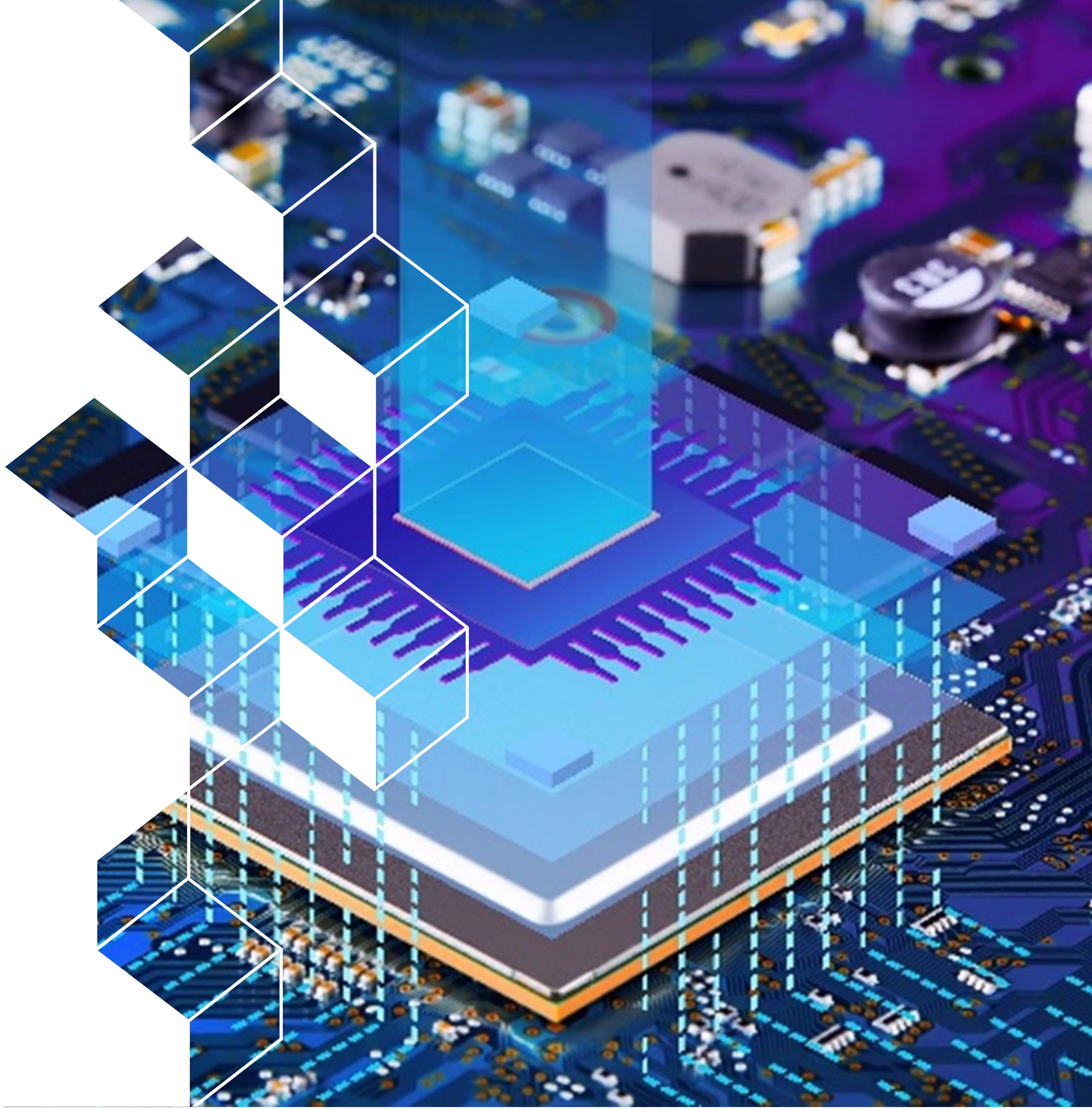
Memory-centric computing

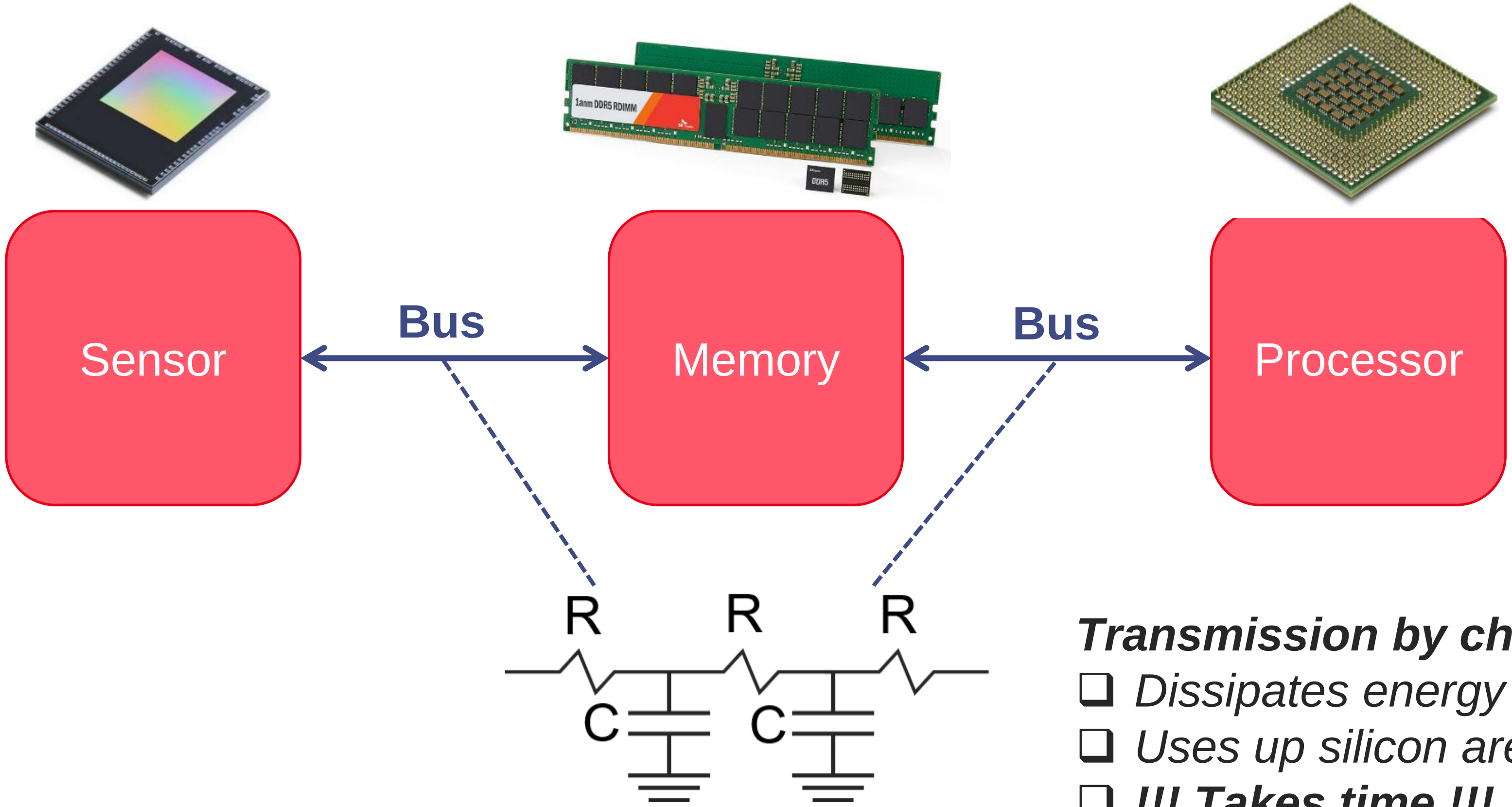
Thomas DALGATY

CEA Grenoble

thomas.dalgaty@cea.fr



Computing’s data transfer problem



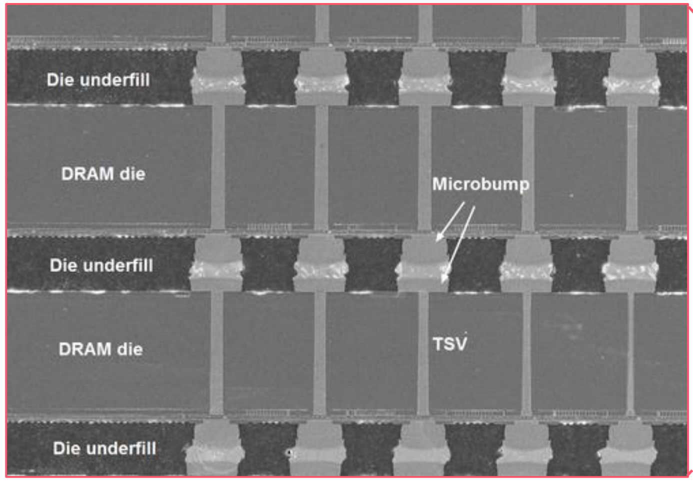
Transmission by charging metal wires @ GHz

- ❑ *Dissipates energy*
- ❑ *Uses up silicon area*
- ❑ ***!!! Takes time !!!***

Lets talk about GPUs

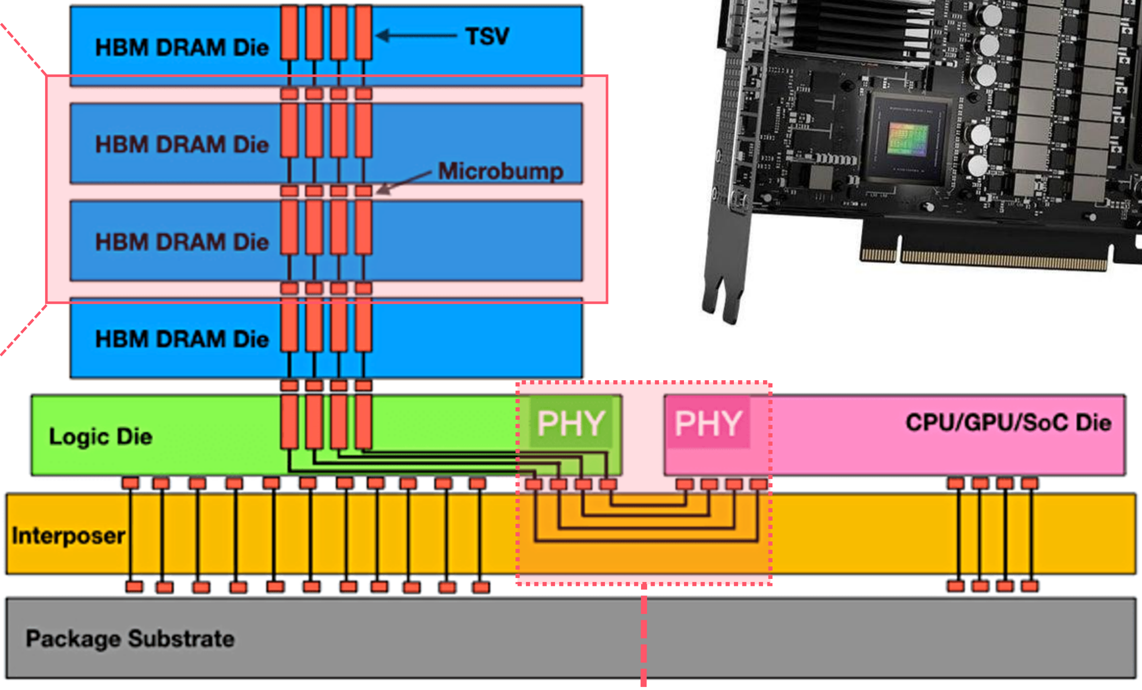
NVIDIA H100 PCIe GPU

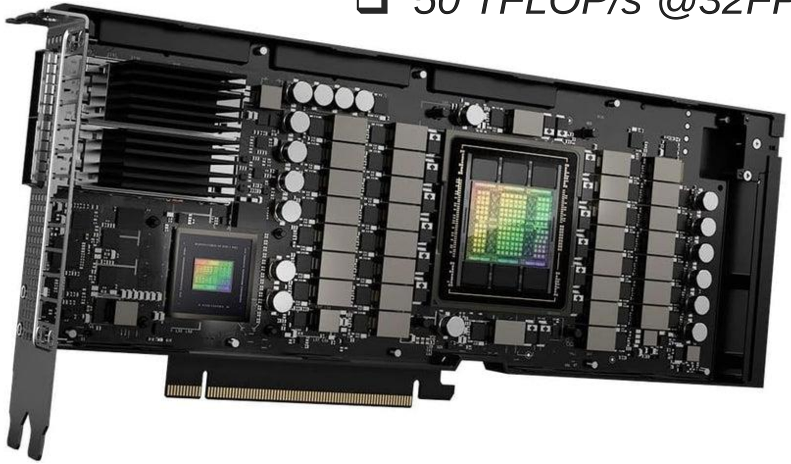
- ❑ *80GB memory*
- • *only 50MB is “on-chip”*
- ❑ *15k CUDA cores*
- ❑ *50 TFLOP/s @32FP*



