Fibonacci – a Scalable Approach to Embedded Machine-learning

E. Azarkhish, P. Jokic, R. Cattenoz, C. Arm, S. Emery

Fibonacci is CSEM's state-of-the-art machine-learning (ML) system-on-chips (SoC) series, designed based on the principle of hierarchical scalability: The SoC can dynamically increase its computational performance by adding accelerator resources based on the application's needs, inspired by the Fibonacci number series. Its heterogenous architecture features a plurality of ML accelerators for temporal and spatial data analysis, energy-optimized on-chip memories, a flexible RISC-V microcontroller core, and a rich set of peripherals. Trained models can be deployed through CSEM's ML compiler, supporting most common formats (e.g., ONNX). Fibonacci targets the power consumption range of 10 uW-100 mW at performance of up to 160 GOPS non-sparse throughput and 2 TOPS/W efficiency.

A conceptual diagram of the Fibonacci SoC is illustrated below.

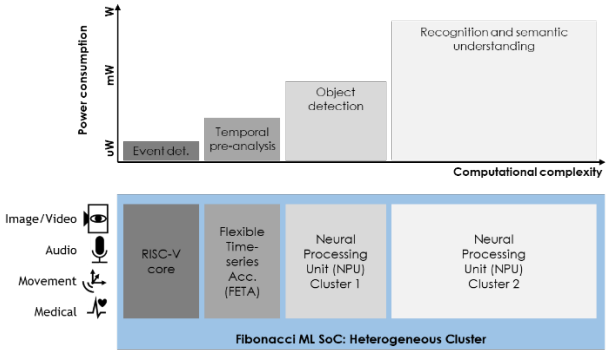

Figure 1: Hierarchical scalability offered by Fibonacci SoC for application-aware energy performance trade-off.

Fibonacci features a multi-cluster architecture with simultaneous support for spatial (e.g., CNN, ResNets) and temporal (RNN) ML network topologies. The Flexible Time-series Accelerator (FETA) cluster focuses on time-series signals, and the Neural Processing Unit (NPU) clusters target spatial models. The two NPU clusters offer similar functionalities at different power and performance targets, allowing the application designer to choose one/both based on the needs.

One key feature of the Fibonacci SoC is the availability of multiple sensing modalities and various data streams to facilitate at-edge data fusion, and multi-modal inference. The embedded transceivers include: I2S (4x), SPI (2x), I2C, OctoSPI, DCMI, UART, JTAG, GPIO, each providing dedicated data streams through various interconnection matrices and DMA engines (2x two channels) to the accelerators. Plus, a complete audio front-end from digital microphones including activity detection and Mel Frequency Cepstral Coefficients (MFCC) is implemented in Smart Front-end (SFE) unit, which can be reprogrammed for other types of time-series signals as well.

The memory hierarchy (illustrated in Figure 2) includes 4 MB of SRAM L2 memory, organized in multiple banks with a wide data bus interface. The accelerators use the same wide-bus interface for higher bandwidth access, but the processor and DMA engines access L2 memory through micro-caches (UC), small write-back caches with software managed coherence, to minimize energy and latency penalties. Each accelerator has small scratchpads (L1) to facilitate intermediate computations.

Power management in Fibonacci follows a hierarchical scheme as well, with a master finite state machine residing in an always-on domain, controlling, and coordinating power up/down sequences of six other domains.

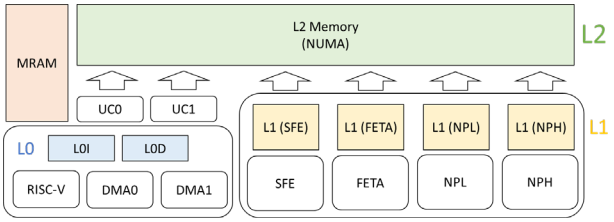

Figure 2: Memory hierarchy of the Fibonacci SoC.

512 kB of MRAM [1] is accessible to the RISC-V [2] cluster to enable zero-leakage weight or program storage for heavily duty-cycled application scenarios.

A mix of different voltages, standard cells and memory flavors are used to further optimize each power domain (a.k.a. cluster) to its range of responsibilities. The always-on cluster is implemented using ultra-low leakage (ULL) transistors, while the other domains use a mix of low VT (LVT) and super-low VT (SLVT) ones. On-chip clock generation and distribution are handled (thanks to existing internal IPs), with frequencies ranging from a few kHz to 400 MHz. To minimize the latency and energy overheads of inter-cluster communications, yet benefit from independent performance settings for individual clusters, a special qualifier-based handshaking mechanism along with a source-synchronous clocking approach are implemented across the chip.

The Fibonacci SoC aims to support multiple classes of applications in the fields of IoT edgeML and smart sensing, such as:

- Multi-modal concurrent data analysis from different sensor types (e.g., audio-visual sensor fusion).
- Multi-stage evaluation, selective/hierarchical execution with increasing complexity, and early exit to reduce energy consumption (e.g., activity tracking and analysis).
- Ultra-low power edge processing, down to µW power budgets with heavy duty cycling (e.g., condition monitoring).
- Spatial and time-series signal analysis (e.g., audio analysis).

The chip will be implemented in GF22FDX technology. A demonstrator is planned to target different application scenarios from the classes listed above.

[1] https://gf.com/blog/making-new-memories-22nm-emram-ready-displace-eflash/

[2] www.csem.ch/en/news/low-power-risc-v-integration-customization-and-soc