Project Trillium: A New Suite of Machine Learning IP from Arm

arm

Steve Steele, Director Product Marketing, ML Group, Arm

May 23rd, 2018

AI Presents a Significant Opportunity for Innovation

VR/MR

Robotics

Drones

Shipping & Logistics

Automotive

IoT

Home, Surveillance

Medical

Mobile

Servers

The Smartphone is the World's Most Popular AI Device

90% of AI today runs on smartphones*; 95% of the world's smartphones run on Arm

** IDC WW Embedded and Intelligent Systems Forecast, 2017-2022 and Arm forecast*

Why is ML Moving to the Edge?

Bandwidth

Power

Cost

Latency

Reliability

Security

Driven by: the laws of physics, the laws of economics and the laws of the land

Some Example Use Cases

Amiko
Smart Inhaler

EZVIZ C5Si
Camera

Nauto
for Safe Driving

DriveCore
Platform

Amazon
Alexa

Kissfly Drones

Huawei Mate 9

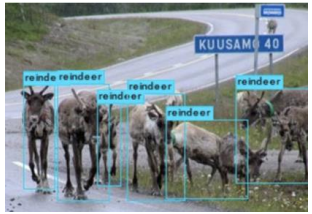
Buddyguard
Flare

Hive View
Camera

BRAGI
'The Dash PRO'

Elli Q

ReindeerCam



Project Trillium: the Arm ML Computing Platform

A suite of Arm ML IP, designed for unmatched versatility and scalability:

The recently announced products

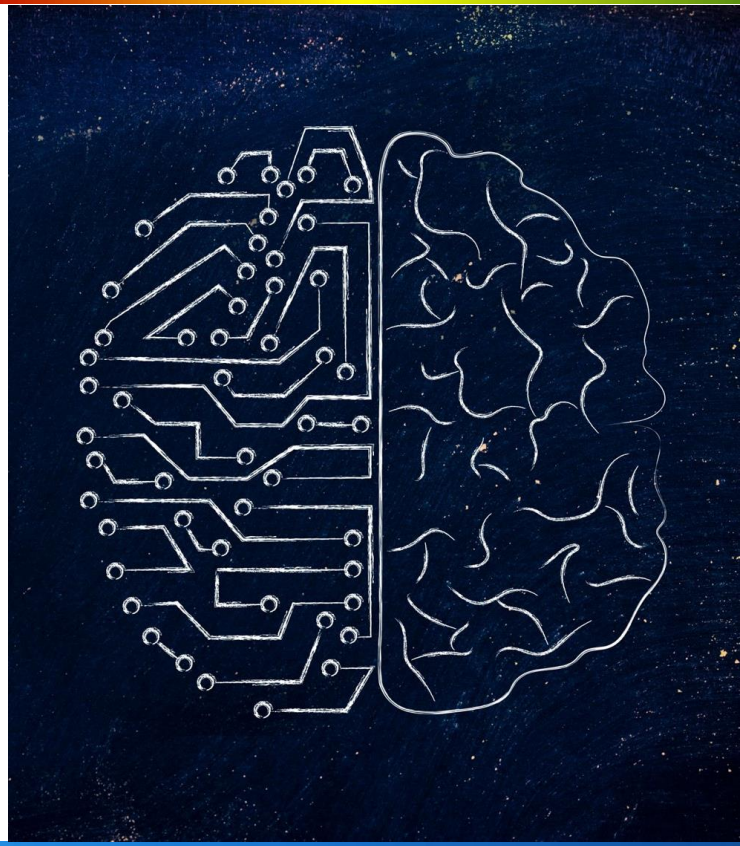
- Machine Learning (ML) processor
- Object Detection (OD) processor
- Neural Network (NN) software libraries

