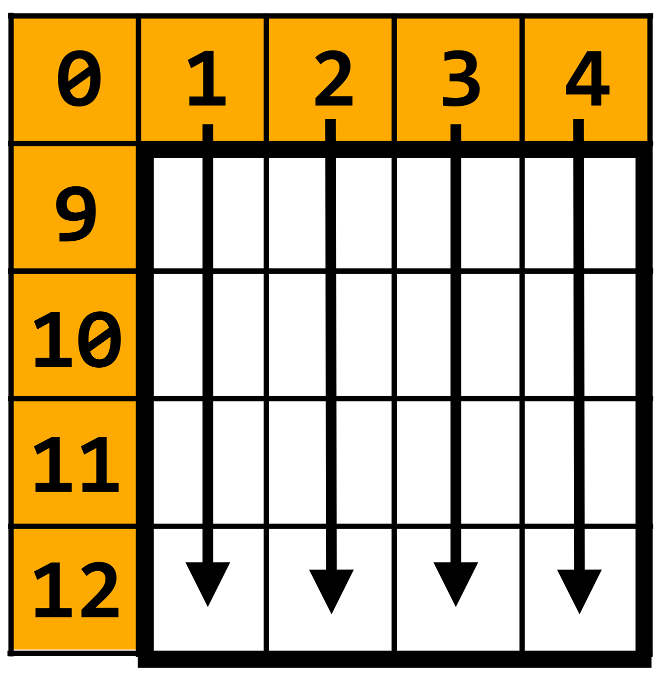


Intra_16x16

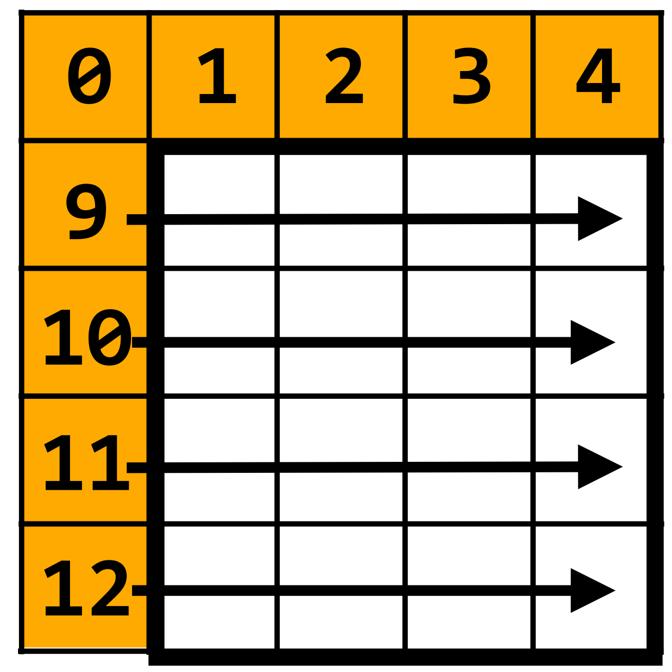
* An additional 8x8 mode exists in the H.264 High Profile

Intra_4x4 prediction modes

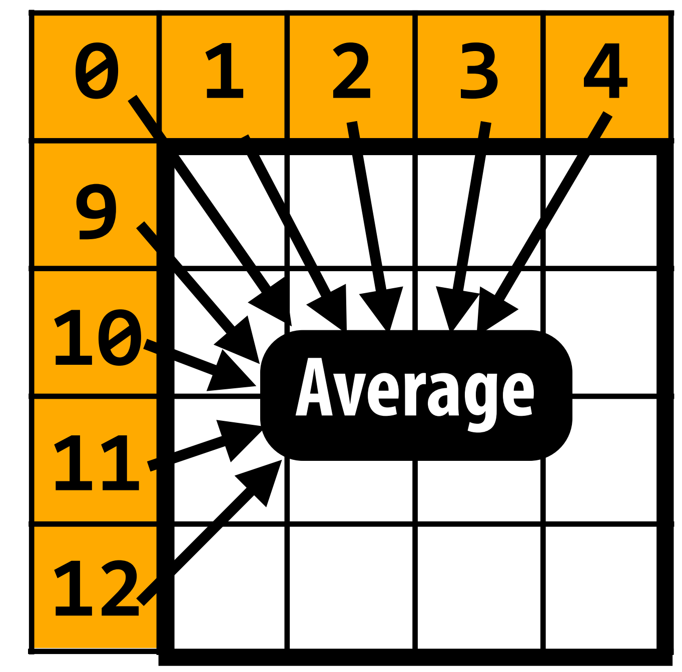
- Nine prediction modes (6 shown below)
  - Other modes: horiz-down, vertical-left, horiz-up



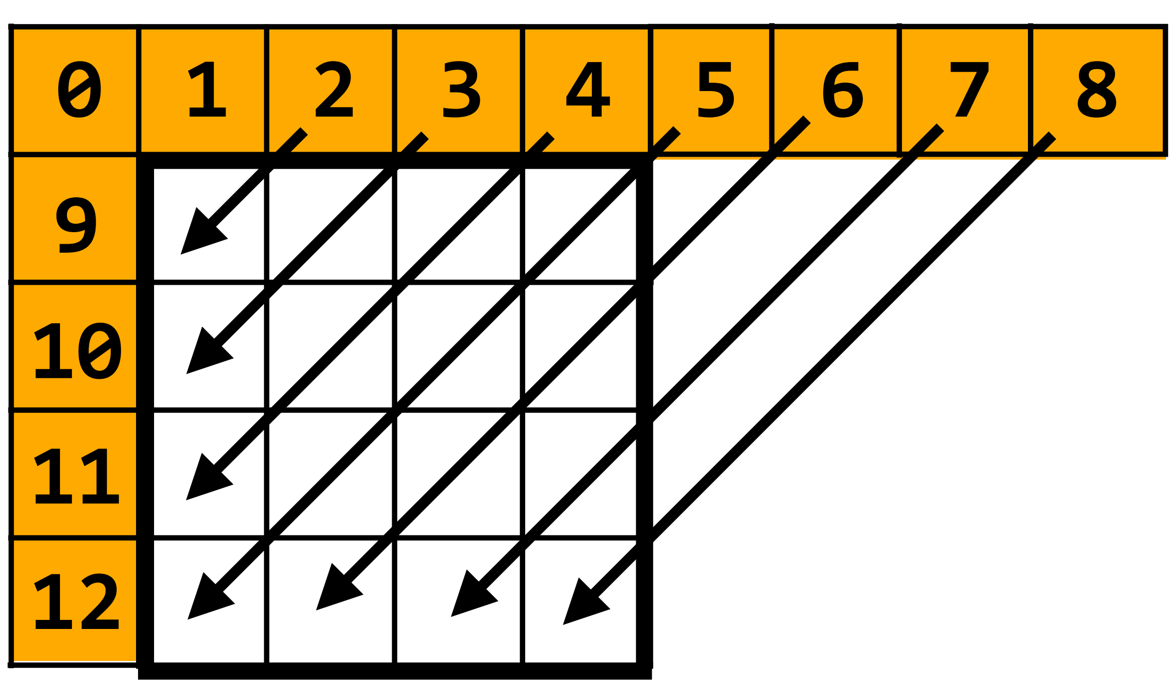
Mode 0: vertical
(4x4 block is copy of above row of pixels)



