Seamless Deployment of Multimedia & Machine Learning Applications at the Edge

Megha Daga

Sr Director, Product Management, AIoT

Qualcomm Technologies, Inc.

Intelligent Edge use cases growing at exponential scale

Latency/Speed

Bandwidth/Load Balancing/Scale

Reliability

Cost

Privacy/Security

Retail

Intelligent Inventory Management:
Latency driving productivity
Secured & Bandwidth
driven last mile delivery

Manufacturing

Robot Navigation & Management:
Latency driven real-time
defect inspection on
conveyor belts

Healthcare

Latency critical Remote Robotics:
Assisted Surgery,
Immersive training AR/VR,
Image diagnostics,
OR monitoring

Smart City

Smart AI Cameras:
Bandwidth & Latency
critical traffic flow across
connected cameras
Secured building management
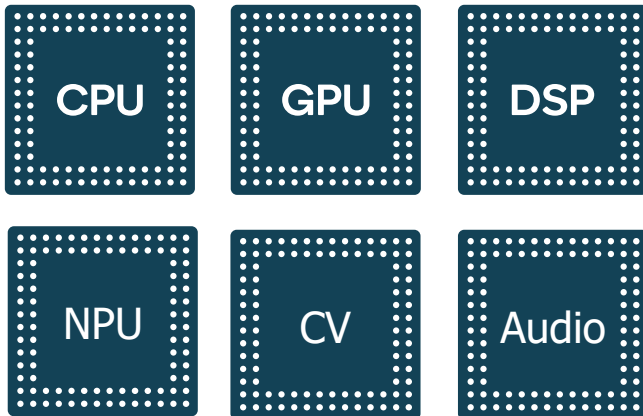across connected sensors

Warehouse & Logistics

People Navigation:
Latency driven quickest route
Real-time package sorting

Qualcomm

Exponential increase and diversity in use cases leading to fragmentation across Hardware & Software
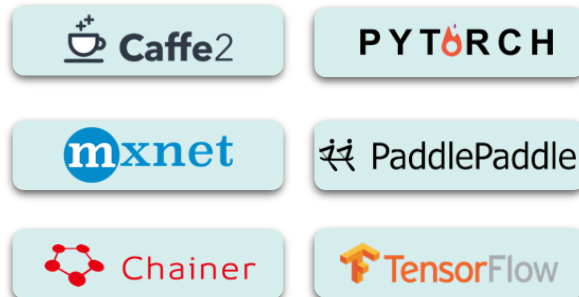

