基于Intel架构的实时视频流分析系统的设计与优化

翟磊 英特尔开源技术中心
2018.10.19-20 北京丽亭华苑酒店

EMERGENCY RESPONSE

FINANCIAL SERVICES

CITIES/TRANSPORTATION
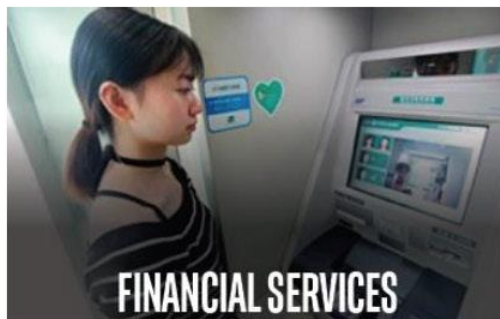
VIDEO: THE "EYE OF IOT"

USE OF VIDEO, COMPUTER VISION, AND DEEP LEARNING IS GROWING RAPIDLY

AUTONOMOUS VEHICLES

RESPONSIVE RETAIL

MANUFACTURING

基于深度学习的视频分析需求增长迅猛

Deep learning revenue is estimated to grow from $655M in 2016 to $35B by 2025[1]

Market Opportunities + Advanced Technologies have Accelerated Deep Learning Adoption

实时视频流分析典型流程 (人脸识别)

Store meta data

Full resolution → Crop, Scale → **FaceRecognition**

Decode → CSC, Scale → **FaceDetection**

ROI coordinates

Crop, Scale

Full resolution → Composition → Display / Encode

- Up to 30fps
- Up to 4K
- Typically FHD/HD
- Mostly H.264
- HEVC demand on the rise
- Input could be raw video

- CSC to planar RGB/BGR
- Scale to match NN input res.

- E.g. intel-facedetection-retail in OpenVINO
- Potentially dozens of faces/frame

- Crop ROI coordinates
- Scale to match NN input res.

- Multiple instances
- One per detected face
- FaceNet

- Optional
- Draw bounding box(es), text

- Optional
- H.264 today
- HEVC needed

| Complexity | Performance | Scalability |

Intel平台对视觉计算全方位支持

LiveVideo StackCon
音视频技术大会

EXPERIENCES

TOOLS

Intel® Media SDK/Media Server Studio   OpenVINO™ toolkit
Intel® System Studio   Intel® SDK for OpenCL™ Applications
Intel® Parallel Studio XE

UNLEASH FULL POTENTIAL

FRAMEWORKS

Spark
MLib BigDL
TensorFlow
mxnet
theano
neon
torch
Caffe

LIBRARIES

Intel® Data Analytics Acceleration Library

Intel® Distribution for python
Intel® Math Kernel Library

Intel® Nervana™ Graph

saffron TECHNOLOGY

Movidius Stack

HARDWARE

intel ATOM inside   intel QUARK inside   intel CORE i7 inside   intel XEON inside   intel ARRIA 10 inside

Compute   Memory & Storage   Networking   Visual Intelligence

intel REALSENSE   intel Movidius

OpenVINO 工具包
Open Visual Inference & Neural Network Optimization

Enables deep learning on hardware accelerators and easy heterogeneous execution across Intel platforms

Intel® Deep Learning Deployment Toolkit

Model Optimizer
Convert & Optimize

IR

Inference Engine
Optimized Inference

20+ Pre-trained Models

Computer Vision Algorithms

Samples

IR = Intermediate Representation file

Traditional Computer Vision Tools & Libraries

Optimized Libraries

OpenCV*

OpenVX*

Photography Vision

Code Samples

For Intel® CPU & CPU with integrated graphics

Increase Media/Video/Graphics Performance

Intel® Media SDK
Open Source version

OpenCL™ Drivers & Runtimes

For CPU with integrated graphics

Optimize Intel® FPGA

FPGA RunTime Environment
(from Intel® FPGA SDK for OpenCL™)

Bitstreams

FPGA – Linux* only

OS Support CentOS* 7.4 (64 bit) Ubuntu* 16.04.3 LTS (64 bit) Microsoft Windows* 10 (64 bit) Yocto Project* version Poky Jethro v2.0.3 (64 bit)

