Energy-Efficient AI Video Analytics with MPEG-5 LCEVC: GPU-Optimised Results

Manzoor Razaak, V-Nova Ltd., London
Guendalina Cobianchi, V-Nova Ltd., London
Ben Cope, Intel Corp.
Palanivel Guruvareddiar, Intel Corp.

Video Formats vs. AI Requirements

Human Consumption

Aimed at perceptual fidelity

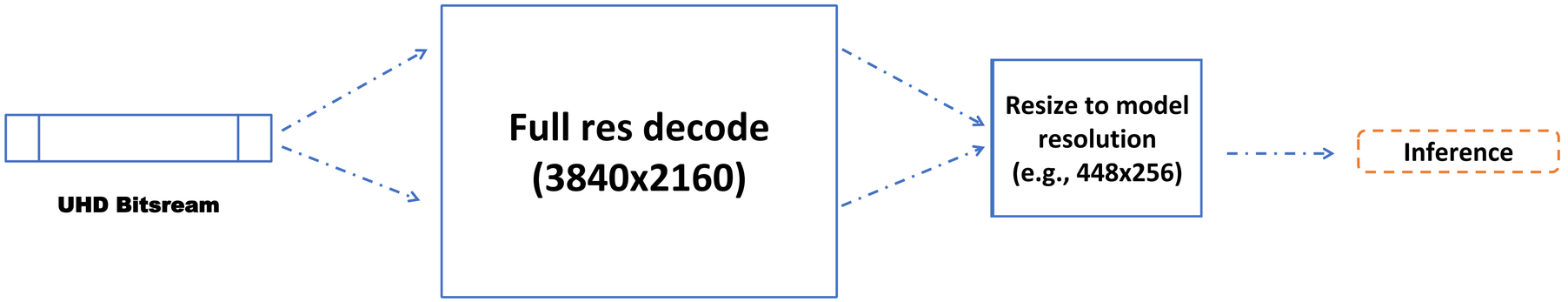
e.g., full resolution, 60 fps, 4:2:0 color

Machine Vision

Aimed at Feature Extraction

e.g., low resolution, sparse access, ROI focus

Decoding Driven Inefficiency in Video Analytics



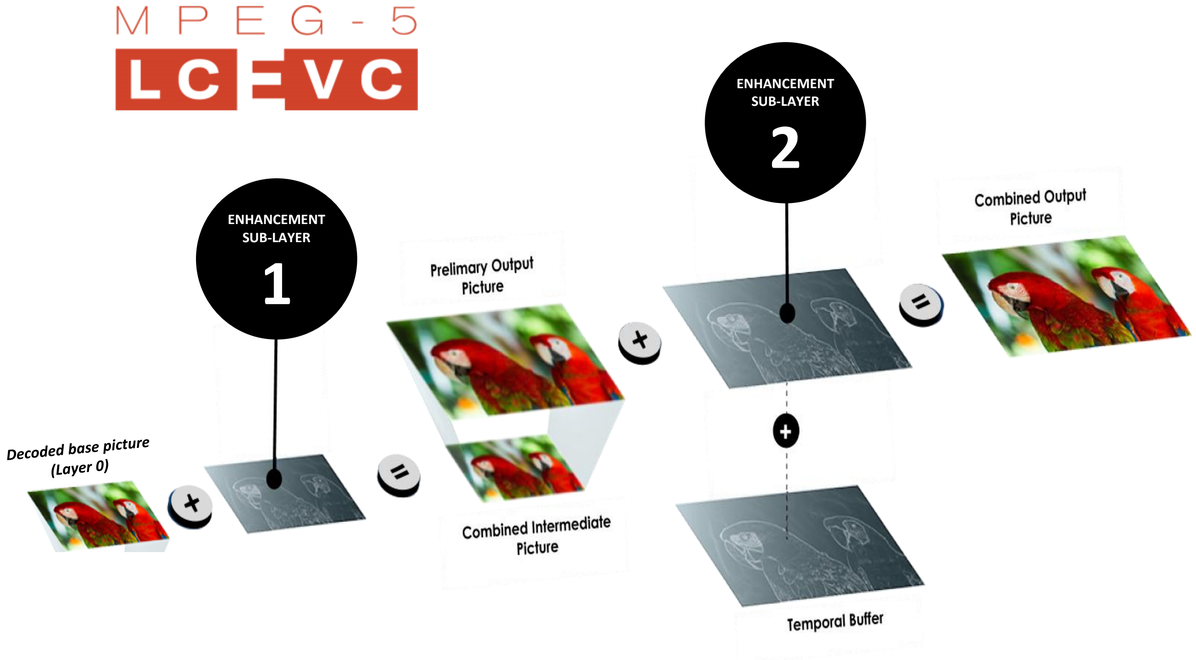
Significant optimizations are made in AI inference models, yet a fundamental inefficiency remains

