Overview

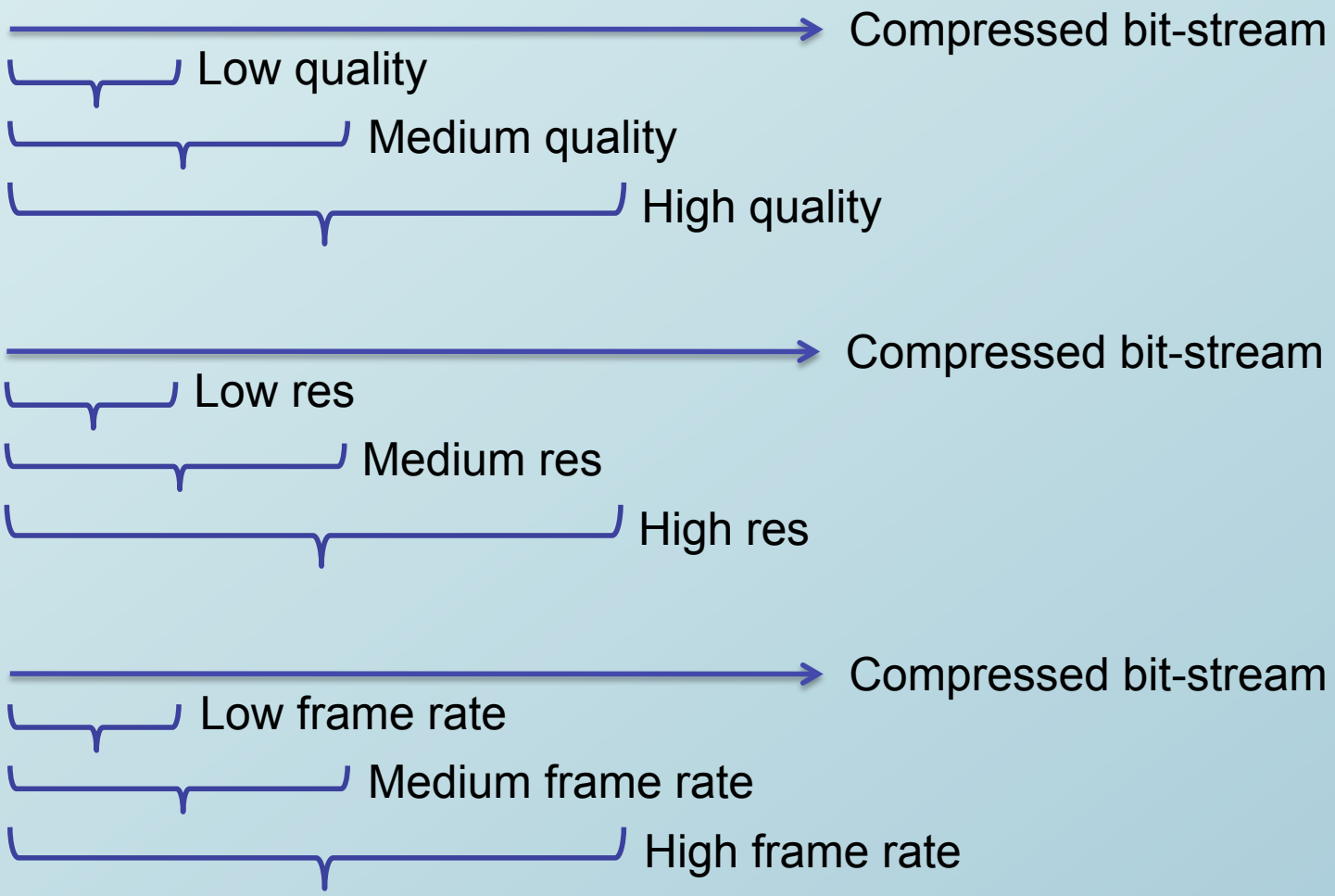
- Introduction to scalable media compression
  - emerging trends
  - scalability and accessibility
  - things that work well
- The SVC extension to H.264
- Beyond prediction
  - motion compensated temporal transforms and their merits
- Spatio-temporal transform structures for scalable video
  - wavelets, pyramids and lifting structures
- Beyond video
  - other media types
- Motion models for scalable video
  - important properties
  - block-based and block-free motion schemes
  - scalable compression of sparse innovations (discontinuities)
- Related research directions and themes

The Changing Landscape of Video

- Video formats
  - QCIF (25 Kpel), CIF (100 Kpel),
    4CIF/SDTV (½ Mpel), HDTV (2 Mpel)
    UHDTV 4K (10 Mpel) and 8K (32 Mpel) – ITU, June 2012
  - Cinema: 24/48/60 fps; UHDTV: potentially up to 120 fps
- Displays
  - "retina" resolutions (200 to 400 pixels/inch)
  - what resolution video do I need for an iPad? (2048x1536?)
- Internet and mobile devices
  - ~80% of internet traffic is video
  - YouTube 2nd most popular web-site - 4 billion views/day
  - Global mobile TV subscribers to reach ~800M by 2014
- New media: multi-view video, 2.5D (texture+depth)

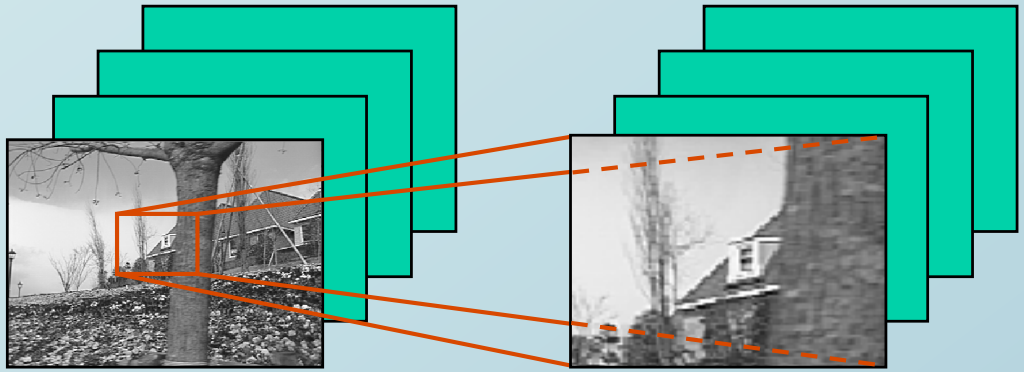
Scalability – degrees of interest

- Usually implies embedding

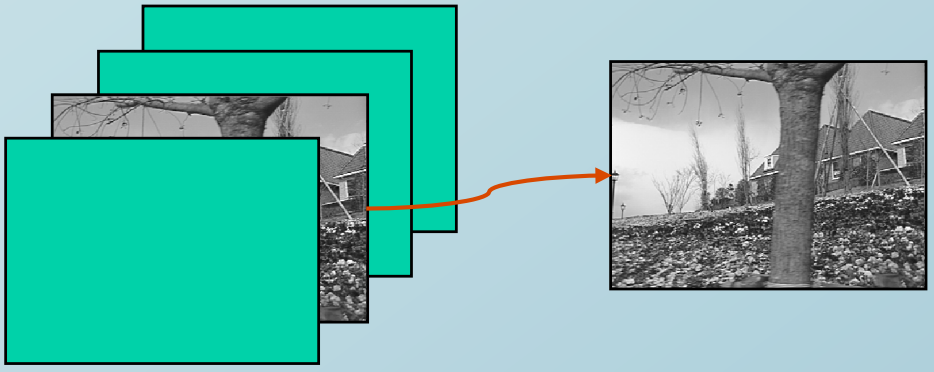


Accessibility – disjoint subsets of interest

- Spatial region of interest



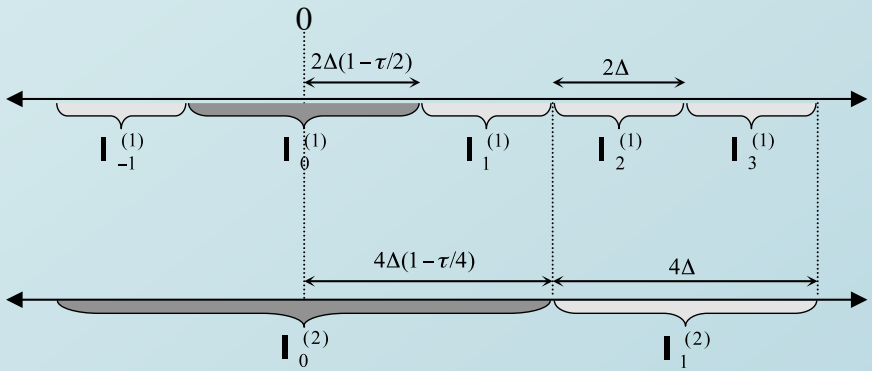
- Temporal region (or frames) of interest

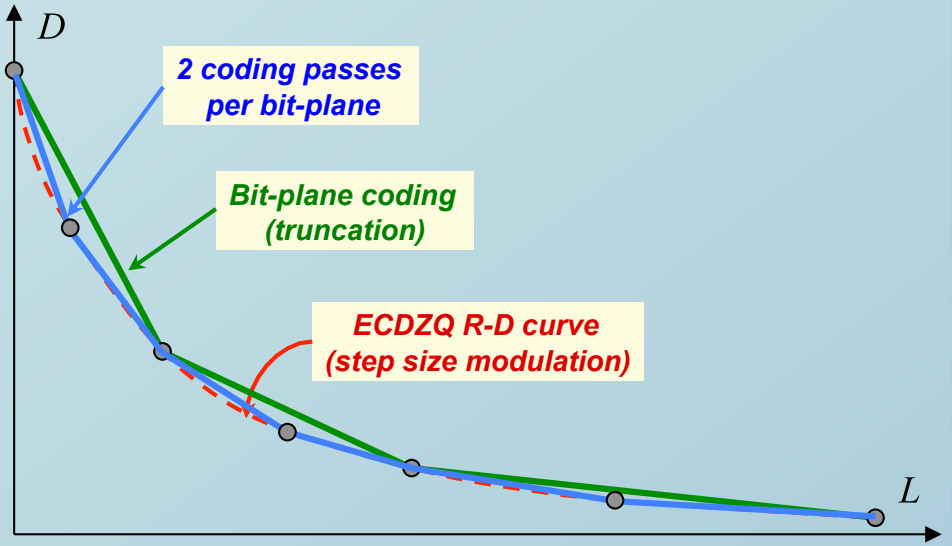


- Implications:
  - need to break or localize dependencies

Scalable images – things that work well

- Multi-resolution transforms
  - 2D wavelet transforms work well
- Embedded coding
  - Successive refinement through bit-plane coding
  - Multiple coding passes/bit-plane improve embedding

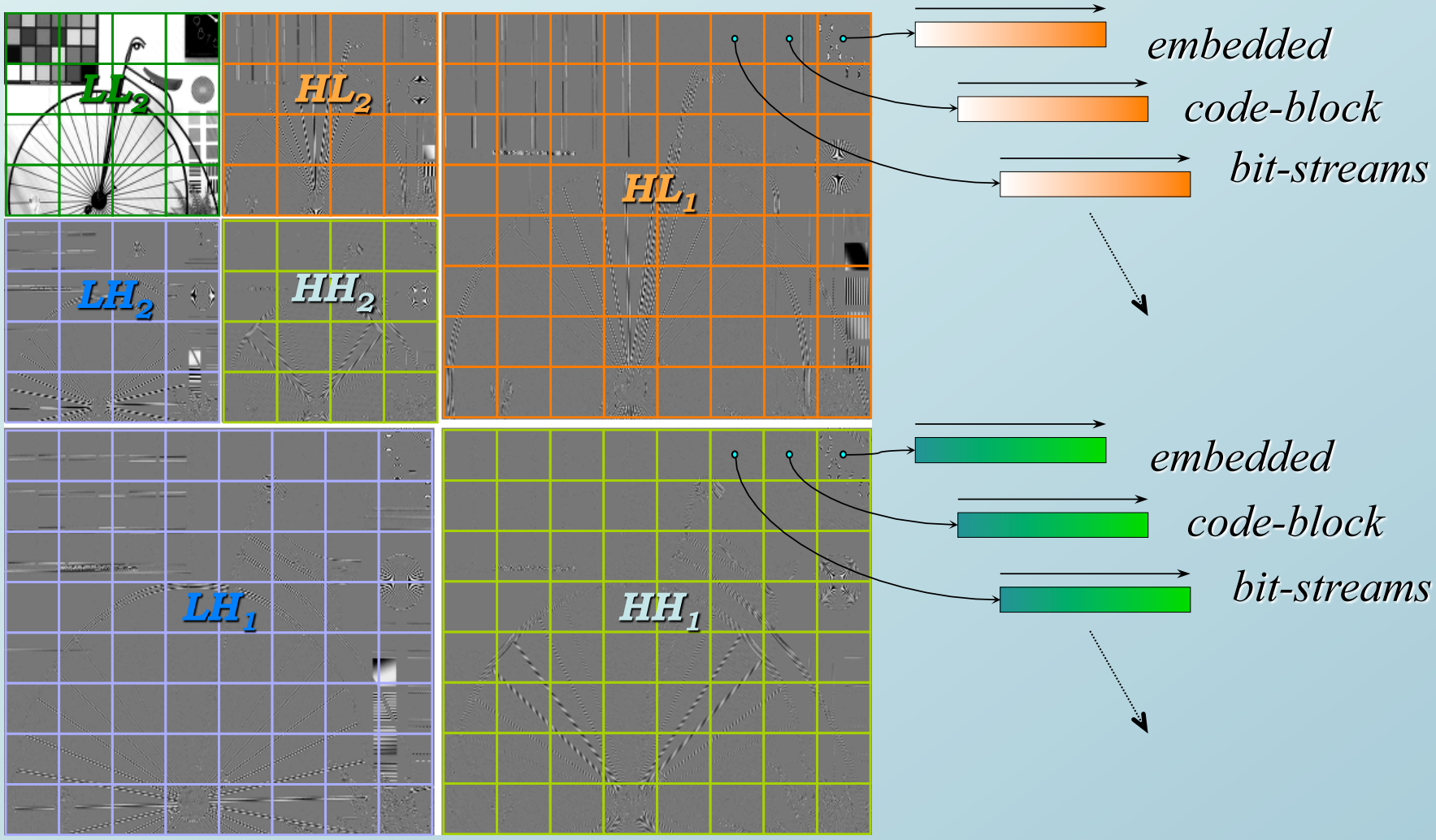




- Accessibility through partitioned coding of subbands
  - Region of interest access without any blocking artefacts

