Shaheen: An Open, Secure, and Scalable RV64 SoC for Autonomous Nano-UAVs

University of Bologna

L. Valente, A. Veeran, M. Sinigaglia, Y. Tortorella, A. Nadalini, N. Wistoff, B. Sá, A. Garofalo, R. Psiakis, M. Tolba, A. Kulmala,N. Limaye, O. Sinanoglu, S. Pinto, D. Palossi, L. Benini, B. Mohammad, D. Rossi

luca.valente@unibo.it

PULP Platform
Open Source Hardware, the way it should be!

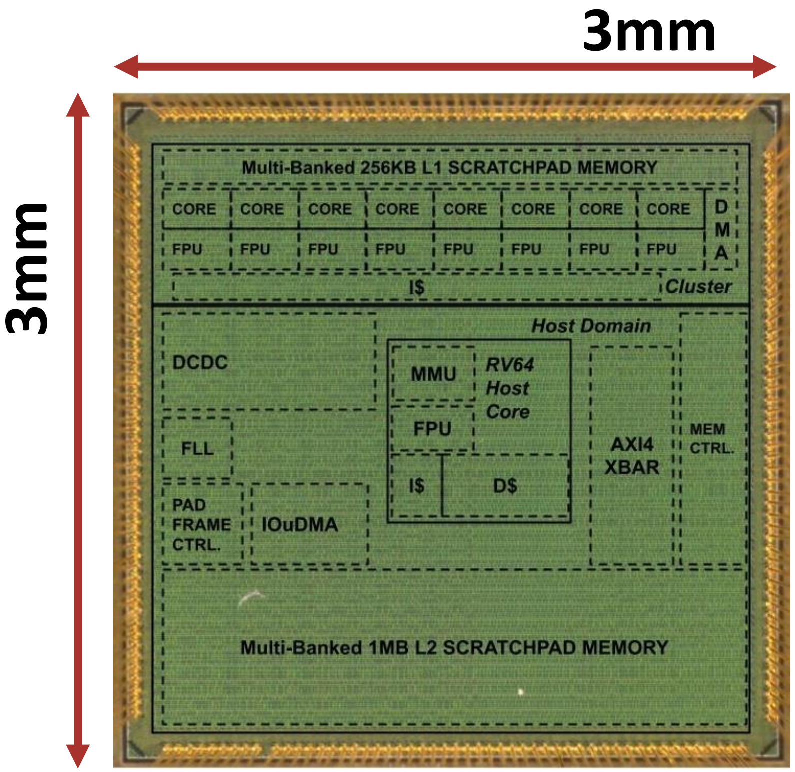
@pulp_platform
pulp-platform.org
youtube.com/pulp_platform

Autonomous Nano-UAVs

- **Versatility, safety, and cost-effet:**
  - small and agile
  - ideal for accessing hard-to-reach areas or tight spaces (inspection/maintenance)
  - relatively inexpensive to produce and operate
- **Requirements for future generation of nano-UAVs:**
  - Run increasingly complex multi-tasking workloads with large memory footprint
  - Within a few hundred mW power budget
  - Support for virtualization and secure operations in uncontrolled/hostile scenarios

