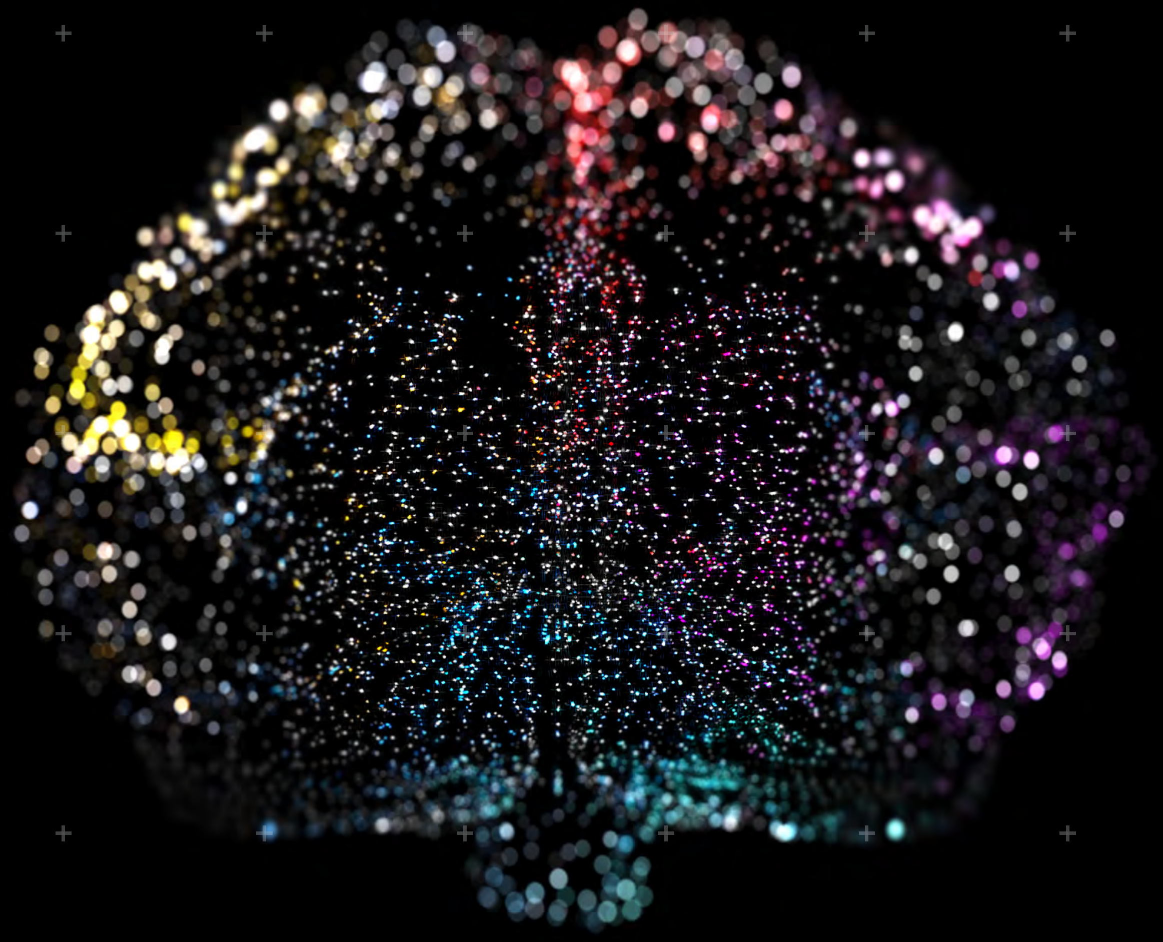
White Paper

What's Powering Artificial Intelligence?

To scale artificial intelligence (AI) and machine learning (ML), hardware and software developers must enable AI/ML performance across a vast array of devices. This requires balancing the need for functionality alongside security, affordability, complexity and general compute needs. Fortunately, there's a solution hiding in plain sight.

By Rene Haas, President, IPG Group,
and Jem Davies, GM of Machine Learning, Arm.



AI compute is increasingly moving out of the traditional cloud and closer to where the data is being gathered. This shift is driving huge innovation in AI-enabling hardware and software as designers stretch the boundaries of what edge AI is capable of.

