Frugal Artificial Intelligence for Edge Devices

Thomas Signamarcheix

Vice President Strategic Development, CEA-Leti

The Challenge: Capitalize on hardware and software advances to master global digitization and preserve the planet

2010-2020:

› A data-driven era

Internet users

+125%

World population

+10%

Mobile data

+20316%

Internet traffic

+1170%

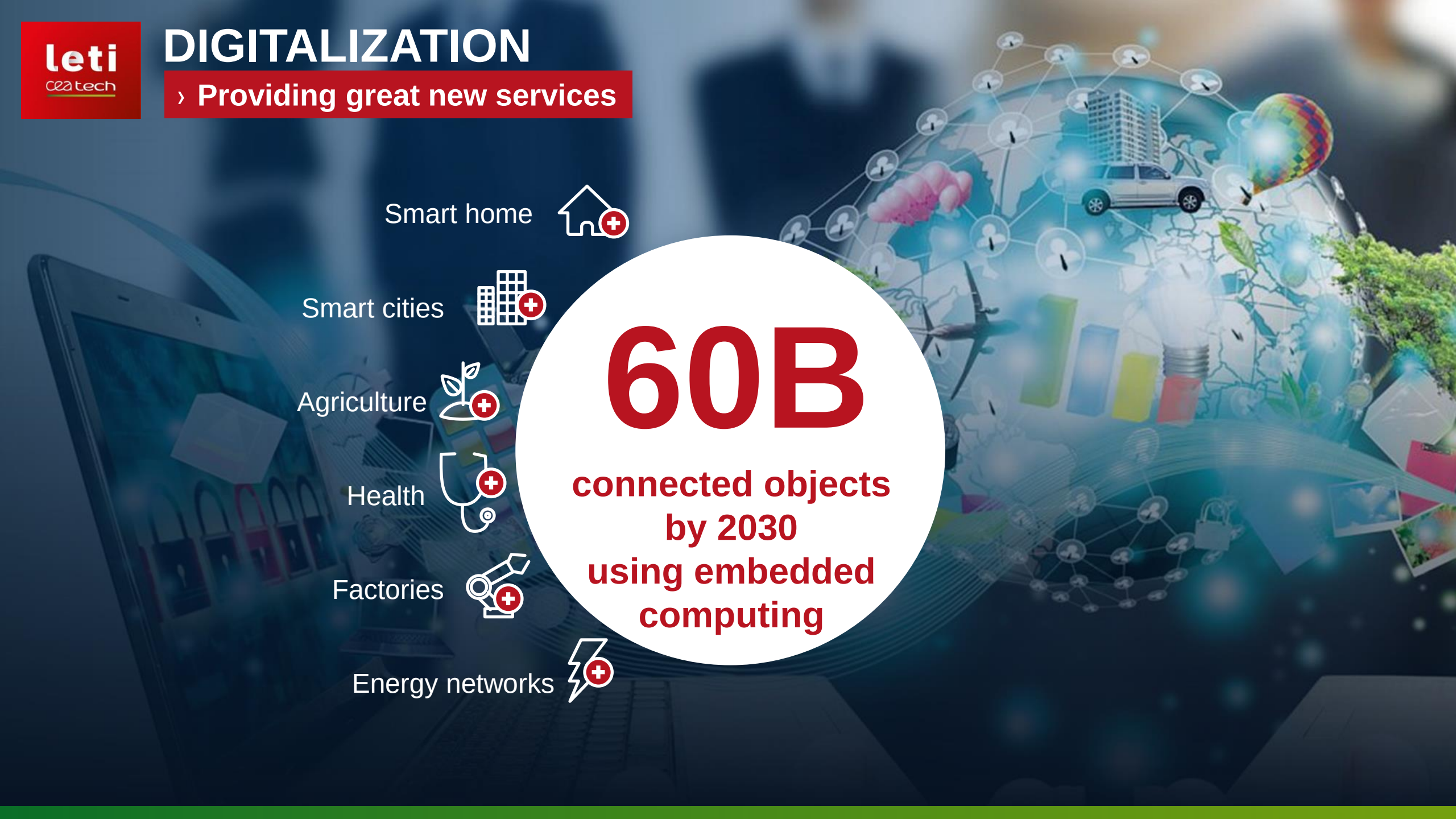
Electricity

+22%

Sources: Worldometers 2021, enerdata 2021, ourworldindata.org
Internet 2021, Cisco Visual Networking Index 2011 & 2020, Ericsson Mobility report 2021

