COIN-3D

Collaborative Innovation in 3D VLSI Reliability

Machine Learning Applications in EDA for Chiplet Reliability

University of Glasgow

Task 2.3: Expert Webinars

M05-M18

Date: Fri, 31 Jan 2025 11:00 - 12:10 GMT

Funded by the European Union

Agenda

- **Introduction :** *Dr George Floros*
- **ML/DL/AI Introduction :** *Dr Tahani Aladwani*
- **Thermal Challenges in EDA: From Traditional to AI-Driven Solutions:** *Dr Yixian Shen*

Q&A/Break

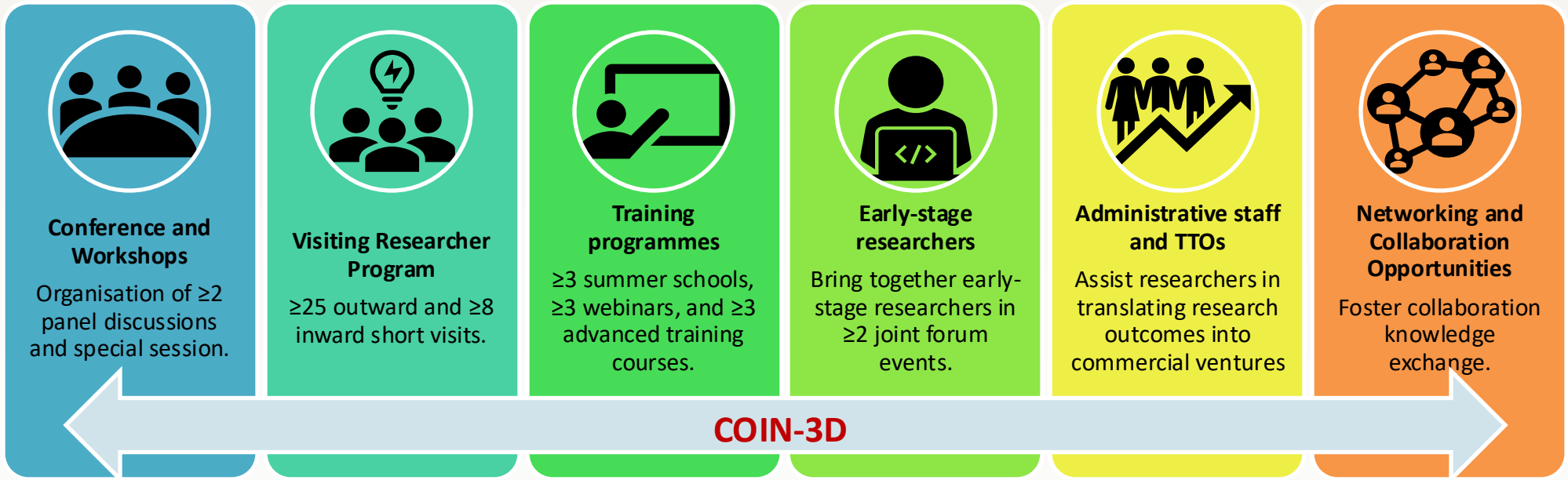
- **AI-driven Design Optimization :** *Dr Tahani Aladwani*
- **AI-driven Lithographic Hotspot Detection :** *Dr Tahani Aladwani*
- **ML-Assisted Approach for Accelerated Thermal & IR Drop Analysis:** *Dr Olympia Axelou*
- **Swarm Intelligence for Electromigration-Resilient Design:** *Dr Olympia Axelou*
- **Unified thermal scheduler for DNN in 3D-stacked systems :** *Dr Yixian Shen*

Q&A

Funded by
the European Union

COIN-3D Overview

- COIN-3D goal:
  - Research and knowledge exchange activities in the area of 3D VLSI



Funded by the European Union

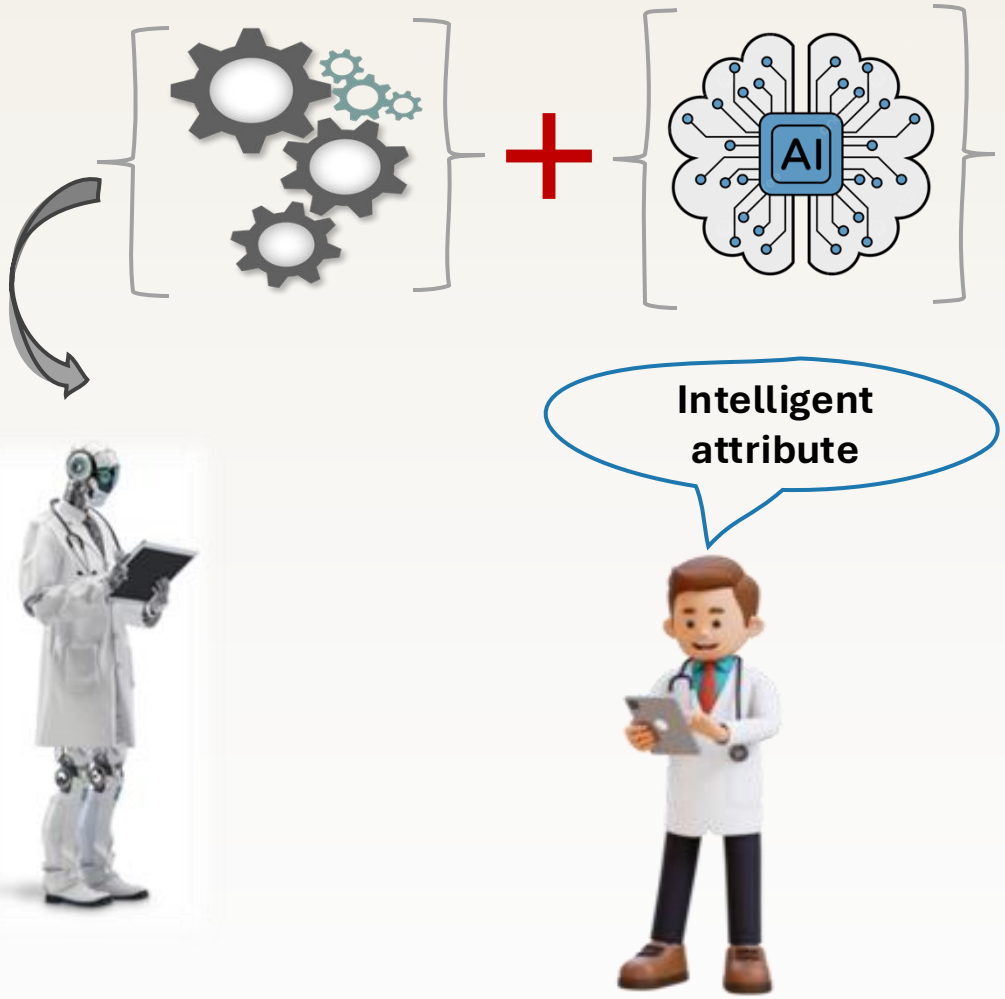
COIN-3D: Partners

Funded by
the European Union

ML/DL/AI Introduction

What Is Artificial Intelligence (AI) ?

AI The science of making things smart.



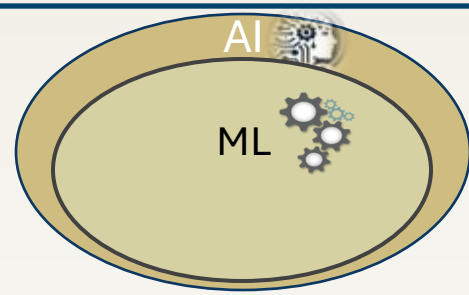
AI is a technique which enables **machines** to mimic human behaviors to **learn** and solve problem.

Funded by the European Union

There are many way to make AI done

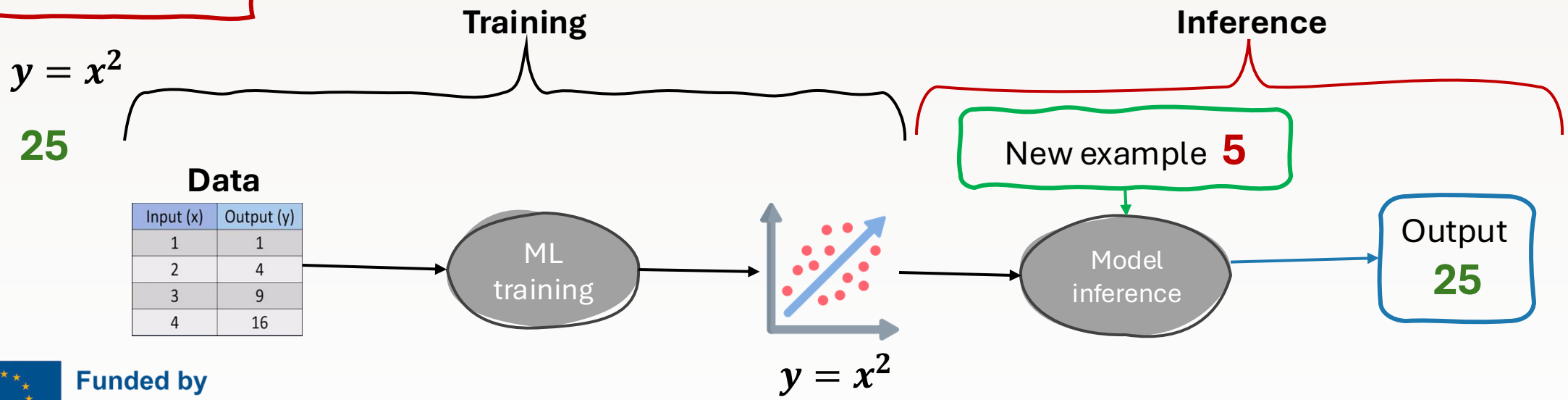
Input (x)	Output (y)
1	1
2	4
3	9
4	16
5	?

How we got this



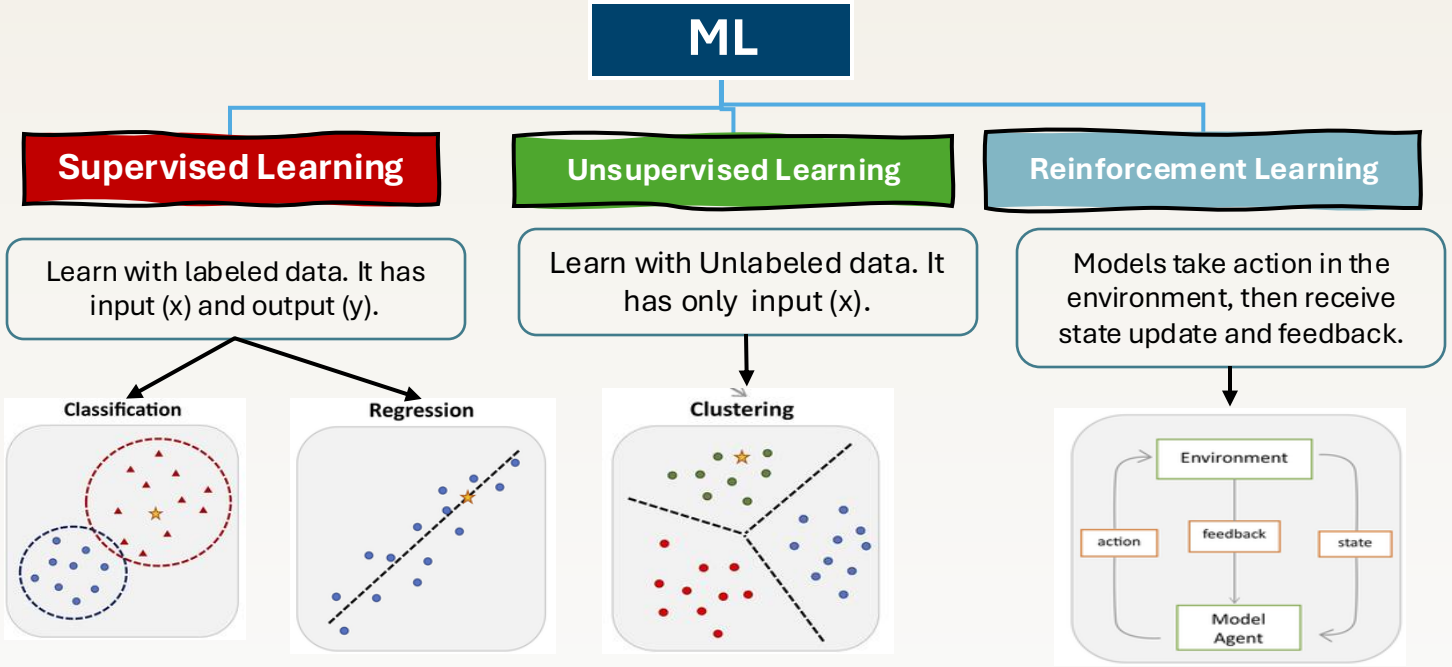
We observed the x and y values to extract some patterns.

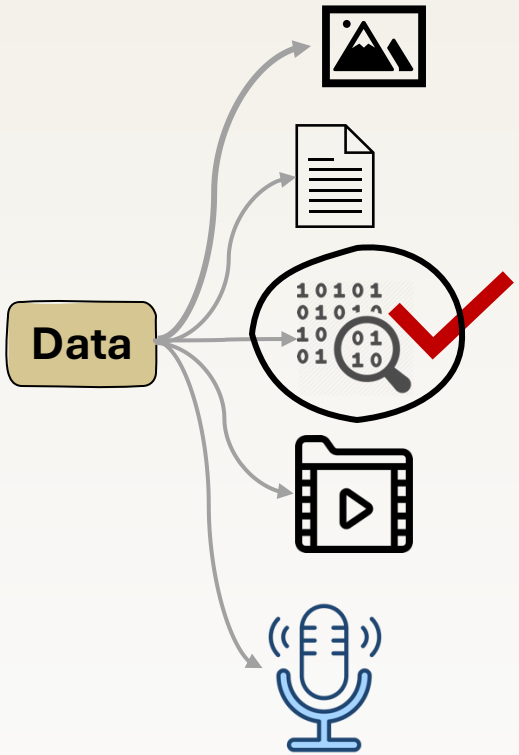
To solve this problem by machines, they needs to learn to **extract** these **patterns** by applying set of programs or algorithms.



Funded by the European Union

Machine learning ML





ML requires features from humans and structured data.

Smaller datasets

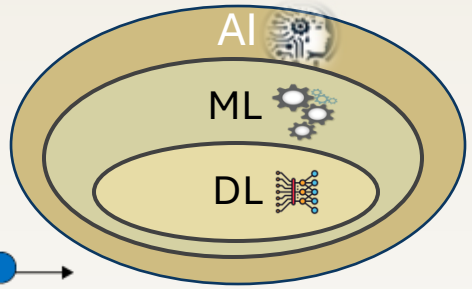
Complex data needs complex tourniquets.

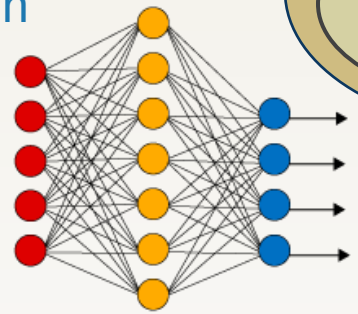
Funded by the European Union

Deep learning (DL)

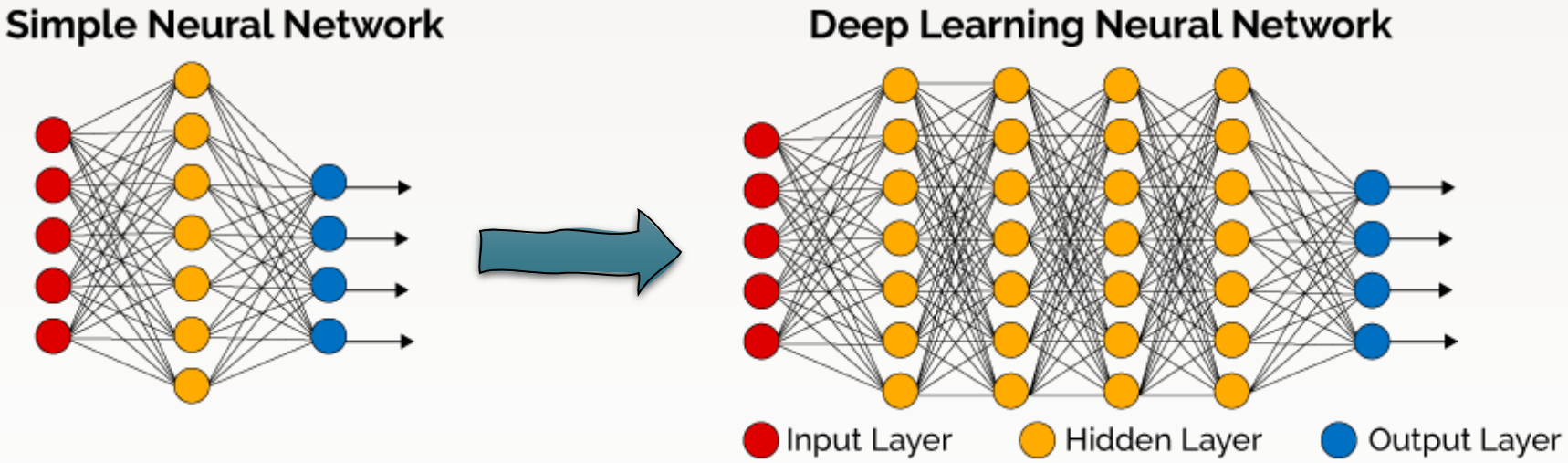
A technique for Implementing Machine Learning, known as **deep neural networks** as well.

What we mean by **neural networks (NN) ? Way** to mimic how neurons in your brain work.