https://www.eetimes.com/hats-off-to-hynix-inside-1st-high-bandwidth-memory/3/

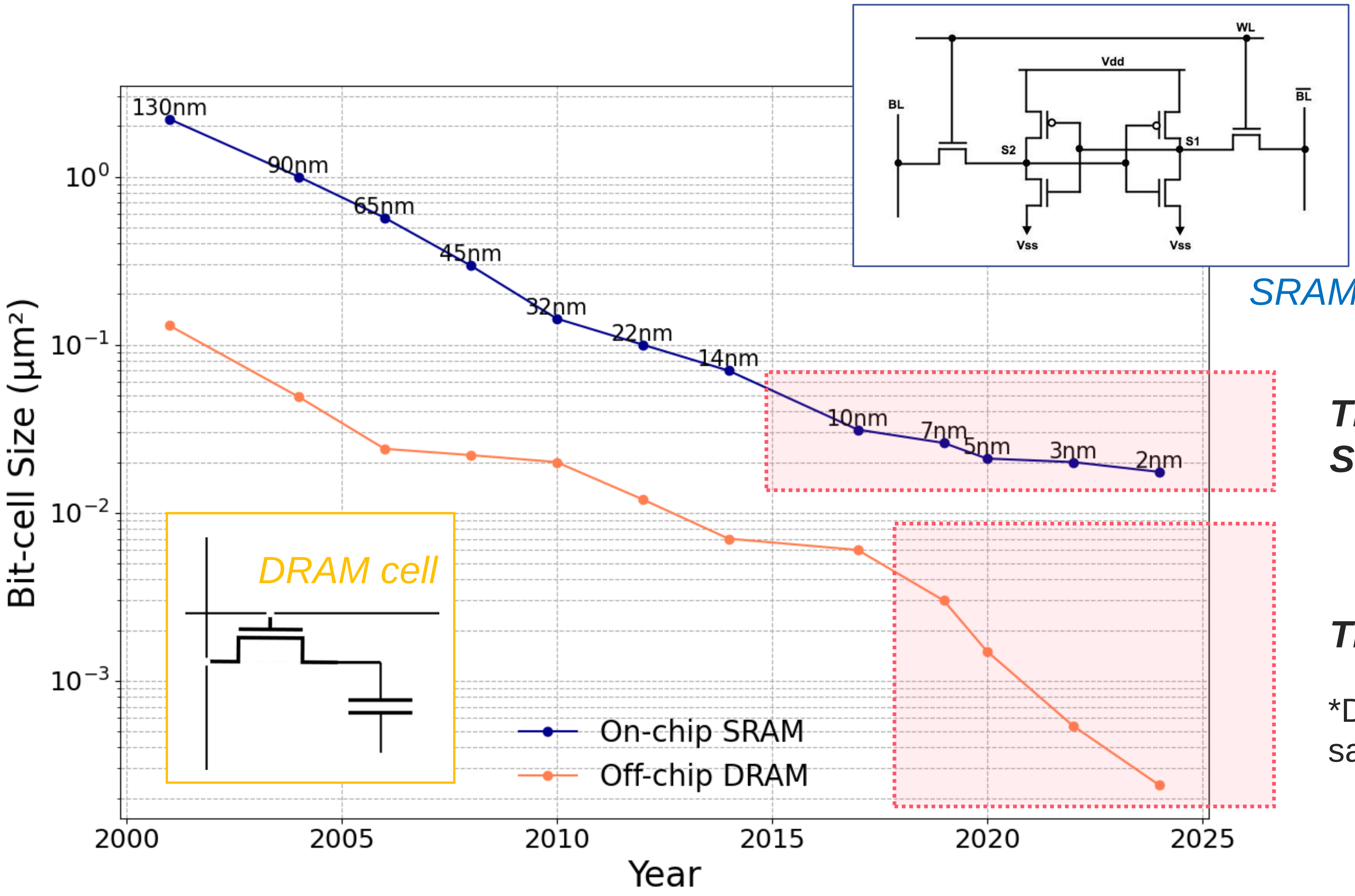
**very old HBM1 TEM*





***3TB/s** memory bandwidth (>5k metal wires)*

On- and off-chip memory scaling



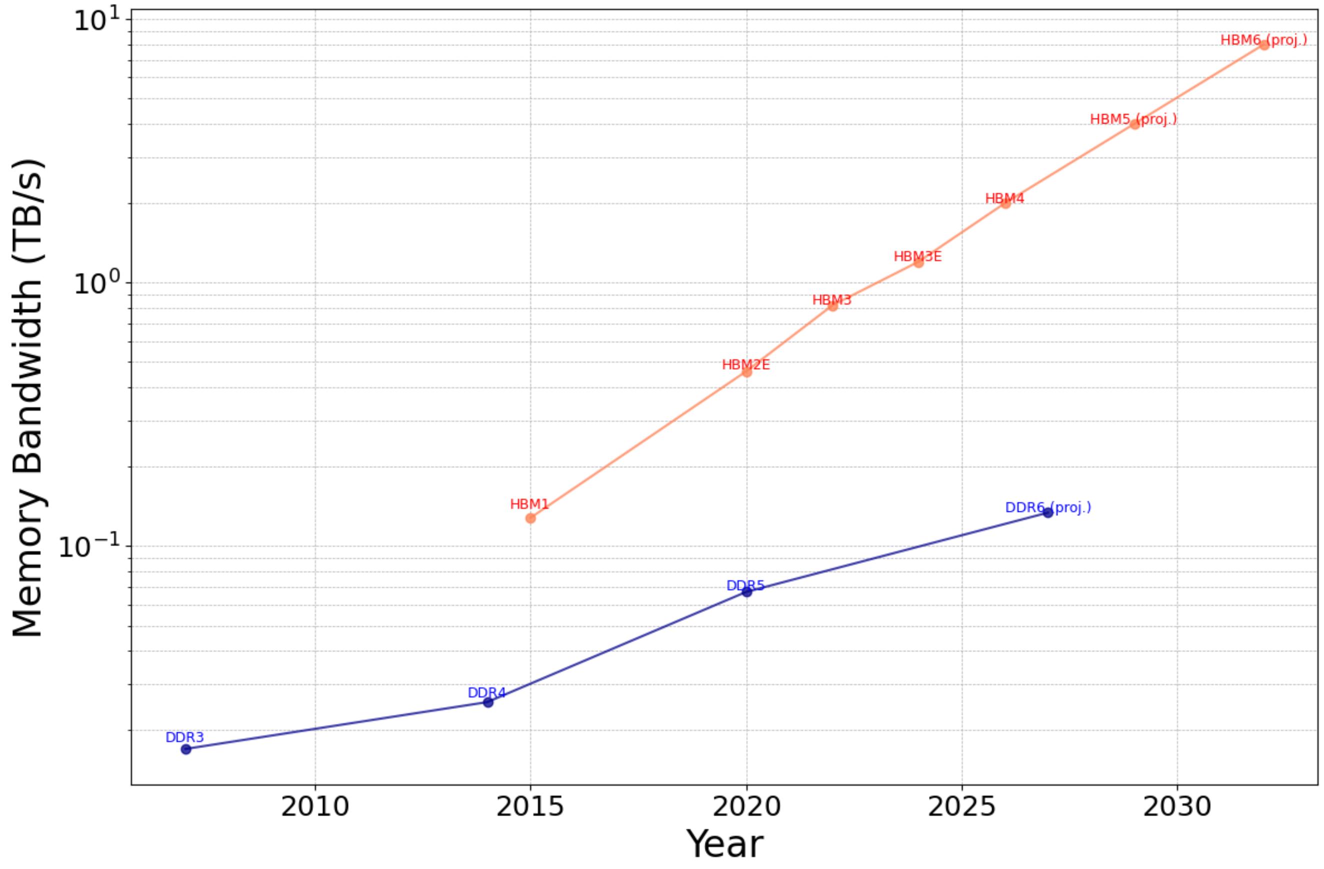
SRAM cell

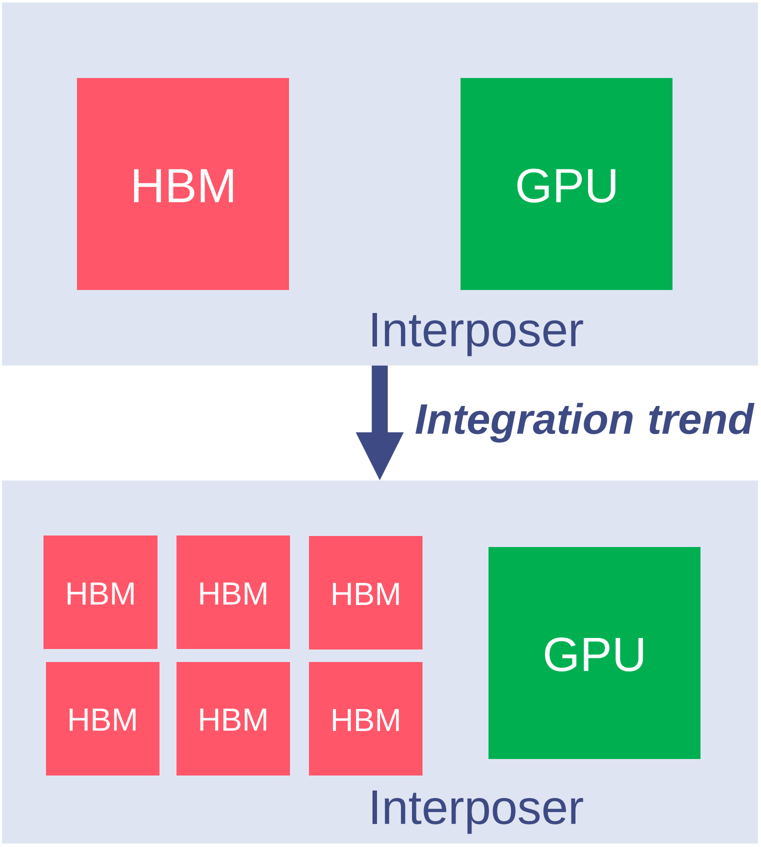
The end of SRAM scaling

The HBM era*

*DRAM bit-cell not in same process as GPU

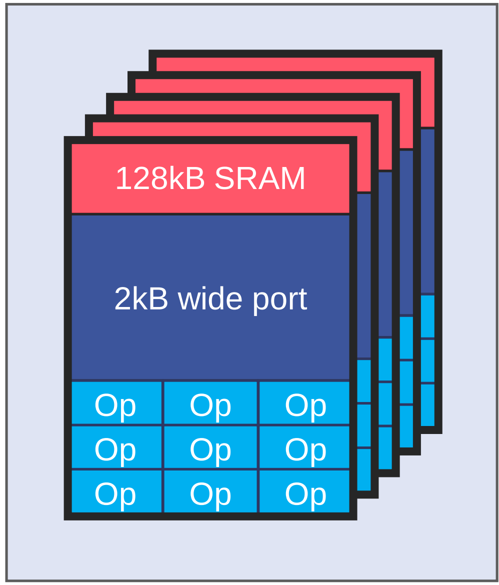
HBM scaling



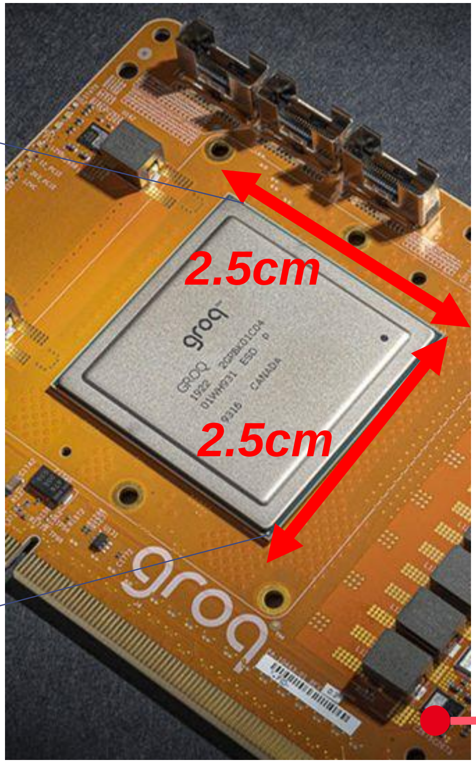


Trend towards multi-stacked HBM for ~40TB/s bandwidth per GPU by 2030

“You can’t bring all memory on-chip”



❑ *Ultra-wide SRAM ports enable high bandwidth between SRAM cuts and operator array*



❑ *220MB on-chip SRAM per 650mm^2 14nm FinFET chip*

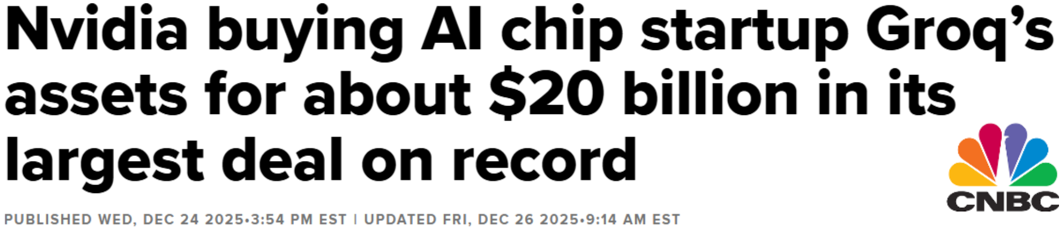
❑ *576 chips per cluster => 128GB of memory*

❑ ***80TB/s** memory bandwidth per chip*

❑ *Up to **46PB/s*** of bandwidth per cluster*

https://groq.com/

****theoretical limit, not accounting for inter-chip communication bottleneck***

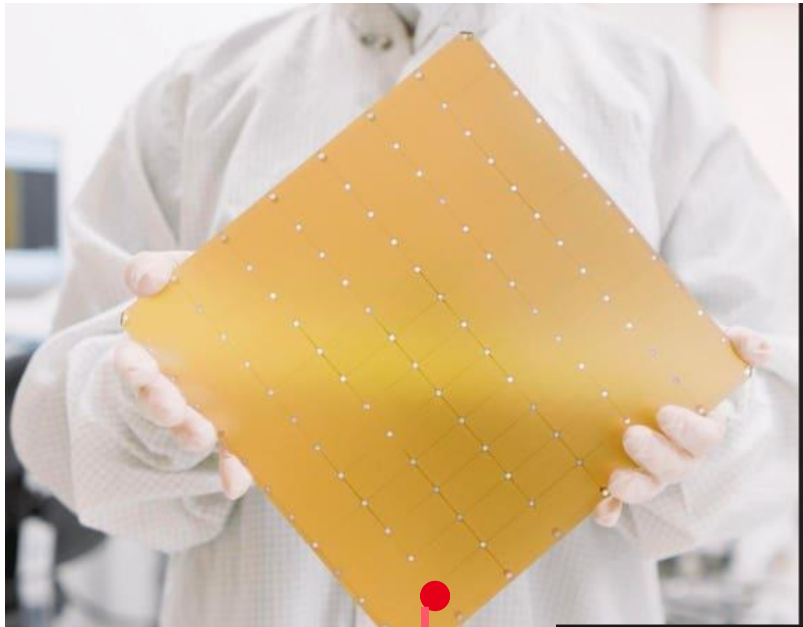


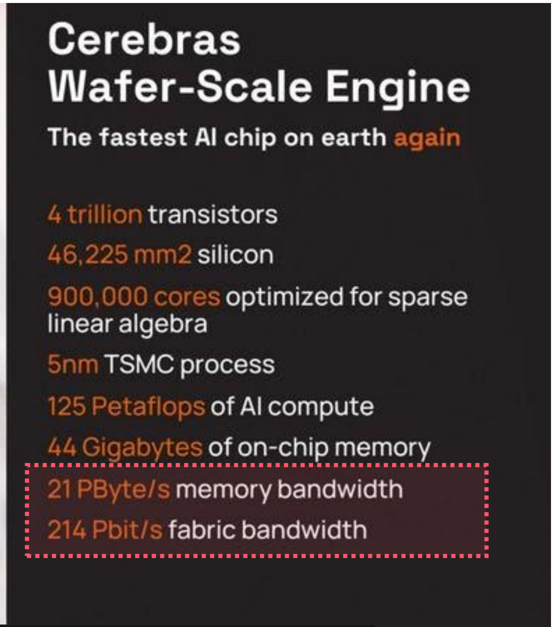
Nvidia buying AI chip startup Groq’s assets for about $20 billion in its largest deal on record

PUBLISHED WED, DEC 24 2025·3:54 PM EST | UPDATED FRI, DEC 26 2025·9:14 AM EST

https://www.cnbc.com/2025/12/24/nvidia-buying-ai-chip-startup-groq-for-about-20-billion-biggest-deal.html

“You can’t make a wafer-scale processor”





Cerebras Wafer-Scale Engine

The fastest AI chip on earth again

4 trillion transistors

46,225 mm2 silicon

900,000 cores optimized for sparse linear algebra

5nm TSMC process

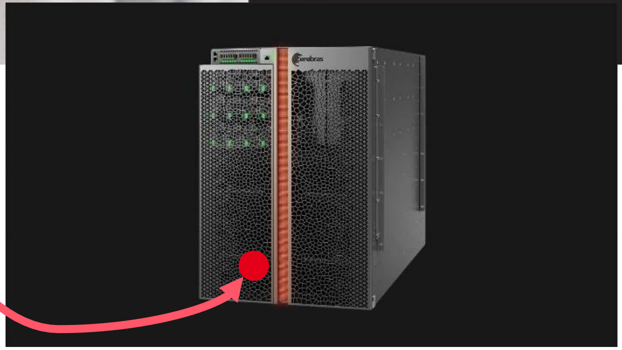
125 Petaflops of AI compute

44 Gigabytes of on-chip memory

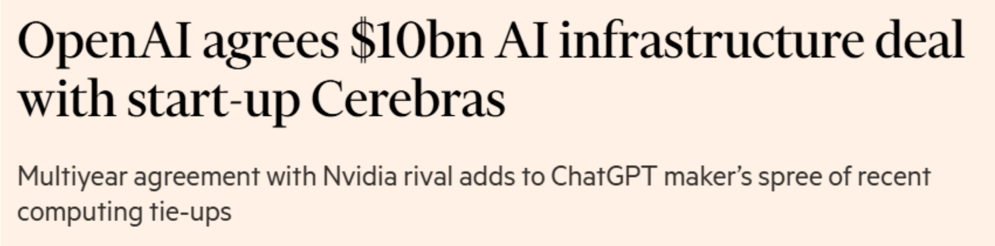
21 PByte/s memory bandwidth

214 Pbit/s fabric bandwidth

https://www.cerebras.ai/chip



- ❑ *192 wafers per GS2*
- ❑ ***4 EXA B/s** bandwidth*



OpenAI agrees $10bn AI infrastructure deal with start-up Cerebras

Multiyear agreement with Nvidia rival adds to ChatGPT maker’s spree of recent computing tie-ups

https://www.ft.com/content/2f7e17fb-5892-4a41-b612-8e384f3dc131

Groq and Cerebras have shown that on-chip memory still has a role to play in the race for tomorrow’s compute

Memory-centric computing

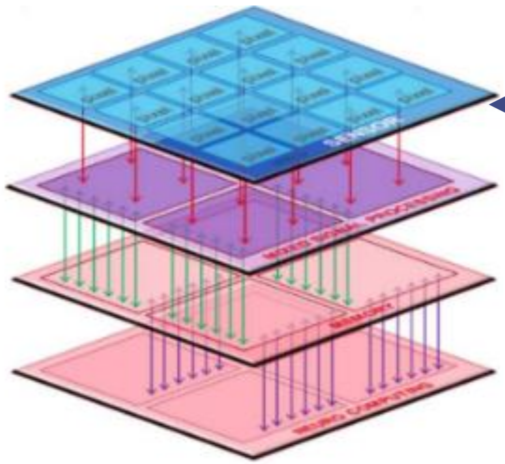
1. Memory-frugal algorithmic design (beyond quantization)

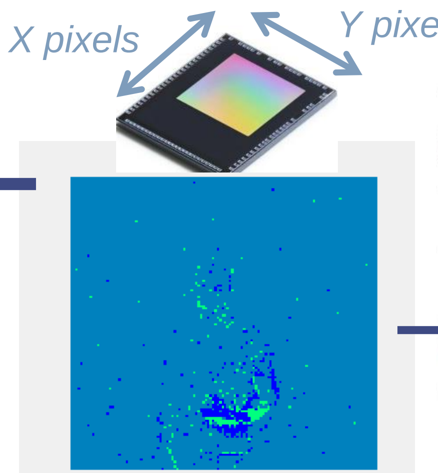
2. New on-chip memory concepts (replacing SRAM)

3. In-memory computing (“neuromorphic” computing)

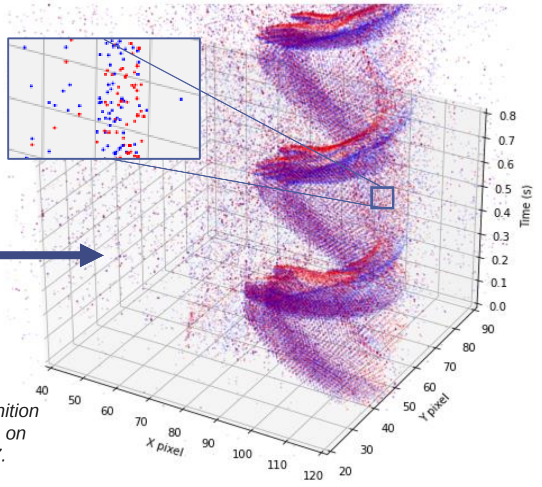
4. The next big bet for memory-centric computing ?

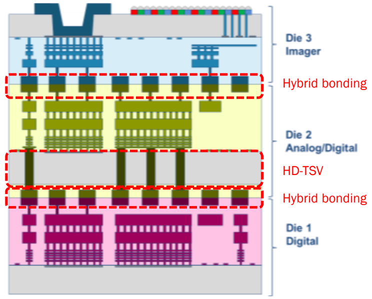
3D near-sensor computing with event-graphs

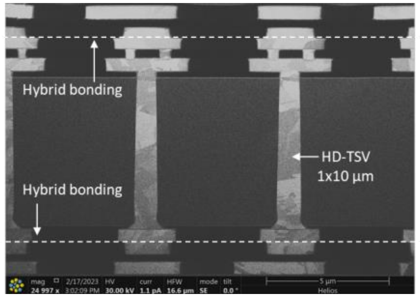




A low power, fully event-based gesture recognition system." Proceedings of the IEEE conference on computer vision and pattern recognition. 2017.



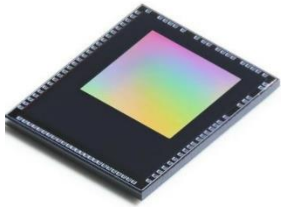




- *320x320 array of 5µm pixels* ***=> ~1mm^2 chip in 22nm***
- ***1.25MB*** *if the chip is only memory*

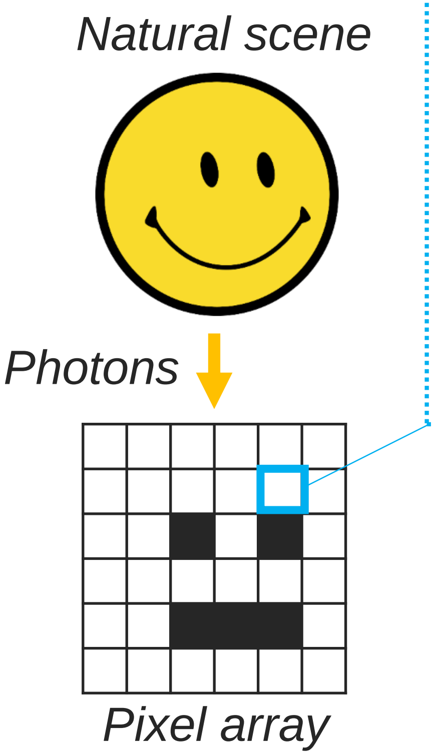
- *State of the art uses CNNs between 10 -> 100MB*
- *How to treat data fast (hundreds of µs) with so little memory ?*

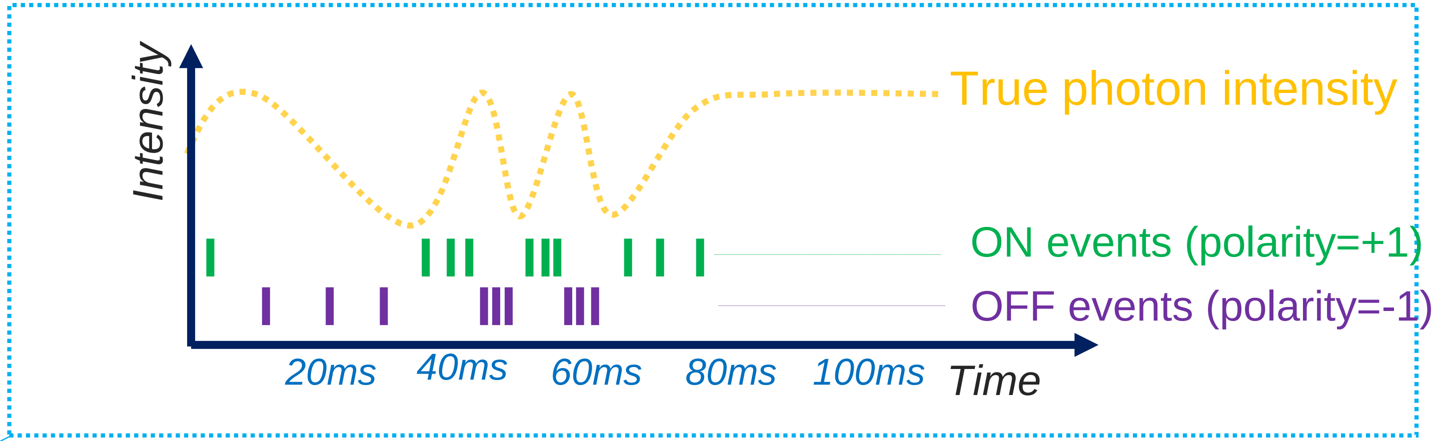
Introduction to event-based sensing



Prophesee GenX 320 Event-camera

***Event-based** camera principle*





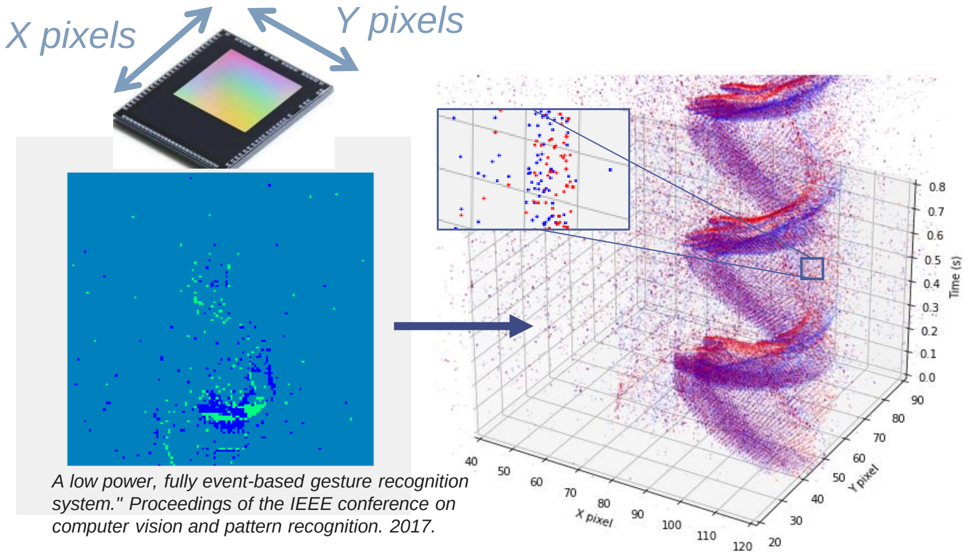
Record relative change events (temporal contrast)