DIGITALIZATION

› Providing great new services



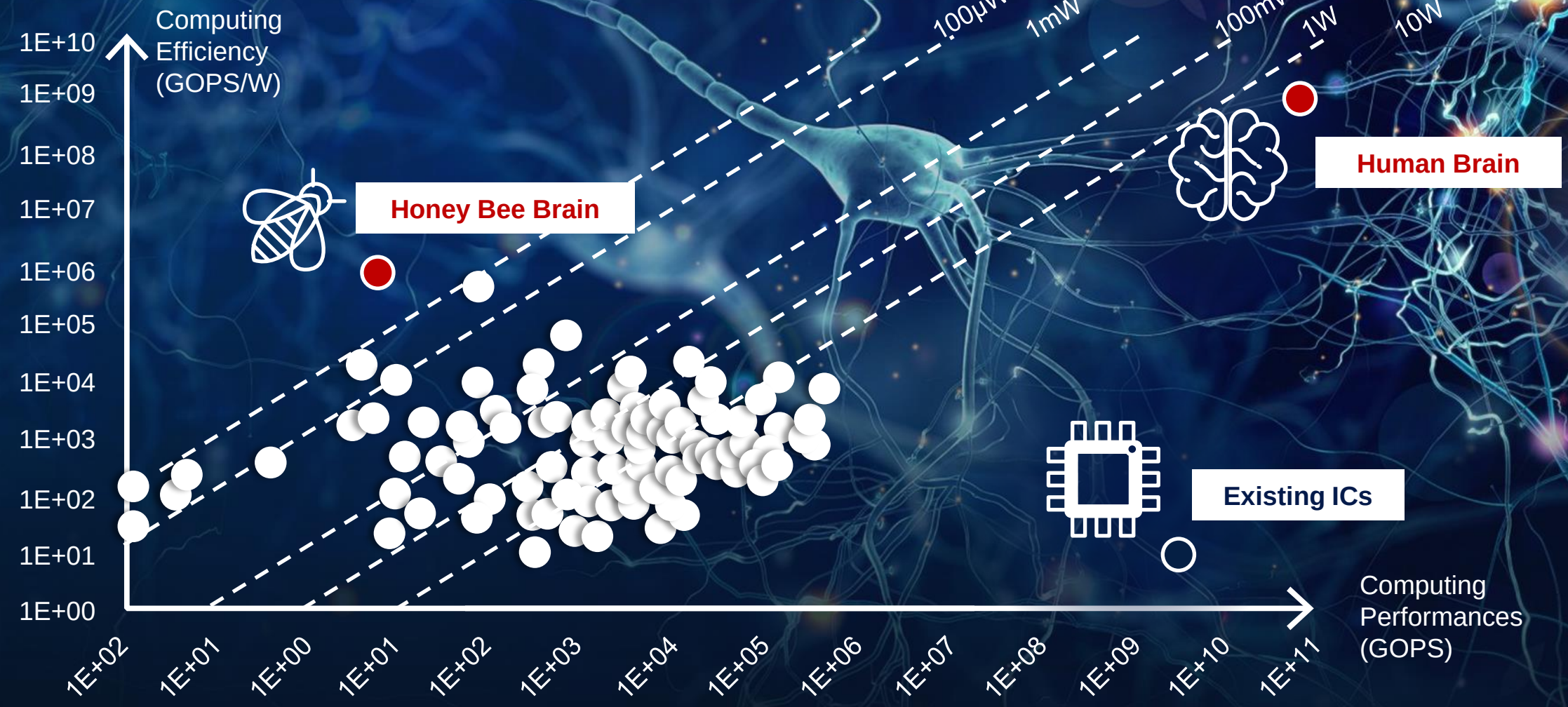
CHALLENGE

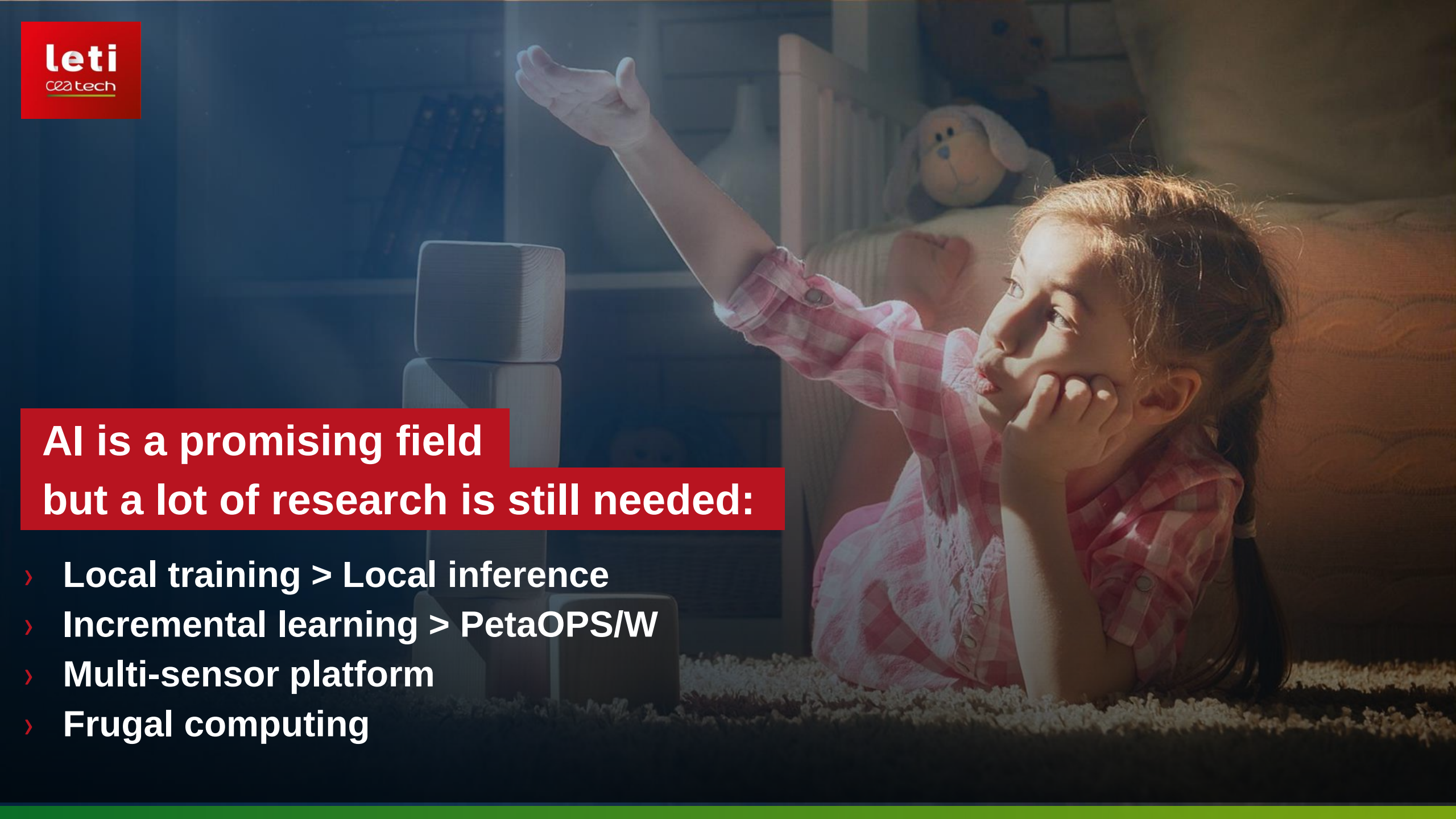
INNOVATION WILL BE KEY FOR ENVIRONMENTAL SUSTAINABILITY


leti
ceatech
Committed to innovation,
CEA-Leti’s dedicated teams
pioneer micro-nanotechnologies
enabling smart, energy-efficient
and secure solutions for industry

NATURE

› A source of inspiration for semiconductors





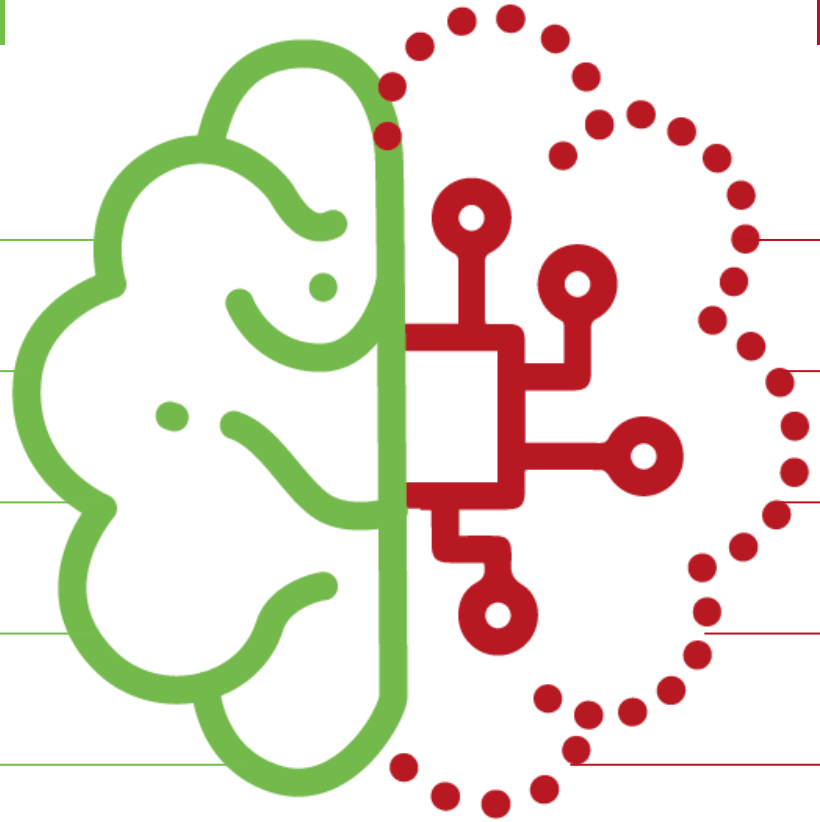
AI is a promising field but a lot of research is still needed:

- Local training > Local inference
- Incremental learning > PetaOPS/W
- Multi-sensor platform
- Frugal computing

FILLING THE GAP WITH NATURAL INTELLIGENCE

› Bio inspiration

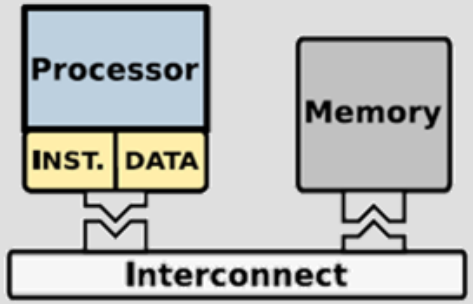
Biology	Technology choices
Asynchronous communication	Spike coding
Plasticity	Re-configurability (routing)
Sensing	Smart sensors
Brain is not flat, it is 3D	Dense 3D integration
Lifelong learning	New technologies coupled with algorithms



RESISTIVE MEMORIES

› Efficient back-end of line implementation

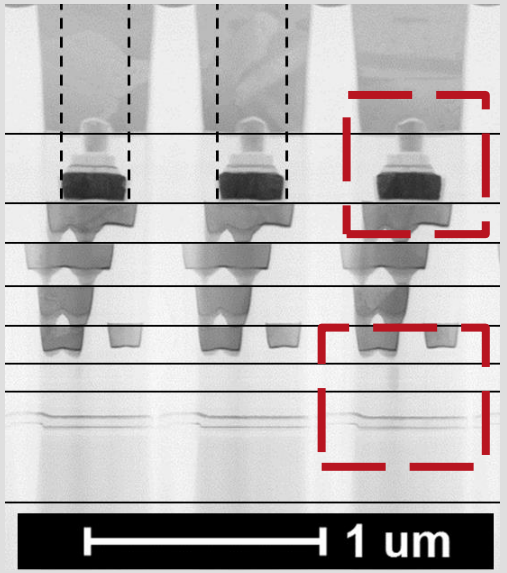
How data consumes energy



Operation	Energy
Addition of data (fixed point)	1×
Accessing data (onchip cache)	60×
Accessing data (offchip RAM)	3500×

Data movement between
storage and processing units can reach
90% of the overall energy consumption

