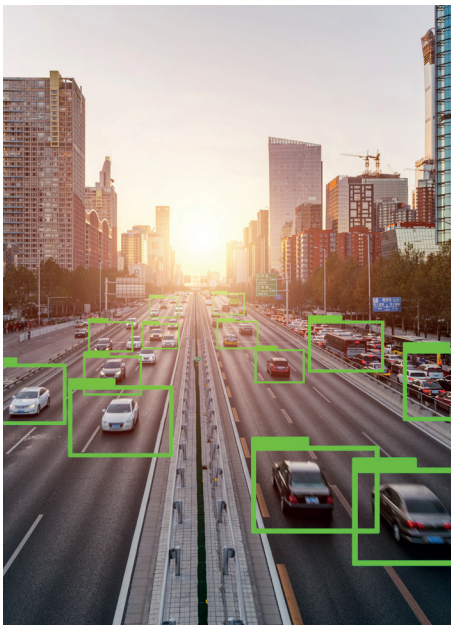
Scalable Machine Learning Solutions for Every Application

arm

Solution Brief



Automotive
Machine learning in automotive

What is Driving ML to the Edge

- Bandwidth
- Power
- Cost
- Latency
- Reliability
- Security

Artificial intelligence (AI) represents the biggest inflection point in computing for more than a generation. As a core enabler for AI, machine learning (ML) has quickly moved from experimental tasks, such as identifying pictures of cats, to solving real-world problems in areas such as healthcare, food production, automotive and retail. Few sectors will remain untouched by its transformative power and arguably few devices – from Internet of Things (IoT) endpoints to servers.

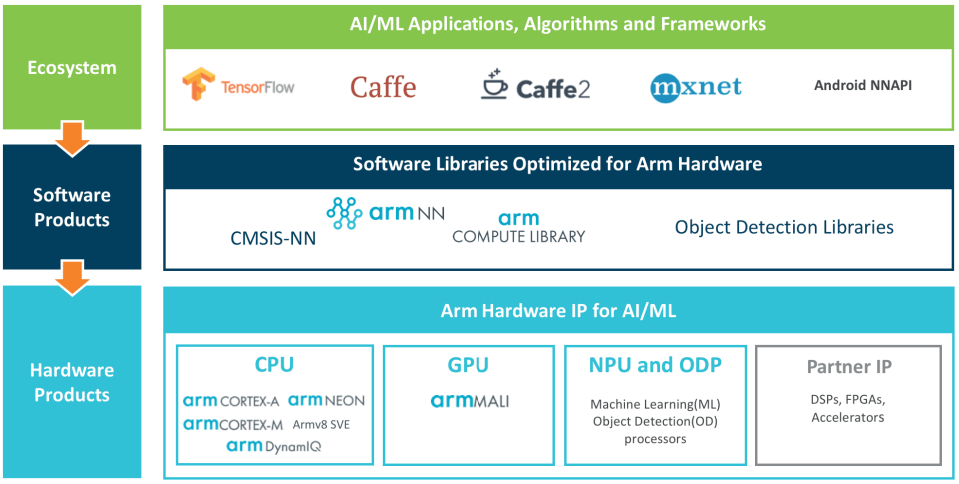
The Unstoppable March of ML to Edge Devices

Replicating the learning and decision-making functions of the human brain starts with algorithms that often require intensive compute power. However, a cloud-centric approach is not an optimal long-term solution. The power and cost required to shift massive amounts of data back and forth to the cloud can be prohibitive and produce a noticeable lag or delay in response – something that time-critical applications simply cannot tolerate, and users often find frustrating.

Today, advances in processing power and ML algorithms have pushed applications, training, and inference down from the cloud, with an increasing number of workloads now performed on devices at the edge. In addition to helping reduce costs and increase efficiency, this approach maximizes security as it limits the number of times sensitive data is shifted between cloud and device.