Intel® Architecture-Based Platforms Support

(intel) CELERON inside

(intel) ATOM inside

(intel) CORE inside

(intel) XEON inside

(intel) ARRIA 10 inside

(intel) Movidius

(intel) IRIS Pro GRAPHICS

Intel深度学习部署工具包
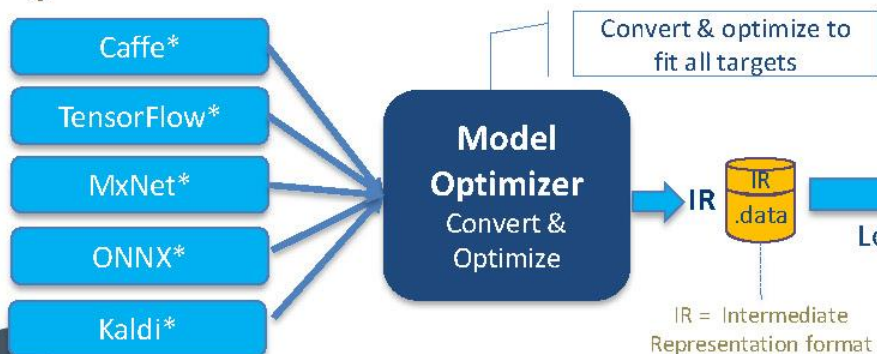## Intel Deep Learning Deployment Toolkit

Model Optimizer

- **What it is**: Preparation step -> imports trained models

- **Why important**: Optimizes for performance/space with conservative topology transformations; biggest boost is from conversion to data types matching hardware.
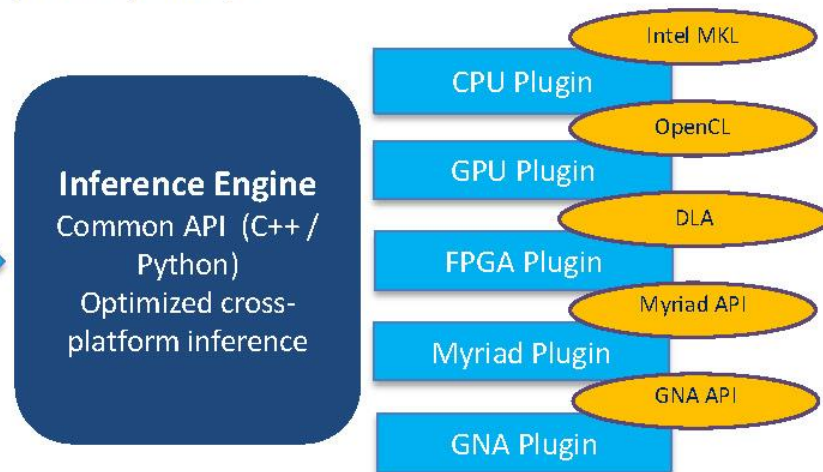
Inference Engine

- **What it is**: High-level inference API

- **Why important**: Interface is implemented as dynamically loaded plugins for each hardware type. Delivers best performance for each type without requiring users to implement and maintain multiple code pathways.

Trained Model

Caffe*

TensorFlow*

MxNet*

ONNX*

Kaldi*

Convert & optimize to fit all targets

Model Optimizer
Convert & Optimize

IR
.data

IR = Intermediate Representation format

Load, infer

Inference Engine
Common API (C++ / Python)
Optimized cross-platform inference

Intel MKL

CPU Plugin

OpenCL

GPU Plugin

DLA

FPGA Plugin

Myriad API

Myriad Plugin

GNA API

GNA Plugin

GPU = Intel CPU with integrated graphics processing unit/Intel® Processor Graphics

OpenVINO工具包优势

ACCELERATE PERFORMANCE

Access Intel computer vision accelerators. Speed code performance.
Supports heterogeneous processing & asynchronous execution.

INTEGRATE DEEP LEARNING

Unleash convolutional neural network (CNN) based deep learning inference
using a common API, pre-trained models & computer vision algorithms.

Up to 19.9x increase[1]

SPEED DEVELOPMENT

Reduce time using a library of optimized OpenCV* & OpenVX* functions, & 15+ samples.

Develop once, deploy for current & future Intel-based devices.

INNOVATE & CUSTOMIZE

Use the increasing repository of OpenCL™ starting points in OpenCV* to add your own unique code.