- Typical analytics pipeline: full-frame decode → color convert / scale → inference
- Inference operates on reduced representations (downscaled inputs and/or sparse frame sampling)
- Example: 3840×2160 → 448×256 model resolution – the AI model works on **~72x reduced spatial size**

This mismatch shifts bottlenecks to decode/scale operations and memory movement, leading to **accelerator under-utilisation** when inference cannot be fed at line rate

Hierarchical Codecs Enable 'Compute-Aware' Data Access

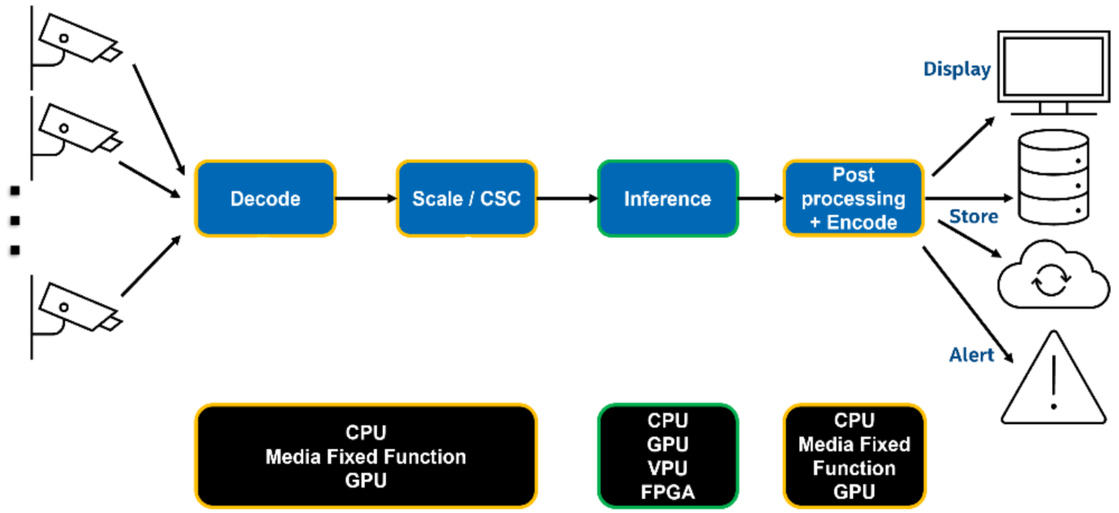
- **Conventional video codecs** (e.g. AVC, HEVC) encode and decode video at a single target resolution, requiring **full-resolution processing** regardless of downstream use
- **MPEG-5 LCEVC** enhances an existing base codec with a lightweight enhancement layer that carries residual information
- **Hierarchical representation**: frames are encoded across multiple spatial resolution layers.
- Flexible decoding strategies enabled:
  - **Base-first decoding** for coarse or early-stage AI analysis with reduced compute.
  - **Selective enhancement** decoding to access higher-resolution detail only when required



MPEG-5 LCEVC (ISO/IEC 23094-2)