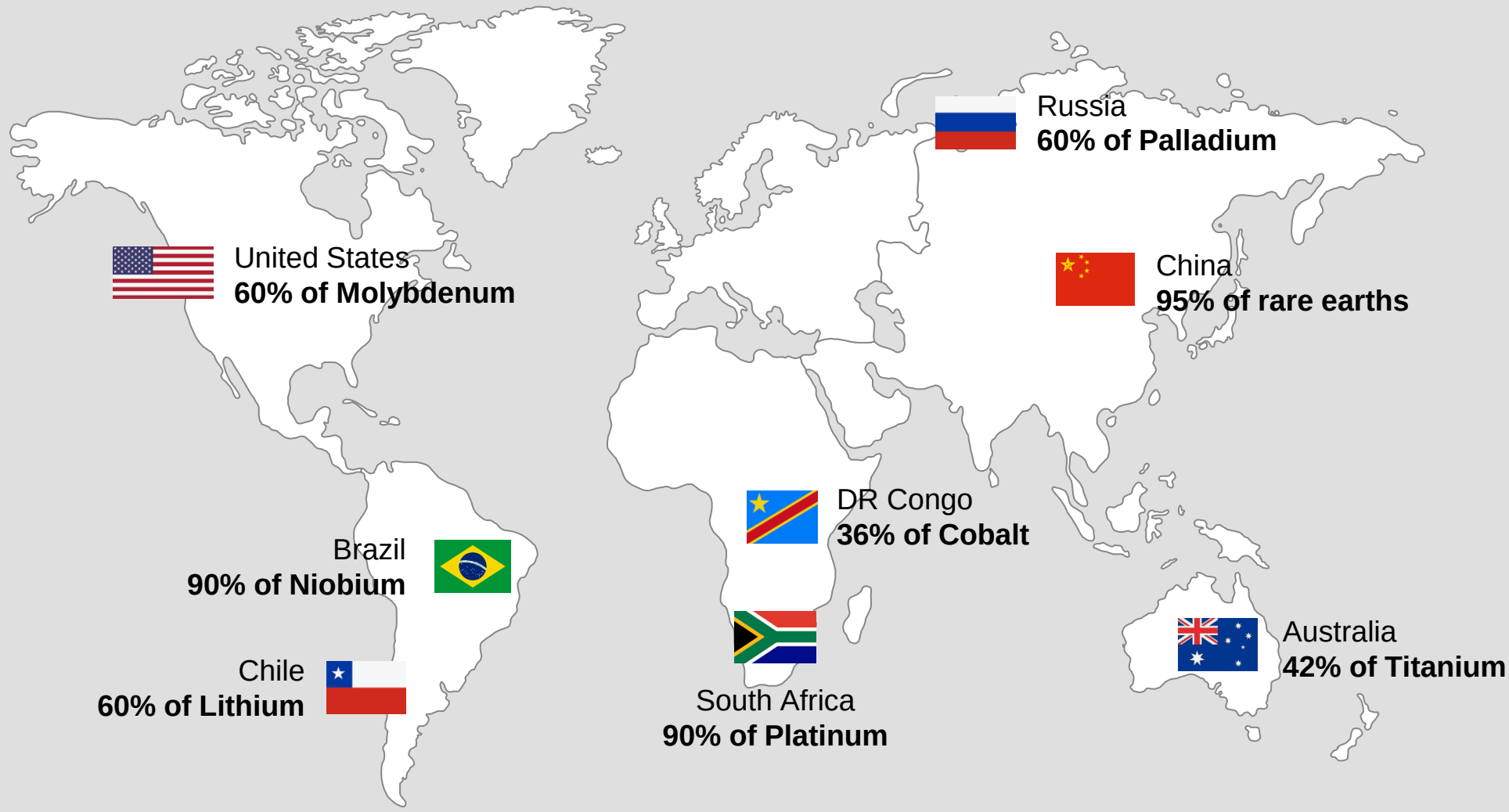
Non-volatile memories

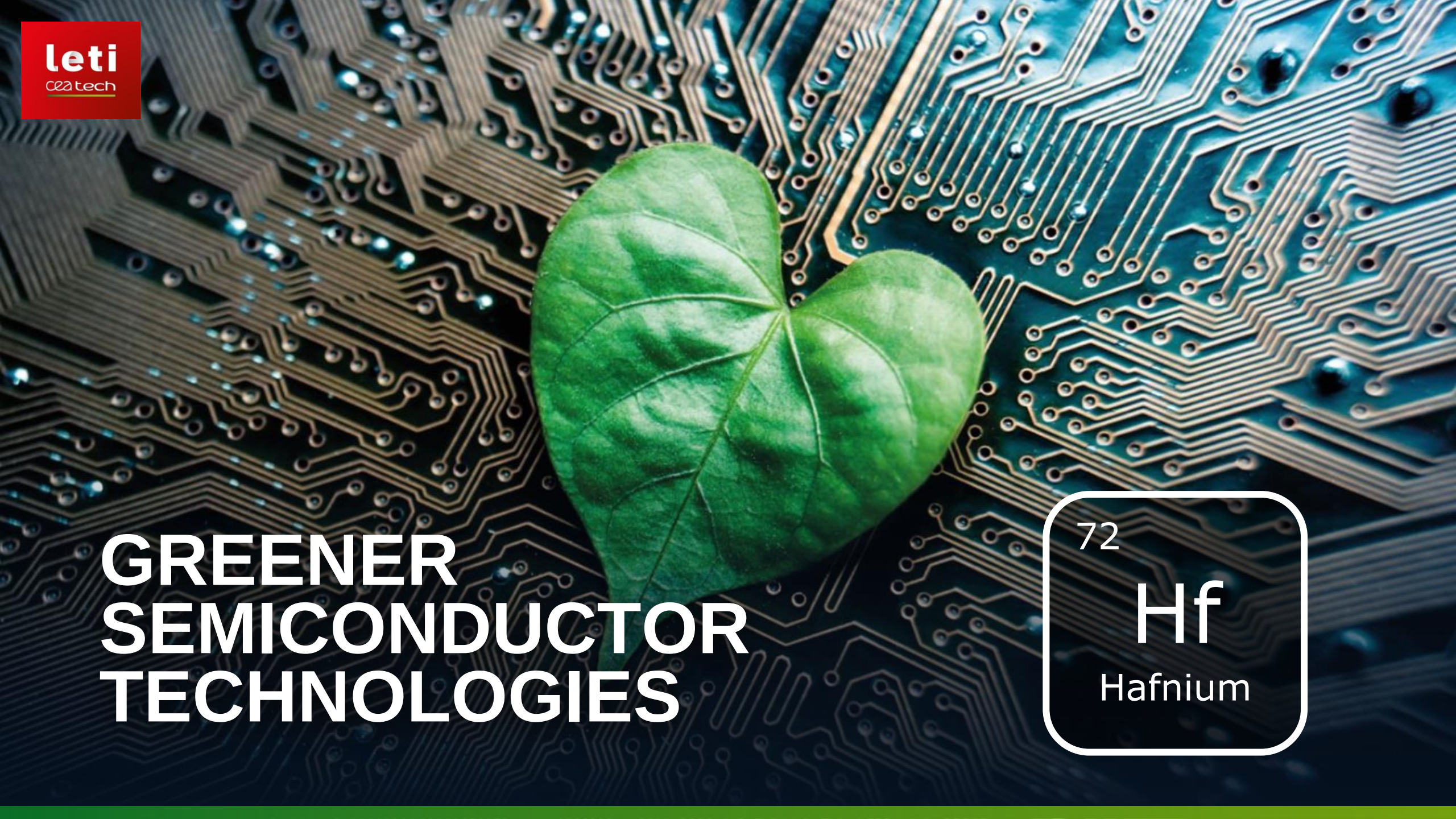


Need for high density on-chip resistive memories

RARE EARTHS AND MINERALS

› A small number of countries control the production




leti
ceatech
GREENER
SEMICONDUCTOR
TECHNOLOGIES
72
Hf
Hafnium

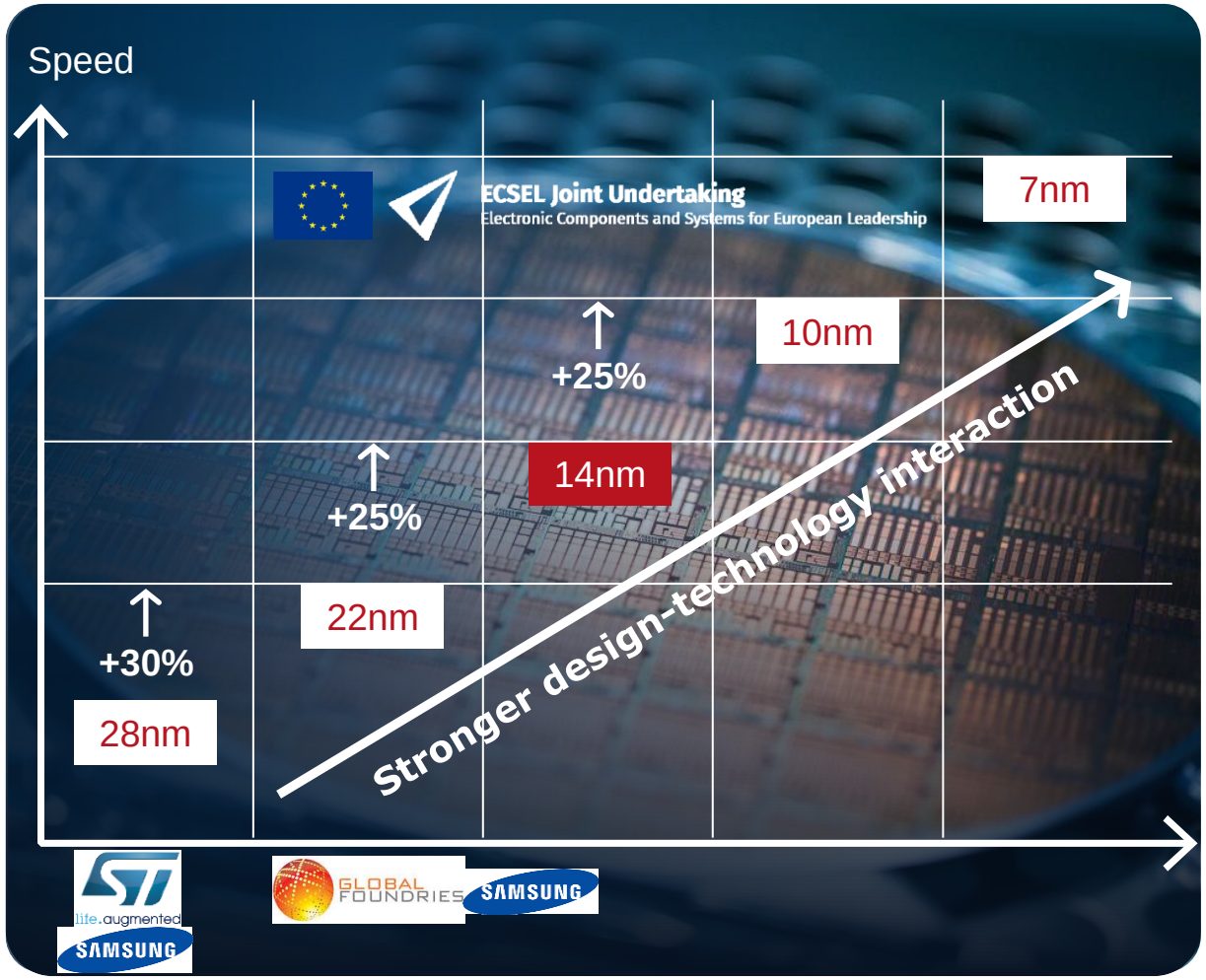
DIFFERENT TYPES OF MEMORIES

Programming power **reduction ×20,000**

	FLASH	ReRAM (HfO_2)	FeRAM (HfO_2)
Programming power	**~200pJ/bit**	~100pJ/bit	**~10fJ/bit**
Write speed	20 µs	10-100 ns	14ns @ 2.5V
Endurance	10^5 - 10^6	10^5 - 10^6	**> 10^{11}**
Retention	> 125°C	> 125°C	85°C
Extra masks	Very high (>10)	Low (2)	Low (2)