Add to the existing ML capabilities of

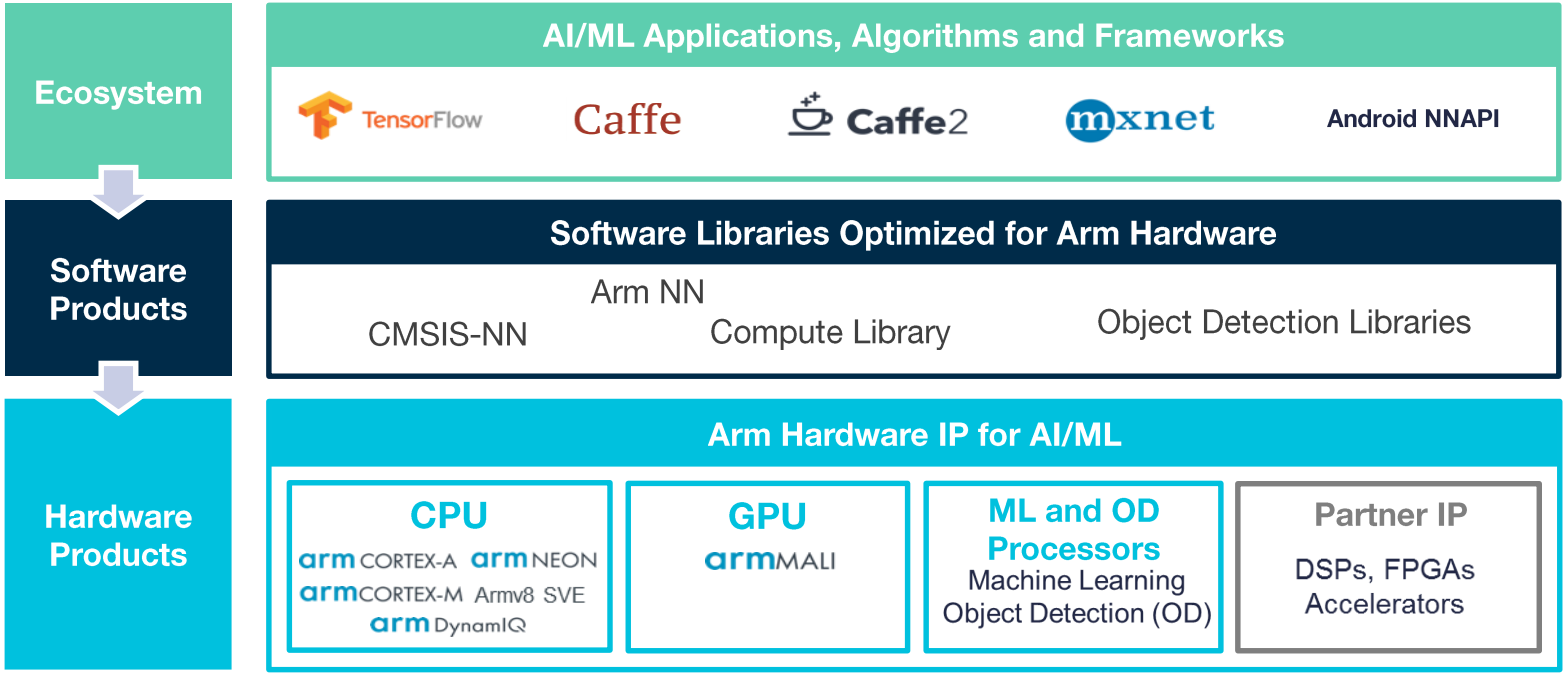
- Cortex-A and Cortex-M CPUs
- Mali GPUs

Market growth in units (today to 2028):

- Mobile – 1.7Bn to 2.2Bn
  (source: Strategy Analytics and Arm forecast)
- Smart IP Cameras – 160M to 1.3Bn
  (source: Gartner and Arm forecast)
- AI-enabled devices – 300M to 3.2Bn
  (source: IDC WW Embedded and Intelligent Systems Forecast, 2017-2022 and Arm forecast)



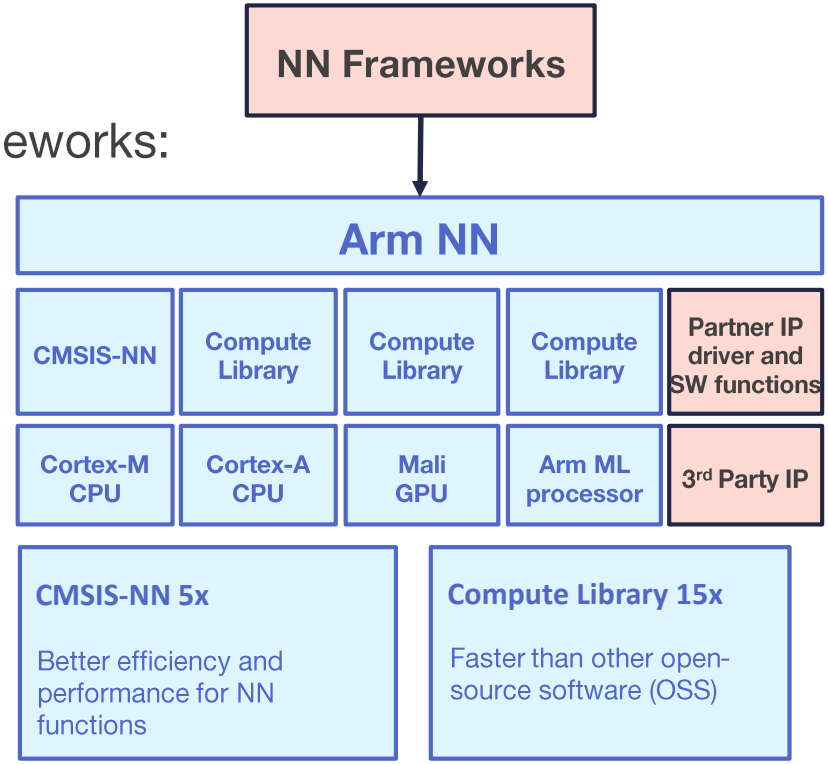
Project Trillium: the Arm ML Computing Platform