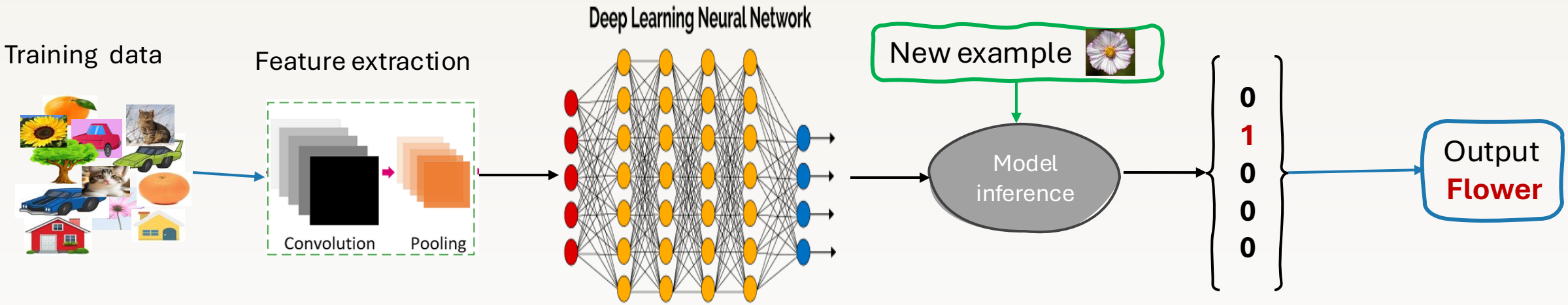
The word **deep** refers to the use of multiple layers in a neural network.



Funded by
the European Union

Deep Learning

How DL works



Funded by the European Union

Deep learning (DL)

DL

CNN

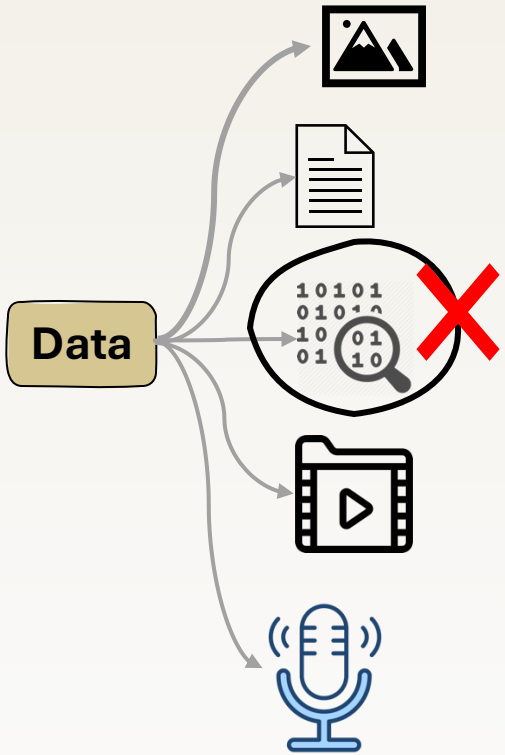
Convolutional Neural Networks (CNN).
Effective for image recognition and processing.

RNN

Recurrent Neural Networks. well-suited to sequence data such as time series or natural

LSTM

An advanced RNN architecture that can learn long-term dependencies in sequence data.



DL does not requires features from humans and It works well with unstructured data.

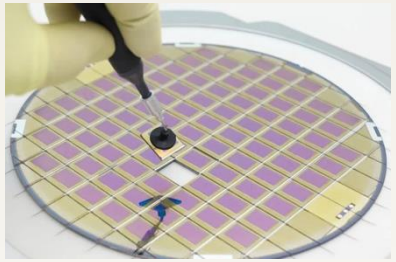
How this can be helpful in the chiplet design?

Ideal for complex problems, such as LLM, and image recognition.

large datasets

Funded by the European Union

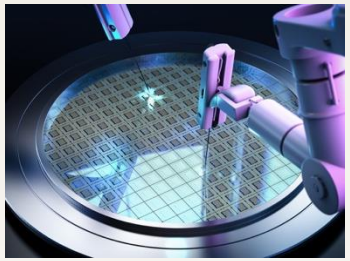
AI is Transforming Semiconductor Design Processes



Traditional design methodologies

- ✓ Requires significant manual effort.
- ✓ Often labor-intensive and time consuming.
- ✓ Increasingly constrained by human expertise and iterative processes.
- ✓ Less accuracy
- ✓ Less efficiency, and more vulnerable to defects.

Vs



AI design methodologies

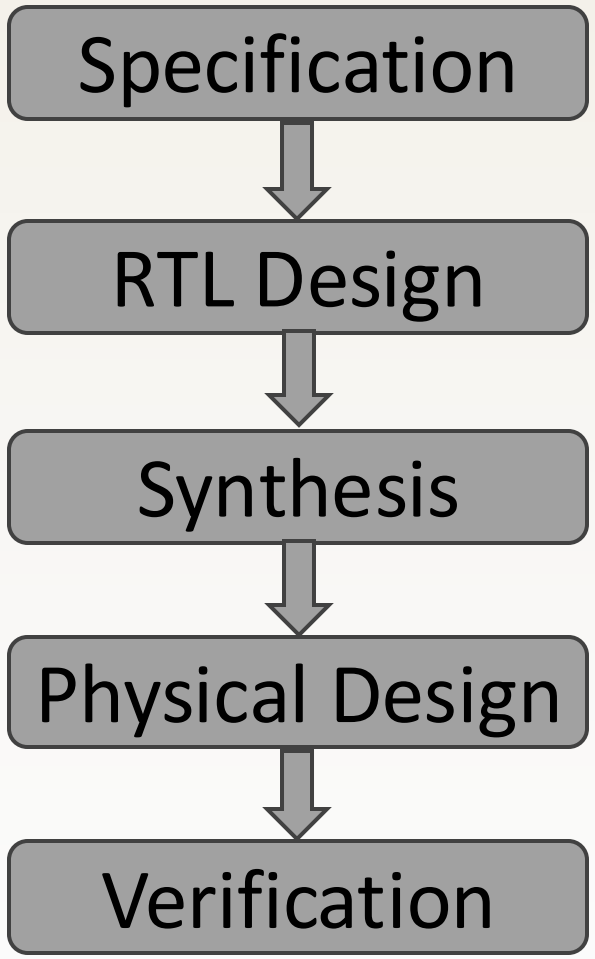
- ✓ AI automates repetitive tasks like component placement, routing, and verification.
- ✓ Reduce time-to-market, improving the accuracy of semiconductor fabrication processes.
- ✓ Increase design efficiency and reducing defects.
- ✓ AI utilizing advanced models such as Generative Adversarial Networks (GANs) and DNN, CNN, and RL offers innovative approaches to automate and optimize various stages of chip design.

Funded by
the European Union

Thermal Challenges in EDA: From Traditional to AI-Driven Solutions.

Funded by
the European Union

Overview of Traditional Chip Design Flow



Define chip functionality (e.g., thermal constraints, power targets).

Describe behavior using Verilog/VHDL.

Convert RTL into gate-level netlist.

Place gates, route wires, and optimize for timing/power.

Ensure the design works as intended.

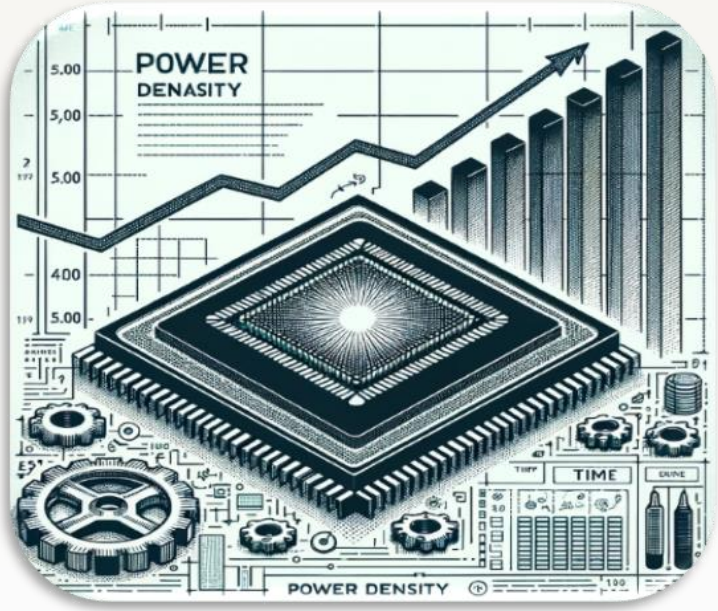
Funded by
the European Union

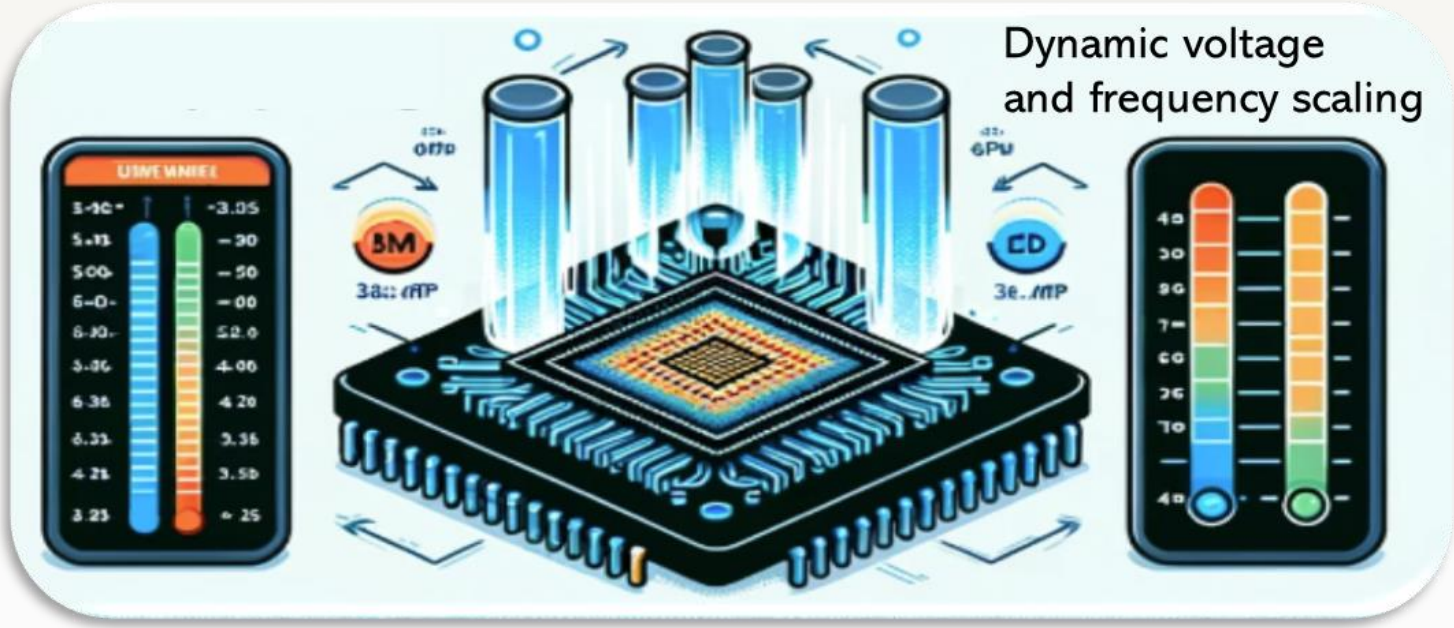
Key Challenge: Power & Thermal Issues

More transistors

Higher power density

Overheating and reliability risks.

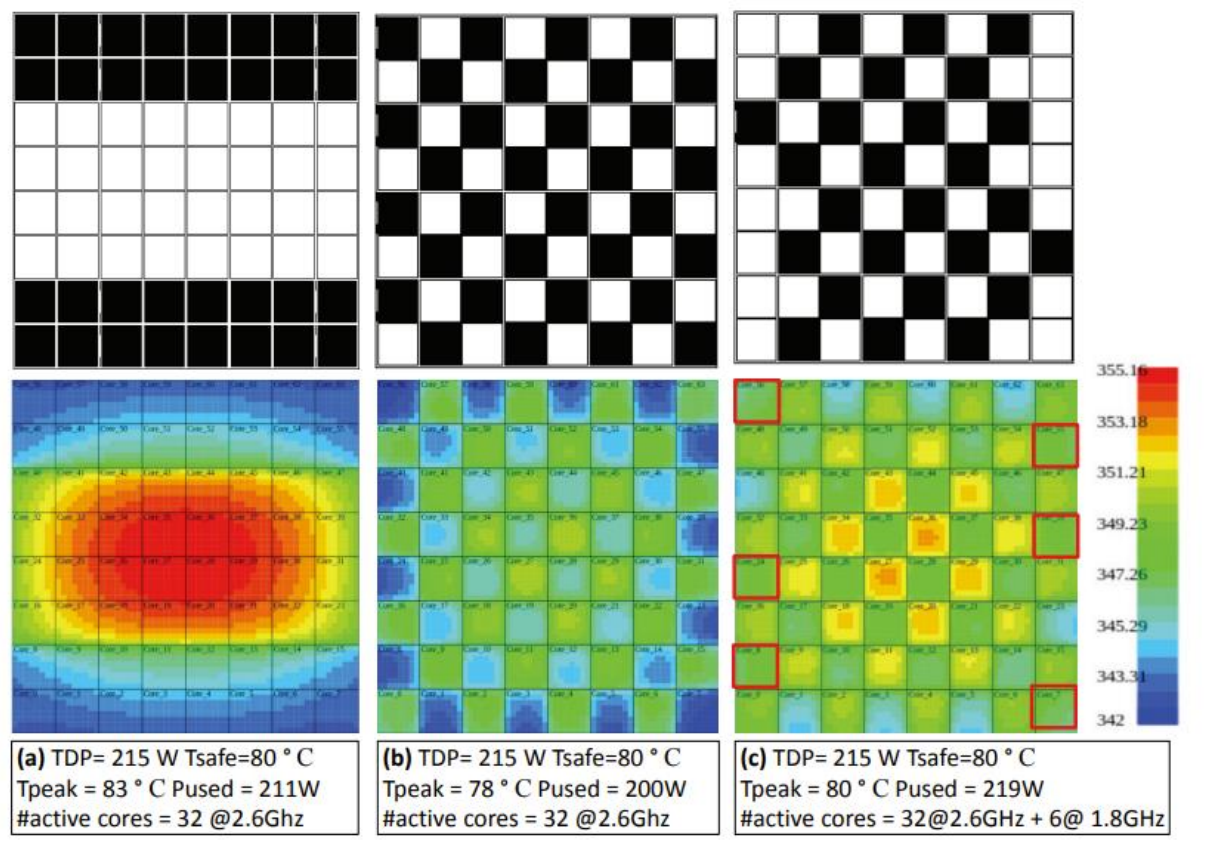




Funded by
the European Union

Dark Silicon

Only a fraction of cores can be **active** simultaneously

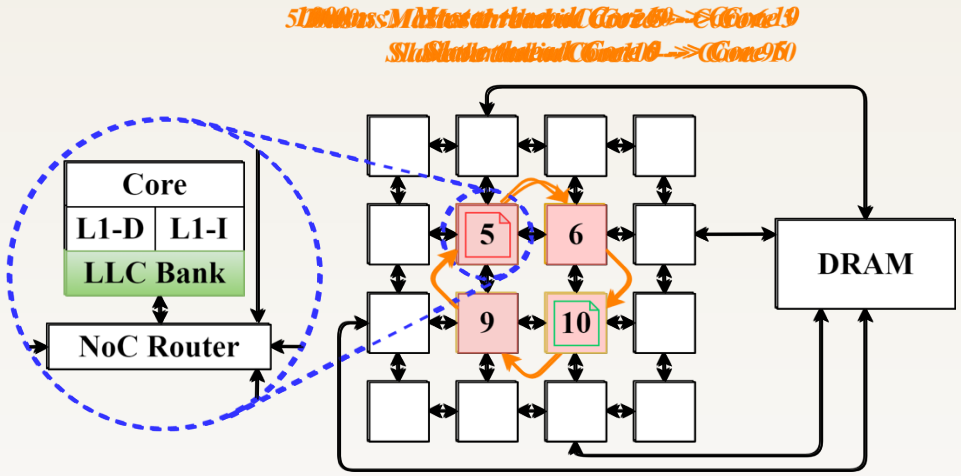


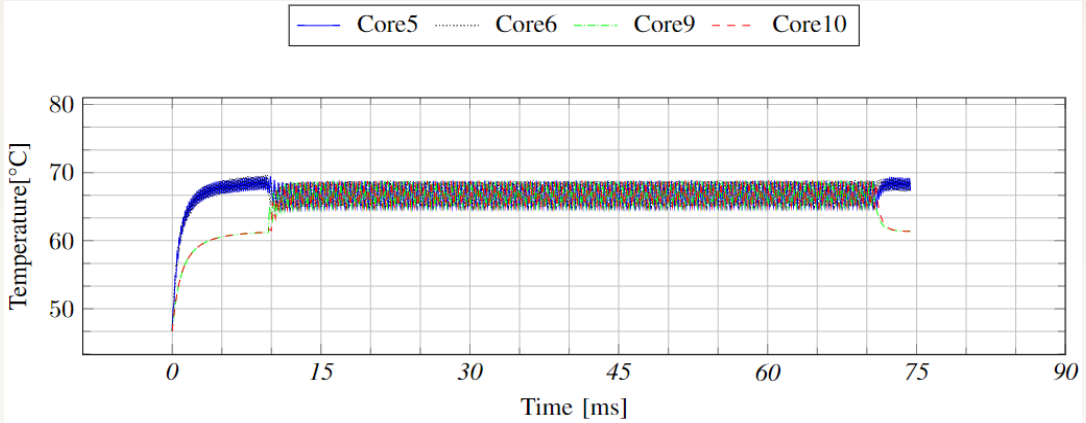
Shafique, Muhammad, et al. "**Dark silicon as a challenge for hardware/software co-design:** Invited special session paper." Proceedings of the 2014 International Conference on Hardware/Software Codesign and System Synthesis. 2014..