Introduction

Artificial intelligence (AI) and machine learning (ML) are now rapidly gaining commercial traction. In the past six years, the amount of compute used in the largest AI training models has, on average, doubled every hundred days – contributing to a 300,000-fold increase in AI computing, according to a GigaOm report, 'AI at the Edge: A GigaOm Research Byte'.

These and other AI headlines have prompted feverish development in the engineering and developer community, as individuals and companies compete to deliver life- and business-changing AI products and services. However, amid this growth frenzy, design teams face a multitude of possible design approaches, mainly around the choice or combination of processor technology to run their AI workloads on. This white paper's objective is to unravel some of those technical knots and offer guidance on how best to approach designing AI compute that prioritizes practicality, performance and cost.

Evolving landscape

To date, much of the ML focus has been on the Cloud, on huge centralized compute farms. However, ML compute is increasingly moving out of the traditional cloud and closer to where the data is being gathered. This is for reasons including efficiency, speed, privacy and security. This approach is being accelerated by the emergence of new connected devices in areas such as advanced and autonomous cars, healthcare, smart cities and the Industrial Internet of Things (IIoT). As a recent PCmag article put it:
'When the Cloud Is Swamped, It's Edge Computing, AI to the Rescue'.

This drive to democratize AI compute is evidenced by the move to create more intelligence in the world's data networks, notably with increasing numbers of edge servers being shipped. Indeed, AI compute is a key focus of Arm's Neoverse family of technologies now adopted by the likes of Amazon Web Services. According to Santhosh Rao, a principal research analyst in this area at Gartner: "Currently, around 10% of enterprise-generated data is created and processed outside a traditional centralized data center or cloud. By 2022, Gartner predicts this figure will reach 50%."

AI at the Edge: A GigaOm Research Byte
https://pages.arm.com/ai-edge.html

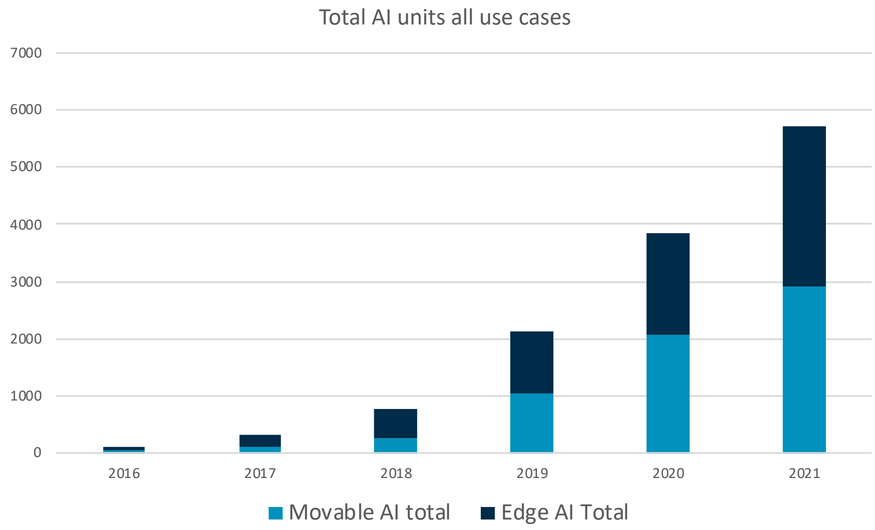
When the Cloud Is Swamped, It's Edge Computing, AI to the Rescue
https://uk.pcmag.com/features/94303/when-the-cloud-is-swamped-its-edge-computing-ai-to-the-rescue

Arm Neoverse
https://www.arm.com/company/news/2019/02/our-next-step

"Where should you perform the calculations to do inference? The short answer is that for many – if not most – applications, inference in the future will be done at the (device) edge, that is, where the data is collected. This will have a huge impact on how ML will develop."

– The GigaOm study 'AI at the Edge'

It's too early to put reliable figures on how much AI compute is being done inside edge servers but there are robust figures for edge devices. Data we have seen from IDC suggests 324 million edge devices used some form of AI (inference and training) in 2017. Two thirds of those were edge-compute systems, such as Internet of Things (IoT) applications, and one third were mobile devices.



Outside the cloud, we see that the vast majority of AI compute growth now is in inference, comparing real-world data to a trained AI model. By 2020, IDC forecasts that the combination of edge and mobile devices doing inferencing will total 3.7 billion units, with a further 116 million units doing training.

