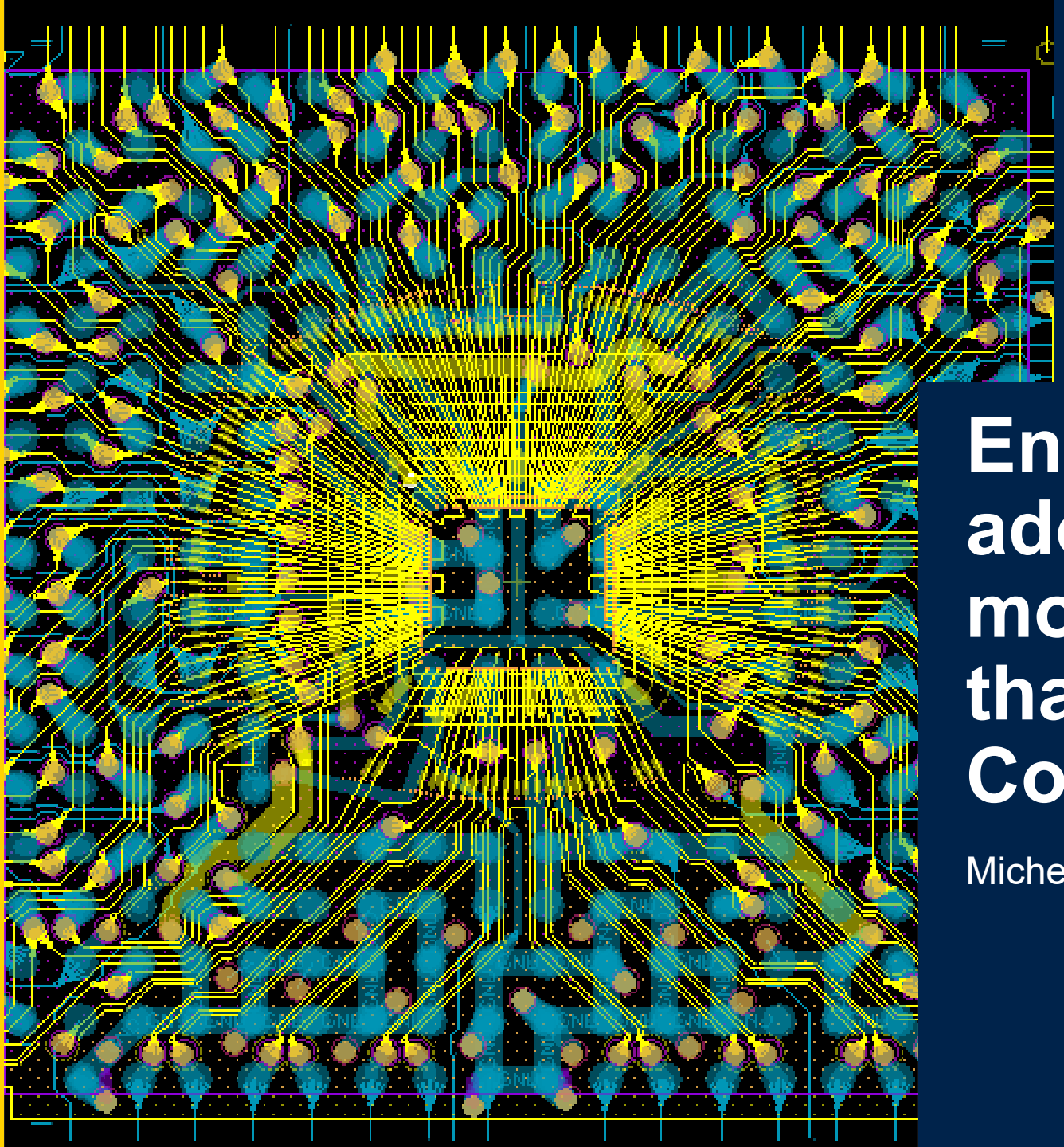
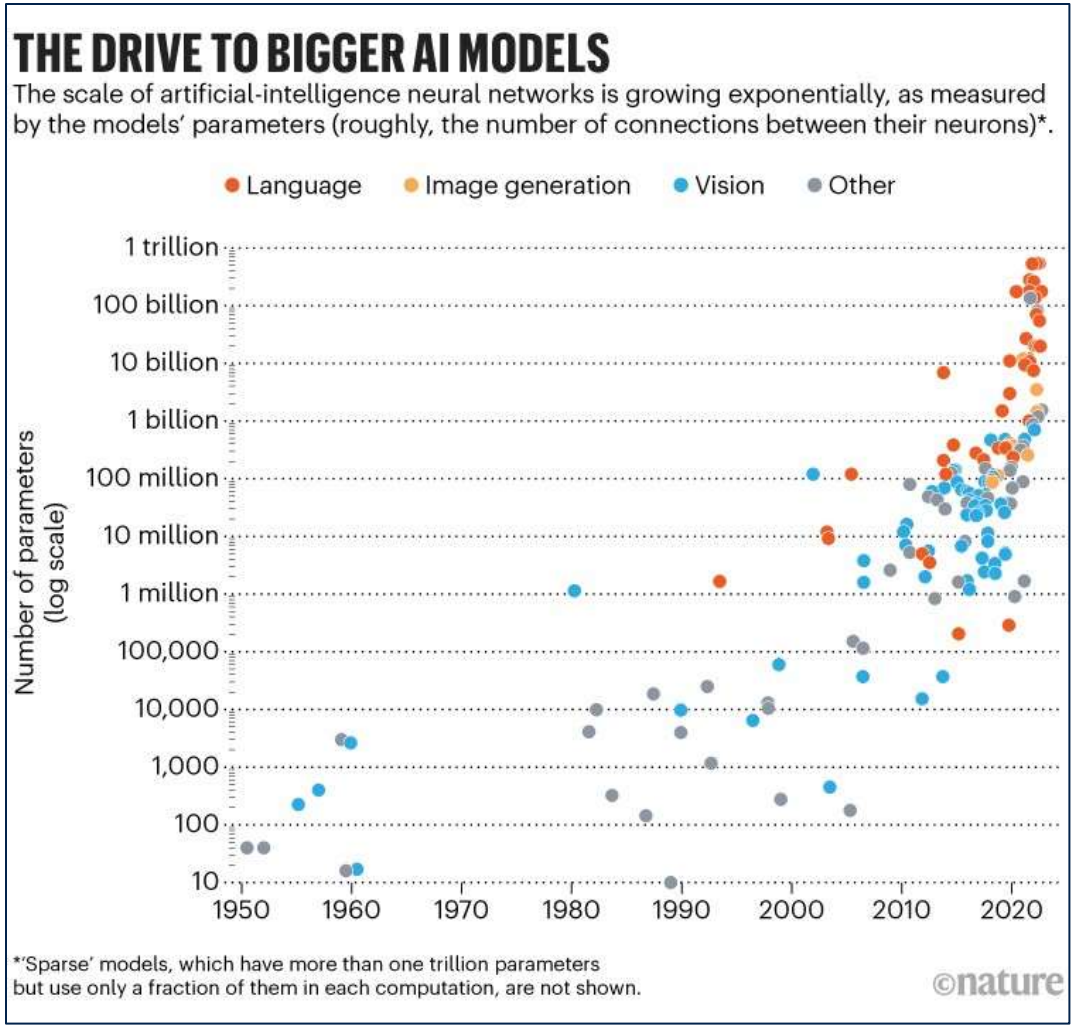