Funded by
the European Union

Analytical Model for thermal management

- **Thread rotations for blacksholes (Master+Slave threads)**





- **The inherent heterogeneity of many-core systems**

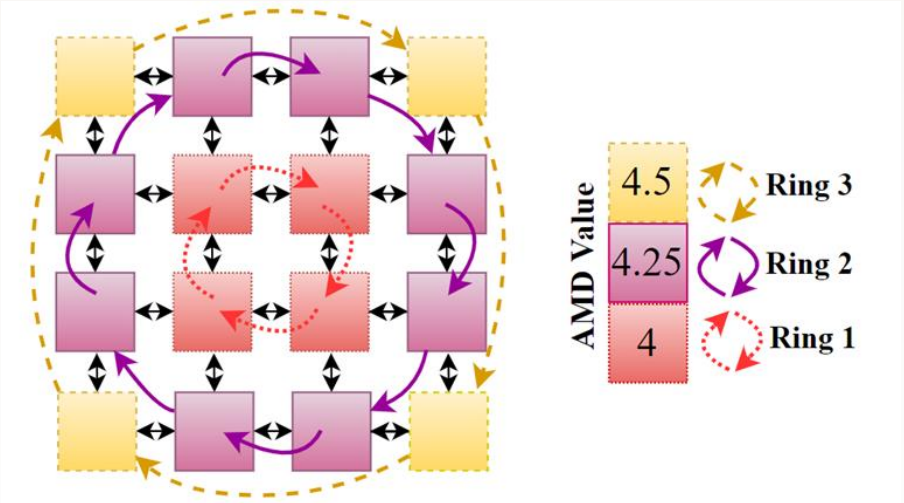


Fig. 3: Abstraction for concentric AMD-based rotation rings.

- Access cache latency

  Ring1 < Ring2 < Ring3

- Thermal dissipation condition

  Ring1 < Ring2 < Ring3

Requiring extensive analysis

- **A theoretical peak temperature calculation**

$$\mathbf{T}_{(d\delta+1)\tau} = \mathbf{V} \cdot \mathbf{diag}\left(\frac{1}{1-e^{\delta\lambda_1\tau}}, \frac{1}{1-e^{\delta\lambda_2\tau}}, \cdots, \frac{1}{1-e^{\delta\lambda_N\tau}}\right) \cdot \mathbf{V}^{-1}\mathbf{w}\mathbf{P}_{\tau}$$
$$+ \mathbf{V} \cdot \mathbf{diag}\left(\frac{e^{(\delta-1)\lambda_1\tau}}{1-e^{\delta\lambda_1\tau}}, \frac{e^{(\delta-1)\lambda_2\tau}}{1-e^{\delta\lambda_2\tau}}, \cdots, \frac{e^{(\delta-1)\lambda_N\tau}}{1-e^{\delta\lambda_N\tau}}\right) \cdot \mathbf{V}^{-1}\mathbf{w}\mathbf{P}_{2\tau} + \cdots$$
$$+ \mathbf{V} \cdot \mathbf{diag}\left(\frac{e^{\lambda_1\tau}}{1-e^{\delta\lambda_1\tau}}, \frac{e^{\lambda_2\tau}}{1-e^{\delta\lambda_2\tau}}, \cdots, \frac{e^{\lambda_N\tau}}{1-e^{\delta\lambda_N\tau}}\right) \cdot \mathbf{V}^{-1}\mathbf{w}\mathbf{P}_{\delta\tau} \quad (10)$$

Auxiliary matrix using floorplan-based constants

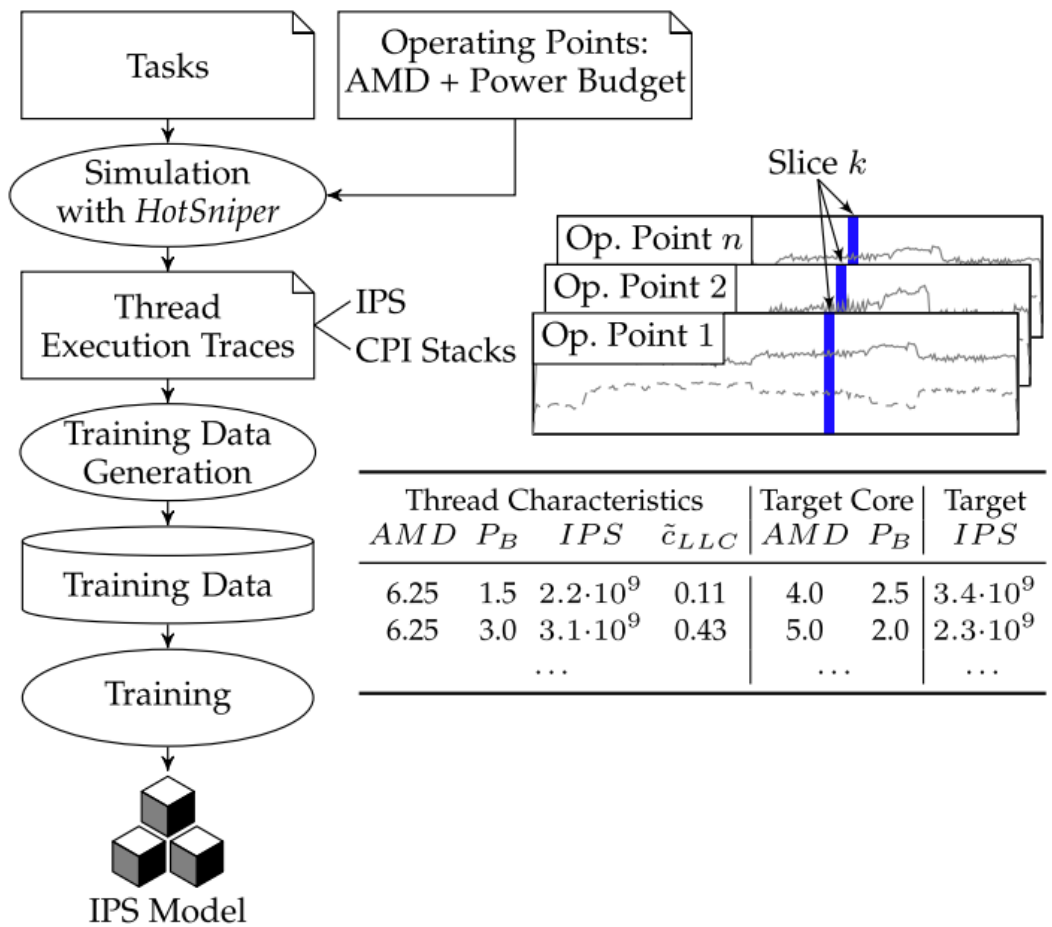
Rotation speed

Rotation period

Power traces

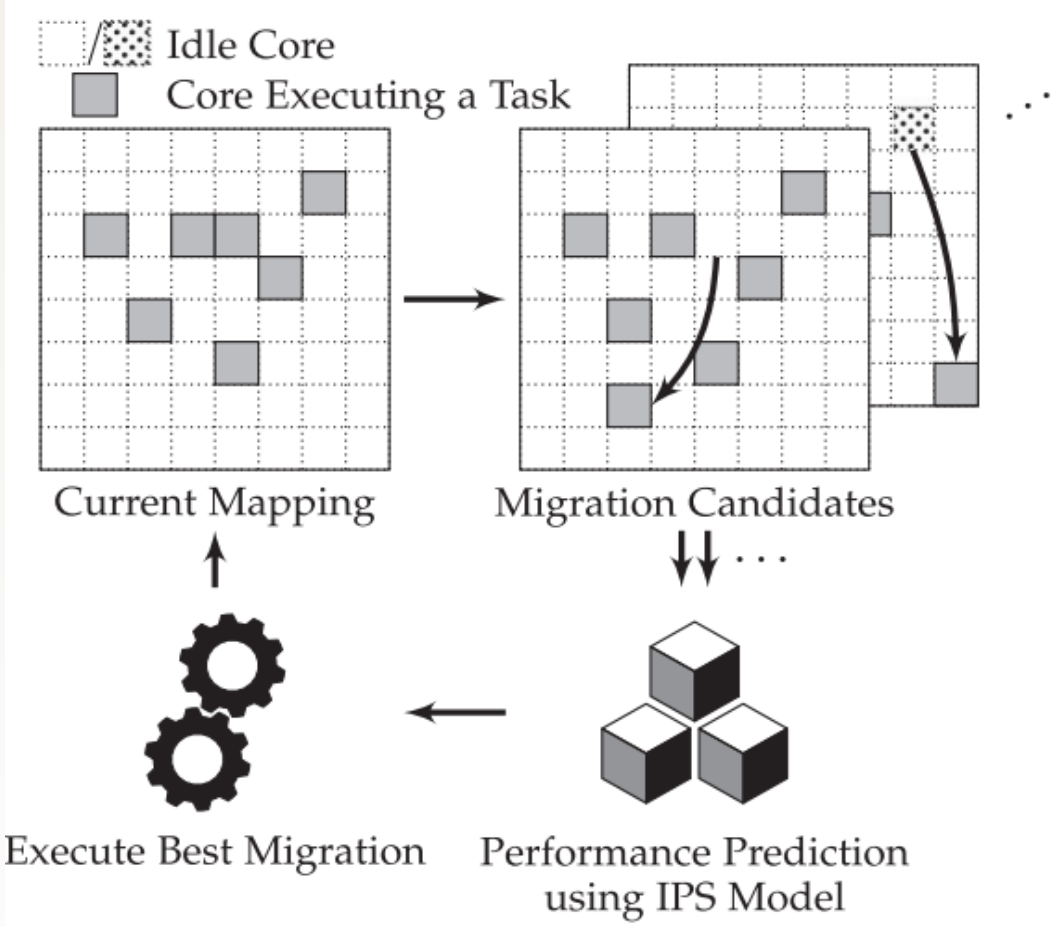
Funded by the European Union

Machine Learning-based Thermal Management



Thread Characteristics				Target Core		Target
AMD	P_B	IPS	$\tilde{c}_{LLC}$	AMD	P_B	IPS
6.25	1.5	$2.2{\cdot}10^9$	0.11	4.0	2.5	$3.4{\cdot}10^9$
6.25	3.0	$3.1{\cdot}10^9$	0.43	5.0	2.0	$2.3{\cdot}10^9$
		...		...		...

Training Time



Inference Time

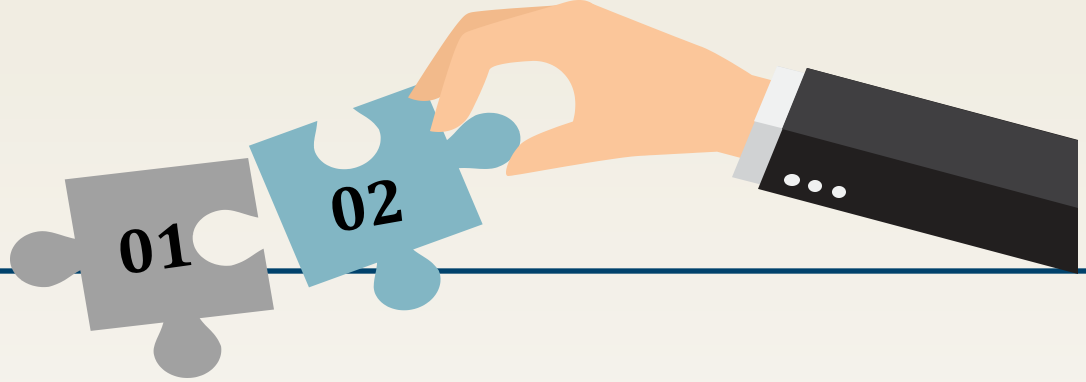
Funded by the European Union

Q&A/Break

Funded by
the European Union

Lithographic hotspots (HSs) Detection

Funded by
the European Union



STEP 01

Filed

Manufacturing Precision: Lithographic hotspots (HSs) Detection

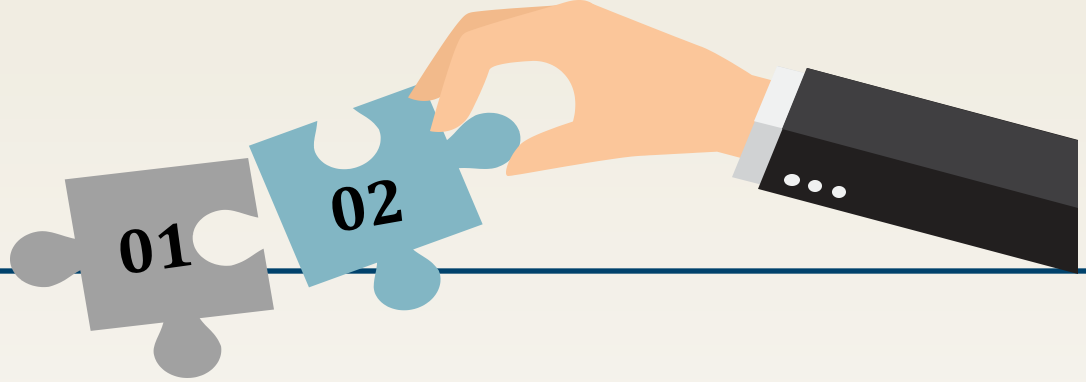
STEP 02

Problem

Smart devices nowadays such as (Apple iPhone , Watch , Laptops) use modern chips with advanced capabilities (like 5 nm feature size or even smaller in some cases).

Chips can have tens of billions of transistors, enabling advanced features like AI acceleration, high-performance gaming, multi-core computing, lower power consumption, and more capabilities (e.g., AI, 5G, etc.). However, as ***the features shrink***, the risk of **manufacturing defects** increase significantly.

Funded by
the European Union



STEP 02

Problem

Lithography is used to create extremely tiny structures on semiconductor wafers, which are the base materials for electronic chips.

Due to the effect of ***neighboring patterns***, there is a chance of ***bridging*** or ***pinching*** of the structures. This bridging or pinching of the structures is known as ***lithographic hotspots.***

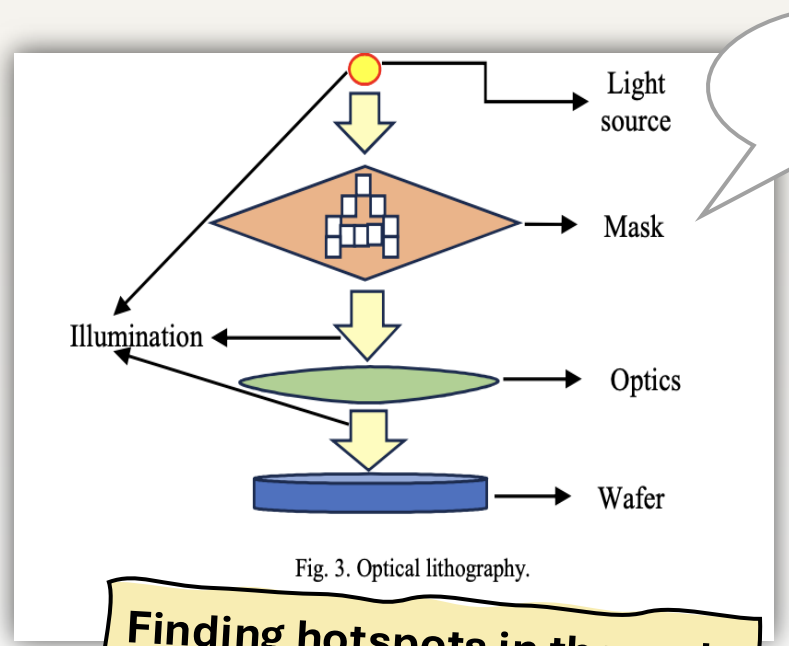


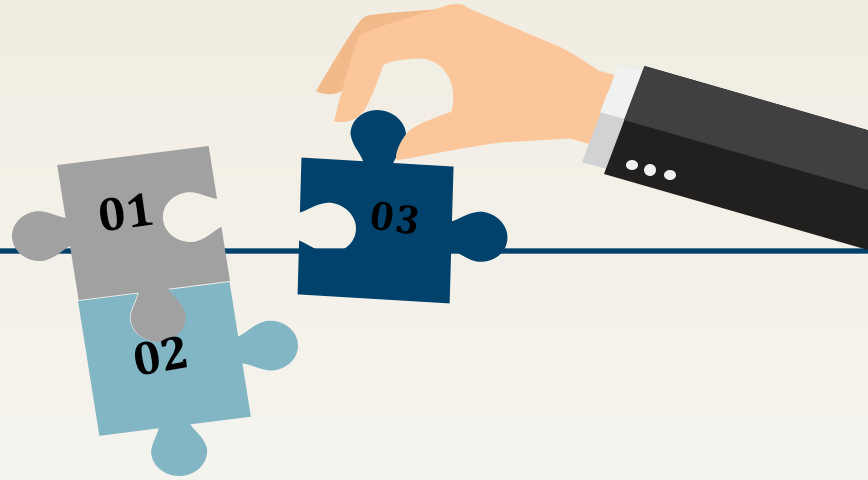
Fig. 3. Optical lithography.

Finding hotspots in the early design stage is mandatory to avoid undesirable layout patterns.