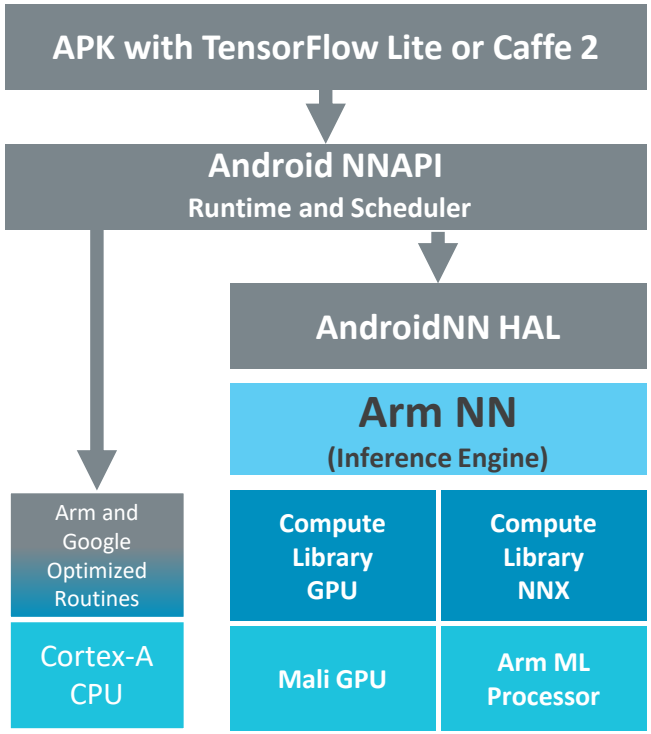
Optimum ML Performance on Arm for any Application

Arm NN software translates existing NN frameworks:

- TensorFlow, Caffe, Android NNAPI, MXNet, etc
- Developers maintain their existing workflow and tools
- Reduces overall development time
- Abstracts away the complexities of the underlying hardware



Arm NN for Android & Linux: Overview



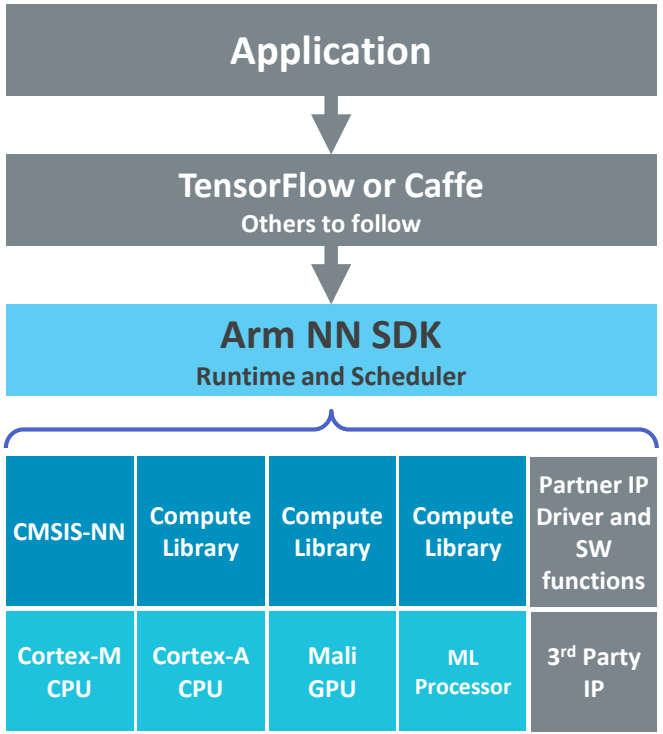
Arm NN providing support for Cortex-A CPUs and Mali GPUs under embedded Linux