Shaheen: an Open, Secure, and Scalable RV64 SoC for Autonomous Nano-UAVs

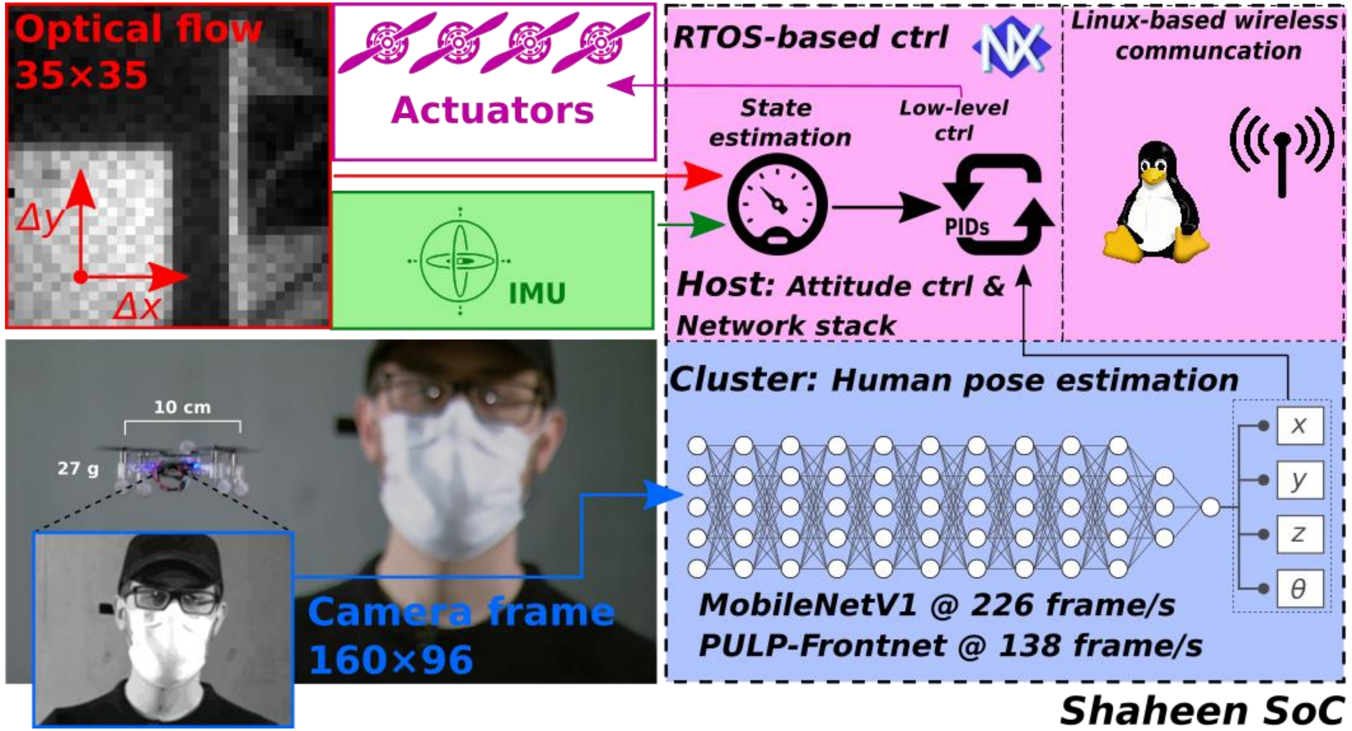
- **9mm^2 SoC in 22nm FDSOI technology with:**
  - A **RV64** Linux-capable CPU enhanced with
    - **Hypervisor** support
    - **Timing-channels mitigation**
  - An **energy efficient** programmable multi-core accelerator (PMCA) based on **8 RV32 cores with ML and DSP extensions**
  - Up to **512MB** of low-power off-chip **main memory**
  - **Logic locking** on key IPs within the architecture
  - **200mW power envelope**



Let’s dive in!