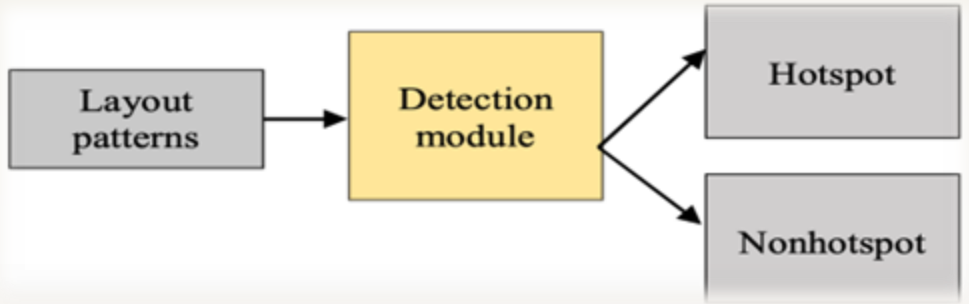
Funded by the European Union

STEP 03

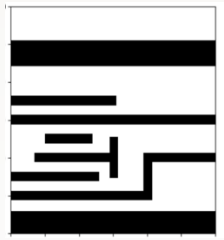
Proposed solution

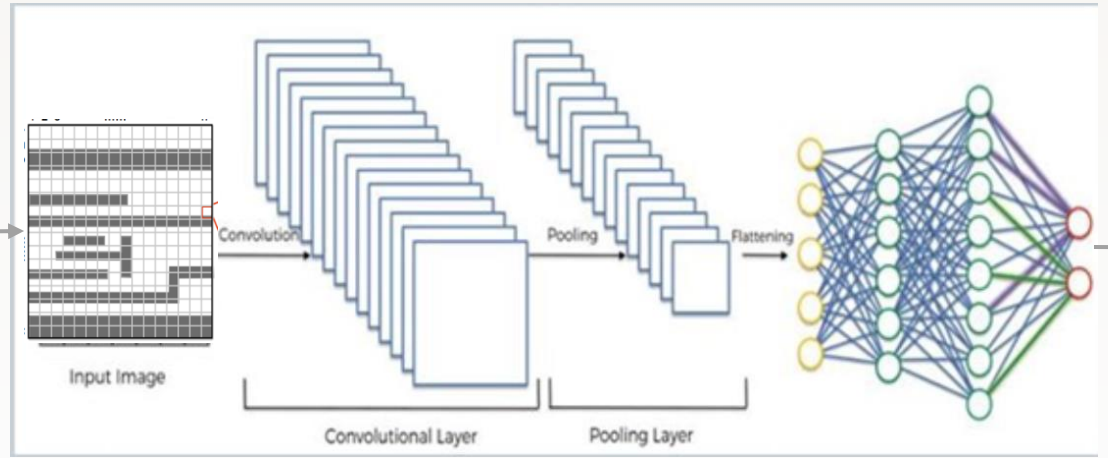


Lithographic hotspot detection can be done effectively using convolutional neural networks (CNNs).



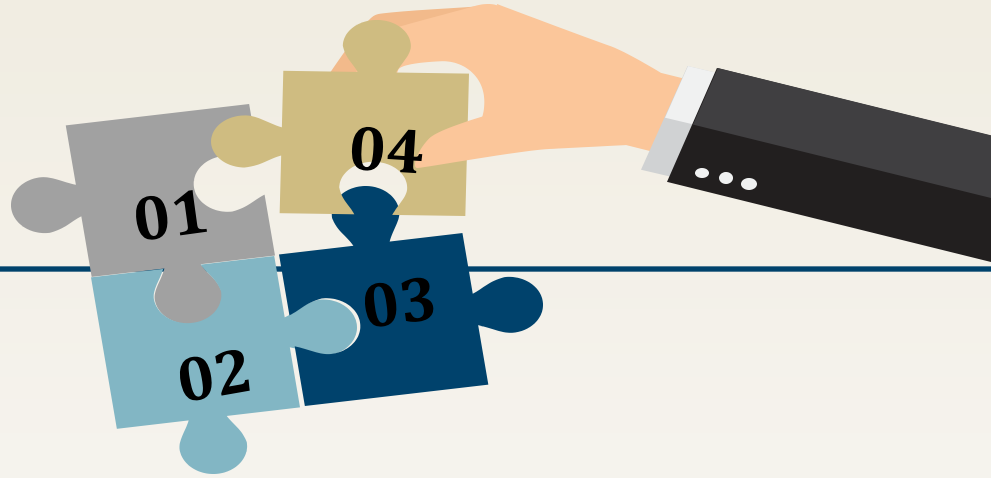
Lithographic hotspot images





1 **Hotspot**
0 **Not hotspot**

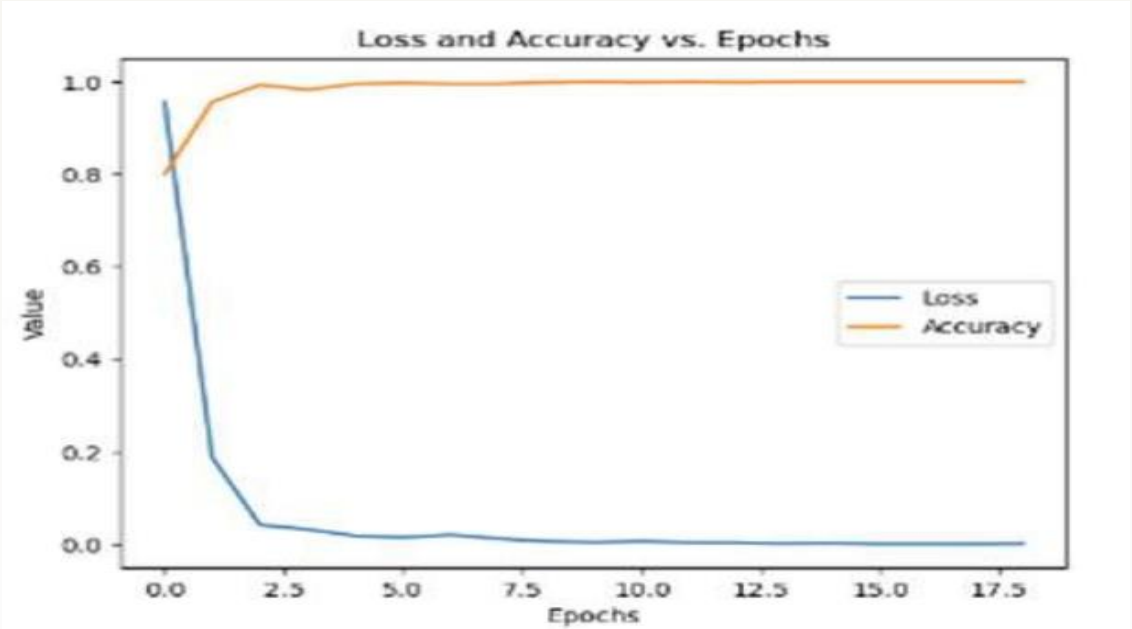
Funded by the European Union



STEP 04

Results

Introduced an accurate, fast , cheaper , and reusable hotspot detection framework.

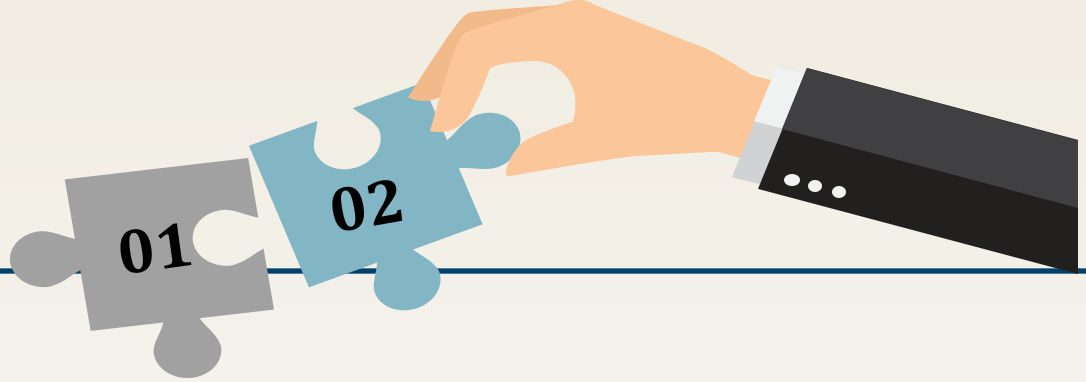


Funded by
the European Union

BeGAN

Funded by
the European Union

BeGAN



STEP 01 **Filed**

Design Optimization: BeGAN

STEP 02 **Problem**

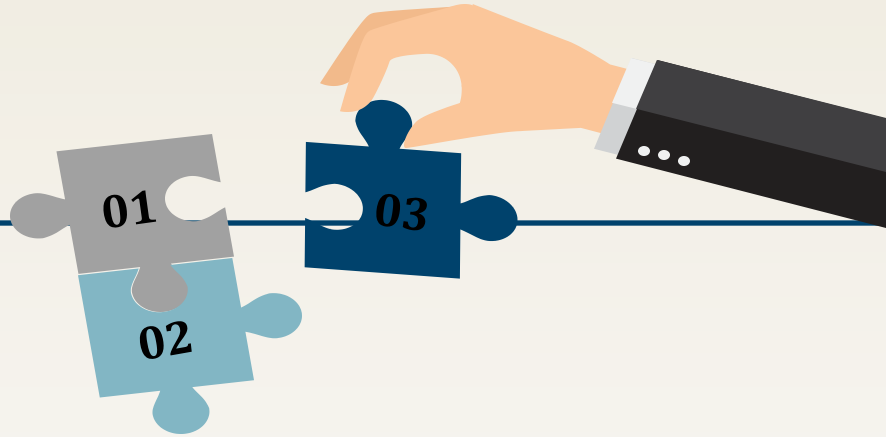
Benchmarks allow researchers and developers to validate the performance of new EDA algorithms by comparing them to prior approaches.

For example, benchmarks help test whether an algorithm optimizes power delivery networks (PDNs) effectively or improves upon existing methods.

Benchmarks are difficult to create due to obsolescence, confidentiality, labor intensity, and data scarcity.

Funded by
the European Union

BeGAN



STEP 03

Proposed solution

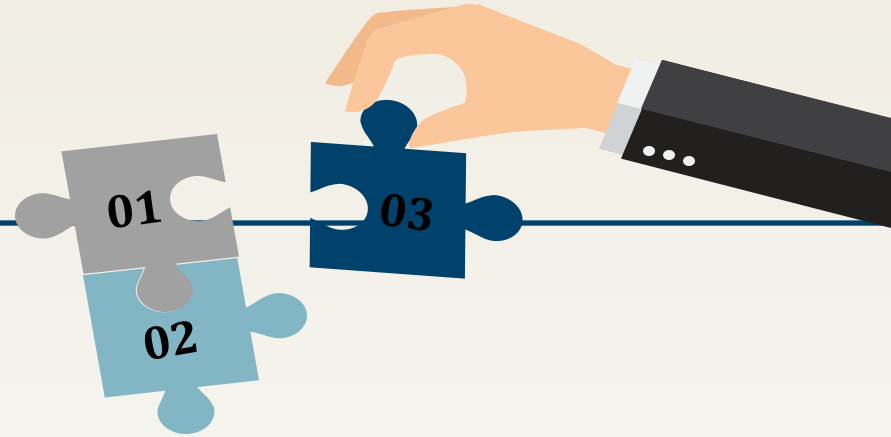
BeGAN, a GAN-based approach to synthetically ***generate realistic benchmarks for PDN design***, addressing the limited availability of real benchmarks.

PDN Benchmarks: Test tools that optimize the power delivery network for issues like IR drop and electromigration.

BeGAN employs **generative adversarial networks (GANs)** to extract the characteristics of real current maps (CMs) in a training set and generate a large set of realistic synthetic PDN benchmarks (CMs and PDNs.

Funded by
the European Union

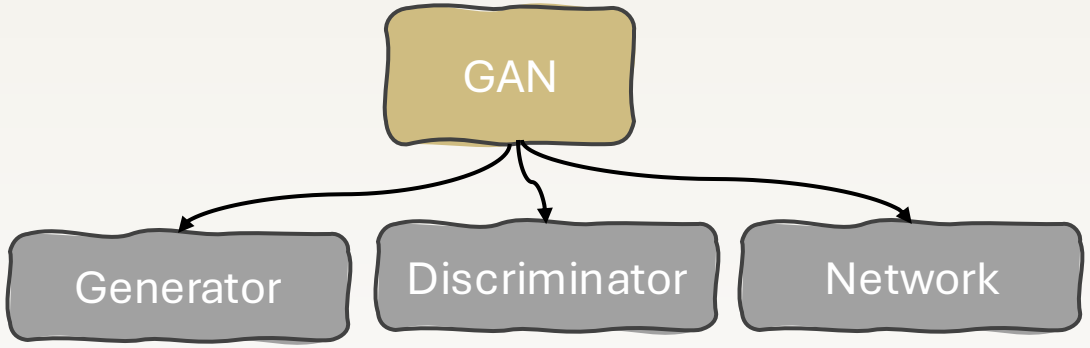
BeGAN



STEP 03

Proposed solution

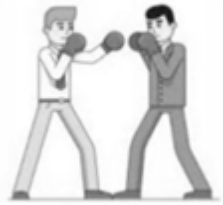
What Is GAN?

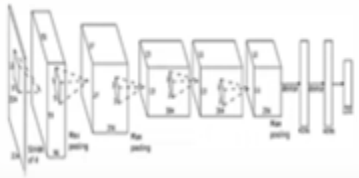


Generate data
Create fake images

Try to distinguish between real and fake images.

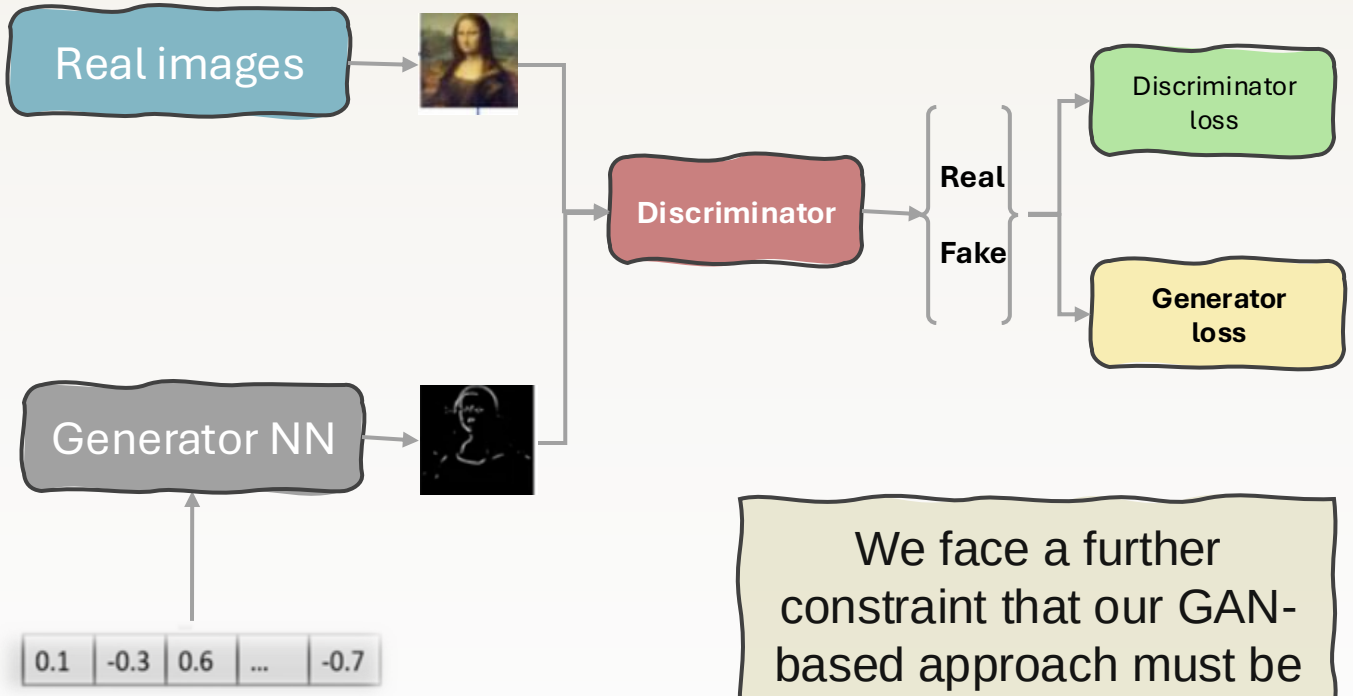
Network model, can be CNN.





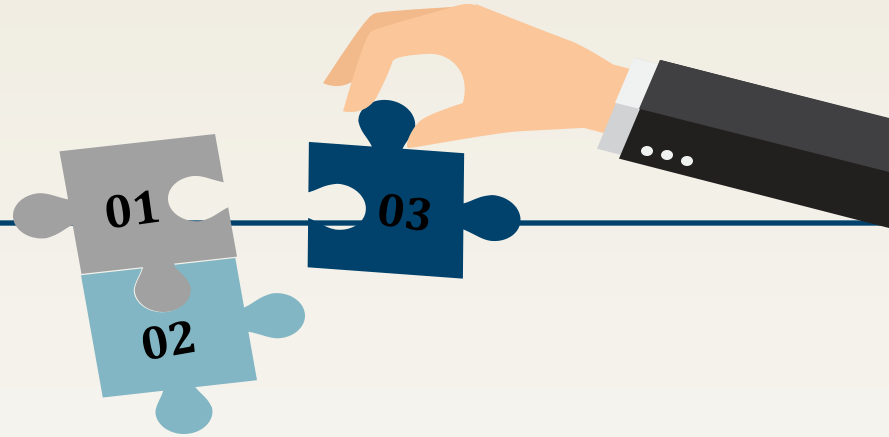
Funded by the European Union

How GAN works?



We face a further constraint that our GAN-based approach must be trained on limited data.

BeGAN steps



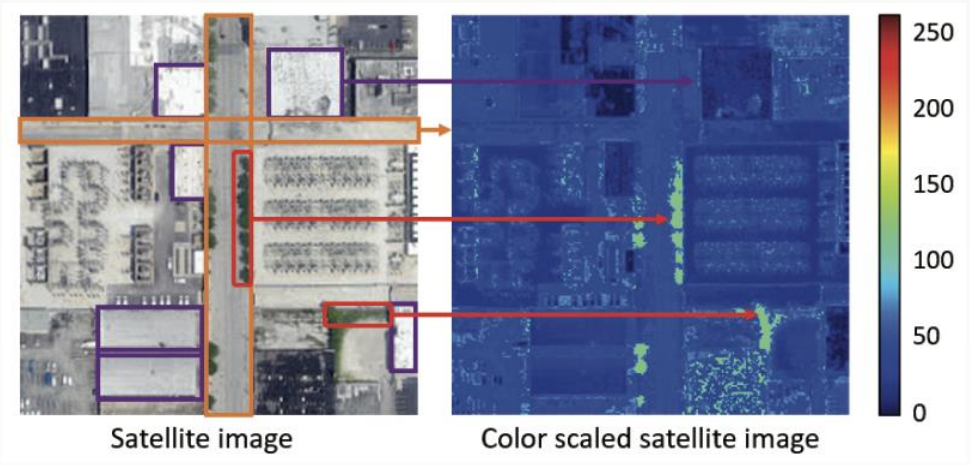
STEP 03

Proposed solution

How we can have enough data to train GAN model ?

