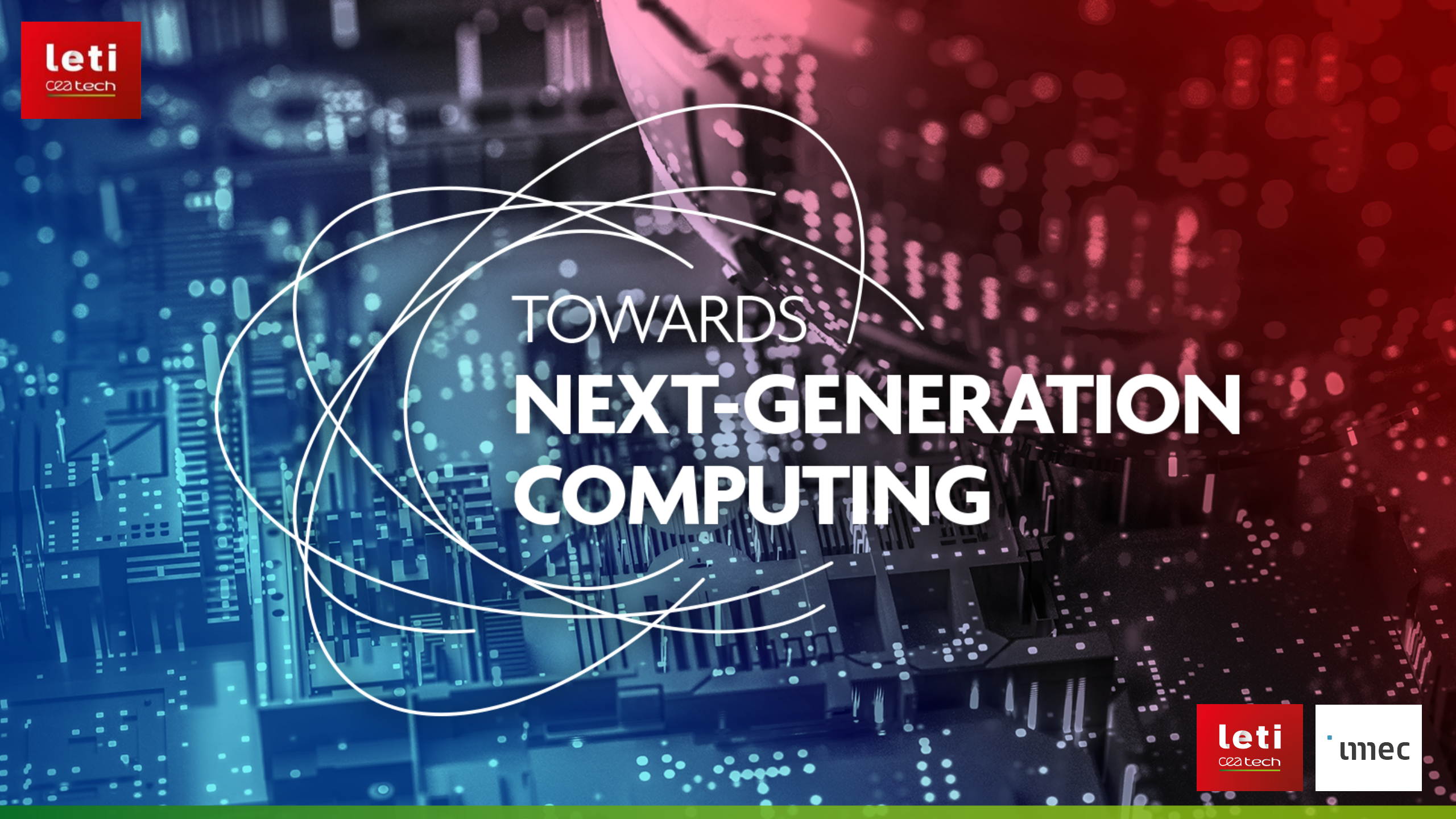
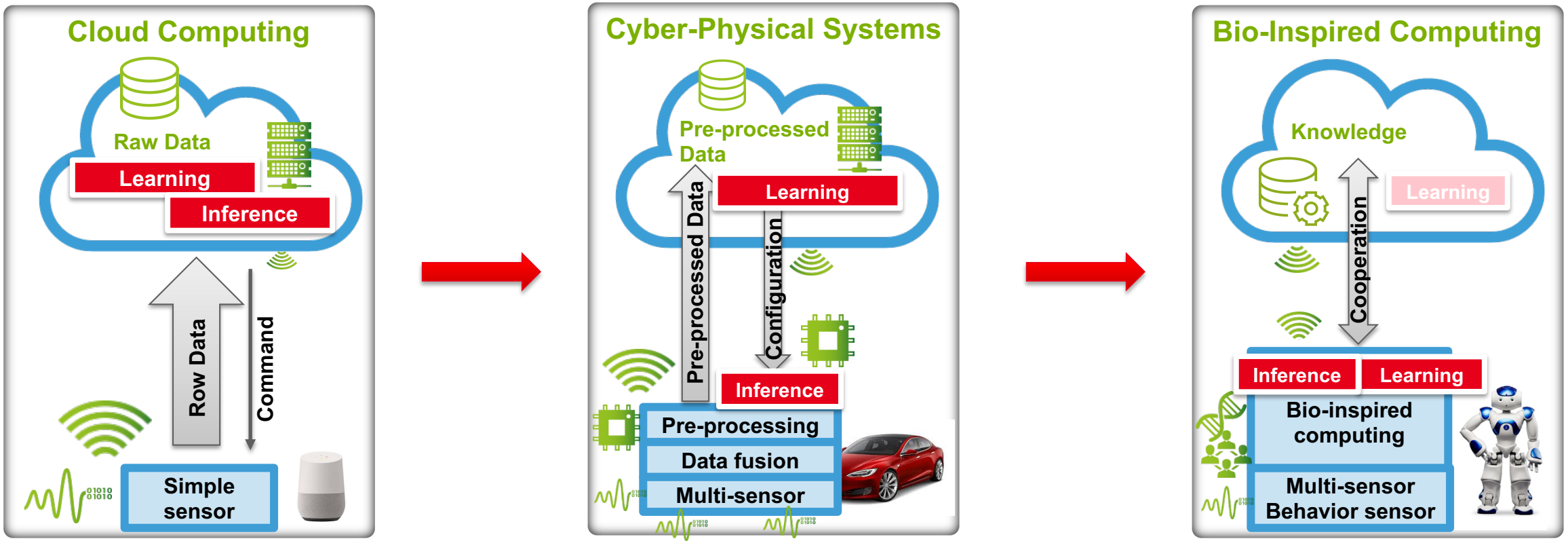