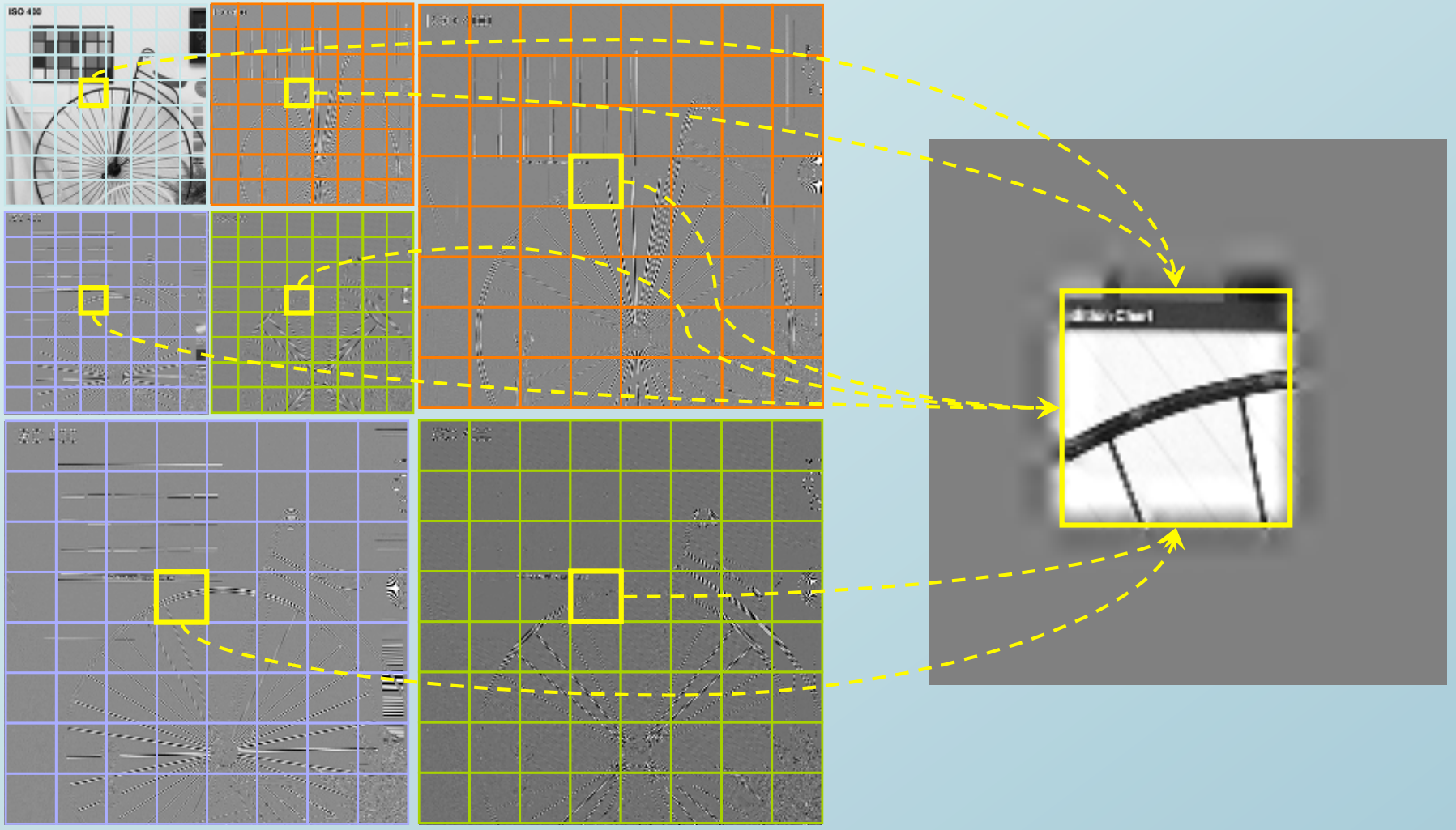
JPEG2000 – more than compression

Decoupling and embedding



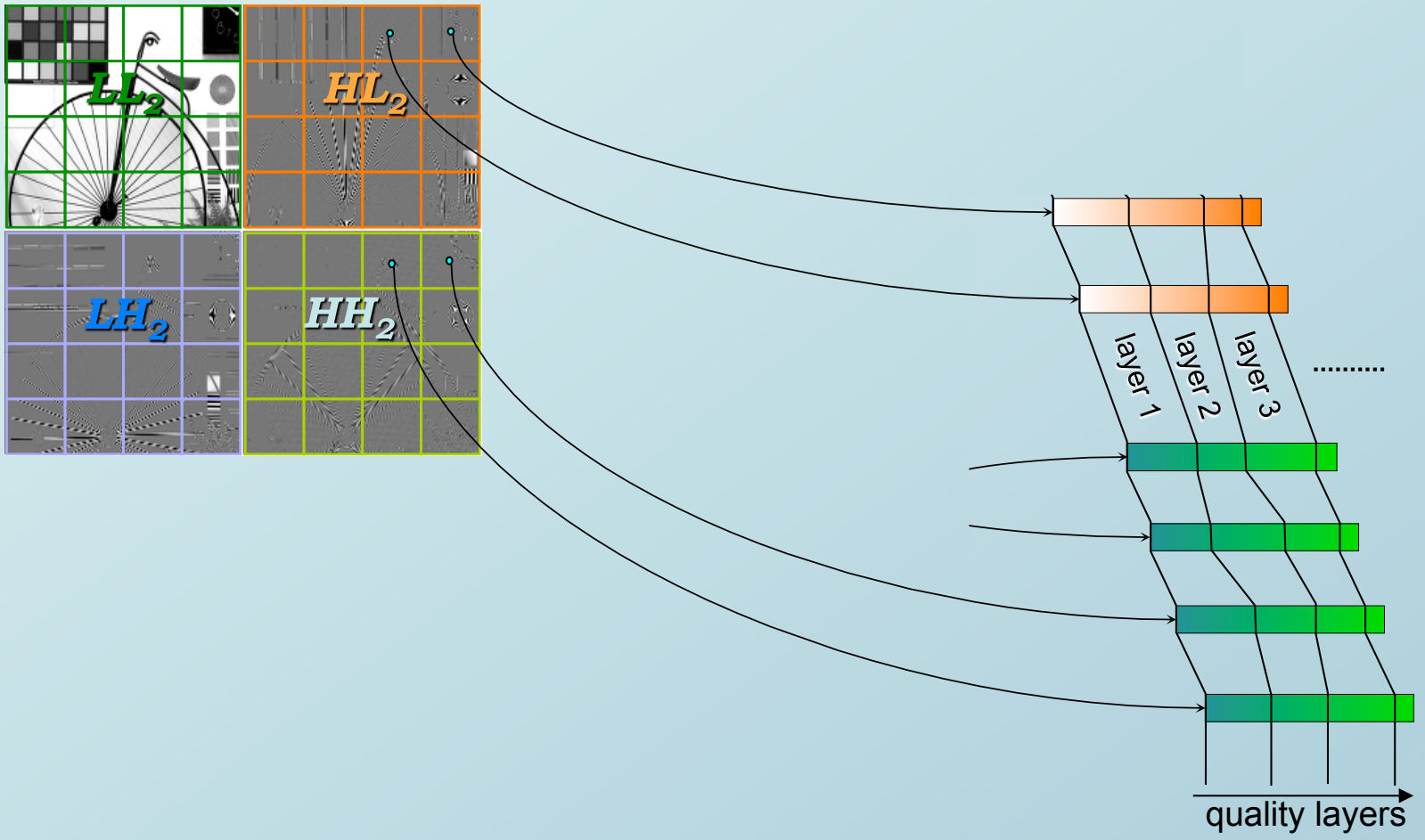
JPEG2000 – more than compression

Spatial random access

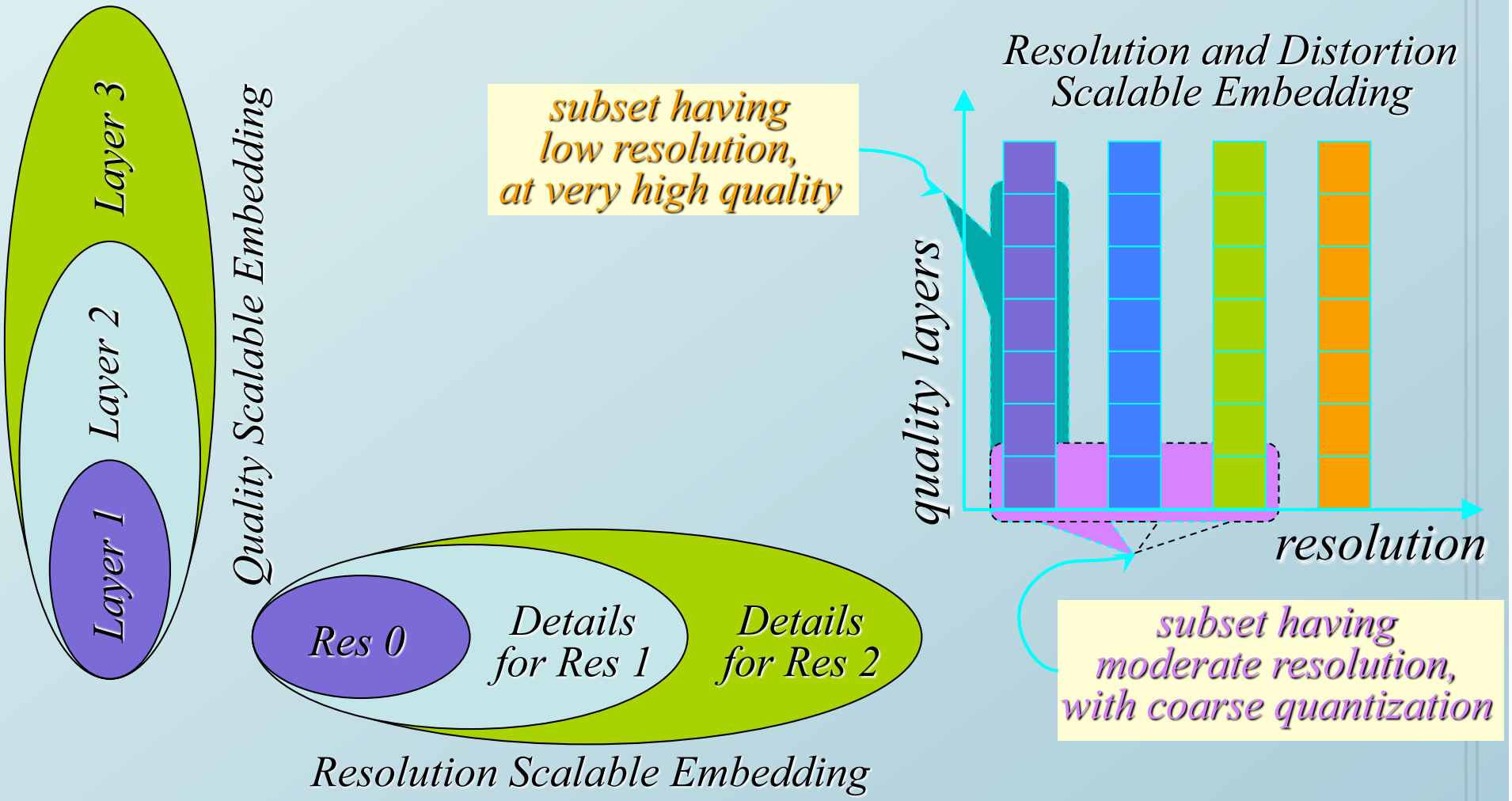


JPEG2000 – more than compression

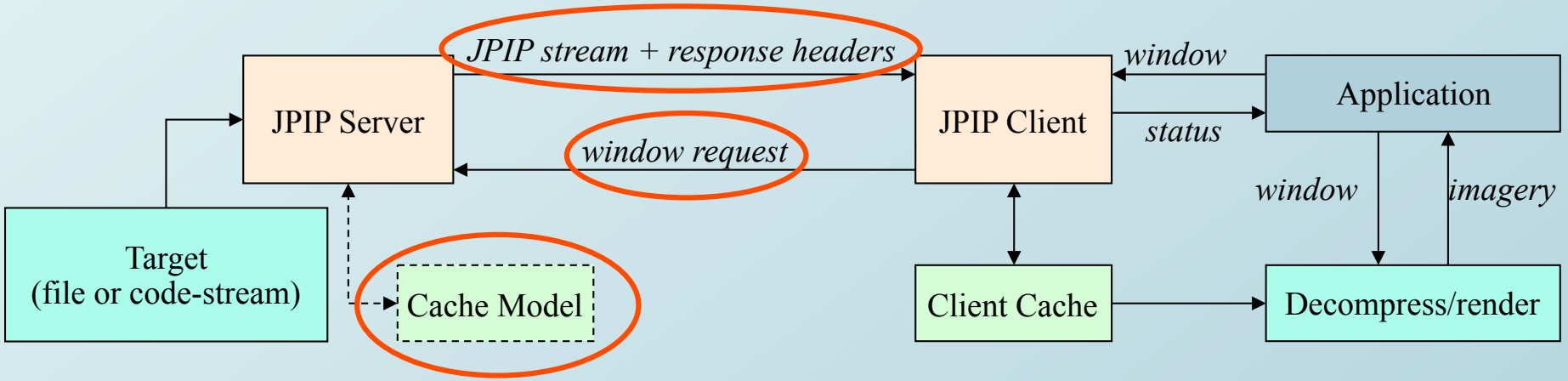
Quality and resolution scalability



JPEG2000 – dimensions of scalability



JPEG2000 – JPIP interactivity (IS15444-9)



- Client sends “window requests”
  - spatial region, resolution, components, …
- Server sends “JPIP stream” messages
  - self-describing, arbitrarily ordered
  - pre-emptable, server optimized data stream
- Server typically models client cache
  - avoids redundant transmission

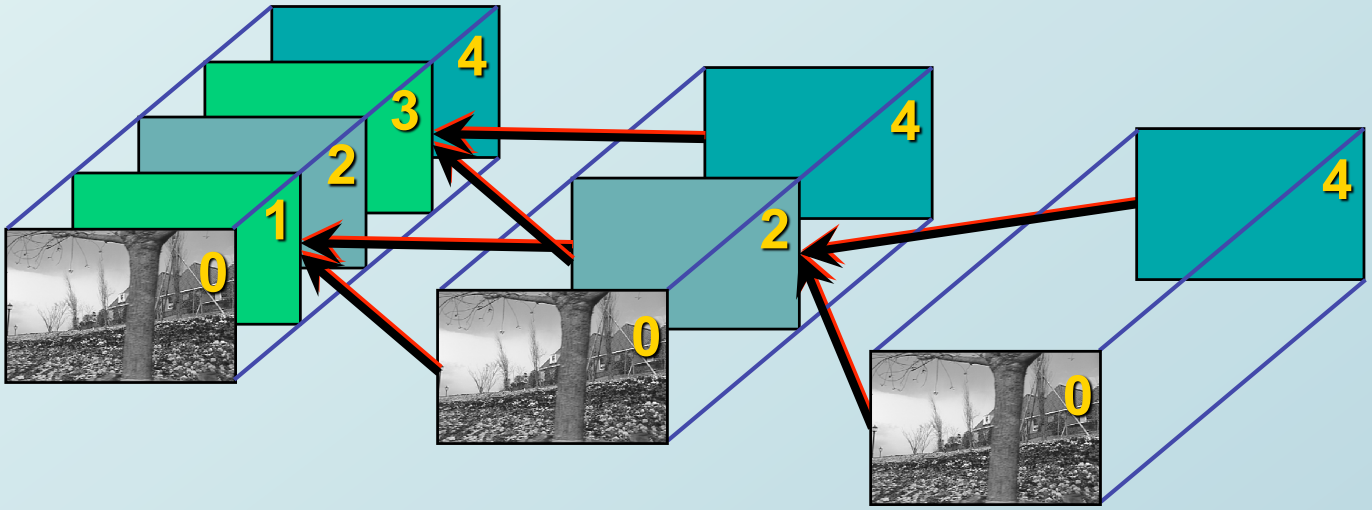
What can you do with JPIP?

- Highly efficient interactive navigation within
  - large images (giga-pixel, even tera-pixel) **Aerial Demo**
  - medical volumes **Catscan Demo**
  - virtual microscopy
  - window of interest access, progressive to lossless
  - interactive metadata **Album Demo** **Campus Demo**
- Interactive video
  - frame of interest **Panoramic Video Demo**
  - region of interest
  - frame rate and resolution of interest
  - quality improves each time we go back over content

Scalable video standardization

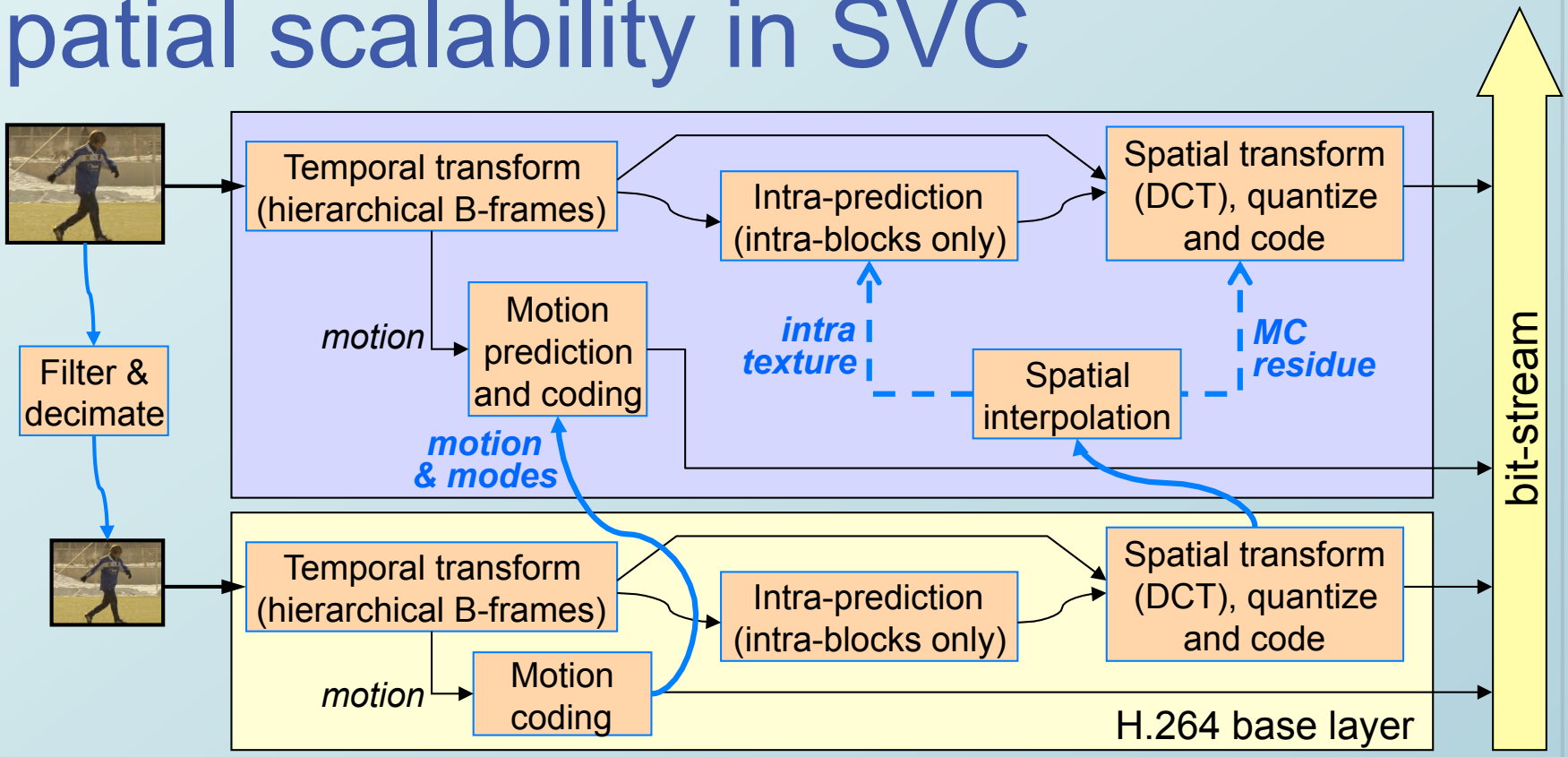
- SVC extension to H.264/AVC
- Lots of prediction
  - good adaptation of the prediction strengths of H.264
  - new macro-block modes and slice options
- Supports temporal, spatial and quality scalability
  - also supports combinations of these scalabilities
- Key design objectives
  - relatively small set of defined “access layers”
  - minimal increase in decoding complexity w.r.t. H.264
  - minimal loss in coding efficiency from scalability
    - has to be much better than “simulcast”