life.augmented

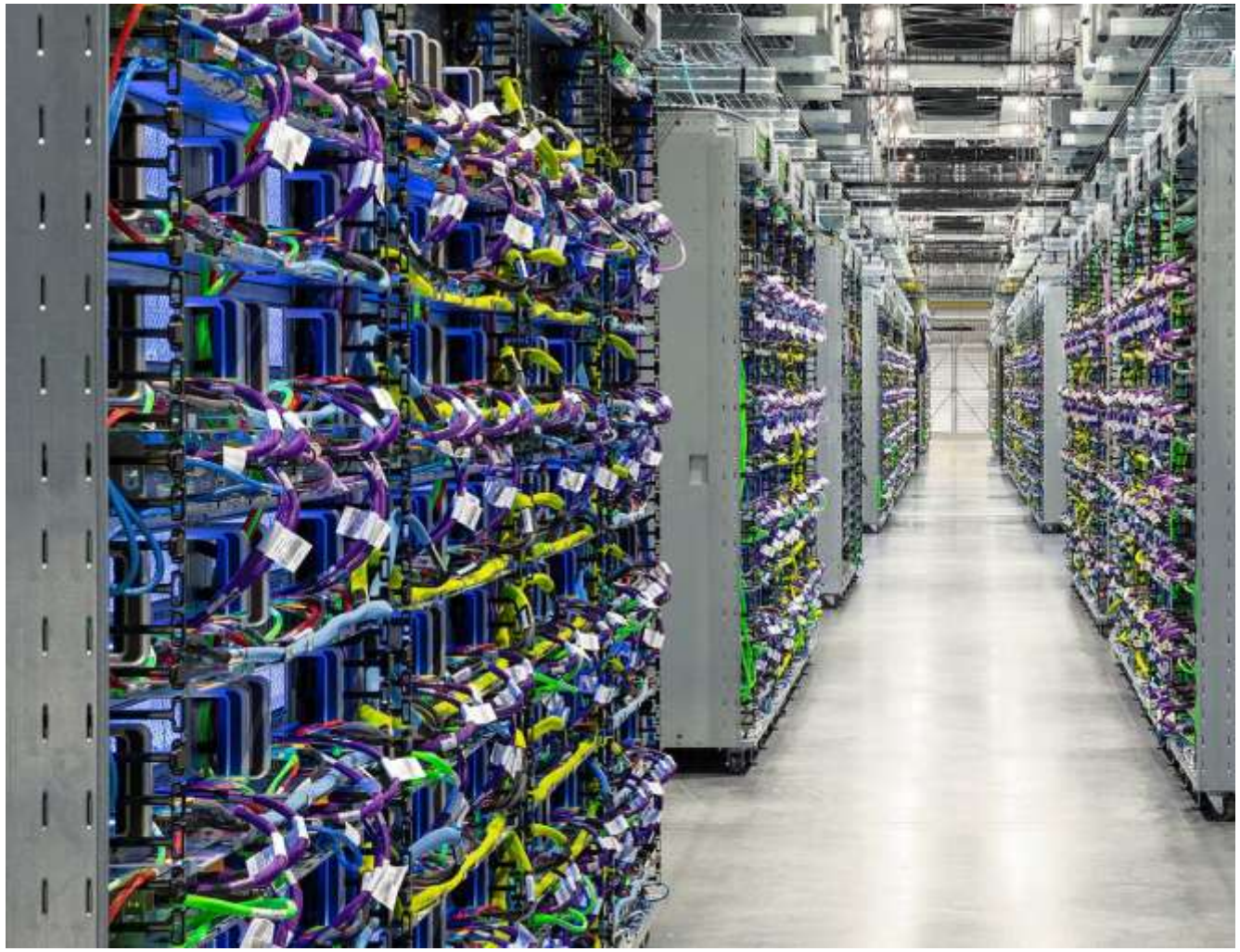
Enabling massive adoption of TinyML models in edge devices thanks to In-Memory Computing

Michele Rossi

Why Edge AI?



https://www.nature.com/articles/d41586-023-00777-9

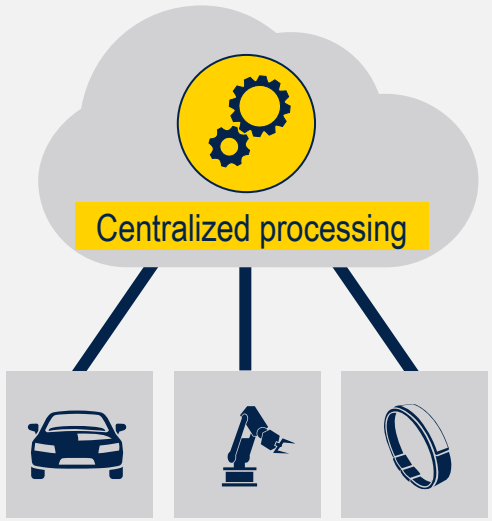


https://cloud.google.com/blog/products/compute/announcing-cloud-tpu-v5e-and-a3-gpus-in-ga

life.augmented

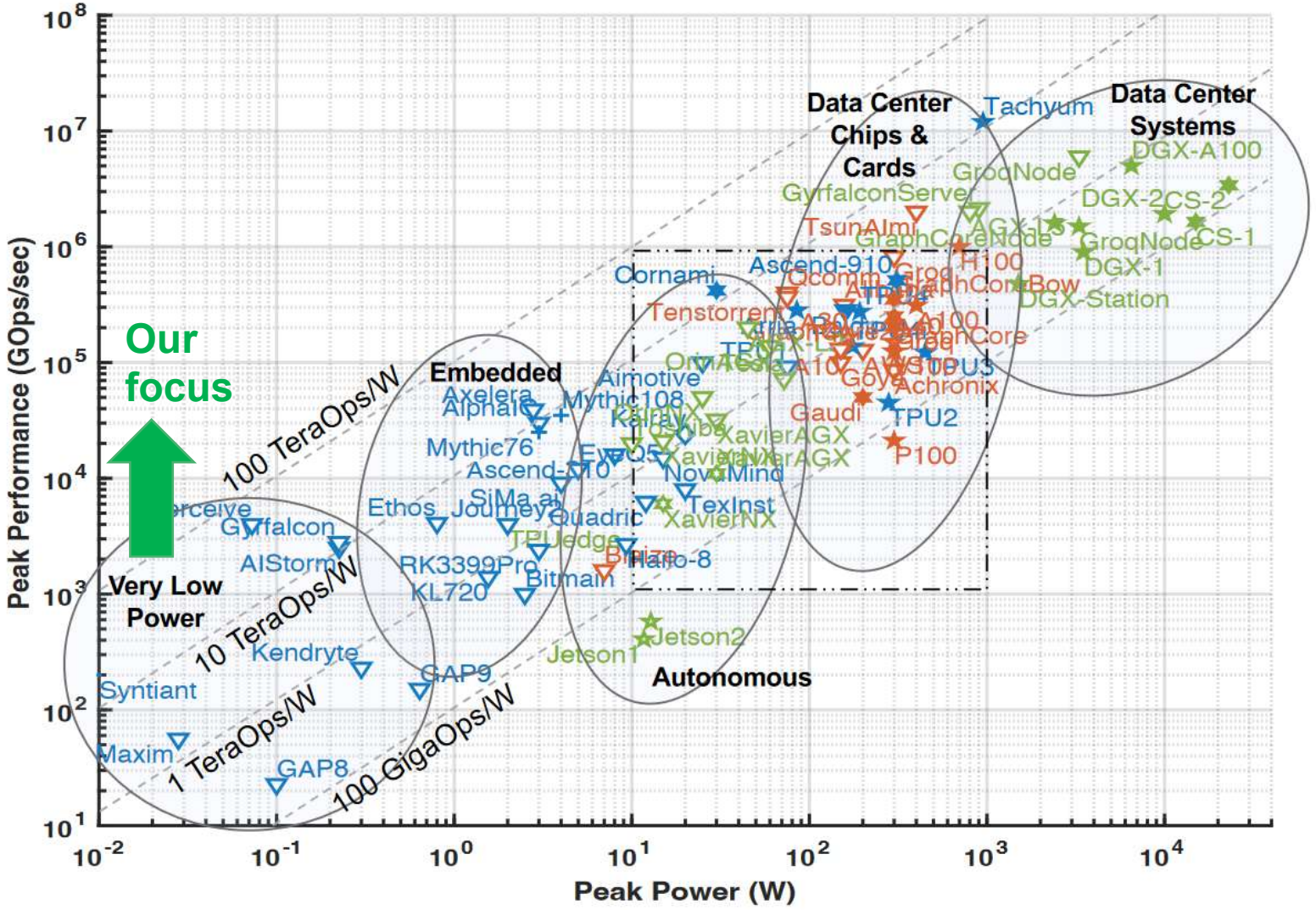
Cloud vs. Edge

Cloud computing	Edge computing
Expensive	Affordable
Inefficient	Efficient
Remote	On-prem
Privacy at risk	Privacy granted
Multiple points of failure	Reliable



life.augmented

From the Cloud to the Edge



Limits of edge devices:

