StreamingAI

AI-Energy-Awareness for the digital transformation of Austrian Industries.

Pro²Future

Alois Ferscha[1,2], Bernhard Anzengruber-Tanase[1], Michael Haslgrübler[1], Ekaterina Sysoykova[1], Georgios Sopidis[1], Behrooz Azadi[1], Michael Siegl[1], Miguel Vazquez[2], Patrick Denzler[2], Sepp Hochreiter[3]

Pro2Future GmbH[1], JKU-IPC (Institute of Pervasive Computing)[2] , JKU-IML (Institute of Machine Learning)[3]

[1] Science Park 4, Altenberger Strasse 69, 4040 Linz

[2] Science Park 3, Altenberger Strasse 69, 4040 Linz

[3] Science Park 3, Altenberger Strasse 69, 4040 Linz

MOTIVATION & GOALS

Streaming AI aims to drive low TRL, foundational research to develop AI for industrial applications. In contrast to conventional pre-trained, holistic, and resource-intensive AI,

i. **streaming machine learning methods**
ii. **on-device machine learning methods** are to be introduced,

thereby reducing dependence on mass training data and supporting ecological sustainability.

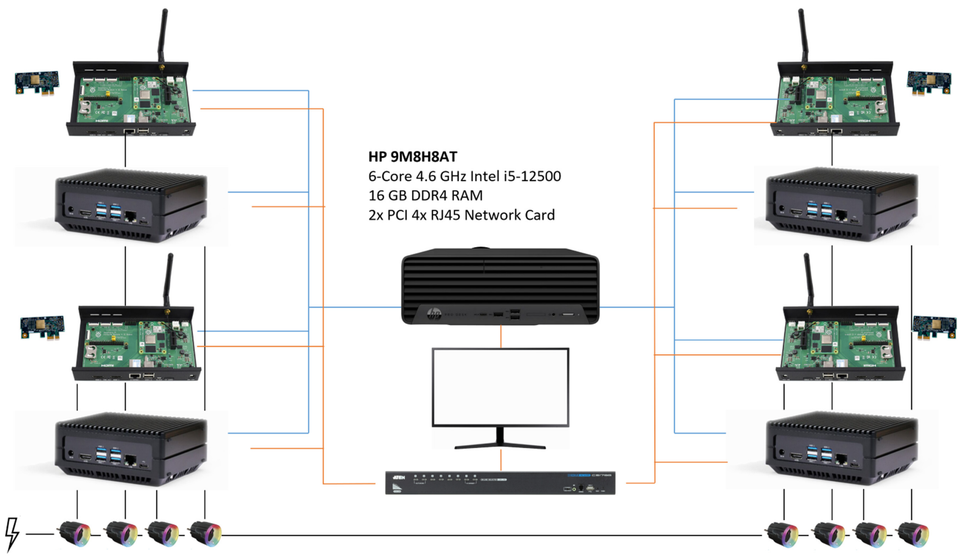
Project FactBox

Project Name StreamingAI
Project ID -
Duration 18 Months

Area 1
Area Perception

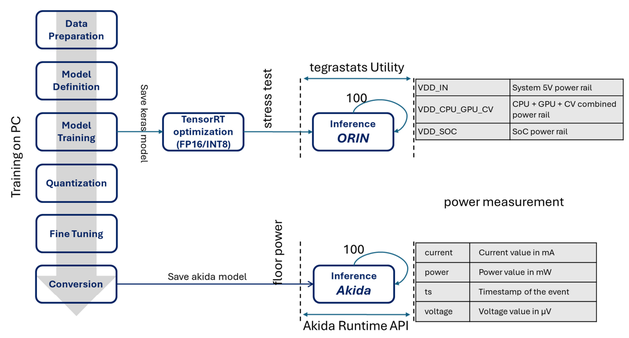
Project Lead
Dr. Bernhard Anzengruber-Tanase

TESTBED and AVAILABLE HARDWARE



Experiment Design & Approach

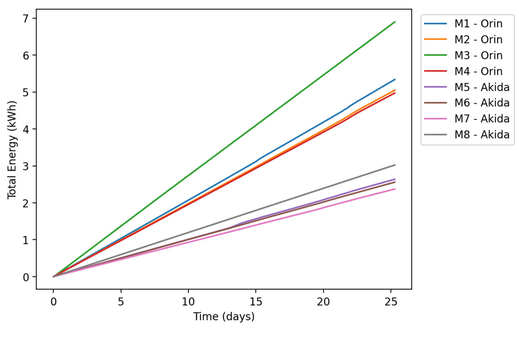
A trained vision model was converted to be deployed on both Nvidia Jetson Orin, as well as Brainchip Akida devices for inference. The energy cost of the entire compute boards was measured with a smart plug of the testbed.



