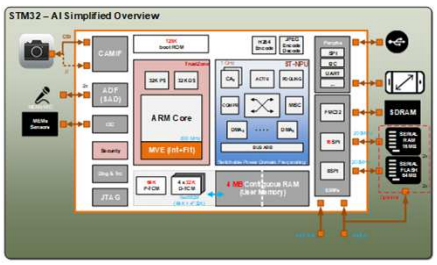
ANDANTE

SoC 1.1 — STM32 with AI

Overview

- The objective for STM is to add the hardware support of Artificial Intelligence to the STM32 family of microcontrollers.
- SoC1.1 is a high performance STM32 MCU with our internally developed Neural Processing Unit (NPU) providing an order of magnitude benefit in both inference/w and inference/$ against alternative MPU solutions
- Soc1.1 will deliver MPU AI workloads at the cost and the power consumption of MCU. This a complete game changer that will open new ranges of applications for our customers and allow them to democratise AI at the edge.

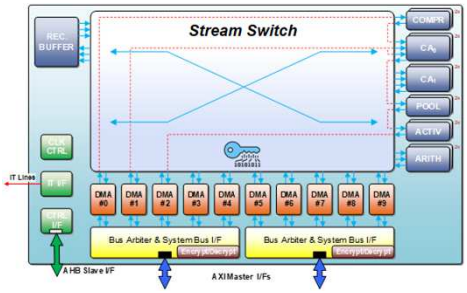


Digital cores

- The main CPU is a cortex M55 clocked at 800MHz. The micro architecture is new. It can be seen as an upgraded version of the Cortex M33 pipeline. Allowing it to push to higher frequencies.
- The real benefit comes from Helium or MVE the M-Profile Vector Extension , brings in up to roughly 150 new instructions and is aimed at rapidly executing advanced DSP and machine-learning code
- Effectively bringing Neon-style SIMD tech from the Cortex-A line to its Cortex-M cousins. Helium is engineered to be as small and energy-efficient as possible.
- It is foreseen from local caches and tightly coupled memories for data and instructions

NPU

- The Neural-ART is an internally developped IP that targets EDGE AI applications
  - It provides 600Gops of AI processing power at an average of 3TOPS/W energy efficiency. It is clocked at 1GHz
  - The Neural ART accelerator has ben architected for best integration into the system
  - To optimize power consumption and memory usage there's a tigh coupling of the NPU with the camera pipeline
  - Memory accesses to the SoC central memory or to external memories are optimized. Data is kept as much as possible internal to the accelerator. Again to reduce power consumption and optimize performance.



Technology and methodology

- SoC1.1: ARM core Hardening
- RTL synthesis for remaining digital cores
- Technology 16 FinFET

Use Cases

- Use Cases deployed (see demos)

Results and SotA

- Arm cortex®- M55 core
  - 1280 DMIPS / 3360 CoreMark
    New DSP extensions (MVE)
- ST YoloLC Neural Network: STM32 AI
  - 75x in performance versus STM32H7 @ 480 MHz
  - 25x in perf. versus STM32MP1 Dual A7 @ 800MHz

Lessons learned

- Capitalizing on more than 10 years of investment in edge AI to enable the rise of embedded AI with a comprehensive solution that covers a wide range of use cases

ANDANTE

AI for New Devices And Technologies at the Edge

andante-ai.edu linkedin.com/company/andante-ai mario.diaznava@st.com (coordinator)

"This project has received funding from the ECSEL Joint Undertaking (JU) under grant agreement № 876925. The JU receives support from the European Union's Horizon 2020 research and innovation programme and France, Belgium, Germany, Netherlands, Portugal, Spain and Switzerland"

Scan Me to visit website