- memory
- area
- performances
- battery-powered

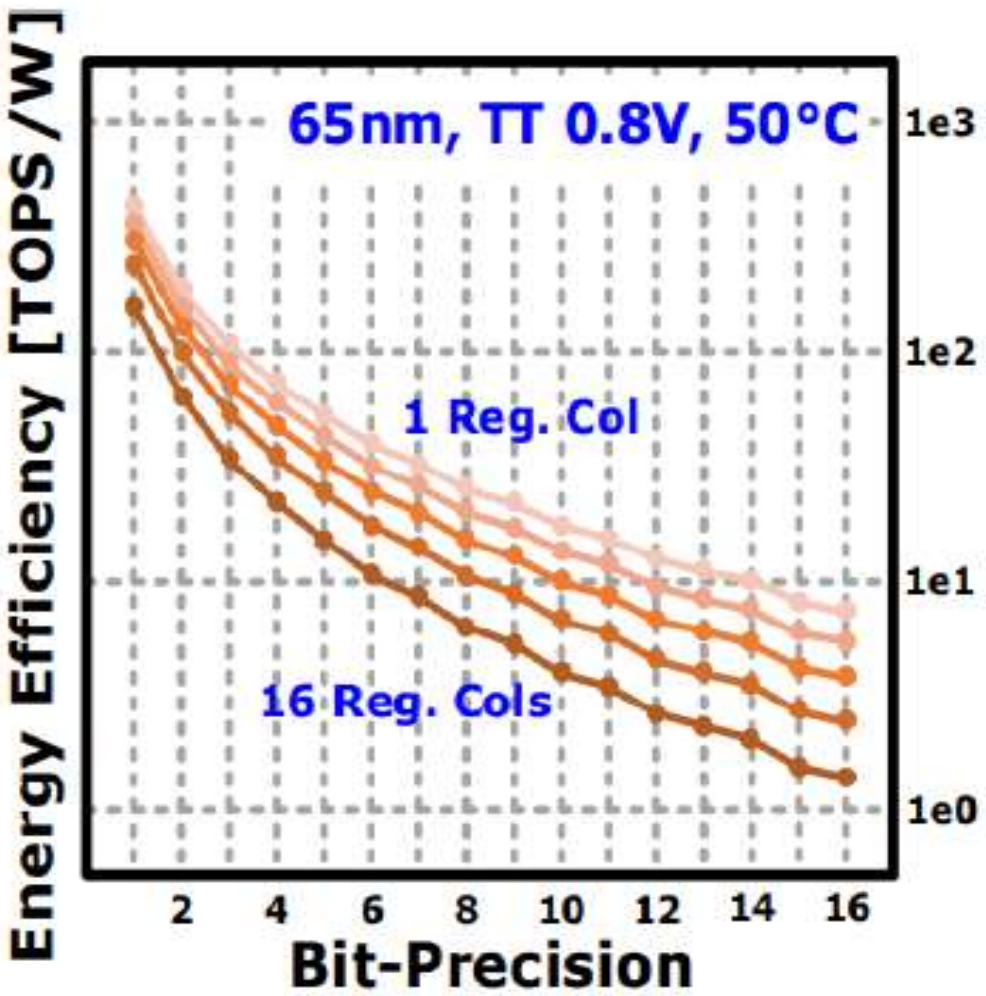
Solutions:

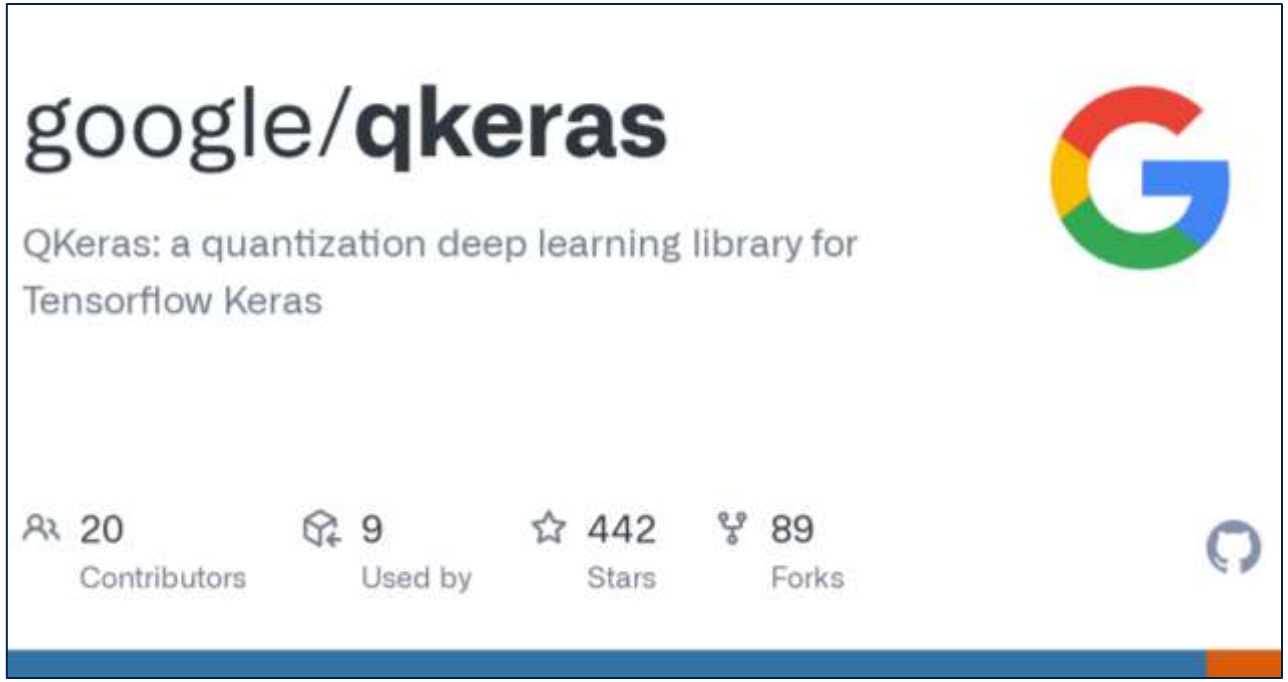
- TinyML
- NPUs
- New technologies

life.augmented

A. Reuther, P. Michaleas, M. Jones, V. Gadepally, S. Samsi, and J. Kepner, "AI and ML Accelerator Survey and Trends", in *2022 IEEE High Performance Extreme Computing Conference (HPEC)*, Sep. 2022