Project Trillium

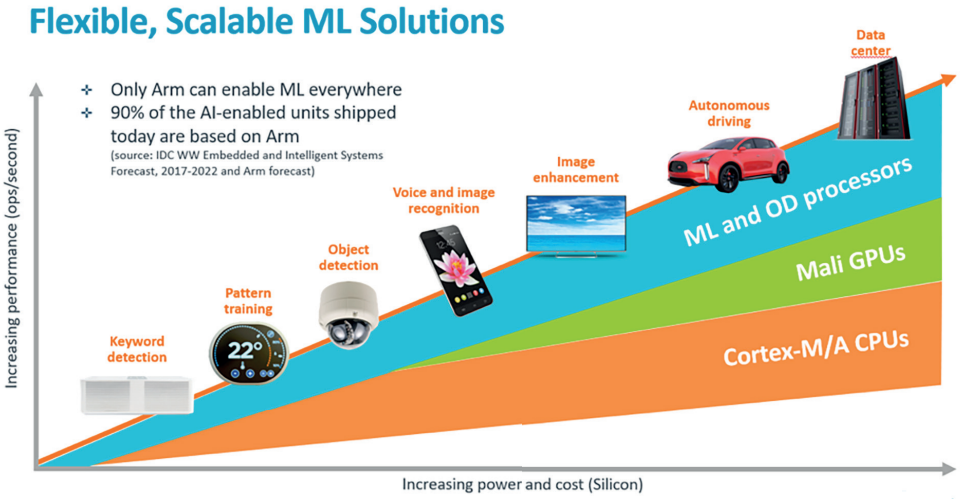
Designed for unmatched versatility and scalability, Project Trillium is Arm's heterogeneous ML platform. Project Trillium is advancing a new era of ultra-efficient ML inference at the edge, providing a range of performance options through a suite of products based on the world's most innovative and advanced technologies.



The power to transform

Project Trillium provides the power to transform computing across all sectors and devices. The only complete, heterogeneous compute platform for ML, it includes a new, highly scalable processor line that is compatible with all programmable Arm IP.

Future derivatives of this architecture will scale from as low as 2 GOPs for IoT and always-on devices to over 150 TOPs for server-type applications. This flexibility to address all applications is unique to Arm.



Arm Cortex CPUs and Mali GPUs

Cortex CPUs and Mali GPUs are already well established in high volume across a wide range of edge devices, and include a number of dedicated features to enhance ML performance:

- Cortex-A76 CPU delivers 4x the ML performance of the previous generation processor.
- Mali GPU offers an inherently large ML compute capability.
- Mali-G76 GPU provides a 3x increase in ML performance over the previous generation.

Arm Machine Learning processor

Specifically designed for inference at the edge, the ML processor is capable of an industry-leading performance of 4.6 TOPs, with a stunning efficiency of >3 TOPs/W for mobile devices and smart IP cameras.

Arm Object Detection (OD) processor

The Arm OD processor is the most efficient way to detect people and objects on mobile and embedded platforms. It continuously scans every frame to provide a list of detected objects, along with their location in the scene, and a number of other object characteristics.

Software Libraries

Project Trillium includes Arm NN, a software framework for the efficient translation of existing neural networks, to support ML workloads across all Arm programmable IP. The software also provides support for Arm Cortex-A CPUs, Arm Mali GPUs and the ML processor via the Compute Library; and for Cortex-M CPUs via CMSIS-NN.

Benefits of Project Trillium

- The only complete heterogeneous compute platform for ML
- Highly scalable, from 2 GOPs to over 150 TOPs
- Flexible support for ML workloads across all Arm programmable IP
- Forward-compatible with future Arm IP

Why Arm ML?

- Technology solutions from a trusted company
- Flexible, scalable and power-efficient solution to tackle a wide range of applications
- Support for a range of devices from ultra-constrained to servers
- Broad AI support from a diverse set of ecosystem partners

Today, the technologies within Project Trillium are optimized for the mobile and smart IP camera markets in response to current demands for edge ML performance. But as demands to deploy ML across a diverse range of mainstream markets increase, Arm's ML solution is flexible and scalable enough to meet almost any requirement or use case.

Whether your focus is increasing efficiency and performance or minimizing silicon cost, Project Trillium provides a solution for any ML workload.

To find out more visit: **www.arm.com/ai**

All brand names or product names are the property of their respective holders. Neither the whole nor any part of the information contained in, or the product described in, this document may be adapted or reproduced in any material form except with the prior written permission of the copyright holder. The product described in this document is subject to continuous developments and improvements. All particulars of the product and its use contained in this document are given in good faith. All warranties implied or expressed, including but not limited to implied warranties of satisfactory quality or fitness for purpose are excluded. This document is intended only to provide information to the reader about the product. To the extent permitted by local laws ARM shall not be liable for any loss or damage arising from the use of any information in this document or any error or omission in such information.

© ARM Ltd. 09.12.2018