Applying AI/ML techniques for Live Video Transcoding

Ramdas Satyan

AGENDA

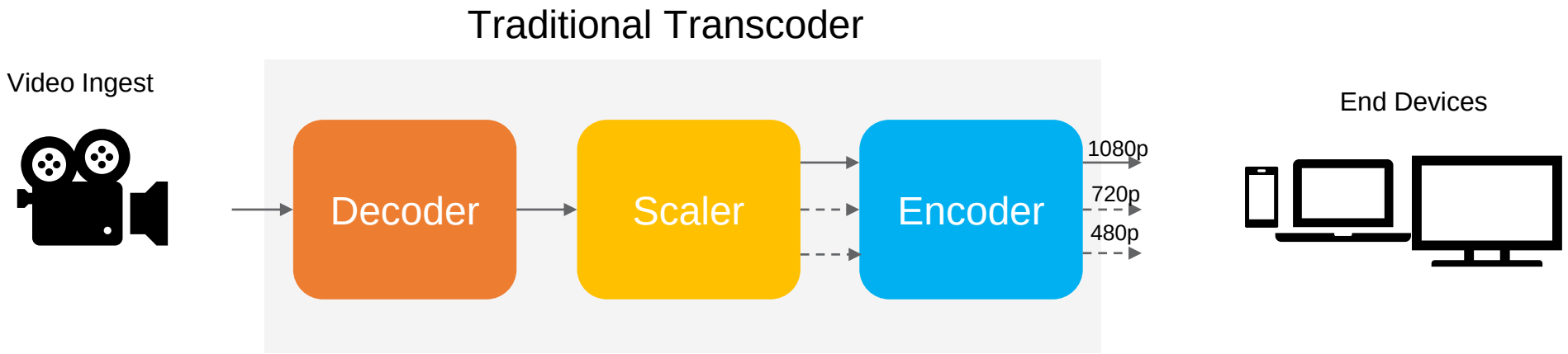
- Introduction
- Requirements for live video transcoding at scale
  - AMD Alveo MA35D transcode card
- AI/ML techniques for live video transcoding
- ML tools featured in this talk
  - Region of Interest (ROI)
  - Content Aware Encoding (CAE)
- Results

Live Streaming

- Reshaping the way we engage, communicate and entertain
- Gaining huge traction among people seeking real-time experiences and interactions
- It is thrilling to be part of live events from anywhere in the world
- Platforms like Twitch, YouTube Live, Facebook Live and others are go-to destinations
- 70% of the global streaming market is now Live (from a 2021 survey)

Real-time transcoding

- Plays a crucial role in delivering high-quality, compatible and optimized video content
- To cater to audiences with a range of devices and bandwidth constraints



Requirements of live transcoding at scale

- Everyone is a streamer!
  - Many Ingress and Egress streams (many to many)
- Live streaming platforms are generating videos at a massive scale (many petabytes a day)
- Compute is increasing exponentially
  - To support next generation standards (AV1, AV2, VVC etc.)
  - To cater to higher resolutions (8K) and frame rates (120fps)
- These interactive media applications require very low latency

➢ Use of Traditional CPUs: Economically unfeasible
  - The sheer number of servers needed would blow up the CAPEX and OPEX
  - Power and cooling is equally a big concern

➢ Solution: Video acceleration
  - Offload the entire video pipeline using an ASIC to cater to each of the above requirements

AMD Alveo MA35D Media accelerator

- Video transcode card for the data center
- 5nm ASIC architecture
  - Optimized for high channel density
  - Designed to prioritize high video quality while maintaining power efficiency
- Supports encoding and decoding of H.264, H.265/HEVC and AV1 standards
- Striving to deliver software-encoder-level video quality
- Video Quality is comparable to
  - H.264: x264 medium preset
  - H.265/HEVC: x265 medium preset
  - AV1: x265 slow preset

# of channels: Up to 32x 1080p60
Power: 1 watt per 1080p60 channel
Latency: 4K encode at 8ms max latency
AI processer: 22 TOPS compute