RV64 and custom RV32: the best of both worlds

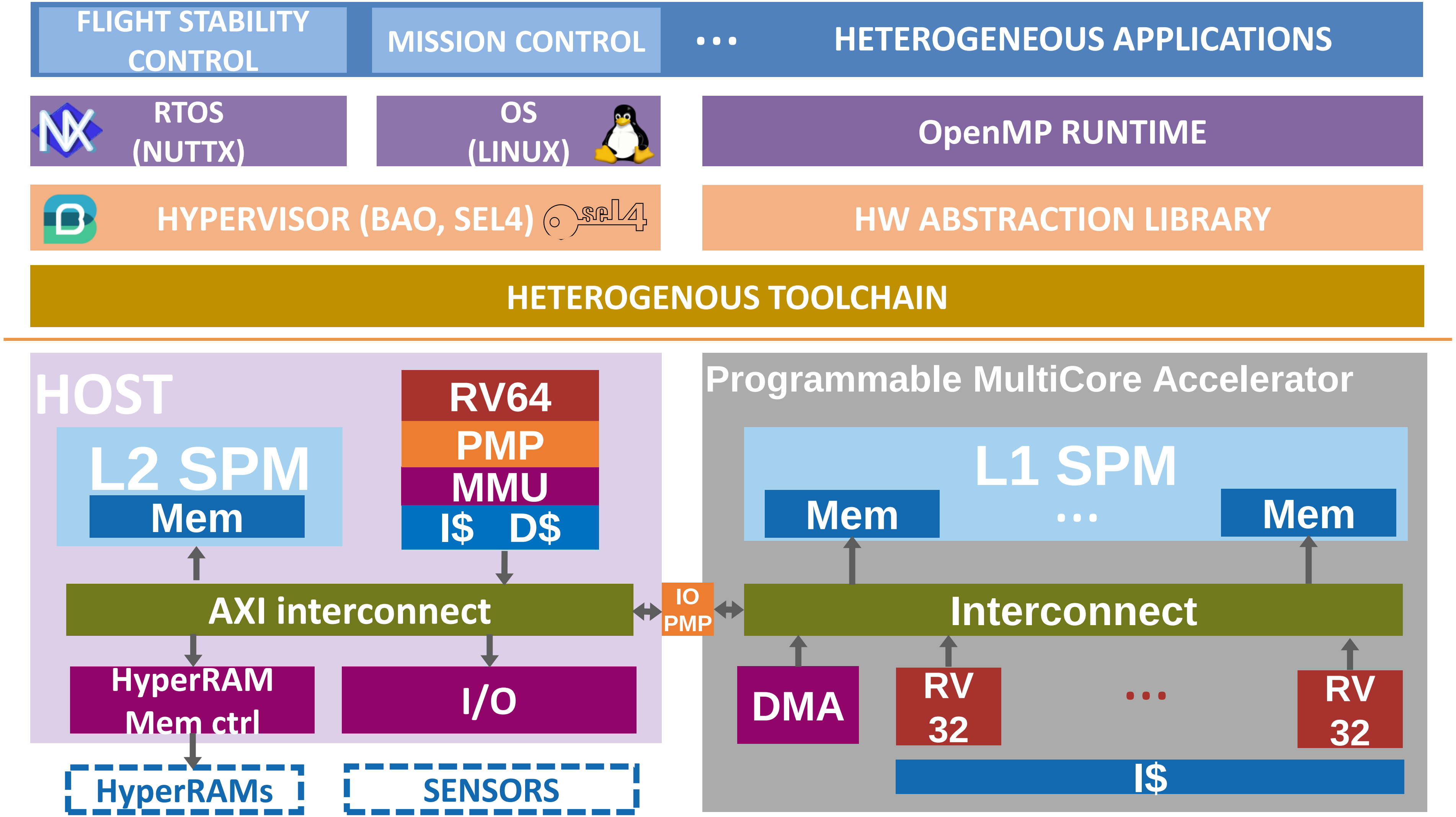
- **Different cores serve different parts of the target application**
- **Host:**
  - On top of the hypervisor:
    - Attitude control (**RTOS-based**)
    - **Linux-based** legacy software such as wireless network stack.
- **PMCA:**
  - The PMCA runs the **CNN-based** pose estimation task [1] fed by a low-resolution front-looking camera.



Optical flow
35×35
Δy
Δx
Actuators
IMU
RTOS-based ctrl
State estimation
Low-level ctrl
PIDs
Host: Attitude ctrl & Network stack
Linux-based wireless communcation
10 cm
27 g
Camera frame
160×96
Cluster: Human pose estimation
x
y
z
θ
MobileNetV1 @ 226 frame/s
PULP-Frontnet @ 138 frame/s
Shaheen SoC

[1] Cereda et.al. "Deep Neural Network Architecture Search for Accurate Visual Pose Estimation aboard Nano-UAVs", arXiV

Shaheen’s heterogeneous HW-SW stack



ALMA MATER STUDIORUM UNIVERSITÀ DI BOLOGNA · ETH zürich · جامعة خليفة Khalifa University · Universidade do Minho · جامعة نيويورك أبوظبي NYU ABU DHABI · USI/SUPSI IDSIA · TII Technology Innovation Institute