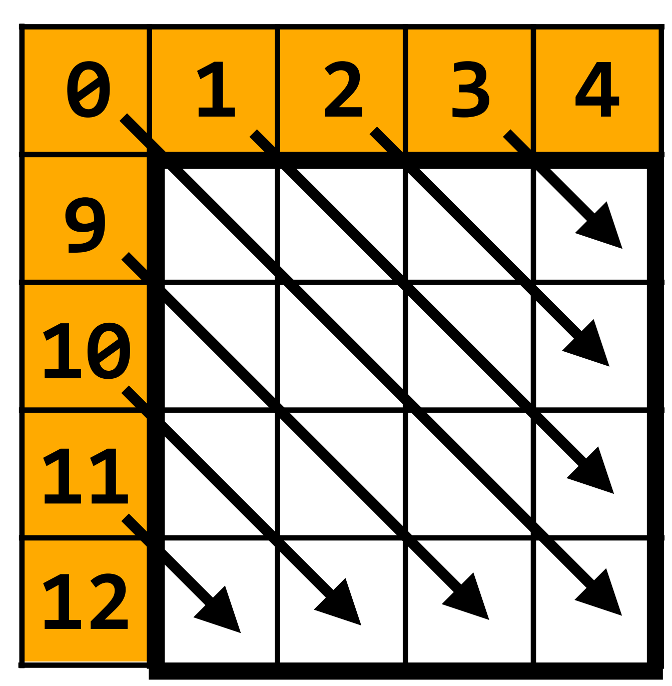
Mode 1: horizontal
(4x4 block is copy of left col of pixels)



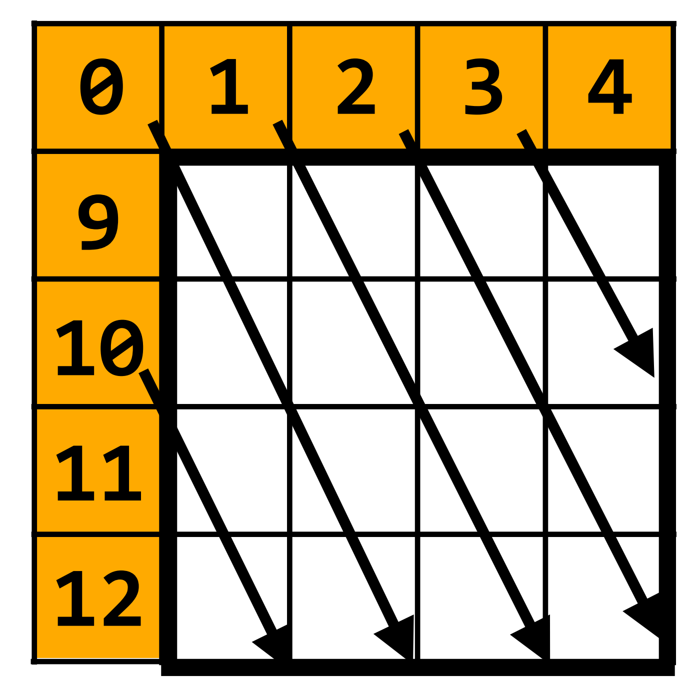
Mode 2: DC
(4x4 block is average of above row and left col of pixels)



Mode 3: diagonal down-left (45°)



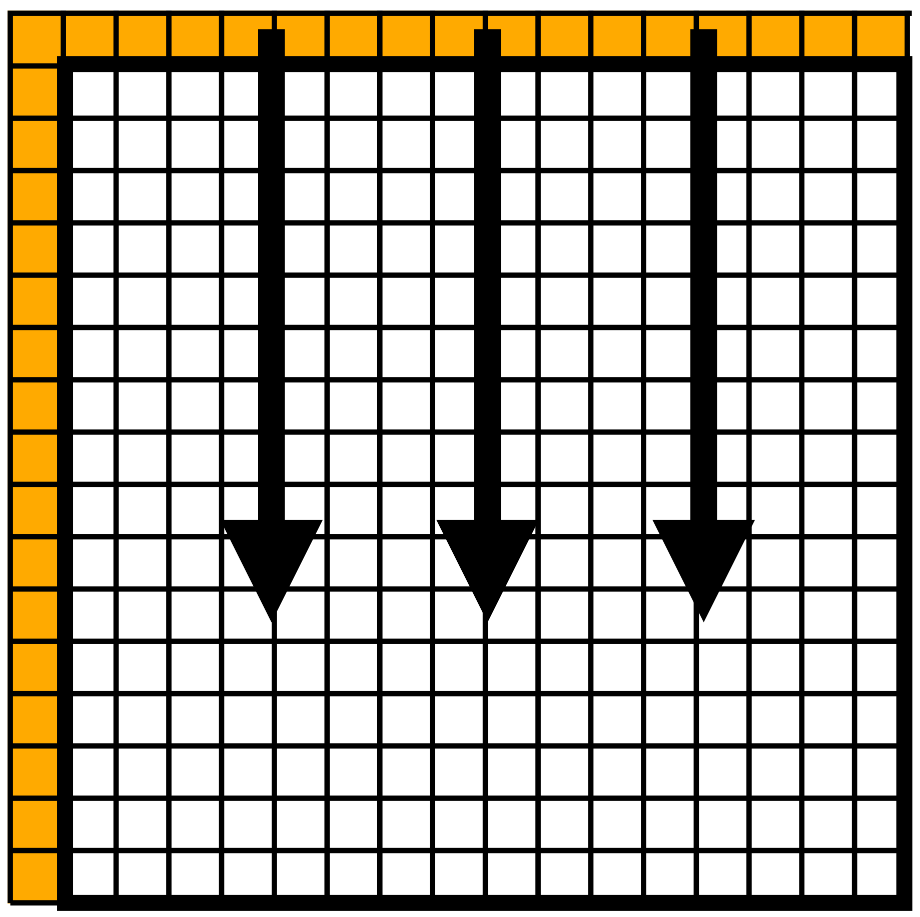
Mode 4: diagonal down-right (45°)