The GigaOm study sums up this evolution succinctly:
"Where should you perform the calculations to do inference? The short answer is that for many – if not most – applications, inference in the future will be done at the (device) edge, that is, where the data is collected. This will have a huge impact on how ML will develop."

Technological drivers

There are a number of trends pushing AI and ML to the edge and driving the technology to the scale and transformative impact that many anticipate.

First, there is the change in the nature of data. The explosion of connected devices at the edge – particularly in IoT applications – is creating zettabytes of new data in the AI-driven world. There isn't enough bandwidth in the world to support sending all that data to server farms for processing. This means that edge-based AI won't be a choice but a necessity. For example, Google has said that if everyone in the world used Google's voice assistant for three minutes per day, the company would have to double the amount of compute servers they have.

Then, there's reliability. The latency that the traditional cloud computing model introduces into systems will not work for many edge applications, such as autonomous vehicles. It's inefficient to send vehicle data to the cloud and back, so the imperative is to do as much onboard processing as possible.

“Of the 350 respondents to an Arm survey, 42 percent said they are using a CPU for the bulk of their AI computation.”

Related to this is power. It takes power to send data to the cloud and back and significant power inside compute farms to compute the data.

An increasingly influential trend is privacy. Consumers are more sensitive today to their personal data being sent to the cloud.

Connectivity plays a role as well. While billions of devices, mainly smartphones, are capable of running AI apps, most cannot do that without connectivity – potentially limiting their use.

Unlocking affordable AI at the edge

Amid these technological AI forces, there are further factors. Specifically, engineers and developers must deal with conflicting information that could make the difference between the commercial success or failure of an AI project. That is, which processor technology is most appropriate for these AI workloads? The answer is the workhorse that's already the default processor for AI computing: the CPU. The CPU is central to all AI systems, whether it's handling the AI entirely or partnering with a co-processor, such as a GPU or an NPU for certain tasks.

As such, the CPU will remain the workhorse for ML workloads because it benefits from common software and programming frameworks that have been built up over many years during the mobile compute revolution. Additionally, it plays a vital role as mission control in more complex systems that leverage various accelerators via common and open software interfaces.

Indeed, Facebook researchers reported that when it comes to Facebook apps most inference today is already done at the edge on mobile-device CPUs – most of it being computed on processors introduced years ago.

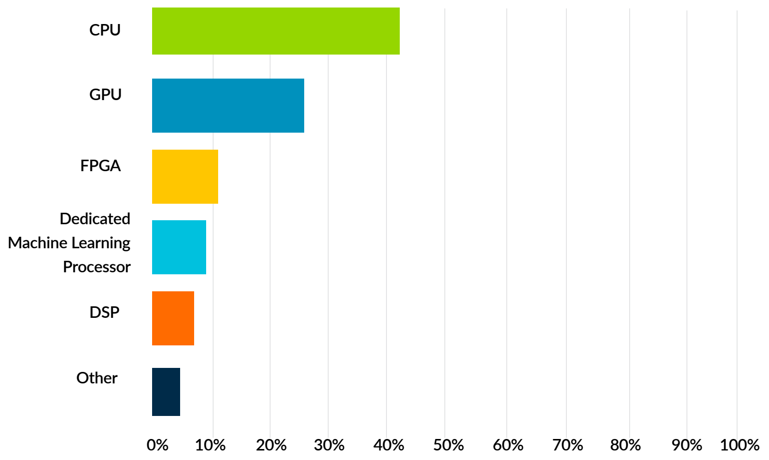
“System diversity makes porting code to co-processors, such as DSPs, challenging. We find it more effective to provide general, algorithmic-level optimizations that can target all processing environments,” the authors wrote.

To test that thinking, Arm commissioned a research survey among chip and AI product designers in the global Arm ecosystem. They were drawn from all sectors now using AI-enabled technologies, including the IoT (54 percent of respondents), industrial (27 percent), automotive (25 percent) and mobile computing (16 percent). The survey results paint a clear picture.