Innovation in Video Processing

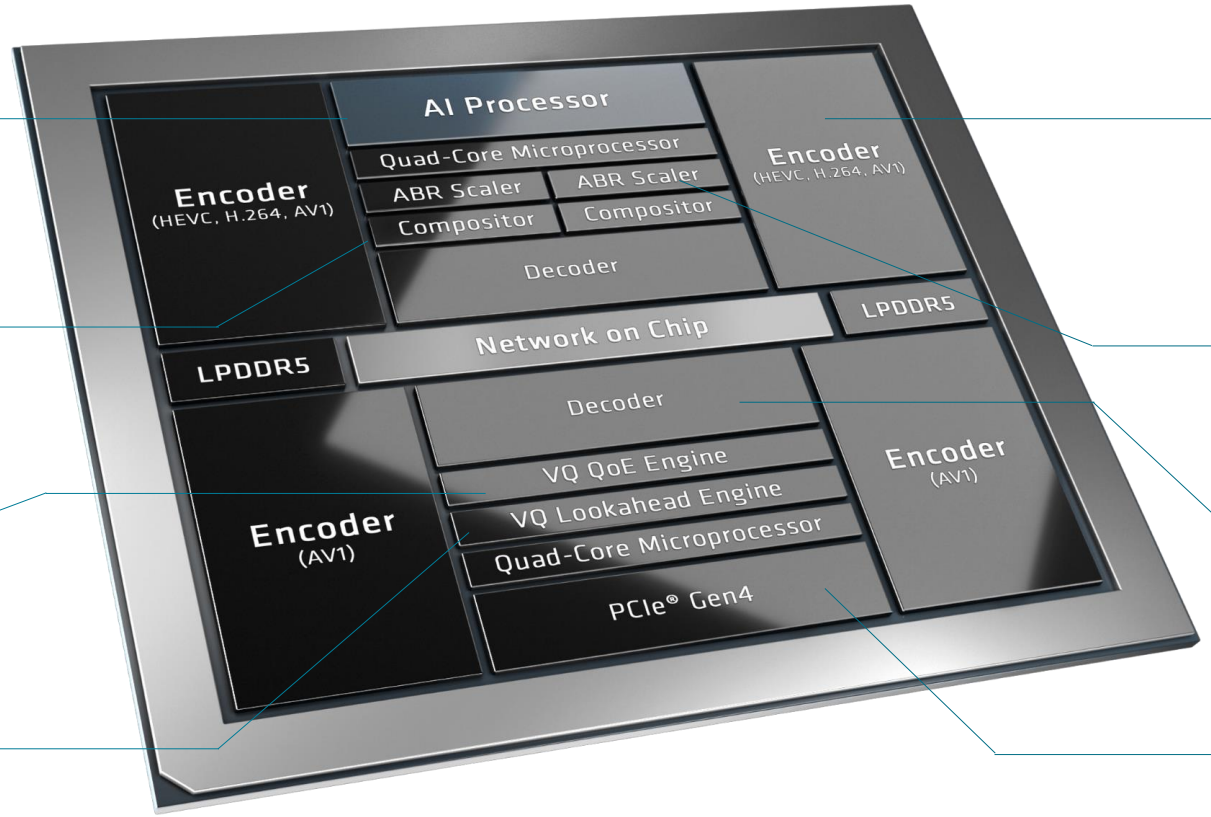
Targeted ASIC for Interactive Streaming

AI Processor
- Optimize for 'perceived' visual quality
- Improve compression efficiency

Compositor Engine
- For multi-screen, multi-layering
- Immersive experiences

VQ QoE Engine
- For consistent visual quality
- Maintains QoS for customer experience

VQ Look-Ahead
- Analysis of motion and content
- Improves compression efficiency



Encoders
- Supporting mainstream standards
- Next generation AV1 standard

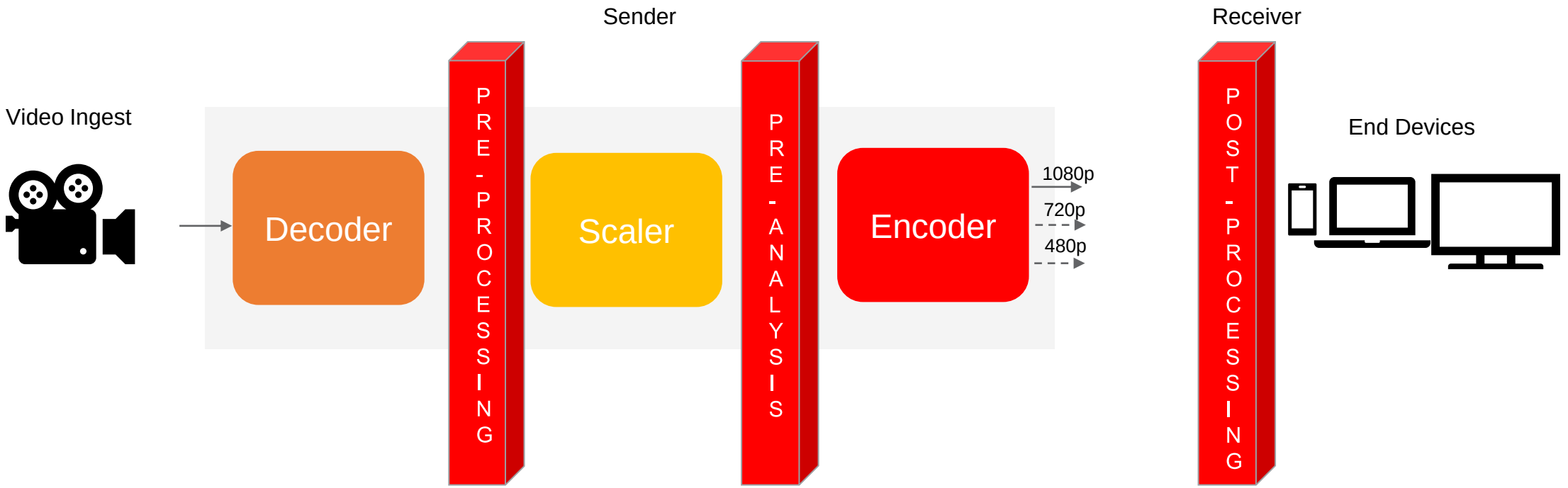
Adaptive Bitrate Scalers
HW accelerated ABR scalers for multiple resolutions for diverse endpoints