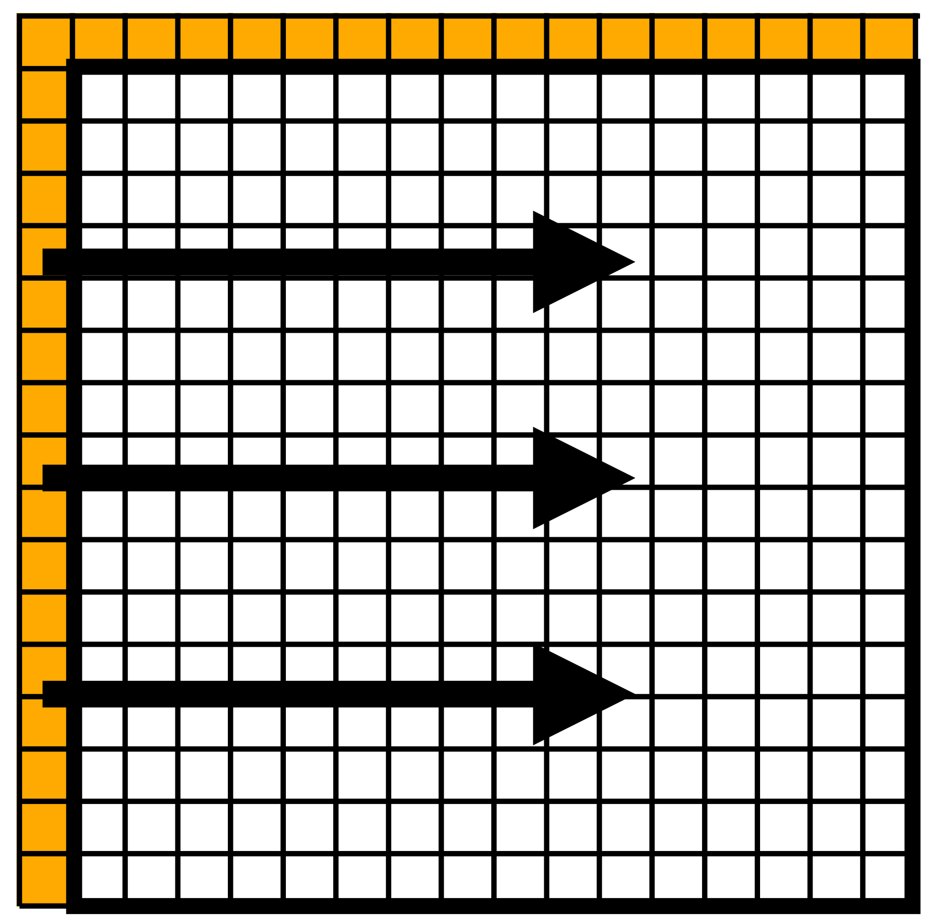
Mode 5: vertical-right (26.6°)

Intra_16x16 prediction modes

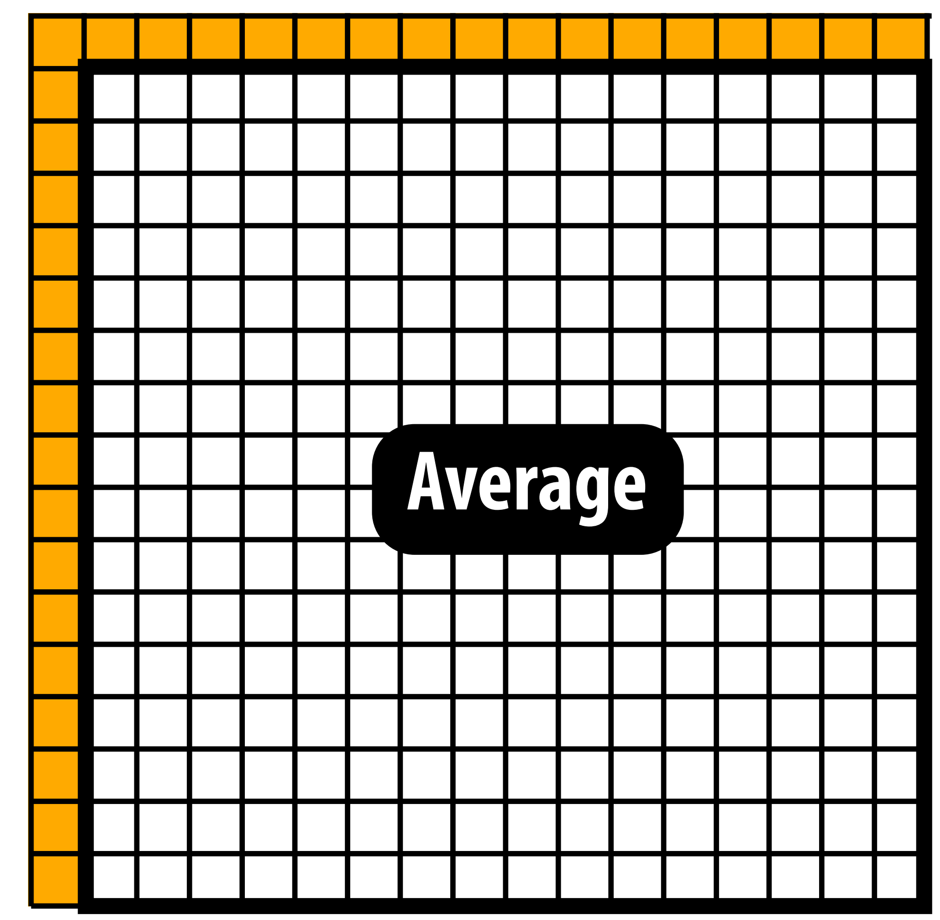
- **4 prediction modes: vertical, horizontal, DC, plane**



