GenAI and the Transformation of Edge Computing: Leveraging Heterogeneous Circuits for Innovation

Jose Miranda

EPFL - Embedded Systems Laboratory

jose.mirandacalero@epfl.ch

03/27/2024
Outline

- The Edge
  - Current technological landscape

- GenAI into the Edge
  - Computation and Energy
  - Domain-specific accelerators and IMC

- Deeply Heterogeneous SoCs
  - Open-source frameworks

- Overcoming Computational and Energy constraints

- Future directions
The Edge: Current technological Landscape

- Automation and Efficiency
- Privacy-centric or Privacy by Design
- Environmental stepwardship
- 5G Adoption
- Deployment Optimization Frameworks
- Integration of GenAI into the Edge
- EaaS
GenAI into the Edge

$76 billion by 2028
(TIRIAS RESEARCH)

20% Offload to Edge

- $15 billion
- 800 MW

On-device AI will help democratize GenAI and ensure inclusion in the economy

SoCs
Performance improvements

Parameter growth

Average Inference Model Parameter Growth
Running On Edge Devices

Source: TIRIAS Research
GenAI into the Edge: Computation and Energy

- **GOAL**
  - 300 TOPS/W
  - 20 TOPS/mm²

- Two technical challenges:
  - Achieving a small, power efficient Multiply **Compute** Elements
  - Achieving a small, scalable, and power-efficient Multiply Accumulate **Storage**

**E**(memory access) >> **E**(computations)
GenAI into the Edge:
Domain-Specific Accelerators and IMC

GenAI into the Edge: Domain-Specific Accelerators and IMC

One of the best DSA providing energy efficient inference with transformers (BERT)
95.6 TOPS/W – 1711 inferences/s/W – 0.7% ACC loss

B. Keller et al., "A 17–95.6 TOPS/W Deep Learning Inference Accelerator with Per-Vector Scaled 4-bit Quantization for Transformers in 5nm," 2022 IEEE Symposium on VLSI Technology and Circuits (VLSI Technology and Circuits), Honolulu, HI, USA, 2022, pp. 16-17.
GenAI into the Edge: Domain-Specific Accelerators and IMC

Heterogeneous integrations of domain-specific accelerators: KRAKEN by PULP

- RISC-V Cluster
- SNE – Spiking NN accelerator
- CUTIE – Ternary Neural Network
  - > 1 PetaOps/s/W

M. Scherer et al., "A 1036 T0p/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
GenAI into the Edge: Domain-Specific Accelerators and **IMC**

**Traditional Digital Accelerators** (GPU, TPU, FPGA)

**Problem #1:** Bit-by-bit movement of lots of data

**Problem #2:** Digital MAC (<5-10 TOPS/W)

**Current-based Analog IMC** (Transistors, NVM, Spintronics)

- Memory & Compute
- Array size limited by reduced SNR
- Matrix multiply output (compute results over some bits simultaneously)

**EnCharge AI Analog IMC** (Standard CMOS Capacitors)

- Memory & Compute
- Matrix multiply output (compute results over all bits simultaneously)

https://www.enchargeai.com/technology
Deeply Heterogeneous SoCs: Open source Frameworks

Configurability

1. RISC-V core
2. Coprocessor interface
3. Peripherals
4. Interrupt controller
5. Accelerator interface
6. Power manager
7. Bus topology
8. Number of banks

https://x-heep.epfl.ch/

https://www.esp.cs.columbia.edu/

https://www.dolphin-design.fr/chameleon-mcu-subsystem/
Overcoming computational and energy constraints

HEEPocrates tape-out
Overcoming computational and energy constraints

- **in-SRAM computing**
  - Exploiting bit-line computing $\rightarrow$ SIMD enabler
  - Blade bit-line computing
    - **New IMC architecture developed at ESL**: BLADE is an in-SRAM computing architecture that utilizes local word-line groups to perform computations at a frequency 2.8x higher than state-of-the-art in-SRAM computing architectures.
Future directions:

- Need for next-generation end-to-end and HW-aware deployment frameworks for deeply heterogeneous processor and SoC architectures
Thank you!

Jose Miranda

EPFL - Embedded Systems Laboratory
jose.mirandacalero@epfl.ch