ColibriES: End-to-End Efficiency for Neuromorphic Processing at the Edge

Georg Rutishauser¹, Robin Hunziker¹, Affio di Mauro¹, Sizhen Bian¹, Michele Magno², Luca Benini¹,³
¹Integrated Systems Lab, ETH Zurich; ²Project-Based Learning Center, ETH Zurich; ³Università di Bologna

Introduction

Event-driven paradigm: improve sensing and computing efficiency by only processing events describing relevant changes in the input:

<table>
<thead>
<tr>
<th>Domain</th>
<th>Efficiency Benefit</th>
<th>Technology</th>
</tr>
</thead>
<tbody>
<tr>
<td>Sensing</td>
<td>Avoid redundant data collection &amp; transmission</td>
<td>Dynamic Vision Sensors (DVS/Event Cameras)</td>
</tr>
<tr>
<td>Computing</td>
<td>Avoid processing of irrelevant data</td>
<td>Neuromorphic Computing: Spiking Neural Networks (SNNs)</td>
</tr>
</tbody>
</table>

Existing platforms fail to harness the potential of event-driven processing for ultra-low-power edge applications due to overheads in communication and the lack of an efficient and versatile host platform. ColibriES closes this gap, bringing end-to-end efficiency in neuromorphic and conventional algorithms to the edge.

ColibriES: Enabling End-to-End Efficiency in Neuromorphic & Conventional Applications

ColibriES unites event-based neuromorphic, DNN-based and general-purpose computing in an ultra-efficient edge system. End-to-end efficiency is achieved through integration of sensor interfaces with efficient heterogeneous processing and extensive control capabilities in the Kraken RISC-V SoC (see below). The ColibriES evaluation PCB connects Kraken to the outside world:

- Kraken: PULP SoC
- Event Camera (DVS) Interface
- Configurable PSUs for run-time DVS
- External Memory/Storage: HyperFlash/GRAM + QSPI Flash
- RGB Camera Interface (CPI)
- Arduino Headers for flexible extensibility

Kraken: The RISC-V SoC at the Heart of ColibriES

Kraken is a multi-core RISC-V based SoC from the PULP (Parallel Ultra Low Power) family. It offers the following features:

- Rich peripheral set, including DVS and RGB camera interfaces
- Power management: power gating of unused blocks
- Ultra-efficient processing units for multi-paradigm computing:

<table>
<thead>
<tr>
<th>Paradigm</th>
<th>Processing Unit</th>
<th>Algorithms</th>
</tr>
</thead>
<tbody>
<tr>
<td>Event-Driven</td>
<td>SNE [1]</td>
<td>Spiking Neural Networks</td>
</tr>
<tr>
<td>Frame-Based</td>
<td>CUTIE [2]</td>
<td>Ternary Neural Networks</td>
</tr>
<tr>
<td>General-Purpose</td>
<td>8-Core PULP Cluster</td>
<td>Arbitrary</td>
</tr>
</tbody>
</table>

Results and SoA Comparison

We evaluate ColibriES’s efficiency on the application of 11-class gesture recognition from DVS data with a 7-layer spiking CNN (SCNN), using the DVS128 dataset from [2]. We evaluate the latency and energy consumption of the end-to-end pipeline of data acquisition, SNN inference on SNE and actuation of a PWM output.

<table>
<thead>
<tr>
<th>Work/Platform</th>
<th>Network</th>
<th>Accuracy (%)</th>
<th>End-to-End</th>
<th>$P_{actuation}$ (mW)</th>
<th>$P_{perception}$ (mW)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Intel Loihi</td>
<td>5-layer SCNN</td>
<td>90.5%</td>
<td>✔</td>
<td>29.2</td>
<td>17.7</td>
</tr>
<tr>
<td>IBM TrueNorth</td>
<td>16-layer SCNN</td>
<td>86.5%-94.6%</td>
<td>✔</td>
<td>68.8-134.4</td>
<td>17.7</td>
</tr>
<tr>
<td>Ours/ColibriES</td>
<td>7-layer SCNN</td>
<td>83%</td>
<td>✔</td>
<td>85.1-178.8</td>
<td>35.6</td>
</tr>
</tbody>
</table>

- 2.5x lower inference energy than TrueNorth
- Further improvement potential from improved preprocessing
- Ultra-efficient data acquisition with native DVS interface

Conclusion

With ColibriES, we have presented a fully embedded, low-power heterogeneous edge computing system. ColibriES:

- Brings end-to-end event-driven computing to the edge
- Offers a wide range of peripherals, including for DVS cameras and RGB cameras, enabling novel low-power sensor fusion approaches
- Unites ultra-efficient accelerators in one versatile platform:
  - SNE for SNN inference
  - CUTIE for ternary neural networks
  - 8-core PULP Cluster for arbitrary compute tasks
- Achieves SoA efficiency on end-to-end DVS-based gesture recognition (IBM DVS-Gesture dataset):
  - DVS-to-label energy consumption of 7.7 mJ
  - SNN inference energy of 1.4 mJ on SNE
  - 35 mW average inference power

References & Links