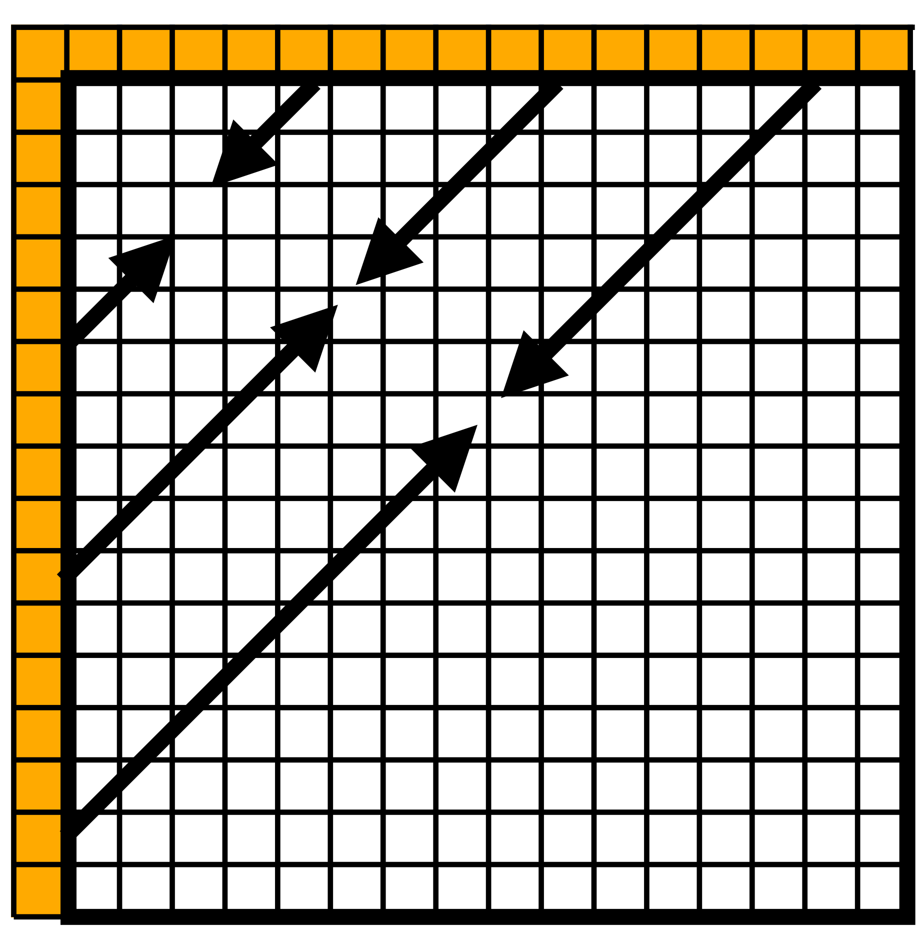
Mode 0: vertical



Mode 1: horizontal



Mode 2: DC



Mode 4: plane

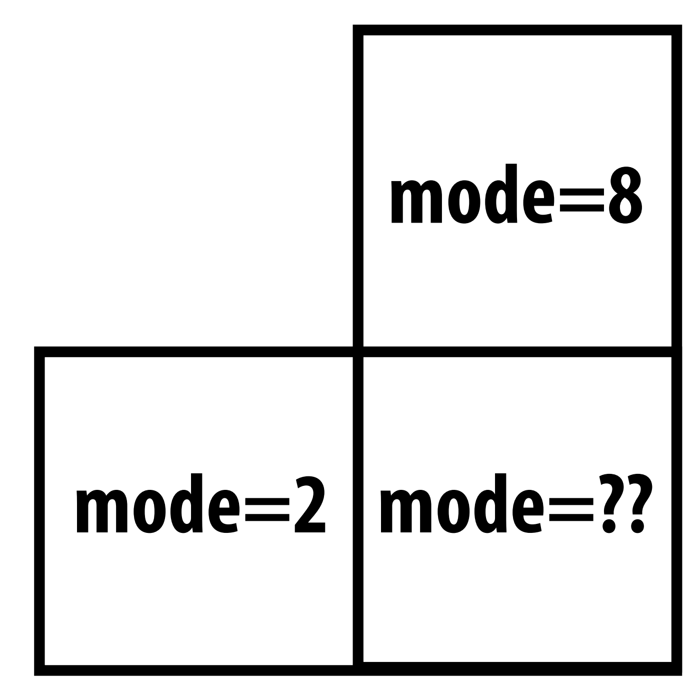
P[i,j] = Ai * Bj + C

A derived from top row, B derived from left col, C from both

Further details

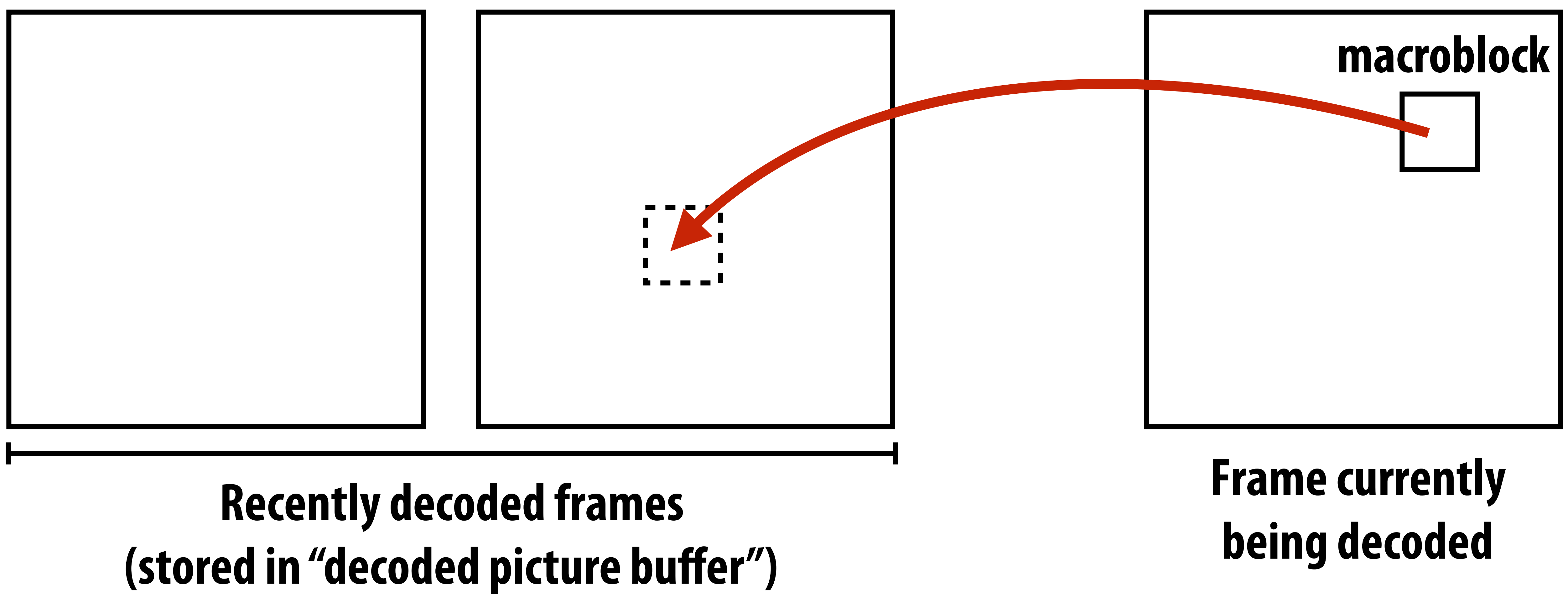
- Intra-prediction of chroma (8x8 block) is performed using four modes similar to those of intra_16x16 (except reordered as: DC, vertical, horizontal, plane)

- Intra-prediction scheme for each 4x4 block within macroblock encoded as follows:
  - One bit per 4x4 block:
    - if 1, use <u>most probable</u> mode
      - Most probable = lower of modes used for 4x4 block to left or above current
    - if 0, use additional 3-bit value `rem_intra4x4_pred_mode` to encode one of nine modes
      - if `rem_intra4x4_pred_mode` is smaller than most probable mode, use mode given by `rem_intra4x4_pred_mode`
      - else, mode is `rem_intra4x4_pred_mode+1`



Inter-frame prediction (P-macroblock)

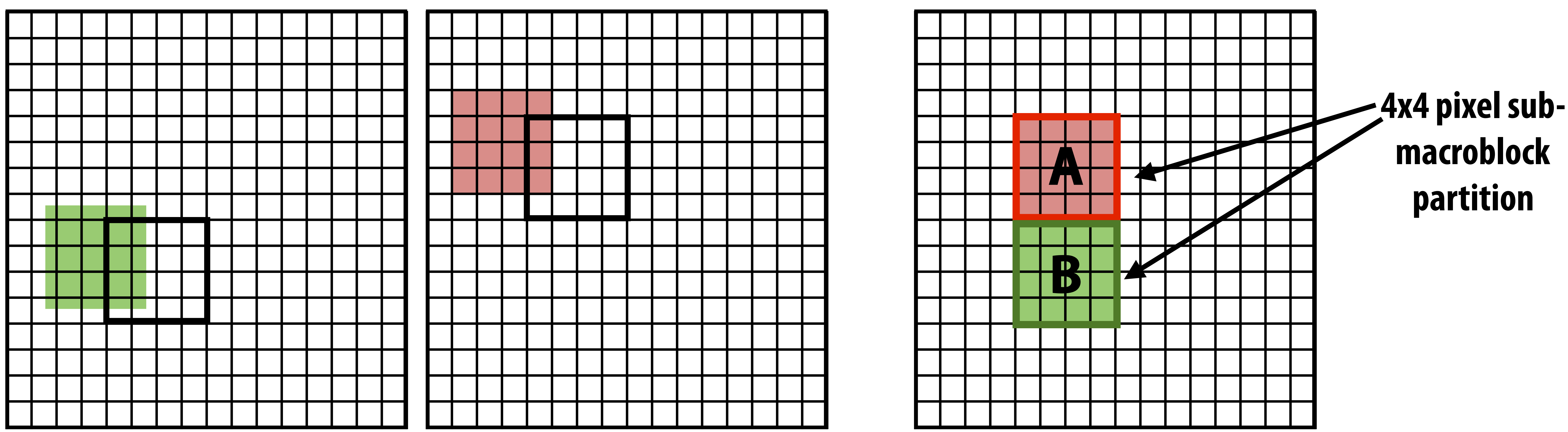
- Predict sample values using values from a block of a previously decoded frame *
- Basic idea: current frame formed by translation of pixels from temporally nearby frames (e.g., object moved slightly on screen between frames)
  - "Motion compensation": use of spatial displacement to make prediction about pixel values