leti
ceatech
TOWARDS
NEXT-GENERATION
COMPUTING
leti
ceatech
imec

TECHNOLOGICAL SOLUTIONS FOR IMPROVING THE FIGURES OF MERIT OF EDGE AI APPLICATIONS

Alexandre VALENTIAN

TRENDS IN AI COMPUTING

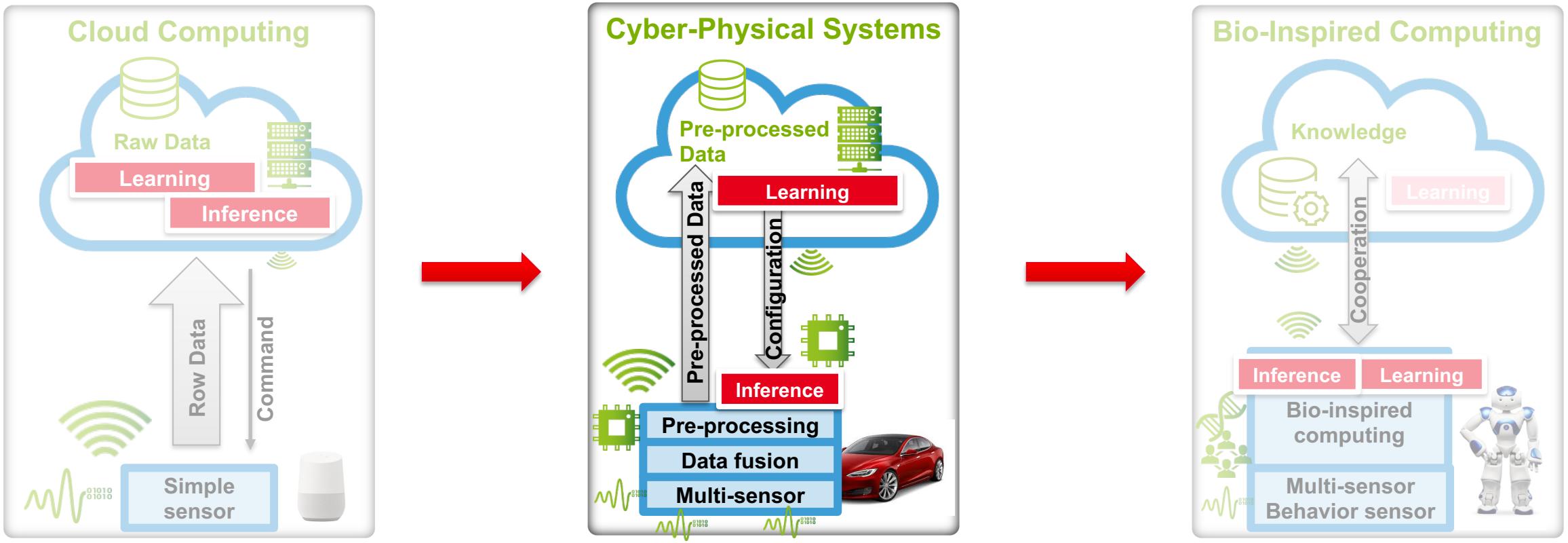


2012 — AlexNet

2019 — Edge TPU, Movidius …

2025 — ?