Support for Cortex-M in development

Support for ML Processor available on release

arm NN

Arm NN providing support for Mali GPUs under Android NNAPI



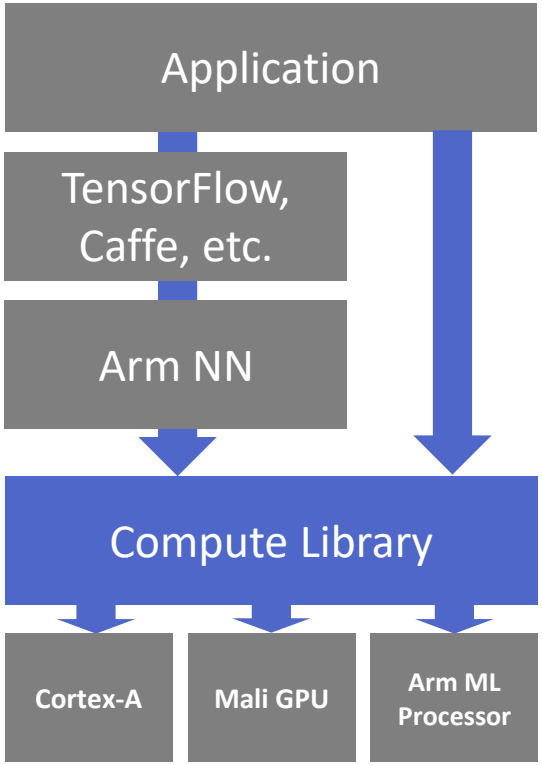
Compute Library

Optimized low-level functions for CPU and GPU
- Most popular CV and ML functions
- Supports common ML frameworks
- Over 80 functions in all
- Up to 15x performance improvement
- Quarterly releases
- CMSIS-NN separately targets Cortex-M

Enable faster deployment of CV and ML
- Targeting CPU (NEON) and GPU (OpenCL)
- Significant performance uplift compared to OSS alternatives

Publicly available now (no fee, MIT license)

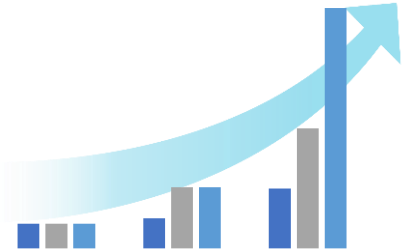
developer.arm.com/technologies/compute-library



ML Support in Cortex CPUs & Mali GPUs

Cortex-A CPU

- 10x SIMD performance improvement in two generations
- Cortex-A v8.2 instruction set with efficient FP16 and 8-bit dot product operation
- Future SVE ISA for general ML performance expansion



Cortex-M CPU

- Optimized Compute Library and CMSIS-NN to improve ML compute
- Small area and power profile with enhanced compute capability for embedded devices

Mali GPU

- Parallel architecture with large compute processing capacity for higher ML performance
- Further improvements for ML planned

Arm ML Processor

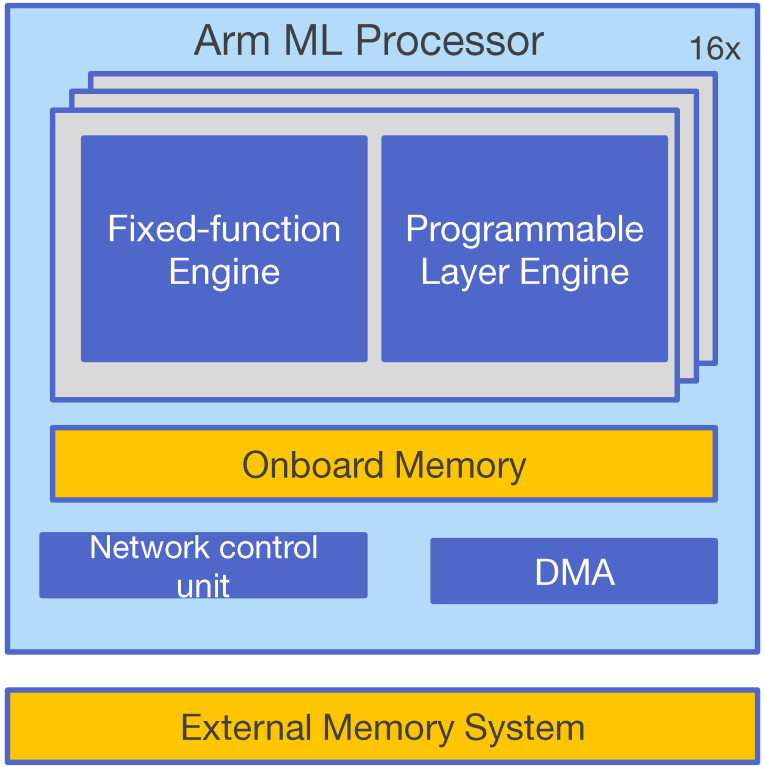
Network control unit
- Overall programmability and high level control flow