* Note: "previously decoded" does not imply source frame must come before frame in the video sequence. (H.264 supports decoding out of order.)

P-macroblock prediction

- Prediction can be performed at macroblock or sub-macroblock granularity
  - Macroblock can be divided into 16x16, 8x16, 16x8, 8x8 “partitions”
  - 8x8 partitions can be further subdivided into 4x8, 8x4, 4x4 sub-macroblock partitions
- Each partition predicted by sample values defined by:
  (reference frame id, motion vector)



Decoded picture buffer: frame 1

Decoded picture buffer: frame 0

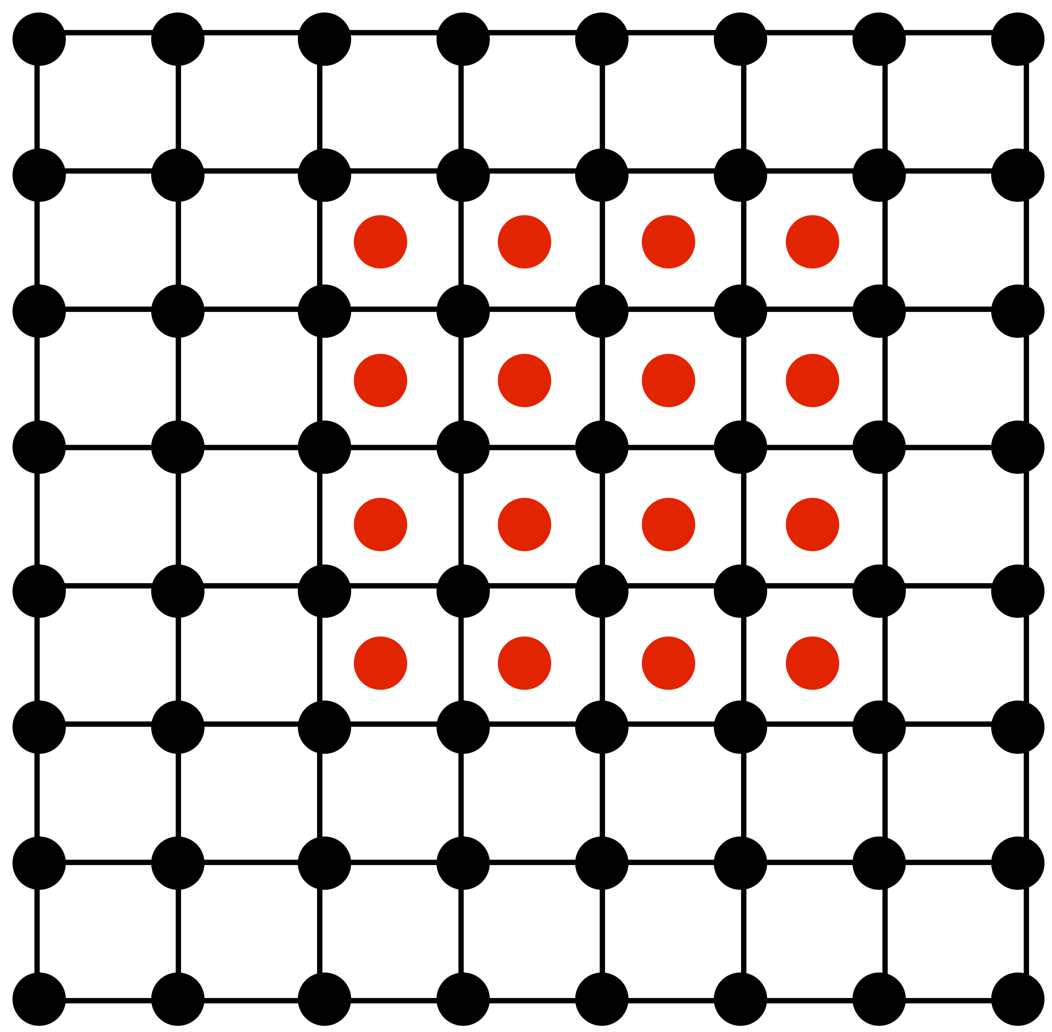
Current frame

Block A: predicted from (frame 0, motion-vector = [-3, -1])

Block B: predicted from (frame 1, motion-vector = [-2.5, -0.5])

