ETH zürich

Integrated Systems Lab

ColibriES: End-to-End Efficiency for Neuromorphic Processing at the Edge

Georg Rutishauser[1],Robin Hunziker[1], Alfio di Mauro[1], Sizhen Bian[1], Michele Magno[2], Luca Benini[1,3]
[1]Integrated Systems Lab, ETH Zurich; [2]Project-Based Learning Center, ETH Zurich; [3]Università di Bologna

Introduction

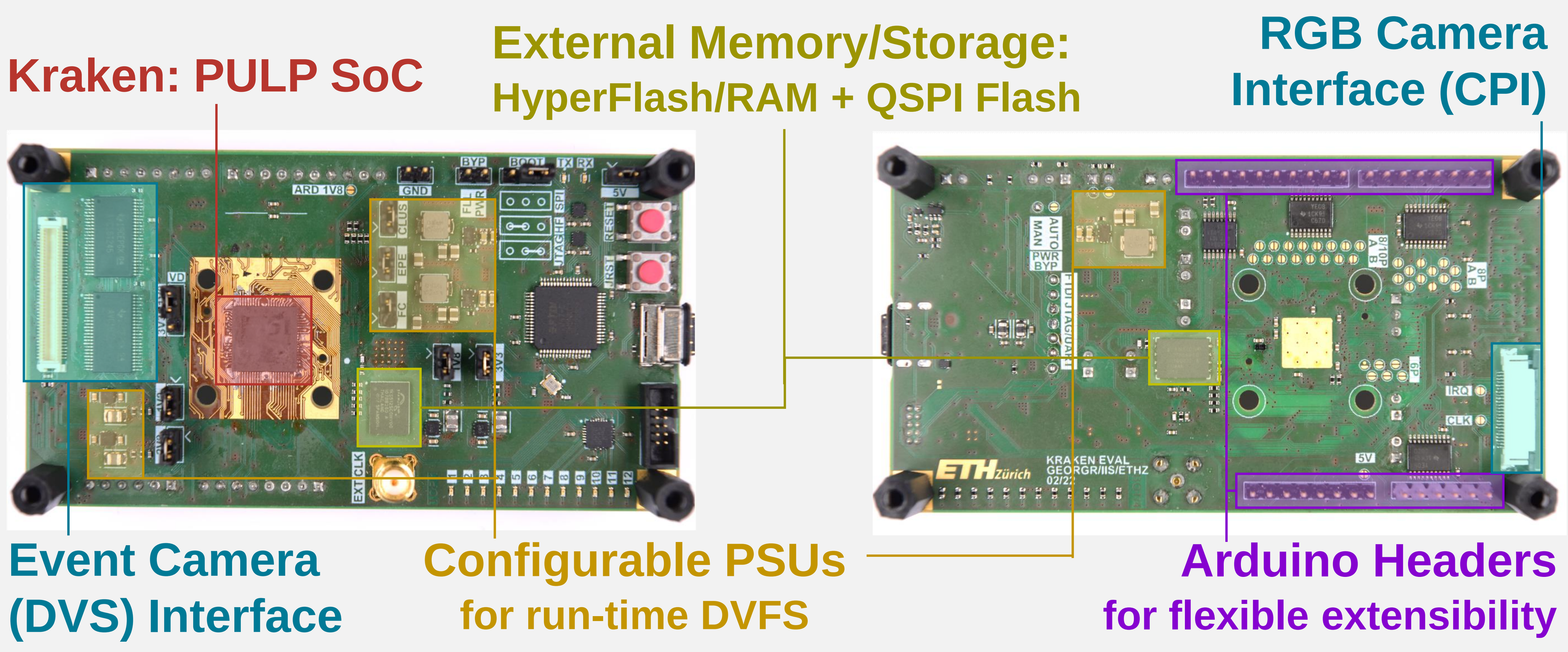
Event-driven paradigm: **improve sensing** and **computing efficiency** by only processing events describing relevant changes in the input:

Domain	Efficiency Benefit	Technology
Sensing	Avoid redundant data **collection & transmission**	**Dynamic Vision Sensors** (DVS/Event Cameras)
Computing	Avoid **processing** of irrelevant data	**Neuromorphic Computing:** Spiking Neural Networks (SNNs)

Existing platforms fail to harness the potential of event-driven processing for **ultra-low-power edge applications** due to **overheads in communication** and the **lack of an efficient and versatile host platform**. **ColibriES closes this gap**, bringing **end-to-end efficiency** in **neuromorphic and conventional algorithms** to the edge.