+ Captures the fine temporal detail of motion

+ Compression by eliminating redundant information

Event-data format

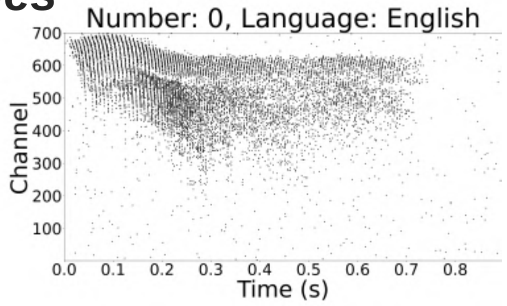
Event-based cameras



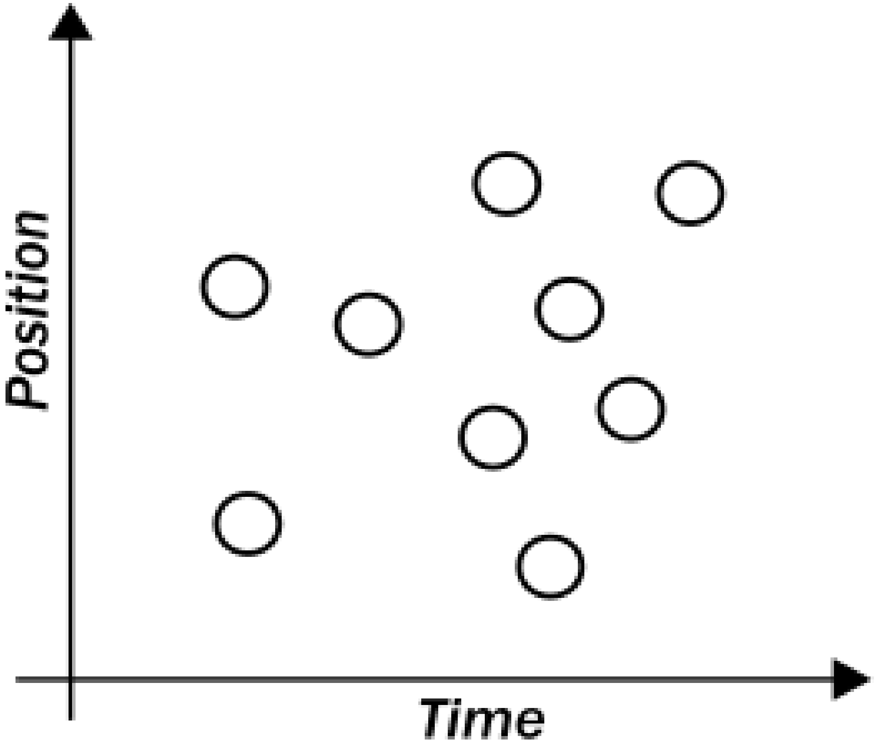
A low power, fully event-based gesture recognition system." Proceedings of the IEEE conference on computer vision and pattern recognition. 2017.

Event-based microphones

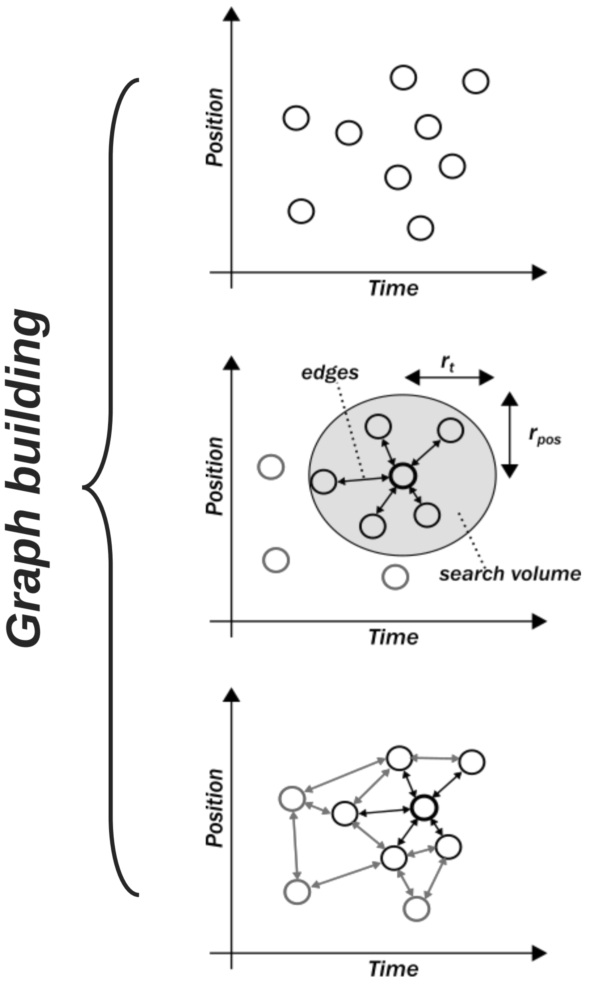
https://inilabs.com/products/dynamic-audio-sensor/

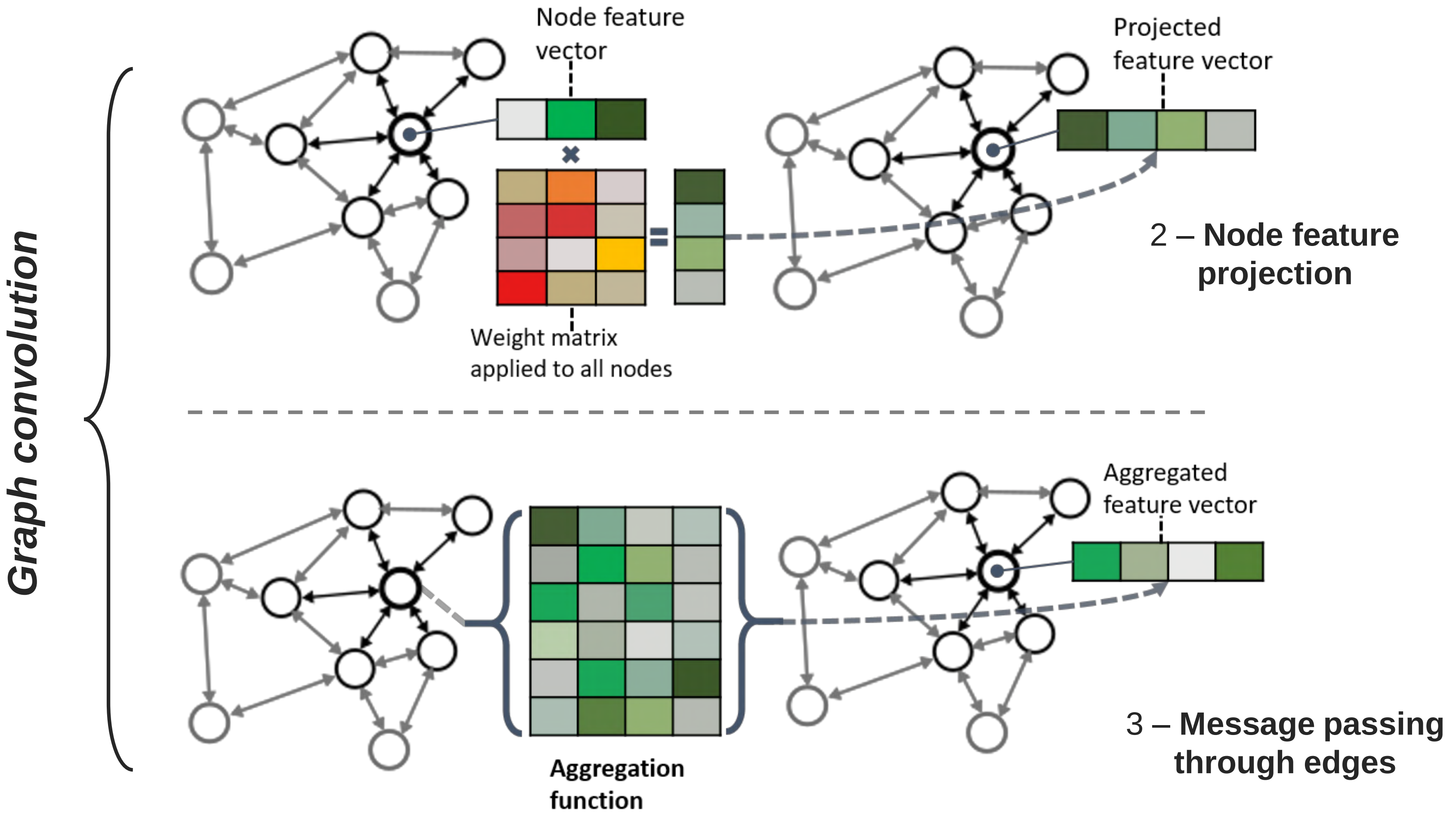


"Hardware-accelerated event-graph neural networks for low-latency time-series classification on soc fpga." ARC: Springer Nature Switzerland, 2025.



Idea : event-graph neural networks

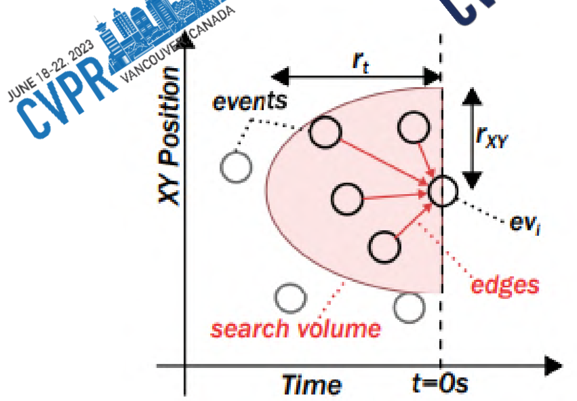



Graph convolution
Node feature vector
Weight matrix applied to all nodes
Projected feature vector
2 – Node feature projection
Aggregation function
Aggregated feature vector
3 – Message passing through edges

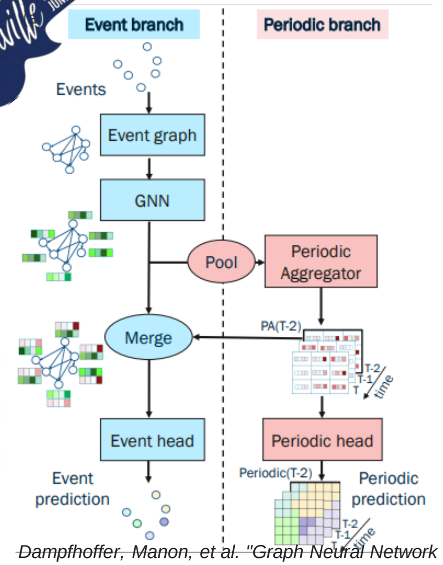
Dalgaty, Thomas, et al. "The cnn vs. snn event-camera dichotomy and perspectives for event-graph neural networks." 2023 Design, Automation & Test in Europe Conference & Exhibition (DATE). IEEE, 2023.

Event-graphs memory saving

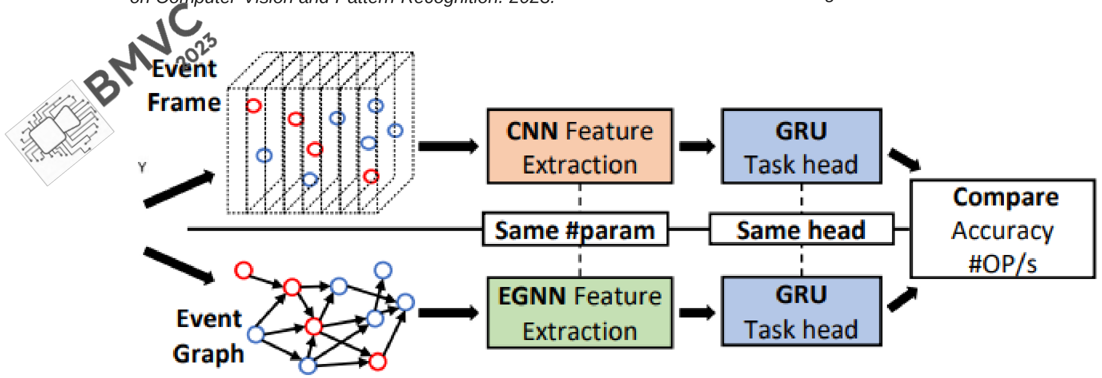
HUGnet-1 (2023) & HUGnet-2 (2025)



Dalgaty, Thomas, et al. "Hugnet: Hemi-spherical update graph neural network applied to low-latency event-based optical flow." Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2023.



Dampfhoffer, Manon, et al. "Graph Neural Network Combining Event Stream and Periodic Aggregation for Low-Latency Event-based Vision." Proceedings of the Computer Vision and Pattern Recognition Conference. 2025.



Mesquida, Thomas, et al. "G2N2: lightweight event stream classification with GRU graph neural networks." BMVC 2023-The 34th British Machine Vision Conference. 2023.

- ❑ ***10000x faster*** *optical-flow updating latency vs. convolutional neural network*

- ❑ ***100x fewer OPS/s*** *vs. a CNN for the same performance*

- ❑ ***100x fewer parameters*** *than a CNN for the same performance (i.e., 10-100s of KB)*

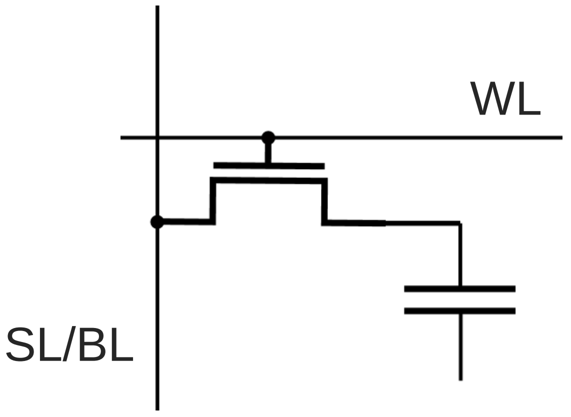
- ❑ *Better AI models that map onto inherent structure of data can ease on-chip memory requirements*

Memory-centric computing

1. Memory-frugal algorithmic design (beyond quantization)
2. New on-chip memory concepts (replacing SRAM)
3. In-memory computing (“neuromorphic” computing)
4. The next big bet for memory-centric computing ?

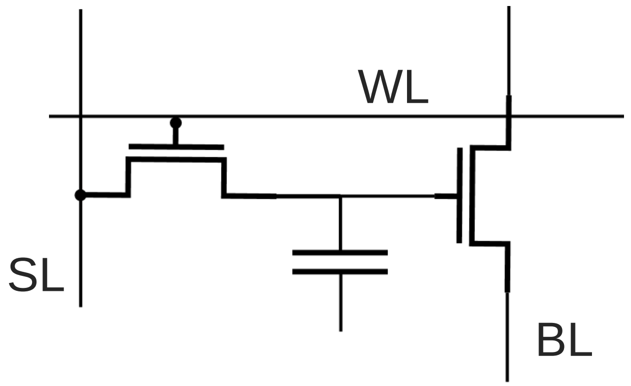
Embedded DRAM (eDRAM)

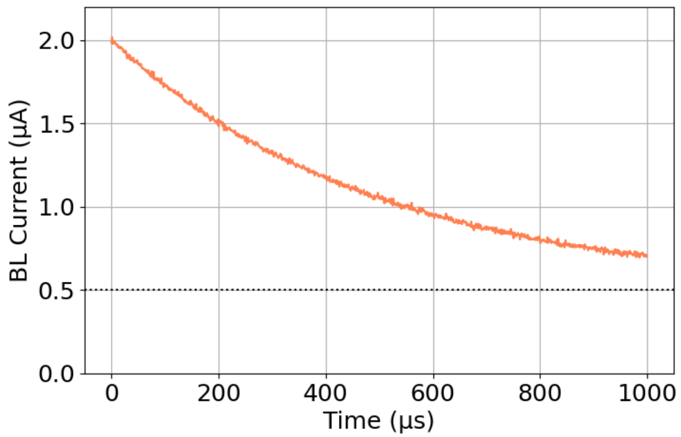
Off-chip DRAM cell



- ❑ *Store a charge on a big deep-trench capacitor (10s of fF)*
- ❑ *~50ns latency*
- ❑ *65ms refresh time*

<u>On-chip</u> eDRAM "gain-cell"

