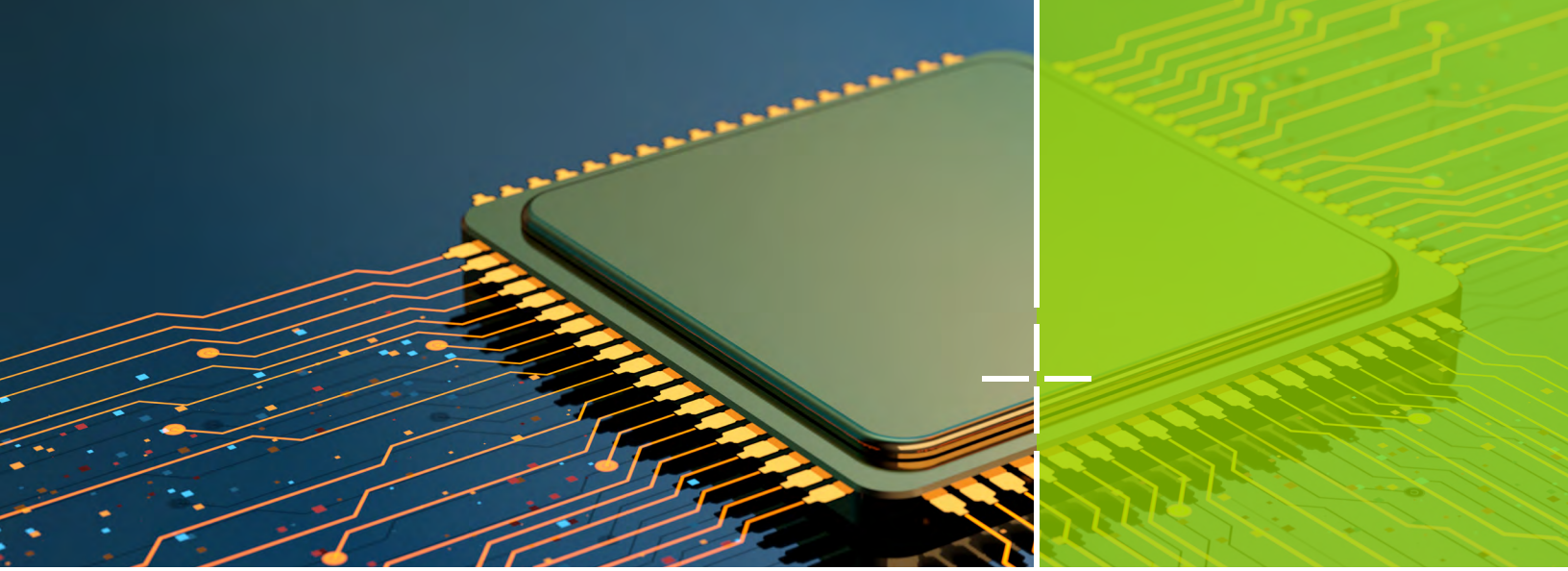


SOLUTION BRIEF

Faster, Smarter, More Efficient AI in the Real World

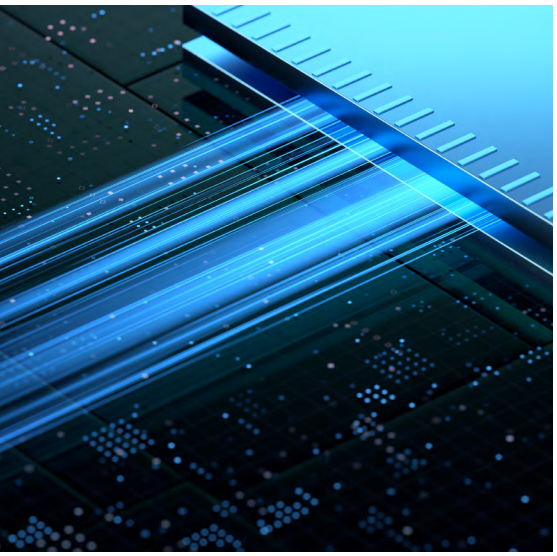
UNTETHER AI

ARM IP

+ Cortex-A

OVERVIEW & GOAL

Untether AI was founded to radically rethink how computation for machine learning is accomplished. In current architectures, 90 percent of the energy for AI workloads is consumed by data movement, transferring the weights and activations between external memory, on-chip caches, and finally to the computing element itself. By focusing on the needs for energy-centric AI compute, Untether AI provides highly performant and energy-efficient hardware, like the runAI® and speedAI® family of devices, tsunAImi® family of PCIe cards, and enabling imAIgine® software that supports a wide variety of AI workloads. Untether AI's hardware accelerators work in conjunction with host processors based on Arm Cortex-A. The power and energy efficiency of Arm Cortex-A CPUs, coupled with Untether AI at-memory compute provides an industry-leading combination for energy efficient AI inference across a wide variety of workloads.



APPLICATION AREA

- Artificial Intelligence
- Autonomous Vehicles
- Smart Cities
- Vision

CHALLENGE

Early approaches to neural networks and learning have revolved around massively parallel GPUs or expensive custom neural processors crunching vast amounts of data to learn and act as quickly as possible. The biggest bottleneck in scaling these technologies lies in memory. The traditional von Neumann architecture is not very well suited for the compute requirements of neural net inference. In current architectures, 90 percent of the energy for AI workloads is consumed by data movement. This means the energy used inside a chip has transitioned from being dominated by the transistors doing the computation to the wires that get the data to them. To ensure system-wide gains in energy efficiency, it's important that not only the accelerator but the host processor is efficient, whether it be autonomous applications where battery life is paramount, or datacenters where for every 1 W of power consumed, an additional 1 W is used for cooling.

SOLUTION & BENEFITS

Compute workloads in the automobile have changed radically in the past several years driven by the industry shift to centralized compute architectures, the adoption of ADAS (advanced driver-assistance system) sensor networks with a greater number of sensors with higher resolution, and the move toward higher levels of ADAS. Additionally, the introduction of newer neural networks that offer greater accuracy and lower latency also adds to the rising workload performance and complexity.

This trend is giving rise to the need for heterogeneous computing employing domain-specific architectures that consist of AI accelerators, Arm-based CPUs, memory, and networking technologies.

Through close collaboration with Arm in the joint definition and development of solutions that are optimized for these complex, evolving workloads in the car, we deliver proven, best-in-class solutions that achieve the highest level of performance and accuracy with unmatched energy efficiency.

ARM IP IN USE:

We provide a runtime driver for 64-bit Arm processors that allows us to act as an accelerator for an Arm host.