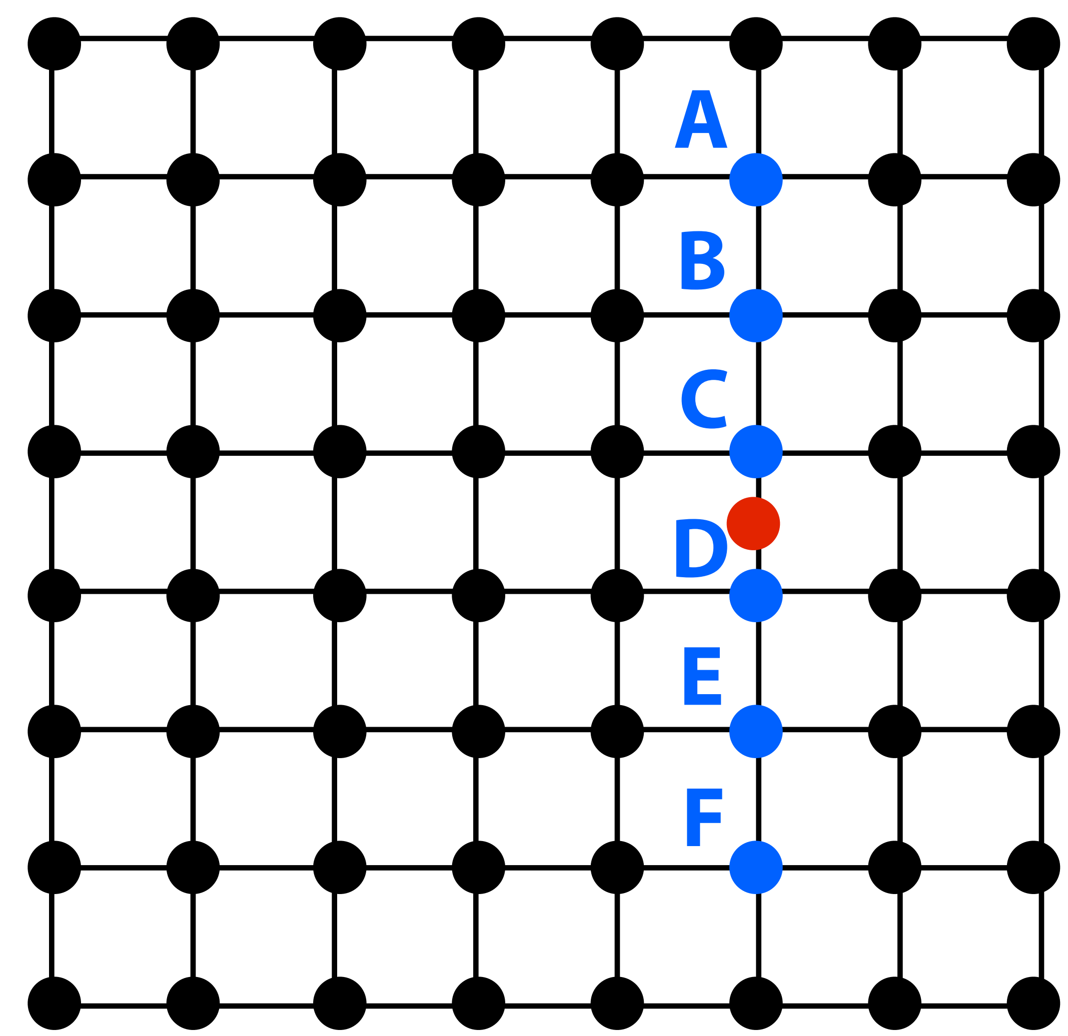
Note: non-integer motion vector

Non-integer motion vectors require resampling



Example: motion vector with 1/2 pixel values.

Must resample reference block at positions given by red dots.



Interpolation to 1/2 pixel sample points via 6-tap filter:

half_integer_value = `clamp((A - 5B + 20C + 20D - 5E + F) / 32)`

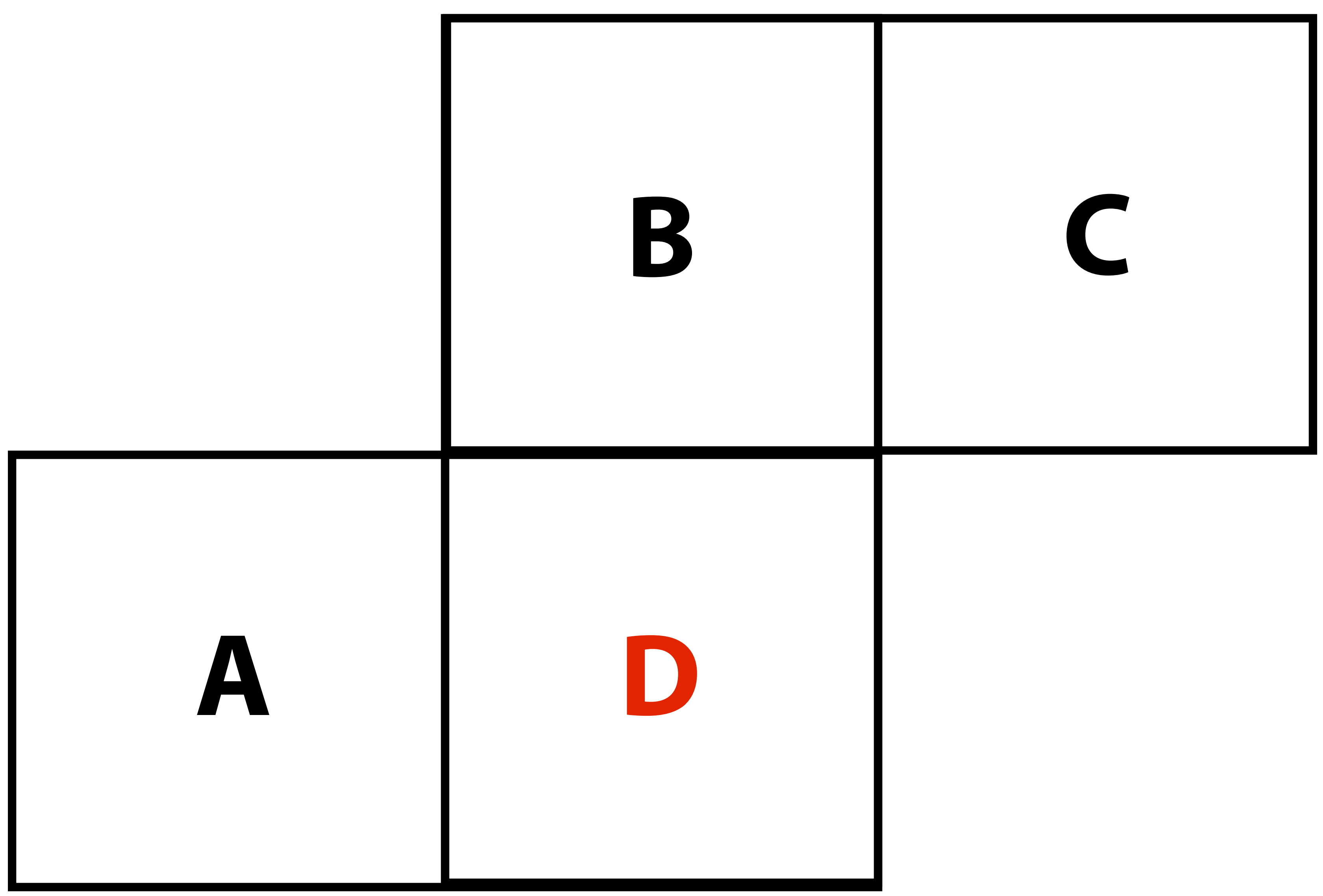
H.264 supports both 1/2 pixel and 1/4 pixel resolution motion vectors

1/4 resolution resampling performed by bilinear interpolation of 1/2 pixel samples

1/8 resolution (chroma only) by bilinear interpolation of 1/4 pixel samples

Motion vector prediction

- Problem: per-partition motion vectors require significant amount of storage
- Solution: predict motion vectors from neighboring partitions and encode residual in compressed video stream
  - Example below: predict D's motion vector as average of motion vectors of A, B, C
  - Prediction logic becomes more complex when when partitions of neighboring blocks are of different size