26/08/2023

HOT CHIPS

Timing-channel mitigation

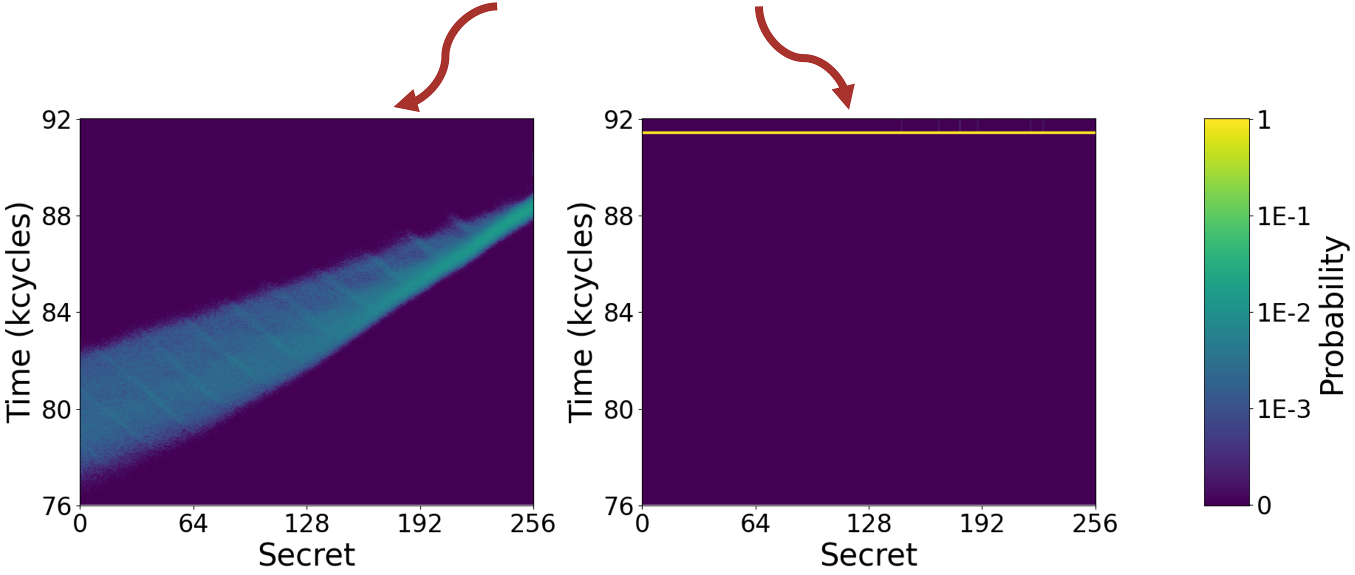
- **The 64-bit core implements the temporal fence instruction "*fence.t*"[2]:**
  - capability of clearing vulnerable microarchitectural states
  - enables a history-independent context-switch latency
  - low implementation effort (<1%)
  - low performance impact
  - negligible hardware costs

[2] Wistoff et. al. "Systematic Prevention of On-Core Timing Channels by Full Temporal Partitioning", IEEE Transactions on Computers, 2022

Timing-channel mitigation: prime and probe attacks

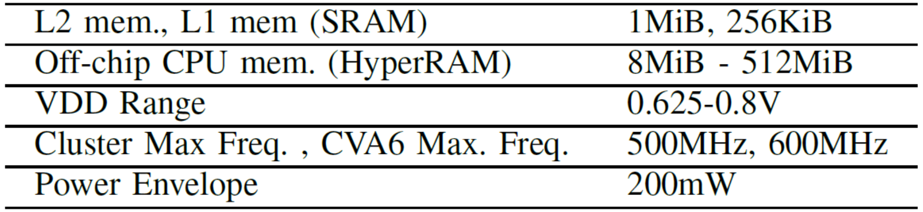
- **Prime and probe attacks:**
  - The spy brings the target HW into a known state (*prime*)
  - The OS switches to an applications containing a Trojan, accessing a subset of the HW resources to encode a secret
  - The execution switches back to the spy, which *probes* the execution time, correlated with the encoded secret.