HARDWARE
Fragmentation

CPU	GPU	DSP
NPU	CV	Audio

SOFTWARE
Fragmentation

Framework

Caffe2   PYTORCH

mxnet   PaddlePaddle

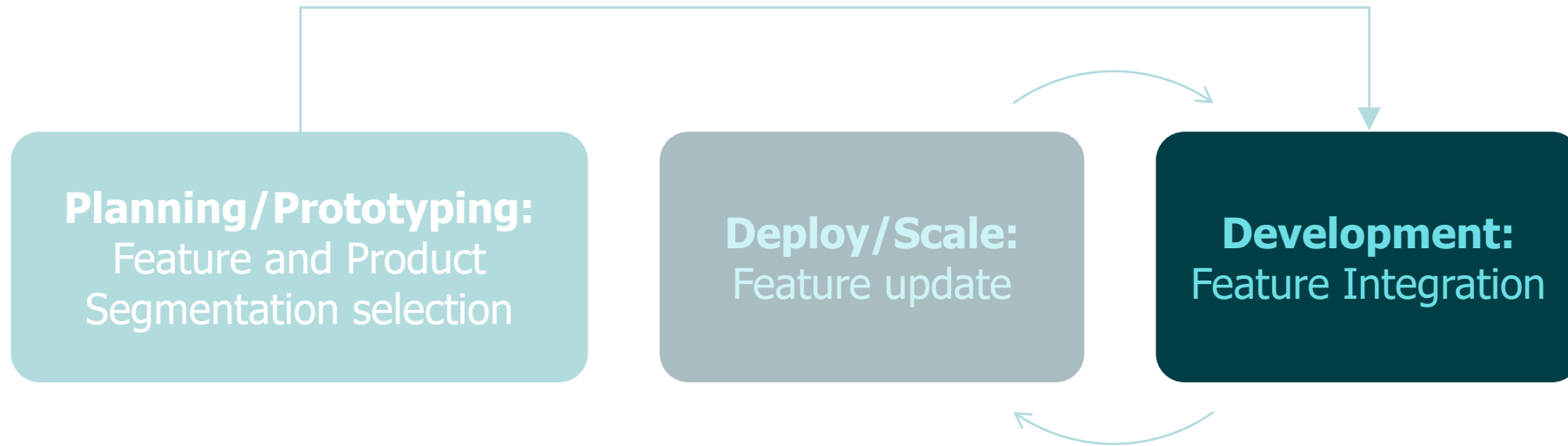Chainer   TensorFlow

Tools

Profiler   Debugger

Deployment

Qualcomm

Developer challenge around efficient end-to-end MM/ML pipeline deployment

Iterative process on embedded:

Planning/Prototyping:
Feature and Product Segmentation selection
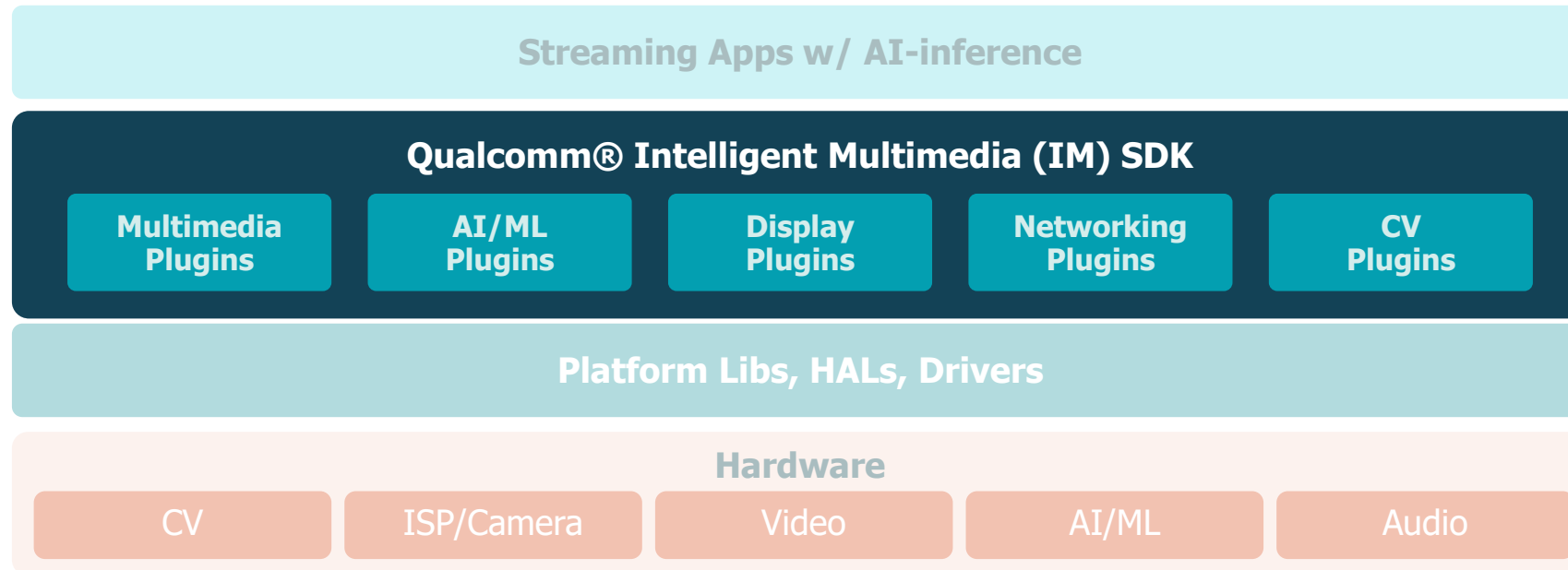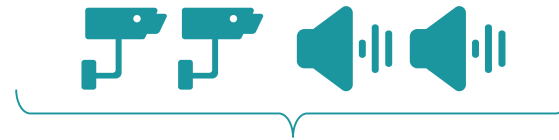
Deploy/Scale:
Feature update