Quantization

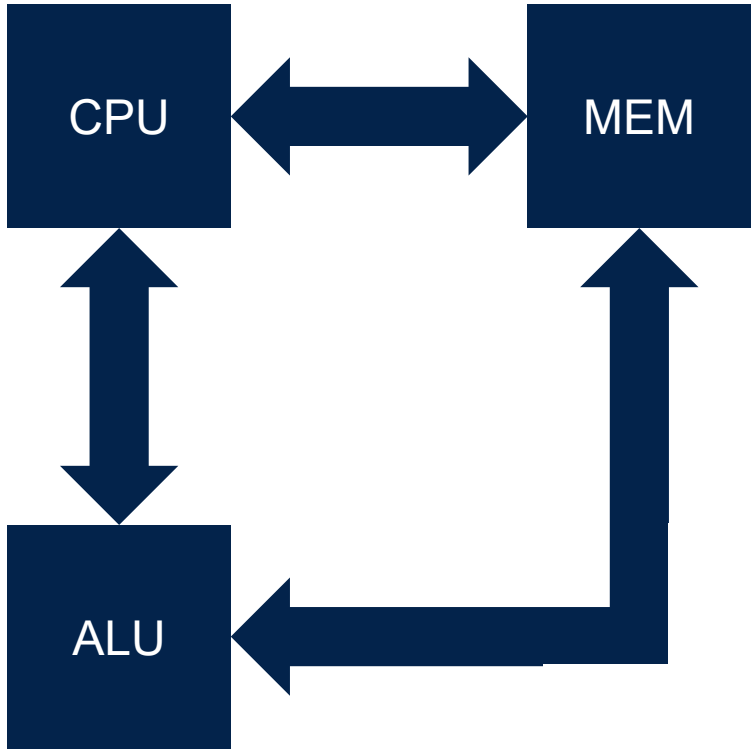




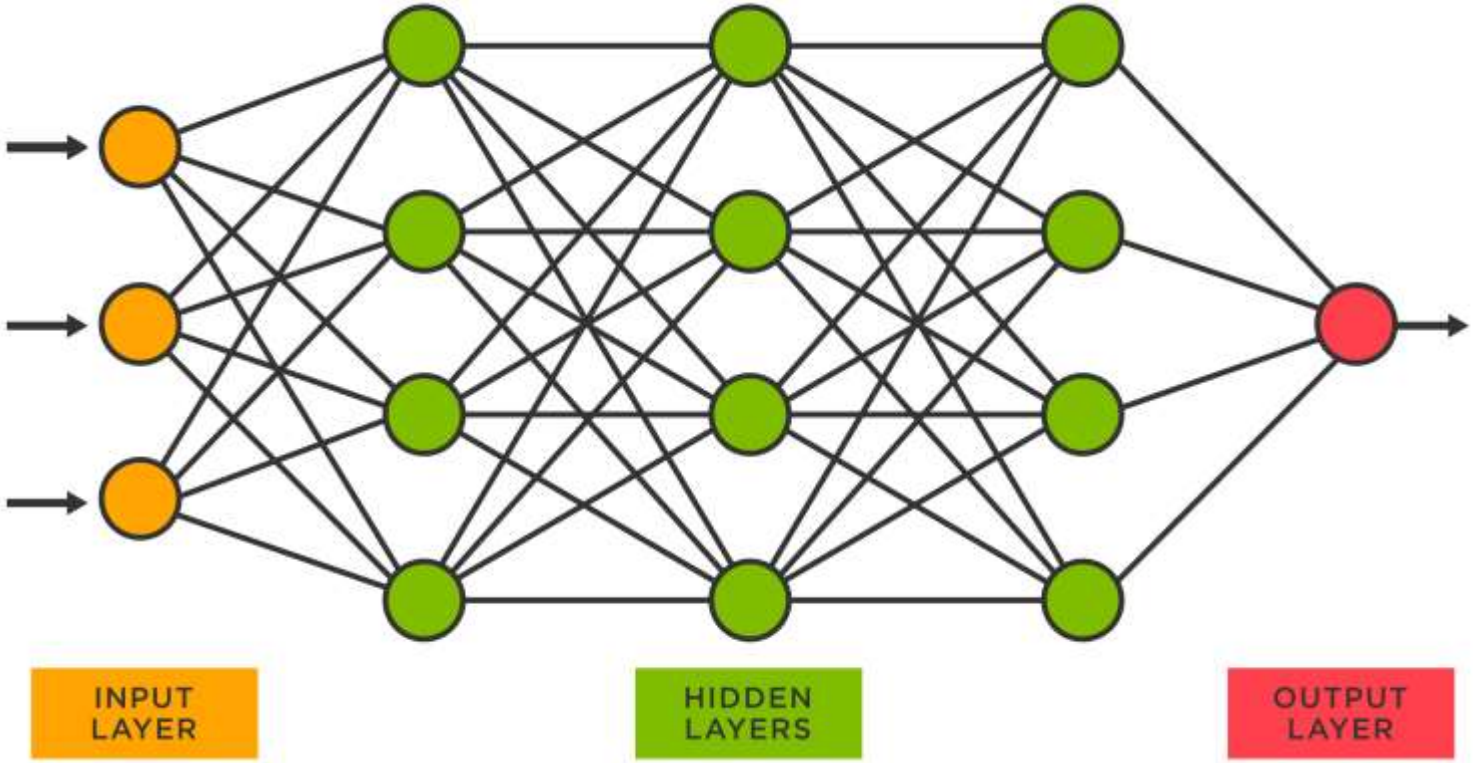
https://github.com/google/qkeras

Von Neumann Bottleneck

Von Neumann Architecture

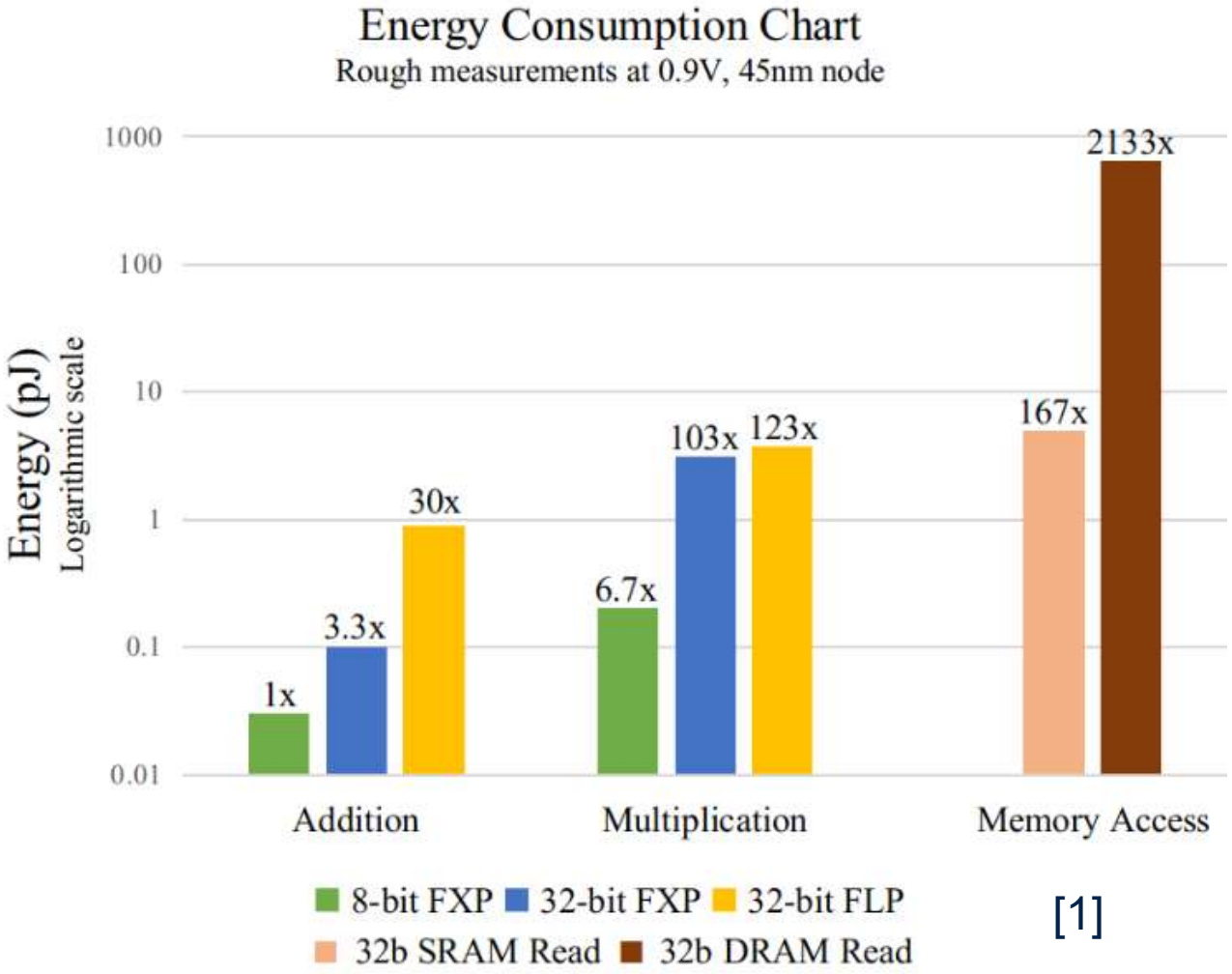


Neural Network Architecture

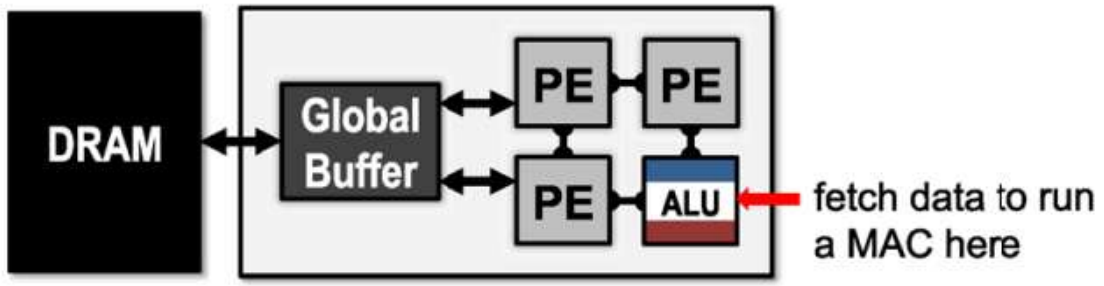


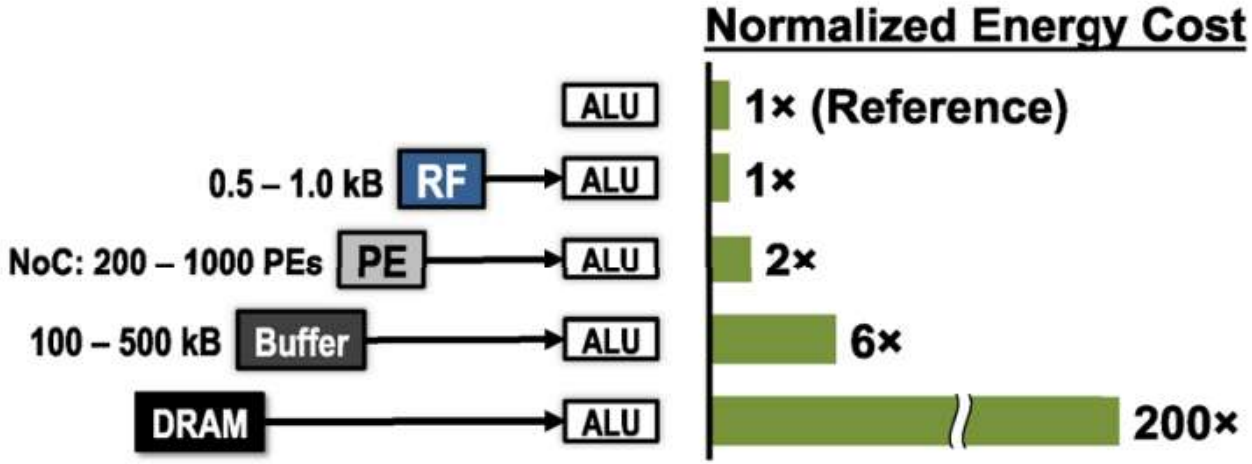
life.augmented

Memory Wall



[1]