Temporal scalability in SVC



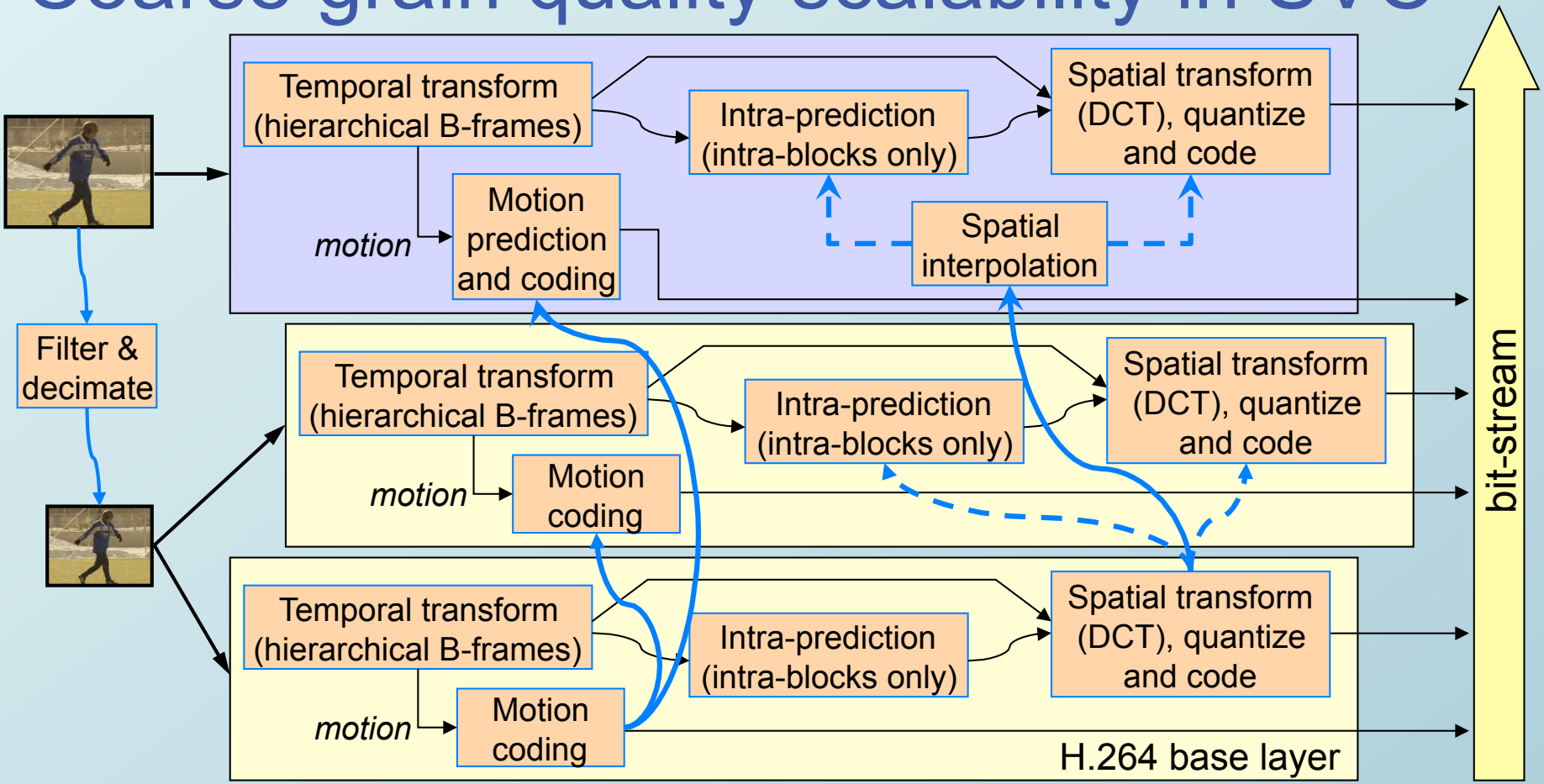
- Essentially hierarchical B-frames
  - Temporal prediction only: no temporal update steps
  - Not limited to the B-frame structure
    - use prev coded frames at the same or a coarser temporal level
    - allows non-dyadic frame-rates
- Encoding typically not open-loop
  - prediction residuals based on quantized reference frames

Spatial scalability in SVC



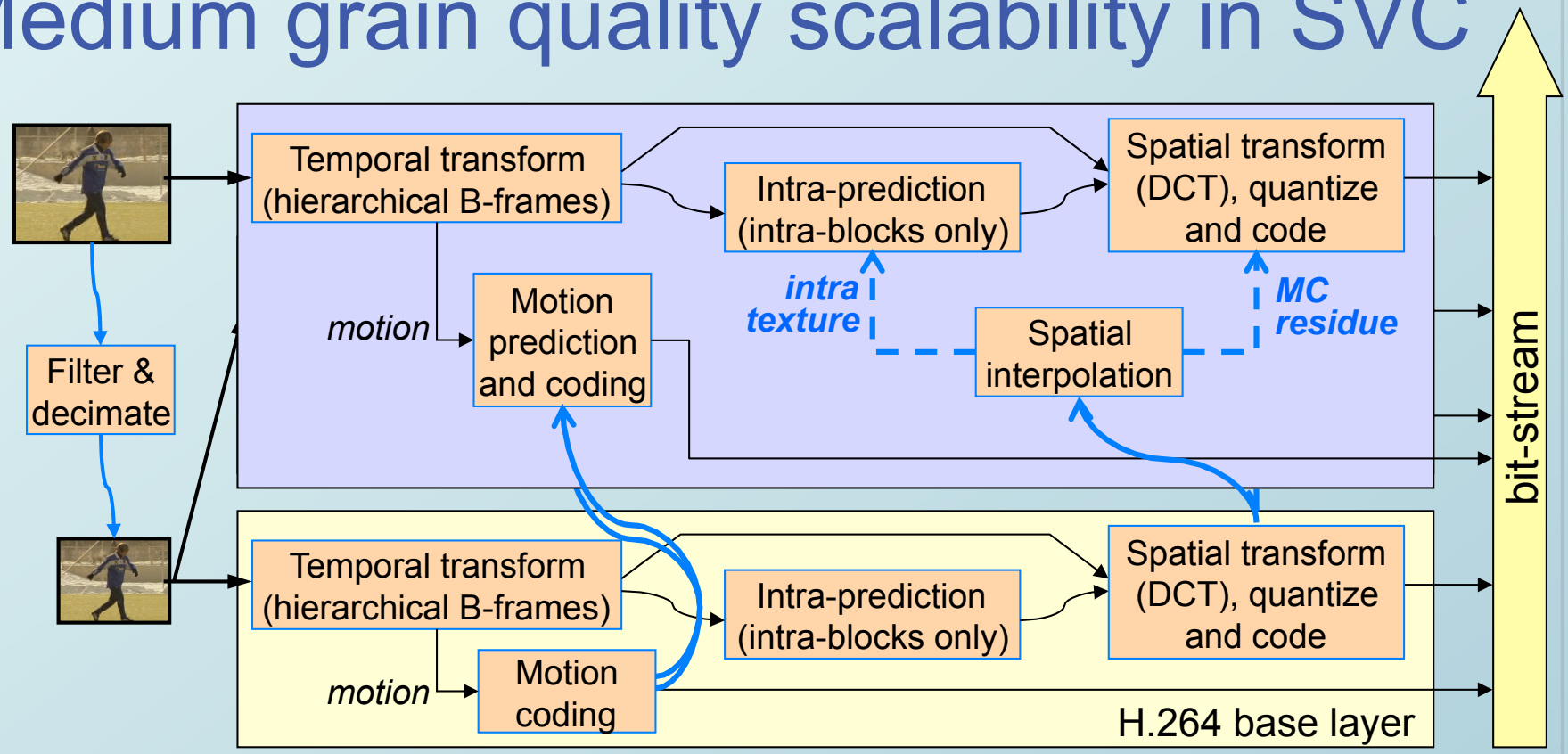
- Multi-resolution pyramid – redundant sampling
- Macro-block modes allow optional re-use of:
  - motion and macro-block modes from lower layer
  - intra-coded samples from lower layer (for intra-blocks)
  - prediction residues from lower layer (for non-intra blocks)
  - decoder runs only one motion compensation loop

Coarse grain quality scalability in SVC



- Extra "spatial" layers the same resolution
  - SVC uses the term "dependency layer"
  - Each higher layer depends on one specific lower layer
    - not fully embedded

Medium grain quality scalability in SVC



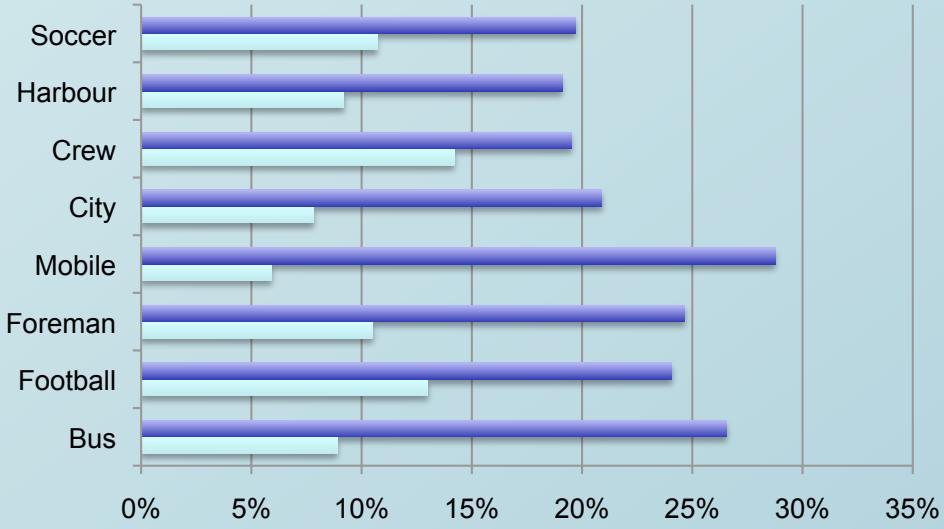
- Similar to CGS, but
  - One dependency layer formed from multiple quality layers
    - Dependent (higher spatial resolution) layers use highest available quality for prediction
    - Except where use of lower quality forced by “key frames”
  - Decoder still runs one motion compensation loop

SVC Efficiency

- Can come within ~10% of H.264/AVC bit-rate

  (Schwarz, Marpe & Wiegand, 2007)

  – depends on number of quality layers
- Not easy to optimize at encoder
  – multiple coupled closed-loop encoders: one per layer
  – bottom-up approach (layer by layer) not optimal
- Performance of spatial scalability



(Segall & Sullivan, 2007)

2 layers only, JVT test bit-rates

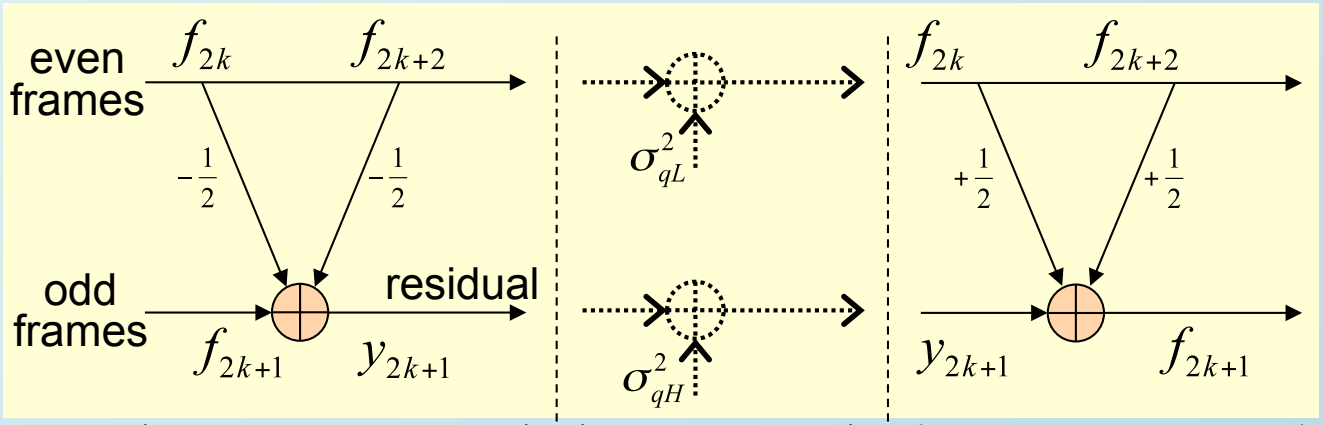
SVC rate reduction relative to simultcast

Single layer rate reduction relative to simultcast

Limitations of SVC

- Prediction only solution – inherently sub-optimal
  - in time (hierarchical B-frame prediction)
  - in space (prediction across scales)
  - in motion (prediction across scales) – least effective
- Redundant sampling with multi-resolution pyramids
- Not fully embedded
  - high res stream includes only some low res info
    - depends on relative quality (SNR) of low res layers
    - partially simulcast
  - having high quality content for low spatial resolution
    - may not help a lot if we then decide we want high resolution
  - oriented toward provision of a small set of access layers
    - as opposed to progressive build-up during interactive browsing
- Block-based motion modeling is not physical
  - motion does not scale well

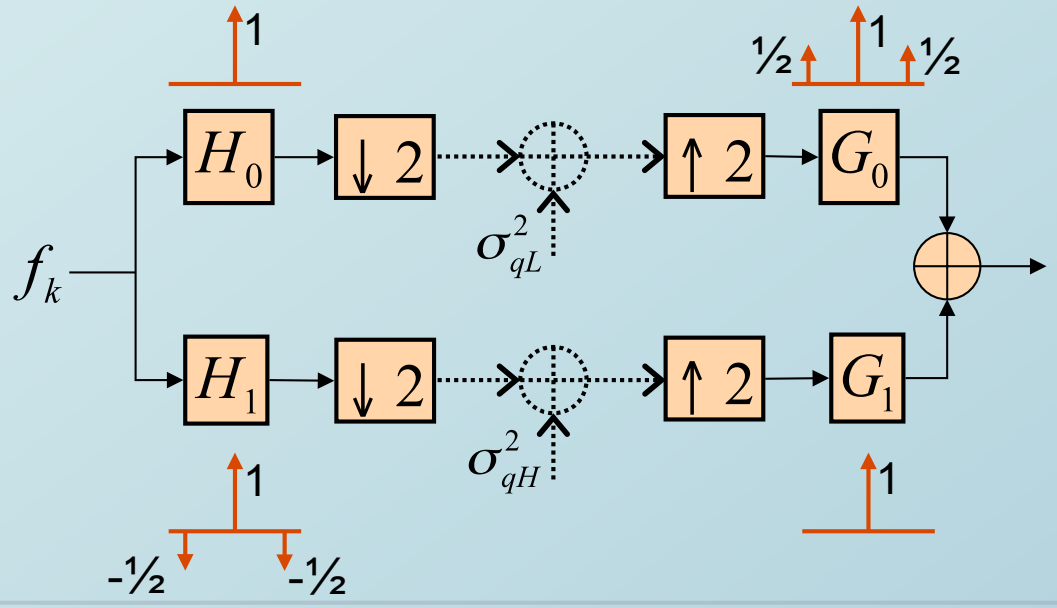
Temporal transforms: Why prediction alone is sub-optimal

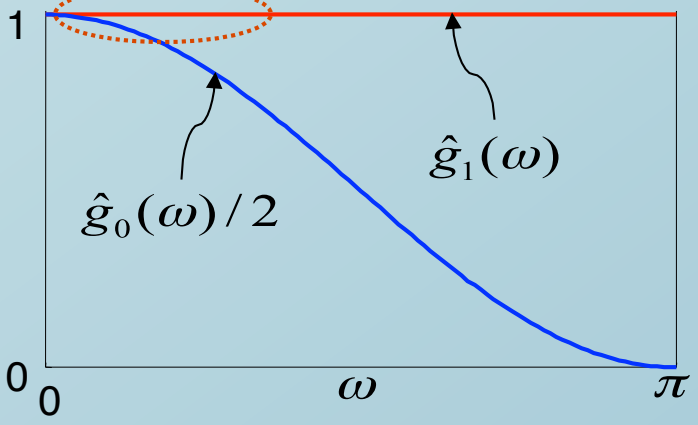


Bi-directional prediction

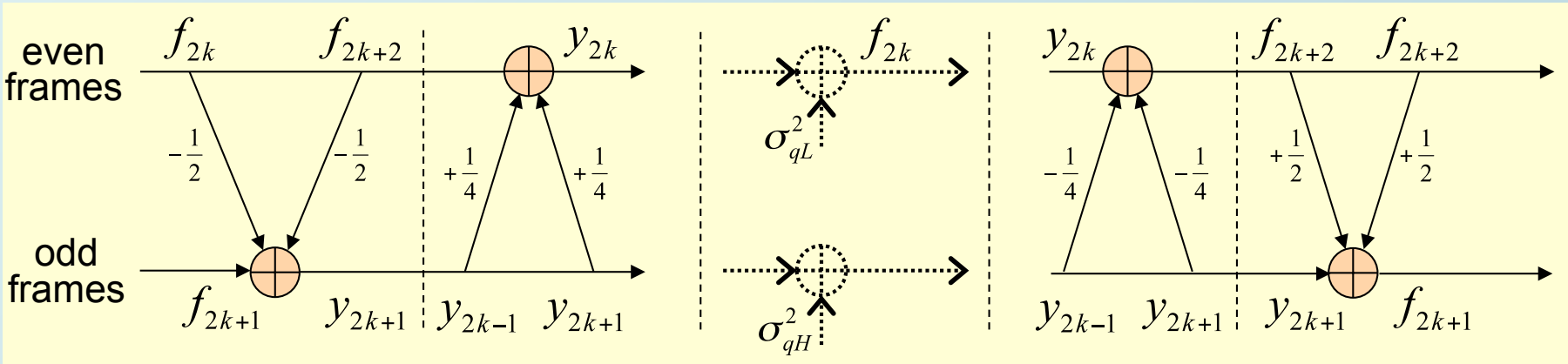
Redundant spanning of low-pass content by both channels ⇒

High-pass quantization noise has unnecessarily high energy gain.