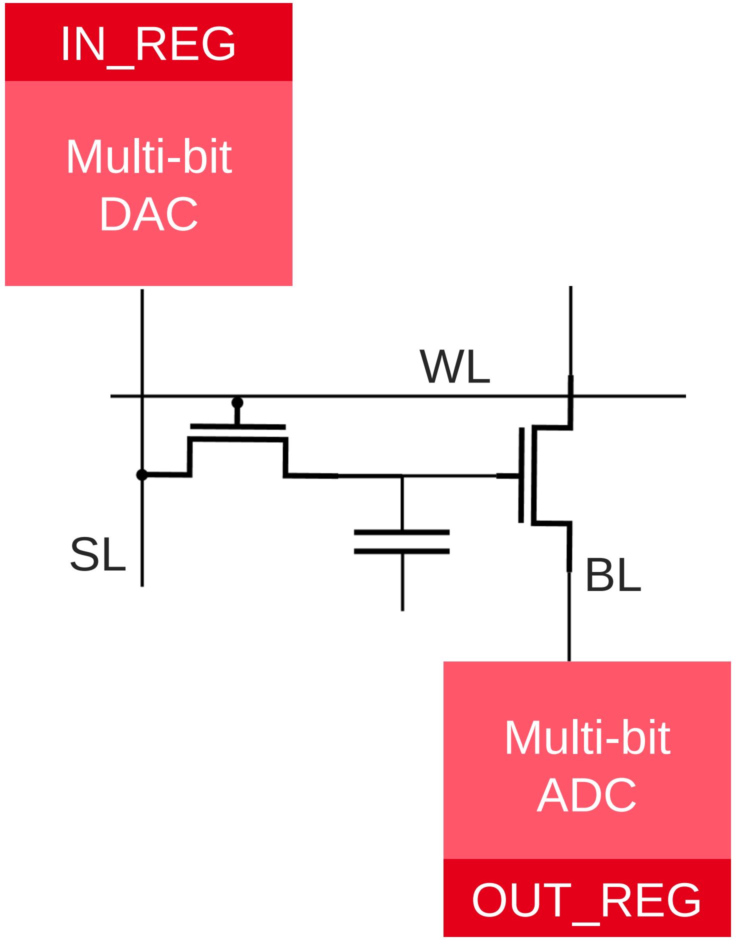


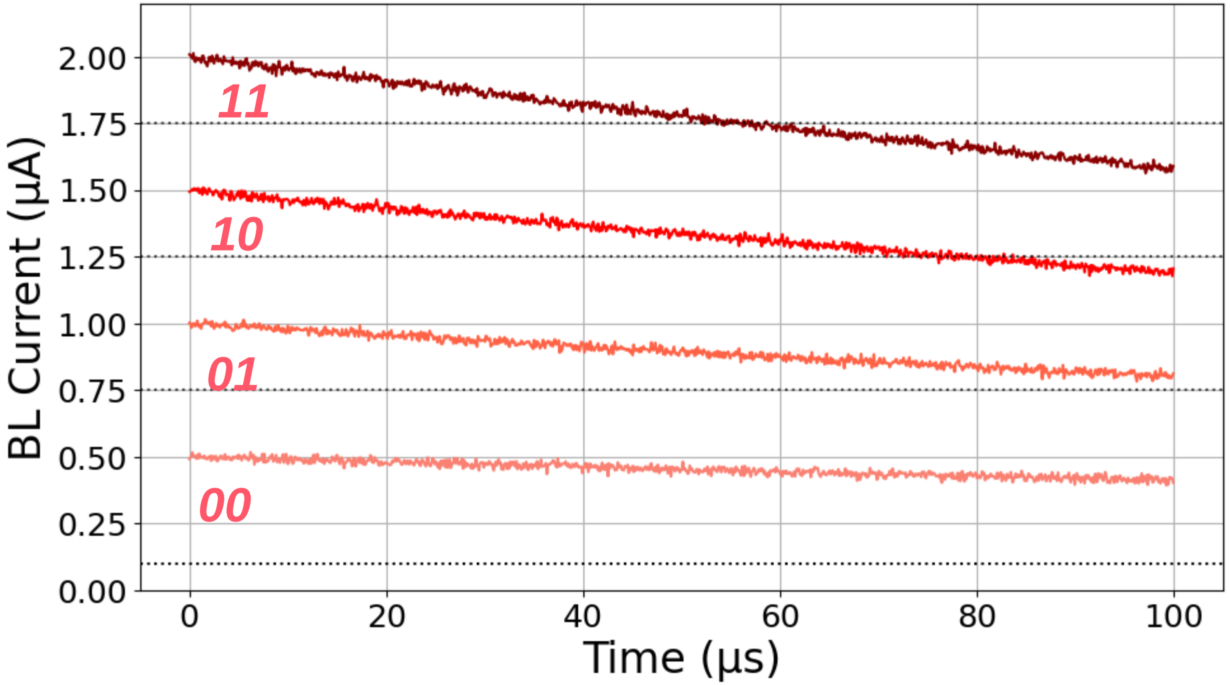


- ❑ *Store a "current" on a small MIM/MOM capacitor (< 1fF)*
- ❑ *~1-2ns latency - SRAM-like speed*
- ❑ ***2-3x denser** than an SRAM cell (0.2um2 vs. 0.6um2 in 14nm)*
- ❑ *100μs refresh time*

Bonetti, Andrea, et al. "Gain-cell embedded DRAMs: Modeling and design space." IEEE Transactions on Very Large Scale Integration (VLSI) Systems 28.3 (2020): 646-659.

Multi-bit eDRAM



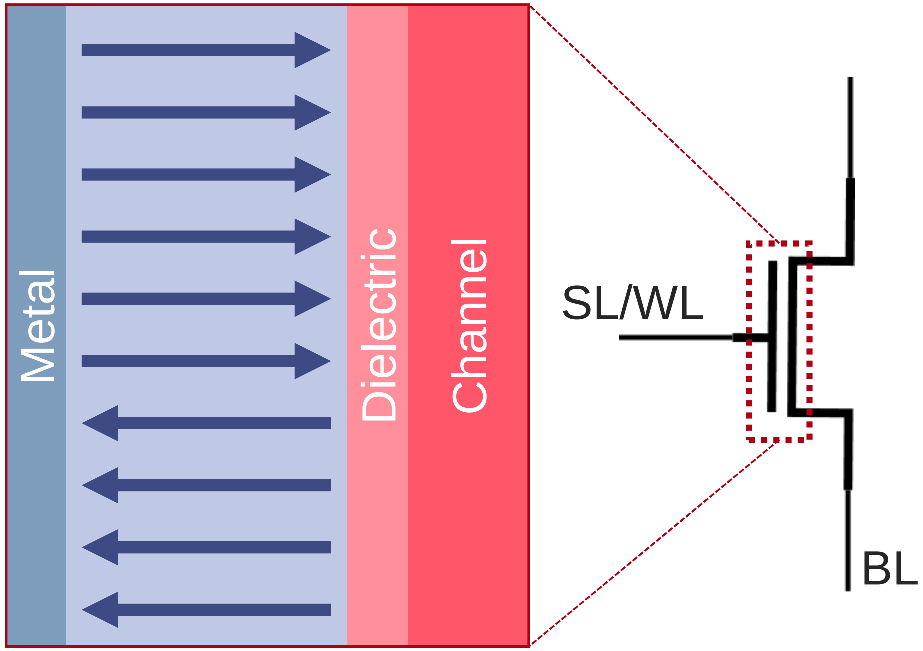


- ❑ *Voltage DAC & current ADC can permit multi-bit operation*
- • *But, at the expense of slower reading*
- ❑ *Ultra-low leakage transistors (<1 aA) are arriving to the market*
- ❑ *Example : 4-bit eDRAM could increase density by 12x vs. SRAM.* ***Groq chip goes from 220MB up to 2,6GB of on-chip memory****

****theoretical limit, not considering read-port size and side-effects***

Non-volatile multi-bit memory - FeFET

Ferro-electric polarisation domains



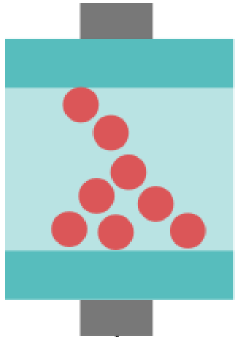
- ❑ Electric field determines the polarization of ferroelectric domains
- ❑ The sum of all domains modifies the transconductance and therefore BL current read from the transistor
- ❑ Projections
  - ❑ **10x higher density** than SRAM in equivalent nodes
  - ❑ Up to **4-6-bit multi-level** storage (16-64 distinct current levels)
  - ❑ On the market after 2030
- ❑ For Groq : from **220MB up to 10GB*** of on-chip memory

theoretical limit, not considering read-port size and side-effects

Ali, T., et al. "High endurance ferroelectric hafnium oxide-based FeFET memory without retention penalty." IEEE Transactions on Electron Devices 65.9 (2018): 3769-3774.

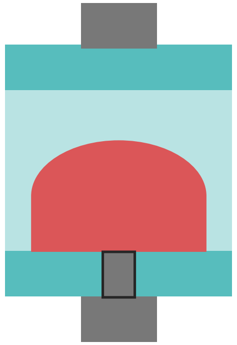
Two terminal non-volatile multi-bit memory

Filamentary ReRAM (OxRAM)



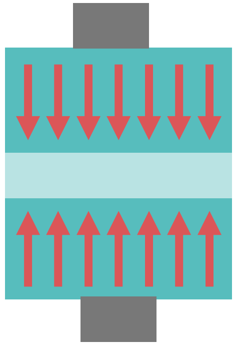
□ *Programming modifies conductance of filament*

Phase-change memory (PCM)



□ *Modulate conductance by heating/cool crystalline mushroom*

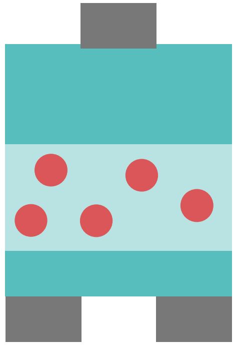
Magnetic spin transfer torque memory (MRAM)



□ *Set magneto-tunneling resistance by flipping magnetic spin*

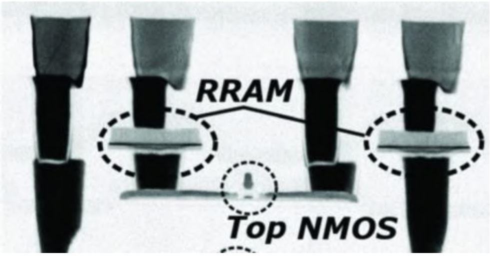
Electro-chemical memory (ECRAM)

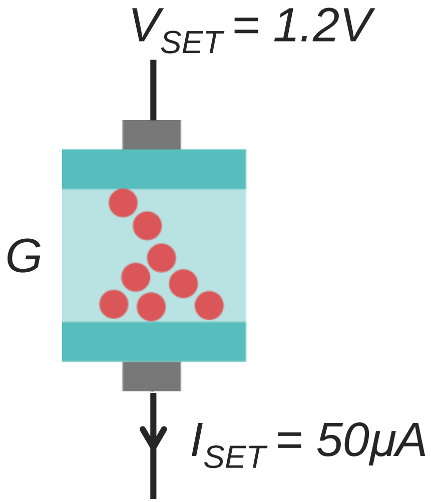
Not yet on the market

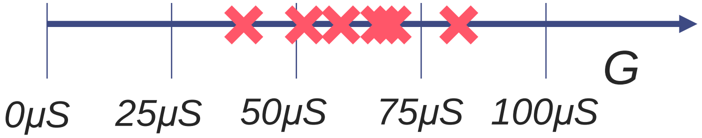


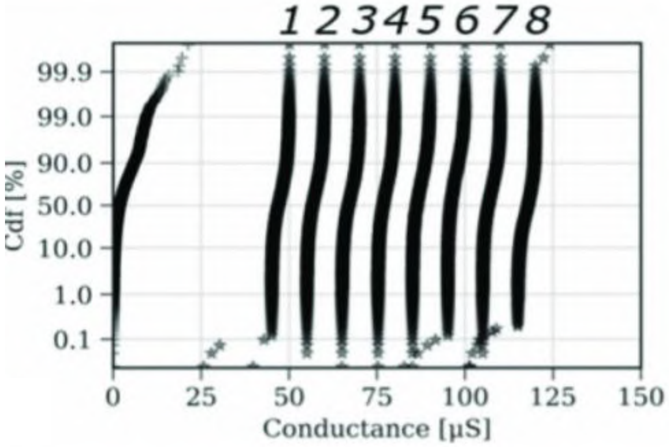
□ *Modulate channel conductance via chemical reactions with a reservoir*

Multi-bit limitation from programming variability









Esmanhotto, E., Brunet, L., Castellani, N., Bonnet, D., Dalgaty, T., Grenouillet, L., ... & Vianello, E. (2020, December). High-density 3D monolithically integrated multiple 1T1R multi-level-cell for neural networks. In 2020 IEEE International Electron Devices Meeting (IEDM) (pp. 36-5). IEEE.

❑ *It is challenging to go beyond 3-4 bits of precision due to intrinsic variability*

❑ *Is it still worth it considering increased read-out complexity ?*

Memory-centric computing

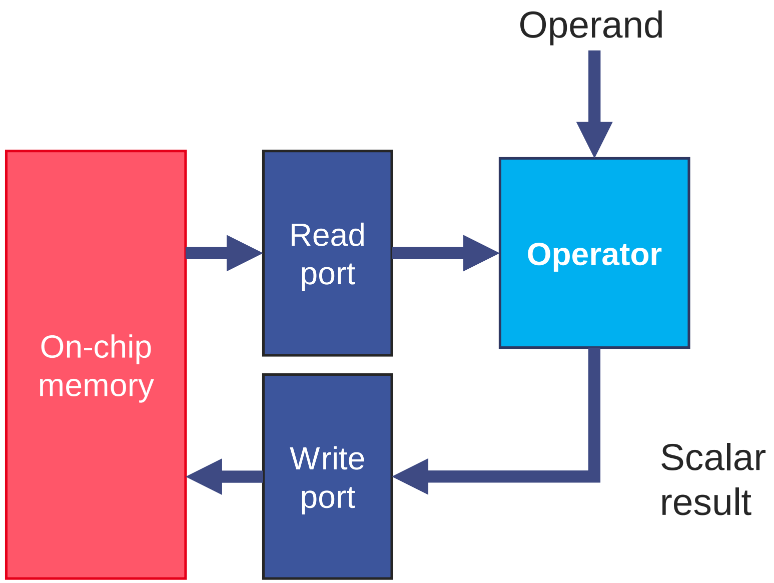
1. Memory-frugal algorithmic design (beyond quantization)

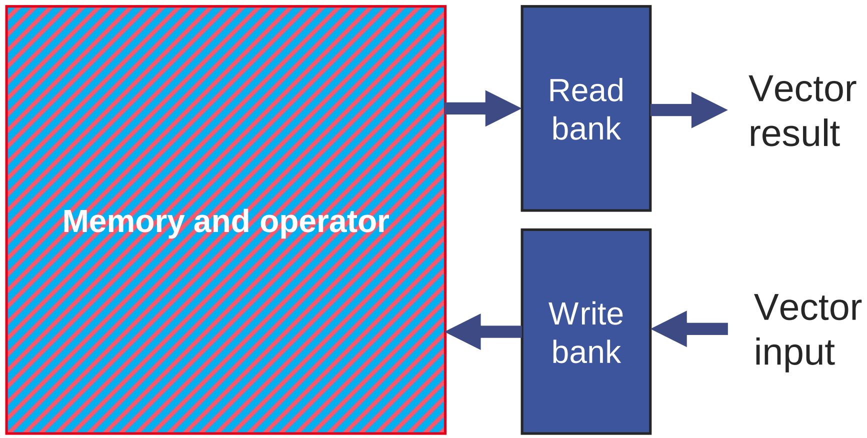
2. New on-chip memory concepts (replacing SRAM)

3. In-memory computing (“neuromorphic” computing)

4. The next big bet for memory-centric computing ?

How to further increase bandwidth ?





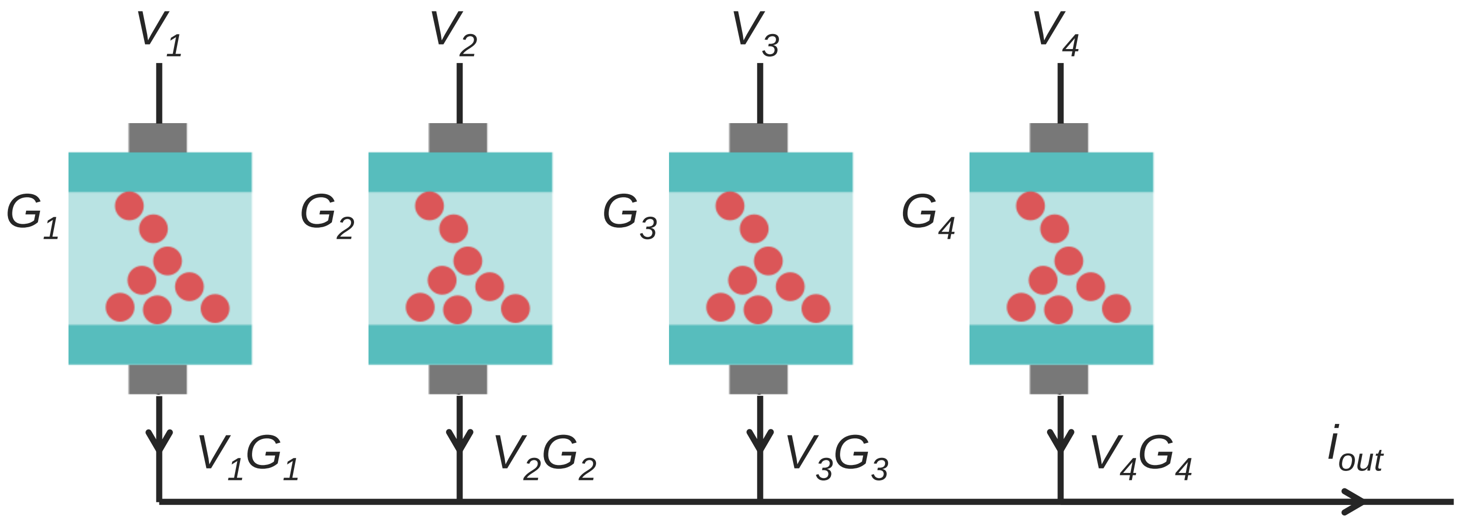
❑ Will ultimately be limited by:

- *Memory size*
- *Operator size*
- *Port size*
- *Multi-bit resolution*
- *Read/write speed*

❑ Fuse memory and operator in same circuit

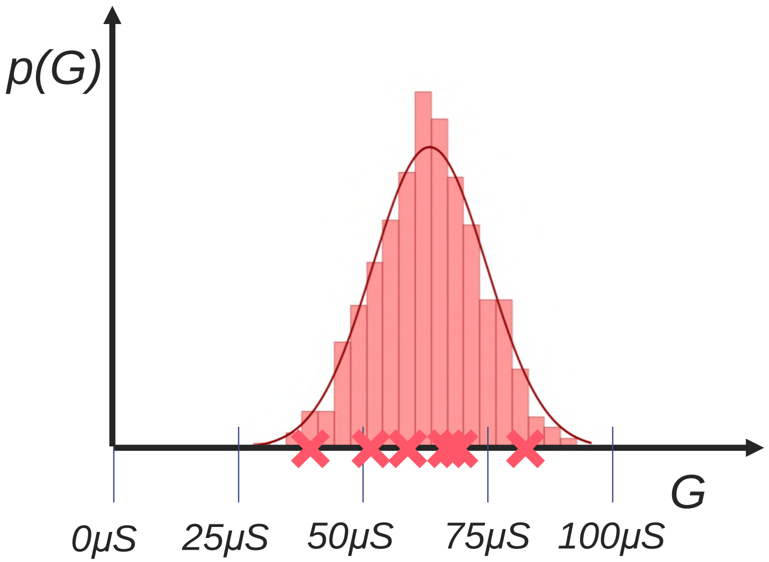
❑ Read/write banks relay input and output of vectorial computation

In-memory computing



$$i_{out} = V_1G_1 + V_2G_2 + V_3G_3 + V_4G_4 = \boldsymbol{V \cdot G}$$

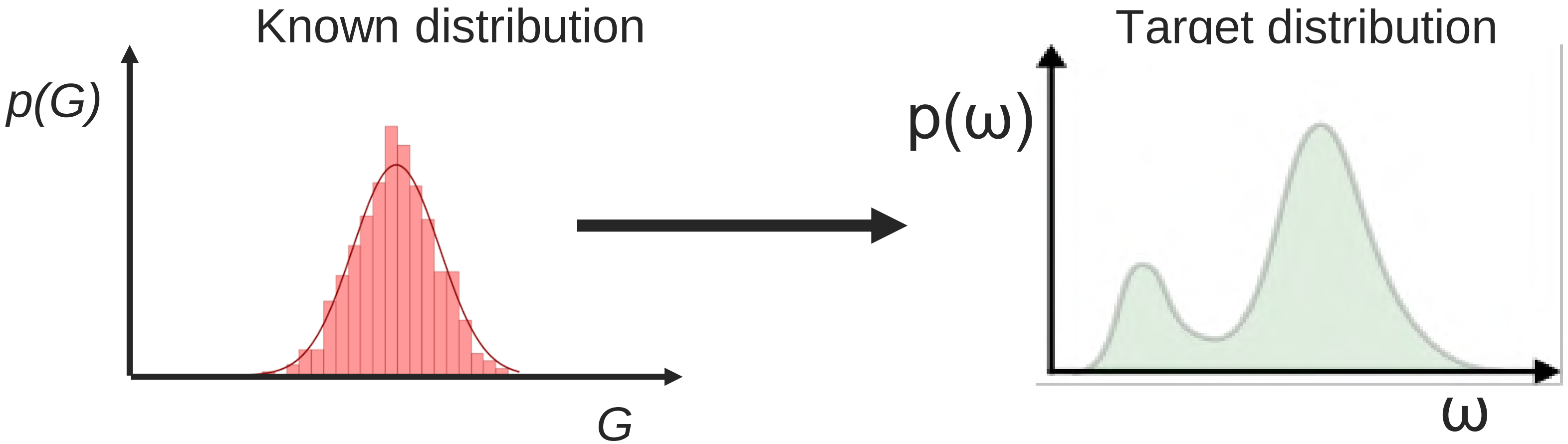
Change of perspective !
Programming draws a sample from a precise probability distribution in a physical precision



