The case for scalability support in version 1 of Future Video Coding

Stephan Wenger
Vidyo, Inc.

The case for

multiple representations of the same content,
at the same presentation time,
decodable at different computational complexity / bitrate points,
that use prediction between the various representations

in version 1 of Future Video Coding

Don’t disregard Video Conferencing

- Multi-billion $$$ industry (yawn)
- Video conferencing client on every smartphone, tablet, PC, dedicated Vconf unit, …
- And, those clients are being used!
  - Skype, Hangouts, enterprise uses
- Multipoint also more and more common
- Not justifiable anymore to call Video Conferencing a “niche”

Requirements for Video Conferencing

- Error prone environment (Internet over heterogeneous access links)
- Low glass-to-glass delay
  - 100ms desirable, more than 250ms starts hurting
- Requires source coding based error resilience
  - End-to-end repair of IPPP stream not possible
  - Temporal scalability is tool of choice today
  - (allows repair of temporal base layer through retransmission/FEC with reasonable overhead)

Multipoint Video Conferencing

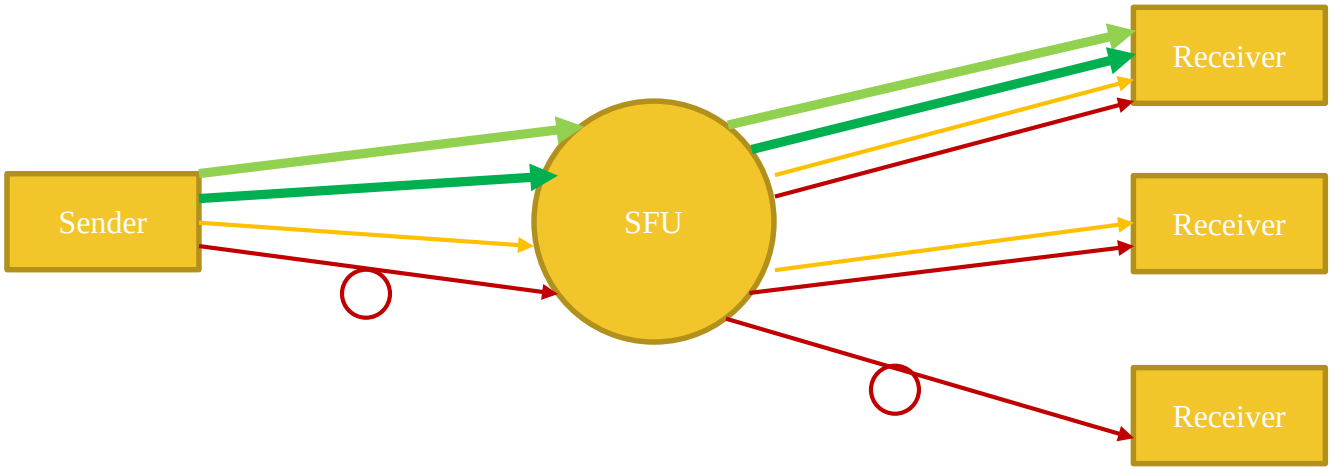
Vidyo

- Heterogeneous endpoint population
  - Screen size (cellphone through Telepresence room)
  - Connectivity (crappy 3G, and hotel room DSL through Gigabit link to backbone, roaming)
  - Error rates (practically error free through 5+%)

- Need different representations
  - Resolution/Fidelity to adapt to connectivity, computational complexity, and screen size
  - Possibly also to adapt to error rates?

Vidyo’s architecture

- (Not only ours... centralized with SFU quite common now)

