Open Platforms for the Embodied AI era

Luca Benini <luca.Benini@unibo.it, lbenini@ethz.ch>
Embodied AI

Path Towards Full Autonomy

On-car Computing
PMAX < 1.5KW

Energy Efficiency
\[
\left(\frac{1}{\text{Power-Time}}\right)
\]

10x/12Y by scaling vs. model complexity
10x/2Y

[SCR’23]

Efficient

Safe
Real-time
Secure
Start Small: Open Platform for Autonomous Nano-Drones

Advanced autonomous drone


- 3D Mapping & Motion Planning
- Object recognition & Avoidance
- 0.06m² & 800g of weight
- Battery Capacity 5410mAh

Nano-drone

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

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

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

https://www.bitcraze.io/products/crazyflie-2-1
Achieving True Autonomy on Nano-UAVs

Multiple, complex, heterogeneous tasks at high speed and robustness fully on board

Obstacle avoidance & Navigation

Environment exploration

Object detection

Multi-GOPS workload at extreme efficiency $\rightarrow P_{\text{max}} 100\text{mW}$
Multiple Heterogeneous Accelerators

*Brain-inspired*: Multiple areas, different structure different function!

1. Higher Mental Functions
   - Concentration
   - Planning
   - Judgment
   - Emotional expression
   - Creativity
   - Inhibition - Ability to control self

2. Motor Function Area
   - Eye movement and placement of eyes

3. Broca’s Area
   - Ability to talk
   - Ability to write

4. Motor Function Area
   - Ability to move muscles

5. Association Area
   - Short-term memory
   - Emotion

6. Sensory Area
   - Touching and feeling

7. Auditory Area
   - Hearing

8. Wernicke’s Area
   - Written and spoken language understanding

9. Somatosensory Association Area
   - Understanding of weight, texture, temperature, etc. for recognizing and comprehending an object

10. Visual Areas
    - Sight
    - Ability to recognize pictures
    - Awareness of size and shape

11. Functional Areas of the Cerebellum
    - Motor Functions
      - Coordination of movement
      - Balance
      - Posture

ETH Zürich
Multiple Heterogeneous Accelerators

The **Kraken**: an “Extreme Edge” Brain

- RISC-V Cluster (8 Cores + 1)
- CUTIE – dense ternary neural network accelerator
- SNE – energy-proportional spiking neural network accelerator

<table>
<thead>
<tr>
<th>Technology</th>
<th>22 nm FDSOI</th>
</tr>
</thead>
<tbody>
<tr>
<td>Chip Area</td>
<td>9 mm$^2$</td>
</tr>
<tr>
<td>SRAM SoC</td>
<td>1 MB</td>
</tr>
<tr>
<td>SRAM Cluster</td>
<td>128 KB</td>
</tr>
<tr>
<td>VDD range</td>
<td>0.55 V - 0.8 V</td>
</tr>
<tr>
<td>Cluster Freq</td>
<td>~370MHz</td>
</tr>
<tr>
<td>SNE Freq</td>
<td>~250MHz</td>
</tr>
<tr>
<td>CUTIE Freq</td>
<td>~140MHz</td>
</tr>
</tbody>
</table>

![Diagram of Kraken with RISC-V, CUTIE, SNE domains and FLLs](image)
CUTIE: Perception from Nyquist (Sampled) Sensors

- **Completely Unrolled Neural Inference Engine**: KxK window, all input channels, cycle-by-cycle sliding
- One OCU computes one output activation per cycle!
- Zeros in weights and activations, spatial smoothness of activations reduce switching activity
- 96 OCUs, 96 Input channels, 3x3 kernels: 96 * 96 * 3 * 3 = 82,944 TMAC/cycle (~1fJ/MAC)

Aggressive quantization and full specialization
SNE: Perception on Event Sensors

Event Sensors:
- DVS
- Ultra-low latency
- Energy-proportional interface

SNE works seamlessly with DVS (event-based) sensors

Leaky Integrate & Fire (LIF) neurons

Weight memory (~1.1kB)
256 slots of 9 4bits weights

Neuron Sequencer

State memory (4kB)
64x1632bit states

(85 SOP/cycle)

Event router

Spike event in

16 NGs

Spike event out

Spiking Neural Engine (SNE)

[Di Mauro et al. DATE22]
General Purpose PE: Domain-Specialized RV32 Core

8-bit Convolution

Vanilla

RISC-V core

Specialized for AI

Init NN-RF (outside of the loop)
lp.setup
pv.nnsdotup.h s0,ax1,9
pv.nnsdotsp.b s1, aw2, 0
pv.nnsdotsp.b s2, aw4, 2
pv.nnsdotsp.b s3, aw3, 4
pv.nnsdotsp.b s4, ax1, 14
end

15x less instructions than Vanilla!