❑ Analogue computing using physics
❑ Extend into large multi-row memory arrays
❑ Access bandwidth by write and read bank size ~ 128 DAC/ADC per bank
❑ Massively parallel access brings **bandwidth >> 100TB/s** for a full chip

❑ But we are still limited to 4-6 bits
❑ This also restricts application scope

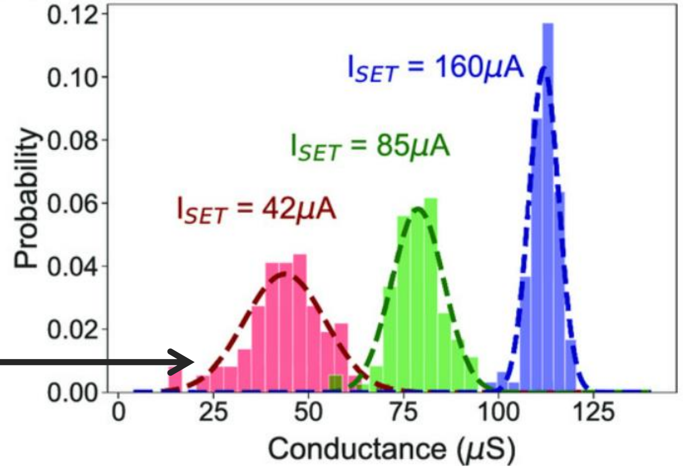
Probabilistic in-memory computing



□ Probabilistic inference uses samples from a known distribution to characterize complex target distributions

□ Memory cells become the sampling operator and multiply/add operator

$$p(G|I_{SET}) = N(\mu, \sigma^2)$$



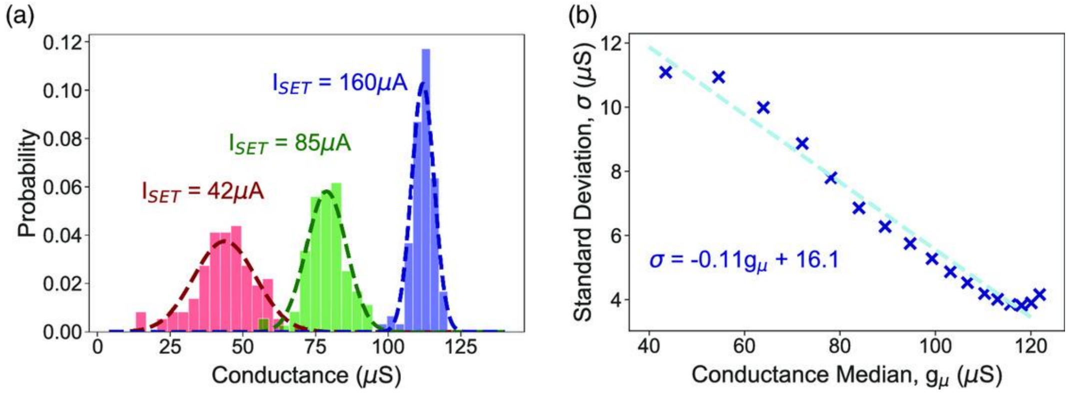
Dalgaty, Thomas, et al. "Ex situ transfer of bayesian neural networks to resistive memory-based inference hardware." Advanced Intelligent Systems 3.8 (2021): 2000103.

□ Data stored in “physical precision”

- Conservatively > 16-bits per cell

Generalization via convolution-closed mixing

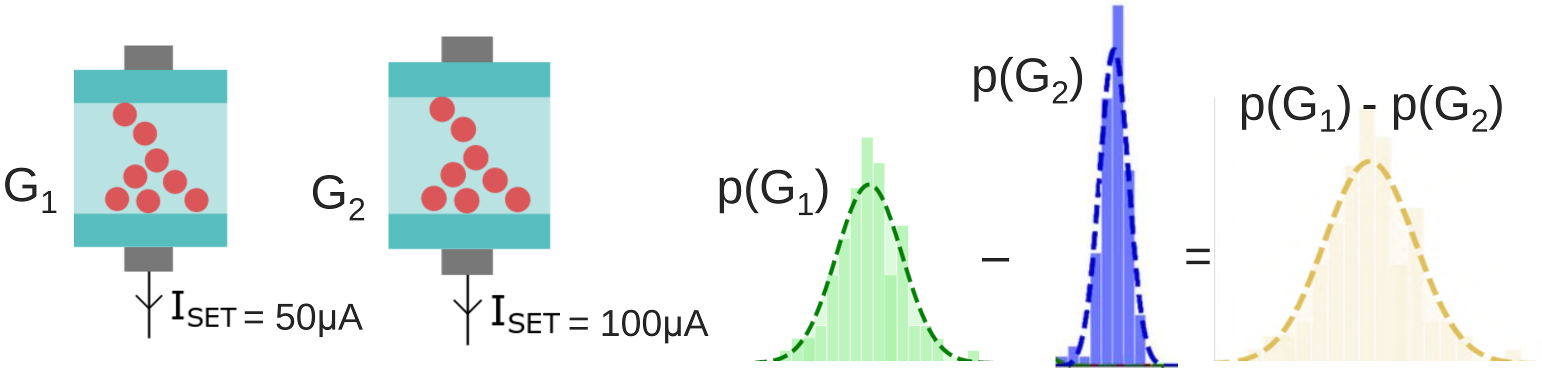
- ❑ Samples from a single memory may have correlated mean and variance

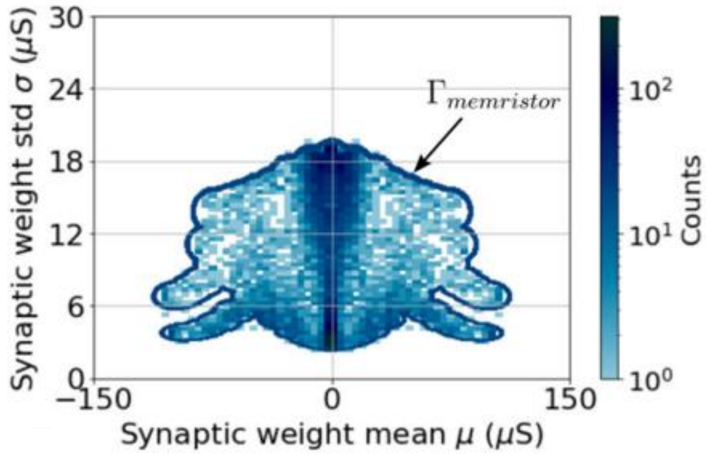


Dalgaty, Thomas, et al. "Ex situ transfer of bayesian neural networks to resistive memory-based inference hardware." Advanced Intelligent Systems 3.8 (2021): 2000103.

- ❑ The difference of Gaussian samples is also Gaussian (“convolution-closed”)

$$p(G1) - p(G2) = N(\mu_1, \sigma_1{}^2) - N(\mu_2, \sigma_2{}^2) = N(\mu_1 - \mu_2, \sigma_1{}^2 + \sigma_2{}^2)$$

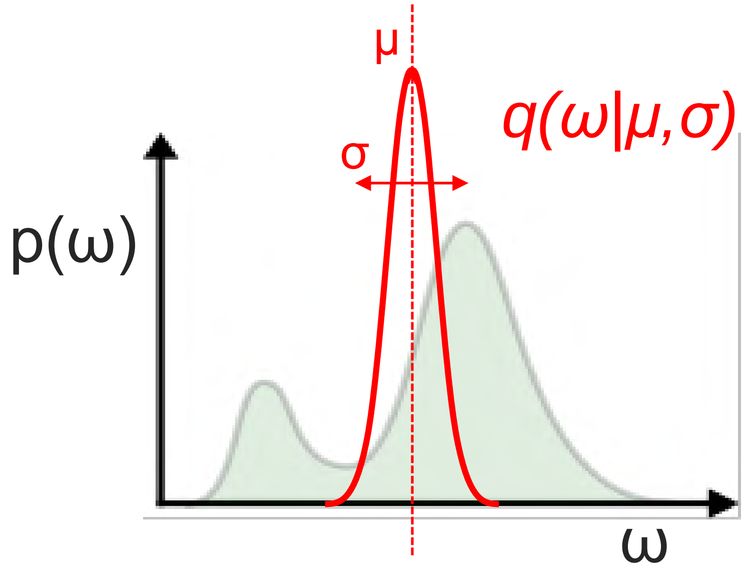




Bonnet, Djohan, Tifenn Hirtzlin, Atreya Majumdar, Thomas Dalgaty, Eduardo Esmanhotto, Valentina Meli, Niccolo Castellani et al. "Bringing uncertainty quantification to the extreme-edge with memristor-based Bayesian neural networks." *Nature Communications* 14, no. 1 (2023): 7530.

- ❑ Differential samples provides a larger domain of distributions

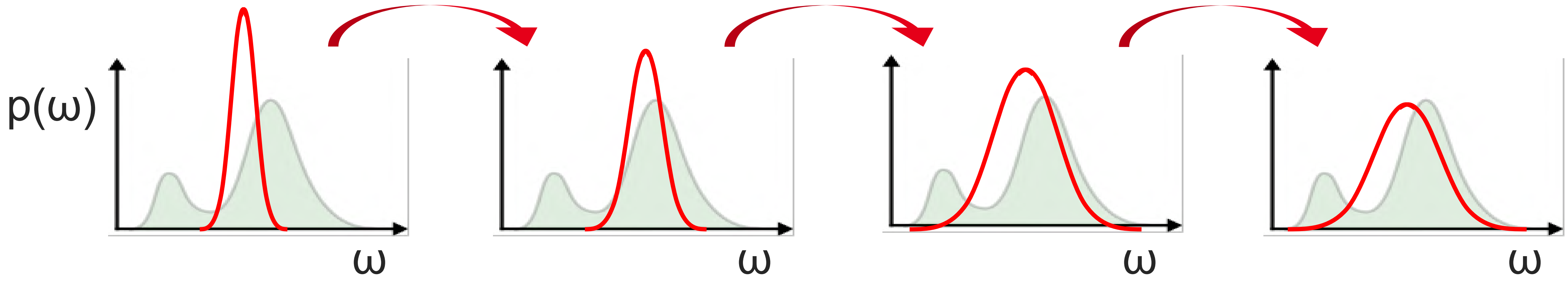
Principle of variational inference



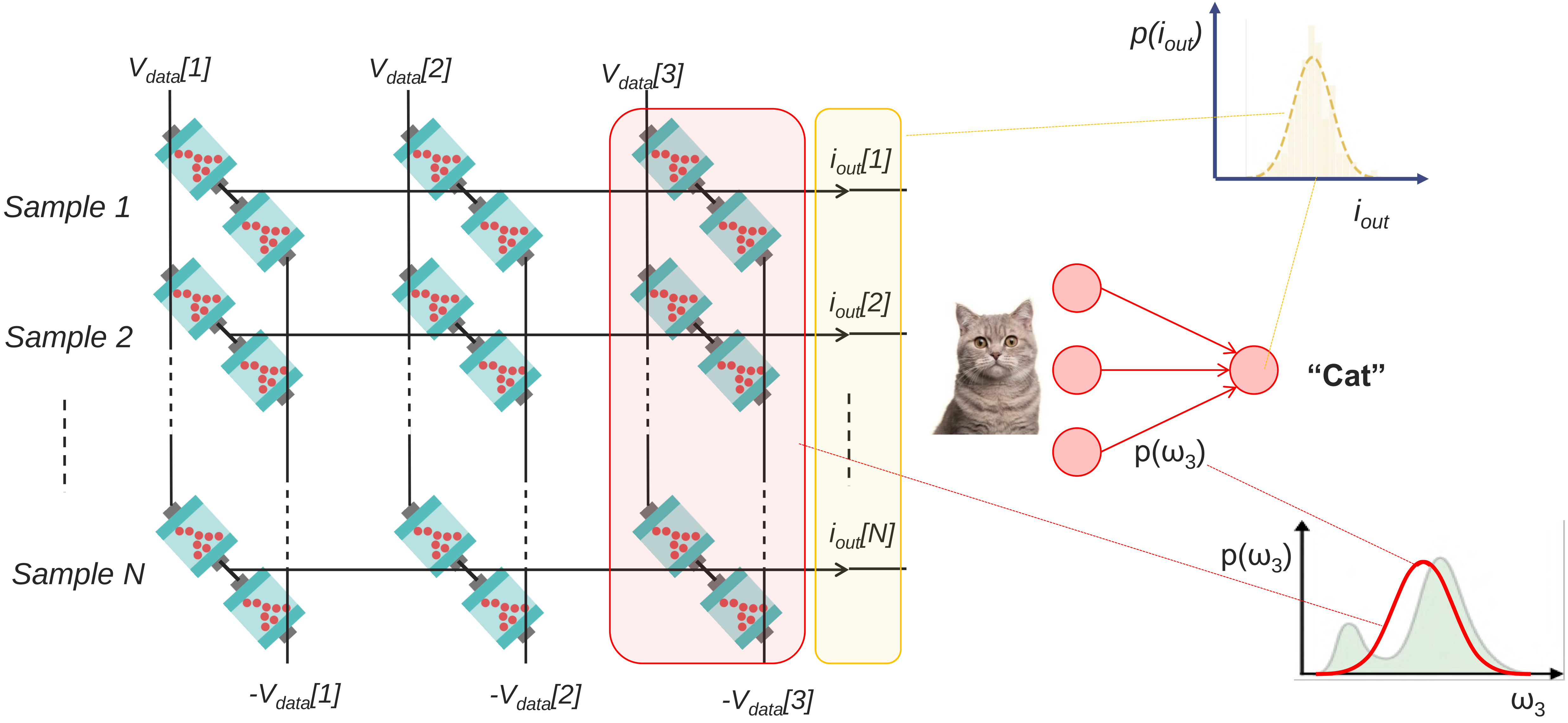
❑ Minimise the Kullback-Leibler between p(ω) and q(ω|μ,σ)

1. Draw samples from q(ω|μ,σ)
2. Calculate expectation of log-likelihoods*
3. Calculate gradients w.r.t. μ and σ (with backpropagation)
4. Update μ and σ
5. Goto 1

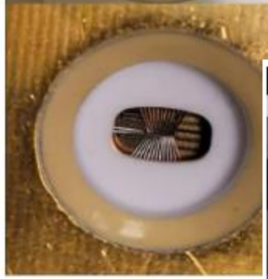
**And account for prior*

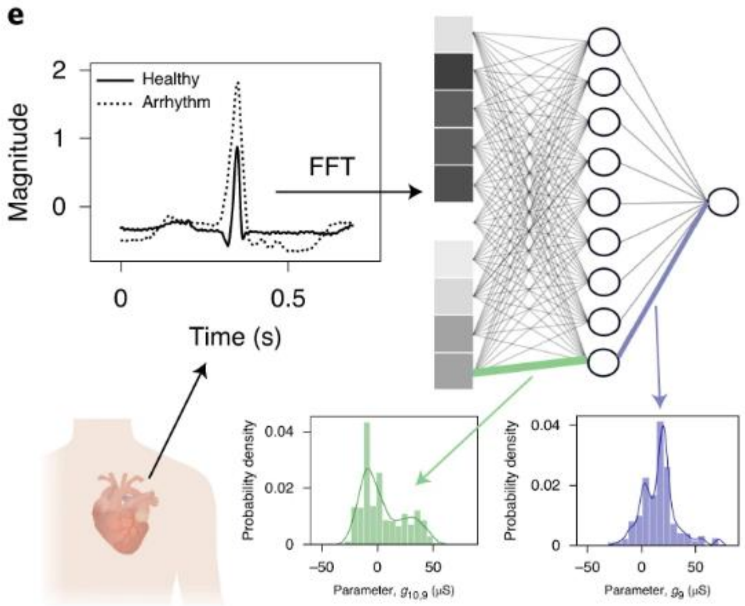


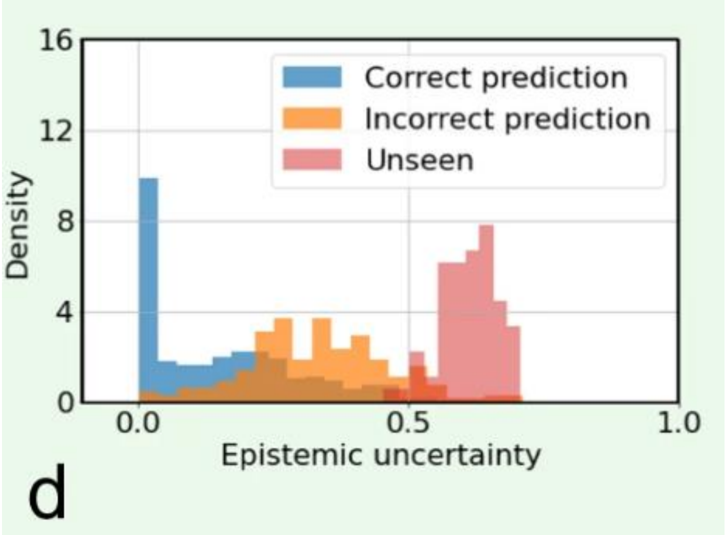
AIMC variational inference



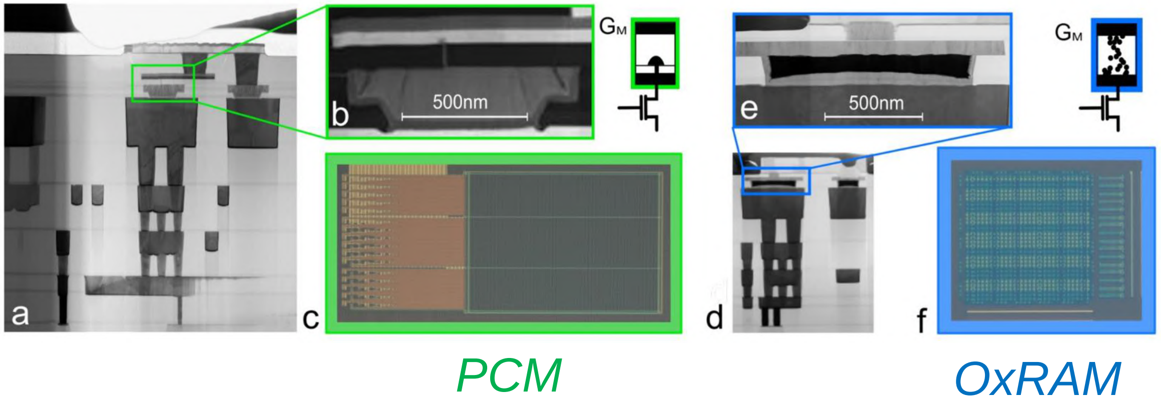
Experimental proof of concept







Bonnet, Djohan, Tifenn Hirtzlin, Atreya Majumdar, Thomas Dalgaty, Eduardo Esmanhotto, Valentina Meli, Niccolo Castellani et al. "Bringing uncertainty quantification to the extreme-edge with memristor-based Bayesian neural networks." Nature Communications 14, no. 1 (2023): 7530.