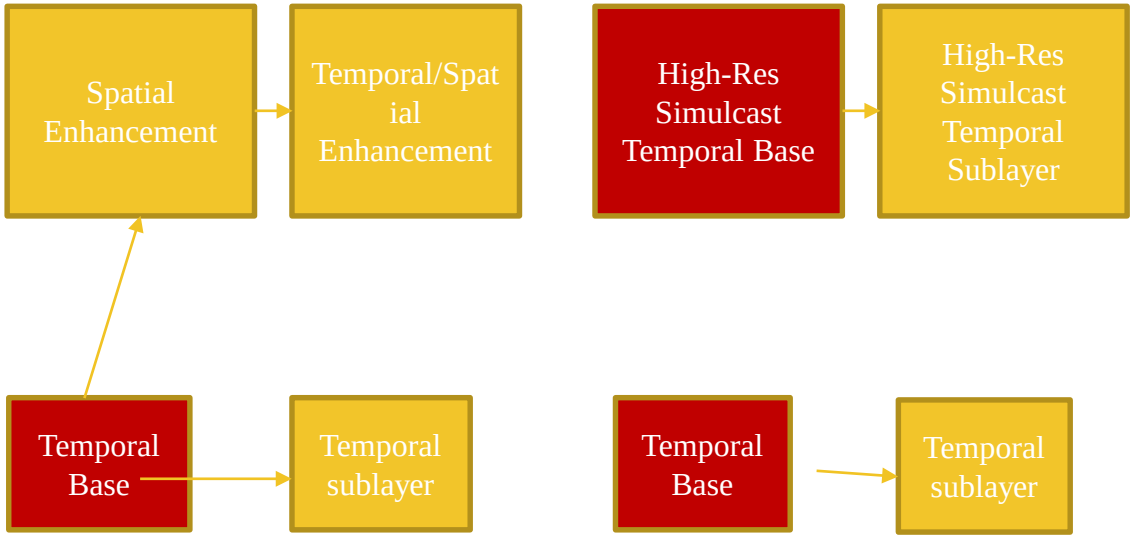


Technical Requirements

- **The case for multiple representations of the same content**
  - Needed to support heterogeneous receiver population
- **at the same capture / presentation time**
  - Bitstream requirement on the sender side
  - Does not necessarily imply same coding time
- **decodable at different computational complexity / bitrate points,**
  - Needed to support multipoint heterogeneous receiver population
- **that use prediction between the various representations**
  - Needed to keep bitrate on the sender link within reason

Prediction between various representations...

- Well-known benefit of inter-layer prediction in error-free case
  - Bandwidth savings, especially for intra
- Additional benefits for error prone use cases when used in combination w/ temporal scalability