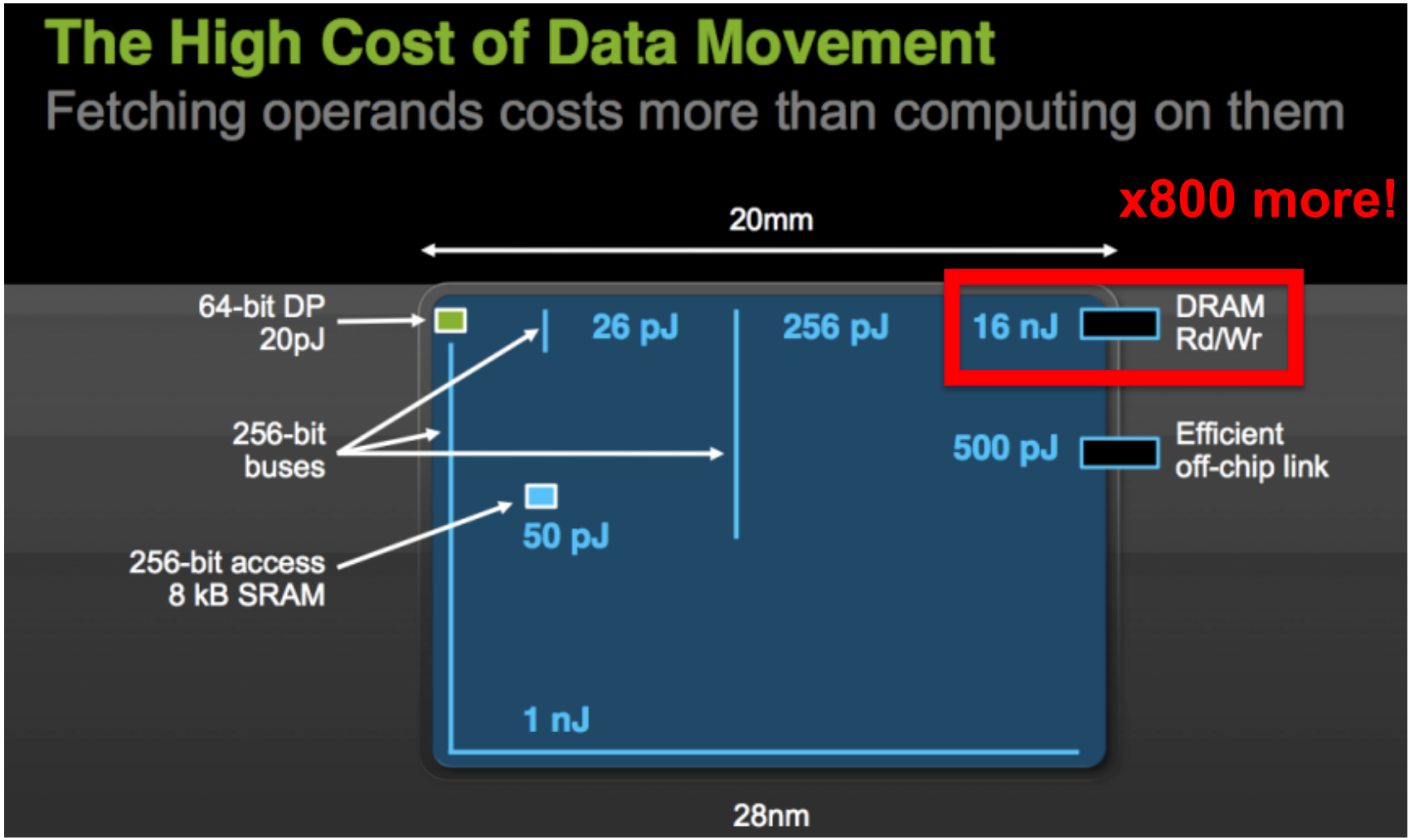
TRENDS IN AI COMPUTING



2012 — **2019** — **2025**

AlexNet — **Edge TPU, Movidius …** — **?**

SOLVING THE ENERGY CHALLENGE: COST OF MOVING DATA



Bill Dally, *"To ExaScale and Beyond"*, 2010

TRENDS IN EDGE COMPUTING

Increased computing efficiency

Weight quantization

Reduced bit accuracy

- Smaller memory footprint
- Lighter operations

Variable bit precision

Handling higher bit accuracy when needed

- For higher inference precision

Sparsity

Skip MAC operations

- When weight or intermediate result is 0