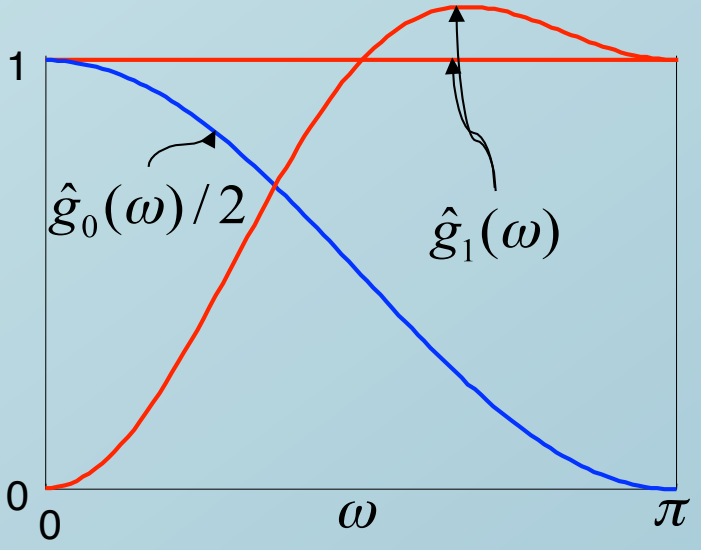




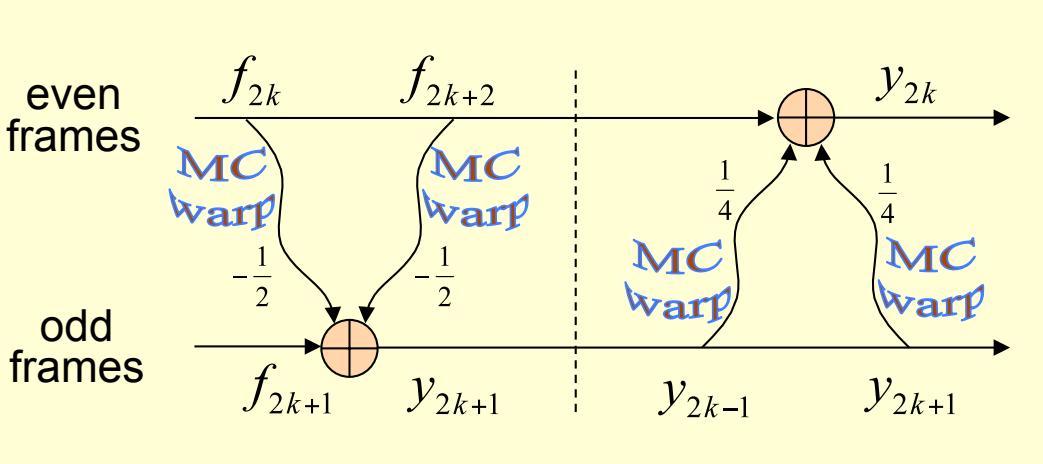
Reduced noise power through lifting



- Inject –ve fraction of high band into low band synthesis path
  - removes low freq. noise power from synthesized high band
- Add compensating step in the forward transform
  - does not affect energy compacting properties of prediction

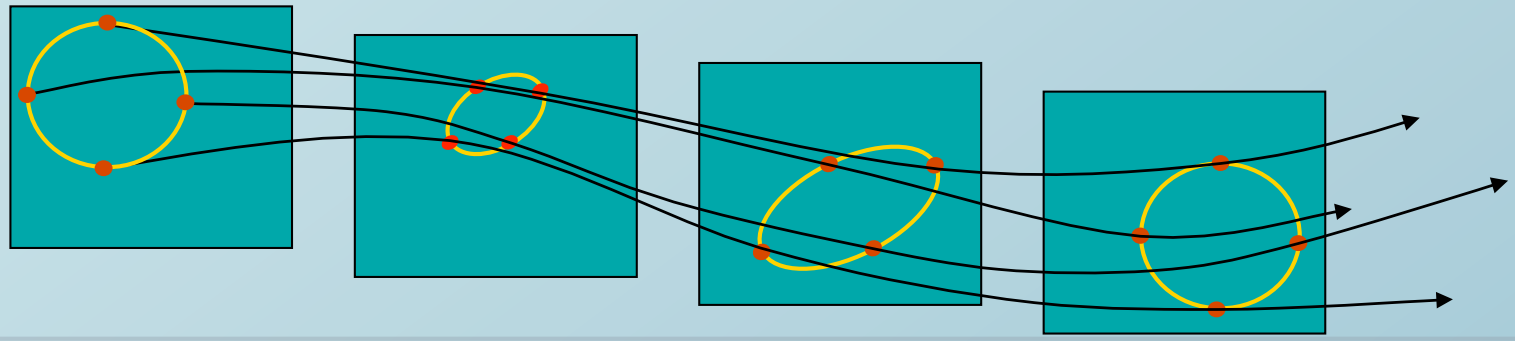


Motion compensated lifting

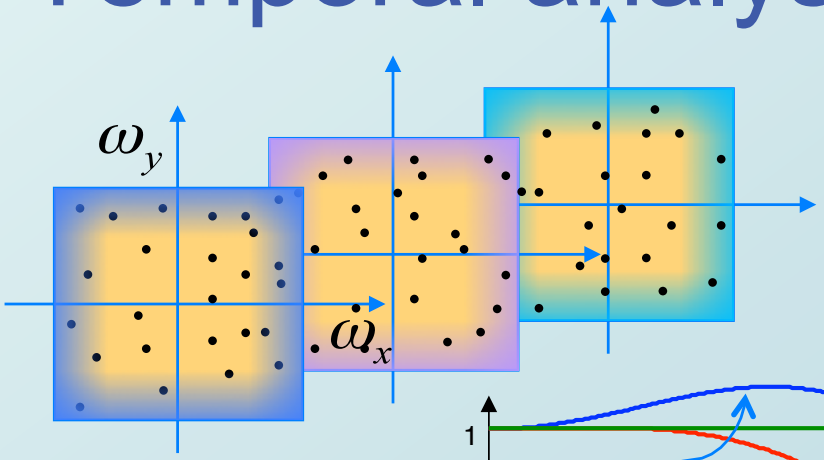


- Motion compensate each lifting step
  - transform remains reversible
- Proposed in 2001:
  (Pesquet-Popescu & Bottreau)
  (Secker & Taubman)
  (Luo, Li, Li, Zhuang, Zhang)

- All FIR subband/wavelet transforms have lifting factorizations
- MC warped lifting steps ⇒ xform is applied along motion trajectories:
  - provided trajectories exist (motion model is invertible);
  - strictly true only for spatially continuous frames (Secker & Taubman)



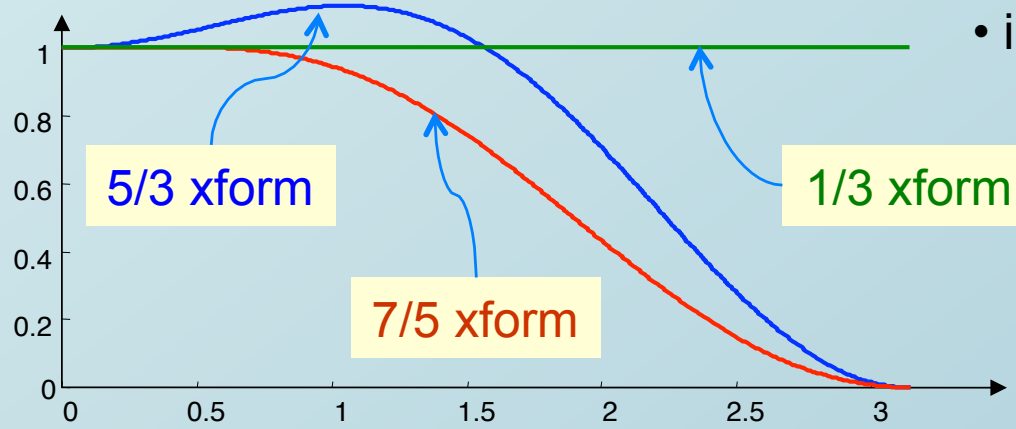
Temporal analysis effects



True scene spatial content:
- coherent across motion trajectories

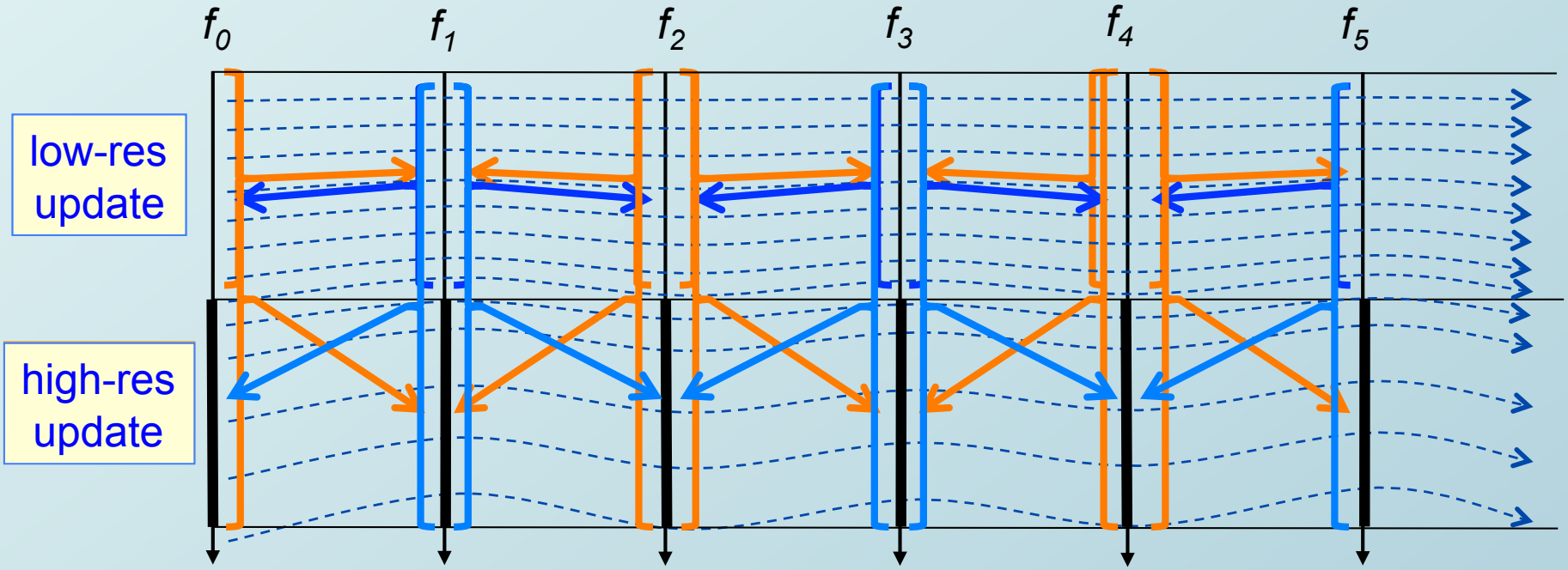
Sampling noise:
- incoherent

Spatial aliasing:
- incoherent



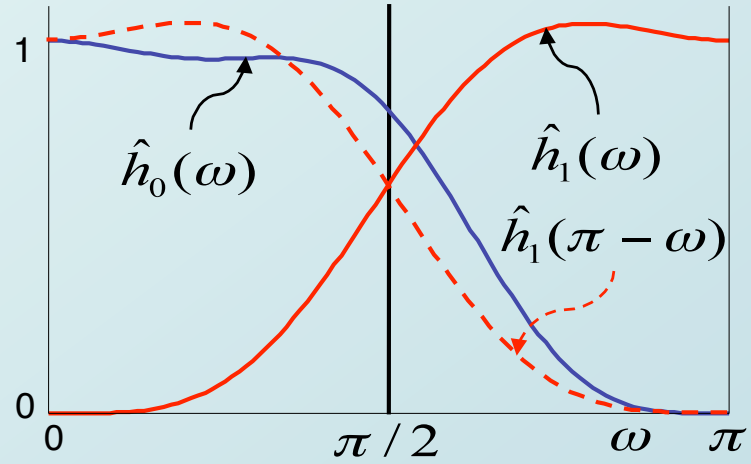
- Temporal analysis reduces noise & aliasing power
  - Improves energy compaction in next level of temporal transform
  - Improves visual appearance at reduced temporal resolutions

Spatial scalability – 2D+t



- Start with spatial multi-resolution transform
- Apply temporal transform to each spatial resolution
  - use only information from same or lower resolution
    (Andreopoulis, Van der Schaar, Munteaneau, Barbarien, Schelkens, Cornelis – 2003)
- Frequency leakage limits low-res energy compaction
- Each frame contributes its own aliasing at low-res

Wavelet transforms – critically sampled



Analysis filter responses of the popular 9/7 wavelet transform

Fundamental constraint:

(for perfect reconstruction)

$$\underbrace{\hat{h}_0(\omega)\hat{h}_1(\pi-\omega)}_{\text{half-band filter}} + \hat{h}_0(\pi-\omega)\hat{h}_1(\omega) = 1$$

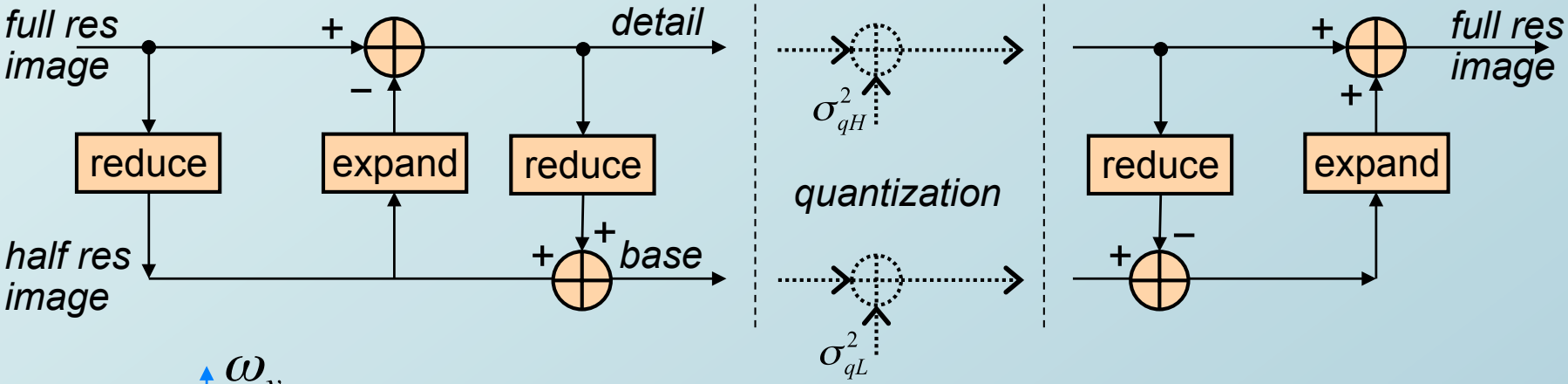
Extract LL subband

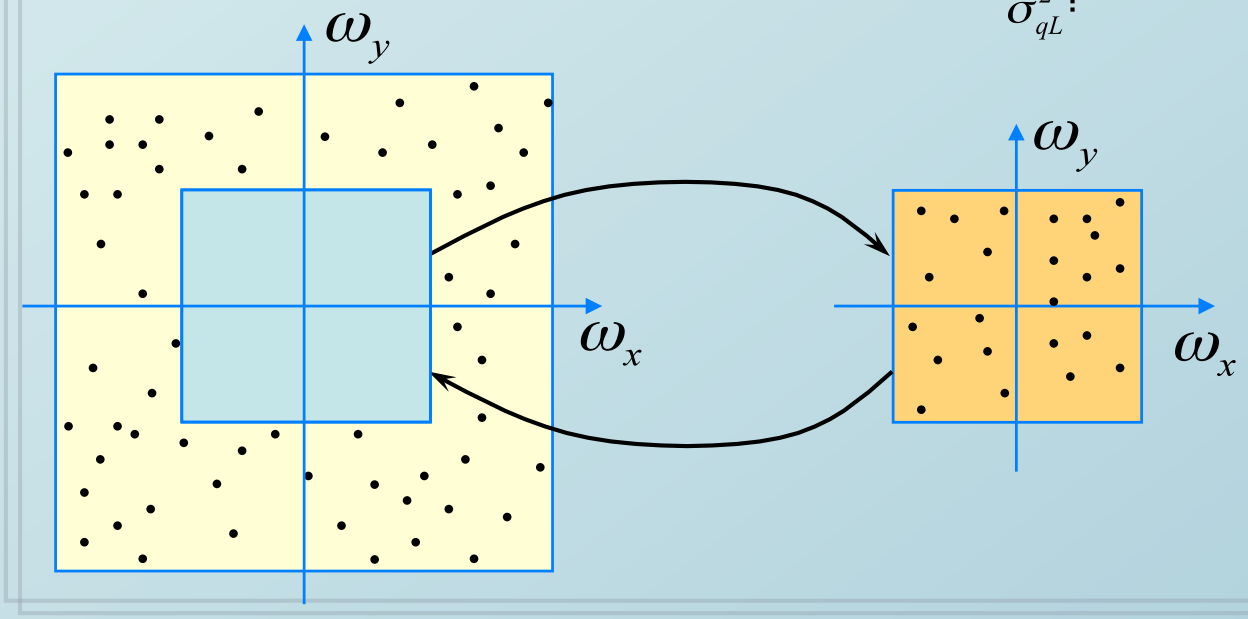
Spatial aliasing

Lifted pyramid transforms

– for improved quality scalability

(Flierl & Vandergeist, 2005)

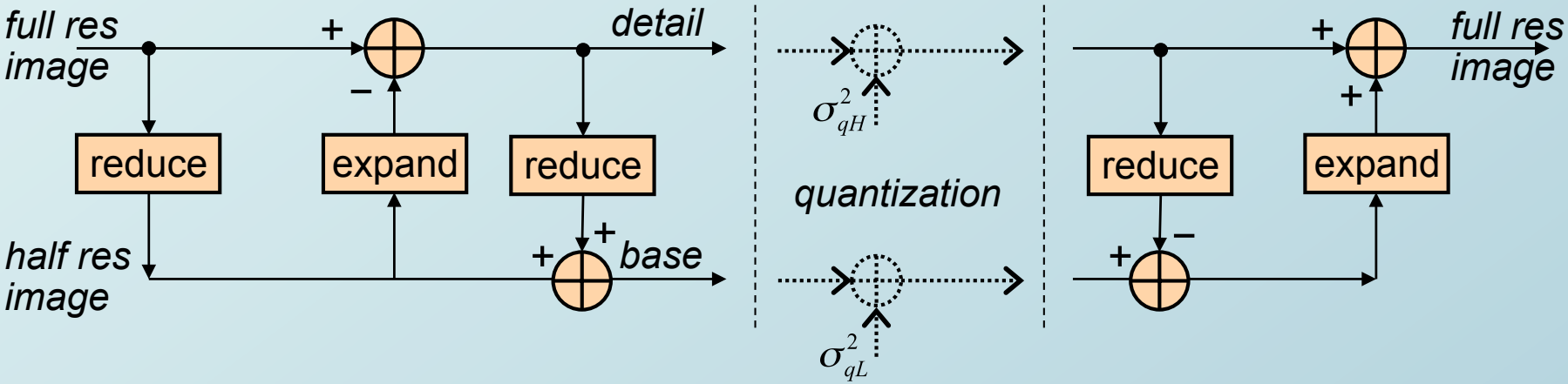




Prediction alone is sub-optimal!