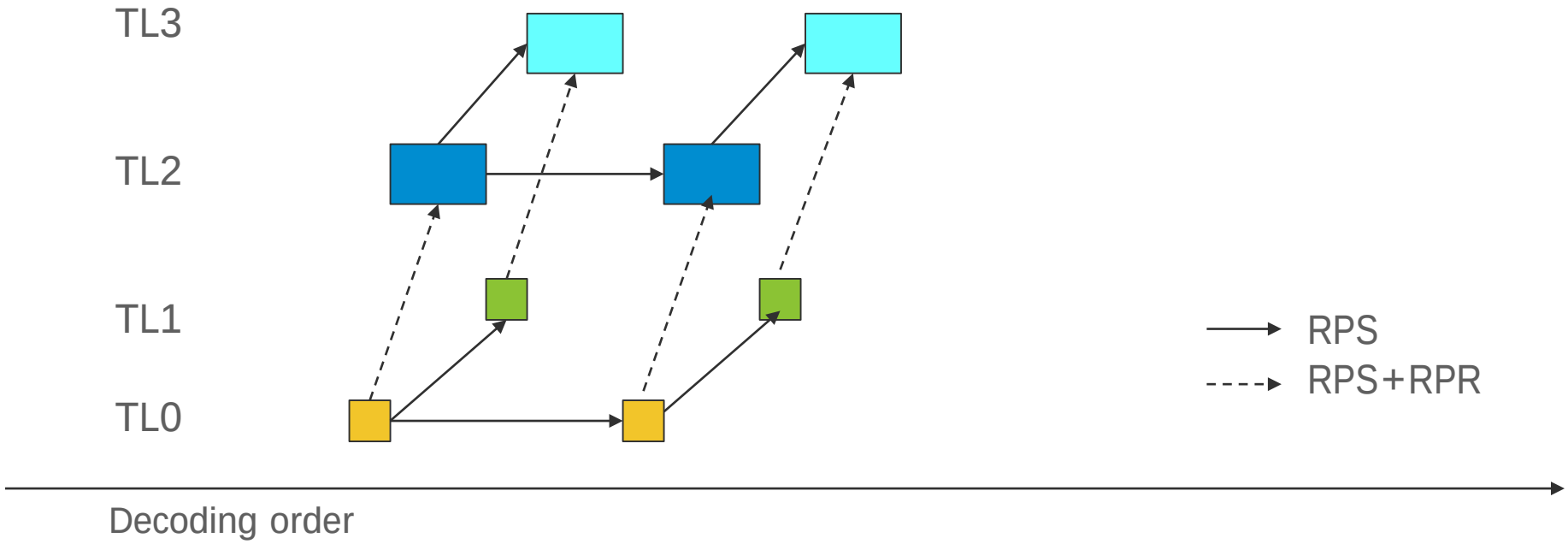


- Yellow: unprotected
- Red: protected by FEC/retransmission

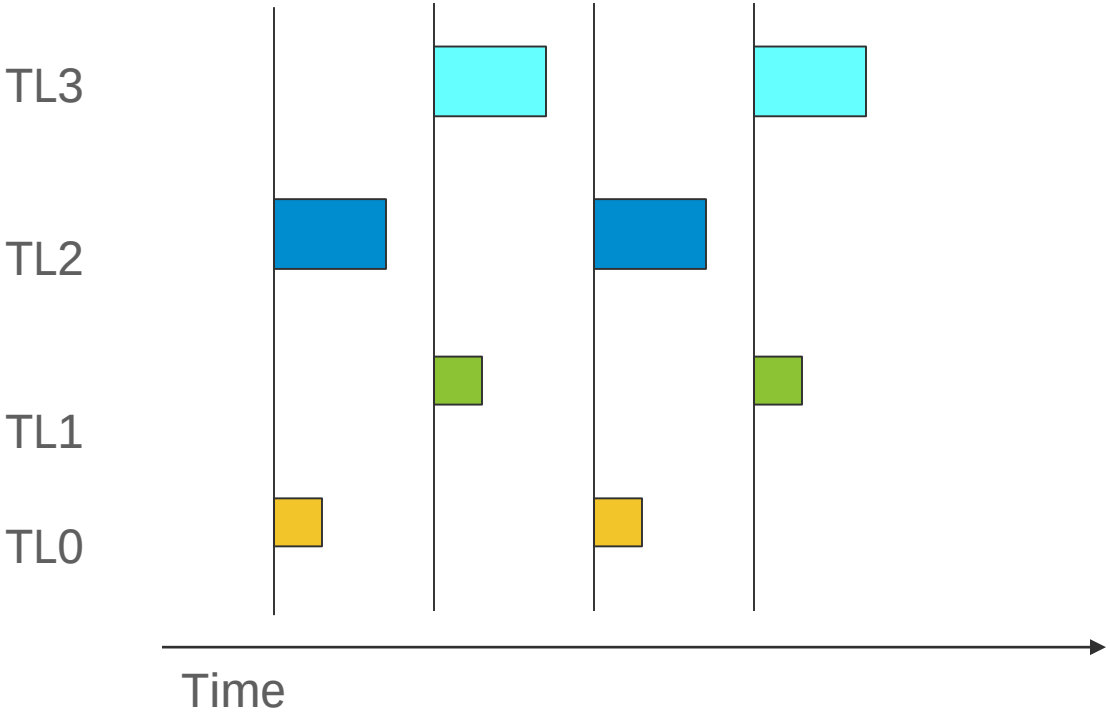
Technologies that fulfill requirements

- ~~Classic spatial scalability~~
- Combination of temporal scalability and reference picture resampling (RPR)
  - RPR allows for creating different representations w/ different resolutions, and is known from H.263 Annex P, JCT-VC proposals by Cisco among others
  - Temporal scalability (reference picture selection) allows to predict between pictures of the different representations, and also from other representations
  - Need metadata (SEI or normative) to identify pictures of different representations
  - Can be implemented in single traditional coding loop
  - Implemented by Vidyo in context of VP9 and its payload format
- There may be others, let's be inventive :-)