Onboard memory
- Central storage for weights and feature maps

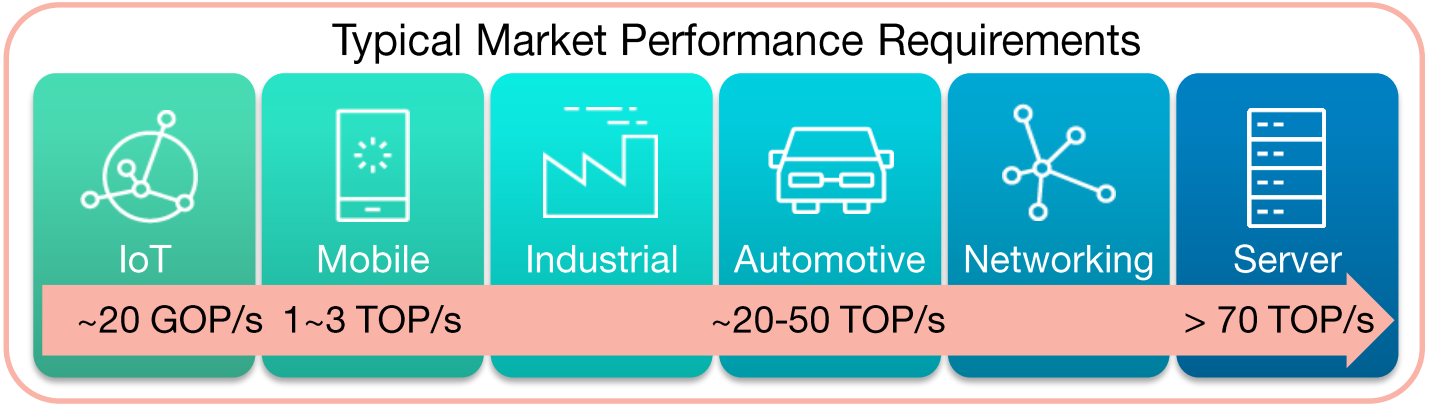
DMA
- Move data in and out of main memory

Fixed-function engines
- Main fixed-function compute engines

Programmable layer engines
- Programmable engines for future proofing



Targeting Multiple Markets with Scalable Architecture



Arm Scalable Machine Learning Processor Architecture

Phase 1

Sensors (2 GOP/s) — Servers 150 TOP/s

Scalable

Compatible

Programmable

Trillions of Operations per Second for Mobile

The Arm ML processor is built on a highly versatile and scalable architecture

The first generation targets the Mobile market for inference at the edge:

- Highest performance per mm^2 in the market
  - Typical mobile performance of >4.6 TOP/s
  - Optimizations provide a further uplift of 2x to 4x in real-world use
- Unmatched performance in thermal- and cost-constrained environments
  - Efficiency of 3 TOP/W[1]
- First IP available to partners mid 2018

[1]*Based on 7nm implementation*

Arm OD Processor

Object Detection processor:

- Second-generation OD processor
- Detects in real time with Full HD @ 60fps
- Object sizes from 50x60 pixels upwards
- Virtually unlimited objects per frame

Object detection with rich characterization:

- Direction people are facing
- Trajectory through robust inter-frame tracking
- Gesture and pose