Intel Media SDK & Intel Media Server Studio

Deliver fast, high quality video transcoding from camera to cloud

What it is: An API to access Intel® Quick Sync Video hardware-accelerated encode/decode & processing

- H.265 (HEVC)
- H.264 (AVC)
- MPEG-2, VP9 & more

- Resize, Scale, Deinterlace
- Color Conversion, Composition
- Denoise, Sharpen & more

Benefits

- **Boost media and video application performance** with hardware-accelerated codecs & programmable graphics on Intel® processors.**

- **Speed transition to higher frame rates & resolutions.**

- **Improve video quality, innovate cloud graphics & media analytics.**

- **Reduce infrastructure & development costs.**

FREE Downloads
software.intel.com/tools-by-segment/media ▶

Intel Media Server Studio卓越媒体转码性能

Intel® Media Server Studio for Linux*

Multistream Performance (1xRT=30fps)		Number of Real-time (30fps) streams	Number of Real-time (60fps) streams
1080p-to-1080p	**AVC-to-HEVC**	**15**	**7**
	HEVC-to-HEVC	**8**	**4**
4K-to-4K	**AVC-to-HEVC**	**4**	**2**
	HEVC-to-HEVC	**2**	**1**

Up to 2 Real-time HEVC streams per processor[1]
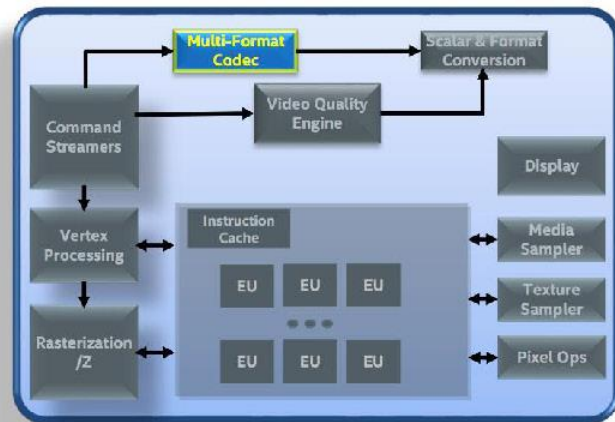
E3-1500 v5 HEVC is fully accelerated targeting 4K60 capability

Intel Collaboration Suite for WebRTC 实时视频分析实践

An end-to-end real-time media delivery platform based on WebRTC http://webrtc.intel.com
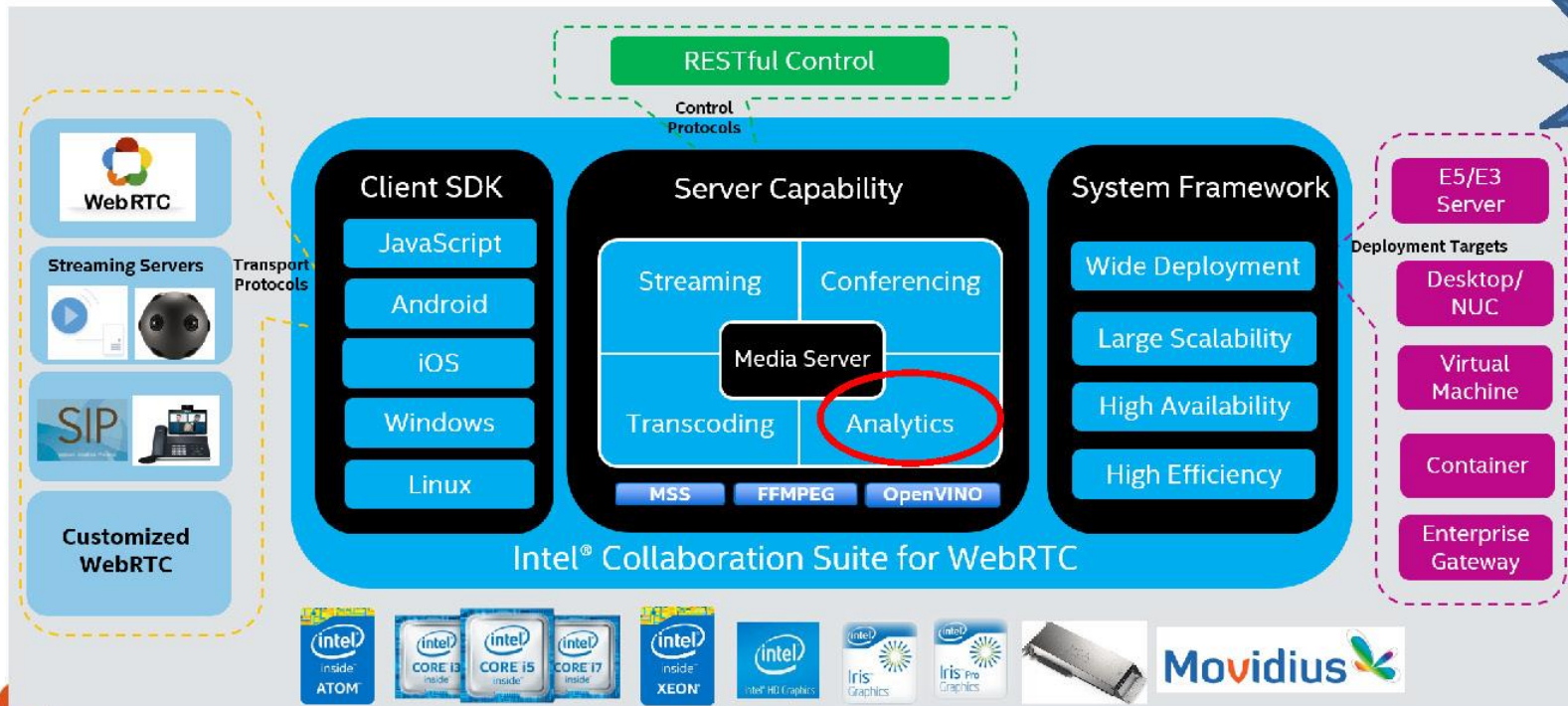
Intel CS for WebRTC 典型应用场景