Device	Code	Active Power (W)	Total Energy (Wh)	Temperature (C°)
ORIN	M1	10.38	0.49	43.30
	M2	9.66	0.49	44.91
	M3	12.69	0.60	45.16
	M4	-	-	-
Akida	M5	5.30	0.03	45.20
	M6	5.10	0.02	45.50
	M7	4.53	0.03	42.87
	M8	-	-	-

Inference Energy

The values measured by shelly plugs during inference demonstrate that neuromorphic devices consume notably lower amounts of energy. Our experiment yielded ~16-20 times lower energy consumption.



Cumulative Energy Use

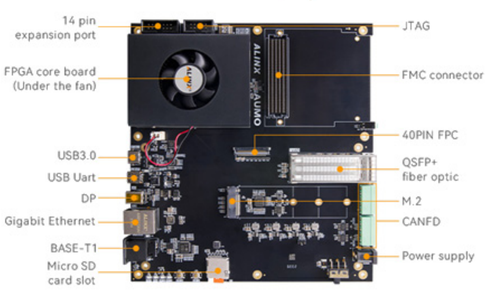
Energy consumption of each machine over 25 days shows that Orin devices exhibit a steeper consumption slope due to their higher active power draw, resulting in almost twice the energy usage compared to the investigated neuromorphic systems.

MPSoC/SOM :: AMD/Xilinx Zynq™ Ultrascale+™ MPSoC Z7-A

Compute:
- ARM® quad-core Cortex™-A53 up to 1.3 GHz
- ARM® dual-core Cortex™-R5F up to 533 MHz
- Mali-400 MP2 GPU
- 16nm FinFET+ FPGA fabric

Memory:
- 4 GByte (64-bit) DDR4 SDRAM on PS
- 4 GByte (64-bit) DDR4 SDRAM on PL

Storage:
- 64 MB QSPI Flash
- 32 GB eMMC Flash

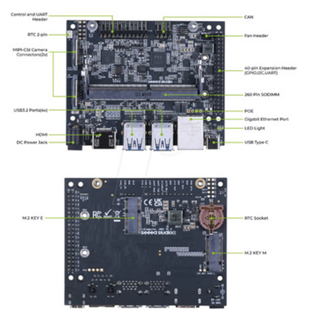


SOM :: Nvidia Orin NX

Compute:
- 8-core Arm® Cortex®-A78AE x 2 GHz
- NVIDIA Ampere DL up 100 TOPS

Memory:
- 16 GB 128-bit LPDDR5

Storage:
- 128 GB

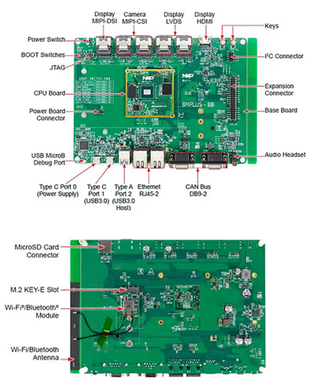


SoC :: NXP Semiconductors 8MPLUSLPD4-EVK

Compute:
- 4 x ARM® Cortex-A53 von bis zu 1,8 GHz
- 1 x ARM® Cortex-M7 von bis zu 800 MHz
- Neural Processing Unit (NPU) 2.3 TOPS

Memory:
- 6 GB LPDDR4

Storage:
- 32 GB eMMC 5.1
- 32 MB QSPI NOR

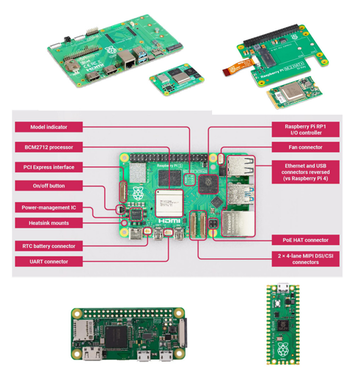


SBC :: Raspberry PI 5

Compute:
- 4x 64-bit Arm Cortex-A76 @2.4GHz
- VideoCore VII GPU
- Neural Processing Unit (NPU) 2.3 TOPS
- AI-Hat Extension with 26/13TOPS

Memory:
- Up to 16 GB LPDDR4X-4267 SDRAM

Storage:
- microSD card slot



Contact: Dr. Behrooz Azadi, Pro2Future GmbH, behrooz.azadi@pro2future.at, +43 732 2468 - 9469

Acknowledgement: This work was supported by Pro²Future II (FFG, 911655) and the Province of Upper Austria (Land OÖ).