❑ **Bayesian neural network chip to identify medical prediction errors and unseen disease types**

Probabalistic learning with Langevin Monte Carlo

Stochastic Gradient Langevin Dynamics (SGLD)

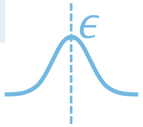
Welling, Max. "Bayesian learning via stochastic gradient Langevin dynamics." ICML. 2011.

$$\Delta\omega = \tau\nabla_{\omega}L(\omega) + \sqrt{2\tau}\epsilon$$

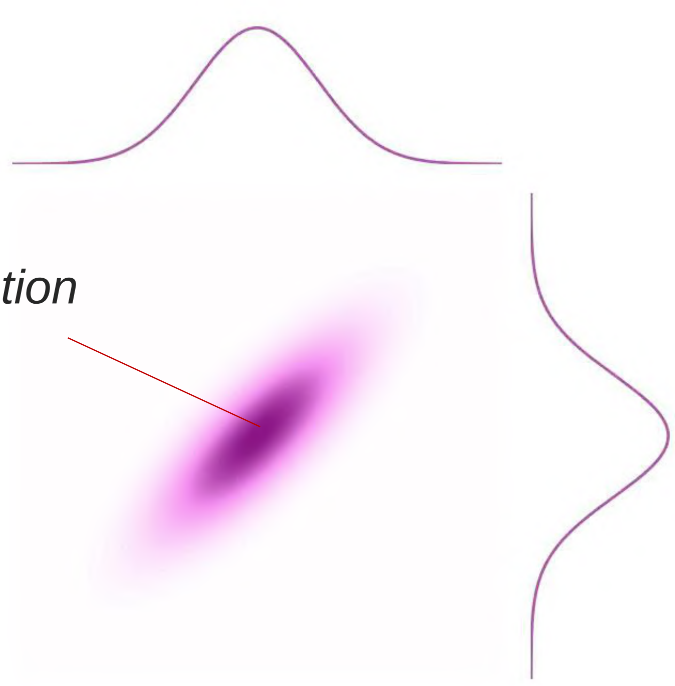
✗ *Model sample*

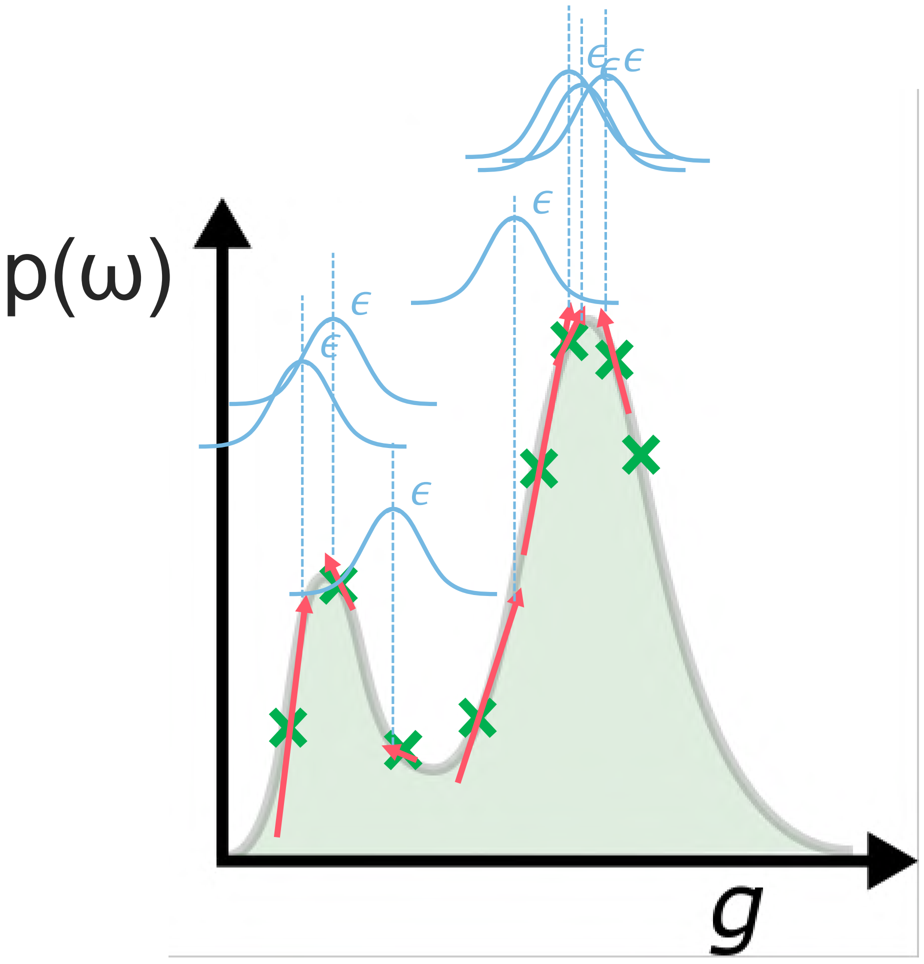
Apply Langevin

Perturb g with noise

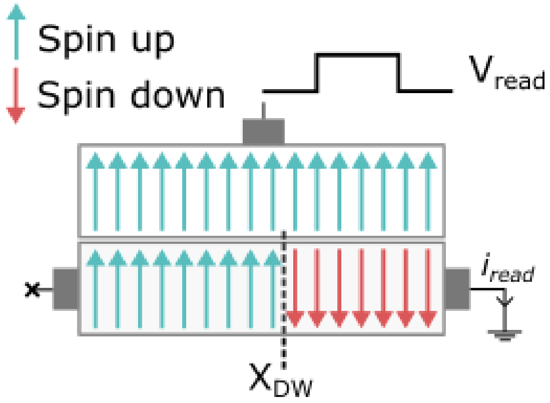


Weights distribution



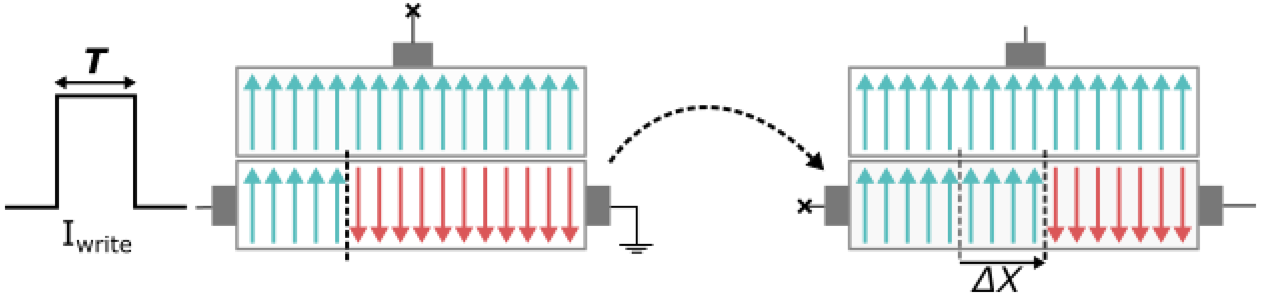


Introducing magnetic domain-wall memory



*The domain-wall position (X_{DW}) determines the **tunneling magnetoresistance* (TMR)** between magnetic layers*

*Spin-torque transfer current pulses move the domain-wall and **change cell conductance***



** M. Julliere (1975). "Tunneling between ferromagnetic films". Phys. Lett. **54A** (3): 225–6*
& T. Miyazaki (1995). "Giant magnetic tunneling effect in Fe/Al2O3/Fe ". Magn. Mater. 139 (3).

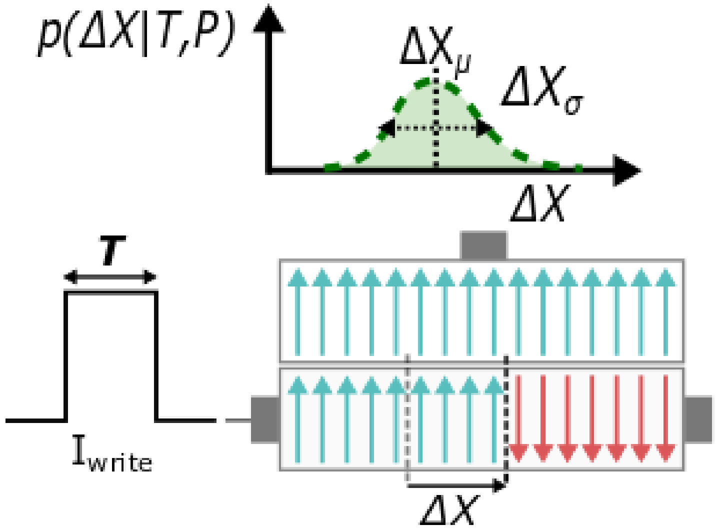
NEURAL INFORMATION PROCESSING SYSTEMS

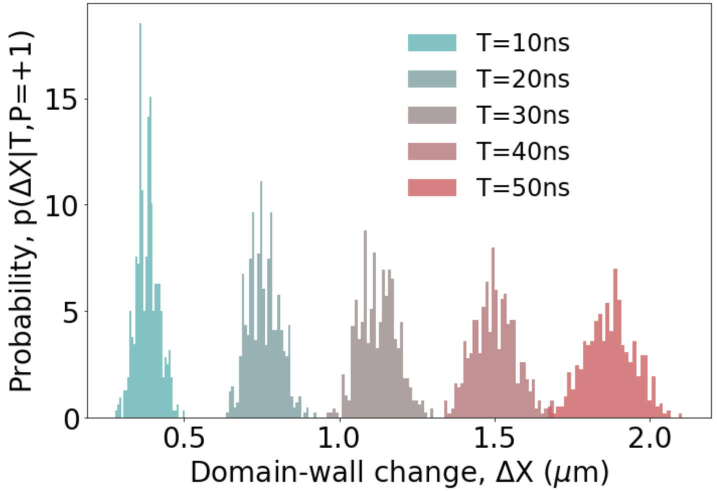
Best paper award

TDK

Domain-wall position change variability

*Physical phenomena dictate that domain-wall movement is stochastic – **position change is sampled from p(ΔX|T,P)***



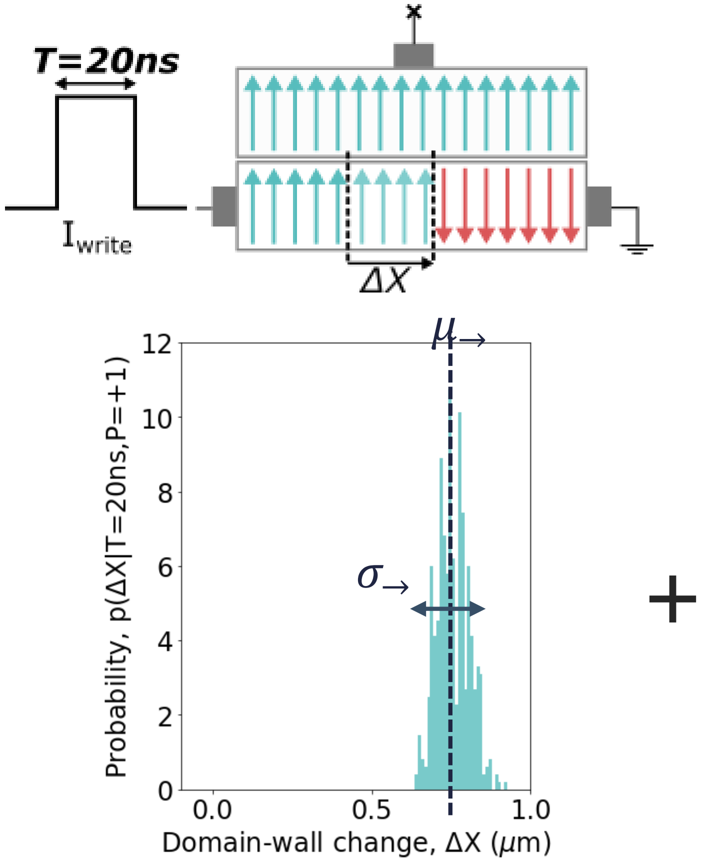


*The **current pulse-width** determines the mean (X_μ) and variance (X_σ) of the change*

*Dalgaty, Thomas, et al. "Scaling-up Memristor Monte Carlo with magnetic domain-wall physics." MLNCP2023-37th NeurIPS Machine Learning with New Compute Paradigms workshop. 2023. (**best paper**)*

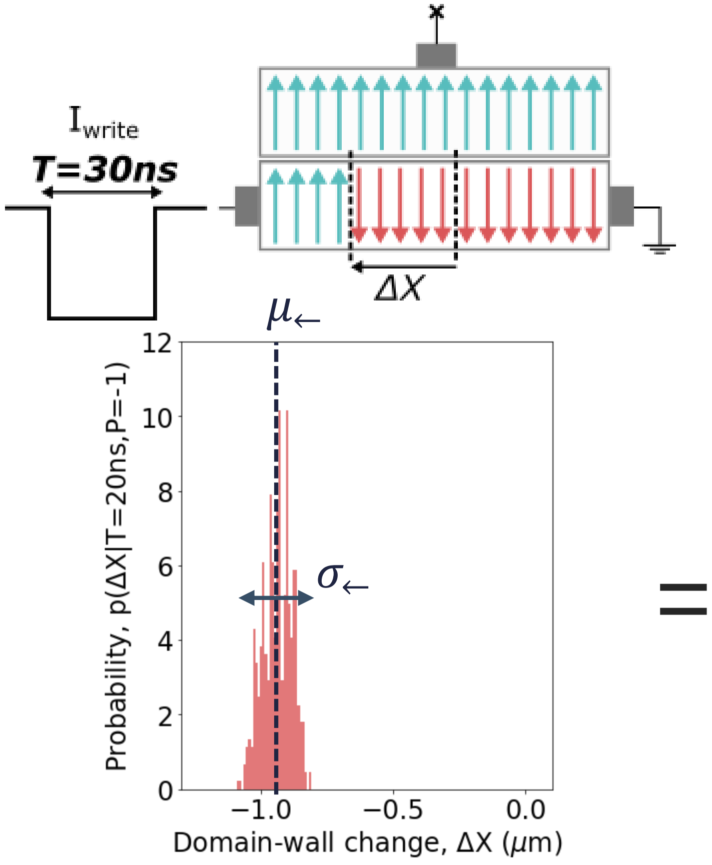
Domain-wall motion as Langevin updates

Push $(T_{\rightarrow})$



+

Pull $(T_{\leftarrow})$

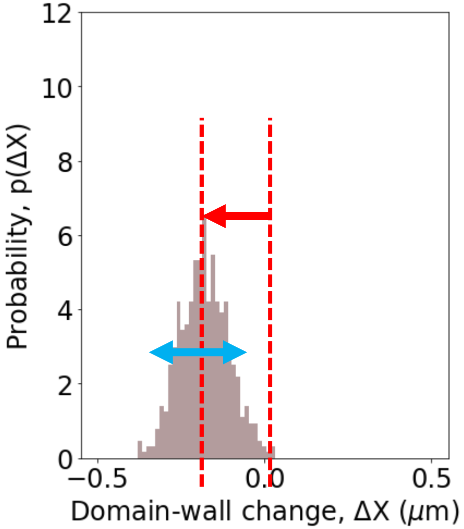


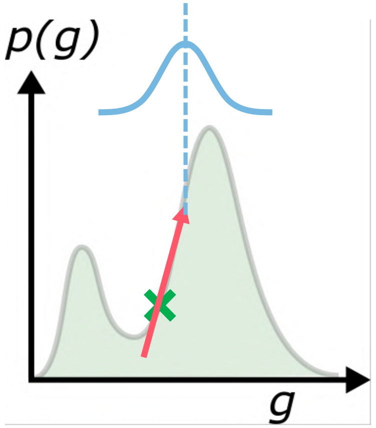
=

Normal probability law

- ❑ *Variances sum*
- ❑ *Difference of means*
- ❑ $p(\Delta X|\boldsymbol{T}_{\rightarrow},\boldsymbol{T}_{\leftarrow}) = N(\mu_{\rightarrow} - \mu_{\leftarrow}, \sqrt{{\sigma_{\rightarrow}}^2 + {\sigma_{\leftarrow}}^2})$

$$\Delta g = \tau \nabla_g L(g) + \sqrt{2\tau\epsilon}$$

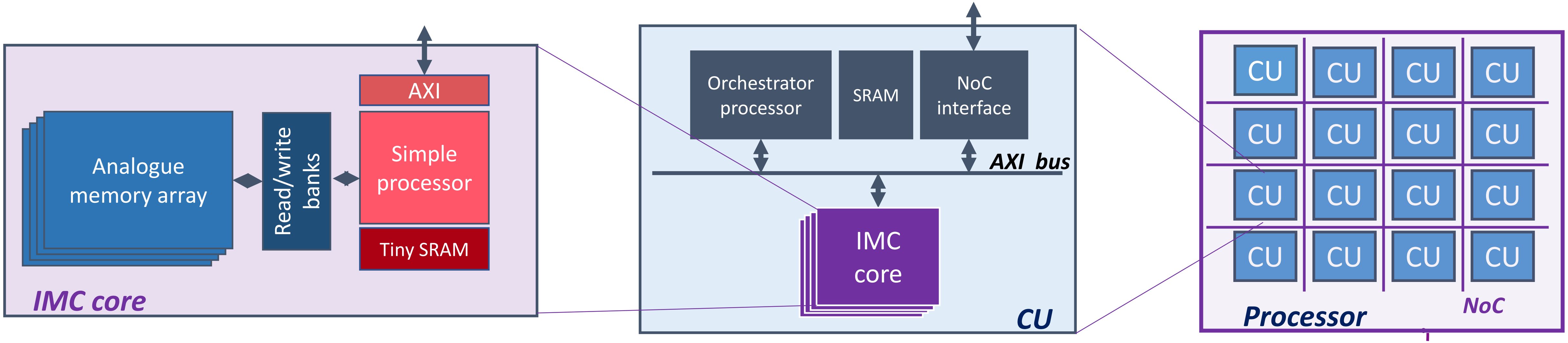




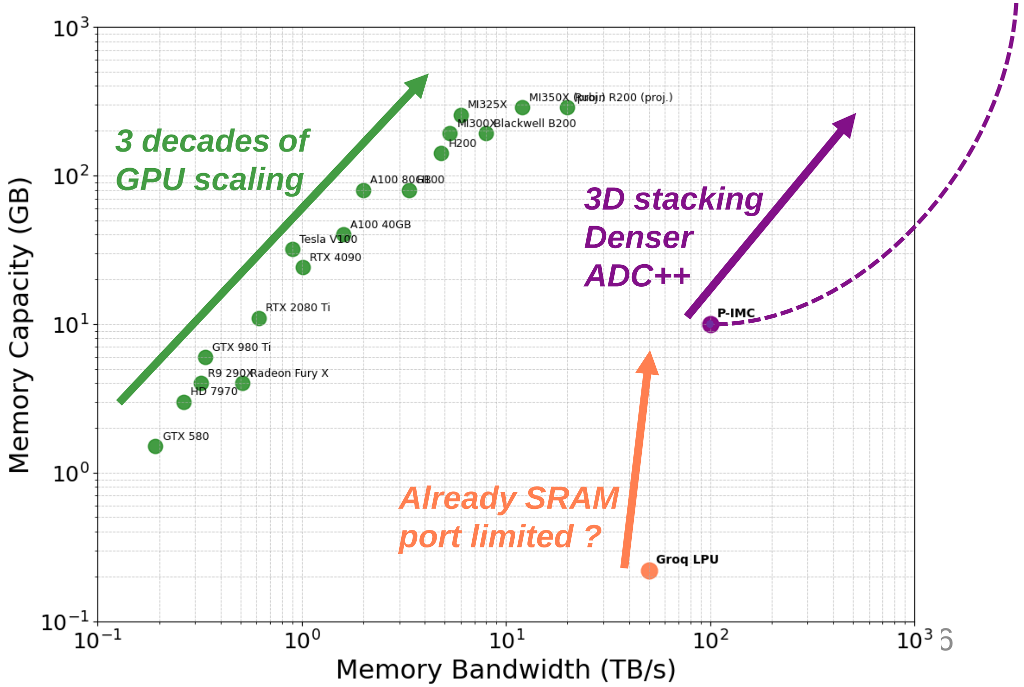
Applied to train a ResNet-50 on CIFAR-10 with floating-point software equivalent accuracy