Decoders
AV1, HEVC, H.264, VP9

PCIe® Gen4
- Host CPU communication
- Bifurcated x4x4

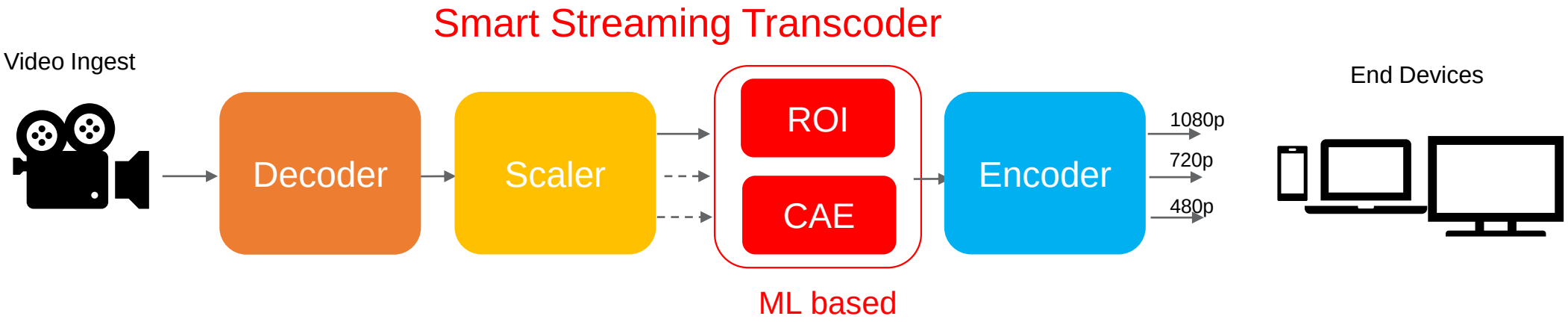
5nm Technology
(2x Devices per Card)

AI/ML techniques for video transcoding



Changing input pixels to make it compression friendly	Pre-analyzing input to identify areas which needs better fidelity	Driving the encoder decisions based on data instead of classical approaches	Post processing of pixels to improve visual quality before display