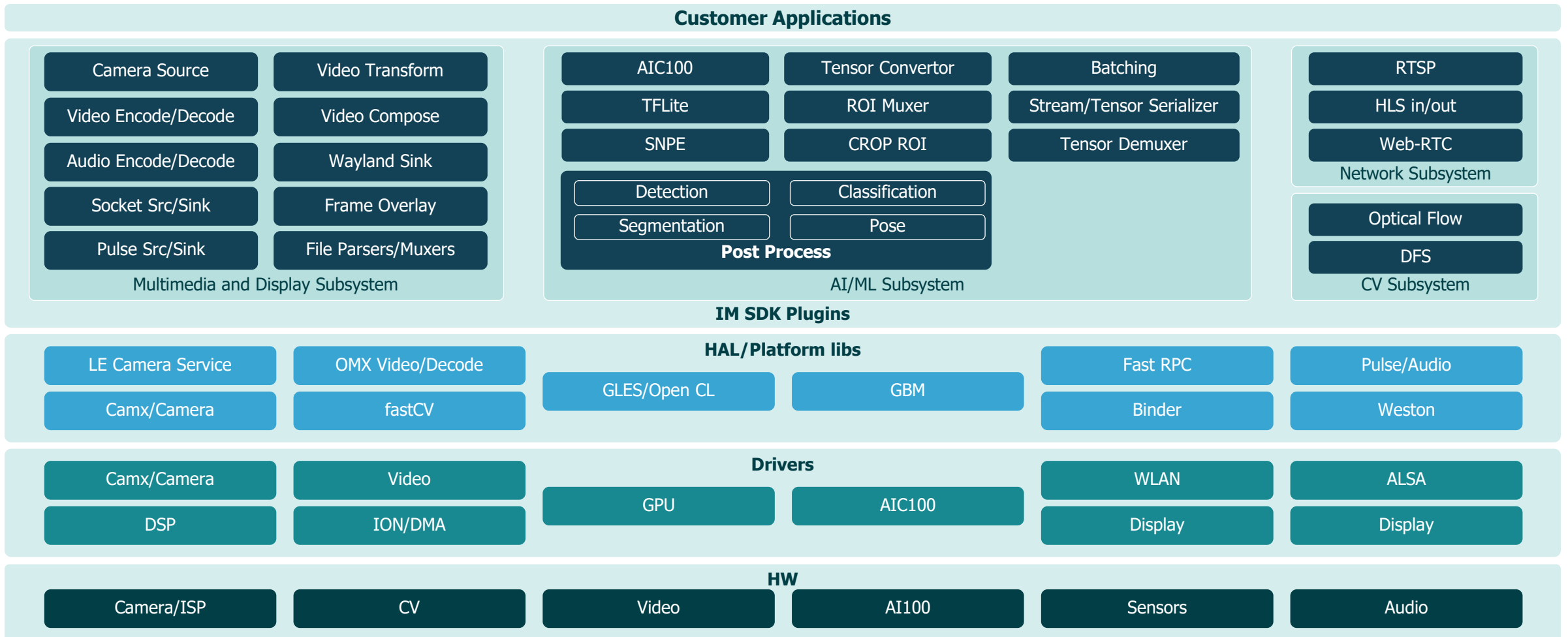**Development:**
Feature Integration

Pipeline creation (end to end application development) | **Memory management** | **Debugging/Profiling/Optimization**