*Dalgaty, Thomas, et al. "Scaling-up Memristor Monte Carlo with magnetic domain-wall physics." MLNCP2023-37th NeurIPS Machine Learning with New Compute Paradigms workshop. 2023. (**best paper**)*

Probabilistic IMC processor



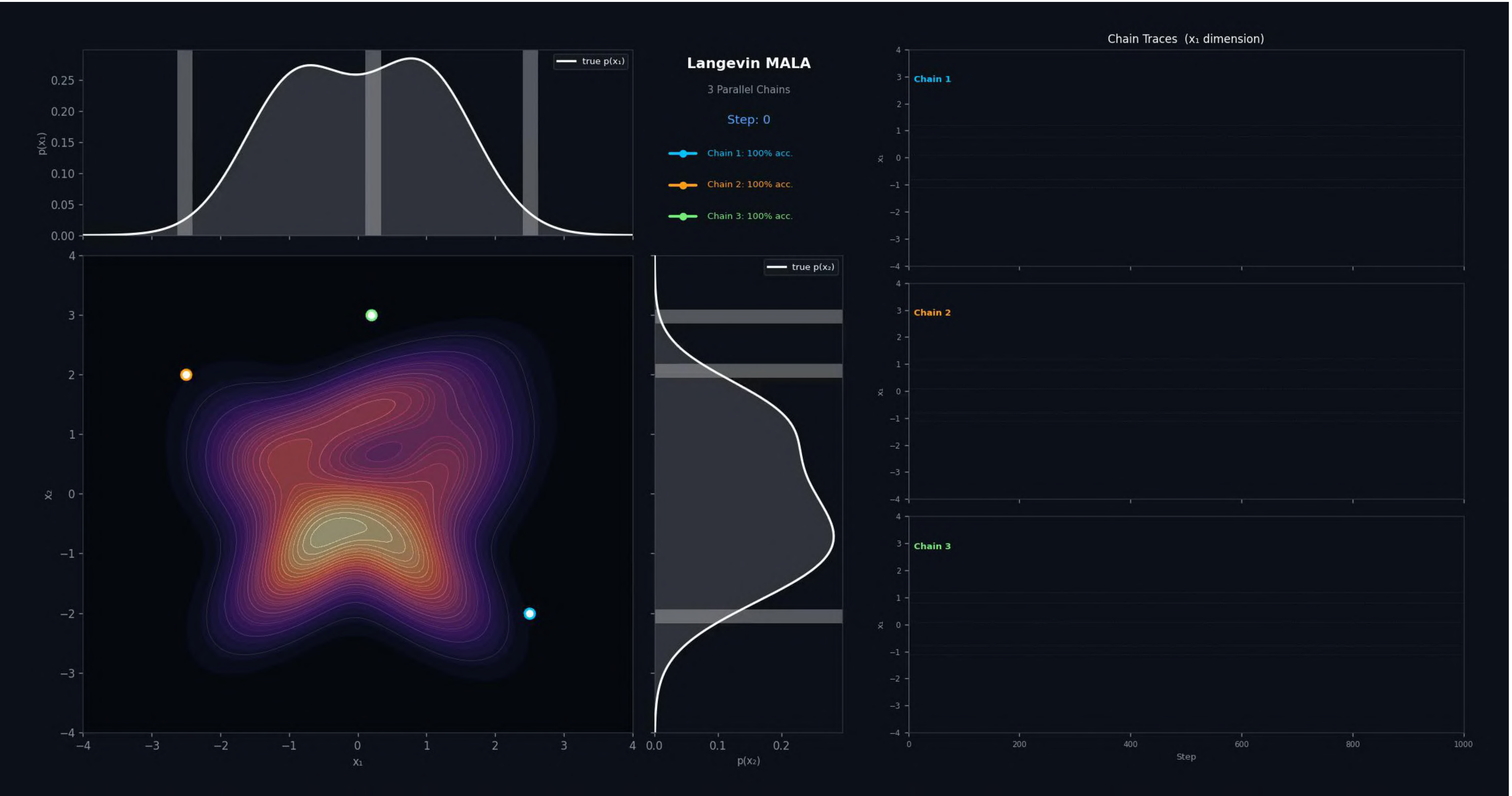
- ❑ **Key:** how much memory can I access as fast as possible ?
- ❑ For the class of probabilistic algorithms, IMC processor finds the best trade-off



Memory-centric computing

1. Memory-frugal algorithmic design (beyond quantization)
2. New on-chip memory concepts (replacing SRAM)
3. In-memory computing (“neuromorphic” computing)
4. The next big bet for memory-centric computing ?

What are probabilistic algorithms ?



What do probabilistic algorithms need ?

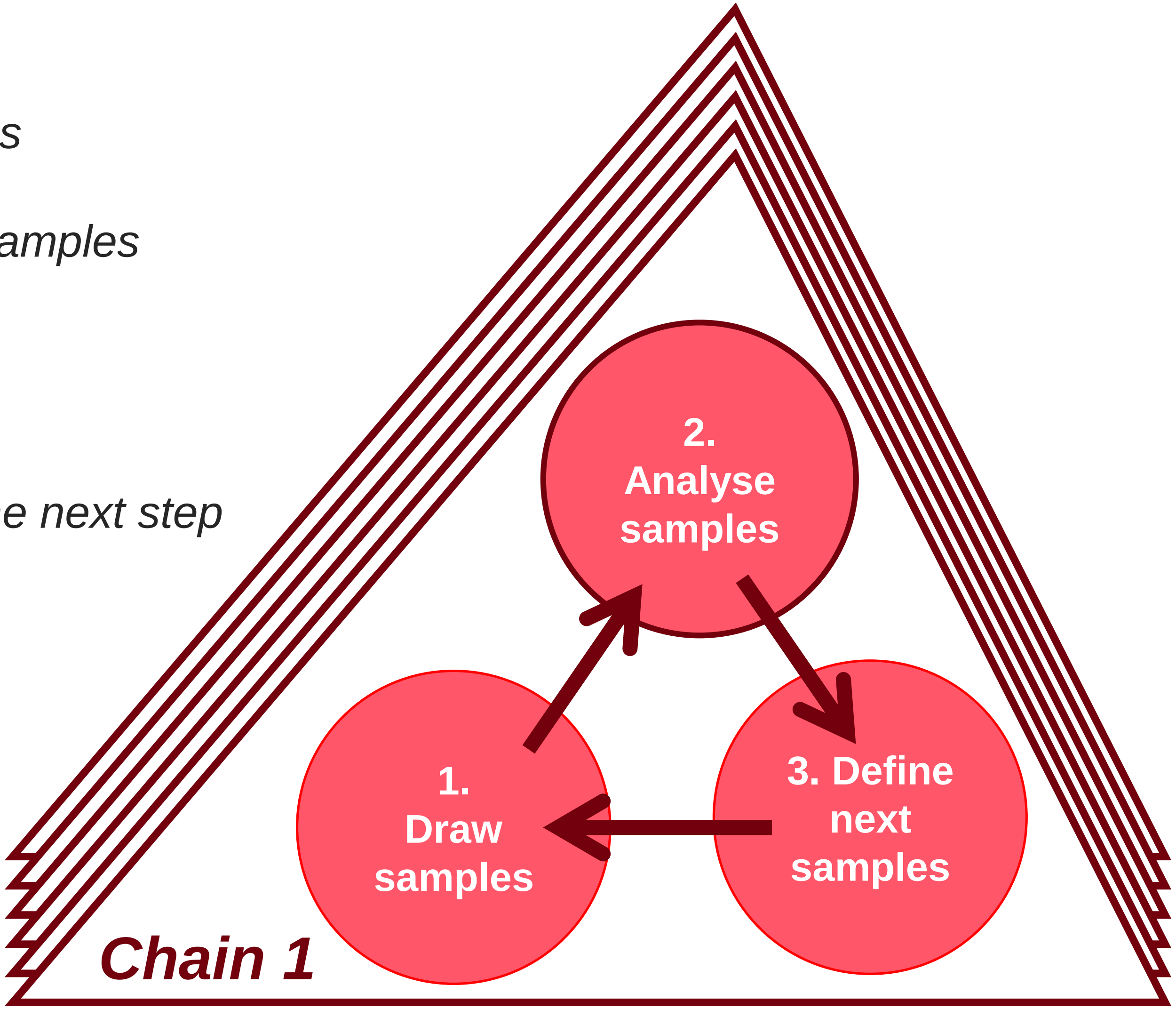
1. *Generate and store massive quantities of samples*

2. *Perform vectorial computation to analyze these samples*
   - ❑ *Langevin gradients*
   - ❑ *Hamiltonians*
   - ❑ *Acceptance probabilities*

3. *Define the probability distributions to sample in the next step of the algorithm*

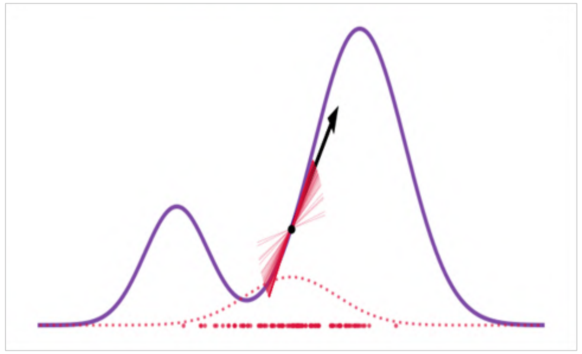
In parallel for thousands to millions of chains

Probabilistic algorithms need the memory capacity and bandwidth of in-memory computing processors



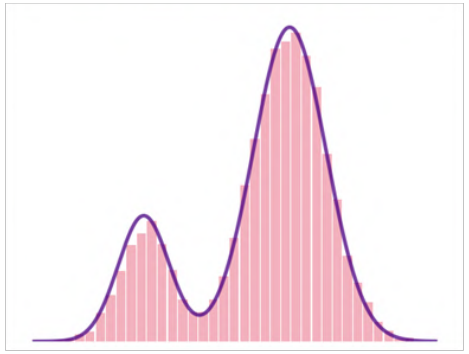
What are probabilistic algorithms used for ?

1. Calculate intractable functions



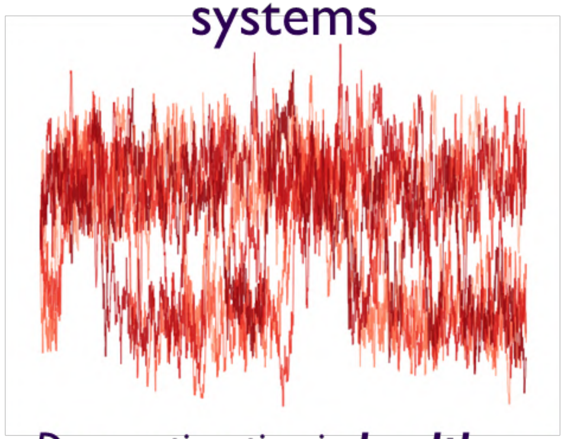
- ***Financial*** *pricing*
- *Integrals for* ***physical modelling***
- *Score functions for* ***Gen-AI***

2. Approximate probability distributions



- *Bayesian & probabilistic* ***AI***
- *Demand* ***forecasting***
- ***Insurance*** *risk quantification*
- *Strategy in* ***defense***

3. Simulate complex systems



- *Dose estimation in* ***healthcare***
- ***Logistics*** *optimization*
- ***Engineering*** *analysis*
- ***Nuclear*** *reactor simulation*

□ Nearly all existing industrial use cases are severely latency constrained

□ Probabilistic processor market estimated to be worth 5 billion USD in 2025

AI may need IMC to overcome its limitations

Mila ***most cited living scientist**

*"How can we design an AI that will be highly capable and not harm humans? Only probabilistic AI, by reasoning over multiple hypotheses, can avoid wrong predictions and prioritize human safety." **Youshua Bengio***

UNIVERSITY OF CAMBRIDGE

*"Probabilistic machine learning is the natural language for human-like reasoning under uncertainty for AI systems." **Zoubin Ghahramani***

Probabilistic Computing Project

https://probcomp.csail.mit.edu/

uai2025

https://www.auai.org/

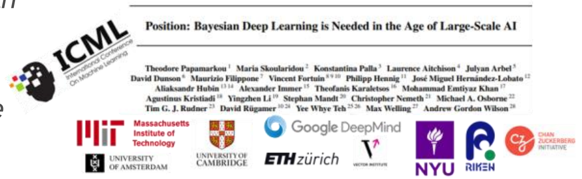
https://www.stancon2026.org/

Frontiers in Probabilistic Inference: Sampling Meets Learning @ ICLR 2025

https://sites.google.com/view/fpiworkshop/about

LawZero Safe AI for Humanity

https://lawzero.org/en/research



Position: Bayesian Deep Learning is Needed in the Age of Large-Scale AI

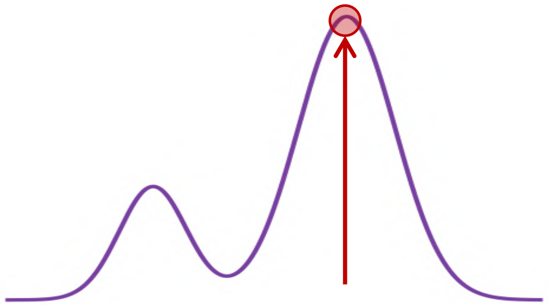
Theodore Papamarkou[1] Maria Skoularidou[2] Konstantina Palla[3] Laurence Aitchison[4] Julyan Arbel[5] David Dunson[6] Maurizio Filippone[7] Vincent Fortuin[8 9 10] Philipp Hennig[11] José Miguel Hernández-Lobato[12] Aliaksandr Hubin[13 14] Alexander Immer[15] Theofanis Karaletsos[16] Mohammad Emtiyaz Khan[17] Agustinus Kristiadi[18] Yingzhen Li[19] Stephan Mandt[20] Christopher Nemeth[21] Michael A. Osborne[22] Tim G. J. Rudner[23] David Rügamer[10 24] Yee Whye Teh[25 26] Max Welling[27] Andrew Gordon Wilson[28]

PyMC

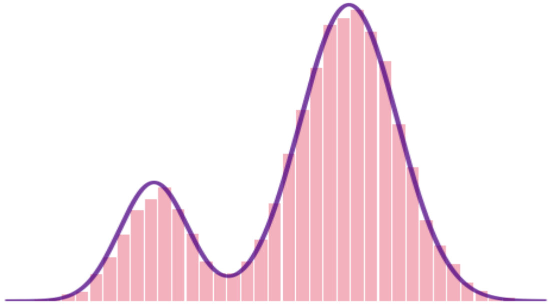
https://www.pymc.io/welcome.html

intel Probabilistic Computing A Pathway Through Real-World Uncertainty

https://www.intel.com/content/www/us/en/research/news/probabilistic-computing.html



Deterministic estimate

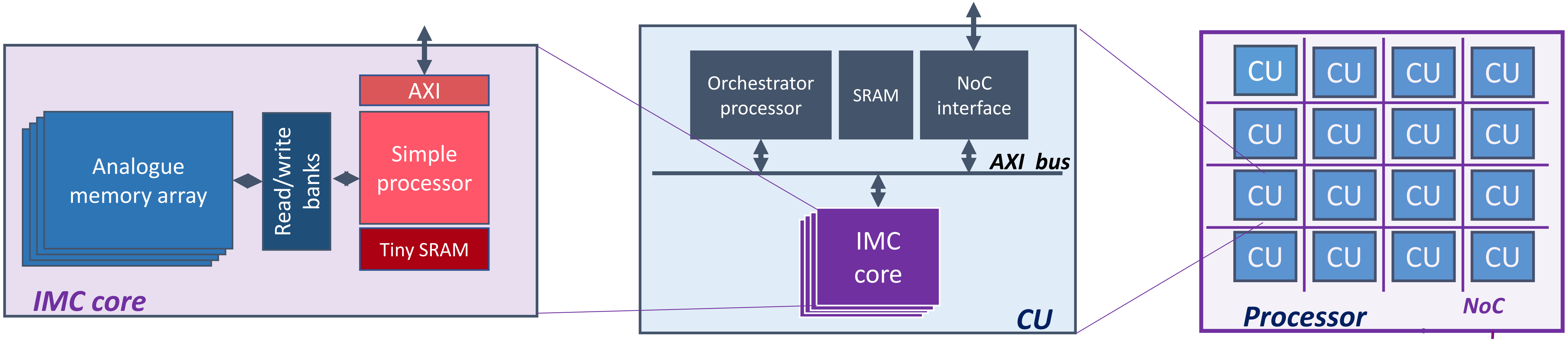


Probability distribution

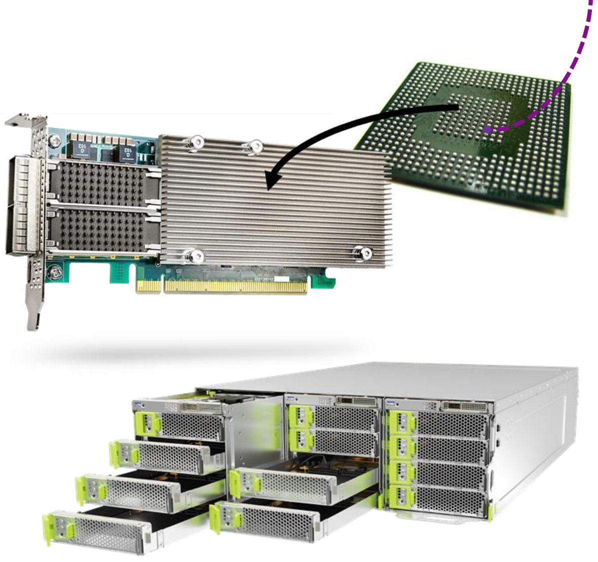
Certainty ⟷ ***Uncertainty***

Modelling spectrum

CEA spin-off to seize this opportunity



entropie



- ❑ A fabless probabilistic processor spin-off of the CEA
- ❑ To be created Q3 2026
- ❑ A solution for the latency constrained market of today and tomorrow

Thanks to all colleagues

CEA-List: *François Rummens, Thomas Mesquida, Marc Duranton, Anca Molnos, Phan-Anh Vu,, Antoine Reynaert, Manon Dampfhoffer*

CEA-Leti: *Elisa Vianello, Tifenn Hirtzlin, Louis Hutin, Djohan Bonnet, Niccolo Castellani, Kevin Garello,, Mathieu Faye, Emmanuel Hardy*

TDK corporation: *Shogo Yamada, Yukio Terasaki, Tatsuo Shibata, Tomoyuki Sasaki, Yukihiro Urakawa, Toshiki Gushi, Mai Takeuchi*