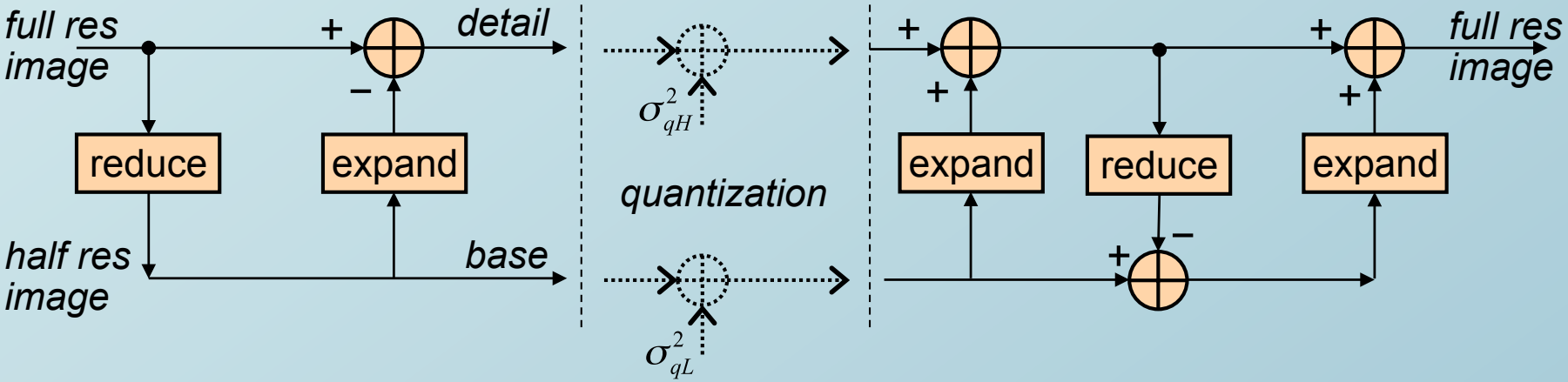
Lifted Pyramid transforms – variations

(Flierl and Vandergeist, 2005)



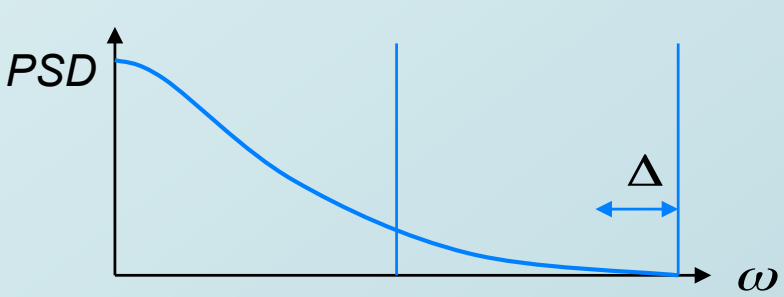
(Liu, Gan and Tran, 2008)

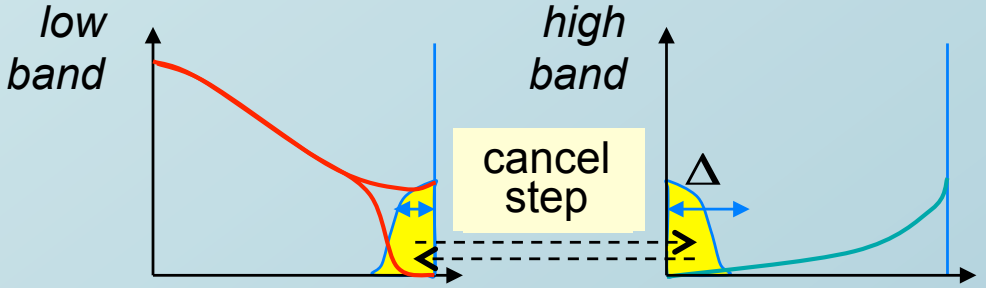
Similar compression performance,
more control over low-pass anti-aliasing filter



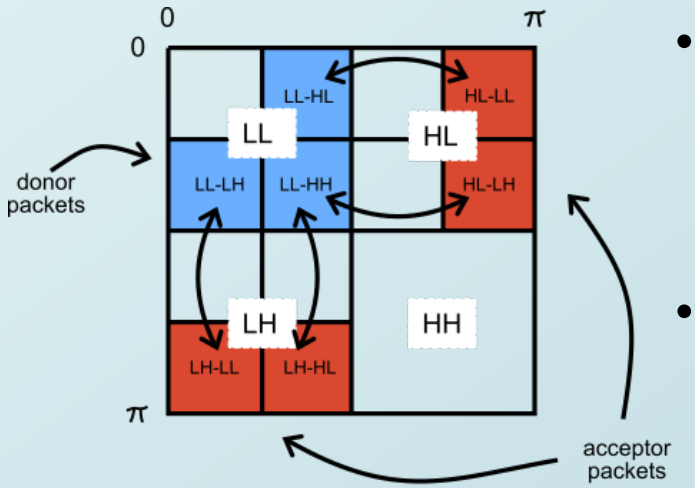
Wavelets with energy exchange – 2D+t

- Key: video spectrum rolls off quickly with frequency
  - Property not preserved by DWT at reduced resolution
- Modulated lifting steps can move spectral content
  (Gan and Taubman, 2007)



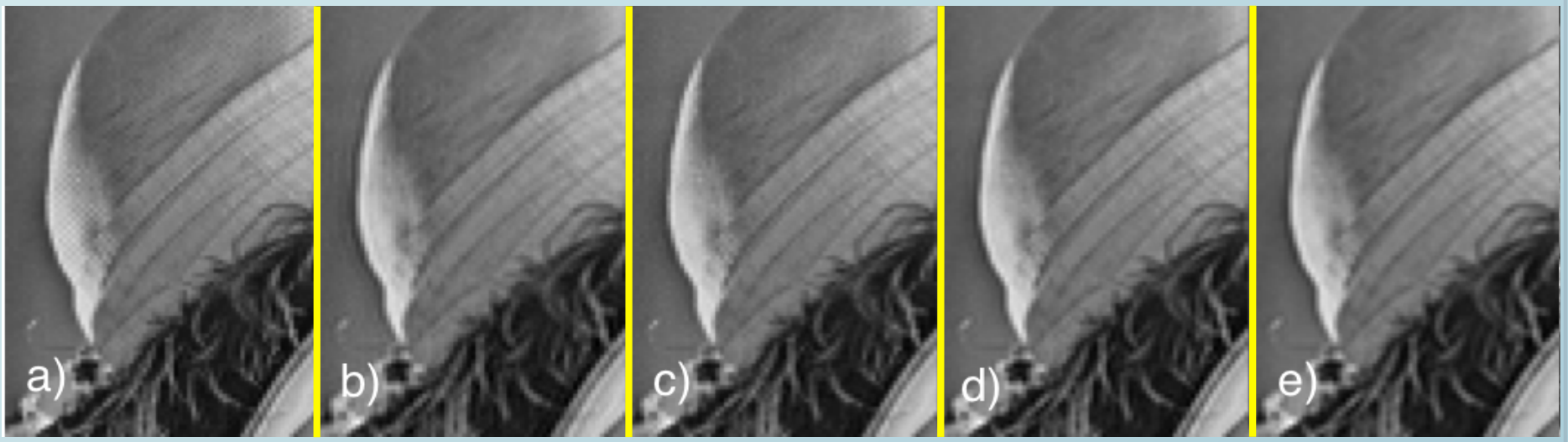


Wavelet energy exchange – variations



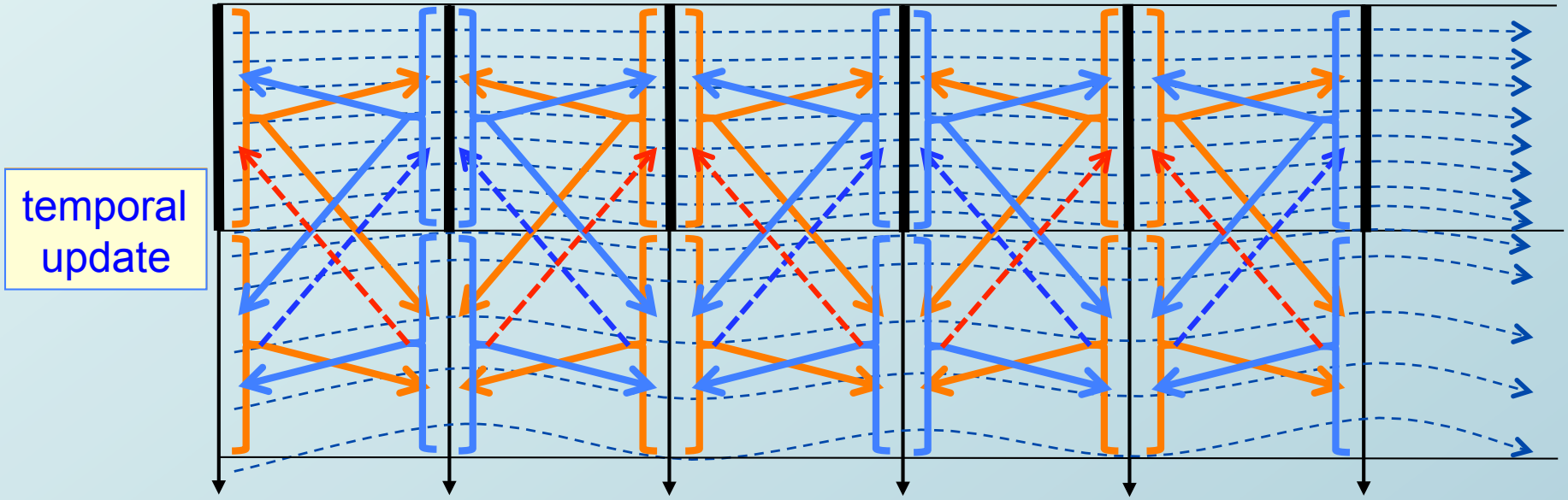
- Start with Packet Wavelet transform
  - transfer step moves aliased content to “acceptor packets”
  - cancel step cancels aliased content in “donor packets”
- Can make transfer step adaptive
  - modulate by local estimate of aliasing energy in donor packets

(Gan and Taubman, 2009)



DWT | Packet Lift | Adaptive Packet Lift | 3/2 Lifted Pyramid | 2/3 Lifted Pyramid

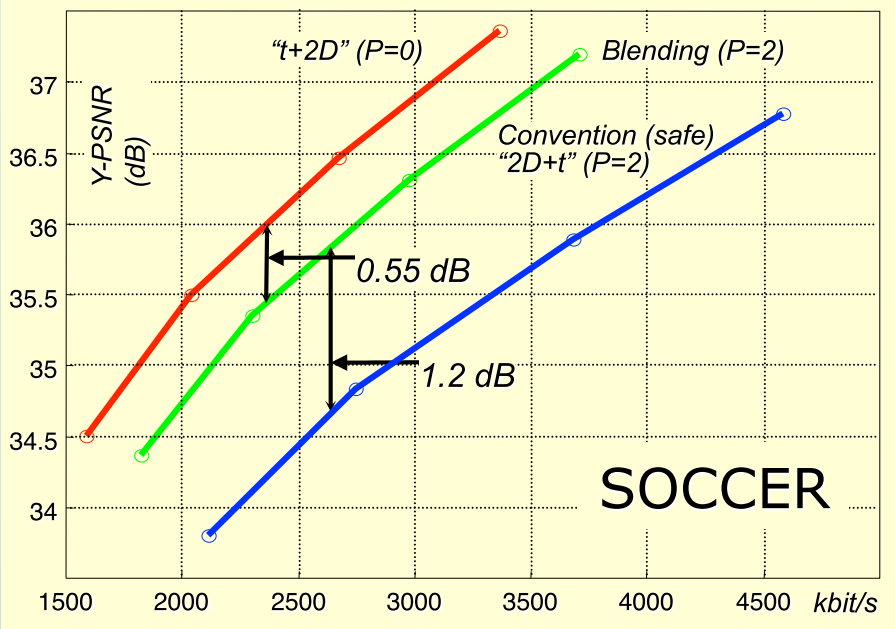
Spatial scalability – t+2D



- Temporal transform uses full spatial resolution
- At reduced spatial resolution
  - Temporal synthesis steps missing high-resolution info
  - If motion trajectories wrong/non-physical ➔ ghosting
  - If trajectories valid ➔ temporal synthesis reduces aliasing
    - less aliasing power (relative to 2D+t case)
    - aliasing content changes slowly over time

Adaptive Schemes – t+2D vs 2D+t

- Adaptively use hi-res info in low-res temporal lifting
  - to the extent that this is “safe” (from ghosting)



(Mehrseresht & Taubman, 2004)

- Could further reduce aliasing effects
  - by combining with adaptive energy exchange schemes

Aliasing suppression – t+2D

(Wu & Woods, 2007)

- Temporal transform performed at full res
- Spatial DWT applied to temporal subband frames
- High-pass subband samples “attenuated”
  - attenuation undone to reconstruct higher resolutions
  - reduces aliasing effects in low-res reconstructions
  - no loss of full-res coding efficiency
    - done through bit-plane shifting
- Attenuated subband samples get less bits
  - not just decoder-side post-processing
- Best with wavelet packet transforms
  - more control over frequency roll-off produced by subband sample attenuation

Summary of transform effects

- 2D+t pyramid schemes are simplest
  - but, redundant sampling hurts performance
    - especially at high bit-rates
  - lifting important for open loop pyramids
  - wavelets with energy exchange present an interesting alternative
- t+2D schemes always the most efficient
  - full resolution motion compensation
  - can produce ghosting at reduced spatial resolutions
  - t+2D DWT schemes produce aliasing at reduced resolutions
    - reduced by good motion models
    - still not clear that this is a real issue in practice
- 5/3 temporal transform superior to hierarchical B-frames
  - reduced quantization noise power
  - less noise/artefacts passed to lower temporal resolutions
  - but, dangerous with some 2D+t schemes
    - can damage low-temporal, high-spatial resolution
  - high quality motion is very important
  - adaptive schemes required to reep benefits “safely”

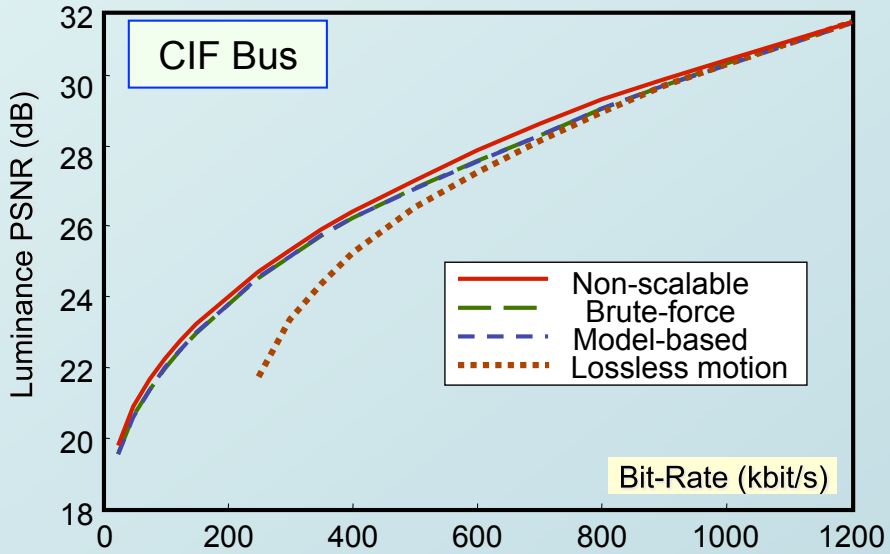
Beyond video

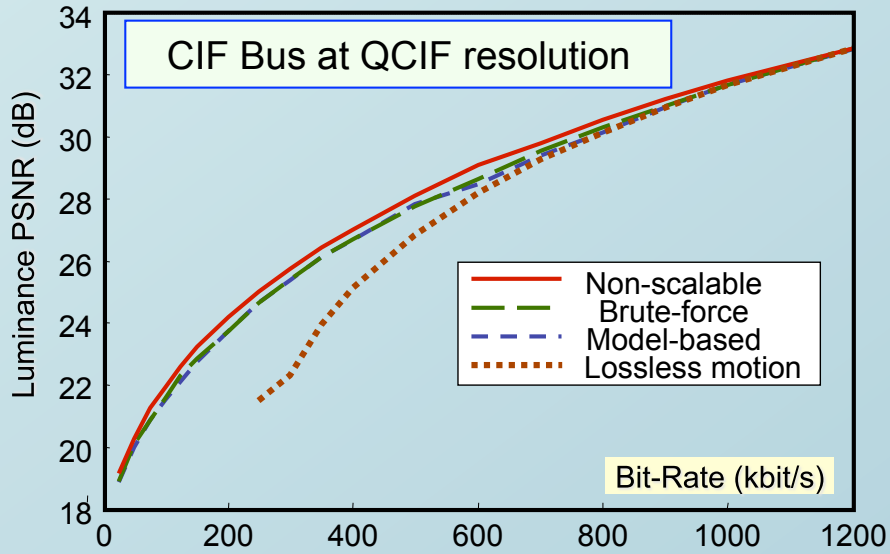
- Object-based video
  - MC shape-adaptive lifting (Liu, Ngan, Wu, 2007 & 2008)
- Scalable volume compression
  - MC lifting on slices (Taubman, Leung & Secker, 2002)
  - DC (disparity comp) lifting on volume views (Marcellin, Bilgin et al. 2008)
- Worth noting:
  - above schemes generally based on 3/4D DWT with 5/3 "temporal/ inter-view" lifting, using motion/geometry compensation
  - competitive with H.264, especially when motion/geometry smooth
- Light fields and free view-point video
  - DC (disparity comp) lifting on scene views (Girod, Chang, et al. 2003)
  - MC/DC lifting on views (Garbas, Fecker, Troger & Kaup 2006) (Garbas, Pesquet-Popescu & Kaup 2011)
- Scalable depth fields for 2.5D media
  - closely related to motion compression; see later

