Kraken: A Direct Event/Frame-Based Multi-sensor Fusion SoC for Ultra-Efficient Visual Processing in Nano-UAVs

Alfio Di Mauro (adimauro@ethz.ch)
Moritz Scherer (scheremo@ethz.ch)
Davide Rossi (davide.rossi@unibo.it)
Luca Benini (lbenini@ethz.ch)

PULP Platform
Open Source Hardware, the way it should be!
Toward nano and pico-size form factor UAVs

Advanced autonomous drone
[1] A. Bachrach, “Skydio autonomy engine: Enabling the next generation of autonomous flight,” IEEE Hot Chips 33 Symposium (HCS), 2021

Applications
• Search & rescue
• Post-disaster inspection
• Surveillance
• maintenance
deployment in tight space constraints?

Nano-drone
https://www.bitcraze.io/products/crazyflie-2-1

Can we fit sufficient “intelligence” in a 30X smaller payload and 20X lower energy budget?

• 3D Mapping & Motion Planning
• Object recognition & Avoidance
• 0.06m2 & 800g of weight
• Energy Capacity (Battery) 5410mAh

• Smaller form factor of 0.008m2
• Weight of 27g (30X lighter)
• Battery capacity of 250mAh (20X smaller)

https://www.skydio.com/skydio-2-plus

34cm
27cm

9.2cm
9.2cm

23cm

Achieving true autonomy on nano-UAVs

Execute complex visual task at high speed and robustness fully on board

Obstacle avoidance & Navigation

Environment exploration

Object detection
The Kraken
Kraken SoC Architecture

- 1 RISC-V Core (Fabric Controller)
- 1MB L2 Memory
- IO peripherals
- 128kB L1 Memory
- 8 RISC-V Cores (Compute Cluster)
- ML Accelerators
Autonomous navigation building blocks deployable on Kraken

RISC-V FC:
- RGB frames sent to CUTIE and the RISC-V Cluster
- Event-Frames streamed to SNE

RISC-V Cluster:
- “DroNet” Obstacle avoidance network [2]

SNE:
- “LIF-FireNet” Low-Latency Optical flow spiking network [3]

CUTIE:
- CIFAR10 Accurate Object recognition ternary network [4]

[2] D. Palossi et al., “A 64-mw dnn-based visual navigation engine for autonomous nano-drones,” IEEE Internet of Things Journal, 2019
[3] J. Hagenaars et al., “Self-supervised learning of event-based optical flow with spiking neural networks,” in Advances in Neural Information Processing Systems, 2021
[4] 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), 2022
Multi-Sensor direct data flow towards accelerators

- Autonomous IO subsystem
- Support for many protocols:
  - HyperBus, (4 x) I2C, QSPI, UART
- Support for visual sensors:
  - 1 x Event-Camera IF (DVSI)
  - 1 x RGB Camera IF (CPI)
Direct data processing
Processing event-frames on Kraken’s neuromorphic accelerator

A more complex dynamic than conventional DNNs neurons:

- Membrane Potential Accumulation/Activation: \[ 1 \text{ SynAcc} = 1 \text{ 4b-ADD} + 1 \text{ 8b-COMPARE} \]
- Membrane Potential decay: \[ 1 \text{ SynDec} = (1 \text{ 8b-MUL}) + (1 \text{ 8b-MUL} + 1 \text{ 8b-ADD}) \]

DVS stream of event (Event-frame)

- 3x3xT Event-Frame patch
- List of Coordinate (COO)

Synaptic connection

- Kraken’s Leaky Integrate & Fire (LIF) neuron

- Membrane Potential (State)

Output event

1x1xT tensor per output neuron

SNN’s neuron processing element (PE)
Mapping full neural networks on SNE

Neuromorphic Engine

- Neuron Sequencer
- Event Dispatcher
- Weight Memory 256 slots of 9 4bits

Patch 0
Patch 1

DVS stream of event (Event-frame)
L2 Memory Ports

Layer In Layer 1 Layer 2 Layer Out

Data Streamer Data Streamer

Engine Engine Engine Engine Engine Engine

8 Neuromorphic engines
- 16 Processing elements
- 64 Leaky Integrate & Fire (LIF) neurons per PE

Spiking Neural Engine (SNE) 522 SynDec & 24 SyAcc OP/cycle

Alfio Di Mauro - adimauro@ethz.ch

ETH Zürich

08/22/2022
Processing RGB frames on CUTIE ternary engine

- KxK window on all input channels unrolled, cycle-by-cycle sliding
- All weights for an output channel are held stationary in local buffer (latch-based)
- Completely unrolled inner products vs. systolic MAC \(\rightarrow\) one output activation per cycle!