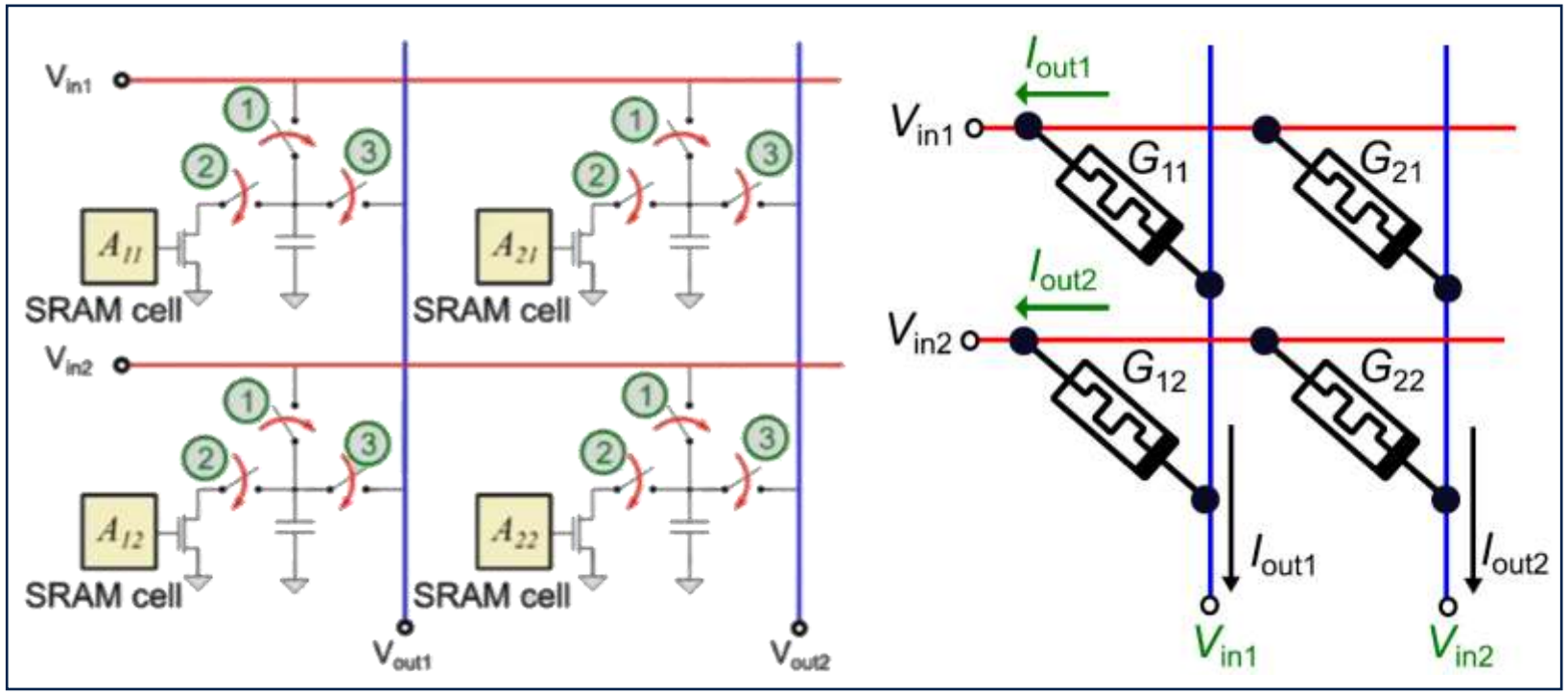
[2]

life.augmented

[1] MIT Eyeriss V2 Architecture, Fig. from "Custom Hardware Architectures for Deep Learning on Portable Devices: A Review", *IEEE Transactions on Neural Networks and Learning Systems*, Zaman et al.
[2] "Efficient processing of deep neural networks: A tutorial and survey", *Proceedings of the IEEE*, 2017, V. Sze et al.