FULLY DEPLETED SOI

› Unic technology for edge devices

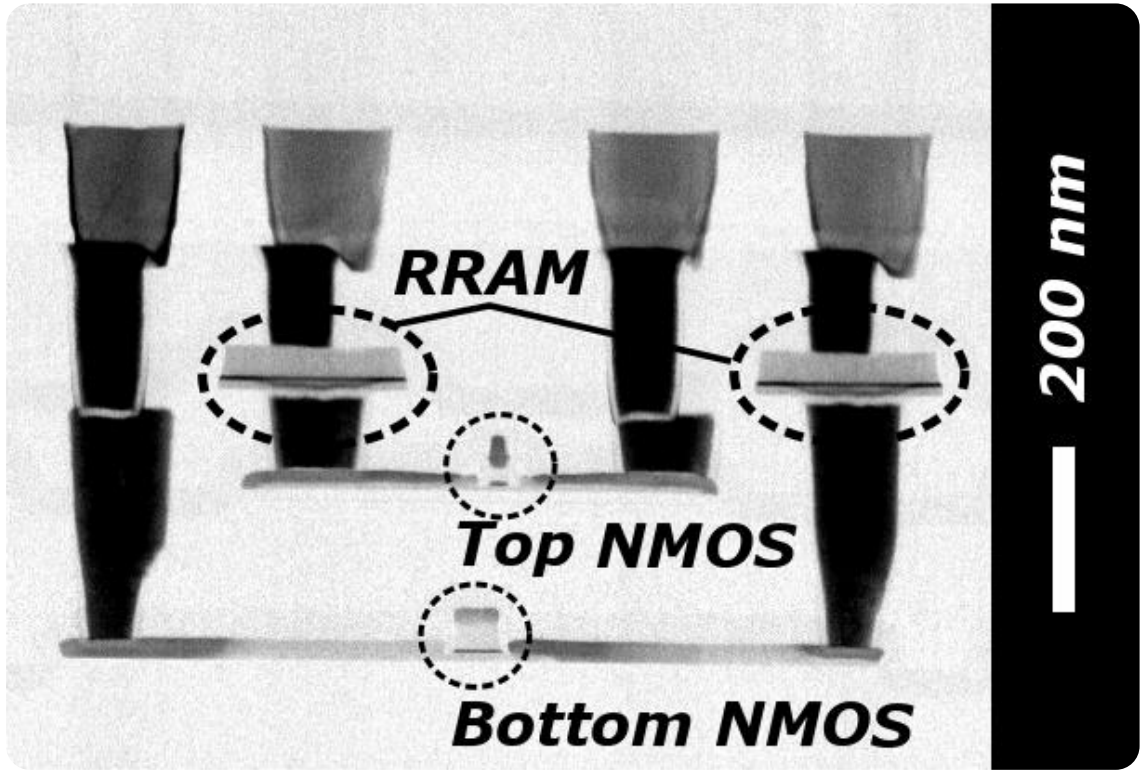


Scaling the FD-SOI technology is becoming indispensable

- ultra-low power IoT devices,
- automotive,
- RF,
- Edge AI,
- 5G-6G

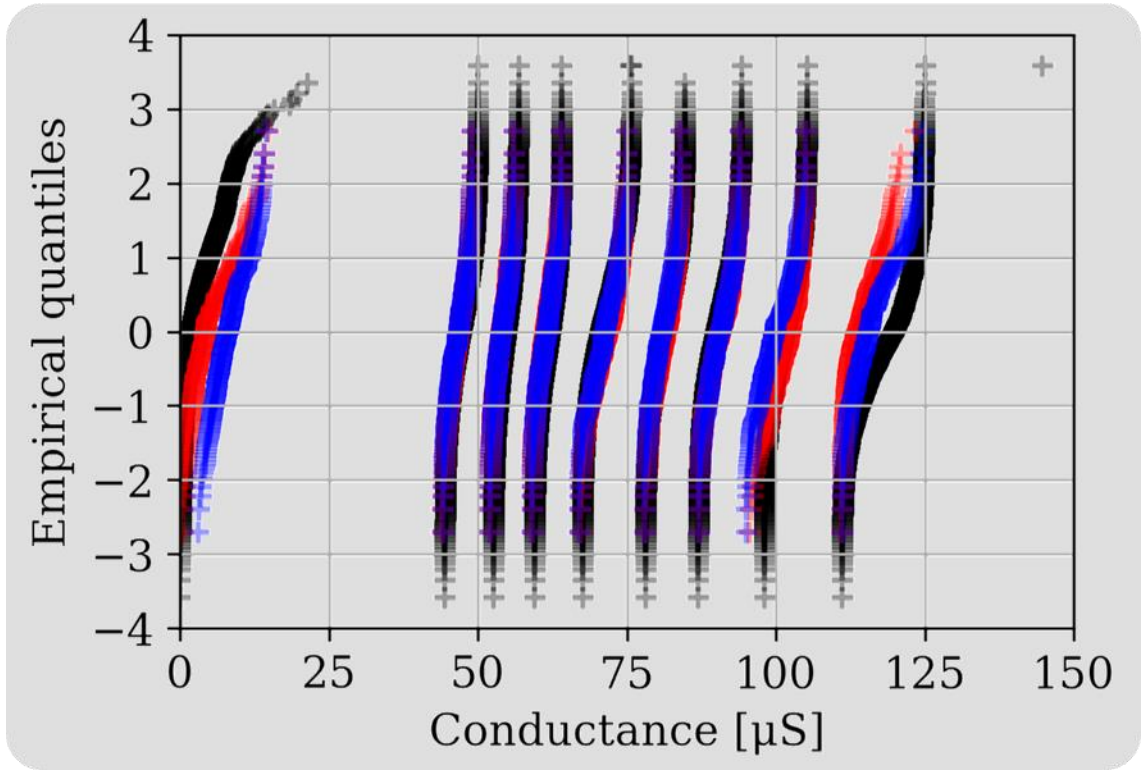
RESISTIVE MEMORIES

› Analog behavior of resistive RAM



Multiple 1T1R

1.5 × area gain

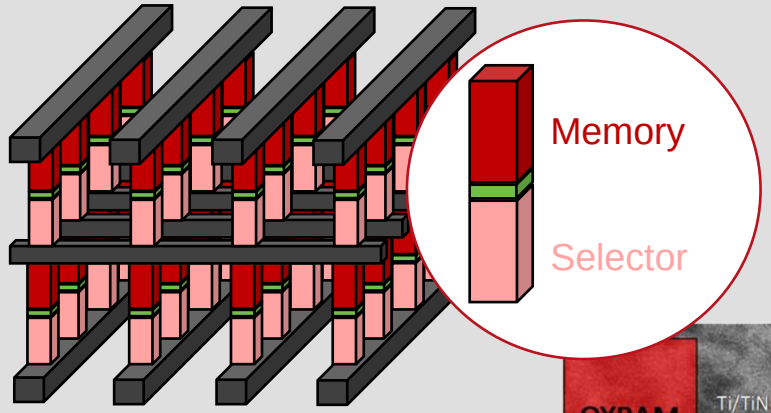


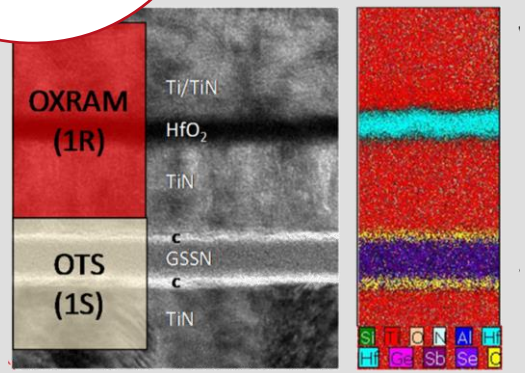
Multi-Level-Cell

3.17 bit per RRAM

HIGH DENSITY MEMORIES

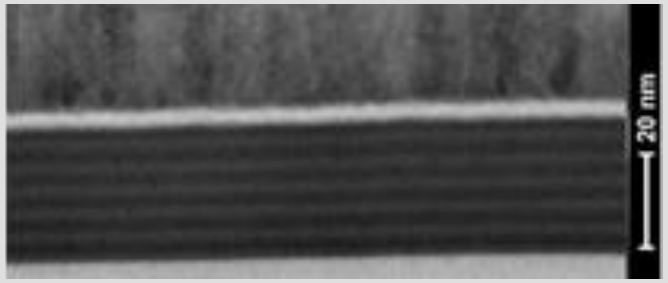
› Back-end selectors

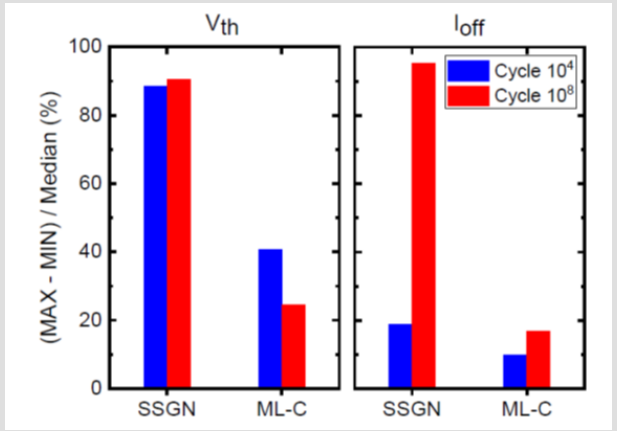