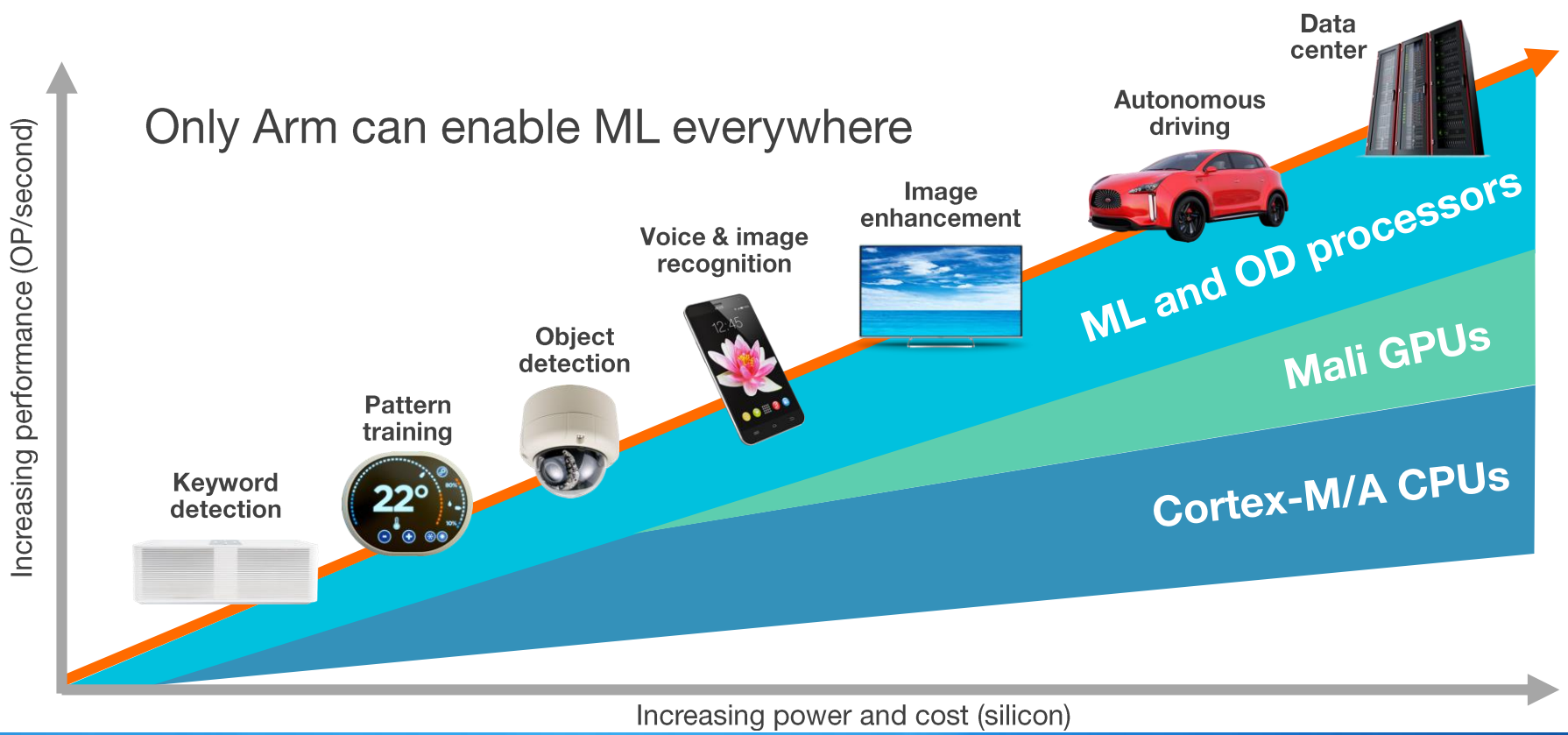
The first-generation OD processor powers Hive and Hikvision security cameras

OD plus ML Processors: a Better User Experience

Combined Arm solution:

- Better user experience with high-resolution, real-time face recognition
- OD processor isolates areas of interest in real time with Full HD @ 60fps
- ML processor analyzes fewer pixels for faster, fine-grain object recognition
- Leads to a new class of smart camera and other vision-based devices

Flexible, Scalable ML Solutions



Mobile Experience: Insight from Advanced Compute

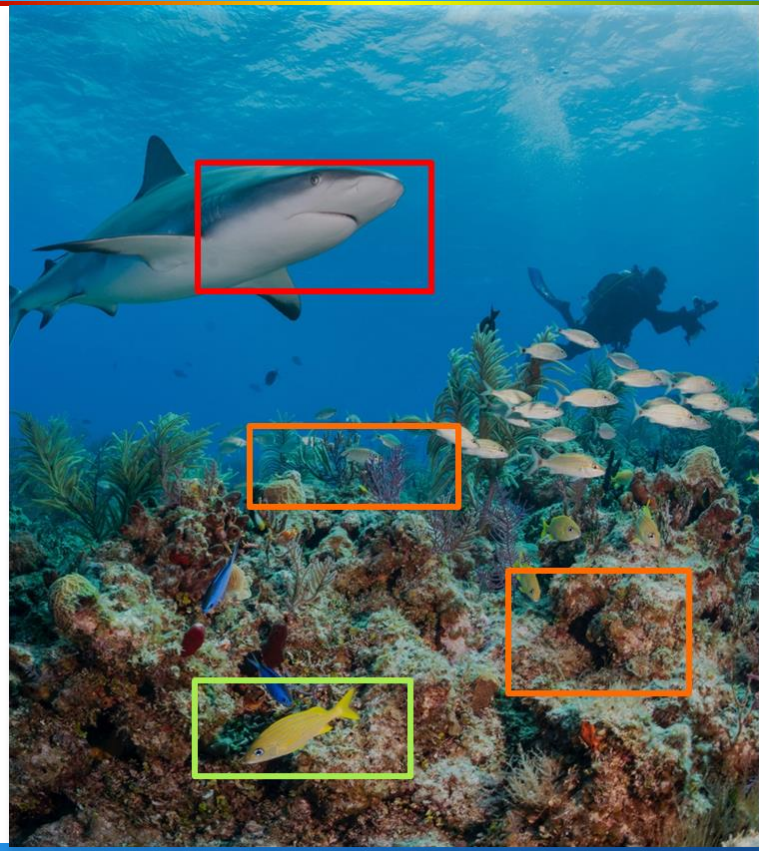
ML and OD processors enable smartphone to be linked to any screen for awareness and protection
(e.g. sunglasses, ski goggles, dive masks)