In-Memory Computing

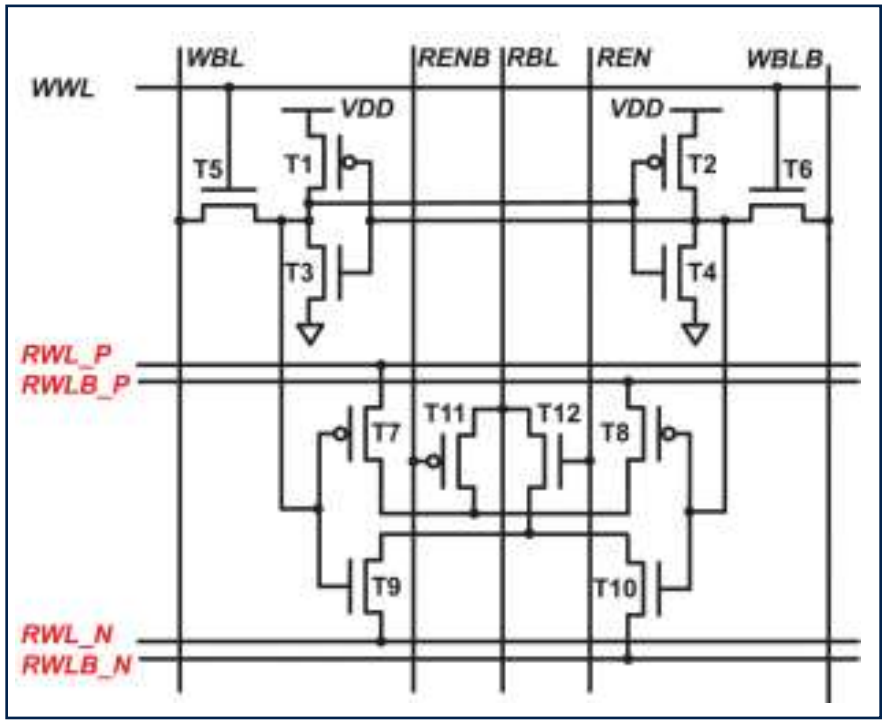
Analog



Capacitor-based [4]

Resistive [4][5]

Digital



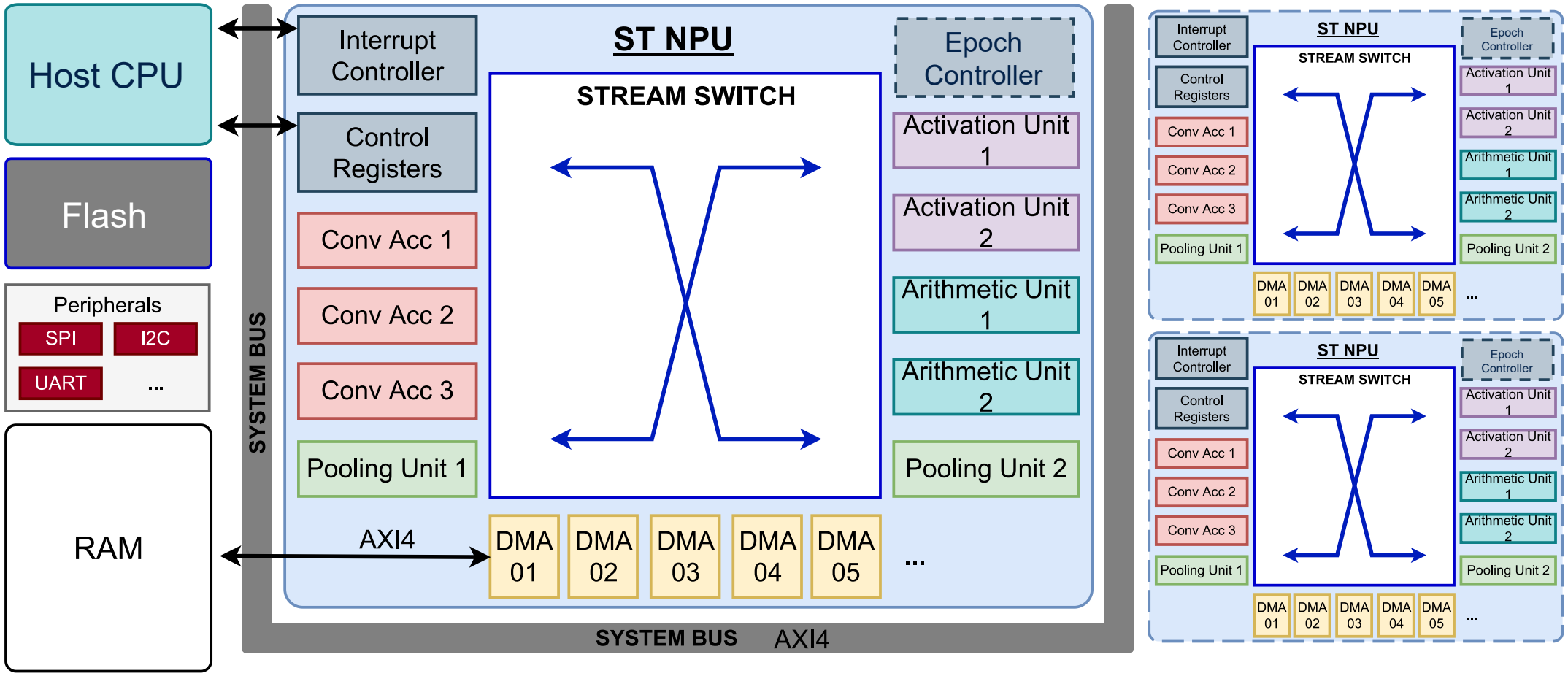
SRAM [6]

[4] "A 64-Tile 2.4-Mb In-Memory-Computing CNN Accelerator Employing Charge-Domain Compute", Valavi et al., in *IEEE Journal of Solid-State Circuits* (2019)

[5] "In memory computing: Memory devices and applications", Abu Sebastian (IBM Research)