IEDM 2019

Ovonic Threshold Switch





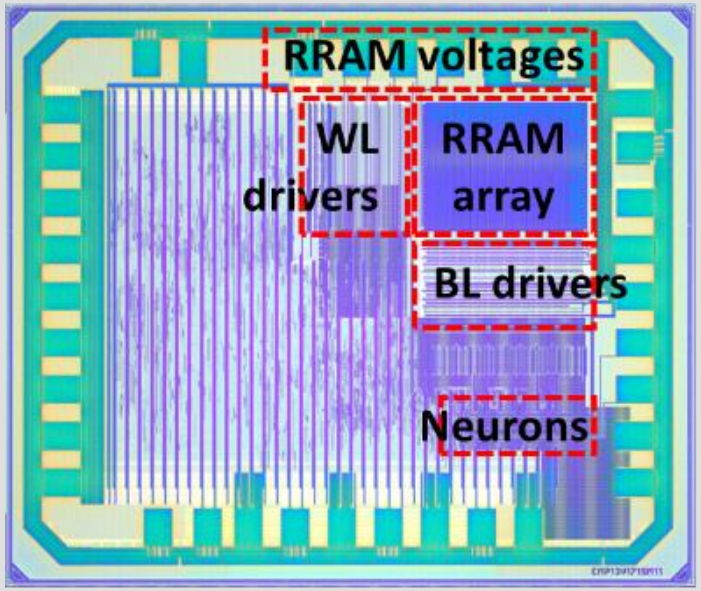
imw 2020

Ovonic Threshold Switch Multilayer Architecture

SPIKE CODING

› Toward sub pJ / event efficiency

SPIRIT



2019

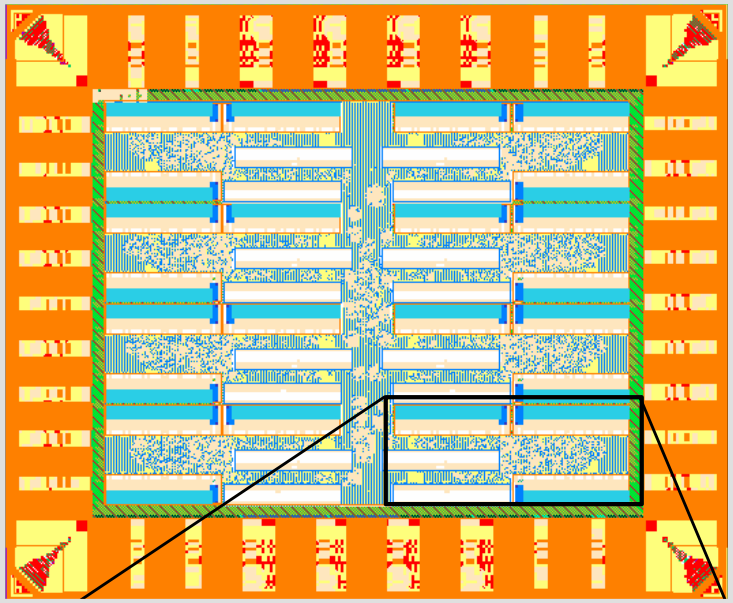
NeuRAM3

CMOS node: 130nm

10 neurons & 144 synapses

3.6 pJ /spike

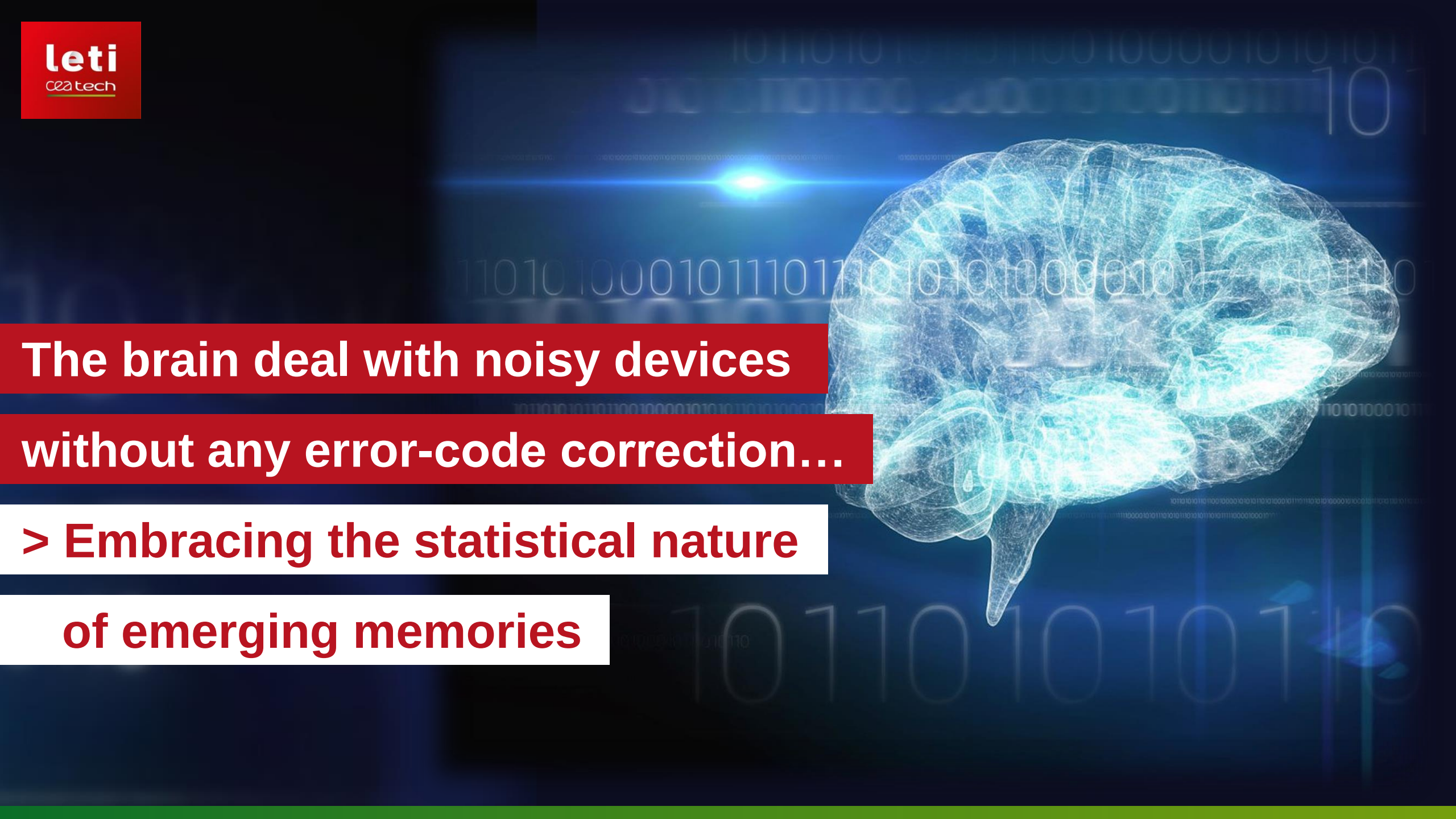
LARGO



CMOS node: 28nm (FD-SOI)

131k neurons & 75M synapses

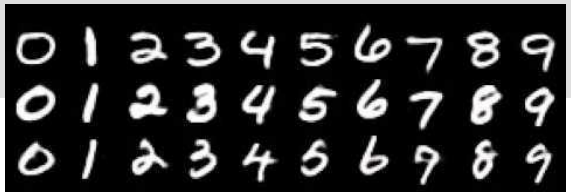
0.5pJ / spike


leti
ceatech
The brain deal with noisy devices
without any error-code correction…
> Embracing the statistical nature
of emerging memories