Specialization Cost: Power, Area: 1.5x↑ but Time 15x↓ → E = PT 10x↓
Parallel, Ultra-Low Power (PULP) PE Cluster

- As VDD decreases, operating speed decreases
- However efficiency increases → more work done per Joule
- Run parallel to get performance and efficiency!

AI is parallel and scales
More paralle with NN size

Better have N PEs at optimum
Energy efficiency than 1 PE
running fast at low efficiency
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]

CUTIE, SNE can work concurrently for SNN + TNN “fused” inference (never done so far)

From Drones to Cars: Stepping up

- **Microcontroller class of devices**
  - Infineon AURIX Family MCUs
  - Control tasks, low-power sensor acquisition & data processing
    Features: lockstepped 32-b HP TriCore CPU, HW I/O monitor, dedicated accelerators

- **Powerful real-time architectures**
  - ST Stellar G Series (based on ARM Cortex-R cores)
  - Domain controllers and zone-oriented ECUs
    Features: HW-based virtualization, Multi-core Cortex-R52 (+NEON) cluster in split-lock, vast I/Os connectivity

- **Application class processors**
  - NXP i.MX 8 Family
  - ADAS, Infotainment
    Features: Cortex-A53, Cortex-A72, HW Virtualization, GPUs
Carfield: Efficiency + Safety, Security, RT-Predictability

Main Computing and I/O System

Accelerators Domain

### SAFETY ISLAND
- DATA SPM
- INSN SPM
- CV32E4
- CV32E4
- CV32E4
- TRI-LOCKSTEP
- CLIC INTC

### SECURITY ISLAND
- SPI
- JTAG
- SRAM wECC
- BOOT ROM
- IBEX
- RV-PLIC
- SHA2
- KEYS
- TRNG
- AES128
- OTPs
- K-H-MAC
- WATCHDOG
- MAILBOXes
- L2 MULTI-BANK SPM
- BK
- BK
- BK
- BK
- BK
- BK
- ECC
- MULTI-PORT (AXI)

### PREDICTABLE AXI INTERCONNECT

### I/Os AND PERIPHERALS
- UART
- QSPI
- Serial Link
- CAN
- GPIOs
- ETH
- WATCH DOG
- I2C
- SPI
- TIMERS

### HYPERBUS MEMORY CONTROLLER
- LAST LEVEL CACHE (LLC)

### HOST SUBSYSTEM
- MMU
- D$/I$ wECC
- CVA6 wH-EXT
- FPU
- CLIC INTC

### LOCAL INTERCONNECT
- DMA
- RV0
- RV1
- RV (N-1)
- Tensor Core

### FP VECTOR CLUSTER (SPATZ)
- L1 MULTI-BANKED SPM
- LOCAL INTERCO
- CTRL CC
- DMA
- PE0
- CC0
- VRF
- PE1
- CC0
- VRF
- I$
How Do We Handle Safety-Critical and Real-Time Tasks?

- Protection against transient faults (safety)
- Predictable On-Chip Communication (RT)
- Reduced contentions to access critical shared memory resources (RT)
- Safety-critical applications running on a RTOS
- **Three CV32E40 cores** physically isolated operating in **lockstep** (single HART) and **fast HW/SW recovery** from faults
- **ECC protected scratchpad memories** for instructions and data
- **Fast and Flexible Interrupts Handling** through RISC-V compliant CLIC controller
- **AXI-4 port** for in/out communication
Predictable On-Chip Communication (AXI RT)

- AXI4 inherently unpredictable
- Minimally Intrusive Solution
  - No huge buffering, limited additional logic
  - Solution verified in systematic worst-case real-time analysis
- AXI Burst Splitter
  - Equalizes length of transactions to avoid unfair BW distribution in round-robin scheme
- AXI Cut & Forward
  - Configurable chunking unit to avoid long transaction delays influencing access time to the XBAR
- AXI Bandwidth Reservation Unit
  - Predictably enforces a given max nr of transactions per time period (to each master)
  - Per-address-range credit-based mechanism
  - Periodically refreshed (or by user)

[Restuccia et al. DAC 2020]
[Pagani et al. ECRTS 2019]
- **Host [Cheshire]**
  - Dual-Core 64-bit RISC-V processor; 2.45 mm²; 600 MHz;

- **Security Island**
  - Low-power secure monitor; 1.94 mm²; 100 MHz;

- **Safety Island**
  - 0.42 mm²; 500 MHz

- **Re-configurable L2 Memory Subsystem**
  - 1MB; 2.33 mm²; 500 MHz

- **HMR Integer Cluster**
  - 1.17 mm²; 500 MHz;