ML tools presented in this talk



Value proposition

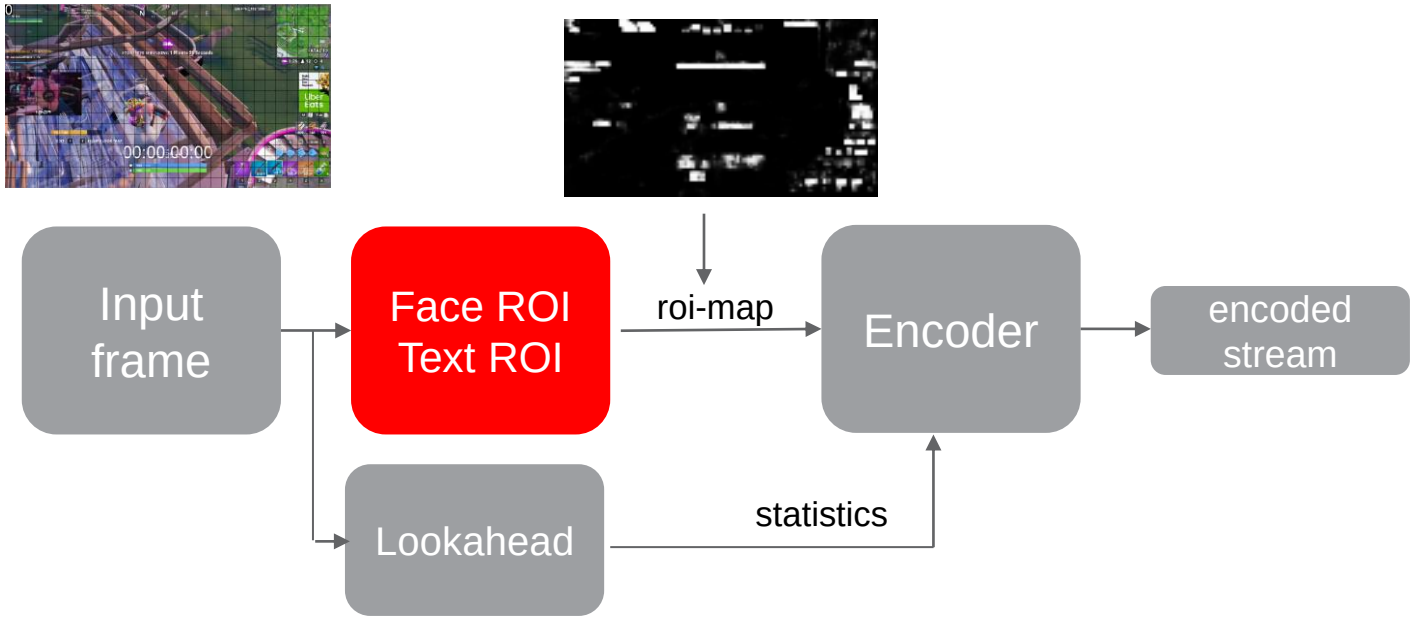
1.Region of Interest (ROI) improves visual quality by redistributing bits to certain important areas within a frame

2. Content Aware Encoding (CAE) allows content distributors to stream video at lower bit-rate without compromising on video quality (reduces streaming costs)

ROI

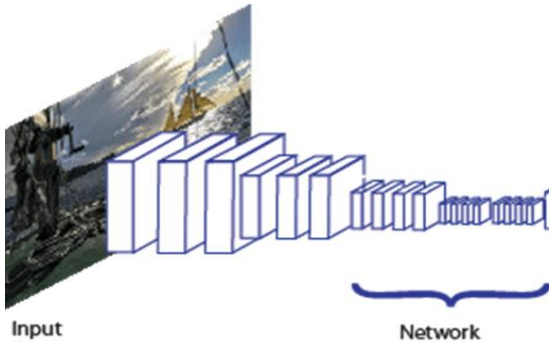
ML based ROI

- In some applications, certain regions in a frame are more important
- Two such regions are
  - Face
  - Text
- ML is used to detect such regions of interest.
- Encoder smartly redistributes bits on regions of interest by adjusting block level quantization parameter (delta-QP) using a ROI map
- Results in improved visual quality in ROI
- Can also be used to reduce the bit-rate by maintaining same video quality as before to achieve additional bandwidth savings



Model Architecture

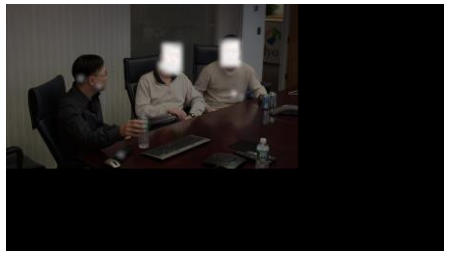
- Customized network architecture that meets the following requirements
  - Easily fits MA35D ML engine and achieves high throughput
  - Maintains high accuracy
  - 10x less compute over standard MobileNet-V1
  - Model operates on native HD/SD resolution
    - This avoids input distortions and pre-processing
  - Directly produces a block level (16x16) ROI map
    - This avoids heavy post processing

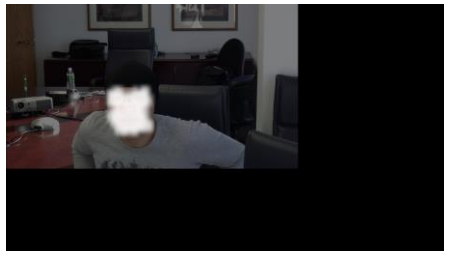


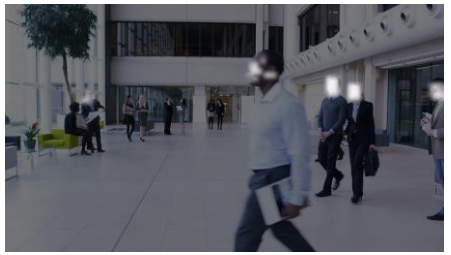
Text ROI
- Trained on COCO-Text dataset
- Good accuracy for small and medium text (receptive field is 80x80 pixels)
- Model can detect both English and non-English (Chinese) text
- Large text ignored as the encoder does a good job

Face ROI
- Trained on WIDER dataset
- Good accuracy for small and medium faces
- Side face detection needs some improvement
- Part of Large face detected (beyond receptive field)

Examples of ROI detection







ROI ML Inference Pipeline

Pre-Processing

Scale & Tile to Canvas

(1920x1080x3)

High resolution input image required for small objects

Infer

ML Model

Post-Processing

Non-Maximal Suppression

(120x68x1)

Score, Sigma

Threshold

Objects which exceed receptive field have weak activations. Use NMS to suppress potential false positives.