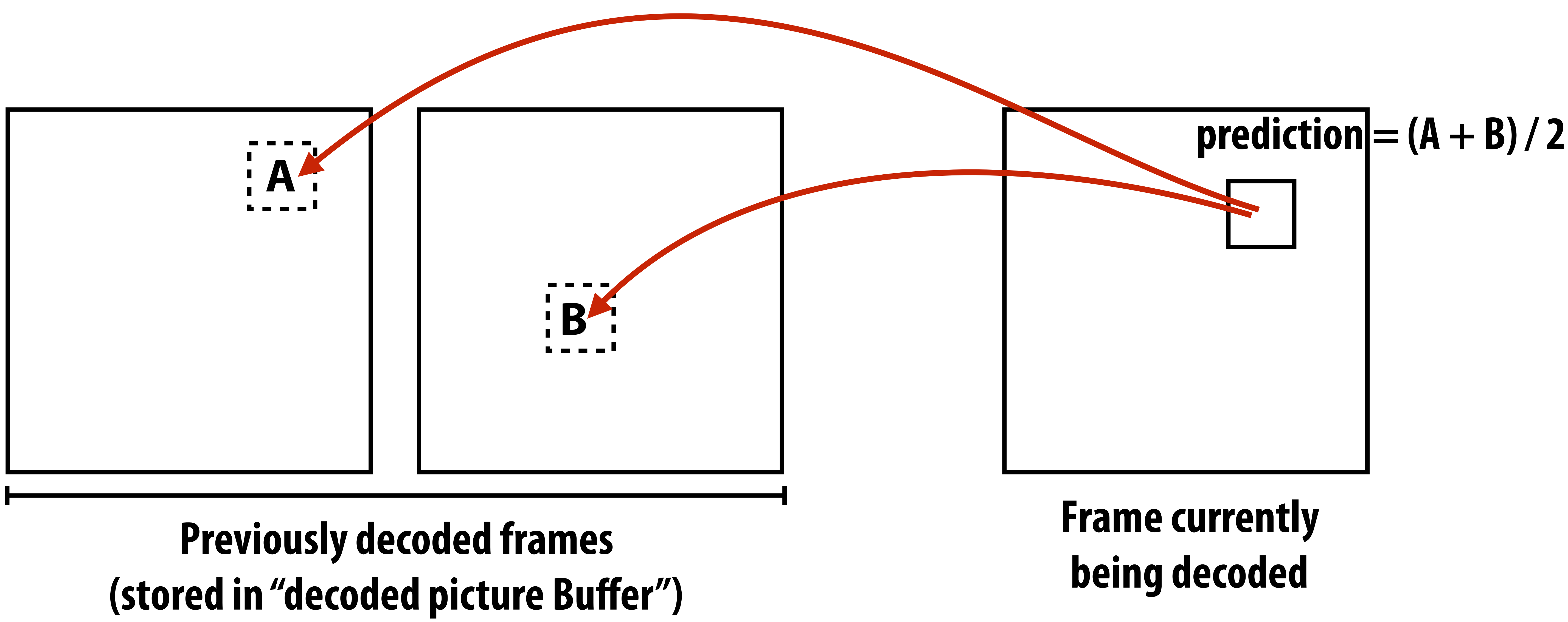
Question: what partition size is best?

- Smaller partitions likely yield more accurate prediction
  - Fewer bits needed for residuals

- Smaller partitions require more bits to store partition information (diminish benefits of prediction)
  - Reference picture id
  - Motion vectors (note: motion vectors are more coherent with finer sampling, so they likely compress well)

Inter-frame prediction (B-macroblock)

- Each partition predicted by up to two source blocks
  - Prediction is the average of the two reference blocks
  - Each B-macroblock partition stores two frame references and two motion vectors (recall P-macroblock partitions only stored one)



Additional prediction details

- **Optional weighting to prediction:**
  - Per-slice explicit weighting (reference samples multiplied by weight)
  - Per-B-slice implicit weights (reference samples weights by temporal distance of reference frame from current frame in video)
    - Idea: weight samples from reference frames nearby in time more
- **Deblocking**
  - Blocking artifacts may result as a result of macroblock granularity encoding
  - After macroblock decoding is complete, optionally perform smoothing filter across block edges.

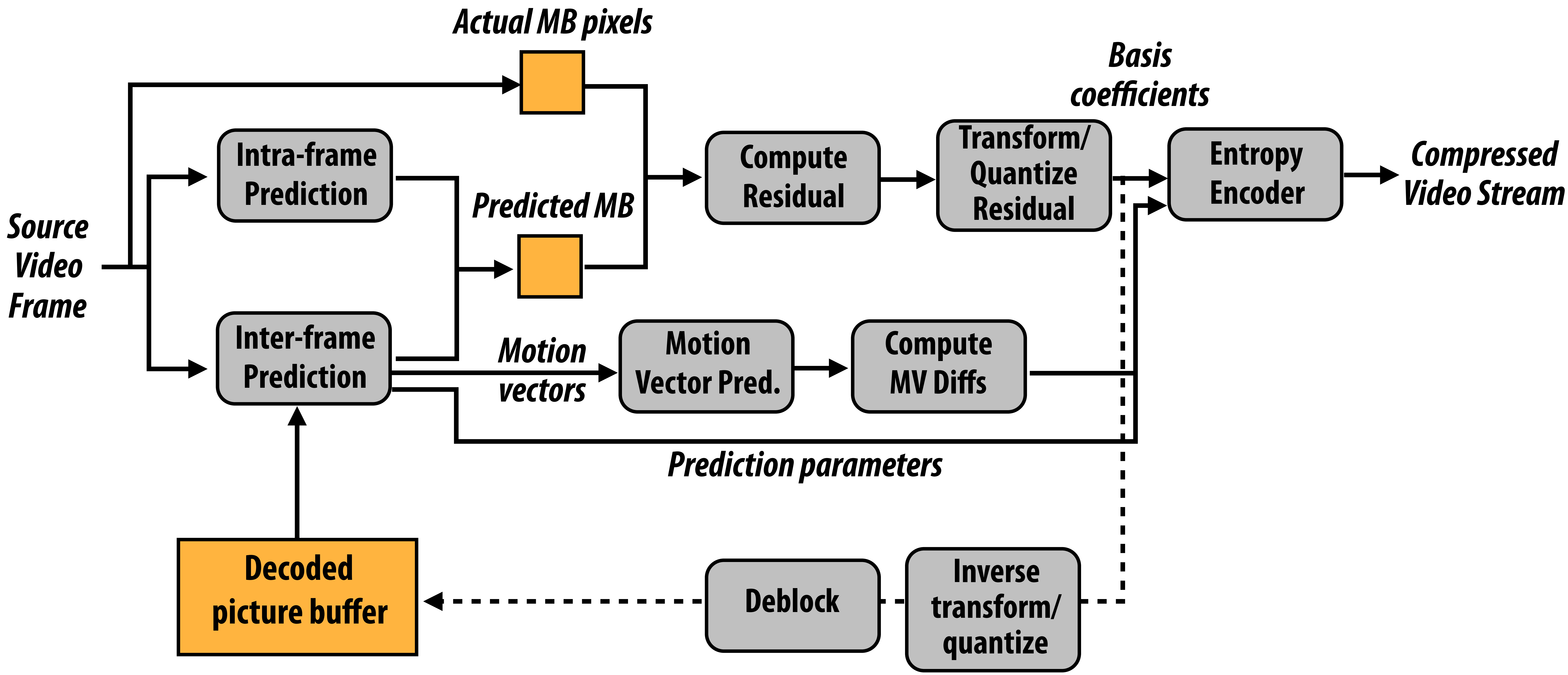
Putting it all together: encoding an inter-predicted macroblock