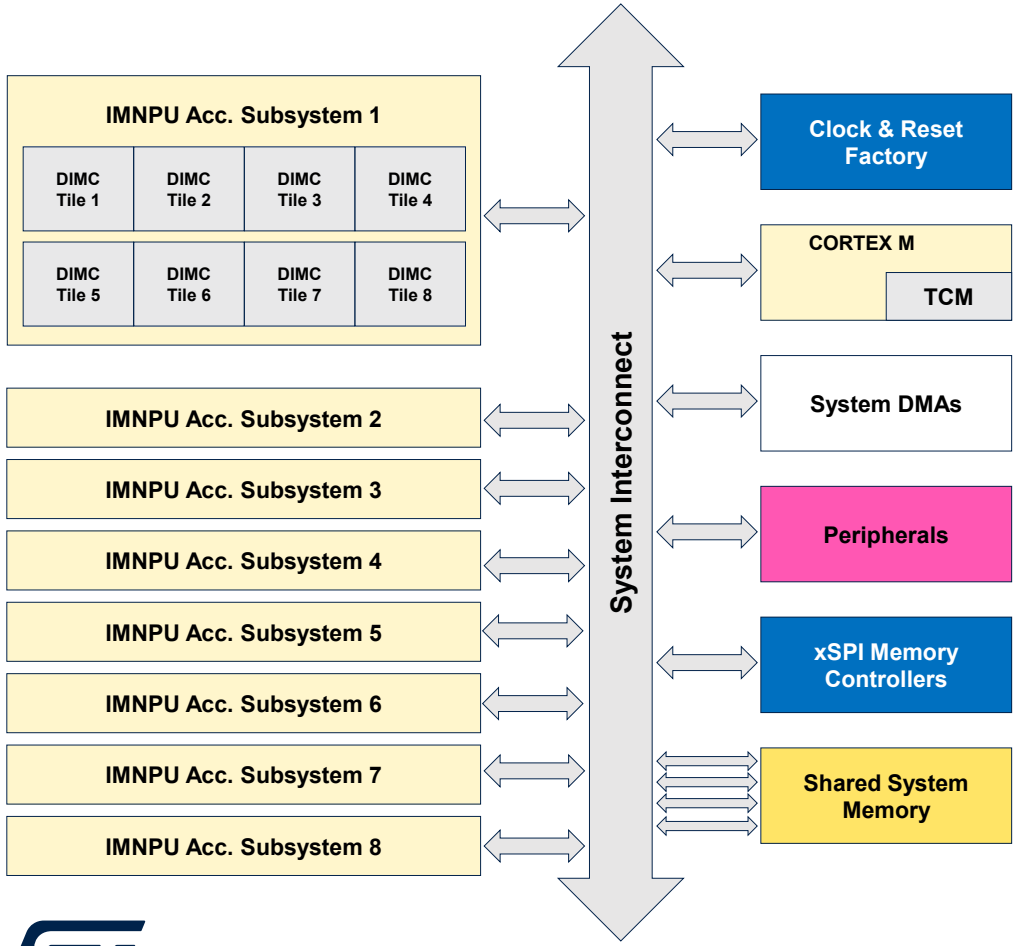
[6] "XNOR-SRAM: In-Memory Computing SRAM Macro for Binary/Ternary Deep Neural Networks", S. Yin et al., in *IEEE Journal of Solid-State Circuits*, vol. 55, June 2020