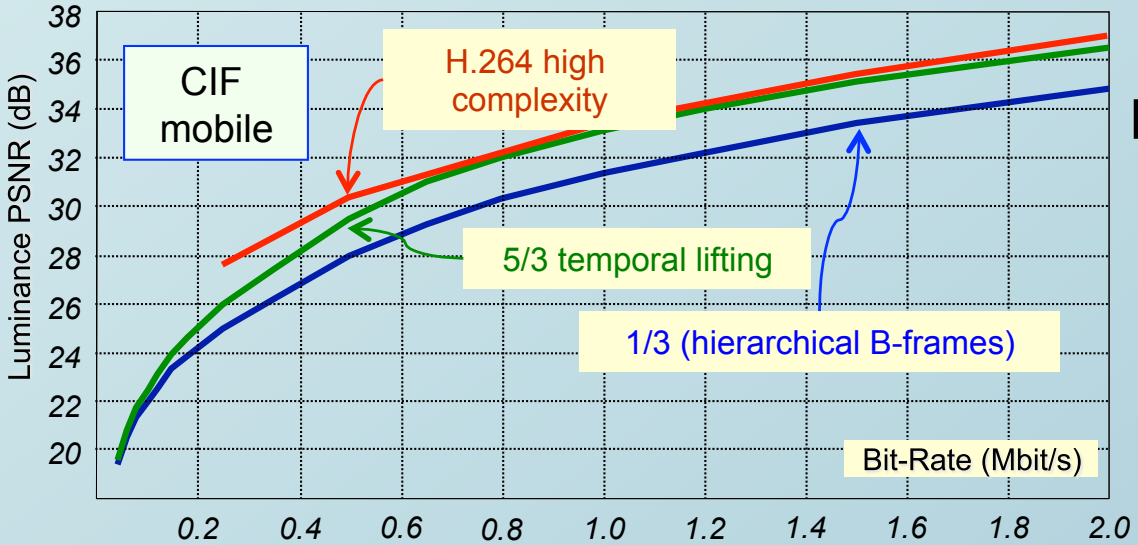
Motion for Scalable Video

- Fully scalable video requires scalable motion
  - reduce motion bit-rate as video quality reduces
  - reduce motion resolution as video resolution reduces
- First demonstration (Taubman & Secker, 2003)
  - 16x16 triangular mesh motion model
  - Wavelet transform of mesh node vectors
  - EBCOT coding of mesh subbands
  - Model-based allocation of motion bits to quality layers
  - Pure t+2D motion-compensated temporal lifting

Scalable motion – very early results







H.264 results

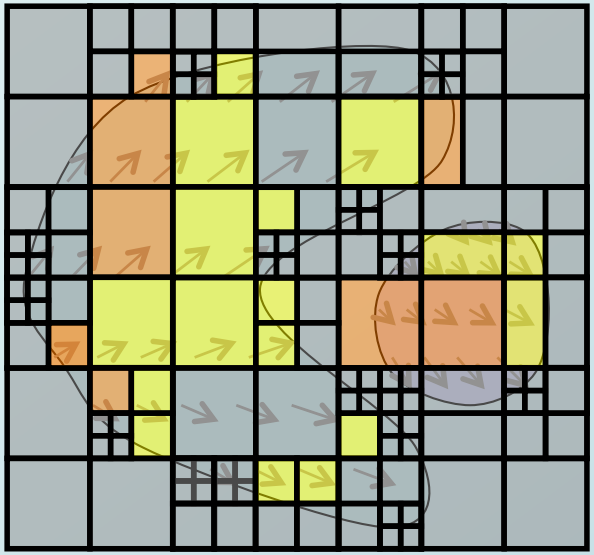
- CABAC
- 5 prev, 3 future ref frames
- multi-hypothesis testing
- (courtesy of Marcus Flierl)

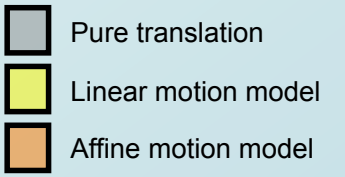
On the road to better motion

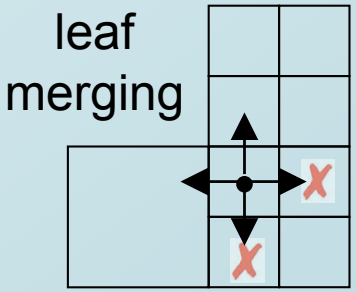
- Issues:
  - smooth motion fields scale well
    - mesh is guaranteed to be smooth and invertible everywhere
  - but, real motion fields have discontinuities
- Hierarchical block-based schemes
  - produce a massive number of artificial discontinuities
  - not invertible – i.e., there are no motion trajectories
  - non-physical – hence, not easy to scale
  - but, easy to optimize for energy compaction
    - particularly effective at lower bit-rates
- Objectives
  - minimize artificial discontinuities
  - encourage smooth models wherever possible
    - pure translation not generally sufficient

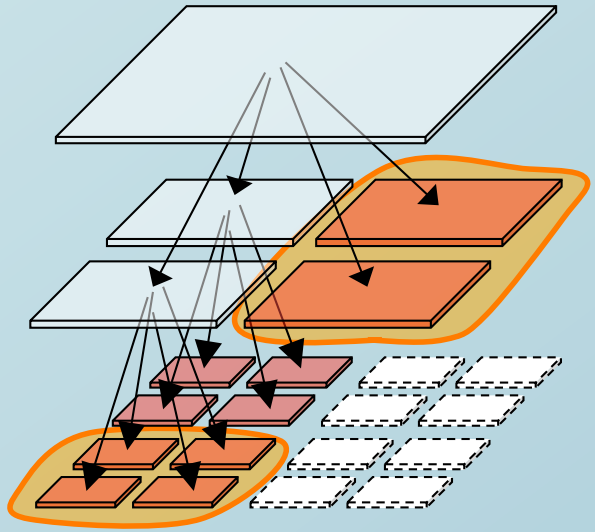
Block-based schemes with merging

(Mathew and Taubman, 2006)



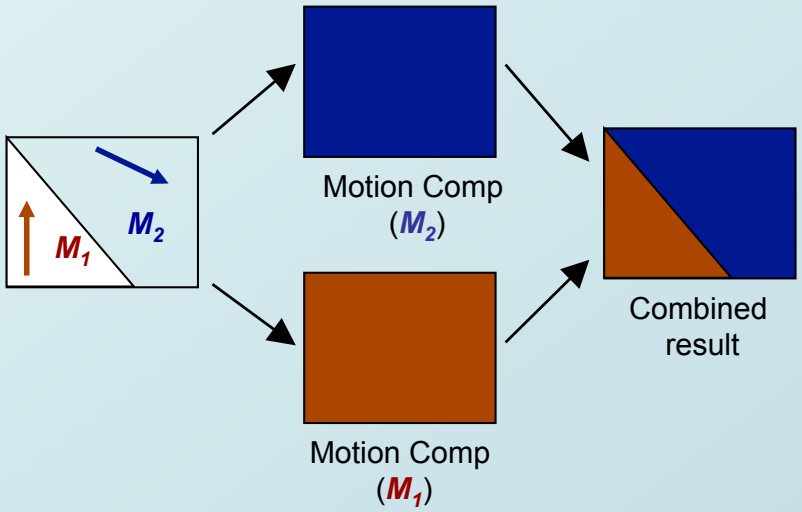






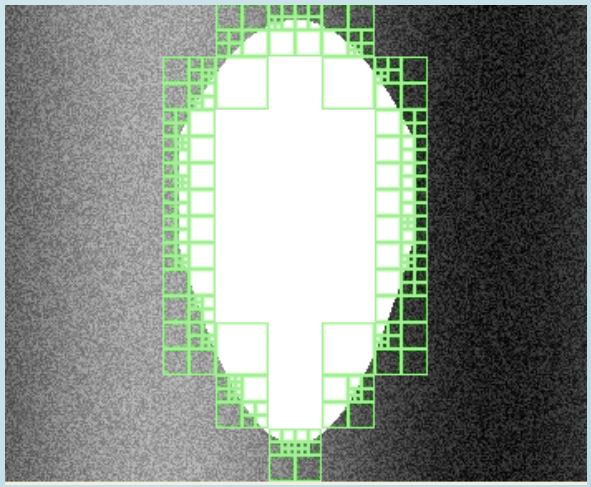
- Linear & affine models
  - encourages larger blocks
- Merging of quad-tree nodes
  - encourages larger regions and improves efficiency
  - merging approach later picked up by the HEVC standard
- Hierarchical coding
  - works very well with merging; provides resolution scalability

Boundary geometry and merging

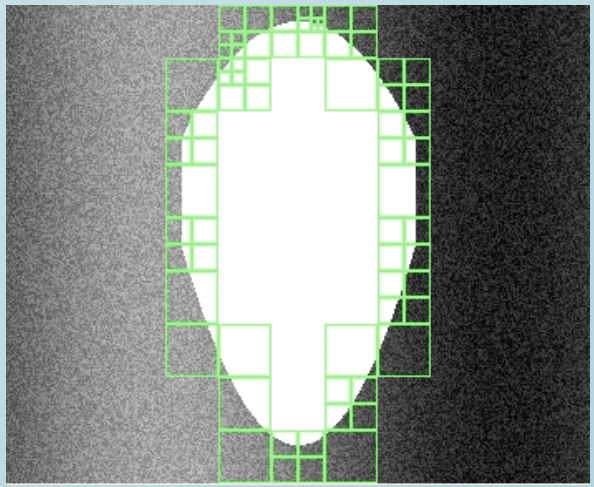


- Model motion & boundary
- No merging
  - Hung et al. (2006)
  - Escoda et al. (2007)
- With merging
  - Mathew & Taubman (2007)
  - separate quad-trees (2008)

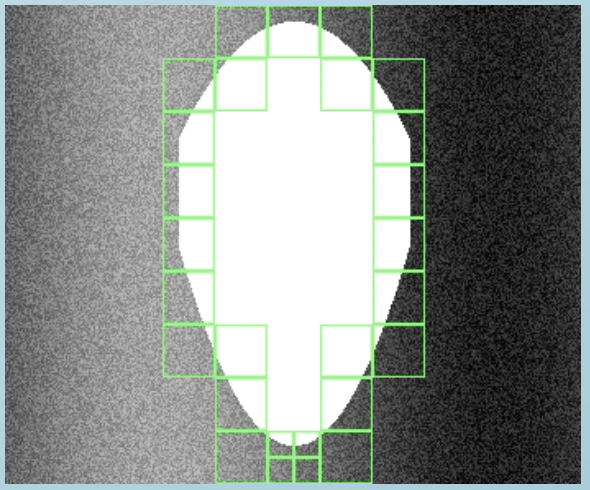
Motion only



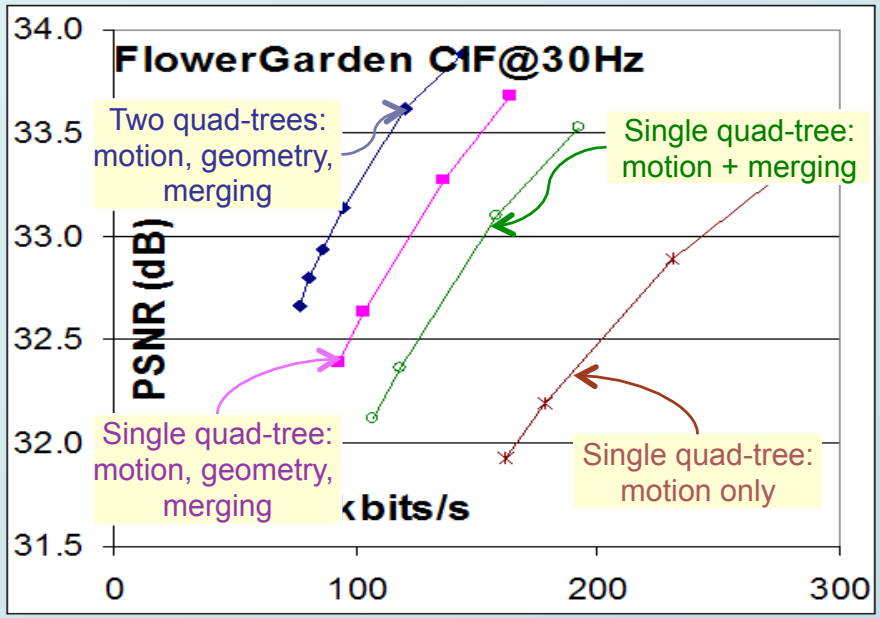
Motion + boundary

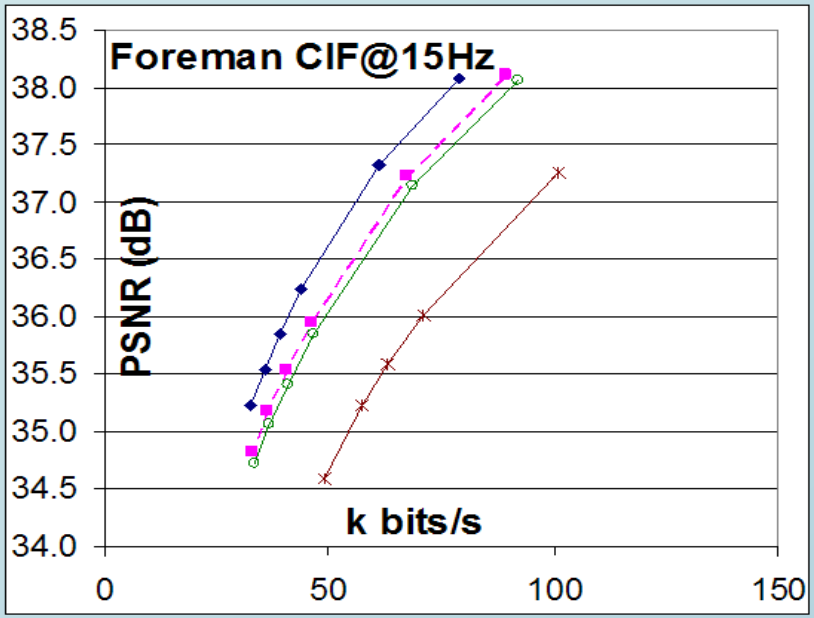


Separate quad-trees



Indicative Performance





- Things that reduce artificial discontinuities:
  - modeling geometry as well as motion
  - separately pruned trees for geometry and motion
  - merging nodes from the pruned quad-trees
- These schemes are practical and resolution scalable
  - readily optimized across the hierarchy

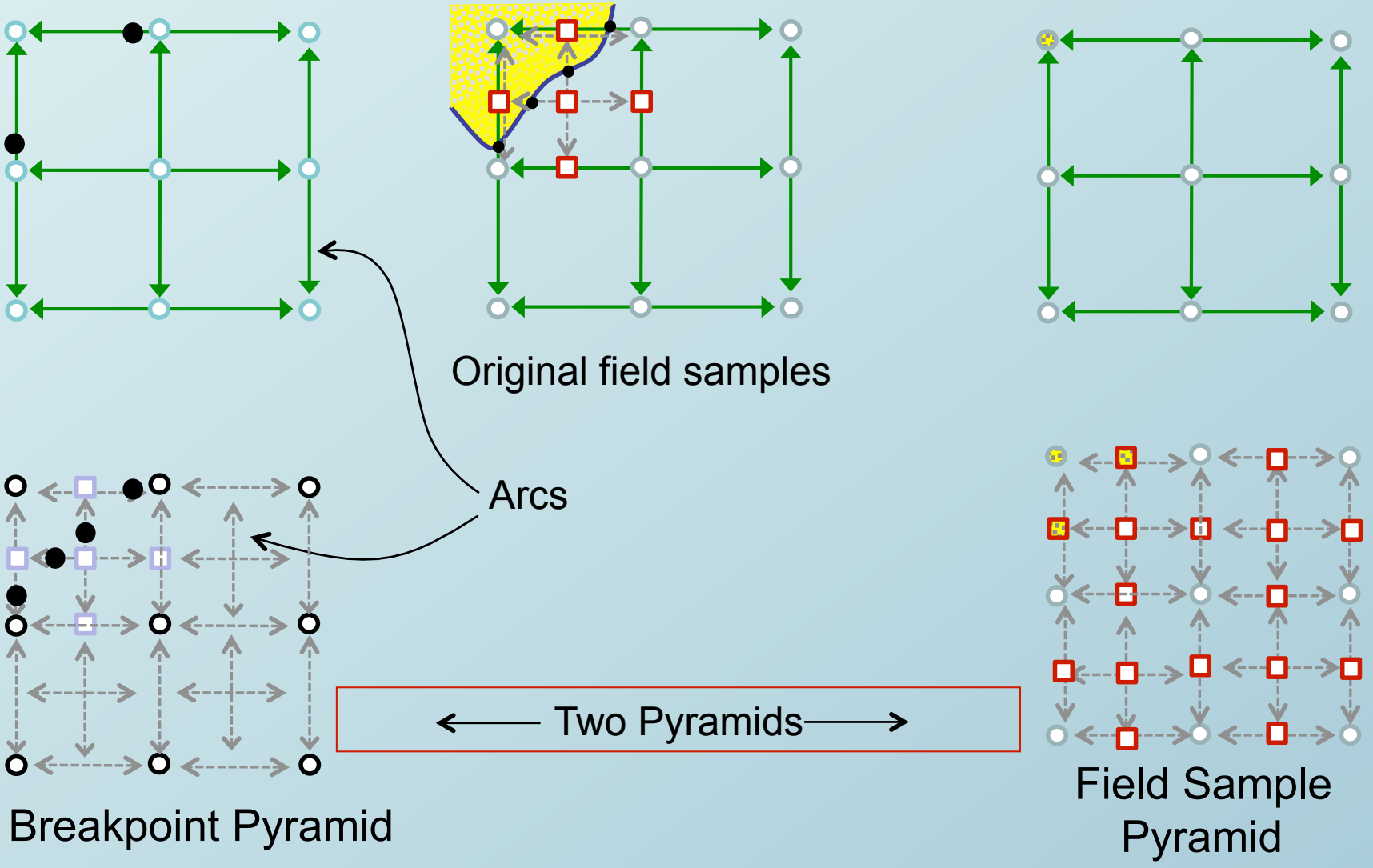
A new approach – currently implemented only for depth maps; very similar to motion maps

Highly scalable depth/motion coding

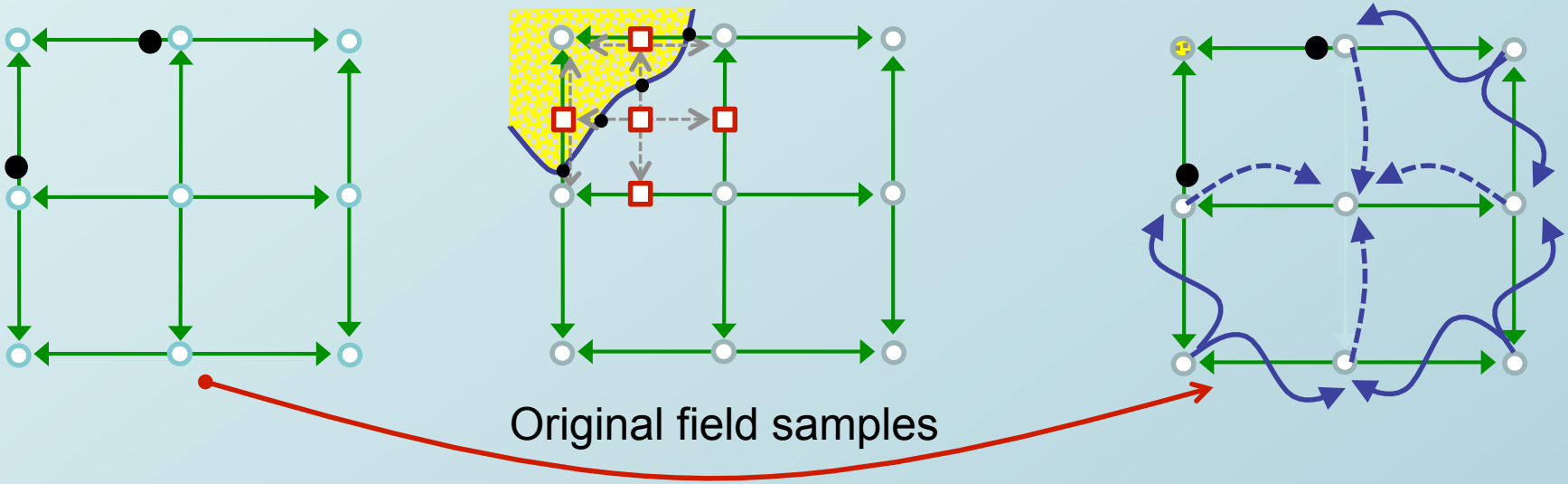
(Mathew, Taubman, Zanuttigh, 2012)