Base-First Decoding for AI (HD), selectively pull UHD if needed

Methodology: Benchmarking on Intel® Core™ Ultra 7

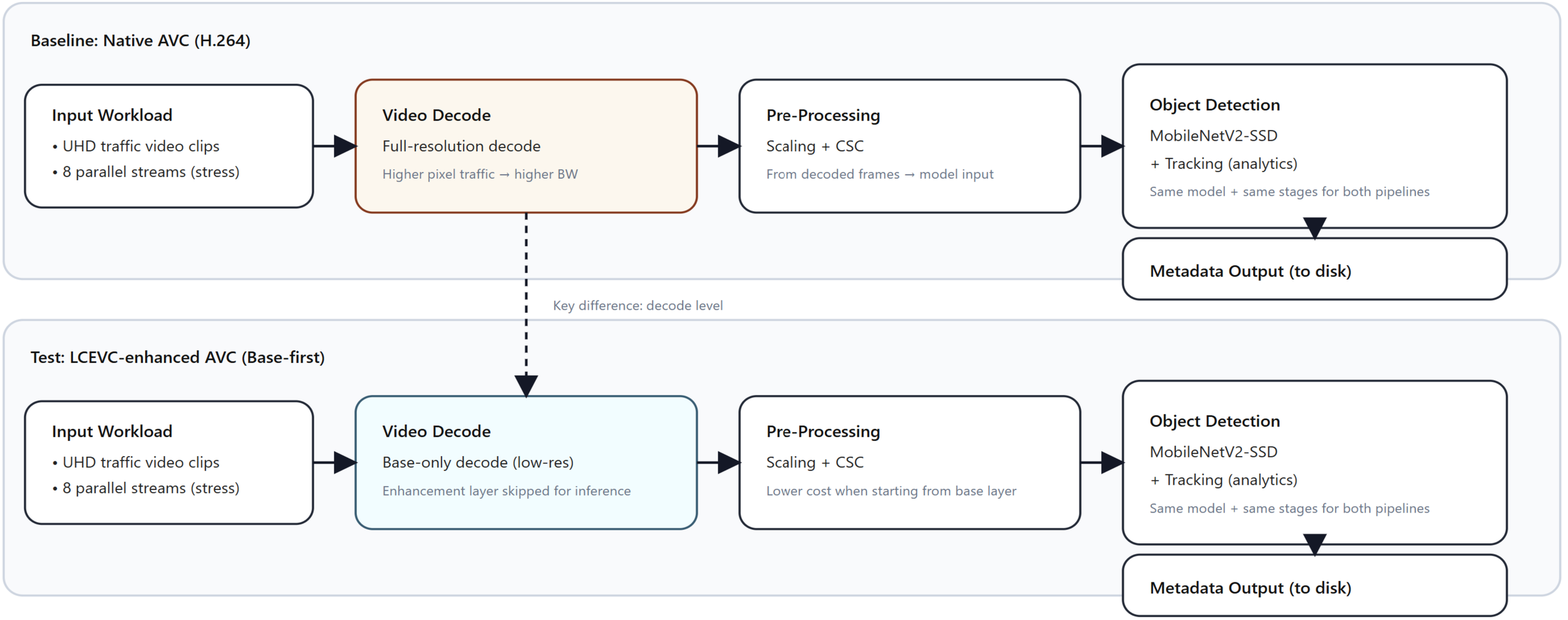


Specs Table:

- **Framework**: Intel's Deep Learning Streamer (DL Streamer), an open-source streaming media analytics framework
- **Hardware**: Intel® Core™ Ultra 7 258V (Lunar Lake) w/ Arc 140V integrated GPU. Hardened AVC decoders.
- **Test Codec**: LCEVC-enhanced AVC (base-first operation)
- **Dataset**: VisDrone - UHD Traffic Sequences – 4 videos
- **Inference Model**: SSD MobileNet V2 for object detection (Model resolution: 448x256 input)
- **Evaluation** focuses on system-level throughput using standardized, reproducible hardware and software stacks without impacting inference accuracy

https://docs.openedgeplatform.intel.com/dev/edge-ai-libraries/dlstreamer/index.html

DLStreamer Pipeline: AVC (H.264) vs. LCEVC-AVC



Platform & Measurement

Intel Core Ultra 7 258V (integrated GPU, unified memory) • Intel DL Streamer • GPU-only measurements