ROI Map

(120x68x1)

Each pixel in the ROI map corresponds to a 16x16 block of the input image

Face ROI results

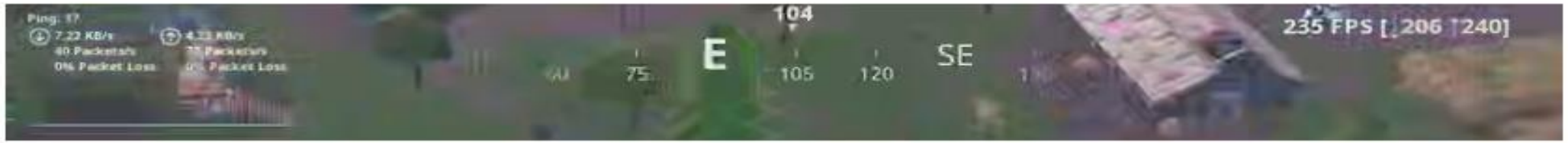
MA35D AV1: FourPeople@125 Kbps, ~VMAF 75

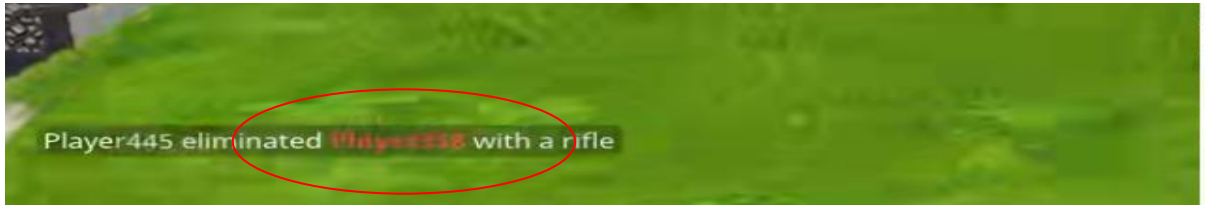
Without ROI Coding

With ROI Coding

Text ROI results

MA35D AV1: Ninja_720p@2M, VMAF= ~70



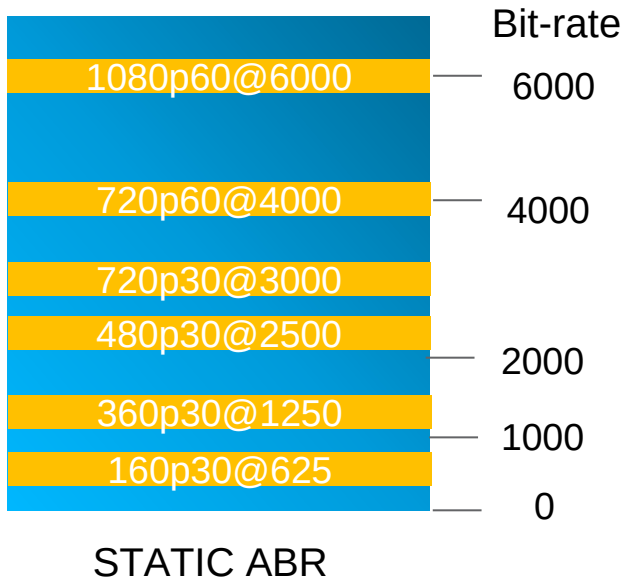


MA35D HEVC: Fornite_720p@2M, VMAF= ~70

CAE

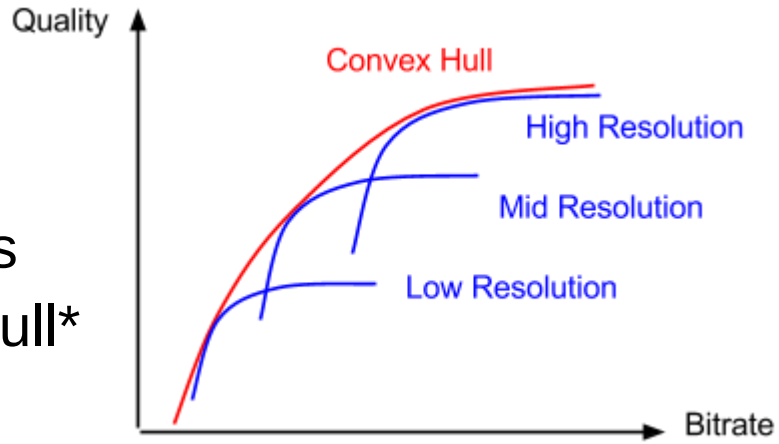
Static ABR ladder

- Pre-determined set of bit-rate and resolution combinations are used
- This one size fits all approach has these limitations
  - Inefficient use of bandwidth
  - Sub optimal video quality