Qualcomm® Intelligent Multimedia SDK for end to end MM/ML pipeline deployment

Streaming Apps w/ AI-inference

Qualcomm® Intelligent Multimedia (IM) SDK

| Multimedia Plugins | AI/ML Plugins | Display Plugins | Networking Plugins | CV Plugins |

Platform Libs, HALs, Drivers

Hardware

| CV | ISP/Camera | Video | AI/ML | Audio |

Modular | Scalable | Optimized

Qualcomm IM SDK is a product of Qualcomm Technologies, Inc. and/or its subsidiaries.

Qualcomm IM SDK Architecture

Customer Applications

Multimedia and Display Subsystem

Camera Source	Video Transform
Video Encode/Decode	Video Compose
Audio Encode/Decode	Wayland Sink
Socket Src/Sink	Frame Overlay
Pulse Src/Sink	File Parsers/Muxers

AI/ML Subsystem

AIC100	Tensor Convertor	Batching
TFLite	ROI Muxer	Stream/Tensor Serializer
SNPE	CROP ROI	Tensor Demuxer

Post Process

Detection	Classification
Segmentation	Pose

Network Subsystem

RTSP
HLS in/out
Web-RTC

CV Subsystem

Optical Flow
DFS

IM SDK Plugins

HAL/Platform libs

LE Camera Service	OMX Video/Decode
Camx/Camera	fastCV

GLES/Open CL	GBM

Fast RPC	Pulse/Audio
Binder	Weston

Drivers

Camx/Camera	Video
DSP	ION/DMA

GPU	AIC100

WLAN	ALSA
Display	Display

HW

Camera/ISP	CV	Video	AI100	Sensors	Audio

Qualcomm IM SDK is a unified SW SDK across our IoT chipsets

LOW

MID
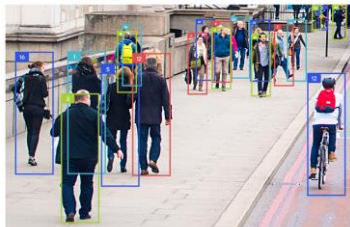
HIGH

PREMIUM

- 1-2 camera stream
- <1TOPs/stream

- Upto 4 camera streams
- 1-2 TOPs/stream
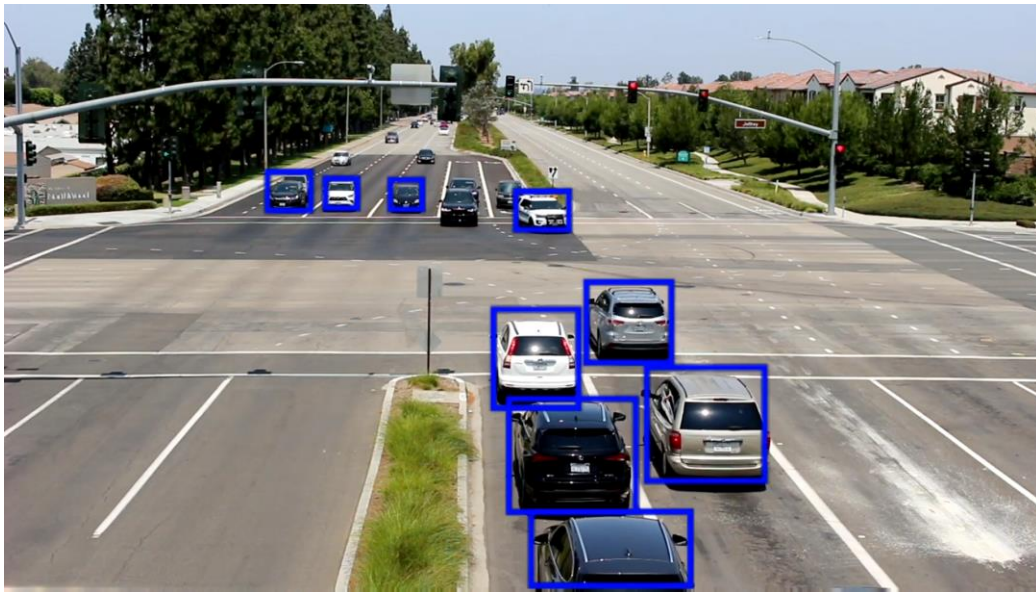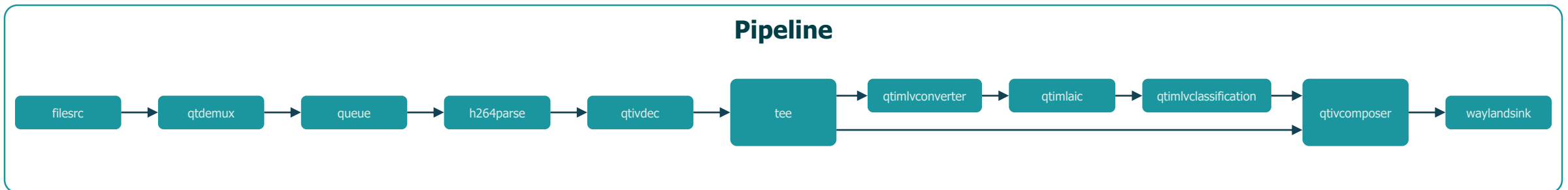
- Upto 16 camera streams
- 2-4 TOPs/stream

- >16 streams
- >4 TOPs/stream

Unified across our IoT chipsets

Qualcomm

Live Camera Inference: object detection

- Runs Inference on live stream to find objects in frame and highlight with bounding boxes.

- Returns multiple bounding boxes (bbox) with score and label name