Machine Learning at Facebook: Understanding Inference at the Edge
https://research.fb.com/publications/machine-learning-at-facebook-understanding-inference-at-the-edge/

More than 40 percent of the nearly 350 respondents said they are using a CPU for the bulk of their AI computation. A quarter are using GPUs and the remainder are using FPGAs (12 percent), dedicated machine learning processors (8.6 percent) or DSPs (7.5 percent).

Q5 Thinking about current products or designs projects, what hardware does the bulk of computation of AI/ML inside your design or for your app?



The Arm AI processor workload study was fielded in April 2019 with nearly 350 responses received from the semiconductor and broad technology product sectors.

The smartphone as a transformative force for ML

The CPU has grown into the go-to solution for AI and ML because developers and engineers have adapted lessons from the rise of the mobile compute ecosystem. Mobile scaled so quickly and massively precisely because it had its CPU anchor point and could handle any compute workload. Then, firmware and software built around the foundational CPU helped to give rise to millions of products and apps.

This legacy has propelled the smartphone to become the first-mover device for running AI and ML, especially for inference. It is a device that combines efficiency, portability, cost, power consumption and cybersecurity. These are all critical factors in putting AI compute into the maximum number of hands.

CPU: Mission control for AI

A quick guide to the difference between CPU, GPU and NPU compute for AI:

CPUs: The main advantage of CPUs is they already sit at the center of the system, and they are the only processors flexible enough to run any type of ML workload, today or tomorrow. In addition, they scale easily and can support any programming frameworks or languages including C/C++, Scala, Java, Python or many others. They are often used in the cloud for AI inference (comparing models to real world data) and are the only ubiquitous solution for AI on edge devices. CPUs are perfect for running complex workloads and are often used as first-choice AI processors even in AI bespoke designs across the IoT and mobile computing arenas.

GPUs: Are specialized processors that are more complex compared to CPUs and offer more focused functionality. They are typically designed only to support graphics or small amounts of highly parallel, fixed workloads, such as processing image pixels or videos. GPUs are programmed in relatively few languages so provide limited flexibility compared to CPUs. They are used today in AI applications such as prototype autonomous cars and training in servers.

NPUs: The need for more specialized and hyper-efficient AI compute has led to the design of neural processors (NPUs) that are highly task-specific. NPUs are designed from the bottom up for faster execution of AI applications and are lightning quick at performing the dense vector and matrix computations typical with machine learning workloads but are less flexible than either CPUs or GPUs. They must be used in conjunction with a CPU to handle general compute functions.

Another reason for running AI workloads on the CPU lies in processor history: CPUs have always executed and managed system and application software. In an age of increasing system complexity, this legacy and familiarity is a powerful lure for designers who are always concerned over time-to-market.

Today, CPUs are used for optimized ML and ML-related tasks such as:

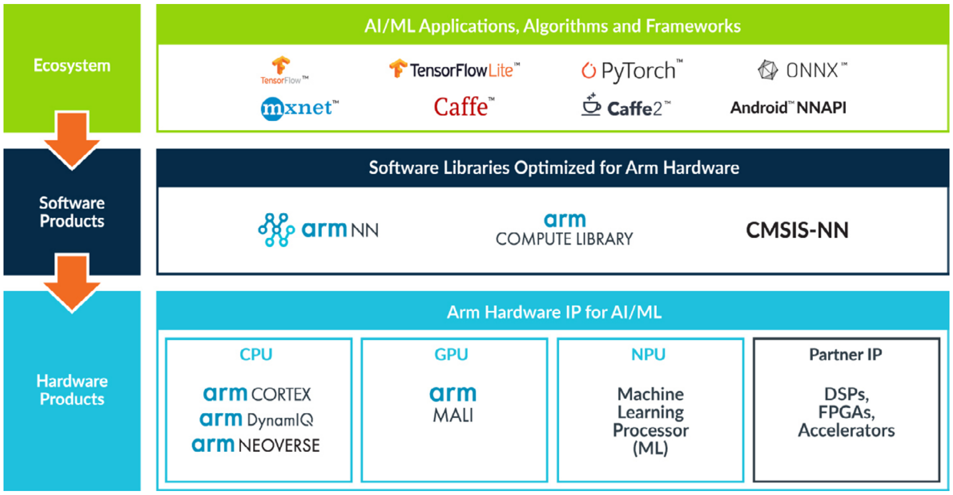
- Simple natural language processing (NLP): Robots and home devices can use NLP to recognize requests, particularly complex or compound statements, when network connectivity is spotty or unavailable.
- Face recognition/identification: The system can manage access to shared areas or record participants of a meeting, while maintaining privacy by keeping data local.
- Simultaneous location and mapping (SLAM) functions for autonomous movement: In systems with constrained or slow movement (not fast-moving cars or drones), a CPU is often sufficient for running SLAM functions.

