RESEARCH NOTE

ARM ETHOS-U65 NPU

STEVE MCDOWELL, CHIEF ANALYST
APRIL 10, 2024

CONTEXT

Arm introduced its new Ethos-U65 microNPU (Neural Processing Unit). This state-of-the-art AI accelerator facilitates machine learning (ML) inference in many embedded systems and high-performance devices.

BACKGROUND: WHAT IS AN NPU?

An NPU, or Neural Processing Unit, is a specialized hardware accelerator designed to accelerate AI/ML computations. NPUs efficiently handle the types of operations common in AI workloads, such as tensor operations, matrix multiplications, and neural network inferences, all fundamental to tasks like image and speech recognition, natural language processing, and autonomous driving, among others.

Key Characteristics of NPUs include:

- **Optimized for AI Workloads:** NPUs perform AI-related computations more efficiently than general-purpose CPUs, focusing on accelerating deep learning algorithms.
- **Parallel Processing Capabilities:** These accelerators often feature many parallel processing capabilities, allowing them to handle multiple computations simultaneously.
- **Energy Efficiency:** NPUs maximize performance per watt, providing the computational power needed for AI tasks while minimizing energy consumption.
- **Low Latency:** By processing AI workloads directly on the device (on-device AI), NPUs can reduce the latency in sending data back and forth to the cloud or core data centers for further processing.

NPUs can be standalone processors, part of a SOC alongside other components like CPUs and GPUs, or integrated within CPUs or GPUs as dedicated AI acceleration cores.

NEW: ETHOS-U65 NPU

The Arm Ethos-U65 extends the power and efficiency benefits to more powerful Arm Cortex-A, Cortex-R, and Arm Neoverse-based systems, targeting a broader range of applications, from smart cameras to infrastructure management subsystems.

Here's a closer look at the Ethos-U65 and its key features:

- **Performance:** The Ethos-U65 delivers 1 TOPS (tera-operations per second), providing a significant performance boost for on-device ML inference.
- **Efficiency:** Despite its enhanced performance, the Ethos-U65 maintains the power efficiency hallmark of the Ethos-U55, ensuring that devices can handle more complex ML tasks without compromising on battery life or thermal constraints.
- **Versatility:** The Ethos-U65 is compatible with a broader range of Arm processors, including the more powerful Cortex-A and Neoverse architectures.
- **MicroNPU Architecture:** The Ethos-U65 builds upon the microNPU architecture introduced with the Ethos-U55, which is optimized for executing neural networks in extremely low area and with low power consumption.
- **Dual AXI Interface:** It features a dual AXI interface, enhancing the bandwidth for weight-bound networks and ensuring efficient data flow within the system.
- **SRAM and DRAM Support:** The Ethos-U65 accommodates the needs of more complex systems by supporting both SRAM (128-bit AXI) and DRAM (128-bit AXI). It offers an average increase in network performance of 150% over the Ethos-U55.
- **MACs/Cycle:** The Ethos-U65 provides two different configurations, with 256 and 512 MACs (multiply-accumulate operations) per cycle, allowing a balance between performance and area for their specific application.
- **Memory System Support:** It is designed for use with DRAM-based systems, leading to higher bandwidth availability, and can be used in both high-performance Cortex-A/Neoverse-based SoCs and low-power Cortex-M-based SoCs.
- **Unified Software and Tools:** The Ethos-U65 leverages the same software and tools as the Ethos-U55, offering developers a consistent and familiar development environment. This simplifies the integration of AI capabilities into a wide range of products.

ANALYSIS

By extending the power efficiency and performance of its Ethos line from the Cortex-M to the more powerful Cortex-A and Neoverse architectures, Arm is democratizing AI, making it accessible to a broader range of applications.

The Ethos-U65's delivery of 1 TOPS (tera-operations per second) while maintaining energy efficiency addresses two critical constraints in AI deployment: computational power and power consumption. This balance is crucial for enabling advanced AI functionalities in devices where battery life and thermal management are important considerations.

The new Ethos-U65 microNPU is Arm delivering the IP necessary for the burgeoning demand for on-device AI. Arm positions itself to drive the next wave of AI innovation by focusing on power efficiency, performance, and broad applicability.

The information contained within this document was believed by NAND Research to be reliable and is provided for informational purposes only. The content may contain technical inaccuracies, omissions, or typographical errors. This document reflects the opinions of NAND Research, which is subject to change. NAND Research does not warranty or otherwise guarantee the accuracy of the information contained within.

NAND Research is a technology-focused industry analyst firm providing research, customer content, market and competitive intelligence, and custom deliverables to technology vendors, investors, and end-customer IT organizations.

Contact NAND Research via email at info@nand-research.com or visit our website at nand-research.com.