- No explicit segmentation
- No parametric models of boundaries
- Explicit signalling of discontinuities along “arcs”
  - Spatial hierarchy of arcs that may contain breakpoints
    - introduces resolution scalability to discontinuity field
  - Position of breakpoints on arcs successively refined
    - introduces quality scalabiity to discontinuity field
  - Breakpoint adaptive DWT of depth/motion field values

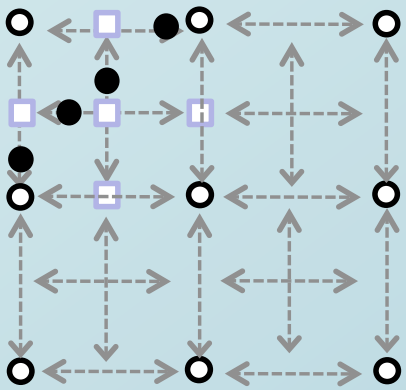
Field samples & Breakpoint pyramids



Breakpoint adaptive DWT – sequence of non-separable 2D lifting steps

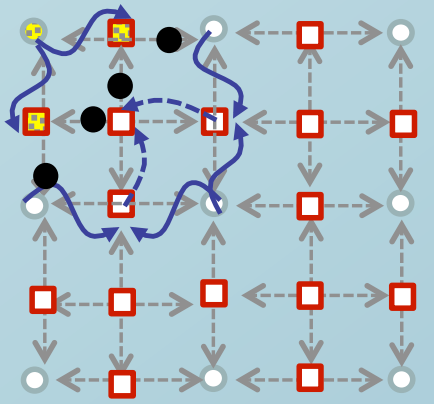


Original field samples



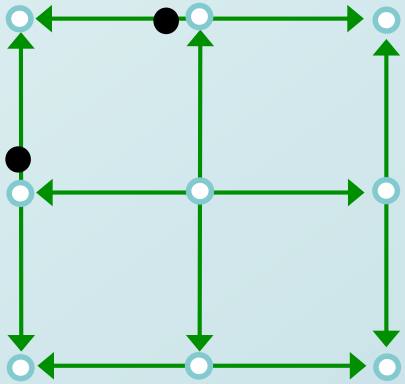
Arc Breakpoint Pyramid

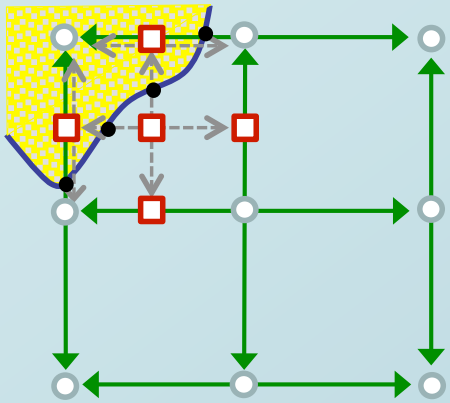
- Breakpoints drive an adaptive DWT
- Basis functions do not cross discontinuity along an arc
- Max of one breakpoint per arc
- Adaptive transform well defined



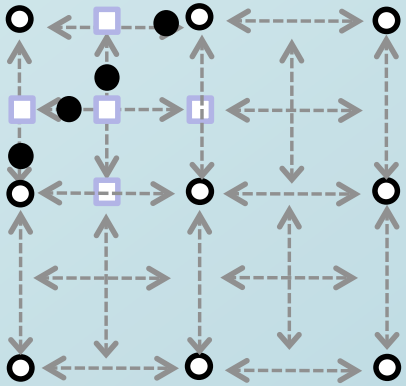
Field Sample Pyramid

Vertices & Induced Breakpoints





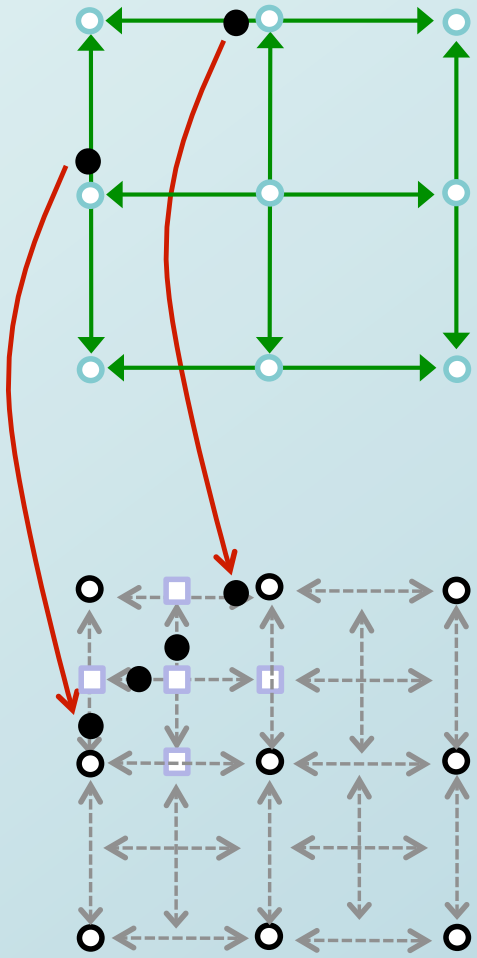
Original field samples

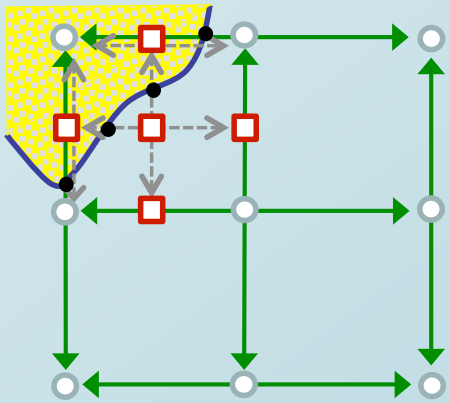


Arc Breakpoint Pyramid

- Only a subset of breakpoints communicated
- We call these "vertices"
- Vertices induce remaining breakpoints
- Breakpoints at a coarser resolution level can induce breakpoints on arcs at finer levels (recursive)

Vertices & Induced Breakpoints





Original field samples

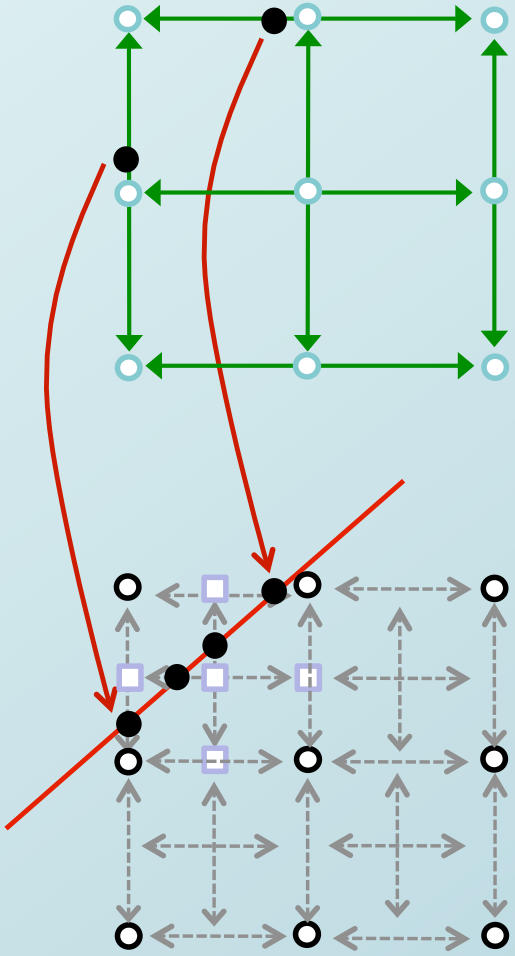
Inducing Policy

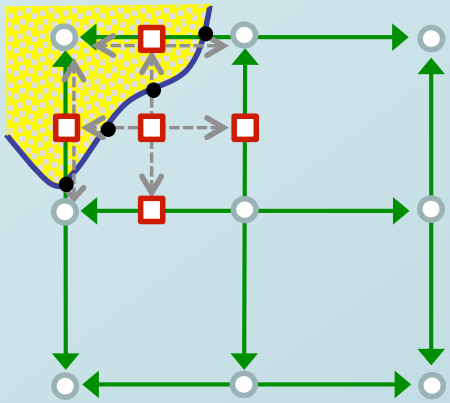
1. **Parent to child arc**
2. Inferred edge

- Arc breakpoint can induce breakpoints on its sub-arc
- Vertex on an arc overrides any induced breakpoints

Arc Breakpoint Pyramid

Vertices & Induced Breakpoints





Original field samples

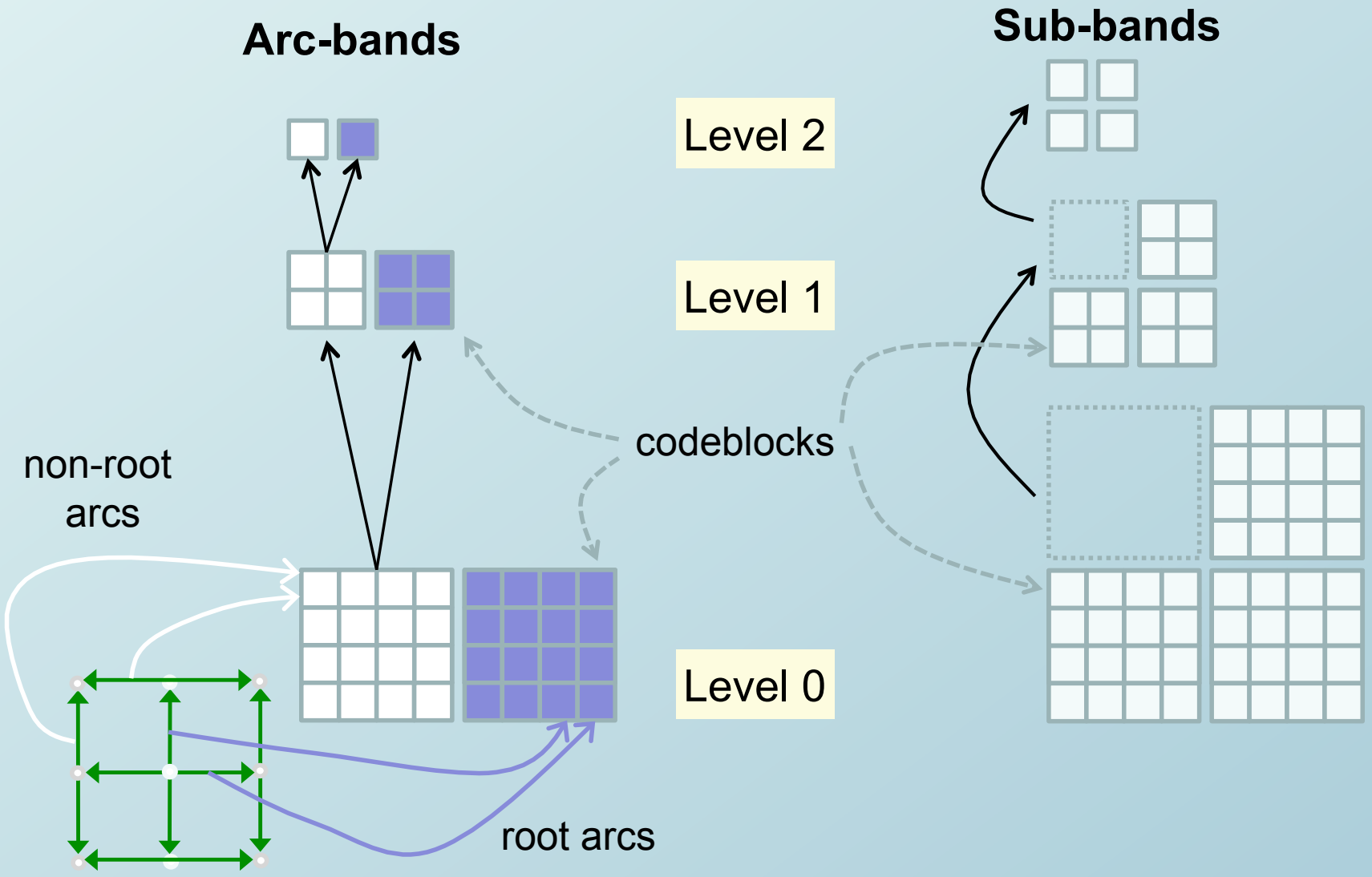
Inducing Policy

1. Parent to child arc
2. **Inferred edge**

- ❑ Breakpoints induced on “root arcs”
- ❑ Vertex on a root arcs overrides induced breakpoints
- ❑ Good for compression & scalable decoding

Arc Breakpoint Pyramid

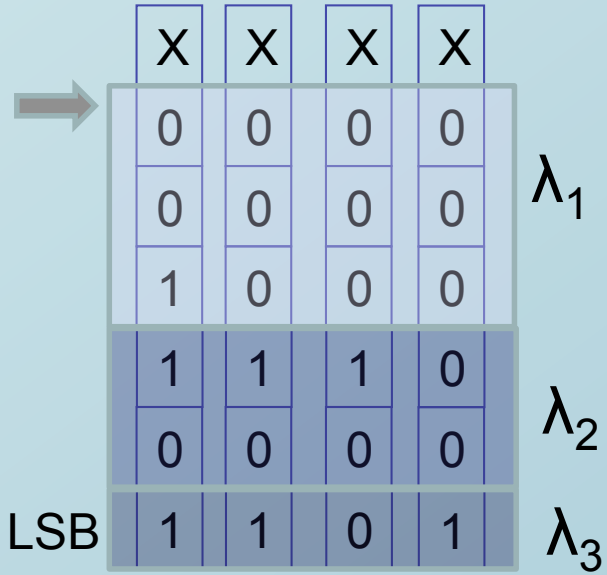
Sub-bands and Arc-bands



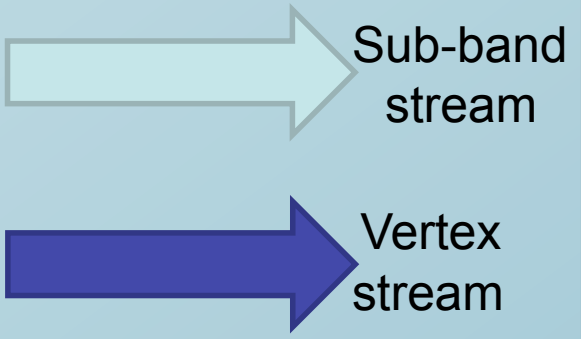
Embedded Block Coding

– for scalability and ROI accessibility

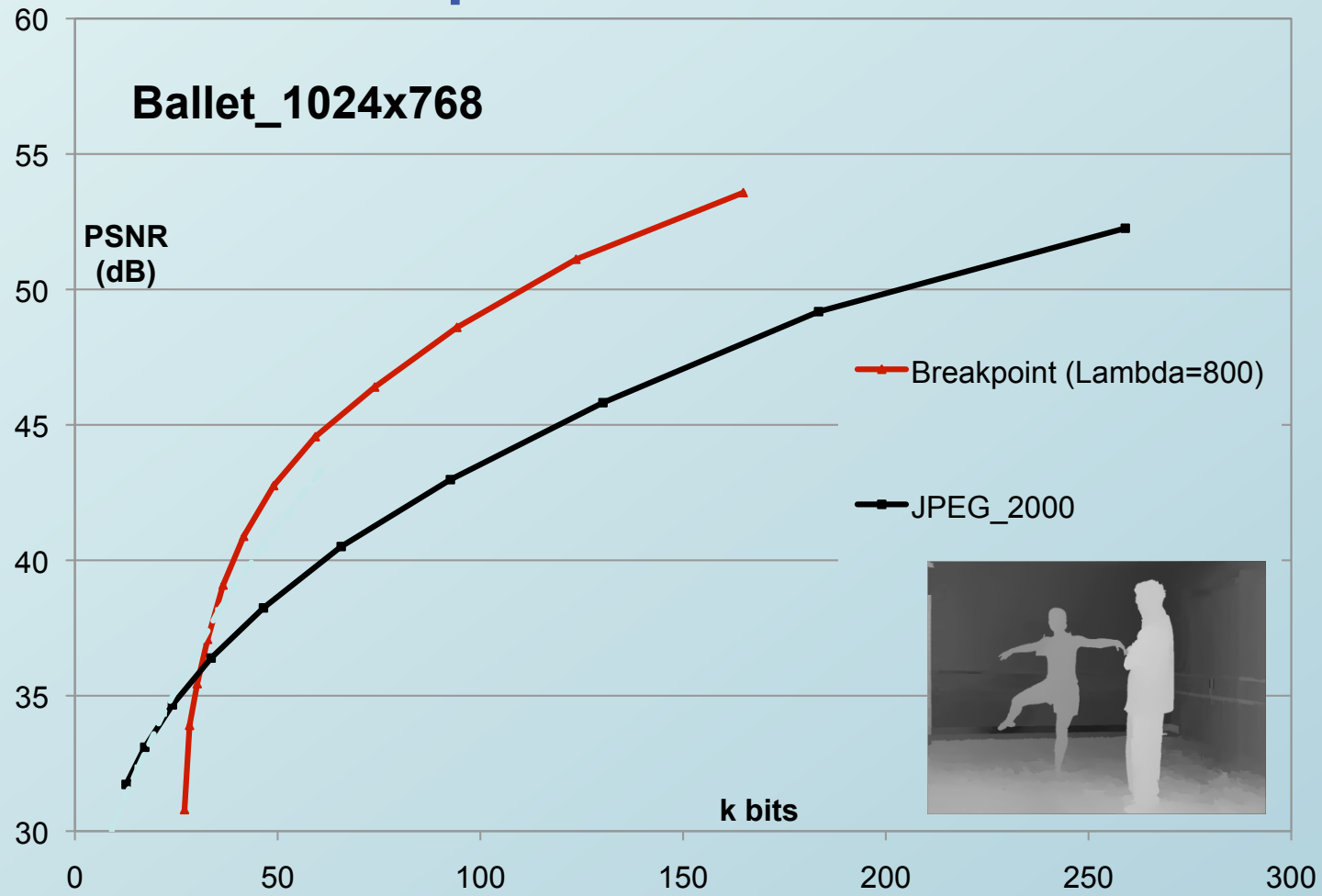
- Sub-band stream (field samples)
  - Sub-bands divided into code blocks
  - Coded using EBCOT (JPEG2000)
  - Bitplanes assigned to quality layers