Implications for software designers

Whatever variety of AI system a designer is looking to create, he or she should always think of the software developer who wants to write for as many products as possible. Solutions have emerged to enable that. For example, Arm NN makes it much easier to build and run ML applications on power-efficient, Arm-based platforms.

Arm NN
https://www.arm.com/products/silicon-ip-cpu/machine-learning/arm-nn

The open-source Arm NN software and Arm Compute library (see graphic below) provide a bridge between existing neural network frameworks – such as TensorFlow or Caffe – and the underlying processing hardware, such as CPUs, GPUs or any Arm machine learning processor.



Ecosystem

AI/ML Applications, Algorithms and Frameworks

TensorFlow™ | TensorFlow Lite™ | PyTorch™ | ONNX™
mxnet™ | Caffe™ | Caffe2™ | Android™ NNAPI

Software Products

Software Libraries Optimized for Arm Hardware

arm NN | arm COMPUTE LIBRARY | CMSIS-NN

Hardware Products

Arm Hardware IP for AI/ML

CPU: arm CORTEX, arm DynamIQ, arm NEOVERSE

GPU: arm MALI

NPU: Machine Learning Processor (ML)

Partner IP: DSPs, FPGAs, Accelerators

While the CPU is the first choice for ML workloads today, it will remain an anchor technology in the world of heterogenous, accelerator-enhanced systems that are coming. That's due in part to the nature of ML workloads. We typically see ML as a computing 'burst', with processing set up by the CPU and then done on the CPU or offloaded onto an NPU or GPU.

How to tackle 'Tiny ML'

The growing arena of so-called Tiny ML is a major expansion area mainly because of the big drive in the often cost- and power-constrained world of IoT. Thanks to the relentless capability improvement in MCUs, we now see plenty of ML and ML-related tasks such as:

- **Power management and scheduling:** The MCU can adjust operating points, reassign tasks to different cores and balance throughput against efficiency.
- **Security auditing:** The system can monitor its own behavior by observing normal patterns and quickly identifying abnormalities.
- **Object detection:** The system can enable low-power modes when undertaking complex actions, such as unlocking a device or taking a camera image, without draining the battery.
- **Keyword spotting:** The system can respond to keywords or phrases as part of normal operation.

ML at your fingertips today

The AI technical capabilities of CPUs and MCUs are already influencing design choices in smartphones and edge devices. ML workloads are commonly used for face and fingerprint recognition and even medical diagnoses. This has spawned new classes of products and features, such as intelligent asthma inhalers, facial recognition and facial-unlock capabilities, wake- and voice-sensing, and more:

Flexible edge ML on MCUs

Studies have shown that despite memory and compute constraints, many designs have deployed complex neural networks on MCUs. Techniques such as pruning and quantization help to compress models for deployment on MCUs, as well as being put to the same use for CPUs, NPUs and GPUs.

The neural network must also be updated securely over time, often remotely. A team of Arm engineers in San Jose, Austin and Taiwan demonstrated scalable and secure remote updates from the cloud to neutral nets deployed on MCUs. The system used Google TensorFlow, an Mbed OS-enabled Arm Cortex-M4 running at 100MHz and the Arm Pelion Device management platform for updating.

Breathing easier

The Amiko Respiro provides smart inhaler technology that helps asthma sufferers breathe more easily. At the center of Respiro's sensor module is an ultra-low-power Arm Cortex-M processor. The solution uses ML to interpret vibration data from the inhaler. The sensor is trained to recognize the patient's breathing pattern and inhalation time. The processor allows the Respiro device to run real-time ML algorithms that recognize behavior patterns and interpret data within the sensor module itself. When the user presses the trigger, the module instantly recognizes the breath data pattern and provides fast, private user feedback.