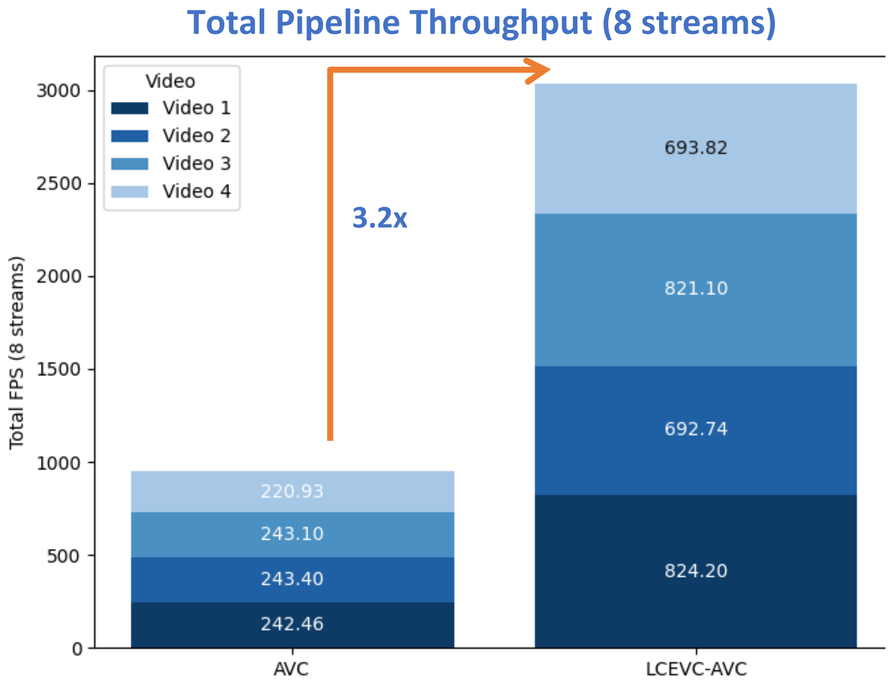
Metrics Collected

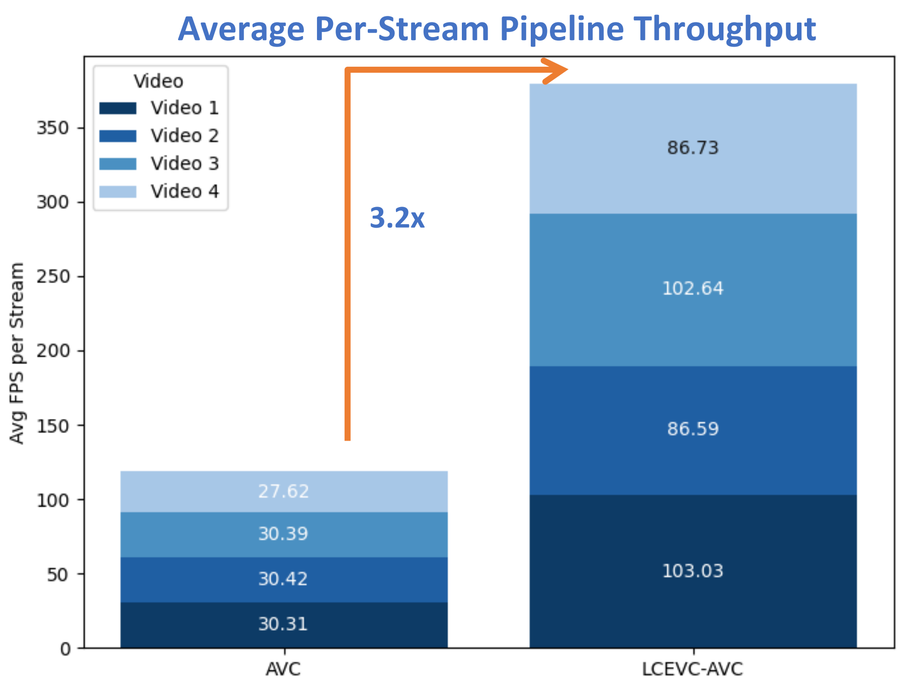
End-to-end FPS • Avg power/stream • Memory BW/stream

LCEVC Acceleration Results – Pipeline Throughput

Throughput (FPS) comparison

AVC vs. LCEVC AVC (hardened AVC codec support)





Key insights

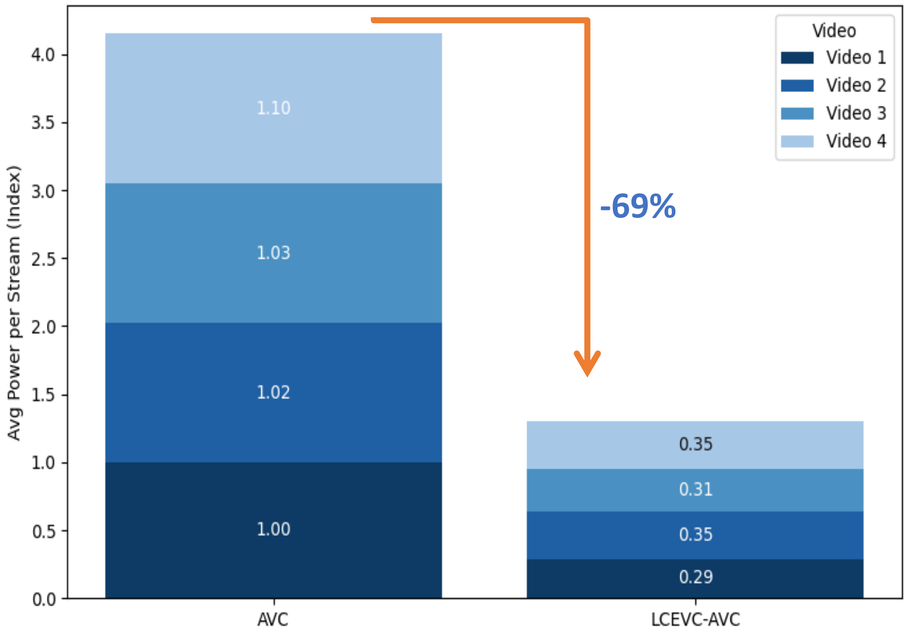
- LCEVC-AVC achieves faster pipeline execution compared to native AVC
- Upto **3.4x** higher throughput in terms of FPS
- GPU decode load decreases ~4x, inference accelerated thanks to lighter scaling to model resolution