Architecture Template

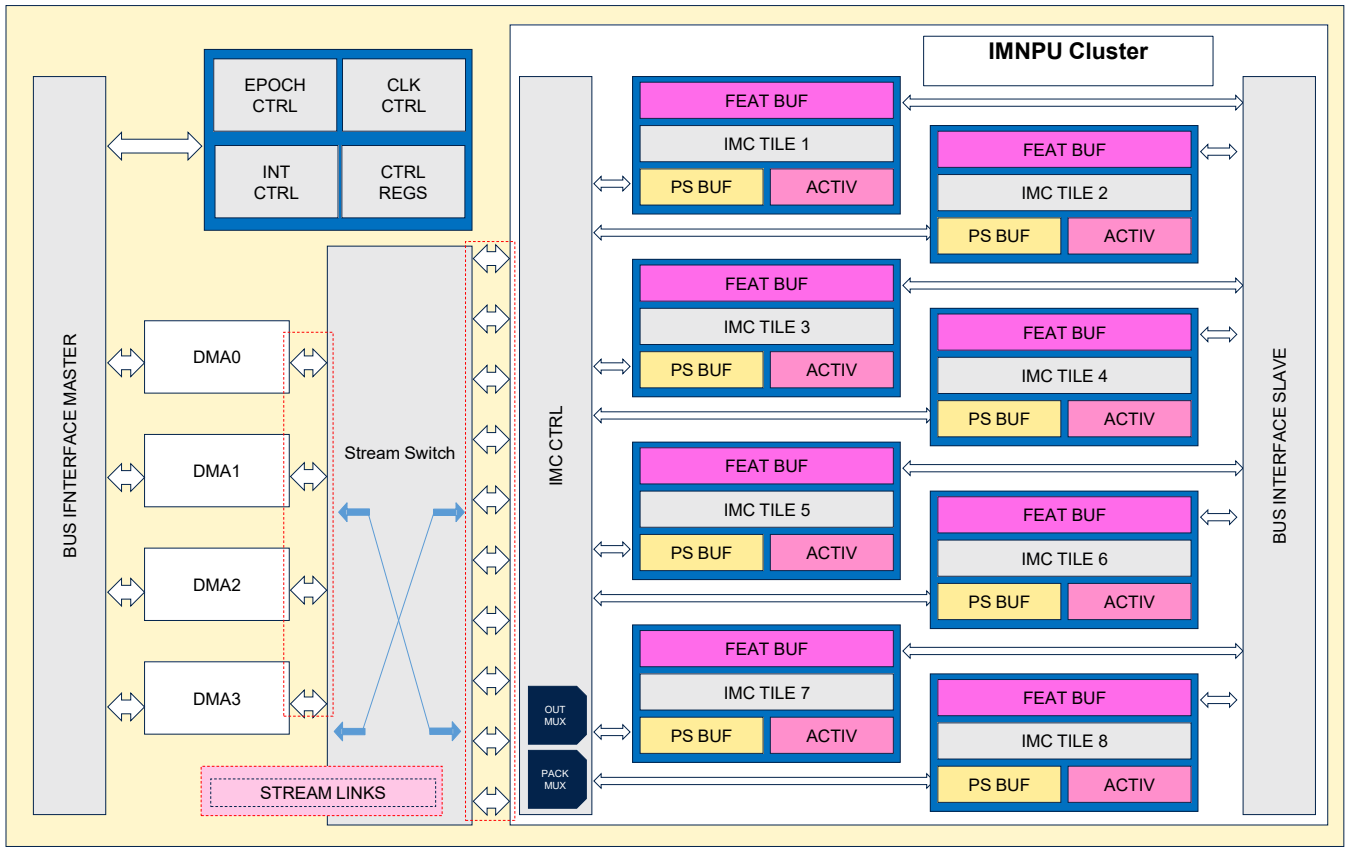


IMNPU System-on-Chip

IMNPU System-on-Chip

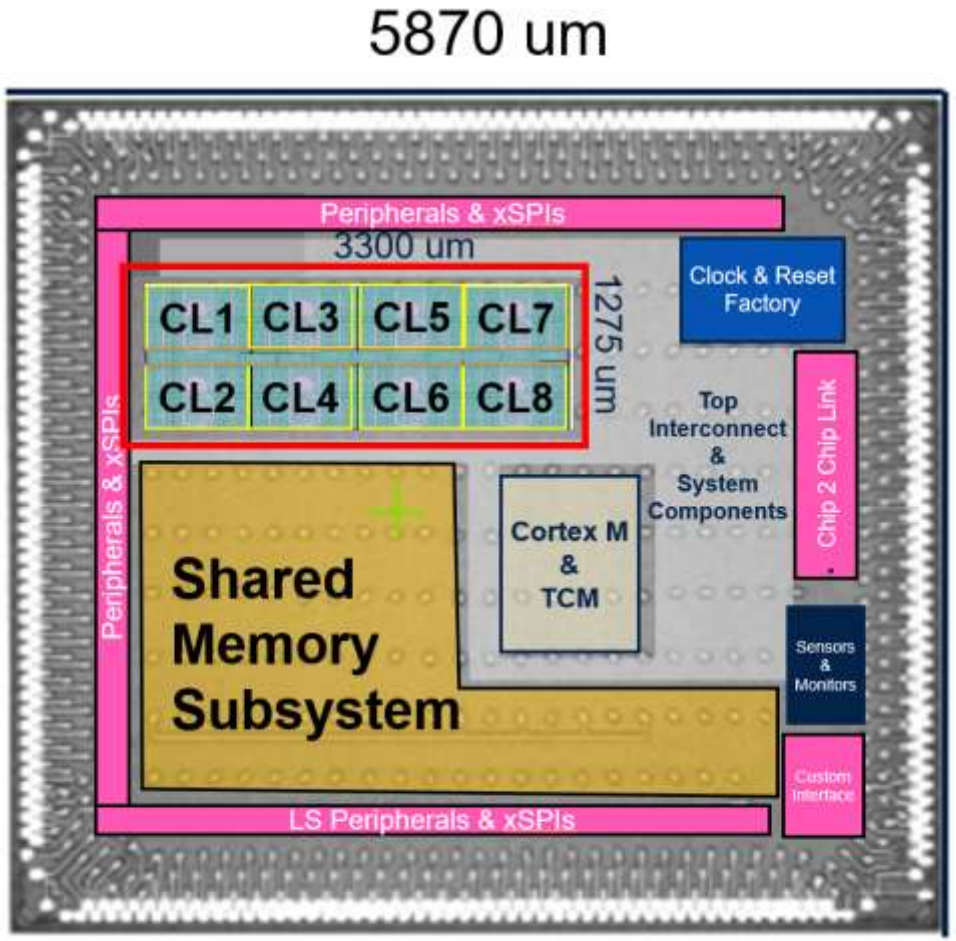


IMNPU Accelerator Subsystem



life.augmented

Prototype Chip



CL : IMNPU Cluster

Technology 18nm FDSOI
Multi-Cluster IMNPU along with system interconnect: 4.2 mm²
Voltage range: 0.525-1.0V, FBB 0-1.5V
IMC Capacity 2 Mb
Computation: Deterministic
Precision Mode: 1-4 bits
229 TOPS (Peak Performance) 1 bit Weight - 1bit Feature 57 TOPS (Peak Performance) 4bit Weight - 4bit Feature
310 TOPS/W (1 bit) 77 TOPS/W (4 bit)
54 TOPS/mm2 (1 bit) 13.6 TOPS/mm2 (4 bit)
CNN, LSTM, RNN

life.augmented

Measurements

Ultra-low-power always-on staged inference application example

Endurance for a battery-operated device for video surveillance, a single cluster is always on clocked @ 10 MHz running a VGG16-like network (as in Fig. 16.7.4) while the rest of the SoC is powered down and selectively activated @ 400 MHz to process a 10x complexity network (e.g., a ResNet152) with a duty cycle of 1 to 100 at 0.525 VDD

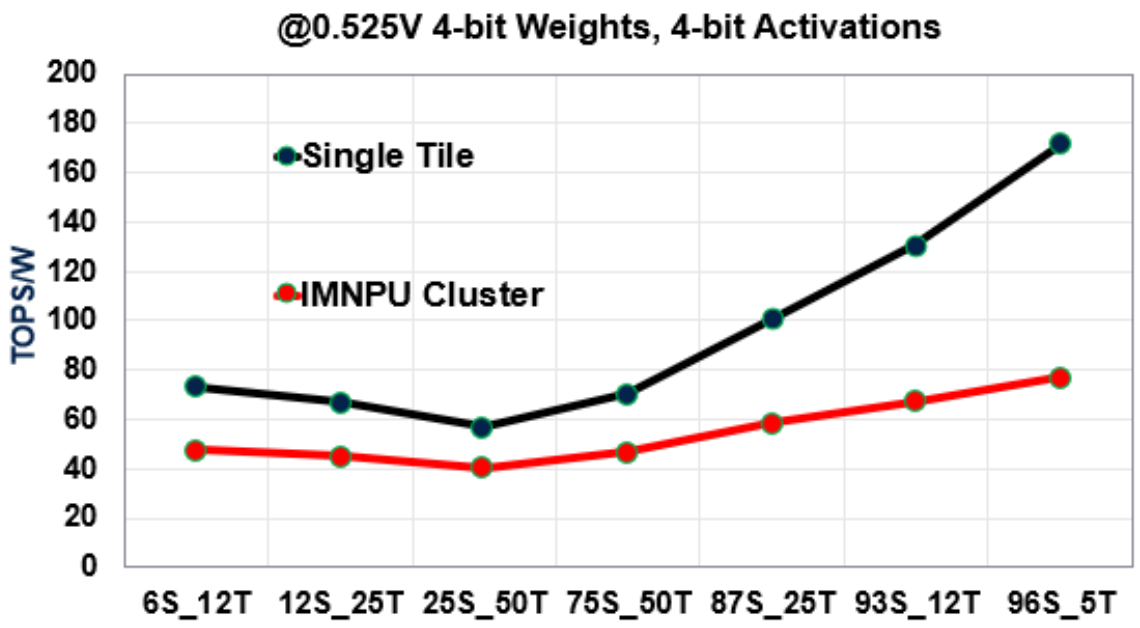
configuration	MACSS/inf	Inf/sec	IMNPU Power	Total Power[1]	Battery endurance[2] (1/100 duty cycle)
1 cluster @ 10 MHz, 0.0V FBB	1.25 GOPS	10	267 uW	567 uW	363 days
8 clusters @ 400MHz, 0.3V FBB	12.5 GOPS	30	8.0 mW	12.0 mW	

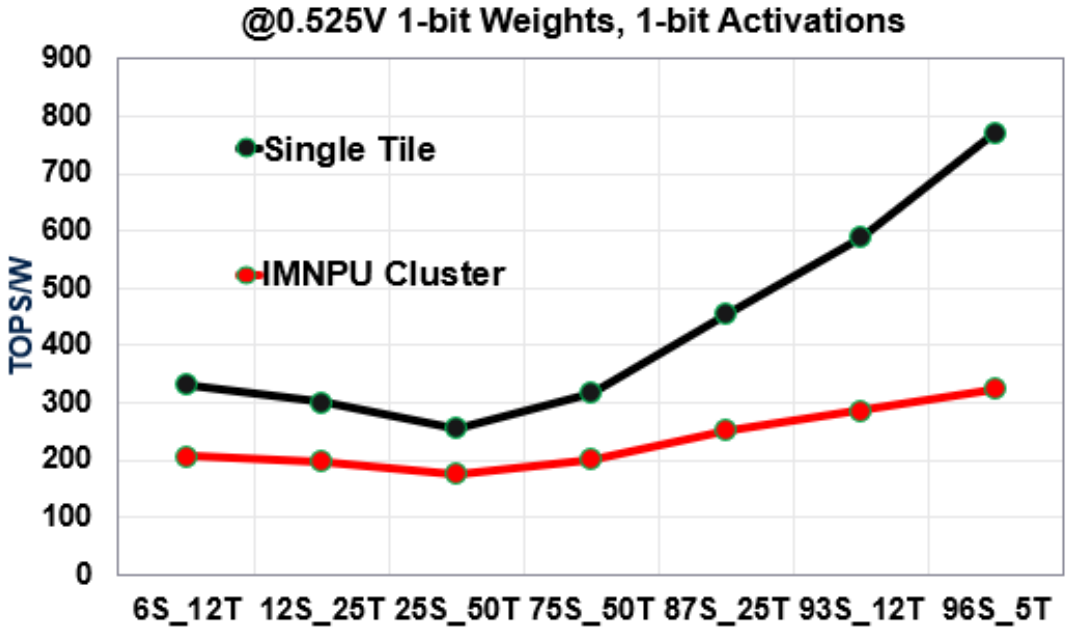
(1) Estimated power includes a portion of shared memory, IOs, clock, and external sensor interface
(2) 6000 mA/h battery capacity assumed (e.g., 2 AA 1.5v batteries)

life.augmented

Measurements

Energy efficiency for tile and IMNPU





(S)parsity (% of zeros) & (T)ransitions (% bit toggling across adjacent rows)

Charts show the trend of diminishing return of TOPS/W gains for a single DIMC instance not translating to proportiona TOPS/W improvement at the IMNPU cluster level due to the scalar processing and data movement overheads.

life.augmented

Our technology starts with You

Find out more at www.st.com

life.augmented