Execution time depending on the encoded secret, without and with fence.t [2]

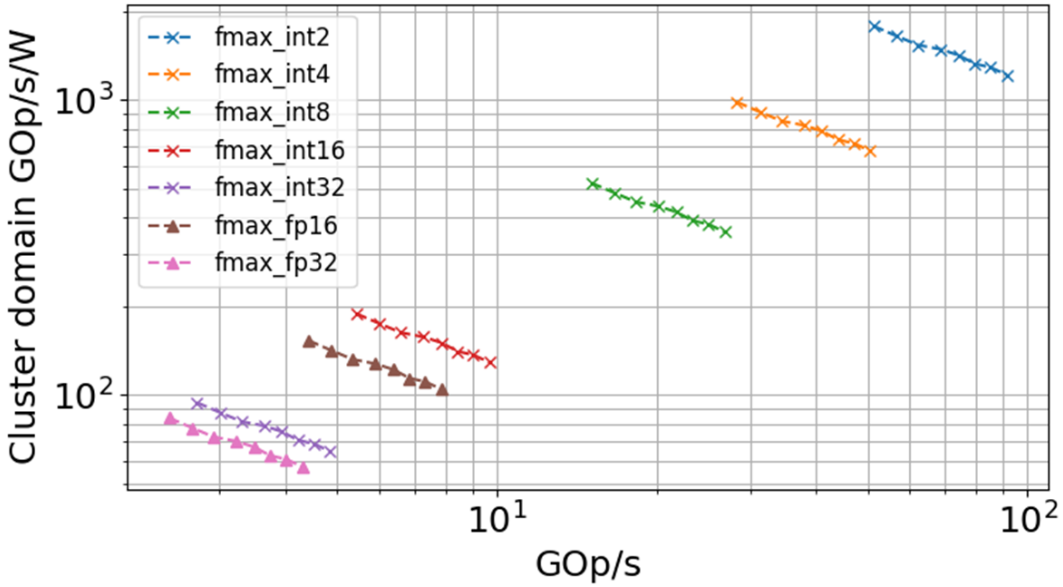


Physical implementation details: performance

- **Overview:**
  - **1MB+256kB** of scratchpad memory
  - **200mW** (120mW Host domain+ 80mW PMCA)
  - RV-32 cluster's cores aggressively optimized for FP-DSP and integer QNN inference [3]
  - The cluster can deliver up to:
    - **7.9GFLOp/s** on **16-bit** FP data
    - up to **90GOp/s** on 2-bit integer data @1.2TOPs/s/W (high-throughput mode)
    - up to 50GOp/s on 2-bit integer data @**1.8TOPs/s/W** (energy-efficient mode)



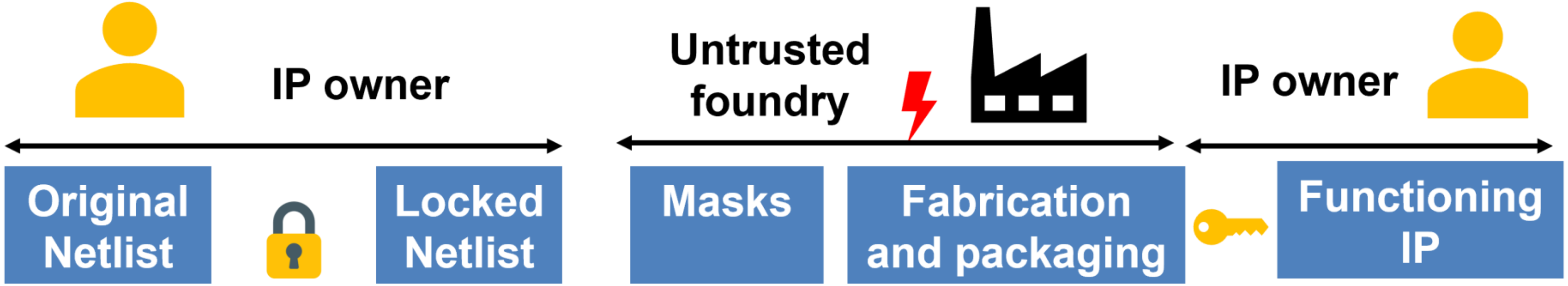
L2 mem., L1 mem (SRAM)	1MiB, 256KiB
Off-chip CPU mem. (HyperRAM)	8MiB - 512MiB
VDD Range	0.625-0.8V
Cluster Max Freq. , CVA6 Max. Freq.	500MHz, 600MHz
Power Envelope	200mW