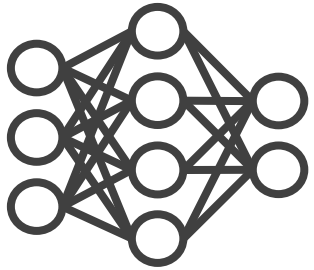
SPIKE CODING

› Bayesian neural networks

The “small-data” world has a lot of uncertainty

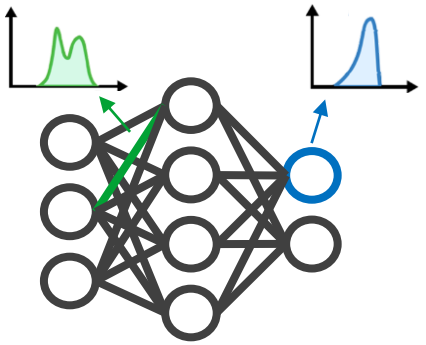


Deterministic model



‘You have input a 3’

Bayesian model



‘The input looks most like a 3… but I am very uncertain about that’

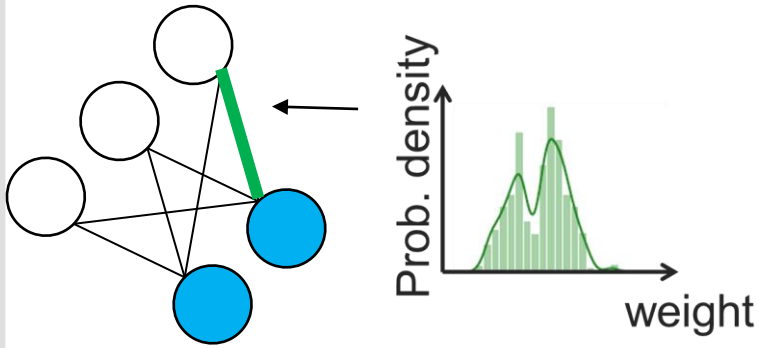
SPIKE CODING

› Bayesian neural networks

nature electronics
2021

Bayesian model

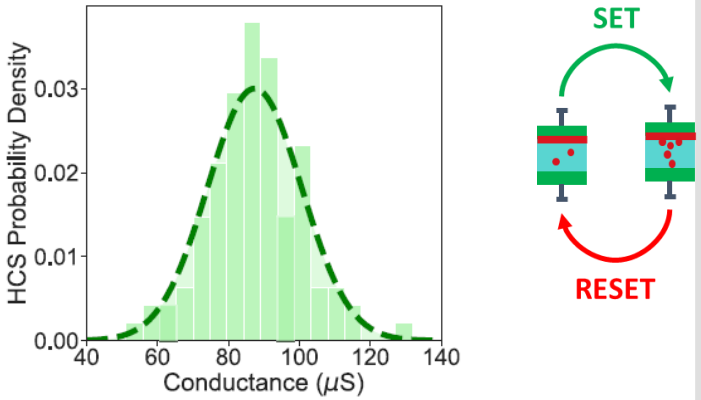
to face limited / incomplete input data



models = random variables

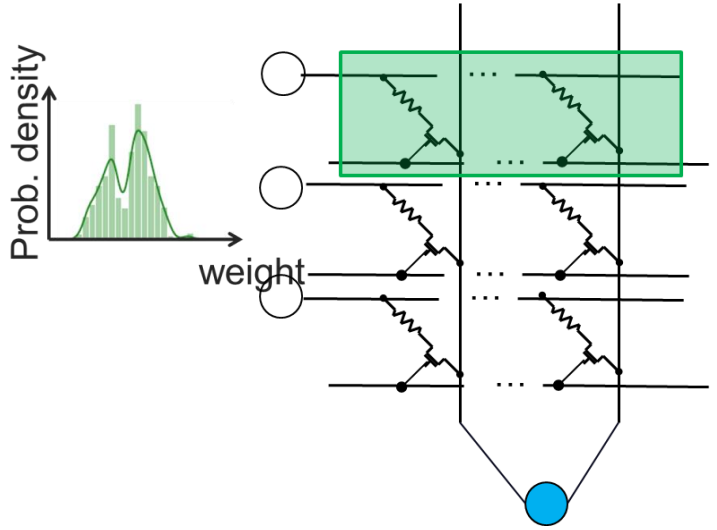
Entropy source

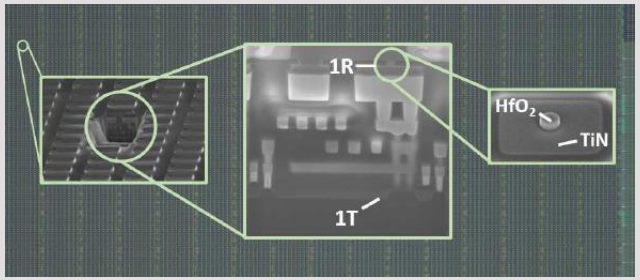
cycle-to-cycle variability

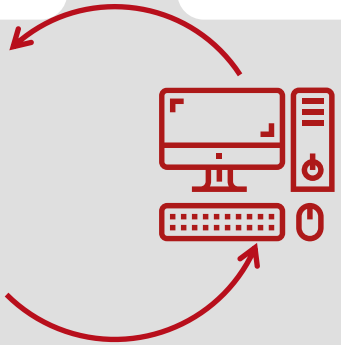


Circuit

in-memory multiply and accumulate







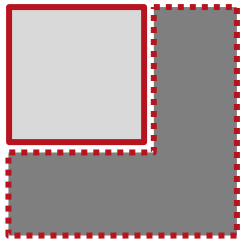
In-situ learning (Markov Chain Monte Carlo Sampling)

