moduleqnc
quantum & neuromorphic computing

IC-Design Automation Ecosystem

Rapid Full Custom IC-Design for the **A**nalog **De**ep **L**earning **I**nference **A**ccelerator – ADeLIA – using UniLib Plus

1 Edge AI: Design Challenge

Situation:

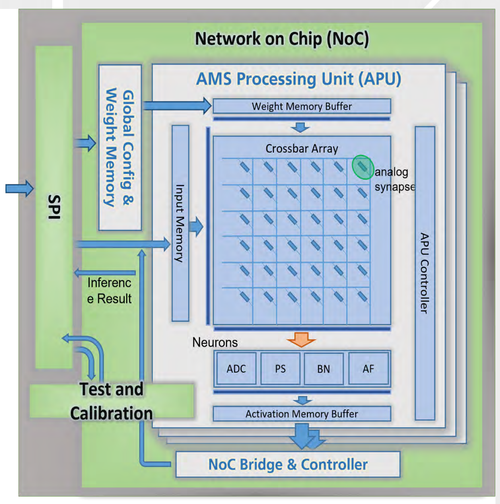
- **Cloud Edge Computing** and **Edge AI** are characterized by a large array of **heterogeneous system requirements**, driven by the **diversity of applications and system components**.
- **Analog in-memory computing** (IMC) offers a solution for applications demanding **ultra-low power and latency** with constrainted form factor (battery size)
  - Analog IMC extends battery lifetime and/or enables new use cases where micro-controller or DSPs cannot.

Challenge:

- Flexible and scalable analog/mixed-signal circuit design for Edge AI with reduced time to market

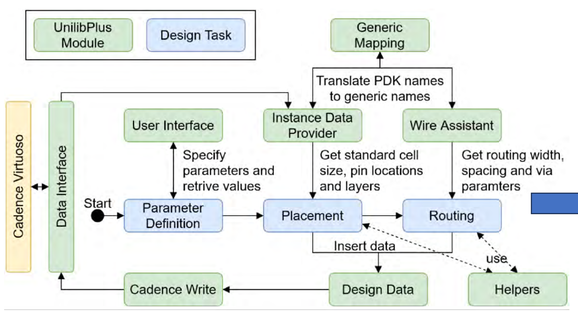
2 Template: ADELIA Gen 2 Architecture

- scalable analog/mixed-signal processing cores (APUs) connected via cyclic bus fabric
- **field programmable analog in-memory compute** for fast and efficient multiply and accumulate (MAC)
- non-linear activation function: ReLU
- ADCs between APUs additional flexibility and efficiency
- low power serial interface
- user selectable weight and input precision: up to 8 bits for each

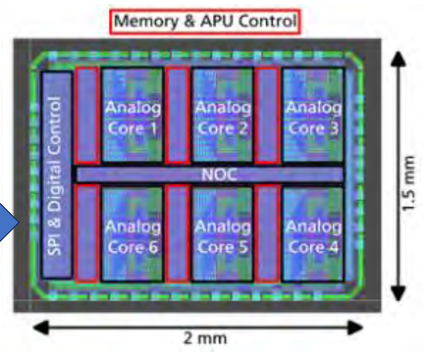


3 Solution: Automated Circuit Generation

- Design Automation for Analog Cells (IMC Crossbar Array and Neurons) not standardized
- **UnilibPlus**: **custom developed automation framework** compatible with Cadence Virtuoso analog design flow
- From custom design parameterization (e.g. No. of cores and neurons) to schematic and and layout in **10 mins design time**
- **Full custom verification:** Automated verification flow and simulation reports, with pass/fail testing



Flow chart: Module interaction between Cadence Virtuoso, UnilibPlus and a Design Task



Full chip layout snapshot

4 Results: VAD Example Use Case

KPI Name	Target	Measured
Accuracy	> 80 %	82.5%
Latency	< 3 ms	1.8 ms
Inference time	< 3 ms	1 ms
Power	< 1 mW	**260 µW**
Energy per Inference		**0.2 µJ**
Power Eff. 3b x 8b	1-5 TOPS/W	5.5 TOPS/W

- **Proven Silicon** with Demonstrator
- **first time right**
- Voice Activity Detection (VAD) Use Case
- Raspberry Pi host controller
- Detect Human Speech in Audio Signal (binary decision)
- VAD may be used as wake-up for “power-hungry” speech recognition algorithms
- Other neural networks in testing:
  - ECG monitoring
  - EMG monitoring
  - Keyword spotting
  - Gas sensor fingerprinting (smelling)

5 Outlook

- Design automation
  - **Chiplet version** with interface for hetero-integration (AHSI)
  - Improved verification flow
  - Continuous Integration pipeline for code (RTL + UniLib + Templates)
- ADeLIA template evolution
  - support LSTM-type neural networks, e.g. **RNN**
  - **support embedded non-volatile memory**, e.g. MRAM
  - support image object recognition

6 Impact

- 2 dissertations, 3 master thesis, 1 peer reviewed conference, 2 book chapters, 1 Patent
- Progress beyond SoA
  - Designed, verified and trained with full analog PVT variations
  - Execution of complete NNs
  - 200 nJ inference
- Reputable eNVM IC and IP provider interested in joint product
- Physical verification and lab evaluation with FMD equipment and EDA tool licenses

GEFÖRDERT VOM
Bundesministerium für Bildung und Forschung