High Quality Video Transcoding in Data Center

Jensen Zhang

Sep. 2019

High Quality Video Transcoding in Data Center

What's the current Status of Data Center for video?

- Explosive growth of different kinds of video streams
- Compute requirements skyrocketing
  - More complexity video codecs formats, higher video resolutions
  - CPUs are too slow for video transcoding by software, especially for live video
- Huge Demands for better economics

Video Acceleration Overview

- Today market is dominated by high-powered x86 servers for video processing , servers struggle with video apps /new codecs and high resolution

- Huge growth video PUSH forward alternate architectures, but still not saving enough
  - NVidia NVENC/NVDEC - Hardware based codec engine
  - Intel Hardened (QSV)- using consumer GPU with hardened video engine to achieve higher density , Intel VCA2 PCIE Card
  - Xilinx VU9P PCIe card- FPGA integrates H264/H265/VP9 codecs

- Giant SNS company like FB Requires ASIC to save much more cost !!!

Huge demands require ASIC solution to solve the troubles

▲ Huge demands require ASIC solution

- ▶Strong requirements by internet company, can't to wait
- ▶Server company, including chip design, OEMs
- ▶FPGA company, AI company , etc.

▲VeriSilicon build up Video Transcoding Solution to solve the troubles

- ▶Excellent codec IPs work for Data center and Edge Server
- ▶Total solution with BOTH HW and SW

VeriSilicon leading video transcoding IP & customized ASIC

CPU vs Video transcoding ASIC

6X HEVC 4K Processing

1/13 Power Consumption

Much Smaller Size

World Leading Video Product

VeriSilicon

Hantro Video IP Track Record

Hantro
Multi-Format Video Codec

- **Multi-generations of Hantro encoders and decoders**
  - More than 100 licensees
  - Billions of shipped devices
- **Market leader with success in multiple market segments:**