- Inference results - draw bbox around objects with text per frame

Pipeline

filesrc → qtdemux → queue → h264parse → qtivdec → tee → qtimlvconverter → qtimlaic → qtimlvclassification → qtivcomposer → waylandsink

Creating a reference application

Creating a reference application


```
sh-4.4# export XDG_RUNTIME_DIR=/usr/bin/weston_socket && gst-launch-1.0
```

Launch GStreamer   Identify all required plugins & configure parameters   Connect plugins & Launch

Creating a reference application

```
sh-4.4# export XDG_RUNTIME_DIR=/usr/bin/weston_socket && gst-launch-1.0 qtivcomposer name=mixer
sink_1::dimensions="<1920,1080>" ! queue ! waylandsink x=100 y=100 width=1280 height=720 enable-
last-sample=false sync=true filesrc location=/data/video.mp4
```

Launch GStreamer   **Identify all required plugins & configure parameters**   **Connect plugins & Launch**

Creating a reference application


```
sh-4.4# export XDG_RUNTIME_DIR=/usr/bin/weston_socket && gst-launch-1.0 qtivcomposer name=mixer
sink_1::dimensions="<1920,1080>" ! queue ! waylandsink x=100 y=100 width=1280 height=720 enable-
last-sample=false sync=true filesrc location=/data/video.mp4 ! qtdemux ! queue ! h264parse ! qti
vdec ! queue ! tee name=t ! queue ! mixer. \ t. ! queue ! qtimlvconverter ! queue ! qtimlaic dev
ices="<0>" activations=1 model=/data/yolov5m/programqpc.bin ! queue ! qtimlvdetection threshold=
60.0 results=10 module=ssd-yolo labels=/data/yolov5_detection.labels ! video/x-raw,width=480,hei
ght=270 ! queue ! mixer.
```

Launch GStreamer **Identify all required plugins & configure parameters** **Connect plugins & Launch**
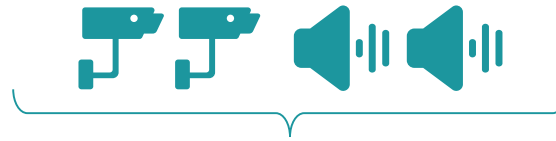
Multi-Inferencing

- Runs multiple inferences on 16 live streams to find objects in frame and highlight with bounding boxes.