Detection of arrhythmic heartbeats

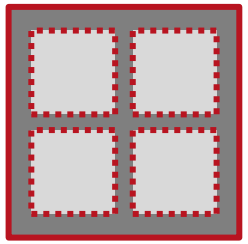
91% test accuracy

DENSE 3D INTEGRATION

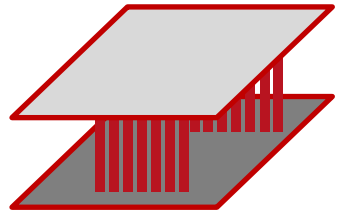
› Enhancing edge devices architectures



Limited area while increasing the memory



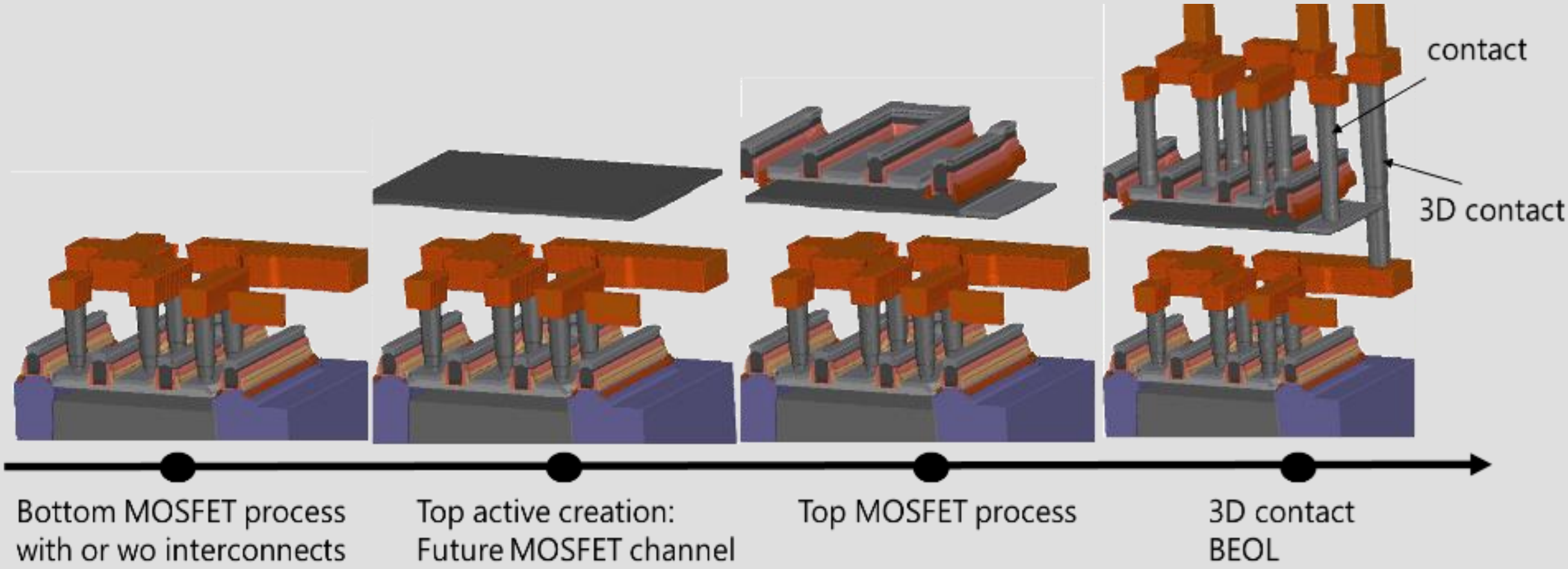
Flexible computing and analog units



Highly interconnected circuits

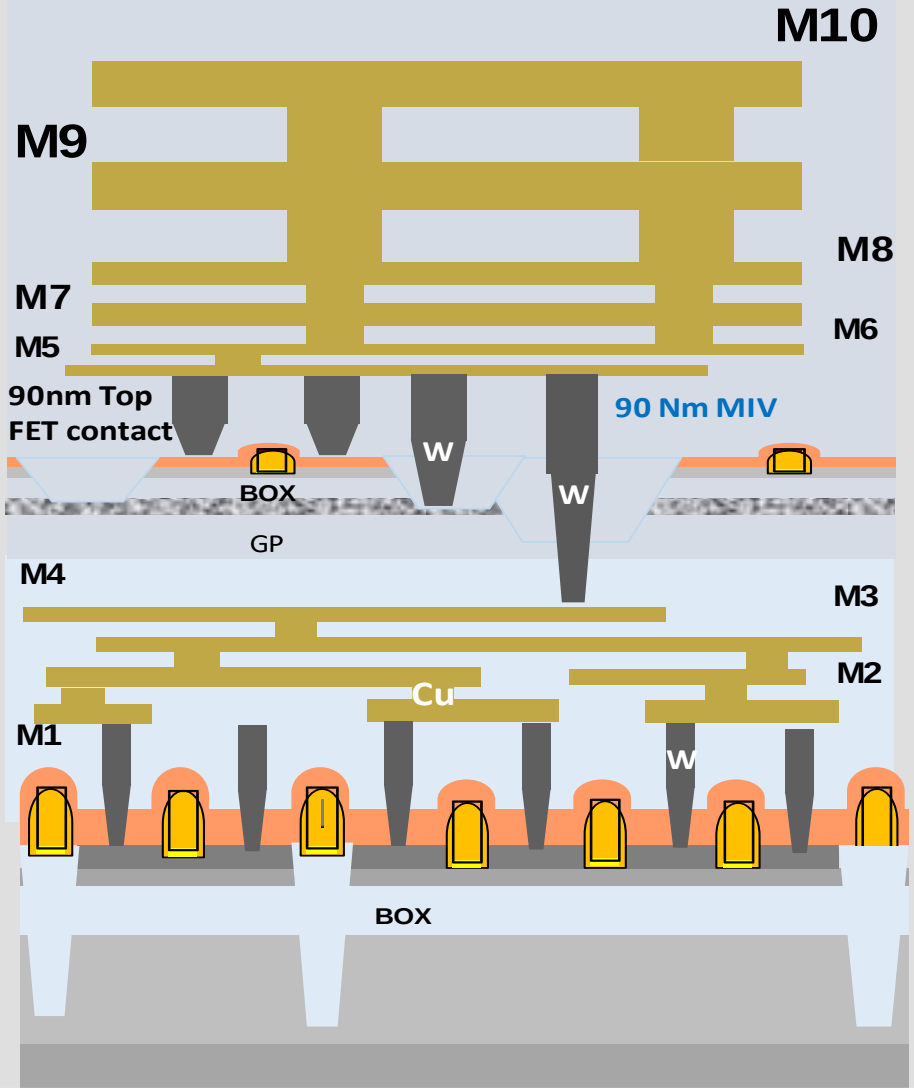
MONOLITHIC 3D STRUCTURES

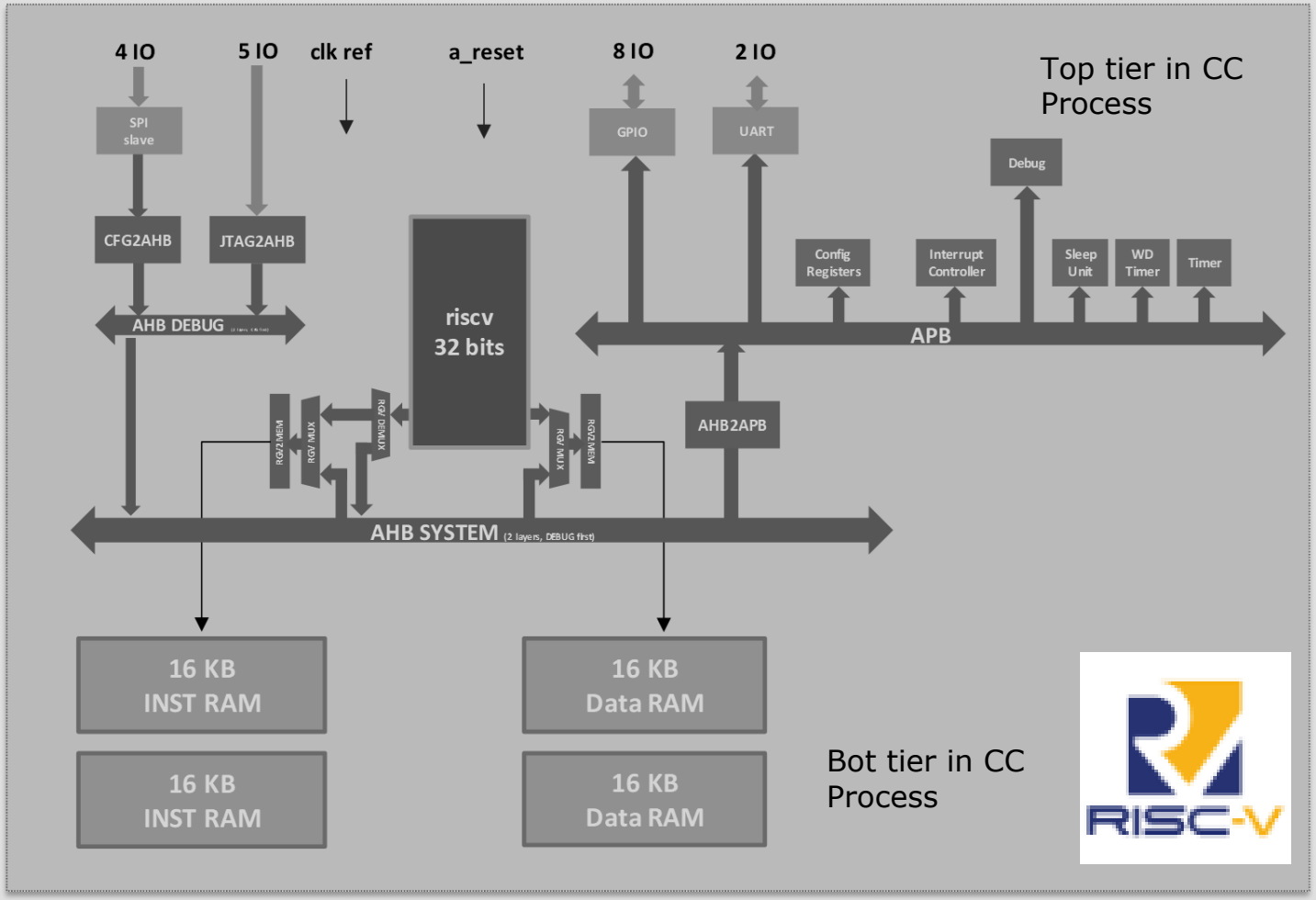
› Improve energy efficiency ×2



DENSE 3D INTEGRATION

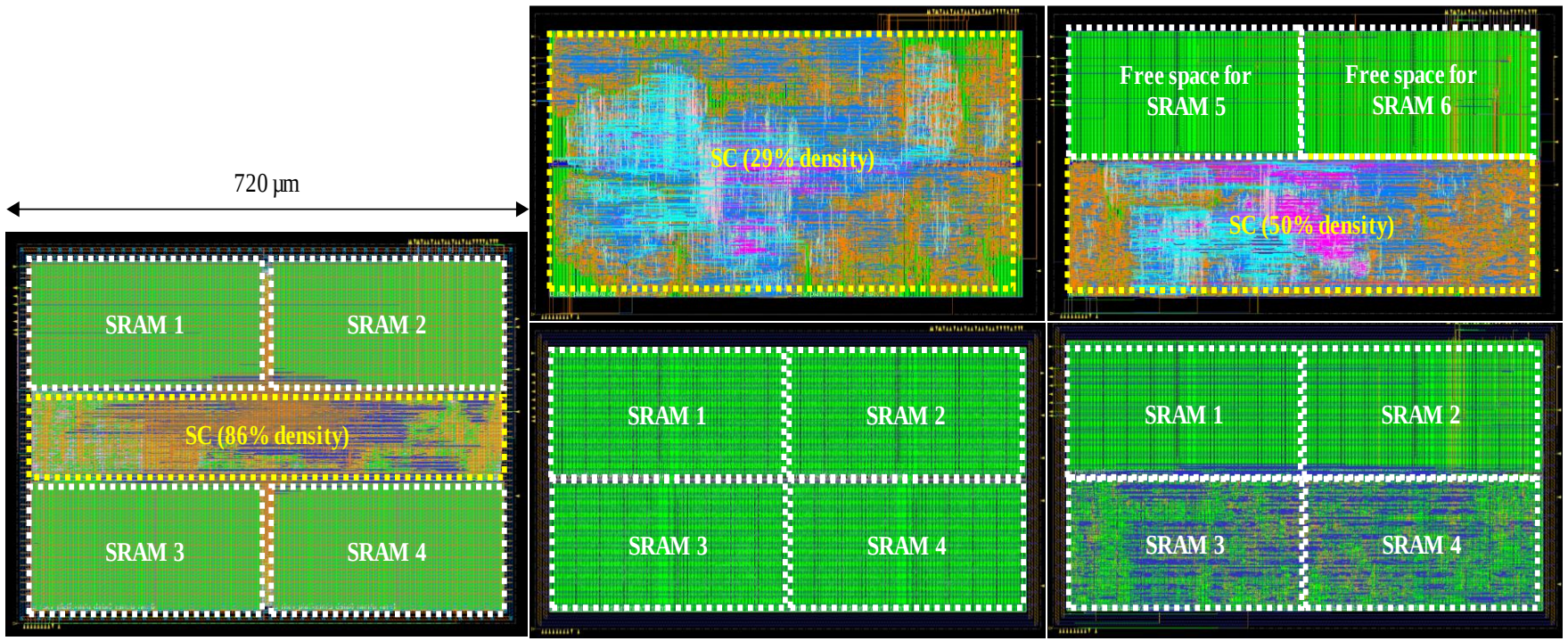
› Enhancing / Optimizing circuits footprint