ColibriES: Enabling End-to-End Efficiency in Neuromorphic & Conventional Applications

ColibriES unites event-based neuromorphic, DNN-based and general-purpose computing in an **ultra-efficient edge system.** **End-to-end efficiency** is achieved through **integration** of **sensor interfaces** with **efficient heterogeneous processing** and **extensive communication and control capabilities** in the **Kraken** RISC-V SoC (see below). The **ColibriES evaluation PCB** connects Kraken to the outside world:

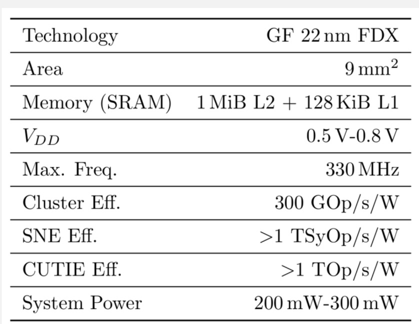


Kraken: The RISC-V SoC at the Heart of ColibriES

Kraken is a multi-core RISC-V based SoC from the PULP (**P**arallel **U**ltra **L**ow **P**ower) family. It offers the following features:

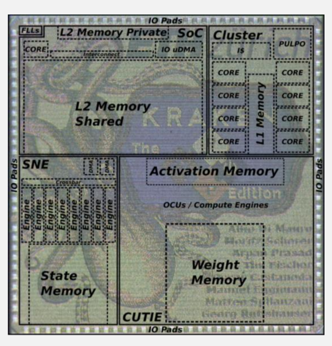
- **Rich peripheral set**, including **DVS and RGB camera** interfaces
- **Power management:** power gating of unused blocks
- **Ultra-efficient processing units** for multi-paradigm computing:

Paradigm	Processing Unit	Algorithms
Event-Driven	SNE [1]	**Spiking Neural Networks**
Frame-Based	CUTIE [2]	**Ternary Neural Networks**
General-Purpose	8-Core PULP Cluster	**Arbitrary**

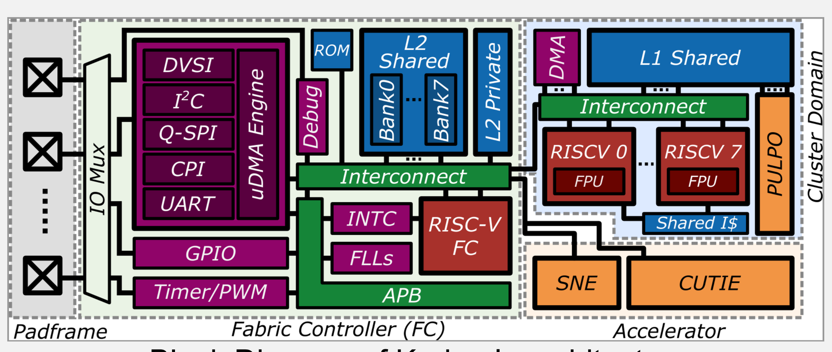


Technology	GF 22 nm FDX
Area	9 mm²
Memory (SRAM)	1 MiB L2 + 128 KiB L1
V_{DD}	0.5 V-0.8 V
Max. Freq.	330 MHz
Cluster Eff.	300 GOp/s/W
SNE Eff.	>1 TSyOp/s/W
CUTIE Eff.	>1 TOp/s/W
System Power	200 mW-300 mW