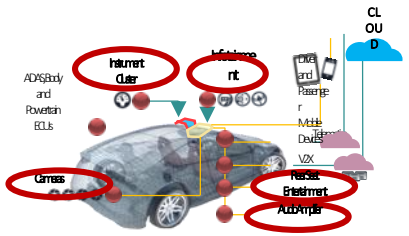
VeriSilicon

VeriSilicon Technology in Edge Device, Edge Server and Cloud

Edge Server

Video Transcoding
Pixel Compression
High Performance Computing

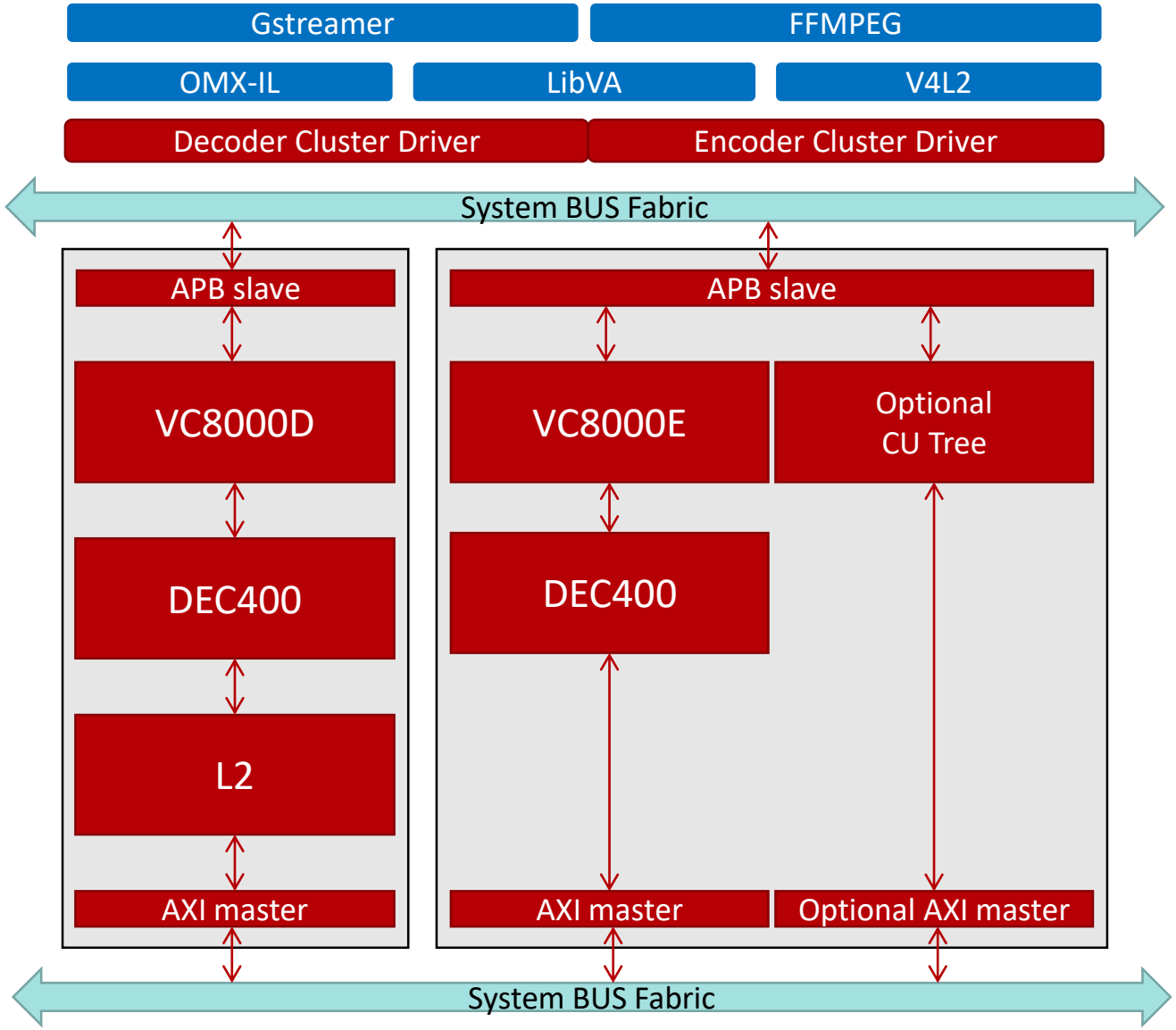
Edge Device

Surveillance	AR/VR Wearables	Smart Home, Vision, Voice	Automotive
		 	

Strengths of the Solution

VeriSilicon

Easy Integration as a Whole Solution



Integrated Decoder Cluster

Ready software and hardware integration and configuration

VC8000D: VeriSilicon multi-format decoder IP: H.264, H.265, VP9, AVS2, JPEG and legacy formats
DEC400: VeriSilicon system-adaptive frame compression IP
L2: Data cache and burst shaper for DRAM efficiency

Integrated Encoder Cluster

Ready software and hardware integration and configuration

VC8000E: VeriSilicon multi-format encoder IP: H.264, H.265 and JPEG
DEC400: VeriSilicon system-adaptive frame compression IP
CU Tree: Optional hardware for 2-pass encoding analysis

Transcoding Slice

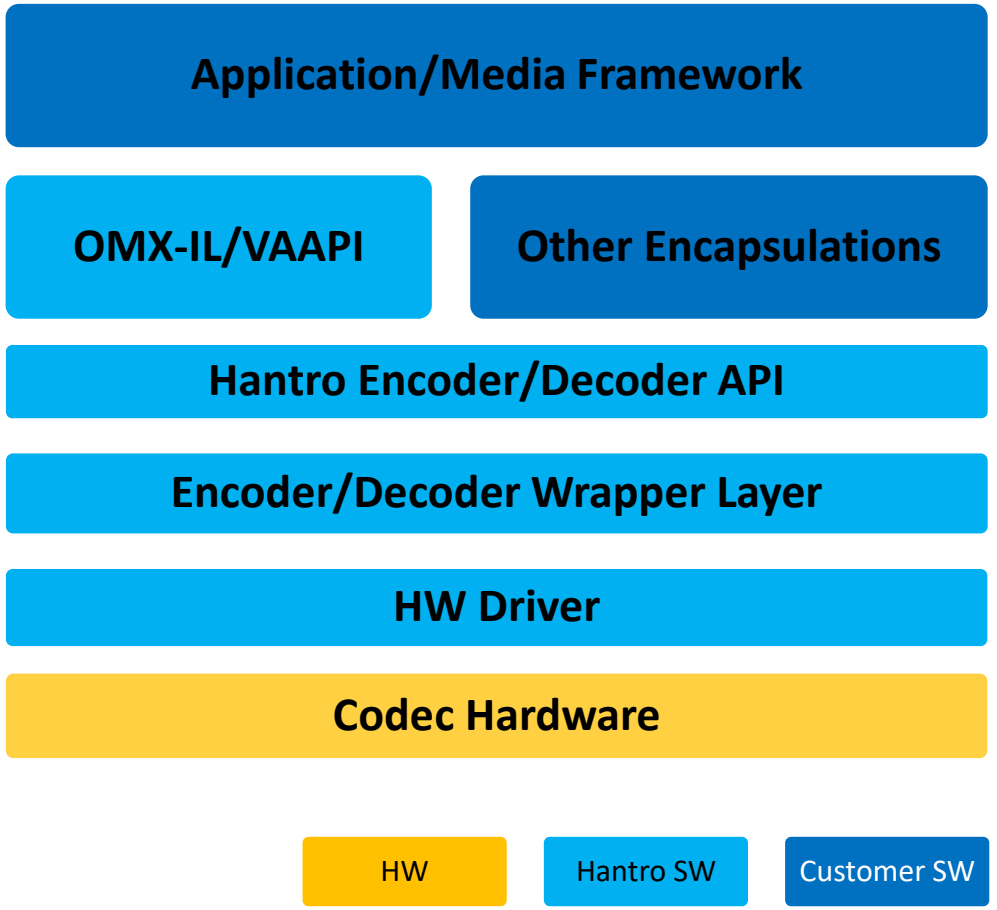
Decoder cluster + encoder cluster optimized for transcoding

- Optimized transoding data paths
- Optimized transcoding operations
- FFMPEG and Gstreamer ready solution

VeriSilicon

Ready Software library support

- Native Encoder/Decoder API are provided to fully explore the HW features;
  - Small CPU load for full HW algorithm.
  - Porting to different CPU: ARM, MIPS, PowerPC, C51.
  - Optimized according to HW flow.
  - Multi-core supported.
  - Multi-Instance support of interleave working for different format or resolutions.
- OMX-IL or VAPPI(libva/libdrm) components provide standard interface to help media framework integration easily;
- All software is provided as source code.