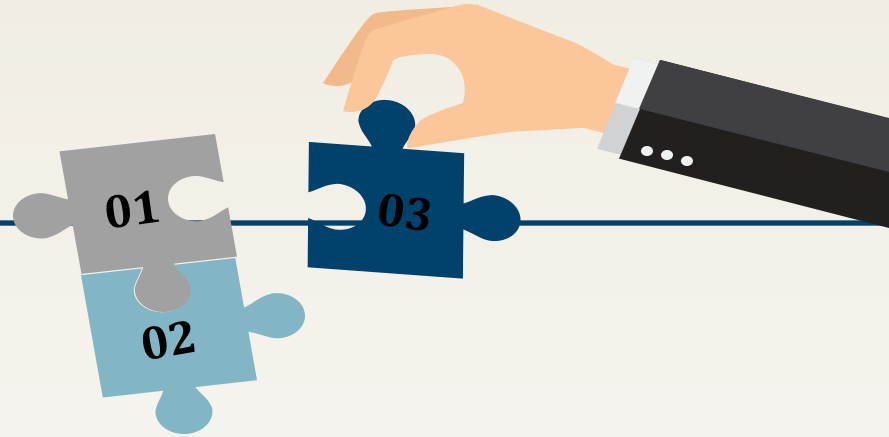
TL to training the GAN on a **source data** and then **fine-tuning** the model on the target datasets.
We aim to synthetically ***generate a set of source benchmarks*** that are sufficiently similar to the target datasets.

Source benchmarks:
A large dataset (24,000 datapoints) of urban satellite images from Google Maps, filtered down manually to 4,167 images.
Target datasets:
A small dataset of 10 circuit CMs.



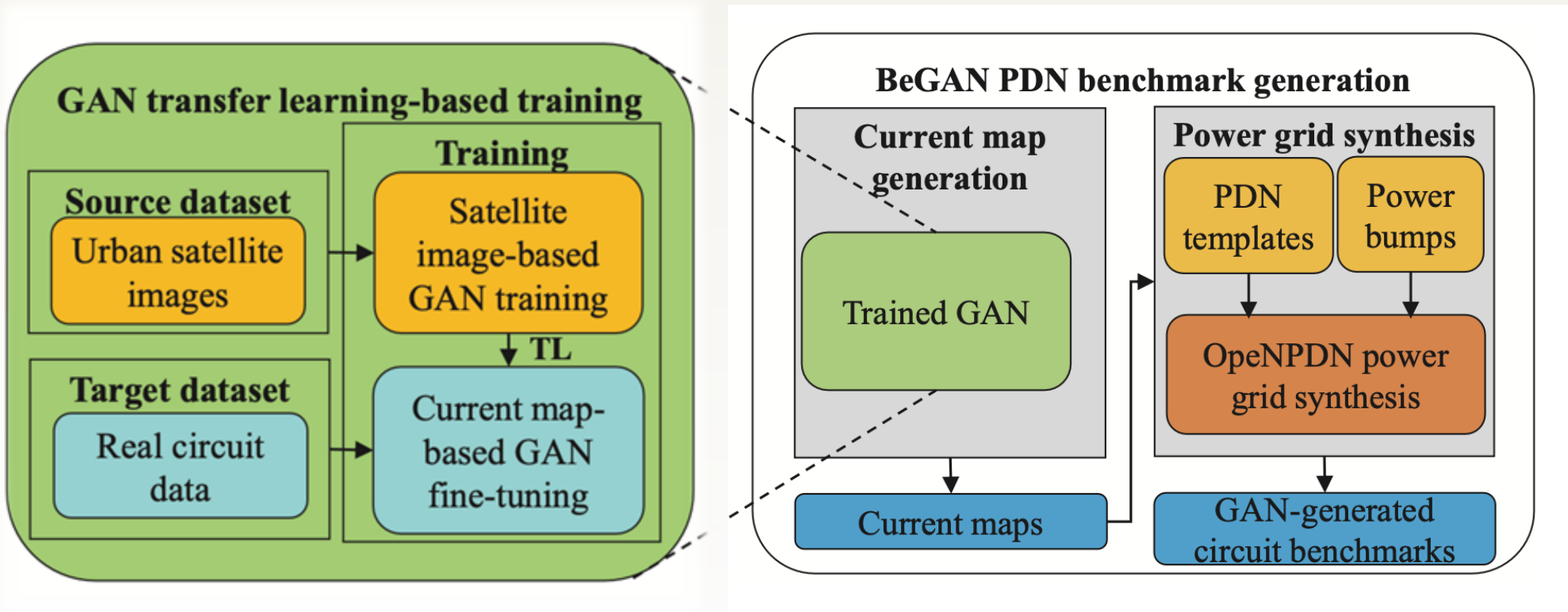
Funded by
the European Union

BeGAN



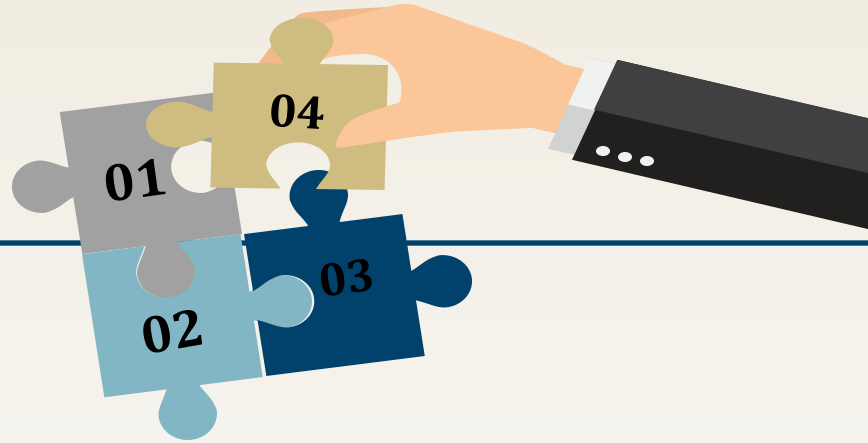
STEP 03

Proposed solution



Funded by the European Union

BeGAN



STEP 04

Results

The BeGAN methodology uses Generative Adversarial Networks (GANs) to:

✓ Overcome Data Scarcity:

Instead of requiring thousands of real circuits, BeGAN learns from a limited dataset and generates many realistic synthetic benchmarks.

✓ Enable Machine Learning in EDA:

BeGAN produces large, realistic datasets needed for training and testing ML-based EDA tools.

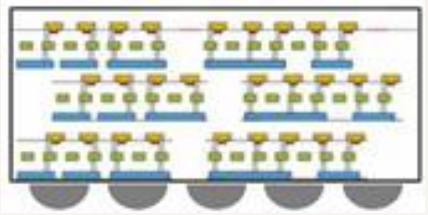
Funded by the European Union

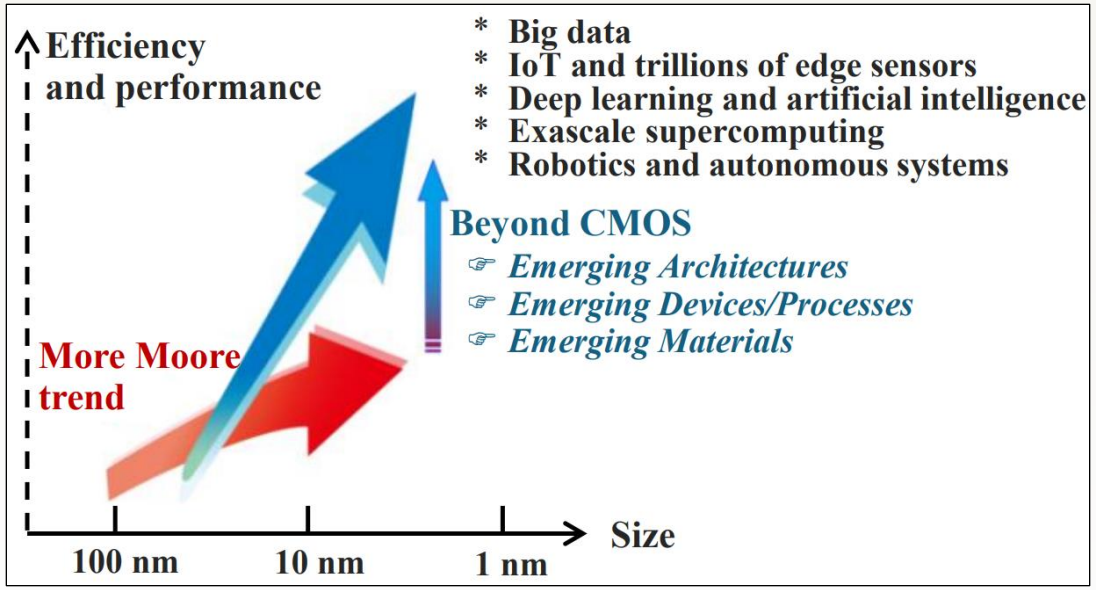
ML-Assisted Approach for Accelerated Thermal & IR Drop Analysis

Funded by
the European Union

The Challenges in VLSI Design

- Increasing design complexity
  - IoT (Medical, Automotive, …)
  - Miniaturization in VLSI design
  - Chiplet-to-Chiplet 3D ICs
- Criticality of hotspots and IR drops
  - Long-term reliability (e.g. EM)
- Traditional simulation bottlenecks…
- Time to Market



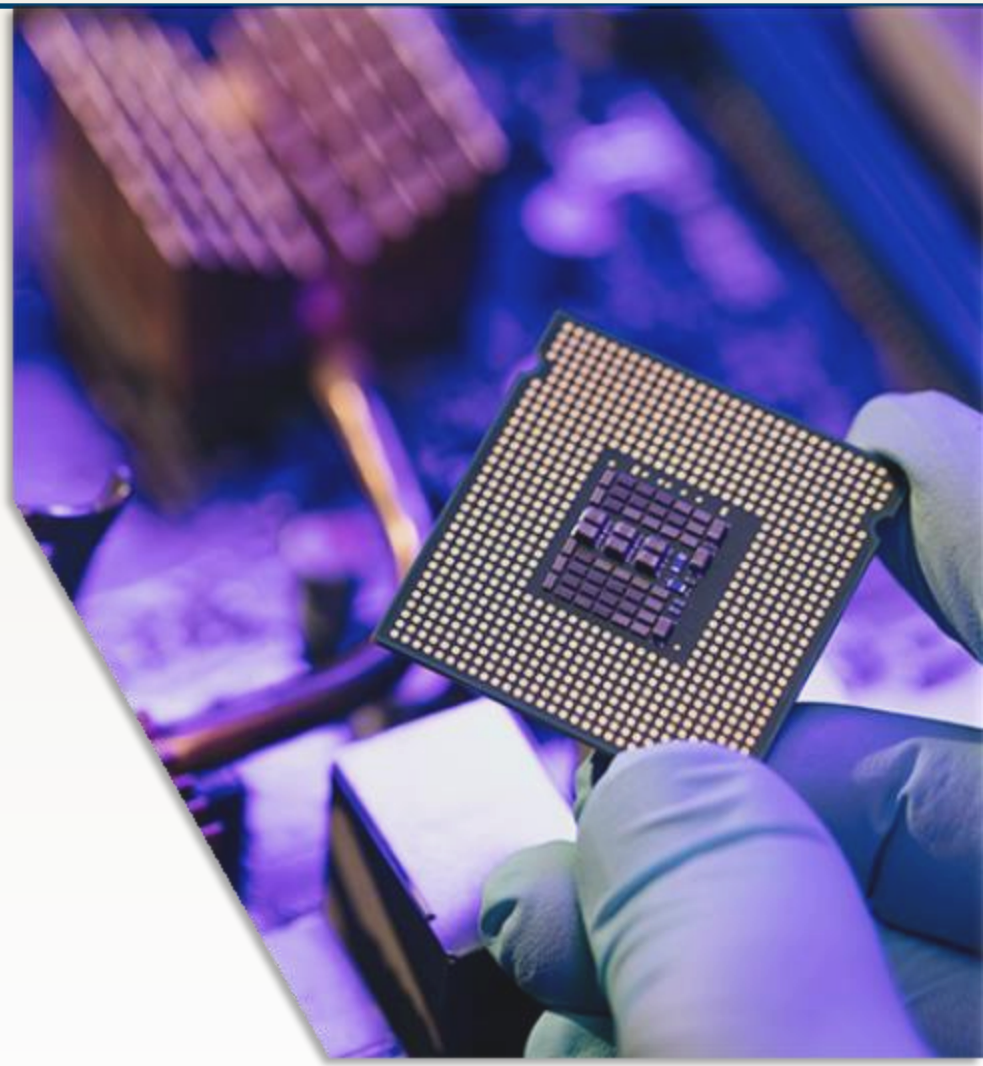


Source: IEEE IRDS 2023

Funded by
the European Union

Why Traditional Methods Fall Short

- Computational inefficiency
  - Methods like FEA and equivalent thermal RC network are computationally expensive
- Accuracy VS Scalability
  - Becomes difficult as chip sizes grow and designs become more complex
  - Simulations can take hours to complete…
- Iterative approaches
  - “*What if*” scenarios?



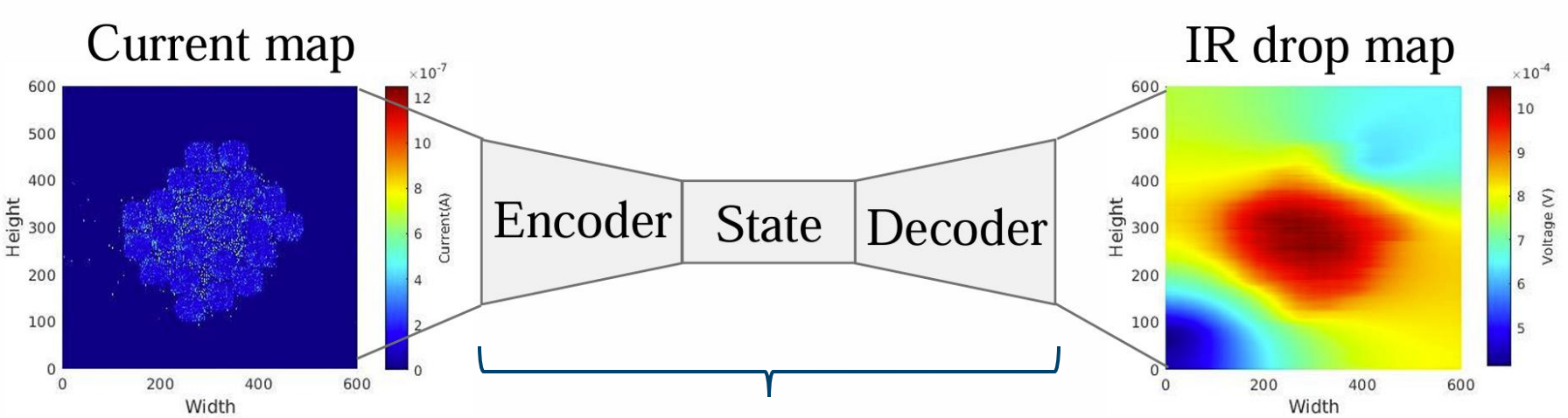
Funded by
the European Union

Transforming Chip Simulations with ML

Source: V. A. Chhabria, et. al. “Thermal and IR Drop Analysis Using Convolutional Encoder-Decoder Networks”, ASP-DAC, 2021.

Image-to-image or *sequence-to-sequence* translation

- ✓ Inference time in milliseconds
- ✓ Competitive results w.r.t. commercial tools



Convolutional Encoder-Decoder networks (e.g., U-Net, ConvLSTM)

Funded by the European Union

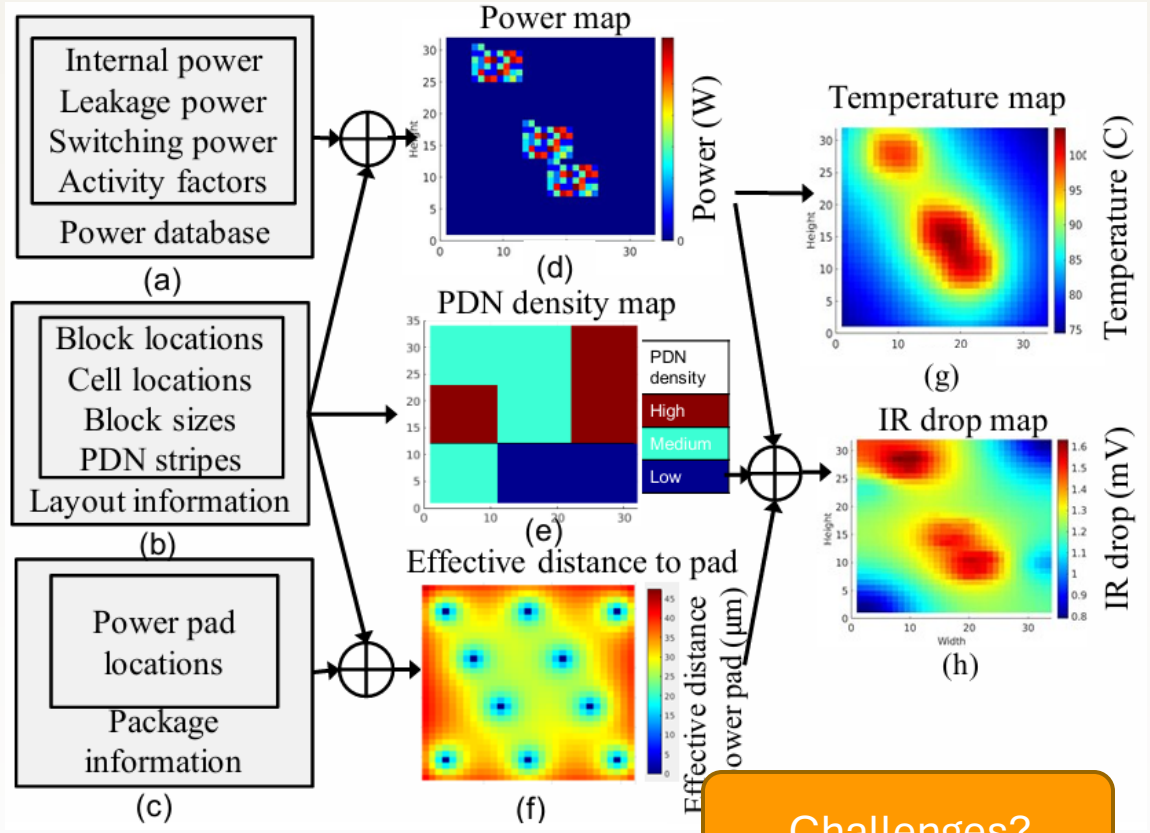
Thermal and IR Drop Analysis using ML

Source: V. A. Chhabria, et. al. “Thermal and IR Drop Analysis Using Convolutional Encoder-Decoder Networks”, ASP-DAC, 2021.