[3] Nadalini et.al. "A 3 TOPS/W RISC-V Parallel Cluster for Inference of Fine-Grain Mixed-Precision Quantized Neural Networks", ArXiv, 2023

Physical implementation details: logic locking

- **Logic Locking:**
  - Consists in modifying a hardware IP to add a new input (“*logic locking key*”) to be applied to unlock the original IP functionality. Without the proper logic locking key, the chip is non-functional [4].
  - Between the interconnect and the memory controller
  - Between the interconnect and the PMCA



[4] Limaye et.al.”Thwarting all logic locking attacks: Dishonest oracle with truly random logic lockin” IEEE TCAS, 2020

Advancing the SoA

- **Overview:**
  - Match best in class (AI-IoT) SW performance
  - Only SoC for autonomous UAVs (within 200mW) with Hypervisor+Linux support
  - Advanced security features

	STM32-H7[5]	STM32-F4[6]	GAP8[7]	Vega[8]	Kraken[9]	Shaheen
Target board	Pixhawk	Crazyflie	AIDeck	AIDeck	AIDeck	**AIDeck / Pixhawk**
Technology	40nm	90nm	55nm	22nm FDSOI	22nm FDSOI	**22nm FDSOI**
Die Size	-	-	10mm2	12mm2	9mm2	**9mm2**
CPU	Cortex M7	Cortex M4	9x RI5CY	10x RI5CY-NN	9x RI5CY-XNN	**CVA6 + 8x FLEX-V**
Supported OS	RTOS	RTOS	RTOS	RTOS	RTOS	**Linux/RTOS/Hypervisor**
Host-compute FP support	SP-FPU, DP-FPU	-	-	SP-FPU	SP-FPU	**SP-FPU, DP-FPU**
Security Features	Crypto/hash accelerators	-	-	-	-	**Side-channel protection, Logic Locking, IOPMP**
Peak SW Performance	240MFOp/s(FP32) 390MOp/s (8b)	72MOp/s (8b)	6 GOp/s (8b)	7GFOp/s(FP16) 15,6GOp/s(8b)	3,12GFLOPs(FP32) 85GOp/s(2b)	**7.9GFOp/s(FP16)** **90 GOp/s(2b)**

[5] https://www.st.com/en/microcontrollers-microprocessors/stm32h7-series.html [6] https://www.st.com/en/microcontrollers-microprocessors/stm32f4-series.html [7] https://greenwaves-technologies.com/low-power-processor/ [8] Rossi et al ”Vega: A Ten-Core SoC for IoT Endnodes With DNN Acceleration and Cognitive Wake-Up From MRAM-Based State-Retentive Sleep Mode”, IEEE JSSC, 2021 [9] Di Mauro et al ”Kraken: A Direct Event/Frame-Based Multi-sensor Fusion SoC for Ultra-Efficient Visual Processing in Nano-UAVs”, HotChips 34, 2022

PULP Platform
Open Source Hardware, the way it should be!

Luca Valente luca.valente@unibo.it

Institut für Integrierte Systeme – ETH Zürich
Gloriastrasse 35
Zürich, Switzerland

DEI – Universitá di Bologna
Viale del Risorgimento 2
Bologna, Italy

@pulp_platform
pulp-platform.org
youtube.com/pulp_platform

ETH zürich
ALMA MATER STUDIORUM
UNIVERSITÀ DI BOLOGNA