Prophesee: Christophe Posch, Amos Sironi, Damien Joubert, Cyril Soubeyrat

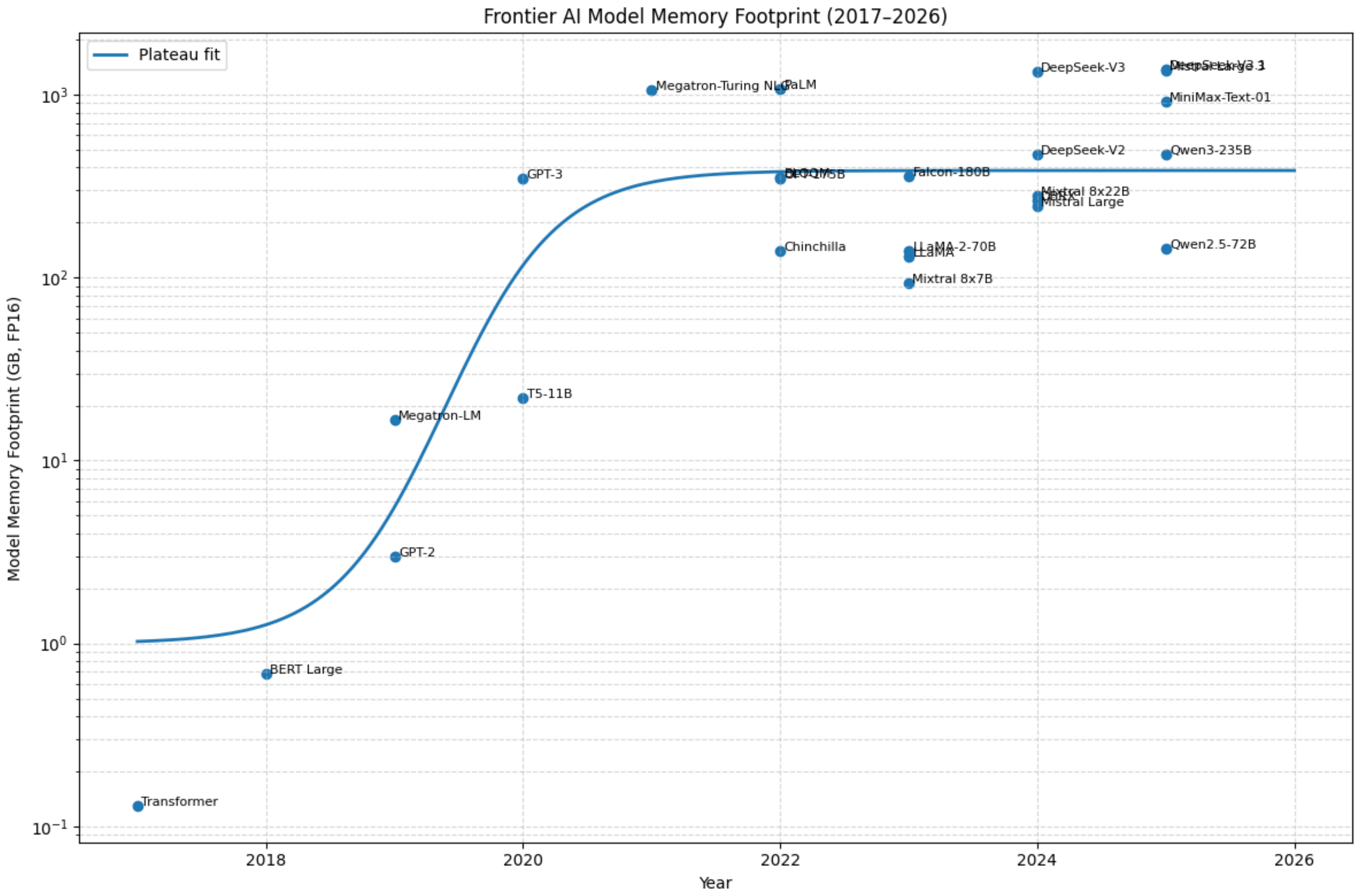
CEA-Neurospin: *Eiji Kawasaki, Florent Meyniel, Philippe Ciuciu*

C2N & IRT-nanolec: *Damien Querlioz, Kamel-Eddine Harabi, Clement Turck, Sebastian Thuries, Romain Lemaire, Laurent Alacoque*

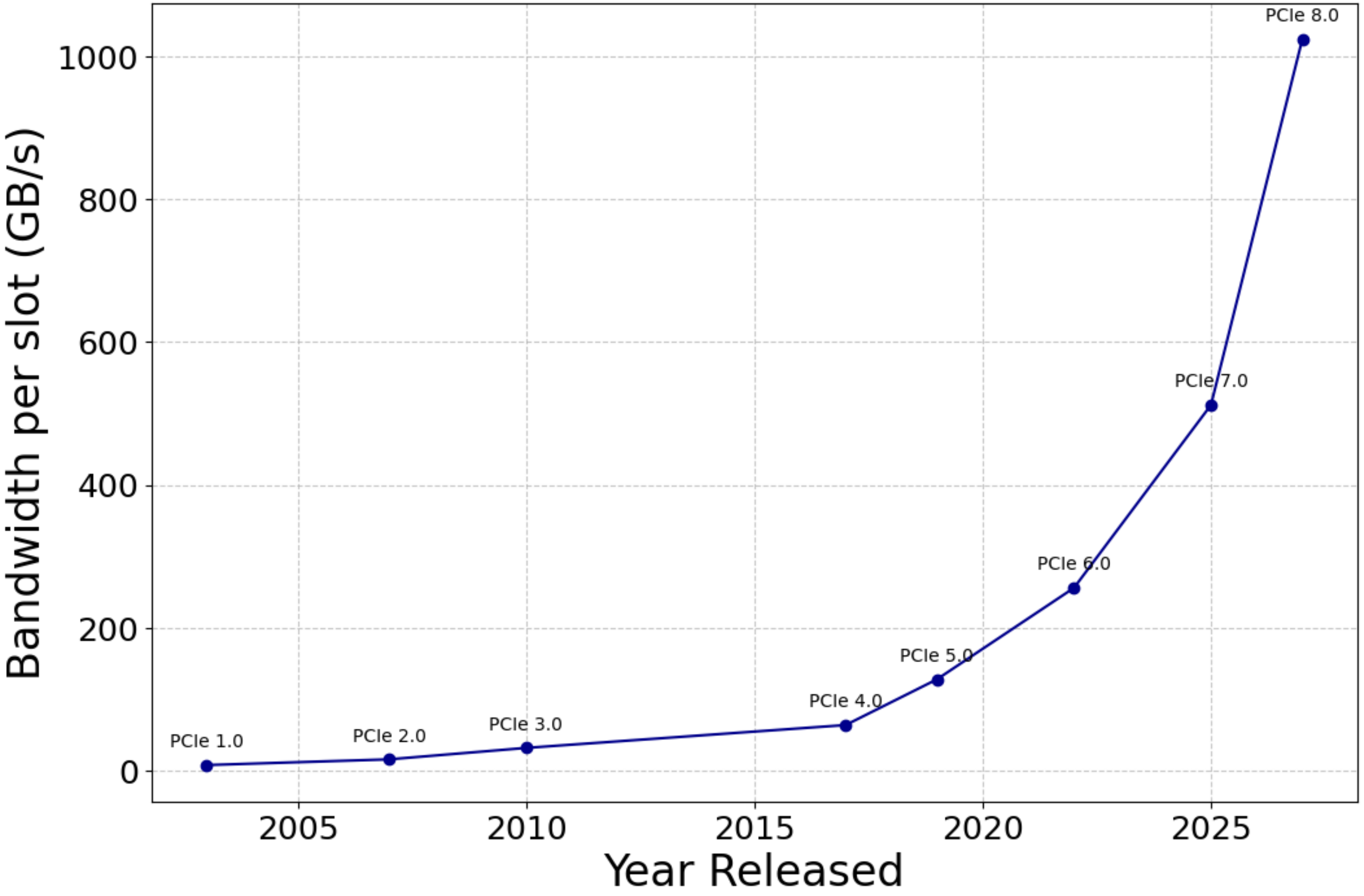
Keio & AGH-Krakow Universities: Hiroshi Nakano, Hiroaki Nishi, Krzysztof Blachut, Kamil Jeziorek, Piotr Wzorek, Tomasz Kryiak

Funding : *The TDK corporation, Prophesee, PEPR France Emergence, BlueSky digital brain, Horizon Europe dAIedge, KDT Edge-AI, ERC DIVERSE, PEPR BEP, IRT Smart Imager program*

Support slides

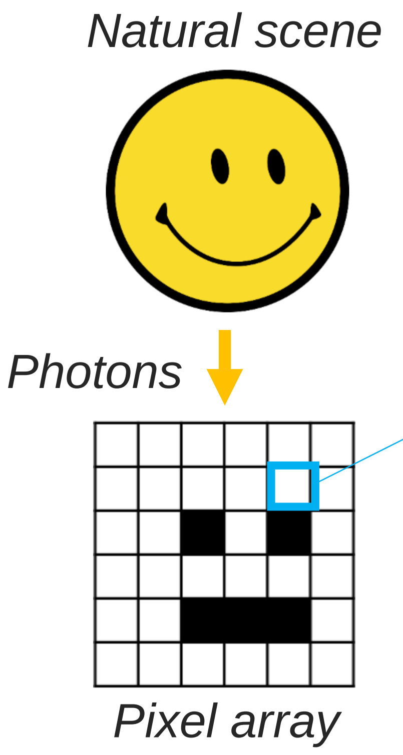

Frontier AI Model Memory Footprint (2017–2026)
Plateau fit
Model Memory Footprint (GB, FP16)
Year
10^{-1}
10^{0}
10^{1}
10^{2}
10^{3}
2018
2020
2022
2024
2026
Transformer
BERT Large
GPT-2
Megatron-LM
T5-11B
GPT-3
Megatron-Turing NLG
PaLM
BLOOM
OPT-175B
Chinchilla
Falcon-180B
LLaMA-2-70B
LLaMA
Mixtral 8x7B
DeepSeek-V3
DeepSeek-V2
Mixtral 8x22B
DBRX
Mistral Large
DeepSeek-V3.1
Mistral Large 3
MiniMax-Text-01
Qwen3-235B
Qwen2.5-72B

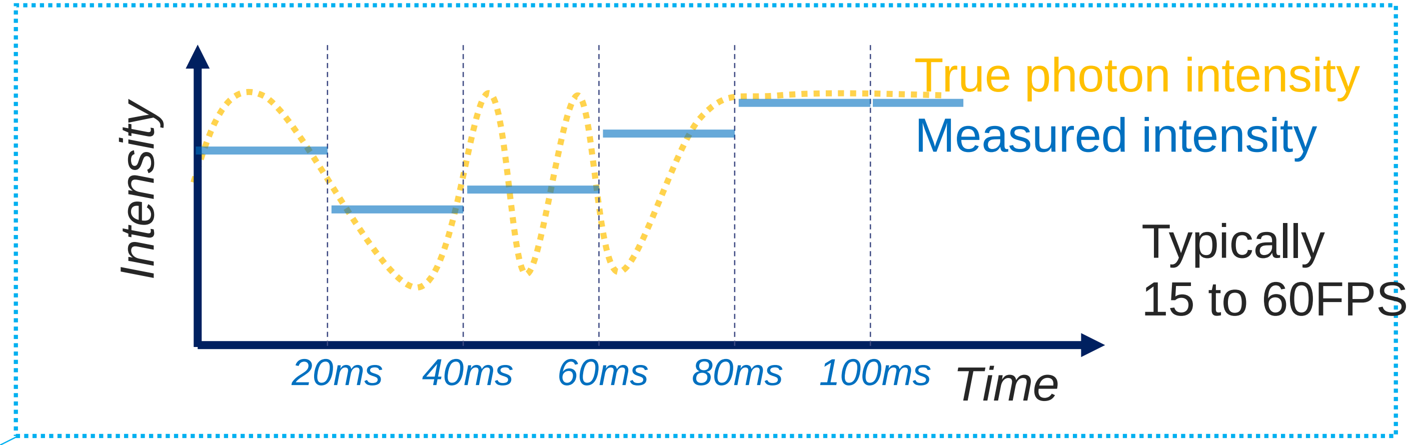
PCIe scaling



Introduction to event-based sensing

Frame-based camera principle

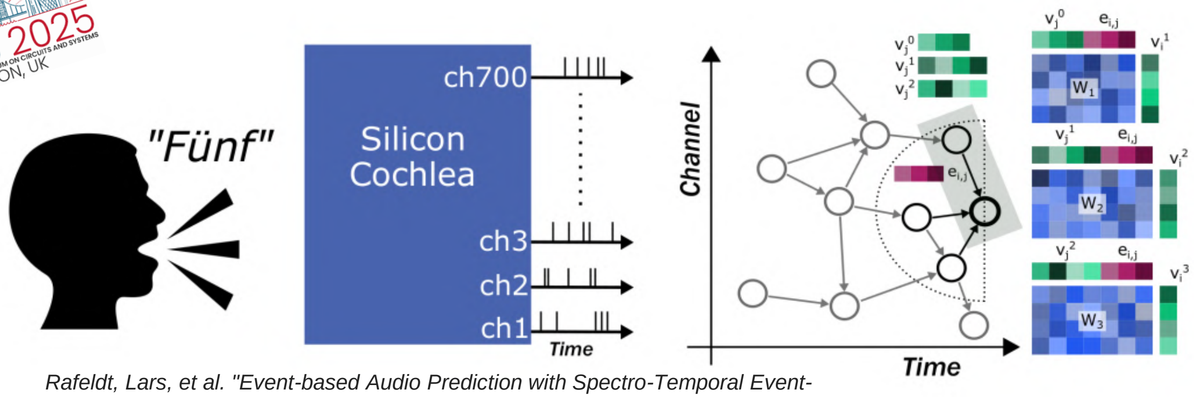




At regular exposure intervals, record the average light intensity at each pixel

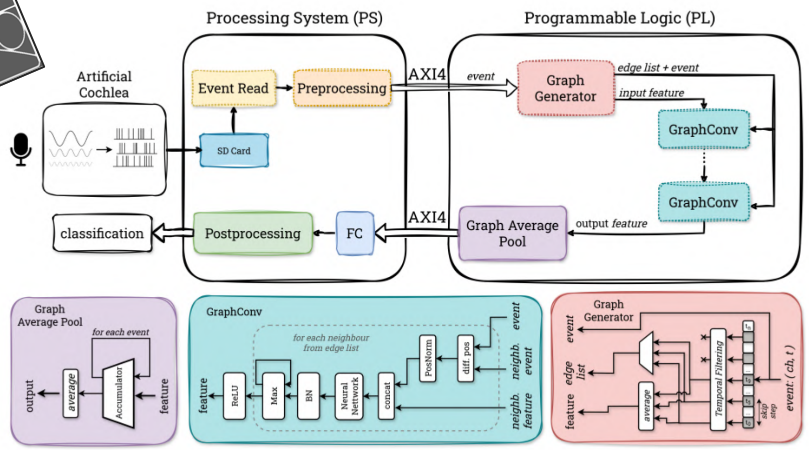
- Fine temporal detail lost between frames
- Redundant measurement in absence of change

Event-graphs for audio classification



Rafeldt, Lars, et al. "Event-based Audio Prediction with Spectro-Temporal Event-Graphs." 2025 IEEE International Symposium on Circuits and Systems (ISCAS). IEEE, 2025.

- ❑ *Classifies keywords better than a Spiking Neural Network* ***with >10x less weight parameters***

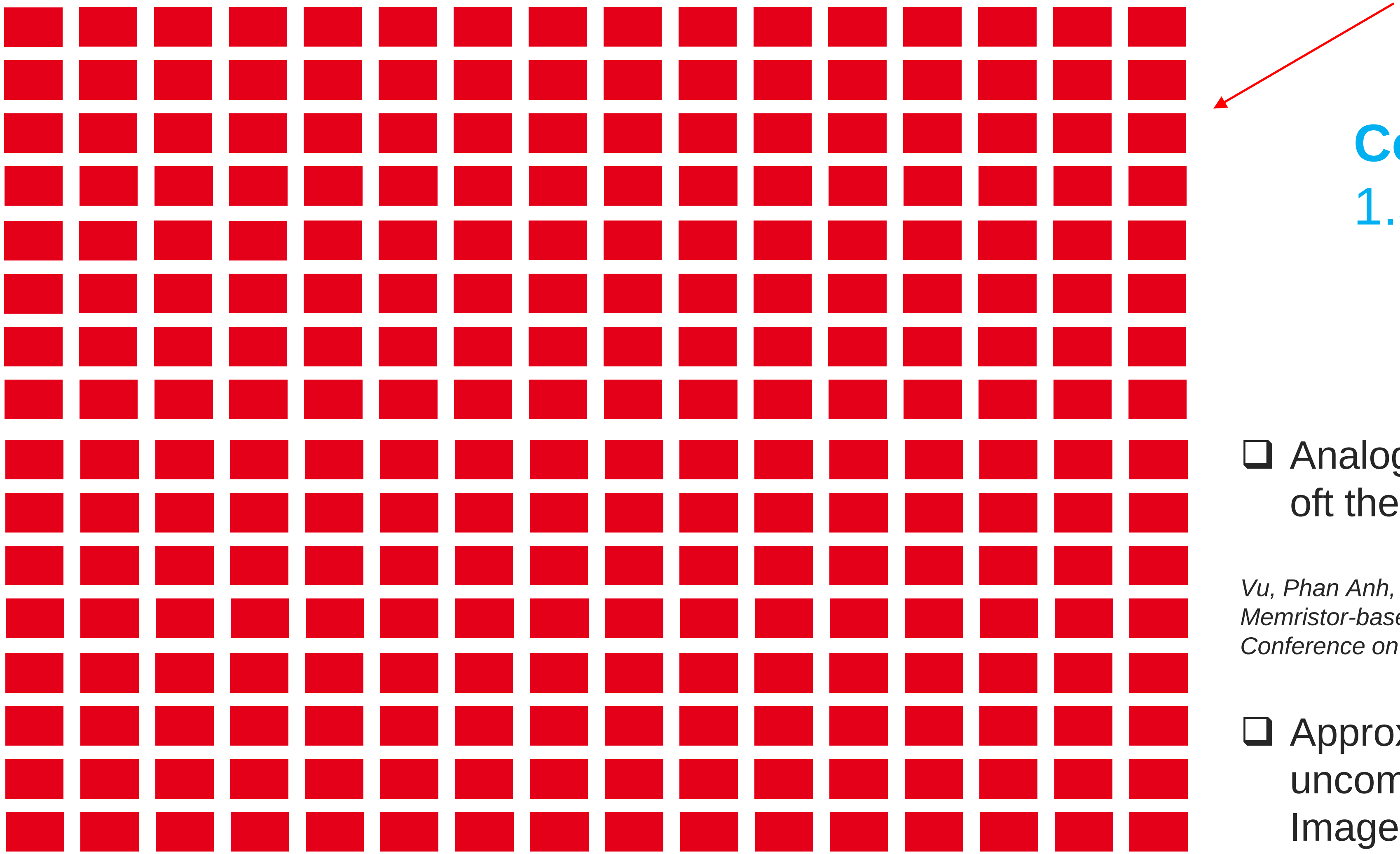


Nakano, Hiroshi, et al. "Hardware-accelerated event-graph neural networks for low-latency time-series classification on soc fpga." International Symposium on Applied Reconfigurable Computing. Cham: Springer Nature Switzerland, 2025.

- ❑ *FPGA implementation*
  - ❑ *Event-graph* ***20% higher accuracy*** *(92% vs. 72%) than FPGA SNN*
  - ❑ *Event-graph is* ***70x lower latency*** *than SNN (8µs vs. 540µs)*

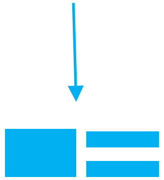
Compressing Bayesian in-memory computing

***Scenario**: Bayesian Transformer layer of size 1024 neurons and 256 samples*



Uncompressed
537 million ANVM

Compressed
1.57 million Memristors



- ❑ Analogue MAC in a compressed space oft the weights

Vu, Phan Anh, et al. "Compressed vector-matrix multiplication for Memristor-based ensemble neural networks." 2024 IEEE International Conference on Rebooting Computing (ICRC). IEEE, 2024.

- ❑ Approximately same performance as uncompressed Bayesian ViT on ImageNet