- Inputs:
  - Power distribution (d), PDN density (e), effect. distance to power pads (f)
- Outputs:
  - High resolution thermal maps (g) and IR drop contours (h)

✓ Reduces analysis bottlenecks

✓ Generalizable across designs with same resolution and technology node

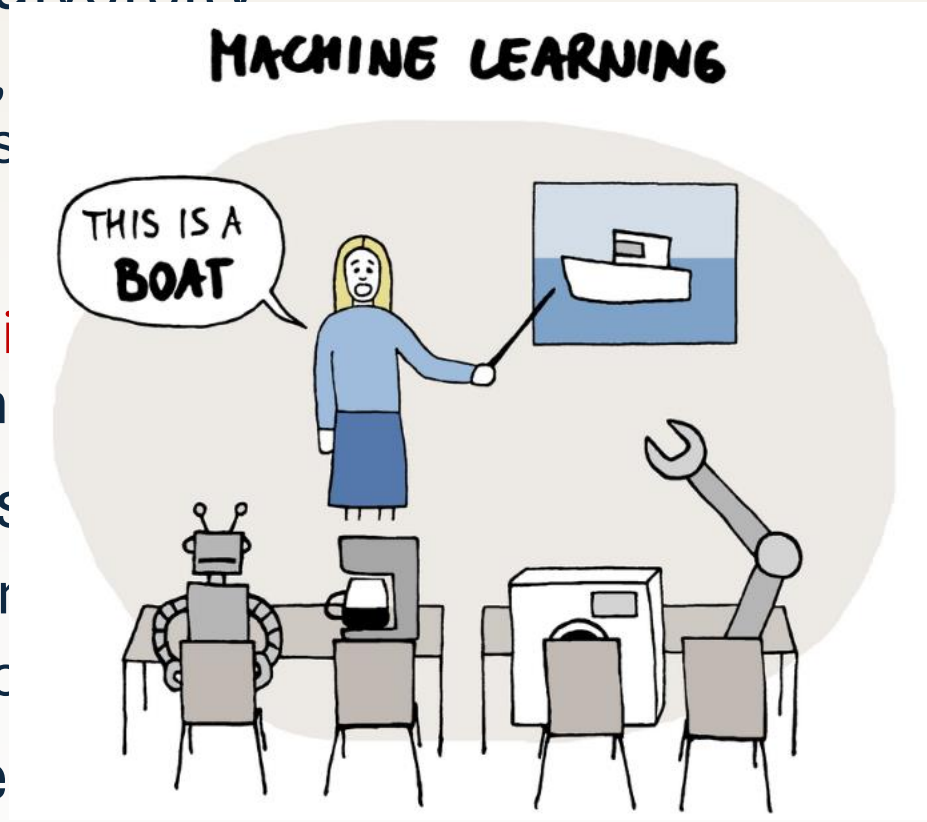


Challenges?

Funded by the European Union

Shaping the Future of ML in EDA

- Enhance dataset diversity
  - Robust, scalable, cross different technology nodes
- Democratizing Ha
  - Fostering innovati the EDA commun
- Improve AI models
  - Transferability acr s
  - Expand applicatic
- (Further) integrate



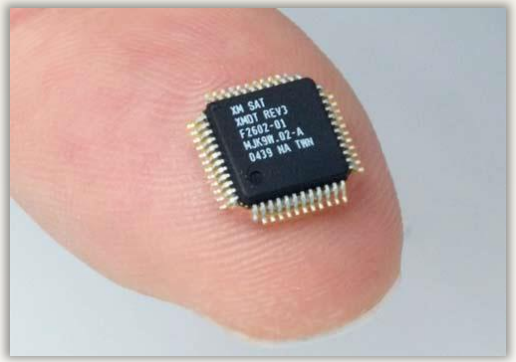
> **OpenROAD**
>
> **Democratizing Hardware Design**
>
> The OpenROAD™ project attacks the barriers of Cost, Expertise and Uncertainty (i.e., Risk) that block the feasibility of hardware design in advanced technologies.

Funded by the European Union

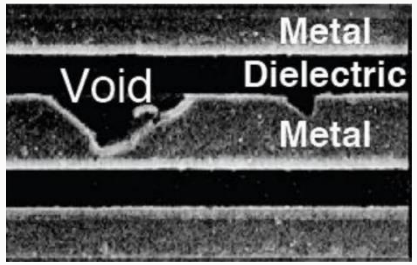
Swarm Intelligence for Electromigration-Resilient Design

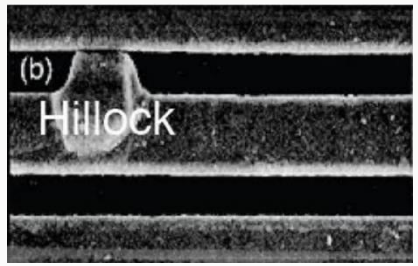
Funded by
the European Union

Long-Term Reliability of Power Grids



- Miniaturization in VLSI design
  - Higher current densities exacerbate reliability issues
- Phenomenon of Electromigration (EM)
  - Causes migration of metal atoms, resulting in voids (tensile stress), hillocks (compressive stress) and ultimately open circuits
  - Performance issues like IR drop and signal degradation
- The need for smarter solutions in earlier stages
  - Replace the conventional empirical models





Funded by
the European Union

Challenges in EM-Aware Design Optimization

Problem #1

Black's model for EM

$$MTTF = \frac{A}{j^n} e^{\frac{E_a}{k_b T}}$$

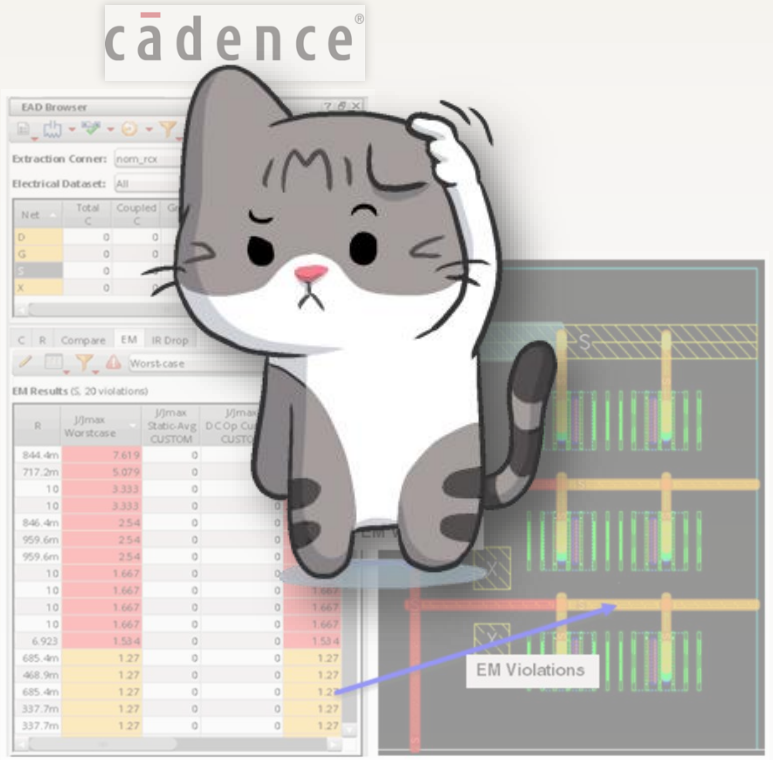
Given the mission profile

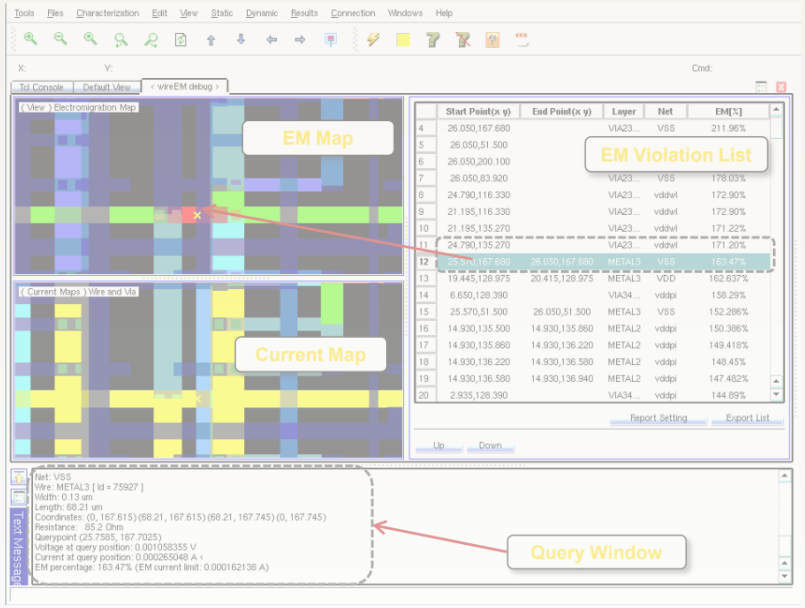
J_{max}

The **maximum allowed current dens.** in the design

Korhonen's model:
Physics-based PDE for EM stress

Problem #2

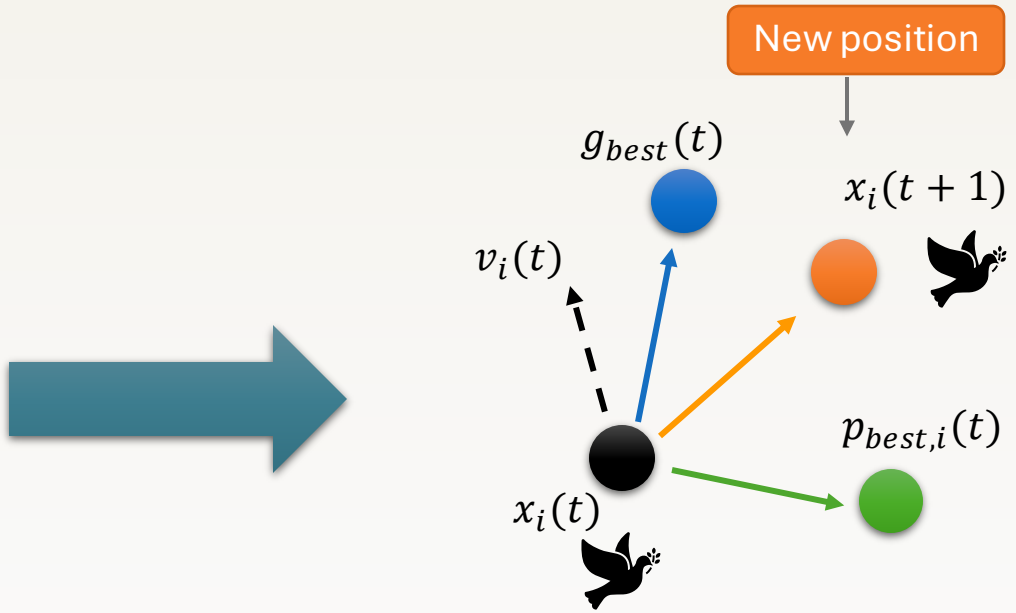




Funded by the European Union

Swarm Intelligence for EM-Resilient Design

Source: O. Axelou, et. al. “An Electromigration-Aware Wire Sizing Methodology via Particle Swarm Optimization”, GLS-VLSI, 2024.



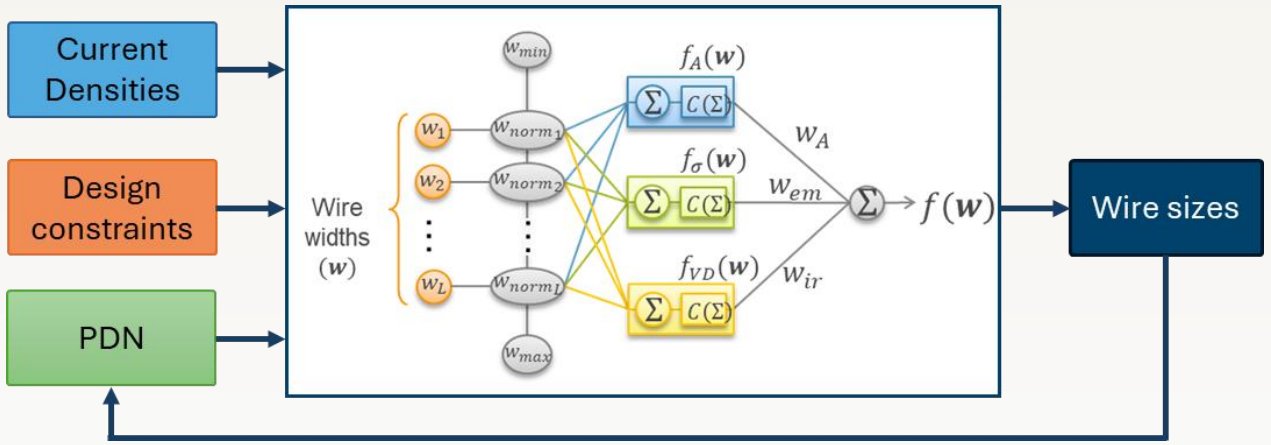
✓ Well-suited for high-dimensional, non-linear design spaces

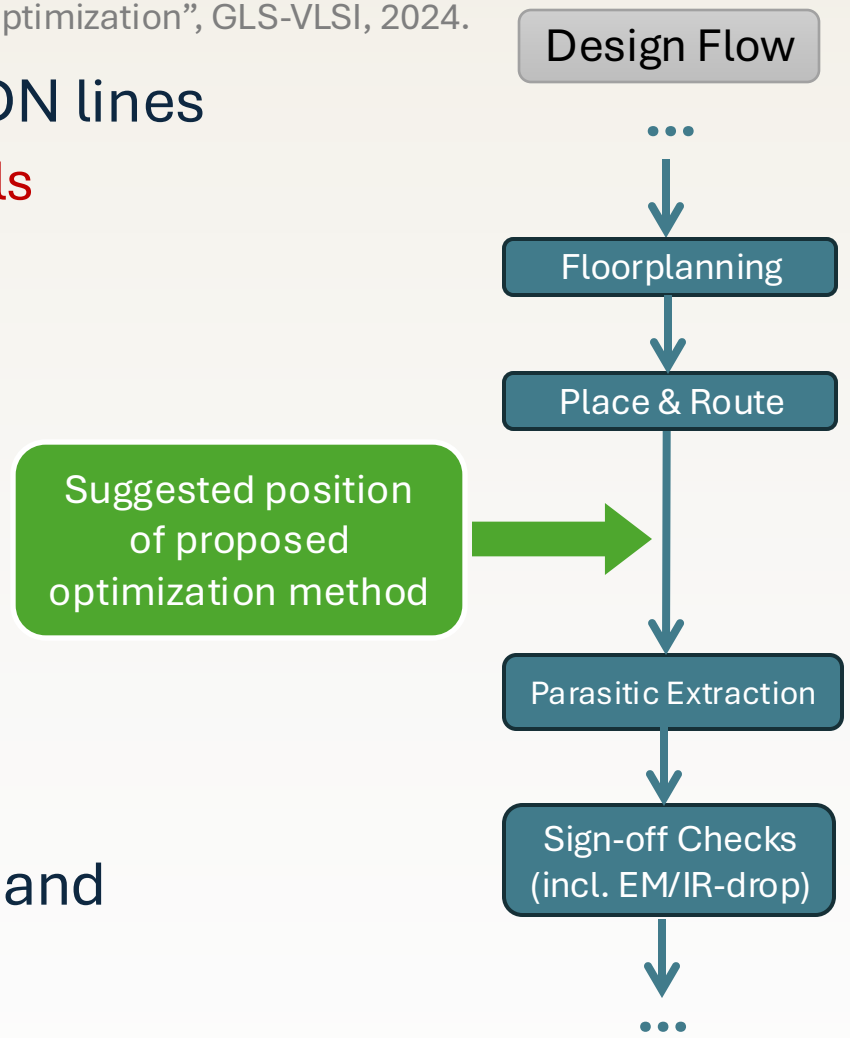
Funded by
the European Union

Incorporating Swarm Intelligence into Design Flow

Source: O. Axelou, et. al. “An Electromigration-Aware Wire Sizing Methodology via Particle Swarm Optimization”, GLS-VLSI, 2024.