DENSE 3D INTEGRATION

› Enhancing / Optimizing circuits footprint



(a) 2D GDS

(b) full M3D RI5C-V SoC GDS (up) and M3D RI5C-V SoC SRAM bottom tier GDS only (down)

(c) full M3D RI5C-V SoC GDS with 2x Instr Mem added (up) and M3D RI5C-V SoC SRAM bottom tier GDS only (down)

› 23% area reduction

› 7% performances improvement

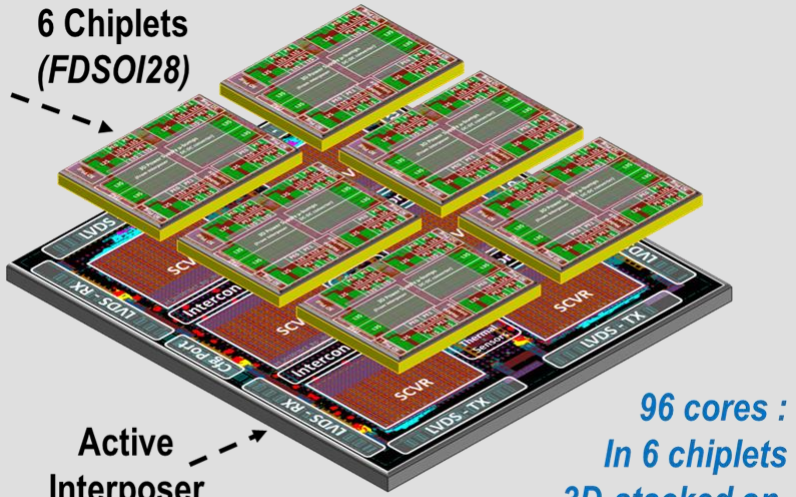
› 8k-415k Intermediate Vias

› 2x local inst. memory Increase

DENSE 3D INTEGRATION

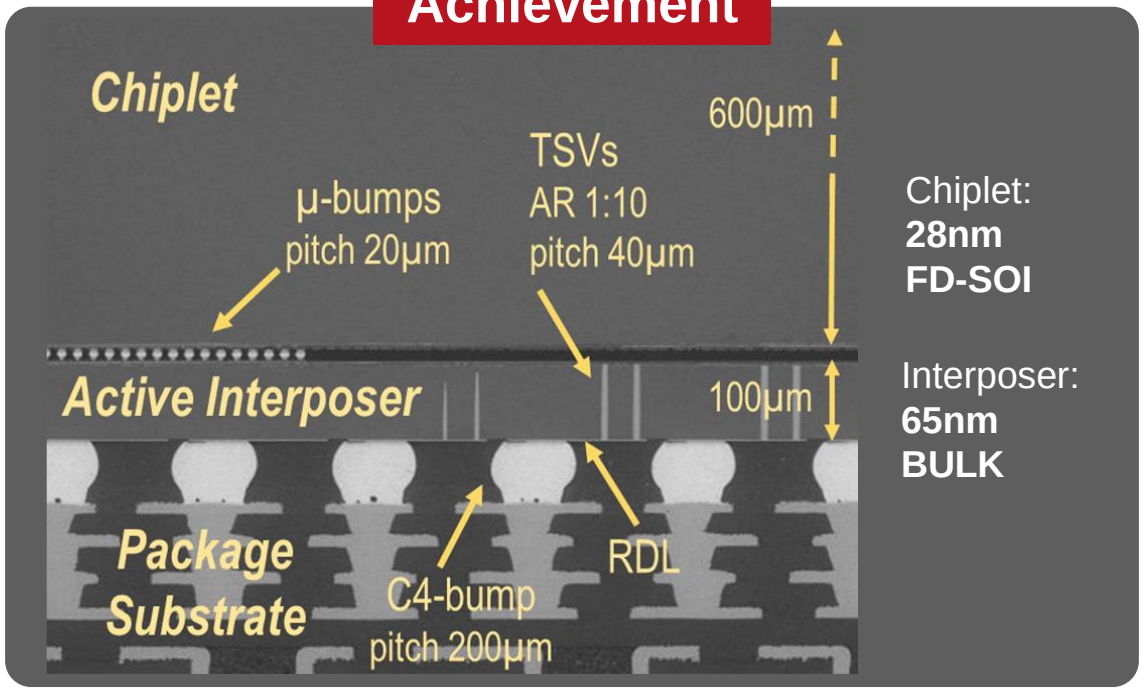
› Flexible computing and analog units

Concept



Improve parallelism, power performance, versatility and cost with a modular architecture based on smaller chips

Achievement



The power of 10 laptops with a surface of only 200mm²

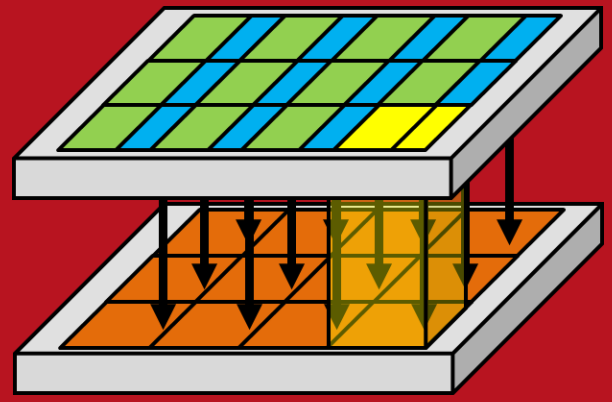
100GOPS, 10GOPS/Watt

DENSE 3D INTEGRATION

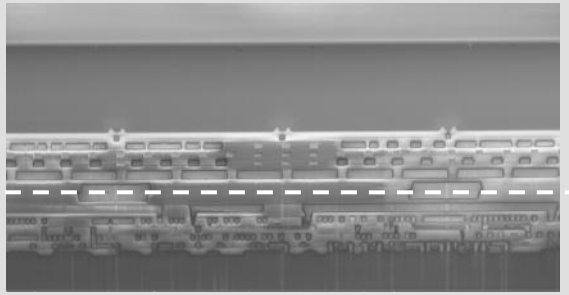
› Highly interconnected circuits

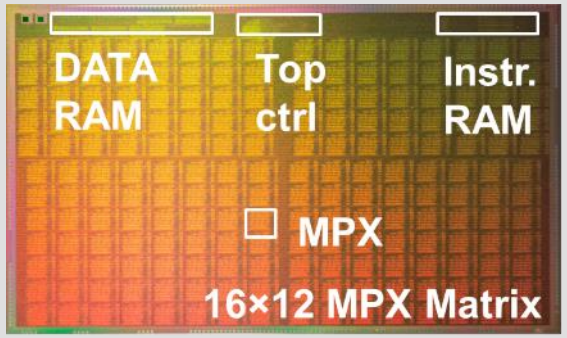
Concept

Pixels + ADC Array



Memory + processing





Achievements

Imager	Node	**130nm**
	Integration	**Stacked BSI***
	Power cons.	**720mW @9b 5500fps** **230mW @8b 340fps**
	Pixel size	**12um**x**12um (subpixel)**
	Resolution	**1024**x**768**
	Sensitivity	**15V/lx.s**
	Dynamic range	**54dB**
	ADC resol.	**10b**
	Frame rate	340fps @0.78Mpixels 10bits 1500fps @0.05Mpixel 10bits 5500fps @0.05Mpix 9bits
Computing	Parallelism	**matrix**
	PE array	**3072**
	Data memory	**73kB+98kB**
	Instr. Memory	**65kB**
	Clock freq	**80MHz**
	Performance	**61Gops @8b img**

NEW PARADIGM IS NEEDED TO FAVOR SOBRIETY/FRUGALITY VS. DECLINISM

We need to drastically reduce the energy and environmental footprint of electronic devices

Declinism
pessimist's approach
Reducing or limiting performance

Sobriety
athlete's approach
Maximize performance for a given resource

If you share the same vision, Join us!

thomas.signamarcheix@cea.fr

CEAleti @CEA-leti @CEA_leti