Increased storage efficiency

Near memory computing

Avoid external memory accesses

Weights

- Embedded Non-Volatile Memory

Intermediate results

- SRAM or Embedded DRAM

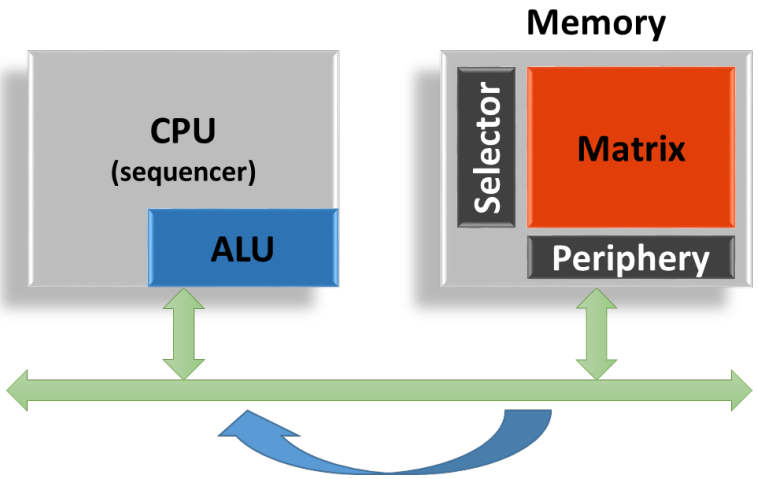
In-Memory computing

SRAM or Embedded NVM

Digital or analog

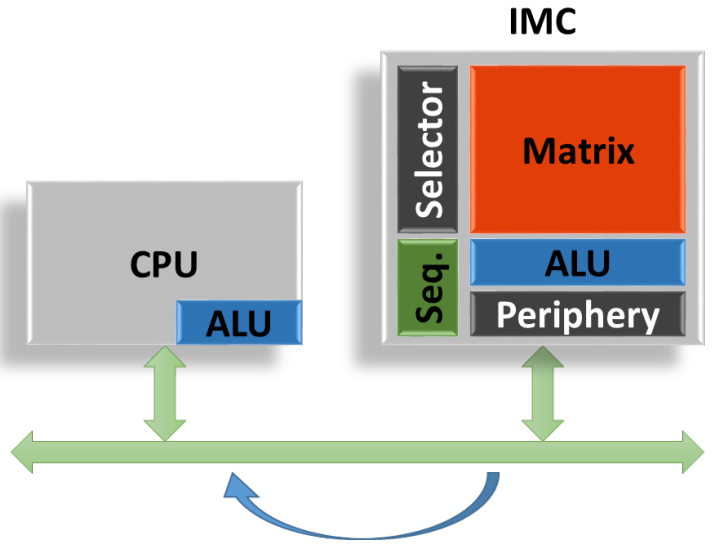
ACTIVATIONS – IN-MEMORY COMPUTING

Von Neumann architecture



High data transfer

In-Memory Computing (IMC) architecture



Low data transfer

Matrix Multiplication

	Clock Cycle	Energy
Matrix Size	Scalar vs. CSRAM	Scalar vs. CSRAM
8x8 8-bit integer (512 bits)	x256	X9