- Concept: Proactive EM/IR-Drop optimization of PDN lines
  - Incorporates highly accurate physics-based models





- Inputs: PDN, current densities, design constraints
- Outputs: Optimized wire sizes for better reliability and minimal area expansion

Funded by the European Union

Next Steps in EM-Resilient Design

- Scalable solutions for future technology nodes
  - Need for more publicly available power grid benchmarks
- Industry Adoption
  - Create user-friendly plugins for existing EDA tools
  - Standardized workflows into early design phases
- Foster collaboration
  - Build interdisciplinary teams across Academia, Industry, and ML scientists

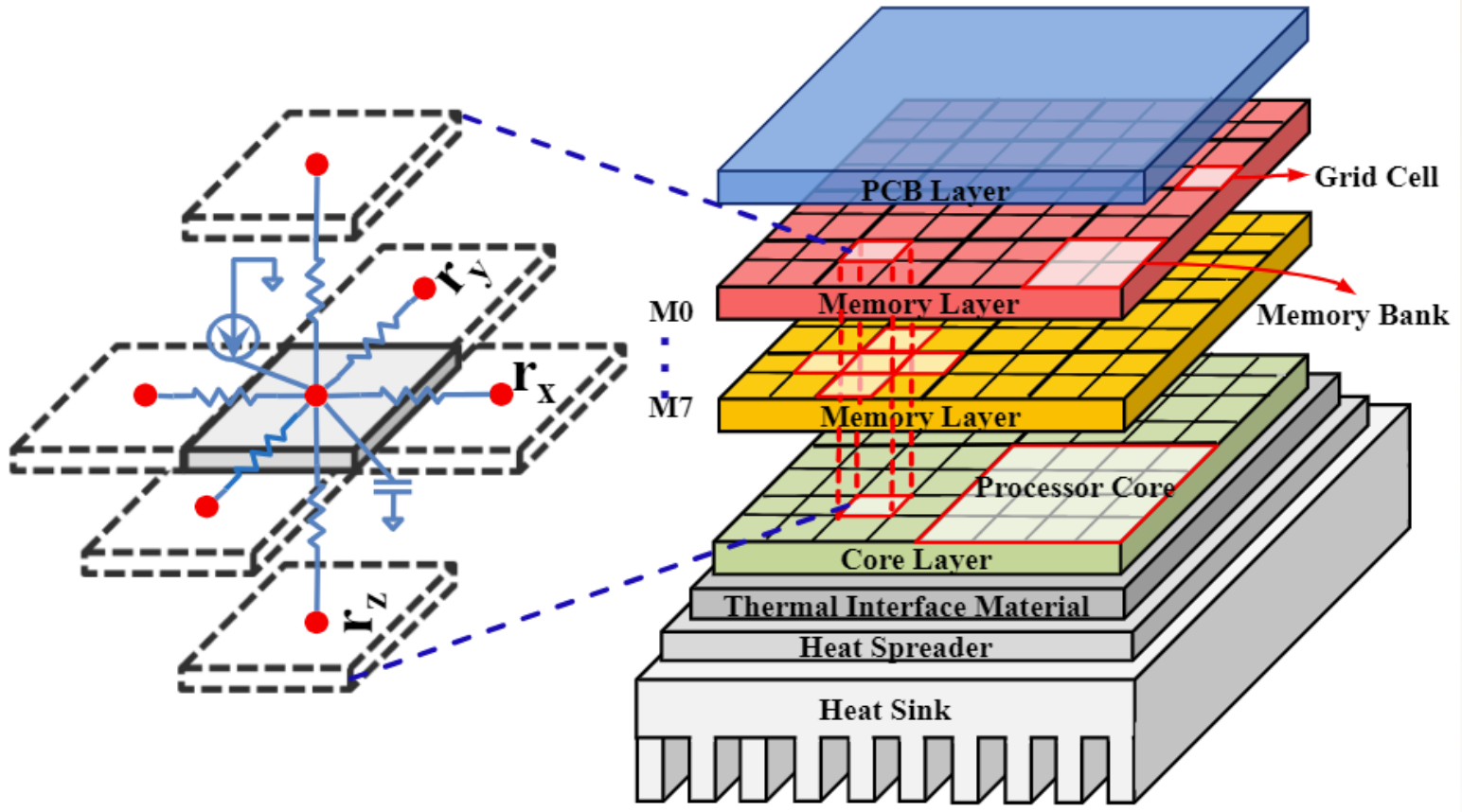
Funded by
the European Union

Unified thermal scheduler for DNN in 3D-stacked systems

Funded by
the European Union

Research background→3D-stacked systems

❑ 3D-stacked Processor-Memory Architecture

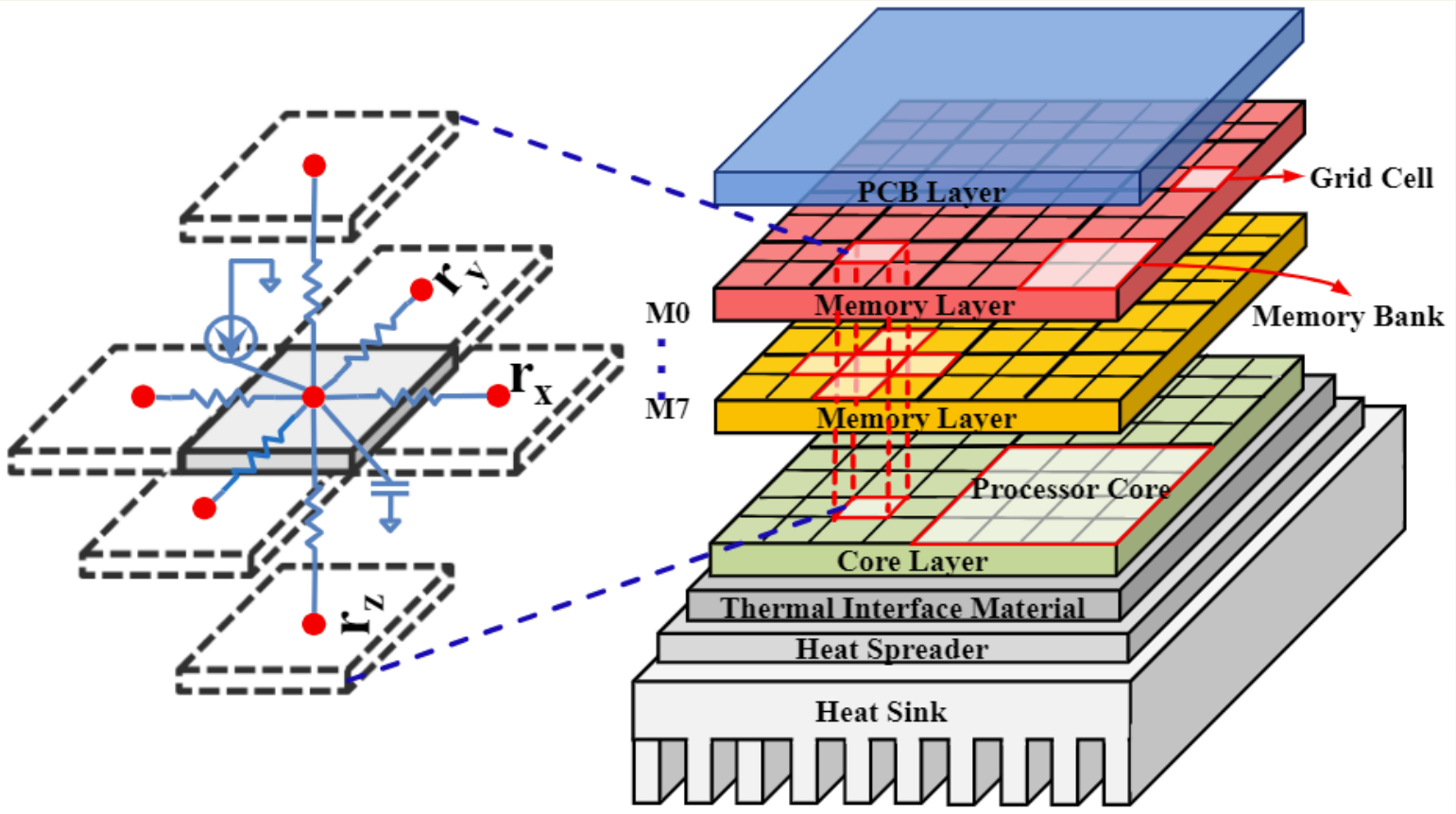


- Allow Heterogenous Integration
- Stacked (Logic/DARM) Dies
  - TSV(Through Silicon Via) interconnection
  - More compact material
- Higher Bandwidth via TSV
- Lower Latency

Funded by the European Union

Research background→3D-stacked systems

❑ 3D-stacked Processor-Memory Architecture

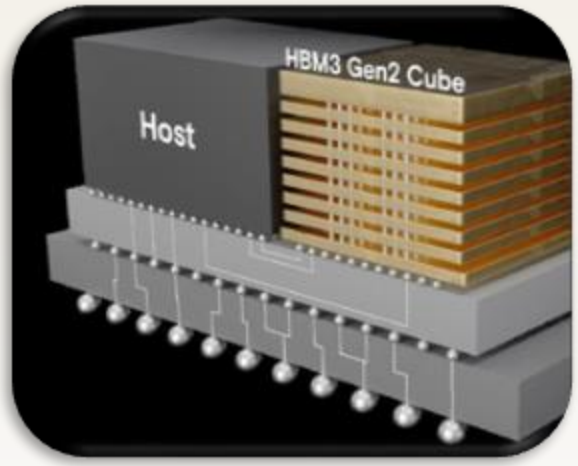


- Higher cost
  - Simulation needed
  - CoMeT
- Higher power density

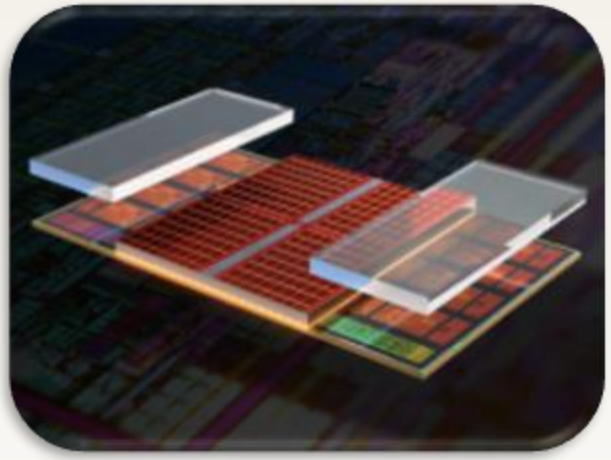
Funded by the European Union

Research background→3D-stacked systems

- 3D-stacked chips in industry

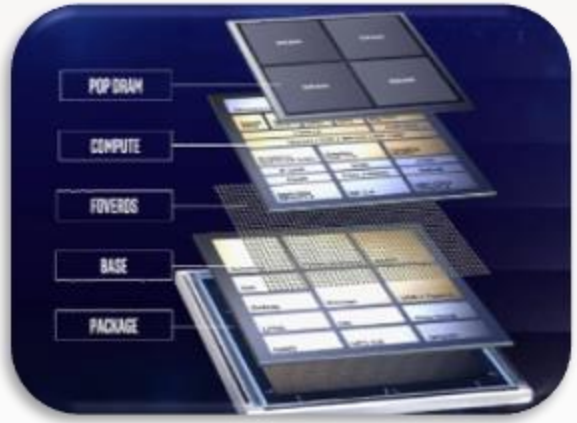


Micron HBM

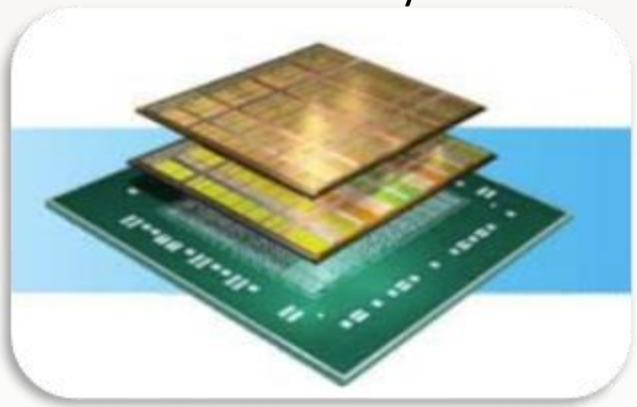


AMD Zen 3 Ryzen

Apple M2



Intel Lakefield



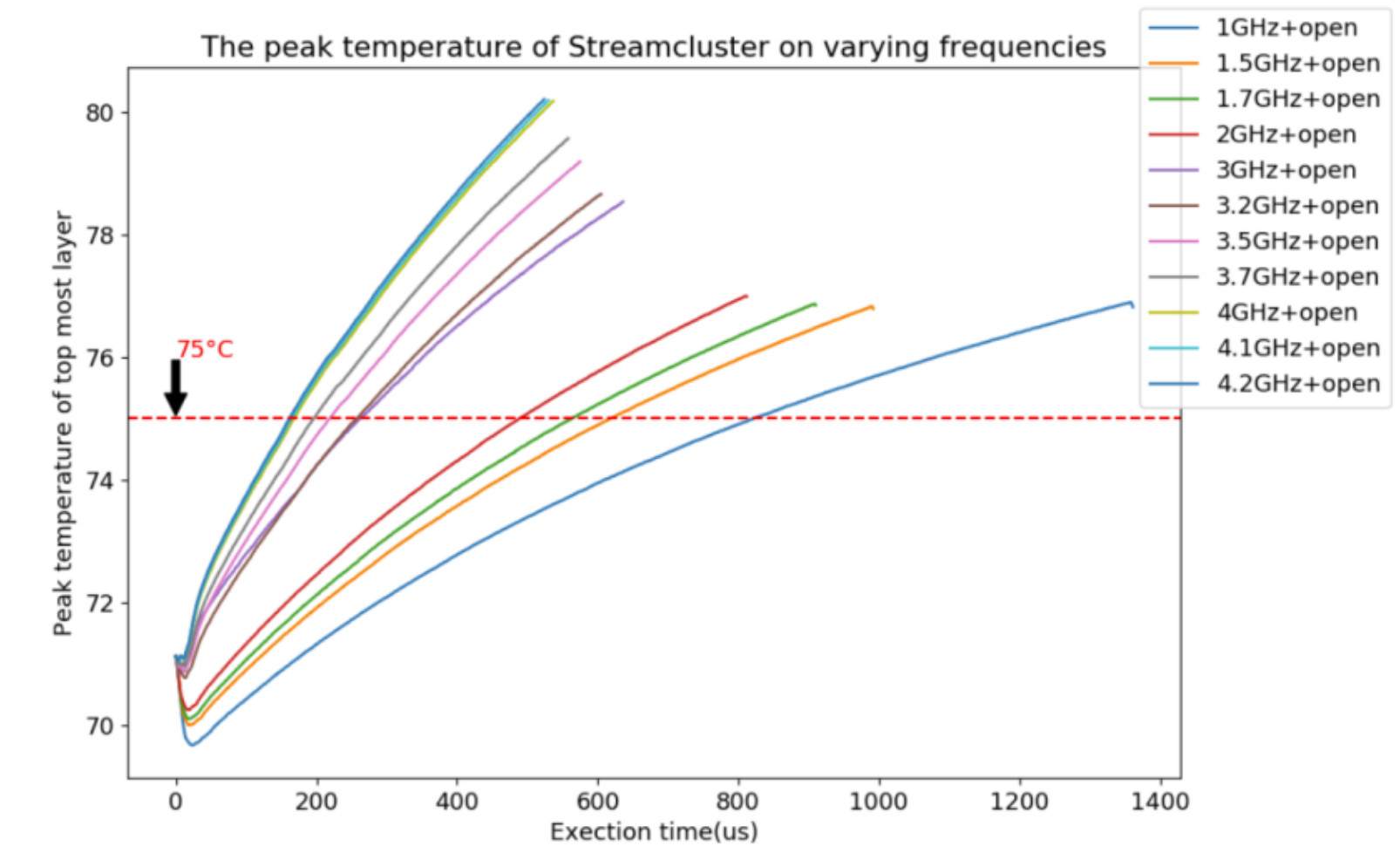
Xilinx Virtex Ultrascale

Samsung HBM2E Flashbolt

Funded by
the European Union

Research background→3D-stacked systems

❑ 3D-stacked Processor-Memory Architecture

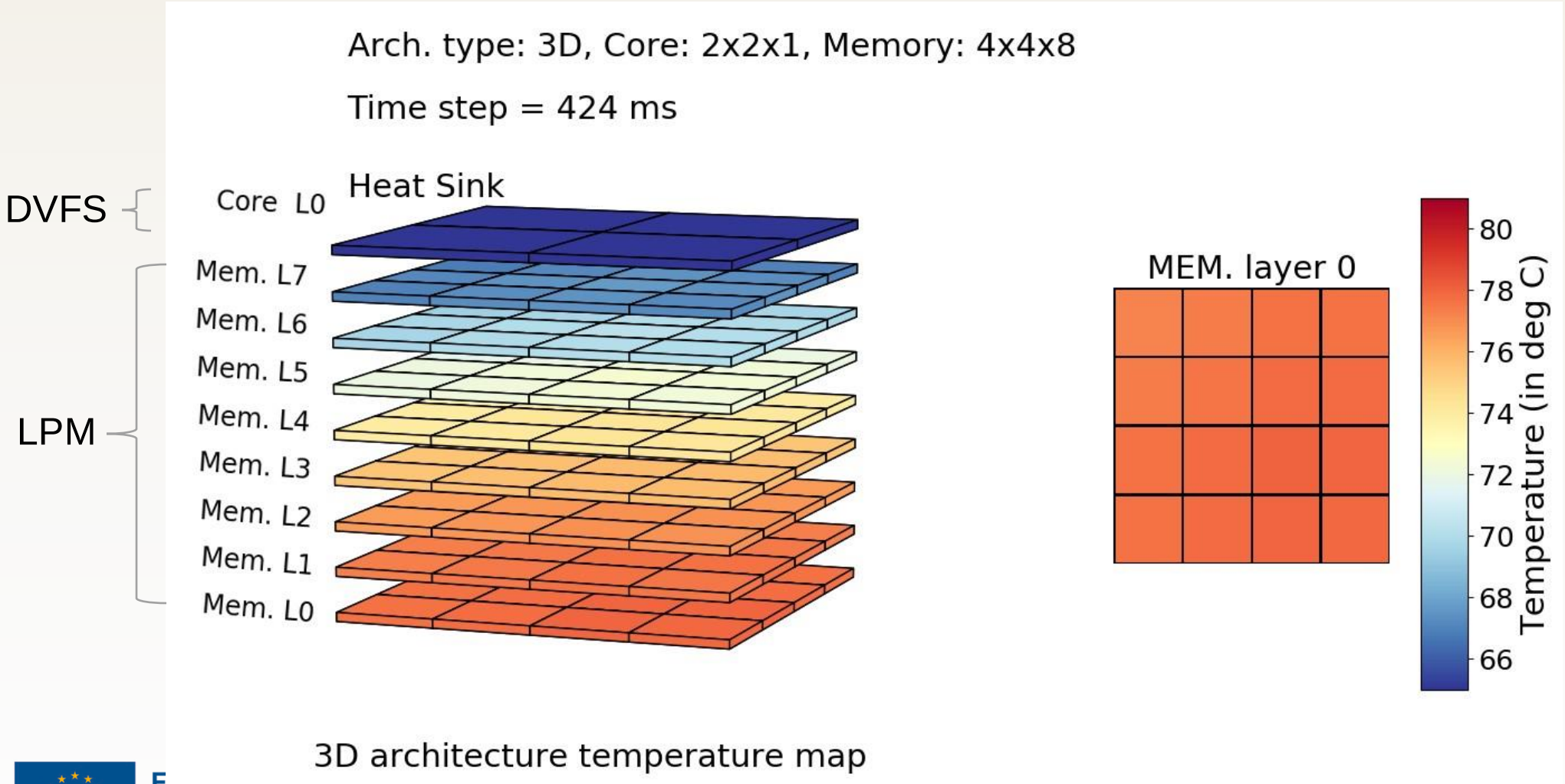


- Higher cost
  - Simulation needed
  - CoMeT
- Higher power density

Funded by
the European Union

Research background→3D-stacked systems

- ❑ Thermal behavior in 3D-stacked systems(running blacksholes)