Power & Area Efficient ASIC Solution

- 100% ASIC design in the high Performance Decoding & Encoding video IP products

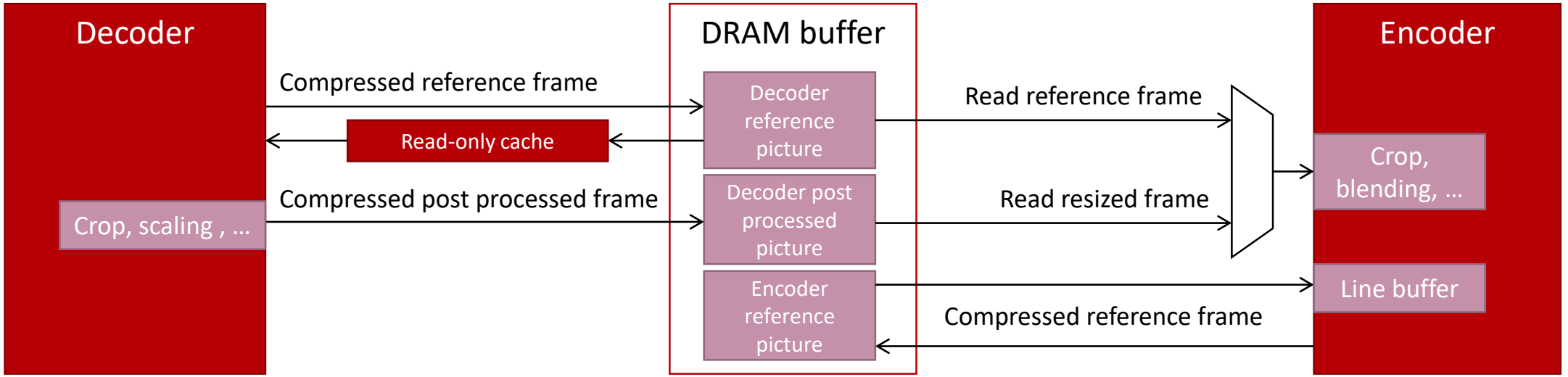
- Low area cost

4K60 10-bit H.264 & H.265 configuration area at 16 nm (mm^2)							
Decoder Cluster				Encoder Cluster			Transcoder
VC8000D	DEC400	L2	Total	VC8000E	DEC400	Total	Total
1.05	0.16	0.23	1.44	3.39	0.12	3.51	4.95

- Low power consumption

4K60 10-bit H.264 & H.265 configuration power consumption at 16 nm (mW)							
Decoder Cluster				Encoder Cluster			Transcoder
VC8000D	DEC400	L2	Total	VC8000E	DEC400	Total	Total
230	12	22	264	532	11	543	807