Kraken Key Figures of Merit



Die Micrograph of Kraken



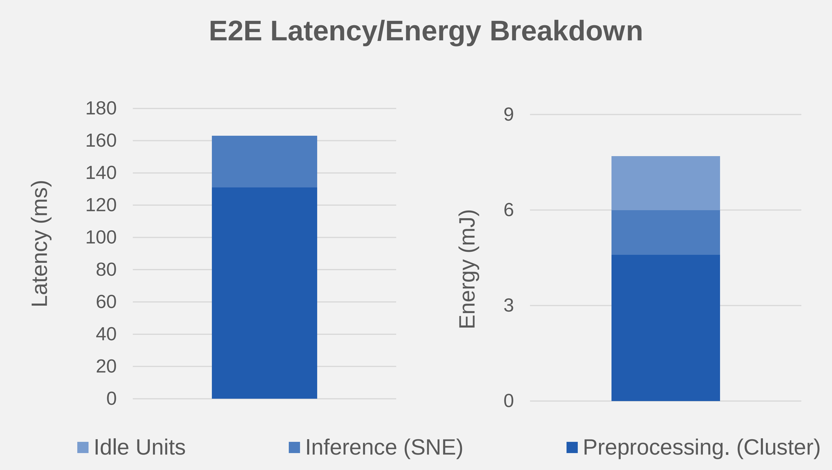
Block Diagram of Kraken's architecture

Results and SoA Comparison

We evaluate ColibriES's efficiency on the application of 11-class **gesture recognition from DVS data** with **a 7-layer spiking CNN (SCNN)**, using the DVS128 dataset from [2]. We evaluate the **latency** and **energy consumption** of the **end-to-end pipeline** of **data acquisition, SNN inference on SNE** and **actuation of a PWM output**.

Work/Platform	[3]/Intel Loihi	[4]/IBM TrueNorth	Ours/ColibriES
Network	5-layer SCNN	16-layer SCNN	7-Layer SCNN
Accuracy	90.5%	86.5%-94.6%	83%
End-to-End?	✗	✓	✓
$P_{proc,idle}$ (mW)	29.2	68.8-134.4	**17.7**
$P_{proc,inf}$ (mW)	N/A	88.5-178.8	**35.6**
$E_{proc,inf}$	N/A	28.8	**7.7**

- **2.5x lower inference energy** than TrueNorth
- **Further improvement potential** from improved preprocessing
- **Ultra-efficient data acquisition** with **native DVS interface**



Conclusion

With **ColibriES**, we have presented a fully embedded, low-power heterogeneous edge computing system. ColibriES:

- Brings **end-to-end event-driven computing** to the edge
- Offers a **wide range of peripherals**, including for **DVS cameras and RGB cameras**, enabling novel low-power sensor fusion approaches
- Unites **ultra-efficient** accelerators in one versatile platform:
  - **SNE** for SNN inference
  - **CUTIE** for ternary neural networks
  - **8-core PULP Cluster** for arbitrary compute tasks
- Achieves **SoA efficiency** on **end-to-end DVS-based gesture recognition** (IBM DVS-Gesture dataset):
  - **DVS-to-label energy** consumption of **7.7 mJ**
  - **SNN inference energy** of **1.4 mJ** on SNE
  - **35 mW** average inference power

References & Links

1. A. Di Mauro et al.: "SNE: An Energy-Proportional Digital Accelerator for Sparse Event-Based Convolutions", 2022 Design, Automation & Test in Europe Conference & Exhibition (DATE)
2. M. Scherer et al.: "A 1036 TOp/s/W, 12.2 mW, 2.72 μJ/Inference All-Digital TNN Accelerator in 22 nm FDX Technology for TinyML Applications", 2022 IEEE Symposium in Low-Power and High-Speed Chips (COOL CHIPS)
3. R. Massa et al.: "An Efficient Spiking Neural Network for Recognizing Gestures with a DVS Camera on the Loihi Neuromorphic Processor", Proceedings of the 2020 International Joint Conference on Neural Networks
4. A. Amir et al.: "A Low Power, Fully Event-Based Gesture Recognition System", CVPR 2017