Content Aware Encoding (CAE)

- VOD
  - Netflix pioneered per-shot encoding
  - Each shot is encoded multiple times in many resolutions
  - Final encoder decisions are made based on a convex hull*
- Live
  - No luxury of “**infinite**” latency
  - **Real-time** Transcoding at **scale** => limits processing capacity

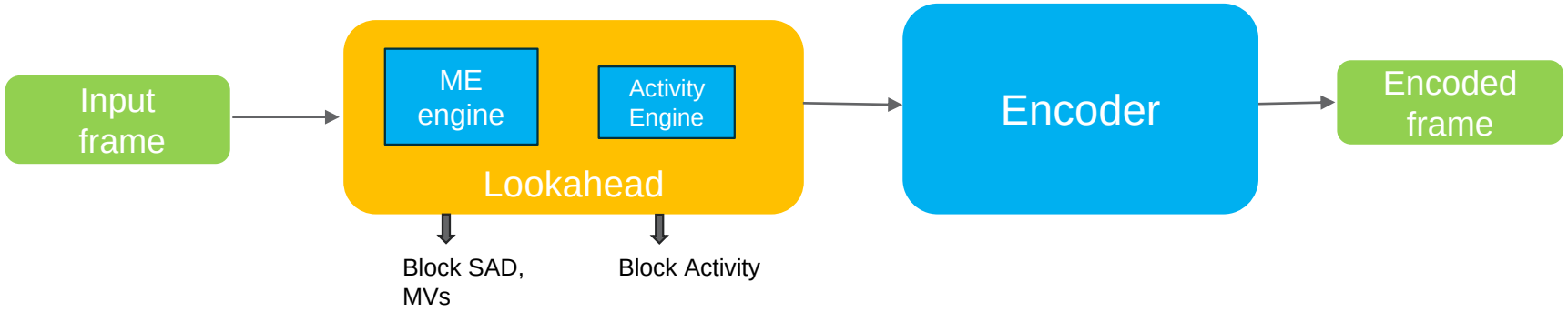


* https://netflixtechblog.com/per-title-encode-optimization-7e99442b62a2

ML based CAE for live streaming

- Problem definition
  - How to find the best possible quality/bit-rate trade-off in real-time with limited compute while maintaining the latency?
- What do we have in our arsenal?
  - Some knowledge of incoming content (limited lookahead)
  - Deep learning toolbox
- Our approach
  - Using the information from the incoming input, train deep neural networks to predict “optimum” encoder parameters
    - Dynamically adapt bitrate based on content to optimize quality per resolution