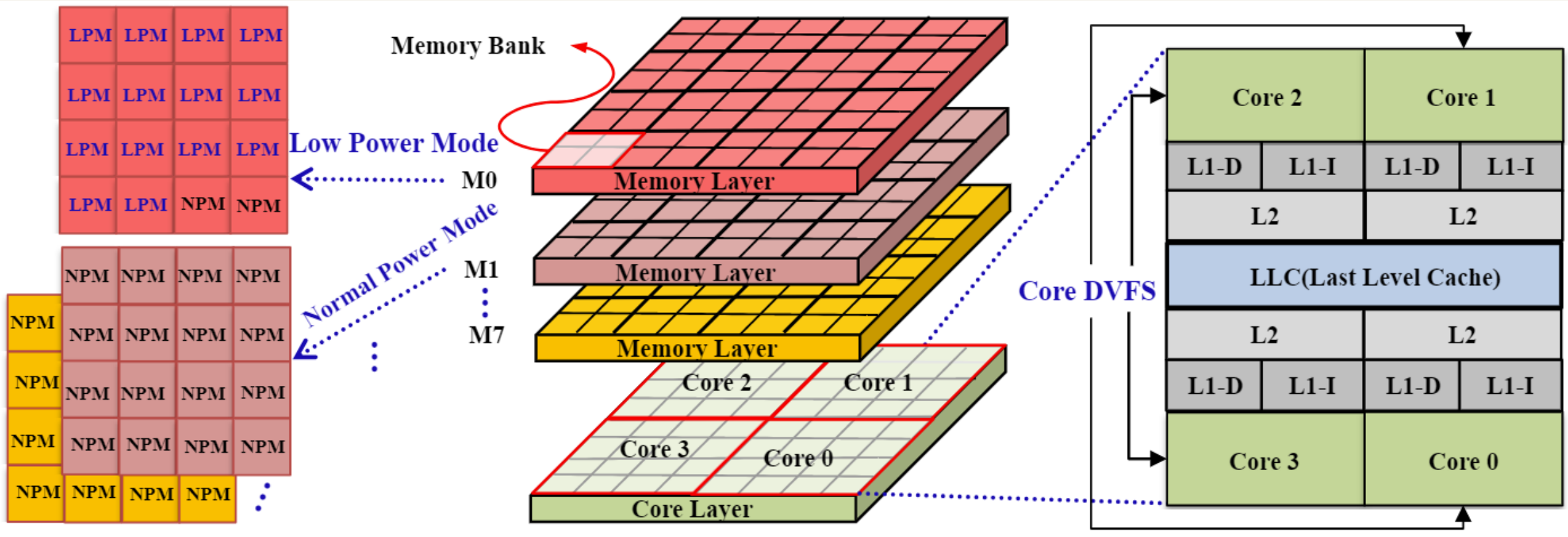


3D architecture temperature map

Funded by the European Union

A motivational example

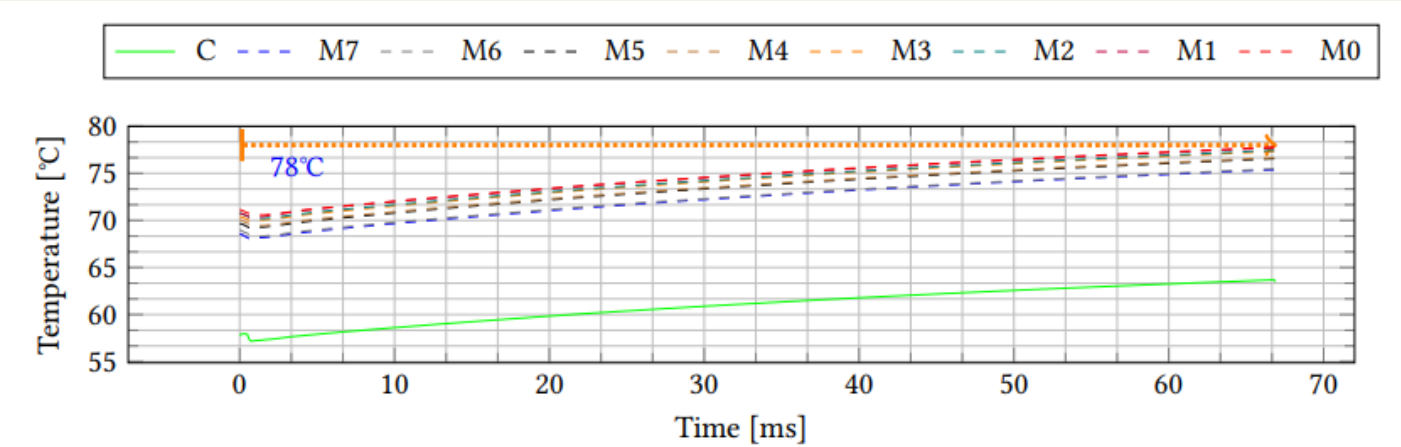
- 3D-stacked processor-memory systems with DVFS and LPM



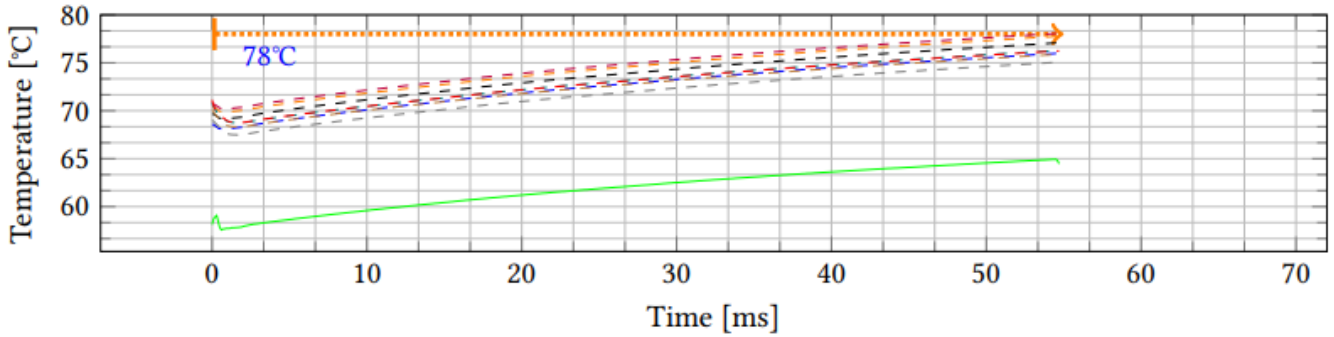
Funded by
the European Union

A motivational example

- Layer-wised peak temperature for different layers



(a) 1st Scenario: 2.7 GHz core(s) frequency with 0 memory banks in LPM and 128 memory banks in NPM.



(b) 2nd Scenario: 4 GHz core(s) frequency with 14 memory banks in LPM and 114 memory banks in NPM.

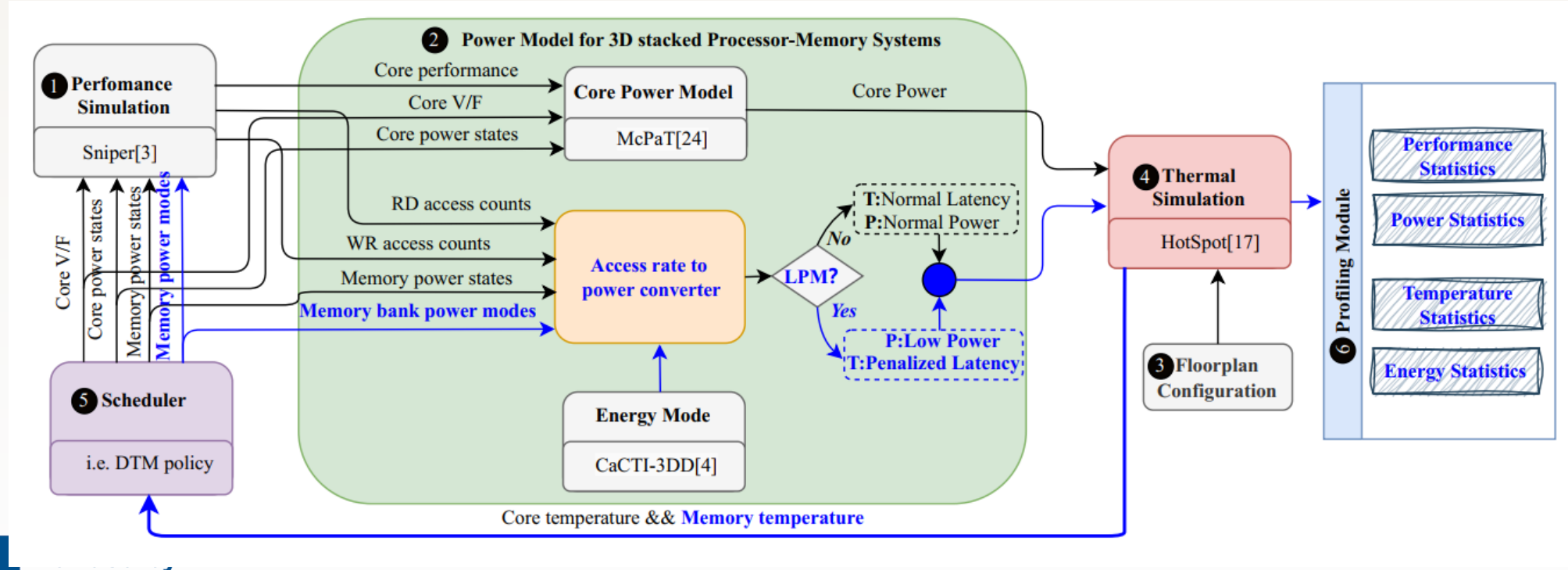
- LPM+4GHz is 18.33% faster than No LPM+2.7GHz
- Computer-intensive benchmarks benefit more from higher core frequency than from LPM penalties

Funded by the European Union

Low power mode(LPM) implementation

❑ Key features of LPM

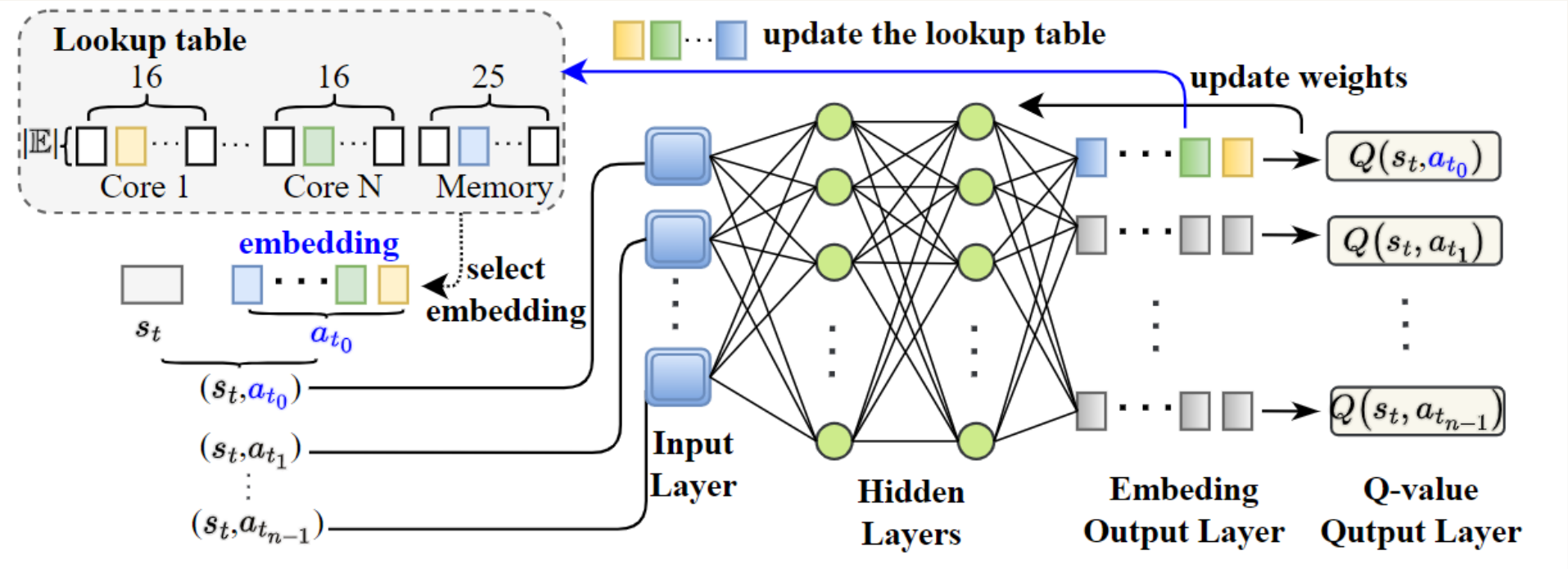
- Data preserved in situ in LPM, avoiding migration
- Data inaccessible in LPM mode
- Re-accessing data requires toggling to NPM



3QUTM: a unified thermal scheduler via Deep Q-learning

❑ The design of the Q-network

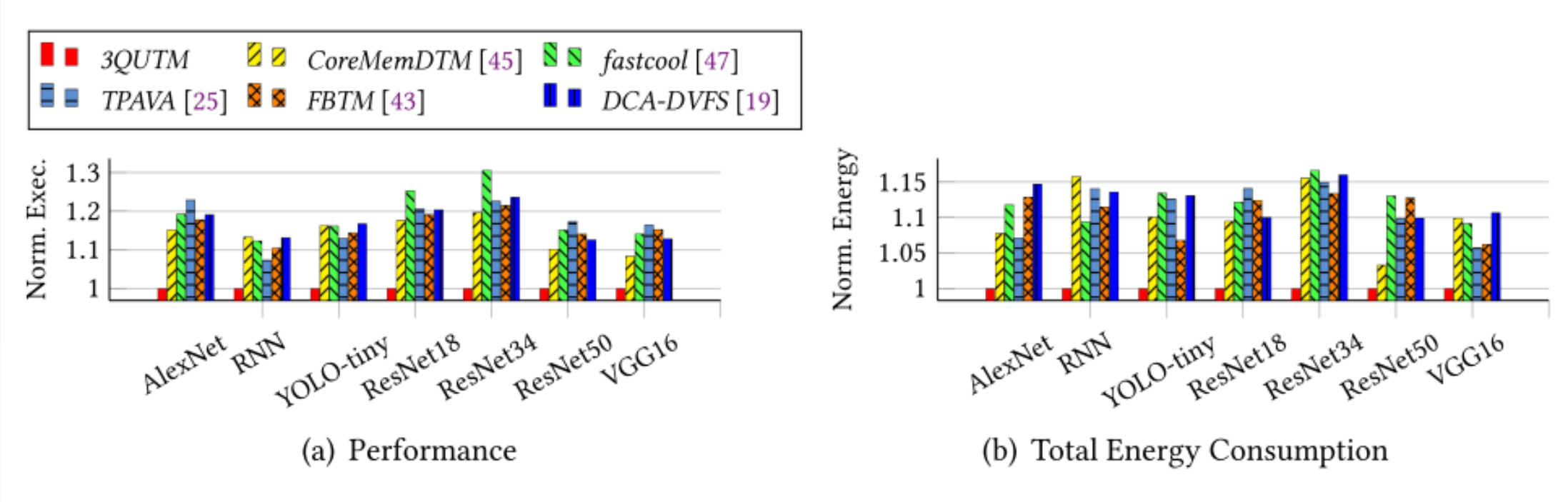
- Parameterize each subaction(each core and each DRAM action)
- Embedding the parameterized subaction



Funded by the European Union

Experimental results

- Performance and Energy Consumption for DNN Workloads in Inference



Funded by
the European Union

Q&A

Funded by
the European Union