- Inputs:
  - Current state of decoded picture buffer (state of the decoder)
  - 16x16 block of input video to encode
- General steps: (need not be performed in this order)
  - Resample images in decoded picture buffer to obtain 1/2, and 1/4, 1/8 pixel resampling
  - Choose prediction type (P-type or B-type)
  - Choose reference pictures for prediction
  - Choose motion vectors for each partition (or sub-partition) of macroblock
  - Predict motion vectors and compute motion vector difference
  - Encode choice of prediction type, reference pictures, and motion vector differences
  - Encode residual for macroblock prediction
  - Store reconstructed macroblock (post deblocking) in decoded picture buffer to use as reference picture for future macroblocks

Coupled decisions

H.264/AVC video encoding

MB = macroblock
MV = motion vector

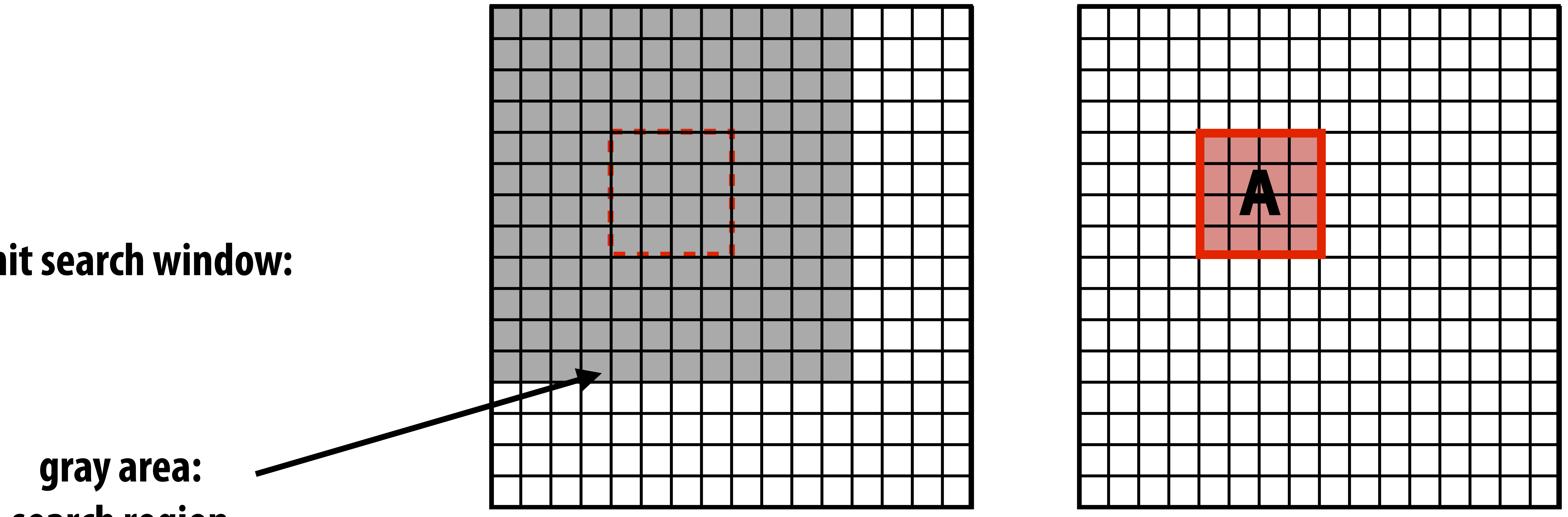


Credit: Figure derived from H.264 Advanced Video Compression Standard, I. Richardson, 2010

Motion estimation

- Encoder must find reference block that predicts current frame's pixels well.
  - Can search over multiple pictures in decoded picture buffer + motion vectors can be non-integer (huge search space)
  - Must also choose block size (macroblock partition size)
  - And whether to predict using combination of two blocks
  - Literature is full of heuristics to accelerate this process
    - Remember, must execute motion estimation in real-time for HD video (1920x1080), on a low-power smartphone

Limit search window:

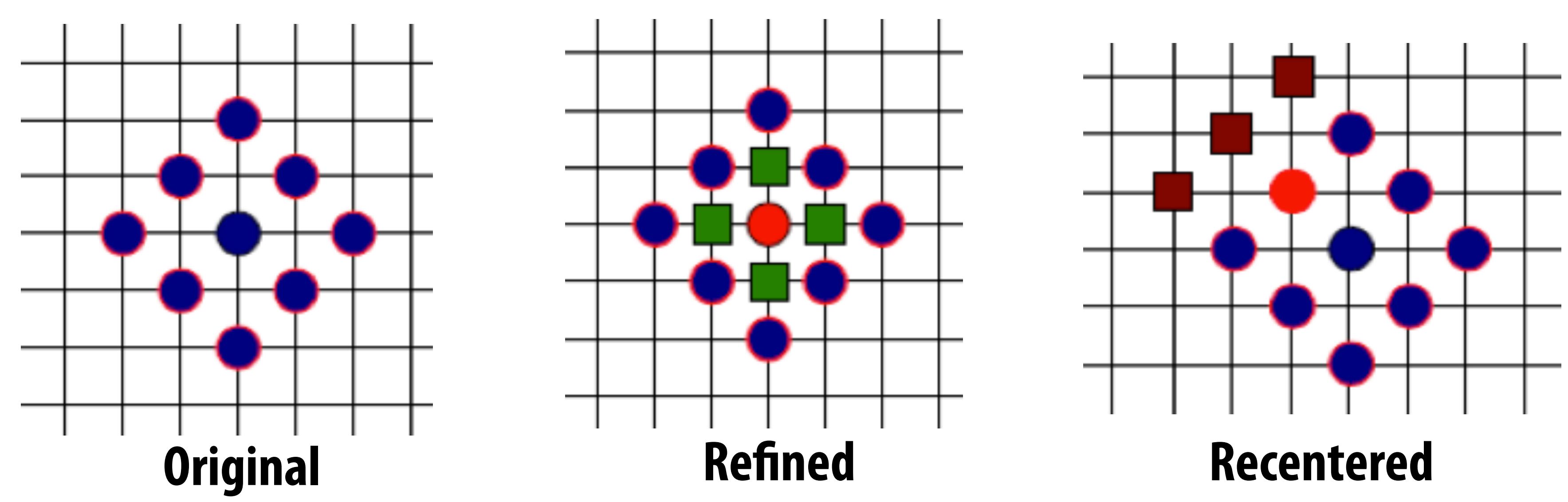


Decoded picture buffer: frame 0

Current frame

Motion estimation optimizations

- Coarser search:
  - Limit search window to small region
  - First compute block differences at coarse scale (save partial sums from previous searches)
- Smarter search:
  - Guess motion vectors similar to motion vectors used for neighboring blocks
  - Diamond search: start by test large diamond pattern centered around block
    - If best match is interior, refine to finer scale
    - Else, recenter around best match



- Early termination: don't find optimal reference patch, just find one that's "good enough": e.g., compressed representation is lower than threshold
  - Test zero-motion vector first (optimize for non-moving background)
- Optimizations for subpixel motion vectors:
  - Refinement: find best reference block given only pixel offsets, then try 1/2, 1/4-subpixel offsets around this match

H.265 (HVEC)

- Standard ratified in 2013
- Goal: ~2X better compression than H.264
- Main ideas:
  - Macroblock sizes up to 64x64
  - Prediction block size and residual block sizes can be different
  - 35 intra-frame prediction modes (recall H.264 had 9)
  - …