- Inference results - draw bounding boxes around objects with different colors for different objects

- Runs optical flow concurrently on a stream & display motion vector drawing

Multi-Inferencing

Conclusion

Streaming Apps w/ AI-inference

Qualcomm IM SDK

| Multimedia Plugins | AI/ML Plugins | Display Plugins | Networking Plugins | CV Plugins |

Platform Libs, HALs, Drivers

Hardware

| CV | ISP/Camera | Video | AI/ML | Audio |
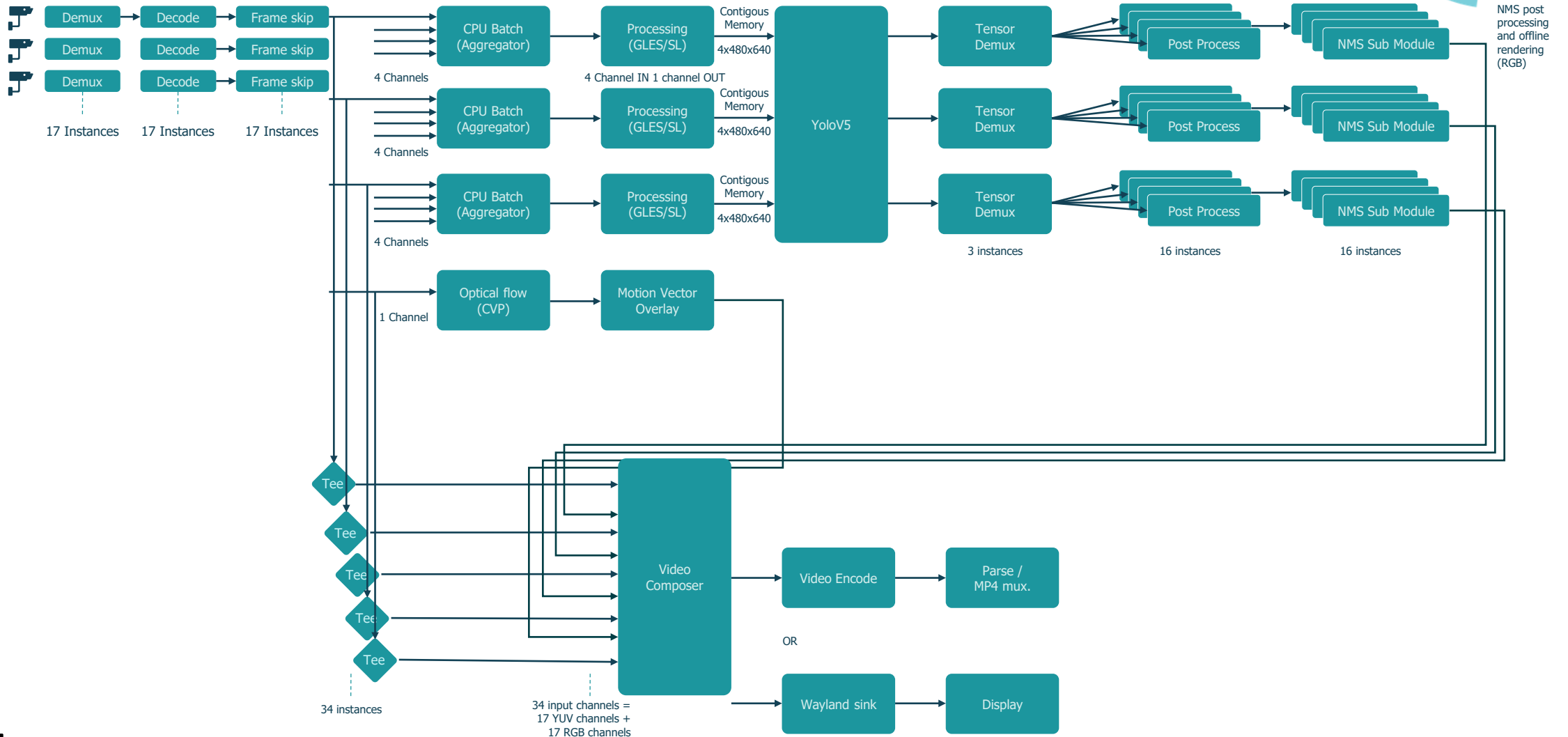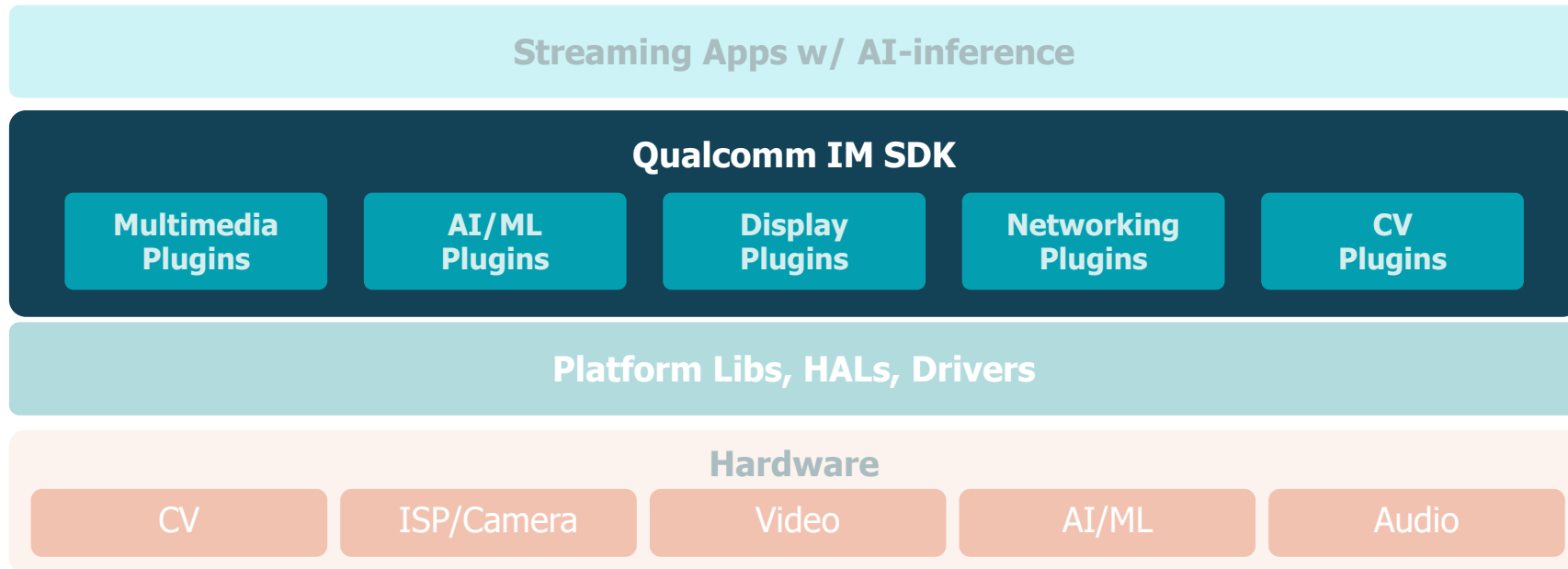
To resolve fragmentation problem, need for "IoT as a simple software stack" is eminent.

Qualcomm IM SDK act as the major building block of the software stack for seamless end to end application deployment.

Qualcomm

Thank you

Qualcomm