VIDEO CONFERENCE

MEDICAL

EDUCATION

AUTONOMOUS DRIVING

ONLINE BROADCASTING

SOCIAL MEDIA

CLOUD

TELECOM
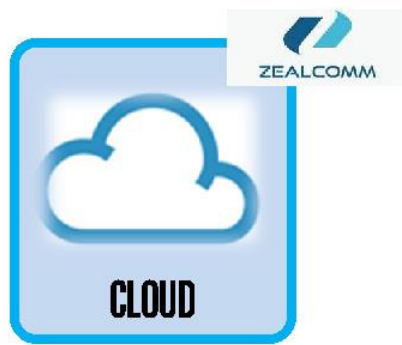
Intel CS for Media Analytics实时视频流分析系统架构

ICS for WebRTC

| Streaming | Transcoding | Conferencing |

Video Stream for Analysis

New composed video Stream
Detected Events

ICS for Media Analytics

WebRTC Accessing

| Decoding | Image Pre-processing | Analytics Plugin System | Image Post-processing | Encoding |

Intel Media Server Studio | OpenVINO | FFmpeg

Session Management

Node Scheduling
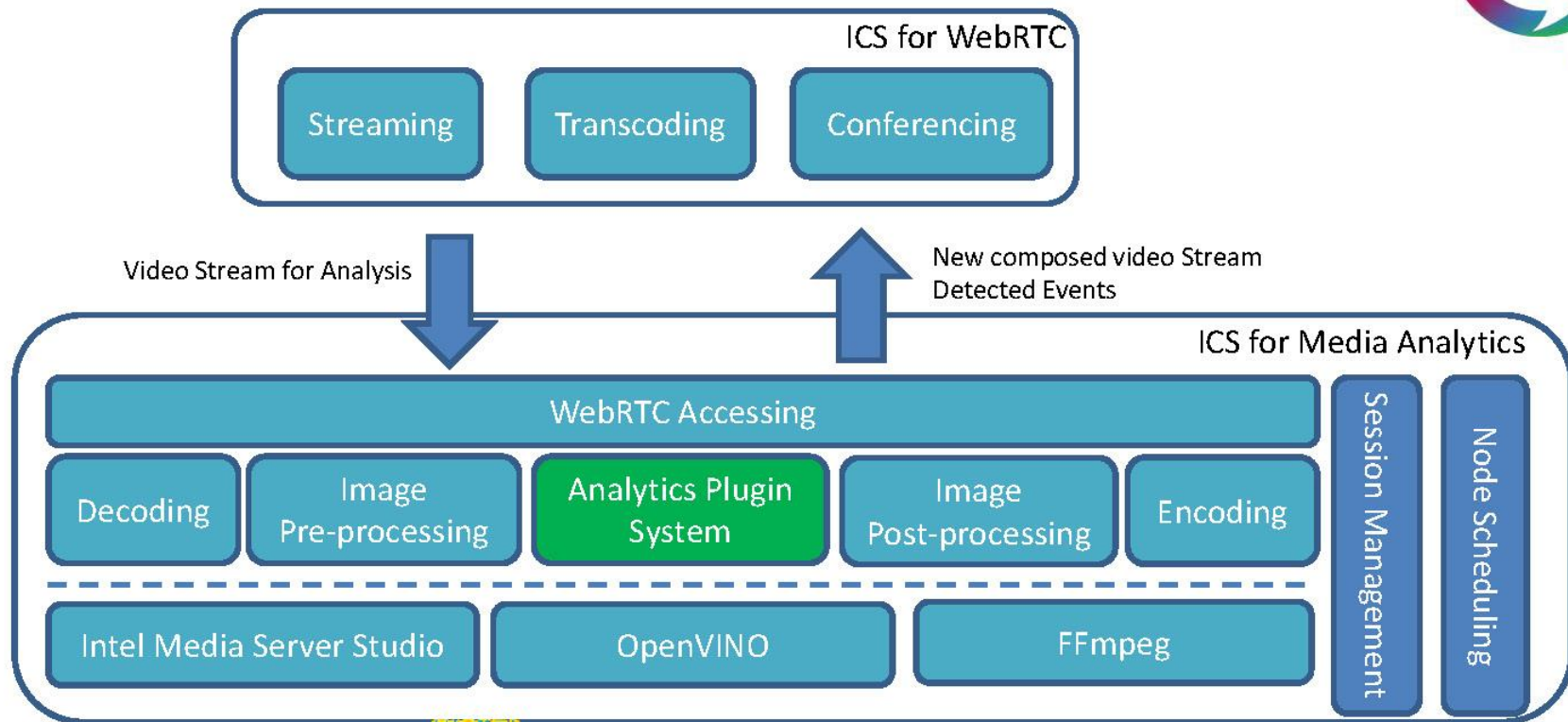
Intel CS for Media Analytics Plugin设计
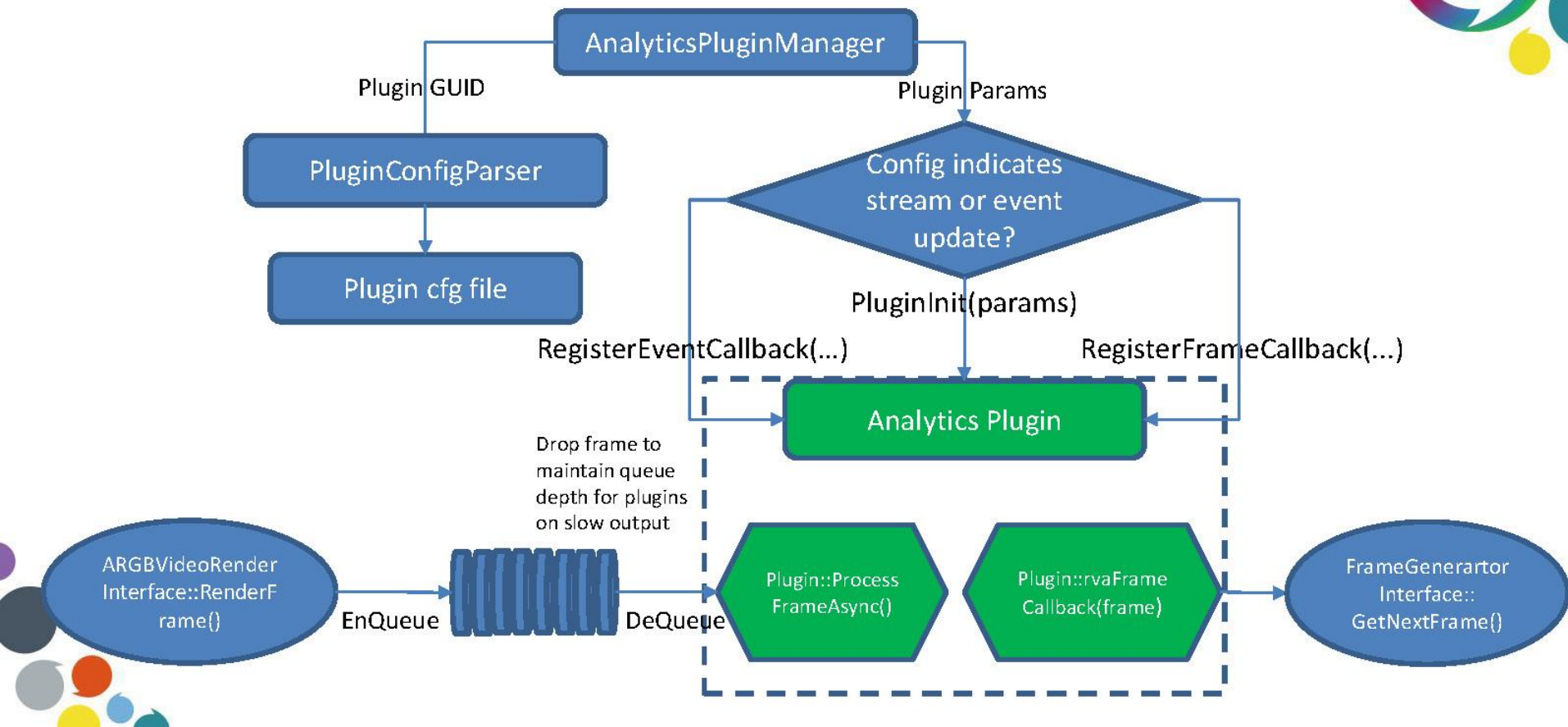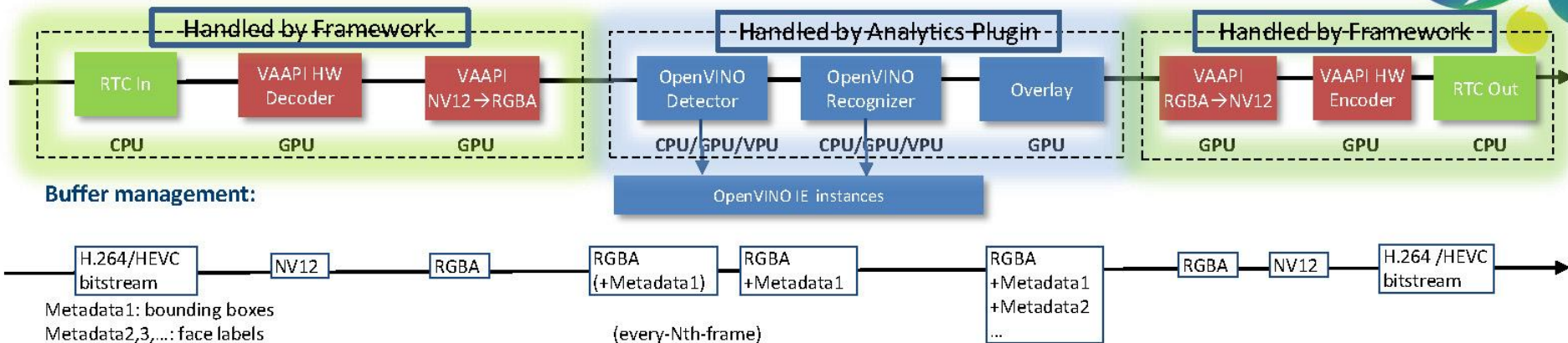
Intel CS for Media Analytics分析全流程GPU加速