Input stats used



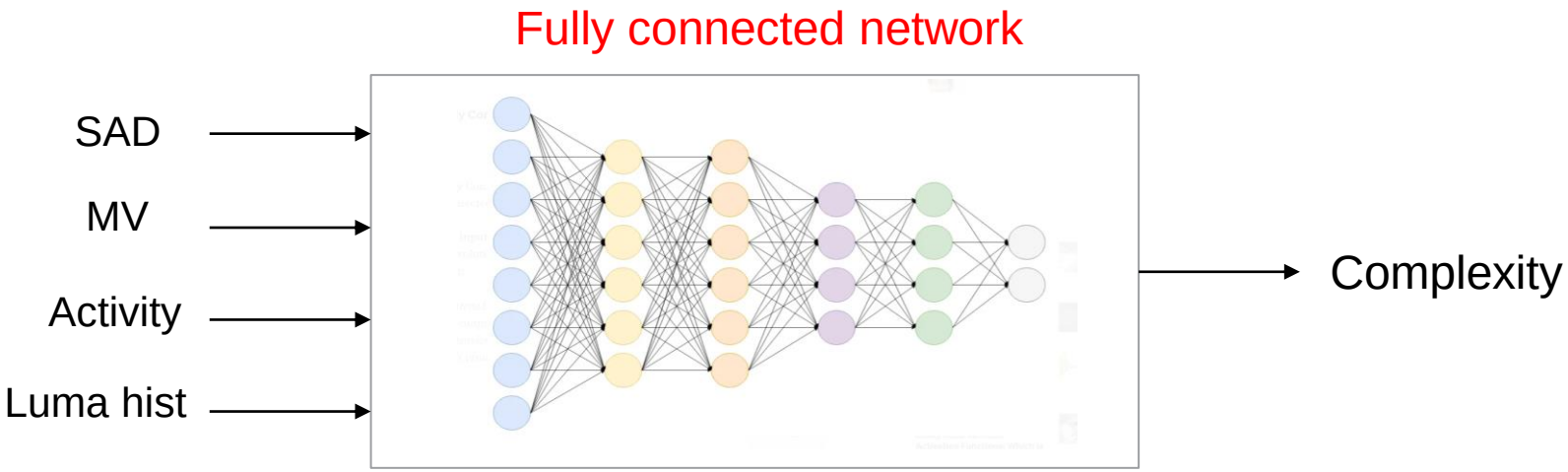
Lookahead stats used

- Frame SAD Ref0 Histogram (16,1)
- Frame MV16 Ref0 Histogram (16,1)
- Frame Activity Histogram (512, 1)
- Frame Luma Histogram (256, 1)

Create Input vectors (X)

- Generate segment wise vector sum of SAD, MV, Activity and Luma
- Stacking sum of SAD, MV and Activity vector to create (800,1) vector
- Normalize them separately

DNN based Complexity predictor



- Fully connected regression network
- 800 input nodes
- 11 hidden layers
- 1 output number suggesting complexity
- Loss function is MSE
- Optimizer is RMS prop

Ground Truth

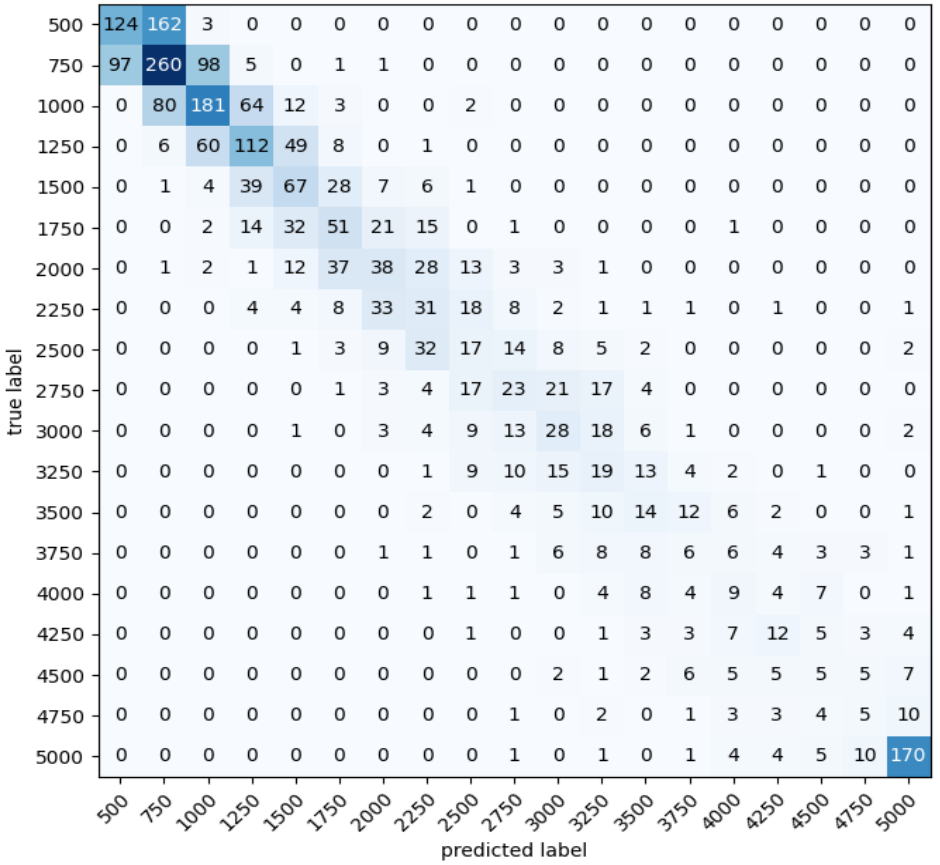
- We believe a minimum quality bar of VMAF 90 is acceptable for live streaming
- Model is learning about complexity as follows
  - Easy clips will reach high quality at lower bitrates in comparison with complex clips
- Training
  - Our video dataset is broken down to segments (2sec intervals)
  - For example: 1080p clips are encoded from 500kbps to 10Mbps (interval of 250kbps)
  - Y = minimum bitrate where VMAF >= 90