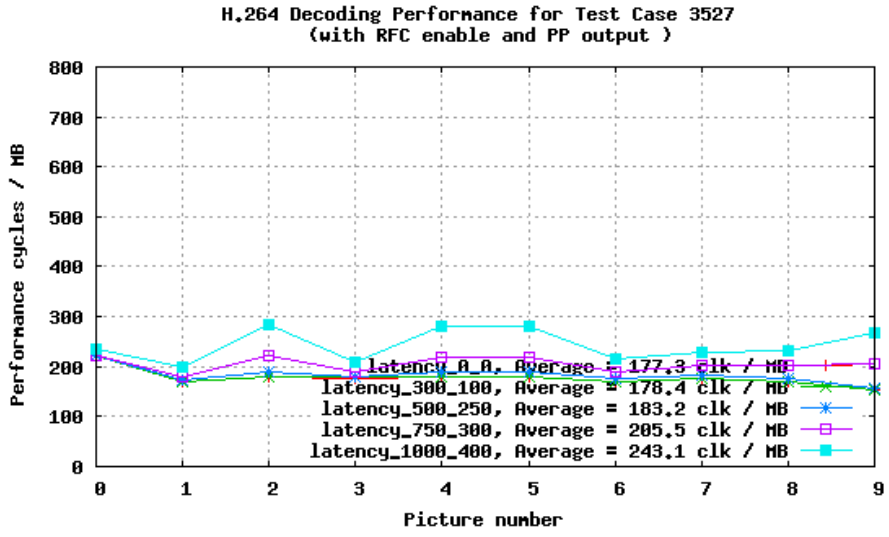
Low DRAM Bandwidth Requirements

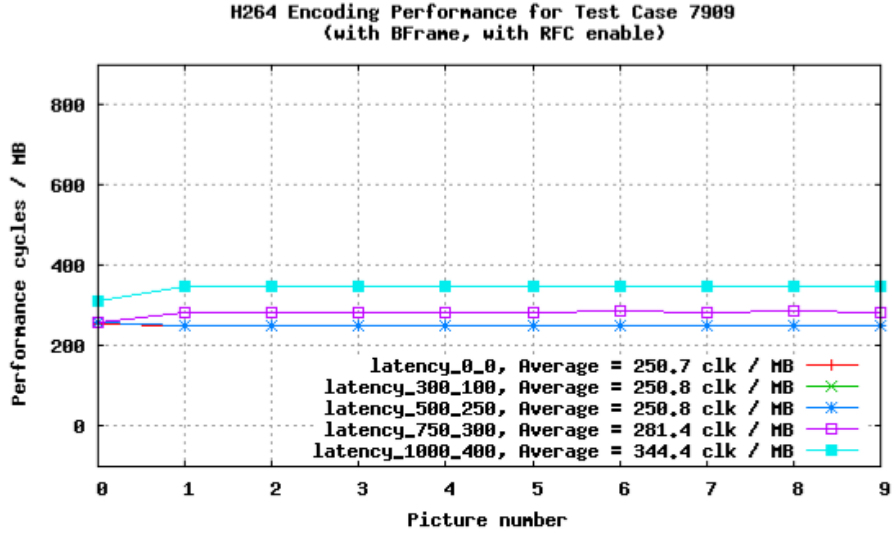


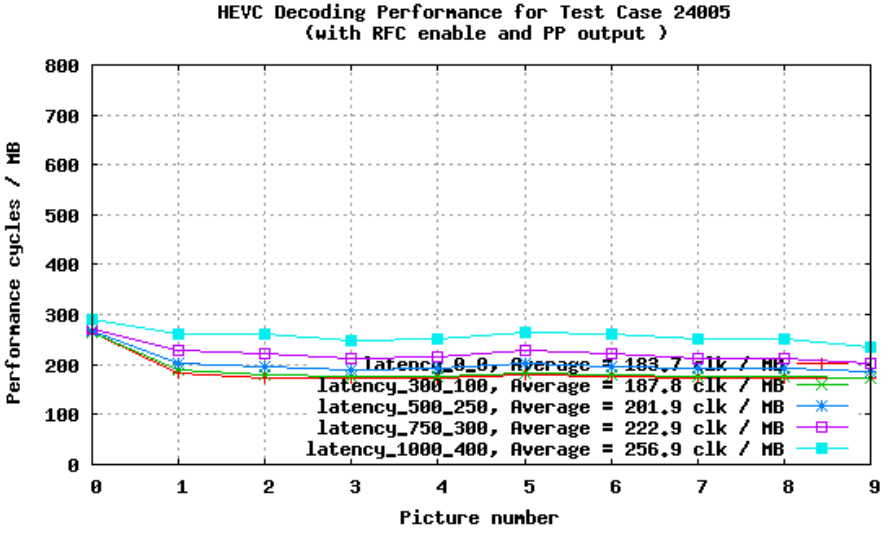
Bandwidth saving technology applied everywhere		
Decoder Cluster	Encoder Cluster	Transcoder
• All frames are compressed: saving 45~55% • >90% bursts are aligned: NO overhead • Configurable L2 cache size for reference frame, saving 0.8 ~ 1.6 GB/s	• All frames are compressed: saving 45~55% • >90% bursts are aligned: NO overhead • Configurable line buffer for reference frame, saving 1.2 ~ 2.4 GB/s	• Encoder directly read decoder reference frame • Crop and down scaled output from decoder • Blending in encoder input
Typical bandwidth: 2.2 GB/s	Typical bandwidth: 3.49 GB/s	Typical bandwidth: 5.69 GB/s
Ultra saving bandwidth: 1.4 GB/s	Ultra saving bandwidth: 2.4 GB/s	Ultra saving bandwidth: 3.8 GB/s

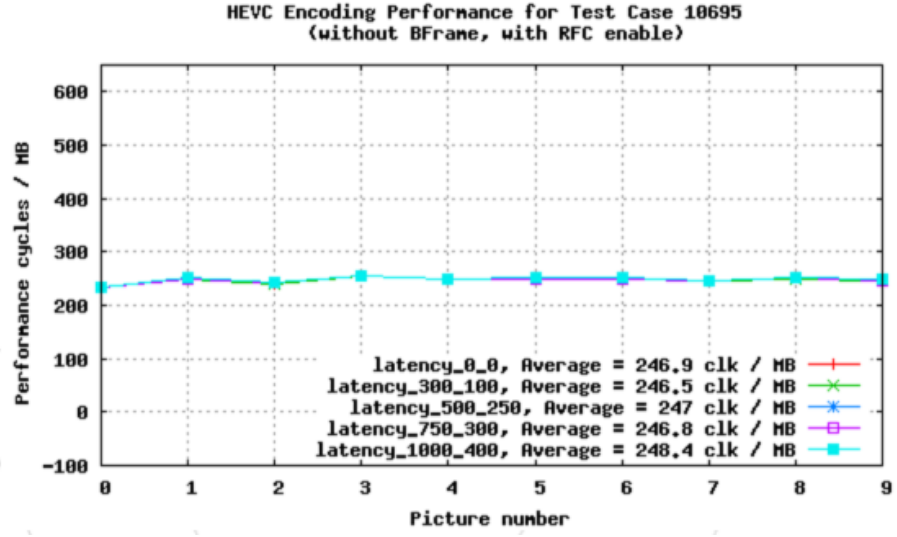
High BUS Latency Tolerance

▲Provide enough performance even in SoC with high BUS latency (up to 700 cycles)









Cycles/MB budget at 500 MHz:

4096x2160@60fps: 258 cycles/MB
3840x2160@60fps: 242 cycles/MB

VeriSilicon

Low DRAM Footprint

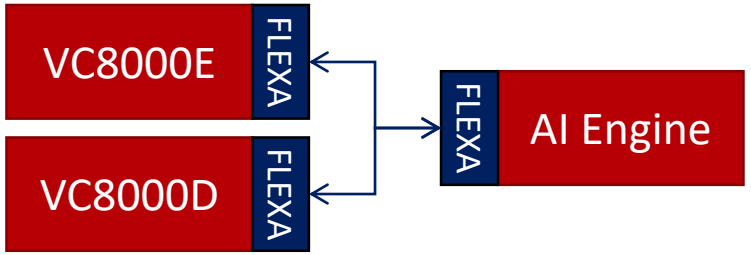
- **Use packed storage in DRAM for 10-bit data**
  - Our solution: 64 MB DRAM size for one 8K 10-bit picture ☺
  - Unpacked 16-bit: 102 MB DRAM size for one 8K 10-bit picture ☹
- **Allocate frame buffer on demand**
- **Direct reading decoder reference frame buffer which eliminates up to 10 frames of buffer from extra decoder output**

Robust Decoding and Encoding

- **Silicon proved video IP**
- **Rich test pattern database including multiple commercial test streams, streams from customers, compatibility streams, and self generated random error streams.**
- **Strong error handling**
  - Stream error detection in decoder
  - BUS error detection
  - Frame compression error concealment
- **Complex transcoding runs stably in hundreds of hours real product test**

Flexible Controllability by FLEXA API Video

▲FLEXA API Video is a Software & hardware interface enables VC8000E and VC8000D to cooperate with an AI engine



