Arm AI Platform Solutions Brief

CPU, GPU, NPU, Software and Tools

Solution Brief

AT A GLANCE

As the demand for powerful compute accelerates, Arm's heterogeneous architecture supports diverse AI applications wherever they run—from the mega compute in the data center, to power-efficient microcontrollers, or constrained battery-powered endpoint device such as wearable sensors. Arm delivers hardware, software and services to meet the demand.

Create Diverse AI Solutions from Cloud to Edge to Endpoint on the Arm AI Platform

Artificial intelligence (AI) represents the biggest transformation in computing in over a generation. It 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–from mobile devices, Internet of Things (IoT) endpoints to servers (edge).

Arm's AI Platform includes the Cortex-A and Cortex-M CPUs, Mali GPUs and the Ethos-N NPUs and Ethos-U microNPUs. It provides the power to transform computing across all sectors and devices. Arm's AI Platform is only complete, heterogeneous compute platform for AI that is compatible with all programmable Arm IP.

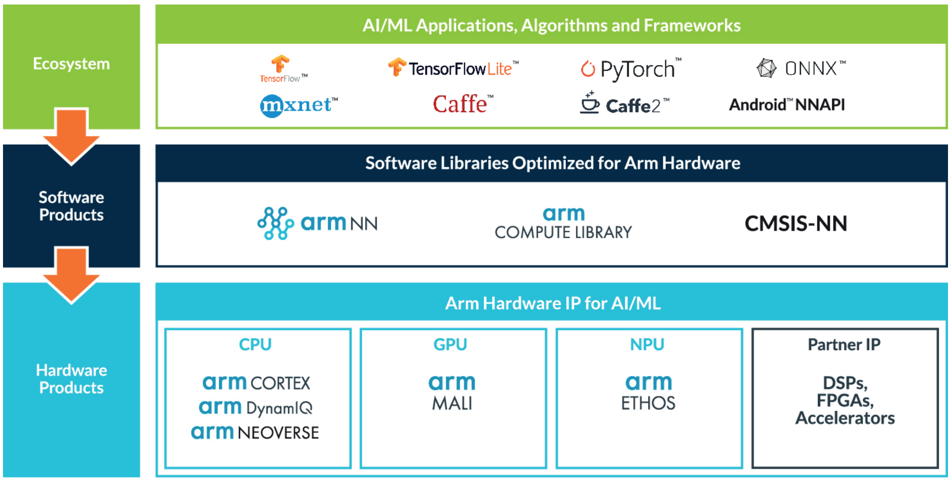
Coupled with one of the world's largest AI partner ecosystems, Arm provides the essential best-in-class tools, algorithms and applications needed for the deployment of next-generation intelligent edge solutions and endpoint devices.

Futureproof Operator Coverage

The MAC engine flexibly decomposes arbitrarily sized kernels with stride and dilation modes including convolution, deconvolution, depthwise separable, and vector product. Programable Layer Engines execute layers not supported by the MAC engine, supporting various primitives, activation functions and future operators.

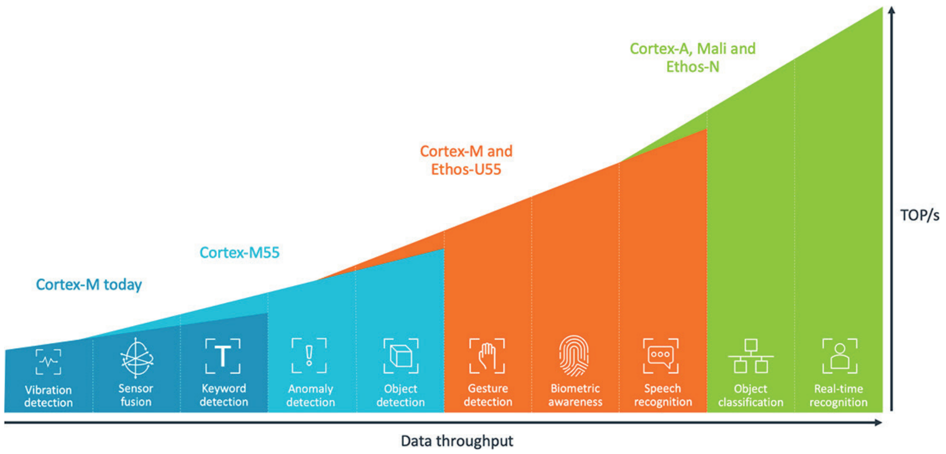
WHY THE ARM AI PLATFORM?

- The only complete heterogeneous compute platform for AI
- Highly scalable, from 2 GOPs to over 640 TOP/s
- Flexible support for AI workloads across programable IP
- Forward compatible with future IP
- One of the world's largest AI ecosystems



Flexible, Scalable AI Solutions

Only Arm has the broadest range of ML-optimized processing solutions



DEFINITIONS

Cloud data center computing is the relatively centralized availability of computer system resources, especially data storage and computing power, usually located in large internet accessible data centers located great distances from the user.

Edge computing starts where endpoint devices plug into the network and ending at cloud data centers. In many ways, the edge becomes an offload layer from the cloud. It's closer to the endpoint devices (so latency remains lower) but contains powerful enough compute to handle most workloads.

Endpoints are where data is harvested and primary compute frequently happens. The right type of compute is required on the endpoint to extract information locally from the harvested data while satisfying the operational requirements of a device.

Arm Cortex CPUs

Arm Cortex CPUs are well established across a wide range of devices. The processors address the performance, power and cost requirements for many applications from the cloud to the edge and endpoint devices. They have the flexibility to run any ML-based workload or task and come with many features that enhance ML performance. Cortex CPUs also enable ML use cases on devices versus sending workloads to the cloud, which provide greater privacy and security to users.

Arm Mali GPUs

Arm Mali GPUs with ML extensions and high floating-point throughput provide high performance across a variety of immersive entertainment experiences. This includes visual enhancements to gaming, images and video, as well as enabling image and video recognition features like depth sensing. Mali GPUs also enable longer battery life on mobile devices, providing longer untethered entertainment experiences for users.

Arm Ethos NPUs

Arm Ethos NPUs enables you to target multiple market segments with the same software framework using Ethos-U and Ethos-N processors. The NPUs provide the highest throughput and efficiency in the lowest area for ML inference. Ethos' unmatched scalability from 10s of TOP/s to over 640 TOPs enables new AI solutions in the cloud, edge or endpoint.

Opensource Software and Tools

Arm AI Platform supports all popular frameworks and operators and contributes up-stream optimizations to popular industry frameworks. Arm software allows seamless integration with existing NN frameworks, such as TensorFlow, Caffe, and Android NN for efficient AI deployment. Arm IP are easy to debug and profile using the fully integrated Arm Development Studio 5 Streamline showing side by side performance analysis of CPU, GPU, NPU.

Learn More About the Arm AI Platform Solutions

Whether your focus is increasing efficiency and performance or minimizing silicon cost, the Arm AI platform provides a solution for any AI 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.