Model details

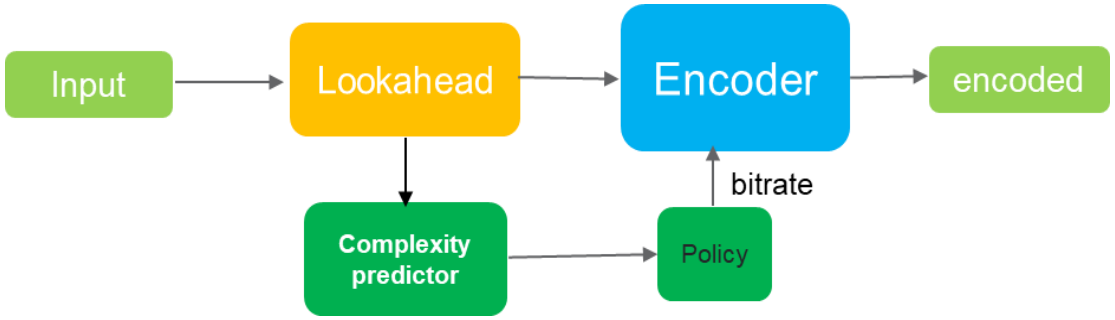
Total dataset: 10451 segments
70% training, 20% testing, 10% validation

Accuracy: 88.75%
MAE: 160 (kbps)

Confusion matrix (Actual vs Predicted bitrate)



How is this model used?



- Model is used to make prediction for every frame
- Minimum lookahead of 4 frames needed to achieve good accuracy
- Policy block determines the final bitrate
  - Customers have control based on their use case and requirements
  - Sample policy used
    - Primary motive is to preserve quality and bit-rate changes are conservative
    - Lower bitrate slowly over segments (when complexity decreases)
    - Increase bitrate immediately (when complexity increases)

Results: CAE bitrate savings

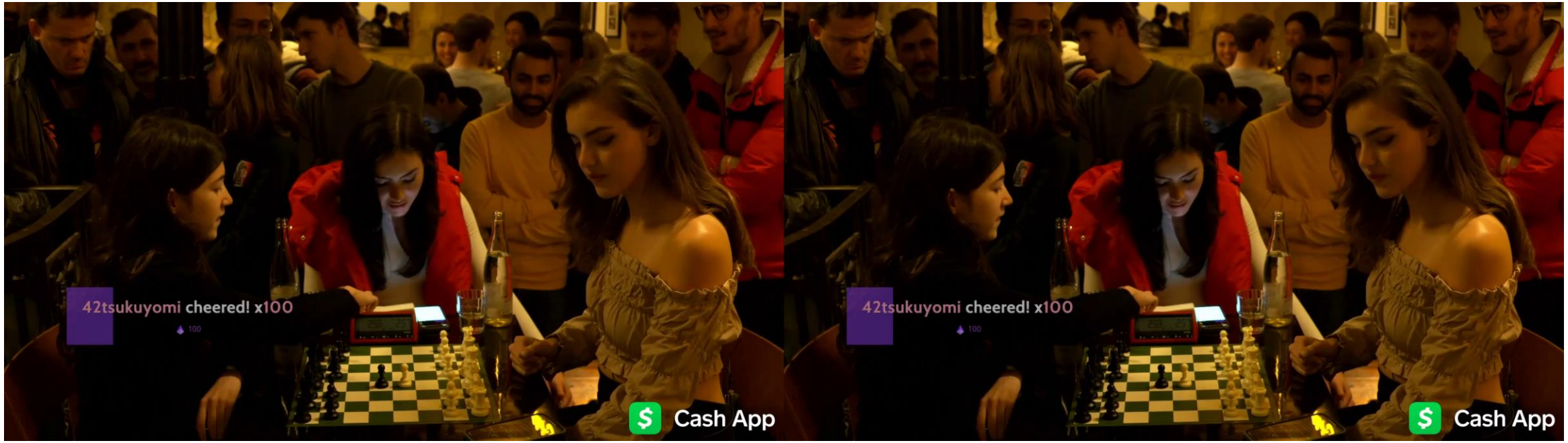
- ~ 100 popular 1080p clips were downloaded from twitch.tv
- Encoder used: AMD MA35D AV1 encoder
- Comparison: CBR @4Mbps and CAE @recommended bitrate

Complexity	# of streams	Avg VMAF CBR	Avg VMAF CAE	Bit-rate Savings
low	27	96.38	95.63	35.08%
med	32	93.41	92.92	9.12%
high	43	82.79	82.66	1.35%

Subjective assessment of CAE
Resolution:1920x1080
Encoder: **AMD MA35D AV1 encoder**

Fixed: 3.90Mbps
VMAF: 94.36

CAE: 2.26Mbps
VMAF: 92.0



CAE savings: ~42%

Thank you


AMD