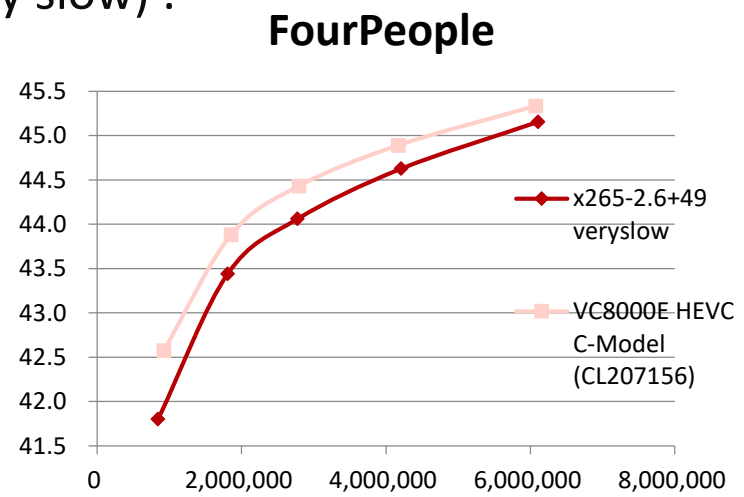
▲FLEXA API Video Examples

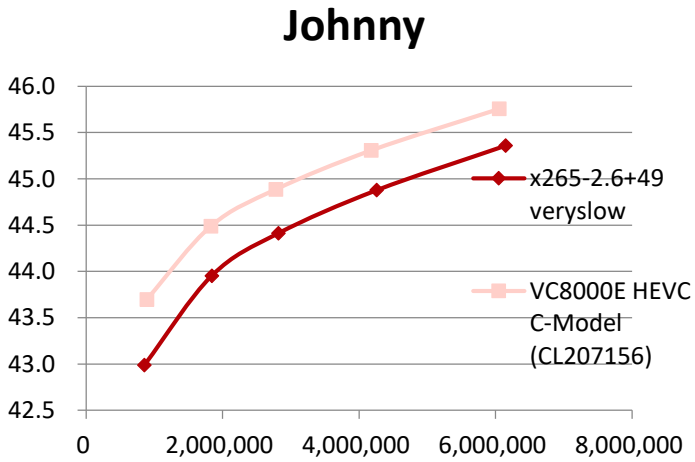
VC8000E	VC8000D
• **Various GOP structure setting**: hierarchal B, IDR, long term etc. • **Rate control setting**: Frame level and coding block level • **ROI map**: coding control down to 8x8 block such as qp and coding mode • **Special coding area**: Intra area, ROI area, IPCM area • **RDO level**: trade off between quality and performance • **Other controls**: Global MV, GDR, CIR etc. • Coding information output to DRAM • PSNR and SSIM report	• Coding information output to DRAM • Multiple down scaled frames

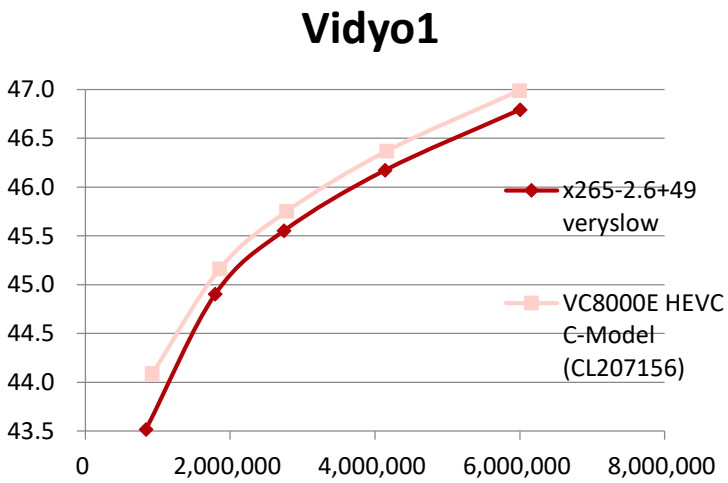
VeriSilicon

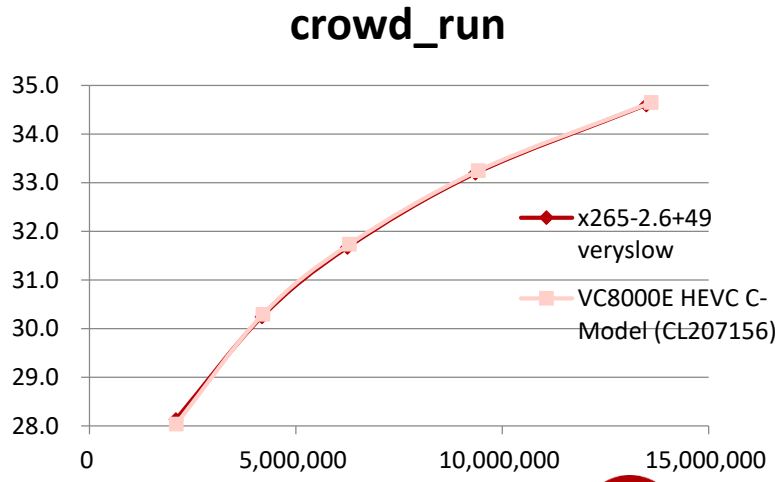
High Quality Video Encoding

- HEVC encoding quality achieves similar quality as x265(preset=very slow) .
  - Compare PSNR with x265-2.6+49:
  - Quality tuning based on JCTVC streams.
- H.264 encoding quality achieves similar to x264 medium.