WEIGHTS – NON-VOLATILE MEMORIES

Multi valued RRAM technology compatible with advanced logic

- **Cell density of 40F² is available at 28nm**
- **And compatible to sub-20nm**

Roadmap from single bit to multi-bit* and beyond

- **From 1 bit to 4 bit and beyond**

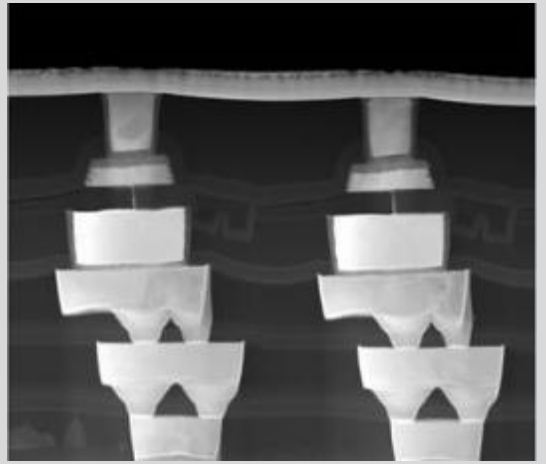
** [T. Wu, ISSCC 2019]*

Roadmap for increasing embedded cell density

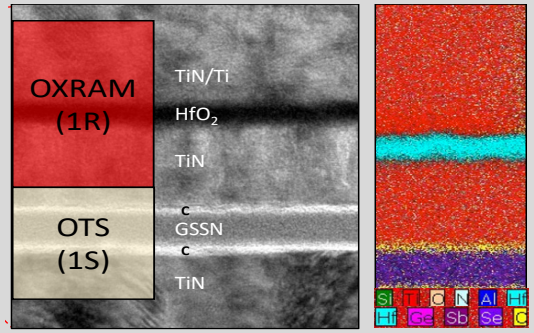
- **From 40F² down to 4F² thanks to new selector technology**

Examples with technology available today

- **ResNet50 (74 MB of weights) → 15 mm² of memory**
- **YoloV3 (101MB of weights) → 20 mm² of memory**

