Electronic Design

ALAN R. EARLS, Contributing Editor

As AI Takes Off, Chipmakers Pump Up Performance

Microcontrollers and other small systems are the latest to embrace AI, and hardware vendors are leading the charge with fresh innovations.

It's hard to miss the fact that AI is trending strongly across the industry. Demand for it is on the rise everywhere, including the edge, leading to upgrades of existing semiconductor products. Suppliers are targeting new data types and new instruction sets as well as introducing new products.

Each of these initiatives aims to squeeze more computing, optimized for the needs of AI, from small devices that generally must use the bare minimum of power served up at a rock bottom price point. A big challenge is that inferencing used in more advanced AI depends greatly on large language models (LLMs) that require enormous computing resources for training. It also, for the most part, forces edge devices to rely on high-bandwidth communication with more powerful resources in the cloud or a data center, which can be inconvenient and power-intensive.

This drives some of the developments in AI accelerators, which can appear as co-processors in consumer devices and smartphones and as an element of some system-on-chip (SoC) products. Because efficiency is so critical when attempting the difficult task of advanced AI, implementations are often very application- or market-specific and must consider what kind of processing (e.g., floating-point inference vs. fixed point) is being used and the kind of network to which the device will connect.

On the other hand, there's still the need to have adaptable, general-purpose systems that can potentially survive in the field for a decade or more, and are ready, in some cases, to significantly shift functionality through software updates. Vendors are responding with a spectrum of new or updated processors, accelerators, and more.**Progress Reports on AI-Enhanced Products**

Offerings from some of the growing number of product suppliers with new AI-oriented capabilities or enhancements are summarized below.

AMD

Edge AI is the focus of AMD with its Versal AI Edge SoC. The Versal ACAP (Adaptive Compute Acceleration Platform) architecture scales from sub-10-W edge devices to >100-W supercomputer systems.

According to the company, the chips consist of three main parts: scalar engines that include two dual-core Arm processors for running Linux-class applications plus safety-critical code; adaptable engines that can handle determinism and parallelism to process data from sensors; and intelligence engines that can run edge AI workloads such as AI inference, image processing and motion control.

AMD said it offers competitive performance per watt compared to GPUs for roles like real-time systems in autonomous vehicles, healthcare systems, factories, and aircraft. For developers, its Vitis unified software platform provides open-source libraries, a model zoo, a single programming model for developing applications on all of AMD's chip architectures, and a video-analytics software development kit.

Arm

Power-efficient 32-bit CPU cores from Arm, based on the company's Helium technology, target IoT-type markets. The Cortex-M52 is intended to deliver machine-learning (ML) capabilities for small, battery-powered IoT devices.

Arm claims that the Cortex-M52 CPU core can provide 5.6X the performance of its predecessors for AI inference tasks, while avoiding the need for a separate neural process-

The self-driving car is a prime use case for more advanced AI and the semiconductors that can support it.

ing unit (NPU). It also incorporates the company's TrustZone technology to ensure security.

According to Arm, Helium is an optional extension in the Armv8.1-M architecture that improves ML and digital-signal-processing performance.

Expedera

Expedera makes an NPU that's pitched for SoC designs. It includes a unified compute pipeline designed to reduce memory bottlenecks and meet performance needs of AI applications.

Origin is a neural engine IP line of products that reduces memory requirements and overhead to deliver performance and power efficiency. In particular, Origin's hardware takes on software burdens, yielding a simplified software stack that allows TensorFlow to execute directly in hardware. Expedera claims a sustained single-core performance of up to 128 TOPS with typical utilization rates of 70% to 90%.

MIPS

The main target of MIPS's eVocore RISC-V CPUs is automotive systems. Within eVocore CPUs, multithreading enhances performance efficiency, and a coherent fabric optimizes data movement to different compute engines. The company said that multi-cluster support provides a high level of system scalability.

NeoLogic

According to NeoLogic, its new post-CMOS-VLSI technology—the patent-pending Quasi-CMOS—can slash the transistor count of cores by up to 3X, cutting power dissipation in half and sharply reducing area requirements. The upshot of all of this, said the company, is a significant improvement in performance per watt. It targets the needs of video, AI/ML, and data-analytics applications both in the cloud and at the edge.

Quadric

The processor architecture of the Chimera general-purpose neural processing unit (GPNPU) has been designed to deliver artificial-intelligence computing "on device," said Quadric. The Chimera can provide strong ML inference performance while also running traditional C++ code.

There's no need for a partition code between multiple kinds of processors. The GPNPU uses a single pipeline to handle matrix and vector operations and scalar (control) code.

According to the company, the GPNPU is a licensable processor that scales from 1 to 16 TOPS and can run all types of ML networks, including backbones, vision transformers, and LLMs. The attributes of the GPNPU, said Quadric, result in faster porting for ML models and thus a faster time-to-market.

Quadric also offers DevStudio, which provides simulation of AI software and visualization of SoC design choices.

Synthara AG

The ComputeRAM from Synthara is primarily geared toward ASIC makers to support in-memory computing in existing technology. The company claims its technology can increase performance by up to 50X without altering architecture. It can also be incorporated in many types of processors, including x86, Arm, and RISC-V. Moreover, the software supports both existing and new algorithms.

In addition to the above vendors, a number of startups are entering the field, including Red Semiconductor in the U.K. and U.S.-based AI-chip startups such as Graphcore, iDEAL Semiconductor, and Kneron. Many more will likely enter the fray as this sector continues to attract attention and investment.

References

"AI Accelerator Architectures Poised For Big Changes," Semiconductor Engineering.

"Processor Tradeoffs for AI Workloads," Semiconductor Engineering.