- **Vectorial FP Cluster**
  - 1.14 mm²; 600 MHz;

- **Hyperbus**
  - 2 PHY, 2 Chips; 200 MHz; Max BW 400 MB/s

---

Modules marked with (*) are not in scale
**GF12, target 1GHz (typ)**

- **2 AXI NoCs (multi-hierarchy)**
  - 64-bit
  - 512-bit with “interleaved” mode

**Peripherals**

- **Linux-capable manager core CVA6**

**6 Quadrants: 216 cores/chiplet**

- 4 cluster / quadrant:
  - 8 compute +1 DMA core / cluster
  - 1 multi-format FPU / core (FP64,x2 32, x4 16/alt, x8 8/alt)

**8-channel HBM2e (8GB) 512GB/s**

- **D2D link (Wide, Narrow) 70+2GB/s**

- **System-level DMA**

- **SPM (2MB wide, 512KB narrow)**

Peak 384 GDPflop/s per chiplet
Occamy: RISC-V goes HPC Chiplet!

- Off-die Serial Link: 8 GB/s
- System-level DMA: 8 GB/s
- Die-to-Die Serial Link: 64 GB/s
- C2C: 64 GB/s
- C2Mem: 384 GB/s
- HBM2e DRAM: <410 GB/s
- HBM2e PHY: 512 GB/s
- NoC(P)
- Group-to-Group: 384 GB/s

**Periph**
- 64bit CVA6 Host
- 512KB SPM 64bit
- 1MB SPM 512bit
- ZeroMem 8GB / 512bit
- SPI
- I2C
- UART
- GPIO
- Timers
  - Runs Linux
  - Peripheral Manager
  - <1% traffic

**Global NoC**
- 32b
- 64b
- 512 b
- 64b
- 512 b

**Multi-cluster Multi-core Compute**
- 6 groups of each 4 clusters
- Each cluster has 8 compute cores + 1 DMA core

**Total of 216x Snitch cores with Multi-precision FPU (64 to 8)**
What’s Next? The era of Foundation Models

- Versatility and Multi-modality
  - Natural language processing, computer vision, robotics, biology, …

- Homogenization of models
  - Transformers as foundation models

- Self-supervision, Fine-tuning
  - Self-supervised training on large-scale unlabeled dataset
  - Fine-tune (few layers) on specific tasks with smaller labeled datasets.

- Zero-shot specialization
  - Prompt engineering for new tasks

Challenges in Attention

- Attention matrix is a square matrix of order input length
  - Computational complexity
  - Memory requirements
- MatMul & Softmax dominate
- Linear layers are next
Matmul Benefits from Large Shared-L1 clusters

- **Why?**
  - Better global latency tolerance if $L1_{\text{size}} > 2 \times L2_{\text{latency}} \times L2_{\text{bandwidth}}$ (Little’s law + double buffer)
  - Smaller data partitioning overhead
  - Larger Compute/Boundary bandwidth ratio: $N^3/N^2$ for MMUL grows linearly with $N$

- **A large “MemPool”**
  - 256+ cores
  - 1+ MiB of shared L1 data memory
  - $\leq$ 10 cycle latency (Snitch can handle it)

- **Physical-aware design**
  - WC Frequency > 700+Mhz
  - Targeting iso-frequency with small cluster

*Butterfly Multi-stage Interconnect 0.3 req/core/cycle, 5 cycles*
MemPool + Integer Transformer Accelerator (ITA)

Boosting dot product & matmul
MemPool + Integer Transformer Accelerator (ITA)

Transformer Accelerator
- INT8 operand precision
- Builtin data marshaling & pipelined operation
- Streaming Softmax (support in QuantLib)
- Last layer for MH-Attention, the head accumulation computed in cores

High-performance multi-core system
- Flexible, programmable snitch-based architecture
- 192 cores split into 48 tiles
- Support convolutions and “exotic” operators
Offloading Attention Operation to ITA

- **Performance** increase of 15x
- **Energy Efficiency** increase of 36x
- **Area Efficiency** increase of 74x

![Bar chart showing performance, energy efficiency, and area efficiency increases for different configurations of ITA.](chart.png)
Embodied AI vision: Transformers everywhere?

Software 1.0 tools “classical computer”

- Disk
- File system (+embeddings)
- Calculator
- Python interpreter

Peripheral devices I/O

- video
- audio

CPU

Ethernet

Browser

Context window

Efficient

Safe

Real-time

Secure
Research on open-source energy-efficient computing architectures

Started in 2013, we are celebrating 10 years of our project this year

Led by Luca Benini

Involves ETH Zürich (Switzerland) and University of Bologna (Italy)

Large group of almost 100 people

Thank You!