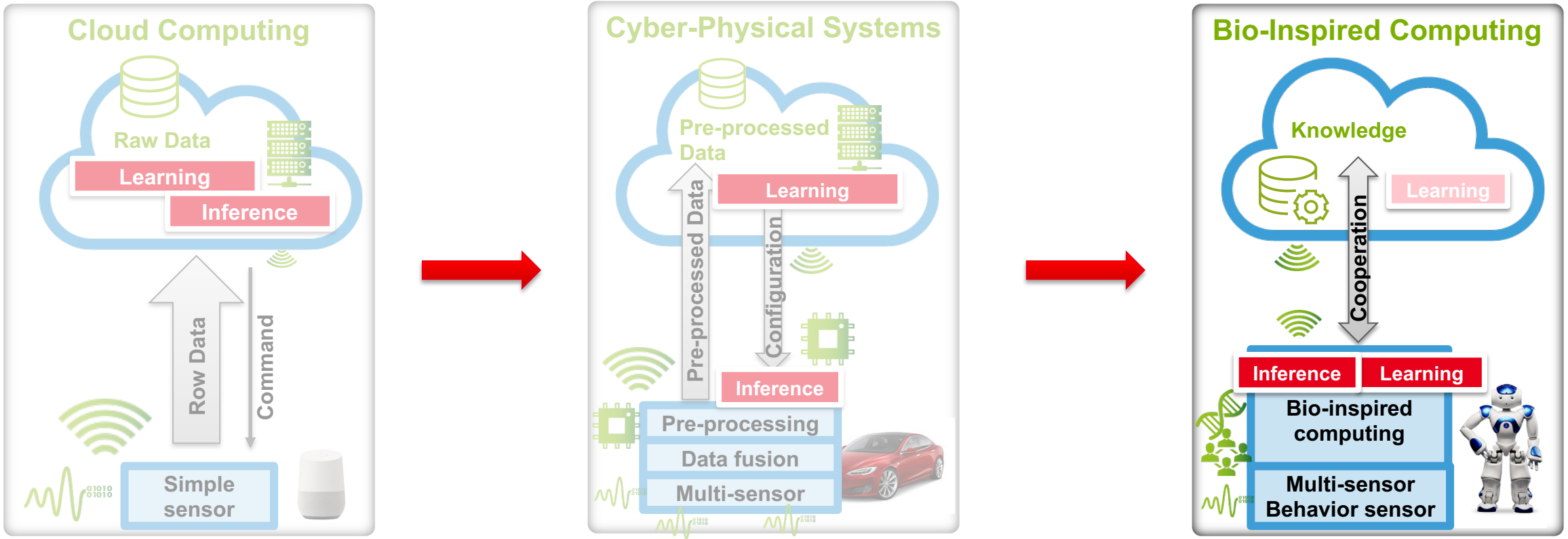


28nm RRAM integration



Selector and RRAM integration
[IEDM 2019]

TRENDS IN AI COMPUTING



2012 — AlexNet

2019 — Edge TPU, Movidius …

2025 — ?

CHALLENGE OF ONLINE LEARNING

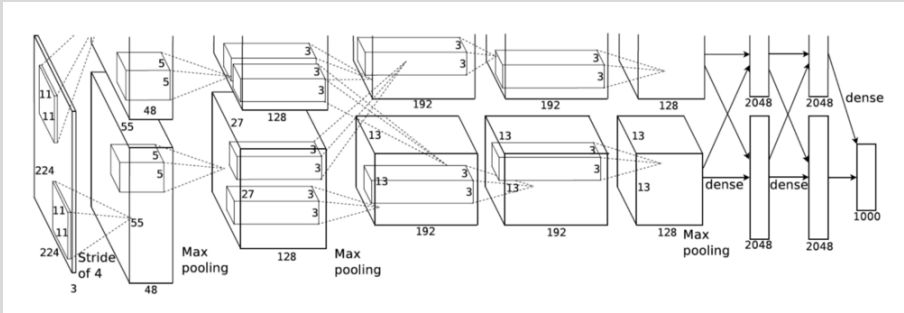
Back-propagation algorithm

- **Necessitates to keep all intermediate results (activations)**
- **With a batch size of more than one**
  - To not cycle too much the non-volatile memories

This requires a tremendous amount of activation memory

- **Example – YoloV3**
  - A batch of 20 images requires 800MB of memory

Feedforward



Back-propagation

TECHNOLOGY SOLUTION – 3D INTEGRATION

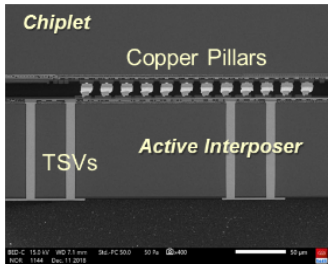
Advantages

- Increasing computing & memory capacity

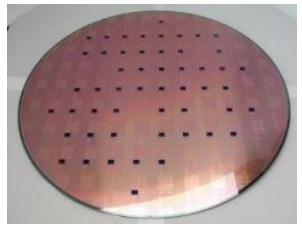
Trends

- « Denser Integration » : tight memory ⇔ logic computing paradigm
- « Chipletization » : Generic computing templates, Heterogeneous technologies

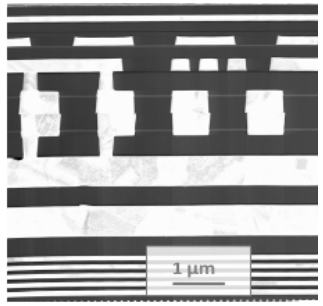
Roadmap



TSV & µ-bumps, pitch 20 µm
[P. Coudrain, ECTC 2019]
[P. Vivet, ISSCC 2020]



Hybrid Bonding
CoW, pitch 3-5 µm
Including K-G-D
[A. Jouve, 3DIC 2019]



Hybrid Bonding
WoW, pitch 1.44 µm
[A. Jourdan, IEDM 2018]