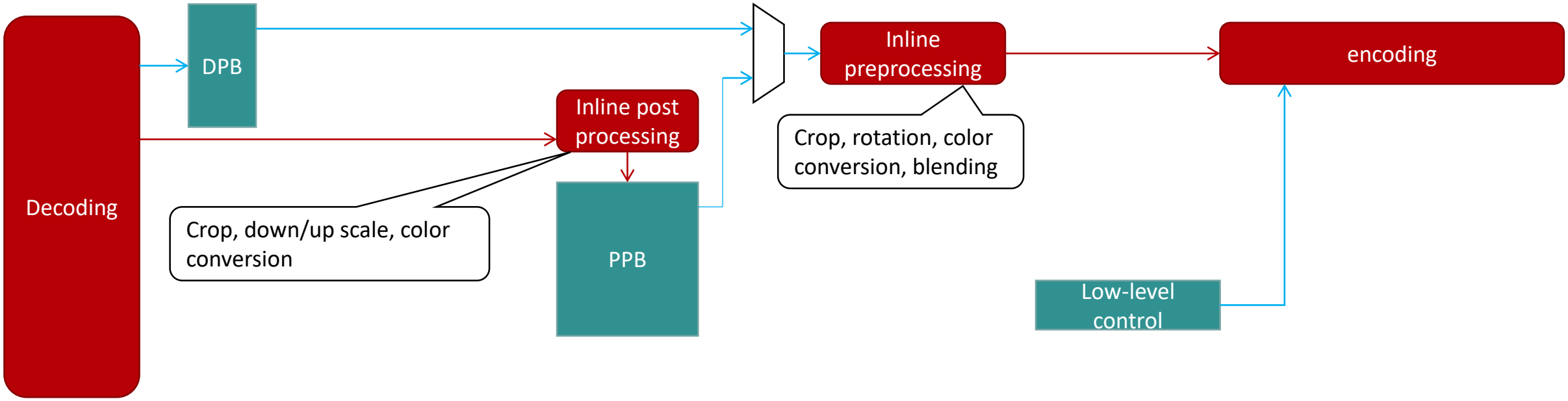


VeriSilicon

Video Transcoding

VeriSilicon

Basic Transcoding Flow



DPB: decoded picture buffer (reference picture buffer)
PPB: post processed picture buffer
Low-level control: Encoding control in a picture including QP map, ROI map, IPCM

Resize, Blending, and Multicast

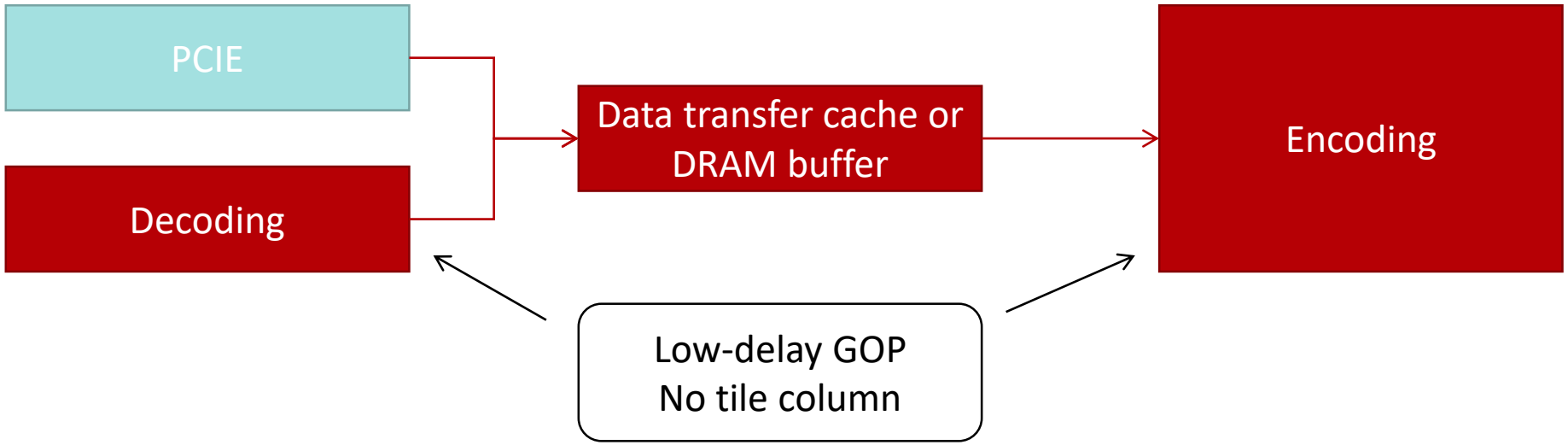
Decoder can support up to 4 resized outputs (down scaling and up scaling)
Encoder can support blending between 1 video layer and 1 another layer, sparsely with up to 8 regions
One input stream can have multiple output streams resized and with/without blending
Specific operation between decoding and encoding can be discussed, such as de-watermark

Multi-stream transcoding

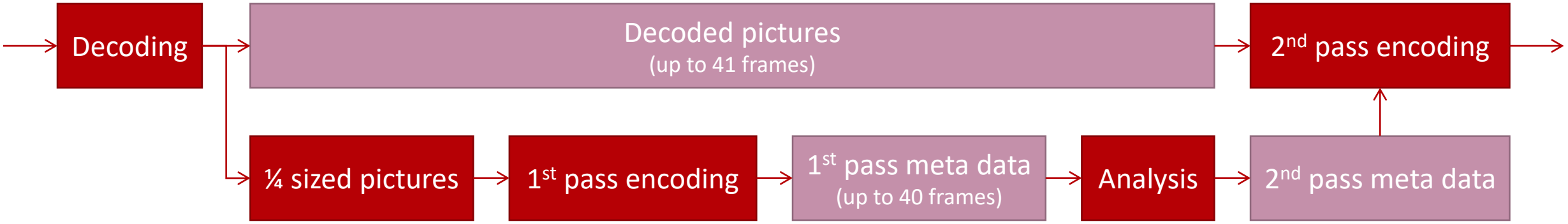
- **Scalable multi-stream transcoding**
  - Proportional CPU load increase
  - Proportional performance scaling
- **Concurrent video and JPEG transcoding is available by standalone JPEG only hardware**
- **Job switch at picture level and are flexibly scheduled by software driver**
  - Maximize the overall throughput
  - Ensure latency by priority management

Transcoding latency control

- Transcoding latency typically is less than 100 ms
- When the application requires, several ms ultra low-latency transcoding is possible
  - Sub picture level synchronization
  - On-chip SRAM for data transfer minimizes DDR traffic
  - Low-latency encoding or transcoding