LCEVC Acceleration Results – GPU Performance

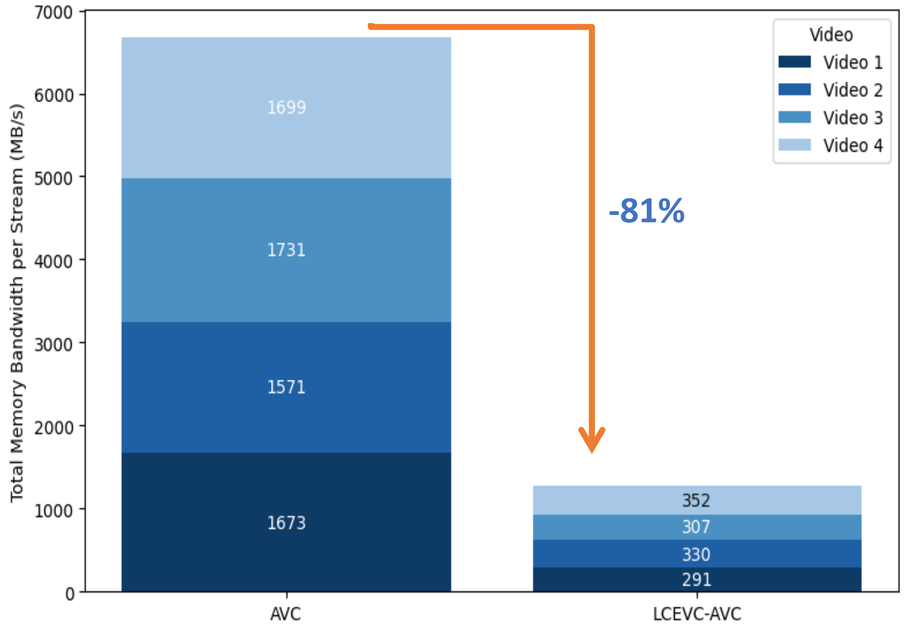
GPU Performance comparison

AVC vs. LCEVC AVC (hardened AVC codec support)

Average Power Consumption per Stream (W)



Memory Bandwidth utilization per stream (Mb/s)



Key insights

- Power per stream reduced by **~70%**
- Memory bandwidth utilization reduced by **~80%**
- Efficiency gains are consistent across content
- LCEVC not only improves throughput but also fundamentally reduces the data and energy cost of getting pixels to the GPU.

Impact of Base-Layer Decoding on Inference Accuracy

Native H.264

LCEVC H. 264

Native H.264

LCEVC H. 264

*** Visualization at display quality, as outputted by DL Streamer**

- **Inference model`s detection** remains consistent when on operating on LCEVC base layer
- **Both inputs are downscaled to 448×256** before inference; the comparison therefore reflects the effect of decoding full-resolution frames versus decoding only the base layer before scaling

Implications for Sustainable AI Infrastructure

- **Higher utilisation of fixed compute budgets.** More concurrent analytics workloads on the same edge or server hardware, without increasing silicon capability
- **Lower pressure on memory and I/O subsystems.** Base-layer and selective access reduce data movement across memory hierarchies and storage interfaces, which are dominant bottlenecks
- **Efficient sparse analytics at scale.** Selective resolution and frame access support sparse inference patterns, avoiding unnecessary decode and processing
- **Energy efficiency as a system-level outcome**. Reduced decode, memory traffic, and data transfer translate into lower power per processed stream, as confirmed by complementary measurements on the same platform*

Conclusion & Future Work

Conclusions

- **Codec choice** is a system-level optimisation lever for video analytics pipelines, not only a storage or bitrate decision.
- LCEVC enables **base-first AI inference**, aligning video decoding with actual model input requirements
- **3.4x higher inference throughput** achieved on integrated GPU hardware without model changes
- **Per-stream efficiency gains are substantial: ~70%** lower power and **~80%** lower memory bandwidth

Limitations and next steps

- Evaluation focuses on system-level throughput; ground-truth accuracy metrics (e.g. recall, F1) were not reported.
- **Future work**: extend to labelled datasets, additional tasks (e.g. tracking, segmentation), and GPU-accelerated VVC decoding for direct comparison with AVC.

Key Takeaway:

Optimising AI pipelines is not only about better AI models, but also about smarter data architecture

Thank you!