- Vertex Stream
  - Arc-bands divided into code blocks
  - Coding scheme similar to EBCOT
  - Bitplanes refine vertex locations
  - Bitplanes assigned to quality layers

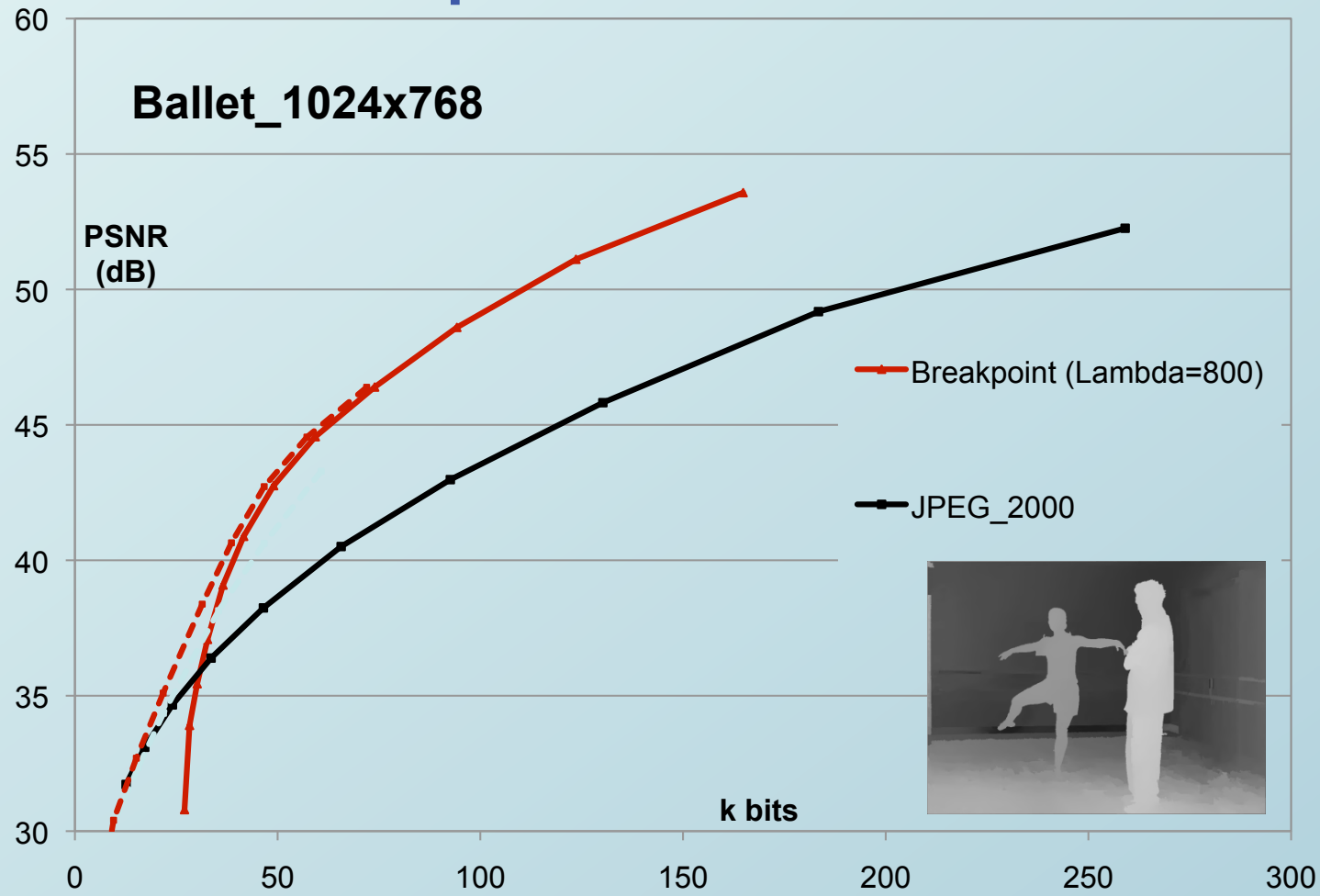


Indicative performance – depth coding



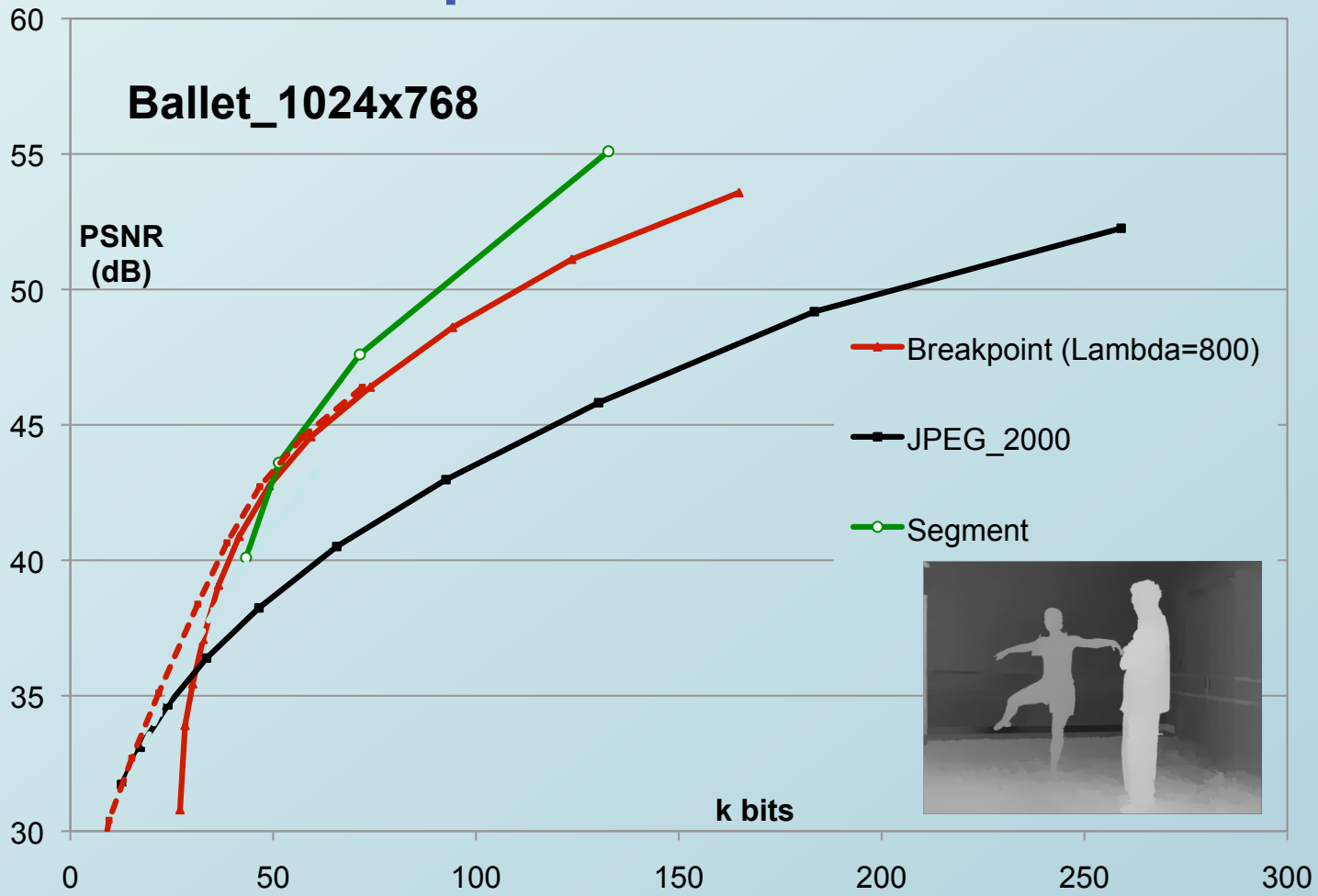
- Scaled by discarding sub-band quality layers only
  - vertex coding cost hurts low bit-rate performance

Indicative performance – depth coding



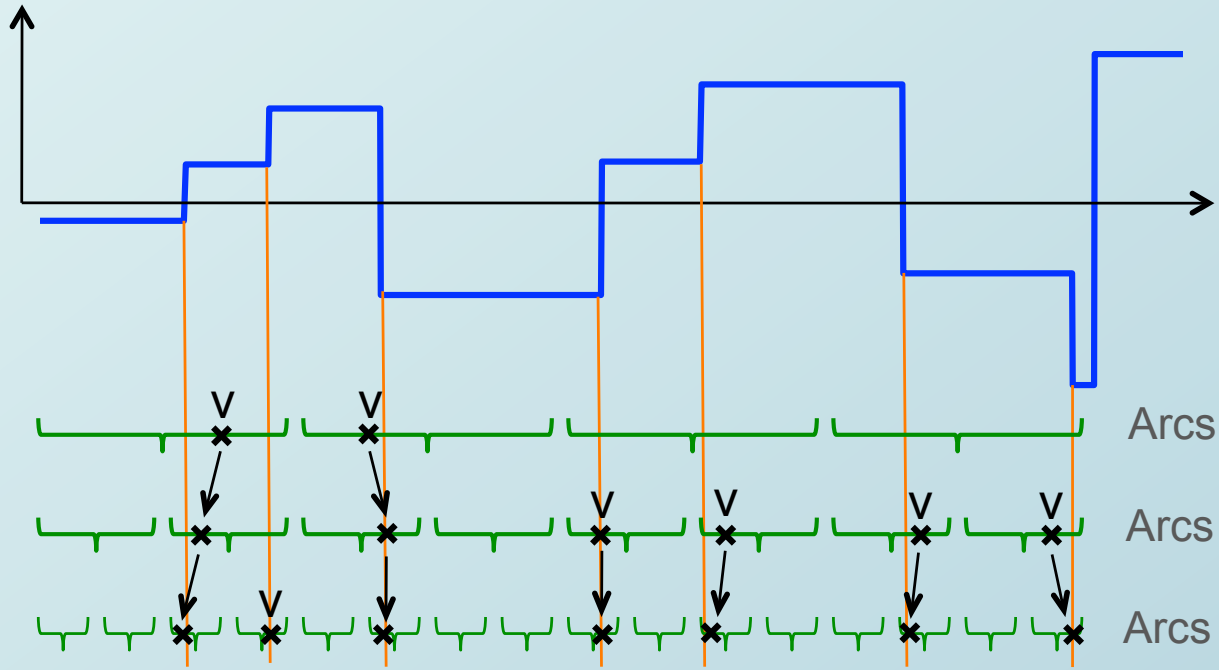
- Scaled by discarding sub-band and arc-band quality layers
  - fully automatic model-based quality layer formation
  - model-based interleaving of all quality layers for optimal embedding

Indicative performance – depth coding



- Compared with segmentation based approach (Zanuttigh & Cortelazzo, 2009)
  - not scalable; sensitive to initial choice of segmentation complexity

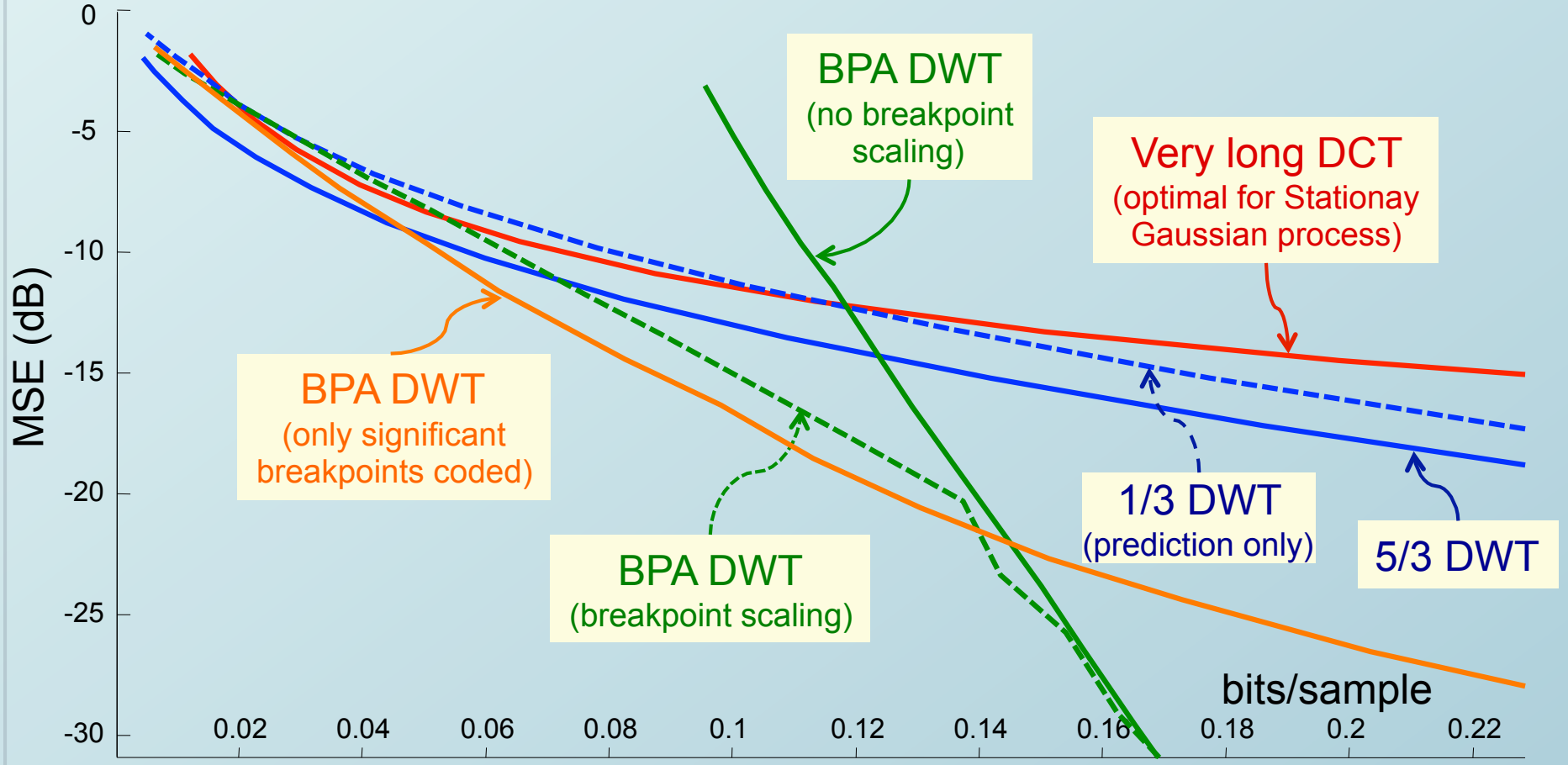
Scalable coding of sparse data



Model 1D process
- Stationary
- Marginally Gaussian
- Innovations (jumps)
  - memoriless
  - sparse

- Breakpoint adaptive DWT simple in 1D
- Breakpoints coded at vertices (v)
  - Successive bit-planes refine accuracy of breakpoint
- Model based quality layering of vertex bit-planes
  - Discard layers at low bit-rates based on D-R slope

Scalable coding of sparse data



- High rate asymptotic behaviour affected by sparsity preservation
- Low rate behaviour dominated by breakpoint discard process
  - can be shown to have comparable R-D properties to 1/3 DWT

Related research

- Motion compensated orthogonal transforms

  (Flierl & Girod, 2006 & 2007) (Flierl 2009) (Liu & Flierl 2012)

  – build temporal transform from a sequence of stages
    - each stage transforms a small set of pixels (e.g., 2 or 3)
    - stages incrementally orthogonalized, based on motion field
  – follow with MCOT-adapted spatial "wavelet" transform and EBCOT (as in JPEG2000)
- Lifting transforms on graphs for video coding

  (Martinez Enriquez & Ortega, 2011)

  – model video as graph with temporal & spatial weights
  – "wavelet-like" lifting on paritioned graph
- Above schemes support quality scalability
  – but visual properties of reduced scales not considered

Summary

- Scalable image compression is very effective
  - fully embedded, no loss in efficiency, extremely flexible
- Prediction alone is sub-optimal for video
  - produces more quantization noise than transform approach
  - fails to progressively clean noise from high res, high fps content
- SVC standard
  - has probably reduced the intensity of research
  - but many fundamental issues remain to be explored
- Lots of interesting tools have been developed
  - motion-compensated lifting; lifted spatial pyramids; adaptive inter-resolution blending; motion compensated orthogonal transforms; …
- Breaking away from block-based motion is key
  - need to understand discontinuities as innovation process
    - scalability needs to address the R-D properties of this process
    - block models are riddled with artificial discontinuities

Dependent research directions

- Perceptual models for scalable video
  - perceptually optimize allocation of bandwidth
    - e.g., (Leung & Taubman, CSVT 2009)
  - spatial details vs. temporal details vs. quant. artefacts
  - conclusions are codec dependent
    - see, e.g., (Lee, De Simone, Ramzan, Zhao, Kurutepe, Sikora, Ostermann, Izquierdo & Ebrahimi, ACM Multimedia 2010)
    - room for much more research, inc development of good models
- Robust communication of scalable video
  - Lossy channels, real-time constraints
  - Explored in many different contexts
  - PET-based schemes are appealing for open-loop scalable coders with packet erasure channels
    - e.g., “Limited-Retransmission-PET” (Taubman & Thie, 2005)