Facial recognition

An Arm team developed a proof-of-concept to run facial verification. A low-resolution image processed on an Arm Cortex-M microcontroller activates the system, which triggers the high-resolution camera. Face detection and anti-spoof operations run on an Arm Cortex-A CPU using low-level and optimized software functions supported by the Arm Compute Library.

ADAS and vision systems

At CES 2019, Israel-based Brodmann17 demonstrated its solution for advanced driver assistance systems (ADAS): A deep-learning vision software solution that works at a fraction of current computing power needs yet delivers the highest accuracy levels recorded for ADAS and autonomous driving vehicles. The ADAS system runs on an Arm Cortex-A72, using only a single core while maintaining high performance and accuracy.

Amiko Respiro case study
http://bit.ly/AR-case-study

Arm Compute Library
https://www.arm.com/why-arm/technologies/compute-library

> “The CPU and MCU will remain preeminent, even as innovative companies continue to develop specialized solutions that will optimize AI projects for cloud-based or edge-based systems.”

An ecosystem for ML development

These and other innovative CPU- and MCU-powered AI edge deployments are enabled by a thriving design and development ecosystem that has extended its reach from smartphone design and app development to ML. Consider Arm’s Project Trillium approach. Project Trillium provides flexible support for ML workloads across all programmable Arm IP as well as partner IP. It allows seamless integration with existing neural network frameworks, such as TensorFlow, Caffe, and Android NN and is supported by a vibrant and diverse ecosystem, driving innovation and choice.

The way forward

The characteristics that have driven the ubiquity of the mobile CPU and the relentless innovation around it will continue to bolster the CPU’s central role for ML applications. That is especially true when we consider microarchitectural enhancements, which bring a fundamental shift in the AI capability of future Arm Cortex-A applications processors and Cortex-M devices, including:

- New dedicated processor instructions in Cortex-A and Mali graphics processors designed to boost AI performance.
- Continuous innovation and evolution of the Arm architecture and microarchitecture, which includes supporting newer data types and extended and faster vector processing supporting both PC/cloud and edge compute.
- Helium, introduced with the Armv8.1-M architecture, which is a new optimized vector architecture for the Cortex-M architecture that delivers up to 15x performance improvement to AI performance.

AI and ML design will get further help as Arm continues to evolve its GPU and NPU technologies. These hardware innovations, together with Arm’s software solutions, offer a powerful choice for developers who build transformative AI and ML products and services.

The edge is where the next phase of AI technology evolution is happening. Economics and sheer bandwidth constraints mean it doesn’t make sense to send data to the cloud to get AI-enabled decisions.

Development will track the mobile computing evolution, leveraging standardized approaches, software and tools frameworks, and ecosystems that will deliver reliable, powerful and efficient systems quickly.

The CPU and MCU will remain preeminent, even as innovative companies continue to develop specialized solutions that will optimize AI/ML projects for cloud-based or edge-based systems. Increasingly heterogenous systems will rely even more on the CPU for mission control, while enabling the addition of processors, such as the NPU, to take on task-specific workloads.

Arm Project Trillium
https://www.arm.com/products/silicon-ip-cpu/machine-learning/project-trillium

With CPUs as the primary AI compute programming target, developers are able to focus code on the one processor that will feature in all ML-based systems.

But while the CPU is a natural and powerful choice for AI mission control, we already see the emergence of new hybrid processing environments where the CPU takes advantage of co-processors such as NPUs, GPUs, VPUs and DPUs for certain specific workload tasks.

In fact, nearly a third of the respondents to the Arm AI survey indicated that for future projects they want to leverage dedicated ML devices, such as NPUs, as their primary ML engine. Arm continues to invest heavily in developing NPUs and GPUs to meet the diverse needs of the AI development community.

These CPU-centric or hybrid AI environments will raise the bar on what we can do with data and lead to far greater compute possibilities, actions and insights than is possible today. But as the value of what AI can do for data rises so does the cybersecurity threat, so utilizing the decades of honed security capabilities of CPUs will be vitally important.

We hope this paper has set out a sensible consideration path for designers looking to build their AI compute-based systems, as well as given some insight into what today's AI developers are already doing.

Our simple message is that CPUs are ever more AI-capable, so focus your system on the CPU for AI mission control. It will give you the system performance, security and cost efficiency any mass market product requires. At the same time, consider the balance of needs and look at hybrid systems where that makes sense.