Single-channel source, face detection inference followed by batching face recognition inference according to detection result

Buffer management:

Metadata1: bounding boxes
Metadata2,3,…: face labels

(every-Nth-frame)

Legend:
- RTC components
- MSS components
- User Analytics Plugins
- Component boundary

Input video size: 1080p(tested on E3-1585L)

27% boost

SW w/o buffer sharing   HW w/ shared GPU mem
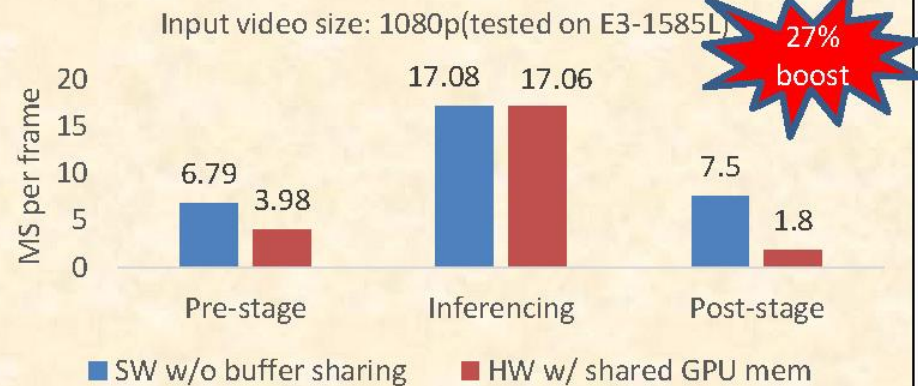
Intel CS for Media Analytics 系统负载优化
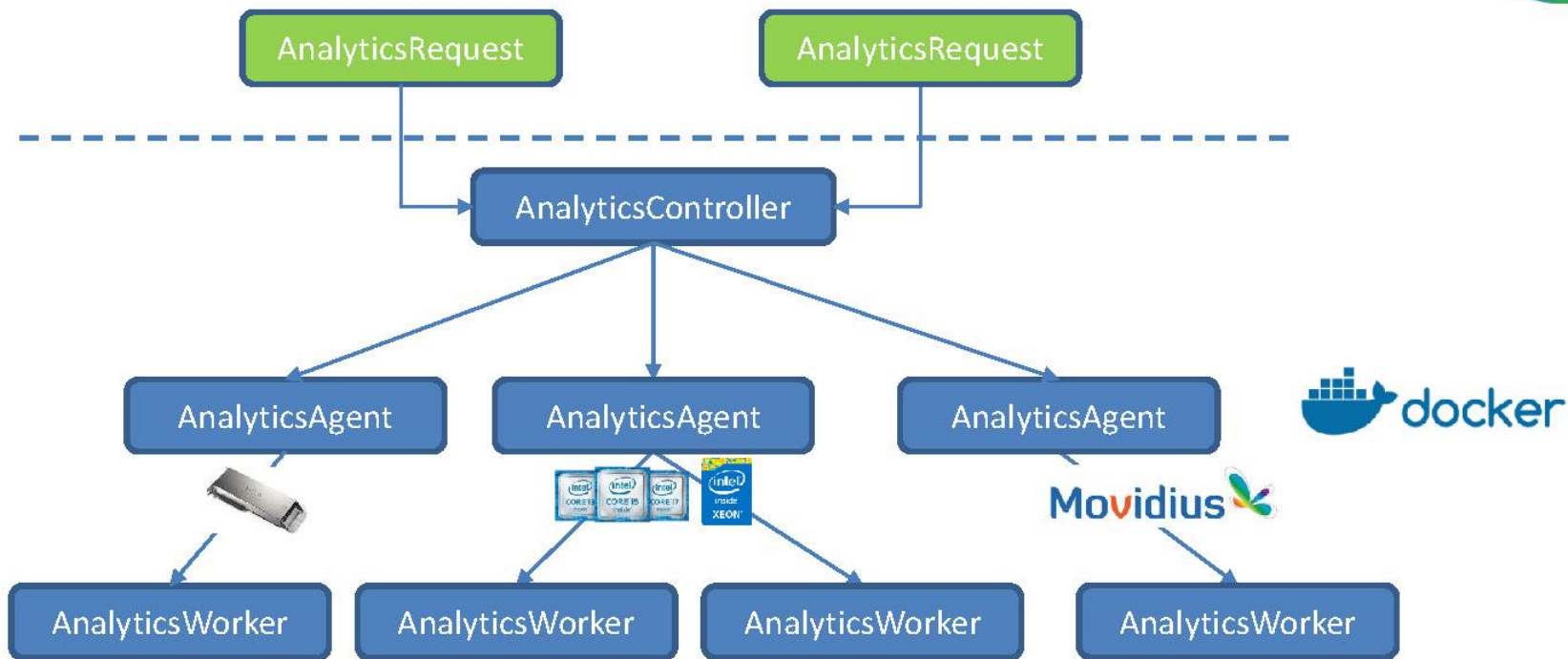
Analytics Agent Docker 节点Intel GPU加速实践