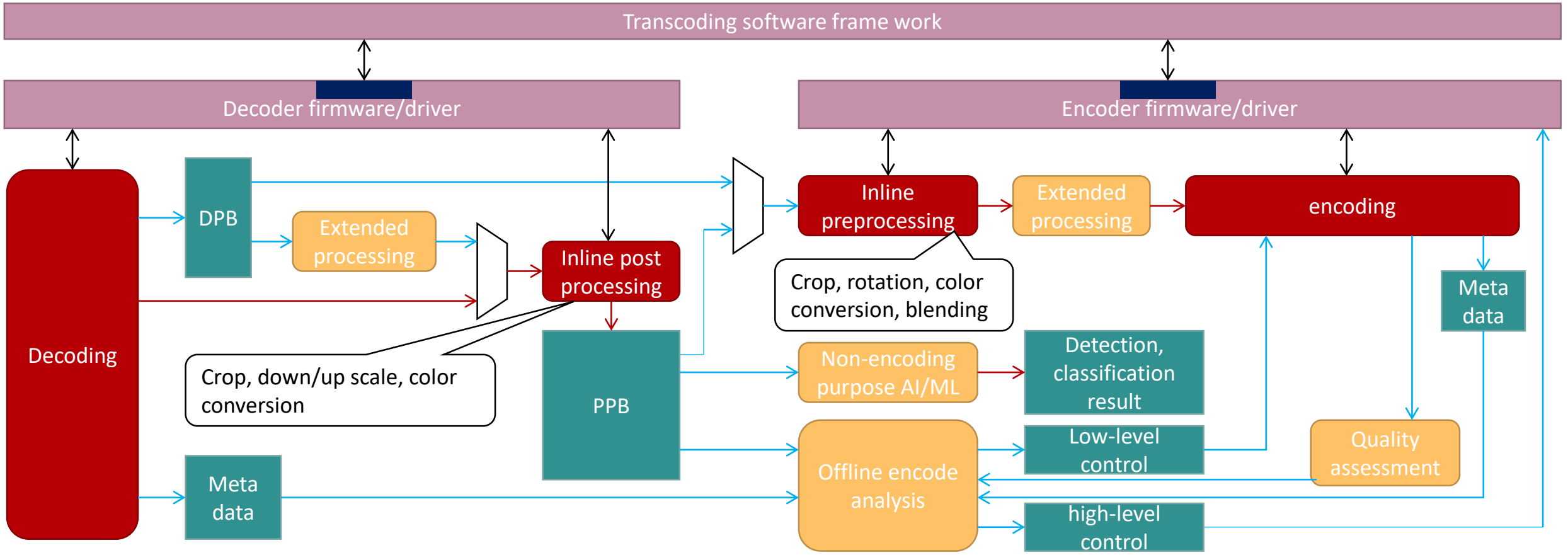
Hardware accelerated 2-pass Encoding



- The whole look-ahead 2-pass encoding process is hardware accelerated
- Support up to 40 frames look ahead
- The 2nd pass analysis hardware is a configurable module
  - Adds 0.8 GB/s bandwidth for 1 4K60 2-pass encoding
  - Saves 4200 MHz from meta data processing by CPU (18-frame look ahead)
- The first pass encoding is performed on ¼ sized picture. For example, when input is 4K, 1st pass encoding picture size is 1080p

Build Comprehensive Transcoding System by FLEXA API Video

A possible comprehensive transcoding system enabled by VeriSilicon transcoding solution