20 µm — 5 µm — 1 µm

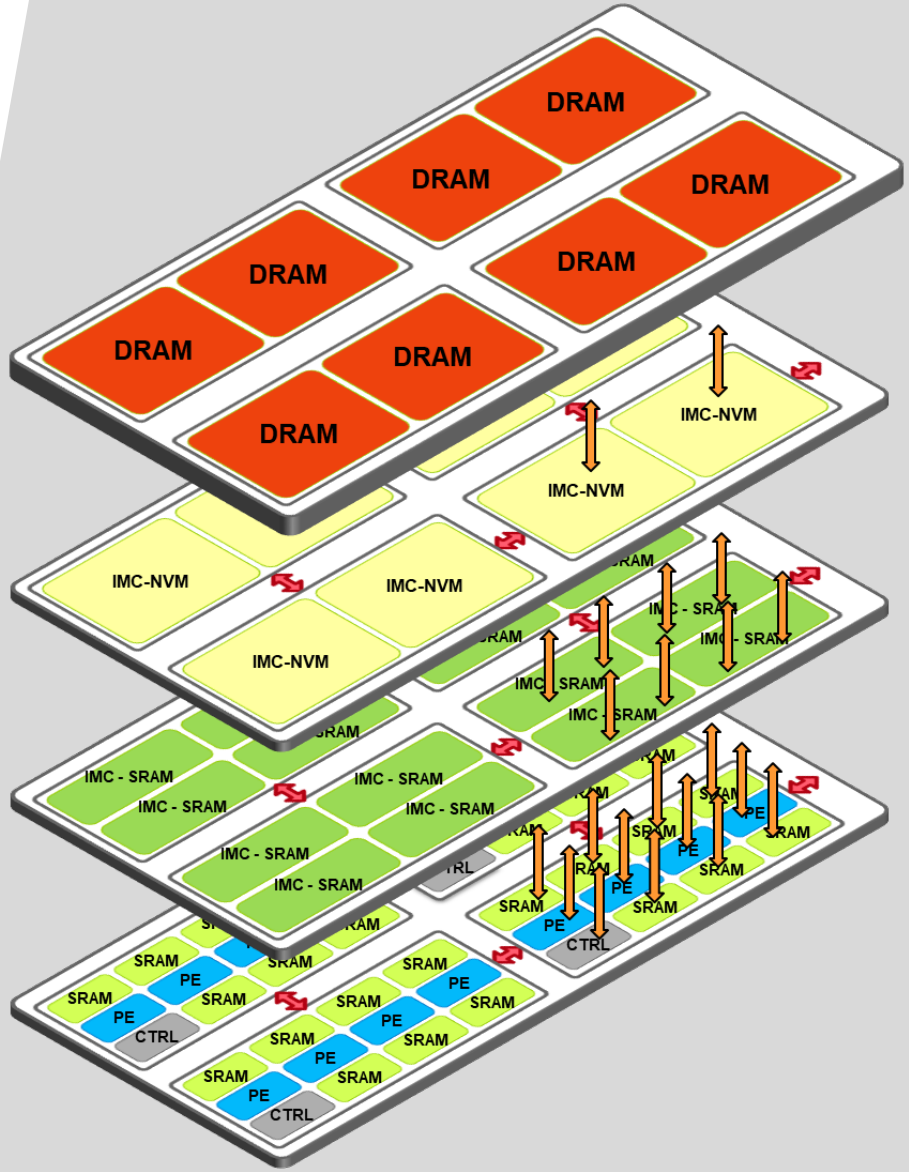
DISTRIBUTED MEMORY-CENTRIC EDGE AI COMPUTING ARCHITECTURE

Memory-Centric architecture

- No more global buffers
- No more power-hungry caches
- Fully distributed memory and control
- Energy efficient use of memory using 3D technology

Edge AI architecture, using

- Generic PE engines
- Vertically and horizontally connected computing clusters
- In-Memory Computing tiles (IMC)
- Dense NVM for storage
- DRAM for online learning



CONCLUSION

Trends in Edge AI applications

- Inference first
- Then lifelong local learning

Main challenge is to reduce data movement

This can be solved thanks to a combination of architecture and technology

- Combination of In-Memory Computing
- Non-volatile memory for synaptic weights
- 3D technology for heterogeneous integration

LETI is working to advance those technologies

THANK YOU FOR YOUR ATTENTION