**Ternary Weights (2bits)**

**Ternary Activations (2bits)**

**864 Ternary Mult.**

**2 Popcount units**

**Pooling**

**Threshold**

**Output channel compute unit (OCU)**
Kraken`s CUTIE Implementation

- Data in 1.6 bits (Ternary value) with Comp/Decomp on the fly
- Configuration in Kraken
  - 96 channels (OCUs)
  - 3x3 kernels
  - 64 x 64 pixels feature maps (158 KB)
  - 9 layers of weights (117 KB)
- Lots of TMAC/cycle
  - 96 OCUs, 96 Input channels, 3x3 kernels:
    - $96 \times 96 \times 3 \times 3 = 82'944$ Ternary-MAC/cycle
Silicon prototype
Physical implementation

- GlobalFoundries 22nm FDX technology
- QFN88 chip package, 9mm² chip area
- 0.5V to 0.9V operating voltage
- Cluster Max Freq: 370MHz
- CUTIE Max Freq: 140MHz
- SNE Max Freq: 220MHz
- Independent clock/power domain:
  - RISC-V Cluster
  - SNE
  - CUTIE
RISC-V Cluster Power/Performance tradeoff

Parallel Convolutional Benchmark (8 Cores)

- SIMD operation to maximize power/performance
- Wide range of numerical precision (32bits to 2bits)
- peak throughput of 0.98 MAC/cycle/core
- High throughput mode
  - 380MHz @ 0.9V (118mW)
  - 90GOP/s @ 750 GOP/s/W (2bit)
- High efficiency mode
  - 130MHz @ 0.5V (15mW)
  - 30GOP/s @ 1.9TOP/s/W (2bit)

DroNet [2]
Obstacle avoidance: 28 inf/s

[2] D. Palossi et al., “A 64-mw dnn-based visual navigation engine for autonomous nano-drones,” IEEE Internet of Things Journal, 2019
SNE Power/Performance tradeoff

Parallel 5-layers SNN inference benchmark (8 SNE engines)

- High throughput mode
  - 220 MHz @ 0.8V (98mW)
  - 55 GSyOP/s @ 0.4 TSyOP/s/W
- High efficiency mode
  - 90MHz @ 0.5V (23mW)
  - 18GSyOP @ 1.1 TSyOP/s/W

LIF-Firenet [2] Optical flow:
- 20k inf/s @ 8uJ/inf (1% activity)
- 1k inf/s @ 170uJ/inf (20% activity)

[3] J. Hagenaars et al., “Self-supervised learning of event-based optical flow with spiking neural networks,” in Advances in Neural Information Processing Systems, 2021
CUTIE Power/Performance tradeoff

Neural network inference benchmark

- High throughput mode (0.85V)
  - 55 TOp/s @ 450 TOp/s/W
- High efficiency mode (0.5V)
  - 15 Top/s @ 1036 TOp/s/W

CIFAR-10 – Ternary, Object detection [4]

Accuracy: 86%
Energy: 2.72µJ/inf

[4] 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), 2022
Advancing the SOA on all tasks

RISC-V Cluster

- Comparable 32bits-8bits SOA Energy efficiency to other PULPs [7]
- The highest energy efficiency on sub-byte SIMD operations (4b-2b)

SNE

- 1.7X higher than SOA [5] energy/efficiency

CUTIE

- 2X higher energy efficiency improvement over SOA [6]

[5] L. Deng et al., “Tianjic: A unified and scalable chip bridging spike-based and continuous neural computation,” IEEE Journal of Solid-State Circuits 2020
[6] B. Moons et al., “Binareye: An always-on energy-accuracy-scalable binary cnn processor with all memory on chip in 28nm cmos,” in Proc. IEEE CICC, 2018
[7] D. 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 Journal of Solid-State Circuits, 2022.
In conclusion
Kraken [10] can solve three complex visual tasks on-chip
Enable autonomous navigation on nano-UAVs!
✓ Optical flow from Event-Frames ➔ SNE [8]
✓ Obstacle avoidance from RGB frames ➔ RISC-V
✓ Object detection from RGB frames ➔ CUTIE [9]
✓ Vertical software stack to deploy applications

Next steps:
• Design a nano-drone form factor Kraken PCB
• Mount it on a Crazyflie drone platform

[10] Kraken: A Direct Event/Frame-Based Multi-sensor Fusion SoC for Ultra-Efficient Visual Processing in Nano-UAVs
https://www.research-collection.ethz.ch/handle/20.500.11850/565105

Thanks!