Analytics Worker A
OpenVINO

Analytics Worker B
OpenVINO

Docker

/dev/dri/card0
/dev/dri/renderD128

Host Operating System

(intel) IRIS Pro GRAPHICS Infrastructure
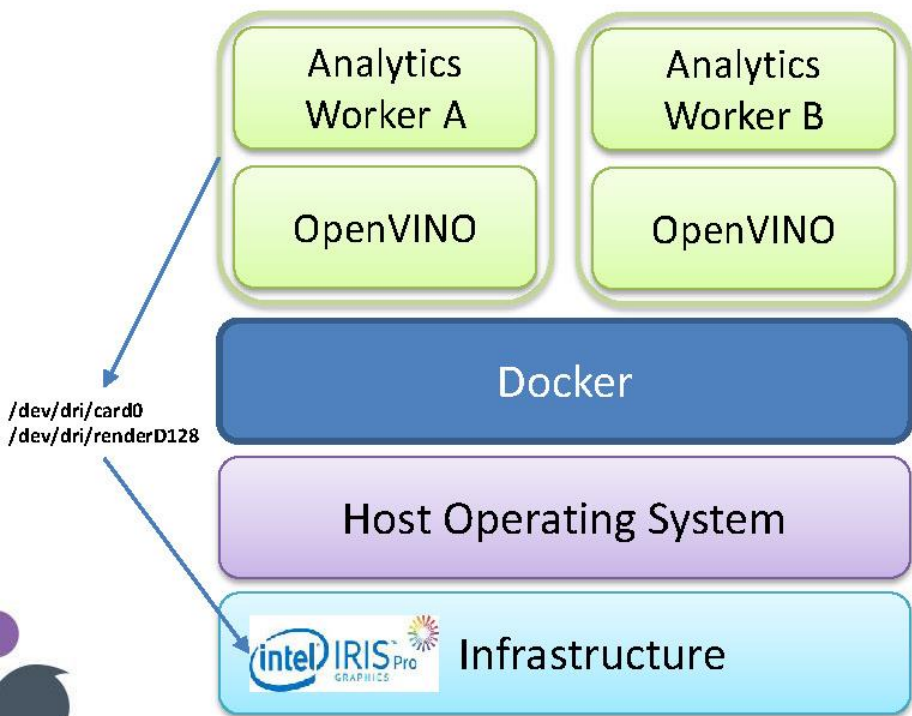
- Docker pull Ubuntu:16.04 image
- **Run docker image w/ GPU enabled,**

  --device=/dev/dri/card0 --device=/dev/dri/renderD128
- Install dependency in container, *apt install –y cpio lsb*
- Install OpenCL™ NEO driver to enable OpenVINO GPU
- Install OpenVINO 2018R3 as *root*, see Installation Guide
- Run demo

  - *source /opt/intel/computer_vision_sdk/bin/setupvars.sh*
  - *cd /opt/intel/computer_vision_sdk/deployment_tools/demo*
  - add param *"-no_show"* to "./security_barrier_camera_sample" in *demo_security_barrier_camera.sh*
  - *./ demo_security_barrier_camera.sh –d GPU*

LiveVideo
StackCon
音视频技术大会

Intel CS for Media Analytics示例 (智能课堂点名)

Take the face recognition in classroom for example, to enable auto-call-the-roll functionality, a few challenges:

- Resource intensive due to dozens of faces

- Loss of accuracy due to scaling

- Faces at sub-picture boundaries

Raw Image source: http://beiwoo.cn/yangyanji/dabaoyanfu/2012/0829/16172_2.html

Intel CS for Media Analytics示例 (智能课堂点名)

Face Detection in sub-region

Enlarge ROI on boundary and perform detection again

Mapping back to original video

Face recognition on each sub-region

总结

Intel provided comprehensive platform and toolkit for real-time video analytics based on deep learning

Intel CS for Media Analytics + Intel Media Server Studio + OpenVINO established easy-to-use, efficient and scalable real-time video analytics framework

Intel is driving open source strategy portfolio to drive visual cloud computing industry based on IA platform
https://www.intel.com/content/www/us/en/cloud-computing/visual-cloud.html

Thank you

LiveVideoStack
音视频技术社区

CSDN