Blue shark
- Anti-shark suit electric current active
- Dive boat alerted !

Sea anemone
- Poisonous, only touch with gloves !

Beware hole
- Could be Moray eel hideaway !

Bigeye snapper
- Not protected…



Living: Interpreting Data for Smart City Planning

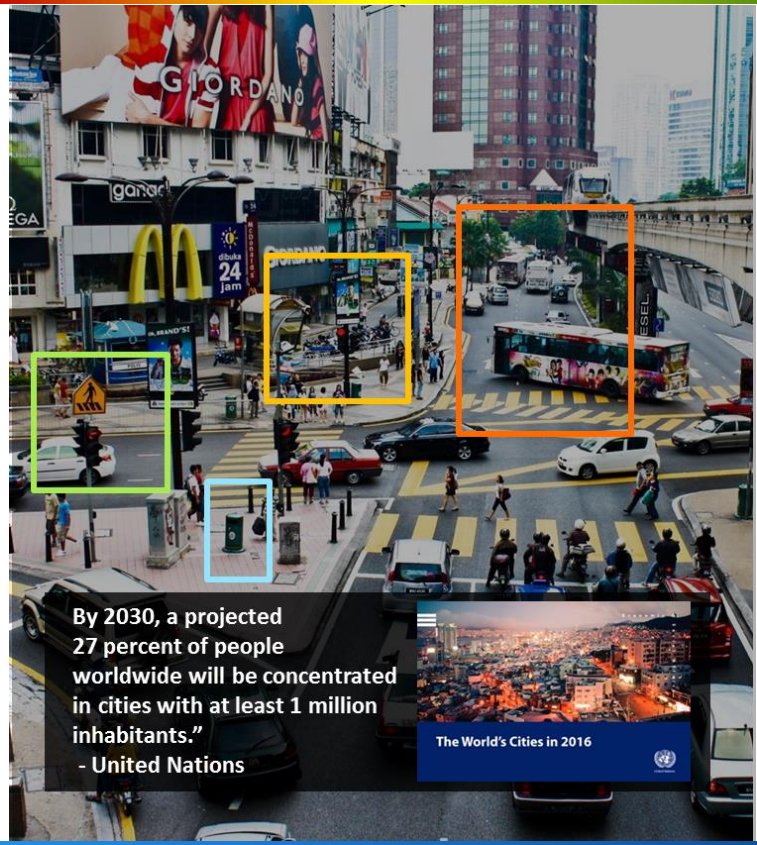
ML and OD processors embedded in city camera systems for real-time information and control