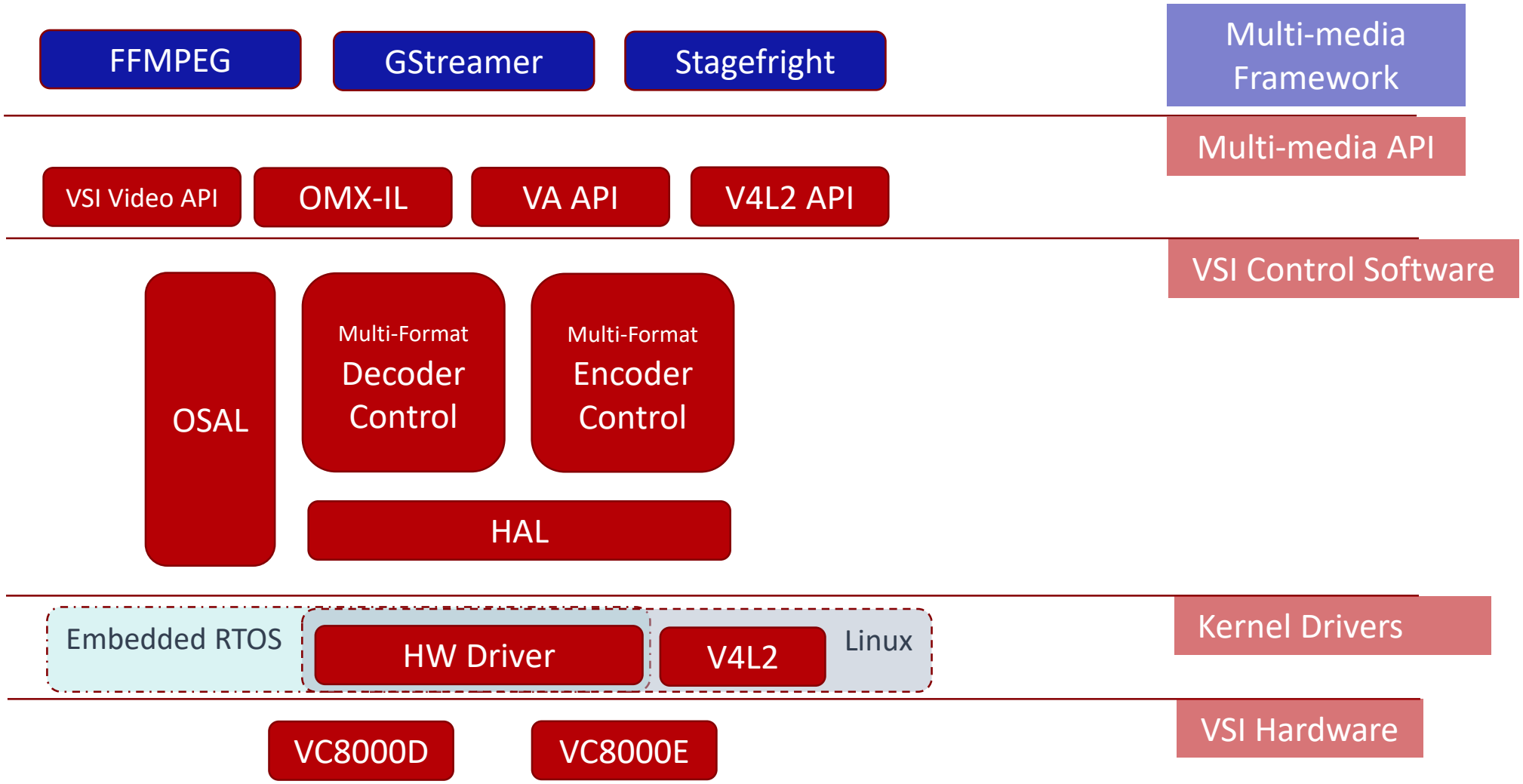


Software interface (API) in FLEXA API Video, providing ability to flexibly configure and control the encoding and decoding

AI and 3rd party computing processors cooperate with the encoder and decoder through hardware/buffer interface in FLEXA API video

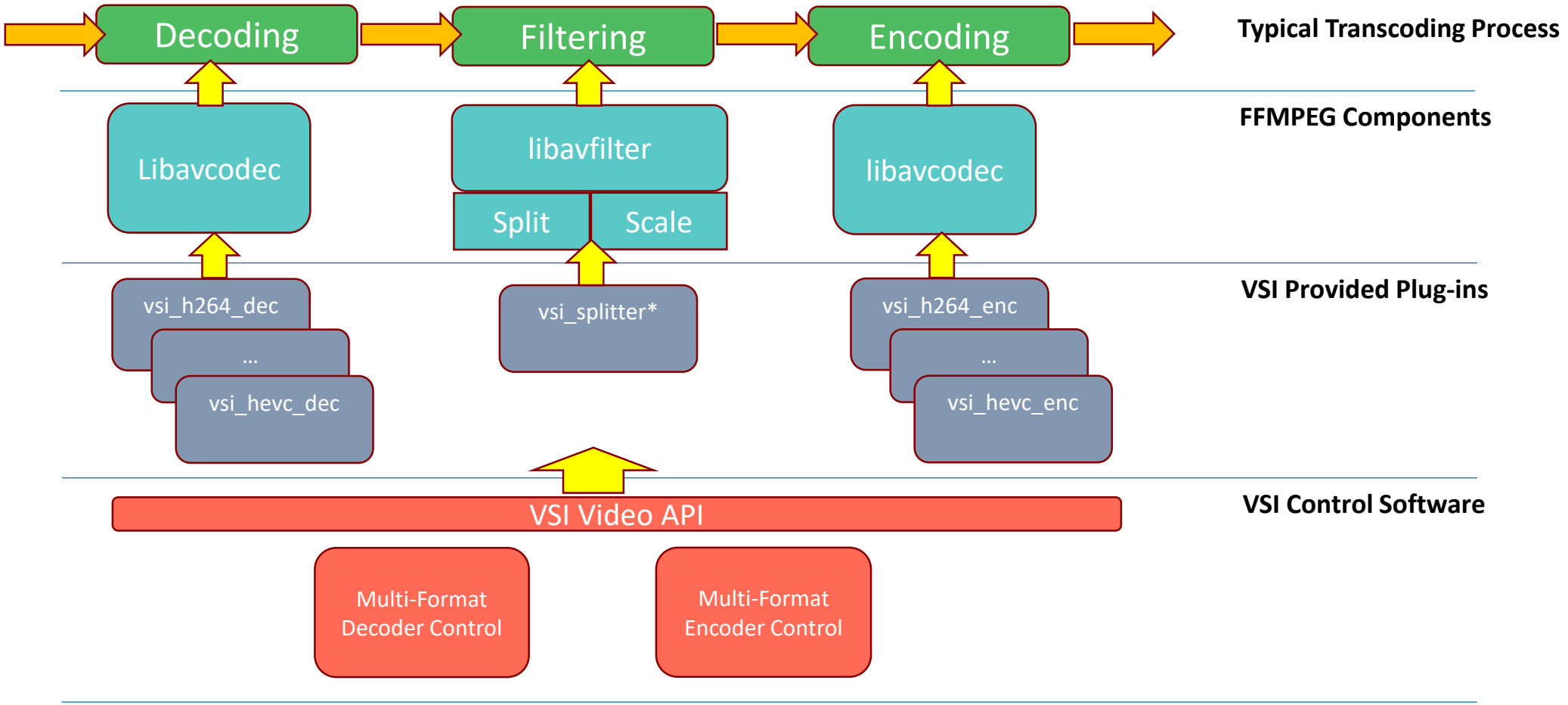
VeriSilicon

Complete Software Stack – Embedded to your framework easily.



VeriSilicon

Seamless Integrate with Industry standard FFmpeg framework



vsi_splitter*: VSI hardware decoder has inside Post-Processors that is capable of scaling. Using the splitter filter mechanism to simply "copied" these scaled video.

VeriSilicon


VeriSilicon