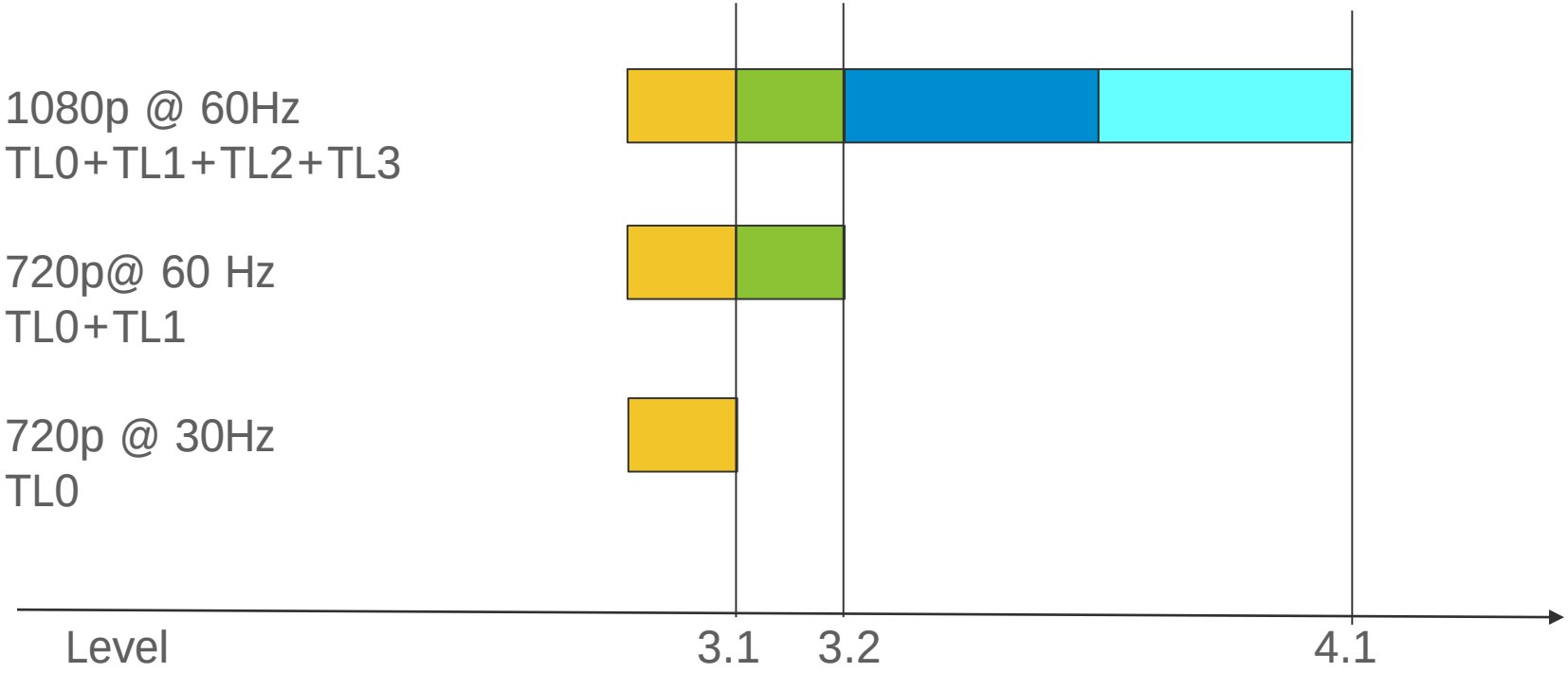
Prediction Structure



Capture/Presentation time



Levels to signal bitstream complexity



Why version 1?

- Video conferencing early adopter, mass user, and not niche
- Consensus for inclusion of traditional spatial scalability into version 1 is admittedly unlikely, and we are NOT asking for it
- There seems to be at least one way (as outlined on the previous slide) that does not require the allegedly onerous aspects of spatial scalability
  - We are, of course, open for other ways
- All we ask now is NOT to specifically rule out version 1 work towards the support of our application (as defined by our technical requirements)

Thank you