Pedestrians

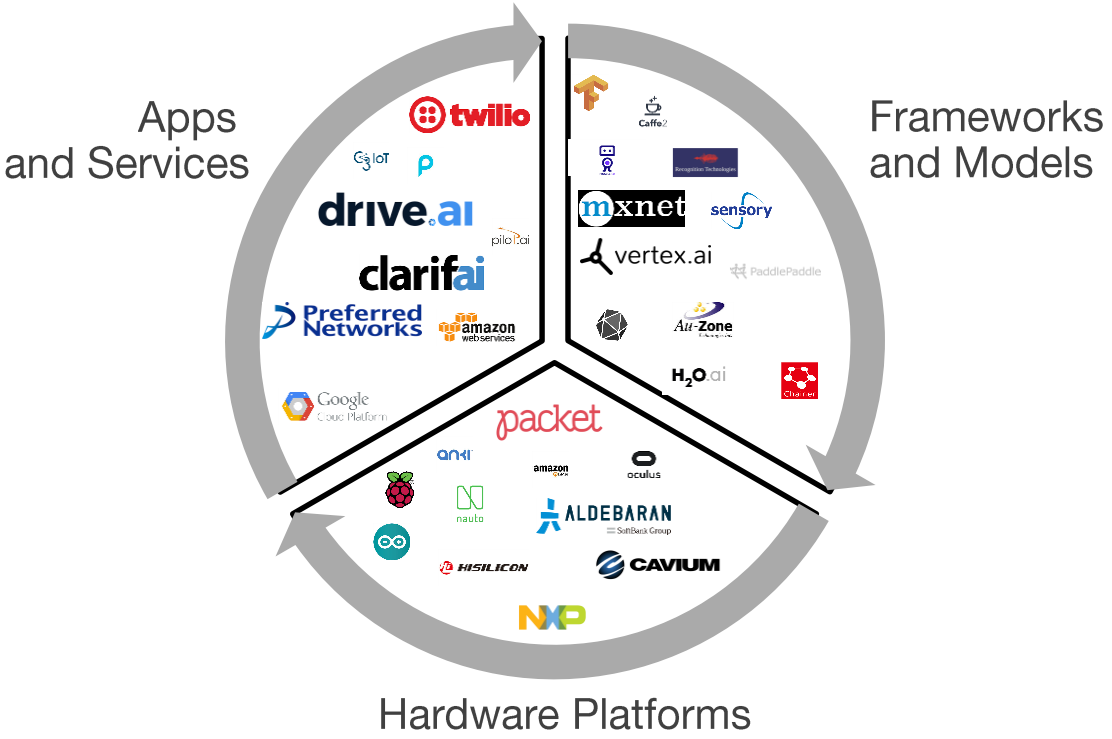
– Detecting pedestrian impedance
– Congestion, overcrowding
– Time-critical safety issues (e.g. lost child)

Traffic

– Recognizing road obstructions
– Reporting accidents
– Enhance with additional information (GPS, Internet)



Machine Learning Ecosystem and Developer Resources



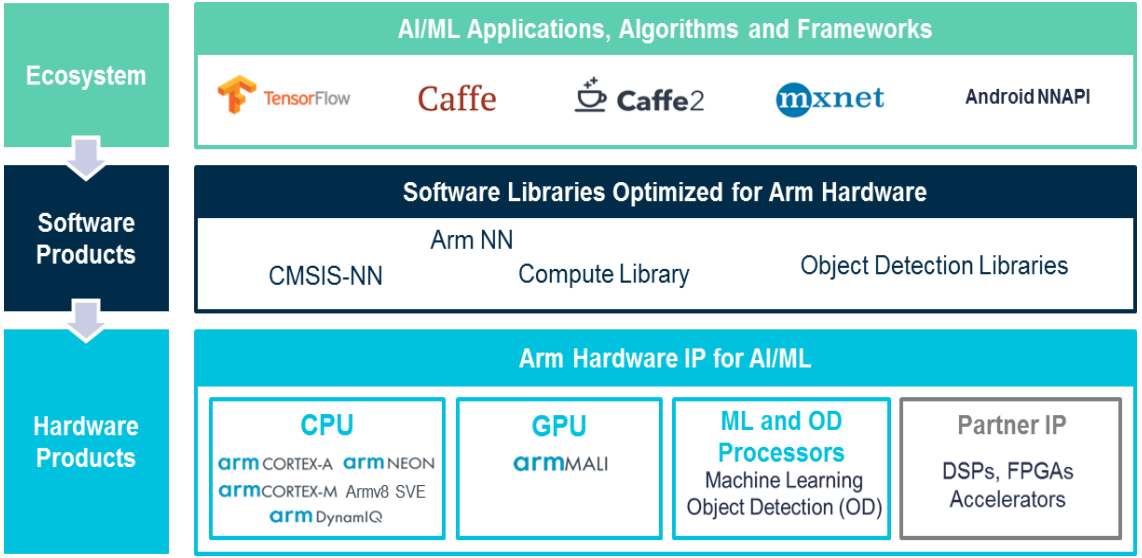
New Developer Site

- Tutorials
- ML in devices
- Blogs
- Videos and webinars
- Arm product details
- ML research and publications

Visit: developer.arm.com/mlcommunity

Project Trillium: Unleashing Innovation for ML on Arm

- ML processor delivers performance of >4.6 TOP/s with efficiency of 3 TOP/W
- OD processor provides object detection and rich characterization in real time with Full HD @ 60fps
- Full suite of Arm NN software supports leading NN frameworks
- Targets mobile and smart camera markets first and then scaling to all devices



Resources:

Developer resources - https://developer.arm.com/mlcommunity
Arm NN - https://github.com/Arm-software/armnn
Compute Library - https://github.com/ARM-software/ComputeLibrary
CMSIS-NN - http://www.keil.com/pack/doc/CMSIS_Dev/NN/html/index.html

Thank You!
Danke!
Merci!
谢谢!
ありがとう!
Gracias!
Kiitos!
감사합니다
धन्यवाद